nature materials

Perspective

https://doi.org/10.1038/s41563-025-02272-0

Accelerated data-driven materials science with the Materials Project

Received: 13 June 2024

Accepted: 16 May 2025

Published online: 03 July 2025

Check for updates

Matthew K. Horton[1,2], Patrick Huck[1], Ruo Xi Yang[1], Jason M. Munro[1], Shyam Dwaraknath[1], Alex M. Ganose[3], Ryan S. Kingsbury[4,5], Mingjian Wen[6], Jimmy X. Shen[7], Tyler S. Mathis[1], Aaron D. Kaplan[1], Karlo Berket[8], Janosh Riebesell[1,9], Janine George[10,11], Andrew S. Rosen[1,2], Evan W. C. Spotte-Smith[12], Matthew J. McDermott[1], Orion A. Cohen[1,2], Alex Dunn[13], Matthew C. Kuner[1,2], Gian-Marco Rignanese[14], Guido Petretto[15], David Waroquiers[15], Sinead M. Griffin[1,16], Jeffrey B. Neaton[1,17], Daryl C. Chrzan[1,2], Mark Asta[1,2], Geoffroy Hautier[18], Shreyas Cholia[8], Gerbrand Ceder[1,2], Shyue Ping Ong[19], Anubhav Jain[13] & Kristin A. Persson[1,2]

The Materials Project was launched formally in 2011 to drive materials discovery forwards through high-throughput computation and open data. More than a decade later, the Materials Project has become an indispensable tool used by more than 600,000 materials researchers around the world. This Perspective describes how the Materials Project, as a data platform and a software ecosystem, has helped to shape research in data-driven materials science. We cover how sustainable software and computational methods have accelerated materials design while becoming more open source and collaborative in nature. Next, we present cases where the Materials Project was used to understand and discover functional materials. We then describe our efforts to meet the needs of an expanding user base, through technical infrastructure updates ranging from data architecture and cloud resources to interactive web applications. Finally, we discuss opportunities to better aid the research community, with the vision that more accessible and easy-to-understand materials data will result in democratized materials knowledge and an increasingly collaborative community.

In 2011, the Materials Project (MP) was launched to accelerate materials discovery by leveraging open science principles, such as open data, source code and collaboration, aided by advances in computational power and first-principles methodology. The vision culminated in the MP website and application programming interface (API) that offer freely accessible, easy-to-understand property data for individual materials and bulk downloads for large datasets, empowering scientists to make better informed decisions and advance materials science research. A decade later, the MP has evolved into a comprehensive repository of materials data, encompassing a diverse array of properties calculated for various materials chemistries and structures. This has led to an impressive growth, eclipsing 600,000 registered users and growing exponentially (see top of Fig. 1). The MP is now a widely recognized and routinely used platform in various materials science domains, alongside other databases such as NOMAD[1], OQMD[2,3], AFLOW[4,5], JARVIS[6] and Materials Cloud[7], where each offers a diverse range of structures, properties and software tools[8]. This shift signifies a broader transition towards data-driven science, where computation and machine learning (ML) provide opportunities for further acceleration.

A full list of affiliations appears at the end of the paper. e-mail: kapersson@lbl.gov

In this Perspective, we discuss the main advances in the MP over the past decade, contextualized within the materials science community, and we outline the remaining challenges and the philosophy with which the MP hopes to face them.

The MP database

The primary output of the MP is its database of inorganic crystal structures and their associated properties. In addition, it hosts databases of molecules[9–11], synthesis recipes extracted from the literature[12–14], the properties of battery materials[15–18] and metal–organic frameworks[19], and catalysis datasets[20] that are developed in collaboration with the MP or that have been developed externally.

The main MP database can be conceptualized along two axes: breadth and depth (see bottom of Fig. 1). Breadth reflects the number of distinct materials, whereas depth indicates the range of properties and metadata known for each material. Each axis presents unique challenges and benefits. Over the past decade, the depth of the MP database has increased substantially through the implementation of automated workflows. Every workflow and the data produced undergoes rigorous benchmarking and validation against available experimental data, and is typically documented in a peer-reviewed publication. These workflows enable the MP to autonomously and continuously calculate specific properties across various structures and chemistries.

Table 1 and Fig. 1 summarize the properties that have been calculated via the MP in high throughput since its inception. These include both additional properties and enhancements to existing capabilities as methods mature. State-of-the-art first-principles methodology is constantly evolving and necessitates that properties are recalculated and updated as more accurate methods that are compatible with high-throughput workflows emerge. For example, improvements in the predicted thermodynamic stability, particularly the formation enthalpy, were achieved through empirical correction schemes and the adoption of the r^{2}SCAN density functional[21–23]. This continued assessment and evaluation against experimental gold standards is vital for ensuring data integrity, user trust and remaining at the forefront of materials informatics.

The breadth of the MP refers to its range of unique crystal structures, termed 'materials' within the MP, while acknowledging that the scope of real materials within materials science is broader. Today, it covers a total of 178,627 materials across 51,298 chemical systems and 228 space groups throughout the majority of the elemental space (Fig. 1). The MP's structure and compound coverage has drawn from established experimental databases such as the Inorganic Crystal Structure Database (ICSD)[24], the Pauling File[25] and the Crystallography Open Database[26]. As a result, the MP has been able to compute the majority of unique stoichiometric inorganic materials found in the ICSD. Furthermore, owing to the vastness of the literature, text mining using natural language processing has helped to identify compositions not recorded in the ICSD[27].

By seeking breadth, the MP does not specialize in specific categories of materials, instead emphasizing properties and characteristics across diverse chemical systems and structures. Specialized materials databases, such as those for perovskites[28], organic crystals[29] or the subsets of the Computational Materials Repository[30], may better serve specific applications. Yet, not all possible materials have been synthesized, rendering incomplete any database that relies solely on synthesized materials. Databases that include a large proportion of hypothetical materials, such as AFLOW[4] and Alexandria[31], can aid in mapping the energy landscape of polymorphs.

Thus, innovative strategies are needed to broaden the database into unexplored chemical spaces. To this end, the MP has applied structure-prediction techniques[32] using data-mined statistical analysis to provide likely ionic substitutions for existing materials prototypes. Similar efforts from AFLOW include the release of materials prototypes containing stoichiometry and geometries[33]. Since then, ML has transformed structure searching through techniques such as graph neural networks and active learning[34]. These methods, often trained on accurate and diverse materials data from the MP, underscore the importance of curated and accessible data for rapid learning.

The second aspect of expanding breadth is increasing the coverage of disordered and non-stoichiometric compounds motivated by enhancing the representations of real materials that are often disordered and imperfect. As the underlying density functional theory (DFT) methods used by the MP require fully ordered unit cells, many experimentally known disordered compounds require the construction of ordered approximations before they can be added to the MP. This process has been used continuously for many disordered materials, but selecting an ordered approximation is not unique and may overlook crucial details such as solid solubility. Structures predicted by ML methods are often approximated into ordered forms[35]; this may not match experimental observations of a material synthesized at high temperatures where multiple disordered chemical arrangements may coexist. Thus, the MP is establishing methods to classify these ordered approximations[36], for better database construction and enhanced simulation accuracy through cluster expansion techniques[37]. Effective stochastic methods that can simulate disorder in large-scale systems and correlate them to the as-synthesized disordered materials will greatly advance the gap between approximated theoretical models and experiments. With these limitations in mind, ordered approximations are still useful for explaining average behaviour, short-range orders and associated energetics.

Another critical aspect for translating materials informatics across simulation codes is simulation metadata[38]. From the start, the MP has made substantial infrastructure efforts to preserve structured simulation metadata wherever possible. For example, calculation metadata are available both on the MP website and via programmatic download. Retaining these metadata—for example, the charge densities, structure provenance and relaxation trajectories—has enabled continuous data validation against the latest methods and current development of universal ML potentials[39,40].

