Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Digital Discovery*, 2026, **5**, 1114

MC3D: the Materials Cloud computational database of experimentally known stoichiometric inorganics

Sebastiaan P. Huber, †[a] Michail Minotakis, †[b] Marnik Bercx,†[ab] Timo Reents,[b] Kristjan Eimre,[ab] Nataliya Paulish,[b] Nicolas Hörmann,[c] Martin Uhrin,[d] Nicola Marzari[abe] and Giovanni Pizzi *[ab]

Received 15th September 2025
Accepted 1st February 2026

DOI: 10.1039/d5dd00415b

rsc.li/digitaldiscovery

Density-functional theory (DFT) is a widely used method to compute properties of materials, which are often collected in databases and serve as valuable starting points for further studies. In this article, we present the Materials Cloud Three-Dimensional Structure Database (MC3D), an online database of computed three-dimensional (3D) inorganic crystal structures. Close to a million experimentally reported structures were imported from the COD, ICSD and MPDS databases; these were parsed and filtered to yield a collection of 72 589 unique and stoichiometric structures, of which 95% are, to date, classified as experimentally known. The geometries of structures with up to 64 atoms were then optimized using DFT with automated workflows and curated input protocols. The procedure was repeated for different functionals and computational protocols, generating three methodology-based MC3D subdatabases: PBE-v1, PBEsol-v1, and PBEsol-v2, with the latest containing 32 013 unique structures. All subdatabases of the MC3D are made available on the Materials Cloud portal, which provides a graphical interface to explore and download the data. The database includes the full provenance graph of all the calculations driven by the automated workflows, thus establishing full reproducibility of the results and more-than-FAIR procedures.

Introduction

The paradigm of computational materials discovery based on quantum-mechanical approaches has found significant adoption in recent years.[1,2] Computational studies have grown in number and scale, in particular those based on DFT,[3,4] a powerful first-principles method to compute the electronic ground state of materials, and routinely used to predict their properties.[5,6] The predictive power of DFT, together with the growing availability of powerful computational resources, has driven high-throughput computational materials discovery,[7] aiming to screen or discover materials with optimal properties. This has in turn spurred the development of several workflow systems to manage the large amount of calculations and the data they produce,[8–20] leading to the creation of multiple publicly available databases of computed materials properties.[21–32] Recently, these databases have been a catalyst for the adoption of machine-learning methods in computational materials science, as the basis to train models,[33–37] to predict the properties of known materials[38–43] or even the existence of new ones.[44–49] To ease the ingestion of data from these databases, a community effort has resulted in the OPTIMADE universal API[50,51] that enable users of the participating databases to use a common query language and obtain results in a common format.

Nevertheless, computational databases might not always use a consistent setup to compute the properties of all materials. This, in turn, leads to potential small inconsistencies in the data across the periodic table, making more challenging to obtain accurate machine-learning models to predict properties of materials not present in the database, as already noted in ref. 52 and recently demonstrated by the accuracy of the PET-MAD model.[35,53]

In this article, we first present a set of automated workflows that import crystal structures from the Crystallographic Open Database (COD),[54] the Inorganic Crystal Structure Database (ICSD),[55] and the Materials Platform for Data Science (MPDS).[56] Using these workflows, we import all entries (close to a million structures, mainly reported from experiments) and reduce them to a set of 72 589 unique bulk stoichiometric inorganic crystal structures, after several filtering steps that we describe in detail below. For the subset of structures not including lanthanides or

[a] *Theory and Simulation of Materials (THEOS), National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland*

[b] *PSI Center for Scientific Computing, Theory, and Data, 5232 Villigen PSI, Switzerland. E-mail: giovanni.pizzi@psi.ch*

[c] *Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany*

[d] *Université Grenoble Alpes, MIAI Cluster IA, SIMaP, 38000 Grenoble, France*

[e] *Bremen Center for Computational Materials Science, MAPEX Center for Materials and Processes, University of Bremen, 28359 Bremen, Germany*

† These authors contributed equally to this work.

actinides, and including at least all systems with up to 64 atoms, we compute the electronic ground state of their optimized geometries *via* DFT using the open-source Quantum ESPRESSO (QE)[57,58] code, powered by the SIRIUS library.[59] The computational inputs for the DFT simulations executed by the workflows are determined by fully automated protocols, requiring only the input structure as mandatory input, therefore enabling automated high-throughput execution of the workflows. The protocols are tuned to select input parameters that optimize the balance between precision and computational cost. Since the workflows are implemented in AiiDA,[8,9,60] the entire provenance of each optimized geometry, including all raw simulation inputs and outputs, is automatically preserved in the database.

The calculations use both PBE[61] and PBEsol,[62] and different protocols for the input parameters. We refer to the collection of resulting databases of optimized geometries as the Materials Cloud Three-Dimensional Structure Database (MC3D). In particular, when considering the PBEsol functional and the most recent protocol `PBEsol-v2`, we provide 32 013 unique structures computed and relaxed with DFT.

We make the MC3D database fully and publicly available through the Materials Cloud Archive.[63] Moreover, for easier accessibility, we provide a dedicated web application in the Materials Cloud[64] platform (https://www.materialscloud.org/mc3d), through which all data can be inspected interactively, as well as an OPTIMADE-compliant endpoint (implemented using `optimade-python-tools`[65]) to access the data *via* a standardized API. In the following, we describe the data import and processing pipeline, the computational workflows, the content of the MC3D, as well as its web interface.

Results

Our pipeline started from importing 901 210 crystal structures from three of the largest experimental inorganic crystal structure databases: the COD[54] (revision 213 553), the ICSD[55] (version 2017.2), and the MPDS[56] (version 1.0.2).

Fig. 1(a) shows a graphical representation of the pipeline that reduced this initial collection to 72 589 unique stoichiometric crystal structures. In the first step of the pipeline, the input structures—obtained in the Crystallographic Information File (CIF) format[66]—were parsed and validated. A number of CIF files had to be discarded due to invalid syntax or inconsistent information. From the 723 165 valid CIFs, the crystal structure was successfully parsed and the crystal lattice was normalized reducing it to the primitive cell (see the Methods for details on the validation and parsing).

The next step in the pipeline filtered out 244 962 structures that are not stoichiometric, *i.e.*, those that contain partial occupancies. These structures are beyond the scope of the current study, due to the requirement of considering appropriately chosen large supercells to accommodate partial occupancies. From the remaining 478 203 crystal structures, 197 105 were found to be duplicates (details on how uniqueness is determined are given in the Methods section). The last filtering step in the pipeline removed 208 509 structures that include hydrogen and are available only in the COD database. As detailed in the Methods section, we apply this filter to effectively exclude molecular crystals, included in COD but not in ICSD nor in MPDS, since our study focuses on inorganic compounds.

The final collection of starting crystal structures, labeled MC3D-source, consists of 72 589 unique three-dimensional crystal structures. Of these, 3305 structures are explicitly reported to have a theoretical origin by the source database. Therefore, we identify no more than 69 284 experimentally known stoichiometric inorganics (see Methods for more details), a much smaller number than typically quoted. This first result already highlights the relatively small size of the space of experimentally known inorganic stoichiometric crystal structures. This might point to the actual size of this space, as well as the complexity of synthesizing and characterizing complex systems,

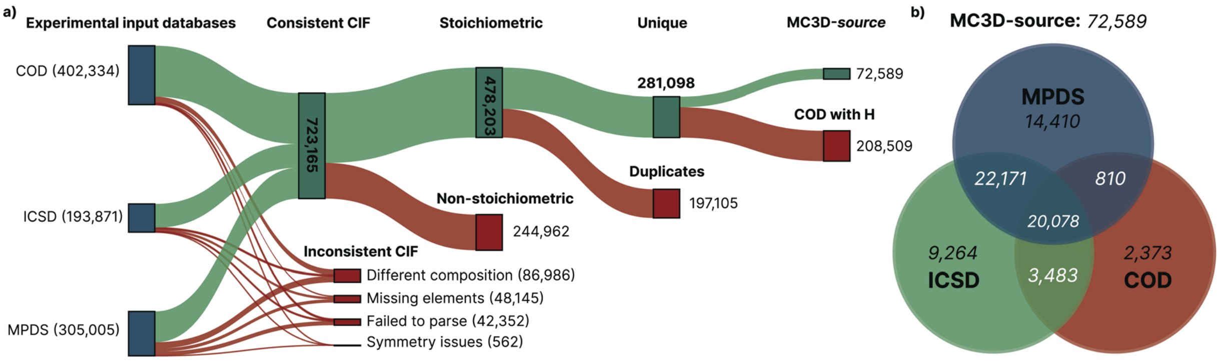


Fig. 1 (a) Sankey diagram visualizing the pipeline that filtered the 901 210 CIF files, as imported from the COD, ICSD, and MPDS databases, down to the MC3D-source collection of 72 589 unique stoichiometric inorganic crystal structures. Red branches indicate structures that were discarded, while green branches correspond to structures that made their way into the following filtering step. The first step discarded CIF files that contained invalid syntax or inconsistent information, or could otherwise not be parsed to yield a valid crystal structure definition. Non-stoichiometric structures are discarded in the second step. Out of the remaining structures, 197 105 were determined to be duplicates in the third step. In the last step, 208 509 structures only available in COD and including hydrogen atoms were not considered for further analysis because they typically correspond to molecular crystals. (b) Venn diagram of the distribution of unique structures in the MC3D-source collection from the three source databases.

especially those including several chemical elements (ternaries, quaternaries and beyond), as already highlighted in ref. 67.

Fig. 1(b) shows the Venn diagram of the distribution of structures in the MC3D-source from the three source databases. Out of the total 72 589 structures, 20 078 structures are present in all three databases, and 46 542 structures are present in at least two of the three databases. The COD, ICSD and MPDS each contain 2 373, 9264 and 14 410 structures that are unique to them, respectively (or 210 882 for COD, if we consider also structures with hydrogen, excluded from MC3D-source in our last filtering step as discussed above).

A subset of the 72 589 unique crystal structures of the MC3D-source was then passed through our automated workflow to optimize their geometry using DFT. First, structures containing lanthanides or actinides were not considered. This is because, on one hand, a physically accurate description of their electronic structure (in particular of the f electrons) would require a more sophisticated treatment than standard DFT. On the other hand, the pseudopotentials currently available for these elements, when available, have limited precision (see, *e.g.*, Fig. S9.7 in the Supplementary Information of ref. 68) and also often lead to numerical instabilities and a significant failure rate. The subset of the MC3D-source database that excludes lanthanides and actinides contains 46 964 structures. Furthermore, in order to better exploit the available computational budget, we prioritized processing structures with smaller number of atoms in the unit cell. While also some larger structures have already been computed, in this paper we focus the analysis on the structures with up to 64 atoms, all of which were submitted to the ground-state atomic and geometric relaxation workflow.

We considered two different DFT functionals (PBE and PBEsol) and, in the PBEsol case, we run the workflow with two different versions of the input parameter protocols, the second one (v2) being refined and whose numerical precision was verified *via* extensive tests.[69] The resulting optimized geometries form three subdatabases of MC3D, identified by the methodology labels `PBE-v1`, `PBEsol-v1`, and `PBEsol-v2`. An overview of the differences in physical methods, input parameters, and pseudopotential libraries for these methodology labels is provided in the Methods section. The rest of this article focuses on the results of the most accurate and precise subdatabase, `PBEsol-v2`, but the results of all three subdatabases are publicly available.

Out of the 38 739 structures that have been processed, the relaxation workflow successfully completed 33 142 structures, corresponding to a success rate of 85.5%. In the remaining cases, the workflow encountered errors that could not be solved by the automatic error handling mechanisms implemented (see the section on Automated error handling in the Methods for more details). These results are represented visually in Fig. 2, where the workflow results are plotted as a histogram as a function of (a) the number of elemental species and (b) the number of atomic sites in the structure.

