[Research Paper] 대한금속 · 재료학회지 (*Korean J. Met. Mater.*), Vol. 58, No. 10 (2020) pp.728-739
DOI: 10.3365/KJMM.2020.58.10.728

# KIST-NOMAD
# - a Repository to Manage Large Amounts of Computational Materials Science Data

**Samuel Boateng[1,3,†], Kwang Ryeol Lee[2,†], Deepika[2], Haneol Cho[2], Kyu Hwan Lee[3,*], and Chansoo Kim[2,*]**

[1]*Department of Nanomaterial Science and Engineering, University of Science and Technology, Daejeon 34113, Republic of Korea*
[2]*Computational Science Research Center, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea*
[3]*Innovation Enterprise Cooperation Center, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea*

**Abstract:** We introduce the Korea Institute of Science and Technology-Novel Materials Discovery (KIST-NOMAD) platform, a materials data repository. We describe its functionality and novel features from an academic viewpoint. It is a data repository designed for computational material science, especially focusing on managing and sharing the results of molecular dynamics simulation results as well as quantum mechanical computations. It consists of three main components: a database, file storage, and web-based front end. The database hosts material properties, which are extracted from the computational results. The front end has a graphical user interface and an open application programming interface, which allow researchers to interact with the system more easily. KIST-NOMAD's panel displays the searched results on a well-organized and research-oriented web page. All the open access data and files are available for downloading in comma-separated value format as well as zipped archives. This automated extraction function was developed by utilizing database parsers and JSON scripts. KIST-NOMAD also has an efficient option to download simulation and computation results on a large-scale. All of the above functions are designed to satisfy academic and research demands, and make high-throughput screening available, while incorporating machine learning for computational material engineering. We finally stress that the repository platform is user-driven and user-friendly. It is clearly designed to follow the modern big-data architecture and re-use principles for scientific data, such as being findable, accessible, and interoperable.

(Received June 25, 2020; Accepted August 6, 2020)

**Keywords:** repository for computational material science, open-access data sharing, scientific data, data extraction

## 1. Introduction

Computational material science (CMS) is a key component of modern materials science. It has advanced over the last few decades thanks to the improvements in computational capabilities as well as the development and availability of commercial and open source codes. Program codes such as VASP [1], quantum espresso [2] and TURBOMOLE [3] enable materials science researchers to perform intensive calculations on high performance computers (HPCs) using CPUs and GPUs. The aim of these calculations, among others, is to identify new high-performance materials, discover new/improved materials, reveal new properties of materials as well as confirm empirical results.

CMS is well suited for very complex and large-scale problems that are far beyond the capabilities of experimental materials science [4]. They are high-throughput simulations, which use quantum-mechanical, density-functional theory and molecular dynamics approaches to solve problems in materials science. They inevitably generate large amounts of input and output files. These files, which accumulate over time, are usually stored on users' local computers and are then discarded.

However, recent high level of interest in data-driven [9-13] and interdisciplinary research [14] efforts have proven that

[†]These authors contributed equally to this work.
- Samuel Boateng: 학생, 이광열·이규환·김찬수: 연구원, Deepika·조한얼: 박사 후 연구원
**Corresponding Author:* Chansoo Kim
[Tel: +82-2-958-6448, E-mail: eau@kist.re.kr]
**Corresponding Author:* Kyu Hwan Lee
[Tel: +82-2-958-6759 , E-mail: biometal@kist.re.kr]

these discarded data files could contain valuable information. That information can be used by not only materials scientists to advance materials research but also researchers in related fields such as data science, bioinformatics, biomaterials and nano-informatics. The data generated from material simulations could be used in different contexts beyond the original scope of the data. This makes it imperative to collect these scattered results files, securely store them in repositories and make them freely available.

The creation of materials data repositories, and research efforts that utilize such data, has given rise to a fourth paradigm in material science research, the he so-called "big data-driven research." With big data-driven materials science, researchers will have almost limitless use of the information and knowledge that can be extracted from stored data files. Machine learning (ML) algorithms such as Artificial Neural Networks (ANN) are some of the leading and reliable approaches to conducting data-driven research in materials science. ML algorithms are capable of extracting knowledge and predicting the materials properties of complex systems in a fast and efficient way, and often produces very accurate results.

The success stories of materials big data repository efforts over the last decade include Automatic – FLOW for Materials Discovery (AFLOWlib) [5], Open Quantum Materials Database (OQMD) [6], Materials Project [7], Materials Cloud [28] and Novel Materials Discovery (NOMAD) [8]. These are multipurpose repositories and include innovative tools such as ML algorithms to manipulate data. Many ML codes have been developed for materials science research, such as the Atomic Energy Network (aenet) [16], DeepMD-Kit [17] and Atomistic Machine-learning Package (AMP) [18]. The availability of materials repositories and the rapid development of reliable and efficient ML tools have increased the popularity of the *'fourth paradigm'* amongst materials scientists. This is supported by the increasing number of data-driven materials research publications [15] and the trend is expected to continue. In this work we explain a novel repository system, KIST-NOMAD, and describe its applicability for CMS. Section 2 introduces its structure, and we describe its utilities in Section 3. In Section 4, there are examples showing its applicability. We summarize in Section 5.

## 2. The Structure of KIST-NOMAD

KIST-NOMAD is a web-based materials data infrastructure designed to support CMS data sharing. All the data in the repository are freely contributed by researchers in the field of computational materials science. The repository is designed to conform to the findable, accessible, interoperable and re-usable (FAIR) [20] principles of data sharing. The data and files in the repository are the results of calculations from codes such as VASP, Gaussian [21], Exciting [29], Octopus [22] and FHI-aims [19]. The repository hosts a database and data files, which are accessed via a web-based front end. Resultant files are stored in two folders and the front end has Graphical User Interfaces (GUIs) and an Application Programming Interface (API). The components of KIST-NOMAD are illustrated in Fig. 2.

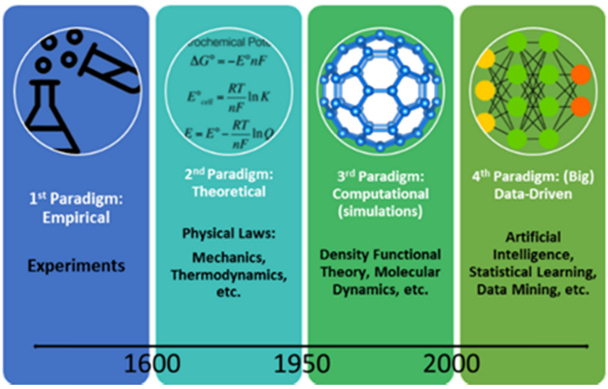


**Fig. 1.** The four paradigms of materials science research: It include the empirical, theoretical, computational and data-driven approaches (adapted from [15]).

