Computational Materials Science 187 (2021) 110103

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

# MATCOR, a program for the cross-validation of material properties between databases

Jorge Marquez Chavez, Boris Kiefer *

*Department of Physics, New Mexico State University, MSC 3D, 1255 N Horseshoe Dr., Las Cruces, NM 88003, USA*

ARTICLE INFO

*Keywords:*
Materials databases
Data mining
Cross-correlation
Materials
Material properties

ABSTRACT

Data analytics approaches are increasingly often used to facilitate property-specific materials discovery. The uncertainties in these approaches can be greatly affected by the fidelity of the data sets that are used to train the data models. Therefore, data curation is an essential step for obtaining well-constrained model predictions. This can be a challenging task, especially for data sets that are too large for human quality control. We developed MATCOR, an open source, user-friendly, easily adaptable software to facilitate the data curation process. MATCOR processes lists of material identifiers in either AFLOW or Materials Project format and searches for the best matching materials entry in the other database. This is a non-trivial task due to differences in labeling and/or non-unique usage of material labels. MATCOR uses a combination of characteristics such space group, compound formula, crystal structure and use of Hubbard-U to provide the best possible comparison between databases. The capabilities of MATCOR are demonstrated for density, elastic properties, magnetic properties, and band gap correlations between AFLOW and Materials Project. We find that density shows the highest correlation among the tested properties, 93% of verified densities agree to within ±2%. Bulk- and shear-moduli showed deviations of less than ±10% for 80.6% and 65.1% of the materials, respectively. The classification of materials as non-magnetic/paramagnetic and metallic/gapped are consistent among the two databases for 91% and 69% of the materials, respectively. These examples show that MATCOR can be used to automate and thereby accelerate the data curation process prior to materials discovery through data analytical models.

## 1. Introduction

The advance of science and engineering is accompanied by an ever-increasing amount of data. This data is preserved and archived in databases and dictionaries. In materials science, for example, databases such as the Crystallography Open Database [1] and ICSD [2] provide mainly experimentally obtained crystallographic information for a wide range of materials. This mainly experimentally driven effort is mirrored by computational efforts to provide libraries of material properties for experimentally observed, as well as theoretically predicted, materials. Open-source databases include AFLOW (Automatic FLOW for Materials discovery, [3]), MP (Materials Project, [4]), OQMD (Open Quantum Materials Database, [5]), and NOMAD (Novel Materials Discovery, [6]). These repositories provide a convenient starting point for materials design and discovery through data mining [7,8], machine learning [9], and materials informatics [10]. However, as time progresses and experimental, computational/theoretical approaches and analysis techniques advance, one may find several data entries in a database for the same material or (not unexpectedly) entries for a material in different databases. If large datasets are retrieved from these databases, it is unlikely that every single datum can be validated individually, yet material verification errors may be significant. For smaller datasets intra- and inter-database biases are amplified and may affect the reliability of predictions as obtained from data mining and data analytics more immediately. In this contribution, we address the challenging issue of data curation and data cross-validation between databases, and introduce MATCOR (MATerials CORrelation), a software to automate inter-database property validation by matching materials meta-information and crystal structure. For a reference database material identifier, the software returns the corresponding material identifier and property value of the structurally best match in the target database. The output also reports the pool of candidate materials that were considered in the verification process. Section 2 describes details of the MATCOR implementation and in Section 3 we provide examples and statistics of material correlations between MP and AFLOW for lists of 100 input materials for density, elasticity, magnetism, and bandgap.

* Corresponding author.
*E-mail addresses:* jorgemar@nmsu.edu (J. Marquez Chavez), bkiefer@nmsu.edu (B. Kiefer).

https://doi.org/10.1016/j.commatsci.2020.110103
Received 25 June 2020; Received in revised form 8 October 2020; Accepted 9 October 2020
Available online 20 October 2020

## 2. Methods

Our MATerials CORelation software (MATCOR) integrates available online query tools for the automated cross-correlation/validation of material properties between databases. For the purpose of illustration of the capabilities of MATCOR we focus on the widely used AFLOW and MP databases and their respective remote access and search capabilities [11,12]. However, MATCOR is designed for adaptation for cross-correlation of other pairs of databases as long as they follow REST protocols [13,14] for keyword-driven queries. Ideally, each material should possess a unique identifier, such as an ICSD structure tag, and it should be a trivial task to compare material properties between databases. However, this brute-force strategy fails since several database entries may refer to the same ICSD identifier, or the desired identifier is not easily accessible. The former observation is illustrated in Table 1, for an array of possible ICSD identifiers that is returned for a single material. Therefore, the canonical strategy to use cross-database compatible identifiers fails at present. As a corollary, database entries and ICSD identifiers are no longer uniquely related. For almost every entry, MP provides an ICSD identifier (MP ID) for the crystal structure from which the database entry is derived. However, when an MP ID is queried, generally an array of ICSD identifiers for structurally similar entries is returned (Table 1). Similarly, AFLOW contains many additional materials entries beyond the ICSD catalog and ICSD IDs may not be available or are stored in different sections of the AFLOW library.

Moreover, AFLOW does not provide the same type of post-processing for each material entry, hindering a direct comparison with MP material properties. This means that the post-processing of a material entry may lack certain properties (i.e. elasticity); yet, these properties may be stored and accessible through another entry with the same structure. Therefore, it is necessary to employ a more complex validation strategy. Since crystallographic information is used to generate the input structures, a suitable starting point is querying material attributes such as space group and number of atoms in the unit cell. This strategy, however, unfortunately fails as well. For example, $FeNi_3$ crystallizes in space group Pm-3m; #221 [15], a space group that has 14 site symmetries *(a-n)* [16], where some of the sites have the same multiplicities, in the present example: *12 h*, *12i*, and *12j*, providing the same composition for the same space group but different structures. Thus, individual space group related tags alone, while having merits, are insufficient to consistently enforce unique structure comparisons. Similarly, database-specific digital object identifiers are provided for each database entry, but it remains unclear how to translate those between databases. Therefore, MATCOR facilitates the identification of outliers and common trends among databases that may otherwise bias or hinder materials discovery.

