The Materials Project: Accelerating Materials Design Through Theory-Driven Data and Tools

Anubhav Jain, Joseph Montoya, Shyam Dwaraknath, Nils E. R. Zimmermann, John Dagdelen, Matthew Horton, Patrick Huck, Donny Winston, Shreyas Cholia, Shyue Ping Ong, and Kristin Persson

Contents

A. Jain (✉) · J. Montoya · S. Dwaraknath · N. E. R. Zimmermann · M. Horton · P. Huck · D. Winston · S. Cholia
Lawrence Berkeley National Laboratory, Berkeley, CA, USA
e-mail: ajain@lbl.gov; montoyjh@lbl.gov; shyamd@lbl.gov; nerz@lbl.gov; mkhorton@lbl.gov; phuck@lbl.gov; dwinston@lbl.gov; scholia@lbl.gov

J. Dagdelen
University of California, Berkeley, CA, USA
e-mail: jdagdelen@berkeley.edu

S. P. Ong
Department of NanoEngineering, University of California, San Diego, La Jolla, CA, USA
e-mail: ongsp@eng.ucsd.edu

K. Persson
Lawrence Berkeley National Laboratory, Berkeley, CA, USA

University of California, Berkeley, CA, USA
e-mail: kapersson@lbl.gov

W. Andreoni, S. Yip (eds.), *Handbook of Materials Modeling*,
https://doi.org/10.1007/978-3-319-42913-7_60-1

Abstract

The Materials Project (MP) is a community resource for theory-based data, web-based materials analysis tools, and software for performing and analyzing calculations. The MP database includes a variety of computed properties such as crystal structure, energy, electronic band structure, and elastic tensors for tens of thousands of inorganic compounds. At the time of writing, over 40,000 users have registered for the MP database. These users interact with this data either through the MP web site (https://www.materialsproject.org) or through a REpresentational State Transfer (REST) application programming interface (API). MP also develops or contributes to several open-source software libraries to help set up, automate, analyze, and extract insight from calculation results. Furthermore, MP is developing tools to help researchers share their data (both computational and experimental) through its platform. The ultimate goal of these efforts is to accelerate materials design and education by providing new data and software tools to the research community. In this chapter, we review the history, theoretical methods, impact (including user-led research studies), and future goals for the Materials Project.

1 History and Overview of the Materials Project

Materials scientists and engineers have always depended on materials property data to inform, guide, and explain research and development. Traditionally, such data originated almost solely from experimental studies. In the past 10–15 years, it has become possible to rapidly generate reliable materials data using scalable computer simulations of the fundamental equations of physics such as the Schrödinger equation. This paradigm shift was induced by a combination of theoretical advances, most notably the development of density functional theory (DFT), algorithmic improvements, and low-cost computing.

The Materials Project (MP, or “The Project”) was founded in 2011 as a collaborative effort to leverage ongoing advances in theory and computing to accelerate the research and design of new materials. The Project rests on a comprehensive database of predicted properties of materials that is the result of executing millions of DFT simulations on supercomputing resources. At the time of writing, this database includes >69,000 inorganic materials with crystal structures and total energies, >57,000 materials with electronic band structures, >48,000 with electronic transport properties (Fig. 1) (Ricci et al. 2017), >30,000 with XANES

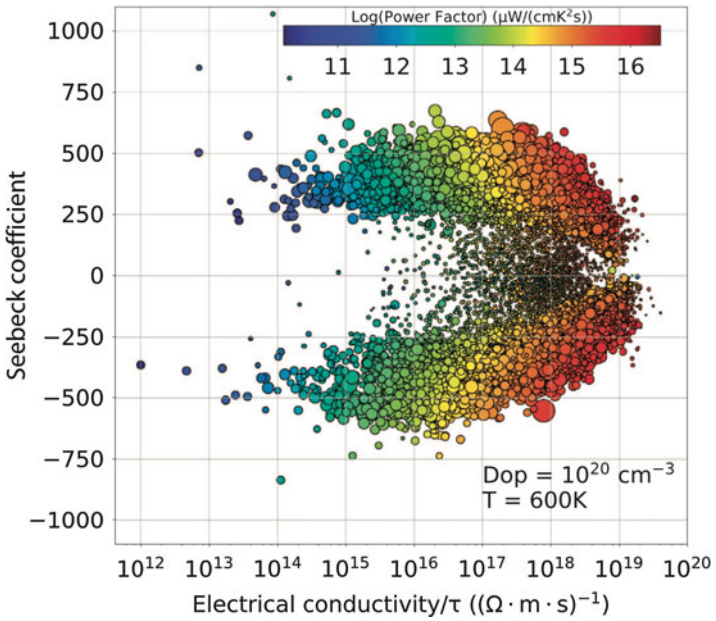
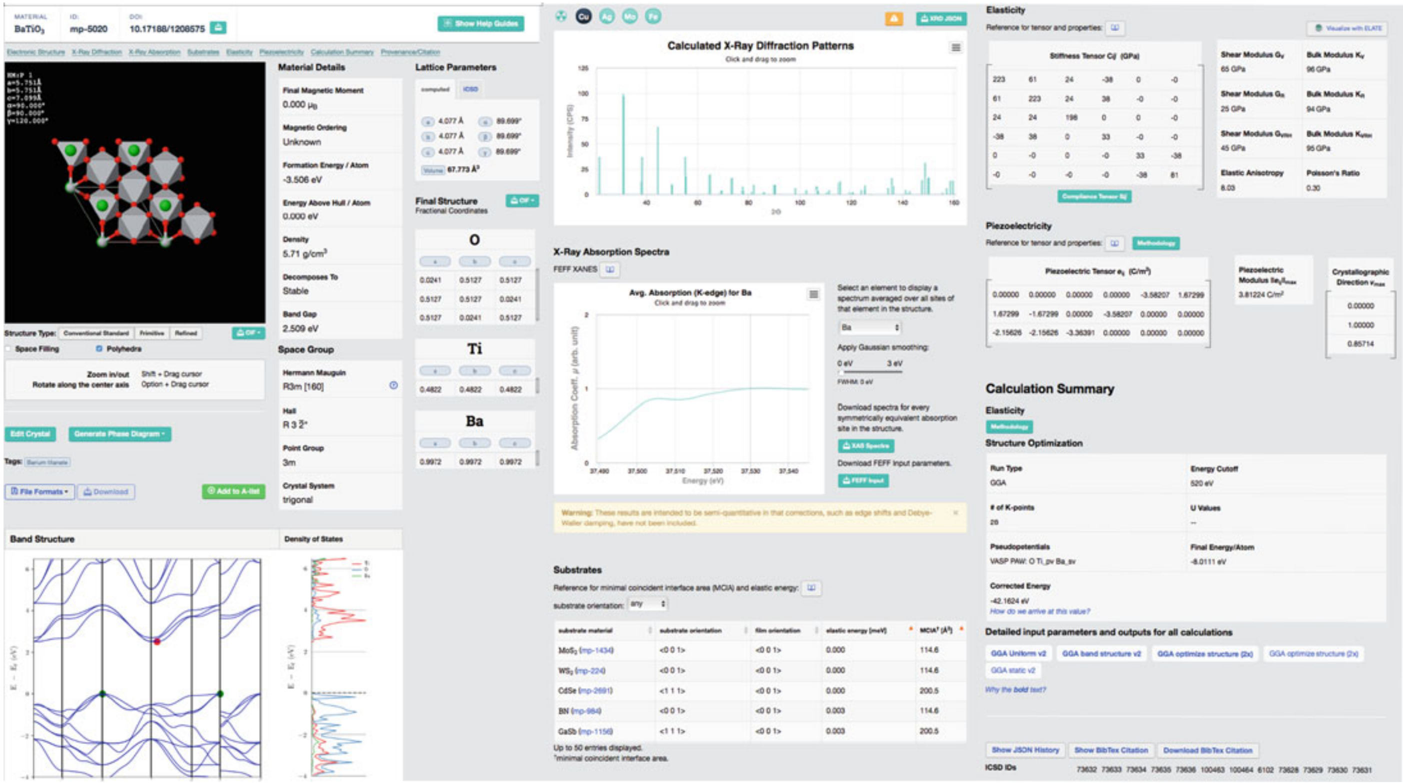


Fig. 1 Example of a large electronic transport data set in MP generated through computations. Each point represents one compound, with Seebeck coefficient versus electron conductivity (divided by τ) plotted. The color represents the thermoelectric power factor ($S^2\sigma$), and the point size is proportional to the bandgap (Ricci et al. 2017). This data set is available through the MPContribs platform (see Section 6.2) at: https://materialsproject.org/mpcontribs/boltztrap

k-edge spectra (Dozier et al. 2017), >15,000 with conversion battery properties, >6000 with elastic tensors (de Jong et al. 2015a), >3,000 with intercalation battery properties, >1,000 with piezoelectric tensors (de Jong et al. 2015b), >1,000 with dielectric tensors (Petousis et al. 2017), and > 1000 elemental surface energies (Tran et al. 2016). This database is continually expanding with more materials and more properties (see Fig. 2 for an example of properties listed in the current iteration).

The Project launched its publicly accessible web site in October 2011 and has since grown into a multi-institution collaboration as part of the US Department of Energy Office of Basic Energy Sciences (BES). The web site provides access to the database as well as applications (or “apps”) that combine and visually present the data for specific analyses such as phase diagram generation or battery electrode evaluation. The MP web site hosts more than 40,000 registered users worldwide consisting of a diverse set of researchers and students from academia, industry, and educational institutions (Figs. 3 and 4). The diversity of the audience base highlights the usefulness of a theory-based materials database across the spectrum of education, research, and development activities.

Apart from the core data and web site, MP helps develop and maintain a set of open-source software libraries for setting up, executing, analyzing, and deriving



MATERIAL BaTiO$_3$ | ID: mp-5020 | DOI: 10.17188/1208575

Show Help Guides

Electronic Structure X-Ray Diffraction X-Ray Absorption Substrates Elasticity Piezoelectricity Calculation Summary Provenance/Citation

Structure Type: Conventional Standard | Primitive | Refined

☐ Space Filling ☑ Polyhedra

Zoom in/out: Shift + Drag cursor
Rotate along the center axis: Option + Drag cursor

Edit Crystal | Generate Phase Diagram

Tags: Barium titanate

File Formats | Download | Add to A-list

Material Details

Final Magnetic Moment	0.000 μ_B
Magnetic Ordering	Unknown
Formation Energy / Atom	-3.506 eV
Energy Above Hull / Atom	0.000 eV
Density	5.71 g/cm^3
Decomposes To	Stable
Band Gap	2.509 eV

Space Group

Hermann Mauguin	R3m [160]
Hall	R 3 2"
Point Group	3m
Crystal System	trigonal

Lattice Parameters

computed | ICSD

a	4.077 Å	α	89.699°
b	4.077 Å	β	89.699°
c	4.077 Å	γ	89.699°

Volume 67.773 Å^3

Final Structure
Fractional Coordinates

O

a	b	c
0.0241	0.5127	0.5127
0.5127	0.5127	0.0241
0.5127	0.0241	0.5127

Ti

a	b	c
0.4822	0.4822	0.4822

Ba

a	b	c
0.9972	0.9972	0.9972

Band Structure | **Density of States**

Cu Ag Mo Fe | XRD JSON

Calculated X-Ray Diffraction Patterns
Click and drag to zoom

X-Ray Absorption Spectra

FEFF XANES

Avg. Absorption (K-edge) for Ba
Click and drag to zoom

Select an element to display a spectrum averaged over all sites of that element in the structure.

Ba

Apply Gaussian smoothing:
0 eV 3 eV
FWHM: 0 eV

Download spectra for every symmetrically equivalent absorption site in the structure.

XAS Spectra

Download FEFF input parameters.

FEFF Input

Warning: These results are intended to be semi-quantitative in that corrections, such as edge shifts and Debye-Waller damping, have not been included.

Substrates

Reference for minimal coincident interface area (MCIA) and elastic energy:

substrate orientation: any

substrate material	substrate orientation	film orientation	elastic energy (meV)	MCIA* (Å^2)
MoS$_2$ (mp-1434)	<0 0 1>	<0 0 1>	0.000	114.6
WS$_2$ (mp-224)	<0 0 1>	<0 0 1>	0.000	114.6
CdSe (mp-2691)	<1 1 1>	<0 0 1>	0.000	200.5
BN (mp-984)	<0 0 1>	<0 0 1>	0.003	114.6
GaSb (mp-1156)	<1 1 1>	<0 0 1>	0.003	200.5

Up to 50 entries displayed.
*minimal coincident interface area.

