Structural Representations of Materials for Machine Learning using the Novel Materials Discovery Big-Data Analytics Platform

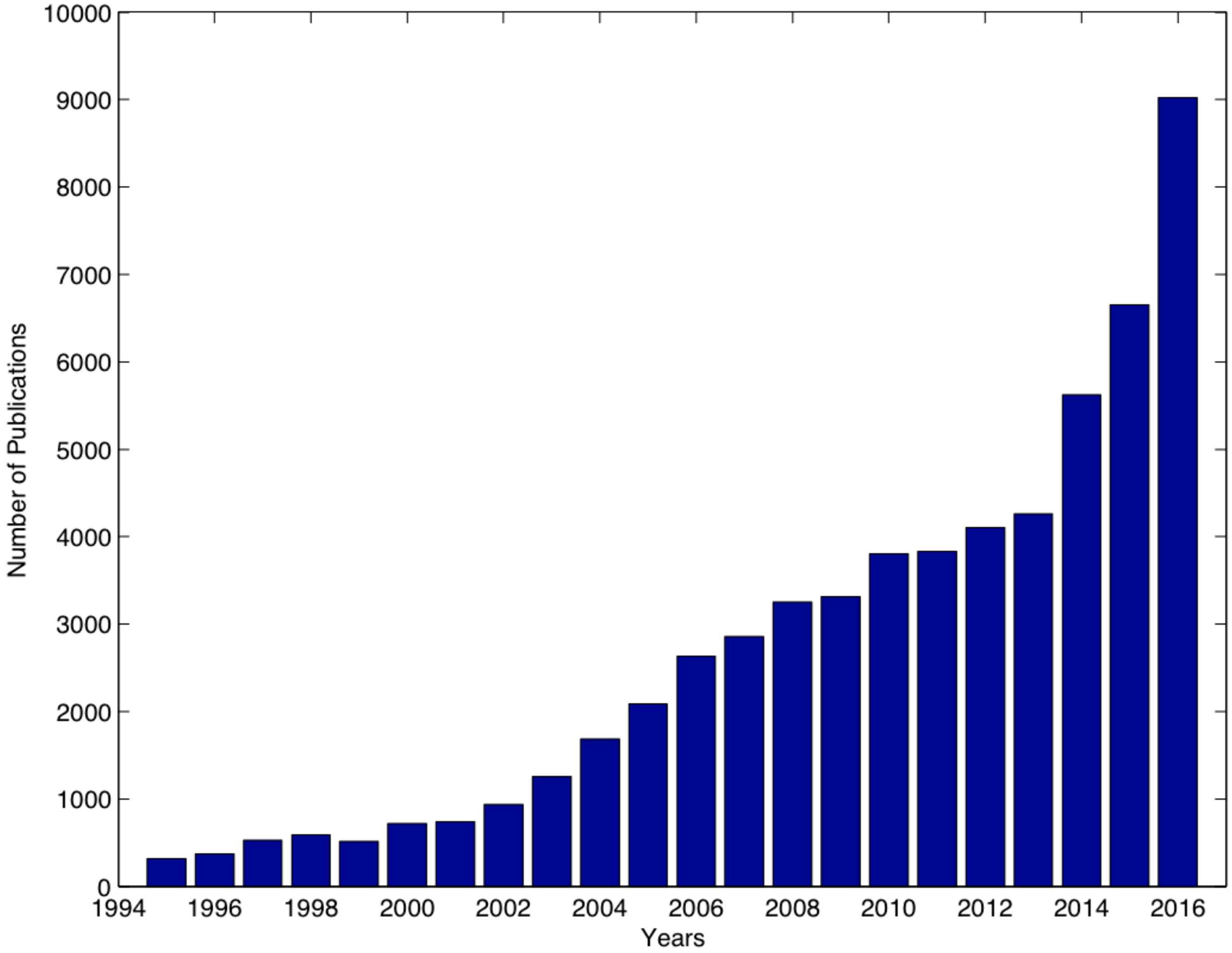
WARWICK

Berk Onat

School of Engineering
University of Warwick

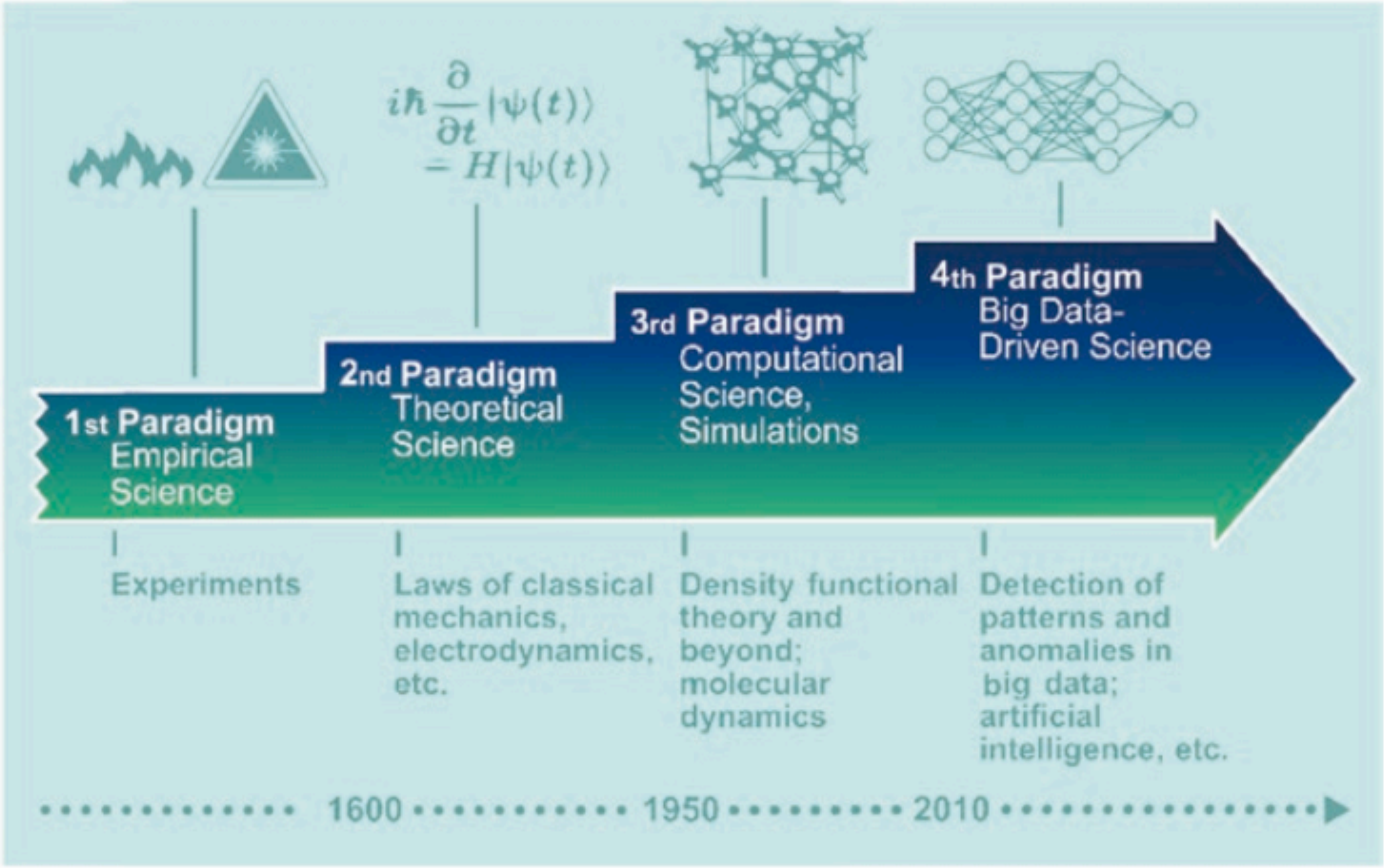
WCPM Warwick Centre for Predictive Modelling

Why Machine Learning is popular now?



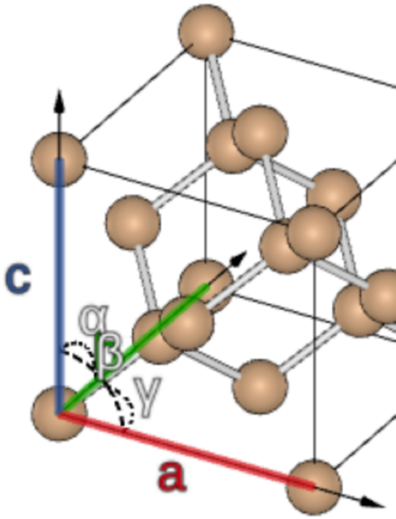
Number of publications per year between 1994 to 2016 (Web of Science)

Paradigms of Material Science and Engineering

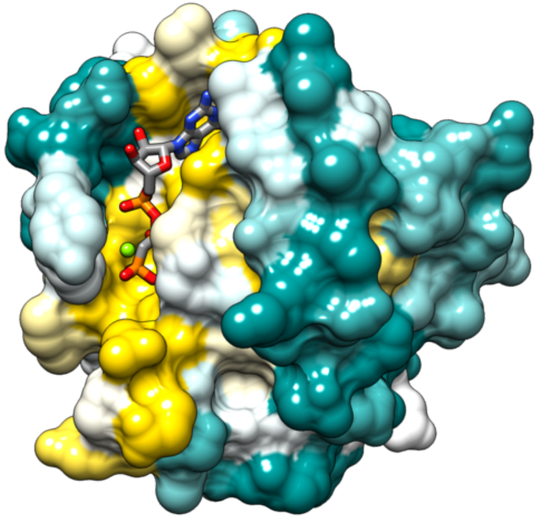

$i\hbar \frac{\partial}{\partial t}|\psi(t)\rangle = H|\psi(t)\rangle$
1st Paradigm
Empirical Science
2nd Paradigm
Theoretical Science
3rd Paradigm
Computational Science, Simulations
4th Paradigm
Big Data-Driven Science
Experiments
Laws of classical mechanics, electrodynamics, etc.
Density functional theory and beyond; molecular dynamics
Detection of patterns and anomalies in big data; artificial intelligence, etc.
1600
1950
2010

Ref: Claudia Draxl and Matthias Scheffler, “NOMAD: The FAIR concept for big data-driven materials science”, MRS Bulletin, Volume 43, (2018) 676-682.

