Title

User applications driven by the community contribution framework MPContribs in the Materials Project

Permalink
https://escholarship.org/uc/item/8tv9482g

Journal
Concurrency and Computation Practice and Experience, 28(7)

ISSN
1532-0626

Authors
Huck, P
Gunter, D
Cholia, S
et al.

Publication Date
2016-05-01

DOI
10.1002/cpe.3698

Peer reviewed

CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. (2015)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cpe.3698

User Applications Driven by the Community Contribution Framework MPContribs in the Materials Project

P. Huck*, D. Gunter, S. Cholia, D. Winston, A. T. N'Diaye and K. Persson[1]

[1]*Lawrence Berkeley National Laboratory, Berkeley, California 94720*

SUMMARY

This work discusses how the MPContribs framework in the Materials Project (MP) allows user-contributed data to be shown and analyzed alongside the core MP database. The Materials Project is a searchable database of electronic structure properties of over 65,000 bulk solid materials that is accessible through a web-based science-gateway. We describe the motivation for enabling user contributions to the materials data and present the framework's features and challenges in the context of two real applications. These use-cases illustrate how scientific collaborations can build applications with their own "user-contributed" data using MPContribs. The Nanoporous Materials Explorer application provides a unique search interface to a novel dataset of hundreds of thousands of materials, each with tables of user-contributed values related to material adsorption and density at varying temperature and pressure. The Unified Theoretical and Experimental x-ray Spectroscopy application discusses a full workflow for the association, dissemination and combined analyses of experimental data from the Advanced Light Source with MP's theoretical core data, using MPContribs tools for data formatting, management and exploration. The capabilities being developed for these collaborations are serving as the model for how new materials data can be incorporated into the Materials Project website with minimal staff overhead while giving powerful tools for data search and display to the user community. Copyright © 2015 John Wiley & Sons, Ltd.

Received 21 July 2015; Revised 24 August 2015; Accepted 25 August 2015

KEY WORDS: Science Gateways; Databases; User Contributed Data; Materials Science

1. INTRODUCTION

In recent years, an exciting trend has begun to emerge - the integration of computational materials science with information technology (e.g., web-based dissemination, databases, data-mining) to go beyond the confines of any single research group. This development has expanded access to computed materials datasets to new communities and spurred new collaborative approaches for materials discovery. The next step is to leverage open-source development and interactive web-based technologies to enable user contributions back to the community by reporting problems, coding new types of analyses and apps, and suggesting the next set of high potential materials for computation. All the pieces are ready to change the paradigm by which materials are designed.

Materials Project (MP) [1] is a component of the Materials Genome Initiative [2] that leverages the power of high-throughput computation and best practices from the information age to create an open, collaborative, and data-rich ecosystem for accelerated materials design. Started in October of 2011 as a joint collaboration between the Massachusetts Institute of Technology and Lawrence Berkeley National Laboratory, the Materials Project today has partners in more than ten institutions

*Correspondence to: Lawrence Berkeley National Laboratory, 1 Cyclotron Road, M/S 33R0146, Berkeley, CA 94720. E-mail: phuck@lbl.gov worldwide. The Materials Project web site currently receives several hundred unique page views a day and has registered more than 13,000 users in academia, government, and industry.

MP uses high-performance computing (HPC) to determine structural, thermodynamic, electronic, and mechanical properties of over 65,000 inorganic compounds by means of high-throughput ab-initio calculations. The calculation results and analysis tools are disseminated to the public via modern web and application interfaces. However, the materials science research community has a continually increasing supply of experimental and theoretical material properties that are either not yet calculated by MP or outside the scope of MP's data generation efforts. It becomes increasingly important for MP and similar scientific platforms to also enable community-driven submissions, which would extend the scope of the possible applications and improve the integrity/quality of the provided datasets, hence enhancing its value to the user community.

In this paper, we describe how the MP web portal manages user contributions and dissemination of data. In particular, we discuss how our framework for user contributed data, called MPContribs [3], is being extended to handle two major new use-cases: (1) simulation results for large numbers of new objects (i.e., crystals) with very different properties of interest and (2) experimental results that augment the simulation data for existing objects. Both these use-cases have general aspects that recur in other scientific projects, but here we describe the challenges and solutions in the context of two specific projects: for (1), incorporation of millions of nanoporous materials from the Nanoporous Materials Genome Project (NMGC) and for (2), analysis of high throughput spectroscopy data from the Advanced Light Source (ALS) at LBNL. Each use-case illustrates different aspects of the flexibility and performance of the MPContribs framework.

Most of the tools described here (including MPContribs and MPContribsUsers) are released as open source software and are available on the MP Github repository at https://github.com/materialsproject for download.

2. MATERIALS PROJECT PORTAL

The Materials Project portal consists of a web portal that interfaces with a large database of materials, stored in a MongoDB NoSQL database [4]. We chose MongoDB for its rich and flexible data model, which allows us the flexibility to evolve our data structures and schemas as we add new materials and materials properties from different sub-domains of the field [5].

