Bachelorarbeit
Zur Erlangung des akademischen Grades Bachelor of Science (B.Sc.)

Thema der Arbeit:

Electronic density-of-states fingerprints for similarity analysis

eingereicht von:
Šimon Gabaj

Gutachter/innen: Prof. Claudia Draxl
Prof. Christoph T. Koch

Eingereicht am Institut für Physik der Humboldt-Universität zu Berlin am: 02. 07. 2021

Contents

1 Introduction

Designing new materials or finding novel applications of existing materials is a difficult task. Solving this task is, however, necessary, as new technological devices, such as solar cells or thermoelectric generators, require materials with specifically tailored properties. These materials should exhibit properties that are advantageous without drawbacks. For example, for energy harvesting, there are many promising candidate materials, which have excellent electronic and mechanical properties, they may however be toxic, such as CdTe, or unstable, such as halide perovskites. Also, cheap production is an important criterion. With rising complexity of technological solutions for problems of todays society, novel materials are required to enable future developments. To enable these developments, the design of materials should be efficient.

Experimental design of materials is constrained by high costs and long time scales, mostly due to challenges in synthesis and experimental characterisation. To accelerate this process, theoretical solid-state models and theories are employed to predict material properties. The work horse of materials science is density-functional-theory (DFT), which achieves reliable results [1] from first principles. DFT calculations however still require substantial computational resources to produce accurate results. The recent establishment of public databases for materials data reduces this cost, as scientists are able to share and re-use data published by others. Thus, the repetition of calculations, that have been performed previously, can be avoided, and calculations can be re-purposed for analysis that may not have been intended by the original authors of the data. There are multiple material databases readily available, e.g. AFLOW [2], The Materials Project [3], or the NOMAD Laboratory [4]. The NOMAD Laboratory provides the largest collection of DFT calculations, more than 100 milion individual calculation runs from more than 40 different DFT codes (specific implementations of the DFT formalism).

This data, coming from varying sources, is very inhomogeneous. It differs not only in author, but also in the purpose of the calculation, the DFT code used or the specific computational parameters. For example, a DFT calculation for elastic properties of a material may employ computational parameters that are not suitable for a reliable representation of the electronic band-structure. Thus, the unification of this data in a single database rises the question of data quality and error estimates.

To address this question, that data must be analysed to determine their quality. However, the large amount of available data makes manual inspection infeasible. Therefore, (semi-) automatic workflows must be employed to identify the most reliable results. In this thesis, this is addressed by introducing a quantitative measure of materials similarity. This measure is then used to evaluate the impact of computational parameters on the electronic structure of materials. This analysis can be used to study the reliability of scientific results by comparing the similarity in cases where high similarity between different calculations is expected, e.g. for different calculations of the same material.

The quantitative evaluation of materials similarity requires a numerical representation of materials, a so-called fingerprint. Material fingerprints represent materials in a specialised manner, usually concentrating on selected materials. In this work, the similarity of materials is studied in terms of their electronic density-of-states (DOS). DOS fingerprints were reported in literature by O. Isayev *et al.* [5], representing the DOS in a fixed energy interval as a vector of 256 real numbers. In this work, a binary-valued representation of the DOS is employed, which allows to emphasise selected energy regions in the spectrum. This allows users to tailor the evaluation of materials similarity to specific features in the DOS. The construction of the DOS fingerprint is

customisable making it suited to a broader range of applications.

The correlation of the DOS similarity based on selected parameters is studied by calculating the similarity between different calculations for individual materials. For each material, results obtained with different parameters (e.g. cell volume) are compared to each other, generating similarity scores. These scores are visualised in so-called similarity matrices. In these matrices, structures from sets of calculations emerge, which are highly similar to one another. Their analysis gives insight into the correlation between DOS similarity and material parameters, while also identifying parameters affecting the DOS and making systematic deviations between calculations visible.

2 Density Functional Theory

2.1 Kohn-Sham Equations

The basis for DFT is laid by the Hohenberg-Kohn theorem [6] which states, that the ground state energy E of an interacting electron gas in an external potential $v(\vec{r})$ is a unique functional of the electron density $n(\vec{r})$:

$$E[n] = F[n] + \int d\vec{r}\, v(\vec{r})\, n(\vec{r}), \tag{1}$$

where $F[n]$ is a functional dependent only on the electron density. The energy $E[n]$ is minimal for the electron density $n(\vec{r})$ of the system.

Approaching the quantum-mechanical many-body problem from the electron density $n(\vec{r})$ simplifies it from a 3N-dimensional to a 3-dimensional problem. Solving for the energy is not straightforward, as the form of the term $F[n]$ is in general unknown and has to be approximated.

A system of many interacting electrons can be uniquely mapped to an auxiliary system of non-interacting electrons, with the same the electron density as the interacting electron system. The basis for this approach was set by W. Kohn and L. J. Sham [7]. The auxiliary system of non-interacting electrons is described by single particle Schrödinger equations:

$$\left[-\frac{1}{2}\nabla^2 + v_{eff}(\vec{r})\right] \psi_i(\vec{r}) = \epsilon_i\, \psi_i(\vec{r}), \tag{2}$$

where each solution $\psi_i(\vec{r})$ to Eq. (2), corresponds to a single non-interacting electron of the auxiliary system, with an eigenvalue ϵ_i. The effective potential $v_{eff}(\vec{r})$ can be split into a sum of interactions:

$$v_{eff} = v_{ext} + v_H + v_{XC}. \tag{3}$$

v_{ext} is the external potential acting on the electrons, v_H is the Hartree potential, describing the Coulomb interaction of electrons and v_{XC} accounts for exchange and correlation effects. The exchange-correlation term v_{XC} is defined as the functional derivative of the exchange-correlation energy with respect to the electron density $n(\vec{r})$:

$$v_{XC} = \frac{\delta E_{XC}}{\delta n(\vec{r})}. \tag{4}$$

The electron density $n(\vec{r})$ is calculated from the single particle wave-functions of the auxiliary system:

$$n(\vec{r}) = \sum_{i=1}^{N} |\psi_i(\vec{r})|^2. \tag{5}$$

The general form of the exchange-correlation energy E_{XC} is unknown and has to be approximated. One possible approximation for the exchange-correlation energy E_{XC} is the local density approximation (LDA), where the exchange-correlation effects are approximated to those of a uniform electron gas with the same electron density $n(\vec{r})$.

$$E_{XC}^{LDA}[n] = \int d\vec{r}\, \epsilon_{XC}^{unif}(n)\, n(\vec{r}), \tag{6}$$

where ϵ_{XC}^{unif} is the exchange-correlation energy per particle of the uniform electron gas with an electron density $n(\vec{r})$. Other approaches are possible, such as the generalized gradient approximation (GGA), where the gradient of the electron density is taken into account.

$$E_{XC}^{GGA}[n] = \int d\vec{r}\, f[n(\vec{r}), \nabla n(\vec{r})], \tag{7}$$

Going one step further, the second derivative, the Laplacian of the electron density, is considered in meta-GGA exchange-correlation functionals. Another step further, hybrid functionals, combine semi-local functionals with the Hartree-Fock approach. More accurate and computationally efficient ways of formulating the exchange-correlation energy are a topic of ongoing research.

The Kohn-Sham equations have to be solved self-consistently, as the exact form of the electron density and of the XC potential are unknown. The problem is solved by starting from an initial guessed electron density $n_0(\vec{r})$. The exchange-correlation potential is then computed according to Eq. 4 and the Kohn-Sham Equations are mapped to a linear eigenvalue problem which can be solved numerically for the eigenvalues ϵ_i and eigenvectors ψ_i. The initial guess $n_0(\vec{r})$ is updated by the new electron density $n_j(\vec{r})$. This process is repeated until the energy (and possibly other properties) is converged, i.e. does not change with new iterations.

With a converged electron density $n(\vec{r})$ the total energy E_{TOT} of the system can be computed, and from the single particle eigenvalues ϵ_i the Kohn-Sham band structure and the DOS can be computed. Different DFT codes employ different implementation of the method of solving the Kohn-Sham Equations. There are all-electron codes, where every electron of the system is included in the calculation. An example of such a code is FHI-aims [8] or `exciting` [9]. Contrary to that approach, some codes make use of pseudopotentials to approximate the low lying core electrons, which leads to a drastic reduction in the computational effort. One example of such popular DFT code is VASP [10].

2.2 Density of states (DOS)

The electronic DOS describes the number of available electronic states as a function of energy. For a system of volume V with energy levels $E_n(\vec{k})$, the DOS is defined as:

$$DOS(E) = \frac{1}{V} \sum_n \sum_{\vec{k}} \delta(E - E_n(\vec{k})), \tag{8}$$

where $\vec{k}$ is the reciprocal space vector, n is the band index. The energy eigenvalues $E_n(\vec{k})$ are obtained from the Kohn-Sham equations. The DOS is approximated by discretizing the k-space and computing the sum of Eq. (8) using this k-point mesh. The k-space needs to be sampled densely enough until convergence with respect to (w.r.t) the k-point mesh is achieved. Convergence is reached, when the DOS does not further change with increasing number of k-points.

3 Material similarity

The DFT formalism allows the computation of material properties from first principles, requiring only the atomic positions and species of the involved chemical elements. In this work, materials are defined to be similar when they resemble each other in one or some of their properties. To compute the similarity between two materials, the electronic structure can be utilised. There are two common representations of the electronic structure, the band structure and the DOS. The band structure is not suitable as it is symmetry dependent and not all high-symmetry points are shared between materials. In contrast to the band structure, the DOS is an integrated property, making it comparable for all materials. To compute the similarity between materials, the DOS needs to be numerically represented. For this purpose a DOS fingerprint is defined.

3.1 Material fingerprints

In this thesis, a DOS fingerprint based on the work of Kuban *et al.* [11] is used, which encodes a selected region of the DOS while also allowing for feature selection inside of this region. This means that a specified energy interval of the DOS (further referred to as a feature region) is dominant in the fingerprint. Energy regions outside of this interval are still present in the fingerprint, their influence on the fingerprint is, however, suppressed in comparison to the feature region.

The DOS fingerprints are generated by the following procedure: The energy interval $[E_0, E_1]$, in which the DOS is defined, is partitioned into subintervals $[E_i, E_i + \Delta E_i]$ with variable width ΔE_i. The width of these subintervals, ΔE, is modulated by a the following, empirically chosen, function [11]:

$$\Delta E(x) = \left\lfloor \frac{1}{\delta\epsilon}(1 + \delta\epsilon - e^{-\frac{1}{2}\frac{(x-E_{ref})^2}{\sigma^2}}) \right\rfloor \delta\epsilon, \tag{9}$$

where $\lfloor\cdot\rfloor$ denotes the round down operator and the energy interval width ΔE changes in integer multiples of $\delta\epsilon$. The DOS is then integrated in every subinterval $[E_i, E_i + \Delta E_i]$:

$$f_i = \int_{E_i}^{E_i+\Delta E(E_i)} DOS(E)dE, \tag{10}$$

where f_i represent number of electronic states in the interval $[E_i, E_i + \Delta E_i]$.
A grid, whose bin width and height is given by Eq. (9), is laid on top of the DOS histogram. Individual rectangles of the grid are referred to as cells and a column of cells is referred to as a bin. The gaussian distribution of Eq. (9) is centered at the reference energy E_{ref}. The energy intervals ΔE near E_{ref} are smaller, therefore the DOS is better resolved in this region and this particular region makes up the major part of the total fingerprint. How rapidly the interval width ΔE changes is determined by the parameter σ. The following parameters are chosen for the grid: $E_{ref} = -2$ eV, $\sigma = 7$, $\delta\epsilon = 0.05$, the fingerprint is calculated in an energy interval $(-10, 5)$ eV. These parameters have been chosen empirically to emphasise the DOS near the upper edge of the valence band, while putting less emphasis on the unoccupied states and ignoring the deep lying valence states.

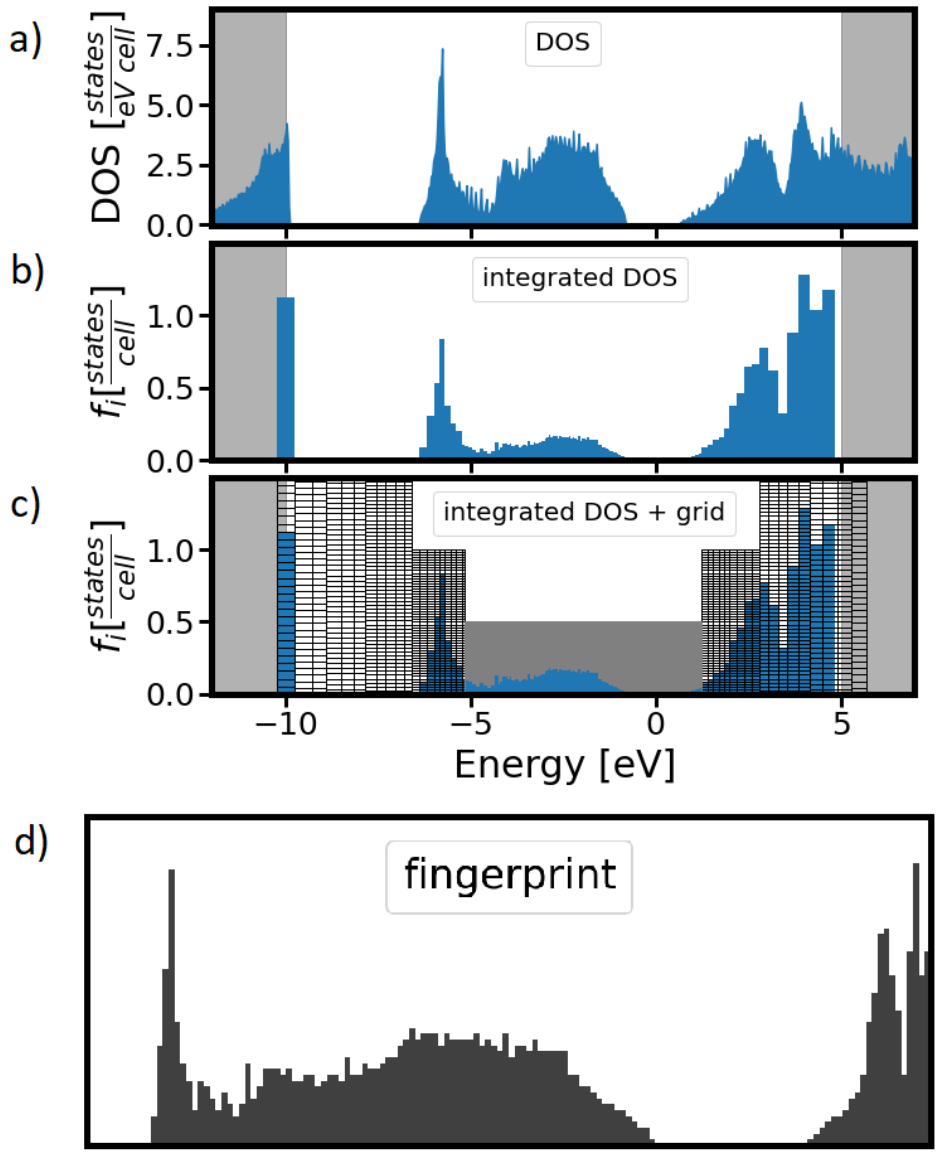


Figure 1: *Scheme of the fingerprint generation. The light grey areas in panels a)-c) indicate the borders of the fingerprint. The DOS (panel a)) is integrated into smaller portions according to Eq. 9. This generates a histogram (panel b)) over which a grid is laid (panel c)). Only those bins are kept, which are entirely within the fingerprint borders. The bins which are kept are visualised (panel d)), making up the fingerprint.*

In Fig. 1 the process of converting the DOS to a fingerprint is described. The DOS is integrated into a histogram, over which a grid is laid. Only those bins are kept, which are entirely within the fingerprint borders. The resulting fingerprint does not include the left most bin and the 2 right most bins in panel c) of Fig.1. Note that the region near -2 eV is resolved very well compared to the regions near the edge of the energy interval. Each column of the grid is made of individual cells. If a cell of the grid is fully covered it is assigned a value of one, zero otherwise. The cells of the grid are stacked on to of each other from left to right, constructing a binary vector of dimension N, where N is the number of cells in the fingerprint grid. Each bin (containing several cells) represents the amount of electronic states in an energy interval $[E_i, E_i + \Delta E(E_i)]$.

The feature region, which is dominant in the DOS fingerprint, is defined as the range where

the Gaussian distribution $G(E, E_{ref}, \sigma)$ satisfies the inequality $G(E, E_{ref}, \sigma) \geq \frac{1}{e}$. From this inequality, the feature region corresponds to the energy interval: $(E_{ref} - \sqrt{2}\sigma, E_{ref} + \sqrt{2}\sigma)$. The energy interval, to which the DOS fingerprint is restricted, may be smaller or larger than the feature region.

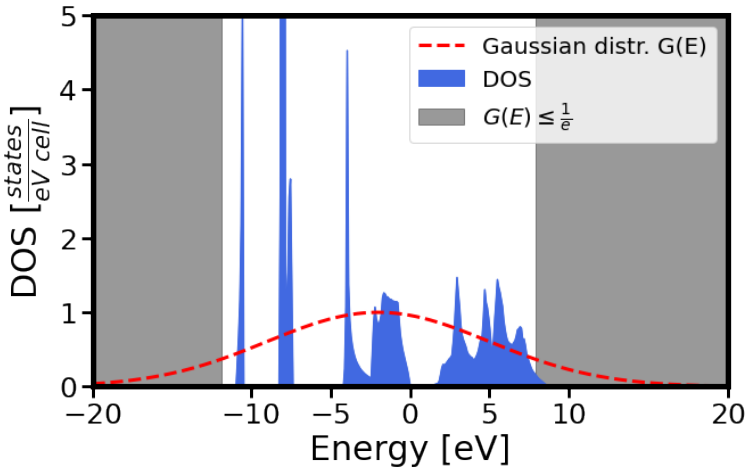


Figure 2: *Construction of the feature region. The energy interval satisfying the ineqaulity $G(E, E_{ref}, \sigma) \geq \frac{1}{e}$ is the feature region. In blue an example DOS is shown. In grey, energies outside of the feature region are highlighted. Values $E_{ref} = 2$ eV and $\sigma = 7$ are used.*

In Fig. 2, the definition of the feature region is demonstrated. In grey, energies outside of the feature region are highlighted. Note, that the feature region may make up the full fingerprint. As already stated, fingerprints in this thesis, unless stated otherwise, are restricted to the energy interval $(-10, 5)$ eV and use the parameters $E_{ref} = 2$ eV and $\sigma = 7$. In this case the entire fingerprint is in the feature region.

To treat scalar properties of materials on equal footing with the DOS, one can formally define a so-called property (PROP) fingerprint. It connects the differences of scalar properties with a similarity value. The PROP fingerprint is defined for each scalar property. The following parameters of materials are expressed in a PROP fingerprint: unit cell volume V_{UC} (in m^3), number of sampled k-points N_{kpt}(unitless), and DFT code and XC functional type in a so-called code-functional value (unitless). The code-functional value (cf-value) is an artificial score, with no physical meaning, describing the DFT code and XC functional combination of a DFT calculation. The cf-value is constructed by assigning each DFT code and XC functional combination a unique (but arbitrarily chosen) value. It is only used for the purpose of analysis and visualisation (see Chap. 5.2). The following values were chosen for DFT codes: VASP: 5.0, FHI-aims: 1.0, and for XC functionals: GGA: 0.1, LDA: 0.9. The sum of the DFT and XC functional values build the cf-value.

3.2 Similarity measures

With the materials represented in material fingerprints, their similarities are computed by a similarity coefficient S. There are many similarity coefficients in use, all of which have their advantages and disadvantages. An overview of is given by Hubálek [12]. Here, the Tanimoto coefficient [13] $T_c(\mathbf{a}, \mathbf{b})$ is used to compute the similarity between two DOS fingerprints. The Tanimoto coeffi-

cient $T_c(\mathbf{a}, \mathbf{b})$ between binary vectors $\mathbf{a}$ and $\mathbf{b}$ and is given as follows:

$$T_c(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}||^2 + ||\mathbf{b}||^2 - \mathbf{a} \cdot \mathbf{b}}, \tag{11}$$

where $\cdot$ is the vector scalar product and $|| \cdot ||$ is the vector norm. The similarity $T_c(\mathbf{a}, \mathbf{b})$ of two DOS fingerprints $\mathbf{a}$ and $\mathbf{b}$, is a real number in the interval $T_c \in [0, 1]$, $T_c = 1$ signifying identical materials, and $T_c = 0$ completely different materials. The scalar product between two binary vectors can be interpreted as the number of shared components with value 1. For DOS fingerprints, this translates to the number of shared cells fully covered by the DOS histogram. This gives meaning to $S(\mathbf{a}, \mathbf{b}) = 0$, signifying no overlap between two DOS fingerprints. The following properties are also fulfilled of the Tanimoto coefficient of binary vectors:

$$T_c(\mathbf{a}, \mathbf{a}) = 1 \tag{12}$$

$$T_c(\mathbf{a}, \mathbf{b}) = T_c(\mathbf{b}, \mathbf{a}) \tag{13}$$

$$T_c(\mathbf{a}, \mathbf{c}) \geq T_c(\mathbf{a}, \mathbf{b}) + T_c(\mathbf{b}, \mathbf{c}) - 1. \tag{14}$$

To establish a similarity measure for the PROP fingerprint, one can formally define for two scalar values a and b:

$$S_{PROP}(a, b) = \frac{1}{|a - b| + 1}. \tag{15}$$

While this definition allows the DOS and PROP similarities to be treated on equal footing, it needs to be noted that the value of S_{PROP} depends on the spread of the property values. Because of this, the proper choice of the unit of a and b is necessary to recieve interoperable results. It should also be stated that the case $S_{PROP}(a, b) = 0$ is not achievable. This property similarity function S_{PROP} is used to only help visualising and understanding the DOS similarity between materials and connecting it to the similarity of material properties.