Big-Data Analytics in Material Science and Engineering



Si Diamond Crystal



HRAS Protein (~7000 Atoms) Ref: Wikipedia / PDB

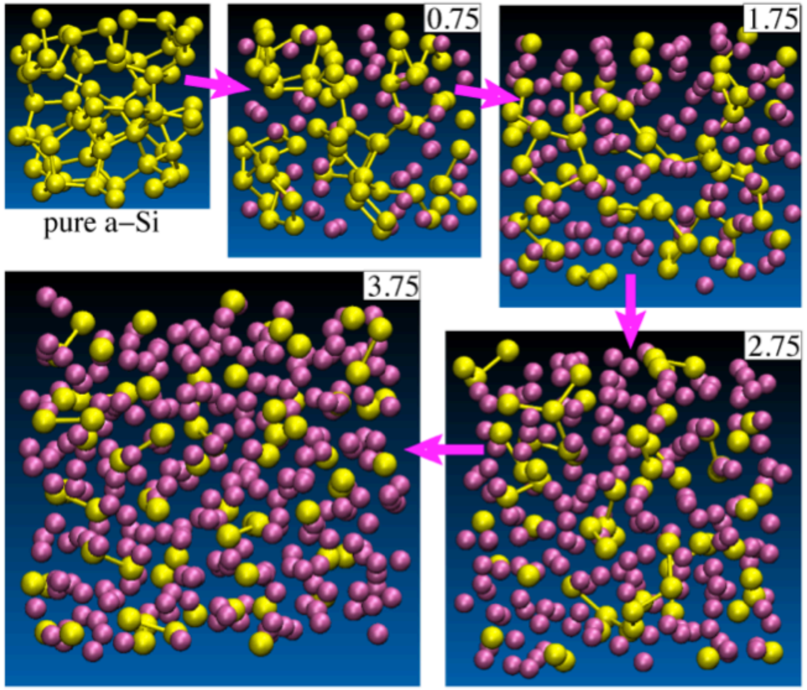
Prediction

Electronic or Structural Properties:

Energies
Forces
Atomic Charges
Magnetization
Band Structure
Band Gap
Density of States
…

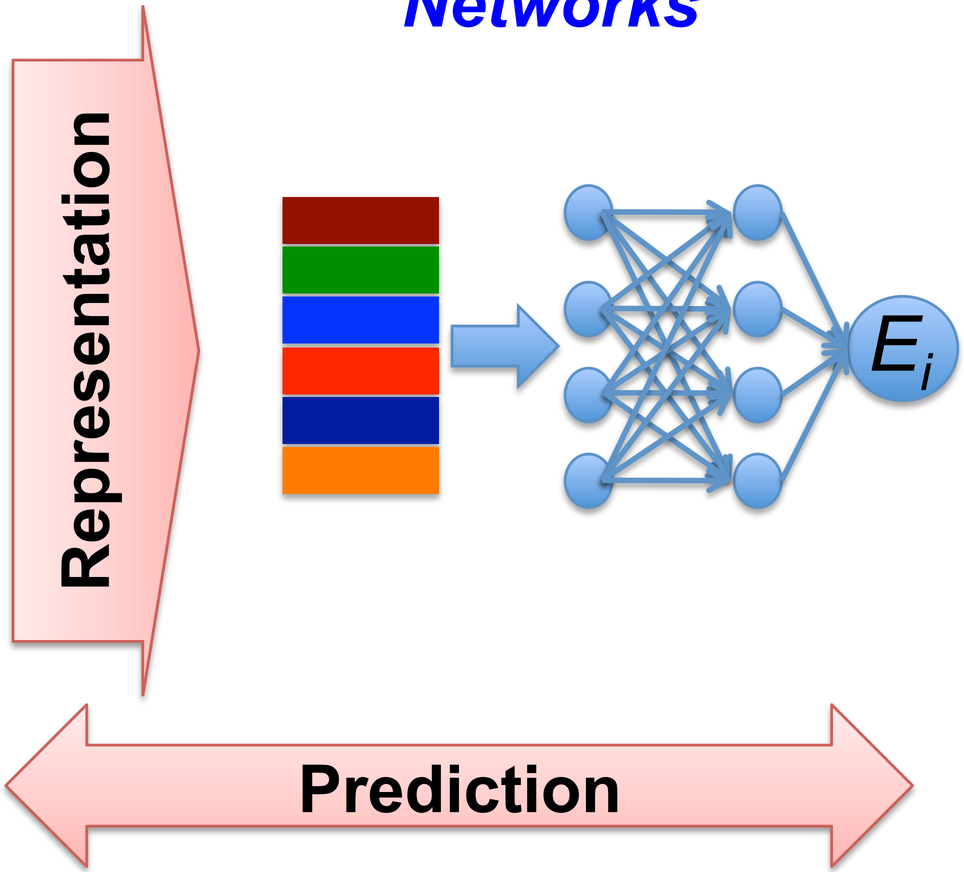
Representation of Potential Energy Surface of Li-Si

Lithiation of Amorphous Si



19000 Structures
SIESTA calculations

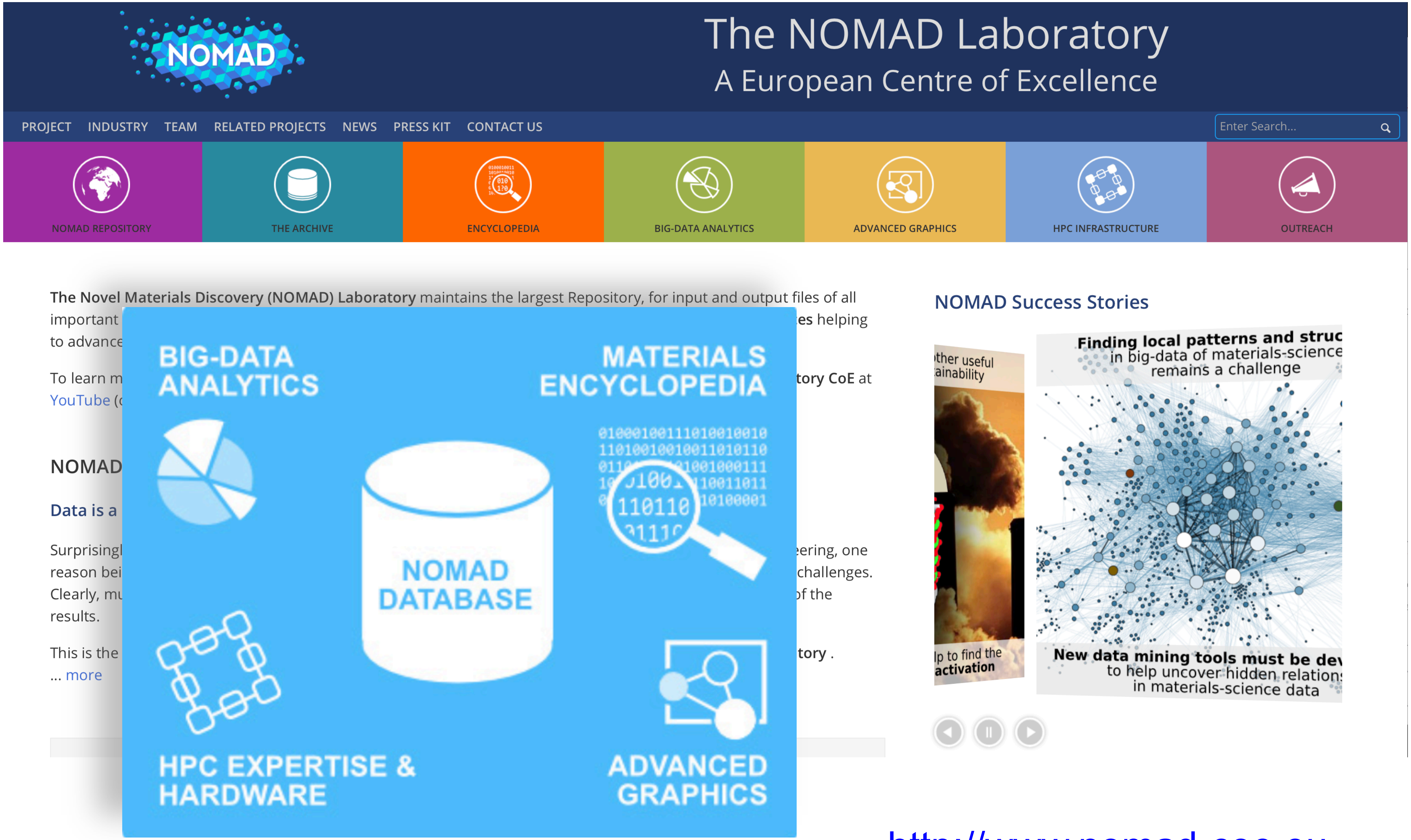
Implanted Neural Networks


Representation

Prediction

Ref: Berk Onat, Ekin D. Cubuk, Brad D. Malone, and Efthimios Kaxiras,
PRB 97, 094106 (2018)

Novel Materials Discovery (NOMAD) Lab



The NOMAD Laboratory

A European Centre of Excellence

The Novel Materials Discovery (NOMAD) Laboratory maintains the largest Repository, for input and output files of all important … es helping to advance …

To learn m… tory CoE at YouTube (…

NOMAD…

Data is a…

Surprisingl… eering, one reason bei… challenges. Clearly, mu… of the results.

This is the … tory .

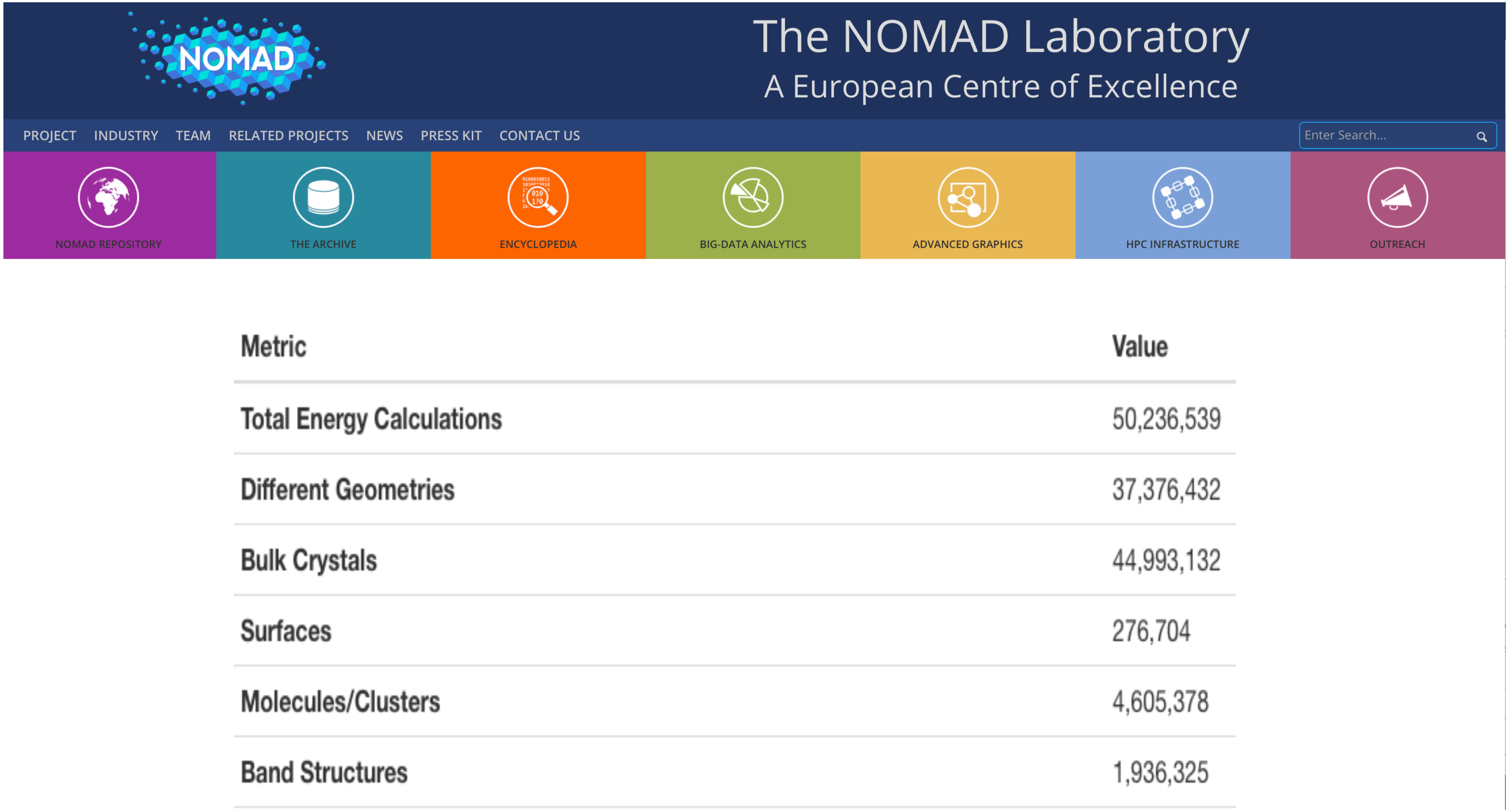
... more

NOMAD Success Stories

…ther useful …ainability
…lp to find the activation
Finding local patterns and struc… in big-data of materials-science remains a challenge
New data mining tools must be dev… to help uncover hidden relation… in materials-science data
BIG-DATA ANALYTICS
MATERIALS ENCYCLOPEDIA
NOMAD DATABASE
HPC EXPERTISE & HARDWARE
ADVANCED GRAPHICS

http://www.nomad-coe.eu

Novel Materials Discovery (NOMAD) Lab



Metric	Value
Total Energy Calculations	50,236,539
Different Geometries	37,376,432
Bulk Crystals	44,993,132
Surfaces	276,704
Molecules/Clusters	4,605,378
Band Structures	1,936,325

