“High-throughput computing for materials databases and materials design”.

Open Science Grid User School

July 29, 2016

Univ. of Wisconsin – Madison, WI

High-throughput computing for materials databases and materials design

Dane Morgan

Tam Mayeshiba, Henry Wu, and many others …

Open Science Grid User School

July 29, 2016

Univ. of Wisconsin – Madison, WI

THE UNIVERSITY of WISCONSIN MADISON

Computational Materials Group

Funding Acknowledgements

Software Infrastructure
for Sustained Innovation
(SI^2) award No. 1148011

CENTER FOR
HIGH THROUGHPUT
COMPUTING

National Energy Research
Scientific Computing Center

Outline

High-throughput Molecular Simulation

Alloy Diffusion Database

Outline

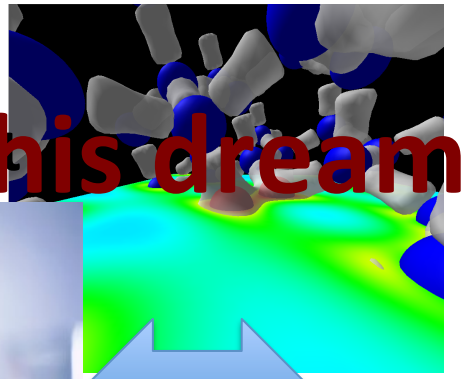
High-throughput Molecular Simulation

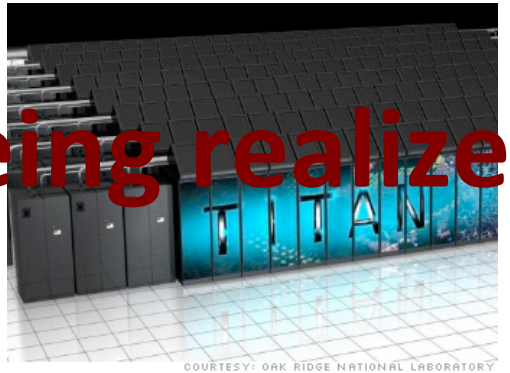
Alloy Diffusion Database

The Dream of Molecular Computational Materials Science

Atomic Understanding

Computation





Experiment

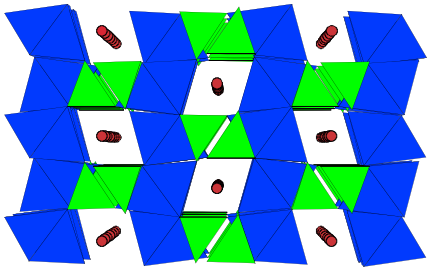
This dream is being realized

This is a major transformation

Understand, Optimize, Discover Materials

Ab Initio Methods

Composition
Structure



$$H\psi = \varepsilon\psi$$

$$H = \left[\sum_{i,I} \frac{Z_I}{|r_i - R_I|} - \sum_i \frac{1}{2}\nabla_i^2 + \sum_{i,i'} \frac{1}{|r_i - r_{i'}|} \right]$$

Materials Properties

Electronic Structure
Band structure, magnetism, …

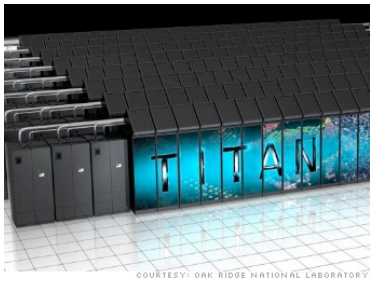
Energies, Forces
Atomic positions, phase stability, …

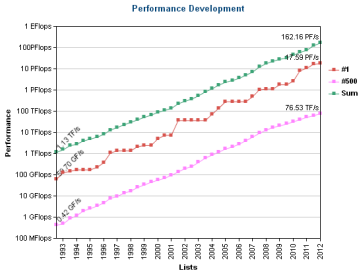
Drivers for Transformation

Fundamental theory

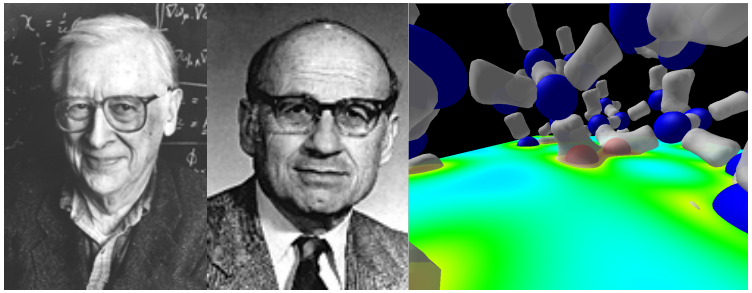
$$\hat{H} = \left[\begin{array}{c} -\sum_I \frac{1}{2}\nabla_I^2 + \sum_{j,j'} \frac{Z_j Z_{j'}}{|R_j - R_{j'}|} + \sum_{i,j} \frac{Z_j}{|r_i - R_j|} \\ -\sum_i \frac{1}{2}\nabla_i^2 + \sum_{i,i'} \frac{1}{|r_i - r_{i'}|} \end{array} \right]$$

