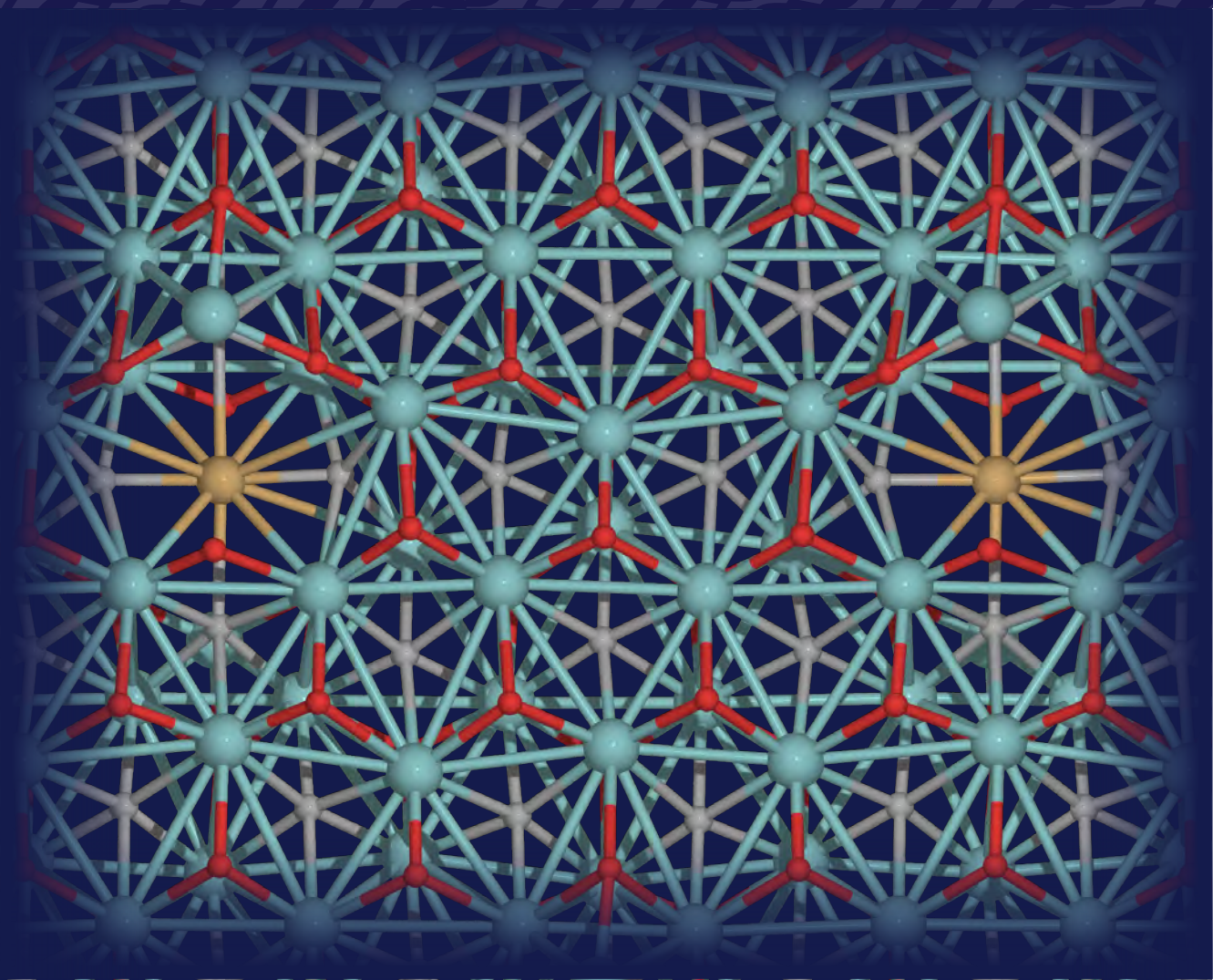Ph.D. Thesis

Computational Design of Quantum Defects in Low-Dimensional Semiconductors

Fabian Bertoldo, May 2022

Ph.D. Thesis

Computational Design of Quantum Defects in Low-Dimensional Semiconductors

Fabian Bertoldo

Kongens Lyngby, 2022

Computational Design of Quantum Defects in Low-Dimensional Semiconductors

Ph.D. Thesis
Fabian Bertoldo
May 14th, 2022

Preface

This thesis is submitted in candidacy for a Ph.D. degree in Physics from the Technical University of Denmark (DTU). The Ph.D project was carried out at the Center for Atomic-scale Materials Design (CAMD) at the DTU Department of Physics between May 2019 and May 2022. The project was supervised by Main Supervisor Professor Kristian Sommer Thygesen and Co-supervisor Associate Professor Thomas Olsen. The Ph.D. project was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program Grant No. 773122: Controlling Light-Matter Interactions by Quantum Designed 2D Materials (LIMA).

Some of the work was conducted as part of an external stay hosted by Professor Igor Abrikosov at Linköping University during the months of October and November 2021. In addition, the thesis includes work carried out in contribution to joint projects of the CAMD section.

Fabian Bertoldo
Kongens Lyngby, May 14th, 2022

Abstract

Defects are unavoidable entities in any material and disturb the behavior of a given pristine host crystal in one way or another. Depending on the application of interest, defects can either induce detrimental or advantageous property changes to the unperturbed system. The former might for example lead to defect induced reductions in the photovoltaic efficiencies or limitations due to back-scattering in transport applications whereas the latter can be a tool to create single-photon sources, intrinsically n- or p-type conducting materials, and qubits. Therefore, understanding the processes involved in the formation of defects is of utmost importance for the upcoming challenge of utilizing novel applications mitigated by defects. The field of two-dimensional (2D) materials represents a promising platform for investigations due to easier control of point defects as compared to their bulk counterparts.

In this thesis, we systematically study the impact of point defects on the pristine properties of two-dimensional materials by means of density functional theory. At the heart of our studies lies the evaluation of thermodynamic properties which gives fundamental insight into the stability and intrinsic behavior of a given defect system. We apply our methods to intrinsic and extrinsic defect entities, such as vacancies, substitutional defects, interstitials, and adatoms to understand the limitations posed by disruptions to various 2D host crystals. From a computational point of view, we develop high-throughput functionality for defect calculations in materials science to automatically set up and characterize a large number of defect systems simultaneously.

As a starting point, for the specific host material of $2H$-MoS_2, we use first-principle calculations as a useful tool to identify different intrinsic defect types in experimentally grown MoS_2 samples and establish the importance of vacancy, antisite, and grain boundary defects in real systems. We then build the bridge to a detailed property characterization of 500 simple intrinsic defects within a host crystal set of over 80 two-dimensional semiconductors. Our results give valuable insight into the formation processes of vacancy and antisite defects, analyze the nature of occurring defect states in the pristine band gap, and quantify intrinisic dopabilities as well as defect tolerances. The results are published in the open-access Quantum Point Defects in 2D Materials (QPOD) database. Subsequently, we turn our attention to extrinsic doping (*via* the incorporation of interstitial defects and adatoms) in experimentally known monolayer materials. We address the question whether a given dopant atom within a specific host material favors adsorption over absorption or *vice versa* and find out that doping 2D materials in interstitial positions is a generally challenging task. Lastly, we briefly discuss our contribution to two joint efforts of the CAMD section, namely the recent progress of the Computational 2D Materials Database (C2DB) and the Atomic Simulation Recipes (ASR), a Python framework for automated workflows.

Overall, the obtained results stress the importance of defects in any in-depth characterization of semiconducting host materials, build knowledge towards understanding the fundamental formation processes for defects in monolayer materials, and identify trends that should be factored into the planning for upcoming studies. The computational frameworks developed within this thesis can be used as the building blocks for future investigations within the field of point defects in two-dimensional materials. The publication of open-access datasets focusing on point defects, *e.g.* the QPOD database, is a valuable resource for subsequent machine learning studies which could help in circumventing the computationally demanding task of supercell defect calculations by means of accurate first principle methods for screening purposes.

Resumé

Defekter er uundgåelige enheder i et hvilket som helst materiale, og de påvirker det oprindelige værtsmateriales opførsel på adskillige måder. Afhængigt af anvendelsesom-rådet, defekter kan enten inducere skadelige eller fordelagtige ændringer af egenskaber for det uforstyrrede system. Det første kan for eksempel lede til defektinducerede re-duktioner af den fotovoltaiske effekt eller begrænsninger grundet tilbagespredning i transport-anvendelser, hvorimod det sidste kan være et værktøj til at lave enkelt-foton kilder, intrinsiske n- eller p-type ledende materialer og qubits. At forstå processerne involveret i dannelse af defekter er derfor af stor betydning for den kommende opgave med at udnytte anvendelsesmulighederne skabt af defekter. 2D-materialer repræsen-terer en lovende platform for undersøgelser grundet en lettere kontrol af punkt-defekter i forhold til deres tilsvarende 3D-materialer.

I denne afhandling undersøger vi systematisk effekten af punkt-defekter på grun-degenskaberne for 2D-materialer ved brug af tætheds-funktional-teori. I kernen af vores studier er evalueringen af termodynamiske egenskaber, som giver fundamentale indsigter i stabiliteten og den intrinsiske opførsel af et defekt-system. Vi anvender vores metoder på intrinsiske og ekstrinsiske defekt-enheder som vacancy-, substitutions-, interstitial- og adsorptions-defekter for at forstå begrænsningerne fra forstyrrelser af 2D værtskrystaller. Fra et beregningsmæssigt synspunkt udvikler vi stor-skala funk-tionaliteter for defekt-beregninger indenfor materialevidenskab til automatisk opsæt-ning og simultan karakterisering af adskillige defekt-systemer.

Med værtsmaterialet 2H-MoS_2 som udgangspunkt bruger vi ab initio beregninger til at identificere forskellige typer af intrinsiske defekter i eksperimentelt fremstillede PLD-groede MoS_2-prøver, samt til at fastslå betydningen af vakance-, antisite- og korngrænse-defekter i virkelige systemer. Dette videreføres derefter til en detaljeret karakterisering af egenskaber for 500 simple intrinsiske defekter i et sæt af værtskrys-taller bestående af mere end 80 to-dimensionale halvledere. Vores resultater giver værdifulde indsigter i formationsprocesserne af vakance- og antisite-defekter, og bruges til at analysere årsagerne til defekt-tilstande i det oprindelige båndgab samt kvan-tificere intrinsiske donationseffekter og defekttolerancer. Resultaterne er publiceret i Quantum Point Defects in 2D Materials (QPOD) databasen med en tilhørende web-applikation. Dernæst ledes opmærksomheden på ekstrinsisk donation (*via* tilføjelse af indsættelses- og adsorptions-defekter) i eksperimentelt kendte enkeltlags-materialer. Vi undersøger spørgsmålet, om et givent donationsatom i et specifkt værtsmateriale fore-trækker absorption frem for adsorption, og påviser, at det for 2D-materialer generelt er udfordrende at donere atomer i interstitielle positioner.

Til sidst diskuterer vi kort vores bidrag til to fælles projekter i CAMD-sektionen i form af den seneste fremgang for Computational 2D Materials Database (C2DB) og

Atomic Simulation Recipes (ASR), hvilket er et Python værktøj til automatiserede processer indenfor materialevidenskab.

De opnåede resultater tydeliggør vigtigheden af defekter i en dybdegående karakterisering af halvledende værtsmaterialer, opbygger viden til forståelsen af de fundamentale processer i skabelsen af defekter i enkeltlags-materialer samt identificerer tendenser, der skal medregnes i planlægningen af fremtidige studier. Beregningsværktøjerne, der er udviklet i forbindelse med denne afhandling, kan bruges som fundament for fremtidige undersøgelser af punkt-defekter i 2D materialer. Publiceringen af frit tilgængelige datasæt med fokus på punkt-defekter, fx QPOD-databasen, er en værdifuld ressource for efterfølgende maskinlæringsstudier, som kan hjælpe med at omgå de nøjagtige men krævende ab initio beregninger af supercelle-defekter for screening-formål.

Acknowledgements

Has it been perfect? Hardly. You will see in this thesis that nothing is ever really perfect. But that's okay: even if things are not perfect, they can still be amazing and great. The past three years at DTU and CAMD have been great, and many are responsible for making this period such an enjoyable time for me.

First, I'd like to thank Kristian for giving me the opportunity to pursue my Ph.D. here at DTU and for guiding me through the last three years. My project certainly had its ups and downs, but when the downs came, when things really counted, we were able to get our head down and overcome quite some challenges. I'm also very grateful for the hard efforts from CAMD's very own defect complex evolving around Simone, Sajid, and Fredrik. It's been fun to make materials better by making materials worse.

Second, a big thank you goes out to the people at LiU, in particular Joel, Rickard, and Igor, as well as all of my collaborators. Planning an external stay in the times of Corona is certainly a challenging task but you guys from Linköping made it quite easy for me. I would also like to acknowledge all of the work and collaborations that unfortunately didn't make it into the thesis in the end. It has been fun nonetheless.

Third, thanks to Ole, JJ, Morten, Ask and their amazing work to supply the researchers at CAMD with a powerful hardware and software infrastructure. To the professors Kristian, Karsten, Jakob, and Thomas for transforming a heterogeneous collection of scientists into a single research group. Thanks to Bettina and Lone for always being helpful when handling the non-scientific tasks. The main outcome of a Ph.D. is of purely scientific nature but what makes this outcome possible is the people you work with. I want to thank Mark, Matthew, Asbjørn, Estefania, Thorbjørn, Mads, Stefano, Nikolaj, Hadeel, Stefano, Daniele, and Anders for the great times we had along the way, as well as Cuau, Jiban, Sahar, Urko, Sami, Mikael, Alireza, Luca, Joachim, Frederik, Casper, Peder, Peter for making CAMD truly amazing (past and present). It's a miracle that I was able to finish the Ph.D. even though I was 'fired' on an almost daily basis. And to Mark: never forget that Zinc has the upper hand!

To my family, in German: Diese Arbeit ist nicht nur ein Resultat der letzten drei Jahre, sondern das Ergebnis eines viel längeren Zeitraums. Und ohne Euch wäre dieses Ergebnis nie zustandegekommen. Ich konnte immer auf eure Unterstützung zählen, egal wie unerwartet oder spontan meine Entscheidungen manchmal ausfielen. DANKE! And @PauliPauliPauli: you almost made it onto the cover of the thesis.

Finally, to Olga: thank you for the most unexpected cup of tea which turned out to be one of the most important ones I ever had. The cup of tea that changed things for the better. The best, really.

List of Publications

[I] **F. Bertoldo**, R. R. Unocic, Y. Lin, X. Sang, A. A. Puretzky, Y. Yu, D. Miakota, C. M. Rouleau, J. Schou, K. S. Thygesen, D. B. Geohegan, and S. Canulescu, "Intrinsic Defects in MoS_2 Grown by Pulsed Laser Deposition: From Monolayers to Bilayers,"

ACS Nano 2021, **15**, 2858−2868.

Printed in copy from page 59.

[II] **F. Bertoldo**, S. Ali, S. Manti, and K. S. Thygesen, "Quantum Point Defects in 2D Materials - the QPOD Database,"

npj Comput. Mater. 2022, **8**, 56.

Printed in copy from page 71.

[III] **F. Bertoldo**, J. Davidsson, R. Armiento, and K. S. Thygesen, "High-Throughput Doping of 2D Materials: Absorption *versus* Adsorption,"

in preparation.

Printed in copy from page 89.

[IV] M. Gjerding, A. Taghizadeh, A. Rasmussen, S. Ali, **F. Bertoldo**, T. Deilmann, N. R. Knøsgaard, M. Kruse, A. H. Larsen, S. Manti, T. G. Pedersen, U. Petralanda, T. Skovhus, M. K. Svendsen, J. J. Mortensen, T. Olsen, and K. S. Thygesen, "Recent Progress of the Computational 2D Materials Database (C2DB),"

2D Mater. 2021, **8**, 044002.

Printed in copy from page 101.

[V] M. Gjerding, T. Skovhus, A. Rasmussen, **F. Bertoldo**, A. H. Larsen, J. J. Mortensen, and K. S. Thygesen, "Atomic Simulation Recipes – a Python Framework and Library for Automated Workflows,"

Comp. Mater. Sci. 2021, **199**, 110731.

Printed in copy from page 131.

Contents

CHAPTER 1

Introduction

The real world is never perfect. Minor and major disruptions to the ideal have been, and always will be, present in our life: the train that is delayed by more than an hour, the scratch in the display of our new phone, a flat spot in the bike tire. Each and every single one of us faces some disrupted situations of various nature on a daily basis. Obviously, I do not intent to talk about the annoyances of everyday life in this thesis but rather about disruptions to perfect crystals, *i.e.* defects, in solid state systems.

Historically, disruptions to pristine materials played an important role early on: around 5000 to 6000 years ago[1], societies started utilizing copper for various tools. A few thousand years down the road, mankind realized that adding tin to copper, which creates bronze, improves the unperturbed properties of the material and results in a harder material that is easier to work with[2, 3]. If we fast-forward to more recent history, the early 20th century to be precise, semiconducting silicon was set to revolutionize the technological advances of the upcoming century. One of the first semiconducting devices based on silicon surfaced as early as 1906[4]. By doping the pristine silicon semiconductor, *i.e.* deliberately adding extrinsic defects to the material, one can effectively tune them to be of *n*- or *p*-type nature and introduce freely available electron and hole carriers into the previously insulating material. These *n*- or *p*-type semiconductors can in turn be combined to form *p*-*n*-junctions[5] and these eventually became the basis for diodes and transistors to drive the age of information technology. However, not all of the effects induced by defects are beneficial: for example, deep states induced by point defects in photovoltaic (PV) materials lead to carrier trapping and non-radiative recombination can be a source of decreased efficiencies in PV devices[6].

Scientific progress in the early stages of the twentieth century pioneered by the likes of de Broglie, Planck, Einstein, Heisenberg, Born, Jordan, and Schrödinger (to only name a few) paved the way towards the formulation of what we know as quantum mechanics. Thanks to quantum mechanics, it was now possible to describe matter on the atomic level, opening up completely new fields of research. These drastic improvements would at a later stage build the foundation to characterize single entities of point defects in solid state systems which marks another big step in understanding materials in their entirety.

This brings us to the present where we are facing the next transition from classical information technologies to quantum technologies[7]. The big question is: what will be a useful platform for those novel quantum technologies in practice? So far, quantum dots[8], atoms and photons[9], as well as superconducting circuits[10] are among some of the most promising examples for practical realizations of quantum technologies. And once again, point defects as manifestations of imperfections in a material might just

be the useful entity to advance the progress[11] in the form of qubits for quantum computation[12, 13] or single-photon emitters (SPEs) in the form of color centers[14].

Thanks to the capabilities supplied by computational materials design, finding the needle (which would be the defect system with properties perfectly suited for the specific quantum applications) in the haystack no longer seems like an impossible task. Numerous point defects in bulk materials have already been characterized and their suitability for quantum applications has been established. Common examples are the NV center in diamond[15–18], color centers in SiC[19–22], with a few promising counterparts in the realm of two-dimensional materials not far behind[23–25]. Extending this relatively small set of defect systems with technological impact is the current challenge that needs to be addressed. But one should not forget: just because finding something is possible does not mean that it is easy. The pool of existing materials is estimated to be around the order of 80,000[26] which makes our previously mentioned haystack seem more like a proper mountain. And that mountain is just containing the pristine materials.[1] Adding defects to the equation opens up a whole new dimension in the materials space as each pristine material might have tens, hundreds, thousands of possible defects depending on their respective complexity. And before we know it, our mountain has grown to be quite the peak.

Researchers within the defect physics community currently investigate various systems to eventually find exactly those defects that will end up being essential building blocks for future applications. Future applications that will ideally make a positive impact on society and technology, for instance by addressing the challenges posed by the Sustainable Development Goals (SDGs)[27]. For example, quantum computers that are able to run efficient quantum algorithms could be based on defect qubits[13]. These computers and algorithms can contribute to reducing computational time and energy used for answering new upcoming computational challenges. More sustainable photovoltaic materials might take over to drive the transition towards renewable energies. Defects play a role when thinking about those applications and understanding those disruptive entities better is a vital step for utilizing them in applications.

The first realization of graphene monolayers[28] in 2004 was a defining moment in materials science and marked the starting point for research in the world of atomically thin two-dimensional (2D) materials[29, 30]. Various phenomena in 2D materials are fundamentally different compared to their 3D (bulk) counterparts, *e.g.* large exciton binding energies[31], the implications of van der Waals forces between monolayers[28], and the ability to combine 2D materials both horizontally[32] and vertically[30] to tune their properties according to your wishes and design new 2D materials[31]. Just like for their 3D counterparts, two-dimensional materials can be subject to defect engineering[33–35]. However, on top of the usual substitutional and interstitial defects, the special nature of 2D monolayers as well as homo- and hetero-multilayers also allows for manipulation of the pristine material by means of adsorbates[36, 37] and intercalated structures[38–40]. In recent years, 2D materials have also gained importance in the context of valleytronics[41], optoelectronics[42], and quantum information science[43] as can for example be seen in the advances of single photon sources in atomically thin

---

[1]This number even increases with new hypothetical materials.

materials[44].

This PhD thesis will describe the development of computational tools and the creation of data resources for defect calculations in materials science, as well as the investigation of the effect of point defects in two-dimensional materials by means of density-functional theory with respect to their thermodynamic properties. In summary, we (i) supported the experimental identification and interpretation of intrinsic point defects in MoS_2 monolayers by utilizing our computational tools, (ii) conducted a high-throughput study on intrinsic point defects in two-dimensional materials and published the Quantum Point Defects in 2D Materials (QPOD) database[2], and (iii) systematically studied the effect of interstitial and adsorption site doping for experimentally synthesised monolayers in over 11,000 different defect systems.

We note, that this thesis is not at all meant to give a full and in-depth discussion of the physical concepts needed for the characterization of point defects. However, we try to refer to as much relevant literature as possible along the way to give the interested reader the chance to dive into the specifics of the presented topics.

Thesis Outline

We conclude the introduction with the thesis outline:

- **Chapter 2** introduces and reviews the existing basic theoretical concepts needed to characterize materials at the atomic level. In particular, we build the bridge from the many-body Schrödinger equation to density-functional theory as a powerful tool to reduce the computational complexity of systems containing many electrons. We further introduce Janak's theorem[45] which will be of high importance in the upcoming chapter about defect thermodynamics.

- **Chapter 3** dives into the world of point defects and explains how relevant thermodynamic properties like formation energies, charge transition levels, self-consistent Fermi-levels as well as equilibrium defect and carrier concentrations can be calculated from first principles. Particular emphasis is put on the utilization of Slater-Janak transition state theory for defect calculations which takes a different approach than the standard treatment of defect thermodynamics which is mostly based on total energies. In addition, this chapter briefly mentions concepts important when looking at defects beyond the thermodynamic characterization.

- **Chapter 4** takes a more abstract route to give an excursion into high-throughput methods in materials science. We outline the concept of automation, present existing high-throughput frameworks, and discuss the general structure of a scientific computational workflow. We discuss how the introduction of defects into a material opens up new dimensions in the materials space to motivate subsequent studies. Throughout the chapter, we try to relate the abstract concepts to our specific use cases.

---

[2]Find the database here: https://cmr.fysik.dtu.dk/qpod/qpod.html.

- **Chapter 5** is formulated in very practical terms and presents the computational tools that have been implemented as part of this thesis. It is mainly split up in two contributions: (i) an overview of ASR Recipes for both defect calculations and general tasks in materials science and (ii) an explanation of the ASE `DefectBuilder`, an easy setup tool for point defect structures in 3D and 2D host crystals within the Atomic Simulation Environment (ASE)[46].

- **Chapter 6** summarizes the main results obtained in the studies leading to Publications [I] to [V]. The amount of detail discussing each individual study is very limited and we aim to give a rough overview of what happened within each publication. The specifics can be found in the included copies of the publications which are collected in Chapter 9 and should be read in conjunction with Chapter 6. We note that Publications [I], [II], and [III] are regarded as the main contributions in terms of this thesis whereas Publications [IV] and [V] discuss big collaborative projects where I only played a minor role and contributed a small part to the entirety of the study.

- **Chapter 7** and **Chapter 8** are the concluding chapters of this thesis and first summarize the content of this thesis before giving a glance over which scientific questions should be addressed in future studies within the field of first principle defect calculations.

- **Chapter 9** collects the copies of Publications [I] to [V] which are linked to the brief summaries for each publication given in Chapter 6. We note that Publication [III] is in preparation and not submitted yet.

Theory: The Many-Body Problem and Density-Functional Theory

The goal of this thesis is to characterize and simulate the behavior of point defects in crystalline materials. To do so, one is in need of sophisticated methods to compute materials properties at the microscopic level. At the heart of it lies quantum mechanics and its most famous and powerful description with the Schrödinger equation[47], which established a new era of quantum physics in 1926. It was now possible to solve simple model systems like the infinite quantum well as well as real systems like the Hydrogen atom.

However, describing complex systems inevitably entails being able to treat systems with many atoms (and electrons) which makes solving Schrödinger's equation very challenging. Alternative methods to the 'standard' quantum mechanical many-body treatment arose and nowadays play an essential part in theoretical physics, computational physics, and materials science. Density-functional theory (DFT) as a formally exact theory marks one of those efforts to find alternative approaches, resulting in a powerful tool to compute properties of many-body systems in a computationally efficient manner.

In the following, we will present the many-body problem with its intrinsic challenges, introduce the concept of density-functional theory, discuss its most central theorems, and explain how DFT can help tackle the description of systems consisting of many atoms in practice. Personally, I first came into contact with DFT by reading Capelle's[48] *Bird's-Eye View of Density-Functional Theory* which gave me a great introduction into the topic. Therefore, throughout this chapter, we will follow Capelle's notation and explanations to get an overview of relevant concepts and implications of DFT.

2.1 The Many-Body Problem

At the heart of every characterization of a material's microscopic properties lies quantum mechanics. Given a system that comprises the positions of all the nuclei and electrons, the Schrödinger equation[47] makes it possible to calculate the corresponding wavefunction Ψ. In its most general form, Ψ is a function of the positions of all

nuclei and electrons of that particular system. Once the wavefunction is obtained, one can calculate physical observables of a system and get insight into its behavior. Within the Born-Oppenheimer approximation[49], the atomic positions of the nuclei are treated as static, which results in electronic and nuclear degrees of freedom being decoupled. In this decoupled form, for a single electron moving inside a potential $\nu(\mathbf{r})$, one can write down Schrödinger's equation as

$$\left[-\frac{\hbar^2 \nabla^2}{2m} + \nu(\mathbf{r}) \right] \Psi(\mathbf{r}) = \varepsilon \Psi(\mathbf{r}). \tag{2.1}$$

The solution for $\Psi(\mathbf{r})$ and its eigenvalues ε can be computed analytically. However, for a more general system consisting of N electrons, the previous equation becomes more cumbersome:

$$\left[\sum_i^N \left(-\frac{\hbar^2 \nabla_i^2}{2m} + \nu(\mathbf{r}_i) \right) + \sum_{i<j} U(\mathbf{r}_i, \mathbf{r}_j) \right] \Psi(\mathbf{r}_1..., \mathbf{r}_N) = E\Psi(\mathbf{r}_1..., \mathbf{r}_N). \tag{2.2}$$

The many-body wavefunction $\Psi(\mathbf{r}_1..., \mathbf{r}_N)$ is now a function of $3N$ spatial coordinates, and one has to take the interaction $U(\mathbf{r}_i, \mathbf{r}_j)$ between electrons i and j into account. In general, one can solve Equation (2.2), obtain the many-body wavefunction $\Psi(\mathbf{r}_1..., \mathbf{r}_N)$, and get the desired observables by computing the expectation value of the respective operator. For example, the electron density can be calculated in the following way:

$$n(\mathbf{r}) = N \int d^3r_2... \int d^3r_N \Psi^*(\mathbf{r}, \mathbf{r}_2..., \mathbf{r}_N)\Psi(\mathbf{r}, \mathbf{r}_2..., \mathbf{r}_N). \tag{2.3}$$

However, since $\Psi(\mathbf{r}_1..., \mathbf{r}_N)$ is dependent on $3N$ variables, solving the many-body Schrödinger equation becomes very challenging and alternative approaches have to be considered. In the following sections, we will discuss how the electron density $n(\mathbf{r})$ marks the foundation of density-functional theory and how it builds the framework to efficiently solve the many-body problem.

2.2 The Wavefunction *versus* Density-Functional Theory Approach

We saw that obtaining the many-body wavefunction *via* the Schrödinger equation is a challenging task for systems consisting of many electrons. Hence, alternative approaches are needed. Density-functional theory is a powerful and versatile method to compute properties of quantum mechanical systems and will be the method of choice for solving the electronic structure problem throughout this thesis.

In the previously introduced wavefunction approach, a given potential $\nu(\mathbf{r})$ is inserted into Equation (2.2), the many-body wavefunction is computed, and all desired observables can be deducted from it. DFT, a formally exact theory, takes a different pathway and tackles the given problem in a somewhat inverted manner by introducing

the ground state density $n_0(\mathbf{r})$ as its central quantity. All observables that are accessible through the many-body wavefunction $\Psi(\mathbf{r}_1..., \mathbf{r}_N)$ can be deduced from $n_0(\mathbf{r})$ as well. Thereby, density-functional theory effectively maps the many-body problem onto a single-body problem making it a powerful method for treating systems with many electrons, in particular from a computational point of view.

In the following section, we will present two fundamental theorems of DFT that pave the way towards efficiently solving the many-body problem based on the electron density. Before doing so, let us first define the electronic Hamiltonian consisting of kinetic energies $\hat{T}$, electron-electron interaction $\hat{U}$, and potential $\hat{V}$ as

$$\hat{H} = \hat{T} + \hat{U} + \hat{V} = \sum_i \left(-\frac{\hbar^2 \nabla_i^2}{2m} \right) + \sum_{i<j} U(\mathbf{r}_i, \mathbf{r}_j) + \sum_i \nu(\mathbf{r}_i). \tag{2.4}$$

We note that both $\hat{T}$ and $\hat{U}$ are the same for any non-relativistic system and they are independent of the potential.

2.3   The Hohenberg-Kohn Theorems

The ground state density $n_0(\mathbf{r})$ is the central quantity in density-functional theory. But how can a quantity only depending on three spatial variables be equivalent (at least in its use) to the full many-body wavefunction with $3N$ variables? Famously, in 1964, Hohenberg and Kohn[50] presented two fundamental theorems that ultimately link the many-body wavefunction $\Psi(\mathbf{r}_1..., \mathbf{r}_N)$ to the electronic ground state density $n_0(\mathbf{r})$. We will now present an interpretation of the Hohenberg-Kohn (HK) theorems and briefly discuss their content.[1]

Theorem 1 *Equation (2.3) can be inverted. Thus, the ground state wavefunction $\Psi_0(\mathbf{r}_1..., \mathbf{r}_N)$ (and consequently the potential $v(\mathbf{r})$) is obtainable from the so-called ground state density $n_0(\mathbf{r})$. The wavefunction and the potential are a functional of $n_0(\mathbf{r})$:*

$$\Psi_0(\mathbf{r}_1..., \mathbf{r}_N) = \Psi \left[n_0(\mathbf{r}) \right]. \tag{2.5}$$

In principle, one can compute ground state observables from the ground state density.

Theorem 2 *The ground state density that corresponds to the ground state wavefunction minimizes the energy of a given system with potential $\nu(\mathbf{r})$. This energy is called the ground state energy $E_{\nu,0}$ and can be written as*

$$E_{\nu,0} = \min_{\Psi \to n_0} \langle \Psi | \hat{T} + \hat{U} + \hat{V} | \Psi \rangle. \tag{2.6}$$

---

[1]We note that this is only an introduction to the topic to put the present work into perspective and is not at all covered in its entirety. For an in-depth explanation of the topic we direct the interested reader to many excellent works on this, *e.g.* Ref. [51].

The validity of Theorem 1 and Theorem 2 has been proven by Hohenberg and Kohn themselves[50] as well as others[52, 53]. One can define the total energy functional by obtaining the expectation value of the individual operators in Equation (2.2) for a general density $n(\mathbf{r})$ as

$$E_\nu[n] = T[n] + U[n] + V[n] = F[n] + V[n] = F[n] + \int d^3r\, n(\mathbf{r})\nu(\mathbf{r}). \tag{2.7}$$

Kinetic and interaction energies $F[n] = T[n] + U[n]$ are universal functionals and independent of potential $\nu(\mathbf{r})$.

To summarize, we have now seen that there is a one-to-one correspondence between ground state density $n_0(\mathbf{r})$ and the ground state many-body wavefunction $\Psi[n_0(\mathbf{r})]$. In addition, $n_0(\mathbf{r})$ minimizes the energy of a given system and will further be referred to as the ground state energy E_0. Any ground state observable $\hat{O}$ can be calculated by using the ground state density *via*

$$O_0 = O[n_0] = \langle \Psi[n_0]|\hat{O}|\Psi[n_0]\rangle. \tag{2.8}$$

However, being able to calculate those observables in principle does not tell us how to calculate them in practice. The Hohenberg-Kohn theorems and their implications are a very general and abstract consideration of the problem. Therefore, we will now address how one can efficiently obtain the ground state density in practice.

2.4   The Kohn-Sham Equations

One major setback of the formulation introduced in the previous section is that it is very abstract and does not give an idea of practical ways to solve the many-body problem that can now be expressed in terms of the density. To address this issue, Kohn and Sham[54] presented a computationally feasible strategy to translate the abstract HK theorems into practical terms. It is based on the idea that electrons in a many-body system are treated as single, non-interacting particles moving inside an effective potential $\nu_s[n](\mathbf{r})$.[2] The many-body interacting Schrödinger equation in Equation (2.2) can be decoupled into a set of single particle, non-interacting Schrödinger equations. As a result, from minimizing the energy functional in terms of single particle orbitals $\phi_i(\mathbf{r})$, one ends up with the famous Kohn-Sham (KS) equations[3]

$$\varepsilon_i \phi_i(\mathbf{r}) = \left[-\frac{\hbar^2 \nabla^2}{2m} + \nu_s[n](\mathbf{r}) \right] \phi_i(\mathbf{r}), \tag{2.9}$$

$$\nu_s[n](\mathbf{r}) = \nu(\mathbf{r}) + \nu_{\mathrm{H}}[n](\mathbf{r}) + \nu_{xc}[n](\mathbf{r}), \tag{2.10}$$

$$n(\mathbf{r}) = \sum_i f_i |\phi_i(\mathbf{r})|^2. \tag{2.11}$$

---

[2]This effective potential is a functional of the density $n(\mathbf{r})$.

[3]The actual derivation of the KS equations is well established and can be found in the literature, *e.g.* Ref. [48].

Here, f_i is the occupation number for orbital $\phi_i(\mathbf{r})$, $\nu(\mathbf{r})$ is the external potential and $\nu_{\mathrm{H}}[n](\mathbf{r})$ the Hartree potential. One can now start with an initial guess for the electron density $n(\mathbf{r})$ and solve the problem self-consistently to obtain the ground state density of a given system. Formally, $\nu_{xc}[n](\mathbf{r})$ incorporates all exchange and correlation effects and is obtained through the xc functional E_{xc} and the relation

$$\nu_{xc}[n](\mathbf{r}) = \frac{\delta E_{xc}}{\delta n(\mathbf{r})}. \tag{2.12}$$

However, no universal analytical form for the xc functional E_{xc} is known and $\nu_{xc}[n](\mathbf{r})$ can only be obtained when E_{xc} is approximated. Possible approximations, their advantages and limitations will be discussed in the next section.

2.5   Exchange Correlation Functionals

Finding suitable approximations for E_{xc} is vital to compute the ground state density *via* the KS equations. In the last decades, hundreds of xc functionals have been developed, every one of them serving specific purposes. Most of those functionals can be grouped into three categories: local, gradient-dependent and nonlocal functionals. In this section, we give an overview of the most prominent examples, as well as the most relevant implementations for the present thesis.

Local Density Approximation (LDA)   In the early stages of DFT, thanks to the efforts of Kohn and Sham[54], the local density approximation helped in finding one of the first realizations of a xc functional. LDA is local in the way that the exchange correlation functional only depends on one point $\mathbf{r}$ in space. The motivation behind this local approach was to mimic the behavior of the homogeneous electron gas for the exchange and correlation part of E_{xc}. The LDA xc functional can be written as

$$E_{xc}^{\mathrm{LDA}}[n] = \int d^3 r\, e_{xc}^{\mathrm{hom}}(n(\mathbf{r})) \tag{2.13}$$

in its most general form, where $e_{xc}^{\mathrm{hom}}(n(\mathbf{r}))$ is the exchange-correlation energy per volume. The exchange part, *i.e.* $e_x^{\mathrm{hom}}(n(\mathbf{r}))$, has an analytical form for the homogeneous electron gas[55, 56] while the correlation part, *i.e.* $e_c^{\mathrm{hom}}(n(\mathbf{r}))$, can be approximately calculated with quantum Monte Carlo methods[57–60]. In spite of its conceptual simplicity compared to its semi-local and nonlocal counterparts, LDA proves to produce reliable results for systems of slowly varying density[61] and metals. However, a major drawback comes into play when describing localized systems or predicting optical band gaps[62].

Generalized Gradient Approximation (GGA)   A natural evolution of local xc functionals is the incorporation of gradient corrections. In the generalized gradient approximation[63], E_{xc} takes the general form

$$E_{xc}^{\mathrm{GGA}}[n] = \int d^3 r\, f(n(\mathbf{r}), \nabla n(\mathbf{r})). \tag{2.14}$$

The integrand of Equation (2.14) now depends both on the density $n(\mathbf{r})$ and its gradient $\nabla n(\mathbf{r})$. Many different flavors of GGA functionals exist and they mainly differ in the choice of the form of $f(n(\mathbf{r}), \nabla n(\mathbf{r}))$. Perdew, Burke and Enzerhof[64] developed the famous PBE xc functional which is commonly used for various purposes due to its possibility to be applied to a broad range of problems. However, GGA functionals generally suffer from an underestimation of band gaps which is of particular relevance for defects in semiconductors. PBE will be the functional of choice for most of the calculations conducted throughout this thesis as it represents a very reasonable trade-off between accuracy and computational cost.

Hybrid Functionals Different from the previous approaches of LDA and GGA, hybrid functionals combine local density-functional theory and Hartree-Fock theory[65]. As the exchange in Hartree-Fock is no longer just a property of the density but also the Slater determinant of single electron orbitals, the computational effort needed for hybrid functionals is larger compared to its simple LDA and GGA counterparts. Famous examples of hybrid functional realizations are PBE0[66, 67] and HSE[68]. In HSE, the xc functional takes the form[68]

$$E_{xc}^{\omega \mathrm{PBEh}} = a E_x^{\mathrm{HF,SR}}(\omega) + (1-a) E_x^{\mathrm{PBE,SR}}(\omega) + E_x^{\mathrm{PBE,LR}}(\omega) + E_c^{\mathrm{PBE}}. \tag{2.15}$$

The adjustable parameter ω defines the extend of short range (SR) interactions, whereas a is the mixing parameter for the exact exchange with Hartree-Fock.[4] The full long range (LR) exchange, as well as the correlation part are simply taken from PBE. In the limit of $\omega \to \infty$, Equation (2.15) becomes PBE. Hybrid functionals in general and the HSE functional in particular are known to improve on PBE's underestimation of band gaps resulting in calculated values closer to experiment. This is of high relevance for defect calculations as applications in defect physics are often based on optical excitations where accurate band gaps are desirable.

2.6   The Slater-Janak Theorem

We have discussed why the transition from conventional many-body quantum mechanics to alternative approaches such as density-functional theory is needed, how DFT is introduced and how it embodies an efficient computational approach for treating systems consisting of many atoms. Let us now go into more detail and focus on a specific theorem that will be linked to the calculation of charge transition levels for defect systems in Chapter 3. In 1978, Janak[45] made a useful observation about the variation of the total energy in density-functional theory. The most important finding marks part one of Janak's theorems, *i.e.*

$$\frac{\partial E[N]}{\partial \eta_i} = \varepsilon_i, \tag{2.16}$$

---

[4]Typical values for the two parameters are $\omega = 0.207$ Å^{-1} and $\alpha = 0.25$[69] which is commonly referred to as HSE06.

where $E[N]$ is the total energy of a system with N electrons, η_i the occupation number for state i ($\eta_i = 1$ for $i \leq N$ and $\eta_i = 0$ for $i > N$) and ε_i its respective eigenenergy. Interestingly, Equation (2.16) is independent of the specific choice of the xc functional[45]. For the second part of Janak's theorem we can now use Equation (2.16) together with η electrons ($0 \leq \eta \leq 1$) added to the lowest unoccupied level of the system to write down the ground states of a N and $N+1$ electron system as[45]

$$E_{N+1} - E_N = \int_0^1 \varepsilon_H(\eta)d\eta, \tag{2.17}$$

where $\varepsilon_H(\eta)$ is the energy for the highest occupied orbital of the system with occupation number η. From Equation (2.17), one can obtain the transition energy between the $N+1$ and N systems. Figure 2.1 shows the total energy as a function of occupation number. The exact DFT value for the total energy is piecewise linear[70], whereas $E[N]$ shows curved behavior for GGA and LDA xc functionals[71, 72]. This observation will be of importance when we address the accuracy of SJ theory in connection with defect thermodynamics in Chapter 3.

However, as no analytical form for $\varepsilon_H(\eta)$ is known, one needs to apply approximations. The lowest-order approximation (working well when assuming that the eigenenergy of the lowest unoccupied level is varying linearly with occupation number) is to simply evaluate $\varepsilon_H(\eta)$ at $\eta = \frac{1}{2}$[73]. We will further refer to this approximation as the Slater-Janak (SJ) approximation and use it for evaluating charge transition levels for defects in 2D materials. The specifics will be discussed in Chapter 3. We note, that Williams, de Groot, and Sommers[74] presented a generalization of the transition state concept for higher order approximations to evaluate Equation (2.17).
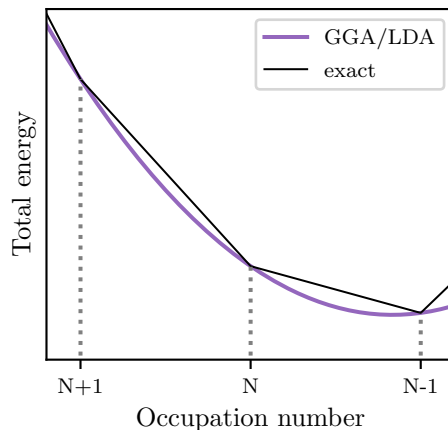
Figure 2.1: Exact and GGA/LDA total energy as a function of the number of electrons. Generally, the exact total energy is linear between points of integer numbers of electrons (black curve). For local or semi-local functionals like GGA and LDA, the curve deviates from the linear behavior for total energies of partially occupied systems.

2.7   Computational Details

The previous considerations with respect to density-functional theory were conducted on a fundamental level. In practice, one is in need of a computational scheme to actually solve the self-consistency KS problem to compute materials properties. Throughout this thesis and its corresponding publications, we used the GPAW[75] electronic structure code for our calculations.[5] Unless specified differently, all conducted calculations have been converged with respect to plane wave cutoff and k-point sampling, as well as supercell sizes for defect calculations. The details about xc functionals, choice of system sizes, and exact computational parameters are specified in the respective publications and the corresponding results sections.

---

[5]Except for Publication [III], where the Vienna Ab initio Simulation Package (VASP)[76] was used for the DFT calculations.

Theory: Thermodynamics of Point Defects in Semiconductors

The physical properties of materials are mainly determined by the structure and composition of the respective host crystal. However, there is always one entity that complicates the situation when considering a material just by looking at its unperturbed pristine form: defects. No matter how much care is taken in producing 'clean' materials, defects always seem to be present. They can occur in the form of (3D) bulk defects, (2D) surface defects, (1D) line defects, and (0D) point defects. No matter what form they have, defects always alter the properties of the host crystal which can either be detrimental or useful for specific applications.

The focus of the present work will be the characterization of point defects, *i.e.* 0D manifestations like vacancies, substitutional defects, interstitials, or single adsorbates. Over the course of recent decades, promising realizations of point defects in semiconductors suitable for quantum applications have been identified and characterized. Nowadays, not only bulk examples like the NV centre in diamond[15–18] or color centers in SiC[19–22], but also defects in 2D materials like hBN[23, 24] and MoS_2[25] play a central role in quantum information technology.

Therefore, predicting and computing the impact of point defects in crystalline solids is of utmost importance. In the following, we will present the most relevant concepts in the evaluation of thermodynamic quantities for point defects. We start by introducing the supercell approach, outline the standard approach to compute defect formation energies and charge transition levels, present Slater-Janak transition state theory for defects which is our alternative approach to defect formation energies and charge transition levels in 2D, and discuss properties of the thermodynamic equilibrium for defect systems.

Once again, we will limit ourselves to the most essential points rather than discussing everything in detail. In particular, since this thesis is mainly focusing on thermodynamic properties, we will put most emphasis on that specific aspect of point defect calculations. Freysoldt *et al.*[77] published an excellent review of the theory behind first-principle calculations of point defects which we will use as inspiration throughout this chapter. As the focus of this work is solely point defects, we will refer to 'point defects' simply as 'defects' from now on.

3.1   The Supercell Approach

Ideally, one wants to calculate defect properties for the purely isolated case, *i.e.* where a given defect is unaffected by any other defects close by. The supercell (SC) approach[78] tries to mimic this situation and to achieve mainly two things: (i) create a large distance between periodic images of a given defect and (ii) maintain periodic boundary conditions to efficiently solve the self-consistency problem with DFT. Figure 3.1 schematically shows the concept of the supercell approach. The closest distance between periodic repetitions of defects is called defect-defect distance and will be abbreviated as $d_{\mathrm{def-def}}$.

In practice, choosing the right supercell often comes with some kind of trade-off. On one hand, increasing the supercell size as much as possible is closer to the physical situation of simulating the isolated defect. On the other hand, it is advantageous to keep the supercell size (and therefore the number of atoms) small to minimize the computational effort for the DFT calculations. This is particularly important when using computationally expensive hybrid functionals. During numerous benchmark studies throughout the course of this thesis, we found that supercells with a minimum defect-defect distance of at least 15 Å typically produce well converged results for defects in two-dimensional materials as will be further discussed in the publications attributed to this thesis.
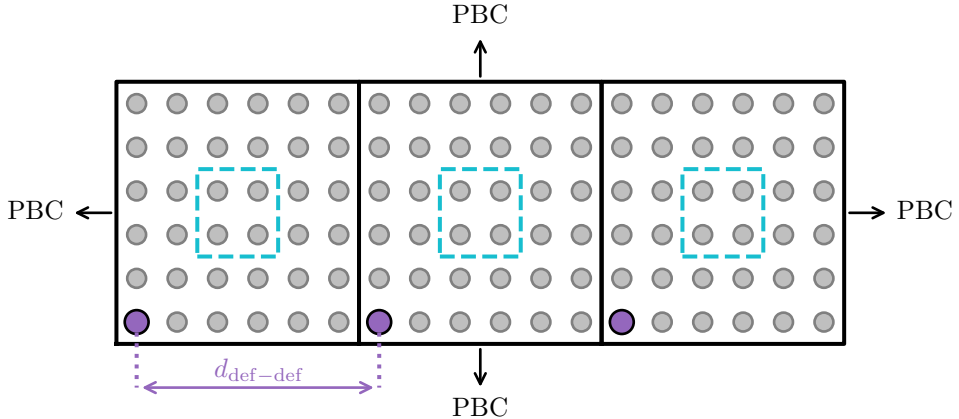
Figure 3.1: Schematic picture of defects in the supercell approach. Starting from a primitive unit cell (light blue dashed) and a specific host species (grey points), one creates a supercell, *i.e.* a 3x3 repetition (see black cell), sets up the defect in the structure (purple points), and applies periodic boundary conditions (PBC) for the correct DFT-treatment of the system. The distance between mirroring images of the defect is called defect-defect distance and abbreviated as $d_{\mathrm{def-def}}$.

3.2   Defect Formation Energies in the Total Energy Approach

Now that we have introduced the concept of the supercell approach, we can start look-ing at properties of isolated defects. First and foremost, the thermodynamic stability of a defect can give an estimate of how likely it is for a given defect to form in a compound. Let us introduce a defect X into a host crystal. For a semiconductor, the defect can occur in different charge states that will be labelled by q. If $q = 0$, the defect carries no charge, whereas for $q = -1$ ($q = +1$) an electron was added (removed). One can now write down the so-called formation energy of X^q (defect X in charge state q) as[79, 80]

$$E^f[X^q] = E_{\text{tot}}[X^q] - E_{\text{tot}}[\text{pristine}] - \sum_i n_i \mu_i + q E_{\text{F}} + E_{\text{corr}}. \tag{3.1}$$

Here, $E_{\text{tot}}[X^q]$ is the total energy of the defect in the supercell and $E_{\text{tot}}[\text{pristine}]$ the energy of the host crystal in the same supercell. In order to create the defect, one needs to add ($n_i > 0$) and/or remove ($n_i < 0$) n_i atoms of species i coming from the respective chemical potential μ_i to and/or from the pristine material. For charged defects, the chemical potential of the electrons, *i.e.* $\mu_e = E_{\text{F}}$, will make the formation energy dependent on the Fermi energy. In general, due to the electrostatic interactions between periodic repetitions of charged defects, one needs to add a correction term E_{corr} which we will discuss shortly.