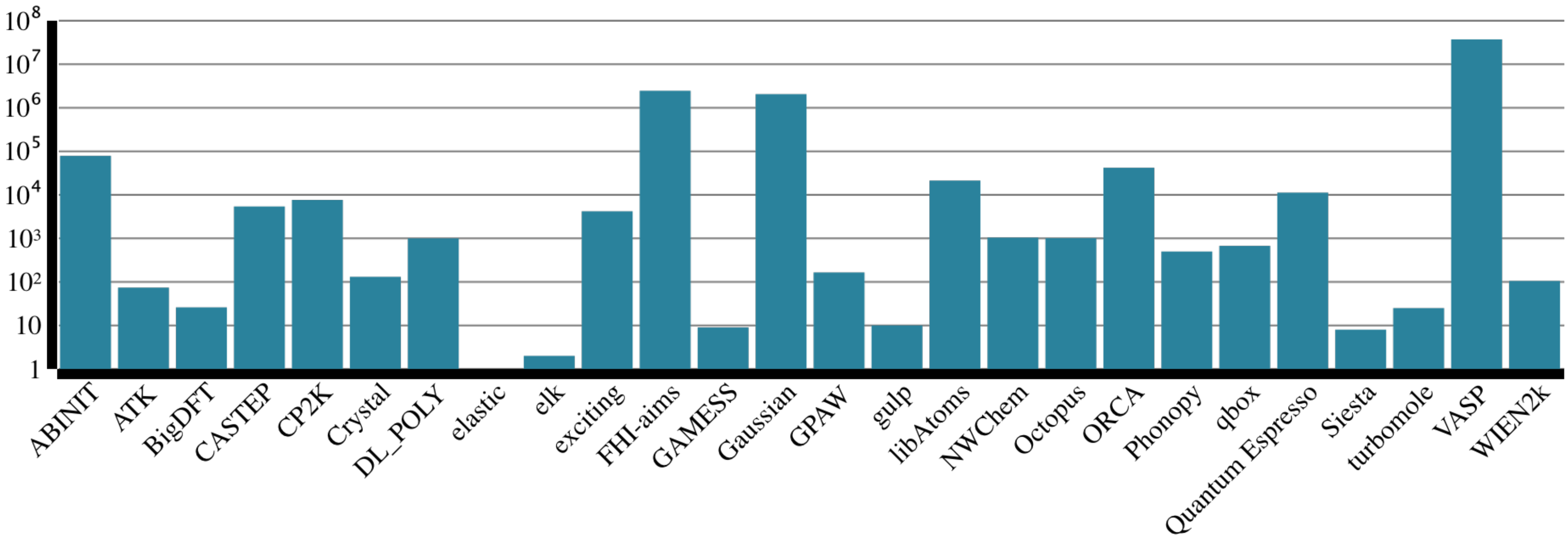
NOMAD Archive as of ***March 2018***.

http://www.nomad-coe.eu

Novel Materials Discovery (NOMAD) Lab

The NOMAD Laboratory

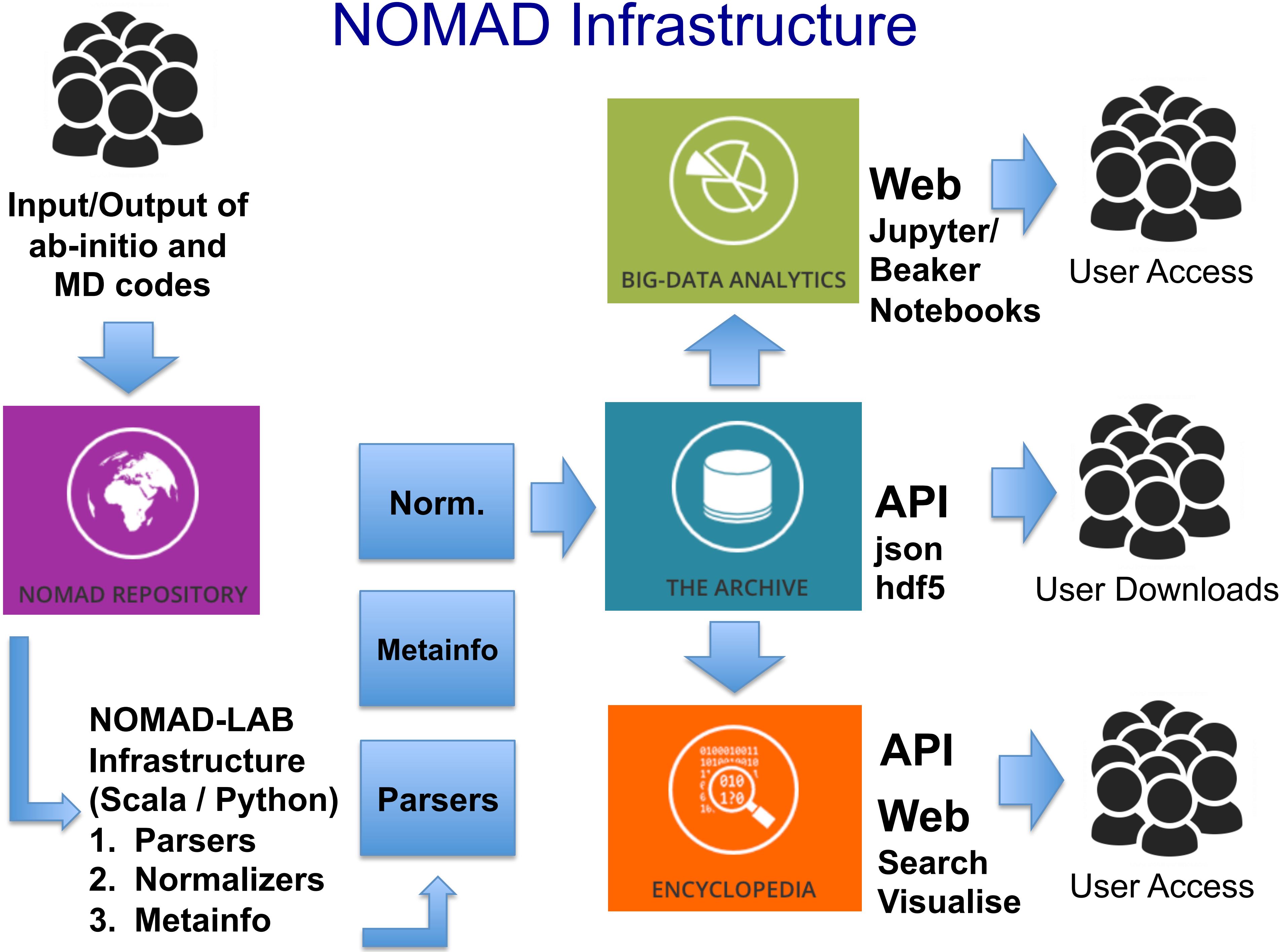
A European Centre of Excellence



Codes with more than 100 uploads to NOMAD Archive as of ***March 2018***.

http://www.nomad-coe.eu

NOMAD Infrastructure



Input/Output of ab-initio and MD codes

NOMAD REPOSITORY

NOMAD-LAB Infrastructure (Scala / Python)

1. Parsers
2. Normalizers
3. Metainfo

Parsers

Metainfo

Norm.