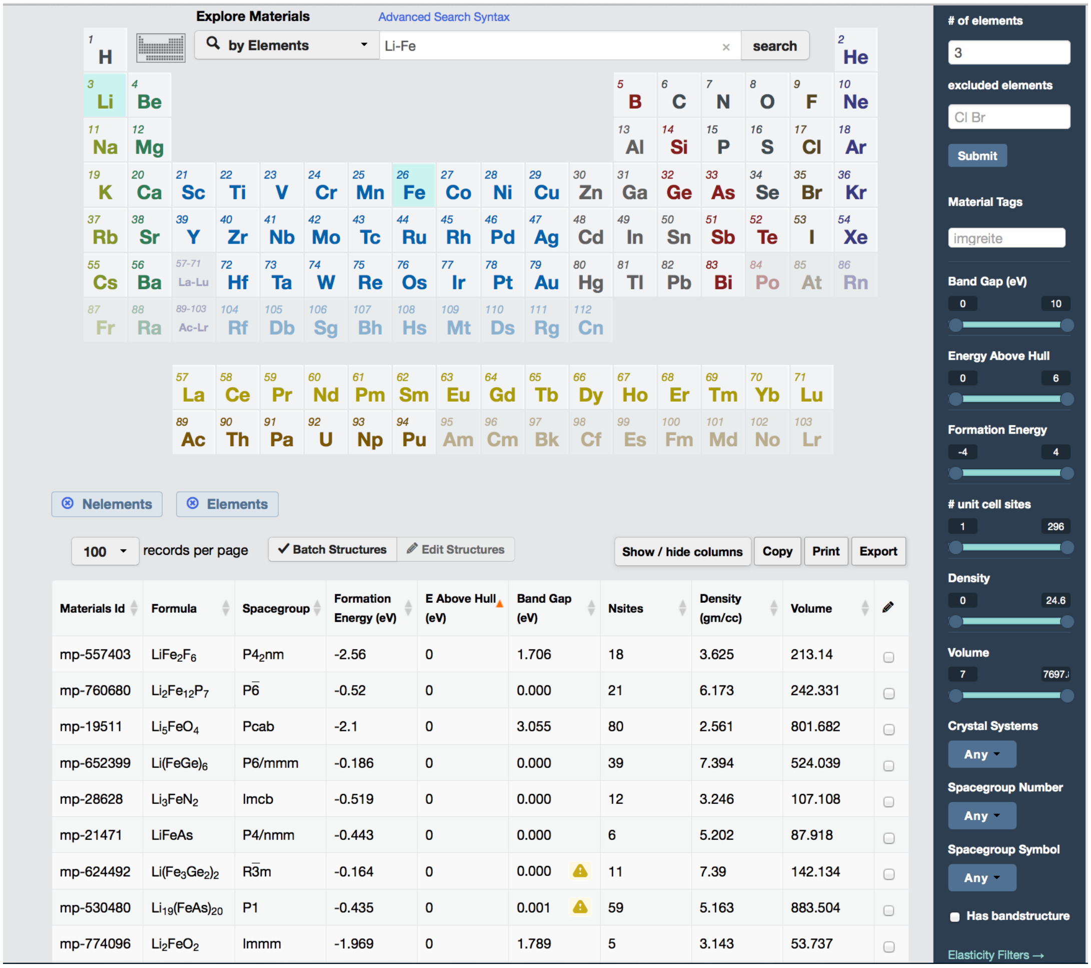
The portal is built using the Python Django web framework. It is hosted at the National Energy Research Scientific Computing (NERSC) Center of LBNL. The data are generated through a series of high-throughput computational runs at NERSC by performing density-functional theory calculations on a large space of possible materials. The results of these calculations are aggregated and indexed in MongoDB, which can then be queried using the web portal or through a REST API [6]. Figure 1 shows the search interface to the MP data - users enter search parameters such as the elements they are interested in, along with specific materials properties and the portal returns a set of matching results. Users can then drill down and look at the data for a specific material that they might be interested in.

The MP portal supports a series of applications that lets users explore materials data. Along with the aforementioned Materials Explorer, there is a Lithium Battery Explorer which adds application-specific search criteria for lithium-ion battery electrode materials. Additional web applications allow users to interactively analyze the dataset. For instance, the Phase Diagram App constructs low-temperature phase diagrams for any chemical system. The Reaction Calculator balances reactions and computes reaction energies between sets of compounds in the database. Other applications include the Pourbaix Diagram viewer, the Crystal Toolkit, the Structure Predictor and the Nanoporous Materials application (which we discuss in detail in Section 4).

MP also supports a REST API, which allows more sophisticated users to generate custom queries, perform bulk downloads of data and interface with the data in a programmatic fashion.

A common concern in scientific collaboration is the sharing of data among privileged collaborators prior to its wider dissemination after validation/publication. MP supports such "sandboxed" data exploration through its REST API as well as with additional web applications

DOI: 10.1002/cpe.3698


Explore Materials
Advanced Search Syntax
by Elements
search
1 H
2 He
3 Li 4 Be 5 B 6 C 7 N 8 O 9 F 10 Ne
11 Na 12 Mg 13 Al 14 Si 15 P 16 S 17 Cl 18 Ar
19 K 20 Ca 21 Sc 22 Ti 23 V 24 Cr 25 Mn 26 Fe 27 Co 28 Ni 29 Cu 30 Zn 31 Ga 32 Ge 33 As 34 Se 35 Br 36 Kr
37 Rb 38 Sr 39 Y 40 Zr 41 Nb 42 Mo 43 Tc 44 Ru 45 Rh 46 Pd 47 Ag 48 Cd 49 In 50 Sn 51 Sb 52 Te 53 I 54 Xe
55 Cs 56 Ba 57-71 La-Lu 72 Hf 73 Ta 74 W 75 Re 76 Os 77 Ir 78 Pt 79 Au 80 Hg 81 Tl 82 Pb 83 Bi 84 Po 85 At 86 Rn
87 Fr 88 Ra 89-103 Ac-Lr 104 Rf 105 Db 106 Sg 107 Bh 108 Hs 109 Mt 110 Ds 111 Rg 112 Cn
57 La 58 Ce 59 Pr 60 Nd 61 Pm 62 Sm 63 Eu 64 Gd 65 Tb 66 Dy 67 Ho 68 Er 69 Tm 70 Yb 71 Lu
89 Ac 90 Th 91 Pa 92 U 93 Np 94 Pu 95 Am 96 Cm 97 Bk 98 Cf 99 Es 100 Fm 101 Md 102 No 103 Lr
Nelements
Elements
records per page
Batch Structures
Edit Structures
Show / hide columns
Copy
Print
Export
Submit
Elasticity Filters →

Li-Fe

100 records per page

# of elements: 3

excluded elements: Cl Br

Material Tags: imgreite

Band Gap (eV): 0 – 10

Energy Above Hull: 0 – 6

Formation Energy: -4 – 4

# unit cell sites: 1 – 296

Density: 0 – 24.6

Volume: 7 – 7697.

