Crystal Structure Prediction

Fawzi Mohamed, Emre Ahmetcik, Angelo Ziletti, Luca Ghiringhelli and Matthias Scheffler

Fritz Haber Institute Berlin
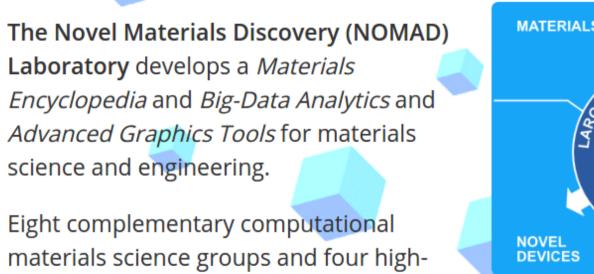
NOMAD: Novel Material Discovery

THE NOMAD LABORATORY — A EUROPEAN CENTRE OF EXCELLENCE

The Novel Materials Discovery (NOMAD) Laboratory develops a *Materials Encyclopedia* and *Big-Data Analytics* and *Advanced Graphics Tools* for materials science and engineering.

Eight complementary computational materials science groups and four high-performance computing centers form the synergetic core of this Centre of Excellence.

http://nomad-coe.eu
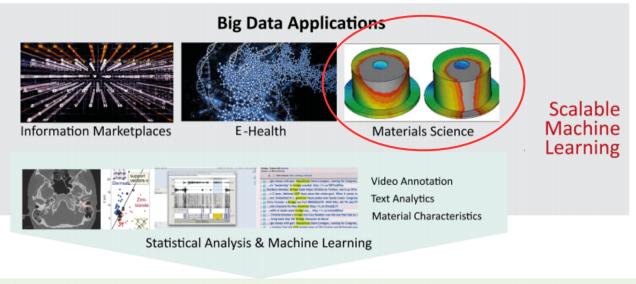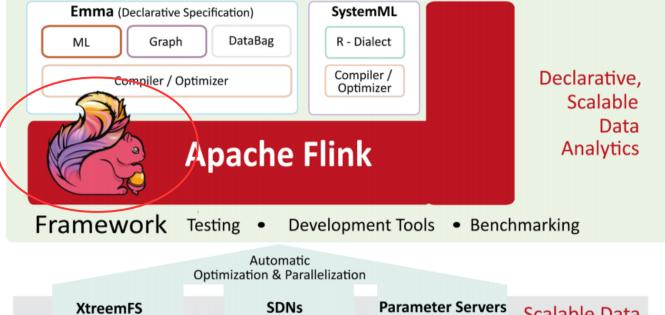
We develop and implement methods that identify correlations and structure in big data of materials. This will enable scientists and engineers to decide which materials are useful for specific applications or which new materials should be the focus of future studies.

The first step to enable this is to **make** the **data available** for analysis.

Then we want to provide interactive tools to find trends and anomalies to discover novel materials.
Here the **BBDC** knowledge in **scaling big data analysis** enriches the expertise of the NOMAD Center of Excellence