THE ARCHIVE

BIG-DATA ANALYTICS

ENCYCLOPEDIA

Web
Jupyter/
Beaker
Notebooks

User Access

API
json
hdf5

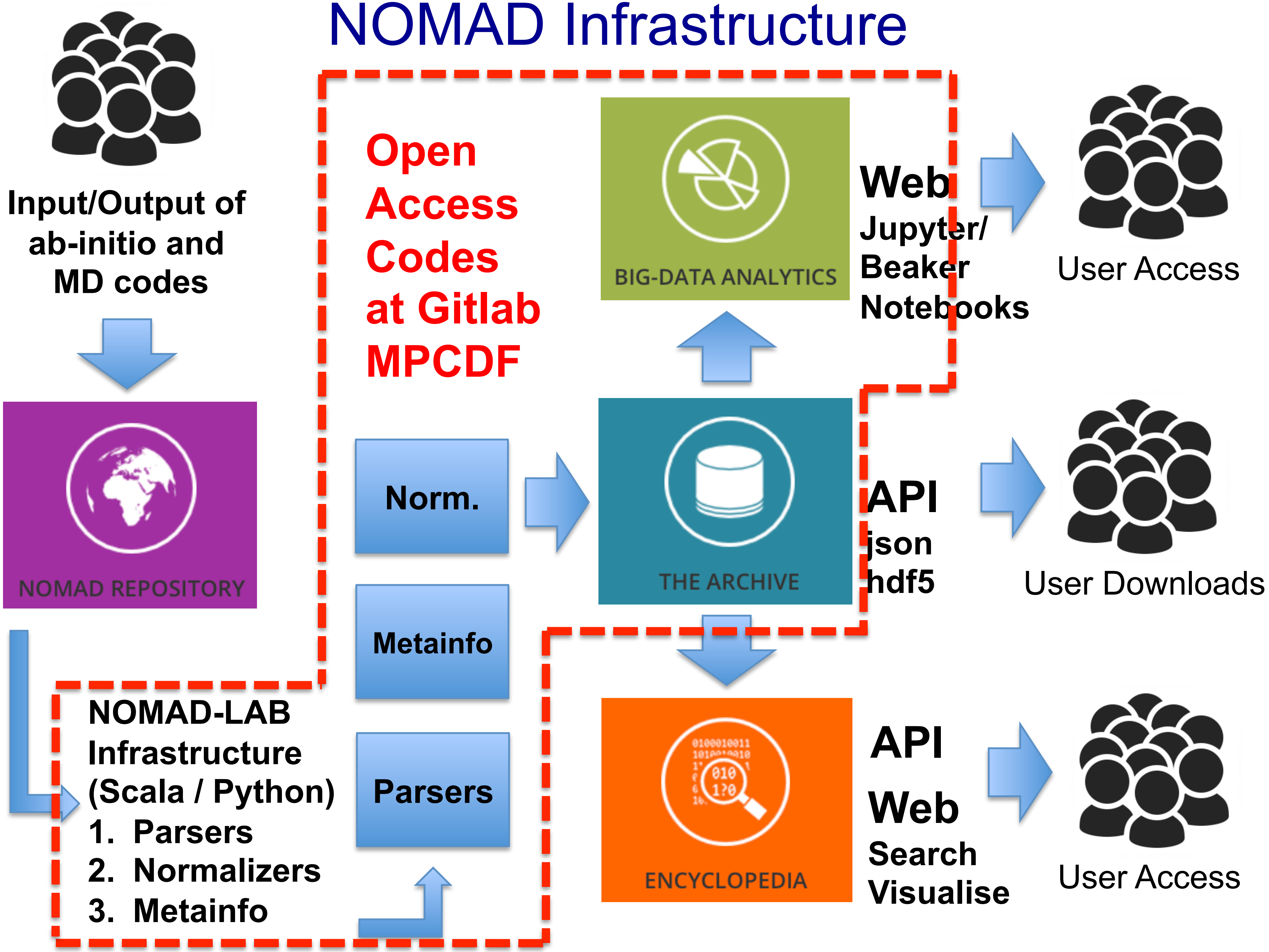
User Downloads

API

Web
Search
Visualise

User Access

NOMAD Infrastructure



Input/Output of ab-initio and MD codes

NOMAD REPOSITORY

NOMAD-LAB Infrastructure (Scala / Python)

1. Parsers
2. Normalizers
3. Metainfo

Open Access Codes at Gitlab MPCDF

Norm.

Metainfo

Parsers

BIG-DATA ANALYTICS

THE ARCHIVE

ENCYCLOPEDIA

Web
Jupyter/
Beaker
Notebooks

API
json
hdf5

API

Web
Search
Visualise

User Access

User Downloads

User Access

NOMAD Parsers

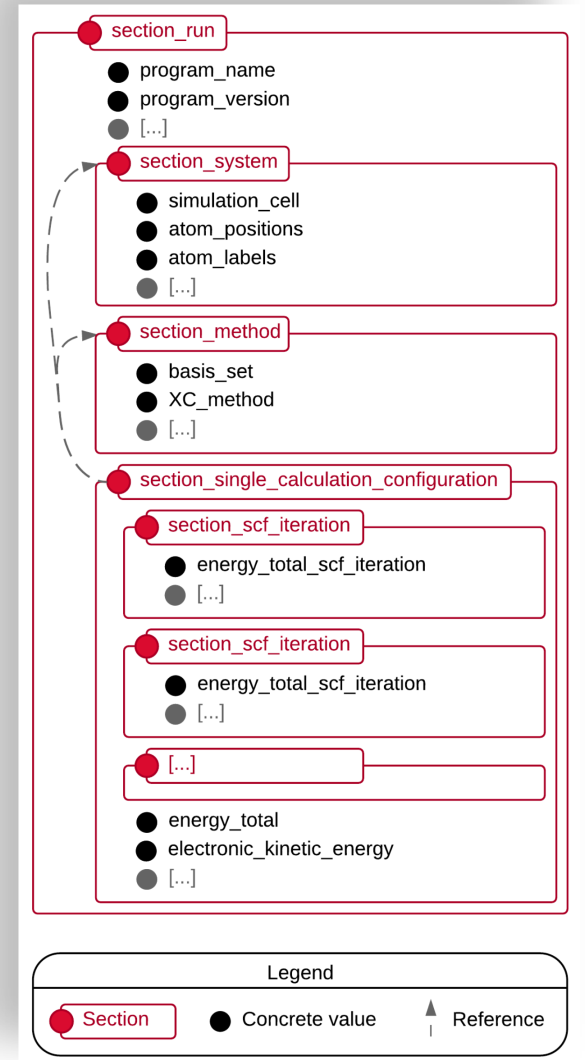
Most cited 15 codes:

Code	Citations (2013-17)	Type	Search Name
🟩 Gaussian	19100	DFT	Frisch
🟩 VASP	17900	DFT	Kresse
🟦 Gromacs	11200	FF	Lindahl
🟩 LAMMPS	10300	FF	Plimpton
🟦 Amber	9440	FF	Kollman
🟦 NAMD	7110	FF	Schulten
🟦 GROMOS	7080	FF	Van Gunsteren
🟩 Quantum Espresso	6960	DFT	Giannozzi
🟩 ASE/ASAP	6650	FF	Jacobsen
🟦 CHARMM	6250	FF	Karplus
Discovery Studio	6240	DFT, FF	*Accelyrs*
🟩 GAMESS	5780	DFT	Gordon
🟩 WIEN2k	5570	DFT	Blaha
🟩 CASTEP	5330	DFT	Payne
Molpro	4440	DFT	Werner

🟦 **Parser codes developed in our group.**

🟩 **Parsers from other groups in NOMAD.**

Standard Metadata



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
[...]

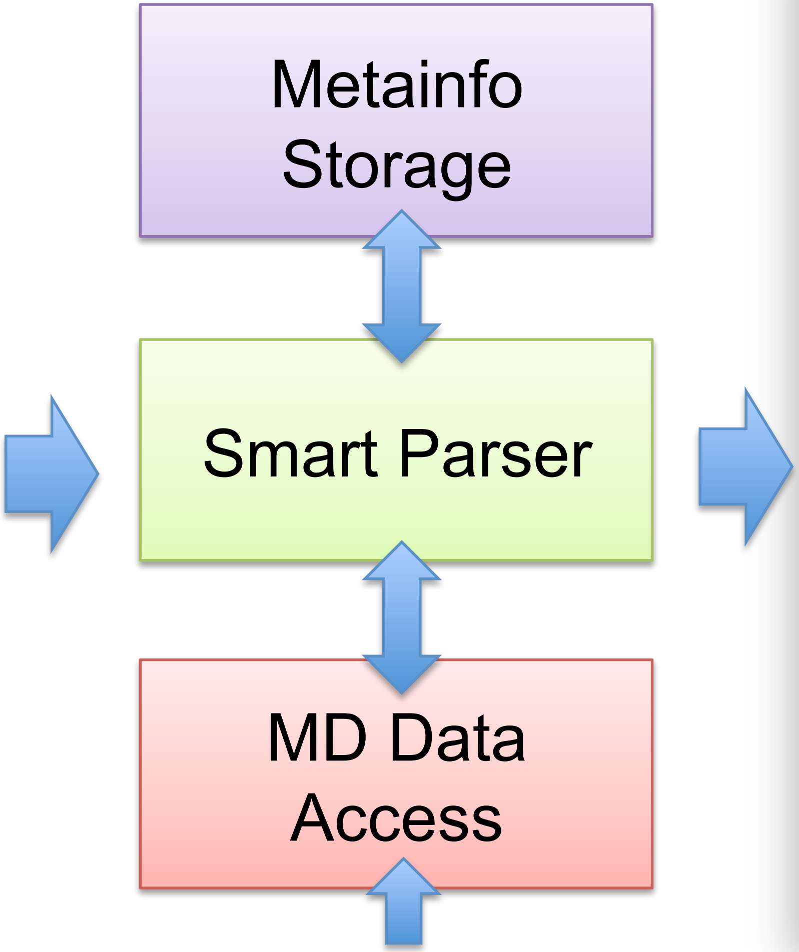
Legend
Section
Concrete value
Reference

Ref: L.M. Ghiringhelli, C. Carbogno, S. Levchenko, F. Mohamed, G. Huhs, M. Lueders, M. Oliveira, M. Scheffler, arXiv:1607.04738

Metainfo for MD Codes

https://gitlab.mpcdf.mpg.de/nomad-lab/python-common
https://gitlab.mpcdf.mpg.de/nomad-lab/pymolfile

