Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2024, **3**, 1509

Developments and applications of the OPTIMADE API for materials discovery, design, and data exchange†

Matthew L. Evans, [a b] Johan Bergsma, [‡c] Andrius Merkys, [d] Casper W. Andersen, [e] Oskar B. Andersson, [f] Daniel Beltrán, [g] Evgeny Blokhin, [h i] Tara M. Boland, [j] Rubén Castañeda Balderas, [k] Kamal Choudhary, [l] Alberto Díaz Díaz, [k] Rodrigo Domínguez García, [k] Hagen Eckert, [m n] Kristjan Eimre, [o] María Elena Fuentes Montero, [p] Adam M. Krajewski, [q] Jens Jørgen Mortensen, [j] José Manuel Nápoles Duarte, [p] Jacob Pietryga, [r] Ji Qi, [s] Felipe de Jesús Trejo Carrillo, [k] Antanas Vaitkus, [d] Jusong Yu, [o ah] Adam Zettel, [m n] Pedro Baptista de Castro, [t] Johan Carlsson, [u] Tiago F. T. Cerqueira, [v] Simon Divilov, [m n] Hamidreza Hajiyani, [§u] Felix Hanke, [w] Kevin Jose, [x] Corey Oses, [y] Janosh Riebesell, [x z] Jonathan Schmidt, [aa] Donald Winston, [ab] Christen Xie, [s] Xiaoyu Yang, [ac ad ae] Sara Bonella, [c] Silvana Botti, [af] Stefano Curtarolo, [m n] Claudia Draxl, [ag] Luis Edmundo Fuentes Cobas, [k] Adam Hospital, [g] Zi-Kui Liu, [q] Miguel A. L. Marques, [af] Nicola Marzari, [o ah] Andrew J. Morris, [ai] Shyue Ping Ong, [s] Modesto Orozco, [g] Kristin A. Persson, [z aj] Kristian S. Thygesen, [j] Chris Wolverton, [r] Markus Scheidgen, [ag] Cormac Toher, [n ak] Gareth J. Conduit, [x an] Giovanni Pizzi, [o ah] Saulius Gražulis, [d al] Gian-Marco Rignanese [*a b am] and Rickard Armiento [*f]

The Open Databases Integration for Materials Design (OPTIMADE) application programming interface (API) empowers users with holistic access to a growing federation of databases, enhancing the accessibility and discoverability of materials and chemical data. Since the first release of the OPTIMADE specification (v1.0), the API has undergone significant development, leading to the v1.2 release, and has underpinned multiple scientific studies. In this work, we highlight the latest features of the API format, accompanying software tools, and provide an update on the implementation of OPTIMADE in contributing materials databases. We end by providing several use cases that demonstrate the utility of the OPTIMADE API in materials research that continue to drive its ongoing development.

Received 2nd February 2024
Accepted 15th April 2024

DOI: 10.1039/d4dd00039k

rsc.li/digitaldiscovery

1 Introduction

Industrial chemicals and materials underpin the global economy: for example, chemicals alone contribute $6.4 trillion[24] annually to the global economy. Industrial chemicals and materials companies are under significant pressure to improve the environmental, social, and governance impact of their business,[10] with a particular focus on reducing the carbon footprint. Unfortunately, the discovery of chemicals and drugs is traditionally a time-consuming and expensive process driven by experiment-led trial-and-improvement, delaying the response to the climate crisis. However, in the last few years, high-throughput calculations have led to an explosion in the volume of available materials data.[41,57,77] Machine learning has emerged as a pivotal tool to exploit this data,[125] accelerating the discovery of chemicals and drugs that meet the challenges faced today.

A core requirement for the data and machine-learning revolution is data availability and interoperability. Therefore, the Open Databases Integration for Materials Design (OPTIMADE) universal application programming interface (API) was created to empower users with programmatic access to many leading materials databases. By organising under an open federation, and emphasising the interoperability of search as well as access, OPTIMADE improves the discoverability of materials data, especially from smaller, less known databases. As we move into the era of autonomous laboratories (both computational and experimental), the technical approach taken renders all OPTIMADE APIs machine actionable, allowing for automated serendipitous discovery of newly added data entries in a given materials space without needing to specify which

databases to access. Such an extended data availability requires explicit clarification of data permissions and ownership, which can be different for each database.

Since the first release of the OPTIMADE specification (v1.0)[2] with accompanying article,[3] the OPTIMADE API format has enjoyed significant adoption, with 22 registered providers[130,131] of 25 interoperable databases serving over 22 million crystal structures with associated properties. In the v1.2 release,[5] the specification has undergone significant extensions and enhancements that enable novel use cases whilst making the format accessible to both users and developers. This gives users access to data from both large and well-known sources, and many specialist datasets focused on a family of materials of particular interest. The combination of a general overview of all possible materials and detailed knowledge of particular materials enables novel discovery and deep insights with for example machine learning.

In this paper, we highlight recent developments to and uses of the OPTIMADE API. First, in Section 2, we provide an overview of the OPTIMADE API format, and of the latest features. Next, in Section 3, we highlight the efforts of leading materials databases to provide access through the OPTIMADE API format. In Section 4, we show how the OPTIMADE API has been used in computational screening and for the creation of machine learning datasets. Section 5 discusses future plans for OPTIMADE, both in terms of new technical frontiers, and of the sustainability of the ecosystem and community. Finally, in Section 6, we summarise and look ahead to the future of materials databases.

2 Overview of the OPTIMADE API

The OPTIMADE API is well-positioned to set the standard to enable search, retrieval and annotation in a common way for all databases. Crystal structure data has benefited from decades of standardisation work in the form of the Crystallographic Information File (CIF)[13,51] and related initiatives, which heavily inspired the crystal structure representation employed by the OPTIMADE API format. By building a standardised, open format, both proprietary and open data can be aggregated and used on the same footing.

Building on these seminal standardisation efforts, OPTIMADE goes considerably further than standardising the representation of crystal structure data, by including: the means for

[a]UCLouvain, Institut de la Matière Condensée et des Nanosciences (IMCN), Chemin des Étoiles 8, Louvain-la-Neuve 1348, Belgium. E-mail: gian-marco.rignanese@uclouvain.be
[b]Matgenix SRL, 185 Rue Armand Bury, 6534 Gozée, Belgium
[c]Centre Européen de Calcul Atomique et Moléculaire (CECAM), École Polytechnique Fédérale de Lausanne, Avenue de Forel 3, 1015 Lausanne, Switzerland
[d]Institute of Biotechnology, Life Sciences Center, Vilnius University, Saulėtekio av. 7, LT-10257 Vilnius, Lithuania
[e]SINTEF, P.O. Box 4760 Torgarden, NO-7465 Trondheim, Norway
[f]Materials Design and Informatics Unit, Department of Physics, Chemistry and Biology, Linköping University, Sweden. E-mail: rickard.armiento@liu.se
[g]Institute for Research in Biomedicine (IRB Barcelona), Baldiri i Reixac 10-12, 08028 Barcelona, Spain
[h]Tilde Materials Informatics, Straßmannstraße 25, 10249, Berlin, Germany
[i]Materials Platform for Data Science, Sepapaja 6, 15551, Tallinn, Estonia
[j]Computational Atomic-Scale Materials Design, Technical University of Denmark, Kgs. Lyngby, Denmark
[k]Centro de Investigación en Materiales Avanzados, S.C. (CIMAV), Av. Miguel de Cervantes 120, Complejo Industrial Chihuahua, 31136, Chihuahua, Chih., Mexico
[l]Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, 20899, USA
[m]Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708, USA
[n]Center for Extreme Materials, Duke University, Durham, NC 27708, USA
[o]Theory and Simulation of Materials (THEOS), and National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[p]Universidad Autónoma de Chihuahua, Facultad de Ciencias Quimicas, 31125 Chihuahua, Mexico
[q]Department of Materials Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA
[r]Department of Materials Science and Engineering, Northwestern University, Evanston, IL 60208, USA
[s]Department of NanoEngineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, California, 92093-0448, USA
[t]National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki, 305-0047, Japan
[u]Dassault Systèmes Germany GmbH, Am Kabellager 11-13, 51063 Cologne, Germany
[v]CFisUC, Department of Physics, University of Coimbra, Rua Larga, 3004-516 Coimbra, Portugal
[w]Dassault Systèmes, 22 Science Park, CB4 0FJ, UK
[x]Theory of Condensed Matter, Cavendish Laboratory, Cambridge, UK
[y]Department of Materials Science and Engineering, Johns Hopkins University, Baltimore, MD 21218, USA
[z]Lawrence Berkeley National Lab, Berkeley, CA, USA
[aa]Materials Theory, ETH Zürich, Wolfgang-Pauli-Strasse 27, 8093 Zurich, Switzerland
[ab]Polyneme LLC, New York, NY 10038, USA
[ac]Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100083, China
[ad]University of Chinese Academy of Sciences, Beijing, 101408, China
[ae]Beijing MaiGao MatCloud Technology Co. Ltd, Beijing, 100149, China
[af]Research Center Future Energy Materials and Systems of the University Alliance Ruhr and Interdisciplinary Centre for Advanced Materials Simulation, Ruhr University Bochum, Universitätsstraße 150, D-44801 Bochum, Germany
[ag]Humboldt-Universität zu Berlin, Institut für Physik and IRIS Adlershof, 12489 Berlin, Germany
[ah]Laboratory for Materials Simulations (LMS), Paul Scherrer Institute (PSI), 5232 Villigen PSI, Switzerland
[ai]School of Metallurgy and Materials, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK
[aj]Department of Materials Science and Engineering, UC Berkeley, Hearst Mining Memorial Building, Berkeley, 94720 CA, USA
[ak]Department of Materials Science and Engineering and Department of Chemistry and Biochemistry, The University of Texas at Dallas, Richardson, TX 75080, USA
[al]Institute of Computer Science, Faculty of Mathematics and Informatics, Vilnius University, Naugarduko g. 24, LT-03225 Vilnius, Lithuania
[am]School of Materials Science and Engineering, Northwestern Polytechnical University, Xi'an, Shaanxi, 710072, China
[an]Intellegens Ltd, French's Rd, Cambridge, CB4 3NP, UK

† Electronic supplementary information (ESI) available: Copy of Table 1 with web links. See DOI: https://doi.org/10.1039/d4dd00039k

‡ Present address: Van't Hoff Institute for Molecular Sciences, University of Amsterdam, PO Box 94157, 1090 GD Amsterdam, Netherlands.
§ Present address: Henkel, AID/Digital Twins & Data Analytics, 40589 Düsseldorf, Germany.

filtering entries (*via* the OPTIMADE filter grammar), a standard for laying out resources on the web (by providing rules and expectations of URL formats), a means for introspectively defining additional properties and entry types per-database, and the creation of a decentralised federation of compatible databases. These additional aspects are what maximises the impact of the OPTIMADE API and enable new scientific applications. The OPTIMADE API is registered with FAIRsharing.org as a data standard,[38] and releases are archived on Zenodo,[2] with ongoing development occurring openly under the Materials-Consortia banner on GitHub (**https://github.com/Materials-Consortia**).[1,29] The initial motivation for OPTIMADE, and a discussion of the previously existing materials API formats and filter mechanisms can be found in ref. 3 which described the first release.

The process for a user to access OPTIMADE compliant data is as follows. The user starts with the base URL defined by the database provider (available from the federated OPTIMADE provider list[130,131]), and then appends a common string describing the entry type to query, plus any filter or implementation parameters, which is submitted as an HTTP GET request. For example, to probe the Materials Project[66] for materials containing SiO_2, we make a GET request to the following URL:

base URL endpoint

```
https://optimade.materialsproject.org/v1/structures
?filter=chemical_formula_reduced="O2Si"
```

OPTIMADE filter

This delivers the response in Box 1 that contains entries where oxygen and silicon occur in a 2 : 1 ratio. The power of the OPTIMADE API is that the same generic request can be appended to the base URL of any other database, and its matching entries will be returned. This is a non-trivial step for database providers, who must convert the OPTIMADE filter grammar into the corresponding query for their own database engine. The benefit is that client code can then be written to unify the results from multiple databases, allowing users to receive the most comprehensive results for the query. Such an extended data availability requires explicit clarification of data permissions and ownership, which can be different for each database or even each entry.

Box 1: An excerpt of the JSON response showing the material attributes for one of the returned entries.


```
{
 "data": [
  {
   "id": "mp-7000",
   "type": "structures",
   "attributes": {
    "immutable_id": "645d2ba4bcd30f748b475981",
    "last_modified": "2023-03-11T14:56:30Z",
    "elements": ["O", "Si"],
    "nelements": 2,
    "elements_ratios": [0.3333333333333333, 0.6666666666666666],
    "chemical_formula_descriptive": "O6Si3",
    "chemical_formula_reduced": "O2Si",
    "chemical_formula_hill": "O6Si3",
    "chemical_formula_anonymous": "A2B",
    "dimension_types": [1, 1, 1],
    "nperiodic_dimensions": 3,
    "lattice_vectors": [
      [4.914966, -1e-8, 0],
      [-2.45748252, 4.2564861, 0],
      [0, 0, 5.43130114]
    ],
    "nsites": 9,
   "species_at_sites": ["Si", "Si", "Si", "O", "O", "O", "O", "O", "O"],
```


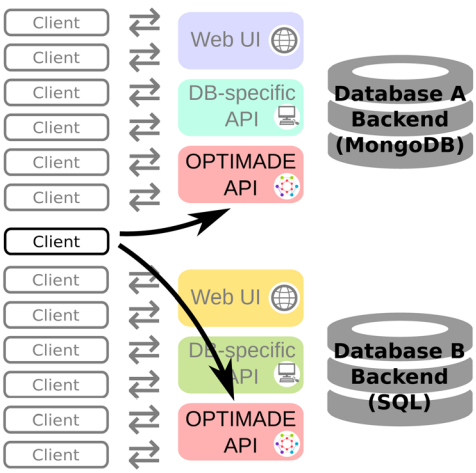

Fig. 1 Schematic for how databases with different backends may provide their own web-based user interfaces and database-specific APIs alongside the OPTIMADE API. A single client can then interact with the databases *via* the OPTIMADE API without having to be aware of these differences.

2.1 OPTIMADE core design principles

A materials database provider that implements the OPTIMADE API will have a database backend and one or more interfaces available to clients. These interfaces include the OPTIMADE API, but can also provide access by other means, *e.g.*, a database-specific API or web-based graphical user interface. Fig. 1 serves as a schematic illustration of this point.

The providers that presently support the OPTIMADE API represent a wide range of underlying backends. The primary backend component is, usually, a database engine, but the backend covers all parts of the system that manages the stored data. Relevant backends for OPTIMADE implementations range from simple flat files in a filesystem, to sophisticated setups with load-balanced distributed cloud hosting of relational database engines (*e.g.*, Structured Query Language, SQL-based) or non-relational key-value, document or graph database engines (so-called NoSQL).

There exists a multitude of APIs for data access and storage (both for materials databases and more generally) that have been designed for a specific database backend. These APIs are typically designed around specific features the backend provides in terms of, *e.g.*, browse, search, and retrieval, and, crucially, how these features are implemented by the backend. The API then typically becomes a thin wrapper for these features, which exposes the functionality of the backend implementation. There is thus

a crucial need for a generic database interface, which should be based on the central principles:

• A core feature set that any reasonable materials database backend can implement *via* cheap, single-pass, on-the-fly translations in the API layer. Furthermore, the translations should be possible without modifying the underlying backend – *i.e.*, participating databases should not be required to reformat or amend the stored data, software features, *etc.* to provide these core OPTIMADE features.

• Extended features beyond the core features that are shared among multiple (but not all) backends should be standardised as optional features. It may appear that this design works against interoperability, as it may lead to query Q working for databases A and B, but not for database C. However, if the database backend for C cannot support the type of query that Q represents (for example, a database containing molecular dynamics calculations of proteins cannot sensibly be searched on the chemical formula in the simulation cell), it means that there exists no way to make that query interoperable across A, B, and C (without altering the backend of C). Hence, the highest level of interoperability is achieved by standardising Q as an optional feature.

• For standardised optional features, multiple overlapping representations of the same features and/or data should be avoided. The reason is that, for two different ways of providing feature Q as Q_1 and Q_2, we may end up with database A supporting only Q_1 and database B supporting only Q_2. Hence, the client either has to tailor the query differently depending on the destination or provide multiple versions of the query in the request, which is an undesirable burden to put on clients and works against an interoperable design. What is strictly standardised is how a database that does not implement a certain feature should respond should that feature be requested.

• Those features and data that only occur in a particular database should be provided in isolated database-specific namespaces that other databases can recognise and handle appropriately. The OPTIMADE specification describes how references to such fields should be announced by the particular database and also how clients making multi-provider queries should handle in the response in various situations for optimal interoperability. The latest version of OPTIMADE even outlines a mechanism for creating sub-specifications that allow multiple database providers to collaborate on custom communal definitions, as shall be discussed later.