Our complete relaxation workflow to compute the optimized geometry of a crystal structure and its ground-state electronic charge density comprises multiple steps executing the `pw.x` code of the SIRIUS-accelerated[59] QE package. However, the `pw.x` code can encounter a variety of errors, ranging from issues with the compute hardware on which the code is run itself (*e.g.*, insufficient job resources or node failures) to errors during code execution (*e.g.*, numerical instabilities or convergence issues). These errors need to be handled as much as possible automatically by the workflow in order to be robust and scalable. Fig. 3 shows a visual representation of how often a workflow needed to restart the calculation to achieve successful completion, and of the most commonly observed errors. We first

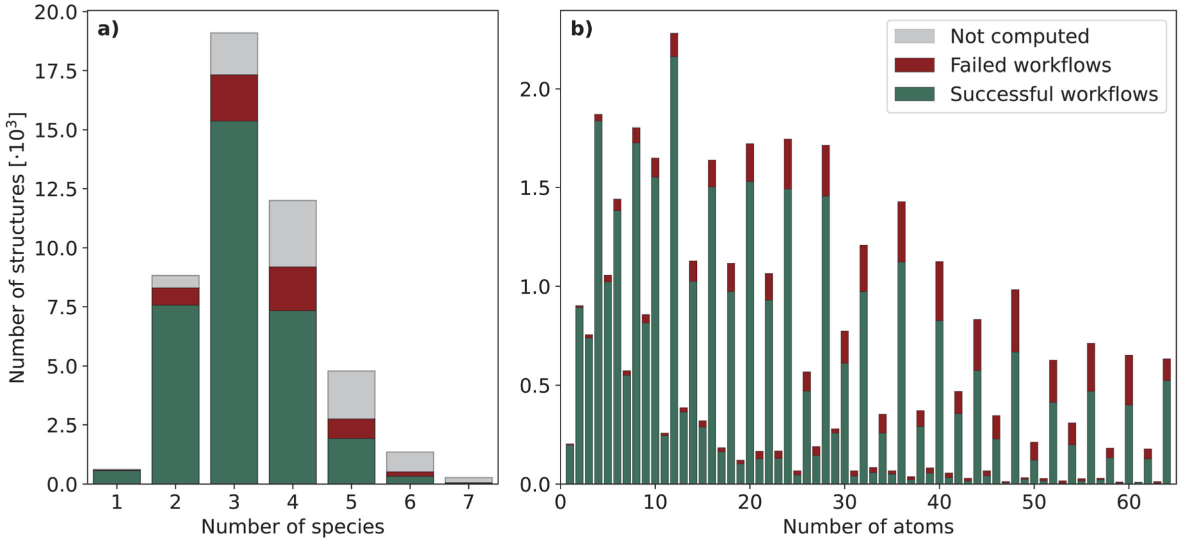


Fig. 2 Histograms of the results of the geometry-optimization workflows executed on the subset of MC3D-source excluding lanthanides and actinides (46 964 structures) as a function of (a) the number of species (*i.e.*, distinct chemical elements) in the structure, and (b) the number of atoms in the primitive cell. Green represents successful workflows, whereas red indicates workflows that failed with an unrecoverable error. The gray bars in (a) correspond to structures with more than 64 atoms that have not been computed. In panel (a), the 34 structures in MC3D-source with 8 or more species are not shown in the plot.

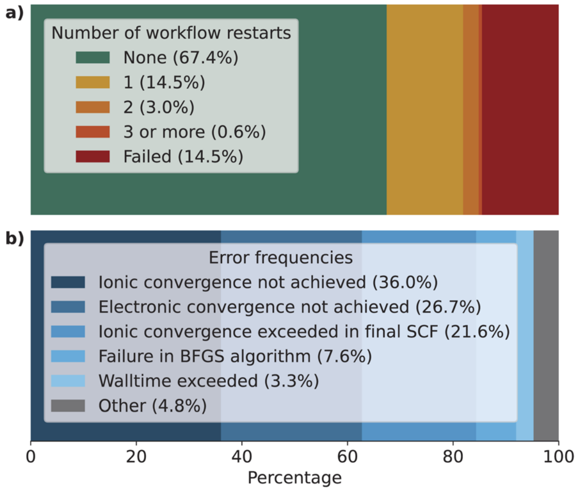


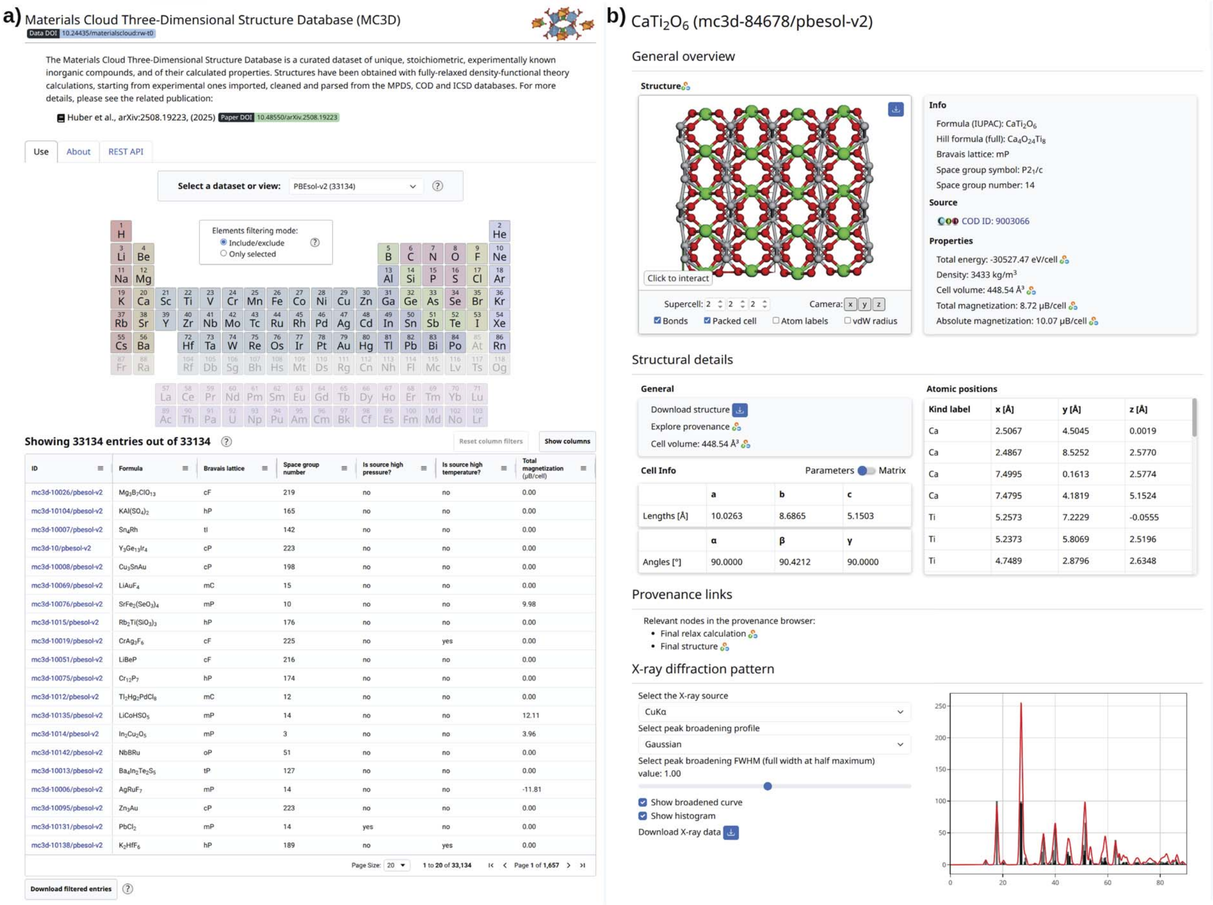
Fig. 3 Bar charts of (a) the number of restarts for each geometry optimization workflow and (b) the frequency of the 5 most common observed errors (including those that were recovered). The vast majority of workflows finish after the first iteration (67.4%). Workflows that need 1, 2, or 3 or more restarts are less common with 14.5%, 3.0%, and 0.6%, respectively. About 14.5% of workflows fail to complete successfully despite the error handling mechanism. The most common observed errors are the failure of the `pw.x` simulation code to achieve ionic or electronic convergence within the maximum allowed number of iterations.

highlight that over 67% of the workflows completed successfully after the first execution of `pw.x` (green bar in Fig. 3(a)) without triggering any error handling. We stress that this success rate could be achieved also thanks to the appropriate choices of input parameters defined by our protocols, that prevent certain failure modes, and by the use of advanced direct minimization algorithms for electronic convergence beyond the standard iterative self-consistent field (SCF) approach, such as direct-minimization strategies[70,71] as implemented in the `nlcglib`[72] plugin of SIRIUS library.[59] The vast majority of errors come from the failure to converge to the required precision either the geometry optimization or the self-consistent field calculation (within the maximum allowed number of iterations). Notably, the error handling mechanism implemented in our workflows (see the Methods section for more details) manages to recover from these errors in over half of the cases in which a single `pw.x` execution fails (orange bars in Fig. 3(a), with one or more workflow restarts). We emphasize that the error handler can decide to change the input of the calculation before resubmission, depending on the type of error reported in the code output. However, such changes do not include physical parameters of the calculation that would affect the results, but are limited only to those numerical aspects that can increase the chance of convergence (*e.g.*, reducing the mixing parameters of the self-consistent cycle or changing diagonalization algorithm). Nevertheless, despite our error-handling mechanisms, for a number of crystal structures (14.5% of the total, corresponding to the rightmost red segment in Fig. 3) the workflow failed to complete after exhausting the configured maximum number of retries. Work on improving the error handling and robustness of the workflows is ongoing, and we expect to further increase the success rate in future versions of the MC3D. We also stress that, because of the nature of the errors, achieving a significant increase in convergence rate cannot be obtained solely by improving workflow error-handling mechanisms, but also requires enhancements to the underlying algorithms in the `pw.x` code, as already addressed in this work by using the advanced convergence algorithms implemented in SIRIUS[59] discussed above.

To make the MC3D easily accessible and explorable, we make it available as a Discover section (https://mc3d.materialscloud.org/) on the Materials Cloud[64] portal at https://www.materialscloud.org/mc3d. Fig. 4 shows screenshots of the web application. The landing page allows the user to select a subdatabase of the MC3D and filter structures based on their composition *via* a periodic table. The matched structures are displayed in an index table that provides a link to a detail page together with several structural and material properties. The user can interactively edit the list of visible columns, adjust the row sorting, and apply column-based filtering. The detail pages give a detailed overview of the structural properties of the crystal as well as any other computed property. In particular, we provide powder X-ray diffraction (XRD) patterns computed using `pymatgen`[73] based on Bragg's law and atomic scattering factors. Reflections can be plotted using Gaussian or Lorentzian peaks with an area corresponding to peak intensity, and the full width at half maximum (FWHM) parameter can be set by the user. Notably, the full data and calculation provenance is directly accessible from the web interface. Computed properties are decorated with an AiiDA icon linking to the source calculation in the Materials Cloud Explore section (https://www.materialscloud.org/explore/mc3d-pbesol-v2/). There, users can browse the full provenance graph of MC3D, extract raw inputs and outputs, and reconstruct how the optimized geometries were obtained. A more detailed description of the web platform is provided in the Methods section.

