NOMAD NOVEL - MATERIALS DISCOVERY LABORATORY

FRITZ-HABER-INSTITUT MAX-PLANCK-GESELLSCHAFT

IRIS ADLERSHOF

HUMBOLDT-UNIVERSITÄT ZU BERLIN

The NOMAD Artificial-Intelligence Toolkit: Web-Based FAIR-Data-Driven Materials Science

Luigi Sbailò

Humboldt –Universität zu Berlin

Workshop on NFDI tools / services / synergies between physics-related consortia and others • April 5, 2022

A web based Artificial-Intelligence toolkit

The NOMAD Archive includes data from AFLOW, Materials Project, OQMD, and more, and from numerous individual researchers.

A web based Artificial-Intelligence toolkit

The NOMAD Archive includes data from AFLOW, Materials Project, OQMD, and more, and from numerous individual researchers.

Findable **A**ccessible **I**nteroperable **R**eusable (**FAIR**) data

A web based Artificial-Intelligence toolkit

The NOMAD Archive includes data from AFLOW, Materials Project, OQMD, and more, and from numerous individual researchers.

➡ **F**indable **AI** **R**eady data

A web based Artificial-Intelligence toolkit

The NOMAD Archive includes data from AFLOW, Materials Project, OQMD, and more, and from numerous individual researchers.

➡ **F**indable **AI R**eady data

The *NOMAD Archive* can be queried from the *Artificial-Intelligence Toolkit*, and the data retrieved can be analysed on the fly with the desired AI tools.

A web based Artificial-Intelligence toolkit

The NOMAD Archive includes data from AFLOW, Materials Project, OQMD, and more, and from numerous individual researchers.

➡ **F**indable **AI R**eady data

The *NOMAD Archive* can be queried from the *Artificial-Intelligence Toolkit*, and the data retrieved can be analysed on the fly with the desired AI tools.

The *Artificial-Intelligence Toolkit* hosts tutorials for hands-on learning of the most popular AI tools.

A web based Artificial-Intelligence toolkit

The NOMAD Archive includes data from AFLOW, Materials Project, OQMD, and more, and from numerous individual researchers.

➡ **F**indable **AI R**eady data

The *NOMAD Archive* can be queried from the *Artificial-Intelligence Toolkit*, and the data retrieved can be analysed on the fly with the desired AI tools.

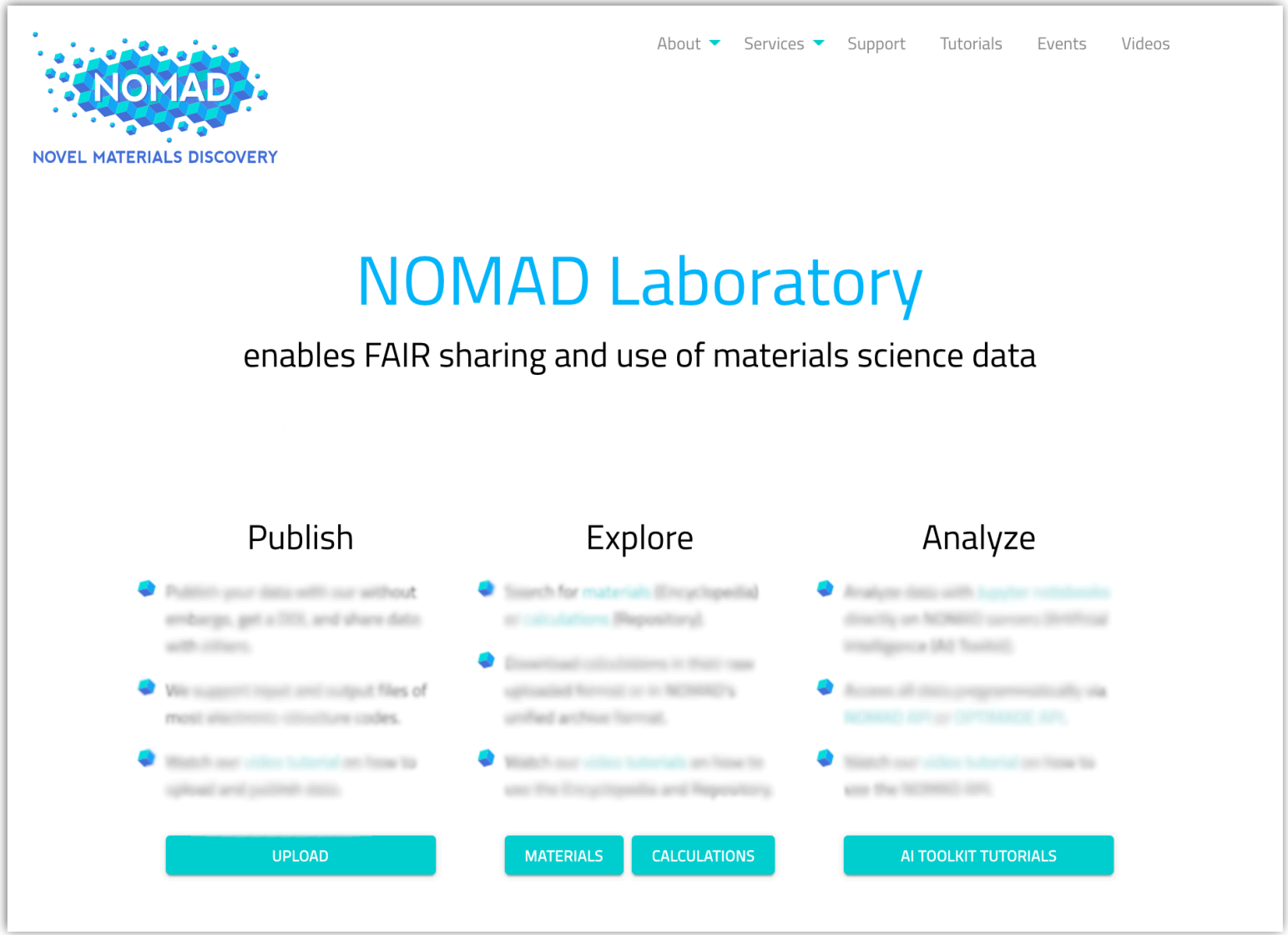
The *Artificial-Intelligence Toolkit* hosts tutorials for hands-on learning of the most popular AI tools.

In the *NOMAD Artificial-Intelligence Toolkit*, we maintain several notebooks demonstrating the latest applications of AI to materials science data.

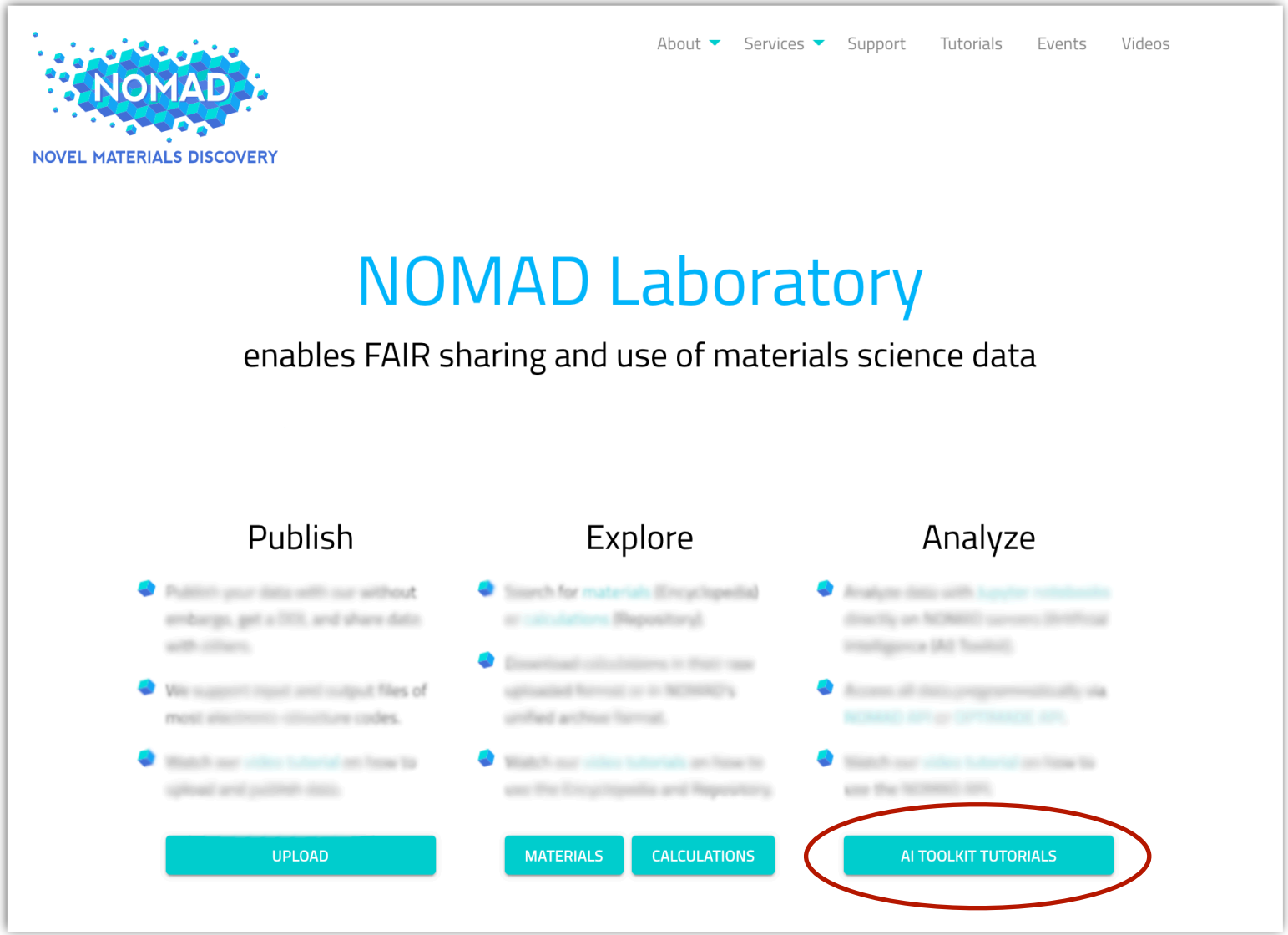
Technology

- Jupyter Notebook is used as environment for tutorials.
- Machine-learning packages, i.e. scikit-learn, tensorflow, …, can be imported within the environment for immediate deployment.
- Servers deployed by the NOMAD AI toolkit are hosted by the Max Planck Computer & Data Facility in Garching.

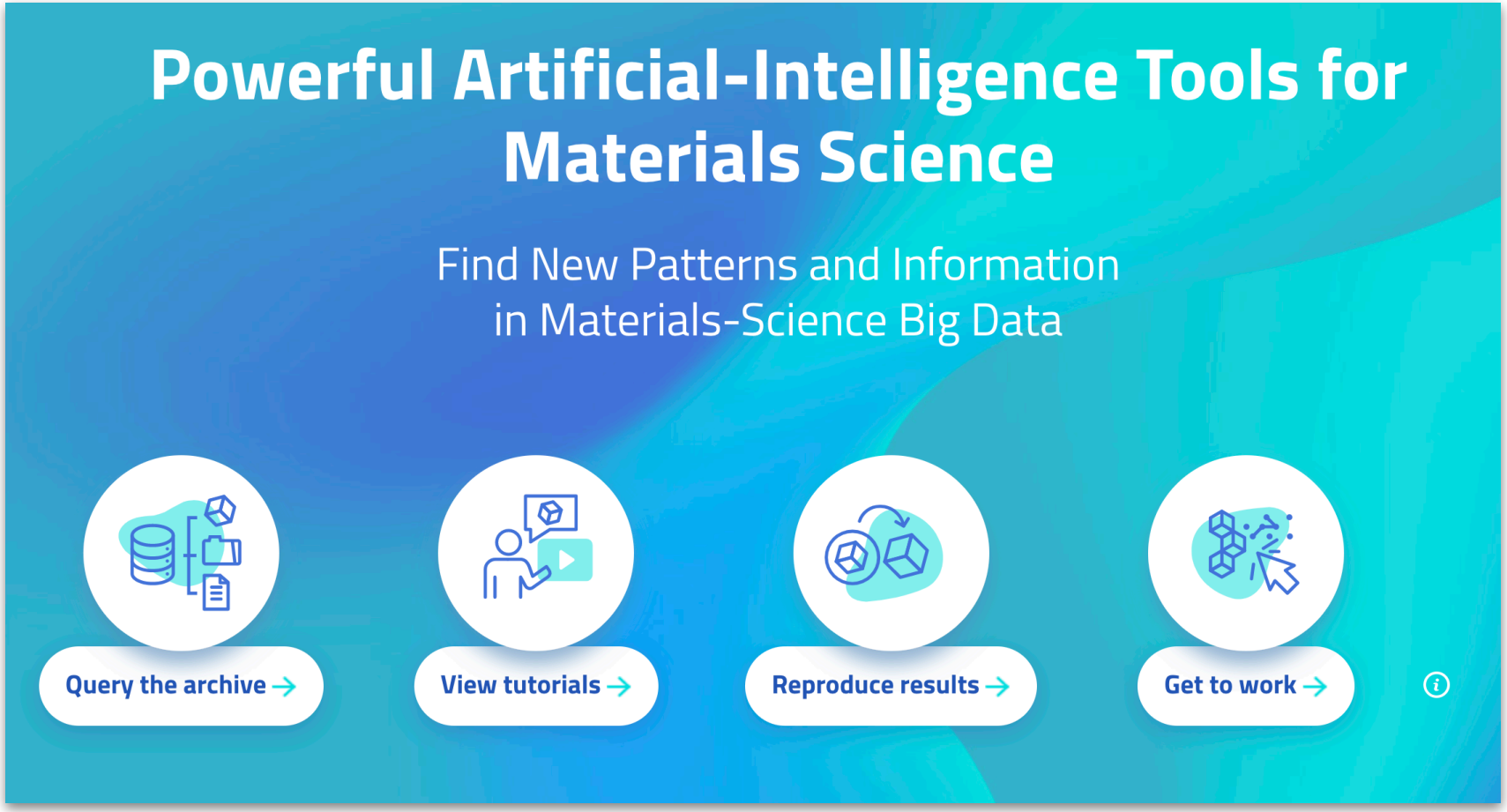
nomad-lab.eu


NOMAD
NOVEL MATERIALS DISCOVERY
About
Services
Support
Tutorials
Events
Videos
NOMAD Laboratory
enables FAIR sharing and use of materials science data
Publish
Explore
Analyze
UPLOAD
MATERIALS
CALCULATIONS
AI TOOLKIT TUTORIALS