The implementation of MATCOR overcomes these challenges by first searching meta-information such as composition, space group, and consistent Hubbard-U usage to identify a pool of candidate materials in the entire target database. After ranking the materials in this pool by magnitude difference of the property of interest, the pymatgen structure matcher utility [12] is used to identify the structure with the best matching property. The materials in the pool for each queried material as well as the best match are reported in the result file and tagged for easy filtering to extract desired information. Fig. 1 shows the typical output of MATCOR for a successful materials comparison for density. The meta-tagging of the result section allows different levels of filtering and provides information about Hubbard-U usage and structure matching. In the present case, we queried *mp-570673* for the material in AFLOW with the best matching density. The result file shows its space group and chemical formula (line 1), and a pool of two candidate materials with filtering details (lines 2 to 3). The best match is displayed in line 5 together with the result of structure matching (SV//NSV), the property difference, and indicates whether Hubbard-U (hubbard_U) was used in the calculation, including the following possibilities: HUB-HUB (yes-yes), DFT-DFT (no-no), HUB-DFT (yes–no), DFT-HUB (no-yes). In summary, for the present case, the pool of candidate structures consisted of two materials. For the best matching material, neither MP nor AFLOW used Hubbard-U in the calculation (DFT-DFT), and the reference and target structure are found to be structurally identical (SV). The property difference between reference and best matching material is 0.0367 g/$cm^3$.

**Table 1**
Multiple ICSD entries for single materials, such as β-$AuTi_3$ in MP.

| Materials Project | |
|---|---|
| Material Identifier | mp-1786 |
| Available ICSD | 612419, 612405, 58605, 612417, 612420, 612418 |

### 2.1. Design of the MATCOR data cross correlation/validation python module

The developed Python 3.8+ modules take as command line input the name of a file that contains a list of valid material identifiers in either AFLOW or MP format (one per line, Table 2) and determines the materials ID and property values of candidate and best-matching entry in the other database. The name of the result file must be specified as command line argument #2.

All other runtime parameters are provided in one location at the beginning of the main script (Fig. 2): The user must provide the MP API ID, and due to the possible non-uniqueness of the cross-validation the matching "*base_property*" keywords for the two databases must be provided. For MP, it may be necessary for the user to specify two base properties ("*base_property_MP_1*" and "*base_property_MP_2*"). This is because, for certain properties, MP may return a dictionary (i.e. elasticity) of several "sub- properties" (G_VRH, K_VRH, etc.); in order to access a sub-property (corresponding to *base_property_MP_2*), a *base_property_MP_1* must be specified first. The program provides an option for the user to automatically send an email to notify MP and AFLOW database maintainers if the cross-validated property differs in magnitude by more than a user-specified difference "*max_difference.*" By default, this option is commented out. Moreover, an "OUTLIER" tag is added to the entries for which its cross-validated difference exceeds the *max_difference.* The user may also change the tolerance parameters ("*tol_in*", "*stol_in*", "*angle_tol_in*") for the pymatgen structure matcher; by default, these are set to 0.2, 0.3, and 5°.

After these minimal adjustments, it is easiest to execute MATCOR in the MP pymatgen [12] environment:

```
python3 MATCOR.py reference_id_list result_file
```

During the execution of the program, MATCOR first searches the entire target database to identify pools of candidate materials with matching chemical formula, space group, and the use of Hubbard-U filtering (yes/yes, or no/no). The materials in the pool are tagged and reported in the result file (Fig. 1). The result file clearly identifies if verification is based on a consistent Hubbard-U usage (yes/yes or no/no). If no Hubbard-U consistent verification is possible, MATCOR ignores Hubbard-U and allows for yes/no and no/yes comparisons. Both possibilities are clearly tagged in the result file. After ranking materials in the pool by magnitude difference from the reference property value, the pymatgen structure matcher is used as a final verification step in the comparison. If the structures are identical (within tolerances), the result is reported as the best match and tagged as "SV" (Fig. 1). If the structures are not identical, the next best structure in the pool is investigated until a match is achieved. If no structural match is found in the pool, the entry is reported as "NSV" in the result file. Depending on the user-defined tolerances (Fig. 2), the structure matching between entries and the cross-validation of properties itself can vary. The obtained results are not necessarily unique, but define a pool of materials for property matching, that is tagged and reported in the result file. If the *base_property* selected

```
pool_0_0: mp-570673                P6/mmm #191                          None          ScNi5
pool_0_1: aflow:69d3f9ef55dcd065   P6/mmm #191,P6/mmm #191              None
pool_0_2: aflow:14bcd4bcafc55bf7   P6/mmm #191,P6/mmm #191              None
input id                           hubbard_U   SV//NSV   property_py   output material          property_AFLOW   min_difference
pool_best_0: ScNi5   mp-570673     DFT-DFT     SV        7.3353       aflow:69d3f9ef55dcd065   7.2986           0.0367
```

**Fig. 1.** Sample output of materials verification through MATCOR.

**Table 2**
Example MATCOR input materials ID list.

| AFLOW | MP |
|---|---|
| aflow:a94cbd0169e5270f | mp-1017543 |
| aflow:7cb8e9e9facd592b | mp-1358 |
| aflow:0de15907838a9be4 | mp-1761 |
| aflow:70189be7ddab0598 | mp-1010071 |

```
materials = ['mp-1007855']
mpr = MPRester('your API ID')
base_property_MP _1= 'base_property'
base_property_MP _2= ''
base_property_AFLOW = 'base_property'
max_difference = 0.0
ltol _in= 0.2
stol _in= 0.3
angle_tol_in = 5
```

**Fig. 2.** Input deck for correlation analysis.

by the user or provided by either database is not a number, the filtering consists of composition, space group, Hubbard-U and structure matching. Cases where no structure, composition, or space group matches are found at all within the pool of candidate materials are reported as "NO_MATCH" in the result file. The general MATCOR workflow is outlined in Fig. 3.

In both search directions, the structure matcher function of pymatgen [12] is used. MATCOR verifies that structures are formatted correctly for the pymatgen structure matcher utility and fixes formatting issues prior to structure matching, mainly due to a possibly missing line of chemical species on line 6 of the VASP formatted structure files (CONTCAR/POSCAR).