3.3 Visualisation and Analysis

The similarities between all pairs of N materials of a dataset are stored in an N x N matrix, the similarity matrix. The similarity matrix contains the similarity between material n and m in the n x m-th entry of the matrix. The matrix has 1's on the diagonal, as the comparison of a fingerprint to itself returns a similarity score of $S = 1$. The similarity measures in this thesis, Eqs. (11), (15) are symmetric, and the similarity matrix is therefore also symmetric.

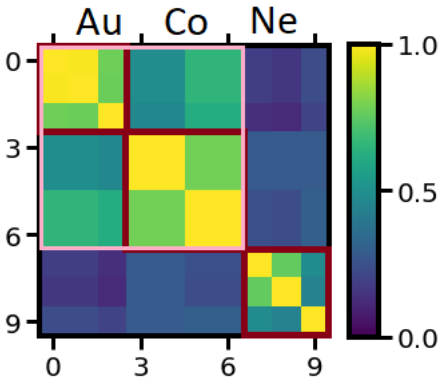


Figure 3: *Similarity matrix for a set of 10 materials, indexed 0-9. The x and y coordinates of the 10x10 matrix are arbitrary indices of the materials. The n x m-th entry of the matrix is the similarity between the materials n and m, as indicated by the color of the matrix entry. The red squares indicate groups of similar calculations. On top of the matrix, the atomic composition of the corresponding calculations is shown.*

In Fig. 3, the similarity matrix for a set of 10 materials is visualised. The x and y axes are arbitrary indices of the materials. The colour code describes the similarity between pairs of materials. We see that groups of calculations have formed which are more similar to each other than to other calculations. These groups, which will further be referred to as clusters, are marked by red and pink squares. To isolate the property which is causing the formation of clusters the similarity matrices are sorted with respect to a selected property. In Fig. 3, this is the atomic composition. We realise that each cluster corresponds to a specific material. The first three materials are elemental solids of Au, the second cluster (materials 3-6) contains elemental solids of Co. The lowest cluster contains elemental solids of Ne. The similarity between metals and noble gases is lower, as neon is a classical example of a noble gas, while gold and cobalt are metals exhibiting nearly free electron behaviour.

In the following, the impact of used methods and computational parameters on the DOS is studied. The data of the NOMAD Encyclopedia is analysed for each material by sorting the similarity matrix w.r.t selected properties. Each material is analysed separately, i.e. each similarity matrix contains only calculations corresponding to one material. In the ideal case it is expected, that the similarity between all calculations of a material will be 1, i.e. the DOS are identical. This is not necessarily the case, as DFT data is computed and converged with an objective in mind. If this objective is not the DOS (but for example the elastic properties), then the computed spectrum may not give fully converged results (while the elastic properties will be reliable). Thus, only if two DOS are computed in the exact same way, will their similarity be 1. If some parameters are different, this may lead to differences in the results which will be reflected in the similarity. Obviously, the similarity matrix will also reflect this change. The sorting of the similarity matrix is used to inspect the correlation of the DOS w.r.t. the numerical parameters introduced in Sec. 3.1.

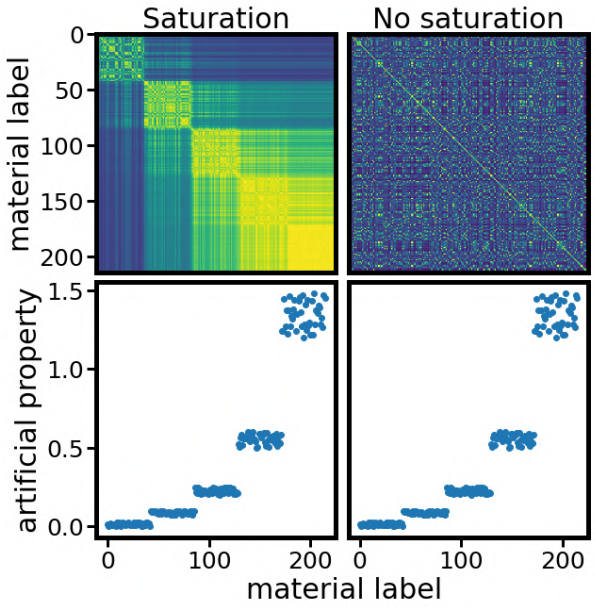


Figure 4: *Similarity matrices generated using artificial data. Similarity matrices are sorted by increasing value of the numerical value of the artifical property (shown below each matrix). Two behaviours of the similarity matrix are shown, 'Saturation' where the similarity correlates with the numerical property and 'No saturation', where no such correlation exists.*

Fig. 4 schematically presents the expected appearance of similarity matrices under different circumstances. The 'Saturation' data demonstrates the expected behaviour of a similarity matrix, which correlates with a numerical value. The similarity between calculations increases going down the diagonal, eventually becoming 1 for all materials in the lower right corner of the matrix, which shows convergence w.r.t. the numerical value. This indicates that the numerical value has influence on the similarity, and sorting of the matrix w.r.t. the numerical value reveals clusters in the matrix. In the second case ('No saturation'), the sorting of the similarity matrix w.r.t. the numerical value reveals no clusters in the matrix and conversely no correlation between similarity and the numerical value is observed.

4 Data

The data that is analysed in this thesis is provided by the NOMAD Laboratory [4]. Within this project, materials data from more than 100 million individual code runs is made available to the scientific public. This is achieved by requesting input and output files for the upload to the data repository. Based on the information contained in these files, a calculation is fully characterised. To allow for standardised access to this information, the data contained in the uploaded files is parsed. Parsing is the automatized processing by a special computer program, the parser, which reads specific information (like atomic positions or the band gap) from these files and writes them to a maschine readable dictionary. A single calculation is a single run of a DFT code with all of its results. Within the NOMAD Laboratory, materials are identified by calculating a hash value which is composed by the atomic positions, chemical species, unit cell and periodic boundary conditions.

The DOS from DFT codes are computed through a discrete sampling of the Brillouin zone. This finite sampling may introduce numerical instabilities or give results which do not describe the real system well. To circumvent this problem, a smoothing function is used to occupy the states of the system. There are many smearing methods used, such as Methfessel-Paxton smearing [14], Fermi smearing or Gaussian smearing [15], to name a few.

The DOS data used here originate from the NOMAD Encyclopedia, along with all other information, with one exception, the number of sampled k-points, which originates from the NOMAD Archive.

Search query	Number of materials
DOS	1 907 108
DOS + VASP + GGA	1 907 086
DOS + FHI-aims + LDA	160
DOS + FHI-aims + GGA	179
DOS + exciting + LDA	14

Table 1: *Results of search queries of the NOMAD Encyclopedia as of February 2021, Show the number of materials containing at least a single DOS calculation satisfying the search query.*

In total there are almost 2 million materials with DOS calculations available in the NOMAD Encyclopedia, with the majority of these materials having at least one VASP calculation using a GGA XC functional. There are less than 400 materials, which have a DOS calculation with FHI-aims and less than 20 materials from `exciting`.

To restrict the dataset size, cubic bulk materials with more than 300 individual DOS are selected. Further, only VASP and FHI-aims calculations with GGA or LDA are considered, as these DFT codes and XC functionals are most common in the NOMAD Encyclopedia. The full DFT data is accessible in the NOMAD Archive, using the search engine and the upload ID of the calculations. The upload ID is a global and unique identifier, assigned to the upload of the data to the Archive. The upload IDs are given in the Appendix for each material individually. By sorting DOS similarity matrices by DFT code, XC functional, the number of sampled k-points and the unit cell volume, clusters form which show the correlation of the DOS with the respective parameters.

Space group number	Materials
221	Po
225	Ag, Al, Cu, Ir, Mn, Ni, Pb, Pd, Pt, Rh
227	Ge, Si, Sn
229	Cr, Fe, Mo, Nb, Ta, V, W

Table 2: *Bulk elemental solids, which are studiedin this work. Each of these materials has a minimum of 300 individual DOS.*

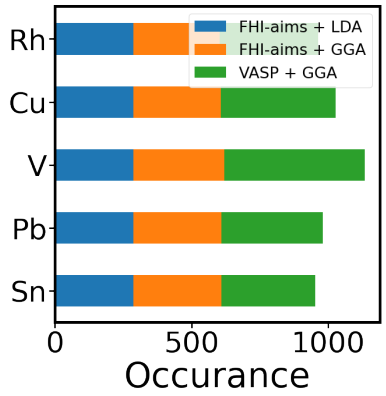


Figure 5: *Selected materials from Table 2, showing the distribution of XC functionals and codes for individual materials. For information on all materials of Table 2, see Appendix.*

5 Results

5.1 Feature selection

Different energy ranges of the DOS have different importance for specific applications. This is reflected by the proportion of the material fingerprint which describes this region. The grid implemented in the generation of the DOS fingerprint, is used to influence this proportion. It can be adjusted, to reflect a specific energy region more accurately using the parameters E_{ref} and σ. By setting σ to a large number a homogenous grid can be produced and by setting σ to a small number the fingerprint will reflect a specific feature near E_{ref}. What is considered a feature of the DOS is difficult to define as it is context dependent. A feature may have many forms (a peak, a plateau, etc.), depending on the material in question.

In the following example of CdTe, the feature selection of the DOS fingerprint is demonstrated.

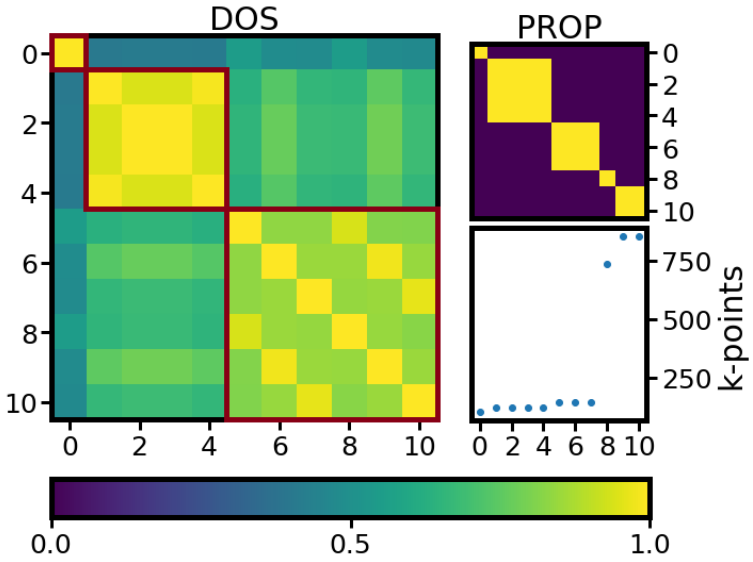


Figure 6: *On the left the DOS similarity matrix (computed from Eq.* (11)*) of 11 different DFT calculations for CdTe. On the righ the PROP similarity matrix visualises the similarity of the sampled k-points(using Eq.* (15)*). The scatter plot below the PROP matrix shows the number of sampled k-points for each calculation. The x axis shows the material index. In the similarity matrices the y axis is the same material index, in the scatter plot the y axis shows the number of sampled k-points. The similarity of pairs of calculations is colour coded, blue indicating similarity 0 and yellow similarity 1.*

In Fig. 6, the similarity analysis of 11 DFT calculations of CdTe (space group number (SPG) 216) is presented. The data are sorted by increasing number of sampled k-points. The sorting produces clusters in the DOS similarity matrix (marked by red squares). The similarity of pairs of DOS is colour coded, shown at the bottom. The entries on the diagonal have naturally similarity 1 (i.e. identical DOS). The same color coding is used for the PROP matrix, where the similarity in terms of the sampled number of k-points is visualised. The number of sampled k-points is shown in the scatter plot below. One can see that the first two clusters are reproduced in the PROP matrix. This hints at a strong correlation between the DOS similarity and the number of sampled k-points. The third DOS cluster does not have a correspondance to the clusters in the PROP matrix. In order to resemble the PROP matrix, the last cluster would need to split into 3 subclusters. This is, however, not the case, and shows, that these calculations are not sensitive to the variation in the number of sampled k-points, above a certain threshold. This suggests, that the DFT calculations inside this

cluster are converged w.r.t. the number sampled k-points. In this cluster, off-diagonal entries of high similarity have formed. As is shown below, in Fig. 7, they are formed because of a certain feature in the DOS.

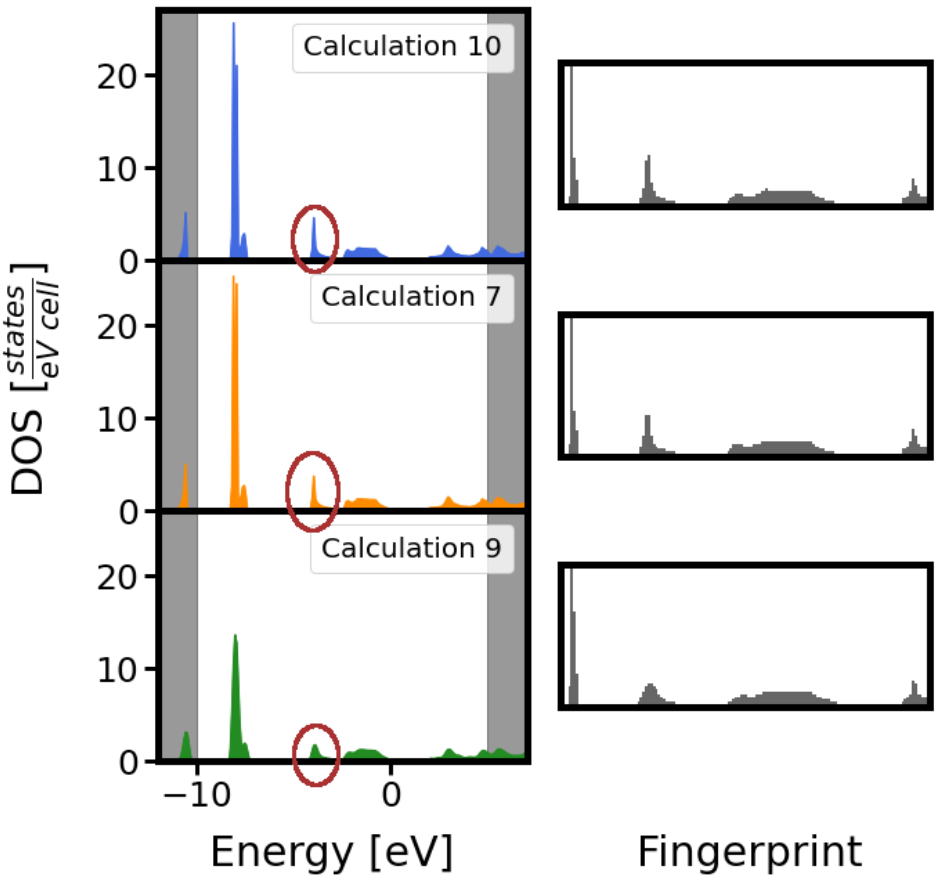


Figure 7: *DOS of calculations 10, 9 and 7 from Fig. 6. The grey area signifies the borders of the fingerprint. On the right, the fingerprint of the respective material is visualised. A feature of the DOS is marked in red ellipses.*

In Fig. 7, 3 spectra from the third cluster of Fig. 6 are shown. Calculations 10 and 7 form an off diagonal entry while calculations 10 and 9 do not. For calculations 10 and 7, the peak at -4 eV (marked by the red ellipse) has similar height and width. Calculation 9, while having this peak as well, has approximately half the magnitude, and is broader. This difference is reflected in the fingerprints of the respective calculations and it is concluded that this peak is responsible for the formation of the off-diagonal entries.

The fingerprints of Fig. 7 have been generated by a grid, which is centered at $E_{ref} = -2$ eV, $\sigma = 7$, the feature region being $(-2 \pm \sqrt{2} \cdot 7)$ eV. To show, that the peak at approx. -4 eV is responsible for the creation of the secondary diagonal in the DOS matrix in Fig. 6, the feature region is now restricted to only include this peak. This is realised by using the parameters $E_{ref} = -4$ eV and $\sigma = 1$, the feature region is then $(-4 \pm \sqrt{2})$ eV. The feature region is now only a small part of the fingerprint.

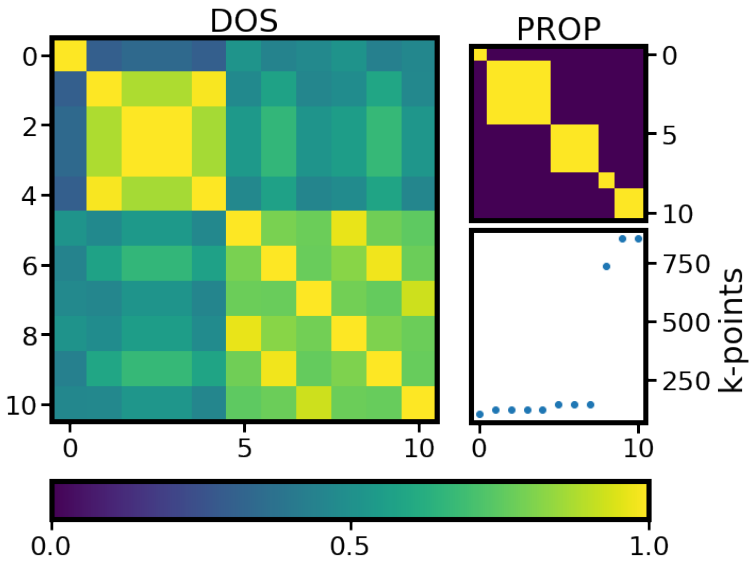


Figure 8: *Same as Fig. 6, but with the fingerprint generated using different grid parameters:* $E_{ref} = -4$ *eV,* $\sigma = 1$.

The general form of the DOS matrix does not change (in comparison to Fig. 6), i.e. there are still 3 clusters present. However, by modifying the feature region to include only the peak, showing differences between individual DOS, the secondary diagonal of high similarity in the third cluster is enhanced, as shown in Fig. 8.

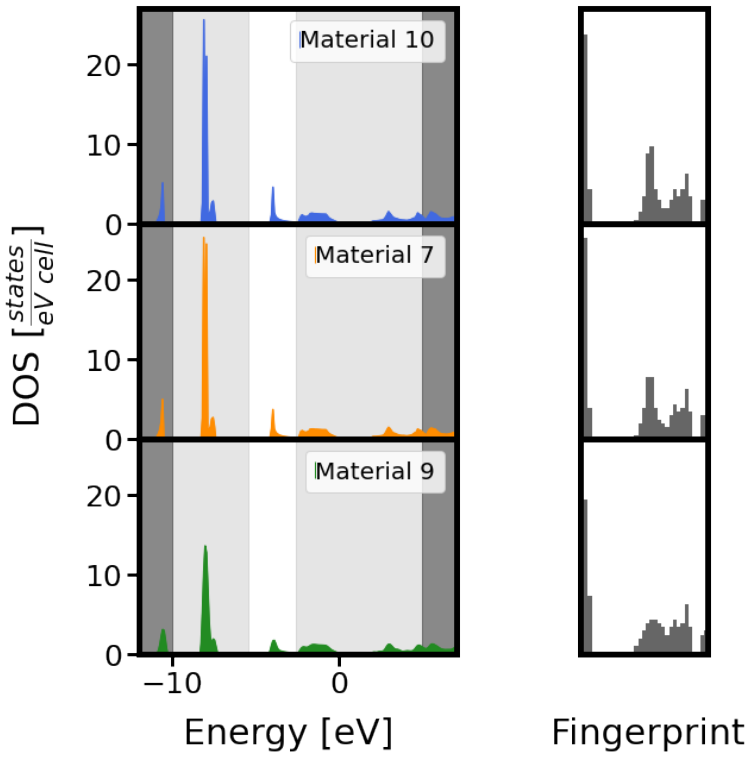


Figure 9: *Same as Fig. 7, but fingerprint generated using different grid parameters:* $E_{ref} = -4$ *eV,* $\sigma = 1$. *The feature region is indicated by white background, in light grey the fingerprint region excluding the feature region is shown, dark grey signifies DOS outside of the fingerprint.*

Fig. 9 shows the same DOS as Fig. 7, but with a different fingerprint specification. In Fig. 9, the feature region only includes a single peak. On the right, the fingerprint of the respective DOS is shown. Note, that a large proportion of the fingerprint is composed by the peak in the feature region. In contrast to Fig. 6, Fig. 8 shows an enhanced secondary diagonal in the last cluster. Based on this, it is concluded, that this peak is responsible for the off-diagonal. In the Appendix, a similarity matrix is available, where the feature region is shifted away from this peak. In this case the secondary diagonal in the third cluster of Figs. 7 and 9 is not observable.