The preparation, synthesis, and characterization of new materials is a complex and costly aspect of materials design. About 200,000 materials are "known" to exist, but the basic properties (e.g., optical gap, elasticity constants, plasticity, piezoelectric tensors, conductivity, etc.) have been determined for very few of them. Considering organic and inorganic materials, surfaces, interfaces, and nanostructures, as well as inorganic/organic hybrids, the number of possible materials is practically infinite. It is therefore highly likely that new materials with superior (but currently unknown) properties exist but still have yet to be identified, which could help address fundamental issues in a number of widespread fields such as energy storage and transformation, mobility, safety, information, and health.
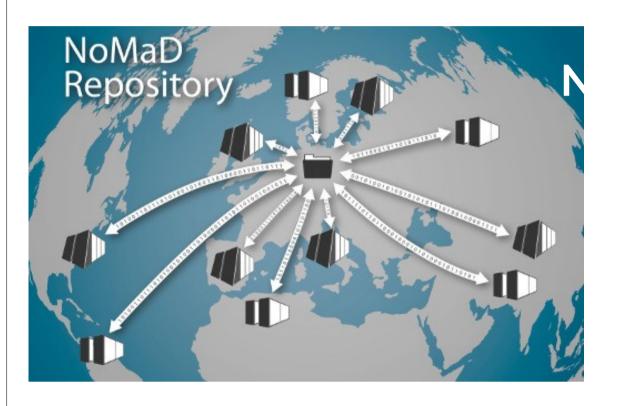
Despite a huge number of possible materials, we note that "the chemical compound space" is sparsely populated when the focus is on selected properties or functions. Our aim is to develop big-data analytics tools that will help to sort all of the available materials data to identify trends and anomalies.
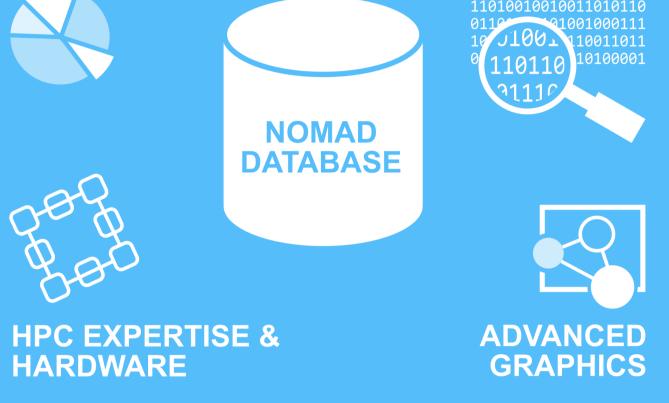
Data: Theoretical Material Science Calculations

http://nomad-repository.eu

- Source of our data
- Established to host organize and share materials data
- Keeps data for at least 10 Years
- Open access and restricted data
  - >3M entries, >2.8M open access

- We use only open access data
- Joint effort by the groups of
  Matthias Scheffler, FHI Berlin and Claudia Draxl, HU Berlin
  Max Planck Computer & Data Facility (MPCDF)

- Parsers **interpret** all calculation data
- **Organize** it according to the metadata structure
- Data not extracted is invisible
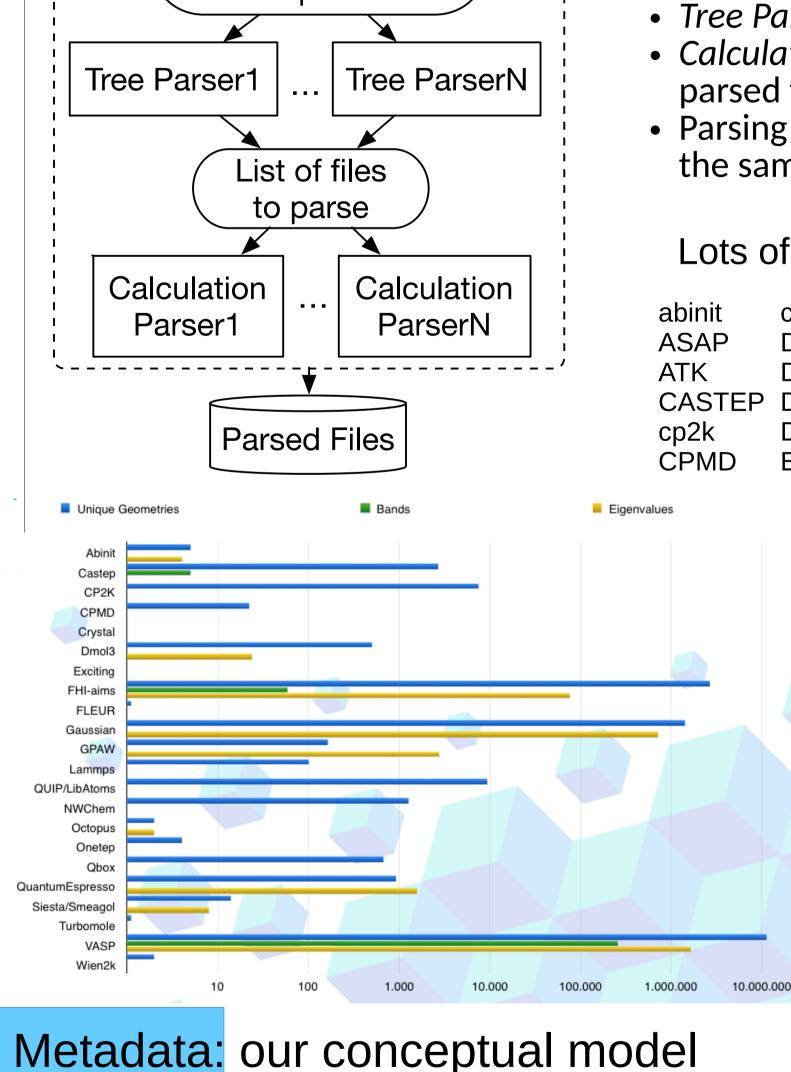- Writing a parser cannot be automatized and requires a person with *scientific knowledge*