### 2.1.1. Data-correlation: AFLOW to MP

MATCOR takes a list of AFLOW identifiers (Table 2), searches for the closest consistent property match in the MP database (schematic workflow is shown in Fig. 4) and writes the results of the comparisons to a file with a name that is specified as command line argument #2. First, the module determines the number of atoms and chemical formula for an AFLOW material ID (*auid*), and determines Hubbard-U usage (yes: HUB; no: DFT). Once these attributes have been obtained, the chemical formula is reduced to the smallest meaningful unit. For example, $Fe_8Re_{12}$ (*auid: aflow:340c69078949ff30*) is reduced to [Fe2, Re3]. Then, the program generates all possible combinations of the elements/abundances, to cope with database-related differences in the ordering of chemical species. In the example, a list ['Fe2Re3', 'Re3Fe2'] with two entries is generated; each of these candidate formulas is queried in MP until it returns a match (if no match is found, the AFLOW structure has no equivalent in MP and after reporting this finding in the result file, MATCOR moves to the next entry in the input list). Once a match has been obtained, the "*pretty_formula*" tag in MP is generated and used to query material properties. Afterwards, this module searches for an entry in MP with the same space group as the input *auid* and utilizes Hubbard-U filtering. It is important to note that AFLOW uses two sets of space groups; "sg" and "sg2" for the space group of the compound and space group of the relaxed compound, and each set contains three possible options: the space group of the material before the first, after the first, and after the last relaxation of the calculation. In order to maximize the cross-validation and considering that space groups differ for some materials, MATCOR considers if the space group of any compound in MP matches the space group before the first or after the last relaxation for both sets. Finally, the reference and target structures are reported in the result file and tagged as "SV" if the structures match. If more than one material is found, MATCOR selects the material with the best matching *base_property*

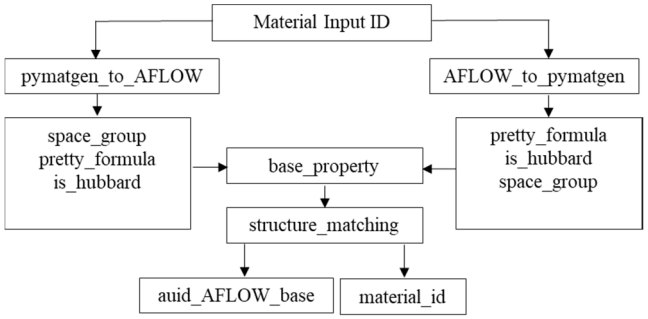


**Fig. 3.** Schematic workflow for MATCOR database cross validation software.

### 2.1.2. Data-correlation: MP to AFLOW

MATCOR takes a list of identifiers in MP format from a file specified as command line argument #1 (Table 2) and queries for the closest matching material in AFLOW (schematic workflow is shown in Fig. 5). The results are written to a file with a name specified as command line argument #2. The main difference from the AFLOW to MP search (Section 2.1.1) is the order in which the pool of possible matches is filtered; the entries are narrowed down first by space group, then by formula, and finally by Hubbard-U.

We added a quick search option that allows to limit correlation to a specific AFLOW library (i.e ICSD). This can be accomplished by adding "catalog = desired catalog" within the MP to AFLOW module inside the "*search()*" function of the "*result_aflow_rep*" variable. The location of the needed change is clearly specified and by default commented out.

If more than one entry in AFLOW is obtained from the initial filtering method, the program follows the same process as in the AFLOW to MP module by searching for the *auid* with the closest matching *base_propety* value to the MP reference entry and uses the pymatgen structure matcher to verify matching crystal structures.

One exception is considered for this search direction: deprecated or updated materials identifiers. The material identifiers in MP can change over time. If an old and hence unrecognizable material entry is encountered, MATCOR will output a warning and check the next entry in the input materials ID list.

One final exception concerns the recovery of the original specified material IDs for MP to AFLOW followed by AFLOW to MP. In principle, the results should be the same identifier as specified in the original MP list. This is correct if material entries are uniquely correlated between the two databases through the *base_property*. However, this may not be the case. The following example demonstrates this observation: for material *mp-972808* ($E_{gap}$ = 5.568 eV) and using band gap as the *base_property*, MATCOR finds *aflow:170838e139988cb3* ($E_{gap}$ = 5.39 eV). If this *auid*, is used as the input ID, *mp-972808* is correctly found in the pool of possible materials, but MATCOR will ultimately determine *mp-*

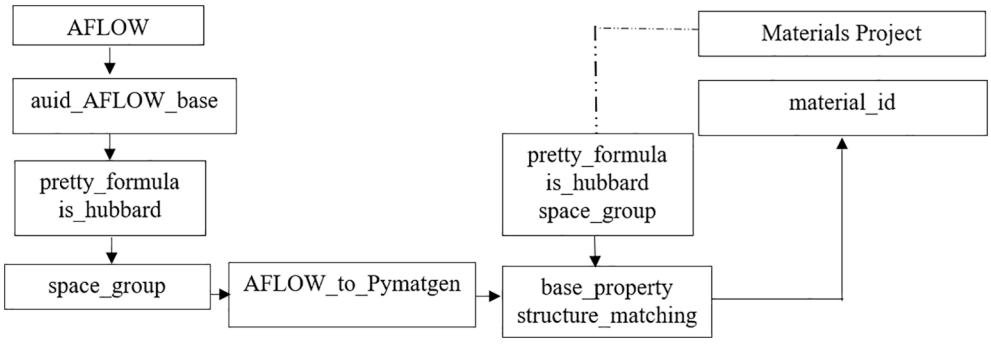


**Fig. 4.** Schematic workflow for AFLOW to MP cross-correlation.

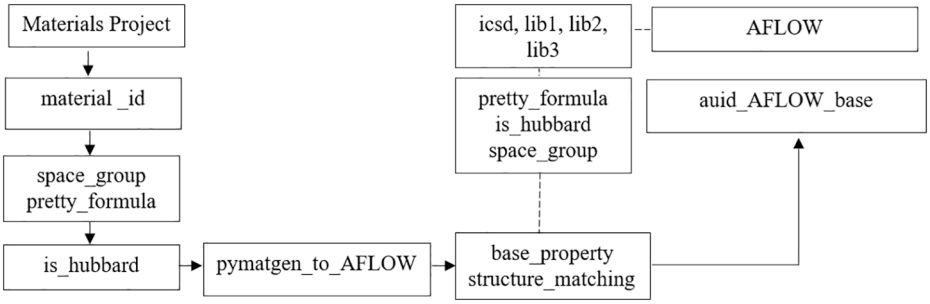


**Fig. 5.** Schematic workflow for MP to AFLOW cross-correlation.

*555891* ($E_{gap}$ = 5.397 eV) as the best match because this material has a band gap value closest to the input *auid*.