To summarise, comparison of Figs. 7 and 9, shows that features of the DOS are reflected in the fingerprint, especially those in the feature region. Through the selection of the feature region, the focus can be put on specific parts of the DOS. Figs. 6 and 8 show that the similarity is sensitive to this focus. It is observed, that within a cluster, a secondary diagonal can form, which highlights a feature shared by pairs of calcuations, as demonstrated on the example of CdTe. Therefore, this clustering can be used to detect features in calculations.

5.2 Impact of DFT code and XC functional

In 2016, reproducibility of DFT results was studied by K. Lejaeghere *et al.*, reporting that energies computed by most DFT codes agree very well [16]. In this section we ask the question, how different DFT codes (VASP or FHI-aims) and XC functionals (LDA or GGA) influence the density-of-states.

To apply the method described in Chapter 3.3, a cf-value, describing the combination of code and functional, is defined. This value has no physical meaning and is used only for visualisation. For details on the construction of the cf-value, see Chapter 3.1.

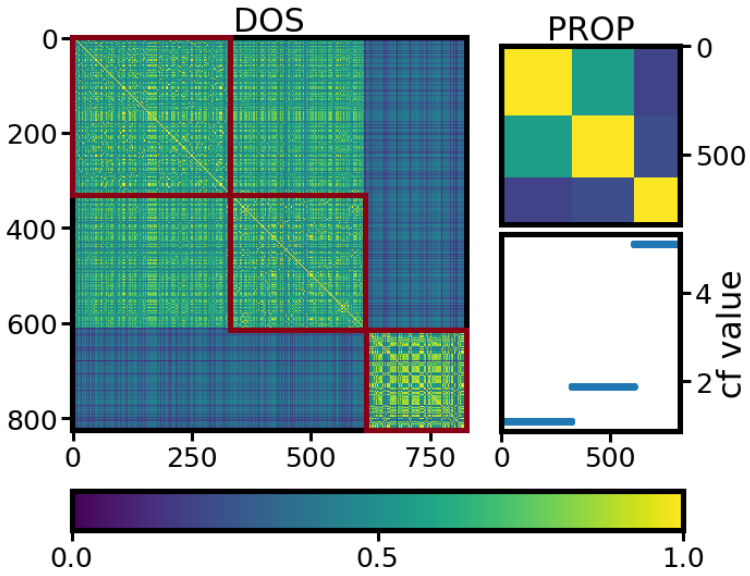


Figure 10: *Similarity matrices, including more than 800 DOS calculations of Al (SPG 225), obtained from XC functionals LDA and GGA, and from DFT codes FHI-aims and VASP. On the left, the DOS similarity matrix is shown. On the right, the PROP matrix visualises the similarity of the cf-value. Clusters corresponding to different DFT code and XC functionals are marked by red squares in this order from to to bottom: FHI-aims + GGA, FHI-aims + LDA, VASP + GGA.*

In Fig. 10 the DOS similarity of more than 750 DOS calculations for Al is studied. The data is sorted by DFT code and XC functional, which is encoded in the cf-value. In the PROP matrix, three clusters, corresponding to DFT code and XC functional combinations, formed. The three groups of materials are marked in the DOS matrix. However, only two clusters formed in the DOS matrix, corresponding to FHI-aims and VASP. The largest (FHI-aims) cluster of the DOS matrix, has no subclusters, even though it contains DOS from FHI-aims + LDA and FHI-aims + GGA (each marked by red squares). From this it is concluded, that GGA or LDA with FHI-aims makes no significant difference in the DOS. The similarity between the (VASP) cluster and the FHI-aims cluster is low. Such behaviour is unexpected as different DFT codes should give similar results, if converged. This shows systematic differences in the DOS spectra computed by FHI-aims and VASP. The same behaviour was observed for all materials studied in this work.

In the following, 3 DOS are chosen for further analysis, each representing one DFT code and XC functional combination.

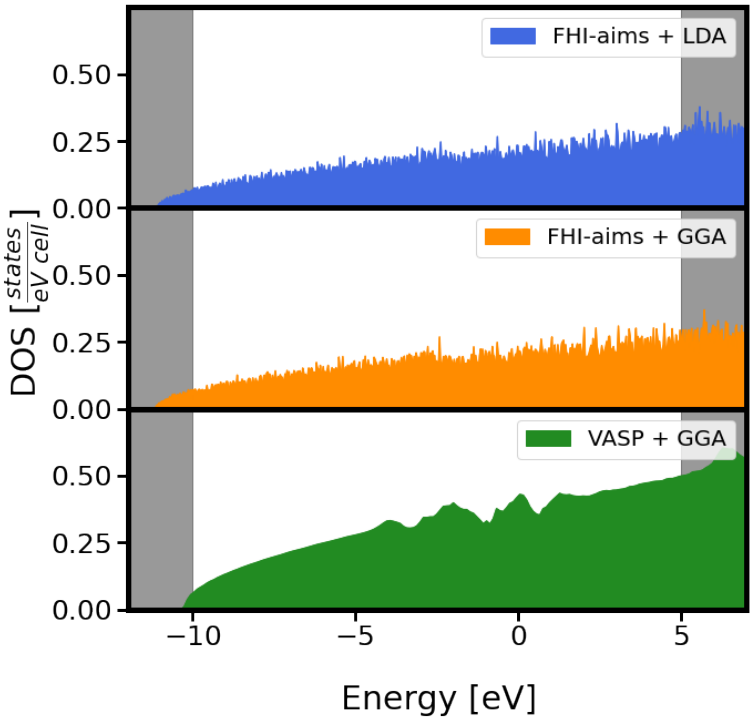


Figure 11: *3 DOS chosen from Fig. 10, each representing a combination of DFT code and XC functional. In grey, the borders of the DOS fingerprint are visualised.*

The shapes of the spectra in Fig. 11 are similar, all three spectra resemble the DOS for a free electron gas. The VASP spectrum shows some features in the energy interval $(-5, 2.5)$ eV, which are difficult to observe in the noisy FHI-aims data. A possible explanation for the noisiness of the FHI-aims spectra is low smearing. The FHI-aims results exhibit finite values below the onset of the VASP data at approx. -11 eV. All three spectra resemble the DOS for a free electron gas. The FHI-aims results, however, have only half the magnitude of the VASP counterparts. This systematic difference indicates that the two codes use different conventions during the normalisation of the results w.r.t. the spin treatment. Based on this an error report has been sent to the NOMAD developers, indicating a problem with the normalisation of the data. The above analysis is repeated, with FHI-aims spectra multiplied by a factor of 2, which systematically improves the similarity between the two codes, as shown below.

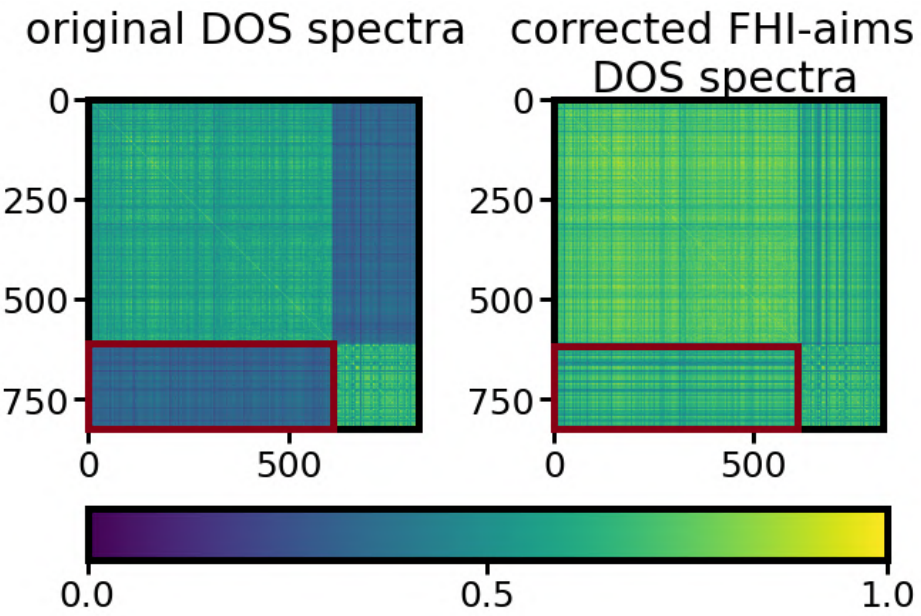


Figure 12: *DOS data from Fig. 10, for Al. The left matrix is based on the original data, the right matrix on the same data, but the FHI-aims data is multiplied by a factor of two. The matrix elements indicating the DOS similarity of different codes are marked in red.*

Fig. 12 demonstrates, that the correction of the FHI-aims DOS systematicaly increases the similarity between them and VASP spectra. Distribution of similarity values inside the marked regions is presented in the Appendix. The average DOS similarity between the codes rises from 0.33 before the correction to 0.62 after. The similarity between VASP and FHI-aims (red square in Fig. 12) is still not perfect. There are still some differences between the DOS from the two DFT codes. Similar behaviour is observed for all materials studied in this work. To summarise: two different DFT codes should not give systemtically different results if converged. Based on this result, all further analysis is done with corrected FHI-aims DOS.

5.3 Impact of k-point mesh

The convergence of the DOS is assessed with the number of sampled k-points in terms of the DOS similarity. If the k-point mesh is chosen appropriately (convergence is achieved) then denser k-point meshes will no longer improve the DOS. The appropriate choice of the k-point mesh depends on the material and on the required precision. The k-point mesh is studied in this chapter, resulting in an approximate number of sampled k-points for a reasonably converged DOS.

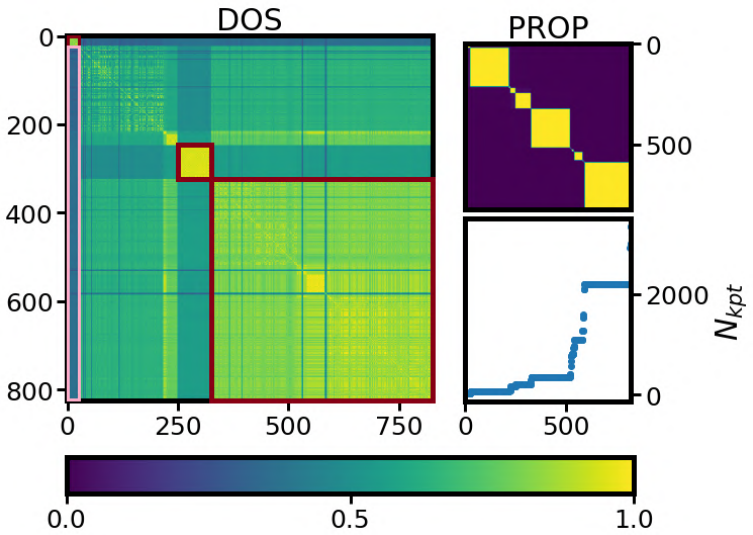


Figure 13: *Similarity matrix, including more than 800 DOS of Al (SPG 225), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. The PROP similarity matrix shows the similarity of the number of sampled k-points. The scatter plot below shows the number of sampled k-points for each calculation.*

In Fig. 13, the DOS similarity of Al (SPG 226) is visualised, sorted by increasing number of k-points N_{kpt}, with calculations using between 32 to 3339 k-points. Starting from the top of the matrix, calculations with 32 sampled k-points, marked by first red square, are not similar to any other calculation (as evident from the similarities marked by pink rectangle), including those with thousands of sampled k-points. Based on this observation it is concluded, that 32 k-points are insufficient to converge the DOS of Al. The DOS clusters correspond to the clusters of the PROP matrix, up to calculation 326. In this region, a cluster of calculations with an unexpectedly low similarity to all of the other calculations has formed, marked by second red square. From material 326 onwards (marked by third red square), the DOS calculations are no longer sensitive to the increasing N_{kpt}. This shows, that the threshold for convergence w.r.t. N_{kpt} is reached. Selected DOS from the DOS matrix are now studied, to confirm these results, and to inspect the abnormal low similarity cluster.

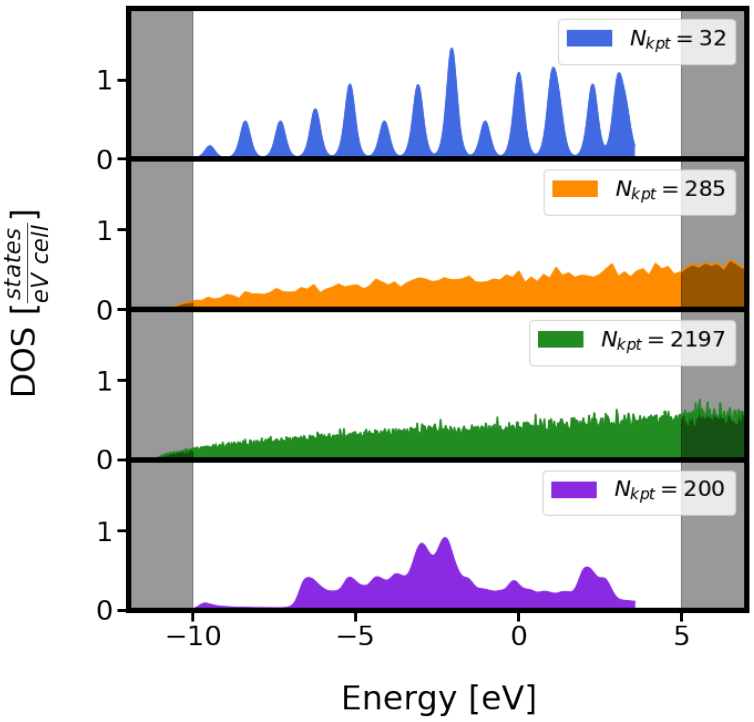


Figure 14: *The DOS of selected calculations from Fig. 8. The grey area signifies the borders of the fingerprint. Blue, orange and purple DOS are from VASP + GGA, green DOS is from FHI-aims + LDA. The selected calculations are from the marked clusters from Fig. 13: blue from first red square, purple from second, the rest from the third red square.*

In Fig. 14, selected DOS calculations are shown to demonstrate the influence of the number of sampled k-points. It is observed, that 32 k-points are insufficient for a converged DOS calculation, and delta like peaks are observed. With 285 k-points, the DOS no longer shows delta peaks and resembles the DOS with a high N_{kpt}. The similarity coefficient between the orange and green DOS is 0.83, whereas between blue and green only 0.37. In Fig. 13 an unexpected island, marked by the second red square, has formed. This island contains only VASP calculations, which are not similar to any other calculations. The purple DOS originates from this cluster. DOS calculations of this cluster may use some other parameters, which strongly change the DOS, as it does not resemble any of the other DOS. What these parameters may be is however at the moment unknown.

To summarise, for well-converged results a high number of k-points is required, while less dense meshes only provide a rough estimate at a lower computational cost. The formation of clusters can be used to identify an approximate minimum value, which gives a reasonably converged DOS. By the above argumentation, for Al (SPG 225), the minimum number of k-points needed for a reasonably converged DOS caluclation is $N_{kpt} = 285$. At the same time, the clustering of the DOS has revealed a cluster of DOS, which shows abnormal differences to all the other DOS, further illustrating the use of the methods of this work as a tool for evaluating data reliability.

5.4 Impact of unit cell volume

The volume of the unit cell has a complex effect on the electronic states and on the DOS. By using a non-relaxed volume, pressure is introduced to the unit cell and the electronic states are perturbed, which changes the DOS. If the bonds are strong, then a change in bond length will perturb the electronic states strongly. The magnitude of this perturbation is material dependent

and it is studied in following section. The classification of materials w.r.t. their hardness is driven by the bulk modulus B [17] [2] [3].

To study this behaviour the DOS similarity matrices are sorted by the unit cell volume and compared to the PROP matrices, containing the unit cell volume similarities. This is demonstrated on two materials, one with a small bulk modulus and one with a large bulk modulus.

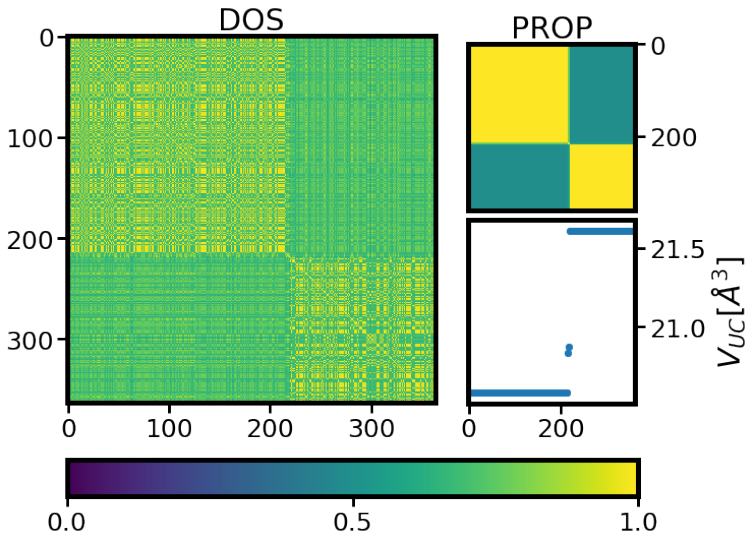


Figure 15: *Similarity matrix, including more than 300 DOS of OsB (SPG 187, $B = 354.632$ GPa [18]), obtained from XC functionals LDA and GGA and DFT codes FHI-aims and VASP. The PROP similarity matrix shows the similarity of unit cell volumes. The correlation between the DOS similarity and the unit cell volume is shown by comparing the DOS and the PROP matrices.*

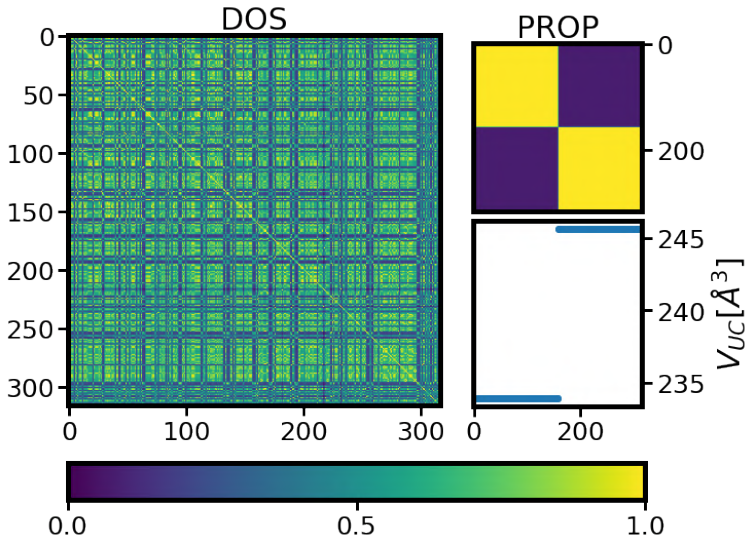


Figure 16: *Similarity matrix, including more than 300 DOS of Cs (SPG 229, $B = 2.27$ GPa [19]), obtained from XC functionals LDA and GGA and DFT codes FHI-aims and VASP. The PROP similarity matrix shows the similarity of unit cell volumes. The correlation between the DOS similarity and the unit cell volume is analysed.*

Similarity matrices of two materials are shown, each containing a large number of DOS, calculated for different unit cell volumes: OsB ($B = 354.632$ GPa [18]) and Cs ($B = 2.27$ GPa [19]). Cs is the only non-cubic material studied in this work. Each material has at least 150 individual DOS spectra available for two unit cell volumes respectively, which differ by 4%. It is important to have a comparable unit cell volume change between the two materials, otherwise the comparison is not meaningful.
Fig. 15 shows the DOS similarity matrix, sorted by the unit cell volume, of bulk OsB ($K = 354.632$ GPa [18]). The sorting has produced two clusters, each corresponding to different unit cell volumes, as can be seen by comparing the DOS and PROP matrix. This correlation shows a strong influence of the unit cell volume on the DOS. Fig. 16 shows the DOS similarity matrix of bulk Cs ($B = 2.27$ GPa [19]), sorted by unit cell volume. In this case, for a material with a small bulk modulus, no clustering is observed in the DOS matrix. This shows, that for Cs, no correlation is found between the DOS similarity and the unit cell volume.

The DOS similarity of hard materials correlates with the unit cell volume, while soft materials show a weaker, or no correlation at all. The clustering of DOS with respect to the unit cell volume was observed for majority of the materials studied in this work. Most of the materials studied are metals and have a bulk modulus B larger than 100 GPa. The material with the lowest bulk modulus is Sn (SPG 141) with $B = 38$ GPa [20]. For this material, no correlation between DOS similarity and unit cell volume was observed. The material with the highest bulk modulus is Ir ($B = 345$ GPa) [21]. For Ir, a correlation between DOS similarity and cell volume was observable, it was however not as strong as in OsB. The matrices for these, and all other studied materials, are in the Appendix. For each material, at least two unit cell volumes were available with a large number of representatives. The two volumes available for each material differed by 5%, with the only exception being Si, where the volumes differed by 10.25%.

