



Similarity of materials and data-quality assessment by fingerprinting

Martin Kuban, Šimon Gabaj, Wahib Aggoune, Cecilia Vona, Santiago Rigamonti, and Claudia Draxl*^{ID}

Identifying similar materials (i.e., those sharing a certain property or feature) requires interoperable data of high quality. It also requires means to measure similarity. We demonstrate how a spectral fingerprint as a descriptor, combined with a similarity metric, can be used for establishing quantitative relationships between materials data, thereby serving multiple purposes. This concerns, for instance, the identification of materials exhibiting electronic properties similar to a chosen one. The same approach can be used for assessing uncertainty in data that potentially come from different sources. Selected examples show how to quantify differences between measured optical spectra or the impact of methodology and computational parameters on calculated properties, like the density of states or excitonic spectra. Moreover, combining the same fingerprint with a clustering approach allows us to explore materials spaces in view of finding (un)expected trends or patterns. In all cases, we provide physical reasoning behind the findings of the automatized assessment of data.

Introduction

Finding materials for a specific application may be a long and tedious process. Typical time spans from the scientific invention to the market are 20 years, or even longer.¹ Thus, there is urgent demand for speeding up the process to identify candidate materials that exhibit desired properties. This is particularly so in view of the enormous challenges arising from the world's tremendously increasing energy consumption and environmental problems. Overall, there is hardly any part of our society that is not concerned with materials, and the materials science community is dealing with a wide variety of materials classes, their properties, and functions.

For identifying materials with desired properties, data-centric approaches² are

attracting a lot of attention. For being successful on a large scale, a fundamental requirement is to make use of the research data from the entire community. Indeed, more and more publicly available data collections are being established, and efforts toward “FAIRification”³ are growing worldwide. This raises in particular the question of interoperability, the “I” in FAIR. Bringing together data from various sources implies heterogeneity. In other words, variety and veracity, two of the 4V of Big Data are becoming an issue. Therefore, it is crucial to assess and control the uncertainty in data.⁴ On the experimental side, variety concerns different methods to measure a physical property, and—within any such method—diverse instruments

Martin Kuban, Institut für Physik and IRIS Adlershof, Humboldt-Universität zu Berlin, Berlin, Germany
 Šimon Gabaj, Institut für Physik and IRIS Adlershof, Humboldt-Universität zu Berlin, Berlin, Germany
 Wahib Aggoune, Institut für Physik and IRIS Adlershof, Humboldt-Universität zu Berlin, Berlin, Germany
 Cecilia Vona, Institut für Physik and IRIS Adlershof, Humboldt-Universität zu Berlin, Berlin, Germany
 Santiago Rigamonti, Institut für Physik and IRIS Adlershof, Humboldt-Universität zu Berlin, Berlin, Germany
 Claudia Draxl, Institut für Physik and IRIS Adlershof, Humboldt-Universität zu Berlin, Berlin, Germany; claudia.draxl@physik.hu-berlin.de
 *Corresponding author
 doi:10.1557/s43577-022-00339-w

Impact statement

To predict novel materials with desired properties, data-centric approaches are in the process of becoming an additional fundament of materials research. Prerequisite for their success are well-curated data. Ideally, one can make use of multiple data collections. Bringing data from different sources together, poses challenges on their interoperability which are routed in two out of the 4V of Big Data. These are the uncertainty of data quality (veracity) and the heterogeneity in form and meaning of the data (variety). To overcome this barrier, universal and interpretable measures must be established, which quantify differences between data that are supposed to have the same meaning. Here, we show how a spectral fingerprint in combination with a similarity metric can be used for assessing spectral properties of materials. Our approach allows for tracing back in computed as well as measured data, differences stemming from various aspects. It thus paves the way for automatized data-quality assessment toward interoperability. Based on this, in turn, materials exhibiting similar features can be identified.



together with a possible selection of measurement modes. Likewise, computing a specific property can be done by different methodologies and possible approximations, utilizing different software packages. Uncertainty in computed data, in turn, is related to algorithms, implementation, and computational parameters. These correspond to different resolution and measurement conditions (e.g., pressure, temperature, environment) on the experimental side. The latter is also concerned with the quality and growth condition or treatment of the materials sample. All of these uncertainties need to be understood, and ideally be quantified, to allow for unrestricted interoperability. Obviously, this is an overwhelming task. It also illustrates the urgent need for benchmark data to quantify deviations from the ideal results. For the computational materials side, the topics of reproducibility⁵ and benchmarks for solids^{6–9} are being pursued only in the last few years.

For the identification of materials with specific features, similarity is an important concept. Materials of interest for a given application should share specific properties (i.e., being similar in some aspects while they may be very different in others). To assess similarity in general and, in view of data-centric approaches in particular, one needs to introduce adequate descriptors and similarity measures^{10–15} that go beyond that of the atomic structure.¹⁶

In this article, we show how both aspects—identifying similar materials and quantifying uncertainties—can be addressed with methods and tools measuring similarity. Specifically, we assess similarity in electronic properties in terms of a spectral fingerprint. We demonstrate our approach by various scenarios such as identifying materials that are similar to a chosen one, measuring the impact of structural features as well as methodology or computational parameters on the accuracy and precision of computed properties, or highlighting differences in sample quality or measurement details on experimental results. The same descriptor can also be combined with unsupervised learning to identify and analyze trends in large data sets as demonstrated recently.¹⁵ In all of our examples, we provide physical reasoning behind the data-based observations. Finally, we discuss how this approach can be used to enhance large-scale data collections.

Results

The following examples focus on exploring and understanding data spaces on the one hand, and on highlighting effects that could potentially lead to veracity on the other hand. We emphasize that all calculations shown here are perfectly valid, but may differ in certain numerical aspects as each of them has been created for a specific purpose. The examples are neither chosen such to represent the best possible calculations nor to showcase inconsistencies, but only to highlight the respective scenario that we wish to demonstrate. They help to illustrate where differences between data that should mean or describe

the same, may come from and how these differences can be quantified. This quantification is carried out by combining the spectral fingerprint with the Tanimoto coefficient (T_c) as a similarity measure,¹⁷ as described in the “Methods” section. In short, T_c varies between 0 (completely different) to 1 (identical).