Crystal Systems: Any

Spacegroup Number: Any

Spacegroup Symbol: Any

□ Has bandstructure

Materials Id	Formula	Spacegroup	Formation Energy (eV)	E Above Hull (eV)	Band Gap (eV)	Nsites	Density (gm/cc)	Volume	
mp-557403	$LiFe_2F_6$	$P4_2nm$	-2.56	0	1.706	18	3.625	213.14	□
mp-760680	$Li_2Fe_{12}P_7$	$P\bar{6}$	-0.52	0	0.000	21	6.173	242.331	□
mp-19511	Li_5FeO_4	Pcab	-2.1	0	3.055	80	2.561	801.682	□
mp-652399	$Li(FeGe)_6$	P6/mmm	-0.186	0	0.000	39	7.394	524.039	□
mp-28628	Li_3FeN_2	Imcb	-0.519	0	0.000	12	3.246	107.108	□
mp-21471	LiFeAs	P4/nmm	-0.443	0	0.000	6	5.202	87.918	□
mp-624492	$Li(Fe_3Ge_2)_2$	$R\bar{3}m$	-0.164	0	0.000 ⚠	11	7.39	142.134	□
mp-530480	$Li_{19}(FeAs)_{20}$	P1	-0.435	0	0.001 ⚠	59	5.163	883.504	□
mp-774096	Li_2FeO_2	Immm	-1.969	0	1.789	5	3.143	53.737	□

Figure 1. The Materials Explorer application provides an intuitive interface to search for materials in the Materials Project (MP) database. Users select elements from a periodic table and refine their queries to look for properties within specific ranges.

such as the Molecule Explorer for battery electrolyte materials [7], which is not currently available to all MP users. However, another pattern of data sharing, whereby contributors wish not to maintain a privileged sandbox but rather intend to make data publicly available – perhaps after a short incubation period for inspection – is the aim of the MPContribs framework.

3. MPCONTRIBS FRAMEWORK

The current pipeline consists of calculations run directly by MP staff, at NERSC and other computing centers, the results of which are fed directly into the core database using a set of automated builder scripts. MP represents a growing community of materials scientists, and one of the key features towards making this project sustainable is the ability to integrate user contributed data directly into MP. The longer term vision of MP is to be a hub for datasets that come from several different contributors, and are integrated under a common platform.

Towards this end, we have developed a framework called MPContribs that enables users to add their own data to MP, and use the underlying tools to interact with and reference that data.

We briefly summarize the MPContribs framework here; details on the format and tools provided by the MPContribs framework are described in [3] and available online [8, 9]. The three major

DOI: 10.1002/cpe.3698

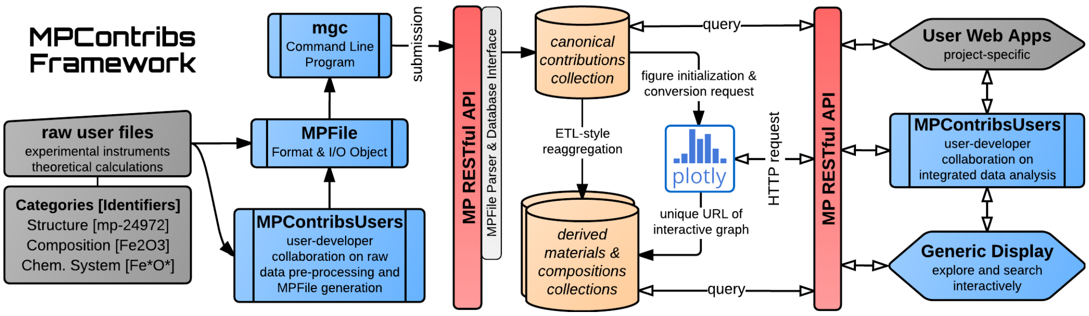


Figure 2. Processing workflow of `MPContribs` framework. The pre-submission stages, with category identification and `MPFile` generation, are followed by internal parser and builder phases and the dissemination to the interactive portal. Contribution submission as well as queries to the databases (orange) and Plotly requests operate through the MP's REST API (red). Blue boxes and the processing behind the REST API "wall" are provided by the `MPContribs` framework. Grey boxes denote tasks and tools under the user's responsibility. For more details, see text.

features of the framework are (1) a text format and associated parsing tools that can flexibly incorporate user-defined data, (2) a RESTful API that can receive and update records in this format to a back-end database, and (3) a display framework for showing the data in the context of MP core databases. The workflow connecting these components is shown in Figure 2. First, the raw user output is identified and gathered. Examples of such data include the nanoporous materials and x-ray absorption (XAS) spectra discussed below, as well as diffusivities computed for various temperatures and solutes, band gaps calculated under alternative conditions, and electronic structure calculations for photovoltaics. Next, the raw data are converted into a concise summary using the MPFile syntax. The result of this phase is submitted, by the user, via the command line program or web interface, both of which operate through the RESTful MP API. The received data are parsed, and the resulting objects are correlated with existing materials data via suitable identifiers. The results are inserted into the core database. Finally, the data are disseminated through the API—and a graphical UI with interactive graphs, tables and tree structures—for exploration and analysis.

As a collaborative extension to the MPContribs core, we created MPContribsUsers [9], a codebase hosted on Github. It contains reusable routines serving a dual purpose: *i*) pre-submission processing to help with the conversion of raw data into MPFiles, and *ii*) analyzing and post-submission processing of contributed data in an integrated manner with MP core data. By hosting this dual-purpose software on Github, the standard open-source contribution workflow[†] can be employed easily with our collaborators which sets up MPContribsUsers as a designated central area separated from the MPContribs core. It consequently allows for the collaborative involvement of the framework developer while simultaneously establishing an archive of example files and scripts to support future contributors. It also enables coordinating between users and reducing duplication of code that deals with similar or identical formats in different domains of materials science.

The Nanoporous Materials Explorer (Section 4) and the X-ray Spectroscopy Application (Section 5) sections illustrate how this is used in practice.

4. NANOPOROUS MATERIALS EXPLORER

The Nanoporous Materials Genome Center (NMGC) [10] discovers and explores microporous and mesoporous materials, including metal-organic frameworks (MOFs), zeolites, and porous polymer networks (PPNs). These materials are important for many applications, including separation media and catalysts in many energy-relevant processes. The NMGC provides a repository of experimental

[†] fork repo, make edits & issue a pull request for merging w/ master repo

DOI: 10.1002/cpe.3698

and predicted structures and associated properties for the rapidly growing scientific communities that are interested in using these materials in energy-relevant technologies.

The Materials Project has integrated porous material data from the NMGC with its software infrastructure. These data are accessible through the production MP website via a *porous materials "app"* [11] that allows for interactive browsing of the multidimensional space of porous materials properties. This graphical search interface allows the user to plot the values of pairs of properties, pan and zoom on the graph, and then list all materials in a region of interest, and select from that list for a detailed per-material view. This interactive view of potentially millions of materials requires its own database of discretized values in each dimension on the website back-end.