2.2 Recent OPTIMADE improvements

Since the initial v1.0 release[2] of the OPTIMADE API in July 2020, many features have been added, driven by user feedback and use cases. The major enhancements to the specification introduced in versions 1.1 (ref. 4) and 1.2 are discussed below, with the full changelog (**https://github.com/Materials-Consortia/OPTIMADE/blob/master/CHANGELOG.md**) and specification text available online on GitHub (**https://github.com/Materials-Consortia/OPTIMADE**).[129]

2.2.1 Property definitions. In previous versions of OPTIMADE, the information served through introspection for a provided property was limited to a single specification of an OPTIMADE type, which is far from sufficient for a client to use the data. For standard fields, clients were referred to the human readable descriptions in the OPTIMADE specification, and for database-specific fields to side channel information, *e.g.*, at the website of the database.

OPTIMADE v1.2 includes a full schema format, based on – and compatible with – JSONSchema,[99] capable of fully describing in a machine-readable way what a property is (including definition of sophisticated data structures with multiple layers of lists and dictionary subfields). All properties are given clear versioned stable identifiers (URIs) that can be used to identify that multiple databases refer to the exact same property. A simple example is provided in Box 2, which displays the definition of the `nsites` field. Furthermore, a browsable interface to these definitions for the OPTIMADE standard properties is available at **https://schemas.optimade.org/**.

There is related work outside of the specification on providing an enlarged set of shared definitions that can be incrementally adopted by the subsections of the community. This will allow databases to point to and share the same property definition without requiring the slow consensus-building step for the relevant fields to be promoted to the main specification. This is especially useful in cases where this shared information is the defining feature of a given database; for example, the first target namespace is that of stability predictions arising from density-functional theory calculations, the key property required to truly enable materials discovery applications with OPTIMADE (see Section 4.1.2 later for more information).

2.2.2 Streaming of partial data. So far, OPTIMADE has been based on the JSON format, which has the advantage that it is relatively human readable and well supported by most programming languages. Most tools, however, can only process an entire JSON file and thus do not support streaming processing,

```
{
    "$id": "https://schemas.optimade.org/defs/v1.2/properties/optimade
    ↪ /structures/nsites",
    "$schema": "https://schemas.optimade.org/meta/v1.2/optimade
    ↪ /property_definition.json",
    "title": "number of sites",
    "x-optimade-type": "integer",
    "x-optimade-definition": {
        "label": "nsites_optimade_structures",
        "kind": "property",
        "version": "1.2.0",
        "format": "1.2",
        "name": "nsites"
    },
    "type": [
        "integer",
        "null"
    ],
    "description": "An integer specifying the length of the
    ↪ `cartesian_site_positions`
    ↪ property.\n\n**Requirements/Conventions**:\n\n- MUST be equal to
    ↪ the lengths of the list properties elements and elements_ratios,
    ↪ if they are provided.\n\n**Query examples**:\n\n- Match only
    ↪ structures with exactly 4 sites: `nsites=4`\n- Match structures
    ↪ that have between 2 and 7 sites: `nsites>=2 AND nsites<=7`",
    "examples": [
        42
    ],
    "x-optimade-unit": "dimensionless"
}
```

which makes it difficult to handle large JSON files. In addition, JSON does not support binary data: a response with large amounts of numerical data needs to be encoded (*e.g.*, in `base64` or similar encodings), significantly increasing the size of the data to be transferred over the network. Adding support for a format like BSON, which does support binary data, or JSON variants which do support streamed processing (such as JSON Lines) could therefore improve the data transfer rates for OPTIMADE requests, and allow for immediate visualisation of partial results.

Box 2: An example OPTIMADE property definition for the `nsites` property.

In previous versions of OPTIMADE, the size of property data was limited so that a single database entry could fit in a single HTTP response. In the most recent version of OPTIMADE, the implementation can decide to defer large properties to be communicated over a separate simplified streamable protocol, which in practice can be implemented by serving large static files over HTTP. This addition was driven by the future application of serving trajectory data with OPTIMADE (see Section 5 for more information) where individual "properties", such as atomic positions in each frame of a trajectory, can individually be prohibitively large.

2.2.3 Other technical and scientific enhancements. There have been several other extensions of the format since version 1.0:

Symmetry information for structures. The latest OPTIMADE release includes a set of standardised and comprehensive descriptions of structural symmetry. This has been achieved with 5 fields that can be variously used for filtering by symmetry and for reconstructing atomic site positions, building on existing standards laid out by the IUCr and others:

- `space_group_symmetry_operations_xyz`,
- `space_group_symbol_hall`,
- `space_group_symbol_hermann_mauguin`,
- `space_group_symbol_hermann_mauguin_extended`,
- `space_group_it_number`.

`/files` *endpoint.* The OPTIMADE specification has been extended to enable the description of files and their precise relationships to other OPTIMADE entries. To implement this, a new `/files` entry endpoint has been defined. This addition allows, for example, linking `/structures` entries with their representations in widely used structural data formats such as CIF, POSCAR, and SDF, or linking directly to the raw input or output files associated with a calculation involving a structure.

Per property and per entry metadata. We plan to create a mechanism to provide metadata per property for each individual entry. This metadata could, for example, be confidence intervals, or information about how a property was calculated.

Request delay. In order not to overload a particular OPTIMADE server, the metadata property `request_delay` appears in the latest release to allow an implementation to suggest a specific back-off time delay between subsequent requests. It is up to the given server implementation to decide what to do with clients violating the requested delay: refuse to serve, intentionally delay the response, or ignore.

Licensing. The possibility of unsupervised database harvesting raises a need for machine-readable definition of data licenses. Whilst OPTIMADE is an open format, it can be used to serve proprietary or otherwise restricted data, the usage terms of which must be described. The latest OPTIMADE release addresses this issue by introducing the metadata fields `license`, `available_licenses`, and `available_licenses_for_entries`. The property `license` is intended for databases to link to human-readable licensing terms, while `available_licenses` and `available_licenses_for_entries` allow specifying machine-readable license identifiers following the Software Package Data Exchange (SPDX) standard, making interpreting requirements easier for automated clients and web crawlers. This standardised way to announce that either the whole database, or the individual entries, are explicitly available under certain licenses enables machine-actionable licensing decisions, *e.g.* for commercial re-use and republishing of data.

Substring comparisons on list elements. The OPTIMADE filter grammar defines substring comparison operators to match either the start, end or any part of a property value. However, until version 1.2, such comparisons could not be carried out on list elements. The latest version of OPTIMADE now explicitly supports such substring queries on elements of list properties.

Boolean values. Although there are no Boolean fields in the main OPTIMADE specification, the latest version of the OPTIMADE specification includes support in the filter grammar for defining and filtering on such custom fields defined by data providers.

2.3 Associated software tools

2.3.1 Optimade-python-tools. `optimade-python-tools` is an open-source software (MIT license) package that provides tooling for serving, validating and consuming OPTIMADE APIs in Python,[36] available on GitHub (**https://github.com/Materials-Consortia/optimade-python-tools**). Now in version 1.0, it provides a highly extensible reference server implementation of an OPTIMADE API, with support for different database backends. This server is provided as a Docker container for easy deployment and can be configured to use an existing database, or generate one from scratch in the OPTIMADE format. Existing databases wanting to make use of the library need to provide mappings to and from their existing data format and query mechanisms. The package contains isolated modules for various OPTIMADE-related functionalities, for example, a grammar and parser for the OPTIMADE filter language, mappers for querying different database backends, and a fuzzy validator that can dynamically generate requests to an OPTIMADE server to assess its compliance with the standard (which is also used to generate the OPTIMADE provider dashboard: **https://www.optimade.org/providers-dashboard/**).[130] In addition to server-focused functionalities, the package includes reusable code that can help OPTIMADE consumers and clients, including adapters for converting OPTIMADE entries into common formats used in the community, such as ASE `Atoms`,[72] pymatgen `Structures`[90] and AiiDA nodes.[60]

The package also contains an advanced asynchronous HTTP client that can be used within Python code or at the command line to concurrently filter multiple OPTIMADE databases with

multiple queries, paginating and validating the results, as well as searching across databases for supported properties. Users can provide callbacks to store the results in local secondary databases for re-use in other projects.

`optimade-python-tools` is fully documented online at **https://www.optimade.org/optimade-python-tools/latest/** with guides for setting up and validating an API, deploying a server and using the client. In this way, `optimade-python-tools` significantly lowers the barrier to retrieving data from OPTIMADE APIs, and to the development of new server implementations. Future developments will focus on extending the use cases of `optimade-python-tools` to operating on static data, so that it can be embedded within archival infrastructure, such as Materials Cloud,[127] to serve user data without any additional input.

2.3.2 Optimade-gateway. `optimade-gateway` is an open-source software (MIT license) package implementing a RESTful API in Python for querying OPTIMADE databases, available on GitHub at the Materials-Consortia gateway (**https://github.com/Materials-Consortia/optimade-gateway**). This provides a powerful yet straightforward tool to allow users of the Python programming language (that is widespread in machine learning and other branches of data science) to access material database results. `optimade-gateway` supports both synchronous and asynchronous searches *via* HTTP GET and POST requests, respectively. It can return search results in the standard OPTIMADE data format, as well as a custom OPTIMADE-inspired data format. The main purpose and goal of the deployed service is to be a client backend. A gateway to version 0.4 is running at **https://mmp-optimade-gateway.materialscloud.io/redoc** as well as at **https://optimade-gateway.fly.dev/redoc**. Both services utilise a MongoDB database for time-limited caching of the query results, increasing response speeds for common queries.

2.3.3 Optimade.Science. Optimade.Science (**https://optimade.science/**) is a minimalist in-browser OPTIMADE aggregator, written in the TypeScript language on top of the Svelte frontend framework.[126] It fetches the official OPTIMADE providers list, looks for the structure endpoints, and allows simultaneous querying against all of them, collecting the results together on a single webpage. Technically, this is just the single file `index.html` and is thus highly-portable, can be opened from anywhere, on any environment (*e.g.*, on a smartphone or locally from a USB stick). To increase ease-of-use, a simple pattern-matching library was developed (**https://github.com/mpds-io/optimade-mpds-nlp**), transforming the free-text user input into a standard OPTIMADE query (*e.g.*, a keyword `ternary` is transformed into `nelements=3`). A separate Svelte user interface kit (**https://github.com/basf/svelte-spectre**) was developed offering a range of the modular GUI components, willingly accepted by the frontend community and already re-used in many other web-projects, including commercial ones. A standalone OPTIMADE client written in TypeScript was employed, being fully isomorphic (that is, the same code can be used inside the web browser and on the web server).

3 Contributing databases

The burgeoning community of materials databases are the core that underpins the OPTIMADE consortium. Therefore, below we discuss the key features of the major materials databases that make data available through the OPTIMADE API. We first briefly introduce each database and its offering, and we then discuss its particular OPTIMADE implementation. Finally, we provide an updated table from our previous work[3] that compares the amount of compliant data available in different databases.

3.1 AFLOW

Database. One of the largest open-access databases for inorganic materials, with 4 million compounds and 800 million associated properties,[35,92] which also includes the 2000+ entries of the AFLOW encyclopedia of crystallographic prototypes.[55,56,83] The data has been employed for the discovery of new permanent magnets,[107] superalloys,[89,104] high-entropy high-hardness plasmonic carbides,[1] super-hard disordered carbides,[108] borides and carbo-nitrides,[30] and phase-change memory compositions,[71] and has also been used to study bulk metallic glasses,[39,98] superconductors,[62,124] and thermoelectrics.[135] The data can be retrieved conveniently through the AFLUX search API[105] with a minimal, flexible, and human-readable query language.

OPTIMADE implementation. The AFLOW OPTIMADE API builds on AFLUX to offer a common query syntax across multiple materials databases, mapping AFLOW property labels to that of OPTIMADE while still offering access to AFLOW-specific properties with the `_aflow_` prefix. A full list of keywords available to use with OPTIMADE to query AFLOW are available at the info endpoint `/info/structures`.

Base URL: **https://aflow.org/API/optimade**

3.2 Alexandria

Database. Comprises both hypothetical and existing compounds, which have been relaxed using density functional theory (DFT). Currently, the database contains 5 062 521 entries, spanning nearly the entire periodic table with 89 elements. The database was primarily generated by scanning binary, ternary, and quaternary prototypes to identify stable compounds. This process employed crystal graph attention networks[113,114,117] to predict the stability of all potential compositions for each prototype. Compounds that were found to be close to stability were subsequently confirmed using DFT. Additionally, the database includes compounds obtained from traditional high-throughput searches conducted previously.[112,115,137]

Most entries were calculated using the PBE functional with parameters mostly consistent with those of the Materials Project,[66] and includes 3D (4 489 295 entries), 2D (137 833) and 1D (13 295 entries) compounds. Additionally, a total of 422 098 materials were computed using the PBEsol and SCAN functionals to yield more precise geometries, formation energies, and bandgaps.[116] The PBE version of the Alexandria database, which comprises 115 535 potentially stable materials,

represents the most extensive publicly available DFT convex hull of thermodynamic stability in our knowledge. Furthermore, 771 696 materials lie within a distance of less than 50 meV per atom from the convex hull. The database encompasses various properties, including structure (lattice and atomic positions), energy distance to the convex hull, formation energy and direct as well as indirect bandgaps. Continuous expansion of the database is underway through further ongoing high-throughput searches.

OPTIMADE implementation. Uses the `optimade-python-tools`[36] reference implementation and provides a list of extra properties with the `_alexandria_` prefix. Prior to this development, the Alexandria database was made available solely as a static archive that users had to download in its entirety to explore, but now the OPTIMADE format supports filtering on both composition and the predicted stability of database entries.

Base URL: **https://alexandria.icams.rub.de**

3.3 BioExcel COVID-19

Database. A platform designed to provide online access to atomistic molecular dynamics trajectories for biological macromolecules related to the COVID-19 disease.[12] The project is part of the open access initiatives promoted by the world-wide scientific community to share information about COVID-19 research and integrate technology developed in previous biology related projects.[6,58,59,148]

OPTIMADE implementation. A web-server interface **https://bioexcel-cv19.bsc.es** presents the MD trajectories, with a set of quality control analyses and system information. Using an extension of version 1.1.0 of the OPTIMADE specification, a basic OPTIMADE server based on the `optimade-python-tools` has been set up, which provides the trajectory data at the `/trajectories` endpoint. This server also provides protein specific properties and metadata under database specific fields with the `_bioxl_` prefix. Querying has been partially implemented, but is not yet available for all fields. No atomistic structures are shared, so the `/structures` endpoint is not available.

Base URL: **https://bioexcel-cv19.bsc.es/optimade/**

3.4 Computational Materials Repository (CMR)

Database. CMR is a repository of databases containing calculated atomic structures and basic properties of a broad set of materials. The CMR databases can be browsed online using a simple querying system or downloaded in various formats (**https://cmr.fysik.dtu.dk**). Currently, the CMR holds more than 30 different databases.

The flagship database in CMR is the Computational Two-Dimensional Materials Database (C2DB)[44,50,88] that contains structural, thermodynamic, elastic, electronic, magnetic, and optical properties of more than 15 000 two-dimensional (2D) monolayer materials computed using the GPAW[34,86] package. The core set of materials in C2DB have been obtained by extracting monolayers from experimentally known layered van der Waals crystals. Subsequently, new monolayers have been generated by systematic atom-substitution applied to the core materials, or using deep generative AI models.[78] Recently, the C2DB has been complemented by the BiDB database containing homobilayers formed by stacking 1000 of the most stable monolayers from the C2DB in all possible commensurate configurations.[94]

OPTIMADE implementation. CMR implements the OPTIMADE API through the CAMD-web package (**https://gitlab.com/camd/camd-web**), utilising the `optimade-python-tools` library. At present only C2DB is available *via* an OPTIMADE API, but in the future other CMR databases will also be available *via* OPTIMADE.

Base URL: **https://cmr-optimade.fysik.dtu.dk**

3.5 Crystallography Open Database (COD)

Database. The largest open access collection of experimental crystal structures.[49] It is widely used by the scientific community to explore different material categories such as superconductors,[21] metal–organic frameworks,[11] high entropy alloys,[118] organic molecules[17] as well as for conformer sampling[84] or custom force field generation.[45] Having a set of experimental structures readily available under the same format as is required for computational materials research is highly beneficial since these structures serve as initial points for material property calculations[127] or for the search of new materials.[101] They also serve as experimental points that theoretical computations can be checked against.[138]

OPTIMADE implementation. The COD database currently implements version v1.1.0 of the OPTIMADE standard. The new OPTIMADE version will allow this implementation to be enriched with new features that are required for the faithful representation of experimental data, thus making computations from these data and comparisons of theory and experiments more accurate. Being an experimental structure database, the COD requires a slightly different data presentation than computational material databases. Structures from the COD are evaluated using a set of experimental data quality criteria, established by the IUCr and the chemical crystallography community.[63] An OPTIMADE response within the core features does not contain all of the necessary fields to convey these additional data elements; however, the OPTIMADE standard allows introducing database-specific fields in a regular way. As a result, all established crystallographic quality criteria are included into a COD response as COD-specific fields with the `_cod_` prefix. This allows OPTIMADE to include experimental position and composition disorder information following the Crystallographic Information Framework.[51]

Base URL: **https://www.crystallography.net/cod/optimade**

3.6 Joint Automated Repository for Various Integrated Simulations (JARVIS)

Database. A repository designed to automate materials design using classical force-field, density functional theory (DFT), machine learning calculations and experiments. The JARVIS-DFT originated about 5 years ago and contains millions of properties materials with carefully converged atomic

structures as well as tight convergence parameters and various exchange–correlation functionals. The JARVIS-DFT contains metallic, semiconducting, insulator, superconductor, high-strength, topological, solar, thermoelectric, piezoelectric, dielectric, two-dimensional, magnetic, porous, defect and various other classes of materials.[22,142]

OPTIMADE implementation. Based on the Django Rest Framework and the JARVIS-Tools packages to follow OPTIMADE protocols of filtering and curating data. JARVIS-DFT specific fields are included in the results with the `_jarvis_` prefix.