Figure 3.2 shows the defect formation energy as a function of the Fermi level for a hypothetical defect. We can see that the formation energies of the charged defects vary with the slope of their respective charge states whereas E^f is constant in the neutral state. Assuming that one can freely adjust E_{F}, different charge states are most stable throughout the band gap. For the example of Figure 3.2, the defect is most stable in its $q = +1$ charge state close to the valence band maximum (VBM), and most stable in its $q = -1$ charge state close to the conduction band minimum (CBM). In between, E^f is lowest in the neutral state.

3.2.1   Charge Transition Levels

We saw that a particular defect, depending on the position of E_{F}, can be present in different charge states. Let us once again draw our attention to Figure 3.2 and assume the current position of the Fermi energy is well inside the range where $q = 0$ is the most stable charge state, *i.e.* deep inside the band gap above $\varepsilon(+1/0)$ and below $\varepsilon(0/-1)$.

If we raise the Fermi level up to the point where $q = -1$ becomes the most stable charge state, we pass a special point where the formation energy of the neutral and negatively charged defect are equal. This point is called charge transition level (CTL) between charge states $q = 0$ and $q = -1$ and will be denoted as $\varepsilon(0/-1)$. Physically speaking, in order to increase the charge of a system by one, one needs to occupy an empty state. Therefore, $\varepsilon(0/-1)$ indicates the position of E_{F} corresponding to that empty state. Similarly, by lowering E_{F} in Figure 3.2 down until $\varepsilon(+1/0)$ one will hit an
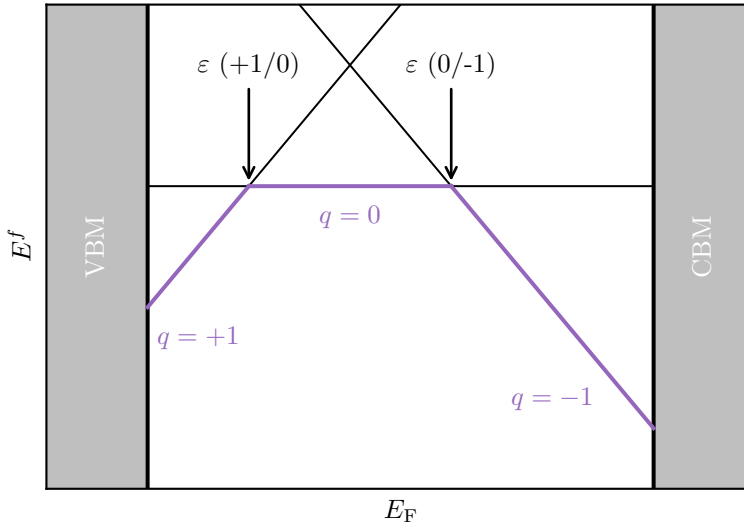
Figure 3.2: Formation energy diagram for a hypothetical point defect in a semi-conductor. Formation energy E^f as a function of the Fermi energy E_F within the pristine band gap (VBM and CBM represent the edges of valence band and conduction band, respectively). Depending on the position of E_F, the defect can exist in different charge states: the single positively charged defect ($q = +1$), the neutral defect ($q = 0$), and the single negatively charged defect ($q = -1$). The purple lines represent the path of the lowest formation energy with the two charge transition levels $\varepsilon(+1/0)$ and $\varepsilon(0/-1)$ indicated by the arrows.

occupied state that will be emptied out, and $q = +1$ becomes the most stable charge state for $E_F < \varepsilon(+1/0)$. One can generally find the CTL for a given charge transition via[77]

$$\varepsilon(q_1/q_2) = \frac{E^f(X^{q_1}; E_F = 0) - E^f(X^{q_2}; E_F = 0)}{q_2 - q_1}. \tag{3.2}$$

3.2.2 Correction Term for Electrostatic Interactions

As commonly known, the electrostatic energy of a system with non-zero charge diverges[77, 81, 82]. Consequently, when calculating properties of charged defects in DFT, one needs to include a background charge into the system to compensate the defect charge and make the system neutral overall. What one ends up with is an unphysical electrostatic interaction between the defect charge, the charges of its periodic repetitions, as well as the background charge. It is vital to systematically correct for this unphysical electrostatic interaction in order to obtain reliable results for charged defect calculations.[1]

---

[1]The interested reader can find a good review of the different types of correction schemes in Ref. [83].

Different useful *a posteriori* correction schemes have been proposed[81, 82, 84].[2] A correction scheme developed by Freysoldt, Neugebauer, and van de Walle (FNV)[87] is based on modelling the charge as a localized Gaussian charge distribution around the defect. To correct for the unphysical electrostatic interaction, one calculates the interaction of the charge distribution with its periodic images inside a dielectric medium. Furthermore, the average electrostatic potential of pristine and defect supercell need to be aligned[87]. The FNV scheme shows good accuracy[88] and is a widely used method in the field of first-principle defect calculations. It can be directly applied to defects in bulk materials, is implemented in the GPAW[75] electronic structure code, and will be used in this thesis whenever defects in 3D bulk materials are considered.

However, complications arise for 2D materials. For instance, the dielectric constant is ill-defined in two dimensions and makes calculating the correction term challenging. Efforts by Komsa *et al.*[89] model the dielectric constant of a two-dimensional material to obtain supercell size independent values for charged formation energies. Unfortunately, we were not able to reproduce supercell size independent values with the charge correction scheme for defects in 2D materials implemented in GPAW[75]. In particular, charged formation energies as a function of out-of-plane supercell size, *i.e.* different amounts of vacuum above and below the monolayer, did not converge. Therefore, in this work, we apply an alternative methodology to compute formation energies and CTLs for monolayer defect systems as will be discussed in Section 3.3 on Slater-Janak transition state theory for defects.

3.2.3 Chemical Potentials

In Equation (3.1) we saw that the formation energy of a defect is dependent on the chemical potential of atomic species of a defect's constituents. But how should one choose the respective chemical potential μ_i? Let us assume that we want to incorporate a defect of type A_B (atom B is replaced by atom A) into a host crystal A_xB_y (x and y define the stoichiometry of the host compound). In thermodynamic equilibrium, neglecting vibrational and entropic contributions, the relation[77, 90]

$$\Delta H_f(A_xB_y) \leq x\Delta\mu_A + y\Delta\mu_B \tag{3.3}$$

holds.[3] Here, $\Delta H_f(A_xB_y)$ denotes the heat of formation of the host crystal and $\Delta\mu_i = \mu_i - \mu_i^{\text{ref}}$, where μ_i^{ref} is the chemical potential of the reference phase for species i. The heat of formation is a negative number for stable compounds and $\Delta\mu_i \leq 0$. For finding the boundary of the phase space defined in Equation (3.3), one can equate the expression and calculate the chemical potentials μ_A, μ_B in the A-rich limit (*i.e.*

---

[2]More recently, methodologies based on altering the actual self-consistent DFT calculation and avoiding *a posteriori* corrections have been developed[85, 86].

[3]The interested reader can find an example of temperature and pressure dependent chemical potentials including vibrational and entropic effects in Ref. [91].

$\Delta\mu_A = 0$) as

$$\mu_A = \mu_A^{\text{ref}},\tag{3.4}$$

$$\mu_B = \frac{\Delta H_f(A_x B_y)}{y} + \mu_B^{\text{ref}}.\tag{3.5}$$

The A-rich condition (*i.e.* μ_A takes its maximum possible value) corresponds to a B-poor condition where μ_B is effectively reduced as both $\Delta H_f(A_x B_y)$ and μ_B^{ref} are negative numbers. Similarly, one can obtain the chemical potential for the B-rich case by simply setting $\Delta\mu_B$ to zero and obtaining μ_A from Equation (3.1). The different choices of chemical potentials result in an effective up/down-shift of the formation energy curve depicted in Figure 3.2 which in turn affects the self-consistent Fermi level position (see Section 3.4). The charge transition levels, however, are unaffected by changes to the chemical potentials.

3.3   Slater-Janak Transition State Theory for Defects

Within the defect community, using Equation (3.1) combined with Equation (3.2) is the standard for obtaining formation energies and CTLs in bulk systems. However, in two dimensions, handling the correction term E_{corr} in Equation (3.1) becomes a very challenging task. We therefore rely on a different strategy, namely Slater-Janak (SJ) transition state theory for defects, a strategy that was previously applied to defect systems[92–95]. Generally, for the concept of SJ, we saw in Equation (2.17) that the transition energy between a N and $N + 1$ electron system can be expressed in terms of an integral over the highest occupied orbital, *i.e.* ε_H, of the system with occupation number η. SJ transition state theory has been the basis for the calculation of thermodynamic properties in the QPOD project[II] and other studies for defects in two dimensional materials.

3.3.1   Slater-Janak Formalism for Charge Transition Levels and Formation Energies

Before diving into the details of the SJ formalism for defects, let us change from the picture of a system's number of electrons N to the actual charge q. We consider the system with N electrons as the 'neutral' starting point with $q = 0$.[4] By exploiting that ε_H varies linearly with the occupation number η and by evaluating the integrand in Equation (2.17) at $\eta = 0.5$, one can write down the formula to obtain the CTL $\varepsilon(q/q')$ between two charge states q and $q' = q \pm 1$ as

$$\varepsilon(q/q') = \begin{cases} \varepsilon_H\left(q + \frac{1}{2}; R_q\right) - \lambda_{q'}, & q' = q + 1 \\[2mm] \varepsilon_H\left(q - \frac{1}{2}; R_q\right) + \lambda_{q'}, & q' = q - 1. \end{cases}\tag{3.6}$$

---

[4] Adding (removing) an electron to (from) the system will effectively change the charge of the system from $q = 0$ to $q = -1$ ($q = +1$).

We see that one can calculate $\varepsilon\left(q/q'\right)$ by extracting the energy of the highest occupied orbital of the system ε_{H} with half integer charge $q \pm 0.5$ at the atomic configuration of charge state q (*i.e.* R_q) and by subtracting (adding) the reorganization energy

$$\lambda_{q'} = E_{\mathrm{tot}}\left(q'; R_{q'}\right) - E_{\mathrm{tot}}\left(q'; R_q\right). \tag{3.7}$$

Here, $E_{\mathrm{tot}}\left(q'; R_q\right)$ represents the total energy of the system with charge q' evaluated at the atomic configuration of charge state q, *i.e.* R_q. In practice, for evaluating $\varepsilon_{\mathrm{H}}\left(q \pm \frac{1}{2}; R_q\right) \mp \lambda_{q'}$, we conduct a self-consistent DFT calculation with half-integer charge. Figure 3.3 schematically visualizes the idea behind the SJ approach for defects and sketches its central components.

Slater-Janak transition state theory hereby enables us to calculate the charge transition levels of defect systems. In addition, for neutral defects one can drop the last two contributions in Equation (3.1), *i.e.* qE_{F} and E_{corr}, and simply compute the neutral formation energy $E^f[\mathrm{X}^{q=0}]$ without any corrections. In a second step, this enables us to combine $E^f[\mathrm{X}^{q=0}]$ with its CTLs to obtain the formation energies for charged defects. To summarize, we look at the general differences in the approach for the total energy scheme vs. the SJ scheme:

- Total energy scheme: $E^f[\mathrm{X}^q] \rightarrow \varepsilon(q/q') \rightarrow$ TD defect properties,

- Slater-Janak: $E^f[\mathrm{X}^{q=0}]$ & $\varepsilon(q/q') \rightarrow E^f[\mathrm{X}^q] \rightarrow$ TD defect properties.

While the total energy scheme relies on the calculation of formation energies for charged defects by comparing total energies of systems with different charges, SJ avoids this step as it uses total energy considerations only for the calculation of the neutral formation energy of the defect (which does not need electrostatic corrections). Thus, everything related to charged states is extracted directly from the KS spectrum of the system.

3.3.2   On the Accuracy of Slater-Janak for Defects

We have established that SJ transition state theory for defects bypasses the comparison of total energies for systems of different net charge. In addition, for calculating the CTLs, we reference the KS eigenstates of the neutral and charged systems to the average electrostatic potential associated with an atom far away from the defect itself. Extracting the average electrostatic potential for the corresponding atom in the pristine supercell makes it possible to obtain the relative position of the defect CTLs with respect to the VBM of the host crystal. Thereby, we are able to bypass the addition of an electrostatic correction to the SJ method. We verified that the resulting charge transition levels converge well with respect to supercell size (both in-plane and out-of-plane) as we showcase as part of the QPOD study[II] (see Section 6.2).
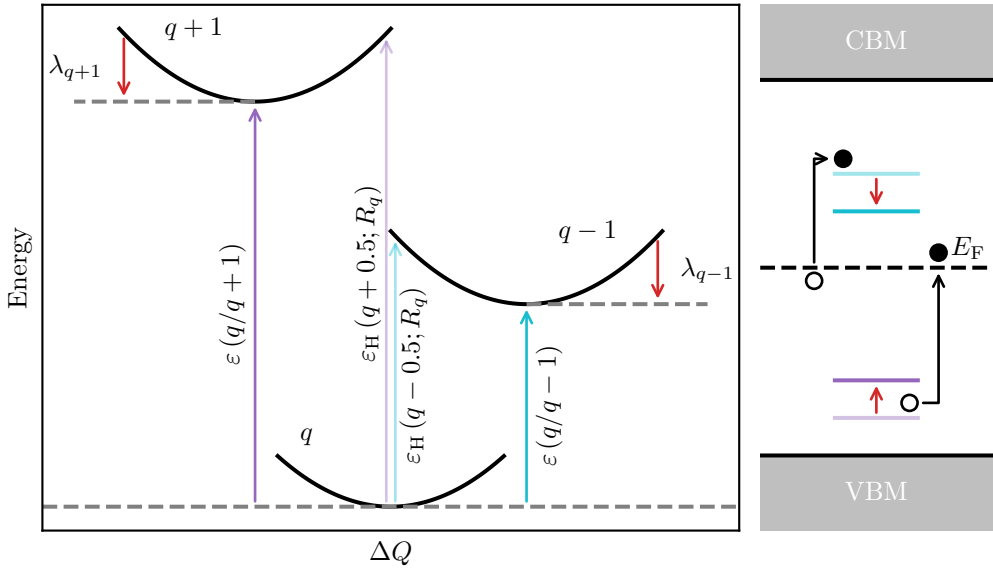
Figure 3.3: The SJ approach for defect systems. Left panel: schematic diagram of the energy *versus* the change in configuration coordinate ΔQ. In SJ, an electron gets promoted (removed) from q to $q \pm 1$ in a vertical transition. The transitions $\varepsilon_{\mathrm{H}}\left(q \pm \frac{1}{2}; R_q\right)$ are the so-called optical charge transition levels (no relaxation effects included). Afterwards, the $q \pm 1$ system undergoes a structural reorganization $\lambda_{q\pm1}$ (red arrows) to relax from the atomic configuration of R_q to its new minimum energy structure. The new 'corrected' charge transition level $\varepsilon\left(q/q \pm 1\right)$ is known as the thermodynamic CTL. Right panel: KS picture where the reorganization (red arrows) effectively raises $(q + 1)$ or lowers $(q - 1)$ the optical CTLs with respect to the VBM. Figure adapted from Publication [II].

There is, however, one source of error included: SJ assumes that the total energy as a function of occupation number N is piecewise linear between integer values of N (see Figure 2.1 of the DFT chapter) which is not the case when using GGA or LDA functionals. We are willing to accept this source of error as we are limited by the band gap underestimation and delocalization error within PBE[96] for our calculations.

3.4 Thermodynamic Equilibrium and Fermi Level

We are now able to compute charge transition levels and defect formation energies for different chemical potential conditions both in 3D bulk (by using the total energy scheme) and in 2D monolayer (by using SJ transition state theory) materials. Just like for a pristine semiconductor, the Fermi level for a defect system in a particular charge state is always located somewhere in between the highest occupied and lowest unoccupied state. In real materials, however, several defects of different types coexist inside the same host crystal and raise the question: how can one evaluate the overall position of the Fermi level for the entire material? The solution is represented by a

self-consistent evaluation of the position of the Fermi level which we will introduce in the following. For our implemented procedure, we took inspiration from Ref. [97].

Let us consider a material with several defects present. For a given defect X in charge state q, the concentration at temperature T is connected to the formation energy *via*

$$C[\mathrm{X}^q] = N_\mathrm{X} g_{\mathrm{X}^q} \exp\left(-E^f\left[\mathrm{X}^q\right]/(k_\mathrm{B}T)\right). \tag{3.8}$$

Here, N_X and g_{X^q} denote the site and state degeneracy of the particular defect, respectively, and k_B is the Boltzmann constant.[5] As one would intuitively expect, a low formation energy corresponds to a high defect concentration and *vice versa*. We further know that the host material including all of its defects has to satisfy the charge neutrality condition

$$\Delta = \sum_\mathrm{X} \sum_q qC[\mathrm{X}^q] - (n_0 + p_0) = 0, \tag{3.9}$$

where n_0, p_0 denote the electron and hole carrier concentrations, respectively, and the sums run over all defect species X in their charge states q. Both n_0 and p_0 are obtained through the local density of states of the host crystal $\rho(E)$ *via*

$$n_0 = \int_{E_\mathrm{gap}}^{\infty} dE f(E)\rho(E), \tag{3.10}$$

$$p_0 = \int_{-\infty}^{0} dE[1 - f(E)]\rho(E). \tag{3.11}$$

We note, that the energy is referenced to the valence band maximum, *i.e.* $E_\mathrm{VBM} = 0$ and $E_\mathrm{CBM} = E_\mathrm{gap}$ (pristine band gap), and the Fermi-Dirac distribution is given by $f(E) = \exp\left[(E - E_\mathrm{F})/(k_\mathrm{B}T)\right]^{-1}$. With the ingredients introduced in Equations (3.8) to (3.11), we are now able to self-consistently search for the Fermi level position E_F^sc at which the charge neutrality in Equation (3.9) is guaranteed. The algorithm to search for the self-consistent Fermi level based on a specific host material and its defects has been implemented into the ASR[V] framework as part of this thesis (more details see chapter 5). Figure 3.4 shows the respective output for MgH_2S_2 (part of the QPOD[II] host material set) in two different chemical potential conditions: S-poor and Mg-poor.

For the S-poor case, the self-consistent Fermi level position E_F^sc is located closer to the conduction band compared to the Mg-poor chemical potential condition. This makes intuitive sense as the lowest lying charged defect, *i.e.* Mg_H (blue solid line) in charge state $q = +1$, transfers a high concentration of charge carriers from the defect into the conduction band, pushing the Fermi level close to the conduction band edge. On the other hand, in the Mg-poor case, Mg_H is slightly higher in energy which results in E_F^sc being pinned further away from the conduction band edge. These Fermi level positions make the material intrinsically *n*-type resulting in electron concentrations as high as 2×10^{11} cm^2 (S-poor) and 7×10^5 cm^2 (Mg-poor). As part of the QPOD project[II], we calculated the self-consistent Fermi-level position, electron and hole carrier concentrations, and equilibrium defect concentrations within different chemical potential conditions for more than 50 host materials.

---

[5]Currently, site and state degeneracy are simply set to one, but in an upcoming version of the implementation we are planning to extract state degeneracies from the symmetries of defect states.
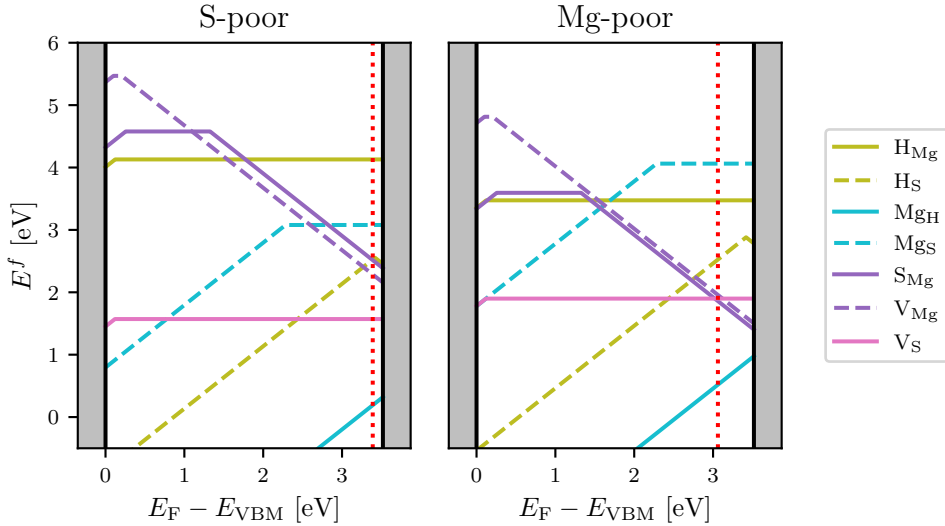
Figure 3.4: Fermi level pinning in MgH$_2$S$_2$ for S- and Mg-poor conditions. Formation energy E^f as a function of the Fermi energy inside the pristine band gap for seven defects inside the MgH$_2$S$_2$ host crystal. The self-consistent Fermi level position $E_{\mathrm{F}}^{\mathrm{sc}}$ is marked with the red dotted line. Left: S-poor conditions, right: Mg-poor conditions.

3.5   Going Beyond Thermodynamic Properties

It is important to stress that there is more to characterizing point defects than just computing their thermodynamic and electronic properties. On the pathway towards utilizing defect systems for quantum technological applications, one needs to also look at properties beyond thermodynamics (which mainly give an insight into a defect's stability, and can give an intuition of opto-electronic and transport behavior). To conclude this chapter, we will give a brief overview of some of those other properties and discuss their relevance in terms of applications.

First, to correctly address the optical properties of a defect, which is of high relevance for applications related to single-photon emitters (SPEs)[98], one has to take the effects of electron-phonon coupling into account. Finding a good candidate for a single-photon source is typically tied to some prerequisites. A SPE needs to (i) be stable over time, (ii) have a high emission rate, (iii) emit photons of a well-defined wavelength, and (iv) be indistinguishable.[98] The effects of electron-phonon coupling define the emission spectrum[99], as well as the radiative and non-radiative recombination rates[100] and are therefore vital in characterizing any SPE source. A detailed review of the processes involved is given in Refs. [101, 102] as well as in the PhD thesis of S. Manti[103].

Second, within the field of quantum information technology, it is commonly known that high-spin defects can act as qubits[104–108]. This makes a defect characterization in terms of magneto-optical properties essential. For instance, the hyperfine interac-

tion makes it possible to control qubits[18, 109]. Third, carrier recombination often limits photovoltaic (PV) efficiencies making the accurate description of carrier capture kinetics particularly important for PV applications[110].

An in-depth characterization of point defects in terms of these 'additional' defect properties is beyond the scope of this thesis. However, our high-throughput implementations for defect studies (see Chapter 5) make a natural extension towards a more detailed framework possible. Extending the existing developments should be the goal for future projects.

Excursion: High-Throughput Methods in Materials Science

Generally speaking, the field of physics can roughly be split up into three areas: experimental physics, theoretical physics and computational physics (see Figure 4.1). While experiments give valuable insight by actually measuring physical quantities, the theoretical counterpart tries to build models and theories to understand the underlying concepts in mathematical terms. Computational physics, marking the youngest of the three areas, solves numerical models with simulations and tries to correctly predict the behavior of specific systems.

In this chapter, we will build the bridge from computational physics to automation and high-throughput (HT) methods in materials science. We will discuss the steps needed to achieve automation, present common tools, and introduce the Atomic Sim-

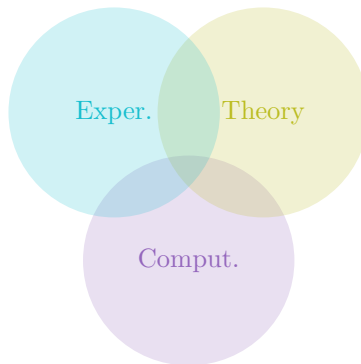**Figure 4.1: The three areas of physics: experiment, theory, and computations.** The work for this thesis can be grouped into the computational branch but through our findings and developments we hope to create synergy between the different fields.

ulation Recipes (ASR)[V], a joint effort of the CAMD section to create an easy-to-use Python framework to conduct HT simulations in materials science. Throughout the chapter, we will try to translate the general concepts to the specific application of defect calculations.

4.1 Automation and High-Throughput Methods in Materials Science

Technological advances have always been closely tied to materials: the stone age, the bronze age and the iron age are prime examples of prehistoric eras where the underlying material made it possible to completely transform the way of living. More recently, silicon and its semiconducting behavior shaped the development of information and communication technologies. Throughout the different eras, there was always one common denominator: with new technological challenges arising, the need for new materials is growing. But how to go about finding new materials with properties of technological impact? The space of possible candidates is vast and characterizing each and every one of them experimentally is impossible.

Computational materials design enters the stage and is one possible approach to tackle the challenge of finding new promising materials. While computational methods become more and more efficient, computer resources become more and more powerful, and enable researchers to characterize an increasing number of systems. However, in spite of the recent computational improvements, the number of possible candidates is simply too large to characterize them manually and one by one. Consequently, materials science is in need to automate the computational characterization process and to incorporate high-throughput methods. In the last decades, several successful examples of computational high-throughput studies based on first-principle calculations have emerged. Examples range from cathode materials for lithium batteries[111] to the calculation of thermoelectric properties[112] and the search of new topological materials[113] to only name a few.[1]

4.2 Tools for High-Throughput Materials Science

With emerging high-throughput studies, one inevitably is in need of suitable frameworks and tools to easily and efficiently control and execute vast studies focusing on thousands of materials at the same time. Within recent years, several efforts successfully made materials science more reproducible and paved the way towards full data provenance by developing powerful high-throughput frameworks. Some common examples with particular relevance for materials science are AiiDA[115, 116], FireWorks[117], AFLOW[118], *httk*[119, 120], or GC3Pie[121].

---

[1]Many more examples can be found in the literature. Curtarolo *et al.*[114] review the *high-throughput highway to computational materials design* in great detail.

The Atomic Simulation Recipes [V] are of particular importance for the results high-lighted in the present thesis. ASR is an open source Python framework that aims to simplify atomistic simulations and enable users to conduct calculations in a high-throughput manner. In contrast to existing powerful frameworks like AiiDA[115, 116] and FireWorks[117], ASR's design philosophy evolves around simplicity and flexibility. In ASR, a central building block is the concept of a Recipe. A Recipe has a specific unique task (*e.g.* conduct a ground state calculation, obtain the bandstructure, calculate the density of states, *etc.*) and ASR represents a library of numerous useful Recipes combined with core functionalities to ensure data provenance and reproducibility. The details of ASR can be found in Publication [V] (see page 131), its most important aspects and applications are discussed in Section 6.5.

In the context of this thesis, where the focus lies on the properties of point defects, it is important to mention the existing efforts in the field of high-throughput tools for defect calculations. Several useful frameworks exist and examples like PyCDT[122], PyDEF[123], ADAQ[124], and others[125, 126] make it possible, at least in general, to conduct defect calculations for hundreds and thousands of defects. We tried to add to those efforts by extending the ASR library with numerous Recipes specifically designed for defect calculations. An overview and summary of those specific Recipes is given in Section 5.1. We note, that a large amount of work for this thesis went into designing, implementing, and maintaining the defect Recipes in ASR.

4.3   High-Throughput Workflows

High-throughput tools translate abstract concepts of automation into versatile frameworks applicable to various types of simulations. However, in order to compute specific properties one needs to tell the high-throughput tools what and how they should execute tasks on a given computer resource. The concept of workflows builds the bridge between a user, that wants to address a specific computational problem, and the HT frameworks, that make it possible to orchestrate thousands and millions of specific tasks. A workflow incorporates all of the information that can not be abstracted inside the HT framework itself.

Figure 4.2 visualizes a simplified representation of how a user, workflow, workflow manager (*i.e.* high-throughput framework), and the computing resource generally interact: the developer implements a low-level logic that computes certain properties utilizing an underlying simulation code. The resulting separate building blocks should be designed in a modular way such that they can be reused in many different contexts and are easy to test. Once these modular building blocks are in place, the developer can translate the overall scientific question into a workflow that utilizes the separate building blocks and combines them in a high-level logic. Subsequently, the workflow manager and HPC scheduler facilitate the execution of tasks on the computer resource and stores the results, typically in a database.

Specifically within the ASR framework, the modular building blocks in Figure 4.2 correspond to the ASR Recipes. Once the logic behind the desired workflow is established and implemented, MyQueue[127], a task and workflow scheduling system, connects to
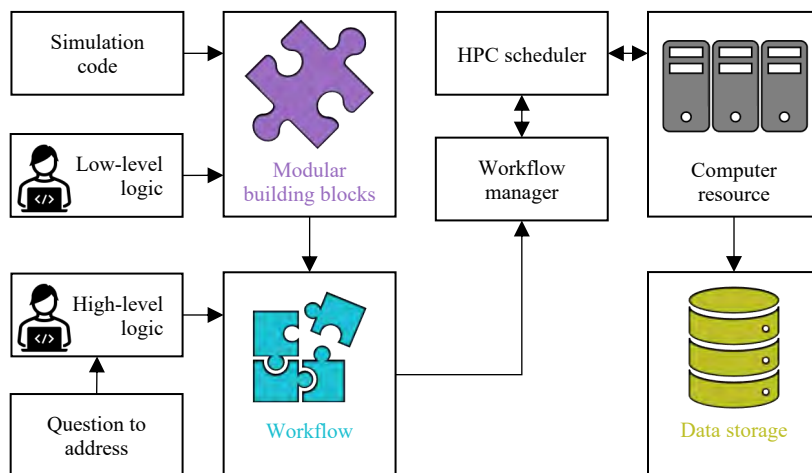
Figure 4.2: Schematic visualization of workflow-centered frameworks in high-throughput projects. The developer can create low-level modular building blocks by wrapping simulation code and his own logic for a specific task. The numerous building blocks and some kind of high-level logic can be combined into a workflow that can be executed on a computer resource. The execution of jobs on the computer resource is handled by workflow manager and high performance computing (HPC) scheduler. Finally, the data is stored, for example in a database.

the HPC scheduler and handles the correct execution of tasks. The resulting data is collected and stored using core functionalities of ASR. We have now seen how abstract concepts within HT frameworks can be translated to address a specific computational question *via* workflows. Next, we will discuss the importance of making data accessible in databases, in particular in the context of defect physics.

4.4   Databases and Machine Learning

We have now introduced all the necessary concepts and tools needed to systematically generate high-throughput data for materials science. During the last decades, numerous high-throughput studies were conducted and resulted in the generation of large materials databases with thousands to millions of entries. The most prominent and largest examples are Materials Project[128], AFLOW[118], as well as the open quantum materials database (OQMD)[129]. Furthermore, studies to generate high-quality data for 2D materials have been published[130, 131]. Introducing defects into a given system opens up a new dimension in the material space as we will discuss in the subsequent section. Even though numerous theoretical *ab initio* characterizations of defects exist, high-throughput studies on the topic are scarce, especially in the realm of two-dimensional materials.

We therefore made it our goal to drive efforts towards systematically studying defects

in low-dimensional materials and, by means of high-throughput methods, to generate tools that make cumbersome defect calculations easily accessible, and to establish publicly available DFT defect data in the form of open-access databases. These efforts can be particularly useful in connection with machine learning (ML) techniques: as defect calculations in the supercell approach are computationally demanding, one central goal of the defect community needs to be to work towards supplying a sufficient amount of data for enabling machine learning methods to step in. With ML, DFT calculations could be skipped to screen for promising and/or particularly important defects in the vast material space. Subsequently, those particular systems can be subject to detailed and in-depth characterizations by means of accurate DFT calculations to eventually find new defect candidates of technological relevance.

4.5   The Materials Space for Defect Calculations

The goal of the present thesis is to systematically characterize defects in semiconductors. Defects introduce a new dimension to the materials space. For every single material inside that space, one can create numerous defects, and each individual defect has to be characterized in its occurring charge states. The exact number depends on how complex the defects are allowed to be.

To give an example, let us focus on the MoS_2 crystal in Figure 4.3: the simplest types of defects are represented by single intrinsic point defects, namely vacancies and substitutions of intrinsic element type (V_S, V_{Mo}, Mo_S, S_{Mo}). This class of defects has been the basis for the QPOD study[II] (details see Section 6.2). Next, one can extend those single intrinsic point defects to extrinsic and double defects. Numerous possibilities to choose extrinsic dopant elements, as well as combinatorial freedom in the choice of double defects increases the materials space to a large extent. For instance, when doping MoS_2 with 25 extrinsic elements[2], only 54 out of 493 (roughly 11 %) of the generated defects are single defects and the rest can be attributed to double defects. The four simple intrinsic defects V_S, V_{Mo}, Mo_S, and S_{Mo} make up less than 1 % of that space!

The last type of defects, namely interstitials, are somewhat special: in contrast to vacancies and substitutions, which are uniquely defined by the atomic positions of the host crystals, interstitials do not have a well defined position as they are generally located at 'free spaces' inside the host crystal. As a result, there are many possibilities to place an interstitial defect. As part of this thesis we implemented a systematic strategy to choose reasonable characterizations of interstitials which will be discussed in Section 5.2 and Publication [III]. There's no limit as to how complex the defects can become[3], but since one eventually needs to run expensive DFT calculations for each one of those defects in their respective supercells, it is reasonable to focus on the most important ones.

---

[2]Defect pairs with a distance of 1.7 times the sum of covalent radii of defect species involved are accepted as new double defects. We refer to the source code of `asr.setup.defects` for more details: https://gitlab.com/asr-dev/asr/-/blob/master/asr/setup/defects.py.

[3]One needs to be careful, however, to not let the defect complexes become to large as the extension

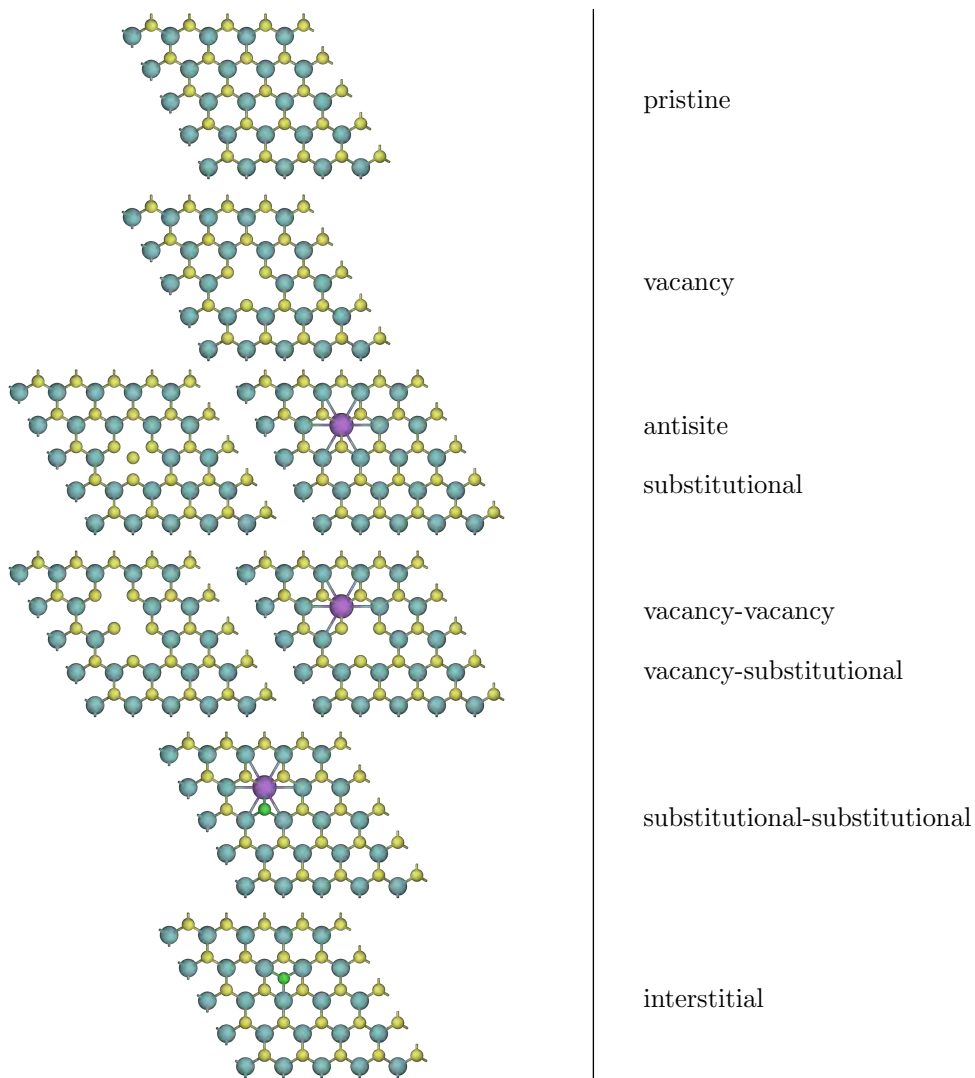**Figure 4.3: Overview of single and double point defects in MoS$_2$.** From top to bottom: pristine MoS$_2$ crystal in a 5x5x1 supercell, single vacancy defect, antisite defect (left) and an extrinsic single substitutional defect (right), vacancy-vacancy (left) and vacancy-substitutional double defects (right), substitutional-substitutional double defect and, lastly, an example for an interstitial defect.

In many cases only looking at intrinsic defects is a good first approximation as those are typically not controllable and unavoidable due to the constant presence of intrinsic elements to form the defect. As a result, when looking at the impact of defects on a specific host material, one should not exclude intrinsic defects. This hypothesis was confirmed by the outcome of the QPOD study[II] and the presence of intrinsic defects, whether desired or not, has to be taken into account.

In addition, it became evident that finding suitable defects for quantum applications within the set of intrinsic defects is challenging. Extrinsic defects can be partly controlled by the synthesis conditions and define a suitable platform to tune host crystal properties by deliberate doping. In a spin-off study to the QPOD project[II], which is beyond the scope of the present thesis, we try to use this strategy to screen for promising defect systems in 2D materials in the context of quantum applications. In particular, we took an initial host crystal set of experimentally known monolayers and systematically incorporated single and double intrinsic and extrinsic defects. Even though this study is still running and not complete, we already obtained the preliminary result that including double defects yields a higher relative number of triplet systems that might be of high importance for quantum applications.

Adatoms define yet another type of point defects for the specific case of surfaces and 2D materials. They are similar to interstitial defects as they are located at positions that were not previously occupied in the pristine host crystal. Using the functionalities of the ASE `DefectBuilder`, which will be introduced in Section 5.2, we conducted a high-throughput study on extrinsic doping in interstitial and adsorption sites in experimentally known 2D monolayers. The details and outcomes of this study will be discussed in Section 6.3 as well as in Publication [III].

We now know what the materials space for point defects in crystalline systems looks like in theory. The next chapter changes gears and focuses on practical solutions of defect system setup and property characterization.

---

of the defect should be small compared to the size of the supercell.

CHAPTER 5

Tutorial: A Short Guide on Defect Utilities

In the previous chapter we saw how high-throughput methods can generate large amounts of data to advance materials discovery, specifically in the context of defect calculations. Thus, every high-throughput study should be built on top of a stable workflow which combines the small modular building blocks and adds some kind of logic on top of it. Developing these small building blocks in the most general way is of utmost importance to ensure their successful and fault-free execution within a computational workflow.

Due to the modular design of the individual building blocks, one can use them for HT purposes, as well as for applying them on single systems. In this chapter, we will present some of the building blocks that have been developed as part of this thesis. We will discuss the underlying logic, describe possible use-cases, and showcase how they helped in conducting studies of diverse nature. Specifically, we will give an overview of defect-related functionalities within ASR in Section 5.1 which have been the basis for the QPOD study[II]. Furthermore, we present the ASE `DefectBuilder`, a useful module to set up various defect systems within the Atomic Simulation Environment (ASE)[46], which helped us to generate the set of materials for Publication [III] and to draw valuable conclusions on adsorption and absorption processes in doped 2D materials.

5.1 Defect Recipes within the ASR Framework

Within ASR, the aforementioned building blocks for a workflow are represented by ASR Recipes. A major effort for the present thesis was the development, implementation, and maintenance of Recipes designed for the usage on defect systems (we will call those 'defect Recipes' in the following). The defect Recipes are summarized in Table 5.1.

For each Recipe, which is basically a Python module, the user can pass certain options and arguments to conduct the desired task in a particular fashion. After correct execution of the Recipe, a visual representation of the results is created *via* built-in ASR functionalities (*e.g.* formation energy diagram for `asr.sj_analyze`, states in the gap with symmetry labels for `asr.defect_symmetry`, formation energy diagrams with self-consistent Fermi-level in different chemical potentials for `asr.charge_neutrality`).[1]

---

[1]The source code of the Recipes as well as usage instructions can be found in the ASR repository: https://gitlab.com/asr-dev/asr/-/tree/old-master.

Table 5.1: List of ASR defect Recipes. The Recipes are grouped into four application categories: defect setup, thermodynamic properties, defect states and symmetries, magneto-optical properties, and metadata helper utilities. Recipe names always start with the `asr.-` prefix.

Recipe name	Description
Structure setup	
`asr.setup.defects`	Set up defect structures in a supercell.
Thermodynamics	
`asr.sj_analyze`	Calculate form. en. & CTLs with SJ theory.
`asr.charge_neutrality`	Calculate SC Fermi level for defect systems.
States and symmetries	
`asr.defect_symmetry`	Analyze structural and defect state symmetries.
`asr.get_wfs`	Save WFs of defect states in the gap.
`asr.tdm`	Calculate transition dipole moments.
Magneto-optical	
`asr.hyperfine`	Calculate hyperfine coupling.
`asr.zfs`	Calculate zero-field-splitting.
Helper utilities	
`asr.defectinfo`	Collect information on defect system.
`asr.defectlinks`	Create hyperlinks between related DB entries.

We will now give a brief overview of the different defect Recipes and present their most important functionalities:

`asr.setup.defects`   Collection of utilities to set up defect structures in a supercell. The Recipe takes a primitive (defect free) structure as an input, creates numerous kinds of point defects in different charge states, and sets up a folder structure that makes it possible to automatically characterize the defect system within the ASR framework. Currently, the possible types of defects are: single vacancies, antisites, single substitutional defects, double vacancies, substitutional defect-vacancy pairs, substitutional defect pairs. A list of dopant atoms can be given as an input to include extrinsic defect elements. Additionally, the Recipe automatically sets up suitable defect supercells with (i) a simple $N_1xN_2xN_3$ repetition of the primitive cell, (ii) a NxNxN repetition[2] that ensures a given input minimum defect-defect distance $d_{\mathrm{def-def}}$, or (iii) a symmetry-broken supercell (details discussed in Publication [II], page 73).

`asr.sj_analyze`   Calculates formation energies and charge transition levels by utilizing Slater-Janak transition state theory. Once the structures of the defect in different charge states are relaxed, and the ground states for the half-integer calculations are

---

[2]NxNx1 for two-dimensional systems.

found, this Recipe collects the correct GPAW[75] outputs, computes the charge transition levels for a given defect according to Equation (3.6) as well as its neutral formation energy. Combining CTLs and neutral formation energy finally yield the charged formation energies in the Slater-Janak approach. From the results, the Recipe automatically plots the formation energy diagram for the defect of interest.

asr.charge_neutrality   Evaluate self-consistent Fermi level for a host crystal with numerous defects. The Recipe takes the formation energies of a list of defects as an input (either manually or from the results of `asr.sj_analyze`), together with a converged ground state density of the pristine host crystal. It then self-consistently obtains the equilibrium Fermi level position at a given temperature according to the methodology described in Section 3.4. This procedure is conducted for the different limiting chemical potential conditions. A full formation energy diagram with all defects, as well as the self-consistent Fermi-level is produced automatically.

asr.get_wfs   Identifies defect states in the gap and writes their wavefunctions to files. This Recipe references the eigenstates of the defect system to the band edges of the pristine supercell. In addition, for applications outside the defect context, the Recipe can write out wavefunctions from a given input energy range around the Fermi level making it a useful extension to ASR for numerous purposes.

asr.defect_symmetry   Analyzes the structural and electronic symmetry of a defect system. Using the defect wavefuntions obtained from `asr.get_wfs`, the Recipe analyzes the symmetry of the defect states themselves by finding the irreproducible representations of the Kohn-Sham states. We note that the theory behind the implementation of the symmetry analysis is not a part of the present thesis and the conceptual work was carried out in the PhD thesis of S. Manti [103]. For each defect state, the Recipe outputs its irreproducible representation, localization ratio and accuracy for the symmetry mapping. Further details can be found in Publication [II] (see page 75).

asr.tdm   Calculates the transition dipole moment between Kohn-Sham eigenstates. This Recipe can be used (i) in defect-mode, where it includes all of the transitions between defect states inside the pristine band gap, or (ii) in standard-mode, where it simply includes a set of input states as the basis for possible transitions. The Recipe is built on top of an existing functionality in GPAW[75], more details can be found in Publication [II] (see page 76).

asr.hyperfine   Calculates the hyperfine coupling for a given system. This Recipe wraps existing GPAW[75] functionality into the ASR framework to calculate isotropic and anisotropic hyperfine coupling paramters. The conceptual work was done by S. Ali as part of Publication [II] (see page 76) and further details about the theory behind the hyperfine implementation can be found in Refs. [132, 133].

asr.zfs   Calculates the zero-field-splitting for a given system. This Recipe wraps existing GPAW[75] functionality into the ASR framework. The conceptual work for

the theory behind the implementation was done by S. Ali as part of Publication [II] (see page 77).

`asr.defectinfo` & `asr.defectlinks`   Recipes to link entries of ASR generated defect databases, *e.g.* the QPOD[II] database, and to collect general information about defect systems and their respective pristine host crystals. These functionalities are mostly relevant for creating an easily browsable web-interface of defect databases inside the ASR framework.

5.2   The ASE DefectBuilder: a Setup Tool for Defect Structures

The initial step in characterizing defect structures by means of first principle methods is always linked to the structures themselves. For individual systems it is convenient to just create and initialize the structures manually. However, when focusing on multiple defect systems it is advantageous to utilize tools that take care of the structure creation with as little manual work as possible. Therefore, we implemented a setup tool for point defects within the Atomic Simulation Environment (ASE)[46]: the ASE `DefectBuilder`. In this section we present its general structure, dive into the details of the interstitial site creation based on Voronoi tessallation, and give a reference of how to use the `DefectBuilder` in practice.

5.2.1   General Structure

The `DefectBuilder` was designed to create an easy-to-use computational setup tool for defect structures. Its general structure is outlined in Figure 5.1: the setup tool is initialized starting from an input structure in the form of an ASE `Atoms` object. Single point defects like vacancies and substitutions, interstitial defects and adsorption sites are accessible *via* methods of the `DefectBuilder` class. As supercells are typically essential when characterizing any point defect using density-functional theory, the setup tool also embeds the desired defect into a supercell. Finally, the setup tool returns the defect supercell structures again in the ASE `Atoms` object format. The source code for the ASE `DefectBuilder` is freely available[3] and detailed usage examples for the different fucntionalities are given in the documentation[4].

We note that setting up vacancies and substitutional defects is already possible using functionalities from `asr.setup.defects` inside the ASR framework. The creation of the more advanced interstitial and adsorption sites, however, is a new feature that is not part of the ASR framework.[5] Therefore, we will go into detail and explain how one can create the latter two types of defects from a systematic point of view.

---

[3]See https://gitlab.com/ase/ase/-/blob/defect-setup-utils/ase/build/defects.py.
[4]See https://gitlab.com/ase/ase/-/blob/defect-setup-utils/doc/ase/build/defects.rst
[5]Advanced in the way that interstitials and adsorption sites can not simply be set up by removing or substituting existing occupied atomic sites inside the pristine host crystal.
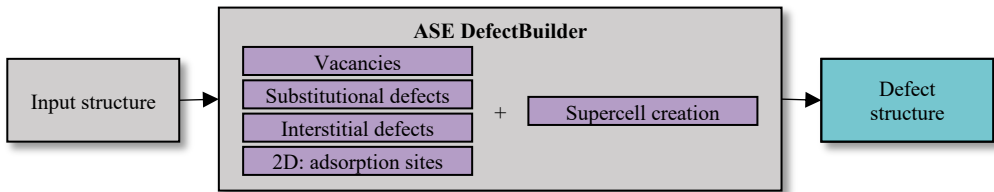
Figure 5.1: General structure of the ASE `DefectBuilder`. A given input structure in the primitive cell is used to initialize the `DefectBuilder`. Depending on the user's desire vacancy, substitutional or interstitial defects will be set up and incorporated into a supercell. The `DefectBuilder` returns the corresponding defect structures in their supercells. For the special case of 2D materials it is also possible to set up adsorption sites.

5.2.2 Background: Voronoi Construction

Before we dive into the logic behind the creation of interstitial sites in solids we take a step back and introduce the basic concept of Voronoi tessallations (or Voronoi constructions). Let us restrict ourselves to the special case of two dimensions and draw our attention to Figure 5.2. The starting point is a set of atomic positions. We can now construct polygons around each single position such that every point inside a particular polygon has to be closer to the corresponding atom than to any other atom outside of the polygon. This strategy is called Voronoi tessallation.