The current database and porous materials app were created by custom loading scripts. The indexing of the database for efficient graphical search is performed by a post-processing script in the Materials Project database "builder" framework, which is used for post-processing of electronic structure and other properties. The rest of this section will describe how we are moving from this custom import of NMGC data to using the MPContribs framework to regard all the porous material properties as "contributed data" on the base crystal structure, and to automate the post-processing through the MPContribs maintenance and analysis tools. This allows updates to the data to be performed independently by NMGC collaborators through the increasingly sophisticated interfaces provided by the MPContribs framework. In the longer view, it also will allow multiple types of data to be added to porous materials in the same way they are added to MP's "core" materials. From a devops perspective, this lessens the maintenance burden of keeping up two different databases.

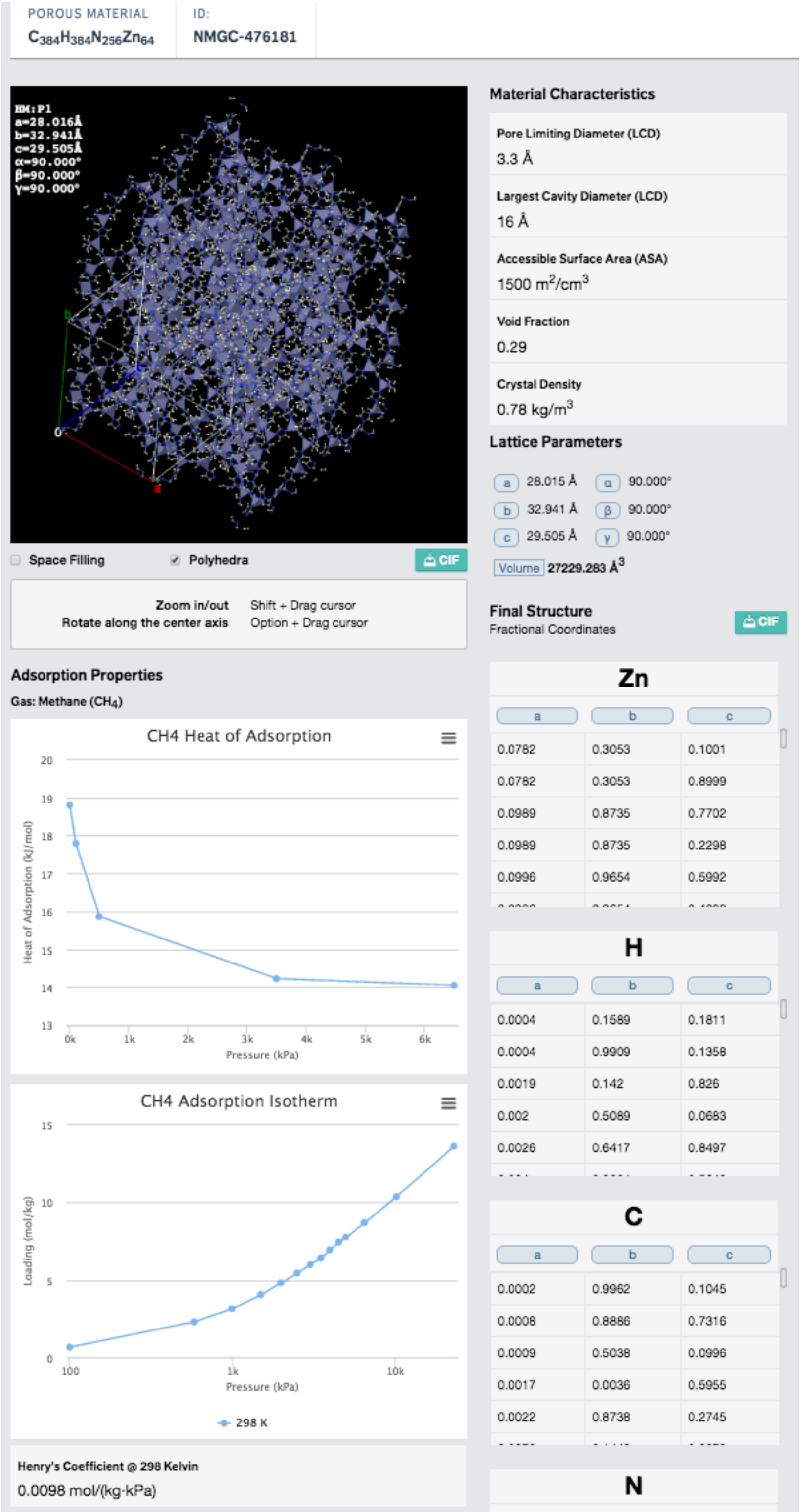
The NMGC data provides an example of how a well-structured custom data format can easily be adapted to MPContribs. In this case, the input data has a crystal structure (giving the position of each atom in the unit cell) that corresponds exactly to core MP crystal structures, and then a set of properties relevant to its application such as the *heats of adsorption*, *crystal density*, *Henry coefficient*, and *isotherms* for the loading of the material for a given gas molecule at varying pressure. All of these data types are easily represented as multi-column tables with attached metadata, and therefore we are able to perform the conversion from our existing data to MPContribs files automatically. The identifiers for the materials will be of the form `<prefix>-<nativeId>`, where the prefix is specific to porous materials (e.g., "por") and the `nativeId` is the identifier that is consistently used by the software that generates the porous materials properties. This allows updates to use these native material identifiers.

An important requirement of the MPContribs integration is that the porous materials remain separate from the core set of materials for the purposes of searching and display, that is, they can only be accessed through the porous materials app. In the short term, we are doing this via some ad-hoc routing using our pre-assigned material identifiers; the NMGC materials will remain in their separate database. But the long-term goal is to avoid special databases for different sets of materials properties, so we are looking into using the existing Materials Project "sandboxing" capability, which can restrict certain materials properties to specific groups. In this case a pseudo-user (i.e. porous materials app), can be used to protect the porous materials data even if they are in the core database. Another potential advantage of this deeper integration is that the core Materials Project explorer app could also, if desired, include porous materials in searches by selected users, by simply changing group memberships.

Another important aspect of the NMGC data that must be preserved is the post-processed discretization of the data values that allows for efficient retrieval and zooming in the search interface. This is easily created with existing scripts for the initial data load, but for updates of the data the scripts will become part of MPContribsUsers to be co-maintained with the framework developer using pull requests (see Section 3). Once the pull request is merged, the framework will call the user's post-processing methods automatically on a set of materials for which he has sufficient permissions. This is achieved by a REST API endpoint exposing routines in the MPContribs core that allow the user to run limited post-processing algorithms based on abstract classes in the core library.