Base URL: **https://jarvis.nist.gov/optimade/jarvisdft**

3.7 Materials Cloud

Database. A platform created to enable sharing and dissemination of resources in computational materials science.[127] A major service offered is the archiving and publishing of research data for the community *via* the open Materials Cloud Archive service (**https://archive.materialscloud.org**). Moreover, several databases that are generated within the AiiDA[60,100,133] framework are published in the Materials Cloud Explore section, which enables users to interactively browse the data and its provenance. Curated visualisations of these databases are also provided in the Materials Cloud Discover section. These databases are accessible *via* the OPTIMADE RESTful API. The databases are divided into flagship and contributed databases. The current flagship databases are MC3D and MC2D[15,87] hosting over 34 000 3D crystals and 3000 2D crystals, respectively, providing properties of experimentally-known inorganic compounds obtained *via* DFT simulations. The contributed databases include 2D topological insulators, pyrene-based metal organic frameworks, high-throughput Wannierisation, $SrTiO_3$–CeO_2 interfaces, tail-corrections in the molecular simulations of porous materials, hidden spontaneous polarisation in the chalcohalide photovoltaic $Sn_2SbS_2I_3$, and the CURATED covalent organic frameworks database. The data has been used to investigate transport properties such as mobility,[122,123] to search for $\mathbb{Z}_2$ topological order,[80] to screen and discover quantum spin-Hall insulators[47,79] and Weyl semimetals,[48] to obtain tight-binding-like Wannier Hamiltonians in a fully automated fashion,[103] and to develop machine-learning methods for fast identification of low-dimensional materials.[134]

Current developments are focusing on a second "Materials Cloud Archive" provider, allowing submissions to Materials Cloud Archive to specify whether and how data contributed by users should be served *via* an OPTIMADE API, to enable advanced federated search over archived data. The first proof-of-concept of this integration is a recently published dataset of novel electride materials.[139,140] The full list of OPTIMADE databases served by the Materials Cloud can be explored at the `/main/v1/links` endpoint, with a similar list for the Materials Cloud Archive available at `/archive/v1/links`. A landing page for both OPTIMADE providers is available at **https://www.materialscloud.org/optimade**.

OPTIMADE implementation. The data on the backend of Materials Cloud is managed *via* AiiDA. Along with a custom REST API, AiiDA can serve data in the OPTIMADE format thanks to the AiiDA-OPTIMADE (**https://github.com/aiidateam/aiida-optimade**) plugin, that is thus also used to serve the main Materials Cloud data. The OPTIMADE implementation of the Materials Cloud Archive provider is instead based directly on the `optimade-python-tools`[36] package. In addition to the server implementations, Materials Cloud also offers users several web applications that include clients of the OPTIMADE API, as we describe in more detail in Section 4.2.1.

Index base URL: **https://www.materialscloud.org/optimade/main**

Index base URL: **https://www.materialscloud.org/optimade/archive**

3.8 Materials Platform for Data Science

3.8.1 Database. Materials Platform for Data Science (MPDS) serves the Pauling File dataset.[95] Started in 1993, Pauling File is the oldest privately funded initiative for the curation and standardisation of the published inorganic chemistry data.¶ Data is drawn from nearly 400 thousand publications and backs up such commercial products as Springer Materials, ICDD PDF 4+, ASM's Alloy Phase Diagram Database and Pearson's Crystal Data, MedeA, and AtomWork Advanced.

OPTIMADE implementation. MPDS presents curated experimental data of three types: crystalline structures (`/structures` endpoint), physical properties (`/extensions/properties` endpoint), and phase diagrams (`/extensions/phase_diagrams` endpoint). These three data types are inter-linked into about 200 thousand distinct phases (`/extensions/phases` endpoint). Any distinct phase is uniquely determined by the chemical formula, space group, and Pearson symbol. Furthermore, each distinct phase has the permanent integer identifier `phase_id`, *e.g.*, see brookite (**https://mpds.io/#phase_id/27712**). The MPDS OPTIMADE implementation is specifically designed for the low response time and high retrieval speed, therefore some expensive operators (`ANY`, `OR`) are currently not supported (*cf.* Table 1†).

Base URL: **https://api.mpds.io**

3.9 Materials Project

Database. This multi-institution, multi-national effort[66] aims at computing the properties of all inorganic materials and providing the data and associated analysis algorithms for every materials researcher free of charge. Currently, over 172k molecules and over 154k inorganic compounds are included in the database. The project was established in 2011 with an emphasis on battery research, but includes property calculations for many areas of clean energy systems such as photovoltaics, thermoelectric materials, and catalysts.

¶ The double awarded Nobel laureate Linus Pauling personally endorsed this project and gave an explicit written permission to use his name. In 2019, the Pauling File's founder Pierre Villars was acknowledged with the NIMS Award (Tsukuba, Japan) for the fundamental research for data-driven materials development.

Table 1 The table from ref. 3 recreated in March 2024, with new providers[a]

Provider	N_1	N_2	N_3	N_{tot}
AFLOW	704 302 (700 192)	63 017 (62 293)	413 797 (382 554)	3 530 330
Alexandria*	939 084	48 510	437 768	5 055 842
COD	458 249 (416 314)	4082 (3896)	34 739 (32 420)	512 282
CMR	2811	386	0	16 789
JARVIS-DFT	9017	1426	8084	77 096
Materials cloud*	961 564	4218	136 176	4 515 120
Materials project	34 424 (27 309)	3750 (3545)	11 861 (10 501)	154 387
MPDD	811 136	80 195	490 900	3 975 666
MPOD	91	8	16	401
MPDS	—	—	—	507 178
NOMAD	4 451 056 (3 359 594)	587 923 (532 123)	2 092 989 (1 611 302)	12,116 021
odbx*	125 648 (55)	3179 (54)	17 009 (0)	523 216
omdb	58 718 (58 718)	690 (690)	7428 (7428)	68 566
OQMD	261 400 (153 113)	15 375 (11 011)	81 673 (70 252)	1 226 781
TCOD	7161 (2631)	296 (296)	662 (660)	7452
2DMatpedia	1172	739	255	6351

[a] The final column indicates the total number of structures served by each OPTIMADE API. Providers that serve multiple databases are indicated with *. Results for Materials Cloud and Materials Cloud Archive have been aggregated under the same title. The corresponding values from the 2021 paper[3] are provided in brackets, where appropriate.

OPTIMADE implementation. The Materials Project (MP)[66] makes use of the reference server implementation provided by the `optimade-python-tools`.[36] Since May 2022, Materials Project has been serving formation energy data *via* its OPTIMADE API. In June 2023, MP started exposing additional thermodynamic stability in the form of energy distance to the convex hull *via* OPTIMADE for all 154k materials in its core database. The convex hull distance to the Materials Project is one of the most important properties of theoretical structures for experimentalists and simulators alike, as it indicates whether a postulated material is potentially synthesisable.

The MP OPTIMADE integration is complemented by a convenient open-source pymatgen[90] interface in the form of the `OptimadeRester` class (**https://github.com/materialsproject/pymatgen/blob/ec750ca15d02cdd51b0c0a7a4408af8e0d259223/pymatgen/ext/optimade.py#L31**), designed to streamline access to these resources for existing users of pymatgen and the Materials Project. Additionally, efforts are underway to further expose the full set of MP `summary` data *via* the OPTIMADE endpoint under the `_mp_` namespace, mirroring the complete set of data recorded in the `emmet` SummaryDoc (**https://github.com/materialsproject/emmet/blob/bf8a4ef09a0d9f91bb6e9fe3e2fca0acd3582306/emmet-core/emmet/core/summary.py#L137**).

Base URL: **https://optimade.materialsproject.org**

3.10 Material-Property-Descriptor Database

Database. Material-Property-Descriptor Database (MPDD) is an extensive database (4M+) of *ab initio* relaxations of 3D crystal structures, combined with an infrastructure of tools allowing efficient descriptor calculation (featurization), as well as the deployment of ML models.[68] The most critical feature of the MPDD is the retention of intermediate modelling data, including structure-informed descriptors, which typically cost orders of magnitude more computational time than any of the other steps performed during ML model deployment.[69] Thus, many ML models can be run at a small fraction of the original cost if the same descriptor (or, more commonly, a subset chosen through feature selection) is used. This benefit applies regardless of whether a model is just another iteration, *e.g.*, fine-tuned to a specific class of materials like perovskites, or an entirely new model for a different property. Furthermore, MPDD's access to stored atomic structures and associated metadata has been shown to be useful, for instance, in the fully data-driven prediction of atomic structures (validated with DFT and experiments), allowing quick identification of unknown structures in Nd–Bi[61] and Al–Fe[119] systems.

OPTIMADE implementation. MPDD has a stable OPTIMADE API that serves the entire core MPDD dataset, fully implementing v1.1.0 of the OPTIMADE standard through a server based on `optimade-python-tools`.[36] Making the MPDD available *via* OPTIMADE was initially challenging, as MPDD stores and exchanges data in a way that prioritises high throughput and low storage requirements, including binary data, making it difficult or slow to make MPDD queryable as an OPTIMADE API on-the-fly. However, issues have been resolved by establishing a self-updating mirror of the dataset where structures are made OPTIMADE-compliant during transfer and with most associated MPDD-specific data available under the `_mpdd_` namespace, including dictionaries of metadata, properties, and descriptors.

Base URL: **http://optimade.mpdd.org**

3.11 Materials Properties Open Database

Database. The Material Properties Open Database (MPOD)[40,97] is a web-based, open access repository of experimentally determined quantitative information about the physical properties of crystalline materials. MPOD is oriented at design engineers, scientists, science teachers and students.

Properties are generally treated as tensor magnitudes. In MPOD the compact matrix notation is applied. To bring an intuitive view of tensor properties, so-called longitudinal properties surfaces are displayed. 3D printing of properties surfaces is implemented *via* creation of STL files. A dictionary of properties definitions is included. Eventually, comments are added. Syntax and notation in MPOD files are oriented towards matching IUCr standards and so tries to comply with CIF format.

OPTIMADE implementation. The integration of OPTIMADE with MPOD encompassed two distinct phases. Initially, the process entailed migrating all data from the MySQL database to MongoDB. This was followed by the mapping of MPOD objects to OPTIMADE, utilizing `optimade-python-tools`.[36] In this context, the prefix `_mpod_` was employed to delineate specific database fields.

Base URL: **http://mpod_optimade.cimav.edu.mx**

3.12 Materials Resource Registry

Database. A federated, decentralised registry of resources in the domain of materials science. It exposes these resources to users and machines *via* XML and OAI-PMH APIs.[102]

OPTIMADE implementation. OPTIMADE has been added to the Materials Resource Registry as an API format that other services and datasets can link to, to indicate their own compliance. Materials Resource Registry's rich semantic description of databases, with regards to their scientific content, techniques, and material focus,[82] as curated by the provider, enables users to make expressive queries over OPTIMADE providers, to narrow down which databases may be of interest to them. This makes it much easier to discover data and direct clients to the resources that are scientifically the most relevant to them.

3.13 Matterverse

Database. A database of yet-to-be-synthesised materials predicted using state-of-the-art ML models, currently comprised of 31 664 858 hypothetical materials. The current structures were generated by combinatorial isovalent ionic substitutions on 5283 binary, ternary, and quaternary structural prototypes from the 2019 version of the ICSD database. A critical enabler for this database is the Materials 3-body Graph Network (M3GNet) universal interatomic potential encompassing 89 elements of the periodic table.[18] Along with the information of lattice parameters, atom coordinates and E_{hull}, matterverse.ai (**https://matterverse.ai/**) also provides the predicted formation energies, bandgaps (of multiple fidelities, including PBE, HSE and experimental), and bulk and shear moduli. As an ongoing effort, matterverse.ai is growing in two directions, (i) increasing the number of hypothetical materials *via* various structure generation strategies, and (ii) increasing the number of ML-predicted properties.

OPTIMADE implementation. Support of the OPTIMADE API is under active development, with so far successful mapping of data to the OPTIMADE format using the `optimade-python-tools`.[36]

Base URL: **https://optimade.matterverse.ai**

3.14 NOMAD

Database. An open-source software and free service for managing and publishing FAIR[110] materials science data. NOMAD[31,32] was made publicly available in 2014; it provides over 12 million data entries from over 500 researchers.[111] Originally, NOMAD focused on *ab initio* codes based on density-functional theory (DFT), automatically extracting data and metadata from input and output files. Meanwhile, NOMAD was significantly expanded in scope by the consortium FAIRmat (**https://www.fairmat-nfdi.eu/fairmat/**). It now supports file types from over 60 simulation codes, it encompasses advanced many-body calculations, including GW, the Bethe–Salpeter equation (BSE), and dynamical mean-field theory (DMFT), and classical molecular dynamics simulations. It can cope with different types of experimental data. For instance, it provides support for electronic lab notebooks and the NeXus format. NOMAD can track data provenance in complex simulation and experiment workflows.

NOMAD enables individual researchers to make their data available to a wide range of possible clients and applications. All data is formally described through rich metadata schema[41] and can be analyzed with build in containerised tools and notebooks.[109] Data in NOMAD is provided through the OPTIMADE API, NOMAD specific APIs, and a rich graphical user interface with faceted search, (meta)data explorer, and visualisations for material properties. API access is particularly important for re-use of the data, *e.g.* with artificial-intelligence (AI) methods. A collection of AI tools is available in the NOMAD AI Toolkit.[109] Besides supporting the community with the central data infrastructure, NOMAD offers the same software[111] for local installation through NOMAD Oasis, which allows research groups to manage and provide their own research data individually and customise the software accordingly.

OPTIMADE implementation. NOMAD supports a full OPTIMADE API implementation based on the[36] using the Elasticsearch database engine, and a web-based search interface that allows users to formulate queries based on the standardised OPTIMADE query strings. Furthermore, NOMAD users can search for related resources from all other OPTIMADE database providers in the OPTIMADE provider list.

Base URL: **https://nomad-lab.eu/prod/rae/optimade**

3.15 Open Database of Xtals (odbx)

Database. A small database serving selected phase diagrams studied with *ab initio* crystal structure prediction techniques.[37,52] Recently, odbx has been used to ingest new materials discovery datasets into the OPTIMADE ecosystem as part of the optimade-misc sub-database (**https://optimade-misc.odbx.science/**),[46,73,147] as well as the GNome dataset[85] at **https://optimade-gnome.odbx.science/**, as will be discussed in Section 4.1.2.

OPTIMADE implementation. odbx was created using the matador[37] and `optimade-python-tools`[36] packages. As well as serving the standard OPTIMADE properties, odbx also serves stability data (hull distances, formation energies) and the DFT parameters used to relax the structures under the `_odbx_`

namespace. odbx serves multiple distinct datasets; the `links` endpoint of the index base URL below can be used to retrieve them.

Index base URL: **https://optimade-index.odbx.science**

3.16 Open Materials Database (*omdb*)

Database. *omdb* provides materials properties and is maintained by the developers of the High-Throughput Toolkit (*httk*).[7] It contains 205 264 structures for access *via* programmatic interaction using this toolkit. The structures are also accessible *via* a web interface. Recently, it is being integrated in the broader database effort Anyterial (**https://www.anyterial.se/**), which also includes the ADAQ database of point defects.[27]

OPTIMADE implementation. *omdb* uses the built-in implementation of the OPTIMADE API provided in *httk*. This implementation is written in Python using no dependencies beyond the Python standard library. Work is currently ongoing to extend the implementation to fully support version v1.2.0.

Base URL: **https://optimade.openmaterialsdb.se**

3.17 Open Quantum Materials Database

Database. The Open Quantum Materials Database (OQMD) holds over 1 million materials, consisting of both experimental and hypothetical compounds.[106,120] The overarching interest of OQMD is to understand the competing stability between known and unknown compounds by generating large-scale convex hulls – a stable yet-to-be-synthesised material should fall along or close to this convex hull. In addition, OQMD grows and develops organically with interests in Wolverton group, including calculations targeting thermoelectrics,[53] battery materials,[8] and high-strength alloys.[67]

OPTIMADE implementation. OQMD currently utilises v1.0.0 of the OPTIMADE standard and will adopt newer versions of OPTIMADE to replace OQMD's current qmpy API. OQMD offers database specific properties through the `_oqmd_` prefix, including formation energies, bandgap, and stabilities of compounds.