Discussion

Our main goal is to provide a curated set of crystal structures, optimized with DFT using a consistent protocol, focusing on experimentally known structures. In this way, if structures from MC3D are used as a starting point for further screening studies, optimal candidates are already available for additional experiments. We highlight however that, while the source databases typically focus on experimental structures, they also report theoretical structures in some cases. By inspecting the metadata reported by the source databases, we could flag 2586 structures in MC3D `PBEsol-v2` that originate from structures tagged as theoretical (see details in the Methods section). This information can be accessed by the corresponding column in the Materials Cloud web interface. We stress that, however, we need to rely on the information provided by the source databases which might not always be accurate or complete. Therefore, users should be aware that, in some rare cases, some of the



Materials Cloud Three-Dimensional Structure Database (MC3D)

Data DOI 10.24435/materialscloud:rw-t0

The Materials Cloud Three-Dimensional Structure Database is a curated dataset of unique, stoichiometric, experimentally known inorganic compounds, and of their calculated properties. Structures have been obtained with fully-relaxed density-functional theory calculations, starting from experimental ones imported, cleaned and parsed from the MPDS, COD and ICSD databases. For more details, please see the related publication:

Huber et al., arXiv:2508.19223, (2025) Paper DOI 10.48550/arXiv.2508.19223

Select a dataset or view: PBEsol-v2 (33134)

Elements filtering mode:
● Include/exclude
○ Only selected

Showing 33134 entries out of 33134

ID	Formula	Bravais lattice	Space group number	Is source high pressure?	Is source high temperature?	Total magnetization (μB/cell)
mc3d-10026/pbesol-v2	$Mg_3B_7ClO_{13}$	cF	219	no	no	0.00
mc3d-10104/pbesol-v2	$KAl(SO_4)_2$	hP	165	no	no	0.00
mc3d-10007/pbesol-v2	Sn_4Rh	tI	142	no	no	0.00
mc3d-10/pbesol-v2	$Y_3Ge_{13}Ir_4$	cP	223	no	no	0.00
mc3d-10008/pbesol-v2	Cu_3SnAu	cP	198	no	no	0.00
mc3d-10069/pbesol-v2	$LiAuF_4$	mC	15	no	no	0.00
mc3d-10076/pbesol-v2	$SrFe_2(SeO_3)_4$	mP	10	no	no	9.98
mc3d-1015/pbesol-v2	$Rb_2Ti(SiO_3)_3$	hP	176	no	no	0.00
mc3d-10019/pbesol-v2	$CrAg_3F_6$	cF	225	no	yes	0.00
mc3d-10051/pbesol-v2	LiBeP	cF	216	no	no	0.00
mc3d-10075/pbesol-v2	$Cr_{12}P_7$	hP	174	no	no	0.00
mc3d-1012/pbesol-v2	$Tl_2Hg_2PdCl_8$	mC	12	no	no	0.00
mc3d-10135/pbesol-v2	$LiCoHSO_5$	mP	14	no	no	12.11
mc3d-1014/pbesol-v2	$In_2Cu_2O_5$	mP	3	no	no	3.96
mc3d-10142/pbesol-v2	NbBRu	oP	51	no	no	0.00
mc3d-10013/pbesol-v2	$Ba_4In_2Te_2S_5$	tP	127	no	no	0.00
mc3d-10006/pbesol-v2	$AgRuF_7$	mP	14	no	no	-11.81
mc3d-10095/pbesol-v2	Zn_3Au	cP	223	no	no	0.00
mc3d-10131/pbesol-v2	$PbCl_2$	mP	14	yes	no	0.00
mc3d-10138/pbesol-v2	K_2HfF_6	hP	189	no	yes	0.00

Page Size: 20 1 to 20 of 33,134 Page 1 of 1,657

$CaTi_2O_6$ (mc3d-84678/pbesol-v2)

General overview

Structure

Supercell: 2 2 2
☑ Bonds ☑ Packed cell □ Atom labels □ vdW radius

Info
Formula (IUPAC): $CaTi_2O_6$
Hill formula (full): $Ca_4O_{24}Ti_8$
Bravais lattice: mP
Space group symbol: $P2_1/c$
Space group number: 14

Source
COD ID: 9003066

Properties
Total energy: -30527.47 eV/cell
Density: 3433 kg/m^3
Cell volume: 448.54 Å^3
Total magnetization: 8.72 μB/cell
Absolute magnetization: 10.07 μB/cell

Structural details

General
Download structure
Explore provenance
Cell volume: 448.54 Å^3

Cell Info Parameters Matrix

	a	b	c
Lengths [Å]	10.0263	8.6865	5.1503

	α	β	γ
Angles [°]	90.0000	90.4212	90.0000

Atomic positions

Kind label	x [Å]	y [Å]	z [Å]
Ca	2.5067	4.5045	0.0019
Ca	2.4867	8.5252	2.5770
Ca	7.4995	0.1613	2.5774
Ca	7.4795	4.1819	5.1524
Ti	5.2573	7.2229	-0.0555
Ti	5.2373	5.8069	2.5196
Ti	4.7489	2.8796	2.6348

Provenance links

Relevant nodes in the provenance browser:
- Final relax calculation
- Final structure

X-ray diffraction pattern

Select the X-ray source: CuKα
Select peak broadening profile: Gaussian
Select peak broadening FWHM (full width at half maximum) value: 1.00
☑ Show broadened curve
☑ Show histogram
Download X-ray data

Fig. 4 Screenshot of the web application, integrated into the Materials Cloud, that provides a visual interface to explore the structures of the MC3D and the associated computed properties. (a) The landing page of the web application, providing an intuitive interface to select the desired version of the MC3D and a periodic table to filter for materials based on their composition. Moreover, the index table below gives an overview of the filtered structures together with a number of structural properties, and a link to a detail page. (b) The detail page for a specific material, showing detailed information and various computed properties, as well as links to the source calculation in the AiiDA provenance graph, providing direct access to the full provenance of how the property was obtained.

structures in the MC3D might not be experimentally known, even if they are not flagged as theoretical.

For the 33 142 successfully completed workflows, the optimized geometries were validated by comparing the difference in volume between the initial and optimized crystal structures, as large volume changes could be an indication of problems with either the initial or optimized structure. Fig. 5 shows a histogram of the relative volume change in percent between the initial and optimized geometry for the PBE and PBEsol functionals. The majority of optimized geometries have a change in volume within ±5% (78.1% and 62.0% for PBEsol and PBE, respectively), highlighting that the optimized geometry (especially when using the PBEsol functional) is not too far away from the experimentally determined source crystal structure.

The changes in volume between the source and optimized geometry are due to various factors. First of all, the source databases include thousands of layered or lower-dimensional structures.[74,75] However, since in this work we are not including van der Waals (vdW) corrections in the DFT calculations, the optimized geometries of these layered structures will typically have a potentially much larger cell volume than the source structure.[74] To further investigate this aspect, we applied the `lowdimfinder` code,[76] already used to generate the MC2D two-dimensional database,[74,75,77] to determine the dimensionality of disconnected substructures in each crystal structure in the `PBEsol-v2` dataset. We define the dimensionality of a material as the largest dimensionality that is found among any of the substructures. Using this classification, we indeed find that the 90% confidence interval of the relative volume changes is (−8.7%, 7.4%), (−10.2%, 21.0%), (−8.6%, 16.3%) and (−7.2%, 26.7%) for 3D, 2D, 1D and 0D structures, respectively, reflecting our expectation that the non-3D structures exhibit larger volume changes when relaxed without vdW corrections, especially for volume expansion.

Moreover, some source structures may have been characterized at high pressure and/or high temperature conditions (whereas our computational methods are limited to a temperature of 0 K and we considered a target zero pressure), which can also highly influence cell volume.

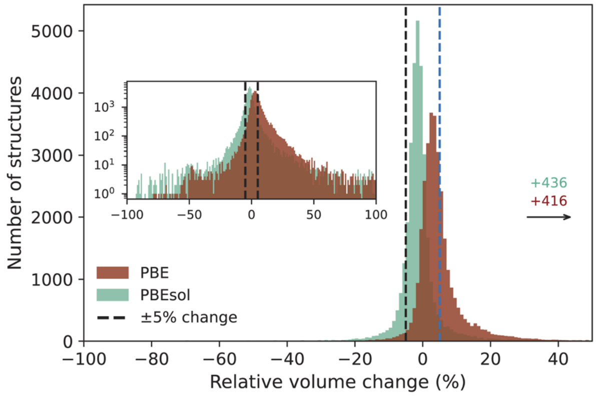


Fig. 5 Histogram of the relative differences between the cell volumes of the input and optimized structures for the `PBE-v1` and `PBEsol-v2` MC3D subdatabases. We note that the histograms for the two PBEsol versions (`PBEsol-v1` and `PBEsol-v2`) are visually almost indistinguishable, so we report only one of them for clarity. The bin size is 1%. A positive (negative) difference indicates that the optimized structure expanded (compressed) compared to the input structure. The majority of structures (78.1% and 62.0% for PBEsol and PBE, respectively) falls within the range of −5% to 5%, indicated by the vertical dashed lines. There are 436 and 416 structures for PBEsol and PBE, respectively, that have a relative volume change greater than 50% and are not shown in the image.

To account for these factors, we extracted the metadata from the source databases regarding the experimental conditions of the source structures, *i.e.*, the temperature and pressure. Unfortunately, the conditions at which structures have been characterized are only very rarely accurately reported in the CIF source, making it difficult to reliably classify the structures. For those where such data is available, we classify them as high pressure (high temperature) if the pressure (temperature) is reported and above 5000 atm (100 °C). In Fig. 6(a) we show a histogram of the relative volume change for structures classified as above, only including those structures that report explicitly both pressure and temperature, highlighting that the vast majority of the outliers are structures that are either high pressure or high temperature.

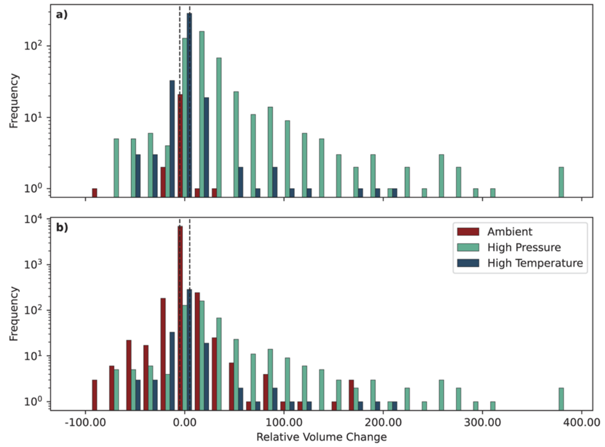


Fig. 6 Histogram of the relative volume change of structures, sorted into the three categories "Ambient", "High Pressure" and "High Temperature". (a) Structures with explicit ambient conditions: only structures where both pressure and temperature are reported in the source database are included in the plot. See main text for the thresholds defining high pressure and temperature. Structures that are both high pressure and high temperature are counted in both categories. (b) Structures with implicit ambient conditions: also structures where only one between pressure and temperature is reported in the source database, and this condition is below our thresholds, are included in the plot (and included in the "Ambient" group).

However, this plot includes very few structures, and in particular only very few (21) explicitly report "ambient" conditions (according to our definition of both pressure and temperature below 5000 atm and 100 °C, respectively). We therefore also plot in Fig. 6(b) the distribution of implicitly defined ambient structures, *i.e.*, those that only report one of the two conditions (pressure or temperature), and that condition is below our thresholds (*i.e.*, we assume that the other condition is also ambient). Although the majority of the most extreme outliers are still high pressure or high temperature, there are more ambient structures whose volume changes significantly. To better understand these outliers, we investigated the source papers of the 7 explicit ambient outliers outside the [−5%, 5%] range, as well as the 43 implicit ambient outliers outside the [−50%, 50%] range. Among the explicit ambient structures, we found that 6 were actually extracted from a high pressure/temperature study (but this was not correctly reported in the source databases), and one was a layered structure. Among the implicit ambient structures, also in this case about half were part of high pressure/temperature studies, layered structures, or molecular crystals. For other structures, the atomic occupations were most likely not fully stoichiometric, even if reported as such in the CIF file. Therefore, although we cannot explain all ambient outliers, it is clear that many of them are related to a possible incorrect reporting of the conditions or of the crystal structure, thus highlighting the importance of careful curation of the source databases.

The 33 142 successfully optimized geometries were analyzed for uniqueness, using the same method used to determine the unique structures of all source structures imported from the COD, ICSD and MPDS. The analysis found that, after the geometry optimization, another 1129 structures had become duplicates and the final set contains 32 013 unique optimized geometries.

Finally, we compared this set of unique structures with two other computational materials databases: the Materials Project[23] and the OQMD.[25] These databases mostly use the ICSD as the source of input crystal structures, with the Materials Project recently also including in part structures from the MPDS, but none have considered structures from the COD yet. As already mentioned, another key difference with this work is that instead of QE as the main DFT code, the Materials Project and the OQMD typically use VASP.[78,79] Our comparison shows that MC3D contains 3328 novel structures. Table 1 gives a detailed overview of the number of structures that have a space group or composition that does not occur in the Materials Project and/or in the OQMD (see the Methods section for more details on how structures are matched across the different databases).

Table 1 Number of newly contributed unique crystal structures according to pymatgen's StructureMatcher compared to the reference databases OQMD and Materials Project (and their union). We distinguish between three categories of new structures: Subtype A: number of structures in the MC3D with a composition that is not found in the reference. Subtype B: number of structures in the MC3D with an existing composition but a different space group with respect to the reference. Subtype C: number of structures in the MC3D with existing composition and space group combination with respect to the reference, but non-duplicate according to the StructureMatcher

New in MC3D: reference	Subtype A	Subtype B	Subtype C	Total new
Materials project[a]	2647	2704	1139	6490
OQMD[b]	2732	1731	1334	5797
Materials project & OQMD	1245	1065	1018	3328

[a] As downloaded using the Materials Project REST API in May 2025.
[b] Version 1.6, downloaded in May 2025.