In Chapter 5.3, it was shown, that the DOS matrix, sorted by the number of sampled k-points produces clusters. These clusters, however, do not show the information of the unit cell volume. To show the information about the unit cell volume V_{UC} and the number of sampled k-points N_{kpt} a composite property ρ_k is constructed. This composite property ρ_k is defined as the product of the unit cell volume V_{UC} and the number of sampled k-points ρ_k follows:

$$\rho_k = N_{kpt} \cdot V_{UC} \ . \tag{16}$$

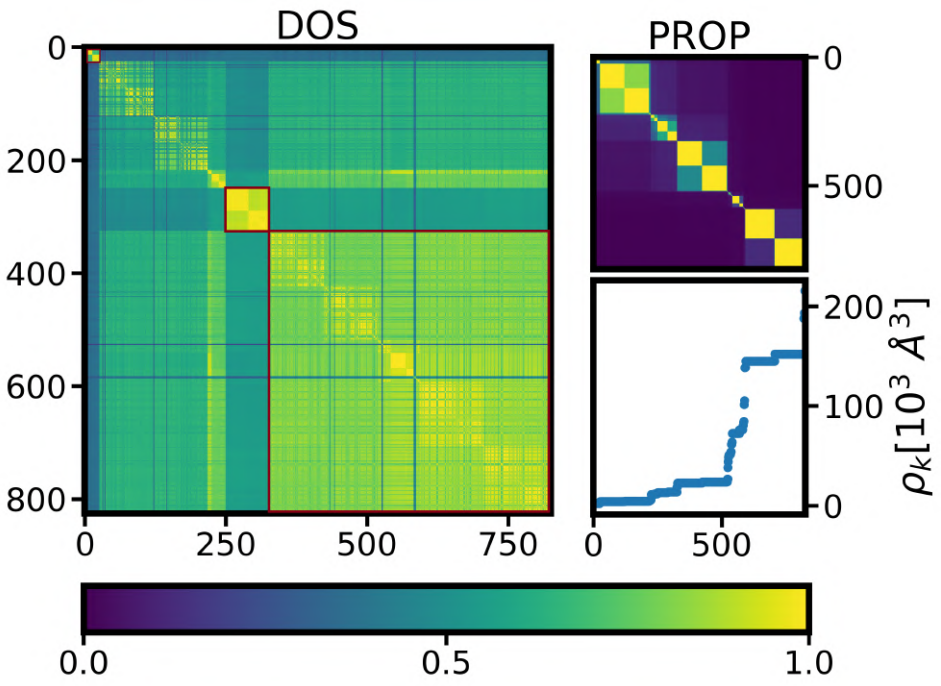


Figure 17: *Similarity matrix, including more than 800 DOS of Al (SPG 225), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. The PROP similarity matrix shows the similarity of the composite property ρ_k. The scatter plot below shows the value of the composite property ρ_k for each calculation. In red, the marked clusters from Fig. 13 are highlighted.*

Fig. 17 shows the same data as in Fig. 13, but instead of N_{kpt}, the data is sorted by the composite property ρ_k. In red, the same clusters as in Fig. 13 are marked. We observe, that the composite property ρ_k causes a better clustering of the similarity matrix. This is especially well seen in the marked clusters, which are now split into subclusters, each corresponding to a unit cell volume. The composite property ρ_k is more suitable to evaluate the convergence of the DOS than the total number of k-points N_{kpt} as it captures information about the unit cell volume and the sampled k-points.

To summarise, by introducing pressure to the unit cell of hard materials, the electronic states are perturbed, changing the DOS. The strength of this perturbation is material dependant and it is shown, that the DOS similarity correlates with the strain of the unit cell for hard materials (characterised by the bulk modulus B). For soft materials such a correlation was not found. By constructing composite properties, containing more information about the studied calculations, a finer clustering of the DOS is achieved.

6 Conclusion

To assess data quality and to ensure the interoperability of data from different sources, it is crucial to identify which parameters affect the material properties and to identify outliers in the data. In this work, the correlation between the DOS similarity and similarity of different material properties and parameters was studied. To numerically represent materials, a special fingerprint was employed. The construction of the DOS fingerprint makes it possible to reflect selected features more strongly, i. e. allows for feature selection. For investigating the pairwise similarity between materials, based on this fingerprint, similarity matrices were constructed. A similarity matrix contains all pairwise similarities between DOS and makes the analysis of these similarities more approachable. By sorting the similarity matrix, correlations between the DOS similarity and different parameters are uncovered. As a result, similarity matrices allow for an intuitive understanding, visualisation and analysis of large amounts of data.

The data used in this thesis originates from the NOMAD Encyclopedia [4]. DOS for a selection of elemental cubic solids were studied. The DOS data is computed by DFT, from DFT codes FHI-aims and VASP, using semi-local XC functionals LDA and GGA.

The correlation of the DFT code and XC functional and the DOS was studied to examine the influence of semi-local XC functionals on the DOS. The methods described in this work, can be used to analyse the influence of any XC functionals or DFT code on the DOS, the only requirement is enough data for the particular XC functional or DFT code. For systems under investigation, no influence of the functional was observed. Unexpected behaviour was observed, when including data from different codes.

To assess the convergence of data, the number of sampled k-points were used. It was obsreved, that the DOS similarity correlates with the k-point similarity. As the number of sampled k-points increases, the similarity between DOS calculations also increases. Once a sufficient number of k-points is sampled, the correlation between number of sampled k-points and DOS similarity disappears, indicating, that a sufficient number of k-points was sampled. An estimate for the minimum number of k-points to reasonably converge a DOS is obtainable from the analysis of the DOS similarity matrix. During this analysis, a subset of abnormal data was made visible by the sorting of the DOS matrix.

The influence of the cell volume on the DOS was also investigated. For each material, a large amount of DOS are available with different unit cell volumes. These varied by 5 %, with only silicon reaching a change of 10%. It was observed, that for hard materials (classified by their bulk modulus B), the similarity of the DOS correlates with the cell volume. This effect was not observed for soft materials, however, only a small amout of soft materials were studied. It was possible to demonstrate, that with a comparable change in unit cell volume, hard materials have their electronic states perturbed more strongly by strain than soft materials. Further, a composite property was constructed, which contains information about the unit cell volume and the number of sampled k-points. It was shown, DOS matrices sorted by this property exhibit a finer clustering.

6.1 Outlook

The analysis of the NOMAD Encyclopedia DOS data was carried out on only a few selected properties and by hand. The effect of further properties may be studied, for example the effect of smearing of the k-point mesh to get a better estimate for the parameters which results in convergence of the DOS w.r.t. the k-point mesh.

Further analysis of the correlations between the similarity of the DOS and the material parameters may be carried out using this method, as it is computationally inexpensive, as the DFT data is already available in databases. More complex inspection of the properties analysed here is also possible. In this work, it was possible to characterize the pressure by only using the unit cell volume. For example, strained systems may also be studied on the basis of the stress tensor. Further, the k-point mesh may be better analysed by considering the actual positions of the k-points the Brillouin zone, instead of just the total number of them.

The automatisation of the presented analysis methods would allow for real time data analysis and error estimation inside the database, with the possibility of making computational parameters which give best results, available to the scientific community.

Appendix

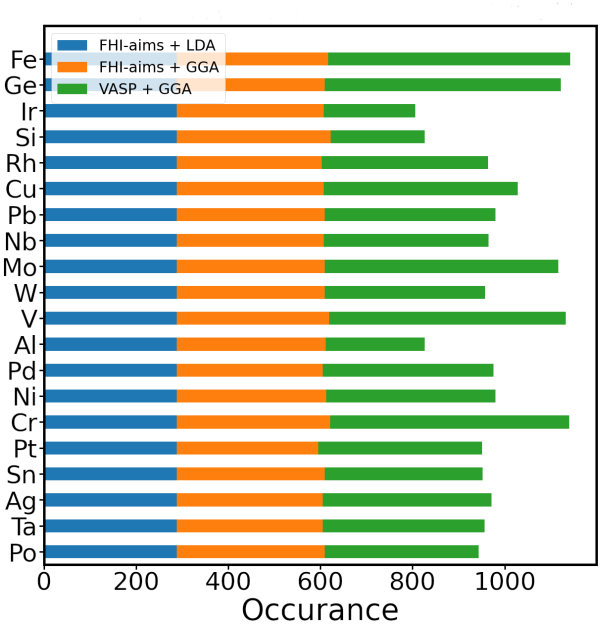


Figure 18: *All materials from Tab. 2, showing the distribution of DFT codes and XC functionals of calculations for individual materials.*

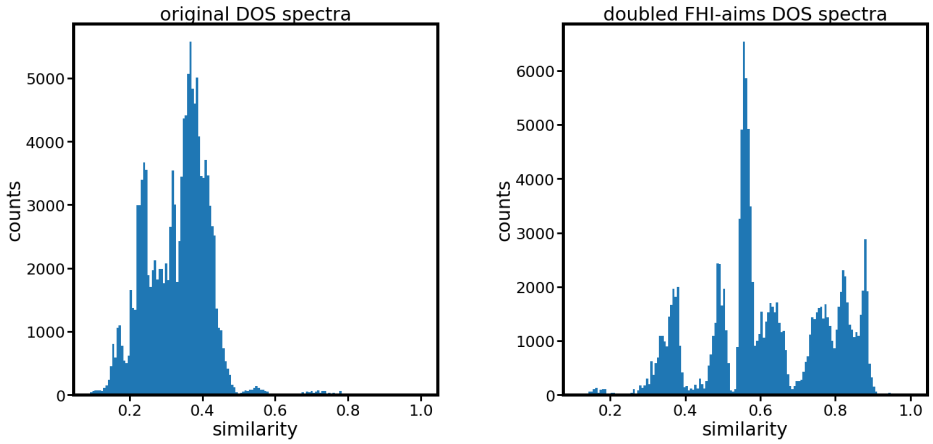


Figure 19: *Distribtuion of similarites in the marked regions of Fig. 12.*

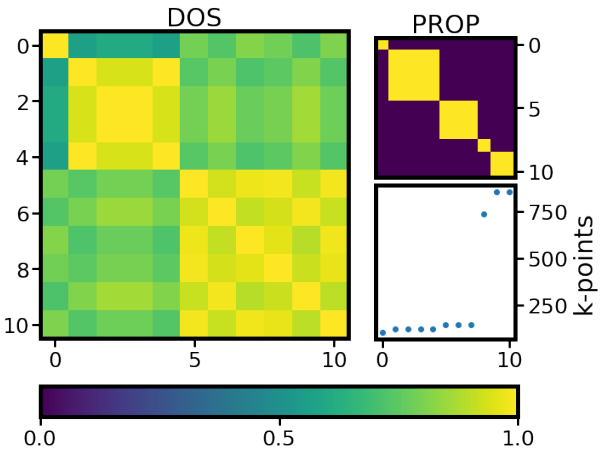
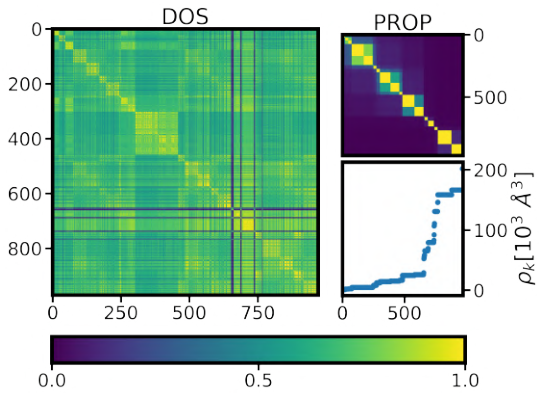
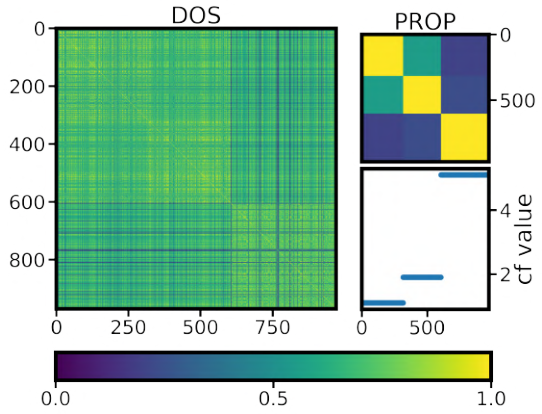


Figure 20: *Same as Fig. 6, but fingerprint generated using parameters $E_{ref} = -1$ eV and $\sigma = 1$, in an interval $(-10, 5)$ eV. The secondary diagonals in the last cluster, seen in Fig. 6, are no longer observable.*

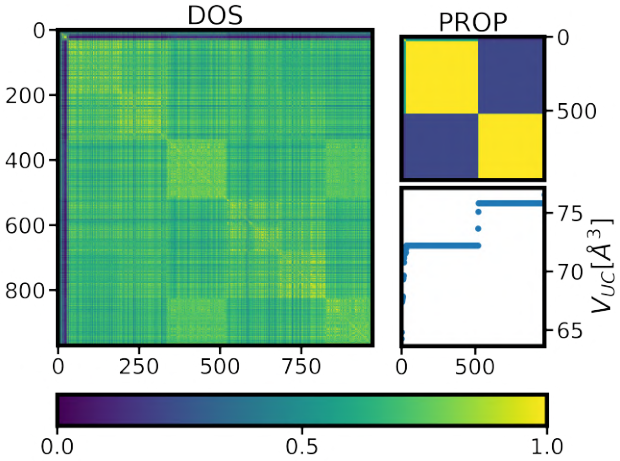
Ag, SPG 225



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



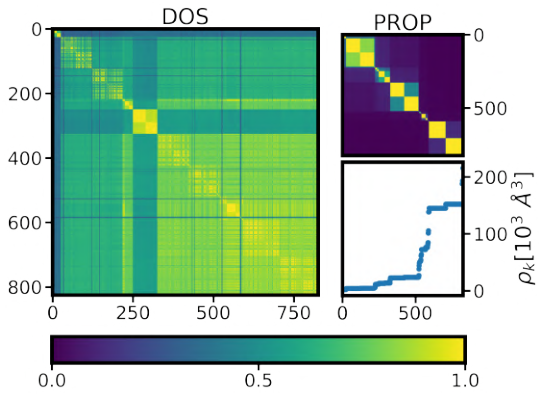
(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 21: Similarity matrices, for data of Ag (SPG 225), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

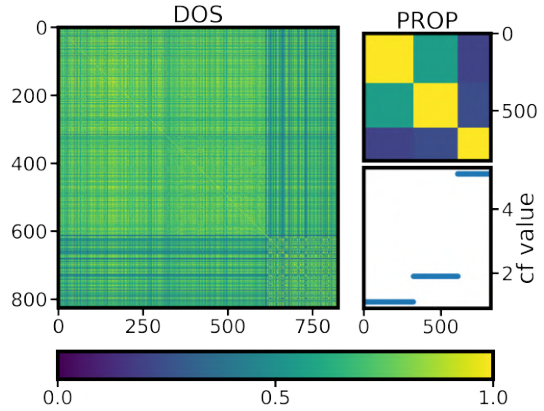
Upload IDs of the calculations: ’0-sUPGbCT4WTrNYF4-xkZQ’, ’1CFTsyFoRHuDrpYmuEwYCw’, ’1sJnuLHzTgqKJTy2ED4Ccg’, ’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’, ’5A1BV5LBTbGfuvTFh6F0GQ’, ’7jo7hwe4TZWwSCP_V6n1lA’, ’AGeD22S7TuCM0TGpN8uAoQ’, ’ATy4zekAQgW-Nlu-0-87Vg’, ’CbcV5E5sT56tBx1zm02_sQ’, ’F9l8brkjQTakEYsqQMI7iQ’, ’FjiYy5-ETRCrs5ktzJA55w’, ’HNZ153qjRRuc0h6A7fdxjg’, ’HrrEtyp8SaSZphJRpKgkXA’, ’L_LSn0amQlKS6Q1DCni4nQ’, ’MF3kN7zlT-q942wC1G33Jw’, ’SgUOMRORQ4m6FIm–Z56UA’, ’_2nyO9ClQEWCl30Uad7Rpg’, ’_ijarO1hQCqsTpOyb62WXw’, ’a30kw7QmQduZzCPtgvAFFA’, ’b5BQDtBcQFKChE0PsrjIAw’, ’e-UOMrI-QDGModmBnxIaBA’, ’eJexi2o0RGmZ-HvagQ0DMg’,

'ff6C4prETyGImi8aumXLOg', 'gH4QxKwDRtWU1c5offnNNA',
'gp_Q_YgVS06Ape6wZkQSaA', 'gr1xwRnQQ8yBfXUxkg3clw', 'hXjxFrSARi6vr8ZsckYj7Q',
'hiSoHoH4QOel0HmSbYs58Q', 'nm5KPjFuQTugZfXcR73IYA',
'uCSlNMEWQfOS5TMvsQXz1A'

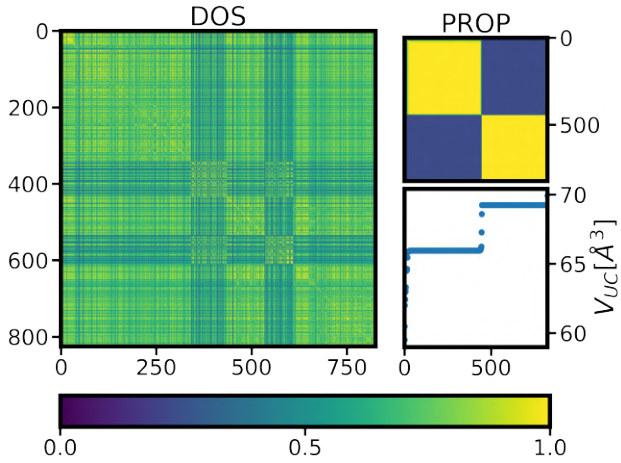
Al, SPG 225



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



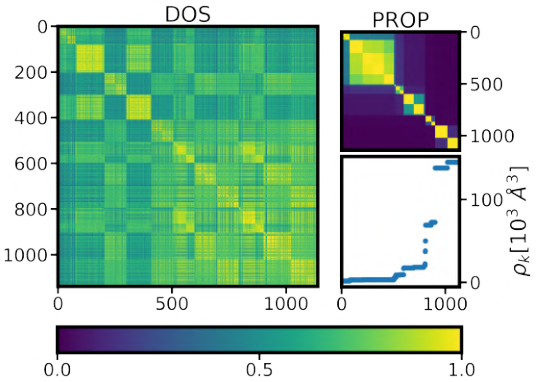
(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 22: Similarity matrices, for data of Al (SPG 225), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

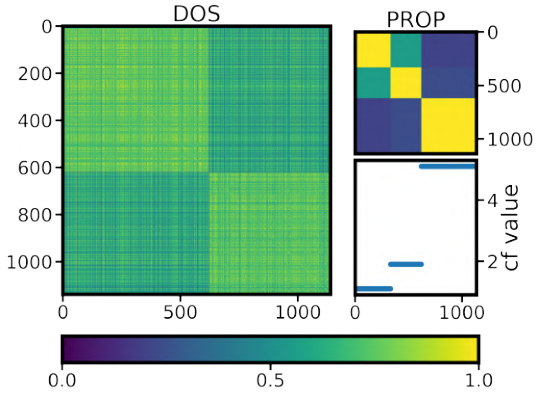
Upload IDs of the calculations:'0-sUPGbCT4WTrNYF4-xkZQ',
'1CFTsyFoRHuDrpYmuEwYCw', '1sJnuLHzTgqKJTy2ED4Ccg',
'3KZBVXpdQ2au-USEG3FwKw', '4AKh_pFmSbyUoWphewbpqQ',
'5A1BV5LBTbGfuvTFh6F0GQ', '7jo7hwe4TZWwSCP_V6n1lA',
'AGeD22S7TuCM0TGpN8uAoQ', 'ATy4zekAQgW-Nlu-0-87Vg',
'Aoh5s6VqSN-zvF9w6G3F4g', 'CbcV5E5sT56tBx1zm02_sQ',

'EkKGNHKNSNmX8TqxEh7POw', 'F6FpfGPiRSaYxoQqgfNL3A',
'FjiYy5-ETRCrs5ktzJA55w', 'HNZ153qjRRuc0h6A7fdxjg', 'L_LSn0amQlKS6Q1DCni4nQ',
'MF3kN7zlT-q942wC1G33Jw', 'SgUOMRORQ4m6FIm–Z56UA',
'_2nyO9ClQEWCl30Uad7Rpg', '_ijarO1hQCqsTpOyb62WXw',
'a30kw7QmQduZzCPtgvAFFA', 'b5BQDtBcQFKChE0PsrjIAw',
'e-UOMrI-QDGModmBnxIaBA', 'eJexi2o0RGmZ-HvagQ0DMg',
'ff6C4prETyGImi8aumXLOg', 'gH4QxKwDRtWU1c5offnNNA',
'gr1xwRnQQ8yBfXUxkg3clw', 'hBSz9S3mRMWJaIvF-l0jqg', 'hXjxFrSARi6vr8ZsckYj7Q',
'hiSoHoH4QOel0HmSbYs58Q', 'uCSlNMEWQfOS5TMvsQXz1A'