Computational power

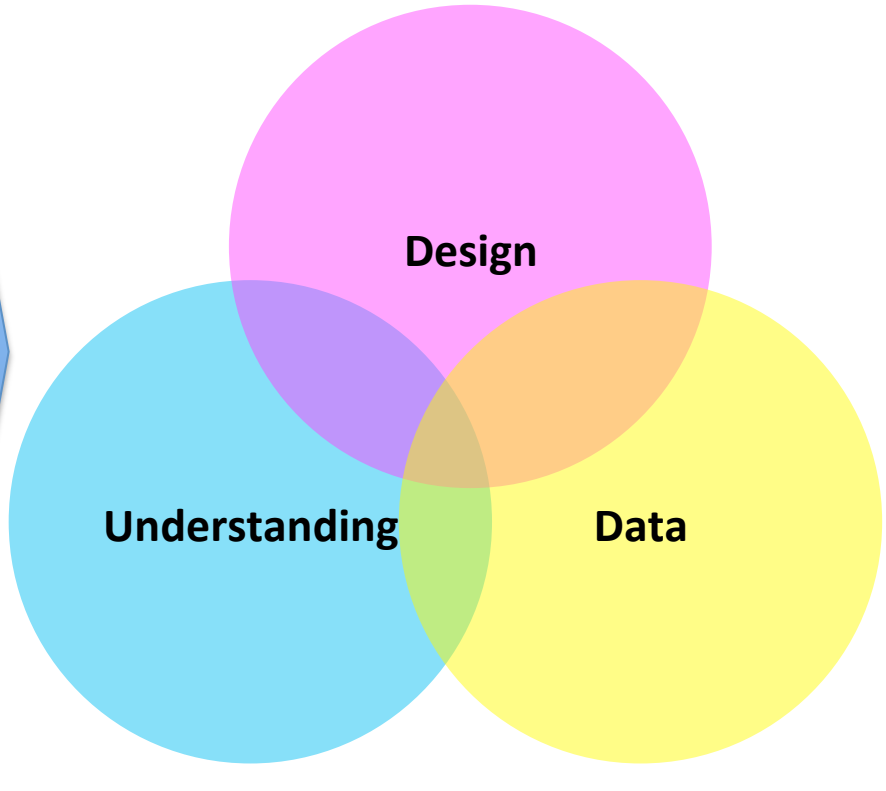




Modeling methods



Unprecedented transformation in



A Simple but Powerful Message

Computation Is Scalable

A Simple but Powerful Message

If you can compute it once

Then with some automation

You can compute it a lot

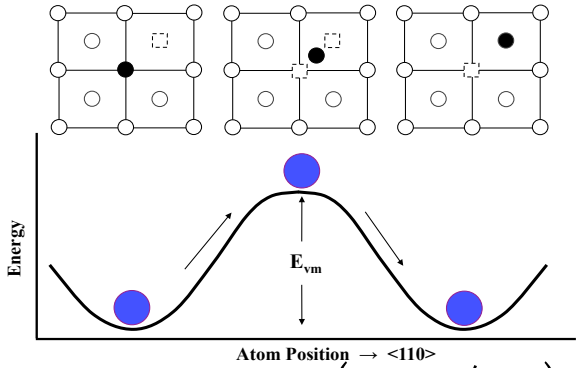
Outline

High-throughput Molecular Simulation

Alloy Diffusion Database

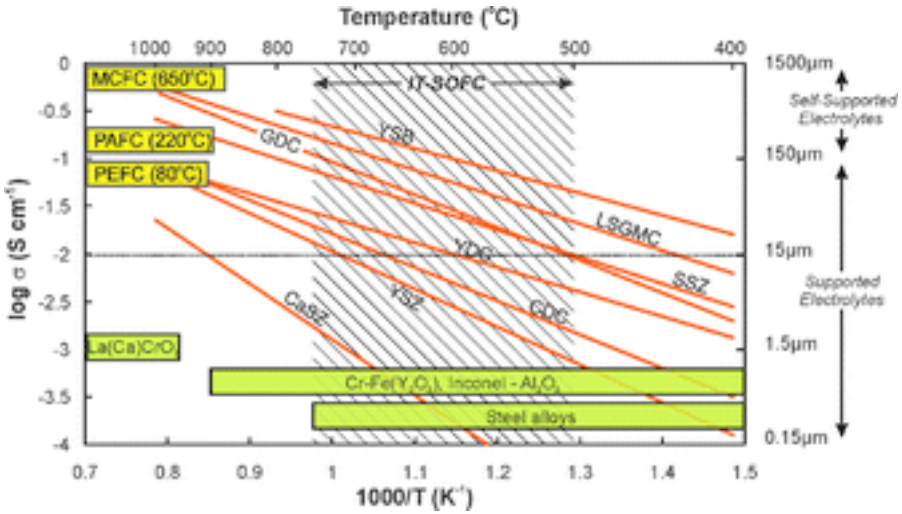
Ab Initio Methods and Diffusion in Solids

- Diffusion typically occurs by jumps between stable sites
- Jump rates depends on attempt rates and migration barriers, which can be calculated ab initio



$$Rate = \nu \exp\left(-E_m / kT\right)$$