- **Parallel** execution
- *Tree Parser* identifies the files
- *Calculation Parser* performs the parsing and generates the parsed files
- Parsing is **pure**: the same version on the same data should give the same result

Lots of codes and formats

abinit	crystal	exciting	GPAW	octopus	SIESTA	VASP
ASAP	DFTB+	FHI-aims	GULP	onetep	Smeagol	WIEN2k
ATK	DFTB+	FLEUR	LAMMPS	ORCA	turbomole	
CASTEP	DL_POLY	FPLO	Molcas	Qbox		
cp2k	Dmol3	GAMESS	MOPAC	Quantum Espresso		
CPMD	ELK	Gaussian	NWChem	QUIP /libatoms/GAP		

Molecule 27 %
Bulk 73 %

Atom	33.663
Molecule	5.864.435
Surface	37.632
Bulk	15.939.034

Semiconductors or Insulators — Non Metals 59% — Metals 41%

Semiconductors or Insulators: Elemental 2%, Quaternary 1%, Binary 35%, Ternary 62%

Metals: Elementary 10%, Binary 7%, Ternary 83%

Metadata: our conceptual model

- section_run
  - program_name
  - program_version
  - [...]
  - section_system
    - simulation_cell
    - atom_positions
    - atom_labels
    - [...]
  - section_method
    - basis_set
    - XC_method
    - [...]
  - section_single_calculation_configuration
    - section_scf_iteration
      - energy_total_scf_iteration
      - [...]
    - section_scf_iteration
      - energy_total_scf_iteration
      - [...]
    - [...]
    - energy_total
    - electronic_kinetic_energy
    - [...]

Legend: Section, Concrete value, Reference

NOMAD Meta Info

section_single_configuration_calculation

Type: Section

Description: Every section_single_configuration_calculation contains the values computed during a single configuration calculation, i.e., a calculation performed on a given configuration of the system (as defined in section_system) and a given computational method (e.g., exchange-correlation method, basis sets, as defined in section_method). The link between the current section_single_configuration_calculation and the related section_system and section_method is established by the values stored in single_configuration_calculation_to_system_ref and single_configuration_calculation_to_method_ref, respectively. The reason why information on the system configuration and computational method is stored separately is that several single configuration calculations can be performed on the same system configuration, viz. several system configurations can be evaluated with the same computational method. This storage strategy avoids redundancies.
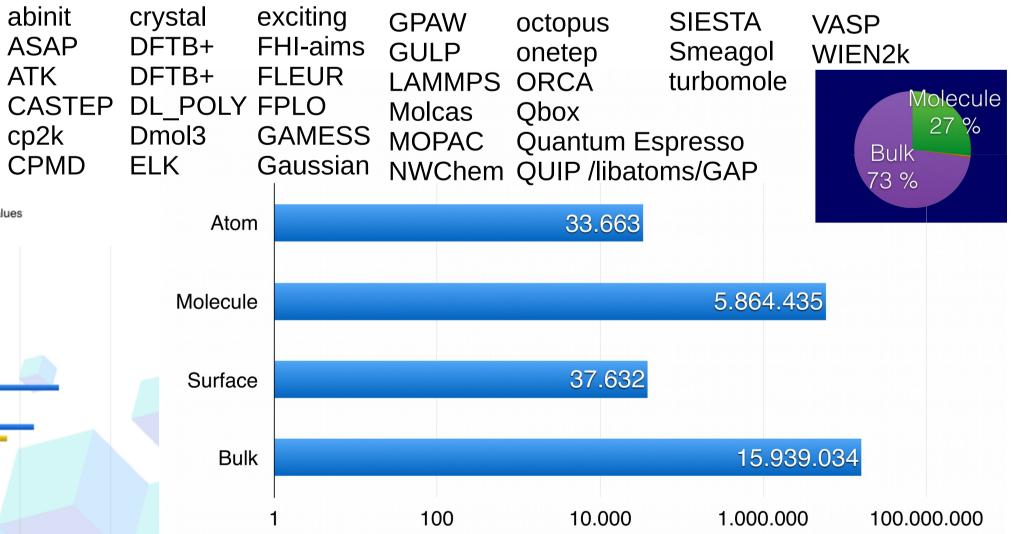
Tool: Crystal Structure Prediction