## 3. Discussion and examples

Using the MATCOR search strategy and exception handling, we can efficiently query entries across databases and identify best matching materials as part of a data curation process. Here we demonstrate the simplicity and versatility of MATCOR with a few examples. We selected specific properties that are available in both databases. In examples 1–3, we queried a list of 100 bimetallic alloys from MP in AFLOW and recovered their density, mechanical, and magnetic properties. In example 4, we compared the band gaps of 100 oxides. The complete listing of compounds and validation results as well as the source code can be found on the GitHub MATCOR page.

For all four examples, we follow the general MATCOR workflow (Figs. 3–5), with minor modifications of the source code to adjust the *base_property* (Fig. 2). The input list of material identifiers is provided as a command-line argument (one materials ID per line; examples are provided in Tables 1–3) and the name of the result file is specified as the second command line argument, as explained above.

### 3.1. Example 1: density

The first example uses MATCOR to compare densities, a property that is found in AFLOW and MP for every material, since structure optimization is part of the computational protocols used to generate the databases [3,4]. Density is fundamental to the identification of lightweight materials, with applications in aerospace [17] and in the transportation sector to reduce fuel consumption and to reduce greenhouse gas emissions [18].

All 100 input materials could be found in AFLOW. The recovered densities cluster around the 45° equal value line (Fig. 6) with a correlation coefficient of R = 0.99, showing the excellent agreement between

**Table 3**
Example outputs.

| Material | MP | Hubbard | Property | AFLOW | Hubbard | Property |
|---|---|---|---|---|---|---|
| Density (g/cm$^3$) | | | | | | |
| $FeNi_3$ | mp-1007855 | DFT | 8.71 | aflow:69e41e3bdec0a652 | DFT | 8.71 |
| MgO | mp-1265 | DFT | 3.47 | aflow:b0bd25030f17306c | DFT | 3.47 |
| $SiO_2$ | mp-7000 | DFT | 2.49 | aflow:300974e2e8b1b7c4 | DFT | 2.49 |
| $Rb_2O$ | mp-1394 | DFT | 3.82 | aflow:9df378b18977f7f4 | DFT | 3.82 |
| $HfO_2$ | mp-550893 | DFT | 10.70 | aflow:a5b27b615a8a8211 | DFT | 10.70 |
| $PaO_2$ | mp-2364 | DFT | 10.75 | aflow:613f851a6f86dd5b | DFT | 10.75 |
| Elasticity ($G_{VRH}$ and $K_{VRH}$ in GPa) | | | | | | |
| $Ti_3Au$ | mp-1786 | DFT | $G_{VRH}$ = 71 | aflow:6809e154de355a4c | HUB | $G_{VRH}$ = 52.3 |
| $Ti_3Au$ | mp-1786 | DFT | $K_{VRH}$ = 139 | aflow:6809e154de355a4c | HUB | $K_{VRH}$ = 78.2 |
| Magnetism (M in $\mu_B$) | | | | | | |
| $Fe_9Co_7$ | mp-601842 | DFT | M = 36.7 | aflow:7f73ffebb07b2d90 | DFT | M = 36.6 |
| Band gap ($E_g$ in eV) | | | | | | |
| SnO | mp-2097 | DFT | $E_g$ = 0.63 | aflow:6da067cb48a0edde | DFT | $E_g$ = 0.32 |
| MgO | mp-1265 | DFT | $E_g$ = 4.64 | aflow:e590f0da1597c30f | DFT | $E_g$ = 4.50 |

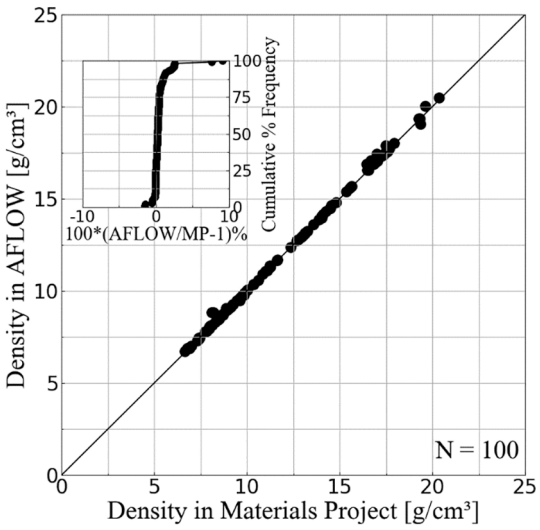


**Fig. 6.** Density correlations between AFLOW and MP. Point-by-point comparison of data points. Inset: cumulative statistics of point-to-point percentage differences.

the recovered material densities in the two databases. All percentage differences are within −2% and 10%, (Fig. 6, inset), and 93% deviate by less than ±2%. All structures were structurally verified. Hubbard-U consistent (HUB-HUB, DFT-DFT) was successful for 92% of the tested materials. The remaining 8% were compared for inconsistent Hubbard-U usage (HUB-DFT and DFT-HUB).

### 3.2. Example 2: elastic properties

Elastic properties are of broad interest, i.e. some models of ferroelectricity rely on knowledge of (soft) elastic directions to rationalize spontaneous, reversible charge displacement [19], and combinations of aggregate shear and bulk moduli have been used to predict new superhard compounds [20]. Here, we use the Voigt-Reuss-Hill average shear modulus ($G_{VRH}$) and Voigt-Reuss-Hill average bulk modulus ($K_{VRH}$) as examples [21]. The necessary changes can be accomplished by editing the property selection (Fig. 2). In this case, 'elasticity' is set as the *base_property_MP_1* and either '*G_VRH*' or '*K_VRH*' as *base_property_MP_2*. Similarly, either '*ael_bulk_modulus_vrh*' or '*ael_bulk_modulus_vrh*' is used for *base_property_AFLOW*. A sample output is shown in Table 3.