Base URL: **https://oqmd.org/optimade**

3.18 Comparison of data available

An important benefit of a universal API format such as OPTIMADE is the ability to simultaneously request and unify results from different databases. While the key features of the databases are highlighted under the subsections dedicated to the respective providers in Section 3, a summarizing list of the implementations tested and confirmed to support the OPTIMADE API is shown in Table 1.† These databases are all openly accessible and provide users with a broad range of materials classes, applications and modalities.

The table shows the return from three requests that explore materials that contain at least one element from Group 14 (N_1), and then constrains the search to cover only binary materials (N_2), and only ternary materials without toxic lead (N_3):

```
/v1/structures?filter=elements HAS ANY "C",
"Si","Ge","Sn","Pb"
```

```
/v1/structures?filter=elements HAS ANY "C",
"Si","Ge","Sn","Pb" AND nelements=2
```

```
/v1/structures?filter=elements HAS ANY "C",
"Si","Ge","Sn" AND NOT elements HAS "Pb" AND
elements LENGTH 3
```

These queries directly duplicate the three detailed in the 2021 OPTIMADE paper.[3] The ability to repeat the query attests to how the OPTIMADE API helps with reproducibility in research. In addition, we provide the total number of structures served by each OPTIMADE API in the N_{tot} column.

Comparison of this table to that in ref. 3 bears witness to the growth and impact of OPTIMADE. We see several additional providers (Alexandria, BioExcel, CMR, JARVIS, MPOD, MPDD and 2DMatpedia) that now support OPTIMADE, and several new databases hosted by pre-existing providers. Furthermore, among the databases that did support OPTIMADE in 2021, there has been an impressive growth in the volume of returned data, reflecting their continued efforts to assimilate further data.

4 Application of OPTIMADE to real-life problems

A key goal for the OPTIMADE API is for it to act as an enabling technology for materials discovery, design and other new research avenues. Feedback from users is crucial to motivate the future development of the API. Therefore, in the following sections, we spotlight several use cases of the application of the OPTIMADE API to real-life systems, firstly in Section 4.1 by supplying data for machine learning, and secondly in Section 4.2 by providing data for screening and other studies. We highlight examples which benefit from access to the wealth of data available in large databases (*e.g.*, the hard-coating alloys database discussed immediately below), and examples that benefit from access to specialist data available only in the small and focused databases (*e.g.*, Section 4.1.1).

Furthermore, there has been additional use of OPTIMADE in the literature: firstly how OPTIMADE has been a central tool to access materials data for materials discovery, secondly as a template for materials data curation and access, and thirdly through online web-based interfaces:

4.1 Discovery

The hard-coating alloys database (HADB)[74] exploited the OPTIMADE API to rapidly and easily provide the browser-based graphical web interface as well as a RESTful API. A second application of the OPTIMADE API was to query and retrieve an unprecedented volume of data to train an attention-based crystal graph convolutional neural network to accurately predict the formation energy, total energy, bandgap, and Fermi energy of a broad range of crystals.[136] Finally, OPTIMADE has found application in materials discovery, where it was used in ref. 54 to assess the novelty of predicted structures in a high-

throughput study on quaternary mixed metal chalcohalide perovskites using the `optimade-python-tools` client.[36] As these structures were ingested into an OPTIMADE-compliant database, in this case NOMAD,[111] any future OPTIMADE queries for novelty in this chemical space will yield the results of this study.

4.2 Template

Development of the OPTIMADE API has motivated and guided data access in other ongoing projects. For example, firstly OPTIMADE API collaborates with, and is being used in, the development of the FAIRmat metadata, dictionaries, and materials ontology.[110] The inclusion in other community efforts reflects the maturity and uptake of the OPTIMADE API. A second example is the BIG-MAP project, where the consortium plans to use the OPTIMADE API to guide the access of the data gathered in the Battery Interface Genome.[16]

4.3 Interfaces and integrations

The MarketPlace Project[128] has integrated the OPTIMADE Gateway (Section 2.3.2) into the platform, which will make it possible to perform OPTIMADE queries through its global search functionality. OntoTrans[91] has developed the Open Translation Environment to perform ontology-driven data pipelines to retrieve, parse, map, and transform data. As part of the project, an OPTIMADE plugin has been developed for the system, making it possible to request and digest OPTIMADE resources. The next steps include semantic mappings for the OPTIMADE data models for true semantic data interoperability.

4.4 Machine learning

Machine learning is a promising tool that is already having a significant impact in the materials sciences. Machine learning starts from already computed data about a system and trains a model to capture trends. The machine learning model can then make predictions and design materials quicker and more cost effectively than performing additional experiments.

Machine learning relies on having a pool of historical data available. This is where the OPTIMADE API offers a significant boost, by opening access to a wide range of materials databases that hold complementary data. We highlight the help offered by the OPTIMADE API to machine learning with two case studies.

4.4.1 High-entropy alloys. High-entropy alloys are comprised of roughly equal parts of five or more elements. This endows the alloy with a high entropy of mixing, which in turn delivers excellent high-temperature properties, such as strength-to-weight ratio, corrosion resistance, and fracture resistance. These favourable properties have driven an acceleration in research into high-entropy alloys over the last decade, but this means that there is still relatively little historical data available for methods such as machine learning.

We use the OPTIMADE API to retrieve high-entropy alloy materials using the filter query

```
elements HAS ANY "W","Al","Cd","Zn" AND NOT
elements HAS ANY "B","Cl","F","H","N","O","S" AND
nelements>=5
```

from three different providers (P1, P2, P3).

The complete dataset obtained using OPTIMADE is split into training and testing sets (4 : 1 ratio) while ensuring that ratio of entries from each provider is the same in both training and testing sets. The training set is used to train the "combined" model M-C. Data points from the providers P1, P2, and P3 that appear in the training set are used to train the models M-P1, M-P2, M-P3, respectively. The predictive power of the models is assessed by calculating the R^2 on the same test set.

We choose Random Forest Regressor (with default parameters, from the `scikit-learn` 1.2 package in Python[96]) to construct the machine learning models for our example. Standard structural entries of the OPTIMADE specification, `species_at_sites` and `lattice_vectors`, are used to construct vectors codifying the composition of each material and also to calculate the density of each material. The 'composition vectors', described above, are used as input to machine learning models that are trained to predict the densities (output). Models that are trained on data from only one provider (M-P1, M-P2, and M-P3) perform poorly when tested on data from all the providers ($R^2 = 0.316$, 0.104, −2.79 for M-P1, M-P2, and M-P3 respectively). Meanwhile, the "combined" model that is trained on data from all providers (M-C) performs very well ($R^2 = 0.995$). A comparison of the R^2 values is shown in the top-left of Fig. 2.

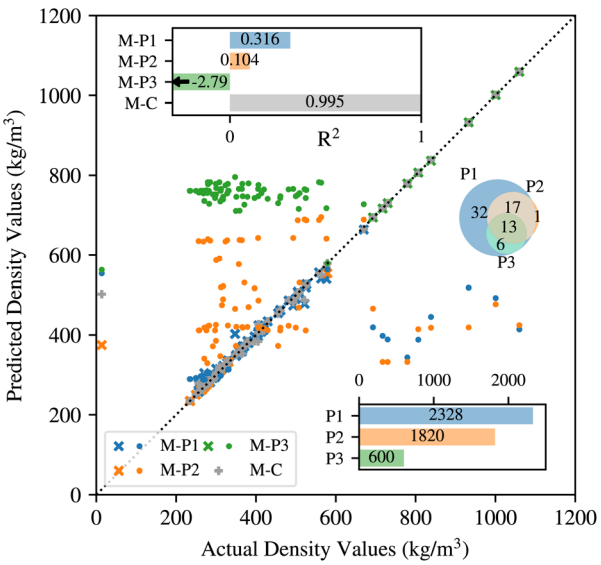

Fig. 2 Scatter plot showing the comparison between actual density values and those predicted by the models trained on data obtained using OPTIMADE (M-P1, M-P2, M-P3 for three providers, and M-C trained on all data). For a particular provider, the × symbols are those points validated against blind data from the same provider, the ● symbols those validated against blind data from a different provider. The top inset shows the R^2 values for each of the models. The bottom inset shows the number of entries returned by each provider. The inset on the centre-right shows the Venn diagram of unique elements in the entries returned by each of the three providers.

We can get a better insight into the benefits of leveraging data from multiple providers by looking at comparison of actual density values and those predicted by the machine learning models for small random sampling of materials (scatter plot in Fig. 2). For models trained only on data from a single provider (M-P1, M-P2, and M-P3), the prediction is quite accurate when tested on data from the same provider (indicted by 'cross' markers). However, most of the predictive power is lost when tested on data from another provider ('dot' markers). This explains their poor R^2 values. Meanwhile, the model which is trained on data from all providers (M-C) retains its predictive power when tested on data from all the providers. The number of materials returned by each provider is shown in the bar-graph on the bottom right of Fig. 2. A Venn diagram of unique elements that appear in the materials from each provider are shown in the centre right in Fig. 2. Therefore, the OPTIMADE API offers the significant benefit to merge the information from the datasets together.

4.4.2 Materials discovery and accelerated design. Recent advances in AI-driven materials discovery have created an abundance of hypothetical crystal structures that are expected to be stable.[18,46,73,85,137,147] New datasets targeted towards materials discovery have been ingested and made available as OPTIMADE APIs within the odbx provider.[37,52] Typically, these datasets would only be explored by other materials discovery specialists, at least until a more established DFT database ran them through their pipelines. With OPTIMADE, this dissemination process can be automated and greatly accelerated. Anyone can register as a provider and have their novel crystal structures appear in searches by other data-driven applications, such as X-ray diffraction phase identification to be discussed below. As OPTIMADE is not limited to purely theoretical crystal structures, any future experimental confirmation of a structure could also be served through OPTIMADE and used to further improve generative models for materials discovery.

Another possible application is to repeat previous high-throughput materials design campaigns on the wider set of structures now available through OPTIMADE. As each structure is time-stamped, active or ongoing workflows can be implemented to constantly monitor and screen new crystal structures against the search criteria of the campaign, to avoid having to redo such searches from scratch. Structure-based property prediction models, *e.g.*, MODNet (for small structure–property datasets)[28,29,76] or graph-based models (where larger structure–property datasets are available),[19,20,143] can be leveraged to sift through huge swathes of available crystal structures, with the results being used to prioritise future calculations, attempts at synthesis, characterisation experiments or model retraining, by selecting for structures with combinations of particular target properties.

4.5 Data provision

The OPTIMADE API has also been used to provide data for a rich variety of other scientific analysis approaches, below we highlight three projects that take advantage of the comprehensive range of data offered by OPTIMADE.

4.5.1 OPTIMADE client: a web-based GUI to find and import structures. The primary input to a first-principles materials calculation is the structure of the system. Experimental or computational crystal databases are commonly used as sources for the input structures of first-principles simulation software. To find the target structure in these databases, a query with filter conditions needs to be prepared, and then the structure needs to be downloaded, inspected, possibly converted into a different format, and finally used in simulations.

To facilitate this goal, Materials Cloud[127] provides the OPTIMADE client, a web application to perform the structure search task *via* a unified and user-friendly GUI, empowering users to not only generate and execute complex OPTIMADE queries, but also to provide immediate graphical access and visualisation of the resulting structures. The OPTIMADE client can be embedded in other applications or used as a standalone tool, which is hosted on Materials Cloud at **https://optimadeclient.materialscloud.io/**. The filtering section of the GUI is shown in Fig. 3. A dropdown (at the top) is provided to select any of the known and automatically discovered OPTIMADE database providers. A periodic table widget allows users to select which elements need to be included (green) or excluded (red) in the compounds; additional filtering tools are also provided, such as for the number of elements and of sites, and for the dimensionality. An OPTIMADE query string is then produced (which can be optionally manually modified). After the search button is clicked, the OPTIMADE query is sent to the selected database provider. The results are curated and shown in the results widget. Here, the structures are visualised and can be downloaded.

Materials Cloud also provides various tools that leverage the power of the OPTIMADE client. One example is the Quantum ESPRESSO input generator, available as a tool at **https://www.materialscloud.org/work/tools/qeinputgenerator**. A structure can be sent directly to this tool using a button in the OPTIMADE client (shown at the bottom of Fig. 3). The Quantum ESPRESSO input generator enables any user to obtain a working input file for the Quantum ESPRESSO DFT code,[42,43] including an automated selection of all numerical parameters, by just specifying a crystal structure (either by uploading it, or by selecting it from OPTIMADE). A second example is the Quantum ESPRESSO app (**https://aiidalab-qe.readthedocs.io**) developed within the AiiDAlab platform.[144] It allows users to run complex computational workflows from the web browser, using straightforward graphical user interfaces for structure selection (including *via* OPTIMADE), parameter selection and inspection of the results.

4.5.2 Automatic phase identification from X-ray diffraction. The Xerus (X-ray Estimation and Refinement Using Similarity)[9,64,93,132] software package implements procedures to refine and screen measured X-ray diffraction patterns of inorganic crystals against databases of crystal structures reported in the literature and beyond. By querying for all possible structures in a given chemical space, Xerus excels at multiphase fits and

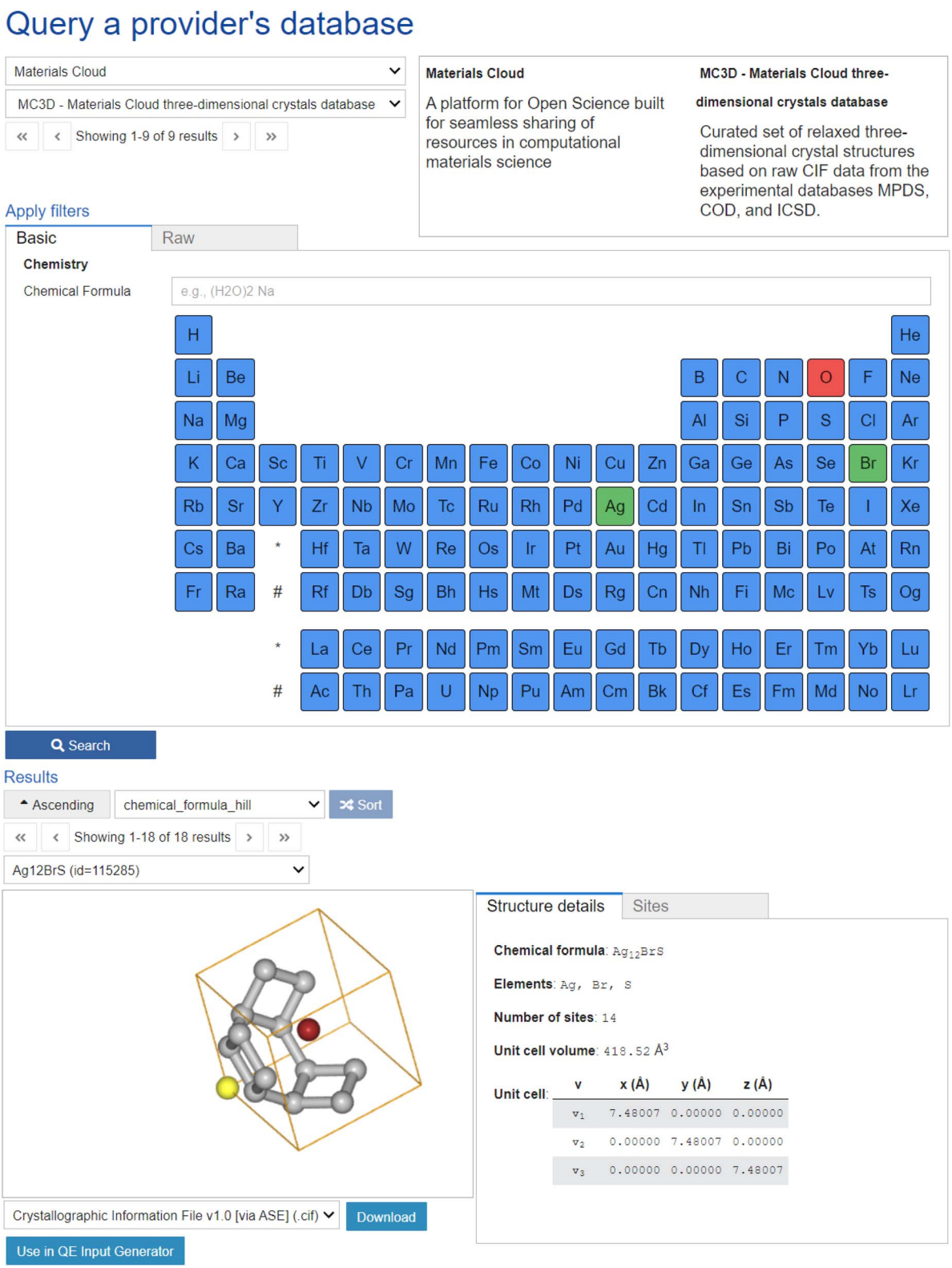

Fig. 3 The search interface of the standalone OPTIMADE client. The user can select any OPTIMADE-compliant database and make a query based on filters specified through GUI elements.

performs competitively against more specialised and compute-intensive models constructed with machine learning.

Xerus uses a straightforward OPTIMADE interface to connect to the multiple databases hosted by members of the OPTIMADE consortia. The dynamic OPTIMADE providers list allows new databases to be automatically included in Xerus search results. Additional filtering parameters can be used to refine the searches towards materials stable (or predicted to be stable) at

the experimental conditions (*e.g.*, low temperature or high pressure).