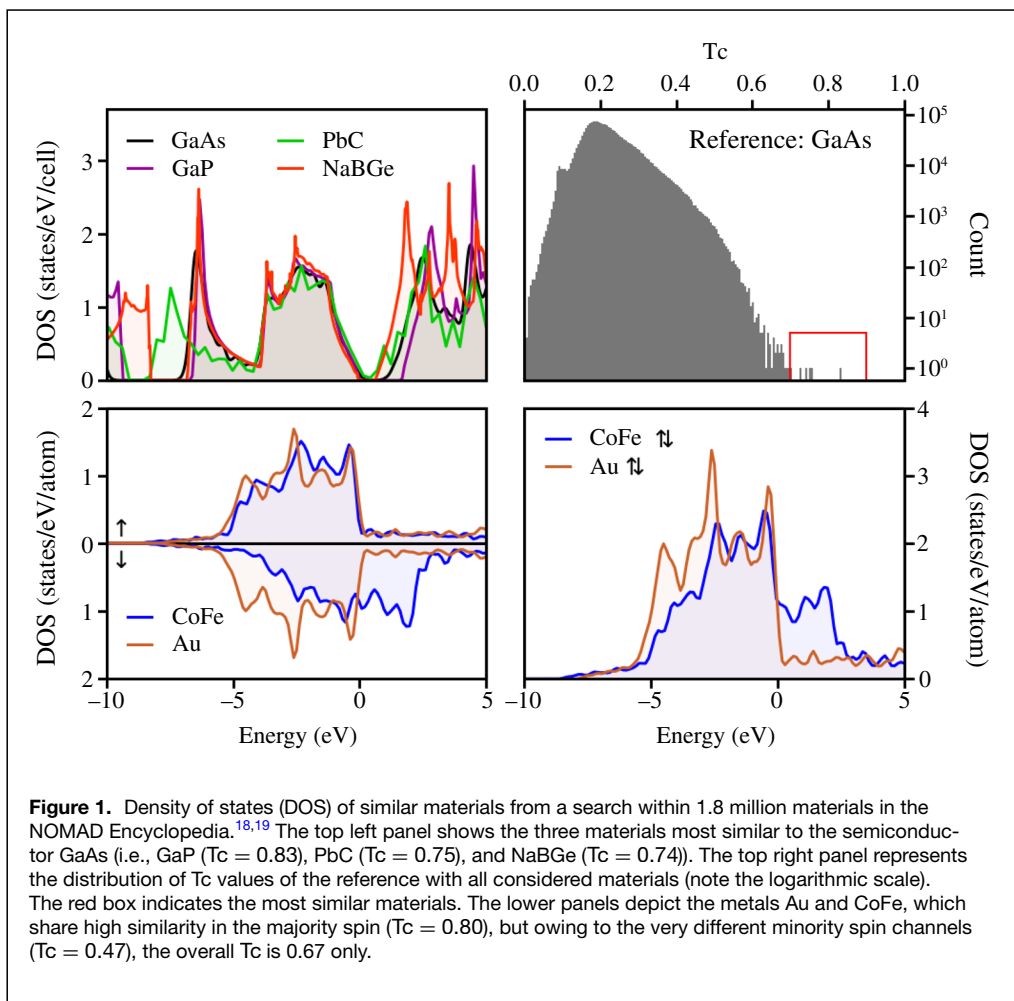
Identifying similar materials

Our first case is about finding materials that share certain characteristics with a selected reference. Replacing, for instance, toxic or rare elements or components with less harmful or abundant ones, is one important aspect in the search of new materials. Here, we demonstrate how to find materials that share features of their electronic structure. Using the spectral fingerprint (see “Methods” section), we have searched about 1.8 million materials in the NOMAD Encyclopedia (corresponding to about 93% of the materials available today) in order to find materials that have a similar electronic density of states (DOS). **Figure 1** presents two selected results of this investigation. The top left panel shows the most similar materials to the semiconductor GaAs. We observe that all four compounds share the higher-lying valence band structure and differ either in the lower valence region or the conduction bands. As could be expected from the valence configurations of atoms from the same column of the periodic table of elements (PTE), GaP—also sharing the same crystal structure—is a good candidate that indeed, turns out to be the material most similar to GaAs ($T_c = 0.83$). Maybe less expected are PbC ($T_c = 0.75$) and NaBGe ($T_c = 0.74$). To obtain a better understanding about what the observed T_c values mean, we provide a histogram (top right) that shows the distribution of Tanimoto coefficients of the reference DOS with that of all 1.8 million materials. Strikingly, materials with high similarity coefficients are exceptional. The red box indicates those with $T_c > 0.7$. We note, however, that the distribution depends on the considered energy range (here from -10 to 5 eV). Choosing a narrower energy window to focus on specific features of the spectra may lead to a larger number of materials that are similar to the reference.

The lower panels of **Figure 1** demonstrate how (dis)similar the two metals Au and CoFe are. Here, the spin properties govern the behavior. The majority spin exhibits a high similarity of $T_c = 0.80$, where the DOS reflects the akin character of the occupied $5d$ Au and $3d$ bands of Cu, respectively. In contrast, the minority spins ($T_c = 0.47$) differ mainly due to a rigid shift of the partially filled Cu $3d$ minority band by about 2.5 eV such that the overall similarity is only moderate ($T_c = 0.67$).

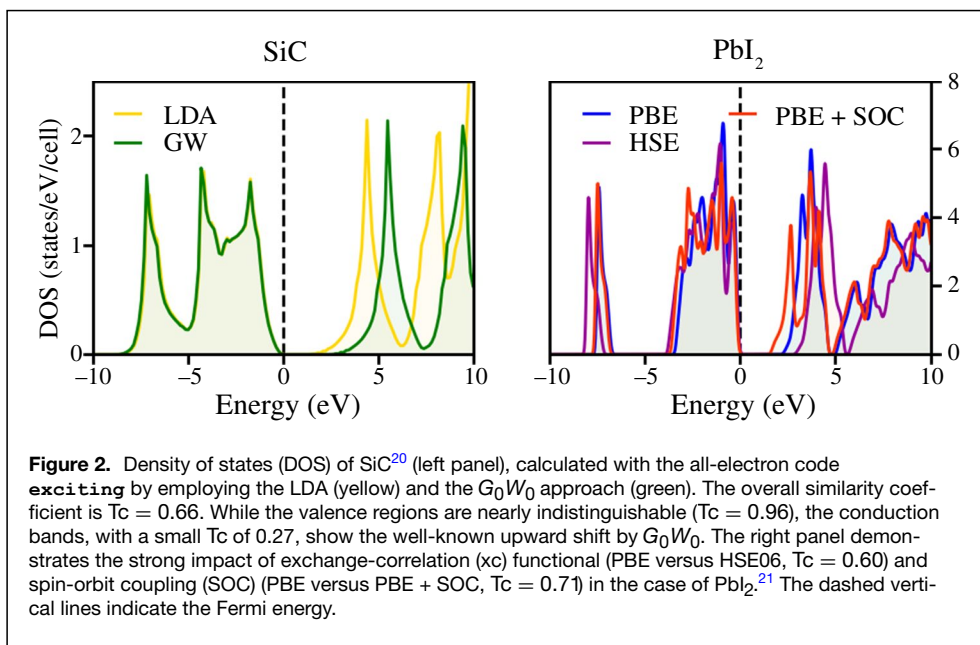
Impact of methodology

Figure 2 compares the density of states of SiC,²⁰ calculated by the local-density approximation (LDA) and the G_0W_0 approximation of many-body perturbation theory. Both



results are obtained with the `exciting` code,²² using the same set of computational parameters. Hence, the difference in the two results can be solely assigned to the method. Only assessing the valence-band region, gives a similarity coefficient of Tc = 0.96. In contrast, the conduction region reveals the well-known opening of the electronic bandgap by G_0W_0 , leading to a small similarity coefficient of 0.27 only. The overall similarity covering the entire energy range is moderate (i.e., Tc = 0.66).