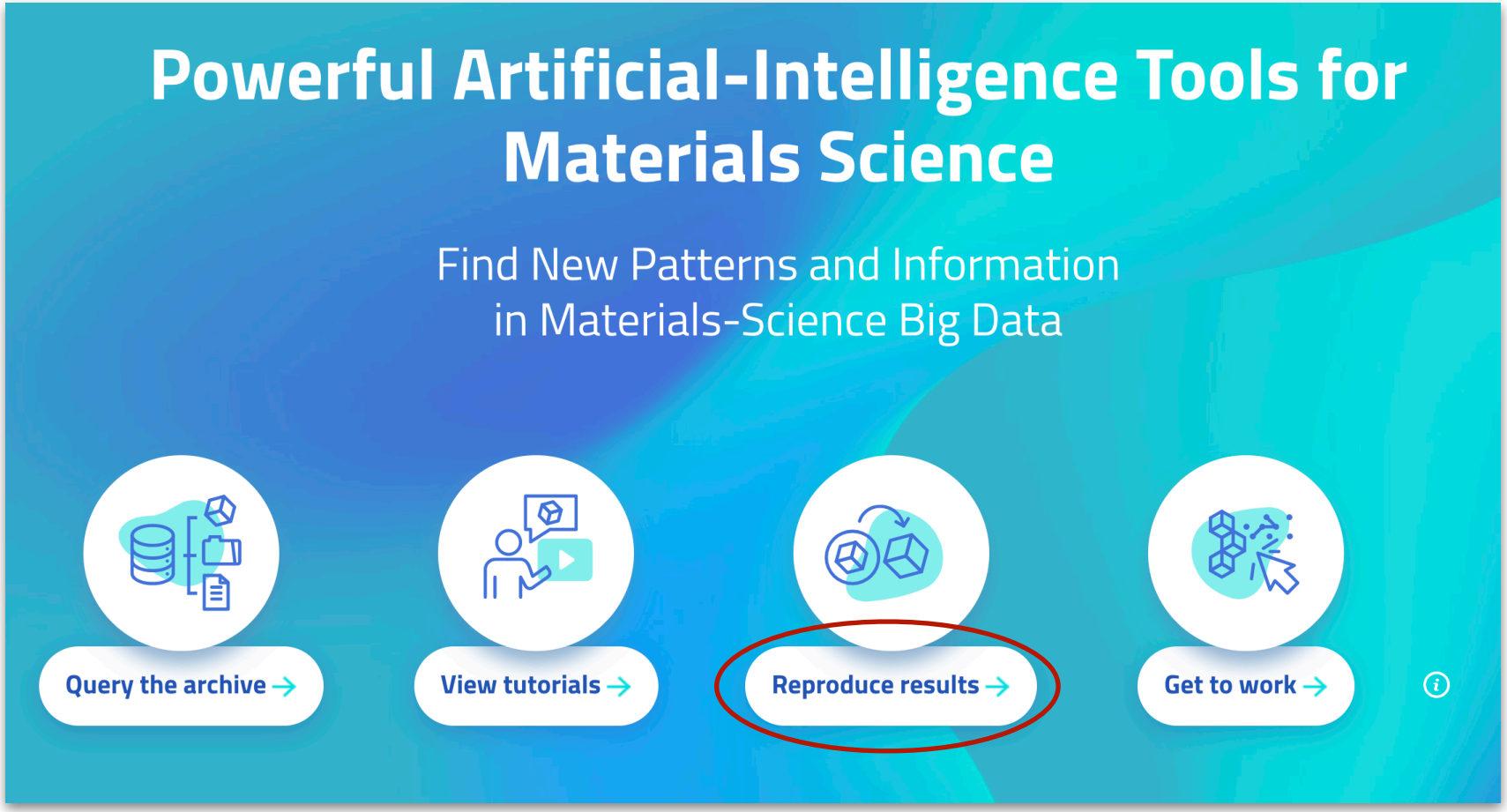
nomad-lab.eu


NOMAD
NOVEL MATERIALS DISCOVERY
About
Services
Support
Tutorials
Events
Videos
NOMAD Laboratory
enables FAIR sharing and use of materials science data
Publish
Explore
Analyze
UPLOAD
MATERIALS
CALCULATIONS
AI TOOLKIT TUTORIALS

The NOMAD Artificial-Intelligence toolkit



The NOMAD Artificial-Intelligence toolkit



Reproducing published results

Identifying domains of applicability of machine-Learning models for materials science

Authors: Mohammad-Yasin Arif | Luigi Sbailò | Luca M. Ghiringhelli

Discovery of new topological insulators in alloyed tetradymites

Authors: Luigi Sbailò | Thomas A. R. Purcell | Luca M. Ghiringhelli | Matthias Scheffler

Learn how to find descriptive parameters (short formulas) that predict whether alloyed materials are topological or trivial insulators, using the example of tetradymites. This notebook is based on the algorithm 'sure independence screening and sparsifying operator' (SISSO) that enables to search for optimal descriptor by scanning huge feature spaces.

ACCESS TUTORIAL

AI methods:
Supervised learning | Classification | Symbolic regression | Features selection | Atomic features | SISSO

System:
Tetradymites | Topological insulators

Additional Resources:
DOI | PDF

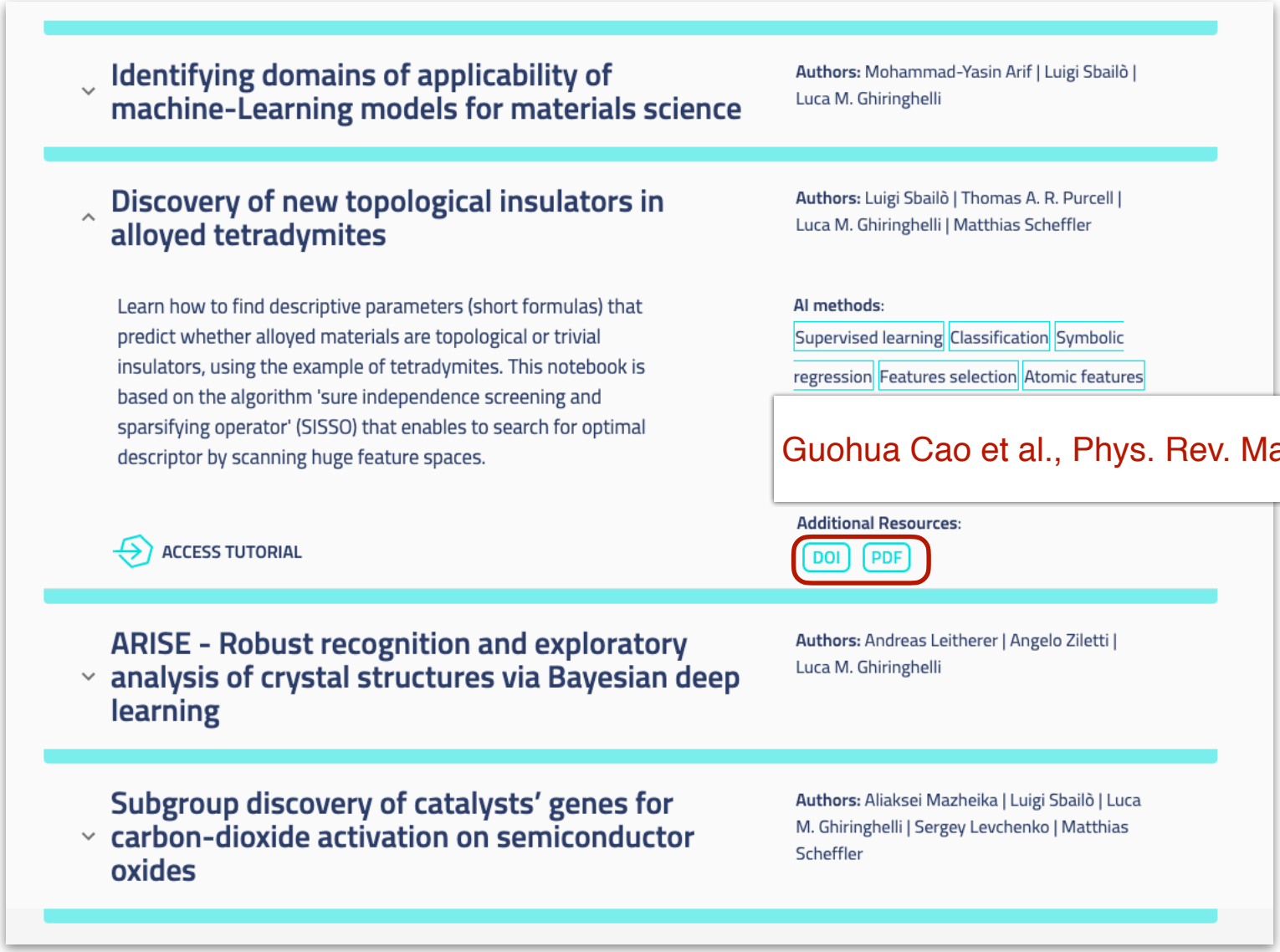
ARISE - Robust recognition and exploratory analysis of crystal structures via Bayesian deep learning

Authors: Andreas Leitherer | Angelo Ziletti | Luca M. Ghiringhelli

Subgroup discovery of catalysts' genes for carbon-dioxide activation on semiconductor oxides

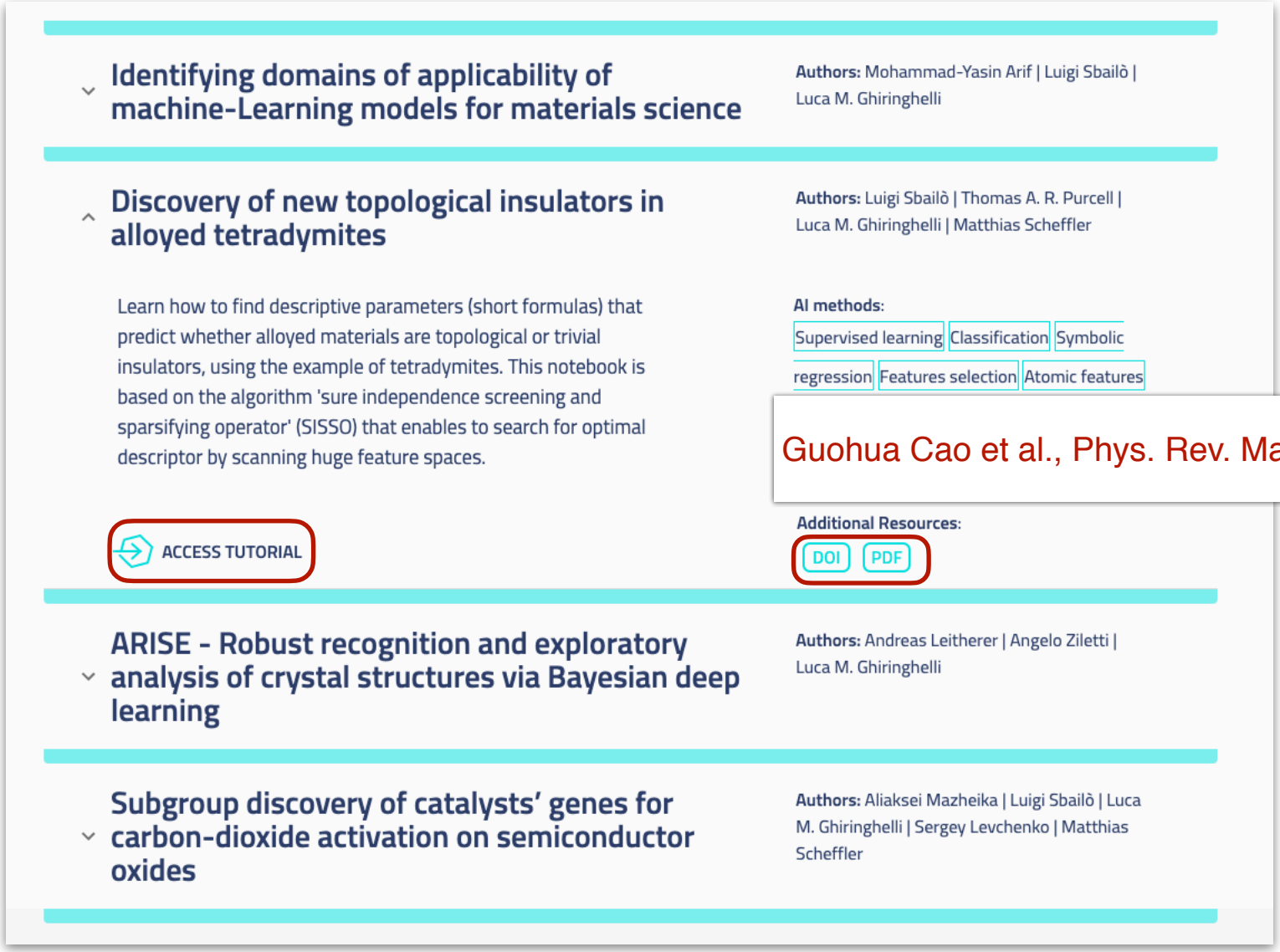
Authors: Aliaksei Mazheika | Luigi Sbailò | Luca M. Ghiringhelli | Sergey Levchenko | Matthias Scheffler