A major advantage of using MPContribs for this data is the ability to use the MPContribs generic display interface, in lieu of the current custom display, for the details of each material. Figure 3

DOI: 10.1002/cpe.3698



POROUS MATERIAL

$C_{384}H_{384}N_{256}Zn_{64}$

ID:

NMGC-476181

□ Space Filling ☑ Polyhedra

Zoom in/out Shift + Drag cursor

Rotate along the center axis Option + Drag cursor

Adsorption Properties

Gas: Methane (CH_4)

Henry's Coefficient @ 298 Kelvin

0.0098 mol/(kg·kPa)

Material Characteristics

Pore Limiting Diameter (LCD)

3.3 Å

Largest Cavity Diameter (LCD)

16 Å

Accessible Surface Area (ASA)

1500 m^2/cm^3

Void Fraction

0.29

Crystal Density

0.78 kg/m^3

Lattice Parameters

a	28.015 Å	α	90.000°
b	32.941 Å	β	90.000°
c	29.505 Å	γ	90.000°

Volume **27229.283 $Å^3$**

Final Structure

Fractional Coordinates

Zn

a	b	c
0.0782	0.3053	0.1001
0.0782	0.3053	0.8999
0.0989	0.8735	0.7702
0.0989	0.8735	0.2298
0.0996	0.9654	0.5992

H

a	b	c
0.0004	0.1589	0.1811
0.0004	0.9909	0.1358
0.0019	0.142	0.826
0.002	0.5089	0.0683
0.0026	0.6417	0.8497

C

a	b	c
0.0002	0.9962	0.1045
0.0008	0.8886	0.7316
0.0009	0.5038	0.0996
0.0017	0.0036	0.5955
0.0022	0.8738	0.2745

N

Figure 3. Current detail view for a porous material. The view template was custom-built to display porous material data. Note that the data on display, beyond those pertaining to the crystal structure, are a set of key-value pairs (tree-like data) and simple 2D plots (derived from tabular data). Such a display may be prototyped by a contributor using the generic front-end of MPContribs prior to optional customization.

DOI: 10.1002/cpe.3698

shows the current detail view for a porous material. The view template was custom-built to display porous material data, but note that the data on display, beyond those pertaining to the crystal structure, are a set of key-value pairs (tree-like data) and simple plots (derived from tabular data). Such a display may be prototyped by a contributor using the generic front-end of MPContribs, making customization optional. A screenshot of a prototype *Contribution Details Page* is shown in [3] featuring hierarchical and tabular data along with interactive and customizable graphs as well as search capabilities.

An important challenge introduced by these materials is bulk operations. The porous materials are generated by a method very different from the core MP database, which is based on a database (the ICSD [12]) culled from many years of theory and experiment. The porous materials can be generated and analyzed very quickly, so that hundreds of thousands of new materials, or sets of new properties on existing materials, can be created. This poses a challenge to the current MPContribs interfaces, which operate on a single material at a time. We are working on a bulk upload and bulk modify interface that will remove the round-trips for large numbers of materials.

5. UNIFIED THEORETICAL & EXPERIMENTAL X-RAY SPECTROSCOPY APPLICATION

As discussed in Section 3, the MPContribsUsers repository collects code (*i*) for data generation to transform output files from common simulation software or from experimental instruments into the MPFile format, and (*ii*) for integrated data analyses to unify MP core data and MPFile contributions.

An example of codes for data generation that would naturally belong in this repository are domain-specific multi-level directory parsers to compile the extracted and derived data into MPFiles. With directories containing electronic calculations based on crystal structures, for instance, the parser can use *pymatgen*'s [13] powerful `Borg` module for high-throughput data assimilation as well as its structure matcher code to match the input structures with existing MP structures of the core dataset. This allows automatic assignment of MP materials identifiers from general-purpose crystal structure definitions and its accompanying calculations.

In the remainder of this section, we discuss the contribution and analysis of measured x-ray absorption spectra as an example which uses both aspects of MPContribsUsers - data generation and data analysis. The result is a MPFile-based and pipelined dataflow as well as analysis workflow established in collaboration with our experimental colleagues.

Soft x-ray absorption spectroscopy (XAS) is a characterization technique which probes the electronic structure of materials. It yields element-, valence-, and lattice site-specific information. At beamline 6.3.1 at the Advanced Light Source (ALS), XAS is being implemented as a high throughput technique [14] yielding datasets with a few hundred spectra per measurement.

Each measurement produces a text file containing data which corresponds to several spectra taken on a certain material and with instrument specific technical information (meta-data). This set of associated instrumental output files is combined with their shared meta-data from the experimentalist's lab book using the MPFile format, into a single output file. This output file also serves as a configuration/input file to apply a sequence of automated pre-processing routines to clean and refine the raw data for a large selection of measured compositions. These pre-processing routines embed their results into the MPFile using composition, e.g. $Ni_{20}Fe_{80}Pt_{10}$, as universal cross-domain identifier. The embedded data can be tabulated datasets, such as XAS spectra or sets of numbers, such as lattice parameters or magnetic moments. Example 5.0.1 shows an excerpt of a MPFile containing meta-data, processing instructions and embedded spectral data measured on Permalloys to study their change in magnetic properties with increasing Platinum doping.

Before official submission of the final MPFile to MP and dissemination of the contributed data to its users, the contained data can be inspected locally using a software tool called *MPFileViewer*. This tool is a Flask web application that simulates a submission using the same internal methods on the local machine. The ability to debug results locally in a realistic environment is integral in the preparation of data for its submission to a well-maintained community database. A screenshot of the XAS data rendered by the MPFileViewer is depicted in Figure 4. The contribution can be