In summary, we have presented MC3D, a database of relaxed geometries of three-dimensional stoichiometric inorganic structures computed using DFT. The database is built starting from almost a million experimentally known crystal structures imported from the COD, ICSD and MPDS databases. The structures were parsed and filtered to yield a collection of 72 589 unique stoichiometric inorganic crystal structures, which we refer to as the MC3D-source. We note that at most 69 284 of these are experimentally known, highlighting the small size of the space of experimentally known inorganic stoichiometric crystal structures. The geometries of a large fraction of these structures were then optimized with different functionals (PBE and PBEsol) and numerical protocols using a SIRIUS-enabled version of QE. The most accurate and precise subdatabase, PBEsol-v2, contains 32 013 unique structures. These optimizations were performed using fully automated workflows implemented in AiiDA, automatically handling errors that occurred during the calculations, and preserving the full provenance of all data produced. We demonstrate the workflow robustness by showing that they successfully complete for over 85% of the structures, significantly improving upon the success rate without any automated error handling. The resulting database of optimized geometries as well as their underlying provenance is made available on the Materials Cloud Archive[63] and can be browsed interactively on the Materials Cloud[64] portal at https://www.materialscloud.org/mc3d, including interactive access to the full provenance graph of the calculations. The key distinguishing features of MC3D include: (1) the inclusion of structures from several source databases, thus extending the coverage of materials; (2) its focus on experimentally known compounds, so that top-performing materials identified from a screening study of MC3D are easily available; (3) the use of a consistent set of computational methods and input protocols to compute optimized geometries, providing a solid foundation for the creation of training sets for machine-learning applications; (4) the use of open-source software in the whole pipeline and in particular of QE (and SIRIUS) as the DFT engine, at variance with many other existing databases,[22,23,25] thus providing the possibility to cross-validate results with different DFT implementations; and (5) the preservation of the full provenance of all data produced by the workflows, improving the reliability of derivative work based on MC3D and allowing other researchers to easily reproduce the results. MC3D thus constitutes a valuable resource for the materials science community complementing existing databases, and we expect it to be useful for applications including machine learning, materials discovery, and computational materials research.

Methods

Importing crystallographic data

The initial crystallographic data was imported from three external databases: the COD,[54] the ICSD,[55] and the MPDS.[56] All three databases provide an application programming interface (API) to query and download crystal structures in the CIF[66] format, although for the ICSD the provided MySQL database was accessed directly, without use of the API. AiiDA provides an interface, the DbImporter class, to implement functionality that allows importing structure from an external database. An implementation for the COD and ICSD already existed in the aiida-core[80] package, and the implementation for the MPDS was contributed for this work. A command line interface (CLI) was developed to download crystal structures through these importers and store them in an AiiDA[8] database with associated metadata, such as the Uniform Resource Locator (URL) and unique identifier used by the source database. The code is made available through the aiida-codtools[81] Python package v3.1.0.

The MPDS reports all of their structures as experimentally observed, and the experimental temperature and pressure condition data was retrieved from the condition field of the API responses. For the COD, the OPTIMADE compliant API was used, and structures were flagged as theoretical if _cod_method field equaled theoretical. Pressure and temperature data was retrieved from the fields _cod_cellpressure, _cod_diffrpressure, _cod_celltemp, and _cod_diffrtemp. In the case of the ICSD, the standard_remarks table in the 2020 MySQL version was parsed. Remarks "THE", "ZTHE", "ABIN", "DFT" were flagged as theoretical, pressure information was parsed from the "PRE", "ZPRE" fields, and temperature was parsed from the "TEM", "ZTEM" fields.

Using these criteria, out of the 72 589 structures in MC3D-source we flag 3305 structures as theoretical, leaving out 69 284 structures that might be experimentally known. We note that 2214 of these have no flag, as *e.g.* the 2020 version of the ICSD database does not report anymore some of the structures originally imported from ICSD version 2017.2; these might also need to be flagged as deprecated in the future, as this might point to structures that have been retracted or corrected in later versions of the database. Because of the nature of the three source databases, we expect the vast majority of these structures to be experimentally known, even if we cannot exclude that some of them might be theoretical structures not correctly flagged as such in the source databases.

Cleaning and parsing crystallographic data

The imported CIF files were cleaned using the `cod-tools` package v2.1.[82,83] The `cif-filter` and `cif-select` scripts were used to correct any incorrect syntax and remove unnecessary tags, respectively. The `cif-filter` script was run with the flags `-fix-syntax-errors`, `-use-c-parser` and `-use-datablocks-without-coordinates`. The `cif-select` script was run with the flags `-canonicalize-tag-names`, `-dont-treat-dots-as-underscores`, `-invert`, `-use-c-parser`, and `-tags='_publ_author_name,_citation_journal_abbrev'`. The cleaned CIF files were subsequently parsed using the `pymatgen` library v2018.12.12 (ref. 73) to extract the crystal structure defined through the cell basis vectors and atomic positions. Here a tolerance of 5×10^{-4} was employed in fractional coordinate distances to determine overlapping sites. The parsed crystal structure was then normalized and converted to primitive cell using `SeeK-path` v1.8.1,[84] using a symmetry precision `symprec` of 5×10^{-3} for the underlying `spglib` symmetry-detection code.[85] The entire cleaning, parsing, and normalizing process is implemented in a single AiiDA workflow, called the `CifCleanWorkChain`., which is available from the `aiida-codtools`[81] Python package. Fig. 7 shows the provenance graph produced by the execution of a `CifCleanWorkChain`.

After importing the CIF files with the `CifCleanWorkChain`, we noticed a significant number of parsed structures had a reduced formula that is inconsistent with the formula reported in the original CIF. To address this, a post-processing step was executed that compares the chemical formula in the CIF with the chemical formula of the parsed structure. Structures with inconsistent formulae were flagged in the database, and not considered for the next step in filtering procedure described in Fig. 1 of the Results section. Table 2 gives a complete overview of the cleaning results for all structures imported from the three external databases.

The last row shows the total number of files that have been imported and analyzed with the `CifCleanWorkChain`. The bold rows indicate the corresponding (aggregated) category in the Sankey diagram in Fig. 1.

Structure uniqueness analysis

The structures that were successfully parsed by the `CifCleanWorkChain` contain many duplicates. To determine the set of structures that are unique across all three databases, first the unique set of structures within each database was determined. The procedure first maps all the structures of a database on their chemical formula in the reduced Hill notation.[86] One then only has to search for duplicates within each set of structures that share the same chemical formula, drastically reducing the computational complexity of the problem.

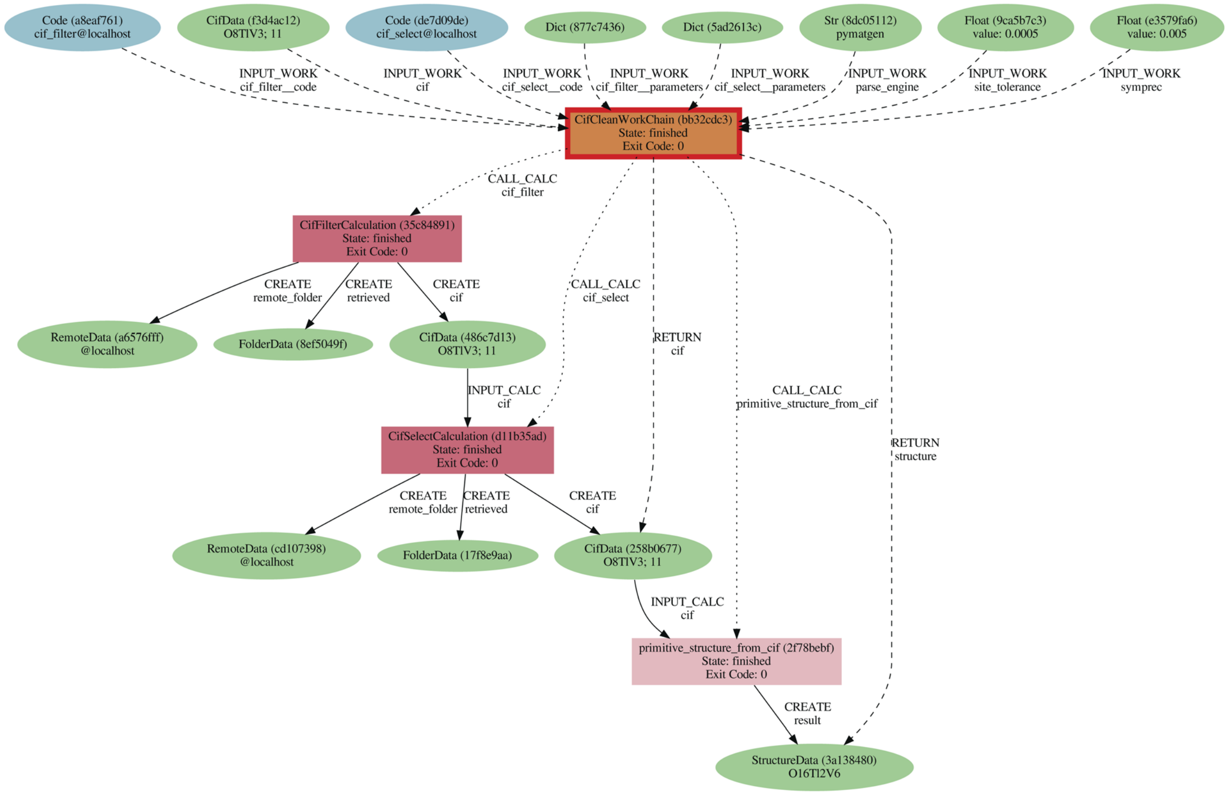


Fig. 7 Provenance graph of an execution of the `CifCleanWorkChain`. It takes a CIF file, represented by a `CifData` node, as input, and uses the `CifFilter` and `CifSelect` code plugins to correct and clean the file. The cleaned `CifData` is then parsed with `pymatgen` to extract the structure definition (cell vectors and atom coordinates), which is then normalized and converted to primitive cell with `SeeK-path` into a `StructureData`.

Table 2 The results of the `CifCleanWorkChain` workflows and post-processing formula check. Each row is a potential outcome of the process, where the first column describes the outcome and the remaining columns give the number of occurrences for CIFs imported from the COD, ICSD and MPDS, respectively. The "Different composition" and "Missing elements" outcomes are a result of the formula check, the other issues are direct failure modes (exit codes) of the `CifCleanWorkChain`.

	Occurrences			
Results of the CIF cleaning and parsing	COD	ICSD	MPDS	Total
Successfully cleaned, parsed and normalized the crystal structure	347 802	163 421	211 942	723 165
Problematic CIFs				
Different composition	**36 358**	**7809**	**42 819**	**86 986**
Inconsistent formula: different compositions	36 358	7809	42 819	86 986
Missing elements	**9743**	**16 590**	**21 812**	**48 145**
Inconsistent formula: missing hydrogen	7426	15 249	18 338	41 013
Inconsistent formula: missing lithium	42	77	133	252
Inconsistent formula: missing other	2275	1264	3341	6880
Failed to parse	**8314**	**5978**	**28 060**	**42 352**
The CIF had invalid syntax that could not be corrected	0	23	0	23
The cleaned CIF defines no atomic sites	1037	0	20 645	21 682
The cleaned CIF defines sites with invalid atomic occupancies	3136	3141	7415	13 692
The cleaned CIF contains sites with unknown species	1725	2814	0	4539
The cleaned CIF defines sites with attached hydrogens with incomplete positional data	2416	0	0	2416
Symmetry issues	**117**	**73**	**372**	**562**
Failed to determine the primitive structure	34	70	8	112
Detected inconsistent symmetry operations	83	3	364	450
Total	402 334	193 871	305 005	901 210

The second step is to separate the structures with the same formula by space group. This is mainly to avoid cases where the structure similarity algorithm is unable to properly differentiate distinct phases with very similar unit cells. The space group is determined using the `spglib` package with a `symprec` of 0.005. Using this strict approach, we give preference to cases where we consider two structures to be different that, in fact, should be flagged as similar. This is acceptable, since it is very likely that such cases will be detected as duplicates after the geometry has been optimized in the next step. As a consequence, we will be performing a geometry optimization for more structures than strictly necessary, but we will be less likely to discard distinct but similar crystal phases that should be considered independently.