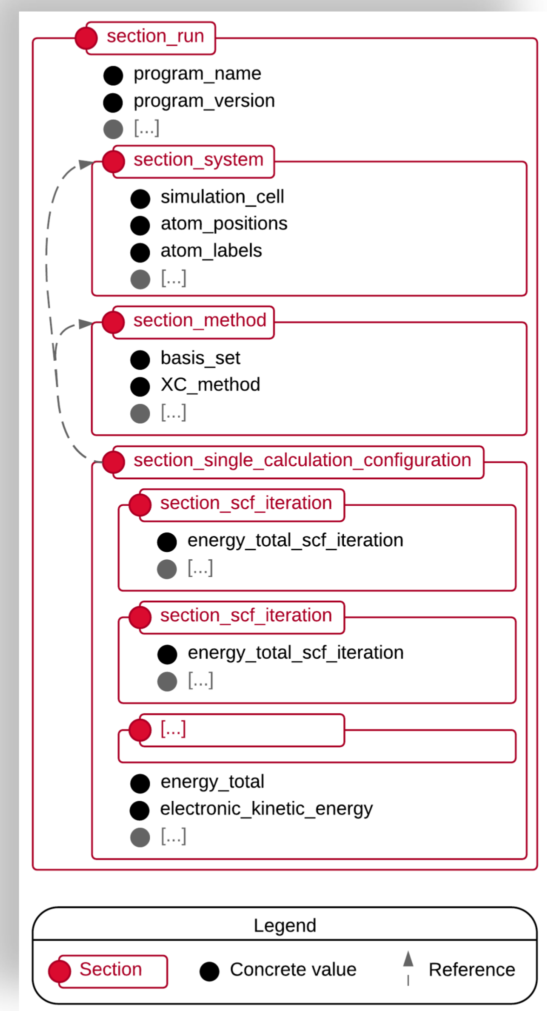
Amber
CHARMM
Gromacs
GROMOS
NAMD
Tinker



Metainfo Storage

Smart Parser

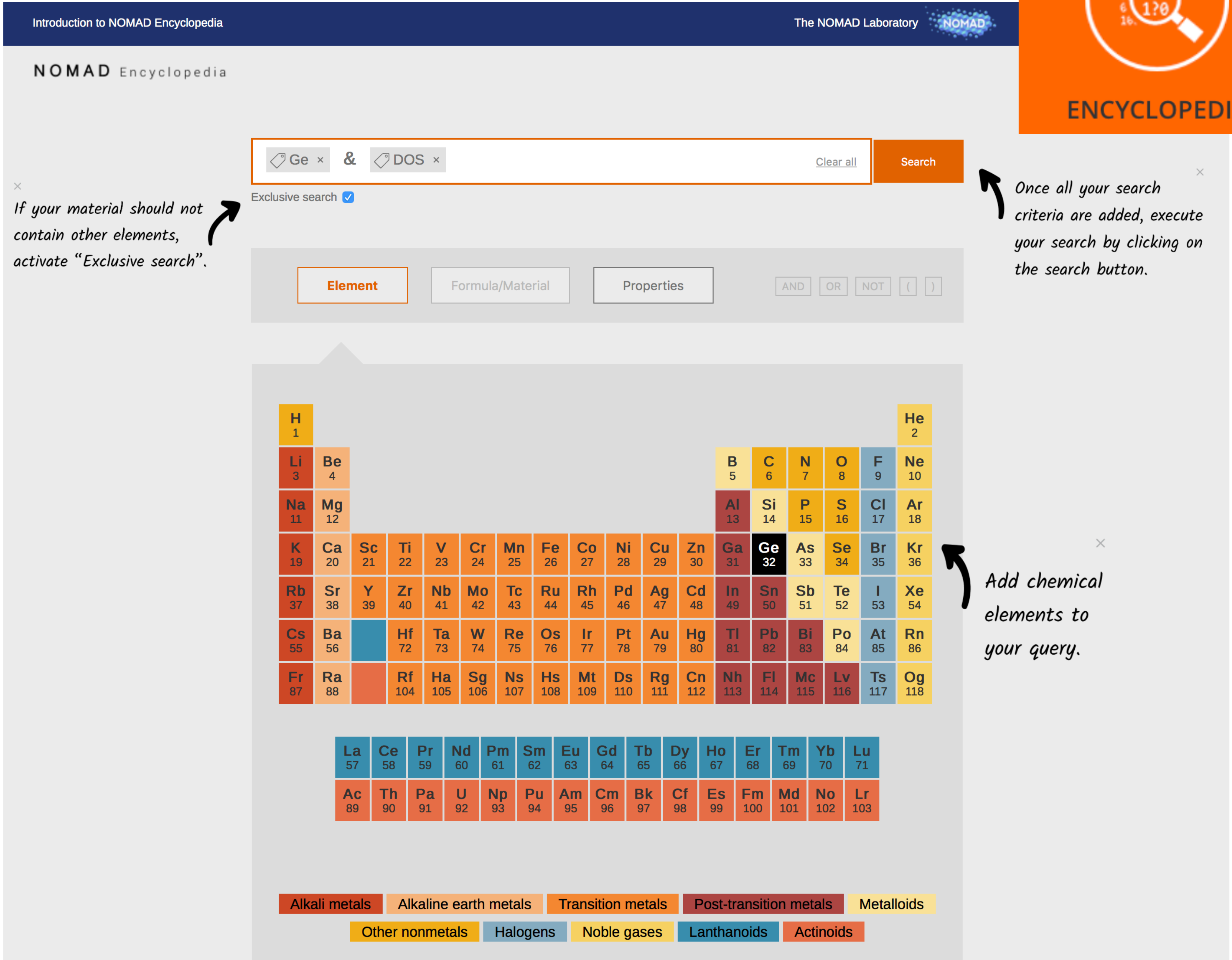
MD Data Access



Supports ***128*** Topology/Trajectory : Formats

ASE, Mdtraj, MdAnalysis, **ParmEd**,
GROMOSTopo, CHARMM_Reader,
Pymolfile (VMD plugins)

NOMAD Encyclopedia



Introduction to NOMAD Encyclopedia

The NOMAD Laboratory NOMAD

NOMAD Encyclopedia

Ge × & DOS × Clear all Search

☑ Exclusive search

If your material should not contain other elements, activate "Exclusive search".

Once all your search criteria are added, execute your search by clicking on the search button.

Element | Formula/Material | Properties | AND | OR | NOT | (|)

Add chemical elements to your query.

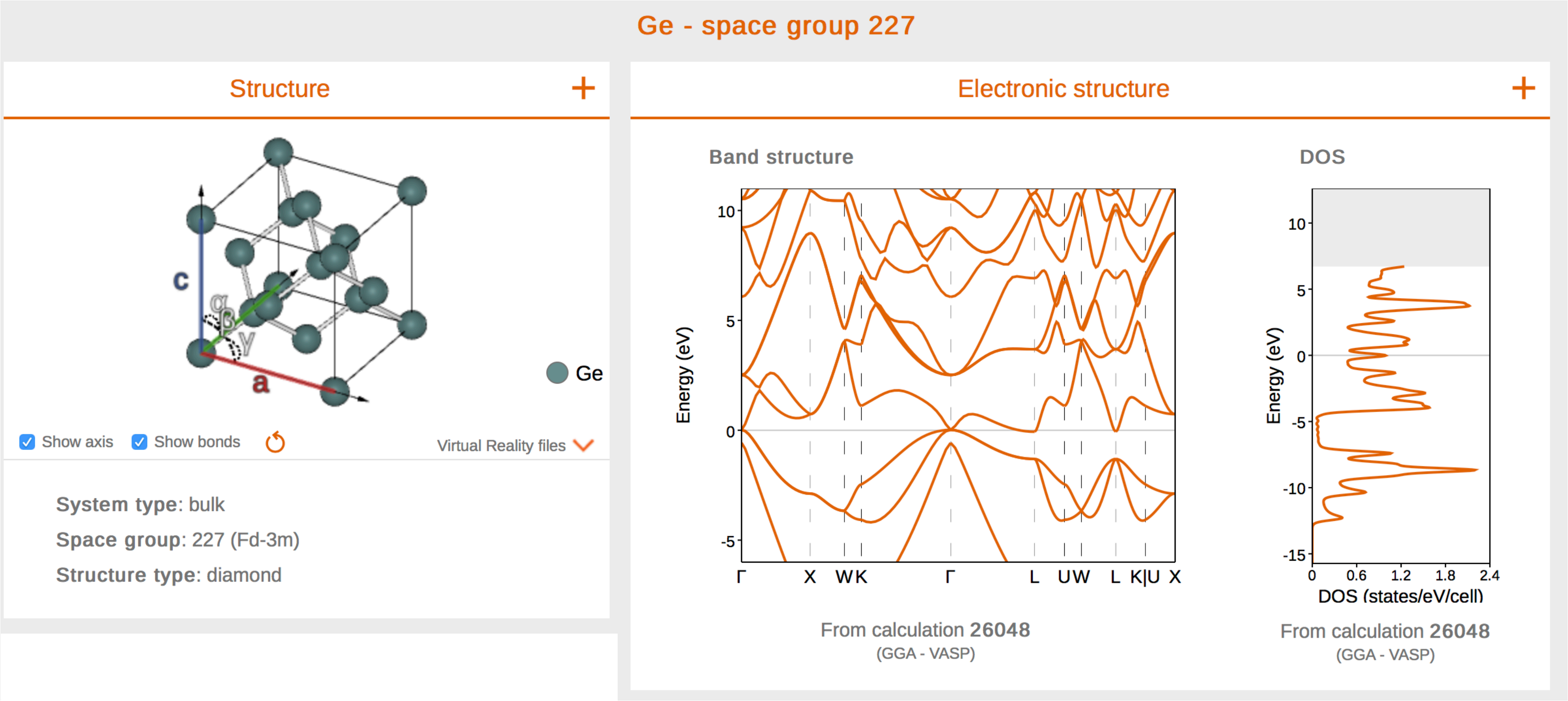
H 1																	He 2
Li 3	Be 4											B 5	C 6	N 7	O 8	F 9	Ne 10
Na 11	Mg 12											Al 13	Si 14	P 15	S 16	Cl 17	Ar 18
K 19	Ca 20	Sc 21	Ti 22	V 23	Cr 24	Mn 25	Fe 26	Co 27	Ni 28	Cu 29	Zn 30	Ga 31	Ge 32	As 33	Se 34	Br 35	Kr 36
Rb 37	Sr 38	Y 39	Zr 40	Nb 41	Mo 42	Tc 43	Ru 44	Rh 45	Pd 46	Ag 47	Cd 48	In 49	Sn 50	Sb 51	Te 52	I 53	Xe 54
Cs 55	Ba 56		Hf 72	Ta 73	W 74	Re 75	Os 76	Ir 77	Pt 78	Au 79	Hg 80	Tl 81	Pb 82	Bi 83	Po 84	At 85	Rn 86
Fr 87	Ra 88		Rf 104	Ha 105	Sg 106	Ns 107	Hs 108	Mt 109	Ds 110	Rg 111	Cn 112	Nh 113	Fl 114	Mc 115	Lv 116	Ts 117	Og 118

La 57	Ce 58	Pr 59	Nd 60	Pm 61	Sm 62	Eu 63	Gd 64	Tb 65	Dy 66	Ho 67	Er 68	Tm 69	Yb 70	Lu 71
Ac 89	Th 90	Pa 91	U 92	Np 93	Pu 94	Am 95	Cm 96	Bk 97	Cf 98	Es 99	Fm 100	Md 101	No 102	Lr 103

Alkali metals | Alkaline earth metals | Transition metals | Post-transition metals | Metalloids

Other nonmetals | Halogens | Noble gases | Lanthanoids | Actinoids

NOMAD Encyclopedia



Ge - space group 227

Structure +

☑ Show axis ☑ Show bonds

Virtual Reality files

System type: bulk

Space group: 227 (Fd-3m)

Structure type: diamond

Electronic structure +

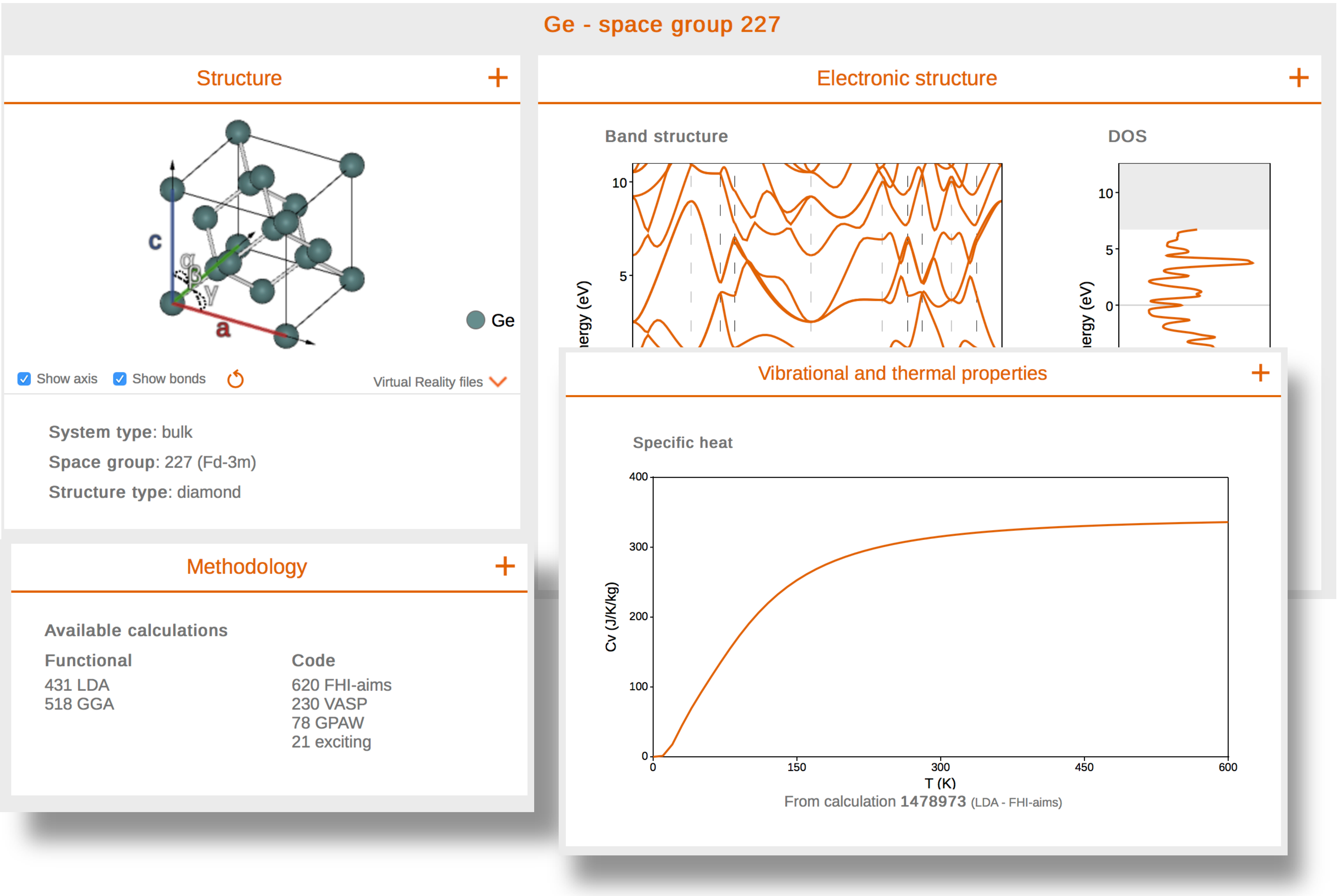
Band structure

From calculation **26048**
(GGA - VASP)