The panel on the right draws a similar picture for PbI_2 . Here, we explore on the one hand how accounting for spin-orbit coupling (SOC) changes the results. In essence, by decreasing the bandgap by as much as 0.68 eV, SOC showing up in a modest Tanimoto coefficient of Tc = 0.71 when considering the entire energy range of -10 to 10 eV. On the other hand, going from PBE to HSE06, the gap opens by 0.69 eV, blueshifting the DOS by this amount and leading, with PBE as the counterpart, to Tc = 0.60. Overall, the competition of the two effects, giving rise to cancellation of errors, explains why the Kohn–Sham gap obtained by PBE without SOC is rather close to experiment.^{23–25} Considering the valence region only, both SOC and HSE have a significant impact, giving rise to Tc values of 0.75 (PBE versus PBE + SOC) and 0.73



(PBE versus HSE). The conduction region is characterized by an upward shift of all bands by HSE ($T_c = 0.45$ for PBE versus HSE) while SOC only affects the bands up to 5 eV ($T_c = 0.67$ for PBE versus PBE + SOC), thus changing the shape of the DOS.

Obviously, these examples confirm in an automated way what we usually observe by visual inspection. At the same time, it illustrates a workflow that can be applied to large data sets, when visual inspection becomes unfeasible. Given very many calculations for many materials, our fingerprint will eventually allow us to extract knowledge that we cannot obtain from individual calculations or publications. The longer-term goal here is to verify such observations on large data sets, and provide and implement automated tools to simplify the here presented analysis. This will enable us to detect, on a large scale, for which material class what level of methodology is needed to provide a reliable result. In other words, we will learn what one can expect from the performance of a method for a given material or property, and what we can recommend to a novice user. In the following examples, we will take the idea

further to also include computational parameters in these considerations.

Excitonic spectra

In **Figure 3**, we display how the optical spectra of h-BN in A'A stacking evolve as a function of the k -grid used in the calculations. The spectra are characterized by an absorption onset (solid vertical lines) far below the bandgap (dashed line), indicative of strong excitonic effects.²⁶ The excitonic binding energy decreases with improved sampling (i.e., the onset experiencing a blueshift). This seems in contradiction to the corresponding exciton wave functions displayed on the right that one would expect to become more delocalized. The latter is indeed the case. Looking, for instance, at the behavior in c direction, the reason for having the same electron distribution in every second plane is simply explained by the fact that the unit cell contains two such planes and, restricting ourselves to 1 k -point in this direction, we obtain replica of the same wave function in every other plane. In other words, the exciton is

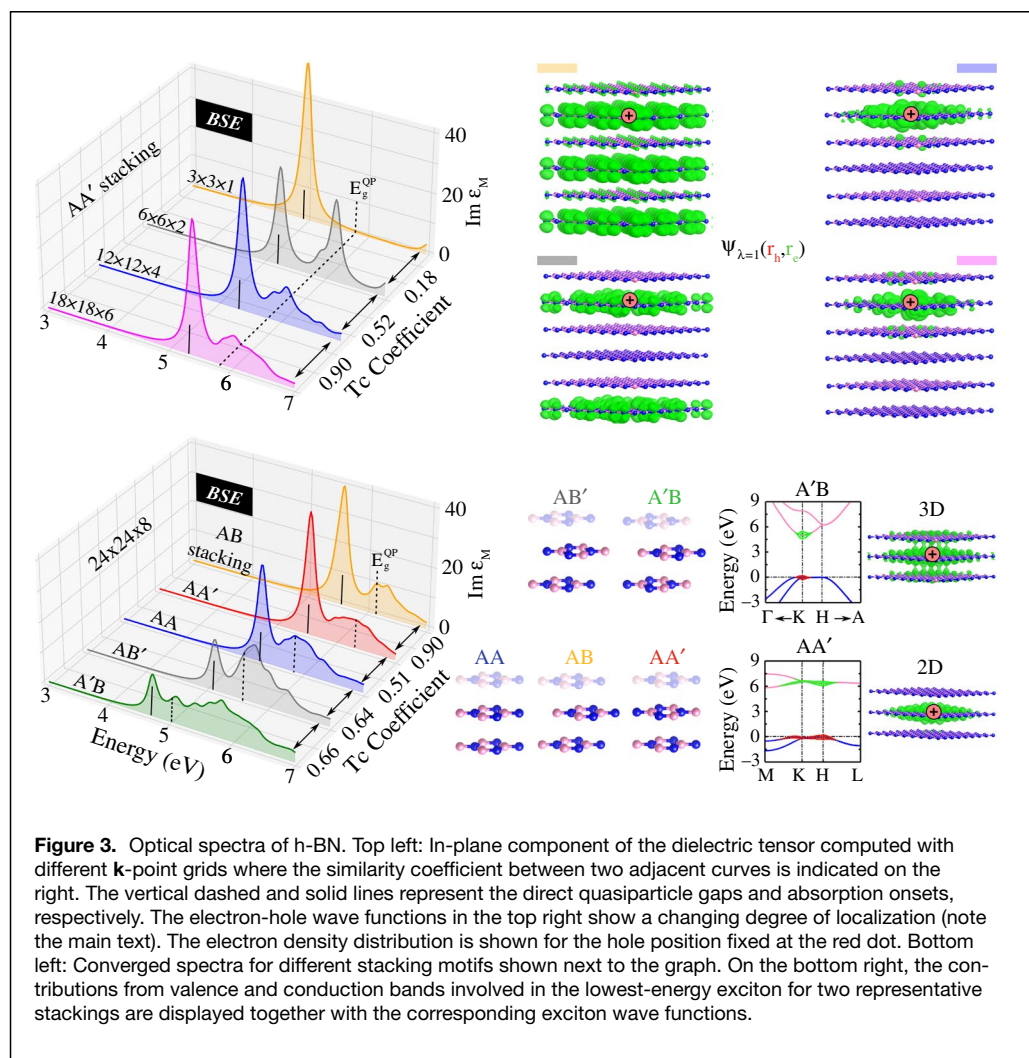


Figure 3. Optical spectra of h-BN. Top left: In-plane component of the dielectric tensor computed with different k -point grids where the similarity coefficient between two adjacent curves is indicated on the right. The vertical dashed and solid lines represent the direct quasiparticle gaps and absorption onsets, respectively. The electron-hole wave functions in the top right show a changing degree of localization (note the main text). The electron density distribution is shown for the hole position fixed at the red dot. Bottom left: Converged spectra for different stacking motifs shown next to the graph. On the bottom right, the contributions from valence and conduction bands involved in the lowest-energy exciton for two representative stackings are displayed together with the corresponding exciton wave functions.