For all structures that have the same reduced formula and space group, the similarity was analyzed using the `StructureMatcher` utility of the `pymatgen`[73] package (v2023.5.10), with a fractional length tolerance of 0.2, a site tolerance of 0.3 and an angle tolerance of 5°. The options to convert to primitive cell or to attempt to match a supercell were disabled, since all structures had already been normalized using `spglib` in the previous step of the pipeline. Finally, within each family of duplicate structures, a single representative structure was selected as the prototype. Preference was given to structures originating from databases with more permissive usage terms, *i.e.* we selected from the COD first, followed by ICSD and MPDS.

Removing organic molecular crystals

The successfully parsed structures were filtered a last time to remove any organic molecular crystal structures, as they are beyond the scope of this work. A well-defined heuristics-based algorithm to distinguish organic from inorganic crystal structures is complex to implement. However, we note that from all three source database, the COD contains a considerable amount of organic molecular crystal structures. Therefore, as a simple way to filter most of these structures, we remove all hydrogen-containing-structures originating from the COD that are unique to the COD. This method does not guarantee catching all organic molecular structures and it potentially also removes some inorganic crystal structures, *e.g.*, metal hydrides (if they are only included in COD), but this can be remedied in future versions of MC3D by including again a selection of the structures that have been discarded here. In this final filtering step, 208 509 structures originating from the COD were removed.

Geometry optimization

After all cleaning and filtering steps, a collection of 72 589 experimental unique inorganic crystal structures remains, which is referred to as the MC3D-source. The structures of the MC3D-source form the starting point of the next part of the work that seeks to computationally optimize their geometries.

The geometry optimization is performed in several steps, implemented as an AiiDA workflow, as shown in Fig. 8. The first step performs an initial geometry optimization with looser convergence parameters. This helps with two aspects: (1) it is less costly to obtain a reasonable first ground state geometry, improving the efficiency and (2) the less strict parameters makes it easier to converge for geometries far removed from the ground state, improving robustness. If the initial geometry optimization completes successfully, the optimization workflow

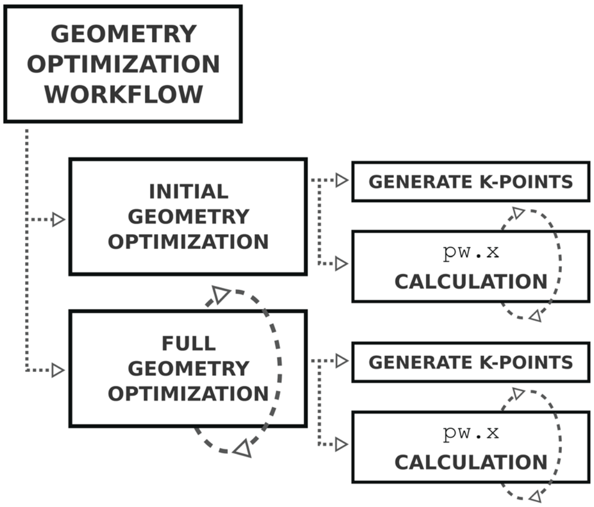

GEOMETRY OPTIMIZATION WORKFLOW
INITIAL GEOMETRY OPTIMIZATION
FULL GEOMETRY OPTIMIZATION
GENERATE K-POINTS
pw.x CALCULATION
GENERATE K-POINTS
pw.x CALCULATION

Fig. 8 Schematic diagram of the steps of the geometry optimization workflow, as described in the text. A dashed circular arrow around a block indicates the step can be repeated.

is run again but this time with the convergence parameters as determined by the input protocol.

Both the initial and full optimization steps are instances of the same workflow, run with different input parameters. Each workflow starts with generating the *k*-point mesh used to sample the Brillouin zone based on the input geometry, such that the *k*-point density in reciprocal space matches the minimum required density specified by the input protocol. The geometry is then optimized using QE's `pw.x` code[57,58] employing the SIRIUS library[59] to allow running efficiently on Graphics Processing Unit (GPU) compute nodes of different architectures. The workflow automatically parses the output of the calculation and, in case of an error or a failure to converge, the input parameters are adjusted and restarted up to a maximum of 5 restarts. The logic for this automated workflow is described in the "Automated error handling" section of the Methods. The final optimization step is repeated until the following two conditions are met:

(1) The stresses in the final self-consistent field (SCF) performed by `pw.x` are below the selected convergence threshold: during the geometry optimization, the basis sets used (*i.e.* the list of reciprocal-space *G* vectors), that depend not only on the energy cutoffs but also on the crystal geometry, are not updated at each step. The final SCF step that is performed at the end of the optimization recomputes the basis sets on the optimized unit cell. Large stresses in the final SCF indicate that the optimized geometry differs significantly from the initial geometry and the static basis sets used towards the end of the optimization were no longer converged enough. A restart of the optimization at the latest geometry is therefore performed.

(2) The *k*-point mesh corresponds to a lower density than the one dictated by the protocol: since the unit cell can change during the optimization, it is possible that, towards the end of the optimization cycle, the initial *k*-point mesh no longer satisfies the minimum required *k*-point density specified by the input protocol. In this case, a new *k*-point mesh is generated and another geometry optimization is performed.

Input parameter protocol

The geometry optimization workflow takes a number of input parameters that control the precision of the calculations performed. For the `PBEsol-v2` subdatabase of the MC3D (see next subsection for the `PBE-v1` and `PBEsol-v1` subdatabases), pseudopotentials were taken from the Standard Solid State Pseudopotentials (SSSP) PBEsol[61,62] efficiency v1.3 (ref. 87 and 88) library, which collects pseudopotentials from a number of libraries.[89–95] SSSP provides a set of rigorously tested values for the recommended wavefunction and charge density energy cutoffs for each pseudopotential. For each structure, the highest value among those recommended for the elements in the composition of the material was selected. It should be noted here that recent tests *versus* an all-electron benchmark[68] have found a suboptimal precision for the Hg and Au pseudopotentials provided by the SSSP v1.3 efficiency. Structures containing these elements are currently being recomputed with improved pseudopotentials and cutoffs, which will be added to the online database at a later point in time.

The magnetic and electronic properties cannot be reliably determined by just inspecting the geometry and composition of a crystal structure. Therefore, by default, all structures are considered to be magnetic and metallic in behavior in the initial geometry optimization. All calculations in the geometry optimization workflow are spin-polarized and are performed with smeared electronic occupations using the Marzari–Vanderbilt[96] cold-smearing method with a broadening of 0.02 Ry ($\approx 2.72 \times 10^{-1}$ eV). Each calculation is initialized in a high-spin ferromagnetic configuration, where elements with partially occupied *d* or f orbitals are assigned a magnetic moment of $5\mu_B$ or $7\mu_B$, respectively. For all other elements, the electrons are initialized to have a 10% surplus in the spin-up channel. To avoid an erroneous magnetic configuration being passed from the initial geometry optimization to the full one, magnetic moments are reinitialized in the second part of the workflow.

The Brillouin zone is sampled at *k*-points that are defined by a Monkhorst–Pack[97] mesh including the Γ-point, where the distance between *k*-points in each reciprocal-space direction is at most 0.15 Å^{-1} (*i.e.*, the algorithm chooses the smallest *k*-point mesh with at least the density required by the specified *k*-point distance). These values correspond to the extensively tested "balanced" protocol described in detail by de Miranda Nascimento *et al.*[69]

The threshold for electronic convergence in the SCF calculation is set to 0.2×10^{-9} Ry ($\approx 2.72 \times 10^{-9}$ eV) per atom. Geometry optimizations were stopped when the energy difference between two ionic steps was smaller than 10^{-5} Ry ($\approx 1.36 \times 10^{-4}$ eV) per atom, all forces on all atoms were smaller than 10^{-4} Ry bohr^{-1} ($\approx 2.57 \times 10^{-3}$ eV Å^{-1}) and cell stress was smaller than 0.5 kbar (0.05 GPa).

MC3D subdatabases and versioning

The structures of the MC3D-source have been optimized using three combinations of physical method and computational protocol version, yielding three methodology-based subdatabases. These are identified by the methodology labels

Table 3 Differences in geometry optimization workflow and the input parameter protocol for the three different subdatabases of the MC3D. The input protocols only differ in terms of the cold smearing value used for the BZ sampling and the version of the SSSP library. The version of the geometry optimization workflow corresponds to the major version of the aiida-quantumespresso[98] plugin package that publishes the workflow

	PBE-v1	PBEsol-v1	PBEsol-v2
Marzari–Vanderbilt smearing [Ry]	0.01	0.01	0.02
Pseudopotential library	SSSP v1.2 PBE efficiency	SSSP v1.2 PBEsol efficiency	SSSP v1.3 PBEsol efficiency
Geometry optimization workflow version	v4	v4	v5

PBE-v1, PBEsol-v1, and PBEsol-v2. The results discussed in this article focus on the most recent subdatabase, PBEsol-v2. Table 3 provides an overview of the differences among the three methodology labels.

The geometry optimization workflow, which for PBE-v1 and PBEsol-v1 used the workflows from aiida-quantumespresso version 4, was changed for PBEsol-v2 to the workflows of aiida-quantumespresso version 5, with the main differences being:

- For the MC3D v1 databases, the workflow ran a dedicated initial calculation of the electronic charge density to determine the electronic and magnetic character of the structure. This was dropped in PBEsol-v2 in favor of reinitializing the magnetic configuration for each geometry optimization;
- For the MC3D v1 databases, the workflow did not perform an initial geometry optimization with looser convergence parameters;
- For the MC3D v1 databases, the workflow only considered the volume difference between sequential geometry optimization runs in the convergence criteria;

As mentioned before, each methodology (physical method and computational protocol version) has been applied to (a subset of) the structures of the MC3D-source, which results in different optimized versions of these source structures. A naming convention was developed and adopted to unambiguously identify source structures of the MC3D-source and their optimized variants of the MC3D. All structures in the MC3D are given a unique identifier of the form mc3d-<source_id>. Optimized geometries derived from MC3D-source structures are given a unique identifier of the form mc3d-<source_ id>/<physical_method>-<protocol_version>. Here the <physical_method> and <protocol_version> refer to the name of the main distinguishing physical methodology used in the optimization workflow and version of the corresponding computational protocol. These two concepts are separated because a change in physical methodology does not necessarily mean that its results supersede those of a different subdatabase. However, for a given physical method, multiple versions of the computational protocol may be used. These typically represent improvements, and the resulting optimized geometries should, in principle, supersede those produced by earlier protocol versions.

Automated error handling

The geometry optimization workflow consists largely of runs of the pw.x simulation code. As presented in the Results section, these calculations have various ways of failing that need to be dealt with automatically by the workflow as much as possible for the workflow to be scalable to tens of thousands of structures.

Therefore, a logical workflow was developed to wrap the calculation in a loop and run it until it completes successfully, whose logic is schematically represented in Fig. 9. Each time a calculation fails, predefined error handlers are called that inspect the results of the failed calculation and determine whether to abort the workflow or to perform an operation, such as changing selected input parameters, before restarting the calculation. The workflow automatically aborts if two calculations fail consecutively without the intervention of an error handler, or the number of iterations exceeds a predefined maximum to prevent the workflow from running indefinitely. The list of main error handlers includes, in order of executed priority:

- Insufficient bands: this handler performs a post-calculation check on successfully converged calculations to check if for any spin channel and *k*-point the highest band has an occupation >0.005, which indicates that insufficient bands have been used in the calculation. In this case, increase the number of bands (nbnd) by 5% with a minimum of 4 bands.
- Diagonalization error: when the diagonalization algorithm fails, this handler systematically tries alternative algorithms in order of increasing robustness and computational cost (*i.e.*, starting with the fastest but potentially less robust algorithm). The default algorithm is david, and alternatives are tried in the order: ppcg, paro, and finally cg. If all algorithms have been tried, the calculation aborts.
- Out of walltime: handles calculations that exceeded the allocated walltime but shut down cleanly. Performs a full restart from the final structure (in case of a relaxation), charge density and wave functions.
- BFGS history failure: when the BFGS minimization algorithm fails, switches to the more robust damp dynamics algorithm (or reduces the trust radius for vc-relax calculations), then restarts from the previous structure and charge density.
- Ionic convergence: handles other cases where ionic convergence thresholds were not met but the calculation shut down cleanly with a usable output structure. Restarts from the previous charge density using the last output structure.
- Electronic convergence: handles electronic convergence failures by reducing charge density mixing (mixing_beta) by a factor of 0.8.