DOS

From calculation **26048**
(GGA - VASP)

NOMAD Encyclopedia



Ge - space group 227

Structure

☑ Show axis ☑ Show bonds

System type: bulk

Space group: 227 (Fd-3m)

Structure type: diamond

Methodology

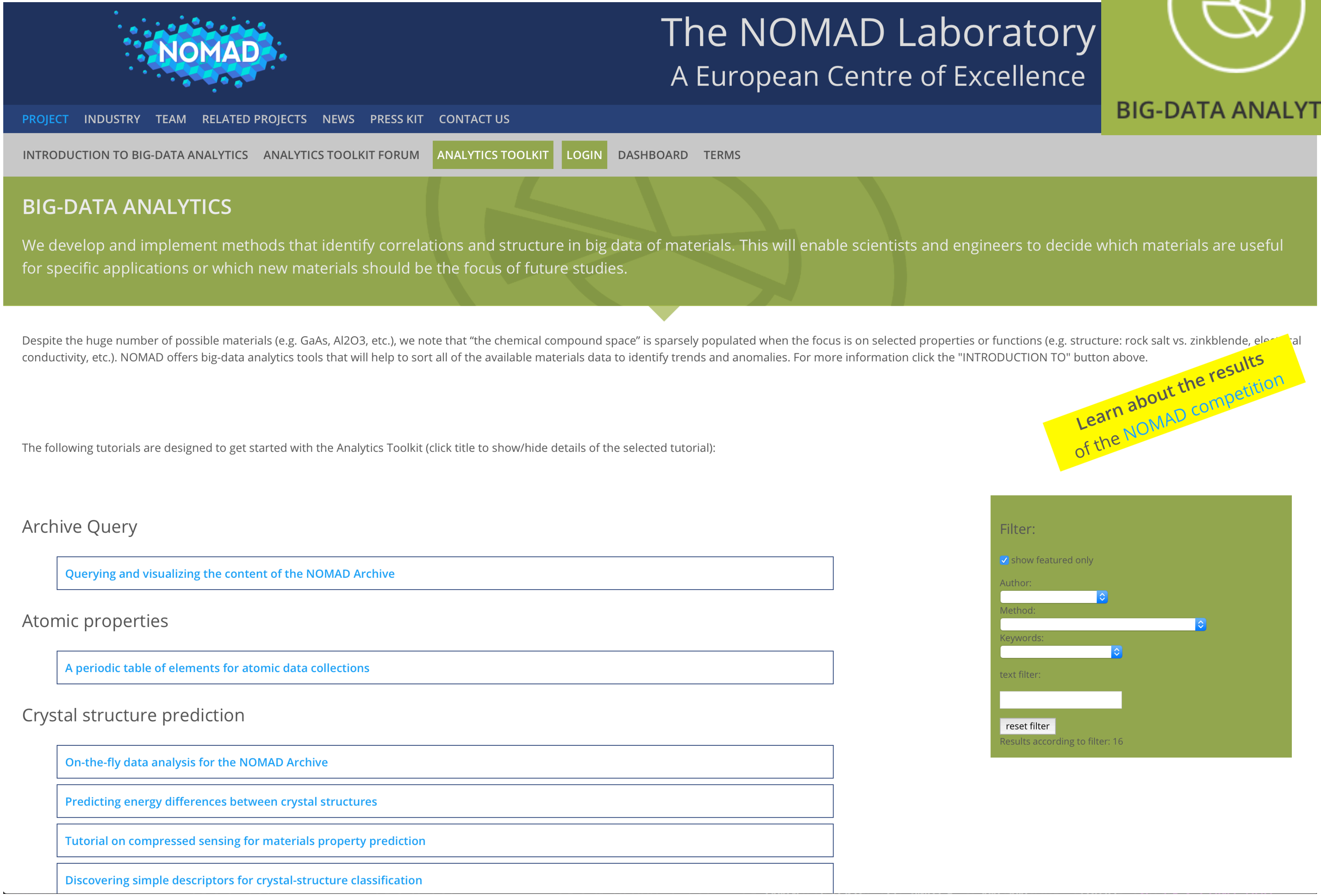
Available calculations

Functional	Code
431 LDA 518 GGA	620 FHI-aims 230 VASP 78 GPAW 21 exciting

From calculation **1478973** (LDA - FHI-aims)

NOMAD Big-Data Analytics & Query

BIG-DATA ANALYTICS



The NOMAD Laboratory

A European Centre of Excellence

BIG-DATA ANALYTICS

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.

Despite the huge number of possible materials (e.g. GaAs, Al2O3, etc.), we note that “the chemical compound space” is sparsely populated when the focus is on selected properties or functions (e.g. structure: rock salt vs. zinkblende, electrical conductivity, etc.). NOMAD offers big-data analytics tools that will help to sort all of the available materials data to identify trends and anomalies. For more information click the "INTRODUCTION TO" button above.

Learn about the results of the NOMAD competition

The following tutorials are designed to get started with the Analytics Toolkit (click title to show/hide details of the selected tutorial):

Archive Query

- Querying and visualizing the content of the NOMAD Archive

Atomic properties

- A periodic table of elements for atomic data collections

Crystal structure prediction

- On-the-fly data analysis for the NOMAD Archive
- Predicting energy differences between crystal structures
- Tutorial on compressed sensing for materials property prediction
- Discovering simple descriptors for crystal-structure classification

Filter:

☑ show featured only

Author:

Method:

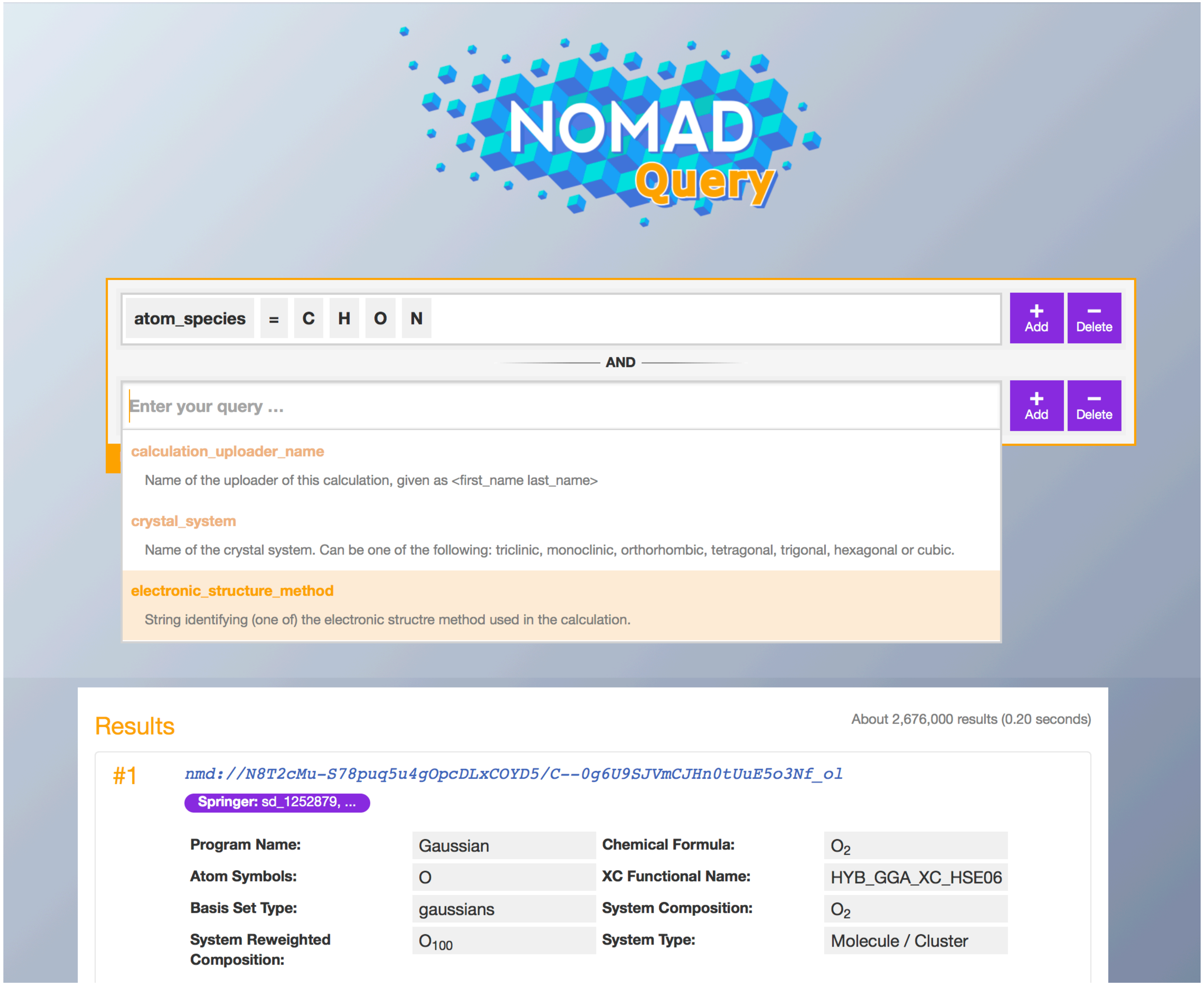
Keywords:

text filter:

reset filter

Results according to filter: 16

NOMAD Big-Data Archive and Query



calculation_uploader_name

Name of the uploader of this calculation, given as <first_name last_name>

crystal_system

Name of the crystal system. Can be one of the following: triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal or cubic.

electronic_structure_method

String identifying (one of) the electronic structre method used in the calculation.