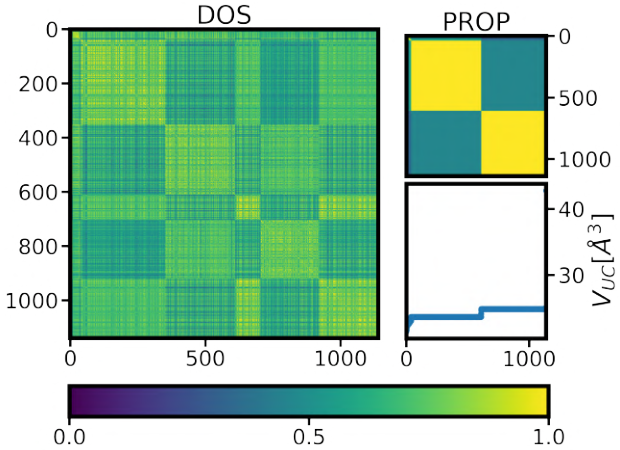
Cr, SPG 229



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



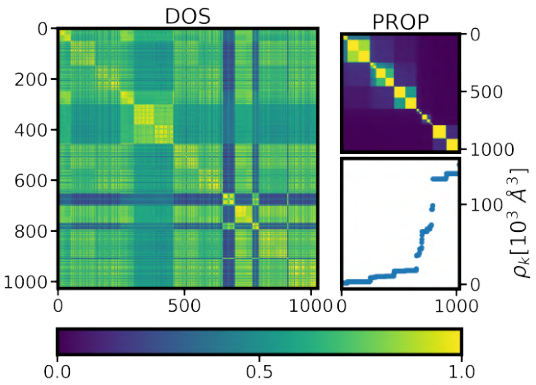
(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 23: Similarity matrices, for data of Cr (SPG 229), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

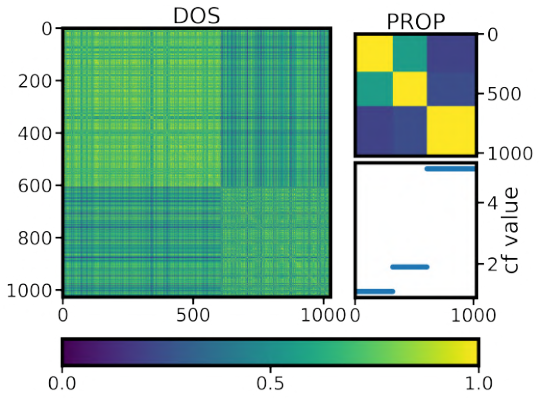
Upload IDs of the calculations: '0-sUPGbCT4WTrNYF4-xkZQ',
'1CFTsyFoRHuDrpYmuEwYCw', '1sJnuLHzTgqKJTy2ED4Ccg',

’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’,
’5A1BV5LBTbGfuvTFh6F0GQ’, ’7jo7hwe4TZWwSCP_V6n1lA’,
’AGeD22S7TuCM0TGpN8uAoQ’, ’ATy4zekAQgW-Nlu-0-87Vg’,
’CbcV5E5sT56tBx1zm02_sQ’, ’HNZ153qjRRuc0h6A7fdxjg’, ’L_LSn0amQlKS6Q1DCni4nQ’,
’MF3kN7zlT-q942wC1G33Jw’, ’QloUnHliS1aFU_Chsg1SUA’,
’SgUOMRORQ4m6FIm–Z56UA’, ’ZFNMJDbJQ9aUq3o-aMq7lw’,
’_2nyO9ClQEWCl30Uad7Rpg’, ’_6Tx47cFRK62jBx4fsvfrw’, ’_ijarO1hQCqsTpOyb62WXw’,
’a30kw7QmQduZzCPtgvAFFA’, ’b5BQDtBcQFKChE0PsrjIAw’,
’e-UOMrI-QDGModmBnxIaBA’, ’eJexi2o0RGmZ-HvagQ0DMg’,
’ff6C4prETyGImi8aumXLOg’, ’gH4QxKwDRtWU1c5offnNNA’,
’gp_Q_YgVS06Ape6wZkQSaA’, ’gr1xwRnQQ8yBfXUxkg3clw’, ’hXjxFrSARi6vr8ZsckYj7Q’,
’hiSoHoH4QOel0HmSbYs58Q’, ’mAAfBZKmTQWDVL1ZPRsrVQ’,
’rpNkNxF-TGS9e2-iKdC6bg’, ’uCSlNMEWQfOS5TMvsQXz1A’,
’w5wwrFTNQvyULhZF4ISHPQ’

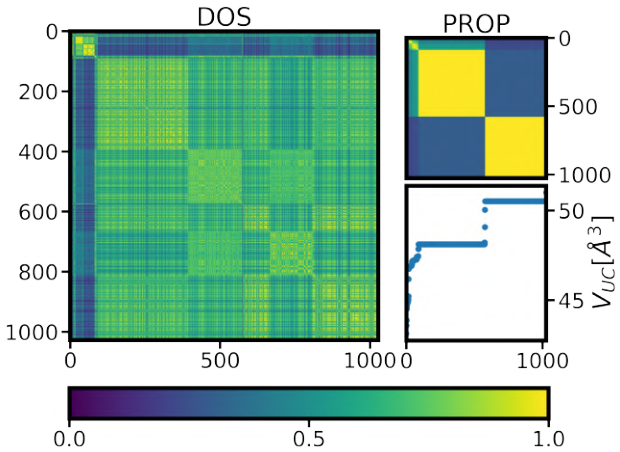
Cu, SPG 225



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



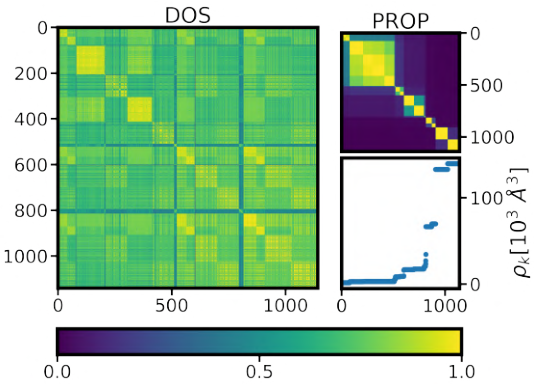
(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 24: Similarity matrices, for data of Cu (SPG 225), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

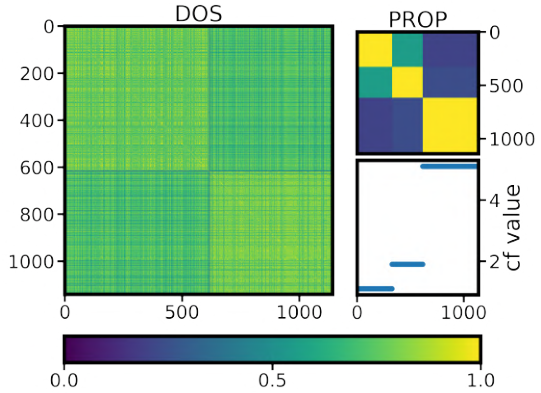
Upload IDs of the calculations:'0-sUPGbCT4WTrNYF4-xkZQ',
'1CFTsyFoRHuDrpYmuEwYCw', '1sJnuLHzTgqKJTy2ED4Ccg',
'3KZBVXpdQ2au-USEG3FwKw', '4AKh_pFmSbyUoWphewbpqQ',
'5A1BV5LBTbGfuvTFh6F0GQ', '7jo7hwe4TZWwSCP_V6n1lA',
'AGeD22S7TuCM0TGpN8uAoQ', 'ATy4zekAQgW-Nlu-0-87Vg',
'CGEPEw01RSiNbkmoIHmanA', 'CbcV5E5sT56tBx1zm02_sQ', 'F9l8brkjQTakEYsqQMI7iQ',
'FjiYy5-ETRCrs5ktzJA55w', 'HNZ153qjRRuc0h6A7fdxjg', 'I84GYdt4R5yEDBQJqZm9_g',
'L_LSn0amQlKS6Q1DCni4nQ', 'MF3kN7zlT-q942wC1G33Jw', 'N6rQDIFJT36PituBeUrqEA',
'SgUOMRORQ4m6FIm–Z56UA', 'XCP1UWuXT0SpnwhxEiDrow',
'_2nyO9ClQEWCl30Uad7Rpg', '_6Tx47cFRK62jBx4fsvfrw', '_ijarO1hQCqsTpOyb62WXw',
'a30kw7QmQduZzCPtgvAFFA', 'b5BQDtBcQFKChE0PsrjIAw',
'dcDrKssEQVCwXF3aAEqaVA', 'e-UOMrI-QDGModmBnxIaBA',

’eJexi2o0RGmZ-HvagQ0DMg’, ’ff6C4prETyGImi8aumXLOg’,
’gH4QxKwDRtWU1c5offnNNA’, ’gp_Q_YgVS06Ape6wZkQSaA’,
’gr1xwRnQQ8yBfXUxkg3clw’, ’hXjxFrSARi6vr8ZsckYj7Q’, ’hiSoHoH4QOel0HmSbYs58Q’,
’igbLK6TKSQqFKWRy2P7ZEQ’, ’mAAfBZKmTQWDVL1ZPRsrVQ’,
’mKzhnavVRtS2prPAzg9_Ug’, ’nm5KPjFuQTugZfXcR73IYA’, ’sr_exxOiRryJwEWhe-Gcww’,
’uCSlNMEWQfOS5TMvsQXz1A’, ’vw0_sd90QjuaLQ1KxzitwQ’

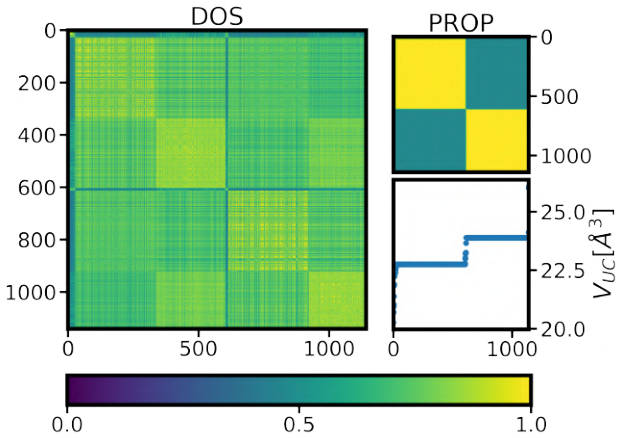
Fe, SPG 229



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



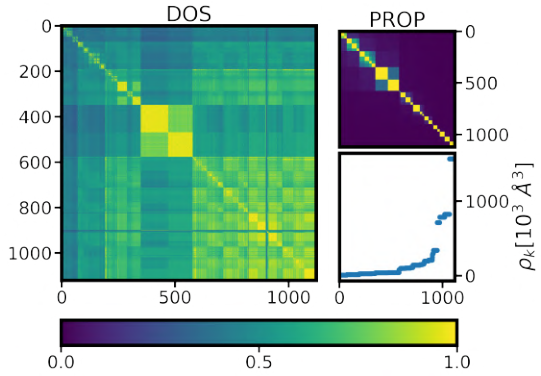
(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 25: Similarity matrices, for data of Fe (SPG 229), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

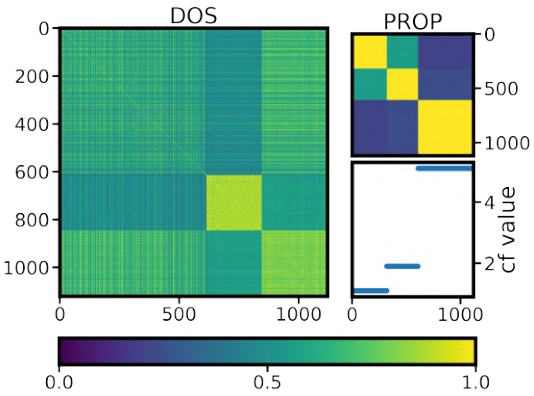
Upload IDs of the calculations: ’0-sUPGbCT4WTrNYF4-xkZQ’,
’1CFTsyFoRHuDrpYmuEwYCw’, ’1sJnuLHzTgqKJTy2ED4Ccg’,
’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’,
’5A1BV5LBTbGfuvTFh6F0GQ’, ’6L_W9ZGYTJ2-KBRJrRJXmQ’,
’7jo7hwe4TZWwSCP_V6n11A’, ’AGeD22S7TuCM0TGpN8uAoQ’,

'ATy4zekAQgW-Nlu-0-87Vg', 'Aoh5s6VqSN-zvF9w6G3F4g', 'CbcV5E5sT56tBx1zm02_sQ', 'FjiYy5-ETRCrs5ktzJA55w', 'GFwcEYgBR2SkvxgWt4-3mw', 'HNZ153qjRRuc0h6A7fdxjg', 'L_LSn0amQlKS6Q1DCni4nQ', 'MF3kN7zlT-q942wC1G33Jw', 'SgUOMRORQ4m6FIm–Z56UA', '_2nyO9ClQEWCl30Uad7Rpg', '_ijarO1hQCqsTpOyb62WXw', 'a30kw7QmQduZzCPtgvAFFA', 'b5BQDtBcQFKChE0PsrjIAw', 'e-UOMrI-QDGModmBnxIaBA', 'eJexi2o0RGmZ-HvagQ0DMg', 'ff6C4prETyGImi8aumXLOg', 'gH4QxKwDRtWU1c5offnNNA', 'gr1xwRnQQ8yBfXUxkg3clw', 'hXjxFrSARi6vr8ZsckYj7Q', 'hiSoHoH4QOel0HmSbYs58Q', 'nm5KPjFuQTugZfXcR73IYA', 'uCSlNMEWQfOS5TMvsQXz1A'

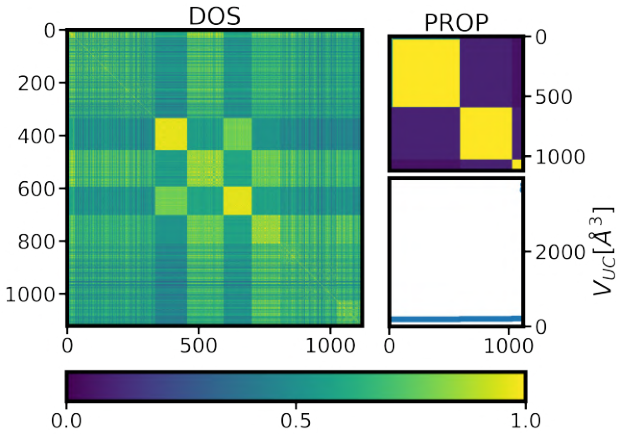
Ge, SPG 227



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.

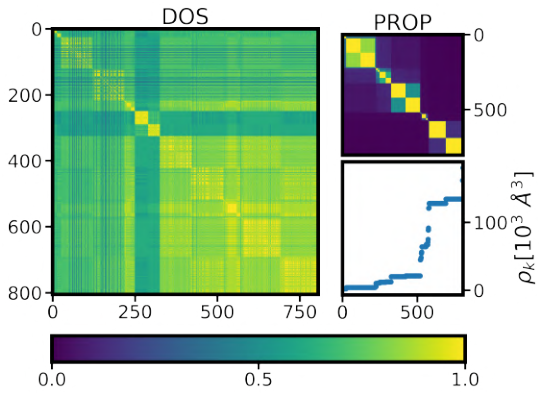


(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

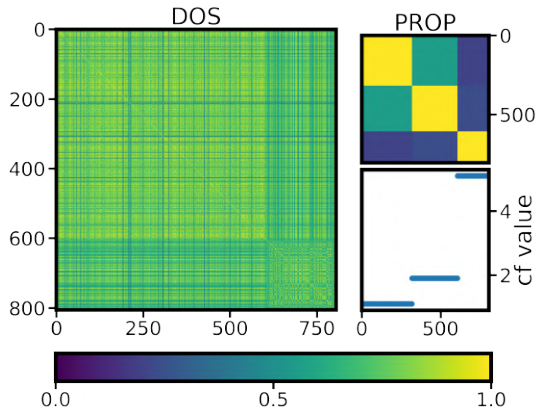
Figure 26: Similarity matrices, for data of Ge (SPG 227), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

Upload IDs of the calculations: ’1CFTsyFoRHuDrpYmuEwYCw’,
’1sJnuLHzTgqKJTy2ED4Ccg’, ’2RbfuC9SQjaeKYbaxNPZAw’,
’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’,
’5A1BV5LBTbGfuvTFh6F0GQ’, ’6ITh58J5QcGWUDsmpR9t2A’,
’7jo7hwe4TZWwSCP_V6n1lA’, ’A42JdZXMSM6ZvFkyAy3k1g’,
’AGeD22S7TuCM0TGpN8uAoQ’, ’ATy4zekAQgW-Nlu-0-87Vg’,
’CGEPEw01RSiNbkmoIHmanA’, ’CbcV5E5sT56tBx1zm02_sQ’,
’E8Ke1HW5S8qRcqXQZDs-bg’, ’EkKGNHKNSNmX8TqxEh7POw’,
’F6FpfGPiRSaYxoQqgfNL3A’, ’FjiYy5-ETRCrs5ktzJA55w’, ’HNZ153qjRRuc0h6A7fdxjg’,
’L_LSn0amQlKS6Q1DCni4nQ’, ’MF3kN7zlT-q942wC1G33Jw’,
’ZFNMJDbJQ9aUq3o-aMq7lw’, ’_2nyO9ClQEWCl30Uad7Rpg’,
’_ijarO1hQCqsTpOyb62WXw’, ’b5BQDtBcQFKChE0PsrjIAw’,
’e-UOMrI-QDGModmBnxIaBA’, ’eJexi2o0RGmZ-HvagQ0DMg’,
’gH4QxKwDRtWU1c5offnNNA’, ’gr1xwRnQQ8yBfXUxkg3clw’,
’hiSoHoH4QOel0HmSbYs58Q’, ’uCSlNMEWQfOS5TMvsQXz1A’

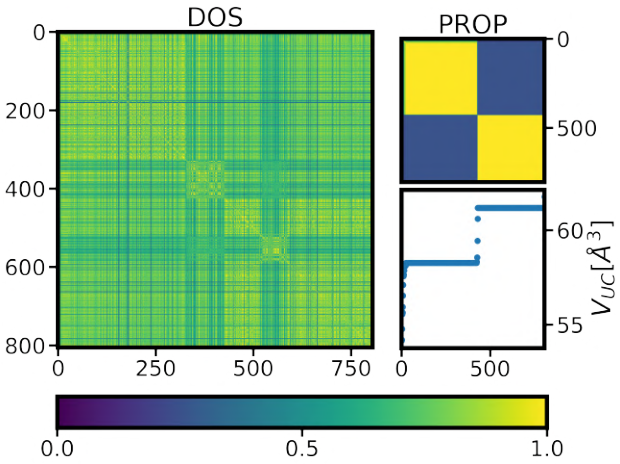
Ir, SPG 225



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



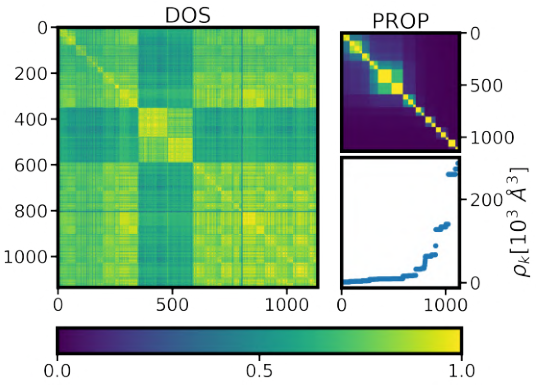
(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 27: Similarity matrices, for data of Ir (SPG 225), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

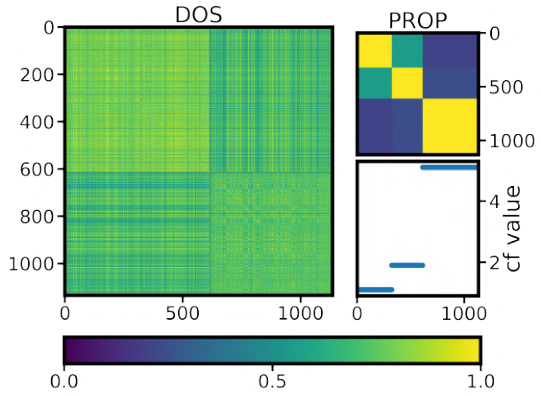
Upload IDs of the calculations: ’0-sUPGbCT4WTrNYF4-xkZQ’, ’1CFTsyFoRHuDrpYmuEwYCw’, ’1sJnuLHzTgqKJTy2ED4Ccg’, ’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’, ’5A1BV5LBTbGfuvTFh6F0GQ’, ’6ITh58J5QcGWUDsmpR9t2A’, ’7jo7hwe4TZWwSCP_V6n1lA’, ’AGeD22S7TuCM0TGpN8uAoQ’, ’ATy4zekAQgW-Nlu-0-87Vg’, ’CbcV5E5sT56tBx1zm02_sQ’, ’FjiYy5-ETRCrs5ktzJA55w’, ’L_LSn0amQlKS6Q1DCni4nQ’, ’MF3kN7zlT-q942wC1G33Jw’, ’QAetFPbFTpOkSRmNQEUfnA’, ’_2nyO9ClQEWCl30Uad7Rpg’, ’_ijarO1hQCqsTpOyb62WXw’, ’b5BQDtBcQFKChE0PsrjIAw’, ’e-UOMrI-QDGModmBnxIaBA’, ’eJexi2o0RGmZ-HvagQ0DMg’, ’gH4QxKwDRtWU1c5offnNNA’, ’gp_Q_YgVS06Ape6wZkQSaA’,