spatially *not* much delocalized. The same applies to the in-plane directions where we also observe a seeming delocalization. Only by considering a denser k -grid, the localized character becomes apparent.

Making use of the fingerprinting in the exact same manner as for the DOS, one can measure the convergence behavior. The similarity coefficient between the $3 \times 3 \times 1$ and $6 \times 6 \times 2$ results is 0.18, increasing to 0.52 between $6 \times 6 \times 2$, and $12 \times 12 \times 4$ and to 0.90 between $12 \times 12 \times 4$ and $18 \times 18 \times 6$. The latter two spectra are spot on in the onset region and only slightly differ at higher energies, causing the deviations from the optimal T_c value (ideally being 1) as it is measured over the entire energy range. The gradual increase in



similarity indeed reflects the behavior of the exciton wave function. On the longer run, based on a large-enough data pool, we expect to learn optimal parameter settings for a given material class such to provide recommendations to users who want to carry out such computationally expensive calculations with high numerical precision at least possible effort.

The lower half of Figure 3 shows the impact of structural features on the optical spectra of h-BN. On the left, the BSE spectra of five different stackings are shown that are displayed next to the figure. From a first inspection, we can discriminate between two groups: (1) The spectra of the AB, AA' and AA stackings have similar shape and intensity while (2) those of AB' and A'B exhibit lower-energy excitations with smaller oscillator strength. As evident from the exemplary exciton wave function on the right, the intense peak at the absorption onset of the first group is characterized by a localized exciton with 2D distribution. In these cases, the degeneracy of the valence- and conduction-band edges is governed by how the B and N atoms are aligned in the out-of-plane direction.²⁶ The resulting character of the initial and final states of the excitation leads to its 2D distribution.²⁶ The similarity coefficient between AB and AA' results is 0.9 and only 0.51 between AA' and AA. Despite the similar shape of the spectra, the absorption onset of the latter is lower in energy due to its smaller quasiparticle gap (dashed lines). The moderate similarity coefficients between the spectra of the first and the second group reflects the differences in their band structure. The valence bands (conduction bands) in the AB' (A'B) stackings are split, which is a consequence of the vertical alignment of atoms of the same species. This impacts the exciton distribution which is of 3D character in these cases and exhibit lower binding energy compared to the first group. Overall, this example shows the direct relationship between structural arrangement, band structure, absorption spectrum, and exciton character. The similarity coefficients are able to capture (or measure) the main differences between various results. On the longer term, based on the analysis of many spectra, one may be able to provide recommendations which configurations may generate bound excitons with localized/delocalized character.

Interoperability of experimental spectra

Interoperability of experimental data is even a bigger issue compared to computational results. For instance, one and the same physical property can be measured by different methods, giving rise to variety in the data. The dielectric function is a good example for this, as it can be obtained by several experimental probes, being optical absorption or reflection spectroscopy, ellipsometry, as well as electron-loss spectroscopy. None of the measurements yields this property directly but only after some transformation steps or modeling behind, which adds a veracity problem. Overall, as pointed out in the Introduction, a measurement may depend on a huge number of parameters that need to be captured by metadata.

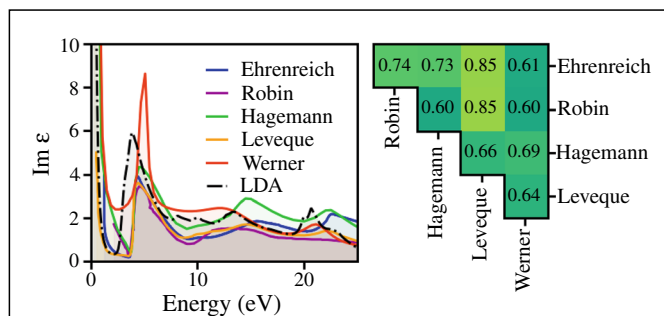


Figure 4. Optical spectra of the elemental solid silver from different sources and measurements.^{27–31} The background of the similarity matrix on the right shows the same color code as Figure 5. Due to symmetry, only the upper off-diagonal elements of the matrix are shown. For comparison, a calculation within the independent-particle approximation based on LDA is shown. Its Tc value with the spectra from Werner et al. is 0.78; with that of Robin, it is 0.68 only.

So, we may ask whether it is surprising or not that the optical spectra of silver displayed in Figure 4 exhibit so low similarity among each other. In the ideal world, one would expect spectra of an elemental, noble metal to be spot on. Reality tells us that this is not the case. While the measurements displayed here all agree on the absorption onset of inter-band transitions, the peak positions at higher energies differ up to several eV, and the oscillator strengths are within roughly a factor of two. Without further information, it is impossible to tell which of the measurements may be superior to others and why. Obviously, both sample quality as well as measurement conditions have a strong impact, and a full annotation of the measured data with metadata are key for a fair comparison of experimental data.

Veracity is also an issue when comparing measured and calculated data. It is needless to say that the dielectric function can be computed with various theoretical methods of different levels of sophistication. Here, we show an LDA result within the independent-particle approximation.³¹ Measuring its similarity to the experimental data of Robin²⁸ and Werner et al.,³¹ delivers Tc values of 0.63 and 0.78, respectively. The better agreement with the latter—the most recent of the experimental data sets—has been discussed elsewhere.³¹ We note that, to a large extent, the disagreement comes from the underestimated absorption onset by LDA, owing to the too high position of the Ag-*d* bands.

This example shows how heterogeneous experimental data of even very simple and well understood materials can be. This situation imposes the need for detailed annotation of measured data, and it also highlights the urgent need for a large variety of benchmark results that serve as reference data. Otherwise, an in-depth assessment of data quality on a large scale will remain difficult.