DOI: 10.1002/cpe.3698

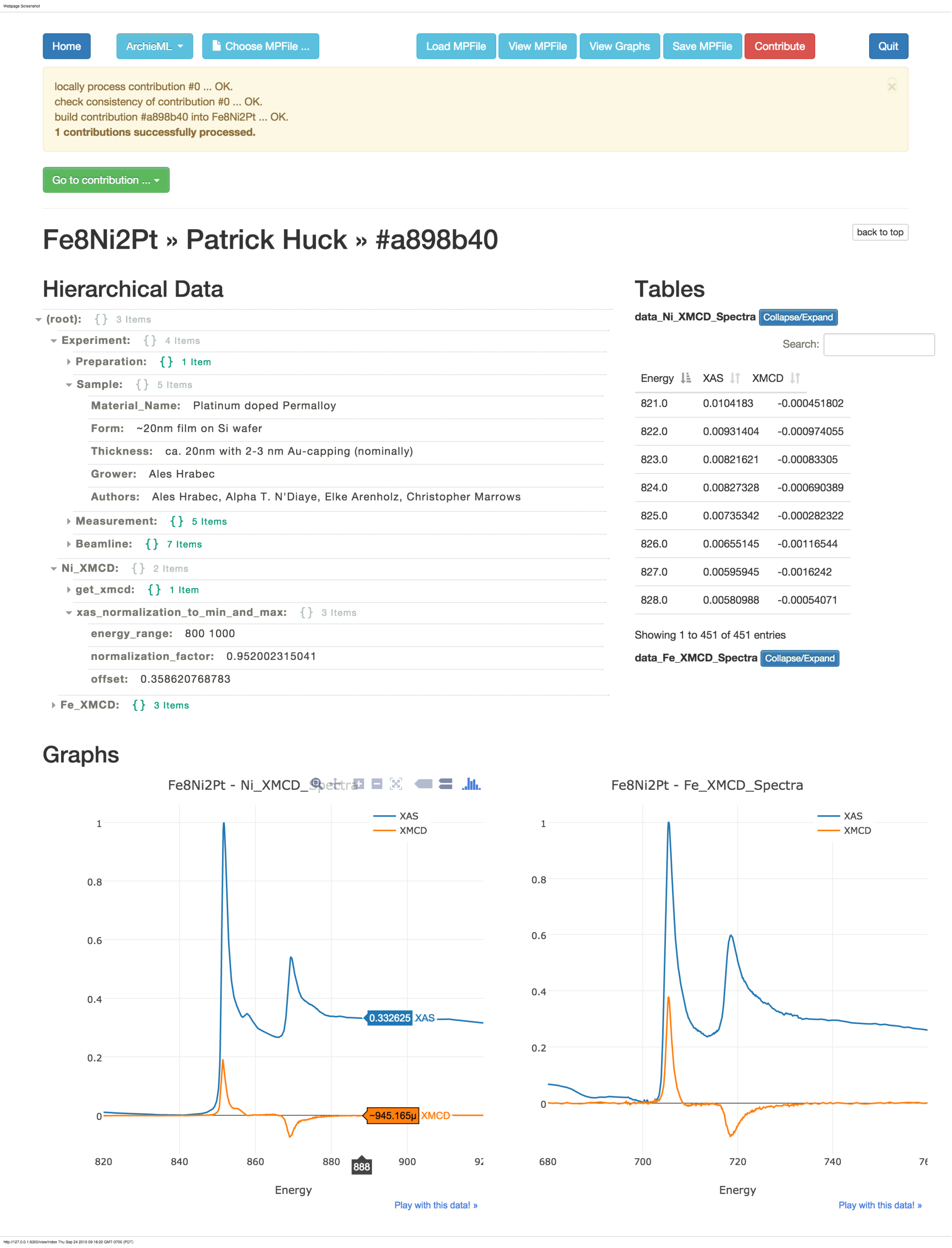
explored via collapsable tree and table structures as well as interactive graphs. The submittors can subsequently use the MP web portal with its dedicated *Contribution Detail Pages* for further analysis and to confirm links to existing projects or update the dataset after additional local analysis steps.

Example 5.0.1: Sample `MPFile` content for measured XAS data on composition Ni20Fe80Pt10. Line numbers and bold section headers were added for presentation purposes. The representation of the MPFile in the `MPFileViewer` is shown in Figure 4. The content of the `GENERAL` section is embedded into all other root-level sections to enable common/shared meta-data. Lines in italic font are added by the pre-processing scripts via `MPContribs` tools based on a skeleton input `MPFile` and instrumental output files (see Section 5).

```
>>> GENERAL
>>>> Experiment
>>>>> Preparation
Description: Sputter deposition
>>>>> Sample
Material Name: Platinim doped Permalloy
Form:  20nm film on Si wafer
Thickness: ca. 20nm with 2-3 nm Au-capping (nominally)
Grower: Ales Hrabec
Authors: Ales Hrabec, Alpha T. N'Diaye, Elke Arenholz, Christopher Marrows
>>>>> Measurement
Detection: total electron yield
Temperature: RT
Orientation: 30 grazing incidence
Date: 2015-06-24
Measured by: Alpha T. N'Diaye
>>>>>> Beamline
Beamline: ALS-6.3.1
Method: Soft x-ray XAS and XMCD
Polarization: circular, positive (ca. 60%)
Magnet Field: 0.8T switching point by point, parallel to x-ray beam
Count Time: 1s
Delay Time: 0.5s
>>>>>>> Monochromator
Exit Slit: 20m
Grating: 600l/mm

>>> Ni20Fe80Pt10
>>>> Ni XMCD
>>>>> get xmcd
energy range: 800 1000
>>>>> xas normalization to min and max
energy range: 800 1000
normalization factor: 0.952002315041
offset: 0.358620768783
>>>> Ni XMCD Spectra
Energy,XAS,XMCD
820,0.0104944,-0.00140602
821,0.0104183,-0.000451802
822,0.00931404,-0.000974055
...
>>>> Fe XMCD
>>>>> get xmcd
energy range: 600 800
>>>>> scaling preedge to 1
preedge range: 690 700
xas- factor: 0.348231766387
xas+ factor: 0.349333591384
>>>>> xas normalization to min and max
energy range: 600 800
normalization factor: 1.00964185927
offset: 0.984095999176
>>>> Fe XMCD Spectra
Energy,XAS,XMCD
680,0.0670848,0.000905727
681,0.0659347,-0.00085033
682,0.0631599,-8.87504e-05
...
```

DOI: 10.1002/cpe.3698



locally process contribution #0 ... OK.
check consistency of contribution #0 ... OK.
build contribution #a898b40 into Fe8Ni2Pt ... OK.
1 contributions successfully processed.

