



Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Recent progress of the computational 2D materials database (C2DB)

Gjerding, Morten Niklas; Taghizadeh, Alireza; Rasmussen, Asbjørn; Ali, Sajid; Bertoldo, Fabian; Deilmann, Thorsten; Knøsgaard, Nikolaj Rørbæk; Kruse, Mads; Larsen, Ask Hjorth; Manti, Simone; Pedersen, Thomas Garm; Petralanda, Urko; Skovhus, Thorbjørn; Svendsen, Mark Kamper; Mortensen, Jens Jørgen; Olsen, Thomas; Thygesen, Kristian Sommer

Published in:
2D materials

DOI (link to publication from Publisher):
[10.1088/2053-1583/ac1059](https://doi.org/10.1088/2053-1583/ac1059)

Creative Commons License
CC BY 4.0

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

[Link to publication from Aalborg University](#)

Citation for published version (APA):

Gjerding, M. N., Taghizadeh, A., Rasmussen, A., Ali, S., Bertoldo, F., Deilmann, T., Knøsgaard, N. R., Kruse, M., Larsen, A. H., Manti, S., Pedersen, T. G., Petralanda, U., Skovhus, T., Svendsen, M. K., Mortensen, J. J., Olsen, T., & Thygesen, K. S. (2021). Recent progress of the computational 2D materials database (C2DB). *2D materials*, 8(4), Article 044002. <https://doi.org/10.1088/2053-1583/ac1059>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

PAPER • OPEN ACCESS

Recent progress of the Computational 2D Materials Database (C2DB)

To cite this article: Morten Niklas Gjerding *et al* 2021 *2D Mater.* **8** 044002

View the [article online](#) for updates and enhancements.

You may also like

- [High throughput computational screening for 2D ferromagnetic materials: the critical role of anisotropy and local correlations](#)
Daniele Torelli, Kristian S Thygesen and Thomas Olsen
- [Exploring key aspects of art documentation and examination in curation technology-using Cheng-Shiu university conservation center's project as a case study](#)
L C Lin and I C Li
- [Machine learning enabled discovery of application dependent design principles for two-dimensional materials](#)
Victor Venturi, Holden L Parks, Zeeshan Ahmad et al.

2D Materials



PAPER

Recent progress of the Computational 2D Materials Database (C2DB)

OPEN ACCESS

RECEIVED
18 January 2021

REVISED
31 May 2021

ACCEPTED FOR PUBLICATION
30 June 2021

PUBLISHED
15 July 2021

Original Content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#).

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.



Morten Niklas Gjerding¹ , Alireza Taghizadeh^{1,2} , Asbjørn Rasmussen¹ , Sajid Ali¹ , Fabian Bertoldo¹ , Thorsten Deilmann³ , Nikolaj Rørbæk Knøsgaard¹ , Mads Kruse¹ , Ask Hjorth Larsen¹ , Simone Manti¹ , Thomas Garm Pedersen² , Urko Petralanda¹ , Thorbjørn Skovhus¹ , Mark Kamper Svendsen¹ , Jens Jørgen Mortensen¹ , Thomas Olsen¹ and Kristian Sommer Thygesen^{1,*}

¹ Computational Atomic-scale Materials Design (CAMD), Department of Physics, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

² Department of Materials and Production, Aalborg University, 9220 Aalborg Øst, Denmark

³ Institut für Festkörpertheorie, Westfälische Wilhelms-Universität Münster, 48149 Münster, Germany

* Author to whom any correspondence should be addressed.

E-mail: thygesen@fysik.dtu.dk

Keywords: 2D materials, high-throughput, ab-initio, database, density functional theory

Abstract

The Computational 2D Materials Database (C2DB) is a highly curated open database organising a wealth of computed properties for more than 4000 atomically thin two-dimensional (2D) materials. Here we report on new materials and properties that were added to the database since its first release in 2018. The set of new materials comprise several hundred monolayers exfoliated from experimentally known layered bulk materials, (homo)bilayers in various stacking configurations, native point defects in semiconducting monolayers, and chalcogen/halogen Janus monolayers. The new properties include exfoliation energies, Bader charges, spontaneous polarisations, Born charges, infrared polarisabilities, piezoelectric tensors, band topology invariants, exchange couplings, Raman spectra and second harmonic generation spectra. We also describe refinements of the employed material classification schemes, upgrades of the computational methodologies used for property evaluations, as well as significant enhancements of the data documentation and provenance. Finally, we explore the performance of Gaussian process-based regression for efficient prediction of mechanical and electronic materials properties. The combination of open access, detailed documentation, and extremely rich materials property data sets make the C2DB a unique resource that will advance the science of atomically thin materials.

1. Introduction

The discovery of new materials, or new properties of known materials, to meet a specific industrial or scientific requirement, is an exciting intellectual challenge of the utmost importance for our environment and economy. For example, the successful transition to a society based on sustainable energy sources and the realisation of quantum technologies (e.g. quantum computers and quantum communication) depend critically on new materials with novel functionalities. First-principles quantum mechanical calculations, e.g. based on density functional theory (DFT) [1], can predict the properties of materials with high accuracy even before they are made

in the lab. They provide insight into mechanisms at the most fundamental (atomic and electronic) level and can pinpoint and calculate key properties that determine the performance of the material at the macroscopic level. Powered by high-performance computers, atomistic quantum calculations in combination with data science approaches, have the potential to revolutionise the way we discover and develop new materials.

Atomically thin, two-dimensional (2D) crystals represent a fascinating class of materials with exciting perspectives for both fundamental science and technology [2–5]. The family of 2D materials has been growing steadily over the past decade and counts about a hundred materials that have been realised

in single-layer or few-layer form [6–10]. While some of these materials, including graphene, hexagonal boron nitride (hBN), and transition metal dichalcogenides (TMDCs), have been extensively studied, the majority have only been scarcely characterised and remain poorly understood. Computational studies indicate that around 1000 already known layered crystals have sufficiently weak interlayer (IL) bonding to allow the individual layers to be mechanically exfoliated [11, 12]. Supposedly, even more 2D materials could be realised beyond this set of already known crystals. Adding to this the possibility of stacking individual 2D layers (of the same or different kinds) into ultrathin van der Waals (vdW) crystals [13], and tuning the properties of such structures by varying the relative twist angle between adjacent layers [14, 15] or intercalating atoms into the vdW gap [16, 17], it is clear that the prospects of tailor made 2D materials are simply immense. To support experimental efforts and navigate the vast 2D materials space, first-principles calculations play a pivotal role. In particular, FAIR⁵ [18] databases populated by high-throughput calculations can provide a convenient overview of known materials and point to new promising materials with desired (predicted) properties. Such databases are also a fundamental requirement for the successful introduction and deployment of artificial intelligence in materials science.

Many of the unique properties exhibited by 2D materials have their origin in quantum confinement and reduced dielectric screening. These effects tend to enhance many-body interactions and lead to profoundly new phenomena such as strongly bound excitons [19–21] with nonhydrogenic Rydberg series [22–24], phonons and plasmons with anomalous dispersion relations [25, 26], large dielectric band structure renormalisations [27, 28], unconventional Mott insulating and superconducting phases [14, 15], and high-temperature exciton condensates [29]. Recently, it has become clear that long range magnetic order can persist [30, 31] and (in-plane) ferroelectricity even be enhanced [32], in the single layer limit. In addition, first-principles studies of 2D crystals have revealed rich and abundant topological phases [33, 34]. The peculiar physics ruling the world of 2D materials entails that many of the conventional theories and concepts developed for bulk crystals break down or require special treatments when applied to 2D materials [26, 35, 36]. This means that computational studies must be performed with extra care, which in turn calls for well-organised and well-documented 2D property data sets that can form the basis for the development, benchmarking, and consolidation of physical theories and numerical implementations.

The Computational 2D Materials Database (C2DB) [6, 37] is a highly curated and fully open database containing elementary physical properties of around 4000 2D monolayer crystals. The data has been generated by automatic high-throughput calculations at the level of DFT and many-body perturbation theory as implemented in the GPAW [38, 39] electronic structure code. The computational workflow is constructed using the atomic simulation recipes (ASR) [40]—a recently developed Python framework for high-throughput materials modelling building on the atomic simulation environment (ASE) [41]—and managed/executed using the MyQueue task scheduler [42].

The C2DB differentiates itself from existing computational databases of bulk [43–45] and low-dimensional [11, 12, 46–50] materials, by the large number of physical properties available, see table 1. The use of beyond-DFT theories for excited state properties (GW band structures and Bethe–Salpeter equation (BSE) absorption for selected materials) and Berry-phase techniques for band topology and polarisation quantities (spontaneous polarisation, Born charges, piezoelectric tensors), are other unique features of the database.

The C2DB can be downloaded in its entirety or browsed and searched online. As a new feature, all data entries presented on the website are accompanied by a clickable help icon that presents a scientific documentation (‘what does this piece of data describe?’) and technical documentation (‘how was this piece of data computed?’). This development enhances the usability of the database and improves the reproducibility and provenance of the data contained in C2DB. As another novelty it is possible to download all property data pertaining to a specific material or a specific type of property, e.g. the band gap, for all materials thus significantly improving data accessibility.

In this paper, we report on the significant C2DB developments that have taken place during the past two years. These developments can be roughly divided into four categories: (1) General updates of the workflow used to select, classify, and stabilise the materials. (2) Computational improvements for properties already described in the 2018 paper. (3) New properties. (4) New materials. The developments, described in four separate sections, cover both original work and review of previously published work. In addition, we have included some outlook discussions of ongoing work. In the last section we illustrate an application of statistical learning to predict properties directly from the atomic structure.

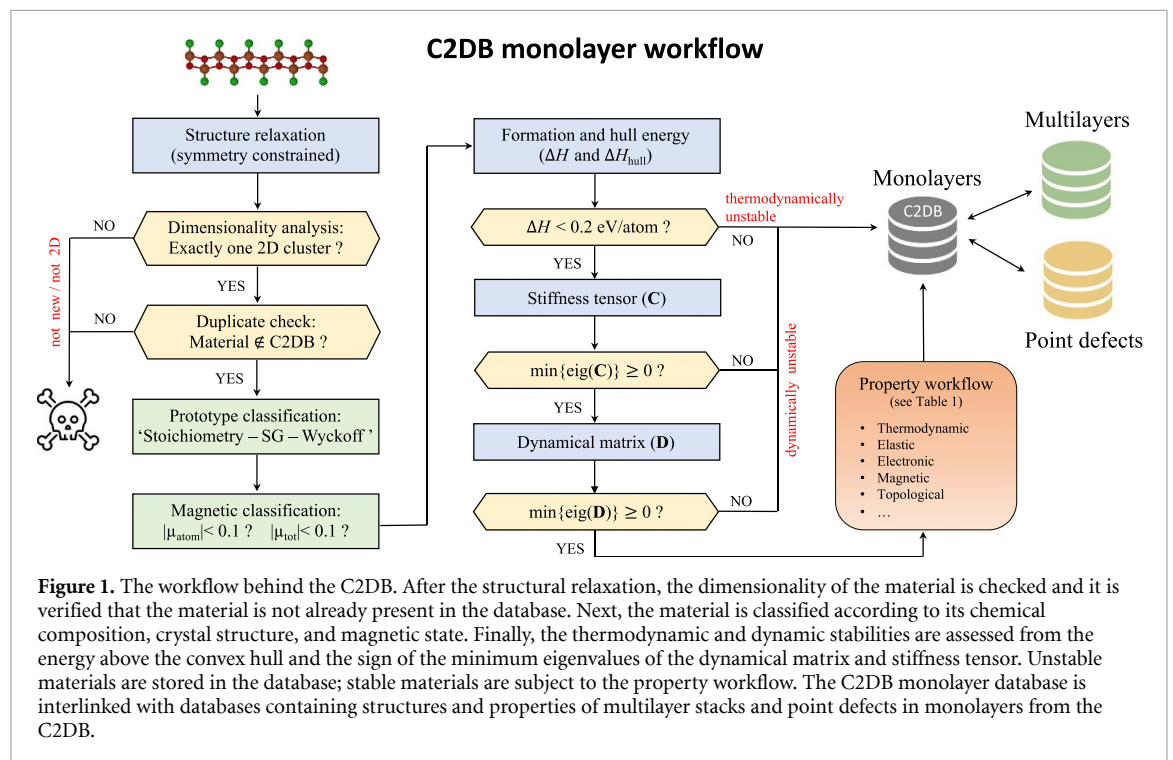
2. Selection, classification, and stability

Figure 1 illustrates the workflow behind the C2DB. In this section we describe the first part of the workflow

⁵ FAIR data are data which meet principles of findability, accessibility, interoperability, and reusability.

Table 1. Properties calculated by the C2DB monolayer workflow. The computational method and the criteria used to decide whether the property should be evaluated for a given material is also shown. A ‘*’ indicates that spin–orbit coupling (SOC) is included. All calculations are performed with the GPAW code using a plane wave basis except for the Raman calculations, which employ a double-zeta polarised basis of numerical atomic orbitals [51].

Property	Method	Criteria	Count
Bader charges	PBE	None	3809
Energy above convex hull	PBE	None	4044
Heat of formation	PBE	None	4044
Orbital projected band structure	PBE	None	2487
Out-of-plane dipole	PBE	None	4044
Phonons (Γ and BZ corners)	PBE	None	3865
Projected density of states	PBE	None	3332
Stiffness tensor	PBE	None	3968
Exchange couplings	PBE	Magnetic	538
Infrared polarisability	PBE	$E_{\text{gap}}^{\text{PBE}} > 0$	784
Second harmonic generation	PBE	$E_{\text{gap}}^{\text{PBE}} > 0$, non-magnetic, non-centrosymmetric	375
Electronic band structure PBE	PBE*	None	3496
Magnetic anisotropies	PBE*	Magnetic	823
Deformation potentials	PBE*	$E_{\text{gap}}^{\text{PBE}} > 0$	830
Effective masses	PBE*	$E_{\text{gap}}^{\text{PBE}} > 0$	1272
Fermi surface	PBE*	$E_{\text{gap}}^{\text{PBE}} = 0$	2505
Plasma frequency	PBE*	$E_{\text{gap}}^{\text{PBE}} = 0$	3144
Work function	PBE*	$E_{\text{gap}}^{\text{PBE}} = 0$	4044
Optical polarisability	RPA@PBE	None	3127
Electronic band structure	HSE06@PBE*	None	3155
Electronic band structure	G ₀ W ₀ @PBE*	$E_{\text{gap}}^{\text{PBE}} > 0$, $N_{\text{atoms}} < 5$	357
Born charges	PBE, Berry phase	$E_{\text{gap}}^{\text{PBE}} > 0$	639
Raman spectrum	PBE, LCAO basis set	Non-magnetic, dyn. stable	708
Piezoelectric tensor	PBE, Berry phase	$E_{\text{gap}}^{\text{PBE}}$, non-centrosym.	353
Optical absorbance	BSE@G ₀ W ₀ *	$E_{\text{gap}}^{\text{PBE}} > 0$, $N_{\text{atoms}} < 5$	378
Spontaneous polarisation	PBE, Berry phase	$E_{\text{gap}}^{\text{PBE}} > 0$, nearly centrosym. polar space group	151
Topological invariants	PBE*, Berry phase	$0 < E_{\text{gap}}^{\text{PBE}} < 0.3$ eV	242



until the property calculations (red box), focusing on aspects related to selection criteria, classification, and stability assessment, that have been changed or updated since the 2018 paper.

2.1. Structure relaxation

Given a prospective 2D material, the first step is to carry out a structure optimisation. This calculation is performed with spin polarisation and with the symmetries of the original structure enforced. The latter is done to keep the highest level of control over the resulting structure by avoiding ‘uncontrolled’ symmetry breaking distortions. The prize to pay is a higher risk of generating dynamically unstable structures.

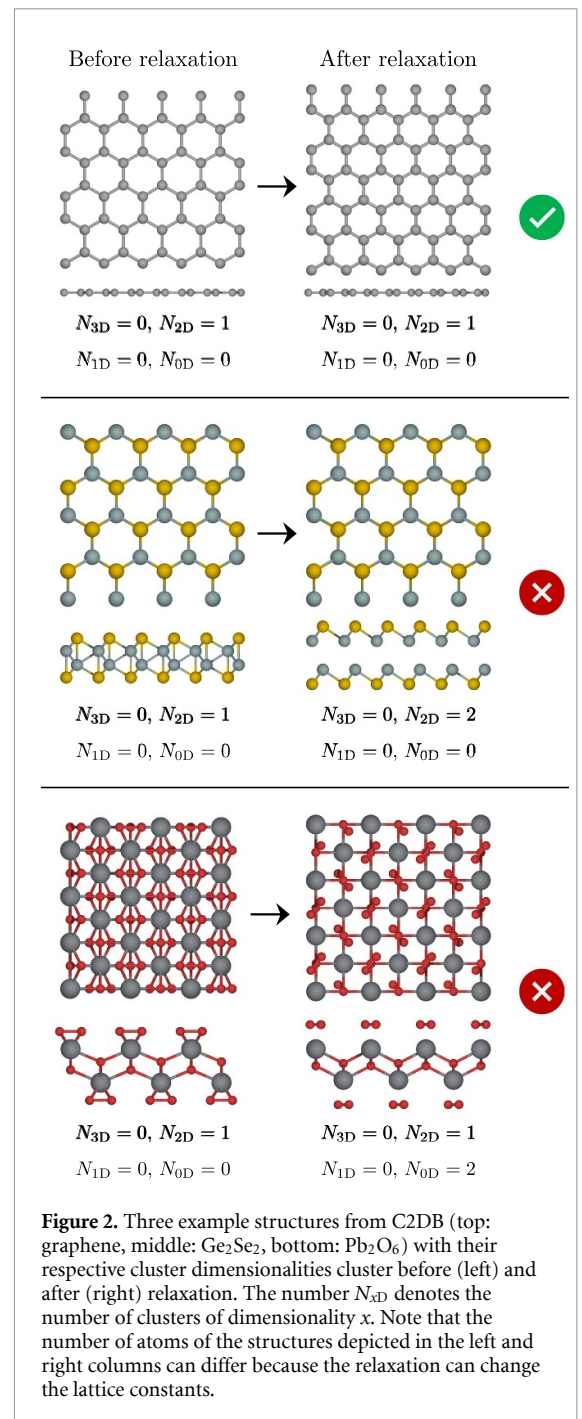
2.2. Selection: dimensionality analysis

A dimensionality analysis [52] is performed to identify and filter out materials that have disintegrated into non-2D structures during relaxation. Covalently bonded clusters are identified through an analysis of the connectivity of the structures where two atoms are considered to belong to the same cluster if their distance is less than some scaling of the sum of their covalent radii, i.e. $d < k(r_i^{\text{cov}} + r_j^{\text{cov}})$, where i and j are atomic indices. A scaling factor of $k = 1.35$ was determined empirically. Only structures that consist of a single 2D cluster after relaxation are further processed. Figure 2 shows three examples (graphene, Ge_2Se_2 , and Pb_2O_6) of structures and their cluster dimensionalities before and after relaxation. All structures initially consist of a single 2D cluster, but upon relaxation Ge_2Se_2 and Pb_2O_6 disintegrate into two 2D clusters as well as one 2D and two 0D clusters, respectively. On the other hand, the relaxation of graphene decreases the in-plane lattice constant but does not affect the dimensionality. According to the criterion defined above only graphene will enter the database.