Influence of code and computational parameters

To reveal the effects arising from computational parameters and tools, we study fcc Al. For this material, we find more than 800 calculations in the NOMAD Repository.^{18,19} They have been carried out by different people with different codes and computational settings, depending on the individual purposes. Making use of the similarity coefficient, we can gather useful information about the impact of these parameters on the DOS. Again, in the ideal world, all calculations based on the same geometry should be identical. **Figure 5** displays the corresponding similarity matrix that shows that this is obviously not the case. On the left, the matrix is sorted according to the combination of DFT code and xc functional. Two clusters become apparent, basically distinguishing between two different codes. The one stemming from FHI-aims³² calculations contain both LDA and GGA results. Their similarity suggests that for fcc Al, the choice of the semi-local functional has no notable effect on the DOS. In contrast, there is overall less similarity between the VASP³³ and FHI-aims blocks. This behavior is rather unexpected and could be assigned to differences in the calculation of the DOS itself rather than in the electronic structure. Note that the pattern inside each block reflects all other differences in the calculations, such as volume, **k**-grid, computational parameters, etc.

The right panel of the figure shows the same data, now sorted by the total number of sampled **k**-points, N_{kpt} . We see that the first about 350 entries are characterized by low similarity coefficients with all other data, with the exception of two smaller blocks that are formed by pure VASP and FHI-aims calculations, respectively. Above this index, the DOS of all calculations are very similar to each other, forming a significant cluster. We conclude that at this threshold, convergence

with respect to N_{kpt} is reached. Similar to the other example, here the patterns inside the bigger blocks stem from differences other than the **k**-grid. This explains the dark lines inside the block and the two features previously mentioned.

To conclude from this example, our framework quantifies and visualizes the effect of computational parameters on the DOS, which allows us to identify outliers and helps in choosing representative parameters. We note in passing that the data in the NOMAD Repository are normalized (i.e., brought to the same file formats and units, and having the same energy zero ($E_F = 0$ eV)). Nevertheless, they still reflect individual implementations. For example, applying additional smearing to the here presented data, enhances their similarity.

Finding patterns in data

Applied to well-curated data sets, the spectral fingerprint can be used to explore materials spaces in an unbiased manner when being combined with unsupervised learning. Adopting a clustering approach for the learning task, this has been recently demonstrated¹⁵ using data from the C2DB database.¹³ Here, we provide another example of this kind to round off the potential of our descriptor and demonstrate how compact sets of materials can be found that are more similar to each other than a defined similarity threshold. The selected cluster was identified from a data set consisting of 3491 materials. It comprises three materials, Hf_2Te_6 , Zr_2Te_6 , and Zr_2NSe_2 . Their DOS and unit cells are depicted in **Figure 6**. Inspecting their crystal structures and electronic configurations, we see that this cluster exhibits expected as well as unexpected trends. The cluster center (i.e., the material most similar to both others) is Hf_2Te_6 . On the one hand, the latter and Zr_2Te_6 share the crystal symmetry (space group 59); also, Hf and Zr are isoelectronic. Thus, their high similarity of $T_c = 0.85$ in terms of the electronic structure could be anticipated. On the other hand, Zr_2NSe_2 neither shares one nor the other characteristics with the other cluster members, nevertheless exhibits high similarity values of $T_c = 0.79$ and $T_c = 0.78$ with Hf_2Te_6 and Zr_2Te_6 , respectively. Thus, it can be considered as an outlier. Outliers are particularly interesting in view of finding candidate materials with specific features in materials spaces where

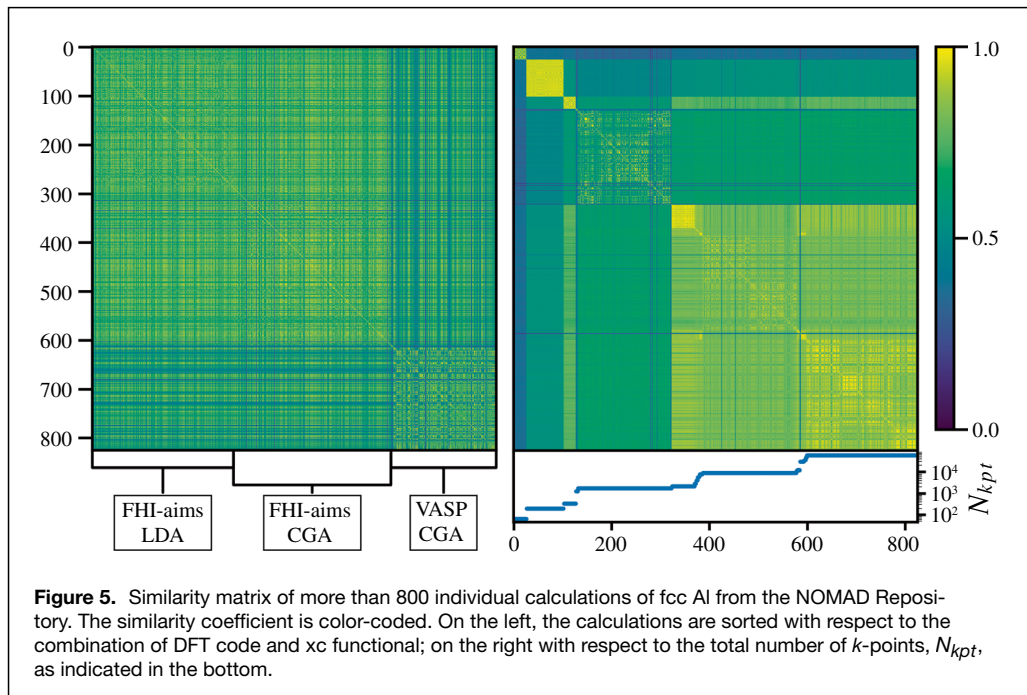
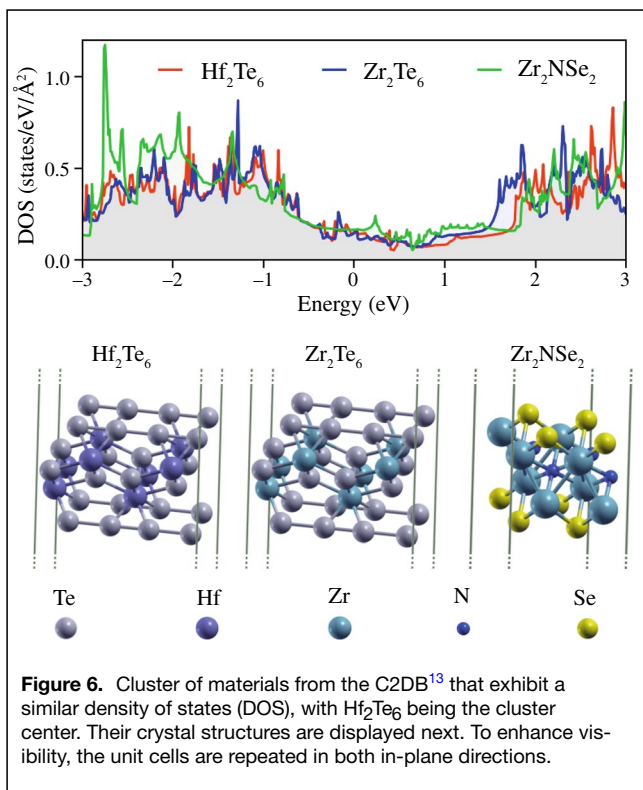