Reproducing published results



Guohua Cao et al., Phys. Rev. Materials **4**, 034204 (2020)

Reproducing published results



Guohua Cao et al., Phys. Rev. Materials **4**, 034204 (2020)

Reproducing published results

Discovery of new topological insulators in alloyed tetradymites

created by: Luigi Sbailò, Thomas A. R. Purcell, Luca M. Ghiringhelli, and Matthias Scheffler

Fritz Haber Institute of the Max Planck Society, Faradayweg 4-6, D-14195 Berlin, Germany
[Last updated: Sep 29, 2020]

MAX-PLANCK-GESELLSCHAFT

NOMAD
NOVEL MATERIALS DISCOVERY

Introduction

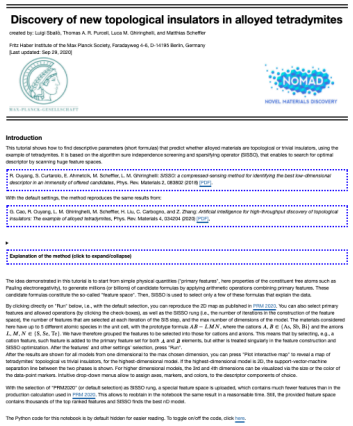
This tutorial shows how to find descriptive parameters (short formulas) that predict whether alloyed materials are topological or trivial insulators, using the example of tetradymites. It is based on the algorithm sure independence screening and sparsifying operator (SISSO), that enables to search for optimal descriptor by scanning huge feature spaces.

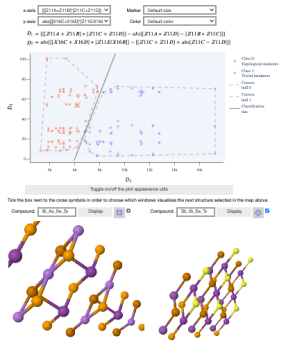
R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L. M. Ghiringhelli: SISSO: *a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates*, Phys. Rev. Materials 2, 083802 (2018) [PDF].

With the default settings, the method reproduces the same results from:

G. Cao, R. Ouyang, L. M. Ghiringhelli, M. Scheffler, H. Liu, C. Carbogno, and Z. Zhang: *Artificial intelligence for high-throughput discovery of topological insulators: The example of alloyed tetradymites*, Phys. Rev. Materials 4, 034204 (2020) [PDF],

Reproducing published results





Operations: $x+y$, $x-y$, $|x-y|$, $x\cdot y$, x/y, $\exp(x)$, $\exp(-x)$, $1/x$, x^2, x^3, $\sqrt{x}$, $\sqrt[3]{x}$, $\log(x)$, $|x|$

Features: $Z_{cations}$, $\chi_{cations}$, $\lambda_{cations}$, Z_{anions}, χ_{anions}, λ_{anions}

Settings:

SISSO rung: PRM2020

To unfreeze the feature selection, please select any rung other than PRM2020.

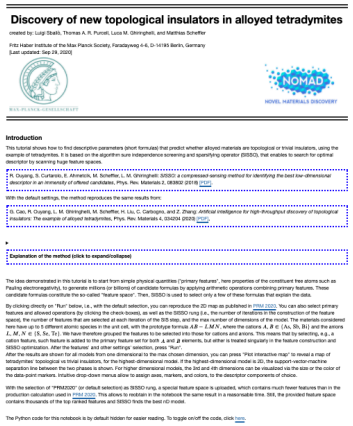
Number of selected features per SIS iteration: 50

Maximum number of dimensions: 2

Default selection | Run | Plot interactive map

```
Number of features generated: 15000

1D model
# misclassified: 25
SVM dividing line: c0 + a0*[[[Z11A*Z11B]/sqrt[X16E]]/[[Z11A/Z11C]+[Z11B/Z11D]]] = 0
c0:8.968   |   a0:-0.01492

2D model
# misclassified: 0
SVM dividing line: c0 + a0*abs[[[X16C+X16D]*[Z11E/X16B]]-[[Z11C+Z11D]+abs[Z11C-Z11D]]] + a1*[[[Z11A+Z11B]*[Z11C+Z11
D]]-abs[[Z11A*Z11D]-[Z11B*Z11C]]] = 0
c0:25.55   |   a0:0.1313   |   a1:-0.004255
```

Reproducing published results


Discovery of new topological insulators in alloyed tetradymites

created by: Luigi Sbailò, Thomas A. R. Purcell, Luca M. Ghiringhelli, and Matthias Scheffler

Fritz Haber Institute of the Max Planck Society, Faradayweg 4-6, D-14195 Berlin, Germany

[Last updated: Sep 29, 2020]

Introduction

This tutorial shows how to find descriptive parameters (short formulas) that predict whether alloyed materials are topological or trivial insulators, using the example of tetradymites. It is based on the algorithm sure independence screening and sparsifying operator (SISSO), that enables to search for optimal descriptor by scanning huge feature spaces.

P. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L. M. Ghiringhelli: SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, Phys. Rev. Materials 2, 083802 (2018) [PRD]

With the default settings, the method reproduces the same results from:

S. Cao, R. Ouyang, L. M. Ghiringhelli, M. Scheffler, H. Liu, C. Carbogno, and Z. Zhang: Artificial intelligence for high-throughput discovery of topological insulators: The example of alloyed tetradymites, Phys. Rev. Materials 4, 034204 (2020) [PCD]

Explanation of the method (click to expand/collapse)

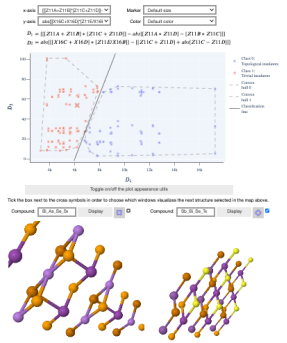
The idea demonstrated in this tutorial is to start from simple physical quantities ("primary features", here properties of the constituent free atoms such as Pauling electronegativity), to generate millions (or billions) of candidate formulas by applying arithmetic operations combining primary features. These candidate formulas constitute the so-called "feature space". Then, SISSO is used to select only a few of these formulas that explain the data.

By clicking directly on "Run" below, i.e., with the default selection, you can reproduce the 2D map as published in PRM 2020. You can also select primary features and allowed operations (by clicking the check-boxes), as well as the SISSO rung (i.e., the number of iterations in the construction of the feature space), the number of features that are selected at each iteration of the SIS step, and the max number of dimensions of the model. The materials considered here have up to 5 different atomic species in the unit cell, with the prototype formula $AB-L_2M_2X$, where the cations $A, B \in \{As, Sb, Bi\}$ and the anions $L, M, X \in \{S, Se, Te\}$. We have therefore grouped the features to be selected into those for cations and anions. This means that by selecting, e.g., a cation feature, such feature is added to the primary feature set for both A and B elements, but either is treated singularly in the feature construction and SISSO optimization. After the features' and other settings' selection, press "Run".

After the results are shown for all models from one dimensional to the max chosen dimension, you can press "Plot interactive map" to reveal a map of tetradymites' topological vs trivial insulators, for the highest-dimensional model. If the highest-dimensional model is 2D, the support-vector-machine separation line between the two phases is shown. For higher dimensional models, the 3rd and 4th dimensions can be visualized via the size or the color of the data point markers. Intuitive drop-down menus allow to assign axes, markers, and colors, to the descriptor components of choice.

With the selection of "PRM2020" (or default selection) as SISSO rung, a special feature space is uploaded, which contains much fewer features than in the production calculation used in PRM 2020. This allows to reobtain in the notebook the same result in a reasonable time. Still, the provided feature space contains thousands of the top ranked features and SISSO finds the best 1D model.

The Python code for this notebook is by default hidden for easier reading. To toggle on/off the code, click here.



Operations: Features: Settings:

- $x + y$, $x - y$, $|x - y|$, $x \cdot y$, x/y, $\exp(x)$, $\exp(-x)$, $1/x$, x^2, x^3, $\sqrt{x}$, $\sqrt[3]{x}$, $\log(x)$, $|x|$
- $Z_{cations}$, $\chi_{cations}$, $\lambda_{cations}$, Z_{anions}, χ_{anions}, λ_{anions}

SISSO rung: PRM2020

To unfreeze the feature selection, please select any rung other than PRM2020.

Number of selected features per SIS iteration: 50

Maximum number of dimensions: 2

Default selection

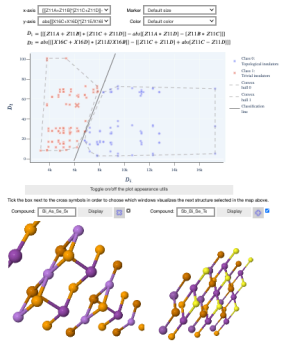
Run

Plot interactive map

Input parameters

```
Number of features generated: 15000

1D model
# misclassified: 25
SVM dividing line: c0 + a0*[[[Z11A*Z11B]/sqrt[X16E]]/[[Z11A/Z11C]+[Z11B/Z11D]]] = 0
c0:8.968   |   a0:-0.01492

2D model
# misclassified: 0
SVM dividing line: c0 + a0*abs[[[X16C+X16D]*[Z11E/X16B]]-[[Z11C+Z11D]+abs[Z11C-Z11D]]] + a1*[[[Z11A+Z11B]*[Z11C+Z11
D]]-abs[[Z11A*Z11D]-[Z11B*Z11C]]] = 0
c0:25.55   |   a0:0.1313   |   a1:-0.004255
```

Reproducing published results



Operations: Features: Settings:

Default selection

Run

Plot interactive map

$x + y$ $Z_{cations}$

$x - y$ $\chi_{cations}$ SISSO rung: PRM2020

$|x - y|$ $\lambda_{cations}$ To unfreeze the feature selection, please select any rung other than PRM2020.

$x \cdot y$ Z_{anions}

x/y χ_{anions} Number of selected features per SIS iteration: 50

$\exp(x)$ λ_{anions}

$\exp(-x)$ Maximum number of dimensions: 2

$1/x$

x^2

x^3

$\sqrt{x}$

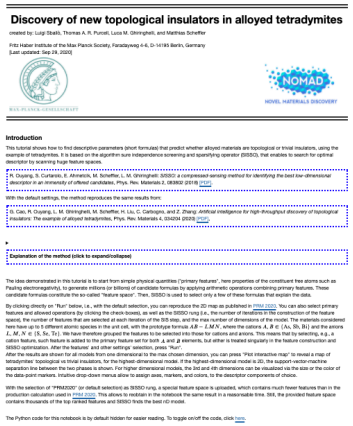
$\sqrt[3]{x}$

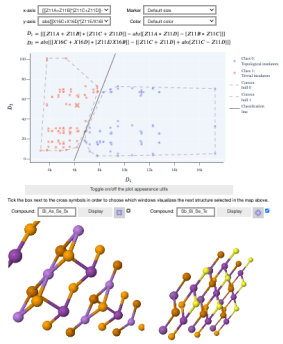
$\log(x)$

$|x|$

```
Number of features generated: 15000

1D model
# misclassified: 25
SVM dividing line: c0 + a0*[[[Z11A*Z11B]/sqrt[X16E]]/[[Z11A/Z11C]+[Z11B/Z11D]]] = 0
c0:8.968   |   a0:-0.01492

2D model
# misclassified: 0
SVM dividing line: c0 + a0*abs[[[X16C+X16D]*[Z11E/X16B]]-[[Z11C+Z11D]+abs[Z11C-Z11D]]] + a1*[[[Z11A+Z11B]*[Z11C+Z11
D]]-abs[[Z11A*Z11D]-[Z11B*Z11C]]] = 0
c0:25.55   |   a0:0.1313   |   a1:-0.004255
```

Reproducing published results





Operations: $x+y$, $x-y$, $|x-y|$, $x\cdot y$, x/y, $\exp(x)$, $\exp(-x)$, $1/x$, x^2, x^3, $\sqrt{x}$, $\sqrt[3]{x}$, $\log(x)$, $|x|$

Features: $Z_{cations}$, $\chi_{cations}$, $\lambda_{cations}$, Z_{anions}, χ_{anions}, λ_{anions}

Settings:

SISSO rung: PRM2020

To unfreeze the feature selection, please select any rung other than PRM2020.

