

Project: Materials Genomics

The methods available to scientists and engineers seeking out new materials that meet emerging functional and cost requirements for technological innovation are expanding. Systematic computational surveys across spaces of possible materials are building large scale databases that allow deep search and discovery tools that shorten the time between recognized need and deployed solution. The Materials Project (MP), www.materialsproject.org, has developed an informatics approach to materials discovery and shown successful application to areas such as energy storage, functional electronic, and other emerging materials science challenges. It has done so in a data-driven and highly-collaborative way. The HPC horsepower behind MP is controlled by a database that is feed by community needs and input. The data products are systematically organized in ways that allow science-specific web applications to be built on top of this resource. Knowledge discovery is significantly advanced through this highly discoverable, sharable, and scalable science resource. MP has attracted 3500 science users in the first two years of its operation. Other projects have applied similar approaches toward materials discovery. Many new areas of materials and new applications build on “mashing up” materials science data remain open to exploration. Here are some of the research components of materials genomics:

Flexible and scalable data stores

Materials genomic data is both the scientific results and the provenance of how those results were derived. As knowledge, methodology, and scientific goals evolve, both of these can change. The data store should "get out of the way" of this evolution, changing as easily and quickly as possible to accommodate new information. The main advantage of no-SQL key-value-pair data stores is that it is able to accommodate changes in the data and data structure quickly and relatively simply, which allows a small team of developers to maintain and update the database without a dedicated database administrator

Documented data formats

The process of collaborating with data must be architected so that individual explanations are not necessary for every new data producer and consumer. There are many methods for doing this, but they all start and end with documentation. The meaning, in scientific terms, of every element and relationship in the data must be described so that other people, who are writing the tools and code that consume this data, can understand it. There are several low-hanging fruits that can form a base for understanding: self-documenting, semi-structured data formats like the JavaScript Object Notation (JSON) and standard unit notation such as UCAR's UDunits. Formal schemata and ontologies can also be applied.

Usable and powerful user interfaces

UIs are where scientists meet their instruments and data. Their success in doing so and thus ultimately the value of these resources is bracketed on their human-computer interaction. Good UIs are crucial to success. Of course, fundamentally good science data must lie behind the interface. But confusing and poorly designed interfaces can silently turn away many potential collaborators. Scientists cannot appreciate the power of data that they cannot access, and a good UI allows the "first contact" between the scientist and the data to enable a deep and broad exploration.

Usable and powerful web APIs

Scientists need to ask deep questions involving large and/or complex portions of data sets. They often need to feed these results to analysis programs or visualizations. To enable this type of interaction, the MGI stack needs a web API that can provide programmatic access and transformations of the remote data, which is then integrated into the community analysis tools. For example, in the Materials Project the web API can search, download, or upload new data. This web API, which is integrated into the open-

source *pymatgen* analysis toolkit, has enabled collaborators to download and analyze much larger data sets with their own tools, thus growing the usability of the data beyond the vision of its producers. The Internet revolution has taught us that if you put data in the hands of people, they do unexpected things with it. In the same sense, the Materials Project disseminates its data both through 1) its user-friendly (but ultimately ‘pre-packaged’) web interface and 2) its raw data form to cater to the different needs of the scientific community.

Community analysis tools

One of the goals of materials genomics is to enable users to build a rich set of tools to analyze materials data. The calculations occurring in these tools are as important as the calculations on supercomputers or experiments at light sources to the end analysis. Open-source toolkits and custom web apps form an ecosystem around the data that allow collaborators to add their own analysis capabilities. An evolving set of community efforts based on a common foundation allows data add value over time, become more interconnected with other data resources, and expand the leverage of materials genomics.

Integrated data provenance and versioning

Materials genomics data, like all scientific data, is bound up with the idea that results relating to the same phenomenon should be comparable. The use of computing resources as a principal medium for simulating and analyzing materials genomics data provides a challenge, in the complexity and mutability of the conditions and parameters for data derivation, but also an opportunity to automate and integrate the provenance and versioning of the data in a fundamental way -- to answer crucial questions about data validity and change. Addressing this challenge in a useful way depends on domain knowledge of which conditions and parameters are significant, what granularity would make sense for versioning of the data, etc. The interconnectedness of materials genomics data almost requires that this issue be addressed early, and consistently across projects.

Community Requirements & Needs	CI Components
High throughput workflows	Software that enables small teams to engage in simulation surveys that demands $O(10^6)$ tasks to be marshaled through an HPC batch queue. Nimble HTC as opposed to HEP long term large-team infrastructures.
An application ecosystem that is community accessible.	Data standards and data analysis infrastructure that allows widespread development of interoperable mat sci applications based on the same data. An “app store” for Mat Sci.
Strategy, plans and resources to foster and coordinate the intercomparison of simulation and experimental data	Business process expertise. CI enterprise research. Outreach to experimental science.
Flexible high-performance data store	XL databases, graph and no-sql data bases that are analytics friendly (bring analysis into the DB)

Extreme scale computing	Exascale computing resources to drive simulation surveys. Reliable workflow management, checkpointing, preemption, etc.
Web APIs for science	R&D that make science data more web accessible through standards, formats, and new APIs
Anomaly detection, data QA, collaborative data enhancement	Algorithms and collaborative methods that improve data, finding gaps and inconsistencies in large complex data sets.
Federated identity management, VO access models to simulation and data.	Community-wide authentication and access
Supported, networked community of professionals	Sustainable ASCR/Gvmt. Support

We propose the following research opportunity as an initial research project on which to construct a foundation of cyberinfrastructure for the materials genomics research community.

PI: Kristin Persson

Sponsor/Champion: DOE Office of Science BES, LBNL

Program: Basic Energy Science

Challenge: Collaboration in engineering co-design, integrating hardware, software and applications for new computing platforms. Need a collaborative environment that would enable modelers worldwide to use and compare both simulation data from exascale machines and from experiments.

Need: Cyberinfrastructure that balances large scale computing with data and knowledge systems. Need to effectively coordinate community driven simulation surveys in an HPC context, refine the resulting data into actionable knowledge models, make these models widely accessible.

Partners: JCESR Energy Innovation Hub. LBL, ANL.

Effort: Re-engineering engineering codes to run on new systems, both in terms of new OS and applications operating in-situ to minimize data movement.

Collaboration Goal: Reproduce the success of Materials Project in new science areas, at larger scale, and

Project Requirements and Goals:

1. Identify opportunities to extend infromatic approaches to materials discovery;
2. Identify the necessary CI elements (hardware, data management, algorithms, software, experimental facilities, people, support, etc.) through retrospective and future architecture perspectives.
3. Identify CI that is needed for leveraging MG data at increasing scales and increasingly levels of interconnectedness and depth.

4. Identify human, cultural, institutional, and policy challenges in the socio-technical space of data sharing between simulation scientists, experimentalists, user communities, and the private sector.
5. Estimate the resources (funding, manpower, facilities) needed to provide stable, long-term MG capabilities;
6. Recommend a plan for enhanced exploitation of CI for MG research, leveraging emerging CI and DOE HPC facilities plans.