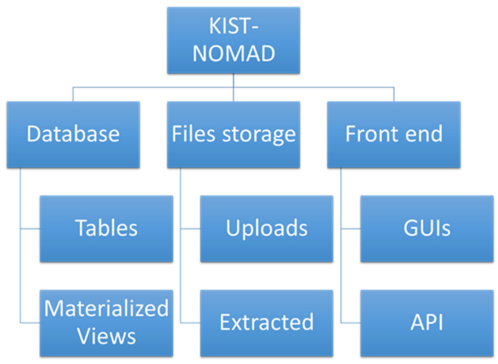


**Fig. 2.** Main components of KIST-NOMAD. Database is weaved as tables in a view of material engineering. The front end consists of GUIs and API.

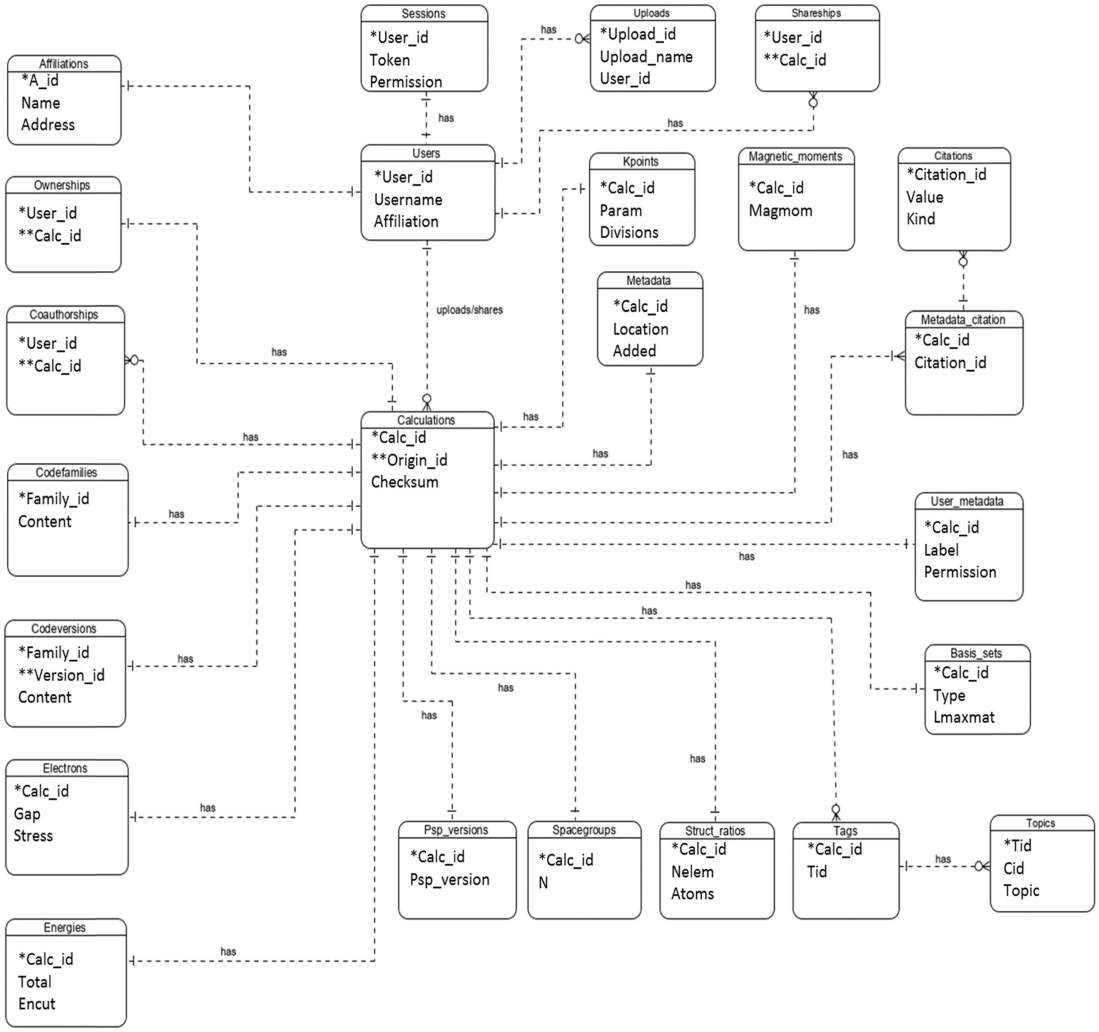


**Fig. 3.** Entity Relationship Diagram of the database. The 24 tables include one-to-many relationships between tables such as users and calculations or one-to-one relationships such as calculations and electrons.

## 2.1. The database

The database is one of the three main components of KIST-NOMAD. It is implemented in PostgreSQL [23] and enables the storage, retrieval, modification and deletion of data. The database has a total of 24 normalized tables and 8 materialized views. The normalized tables ensure that no specific information is saved in more than one table, and helps reduce data redundancy and improve data integrity. Central to the database is the calculations table with the calculation identity (calc_id) primary key. The calc_id is a unique identifier for each calculation, and it is used as a foreign key in most of the other tables to identity other data items that belong to the same calculation. Each row or record in a table represents a unique instance of data. The logical structure of the database is illustrated in an Entity Relationship Diagram (ERD) as in Fig. 3, which shows the database tables and the relationships between them.

*'Materialized views'* are the result of a database query, which is stored as a table. It provides efficient data access and enables faster performance (in turnaround time) for queries. Materialized views have faster performance, because the data in the view is not refreshed at the time of the query request and all of the data presented in the result set are in one materialized view, giving efficient access to requested data.

Details of table names and the data they store are in Table 1. Values of the each table are normalized to avoid redundancy and improve data integrity.

