HoW exciting 2023

Data analysis and quality assessment by fingerprinting

MARTIN KUBAN

HU BERLIN & IRIS ADLERSHOF

06.08.2023

Introduction

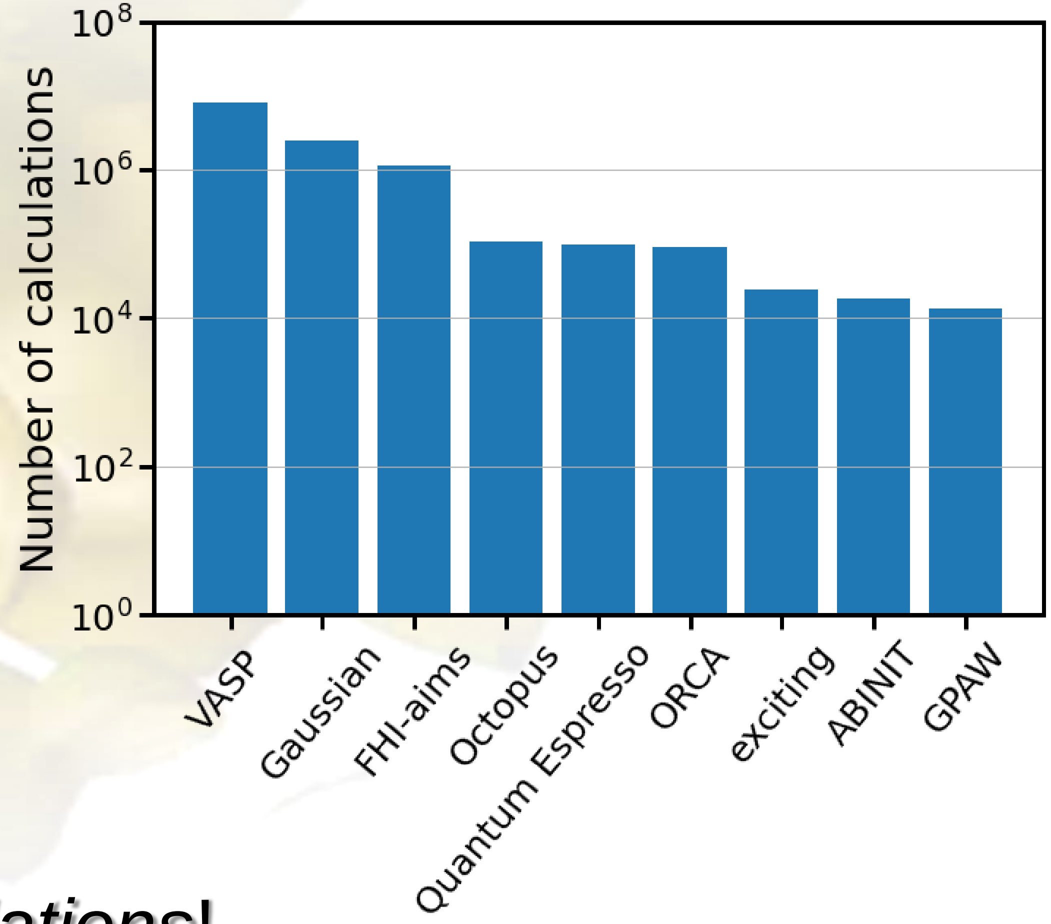
NOMAD Repository:

> 100 million calculations

> 40 codes

Interoperability?

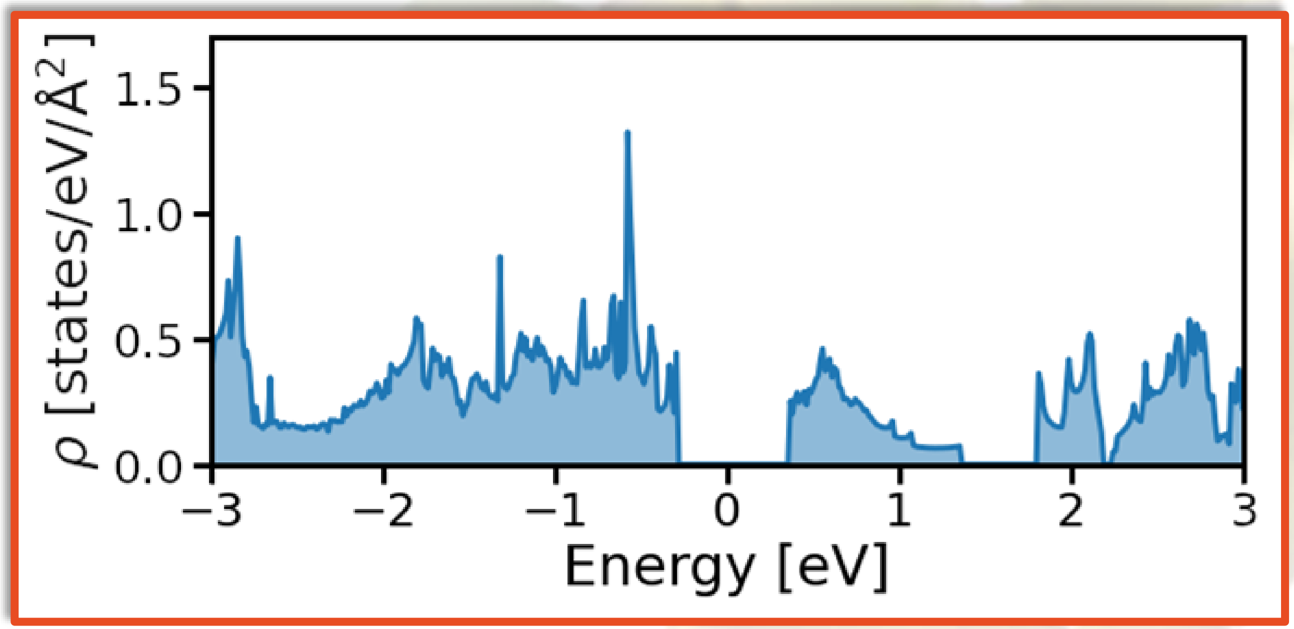
Findability?



09.10.2021

Analyze data using *similarity relations*!

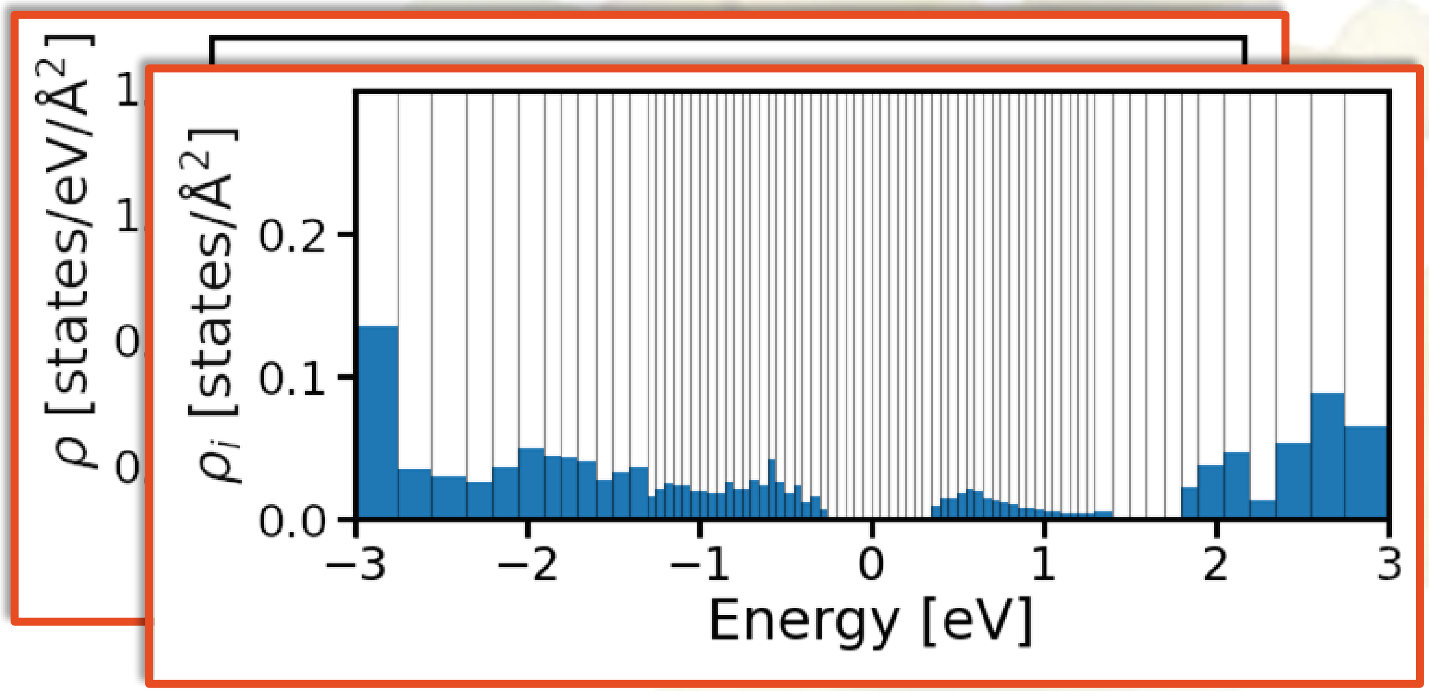
Spectrum fingerprints



M. Kuban *et al. Sci Data* **9**, 646 (2022).

Inspired by O. Isayev, et al., Chem. Mater. 27, 735 (2015)

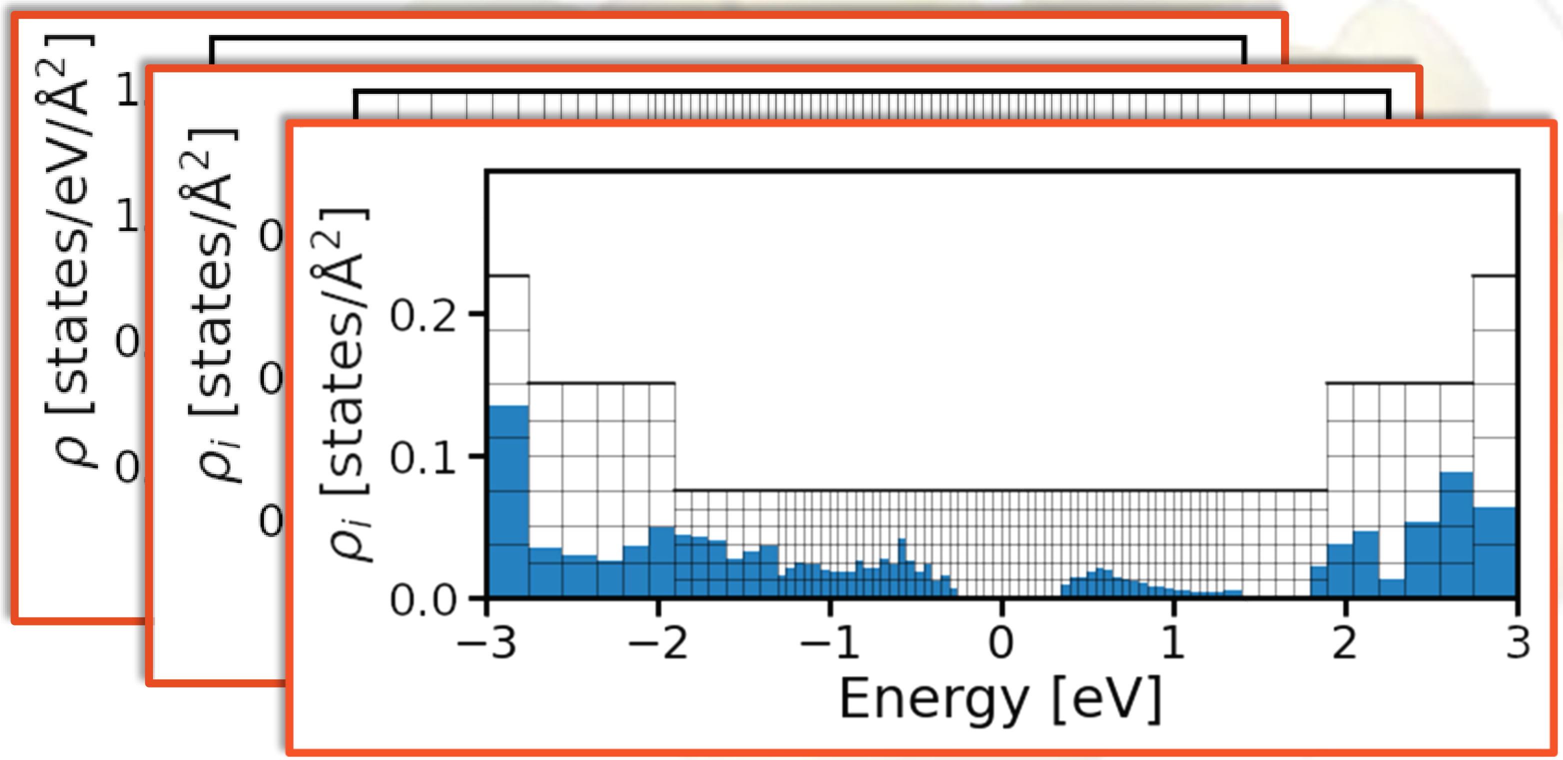
Spectrum fingerprints



M. Kuban *et al. Sci Data* **9**, 646 (2022).

Inspired by O. Isayev, et al., Chem. Mater. 27, 735 (2015)

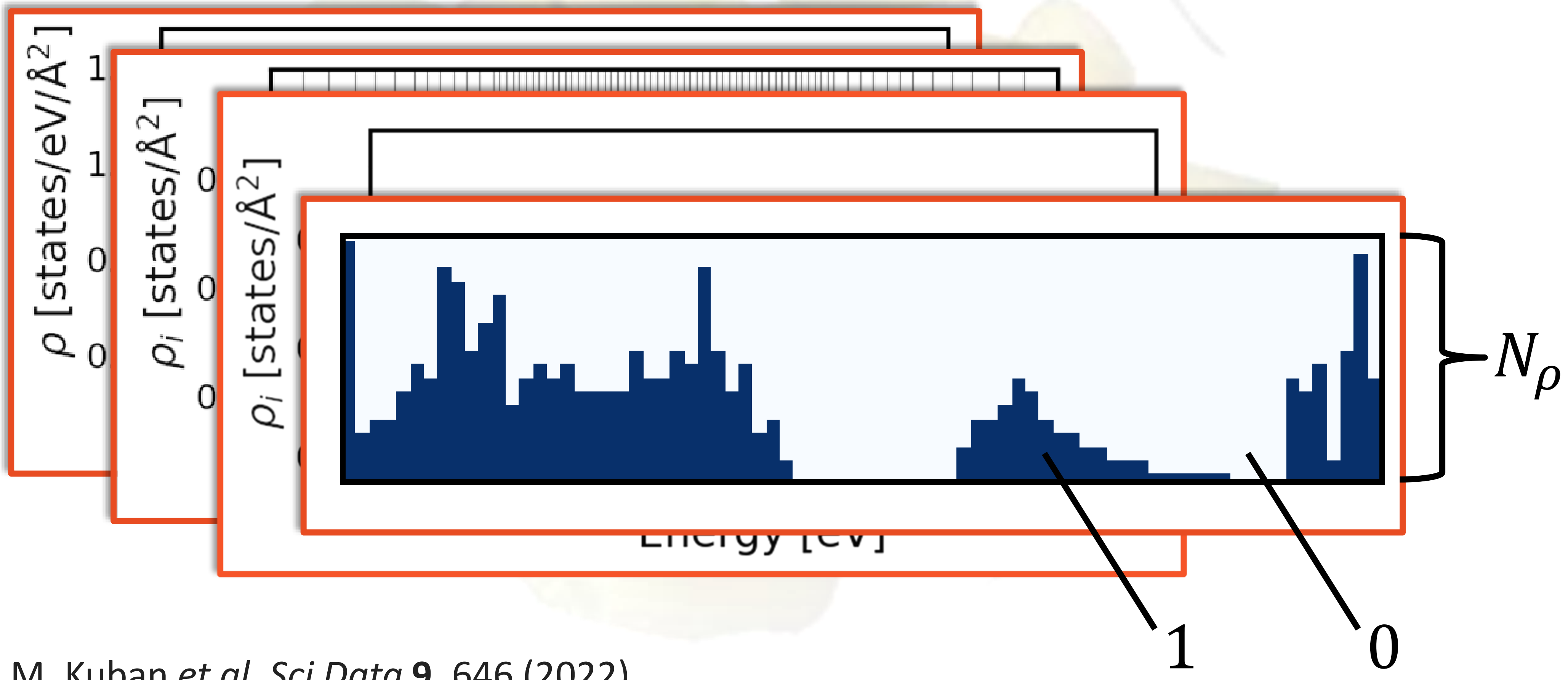
Spectrum fingerprints



M. Kuban *et al. Sci Data* **9**, 646 (2022).

Inspired by O. Isayev, et al., Chem. Mater. 27, 735 (2015)