Figure 5. Similarity matrix of more than 800 individual calculations of fcc Al from the NOMAD Repository. The similarity coefficient is color-coded. On the left, the calculations are sorted with respect to the combination of DFT code and xc functional; on the right with respect to the total number of *k*-points, N_{kpt} , as indicated in the bottom.



one would not expect them. In fact, one of the great hopes is that data-driven analysis can guide researchers toward possibly interesting materials classes.

Conclusions and outlook

We have shown how a spectral fingerprint can be used in various conditions to measure the similarity of materials in terms of their electronic and spectroscopic properties. Our examples comprise a broad range of applications. We have, for instance, identified materials that share the electronic properties of their valence electrons or the DOS of one spin-component. Moreover, we could demonstrate that the same approach allows for assessing data quality. More specific, with the examples of SiC and PbI_2 , it was illustrated how different treatment of exchange-correlation effects affect the electronic properties of semiconducting materials. More than 800 calculations of fcc Al were used to demonstrate the impact of computational parameters on the numerical precision of computed properties. Likewise, we probed the effect of structural features on the excitonic spectra in terms of the layer stacking in h-BN. All these examples not only allow us to determine the degree of interoperability of materials data but also help us to quantify more generally, how approximations, implementation details, computational parameters, or structural features affect calculations that are routinely carried out in numerous research groups all over the world. Automatizing such assessments would greatly enhance the understanding of computed results as they help in analyzing many more materials calculations

than what can be done in case-by-case investigations and single publications. The resulting findings will pave the way to incorporate physical reasoning into materials science data and annotate the data accordingly.

With the example of the noble metal silver, we also confronted various measured optical spectra with each other. Finally, we have shown how the spectral fingerprint can be combined with unsupervised learning to explore large data spaces and identify clusters of similar materials.

What are the next steps for using and/or enhancing our tools? There are several open issues, and we address a few of them:

- Regarding the search for similar materials, one question concerns the choice of the reference calculation. What is a representative calculation? Ideally, one would choose (the) one carried out with the highest-level methodology, ensuring good convergence with respect to the computational parameters. Obviously, then also the calculations to compare with should fulfill the same criteria which in most cases would restrict the data space significantly. Thus, a milder criterion would be to use lower-level calculations as long as all calculations are from the same level of sophistication. To judge the impact of the method and parameters, in turn, the tools presented in sections “[Impact of methodology](#),” “[Excitonic spectra](#),” or “[Influence of code and computational parameters](#)” can guide the choice.
- Obviously (see also the section “[Identifying similar materials](#)”), the similarity coefficient depends on where we place the focus on. Thus, having a certain function in mind, one could focus on certain energy regions or features. They might be as different as the band edges in semiconductors, the absorption onset in optical spectra, or the DOS at the Fermi level in metals or superconductors. The construction of the fingerprint is flexible and prepared for all of this.
- Methodology-wise, we plan to consider additional similarity metrics. Let us exemplify the need with the example of excitons. Weakly bound excitons are characterized by a redistribution of oscillator strength when compared to the independent-particle spectrum. This effect is well captured by the metric presented here. On the other hand, the rigid redshift of a strongly bound exciton is better captured by the Earth Mover’s Distance,³⁴ also known as Wasserstein Distance. Therefore, a more complex metric is required that allows one to capture both types of excitonic effects or tell them apart.

To conclude, the here presented examples demonstrate that the spectral fingerprint can serve as a descriptor for a variety of scenarios. Besides the previously described enhancements, for the future, we propose and plan to define similar criteria for other physical properties toward data analysis and quality assurance of computed and measured results. They could, for instance, bring to light differences in sample preparation,

composition, dimensionality or other structural features, measurement conditions, and more. On the other hand, exploring interoperable data by unsupervised learning will help in finding either anticipated or unexpected trends in the data and may enable discoveries.

Methods

In order to identify materials that share favorable properties in a large number of materials, a numeric representation (i.e., a descriptor) of the investigated property is required. In this work, we focus on energy-dependent spectra that are here represented by a spectral fingerprint, analogous to the DOS fingerprint of Reference 15. It consists of a 2D raster image obtained through a special discretization of a spectrum, $s(E)$: First, the spectrum is converted to a histogram, s_i , with energy bins of variable width

$$\Delta\varepsilon_i = n(\varepsilon_i, W, N) \Delta\varepsilon_{\min}, \quad 1$$

with s_i being the integrated spectrum in the interval $[E_i, E_{i+1})$, with $E_i = E_{\text{ref}} + \varepsilon_i$, $E_{i+1} = E_i + \Delta\varepsilon_i$, and $\varepsilon_0 = 0$. The integer-valued function $n(\varepsilon_i, W, N)$ is equal to 1 for $\varepsilon_i = 0$ and approaches N monotonously for $|\varepsilon_i| > W$. Defined in this way, the integer $N \geq 1$ and the real number $W > 0$ allow for increased resolution of the fingerprint in a region of width W around E_{ref} . If $N = 1$, a uniform resolution is achieved. Finally, a grid of pixels is obtained by discretizing every column of the histogram in a grid of N_ρ intervals of height $n(\varepsilon_i, W_H, N_H) \Delta\rho_{\min}$. For the definition of the function n and additional details on the fingerprint, see Reference 15, where the spectral fingerprint is applied to the case of the electronic DOS.

To quantify the similarity between two spectra A and B , we make use of the Tanimoto coefficient:¹⁷

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

which, for dichotomous descriptors like the spectral fingerprint, is a metric that ranges from 0 (not similar) to 1 (identical). For an overview of other similarity measures (e.g., the Dice and cosine similarity), which are commonly applied in chemistry research, we refer to Reference 17. For our specific application, we note that the Tanimoto coefficient has several advantages. First, it is highly interpretable, as it can be understood as the intersection of the two DOS, divided by their union. Additionally, in contrast to, for example, the Dice coefficient, the Tanimoto coefficient obeys the triangle inequality, which is a favorable property for applications such as clustering. Furthermore, it is computationally cheap, as it can be calculated using only binary operations and bit counts, in contrast to, for example, the cosine similarity, that requires the calculation of a square root.