’hiSoHoH4QOel0HmSbYs58Q’, ’pyUZ9lChTd-4CYT93YNl1A’,
’uCSlNMEWQfOS5TMvsQXz1A’

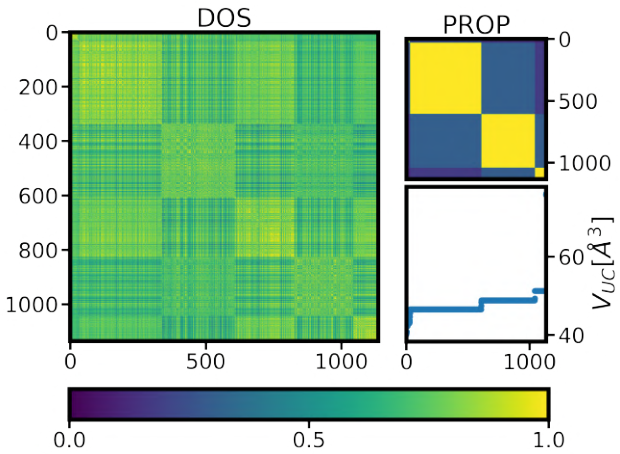
Mn, SPG 225



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



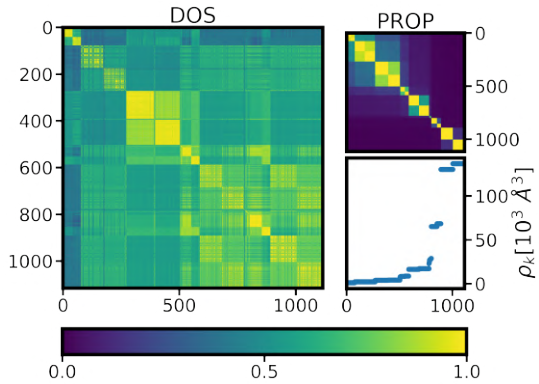
(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 28: Similarity matrices, for data of Mn (SPG 225), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

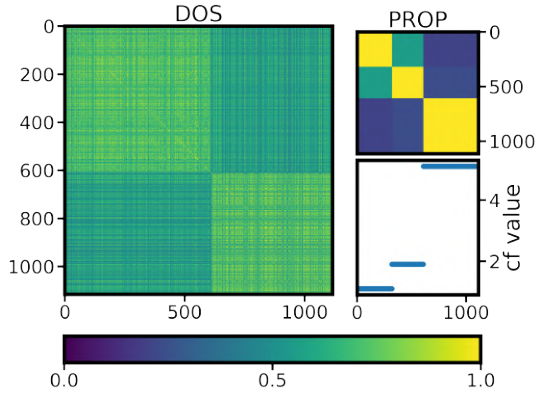
Upload IDs of the calculations: ’0-sUPGbCT4WTrNYF4-xkZQ’,
’1CFTsyFoRHuDrpYmuEwYCw’, ’1sJnuLHzTgqKJTy2ED4Ccg’,
’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’,
’4LdiZPvqRzSTS1Y-gI-WKA’, ’5A1BV5LBTbGfuvTFh6F0GQ’,
’6ITh58J5QcGWUDsmpR9t2A’, ’7jo7hwe4TZWwSCP_V6n1lA’,
’AGeD22S7TuCM0TGpN8uAoQ’, ’CbcV5E5sT56tBx1zm02_sQ’,
’FjiYy5-ETRCrs5ktzJA55w’, ’HNZ153qjRRuc0h6A7fdxjg’, ’L_LSn0amQlKS6Q1DCni4nQ’,
’MF3kN7zlT-q942wC1G33Jw’, ’PB4nB-jBRa-BrN896n0fWw’, ’_ijarO1hQCqsTpOyb62WXw’,
’b5BQDtBcQFKChE0PsrjIAw’, ’e-UOMrI-QDGModmBnxIaBA’,

’eJexi2o0RGmZ-HvagQ0DMg’, ’gH4QxKwDRtWU1c5offnNNA’,
’gp_Q_YgVS06Ape6wZkQSaA’, ’gr1xwRnQQ8yBfXUxkg3clw’,
’hiSoHoH4QOel0HmSbYs58Q’, ’pyUZ9lChTd-4CYT93YNl1A’,
’uCSlNMEWQfOS5TMvsQXz1A’, ’ydzASitvSpiyji3lsQg1gw’

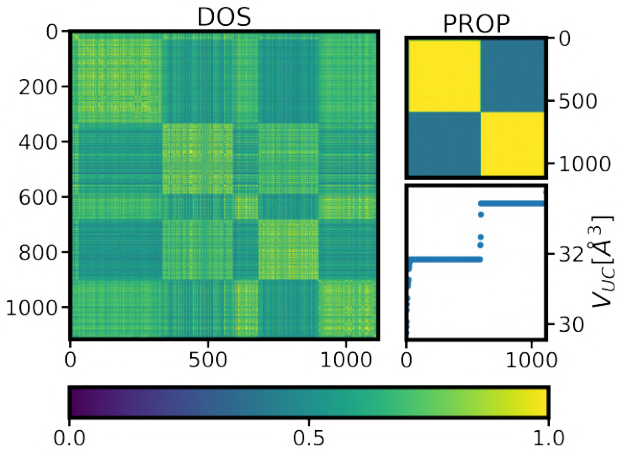
Mo, SPG 229



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



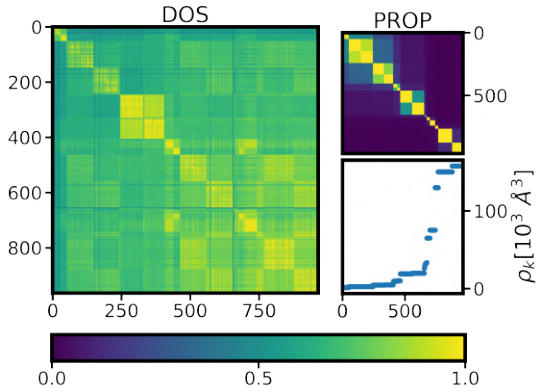
(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 29: Similarity matrices, for data of Mo (SPG 229), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

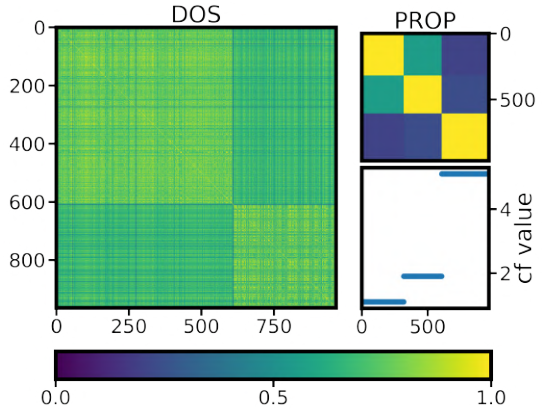
Upload IDs of the calculations: ’0-sUPGbCT4WTrNYF4-xkZQ’,
’1CFTsyFoRHuDrpYmuEwYCw’, ’1sJnuLHzTgqKJTy2ED4Ccg’,
’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’,
’4wllXGjiRsaP2PmYX_y3dQ’, ’5A1BV5LBTbGfuvTFh6F0GQ’,
’7jo7hwe4TZWwSCP_V6n1lA’, ’AGeD22S7TuCM0TGpN8uAoQ’,
’ATy4zekAQgW-Nlu-0-87Vg’, ’CbcV5E5sT56tBx1zm02_sQ’, ’FjiYy5-ETRCrs5ktzJA55w’,
’HNZ153qjRRuc0h6A7fdxjg’, ’L_LSn0amQlKS6Q1DCni4nQ’, ’MF3kN7zlT-q942wC1G33Jw’,

’SgUOMRORQ4m6FIm–Z56UA’, ’_2nyO9ClQEWCl30Uad7Rpg’,
’_ijarO1hQCqsTpOyb62WXw’, ’b5BQDtBcQFKChE0PsrjIAw’,
’e-UOMrI-QDGModmBnxIaBA’, ’eJexi2o0RGmZ-HvagQ0DMg’,
’gH4QxKwDRtWU1c5offnNNA’, ’gr1xwRnQQ8yBfXUxkg3clw’,
’hiSoHoH4QOel0HmSbYs58Q’, ’mAAfBZKmTQWDVL1ZPRsrVQ’,
’pyUZ9lChTd-4CYT93YNl1A’, ’t8gilx1XSMecneUVshmf5Q’,
’uCSlNMEWQfOS5TMvsQXz1A’

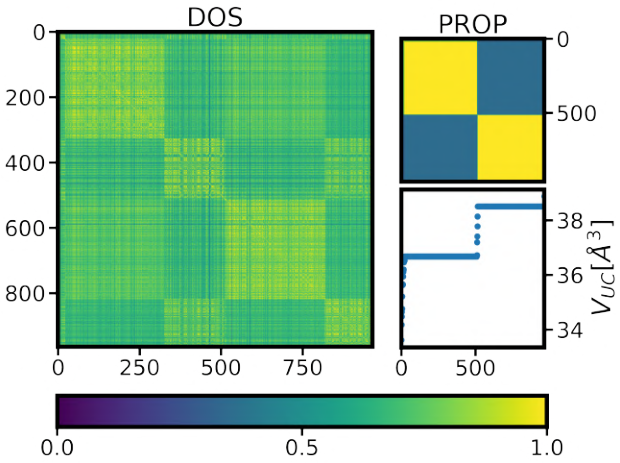
Nb, SPG 229



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



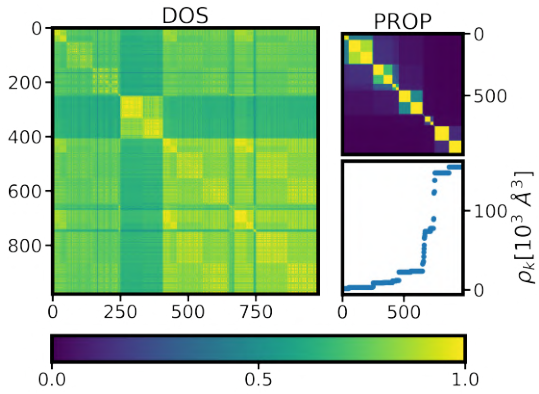
(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 30: Similarity matrices, for data of Nb (SPG 229), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

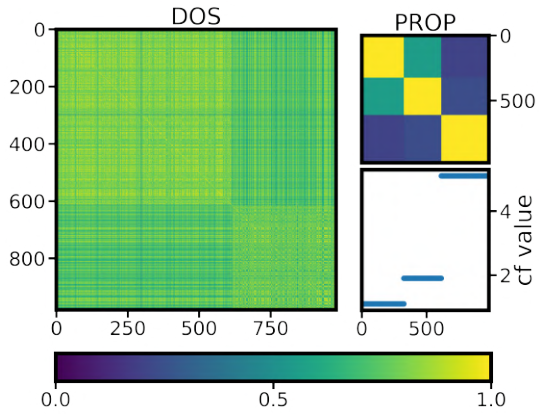
Upload IDs of the calculations: ’0-sUPGbCT4WTrNYF4-xkZQ’,
’1CFTsyFoRHuDrpYmuEwYCw’, ’1sJnuLHzTgqKJTy2ED4Ccg’,
’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’,

'5A1BV5LBTbGfuvTFh6F0GQ', '7jo7hwe4TZWwSCP_V6n11A',
'AGeD22S7TuCM0TGpN8uAoQ', 'ATy4zekAQgW-Nlu-0-87Vg',
'Aoh5s6VqSN-zvF9w6G3F4g', 'CbcV5E5sT56tBx1zm02_sQ', 'FjiYy5-ETRCrs5ktzJA55w',
'HNZ153qjRRuc0h6A7fdxjg', 'L_LSn0amQlKS6Q1DCni4nQ', 'MF3kN7zlT-q942wC1G33Jw',
'VshkgA9fRVaBUM1A99_pTA', '_2nyO9ClQEWCl30Uad7Rpg',
'_ijarO1hQCqsTpOyb62WXw', 'b5BQDtBcQFKChE0PsrjIAw',
'e-UOMrI-QDGModmBnxIaBA', 'eJexi2o0RGmZ-HvagQ0DMg',
'gH4QxKwDRtWU1c5offnNNA', 'gr1xwRnQQ8yBfXUxkg3clw',
'hiSoHoH4QOel0HmSbYs58Q', 'oyEFPKzsQ8-zFtlNPcSXYw',
'uCSlNMEWQfOS5TMvsQXz1A'

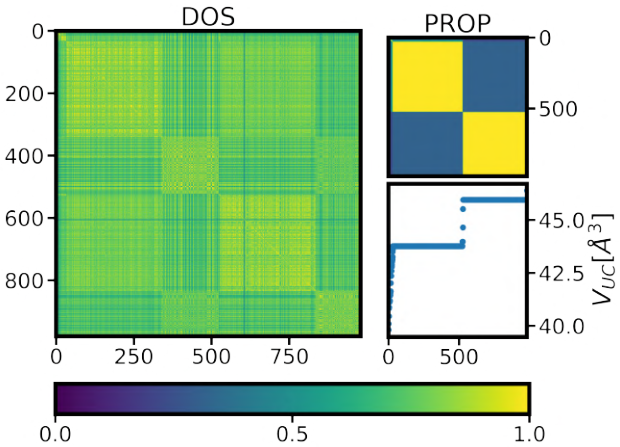
Ni, SPG 225



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.

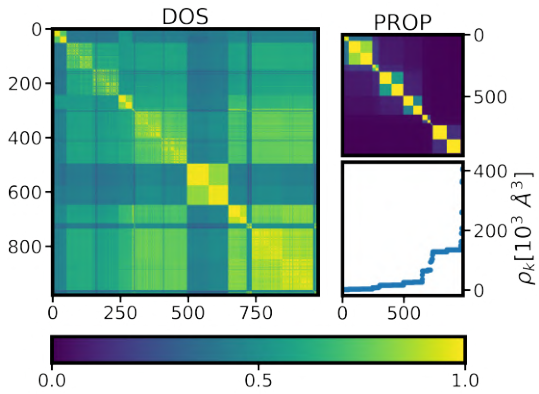


(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

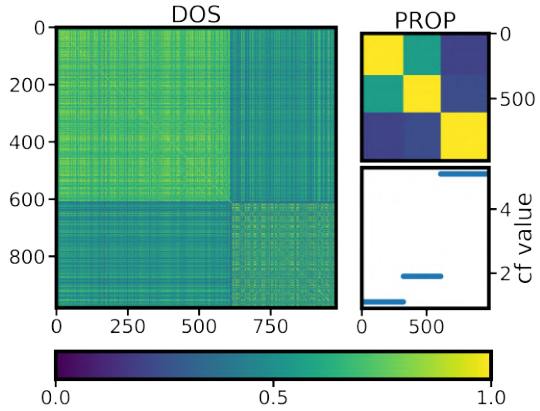
Figure 31: Similarity matrices, for data of Ni (SPG 225), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

Upload IDs of the calculations: ’0-sUPGbCT4WTrNYF4-xkZQ’,
’1CFTsyFoRHuDrpYmuEwYCw’, ’1sJnuLHzTgqKJTy2ED4Ccg’,
’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’,
’5A1BV5LBTbGfuvTFh6F0GQ’, ’6ITh58J5QcGWUDsmpR9t2A’,
’7jo7hwe4TZWwSCP_V6n1lA’, ’AGeD22S7TuCM0TGpN8uAoQ’,
’ATy4zekAQgW-Nlu-0-87Vg’, ’CbcV5E5sT56tBx1zm02_sQ’, ’FjiYy5-ETRCrs5ktzJA55w’,
’HNZ153qjRRuc0h6A7fdxjg’, ’L_LSn0amQlKS6Q1DCni4nQ’, ’MF3kN7zlT-q942wC1G33Jw’,
’_2nyO9ClQEWCl30Uad7Rpg’, ’_ijarO1hQCqsTpOyb62WXw’,
’b5BQDtBcQFKChE0PsrjIAw’, ’e-UOMrI-QDGModmBnxIaBA’,
’eJexi2o0RGmZ-HvagQ0DMg’, ’gON45FEiRfCVg8YD0IHwvQ’,
’gp_Q_YgVS06Ape6wZkQSaA’, ’gr1xwRnQQ8yBfXUxkg3clw’,
’hiSoHoH4QOel0HmSbYs58Q’, ’uCSlNMEWQfOS5TMvsQXz1A’

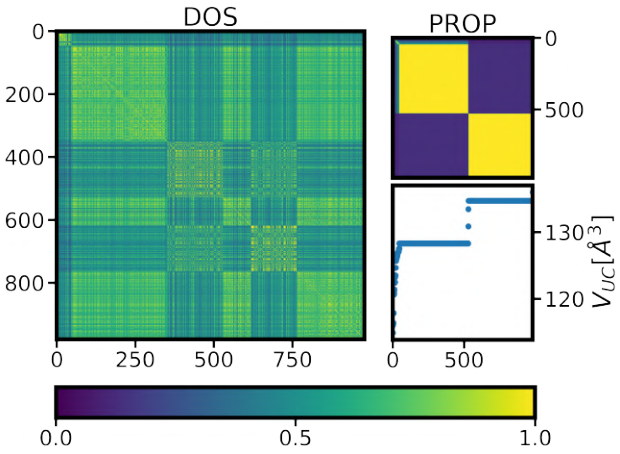
Pb, SPG 225



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



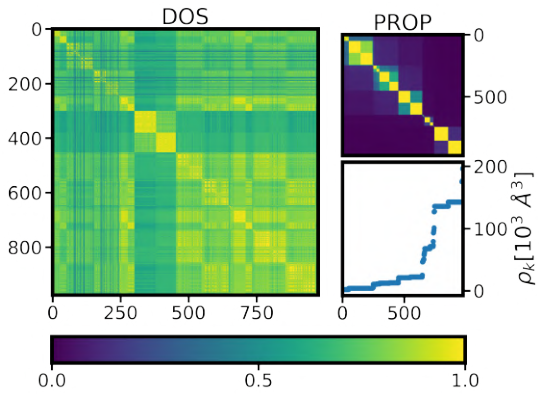
(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 32: Similarity matrices, for data of Pb (SPG 225), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

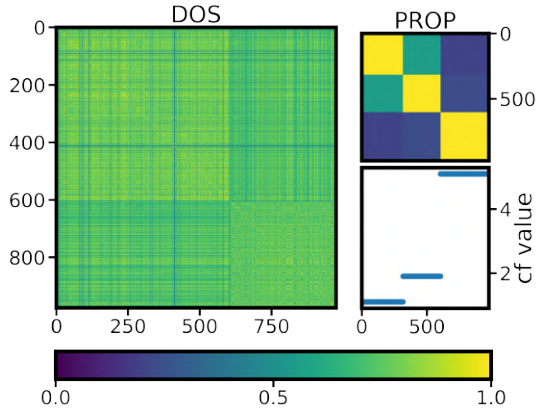
Upload IDs of the calculations: ’0-sUPGbCT4WTrNYF4-xkZQ’, ’1CFTsyFoRHuDrpYmuEwYCw’, ’1sJnuLHzTgqKJTy2ED4Ccg’, ’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’, ’5A1BV5LBTbGfuvTFh6F0GQ’, ’7jo7hwe4TZWwSCP_V6n11A’, ’AGeD22S7TuCM0TGpN8uAoQ’, ’ATy4zekAQgW-Nlu-0-87Vg’, ’CbcV5E5sT56tBx1zm02_sQ’, ’FjiYy5-ETRCrs5ktzJA55w’, ’L_LSn0amQlKS6Q1DCni4nQ’, ’MF3kN7zlT-q942wC1G33Jw’, ’N6lSJSvYTo6nSBOMq5K5IA’, ’PB4nB-jBRa-BrN896n0fWw’, ’ROc3S3MuRsyMxvN1LTRF1Q’, ’SgUOMRORQ4m6FIm–Z56UA’, ’VzIO2WGIQ1aHwevmKvjjyQ’, ’_2nyO9ClQEWCl30Uad7Rpg’, ’_ijarO1hQCqsTpOyb62WXw’, ’b5BQDtBcQFKChE0PsrjIAw’, ’bwWmKAYoTR6kfCLL6bBxtA’, ’e-UOMrI-QDGModmBnxIaBA’, ’eJexi2o0RGmZ-HvagQ0DMg’,