**Table 1.** List of tables in the KIST-NOMAD database and the kind of data stored in them.

| Table Name | Table data |
|---|---|
| Affiliations | A list of research institutes and universities |
| Basis_sets | A list of basis set types |
| Calculations | The ID for each calculation is generated and saved in this table. |
| Citations | A list of calculations citations/references |
| Coauthorships | Details of co-authors |
| Codefamilies | List of materials computation software |
| Codeversions | The versions of the various software applications |
| Electrons | Band gap, band gap type and cell optimization information |
| Energies | Total energy, cut off energy and fermi energy |
| Kpoints | K-points information |
| Magnetic_moments | Magnetic moment information |
| Metadata | Information of the location (full path) of various calculations on disk and other calculation details |
| Metadata_citations | Information of the calculations and their corresponding citations |
| Ownerships | Details of who 'owns' a specific calculation |
| Psp_versions | All calculations pseudo potential versions |
| Sessions | User sessions information |
| Shareships | Shared calculations information |
| Spacegroups | All calculated space groups |
| Struct_ratios | The chemical formula, total number of atoms, number of elements, formula units and volume (if available) |
| Tags | This table contains information of the various tags or labels for each calculation. |
| Topics | All chemical symbols, system types, crystal systems etc. |
| Uploads | Details of all uploaded calculations such as upload name, date of upload, user who uploaded it etc. |
| Users | All user information including user names, passwords and emails |
| User metadata | Calculations access information (restricted or open access) |

### 2.2 The data files

KIST-NOMAD has a large amount of storage. The data files come from the most popular computational science codes. These files are kept in two folders, uploads and extracted. The uploads folder keeps compressed files in the tar.gz format. The compressed files are later extracted and moved to the extracted folder. All of the files in both folders are available for download.

### 2.3. The browser-based front end

KIST-NOMAD has a web implementation interface. It provides a convenient medium for interacting between the database and data files. It has embedded functions and methods to perform various database and file operations. The data search GUI, results GUI, shared calculations, data upload GUI and API are included in this front end.

## 3. KIST-NOMAD Utilization

In this section, we highlight the functionality of KIST-NOMAD as well as its importance. The search GUI, search API, data and files download, and GUI data search and search results are described in detail with examples.

### 3.1 The data search GUI

KIST-NOMAD provides a neat and intuitive data search GUI design with a well laid out sequence of information, which ensures easy navigation. It has both selectable and text input functions. The search GUI has two main sections, the *'Chemical Elements'* and the *'Search Conditions.'*

**1. Chemical Elements** – This shows the periodic table. Any clicked element will fill the element text box in the search conditions section.

**Table 2.** List of crystal systems and their space group numbers. This is based on the symmetry class of the Hermann-Mauguin notation [25].

| Crystal System | Space group number |
|---|---|
| Cubic | 195-230 |
| Hexagonal | 168-194 |
| Monoclinic | 3-15 |
| Orthorhombic | 16-74 |
| Rhombohedral(Trigonal) | 143-167 |
| Tetragonal | 75-142 |
| Triclinic | 1 and 2 |

**Table 3.** The compound types defined in KIST-NOMAD and the number of elements. Currently, calculation results in KIST-NOMAD are from 'Unary' to 'Septenary'

| Compound Type | Number of Elements |
|---|---|
| Unary | 1 |
| Binary | 2 |
| Ternary | 3 |
| Quaternary | 4 |
| Quinary | 5 |
| Senary | 6 |
| Septenary | 7 |
| Octonary | 8 |
| Novenary | 9 |
| Decanary | 10 |

**2. Search Conditions** – These are carefully selected search conditions based on familiar materials properties and aimed at giving users the best options when searching for a data. The search conditions are as follows.

**a. Element** – The selected element(s) from the periodic table will appear in this box. It also allows for direct user input.

**b. Crystal System** – The specified crystal system is based on the various classes of space groups.

**c. System Type –** The available system type options are 0D/Cluster, 1D, 2D/Surface-Adsorption, 3D/Bulk and Atom/Molecule.

**d. Method** – The method is a list of computational codes. The available options are Abinit, BigDFT, Quantum Espresso and VASP.

**e. Basis Set Type** – This includes the basis set. In the search query, it is one option. The available options are Plane Waves, Gaussian and Wavelets.

**f. XC Functional** – Using this we select data with a specific exchange correlation function. The available options include GGA, DFT+U and Hybrid.

**g. Authors** – This is a list of users who have uploaded data to the repository.

**h. Compound Type** – The compound type option is based on the number of elements present in each calculation. The compound types and their corresponding number of elements are shown in the Table 3.

**i. Access Type** – **Restricted** or **Open Access** permission. Open Access is the default option, which means that user can download both data files and search results.

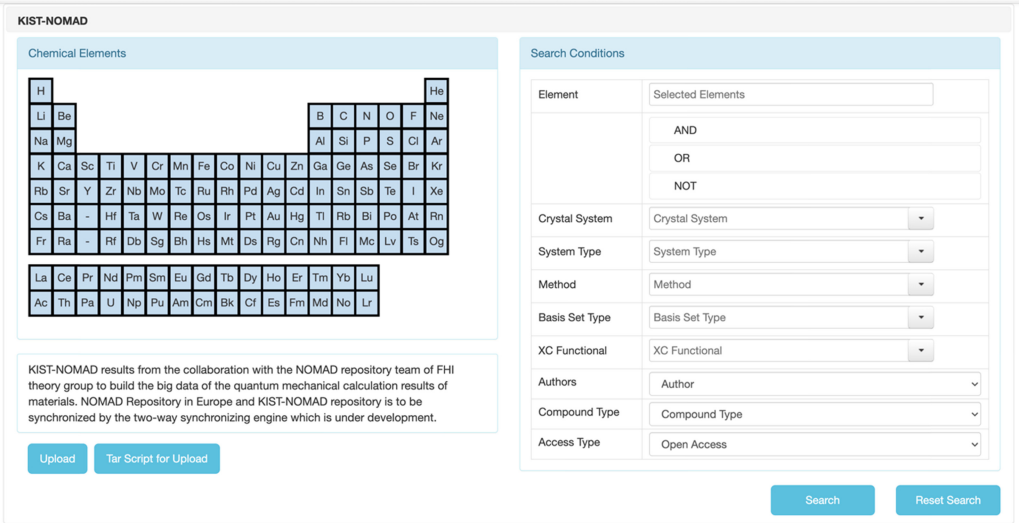


**Fig. 4.** KIST-NOMAD data search GUI. It shows the two main sections 'Chemical Elements' and 'Search Conditions'. In 'Search Conditions', there are buttons such as 'Upload', 'Search' and 'Reset Search'.