Number of selected features per SIS iteration: 50

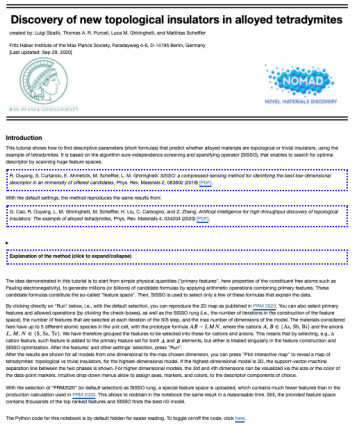
Maximum number of dimensions: 2

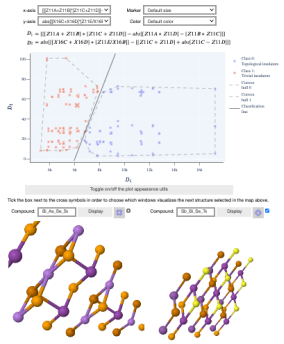
Default selection | Run | Plot interactive map

Output

```
Number of features generated: 15000

1D model
# misclassified: 25
SVM dividing line: c0 + a0*[[[Z11A*Z11B]/sqrt[X16E]]/[[Z11A/Z11C]+[Z11B/Z11D]]] = 0
c0:8.968   |   a0:-0.01492

2D model
# misclassified: 0
SVM dividing line: c0 + a0*abs[[[X16C+X16D]*[Z11E/X16B]]-[[Z11C+Z11D]+abs[Z11C-Z11D]]] + a1*[[[Z11A+Z11B]*[Z11C+Z11
D]]-abs[[Z11A*Z11D]-[Z11B*Z11C]]] = 0
c0:25.55   |   a0:0.1313   |   a1:-0.004255
```

Reproducing published results





Operations: Features: Settings:

Default selection

Run

Plot interactive map

SISSO rung: PRM2020

To unfreeze the feature selection, please select any rung other than PRM2020.

Number of selected features per SIS iteration: 50

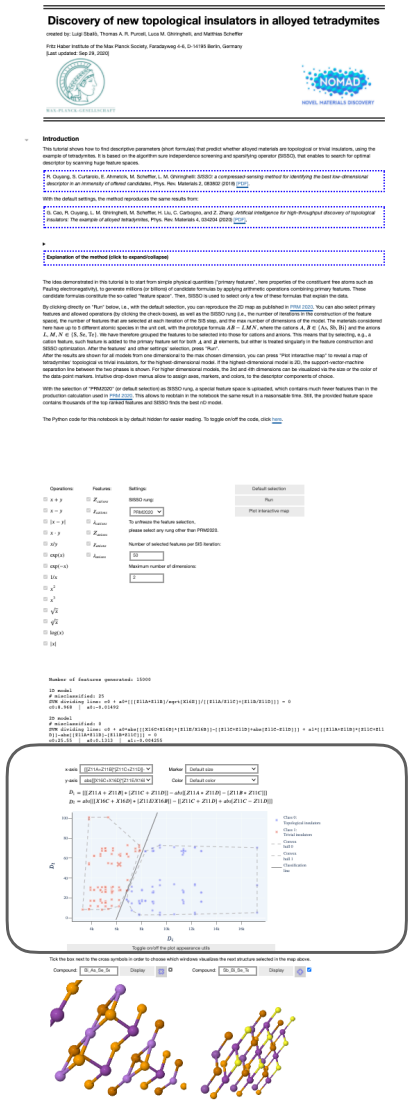
Maximum number of dimensions: 2

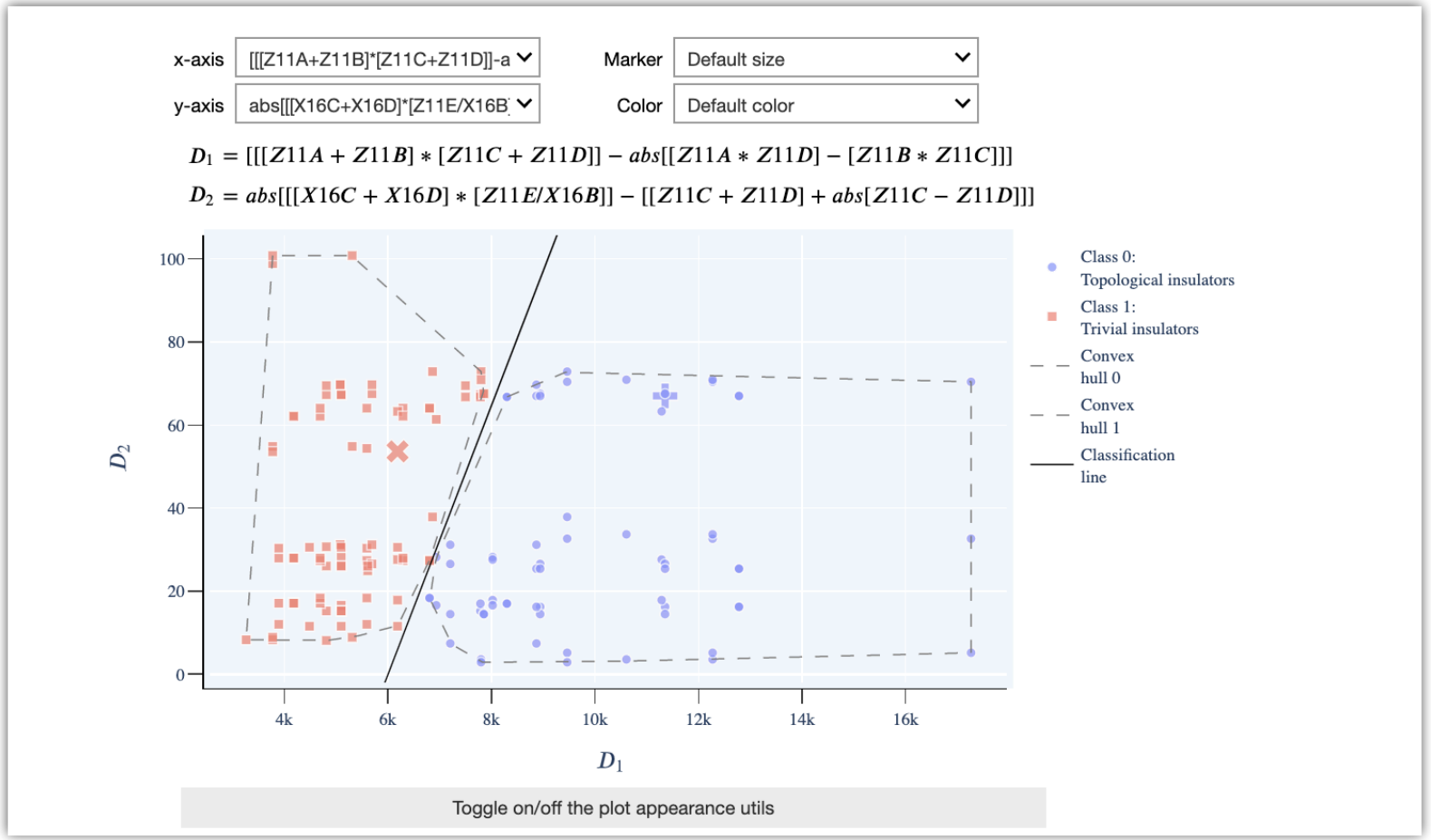
$x + y$ $x - y$ $|x - y|$ $x \cdot y$ x/y $\exp(x)$ $\exp(-x)$ $1/x$ x^2 x^3 $\sqrt{x}$ $\sqrt[3]{x}$ $\log(x)$ $|x|$

$Z_{cations}$ $\chi_{cations}$ $\lambda_{cations}$ Z_{anions} χ_{anions} λ_{anions}

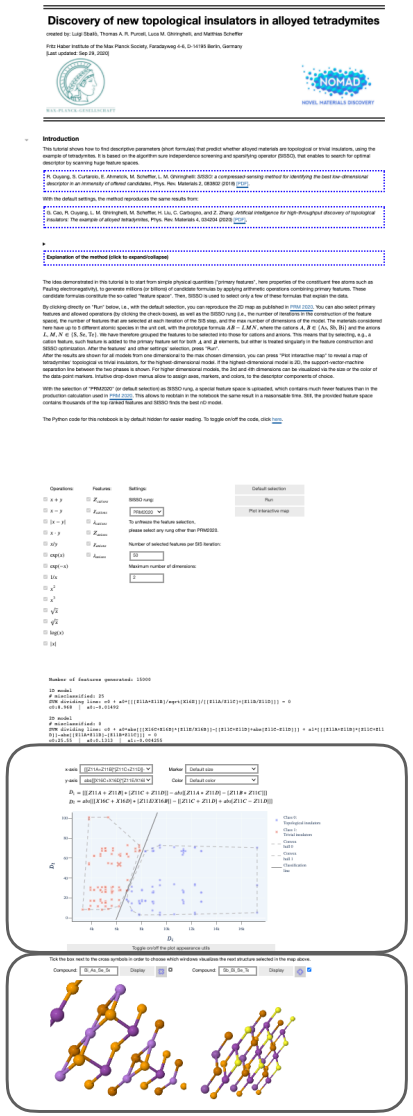
```
Number of features generated: 15000

1D model
# misclassified: 25
SVM dividing line: c0 + a0*[[[Z11A*Z11B]/sqrt[X16E]]/[[Z11A/Z11C]+[Z11B/Z11D]]] = 0
c0:8.968   |   a0:-0.01492

2D model
# misclassified: 0
SVM dividing line: c0 + a0*abs[[[X16C+X16D]*[Z11E/X16B]]-[[Z11C+Z11D]+abs[Z11C-Z11D]]] + a1*[[[Z11A+Z11B]*[Z11C+Z11
D]]-abs[[Z11A*Z11D]-[Z11B*Z11C]]] = 0
c0:25.55   |   a0:0.1313   |   a1:-0.004255
```