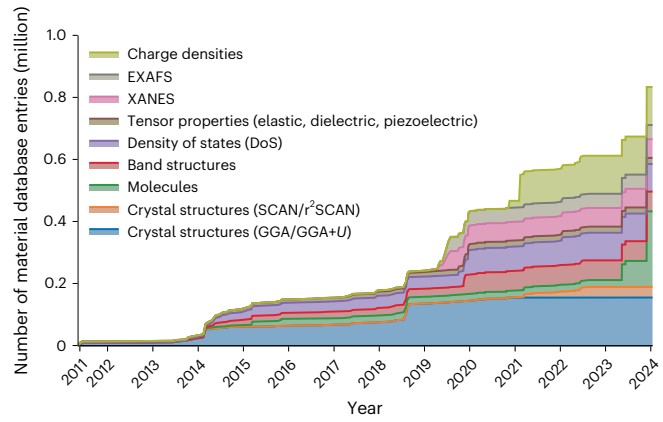
Despite the increase in throughput enabled by workflow tools, data production faces a human rate limit as automated workflows still require considerable intervention. Recognizing its limited personnel resources, the MP emphasizes expanding coverage in both breadth and depth through active engagement with the community.

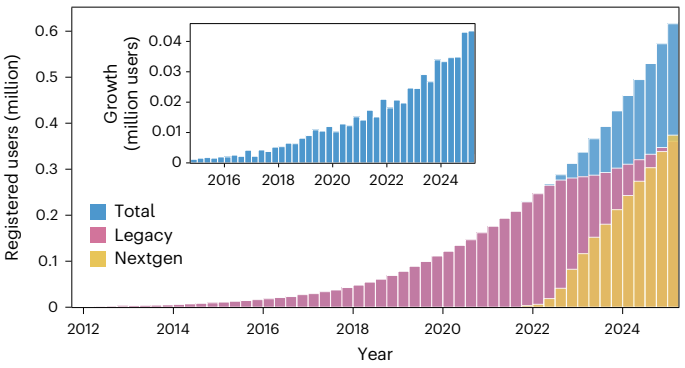
The MP platform

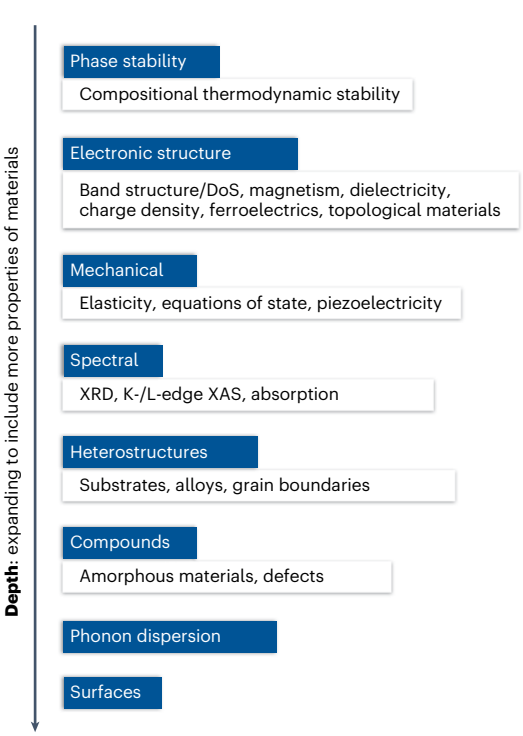
Besides curating a materials database, the MP also serves as a platform for sharing and analysing data with open-source software. Today, the MP is powered by a large collection of software libraries that encompasses a complete data pipeline: from high-throughput workflow management to data extraction, transformation and loading, data analysis and visualization, and its website and API (Fig. 2)[41–44]. These software packages transform raw outputs from first-principles electronic structure codes into interpretable and query-able data, further enabling informative and interactive figures. Created within the MP, these codes have thrived through the combined endeavours of many contributors who applied their scientific domain knowledge with modern software engineering. This combined effort in sustained open-source software development and open data infrastructure has scaled up computational materials science and has helped drive the field towards open science that is shareable and reproducible.

The production workflows in the MP use the commercial Vienna ab initio simulation package (VASP)[45,46] due to its robustness, speed and general acceptance within the broader community. However, our most recent production workflow software, atomate2 (refs. 43,44), enables users to run comparable workflows with other codes. Box 1 discusses both verification and validation of the MP's electronic structure workflows.

https://doi.org/10.1038/s41563-025-02272-0







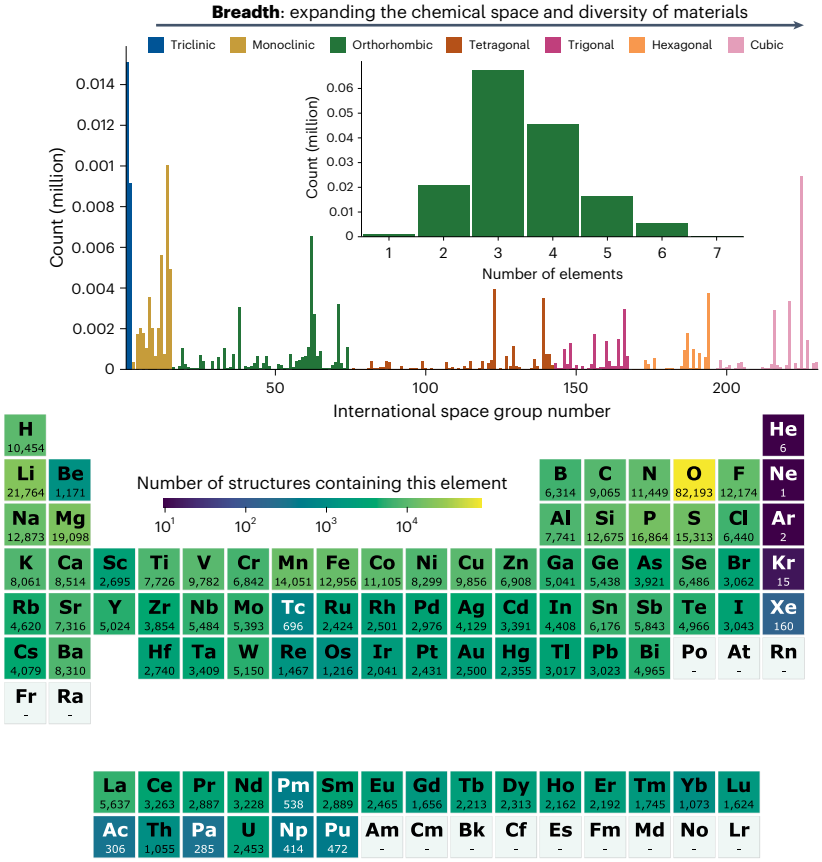


Fig. 1 | Growth of the MP in materials properties and users since its inception. Top: growth of the MP in materials properties (left) and users (right). Registered users are those who provide email addresses, and do not include anonymous users. Depicted are the growth of both the legacy and new next-generation ('nextgen') websites, and the de-duplicated total number of the two combined. The inset shows the growth by quarter since 2015. Bottom: illustration of the depth (left) and breadth (right) axes for expansion of the MP database. Depth increases through the number of properties available for each material (compare Table 1). Breadth increases through the number of materials predicted to be thermodynamically stable, their diversity in space group, elemental complexity and chemical composition. DoS, density of states; EXAFS, extended X-ray absorption fine structure; GGA, generalized gradient approximation; SCAN, strongly constrained and appropriately normed (functional); r²SCAN, regularized–restored SCAN (functional); *U*, Hubbard correction; XANES, X-ray absorption near-edge spectroscopy; XAS, X-ray absorption spectroscopy; XRD, X-ray diffraction. The periodic table was prepared using the open-source Python package pymatviz, which is co-developed by J.R.

In addition to software development, the MP has spearheaded algorithmic advances in many distinct areas, such as crystal structure analysis (for example, CrystalNN[47], ChemEnv[48] and robocrystallographer[49]), inorganic synthesizability analysis (reaction network analysis[50,51] and text-mined recipes[13]) and cathode discovery (charge density analysis[52,53] and migration analysis workflows[54,55]). These tools are continuously being developed and improved by the MP collaboration as part of the MP's software ecosystem. Its success is owed to the efforts of numerous contributors, most of whom are not funded by the MP, enabling the MP to surpass the achievements that are possible for a single research group and showcasing the power of open-source principles in science.