2.3. Selection: ranking similar structures

Maintaining a high-throughput database inevitably requires a strategy for comparing similar structures and ranking them according to their relevance. In particular, this is necessary in order to identify different representatives of the same material e.g. resulting from independent relaxations, and thereby avoid duplicate entries and redundant computations. The C2DB strategy to this end involves a combination of structure clustering and Pareto analysis.

First, a single-linkage clustering algorithm is used to group materials with identical reduced chemical formula and ‘similar’ atomic configurations. To quantify configuration similarity a slightly modified version of PyMatGen’s [53] distance metric is employed where the cell volume normalisation is removed to make it applicable to 2D materials surrounded by vacuum. Roughly speaking, the metric measures the maximum distance an atom must be moved (in units of Å) in order to match the two



atomic configurations. Two atomic configurations belong to the same cluster if their distance is below an empirically determined threshold of 0.3 Å.

At this point, the simplest strategy would be to remove all but the most stable compound within a cluster. However, this procedure would remove many high symmetry crystals for which a more stable distorted version exists. For example, the well known T-phase of MoS_2 would be removed in favour of the more stable T'-phase. This is undesired as high-symmetry structures, even if dynamically unstable at $T = 0$, may provide useful information and might in fact become stabilised at higher temperatures [54]. Therefore, the general strategy adopted for the C2DB,

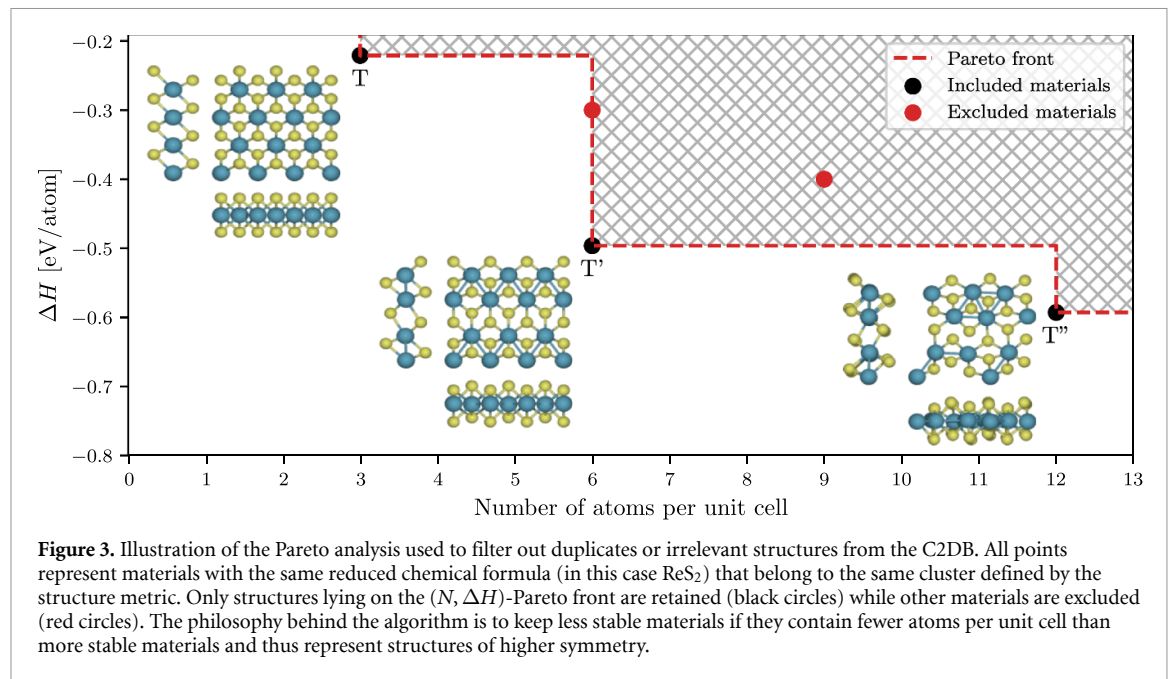


Figure 3. Illustration of the Pareto analysis used to filter out duplicates or irrelevant structures from the C2DB. All points represent materials with the same reduced chemical formula (in this case ReS_2) that belong to the same cluster defined by the structure metric. Only structures lying on the $(N, \Delta H)$ -Pareto front are retained (black circles) while other materials are excluded (red circles). The philosophy behind the algorithm is to keep less stable materials if they contain fewer atoms per unit cell than more stable materials and thus represent structures of higher symmetry.

is to keep a material that is less stable than another material of the same cluster if it has fewer atoms in its primitive unit cell (and thus typically higher symmetry). Precisely, materials within a given cluster are kept only if they represent a defining point of the $(N, \Delta H)$ -Pareto front, where N is the number of atoms in the unit cell and ΔH is the heat of formation. A graphical illustration of the Pareto analysis is shown in figure 3 for the case of ReS_2 .

2.4. Classification: crystal structure

The original C2DB employed a *crystal prototype* classification scheme where specific materials were promoted to prototypes and used to label groups of materials with the same or very similar crystal structure. This approach was found to be difficult to maintain (as well as being non-transparent). Instead, materials are now classified according to their *crystal type* defined by the reduced stoichiometry, space group number, and the alphabetically sorted labels of the occupied Wyckoff positions. As an example, MoS_2 in the H-phase has the crystal type: AB2-187-bi.

2.5. Classification: magnetic state

In the new version of the C2DB, materials are classified according to their magnetic state as either *non-magnetic* or *magnetic*. A material is considered magnetic if any atom has a local magnetic moment greater than $0.1 \mu_B$.

In the original C2DB, the *magnetic* category was further subdivided into ferromagnetic (FM) and anti-ferromagnetic (AFM). But since the simplest anti-ferromagnetically ordered state typically does not represent the true ground state, all material entries with an AFM state have been removed from the C2DB and replaced by the material in its FM state. Although the latter is less stable, it represents a

more well defined state of the material. Crucially, the nearest neighbour exchange couplings for all magnetic materials have been included in the C2DB (see section 5.8). This enables a more detailed and realistic description of the magnetic order via the Heisenberg model. In particular, the FM state of a material is not expected to represent the true magnetic ground if the exchange coupling $J < 0$.

2.6. Stability: thermodynamic

The heat of formation, ΔH , of a compound is defined as its energy per atom relative to its constituent elements in their standard states [55]. The thermodynamic stability of a compound is evaluated in terms of its energy above the *convex hull*, ΔH_{hull} , which gives the energy of the material relative to other competing phases of the same chemical composition, including mixed phases [6], see figure 4 for an example. Clearly, ΔH_{hull} depends on the pool of reference phases, which in turn defines the convex hull. The original C2DB employed a pool of reference phases comprised by 2807 elemental and binary bulk crystals from the convex hull of the Open Quantum Materials Database (OQMD) [55]. In the new version, this set has been extended by approximately 6783 ternary bulk compounds from the convex hull of OQMD, making a total of 9590 stable bulk reference compounds.

As a simple indicator for the thermodynamic stability of a material, the C2DB employs three labels (low, medium, high) as defined in table 2. These indicators are unchanged from the original version of the C2DB. In particular, the criterion $\Delta H_{\text{hull}} < 0.2 \text{ eV atom}^{-1}$, defining the most stable category, was established based on an extensive analysis of 55 experimentally realised monolayer crystals [6].

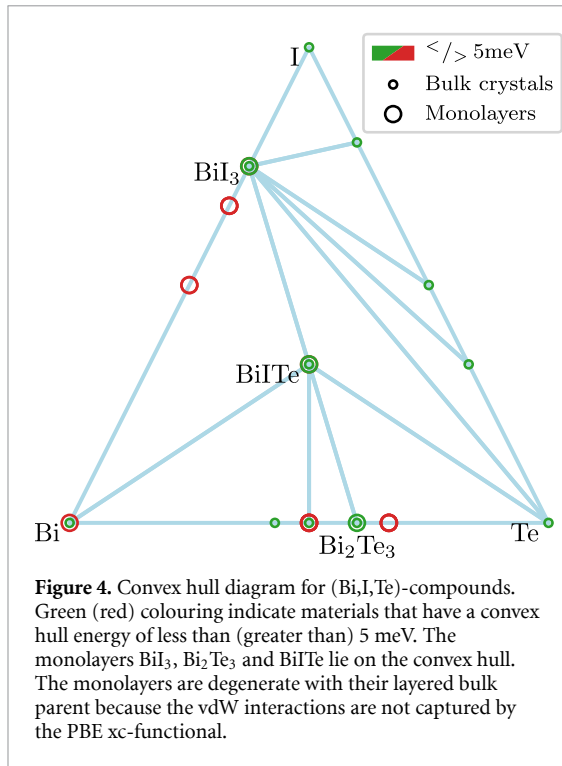


Figure 4. Convex hull diagram for (Bi,I,Te)-compounds. Green (red) colouring indicate materials that have a convex hull energy of less than (greater than) 5 meV. The monolayers BiI_3 , BiI_3Te_3 and BiI_3Te_5 lie on the convex hull. The monolayers are degenerate with their layered bulk parent because the vdW interactions are not captured by the PBE xc-functional.

Table 2. Thermodynamic stability indicator assigned to all materials in the C2DB. ΔH and ΔH_{hull} denote the heat of formation and energy above the convex hull, respectively.

Thermodynamic stability indicator	Criterion (eV atom^{-1})
Low	$\Delta H > 0.2$
Medium	$\Delta H < 0.2$ and $\Delta H_{\text{hull}} > 0.2$
High	$\Delta H < 0.2$ and $\Delta H_{\text{hull}} < 0.2$

It should be emphasised that the energies of both monolayers and bulk reference crystals are calculated with the Perdew-Burke-Ernzerhof (PBE) xc-functional [56]. This implies that some inaccuracies must be expected, in particular for materials with strongly localised d -electrons, e.g. certain transition metal oxides, and materials for which dispersive interactions are important, e.g. layered van der Waals crystals. The latter implies that the energy of a monolayer and its layered bulk parent (if such exists in the pool of references) will have the same energy. For further details and discussions see reference [6].

2.7. Stability: dynamical

Dynamically stable materials are situated at a local minimum of the potential energy surface and are thus stable to small structural perturbations. Structures resulting from DFT relaxations can end up in saddle point configurations because of imposed symmetry constraints or an insufficient number of atoms in the unit cell.

In C2DB, the dynamical stability is assessed from the signs of the minimum eigenvalues of (1) the stiffness tensor (see section 3.1) and (2) the Γ -point

Hessian matrix for a supercell containing 2×2 repetitions of the unit cell (the structure is not relaxed in the 2×2 supercell). If one of these minimal eigenvalues is negative the material is classified as dynamically unstable. This indicates that the energy can be reduced by displacing an atom and/or deforming the unit cell, respectively. The use of two categories for dynamical stability, i.e. stable/unstable, differs from the original version of the C2DB where an intermediate category was used for materials with negative but numerically small minimal eigenvalue of either the Hessian or stiffness tensors.

3. Improved property methodology

The new version of the C2DB has been generated using a significantly extended and improved workflow for property evaluations. This section focuses on improvements relating to properties that were already present in the original version of the C2DB while new properties are discussed in the next section.

3.1. Stiffness tensor

The stiffness tensor, C , is a rank-4 tensor that relates the stress of a material to the applied strain. In Mandel notation (a variant of Voigt notation) C is expressed as an $N \times N$ matrix relating the N independent components of the stress and strain tensors. For a 2D material $N = 3$ and the tensor takes the form:

$$C = \begin{bmatrix} C_{xxxx} & C_{xxyy} & \sqrt{2}C_{xxxy} \\ C_{xxyy} & C_{yyyy} & \sqrt{2}C_{yyxy} \\ \sqrt{2}C_{xxxy} & \sqrt{2}C_{yyxy} & 2C_{xyxy} \end{bmatrix}, \quad (1)$$

where the indices on the matrix elements refer to the rank-4 tensor. The factors multiplying the tensor elements account for their multiplicities in the full rank-4 tensor. In the C2DB workflow, C is calculated as a finite difference of the stress under an applied strain with full relaxation of atomic coordinates. A negative eigenvalue of C signals a dynamical instability, see section 2.7.

In the first version of the C2DB only the diagonal elements of the stiffness tensor were calculated. The new version also determines the shear components such that the full 3×3 stiffness tensor is now available. This improvement also leads to a more accurate assessment of dynamical stability [57].

3.2. Effective masses with parabolicity estimates

For all materials with a finite band gap the effective masses of electrons and holes are calculated for bands within 100 meV of the conduction band minimum and valence band maximum, respectively. The Hessian matrices at the band extrema (BE) are determined by fitting a second order polynomial to the PBE band structure including SOC, and the effective masses are obtained by subsequent diagonalisation of the Hessian. The main fitting-procedure is unaltered

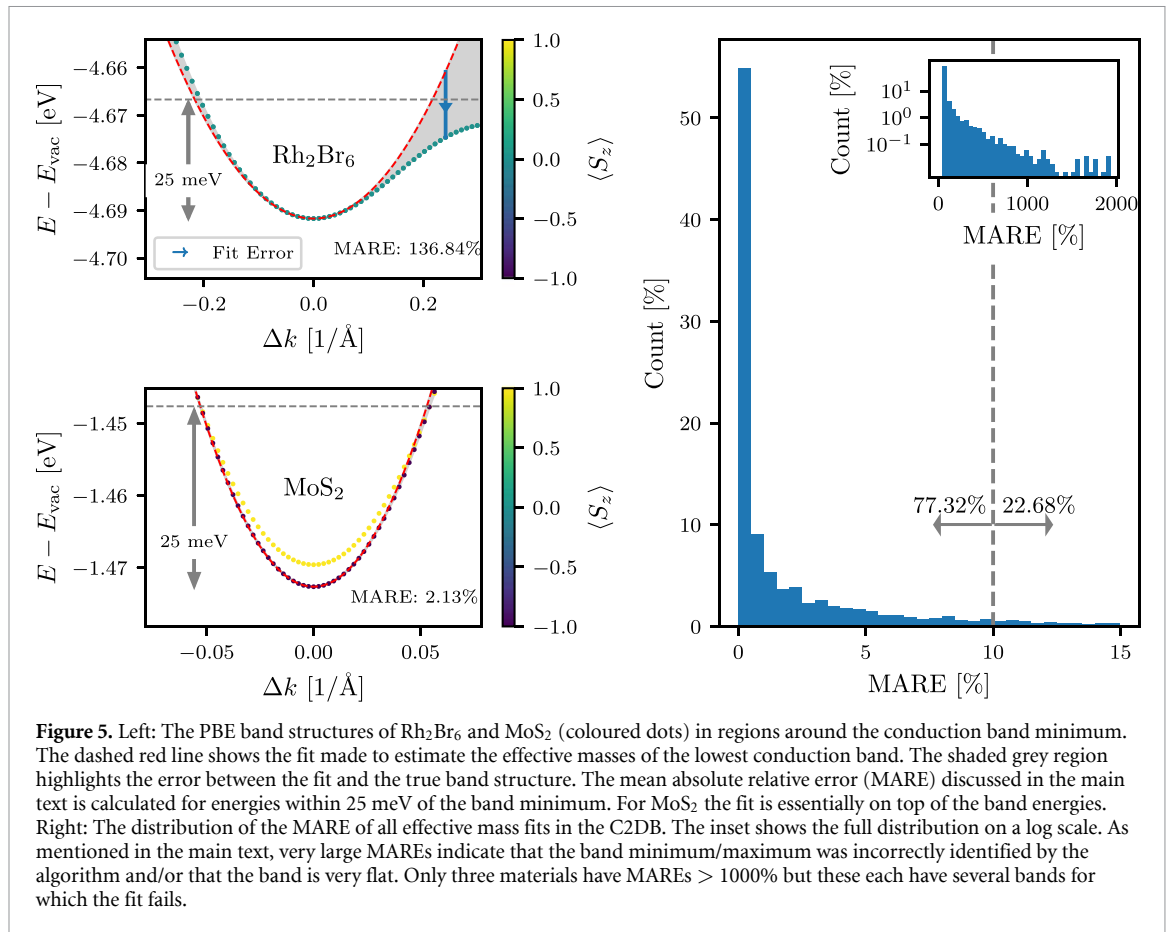


Figure 5. Left: The PBE band structures of Rh_2Br_6 and MoS_2 (coloured dots) in regions around the conduction band minimum. The dashed red line shows the fit made to estimate the effective masses of the lowest conduction band. The shaded grey region highlights the error between the fit and the true band structure. The mean absolute relative error (MARE) discussed in the main text is calculated for energies within 25 meV of the band minimum. For MoS_2 the fit is essentially on top of the band energies. Right: The distribution of the MARE of all effective mass fits in the C2DB. The inset shows the full distribution on a log scale. As mentioned in the main text, very large MAREs indicate that the band minimum/maximum was incorrectly identified by the algorithm and/or that the band is very flat. Only three materials have MAREs $> 1000\%$ but these each have several bands for which the fit fails.

from the first version of C2DB, but two important improvements have been made.

The first improvement consists in an additional k -mesh refinement step for better localisation of the BE in the Brillouin zone. After the location of the BE has been estimated based on a uniformly sampled band structure with k -point density of 12 \AA , another one-shot calculation is performed with a denser k -mesh around the estimated BE positions. This ensures a more accurate and robust determination of the location of the BE, which can be important in cases with a small but still significant spin-orbit splitting or when the band is very flat or non-quadratic around the BE. The second refinement step is the same as in the first version of C2DB, i.e. the band energies are calculated on a highly dense k -mesh in a small disc around the BE, and the Hessian is obtained by fitting the band energies in the range up to 1 meV from the BE.

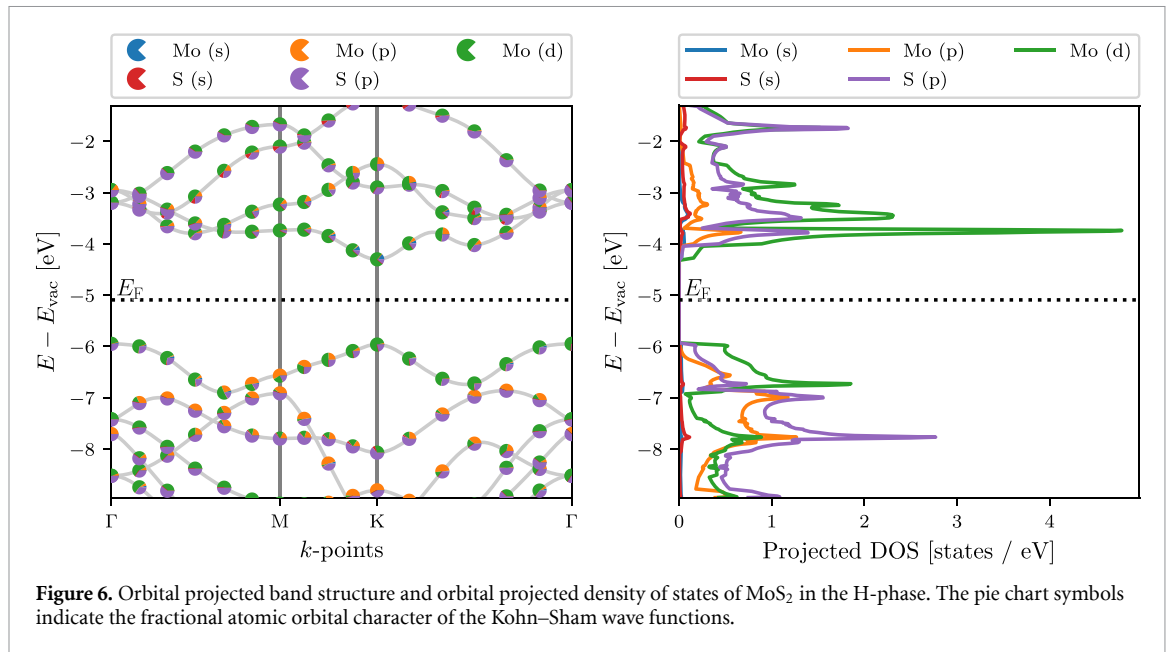
The second improvement is the calculation of the mean absolute relative error (MARE) of the polynomial fit in a 25 meV range from the BE. The value of 25 meV corresponds to the thermal energy at room temperature and is thus the relevant energy scale for many applications. To make the MARE independent of the absolute position of the band we calculate the average energy of the band over the 25 meV and compare the deviation of the fit to this energy scale. The MARE provides a useful measure of the parabolicity

of the energy bands and thus the validity of the effective mass approximation over this energy scale.