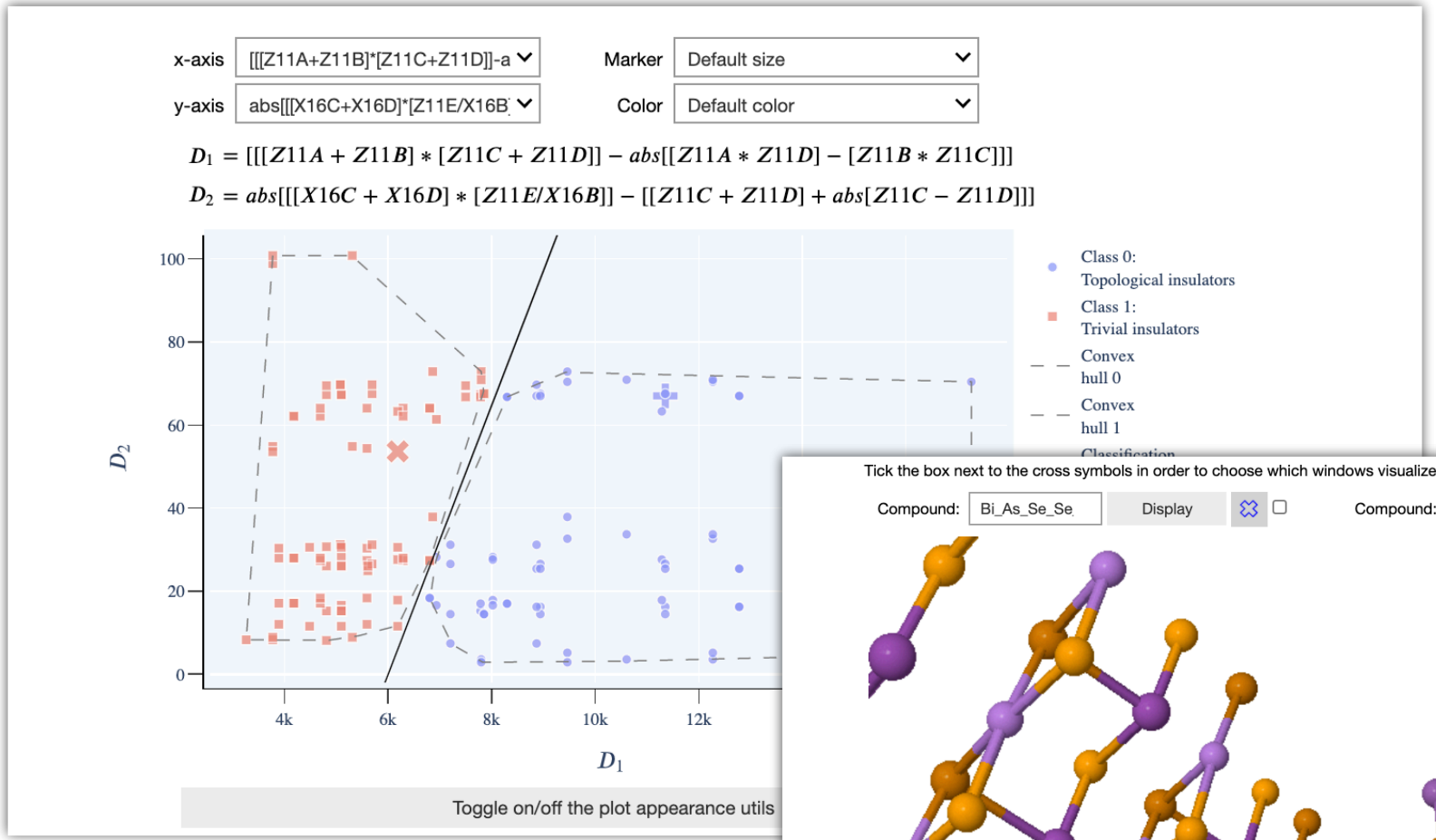
Reproducing published results

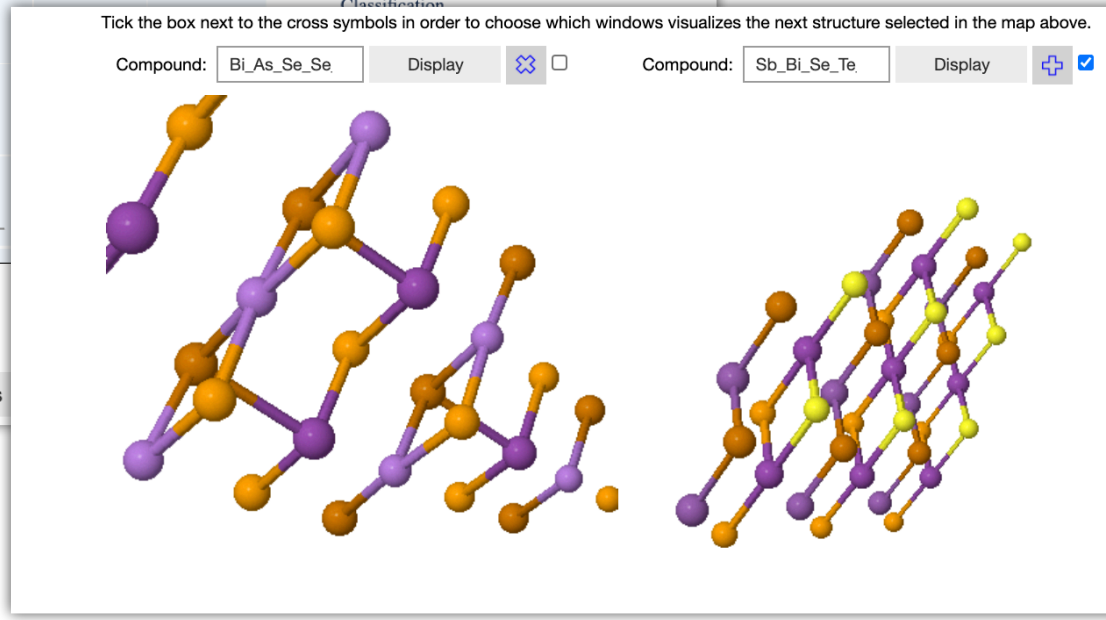




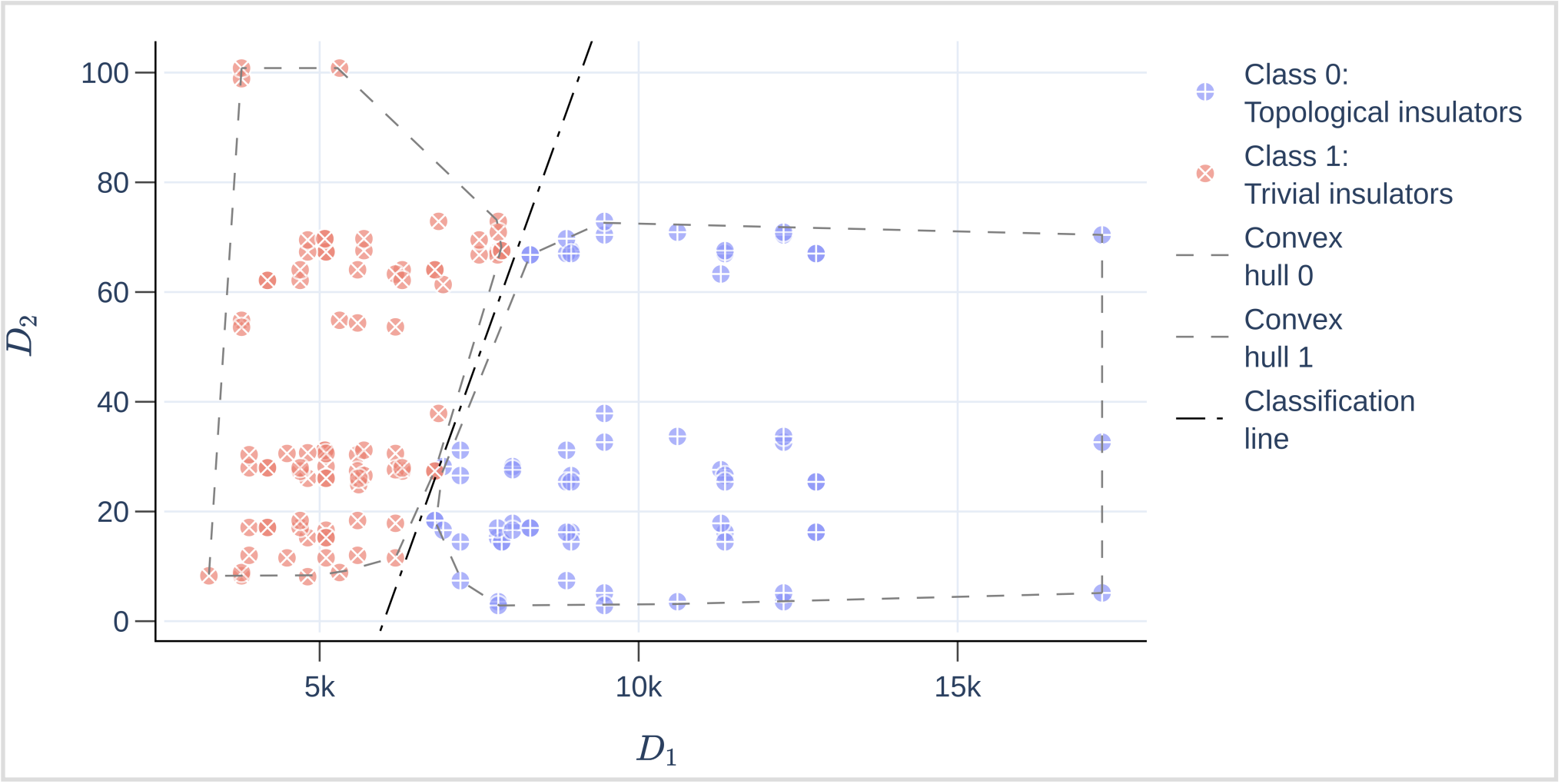
Reproducing published results



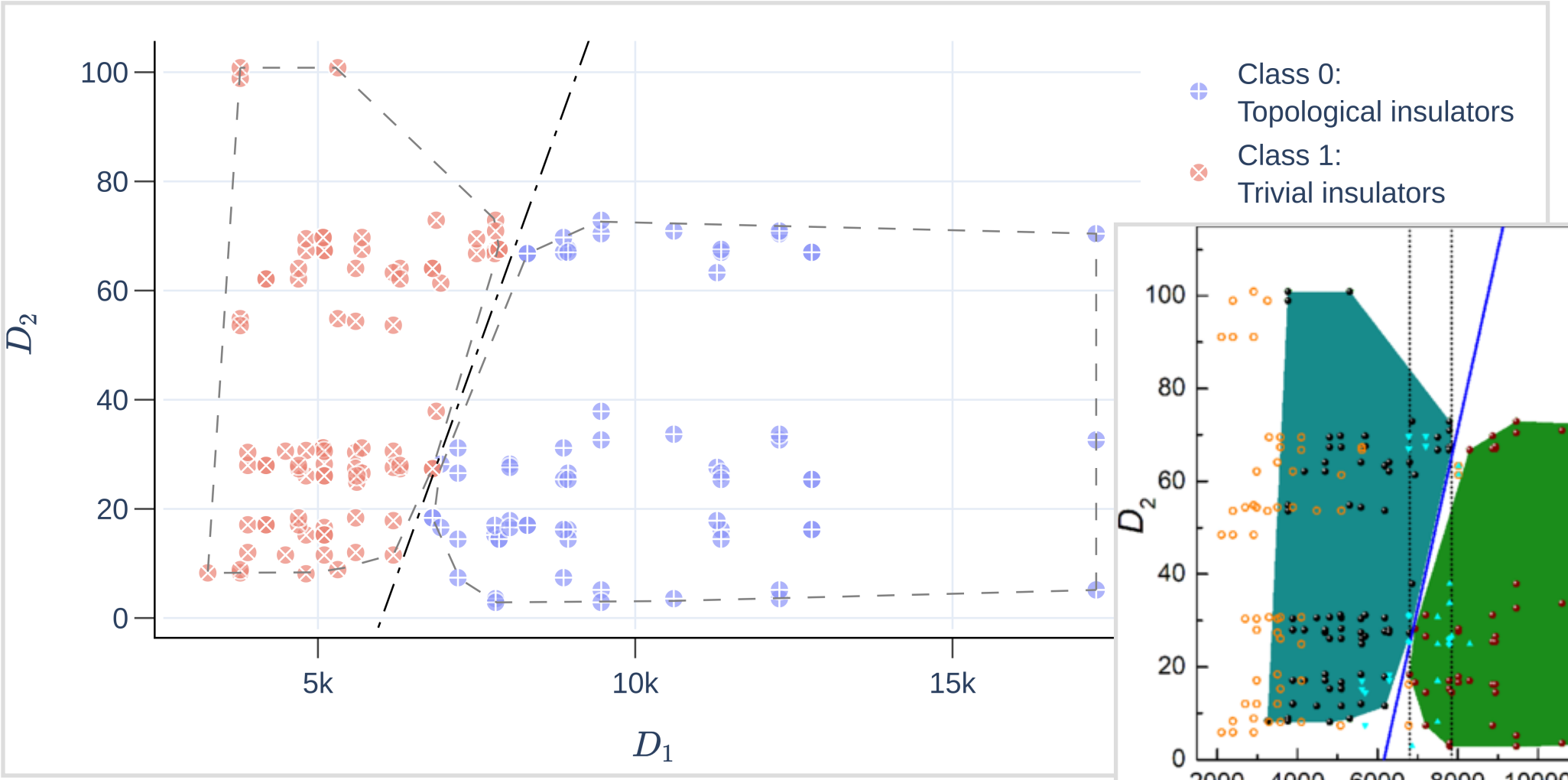




Reproducing published results

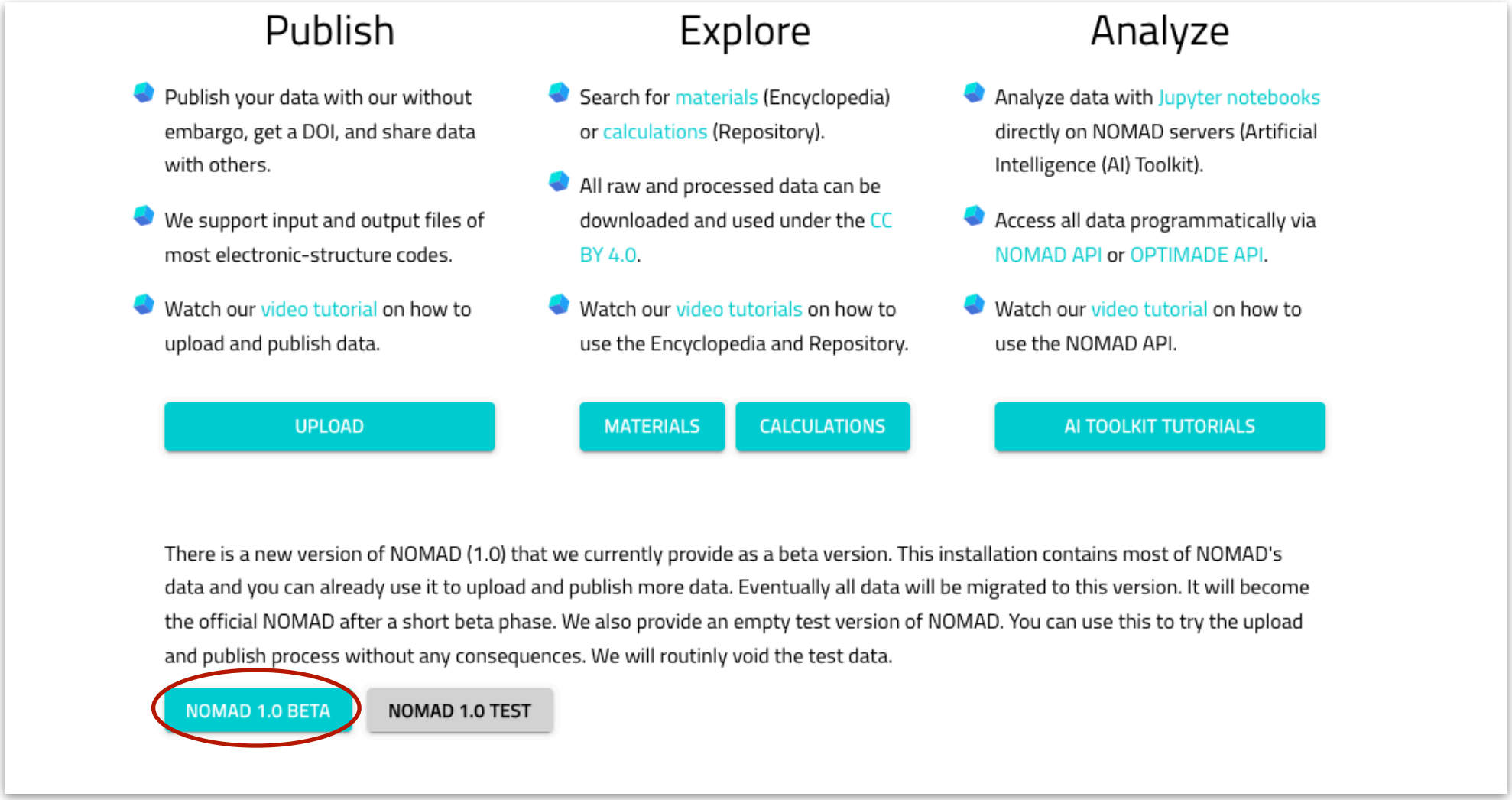


Reproducing published results

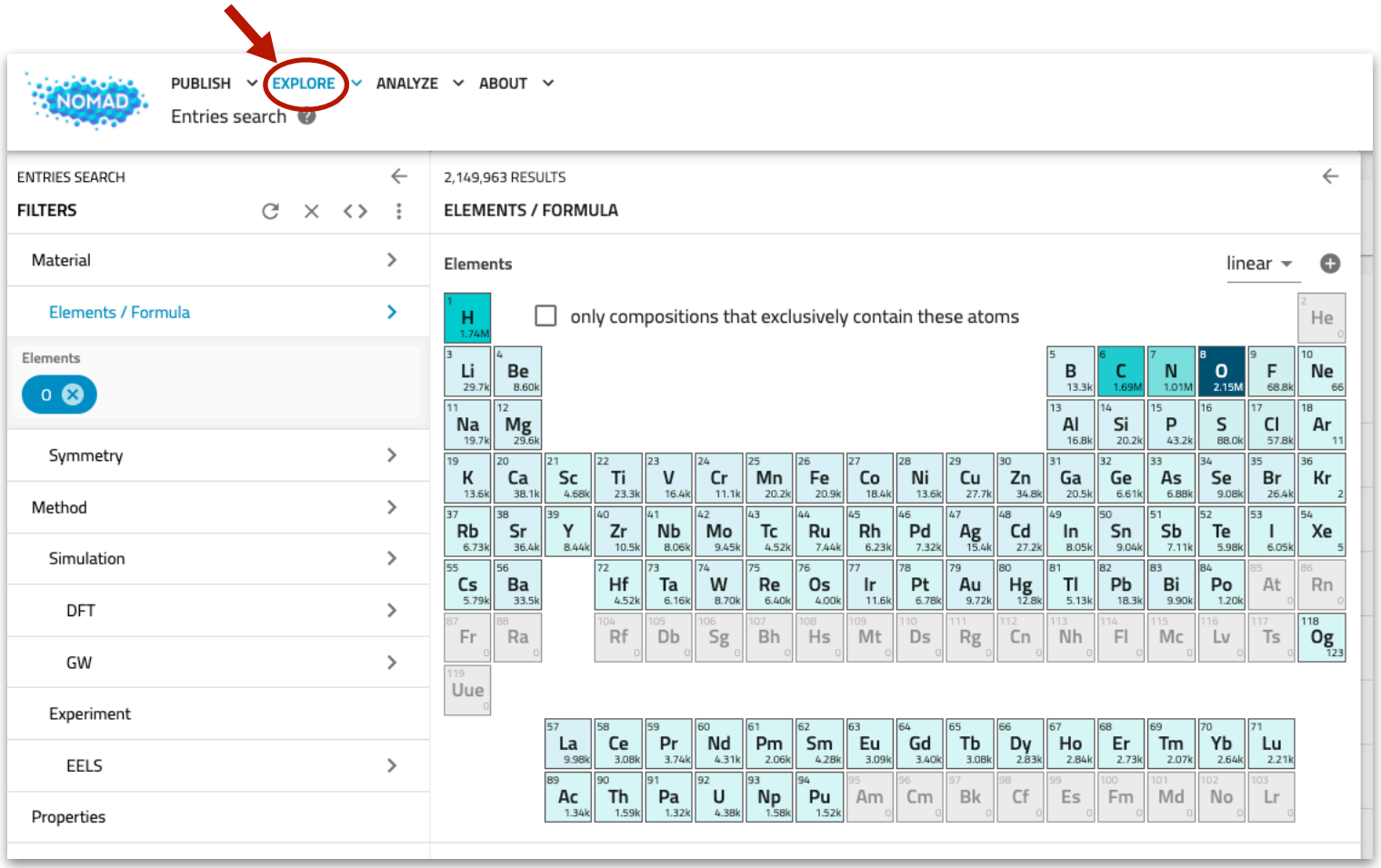


Guohua Cao et al., Phys. Rev. Materials **4**, 034204

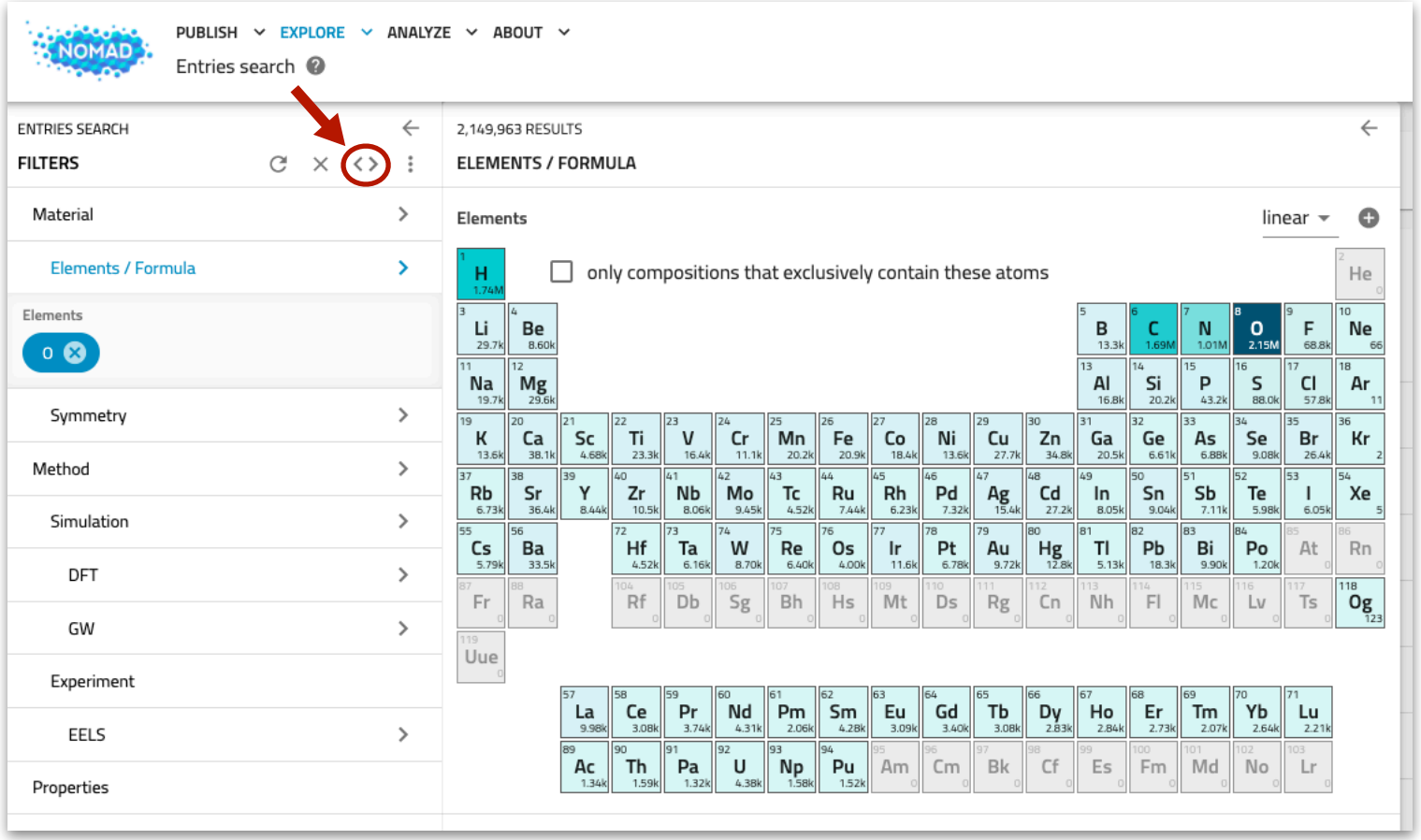
nomad-lab.eu


Publish
Publish your data with our without embargo, get a DOI, and share data with others.
We support input and output files of most electronic-structure codes.
Watch our video tutorial on how to upload and publish data.
UPLOAD
Explore
Search for materials (Encyclopedia) or calculations (Repository).
All raw and processed data can be downloaded and used under the CC BY 4.0.
Watch our video tutorials on how to use the Encyclopedia and Repository.
MATERIALS
CALCULATIONS
Analyze
Analyze data with Jupyter notebooks directly on NOMAD servers (Artificial Intelligence (AI) Toolkit).
Access all data programmatically via NOMAD API or OPTIMADE API.
Watch our video tutorial on how to use the NOMAD API.
AI TOOLKIT TUTORIALS
There is a new version of NOMAD (1.0) that we currently provide as a beta version. This installation contains most of NOMAD's data and you can already use it to upload and publish more data. Eventually all data will be migrated to this version. It will become the official NOMAD after a short beta phase. We also provide an empty test version of NOMAD. You can use this to try the upload and publish process without any consequences. We will routinly void the test data.
NOMAD 1.0 BETA
NOMAD 1.0 TEST