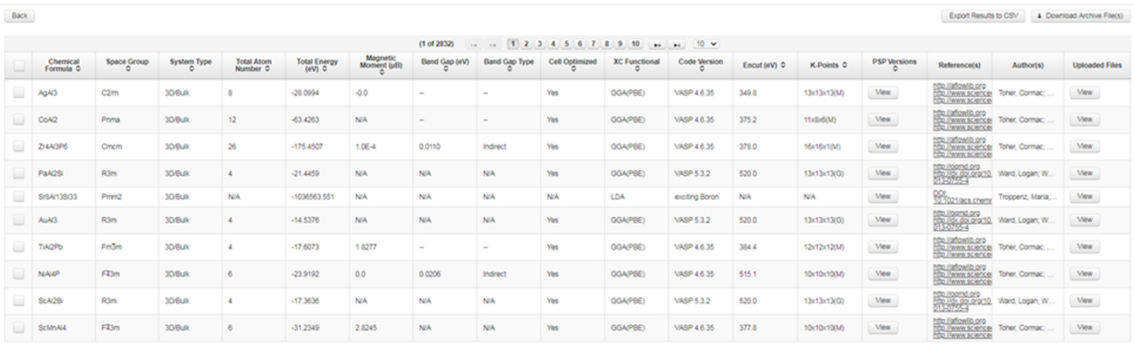

**Fig. 5.** The first results of a GUI data search. Each column is a specific materials property. The columns can be sorted in ascending or descending order.

### 3.1.2 Data search

In a data repository, the primary activity is searching for data. KIST-NOMAD implements a reliable and efficient search algorithm capable of handling all user requests in the shortest possible time. Searches can be performed when just a chemical element or formula is specified. The web implementation uses Java Persistence Query Language (JPQL) [24] to form a query from the selected elements and search conditions.

select statement also uses regular expressions to define a search pattern in the query. The defined search pattern helps to retrieve the exact requested data. The default data access type is always added to the query. This query is converted into Structured Query Language (SQL)'s select statements. Then it is parsed to the database to retrieve data from the materialized views. SQL is used to communicate with relational databases and perform tasks such as select, update and delete.

For example, to retrieve all the Aluminum based computation results, we would use a JPQL query such as SELECT e FROM new_view_grouped e WHERE ((e.chemicalFormula = 'Al' OR e.chemicalFormula REGEXP 'Al[0-99].*')) AND e.permission = :accesstype. This means *'select all aluminum computational results that have open access permission.'* In this command, the most important part is the regular expression 'Al[0-99].*'. This will ensure that the query retrieves any data with 'Al' with any number between '0-9' and any one character after the number and any other character. The first ten records of the searched results using the above query are shown in Fig. 5.

Chemical Formula, Space Group, Total Atom Number, Total Energy, Magnetic Moment, Band Gap, Band Gap Type, Cell Optimized, XC Functional, Code Versions, Encut, KPoints and PSP Versions are materials properties extracted from the uploaded calculations files with parsers and scripts. System Type is selected during upload. References to any published work are hyperlinked in the references column. Author(s) information is the name of the user (who uploaded the calculation files). It is automatically added to the calculation. Where there are 'coauthors', they are added by the user. All uploaded files for a calculation can be viewed in the Uploaded Files column.

### 3.1.3. GUI data search result

The results of GUI data search are presented in a table format and displayed on the results GUI. The results set is a carefully selected set of materials properties which are descriptive of the calculation they represent. The results among other things also allow for the quantitative comparison of calculation data. Each column in the table presents a specific materials property as defined in the database. Any column with N/A means data is not available.

The total energy column displays the total energy of the calculation in electron volts (eV) at temperature 0 K. This is the final energy(sigma → 0) value in the VASP OUTCAR file. The command (sigma → 0) means the SIGMA value, which is used to maintain the rise in temperature for VASP calculations being extrapolated to zero, hence the energy (sigma → 0) is equal to the energy at 0 K.

In the bad gap column, there are three types of values such as --, N/A and a value such as 0.007. If the calculated band gap is less than 0.005, it is represented as '--' in the result set.

Any other band gap value greater than or equal to 0.005 is presented together with the band gap type.

For VASP calculations, the condition for calculating the band gap is that the sum of the total drift in the final relaxation step be less than or equal to 0.001 (≤0.001). If this condition is not met, the band gap value is marked as N/A.

Magnetic moments values are only retrieved for spin calculations. N/A is presented for calculations with no SPIN. Cell optimized is determined by the value of the Pullay Stress. Yes is for Pullay Stress with 0.0 kB, while No is for any other value. Space group is presented in the Hermann-Mauguin notation [25]. The defined space group is a combination of an uppercase letter for the lattice type and symbols identifying the symmetry elements. For example, in space group Pmmm, P is the lattice type and mmm is for the symmetry elements.

The K-Points column displays 3 kinds of values. Two types of values are for non-band-structure calculations, for example 8x8x8(M) and 8x8x8(G). The (M) represents Monkhorst-Pack, and (G) is for Gamma. The third type of value is for band-structure calculations, for example Line-mode(20). Line-mode indicates the calculation is for band-structures and the (20) is the number of steps. When Line-mode(20) is selected, the content of the KPOINTS file is displayed.

#### 3.1.4 The API data search

KIST-NOMAD also provides a restful application programming interface (API) with functions that allow the search, retrieval of data and downloading of archive data files. APIs help in data exchange between two applications. A user sends a data retrieval request to the database though the API. The database retrieves application retrieves the data and performs any necessary actions and presents the results to the user in JavaScript Object Notation (JSON) format [26]. The returned result does not include any materials properties but rather URLs to the calculation archive files, as shown in Fig. 6.

The given URL for KIST-NOMAD API is *http://nomad.kist.re.kr:8080/nomad/rest/api/search.*.