Figure 5 shows two examples of band structures with the effective mass fits and corresponding fit errors indicated. Additionally, the distribution of MARE for all the effective mass fits in the C2DB are presented. Most materials have an insignificant MARE, but a few materials have very large errors. Materials with a MARE above a few hundreds of percentages fall into two classes. For some materials the algorithm does not correctly find the position of the BE. An example is Ti_2S_2 in the space group $C2/m$. For others, the fit and BE location are both correct, but the band flattens away from the BE which leads to a large MARE as is the case for Rh_2Br_6 shown in the figure or Cl_2Tl_2 in the space group $P-1$. In general a small MARE indicates a parabolic band while materials with large MARE should be handled on a case-by-case basis.

3.3. Orbital projected band structure

To facilitate a state-specific analysis of the PBE Kohn-Sham wave functions, an orbital projected band structure (PBS) is provided to complement the projected density of states (PDOS). In the PAW methodology, the all-electron wave functions are projected onto atomic orbitals inside the augmentation spheres centred at the position of each atom. The PBS resolves these atomic orbital contributions to the



wave functions as a function of band and k -point whereas the PDOS resolves the atomic orbital character of the total density of states as a function of energy. The SOC is not included in the PBS or PDOS, as its effect is separately visualised by the spin-PBS also available in the C2DB.

As an example, figure 6 shows the PBS (left) and PDOS (right) of monolayer MoS₂ calculated with PBE. The relative orbital contribution to a given Bloch state is indicated by a pie chart symbol. In the present example, one can deduce from the PBS that even though Mo- p orbitals and S- p orbitals contribute roughly equally to the DOS in the valence band, the Mo- p orbital contributions are localised to a region in the BZ around the M -point, whereas the S- p orbitals contribute throughout the entire BZ.

3.4. Corrected G_0W_0 band structures

The C2DB contains G_0W_0 quasiparticle (QP) band structures of 370 monolayers covering 14 different crystal structures and 52 chemical elements. The details of these calculations can be found in the original C2DB paper [6]. A recent in-depth analysis of the 61.716 G_0W_0 data points making up the QP band structures led to several important conclusions relevant for high-throughput G_0W_0 calculations. In particular, it identified the linear QP approximation as a significant error source in standard G_0W_0 calculations and proposed an extremely simple correction scheme (the *empirical Z* (empZ) scheme), that reduces this error by a factor of two on average.

The empZ scheme divides the electronic states into two classes according to the size of the QP weight, Z . States with $Z \in [0.5, 1.0]$ are classified as QP consistent (QP-c) while states with $Z \notin [0.5, 1.0]$ are classified as QP inconsistent (QP-ic). With this definition, QP-c states will have at least half of their spectral weight in the QP peak. The distribution of

the 60.000+ Z -values is shown in figure 7. It turns out that the linear approximation to the self-energy, which is the gist of the QP approximation, introduces significantly larger errors for QP-ic states than for QP-c states. Consequently, the empZ method replaces the calculated Z of QP-ic states with the mean of the Z -distribution, $Z_0 \approx 0.75$. This simple replacement reduces the average error of the linear approximation from 0.11 to 0.06 eV.

An illustration of the method applied to MoS₂ is shown in figure 7. The original uncorrected G_0W_0 band structure is shown in blue while the empZ corrected band structure is shown in orange. MoS₂ has only one QP-ic state in the third conduction band at the K -point. Due to a break-down of the QP approximation for this state, the G_0W_0 correction is greatly overestimated leading to a local discontinuity in the band structure. The replacement of Z by Z_0 for this particular state resolves the problem. All G_0W_0 band structures in the C2DB are now empZ corrected.

3.5. Optical absorbance

In the first version of the C2DB, the optical absorbance was obtained from the simple expression [6]

$$A(\omega) \approx \frac{\omega \text{Im}\alpha^{2D}(\omega)}{\epsilon_0 c}, \quad (2)$$

where α^{2D} is the long wavelength limit of the in-plane sheet polarisability density (note that the equation is written here in SI units). The sheet polarisability is related to the sheet conductivity via $\sigma^{2D}(\omega) = -i\omega\alpha^{2D}(\omega)$. The expression (2) assumes that the electric field inside the layer equals the incoming field (i.e. reflection is ignored), and hence, it may overestimate the absorbance.

In the new version, the absorbance is evaluated from $A = 1 - R - T$, where R and T are the reflected and transmitted powers of a plane wave at normal

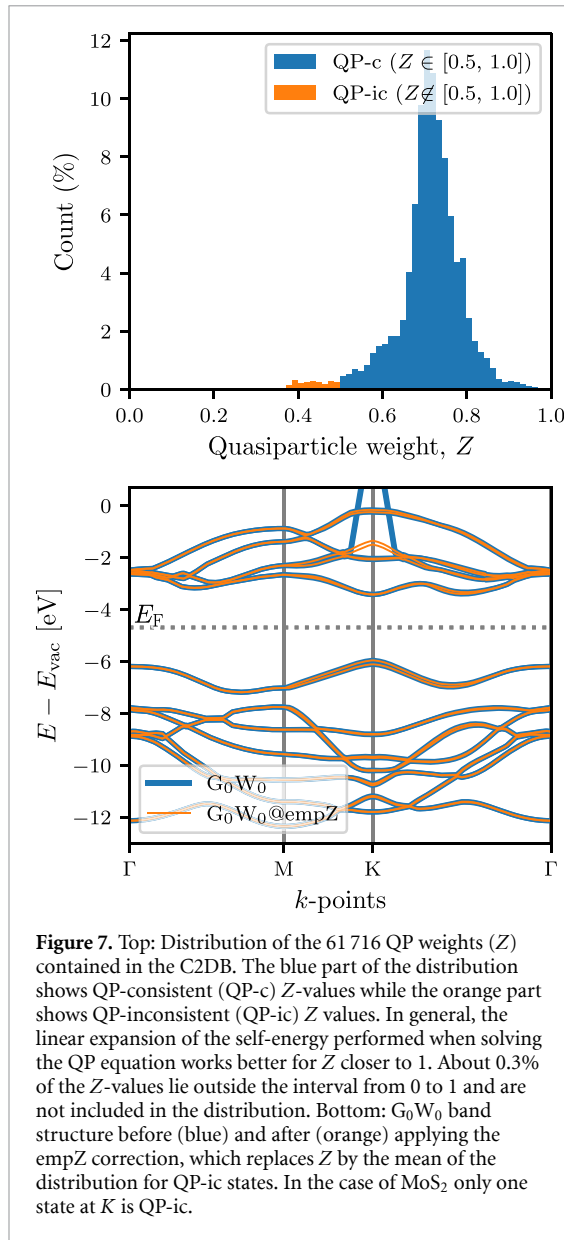


Figure 7. Top: Distribution of the 61 716 QP weights (Z) contained in the C2DB. The blue part of the distribution shows QP-consistent (QP-c) Z -values while the orange part shows QP-inconsistent (QP-ic) Z values. In general, the linear expansion of the self-energy performed when solving the QP equation works better for Z closer to 1. About 0.3% of the Z -values lie outside the interval from 0 to 1 and are not included in the distribution. Bottom: G_0W_0 band structure before (blue) and after (orange) applying the empZ correction, which replaces Z by the mean of the distribution for QP-ic states. In the case of MoS_2 only one state at K is QP-ic.

incidence, respectively. These can be obtained from the conventional transfer matrix method applied to a monolayer suspended in vacuum. The 2D material is here modelled as an infinitely thin layer with a sheet conductivity. Alternatively, it can be modelled as quasi-2D material of thickness d with a ‘bulk’ conductivity of $\sigma = \sigma^{2D}/d$ [58], but the two approaches yield very similar results, since the optical thickness of a 2D material is much smaller than the optical wavelength. Within this model, the expression for the absorbance of a suspended monolayer with the sheet conductivity σ^{2D} reads:

$$A(\omega) = \text{Re} \left\{ \sigma^{2D}(\omega) \eta_0 \right\} \left| \frac{2}{2 + \sigma^{2D}(\omega) \eta_0} \right|^2, \quad (3)$$

where $\eta_0 = 1/(\epsilon_0 c) \approx 377 \Omega$ is the vacuum impedance.

If the light–matter interaction is weak, i.e. $|\sigma^{2D} \eta_0| \ll 1$, equation (3) reduces to equation (2).

Nonetheless, due to the strong light–matter interaction in some 2D materials, this approximation is not reliable in general. In fact, it can be shown that the maximum possible absorption from equation (3) is 50%, which is known as the upper limit of light absorption in thin films [59]. This limit is not guaranteed by equation (2), which can even yield an absorbance above 100%.

As an example, figure 8 shows the absorption spectrum of monolayer MoS_2 for in- and out-of-plane polarised light as calculated with the exact equation (3) and the approximate equation (2), respectively. In all cases the sheet polarisability is obtained from the BSE to account for excitonic effects [6]. For weak light–matter interactions, e.g. for the z -polarised light, the two approaches agree quite well, but noticeable differences are observed in regions with stronger light–matter interaction.

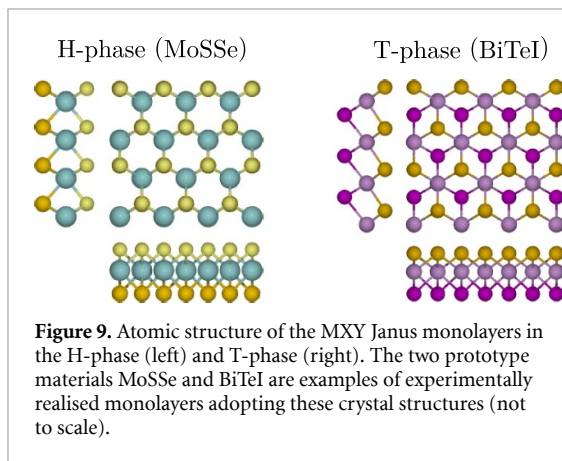
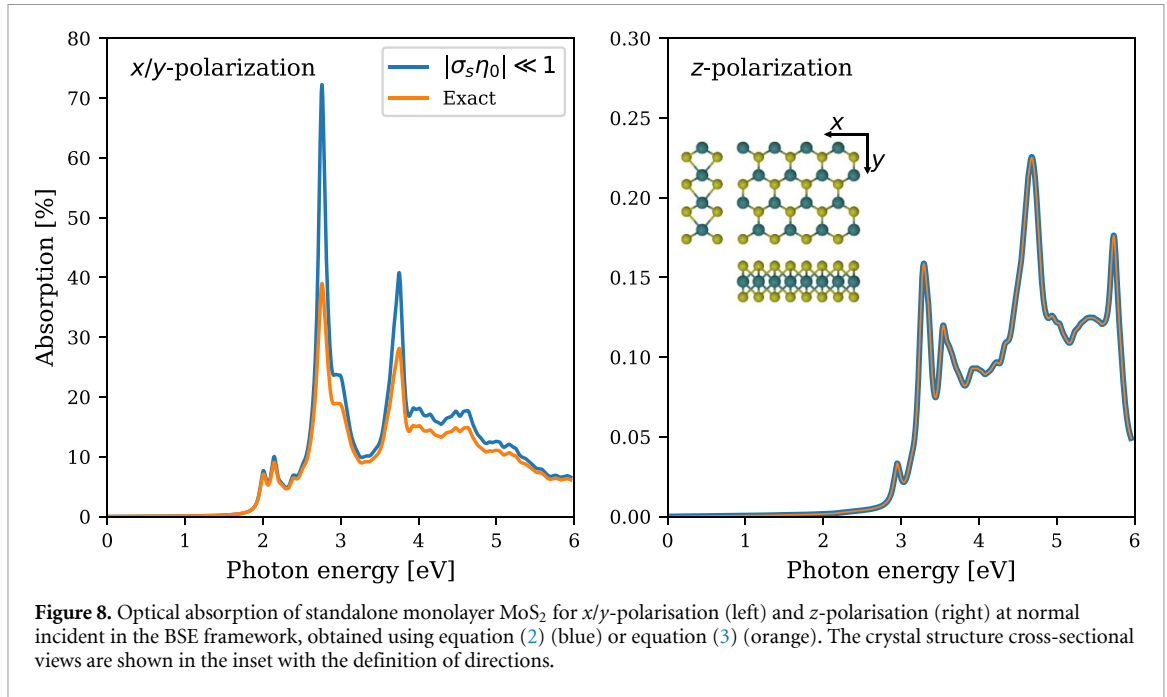
4. New materials in the C2DB

In this section we discuss the most significant extensions of the C2DB in terms of new materials. The set of materials presented here is not complete, but represents the most important and/or well defined classes. The materials discussed in sections 4.1 and 4.2 (MXY Janus monolayers and monolayers extracted from experimental crystal structure databases) are already included in the C2DB. The materials described in sections 4.3 and 4.4 (homo-bilayers and monolayer point defect systems) will soon become available as separate C2DB-interlinked databases.

4.1. MXY Janus monolayers

The class of TMDC monolayers of the type MX_2 (where M is the transition metal and X is a chalcogen) exhibits a large variety of interesting and unique properties and has been widely discussed in the literature [60]. Recent experiments have shown that it is not only possible to synthesise different materials by changing the metal M or the chalcogen X , but also by exchanging the X on one side of the layer by another chalcogen (or halogen) [61–63]. This results in a class of 2D materials known as MXY Janus monolayers with broken mirror symmetry and finite out-of-plane dipole moments. The prototypical MXY crystal structures are shown in figure 9 for the case of MoSSe and BiTeI , which have both been experimentally realised [61–63]. Adopting the nomenclature from the TMDCs, the crystal structures are denoted as H- or T-phase, depending on whether X and Y atoms are vertically aligned or displaced, respectively.

In a recent work [64], the C2DB workflow was employed to scrutinise and classify the basic electronic and optical properties of 224 different MXY Janus monolayers. All data from the study is available in the C2DB. Here we provide a brief discussion of the Rashba physics in these materials and refer the



interested reader to [64] for more details and analysis of other properties.

A key issue when considering hypothetical materials, i.e. materials not previously synthesised, is their stability. The experimentally synthesised MoSSe and BiTeI are both found to be dynamically stable and lie within 10 meV of the convex hull confirming their thermodynamic stability. Out of the 224 initial monolayers 93 are classified as stable according to the C2DB criteria (dynamically stable and $\Delta H_{\text{hull}} < 0.2 \text{ eV atom}^{-1}$). Out of the 93 stable materials, 70 exhibit a finite band gap when computed with the PBE xc-functional.

The Rashba effect is a momentum dependent splitting of the band energies of a 2D semiconductor in the vicinity of a band extremum arising due to the combined effect of spin–orbit interactions and a broken crystal symmetry in the direction perpendicular to the 2D plane. The simplest model used to describe the Rashba effect is a 2D electron gas in a perpendicular electric field (along the *z*-axis). Close to

the band extremum, the energy of the two spin bands is described by the Rashba Hamiltonian [65, 66]:

$$H = \alpha_R (\boldsymbol{\sigma} \times \mathbf{k}) \cdot \hat{\mathbf{e}}_z, \quad (4)$$

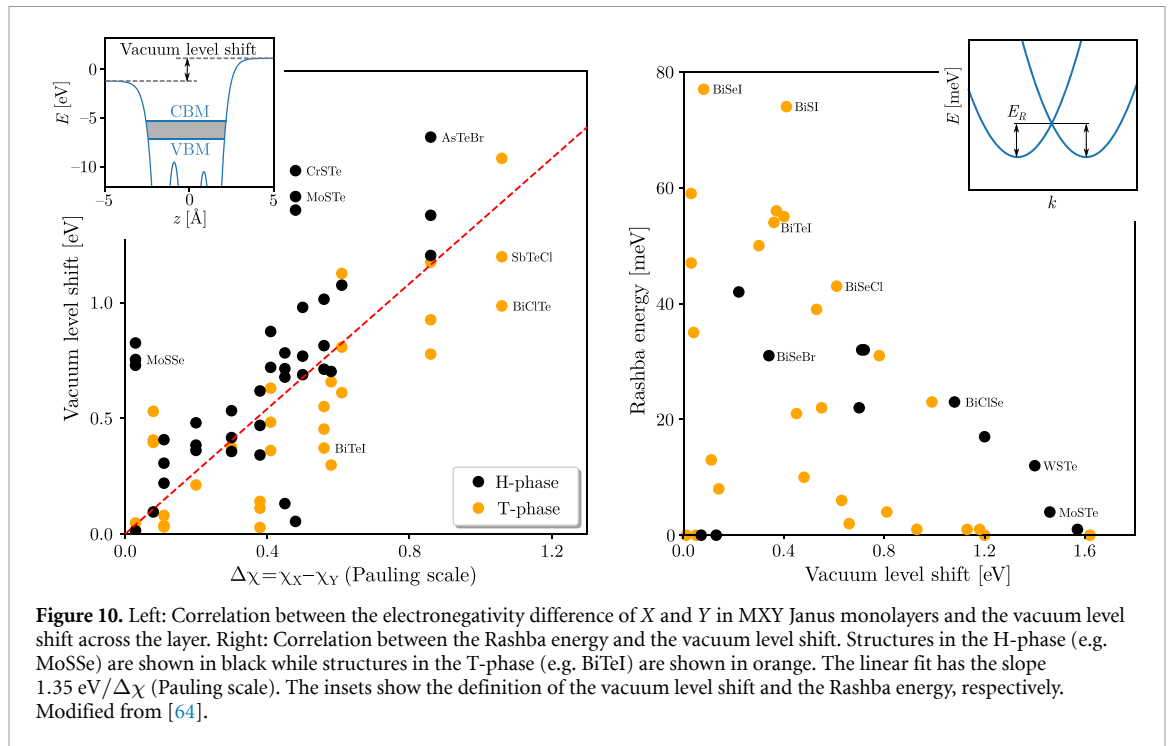
where $\boldsymbol{\sigma}$ is the vector of Pauli matrices, $\mathbf{k} = \mathbf{p}/\hbar$ is the wave number, and the Rashba parameter is proportional to the electric field strength, $\alpha_R \propto E_0$.

Although the Rashba Hamiltonian is only meant as a qualitative model, it is of interest to test its validity on the Janus monolayers. The electric field of the Rashba model is approximately given by $E_0 = \Delta V_{\text{vac}}/d$, where ΔV_{vac} is the shift in vacuum potential on the two sides of the layer (see left inset of figure 10) and d is the layer thickness. Assuming a similar thickness for all monolayers, the electric field is proportional to the potential shift. Not unexpected, the latter is found to correlate strongly with the difference in electronegativity of the X and Y atoms, see left panel of figure 10.