Today, thousands of people visit our website, and millions of data records are downloaded each day. To support increasingly heavy use, the MP has been modernizing its services over the past few years. Specifically, this includes overhauling how data are searched and viewed

Table 1 | Properties available in the MP, including recent advances and roadmaps for future capabilities

Property (year)	Capability	Number of entries
Thermodynamic stability (2011–2020)	Formation energies based on mixing schemes[23,109] of GGA, GGA+*U* and r^{2}SCAN data	341,314
Piezoelectricity (2015)	Piezoelectric tensors calculated using density functional perturbation theory[110]	3,292
Elastic constants (2015)	Elastic constants, compliance tensors and stiffness tensor[111]	12,128
Surfaces (2016)	Surface energies and work functions[112]	134
Dielectric tensors (2017)	Dielectric tensor calculated via density functional perturbation theory, including the ionic component and static component[113]	7,277
Phonon dispersion (2018)	Phonon band structures. Roadmap: add finite temperature phonons and expand to metallic materials	1,521
X-ray absorption spectra (2018)	K- and L-edge calculations including XANES, EXAFS, XAFS[114]	500,000
Equations of state (2018)	Energy–volume relationship across eight thermodynamic equations of state[115]	233
Aqueous stability (2019)	Pourbaix diagrams by combining calculations with experimental aqueous ion references[116]	52,082
Grain boundaries (2020)	Grain boundary energies and work of separation[117]	327
Electronic structure (2020)	Band structures calculated using DFT. Roadmap: improved levels of theory for bandgaps, providing effective masses and Fermi surfaces	70,451
Ferroelectrics (2020)	Screening and workflow for identifying candidate ferroelectric materials, with the external dataset released on MPContribs	413
Magnetism (2020)	Magnetic orderings within the framework of collinear spin-polarized DFT+*U* (ref. 81)	27,000
Optical absorption (2023)	Optical absorption spectra calculated at independent-particle approximation	940
Amorphous materials (2023)	Ab initio molecular dynamics of amorphous structures[91]	5,120
Chemical environment (2023)	Coordination environments with tolerance to small structural distortions	154,302

Details of each property are available via the public documentation for the MP at https://docs.materialsproject.org. A dataset of defect calculations generated using new workflows (based around the pyCDT and pymatgen-analysis-defects python packages) is in progress.

in the explorer and analysis apps, displaying contributed data from the community and developing an easy-to-use Python client for querying data programmatically—all key enablers for data-driven studies.

Since 2015, the MP has put major efforts into making its data on inorganic materials citable and discoverable. Today, the MP has minted over 145,000 digital object identifiers (or DOIs) through a partnership with the US Department of Energy (DOE) Office of Scientific and Technical Information. The DOI metadata contain materials descriptions generated with robocrystallographer, which has made MP materials discoverable in search engines such as Google Dataset Search and DOE Data Explorer.

The MP's mission demands an IT infrastructure that is capable of scaling with evolving requirements while maintaining peak performance, security and efficiency. The MP transitioned to a microservices-based network architecture in 'the cloud' and uses containerized environments to ensure uninterrupted service. A cloud-native observability platform unifies end-to-end visibility across the MP's elastic cloud resources. As our dataset grows in scale and complexity, the MP is transitioning the entire data pipeline to the cloud, while taking advantage of resources provided by industry partners such as the Open Data Sponsorship Program of Amazon Web Services. In doing so, the lean development team brings the infrastructure of the MP to industry standards of scalability and maintainability, providing researchers with a smoother experience.

Recognizing the need for a data-sharing platform, another key shift of the MP is enabling the publishing and sharing of user data. MPContribs, the data contribution platform of the MP, allows materials data of the experimental and computational community to be stored, accessed and searched[42]. The dissemination of these data to the MP's more than half a million users and the option to cite this electronic version of their datasets is an opportunity for scientists to expose their research to a global audience. Data on MPContribs[56] (see https://mpcontribs.org for further documentation) are organized as annotations and expansions to existing MP materials, and are exposed to the MP's user base via the materials detail pages, while leaving full ownership and control over the data with the contributors. The user-submitted entries in MPContribs are organized in strict components to permit fast querying but are otherwise agnostic to whether the data are experimental or simulated. This model places more emphasis on the analysis and curation of materials data over the submission of unstructured raw simulation files, in line with modern data management plans.

Besides data, the MP is building infrastructure that enables collaborators to contribute interactive 'apps'. This infrastructure increases data visibility and usage, and enables community-uploaded data to be discoverable and explorable. The MP also engages in consulting that is scientist-facing, to serve as a user-data facility and compound its impact.

The MP for materials discovery

Whereas the MP is routinely used in data-driven materials design, a key metric of success lies in bringing materials to market and enabling emerging technologies. Despite the typically sluggish pace of materials development and commercialization, the outputs of the MP have directly contributed to synthesizing and characterizing purposefully designed materials and identifying existing ones for novel applications. These achievements are partially credited to the enhanced understanding of materials gained from extensive and precise property data. Consequently, the MP, along with other databases[2,4], has played a pivotal role in advancing the fourth paradigm of materials research[57].

In materials discovery, a common approach involves screening databases on the basis of specific criteria or descriptors to identify and filter candidate materials for further investigation. This approach has been widely adopted for a diverse range of applications, including batteries[58], optoelectronics[59], phosphors[60], thermoelectrics[61], piezoelectrics[62], photocatalysts[63,64], electrides[65], carbon capture[66], magnetocalorics[67], quantum materials such as two-dimensional systems[68], topological insulators[69,70] and semimetals[71]. A recent direction involves 'inverse design' approaches in which atomic composition

https://doi.org/10.1038/s41563-025-02272-0

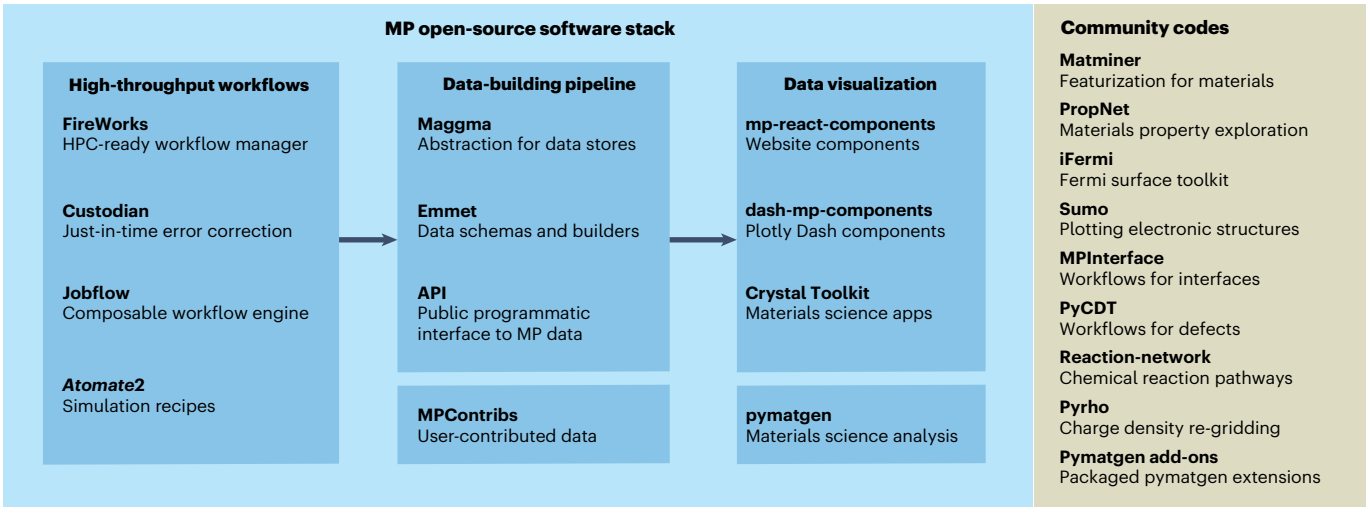


Fig. 2 | The MP ecosystem of open-source software libraries. Left, the key components of the MP software stack and how they depend on each other (indicated by the arrows). Right, a list of several codes that are written either by the community or in partnership with the MP and integrated closely with other MP codes and services. The MP website is almost exclusively powered by open-source software, to which any scientist or developer can submit a change for review. HPC, high-performance computing.

and/or configurations are predicted given a target property without the need to solve the Schrödinger equation, but rather with optimization methods, generative models and materials informatics[72].