As in GUI data search, search conditions are also specified when using the API. The following case sensitive keywords can be appended to the URL as search conditions: element, system_type, crystal_system, calculation, basis_set_type, xctreatment, author and compoundType. The conditions can be used individually or combined as required. For example, element=Si is appended to the API URL to retrieve all Silicon computation results in the database such as *http://nomad.kist.re.kr:8080/nomad/rest/api/search?element=Si*

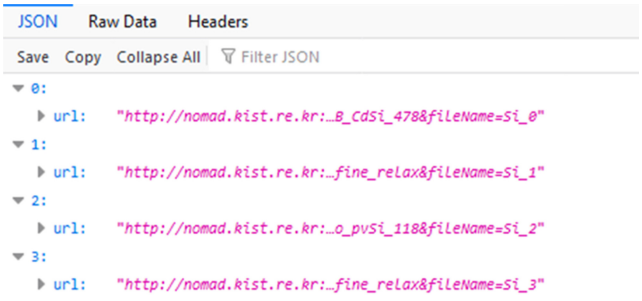


**Fig. 6.** Sample API data search result set for Silicon. The calculation archive file is downloaded when the URL is clicked.

### 3.2 Downloading data and results files

All the KIST-NOMAD open access data and data files are available for download. The download of data and files are made possible by three download functions which are available on the results GUI. The three functions allow the download of (1) Materials data in csv format, (2) Archived files in zipped format, and (3) Individual files also in zipped format.

#### 3.2.1 Materials data download

All the materials properties presented in the result set are downloadable in comma-separated values (csv) format. Materials data in the csv format is useful as input for machine learning and data analysis tasks. The data in the csv file is in the same order and format as the result set from chemical formula to pseudopotential (psp) versions.

The content of the csv file is from chemical formula to pseudopotential versions because these properties are usually used for analysis and machine learning purposes. The user can select up to but not more than 100 results (the maximum number of results per page) for downloading at one time. The formatting of the csv, such as space group, is done by writing the Hermann-Mauguin notation instead of the number in the csv files. This helps to get the csv file content in the same format as the search result set. The CSVWriter of OpenCSV [32] is used in writing the database values into the csv file.

### 3.2.2 Compressed/archived files download

The archived/compressed file for each calculation can be downloaded. These files are stored in KIST-NOMAD's uploads file directory. Downloading the archived files is particularly useful when all the uploaded files for a calculation is needed in bulk or small amounts. The archived files of the selected calculations are placed in a zipped folder during download. The archived files for the entire result set is available for download but only 100 can be downloaded at one time.

### 3.2.3 Individual files download

Additionally, for each calculation result, the individual input and output files such as OUTCAR, POSCAR, KPOINTS, and etc. can be downloaded. These files are stored in the extracted data files directory. This download is useful when only specific calculation files are needed. The uploaded files for a calculation are as shown in Fig. 7. The files can be downloaded from this GUI.

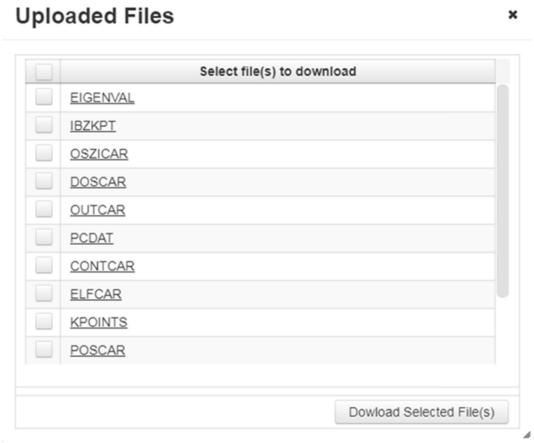


**Fig. 7.** A window displaying all the uploaded files for a given computation. Multiple files can be selected for download or individual files can be downloaded.

### 3.3 Uploading calculation files

Uploading calculations data files to KIST-NOMAD is simplified by the use of an upload GUI. Multiple calculations files in .tar.gz format can be uploaded at a time. During the upload, the system type (2D, 3D, etc.) of the calculations to be uploaded must be selected. A log-in account is required for data files upload.

The uploaded files are first kept in the uploads directory. They are then copied to the extracted folder where they are extracted. Parsers and scripts then automatically extract and calculate all the defined materials properties from the designated files and save them in the database. KIST-NOMAD aims to provide quality and reliable materials data to users, therefore the parsers and scripts are written to produce very accurate results. The user's (uploader) information is also saved in the database and mapped in a one-to-many relationship to their calculations. This process is illustrated in Fig. 8.

All the uploaded calculations details are instantly available to the owner, the user who uploaded them and are read only to all other system users. The owner can grant file and data download access to their calculations when the read only restriction is removed (made open access) or when the calculations are shared with selected individuals or groups. If there is/are any published work based on the uploaded calculations, they can be added. All these functions are available on the upload GUI as shown in Fig. 9.

### 3.4 Adding citations/references

Citations or references can be added to a single or selected calculations on the upload GUI. This is done by selecting the calculation(s) and typing references in the References text box under upload details and Save.

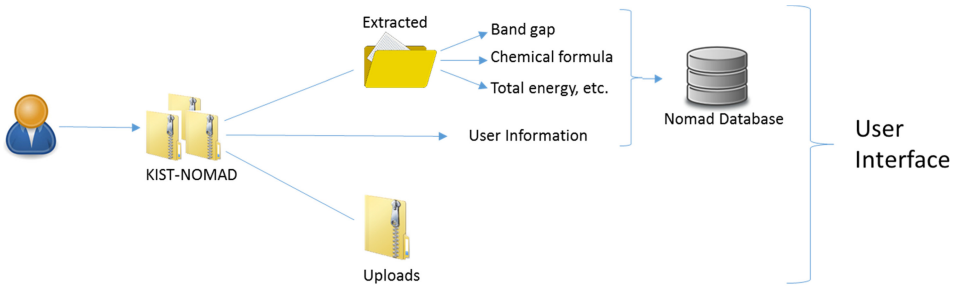


**Fig. 8.** Uploading file. It includes tar.gz files extraction, calculation and retrieval of materials properties and saving materials properties. This series of processes makes files and data available in KIST-NOMAD repository.