The boundaries of the resulting polygons are the Voronoi lines (one dimensional objects) and the corner points of those lines are the Voronoi vortices (zero dimensional objects). Vortices and points along the lines represent the positions that are furthest away from the atoms inside the crystal. In three dimensions, we effectively have one additional dimension and a Voronoi tessallation leads to faces, lines and vortices. We
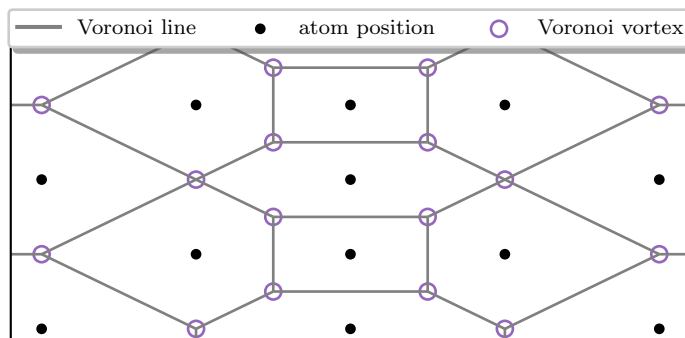

Figure 5.2: Voronoi construction in two dimensions. For a hypothetical 2D system with a set of atomic positions (black points), the construction yields Voronoi vertices (purple circles) and lines (grey lines).

will now see how Voronoi constructions can be a useful tool on the path towards systematically identifying interstitial positions and adsorption sites inside/on top of a given host crystal.

5.2.3   Interstitial Site Creation

Different to vacancy and substitutional defects, the positions of interstitial defects are not uniquely defined. In general, every position that is not occupied by atoms in the host crystal can be considered an interstitial position. But not every possible unoccupied position makes sense physically. We therefore follow our intuition to systematically set up interstitial sites: we expect interstitial sites to form at positions inside the host crystal that are (i) not too close to any of the existing occupied host sites and (ii) located at high-symmetry positions of the host crystal. This systematic approach was inspired by developments inside the ADAQ framework[124, 134] for bulk systems.

By design, the faces, lines and points from a Voronoi construction yield the positions of maximum distance to the existing host atoms. Thus, it is natural to use the Voronoi tessallation as a starting point for the interstitial site creation. Let us now sketch the underlying algorithm encoded in the `DefectBuilder`:

Interstitial site creation algorithm

1. Construct Voronoi faces, lines, and vortices for the host crystal.

2. Extract Wyckoff positions for the pristine host crystal.

3. For each position p generated by the Voronoi construction (start with vortices, then middle points of lines, then centers of faces):

    i. Map position to unoccupied Wyckoff position (ordered from highest symmetry to lowest symmetry).

    ii. Accept new interstitial site if:
       - minimum distance to closest atom is smaller than $d_{\min}$ and
       - no equivalent Wyckoff position is already occupied by a different interstitial site.

    iii. Repeat for all positions p.

We note that the number of interstitial sites that are created with this algorithm can be controlled by the minimum distance parameter $d_{\min}$ which can be passed to the `DefectBuilder` object when initializing it (default value is 1.0 Å). Larger values of $d_{\min}$ result in a lower number of interstitial sites for a given system. The 8 inequivalent and 4 unique interstitial sites for 4H-SiC are shown in Figure 5.3.
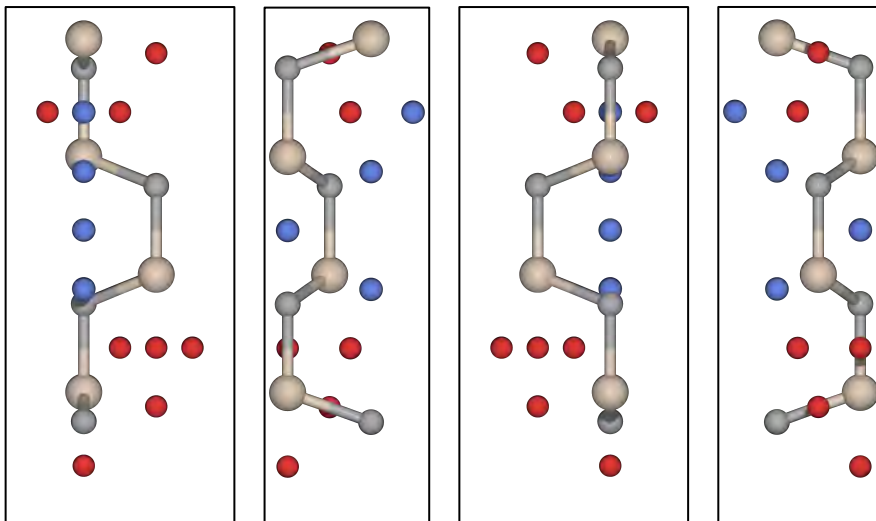
Figure 5.3: Equivalent and inequivalent interstitial sites in 4H-SiC. Interstitial sites (red and blue points) for 4H-SiC generated with the `DefectBuilder` ($d_{\min} = 1.0$ Å). The blue inequivalent sites are unique with each site having a different Wyckoff symbol. The four images correspond to different view angles (0, 90, 180, 270 degrees around the z-axis, from left to right). Red points represent interstitial positions possessing copies that can be mapped back into the same point by symmetry operations of the host crystal.

5.2.4 Adsorption sites in 2D materials

For the special case of 2D materials, the `DefectBuilder` features additional functionalities to set up adsorption sites. As commonly known, adsorbates are often located at so-called top, bridge and hollow sites on top of a surface. Once again, the Voronoi construction can help with identifying these positions.

We simply adjusted the interstitial site creation algorithm in the following way: (i) for a given input 2D material, extract the top- and lowermost layer of atoms. The result is a system of truly 2D nature. (ii) If atomic species on top and bottom are the same, only continue with one, otherwise with both separately.[6] (iii) Conduct the Voronoi construction and Wyckoff mapping in two dimensions. (iv) Translate the newly generated adsorption sites outside of the layer until the minimum distance between adsorbate and host crystal is equal to the sum of covalent radii of adsorbate and nearest neighbor in the host crystal.[7] (v) Add adsorption sites in the top (and/or bottom) position, *i.e.* above (and/or below) an existing atom. Figure 5.4 visualizes the systematically created adsorption sites on top of a 2H-MoS$_2$ layer.

---

[6]This can be of high relevance for example for MXY Janus monolayers.

[7]The criterion can be changed to the sum of van-der-Waals radii if the user chooses the 'physisorption' option in the setup.
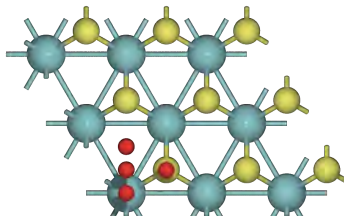
Figure 5.4: Systematically generated adsorption sites on top of 2H-MoS$_2$. For a minimum distance parameter of 0.7 Å, the `DefectBuilder` sets up four adsorption sites: the S- and Mo-top positions, as well as the hollow (center of three Mo-atoms) and bridge (between two S-atoms) adsorption sites.

5.2.5 Summary and Additional Remarks

In summary, the ASE `DefectBuilder` aims to give a convenient and easy tool to set up defect structures of different types. Currently, it features the following single defect types: vacancies, substitutional defects, interstitials, and adsorbates (2D materials). The source code is publicly available in the ASE repository[8] together with a short tutorial on how to work with the `DefectBuilder` in practice, as well as additional remarks on the implementation.[9]

The implemented functionalities should be easy to extend in the future: obvious next steps would be to combine the existing defect types to generate double, triple, and higher order defect complexes. Furthermore, the Voronoi construction and Wyckoff mapping could be generalized to systematically set up intercalated or stacked structures.

---

[8]https://gitlab.com/ase/ase/-/blob/defect-setup-utils/ase/build/defects.py
[9]https://gitlab.com/ase/ase/-/blob/defect-setup-utils/doc/ase/build/defects.rst

Summary of the Results

In this chapter, we will present the most important results that have been obtained throughout the present PhD thesis. In particular, we give a short summary of the following five publications[1]: Publication [I] builds the bridge between experiment and computations to identify intrinsic point defects in monolayer MoS_2 and characterize them by means of simple first-principles calculations. Afterwards, we switch gears and bring our full attention to a high-throughput computational characterization of 500 distinct intrinsic point defects in various two-dimensional materials in Publication [II] which resulted in the creation of the QPOD database. Different to the previous case we now calculate a broader range of properties for the individual systems to gain in-depth insight about the formation processes of defects in 2D materials. Publication [III] utilizes the ASE `DefectBuilder` (see Section 5.2) to investigate adsorption and absorption processes in experimentally synthesized 2D monolayers. Finally, Publications [IV] and [V] are products of the collective efforts of the entire CAMD section to generate the Computational 2D Materials Database (C2DB), a well-curated database with DFT generated data for over 4000 monolayer materials, and the Atomic Simulation Recipes (ASR), a Python framework for high-throughput workflows.

We note, that Publications [I], [II] and [III] can be considered as the main building blocks of this thesis where I gave major contributions and the main emphasis lies on point defects. On the other hand, the contributions for Publications [IV] and [V] were of smaller nature which is important to keep in mind when reading this chapter. We emphasize that the following sections are supposed to only give a general overview of the results obtained. Detailed explanations of the underlying concepts and results, as well as computational parameters, possible benchmark studies, and additional related literature can be found in the respective publications and their corresponding supporting information.

6.1 Publication I: Intrinsic Defects in MoS₂ Grown by Pulsed Laser Deposition: From Monolayers to Bilayers

In Publication [I], we systematically study the atomic structure of point defects, grains, and grain boundaries (GBs) of mono- and bilayer MoS_2 by means of atomic resolution

---

[1]Publication [III] is in preparation, the other four are published already.

images and first principle calculations. The study combines experimental and theoretical efforts, where my contribution was linked to the theoretical/computational part. We will therefore mostly discuss the details of the first principle calculations in this section and limit the amount of detail on the experimental parts. Nevertheless, we will give a quick overview of the experimental techniques needed to make this study possible and refer to detailed explanations in the literature. A copy of Publication [I] can be found on page 59.

Currently, there exist many different experimental techniques to synthesise transition-metal dichalcogenide (TMD) monolayers with chemical vapor deposition (CVD)[135] being among the most common ones. An alternative approach, namely pulsed laser deposition (PLD)[136, 137], lies at the heart of the synthesis process for the MoS$_2$ mono- and bilayers in this work. PLD can be used to grow a wide variety of oxides and other complex materials[136, 138]. In pulsed laser deposition, which takes place inside a vacuum chamber, a pulsed laser beam is used for the ablation of a given target. This process generates vaporized material in the chamber that afterwards deposits onto a given substrate (which is either SiO$_2$/Si for our study). In the context of the growth conditions of MoS$_2$ for this work, an additional external constant sulfur flux was added to the chamber (details see 'Methods' section of Publication [I], page 67).

From an experimental point of view, the emphasis in Publication [I] was put on the synthesis of the MoS$_2$ monolayers, as well as the materials characterization by means of scanning transmission electron microscopy (STEM). My own contribution was of theoretical nature where the existing experimental data was used to identify the experimentally observed point defects with the help of first principle calculations. In particular, we relaxed the atomic structures of numerous point defects in MoS$_2$, *e.g.* V$_S$, V$_{S2}$, Mo$_S$, Mo$_{S2}$, Mo2_{S2}, Mo$_S$Mo$_S$, and generated simulated STEM images with PyQSTEM[139] which is based on the QSTEM[140] software package.[2] An overview of the atomic structures of the most important simulated point defects, together with their experimental and simulated STEM images can be seen in Figure 1 of the manuscript on page 61. Here, the simulated images were compared to their experimental counterparts in order to identify occuring defects in the PLD-grown MoS$_2$ monolayers.

We further deepen the analysis of point defects by calculating the neutral formation energies for varying sulfur and molybdenum chemical potentials (see Figure 6 of Publication [I], page 65). We observed that the simple single defects V$_S$ and Mo$_S$ are among the most stable ones overall. The sulfur vacancy V$_S$ is an energetically favorable defect over a wide range of the chemical potential spectrum, whereas the antisite defect Mo$_S$ is more sensitive to changes to μ_S and μ_{Mo}. One the one hand, this trend agrees with previous findings that V$_S$ is very common in CVD-grown MoS$_2$ monolayers while Mo$_S$ occurs less frequently[141] (related to the fact that CVD synthesis conditions are typically S-rich). On the other hand, for PLD-grown MoS$_2$, both vacancies and antisites are present in a high concentration, which was previously observed in the literature[142]. Lastly, the experimentally observed densities of the different types of intrinsic point defects in the samples qualitatively agree well with the our calculated formation energies.

---

[2]The DFT calulations where conducted with the PBE[64] *xc* functional, computational details as well as parameters for the PyQSTEM simulations see 'Methods' section of Publication [I], page 67.

In the next part of the publication, we move past simple point defects and focus on grain boundaries in the PLD-grown MoS_2 flakes (see page 62). In particular, the experimental samples show the occurrence of mirror twin 180° GBs consisting of 4- and 8-membered rings (see Figure 2a in Publication [I], page 62). With the help of our DFT calculations, we were able to confirm that 4|8 ring motifs are indeed stable for the MoS_2 monolayer. On top of that, a density of states analysis for the 4|8 GB (see Figure 2d in Publication [I], page 62) suggested the appearance of mid-gap states that can be a cause for the generally poor photoluminescence emission[143] in monolayers grown by pulsed laser deposition.

In summary, Publication [I] showcases how simple first principle calculations can help in the identification of point defects in two-dimensional materials. For instance, formation energies give an insight into the formation processes of different point defects and qualitatively link experimentally measured defect concentrations to DFT calculations. The appearance of mid-gap states in semiconductors due to the presence of defects is an important phenomenon which we just briefly discussed in Publication [I]. In the next section we will take the next step and deepen the analysis of defects in terms of their most relevant properties for the QPOD study[II]. On top of that, Publication [II] marks the transition from a manual first principle characterization for defects in a single host material to a systematic and automatic high-throughput approach enabling us to consider hundreds of defects simultaneously.

6.2   Publication II: Quantum Point Defects in 2D Materials - the QPOD Database

In this publication, which we will call the QPOD study in the following, we systematically compute the properties of over 500 unique intrinsic point defects for 82 2D materials. The considered defect types are vacancies and antisite defects. At the heart of the study lie the implementations in the ASR framework [V] which have been discussed in Section 5.1. The ASR Recipes are used as building blocks for the underlying QPOD workflow which is depicted in Figure 9 of Publication [II] (see page 79). A unique feature of QPOD is the collection of results in a database that are visually presented in a web application. With this feature we hope to have generated a platform for easily browsable data within the defect community. Furthermore, with the ASR developments in place, we are confident that the QPOD database can be extended in the future to become an important defect data resource. In this section, we will give an overview of the calculated properties, discuss the main developments needed for the execution of this study, and present the most important conclusions we drew from the data.

In order to conduct a systematic study of hundreds of defects one first needs to choose a suitable set of host materials. For QPOD, we screened the computational 2D materials database (C2DB)[130][IV] for non-magnetic, stable host materials with a PBE bandgap of larger than 1 eV. The resulting set of 281 materials has been further reduced to 82 materials in a second screening step that is described in the Section
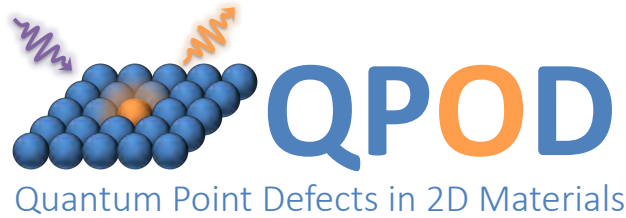
Figure 6.1: Logo of the QPOD database.

'The QPOD workflow' in Publication [II] (see page 78). Second, with the help of the implementations in `asr.setup.defects` in ASR, we were able to automatically set up single intrinsic defects and their respective supercells over the wide range of host materials. Each of the created defect systems (present in several charge states) underwent the QPOD workflow which includes structural relaxations, ground state calculations, and numerous property extractions utilizing the defect Recipes described in Section 5.1.

An essential part of the extracted thermodynamic properties is based on the correct evaluation of charge transition levels and defect formation energies which are obtained by means of Slater-Janak transition state theory (see Section 3.3.1). To benchmark the validity of SJ for defects we established that the CTLs (and therefore the charged formation energies) in 2D materials are independent of in-plane and out-of-plane supercell sizes as can be seen in Figure 6.2. Once the CTLs and formation energies were obtained, we were able to calculate the self-consistent Fermi level positions in different chemical potential conditions for more than 50 of the host materials.

Beyond thermodynamics we systematically analyzed point group symmetries, states in the gap for each of the defects as well as magneto-optical properties. In particular, the wavefunctions of the defect states were investigated in terms of symmetry, spatial localization and transition dipole moments whereas hyperfine interaction and zero-field splitting tensors have been calculated for high-spin defects. Additionally, the QPOD study includes manual excited state calculations for selected systems, resulting in values for zero-phonon lines (ZPLs), Huang-Rhys factors, and PL lineshapes. The calculations were conducted using the DO-MOM[145] method whose theory and implementation is beyond the scope of the present thesis.

With the large dataset we were able to extract some trends and correlations for the formation of intrinsic point defects: generally, most of the defects possess a rather low formation energy which confirms our hypothesis that intrinsic defects are a central part of the characterization of a given host material and should not be excluded in the calculations. Furthermore, vacancies tend to be less detrimental compared to antisites in terms of optoelectronic applications due to a more shallow character of their respective defect states. For bulk materials, it is generally known that finding intrinsically *p*-type conducting materials is challenging[146–149]. Our analysis of intrinsic dopability levels within the QPOD host material set showed the same overall trend in the realm of 2D materials.
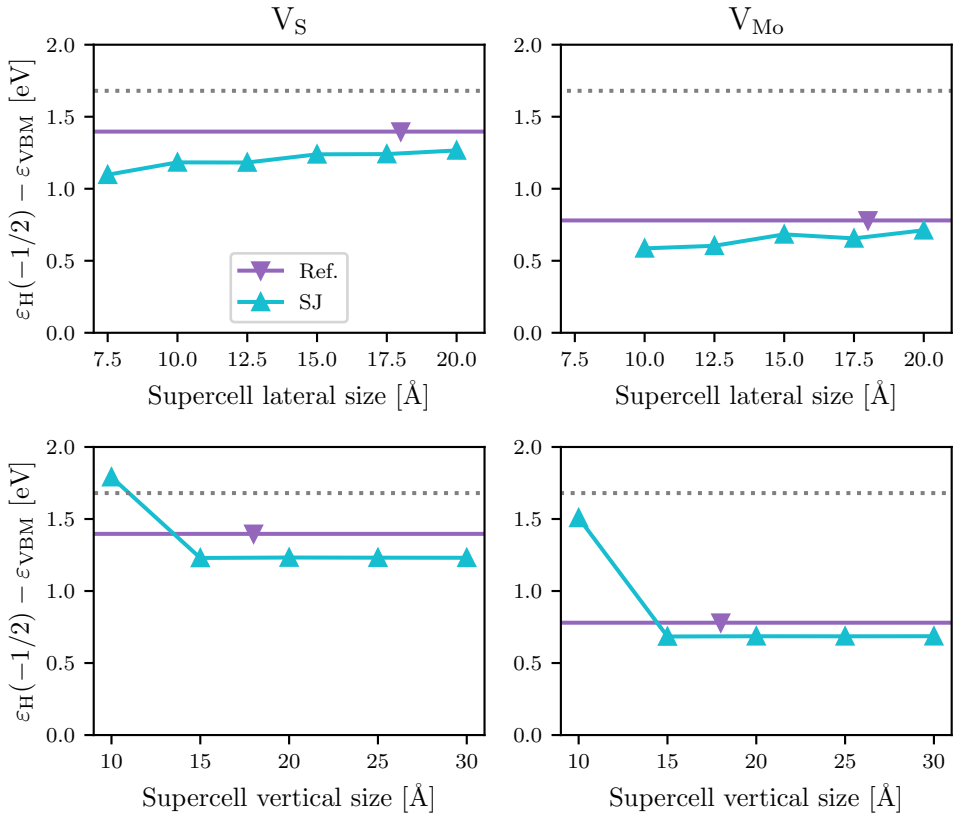
Figure 6.2: Benchmark of Slater-Janak CTLs for different supercell sizes. SJ-calculated (blue, up triangles) vertical charge transition levels $\varepsilon_{\mathrm{H}}(-1/2)$ as a function of lateral (top row) and vertical (bottom row) supercell size for the V_{S} (left column) and V_{Mo} (right column) defect in MoS_2. The purple down triangles show reference values from the literature[144]. The grey dotted line represents the calculated CBM, the zero point on the y-axis is the CBM.

Figure 6.3: The QPOD database web application. Snapshot of the QPOD search interface (top) and an example of the web-panel for charge transition levels and formation energies for the Al_{Se} defect in Al_2Se_2 (bottom). In the search panel a user can query for predefined parameters (host name, defect name, charge state) or any other KVP inside the QPOD database. The 'Formation energies and charge transition level (Slater-Janak)' web-panel shows both the charge transition levels (bottom, left) as well as the formation energies (bottom, right) in a table and in a figure.

Apart from Publication [II] itself, large efforts were put into the development of a web application for the QPOD database. Thanks to the functionalities within ASR, most of the calculated properties have been post-processed and visual representations of the results for each individual defect can be found online[3]. Figure 6.3 shows two snapshots from the web application: (i) the search interface where defects can either be found by querying for their host name, defect name and charge state or by querying for physical properties (*e.g.* point group symmetry, intrinsic dopability, equilibrium carrier concentrations, *etc.*) and (ii) one example of a property web-panel for a specific defect, which shows the CTLs, formation energies and two visual representations of the data.[4] For a more in-depth data analysis, the user also has the option to download the full QPOD database in ASE DB format or to browse the raw data online.

Overall, we drew three major conclusions from the QPOD project: first, simple intrinsic point defects are unavoidable entities in any 2D host crystal. Second, most of those defects will have a detrimental impact on transport applications and optoelectronic applications. The data suggest that different types of intrinsic defects (vacancies and antisites) behave differently in terms of the creation of deep and shallow defect states. Third, it is challenging to find suitable candidates for quantum applications within the set of intrinsic defects. We hypothesize that more complex defects (single extrinsic defects and double defect complexes) define a better platform for finding defect systems with potential for quantum applications. Therefore, we launched a subsequent study where we restrict the set of host materials to experimentally known monolayers and systematically create hundreds of extrinsic and intrinsic double defects for the respective host systems. This subsequent study is work in progress and will be concluded in the future. As part of the project we developed the framework behind the QPOD database which will support future extensions of the database (more defects, more host materials, new properties) and will make QPOD an important resource within the defect community.

6.3   Publication III: High-Throughput Doping of 2D Materials: Absorption *versus* Adsorption

In Publication [III], we investigate the energetics of 53 experimentally synthesized 2D materials that have been doped with single interstitial and adsorption dopants. The effects of doping 2D materials, in particular TMDs with foreign atoms have previously been studied in the literature[150–155]. For example, doping non-magnetic monolayers with certain extrinsic atomic species can induce magnetism. Karthikeyan *et al.*[156] conducted a theoretical investigation of transition metal-doped 2H-MoX$_2$ (X = S, Se, Te) monolayers which inspired our work in Publication [III]. However, to the best of our knowledge, no systematic computational study of interstitial vs. adsorption site doping over a wide range of 2D monolayers has been conducted to date which motivated us to address the topic ourselves.

---

[3]https://cmr.fysik.dtu.dk/qpod/qpod.html

[4]For each defect, several of those panels focusing on different properties are available.
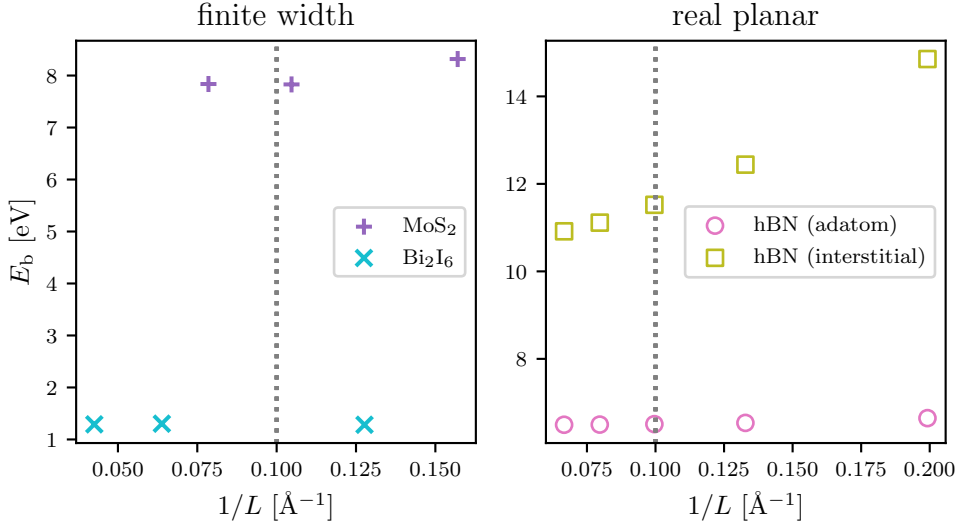
Figure 6.4: Convergence behavior of interstitial and adsorption sites in terms of supercell size. Left: binding energy of an interstitial atom in MoS_2 and Bi_2I_6 (2D materials with finite out-of-plane width) as a function of the inverse supercell size $1/L$ where $L = \frac{L_x + L_y}{2}$. Right: binding energy of an interstitial defect and atom in hBN (real planar system with no out-of-plane width) as function of the inverse supercell size $\frac{1}{L}$.

Before starting the main study, we analyzed the convergence behavior of several doped monolayer systems in terms of in-plane supercell size, shown in Figure 6.4. In-plane supercell sizes larger than 10 Å show well-converged binding energies for mono-layer host systems with a finite width, *e.g.* MoS_2 and Bi_2I_6.[5] However, for real planar systems, *i.e.* monolayer hBN, larger supercells are needed for well-converged values for the binding energy. Additionally, defining interstitial defects in real planar systems is challenging, and buckling effects[157] would give rise to a completely separate study hence we excluded those planar systems (hBN and graphene) from our host material set. An overview of the set of host crystals is given in Table 6.1.

Table 6.1: List of host systems used in Publication [III]. Overview of the 53 host crystals with host formula, space group, pristine band gap. The included materials cover metals, semiconductors, and are all non-magnetic, except for Cr_2I_6.

Host crystal formula	Space group number	Band gap (PBE)
P_4	53	0.90 eV
Ga_2Se_2	123	0.00 eV
$MoSe_2$	187	1.32 eV
W_2Te_4	11	0.00 eV

---

[5]Finite width systems refer to host crystals which are more than a single atom wide whereas real planar systems are only one atom wide and therefore "fully 2D", just like monolayer hBN or graphene.

ZrS_2	164	1.16 eV
Bi_2I_6	147	1.38 eV
Cr_2I_6	162	0.89 eV
Hf_2Te_6	59	0.00 eV
PbI_2	164	1.50 eV
Nb_4C_3	164	0.00 eV
$NiSe_2$	164	0.06 eV
PtS_2	164	1.69 eV
WTe_2	187	0.73 eV
Si_2	164	0.00 eV
$Ti_3C_2H_2O_2$	187	0.00 eV
Ga_2Te_2	187	1.30 eV
SnS_2	164	1.59 eV
TaS_2	164	0.00 eV
Mo_2CO_2	164	0.00 eV
Sn_2	164	0.06 eV
In_2Se_2	123	0.00 eV
$MoTe_2$	187	0.93 eV
$NbSe_2$	187	0.00 eV
SnS_2	187	0.76 eV
TiO_2	164	2.70 eV
$ZrSe_2$	164	0.34 eV
Ge_2H_2	164	0.92 eV
Pd_2Se_4	14	1.31 eV
$SnSe_2$	164	0.76 eV
WS_2	187	1.53 eV
Nb_2CO_2	164	0.00 eV
GaN	187	1.88 eV
MoS_2	164	0.00 eV
$PtSe_2$	164	1.17 eV
Re_4S_8	2	1.27 eV
MoS_2	187	1.58 eV
$BiITe$	156	0.41 eV
As_2	164	1.48 eV
C_2H_2	164	3.46 eV
TiS_2	164	0.02 eV
W_2Se_4	11	0.03 eV
$MoSSe$	156	1.45 eV
HfS_2	164	1.22 eV
$TaSe_2$	164	0.00 eV
Ti_2CO_2	164	0.32 eV
Re_4Se_8	2	1.12 eV
WSe_2	187	1.24 eV
V_2CO_2	164	0.00 eV
C_6N_2	191	0.40 eV
Ti_2S_6	11	0.29 eV

Ge$_2$	164	0.03 eV
HfSe$_2$	164	0.43 eV
NbS$_2$	187	0.00 eV

Another main goal of Publication [III] is to present the ASE `DefectBuilder` (see Section 5.2 for details) as a useful tool for point defect generation in 2D and bulk materials. By doping the 53 host crystals with approximately 80 different atomic species in numerous interstitial and adsorption sites, we ended up with 11579 relaxed defect systems. Each system underwent a simple workflow consisting of ionic relaxations and ground state calculations which was implemented in *httk*[119]. The DFT calculations themselves were conducted with the Vienna Ab initio Simulation Package (VASP)[76, 158] and the PBE[64] *xc* functional, the specific computational parameters can be found in Publication [III] itself (page 92).

For every relaxed structure, we calculated the formation energy and investigated whether interstitial or adsorption site doping is energetically favorable. For MoX$_2$ (X = S, Se, Te) monolayers, we reproduced the trends observed by Karthikeyan *et al.*[156], and generalized those trends to other 2H-TMD host crystals, *e.g.* WX$_2$ (X = S, Se, Te). Interstitial positions are generally favored compared to adsorption sites for TMDs with a relatively large lattice constant, *i.e.* MoTe$_2$, WTe$_2$, and *vice versa* for TMDs with small lattice constants *i.e.* MoS$_2$, WS$_2$. At this stage, we refer to Figure 2 of the related manuscript draft, which is found on page 94.

Next, we shift our focus to general trends in the entire dataset. To get a quantitative measure of how favorable interstitials and adsorption sites are for a given host material H and dopant X, we define:

$$\Delta\left[H, X\right] = E_{\text{int}}^{f,\min}\left[H, X\right] - E_{\text{ads}}^{f,\min}\left[H, X\right]. \tag{6.1}$$

Here, $E_x^{f,\min}\left[H, X\right]$ ($x = \text{ads}, \text{int}$) is the lowest occurring formation energy of a host crystal H and dopant X combination in an adsorption or interstitial site. Negative values of $\Delta\left[H, X\right]$ correspond to interstitial doping being energetically favorable and *vice versa* for adsorption sites. One main observation is that interstitial site doping in 2D materials is generally challenging, in particular for heavy and large dopant elements. $\Delta\left[H, X\right]$ was calculated for the entire dataset, resulting in Figure 3 for Publication [III] (see page 95). Here, we see that the individual $\Delta\left[H, X\right]$ varies for the different hosts and dopant elements. However, there are clear overall trends: elements of the *s*-block in the periodic table (apart from hydrogen) typically have the largest radii in their respective periods and, therefore, are least prone to interstitial doping. For decreasing sizes of the dopant atoms, $\Delta\left[H, X\right]$ decreases as well. For the special case of transition metal dopants, there are many systems for which interstitial sites are favored over adsorption sites. In addition, the data suggest that $\Delta\left[H, X\right]$ is generally lowered when doping inside materials with larger primitive lattice sizes.

Overall, we ended up with many systems that failed or did not reach convergence with respect to the forces in the ionic relaxations. This behavior did not come as a surprise as we (i) simply doped all host crystals with almost the entire periodic table, and (ii) set up systems that might not correspond to physical systems. In Publication

[III], we do not simply throw away the unrelaxed and failed calculations but we analyze the underlying convergence statistics instead. Figures 4, 5, and 6 of Publication [III], on pages 97 and 98, respectively, give a good overview of what we found out. In short, we saw that interstitial positions are generally harder to relax than adsorption sites. This tendency stems from the fact that the free space inside a 2D monolayer is very limited, resulting in higher failure percentages for dopants with larger covalent radii and larger atomic numbers (Figures 4, 5). Furthermore, for the unrelaxed and failed interstitial calculations, we saw that there is a clear correlation between the success rate of a calculation (*i.e.* does a calculation for a given host crystal and dopant atom in a specific interstitial position relax) and the normalized minimum distance parameter $D_{\mathrm{min}}^{\mathrm{norm}}\,(\mathrm{I}, \mathrm{A})$, which we define in Equation (4) of Publication [III] (page 97). For interstitials with small $D_{\mathrm{min}}^{\mathrm{norm}}\,(\mathrm{I}, \mathrm{A})$ it is basically impossible to relax the system, whereas a successful relaxation becomes more and more likely for increasing normalized minimum distance parameters which we show in Figure 6 of the publication on page 98.

Lastly, one main outcome of Publication [III] is the collection of all doped structures in an ASE[46] database. The database contains 17732 structures, of which 11579 are considered fully relaxed and converged according to our convergence criteria outlined in the manuscript.[6] Inside the database, we collected useful key-value pairs for the user to easily query the data. For example, one can query for the depth parameter, extension factor, formation energy, defect type, *etc.*[7] The database will be made available, to make it possible for other researchers to have access to the data, conduct additional or more accurate analyses, use the relaxed/prerelaxed doped structures for subsequent studies and calculations, or use the existing data for machine learning purposes.

In summary, Publication [III] showcases how versatile setup tools like the ASE `DefectBuilder` can help facilitate high-throughput studies focusing on point defects. In particular, we (i) doped experimentally known 2D monolayers with single extrinsic defects in interstitial and adsorption sites and (ii) identified trends and correlations in the stability of interstitial defects and adatoms throughout our material set, (iii) addressed the challenge of relaxing interstitial defects in 2D materials, (iv) gave suggestions for future studies building on top of the existing findings, and (v) collected a database to enable subsequent analyses of the generated dataset.

6.4 Publication IV: Recent Progress of the Computational 2D Materials Database (C2DB)

In Publication [IV], we present the recent developments and updates of the computational 2D materials database (C2DB). In 2018, the first version of C2DB[130] consisting of relaxed crystal structures as well as thermodynamic, elastic, electronic, magnetic, and optical properties of around 1500 two-dimensional materials was released. In the

---

[6]The database is available online through DTU Data *via* the following DOI: https://doi.org/10.11583/DTU.19692238.v1.

[7]More information on which KVPs are stored can be found in the publication itself.

following years, the database evolved thanks to the collective effort of the CAMD section and resulted in the following main additions/improvements which are documented in Publication [IV]: (i) extension of the database to over 4000 2D materials, (ii) addition of new properties and computational improvements for existing properties, (iii) improved provenance for the featured data. My contribution to the C2DB project in terms of the present thesis was of minor nature and was linked to a specific analysis step in the C2DB workflow as well as interlinking C2DB to the QPOD[II] database project. In the following we will briefly mention some of the main general improvements to C2DB and afterwards discuss the two contributions stated above. For detailed explanations we refer to Publication [IV] itself which is printed on page 101.

Figure 6.5 shows the workflow that every material in C2DB has to go through: after the structural relaxation of a possible new monolayer its respective dimensionality is analyzed and duplicate materials are discarded. Next, the monolayer is classified according to symmetry, magnetic structure, as well as thermodynamic and dynamic stability[159]. Lastly, new stable 2D materials are subject to the property workflow. Computational improvements to the already existing properties in the old version of C2DB[130] have been implemented for the stiffness tensor, effective masses, orbital projected band structure, G_0W_0 band structures[160], and optical absorbance (see pages 108 to 111). In addition, a range of new properties like exfoliation energy, Bader charges, spontaneous polarisation, Born charges, infrared polarisability, piezoelectric tensor, topological invariants[161], exchange coupling constants[162, 163], Raman spec-
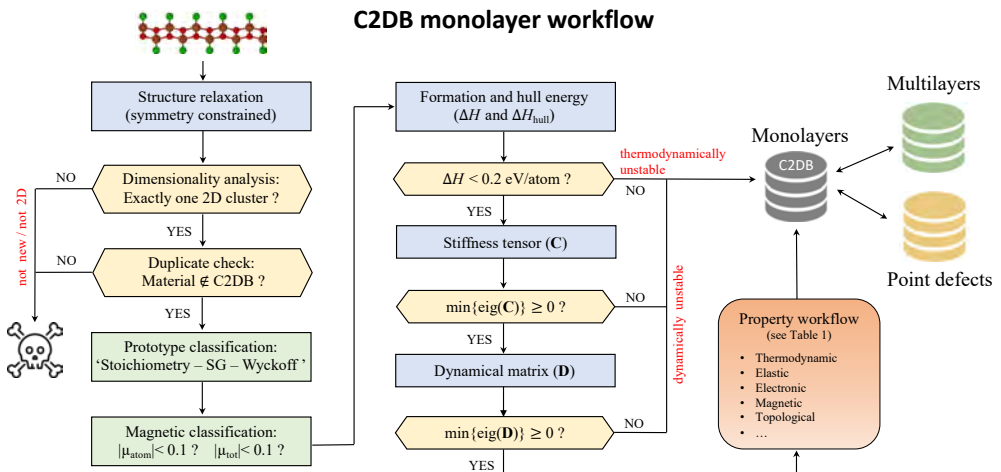
Figure 6.5: The C2DB monolayer workflow. Each material in the C2DB undergoes the same workflow: first, a new possible structure is relaxed. Second, if the structure has the correct dimensionality and is not already present in the database, it is classified in terms of stability. Finally, for all stable materials, numerous properties are calculated with the property workflow. In addition, the C2DB was used as a starting point for specialized studies on mulitlayers and point defects (QPOD database[II]). Figure copied from Publication [IV].

trum[164], and second harmonic generation[165] have been added as described on pages 115 to 124. The number of materials for C2DB increased from roughly 1500 to 4000 2D materials by adding MXY Janus monolayers[166] and exfoliable monolayers[167] from experimentally known bulk materials to the existing pool of materials.

From a personal point of view, contributing to the implementation of an ASR[V] Recipe for the dimensionality analysis based on a scoring parameter approach[168] helped identify structures with exactly one 2D cluster present. Only those 'real' 2D materials are allowed to continue with the subsequent steps of the C2DB monolayer workflow. Furthermore, we contributed in interlinking the QPOD database[II] with C2DB to seamlessly introduce an external data source for defective 2D monolayers.

In conclusion, the computational 2D materials database has been subject to major updates, both in terms of the addition of new monolayer materials and in terms of new/optimized properties added to the underlying workflow. C2DB is a valuable resource for high-quality DFT data of experimentally known and prototypical 2D materials. With its 4000+ entries, the database is a suitable platform for the prediction of materials properties by means of machine learning (ML) methods. For example, in Ref. [169], existing DFT calculations from C2DB were used to predict GW band structures. From the technical side, C2DB is a good example of how ASR[V] is a useful and powerful framework for HT studies in materials science.

6.5   Publication V: Atomic Simulation Recipes: A Python framework and library for automated workflows

In Publication [V], we present the Atomic Simulation Recipes (ASR), an open-source Python framework and library for automated workflows, which is the product of joint efforts of the CAMD section. ASR implements a simple, flexible and easy-to-use high-throughput framework for materials science. As we discussed in Chapter 4, one central part of the ASR framework evolves around the concept of a Recipe. In this section, we will first outline the design concepts behind ASR and afterwards explain the general structure of a Recipe. Lastly, we will showcase how ASR can build the 'backbone' of high-throughput projects at the example of the QPOD project[II]. We note, that this section only gives a very short overview of ASR and simply tries to put the framework into perspective with respect to the studies related to the present thesis. An in-depth discussion of the Atomic Simulation Recipes is given in Publication [V] which can be found on page 131.

ASR distinguishes itself from existing HT frameworks first and foremost by prioritizing "*usability and simplicity over system perfection*" (Ref. [V], page 133). In practice, this results in a high degree of modularity to ensure maximum flexibility for the user while still maintaing data provenance *via* built-in functionalities within ASR. We refer to Figure 1 of Publication [V] (page 135) for a schematic overview of the ASR framework. The central building block of ASR is represented by the Recipes which perform a well defined task (*e.g.* perform a ground state calculation) within material science by combining the desired atomistic simulation code through the ASE[46] interface with

Python scripts. Simply decorating the Recipe scripts with ASR's in-built functionalities will connect the code to the command line interface (CLI) and the caching framework to enable easy Recipe execution and maintain data provenance. After the correct execution of a Recipe, the results are stored *via* the so-called Result objects which in turn can be conveniently collected in an ASR database. Visual representations of the results can be easily generated *via* the command line.

Generally, we define that a Recipe is a Python function/method that performs a particular task, *e.g.* calculate a density of states, relax an atomic structure, *etc.* In practice, ASR comes with functionality to wrap Python functions in a caching layer, resulting in so-called Instructions (see Figures 2 and 3 in Publication [V], pages 135, 136). What is encoded in those Instructions is entirely up to the user and contributes to the 'simplicity' philosophy of ASR. With that in mind, a Recipe is now simply a set of Instructions with automatically cached results. Due to the modular design of the ASR framework, numerous Recipes can be combined into higher-level workflows to perform more complex sequences of calculations. One main contribution in terms of the present thesis was the implementation of several Recipes for the ASR library.[8]

To conclude this section, we will have a look at a specific example of how the ASR framework can be used as the computational backbone of a high-throughput study. In particular, the realization of the QPOD database[II] heavily relies on the core functionalities of ASR: numerous newly developed defect Recipes (see an overview of the corresponding Recipes in Section 5.1) as well as already existing Recipes (`asr.relax` for structural relaxations, `asr.gs` for ground state calculations) have been combined into a MyQueue[127] workflow to compute structural, thermodynamic, electronic, and magneto-optical properties of over 500 unique intrinsic point defect systems in 2D materials (see Publication [II] for details). After the collection of the results, ASR made it possible to conveniently store the data in an database and present the data visually with ASR web panels.

---

[8]The repository can be found here: https://gitlab.com/asr-dev/asr/.

CHAPTER 7

Summary

Characterizing the impact of point defects on its host crystal is an important stepping stone in understanding materials as a whole and in facilitating the utilization of defect systems for applications, *e.g.* in quantum technologies. Efforts within both experiment and theory are needed in order to achieve that goal. This thesis is a part of those efforts and takes the computational route to systematically study point defects in two-dimensional materials. We applied our methodologies to experimentally synthesized as well as theoretically predicted monolayers. As part of the process, we contributed to the development of high-throughput frameworks and tools for defect calculations which should make future studies more straightforward. Generating DFT defect data according to the FAIR principles[170, 171] and making them publicly available through open-access databases, *e.g.* the QPOD database[II][1] or the Adsorption vs. Absorption in 2D Materials database[III][2], is how we want to create synergy between experiment and theory in the discovery of new potential defect systems with impact in modern technology.

The basis for our first principle calculations of point defects is represented by density-functional theory. After reviewing the most important concepts and implications of DFT, we turned our attention to one central extension, namely Janak's theorem, which enables us to connect the total energy of a system to the eigenenergy of a specific state *via* its occupation number. A generalization of the theorem made it possible to calculate charge transition levels for defect systems, which is an important building block in the evaluation of thermodynamic defect properties for our two-dimensional host crystals.

Next, we went more into depth and connected the general concepts of DFT to the actual thermodynamic property evaluation of point defects in two-dimensional materials. In our calculations, every point defect characterization starts with the supercell approach and the resulting defect energetics. We employ Slater-Janak transition state theory for the accurate computation of defect formation energies as well as charge transition levels, and discuss how Slater-Janak is a valid and accurate approach without the need of *a posteriori* electrostatic correction terms. Furthermore, we presented how we self-consistently evaluate the Fermi-level position of crystals in the presence of numerous defects to effectively obtain the intrinsic dopability of materials in different chemical potential conditions, an important property of host materials with respect to transport applications.

The emergence of high-throughput studies in the field of materials science combined

---

[1]Find the database here: https://cmr.fysik.dtu.dk/qpod/qpod.html
[2]The database is available online through DTU Data *via* the following DOI: https://doi.org/10.11583/DTU.19692238.v1.

with increasing computational capabilities created the need for versatile, stable, and powerful frameworks to facilitate, register, manage, and track high-throughput calculations. Throughout this thesis, the Atomic Simulation Recipes have been an important building block in terms of HT infrastructure. The design and implementation of numerous Recipes marked our effort to make HT studies for point defects easier, more reproducible, and enabled us to conduct studies resulting in open-access materials databases, *e.g.* the outcomes of Publications [II], [III], and [IV]. Furthermore, we tried to make the generation of structures with various defect types more accessible by implementing the ASE `DefectBuilder` as a useful setup tool for single point defects within the Atomic Simulation Environment.

From a results point of view, we showed that first principle calculations for point defects in two-dimensional materials is a useful method to support experimental observations, *i.e.* through the identification and characterization of various intrinsic defect types in PLD-grown MoS_2 monolayers. Thanks to the emphasis on implementing the 'backbone' for HT defect calculations in ASR, we were able to systematically study the effect of 500 distinct intrinsic point defects in 82 2D semiconducting materials and publish the open-access QPOD database. We saw that characterizing intrinsic defects is vital to understand a specific host material better. Finding suitable defect candidates for quantum applications within the simple defect set is challenging suggesting that double defects with extrinsic dopant elements might represent a better design platform. Facilitated by the `DefectBuilder` module within ASE, we further gained insight into the adsorption and absorption processes of single extrinsic dopants in two-dimensional monolayers and extracted trends and correlations from a materials set of over 11,000 relaxed interstitial and adsorbate structures. Additional in-depth analysis of the corresponding data, for example by means of machine learning, will be a natural extension of the initial results.

Lastly, through two minor contributions to projects that developed out of the collaborative efforts of the CAMD section, namely the Computational 2D Materials Database (C2DB)[IV] and the Atomic Simulation Recipes (ASR)[V], we made a small contribution (i) to generate high-quality, open-access electronic structure data within the field of first principle calculations on 2D materials and (ii) to develop a Python framework and library for automated workflows.

CHAPTER 8

Outlook

"Defect physics is almost as complex as solid-state physics itself"[102] is what Alkauskas and co-authors write in their tutorial on defects in semiconductors. Due to that inherent complexity and vastness of the field, one single thesis can hardly shed light on all specific aspects. Instead, progressing the specific sub-field of high-throughput characterizations of defects within the realm of two-dimensional materials and developing frameworks to simplify future first principle studies on defects was the goal of this thesis. Along the way we were able to gain insight into the formation processes of intrinsic and extrinsic point defects in monolayer materials. Many questions were answered, others remain untouched, limitations became obvious and future extension need to be addressed.

From a computational point of view, we limited ourselves to the PBE[64] *xc* functional which generally describes thermodynamic properties of defects well[172] but suffers from two major drawbacks: (i) the bandgap underestimation and (ii) delocalization errors[96] which are of high importance for the accurate characterization of defect systems. Naturally, utilizing exchange correlation approximations that are located higher up on Jacob's ladder[173] will improve the accuracy. In particular, studies on point defects utilizing computationally heavy hybrid functionals[174–176], GW[177–180], or many-body perturbation theory combined with configuration interaction methods[181–183] have already been applied to a limited amount of systems. Our implemented functionality can be extended to those more accurate methods also in the high-throughput context once computational capabilities allow for it. Furthermore, a hybrid approach where computationally less demanding functionals like PBE are used for the screening of interesting defect systems before an in-depth characterization by means of HSE06 or GW[184] steps in could be very useful.

In terms of physical properties, a natural extension of the present work would be to include additional properties into the existing HT framework for defect calculations. Examples could be the automatic evaluation of excited states, carrier capture dynamics, photoluminescence spectra, *etc.* to cover a broader range of quantities relevant for applications. With more time and more computational resources, the addition of more host crystals as well as more defect types for the screening of novel defect systems with technological impact is another obvious step. The end result could be a well curated database system which acts as a reference tool for DFT data in the field of defect physics. The availability of such a reference tool is of high importance to both experiment and theory.

The creation of a large amount of publicly available DFT data for point defects in two-dimensional materials is also an important stepping stone in the transition from purely first principle electronic structure characterizations towards machine learning

based predictions of the impact of point defects in monolayer materials. Possible predictions could entail the thermodynamic stability of defect systems (for example by using the QPOD[II] dataset), classification of deep and shallow defect states, as well as addressing the question of segregation[185] of doped 2D materials (using the dataset from Publication [III]). In particular, we would like to advertise the dataset produced as part of Publication [III] with more than 11,000 defect systems as a valuable resource for numerous machine learning applications. We note, that single studies applying machine learning methods to defect systems can be found in the literature already[186–189].

Extending the current work in terms of (i) computational improvements, (ii) property, material, and defect system additions, as well as (iii) machine learning characterizations will be an important step towards understanding the complex field of defect physics better. This thesis was our effort to slightly push the needle for high-throughput characterization of point defects in two-dimensional materials and creating useful platforms/frameworks for future considerations in the field.