Fe8Ni2Pt » Patrick Huck » #a898b40

Hierarchical Data

```
(root): {} 3 Items
  Experiment: {} 4 Items
    Preparation: {} 1 Item
    Sample: {} 5 Items
      Material_Name: Platinum doped Permalloy
      Form: ~20nm film on Si wafer
      Thickness: ca. 20nm with 2-3 nm Au-capping (nominally)
      Grower: Ales Hrabec
      Authors: Ales Hrabec, Alpha T. N'Diaye, Elke Arenholz, Christopher Marrows
    Measurement: {} 5 Items
    Beamline: {} 7 Items
  Ni_XMCD: {} 2 Items
    get_xmcd: {} 1 Item
    xas_normalization_to_min_and_max: {} 3 Items
      energy_range: 800 1000
      normalization_factor: 0.952002315041
      offset: 0.358620768783
  Fe_XMCD: {} 3 Items
```

Tables

data_Ni_XMCD_Spectra Collapse/Expand

Search:

Energy	XAS	XMCD
821.0	0.0104183	-0.000451802
822.0	0.00931404	-0.000974055
823.0	0.00821621	-0.00083305
824.0	0.00827328	-0.000690389
825.0	0.00735342	-0.000282322
826.0	0.00655145	-0.00116544
827.0	0.00595945	-0.0016242
828.0	0.00580988	-0.00054071

Showing 1 to 451 of 451 entries

data_Fe_XMCD_Spectra Collapse/Expand

Graphs

Figure 4. Screenshot of MPContribs' offline MPFileViewer web app showing the contribution with composition Ni20Fe80Pt10.

We require that all the processing routines used to prepare the input (in this case, the experimental x-ray data) be part of the MPContribs libraries. This way, other user contributions can link to each-other without losing provenance or reproducibility of the work. The data on the MP servers can be embedded in the analysis iterations seamlessly, eventually leading to publishable data which is already linked to relevant work in the field. The division of responsibility for software development

DOI: 10.1002/cpe.3698

allows independent groups, such as submitters with a focus on x-ray spectroscopy, to add their own routines, while keeping broader tools such as the MPFile API and the associated offline inspection and manipulation tools within the purview of the MP team.

The code in MPContribsUsers can be extended to perform advanced functions. For example, one goal of the x-ray spectroscopy application is to compare the X-ray magnetic circular dichroism (XMCD) signals resulting from the XAS spectra with the corresponding theoretical phase diagrams calculated by MP. To achieve such a unified analysis/view of experimental and theoretical data, the post-submission workflow looks something like this: Gather contributed XAS/XMCD spectra and magnetic moments for related compositions from the MP database; Generate an overview PDF including a heatmap of magnetic moments; Pull the according MP phase diagram; Produce an instructive comparison graph. This can be implemented collaboratively in MPContribsUsers, using either scripts or, more attractively, via a customized dynamic web app.

The depth of information in the Materials Project database will enable exciting future scientific directions for this contributed data. We expect that experimental contributors will be able to study their submitted absorption spectra with respect to FEFF [15] calculations to relate the measured compositions to existing crystal structures.

6. CONTRIBUTION PROCESS

The purpose of MPContribsUsers repository is to collect pre- and post-submission codes primarily running on the user's local infrastructure. Since the code would live in MPContribsUsers it is up to the core developer to make use of it (pull-based approach). As such, any new code would need to be vetted if it is merged back into the core. Customized web apps should also be deployed on the user's infrastructure. MPContribs, however, would provide the tools to easily start up web apps that are already integrated with core and contributed MP data.

Currently, we limit contributions to known collorators who are trusted domain experts to produce curated, high-quality datasets. At this stage, our goal is to provide tools for efficient manual data inspection. For instance, we have a staging area for data contributions (MPFileViewer), which allows for human verification before release on the live web site. We have also integrated the plotting tool ”Plotly” with our service, to visualize results and reveal outliers or incorrect data.

In the near future, we plan to move to automated self-verification of the data. This includes a data access API, along with instructions on how users can employ the API to develop unit testing and continuous integration (CI) frameworks. This type of CI functionality will greatly improve the robustness and reliability of user-submitted datasets. We are also investigating other possiblities, like user ratings for datasets. This would enable mechanisms like flagging, removal or black-listing of consistently low-quality datasets/users.

Note, that the materials data format used by MP already has fields for citations and references, thus providing provenance and source tracking for contributed data.

7. RELATED WORK

There are a number of community repositories in Materials Science and related disciplines. Traditionally, these repositories have been limited in scope to crystal structures of measured compounds (acting analogously to the Protein Data Bank [16] for biology), and include efforts such as the Inorganic Crystal Structure Database (ICSD) [17] and Pauling File [18]. More recent efforts, such as AFLOWlib [19] and others [20, 21, 22, 23, 24], typically include additional property data (generally computed) but do not currently have a contribution framework. Some repositories, such as NoMaD [25], iochem-bd [26], and the Materials Data Facility (MDF) [27], provide a shared repository of data, with potentially large experimental and structural data items, but many of these are in the early stages of development and without integrated data analysis capabilities over a core dataset. Our approach, by contrast, focuses on relatively small data contributions that can be joined with the carefully curated core dataset, and extended with an integrated set of data analysis tools to provide a materials design environment.

A number of projects in other sciences, particularly life sciences, share the goals of community-enabling data contributions and analysis in the context of reference datasets. The iPlant Collaborative [28] combines a ”Data Store” with a “Discovery Environment” with the goal of connecting public and private datasets with analysis capabilities; the analysis capabilities include workflows running on remote clusters and private Virtual Machines. The DOE’s Systems Biology Knowledgebase (KBase) [29] provides a unified and interactive analysis environment built on the IPython notebook [30] and a back-end store that can store users’ analyses together with both their data and standard reference microbial data. The large amount of funding spent by federal agencies (KBase and IPlant each receive over $10M/year) on these large cross-cutting projects points to the importance of the capability. Our work is scoped more narrowly (e.g., it does not require an overarching data model or remote execution engines), and is simplified by the use of a core identifer – the `material id` – that represents a common object, a crystal structure, to organize contributions. A genome, the logical analogue of a crystal in biology, is orders of magnitude more complex and variable and cannot be used easily as a central identifier.

Efforts such as ChemSpider [31] emphasize aggregation of data from many individual sources rather than extension of “core” reference data. However, the success of such efforts depends on widely-used standards for identifiers [32] and does not allow users to contribute and maintain custom datasets. Thus, our framework occupies a unique position in the materials data sharing space.

8. SUMMARY & OUTLOOK

With over 13000 (and growing) users, the Materials Project has been very successful in serving as a hub for the materials science community. As MP moves forward, user contributed data will play an important role in expanding the project towards serving a larger community. By providing a standard framework in the form of MPContribs and a standard file structure in the form of MPFile, we are providing a common interface for users to add data to the project. We believe that this provides the basis for other projects that wish to add data to MP and to build custom web applications on top of this data. As illustrated by the use cases above, this is already being used to actively develop new web applications in the MP portal to serve up custom data.