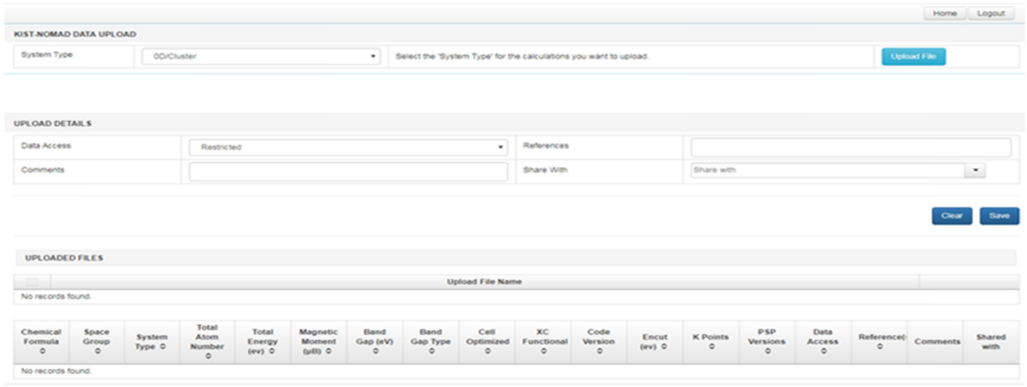


**Fig. 9.** KIST-NOMAD upload GUI. Data upload and all related activities such as uploading data files, changing permissions, adding references and sharing of calculations are available on the GUI.

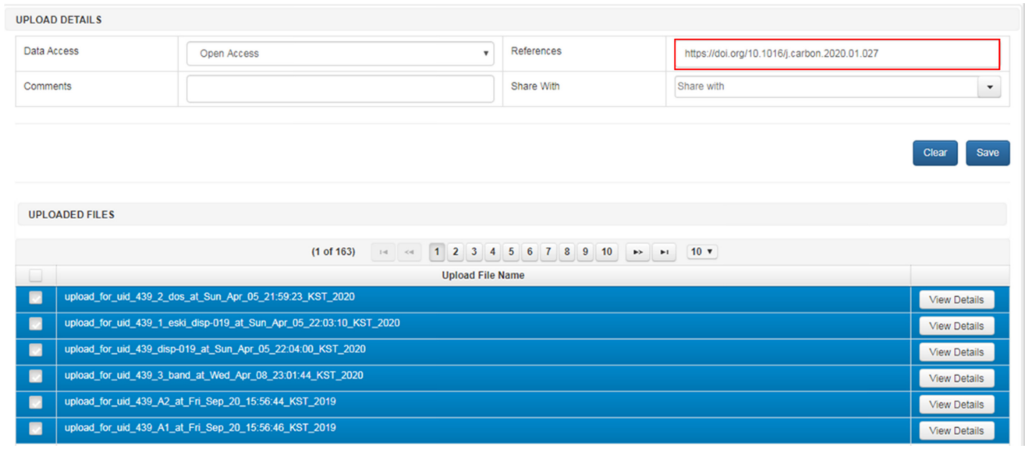


**Fig. 10.** Adding reference. References can be added when uploaded calculations have any published work.

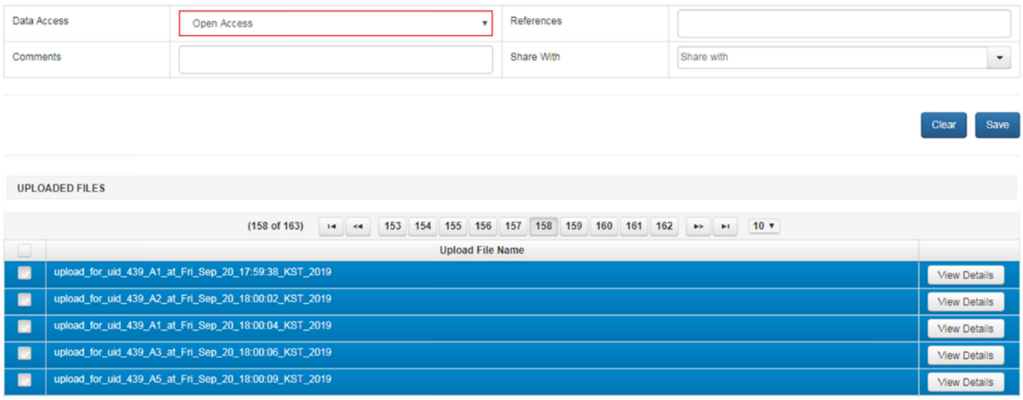


**Fig. 11.** Changing data access permissions for selected calculations. Restricted calculations on the upload GUI are private to the owner.

### 3.5 Changing data access permission

By default all calculations are restricted to the user and a selected group/persons. This restriction is removed when calculation permissions are changed from 'Restricted' to Open Access. For a selected number of calculations, permissions can be changed by choosing another value in the Data Access drop down box and saving.

### 3.6 Accessing shared calculation(s)

Shared calculations are accessed by clicking Shared Calculations

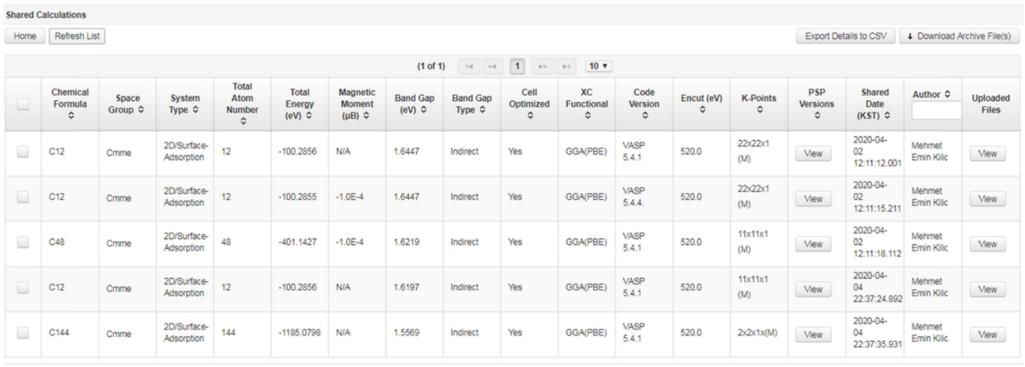


**Fig. 12.** Indicating the shared computation. Two columns Shared Date and Author show its status.

button on the search GUI after log in. When we click Refresh List in the shared calculations the GUI will load all the shared calculations. The shared calculations GUI is shown in Fig. 12.

## 4. Examples

In this section, we provide examples to demonstrate the KIST-NOMAD functions described in Section 3. The given examples include GUI Data search and API search.

### 4.1 API search and download

The example illustrates search and download with API. We searched for all binary Aluminum compounds. The full search URL and the first five results are shown in Fig. 13.

When we click on the first URL, the download dialogue opens up.