Elasticity

Reference for tensor and properties: | Visualize with ELATE

Stiffness Tensor Cij (GPa)

223	61	24	-38	0	-0
61	223	24	38	-0	-0
24	24	198	0	0	-0
-38	38	0	33	-0	-0
0	-0	0	-0	33	-38
-0	-0	-0	-0	-38	61

Compliance Tensor Sij

Shear Modulus G$_V$ 65 GPa	Bulk Modulus K$_V$ 96 GPa
Shear Modulus G$_R$ 25 GPa	Bulk Modulus K$_R$ 94 GPa
Shear Modulus G$_{VRH}$ 45 GPa	Bulk Modulus K$_{VRH}$ 95 GPa
Elastic Anisotropy 8.03	Poisson's Ratio 0.30

Piezoelectricity

Reference for tensor and properties: | Methodology

Piezoelectric Tensor e$_{ij}$ (C/m^2)

0.00000	0.00000	0.00000	0.00000	-3.58207	1.67299
1.67299	-1.67299	0.00000	-3.58207	0.00000	0.00000
-2.15626	-2.15626	-3.36391	0.00000	0.00000	0.00000

Piezoelectric Modulus ‖e$_{ij}$‖$_{max}$
3.81224 C/m^2

Crystallographic Direction v$_{max}$
0.00000
1.00000
0.85714

Calculation Summary

Elasticity

Methodology

Structure Optimization

Run Type GGA	Energy Cutoff 520 eV
# of K-points 28	U Values --
Pseudopotentials VASP PAW: O Ti_pv Ba_sv	Final Energy/Atom -8.0111 eV
Corrected Energy -42.1624 eV How do we arrive at this value?	

Detailed input parameters and outputs for all calculations

GGA Uniform v2 | GGA band structure v2 | GGA optimize structure (2x) | GGA optimize structure (2x) | GGA static v2

Why the bold text?

Show JSON History | Show BibTex Citation | Download BibTex Citation

ICSD IDs 73632 73633 73634 73635 73636 100463 100464 6102 73628 73629 73630 73631

Fig. 2 An example of a "materials detail" page for BaTiO$_3$ on the Materials Project web site. The information available includes crystal structure parameters, thermodynamic properties, electronic band structure and density of states, a simulated x-ray diffraction pattern, simulated x-ray absorption spectra, a substrate matcher, an elasticity tensor, a piezoelectric tensor, a table and links pertaining to calculation methodology details, and metadata regarding the crystal structure source

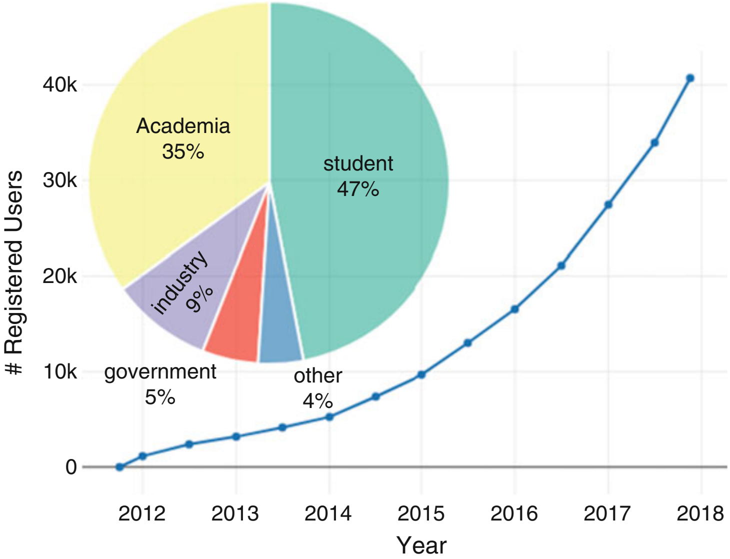


Fig. 3 Total number of registered users since release of the MP web site and fraction of users belonging to various institution types

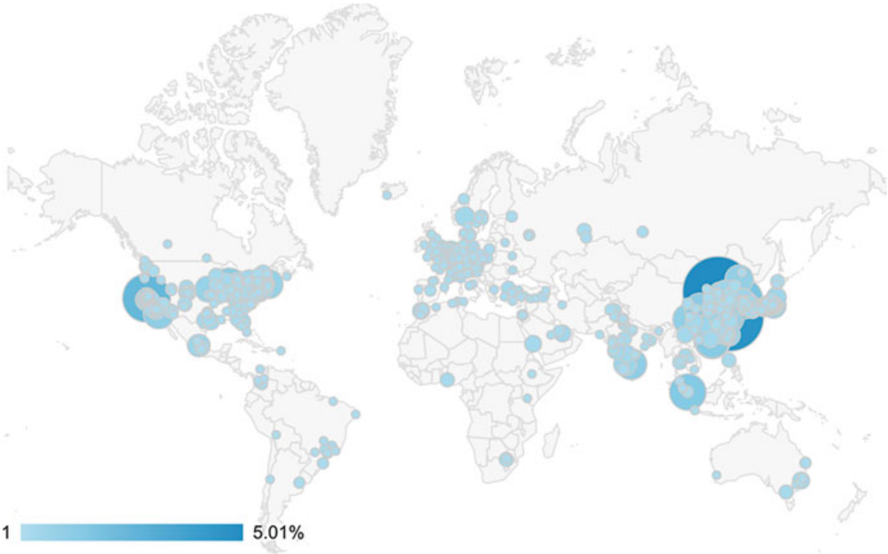


Fig. 4 Amount of Materials Project user sessions by city for the month of October 2017. Sessions originated in 112 countries, 36 of which totaled >100 sessions

insights from calculations. These libraries, which include pymatgen (Ong et al. 2013), custodian, FireWorks (Jain et al. 2015), and atomate (Mathew et al. 2017), have been used by hundreds of researchers worldwide. The newest additions to MP allow users to suggest compounds for computation as well as contribute their own data (theoretical or experimental) to the database. Furthermore, MP hosts educational workshops focused on its online and programmatic infrastructure, and the MP web site has become an integral teaching tool in several materials science courses.

As the state of the art in theory and computing are bound to change, the specifics of MP's data, scope, capabilities, and infrastructure will no doubt change as well. Nevertheless, this chapter summarizes the current state of the Materials Project.

2 Underlying Theoretical Formalism and Development of Materials Design "Apps"

2.1 Theoretical Methods

The Materials Project's core data set consists of results obtained from density functional theory (DFT) calculations on a library of inorganic compounds. DFT is well suited for creating a database of materials properties because it has fewer parameters that require tuning for different materials systems and because the computational cost for small- to medium-sized (approximately 300 atoms or less) systems is manageable. DFT methods have become standardized to a large extent such that various software implementations with slightly different parameters (e.g., pseudopotentials) produce very similar results (Lejaeghere et al. 2016).

Nevertheless, selecting a robust set of parameters for high-throughput computations is still not trivial. It is important to emphasize there is currently no perfect DFT functional as they are all approximations to the complete set of physics that define materials phenomena. For example, strongly correlated systems remain challenging. It is typically possible to treat even complex systems with specialized methods in single studies. However, when constructing a large database with many compounds, such specialized treatment is difficult to achieve practically and would also lead to inconsistent and often incompatible results between various compounds. Additionally, one must more carefully balance computational costs with expected information gain. Whereas a single study may not be noticeably impacted if its calculations are over-converged numerically and use 50% more computing power than necessary, such a situation would severely slow down a high-throughput database project such as MP that consumes tens of millions of CPU hours of computing per year. Thus, MP must make practical compromises that try to maintain the accuracy of a specialized, precise calculation while being completely automatic and computationally efficient and maintaining clarity and consistency of procedure with other calculations.

One of the approaches used by the Materials Project to achieve this balance is to split materials into two classes and apply a different DFT functional to model each class. The first class of compounds are transition metal oxides and sulfides. Standard DFT functionals such as the local density approximation (LDA) (Kohn and Sham 1965) and the generalized gradient approximation (GGA) (Perdew et al. 1996) are not accurate for these compounds due to more pronounced self-interaction error as well as errors in orbital occupation from lack of derivative discontinuity (Zhou et al. 2004; Cococcioni and de Gironcoli 2005). One computationally efficient way to treat these compounds is with the GGA+U framework, in which a Hubbard-like correction is applied to localized d orbitals. The specific U corrections are fitted to formation energy data as described previously (Wang et al. 2006). It is important to note that these same U values may not be optimal for accurately representing other properties such as the electronic band structure. The second class of compounds encompasses all other systems and is treated with the standard GGA-PBE functional (Perdew et al. 1996).

By allowing different compounds to be treated with two different functionals, it is possible to enhance accuracy of the resulting database compared to using only a single functional such as GGA for the entire database. However, one must then additionally design a scheme to mix results (e.g., total energies) obtained from different methods since these results are not directly compatible. In the Materials Project, these adjustments between results from different functionals are made by benchmarking to experimental formation enthalpy data (Jain et al. 2011b). Figure 5 depicts the effects of one instance of this by presenting Fe-P-O phase diagrams using the GGA only, GGA+U only, and mixed GGA and GGA+U total energies. Only the version of the diagram that uses two different functionals (with the mixing adjustment applied) reproduces all known stable phases in this system.

Another practical measure taken by the Materials Project pertains to molecular systems. Although molecular systems and solids can be modeled within the same density functional theory framework (e.g., PBE-GGA with plane-wave basis sets), computed reaction energies that include both molecules and solids typically exhibit high errors because self-interaction errors differ significantly between systems characterized by local (e.g., molecules or highly correlated systems) and nonlocal (e.g., metals) electrons (Grindy et al. 2013; Perdew et al. 1998). Similarly, intermolecular interactions present in gases, 2D materials, and liquids that are not well described by pure GGA functionals present further challenges for constructing a comprehensive thermodynamic framework derived from DFT that avoids such systematic errors.

Rather than calculating the liquid/gas energies directly, MP adjusts the energies of several elements that are liquid or gaseous at room temperature based on experimental reaction enthalpies such as the oxidation of metals (Wang et al. 2006). All of the following compounds have adjusted energies to better reproduce reaction energies with solid phases: O_2, N_2, Cl_2, F_2, and H_2.

Finally, we mention that MP also adjusts certain numerical parameters based on the type of compound. For example, MP uses a denser k-point mesh when calculating metals (as determined from an initial, loose k-point mesh calculation) versus semiconductors and insulators. In addition, the numerical tolerances used by the Materials Project have been growing more precise over time. The parameters used for each calculation are available via the Materials Project web site, and

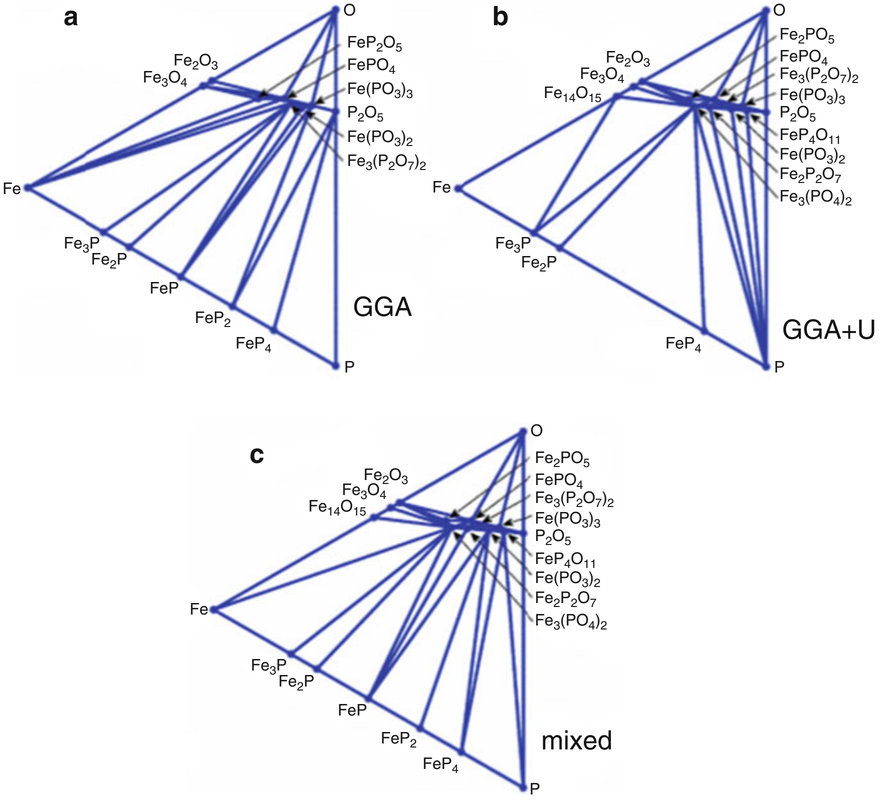


Fig. 5 Fe-P-O ternary phase diagrams built using total energy calculations from (**a**) only GGA, (**b**) only GGA+U, and (**c**) mixing GGA and GGA+U functionals. Only the mixed phase diagram reproduces all known phases as stable on the phase diagram. (Reprinted figure with permission from Jain et al. (2011b). Copyright 2011 by the American Physical Society)

the most current description of parameter settings is provided at https://www.materialsproject.org/docs/calculations.