- Diffusion coefficients (*D*) can be calculated from jump rates analytically
- *D*'s are critical for design of Li ion batteries, solid oxide fuel cells, semiconductor devices, steels, …



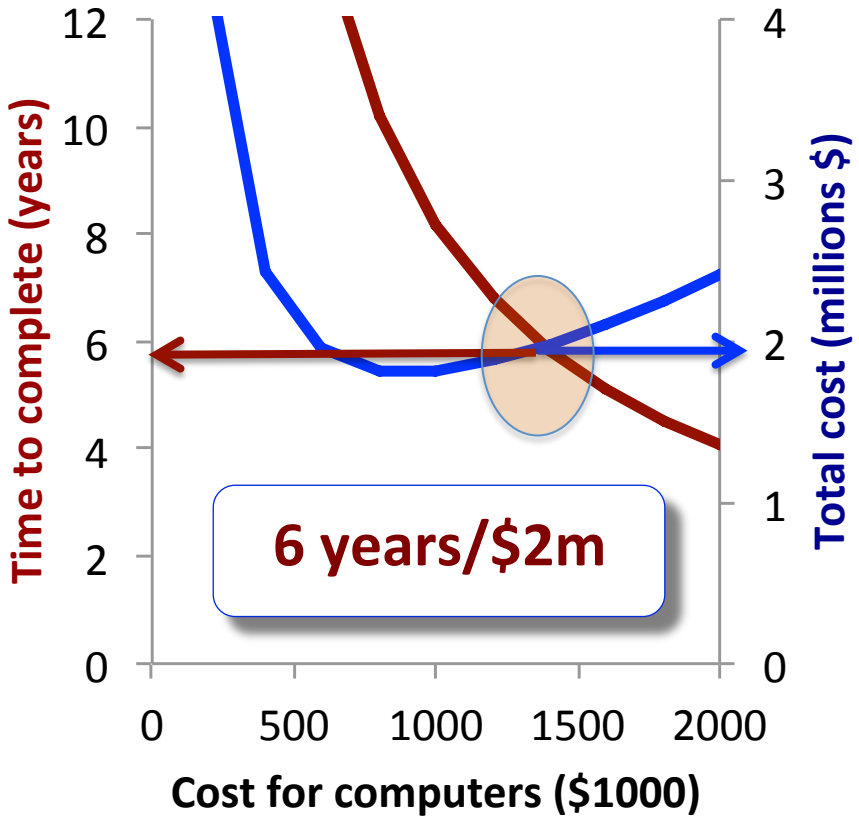
Brett, et al, Chem. Soc. Rev. '08

High-Throughput for Dilute Alloys

Determine D_v* for $A_{1-x}B_x$ (x<<1) for all elements A,B in the common (FCC, BCC, HCP, Diamond) crystal structures

Resource needs

- ~50 viable pure elemental systems in each structure ➔ ~10,000 dilute B in A alloy-structure systems (maybe ~5% known)
- 1 system takes ~20k/core-hours (~9 days on 100 cores (= $20k))
- So need 2×10^8 core-hours or ~23k core-years, 1 postdoc.