'gH4QxKwDRtWU1c5offnNNA', 'gp_Q_YgVS06Ape6wZkQSaA',
'hiSoHoH4QOel0HmSbYs58Q', 'pyUZ9lChTd-4CYT93YNl1A', 'sG-MnEDjSpO07_g94jnc7g',
'uCSlNMEWQfOS5TMvsQXz1A', 'yVWKAcmyREmV2z1ZBKF0SQ',
'ydzASitvSpiyji3lsQg1gw'

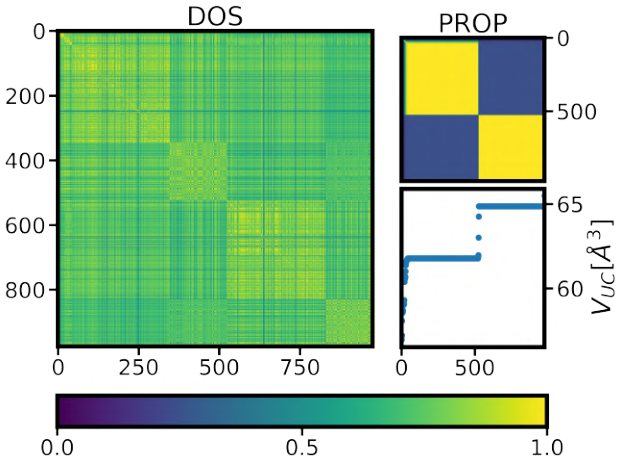
Pd, SPG 225



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



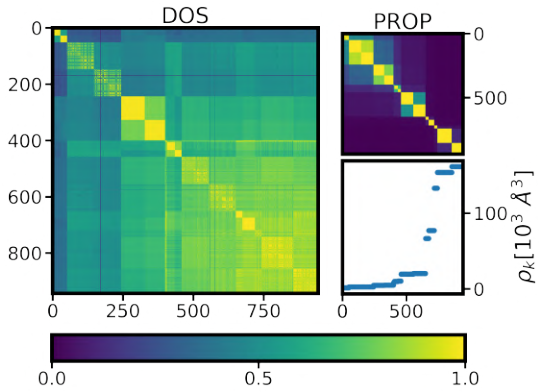
(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 33: Similarity matrices, for data of Pd (SPG 225), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

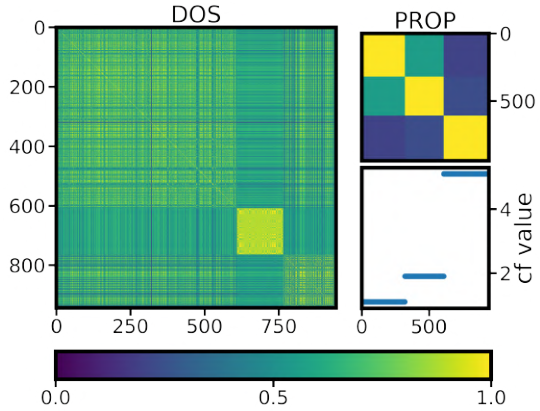
Upload IDs of the calculations: '0-sUPGbCT4WTrNYF4-xkZQ',
'1CFTsyFoRHuDrpYmuEwYCw', '1sJnuLHzTgqKJTy2ED4Ccg',
'3KZBVXpdQ2au-USEG3FwKw', '4AKh_pFmSbyUoWphewbpqQ',
'4wllXGjiRsaP2PmYX_y3dQ', '5A1BV5LBTbGfuvTFh6F0GQ',
'6ITh58J5QcGWUDsmpR9t2A', '7jo7hwe4TZWwSCP_V6n11A',
'AGeD22S7TuCM0TGpN8uAoQ', 'ATy4zekAQgW-Nlu-0-87Vg',

'Aoh5s6VqSN-zvF9w6G3F4g', 'CbcV5E5sT56tBx1zm02_sQ', 'FjiYy5-ETRCrs5ktzJA55w',
'HNZ153qjRRuc0h6A7fdxjg', 'HWoNRk1GRN-6CT6rIVlhlg',
'L_LSn0amQlKS6Q1DCni4nQ', 'MF3kN7zlT-q942wC1G33Jw',
'PB4nB-jBRa-BrN896n0fWw', '_2nyO9ClQEWCl30Uad7Rpg',
'_ijarO1hQCqsTpOyb62WXw', 'b5BQDtBcQFKChE0PsrjIAw',
'e-UOMrI-QDGModmBnxIaBA', 'eJexi2o0RGmZ-HvagQ0DMg',
'gH4QxKwDRtWU1c5offnNNA', 'gp_Q_YgVS06Ape6wZkQSaA',
'gr1xwRnQQ8yBfXUxkg3clw', 'hiSoHoH4QOel0HmSbYs58Q', 'sG-MnEDjSpO07_g94jnc7g',
'uCSlNMEWQfOS5TMvsQXz1A'

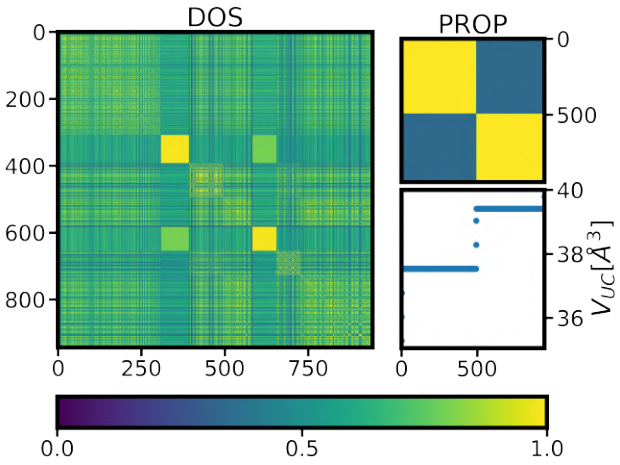
Po, SPG 221



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.

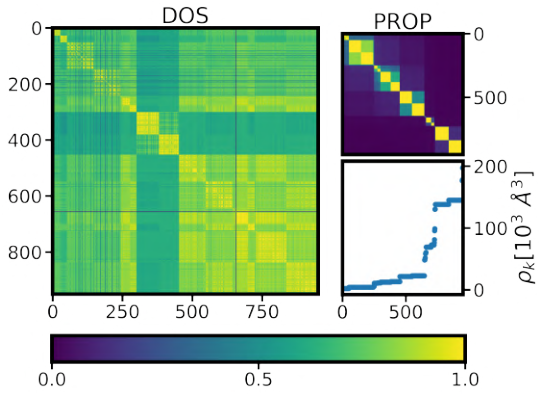


(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

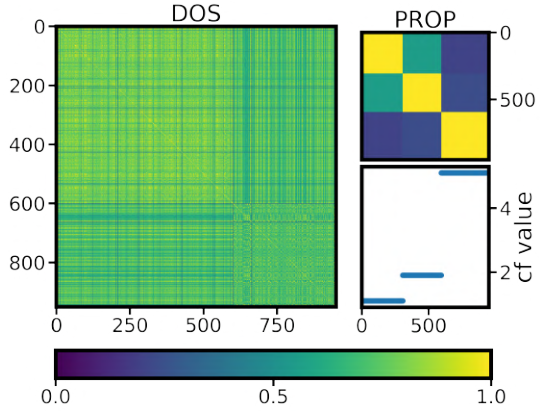
Figure 34: Similarity matrices, for data of Po (SPG 221), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

Upload IDs of the calculations: ’1CFTsyFoRHuDrpYmuEwYCw’, ’1sJnuLHzTgqKJTy2ED4Ccg’, ’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’, ’5A1BV5LBTbGfuvTFh6F0GQ’, ’7jo7hwe4TZWwSCP_V6n1lA’, ’AGeD22S7TuCM0TGpN8uAoQ’, ’CbcV5E5sT56tBx1zm02_sQ’, ’L_LSn0amQlKS6Q1DCni4nQ’, ’MF3kN7zlT-q942wC1G33Jw’, ’_ijarO1hQCqsTpOyb62WXw’, ’b5BQDtBcQFKChE0PsrjIAw’, ’e-UOMrI-QDGModmBnxIaBA’, ’eJexi2o0RGmZ-HvagQ0DMg’, ’gp_Q_YgVS06Ape6wZkQSaA’, ’hiSoHoH4QOel0HmSbYs58Q’, ’jV3NXo1fT0aYQmwv5MNDww’, ’uCSlNMEWQfOS5TMvsQXz1A’

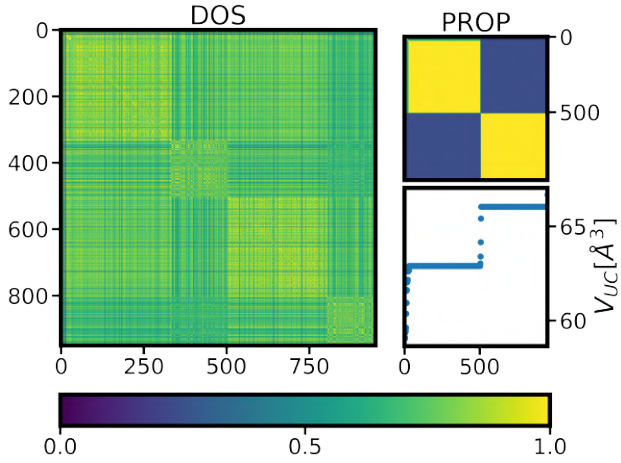
Pt, SPG 225



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.

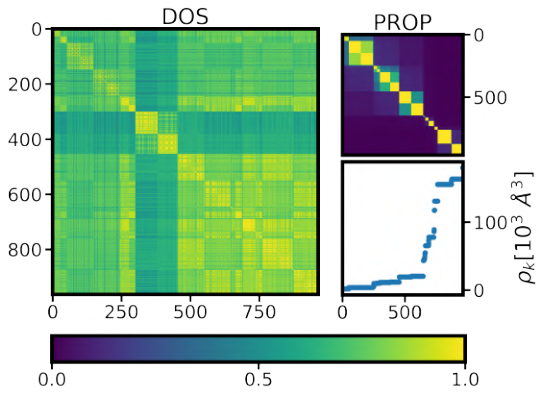


(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

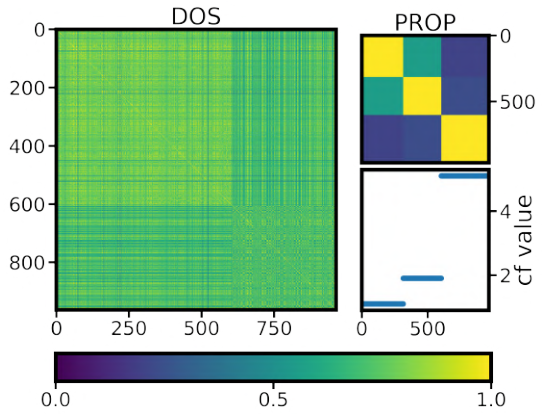
Figure 35: Similarity matrices, for data of Pt (SPG 225), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

Upload IDs of the calculations: ’0-sUPGbCT4WTrNYF4-xkZQ’,
’1CFTsyFoRHuDrpYmuEwYCw’, ’1sJnuLHzTgqKJTy2ED4Ccg’,
’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’,
’5A1BV5LBTbGfuvTFh6F0GQ’, ’6ITh58J5QcGWUDsmpR9t2A’,
’7jo7hwe4TZWwSCP_V6n1lA’, ’AGeD22S7TuCM0TGpN8uAoQ’,
’ATy4zekAQgW-Nlu-0-87Vg’, ’CbcV5E5sT56tBx1zm02_sQ’, ’E9GGbOh1SZGp1Fbf5Ptw_w’,
’FjiYy5-ETRCrs5ktzJA55w’, ’HNZ153qjRRuc0h6A7fdxjg’, ’L_LSn0amQlKS6Q1DCni4nQ’,
’MF3kN7zlT-q942wC1G33Jw’, ’MpyeQdIrSCejxvvDXRINpA’,
’_2nyO9ClQEWCl30Uad7Rpg’, ’_ijarO1hQCqsTpOyb62WXw’,
’b5BQDtBcQFKChE0PsrjIAw’, ’e-UOMrI-QDGModmBnxIaBA’,
’eJexi2o0RGmZ-HvagQ0DMg’, ’hiSoHoH4QOel0HmSbYs58Q’,
’pyUZ9lChTd-4CYT93YNl1A’, ’sG-MnEDjSpO07_g94jnc7g’,
’uCSlNMEWQfOS5TMvsQXz1A’

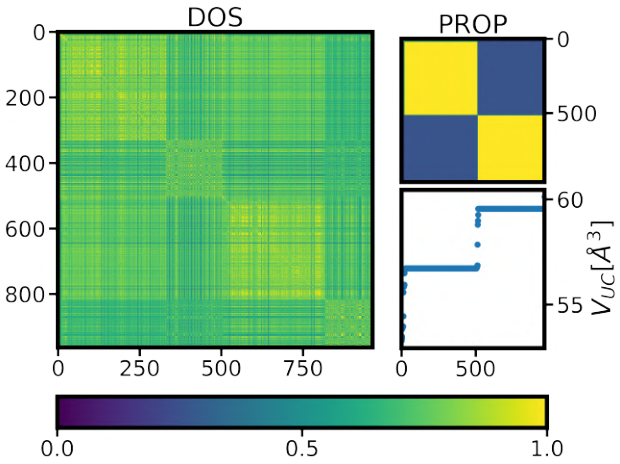
Rh, SPG 225



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



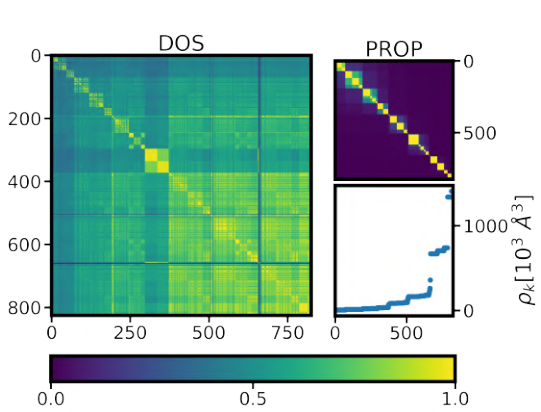
(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 36: Similarity matrices, for data of Rh (SPG 225), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

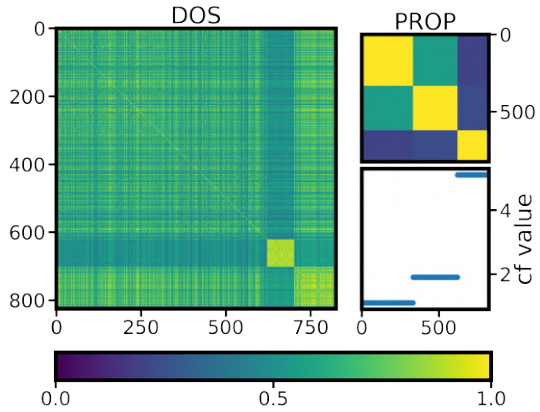
Upload IDs of the calculations: ’0-sUPGbCT4WTrNYF4-xkZQ’, ’1CFTsyFoRHuDrpYmuEwYCw’, ’1sJnuLHzTgqKJTy2ED4Ccg’, ’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’, ’4wllXGjiRsaP2PmYX_y3dQ’, ’5A1BV5LBTbGfuvTFh6F0GQ’, ’6ITh58J5QcGWUDsmpR9t2A’, ’7jo7hwe4TZWwSCP_V6n1lA’, ’AGeD22S7TuCM0TGpN8uAoQ’, ’ATy4zekAQgW-Nlu-0-87Vg’, ’CbcV5E5sT56tBx1zm02_sQ’, ’FjiYy5-ETRCrs5ktzJA55w’, ’HNZ153qjRRuc0h6A7fdxjg’, ’L_LSn0amQlKS6Q1DCni4nQ’, ’MF3kN7zlT-q942wC1G33Jw’, ’Mfeibg97TeyfmIaqseUM_Q’, ’SgUOMRORQ4m6FIm–Z56UA’, ’_2nyO9ClQEWCl30Uad7Rpg’, ’_ijarO1hQCqsTpOyb62WXw’, ’b5BQDtBcQFKChE0PsrjIAw’, ’e-UOMrI-QDGModmBnxIaBA’, ’eJexi2o0RGmZ-HvagQ0DMg’,

'gH4QxKwDRtWU1c5offnNNA', 'gp_Q_YgVS06Ape6wZkQSaA',
'gr1xwRnQQ8yBfXUxkg3clw', 'hiSoHoH4QOel0HmSbYs58Q',
'uCSlNMEWQfOS5TMvsQXz1A'

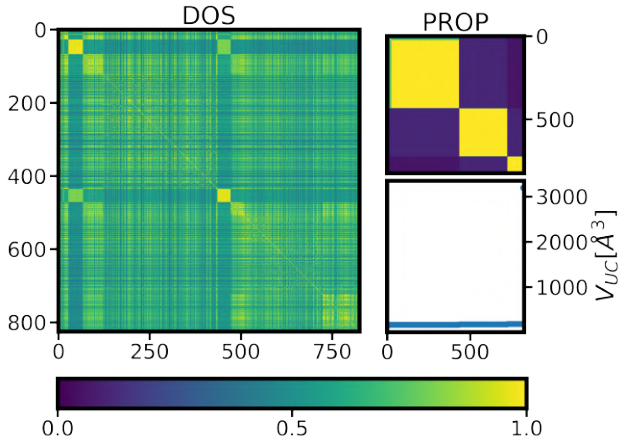
Si, SPG 227



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



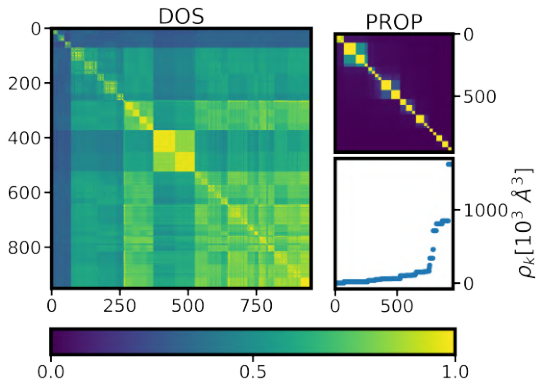
(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 37: Similarity matrices, for data of Si (SPG 227), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

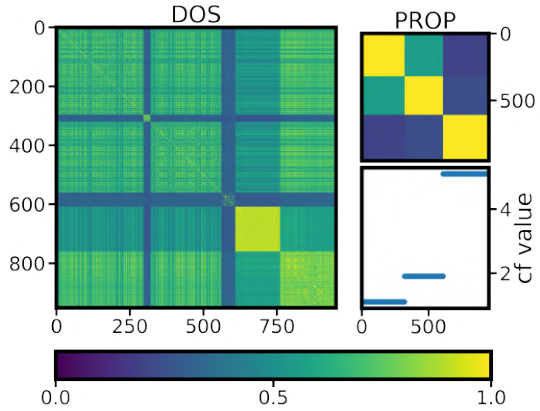
Upload IDs of the calculations: '-6tKmQOSTwe-noXqMyPXBg',
'-n9rmBXfSsKFd8CHgDEIyg', '1CFTsyFoRHuDrpYmuEwYCw',
'1sJnuLHzTgqKJTy2ED4Ccg', '3KZBVXpdQ2au-USEG3FwKw',
'4AKh_pFmSbyUoWphewbpqQ', '5A1BV5LBTbGfuvTFh6F0GQ',
'7jo7hwe4TZWwSCP_V6n1lA', 'AGeD22S7TuCM0TGpN8uAoQ',
'ATy4zekAQgW-Nlu-0-87Vg', 'CbcV5E5sT56tBx1zm02_sQ', 'F6FpfGPiRSaYxoQqgfNL3A',
'FjiYy5-ETRCrs5ktzJA55w', 'HNZ153qjRRuc0h6A7fdxjg', 'IxN_SXI0QlqRv6sdGLULfA',
'L_LSn0amQlKS6Q1DCni4nQ', 'MF3kN7zlT-q942wC1G33Jw',

'QWmDYCHTQpSM3sRylXg8qA', 'SgUOMRORQ4m6FIm–Z56UA',
'U7d-AtBqR6a-my433ppqZg', '_2nyO9ClQEWCl30Uad7Rpg', '_ijarO1hQCqsTpOyb62WXw',
'b5BQDtBcQFKChE0PsrjIAw', 'dcDrKssEQVCwXF3aAEqaVA',
'e-UOMrI-QDGModmBnxIaBA', 'eJexi2o0RGmZ-HvagQ0DMg',
'gH4QxKwDRtWU1c5offnNNA', 'hiSoHoH4QOel0HmSbYs58Q',
'l7r48FczQ6qy6BhRwYyP0g', 'mAAfBZKmTQWDVL1ZPRsrVQ',
'mJGWdEW2RJyE3mIWCknwjw', 'uCSlNMEWQfOS5TMvsQXz1A'