2.2 "Apps" for Data Exploration

Much of the value of the MP data set comes from secondary analyses that are performed on top of the raw data. These secondary analyses often combine multiple data points and can take the form of common diagrams used in materials science (e.g., phase stability diagrams or Pourbaix diagrams), application-specific materials design tools (e.g., evaluating MP compounds as battery electrodes), or simply as additional information (e.g., reporting potential substrates that might form coherent lattices with a target material). Such tools are vital for helping users extract as much value as possible from the data sets.

The Materials Project develops the methodologies to perform many such secondary analyses and releases them both as open-source software implementations (through the pymatgen (Ong et al. 2013) package) and as web applications ("apps"). Apps provide a visual, user-friendly interface to these powerful and often complex analysis routines. In the following example, we describe the underlying methodology as well as the accompanying app for generating and manipulating phase diagrams.

2.2.1 Phase Diagram App

Phase diagrams have multiple applications in materials science. Traditional phase diagrams generated from experiments show not only stable phases but also delineate solubility limits and temperature dependence. In contrast, because MP currently only models materials at zero temperature and pressure and does not model solubility limits, the resulting phase diagrams might be more accurately referred to as phase *stability* diagrams (we use the terms interchangeably here). Nevertheless, such phase stability diagrams show the stable phases in a given chemical system as well as the relevant phase equilibria at various compositional ratios. One major application of such phase diagrams is to serve as a "reality check" for new hypothetical materials. If the energy of that proposed material is low enough to be on or nearly on the phase stability diagram, there is a higher probability that the material will be stable enough to be synthesized in the lab (Sun et al. 2016). Phase stability diagrams are also useful for identifying possible decomposition products that might compete with a target phase.

Generating such computational phase diagrams requires knowledge of the formation energies of all possible materials within a chemical system. For example, calculating a ternary phase diagram requires knowledge of the formation energies of all the relevant unary, binary, and ternary phases in that system. For a typical ternary system, calculating the energy for all known phases would require several dozen calculations. However, because the MP database already contains precomputed energies for most known inorganic compounds, one can now avoid running all these simulations and directly create reasonably complete phase diagrams using the MP data set.

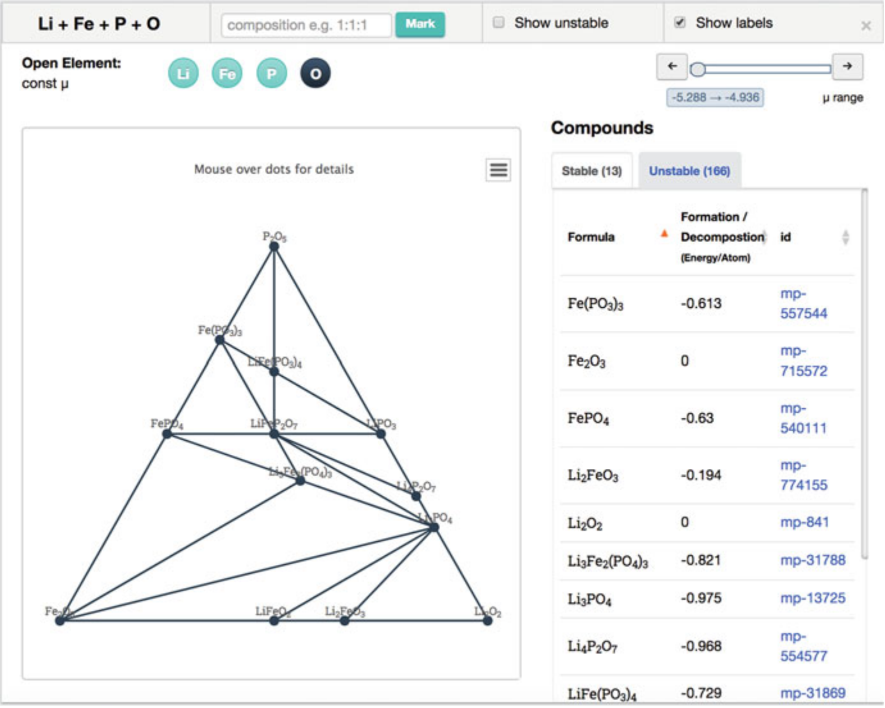
Mathematically, the set of stable points on a phase diagram can be determined using the convex hull construction, which is a method of finding the minima as a function of n degrees of freedom (Barber et al. 1996). By calculating the convex hull for the total energies of various calculated DFT energies, globally stable structures can be found as well as the various tie-lines that connect stable phases. The convex hull construction can be used to construct phase diagrams for an arbitrary number of components.

Many known compounds are not thermodynamically stable, i.e., they do not appear on phase stability diagrams (Sun et al. 2016). An additional metric is then necessary to distinguish the degree of metastability for these compounds. The construction of a convex hull provides an envelope of stability. Compounds on the convex hull are stable, while compounds above the hull in energy are metastable. The energetic distance to the hull at the composition is thus a quantifiable metric and directly related to the metastability of that compound. A lower energy above the hull is typically desirable for synthesis because it implies less of an energy penalty to form the target compound compared to the known stable phase. Many of the known metastable compounds in the Materials Project are within 15 meV/atom of the hull, but depending on chemistry can extend past 60 meV/atom above the hull (Sun et al. 2016). While this analysis focuses on the metastability of known compounds, there is still work needed to quantify the limits of metastability.

Thus far we have described the formalism for closed systems, i.e., ones in which the stoichiometric ratio of elements is fixed, but the same formalism can be equally applied to open systems in which one or more elements are held at a fixed chemical potential rather than held to a fixed amount. For example, experiments may be carried out in air, which essentially serves as an infinite reservoir of atmospheric elements such as oxygen and nitrogen at particular chemical potentials. The same experiment under flowing argon gas would still represent an open system, but one in which the chemical potentials of those elements are greatly reduced. Thus, in environments that are open to a particular element, the relevant control variable is the chemical potential of that element (μ_i) rather than its compositional value. The chemical potential is then treated as an external variable to obtain a grand potential phase diagram.

Users of MP need not be familiar with all the methodological details (Ong et al. 2008) of computational phase diagram construction to generate and use them. The MP web site allows users to simply type (or click on a visual periodic table) the elements for the system they are interested in. This will generate a phase diagram that will graphically display the phase diagram as well as a list of stable and metastable/unstable materials. Figure 6 shows a screenshot of the MP phase diagram app for a grand potential phase diagram for Li-Fe-P-O with an oxygen potential of -5.288 eV. Note that since the oxygen composition is prescribed by the potential, it doesn't exist as a degree of freedom in the phase diagram, collapsing the quaternary phase diagram into a ternary phase diagram (with a slider for controlling the oxygen chemical potential).

Other apps similarly make available powerful underlying methodologies to a broad audience. For example, similar to the grand potential phase diagram, Pourbaix diagrams are projections of global stability into potential-pH space to model electrochemical stability. A methodology for calculating such diagrams by utilizing experimentally measured free energies of aqueous ions and the calculated DFT energies for solid phases available in the Materials Project was previously developed (Persson et al. 2012). This methodology was incorporated into a "Pourbaix app" that allows users to simply select the chemical system of interest, elemental ratio, and concentration of ions in order to generate a familiar Pourbaix diagram that leverages the MP data set and that can be visually and interactively explored by the user. In addition, the stability of individual materials relative to the most stable decomposition product may be generated as a heatmap overlaid on the Pourbaix diagram, providing users with a tool to estimate metastability under aqueous conditions (Singh et al. 2017).



□ Show unstable

☑ Show labels

Open Element: const μ

-5.288 → -4.936 μ range

Formula	Formation / Decompostion (Energy/Atom)	id
$Fe(PO_3)_3$	-0.613	mp-557544
Fe_2O_3	0	mp-715572
$FePO_4$	-0.63	mp-540111
Li_2FeO_3	-0.194	mp-774155
Li_2O_2	0	mp-841
$Li_3Fe_2(PO_4)_3$	-0.821	mp-31788
Li_3PO_4	-0.975	mp-13725
$Li_4P_2O_7$	-0.968	mp-554577
$LiFe(PO_3)_4$	-0.729	mp-31869

Fig. 6 The Li-Fe-P-O ternary grand potential phase diagram open to oxygen as generated by the Materials Project's "phase diagram" app

3 Computation Infrastructure and Software Tools

Developing and maintaining calculation databases such as the Materials Project requires considerable attention to computing and software infrastructure. At the time of this writing, the Materials Project is the result of over one million individual calculations that represent over 100 million central processing unit (CPU) hours of computing time invested. Setting up, executing, analyzing, and managing all these calculations are far from straightforward. Here, we describe the infrastructure of the Materials Project at the time of this writing. However, we note that the economics of computing as well as the optimal choice of software libraries can change very quickly. The Materials Project infrastructure is therefore constantly evolving to apply the latest developments and best practices in computer science and software engineering to the field of materials science.

3.1 Computing Resources

The Materials Project (MP) has employed high-performance computing (HPC) resources at the US National Energy Research Scientific Computing Center (NERSC) and elsewhere, consuming over 100 million of CPU hours to date. Many-task computing workflows (Raicu et al. 2008) are increasingly using HPC environments because these resources typically offer the potential for large amounts of total computing time, good hardware specifications (e.g., moderate to high memory), and adequate storage. However, HPC environments present several challenges for running high-throughput calculations because these environments were originally designed to serve the needs of a small number of large, highly parallel applications that run for predictable times and perform all input/output to disk. In contrast, high-throughput calculations are typically extremely numerous and limited in achievable parallelism and require unpredictable, often very long total run times. In addition, they are often more suited to management by external services rather than solely through flat files on disk. To overcome these challenges, the Materials Project has developed a software library for running high-throughput calculations called “FireWorks” (Jain et al. 2015) that solves many of the computing challenges associated with running high-throughput jobs on HPC resources.

3.2 Choice of Database Software

Many portions of a high-throughput calculation workflow require efficient storage, retrieval, and search of information, including:

- Managing the state of high-throughput calculations
- Storage of the raw calculation results, and
- A searchable set of processed data for data dissemination and analysis

The Materials Project has chosen to use a not-only-SQL (NoSQL) “document store” (Cattell 2011), MongoDB, as its main database technology for these tasks (raw output files are also preserved). We note that this represents a shift from a other SQL-based data management strategies used previously in high-throughput computational materials science (Jain et al. 2011a). This decision was made primarily because MongoDB accommodates both the data heterogeneity and rapid pace of data model development required by the Materials Project. For example, unlike typical SQL relational database management systems (RDBMS) such as MySQL and PostgreSQL, MongoDB does not require a pre-designed, normalized schema between all data types at the beginning of the project. The types of data being stored continually evolve as we add new types of calculations into the project. By choosing MongoDB, MP can adapt quickly to these changes with small changes in application code instead of refactoring complex relational schemata.

Among document-oriented datastores, MongoDB is notable for its simple but powerful query language, ease of administration, and good performance on read-heavy workloads where most of the commonly accessed data (the so-called working set) can fit into memory. Its relative weaknesses for linking disparate data (database "joins") and write-heavy workloads are a reasonable trade-off for MP. A productivity benefit of MongoDB is that both the query language and the native data model are JavaScript Object Notation (JSON) (Bray 2017), which is the standard data format for modern web applications and easily represented and manipulated as native data types in the Python programming language (Van Rossum et al. 2007) in which our other software libraries are written. Thus, users familiar with Python and in particular its "dict" object can adapt quickly to understanding and developing data models with MongoDB. Our experience is that these aspects have allowed many more members of our team to collaborate on database development compared to our historical use of RDBMS in which only one or two members of the team were familiar enough with the system to make changes. More details on our experiences and challenges encountered in deploying a centralized datastore of this type within a scientific HPC ecosystem are described in Gunter et al. (2012).