For the spectra shown in Figure 1, we use a nonuniform grid with $\Delta\varepsilon_{\min} = 0.05$, $\Delta\rho_{\min} \sim 0.001$, $N = 21$, $N_H = 11$, $E_{\text{ref}} = -2$ eV, $W = W_H = 7$ eV, $N_\rho = 256$, and a cutoff of -10 to 5 eV. For the spectra in Figure 2, we define the feature

region by setting $E_{\text{ref}} = 0$ eV, $W = W_H = 10$ eV, $N_\rho = 512$, and the cutoff to -10 to 10 eV. The fingerprints used to compare the optical spectra in Figure 3 employ uniform grids, that is, $N = 1$, with $\Delta\varepsilon_{\min} = 0.02$ eV, $\Delta\rho_{\min} = 3 \times 10^{-3}$, $N_\rho = 256$, and a cutoff of 4 to 8 eV. The fingerprints used for Figure 4 employ uniform grids with $N = 1$, $\Delta\varepsilon_{\min} = 0.02$ eV, $\Delta\rho_{\min} = 1 \times 10^{-3}$, $N_\rho = 256$, and a cutoff of 2.2 to 25 eV.

Acknowledgments

This work received funding by the German Research Foundation (DFG) through the CRC 1404 (FONDA), Project No. 414984028. Partial support from the NFDI consortium FAIRmat, DFG Project No. 460197019, and the European Union's Horizon 2020 research and innovation program under the Grant Agreement No. 951786 (NOMAD CoE) is appreciated. The PbI₂ results have been carried out within the DFG priority program SPP2196 "Perovskite Semiconductors," Project No. 424709454. We thank K. Glantschnig for providing the experimental spectra of Al, extracted from References 27–31.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Data availability

The data used in the section "Identifying similar materials" can be downloaded from the NOMAD Repository^{18,19} through the following links: GaAs,³⁵ GaP,³⁶ PbC,³⁷ NaBGe,³⁸ Au,³⁹ CoFe.⁴⁰ The data used in the section "Impact of methodology" are calculated using the **exciting** package²² and are available in the NOMAD Repository.^{18,19} Specifically, the calculation for SiC can be accessed via Reference 20 and the data set for PbI₂ via Reference 41. The data used in the section "Excitonic spectra" can be accessed in the NOMAD Repository^{18,19} through References 42 and 43. The calculations have been carried out with the **exciting** package²² by solving the Bethe–Salpeter equation of many-body perturbation theory (MBPT) in the linearized augmented plane wave + local-orbital (LAPW+lo) basis.^{44,45} For details of the calculations, we refer to Reference 26. For the experimental spectra analyzed in the section "Interoperability of experimental spectra," we refer to References 27–31. The data used in the section "Influence of code and computational parameters" are obtained from the NOMAD Repository.^{18,19} To access them, one has to query for calculations of materials that have been uploaded before Jan. 1, 2021, containing exclusively Al with space group number 225, and having a DOS. An example of a search query returning these data can be found in Reference 46. The data stem from Curtarolo et al.,^{46–49} Wolverton et al.,^{50,51} Bieniek,⁵² Carbogno and Bieniek,⁵³ Hofmann,⁵⁴ and Lejaeghere.⁵⁵ Additional calculations from Bieniek are not assigned a DOI. Our data analysis of the section "Finding patterns in data" is based on materials from the high-throughput



collection C2DB (Computational 2D Materials Database).^{13,56} The specific data shown here can be accessed via the links provided in References 57–59.