The Rashba energy, E_R , can be found by fitting $E(k) = \hbar^2 k^2 / 2m^* \pm \alpha_R k$ to the band structure (see right inset of figure 10) and should scale with the electric field strength. However, as seen from the right panel of figure 10, there is no correlation between the two quantities. Hence we conclude that the simple Rashba model is completely inadequate and that the strength of the perpendicular electric field cannot be used to quantify the effect of spin–orbit interactions on band energies.

4.2. Monolayers from known layered bulk crystals

The C2DB has been extended with a number of monolayers that are likely exfoliable from experimentally known layered bulk compounds. Specifically, the Inorganic Crystal Structure Database (ICSD) [67] and Crystallography Open Database (COD) [68]



have first been filtered for corrupted, duplicate and theoretical compounds, which reduce the initial set of 585.485 database entries to 167.767 unique materials. All of these have subsequently been assigned a ‘dimensionality score’ based on a purely geometrical descriptor. If the 2D score is larger than the sum of 0D, 1D and 3D scores we regard the material as being exfoliable and we extract the individual 2D components that comprise the material (see also section 2.2). We refer to the original work on the method for details [52] and note that similar approaches were applied in [11, 12] to identify potentially exfoliable monolayers from the ICSD and COD.

The search has been limited to bulk compounds containing less than six different elements and no rare earth elements. This reduces the set of relevant bulk materials to 2991. For all of these we extracted the 2D components containing less than 21 atoms in the unit cell, which were then relaxed and sorted for duplicates following the general C2DB workflow steps described in sections 2.1–2.3. At this point 781 materials remain. This set includes most known 2D materials and 207 of the 781 were already present in the C2DB prior to this addition. All the materials (including those that were already in C2DB) have been assigned an ICSD/COD identifier that refers to the parent bulk compound from which the 2D material was computationally exfoliated. We emphasise that we have not considered exfoliation energies in the analysis and a subset of these materials may thus be rather strongly bound and challenging to exfoliate even if the geometries indicate van der Waals bonded structures of the parent bulk compounds.

Figure 11 shows the distribution of energies above the convex hull for materials derived from

parent structures in ICSD or COD as well as for the entire C2DB, which includes materials obtained from combinatorial lattice decoration as well. As expected, the materials derived from experimental bulk materials are situated rather close to the convex hull whereas those obtained from lattice decoration extend to energies far above the convex hull. It is also observed that a larger fraction of the experimentally derived materials are dynamically stable. There are, however, well known examples of van der Waals bonded structures where the monolayer undergoes a significant lattice distortion, which will manifest itself as a dynamical instability in the present context. For example, bulk MoS₂ exists in van der Waals bonded structures composed of either 2 H-MoS₂ or 1 T-MoS₂ layers, but a monolayer of the 1 T phase undergoes a structural deformation involving a doubling of the unit cell [69] and is thus categorised as dynamically unstable by the C2DB workflow. The dynamically stable materials derived from parent bulk structures in the ICSD and COD may serve as a useful subset of the C2DB that are likely to be exfoliable from known compounds and thus facilitate experimental verification. As a first application the subset has been used to search for magnetic 2D materials, which resulted in a total of 85 ferromagnets and 61 anti-ferromagnets [70].

4.3. Outlook: multilayers

The C2DB is concerned with the properties of covalently bonded monolayers (see discussion of dimensionality filtering in section 2.2). However, multilayer structures composed of two or more identical monolayers are equally interesting and often have properties that deviate from those of the monolayer. In fact, the synthesis of layered vdW structures with a

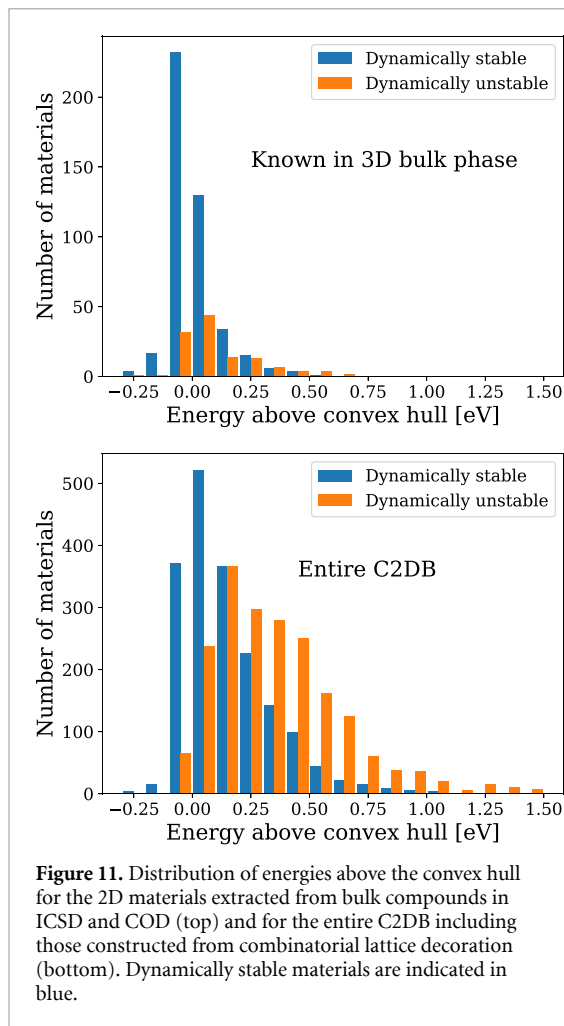


Figure 11. Distribution of energies above the convex hull for the 2D materials extracted from bulk compounds in ICSD and COD (top) and for the entire C2DB including those constructed from combinatorial lattice decoration (bottom). Dynamically stable materials are indicated in blue.

controllable number of layers represents an interesting avenue for atomic-scale materials design. Several examples of novel phenomena emerging in layered vdW structures have been demonstrated including direct-indirect band gap transitions in MoS₂ [71, 72], layer-parity selective Berry curvatures in few-layer WTe₂ [73], thickness-dependent magnetic order in CrI₃ [74, 75], and emergent ferroelectricity in bilayer hBN [76].

As a first step towards a systematic exploration of multilayer 2D structures, the C2DB has been used as basis for generating homobilayers in various stacking configurations and subsequently computing their properties following a modified version of the C2DB monolayer workflow. Specifically, the most stable monolayers (around 1000) are combined into bilayers by applying all possible transformations (unit cell preserving point group operations and translations) of one layer while keeping the other fixed. The candidate bilayers generated in this way are subject to a stability analysis, which evaluates the binding energy and optimal IL distance based on PBE-D3 [77] total energy calculations keeping the atoms of the monolayers fixed in their PBE relaxed geometry, see figures 12 and table 3.

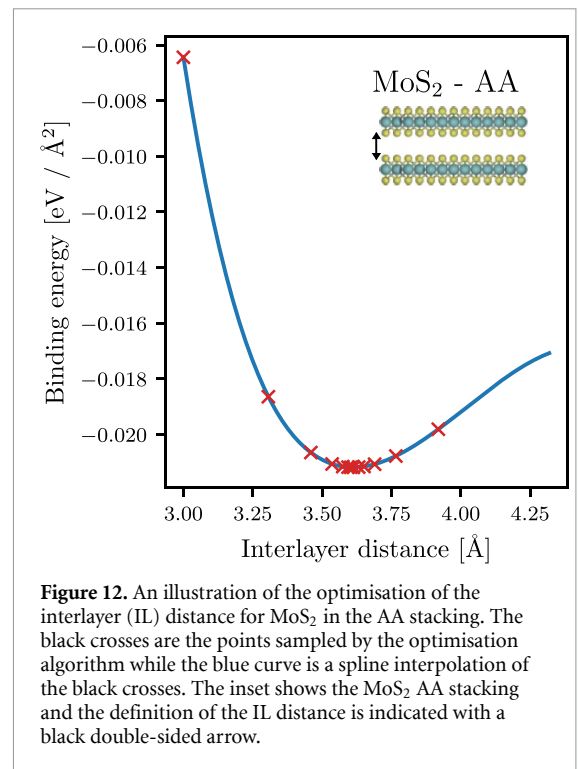


Figure 12. An illustration of the optimisation of the interlayer (IL) distance for MoS₂ in the AA stacking. The black crosses are the points sampled by the optimisation algorithm while the blue curve is a spline interpolation of the black crosses. The inset shows the MoS₂ AA stacking and the definition of the IL distance is indicated with a black double-sided arrow.

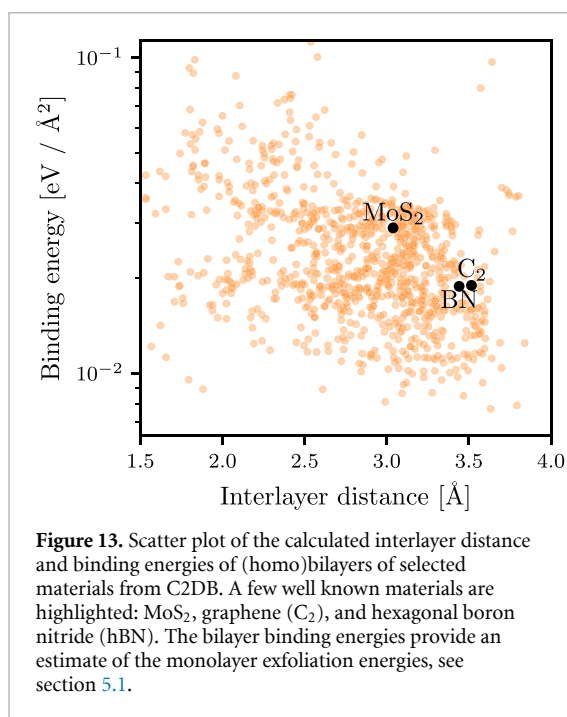
Table 3. Exfoliation energies for selected materials calculated with the PBE+D3 xc-functional as described in section 4.3 and compared with the DF2 and rVV10 results from [11]. The spacegroups are indicated in the column ‘SG’. All numbers are in units of meV Å⁻².

Material	SG	PBE + D3	DF2	rVV10
MoS ₂	P-6m2	28.9	21.6	28.8
MoTe ₂	P-6m2	30.3	25.2	30.4
ZrNBr	Pmmn	18.5	10.5	18.5
C	P6/mmm	18.9	20.3	25.5
P	Pmna	21.9	38.4	30.7
BN	P-6m2	18.9	19.4	24.4
WTe ₂	P-6m2	32.0	24.7	30.0
PbTe	P3m1	23.2	27.5	33.0

The calculated IL binding energies are generally in the range from a few to a hundred meV Å⁻² and IL distances range from 1.5 to 3.8 Å. A scatter plot of preliminary binding energies and IL distances is shown in figure 13. The analysis of homobilayers provides an estimate of the energy required to peel a monolayer off a bulk structure. In particular, the binding energy for the most stable bilayer configuration provides a measure of the *exfoliation energy* of the monolayer. This key quantity is now available for all monolayers in the C2DB, see section 5.1.

4.4. Outlook: point defects

The C2DB is concerned with the properties of 2D materials in their pristine crystalline form. However, as is well known the perfect crystal is an idealised model of real materials, which always contain defects in smaller or larger amounts depending on the intrinsic materials properties and growth conditions. Crystal defects often have a negative impact on



physical properties, e.g. they lead to scattering and life time-reduction of charge carriers in semiconductors. However, there are also important situations where defects play a positive enabling role, e.g. in doping of semiconductors, as colour centres for photon emission [78, 79] or as active sites in catalysis.

To reduce the gap between the pristine model material and real experimentally accessible samples, a systematic evaluation of the basic properties of the simplest native point defects in a selected subset of monolayers from the C2DB has been initiated. The monolayers are selected based on the stability of the pristine crystal. Moreover, only non-magnetic semiconductors with a PBE band gap satisfying $E_{\text{gap}} > 1$ eV are currently considered as such materials are candidates for quantum technology applications like single-photon sources and spin qubits. Following these selection criteria around 300 monolayers are identified and their vacancies and intrinsic substitutional defects are considered, yielding a total of about 1500 defect systems.

Each defect system is subject to the same workflow, which is briefly outlined below. To enable point defects to relax into their lowest energy configuration, the symmetry of the pristine host crystal is intentionally broken by the chosen supercell, see figure 14 (a). In order to minimise defect–defect interaction, supercells are furthermore chosen such that the minimum distance between periodic images of defects is larger than 15 Å. Unique point defects are created based on the analysis of equivalent Wyckoff positions for the host material. To illustrate some of the properties that will feature in the upcoming point defect database, we consider the specific example of monolayer CH₂Si.

First, the formation energy [80, 81] of a given defect is calculated from PBE total energies. Next,

Slater–Janak transition state theory is used to obtain the charge transition levels [82, 83]. By combining these results, one obtains the formation energy of the defect in all possible charge states as a function of the Fermi level. An example of such a diagram is shown in figure 14 (b) for the case of the V_C and C_{Si} defects in monolayer CH₂Si. For each defect and each charge state, the PBE single-particle energy level diagram is calculated to provide a qualitative overview of the electronic structure. A symmetry analysis [84] is performed for the defect structure and the individual defect states lying inside the band gap. The energy level diagram of the neutral V_{Si} defect in CH₂Si is shown in figure 14 (c), where the defect states are labelled according to the irreducible representations of the C_s point group.

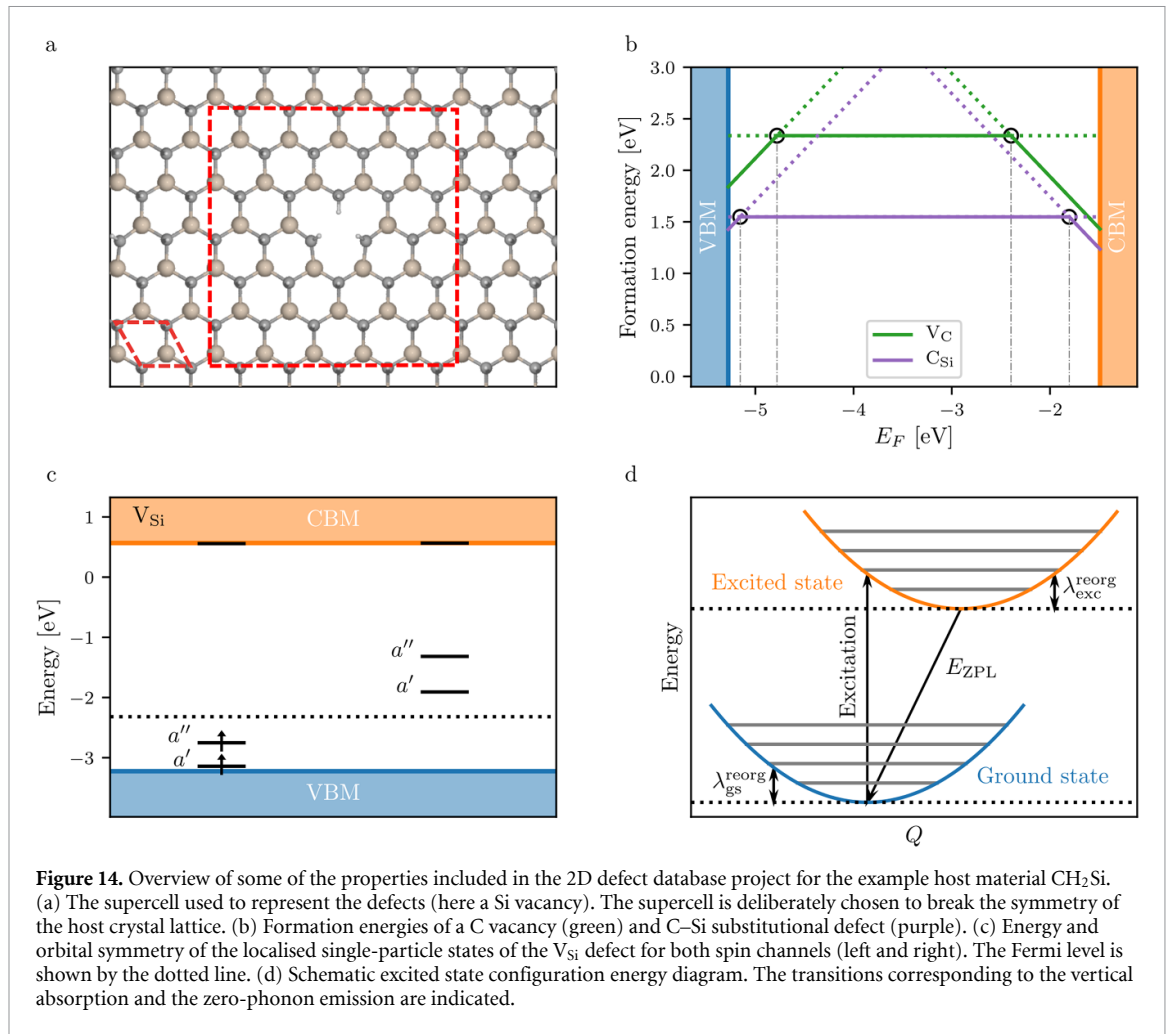
In general, excited electronic states can be modelled by solving the Kohn–Sham equations with non-Aufbau occupations. The excited-state solutions are saddle points of the Kohn–Sham energy functional, but common self-consistent field (SCF) approaches often struggle to find such solutions, especially when nearly degenerate states are involved. The calculation of excited states corresponding to transitions between localised states inside the band gap is therefore performed using an alternative method based on the direct optimisation (DO) of orbital rotations in combination with the maximum overlap method (MOM) [85]. This method ensures fast and robust convergence of the excited states, as compared to SCF. In figure 14 (d), the reorganisation energies for the ground and excited state, as well as the zero-phonon line (ZPL) energy are sketched. For the specific case of the Si vacancy in CH₂Si, the DO-MOM method yields $E_{\text{ZPL}} = 3.84$ eV, $\lambda_{\text{gs}}^{\text{reorg}} = 0.11$ eV and $\lambda_{\text{exc}}^{\text{reorg}} = 0.16$ eV. For systems with large electron–phonon coupling (i.e. Huang–Rhys factor > 1) a one-dimensional approximation for displacements along the main phonon mode is used to produce the configuration coordinate diagram (see figure 14 (d)). In addition to the ZPL energies and reorganisation energies, the Huang–Rhys factors, photoluminescence spectrum from the 1D phonon model, hyperfine coupling and zero field splitting are calculated.

5. New properties in the C2DB

This section reports on new properties that have become available in the C2DB since the first release. The employed computational methodology is described in some detail and results are compared to the literature where relevant. In addition, some interesting property correlations are considered along with general discussions of the general significance and potential application of the available data.

5.1. Exfoliation energy

The exfoliation energy of a monolayer is estimated as the binding energy of its bilayer in the most stable



stacking configuration (see also section 4.3). The binding energy is calculated using the PBE + D3 xc-functional [86] with the atoms of both monolayers fixed in the PBE relaxed geometry. Table 3 compares exfoliation energies obtained in this way to values from Mounet *et al* [11] for a representative set of monolayers.

5.2. Bader charges

For all monolayers we calculate the net charge on the individual atoms using the Bader partitioning scheme [87]. The analysis is based purely on the electron density, which we calculate from the PAW pseudo density plus compensation charges using the PBE xc-functional. Details of the method and its implementation can be found in Tang *et al* [88]. In section 5.4 we compare and discuss the relation between Bader charges and Born charges.

5.3. Spontaneous polarisation

The spontaneous polarisation (\mathbf{P}_s) of a bulk material is defined as the charge displacement with respect to that of a reference centrosymmetric structure [89, 90]. Ferroelectric materials exhibit a finite value

of \mathbf{P}_s that may be switched by an applied external field and have attracted a large interest for a wide range of applications [91–93].