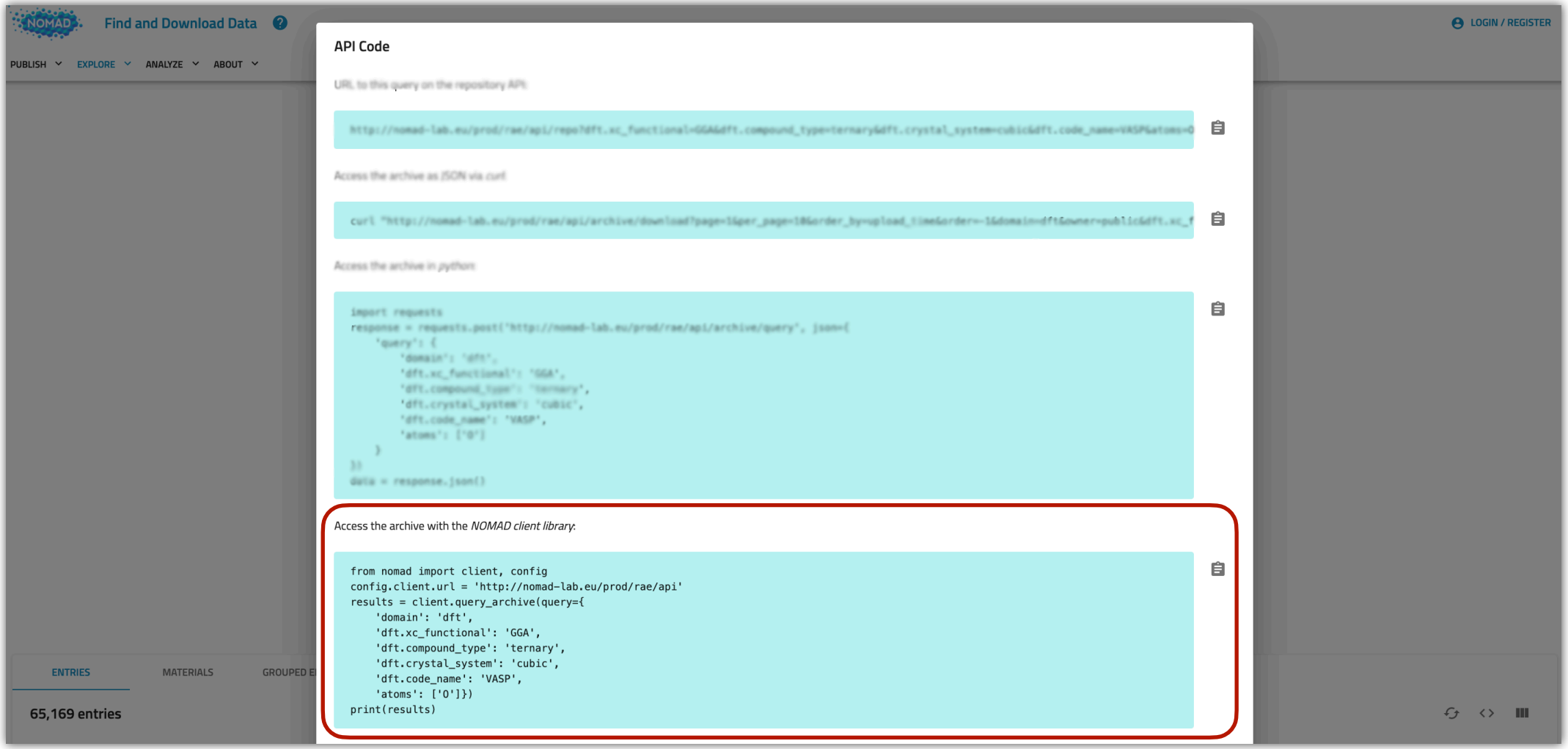
Querying the NOMAD Archive



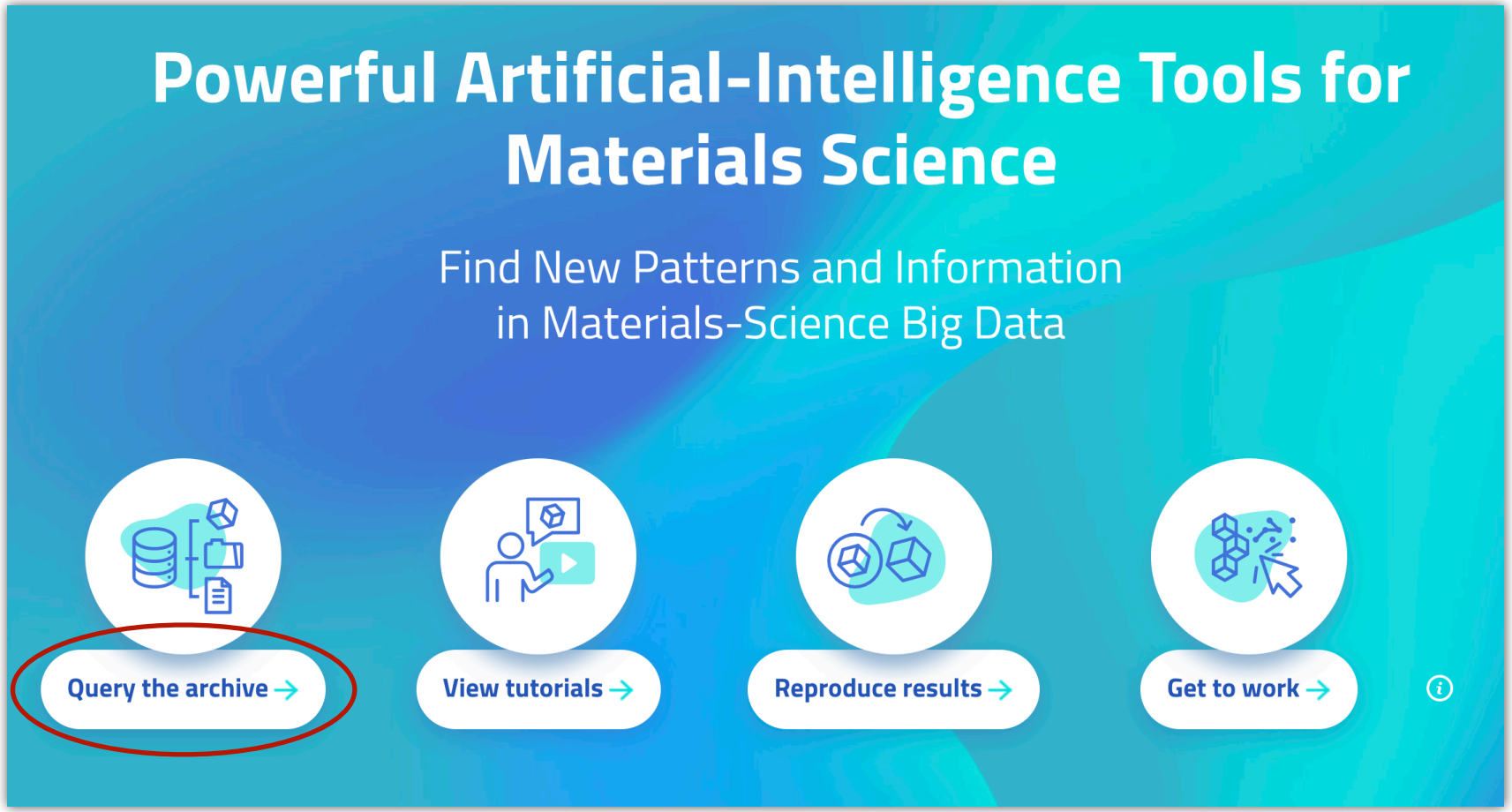
Querying the NOMAD Archive



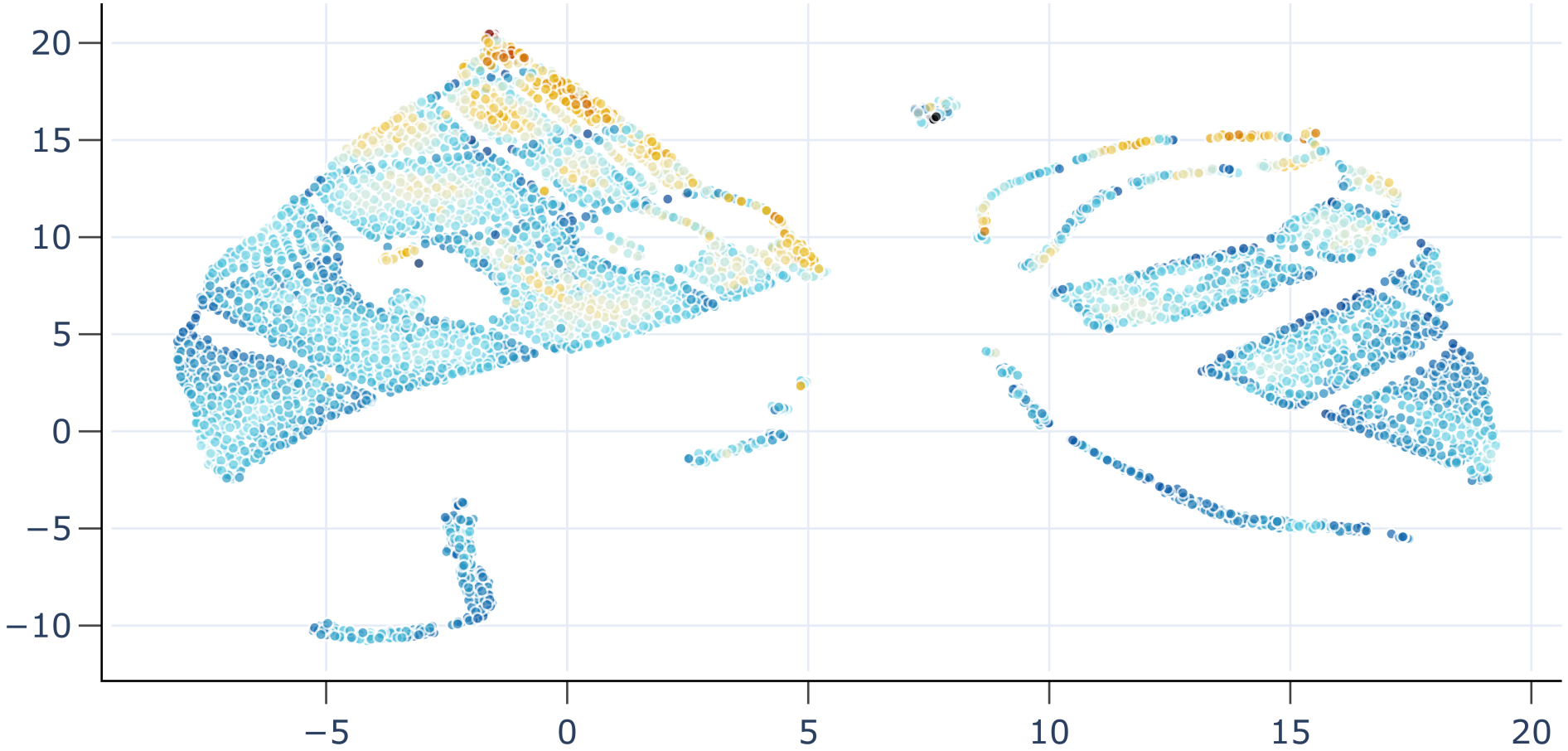
Querying the NOMAD Archive



Querying the NOMAD Archive

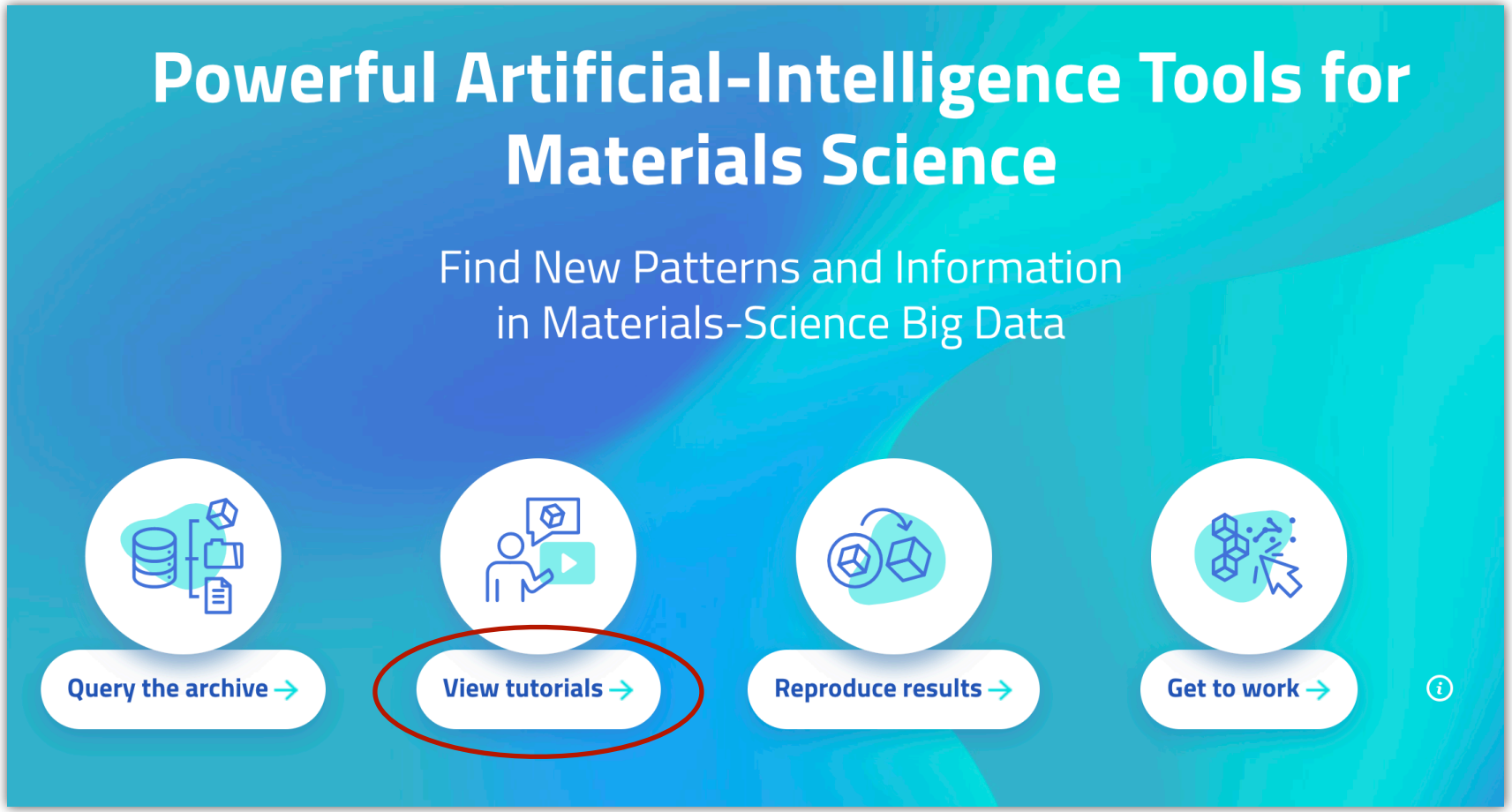


Querying the NOMAD Archive

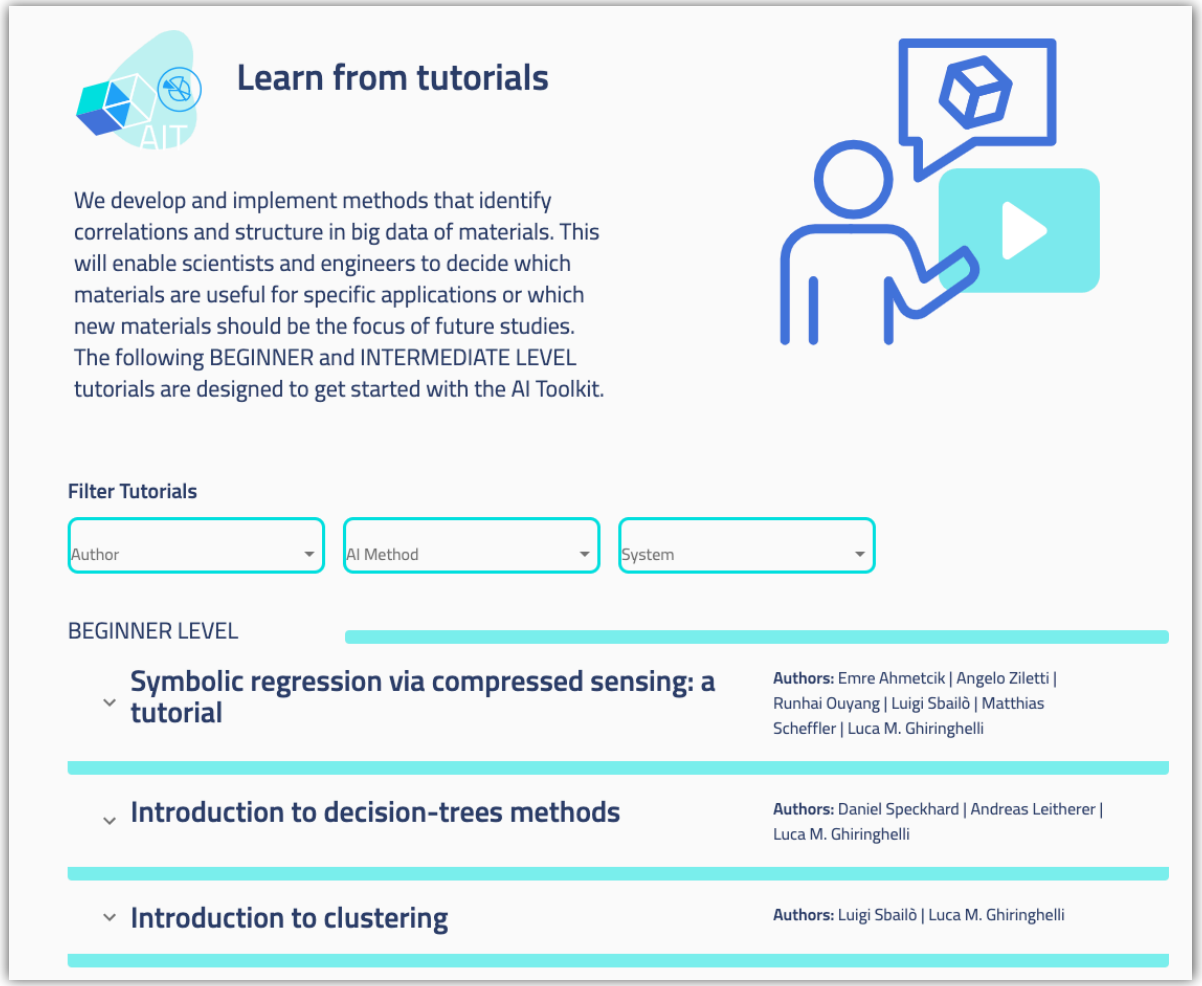


2-dimensional embedding, found with umap. Different colors means different atomic concentration values.

Querying the NOMAD Archive

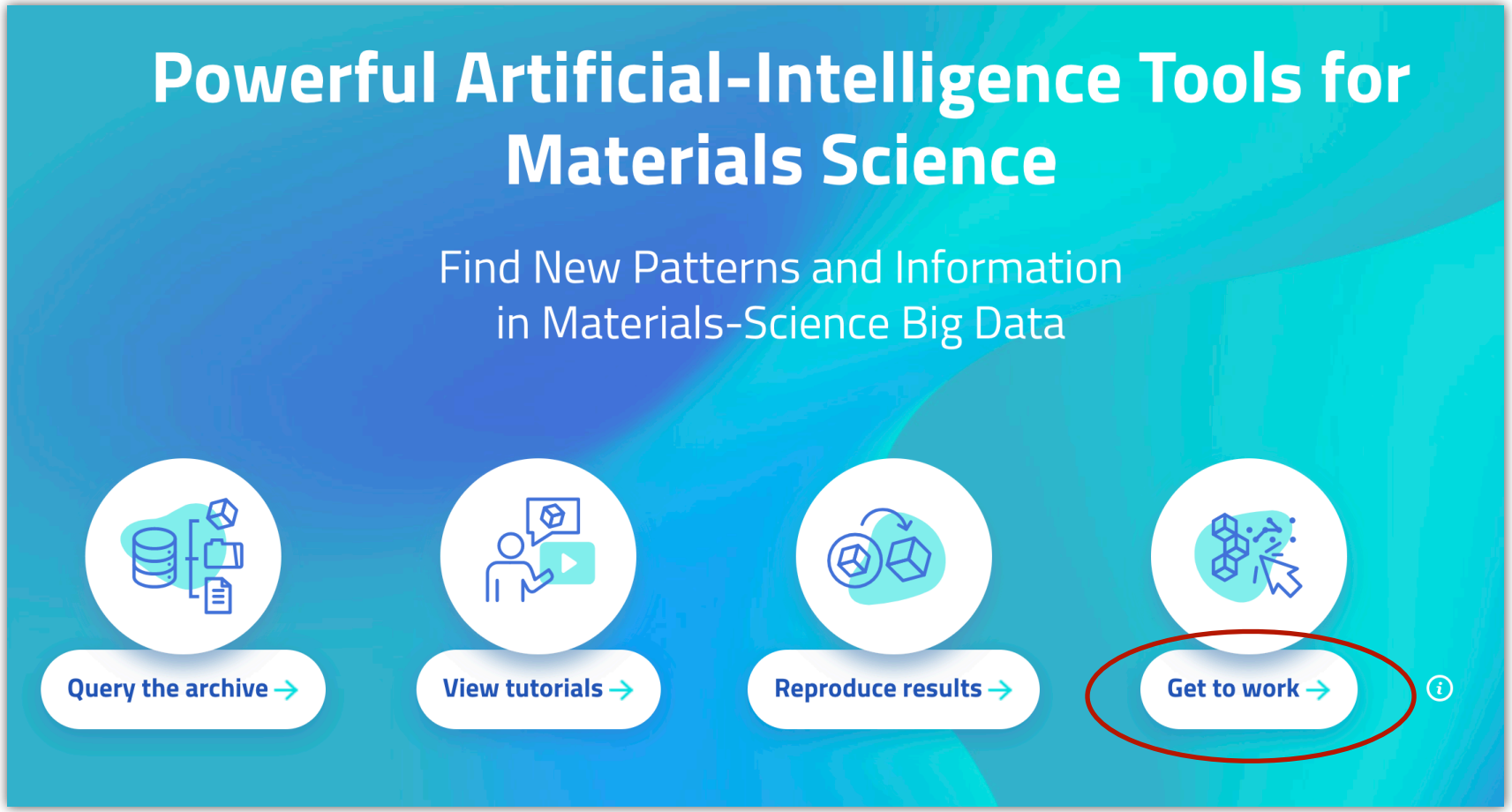


AI tutorials for materials science



- Tutorials span the most popular artificial-intelligence methods, such as clustering, dimension reduction, symbolic regression, decision trees, random forest, kernel ridge regression and deep neural networks.
- Applications to materials science are shown in a preparatory manner.
- Hands on exercises are presented throughout tutorials.
- Introductory videos, prepared in occasion of workshops and lectures held at the NOMAD Laboratory, are available.

Querying the NOMAD Archive



Contribute

The *NOMAD Archive* can be queried from the *Artificial-Intelligence Toolkit*, and the data retrieved can be analysed on the fly with the desired AI tools.

The *Artificial-Intelligence Toolkit* hosts tutorials for hands-on learning of the most popular AI tools.

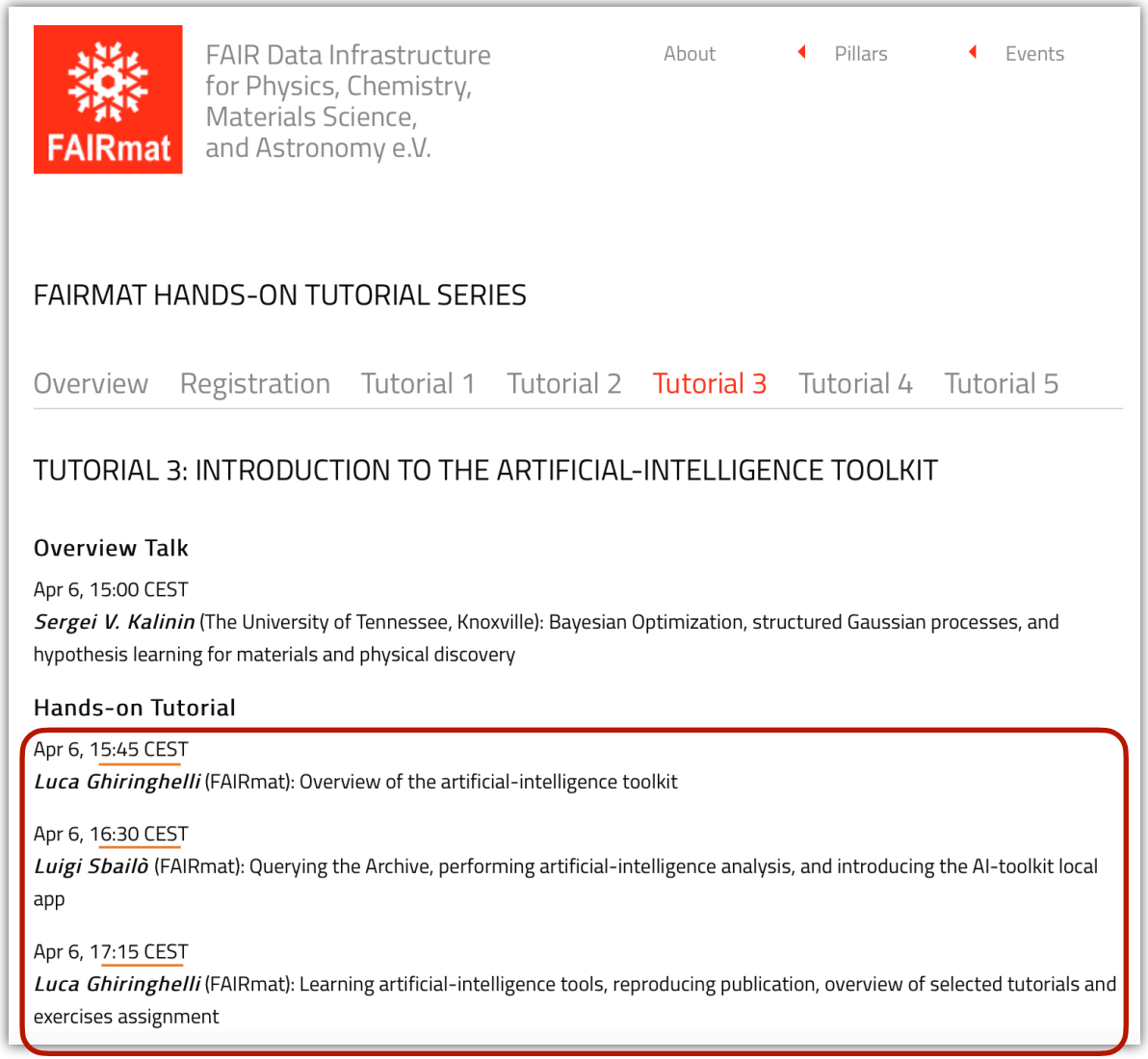
In the *NOMAD Artificial-Intelligence Toolkit*, we maintain several notebooks demonstrating the latest applications of AI to materials science data.