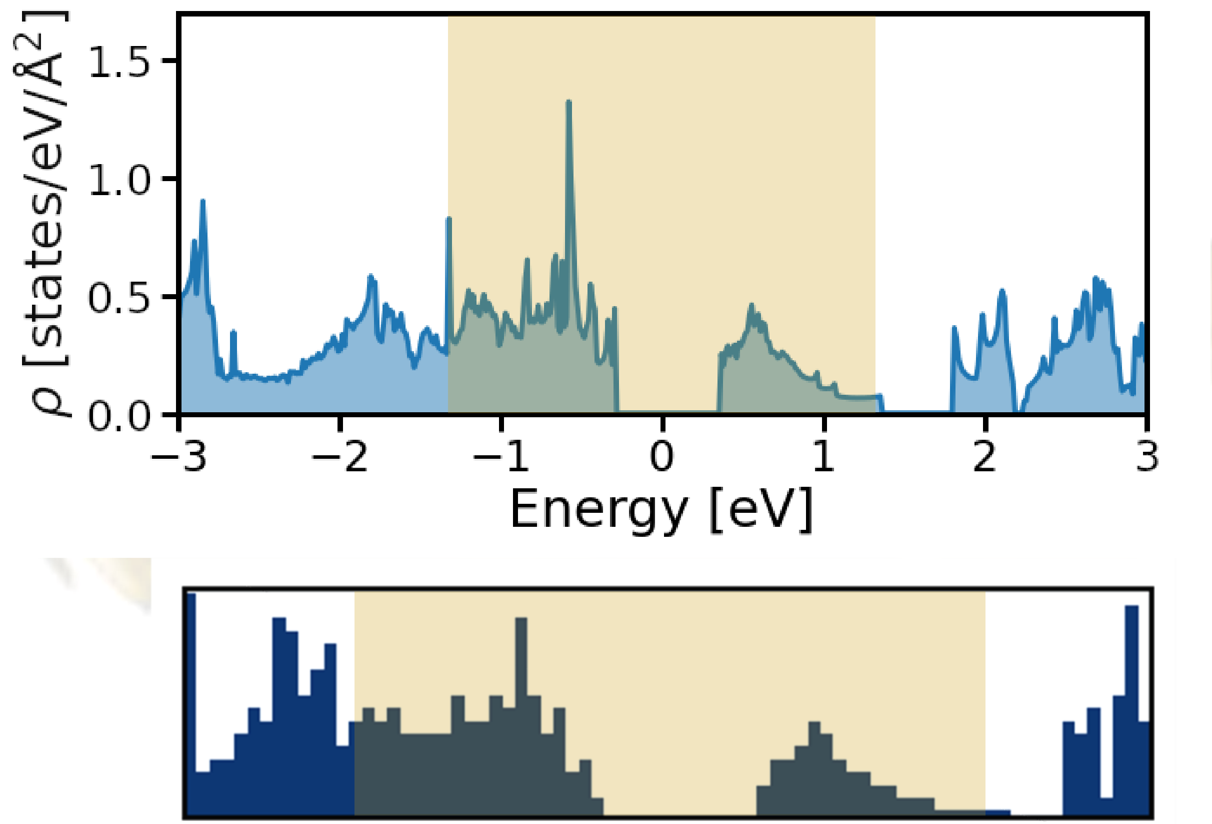
Spectrum fingerprints



M. Kuban *et al. Sci Data* **9**, 646 (2022).

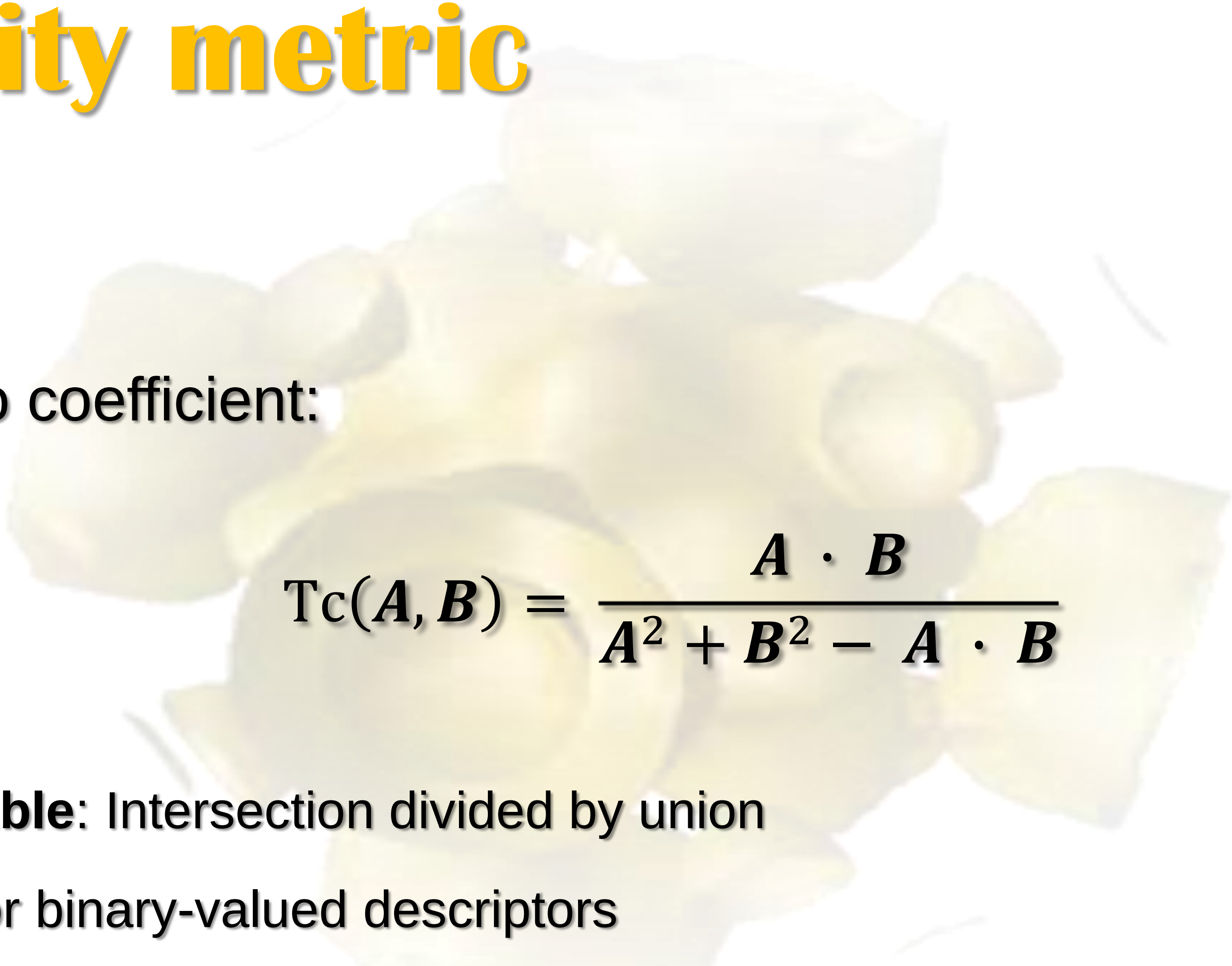
Inspired by O. Isayev, et al., Chem. Mater. 27, 735 (2015)

Fingerprint feature region



M. Kuban *et al.* Sci Data **9**, 646 (2022).

Similarity metric



Tanimoto coefficient:

$$\mathrm{Tc}(\boldsymbol{A},\boldsymbol{B}) = \frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\boldsymbol{A}^2+\boldsymbol{B}^2-\boldsymbol{A}\cdot\boldsymbol{B}}$$

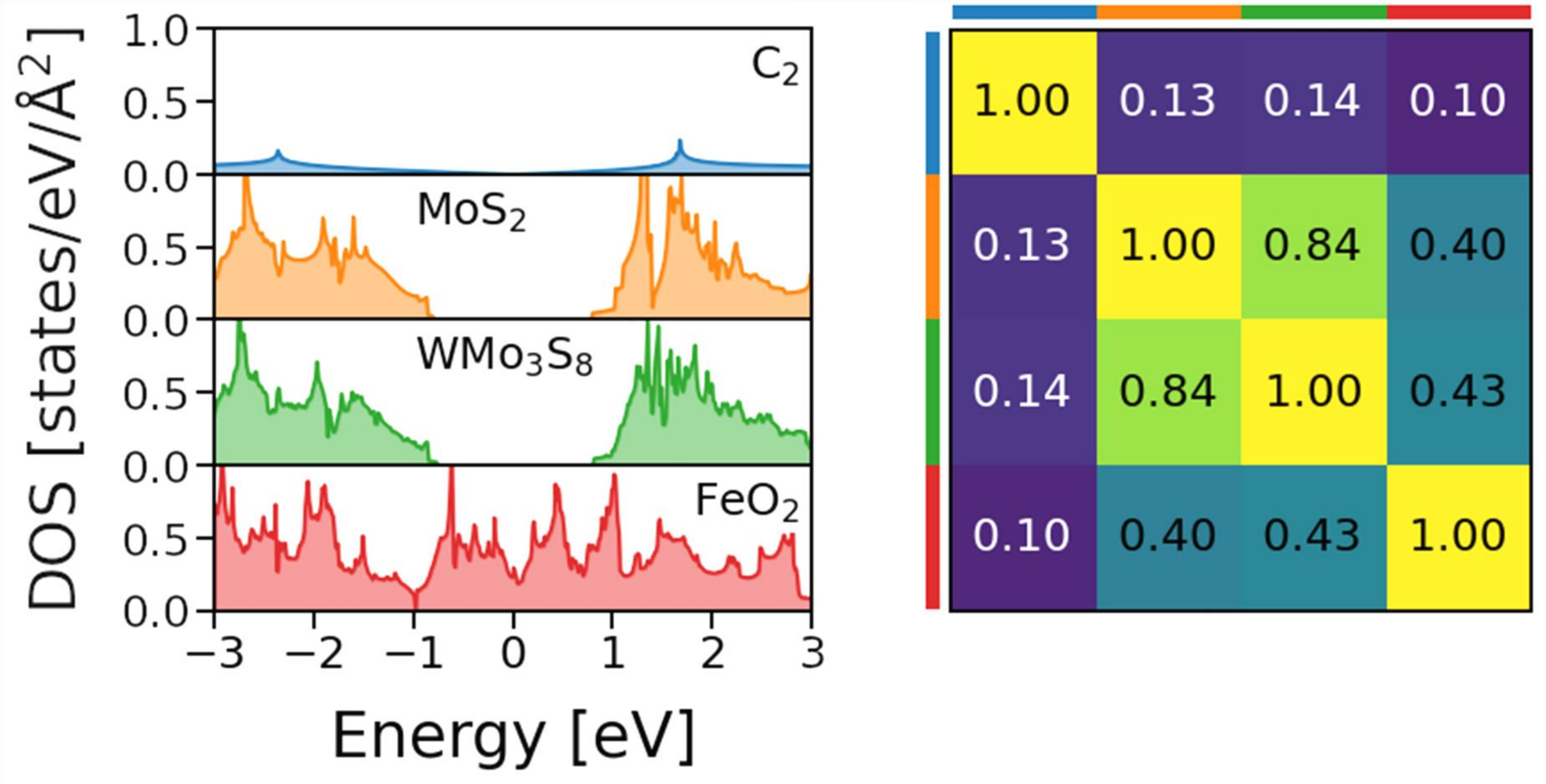
Interpretable: Intersection divided by union

Metric: For binary-valued descriptors

Computationally cheap: Can be described by bitwise operations

Peter Willet et al., *J. Chem. Inf. Comput. Sci.* **38**, 983–996 (1998)

Similarity metric



	🟦	🟧	🟩	🟥
🟦	1.00	0.13	0.14	0.10
🟧	0.13	1.00	0.84	0.40
🟩	0.14	0.84	1.00	0.43
🟥	0.10	0.40	0.43	1.00

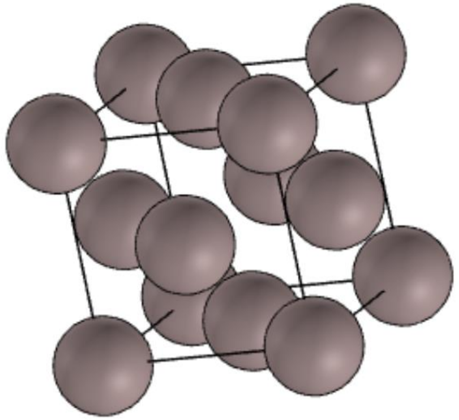
M. Kuban *et al.* Sci Data **9**, 646 (2022).

