Data dissemination and material informatics at LBNL

Anubhav Jain, Energy Technologies Area, Berkeley Lab

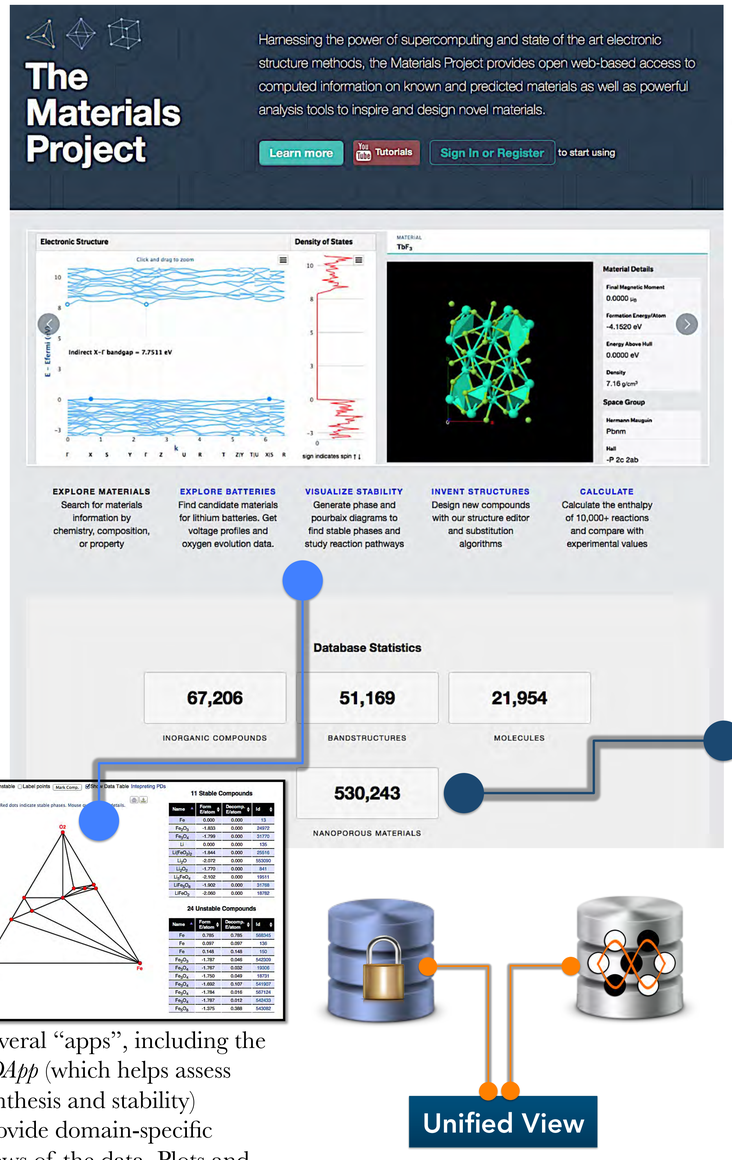
Contact: ajain@lbl.gov

BERKELEY LAB
Lawrence Berkeley National Laboratory

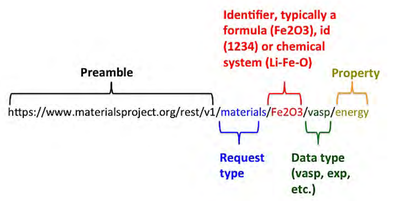
Data Dissemination

The Materials Project

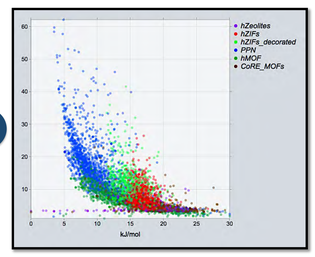
A "science gateway" exposes data to the research community and facilitates knowledge extraction. The Materials Project (www.materialsproject.org) shares simulation data on hundreds of thousands of materials for a community of 22,000 researchers; millions of data points have been downloaded and used in the research articles of its users. A recently introduced feature, MPContribs, accepts data contributions from the user community.