3.3 Software Stack

3.3.1 Software to Perform and Analyze DFT Calculations

At the time of this writing, the Materials Project primarily uses density functional theory as implemented by the Vienna Ab Initio Simulation Package (VASP) (Kresse and Hafner 1994; Kresse and Furthmüller 1996). However, it is likely that other software packages such as ABINIT (Gonze et al. 2016) will play a larger role in MP in the future. Regardless of the choice of DFT implementation, the procedure for performing calculations involves many steps outside the core simulation. These steps include:

- Setting up the geometry for the material or system of interest
- Defining a workflow of calculations to compute the properties of interest
- Executing the calculations and correcting possible errors
- Analyzing, storing, and organizing the output data

The Materials Project has developed a comprehensive suite of software tools to accelerate and assist in these steps (see Fig. 7 for an overview).

Most of the compounds currently in the Materials Project use bulk crystal structure geometries as reported in the Inorganic Crystal Structure Database (ICSD) (Belsky et al. 2002). However, the computation of many properties requires performing algorithmic operations on these geometries. Examples include determining an appropriate ordered cell for sites with partial occupancies, creating appropriate slabs for surface calculations, and performing a series lattice deformations for computing elastic tensors. We have implemented routines for such geometry modifications in Python as part of the pymatgen (Ong et al. 2013) open-source software

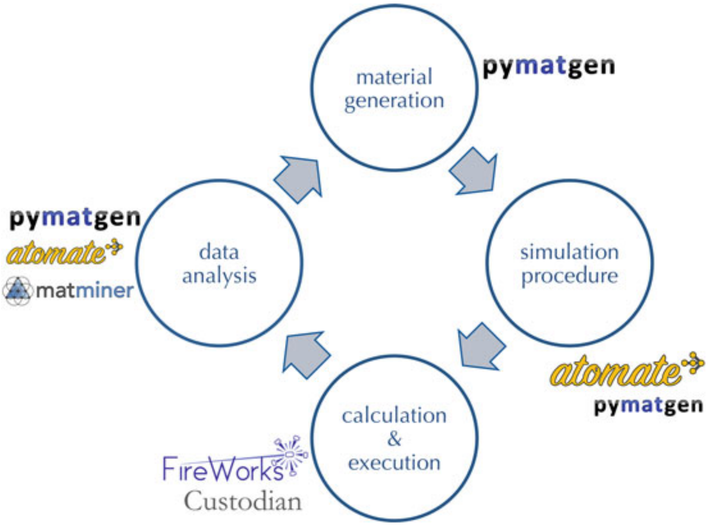


Fig. 7 Various steps involved in data generation and analysis along with the relevant software stack for the Materials Project infrastructure

library. In many cases, these routines are directly implemented in pymatgen, whereas in others we provide an object-oriented Python wrapper to libraries released by the community such as spglib (Togo 2018) and enumlib (Hart and Forcade 2008). We note that the growth of web-based collaboration presents the opportunity for another method of generating new compounds: crowdsourced user suggestions. In this method, crystal structures designed by the user community (either offline or through Materials Project tools) are used as starting points for the calculation with the results reported back to the community. In its first three years of operation, this "MPComplete" service has been employed by over 800 unique users and has resulted in over 8,300 new materials added to MP's public database.

Depending on the property to be studied, a DFT "calculation" may in fact involve a series of individual computations that require data passing and modifications of geometry or input settings between computations. The set of calculations required for obtaining a desired output property, along with the dependencies and data passing requirements between these calculations, define a "workflow." The Materials Project has developed two software libraries in the Python programming language to manage such workflows. The first library, called "FireWorks" (Jain et al. 2015), is a general-purpose workflow library. FireWorks does not contain any materials science or DFT-specific code. Its scope is to provide a framework for users to define arbitrary sequences of calculations, store them in a database, execute them on various types of computing resource, and manage the status of potentially millions of workflows across systems. Thus, FireWorks is compatible with a broad class of scientific computing workflows (although it is best suited for high-throughput applications) and is frequently used outside the field of materials science. The second library for workflow creation, “atomate” (Mathew et al. 2017), contains specific materials science workflows implemented in FireWorks and using pymatgen as a base library. The atomate package can be thought of as providing a library of materials science workflow implementations (e.g., standard workflows for electronic band structure, elastic properties, and piezoelectric, dielectric, and ferroelectric properties). Atomate users can specify an input geometry for a material and the desired workflow type, and atomate will provide a FireWorks-based implementation of that workflow that is ready to execute at supercomputing centers. Furthermore, atomate leverages the pymatgen library to automatically parse calculation outputs and create a database of materials properties that can be queried by the user. Various features in FireWorks and atomate allow for customization of behavior to specific situations, from low-level issues (such as interacting with various queueing systems) to high-level issues (such as running the same workflow with multiple DFT functional choices). Calculation workflows can also automatically adapt their procedure for later calculations based on the results obtained from earlier calculations.

When executing the calculation, it is possible to encounter various errors relating to calculation convergence. The Materials Project has developed a type of job wrapper to simulation software (e.g., VASP Kresse and Hafner 1994; Kresse and Furthmüller 1996 or QChem Kong et al. 2000) called “custodian” (Ong et al. 2014) that automatically monitors the output files of the calculation and automatically fixes errors (by stopping the job, changing the input files, and restarting the job) according to a set of rules. The custodian software can also be used to automate linear sequences of calculations (e.g., a convergence protocol that tightens numerical parameters until no change in output is achieved).

Once the calculation is executed, the results are parsed and stored in various database collections. Raw data is parsed by pymatgen as a component of atomate workflows. We note that pymatgen can parse output files (into structured data or as Python objects with callable functions) and can also perform high-level data analyses such as phase diagram creation or plotting. Separately, we employ code called “builders” that collect, reorganize, and post-process raw data into separate database collections that are more amenable to analysis than raw data collections. For example, a builder might collect together all calculated results on a single material to build a single summary report (a “material” document) for that compound. A builder might also collect together information from multiple compounds, perform an analysis, and store the results in a database. In service of such processes, we develop and use lightweight libraries to automate, simplify, and ultimately streamline the process of creating MongoDB databases. Our general “builder” code could be useful to any project that needs to perform extract-transform-load (ETL) operations with MongoDB. For example, they can be run in parallel without explicit coding of parallelism by the author. This allows CPU-intensive transformations of the data to run much faster on multi-core machines, which includes most modern hardware (integration with the Message Passing Interface (MPI) standard to enable parallelization across supercomputing resources is in development). Furthermore, facilities in our code for incremental building allow successive builds of source

MongoDB collection(s) to only operate on the records added since the last build, which can save significant amounts of computation time. Overall, the builder framework allows for efficient generation and reliable updating of multiple database collections that are tailored for different types of query and usage patterns.

3.3.2 Software for Data Dissemination: The Web Interface and RESTful API

The Materials Project places a strong emphasis on user experience, user interface design, and ensuring that data is efficiently disseminated so that a wide variety of users are able to apply the data for research, development, and education (Jain et al. 2016b). To this end, we have built an interactive web portal (https://www.materialsproject.org) focusing on the scientist as the end user. This web portal is built using the mature Django web framework (Dja 2015) due to its clean separation of front-end views from the back-end business logic. Django is written in the Python programming language, which eases integration with the pymatgen library and the growing scientific software ecosystem in Python. Django also provides a clear structure for organizing a so-called project into "apps," which maps well to our various interactive views across materials data such as compositional phase diagrams, Pourbaix diagrams, or domain-specific applications such as battery electrode searching. Additionally, Django features robust tools for user management, simplifying procedures for authentication (who someone is) and authorization (what a known someone can access/do). These tools are used, for example, to provide prepublication "sandboxes" for certain user groups within which to explore and perform analyses across private data sets prior to public release.

In order to tighten feedback loops for users searching data and using various functionalities that may not be applicable to all users, we organize our front-end code to asynchronously load both data and additional code using standard Asynchronous JavaScript And XML (AJAX) and Asynchronous Module Definition (AMD) protocols. Our choices of specific libraries for the web interface continue to evolve as trade-offs between established best practices (that are attractive for a system intended for continuous and reproducible use over many years) and emerging standards (that simplify ongoing maintenance and adding features).

Although many exploratory research studies are well suited to a graphical interface such as the one described above, other studies require programmatic access to this database. With this in mind, we have chosen to expose our data through an application programming interface (API) called the Materials API (MAPI) (Ong 2015). MAPI allows users to develop computer programs that can query, process, and download Materials Project data through a well-defined interface. To date, the MAPI has served more than 100 million requests for materials data for over 1500 distinct users.

APIs are used extensively throughout technology and software development. They serve to clearly and explicitly define a protocol for communicating with a piece of software or other system that is accessed programmatically. At the time of this writing, the most common framework for APIs that operate over the Internet is REpresentational State Transfer (most commonly referred to as REST). The most simple use case for REST APIs is to map web uniform resource identifiers (URIs) to data (similar to how a computer's file system maps data to directories and filenames). In RESTful systems, information is organized into resources, each of which is uniquely identified via a uniform resource identifier (URI). In the case of Materials Project, each document or object (such as a computational task, crystal structure, or materials property) is represented by a URI (see Fig. 8 for an example) and an HTTP verb that can act on that object (GET, POST, PUT, DELETE, etc.). In most cases, this action returns structured data that represents the object, e.g., in the JavaScript Object Notation (JSON) format. For example, to request energy data (as calculated using VASP) on all Fe_2O_3 compounds in the Materials Project database, the URL shown in Fig. 8 could be constructed according to the protocol specified in the MAPI. We note that since MAPI is a RESTful system, users can interact with the MP database regardless of their computer system or programming language (as long as it supports basic HTTP requests.)

REST APIs allow for more powerful behavior to be seamlessly integrated alongside such basic information retrieval. For example, unique strings of characters associated with specific users (called API keys) can be used to manage access to resources. This is done by implementing the API in such a way that requires users to include their API keys in requests they make to the system and then implementing controls on the back end of the system to handle permissions and activity logging. RESTful APIs can also accept filtering parameters or other variables within requests to give users greater control over what they send or receive from a database. Moreover, URLs can be linked to more than just static resources; they can also point to back-end functions that enable interaction between a user program and MP. An example might be linking a URL such as "https://www.materialsproject.org/rest/v1/materials/snl/submit" to a function registering a request to compute a desired structure embedded in an http POST parameter.

Use of such an API offers a number of advantages. First, users do not have to be concerned with the actual architecture of the database they are interacting with or the details of its implementation since the API serves as a kind of "middleman" in

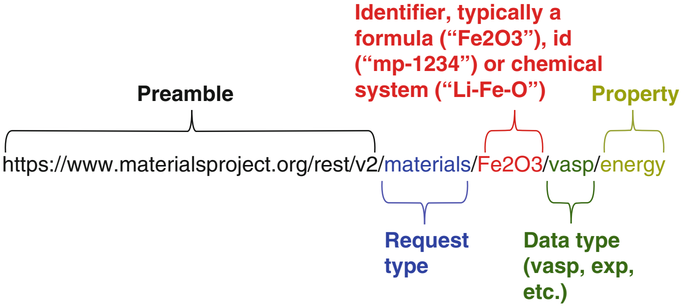

Identifier, typically a formula ("Fe2O3"), id ("mp-1234") or chemical system ("Li-Fe-O")
Preamble
Property
https://www.materialsproject.org/rest/v2/materials/Fe2O3/vasp/energy
Request type
Data type (vasp, exp, etc.)

Fig. 8 An example of the URL structure for the Materials API. (Reprinted from Ong (2015) with permission from Elsevier)

the process. From a user's perspective, the procedure to interact with the database is consistent over time, freeing the development team to make back-end changes without impacting the user's mode of interaction with the data. In addition, access to the database is system-agnostic. Anyone can develop an application in whatever environment they wish on top of the API with the confidence that it will be compatible with the MP database. Moreover, the data that users receive is always up-to-date, with no extra effort on their part, and the capabilities of the API can be seamlessly improved over time to give users access to even more powerful queries and analyses without creating new procedures for their use.