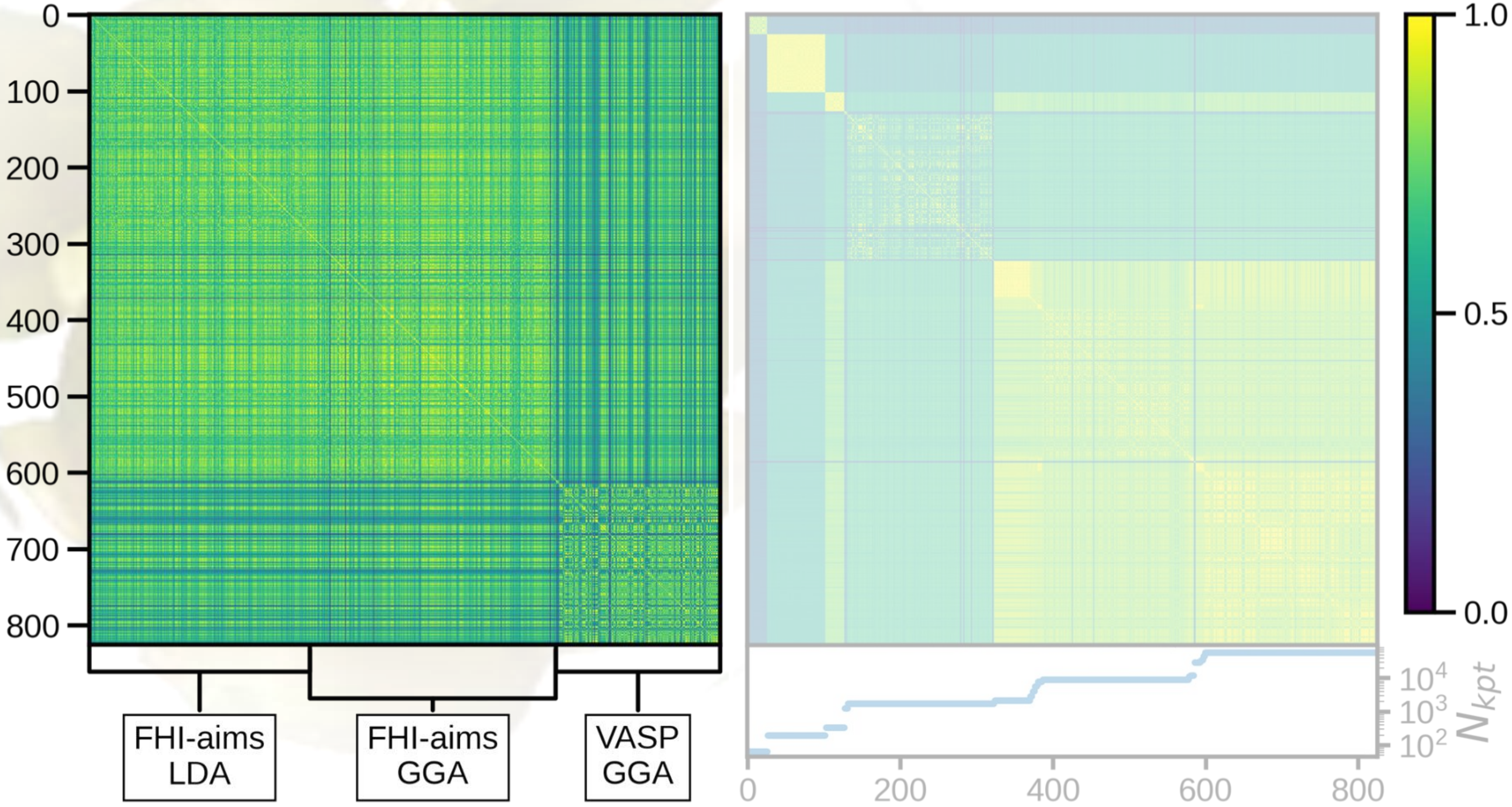
Data quality assessment

Code and computational parameters

NOMAD

fcc Al

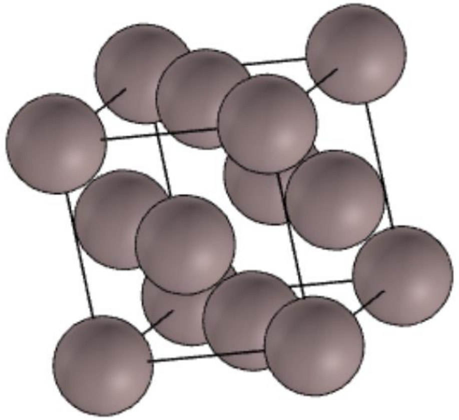


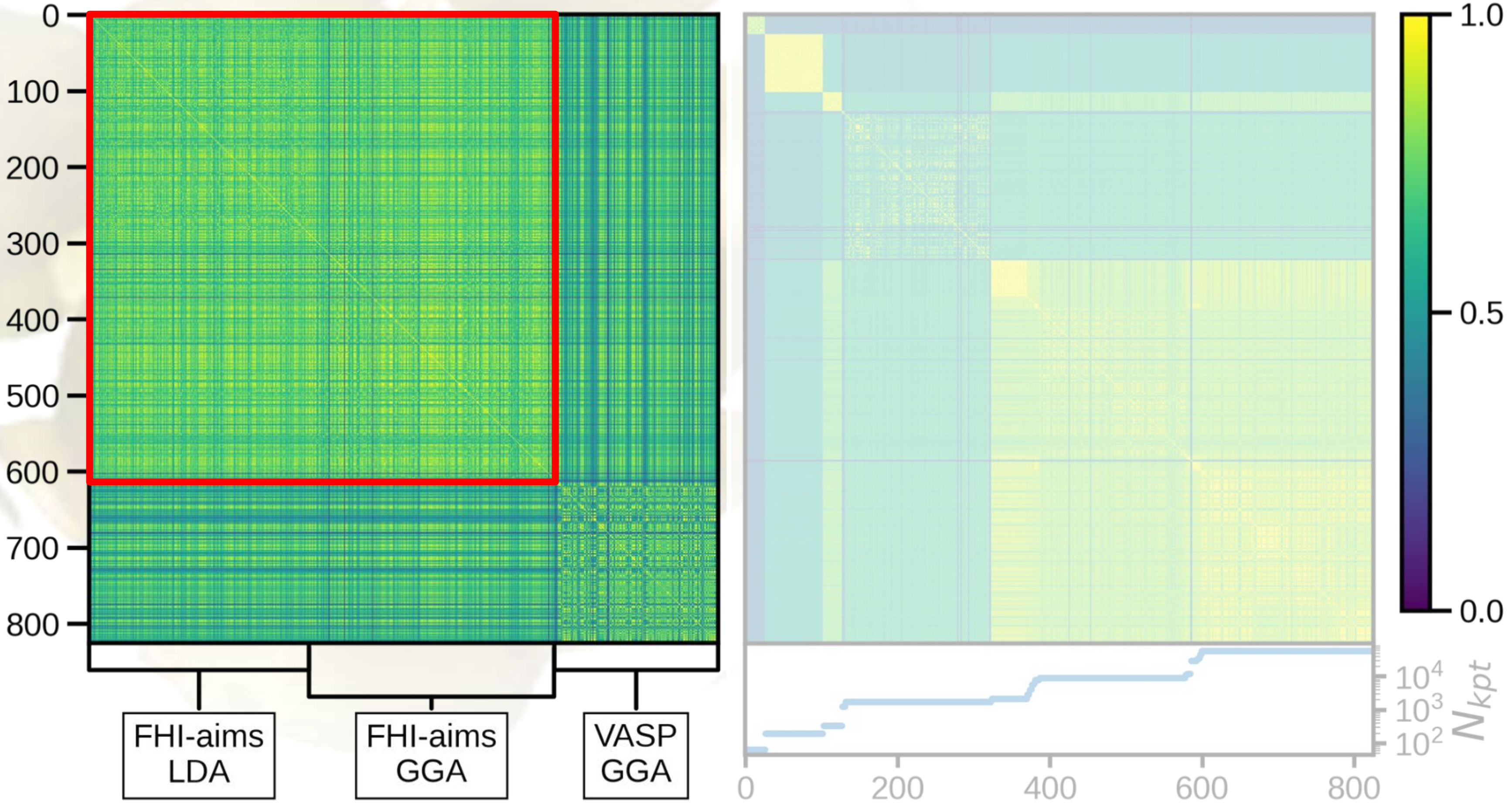


M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Code and computational parameters

fcc Al



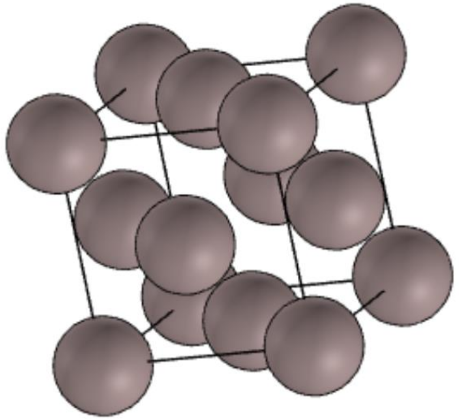


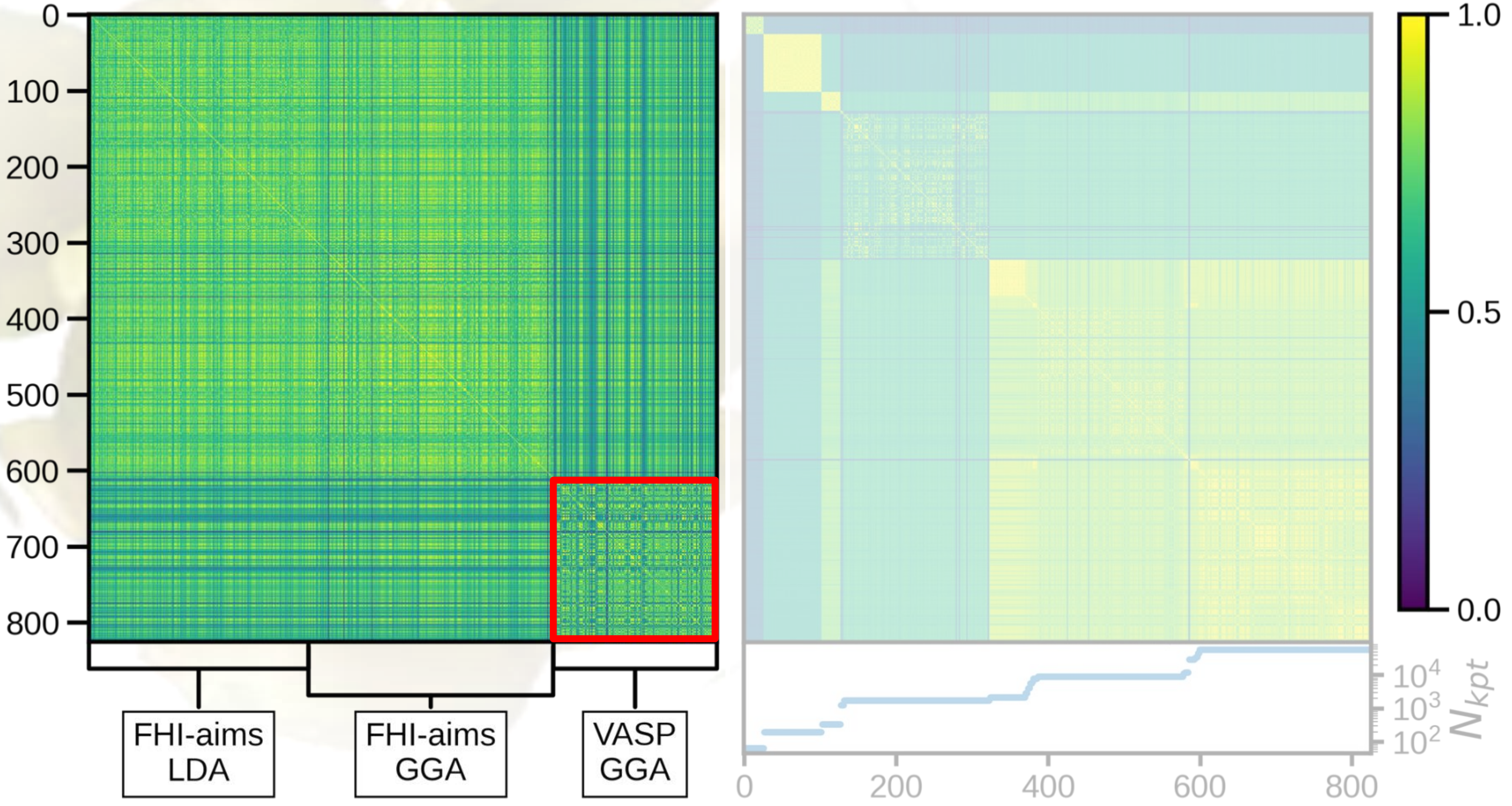
M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Code and computational parameters

NOMAD

fcc Al

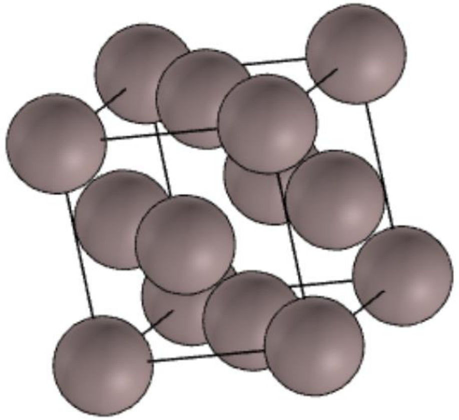


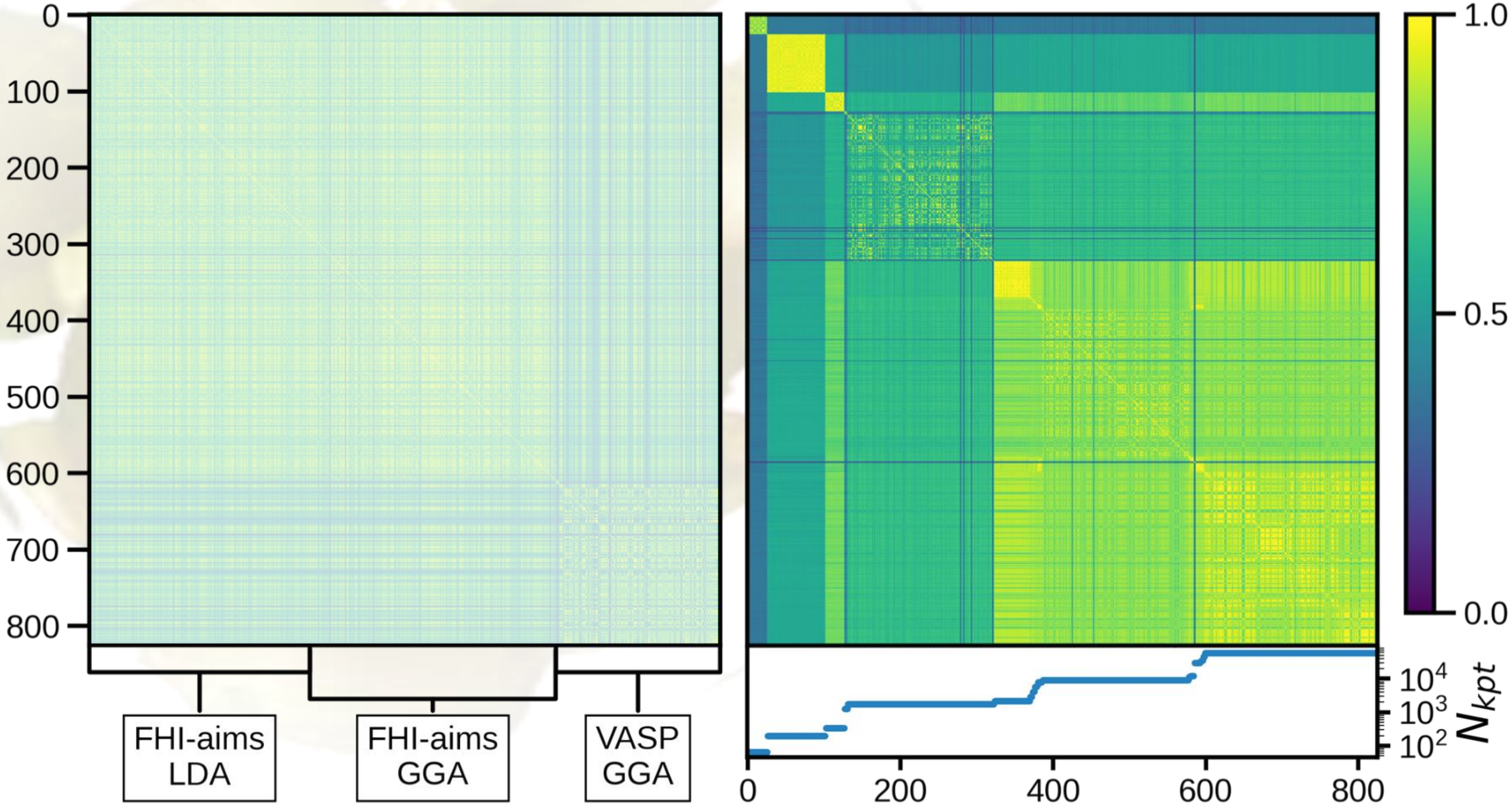


M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Code and computational parameters

fcc Al

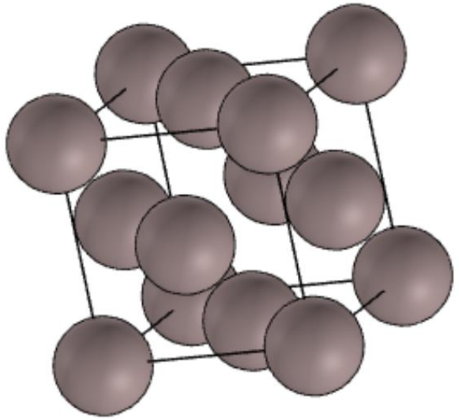


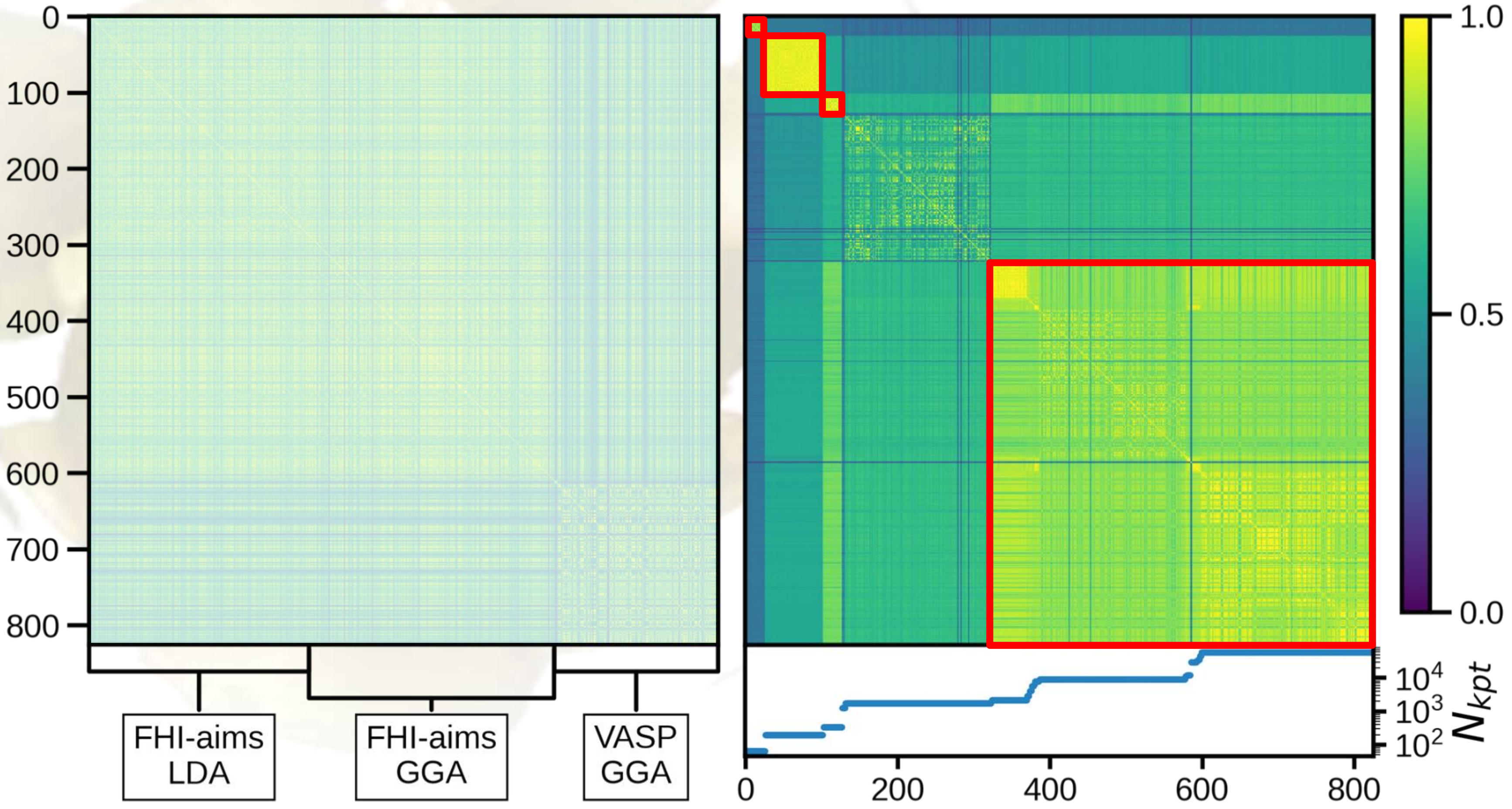


M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Code and computational parameters