A web-based implementation (via notebook) of a data-analysis tool for the recognition of the similarity among crystal structures and for the prediction of the difference in formation energy among them. The tool gathers the data for the analysis by a flink-based query to the NOMAD Archive, that contains several millions of crystal configurations.
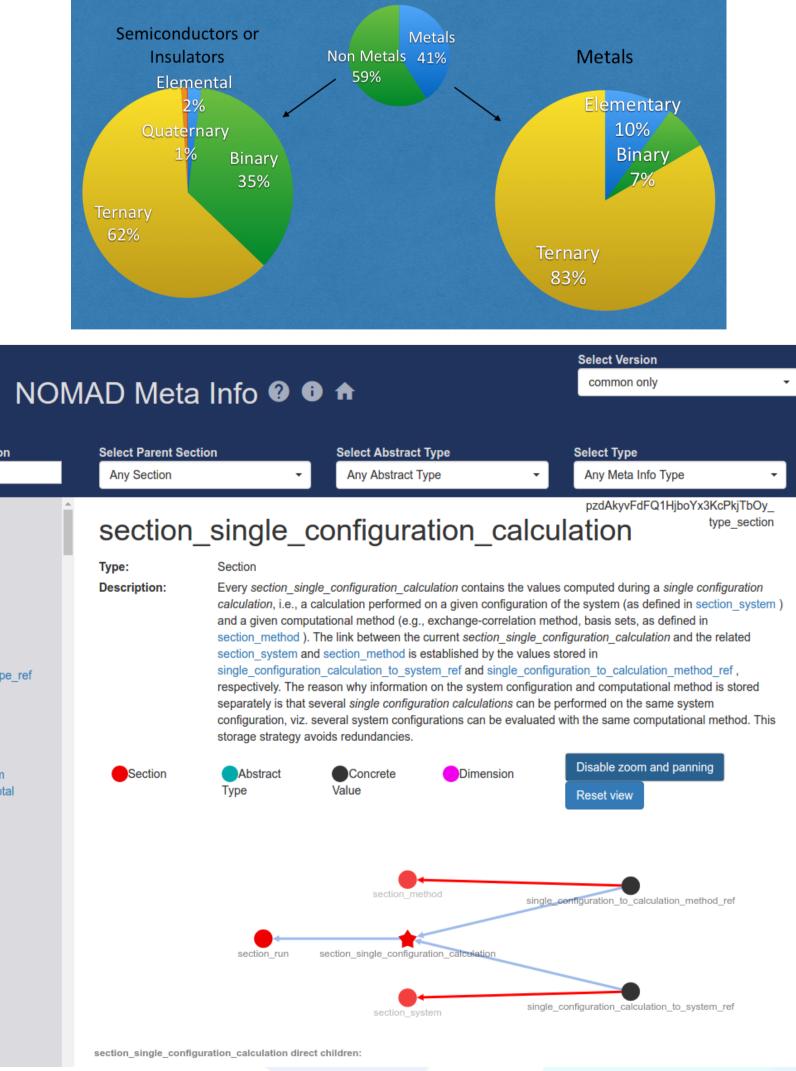The similarity-recognition algorithm, based on descriptors that encode the proper symmetries of a well-behaved physical representation and makes use of linear and non-linear low-dimensional embedding methods, produces a 2-dimensional map that assigns to separate regions perfect and distorted configurations, for given pairs of crystal structures.
The algorithm predicting the difference in formation energies selects the model out of thousands of candidates, by means of a compressed-sensing based method.