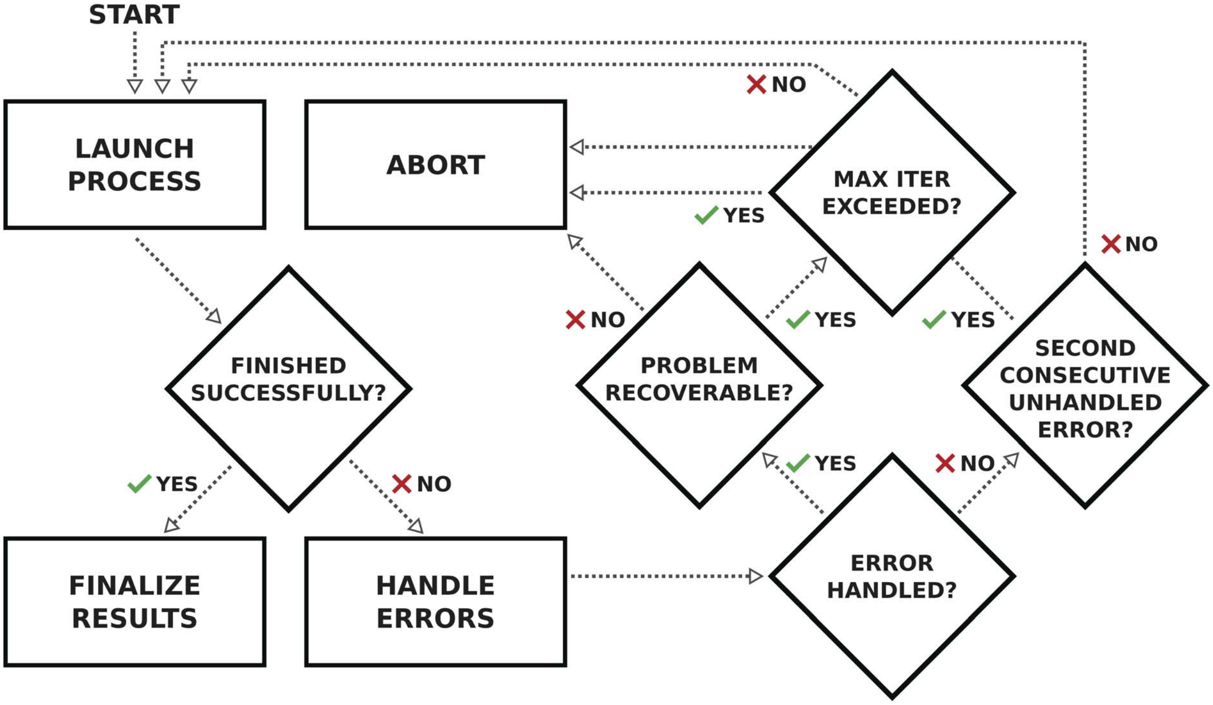


Fig. 9 Flow diagram of the logic implemented by the error handling workflow. The workflow enters the main loop and starts by running the first iteration of the calculation. In the next step, once the calculation has terminated, in case of success the results are reported and the workflow is terminated. Otherwise, in the case of an error, the registered error handlers are called in order. If the error is handled, the calculation is restarted as long as the maximum number of iterations has not been exceeded, otherwise the workflow is aborted. If two consecutive calculations fail without the intervention of an error handler, the workflow is also aborted.

Although specifically developed for this work, the logic of this workflow is generic and can be used in combination with any calculation. It has since been contributed to the AiiDA workflow management system that, as of v1.1, provides it as an integrated component called the `BaseRestartWorkChain` (see AiiDA documentation[99]). There are now many AiiDA plugin packages that provide implementations of the `BaseRestartWorkChain` for a variety of codes.

Structure matching and comparison

In order to determine the amount of newly added unique crystal structures compared to the Materials Project (MP) and Open Quantum Materials Database (OQMD), the same parameters as described in the Structure uniqueness analysis section have been adopted to initialize the `StructureMatcher`, using version `2025.5.2` of `pymatgen`. The only difference with the structure uniqueness analysis approach was that all cells were converted to primitive cell. This was disabled for the uniqueness analysis of the MC3D-source itself as that was already done that in a separate step before the analysis. However, this is not necessarily the case for the structures of the MP and OQMD, but even the structures of the MC3D-source may have changed significantly after the geometry optimization.

To compare structures of the MC3D, MP and OQMD, they were first grouped according to their space group and reduced chemical formula. Afterwards, each structure in the MC3D was compared against all the reference structures in the corresponding subgroup to identify whether any match is detected based on the `StructureMatcher`. To estimate the uncertainty of this analysis, since the assignment of the experimental reference to a certain structure might be handled differently in other databases, a matching was also performed on structures that refer to the same `icsd-id` (the internal identifier of structures in the ICSD, which is the source of the majority of structures in the MP and OQMD). For 657 (2.0%) out of 32 164 common `icsd-ids` (including also tags that were filtered out as duplicates), the final structures reported in the MC3D and MP do not match. However, MP also considers the final structure to determine the reference, whereas we assign the reference database tag based on the initial structures. In 295 cases, the final MC3D structure does not match the initial structure, due to significant changes during geometry optimization, whereas the MP structure does. Further differences might be related to different XC-functionals, *e.g.*, the PBE one adopted in the MP (the r^{2}SCAN calculations are based on PBEsol geometry optimizations). This high agreement justifies the application of the `StructureMatcher` approach to estimate the amount of structures in the MC3D that are not found in the MP and OQMD.

Web platform

Fig. 10 shows a schematic overview of the architecture of the web platform developed for the MC3D. The data of the `PBE-v1`, `PBEsol-v1`, and `PBEsol-v2` subdatabases of the MC3D are uploaded as AiiDA archive files and made publicly available on the Materials Cloud Archive. This data is imported into an AiiDA instance running on a Materials Cloud server, and served *via* the

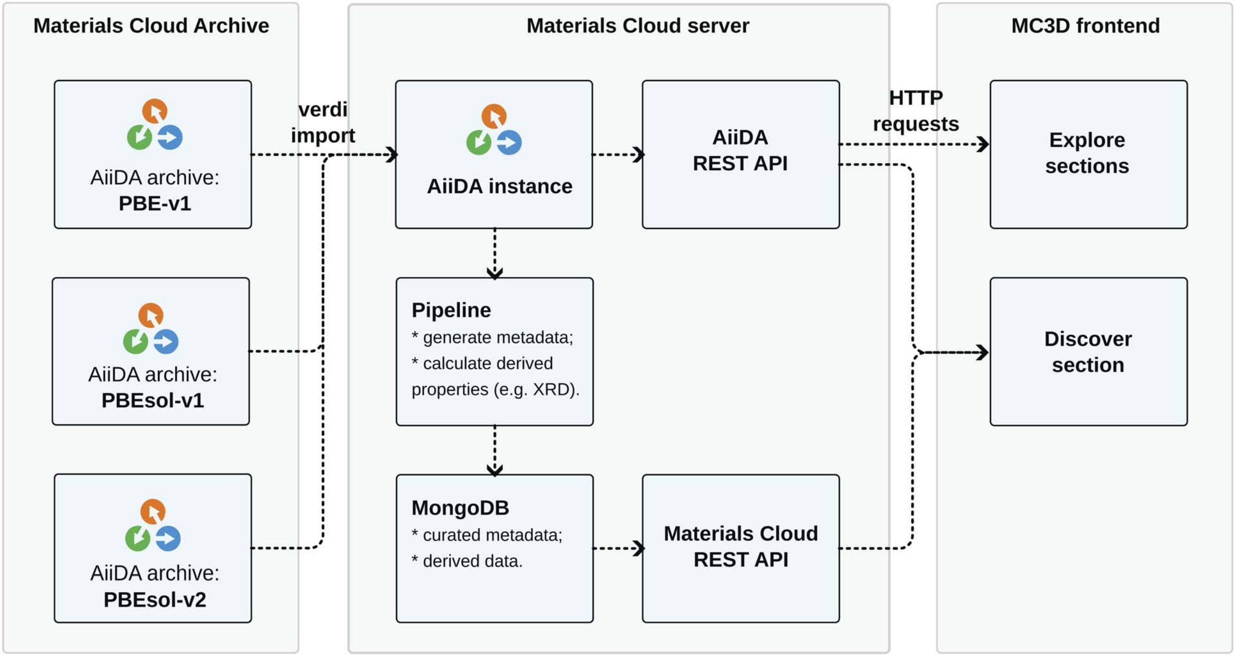


Fig. 10 Schematic overview of the web platform developed for the MC3D. The data of the different versions of the MC3D are uploaded as AiiDA archive files to the Materials Cloud Archive. A Materials Cloud server serves this raw data as well as a reduced set of metadata over two REST APIs, which is consumed by the web application that provides the graphic user interface.

AiiDA REST API, which can be explored through the Explore section of the web application (https://www.materialscloud.org/explore/mc3d-pbesol-v2/). However, while this API is powerful and gives access to the full AiiDA data, it does not provide an intuitive interface to easily find and extract relevant data for non-experts and it does not allow to serve any additional metadata that is not integral part of the AiiDA provenance graph, such as the X-ray diffraction (XRD) data that was calculated as a post-processing step without AiiDA. Therefore, a metadata pipeline was developed which transforms the content of the MC3D into a reduced format that just contains the optimized geometries with their relevant properties and links to the nodes in the AiiDA provenance graph that computed those properties, as well any additional metadata and derived data (such as the XRD data). This metadata is stored in a MongoDB database which is then served *via* the Materials Cloud REST API, which is in turn consumed by the Discover section of the web application.

In addition to the MC3D explore and discover frontends and their related APIs, we also deploy an OPTIMADE[50] compliant API to access the MC3D data. This makes it possible query the MC3D data in a standard manner, and explore it in any of the OPTIMADE-compliant clients available (such as the Materials Cloud OPTIMADE-client). We also provide an example Python script in Listing 1 (Chart 1) that demonstrates how to query the data programmatically, and convert structures to the Atomistic Simulation Environment[12] format *via* the `optimade-python-tools`[65] library. Four example queries are provided, which query structures based on number of elements, chemical composition, the explicit MC3D ID, and the total magnetization.

```
#!/usr/bin/env python
import requests
from optimade.adapters import Structure

BASE_URL = "https://optimade.materialscloud
    .org/main/mc3d-pbesol-v2"

example_queries = [
    "nelements = 6",
    'elements HAS "Mo" AND NOT elements HAS
         "O"',
    '_mcloud_mc3d_id = "mc3d-10011"',
    "_mcloud_total_magnetization > 30.0",
]

def iter_structures(base_url, query):
    url = f"{base_url}/structures?filter={
         query}"
    while url:
        res = requests.get(url)
        pl = res.json()
        for entry in pl.get("data", []):
            yield entry
        url = pl.get("links", {}).get("next
            ")

results = []
for entry in iter_structures(BASE_URL,
    example_queries[0]):
    # convert to ase
    results.append(Structure(entry).as_ase)
print("Returned structures:", len(results))
```

Chart 1 Listing 1: Example script to query MC3D structures via the OPTIMADE API and convert into Atomic Simulation Environment (ase) [12] objects. The only Python dependency is optimade[ase]==1.3.1.

Author contributions

Conceptualization: NM, GP; methodology: SPH, MM, MB, NH, NM, GP; supervision: NM, GP; software: SPH, MM, MB, TR, KE, NP, MU, GP; writing – original draft: SPH, MM, MB, TR, KE; writing – review & editing: SPH, MM, MB, TR, KE, NP, NH, MU, NM, GP; validation: SPH, MM, MB, TR, KE, GP; investigation: SPH, MM, MB, TR, KE, NP; project administration: NM, GP; funding acquisition: NM, GP; formal analysis: SPH, MM, MB, TR, KE; data curation: SPH, MM, MB, TR, KE; visualization: SPH, MM, MB, TR, KE, NP; resources: SPH, MM, MB, NM, GP.

Conflicts of interest

There are no conflicts of interest to declare.