fcc Al



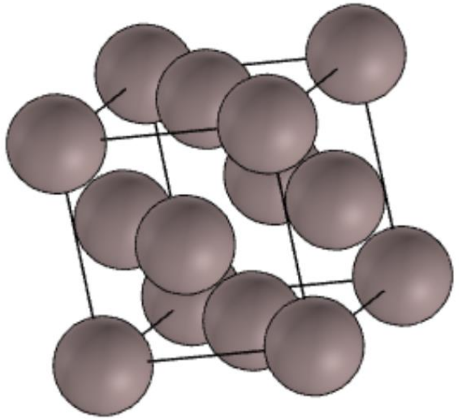


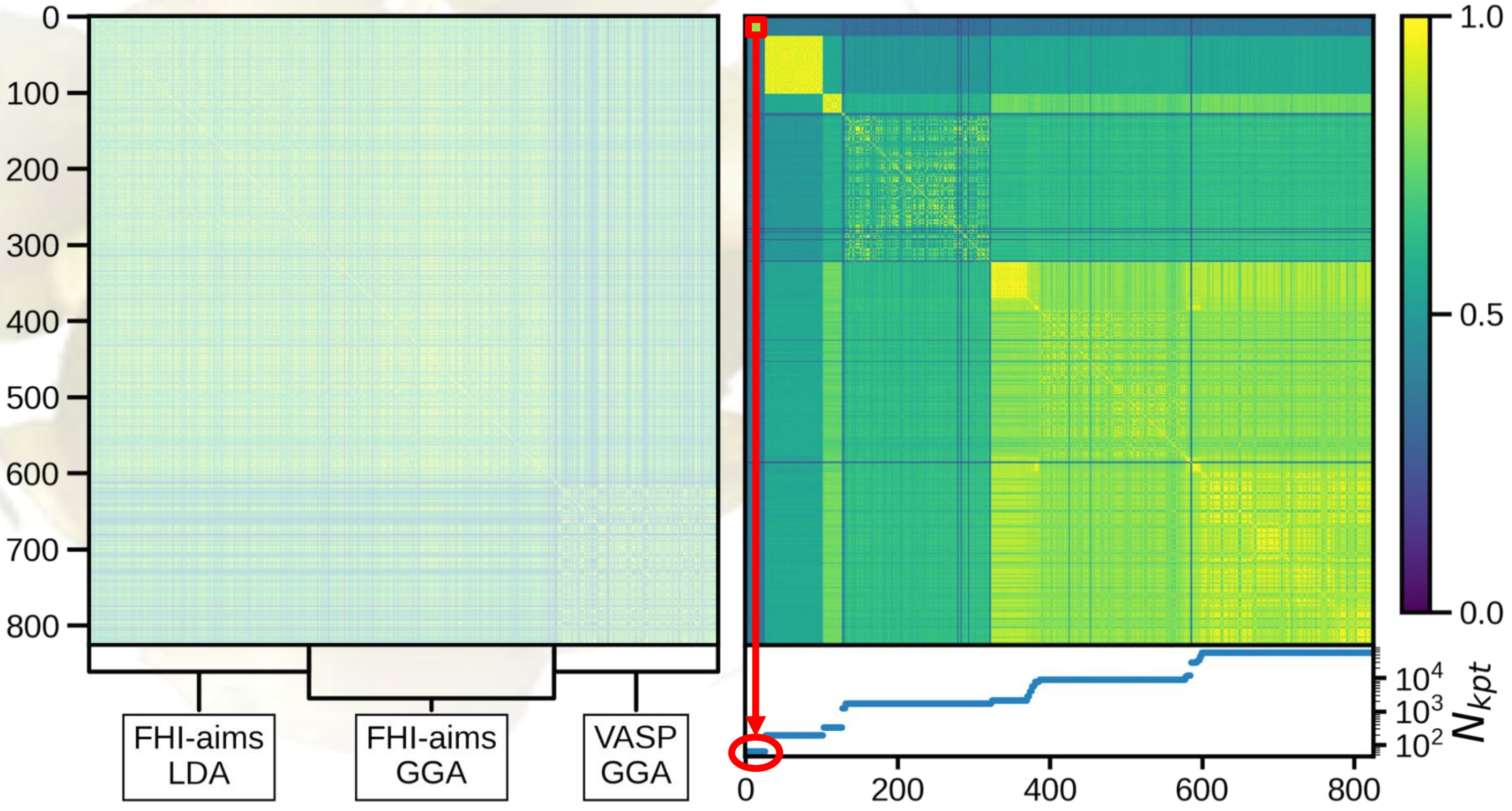
M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Code and computational parameters

NOMAD

fcc Al



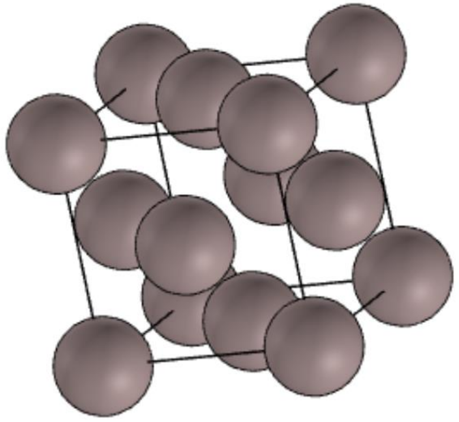


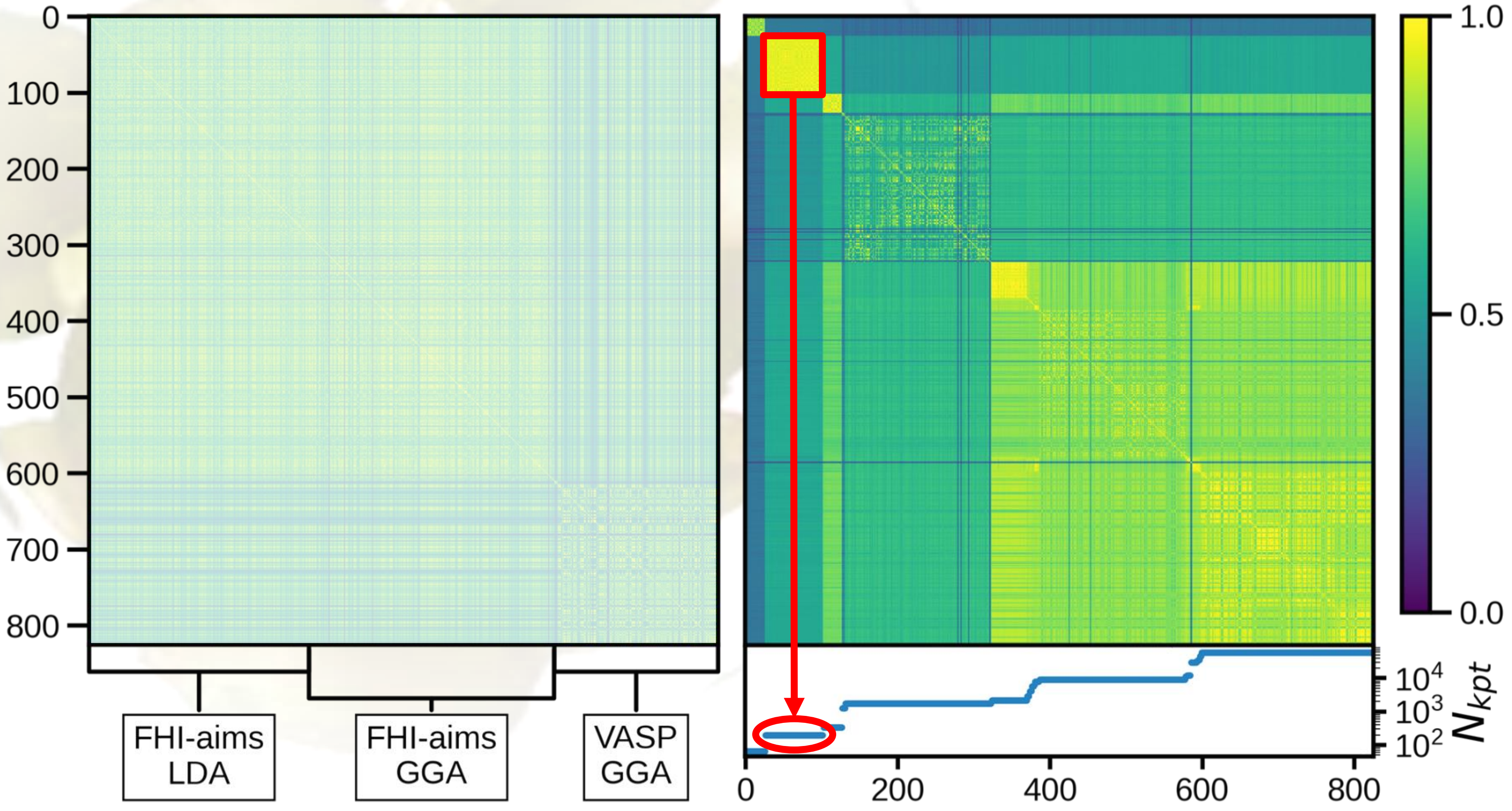
M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Code and computational parameters

NOMAD

fcc Al



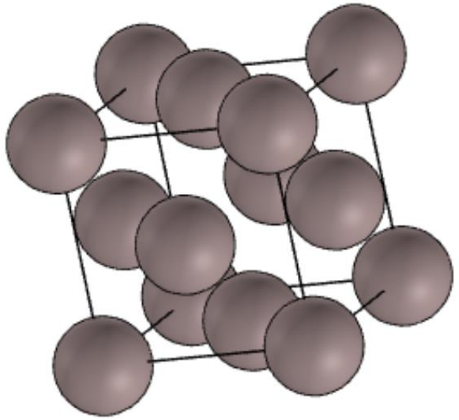


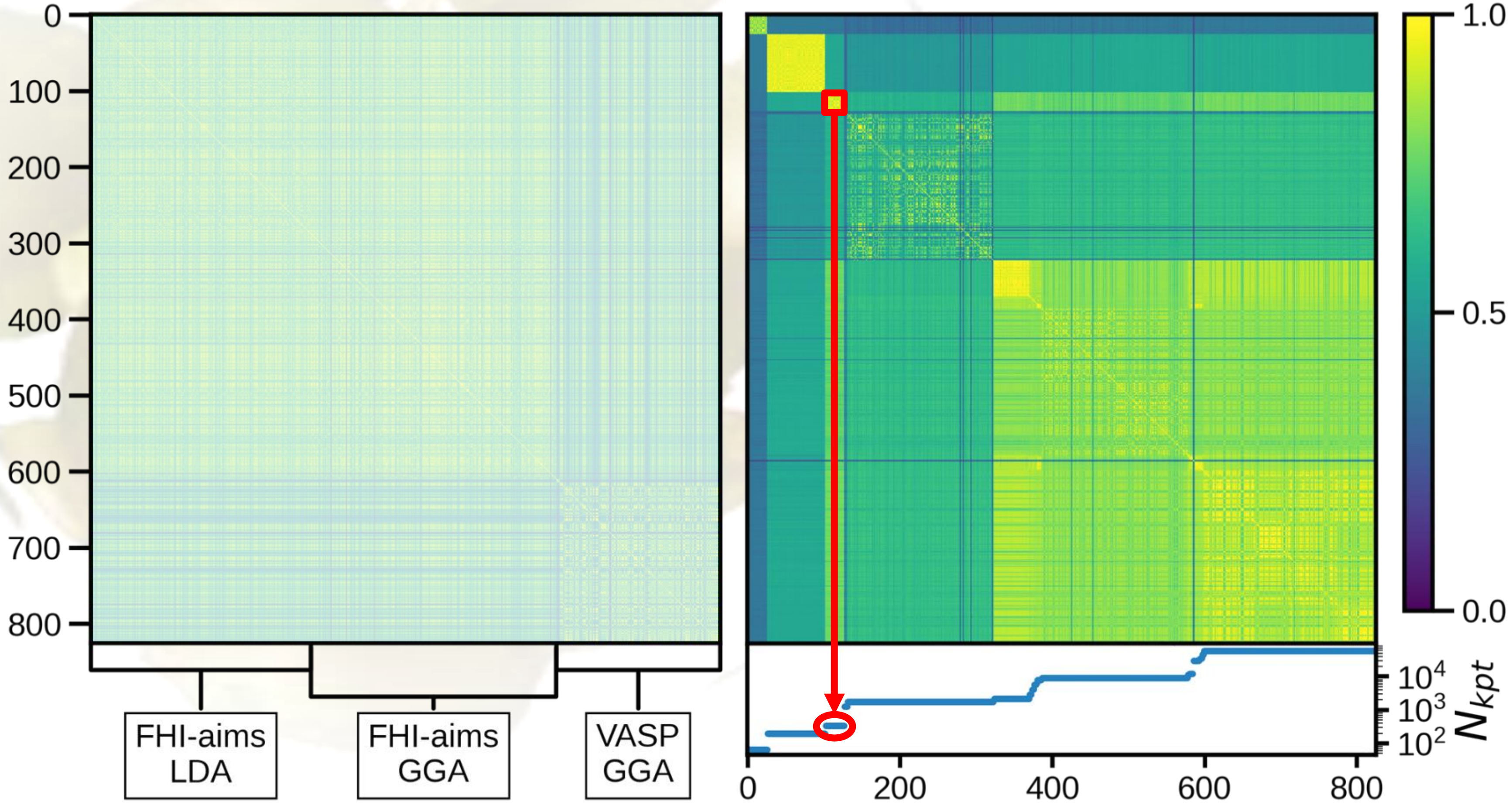
M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Code and computational parameters