Several "apps", including the *PDApp* (which helps assess synthesis and stability) provide domain-specific views of the data. Plots and analyses from such apps are some of the most heavily used and cited aspects of the Materials Project.



A Representational State Transfer Application Programming Interface (REST API) allows users to download large quantities of data using many popular programming languages. This method uses HTTP requests as a mechanism of accessing data and is employed by many of the top internet companies, including Google, Dropbox, and Twitter.

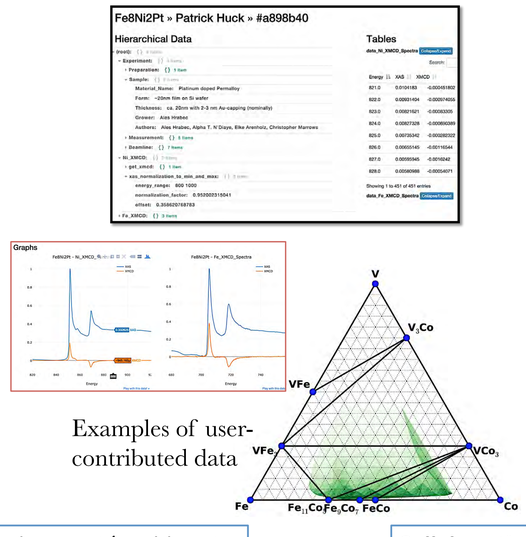


Data on over 500,000 porous materials was contributed by the Nanoporous Materials Genome Center. Users can create custom plots based on properties of interest as well as obtain detailed information about specific materials.

Although the Materials Project is largely an open database, a sandboxing system allows for core data to be stored separately from data generated for proprietary or external projects. Access management is controlled via an API.

MPContribs

Historically, all data in the Materials Project was generated at LBNL via high-throughput computing and density functional theory (DFT) calculations. MPContribs is a new feature that allows users to contribute their own data sets and disseminate them via Materials Project. Online editors and plotting tools allow users to control the look and presentation of their data. In some instances, apps can combine user data with Materials Project core data to create a joint computation-experiment analysis, as shown in the phase diagram on the right.

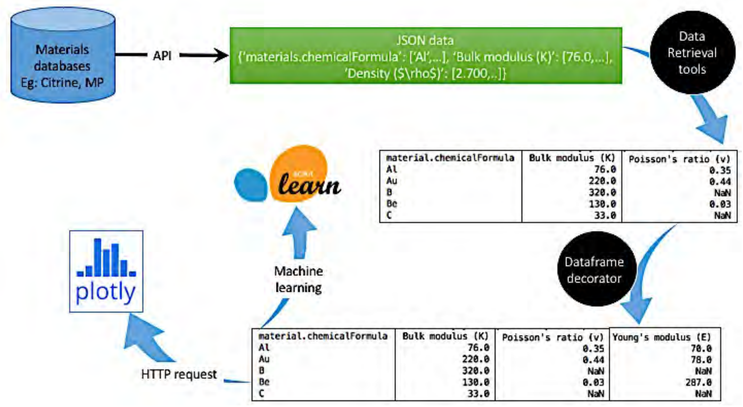


Examples of user-contributed data

Materials Informatics

MIDAS: Materials Informatics & Data Analysis Software

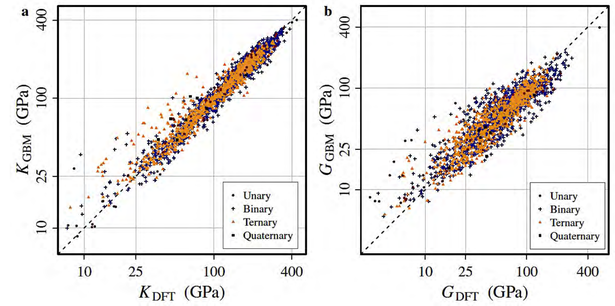
We are building a materials data mining platform called MIDAS (Materials Informatics & Data Analysis Software). MIDAS can retrieve data from several materials databases that have REST APIs or from the JSON format, format that data into a Pandas DataFrame object, automatically generate possible descriptors for the data, and run machine learning algorithms through scikit-learn. Visualization tools are provided through adapters to the plot.ly library. This system will be scaled for "big data analysis" in the future through Apache Spark and similar toolkits such as the FireWorks workflow software developed at LBNL.