Contact us!

ghiringhelli@fhi.mpg.de scheffler@fhi.mpg.de

FAIRMAT HANDS-ON TUTORIAL SERIES

www.fair-di.eu/fairmat-tutorials-3



FAIR Data Infrastructure for Physics, Chemistry, Materials Science, and Astronomy e.V.

About Pillars Events

FAIRMAT HANDS-ON TUTORIAL SERIES

Overview Registration Tutorial 1 Tutorial 2 Tutorial 3 Tutorial 4 Tutorial 5

TUTORIAL 3: INTRODUCTION TO THE ARTIFICIAL-INTELLIGENCE TOOLKIT

Overview Talk

Apr 6, 15:00 CEST

Sergei V. Kalinin (The University of Tennessee, Knoxville): Bayesian Optimization, structured Gaussian processes, and hypothesis learning for materials and physical discovery

Hands-on Tutorial

Apr 6, 15:45 CEST

Luca Ghiringhelli (FAIRmat): Overview of the artificial-intelligence toolkit

Apr 6, 16:30 CEST

Luigi Sbailò (FAIRmat): Querying the Archive, performing artificial-intelligence analysis, and introducing the AI-toolkit local app

Apr 6, 17:15 CEST

Luca Ghiringhelli (FAIRmat): Learning artificial-intelligence tools, reproducing publication, overview of selected tutorials and exercises assignment

Acknowledgments

Luca Ghiringhelli

Matthias Scheffler

Claudia Draxl, Christian Carbogno, Markus Scheidgen, Markus Rampp, Gian-Marco Rignanese, James Kermode, Gábor Csányi, Ádám Fekete, Thomas A. R. Purcell, Mohammad-Yasin Arif, Lauri Himanen, Alvin Noe Lavines, Aakash Ashok Naik, Emre Ahmetcik, Andreas Leitherer, Daniel Speckhard, Marcel F. Langer, Benjamin Regler, Fawzi Mohamed, Angelo Ziletti