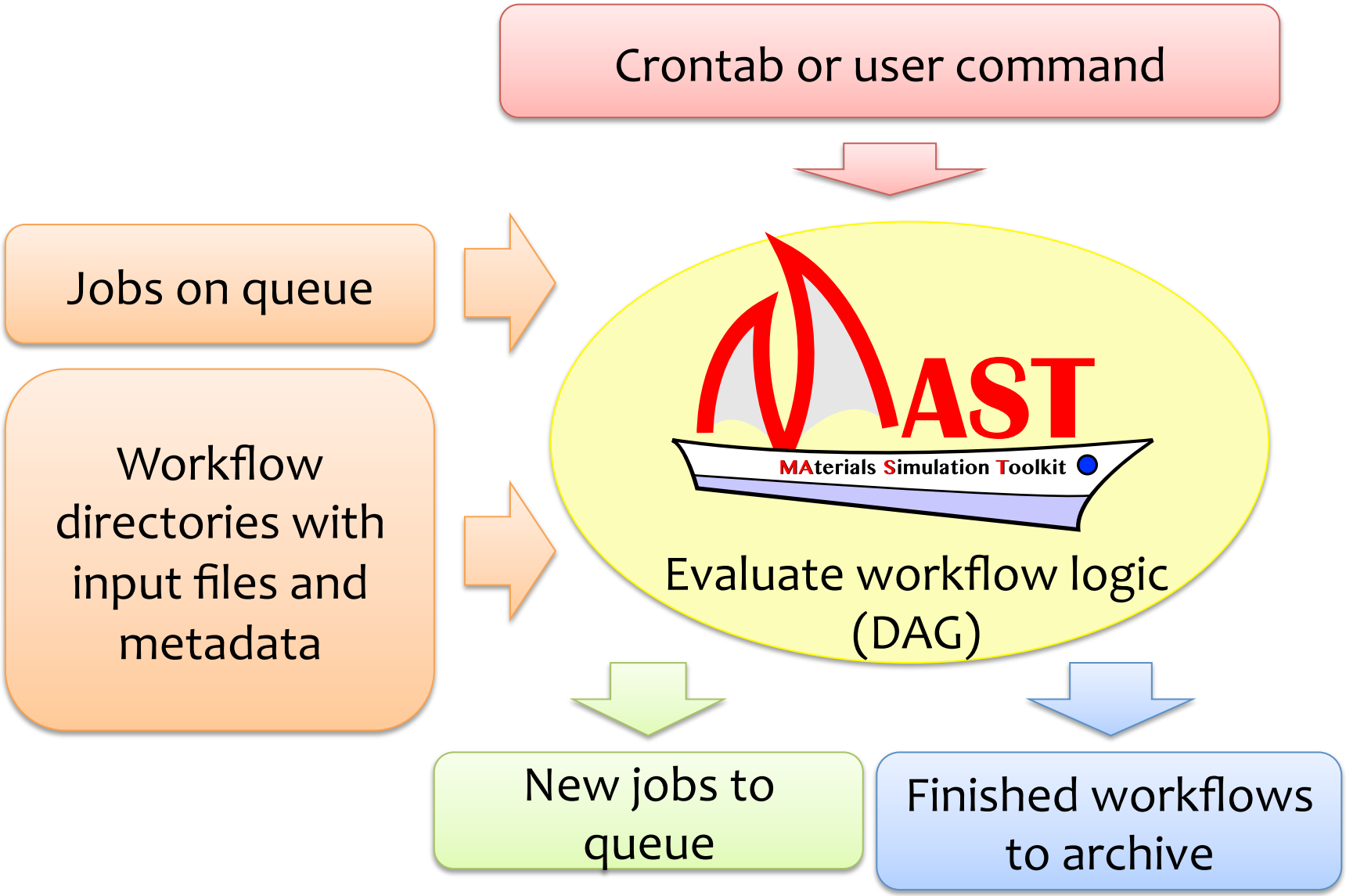


MAterials Simulation Toolkit (MAST)