Data availability

The databases of geometrically optimized structures are made available as AiiDA archive files on the Materials Cloud Archive (https://doi.org/10.24435/materialscloud:rw-t0),[63] together with Python scripts to extract the data from the archives into curated JSON files and then generate the figures provided in this paper. The content of the databases can also be interactively visualized on the dedicated Materials Cloud Discover section (https://www.materialscloud.org/mc3d). For structures originating from the open COD database, the AiiDA archive files contain the complete provenance of all calculations and data, from the initial import of the CIF file from the source database to the final DFT simulations. However, for structures originating from the commercial ICSD and MPDS datasets, we had to remove the original CIF files and the first part of the AiiDA provenance graph in order to avoid redistributing the original crystal structures, as required to comply with the licenses of these databases. Nevertheless, metadata needed to identify the original source structures, such as the source database name and structure ID, are preserved and can be accessed through the AiiDA archive files in the "extras" of the corresponding structures, or *via* the Materials Cloud frontend. The source code to import, parse and clean the CIFs from the COD, ICSD and MPDS are bundled in the `aiida-codtools` package v3.1.0 (https://doi.org/10.5281/zenodo.18406626),[100] which is made available under the MIT open-source license on GitHub at https://github.com/aiidateam/aiida-codtools. The automated workflows and input protocols to compute the optimized ground state electronic structure using Quantum ESPRESSO are bundled with the `aiida-quantumespresso` package (https://doi.org/10.5281/zenodo.18406736),[101] which is made available under the MIT open-source license on GitHub at https://github.com/aiidateam/aiida-quantumespresso. Both AiiDA plugin packages are distributed as installable packages through the Python Package Index, accessible at https://pypi.org/project/aiida-codtools and https://pypi.org/project/aiida-quantumespresso, respectively.

Acknowledgements

We acknowledge funding by the NCCR MARVEL, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant number 205602). We acknowledge support by the European Centre of Excellence MaX "Materials design at the Exascale" (Grant No. 824143). NP was partially supported by a MARVEL INSPIRE Potentials Master's Fellowship. NP and GP acknowledge financial support by the Swiss National Science Foundation (SNSF) Project Funding (grant 200021E_206190 "FISH4DIET"). MM, MB and GP acknowledge financial support by the SwissTwins project, funded by the Swiss State Secretariat for Education, Research and Innovation (SERI). GP acknowledges financial support by the Open Research Data Program of the ETH Board (project "PREMISE": Open and Reproducible Materials Science Research). NM acknowledges support from MaX – Materials design at the Exascale under grant agreement no. 101093374. We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LUMI, hosted by CSC (Finland) and the LUMI consortium through a EuroHPC Extreme Scale Access call (project 465000416) and pilot access to both the LUMI-C and LUMI-G machines under projects 465000028 and 465000170, respectively. This work also was supported by a grant from the Swiss National Supercomputing Centre (CSCS) on Piz Daint under project IDs s836 and s854, and on the Swiss share of the LUMI system under project ID 465000106. We acknowledge PRACE for awarding us access to Piz Daint at the Swiss National Supercomputing Centre (CSCS), Switzerland with the project id pr110. We further acknowledge access to Piz Daint and Alps at the Swiss National Supercomputing Centre, Switzerland under the MARVEL share with the project IDs mr0 and mr32. We acknowledge the use of GH-200 nodes at the Swiss National Supercomputing Centre (CSCS) on the Tödi supercomputer within the Alps infrastructure. This work has been partially supported by MIAI@Grenoble Alpes, (ANR-19-P3IA-0003). We gratefully acknowledge fruitful discussions and support by Lorenzo Bastonero, Dominik Gresch, Gabriel de Miranda Nascimento, Maria Grazia Giuffreda, Michele Kotiuga, Anton Kozhevnikov, Snehal Kumbhar, Bud Macaulay, Tiziano Müller, Elsa Passaro, Simon Pintarelli, Giannis Savva, Leopold Talirz, Joost VandeVondele, Aliaksandr V. Yakutovich, Jusong Yu, and the members of the AiiDA and Materials Cloud team.

References

1 A. R. Oganov, C. J. Pickard, Q. Zhu and R. J. Needs, *Nat. Rev. Mater.*, 2019, **4**, 331.

2 N. Marzari, A. Ferretti and C. Wolverton, *Nat. Mater.*, 2021, **20**, 736.

3 P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864–B871.

4 W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, A1133–A1138.

5 R. Van Noorden, B. Maher and R. Nuzzo, *Nature*, 2014, **514**, 550–553.

6 R. Van Noorden, *Nature*, 2025, **640**, 591.

7 S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito and O. Levy, *Nat. Mater.*, 2013, **12**, 191–201.

8 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, *Sci. Data*, 2020, **7**, 300.

9 M. Uhrin, S. P. Huber, J. Yu, N. Marzari and G. Pizzi, *Comput. Mater. Sci.*, 2021, **187**, 110086.

10 A. S. Rosen, M. Gallant, J. George, J. Riebesell, H. Sahasrabuddhe, J.-X. Shen, M. Wen, M. L. Evans, G. Petretto, D. Waroquiers, G.-M. Rignanese, K. A. Persson, A. Jain and A. M. Ganose, *J. Open Source Softw.*, 2024, **9**, 5995.

11 A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter and K. A. Persson, *Concurrency Comput. Pract. Exper.*, 2015, **27**, 5037.

12 A. Hjorth Larsen, J. Jørgen Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Duł ak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. Bjerre Jensen, J. Kermode, J. R. Kitchin, E. Leonhard Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. Bergmann 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, *J. Phys.: Condens. Matter*, 2017, **29**, 273002.

13 J. Janssen, S. Surendralal, Y. Lysogorskiy, M. Todorova, T. Hickel, R. Drautz and J. Neugebauer, *Comput. Mater. Sci.*, 2019, **163**, 24.

14 C. E. Calderon, J. J. Plata, C. Toher, C. Oses, O. Levy, M. Fornari, A. Natan, M. J. Mehl, G. Hart, M. Buongiorno Nardelli and S. Curtarolo, *Comput. Mater. Sci.*, 2015, **108**, 233.

15 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. B. Nardelli, M. Fornari, Y. Lederer, O. Levy, C. Toher and S. Curtarolo, *Comput. Mater. Sci.*, 2023, **217**, 111889.

16 C. S. Adorf, P. M. Dodd, V. Ramasubramani and S. C. Glotzer, *Comput. Mater. Sci.*, 2018, **146**, 220.

17 A. Rosen, *quacc – the quantum accelerator*, 2025.

18 W. Cunningham, A. Esquivel, C. Jao, F. Hasan, V. Bala, S. Sanand, P. Venkatesh, M. Tandon, O. E. Ochia, A. S. Rosen, dwelsch esi, jkanem, Aravind, HaimHorowitzAgnostiq, R. Li, S. W. Neagle, valkostadinov, A. Ghukasyan, P. U. Rao, S. Dutta, WingCode, A. Hughes, RaviPsiog, Udayan, Akalanka, A. Obasi, D. Singh, and FilipBolt, "Agnostiqhq/covalent: v0.228.0-rc.0", 2023.

19 K. Choudhary and F. Tavazza, *Comput. Mater. Sci.*, 2019, **161**, 300.

20 D. T. Speckhard, C. Carbogno, L. M. Ghiringhelli, S. Lübbeck, M. Scheffler and C. Draxl, *Phys. Rev. Mater.*, 2025, **9**, 013801.

21 D. D. Landis, J. S. Hummelshoj, S. Nestorov, J. Greeley, M. Dulak, T. Bligaard, J. K. Norskov and K. W. Jacobsen, *Comput. Sci. Eng.*, 2012, **14**, 51.

22 S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko and D. Morgan, *Comput. Mater. Sci.*, 2012, **58**, 218.

23 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.

24 J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, *JOM*, 2013, **65**, 1501.

25 S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl and C. Wolverton, *npj Comput. Mater.*, 2015, **1**, 15010.

26 C. Draxl and M. Scheffler, *MRS Bull.*, 2018, **43**, 676.

27 K. Choudhary, K. F. Garrity, A. C. E. Reid, B. DeCost, A. J. Biacchi, A. R. H. 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, *npj Comput. Mater.*, 2020, **6**, 173.

28 F. Q. Wang, K. Choudhary, Y. Liu, J. Hu and M. Hu, *Sci. Data*, 2022, **9**, 59.

29 J. Schmidt, H.-C. Wang, T. F. T. Cerqueira, S. Botti and M. A. L. Marques, *Sci. Data*, 2022, **9**, 64.

30 J. Shen, S. D. Griesemer, A. Gopakumar, B. Baldassarri, J. E. Saal, M. Aykol, V. I. Hegde and C. Wolverton, *J. Phys.: Mater.*, 2022, **5**, 031001.

31 L. Barroso-Luque, M. Shuaibi, X. Fu, B. M. Wood, M. Dzamba, M. Gao, A. Rizvi, C. L. Zitnick, and Z. W. Ulissi, Open materials 2024 (OMat24) inorganic materials dataset and models, *arXiv*, 2024, preprint, arXiv:2410.12771, DOI: 10.48550/arXiv.2410.12771.

32 A. D. Kaplan, R. Liu, J. Qi, T. W. Ko, B. Deng, J. Riebesell, G. Ceder, K. A. Persson, and S. P. Ong, A foundational potential energy surface dataset for materials, *arXiv*, 2025, preprint, arXiv:2503.04070, DOI: 10.48550/arXiv.2503.04070.

33 I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin, F. Berger, N. Bernstein, A. Bhowmik, S. M. Blau, V. Cărare, J. P. Darby, S. De, F. D. Pia, V. L. Deringer, R. Elijošius, Z. El-Machachi, F. Falcioni, E. Fako, A. C. Ferrari, A. Genreith-Schriever, J. George, R. E. A. Goodall, C. P. Grey, P. Grigorev, S. Han, W. Handley, H. H. Heenen, K. Hermansson, C. Holm, J. Jaafar, S. Hofmann, K. S. Jakob, H. Jung, V. Kapil, A. D. Kaplan, N. Karimitari, J. R. Kermode, N. Kroupa, J. Kullgren, M. C. Kuner, D. Kuryla, G. Liepuoniute, J. T. Margraf, I.-B. Magdău, A. Michaelides, J. H. Moore, A. A. Naik, S. P. Niblett, S. W. Norwood, N. O'Neill, C. Ortner, K. A. Persson, K. Reuter, A. S. Rosen,

L. L. Schaaf, C. Schran, B. X. Shi, E. Sivonxay, T. K. Stenczel, V. Svahn, C. Sutton, T. D. Swinburne, J. Tilly, C. v. d. Oord, E. Varga-Umbrich, T. Vegge, M. Vondrák, Y. Wang, W. C. Witt, F. Zills, and G. Csányi, A foundation model for atomistic materials chemistry, *arXiv*, 2023, preprint, arXiv:2401.00096, DOI: 10.48550/arXiv.2401.00096.

34 H. Yang, C. Hu, Y. Zhou, X. Liu, Y. Shi, J. Li, G. Li, Z. Chen, S. Chen, C. Zeni, M. Horton, R. Pinsler, A. Fowler, D. Zügner, T. Xie, J. Smith, L. Sun, Q. Wang, L. Kong, C. Liu, H. Hao, and Z. Lu, MatterSim: A deep learning atomistic model across elements, temperatures and pressures, *arXiv*, 2024, preprint, arXiv:2405.04967, DOI: 10.48550/arXiv.2405.04967.

35 A. Mazitov, F. Bigi, M. Kellner, P. Pegolo, D. Tisi, G. Fraux, S. Pozdnyakov, P. Loche, and M. Ceriotti, Pet-mad, a universal interatomic potential for advanced materials modeling, *arXiv*, preprint, 2025, arXiv:2503.14118, DOI: 10.48550/arXiv.2503.14118.

36 B. Rhodes, S. Vandenhaute, V. Šimkus, J. Gin, J. Godwin, T. Duignan, and M. Neumann, Orb-v3: atomistic simulation at scale, *arXiv*, 2025, preprint, arXiv:2504.06231, DOI: 10.48550/arXiv.2504.06231.

37 X. Fu, B. M. Wood, L. Barroso-Luque, D. S. Levine, M. Gao, M. Dzamba, and C. L. Zitnick, Learning smooth and expressive interatomic potentials for physical property prediction, *arXiv*, 2025, preprint, arXiv:2502.12147, DOI: 10.48550/arXiv.2502.12147.

38 H. Levämäki, F. Tasnádi, D. G. Sangiovanni, L. J. S. Johnson, R. Armiento and I. A. Abrikosov, *npj Comput. Mater.*, 2022, **8**, 17.

39 V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo and I. Takeuchi, *npj Comput. Mater.*, 2018, **4**, 29.

40 O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo and A. Tropsha, *Nat. Commun.*, 2017, **8**, 15679.

41 G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis and T. Lookman, *Sci. Rep.*, 2016, **6**, 19375.

42 B. Póta, P. Ahlawat, G. Csányi, and M. Simoncelli, Thermal conductivity predictions with foundation atomistic models, *arXiv*, 2024, preprint, arXiv:2408.00755, DOI: 10.48550/arXiv.2408.00755.