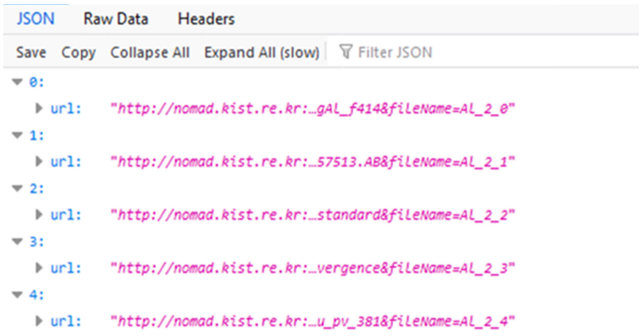


**Fig. 13.** API search results for Binary Aluminum compounds. The Al_2_0 of the first URL represent the element, the compound type and the sequential number of the URL.

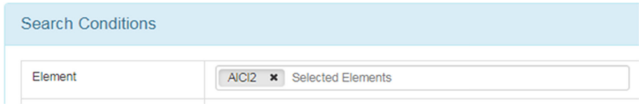


**Fig. 14.** AlCl2 in the Element text box of the search GUI. The text b

### 4.2 GUI data search and csv download

This example will search for $AlCl_2$ compounds and download a csv file for the first ten results. AlCl2 is typed in the Element text box as shown below. A click on Search will

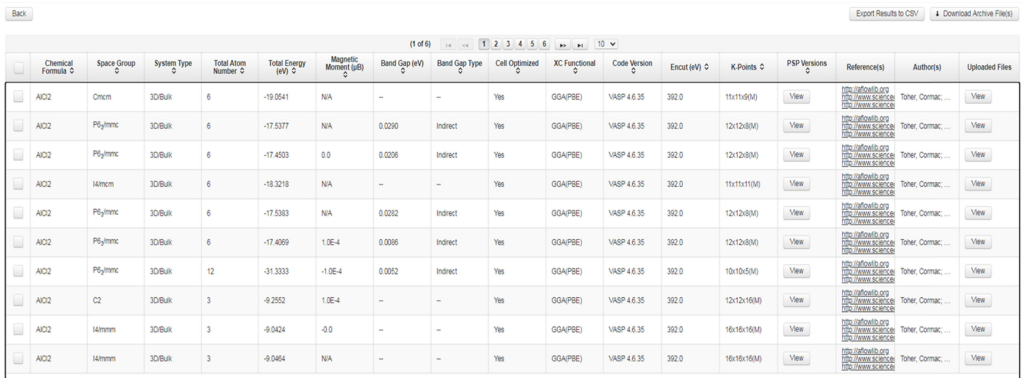


**Fig. 15.** First ten rows of $AlCl_2$ search results. There are six pages of results. The number of rows per page can be increased, up to 100.

| Chemical Formular | Space Group | System Type | Total Atom Number | Total Energy(eV) | Magnetic Moment(ï¼B) | Band Gap(eV) | Band Gap Type | Cell Relaxed | Xc Functional | Code Version | Encut(eV) | K Points | PSP Version(s) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AlCl2 | Cmcm | 3D/Bulk | 6 | -19.0541 | N/A | -- | -- | Yes | GGA(PBE) | VASP 4.6.35 | 392 | 11x11x9(M) | Al 04Jan2001,Cl 17Jan2003, |
| AlCl2 | P63/mmc | 3D/Bulk | 6 | -17.5377 | N/A | 0.029 | Indirect | Yes | GGA(PBE) | VASP 4.6.35 | 392 | 12x12x8(M) | Al 04Jan2001,Cl 17Jan2003, |
| AlCl2 | P63/mmc | 3D/Bulk | 6 | -17.4503 | 0 | 0.0206 | Indirect | Yes | GGA(PBE) | VASP 4.6.35 | 392 | 12x12x8(M) | Al 04Jan2001,Cl 17Jan2003, |
| AlCl2 | I4/mcm | 3D/Bulk | 6 | -18.3218 | N/A | -- | -- | Yes | GGA(PBE) | VASP 4.6.35 | 392 | 11x11x11(M) | Al 04Jan2001,Cl 17Jan2003, |
| AlCl2 | P63/mmc | 3D/Bulk | 6 | -17.5383 | N/A | 0.0282 | Indirect | Yes | GGA(PBE) | VASP 4.6.35 | 392 | 12x12x8(M) | Al 04Jan2001,Cl 17Jan2003, |
| AlCl2 | P63/mmc | 3D/Bulk | 6 | -17.4069 | 1.00E-04 | 0.0086 | Indirect | Yes | GGA(PBE) | VASP 4.6.35 | 392 | 12x12x8(M) | Al 04Jan2001,Cl 17Jan2003, |
| AlCl2 | P63/mmc | 3D/Bulk | 12 | -31.3333 | -1.00E-04 | 0.0052 | Indirect | Yes | GGA(PBE) | VASP 4.6.35 | 392 | 10x10x5(M) | Al 04Jan2001,Cl 17Jan2003, |
| AlCl2 | C2 | 3D/Bulk | 3 | -9.2552 | 1.00E-04 | -- | -- | Yes | GGA(PBE) | VASP 4.6.35 | 392 | 12x12x16(M) | Al 04Jan2001,Cl 17Jan2003, |
| AlCl2 | I4/mmm | 3D/Bulk | 3 | -9.0424 | 0 | -- | -- | Yes | GGA(PBE) | VASP 4.6.35 | 392 | 16x16x16(M) | Al 04Jan2001,Cl 17Jan2003, |
| AlCl2 | I4/mmm | 3D/Bulk | 3 | -9.0464 | N/A | -- | -- | Yes | GGA(PBE) | VASP 4.6.35 | 392 | 16x16x16(M) | Al 04Jan2001,Cl 17Jan2003, |

**Fig. 16.** CSV file of the first ten results of $AlCl_2$ search. The data in the csv is in the same order and format as the result set.

retrieve specified results from the database.

The first ten rows of the results are shown in Fig. 15. Select those ten results and click Export to CSV to download the results in the csv format as in Fig 16.

### 4.3. Machine learning example

The machine learning work described in [27] discussed the interatomic potential energy surface model for silicon oxide, which was used to simulate its molecular dynamics (MD). This approach is an extension of the Behler and Parrinello potential [30], where atomic energy and forces are predicted using atomic configuration information.