The cross-validated data for mechanical properties (Fig. 7) show a larger variation than density (Fig. 6). 63/100 materials could be verified, for the remaining 37 materials only one of the databases provided a property value, simply reflecting that elasticity has not yet been computed for all materials in the databases. We also found one material for which the bulk modulus values differed significantly, discussed below. Based on the remaining bimetallic alloys, $G_{VRH}$ values (N = 63) show a correlation coefficient of R = 0.79 (Fig. 7a), and R = 0.86 for $K_{VRH}$ values (N = 62; Fig. 7b), both lower compared to the correlation coefficient for density. $G_{VRH}$ ($K_{VRH}$) differences varied between −50% to +34% (−45% to +30%). In both cases, percentage differences cover a wider but more symmetric range compared to the density example (Fig. 7 insets). We find that for 65.1% (41/63) and 80.6% (50/62) of the verified materials the magnitude percentage differences for $G_{VRH}$ and $K_{VRH}$ are less than 10%, respectively. We found one case ($UIr_2$, Fd-3 m) where $K_{VRH}$ deviates significantly $K_{VRH}$ (AFLOW) = 545 GPa (*aflow:f958850a678420df*) and $K_{VRH}$ (MP) = 209.9 GPa (*mp-1655*). We explored the origin of this difference with additional VASP [22,23] density-functional-theory computations. Using the same Hubbard-U as AFLOW, we reproduce the elastic constants to within ~5% and uranium finite magnetic moments. In contrast, we could not reproduce the MP elastic constant tensor that was obtained without Hubbard-U usage. We

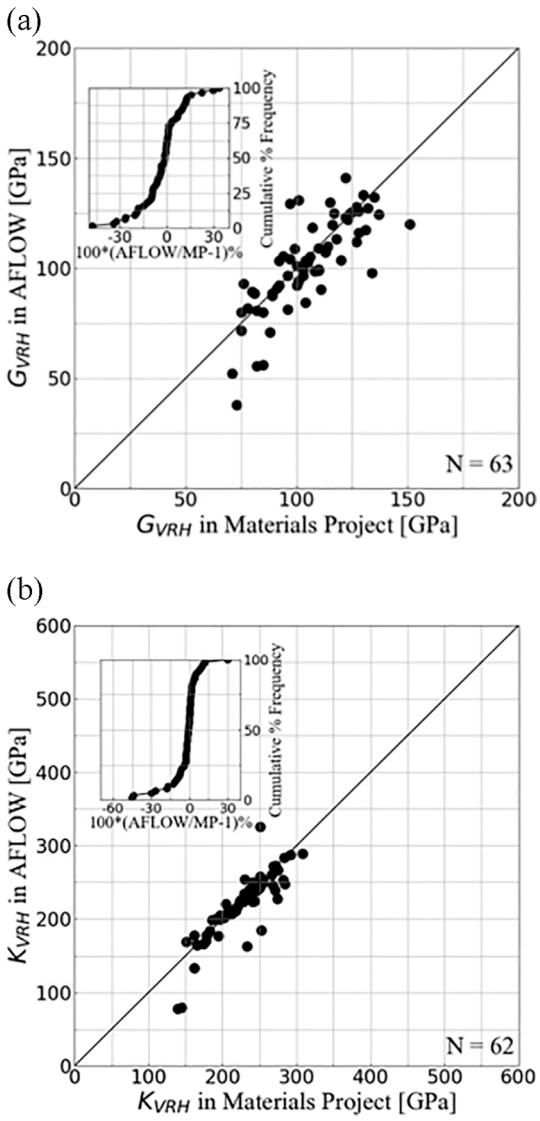


**Fig. 7.** Cross-correlation/validation of aggregate shear/bulk-modulus as obtained from our analysis. a) Shear modulus, GVRH; b) bulk modulus, KVRH. Inset: cumulative statistics of point-to-point percentage differences. Only materials for which AFLOW and MP both provide shear and bulk modulus are shown.

note that the MP equilibrium structure is non-magnetic and tentatively attribute the observed differences in bulk modulus to the presence or absence of magnetism.

### 3.3. Example 3: magnetism

Magnetic properties in materials find a broad range of applications, for example in the areas of multiferroics [24], topological materials [25], skyrmions [26], and merons [27]. Other examples include the design of permanent magnets for electrical motors and generators [28]. Similarly, magnetic materials are researched for cancer treatment [29], including cell isolation enrichment [30], immunoassay [31], and magnetic resonance imaging [32]. As in the previous examples, necessary changes to the MATCOR program are minimal (Fig. 2). '*total_magnetization*' is set as *base_property_MP_1*. Similarly, '*spin_cell*' is used as *base_property_AFLOW*. The sample output is shown in Table 3.

91/100 materials allowed for Hubbard-U consistent property verification (HUB-HUB, DFT-DFT). Eight materials were Hubbard-U inconsistent (HUB-DFT or DFT-HUB) and are clearly tagged in the result file;

all results were structurally verified. The correlation result shows 22% (22/100) of entries for which both databases predict Hubbard-U consistent paramagnetic behavior. Among these 22 materials, we detected one outlier: FeGe: *mp-22478* (M = 0.36 $\mu_B$), *aflow:bf290af0aa9dee39* (M = 3.88 $\mu_B$), which was detected as Hubbard-U-consistent (DFT-DFT) across both databases. The origin of this difference remains unclear but may be related to the choice of pseudopotentials (MP: Ge-d; AFLOW: Ge_h). The cross-validation of the remaining 21/100 entries is shown in Fig. 8. Overall, we find a high correlation of R = 0.99 with magnetic moment differences varying between −76% and 76%, where 76% (16/21) of the materials fall in a narrow band magnetic moment difference of ±4% (Fig. 8 inset). This low number of matching materials is misleading: among the 78 excluded entries, eight are non-magnetic in MP but magnetic in AFLOW and one was listed as non-magnetic in AFLOW but magnetic in MP. The remaining 69% (69/100) entries are characterized as non-magnetic in both databases. Therefore, both databases provide a consistent categorization of magnetic properties for 91% (22% + 69%) of the tested materials.