Today, more than four papers are published every day citing the MP for its data, software or vision. Notably, some of these studies have resulted in the successful synthesis of materials that exhibit target properties. For example, optoelectronic materials, such as rhombohedral-structured Ba_2BiTaO_6 (Fig. 3a), were synthesized as promising transparent p-type conductors[73–75] after screening 3,600 quaternary oxides on the MP for their band structures and mobilities, followed by more accurate *GW* calculations. Phosphor candidate materials Sr_2LiAlO_4 and $Sr_2AlSi_2O_6N$ (Fig. 3e) were determined by leveraging the stability data and structure-prediction algorithms on the MP and, when activated by Ce^{3+} and Eu^{2+}, show broad emission as white-light-emitting diodes[76,77]. For high-temperature carbon-capture applications, Na_3SbO_4 (Fig. 3b) was identified as a promising material by systematically searching the MP's phase diagrams, covering over 1,400 carbonation reactions, whereby five synthesis candidates were extracted and experimentally verified[78]. A screening of tens of thousands of materials with predicted electron-transport properties revealed a family of XYZ_2 compounds as promising thermoelectric candidates, in which $TmAgTe_2$ and $YCuTe_2$ (Fig. 3d) were synthesized and shown to exhibit a low thermal conductivity and a moderate figure of merit[79,80]. $Mn_{1+x}Sb$ (Fig. 3c) was synthesized as a promising magnetocaloric material, suggested by the screening of over 5,000 MP candidate compounds using the 'magnetic deformation' proxy that correlates with entropy change during magnetization[81,82]. Candidate electrides were discovered by scanning the MP for structures with particular free volumes. This search led to the identification of Sr_3CrN_3, marking an example of materials chemistry featuring a partially filled *d*-shell transition metal, thereby expanding the design parameter space for electrides[83,84]. The new lithium–metal–oxyhalide material $LiMOCl_4$ (M = Nb, Ta) (Fig. 3g), with exceptional ionic conductivity for all-solid-state batteries, was discovered via derivation from a structure identified in the MP[85]. An ML model for the Debye temperature was trained to bulk and shear moduli obtained from the MP, and led to the discovery of Eu^{2+}-doped $NaBaB_9O_{15}$ (Fig. 3f) as an efficient, thermally robust inorganic phosphor material with likely high photoluminescent quantum yields, which was verified through subsequent synthesis and measurements[86].

Challenges and outlook

Materials informatics evolves rapidly, incorporating increasingly intricate structural, chemical and physical properties. Whereas early computable properties comprised structure, stability and bandgap data, advancements in computational methods now provide access to complex properties such as dielectric tensors, absorption spectra and electron mobility. Despite these strides, there remain frontiers for enhancement and expansion, notably in refining the accuracy of electronic structure formalisms, expanding property coverage across structural and chemical space and improving our ability to guide synthesis conditions towards more successful outcomes.

Although amenable to high-throughput computations and reasonably accurate ground-state structure identification, first-principles GGA functionals are known to exhibit deficiencies in capturing localized electronic ground and excited states. As modern computing architectures and quantum mechanical codes become increasingly efficient, more accurate methods become tractable. As an example, MP formation energies are partly computed using the r^{2}SCAN method, and together with the GGA/r^{2}SCAN mixing scheme, result in an overall more accurate thermodynamic energy surface[22,23]. The MP has also developed workflows to elucidate the ground-state magnetic order in inorganic crystals by enumerating and calculating a series of plausible magnetic orderings where non-ferromagnetic orderings have been shown to be correctly predicted for 95% of the benchmark cases[81]. High-throughput advances in excited states have been limited by costly and complex methods, and may be pursued as these methods mature. We note that the MP has computed the ground-state frequency-dependent dielectric tensor (Table 1), which serves as the ground-state estimation for optical properties. Ongoing efforts have been directed towards expanding the data coverage of hybrid functionals and other advanced theoretical methods, which typically involve complex processes and demand orders-of-magnitude-higher computational resources. By leveraging efficient workflow management tools[43,44], we aim to considerably enhance our coverage of more accurate electronic structure data.

Finite temperature properties, including phonons and related thermal properties calculated within the harmonic approximation, are continuously incorporated into the database via (GGA-level) density functional perturbation theory[87]. This data enables users to calculate finite temperature free energies (including vibrational entropy effects), thereby aiding in identifying phase-transition temperatures

BOX 1

MP workflow verification and validation

The MP primarily uses VASP[45,46]—a plane-wave, pseudopotential-based electronic structure code—for its DFT calculations. We note that several other plane-wave DFT packages exist, and these have been benchmarked extensively over the past decade for their internal consistency[118] and their agreement with all-electron reference methods. The continued use of VASP within the MP is motivated partly by historical continuity and technical compatibility with the high-throughput infrastructure of the MP. Additional considerations include VASP's demonstrated efficiency on high-performance computing systems (including graphics processing unit (GPU)-accelerated nodes), the rapid integration of new DFT features (such as support for performing relaxations directly using the DFT stress tensor with meta-GGAs such as r^{2}SCAN (ref. 21)) and its robust pseudopotential library, which exhibits a high degree of transferability across diverse chemical environments[118]. Nonetheless, not all of these features are unique to VASP, and comparable capabilities exist or are being incorporated into other widely used codes.

Although the MP does not conduct routine cross-code comparisons for numerical agreement, it has undertaken rigorous internal validation of its computational workflows, many of which are described in separate publications. For example, in the core workflow, comprising structural relaxations and total energy computations using r^{2}SCAN, the MP has benchmarked convergence with respect to plane-wave energy cut-offs and *k*-point densities[22]. Specialized workflows, such as those for computing elastic constants or thermodynamic equations of state, apply additional accuracy criteria (for example, tighter force convergence) and are tested for agreement with experimental data.

The choice of DFT functional within the MP workflows aims to balance accuracy, robustness and computational efficiency in a high-throughput context. The r^{2}SCAN functional was selected for the core workflow because it maintains numerical stability at scale, in contrast to earlier meta-GGAs that were accurate but often failed under automated execution[21,22]. The choice of functional can be property-dependent; for example, to address the known underestimation of electronic bandgaps by (meta-)GGAs, the MP is developing hybrid functional workflows for more accurate band structure predictions.

A forthcoming paper will provide a comprehensive update on the MP's ongoing validation efforts (A.D.K. and K.A.P., manuscript in preparation). Recent developments include improved *k*-point density estimates, improved choices for pseudopotentials (especially for *f*-block elements) and refinement of internal numerical settings. Some updates have been motivated by discrepancies observed in collaborations with experimentalists, particularly in predicting the stability of lanthanide compounds[94].

between solid phases. The phonon data can also be used to calculate the speed of sound, understand dynamically unstable materials (for example, to identify potential ferroelectric phases), calculate heat capacities or identify frequencies of optical modes (used, for example, in electron-scattering calculations). The dataset of harmonic phonons is currently being expanded and will in the near future include tens of thousands of materials.

Workflows that incorporate third- and fourth-order effects beyond the harmonic approximation have recently been implemented. Once in production, such workflows will provide lattice thermal conductivity, thermal expansion coefficients and renormalization of imaginary frequencies at higher temperatures, which will inform finite temperature stability. Furthermore, the detailed force constant matrices determined using such workflows can be used to refine ML interatomic potentials.

An overarching challenge in expanding the materials design space is to connect computation with experiments beyond well-ordered, stoichiometric bulk limits. Owing to the dramatically increased combinatorial phase space, efforts to design alloys using high-throughput computation[67] for applications such as high-entropy alloys, thermoelectrics and magnetocalorics are usually limited to a specific structure and chemistry. Addressing general solid-solution behaviour across all metals, semiconductors and insulators is intractable today. However, the MP has implemented an analysis framework to extract hypothetical alloy pairs between commensurate crystal structures with existing physical laws[36]. Whereas this approach can serve as an initial filter to explore the alloy space, more effort is needed to label and recognize possible alloy-forming compounds.