NOMAD

fcc Al



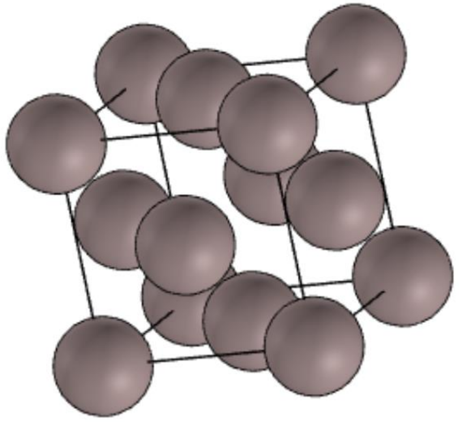


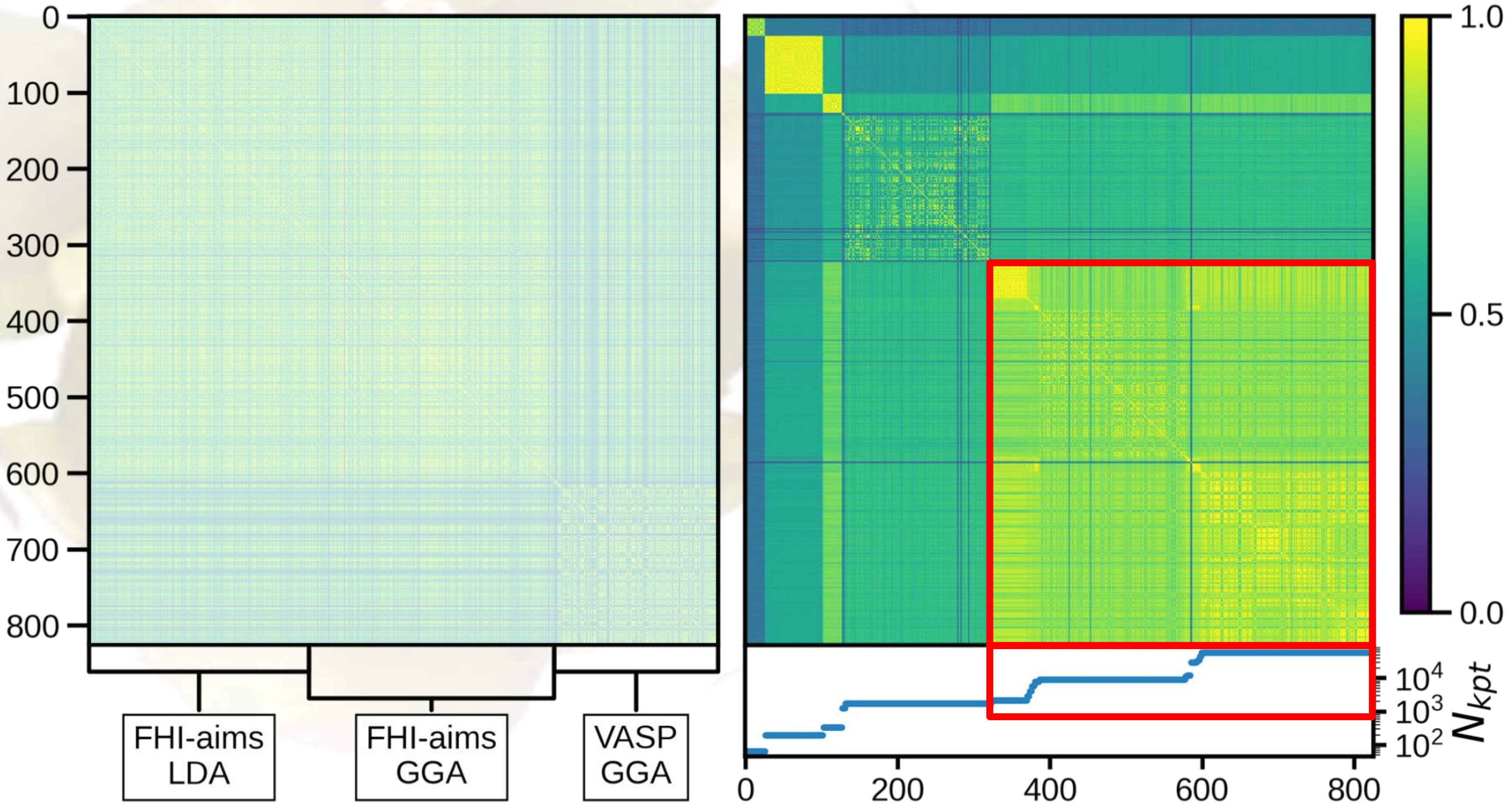
M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Code and computational parameters

NOMAD

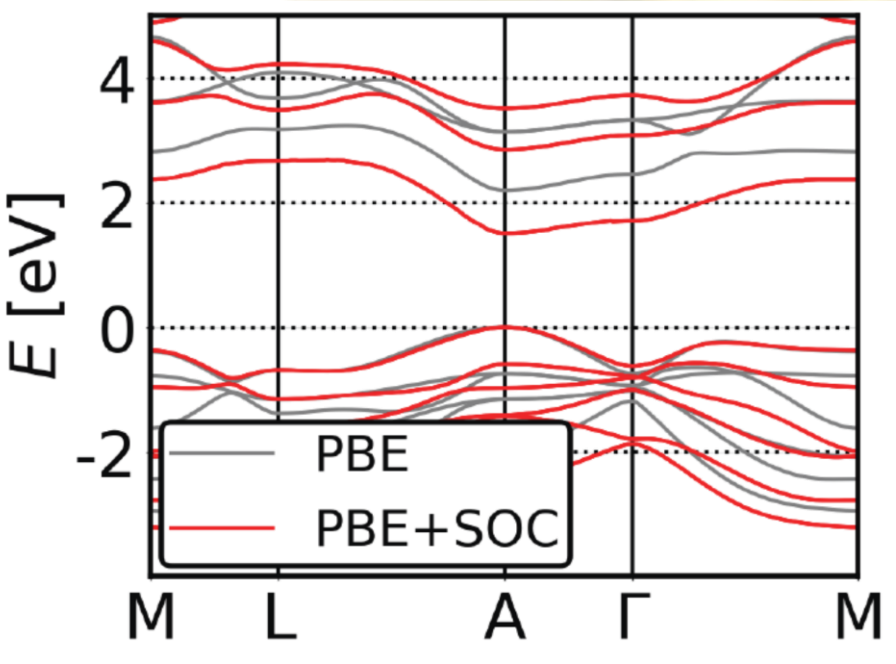
fcc Al

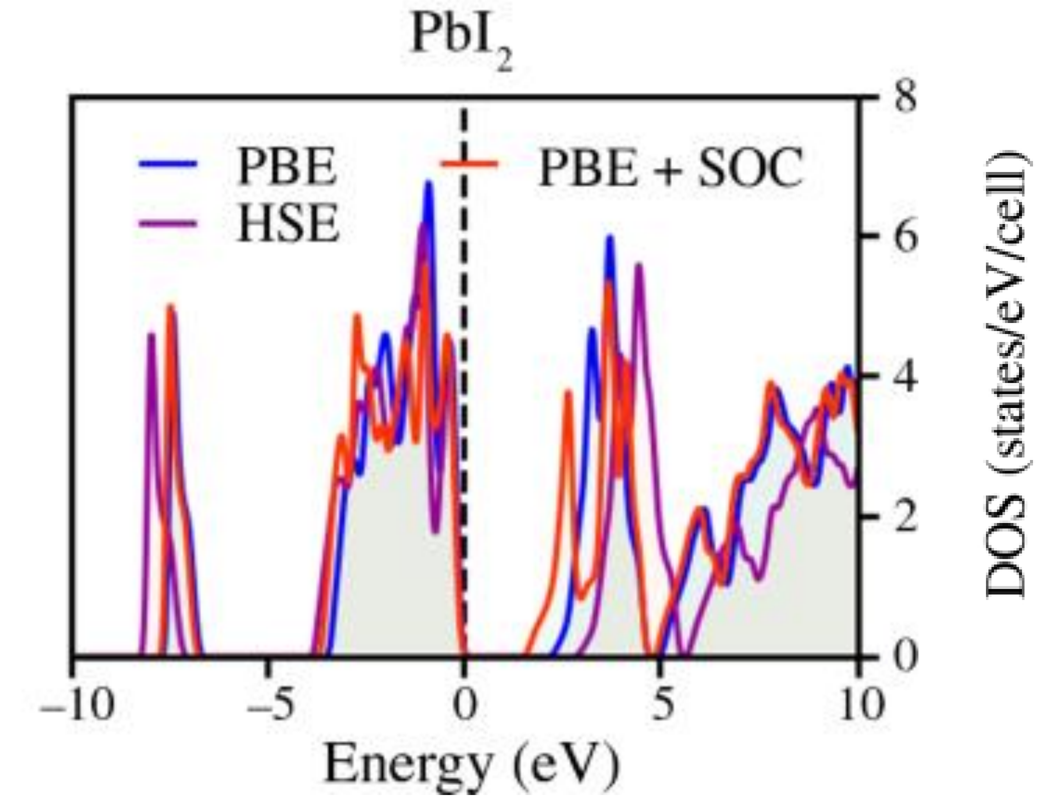




M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Methodology

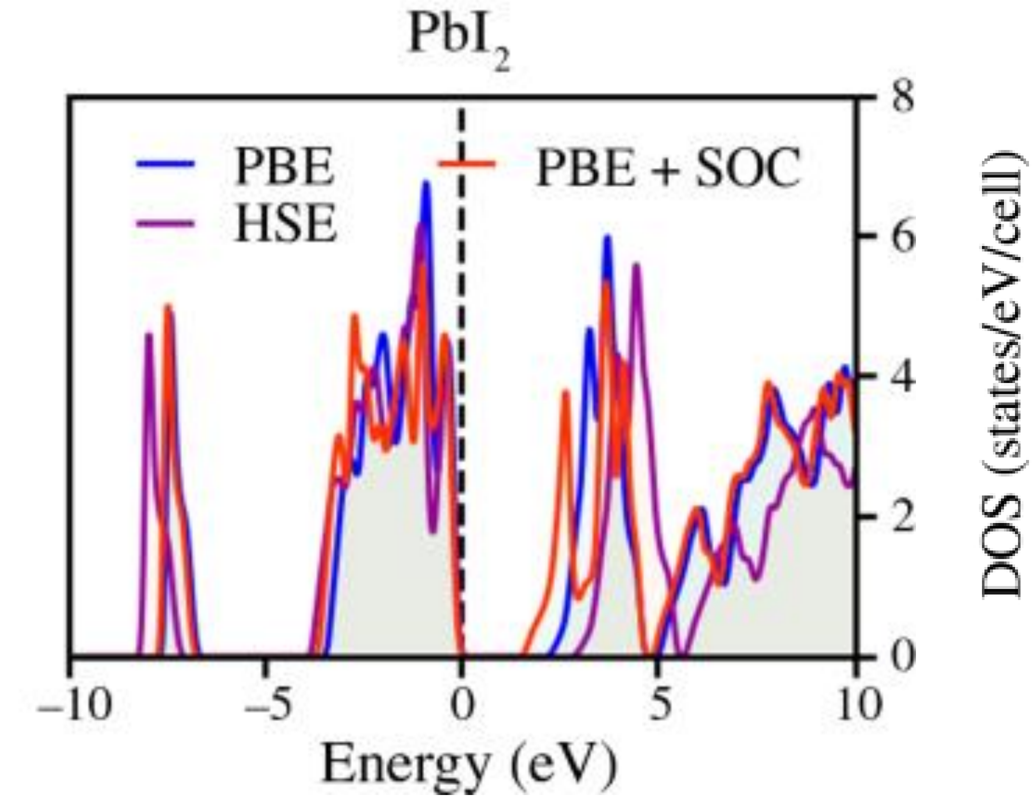




M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Methodology

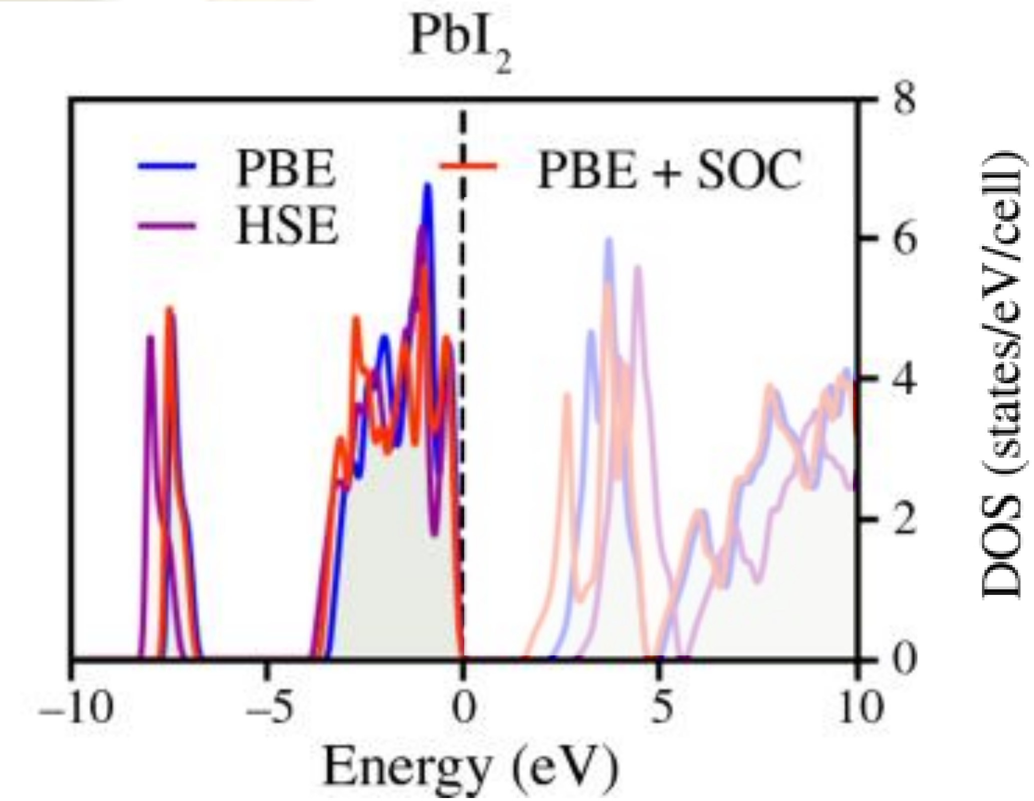
PBE vs.	S_{total}	$S_{valence}$	$S_{conduction}$
PBE + SOC	0.71	0.75	0.67
HSE	0.69	0.73	0.45



M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Methodology

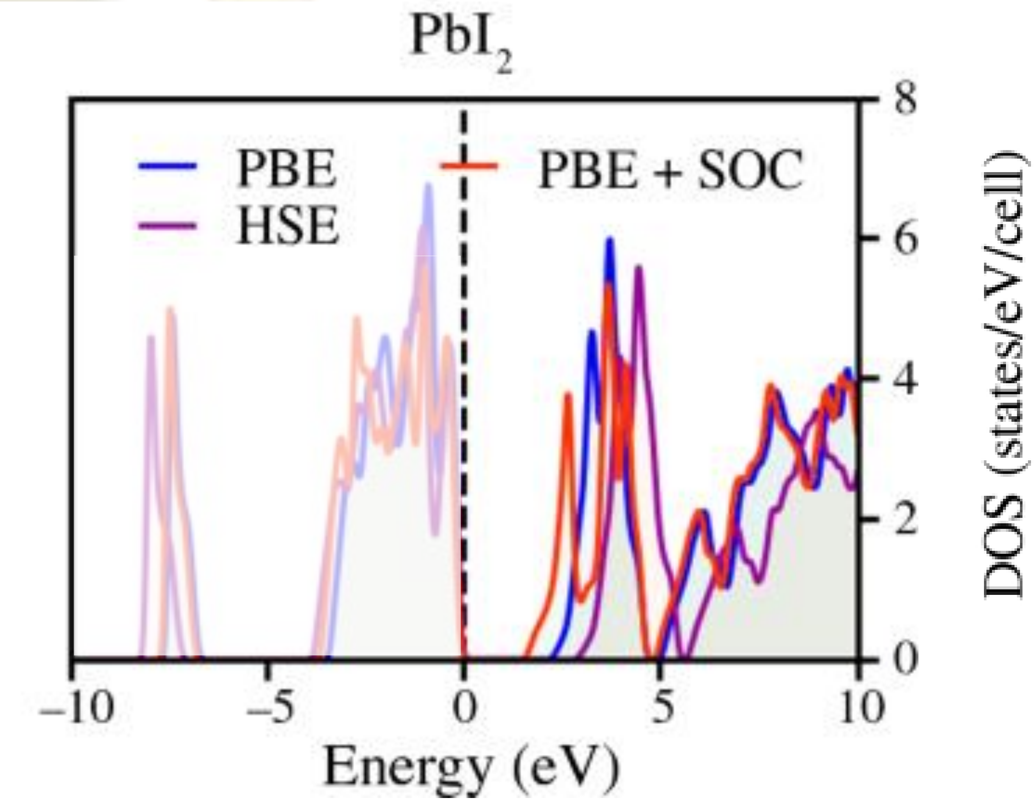
PBE vs.	S_{total}	$S_{valence}$	$S_{conduction}$
PBE + SOC	0.71	0.75	0.67
HSE	0.69	0.73	0.45