The spontaneous polarisation in bulk materials can be regarded as electric dipole moment per unit volume, but in contrast to the case of finite systems this quantity is ill-defined for periodic crystals [89]. Nevertheless, one can define the formal polarisation density:

$$\mathbf{P} = \frac{1}{2\pi} \frac{e}{V} \sum_l \phi_l \mathbf{a}_l, \quad (5)$$

where \mathbf{a}_l (with $l \in \{1, 2, 3\}$) are the lattice vectors spanning the unit cell, V is the cell volume and e is the elementary charge. ϕ_l is the polarisation phase along the lattice vector defined by:

$$\phi_l = \sum_i Z_i \mathbf{b}_l \cdot \mathbf{u}_i - \phi_l^{\text{elec}}, \quad (6)$$

where \mathbf{b}_l is the reciprocal lattice vector satisfying $\mathbf{b}_l \cdot \mathbf{R}_l = 2\pi$ and \mathbf{u}_i is the position of nucleus i with charge eZ_i . The electronic contribution to the polarisation phase is defined as:

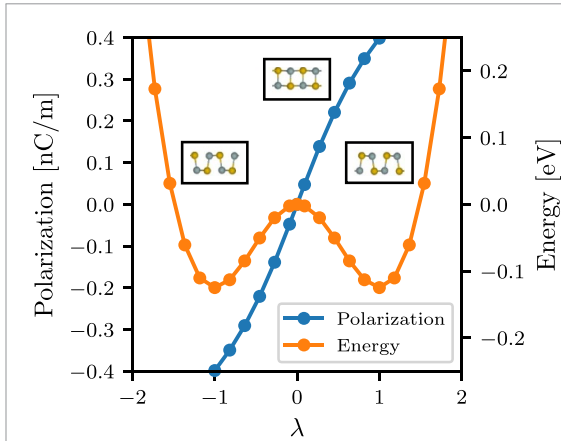


Figure 15. Depicted in the blue plot is the formal polarisation calculated along the adiabatic path for GeSe, using the methods described in the main text. The orange plot shows the energy potential along the path as well as outside. Figure inset: The structure of GeSe in the two non-centrosymmetric configurations corresponding to $-\mathbf{P}_s$ and \mathbf{P}_s and the centrosymmetric configuration.

$$\phi_i^{\text{elec}} = \frac{1}{N_{k \perp \mathbf{b}_i}} \text{Im} \sum_{k \in \text{BZ}_{\perp \mathbf{b}_i}} \times \ln \prod_{j=0}^{N_{k \parallel \mathbf{b}_i} - 1} \det_{\text{occ}} [\langle u_{n\mathbf{k}+j\delta\mathbf{k}} | u_{m\mathbf{k}+(j+1)\delta\mathbf{k}} \rangle], \quad (7)$$

where $\text{BZ}_{\perp \mathbf{b}_i} = \{\mathbf{k} | \mathbf{k} \cdot \mathbf{b}_i = 0\}$ is a plane of \mathbf{k} -points orthogonal to \mathbf{b}_i , $\delta\mathbf{k}$ is the distance between neighbouring \mathbf{k} -points in the \mathbf{b}_i direction and $N_{k \parallel \mathbf{b}_i}$ ($N_{k \perp \mathbf{b}_i}$) is the number of \mathbf{k} -points along (perpendicular to) the \mathbf{b}_i direction. These expressions generalise straightforwardly to 2D.

The formal polarisation is only well-defined modulo $e\mathbf{R}_n/V$ where \mathbf{R}_n is any lattice vector. However, changes in polarisation are well defined and the spontaneous polarisation may thus be obtained by:

$$\mathbf{P}_s = \int_0^1 \frac{d\mathbf{P}(\lambda)}{d\lambda} d\lambda, \quad (8)$$

where λ is a dimensionless parameter that defines an adiabatic structural path connecting the polar phase ($\lambda = 1$) with a non-polar phase ($\lambda = 0$).

The methodology has been implemented in GPAW and used to calculate the spontaneous polarisation of all stable materials in the C2DB with a PBE band gap above 0.01 eV and a polar space group symmetry. For each material, the centrosymmetric phase with smallest atomic displacement from the polar phase is constructed and relaxed under the constraint of inversion symmetry. The adiabatic path connecting the two phases is then used to calculate the spontaneous polarisation using equations (5)–(8). An example of a calculation for GeSe is shown in figure 15 where the polarisation along the path connecting two equivalent polar phases via the centrosymmetric phase is shown together with the total energy. The

spontaneous polarisation obtained from the path is 39.8 nC m^{-1} in good agreement with previous calculations [94].

5.4. Born charges

The Born charge of an atom a at position \mathbf{u}_a in a solid is defined as:

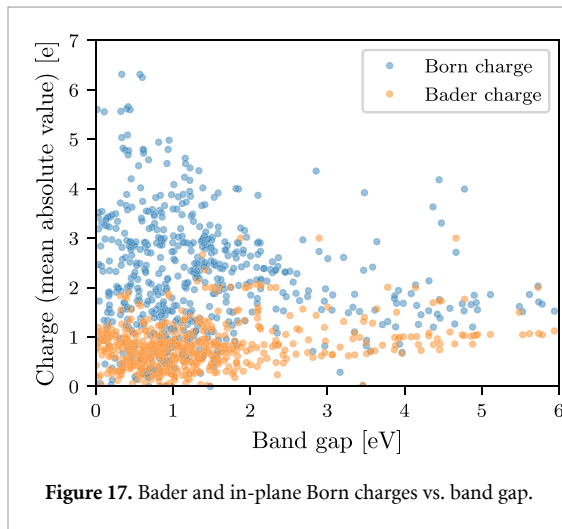
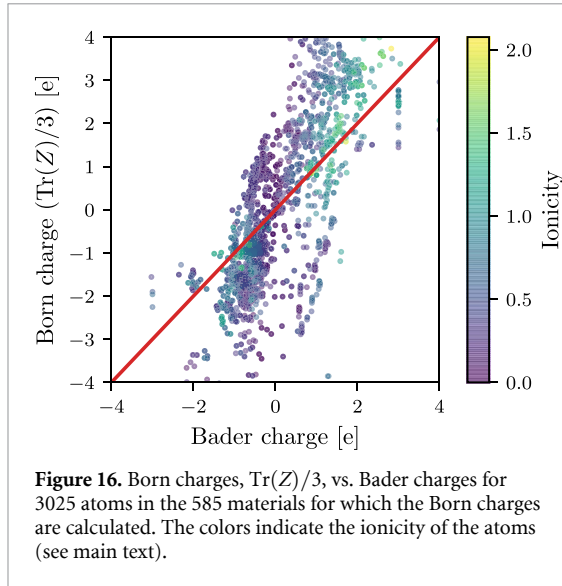
$$Z_{ij}^a = \left. \frac{V}{e} \frac{\partial P_i}{\partial u_{aj}} \right|_{E=0}. \quad (9)$$

It can be understood as an effective charge assigned to the atom to match the change in polarisation in direction i when its position is perturbed in direction j . Since the polarisation density and the atomic position are both vectors, the Born charge of an atom is a rank-2 tensor. The Born charge is calculated as a finite difference and relies on the Modern theory of polarisation [95] for the calculation of polarisation densities, see reference [96] for more details. The Born charge has been calculated for all stable materials in C2DB with a finite PBE band gap.

It is of interest to examine the relation between the Born charge and the Bader charge (see section 5.2). In materials with strong ionic bonds one would expect the charges to follow the atoms. On the other hand, in covalently bonded materials the hybridisation pattern and thus the charge distribution, depends on the atom positions in a complex way, and the idea of charges following the atom is expected to break down. In agreement with this idea, the (in-plane) Born charges in the strongly ionic hexagonal hBN ($\pm 2.71e$ for B and N, respectively) are in good agreement with the calculated Bader charges ($\pm 3.0e$). In contrast, (the in-plane) Born charges in MoS_2 ($-1.08e$ and $0.54e$ for Mo and S, respectively) deviate significantly from the Bader charges ($1.22e$ and $-0.61e$ for Mo and S, respectively). In fact, the values disagree even on the sign of the charges underlining the non-intuitive nature of the Born charges in covalently bonded materials.

Note that the out-of-plane Born charges never match the Bader charges, even for strongly ionic insulators, and are consistently smaller in value than the in-plane components. The smaller out-of-plane values are consistent with the generally smaller out-of-plane polarisability of 2D materials (for both electronic and phonon contributions) and agrees with the intuitive expectation that it is more difficult to polarise a 2D material in the out-of-plane direction as compared to the in-plane direction.

Figure 16 shows the average of the diagonal of the Born charge tensor, $\text{Tr}(Z^a)/3$, plotted against the Bader charges for all 585 materials in the C2DB for which the Born charges have been computed. The data points have been coloured according to the ionicity of the atom a defined as $I(a) = |\chi_a - \langle \chi \rangle|$, where χ_a and $\langle \chi \rangle$ are the Pauling electronegativity of atom a and the average electronegativity of all atoms in the

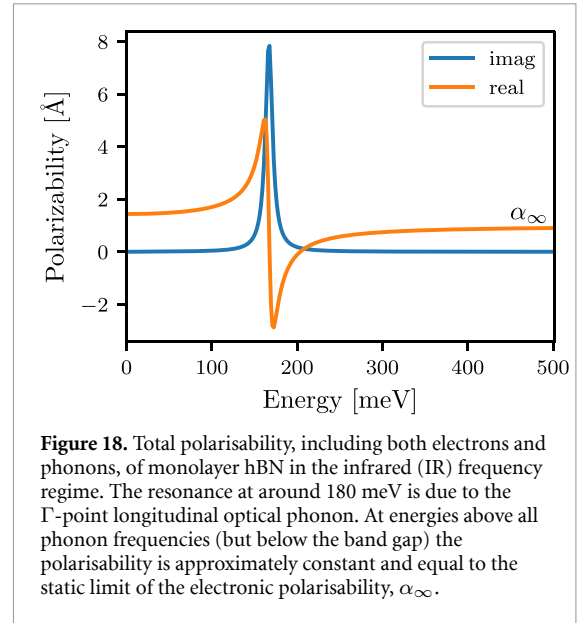


unit cell, respectively. The ionicity is thus a measure of the tendency of an atom to donate/accept charge relative to the average tendency of atoms in the material. It is clear from figure 16 that there is a larger propensity for the Born and Bader charges to match in materials with higher ionicity.

Figure 17 plots the average (in-plane) Born charge and the Bader charge versus the band gap. It is clear that large band gap materials typically exhibit integer Bader charges, whereas there is no clear correlation between the Born charge and the band gap.

5.5. Infrared polarisability

The original C2DB provided the frequency dependent polarisability computed in the random phase approximation (RPA) with inclusion of electronic interband and intraband (for metals) transitions [6]. However, phonons carrying a dipole moment (so-called IR active phonons) also contribute to the polarisability at frequencies comparable to the frequency of optical phonons. This response is described by the IR polarisability:



$$\alpha^{\text{IR}}(\omega) = \frac{e^2}{A} \mathbf{Z}^T \mathbf{M}^{-1/2} \left(\sum_i \frac{\mathbf{d}_i \mathbf{d}_i^T}{\omega_i^2 - \omega^2 - i\gamma\omega} \right) \mathbf{M}^{-1/2} \mathbf{Z}, \quad (10)$$

where \mathbf{Z} and \mathbf{M} are matrix representations of the Born charges and atomic masses, ω_i^2 and d_i are eigenvectors and eigenvalues of the dynamical matrix, A is the in-plane cell area and γ is a broadening parameter representing the phonon lifetime and is set to 10 meV. The total polarisability is then the sum of the electronic polarisability and the IR polarisability.

The new C2DB includes the IR polarisability of all monolayers for which the Born charges have been calculated (stable materials with a finite band gap), see section (5.4). As an example, figure 18 shows the total polarisability of monolayer hexagonal hBN. For details on the calculation of the IR polarisability see reference [96].

5.6. Piezoelectric tensor

The piezoelectric effect is the accumulation of charges, or equivalently the formation of an electric polarisation, in a material in response to an applied mechanical stress or strain. It is an important material characteristic with numerous scientific and technological applications in sonar, microphones, accelerometers, ultrasonic transducers, energy conversion, etc [97, 98]. The change in polarisation originates from the movement of positive and negative charge centres as the material is deformed.

Piezoelectricity can be described by the (proper) piezoelectric tensor c_{ijk} with $i, j, k \in \{x, y, z\}$, given by [99]:

$$c_{ijk} = \frac{e}{2\pi V} \sum_l \frac{\partial \phi_l}{\partial \epsilon_{jk}} a_{li}, \quad (11)$$

which differs from equation (5) only by a derivative of the polarisation phase with respect to the strain tensor

Table 4. Comparison of computed piezoelectric tensor versus experimental values and previous calculations for hexagonal BN and a selected set of TMDCs (space group 187). All numbers are in units of nC/m. Experimental data for MoS₂ is obtained from [102].

Material	Exp.	Theory [101]	C2DB
BN	—	0.14	0.13
MoS ₂	0.3	0.36	0.35
MoSe ₂	—	0.39	0.38
MoTe ₂	—	0.54	0.48
WS ₂	—	0.25	0.24
WSe ₂	—	0.27	0.26
WTe ₂	—	0.34	0.34

ϵ_{jk} . Note that c_{ijk} does not depend on the chosen branch cut.

The piezoelectric tensor is a symmetric tensor with at most 18 independent components. Furthermore, the point group symmetry restricts the number of independent tensor elements and their relationships due to the well-known Neumann's principle [100]. For example, monolayer MoS₂ with point group D_{3h} , has only one non-vanishing independent element of c_{ijk} . Note that c_{ijk} vanishes identically for centrosymmetric materials. Using a finite-difference technique with a finite but small strain (1% in our case), equation (11) has been used to compute the proper piezoelectric tensor for all non-centrosymmetric materials in the C2DB with a finite band gap. Table 4 shows a comparison of the piezoelectric tensors in the C2DB with literature for a selected set of monolayer materials. Good agreement is obtained for all these materials.

5.7. Topological invariants

For all materials in the C2DB exhibiting a direct band gap below 1 eV, the k -space Berry phase spectrum of the occupied bands has been calculated from the PBE wave functions. Specifically, a particular k -point is written as $k_1\mathbf{b}_1 + k_2\mathbf{b}_2$ and the Berry phases $\gamma_n(k_2)$ of the occupied states on the path $k_1 = 0 \rightarrow k_1 = 1$ is calculated for each value of k_2 . The connectivity of the Berry phase spectrum determines the topological properties of the 2D Bloch Hamiltonian [103, 104].

The calculated Berry phase spectra of the relevant materials are available for visual inspection on the C2DB webpage. Three different topological invariants have been extracted from these spectra and are reported in the C2DB: (1) The Chern number, C , takes an integer value and is well defined for any gapped 2D material. It determines the number of chiral edge states on any edge of the material. For any non-magnetic material the Chern number vanishes due to time-reversal symmetry. It is determined from the Berry phase spectrum as the number of crossings at any horizontal line in the spectrum. (2) The mirror Chern number, C_M , defined for gapped materials with a mirror plane in the atomic layer [105]. For such materials, all states may be chosen as mirror

eigenstates with eigenvalues $\pm i$ and the Chern numbers C_{\pm} can be defined for each mirror sector separately. For a material with vanishing Chern number, the mirror Chern number is defined as $C_M = (C_+ - C_-)/2$ and takes an integer value corresponding to the number of edge states on any mirror symmetry preserving edge. It is obtained from the Berry phase spectrum as the number of chiral crossings in each of the mirror sectors. (3) The Z_2 invariant, ν , which can take the values 0 and 1, is defined for materials with time-reversal symmetry. Materials with $\nu = 1$ are referred to as quantum spin Hall insulators and exhibit helical edge states at any time-reversal conserving edge. It is determined from the Berry phase spectrum as the number of crossing points modulus 2 at any horizontal line in the interval $k_2 \in [0, 1/2]$.

Figure 19 shows four representative Berry phase spectra corresponding to the three cases of non-vanishing C , C_M and ν as well as a trivial insulator. The four materials are: OsCl₃ (space group 147)—a Chern insulator with $C = 1$, OsTe₂ (space group 14)—a mirror crystalline insulator with $C_M = 2$, SbI (spacegroup 1)—a quantum spin Hall insulator with $\nu = 1$ and BiTe (spacegroup 156)—a trivial insulator. Note that a gap in the Berry phase spectrum always implies a trivial insulator.

In [106] the C2DB was screened for materials with non-trivial topology. At that point it was found that the database contained 7 Chern insulators, 21 mirror crystalline topological insulators and 48 quantum spin Hall insulators. However, that does not completely exhaust the the topological properties of materials in the C2DB. In particular, there may be materials that can be topologically classified based on crystalline symmetries other than the mirror plane of the layer. In addition, second order topological effects may be present in certain materials, which imply that flakes will exhibit topologically protected corner states. Again, the Berry phase spectra may be used to unravel the second order topology by means of nested Wilson loops [107].

5.8. Exchange coupling constants

The general C2DB workflow described in sections 2.1–2.3 will identify the FM ground state of a material and apply it as starting point for subsequent property calculations, whenever it is more stable than the spin-paired ground state. In reality, however, the FM state is not guaranteed to comprise the magnetic ground state. In fact, AFM states often have lower energy than the FM one, but in general it is non-trivial to obtain the true magnetic ground state. We have chosen to focus on the FM state due to its simplicity and because its atomic structure and stability are often very similar to those of other magnetic states. Whether or not the FM state is the true magnetic ground state is indicated by the nearest neighbour exchange coupling constant as described below.

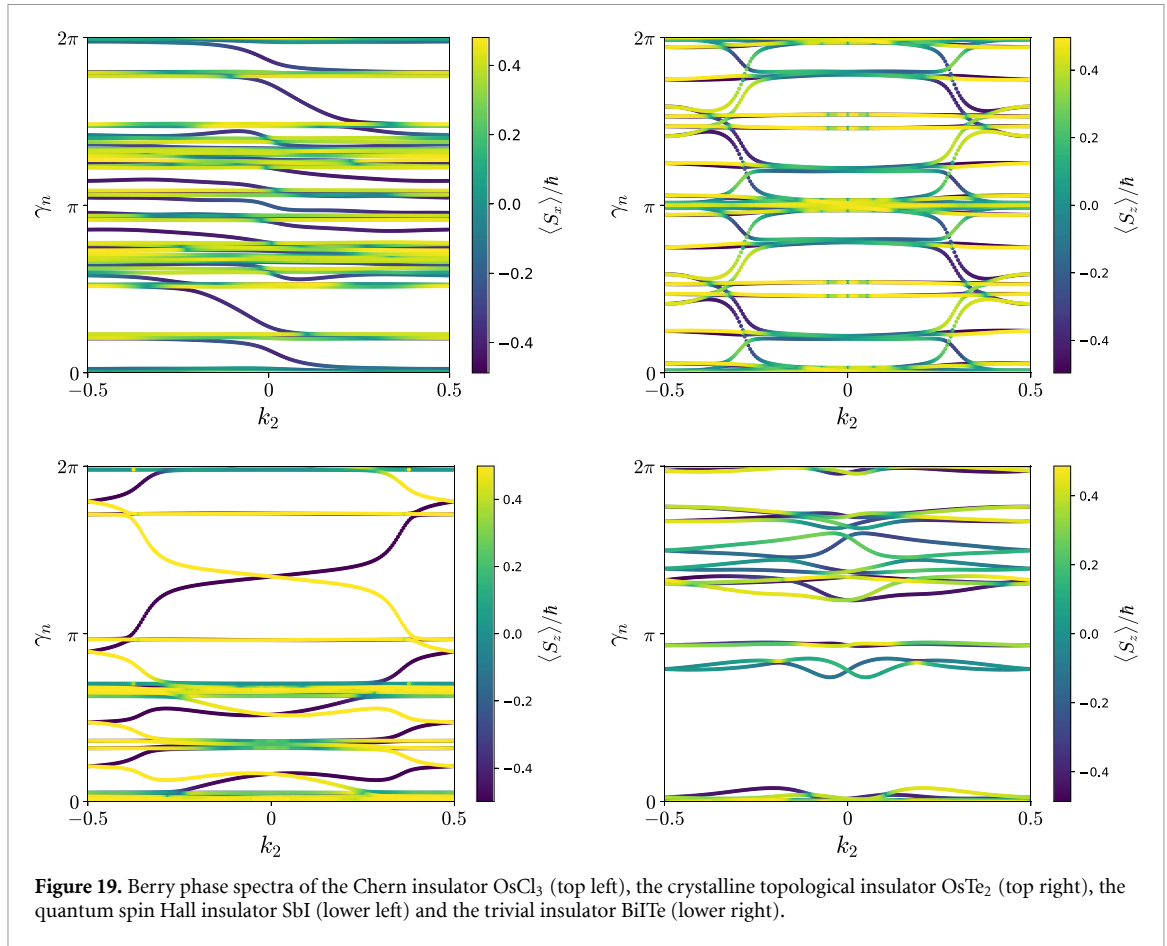


Figure 19. Berry phase spectra of the Chern insulator OsCl₃ (top left), the crystalline topological insulator OsTe₂ (top right), the quantum spin Hall insulator SbI (lower left) and the trivial insulator BiTe (lower right).