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- <https://mgj.gov>
- I. Tanaka, K. Rajan, C. Wolverton, *MRS Bull.* **43**(9), 659 (2018)
- M.D. Wilkinson, M. Dumontier, U.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. Bonino da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hoof, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Salter, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, *Sci. Data* **3**, 160018 (2016)
- M. Scheffler, M. Aeschlimann, M. Albrecht, T. Bereau, H.-J. Bungartz, C. Felser, M. Greiner, A. Groß, C.T. Koch, K. Kremer, W.E. Nagel, M. Scheidgen, C. Wöll, C. Draxl, *Nature* **604**(7907), 635 (2022)
- K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, V. Blum, D. Caliste, I.E. Castelli, S.J. Clark, A. Dal Corso, S. de Gironcoli, T. Deutsch, J.K. Dewhurst, I. Di Marco, C. Draxl, M. Dulak, O. Eriksson, J.A. Flores-Livas, K.F. Garrity, L. Genovese, P. Giannozzi, M. Giantomassi, S. Goedecker, X. Gonze, O. Grånäs, E.K.U. Gross, A. Gulans, F. Gygi, D.R. Hamann, P.J. Hasnip, N.A.W. Holzwarth, D. Iuşan, D.B. Jochym, F. Jollet, D. Jones, G. Kresse, K. Koepnik, E. Küçükbenli, Y.O. Kvashnin, I.L.M. Locht, S. Lubeck, M. Marsman, N. Marzari, U. Nitzsche, L. Nordström, T. Ozaki, L. Paulatto, C.J. Pickard, W. Poelmans, M.I.J. Probert, K. Refson, M. Richter, G.-M. Rignanese, S. Saha, M. Scheffler, M. Schlipf, K. Schwarz, S. Sharma, F. Tavazza, P. Thunström, A. Tkatchenko, M. Torrent, D. Vanderbilt, M.J. van Setten, V. Van Speybroeck, J.M. Walls, J.R. Yates, G.-X. Zhang, S. Cottenier, *Science* **351**(6280), aad3000 (2016)
- A. Gulans, A. Kozhevnikov, C. Draxl, *Phys. Rev. B* **97**, 161105 (2018)
- S.R. Jensen, S. Saha, J.A. Flores-Livas, W. Huhn, V. Blum, S. Goedecker, L. Frediani, *J. Phys. Chem. Lett.* **8**(7), 1449 (2017)
- D. Nabok, A. Gulans, C. Draxl, *Phys. Rev. B* **94**, 035118 (2016)
- T. Rangel, M. Del Ben, D. Varsano, G. Antonius, F. Bruneval, F.H. da Jornada, M.J. van Setten, O.K. Orhan, D.D. O'Regan, A. Canning, A. Ferretti, A. Marini, G.-M. Rignanese, J. Deslippe, S.G. Louie, J.B. Neaton, *Comput. Phys. Commun.* **255**, 107242 (2020)
- R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, *NPJ Comput. Mater.* **3**(1), 54 (2017)
- O. Isayev, D. Fourches, E.N. Muratov, C. Oses, K. Rasch, A. Tropsha, S. Curtarolo, *Chem. Mater.* **27**(3), 735 (2015)
- C.B. Mahmoud, A. Anelli, G. Csányi, M. Ceriotti, *Phys. Rev. B* **102**, 235130 (2020)
- M.N. Gjerding, A. Taghizadeh, A. Rasmussen, S. Ali, F. Bertoldo, T. Deilmann, N.R. Knøsgaard, M. Kruse, A.H. Larsen, S. Manti, U. Pedersen, T. Skovhus, M.K. Svendsen, J.J. Mortensen, T. Olsen, K.S. Thygesen, *2D Mater.* **8**(4), 044002 (2021)
- N.R. Knøsgaard, K.S. Thygesen, *Nat. Commun.* **13**(1), 468 (2022)
- M. Kuban, S. Rigamonti, M. Scheidgen, C. Draxl, Density-of-states similarity descriptor for unsupervised learning from materials data (2022). <https://arxiv.org/abs/2201.02187>
- S. De, A.P. Bartók, G. Csányi, M. Ceriotti, *Phys. Chem. Chem. Phys.* **18**, 13754 (2016)
- P. Willett, J.M. Barnard, G.M. Downs, *J. Chem. Inf. Comput. Sci.* **38**(6), 983 (1998)
- C. Draxl, M. Scheffler, *MRS Bull.* **43**(9), 676 (2018)
- C. Draxl, M. Scheffler, *J. Phys. Mater.* **2**(3), 036001 (2019)
- https://nomad-lab.eu/entry/id/kUqd_BDKGmTWhJcbvY_puqwA9vbi
- C. Vona, D. Nabok, C. Draxl, *Adv. Theory Simul.* **5**(1), 2100496 (2022)
- A. Gulans, S. Kontur, C. Meisenbichler, D. Nabok, P. Pavone, S. Rigamonti, S. Sagmeister, U. Werner, C. Draxl, *J. Phys. Condens. Matter* **26**(36), 363202 (2014)
- Ch. Gähwiler, G. Harbeke, *Phys. Rev.* **185**, 1141 (1969)
- R. Ahuja, H. Arwin, A.F. Da Silva, C. Persson, J.M. Osorio-Guillén, J.S. De Almeida, C.M. Araujo, E. Veje, N. Veissid, C.Y. An, I. Pepe, B. Johansson, *J. Appl. Phys.* **92**(7219), 12 (2002)
- C. Shen, G. Wang, *J. Phys. D Appl. Phys.* **51**(3), 035301 (2018)
- W. Aggoune, C. Cocchi, D. Nabok, K. Rezuouali, M.A. Belkhir, C. Draxl, *Phys. Rev. B* **97**, 241114 (2018)
- H. Ehrenreich, H.R. Philipp, *Phys. Rev.* **128**, 1622 (1962)
- S. Robin, "Propriétés optiques de l'argent et du palladium dans l'ultraviolet lointain," in *Optical Properties and Electronic Structure of Metals and Alloys*, F. Abelès, Ed. (North Holland Publishing, Amsterdam, 1966), p. 202
- H.J. Hagemann, W. Gudat, C. Kunz, *Phys. Rev. B* **65**, 742 (1975)
- G. Leveque, C.G. Olson, D.W. Lynch, *Phys. Rev. B* **27**, 4654 (1983)
- W.S.M. Werner, K. Glantschnig, C. Ambrosch-Draxl, *J. Phys. Chem. Ref. Data* **38**, 1013 (2009)
- V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, M. Scheffler, *Comput. Phys. Commun.* **180**(11), 2175 (2009)
- G. Kresse, J. Furthmüller, *Phys. Rev. B* **54**, 11169 (1996)
- Y. Rubner, C. Tomasi, L.J. Guibas, A Metric for Distributions with Applications to Image Databases, in *Proceedings of the Sixth International Conference on Computer Vision (ICCV)* (1998), pp. 59–66
- <https://nomad-lab.eu/entry/id/zkkMIAPyn40CbdEdW21DZTeretQ3>
- <https://nomad-lab.eu/entry/id/A9cSRMuSt7eF01tmJhBlx4dZ1c>
- https://nomad-lab.eu/entry/id/qLolniorAfKleyQniAudrk_GuKFF
- <https://nomad-lab.eu/entry/id/GtPHkCo0q8VDHmZyJ0BP7mAR1j>
- <https://nomad-lab.eu/entry/id/sYp4jkDZ1VwzFyc7WN0cl5RCFq1s>
- https://nomad-lab.eu/entry/id/m2EbbBqduN-MIEH7U3kaA00_62ic
- <https://doi.org/10.17172/NOMAD/2021.10.26-1>
- <https://doi.org/10.17172/NOMAD/2018.06.05-1>
- <https://doi.org/10.17172/NOMAD/2022.01.23-1>
- S. Sagmeister, C. Ambrosch-Draxl, *Phys. Chem. Chem. Phys.* **11**(22), 4451 (2009)
- C. Vorwerk, B. Aurich, C. Cocchi, C. Draxl, *Electron. Struct.* **1**(3), 037001 (2019)
- https://nomad-lab.eu/prod/rae/gui/search?visualization=elements&dfc.searchable_quantities=electronic_dos&dfc.spacegroup=225&until_time=2021-01-01T00%3A00%3A00.000Z&only_atoms=Al
- <http://afloplib.org>
- R.H. Taylor, F. Rose, C. Toher, O. Levy, K. Yang, M.B. Nardelli, S. Curtarolo, *Comput. Mater. Sci.* **93**, 178 (2014)
- S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G.L.W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, *Comput. Mater. Sci.* **58**, 227 (2012)
- <http://oqmd.org>
- J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, *JOM* **65**(11), 1501 (2013)
- <https://doi.org/10.17172/NOMAD/2017.01.24-1>
- <https://doi.org/10.17172/NOMAD/2020.07.27-1>
- <https://doi.org/10.17172/NOMAD/2020.07.29-1>
- <https://doi.org/10.17172/NOMAD/2016.10.19-1>
- S. Haastrop, M. Strange, M. Pandey, T. Deilmann, P.S. Schmidt, N.F. Hinsche, M.N. Gjerding, D. Torelli, P.M. Larsen, A.C. Riis-Jensen, J. Gath, K.W. Jacobsen, J.J. Mortensen, T. Olsen, K.S. Thygesen, *2D Mater.* **5**(4), 042002 (2018)
- <https://cmrdb.fysik.dtu.dk/c2db/row/HF2Te6-3f5c06f4bf18>
- <https://cmrdb.fysik.dtu.dk/c2db/row/Zr2Te6-8ef6448a7da4>
- <https://cmrdb.fysik.dtu.dk/c2db/row/NSe2Zr2-d567fba5f5ba>