4.5.3 Workflows for automated and simultaneous queries of different databases. BIOVIA Pipeline Pilot[25] is a scientific workflow system that allows users to automate calculations and visualise and report research results by graphically composing a protocol from hundreds of different configurable components. As a technology demonstrator, we investigate the intercalation voltage of a series of cathode materials with the workflow in Fig. 4(a). It uses OPTIMADE to adopt the available structures and energetics from different providers and performs complementary calculations using the CASTEP DFT code.[23]

Results are shown in Fig. 4(b), where we plot the intercalation voltage of Li–Ni–O materials from the extracted VASP[70] data in Materials Project and NOMAD and compare predictions to those from CASTEP. The comparability of results illustrates the functionality of the workflow. Since databases contain different structures, the OPTIMADE API facilitates the process of materials investigations by aggregating the query of all of them.

MatCloud[81] is a cloud-based integrated high-throughput computational materials infrastructure, which is directly connected to computing clusters and material property databases.[145,146] Users worldwide can visually design structures, create and run simulation jobs through workflows, and retrieve crystal structures from multiple databases using OPTIMADE; all the user needs is a web browser. MatCloud provides a Graphical User Interface (GUI)-based environment for users to intuitively create, enact and monitor a workflow. The MatCloud workflow system includes a front-end workflow designer and a back-end workflow engine, and supports the creation of workflows by a drag-and-drop approach.

Fig. 5(a) shows a workflow that retrieves a Si_8 crystal structure from the MPDS database through OPTIMADE, and replaces two Si atoms with Ge to produce structures in the Si–Ge chemical space (Fig. 5(b)). The band structure and density of states (DOS) are then simulated respectively over the chemical space in a high-throughput manner. Fig. 5(c) and (d) show the visualisation results for the band structure and total DOS of the structures.

5 Future of OPTIMADE

The rapid expansion of materials databases, the use cases presented, and the adoption of machine learning motivate the

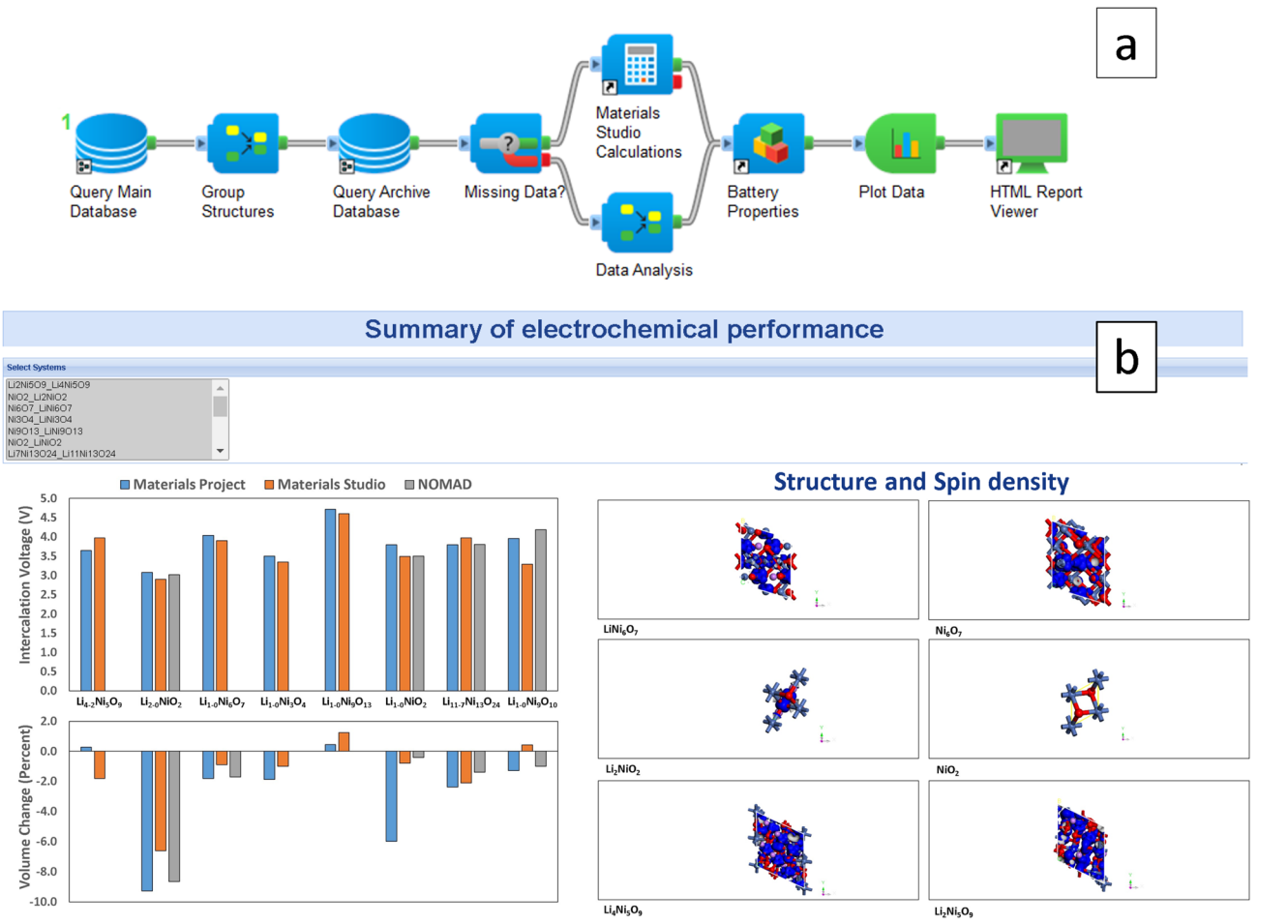

Fig. 4 (a) Workflow architecture for automated simulations using BIOVIA Pipeline Pilot. The workflow queries the structure and energies from the main and archive databases, analyses the available data, and performs CASTEP calculations for the missing data. (b) Screenshot of the app that reports the corresponding electrochemical properties, such as intercalation voltage and crystal structures.

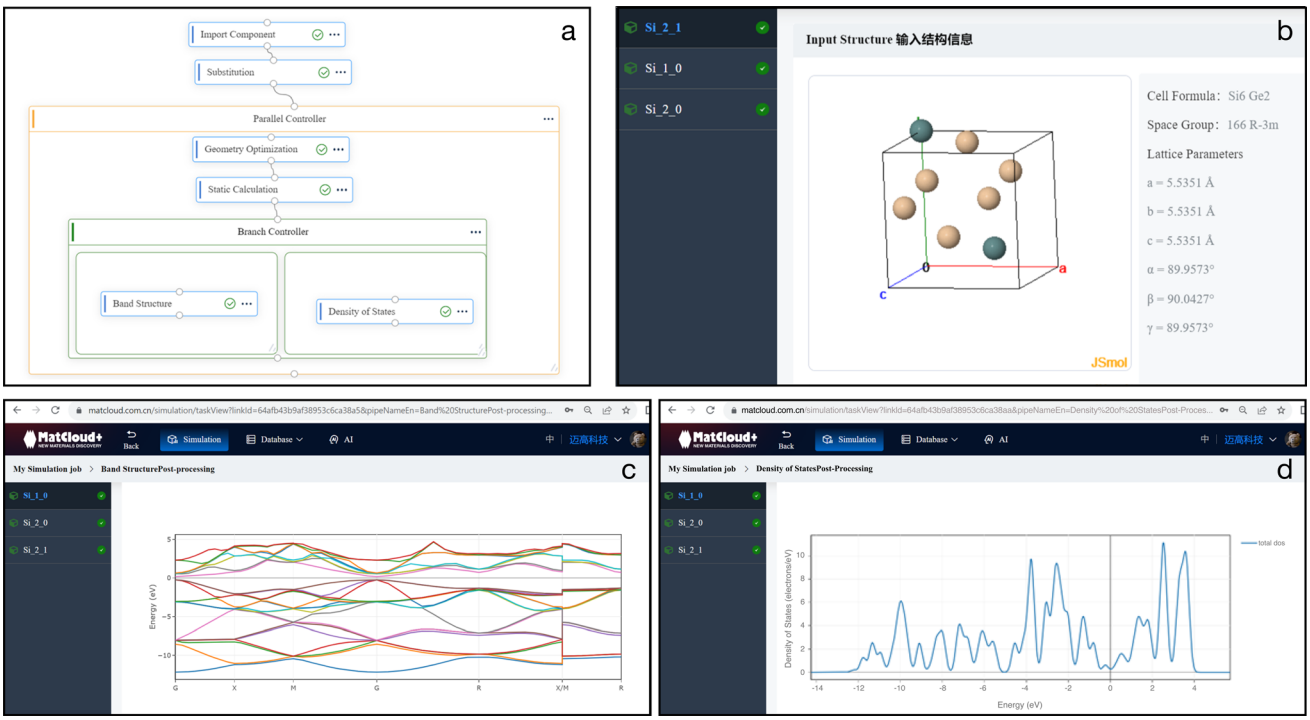

Fig. 5 (a) MatCloud workflow of crystal structure retrieval through OPTIMADE, and job setup. (b) Visualisation of one of the Si–Ge structures. (c) Computed band structure for the Si–Ge structure. (d) Total DOS of the Si–Ge structure.

continued work on the OPTIMADE API. We describe below firstly the ongoing tutorials that introduce new users to OPTIMADE, secondly the workshops that enable the development and promulgation of the OPTIMADE API, and finally the features currently under development coming out of those recent workshops.

5.1 Workshops & tutorials

The OPTIMADE consortium originated from the workshop "Open Databases Integration for Materials Design", held at the Lorentz Center in Leiden, Netherlands in October 2016. There were follow-up workshops held at CECAM in Lausanne, Switzerland annually from 2018 to 2023, with events since 2020 also supporting remote attendees. The workshops in 2022 and 2023 were accompanied by on-site tutorials during the first two days. There was also a partner workshop, "Ontologies for machine learning driven materials design" held at Linköping University, Sweden in 2021. Going forward, the OPTIMADE API will undergo continued development through further annual workshops and publicly advertised monthly video calls. There are continual efforts to reach out to new databases to help accelerate their adoption of the format.

Several tutorial exercises have been developed and delivered at a variety of conferences and topical workshops, outside of the OPTIMADE annual meetings. Eight different exercises developed by the community are now hosted on the Materials-Consortia GitHub repository (**https://github.com/Materials-Consortia/optimade-tutorial-exercises**), ranging from the basics of the OPTIMADE filter syntax and URL structure, to machine-learning pipelines operating on database-specific properties, all the way up to hosting an OPTIMADE API for a new database.

5.2 Upcoming features for future OPTIMADE releases

The ongoing real-life use cases of the OPTIMADE API have identified a series of opportunities to extend the API and make it more applicable to a wide range of materials systems. The features currently under development include:

5.2.1 `/trajectories` endpoint. So far, OPTIMADE can only be used to describe static structures. We are, however, working to expand the OPTIMADE specification, so that OPTIMADE can be used for sharing trajectory data as well. Such data can originate from structural optimisations, or from Monte Carlo and molecular dynamics simulations. These trajectories could be used as a starting point for new simulations, to train machine learning potentials or to extract dynamical properties.

5.2.2 `/collections` endpoint. It has been suggested several times whether it would be possible to have a method to create groups of entries. For example, all the structures that were used to generate a certain machine-learning potential, or all the structures that pertain to a certain research project. We therefore plan to introduce a `/collections` endpoint, which would contain metadata for the collection as a whole, as well as references to all the individual entries that belong to this collection.

5.2.3 Ontologies and semantics. Building on the expanded property definitions outlined above, semantic mappings can be

developed to existing and in-development materials and crystallography domain ontologies. For example, properties of an OPTIMADE structure (such as the specification of periodicity or types of disordered occupation) can be mapped into concepts in the crystallography domain ontology[26] under development within the Elementary Multiperspective Material Ontology (EMMO) ecosystem.[33] This ontology is being created as a collaborative effort that includes members of the OPTIMADE consortium. These mappings of OPTIMADE properties into ontologies facilitate the alignment with other semantic data interoperability frameworks. Examples of such use include the ability to reference properties standardized by OPTIMADE in, *e.g.*, future EMMO-aligned domain ontologies and giving access to property data *via* ontology-based GraphQL server generation.[75]

5.2.4 SMILES property. So far, OPTIMADE has mostly been designed around describing inorganic crystal structures. There are, however, plenty of materials that are (at least partially) comprised of organic constituents. To make it easier to find and select these, we plan to implement a SMILES field[141] and a SMARTS filter[121] for the organic parts of a structure.

5.2.5 Biomolecular fields. One major use case of the upcoming `/trajectories` endpoint is to handle biomolecular structures, which can only be described statistically as a trajectory with multiple configurations. The OPTIMADE specification will therefore be extended to standardise the various dynamical fields and order parameters that are used to describe biomolecular structures. This covers fields that are typically stored in PDB files such as insertion codes, Chain IDs and sequence information of proteins, DNA and RNA.

5.2.6 Proliferation of domain-specific namespaces. As mentioned in Section 2.2, we have already added a mechanism for multiple providers to collaborate on subsets of shared property definitions. It is hoped that many such namespaces will arise to improve coverage of OPTIMADE data across various domains of atomistic science, potentially including the cheminformatics and biomolecular-focused fields above. Other immediate targets include stability information from density-functional theory calculations and magnetic properties.

5.2.7 Large language models (LLMs). LLMs have emerged as an exciting frontier for data science and machine learning.[14,65] We are now considering two uses of LLMs within OPTIMADE. First, a large language model can help a non-expert formulate a query for OPTIMADE, for example the query in section 2 could be found by requesting "tell me the structure of an oxide of silicon"; this could be readily performed by providing the LLM with the specification text or the machine-readable schemas, then constructing the relevant query with in-context learning.[14] A second use is to pass the large language model either textual data or a scan of a page of historical data, which can then be readily parsed to extract out relevant numbers for an OPTIMADE database. The value provided by OPTIMADE here is to give a machine-actionable scaffold that an LLM can be validated and evaluated against, in such a way that the data produced is automatically compatible with other initiatives.

6 Conclusion

The OPTIMADE API provides users with easy access to many of the world leading materials databases. Since the initial release, the OPTIMADE API has not only been adopted by scientists as a tool to drive innovation, but furthermore served as a template for data curation. In this paper, we have provided use cases for how the breadth of data made available through OPTIMADE enables discovery in both academia and industry. The development of the OPTIMADE API has continued apace. Major new support for enhanced property definitions and partial data formats have recently been added and will underpin future work on trajectories and biomolecular data. The concept that sub-consortia of databases are responsible for the definition of new sets of shared properties will accelerate the extension of the OPTIMADE API to other disciplines and fields.

Through monthly meetings, and with the continuing support of CECAM, the developers are continuing to extend the range of properties accessible *via* OPTIMADE APIs. Plans to both expand the format to cover challenges arising from dealing with molecular dynamics data, and continued outreach to support the adoption of the API by additional databases, will further expand the range of scientific use cases that OPTIMADE enables. Increasingly, use cases will take advantage of the unique advantages OPTIMADE has to offer; namely the robust and straightforward aggregation and data unification from the multitude of growing and federated data sources.

Author contributions

These authors were active developers and reviewers of the specification: Johan Bergsma, Matthew L. Evans, and Andrius Merkys. These authors were active developers of implementations for database providers and/or made contributions to specification: Johan Bergsma, Matthew L. Evans, Andrius Merkys, Oskar B. Andersson, Casper W. Andersen, Daniel Beltrán, Evgeny Blokhin, Tara Boland, Rubén Castañeda Balderas, Kamal Choudary, Alberto Díaz Díaz, Rodrigo Domínguez García, Hagen Eckert, Kristjan Eimre, María Elena Fuentes Montero, Adam M. Krajewski, Jens Jørgen Mortensen, José Manuel Nápoles Duarte, Jacob Pietryga, Ji Qi, Felipe de Jesús Trejo Carrillo, Antanas Vaitkus, Jusong Yu, Adam Zettel. These authors primarily contributed to the paper (along with all other authors): Pedro Baptista de Castro, Johan Carlsson, Tiago F. T. Cerqueira, Simon Divilov, Hamidreza Hajiyani, Felix Hanke, Kevin Jose, Corey Oses, Janosh Riebesell, Jonathan Schmidt, Donald Winston, Christen Xie, Xiaoyu Yang. These authors managed individual databases that have implemented the OPTIMADE API: Sara Bonella, Silvana Botti, Stefano Curtarolo, Claudia Draxl, Luis Edmundo Fuentes Cobas, Adam Hospital, Zi-Kui Liu, Miguel Marques, Nicola Marzari, Andrew Morris, Shyue Ping Ong, Modesto Orozco, Kristin A. Persson, Kristian Thygesen, Chris Wolverton. These authors are organisers of the OPTIMADE API and are also senior developers who contributed to code and/or to the specification: Rickard Armiento, Gareth J. Conduit, Saulius Gražulis, Giovanni Pizzi, Gian-Marco Rignanese, Markus Scheidgen, Cormac Toher.