When investigating magnetic materials the thermodynamical properties (for example the critical temperatures for ordering) are of crucial interest. In two dimensions the Mermin–Wagner theorem [108] comprises an extreme example of the importance of thermal effects since it implies that magnetic order is only possible at $T = 0$ unless the spin-rotational symmetry is explicitly broken. The thermodynamic properties cannot be accessed directly by DFT. Consequently, magnetic models that capture the crucial features of magnetic interactions must be employed. For insulators, the Heisenberg model has proven highly successful in describing magnetic properties of solids in 3D as well as 2D [109]. It represents the magnetic degrees of freedom as a lattice of localised spins that interact through a set of exchange coupling constants. If the model is restricted to include only nearest neighbour exchange and assume magnetic isotropy in the plane, it reads:

$$H = -\frac{J}{2} \sum_{\langle ij \rangle} \mathbf{S}_i \cdot \mathbf{S}_j - \frac{\lambda}{2} \sum_{\langle ij \rangle} S_i^z S_j^z - A \sum_i (S_i^z)^2, \quad (12)$$

where J is the nearest neighbour exchange constant, λ is the nearest neighbour anisotropic exchange constant and A measures the strength of single-ion anisotropy. We also neglect off-diagonal exchange coupling constants that give rise to terms proportional to $S_i^x S_j^y$, $S_i^y S_j^z$ and $S_i^z S_j^x$. The out-of-plane direction has

been chosen as z and $\langle ij \rangle$ implies that for each site i we sum over all nearest neighbour sites j . The parameters J , λ and A may be obtained from an energy mapping analysis involving four DFT calculations with different spin configurations [70, 110, 111]. The thermodynamic properties of the resulting ‘first principles Heisenberg model’ may subsequently be analysed with classical Monte Carlo simulations or renormalised spin wave theory [36, 112].

The C2DB provides the values of J , λ , and A as well as the number of nearest neighbours N_{nn} and the maximum eigenvalue of S_z (S), which is obtained from the total magnetic moment per atom in the FM ground state (rounded to nearest half-integer for metals). These key parameters facilitate easy post-processing analysis of thermal effects on the magnetic structure. In [113] such an analysis was applied to estimate the critical temperature of all FM materials in the C2DB based on a model expression for T_C and the parameters from equation (12).

For metals, the Heisenberg parameters available in C2DB should be used with care because the Heisenberg model is not expected to provide an accurate description of magnetic interactions in this case. Nevertheless, even for metals the sign and magnitude of the parameters provide an important qualitative measure of the magnetic interactions that may be used to screen and select materials for more detailed investigations of magnetic properties.

A negative value of J implies the existence of an AFM state with lower energy than the FM state used in C2DB. This parameter is thus crucial to consider when judging the stability and relevance of a material classified as magnetic in C2DB (see section 2.5). Figure 20 shows the distribution of exchange coupling constants (weighted by S^2) of the magnetic materials in the C2DB. The distribution is slightly skewed to the positive side indicating that FM order is more common than AFM order.

The origin of magnetic anisotropy may stem from either single-ion anisotropy or anisotropic exchange and it is in general difficult *a priori* to determine, which mechanism is most important. There is, however, a tendency in the literature to neglect anisotropic exchange terms in a Heisenberg model description of magnetism and focus solely on the single-ion anisotropy. In figure 20 we show a scatter plot of the anisotropy parameters A and λ for the FM materials ($J > 0$). The spread of the parameters indicate that the magnetic anisotropy is in general equally likely to originate from both mechanisms and neglecting anisotropic exchange is not advisable. For ferromagnets, the model (equation (12)) only exhibits magnetic order at finite temperatures if $A(2S - 1) + \lambda N_{nm} > 0$ [113]. Neglecting anisotropic exchange thus excludes materials with $A < 0$ that satisfies $A(2S - 1) + \lambda N_{nm} > 0$. This is in fact the case for 11 FM insulators and 31 FM metals in the C2DB.

5.9. Raman spectrum

Raman spectroscopy is an important technique used to probe the vibrational modes of a solid (or molecule) by means of inelastic scattering of light [114]. In fact, Raman spectroscopy is the dominant method for characterising 2D materials and can yield detailed information about chemical composition, crystal structure and layer thickness. There exist several different types of Raman spectroscopies that differ mainly by the number of photons and phonons involved in the scattering process [114]. The first-order Raman process, in which only a single phonon is involved, is the dominant scattering process in samples with low defect concentrations.

In a recent work, the first-order Raman spectra of 733 monolayer materials from the C2DB were calculated, and used as the basis for an automatic procedure for identifying a 2D material entirely from its experimental Raman spectrum [115]. The Raman spectrum is calculated using third-order perturbation theory to obtain the rate of scattering processes involving creation/annihilation of one phonon and two photons, see reference [115] for details. The light field is written as $\mathcal{F}(t) = \mathcal{F}_{\text{in}} \mathbf{u}_{\text{in}} \exp(-i\omega_{\text{in}}t) + \mathcal{F}_{\text{out}} \mathbf{u}_{\text{out}} \exp(-i\omega_{\text{out}}t) + \text{c.c.}$ where $\mathcal{F}_{\text{in/out}}$ and $\omega_{\text{in/out}}$ denote the amplitudes and frequencies of the input/output electromagnetic fields, respectively. In addition, $\mathbf{u}_{\text{in/out}} = \sum_i u_{\text{in/out}}^i \mathbf{e}_i$ are the corresponding polarisation vectors, where \mathbf{e}_i denotes the unit

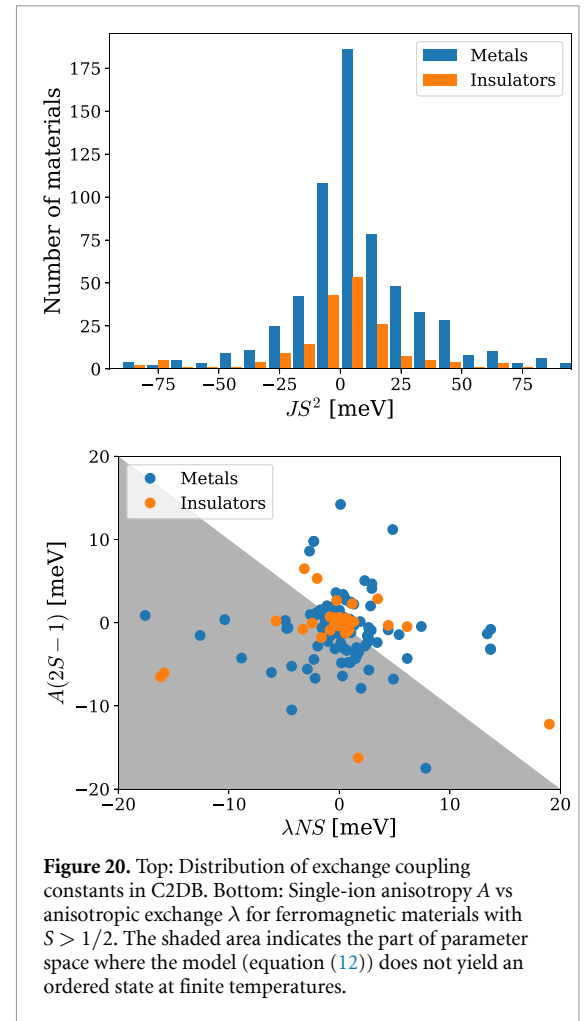
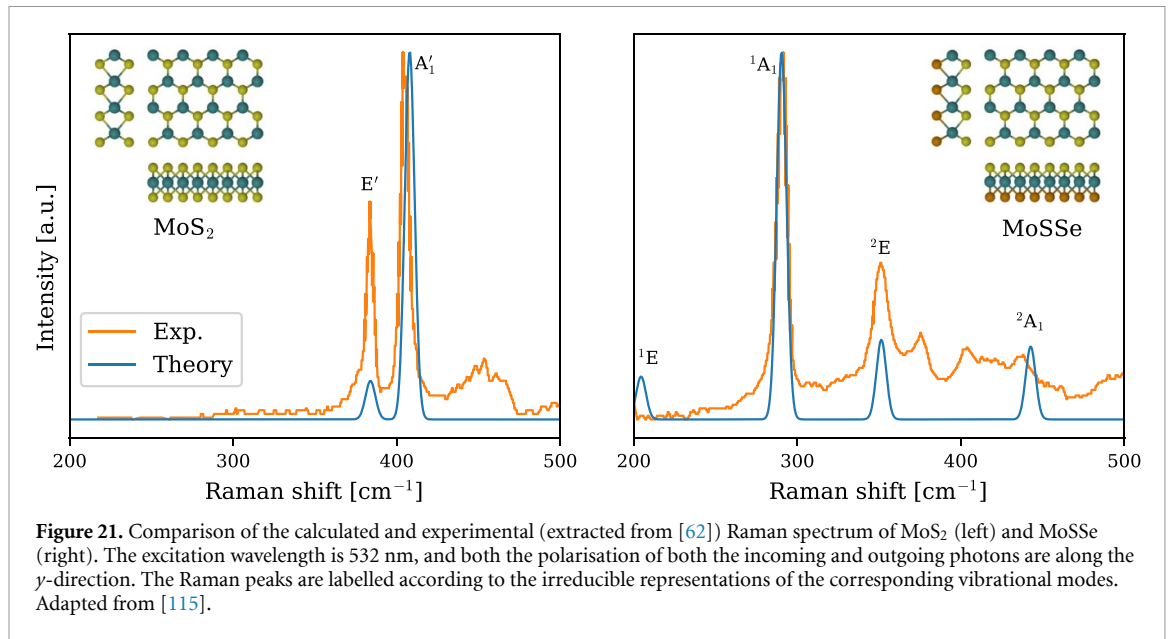


Figure 20. Top: Distribution of exchange coupling constants in C2DB. Bottom: Single-ion anisotropy A vs anisotropic exchange λ for ferromagnetic materials with $S > 1/2$. The shaded area indicates the part of parameter space where the model (equation (12)) does not yield an ordered state at finite temperatures.

vector along the i -direction with $i \in \{x, y, z\}$. Using this light field, the final expression for the Stokes Raman intensity involving scattering events by only one phonon reads [115]:

$$I(\omega) = I_0 \sum_{\nu} \frac{n_{\nu} + 1}{\omega_{\nu}} \left| \sum_{ij} u_{\text{in}}^i R_{ij}^{\nu} u_{\text{out}}^j \right|^2 \delta(\omega - \omega_{\nu}). \quad (13)$$

Here, I_0 is an unimportant constant (since Raman spectra are always reported normalised), and n_{ν} is obtained from the Bose–Einstein distribution, i.e. $n_{\nu} \equiv (\exp[\hbar\omega_{\nu}/k_B T] - 1)^{-1}$ at temperature T for a Raman mode with energy $\hbar\omega_{\nu}$. Note that only phonons at the Brillouin zone center (with zero momentum) contribute to the one-phonon Raman processes due to momentum conservation. In equation (13), R_{ij}^{ν} is the Raman tensor for phonon mode ν , which involves electron–phonon and dipole matrix elements as well as the electronic transition energies and the incident excitation frequency. Equation (13) has been used to compute the Raman spectra of the 733 most stable, non-magnetic monolayers in C2DB for a range of excitation frequencies and polarisation configurations. Note that the Raman shift $\hbar\omega$ is typically expressed in cm^{-1} with



1 meV equivalent to 8.0655 cm^{-1} . In addition, for generating the Raman spectra, we have used a Gaussian [$G(\omega) = (\sigma\sqrt{2\pi})^{-1} \exp(-\omega^2/2\sigma^2)$] with a variance $\sigma = 3 \text{ cm}^{-1}$ to replace the Dirac delta function, which accounts for the inhomogeneous broadening of phonon modes.

As an example, figure 21 shows the calculated Raman spectrum of monolayer MoS₂ and the Janus monolayer MoSSe (see section 4.1). Experimental Raman spectra extracted from reference [62] are shown for comparison. For both materials, good agreement between theory and experiment is observed for the peak positions and relative amplitudes of the main peaks. The small deviations can presumably be attributed to substrate interactions and defects in the experimental samples as well as the neglect of excitonic effects in the calculations. The qualitative differences between the Raman spectra can be explained by the different point groups of the materials (C_{3v} and D_{3h} , respectively), see reference [115]. In particular, the lower symmetry of MoSSe results in a lower degeneracy of its vibrational modes leading to more peaks in the Raman spectrum.

Very recently, the Raman spectra computed from third order perturbation theory as described above, were supplemented by spectra obtained from the more conventional Kramers–Heisenberg–Dirac (KHD) approach. Within the KHD method, the Raman tensor is obtained as the derivative of the static electric polarisability (or equivalently, the susceptibility) along the vibrational normal modes [116, 117]:

$$R_{ij}^{\nu} = \sum_{\alpha l} \frac{\partial \chi_{ij}^{(1)}}{\partial r_{\alpha l}} \frac{v_{\alpha l}^{\nu}}{\sqrt{M_{\alpha}}}. \quad (14)$$

Here, $\chi_{ij}^{(1)}$ is the (first-order) susceptibility tensor, r_{α} and M_{α} are the position and atomic mass of atom

α , respectively, and $v_{\alpha l}^{\nu}$ is the eigenmode of phonon ν . The two approaches, i.e. the KHD and third-order perturbation approach, can be shown to be equivalent [114], at least when local field effects can be ignored as is typically the case for 2D materials [35]. We have also confirmed this equivalence from our calculations. Furthermore, the computational cost of both methods is also similar [115]. However, the KHD approach typically converge faster with respect to both the number of bands and k -grid compared to the third-order perturbation method. This stems from the general fact that higher-order perturbation calculations converge slower with respect to k -grid and they require additional summations over a complete basis set (virtual states) and hence a larger number of bands [118]. Currently, Raman spectra from both approaches can be found at the C2DB website.

5.10. Second harmonics generation

Nonlinear optical (NLO) phenomena such as harmonic generation, Kerr, and Pockels effects are of great technological importance for lasers, frequency converters, modulators, etc. In addition, NLO spectroscopy has been extensively employed to obtain insight into materials properties [119] that are not accessible by e.g. linear optical spectroscopy. Among numerous nonlinear processes, second-harmonic generation (SHG) has been widely used for generating new frequencies in lasers as well as identifying crystal orientations and symmetries.

Recently, the SHG spectrum was calculated for 375 non-magnetic, non-centrosymmetric semiconducting monolayers of the C2DB, and multiple 2D materials with giant optical nonlinearities were identified [120]. In the SHG process, two incident photons at frequency ω generate an emitted photon at frequency of 2ω . Assume that a mono-harmonic electric

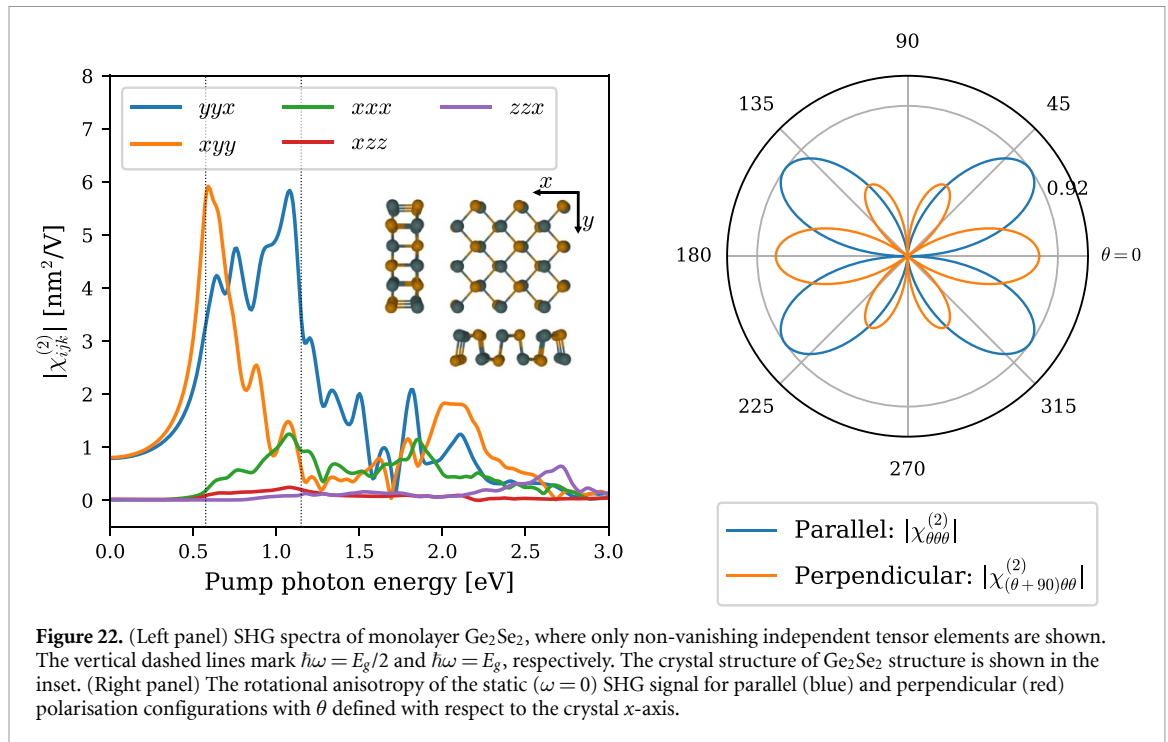


Figure 22. (Left panel) SHG spectra of monolayer Ge_2Se_2 , where only non-vanishing independent tensor elements are shown. The vertical dashed lines mark $\hbar\omega = E_g/2$ and $\hbar\omega = E_g$, respectively. The crystal structure of Ge_2Se_2 structure is shown in the inset. (Right panel) The rotational anisotropy of the static ($\omega = 0$) SHG signal for parallel (blue) and perpendicular (red) polarisation configurations with θ defined with respect to the crystal x -axis.

field written $\mathcal{F}(t) = \sum_i \mathcal{F}_i \mathbf{e}_i e^{-i\omega t} + \text{c.c.}$ is incident on the material, where \mathbf{e}_i denotes the unit vector along direction $i \in \{x, y, z\}$. The electric field induces a SHG polarisation density $\mathbf{P}^{(2)}$, which can be obtained from the quadratic susceptibility tensor $\chi_{ijk}^{(2)}$,

$$P_i^{(2)}(t) = \epsilon_0 \sum_{jk} \chi_{ijk}^{(2)}(\omega, \omega) \mathcal{F}_j \mathcal{F}_k e^{-2i\omega t} + \text{c.c.}, \quad (15)$$

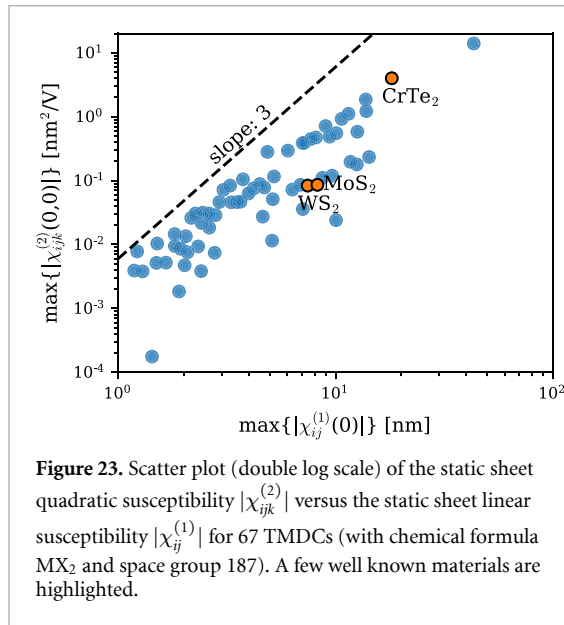
where ϵ_0 denotes the vacuum permittivity. $\chi_{ijk}^{(2)}$ is a symmetric (due to intrinsic permutation symmetry i.e. $\chi_{ijk}^{(2)} = \chi_{jik}^{(2)}$) rank-3 tensor with at most 18 independent elements. Furthermore, similar to the piezoelectric tensor, the point group symmetry reduces the number of independent tensor elements.