Non-crystalline nanomaterials pose other challenging avenues, finding applications in energy storage, catalysis, coatings, aerospace and more. Their complex energy landscape requires extensive computational efforts but offers a richer representation of non-equilibrium structures, valuable for future interatomic ML endeavours[8]. The MP has developed workflows for calculating non-crystalline structures and surfaces[88,89], and the growing database is used to gauge the thermodynamic limit for synthesizability[90,91], for instance. Collaborative efforts such as the Open Catalyst Project[20] have expanded capabilities of the MP by providing extensive surface adsorption data, enabling the training of ML models to predict surface passivation and chemistry.

Ultimately, designing a material with a set of desired properties involves assessing its realizability. Predictive synthesis, a rapidly evolving field, focuses on methods to predict optimal synthesis conditions and precursors. As different synthesis methods operate under a range of environmental parameters with a varying degree of thermodynamic and kinetic influences, descriptors such as energy-above-hull offer approximate guidance on product stability but do not necessitate synthesizability[92]. Recent advancements include setting a rigorous limit on synthesizability[90] and evaluating competing phase space via reaction networks and convex hull shape analysis[50,51]. To address this, the MP collaboration developed web applications for phase diagrams, chemical potential diagrams and Pourbaix diagrams, which depict the thermodynamic landscape's topology and its environmental dependencies.

To harness knowledge in the published literature, MP collaborators used natural language processing to extract synthesis recipes, contributing valuable data to the MP and the wider community. It is important to note, however, that the published literature rarely documents failures and 'dark reactions', whose inclusion is crucial in any data-driven methodology[93]. As automated robotic laboratories become operational, their data infrastructure will offer systematic, standardized and reproducible data on reactants and conditions, irrespective of success[94].

Designing a material from the bottom up using computational methods to realizing it in the laboratory is difficult due to four limitations: the accuracies in ab initio methods; the length- and timescale of a system that can be described by the ab initio method; the lack of full complexity of real-world systems at a finite temperature with defects, disorders, vibrations and kinetic influences; and the consideration of higher-level constraints such as device manufacturing, scale-up, revenue and market size. Despite these limits, we have seen how computation, and its data, are able to reveal trends, guide research efforts and identify and explain physical/chemical phenomena. There are indeed many documented cases where theoretical predictions have led to experimentally confirmed functional materials at the laboratory scale[95,96]. In this regard, ML methods have the potential to massively

https://doi.org/10.1038/s41563-025-02272-0

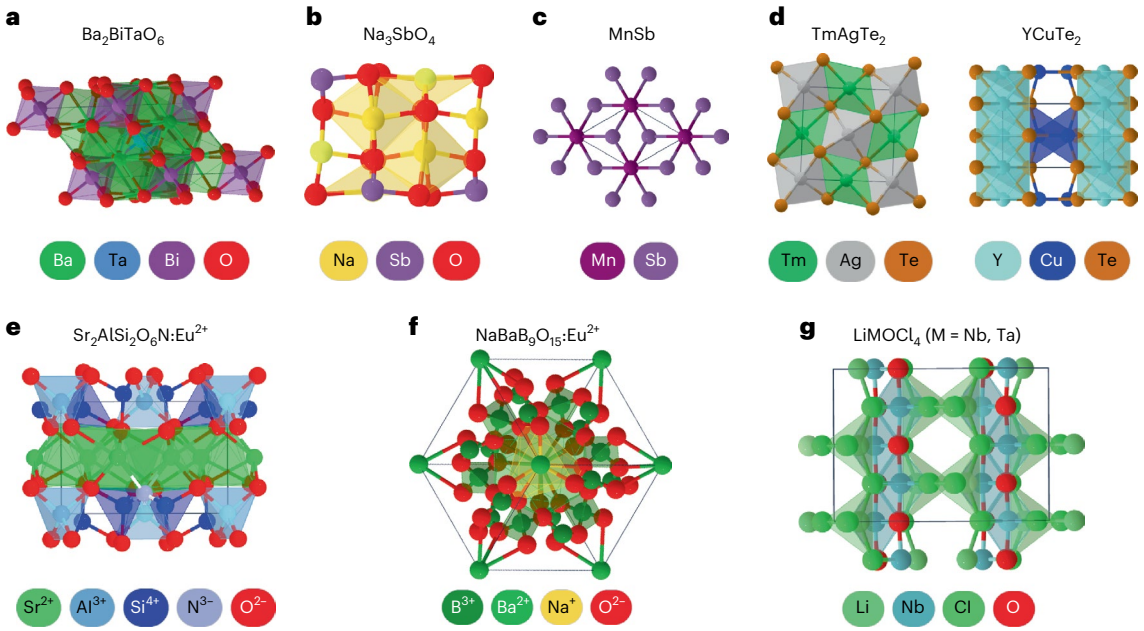


Fig. 3 | Examples of compounds predicted to have target properties by leveraging the data and analysis tools of the MP. These compounds were subsequently synthesized and experimentally confirmed either within the MP collaboration or by external collaborators. The predicted compositions and corresponding screened properties are as follows. **a**, Ba_2BiTaO_6, a p-type transparent conductor screened for bandgap, effective mass and valence band maximum[73–75]. **b**, Na_3SbO_4, a carbon-capture material screened via phase diagram and oxygen reaction potential[78]. **c**, $Mn_{1+x}Sb$, a magnetocaloric material (a magnetic compound for thermal cooling) screened via a magnetic deformation descriptor[81,82]. **d**, $TmAgTe_2$ and $YCuTe_2$, thermoelectric materials screened by electron-transport properties[79,80]. **e**, $Sr_2AlSi_2O_6N:Eu^{2+}$, a phosphor material for energy-efficient lighting, predicted to have extra-broad emission via the screening of thermodynamic stability and bandgaps[76,77]. **f**, $NaBaB_9O_{15}:Eu^{2+}$, a phosphor material predicted to have a high photoluminescent quantum yield by an ML model trained on bulk and shear moduli data in the MP[86]. **g**, $LiMOCl_4$ (M = Nb, Ta), a solid electrolyte (for solid-state batteries), derived via substitution based on the $LiVOF_4$ structure (MP entry: mp-850188)[85]. Figure prepared using the open-source Python package crystal_toolkit, developed and maintained within the MP.

expand the capacities of traditional ab initio methods, an area that the MP has helped to advance.

The MP for artificial intelligence/machine learning

Having amassed and curated vast electronic structure calculation data, the MP has emerged as a key enabler of integrating artificial intelligence (or AI) into materials science, a development unforeseen at its launch.

Taking advantage of a decade's worth of structural and potential energy surface data from MP relaxation trajectories, universal graph deep learning interatomic potentials with coverage of the entire periodic table have been developed. Examples including M3GNet[39] and CHGNet[40] have shown notable accuracy in obtaining fundamental materials properties for a very broad structural and phase space. These models have greatly expanded the space for materials discovery and design, enabling access to scales that are challenging for first-principles computations[97]. Such efforts have further emphasized the importance of expanding the unexplored compositional and configurational space in the datasets, in particular, by adding metastable and non-equilibrium systems.

Besides ML potentials, successful property prediction models include the prediction of elastic tensors, densities of states and X-ray absorption spectra by leveraging the existing data on the MP[98–100], a fruitful outcome of the extensive effort that goes into improving our materials property coverage. For volumetric data, we released an electronic charge density database and provided the code for changing the representation of the charge density[101], expecting it to enable advanced ML studies of materials.

In meeting the community's need for large, reliable and diverse datasets to develop ML algorithms, the MP has also made a specific effort to release standardized materials datasets and provide additional codes to support their use. As an example, the Matbench suite has curated a benchmarking dataset of a diverse range of 13 properties of solid materials, which has become a canonical dataset adopted by the community to test emerging ML architectures[102–104].

Looking ahead, we expect greater integration of ML-derived datasets. Workflows in atomate2 are adapted to swap DFT-based simulations with ML potentials, enabling flexible choice of speed or accuracy. Users will be able to perform virtual materials design across multiple properties with rapid feedback loops. We note that ML and first-principles theories are complementary techniques: ML simulations provide the opportunity to help the MP make efficient use of ab initio data.