M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Methodology

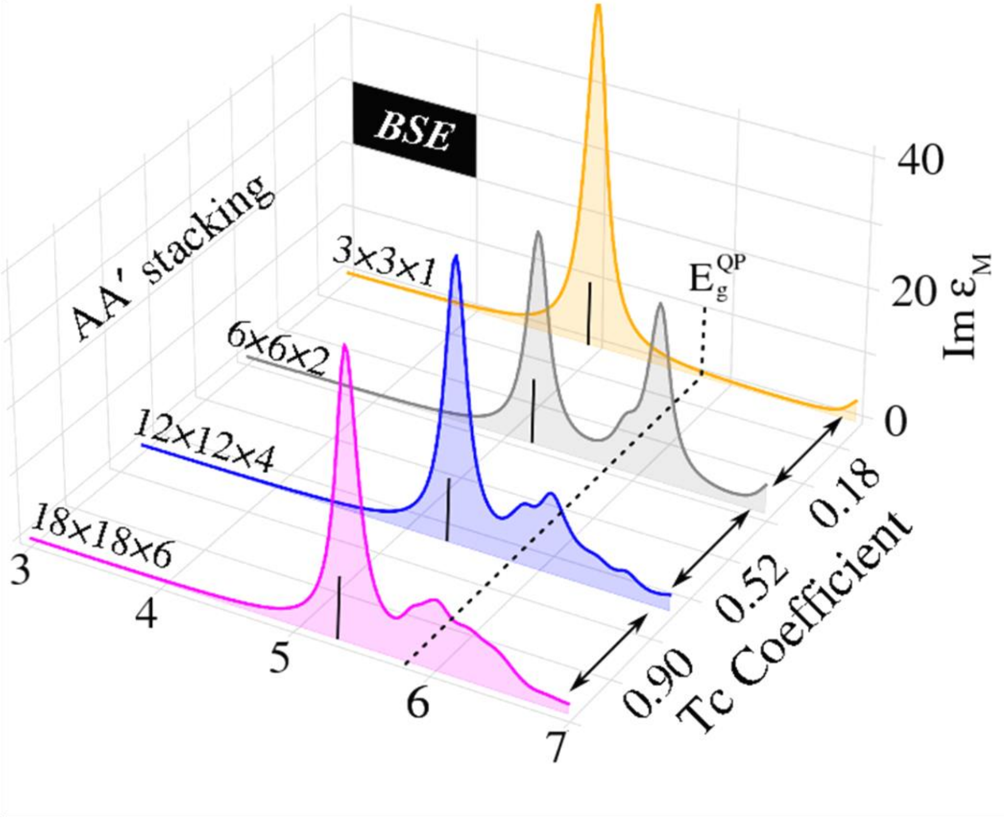
PBE vs.	S_{total}	$S_{valence}$	$S_{conduction}$
PBE + SOC	0.71	0.75	0.67
HSE	0.69	0.73	0.45

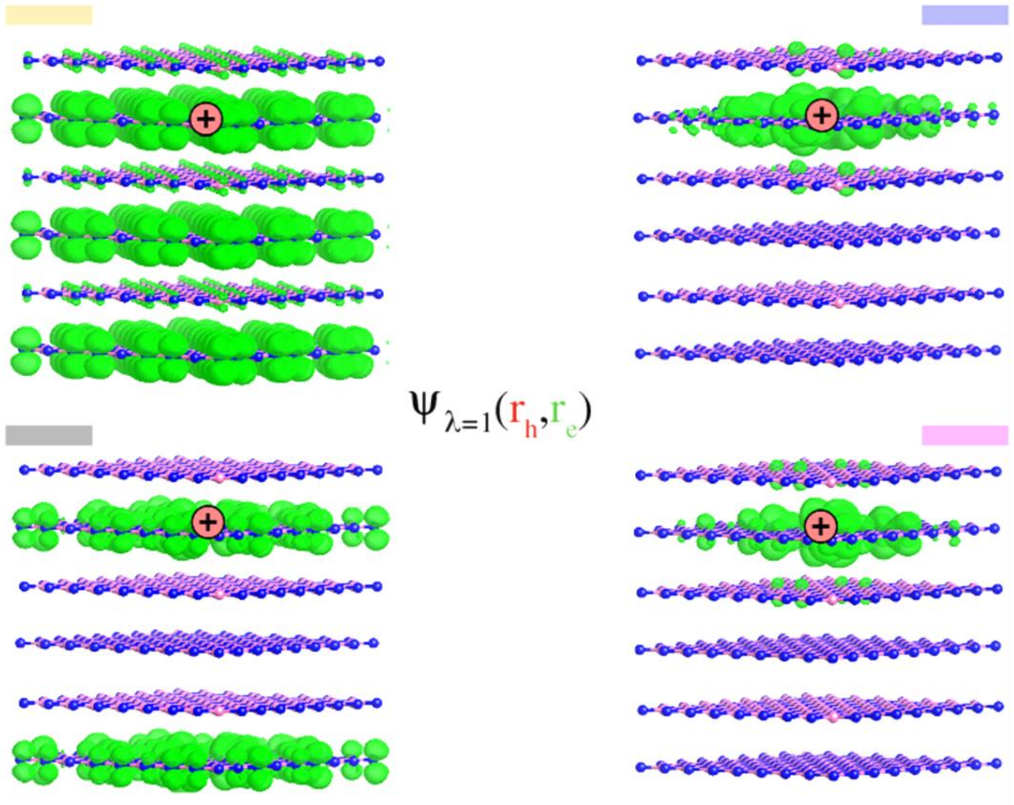


M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Excitonic spectra

$$\phi^{\lambda}(\mathbf{r}_e,\mathbf{r}_h)=\sum A^{\lambda}_{v c \mathbf{k}}\,\psi^{*}_{v\mathbf{k}}(\mathbf{r}_h)\,\psi_{c\mathbf{k}}(\mathbf{r}_e)$$

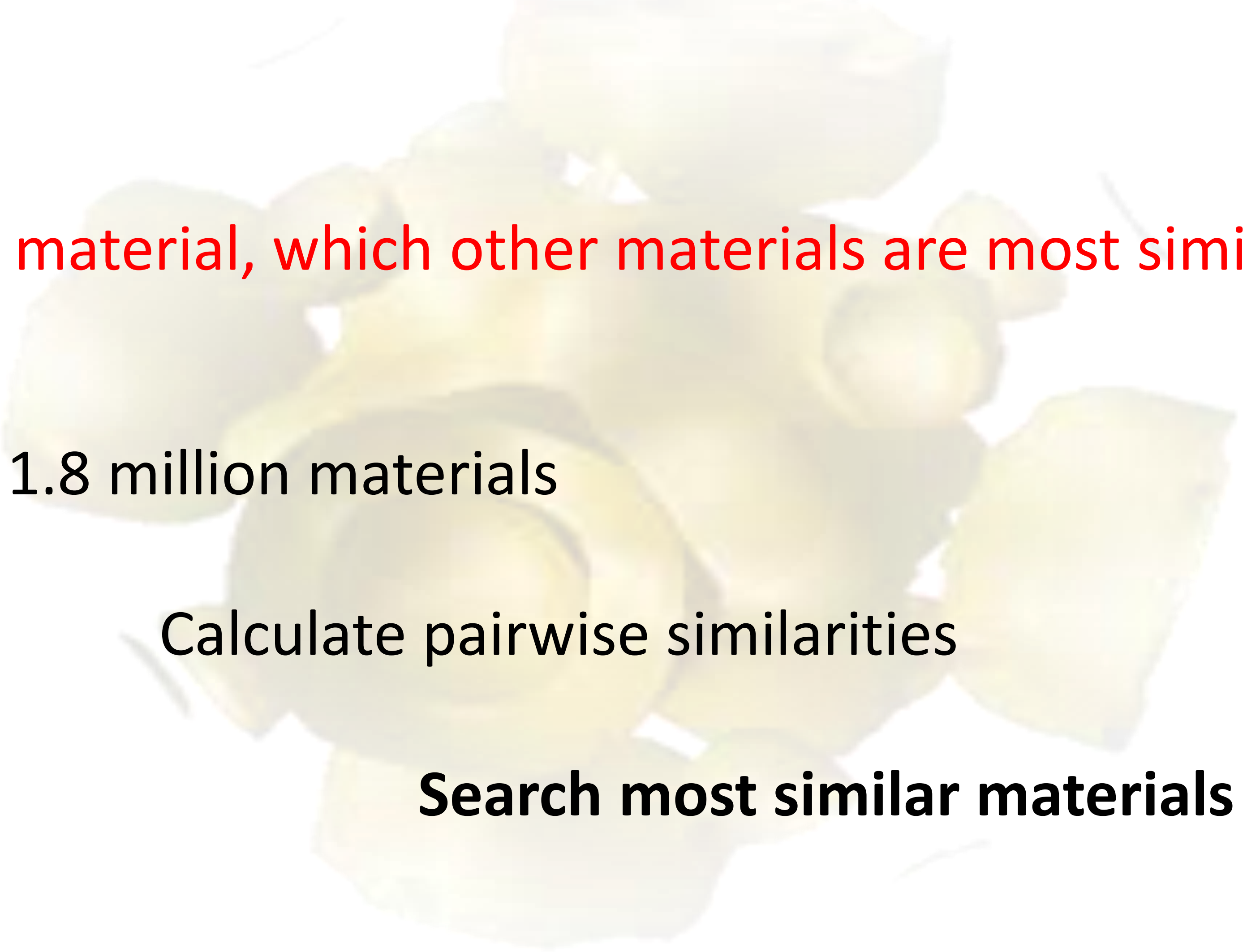




M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Exploring data spaces

Finding similar materials



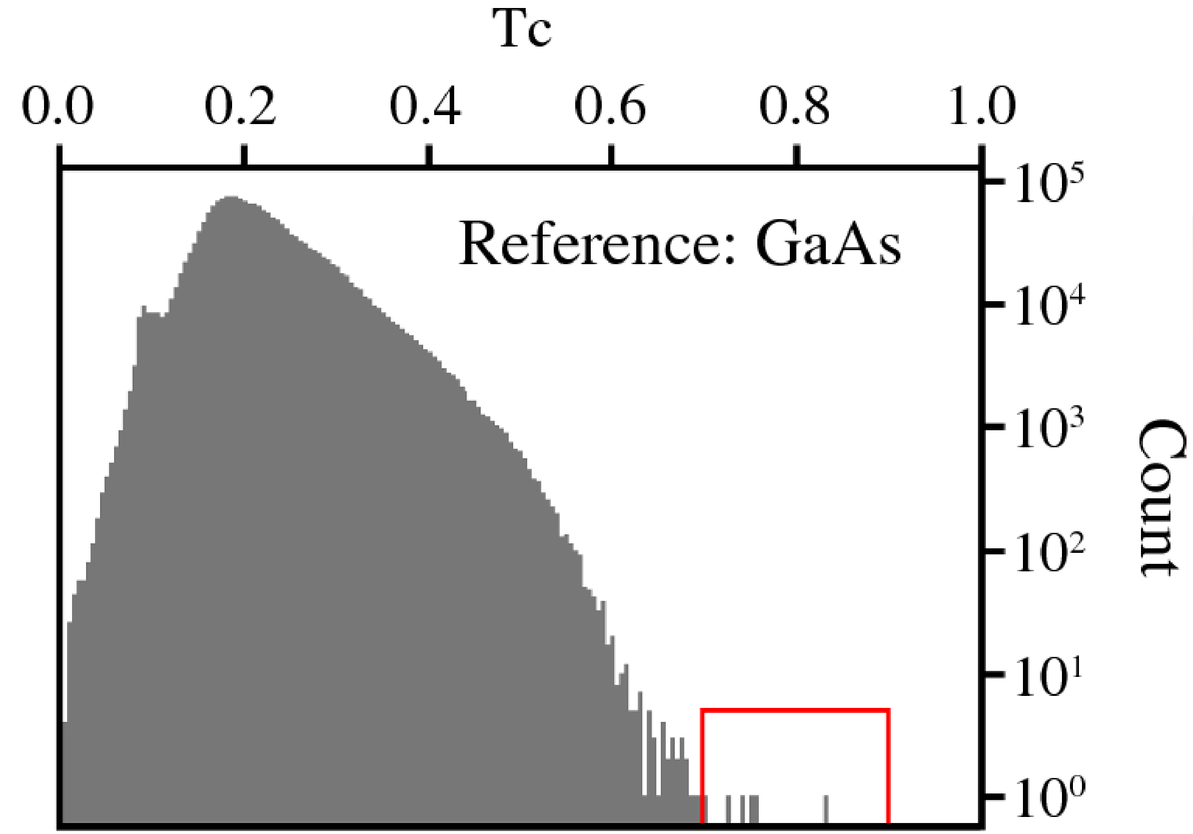
For any material, which other materials are most similar to them?

1.8 million materials

Calculate pairwise similarities

Search most similar materials

Finding similar materials

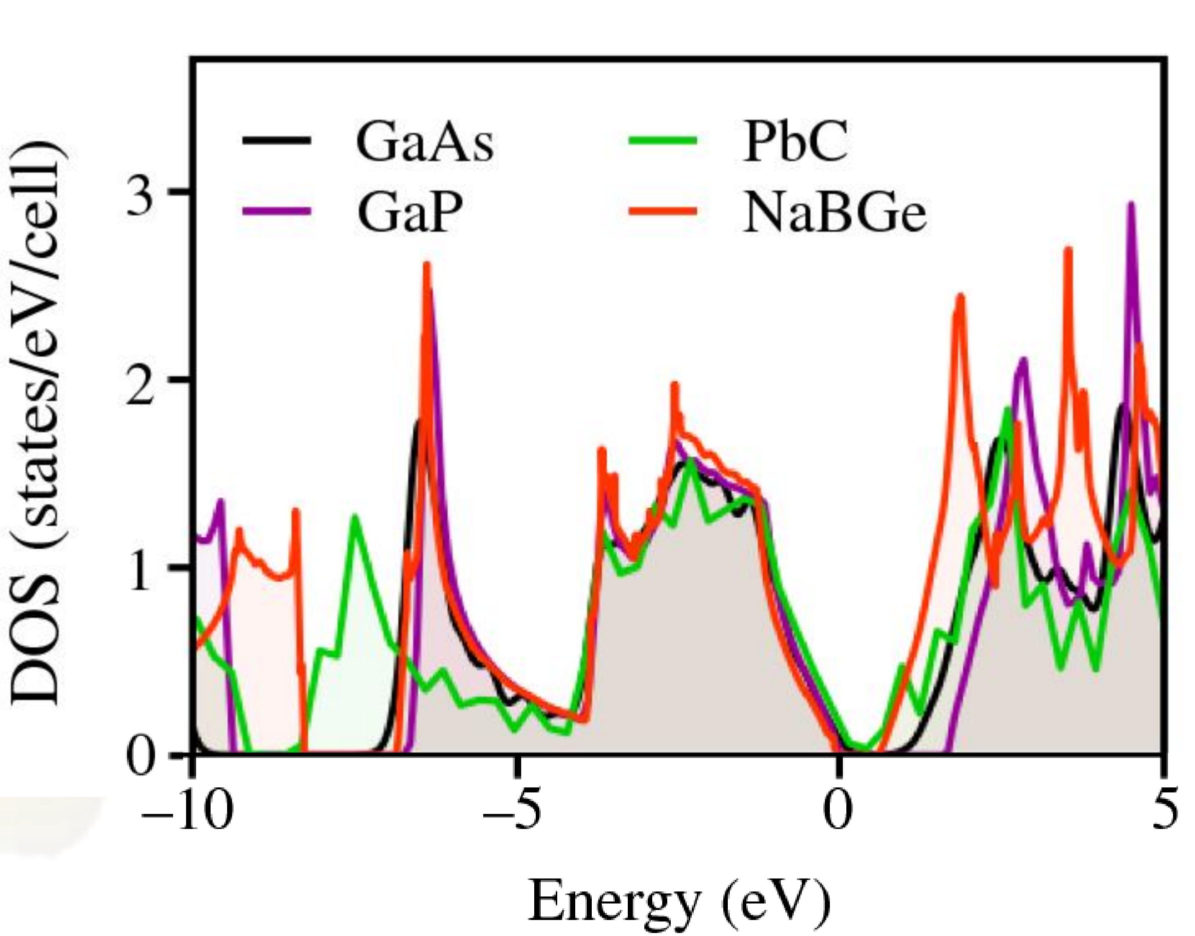


M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Finding similar materials

Similarity of GaAs to:

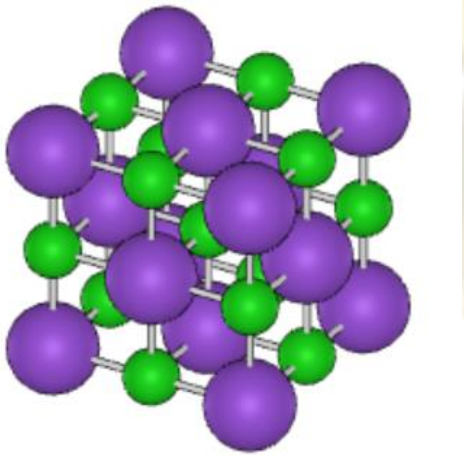
GaP:	0.83
PbC:	0.75
NaBGe:	0.74

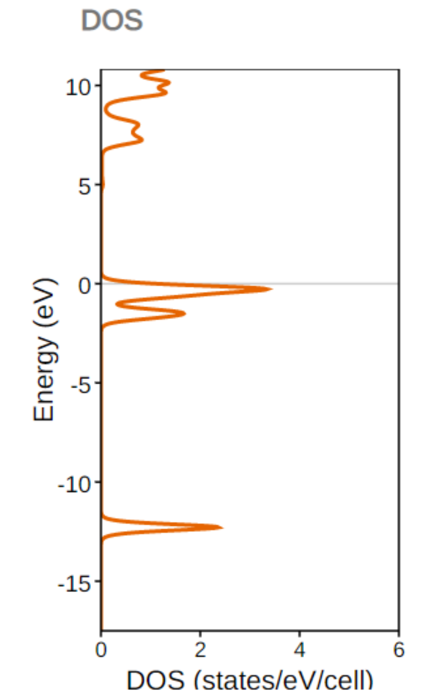


M. Kuban *et al.* MRS Bulletin **47**, 991–999 (2022).

Finding similar materials

NaCl - space group 225

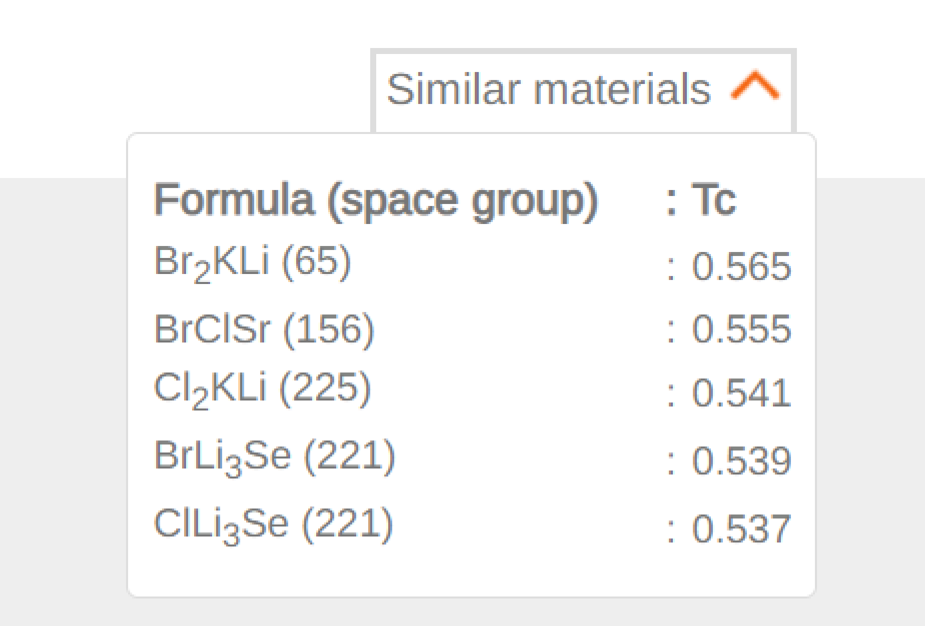




From calculation **zz6kLYmE**

(GGA - VASP)

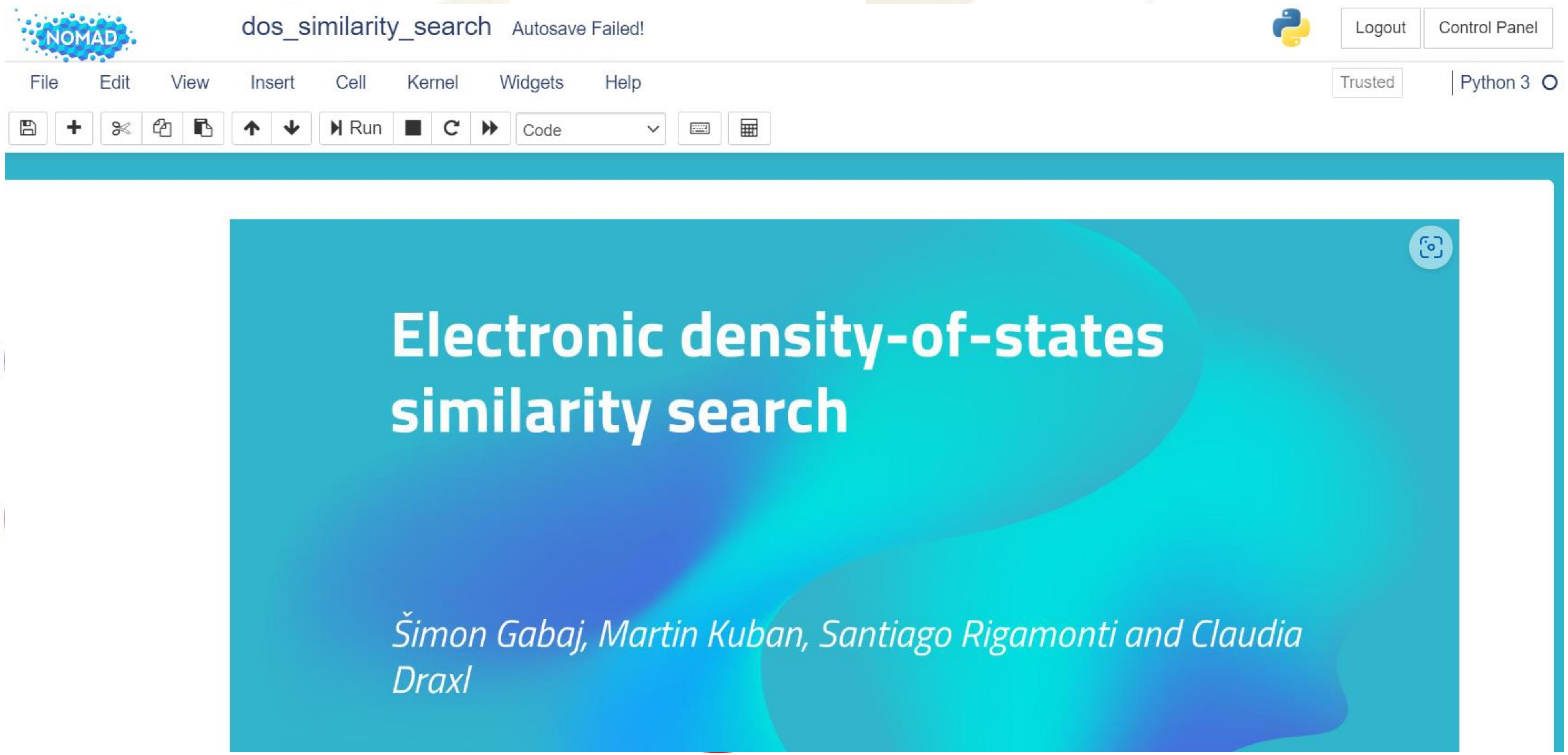
Similar materials



Similar materials

Formula (space group)	: Tc
Br_2KLi (65)	: 0.565
BrClSr (156)	: 0.555
Cl_2KLi (225)	: 0.541
$BrLi_3Se$ (221)	: 0.539
$ClLi_3Se$ (221)	: 0.537

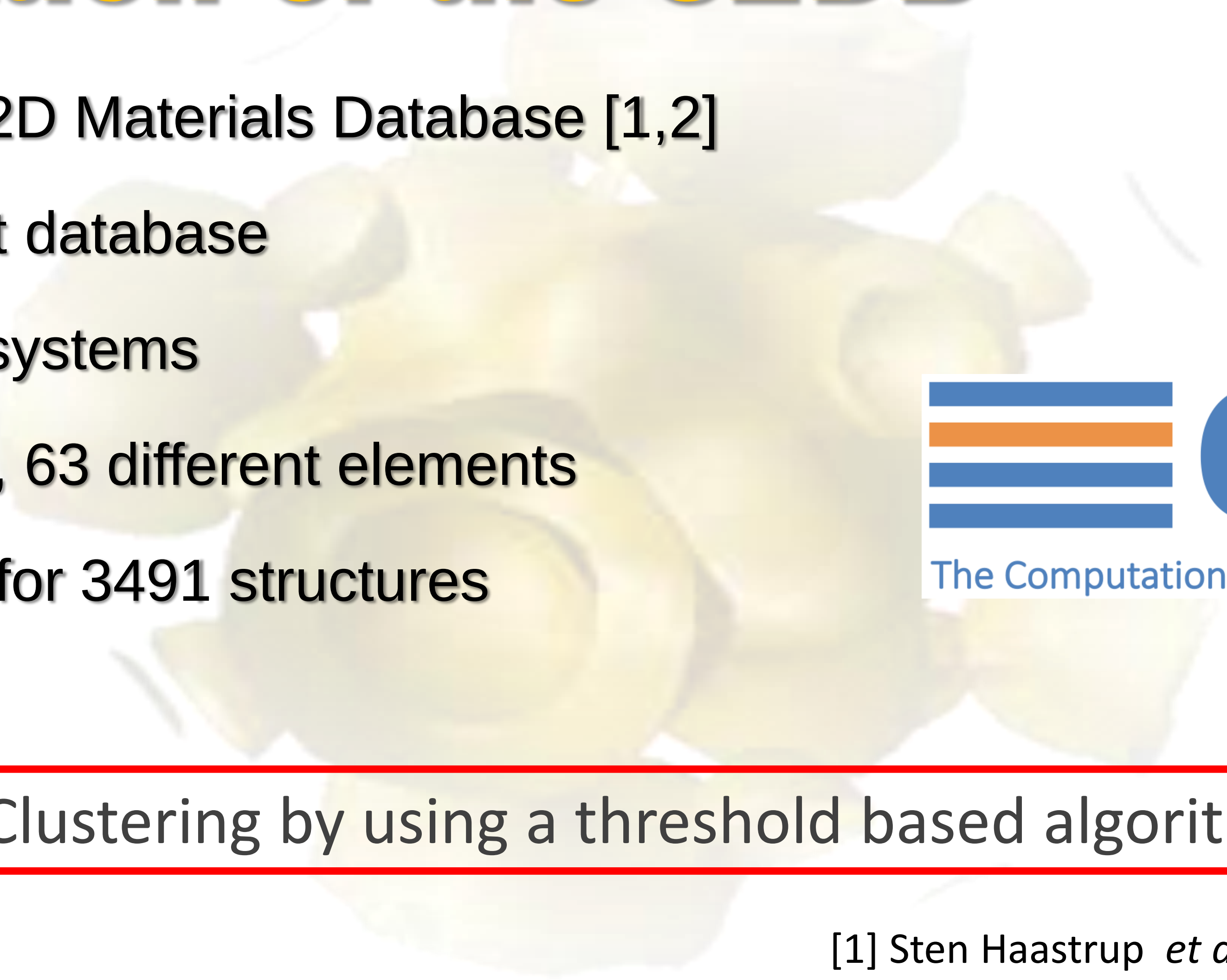
Finding similar materials



Electronic density-of-states similarity search

Šimon Gabaj, Martin Kuban, Santiago Rigamonti and Claudia Draxl

Exploration of the C2DB



Computational 2D Materials Database [1,2]

High-throughput database

Atomically thin systems

4047 structures, 63 different elements

Projected DOS for 3491 structures

C2DB

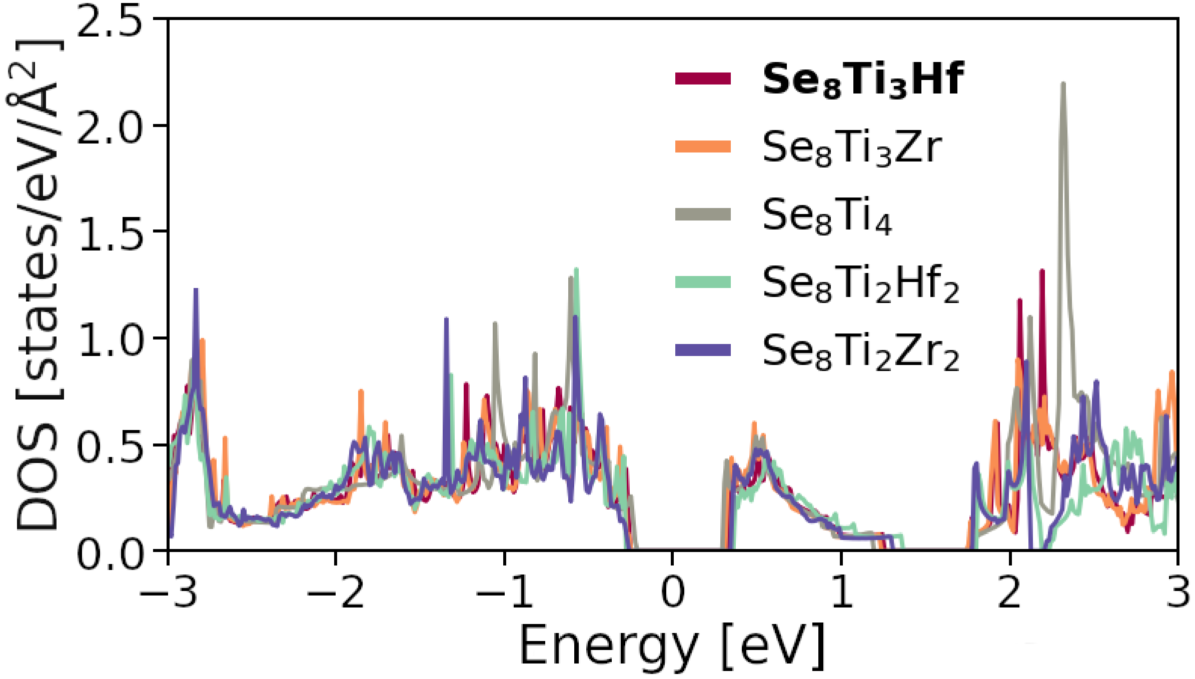
The Computational 2D Materials Database

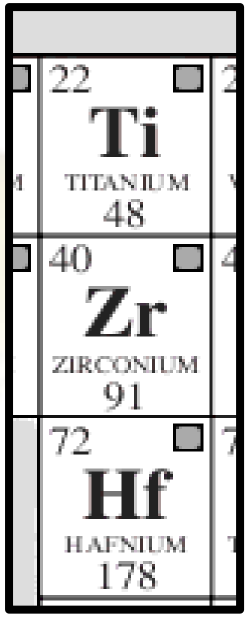
Clustering by using a threshold based algorithm

[1] Sten Haastrup *et al.*, 2D Materials 5, 042002 (2018)

[2] M. N. Gjerding *et al.*, 2D Materials 8, 044002 (2021)