CHAPTER 9

Publications

9.1 Publication I: Intrinsic Defects in MoS$_2$ Grown by Pulsed Laser Deposition: From Monolayers to Bilayers

F. BERTOLDO
R. R. UNICOC
Y. C. LIN
X. SANG
A. A. PURETZKY
Y. YU
D. MIAKOTA
C. M. ROULEAU
J. SCHOU
K. S. THYGESEN
D. B. GEOHEGAN
S. CANULESCU

ARTICLE

Intrinsic Defects in MoS₂ Grown by Pulsed Laser Deposition: From Monolayers to Bilayers

Fabian Bertoldo,* Raymond R. Unocic, Yu-Chuan Lin, Xiahan Sang, Alexander A. Puretzky, Yiling Yu, Denys Miakota, Christopher M. Rouleau, Jørgen Schou, Kristian S. Thygesen, David B. Geohegan, and Stela Canulescu*

ACCESS | 📊 Metrics & More | 📄 Article Recommendations | 🔎 Supporting Information

ABSTRACT: Pulsed laser deposition (PLD) can be considered a powerful method for the growth of two-dimensional (2D) transition-metal dichalcogenides (TMDs) into van der Waals heterostructures. However, despite significant progress, the defects in 2D TMDs grown by PLD remain largely unknown and yet to be explored. Here, we combine atomic resolution images and first-principles calculations to reveal the atomic structure of defects, grains, and grain boundaries in mono- and bilayer MoS₂ grown by PLD. We find that sulfur vacancies and Mo_S antisites are the predominant point defects in 2D MoS₂. We predict that the aforementioned point defects are thermodynamically favorable under a Mo-rich/S-poor environment. The MoS₂ monolayers are polycrystalline and feature nanometer size grains connected by a high density of grain boundaries. In particular, the coalescence of nanometer grains results in the formation of 180° mirror twin boundaries consisting of distinct 4- and 8-membered rings. We show that PLD synthesis of bilayer MoS₂ results in various structural symmetries, including AA′ and AB, but also turbostratic with characteristic moiré patterns. Moreover, we report on the experimental demonstration of an electron beam-driven transition between the AB and AA′ stacking orientations in bilayer MoS₂. These results provide a detailed insight into the atomic structure of monolayer MoS₂ and the role of the grain boundaries on the growth of bilayer MoS₂, which has importance for future applications in optoelectronics.

KEYWORDS: MoS₂, transition-metal dichalcogenides, defects, grain boundaries, moiré patterns, scanning transmission electron microscope, pulsed laser deposition
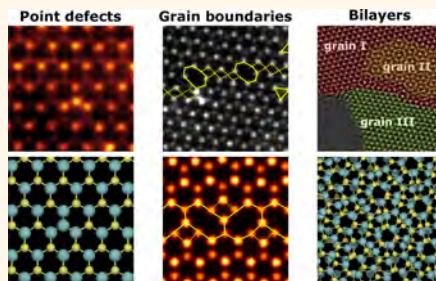
Two-dimensional (2D) transition-metal dichalcogenides (TMDs) and their heterostructures have received much attention in recent years due to their potential as building blocks for next-generation ultracompact optoelectronic devices.[1,2] In monolayer form, the semiconducting TMDs exhibit a photoluminescence (PL) quantum yield that is 4 orders of magnitude higher than their bulk values owing to the indirect-to-direct band gap transition.[3] Moreover, the combination of different monolayers in vertical heterostructures, such as MoS₂/WS₂, are very appealing due to ultrafast electron−hole separation,[4] which provides a platform for studying the emerging field of atomically thin photodetectors or solar cells.[5−7] While a number of well-developed techniques, including chemical vapor deposition (CVD), have shown great potential for the wafer-scale fabrication of TMD monolayers, a general synthetic scheme for the fabrication of heterostructures of 2D-based superlattices is still to be demonstrated.

Pulsed laser deposition (PLD) holds the potential for sequential growth of unlimited combinations of 2D TMDs in vertical heterostructures with clean and atomically flat interfaces.[8,9] The physical processes associated with PLD involve the generation of ions and neutrals with energies of tens of eV in vacuum that can be significantly reduced in background gas[10,11] as well as nonequilibrium processes that occur upon landing on the substrate, which involve diffusion, desorption, and nucleation.[12−14] Recent studies have shown that centimeter-scale WS₂ and MoS₂ films could be achieved using PLD.[15,16] Alternative approaches based on pulsed laser evaporation of Se atoms have been used to reduce Se vacancies in monolayers MoSe₂ nearly to the level of pristine crystals[17] or
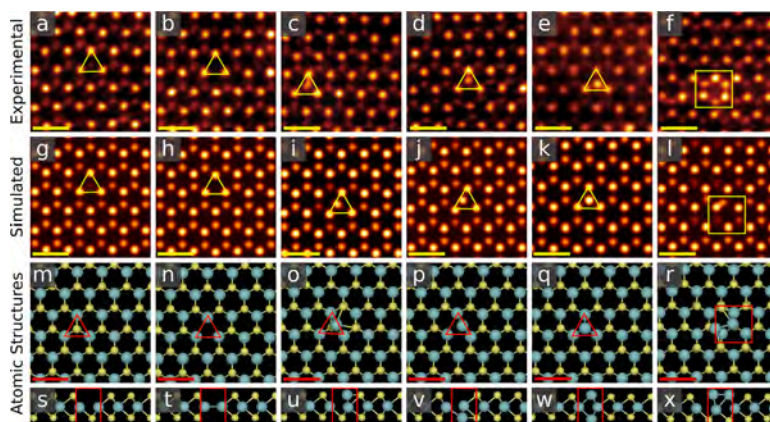
Figure 1. Intrinsic defects in monolayer MoS₂ grown by PLD. (a−f) From left to right: Experimental STEM images of V_S, V_{S2}, Mo_S, Mo_{S2}, $Mo2_{S2}$, and Mo_SMo_S. (g−l) Simulated STEM images using the relaxed structures obtained from DFT calculations using PyQSTEM.[26,27] Top and side views of the relaxed atomic structures are shown in (m−r) and (s−x), respectively. The triangles and rectangles highlight the respective point defects. The scale bars are 5 Å.

as a general scheme for the formation of WSSe Janus structures.[18] In terms of materials properties, the highest attained mobility in films grown by PLD (31 cm⁻² V⁻¹ s⁻¹ for WS₂)[19] is comparable to the values reported for CVD specimens, which typically range from 20 to 90 cm⁻² V⁻¹s⁻¹.[20,21] However, the PL emission of PLD grown monolayers is typically low.[22] The PL quenching in 2D TMDs can be caused by a high density of intrinsic defects, localized mid-gap states at the grain boundaries (GBs)[23] and/or mixed 1T/2H phases.[24] However, to date, the research into the atomic structure of monolayer TMDs synthesized by PLD has been limited, presumably due to challenges in obtaining single-layer crystals. Understanding the types of defects associated with PLD, especially at the monolayer limit, where the effects of defects are profound, will be a critical step in the advancement of the field.

Here, we perform a systematic study of intrinsic defects in TMDs grown using a one-step synthesis PLD with a combination of atomic resolution imaging and first-principles calculations. In our study, we use the MoS₂ monolayer as a prototypical material, and we show that sulfur vacancies, antisite defects, and mirror twin GBs with 4- and 8-membered ring motifs are predominant defects in PLD specimens. Using density functional theory (DFT) calculations, we reveal that the intrisic point defects are thermodynamically favorable under Mo-rich/S-poor conditions specific to the PLD grown films. We further elaborate on the role of the pulsed ablation plasma on the morphology of the grains and GBs and their role on the nucleation of a second layer on top of a polycrystalline continuous monolayer of MoS₂. Finally, we explore the growth of MoS₂ beyond monolayers and reveal that PLD grown bilayers exhibit AA′ and AB stacking orientations as well as moiré patterns with various twist angles.

RESULTS AND DISCUSSION

Point Defects in Monolayer MoS₂. The mono- and bilayers of MoS₂ were grown by a modified PLD approach in which a sulfur cracker was used as an additional source of sulfur during the growth process (see the Methods section for

details). The films grown on either SiO₂/Si or sapphire were transferred onto a TEM grid, and the atomic structure was investigated using atomic-resolution annular dark-field (ADF) imaging on an aberration-corrected scanning transmission electron microscope (STEM). As the intensity profile in a ADF STEM image scales roughly with $Z^{1.7}$, where Z is the atomic number, atomic resolution images were used to identify the type and concentration of defects in monolayer MoS₂.[17,25] Representative Z-contrast STEM images of a monolayer MoS₂ are shown in Figure 1a−f, together with the simulated STEM images (Figure 1g−l) obtained using the DFT relaxed structures (Figure 1m−x). All experimental images were processed using Gaussian and fast Fourier transform (FFT) filters, as discussed in the Supplementary Note 6. A rich variety of intrinsic defects is observed in the monolayer MoS₂, including a single sulfur vacancy (V_S), double sulfur vacancy (V_{S2}), antisite or substitutional defects where one Mo atom occupies a S site (Mo_S), Mo atom occupies two sulfur sites (Mo_{S2}), or two Mo atoms occupy S sites ($Mo2_{S2}$) as well as another type of vacancy complex with two antisite defects in the hexagonal ring (Mo_SMo_S). The other types of antisite defects, such as a S atom occupying the Mo site (S_{Mo}) or pair of S's column occupying the Mo site (S_{2Mo}), not shown here, are less common.

All intrinsic defects were identified through direct analysis of the image intensity and by direct comparison with the image intensity profiles derived from DFT calculations (see Supplementary Note 1 for details). The V_S defect structure can be clearly distinguished from V_{S2} by the dimmer spot associated with single S atom or its absence, respectively, as highlighted by the overlaid triangles in Figure 1a,b. Similarly, the Mo_S antisite defect appears more smeared out as compared to Mo_{S2} due to the contribution from the S atom located in the bottom plane of MoS₂. The Mo_{S2} antisite defect is particularly noticeable in both theoretical and experimental STEM images due its bright image intensity contrast. Moreover, as can be seen in Figure 1, the Mo_S and Mo_{S2} defect structures break the symmetry of the lattice, with the Mo atom substituting one or two S atoms moving closer to one of the nearest-neighbor Mo atoms. This phenomena has been previously attributed to
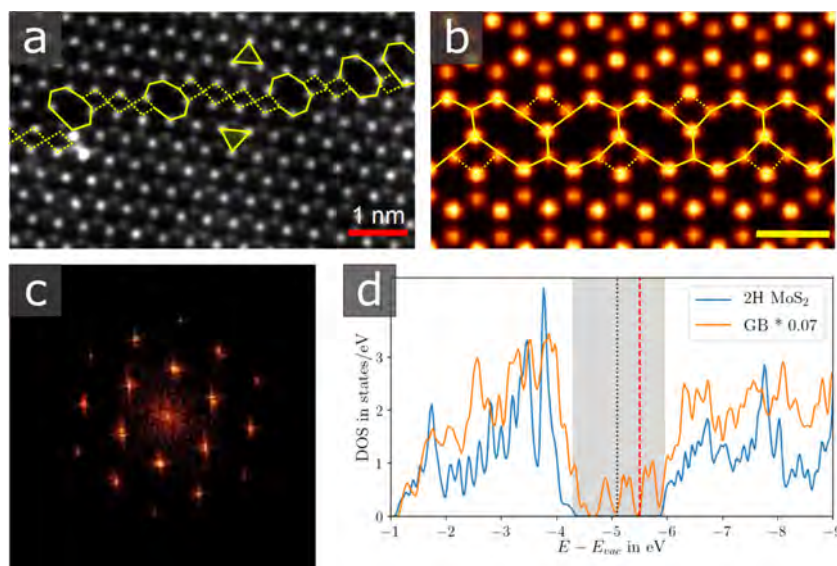
Figure 2. Atomic structure of a mirror GB in PLD-grown MoS$_2$ monolayer. (a) High-resolution ADF STEM image of a 180° twin GB showing a periodic line of defects with 4|8 motif rings. The overlaid triangles denote the twin grains rotated by 180° with respect to each other. (b) Simulated STEM image of a GB relaxed structure reproducing the 4|8 structure. (c) FFT of the image showing one set of 6-fold diffraction spots. (d) The calculated DOS of the pristine 2H-MoS$_2$ (blue line) and 4|8 GB structure (yellow line) as a function of energy with respect to the vacuum level. The black dotted line represents the Fermi energy of the pristine structure, and the red dashed line highlights the Fermi energy of the GB. The energy gap of pristine MoS$_2$ is colored in gray. The yellow scale bar has a length of 5 Å.

symmetry-lowering Jahn−Teller distortion.[28] In addition to the aforementioned defects, Mo and S adatoms were frequently seen in MoS$_2$ MLs, as discussed in the Supplementary Note 2. More importantly, the Jahn−Teller distortion has been theoretically predicted to enhance magnetic anisotropies in defective MoS$_2$ in the presence of transition metals absorbed on the surface, which could provide functionalities and form single adatomic magnets in monolayer MoS$_2$.[29]

Using a large set of STEM images, we have estimated the density of defects in PLD monolayers (see also Supplementary Table 1). Our calculations show that V$_S$, followed by Mo$_S$, are the most predominant points defects in monolayer MoS$_2$, independent of the substrate used for the growth, namely sapphire and SiO$_2$/Si. The qualitative analysis of the STEM images revealed average densities of 0.378 nm^{-2} for V$_S$ vacancies and 0.126 nm^{-2} for Mo$_S$ antisites. The points defects identified in PLD monolayers are significantly different from specimens prepared by mechanical exfoliation or CVD, where single S vacancies are the predominant intrinsic defects with densities up to 0.12 nm^{-2}.[30] We note, however, that a high occurrence of antisite defects was observed in TMDs grown by physical vapor deposition (PVD).[31]

GBs in PLD Monolayers of MoS$_2$. In addition to intrinsic point defects, TMDs are known to host a wide variety of defect rings associated with GBs. Due to the 3-fold symmetry of the MoS$_2$ monolayer, GBs with misorientation angles from 0° to 60° can emerge upon the coalescence of individual crystals in synthetic MoS$_2$.[32] This can lead to formation of dislocation cores with various odd-membered ring motifs, such as 4|4, 4|6, 4|8, or even-membered rings, such as 5|7.[32]

Previous studies based on optical microscopy images of MoS$_2$ and WS$_2$ grown by PLD suggest that 2D crystals are continuous over a large area.[15,33,34] However, a study on the nanoscale structure of PLD monolayers is still lacking. Here, we have used atomic resolution STEM imaging to shed light into the various dislocation core structures that constitute GBs in PLD monolayers. A low-magnification ADF STEM image reveals that PLD monolayers consist of domains of several tens of nanometers in size connected by a high density of dislocation cores (see Figure S3). We argue that this is a common feature of a one-step synthesis of 2D TMDs by PLD[35] and distinct from the majority of reports in CVD, where efforts have concentrated on the slower edge-dominated growth of isolated, triangular, single-crystalline ML domains that are usually tens of micrometers. For comparison, continuous and large-area CVD-synthesized monolayers have crystalline grains varying from tens to hundreds of nanometers.[36] We also find that the edges of the PLD monolayers have zigzag shapes, predominantly along the directions of Mo (seeFigure S3). The edge reconstruction in MoS$_2$ has been well documented elsewhere,[37] suggesting that such edge sites are chemically active and thus particularly responsible for the catalytic performance of TMDs.[38,39] In addition, edge sites have shown to induce magnetism in TMDs.[40]

Figure 2a illustrates the atomic resolution image of a mirror twin 180° GB, commonly observed in PLD monolayers. The mirror twin GB structure consists of edge-sharing 4- and 8-membered rings, with reoccurring 4|8 ring motifs. The ring motifs alternate as 8−4−4−4, 8−4−4−4−4, and 8−4−4 and do not show a regular pattern. Due to the lack of periodicity, it is not straightforward to reproduce the GB in the DFT calculations. Instead, we have shown the structural stability of
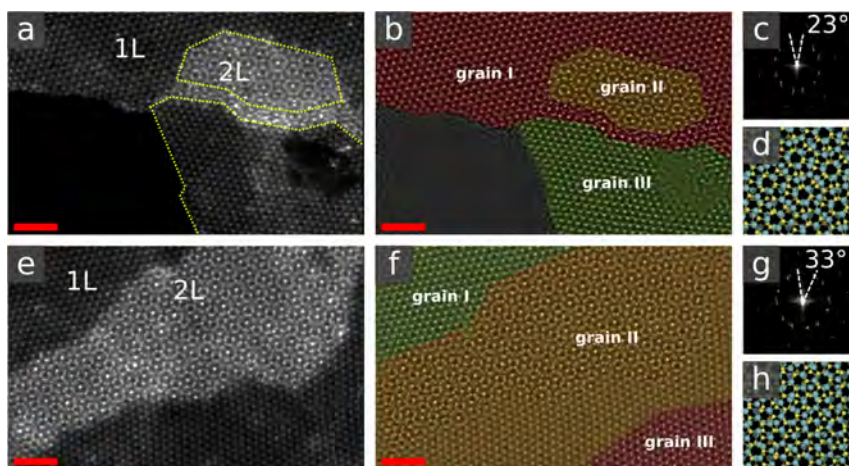
Figure 3. Moiré patterns in bilayer MoS$_2$. (a,e) ADF STEM images of mono−bilayer MoS$_2$ with moiré patterns in the bilayer regions with tilt angles of 23° and 33°, respectively. (b,f) The false-colored FFT-filtered image indicates the three distinct grains in the bottom layer and Moiré fringes that form due to the overlap between grains I and II. (c,g) The FFT patterns of the images shown in (a) and (e). (d,h) Structural models of the moiré patterns. The yellow dashed lines in (a) highlight two distinct grains with a 60° rotation. The scale bars are 2 nm.

8−4 periodical ring motifs (Figure 2b and Figure S4). The mirror twin GB structure emerges when two MoS$_2$ crystals rotate by 180° with respect to each other and Mo and S atoms swap their positions in the hexagonal lattice. In such a configuration, the 4-membered rings are connected by S atoms, whereas GB steps are linked by parallel 8-membered rings. In a 4-membered ring, Mo atoms have the same 6-fold coordination as in the pristine structure, whereas the S atoms change from a 3-fold to 4-fold coordination, similar to prior reports on MoS$_2$.[30] This implies a local change in the chemical composition along the mirror twin structure from the nearly stoichiometric MoS$_2$ to MoS$_{1.5}$.

DFT calculations were performed to obtain insight into the electronic structure of the mirror twin GBs. Figure 2d shows the theoretically predicted density of states (DOS) of the pristine 2H-MoS$_2$ phase (blue line) and GB structure (orange line), while the bandgap of pristine MoS$_2$ is indicated by the gray area. From the DOS of the 4|8 GB structure, it can be observed that the Fermi energy decreases by 400 meV with respect to the pristine structure, and more importantly, localized mid-gap states appear in the calculated DOS. The presence of mid-gap states provides a likely explanation for the typically poor PL emission in the PLD monolayers.[23] Other types of GBs, such as tilt GBs with 5|7 ring motifs, were occasionally observed in this work (not shown).

Atomic Structure of Twisted MoS$_2$ Bilayers. We further elaborate on the role of grains and GBs on the nucleation of domains in PLD using direct atomic resolution imaging. We find that primarily GBs, followed by crystal edges promote the nucleation of a top layer in bilayer MoS$_2$. The nucleation on a flat MoS$_2$ is not commonly observed. In the first stage of the growth, GBs are inherently formed through the coalescence of nanometer size grains. Upon the formation of the complex GBs with distinct dislocation cores, depending on the misorientation angle at the GB, the in-plane growth of MoS$_2$ stops. The second layer nucleates near the GBs or edges, typically retaining a similar crystal orientation as one of the original

grains. Therefore, the orientation of the top domains is directly correlated with the crystal orientation of the bottom MoS$_2$ layer. Our findings reveal that PLD growth of bilayers yields 2H and 3R phases corresponding to the AB to AA′ stacking, respectively, but also moiré patterns with twist angles of 23°, 27°, 33°, and 44°. Their corresponding atomic models are given in Figure S10. Atomic resolution images of mono−bilayer MoS$_2$ with twist angles in the bilayer region of 23° and 33° are shown in Figure 3. Here, the twist angle of a MoS$_2$ bilayer can be readily determined from the FFT spectra of the corresponding stack. The Fourier spectrum shown in Figure 3c,g reveals two distinct sets of diffraction spots, and by selectively filtering each of the diffraction sets, the spatial extension of the top and bottom layers in the overlapping region as well as the in-plane and out-of-plane tilt angles can be resolved (see also Supplementary Figures S5 and S6). The bottom layer of the junction is a continuous monolayer consisting of distinct grains of tens of nanometers in size connected by GBs. Upon the formation of the first layer layer of the junction, the growth continues as one grain extends on top of the neighboring one without any lattice discontinuity. This is shown in Figure 3a, where grain I extends on top of grain II and across the second GB on top of the grain III, resulting in the formation of a 60° GB in MoS$_2$. Here, grains II and III form a 60° GB in MoS$_2$. The ring motif structure of the 60° GB cannot be resolved, but the GB region is evidenced by the discontinuity region in the FFT-filtered image (see Figure S5 for details). In the case of the 33° twisted bilayer shown in Figure 3e, in the absence of line defects associated with the GB, the moiré pattern extends over a large area. In this case, grain I continues to grow on top of grain II, but has not yet reached the mirror GB associated with grain III. The corresponding structural models of the moiré patters with twist angles of 23° and 33° are shown in Figures 3d,h, respectively. More importantly, we observe that the top layer of the bilayer MoS$_2$ extends on top of the neighboring grains and GBs without any lattice discontinuity associated with the GB.
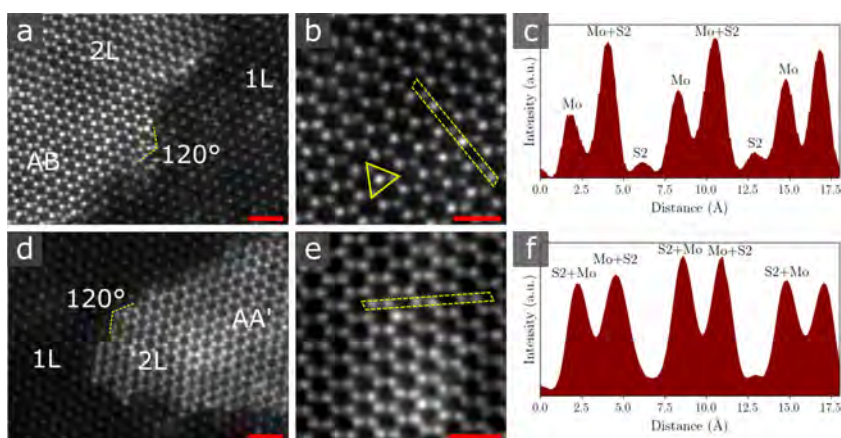
Figure 4. AB and AA′ stacked bilayer MoS$_2$. (a) Atomic resolution ADF STEM image of monolayer (1L)/bilayer (2L) MoS$_2$ with AB stacking configuration. (b) Magnified image of the bilayer region. (c) Intensity line profile measured from the yellow dashed highlighted rectangle in (b). (d) ADF STEM image of a mono−bilayer with AA′ stacked sequence. (e) Magnified image of the AA′ bilayer region with the corresponding the line profile in (f). The yellow dashed lines in (a) and (d) highlight the atomically sharp edges with an angle of 120°. The scale bars are 1 nm.
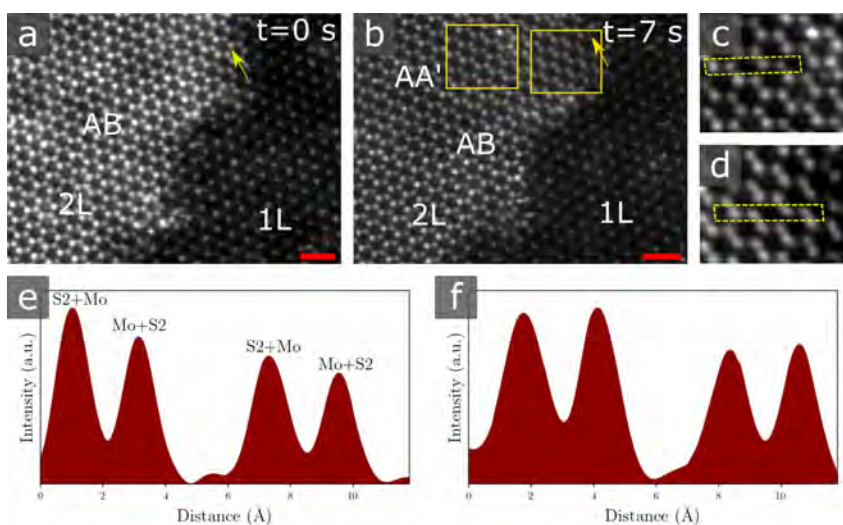

Figure 5. Electron beam-driven transition from AB to AA′ at room temperature. (a) ADF-STEM image of the AB stacked bilayer at time $t = 0$ s. (b) Transition from AB to AA′ orientation upon 7 s of imaging. (c,d) Enlarged images of the areas highlighted in (b). The line profiles for the AA′ stacking (e) and transition zone (f). The scale bars are 1 nm.

The absence of any in-plane chemical bonds between the bottom and top layers of the bilayer MoS$_2$ stems from an interlayer coupling determined by weak van der Waals forces, as expected for layered TMDs.

Figure 4 shows the Z-contrast STEM images of mono−bilayer MoS$_2$ with AB (0° twist angle) and AA′ (60° twist angle) stacking configurations in the bilayer region. The AB and AA′ stacking orientations are the most energetically stable and, thus, commonly observed in synthetic TMDs.[41] A significant image intensity contrast can be observed between the monolayer and bilayer regions. For the AB-stacked bilayer,

the top and bottom crystals have the same lattice orientation, however with a slight translation between them. The corresponding line profiles shown in Figure 4c indicate a regular pattern composed of (i) Mo atoms from the top layer and S2 atoms from the bottom layer (denoted as Mo + S2), (ii) Mo atom from the bottom layer, and (iii) weak intensity associated with the S2 atoms from the top layer. This gives an alternating image intensity contrast of the atoms arranged in hexagonal rings, similar to the 2H-MoS$_2$ monolayer, however, with additional dimmer spots in the middle. In the case of the bilayer with AA′ stacking order, the alternating (i) Mo atoms
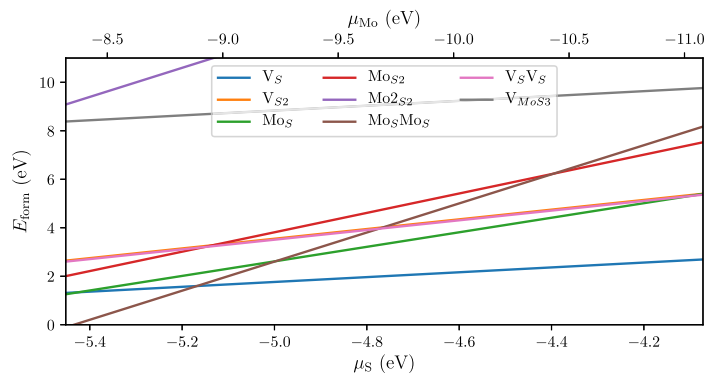
Figure 6. Formation energies of point defects in MoS$_2$. Formation energies of the different point defects in MoS$_2$ as a function of S chemical potential μ_S (bottom X-axis) and Mo chemical potential μ_{Mo} (top X-axis).

from the bottom layer and S2 atoms from the top layer (denoted as Mo + S2) and (ii) S2 atoms from the bottom layer and Mo atoms from the top layer (S2 + Mo) arranged into the hexagons lattice result in a similar intensity contrast (Figure 4d−f). Notably, point defects are observed in the bilayer structure as well, as highlighted by the overlaid triangle in Figure 4b. The line profile analysis of the image intensity enables us to attribute it to the Mo$_S$ antisite defect present in the bottom layer of the bilayer. We find that the top layer of the bilayer MoS$_2$ is of a higher crystal quality as compared the bottom layer and free of GBs. In contrast to the first layer which is grown on SiO$_2$/Si or sapphire, the second layer nucleates on a crystalline 2H-MoS$_2$, which may serve as an optimum template for the van der Waals growth and screen chemically active impurities from the underlying substrate.

Next, we will show that electron beam irradiation of the AB oriented bilayer can induce a transition from the AB to AA′ bilayer stacking (Figure 5). Intriguingly, this transition occurs at room temperature and at a lower accelerating voltage than previously reported (60 keV in this work *versus* 80 keV).[42] Sequential imaging analysis reveals that the AB to AA′ transition occurs within a time frame of 7 s and is mediated by an intermediate phase in which the top and bottom layers of the bilayer MoS$_2$ exhibit a linear interlayer shift, as indicated by the shrinkage of the hexagonal lattice in Figure 5d. The local shrinkage of MoS$_2$ is observed near the edge of the AB bilayer, suggesting that S vacancies are preferentially generated along the crystal edge. While the dominant mechanism for vacancy generation in pristine MoS$_2$ is *via* impact ionization,[43] it is likely that the knock-on threshold of atoms located at the edges is significantly lowered compared to the bulk, such that subsequent S and Mo atoms displacement by knock-on can also play a role.[43] In particular, S atoms are mobile under the electron beam, and interstitial S atoms are observed in the experimental images (Figure 5c). In the presence of S vacancies, the atomic rearrangement leads to the observed shrinkage of the lattice which will inherently induce a local strain in the MoS$_2$ crystal. The atomic arrangement of the shrunk bilayer resembles closely the AA′ stacking. The local build-in strain inside the MoS$_2$ grain is then presumably released by the rotation to a AA′ stacking orientation.

We then examine the edge terminations of the top layer MoS$_2$, because edges can influence the electrical and catalytic properties due to the different bonding between atoms. Atomic

resolution images of bilayer MoS$_2$ indicate that the edges of the second layer of the bilayer are not straight and consist of many atomically sharp zigzag edges. In particular, atomically sharp edges with orientations between them of 120°, highlighted by the yellow dashes line in Figure 4d are observed in the AA′ stacked bilayer MoS$_2$. Two different types of edge configurations are observed in AA′ stacked bilayer MoS$_2$, namely Mo and 50% S−Mo terminated edges (Figure S7). In the case of the 50% S−Mo terminated edges, we observe a different type of edge reconstruction, where the outermost hexagonal ring in the bilayer appears to shrink. The quantitative line profiles indicate that the Mo−Mo distance in the outermost hexagon rings shortens as compared to the pristine layer. This shrinkage can be induced in the presence of single sulfur vacancies at the edge of the bilayer which can change the electrostatic interaction within the hexagon lattice. This can be associated with the lack of one S atom in the outermost bilayer column, which leads to reconstruction of the cage by moving the outermost atoms closer to the inside of the layer.

The formation of intrinsic defects in TMD monolayers grown by a one-step synthesis PLD has a different pathway than that of CVD films. The standard CVD synthesis is performed under a S- or Se-rich environment, in which chalcogen atoms replace oxygen in the metal-oxide precursors.[44] Previous studies have shown that chalcogen vacancies are the predominant intrinsic defects in CVD monolayers. These defects can originate from residual oxygen preserved in the film that can desorb at a later stage, leaving behind chalcogen vacancies in the lattice of TMDs. In the case of PLD monolayers grown by one-step synthesis using sulfides, the presence of a substantial number of chalcogen vacancies (*e.g.*, V$_S$) and antisite defects (*e.g.*, Mo$_S$) hints to a different formation route. To obtain further insight into the formation energy of the observed defects, DFT calculations were performed. Even for materials grown under nonequilibrium conditions, such as PLD, the formation energy is expected to follow a similar correlation with the defect concentration as under thermal equilibrium scenario. The formation energy $E_f[D]$ of a particular defect D was calculated according to

$$E_f[D] = E_{tot}[D] - E_{tot}[\text{pristine}] - \sum_i n_i \mu_i \qquad (1)$$

where $E_f[D]$ is the total energy of the defective supercell and E_{tot}[pristine] the total energy of the pristine supercell.[45,46] The last term accounts for the energy contribution for adding ($n_i >$ 0) or removing ($n_i < 0$) an atom of species i from a reservoir with chemical potential μ_i (details on the calculation of chemical potential limits of point defects are given in the Supplementary Note 1). Our calculations indicate that single S vacancies have a low formation energy (from 1 to 2 eV) under both S-poor and S-rich conditions (Figure 6), suggesting that sulfur vacancies will inherently form under all conditions. Moreover, a similarly low formation energy has been estimated for Mo_S antisite defects under S-poor conditions. The high concentration of these defects in monolayer MoS_2 from PLD synthesis under Mo-rich and S-poor conditions, despite the use of a sulfur cracker source for the synthesis of MoS_2. Indeed, as S has a larger saturated vapor pressure than Mo, after deposition and before crystallization into a stable phase, the sulfur species would be expected to selectively evaporate from the substrate, leading to sulfur vacancies in the MoS_2 lattice. For the same chemical potential range, Mo_{S2} antisite defects and V_{S2} have formation energies larger than V_S and Mo_S.

In addition to intrinsic defects inherited from the synthesis process, defects can be deliberately introduced by the laser-induced plasma. A plasma plume in vacuum consists of energetic ions, neutrals, and clusters propagating in vacuum to the substrate at velocities of 1 cm/μs (corresponding to mean kinetic energies of 48 eV/Mo-atom and 16 eV/S atom) and significantly higher energy tails,[9,11,18,47] which can create defects in the as-grown MoS_2, resulting from implantation or sputtering. The physics of these processes are beyond the scope of this work, especially in the <50 eV/atom regime where few commercial ion sources exist. Recently, however, the effects of Se-cluster impingement on suspended WS_2 and MoS_2 monolayer crystals revealed the regimes for soft landing, implantation, and sputtering all within the <10 eV/Se-atom range.[18] Thus, each PLD pulse delivered in vacuum involves the arrival of highly energetic species that can sputter or crystallize previously grown material, followed by the lower-energy tail of material that can add atoms to heal these defects. This work and a great deal of other research dedicated to intrinsic defects in 2D TMDs suggest that chalcogen vacancies form during fluctuations in the synthesis process[48,49] and can also be deliberately induced by ion/electron bombardment,[50] mild plasma sputtering (H_2 or He),[51] etc. In a simple approximation of a head-on binary collision, the energy E for ion-induced damage shows a dependence on the ion mass given by $E = T_d(m_i + m_t)^2/4m_i m_t$, where T_d is the atom displacement energy, and m_i and m_t are the masses of the ion and target atoms, respectively. Considering T_d(S) of 5.0 eV and T_d(Mo) of 31.7 eV,[52] it can be estimated that it requires an ion energy of 5−6 eV to sputter a S atom and larger than 32 eV to sputter a Mo atom. Moreover, at low ion energies and angles of incidence, single S vacancies are predominately created in the top plane of MoS_2, as more energy is required for the incoming ion to collide with the lattice and remove a S atom in the bottom plane.[52] Interestingly, we have also observed bisulfur vacancies of type V_SV_S consisting of two nearby S vacancies in the top plane of monolayer MoS_2 (see Supplementary Note 9 which includes experimental and simulated STEM images). Notably, the formation energy of V_SV_S is very similar to that of V_{S2} over the whole range of the S chemical potential (Figure 6). Next, we hypothesize that the

availability of single S vacancies can lead to a preferential creation of the observed antisite defects of Mo_S. Under favorable conditions, such defects can be formed by capturing Mo adatoms at the S vacancy sites. These findings allow us to conclude that the observed defects are native in PLD monolayers. Moreover, our calculations show that Mo_SMo_S has the lowest formation energy among all defects under a Mo-rich environment and which increases rapidly from Mo-rich to S-rich conditions. This suggests that the concentration of these cluster defects can be reduced by controlling the synthesis conditions. Moreover, $Mo2_{S2}$ antisite defects were also experimentally observed in this work despite their high formation energy under Mo-rich/S-poor conditions (above 9 eV), suggesting that they may be deliberately induced in PLD, following a similar formation pathway as Mo_S. The vacancy defects V_{MoS3} were not observed in our work. The direct observation of various antisite defects in MoS_2 can provide clues for the poor performance of the PLD-grown monolayers over the CVD-grown monolayers. Previous work on time-domain ab initio calculations have shown that antisite defects and adatoms are more effective than sulfur vacancies in accelerating trapping and nonradiative recombination of charge carriers.[53] Hence, the addition of an extra S reservoir in this work during the growth of MoS_2 is beneficial in reducing S vacancies (and consequently formed antisite defects in MoS_2), while a careful control of adatoms is equally important.

In addition to the aforementioned defects, the PLD specimens host other types of defects, such as edge sites and defects rings constituting the GBs. GBs are common one-dimensional (1D) line defects which inherently form during the growth and which act as a transition from one nm-size grain of a given crystalline orientation to another. The small domain size observed in PLD monolayers can be attributed to the nature of the ablation process. Material transfer from the target to the substrate occurs on a time scale of several microseconds leading to a high instantaneous deposition rate ranging from 0.01 nm/pulse up to 0.1 nm/pulse, which enhances the density of nucleation sites on the substrate. Hence, the high density of nucleation sites in PLD favors the growth of a high density of domains which merge and form GBs where two or more grains meet, resulting in the formation of a quasi-continuous film of 2D nanosheets.[54] In this work, GBs with 4|8 and 5|7 core structures were observed. The formation of dislocation cores appears to be driven by local fluctuations in the Mo/S environment during the growth process and is energetically favorable under a S-poor/Mo-rich environment.[30] We hypothesize that GBs can act as a low-energy barrier for the diffusion of S atoms, which occurs more rapidly than for Mo atoms, leaving behind Mo-rich GB structures. Moreover, the occurrence of an irregular pattern of the 4|8 ring motif constituting the mirror twin GB suggests that the atomistic dynamics across GB are rather complex in nonequilibrium PLD processes involving enhanced mobility of adatoms and high-energy ion-induced modification. The formation of GBs between misoriented crystalline nanosheets as they grow and coalesce, augmented by the effects of hyperthermal impinging species inherent to the PLD process, has not been investigated yet and forms the basis for the atomistic characterization and modeling that follows. Our data reveal a significant change in chemical composition along the GB from pristine MoS_2 to $MoS_{1.5}$. Hence, on a nanometer scale, the GB line defects induce a Mo-rich/S-deficient environment and yield n-type characteristics of MoS_2.[55] The

stoichiometry of the monolayer MoS_2 estimated by X-ray photoelectron spectroscopy (XPS) of 2.02 represents cumulative statistics over a large number of grains, GBs, and crystal edges (see Supplementary Note 7 for details).

While the 60° GBs are predicted to act as 1D metallic wires inside the semiconducting monolayer,[30] the 180° GBs show little or no adverse effect on the electrical transport properties of MoS_2.[56] However, mirror twin GBs have shown to induce a strong PL quenching in 2D TMDs.[23] The 4|8 ring dislocation cores observed in this work are theoretically predicted to introduce midgap states in monolayer MoS_2, which can act as centers for nonradiative recombination and result in quenching of the PL. Moreover, GBs can induce local strain fields in MoS_2 and lower the band gap of the monolayer as compared to the pristine structure.[57] In analogy to the classical semiconductors, band gap tunability at GBs can induce local potential fluctuations at the nanometer scale which can act as traps for charge carriers.

Finally, we elaborate on the growth of bilayer MoS_2, which is captured during the transition growth in the atomic resolution images. The as-grown bilayer crystals are van der Waals coupled and exhibit a wide range of twist angles varying from 0° to 60°. When the growth of the top layer proceeds at crystal edges, the bilayer specimens show exclusively AB to AA′ stacking configurations. On the other hand, moiré stacked bilayers emerge when one grain extends along the GB, growing from this boundary outward on top of the neighboring grain. In this case, the relative orientation of the top and bottom layers is determined by the interdomain misorientation angle. Theoretical calculations have indicated that AB to AA′ configurations have the lowest formation energy in bilayer MoS_2, while moiré bilayers with twist angles between 0° and 60° exhibit a formation energy which is 40 meV/unit cell larger that the AB/AA′ references and which is independent of the twist angle.[58] DFT calculations have predicted that an indirect bandgap will always emerge in bilayer MoS_2[58] and that the bandgap of the bilayer stack widens by 200 meV from the AA- and AB-stacked bilayers to the turbostratic moiré patterns. Our atomic resolution images reveal the presence of bilayer stacks with stacking angles varying significantly on a length of tens of nanometers. Based on these findings, we anticipate that band gap fluctuations are expected in bilayer stacks. Additionally, we have shown electron beam irradiation of MoS_2 bilayers can readily create sulfur vacancies at crystal edges, while Mo atoms remain fairly stable under the beam exposure. As a result, room-temperature electron beam-driven transformation from the AB and AA′ stacking orientation was observed in this study. This unexpected result suggests that the difference in the energy barrier between the AB and AA′ stacking is reduced in the presence of intrinsic point defects, such as S vacancies, and that metastable phases are achieved in PLD. Our observations suggest that electron beam irradiation can be explored as an alternative pathway to engineer a stacking sequence in bilayer MoS_2 at the nanoscale.

CONCLUSIONS

In summary, we have provided a detailed picture of the atomic structure of grains and GBs in monolayers and mono−bilayers MoS_2 grown by pulsed laser deposition (PLD). Direct atomic resolution imaging indicates that sulfur vacancies and Mo_S antisites are the most abundant intrinsic point defects in monolayers MoS_2. The coalescence of the nanometer size grains leads to a high density of GBs with distinct in-plane

orientations. In particular, mirror twin 180° GB structures with 4|8 dislocation cores have been identified to be Mo-rich/S-poor and introduce mid-gap states in MoS_2. The bilayers MoS_2 exhibit not only AB and AA′ stacking but also more complex moiré patterns whose orientations depend on the interdomain misorientation angles. Our findings could help to refine synthesis strategies for the controlled growth of TMDs and can be universally applied to other similar 2D systems grown by this nonequilibrium PLD synthesis method.

METHODS

PLD of MoS_2. The MoS_2 layered films were grown using a hybrid approach based on ablation of a stoichiometric MoS_2 target (from Testbourne, purity 99.95%) in vacuum in the presence of a sulfur evaporation source (from Dr. Eberl, MBE-Komponenten GmbH). The sulfur powder (purity 99.999%) was placed in a reservoir and heated to a temperature of 150 °C. The flux of the sulfur vapor entering the deposition chamber was regulated using a valve. First, the chamber was pumped down to a base pressure of 5×10^{-7} mbar and the pressure increased up to 5×10^{-5} mbar in the presence of sulfur. The sulfur flux was directed toward the substrate during the growth. The target was ablated using a KrF excimer laser (laser wavelength λ = 248 nm and pulse duration τ = 20 ns) operating at a frequency of 1 Hz and laser fluence of 1 J/cm². The sapphire and SiO_2/Si substrates were placed at a distance of 5 cm from the target. Typical growth conditions for one monolayer were a substrate temperature of 700 °C and an average number of 35 pulses.

STEM Sample Preparation and Characterization. The PLD-grown MoS_2 samples were transferred onto a TEM grid with the aid of poly(methyl methacrylate) polymer by using a portable transfer method. The samples transferred on the TEM grid were baked at a temperature of 160 °C in vacuum for 8 h before loading them in the STEM chamber for imaging. The STEM measurements were performed using an aberration-corrected Nion UltraSTEM 100 microscope equipped with a cold field emission gun. The images were acquired at an acceleration voltage of 60 kV and a semiconvergence angle of 31 mrad.

DFT Calculations and STEM Simulations. The DFT calculations were performed within the projected-augmented wave (PAW) method implemented in the electronic structure code GPAW.[59] We used a plane wave basis set with a cutoff of 800 eV and included exchange and correlations with the Perdew−Ernzerhof functional.[60] Reciprocal space was sampled with a k-point density of 6 Å⁻¹ including the Γ point. These settings were applied to the 2H phase of MoS_2.

The lattice parameters were a_{2H} = 3.184 Å and c_{2H} = 18.127 Å in a hexagonal lattice. For the DFT calculations on different point defects, a vacuum size of 15 Å was applied as well as a supercell size of $4 \times 4 \times 1$. The atomic positions were relaxed at a fixed cell until forces were $<10^{-2}$ eVÅ⁻¹. The structural setup for the calculations on GB is discussed in part 3 of the Supporting Information.

STEM images were simulated using the relaxed atomic structures from DFT as an input for PyQSTEM,[26] a python interface for quantitative TEM/STEM simulations using QSTEM.[27] We chose simulation parameters in accordance to the ones of the experimental STEM setup with an acceleration voltage of 60 kV and a convergence angle of 31 mrad.

ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsnano.0c08835.

> Analysis of point defects, calculations of chemical potential limits for the formation energies; STEM images of adatoms; grains and GBs; moiré patterns in bilayer MoS_2; edge termination in AA′ stacked bilayer; processing of STEM images; chemical composition

analysis by XPS; overview of moiré stacking; simulated and experimental images of the V_SV_S defects (PDF)

AUTHOR INFORMATION

Corresponding Authors

Fabian Bertoldo − *CAMD and Center for Nanostructured Graphene (CNG), Department of Physics, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark;* ⊙ orcid.org/0000-0002-1219-8689; Email: fafb@dtu.dk

Stela Canulescu − *Department of Photonics Engineering, Technical University of Denmark, 4000 Roskilde, Denmark;* ⊙ orcid.org/0000-0003-3786-2598; Email: stec@fotonik.dtu.dk

Authors

Raymond R. Unocic − *Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States;* ⊙ orcid.org/0000-0002-1777-8228

Yu-Chuan Lin − *Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States;* ⊙ orcid.org/0000-0003-4958-5073

Xiahan Sang − *State Key Laboratory of Advanced Technology for Materials Synthesis and Processing, Wuhan University of Technology, 430070 Wuhan, China;* ⊙ orcid.org/0000-0002-2861-6814

Alexander A. Puretzky − *Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States;* ⊙ orcid.org/0000-0002-9996-4429

Yiling Yu − *Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States;* ⊙ orcid.org/0000-0001-7180-1137

Denys Miakota − *Department of Photonics Engineering, Technical University of Denmark, 4000 Roskilde, Denmark;* ⊙ orcid.org/0000-0002-3723-915X

Christopher M. Rouleau − *Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States*

Jørgen Schou − *Department of Photonics Engineering, Technical University of Denmark, 4000 Roskilde, Denmark*

Kristian S. Thygesen − *CAMD and Center for Nanostructured Graphene (CNG), Department of Physics, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark;* ⊙ orcid.org/0000-0001-5197-214X

David B. Geohegan − *Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States;* ⊙ orcid.org/0000-0003-0273-3139

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsnano.0c08835

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Koppens, F. H. L.; Mueller, T.; Avouris, P.; Ferrari, A. C.; Vitiello, M. S.; Polini, M. Photodetectors Based on Graphene, Other Two-Dimensional Materials and Hybrid Systems. *Nat. Nanotechnol.* **2014**, *9*, 780−793.

(2) Mak, K. F.; Shan, J. Photonics and Optoelectronics of 2D Semiconductor Transition Metal Dichalcogenides. *Nat. Photonics* **2016**, *10*, 216−226.

(3) Ellis, J. K.; Lucero, M. J.; Scuseria, G. E. The Indirect to Direct Band Gap Transition in Multilayered MoS_2 as Predicted by Screened Hybrid Density Functional Theory. *Appl. Phys. Lett.* **2011**, *99*, 261908.

(4) Hong, X.; Kim, J.; Shi, S.-F.; Zhang, Y.; Jin, C.; Sun, Y.; Tongay, S.; Wu, J.; Zhang, Y.; Wang, F. Ultrafast Charge Transfer in Atomically Thin MoS_2/WS_2 Heterostructures. *Nat. Nanotechnol.* **2014**, *9*, 682−686.

(5) Akama, T.; Okita, W.; Nagai, R.; Li, C.; Kaneko, T.; Kato, T. Schottky Solar Cell Using Few-Layered Transition Metal Dichalcogenides toward Large-Scale Fabrication of Semitransparent and Flexible Power Generator. *Sci. Rep.* **2017**, *7*, 11967.

(6) Pospischil, A.; Furchi, M. M.; Mueller, T. Solar-Energy Conversion and Light Emission in an Atomic Monolayer *p-n* Diode. *Nat. Nanotechnol.* **2014**, *9*, 257−261.

(7) Furchi, M. M.; Pospischil, A.; Libisch, F.; Burgdörfer, J.; Mueller, T. Photovoltaic Effect in an Electrically Tunable van der Waals Heterojunction. *Nano Lett.* **2014**, *14*, 4785−4791.

(8) Tabata, H.; Tanaka, H.; Kawai, T. Formation of Artificial $BaTiO_3/SrTiO_3$ Superlattices Using Pulsed Laser Deposition and Their Dielectric Properties. *Appl. Phys. Lett.* **1994**, *65*, 1970−1972.

(9) Lowndes, D.; Geohegan, D.; Puretzky, A.; Norton, D.; Rouleau, C. Synthesis of Novel Thin-Film Materials by Pulsed Laser Deposition. *Science* **1996**, *273*, 898−903.

(10) Canulescu, S.; Papadopoulou, E.; Anglos, D.; Lippert, T.; Schneider, C.; Wokaun, A. Mechanisms of the Laser Plume Expansion during the Ablation of $LiMn_2O_4$. *J. Appl. Phys.* **2009**, *105*, 063107.

(11) Canulescu, S.; Dobeli, M.; Yao, X.; Lippert, T.; Amoruso, S.; Schou, J. Nonstoichiometric Transfer during Laser Ablation of Metal Alloys. *Physical Review Materials* **2017**, *1*, 73402.

(12) Koster, G.; Kropman, B.; Rijnders, G.; Blank, D.; Rogalla, H. Quasi-Ideal Strontium Titanate Crystal Surfaces through Formation of Strontium Hydroxide. *Appl. Phys. Lett.* **1998**, *73*, 2920−2922.