Conflicts of interest

G. J. C. is a shareholder and Director of Intellegens Ltd. G.-M. R. is a shareholder and Chief Innovation Officer of Matgenix SRL. E. B. is a shareholder and Director of Materials Platform for Data Science OÜ.

Acknowledgements

The authors acknowledge support from CECAM in Lausanne (Switzerland) and the Lorentz Center in Leiden (Netherlands) for hosting OPTIMADE workshops; and for the partner workshop at Linköping University in 2021, support from Psi-k, NCCR MARVEL (a National Centre of Competence in Research, funded by the Swiss National Science Foundation, grant no. 205602), and the Swedish e-Science Research Centre (SeRC). G. J. C. would like to acknowledge financial support from the Royal Society. M. L. E. thanks the BEWARE scheme of the Wallonia-Brussels Federation for funding under the European Commission's Marie Curie-Skłodowska Action (Co-fund 847587). CT thanks NSF (DMR-2219788). M. S. and C. D. acknowledge support from the German Research Foundation (DFG) through the NFDI consortium FAIRmat (project 460197019). K. E., J. Y., S. B., N. M., and G. P. acknowledge funding by the NCCR MARVEL (a National Centre of Competence in Research, funded by the Swiss National Science Foundation, grant no. 205602). G. P. acknowledge funding by the Open Research Data Program of the ETH Board (project "PREMISE": Open and Reproducible Materials Science Research, https://open-research-data-portal.ch/projects/open-and-reproducible-materials-science-research/). R. A. and O. B. A. acknowledges funding from Vetenskapsrådet grant no. 2020-05402 and the Swedish e-Science Research Centre. A. M., A. V. and S. G. acknowledge funding under the Programme "University Excellence Initiatives" of the Ministry of Education, Science and Sports of the Republic of Lithuania (Measure No. 12-001-01-01-01 "Improving the Research and Study Environment"). A. J. M. acknowledges support from EPSRC *via* CCP-NC (EP/T026642/1), CCP9 (EP/T026375/1), UKCP (EP/P022561/1) and Baskeville (EP/T022221/1). The work by J. S., D. W., J. Q., S. P. O., and K. A. P. were supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under contract no. DE-AC02-05-CH11231; the Materials Project program (KC23MP). D. B, A. H., and M. O. acknowledge funding by the European Union's Horizon 2020 programme under grant agreements No. 675728 (BioExcel), No. 823830 (BioExcel-2), No. 720270 (Human Brain Project SGA1), No. 785907 (Human Brain Project SGA2), No. 945539 (Human Brain Project SGA3) and by the Horizon Europe Programme under No. 101093290 (BioExcel-3) and No. 101094651 (Molecular Dynamics Data Bank). K. C. acknowledges funding support from the CHIPS Metrology Program, part of CHIPS for America, National Institute of Standards and Technology, U.S. Department of Commerce. X. Y. acknowledges the funding support from National Natural Science Foundation of China (NSFC) under the grant no. 62376258.

References

1 A. Calzolari, C. Oses, C. Toher, M. Esters, X. Campilongo, S. P. Stepanoff, D. E. Wolfe and S. Curtarolo, Plasmonic high-entropy carbides, *Nat. Commun.*, 2022, **13**, 5993, DOI: 10.1038/s41467-022-33497-1.

2 C. W. Andersen, R. Armiento, E. Blokhin, G. J. Conduit, S. Dwaraknath, M. L. Evans, A. Fekete, A. Gopakumar, S. Gražulis, A. Merkys, F. Mohamed, C. Oses, G. Pizzi, G.-M. Rignanese, M. Scheidgen, L. Talirz, C. Toher, D. Winston, R. Aversa, K. Choudhary, P. Colinet, S. Curtarolo, D. Di Stefano, C. Draxl, S. Er, M. Esters, M. Fornari, M. Giantomassi, M. Govoni, G. Hautier, V. Hegde, M. K. Horton, P. Huck, G. Huhs, J. Hummelshøj, A. Kariryaa, B. Kozinsky, S. Kumbhar, M. Liu, N. Marzari, A. J. Morris, A. A. Mostofi, K. Persson, G. Petretto, T. Purcell, F. Ricci, F. Rose, M. Scheffler, D. Speckhard, M. Uhrin, A. Vaitkus, P. Villars, D. Waroquiers, C. Wolverton, M. Wu andX. Yang, *The OPTIMADE Specification*, version 1.0.0, Zenodo, 2020, p. 1, DOI: 10.5281/zenodo.4195051.

3 C. W. Andersen, R. Armiento, E. Blokhin, G. J. Conduit, S. Dwaraknath, M. L. Evans, A. Fekete, A. Gopakumar, S. Gražulis, A. Merkys, F. Mohamed, C. Oses, G. Pizzi, G.-M. Rignanese, M. Scheidgen, L. Talirz, C. Toher, D. Winston, R. Aversa, K. Choudhary, P. Colinet, S. Curtarolo, D. Di Stefano, C. Draxl, S. Er, M. Esters, M. Fornari, M. Giantomassi, M. Govoni, G. Hautier, V. Hegde, M. K. Horton, P. Huck, G. Huhs, J. Hummelshøj, A. Kariryaa, B. Kozinsky, S. Kumbhar, M. Liu, N. Marzari, A. J. Morris, A. A. Mostofi, K. Persson, G. Petretto, T. Purcell, F. Ricci, F. Rose, M. Scheffler, D. Speckhard, M. Uhrin, A. Vaitkus, P. Villars, D. Waroquiers, C. Wolverton, M. Wu and X. Yang, OPTIMADE: an API for exchanging materials data, *Sci. Data*, 2021, **8**, 217, DOI: 10.1038/s41597-021-00974-z.

4 C. W. Andersen, R. Armiento, E. Blokhin, G. J. Conduit, S. Dwaraknath, M. L. Evans, A. Fekete, A. Gopakumar, S. Gražulis, A. Merkys, F. Mohamed, C. Oses, G. Pizzi, G.-M. Rignanese, M. Scheidgen, L. Talirz, C. Toher, D. Winston, R. Aversa, K. Choudhary, P. Colinet, S. Curtarolo, D. Di Stefano, C. Draxl, S. Er, M. Esters, M. Fornari, M. Giantomassi, M. Govoni, G. Hautier, V. Hegde, M. K. Horton, P. Huck, G. Huhs, J. Hummelshøj, A. Kariryaa, B. Kozinsky, S. Kumbhar, M. Liu, N. Marzari, A. J. Morris, A. A. Mostofi, K. Persson, G. Petretto, T. Purcell, F. Ricci, F. Rose, M. Scheffler, D. Speckhard, M. Uhrin, A. Vaitkus, P. Villars, D. Waroquiers, C. Wolverton, M. Wu andX. Yang, *The OPTIMADE Specification*, version 1.1.0, Zenodo, 2021, p. 1, DOI: 10.5281/zenodo.4251947.

5 C. W. Andersen, R. Armiento, E. Blokhin, G. J. Conduit, S. Dwaraknath, M. L. Evans, A. Fekete, A. Gopakumar, S. Gražulis, A. Merkys, F. Mohamed, C. Oses, G. Pizzi, G.-M. Rignanese, M. Scheidgen, L. Talirz, C. Toher, D. Winston, R. Aversa, K. Choudhary, P. Colinet,

S. Curtarolo, D. Di Stefano, C. Draxl, S. Er, M. Esters, M. Fornari, M. Giantomassi, M. Govoni, G. Hautier, V. Hegde, M. K. Horton, P. Huck, G. Huhs, J. Hummelshøj, A. Kariryaa, B. Kozinsky, S. Kumbhar, M. Liu, N. Marzari, A. J. Morris, A. A. Mostofi, K. Persson, G. Petretto, T. Purcell, F. Ricci, F. Rose, M. Scheffler, D. Speckhard, M. Uhrin, A. Vaitkus, P. Villars, D. Waroquiers, C. Wolverton, M. Wu andX. Yang, *The OPTIMADE Specification*, version 1.2.0, 2024, **https://github.com/Materials-Consortia/OPTIMADE/blob/v1.2.0/optimade.rst**.

6 P. Andrio, A. Hospital, J. Conejero, L. Jordà, M. del Pino, L. Codo, S. Soiland-Reyes, C. Goble, D. Lezzi, R. Badia, M. Orozco and J. L. Gelpi, BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows, *Sci. Data*, 2019, **6**, 169–179, DOI: **10.1038/s41597-019-0177-4**.

7 R. Armiento. Database-Driven High-Throughput Calculations and Machine Learning Models for Materials Design, in *Machine Learning Meets Quantum Physics*, Lecture Notes in Physics, ed. K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller, Springer International Publishing, Cham, 2020, pp. 377–395, DOI: **10.1007/978-3-030-40245-7_17**, ISBN 978-3-030-40245-7.

8 M. Aykol, S. Kirklin and C. Wolverton, Thermodynamic aspects of cathode coatings for lithium-ion batteries, *Adv. Energy Mater.*, 2014, **4**(17), 1400690, DOI: **10.1002/aenm.201400690**.

9 P. Baptista de Castro, K. Terashima, M. G. Esparza Echevarria, H. Takeya and Y. Takano, XERUS: An Open-Source Tool for Quick XRD Phase Identification and Refinement Automation, *Adv. Theory Simul.*, 2022, **5**(5), 2100588, DOI: **10.1002/adts.202100588**.

10 A. Baratta, A. Cimino, F. Longo, V. Solina and S. Verteramo, The Impact of ESG Practices in Industry with a Focus on Carbon Emissions: Insights and Future Perspectives, *Sustainability*, 2023, **15**, 6685, DOI: **10.3390/su15086685**.

11 M. Barjasteh, M. Vossoughi, M. Bagherzadeh and K. Pooshang Bagheri, MIL-100(Fe) a potent adsorbent of Dacarbazine: Experimental and molecular docking simulation, *Chem. Eng. J.*, 2023, **452**, 138987, DOI: **10.1016/j.cej.2022.138987**.

12 D. Beltrán, A. Hospital, J. L. Gelpí and M. Orozco, A new paradigm for molecular dynamics databases: the COVID-19 database, the legacy of a titanic community effort, *Nucleic Acids Res.*, 2023, **52**(D1), D393–D403, DOI: **10.1093/nar/gkad991**.

13 H. J. Bernstein, J. C. Bollinger, I. D. Brown, S. Gražulis, J. R. Hester, B. McMahon, N. Spadaccini, J. D. Westbrook and S. P. Westrip, Specification of the crystallographic information file format, version 2.0, *J. Appl. Crystallogr.*, 2016, **49**(1), 277–284, DOI: **10.1107/S1600576715021871**.

14 T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, Language Models are Few-Shot Learners, *arXiv*, 2020, preprint, arXiv:2005.14165, DOI: **10.48550/arXiv.2005.14165**.

15 D. Campi, N. Mounet, M. Gibertini, G. Pizzi and N. Marzari, Expansion of the Materials Cloud 2D Database, *ACS Nano*, 2023, **17**(12), 11268–11278, DOI: **10.1021/acsnano.2c11510**.

16 I. E. Castelli, D. J. Arismendi-Arrieta, A. Bhowmik, I. Cekic-Laskovic, S. Clark, R. Dominko, E. Flores, J. Flowers, K. U. Frederiksen, J. Friis, A. Grimaud, K. V. Hansen, L. J. Hardwick, K. Hermansson, L. Königer, H. Lauritzen, F. L. Cras, H. Li, S. Lyonnard, H. Lorrmann, N. Marzari, L. Niedzicki, G. Pizzi, F. Rahmanian, H. Stein, M. Uhrin, W. Wenzel, M. Winter, C. Wölke and T. Vegge, Data management plans: the importance of data management in the BIG-MAP project, *Batteries Supercaps*, 2021, **4**, 1803–1812, DOI: **10.1002/batt.202100117**.

17 L. Chan, G. R. Hutchison and G. M. Morris, Understanding ring puckering in small molecules and cyclic peptides, *J. Chem. Inf. Model.*, 2021, **61**(2), 743–755, DOI: **10.1021/acs.jcim.0c01144**.

18 C. Chen and S. P. Ong, A universal graph deep learning interatomic potential for the periodic table, *Nat. Comput. Sci.*, 2022, **2**(11), 718–728, DOI: **10.1038/s43588-022-00349-3**.

19 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals, *Chem. Mater.*, 2019, **31**(9), 3564–3572, DOI: **10.1021/acs.chemmater.9b01294**.

20 K. Choudhary and B. DeCost, Atomistic Line Graph Neural Network for improved materials property predictions, *npj Comput. Mater.*, 2021, **7**(1), 1–8, DOI: **10.1038/s41524-021-00650-1**.

21 K. Choudhary and K. Garrity, Designing high-T_c superconductors with BCS-inspired screening, density functional theory, and deep-learning, *npj Comput. Mater.*, 2022, **8**(1), 244, DOI: **10.1038/s41524-022-00933-1**.

22 K. Choudhary, K. F. Garrity, A. C. Reid, B. DeCost, A. J. Biacchi, A. R. Hight Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, A. Davydov, J. Jiang, R. Pachter, G. Cheon, E. Reed, A. Agrawal, X. Qian, V. Sharma, H. Zhuang, S. V. Kalinin, B. G. Sumpter, G. Pilania, P. Acar, S. Mandal, K. Haule, D. Vanderbilt, K. Rabe and F. Tavazza, The joint automated repository for various integrated simulations (jarvis) for data-driven materials design, *npj Comput. Mater.*, 2020, **6**(1), 173, DOI: **10.1038/s41524-020-00440-1**.

23 S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. J. Probert, K. Refson and M. C. Payne, First principles methods using CASTEP, *Z. Kristallogr.*, 2005, **220**(5–6), 567–570, DOI: **10.1524/zkri.220.5.567.65075**.

24 T. B. R. Company, Chemicals global market report 2022, 2022, **https://www.researchandmarkets.com/reports/5598260/chemicals-global-market-report-2022**.