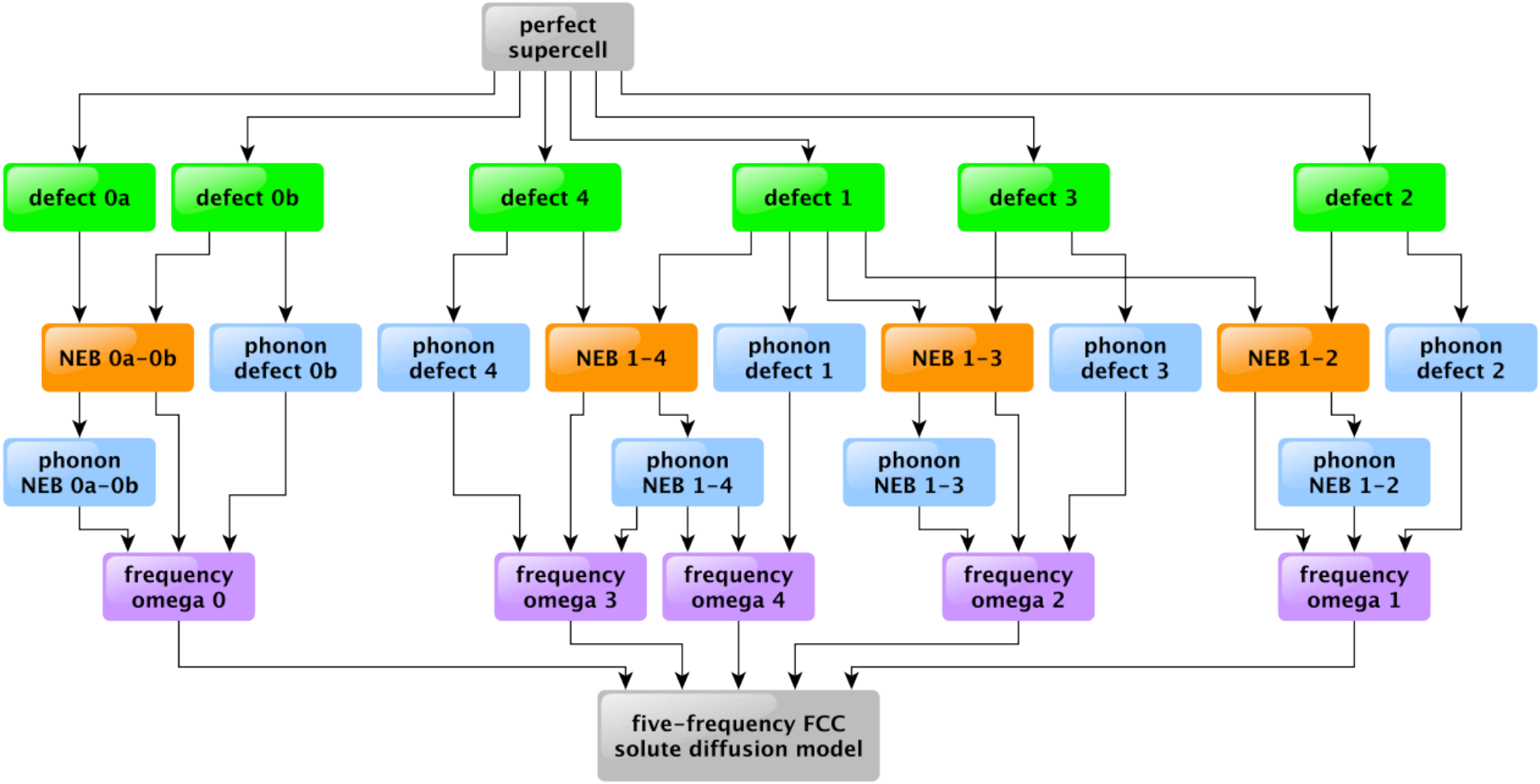
https://pypi.python.org/pypi/MAST

The MAterials Simulation Toolkit (MAST) is an automated workflow manager and post-processing tool primarily designed to perform atomic simulation calculations for diffusion and defect workflows, especially using density functional theory as implemented by the Vienna Ab-initio Simulation Package (VASP).

MAST Workflow Management



Actual diffusion workflow schematic



32 steps (not all steps are shown)

Using Open Science Grid - Problems

- Our typical unit of one diffusion coefficient is ~20k CPU hours – clearly needs to be broken up for OSG
- Single ab initio calculations tend to be significantly parallel (~16-128 cores) and long (10-100h) – poor match for OSG
- MAST workflow manager not initially compatible with OSG (MAST runs from a managing shared home directory)

Using Open Science Grid - Solutions

- Consider the smallest steps in our ~20k CPU hour workflow and build on those (single step calculations).
- Restrict to specific types of nodes with 16-20 parallel cores available on one node.
- Chose materials carefully to be fast (few electrons) so jobs can usually finish within 24h soft limit on OSG machines.
- Manage workflow differently
  - Idea 1: Adapted MAST to CHTC by sending all tools needed on the home directory (MAST, related directories, python language, etc.) to compute node with job. Worked, but had to send a lot back and forth and managing the directories to avoid workflow errors (e.g., overwriting) was very hard.
  - Idea 2: Used MAST to set up workflow DAG and then transcribed to use DAGMAN workflow manager in CONDOR on OSG. Better! But reduces error checking ability.