Although RESTful APIs can be intimidating to novices, they can be made more user-friendly by making the URL scheme explorable and hiding complexity through intermediate software layers. For example, a high-level Python interface to the MAPI called the MPRester is provided in the pymatgen (Ong et al. 2013) code base that allows users to obtain properties like crystal structure or electronic band structure using Python functions rather than explicit HTTP requests. We note that, whenever possible, the main Materials Project web site front end also avoids direct database queries and uses MAPI to query and access data in a way that is more maintainable and less prone to failure than custom interactions with the back-end software.

4 User Applications of the Materials Project to Research and Design Problems

Since its release, users of the Materials Project have used its data and tools in several hundred research studies (as highlighted in a previous review Jain et al. 2016a). In this section, we describe several recent examples and outline general strategies that have emerged in the literature for screening and designing materials for specific applications. While several of these studies involve active MP collaborators (Dagdelen et al. 2017; Yan et al. 2017; Chen et al. 2016; Zimmermann et al. 2017), a large fraction of the most recent studies that we found through a Web of Science search are from users that are not involved in Materials Project (Sendek et al. 2017; Shi et al. 2017; Ashton et al. 2017; Cheon et al. 2017; Choudhary et al. 2017; Lau et al. 2017; Shandiz and Gauvin 2016). This latter class of users perhaps most clearly demonstrates that it is possible to accelerate the research and design of new materials by generating and sharing materials information with the research community.

4.1 Phase Diagrams and Compound Stability

In studies that aim to improve our understanding based on experimental evidence or to synthesize new materials for a given application (Bayliss et al. 2014; Krishnamoorthy et al. 2015; Martinolich and Neilson 2014), generating phase equilibrium data is among the most frequently used MP capabilities . For example, Bayliss et al. conducted a study on a sodium-doped strontium silicate material that had been claimed to feature remarkably high oxide ion conduction (Bayliss et al. 2014). By combining experiments (neutron powder diffraction, two-point AC impedance spectroscopy, time-of-flight secondary ion mass spectrometry) and DFT calculations, they could show that the conductivity was lower than previously reported and that the high energetic cost of oxygen vacancy formation was the underlying reason. Data from MP was used to cross-check the study's results for the DFT-PBEsol-derived major limiting phases of the $SrSiO_3$ decomposition.

Shi and co-workers (Shi et al. 2017) employed a high-throughput DFT screening approach for stable delafossite and related layered phases of composition ABX_2, where A and B are any elements from the periodic table and X a chalcogen (O, S, Se, and Te). From the initial materials set of 15,624 compounds, 285 were found to be within 50 meV/atom from the convex hull. While the majority of these structures are contained within the Materials Project database, the authors highlight that 79 of these stable systems are absent. This underscores that crystal structure databases such as MP still have considerable growth potential in terms of compound completeness and highlights the role that user-based compound submissions (through the MPComplete service, cf., Sect. 6.1) could play in extending such databases.

A similar example is the work by Krishnamoorthy et al. (2015), who used a high-throughput DFT-based screening to identify lead-free germanium iodide perovskites that could be used for light harvesting. The researchers computed the PBE bandgaps of 360 AMX_3 compositions, uncovering 9 interesting candidates. MP phase equilibrium data were used to further reduce the list by requiring that the materials be thermodynamically stable against decomposition to simpler binary phases. Three materials were left from the computational screening, $RbSnBr_3$, $CsSnBr_3$, and $CsGeI_3$, of which the latter was successfully synthesized and characterized. We refer the interested reader to a previous review article Jain et al. (2016b) for further examples of experimental studies conducted using MP-calculated phase diagrams.

4.2 Crystal Structure Analysis, 2D Materials, and Machine Learning

The large corpus of data available in the Materials Project can serve as a test bed for the development of new algorithms for processing of crystallographic data. This is the case for Ashton and co-workers (Ashton et al. 2017) who developed a topology-scaling algorithm to identify the dimensionality of a given crystal structure. They used the algorithm to search the MP database for materials that could be prospective 2D materials; 826 stable layered materials were identified, of which 680 were predicted to be feasible 2D material candidates based on the calculated exfoliation energy.

Similarly, Cheon and co-workers (Cheon et al. 2017) present an algorithm that can identify the dimensionality of weakly bonded subcomponents of a three-dimensional crystal structure. They apply this algorithm to $>$50,000 MP materials and identify 1,173 two-dimensional layered materials as well as 487 weakly bonded one-dimensional molecular chains, representing an order of magnitude increase in the number of identified materials. Furthermore, 325 of these materials were suggested to be piezoelectric monolayers.

Interestingly, by specifically exploiting a weakness in typical DFT calculations that the dispersion forces are not well accounted for and thus lattice parameters of layered materials are often significantly inaccurate, Choudhary and co-workers (Choudhary et al. 2017) were able to identify two-dimensional material candidates. To this end, the authors required that the deviation between lattice constants from experiments and (mainly) MP database be $\geq 5\%$. In order to validate their approach, the authors used an accepted criterion based on the exfoliation energy and found that 88.9% of their predictions met this test.

Many structure-property relationships that form the basis of design rules in materials science are based on information pertaining to the local coordination environment. Therefore, it is highly desirable to have tools that effectively and efficiently identify basic local structural motifs such tetrahedra, octahedra, bcc, fcc, and hcp environments. Zimmermann et al. provided classification criteria for these motifs that are based on local structure order parameters, which were used to automatically identify these motifs in the entire Materials Project database. Additionally, these tools may also lead the way to alternative structure matching avenues (Zimmermann et al. 2017).

The abundance of data in the Materials Project also provides an opportunity to develop new machine learning (ML) techniques for modeling materials properties and for better understanding structure-property relationships. One such example of this appears in work from Faber et al. (2015) aimed at developing representations of periodic systems adaptable to ML models. In this study, 4000 structures from the Materials Project were used to evaluate the generalization error in the predicted formation energy based on three different crystal structure representation schemes and using kernel ridge regression, revealing that a sine matrix approach intended to simulate an infinite Coulomb sum was superior in its efficiency and accuracy.

Similarly, de Jong et al. (2016) demonstrated a machine learning approach to predicting elastic moduli of k-nary compounds that was effective over a highly diverse set of chemistries. More specifically, this study used gradient boosting machine local polynomial regression (GBM-Locfit) over the MP elastic tensor data set to determine a set of relevant descriptors and to derive elastic modulus predictions. Ultimately, this model was leveraged to estimate the Vickers hardness of the entire MP materials library, enabling a rapid search for superhard materials.

In most cases, easily retrievable or computable data such as the space group, composition, and the density are used in order to predict more complex properties such as the formation energy or the elastic tensor. Shandiz and Gauvin pursued the inverse route (Shandiz and Gauvin 2016): the authors conducted a classification study of 339 materials from the MP database that are potential Li-ion silicate cathodes (general composition: Li-Si-(Mn, Fe, Co)-O). In particular, they tested whether or not they could predict the crystal system (monoclinic, orthorhombic, or triclinic) based on features that were derived from both the input crystal structure and DFT outputs: the unit cell volume, the bandgap, the number of sites in the unit cell, the formation energy, and the energy above the convex hull. Pair correlation plots of these features indicated that there was no exploitable direct correlation between any of the features and the crystal system. Decision tree-based methods (random forest and extremely randomized trees) were shown to yield prediction accuracies of up to 75%, and these methods performed better than linear and shrinkage discriminant analysis, respectively, artificial neural networks, support vector machines, and *k*-nearest neighbor classification.

4.3 Screening Materials for Applications

Perhaps the most consistent materials screening strategy that has emerged from the data on the Materials Project is that of filtering materials on successively tighter criteria appropriate to a given application space. In this approach, a filter common to most applications is typically on stability via ΔE_{hull}, which can provide an indicator of whether a compound will be experimentally feasible. As illustrated in Fig. 9, successive filters in turn reduce the number of materials to be considered until it reaches a tractable quantity for follow-up with either more sophisticated calculations or for experimental inquiry.

This was the approach taken by Sendek and co-workers (Sendek et al. 2017) who searched for new candidate materials that could be used as solid-state electrolytes for lithium-ion batteries. The authors screened 12,831 Li-containing compounds from the Materials Project to filter those with high structural and chemical stability, low electronic conductivity, and low cost, thus, eliminating 92.2% of their initial materials. Subsequently, an ionic conductivity classification model, which was

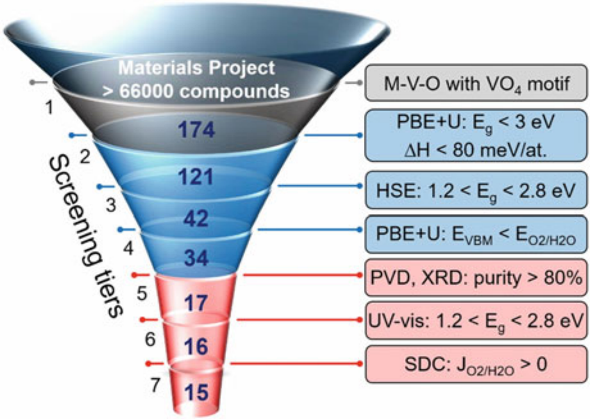


Fig. 9 The "funnel" approach to materials screening through successive criterion filtering as applied to designing materials for solar fuel photoelectrocatalysis by Yan et al. (2017). Such approaches start with a large list of potential candidate materials and use a series of criteria (generally of increasing cost or complexity) to reduce the space of possibilities. (Reproduced from Yan et al. (2017); copyright 2017 National Academy of Sciences)

trained on 40 crystal structures and associated measurements from literature, reduced the list of interesting candidates down to only 21 materials. In the latter step, the consideration of a multi-descriptor model over single-descriptor functions was critical to achieve predictive power. Many of the remaining 21 materials have not yet been studied experimentally, which hence offers new opportunities for experimental electrolyte research.

Identifying structurally similar compounds for the purpose of screening structure-sensitive properties and classifying materials has also begun to emerge as a screening strategy and design paradigm. Dagdelen and co-workers (Dagdelen et al. 2017) demonstrate such a screening procedure for predicting new auxetic materials (compounds with negative Poisson's ratios). The authors systematically screened the entire MP database via the Materials Project's REST API and compared each structure to $\alpha-$cristobalite SiO_2, one of the only inorganic crystalline materials previously known to exhibit a negative homogeneous Poisson's ratio. By coupling pymatgen's structure matching algorithm (which can match structures within a user-defined tolerance irregardless of crystal setting, supercell size, or composition) with more conventional screening strategies, 30 likely candidates were gleaned from over 65,000 structures. The full elastic tensor of each candidate was then calculated and their Poisson's ratios subsequently derived. Of these 30 structures, 3 were found to be homogeneously auxetic, and an additional 9 were found to exhibit near-zero homogeneous Poisson's ratio, with experimental confirmation ongoing.

An example of *in silico* screening with the Materials Project that has led to experimentally confirmed materials discovery was presented by Lau et al. (2017). These authors searched for promising chemical looping air separation (CLAS) materials in the MP database through successive criterion filtering ("funnel" approach). Specifically, the applied search filters included (i) restricting binary and ternary compounds, (ii) identifying compounds that can undergo oxidation reactions (at this step, the phase diagram app was employed), and (iii) restricting the temperature and oxygen partial pressure ranges in which the oxidation reactions would be carried out to sensible limits. The approach resulted in 5,501 tentative compounds and 20,861 relevant redox reactions. Since the reduction enthalpy and the gravimetric O_2 capacity (Fig. 10) did not reveal any exploitable trends, the authors had to employ a more heuristic route to reduce the candidate list. First, they required the reaction complexity and the total number of phases present in the reactions to be minimal, yielding 292 materials. Second, they decreased the number further to 108 by excluding compounds with expensive and toxic materials as well as reactions that involved non-oxides after reduction. From the remaining materials, they picked the ABO_3 perovskites because of their flexibility in oxygen stoichiometry without large structure changes and the ease of synthesizing perovskites in general. They subsequently synthesized and characterized $SrFeO_{3-\delta}$, which has emerged as a promising CLAS candidate due to its thermodynamic and excellent cycling stability as well as its resistance to carbonation over the temperatures of operation (Lau et al. 2017). Further examples of compound discovery with the Materials Project can be found in prior reviews (Hautier et al. 2012; Jain et al. 2016c).