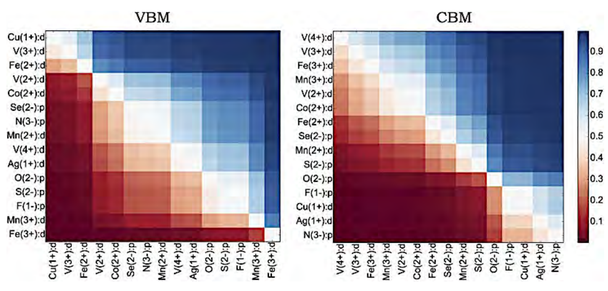
Left: data pipeline for MIDAS
Right: example of an interactive plot that can be generated through MIDAS via the plot.ly toolkit.

Materials data mining

Using data sets from the Materials Project, we have derived structure-property relationships that relate fundamental descriptors such as composition, density, and coordination to output properties such as bulk modulus, shear modulus, and the electronic character of the valence and conduction bands. Crucial to this effort has been the development of relevant descriptor combinations as well as new machine learning approaches.



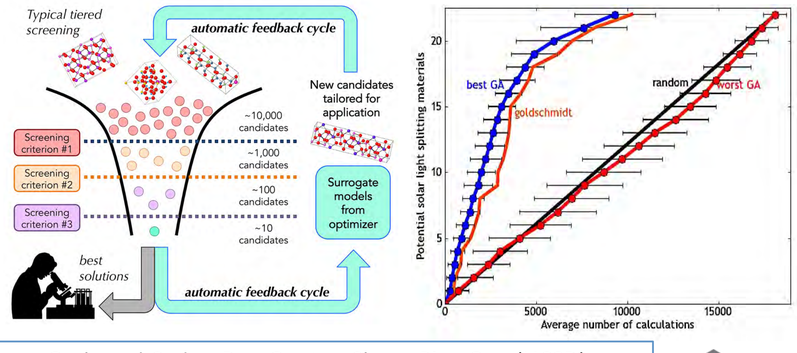
In this machine learning predictor for the bulk and shear moduli of a material, we use Hölder means to improve the predictive power of descriptors and develop a local linear regression method that fits the tails of distributions (i.e., extreme values) accurately without overfitting.



This diagram shows the pairwise likelihood for the electronic state on the y-axis to form a greater contribution to the valence or conduction band character versus the state on the x-axis. For example, *d* orbitals in Cu^{1+} are the most likely to form the VB. For this study, we repurposed algorithms used in ranking sports teams to rank electronic orbitals.

Automated materials design

Materials discovery today is largely performed through targeted experiments based on researcher intuition. However, such intuition is difficult to obtain within the context of high-throughput and combinatorial studies in which tens of thousands of data points may be collected. We are developing automated optimization routines that couple forward models (e.g. density functional theory calculations) with inverse optimizers (e.g., genetic algorithms or Gaussian processes) to build a fully automated materials discovery system.



Left: schematic diagram showing the integration of typical high-throughput screening coupled with an automatic optimization routine.

Right: performance of genetic algorithms (GAs) in uncovering promising perovskite water splitting materials versus random search (black) and chemical rules (orange).

Download this poster: http://www.slideshare.net/anubhavster

Collaborating PIs: Kristin Persson, Gerbrand Ceder, Dan Gunter, Shyue Ping Ong (UCSD), Geoffroy Hautier (U. Catholique de Louvain), Anthony Gamst (UCSD), Karsten Jacobsen (DTU)

Energy Materials Network
U.S. Department of Energy
DuraMAT
Durable Module Materials Consortium