Results

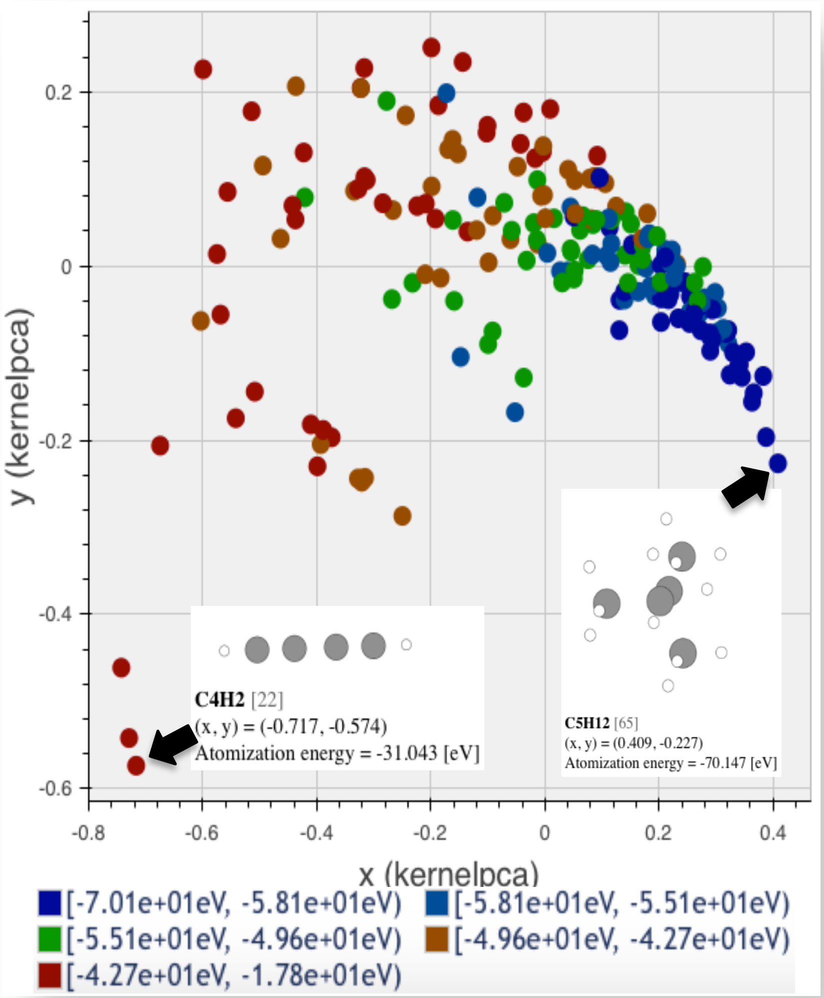
About 2,676,000 results (0.20 seconds)

#1 *nmd://N8T2cMu-S78puq5u4gOpcDLxCOYD5/C--0g6U9SJVmCJHn0tUuE5o3Nf_ol*

Springer: sd_1252879, ...

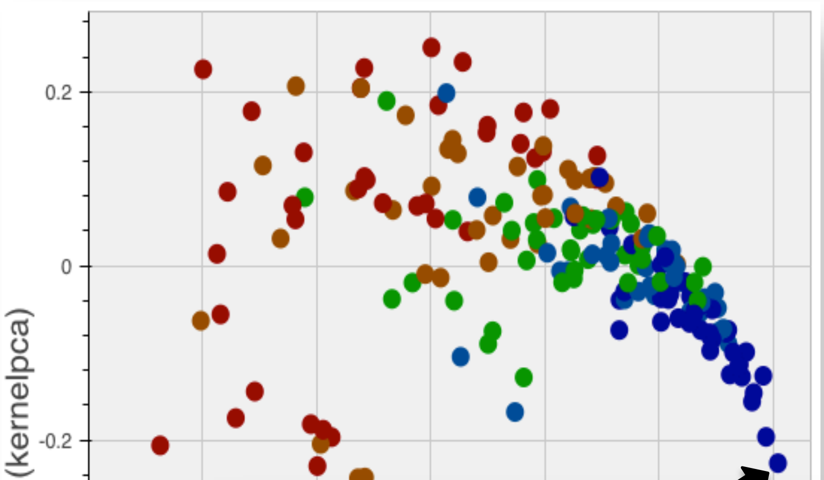
Program Name:	Gaussian	**Chemical Formula:**	O_2
Atom Symbols:	O	**XC Functional Name:**	HYB_GGA_XC_HSE06
Basis Set Type:	gaussians	**System Composition:**	O_2
System Reweighted Composition:	O_{100}	**System Type:**	Molecule / Cluster

Similarity Map with SOAP Representation



[1] C. Poelking, A. Ziletti, L. Ghiringhelli, and G. Csányi, NOMAD. S. De, A. Bartók, G. Csányi, and M. Ceriotti, Phys. Chem. Chem. Phys. (2016)

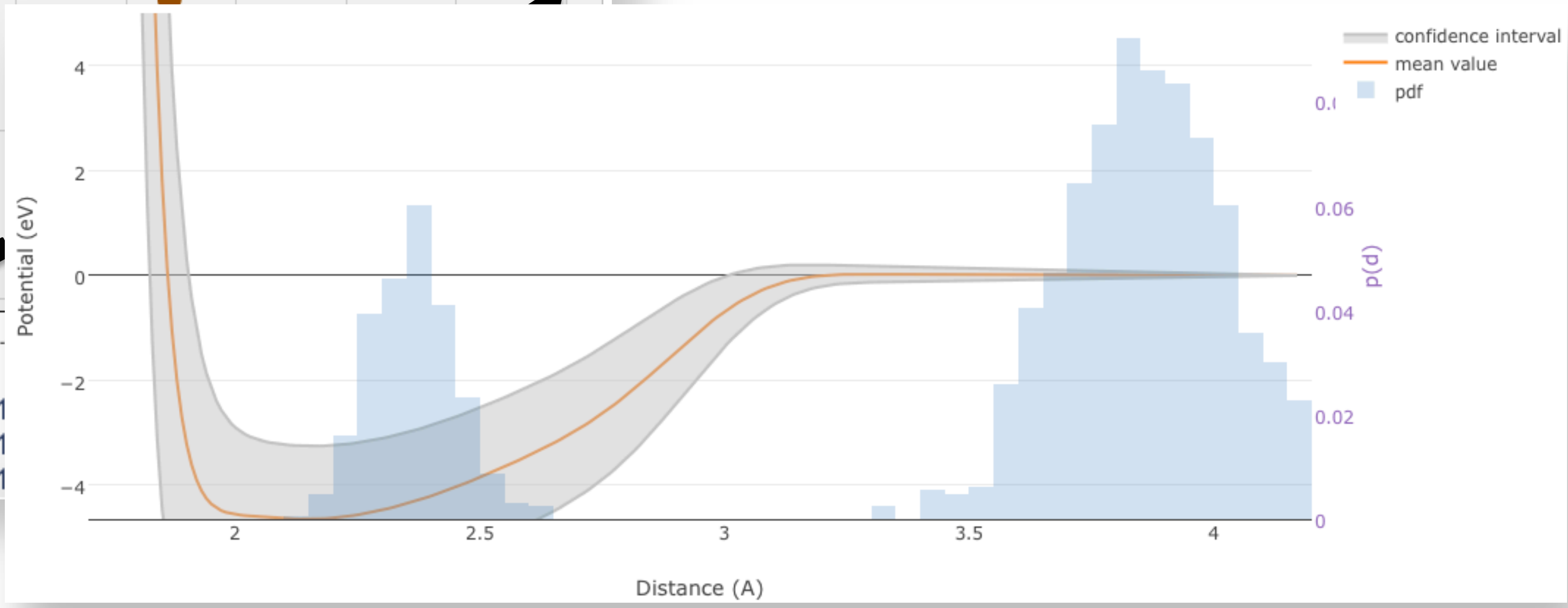
Similarity Map with SOAP Representation



[1] C. Poelking, A. Ziletti, L. Ghiringhelli, and G. Csányi, NOMAD. S. De, A. Bartók, G. Csányi, and M. Ceriotti, Phys. Chem. Chem. Phys. (2016)

[2] Ádám Fekete, Aldo Glielmo, Martina Stella, and Alessandro De Vita, NOMAD Big-Data Analytics

Pair-Potential Predictor for Si



Classification of Representations

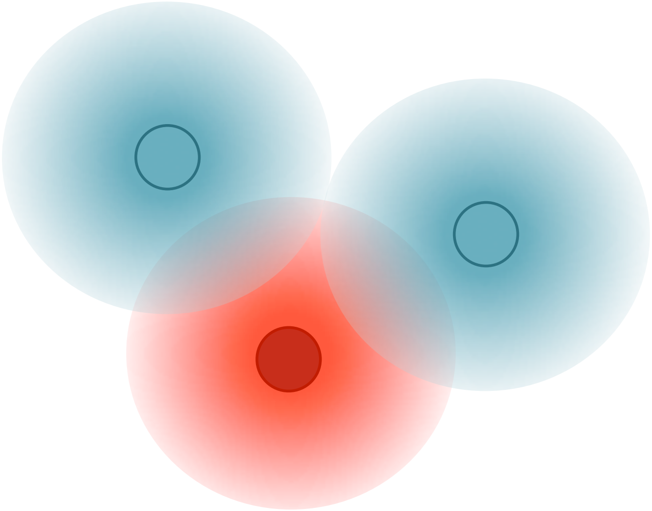
Atomic neighbor density			Histogram		Connectivity (Graphs)		Physical Property
Symmetry Adapted		Direct	Direct		Direct		Direct
Kernel/ Power expension	Functions/ Filters	Expension/Tensor	Structure	Electronic	Static Builds	Hierarchical Builds	Combination of any property with any math. operation
SOAP	Sym. Funcs, Chebyshev expansion, AGNI	MBTR	k-bags, Bag of Bonds/ Angles …	Electronic Band Structure, DOS	n-gram, Graphs, CGCNN,Coulomb Matrix	Building blocks or Ligand/ Residue based Graphs	SISSO

Ref: Berk Onat, James Kermode (2018) ***under preparation***

Smooth Overlap of Atomic Positions (SOAP)

$$\rho_{\mathscr{X}}(\mathbf{r}) = \sum_{i\in\mathscr{X}} \exp\left(-\frac{(\mathbf{x}_i-\mathbf{r})^2}{2\sigma^2}\right)$$

$$\tilde{k}(\mathscr{X},\mathscr{X}') = \int \mathrm{d}\hat{R}\left|\int \rho_{\mathscr{X}}(\mathbf{r})\rho_{\mathscr{X}'}(\hat{R}\mathbf{r})\mathrm{d}\mathbf{r}\right|^n$$



Power Expansion:

$$\rho_{\mathscr{X}}(\mathbf{r}) = \sum_{blm} c_{blm} g_b(|\mathbf{r}|) Y_{lm}(\hat{\mathbf{r}})$$

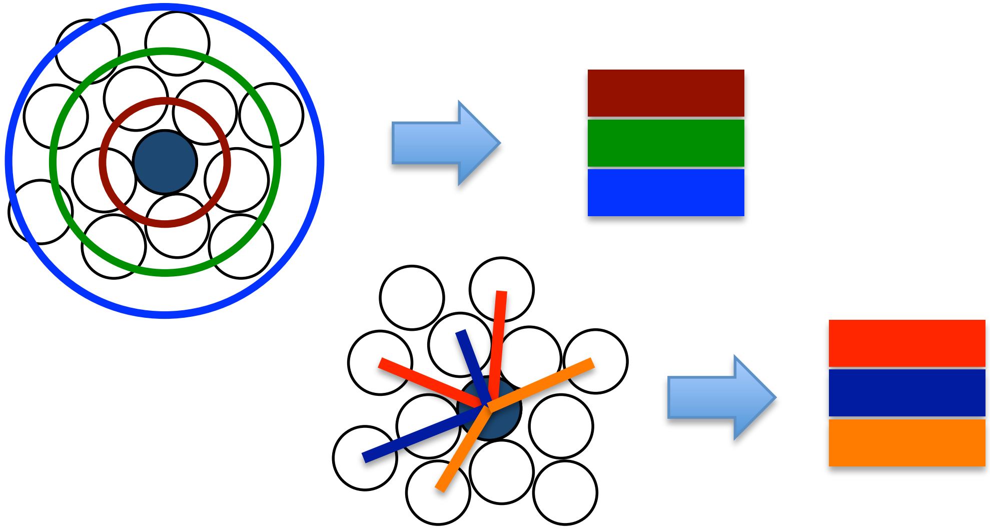
$$p(\mathscr{X})_{b_1b_2l} = \pi\sqrt{\frac{8}{2l+1}}\sum_m (c_{b_1lm})^\dagger c_{b_2lm}$$