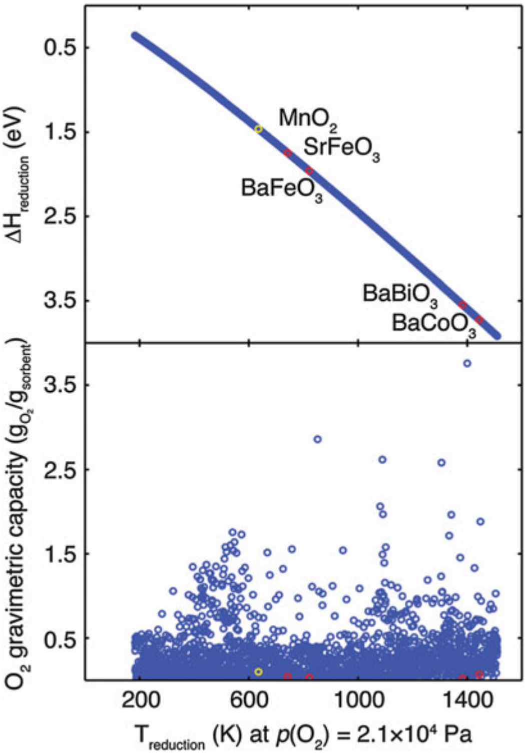


Fig. 10 An example of materials screening for chemical looping air separation application using MP data. Each data point represents a reduction reaction for each distinct compound with the largest μ_{O_2}. The predicted reduction temperature at $p_{O_2} = 2.1 \times 10^4$ Pa for each reaction is plotted against $\Delta H_{\text{reduction}}$ (upper) and O_2 gravimetric capacity (lower). (Reproduced (Adapted or in part) from Lau et al. (2017) with permission of The Royal Society of Chemistry)

5 Outreach

Starting in 2016, the Materials Project has held annual workshops that have hosted more than 100 attendees from around the world. The workshops cover use of the Materials Project web site as well its software stack for performing and analyzing high-throughput calculations. Tutorials for the workshop utilized Jupyter (Ragan-Kelley et al. 2014) notebooks, which are a form of computer document that mixes formatted text, editable code, and interactive plots to illustrate a procedure. Participants were given the option to install the various codes to their own systems or to interact with a pre-installed environment configured using JupyterHub and Docker Swarm. The latter option allowed participants to focus on learning to use the software stack and left the details of individualized setup for later. All tutorials and course materials from these workshops are available online (Mathew et al. 2016; Winston et al. 2017).

Apart from the annual workshops, the Materials Project interacts with users in various ways. For example, MP has created YouTube videos with tutorials on all aspects of the web site, its various apps, and use of the API, which have had a total of over 30,000 views at the time of this writing. The Materials Project web site maintains a general-purpose discussion board (https://discuss.materialsproject.

org/) that has over 100 monthly active users, over 400 posts, and nearly 200 “likes” (whereby users quickly mark the helpfulness of posts) as of this writing. Finally, the MP software stack contains dedicated Google groups and Github issue pages where users and developers of the software can ask questions or get advice on software usage; hundreds of tickets have been resolved thus far.

6 Future of Materials Project

The advances in electronic structure theory, numerical algorithms, computing hardware, and software that have converged to make it possible to develop electronic structure databases are truly stunning. By leveraging these advancements, the Materials Project has computed millions of materials properties (e.g., electronic band structure, thermodynamic properties, mechanical properties, dielectric properties) across tens of thousands of materials, organized that information into searchable databases, and built rich web applications around the data in a way that would not have been possible a decade ago. The future efforts of the Materials Project will concentrate on further empowering the tens of thousands of scientists who design and develop new materials. Here, we describe some possible future developments to enhance property coverage, improve community data import capabilities, and provide an online materials design environment that leverages modern data analytics techniques.

6.1 Data Set Expansion

The Materials Project is continually generating new materials data at a rate of several tens of million CPU hours per year to expand the scope of its database. In the future, the Materials Project will expand in both breadth and depth: a greater variety of materials systems will be investigated, and more information will be calculated about individual materials. In terms of breadth, the Materials Project will expand to more completely encompass crystals with site disorder, i.e., partial site occupancies. The Materials Project will also continue its efforts and partnerships to expand its offerings of data on molecular, i.e., nonperiodic, systems. Finally, the Materials Project expects to play a more active role in not only computationally characterizing known materials but aiding experimentalists in the search for new materials yet to be discovered. In terms of depth, the Materials Project is expanding its library of computational workflows so that more information is available for each material in the database. Active areas of effort include phonon calculations and finite temperature properties, interfaces, spectroscopy, defects, and mapping relations between mechanical, thermal, and electrical effects. Furthermore, the Materials Project will leverage new advances in DFT functionals that make it possible to improve accuracy while still being computationally efficient for high-throughput computation.

This data set expansion will require orders of magnitude more computing resources than is currently employed. The Materials Project will continue to use “crowdsourcing,” i.e., using input from the user community, as a method to prioritize various aspects of data set expansion. For example, the MPComplete service of Materials Project already allows users to both suggest new compounds for calculation and vote on compounds on which to prioritize more computationally expensive workflows. MPComplete then automatically integrates the results of each calculation with MP’s core data set.

6.2 Beyond Simulations: Community-Contributed Materials Data

MP has become a worldwide resource for the materials sciences community, with over 40,000 users who rely on the portal as a trusted source to accelerate their research. This presents an opportunity to broaden the scope of MP’s mission to also include assisting researchers disseminate their own data sets (whether computational or experimental) to the larger community of materials scientists. Thus, MP would serve not only as a hub for centrally generated computational data but would also host and distribute a variety of data sets generated by research groups worldwide. This will also give users of MP a more holistic picture of a compound because they would be presented with both computational and experimental information from a variety of techniques.

For this purpose, we soft-released our general contribution framework, MPContribs (Huck 2016c, 2017; Huck et al. 2015a,b, 2016a,b), as a sustainable solution for well-curated data management, organization, and dissemination in the context of MP. Data as contributed through this framework as well as provenance and citation information for the contributors can be viewed on the MP web site. Early adopters are experimenting with MPContribs as a potential dissemination and hosting platform for their data, expanding the scope of data available through MP.

About a dozen early adopters have released landing pages to their contributed data sets on https://materialsproject.org/mpcontribs. Figure 11 highlights the landing pages for external studies of MnO2 phase selection, GLLB-SC bandgaps, dilute solute diffusion, and Fe-V-Co magnetic thin films. The last of these is based on data measured at the Advanced Light Source at Lawrence Berkeley National Laboratory, whereas the others are computationally derived. These landing pages can serve as interactive versions of the accompanying journal publications and allow research studies to be more easily reproduced and expanded upon.

6.3 MPCite: Citing Materials Data in Publications

The US Department of Energy Office of Scientific and Technical Information (OSTI) (Elliot et al. 2016) provides the E-Link service, which allows researchers to submit information about OSTI products (in form of XML meta-data records) and retrieve persistent digital object identifiers (DOIs) to identify it on the World Wide

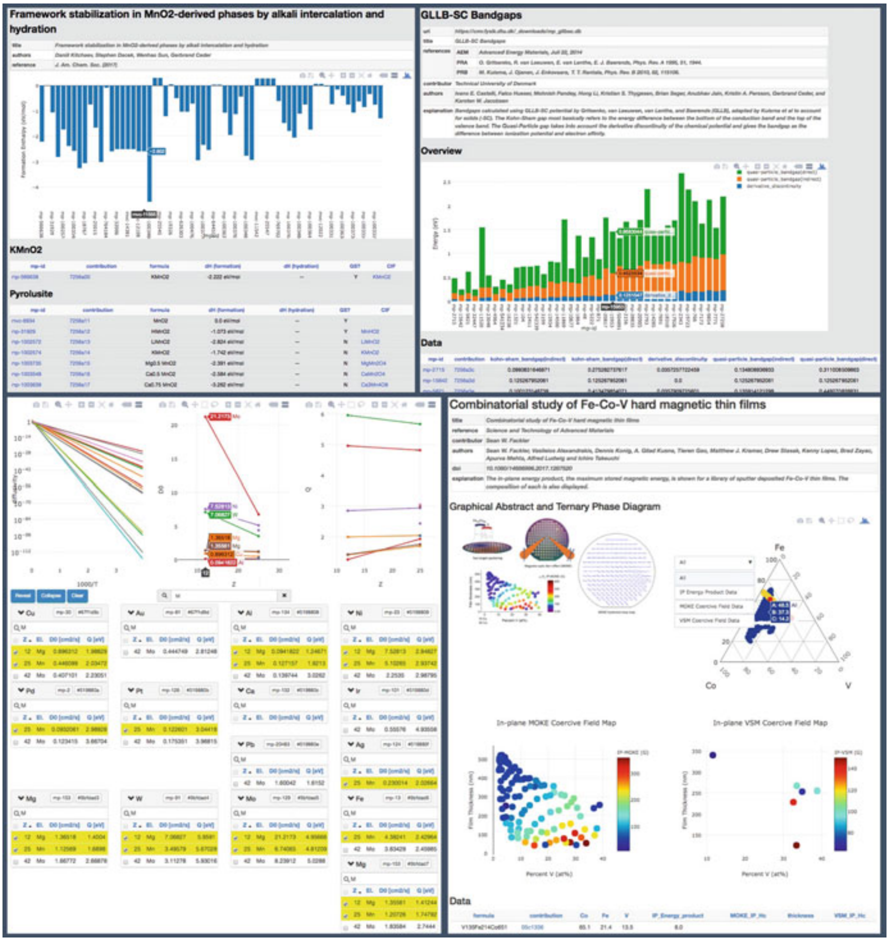

Framework stabilization in MnO2-derived phases by alkali intercalation and hydration
KMnO2
Pyrolusite
GLLB-SC Bandgaps
Overview
Data
Combinatorial study of Fe-Co-V hard magnetic thin films
Graphical Abstract and Ternary Phase Diagram
Fe
Co
V
In-plane MOKE Coercive Field Map
In-plane VSM Coercive Field Map
Percent V (at%)
Data

Fig. 11 Examples of four different landing pages (representing different types of user-generated data sets) submitted to MPContribs

Web. DOIs are most commonly used for referencing and locating journal papers because they provide a unique URL linking to the journal's online landing page with more information about the publication. Our open-source software MPCite (Huck 2016a,b) enables the continuous request, validation, and dissemination of DOIs for all MP compounds. MPCite can also be employed for the assignment of DOIs to non-core database entries such as theoretical and experimental data contributed through MPContribs or user-generated analyses or structural data.

6.4 Data Analytics and Materials Design Environment

The Materials Project aims to not only generate raw data but also to empower users to make the best use of that data. For example, as described previously, we have found that many scientific studies conducted by users employ the "apps" built around the data such as phase diagram plots. As new data capabilities are established, we will continue to build additional apps to enable users to bridge the gap between a simple list of materials data and incorporating that data into a scientific analysis.

The Materials Project will also place additional emphasis on helping users transform underlying data assets into new insights about structure-property relationships. In particular, new capabilities will allow users to formulate complex queries using visual interfaces and perform interactive data analysis and real-time filtering. Users will be able to rapidly iterate on materials design exploration with guidance provided by machine learning algorithms as well as traditional theory calculations. The four components of this vision for a materials design platform are Query, Process, Visualize, and Model/Compute (see Fig. 12). Next, we discuss these components in detail.

Query Today, the Materials Project provides a visual web-based search interface to its underlying databases that is optimal for identifying a set of materials matching a series of constraints. However, many users require more sophisticated data pipelines in which one can visually add or remove filters and inspect the results at multiple points in the analysis or merge results from independent query streams. Such functionality is already possible for those that are capable of writing computer programs to fetch Materials Project data through MAPI, but remains difficult for others. New techniques of allowing users to fetch and interact with the data will be developed in the future so that one is able to call up exactly the desired data using a visual query interface.

Process Once a user has compiled a data set of interest through the query tools, the Materials Project will make it easy for users to add descriptors/features to the data in a way that aids visualization, interpretation, and model building. We envision a system whereby a user can bring up any set of results (e.g., 100 materials of interest) and, by clicking a button, can rapidly generate a library of descriptors such as average electronegativity, local environment type, or polyhedral connection type for every material in the data set. Users will be able to use these descriptors to explore potential structure-property relationships through both conventional data analysis (e.g., visualization, statistical reports) and data mining and machine learning approaches.