(13) Rao, S. S.; Prater, J. T.; Wu, F.; Shelton, C. T.; Maria, J. P.; Narayan, J. Interface Magnetism in Epitaxial $BiFeO_3$-$La_{0.7}Sr_{0.3}MnO_3$ Heterostructures Integrated on Si(100). *Nano Lett.* **2013**, *13*, 5814−5821.

(14) Fluri, A.; Pergolesi, D.; Roddatis, V.; Wokaun, A.; Lippert, T. *In Situ* Stress Observation in Oxide Films and How Tensile Stress Influences Oxygen Ion Conduction. *Nat. Commun.* **2016**, *7*, 10692.

(15) Serna, M. I.; Yoo, S. H.; Moreno, S.; Xi, Y.; Oviedo, J. P.; Choi, H.; Alshareef, H. N.; Kim, M. J.; Minary-Jolandan, M.; Quevedo-Lopez, M. A. Large-Area Deposition of MoS_2 by Pulsed Laser Deposition with *in Situ* Thickness Control. *ACS Nano* **2016**, *10*, 6054−6061.

(16) Siegel, G.; Venkata Subbaiah, Y. P.; Prestgard, M. C.; Tiwari, A. Growth of Centimeter-Scale Atomically Thin MoS_2 Films by Pulsed Laser Deposition. *APL Mater.* **2015**, *3*, 056103.

(17) Mahjouri-Samani, M.; Liang, L.; Oyedele, A.; Kim, Y.-S.; Tian, M.; Cross, N.; Wang, K.; Lin, M.-W.; Boulesbaa, A.; Rouleau, C. M.; Puretzky, A. A.; Xiao, K.; Yoon, M.; Eres, G.; Duscher, G.; Sumpter,

B. G.; Geohegan, D. G. Tailoring Vacancies Far beyond Intrinsic Levels Changes the Carrier Type and Optical Response in Monolayer $MoSe_{2-x}$ Crystals. *Nano Lett.* **2016**, *16*, 5213−5220.

(18) Lin, Y.-C.; Liu, C.; Yu, Y.; Zarkadoula, E.; Yoon, M.; Puretzky, A. A.; Liang, L.; Kong, X.; Gu, Y.; Strasser, A.; Meyer, H. M., III; Lorenz, M.; Chisholm, M. F.; Ivanov, I. N.; Rouleau, C. M.; Duscher, G.; Xiao, K.; Geohegan, D. B. Low Energy Implantation into Transition-Metal Dichalcogenide Monolayers to Form Janus Structures. *ACS Nano* **2020**, *14*, 3896−3906.

(19) Yao, J. D.; Zheng, Z. Q.; Shao, J. M.; Yang, G. W. Stable, Highly-Responsive and Broadband Photodetection Based on Large-Area Multilayered WS_2 Films Grown by Pulsed-Laser Deposition. *Nanoscale* **2015**, *7*, 14974−14981.

(20) Reale, F.; Palczynski, P.; Amit, I.; Jones, G. F.; Mehew, J. D.; Bacon, A.; Ni, N.; Sherrell, P. C.; Agnoli, S.; Craciun, M. F.; Russo, S.; Mattevi, C. High-Mobility and High-Optical Quality Atomically Thin WS_2. *Sci. Rep.* **2017**, *7*, 14911.

(21) Huo, N.; Yang, Y.; Wu, Y.-N.; Zhang, X.-G.; Pantelides, S. T.; Konstantatos, G. High Carrier Mobility in Monolayer CVD-Grown MoS_2 through Phonon Suppression. *Nanoscale* **2018**, *10*, 15071−15077.

(22) Tumino, F.; Casari, C. S.; Passoni, M.; Russo, V.; Li Bassi, A. Pulsed Laser Deposition of Single-Layer MoS_2 on Au(111): From Nanosized Crystals to Large-Area Films. *Nanoscale Advances* **2019**, *1*, 643−655.

(23) van der Zande, A. M.; Huang, P. Y.; Chenet, D. A.; Berkelbach, T. C.; You, Y.; Lee, G.-H.; Heinz, T. F.; Reichman, D. R.; Muller, D. A.; Hone, J. C. Grains and Grain Boundaries in Highly Crystalline Monolayer Molybdenum Disulphide. *Nat. Mater.* **2013**, *12*, 554−561.

(24) Zhu, J.; Wang, Z.; Yu, H.; Li, N.; Zhang, J.; Meng, J.; Liao, M.; Zhao, J.; Lu, X.; Du, L.; Yang, R.; Shi, D.; Jiang, Y.; Zhang, G. Argon Plasma Induced Phase Transition in Monolayer MoS_2. *J. Am. Chem. Soc.* **2017**, *139*, 10216−10219.

(25) Krivanek, O. L.; Chisholm, M. F.; Nicolosi, V.; Pennycook, T. J.; Corbin, G. J.; Dellby, N.; Murfitt, M. F.; Own, C. S.; Szilagyi, Z. S.; Oxley, M. P.; Pantelides, S. T.; Pennycook, S. J. Atom-by-Atom Structural and Chemical Analysis by Annular Dark-Field Electron Microscopy. *Nature* **2010**, *464*, 571−574.

(26) Madsen, J.; Liu, P.; Wagner, J. B.; Hansen, T. W.; Schiøtz, J. Accuracy of Surface Strain Measurements from Transmission Electron Microscopy Images of Nanoparticles. *Adv. Struct. Chem. Imaging* **2017**, *3*, 14.

(27) Koch, C. Determination of Core Structure Periodicity and Point Defect Density along Dislocations. *Ph.D. Dissertation*, Arizona State University, Tempe, AZ, 2002.

(28) González, C.; Biel, B.; Dappe, Y. J. Theoretical Characterisation of Point Defects on a MoS_2 Monolayer by Scanning Tunnelling Microscopy. *Nanotechnology* **2016**, *27*, 105702.

(29) Cong, W. T.; Tang, Z.; Zhao, X. G.; Chu, J. H. Enhanced Magnetic Anisotropies of Single Transition-Metal Adatoms on a Defective MoS_2 Monolayer. *Sci. Rep.* **2015**, *5*, 9361.

(30) Zhou, W.; Zou, X.; Najmaei, S.; Liu, Z.; Shi, Y.; Kong, J.; Lou, J.; Ajayan, P. M.; Yakobson, B. I.; Idrobo, J.-C. Intrinsic Structural Defects in Monolayer Molybdenum Disulfide. *Nano Lett.* **2013**, *13*, 2615−2622.

(31) Hong, J.; Hu, Z.; Probert, M.; Li, K.; Lv, D.; Yang, X.; Gu, L.; Mao, N.; Feng, Q.; Xie, L.; Zhang, J.; Wu, D.; Zhang, Z.; Jin, C.; Ji, W.; Zhang, X.; Yuan, J.; Zhang, Z. Exploring Atomic Defects in Molybdenum Disulphide Monolayers. *Nat. Commun.* **2015**, *6*, 6293.

(32) Ly, T. H.; Perello, D. J.; Zhao, J.; Deng, Q.; Kim, H.; Han, G. H.; Chae, S. H.; Jeong, H. Y.; Lee, Y. H. Misorientation-Angle-Dependent Electrical Transport across Molybdenum Disulfide Grain Boundaries. *Nat. Commun.* **2016**, *7*, 10426.

(33) Yao, J. D.; Zheng, Z. Q.; Yang, G. W. Production of Large-Area 2D Materials for High-Performance Photodetectors by Pulsed-Laser Deposition. *Prog. Mater. Sci.* **2019**, *106*, 100573.

(34) Loh, T. A. J.; Chua, D. H. C.; Wee, A. T. S. One-Step Synthesis of Few-Layer WS_2 by Pulsed Laser Deposition. *Sci. Rep.* **2016**, *5*, 18116.

(35) Xu, X.; Wang, Z.; Lopatin, S.; Quevedo-Lopez, M. A.; Alshareef, H. N. Wafer Scale Quasi Single Crystalline MoS_2 Realized by Epitaxial Phase Conversion. *2D Mater.* **2019**, *6*, 015030.

(36) Yu, Y.; Li, C.; Liu, Y.; Su, L.; Zhang, Y.; Cao, L. Controlled Scalable Synthesis of Uniform, High-Quality Monolayer and Few-Layer MoS_2 Films. *Sci. Rep.* **2013**, *3*, 1866.

(37) Zhao, X.; Fu, D.; Ding, Z.; Zhang, Y.-Y.; Wan, D.; Tan, S. J.; Chen, Z.; Leng, K.; Dan, J.; Fu, W.; Geng, D.; Song, P.; Du, Y.; Venkatesan, T.; Pantelides, S. T.; Pennycook, S. J.; Zhou, W.; Loh, K. P. Mo-Terminated Edge Reconstructions in Nanoporous Molybdenum Disulfide Film. *Nano Lett.* **2018**, *18*, 482−490.

(38) Hu, G.; Fung, V.; Sang, X.; Unocic, R. R.; Ganesh, P. Predicting Synthesizable Multi-Functional Edge Reconstructions in Two-Dimensional Transition Metal Dichalcogenides. *npj Comput. Mater.* **2020**, *6*, 44.

(39) Hu, G.; Fung, V.; Sang, X.; Unocic, R. R.; Ganesh, P. Superior Electrocatalytic Hydrogen Evolution at Engineered Non-Stoichiometric Two-Dimensional Transition Metal Dichalcogenide Edges. *J. Mater. Chem. A* **2019**, *7*, 18357−18364.

(40) Sang, X.; Li, X.; Zhao, W.; Dong, J.; Rouleau, C. M.; Geohegan, D. B.; Ding, F.; Xiao, K.; Unocic, R. R. *In Situ* Edge Engineering in Two-Dimensional Transition Metal Dichalcogenides. *Nat. Commun.* **2018**, *9*, 2051.

(41) Yan, A.; Chen, W.; Ophus, C.; Ciston, J.; Lin, Y.; Persson, K.; Zettl, A. Identifying Different Stacking Sequences in Few-Layer CVD-Grown MoS_2 by Low-Energy Atomic-Resolution Scanning Transmission Electron Microscopy. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2016**, *93*, 041420.

(42) Yan, A.; Ong, C. S.; Qiu, D. Y.; Ophus, C.; Ciston, J.; Merino, C.; Louie, S. G.; Zettl, A. Dynamics of Symmetry-Breaking Stacking Boundaries in Bilayer MoS_2. *J. Phys. Chem. C* **2017**, *121*, 22559−22566.

(43) Zan, R.; Ramasse, Q. M.; Jalil, R.; Georgiou, T.; Bangert, U.; Novoselov, K. S. Control of Radiation Damage in MoS_2 by Graphene Encapsulation. *ACS Nano* **2013**, *7*, 10167−10174.

(44) Dumcenco, D.; Ovchinnikov, D.; Marinov, K.; Lazić, P.; Gibertini, M.; Marzari, N.; Sanchez, O. L.; Kung, Y.-C.; Krasnozhon, D.; Chen, M.-W.; Bertolazzi, S.; Gillet, P.; Fontcuberta i Morral, A.; Radenovic, A.; Kis, A. Large-Area Epitaxial Monolayer MoS_2. *ACS Nano* **2015**, *9*, 4611−4620.

(45) Zhang, S. B.; Northrup, J. E. Chemical Potential Dependence of Defect Formation Energies in GaAs: Application to Ga Self-Diffusion. *Phys. Rev. Lett.* **1991**, *67*, 2339−2342.

(46) Van De Walle, C. G.; Laks, D. B.; Neumark, G. F.; Pantelides, S. T. First-Principles Calculations of Solubilities and Doping Limits: Li, Na, and N in ZnSe. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1993**, *47*, 9425−9434.

(47) Mahjouri-Samani, M.; Gresback, R.; Tian, M.; Wang, K.; Puretzky, A. A.; Rouleau, C. M.; Eres, G.; Ivanov, I. N.; Xiao, K.; McGuire, M. A.; Duscher, G.; Geohegan, D. B. Pulsed Laser Deposition of Photoresponsive Two-Dimensional GaSe Nanosheet Networks. *Adv. Funct. Mater.* **2014**, *24*, 6365−6371.

(48) Gutiérrez, H. R.; Perea-López, N.; Elías, A. L.; Berkdemir, A.; Wang, B.; Lv, R.; López-Urías, F.; Crespi, V. H.; Terrones, H.; Terrones, M. Extraordinary Room-Temperature Photoluminescence in Triangular WS_2 Monolayers. *Nano Lett.* **2013**, *13*, 3447−3454.

(49) Peimyoo, N.; Shang, J.; Cong, C.; Shen, X.; Wu, X.; Yeow, E. K. L.; Yu, T. Nonblinking, Intense Two-Dimensional Light Emitter: Monolayer WS_2 Triangles. *ACS Nano* **2013**, *7*, 10985−10994.

(50) Komsa, H.-P.; Kurasch, S.; Lehtinen, O.; Kaiser, U.; Krasheninnikov, A. V. From Point to Extended Defects in Two-Dimensional MoS_2: Evolution of Atomic Structure under Electron Irradiation. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *88*, 035301.

(51) Tosun, M.; Chan, L.; Amani, M.; Roy, T.; Ahn, G. H.; Taheri, P.; Carraro, C.; Ager, J. W.; Maboudian, R.; Javey, A. Air-Stable *n*-Doping of WSe_2 by Anion Vacancy Formation with Mild Plasma Treatment. *ACS Nano* **2016**, *10*, 6853−6860.

(52) Ghorbani-Asl, M.; Kretschmer, S.; Spearot, D. E.; Krasheninnikov, A. V. Two-Dimensional MoS$_2$ under Ion Irradiation: From Controlled Defect Production to Electronic Structure Engineering. *2D Mater.* **2017**, *4*, 025078.

(53) Li, L.; Long, R.; Prezhdo, O. V Why Chemical Vapor Deposition Grown MoS$_2$ Samples Outperform Physical Vapor Deposition Samples: Time-Domain *ab Initio* Analysis. *Nano Lett.* **2018**, *18*, 4008−4014.

(54) Puretzky, A. A.; Lin, Y.-C.; Liu, C.; Strasser, A. M.; Yu, Y.; Canulescu, S.; Rouleau, C. M.; Xiao, K.; Duscher, G.; Geohegan, D. B. *In Situ* Laser Reflectivity to Monitor and Control the Nucleation and Growth of Atomically Thin 2D Materials. *2D Mater.* **2020**, *7*, 025048.

(55) Yang, J.; Kawai, H.; Wong, C. P. Y.; Goh, K. E. J. Electrical Doping Effect of Vacancies on Monolayer MoS$_2$. *J. Phys. Chem. C* **2019**, *123*, 2933−2939.

(56) Najmaei, S.; Amani, M.; Chin, M. L.; Liu, Z.; Birdwell, A. G.; O'Regan, T. P.; Ajayan, P. M.; Dubey, M.; Lou, J. Electrical Transport Properties of Polycrystalline Monolayer Molybdenum Disulfide. *ACS Nano* **2014**, *8*, 7930−7937.

(57) Huang, Y. L.; Chen, Y.; Zhang, W.; Quek, S. Y.; Chen, C.-H.; Li, L.-J.; Hsu, W.-T.; Chang, W.-H.; Zheng, Y. J.; Chen, W.; Wee, A. T. S. Bandgap Tunability at Single-Layer Molybdenum Disulphide Grain Boundaries. *Nat. Commun.* **2015**, *6*, 6298.

(58) Liu, K.; Zhang, L.; Cao, T.; Jin, C.; Qiu, D.; Zhou, Q.; Zettl, A.; Yang, P.; Louie, S. G.; Wang, F. Evolution of Interlayer Coupling in Twisted Molybdenum Disulfide Bilayers. *Nat. Commun.* **2014**, *5*, 4966.

(59) Enkovaara, J.; Rostgaard, C.; Mortensen, J. J.; Chen, J.; Dułak, M.; Ferrighi, L.; Gavnholt, J.; Glinsvad, C.; Haikola, V.; Hansen, H. A.; Kristoffersen, H. H.; Kuisma, M.; Larsen, A. H.; Lehtovaara, L.; Ljungberg, M.; Lopez-Acevedo, O.; Moses, P. G.; Ojanen, J.; Olsen, T.; Petzold, V.; et al. Electronic Structure Calculations with GPAW: A Real-Space Implementation of the Projector Augmented-Wave Method. *J. Phys.: Condens. Matter* **2010**, *22*, 253202.

(60) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.

9.2   Publication II: Quantum Point Defects in 2D Materials - the QPOD Database

F. BERTOLDO
S. ALI
S. MANTI
K. S. THYGESEN

ARTICLE   OPEN

Quantum point defects in 2D materials - the QPOD database

Fabian Bertoldo [1,2✉], Sajid Ali [1,2✉], Simone Manti[1,2] and Kristian S. Thygesen [1]

Atomically thin two-dimensional (2D) materials are ideal host systems for quantum defects as they offer easier characterisation, manipulation and read-out of defect states as compared to bulk defects. Here we introduce the Quantum Point Defect (QPOD) database with more than 1900 defect systems comprising various charge states of 503 intrinsic point defects (vacancies and antisites) in 82 different 2D semiconductors and insulators. The Atomic Simulation Recipes (ASR) workflow framework was used to perform density functional theory (DFT) calculations of defect formation energies, charge transition levels, Fermi level positions, equilibrium defect and carrier concentrations, transition dipole moments, hyperfine coupling, and zero-field splitting. Excited states and photoluminescence spectra were calculated for selected high-spin defects. In this paper we describe the calculations and workflow behind the QPOD database, present an overview of its content, and discuss some general trends and correlations in the data. We analyse the degree of defect tolerance as well as intrinsic dopability of the host materials and identify promising defects for quantum technological applications. The database is freely available and can be browsed via a web-app interlinked with the Computational 2D Materials Database (C2DB).

INTRODUCTION

Point defects are ubiquitous entities affecting the properties of any crystalline material. Under equilibrium conditions their concentration is given by the Boltzmann distribution, but strong deviations can occur in synthesised samples due to non-equilibrium growth conditions and significant energy barriers involved in the formation, transformation, or annihilation of defects. In many applications of semiconductor materials, in particular those relying on efficient carrier transport, the presence of defects has a detrimental impact on performance[1]. However, point defects in crystals can also be useful and form the basis for applications e.g. in spintronics[2,3], quantum computing[4,5], or quantum photonics[6–9]. For such applications, defects may be introduced in a (semi)controlled manner e.g. by electron/ion beam irradiation, plasma treatment or high-temperature annealing in the presence of different gasses[10].

Over the past decade, atomically thin two-dimensional (2D) crystals have emerged as a promising class of materials with many attractive features including distinct, easily tunable, and often superior physical properties[11]. This holds in particular for their defect-based properties and related applications. Compared to point defects buried deep inside a bulk structure, defects in 2D materials are inherently surface-near making them easier to create, manipulate, and characterise[12]. Recently, single photon emission (SPE) has been observed from point defects in 2D materials such as hexagonal boron-nitride (hBN)[8–10,13], MoS_2[14], and WSe_2[15,16], and in a few cases optically detected magnetic resonance (ODMR) has been demonstrated[17,18]. In the realm of catalysis, defects can act as active sites on otherwise chemically inert 2D materials[19,20].

First-principles calculations based on density functional theory (DFT) can provide detailed insight into the physics and chemistry of point defects and how they influence the structural, electronic and optical properties[21–24]. In particular, such calculations have become an indispensable tool for interpreting the results of experiments on defects, e.g. (magneto-)optical experiments,

which in themselves only provide indirect information about the microscopic nature of the involved defect[10,25,26]. In combination with recently developed tools for high-throughput workflow management[27–30], first-principles calculations have potential to play a more proactive role in the search for defect systems with promising properties. Here, a major challenge is the notorious complexity of defect calculations (even when performed in low-throughput mode) that involves large supercells, local magnetic moments, electrostatic corrections, etc. Performing such calculations for general defects in general host materials, requires a carefully designed workflow with optimised computational settings and a substantial amount of benchmarking[31].

In this work, we present a systematic study of 503 distinct intrinsic point defects (vacancies and antisite defects) in 82 insulating 2D host (monolayer) materials. The host materials were selected from the Computational 2D Materials Database (C2DB)[32,33] after applying a series of filtering criteria. Our computational workflow incorporates the calculation and analysis of thermodynamic properties such as defect formation energies, charge transition levels (CTLs), equilibrium carrier concentrations and Fermi level position, as well as symmetry analysis of the defect atomic structures and wave functions, magnetic properties such as hyperfine coupling parameters and zero-field splittings, and optical transition dipoles. Defects with a high-spin ground state are particularly interesting for magneto-optical and quantum information technology applications. For such defects the excited state properties, zero phonon line energies, radiative lifetime, and photoluminescence (PL) lineshapes, were also calculated.

The computational defect workflow was constructed with the Atomic Simulation Recipes (ASR)[29] and executed using the MyQueue[30] task scheduler frontend. The ASR provides a simple and modular framework for developing Python workflow scripts, and its automatic caching system keeps track of job status and logs data provenance. Our ASR defect workflow adds to other ongoing efforts to automate the computational characterisation of point defects[34–37]. However, to the best of our knowledge, the

[1]CAMD, Computational Atomic-Scale Materials Design, Department of Physics, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark. [2]These authors contributed equally: Fabian Bertoldo, Sajid Ali, Simone Manti. ✉email: fafb@dtu.dk; sajal@dtu.dk

present work represents the first actual high-throughput study of point defects. All of the generated data is collected in the Quantum Point Defect (QPOD) database with over 1900 rows and will be publicly available and accessible via a browsable web-service. The QPOD webpages are interlinked with the C2DB providing a seamless interface between the properties of the pristine host materials and their intrinsic point defects.

The theoretical framework is based entirely on DFT with the Perdew Burke Ernzerhof (PBE) functional[38]. Charge transition levels are obtained using Slater-Janak transition state theory[39] while excited states are calculated with the direct optimisation (DO) and maximum overlap method (MOM) method[40]. We note that PBE suffers from delocalisation errors[41], which may introduce quantitative inaccuracies in the description of some localised defect states, as well as an underestimation of the band gap[42,43]. While range-separated hybrid functionals[44–46] or the GW approach[47–51] represent the state-of-the art methodology for point defect calculations, such a description is currently too demanding for large-scale studies like the current. Moreover, thermodynamic properties of defects are generally well described by PBE[52].

First, we describe the theory and methodology employed at the various computational steps of the workflow. Second, we give a general overview of the workflow, introduce the set of host materials and the considered point defects, and outline the structure and content of the QPOD database web-interface. Afterwards, we present our main results. These include statistical overviews of host crystal and defect system properties, analysis of the effect of structural relaxations on defect formation energies and charge transition levels, an evaluation of intrinsic (equilibrium) doping levels and identification of a small subset of the host materials that are particularly defect tolerant. We also identify a few defect systems with promising properties for spin qubit applications or nanoscale magneto-optical sensing. Finally, we summarise the work and look ahead.

RESULTS AND DISCUSSION

Supercell and defect structures

A supercell in 2D can be created using linear combinations of the primitive unit cell vectors ($\mathbf{a}_1$, $\mathbf{a}_2$). The corresponding supercell lattice vectors ($\mathbf{b}_1$, $\mathbf{b}_2$) are written as:

$$\mathbf{b}_1 = n_1 \mathbf{a}_1 + n_2 \mathbf{a}_2, \tag{1}$$

$$\mathbf{b}_2 = m_1 \mathbf{a}_1 + m_2 \mathbf{a}_2, \tag{2}$$

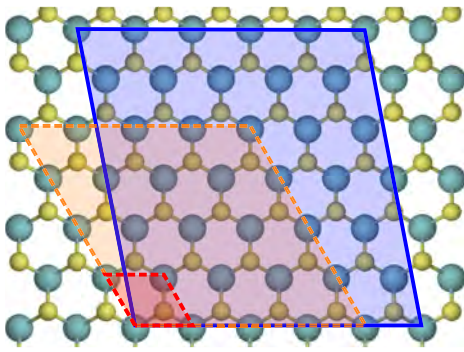where n_1, n_2, m_1, m_2 are integers (see Fig. 1).

Fig. 1 Creation of defect supercells for the example of MoS$_2$. Top view of a primitive unit cell of MoS$_2$ (red, dashed), an example of a conventional $4 \times 4 \times 1$ supercell (orange, dashed), as well as our approach of a symmetry broken supercell (blue, solid).

In this study, we apply an algorithm that finds the most suitable supercell according to the following criteria: (i) Set $n_2 = 0$ and create all supercells defined by n_1, m_1, m_2 between 0 and 10. (ii) Discard combinations where $m_1 = 0$ and $n_1 = m_2$. (iii) Keep only cells where the minimum distance between periodically repeated defects is larger than 15 Å. (iv) Keep supercells containing the smallest number of atoms. (v) Pick the supercell that yields the most homogeneous distribution of defects. We note that step (ii) is conducted in order to break the symmetry of the initial Bravais lattice, and step (iv) minimizes computational cost for a viable high-throughput execution.

Defects are introduced by analysing Wyckoff positions of the atoms within the primitive unit cell. For each non-equivalent position in the structure a vacancy defect and substitutional defects (called antisites in the following) are created (the latter by replacing a specific atom with another atom of a different species intrinsic to the host material). For the example of MoS$_2$ the procedure yields the following point defects: sulfur vacancy V_S, molybdenum vacancy V_{Mo}, as well as two substitutional defects Mo_S and S_{Mo} where Mo replaces the S atom and vice versa. Each defect supercell created by this approach undergoes the workflow which is presented in Section "The QPOD workflow".

Defect formation energy

The formation energy of a defect X in charge state q is defined by[53,54]:

$$E^f[X^q] = E_{tot}[X^q] - E_{tot}[bulk] - \sum_i n_i \mu_i + q E_F \tag{3}$$

where μ_i is the chemical potential of the atom species i and n_i is the number of such atoms that have been added ($n_i > 0$) or removed ($n_i < 0$) in order to create the defect. Equation (3) does not account for vibrational and entropic contributions which are small at room temperature[21]. In this work we set μ_i equal to the total energy of the the standard state of element i. We note, that in general μ_i can be varied in order to represent i-rich and i-poor conditions[55]. On the web page of QPOD one can look up the respective i-poor and i-rich conditions (further information see SI). For finite charge states, the defect formation energy becomes a function of the Fermi energy, E_F, which represents the chemical potential of electrons. In equilibrium, the concentration of a specific defect type is determined by its formation energy, which in turn depends on E_F. Imposing global charge neutrality leads to a self-consistency problem for E_F, which we discuss in Section "Equilibrium defect concentrations".

In general, the lower the formation energy of a particular defect is, the higher is the probability for it to be present in the material. In equilibrium, the defect concentration is given by the Boltzmann distribution,

$$C^{eq}[X^q] = N_X g_{X^q} \exp\left(-E^f[X^q]/(k_B T)\right), \tag{4}$$

where N_X and g_{X^q} specifies the site and defect state degeneracy, respectively, k_B is Boltzmann's constant and T is the temperature.

As an example, Fig. 2 shows the formation energy (blue solid lines) of a sulfur vacancy V_S in MoS$_2$ as a function of the Fermi level position. It follows that this particular defect is most stable in its neutral charge state ($q = 0$) for low to mid gap Fermi level positions. The transition from $q = 0$ to $q' = -1$ occurs for E_F around 1.2 eV. The calculated formation energy of the neutral V_S and its CTLs are in good agreement with reported values from the literature[56]. The differences are below 0.2 eV and can be ascribed to the difference in the employed supercells (symmetric vs. symmetry-broken) and xc-functionals (PBE-D vs. PBE). Moreover, in this work the CTLs are obtained using the Slater-Janak transition state theory (see Section "Slater-Janak transition state theory") while Komsa et al.[56] used total energy differences with electrostatic corrections.
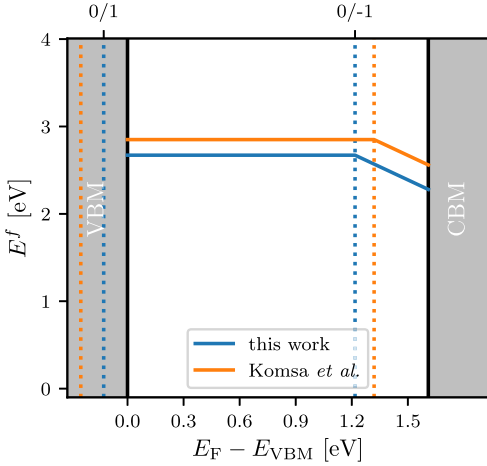
Fig. 2 Formation energy of V_S in MoS$_2$ referenced to the standard states as a function of Fermi energy. Calculated formation energy (blue, solid) as a function of the Fermi energy plotted together with the CTL (blue, dotted). The orange solid lines (orange dotted lines) highlight the PBE-D calculated formation energy (CTL of $(0/-1)$ and $(0/1)$) of V_S in MoS$_2$ taken from Komsa et al.[56].

Slater-Janak transition state theory

The prediction of charge transition levels requires the total energy of the defect in different charge states. In the standard approach, the extra electrons/holes are included in the self-consistent DFT calculation and a background charge distribution is added to make the supercell overall charge neutral. In a post process step, the spurious interactions between periodically repeated images is removed from the total energy using an electrostatic correction scheme that involves a Gaussian approximation to the localised charge distribution and a model for the dielectric function of the material[21]. While this approach is fairly straightforward and unambiguous for bulk materials, it becomes significantly more challenging for 2D materials due to the spatial confinement and non-local nature of the dielectric function and the dependence on detailed shape of the neutralising background charge[57–60].

To avoid the difficulties associated with electrostatic corrections, we rely on the Slater-Janak (SJ) theorem[39], which relates the Kohn-Sham eigenvalue ε_i to the derivative of the total energy with respect to the orbital occupation number n_i,

$$\frac{\partial E}{\partial n_i} = \varepsilon_i(n_i). \tag{5}$$

The theorem can be used to express the difference in ground state energy between two charge states as an integral over the eigenvalue as its occupation number is changed from 0 to 1. This approached, termed SJ transition state theory, has been successfully used to evaluate core-level shifts in random alloys[61], and CTLs of impurities in GaN[62], native defects in LiNbO$_3$[63], and chalcogen vacancies in monolayer TMDs[64]. Assuming a linear dependence of ε_i on n_i (which holds exactly for the true Kohn-Sham system), the transition energy between two localised states of charge q and $q' = q \pm 1$ can be written as

$$\varepsilon(q/q') = \begin{cases} \varepsilon_H(q + \frac{1}{2}; R_q) - \lambda_{q'}, & q' = q + 1 \\ \varepsilon_H(q - \frac{1}{2}; R_q) + \lambda_{q'}, & q' = q - 1. \end{cases} \tag{6}$$

Here, ε_H represents highest eigenvalue with non-zero occupation, i.e. the half occupied state, and R_q refers to the configuration of charge state q. The reorganisation energy is obtained as a total

energy difference between equal charge states

$$\lambda_{q'} = E_{tot}(q'; R_{q'}) - E_{tot}(q'; R_q). \tag{7}$$

Note that the reorganisation energy is always negative. The relevant quantities are illustrated graphically in Fig. 3.

It has been shown for defects in bulk materials that the CTLs obtained from SJ theory are in good agreement with results obtained from total energy differences[61,63]. For 2D materials, a major advantage of the SJ method is that it circumvents the issues related to the electrostatic correction, because it completely avoids the comparison of energies between supercells with different number of electrons. The Kohn-Sham eigenvalues of a neutral or (partially) charged defect supercell are referenced relative to the electrostatic potential averaged over the PAW sphere of an atom located as far as possible away from the defect site (typically around 7 Å depending on the exact size of the supercell). By performing the potential average over an equivalent atom of the pristine 2D layer, we can reference the Kohn-Sham eigenvalue of the defect supercell to the VBM of the pristine material. As an alternative to averaging the potential around an atom, the asymptotic vacuum potential can be used as reference. We have checked that the two procedures yield identical results (usually within 0.1 eV), but prefer the atom-averaging scheme as it can be applied to bulk materials as well.

Equilibrium defect concentrations

According to Eqs. (3) and (4), the formation energy of charged defects, and therefore their equilibrium concentration, is a function of the Fermi level. The Fermi level position of the system in thermal equilibrium is then determined self-consistently from a requirement of charge neutrality[65]

$$\sum_X \sum_q qC[X^q] = n_0 - p_0 \tag{8}$$

where the sum is over structural defects X in charge states q, and n_0 and p_0 are the electron and hole carrier concentrations, respectively. The latter are given by

$$n_0 = \int_{E_{gap}}^{\infty} dE f(E)\rho(E), \tag{9}$$

$$p_0 = \int_{-\infty}^{0} dE[1 - f(E)]\rho(E) \tag{10}$$

where $\rho(E)$ is the local density of states, $f(E) = 1/\exp[(E - E_F)/k_B T]$ is the Fermi-Dirac distribution, and the energy scale is referenced to the valence band maximum of the pristine crystal.

Under the assumption that all the relevant defects, i.e. the intrinsic defects with the lowest formation energies, are accounted for, Equation (8) will determine the Fermi level position of the material in thermal equilibrium. The equilibrium Fermi level position determines whether a material is intrinsically p-doped, n-doped, or intrinsic. The three different cases are illustrated schematically in Fig. 4. For the n-type case (left panel), the most stable defect is D$_1$ in charge state $+1$. Thus, charge carriers are transferred from the defect into the conduction band resulting in a Fermi level just below the CBM. Similarly, for the p-type case (middle panel) the defect D$_1$ in charge state -1 is the most stable. Consequently, charge carriers are promoted from the valence band into the defect resulting in E_F close to the VBM. In the right panel, donor and acceptor states are competing, which results in an effective cancelling of the p- and n-type behavior, pinning the Fermi level in the middle of the band gap. In Section "Intrinsic carrier concentrations" we analyse the intrinsic carrier types and concentrations of 58 host materials where we set g_{X^q} and N_X in Equation (4) to one.
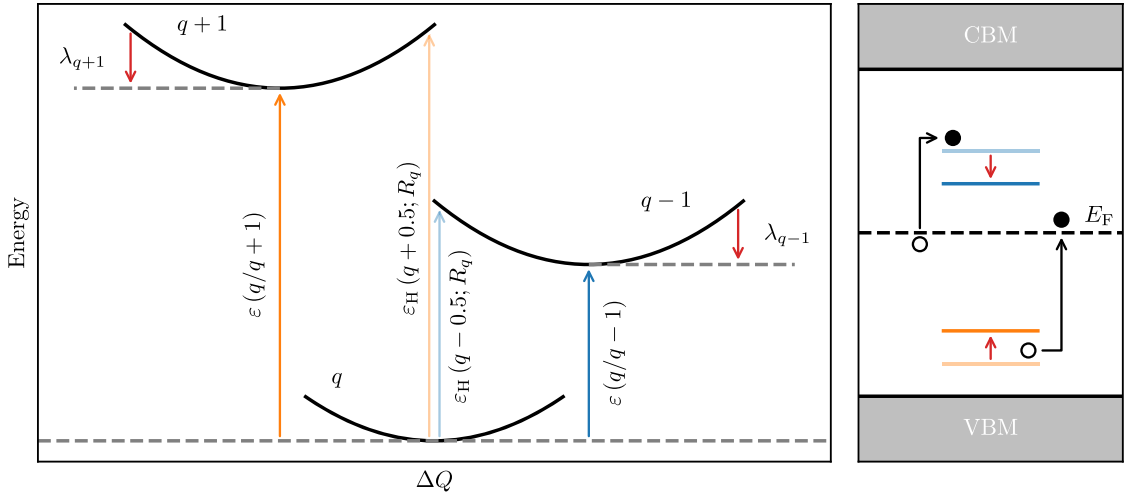
Fig. 3 Schematic view of the SJ approach to calculate charge transition levels. Left: the vertical charge transition levels are calculated starting from a given charge state q by either adding (i.e. $q-0.5$) or removing (i.e. $q+0.5$) half an electron and calculating the HOMO energies at frozen atomic configuration (i.e. $\varepsilon_H(q\pm0.5;R_q)$, light colored arrows in left panel). Due to ionic relaxations one has to correct with the reorganization energies $\lambda_{q\pm1}$ to reduce the energy cost for the transition wrt. E_F. Right: the reorganization energy for adding (removing) an electron has to be added (subtracted) from the vertical transition levels (see red arrows in the right panel) in order to compute the thermodynamic CTLs $\varepsilon(q/q\pm1)$.
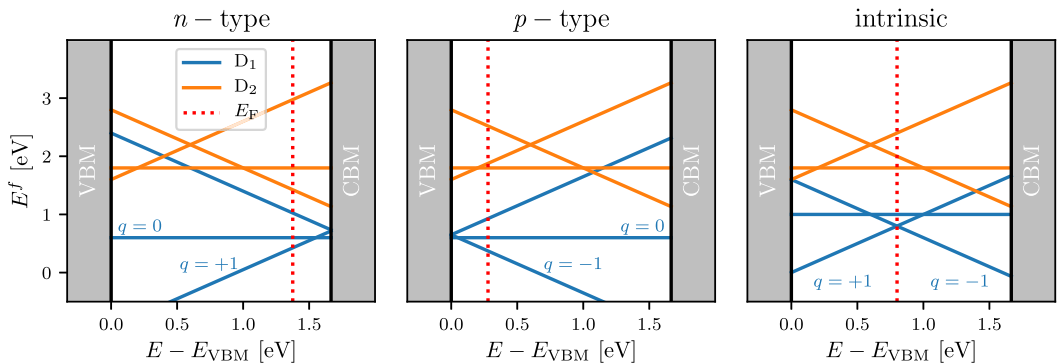

Fig. 4 Intrinsic dopability of defect systems. Formation energies with respect to the standard states as a function of energy for three mock-sytems with two defect types present, respectively. Left: n-type dopable regime (E_F close to CBM) with dominating donor defect. Middle: p-type dopable regime (E_F close to VBM) with dominating acceptor defect. Right: intrinsic material (neither p- nor n-dopable due to the presence of competing acceptor and donor defects).

Symmetry analysis

All states with an energy in the band gap are classified according to the symmetry group of the defect using a generalization of the methodology previously implemented in GPAW for molecules[66]. In a first step, the point group, G, of the defect is determined. To determine G we first reintroduce the relaxed defect into a supercell that preserves the symmetry of the host material, i.e. map the relaxed atomic positions from the symmetry broken calculation into a simple NxNx1 repetition of the primitive unit cell vectors. This is needed for spglib[67] to correctly obtain G as the point group of the symmetry broken supercell. We stress that the high-symmetry supercell is only used to determine G while all actual calculations are performed for the low-symmetry supercell.

The defect states in the band gap are labeled according to the irreducible representations (irrep) of G. To obtain the irrep of a

given eigenstate, $\psi_n(\mathbf{r})$, we form the matrix elements

$$\Gamma(R) = \Gamma_{nn}(R) = \int d\mathbf{r}\ \psi_n(\mathbf{r})^* R\ \psi_n(\mathbf{r}) \qquad (11)$$

where R is any symmetry transformation of G. It follows from the orthogonality theorem[68] that the vector $\Gamma(R)$ can be expanded in the character vectors, $\chi^{(a)}(R)$,

$$\Gamma(R) = \sum_a c_a \chi^{(a)}(R) \qquad (12)$$

where the quantity c_a represents the fraction of ψ_n that transforms according to the irrep a. For any well localised defect state, all the c_a will be zero except one, which is the irrep of the state. In general, less localised states will not transform according to an irrep of G. That is because the wave functions are calculated in the symmetry-broken supercell, see Section "Supercell and defect
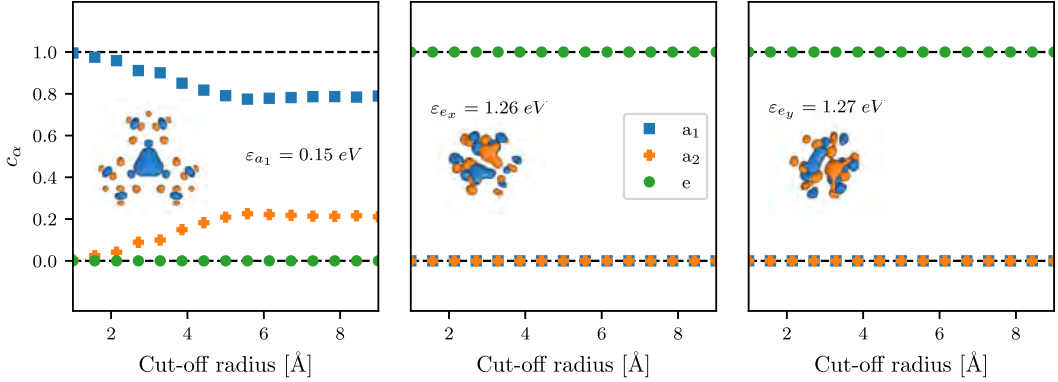
Fig. 5 Orbital symmetry labels for different cutoff radii. All defect states with an energy inside the band gap of the host material are classified according to the irreps of the point group of the defect. The c_a coefficient is a measure of the degree to which a defect state transforms according to irrep a. Performing the symmetry analysis within a radius of a few Å of the center of symmetry leads to a correct classification of the totally symmetric a_1 state and the degenerate e_x and e_y states of the neutral sulfur vacancy in MoS$_2$. The isosurfaces of the orbitals are shown as insets and the energy eigenvalues reference to the VBM.

structures", and therefore will have a lower symmetry than the defect. To exclude such low-symmetry tails on the wave functions, the integral in Equation (11) is truncated beyond a cut-off radius measured from the center of symmetry of the defect.

As an example, Fig. 5 shows the coefficients c_a for the in-gap states of the (neutral) sulfur vacancy in MoS$_2$. This is a well studied prototypical defect with C_{3v} symmetry. The absence of the chalcogen atom introduces three defect states in the gap: a totally symmetric a_1 state close to the VBM and two doubly degenerate mid-gap states (e_x, e_y). The symmetry coefficients c_a correctly captures the expected symmetry of the states[56], at least for small cut-off radii. The effect of employing a symmetry-broken supercell can be seen on the totally symmetric state a_1, which starts to be mixed with the antisymmetric a_2 irrep as a function of the radius, while there is no effect on the degenerate e_x state. Therefore a radius of 2 Å is used in the database to catch the expected local symmetry of the defect.

Transition dipole moment

The transition dipole moment is calculated between all single particle Kohn-Sham states inside the band gap,

$$\mu_{nm} = \langle \psi_n | \hat{\mathbf{r}} | \psi_m \rangle = \frac{i\hbar}{m_e} \frac{\langle \psi_n | \hat{\mathbf{p}} | \psi_m \rangle}{\varepsilon_n - \varepsilon_m}, \tag{13}$$

where $\hat{\mathbf{r}}$ is the dipole operator, $\hat{\mathbf{p}}$ is the momentum operator, and m_e is the electron mass. The transition dipole moment yields information on the possible polarization directions and oscillator strength of a given transition. In this work, the transition dipoles between localised defect states are calculated in real space, i.e. using the first expression in Equation (13), after translating the defect to the center of the supercell.

Radiative recombination rate and life time

Radiative recombination refers to the spontaneous decay of an electron from an initial high energy state to a state of lower energy upon emitting a photon. The rate of a spin-preserving radiative transition between an initial state ψ_m and a final state ψ_n is given by[69]

$$\Gamma_{nm}^{\text{rad}} = 1/\tau_{\text{rad}} = \frac{E_{\text{ZPL}}^3 \mu_{nm}^2}{3\pi\epsilon_0 c^3 \hbar^4}. \tag{14}$$

Here, ϵ_0 is the vacuum permittivity, E_{ZPL} is the zero phonon line (ZPL) energy of the transition (see Section "Excited states"), μ_{nm} is

the transition dipole moment defined in Equation (13) and c is the speed of light. The ZPL energy includes the reorganisation energy due to structural differences between the initial and final states, which can be on the order of 1 eV. Consequently, an accurate estimate of the radiative lifetime requires a geometry optimisation in the excited state. Since this step is not part of our general workflow, but is only performed for selected defects, radiative lifetimes are currently only available for a few transitions.

Hyperfine coupling

Hyperfine (HF) coupling refers to the interaction between the magnetic dipole associated with a nuclear spin, $\hat{\mathbf{I}}^N$, and the magnetic dipole of the electron-spin distribution, $\hat{\mathbf{S}}$. For a fixed atomic nuclei, N, the interaction is written

$$\hat{H}_{\text{HF}}^N = \sum_{i,j} \hat{S}_i A_{ij}^N \hat{I}_j^N \tag{15}$$

where the hyperfine tensor $\mathbf{A}^N$ is given by

$$A_{ij}^N = \frac{2\alpha^2 g_e m_e}{3M_N} \int \delta_T(\mathbf{r})\rho_s(\mathbf{r})d\mathbf{r} + \frac{\alpha^2 g_e m_e}{4\pi M_N} \int \frac{3r_i r_j - \delta_{ij}r^2}{r^5}\rho_s(\mathbf{r})d\mathbf{r}. \tag{16}$$

The first term represents the isotropic Fermi-contact term, which results from a non-vanishing electron spin density $\rho_s(\mathbf{r})$ at the centre of the nucleus. $\delta_T(\mathbf{r})$ is a smeared out δ-function which appears in place of an ordinary δ-function in the non-relativistic formulation of the Fermi-contact term and regulates the divergence of the s-electron wave function at the atomic core[70]. α is the fine structure constant, m_e is the electron mass, M_N is the mass of atom N, and g_e is the gyromagnetic ratio for electrons. The second term represents the anisotropic part of the hyperfine coupling tensor and results from dipole-dipole interactions between nuclear and electronic magnetic moments.

The Fermi-contact term $a = \text{Tr}(\mathbf{A}^N)/3$ depends only on the $l = 0$ component of the spin density at the nucleus. In contrast, the anisotropic term $A_{ij}^N - a$ is sensitive to the $l > 0$ components of the spin density near the nucleus. The hyperfine tensor thus provides direct insight into the electron spin distribution near the corresponding nucleus, and a direct comparison of the calculated HF coupling constants with electron paramagnetic resonance spectroscopy measurements can help to identify the nature of defect centers[26,31]. As an illustration, Fig. 6 shows the iso-surface of the spin density of the V_B^- defect in hexagonal boron nitride.
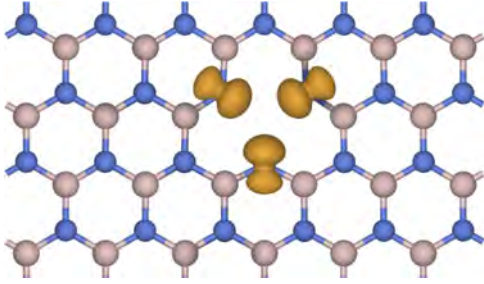
Fig. 6 Spin density for the boron vacancy in hBN. Iso-surface of the calculated spin-density (isovalue of $0.002\ |e|A^{-3}$) for the V_B^- defect in hexagonal boron nitride.

Fig. 7 Possible defect state occupancies for the excited state calculations. The states involved in the excitation are encircled red.

In the present work, the $\mathbf{A}^N$ tensors are calculated for all atoms of the supercell, and its eigenvalues, also known as the HF principal values, are reported in the QPOD database.

Zero field splitting

Zero field splitting (ZFS) refers to the splitting of magnetic sub levels of a triplet defect state due to the magnetic dipole-dipole interaction of the two electron spins that takes place even in the absence of an external magnetic field. A triplet ($S = 1$) defect state can be described by a spin Hamiltonian of the following form

$$\hat{H}_{ZFS} = \sum_{ij} \hat{S}_i D_{ij} \hat{S}_j, \tag{17}$$

where $\hat{\mathbf{S}}$ is the total spin operator and $\mathbf{D}$ is the ZFS tensor given by

$$D_{ij} = \frac{\alpha^2 g_e m_e}{4\pi} \int |\phi_{ij}(\mathbf{r}_1, \mathbf{r}_2)|^2 \frac{3r_i r_j - \delta_{ij} r^2}{r^5} d\mathbf{r}_1 \mathbf{r}_2, \tag{18}$$

where $\phi_{ij}(\mathbf{r}_1, \mathbf{r}_2)$ are the slater determinants of the Kohn-Sham states i and j. After diagonalization, Equation (17) can also be written as

$$\begin{aligned} \hat{H}_{ZFS} &= D_{xx}\hat{S}_x^2 + D_{yy}\hat{S}_y^2 + D_{zz}\hat{S}_z^2 \\ &= D(\hat{S}_z^2 - S(S+1)/3) + E(\hat{S}_x^2 - \hat{S}_y^2), \end{aligned} \tag{19}$$

where $D = 3D_{zz}/2$ and $E = (D_{xx} - D_{yy})/2$ are called axial and rhombic ZFS parameters, respectively. D describes the splitting between the $m_s = \pm 1$ and $m_s = 0$ magnetic sub-levels, while E describes the splitting of the $m_s = \pm 1$ sub-levels. D is generally zero for a spherically symmetric wave function because there is no direction in which electrons of a triplet can move to minimize the repulsive dipole-dipole interaction. However, for non-spherical wave functions, D will be non-zero and lift the degeneracy of magnetic sub-levels $m_s = \pm 1$ and $m_s = 0$. A positive value of D will result from an oblate spin-distribution, while a negative value will result from a prolate spin-distribution. The value of E will be zero for axially symmetric wave functions.