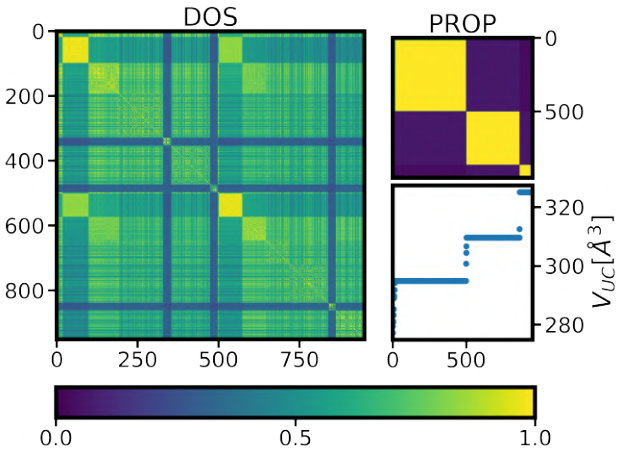
Sn, SPG 227



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



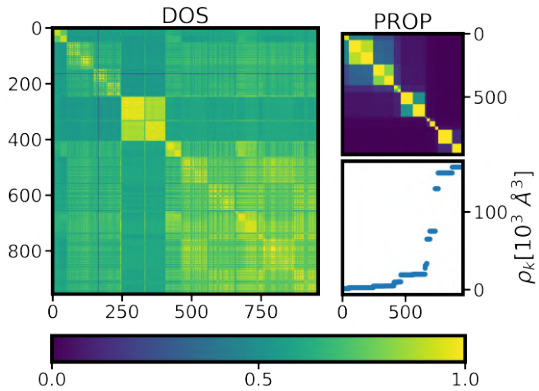
(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 38: Similarity matrices, for data of Sn (SPG 227), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

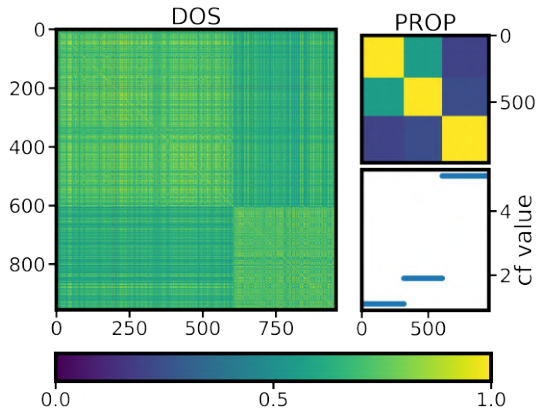
Upload IDs of the calculations: '1CFTsyFoRHuDrpYmuEwYCw',
'1lW8mBSbTii8wIBpaQaRtg', '1sJnuLHzTgqKJTy2ED4Ccg',
'3KZBVXpdQ2au-USEG3FwKw', '4AKh_pFmSbyUoWphewbpqQ',

'5A1BV5LBTbGfuvTFh6F0GQ', '7jo7hwe4TZWwSCP_V6n11A',
'AGeD22S7TuCM0TGpN8uAoQ', 'ATy4zekAQgW-Nlu-0-87Vg',
'CbcV5E5sT56tBx1zm02_sQ', 'F6FpfGPiRSaYxoQqgfNL3A', 'FjiYy5-ETRCrs5ktzJA55w',
'I84GYdt4R5yEDBQJqZm9_g', 'L_LSn0amQlKS6Q1DCni4nQ',
'MF3kN7zlT-q942wC1G33Jw', '_2nyO9ClQEWCl30Uad7Rpg',
'_ijarO1hQCqsTpOyb62WXw', 'b5BQDtBcQFKChE0PsrjIAw',
'e-UOMrI-QDGModmBnxIaBA', 'eJexi2o0RGmZ-HvagQ0DMg',
'gr1xwRnQQ8yBfXUxkg3clw', 'hiSoHoH4QOel0HmSbYs58Q',
'mKzhnavVRtS2prPAzg9_Ug', 'q-yMaQ9IQ3Cdu3hHGmrQtA',
'uCSlNMEWQfOS5TMvsQXz1A'

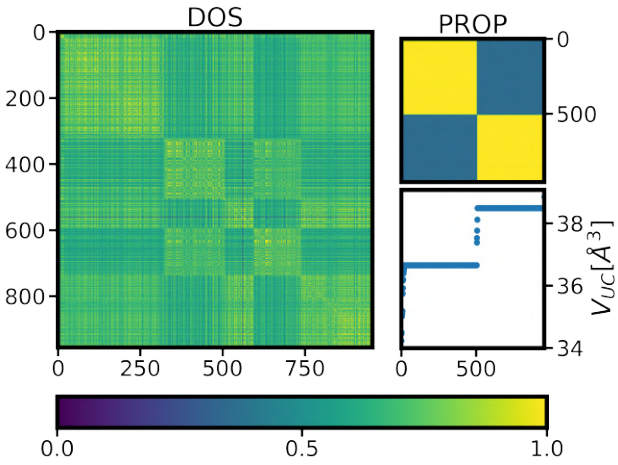
Ta, SPG 229



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.

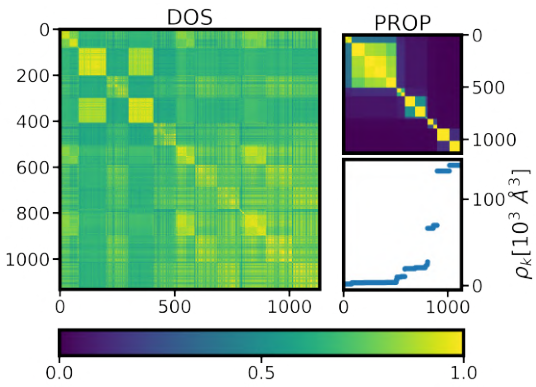


(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

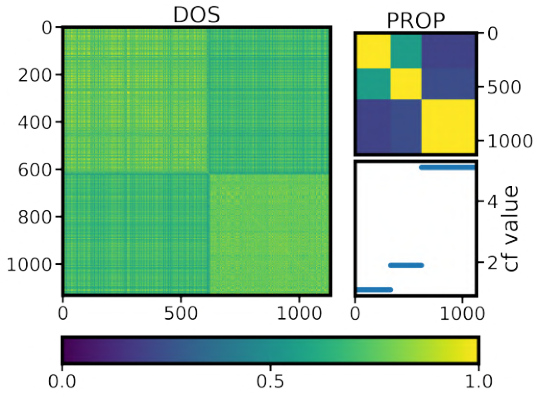
Figure 39: Similarity matrices, for data of Ta (SPG 229), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

Upload IDs of the calculations: ’0-sUPGbCT4WTrNYF4-xkZQ’,
’1CFTsyFoRHuDrpYmuEwYCw’, ’1sJnuLHzTgqKJTy2ED4Ccg’,
’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’,
’5A1BV5LBTbGfuvTFh6F0GQ’, ’7jo7hwe4TZWwSCP_V6n11A’,
’AGeD22S7TuCM0TGpN8uAoQ’, ’CbcV5E5sT56tBx1zm02_sQ’,
’FjiYy5-ETRCrs5ktzJA55w’, ’L_LSn0amQlKS6Q1DCni4nQ’, ’MF3kN7zlT-q942wC1G33Jw’,
’SgUOMRORQ4m6FIm–Z56UA’, ’_2nyO9ClQEWCl30Uad7Rpg’,
’_ijarO1hQCqsTpOyb62WXw’, ’b5BQDtBcQFKChE0PsrjIAw’,
’e-UOMrI-QDGModmBnxIaBA’, ’eJexi2o0RGmZ-HvagQ0DMg’,
’gH4QxKwDRtWU1c5offnNNA’, ’hiSoHoH4QOel0HmSbYs58Q’,
’mAAfBZKmTQWDVL1ZPRsrVQ’, ’qHFCH1d4QZmNMcFCdIJduA’,
’uCSlNMEWQfOS5TMvsQXz1A’

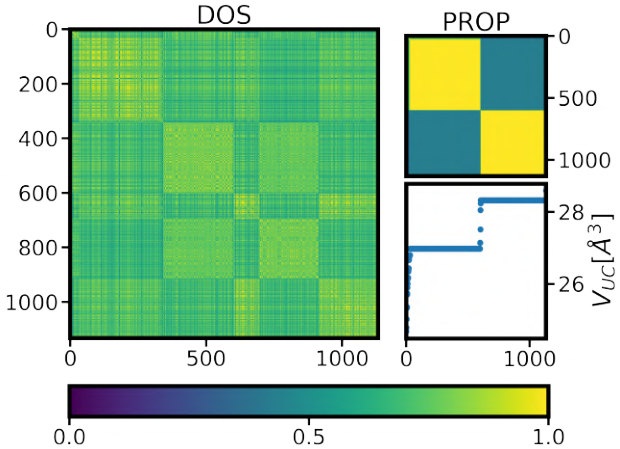
V, SPG 229



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.

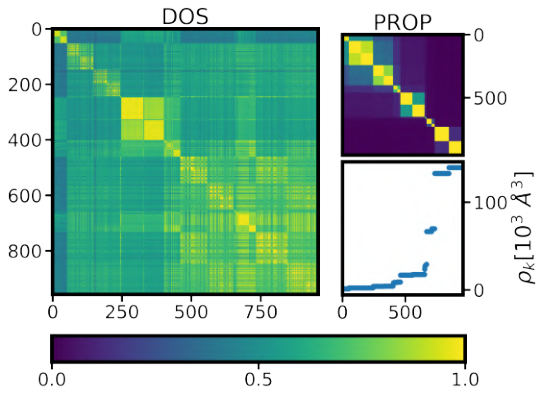


(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

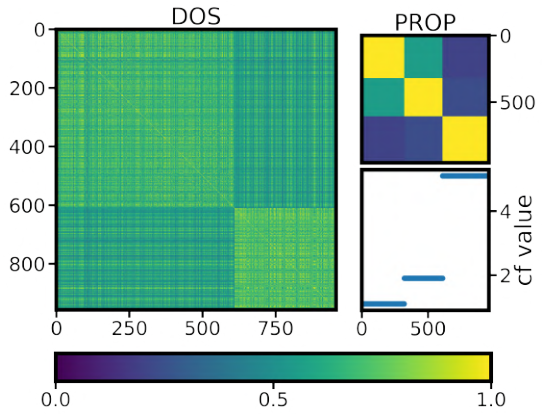
Figure 40: Similarity matrices, for data of V (SPG 229), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

Upload IDs of the calculations: ’0-sUPGbCT4WTrNYF4-xkZQ’,
’1CFTsyFoRHuDrpYmuEwYCw’, ’1sJnuLHzTgqKJTy2ED4Ccg’,
’3KZBVXpdQ2au-USEG3FwKw’, ’4AKh_pFmSbyUoWphewbpqQ’,
’5A1BV5LBTbGfuvTFh6F0GQ’, ’7jo7hwe4TZWwSCP_V6n1lA’,
’AGeD22S7TuCM0TGpN8uAoQ’, ’ATy4zekAQgW-Nlu-0-87Vg’,
’CbcV5E5sT56tBx1zm02_sQ’, ’FjiYy5-ETRCrs5ktzJA55w’, ’HNZ153qjRRuc0h6A7fdxjg’,
’L_LSn0amQlKS6Q1DCni4nQ’, ’MF3kN7zlT-q942wC1G33Jw’,
’SPF7ivRdSQu6Rg3G3GuCDw’, ’Sl2QOupKQ52F2NIqWhcmmg’,
’_2nyO9ClQEWCl30Uad7Rpg’, ’_ijarO1hQCqsTpOyb62WXw’, ’b590n2LESkaPZceushAfAg’,
’b5BQDtBcQFKChE0PsrjIAw’, ’e-UOMrI-QDGModmBnxIaBA’,
’eJexi2o0RGmZ-HvagQ0DMg’, ’gH4QxKwDRtWU1c5offnNNA’,
’gp_Q_YgVS06Ape6wZkQSaA’, ’gr1xwRnQQ8yBfXUxkg3clw’, ’hXjxFrSARi6vr8ZsckYj7Q’,
’hiSoHoH4QOel0HmSbYs58Q’, ’uCSlNMEWQfOS5TMvsQXz1A’

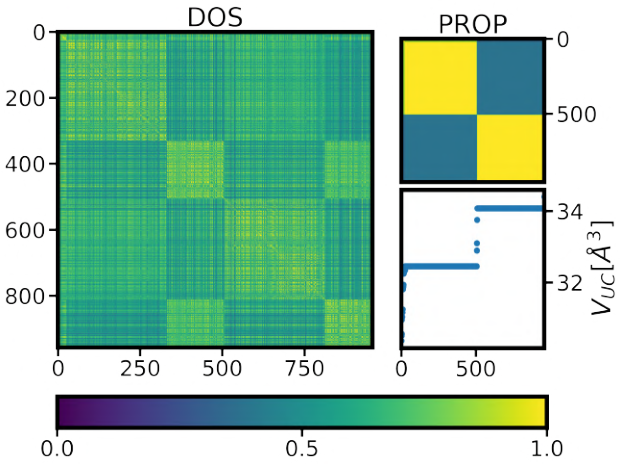
W, SPG 229



(a) Similarity matrices sorted by the composite property ρ_k, showing the correlation between DOS similarity and composite property similarity



(b) Similarity matrices sorted by the cf-value, describing the DFT code and XC functional used to compute each DOS. Similarity matrices show the correlation between the DOS similarity and the cf-value.



(c) Similarity matrices sorted by the unit cell volume, showing the correlation between the DOS similarity and unit cell volume.

Figure 41: Similarity matrices, for data of W (SPG 229), obtained from XC functionals LDA and GGA, and DFT codes FHI-aims and VASP. Each subfigure shows the analysis of an individual property.

Upload IDs of the calculations:'0-sUPGbCT4WTrNYF4-xkZQ',
'1CFTsyFoRHuDrpYmuEwYCw', '1sJnuLHzTgqKJTy2ED4Ccg',
'3KZBVXpdQ2au-USEG3FwKw', '4AKh_pFmSbyUoWphewbpqQ',
'5A1BV5LBTbGfuvTFh6F0GQ', '7jo7hwe4TZWwSCP_V6n1lA',
'AGeD22S7TuCM0TGpN8uAoQ', 'ATy4zekAQgW-Nlu-0-87Vg',
'CbcV5E5sT56tBx1zm02_sQ', 'FjiYy5-ETRCrs5ktzJA55w', 'HNZ153qjRRuc0h6A7fdxjg',
'L_LSn0amQlKS6Q1DCni4nQ', 'MF3kN7zlT-q942wC1G33Jw',
'Oi8t6MmSTFiJaNuovbyG-Q', 'SgUOMRORQ4m6FIm–Z56UA',
'_2nyO9ClQEWCl30Uad7Rpg', '_ijarO1hQCqsTpOyb62WXw',
'b5BQDtBcQFKChE0PsrjIAw', 'e-UOMrI-QDGModmBnxIaBA',
'eJexi2o0RGmZ-HvagQ0DMg', 'gH4QxKwDRtWU1c5offnNNA',

'hXjxFrSARi6vr8ZsckYj7Q', 'hiSoHoH4QOel0HmSbYs58Q', 'mKzhnavVRtS2prPAzg9_Ug', 'uCSlNMEWQfOS5TMvsQXz1A'

7 References

[1] K. Lejaeghere, V. Van Speybroeck, G. Van Oost, and S. Cottenier. Error estimates for solid-state density-functional theory predictions: An overview by means of the ground-state elemental crystals. *Critical Reviews in Solid State and Materials Sciences*, 39(1):1–24, 2014.

[2] Cormac Toher, Corey Oses, David Hicks, Eric Gossett, Frisco Rose, Pinku Nath, Demet Usanmaz, Denise C. Ford, Eric Perim, Camilo E. Calderon, Jose J. Plata, Yoav Lederer, Michal Jahnátek, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kevin Rasch, Roman V. Chepulskii, Richard H. Taylor, Geena Gomez, Harvey Shi, Andrew R. Supka, Rabih Al Rahal Al Orabi, Priya Gopal, Frank T. Cerasoli, Laalitha Liyanage, Haihang Wang, Ilaria Siloi, Luis A. Agapito, Chandramouli Nyshadham, Gus L. W Hart, Jesús Carrete, Fleur Legrain, Natalio Mingo, Eva Zurek, Olexandr Isayev, Alexander Tropsha, Stefano Sanvito, Robert M. Hanson, Ichiro Takeuchi, Michael J. Mehl, Aleksey N. Kolmogorov, Kesong Yang, Pino D'Amico, Arrigo Calzolari, Marcio Costa, Riccardo De Gennaro, Marco Buongiorno Nardelli, Marco Fornari, Ohad Levy, and Stefano Curtarolo. The aflow fleet for materials discovery, 2017.

[3] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin a. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

[4] The NOMAD (Novel Materials Discovery) Center of Excellence (CoE). `https://nomad-coe.eu`.

[5] Olexandr Isayev, Denis Fourches, Eugene N. Muratov, Corey Oses, Kevin Rasch, Alexander Tropsha, and Stefano Curtarolo. Materials cartography: Representing and mining materials space using structural and electronic fingerprints. *Chemistry of Materials*, 27(3):735–743, 2015.

[6] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[7] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.

[8] Volker Blum, Ralf Gehrke, Felix Hanke, Paula Havu, Ville Havu, Xinguo Ren, Karsten Reuter, and Matthias Scheffler. Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications*, 180(11):2175 – 2196, 2009.

[9] Andris Gulans, Stefan Kontur, Christian Meisenbichler, Dmitrii Nabok, Pasquale Pavone, Santiago Rigamonti, Stephan Sagmeister, Ute Werner, and Claudia Draxl. exciting: a full-potential all-electron package implementing density-functional theory and many-body perturbation theory. *Journal of Physics: Condensed Matter*, 26(36):363202, aug 2014.

[10] Vienna ab initio simulation package. `https://www.vasp.at/`.

[11] M. Kuban *et al.* Density-of-states similarity descriptor for unsupervised learning on materials databases. To be published.

[12] Zdenek Hubálek. Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews*, 57:669 – 689, 01 2008.

[13] T. T. Tanimoto. *An elementary mathematical theory of classification and prediction by T.T. Tanimoto*. International Business Machines Corporation New York, 1958.

[14] M. Methfessel and A. T. Paxton. High-precision sampling for brillouin-zone integration in metals. *Phys. Rev. B*, 40:3616–3621, Aug 1989.

[15] Jolyon Aarons, Misbah Sarwar, David Thompsett, and Chris-Kriton Skylaris. Perspective: Methods for large-scale density functional calculations on metallic systems. *The Journal of Chemical Physics*, 145(22):220901, 2016.

[16] Kurt Lejaeghere, Gustav Bihlmayer, Torbjörn Björkman, Peter Blaha, Stefan Blügel, Volker Blum, Damien Caliste, Ivano E. Castelli, Stewart J. Clark, Andrea Dal Corso, Stefano de Gironcoli, Thierry Deutsch, John Kay Dewhurst, Igor Di Marco, Claudia Draxl, Marcin Dułak, Olle Eriksson, José A. Flores-Livas, Kevin F. Garrity, Luigi Genovese, Paolo Giannozzi, Matteo Giantomassi, Stefan Goedecker, Xavier Gonze, Oscar Grånäs, E. K. U. Gross, Andris Gulans, François Gygi, D. R. Hamann, Phil J. Hasnip, N. A. W. Holzwarth, Diana Iuşan, Dominik B. Jochym, François Jollet, Daniel Jones, Georg Kresse, Klaus Koepernik, Emine Küçükbenli, Yaroslav O. Kvashnin, Inka L. M. Locht, Sven Lubeck, Martijn Marsman, Nicola Marzari, Ulrike Nitzsche, Lars Nordström, Taisuke Ozaki, Lorenzo Paulatto, Chris J. Pickard, Ward Poelmans, Matt I. J. Probert, Keith Refson, Manuel Richter, Gian-Marco Rignanese, Santanu Saha, Matthias Scheffler, Martin Schlipf, Karlheinz Schwarz, Sangeeta Sharma, Francesca Tavazza, Patrik Thunström, Alexandre Tkatchenko, Marc Torrent, David Vanderbilt, Michiel J. van Setten, Veronique Van Speybroeck, John M. Wills, Jonathan R. Yates, Guo-Xu Zhang, and Stefaan Cottenier. Reproducibility in density functional theory calculations of solids. *Science*, 351(6280), 2016.

[17] Webelements. `https://www.webelements.com/`.

[18] Osb, spacegroup: 187, aflow page, aflow id: b771a4d99e7ddec1. `http://aflow.org/material/?id=aflow:b771a4d99e7ddec1`.

[19] Cs, spacegroup: 229, aflow page, aflow id: 5acbf212d1783298. `http://aflow.org/material/?id=aflow:5acbf212d1783298`.

[20] Sn, spg 227. `https://materialsproject.org/materials/mp-117`.

[21] Ir, spg 225. `https://materialsproject.org/materials/mp-101/`.

Selbstständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig verfasst und noch nicht für andere Prüfungen eingereicht habe. Sämtliche Quellen einschließlich Internetquellen, die unverändert oder abgewandelt wiedergegeben werden, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, sind als solche kenntlich gemacht. Mir ist bekannt, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuchs bzw. Täuschung eingeleitet wird.

Berlin, 02.07.2021 .....