### 3.4. Example 4: band gaps

Band gaps are fundamental for the design of novel high-performance solar cells. Material informatics and data mining techniques have proven useful for this process [33]. The study of intermediate band gap compounds, for example, can be accomplished by querying through databases to search for compounds with suitable bandgaps [34]. Similarly, the use of data mining methods has been used to predict bandgaps for new compounds with previously untested chemical compositions [35]. Furthermore, machine learning has been used to predict band gaps for the associated fundamental characterization of materials as metallic or gapped [36].

We queried and verified 100 oxides with MATCOR after making small adjustments to the module (Fig. 2); 'band_gap' is set as the MP base_property_MP_1, and 'Egap' is used as the base_property_AFLOW. For 55% (55/100) of the materials queried, both databases showed non-zero band gaps with a correlation coefficient of R = 0.97 (Fig. 9). Among the 55 successfully verified materials, we found two outliers: ZnO: *mp-1986* ($E_{gap}$ = 0.63 eV), *aflow:bbe19cfc7071c717* ($E_{gap}$ = 1.68 eV)\; CoO: *mp-19128* ($E_{gap}$ = 0.71 eV), *aflow:85df97487acd1f14* ($E_{gap}$ = 1.80 eV). In *mp-1986* (ZnO), only AFLOW uses Hubbard-U; in the case of *mp-19128*, (CoO) AFLOW and MP both use Hubbard-U but U(Co, AFLOW) = 6.0 eV is significantly higher than that used in MP: U(Co, MP) = 3.32 eV. Within the 53 remaining entries, the computed percentage differences

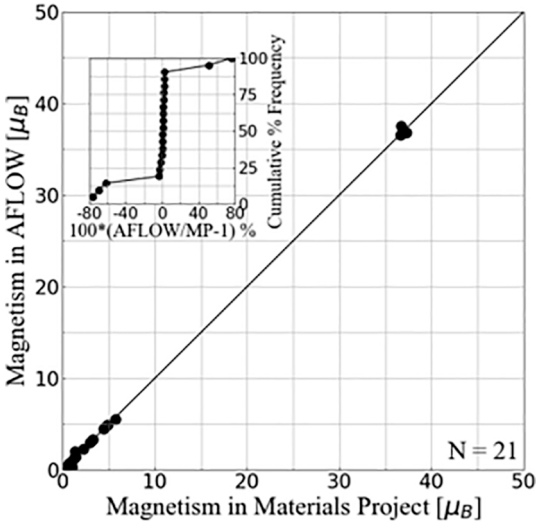


**Fig. 8.** Cross-correlation/validation of cell's magnetic moment. Inset: cumulative statistics of point-to-point percentage differences.

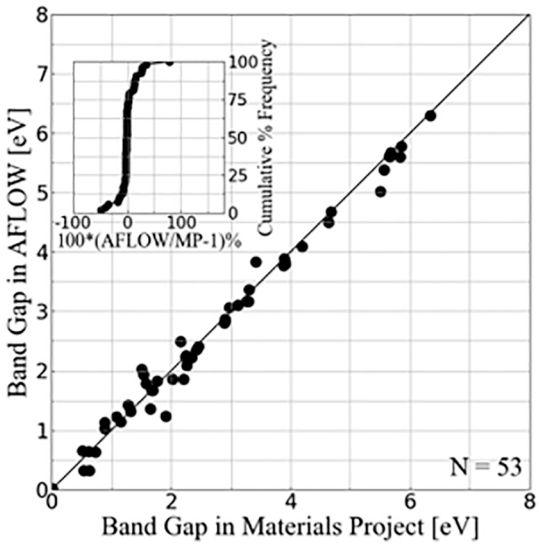


**Fig. 9.** Cross-correlation/validation of band gap. Inset: cumulative statistics of point-to-point percentage differences.

range from −49% to +78%; and 60.4% (32/53) of the gapped materials in both databases differ by less than 5% (Fig. 9 inset). For 14% (14/100) of materials, both databases showed metallic behavior. Therefore, the databases agree for 69% (69/100) of the cases in the classification of a material as gapped or metallic. For 3% (3/100) of the materials either MP or AFLOW predicted inconsistent metallic/gapped behavior. For 23% (23/100) of the entries, MATCOR did not find any matches simply reflecting that the band structure/density of states has not been computed/analyzed in both databases for all materials. Lastly, for 5% (5/100) of materials (*mp-28460, mp-754672, mp-715493, mp-675211, mp-510421*), no structural matches were found in AFLOW ("NSV"), and they are reported as "NO_MATCH".

We would like to emphasize that MATCOR refers to the precision of material properties, *not* their accuracy. For example, for MgO (Table 3) the two databases agree to within 3.03%, but the entries in both databases underestimate the experimentally, well-established larger value of the band gap, $E_{gap}$ = 7.77 eV [37]. The difference between theory and experiment is not an artifact of database design but must be attributed to a deficiency of standard LDA and GGA based DFT [38]. This deficiency can be addressed either empirically [39], or by using more advanced DFT methods such as meta-functionals [40,41], exact-exchange [42,43], or GW quasi-particle approaches [44].

The examples discussed in this section show that material property cross-correlation between databases is a valuable tool for pre-screening datasets for data analytics and machine learning. Using test sets of 100 materials, we find that cross-correlation depends strongly on the queried property. Among the queried material properties, density shows the highest correlation, while elastic, magnetic and optical properties are found to be more database dependent. Therefore, MATCOR can be used to accelerate dataset curation and materials discovery by identifying outliers and common trends between materials in different databases.

## 4. Conclusion

In this contribution, we introduce MATCOR, a new database analysis tool for the cross-correlation/validation of material properties between databases. Major benefits of the software are its user-friendliness and ease of adaptability. For our test sets of 100 compounds, we find that density shows the least variability between the databases. All materials could be verified and for 93% of the materials the densities differed by less than ±2%. Elastic, magnetic and optical properties show a higher variability. For bulk- and shear-modulus, we find that 80.6% and 65.1%

of the values listed in the two databases agree to within ±10%, respectively. A similar analysis for magnetic and optical properties shows that the two databases agree on the categorization of materials as non-magnetic/magnetic and metallic/gapped for 91% and 69% of the materials in the test sets, respectively. Thus, MATCOR can identify outliers and their origins as well as common trends between databases, which can increase the reliability of materials discovery through data analytics and machine learning. The provided cross-correlation is expected to be especially useful in cases where data model predictions are more sensitive to outliers in the input dataset or data sets are too large for the validation of each data entry.