25 D. S. A. Corporation, BIOVIA Pipeline Pilot, 2022, https://www.3ds.com/products-services/biovia/products/data-science/pipeline-pilot/.
26 Crystallography Domain Ontology, https://github.com/emmo-repo/domain-crystallography, 2019.
27 J. Davidsson, V. Ivády, R. Armiento and I. A. Abrikosov, ADAQ: automatic workflows for magneto-optical properties of point defects in semiconductors, *Comput. Phys. Commun.*, 2021, **269**, 108091, DOI: 10.1016/j.cpc.2021.108091.
28 P.-P. De Breuck, M. L. Evans and G.-M. Rignanese, Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODNet, *J. Phys.: Condens. Matter*, 2021, **33**(40), 404002, DOI: 10.1088/1361-648x/ac1280.
29 P.-P. De Breuck, G. Hautier and G.-M. Rignanese, Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet, *npj Comput. Mater.*, 2021, **7**(1), 1–8, DOI: 10.1038/s41524-021-00552-2.
30 S. Divilov, H. Eckert, D. Hicks, C. Oses, C. Toher, R. Friedrich, M. Esters, M. J. Mehl, A. C. Zettel, Y. Lederer, E. Zurek, J.-P. Maria, D. W. Brenner, X. Campilongo, S. Filipovic, W. G. Fahrenholtz, C. J. Ryan, C. M. DeSalle, R. J. Crealese, D. E. Wolfe, A. Calzolari and S. Curtarolo, Disordered enthalpy-entropy descriptor for high-entropy ceramics discovery, *Nature*, 2024, **625**, 66–73, DOI: 10.1038/s41586-023-06786-y.
31 C. Draxl and M. Scheffler, NOMAD: The FAIR concept for big data-driven materials, *MRS Bull.*, 2018, **43**, 676–682, DOI: 10.1557/mrs.2018.208.
32 C. Draxl and M. Scheffler, The NOMAD laboratory: from data sharing to artificial intelligence, *JPhys Mater.*, 2019, **2**(3), 036001, DOI: 10.1088/2515-7639/ab13bb.
33 Elementary Multiperspective Material Ontology (EMMO), 2019, https://github.com/emmo-repo/EMMO.
34 J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. A. Hansen, H. H. Kristoffersen, M. Kuisma, A. H. Larsen, L. Lehtovaara, M. Ljungberg, O. Lopez-Acevedo, P. G. Moses, J. Ojanen, T. Olsen, V. Petzold, N. A. Romero, J. Stausholm-Møller, M. Strange, G. A. Tritsaris, M. Vanin, M. Walter, B. Hammer, H. Häkkinen, G. K. H. Madsen, R. M. Nieminen, J. K. Nørskov, M. Puska, T. T. Rantala, J. Schiøtz, K. S. Thygesen and K. W. Jacobsen, Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method, *J. Phys.: Condens. Matter*, 2010, **22**(25), 253202, DOI: 10.1088/0953-8984/22/25/253202.
35 M. Esters, C. Oses, S. Divilov, H. Eckert, R. Friedrich, D. Hicks, M. J. Mehl, F. Rose, A. Smolyanyuk, A. Calzolari, X. Campilongo, C. Toher and S. Curtarolo, aflow.org: A web ecosystem of databases, software and tools, *Comput. Mater. Sci.*, 2023, **216**, 111808, DOI: 10.1016/j.commatsci.2022.111808.
36 M. L. Evans, C. W. Andersen, S. Dwaraknath, M. Scheidgen, Á. Fekete and D. Winston, optimade-python-tools: a Python library for serving and consuming materials data via OPTIMADE APIs, *J. Open Source Softw.*, 2021, **6**(65), 3458, DOI: 10.21105/joss.03458.
37 M. L. Evans and A. J. Morris, matador: a Python library for analysing, curating and performing high-throughput density-functional theory calculations, *J. Open Source Softw.*, 2020, **5**(54), 2563, DOI: 10.21105/joss.02563.
38 *FAIRsharing.org: OPTIMADE; Open Databases Integration for Materials Design*, 2020, DOI: 10.25504/FAIRsharing.xvfqAC.
39 D. C. Ford, D. Hicks, C. Oses, C. Toher and S. Curtarolo, Metallic glasses for biodegradable implants, *Acta Mater.*, 2019, **176**, 297–305, DOI: 10.1016/j.actamat.2019.07.008.
40 L. Fuentes-Cobas, D. Chateigner, M. Fuentes-Montero, G. Pepponi and S. Grazulis, The representation of coupling interactions in the Material Properties Open Database (MPOD), *Adv. Appl. Ceram.*, 2017, **116**(8), 428–433, DOI: 10.1080/17436753.2017.1343782.
41 L. M. Ghiringhelli, C. Baldauf, T. Bereau, S. Brockhauser, C. Carbogno, J. Chamanara, S. Cozzini, S. Curtarolo, C. Draxl, S. Dwaraknath, Á. Fekete, J. Kermode, C. T. Koch, M. Kühbach, A. N. Ladines, P. Lambrix, M.-O. Himmer, S. V. Levchenko, M. Oliveira, A. Michalchuk, R. E. Miller, B. Onat, P. Pavone, G. Pizzi, B. Regler, G.-M. Rignanese, J. Schaarschmidt, M. Scheidgen, A. Schneidewind, T. Sheveleva, C. Su, D. Usvyat, O. Valsson, C. Wöll and M. Scheffler, Shared metadata for data-centric materials science, *Sci. Data*, 2023, **10**(1), 626, DOI: 10.1038/s41597-023-02501-8.
42 P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A. D. Corso, S. de Gironcoli, P. Delugas, R. A. DiStasio, A. Ferretti, A. Floris, G. Fratesi, G. Fugallo, R. Gebauer, U. Gerstmann, F. Giustino, T. Gorni, J. Jia, M. Kawamura, H.-Y. Ko, A. Kokalj, E. Küçükbenli, M. Lazzeri, M. Marsili, N. Marzari, F. Mauri, N. L. Nguyen, H.-V. Nguyen, A. O. de-la Roza, L. Paulatto, S. Poncé, D. Rocca, R. Sabatini, B. Santra, M. Schlipf, A. P. Seitsonen, A. Smogunov, I. Timrov, T. Thonhauser, P. Umari, N. Vast, X. Wu and S. Baroni, Advanced capabilities for materials modelling with Quantum ESPRESSO, *J. Phys.: Condens. Matter*, 2017, **29**(46), 465901, DOI: 10.1088/1361-648X/aa8f79.
43 P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. Chiarotti, M. Cococcioni, I. Dabo, A. Dal 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. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. Seitsonen, A. Smogunov, P. Umari and R. Wentzcovitch, Quantum ESPRESSO: a modular and open-source software project for quantum simulations of materials, *J. Phys.: Condens. Matter*, 2009, **21**(39), 395502, DOI: 10.1088/0953-8984/21/39/395502.
44 M. N. Gjerding, A. Taghizadeh, A. Rasmussen, S. Ali, F. Bertoldo, T. Deilmann, N. R. Knøsgaard, M. Kruse,

A. H. Larsen, S. Manti, T. G. Pedersen, U. Petralanda, T. Skovhus, M. K. Svendsen, J. J. Mortensen, T. Olsen and K. S. Thygesen, Recent progress of the computational 2D materials database (C2DB), *2D Materials*, 2021, **8**(4), 044002, DOI: **10.1088/2053-1583/ac1059**.

45 V. Gomzi, I. M. Šapić and A. Vidak, ReaxFF force field development and application for toluene adsorption on $MnMO_x$ (M=Cu, Fe, Ni) catalysts, *J. Phys. Chem. A*, 2021, **125**(50), 10649–10656, DOI: **10.1021/acs.jpca.1c06939**.

46 R. E. A. Goodall, A. S. Parackal, F. A. Faber, R. Armiento and A. A. Lee, Rapid discovery of stable materials by coordinate-free coarse graining, *Sci. Adv.*, 2022, **8**(30), eabn4117, DOI: **10.1126/sciadv.abn4117**.

47 D. Grassano, D. Campi, A. Marrazzo and N. Marzari, Complementary screening for quantum spin Hall insulators in two-dimensional exfoliable materials, *Phys. Rev. Mater.*, 2023, **7**, 094202, DOI: **10.1103/PhysRevMaterials.7.094202**.

48 D. Grassano, N. Marzari and D. Campi, High-throughput screening of weyl semimetals, *Phys. Rev. Mater.*, 2024, **8**, 024201, DOI: **10.1103/PhysRevMaterials.8.024201**.

49 S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moeck, R. T. Downs and A. L. Bail, Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration, *Nucleic Acids Res.*, 2012, **40**, D420–D427, DOI: **10.1093/nar/gkr900**.

50 S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, J. Gath, K. W. Jacobsen, J. J. Mortensen, T. Olsen and K. S. Thygesen, The Computational 2D Materials Database: high-throughput modeling and discovery of atomically thin crystals, *2D Materials*, 2018, **5**(4), 042002, DOI: **10.1088/2053-1583/aacfc1**.

51 S. R. Hall, F. H. Allen and I. D. Brown, The crystallographic information file (CIF): a new standard archive file for crystallography, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 1991, **47**, 655–685, DOI: **10.1107/S010876739101067X**.

52 A. F. Harper, M. L. Evans, J. P. Darby, B. Karasulu, C. P. Koçer, J. R. Nelson and A. J. Morris, Ab initio Structure Prediction Methods for Battery Materials: A review of recent computational efforts to predict the atomic level structure and bonding in materials for rechargeable batteries, *Johnson Matthey Technol. Rev.*, 2020, **64**(2), 103–118, DOI: **10.1595/205651320x15742491027978**.

53 J. He, Z. Yao, V. I. Hegde, S. S. Naghavi, J. Shen, K. M. Bushick and C. Wolverton, Computational discovery of stable heteroanionic oxychalcogenides *ABXO* (*A, B*=metals; *X*=S, Se, and Te) and their potential applications, *Chem. Mater.*, 2020, **32**(19), 8229–8242, DOI: **10.1021/acs.chemmater.0c01902**.

54 P. Henkel, J. Li, G. K. Grandhi, P. Vivo and P. Rinke, Screening Mixed-Metal $Sn_2M(III)Ch_2X_3$ Chalcohalides for Photovoltaic Applications, *Chem. Mater.*, 2023, **35**(18), 7761–7769, DOI: **10.1021/acs.chemmater.3c01629**.

55 D. Hicks, M. J. Mehl, M. Esters, C. Oses, O. Levy, G. L. W. Hart, C. Toher and S. Curtarolo, The AFLOW Library of Crystallographic Prototypes: Part 3, *Comput. Mater. Sci.*, 2021, **199**, 110450, DOI: **10.1016/j.commatsci.2021.110450**.

56 D. Hicks, M. J. Mehl, E. Gossett, C. Toher, O. Levy, R. M. Hanson, G. L. W. Hart and S. Curtarolo, The AFLOW Library of Crystallographic Prototypes: Part 2, *Comput. Mater. Sci.*, 2019, **161**, S1–S1011, DOI: **10.1016/j.commatsci.2018.10.043**.

57 L. Himanen, A. Geurts, A. S. Foster and P. Rinke, Data-Driven Materials Science: Status, Challenges, and Perspectives, *Adv. Sci.*, 2019, **6**(21), 1900808, DOI: **10.1002/advs.201900808**.

58 A. Hospital, P. Andrio, C. Cugnasco, L. Codo, Y. Becerra, P. D. Dans, F. Battistini, J. Torrers, R. Goñi, M. Orozco and J. L. Gelpi, BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data, *Nucleic Acids Res.*, 2016, **44**, D272–D278, DOI: **10.1093/nar/gkv1301**.

59 A. Hospital, F. Battistini, R. Soliva, J. L. Gelpi and M. Orozco, Surviving the deluge of biosimulation data, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2020, **10**, e1449, DOI: **10.1002/wcms.1449**.

60 S. P. Huber, S. Zoupanos, M. Uhrin, L. Talirz, L. Kahle, R. Häuselmann, D. Gresch, T. Müller, A. V. Yakutovich, C. W. Andersen, F. F. Ramirez, C. S. Adorf, F. Gargiulo, S. Kumbhar, E. Passaro, C. Johnston, A. Merkys, A. Cepellotti, N. Mounet, N. Marzari, B. Kozinsky and G. Pizzi, AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance, *Sci. Data*, 2020, **7**(1), 300, DOI: **10.1038/s41597-020-00638-4**.

61 S. Im, S. L. Shang, N. D. Smith, A. M. Krajewski, T. Lichtenstein, H. Sun, B. J. Bocklund, Z. K. Liu and H. Kim, Thermodynamic properties of the Nd-Bi system via emf measurements, DFT calculations, machine learning, and CALPHAD modeling, *Acta Mater.*, 2022, **223**, 117448, DOI: **10.1016/J.ACTAMAT.2021.117448**.

62 O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha and S. Curtarolo, Materials cartography: Representing and mining materials space using structural and electronic fingerprints, *Chem. Mater.*, 2015, **27**(3), 735–743, DOI: **10.1021/cm503507h**.

63 IUCr, Data requirements for structures, 2023, **https://journals.iucr.org/c/services/cif/reqdata.html**.

64 Y. Iwasaki, A. G. Kusne and I. Takeuchi, Comparison of dissimilarity measures for cluster analysis of X-ray diffraction data from combinatorial libraries, *npj Comput. Mater.*, 2017, **3**(1), 1–9, DOI: **10.1038/s41524-017-0006-2**.

65 K. M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary, D. Circi, S. Cox, W. A. de Jong, M. L. Evans, N. Gastellu, J. Genzling, M. V. Gil, A. K. Gupta, Z. Hong, A. Imran, S. Kruschwitz, A. Labarre,

J. Lála, T. Liu, S. Ma, S. Majumdar, G. W. Merz, N. Moitessier, E. Moubarak, B. Mouriño, B. Pelkie, M. Pieler, M. C. Ramos, B. Ranković, S. G. Rodriques, J. N. Sanders, P. Schwaller, M. Schwarting, J. Shi, B. Smit, B. E. Smith, J. V. Herck, C. Völker, L. Ward, S. Warren, B. Weiser, S. Zhang, X. Zhang, G. A. Zia, A. Scourtas, K. J. Schmidt, I. Foster, A. D. White and B. Blaiszik, 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon, *Digital Discovery*, 2023, **2**(5), 1233–1250, DOI: **10.1039/d3dd00113j**.

66 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, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.*, 2013, **1**(1), 011002, DOI: **10.1063/1.4812323**.

67 S. Kirklin, J. E. Saal, V. I. Hegde and C. Wolverton, High-throughput computational search for strengthening precipitates in alloys, *Acta Mater.*, 2016, **102**, 125–135, DOI: **10.1016/j.actamat.2015.09.016**.

68 A. M. Krajewski, J. W. Siegel, and Z.-K. Liu. Efficient structure-informed featurization and property prediction of ordered, dilute, and random atomic structures, *arXiv*, 2024, preprint, arXiv:2404.02849, DOI: **10.48550/arXiv.2404.02849**.

69 A. M. Krajewski, J. W. Siegel, J. Xu and Z. K. Liu, Extensible Structure-Informed Prediction of Formation Energy with improved accuracy and usability employing neural networks, *Comput. Mater. Sci.*, 2022, **208**, 111254, DOI: **10.1016/J.COMMATSCI.2022.111254**.

70 G. Kresse and J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**(16), 11169–11186, DOI: **10.1103/physrevb.54.11169**.

71 A. G. Kusne, H. Yu, C. Wu, H. Zhang, J. Hattrick-Simpers, B. DeCost, S. Sarker, C. Oses, C. Toher, S. Curtarolo, A. V. Davydov, R. Agarwal, L. A. Bendersky, M. Li, A. Mehta and I. Takeuchi, On-the-fly closed-loop materials discovery via bayesian active learning, *Nat. Commun.*, 2020, **11**, 5966, DOI: **10.1038/s41467-020-19597-w**.

72 A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, The atomic simulation environment—a Python library for working with atoms, *J. Phys.: Condens. Matter*, 2017, **29**(27), 273002, DOI: **10.1088/1361-648x/aa680e**.

73 J. N. Law, S. Pandey, P. Gorai and P. C. St. John, Upper-Bound Energy Minimization to Search for Stable Functional Materials with Graph Neural Networks, *JACS Au*, 2023, **3**(1), 113–123, DOI: **10.1021/jacsau.2c00540**.

74 H. Levämäki, F. Bock, D. G. Sangiovanni, L. J. S. Johnson, F. Tasnádi, R. Armiento and I. A. Abrikosov, HADB: A materials-property database for hard-coating alloys, *Thin Solid Films*, 2023, **766**, 139627, DOI: **10.1016/j.tsf.2022.139627**.

75 H. Li, O. Hartig, R. Armiento and P. Lambrix, Ontology-based GraphQL server generation for data access and data integration, *Semant. Web*, 2024, 1–37, DOI: **10.3233/sw-233550**.

76 X. Liu, P.-P. De Breuck, L. Wang and G.-M. Rignanese, A simple denoising approach to exploit multi-fidelity data for machine learning materials properties, *npj Comput. Mater.*, 2022, **8**, 233, DOI: **10.1038/s41524-022-00925-1**.

77 Z. Liu, Perspective on materials genome®, *Chin. Sci. Bull.*, 2014, **59**, 1619, DOI: **10.1007/s11434-013-0072-x**.

78 P. Lyngby and K. S. Thygesen, Data-driven discovery of 2D materials by deep generative models, *npj Comput. Mater.*, 2022, **8**(1), 232, DOI: **10.1038/s41524-022-00923-3**.

79 A. Marrazzo, M. Gibertini, D. Campi, N. Mounet and N. Marzari, Prediction of a large-gap and switchable Kane-Mele quantum spin Hall insulator, *Phys. Rev. Lett.*, 2018, **120**, 117701, DOI: **10.1103/PhysRevLett.120.117701**.

80 A. Marrazzo, M. Gibertini, D. Campi, N. Mounet and N. Marzari, Relative abundance of Z_2 topological order in exfoliable two-dimensional insulators, *Nano Lett.*, 2019, **19**(12), 8431–8440, DOI: **10.1021/acs.nanolett.9b02689**.

81 Matcloud, **http://matcloud.com.cn**, accessed 2024.

82 A. Medina-Smith, C. A. Becker, R. L. Plante, L. M. Bartolo, A. Dima, J. A. Warren and R. J. Hanisch, A Controlled Vocabulary and Metadata Schema for Materials Science Data Discovery, *Data Sci. J.*, 2021, **20**(1), 18, DOI: **10.5334/dsj-2021-018**.

83 M. J. Mehl, D. Hicks, C. Toher, O. Levy, R. M. Hanson, G. L. W. Hart and S. Curtarolo, The AFLOW Library of Crystallographic Prototypes: Part 1, *Comput. Mater. Sci.*, 2017, **136**, S1–S828, DOI: **10.1016/j.commatsci.2017.01.017**.

84 J. Mendenhall, B. P. Brown, S. Kothiwale and J. M. BCL, Conf: Improved Open-Source Knowledge-Based Conformation Sampling Using the Crystallography Open Database, *J. Chem. Inf. Model.*, 2020, **61**(1), 189–201, DOI: **10.1021/acs.jcim.0c01140**.

85 A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon and E. D. Cubuk, Scaling deep learning for materials discovery, *Nature*, 2023, **624**(7990), 80–85, DOI: **10.1038/s41586-023-06735-9**.

86 J. J. Mortensen, A. H. Larsen, M. Kuisma, A. V. Ivanov, A. Taghizadeh, A. Peterson, A. Haldar, A. O. Dohn, C. Schäfer, E. Ö. Jónsson, E. D. Hermes, F. A. Nilsson, G. Kastlunger, G. Levi, H. Jónsson, H. Häkkinen, J. Fojt, J. Kangsabanik, J. Sødequist, J. Lehtomäki, J. Heske, J. Enkovaara, K. T. Winther, M. Dulak, M. M. Melander, M. Ovesen, M. Louhivuori, M. Walter, M. Gjerding, O. Lopez-Acevedo, P. Erhart, R. Warmbier, R. Würdemann, S. Kaappa, S. Latini, T. M. Boland, T. Bligaard, T. Skovhus, T. Susi, T. Maxson, T. Rossi, X. Chen, Y. L. A. Schmerwitz, J. Schiøtz, T. Olsen, K. W. Jacobsen and K. S. Thygesen, GPAW: An open

Python package for electronic structure calculations, *J. Chem. Phys.*, 2024, **160**(9), 092503, DOI: **10.1063/5.0182685**.