Excited states

In Kohn-Sham DFT, excited electronic states can be found by solving the Kohn-Sham equations with non-Aufbau occupations of the orbitals. Often this approach is referred to as the delta self-consistent field (Delta-SCF) method[71]. The excited state solutions are saddle points of the Kohn-Sham energy functional. Unfortunately, the Delta-SCF approach often struggle to find such solutions, especially when nearly degenerate states are involved. The Delta-SCF method fails in particular for cases involving charge transfer or Rydberg states[40]. This is due to a significant rearrangement of charge density between orbitals with similar energy.
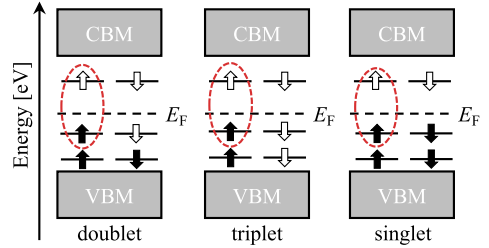
Therefore, in the present work we use an alternative to the conventional Delta-SCF method, namely a method based on the direct optimization (DO) of orbital rotations by means of an efficient quasi-Newton method for saddle points, in combination with the maximum overlap method (MOM)[40]. The MOM ensures that the character of the states is preserved during the optimization procedure. In the DO-MOM method, convergence towards the n-th order saddle point is guided by an appropriate pre-conditioner based on an approximation to the Hessian with exactly n negative eigenvalues. This method ensures fast and robust convergence of the excited states, as compared to conventional Delta-SCF[40] methods.

The DO-MOM method has been previously used for the calculation of excited state spectra of molecules[40]. However, this work is the first application of the method to defect states. We have benchmarked the method for a range of point defects in diamond, hBN, SiC, and MoS_2, and established that the method yields results in good agreement with conventional Delta-SCF calculations.

The most frequently occurring point defect spin configurations are sketched in Fig. 7. For the doublet and triplet spin configuration, both the ground and excited states can be expressed as single Slater determinants. However, for the singlet spin configuration, the ground state is a closed shell singlet, while the excited state, as a result of the single excitation in either spin channel, will result in an open shell singlet state, which cannot be expressed as a single Slater determinant. This open shell singlet state can, however, be written as a sum of two Slater determinants of the form $|a \uparrow, b \downarrow\rangle$ and $|a \downarrow, b \uparrow\rangle$, each of which represent a mixed singlet-triplet state accessible with Delta-SCF. This allows us to obtain the singlet energy as

$$E_s = 2E_{st} - E_t. \tag{20}$$

Here, E_{st} is the DFT energy calculated by setting the occupancy for the open-shell singlet state, while E_s and E_t are the energies of the corresponding singlet and triplet states. Note that the latter can be represented as a Slater determinant. Photoluminescence spectra of selected point defects were calculated using a generating function approach[72] outlined below. First, the mass weighted difference between atomic coordinates in the ground and excited electronic states is computed as follows

$$\Delta Q = \sqrt{\sum_a m_a \Delta R_a^2}, \tag{21}$$

where the sum runs over all the atoms in the supercell. Afterwards, the partial Huang-Rhys factors are computed as

$$S_k = \frac{1}{2\hbar} \omega_k Q_k^2, \tag{22}$$

where Q_k is the projection of the lattice displacement on the normal coordinates of the ground state described by phonon mode k. The electron-phonon spectral function, which depends
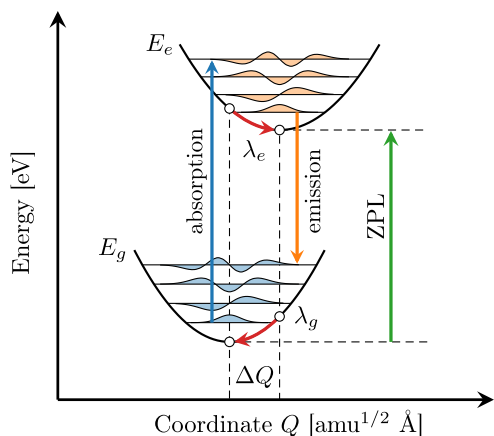
Fig. 8 Schematic CC diagram of a ground state to excited state transition. Absorption to and emission from the excited state (blue, orange) as well as their respective reorganization energies $\lambda_{e,g}$ as a function of energy and configuration coordinate Q. The zero phonon line transition (ZPL) is visualized in green.

on the coupling between lattice displacement and vibrational degrees of freedom, is then obtained by summing over all the modes k

$$S(\omega) = \sum_k S_k \delta(\omega - \omega_k). \tag{23}$$

The integral over the electron-phonon spectral function gives the (total) HR factor of the transition. The above electron-phonon spectral function is fed into a generating function to produce the photoluminescence lineshape[72]. A one dimensional configuration coordinate (CC) diagram illustrating the absorption/emission processes between two defect states is shown in Fig. 8.

The QPOD workflow

The backbone of the QPOD database is represented by a high-throughput framework based on the Atomic Simulation Recipes (ASR)[29] in connection with the MyQueue[30] task and workflow scheduling system. Numerous recipes, designed particularly for the evaluation of defect properties, have been implemented in ASR and have been combined in a central MyQueue workflow to generate all data for the QPOD database.

The underlying workflow is sketched in Fig. 9, and will be described briefly in the following. As a preliminary step the C2DB[32,33] is screened to obtain the set of host materials. Only non-magnetic, thermodynamically ($E_{hull} < 0.2$ eV per atom and $\Delta H_{host} < 0.2$ eV per atom) and dynamically stable materials with a PBE band gap of $E_{gap}^{PBE} > 1$ eV are selected as host materials. These criteria result in 281 materials of which we select 82 from a criterion of $N_e \times V_{supercell} < 0.9$ Å³ combined with a few handpicked experimentally known and relevant 2D materials (MoS_2, hBN, WS_2, $MoSe_2$). It is important to mention, that some host materials exist in different phases (same chemical formula and stoichiometry, but different symmetry), e.g. $CaBr_2$-164 and $CaBr_2$-187. In these cases we only keep the most stable one, i.e. the one with the lowest heat of formation ($CaBr_2$-164 for the previous example).

For each host material, all inequivalent vacancies and antisite defects are created in a supercell as described in Section "Supercell and defect structures". Each defect enters the ground state workflow, which includes relaxation of the neutral defect structure, calculation of a well-converged ground state density, identification and extraction of defect states within the pristine band gap, and SJ calculations with half-integer charges q. If there

are no states within the gap, the defect system directly undergoes the data extraction workflow and is stored in the QPOD database.

For systems with in-gap states above (below) E_F, an electron is added (removed) and the charged structures are relaxed, their ground state calculated and the states within the band gap are examined again up to a maximum charge of $+3/-3$. Once all charge states have been relaxed and their ground state density has been evaluated, the data extraction workflow is executed. Here, general defect information (defect name, defect charge, nearest defect-defect distance, etc.), charge transition levels and formation energies, the equilibrium self-consistent Fermi level, equilibrium defect concentrations, symmetries of the defect states within the gap, hyperfine coupling, transition dipole moments, etc. are calculated and the results are stored in the database. The data is publicly available and easy to browse in a web-application as will be described in Section "The QPOD webapp".

We note, that selected systems have been subject to excited state calculations in order to obtain ZPL energies, PL spectra, HR factors, etc. enabling the identification of promising defect candidates for optical applications as is discussed in Section "Defects for quantum technological applications".

The QPOD database

The QPOD database uses the ASE DB format[73] which currently has five backends: JSON, SQLite3, PostgreSQL, MySQL, MariaDB. An ASE DB enables simple querying of the data via the in-built ase db command line tool, a Python interface, or a webapp (see Section "The QPOD webapp"). With those different possibilities to access and interact with the data, we aim to give users a large flexibility based on their respective technical background and preferences.

Each row of QPOD is defined by defect name, host name, and its respective charge state. Furthermore, the fully relaxed structure as well as all of the data associated with the respective defect is attached in the form of key-value pairs or JSON-formatted raw data.

The QPOD webapp

A defining feature of the QPOD database is its easy accessibility through a web application (webapp). For each row of the database, one can browse a collection of web-panels designed to highlight the various computed properties of the specific defect. Specific elements of the web-panels feature clickable '?' icons with explanatory descriptions of the content to improve the accessibility of the data.

Directing between different entries of the database is either possible by using hyperlinks between related entries, or using the overview page of the database, where the user can search for materials, and order them based on different criteria. Furthermore, we ensure the direct connection to C2DB with hyperlinks between a defect material and its respective host material counterpart in C2DB in case users want to find more information about the defect-free systems.

Overview of host materials

In total, 82 host crystals that were chosen according to the criteria outlined in Section "The QPOD workflow" comprise the basis of our systematic study of intrinsic point defects. The set of host materials span a range of crystal symmetries, stoichiometries, chemical elements, and and band gaps (see Fig. 10). Among them, at least nine have already been experimentally realized in their monolayer form, namely As_2, BN, Bi_2I_6, C_2H_2, MoS_2, $MoSe_2$, Pd_2Se_4, WS_2, and ZrS_2 whereas 15 possess an ICSD[74,75] entry and are known as layered bulk crystals. The PBE band gaps range from 1.02 eV for Ni_2Se_4 up to 5.94 eV for $MgCl_2$ making our set of starting host crystals particularly heterogeneous.

Next, we first present some general illustrations and analyses of the data in QPOD. We then leverage the data to address three
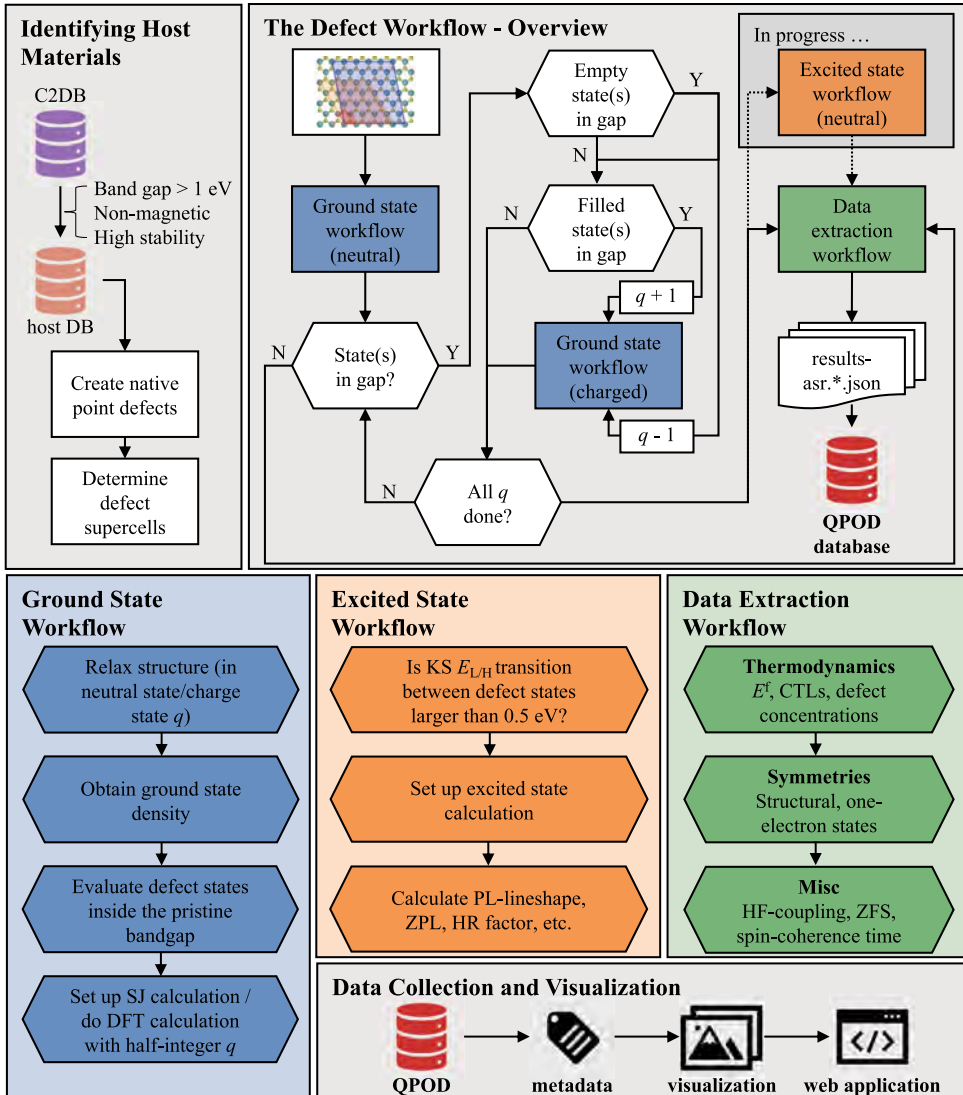
Fig. 9 The workflow behind the defect database. First, starting from C2DB suitable host materials are identified. With the ASR recipe for defect generation the defect supercells are set up and enter the ground state workflow for neutral systems. Afterwards, depending on the nature of the defect states inside the gap, charged calculations are conducted within the charged ground state workflow. If all charged calculations for a specific system are done ($|q| < 3$), physical results are collected using the data extraction workflow and thereafter saved in the defect database. The database is equipped with all necessary metadata, and together with various visualization scripts the browseable web application is created. The excited state workflow is executed for selected systems as discussed in Section Defects for quantum technological applications.

specific scientific problems, namely the identification of: (i) Defect tolerant semiconductors with low concentrations of mid-gap states. (ii) Intrinsically *p*-type or *n*-type semiconductors. (iii) Optically accessible high-spin defects for quantum technological applications.

Relaxation of defect structures

A major part of the computational efforts to create the QPOD went to the relaxation of the defect structures in a symmetry broken supercell. As discussed in Section "Supercell and defect structures" the strategy to actively break the symmetry of the host crystal by the choice of supercell was adopted to enable defects to relax into their lowest energy configuration.

Figure 11 shows the gain in total energy due to the relaxation for the over 1900 vacancy and antisite defects (different charge states included). Not unexpectedly, the relaxation has the largest influence on antisite defects while vacancy structures in general show very weak reorganization relative to the pristine structure as can be seen in the left panel of Fig. 11. The relaxations for charged defects have always been started from the neutral equilibrium
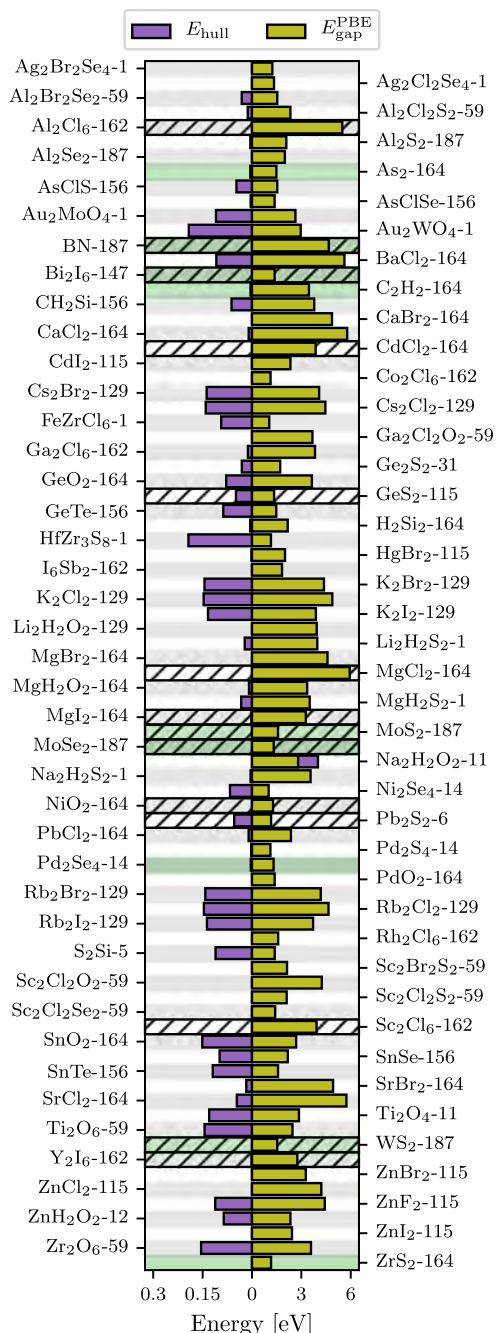
Fig. 10 Overview of host crystals. Energy above convex hull in eV/atom E_{hull} (purple bars, left) and PBE-calculated band gap E_{gap}^{PBE} (yellow bars, right) of the 82 host crystals of QPOD. Host materials which have been realized experimentally (in their monolayer form) are highlighted in green and materials with a known layered bulk phase and related ICSD[74,75] entry are shown with a hatch pattern. The individual names of host crystals are written as: chemical formula-space group.

configuration of the respective defect. As a result, the reorganization energies for charged defects is significantly lower (see right panel of Fig. 11).

Charge transition levels

In Section "Slater-Janak transition state theory" we described how we obtain the CTLs by combining Slater-Janak transition state theory on a static lattice with geometry relaxations in the final state. For 'negative' transitions, e.g. $\varepsilon(0/-1)$, the effect of the relaxation is to lower the energy cost of adding the electron, i.e. the reorganisation energy lowers the CTL. In contrast, the lattice relaxations will produce an upward shift for 'positive' transitions, e.g. $\varepsilon(0/+1)$, because in this case the CTL denotes the negative of the energy cost of removing the electron.

Figure 12 shows the $\varepsilon(0/+1)$ and $\varepsilon(0/-1)$ CTLs for a small subset of defects. Since the energies are plotted relative to the vacuum level, the CTLs correspond to the (negative) ionisation potential (-IP) and electron affinity (EA), respectively. Results are shown both with and without the inclusion of relaxation effects. As expected, the relaxation always lowers the EA and the IP (the $\varepsilon(0/+1)$ is always raised). The reorganisation energies can vary from essentially zero to more than 2 eV, and are absolutely crucial for a correct prediction of CTLs and (charged) defect formation energies.

We notice that the CTLs always fall inside the band gap of the pristine host (marked by the grey bars) or very close to the band edges. This is clearly expected on physical grounds, as even for a defective system the CTLs cannot exceed the band edges (there are always electrons/holes available at the VBM/CBM sufficiently far away from the point defect). However, for small supercells such behavior is not guaranteed as the band gap of the defective crystal could deviate from that of the pristine host material. Thus, the fact that the CTLs rarely appear outside the band gap is an indication that the employed supercells are generally large enough to represent an isolated defect.

When the Fermi level is moved from the VBM to the CBM one expects to fill available defects states with electrons in a stepwise manner, i.e. such that $\varepsilon(q+1/q) < \varepsilon(q/q-1)$. In particular, we expect $-IP < EA$. Interestingly, for a few systems, e.g. V_F in ZF_2, the ordering of IP and EA is inverted. The physical interpretation of such an ordering is that the neutral charge state becomes thermodynamically unstable with respect to positive and negative charge states[21]. This results in a direct transition from positive to negative charge state in the formation energy diagram.

Intrinsic carrier concentrations

For many of the potential applications of 2D semiconductors, e.g. transistors[76], light emitting devices[77], or photo detectors[78], the question of dopability of the semiconductor material is crucial. Modulation of the charge carrier concentration is a highly effective means of controlling the electrical and optical properties of a semiconductor. This holds in particular for 2D semiconductors whose carrier concentrations can be modulated in a variety of ways including electrostatic or ionic gating[79,80], ion intercalation[81], or surface functionalisation[82]. In general, these methods are only effective if the material is not too heavily doped by its ubiquitous native defects, which may pin the Fermi level close to one of the band edges. For applications relying on high carrier conduction rather than carrier control, a high intrinsic carrier concentration may be advantageous - at least if the native defects do not degrade the carrier mobility too much, see Section "Defect tolerant materials".

Figure 13 shows the calculated position of the equilibrium Fermi level (at room temperature) for all the 2D materials considered as defect hosts in this work. It should be noted that the Fermi level position depends on the number of different defect types included in the analysis. Consequently, the results are
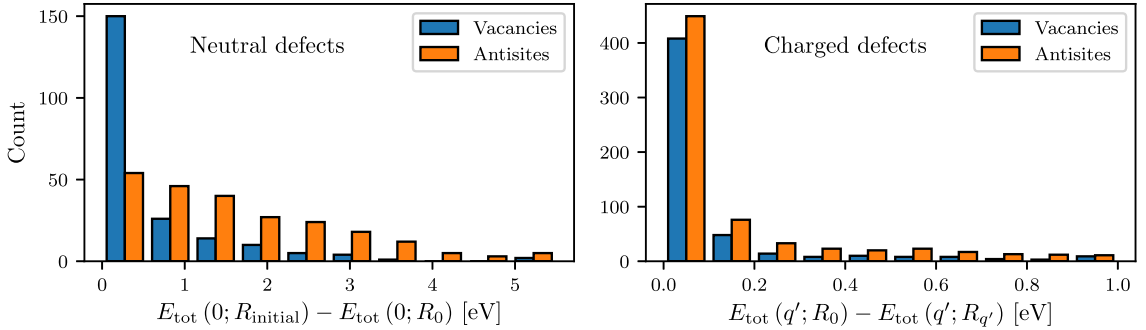
Fig. 11 Relaxation effects on the formation of charged and neutral defects. Left: histogram of the reorganization energy from the initial defect substitution to the neutral equilibrium configuration. Low values on the x-axis correspond to small reorganization of defect structures upon addition of a defect to the pristine host crystal. Right: histogram of the reorganization energy between a neutral defect and its charged counterpart.
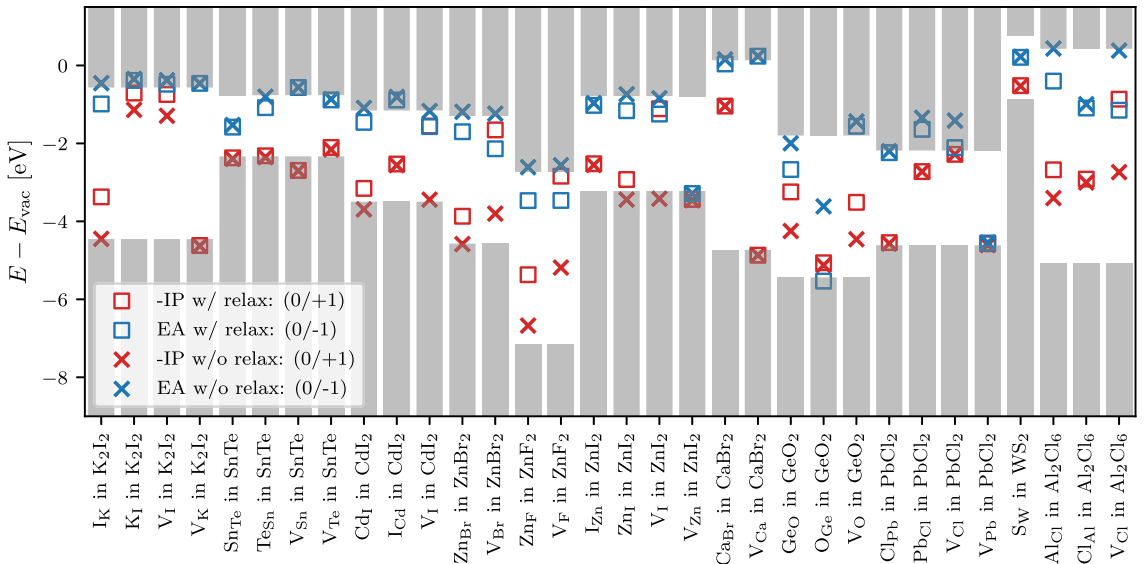

Fig. 12 Relaxation effects for ionisation potentials and electron affinities. Energies of -IP (red symbols) and EA (blue symbols) with and without relaxation effects included (boxes and crosses, respectively). The energies are all referenced to the vacuum level of the pristine crystal and grey bars represent the valence/conduction band of the individual host crystals.

sensitive to the existence of other types of intrinsic defects with formation energies lower than or comparable to the vacancy and antisite defects considered here. Fermi level regions close to the VBM or CBM (indicated by red/blue colors) correspond to p − type and n − type behavior, respectively, while Fermi levels in the central region of the band gap correspond to intrinsic behavior.

Clearly, most of the materials present in the QPOD database show either intrinsic or n-type behavior. An example of a well known material with intrinsic behavior is MoS_2, where native defects pin the Fermi level deep within the band gap resulting in very low electron and hole carrier concentrations in good agreement with previous observations[56,83–85]. As an example of a natural n-type semiconductor we highlight the Janus monolayer AsClSe, which presents an impressive electron carrier concentration of 9.5×10^{11} cm^{-2} at 300 K, making it an interesting candidate for a high-conductivity 2D material. For the majority of the materials, the equilibrium carrier concentrations are in fact

relatively low implying a high degree of dopability. We note that none of the materials exhibit intrinsic p-type behavior. This observation indicates that the challenge of finding naturally p-doped semiconductors/insulators, which is well known for bulk materials[86–89], carries over to the class of 2D materials.

Defect formation energies: Trends and correlations

The formation energy is the most basic property of a point defect. Figure 14 shows the distribution of the calculated formation energies of neutral defects for both vacancy (blue) and antisite (orange) defects. For the chemical potential appearing in Equation (3) we used the standard state of each element. There is essentially no difference between the two distributions, and both means (vertical lines) are very close to 4 eV. Roughly half (44%) of the neutral point defects have a formation energy below 3 eV and 28% are below 2 eV. This implies that many of the defects would
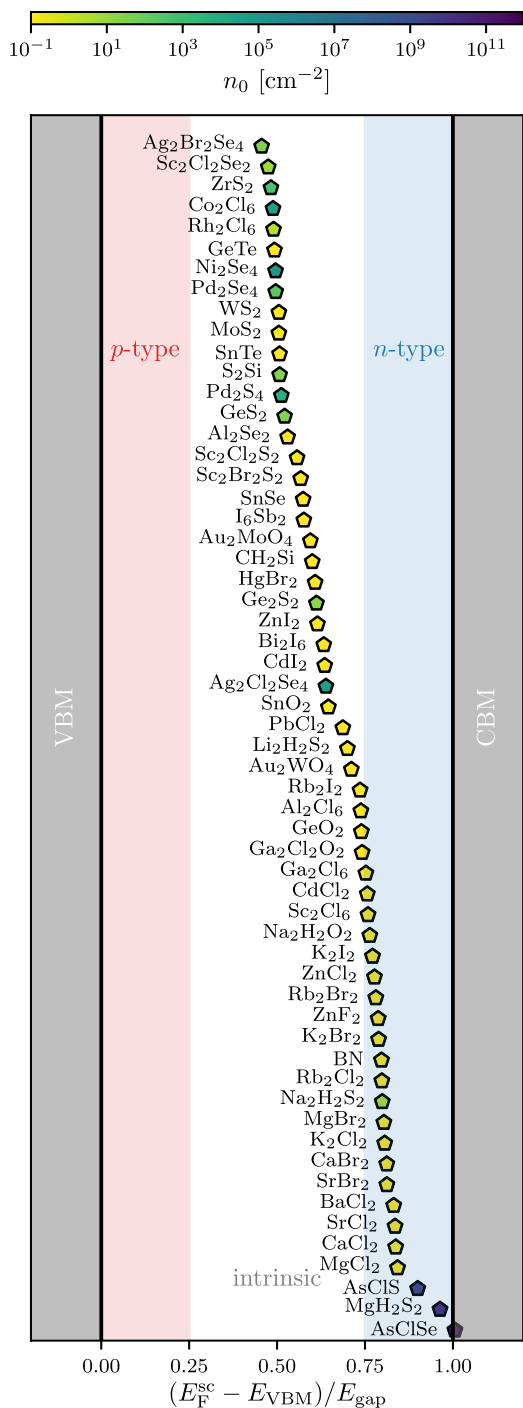
Fig. 13 Self-consistent Fermi-level position for host materials. The energy scale is normalized with respect to the pristine band gap. The regions close to the VBM, CBM correspond to the p-, n-type dopable regimes, respectively. The colorcode of the markers represents the equilibrium electron carrier concentration n_0.
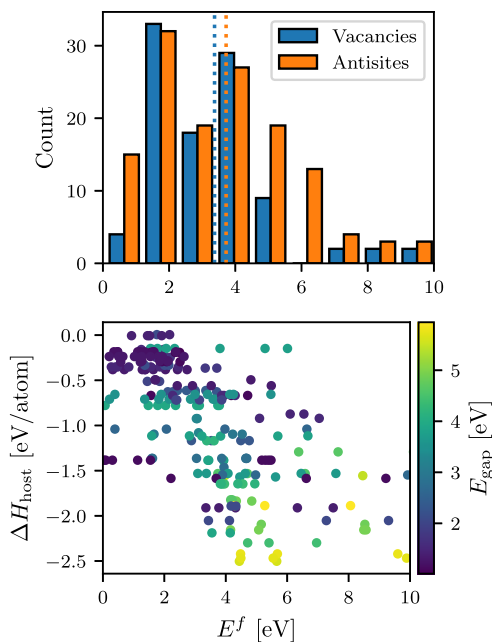

Fig. 14 Distribution of defect formation energies in the neutral charge state. Top: histogram of neutral formation energies E^f wrt. standard states for vacancy and antisite defects. The vertical dotted lines represent the respective mean values. Bottom: heat of formation ΔH_{host} of the pristine monolayer as a function of the neutral formation energies with the pristine PBE-calculated band gap as a color code.

form readily during growth and underlines the importance of including intrinsic defects in the characterization of 2D materials. As a reference, the NV center in diamond shows formation energies on the order of 5 eV to 6 eV[90] (HSE calculations).

The lower panel of Fig. 14 shows the defect formation energy relative to the heat of formation of the pristine host material, ΔH_{host}. There is a clear correlation between the two quantities, which may not come as a surprise since the ΔH_{host} measures the gain in energy upon forming the material from atoms in the standard states. More stable materials, i.e. materials with more negative heat of formation, are thus less prone to defect formation than less stable materials. It is also interesting to note the correlation with the band gap of the host material, indicated by the color of the symbols. A large band gap is seen to correlate with a large (negative) ΔH_{host} and large defect formation energies, and vice versa a small band gap is indicative of a smaller (negative) ΔH_{host} and low defect formation energies. These trends are somewhat problematic as low band gap materials with long carrier lifetimes, and thus low defect concentrations, are required for many applications in (opto-)electronics, while large band gap host materials with high density of (specific types of) defects are required for many color center-based quantum technology applications.

Point defect symmetries

Another basic property of a point defects is its local symmetry, which determines the possible degeneracies of its in-gap electronic states and defines the selection rules for optical transitions between them. Figure 15 shows the distribution of point groups (in Schönflies notation) for all the investigated neutral defect systems. A large fraction of the defects break all the
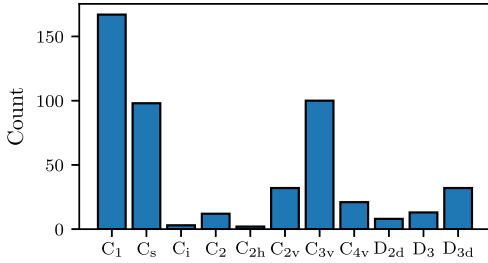
Fig. 15 Distribution of point groups for relaxed defects. Point groups are ordered from the lowest symmetry group (i.e. C_1) to the highest occurring symmetry group (D_{3d}).

symmetries of the host crystal (34% in C_1) or leave the system with only a mirror symmetry (20% in C_s). A non-negligible number (31%) of defects can be characterized by 2, 3, or 4-fold rotation axis with vertical mirror planes (C_{2v}, C_{3v} and C_{4v}). Naturally, C_{3v} defects (20% overall) often stem from hexagonal host structures, some examples being V_S and W_S in WS_2, V_{Se} in SnSe, and Ca_{Br} in $CaBr_2$. Such defects share the symmetry group of the well known NV center in diamond[91] and might be particularly interesting candidates for quantum technology applications. Relatively few defect systems (3%) possess D_3 symmetry, e.g. I_{Sb} in I_6Sb_2.

Adding or removing charge to a particular defect system can influence the structural symmetry as it was previously observed for the negatively charged sulfur vacancy in MoS_2[92]. We find that 10% of the defects in QPOD undergo a change in point group when adding (removing) an electron to (from) the neutral structure and relaxing it in its respective charge state.

Defect tolerant materials

Although defects can have useful functions and applications, they are often unwanted as they tend to deteriorate the ideal properties of the perfect crystal. Consequently, finding defect tolerant semiconductors[64,93], i.e. semiconductors whose electronic and optical properties are only little influenced by the presence of their native defects, is of great interest.

When discussing defect tolerance of semiconductors one should distinguish between two different situations: (i) For transport applications where the system is close to equilibrium, defects act as scattering centres limiting the carrier mobility. In this case, charged defects represent the main problem due to their long range Coulomb potential, which leads to large scattering cross sections. (ii) For opto-electronic applications relying on photo-excited electron-hole pairs, deep defect levels in the band gap represent the main issue as they facilitate carrier capture and promote non-radiative recombination. In the following we examine our set of host materials with respect to type (ii) defect tolerance.

Figure 16 shows the positions of charge transition levels of vacancy (blue) and antisite (orange) defects as a function of the Fermi level normalized to a host material's band gap. The neutral formation energy of the defects is shown in the middle panel, where we have also marked regions of shallow defect states that lie within 10% of the band edges (black dashed lines). A host material is said to be type (ii) defect tolerant if all of its intrinsic defects are shallow or all its deep defects have high formation energy. A number of defect tolerant host materials are revealed by this analysis, including the ionic halides K_2Cl_2, Rb_2I_2, and Rb_2Cl_2. With formation energies lying about 150 meV/atom above those of their cubic bulk structures, these materials may be challenging to realise in atomically thin form. Nevertheless, their defect tolerant nature fits well with the picture of deep defect states having larger tendency to form in covalently bonded insulators
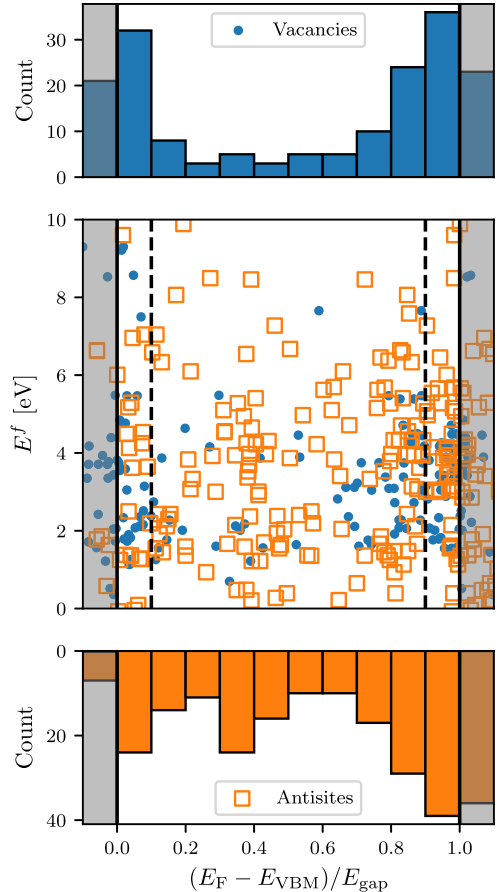
Fig. 16 Defect tolerances for vacancy and antisite defects. Position of defect charge transition levels as a function of Fermi energy (normalized with respect to the band gap) for antisite defects (orange squares) and vacancies (blue dots). Top (bottom): histogram of the occurance of a CTL for vacancies (antisites) within a certain energy range. If a defect's CTL lies above 0.9/below 0.1 (dashed black vertical lines in the middle plot) it corresponds to a CTL not having a detrimental effect on the host's properties. Materials with these CTL are considered defect-tolerant wrt. optical properties. The grey areas on the left and right hand side represent the valence band and conduction band, respectively.

with bonding/anti-bonding band gap types compared to ionic insulators with charge-transfer type band gaps[64].

The data in Fig. 16 suggests that vacancy and antisite defects have different tendencies to form shallow and deep defect states, respectively: While 55% of the vacancy defects form shallow defect states, this only happens for 30% of the antisites. The trend can be seen in the top and bottom histograms of Fig. 16 where the vacancy distribution (top) shows fewer CTLs around the center of the band gap compared to the antisite distribution (bottom). Based on this we conclude that vacancy defects are, on average, less detrimental to the optical properties of semiconductors than antisite defects.

Defects for quantum technological applications

Point defects with a high-spin ground state, e.g. triplets, are widely sought-after as such systems can be spin initialized, manipulated

Fig. 17   An overview of distribution of total magnetic moments for all defect systems. The bin corresponding to systems with triplet ground state is highlighted. A finite Fermi-smearing in our calculations gives rise to spin-contamination and hence the absolute number of entries in each bin is somewhat inaccurate. The inset shows the distribution of Kohn-Sham HOMO-LUMO gaps in both spin-channels for the triplets. The systems with a Kohn–Sham HOMO-LUMO gap of at least 0.5 eV are chosen for excited states calculations.

and read out (the N_V^- centre in diamond is a classical example). In particular, triplet spin systems may be exploited in optically detected magnetic resonance spectroscopy to act as qubits, quantum magnetometers, and other quantum technological applications[4,5].
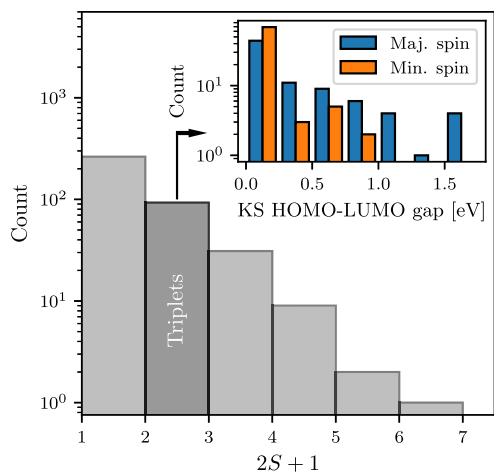
An overview of the distribution of local magnetic moments for all the defect systems is shown in Fig. 17. Although defects with any spin multiplicity are of interest and have potential applications, e.g. as sources of intense and bright single photon emission, we have limited our detailed excited state investigations to the 80 defects with a triplet ground state. These systems are subsequently screened for a (PBE) Kohn–Sham gap of at least 0.5 eV. After this screening, we are left with 33 systems. Out of these 33 defect systems, we further discard those involving heavy atoms, e.g. $PbCl_2$, because of large spin-orbit coupling effects that would hinder their usage in quantum technological applications, leaving us with a total of 25 systems, which are subject to excited state calculations (see Section Excited states). The excited state calculations are performed with the motivation of finding systems exhibiting sharp radiative transitions in the mid infrared to ultraviolet frequency range.

Most of these 25 defects exhibit large structural changes upon optical excitation. After selecting the defects with the smallest ΔQ values, we are left with four systems for which we calculate the PL line shape. A detailed study of these four systems, e.g. their optical cycles, inter-system crossing rates, spin-coherence times, will be the subject of a future study. In the present work we limit ourselves to the calculation of their PL line shape, which is a key characteristic of point defects, which may be used to identify their precise atomic structure[94].

The PL lineshapes calculated using the generating function approach are shown Fig. 18 for the lowest triplet transitions of V_{Si}, C_H^- and Si_H^- in $SiCH_2$ and the lowest singlet transition of Si_H^- in $SiCH_2$. The calculated PBE band gap of pristine $SiCH_2$ is 3.79 eV. The upper and bottom panels show the line shape for the HOMO-
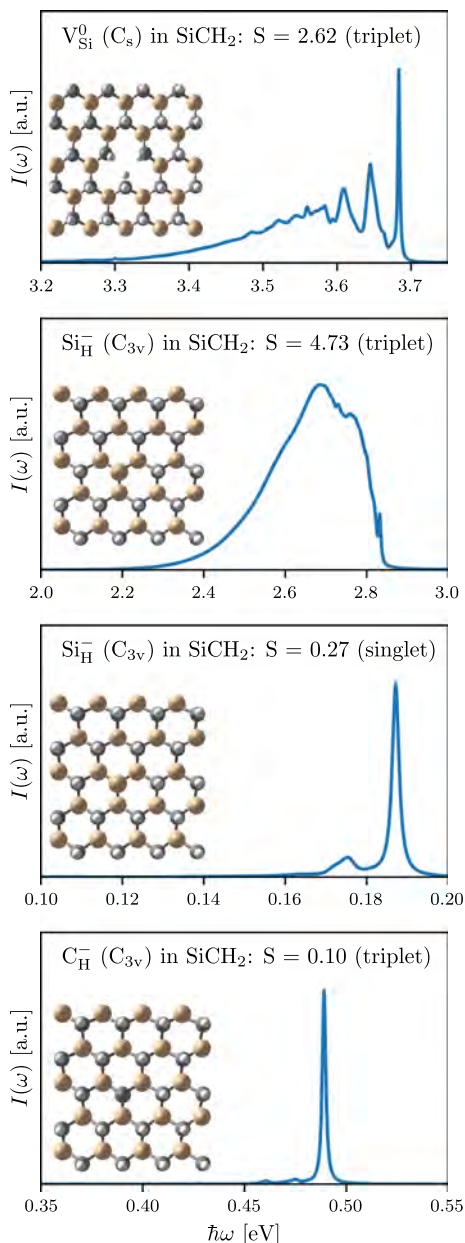
Fig. 18   PL lineshapes for a few selected systems using generating function approach. A few selected systems with small Huang-Rhys factors are shown. The top panel show the PL line shape for HOMO to LUMO transition in the majority spin channel for V_{Si} defect in $SiCH_2$. The middle two panels shows the PL spectra for HOMO to LUMO transition in the minority spin channel (2nd panel) and HOMO to LUMO transition within the singlet manifold (3rd panel) for Si_H^- defect in $SiCH_2$ host material. The bottom panel shows the calculated spectra for C_H^- defect in $SiCH_2$ for HOMO to LUMO transition in the majority spin channel. The calculated radiative life times of these transitions (top to bottom) are 1.3×10^{-3} ns, 4.9×10^{-2} ns, 2.9×10^{-4} ns and 9.0×10^{-4} ns, repectively.

LUMO transition in the majority spin channel for V_{Si} and C_H^-, respectively. While the Huang-Rhys factors for V_{Si} is small, it is extremely small for C_H^-. In fact, radiative transitions with such small line widths have rarely been reported previously in 2D semiconductors[8], which makes these results promising. The second panel shows the PL spectra for the lowest transition in the minority spin channel of Si_H^-, while the third panel shows the PL spectrum for the lowest transition within the singlet manifold of the same defect. The symmetry of this defect, i.e. C_{3v}, and the KS level structure is somewhat similar to the N_V^- centre in diamond. The fact that the calculated Huang-Rhys factor is also close to the one reported for N_V^-[72] makes this defect system particularly interesting as a 2D counter part of the highly useful N_V^- centre.

Discussion

In summary, more than 1900 intrinsic point defects hosted by 82 different 2D semiconductors and insulators have been relaxed and characterised by means of DFT calculations. The 1900+ point defects comprise about 500 structurally distinct vacancy and antisite defects in different charge states. The thousands of DFT calculations were orchestrated by a computational workflow built using the Atomic Simulation Recipes (ASR)[29] and executed with the MyQueue[30] task scheduler.

The ASR defect workflow includes the calculation of formation energies, charge transition levels, equilibrium defect and carrier concentrations, point group symmetry labels of in-gap states, transition dipole moments, hyperfine coupling, and zero field splittings. All DFT calculations were performed using the semi-local PBE xc-functional. While the PBE should yield accurate structural and thermodynamic properties, its tendency to underestimate band gaps can result in high-energy defect levels being missed, if they lie above the PBE band gap. We emphasise, however, that charge transition levels appearing within the PBE band gap should be well described by our PBE-based Slater-Janak transition state approach.

The analysis of defect formation energies showed that many of the investigated defects are likely to be present in the host material in appreciable concentrations thus confirming their relevance and importance for the general properties of the host materials. Based on the thermodynamic and electronic characterisation of single intrinsic defects, we identified a number of defect tolerant ionic insulators among the host materials. While these specific materials may be challenging to synthesize due to competing bulk phases, the results can be used to guide future searches for defect tolerant semiconductors for high-performance opto-electronic 2D devices. Among the 82 host materials, we found several semiconductors with high intrinsic electron carrier concentrations, e.g. MgS_2H_2 and the Janus materials AgClS and AgClSe, whereas no intrinsic p-type materials were found.

Out of the >1900 defects, only around 80 adopt a high-spin ($S > 1/2$) ground state and only a few defects showed low Huang–Rhys factors and correspondingly narrow photoluminescence spectra making them relevant as single-photon emitters. This indicates that the simple defect types considered in this work are not likely to yield useful spin defects for quantum technology applications. In this work, the excited state properties were calculated and analyzed manually. Incorporating this part into the automated workflow will be an important future extension of the current methodology that is critical to enable a systematic and rational design of defects with ideal excited state properties including transition energies, excited state lifetimes/dynamics, and emission line shapes.

The web-based presentation of the QPOD database and its seamless integration with the Computational 2D Materials Database (C2DB) makes a powerful platform for exploring the physics of point defects in 2D materials, which should be useful both as a convenient lookup table and as a benchmark reference for computational studies. Moreover, the possibility to download the entire database makes it applicable for machine learning purposes, which has a large yet untapped potential for establishing structure-property relationships for point defects.

Looking ahead, there are many possible extensions and improvements of the current work. First of all, it would be interesting to move beyond the PBE to more advanced xc-functionals, such as screened hybrids. This would not only enhance the data quality/accuracy but also provide the basis for a systematic and statistically significant assessment of the performance of the PBE for defect calculations. It would also be relevant to expand the set of host materials beyond the 82 materials considered here. The C2DB currently contains about 500 monolayers exfoliated from experimentally known layered van der Waals crystals and a similar number of predicted highly stable monolayers providing ample opportunities for selecting ideal 2D host crystals. As mentioned, the simple defects considered in this work turned out to be mostly non-magnetic. Thus, for applications relying on spinful defect centers, e.g. magnetic field sensors or qubits, it seems important to incorporate more complex defects such as divacancies and vacancy-substitutional defect pairs.

METHODS

Density functional theory calculations

All the DFT calculations (spin-polarized) are performed by the GPAW electronic structure code[95] using a plane wave basis set with 800 eV plane wave cut off, k-point density of 6 Å^{-1} (12 Å^{-1}) for structural relaxations (for ground state calculations) and the PBE xc-functional[38]. The supercell is kept fixed and atoms are fully relaxed until forces are below 0.01 eVÅ^{-1}. We apply a Fermi smearing of 0.02 eV (0.2 eV for relaxations) for all systems and increase that parameter slightly (0.05, 0.08 or 0.1) for systems whose ground state proves difficult to converge. We use the Pulay mixing scheme[96] where total density and magnetisation densities are treated separately. The excited states are calculated, at gamma point, using the same computational parameters as the ground state, and using the DO-MOM method[40], where the maximum step length, ρ_{max}, for the quasi-newton search direction is chosen to be 0.2. The parameters for the Slater-Janak calculations are the same as for the ground state calculations.

An initial benchmarking (for formation energies, CTLS, HF, ZFS, TDM, excited states, etc.) was performed on a subset of defect systems prior to the work highlighted here, in order to ensure that our methods, implementations and calculations yield reasonable results.

DATA AVAILABILITY

The QPOD database will be available upon request. The web-application of the database will be available on the computational materials repository (CMR): https://cmr.fysik.dtu.dk/.

CODE AVAILABILITY

The ASR recipe scripts used in the QPOD workflow are available at: https://gitlab.com/asr-dev/asr/-/tree/old-master/asr.

REFERENCES

1. Polman, A., Knight, M., Garnett, E. C., Ehrler, B. & Sinke, W. C. Photovoltaic materials: present efficiencies and future challenges. Science. **352**, aad4424 (2016).
2. Awschalom, D. D., Bassett, L. C., Dzurak, A. S., Hu, E. L. & Petta, J. R. Quantum spintronics: engineering and manipulating atom-like spins in semiconductors. Science **339**, 1174–1179 (2013).
3. Gomonay, O. Crystals with defects may be good for spintronics. Physics **11**, 78 (2018).
4. Eckstein, J. N. & Levy, J. Materials issues for quantum computation. MRS Bull. **38**, 783–789 (2013).

5. Gardas, B., Dziarmaga, J., Zurek, W. H. & Zwolak, M. Defects in quantum computers. *Sci. Rep.* **8**, 1–10 (2018).

6. Lovchinsky, I. et al. Nuclear magnetic resonance detection and spectroscopy of single proteins using quantum logic. *Science* **351**, 836–841 (2016).

7. Bradley, C. et al. A ten-qubit solid-state spin register with quantum memory up to one minute. *Phys. Rev. X* **9**, 031045 (2019).

8. Sajid, A., Ford, M. J. & Reimers, J. R. Single-photon emitters in hexagonal boron nitride: a review of progress. *Rep. Prog. Phys.* **83**, 044501 (2020).

9. Grosso, G. et al. Tunable and high-purity room temperature single-photon emission from atomic defects in hexagonal boron nitride. *Nat. Commun.* **8**, 1–8 (2017).

10. Fischer, M. et al. Controlled generation of luminescent centers in hexagonal boron nitride by irradiation engineering. *Sci. Adv.* **7**, eabe7138 (2021).

11. Ferrari, A. C. et al. Science and technology roadmap for graphene, related two-dimensional crystals, and hybrid systems. *Nanoscale* **7**, 4598–4810 (2015).

12. Lin, Z. et al. Defect engineering of two-dimensional transition metal dichalcogenides. *2d Mater.* **3**, 022002 (2016).