## CRediT authorship contribution statement

**Jorge Marquez Chavez:** Software, Data curation, Validation. **Boris Kiefer:** Writing - original draft, Writing - review & editing, Supervision, Methodology, Data curation, Validation.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

*Acknowledgement*

This work used the Extreme Science and Engineering Discovery Environment (XSEDE) resource Stampede2 at TACC through allocation TG-DMR110093.

*Data availability*

By the time of submission of the final version, we will make available our repository containing the data used to generate the figures in the manuscript as well as an open-source implementation of the MATCOR software available at GitHub (https://github.com/matcor-mc/MATCOR).

## References

[1] S. Gražulis, D. Chateigner, R.T. Downs, A.F.T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, A. Le Bail, Crystallography Open Database – an open-access collection of crystal structures, J. Appl. Crystallogr. 42 (4) (2009) 726–729, https://doi.org/10.1107/S0021889809016690.

[2] G. Bergerhoff, R. Hundt, R. Sievers, I.D. Brown, The inorganic crystal structure data base, J. Chem. Inf. Model. 23 (2) (1983) 66–69, https://doi.org/10.1021/ci00038a003.

[3] S. Curtarolo, W. Setyawan, G.L.W. Hart, M. Jahnatek, R.V. Chepulskii, R.H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M.J. Mehl, H.T. Stokes, D.O. Demchenko, D. Morgan, AFLOW: an automatic framework for high-throughput materials discovery, Comput. Mater. Sci. 58 (2012) 218–226, https://doi.org/10.1016/j.commatsci.2012.02.005.

[4] A. Jain, P. Ong, G. Hautier, W. Chen, D. Gunter, D. Skinner, G. Ceder, K.A. A. Persson, A materials genome approach to accelerating materials innovation, APL Mater. (2013), https://doi.org/10.1063/1.4812323.

[5] S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, S. Rühl, C. Wolverton, The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies, Npj Comput Mater 1 (1) (2015), https://doi.org/10.1038/npjcompumats.2015.10.

[6] C. Draxl, M. Scheffler, NOMAD: the FAIR concept for big data-driven materials science, MRS Bull. 43 (9) (2018) 676–682, https://doi.org/10.1557/mrs.2018.208.

[7] L. Liu, H.L. Zhuang, Computational prediction and characterization of two-dimensional pentagonal arsenopyrite FeAsS, Comput. Mater. Sci. 166 (2019) 105–110, https://doi.org/10.1016/j.commatsci.2019.04.040.

[8] L. Ward, A. Dunn, A. Faghaninia, N.E.R. Zimmermann, S. Bajaj, Q.i. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K.A. Persson, G. J. Snyder, I. Foster, A. Jain, Matminer: an open source toolkit for materials data mining, Comput. Mater. Sci. 152 (2018) 60–69, https://doi.org/10.1016/j.commatsci.2018.05.018.

[9] V. Stanev, C. Oses, A.G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, I. Takeuchi, Machine learning modeling of superconducting critical temperature, Npj Comput Mater 4 (1) (2018), https://doi.org/10.1038/s41524-018-0085-8.

[10] Y. Umeda, H. Hayashi, H. Moriwake, I. Tanaka, Materials informatics for dielectric materials, Jpn. J. Appl. Phys. 57 (11S) (2018) 11UB01, https://doi.org/10.7567/JJAP.57.11UB01.

[11] F. Rose, C. Toher, E. Gossett, C. Oses, M.B. Nardelli, M. Fornari, S. Curtarolo, AFLUX: The LUX materials search API for the AFLOW data repositories, Comput. Mater. Sci. 137 (2017) 362–370, https://doi.org/10.1016/j.commatsci.2017.04.036.

[12] S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K.A. Persson, G. Ceder, Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis, Comput. Mater. Sci. 68 (2013) 314–319, https://doi.org/10.1016/j.commatsci.2012.10.028.

[13] R.H. Taylor, F. Rose, C. Toher, O. Levy, K. Yang, M. Buongiorno Nardelli, S. Curtarolo, A RESTful API for exchanging materials data in the AFLOWLIB.org consortium, Comput. Mater. Sci. 93 (2014) 178–192, https://doi.org/10.1016/j.commatsci.2014.05.014.

[14] S.P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, K.A. Persson, The Materials Application Programming Interface (API): a simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles, Comput. Mater. Sci. 97 (2015) 209–215, https://doi.org/10.1016/j.commatsci.2014.10.037.

[15] Z. Ahmed, J.C. Bevan, Awaruite, iridian awaruite, and a new Ru-Os-Ir-Ni-Fe alloy from the Sakhakot-Qila complex, Malakand Agency, Pakistan, Mineral. Mag. 44 (334) (1981) 225–230, https://doi.org/10.1180/minmag.1981.044.334.17.

[16] T. Hahn, International Tables for Crystallography, vol. A: Space-group symmetry, 2006, doi: 10.1107/9780955360206000100.

[17] M. Peters, C. Leyens, Aerospace and space materials, Mater. Sci. Eng. (2009).

[18] W.J. Joost, Reducing vehicle weight and improving U.S. energy efficiency using integrated computational materials engineering, JOM 64 (9) (2012) 1032–1038, https://doi.org/10.1007/s11837-012-0424-z.

[19] R.D. King-Smith, D. Vanderbilt, First-principles investigation of ferroelectricity in perovskite compounds, Phys. Rev. B 49 (9) (1994) 5828–5844, https://doi.org/10.1103/PhysRevB.49.5828.

[20] A. Mansouri Tehrani, A.O. Oliynyk, M. Parry, Z. Rizvi, S. Couper, F. Lin, L. Miyagi, T.D. Sparks, J. Brgoch, Machine learning directed search for ultraincompressible, superhard materials, J. Am. Chem. Soc. 140 (31) (2018) 9844–9853, https://doi.org/10.1021/jacs.8b02717.

[21] R. Hill, The elastic behaviour of a crystalline aggregate, Proc. Phys. Soc. A 65 (5) (1952) 349–354, https://doi.org/10.1088/0370-1298/65/5/307.