87 N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I. E. Castelli, A. Cepellotti, G. Pizzi and N. Marzari, Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds, *Nat. Nanotechnol.*, 2018, **13**(3), 246–252, DOI: **10.1038/s41565-017-0035-5**.

88 H. Moustafa, P. M. Larsen, M. N. Gjerding, J. J. Mortensen, K. S. Thygesen and K. W. Jacobsen, Computational exfoliation of atomically thin one-dimensional materials with application to Majorana bound states, *Phys. Rev. Mater.*, 2022, **6**(6), 064202, DOI: **10.1103/PhysRevMaterials.6.064202**.

89 C. Nyshadham, C. Oses, J. E. Hansen, I. Takeuchi, S. Curtarolo and G. L. W. Hart, A computational high-throughput search for new ternary superalloys, *Acta Mater.*, 2017, **122**, 438–447, DOI: **10.1016/j.actamat.2016.09.017**.

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

91 Ontology driven Open Translation Environment (OntoTrans), **https://ontotrans.eu**, 2020, accessed 2023.

92 C. Oses, M. Esters, D. Hicks, S. Divilov, H. Eckert, R. Friedrich, M. J. Mehl, A. Smolyanyuk, X. Campilongo, A. van de Walle, J. Schroers, A. G. Kusne, I. Takeuchi, E. Zurek, M. Buongiorno Nardelli, M. Fornari, Y. Lederer, O. Levy, C. Toher and S. Curtarolo, aflow++: A C++ framework for autonomous materials design, *Comput. Mater. Sci.*, 2023, **217**, 111889, DOI: **10.1016/j.commatsci.2022.111889**.

93 Y. Ozaki, Y. Suzuki, T. Hawai, K. Saito, M. Onishi and K. Ono, Automated crystal structure analysis based on blackbox optimisation, *npj Comput. Mater.*, 2020, **6**(1), 1–7, DOI: **10.1038/s41524-020-0330-9**.

94 S. Pakdel, A. Rasmussen, A. Taghizadeh, M. Kruse, T. Olsen and K. S. Thygesen, High-throughput computational stacking reveals emergent properties in natural van der Waals bilayers, *Nat. Commun.*, 2024, **15**(1), 932, DOI: **10.1038/s41467-024-45003-w**.

95 Pauling File, 2024, **https://paulingfile.com**.

96 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

97 G. Pepponi, S. Gražulis and D. Chateigner, MPOD: A Material Property Open Database linked to structural information, *Nucl. Instrum. Methods Phys. Res., Sect. B*, 2012, **284**, 10–14, DOI: **10.1016/j.nimb.2011.08.070**.

98 E. Perim, D. Lee, Y. Liu, C. Toher, P. Gong, Y. Li, W. N. Simmons, O. Levy, J. J. Vlassak, J. Schroers and S. Curtarolo, Spectral descriptors for bulk metallic glasses based on the thermodynamics of competing crystalline phases, *Nat. Commun.*, 2016, **7**, 12315, DOI: **10.1038/ncomms12315**.

99 F. Pezoa, J. L. Reutter, F. Suarez, M. Ugarte, and D. Vrgoč, Foundations of JSON schema, in *Proceedings of the 25th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2016, pp. 263–273, DOI: **10.1145/2872427.2883029**.

100 G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari and B. Kozinsky, AiiDA: automated interactive infrastructure and database for computational science, *Comput. Mater. Sci.*, 2016, **111**, 218–230, DOI: **10.1016/j.commatsci.2015.09.013**.

101 G. Pizzi, S. Milana, A. C. Ferrari, N. Marzari and M. Gibertini, Shear and breathing modes of layered materials, *ACS Nano*, 2021, **15**(8), 12509–12534, DOI: **10.1021/acsnano.0c10672**.

102 R. L. Plante, C. A. Becker, A. Medina-Smith, K. Brady, A. Dima, B. Long, L. M. Bartolo, J. A. Warren and R. J. Hanisch, Implementing a Registry Federation for Materials Science Data Discovery, *Data Sci. J.*, 2021, **20**(1), 15, DOI: **10.5334/dsj-2021-015**.

103 J. Qiao, G. Pizzi and N. Marzari, Projectability disentanglement for accurate and automated electronic-structure Hamiltonians, *npj Comput. Mater.*, 2023, **9**(1), 208, DOI: **10.1038/s41524-023-01146-w**.

104 F. L. Reyes Tirado, J. Perrin Toinin and D. C. Dunand, $\gamma + \gamma'$ microstructures in the Co-Ta-V and Co-Nb-V ternary systems, *Acta Mater.*, 2018, **151**, 137–148, DOI: **10.1016/j.actamat.2018.03.057**.

105 F. Rose, C. Toher, E. Gossett, C. Oses, M. Buongiorno Nardelli, M. Fornari and S. Curtarolo, AFLUX: The LUX materials search API for the AFLOW data repositories, *Comput. Mater. Sci.*, 2017, **137**, 362–370, DOI: **10.1016/j.commatsci.2017.04.036**.

106 J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD), *JOM*, 2013, **65**(11), 1501–1509, DOI: **10.1007/s11837-013-0755-4**.

107 S. Sanvito, C. Oses, J. Xue, A. Tiwari, M. Žic, T. Archer, P. Tozman, M. Venkatesan, J. M. D. Coey and S. Curtarolo, Accelerated discovery of new magnets in the Heusler alloy family, *Sci. Adv.*, 2017, **3**(4), e1602241, DOI: **10.1126/sciadv.1602241**.

108 P. Sarker, T. Harrington, C. Toher, C. Oses, M. Samiee, J.-P. Maria, D. W. Brenner, K. S. Vecchio and S. Curtarolo, High-entropy high-hardness metal carbides discovered by entropy descriptors, *Nat. Commun.*, 2018, **9**(1), 4980, DOI: **10.1038/s41467-018-07160-7**.

109 L. Sbailò, Á. Fekete, L. M. Ghiringhelli and M. Scheffler, The NOMAD artificial-intelligence toolkit: turning materials-science data into knowledge and understanding, *npj Comput. Mater.*, 2022, **8**(1), 250, DOI: **10.1038/s41524-022-00935-z**.

110 M. Scheffler, M. Aeschlimann, M. Albrecht, T. Bereau, H.-J. Bungartz, C. Felser, M. Greiner, A. Groß, C. T. Koch, K. Kremer, W. E. Nagel, M. Scheidgen, C. Wöll and C. Draxl, FAIR data enabling new horizons for materials research, *Nature*, 2022, **604**(7907), 635–642, DOI: **10.1038/s41586-022-04501-x**.

111 M. Scheidgen, L. Himanen, A. N. Ladines, D. Sikter, M. Nakhaee, Á. Fekete, T. Chang, A. Golparvar, J. A. Márquez, S. Brockhauser, S. Brückner, L. M. Ghiringhelli, F. Dietrich, D. Lehmberg, T. Denell, A. Albino, H. Näsström, S. Shabih, F. Dobener, M. Kühbach, R. Mozumder, J. F. Rudzinski, N. Daelman, J. M. Pizarro, M. Kuban, C. Salazar, P. Ondračka, H.-J. Bungartz and C. Draxl, NOMAD: A distributed web-based platform for managing materials science research data, *J. Open Source Softw.*, 2023, **8**(90), 5388, DOI: **10.21105/joss.05388**.

112 J. Schmidt, L. Chen, S. Botti and M. A. L. Marques, Predicting the stability of ternary intermetallics with density functional theory and machine learning, *J. Chem. Phys.*, 2018, **148**(24), 241728, DOI: **10.1063/1.5020223**.

113 J. Schmidt, N. Hoffmann, H.-C. Wang, P. Borlido, P. J. M. A. Carriço, T. F. T. Cerqueira, S. Botti and M. A. L. Marques, Machine-learning-assisted determination of the global zero-temperature phase diagram of materials, *Adv. Mater.*, 2023, **35**(22), 2210788, DOI: **10.1002/adma.202210788**.

114 J. Schmidt, L. Pettersson, C. Verdozzi, S. Botti and M. A. L. Marques, Crystal graph attention networks for the prediction of stable materials, *Sci. Adv.*, 2021, **7**(49), eabi7948, DOI: **10.1126/sciadv.abi7948**.

115 J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti and M. A. L. Marques, Predicting the thermodynamic stability of solids combining density functional theory and machine learning, *Chem. Mater.*, 2017, **29**(12), 5090–5103, DOI: **10.1021/acs.chemmater.7b00156**.

116 J. Schmidt, H.-C. Wang, T. F. T. Cerqueira, S. Botti and M. A. L. Marques, A new dataset of 175k stable and metastable materials calculated with the PBEsol and SCAN functionals, *Sci. Data*, 2021, **12**(1), 64, DOI: **10.1038/s41597-022-01177-w**.

117 J. Schmidt, H.-C. Wang, G. Schmidt and M. A. Marques, Machine learning guided high-throughput search of non-oxide garnets, *npj Comput. Mater.*, 2023, **9**(1), 63, DOI: **10.1038/s41524-023-01009-4**.

118 H. Schwarz, T. Uhlig, T. Lindner, T. Lampke, G. Wagner and T. Seyller, Hardness Enhancement in $CoCrFeNi_{1-x}(WC)_x$ High-Entropy Alloy Thin Films Synthesised by Magnetron Co-Sputtering, *Coatings*, 2022, **12**(2), 269, DOI: **10.3390/coatings12020269**.

119 S.-L. Shang, H. Sun, B. Pan, Y. Wang, A. M. Krajewski, M. Banu, J. Li and Z.-K. Liu, Forming Mechanism of Equilibrium and Non-equilibrium Metallurgical Phases in Dissimilar Materials: Illustrated With Aluminum/steel (Al-Fe) Joints, *Sci. Rep.*, 2021, **11**, 24251, DOI: **10.1038/s41598-021-03578-0**.

120 J. Shen, S. D. Griesemer, A. Gopakumar, B. Baldassarri, J. E. Saal, M. Aykol, V. I. Hegde and C. Wolverton, Reflections on one million compounds in the open quantum materials database (OQMD), *JPhys Mater.*, 2022, **5**(3), 031001, DOI: **10.1088/2515-7639/ac7ba9**.

121 SMARTS – A Language for Describing Molecular Patterns, **https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html**, accessed 2024.

122 T. Sohier, D. Campi, N. Marzari and M. Gibertini, Mobility of two-dimensional materials from first principles in an accurate and automated framework, *Phys. Rev. Mater.*, 2018, **2**, 114010, DOI: **10.1103/PhysRevMaterials.2.114010**.

123 T. Sohier, M. Gibertini, D. Campi, G. Pizzi and N. Marzari, Valley-engineering mobilities in two-dimensional materials, *Nano Lett.*, 2019, **19**(6), 3723–3729, DOI: **10.1021/acs.nanolett.9b00865**.

124 V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo and I. Takeuchi, Machine learning modeling of superconducting critical temperature, *npj Comput. Mater.*, 2018, **4**, 29, DOI: **10.1038/s41524-018-0085-8**.

125 C. Suh, C. Fare, J. Warren and E. Pyzer-Knapp, Evolving the materials genome: How machine learning is fueling the next generation of materials discovery, *Annu. Rev. Mater. Res.*, 2020, **50**, 1–25, DOI: **10.1146/annurev-matsci-082019-105100**.

126 SVELTE, 2024, **https://svelte.dev**.

127 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, Materials Cloud, a platform for open computational science, *Sci. Data*, 2020, **7**(1), 299, DOI: **10.1038/s41597-020-00637-5**.

128 The MarketPlace Project, **https://materials-marketplace.eu**, 2018, accessed 2023.

129 The OPTIMADE Developers, Open Database Integration for Materials Design (OPTIMADE), **https://github.com/Materials-Consortia/OPTIMADE**, accessed 2023.

130 The OPTIMADE Developers, OPTIMADE Providers Dashboard, **https://www.optimade.org/providers-dashboard/**, accessed 2023.

131 The OPTIMADE Developers, OPTIMADE Providers List, **https://providers.optimade.org**, accessed 2023.

132 B. H. Toby and R. B. Von Dreele, GSAS-II: the genesis of a modern open-source all purpose crystallography software package, *J. Appl. Crystallogr.*, 2013, **46**(2), 544–549, DOI: **10.1107/S0021889813003531**.

133 M. Uhrin, S. P. Huber, J. Yu, N. Marzari and G. Pizzi, Workflows in AiiDA: Engineering a high-throughput, event-based engine for robust and modular computational workflows, *Comput. Mater. Sci.*, 2021, **187**, 110086, DOI: **10.1016/j.commatsci.2020.110086**.

134 M. T. Vahdat, K. V. Agrawal and G. Pizzi, Machine-learning accelerated identification of exfoliable two-dimensional materials, *Mach. Learn.: Sci. Technol.*, 2022, **3**(4), 045014, DOI: **10.1088/2632-2153/ac9bca**.

135 A. van Roekeghem, J. Carrete, C. Oses, S. Curtarolo and N. Mingo, High-throughput computation of thermal conductivity of high-temperature solid phases: The case of oxide and fluoride perovskites, *Phys. Rev. X*, 2016, **6**, 041061, DOI: **10.1103/PhysRevX.6.041061**.

136 B. Wang, Q. Fan and Y. Yue, Study of crystal properties based on attention mechanism and crystal graph convolutional neural network, *J. Phys.: Condens. Matter*, 2022, **34**, 195901, DOI: **10.1088/1361-648X/ac5705**.

137 H.-C. Wang, S. Botti and M. A. L. Marques, Predicting stable crystalline compounds using chemical similarity, *npj Comput. Mater.*, 2021, **7**(1), 1–9, DOI: **10.1038/s41524-020-00481-6**.

138 T. Wang, K. Zhang, J. Thé and H. Yu, Accurate prediction of band gap of materials using stacking machine learning model, *Comput. Mater. Sci.*, 2022, **201**, 110899, DOI: **10.1016/j.commatsci.2021.110899**.

139 Z. Wang, Y. Gong, M. L. Evans, Y. Yan, S. Wang, N. Miao, R. Zheng, G.-M. Rignanese and J. Wang, Machine Learning-Accelerated Discovery of A_2BC_2 Ternary Electrides with Diverse Anionic Electron Densities, *J. Am. Chem. Soc.*, 2023, **145**(48), 26412–26424, DOI: **10.1021/jacs.3c10538**.

140 Z. Wang, Y. Gong, M. L. Evans, Y. Yan, S. Wang, N. Miao, R. Zheng, G.-M. Rignanese and J. Wang, Machine learning-accelerated discovery of A_2BC_2 ternary electrides with diverse anionic electron densities, *Materials Cloud Archive*, 2023, **181**, 26412–26424, DOI: **10.24435/materialscloud:c8-gy**.

141 D. Weininger, SMILES, a chemical languge and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.*, 1988, **28**(1), 31–36, DOI: **10.1021/ci00057a005**.

142 D. Wines, R. Gurunathan, K. F. Garrity, B. DeCost, A. J. Biacchi, F. Tavazza and K. Choudhary, Recent progress in the JARVIS infrastructure for next-generation data-driven materials design, *Applied Physics Reviews*, 2023, **10**(4), 041302, DOI: **10.1063/5.0159299**.

143 T. Xie and J. C. Grossman, Crystal Graph Convolutional Neural Networks for Accurate and Interpretable Prediction of Material Properties, *Phys. Rev. Lett.*, 2018, **120**(14), 145301, DOI: **10.1103/PhysRevLett.120.145301**.

144 A. V. Yakutovich, K. Eimre, O. Schütt, L. Talirz, C. S. Adorf, C. W. Andersen, E. Ditler, D. Du, D. Passerone, B. Smit, N. Marzari, G. Pizzi and C. A. Pignedoli, AiiDAlab – an ecosystem for developing, executing, and sharing scientific workflows, *Comput. Mater. Sci.*, 2021, **188**, 110165, DOI: **10.1016/j.commatsci.2020.110165**.

145 X. Yang, Z. Wang, X. Zhao, J. Song, C. Yu, J. Zhou and K. Li, A high-throughput computational materials infrastructure: Present, future visions and challenges, *Chin. Phys. B*, 2018, **27**(11), 110301, DOI: **10.1088/1674-1056/27/11/110301**.

146 X. Yang, Z. Wang, X. Zhao, J. Song, M. Zhang and H. Liu, Matcloud: A high-throughput computational infrastructure for integrated management of materials simulation, data and resources, *Comput. Mater. Sci.*, 2018, **146**, 319–333, DOI: **10.1016/j.commatsci.2018.01.039**.

147 W. Ye, X. Lei, M. Aykol and J. H. Montoya, Novel inorganic crystal structures predicted using autonomous simulation agents, *Sci. Data*, 2022, **9**(1), 302, DOI: **10.1038/s41597-022-01438-8**.

148 S. Zivanovic, G. Bayarri, F. Colizzi, D. Moreno, J. L. Gelpi, R. Soliva, A. Hospital and M. Orozco, Bioactive conformational ensemble server and database. a public framework to speed up in silico drug discovery, *J. Chem. Theory Comput.*, 2020, **16**, 6586–6597, DOI: **10.1021/acs.jctc.0c00305**.