Open Science Grid Usage

- Used about 2.6m CPU hours over 15m
  - About 1.5m CPU hours dedicated to production runs for diffusion project.
- Ran about 80 diffusion coefficients.
- Integrated with traditional HPC (XSEDE, NERSC) for larger runs.

Diffusion Database

http://diffusiondata.materialshub.org/

https://www.engr.wisc.edu/making-massive-materials-data-sets-tools-accessible/

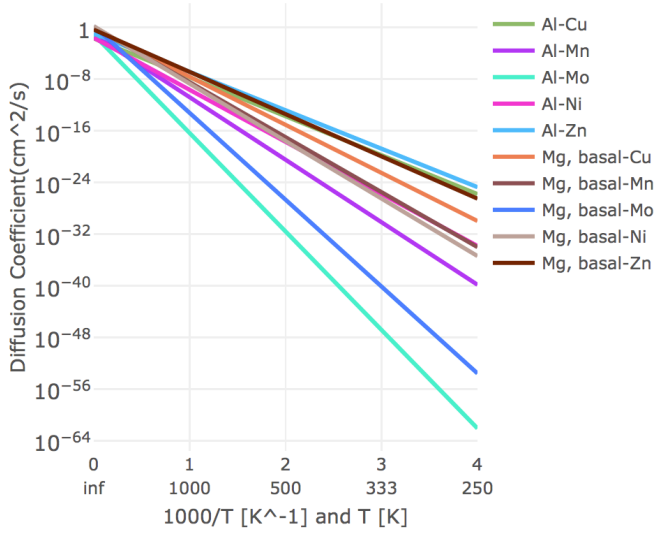
- Impurity diffusion of X in host H for over 350 systems.
- Largest diffusion database from a single group in the world. New science and critical design data.
- Data disseminated through web
  - Web application for plotting and exploring data
  - All data available from figshare with permanent DOI.

Mg, basal × Al ×

Zn × Cu × Mo × Ni × Mn ×

Check-All

Host-Solute	A($cm^2/_{sec}$)	Q(eV)
Al - Cu	0.0207028	1.19609
Al - Mn	0.0771412	1.9213
Al - Mo	0.0802623	3.0262
Al - Ni	0.0258404	1.59959
Al - Zn	0.123356	1.18137
Mg, basal - Cu	0.655208	1.47741
Mg, basal - Mn	1.12569	1.6898
Mg, basal - Mo	1.6(	
Mg, basal - Ni	1.6⁴	
Mg, basal - Zn	0.4⁷	