13. Tran, T. T., Bray, K., Ford, M. J., Toth, M. & Aharonovich, I. Quantum emission from hexagonal boron nitride monolayers. *Nat. Nanotechnol.* **11**, 37–41 (2016).

14. Bertolazzi, S. et al. Engineering chemically active defects in monolayer mos₂ transistors *via* ion-beam irradiation and their healing *via* vapor deposition of alkanethiols. *Adv. Mater.* **29**, 1606760 (2017).

15. Chakraborty, C., Kinnischtzke, L., Goodfellow, K. M., Beams, R. & Vamivakas, A. N. Voltage-controlled quantum light from an atomically thin semiconductor. *Nat. Nanotechnol.* **10**, 507–511 (2015).

16. Koperski, M. et al. Single photon emitters in exfoliated wse 2 structures. *Nat. Nanotechnol.* **10**, 503–506 (2015).

17. Chejanovsky, N. et al. Single-spin resonance in a van der waals embedded paramagnetic defect. *Nat. Mater.* **20**, 1079–1084 (2021).

18. Gottscholl, A. et al. Initialization and read-out of intrinsic spin defects in a van der waals crystal at room temperature. *Nat. Mater.* **19**, 540–545 (2020).

19. Ye, G. et al. Defects engineered monolayer mos2 for improved hydrogen evolution reaction. *Nano Lett.* **16**, 1097–1103 (2016).

20. Xie, J., Yang, X. & Xie, Y. Defect engineering in two-dimensional electrocatalysts for hydrogen evolution. *Nanoscale* **12**, 4283–4294 (2020).

21. Freysoldt, C. et al. First-principles calculations for point defects in solids. *Rev. Mod. Phys.* **86**, 253 (2014).

22. Janotti, A. & Van de Walle, C. G. Native point defects in zno. *Phys. Rev. B* **76**, 165202 (2007).

23. Neugebauer, J. & Van de Walle, C. G. Gallium vacancies and the yellow luminescence in gan. *Appl. Phys. Lett.* **69**, 503–505 (1996).

24. Dreyer, C. E., Alkauskas, A., Lyons, J. L., Janotti, A. & Van de Walle, C. G. First-principles calculations of point defects for quantum technologies. *Annu. Rev. Mater. Res.* **48**, 1–26 (2018).

25. Gupta, S., Yang, J.-H. & Yakobson, B. I. Two-level quantum systems in two-dimensional materials for single photon emission. *Nano Lett.* **19**, 408–414 (2018).

26. Sajid, A., Thygesen, K. S., Reimers, J. R. & Ford, M. J. Edge effects on optically detected magnetic resonance of vacancy defects in hexagonal boron nitride. *Commun. Phys.* **3**, 153 (2020).

27. Jain, A. et al. Fireworks: A dynamic workflow system designed for high-throughput applications. *Concurr. Comput.* **27**, 5037–5059 (2015).

28. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. Aiida: automated interactive infrastructure and database for computational science. *Comput. Mater. Sci.* **111**, 218–230 (2016).

29. Gjerding, M. et al. Atomic simulation recipes: a python framework and library for automated workflows. *Comput. Mater. Sci.* **199**, 110731 (2021).

30. Mortensen, J. J., Gjerding, M. & Thygesen, K. S. Myqueue: task and workflow scheduling system. *J. Open Source Softw.* **5**, 1844 (2020).

31. Sajid, A., Reimers, J. R. & Ford, M. J. Defect states in hexagonal boron nitride: Assignments of observed properties and prediction of properties relevant to quantum computation. *Phys. Rev. B* **97**, 064101 (2018).

32. Haastrup, S. et al. The computational 2d materials database: high-throughput modeling and discovery of atomically thin crystals. *2d Mater.* **5**, 042002 (2018).

33. Gjerding, M. N. et al. Recent progress of the computational 2d materials database (c2db). *2d Mater.* **8**, 044002 (2021).

34. Broberg, D. et al. Pycdt: a python toolkit for modeling point defects in semiconductors and insulators. *Comput. Phys. Commun.* **226**, 165–179 (2018).

35. Péan, E., Vidal, J., Jobic, S. & Latouche, C. Presentation of the pydef post-treatment python software to compute publishable charts for defect energy formation. *Chem. Phys. Lett.* **671**, 124–130 (2017).

36. Goyal, A., Gorai, P., Peng, H., Lany, S. & Stevanović, V. A computational framework for automation of point defect dalculations. *Comput. Mater. Sci.* **130**, 1–9 (2017).

37. Davidsson, J., Ivády, V., Armiento, R. & Abrikosov, I. A. ADAQ: automatic workflows for magneto-optical properties of point defects in semiconductors. *Comput. Phys. Commun.* **269**, 108091 (2021).

38. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).

39. Janak, J. F. Proof that $\frac{\partial e}{\partial n_i} = \varepsilon$ in density-functional theory. *Phys. Rev. B* **18**, 7165 (1978).

40. Levi, G., Ivanov, A. V. & Jonsson, H. Variational calculations of excited states *via* direct optimization of orbitals in DFT. *Faraday Discuss.* **224**, 448–466 (2020).

41. Mori-Sánchez, P., Cohen, A. J. & Yang, W. Localization and delocalization errors in density functional theory and implications for band-gap prediction. *Phys. Rev. Lett.* **100**, 146401 (2008).

42. Sham, L. J. & Schlüter, M. Density-functional theory of the energy gap. *Phys. Rev. Lett.* **51**, 1888 (1987).

43. Perdew, J. P. & Levy, M. Physical content of the exact kohn-sham orbital energies: band gaps and derivative discontinuities. *Phys. Rev. Lett.* **51**, 1884 (1983).

44. Van de Walle, C. G. & Janotti, A. Advances in electronic structure methods for defects and impurities in solids. *pss (b)* **248**, 19–27 (2011).

45. Zhang, X., Turiansky, M. E. & Van de Walle, C. G. Correctly assessing defect tolerance in halide perovskites. *J. Phys. Chem. C* **124**, 6022–6027 (2020).

46. Komsa, H.-P. & Pasquarello, A. Assessing the accuracy of hybrid functionals in the determination of defect levels: application to the as antisite in gaas. *Phys. Rev. B* **84**, 075207 (2011).

47. Chen, W. & Pasquarello, A. Accuracy of gw for calculating defect energy levels in solids. *Phys. Rev. B* **96**, 020101 (2017).

48. Refaely-Abramson, S., Qiu, D. Y., Louie, S. G. & Neaton, J. B. Defect-induced modification of low-lying excitons and valley selectivity in monolayer transition metal dichalcogenides. *Phys. Rev. Lett.* **121**, 167402 (2018).

49. Berseneva, N., Gulans, A., Krasheninnikov, A. V. & Nieminen, R. M. Electronic structure of boron nitride sheets doped with carbon from first-principles calculations. *Phys. Rev. B* **87**, 035404 (2013).

50. Rinke, P., Janotti, A., Scheffler, M. & Van de Walle, C. G. Defect formation energies without the band-gap problem: combining density-functional theory and the gw approach for the silicon self-interstitial. *Phys. Rev. Lett.* **102**, 026402 (2009).

51. Hedström, M., Schindlmayr, A., Schwarz, G. & Scheffler, M. Quasiparticle corrections to the electronic properties of anion vacancies at gaas (110) and inp (110). *Phys. Rev. Lett.* **97**, 226401 (2006).

52. Lyons, J. L. & Van de Walle, C. G. Computationally predicted energies and properties of defects in gan. *Npj Comput. Mat.* **3**, 1–10 (2017).

53. Van de Walle, C. G., Laks, D., Neumark, G. & Pantelides, S. First-principles calculations of solubilities and doping limits: Li, na, and n in znse. *Phys. Rev. B* **47**, 9425 (1993).

54. Zhang, S. & Northrup, J. E. Chemical potential dependence of defect formation energies in gaas: application to ga self-diffusion. *Phys. Rev. Lett.* **67**, 2339 (1991).

55. Van de Walle, C. G. & Neugebauer, J. First-principles calculations for defects and impurities: applications to iii-nitrides. *J. Appl. Phys.* **95**, 3851–3879 (2004).

56. Komsa, H.-P. & Krasheninnikov, A. V. Native defects in bulk and monolayer mos₂ from first principles. *Phys. Rev. B* **91**, 125304 (2015).

57. Komsa, H.-P., Berseneva, N., Krasheninnikov, A. V. & Nieminen, R. M. Charged point defects in the flatland: accurate formation energy calculations in two-dimensional materials. *Phys. Rev. X* **4**, 031044 (2014).

58. Liu, X. et al. Extrapolated defect transition level in two-dimensional materials: the case of charged native point defects in monolayer hexagonal boron nitride. *ACS Appl. Mater. Interfaces* **12**, 17055–17061 (2020).

59. Xia, S. et al. Evaluation of charged defect energy in two-dimensional semiconductors for nanoelectronics: the wlz extrapolation method. *Ann. Phys.* **532**, 1900318 (2020).

60. Xiao, J. et al. Realistic dimension-independent approach for charged-defect calculations in semiconductors. *Phys. Rev. B* **101**, 165306 (2020).

61. Göransson, C., Olovsson, W. & Abrikosov, I. A. Numerical investigation of the validity of the slater-janak transition-state model in metallic systems. *Phys. Rev. B* **72**, 134203 (2005).

62. Sanna, S., Frauenheim, T. & Gerstmann, U. Validity of the slater-janak transition-state model within the lda+u approach. *Phys. Rev. B* **78**, 085201 (2008).

63. Li, Y., Sanna, S. & Schmidt, W. G. Modeling intrinsic defects in linbo₃ within the slater-janak transition state model. *J. Chem. Phys.* **140**, 234113 (2014).

64. Pandey, M. et al. Defect-tolerant monolayer transition metal dichalcogenides. *Nano Lett.* **16**, 2234–2239 (2016).

65. Buckeridge, J. Equilibrium point defect and charge carrier concentrations in a material determined through calculation of the self-consistent fermi energy. *Comput. Phys. Commun.* **244**, 329–342 (2019).

66. Kaappa, S., Malola, S. & Häkkinen, H. Point group symmetry analysis of the electronic structure of bare and protected metal nanocrystals. *J. Phys. Chem. A* **122**, 8576–8584 (2018).

67. Togo, A. Spglib. https://atztogo.github.io/spglib/ (2009).

68. Cornwell, J. F. Group theory in physics: an introduction. Academic press (1997).

69. Stoneham, A. M. Theory of defects in solids: electronic structure of defects in insulators and semiconductors. Oxford University Press (2001).

70. Blöchl, P. E. First-principles calculations of defects in oxygen-deficient silica exposed to hydrogen. *Phys. Rev. B* **62**, 6158 (2000).

71. Gunnarsson, O. & Lundqvist, B. I. Exchange and correlation in atoms, molecules, and solids by the spin-density-functional formalism. *Phys. Rev. B* **13**, 4274 (1976).

72. Alkauskas, A., Buckley, B. B., Awschalom, D. D. & Van de Walle, C. G. First-principles theory of the luminescence lineshape for the triplet transition in diamond nv centres. *N. J. Phys.* **16**, 073026 (2014).

73. Larsen, A. H. et al. The atomic simulation environment – a python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).

74. Bergerhoff, G., Hundt, R., Sievers, R. & Brown, I. The inorganic crystal structure data base. *J. Chem. Inf. Comput. Sci.* **23**, 66–69 (1983).

75. Levin, I. Nist inorganic crystal structure database (icsd). National Institute of Standards and Technology (2018).

76. Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V. & Kis, A. Single-layer mos$_2$ transistors. *Nat. Nanotechnol.* **6**, 147–150 (2011).

77. Withers, F. et al. Light-emitting diodes by band-structure engineering in van der waals heterostructures. *Nat. Mater.* **14**, 301–306 (2015).

78. Koppens, F. et al. Photodetectors based on graphene, other two-dimensional materials and hybrid systems. *Nat. Nanotechnol.* **9**, 780–793 (2014).

79. Wang, Q. H., Kalantar-Zadeh, K., Kis, A., Coleman, J. N. & Strano, M. S. Electronics and optoelectronics of two-dimensional transition metal dichalcogenides. *Nat. Nanotechnol.* **7**, 699–712 (2012).

80. Wang, F. et al. Ionic liquid gating of suspended mos$_2$ field effect transistor devices. *Nano Lett.* **15**, 5284–5288 (2015).

81. Yue, Z. Ionic gating for ion intercalation. *Nat. Rev. Phys.* **3**, 306–306 (2021).

82. Xiang, D. et al. Surface transfer doping induced effective modulation on ambipolar characteristics of few-layer black phosphorus. *Nat. Commun.* **6**, 1–8 (2015).

83. Shang, M.-H. et al. Elimination of s vacancy as the cause for the *n*-type behavior of mos2 from the first-principles perspective. *J. Phys. Chem. Lett.* **9**, 6032–6037 (2018).

84. Splendiani, A. et al. Emerging photoluminescence in monolayer mos2. *Nano Lett.* **10**, 1271–1275 (2010).

85. Noh, J.-Y., Kim, H. & Kim, Y.-S. Stability and electronic structures of native defects in single-layer mos$_2$. *Phys. Rev. B* **89**, 205417 (2014).

86. Fortunato, E., Barquinha, P. & Martins, R. Oxide semiconductor thin-film transistors: a review of recent advances. *Adv. Mater.* **24**, 2945–2986 (2012).

87. Kormath Madam Raghupathy, R., Wiebeler, H., Kühne, T. D., Felser, C. & Mirhosseini, H. Database screening of ternary chalcogenides for *p*-type transparent conductors. *Chem. Mater.* **30**, 6794–6800 (2018).

88. Xu, J., Liu, J.-B., Wang, J., Liu, B.-X. & Huang, B. Prediction of novel *p*-type transparent conductors in layered double perovskites: a first-principles study. *Adv. Funct. Mater.* **28**, 1800332 (2018).

89. Bhatia, A. et al. High-mobility bismuth-based transparent *p*-type oxide from high-throughput material screening. *Chem. Mater.* **28**, 30–34 (2016).

90. Deák, P., Aradi, B., Kaviani, M., Frauenheim, T. & Gali, A. Formation of nv centers in diamond: a theoretical study based on calculated transitions and migration of nitrogen and vacancy related defects. *Phys. Rev. B* **89**, 075203 (2014).

91. Lenef, A. & Rand, S. Electronic structure of the nv-center in diamond: theory. *Phys. Rev. B* **53**, 13441 (1996).

92. Tan, A. M. Z., Freysoldt, C. & Hennig, R. G. Stability of charged sulfur vacancies in 2d and bulk mos 2 from plane-wave density functional theory with electrostatic corrections. *Phys. Rev. Mater.* **4**, 064004 (2020).

93. Walsh, A. & Zunger, A. Instilling defect tolerance in new compounds. *Nat. Mater.* **16**, 964–967 (2017).

94. Sajid, A. & Thygesen, K. S. $V_N c_B$ defect as source of single photon emission from hexagonal boron nitride. *2d Mater.* **7**, 031007 (2020).

95. Enkovaara, J. et al. Electronic structure calculations with gpaw: a real-space implementation of the projector augmented-wave method. *J. Phys. Condens. Matter* **22**, 253202 (2010).

96. Pulay, P. Convergence acceleration of iterative sequences. the case of scf iteration. *Chem. Phys. Lett.* **73**, 393–398 (1980).

AUTHOR CONTRIBUTIONS

F.B., S.A., S.M. contributed equally and are co-first authors. F.B. and K.S.T. developed the initial concept, F.B. developed the workflow and database, F.B., S.A., S.M. conducted initial benchmarks and ran DFT calculations, S.A. and S.M. conducted excited state calculations, S.A. conducted the analysis of magneto-optical properties and the identification of potential defects for quantum technological applications, S.M. conducted the symmetry analysis, F.B. conducted the analysis of relaxation effects, CTLs, dopabilities, defect formation trends, and defect tolerances. K.S.T. supervised the work and helped in interpretation of the results. All authors modified and discussed the paper together.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-022-00730-w.

Correspondence and requests for materials should be addressed to Fabian Bertoldo or Sajid Ali.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

9.3 Publication III: High-Throughput Doping of 2D Materials: Absorption *versus* Adsorption

F. BERTOLDO
J. DAVIDSSON
R. ARMIENTO
K. S. THYGESEN

In preparation.

High-Throughput Doping of 2D Materials: Absorption *versus* Adsorption

Fabian Bertoldo[1,*]**, Joel Davidsson**[2,*,†]**, Rickard Armiento**[2]**, Kristian S. Thygesen**[1]

[1]CAMD, Computational Atomic-Scale Materials Design, Department of Physics, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
[2]Department of Physics, Chemistry and Biology, Linköping University, SE-581 83 Linköping, Sweden
[*]These authors contributed equally: Fabian Bertoldo, Joel Davidsson
[†]Corresponding author: joel.davidsson@liu.se

Abstract. Doping of a two-dimensional (2D) material by foreign atoms can take place via two distinct mechanisms: absorption of the dopants by the 2D crystal or adsorption on its surface. To determine the doping mechanism, we systematically dope 53 experimentally synthesized 2D monolayers by 65 different chemical elements in both absorption and adsorption sites. The resulting 11579 doped monolayer structures were generated using the newly developed ASE `DefectBuilder`—a Python tool to set up point defects in 2D and bulk materials—and subsequently relaxed by an automated high-throughput DFT workflow. We found that interstitial positions are preferred for small dopant atoms with partially filled valence electrons in host materials with large lattice parameters. All relaxed structures and calculated defect formation energies are available in an open database that will help advancing our understanding of defects in low-dimensional materials.

1. Introduction

Atomically thin 2D materials form a unique material platform for building advanced nanoscale devices[1, 2, 3, 4] with unique control of electrons down to the level of individual quantum states[5, 6]. The physical properties of 2D materials can be tuned in a variety of ways, *e.g.* by application of mechanical strain[7, 8] or electric fields[9, 10, 11], stacking of monolayers into multilayers[12], molecular functionalization via their surface[13] or by introduction of dopant atoms. Although the introduction of foreign atoms can have detrimental impact on certain materials properties, such as carrier mobility or lifetime[14], they can also be used to control the amount of charge carriers in a semiconductor or even instill new properties such as localised electron states with distinct emission lines[15, 16, 17, 18], magnetism[19, 20, 21, 22], or active catalytic sites[23, 24, 25, 26].

When a 2D material is doped by foreign atoms, the detailed position of the dopants, in particular whether they are located in the interior of the material or on its surface plane, can be decisive for how they influence the material's properties. This makes it important to establish the relative stability of adsorption *versus* absorption sites for 2D dopants in general. Previous first-principles studies have shown that 2D transition metal dichalcogenides (TMDs) doped by transition metal atoms can favor either internal or surface dopant sites, depending on the dopant species[27]. While first-principles calculations have been widely used to investigate the role of specific dopants in specific 2D host materials[27, 28], there exists to date no systematic study of doping in 2D materials across many different host materials and dopant species.

In this study, we turn to high-throughput calculations to answer whether a given dopant adsorbs, *i.e.* stays on the surface as an adatom, or absorbs, *i.e.* goes into the material as an interstitial, when used to dope a 2D material. We systematically dope 53 experimentally known 2D monolayers from the computational 2D materials database (C2DB)[29, 30] with 65 different atomic species in interstitial positions and adsorption sites. To facilitate the structure setup, we employ the `DefectBuilder` module of the Atomic Simulation Environment (ASE)[31], which is a general tool for constructing point defect structures in bulk and 2D materials. For each crystal-dopant combination, we evaluate the formation energy of a

range of inequivalent absorption and adsorption sites and determine the energetically most favorable. We identify general trends in the dataset of 11579 relaxed defect structures and collect the data in an open-access database, making it a useful resource for future investigations of interstitial and adsorption site doping by machine learning methods.

The paper is organised as follows: Sec. 2 introduces the set of host materials, explains the ASE `DefectBuilder` tool used to setup the initial structures, gives details about the computational workflow, defines key parameters for the interpretation of the results, and presents the resulting database. In Sec. 3, the methodology is first validated against existing data in the literature for the specific class of 2H-MoX$_2$ monolayers, before identifying general trends in the entirety of our system pool. We analyze the convergence behavior for the high-throughput calculations and find interesting correlations to improve future studies of similar nature. Finally, we summarize our findings and look ahead in Sec. 4 to describe how the existing data can be used to gain additional insights into possible future studies.

2. Methodology

2.1. Host materials

The set of host materials was selected by screening the Computational 2D Materials Database (C2DB) [30] for materials that have been previously synthesised in monolayer form. From the resulting 55 monolayers we removed the one-atom thick materials graphene and hexagonal boron-nitride (hBN). These materials were removed because: (i) absorption in interstitial sites is not well defined in such materials. (ii) Our calculations show that interstitials in fully planar systems are particularly challenging to converge with respect to in-plane supercell size (see Supplementary Note 2). (iii) The materials can exhibit a large variety of buckled structures depending on the dopant [32]. An overview of the host materials with their space group number is collected in Supplementary Table 1 of the supporting information. For a selection of host materials we performed convergence tests to determine the minimal supercell size needed for reliable results, see Supplementary Figure 3. Based on these tests, supercells of at least 10 Å were chosen for all calculations.

2.2. DefectBuilder

Large-scale studies of crystal point defect studies rely on tools to automatically define and set up the relevant defect structures. In this work, we introduce the ASE `DefectBuilder`, a useful module within the Atomic Simulation Environment[31] to construct defect structures and supercells. Figure 1 gives an overview of the functionalities currently supported by the `DefectBuilder`: after supplying the pristine crystal as an input structure in one of the numerous ASE structure formats, the `DefectBuilder` can generate single point defects like (i) vacancies, (ii) substitutional defects, (iii) interstitial defects and, for 2D materials, (iv) adsorption sites. For (i) and (ii), the module analyses the Wyckoff positions of the input structure and generates a vacancy and substitution site for each inequivalent position. The user can specify which elements to substitute the respective sites with. Creating interstitial defects is based on the algorithm developed for the ADAQ framework[33] which relies on Voronoi tessellation of the host crystal and subsequent mapping of those Voronoi positions onto Wyckoff positions. A minimum distance between interstitial positions and atomic positions of the host crystal can be given as an input to control the number of generated interstitial sites. A larger minimum distance will produce fewer interstitial sites. More details can be found in the source code and the documentation of the `DefectBuilder` class.[34, 35]

The interstitial site generation algorithm is further generalized to set up adsorption sites: for a given 2D input structure, the algorithm extracts the atoms from the topmost and lowermost part of the monolayer and conducts the Voronoi tessalation and Wyckoff mapping in the 2D subspace. Afterwards, the 2D interstitial positions will be translated out of the layer up to the point where the minimum distance between the adsorbate and the top or bottom of the layer will be equal to the sum of the covalent radius of dopant atom and closest atom in the host monolayer.

In addition, the `DefectBuilder` features simple functionalities to set up suitable defect supercells, *e.g.* by specification of a minimum distance between periodic repetitions of the defect. All input parameters for the `DefectBuilder` are documented in the source code[35] and a detailed description can be found in the documentation[34].

2.3. Workflow

The calculations were carried out using the high-throughput toolkit (*httk*) [36, 37] and the Vienna Ab initio Simulation Package (VASP) [38, 39]. VASP is implemented with the projector augmented wave (PAW)[40, 41] method and performs density functional theory (DFT)[42, 43] calculations. The Perdew, Burke, and Ernzerhof (PBE)[44] exchange-correlation functional was used and all calculations were performed with spin polarization. To speed up the calculations, the Brillouin zone (BZ) was sampled at the Γ-point only, which allowed us to use the Γ complied version of VASP for maximum performance. The default VASP pseudopotentials were used[45] with the plane wave energy cutoff set to 600 eV and kinetic energy cutoff to 900 eV for all elements. Calculations were only performed for defects in their neutral charge states.

To ensure a fast and accurate relaxation of the vast number of defects, we employed a two-stage workflow inspired by ADAQ[33]. The common settings are described above. The different settings for electronic and ionic tolerance as well as the Fast Fourier Transform (FFT) grid between the stages are shown in Table 1. Both stages relax the atom positions and limit the ion relaxation to 20 steps. Hence, a maximum of 40 ionic steps are taken for any given defect system. The defect system does not have to reach the ionic tolerance in the final stage, the runs are saved to the database with the final ionic convergence. For the analysis, we use only structures with a final ionic convergence of $5 \cdot 10^{-2}$ eV or less. Calculations reaching this criterion are from now on referred to as converged for simplicity.

To evaluate the defect formation energy we set to the chemical potential of the dopant element to the energy of the crystalline phases[46]. The elemental crystal structures were relaxed using a workflow similar to the one described above, but with an initial volume relaxation step. To ensure that the plane wave basis is accurate even when the volume is changed, the plane wave energy cutoff is increased to 700 eV and kinetic energy cutoff to 1400 eV. The k-point grid is set up automatically with *httk* using the Monkhorst-Pack method[47]. Additional settings are shown in Table 1. This step is restarted until the energy difference between two steps is less than $5 \cdot 10^{-4}$ eV.

Table 1. Settings for the automatic workflows. The FFT grid column show the size compared to the largest wave vector.

Stage	Electronic tolerance [eV]	Ionic tolerance [eV]	FFT grid
first	10^{-4}	$5 \cdot 10^{-3}$	3/2
second	10^{-6}	10^{-4}	2
chemical potential	10^{-6}	$5 \cdot 10^{-5}$	2

2.4. Classification parameters

In order to classify the relaxed defect structures as adsorption or absorption configurations, we introduce a depth parameter, D. For a given dopant X in a host
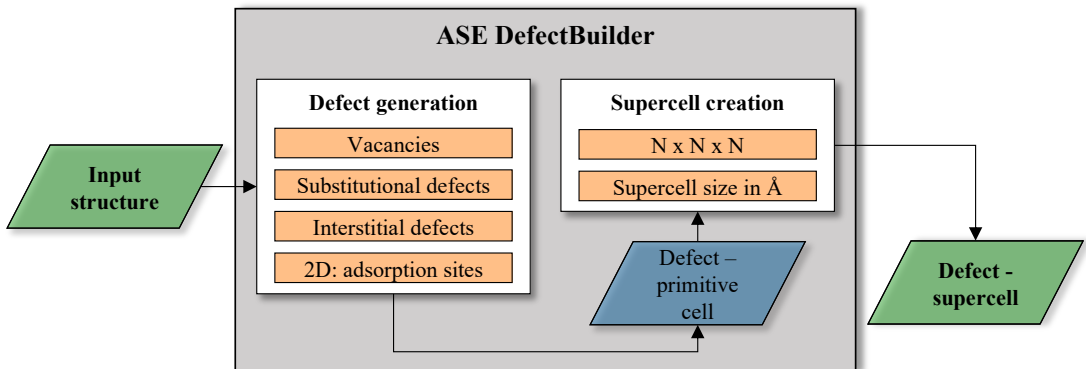
Figure 1. Structure of the `DefectBuilder` class in ASE. The `DefectBuilder` takes an input structure in the commonly known ASE format, generates different kinds of single point defects that are afterwards embedded into a supercell (which can either be set up by simple NxNxN repetitions or a physical minimum defect-defect distance criterion), and finally returns the desired defect structures in their respective supercells.

material H, the depth parameter is defined by

$$D = \frac{2z\,[\mathrm{X}] - \left(z_{\min}\,[\mathrm{H}] + z_{\max}\,[\mathrm{H}]\right)}{z_{\max}\,[\mathrm{H}] - z_{\min}\,[\mathrm{H}]}. \tag{1}$$

Here, $z_{\min}\,[\mathrm{H}]$ ($z_{\max}\,[\mathrm{H}]$) is the lowermost (topmost) z-position in the pristine host structure and $z\,[\mathrm{X}]$ denotes the z-position of the dopant atom. Values $|D| < 1$ correspond to absorption in an interstitial site while $|D| \geq 1$ implies an adsorption site. The sign of D indicates whether the dopant sits above or below the center of the pristine monolayer, and the values -1 and $+1$ correspond to the dopant sitting exactly at the lower or upper boundary of the host crystal.

For some systems the addition of a dopant into the structure can lead to large distortions. Ideally, we would like the defect to only introduce small local changes to the structure and not an entire reorganization of the host crystal. To quantify the amount of distortion introduced by the dopant atom we introduce the expansion factor, $\mathrm{XF}\,[\mathrm{H}, \mathrm{X}]$, as

$$\mathrm{XF}\,[\mathrm{H}, \mathrm{X}] = \frac{d_{\mathrm{unrel}}\,[\mathrm{H}, \mathrm{X}]}{d_{\mathrm{rel}}\,[\mathrm{H}, \mathrm{X}]}, \tag{2}$$

where d_{unrel} (d_{rel}) denotes the thickness of the monolayer plus dopant before (after) relaxation. A large expansion factor, *i.e.* $\mathrm{XF}\,[\mathrm{H}, \mathrm{X}] > 2$, indicate an unphysically large restructuring of the monolayer. This can for example happen when a large atom is introduced in a tight interstitial volume and leads to disintigration the monolayer during relaxation. Not unexpectedly, we find a strong correlation between large expansion factors and unconverged calculations.

2.5. The database

All of the interstitials and adsorption site systems have been subject to the workflow described in Sec. 2.3.

As a result, we created 11579 fully relaxed structures and collected them in an ASE database[31]. Each row of the database contains the relaxed atomic structure of the defect system and is uniquely defined by its host name (`host`), doping site (`site`), and dopant atom (`dopant`). Furthermore, we store numerous key-value pairs (KVPs) for easy querying of the data, *e.g.* formation energy (`eform`), depth-parameter (`depth`), expansion factor (`extension_factor`), spin (`spin`), convergence (`converged`). Furthermore, a web application of the database will be available where users can interactively inspect the relaxed atomic structures of the interstitial and adsorption site doped materials, as well as all of their corresponding KVPs.

3. Results

3.1. Transition metal doping of 2H-MoX$_2$ monolayers

We first perform a validity check of the methodology by comparing to the previous work of Karthikeyan *et al.*[27] who studied the adsorption/absorption of transition metal dopants in MoX_2 (X = S, Se, Te). Figure 2 shows the formation energies of transition metal doped MoX_2 monolayers obtained with our HT workflow. Generally, the transition metal dopants are found to be more stable in adsorption sites for MoS_2 and MoSe_2 whereas interstitial sites become more favorable for MoTe_2. This trend can be explained by the larger lattice constant of MoTe_2, which implies larger spaces to accomodate the dopant in an interstitial site. One also observes a hill-like shape of the formation energy *versus* d-band filling of the dopant, in particular for the $4d$ and $5d$ series. This trend close follows the trend in cohesive energy of the transition metal solids[48] indicating that the behavior
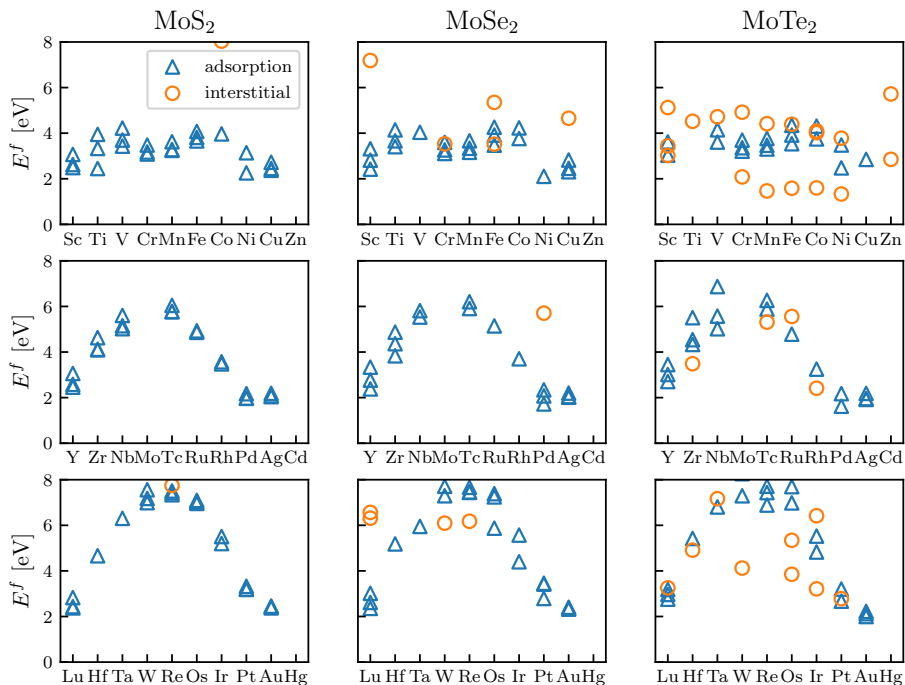
Figure 2. Transition metal doping in transition metal dichalcogenides 2H-MoX$_2$. Formation energies E^f of interstitial (orange circles) and adsorption site (blue triangles) doped MoS$_2$ (left column), MoSe$_2$ (middle column) and MoTe$_2$ (right column) as a function of the doping element.

simply reflects the variation in the chemical potential of the dopants. Both these correlations are in agreement with previous studies on transition metal doping of MoX$_2$ monolayers[27] as well with our analysis on general convergence behavior, which will be discussed in Sec. 3.3. We note that Karthikeyan *et al.*[27] used more accurate computational parameters (*i.e.* 6x6x1 supercells and k-point sampling), as well as different references for the chemical potentials of dopant atoms, explaining small quantitative deviations between Ref. [27] and our results.

Interestingly, we observe the same overall trend for other types of 2H-MX$_2$ host crystals, *i.e.* WX$_2$, NbX$_2$ (see Supplementary Note 1 of the supporting information) which further backs the idea that the stability of interstitial sites is closely tied to the size of the host lattice. The more free space is available inside a 2D monolayer, the more likely it is for interstitial sites to form. In the following sections, we try to confirm this hypothesis by looking at 2D materials of different

prototypes doped with elements of the entire periodic table.

3.2. General trends

After having considered a few specific 2D monolayers and dopants in the previous section, we now explore trends in the entire data set of 53 host crystals and 65 dopants. In particular, we focus on the question: which combinations of host material H and dopant atom X favor interstitial defects over adsorbates. To that end, we introduce the quantity,

$$\Delta\left[H, X\right] = E_{\text{int}}^{f,\min}\left[H, X\right] - E_{\text{ads}}^{f,\min}\left[H, X\right], \qquad (3)$$

where $E_x^{f,\min}\left[H, X\right]$ ($x = \text{ads, int}$) is the minimum formation energy of a dopant X in host crystal H either at an adsorption or interstitial site (as defined by the depth parameter in Eq. (1)). We note that $\Delta\left[H, X\right]$ is only defined if at least one adsorption and interstitial configuration has been converged for the given system (see next section). A negative value of
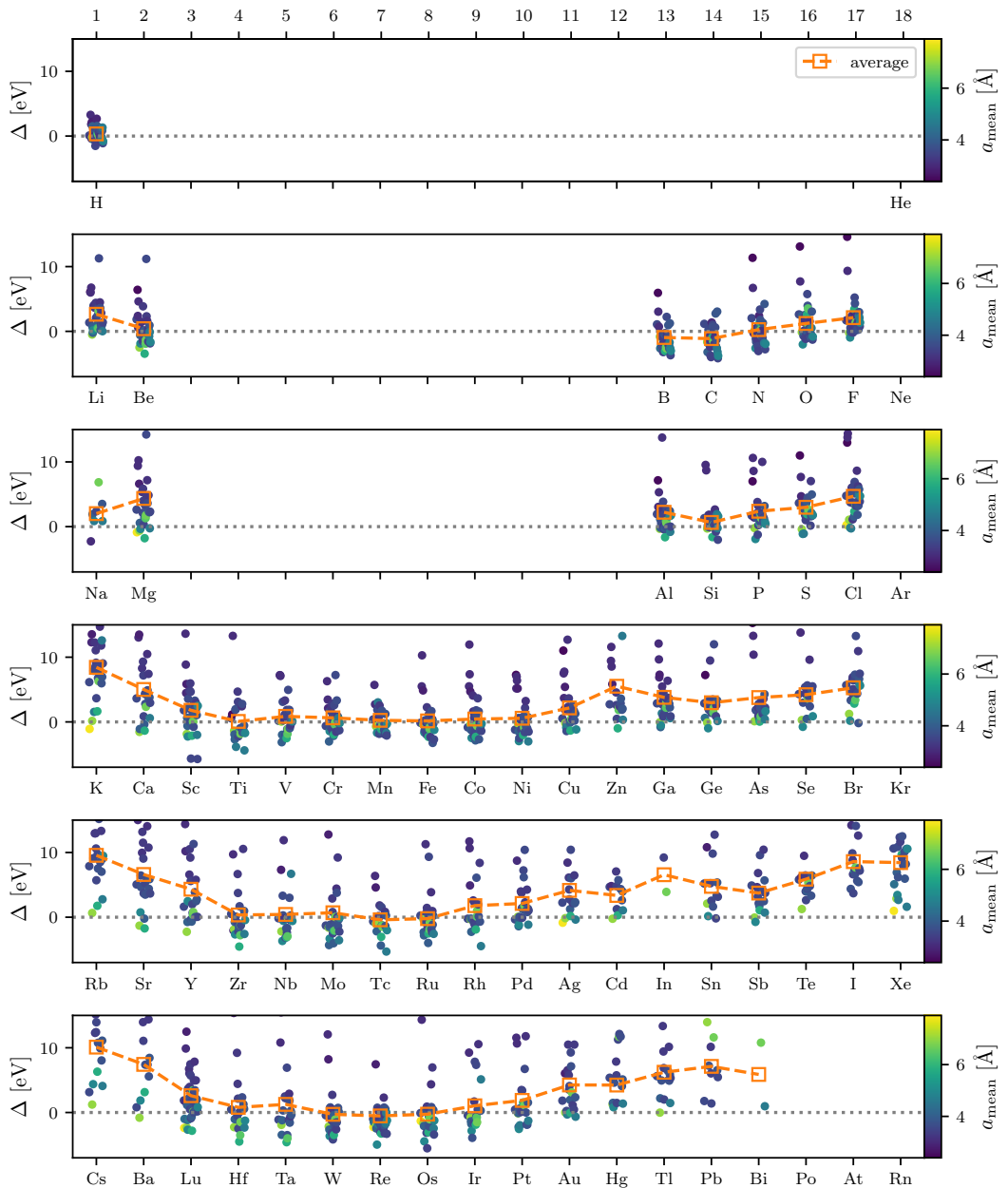
Figure 3. Doping trends for all hosts and dopant elements. $\Delta[\mathrm{H}, \mathrm{X}]$ as a function of the respective dopant element. Negative values correspond to stable interstitials. The different dots in one particular column represent the different host crystals. The average lateral size of the primitive host crystal unit cell is given by the colorcode of the data points. The orange squares visualize the average Δ-value over all host crystals for a given dopant.

$\Delta\,[\mathrm{H}, \mathrm{X}]$ indicates that the interstitial position is more energetically favorable than the adatom position, and *vice versa* for positive values of $\Delta\,[\mathrm{H}, \mathrm{X}]$.

Figure 3 shows the calculated Δ-values for all the considered host crystals and doping elements. The lattice constant (precisely the average of the length of the basis vectors of the primitive unit cell) is indicated by the color code.

For dopants in the first row where only hydrogen was considered, the Δ values are distributed around zero and there is no clear preference for adsorption or absorption. For dopants of the second period we see that Li and F prefer adsorption while B and C prefer absorption. In contrast, Be, N, and O the preference for adsorption/absorption is highly system dependent. Dopants from period 3 have generally larger Δ values and most of the elements have a preference for adsorption. Exceptions occur for Si and, to a lesser extent Al and P, which can also prefer absorption for specific materials.

Dopants from periods 4-6 show very similar trends across the groups of the periodic table, indicating that the preference for adsorption/absorption is mainly dictated by the chemical nature of the dopant atoms. It is clear that adsorption sites are favored for dopant elements from groups 1 and 2, whereas interstitial sites are preferred for the early and middle transition metal dopants with the exception of the group 3 elements (Sc, Y, Lu), which have a slight tendency to prefer adsorption. In contrast, the late transition metals (group 10-12) generally favor adsorption. Beyond the transition metal series, absorption is again preferred with exceptions occurring for very few host material. The same holds for the transition metals with a single d-electron or an almost filled d-shell. In fact, there seems to be a correlation between the number of d-electrons

Not unexpectedly, there is a correlation between Δ and the lattice constant of the host crystal (indicated by the color coding in Figure 3): larger lattice constants are correlated with smaller Δ-values. This observation is a clear indication that the stability of interstitial sites is highly dependent on the available free space inside a monolayer and generalizes the corresponding trend observed in Sec. 3.1. Quantifying these correlations (*e.g.* in terms of machine learning methods) should therefore be the emphasis of future studies.

Even if there are large variations in Δ depending on which host material the dopant is placed in, the general trend across all 2D host materials is clear: small dopants in spacious host materials are preferred. The dopants of the s-block are large and rarely found as interstitials, except H which plays in a league of its own with an average Δ at zero with minimal variation

across host materials. The elements in group 2 are smaller than in group 1, and the average Δ is lower for those elements. Furthermore, the number of valence electrons also plays an important part. Even if the elements in the p-block are getting smaller as the group number increases in each period, there are noticeable dips in the average Δ for the half-filled elements in Figure 3. Group 14 and 15 dopants have a lower average Δ than groups 13, 16, and 17. This trend indicates that not just size is important but also the possibility to form bonds. For the d-block, the elements do not vary noticeably in size and show large variations. Also, the trend of half-filled valence is not clear from the average. Although, the groups 3 and 12 have a higher average Δ compared to the rest. One can see that there are more points below the zero line for the sixth period. Hence, the general trend is small dopants with half-filled valence electrons in spacious host materials are preferred.

3.3. Convergence statistics

Since a number of calculations did not converge within the set limits of 20 ionic steps, it is of interest to check whether such systems could generally be assumed to be unphysical/unstable. For example, systems with a dopant placed in a small interstitial volume is unlikely to result in a stable configuration and is simultaneously expected to require many relaxation steps to reach a local minimum. If such a correlation exists, unconverged calculations would be less of a concern, as they would be unlikely to produce physically relevant information even if they were brought to convergence.

When considering all the calculations that did not converge in their ionic relaxation (in the following referred to as 'failed') we see that around 75% of these represent interstitial configurations while 25% represent adsorption.

Figure 4 shows the percentage of failed calculations as a function of the covalent radius of the dopant atom across all host crystals for interstitial and adsorption positions, respectively. It is clear that the failure rate is higher for interstitial doping. On one hand, the failure rate for interstitials shows a clear correlation with the covalent radius of the dopant atom. In general, it is difficult to converge calculations for interstitial dopants with covalent radius above 1 Å. On the other hand, no such correlation is seen for dopants at adsorption sites, although there are a few outliers with above average failure rate.

In Figure 5, a similar trend is obtained in terms of the failure percentage as a function of dopant type and atomic number of dopant element: the convergence for systems doped at adsorption sites is generally better and not strongly correlated with the atomic number of the dopant element. For interstitials,
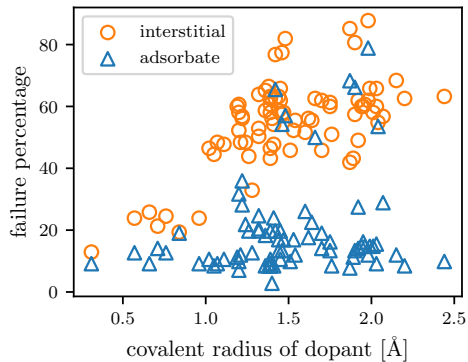
Figure 4. Calculation failure percentage and covalent radii. Percentage of failed calculations as a function of covalent radius of the dopant element for adsorption (blue triangles) and interstitial (orange circles).
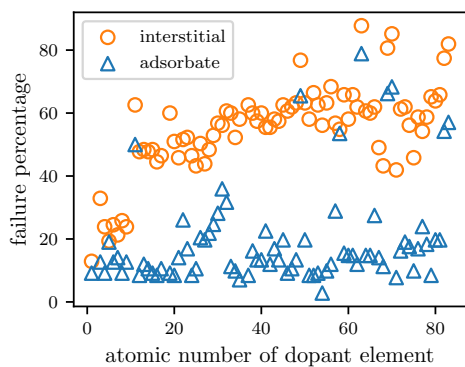

Figure 5. Calculation failure percentage and dopant atomic numbers. Percentage of failed calculations as a function of atomic number of the dopant element for adsorption (blue triangles) and interstitial (orange circles).

however, a low failure rate is only present for dopant elements with atomic numbers lower than 15, whereas heavier dopant elements increase the percentage of unconverged calculations.

Interstitial sites seem to be the bottleneck, and an in-depth analysis of those systems is needed. For a given system doped with an interstitial atom I, there exists one atom A in the host crystal that possesses the minimum distance, $D_{\min}(I, A)$, to the interstitial site. To quantify the convergence behavior of the interstitial systems, we define a normalized minimum distance parameter, $D_{\min}^{\mathrm{norm}}(I, A)$, as

$$D_{\min}^{\mathrm{norm}}(I, A) = \frac{D_{\min}(I, A)}{r_{\mathrm{C}}(I) + r_{\mathrm{C}}(A)}. \tag{4}$$

Here, $r_{\mathrm{C}}(A)$ and $r_{\mathrm{C}}(I)$ denote the covalent radii for the corresponding host atom and interstitial atom, respectively. Figure 6 visualizes the percentage of converged and unconverged calculations for interstitial systems as a function of $D_{\min}^{\mathrm{norm}}(I, A)$. For low normalized minimum distances almost no calculations converge but with increasing $D_{\min}^{\mathrm{norm}}(I, A)$ the percentage of converged calculations increases. For normalized minimum distance values above 0.7, the probability of converging an interstitial position is generally larger than 50%. This trend is confirmed by the cumulative percentage of converged interstitial calculations: 56% for $D_{\min}^{\mathrm{norm}}(I, A) > 0.8$ and only 34% for $D_{\min}^{\mathrm{norm}}(I, A) > 0.3$.

We note that the ionic convergence settings are set relatively low due to the vast number of systems that needed to be calculated. In particular, not to waste computational resources, the ionic steps are limited to 20 since the ions may oscillate back and forth for some systems. Some unconverged structures can be used as a good starting point for a detailed subsequent analysis. Unconverged systems can be found in the database *via* the query: `converged=False`.

In summary, we saw that our unconverged calculations can give insight into the stability of adsorption and interstitial site doping of 2D materials. For interstitial defects, the overall large percentage of unconverged calculations stems from unphysical systems where an interstitial site simply does not have enough space inside the monolayer. For instance, we saw that dopant atoms with large covalent radii are particularly challenging systems. Therefore, in the context of high-throughput calculations, it is important to choose physically reasonable initial configurations for interstitial positions (we suggest normalized minimum distances of at least 0.8) to end up with convergence percentages of above 50%. On top of the structural challenges, our approach of doping the monolayers with most of the elements from the periodic table regardless of the chemical compatibility leads to large percentages of unconverged systems. In a future study, a thorough analysis of the convergence behavior in terms of a chemical compatibility descriptor might be helpful for more insight. The public availability of the dataset with its documentation online should make such an additional study easy to conduct.

4. Conclusions and outlook

In conclusion, we presented the ASE `DefectBuilder` – a powerful and easy-to-use tool for setting up point defect structures within the Atomic Simulation Environment (ASE)[31]. We utilized the `DefectBuilder` to construct almost 18000 interstitial point defects and adsorption structures by combining 65 dopant elements with 53 different 2D materials, which have all been ex-
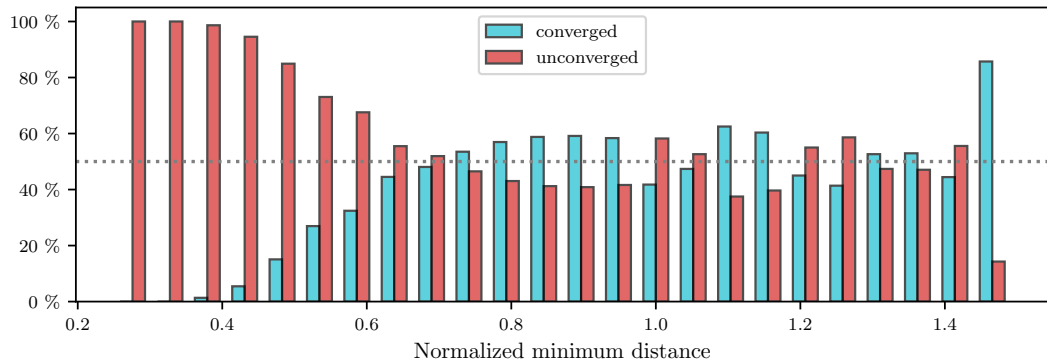
Figure 6. Convergence behaviour and nearest neighbor distances for interstitial sites. Bar diagram of the percentage of converged *versus* unconverged calculations as a function of the normalized minimum distance parameter $D_{\min}^{\mathrm{norm}}$ (I, A). Each pair of bins add up to 100 %.

perimentally realised in monolayer form[29, 30]. Each of the doped structures were subject to a relaxation and ground state workflow implemented within the *httk*[36, 37] high-throughput framework.