In addition to the energy predictions, the force components and the electrostatic charge distribution of each atom are predicted. The mean square deviation in the loss function for the training and test data set is on the order of ~ 0.01 eV/atom.

The considered dataset for training the system is a collection of all the polymorphs of silicon oxide in bulk and cluster form, and ab-initio molecular dynamic calculations including the melting and quenching of the silicon oxide. We confirm that important data for this machine learning comes from KIST-NOMAD using its features such as compressed/archive file download.

## 5. Conclusion

The main features and functions of KIST-NOMAD, a materials data repository, have been presented in the sections above. KIST-NOMAD provides users more materials properties in its results set, allows for the download of materials properties as csv, the bulk download of archive files and API for archive files download. Only open source software and libraries were utilized in the development of KIST-NOMAD. The extraction of materials properties are automated with efficiency and speed. Machine learning and other data exploitation tools are currently being developed for KIST-NOMAD to create a multi-purpose materials data platform.

We mentioned the important role computational materials data repositories have in enhancing and establishing the fourth paradigm of materials research. Some tools and techniques for extracting knowledge in materials data and files are discussed. Collaboration between materials and computer scientists would help create more reliable and powerful tools to take full advantage of materials repositories for new and exciting research discoveries.

## Acknowledgments

The authors acknowledge the support of Prof. Matthias Scheffler and the FHI theory group for their immense assistance and providing the repository data and files. Our research is supported by KIST's Future Convergence Research 2E30460 and Ministry of Science and ICT's Material Platform Research 2N57370.

## REFERENCES

1. G. Kresse and J. Furthmüller. *Comput. Mater. Sci.* **6**, 15 (1996).
2. P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello1, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch. *J. Phys.: Condens. Matter* **21**, 395502 (2009).
3. S. G. Balasubramani, G. P. Chen, S. Coriani, M.

Diedenhofen, M. S. Frank, Y. J. Franzke, F. Furche, R. Grotjahn, M. E. Harding, C. Hättig, A. Hellweg, B. Helmich-Paris, C. Holzer, U. Huniar, M. Kaupp, A. M. Khah, S. K. Khani, T. Müller, F. Mack, B. D. Nguyen, S. M. Parker, E. Perlt, D. Rappoport, K. Reiter, S. Roy, M. Rückert, G. Schmitz, M. Sierka, E. Tapavicza, D. P. Tew, C. van Wüllen, V. K. Voora, F. Weigend, A. Wodyński, and J. M. Yu, *J. Chem. Phys.* **152**, 184107 (2020).
4. B.-J. Lee, *Korean J. Met. Mater.* **56**, 253 (2018).
5. S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, *Comput. Mater. Sci.* **58**, 227 (2012).
6. J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, *JOM* **65**, 1501 (2013).
7. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, *APL Materials* ***1***, 011002 (2013).
8. Claudia Draxl and Matthias Scheffler, *J. Phys. Mater*. **2**, 036001 (2019).
9. J. M. Cole, *Acc. Chem. Res.* **53**, 599 (2020).
10. K. A. Severson, P. M. Attia, J. Norman, N. Jin, N. Perkins, B. Jiang, Z. Yang, M. H. Chen, M. Aykol, P. K. Herring, D. Fraggedakis, M. Z. Bazant, S. J. Harris, W. C. Chueh, and R. D. Braatz, *Nat. Energy* **4**, 383 (2019).
11. T. Lookman, P. V. Balachandran, D. Xue, and R. Yuan, *Npj Comput. Mater.* **5**, 5 (2019).
12. D. You, S. Ganorkar, S. Kim, K. Kang, W. Y. Shin, and D. Lee, *Scr. Mater.* **183**, 1, 2020.
13. S. R. Xie, P. Kotlarz, R. G. Hening, and J. C. Nino, *Comput. Mater. Sci*. **180**, 109690, 2020.
14. R. V. Noorden, *Nature* 525, 306 (2015).
15. G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, and A. Fazzio, *J. Phys. Mater.* **2** 032001 (2019).
16. N. Artrith and A. Urban. *Comput. Mater. Sci*. **114**, 135 (2016).
17. H. Wan., L. Zhang, J. Han, and E. Wienan, *Comput. Phys. Commun.* **228**, 178 (2018).
18. A. Khorshidi and A. A. Peterson, *Comput. Phys. Commun.* **207**, 310 (2016).
19. V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler, *Comput. Phys. Commun.* **180**, 2175 (2009).
20. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, Jan-Willem Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, Susanna-Assunta Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, *Sci Data* **3**, 160018 (2016).
21. J. Pople, *Expanding the limits of computational chemistry*, https://gaussian.com (1970).
22. M. A. L. Marques, A. Castro, G. F. Bertsch, and A. Rubio, *Comput. Phys. Commun*. **151**, 60 (2003).
23. PostgreSQL Global Development Group, *PostgreSQL*, https://www.postgresql.org (1996).
24. Oracle, *JPQL Language Reference,* https://docs.oracle.com /html/E13946_04/ejb3_langref.html (2011).
25. D. Peck and E. Delventhal, *Herman-Maugin Symmetry Symbols*, https://www.mindat.org/article.php/2742/Hermann-Mauguin+Symmetry+Symbols (2020).
26. D. Crockford, *Introducing JSON*, https://www.json.org/ json-en.html (2013).
27. D. S. Boateng, K. R. Lee, *Proc. 5th Int. Conf. on Molecular Simulation*, p. 226, Korean Inst. Metals & Mater., Seoul, Korea (2019).
28. L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, C. S. Adorf, C. W. Andersen, O. Schütt, C. A. Pignedoli, D. Passerone, J. VandeVondele, T. C. Schulthess, B. Smit, G. Pizzi, and N. Marzari, *arXiv: Cond-Mat. Mtrl-Sci*, **2003**, 12510 (2020).
29. A. Gulans, S. Kontur, C. Meisenbichler, D. Nabok, P. Pavone, S. Rigamonti, S. Sagmeister, U. Werner, and C. Draxl. *J. Phys. Condens. Matter* **26**, 363202 (2014).
30. J. Behler and M. Parrinello. *Phys. Rev. Lett.* **98**, 146401 (2007).
31. B. Hollis, *JSONView: Pretty JSON in FireFox and Chrome*, https://jsonview.com (2009).
32. OpenCSV, *Opencsv Users Guide,* http://opencsv. sourceforge.net (2020).