NOMAD analytics toolkit

Tutorial example on Crystal prediction I: The case of octet-binary zincblende-vs.-rocksalt semiconductors

developed by Angelo Ziletti, Ankit Kariryaa, Emre Ahmetcik, Fawzi Mohamed, Luca Ghiringhelli, and Matthias Scheffler. [Last update September 13, 2016]

Machine learning method: Compressed sensing (LASSO performed on a tailor-made feature space, followed by L0-regularized minimization).
Click here for more info on the LASSO+L0 method.

Reference: "Big Data of Materials Science: Critical Role of the Descriptor" L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. 114, 105503 (2015) (Click here for the free access pdf)

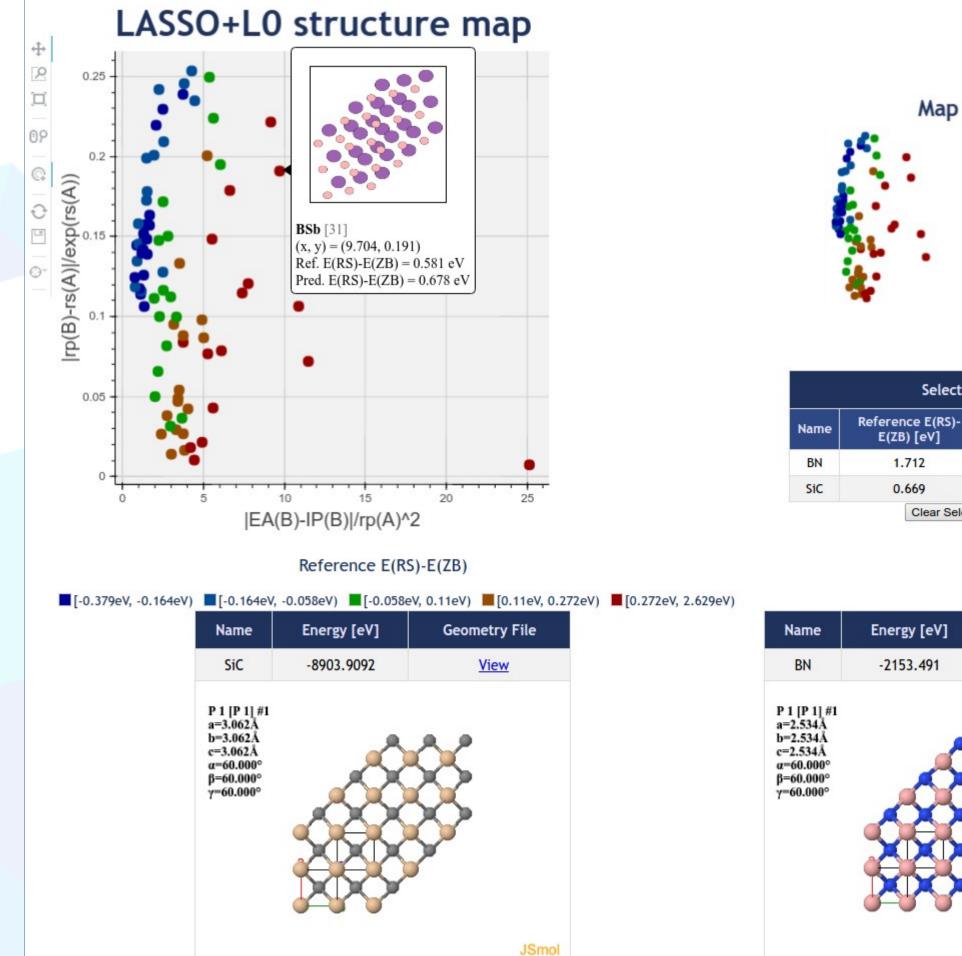**LASSO+L0 structure map**

Map

Bulk [31]
(x, y) = (9.704, 0.191)
Ref. E(RS)-E(ZB) = 0.581 eV
Pred. E(RS)-E(ZB) = 0.678 eV

Selection		
Name	Reference E(RS)-E(ZB) [eV]	Predicted E(RS)-E(ZB) [eV]
BN	1.712	1.618
SiC	0.669	0.562

Reference E(RS)-E(ZB)

Name	Energy [eV]	Geometry File
SiC	-8903.9092	View

Name	Energy [eV]	Geometry File
BN	-2153.491	View