Visualize New software libraries and web frameworks such as Dash by Plotly and Crossfilter are making it easier than ever to produce high-quality charts on the web that can be interactively explored and manipulated. Such libraries can enable users

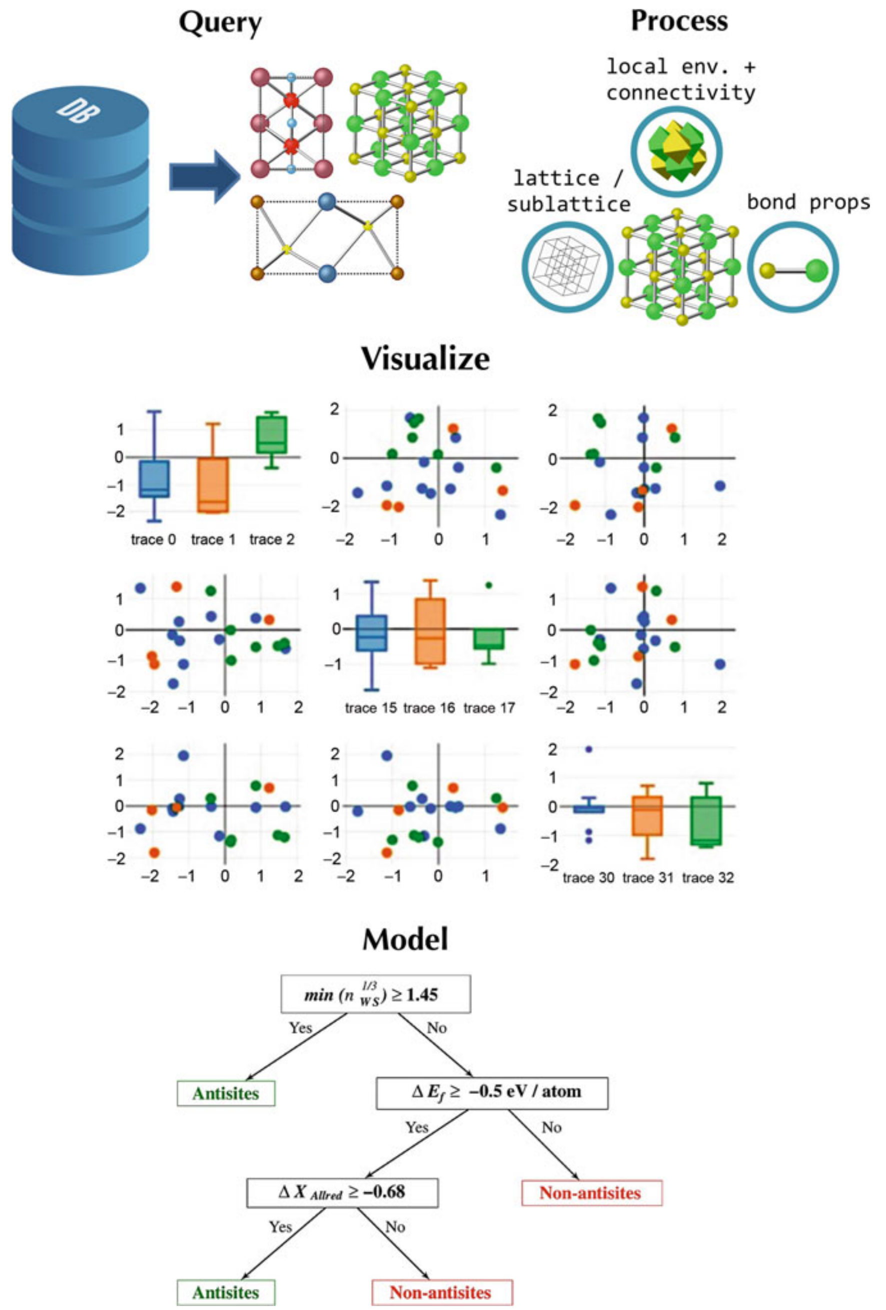


Fig. 12 Four steps in data exploration and modeling for which MP is currently developing new features to assist the user. For a detailed description of these steps, see the main text

to perform a greater fraction of their data exploration on the MP web site itself. As a simple example, a user may decide to produce a standard X-Y scatterplot between two user-chosen properties of interest that additionally allows hovering over specific points to display details of that material. A more complex example would be to include tools for interactive filtering of the data set, with each modification or addition of a filter displaying live histogram charts of the distribution of various materials properties for the materials remaining in the data set. This will allow users to obtain immediate feedback on the distribution of various properties in their data set and visualize how various constraints and filters change this distribution.

Model/Compute After data preparation and exploration, the next step is to take action on the data. One possible action is to produce a model describing the various relationships between materials properties. For example, one may attempt to build a model that relates a structural descriptor such as local environment and a compositional descriptor such as redox active species to a computed output such as battery voltage. Machine learning models are an interesting way forward because, once trained and validated, they can be used to obtain nearly instantaneous feedback on how materials might behave even before any simulations are performed. Thus, such models can serve as surrogates for more complex and time-consuming physics simulations for qualitative estimation and ranking purposes. One can also imagine using these models to guide decisions regarding the computation of new materials.

With these elements in place, a single interactive web session would allow a user to perform sophisticated queries on the data set, automatically generate descriptors that could be useful in forming structure-property relationships, visually explore (and, if necessary, further refine) the data set, produce models that describe the data, and use those models to drive further computations. Indeed, many of these elements are present on the Materials Project today. For example, for materials in which elastic moduli are not yet computed, users can instantaneously obtain an estimate based on machine learning models (de Jong et al. 2016) as well as upvote the full computation based on density functional theory. In the future, this type of mixed usage of both data mining and conventional theory models will become more prevalent and increasingly natural to users.

6.5 Concluding Thoughts

Ab initio simulations have long been powerful tools for understanding and designing materials. With advances in high-throughput computing, it is now possible to create libraries of simulation results that can produce information on materials at a rate far surpassing that possible in the past. Furthermore, advances in software frameworks and web technologies have enabled the dissemination of these results in a barrier-free fashion to thousands of researchers worldwide. The Materials Project is an effort to make use of these advancements to build a valuable resource of materials data as well as software tools that transform the way materials are designed. In addition, the Materials Project aims to make computational materials science a more collaborative process through the development of open-source software and through feedback from experimental groups.

It is an exciting time for theory – never before has there been so much materials data available or the potential of computation to make an impact in materials design higher. Experimentalists and theorists alike have been able to use the Materials Project to conduct scientific and industrial studies in a way that bridges traditional knowledge gaps. These use cases are likely an early sign of a future in which theoretical techniques and large materials databases will be increasingly influential and help to create a new materials design paradigm.

Acknowledgements We thank Professor Gerbrand Ceder, who cofounded the Materials Project and contributed to many of the ideas presented here. We also thank past and present contributors to the Materials Project and the worldwide community of developers that collaborate on the various software libraries that are instrumental to the project.

The Materials Project is funded 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: Materials Project program KC23MP.

We thank the National Energy Research Supercomputing Center (NERSC), a DOE Office of Science User Facility supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231, for providing the primary source of supercomputing time as well as web portal hosting and support. We also thank the San Diego Supercomputing Center for providing additional computing.

Finally, we thank the MP user community for providing feedback and inspiration for the project.

References

Ashton M, Paul J, Sinnott SB, Hennig RG (2017) Topology-scaling identification of layered solids and stable exfoliated 2D materials. Phys Rev Lett 118:106101

Barber CB, Dobkin DP, Huhdanpaa H (1996) The quickhull algorithm for convex hulls. ACM Trans Math Softw 22(4):469–483. http://doi.acm.org/10.1145/235815.235821

Bayliss RD, Cook SN, Scanlon DO, Fearn S, Cabana J, Greaves C, Kilner JA, Skinner SJ (2014) Understanding the defect chemistry of alkali metal strontium silicate solid solutions: insights from experiment and theory. J Mater Chem A 2:17919–17924

Belsky A, Hellenbrandt M, Karen VL, Luksch P (2002) New developments in the inorganic crystal structure database (ICSD): accessibility in support of materials research and design. Acta Crystall Sect B Struct Sci 58(3):364–369

Bray T (2017) The javascript object notation (JSON) data interchange format. STD 90, RFC 8259. https://www.rfc-editor.org/info/rfc8259

Cattell R (2011) Scalable SQL and NOSQL data stores. SIGMOD Rec 39(4):12–27. http://doi.acm.org/10.1145/1978915.1978919

Chen W, Pohls JH, Hautier G, Broberg D, Bajaj S, Aydemir U, Gibbs ZM, Zhu H, Asta M, Snyder GJ, Meredig B, White MA, Persson K, Jain A (2016) Understanding thermoelectric properties from high-throughput calculations: trends, insights, and comparisons with experiment. J Mater Chem C 4:4414–4426

Cheon G, Duerloo KAN, Sendek AD, Porter C, Chen Y, Reed EJ (2017) Data mining for new two- and one-dimensional weakly bonded solids and lattice-commensurate heterostructures. Nano Lett 17:1915–1923

Choudhary K, Kalish I, Beams R, Tavazza F (2017) High-throughput identification and characterization of two-dimensional materials using density functional theory. Sci Rep 7:5179

Cococcioni M, de Gironcoli S (2005) Linear response approach to the calculation of the effective interaction parameters in the LDA + U method. Phys Rev B 71:035105. https://link.aps.org/doi/10.1103/PhysRevB.71.035105

Dagdelen J, Montoya J, de Jong M, Persson K (2017) Computational prediction of new auxetic materials. Nat Commun 8:323

de Jong M, Chen W, Angsten T, Jain A, Notestine R, Gamst A, Sluiter M, Krishna Ande C, van der Zwaag S, Plata JJ, Toher C, Curtarolo S, Ceder G, Persson KA, Asta M (2015a) Charting the complete elastic properties of inorganic crystalline compounds. Sci Data 2:150009. https://doi.org/10.1038/sdata.2015.9, http://www.nature.com/articles/sdata20159

de Jong M, Chen W, Geerlings H, Asta M, Persson KA (2015b) A database to enable discovery and design of piezoelectric materials. Sci Data 2:150053. https://doi.org/10.1038/sdata.2015.53, http://www.nature.com/articles/sdata201553

de Jong M, Chen W, Notestine R, Persson K, Ceder G, Jain A, Asta M, Gamst A (2016) A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds. Sci Rep 6:34256. https://doi.org/10.1038/srep34256, http://www.ncbi.nlm.nih.gov/pubmed/27694824, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5046120

Dja (2015) Django (version 1.8): the web framework for perfectionists with deadlines. https://djangoproject.com

Dozier A, Persson K, Ong SP, Mathew K, Zheng C, Chen C, Kas J, Vila F, Rehr J (2017) Creation of an xas and eels spectroscopy resource within the materials project using feff9. Microscopy Microanalysis 23(S1):208–209

Elliot J, Vowell L, Nelson J, Ensor N, Robinson C, Studwell S, Martin M (2016) U.S. Department of Energy Office of Scientific and Technical Information (OSTI). https://www.osti.gov

Faber F, Lindmaa A, von Lilienfeld OA, Armiento R (2015) Crystal structure representations for machine learning models of formation energies. Int J Quant Chem 115(16):1094–1101. http://doi.wiley.com/10.1002/qua.24917

Gonze X, Jollet F, Araujo FA, Adams D, Amadon B, Applencourt T, Audouze C, Beuken JM, Bieder J, Bokhanchuk A, Bousquet E, Bruneval F, Caliste D, Côté M, Dahm F, Pieve FD, Delaveau M, Gennaro MD, Dorado B, Espejo C, Geneste G, Genovese L, Gerossier A, Giantomassi M, Gillet Y, Hamann D, He L, Jomard G, Janssen JL, Roux SL, Levitt A, Lherbier A, Liu F, Lukacevic I, Martin A, Martins C, Oliveira M, Poncé S, Pouillon Y, Rangel T, Rignanese GM, Romero A, Rousseau B, Rubel O, Shukri A, Stankovski M, Torrent M, Setten MV, troeye BV, Verstraete M, Waroquier D, Wiktor J, Xue B, Zhou A, Zwanziger J (2016) Recent developments in the ABINIT software package. Comput Phys Commun 205:106. https://doi.org/10.1016/j.cpc.2016.04.003, http://www.sciencedirect.com/science/article/pii/S0010465516300923

Grindy S, Meredig B, Kirklin S, Saal JE, Wolverton C (2013) Approaching chemical accuracy with density functional calculations: diatomic energy corrections. Phys Rev B 87(7):075150

Gunter D, Cholia S, Jain A, Kocher M, Persson K, Ramakrishnan L, Ong SP, Ceder G (2012) Community accessible datastore of high-throughput calculations: experiences from the materials project. In: 2012 SC companion: high performance computing, networking storage and analysis, pp 1244–1251. https://doi.org/10.1109/SC.Companion.2012.150