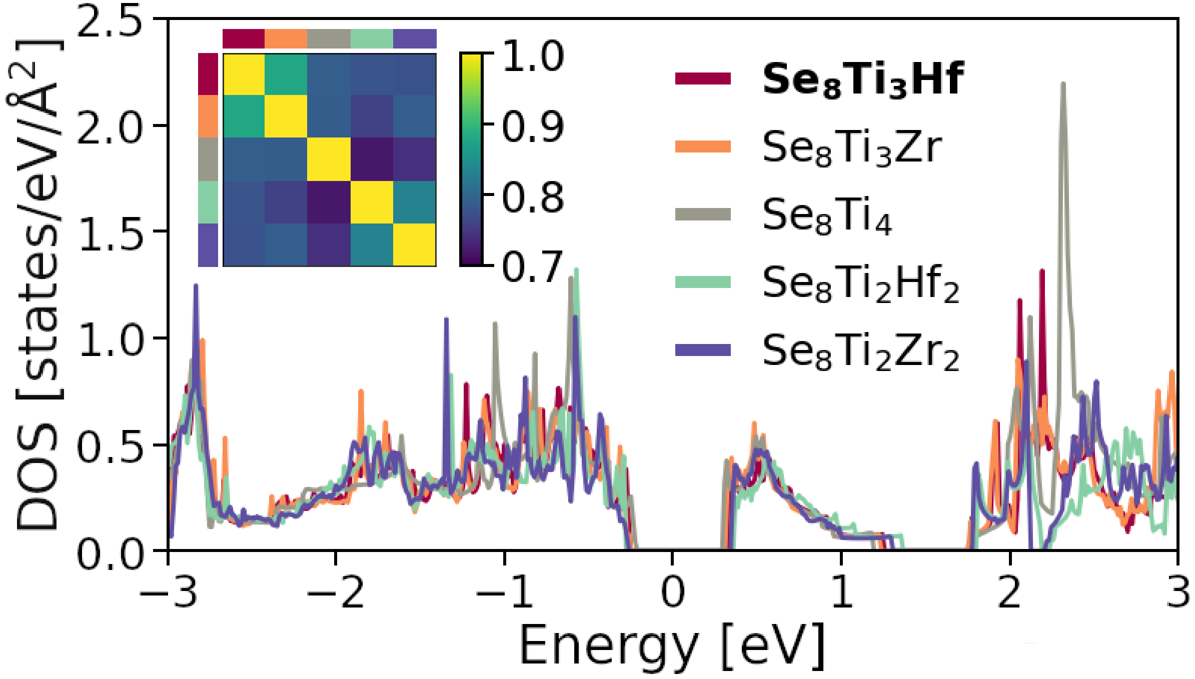
Cluster analysis

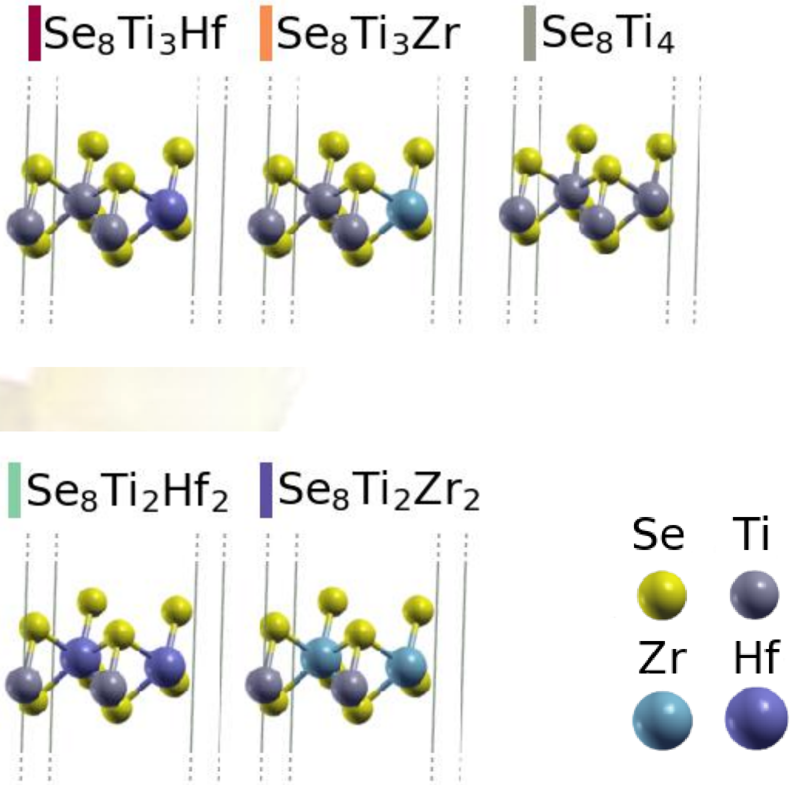




M. Kuban *et al.* Sci Data **9**, 646 (2022).

Cluster analysis





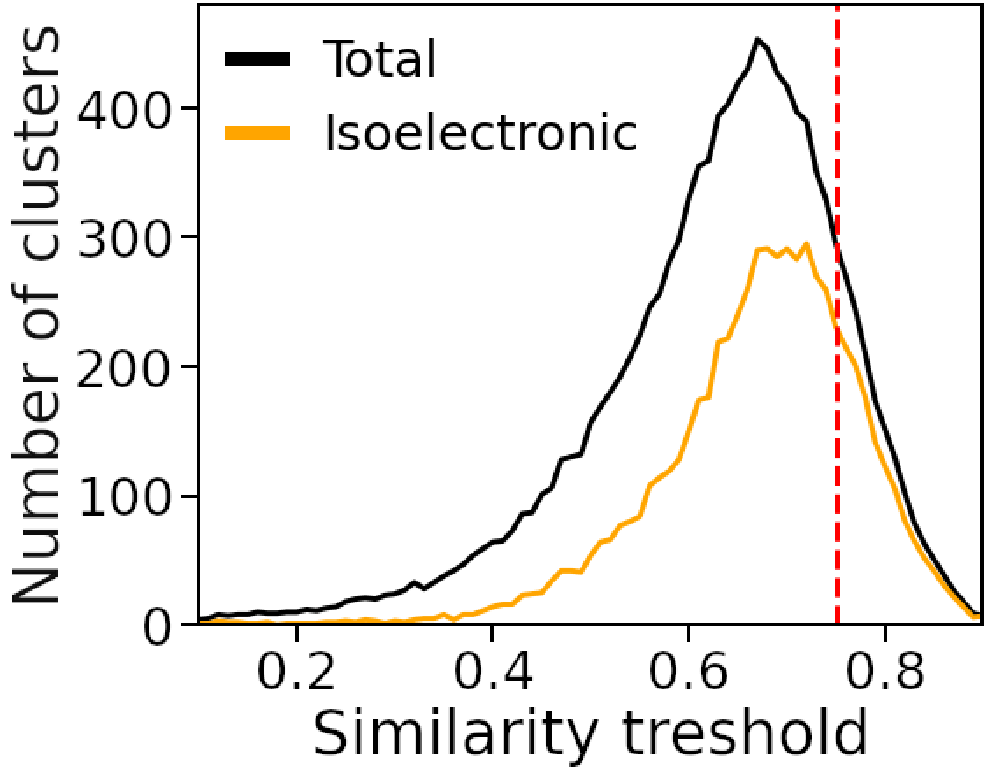
M. Kuban *et al.* Sci Data **9**, 646 (2022).

Isoelectronic clusters

Descriptor for isoelectronic clusters:

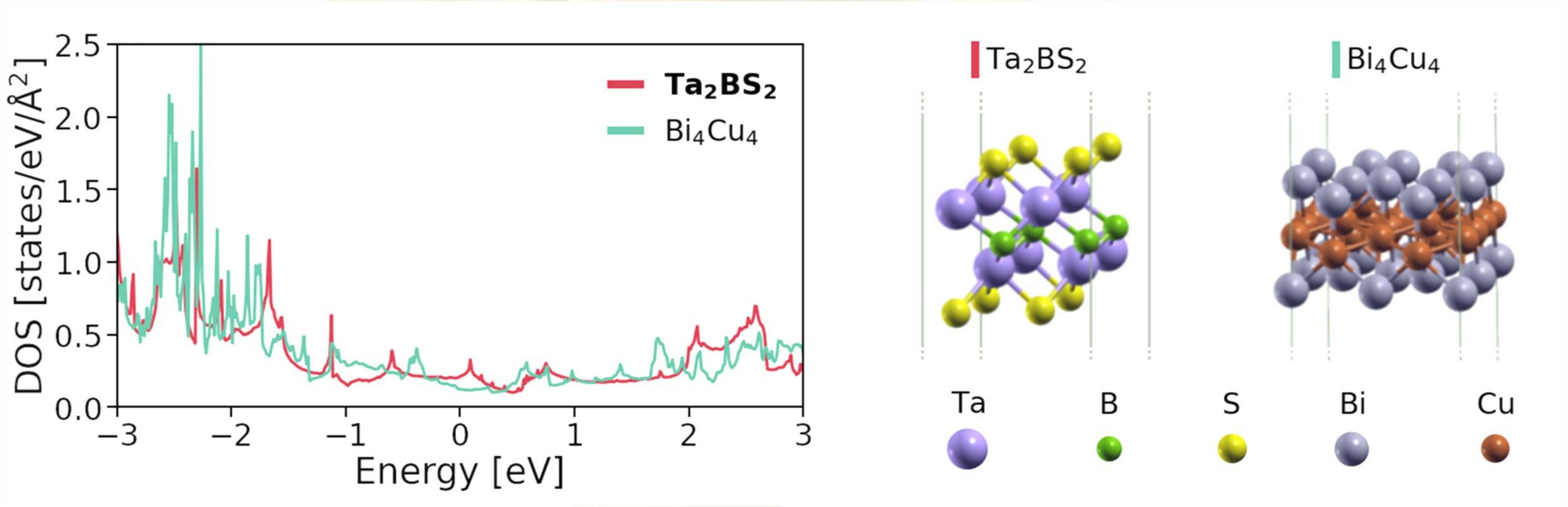
$$\bar{c}_m = \frac{1}{N_{Atoms}} \sum_{i=1}^{N_{Atoms}} c_i$$

230 isoelectronic clusters (78%)



M. Kuban *et al.* Sci Data **9**, 646 (2022).

Outliers



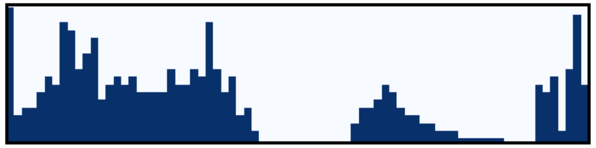
M. Kuban *et al.* Sci Data **9**, 646 (2022).

Conclusions

Spectral fingerprint to quantitatively evaluate the similarity of spectra

Quality assessment:

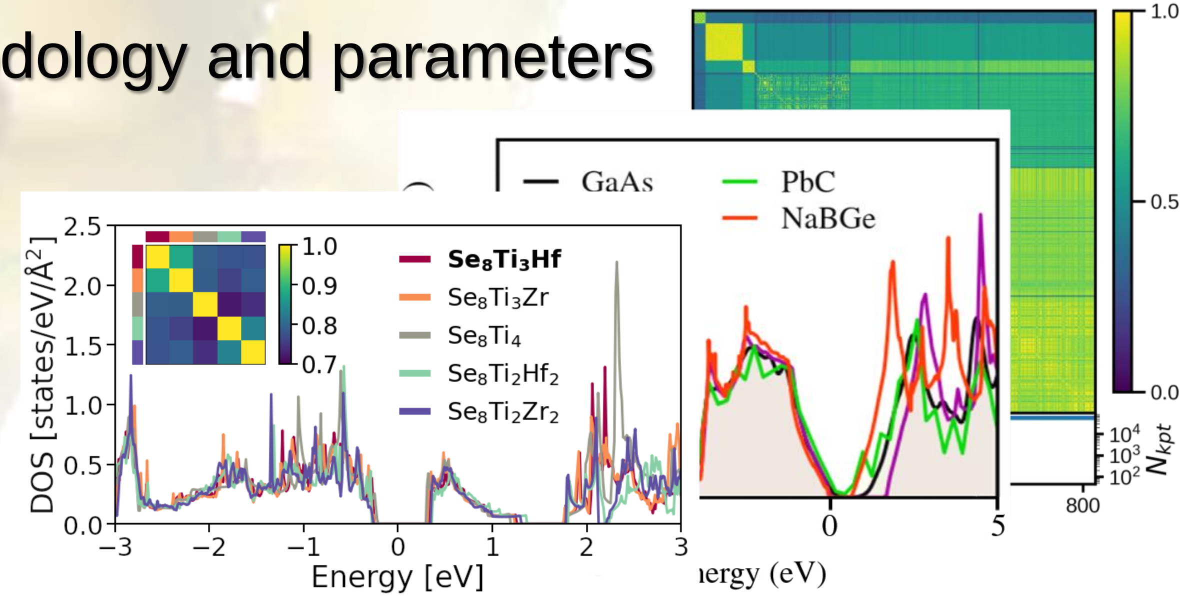
Measuring the impact of methodology and parameters



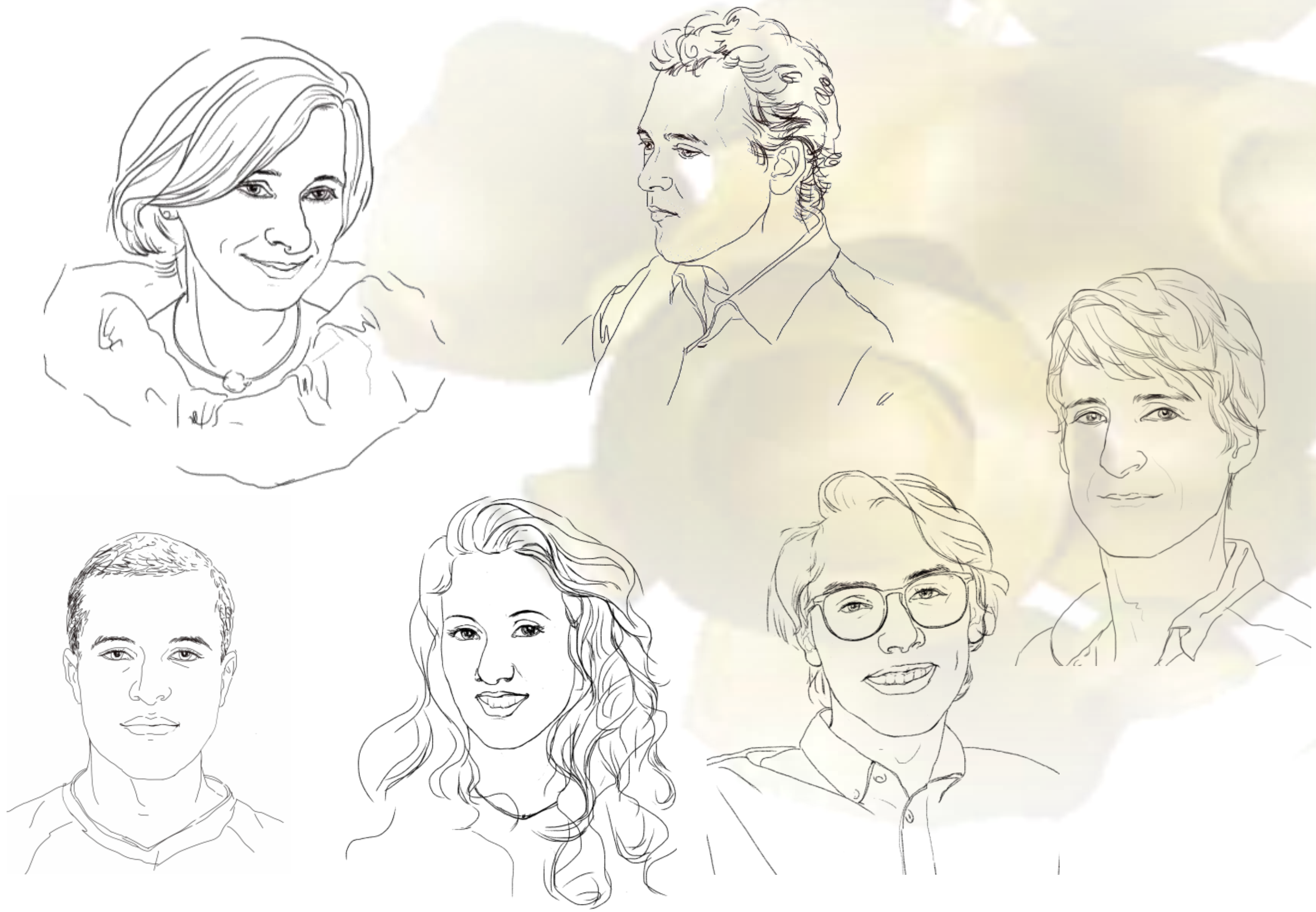
Data analytics:

Finding similar materials

Unsupervised learning



Acknowledgements



FONDA

FAIRmat

NOMAD

NOVEL MATERIALS DISCOVERY

NOMAD

Further reading

Kuban, M., Rigamonti, S., Scheidgen, M., Draxl, C., Density-of-states similarity descriptor for unsupervised learning from materials data. *Sci Data* **9**, 646 (2022)

Kuban, M., Gabaj, Š., Aggoune, W. *et al.* Similarity of materials and data-quality assessment by fingerprinting. *MRS Bulletin* **47**, 991–999 (2022).

NOMAD AI Toolkit Tutorial:

`https://nomad-lab.eu/aitoolkit/tutorial-dos-similarity`