43 A. Loew, D. Sun, H.-C. Wang, S. Botti and M. A. L. Marques, *npj Comput. Mater.*, 2025, **11**, 178.

44 E. Mazhnik and A. R. Oganov, *J. Appl. Phys.*, 2020, **128**, 075102.

45 P. Avery, X. Wang, C. Oses, E. Gossett, D. M. Proserpio, C. Toher, S. Curtarolo and E. Zurek, *npj Comput. Mater.*, 2019, **5**, 89.

46 K. Gu, A. Wang, Z. Liu, Y. Kou, J. Liu, S. Li, C. Yang, G. Shi, T. Wu and S. P. Ong, *npj Comput. Mater.*, 2022, **8**, 142.

47 J. Schmidt, L. Pettersson, C. Verdozzi, S. Botti and M. A. L. Marques, *Sci. Adv.*, 2021, **7**, eabi7948.

48 C. Zeni, R. Pinsler and D. Zügner, A generative model for inorganic materials design, *Nature*, 2025, **639**, 624–632.

49 T. Xie, X. Fu, O.-E. Ganea, R. Barzilay, and T. Jaakkola, "Crystal diffusion variational autoencoder for periodic material generation, *arXiv*, 2021, preprint, arXiv:2110.06197, DOI: 10.48550/arXiv.2110.06197.

50 C. W. Andersen, R. Armiento, E. Blokhin, G. J. Conduit, S. Dwaraknath, M. L. Evans, Á. 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. A. 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, *Sci. Data*, 2021, **8**, 217.

51 M. L. Evans, J. Bergsma, A. Merkys, C. W. Andersen, O. B. Andersson, D. Beltrán, E. Blokhin, T. M. Boland, R. Castañeda Balderas, K. Choudhary, A. Díaz Díaz, R. Domínguez García, H. Eckert, K. Eimre, M. E. Fuentes Montero, A. M. Krajewski, J. J. Mortensen, J. M. Nápoles Duarte, J. Pietryga, J. Qi, F. d. J. Trejo Carrillo, A. Vaitkus, J. Yu, A. Zettel, P. B. de Castro, J. Carlsson, T. F. T. Cerqueira, S. Divilov, H. Hajiyani, F. Hanke, K. Jose, C. Oses, J. Riebesell, J. Schmidt, D. Winston, C. Xie, X. Yang, S. Bonella, S. Botti, S. Curtarolo, C. Draxl, L. E. Fuentes Cobas, A. Hospital, Z.-K. Liu, M. A. L. Marques, N. Marzari, A. J. Morris, S. P. Ong, M. Orozco, K. A. Persson, K. S. Thygesen, C. Wolverton, M. Scheidgen, C. Toher, G. J. Conduit, G. Pizzi, S. Gražulis, G.-M. Rignanese and R. Armiento, *Digital Discovery*, 2024, **3**, 1509.

52 D. Dragoni, T. D. Daff, G. Csányi and N. Marzari, *Phys. Rev. Mater.*, 2018, **2**(1), 013808.

53 A. Mazitov, S. Chorna, G. Fraux, M. Bercx, G. Pizzi, S. De, and M. Ceriotti, Massive atomic diversity: a compact universal dataset for atomistic machine learning, *arXiv*, 2025, preprint, arXiv:2506.19674, DOI: 10.48550/arXiv.2506.19674.

54 https://www.crystallography.net/cod/.

55 https://icsd.fiz-karlsruhe.de/.

56 https://www.mpds.io/.

57 P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari and R. M. Wentzcovitch, *J. Phys.: Condens. Matter*, 2009, **21**, 395502.

58 P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. Buongiorno Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A. Dal 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. Otero-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, *J. Phys.: Condens. Matter*, 2017, **29**, 465901.

59 https://github.com/electronic-structure/SIRIUS.

60 G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari and B. Kozinsky, *Comput. Mater. Sci.*, 2016, **111**, 218–230.

61 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, 77, 3865.

62 J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou and K. Burke, *Phys. Rev. Lett.*, 2008, **100**, 136406.

63 S. Huber, M. Minotakis, M. Bercx, T. Reents, K. Eimre, N. Paulish, N. Hörmann, M. Uhrin, N. Marzari, and G. Pizzi, *Materials Cloud Archive*, 2025, DOI: 10.24435/materialscloud:jn-ac.

64 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, *Sci. Data*, 2020, **7**, 299.

65 M. Evans, C. Andersen, S. Dwaraknath, M. Scheidgen, A. Fekete and D. Winston, *J. Open Source Softw.*, 2021, **6**, 3458.

66 S. R. Hall, F. H. Allen and I. D. Brown, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 1991, **47**, 655.

67 D. W. Davies, K. T. Butler, A. J. Jackson, A. Morris, J. M. Frost, J. M. Skelton and A. Walsh, *Chem*, 2016, **1**, 617.

68 E. Bosoni, L. Beal, M. Bercx, P. Blaha, S. Blügel, J. Bröder, M. Callsen, S. Cottenier, A. Degomme, V. Dikan, K. Eimre, E. Flage-Larsen, M. Fornari, A. Garcia, L. Genovese, M. Giantomassi, S. P. H. Huber, H. Janssen, G. Kastlunger, M. Krack, T. D. Kühne, K. Lejaeghere, G. K. H. Madsen, M. Marsman, N. Marzari, G. Michalicek, H. Mirhosseini, T. M. A. Müller, G. Petretto, C. J. Pickard, S. Poncé, G.-M. Rignanese, O. Rubel, T. Ruh, M. Sluydts, D. E. P. Vanpoucke, S. Vijay, M. Wolloch, D. Wortmann, A. V. Yakutovich, J. Yu, A. Zadoks and B. Zhu, *Nat. Rev. Phys.*, 2024, **6**, 45.

69 G. de Miranda Nascimento, F. J. dos Santos, M. Bercx, D. Grassano, G. Pizzi, and N. Marzari, Accurate and efficient protocols for high-throughput first-principles materials simulations, *arXiv*, 2025, preprint, arXiv:2504.03962[cond-mat.mtrl-sci], DOI: 10.48550/arXiv.2504.03962.

70 N. Marzari, D. Vanderbilt and M. C. Payne, *Phys. Rev. Lett.*, 1997, **79**, 1337–1340.

71 C. Freysoldt, S. Boeck and J. Neugebauer, *Phys. Rev. B*, 2009, **79**, 241103.

72 https://github.com/simonpintarelli/nlcglib.

73 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, *Comput. Mater. Sci.*, 2013, **68**, 314.

74 N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I. E. Castelli, A. Cepellotti, G. Pizzi and N. Marzari, *Nat. Nanotechnol.*, 2018, **13**, 246–252.

75 D. Campi, N. Mounet, M. Gibertini, G. Pizzi and N. Marzari, *ACS Nano*, 2023, **17**, 11268–11278.

76 https://github.com/epfl-theos/tool-ml-layer-finder/blob/master/compute/utils/lowdimfinder.py.

77 M. Tohidi Vahdat, K. Varoon Agrawal and G. Pizzi, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 045014.

78 G. Kresse and J. Furthmüller, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169.

79 G. Kresse and D. Joubert, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1999, **59**, 1758.

80 https://github.com/aiidateam/aiida-core.

81 https://github.com/aiidateam/aiida-codtools.

82 S. Gražulis, A. Merkys, A. Vaitkus and M. Okulič-Kazarinas, *J. Appl. Crystallogr.*, 2015, **48**, 85.

83 A. Merkys, A. Vaitkus, J. Butkus, M. Okulič-Kazarinas, V. Kairys and S. Gražulis, *J. Appl. Crystallogr.*, 2016, **49**, 292.

84 Y. Hinuma, G. Pizzi, Y. Kumagai, F. Oba and I. Tanaka, *Comput. Mater. Sci.*, 2017, **128**, 140.

85 A. Togo and I. Tanaka, Spglib: a software library for crystal symmetry search, *arXiv*, preprint, 2018, arXiv:1808.01590 [cond-mat.mtrl-sci], DOI: 10.48550/arXiv.1808.01590.

86 E. A. Hill, *J. Am. Chem. Soc.*, 1900, **22**, 478.

87 G. Prandini, A. Marrazzo, I. E. Castelli, N. Mounet and N. Marzari, *npj Comput. Mater.*, 2018, **4**, 72.

88 K. Lejaeghere, G. Bihlmayer, T. Bjorkman, P. Blaha, S. Blugel, V. Blum, D. Caliste, I. E. Castelli, S. J. Clark, A. D. Corso, S. de Gironcoli, T. Deutsch, J. K. Dewhurst, I. D. Marco, C. Draxl, M. D. ak, O. Eriksson, J. A. Flores-Livas, K. F. Garrity, L. Genovese, P. Giannozzi, M. Giantomassi, S. Goedecker, X. Gonze, O. Granas, E. K. U. Gross, A. Gulans, F. Gygi, D. R. Hamann, P. J. Hasnip, N. A. W. Holzwarth, D. I. an, D. B. Jochym, F. Jollet, D. Jones, G. Kresse, K. Koepernik, E. Kucukbenli, Y. O. Kvashnin, I. L. M. Locht, S. Lubeck, M. Marsman, N. Marzari, U. Nitzsche, L. Nordstrom, T. Ozaki, L. Paulatto, C. J. Pickard, W. Poelmans, M. I. J. Probert, K. Refson, M. Richter, G.-M. Rignanese, S. Saha, M. Scheffler, M. Schlipf, K. Schwarz, S. Sharma, F. Tavazza, P. Thunstrom, A. Tkatchenko, M. Torrent, D. Vanderbilt, M. J. van Setten, V. V. Speybroeck, J. M. Wills, J. R. Yates, G.-X. Zhang and S. Cottenier, *Science*, 2016, **351**, aad3000.

89 K. F. Garrity, J. W. Bennett, K. M. Rabe and D. Vanderbilt, *Comput. Mater. Sci.*, 2014, **81**, 446.

90 M. Schlipf and F. Gygi, *Comput. Phys. Commun.*, 2015, **196**, 36.

91 A. Willand, Y. O. Kvashnin, L. Genovese, Á. Vázquez-Mayagoitia, A. K. Deb, A. Sadeghi, T. Deutsch and S. Goedecker, *J. Chem. Phys.*, 2013, **138**, 104109.

92 A. D. Corso, *Comput. Mater. Sci.*, 2014, **95**, 337.

93 M. Topsakal and R. Wentzcovitch, *Comput. Mater. Sci.*, 2014, **95**, 263.

94 M. van Setten, M. Giantomassi, E. Bousquet, M. Verstraete, D. Hamann, X. Gonze and G.-M. Rignanese, *Comput. Phys. Commun.*, 2018, **226**, 39.

95 E. Kucukbenli, M. Monni, B. Adetunji, X. Ge, G. Adebayo, N. Marzari, S. de Gironcoli, and A. D. Corso, Projector augmented-wave and all-electron calculations across the periodic table: a comparison of structural and energetic properties, *arXiv*, 2014, preprint, arXiv:1404.3015[cond-mat.mtrl-sci], DOI: 10.48550/arXiv.1404.3015.

96 N. Marzari, D. Vanderbilt, A. D. Vita and M. C. Payne, *Phys. Rev. Lett.*, 1999, **82**, 3296.

97 H. J. Monkhorst and J. D. Pack, *Phys. Rev. B*, 1976, **13**, 5188.

98 https://github.com/aiidateam/aiida-quantumespresso.

99 AiiDA Development Team, AiiDA documentation: How to restart workchains, https://aiida.readthedocs.io/projects/aiida-core/en/v2.7.1/howto/workchains_restart.html, last accessed: 2025-07-24.

100 S. P. Huber, M. Minotakis, M. Bercx, T. Reents, K. Eimre, N. Paulish, N. Hörmann, M. Uhrin, N. Marzari and G. Pizzi, *Version of the 'aiida-codtools' AiiDA plugin used for MC3D: The Materials Cloud computational database of experimentally known stoichiometric inorganics*, 2026, DOI: 10.5281/zenodo.18406626.

101 S. P. Huber, M. Minotakis, M. Bercx, T. Reents, K. Eimre, N. Paulish, N. Hörmann, M. Uhrin, N. Marzari and G. Pizzi, *Version of the 'aiida-quantumespresso' AiiDA plugin used for MC3D: The Materials Cloud computational database of experimentally known stoichiometric inorganics*, 2026, DOI: 10.5281/zenodo.18406736.