[22] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci. 6 (1) (1996) 15–50, https://doi.org/10.1016/0927-0256(96)00008-0.

[23] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B 59 (3) (1999) 1758–1775, https://doi.org/10.1103/PhysRevB.59.1758.

[24] N.A. Spaldin, Multiferroics beyond electric-field control of magnetism, Proc. R. Soc. A. 476 (2233) (2020) 20190542, https://doi.org/10.1098/rspa.2019.0542.

[25] K. Choudhary, K.F. Garrity, F. Tavazza, High-throughput discovery of topologically non-trivial materials using spin-orbit spillage, Sci. Rep. 9 (1) (2019), https://doi.org/10.1038/s41598-019-45028-y.

[26] S.A. Montoya, S. Couture, J.J. Chess, J.C.T. Lee, N. Kent, D. Henze, S.K. Sinha, M.-Y. Im, S.D. Kevan, P. Fischer, B.J. McMorran, V. Lomakin, S. Roy, E.E. Fullerton, Tailoring magnetic energies to form dipole skyrmions and skyrmion lattices, Phys. Rev. B 95 (2) (2017), https://doi.org/10.1103/PhysRevB.95.024415.

[27] X.Z. Yu, W. Koshibae, Y. Tokunaga, K. Shibata, Y. Taguchi, N. Nagaosa, Y. Tokura, Transformation between meron and skyrmion topological spin textures in a chiral magnet, Nature 564 (7734) (2018) 95–98, https://doi.org/10.1038/s41586-018-0745-3.

[28] J.J. Möller, W. Körner, G. Krugel, D.F. Urban, C. Elsässer, Compositional optimization of hard-magnetic phases with machine-learning models, Acta Mater. 153 (2018) 53–61, https://doi.org/10.1016/j.actamat.2018.03.051.

[29] A. Matsumine, K. Takegami, K. Asanuma, T. Matsubara, T. Nakamura, A. Uchida, A. Sudo, A novel hyperthermia treatment for bone metastases using magnetic materials, Int. J. Clin. Oncol 16 (2) (2011) 101–108, https://doi.org/10.1007/s10147-011-0217-3.

[30] B.D. Plouffe, S.K. Murthy, L.H. Lewis, Fundamentals and application of magnetic particles in cell isolation and enrichment: a review, Rep. Prog. Phys. 78 (1) (2015) 016601, https://doi.org/10.1088/0034-4885/78/1/016601.

[31] L. Gao, J. Zhuang, L. Nie, J. Zhang, Y.u. Zhang, N. Gu, T. Wang, J. Feng, D. Yang, S. Perrett, X. Yan, Intrinsic peroxidase-like activity of ferromagnetic nanoparticles, Nat. Nanotech. 2 (9) (2007) 577–583, https://doi.org/10.1038/nnano.2007.260.

[32] C. Bárcena, A.K. Sra, J. Gao, Applications of magnetic nanoparticles in biomedicine, in, Nanoscale Magn. Mater. Appl. (2009), https://doi.org/10.1007/978-0-387-85600-1_20.

[33] C. Ortiz, O. Eriksson, M. Klintenberg, Data mining and accelerated electronic structure theory as a tool in the search for new functional materials, Comput. Mater. Sci. 44 (4) (2009) 1042–1049, https://doi.org/10.1016/j.commatsci.2008.07.016.

[34] D.J.R. Baquião, G.M. Dalpian, Computational screening of bulk materials with intrinsic intermediate band, Comput. Mater. Sci. 158 (2019) 382–388, https://doi.org/10.1016/j.commatsci.2018.11.030.

[35] P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon, K. Rajan, Informatics-aided bandgap engineering for solar materials, Comput. Mater. Sci. 83 (2014) 185–195, https://doi.org/10.1016/j.commatsci.2013.10.016.

[36] Y. Zhuo, A. Mansouri Tehrani, J. Brgoch, Predicting the band gaps of inorganic solids by machine learning, J. Phys. Chem. Lett. 9 (7) (2018) 1668–1673, https://doi.org/10.1021/acs.jpclett.8b00124.

[37] D.M. Roessler, W.C. Walker, Electronic spectrum and ultraviolet optical properties of crystalline MgO, Phys. Rev. 159 (3) (1967) 733–738, https://doi.org/10.1103/PhysRev.159.733.
[38] L.J. Sham, M. Schlüter, Density-functional theory of the band gap, Phys. Rev. B 32 (6) (1985) 3883–3889, https://doi.org/10.1103/PhysRevB.32.3883.
[39] Á. Morales-García, R. Valero, F. Illas, An empirical, yet practical way to predict the band gap in solids by using density functional band structure calculations, J. Phys. Chem. C 121 (34) (2017) 18862–18866, https://doi.org/10.1021/acs.jpcc.7b07421.
[40] B. Patra, S. Jana, L.A. Constantin, P. Samal, Efficient band gap prediction of semiconductors and insulators from a semilocal exchange-correlation functional, Phys. Rev. B 100 (4) (2019), https://doi.org/10.1103/PhysRevB.100.045147.
[41] F. Tran, P. Blaha, Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential, Phys. Rev. Lett. 102 (22) (2009), https://doi.org/10.1103/PhysRevLett.102.226401.
[42] M.K.Y. Chan, G. Ceder, Efficient band gap prediction for solids, Phys. Rev. Lett. 105 (19) (2010), https://doi.org/10.1103/PhysRevLett.105.196403.
[43] J.P. Perdew, W. Yang, K. Burke, Z. Yang, E.K.U. Gross, M. Scheffler, G.E. Scuseria, T.M. Henderson, I.Y. Zhang, A. Ruzsinszky, H. Peng, J. Sun, E. Trushin, A. Görling, Understanding band gaps of solids in generalized Kohn–Sham theory, Proc. Natl. Acad. Sci. USA 114 (11) (2017) 2801–2806, https://doi.org/10.1073/pnas.1621352114.
[44] A. Fleszar, LDA, GW, and exact-exchange Kohn-Sham scheme calculations of the electronic structure of sp semiconductors, Phys. Rev. B 64 (24) (2001), https://doi.org/10.1103/PhysRevB.64.245204.