Note that the contributions format and the standard material database objects already keep track of provenance information to track details like data sources and collaborators. Furthermore, our choice of a document oriented database like MongoDB makes it very easy to add new fields to an object to track additional provenance metadata. We expect provenance tracking to become an important requirement for future efforts, and are well positioned to handle this.

In conclusion, MP has delivered significant value to the materials science community by creating a general purpose web gateway for querying and accesing various materials datasets. As newer datasets get incorporated into the MP database, there is a need to be able to build more specialized applications that target specific classes of materials. With our work we are creating the foundation for a common platform that can be used to host multiple applications and datasets for different user communities.

ACKNOWLEDGEMENTS

We especially thank Anubhav Jain for guidance and help with the manuscript.
This work was intellectually led by the Department of Energy’s Basic Energy Sciences program - the Materials Project - under Grant No. EDCBEE and supported by Center for Next Generation Materials by Design, an Energy Frontier Research Center funded by DOE, Office of Science, BES. Work at Lawrence Berkeley National Laboratory was supported by the Office of Science of the U.S. Department of Energy under Contract No. DEAC02-05CH11231. We thank the National Energy Research Scientific Computing Center for providing invaluable computing resources. Finally, we would like to thank all the users of the Materials Project for their support and feedback in improving the project.

REFERENCES

1. Jain A, *et al.*. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mat. 1* 2013; :011 002doi:10.1063/1.4812323.
2. Materials Genome Initiative. URL www.whitehouse.gov.
3. Huck P, Jain A, Gunter D, Winston D, Persson K. A Community Contribution Framework for Sharing Materials Data with Materials Project. *IEEE Computer Society Press* 2015; **Proceedings of 2015 IEEE 11th International Conference on eScience**. URL http://arxiv.org/abs/1510.05024.
4. MongoDB. URL https://www.mongodb.org/.
5. Gunter D, Cholia S, Jain A, Kocher M, Persson K, Ramakrishnan L, Ong SP. Community accessible datastore of high-throughput calculations: Experiences from the materials project 2012.
6. Ong SP, Cholia S, Jain A, Brafman M, Gunter D, Ceder G, Persson KA. The materials application programming interface (api): A simple, flexible and efficient api for materials data based on representational state transfer (rest) principles. *Computational Materials Science* 2015; **97**:209–215.
7. Qu X, *et al.*. The Electrolyte Genome project: A big data approach in battery materials discovery. *Comp. Mat. Sci.* 2015; **103**:56–67, doi:http://dx.doi.org/10.1016/j.commatsci.2015.02.050.
8. Huck P. Materials Project User Contribution Framework. URL https://github.com/materialsproject/MPContribs.
9. Huck P. MPContribs Collaboration Extension. URL https://github.com/materialsproject/MPContribsUsers.
10. Nanoporous Materials Genome Center. URL http://www.chem.umn.edu/nmgc/.
11. Simon C, Kim J, Gomez-Gualdron D, Camp J, Chung YG, Martin RL, Mercado R, Deem MW, Gunter D, Haranczyk M, *et al.*. The Materials Genome in Action: Identifying the Performance Limits for Methane Storage. *Energy Environ. Sci.* 2015; doi:10.1039/C4EE03515A. URL https://materialsproject.org/#search/porous.
12. Bergerhoff G, Brown I, Allen F, *et al.*. Crystallographic databases. *International Union of Crystallography, Chester* 1987; :77–95.
13. Ong SP, *et al.*. PyMatGen: A robust, open-source Python Library for Materials Analysis. *Comp. Mat. Sci.* 2013; **68**:314–319, doi:http://dx.doi.org/10.1016/j.commatsci.2012.10.028.
14. Idzerda YU, Bhatkar H, Arenholz E. Moment mapping of body-centered-cubic fexmn1-x alloy films on mgo(001). *Journal of Applied Physics* 2015; **117**(17):17A721, doi:http://dx.doi.org/10.1063/1.4914127. URL http://scitation.aip.org/content/aip/journal/jap/117/17/10.1063/1.4914127.
15. Rehr JJ, *et al.*. Parameter-free calculations of X-ray spectra with FEFF9. *Phys. Chem. Chem. Phys.* 2010; **12**:5503–5513, doi:10.1039/B926434E.
16. Berman HM, *et al.*. The Protein Data Bank. *Nucl. Acids Res.* 2000; **28**:235–242, doi:10.1093/nar/28.1.235. URL www.rcsb.org/pdb.
17. Belsky A, *et al.*. New ICSD Developments: Accessibility in Support of Materials Research and Design. *Acta Cryst. B58* 2002; :364–369doi:10.1107/S0108768102006948.
18. Pauling File. URL http://paulingfile.com.
19. AFLOW for Materials Discovery. URL www.aflowlib.org.
20. NIST CCCBDB. URL http://cccbdb.nist.gov.
21. CompES Database. URL http://caldb.nims.go.jp.
22. Ortiz C, *et al.*. Data mining and accelerated electronic structure theory as a tool in the search for new functional materials. *Comp. Mat. Sci.* 2009; **44**(4):1042–1049, doi:http://dx.doi.org/10.1016/j.commatsci.2008.07.016.
23. Clean Energy DB. URL http://cepdb.molecularspace.org.
24. Open Quantum Materials Database. URL http://oqmd.org.
25. NoMaD Repository. URL http://nomad-repository.eu.
26. Alvarez-Moreno M, *et al.*. Managing the Computational Chemistry Big Data Problem: The ioChem-BD Platform. *J. Chem. Inf. Model.* 2015; **55**:95–103, doi:10.1021/ci500593j. URL http://iochem-bd.org.
27. Mater. Data Facility. URL www.nationaldataservice.org.
28. iPlant Collaborative. URL www.iplantcollaborative.org.
29. Systems Biology Knowledgebase. URL https://kbase.us.
30. Pérez F, Granger BE. IPython: a System for Interactive Scientific Computing. *Comp. Sci. Eng.* 2007; **9**(3):21–29, doi:10.1109/MCSE.2007.53.
31. ChemSpider. URL www.chemspider.com.
32. Heller S, *et al.*. InChI - the worldwide Chemical Structure Identifier Standard. *J Chem. Inf. Model.* 2013; **5**:7, doi:10.1186/1758-2946-5-7.

DOI: 10.1002/cpe.3698