In the C2DB, the quadratic susceptibility is calculated using density matrices and perturbation theory [118, 121] with the involved transition dipole matrix elements and band energies obtained from DFT. The use of DFT single-particle orbitals implies that excitonic effects are not accounted for. The number of empty bands included in the sum over bands was set to three times the number of occupied bands. The width of the Fermi–Dirac occupation factor was set to $k_B T = 50$ meV, and a line-shape broadening of $\eta = 50$ meV was used in all spectra. Furthermore, time-reversal symmetry was imposed in order to reduce the \mathbf{k} -integrals to half the BZ. For various 2D crystal classes, it was verified by explicit calculation that the quadratic tensor elements fulfil the expected symmetries, e.g. that they all vanish identically for centrosymmetric crystals.

As an example, the calculated SHG spectra for monolayer Ge_2Se_2 is shown in figure 22 (left panel).

Monolayer Ge_2Se_2 has five independent tensor elements, $\chi_{xxx}^{(2)}$, $\chi_{xyy}^{(2)}$, $\chi_{xzz}^{(2)}$, $\chi_{yyx}^{(2)} = \chi_{yxy}^{(2)}$, and $\chi_{zzx}^{(2)} = \chi_{zxz}^{(2)}$, since it is a group-IV dichalcogenide with an orthorhombic crystal structure (space group 31 and point group C_{2v}). Note that, similar to the linear susceptibility, the bulk quadratic susceptibility (with SI units of mV^{-1}) is ill-defined for 2D materials (since the volume is ambiguous) [120]. Instead, the unambiguous *sheet* quadratic susceptibility (with SI units of $\text{m}^2 \text{V}^{-1}$) is evaluated. In addition to the frequency-dependent SHG spectrum, the angular dependence of the static ($\omega = 0$) SHG intensity at normal incidence for parallel and perpendicular polarisations (relative to the incident electric field) is calculated, see figure 22 (right panel). Such angular resolved SHG spectroscopy has been widely used for determining the crystal orientation of 2D materials. The calculated SHG spectra for all non-vanishing inequivalent polarisation configurations and their angular dependence, are available in the C2DB.

Since C2DB has already gathered various material properties of numerous 2D materials, it provides a unique opportunity to investigate interrelations between different material properties. For example, the strong dependence of the quadratic optical response on the electronic band gap was demonstrated on basis of the C2DB data [120]. As another example of a useful correlation, the static quadratic susceptibility is plotted versus the static linear susceptibility for 67 TMDCs (with formula MX_2 , space group 187) in figure 23. Note that for materials with several independent tensor elements, only the largest is shown. There is a very clear correlation between the two quantities. This is not unexpected as both



the linear and quadratic optical responses are functions of the transition dipole moments and transition energies. More interestingly, the strength of the quadratic response seems to be a very good approximation to be given by a universal constant times the linear susceptibility to the power of three (ignoring polarisation indices), i.e.

$$\chi^{(2)}(0,0) \approx A\chi^{(1)}(0)^3, \quad (16)$$

where A is only weakly material dependent. Note that this scaling law is also known in classical optics as semi-empirical Miller's rule for non-resonant quadratic responses [122], which states that the second order electric susceptibility is proportional to the product of the first-order susceptibilities at the three frequencies involved.

6. Machine learning properties

In recent years, material scientists have shown great interest in exploiting the use of machine learning (ML) techniques for predicting materials properties and guiding the search for new materials. ML is the scientific study of algorithms and statistical models that computer systems can use to perform a specific task without using explicit instructions but instead relying on patterns and inference. Within the domain of materials science, one of the most frequent problems is the mapping from atomic configuration to material property, which can be used e.g. to screen large material spaces in search of optimal candidates for specific applications [123, 124].

In the ML literature, the mathematical representation of the input observations is often referred to as a fingerprint. Any fingerprint must satisfy a number of general requirements [125]. In particular, a fingerprint must be:

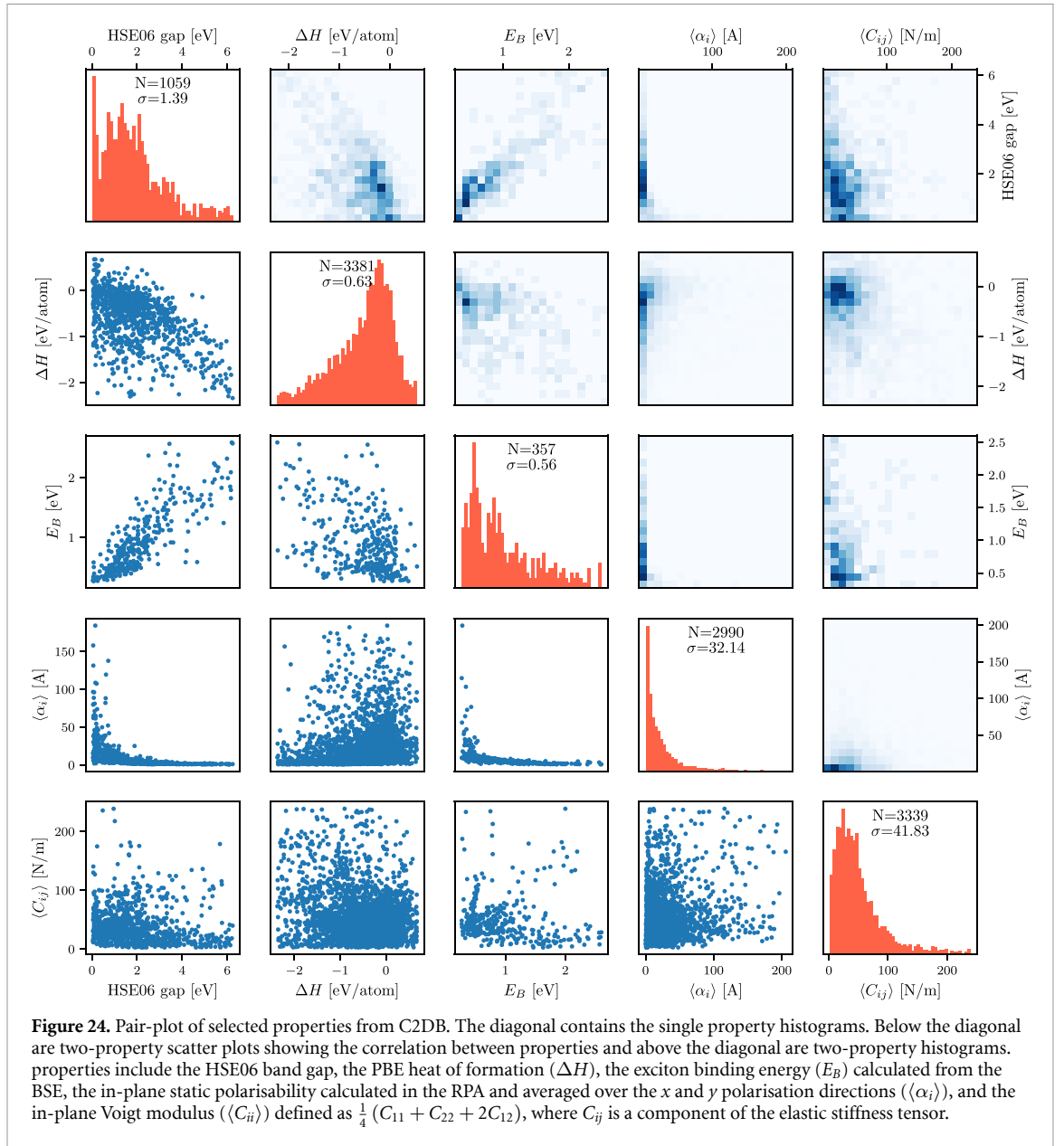
- Complete:* The fingerprint should incorporate all the relevant input for the underlying problem, i.e. materials with different properties should have different fingerprints.
- Compact:* The fingerprint should contain no or a minimal number of features redundant to the underlying problem. This includes being invariant to rotations, translations and other transformations that leave the properties of the system invariant.
- Descriptive:* Materials with similar target values should have similar fingerprints.
- Simple:* The fingerprint should be efficient to evaluate. In the present context, this means that calculating the fingerprint should be significantly faster than calculating the target property.

Several types of atomic-level materials fingerprints have been proposed in the literature, including general purpose fingerprints based on atomistic properties [126, 127] possibly encoding information about the atomic structure, i.e. atomic positions [125, 128, 129], and specialised fingerprints tailored for specific applications (materials/properties) [130, 131].

The aim of this section is to demonstrate how the C2DB may be utilised for ML-based prediction of general materials properties. Moreover, the study serves to illustrate the important role of the fingerprint for such problems. The 2D materials are represented using three different fingerprints: two popular structural fingerprints and a more advanced fingerprint that encodes information about the electronic structure via the PDOS. The target properties include the HSE06 band gap, the PBE heat of formation (ΔH), the exciton binding energy (E_B) obtained from the many-body BSE, the in-plane static polarisability calculated in the RPA averaged over the x and y polarisation directions ($\langle\alpha_i\rangle$), and the in-plane Voigt modulus ($\langle C_{ii}\rangle$) defined as $\frac{1}{4}(C_{11} + C_{22} + 2C_{12})$, where C_{ij} is a component of the elastic stiffness tensor in Mandel notation.

To introduce the data, figure 24 shows pair-plots of the dual-property relations of these properties. The plots in the diagonal show the single-property histograms, whereas the off-diagonals show dual-property scatter plots below the diagonal and histograms above the diagonal. Clearly, there are only weak correlations between most of the properties, with the largest degree of correlation observed between the HSE06 gap and exciton binding energy. The lack of strong correlations motivates the use of ML for predicting the properties.

The prediction models are built using the Ewald sum matrix and many-body tensor representation (MBTR) as structural fingerprints. The Ewald fingerprint is a version of the simple Coulomb matrix fingerprint [128] modified to periodic systems [125]. The MBTR encodes first, second and third order



terms like atomic numbers, distances and angles between atoms in the system [129]. As an alternative to the structural fingerprints, a representation based on the PBE PDOS is also tested. This fingerprint⁶ encodes the coupling between the PDOS at different atomic orbitals in both energy and real space. It is defined as:

$$\rho_{\nu\nu'}(E, R) = \sum_{a \in \text{cell}} \sum_{a'} \rho_{a\nu}(E) \rho_{a'\nu'}(E) G \times (R - |R_a - R_{a'}|), \quad (17)$$

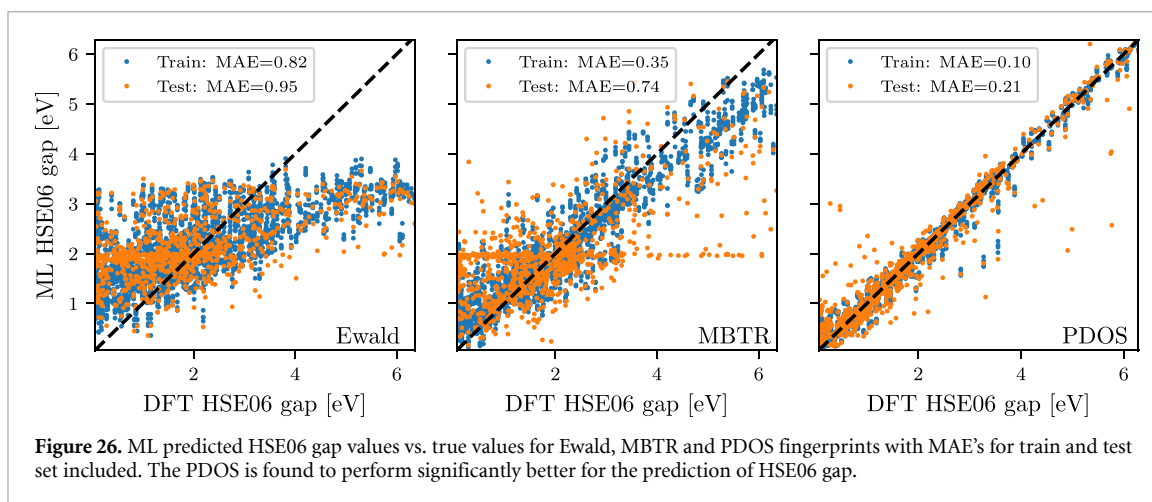
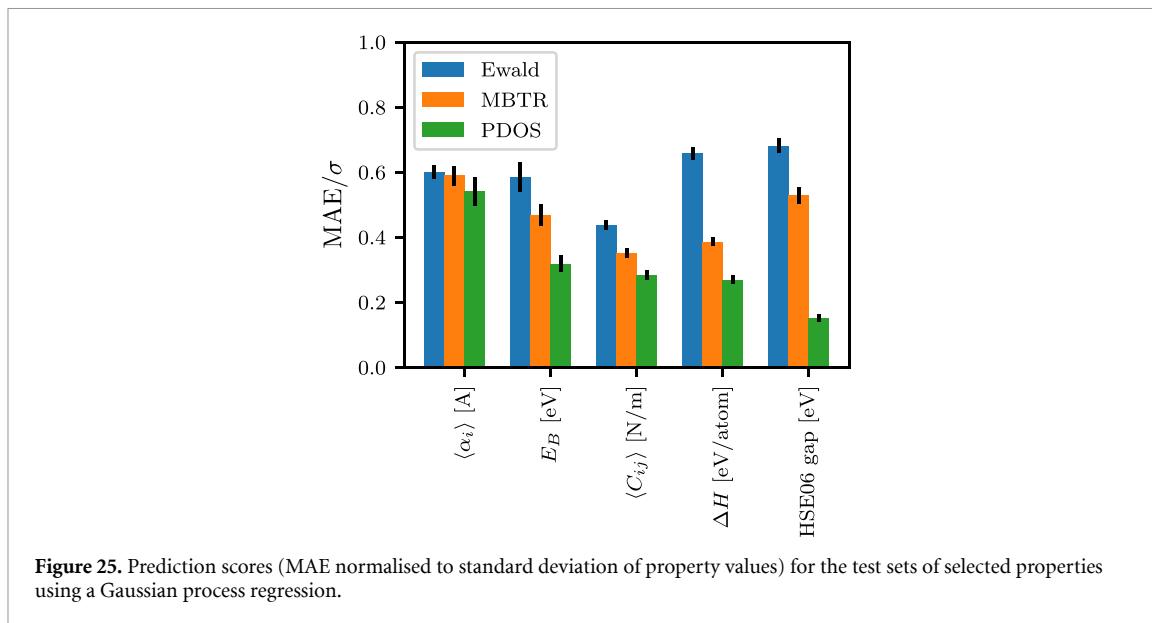
where G is a Gaussian smearing function, a denotes the atoms, ν denotes atomic orbitals, and the PDOS is given by:

$$\rho_{a\nu}(E) = \sum_n |\langle \psi_n | a\nu \rangle|^2 G(E - \epsilon_n), \quad (18)$$

⁶ Details will be published elsewhere.

where n runs over all eigenstates of the system. Since this fingerprint requires a DFT-PBE calculation to be performed, additional features derivable from the DFT calculation can be added to the fingerprint. In this study, the PDOS fingerprint is amended by the PBE band gap. The latter can in principle be extracted from the PDOS, but its explicit inclusion has been found to improve the performance of the model.

A Gaussian process regression using a simple Gaussian kernel with a noise component is used as learning algorithm. The models are trained using 5-fold cross validation on a training set consisting of 80% of the materials with the remaining 20% held aside as test data. Prior to training the model, the input space is reduced to 50 features using principal component analysis (PCA). This step is necessary to reduce the huge number of features in the MBTR fingerprint to a manageable size. Although this is not required for the Ewald and PDOS fingerprints,



we perform the same feature reduction in all cases. The optimal number of features depends on the choice of fingerprint, target property and learning algorithm, but for consistency 50 PCA components are used for all fingerprints and properties in this study.

Figure 25 shows the prediction scores obtained for the five properties using the three different fingerprints. The employed prediction score is the mean absolute error of the test set normalised by the standard deviation of the property values (standard deviations are annotated in the diagonal plots in figure 24). In general, the PDOS fingerprint outperforms the structural fingerprints. The difference between prediction scores is smallest for the static polarisability $\langle\alpha_i\rangle$ and largest for the HSE06 gap. It should be stressed that although the evaluation of the PBE-PDOS fingerprint is significantly more time consuming than the evaluation of the structural fingerprints, it is still much faster than the evaluation of all the target properties. Moreover, structural fingerprints require the atomic structure, which in turns

requires a DFT structure optimisation (unless the structure is available by other means).

The HSE06 band gap shows the largest sensitivity to the employed fingerprint. To elaborate on the HSE06 results, figure 26 shows the band gap predicted using each of the three different fingerprints plotted against the true band gap. The mean absolute errors on the test set is 0.95 and 0.74 eV for Ewald and MBTR fingerprints, respectively, while the PDOS significantly outperforms the other fingerprints with a test MAE of only 0.21 eV. This improvement in prediction accuracy is partly due to the presence of the PBE gap in the PDOS fingerprint. However, our analysis shows that the pure PDOS fingerprint without the PBE gap still outperforms the structural fingerprints. Using only the PBE gap as feature results in a test MAE of 0.28 eV.