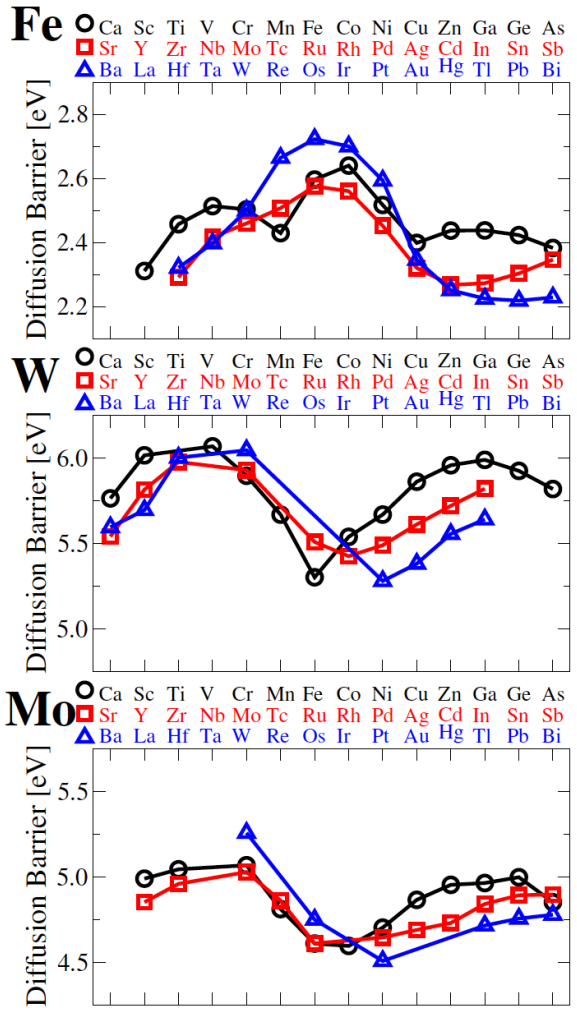
Wu, et al., Scientific Data, '16

Diffusion Database

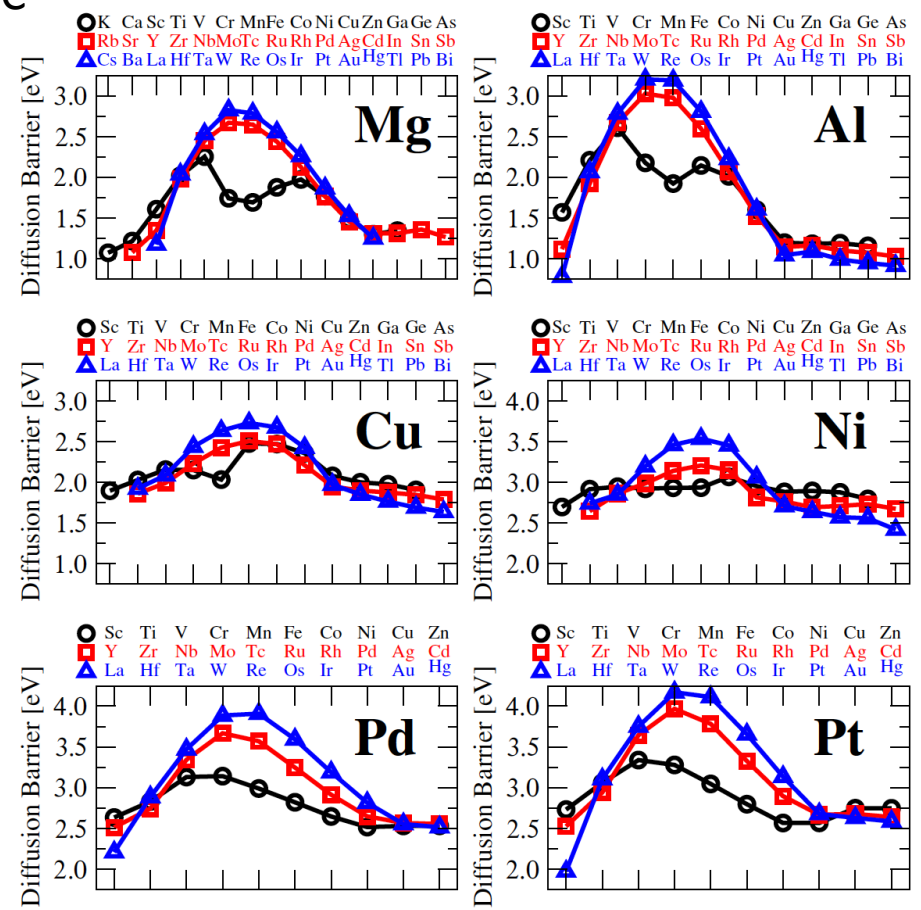
http://diffusiondata.materialshub.org/

https://www.engr.wisc.edu/making-massive-materials-data-sets-tools-accessible/

BCC



FCC



Very different trends between FCC and BCC – need large database to discover this.

Wu, et al., Scientific Data, ‘16

Excess Formation Volume Computation

- Disordered binary mixtures of elements A and B for **BCC** and **FCC** crystal structures at various compositions (A, $A_{0.75}B_{0.25}$, $A_{0.5}B_{0.5}$, $A_{0.25}B_{0.75}$, and B)

- A, B = Al, Co, Cu, Fe, Mg, Mo, Nb, Ni, and Ti
  9 pure elements, and a combined **36 unique elemental pairs**.

- Use **3 different special quasi-random structures (SQS)** for each crystal structure.
  - Each SQS is optimized for all **three mixtures** (25%, 50%, and 75%).