The MP community

Beyond citations and statistics, the most important measure of the success of a project is its widespread use and implied community trust. Over the past decade, the MP has cultivated an evolving community that relies on our data, algorithms and software. The MP platform is also commonly used as a tool in classrooms for educators and students in physical sciences. One important future direction is user education, which is realized via constant efforts in providing clearly documented methods and easily interpretable data and website visualizations, and via community engagement. Through these, we hope that the MP can serve as an accessible platform for sharing the accumulated knowledge on computational materials and educating beginners in the field. Notably, first-principles data, once only interpreted by theorists, are now democratized and used routinely by experimental groups and by industry.

One effort has been establishing the Materials Science Community Discourse at matsci.org, run in partnership with the OpenKIM project[105], that aims to be a shared venue for the materials science community to facilitate discussions around methods, data and tools. It

was established without preference to any one research group or code, and with a process to allow new groups to add their own categories to the forum.

As the number of codes and the diversity of data increases, so does MP's responsibility to train others in how to use them responsibly. To this end, the MP has run workshops for direct training and disseminates the workshop material and videos for free. Documentation, example notebooks and in-line code examples on the new website are provided wherever possible. The intent is to create a pathway for users without previous programming experience to learn how to programmatically use the data and improve the codes for the benefit of the community. To create a welcoming and open environment for participants of all levels, most MP code bases use an open governance model where codes are reviewed and decisions debated in public. All participants, formal or informal, are expected to abide by the code of conduct.

We recently launched the Materials Project Foundation as a form of governance to ensure the sustainability of our software ecosystem, and to recognize and incentivize the important professional service from external contributors. The overarching goal of the Foundation is to promote a community-driven, transparent and inclusive governance model for our public-facing open-source codes. Major development decisions are documented in a public repository to serve as a guide to community members seeking to integrate their work with the MP software stack.

Community involvement extends to user-contributed data which appear clearly marked alongside the MP's generated data with references to the according publications. This service facilitates data sharing among scientists, leveraging the MP's analysis tools and expanding our global user base. The mutual benefit lies in reliable dissemination of third-party data via our infrastructure, enriching the diversity of our dataset. Notably, experimental data remain scarce, underscoring the need for future efforts in gathering more 'gold standard' datasets, a task facilitated by advancements in automated laboratories and high-throughput experiments[106,107].

Industry entities such as the Open Catalyst Project[20] and the GNoME project[35] have partnered with the MP, leveraging its data and software infrastructure. In turn, computed structures from both projects are now hosted on the MP. We hope that these collaborations will increase public awareness of materials research and draw greater interest from the private sector in fundamental materials research for a sustainable future.

Conclusion

Over the past decade, the MP has grown in maturity, accuracy and scope, meeting the expanding needs of the community. By making large inorganic materials data and its calculations FAIR[108]—that is, findable, accessible, interoperable and reusable—the rapid development of ML algorithms has been enabled, further accelerating materials design and entering the fourth paradigm of data-driven science. The MP is a continuously evolving resource, with opportunities and challenges including improvements in first-principles methodology, expanding the scope of materials properties, exploring phase spaces of increasing complexity, and predicting synthesizability. It will continue the trend towards more openness and solicit third-party contributions in data and software, enabling researchers to take advantage of the expanding network effect to discover each other's data and perform ever-more comprehensive data mining and analysis.

The advancement of computational science includes an impressive array of data, tools, algorithms and simulation methods, which requires an equal advancement in the skills of engineers and researchers to take full advantage of them. The adoption of computational tools has to become routine through the democratization of data and underlying methodologies. To achieve this, the data generated by high-throughput efforts and accelerated learning models must be reliable and vetted and—when errors are made—corrected responsibly and openly. Through the implementation of a common software infrastructure, it becomes feasible for data and methods to be rigorously analysed against state-of-the-art methods and examined by a broad audience. With this vision in mind, the MP strives to serve as an engine that not only provides resources for the community but also leads scientific advancement in data-driven materials design.

References

1. Draxl, C. & Scheffler, M. The NOMAD laboratory: from data sharing to artificial intelligence. *J. Phys. Mater.* **2**, 036001 (2019).
2. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the Open Quantum Materials Database (OQMD). *JOM* **65**, 1501–1509 (2013).
3. Kirklin, S. et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
4. Curtarolo, S. et al. AFLOW: an automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
5. Calderon, C. E. et al. The AFLOW standard for high-throughput materials science calculations. *Comput. Mater. Sci.* **108**, 233–238 (2015).
6. Choudhary, K. et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput. Mater.* **6**, 173 (2020).
7. Talirz, L. et al. Materials Cloud, a platform for open computational science. *Sci. Data* **7**, 299 (2020).
8. Yang, R. X. et al. Big data in a nano world: a review on computational, data-driven design of nanomaterials structures, properties, and synthesis. *ACS Nano* **16**, 19873–19891 (2022).
9. Qu, X. et al. The Electrolyte Genome project: a big data approach in battery materials discovery. *Comput. Mater. Sci.* **103**, 56–67 (2015).
10. Cheng, L. et al. Accelerating electrolyte discovery for energy storage with high-throughput screening. *J. Phys. Chem. Lett.* **6**, 283–291 (2015).
11. Spotte-Smith, E. W. C. et al. A database of molecular properties integrated in the Materials Project. *Digit. Discov.* **2**, 1862–1882 (2023).
12. Huo, H. et al. Semi-supervised machine-learning classification of materials synthesis procedures. *npj Comput. Mater.* **5**, 62 (2019).
13. Kononova, O. et al. Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* **6**, 203 (2019).
14. He, T. et al. Similarity of precursors in solid-state synthesis as text-mined from scientific literature. *Chem. Mater.* **32**, 7861–7873 (2020).
15. Zhou, F., Cococcioni, M., Marianetti, C. A., Morgan, D. & Ceder, G. First-principles prediction of redox potentials in transition-metal compounds with LDA+*U*. *Phys. Rev. B* **70**, 235121 (2004).
16. Adams, S. & Rao, R. P. High power lithium ion battery materials by computational design: high power Li ion battery materials by computational design. *Phys. Status Solidi A* **208**, 1746–1753 (2011).
17. Wang, L., Maxisch, T. & Ceder, G. A first-principles approach to studying the thermal stability of oxide cathode materials. *Chem. Mater.* **19**, 543–552 (2007).
18. Ong, S. P., Jain, A., Hautier, G., Kang, B. & Ceder, G. Thermal stabilities of delithiated olivine MPO_4 (M=Fe, Mn) cathodes investigated using first principles calculations. *Electrochem. Commun.* **12**, 427–430 (2010).
19. Rosen, A. S. et al. High-throughput predictions of metal–organic framework electronic properties: theoretical challenges, graph neural networks, and data exploration. *npj Comput. Mater.* **8**, 112 (2022).
20. Chanussot, L. et al. Open Catalyst 2020 (OC20) dataset and community challenges. *ACS Catal.* **11**, 6059–6072 (2021).