Hart GL, Forcade RW (2008) Algorithm for generating derivative structures. Phys Rev B 77(22):224115

Hautier G, Jain A, Ong SP (2012) From the computer to the laboratory: materials discovery and design using first-principles calculations. J Mater Sci 47:7317–7340

Huck P (2016a) Continuous and high-throughput allocation of digital object identifiers for computed and contributed materials data in the materials project – invited talk at reproducibility mini-symposium of SciPy16. https://youtu.be/bHhuO4EOgEw

Huck P (2016b) MPCite GitHub Repository. https://github.com/materialsproject/MPCite

Huck P (2016c) MPContribs GitHub Repository. https://github.com/materialsproject/MPContribs

Huck P (2017) Materials project: a prime case of software engineering in materials sciences. https://youtu.be/rs8b8HaA3_I

Huck P, Gunter D, Cholia S, Winston D, N'Diaye A, Persson KA (2015a) User applications driven by the community contribution framework MPContribs in the materials project. http://arxiv.org/abs/1510.05727

Huck P, Jain A, Gunter D, Winston D, Persson KA (2015b) A community contribution framework for sharing materials data with materials project. http://arxiv.org/abs/1510.05024

Huck P, Gunter D, Persson K, Cholia S, Morgan D, Wu H, Mayeshiba T (2016a) Effective and interactive dissemination of diffusion data using MPContribs. http://sciencegateways.org/wp-content/uploads/2016/09/Patrick-Huck-2016-11-02_Gateways2016-1.pdf

Huck P, Jain A, Gunter D, Cholia S, Winston D, Persson K (2016b) Materials project as analysis and validation hub for experimental and computational materials data. http://www.mrs.org/technical-programs/programs_abstracts/2016_mrs_fall_meeting_exhibit/tc2/tc2_5_3/tc2_5_06_6

Jain A, Hautier G, Moore CJ, Ong SP, Fischer CC, Mueller T, Persson KA, Ceder G (2011a) A high-throughput infrastructure for density functional theory calculations. Comput Mater Sci 50(8):2295–2310

Jain A, Hautier G, Ong SP, Moore CJ, Fischer CC, Persson KA, Ceder G (2011b) Formation enthalpies by mixing gga and gga + u calculations. Phys Rev B 84:045115. https://link.aps.org/doi/10.1103/PhysRevB.84.045115

Jain A, Ong SP, Chen W, Medasani B, Qu X, Kocher M, Brafman M, Petretto G, Rignanese GM, Hautier G, Gunter D, Persson KA (2015) Fireworks: a dynamic workflow system designed for high-throughput applications. Concurr Comput Pract Exp 27(17):5037–5059. https://doi.org/10.1002/cpe.3505, cPE-14-0307.R2

Jain A, Hautier G, Ong SP, Persson K (2016a) New opportunities for materials informatics: resources and data mining techniques for uncovering hidden relationships. J Mater Res 31(08):977–994. https://doi.org/10.1557/jmr.2016.80, http://www.journals.cambridge.org/abstract_S0884291416000807

Jain A, Persson KA, Ceder G (2016b) Research update: the materials genome initiative: data sharing and the impact of collaborative ab initio databases. APL Mater 4(5):053102. http://aip.scitation.org/doi/abs/10.1063/1.4944683

Jain A, Shin Y, Persson KA (2016c) Computational predictions of energy materials using density functional theory. Nat Rev Mater 1:15004

Kohn W, Sham LJ (1965) Self-consistent equations including exchange and correlation effects. Phys Rev 140:A1133–A1138. https://link.aps.org/doi/10.1103/PhysRev.140.A1133

Kong J, White CA, Krylov AI, Sherrill D, Adamson RD, Furlani TR, Lee MS, Lee AM, Gwaltney SR, Adams TR et al (2000) Q-chem 2.0: a high-performance ab initio electronic structure program package. J Comput Chem 21(16):1532–1548

Kresse G, Furthmüller J (1996) Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. Comput Mater Sci 6(1):15–50. https://doi.org/10.1016/0927-0256(96)00008-0, http://www.sciencedirect.com/science/article/pii/0927025696000080

Kresse G, Hafner J (1994) Norm-conserving and ultrasoft pseudopotentials for first-row and transition elements. J Phys Condens Matter 6(40):8245–8257. http://iopscience.iop.org/article/10.1088/0953-8984/6/40/015

Krishnamoorthy T, Ding H, Yan C, Leong WL, Baikie T, Zhang Z, Sherburne M, Li S, Asta M, Mathews N, Mhaisalkar SG (2015) Lead-free germanium iodide perovskite materials for photovoltaic applications. J Mater Chem A 3:23829–23832

Lau CY, Dunstan MT, Hu W, Grey CP, Scott SA (2017) Large scale in silico screening of materials for carbon capture through chemical looping. Ener Env Sci 10:818–831

Lejaeghere K, Bihlmayer G, Björkman T, Blaha P, Blügel S, Blum V, Caliste D, Castelli IE, Clark SJ, Dal Corso A, de Gironcoli S, Deutsch T, Dewhurst JK, Di Marco I, Draxl C, Dułak M, Eriksson O, Flores-Livas JA, Garrity KF, Genovese L, Giannozzi P, Giantomassi M, Goedecker S, Gonze X, Grånäs O, Gross EKU, Gulans A, Gygi F, Hamann DR, Hasnip PJ, Holzwarth NAW, Iuşan D, Jochym DB, Jollet F, Jones D, Kresse G, Koepernik K, Küçükbenli E, Kvashnin YO, Locht ILM, Lubeck S, Marsman M, Marzari N, Nitzsche U, Nordström L, Ozaki T, Paulatto L, Pickard CJ, Poelmans W, Probert MIJ, Refson K, Richter M, Rignanese GM, Saha S, Scheffler M, Schlipf M, Schwarz K, Sharma S, Tavazza F, Thunström P, Tkatchenko A, Torrent M, Vanderbilt D, van Setten MJ, Van Speybroeck V, Wills JM, Yates JR, Zhang GX, Cottenier S (2016) Reproducibility in density functional theory calculations of solids. Science 351(6280). https://doi.org/10.1126/science.aad3000, http://science.sciencemag.org/content/351/6280/aad3000

Martinolich AJ, Neilson JR (2014) Pyrite formation via kinetic intermediates through low-temperature solid-state metathesis. J Am Chem Soc 136:15654–15659

Mathew K, Ong SP, Winston D, Montoya J, Aykol M, Dwaraknath S, Huck P (2016) Assets for the 2016 materials project workshop. https://doi.org/10.5281/zenodo.1040432

Mathew K, Montoya JH, Faghaninia A, Dwarakanath S, Aykol M, Tang H, Heng Chu I, Smidt T, Bocklund B, Horton M, Dagdelen J, Wood B, Liu ZK, Neaton J, Ong SP, Persson K, Jain A (2017) Atomate: a high-level interface to generate, execute, and analyze computational materials science workflows. Comput Mater Sci 139(Supplement C):140–152. https://doi.org/10.1016/j.commatsci.2017.07.030, http://www.sciencedirect.com/science/article/pii/S0927025617303919

Ong (2015) The materials application programming interface (API): a simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles. Comput Mater Sci 97:209–215. https://doi.org/10.1016/j.commatsci.2014.10.037, http://www.sciencedirect.com/science/article/pii/S0927025614007113

Ong SP, Wang L, Kang B, Ceder G (2008) Li- fe- p- o2 phase diagram from first principles calculations. Chem Mater 20(5):1798–1807

Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier VL, Persson KA, Ceder G (2013) Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. Comput Mater Sci 68:314–319. https://doi.org/10.1016/j.commatsci.2012.10.028, http://www.sciencedirect.com/science/article/pii/S0927025612006295

Ong SP, Qu X, Richards W, Dacek S, Jain A, Hautier G, Kitchaev D (2014) Custodian: a simple, robust and flexible just-in-time job management framework in python. https://doi.org/10.5281/zenodo.11714

Perdew JP, Burke K, Ernzerhof M (1996) Generalized gradient approximation made simple. Phys Rev Lett 77:3865–3868. https://link.aps.org/doi/10.1103/PhysRevLett.77.3865

Perdew JP, Ernzerhof M, Zupan A, Burke K (1998) Nonlocality of the density functional for exchange and correlation: physical origins and chemical consequences. J Chem Phys 108(4):1522–1531

Persson KA, Waldwick B, Lazic P, Ceder G (2012) Prediction of solid-aqueous equilibria: scheme to combine first-principles calculations of solids with experimental aqueous states. Phys Rev B 85:235438. https://link.aps.org/doi/10.1103/PhysRevB.85.235438

Petousis I, Mrdjenovich D, Ballouz E, Liu M, Winston D, Chen W, Graf T, Schladt TD, Persson KA, Prinz FB (2017) High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials. Sci Data 4. https://www.nature.com/articles/sdata2016134

Ragan-Kelley M, Perez F, Granger B, Kluyver T, Ivanov P, Frederic J, Bussonnier M (2014) The jupyter/ipython architecture: a unified view of computational research, from interactive exploration to communication and publication. In: AGU fall meeting abstracts

Raicu I, Foster IT, Zhao Y (2008) Many-task computing for grids and supercomputers. In: 2008 workshop on many-task computing on grids and supercomputers, pp 1–11. https://doi.org/10.1109/MTAGS.2008.4777912

Ricci F, Chen W, Aydemir U, Snyder GJ, Rignanese GM, Jain A, Hautier G (2017) Data descriptor: an ab initio electronic transport database for inorganic materials. Sci Data 4:170085

Sendek AD, Yang Q, Cubuk ED, Duerloo KAN, Cui Y, Reed EJ (2017) Holistic computational structure screening of more than 12,000 candidates for solid lithium-ion conductor materials. Ener Env Sci 10:306–320

Shandiz MA, Gauvin R (2016) Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries. Comput Mater Sci 117:270–278

Shi J, Cerqueira TFT, Cui W, Nogueira F, Botti S, Marques MAL (2017) High-throughput search of ternary chalcogenides for p-type transparent electrodes. Sci Rep 7:43179

Singh AK, Zhou L, Shinde A, Suram SK, Montoya JH, Winston D, Gregoire JM, Persson KA (2017) Electrochemical stability of metastable materials. Chemistry of Materials p acs.chemmater.7b03980, http://pubs.acs.org/doi/abs/10.1021/acs.chemmater.7b03980

Sun W, Dacek ST, Ong SP, Hautier G, Jain A, Richards WD, Gamst AC, Persson KA, Ceder G (2016) The thermodynamic scale of inorganic crystalline metastability. Sci Adv 2:e1600225

Togo A, Tanaka I (2018) Spglib: a software library for crystal symmetry search. ArXiv e-prints: 1808.01590. http://adsabs.harvard.edu/abs/2018arXiv180801590T

Tran R, Xu Z, Radhakrishnan B, Winston D, Sun W, Persson KA, Ong SP (2016) Surface energies of elemental crystals. Sci Data 3:160080. https://doi.org/10.1038/sdata.2016.80, http://www.nature.com/doifinder/10.1038/cgt.2016.38, http://www.nature.com/articles/sdata201680

Van Rossum G et al (2007) Python programming language. In: USENIX annual technical conference, vol 41, p 36

Wang L, Maxisch T, Ceder G (2006) Oxidation energies of transition metal oxides within the GGA + U framework. Phys Rev B 73:195107. https://link.aps.org/doi/10.1103/PhysRevB.73.195107

Winston D, Mathew K, Montoya J, Huck P, Dwaraknath S, Dagdelen J, Liu M, Horton M, Jain A (2017) Assets for the 2017 materials project workshop. https://doi.org/10.5281/zenodo.1040436

Yan Q, Yu J, Suram SK, Zhou L, Shinde A, Newhouse PF, Chen W, Li G, Persson KA, Gregoire JM, Neaton JB (2017) Solar fuels photoanode materials discovery by integrating high-throughput theory and experiment. Proc Nat Acad Sci 114(12):3040–3043. https://doi.org/10.1073/pnas.1619940114

Zhou F, Cococcioni M, Marianetti CA, Morgan D, Ceder G (2004) First-principles prediction of redox potentials in transition-metal compounds with LDA + u. Phys Rev B 70:235121. https://link.aps.org/doi/10.1103/PhysRevB.70.235121

Zimmermann NER, Horton MK, Jain A, Haranczyk M (2017) Assessing local structure motifs using order parameters for motif recognition, interstitial identification, and diffusion path characterization. Front Mater 4:34