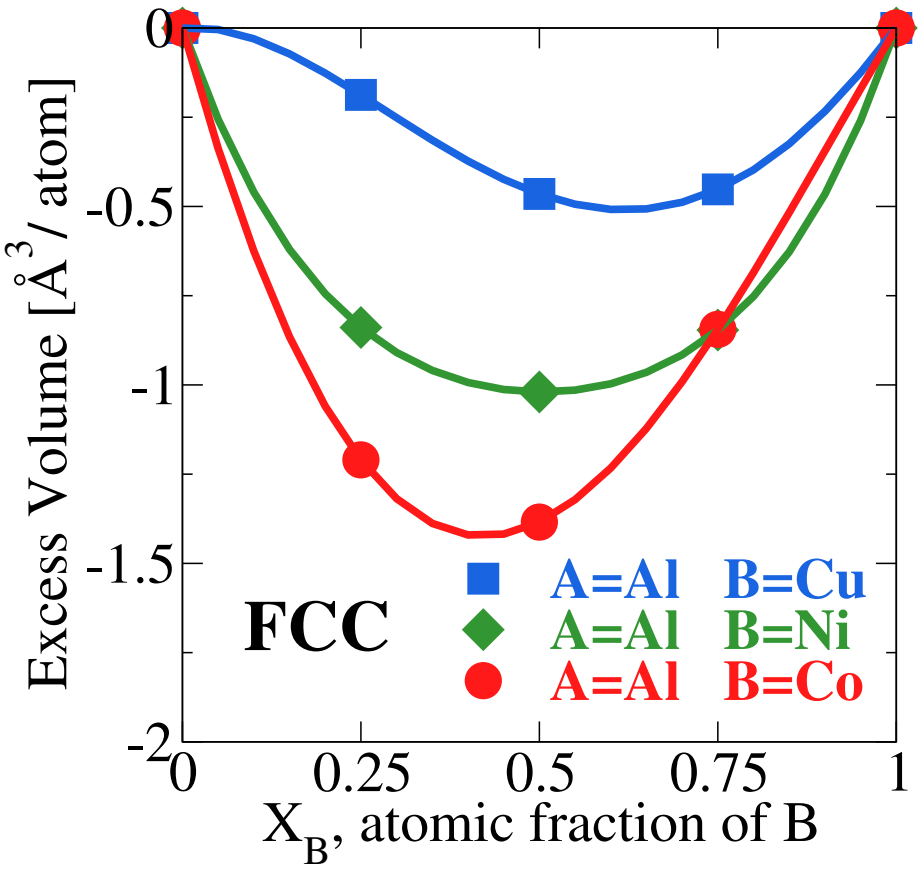
- Calculate formation volume with DFT, spin-polarized:
  - Iterative relaxation between ionic relaxation and volume relaxation.
  - **At least 3 repeats of the above iteration**.

- Total number of calculations:
  - (2 structures)×(36 pairs)×(3 SQS)×(3 compositions)×(6 DFT) =
  - = **3888 DFT calculations**.

Excess Formation Volume Results

Excess volume - fit to 2nd order Redlich-Kister Polynomial

$$V_{excess} = A_0 X_A X_B + A_1 X_A X_B (X_A - X_B) + A_2 X_A X_B (X_A - X_B)^2$$



	Al-Cu	Al-Ni	Al-Co
A_0	-1.8600	-4.0792	-5.5367
A_1	1.4111	0.0389	-1.9511
A_2	0.6178	-1.6611	0.2356

We are able to generate a large amount of accurate data and can extract valuable thermodynamic parameters.

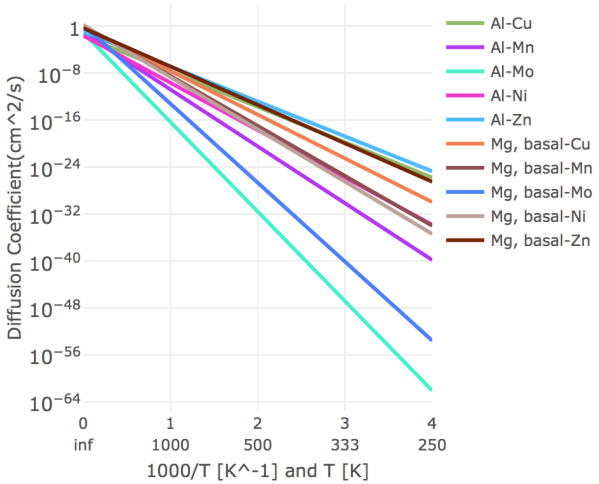
Summary

- We have developed an approach to successfully run large sets of high-throughput ab initio calculations for materials design using OSG.
- We have used over 2.6m CPU hours over the last ~2years to develop the world's largest diffusion database from a single research group.
- Enables many other materials properties calculations which we are exploring, e.g., alloy volumes, oxide defects, etc. …

Mg, basal × Al ×

Zn × Cu × Mo × Ni × Mn ×

Check-All		
Host-Solute	A($cm^2/_{sec}$)	Q(eV)
Al - Cu	0.0207028	1.19609
Al - Mn	0.0771412	1.9213
Al - Mo	0.0802623	3.0262
Al - Ni	0.0258404	1.59959
Al - Zn	0.123356	1.18137
Mg, basal - Cu	0.655208	1.47741
Mg, basal - Mn	1.12569	1.6898
Mg, basal - Mo	1.66772	2.66878
Mg, basal - Ni	1.64189	1.76667
Mg, basal - Zn	0.479027	1.29902



Thank You

Any Questions?