21. Furness, J. W., Kaplan, A. D., Ning, J., Perdew, J. P. & Sun, J. Accurate and numerically efficient r^{2}SCAN meta-generalized gradient approximation. *J. Phys. Chem. Lett.* **11**, 8208–8215 (2020).
22. Kingsbury, R. et al. Performance comparison of r^{2}SCAN and SCAN metaGGA density functionals for solid materials via an automated, high-throughput computational workflow. *Phys. Rev. Mater.* **6**, 013801 (2022).
23. Kingsbury, R. S. et al. A flexible and scalable scheme for mixing computed formation energies from different levels of theory. *npj Comput. Mater.* **8**, 195 (2022).
24. Zagorac, D., Müller, H., Ruehl, S., Zagorac, J. & Rehme, S. Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. *J. Appl. Crystallogr.* **52**, 918–925 (2019).
25. Villars, P. et al. The Pauling File, Binaries Edition. *J. Alloys Compd.* **367**, 293–297 (2004).
26. Gražulis, S. et al. Crystallography Open Database – an open-access collection of crystal structures. *J. Appl. Crystallogr.* **42**, 726–729 (2009).
27. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
28. Jacobsson, T. J. et al. An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. *Nat. Energy* **7**, 107–115 (2021).
29. Borysov, S. S., Geilhufe, R. M. & Balatsky, A. V. Organic materials database: an open-access online database for data mining. *PLoS ONE* **12**, e0171501 (2017).
30. Landis, D. D. et al. The Computational Materials Repository. *Comput. Sci. Eng.* **14**, 51–57 (2012).
31. Schmidt, J. et al. Machine-learning-assisted determination of the global zero-temperature phase diagram of materials. *Adv. Mater.* **35**, 2210788 (2023).
32. Hautier, G., Fischer, C., Ehrlacher, V., Jain, A. & Ceder, G. Data mined ionic substitutions for the discovery of new compounds. *Inorg. Chem.* **50**, 656–663 (2011).
33. Eckert, H. et al. The AFLOW library of crystallographic prototypes: part 4. *Comput. Mater. Sci.* **240**, 112988 (2024).
34. Ye, W., Lei, X., Aykol, M. & Montoya, J. H. Novel inorganic crystal structures predicted using autonomous simulation agents. *Sci. Data* **9**, 302 (2022).
35. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
36. Woods-Robinson, R., Horton, M. K. & Persson, K. A. A method to computationally screen for tunable properties of crystalline alloys. *Patterns* **4**, 100723 (2023).
37. Barroso-Luque, L. et al. smol: a Python package for cluster expansions and beyond. *J. Open Source Softw.* **7**, 4504 (2022).
38. Scheffler, M. et al. FAIR data enabling new horizons for materials research. *Nature* **604**, 635–642 (2022).
39. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).
40. Deng, B. et al. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* **5**, 1031–1041 (2023).
41. Ong, S. P. et al. Python Materials Genomics (pymatgen): a robust, open-source Python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
42. Huck, P. et al. User applications driven by the community contribution framework MPContribs in the Materials Project. *Concurr. Comput.* **28**, 1982–1993 (2016).
43. Ganose, A. et al. Atomate2: modular workflows for materials science. Preprint at https://doi.org/10.26434/chemrxiv-2025-tcr5h (2025).
44. Ganose, A. et al. Atomate2 code repository. *GitHub* https://github.com/materialsproject/atomate2 (2025).
45. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
46. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
47. Zimmermann, N. E. R. & Jain, A. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC Adv.* **10**, 6063–6081 (2020).
48. Waroquiers, D. et al. ChemEnv: a fast and robust coordination environment identification tool. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.* **76**, 683–695 (2020).
49. Ganose, A. M. & Jain, A. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Commun.* **9**, 874–881 (2019).
50. McDermott, M. J., Dwaraknath, S. S. & Persson, K. A. A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. *Nat. Commun.* **12**, 3097 (2021).
51. McDermott, M. J. et al. Assessing thermodynamic selectivity of solid-state reactions for the predictive synthesis of inorganic materials. *ACS Cent. Sci.* **9**, 1957–1975 (2023).
52. Shen, J.-X., Horton, M. & Persson, K. A. A charge-density-based general cation insertion algorithm for generating new Li-ion cathode materials. *npj Comput. Mater.* **6**, 161 (2020).
53. Li, H. H., Shen, J.-X. & Persson, K. A. A rapid lithium-ion cathode discovery pipeline and its exemplary application. *Energy Adv.* https://doi.org/10.1039/D3YA00397C (2024).
54. Shen, J.-X., Li, H. H., Rutt, A., Horton, M. K. & Persson, K. A. Topological graph-based analysis of solid-state ion migration. *npj Comput. Mater.* **9**, 99 (2023).
55. Rutt, A. et al. Expanding the material search space for multivalent cathodes. *ACS Appl. Mater. Interfaces* **14**, 44367–44376 (2022).
56. Huck, P., Jain, A., Gunter, D., Winston, D. & Persson, K. A community contribution framework for sharing materials data with materials project. In *2015 IEEE 11th International Conference on e-Science* 535–541 (IEEE, 2015); https://doi.org/10.1109/eScience.2015.75
57. Bauer, S. et al. Roadmap on data-centric materials science. *Model. Simul. Mater. Sci. Eng.* **32**, 063301 (2024).
58. Aykol, M. et al. High-throughput computational design of cathode coatings for Li-ion batteries. *Nat. Commun.* **7**, 13779 (2016).
59. Luo, S., Li, T., Wang, X., Faizan, M. & Zhang, L. High-throughput computational materials screening and discovery of optoelectronic semiconductors. *WIREs Comput. Mol. Sci.* **11**, e1489 (2021).
60. Luo, X. & Xie, R.-J. Recent progress on discovery of novel phosphors for solid state lighting. *J. Rare Earths* **38**, 464–473 (2020).
61. Gorai, P., Stevanović, V. & Toberer, E. S. Computationally guided discovery of thermoelectric materials. *Nat. Rev. Mater.* **2**, 17053 (2017).
62. Talley, K. R., Sherbondy, R., Zakutayev, A. & Brennecka, G. L. Review of high-throughput approaches to search for piezoelectric nitrides. *J. Vac. Sci. Technol. A* **37**, 060803 (2019).
63. Singh, A. K., Gorelik, R. & Biswas, T. Data-driven discovery of robust materials for photocatalytic energy conversion. *Annu. Rev. Condens. Matter Phys.* **14**, 237–259 (2023).
64. Pan, J. & Yan, Q. Data-driven material discovery for photocatalysis: a short review. *J. Semicond.* **39**, 071001 (2018).
65. Zhao, S., Kan, E. & Li, Z. Electride: from computational characterization to theoretical design. *WIREs Comput. Mol. Sci.* **6**, 430–440 (2016).