The current results show that the precision of ML-based predictions are highly dependent on the type of target property and the chosen material representation. For some properties, the mapping between atomic structure and property is easier to learn while

others might require more/deeper information, e.g. in terms of electronic structure fingerprints. Our results clearly demonstrate the potential of encoding electronic structure information into the material fingerprint, and we anticipate more work on this relevant and exciting topic in the future.

7. Summary and outlook

We have documented a number of extensions and improvements of the C2DB made in the period 2018–2020. The new developments include: (1) A refined and more stringent workflow for filtering prospective 2D materials and classifying them according to their crystal structure, magnetic state and stability. (2) Improvements of the methodology used to compute certain challenging properties such as the full stiffness tensor, effective masses, G_0W_0 band structures, and optical absorption spectra. (3) New materials including 216 MXY Janus monolayers and 574 monolayers exfoliated from experimentally known bulk crystals. In addition, ongoing efforts to systematically obtain and characterise bilayers in all possible stacking configurations as well as point defects in the semiconducting monolayers, have been described. (4) New properties including exfoliation energies, spontaneous polarisations, Bader charges, piezoelectric tensors, IR polarisabilities, topological invariants, magnetic exchange couplings, Raman spectra, and SHG spectra. It should be stressed that the C2DB will continue to grow as new structures and properties are being added, and thus the present paper should not be seen as a final report on the C2DB but rather a snapshot of its current state.

In addition to the above mentioned improvements relating to data quantity and quality, the C2DB has been endowed with a comprehensive documentation layer. In particular, all data presented on the C2DB website are now accompanied by an information field that explains the meaning and representation (if applicable) of the data and details how it was calculated thus making the data easier to understand, reproduce, and deploy.

The C2DB has been produced using the ASR in combination with the GPAW electronic structure code and the MyQueue task and workflow scheduling system. The ASR is a newly developed Python-based framework designed for high-throughput materials computations. The highly flexible and modular nature of the ASR and its strong coupling to the well established community-driven ASE project, makes it a versatile framework for both high- and low-throughput materials simulation projects. The ASR and the C2DB-ASR workflow are distributed as open source code. A detailed documentation of the ASR will be published elsewhere.

While the C2DB itself is solely concerned with the properties of perfect monolayer crystals, ongoing efforts focus on the systematic characterisation

of homo-bilayer structures as well as point defects in monolayers. The data resulting from these and other similar projects will be published as separate, independent databases, but will be directly interlinked with the C2DB making it possible to switch between them in a completely seamless fashion. These developments will significantly broaden the scope and usability of the C2DB+ (+ stands for associated databases) that will help theoreticians and experimentalists to navigate one of the most vibrant and rapidly expanding research fields at the crossroads of condensed matter physics, photonics, nanotechnology, and chemistry.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: <https://doi.org/10.11583/DTU.14616660>.

Acknowledgments

The Center for Nanostructured Graphene (CNG) is sponsored by the Danish National Research Foundation, Project DNR103. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program Grant Agreement No. 773122 (LIMA) and Grant Agreement No. 951786 (NOMAD CoE). T D acknowledges financial support from the German Research Foundation (DFG Projects No. DE 2749/2-1).

ORCID iDs

Morten Niklas Gjerding  <https://orcid.org/0000-0002-5256-660X>

Alireza Taghizadeh  <https://orcid.org/0000-0003-0876-9538>


Asbjørn Rasmussen  <https://orcid.org/0000-0001-7110-9255>

Sajid Ali  <https://orcid.org/0000-0001-7865-2664>

Fabian Bertoldo  <https://orcid.org/0000-0002-1219-8689>

Thorsten Deilmann  <https://orcid.org/0000-0003-4165-2446>

Nikolaj Rørbæk Knøsgaard  <https://orcid.org/0000-0003-3709-5464>

Mads Kruse  <https://orcid.org/0000-0002-0599-5110>

Ask Hjorth Larsen  <https://orcid.org/0000-0001-5267-6852>

Simone Manti  <https://orcid.org/0000-0003-3770-0863>

Thomas Garm Pedersen  <https://orcid.org/0000-0002-9466-6190>

Urko Petralanda  <https://orcid.org/0000-0003-0226-0028>

Thorbjørn Skovhus  <https://orcid.org/0000-0001-5215-6419>

Mark Kamper Svendsen  <https://orcid.org/0000-0001-9718-849X>

Jens Jørgen Mortensen  <https://orcid.org/0000-0001-5090-6706>

Thomas Olsen  <https://orcid.org/0000-0001-6256-9284>

Kristian Sommer Thygesen  <https://orcid.org/0000-0001-5197-214X>

References

- [1] Kohn W and Sham L J 1965 *Phys. Rev.* **140** A1133
- [2] Schwierz F 2010 *Nat. Nanotechnol.* **5** 487
- [3] Novoselov K, Mishchenko A, Carvalho A and Castro Neto A 2016 *Science* **353** 6298
- [4] Ferrari A C et al 2015 *Nanoscale* **7** 4598–810
- [5] Bhimanapati G R et al 2015 *ACS Nano* **9** 11509–39
- [6] Hastrup S et al 2018 *2D Mater.* **5** 042002
- [7] Shivayogimath A et al 2019 *Nat. Commun.* **10** 1–7
- [8] Zhou J et al 2018 *Nature* **556** 355–9
- [9] Anasori B, Lukatskaya M R and Gogotsi Y 2017 *Nat. Rev. Mater.* **2** 1–17
- [10] Dou L et al 2015 *Science* **349** 1518–21
- [11] Mounet N et al 2018 *Nat. Nanotechnol.* **13** 246–52
- [12] Ashton M, Paul J, Sinnott S B and Hennig R G 2017 *Phys. Rev. Lett.* **118** 106101
- [13] Geim A K and Grigorieva I V 2013 *Nature* **499** 419–25
- [14] Cao Y, Fatemi V, Fang S, Watanabe K, Taniguchi T, Kaxiras E and Jarillo-Herrero P 2018 *Nature* **556** 43–50
- [15] Bistrizter R and MacDonald A H 2011 *Proc. Natl Acad. Sci.* **108** 12233–7
- [16] Zhao X et al 2020 *Nature* **581** 171–7
- [17] Wan J, Lacey S D, Dai J, Bao W, Fuhrer M S and Hu L 2016 *Chem. Soc. Rev.* **45** 6742–65
- [18] Wilkinson M D et al 2016 *Sci. Data* **3** 1–9
- [19] Wirtz L, Marini A and Rubio A 2006 *Phys. Rev. Lett.* **96** 126104
- [20] Cudazzo P, Tokatly I V and Rubio A 2011 *Phys. Rev. B* **84** 085406
- [21] Klots A et al 2014 *Sci. Rep.* **4** 6608
- [22] Chernikov A, Berkelbach T C, Hill H M, Rigosi A, Li Y, Aslan O B, Reichman D R, Hybertsen M S and Heinz T F 2014 *Phys. Rev. Lett.* **113** 076802
- [23] Olsen T, Latini S, Rasmussen F and Thygesen K S 2016 *Phys. Rev. Lett.* **116** 056401
- [24] Riis-Jensen A C, Gjerding M N, Russo S and Thygesen K S 2020 *Phys. Rev. B* **102** 201402
- [25] Felipe H, Xian L, Rubio A and Louie S G 2020 *Nat. Commun.* **11** 1–10
- [26] Sohler T, Gibertini M, Calandra M, Mauri F and Marzari N 2017 *Nano Lett.* **17** 3758–63
- [27] Ugeda M M et al 2014 *Nat. Mater.* **13** 1091–5
- [28] Winther K T and Thygesen K S 2017 *2D Mater.* **4** 025059
- [29] Wang Z, Rhodes D A, Watanabe K, Taniguchi T, Hone J C, Shan J and Mak K F 2019 *Nature* **574** 76–80
- [30] Gong C et al 2017 *Nature* **546** 265–9
- [31] Huang B et al 2017 *Nature* **546** 270–3
- [32] Chang K et al 2016 *Science* **353** 274–8
- [33] Olsen T, Andersen E, Okugawa T, Torelli D, Deilmann T and Thygesen K S 2019 *Phys. Rev. Mater.* **3** 024005
- [34] Marrazzo A, Gibertini M, Campi D, Mounet N and Marzari N 2019 *Nano Lett.* **19** 8431–40
- [35] Thygesen K S 2017 *2D Mater.* **4** 022004
- [36] Torelli D and Olsen T 2018 *2D Mater.* **6** 015028
- [37] Rasmussen F A and Thygesen K S 2015 *J. Phys. Chem. C* **119** 13169–83
- [38] Enkovaara J et al 2010 *J. Phys.: Condens. Matter.* **22** 253202
- [39] Mortensen J J, Hansen L B and Jacobsen K W 2005 *Phys. Rev. B* **71** 035109
- [40] Gjerding M, Skovhus T, Rasmussen A, Bertoldo F, Larsen A H, Mortensen J J and Thygesen K S 2021 Atomic simulation recipes—a python framework and library for automated workflows (arXiv:2104.13431)
- [41] Larsen A H et al 2017 *J. Phys.: Condens. Matter.* **29** 273002
- [42] Mortensen J J, Gjerding M and Thygesen K S 2020 *J. Open Source Softw.* **5** 1844
- [43] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 *JOM* **65** 1501–9
- [44] Jain A et al 2013 *APL Mater.* **1** 011002
- [45] Curtarolo S et al 2012 *Comput. Mater. Sci.* **58** 218–26
- [46] Ataca C, Sahin H and Ciraci S 2012 *J. Phys. Chem. C* **116** 8983–99
- [47] Lebègue S, Björkman T, Klintonberg M, Nieminen R M and Eriksson O 2013 *Phys. Rev. X* **3** 031002
- [48] Kormányos A, Burkard G, Gmitra M, Fabian J, Zólyomi V, Drummond N D and Fal’ko V 2015 *2D Mater.* **2** 022001
- [49] Zhou J et al 2019 *Sci. Data* **6** 1–10
- [50] Choudhary K, Kalish I, Beams R and Tavazza F 2017 *Sci. Rep.* **7** 1–16
- [51] Larsen A H, Vanin M, Mortensen J J, Thygesen K S and Jacobsen K W 2009 *Phys. Rev. B* **80** 195112
- [52] Larsen P M, Pandey M, Strange M and Jacobsen K W 2019 *Phys. Rev. Mater.* **3** 034003
- [53] Ong S P et al 2013 *Comput. Mater. Sci.* **68** 314–19
- [54] Patrick C E, Jacobsen K W and Thygesen K S 2015 *Phys. Rev. B* **92** 201205
- [55] Kirklin S, Saal J E, Meredig B, Thompson A, Doak J W, Aykol M, Rühl S and Wolverton C 2015 *npj Computat. Mater.* **1** 1–15
- [56] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865–8
- [57] Maździarz M 2019 *2D Mater.* **6** 048001
- [58] Li Y and Heinz T F 2018 *2D Mater.* **5** 025021
- [59] Hadley L N and Dennison D 1947 *J. Opt. Soc. Am.* **37** 451–65
- [60] Wang G, Chernikov A, Glazov M M, Heinz T F, Marie X, Amand T and Urbaszek B 2018 *Rev. Mod. Phys.* **90** 021001
- [61] Lu A Y et al 2017 *Nat. Nanotechnol.* **12** 744–9
- [62] Zhang J et al 2017 *ACS Nano* **11** 8192–8
- [63] Fülöp B et al 2018 *2D Mater.* **5** 031013
- [64] Riis-Jensen A C, Deilmann T, Olsen T and Thygesen K S 2019 *ACS Nano* **13** 13354
- [65] Bychkov Y A and Rashba E I 1984 *J. Phys. C: Solid State Phys.* **17** 6039
- [66] Petersen L and Hedegård P 2000 *Surf. Sci.* **459** 49–56
- [67] Bergerhoff G, Brown I and Allen F et al 1987 *Int. Union Crystallogr., Chester* **360** 77–95
- [68] Gražulis S et al 2012 *Nucleic Acids Res.* **40** D420–7
- [69] Qian X, Liu J, Fu L and Li J 2014 *Science* **346** 1344–7
- [70] Torelli D, Moustafa H, Jacobsen K W and Olsen T 2020 *npj Comput. Mater.* **6** 158
- [71] Mak K F, Lee C, Hone J, Shan J and Heinz T F 2010 *Phys. Rev. Lett.* **105** 136805
- [72] Splendiani A, Sun L, Zhang Y, Li T, Kim J, Chim C Y, Galli G and Wang F 2010 *Nano Lett.* **10** 1271–5
- [73] Xiao J et al 2020 *Nat. Phys.* **16** 1028–34
- [74] Sivadas N, Okamoto S, Xu X, Fennie C J and Xiao D 2018 *Nano Lett.* **18** 7658–64
- [75] Liu Y, Wu L, Tong X, Li J, Tao J, Zhu Y and Petrovic C 2019 *Sci. Rep.* **9** 1–8
- [76] Yasuda K, Wang X, Watanabe K, Taniguchi T and Jarillo-Herrero P 2020 (arXiv:2010.06600)
- [77] Grimme S, Antony J, Ehrlich S and Krieg H 2010 *J. Chem. Phys.* **132** 154104
- [78] Northrup T and Blatt R 2014 *Nat. Photon.* **8** 356–63
- [79] O’Brien J, Furusawa A and Vuckovic J 2009 Photonic quantum technologies *nat Photonics* **3** 687
- [80] Zhang S and Northrup J E 1991 *Phys. Rev. Lett.* **67** 2339

- [81] Van de Walle C G, Laks D, Neumark G and Pantelides S 1993 *Phys. Rev. B* **47** 9425
- [82] Janak J F 1978 *Phys. Rev. B* **18** 7165
- [83] Pandey M, Rasmussen F A, Kuhar K, Olsen T, Jacobsen K W and Thygesen K S 2016 *Nano Lett.* **16** 2234–9
- [84] Kaappa S, Malola S and Häkkinen H 2018 *J. Phys. Chem. A* **122** 8576–84
- [85] Levi G, Ivanov A V and Jonsson H 2020 *Faraday Discuss.* **224** 448–66
- [86] Grimme S, Antony J, Ehrlich S and Krieg H 2010 *J. Chem. Phys.* **132** 154104
- [87] Bader R F W 1990 *Atoms in Molecules: A Quantum Theory (The Int. Series of Monographs on Chemistry vol 22)* (Oxford: Clarendon)
- [88] Tang W, Sanville E and Henkelman G 2009 *J. Phys.: Condens. Matter.* **21** 084204
- [89] Resta R 1992 *Ferroelectrics* **136** 51–5
- [90] King-Smith R D and Vanderbilt D 1993 *Phys. Rev. B* **47** 3
- [91] Zhang S and Yu F 2011 *J. Am. Ceram. Soc.* **94** 3153–70
- [92] Maeder M D, Damjanovic D and Setter N 2004 *J. Electroceram.* **13** 385–92
- [93] Scott J F 2000 *Ferroelectric Memories* vol 3 (Berlin: Springer)
- [94] Rangel T, Fregoso B M, Mendoza B S, Morimoto T, Moore J E and Neaton J B 2017 *Phys. Rev. Lett.* **119** 067402
- [95] Resta R and Vanderbilt D 2007 *Theory of Polarization: A Modern Approach Phys. Ferroelectr.* vol 105 (Berlin: Springer) pp 31–68
- [96] Gjerding M N, Cavalcante L S R, Chaves A and Thygesen K S 2020 *J. Phys. Chem. C* **124** 11609–16
- [97] Ye Z G 2008 *Handbook of Advanced Dielectric, Piezoelectric and Ferroelectric Materials: Synthesis, Properties and Applications* (Amsterdam: Elsevier)
- [98] Ogawa T 2016 *Piezoelectric Materials* (Croatia: InTech)
- [99] Vanderbilt D 1999 *J. Phys. Chem. Solids* **61** 147–51
- [100] Authier A 2003 *Int. Tables for Crystallography: Volume D: Physical Properties of Crystals* (Dordrecht: Springer)
- [101] Duerloo K A N, Ong M T and Reed E J 2012 *J. Phys. Chem. Lett.* **3** 2871–6
- [102] Zhu H et al 2015 *Nat. Nanotechnol.* **10** 151–5
- [103] Taherinejad M, Garrity K F and Vanderbilt D 2014 *Phys. Rev. B* **89** 115102
- [104] Olsen T 2016 *Phys. Rev. B* **94** 235106
- [105] Fu L 2011 *Phys. Rev. Lett.* **106** 106802
- [106] Olsen T, Andersen E, Okugawa T, Torelli D, Deilmann T and Thygesen K S 2019 *Phys. Rev. Mater.* **3** 024005
- [107] Benalcazar W A, Bernevig B A and Hughes T L 2017 *Phys. Rev. B* **96** 245115
- [108] Mermin N D and Wagner H 1966 *Phys. Rev. Lett.* **17** 1133–6
- [109] Olsen T 2019 *MRS Commun.* **9** 1142–50
- [110] Olsen T 2017 *Phys. Rev. B* **96** 125143
- [111] Torelli D and Olsen T 2020 *J. Phys.: Condens. Matter.* **32** 335802
- [112] Lado J L and Fernández-Rossier J 2017 *2D Mater.* **4** 035002
- [113] Torelli D, Thygesen K S and Olsen T 2019 *2D Mater.* **6** 045018
- [114] Long D A 2002 *The Raman Effect: A Unified Treatment of the Theory of Raman Scattering by Molecules* (Chichester: Wiley)
- [115] Taghizadeh A, Leffers U, Pedersen T G and Thygesen K S 2020 *Nat. Commun.* **11** 3011
- [116] Lee S Y and Heller E J 1979 *J. Chem. Phys.* **71** 4777
- [117] Umari P and Pasquarello A 2003 *J. Phys. Condens. Matter* **15** S1547–52
- [118] Taghizadeh A, Hipolito F and Pedersen T G 2017 *Phys. Rev. B* **96** 195413
- [119] Prylepa A et al 2018 *J. Phys. D: Appl. Phys.* **51** 043001
- [120] Taghizadeh A, Thygesen K S and Pedersen T G 2021 *ACS Nano* **15** 7155
- [121] Aversa C and Sipe J E 1995 *Phys. Rev. B* **52** 14636–45
- [122] Miller R C 1964 *Appl. Phys. Lett.* **5** 17–19
- [123] Schmidt J, Marques M R G, Botti S and Marques M A L 2019 *Computat. Mater.* **5** 83
- [124] Zhuo Y, Mansouri Tehrani A and Brgoch J 2018 *J. Phys. Chem. Lett.* **9** 1668–73
- [125] Faber F, Lindmaa A, von Lilienfeld O A and Armiento R 2015 *Int. J. Quantum Chem.* **115** 1094–101
- [126] Ward L, Agrawal A, Choudhary A and Wolverton C 2016 *Computat. Mater.* **2** 1–7
- [127] Ghiringhelli L M, Vybiral J, Levchenko S V, Draxl C and Scheffler M 2015 *Phys. Rev. Lett.* **114** 105503
- [128] Rupp M, Tkatchenko A, Müller K R and Von Lilienfeld A 2012 *Phys. Rev. Lett.* **108** 058301
- [129] Huo H and Rupp M 2018 Unified representation of molecules and crystals for machine learning (arXiv:1704.06439)
- [130] Jorgensen P B, Mesta M, Shil S, García Lastra J M, Jacobsen K W, Thygesen K S and Schmidt M N 2018 *J. Chem. Phys.* **148** 241735
- [131] Rajan A C, Mishra A, Satsangi S, Vaish R, Mizuseki H, Lee K R and Singh A K 2018 *Chem. Mater.* **30** 4031–8