SOAP Kernel:

$$k(\mathscr{X},\mathscr{X}') = \hat{\mathbf{p}}(\mathscr{X})\cdot\hat{\mathbf{p}}(\mathscr{X}')$$

Ref: **A. P. Bartok, R. Kondor and G. Csanyi, PRB 87, 184115, (2013)**
A. P. Bartok, G. Csanyi, Int. J. Quantum Chemistry 115, 1051–1057, (2015)

Symmetry Functions



Radial Symmetry Functions:

$$G_2^i = \sum_j e^{-\eta (R_{ij} - R_s)^2} \cdot f_c(R_{ij}),$$

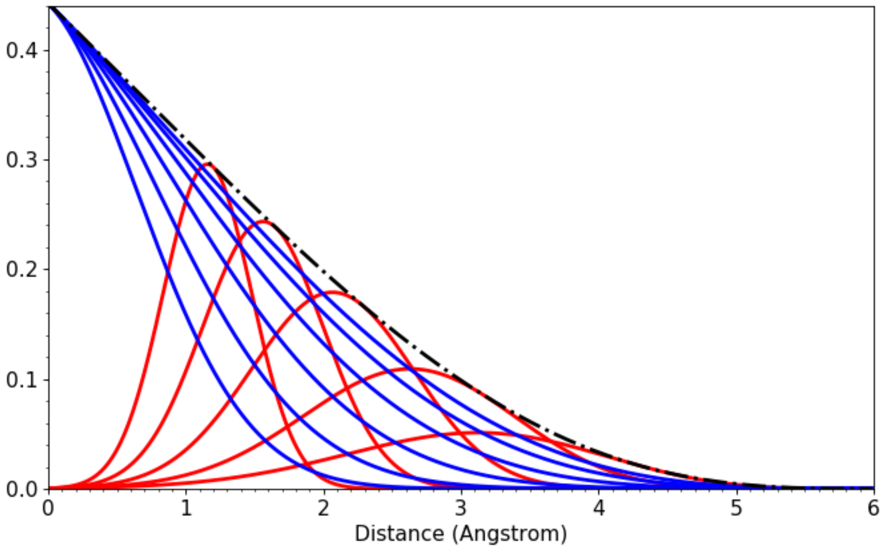
Angular Symmetry Functions:

$$G_3^i = 2^{1-\zeta} \sum_j \sum_{k \neq j} (1 + \lambda \cdot \cos\theta_{ijk})^\zeta$$

$$\cdot\, e^{-\eta (R_{ij}^2 + R_{ik}^2 + R_{jk}^2)} \cdot f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}),$$

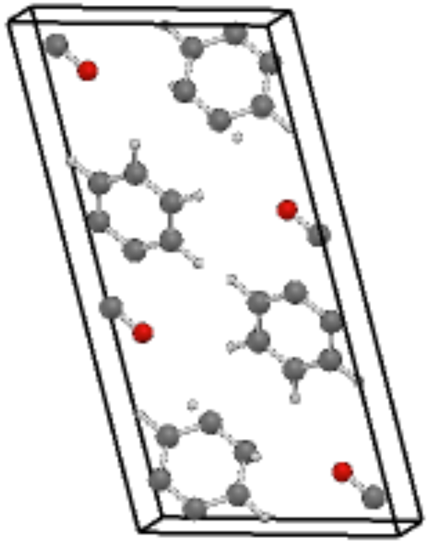
Cutoff function:

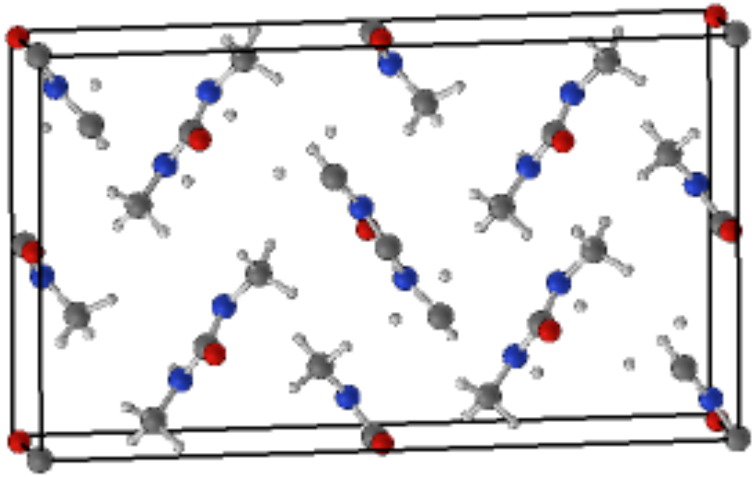
$$f_c(R_{ij}) = \begin{cases} \tanh^3\left[1 - \frac{R_{ij}}{r_c}\right] & \text{for } R_{ij} \le r_c \\ 0.0 & \text{for } R_{ij} > r_c \end{cases}$$

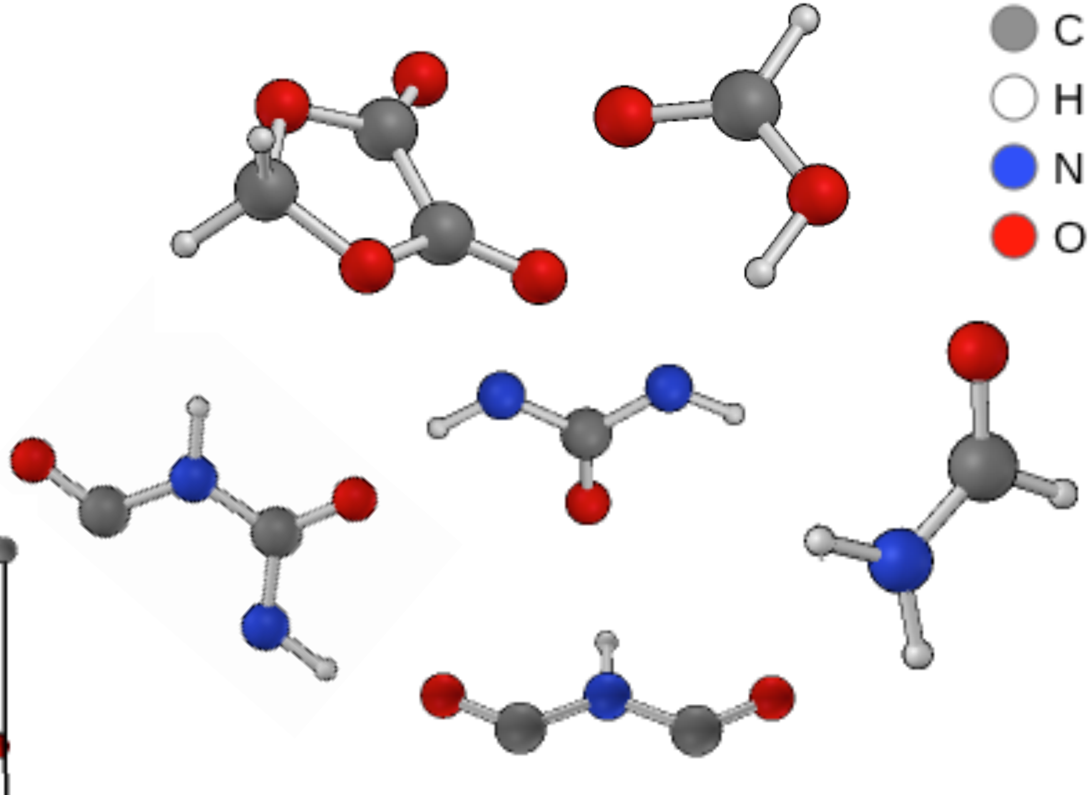


Ref: N. Artrith, B. Hiller and J. Behler, Physica Status Solidi B, 250, 1191 (2013)
G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, M. Ceriotti, J Chem. Phys 148, 241730 (2018)

Compressibility Tests with CUR and FPS Methods on C,H,N,O Structures







2963 Molecular Structures
(Single Molecules, Bulk Molecular Structures)

CUR Decomposition

$$\mathbf{X} \approx \tilde{\mathbf{X}} = \mathbf{C}\,\mathbf{U}\,\mathbf{R},$$

Selecting Columns of X:

1) **SVD Decomposition**
2) **Column scoring with right singular vector *v***

$$\pi_c = \sum_{j=1}^{k} (v_c^{(j)})^2,$$

3) ***Select c from k columns (Here we apply CUR[k=1])***
4) **Orthogonalize all other columns of X to *c***

$$\mathbf{R} = \mathbf{X},$$
$$\mathbf{U} = \mathbf{C}^{+}\mathbf{X}\mathbf{X}^{+},$$

Error estimation:

$$\epsilon = \|\mathbf{X} - \mathbf{CUR}\|_F \;/\; \|\mathbf{X}\|_F$$

Ref: **G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, M. Ceriotti, J Chem. Phys 148, 241730 (2018)**

Farthest Point Sampling (FPS)

Selecting Columns of X:

1) **Select first column randomly**
2) **Select *N* columns according to**

$$k = \mathrm{argmax}(\min_j |X_k - X_j|),$$

3) **Build C matrix using the selected columns**

$$\mathbf{R} = \mathbf{X},$$
$$\mathbf{U} = \mathbf{C}^{+}\mathbf{X}\mathbf{X}^{+},$$

Error estimation:

$$\epsilon = ||\mathbf{X} - \mathbf{CUR}||_F \, / \, ||\mathbf{X}||_F$$

Ref: **G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, M. Ceriotti, J Chem. Phys 148, 241730 (2018)**

Horizon 2020

NOMAD: Novel Materials Discovery

http://www.nomad-coe.eu

Dr. Berk Onat
Research Fellow

Dr. James R. Kermode
Associate Professor

WCPM

Warwick Centre of Predictive Modelling

NOMAD Collaborators

MAX-PLANCK-GESELLSCHAFT
MPCDF
KING'S College LONDON
UNIVERSITY OF CAMBRIDGE
HUMBOLDT-UNIVERSITÄT ZU BERLIN
UNIVERSITAT DE BARCELONA
Aalto University School of Science
CSC
mpsd Max Planck Institute for the Structure and Dynamics of Matter
DTU Technical University of Denmark
BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación
lrz Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften

Acknowledgement

This work is supported in part by **NOMAD Project** and **University of Warwick**.

NOMAD project has received funding from the **European Union's Horizon 2020** research and innovation programme under grant agreement No. 676580.

Horizon 2020