66. Ren, E., Guilbaud, P. & Coudert, F.-X. High-throughput computational screening of nanoporous materials in targeted applications. *Digit. Discov.* **1**, 355–374 (2022).
67. Garcia, C. A. C., Bocarsly, J. D. & Seshadri, R. Computational screening of magnetocaloric alloys. *Phys. Rev. Mater.* **4**, 024402 (2020).
68. Shen, L., Zhou, J., Yang, T., Yang, M. & Feng, Y. P. High-throughput computational discovery and intelligent design of two-dimensional functional materials for various applications. *Acc. Mater. Res.* **3**, 572–583 (2022).
69. Su, Y. et al. High-throughput first-principle prediction of collinear magnetic topological materials. *npj Comput. Mater.* **8**, 261 (2022).
70. Frey, N. C. et al. High-throughput search for magnetic and topological order in transition metal oxides. *Sci. Adv.* **6**, eabd1076 (2020).
71. Gao, J. et al. High-throughput screening for Weyl semimetals with S_4 symmetry. *Sci. Bull.* **66**, 667–675 (2021).
72. Ren, Z. et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter* **5**, 314–335 (2022).
73. Hautier, G., Miglio, A., Ceder, G., Rignanese, G.-M. & Gonze, X. Identification and design principles of low hole effective mass p-type transparent conducting oxides. *Nat. Commun.* **4**, 2292 (2013).
74. Ricci, F. et al. An ab initio electronic transport database for inorganic materials. *Sci. Data* **4**, 170085 (2017).
75. Bhatia, A. et al. High-mobility bismuth-based transparent p-type oxide from high-throughput material screening. *Chem. Mater.* **28**, 30–34 (2016).
76. Wang, Z. et al. Mining unexplored chemistries for phosphors for high-color-quality white-light-emitting diodes. *Joule* **2**, 914–926 (2018).
77. Li, S. et al. Data-driven discovery of full-visible-spectrum phosphor. *Chem. Mater.* **31**, 6286–6294 (2019).
78. Dunstan, M. T. et al. Large scale computational screening and experimental discovery of novel materials for high temperature CO_2 capture. *Energy Environ. Sci.* **9**, 1346–1360 (2016).
79. Zhu, H. et al. Computational and experimental investigation of $TmAgTe_2$ and XYZ_2 compounds, a new group of thermoelectric materials identified by first-principles high-throughput screening. *J. Mater. Chem. C* **3**, 10554–10565 (2015).
80. Aydemir, U. et al. $YCuTe_2$: a member of a new class of thermoelectric materials with $CuTe_4$-based layered structure. *J. Mater. Chem. A* **4**, 2461–2472 (2016).
81. Horton, M. K., Montoya, J. H., Liu, M. & Persson, K. A. High-throughput prediction of the ground-state collinear magnetic order of inorganic materials using density functional theory. *npj Comput. Mater.* **5**, 64 (2019).
82. Cooley, J. A. et al. From waste-heat recovery to refrigeration: compositional tuning of magnetocaloric $Mn_{1+x}Sb$. *Chem. Mater.* **32**, 1243–1249 (2020).
83. Burton, L. A., Ricci, F., Chen, W., Rignanese, G.-M. & Hautier, G. High-throughput identification of electrides from all known inorganic materials. *Chem. Mater.* **30**, 7521–7526 (2018).
84. Chanhom, P. et al. Sr_3CrN_3: a new electride with a partially filled *d*-shell transition metal. *J. Am. Chem. Soc.* **141**, 10595–10598 (2019).
85. Tanaka, Y. et al. New oxyhalide solid electrolytes with high lithium ionic conductivity >10 mS cm^{-1} for all-solid-state batteries. *Angew. Chem. Int. Ed.* **62**, e202217581 (2023).
86. Zhuo, Y., Mansouri Tehrani, A., Oliynyk, A. O., Duke, A. C. & Brgoch, J. Identifying an efficient, thermally robust inorganic phosphor host via machine learning. *Nat. Commun.* **9**, 4377 (2018).
87. Petretto, G. et al. High-throughput density-functional perturbation theory phonons for inorganic materials. *Sci. Data* **5**, 180065 (2018).
88. Aykol, M., Montoya, J. H. & Hummelshøj, J. Rational solid-state synthesis routes for inorganic materials. *J. Am. Chem. Soc.* **143**, 9244–9259 (2021).
89. Sivonxay, E. & Persson, K. A. Density functional theory assessment of the lithiation thermodynamics and phase evolution in Si-based amorphous binary alloys. *Energy Storage Mater.* **53**, 42–50 (2022).
90. Aykol, M., Dwaraknath, S. S., Sun, W. & Persson, K. A. Thermodynamic limit for synthesis of metastable inorganic materials. *Sci. Adv.* **4**, eaaq0148 (2018).
91. Zheng, H. et al. The ab initio non-crystalline structure database: empowering machine learning to decode diffusivity. *npj Comput. Mater.* https://doi.org/10.1038/s41524-024-01469-2 (2024)
92. Bartel, C. J. Review of computational approaches to predict the thermodynamic stability of inorganic solids. *J. Mater. Sci.* **57**, 10475–10498 (2022).
93. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
94. Szymanski, N. J. et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86–91 (2023).
95. Hautier, G., Jain, A. & Ong, S. P. From the computer to the laboratory: materials discovery and design using first-principles calculations. *J. Mater. Sci.* **47**, 7317–7340 (2012).
96. Jain, A., Shin, Y. & Persson, K. A. Computational predictions of energy materials using density functional theory. *Nat. Rev. Mater.* **1**, 15004 (2016).
97. Ko, T. W. & Ong, S. P. Recent advances and outstanding challenges for machine learning interatomic potentials. *Nat. Comput. Sci.* **3**, 998–1000 (2023).
98. Fung, V., Ganesh, P. & Sumpter, B. G. Physically informed machine learning prediction of electronic density of states. *Chem. Mater.* **34**, 4848–4855 (2022).
99. Torrisi, S. B. et al. Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum–property relationships. *npj Comput. Mater.* **6**, 109 (2020).
100. Kong, S. et al. Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings. *Nat. Commun.* **13**, 949 (2022).
101. Shen, J.-X. et al. A representation-independent electronic charge density database for crystalline materials. *Sci. Data* **9**, 661 (2022).
102. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Comput. Mater.* **6**, 138 (2020).
103. Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J. & Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *npj Comput. Mater.* **7**, 77 (2021).
104. Cheng, G., Gong, X.-G. & Yin, W.-J. Crystal structure prediction by combining graph network and optimization algorithm. *Nat. Commun.* **13**, 1492 (2022).
105. Tadmor, E. B., Elliott, R. S., Sethna, J. P., Miller, R. E. & Becker, C. A. The potential of atomistic simulations and the knowledgebase of interatomic models. *JOM* **63**, 17 (2011).
106. Burger, B. et al. A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
107. Szymanski, N. J. et al. Toward autonomous design and synthesis of novel inorganic materials. *Mater. Horiz.* **8**, 2169–2198 (2021).
108. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
109. Jain, A. et al. Formation enthalpies by mixing GGA and GGA+*U* calculations. *Phys. Rev. B* **84**, 045115 (2011).

110. de Jong, M., Chen, W., Geerlings, H., Asta, M. & Persson, K. A. A database to enable discovery and design of piezoelectric materials. *Sci. Data* **2**, 150053 (2015).
111. de Jong, M. et al. Charting the complete elastic properties of inorganic crystalline compounds. *Sci. Data* **2**, 150009 (2015).
112. Tran, R. et al. Surface energies of elemental crystals. *Sci. Data* **3**, 160080 (2016).
113. Petousis, I. et al. High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials. *Sci. Data* **4**, 160134 (2017).
114. Mathew, K. et al. High-throughput computational X-ray absorption spectroscopy. *Sci. Data* **5**, 180151 (2018).
115. Latimer, K., Dwaraknath, S., Mathew, K., Winston, D. & Persson, K. A. Evaluation of thermodynamic equations of state across chemistry and structure in the Materials Project. *npj Comput. Mater.* **4**, 40 (2018).
116. Patel, A. M., Nørskov, J. K., Persson, K. A. & Montoya, J. H. Efficient Pourbaix diagrams of many-element compounds. *Phys. Chem. Chem. Phys.* **21**, 25323–25327 (2019).
117. Zheng, H. et al. Grain boundary properties of elemental metals. *Acta Mater.* **186**, 40–49 (2020).
118. Bosoni, E. et al. How to verify the precision of density-functional-theory implementations via reproducible and universal workflows. *Nat. Rev. Phys.* **6**, 45–58 (2024).

Acknowledgements

This work was intellectually led by the Materials Project program KC23MP, supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under contract no. DE-AC02-05-CH11231. The Materials Project Collaboration includes the authors of this manuscript in addition to current and previous members of the Materials Project program, for example, developers of workflows for the prediction of various properties, plus contributors from the broader community acknowledged in this manuscript. We thank all users of the MP for their support and feedback. We thank all contributors to the MP software stack, without whom the MP would not be possible. A complete and up-to-date list of contributors is publicly available at GitHub (https://github.com/materialsproject#contributors). This research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy User Facility using NERSC award BES-ERCAP 0032604. A.S.R. acknowledges support via a Miller Research Fellowship from the Miller Institute for Basic Research in Science, University of California, Berkeley.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Kristin A. Persson.

Peer review information *Nature Materials* thanks Giulia Galli, Matthias Scheffler and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [2]Department of Materials Science and Engineering, University of California, Berkeley, Berkeley, CA, USA. [3]Department of Chemistry, Imperial College London, London, UK. [4]Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA. [5]Andlinger Center for Energy and the Environment, Princeton University, Princeton, NJ, USA. [6]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China. [7]Lawrence Livermore National Laboratory, Livermore, CA, USA. [8]Scientific Data Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [9]Cavendish Laboratory, University of Cambridge, Cambridge, UK. [10]Federal Institute for Materials Research and Testing, Berlin, Germany. [11]Friedrich Schiller University Jena, Jena, Germany. [12]Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. [13]Energy Storage and Distributed Resources Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [14]Université catholique de Louvain, Louvain-la-Neuve, Belgium. [15]Matgenix SRL, Charleroi, Belgium. [16]Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [17]Department of Physics, University of California, Berkeley, Berkeley, CA, USA. [18]Department of Engineering, Dartmouth University, Hanover, NH, USA. [19]Department of Nanoengineering, University of California San Diego, San Diego, CA, USA. ✉e-mail: kapersson@lbl.gov