The resulting data was first validated for transition metal doped MoX_2 (X = S, Se, Te) monolayers and showed good qualitative agreement with previous studies[27]. In addition, the trends of interstitial and adsorption site stability in MoX_2 monolayers was generalized to other types of 2H-TMDs such as WX_2 and NbX_2 (see supporting information). We saw that interstitial doping is generally very challenging over the entire set of 2D monolayers, especially for doping elements with large covalent radii. However, the combination of light elements like carbon, as well as transition metals in host crystals with relatively large lattice constants are promising systems to realize interstitial doping in 2D materials.

The overall low percentage of converged structures is mostly due to the challenging task of relaxing interstitial defects. We note that a more computationally heavy workflow that allows for a larger number of ionic relaxation steps might improve that number in future studies. Many of the generated initial structures turned out to be unphysical systems, and we suggest to always choose a reasonable normalized minimum distance when setting up interstitial structures. Furthermore, the amount of failed/unconverged calculations is correlated with the atomic species a given system is doped with, in particular for interstitials.

All of the data used for the analysis of this study has been collected in an ASE database and will be made publicly available in a web-application. The database is supposed to mark the starting point for future investigations of interstitial *versus* adsorption site doping in 2D materials. One obvious extension could be the prediction of formation energies outside the existing host material set by means of machine learning. Furthermore, both the relaxed and almost relaxed structures inside the database can be the starting point to apply more accurate calculations which in turn can be the basis for an extended property evaluation, *e.g.* optical transitions, charge transition levels, excited state properties *etc.* The qualitative correlations discussed in the present work could be extended towards a more quantitative analysis.

5. Data availability

The data that support the findings of this study are openly available at the following URL/DOI: `https://doi.org/10.11583/DTU.19692238.v1`.

6. Code availability

The sourcecode for the ASE `DefectBuilder` can be found on gitlab: `https://gitlab.com/ase/ase/-/tree/defect-setup-utils/ase`. Some simple code examples for the setup of defect structures is available at: `https://gitlab.com/ase/ase/-/blob/defect-setup-utils/doc/ase/build/defects.rst`.

References

[1] Ferrari A C, Bonaccorso F, Fal'Ko V, Novoselov K S, Roche S, Bøggild P, Borini S, Koppens F H, Palermo V, Pugno N *et al.* 2015 *Nanoscale* **7** 4598–4810

[2] Schaibley J R, Yu H, Clark G, Rivera P, Ross J S, Seyler K L, Yao W and Xu X 2016 *Nature Reviews Materials* **1** 1–15

[3] Sierra J F, Fabian J, Kawakami R K, Roche S and Valenzuela S O 2021 *Nature Nanotechnology* **16** 856–868

[4] Lin X, Yang W, Wang K L and Zhao W 2019 *Nature Electronics* **2** 274–283

[5] Liu X and Hersam M C 2019 *Nature Reviews Materials* **4** 669–684

[6] Turiansky M, Alkauskas A and Van de Walle C 2020 *Nature materials* **19** 487–489

[7] Dai Z, Liu L and Zhang Z 2019 *Advanced Materials* **31** 1805417

[8] Conley H J, Wang B, Ziegler J I, Haglund Jr R F, Pantelides S T and Bolotin K I 2013 *Nano Letters* **13** 3626–3630

[9] Yu Y J, Zhao Y, Ryu S, Brus L E, Kim K S and Kim P 2009 *Nano Letters* **9** 3430–3434

[10] Leisgang N, Shree S, Paradisanos I, Sponfeldner L, Robert C, Lagarde D, Balocchi A, Watanabe K, Taniguchi T, Marie X *et al.* 2020 *Nature Nanotechnology* **15** 901–907

[11] Peimyoo N, Deilmann T, Withers F, Escolar J, Nutting D, Taniguchi T, Watanabe K, Taghizadeh A, Craciun M F, Thygesen K S *et al.* 2021 *Nature Nanotechnology* **16** 888–893

[12] Novoselov K, Mishchenko o A, Carvalho o A and Castro Neto A 2016 *Science* **353** aac9439

[13] Brill A R, Koren E and de Ruiter G 2021 *Journal of Materials Chemistry C* **9** 11569–11587

[14] Polman A, Knight M, Garnett E C, Ehrler B and Sinke W C 2016 *Science* **352** aad4424

[15] Awschalom D D, Bassett L C, Dzurak A S, Hu E L and Petta J R 2013 *Science* **339** 1174–1179

[16] Gomonay O 2018 *Physics* **11** 78

[17] Eckstein J N and Levy J 2013 *MRS Bull.* **38** 783–789

[18] Gardas B, Dziarmaga J, Zurek W H and Zwolak M 2018 *Sci. Rep.* **8** 1–10

[19] Friend R and Yoffe A 1987 *Advances in Physics* **36** 1–94

[20] Zhao X, Song P, Wang C, Riis-Jensen A C, Fu W, Deng Y, Wan D, Kang L, Ning S, Dan J *et al.* 2020 *Nature* **581** 171–177

[21] Coelho P M, Komsa H P, Lasek K, Kalappattil V, Karthikeyan J, Phan M H, Krasheninnikov A V and Batzill M 2019 *Adv. Electron. Mater.* **5** 1900044

[22] Wang J, Sun F, Yang S, Li Y, Zhao C, Xu M, Zhang Y and Zeng H 2016 *Appl. Phys. Lett.* **109** 092401

[23] Chen Y, Yang K, Jiang B, Li J, Zeng M and Fu L 2017 *J. Mater. Chem. A* **5** 8187–8208

[24] Jia Y, Chen J and Yao X 2018 *Mater. Chem. Front.* **2** 1250–1268

[25] Tang C and Zhang Q 2017 *Adv. Mater.* **29** 1604103

[26] Yan D, Li Y, Huo J, Chen R, Dai L and Wang S 2017 *Adv. Mater.* **29** 1606459

[27] Karthikeyan J, Komsa H P, Batzill M and Krasheninnikov A V 2019 *Nano Lett.* **19** 4581–4587

[28] Fu Z, Yang B, Zhang N, Lu Z, Yang Z and Ma D 2017 *J. Phys. Chem. C* **121** 28312–28316

[29] Haastrup S, Strange M, Pandey M, Deilmann T, Schmidt P S, Hinsche N F, Gjerding M N, Torelli D, Larsen P M, Riis-Jensen A C *et al.* 2018 *2D Mater.* **5** 042002

[30] Gjerding M N, Taghizadeh A, Rasmussen A, Ali S, Bertoldo F, Deilmann T, Knøsgaard N R, Kruse M, Larsen A H, Manti S *et al.* 2021 *2D Mater.* **8** 044002

[31] Larsen A H, Mortensen J J, Blomqvist J, Castelli I E, Christensen R, Dułak M, Friis J, Groves M N, Hammer B, Hargus C *et al.* 2017 *J. Phys. Condens. Matter* **29** 273002

[32] Lehtinen O, Vats N, Algara-Siller G, Knyrim P and Kaiser U 2015 *Nano letters* **15** 235–241

[33] Davidsson J, Ivády V, Armiento R and Abrikosov I A 2021 *Comput. Phys. Commun.* **269** 108091

[34] ASE 2022 `DefectBuilder` documentation URL `https://gitlab.com/ase/ase/-/blob/defect-setup-utils/doc/ase/build/defects.rst`

[35] ASE 2022 `DefectBuilder` sourcecode URL `https://gitlab.com/ase/ase/-/blob/defect-setup-utils/ase/build/defects.py`

[36] Armiento R *et al.* 2019 The high-throughput toolkit (httk) `http://httk.openmaterialsdb.se/` URL `http://httk.openmaterialsdb.se/`

[37] Armiento R 2020 *Database-Driven High-Throughput Calculations and Machine Learning Models for Materials Design* (Cham: Springer International Publishing) pp 377–395 ISBN 978-3-030-40245-7 URL `https://doi.org/10.1007/978-3-030-40245-7_17`

[38] Kresse G and Hafner J 1994 *Phys. Rev. B* **49**(20) 14251–14269 URL `http://link.aps.org/doi/10.1103/PhysRevB.49.14251`

[39] Kresse G and Furthmüller J 1996 *Phys. Rev. B* **54**(16) 11169–11186 URL `http://link.aps.org/doi/10.1103/PhysRevB.54.11169`

[40] Blöchl P E 1994 *Phys. Rev. B* **50**(24) 17953–17979 URL `http://link.aps.org/doi/10.1103/PhysRevB.50.17953`

[41] Kresse G and Joubert D 1999 *Phys. Rev. B* **59**(3) 1758–1775 URL `http://link.aps.org/doi/10.1103/PhysRevB.59.1758`

[42] Hohenberg P and Kohn W 1964 *Phys. Rev.* **136**(3B) B864–B871 URL `http://link.aps.org/doi/10.1103/PhysRev.136.B864`

[43] Kohn W and Sham L J 1965 *Phys. Rev.* **140**(4A) A1133–A1138 URL `http://link.aps.org/doi/10.1103/PhysRev.140.A1133`

[44] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77**(18) 3865–3868 URL `http://link.aps.org/doi/10.1103/PhysRevLett.77.3865`

[45] 2022 Paw potentials `https://www.vasp.at/wiki/index.php/Available_PAW_potentials` accessed: 2022-03-30 URL `https://www.vasp.at/wiki/index.php/Available_PAW_potentials`

[46] Lejaeghere K, Van Speybroeck V, Van Oost G and Cottenier S 2014 *Critical Reviews in Solid State and Materials Sciences* **39** 1–24

[47] Monkhorst H J and Pack J D 1976 *Physical Review B* **13** 5188

[48] Kittel C 1976 *Wiley, New York* **19**

7. Acknowledgments

8. Author contributions

F.B. and J.D. developed the initial concept, designed the code base, F.B. implemented the ASE `DefectBuilder` module, J.D. set up and developed the underlying workflow and ran the calculations, F.B. and J.D. analyzed the data and wrote the manuscript draft. K.S.T. and R.A. supervised the work and helped in interpretation of the results. All authors modified and discussed the paper together.

9. Competing interests

The authors declare no competing interests.

9.4   Publication IV: Recent Progress of the Computational Materials Database (C2DB)

M. N. GJERDING
A. TAGIZADEH
A. RASMUSSEN
S. ALI
F. BERTOLDO
T. DEILMANN
N. R. KNØSGAARD
M. KRUSE
A. H. LARSEN
S. MANTI
T. G. PEDERSEN
U. PETRALANDA
T. SKOVHUS
M. K. SVENDSEN
J. J. MORTENSEN
T. OLSEN
K. S. THYGESEN

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.

2D Materials

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

Morten Niklas Gjerding[1] [iD], Alireza Taghizadeh[1,2] [iD], Asbjørn Rasmussen[1] [iD], Sajid Ali[1] [iD], Fabian Bertoldo[1] [iD], Thorsten Deilmann[3] [iD], Nikolaj Rørbæk Knøsgaard[1] [iD], Mads Kruse[1] [iD], Ask Hjorth Larsen[1] [iD], Simone Manti[1] [iD], Thomas Garm Pedersen[2] [iD], Urko Petralanda[1] [iD], Thorbjørn Skovhus[1] [iD], Mark Kamper Svendsen[1] [iD], Jens Jørgen Mortensen[1] [iD], Thomas Olsen[1] [iD] and Kristian Sommer Thygesen[1,*] [iD]

[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

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, FAIR[5] [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 stability assess 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 evaluation 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	$E_{gap}^{PBE} > 0$	784
Second harmonic generation	PBE	$E_{gap}^{PBE} > 0$, non-magnetic, non-centrosymmetric	375
Electronic band structure PBE	PBE*	None	3496
Magnetic anisotropies	PBE*	Magnetic	823
Deformation potentials	PBE*	$E_{gap}^{PBE} > 0$	830
Effective masses	PBE*	$E_{gap}^{PBE} > 0$	1272
Fermi surface	PBE*	$E_{gap}^{PBE} = 0$	2505
Plasma frequency	PBE*	$E_{gap}^{PBE} = 0$	3144
Work function	PBE*	$E_{gap}^{PBE} = 0$	4044
Optical polarisability	RPA@PBE	None	3127
Electronic band structure	HSE06@PBE*	None	3155
Electronic band structure	G_0W_0@PBE*	$E_{gap}^{PBE} > 0$, $N_{atoms} < 5$	357
Born charges	PBE, Berry phase	$E_{gap}^{PBE} > 0$	639
Raman spectrum	PBE, LCAO basis set	Non-magnetic, dyn. stable	708
Piezoelectric tensor	PBE, Berry phase	E_{gap}^{PBE}, non-centrosym.	353
Optical absorbance	BSE@G_0W_0*	$E_{gap}^{PBE} > 0$, $N_{atoms} < 5$	378
Spontaneous polarisation	PBE, Berry phase	$E_{gap}^{PBE} > 0$, nearly centrosym. polar space group	151
Topological invariants	PBE*, Berry phase	$0 < E_{gap}^{PBE} < 0.3$ eV	242
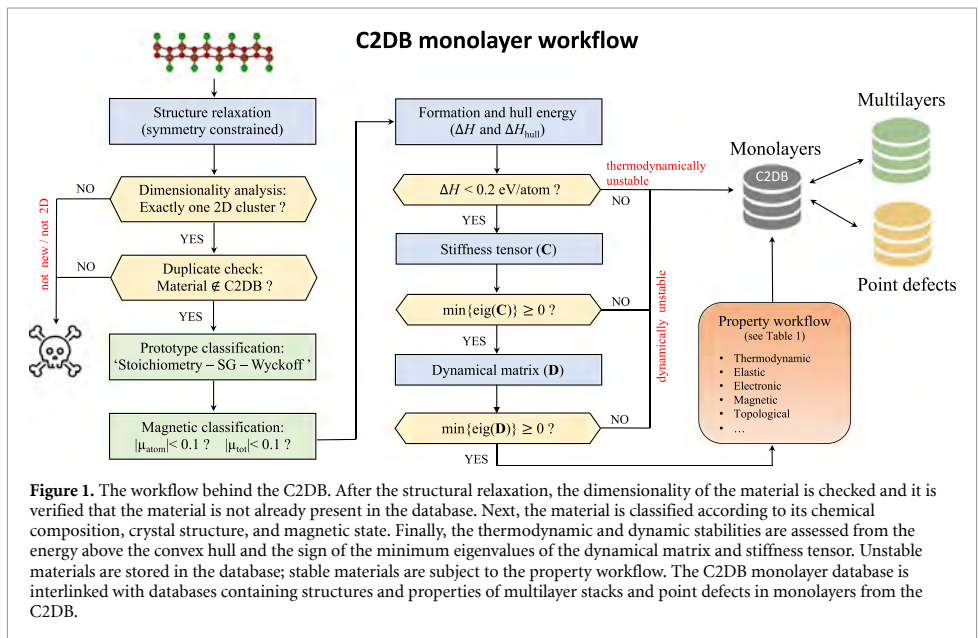
Figure 1. The workflow behind the C2DB. After the structural relaxation, the dimensionality of the material is checked and it is verified that the material is not already present in the database. Next, the material is classified according to its chemical composition, crystal structure, and magnetic state. Finally, the thermodynamic and dynamic stabilities are assessed from the energy above the convex hull and the sign of the minimum eigenvalues of the dynamical matrix and stiffness tensor. Unstable materials are stored in the database; stable materials are subject to the property workflow. The C2DB monolayer database is interlinked with databases containing structures and properties of multilayer stacks and point defects in monolayers from the C2DB.

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. $d < k(r_i^{cov} + r_j^{cov})$, 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

Before relaxation   After relaxation

$N_{3D} = 0, N_{2D} = 1$ $N_{3D} = 0, N_{2D} = 1$
$N_{1D} = 0, N_{0D} = 0$ $N_{1D} = 0, N_{0D} = 0$

$N_{3D} = 0, N_{2D} = 1$ $N_{3D} = 0, N_{2D} = 2$
$N_{1D} = 0, N_{0D} = 0$ $N_{1D} = 0, N_{0D} = 0$

$N_{3D} = 0, N_{2D} = 1$ $N_{3D} = 0, N_{2D} = 1$
$N_{1D} = 0, N_{0D} = 0$ $N_{1D} = 0, N_{0D} = 2$

Figure 2. Three example structures from C2DB (top: graphene, middle: Ge_2Se_2, bottom: Pb_2O_6) with their respective cluster dimensionalities cluster before (left) and after (right) relaxation. The number N_{xD} denotes the number of clusters of dimensionality x. Note that the number of atoms of the structures depicted in the left and right columns can differ because the relaxation can change the lattice constants.

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 *nonmagnetic* or *magnetic*. A material is considered magnetic if any atom has a local magnetic moment greater than 0.1 μ_B.

In the original C2DB, the *magnetic* category was further subdivided into ferromagnetic (FM) and antiferromagnetic (AFM). But since the simplest antiferromagnetically 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 $\Delta H_{hull} < 0.2$ eV atom^{-1}, defining the most stable category, was established based on an extensive analysis of 55 experimentally realised monolayer crystals [6].
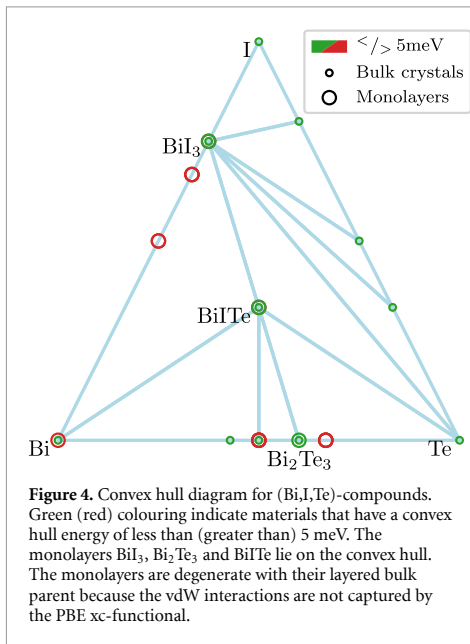
Figure 4. Convex hull diagram for (Bi,I,Te)-compounds. Green (red) colouring indicate materials that have a convex hull energy of less than (greater than) 5 meV. The monolayers BiI_3, Bi_2Te_3 and BiITe lie on the convex hull. The monolayers are degenerate with their layered bulk parent because the vdW 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 $\Delta H_{hull} > 0.2$
High	$\Delta H < 0.2$ and $\Delta H_{hull} < 0.2$

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:

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

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 polynomium 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 Å, 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.
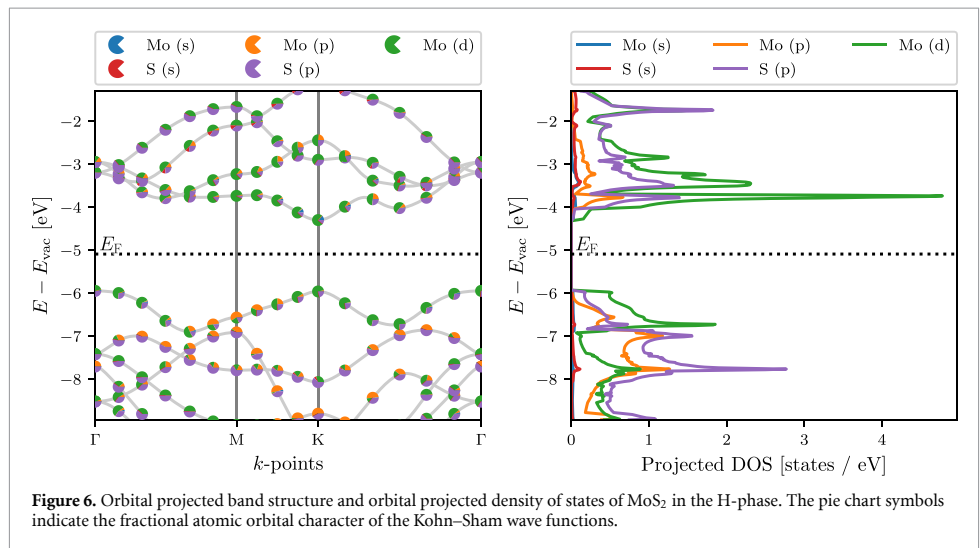
The second improvement is the calculation of the mean absolute relative error (MARE) of the polynomial fit in a 25 meV range from the BE. The value of 25 meV corresponds to the thermal energy at room temperature and is thus the relevant energy scale for many applications. To make the MARE independent 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

Figure 6. Orbital projected band structure and orbital projected density of states of MoS$_2$ in the H-phase. The pie chart symbols indicate the fractional atomic orbital character of the Kohn–Sham wave functions.

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 MoS$_2$ 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$_0$W$_0$ band structures

The C2DB contains G$_0$W$_0$ 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$_0$W$_0$ data points making up the QP band structures led to several important conclusions relevant for high-throughput G$_0$W$_0$ calculations. In particular, it identified the linear QP approximation as a significant error source in standard G$_0$W$_0$ 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 MoS$_2$ is shown in figure 7. The original uncorrected G$_0$W$_0$ band structure is shown in blue while the empZ corrected band structure is shown in orange. MoS$_2$ 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$_0$W$_0$ 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$_0$W$_0$ 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]

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

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
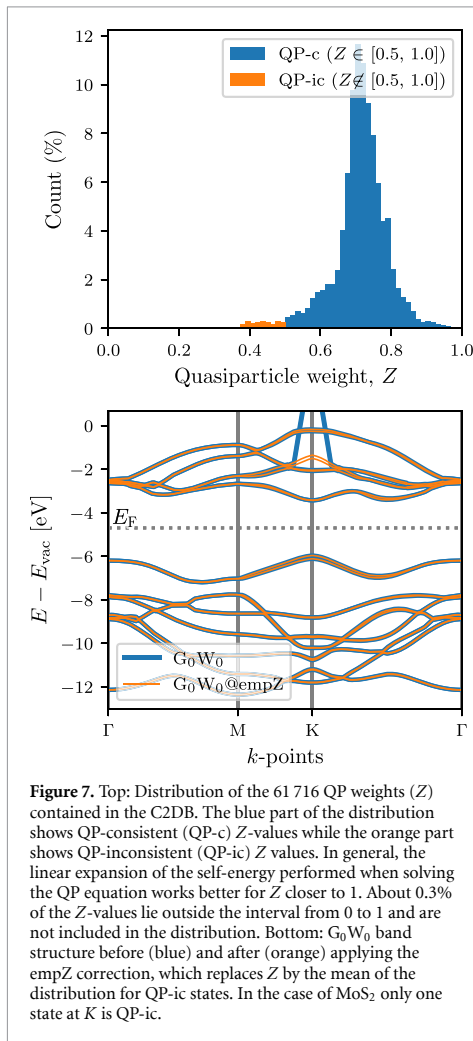
Figure 7. Top: Distribution of the 61 716 QP weights (Z) contained in the C2DB. The blue part of the distribution shows QP-consistent (QP-c) Z-values while the orange part shows QP-inconsistent (QP-ic) Z values. In general, the linear expansion of the self-energy performed when solving the QP equation 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:

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

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 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.
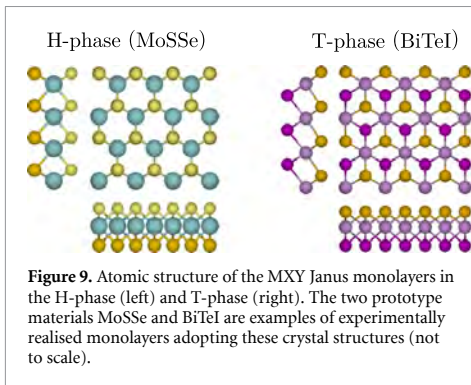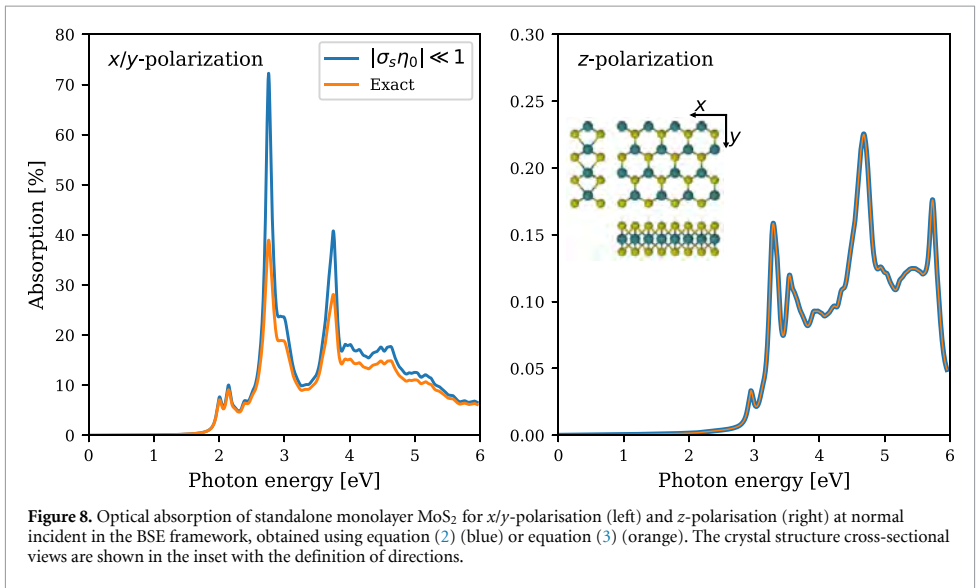
4. New materials in the C2DB

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

Figure 8. Optical absorption of standalone monolayer MoS_2 for x/y-polarisation (left) and z-polarisation (right) at normal incident in the BSE framework, obtained using equation (2) (blue) or equation (3) (orange). The crystal structure cross-sectional views are shown in the inset with the definition of directions.

Figure 9. Atomic structure of the MXY Janus monolayers in the H-phase (left) and T-phase (right). The two prototype materials MoSSe and BiTeI are examples of experimentally realised monolayers adopting these crystal structures (not to scale).

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 $\Delta H_{hull} < 0.2 \, \mathrm{eV \, atom^{-1}}$). 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]:

$$H = \alpha_R(\boldsymbol{\sigma} \times \mathbf{k}) \cdot \hat{\mathbf{e}}_z, \qquad (4)$$
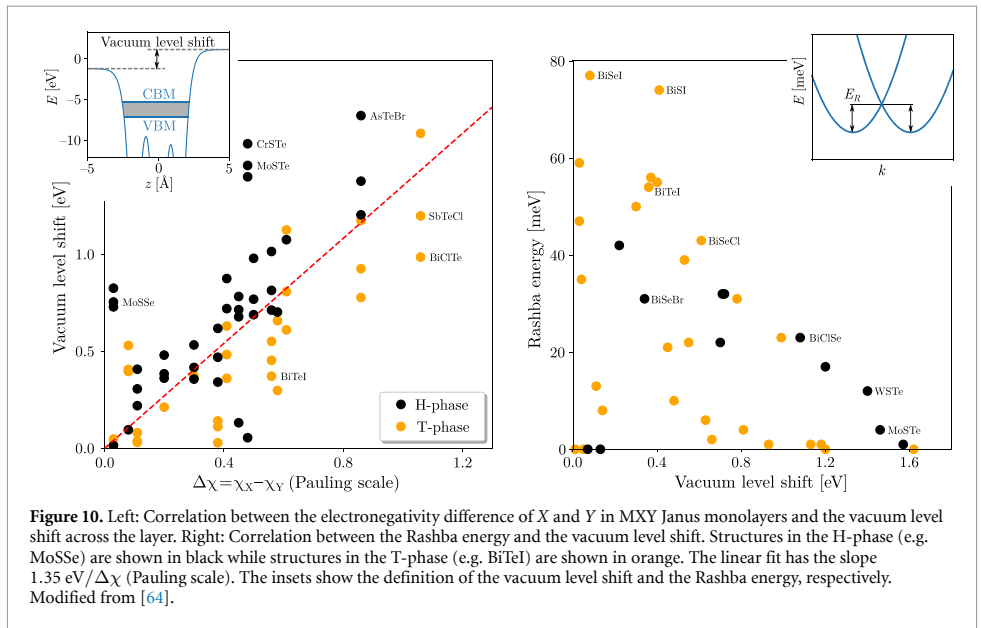
where $\boldsymbol{\sigma}$ is the vector of Pauli matrices, $\mathbf{k} = \mathbf{p}/\hbar$ is the wave number, and the Rashba parameter is proportional to the electric field strength, $\alpha_R \propto E_0$.

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 $E_0 = \Delta V_{vac}/d$, 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]

Figure 10. Left: Correlation between the electronegativity difference of X and Y in MXY Janus monolayers and the vacuum level shift across the layer. Right: Correlation between the Rashba energy and the vacuum level shift. Structures in the H-phase (e.g. MoSSe) are shown in black while structures in the T-phase (e.g. BiTeI) are shown in orange. The linear fit has the slope 1.35 eV/$\Delta\chi$ (Pauling scale). The insets show the definition of the vacuum level shift and the Rashba energy, respectively. Modified from [64].

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.
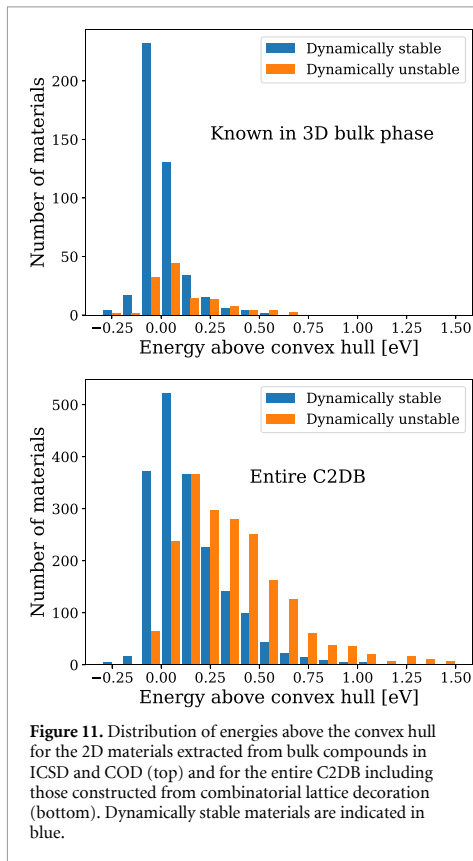
The search has been limited to bulk compounds containing less than six different elements and no rare earth elements. This reduces the set of relevant bulk materials to 2991. For all of these we extracted the 2D components containing less than 21 atoms in the unit cell, which were then relaxed and sorted 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 MoS_2 exists in van der Waals bonded structures composed of either 2 H-MoS_2 or 1 T-MoS_2 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.

Figure 12. An illustration of the optimisation of the interlayer (IL) distance for MoS_2 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 MoS_2 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

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 MoS_2 [71, 72], layer-parity selective Berry curvatures in few-layer WTe_2 [73], thickness-dependent magnetic order in CrI_3 [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.
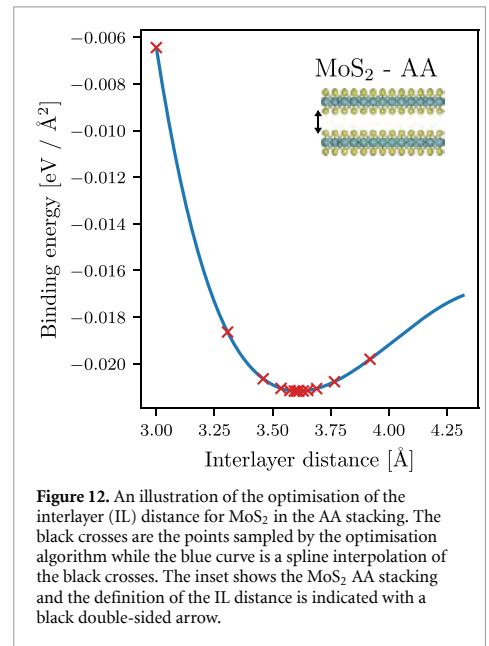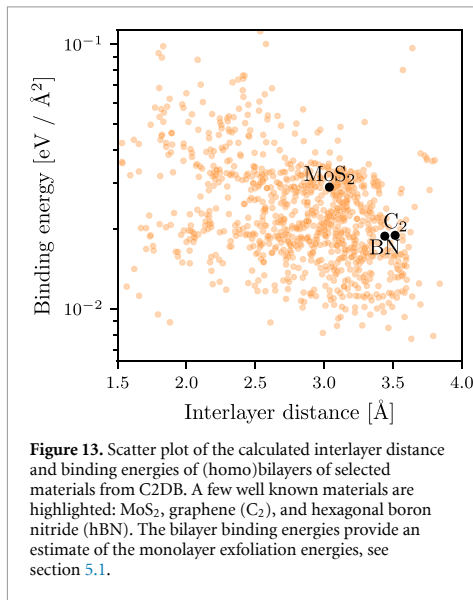
The calculated IL binding energies are generally in the range from a few to a hundred meV $Å^{-2}$ and IL distances range from 1.5 to 3.8 Å. A scatter plot of preliminary binding energies and IL distances is shown in figure 13. The analysis of homobilayers provides an estimate of the energy required 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

Figure 13. Scatter plot of the calculated interlayer distance and binding energies of (homo)bilayers of selected materials from C2DB. A few well known materials are highlighted: MoS_2, graphene (C_2), and hexagonal boron nitride (hBN). The bilayer binding energies provide an estimate of the monolayer exfoliation energies, see section 5.1.

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 $E_{gap} > 1$ 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 CH_2Si.

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 CH_2Si. 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 CH_2Si is shown in figure 14 (c), where the defect states are labelled according to the irreducible representations of the C_s point group.
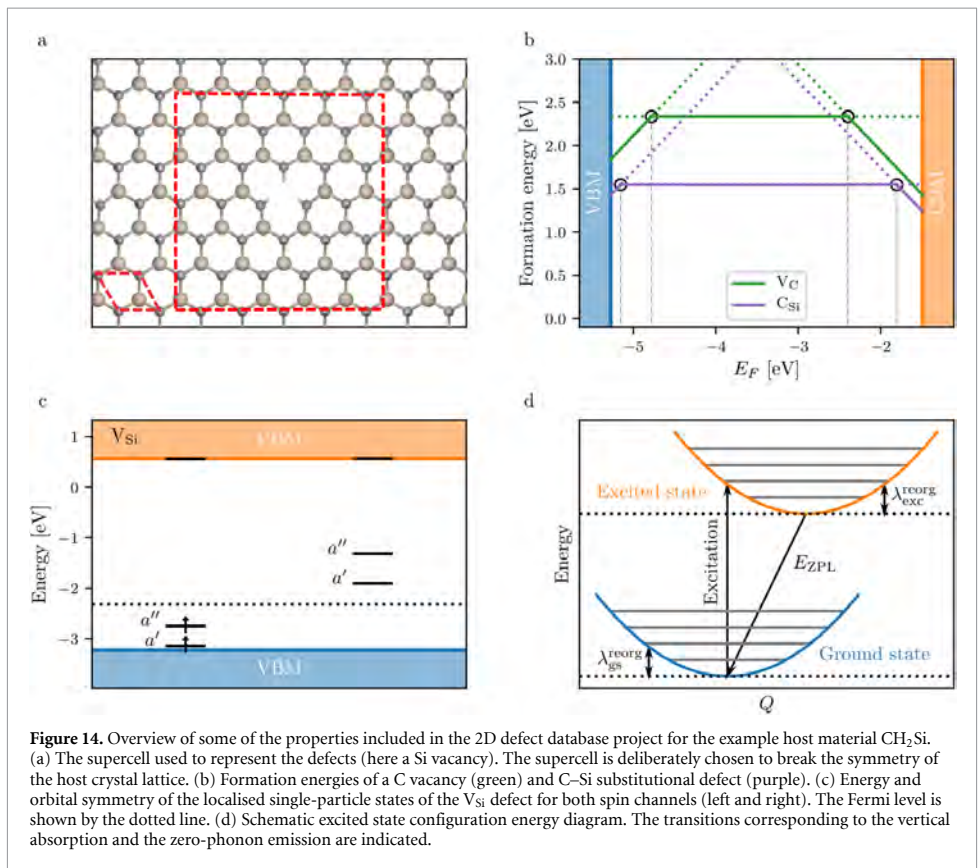
In general, excited electronic states can be modelled by solving the Kohn–Sham equations with non-Aufbau occupations. The excited-state solutions are saddle points of the Kohn–Sham energy functional, but common self-consistent field (SCF) approaches often struggle to find such solutions, especially when 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 CH_2Si, the DO-MOM method yields $E_{ZPL} = 3.84$ eV, $\lambda_{gs}^{reorg} = 0.11$ eV and $\lambda_{exc}^{reorg} = 0.16$ 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

Figure 14. Overview of some of the properties included in the 2D defect database project for the example host material CH_2Si. (a) The supercell used to represent the defects (here a Si vacancy). The supercell is deliberately chosen to break the symmetry of the host crystal lattice. (b) Formation energies of a C vacancy (green) and C–Si substitutional defect (purple). (c) Energy and orbital symmetry of the localised single-particle states of the V_{Si} defect for both spin channels (left and right). The Fermi level is shown by the dotted line. (d) Schematic excited state configuration energy diagram. The transitions corresponding to the vertical absorption and the zero-phonon emission are indicated.

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:

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

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:

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

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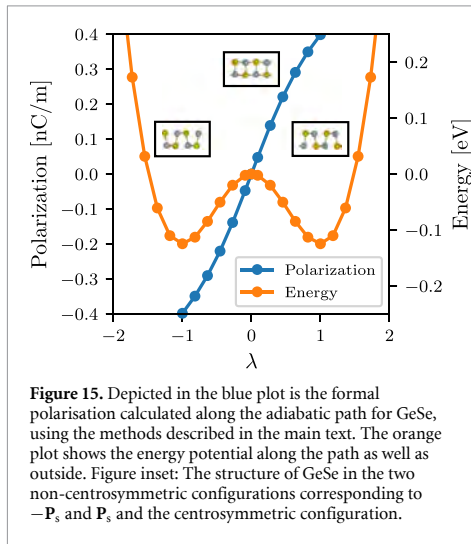
Figure 15. Depicted in the blue plot is the formal polarisation calculated along the adiabatic path for GeSe, using the methods described in the main text. The orange plot shows the energy potential along the path as well as outside. Figure inset: The structure of GeSe in the two non-centrosymmetric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

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

where $\mathrm{BZ}_{\perp \mathbf{b}_l} = \{\mathbf{k}|\mathbf{k}\cdot\mathbf{b}_l = 0\}$ is a plane of $\mathbf{k}$-points orthogonal to $\mathbf{b}_l$, $\delta\mathbf{k}$ is the distance between neighbouring k-points in the $\mathbf{b}_l$ direction and $N_{k\|\mathbf{b}_l}$ ($N_{k\perp\mathbf{b}_l}$) is the number of $\mathbf{k}$-points along (perpendicular to) the $\mathbf{b}_l$ direction. These expression 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:

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

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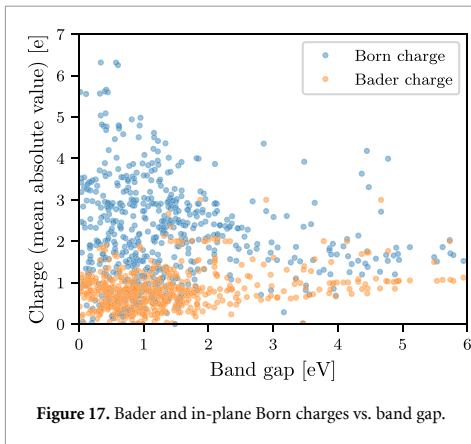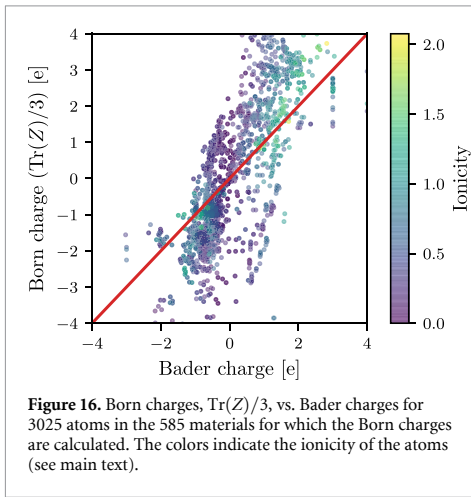
$$
Z_{ij}^a = \frac{V}{e} \frac{\partial P_i}{\partial u_{aj}} \bigg|_{E=0}.
$$
$$(9)$$

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 nonintuitive 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, $\mathrm{Tr}(Z^a)/3$, 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

Figure 16. Born charges, $\mathrm{Tr}(Z)/3$, vs. Bader charges for 3025 atoms in the 585 materials for which the Born charges are calculated. The colors indicate the ionicity of the atoms (see main text).

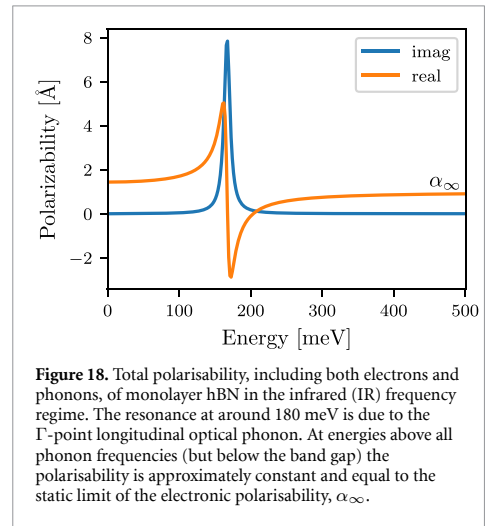**Figure 17.** Bader and in-plane Born charges vs. band gap.

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:

Figure 18. Total polarisability, including both electrons and phonons, of monolayer hBN in the infrared (IR) frequency regime. The resonance at around 180 meV is due to the Γ-point longitudinal optical phonon. At energies above all phonon frequencies (but below the band gap) the polarisability is approximately constant and equal to the static limit of the electronic polarisability, α_∞.

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

$$c_{ijk} = \frac{e}{2\pi V} \sum_l \frac{\partial \phi_l}{\partial \epsilon_{jk}} a_{li},$$
(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 MoS_2 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 MoS_2 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: $OsCl_3$ (space group 147)—a Chern insulator with $C = 1$, $OsTe_2$ (space group 14)—a mirror crystalline insulator with $C_M = 2$, SbI (spacegroup 1)—a quantum spin Hall insulator with $\nu = 1$ and BiITe (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.
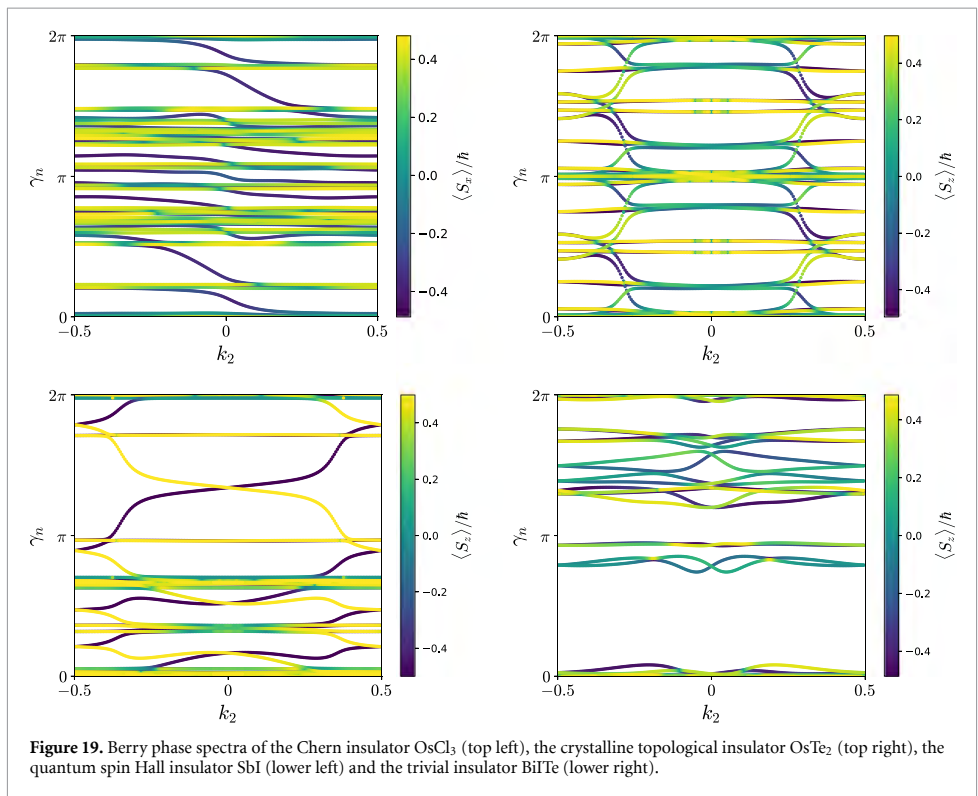
Figure 19. Berry phase spectra of the Chern insulator OsCl$_3$ (top left), the crystalline topological insulator OsTe$_2$ (top right), the quantum spin Hall insulator SbI (lower left) and the trivial insulator BiITe (lower right).

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:

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

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 postprocessing 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 $A(2S-1) + \lambda N_{nn} > 0$ [113]. Neglecting anisotropic exchange thus excludes materials with $A < 0$ that satisfies $A(2S-1) + \lambda N_{nn} > 0$. 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 $\mathcal{F}(t) = \mathcal{F}_{in}\mathbf{u}_{in}\exp(-i\omega_{in}t) + \mathcal{F}_{out}\mathbf{u}_{out}\exp(-i\omega_{out}t) + c.c.$ where $\mathcal{F}_{in/out}$ and $\omega_{in/out}$ denote the amplitudes and frequencies of the input/output electromagnetic fields, respectively. In addition, $\mathbf{u}_{in/out} = \sum_i u_{in/out}^i \mathbf{e}_i$ are the corresponding polarisation vectors, where $\mathbf{e}_i$ denotes the unit
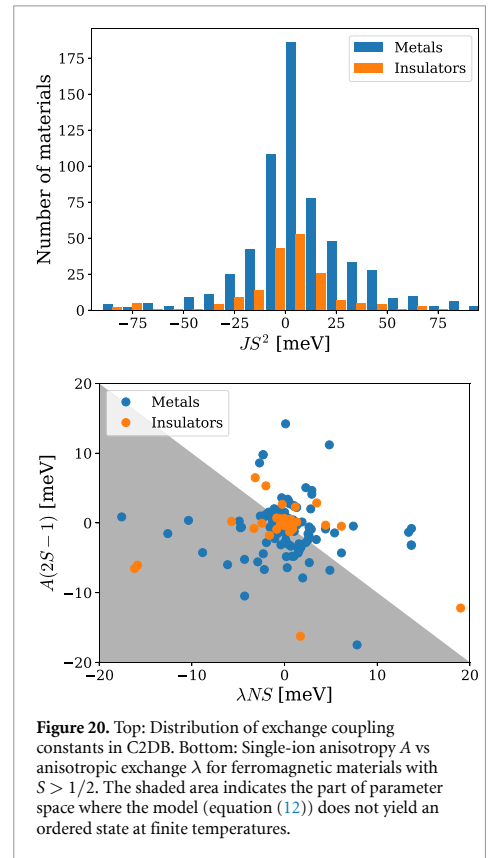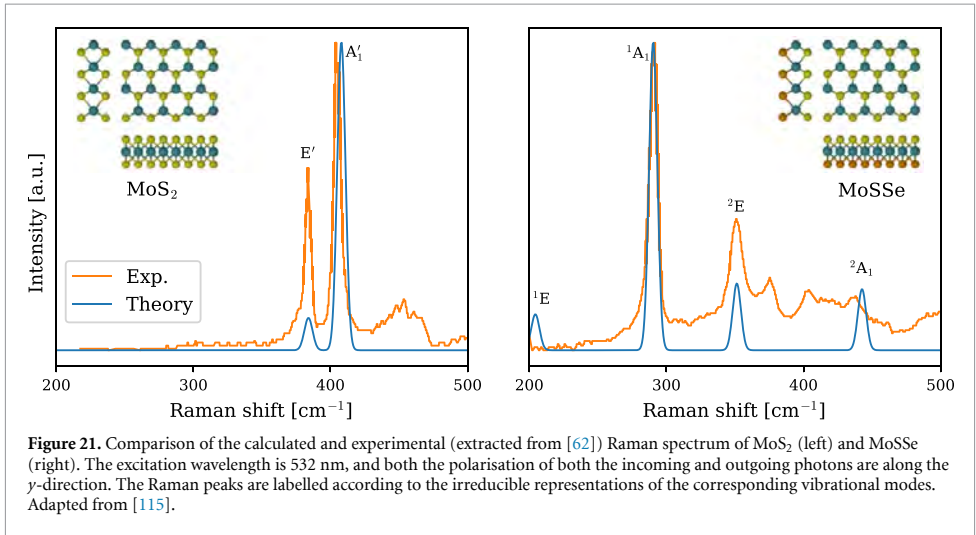

Figure 20. Top: Distribution of exchange coupling constants in C2DB. Bottom: Single-ion anisotropy A vs anisotropic exchange λ for ferromagnetic materials with $S > 1/2$. The shaded area indicates the part of parameter space where the model (equation (12)) does not yield an ordered state at finite temperatures.

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]:

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

(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

Figure 21. Comparison of the calculated and experimental (extracted from [62]) Raman spectrum of MoS₂ (left) and MoSSe (right). The excitation wavelength is 532 nm, and both the polarisation of both the incoming and outgoing photons are along the *y*-direction. The Raman peaks are labelled according to the irreducible representations of the corresponding vibrational modes. Adapted from [115].

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 $\sigma = 3$ 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]:

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

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
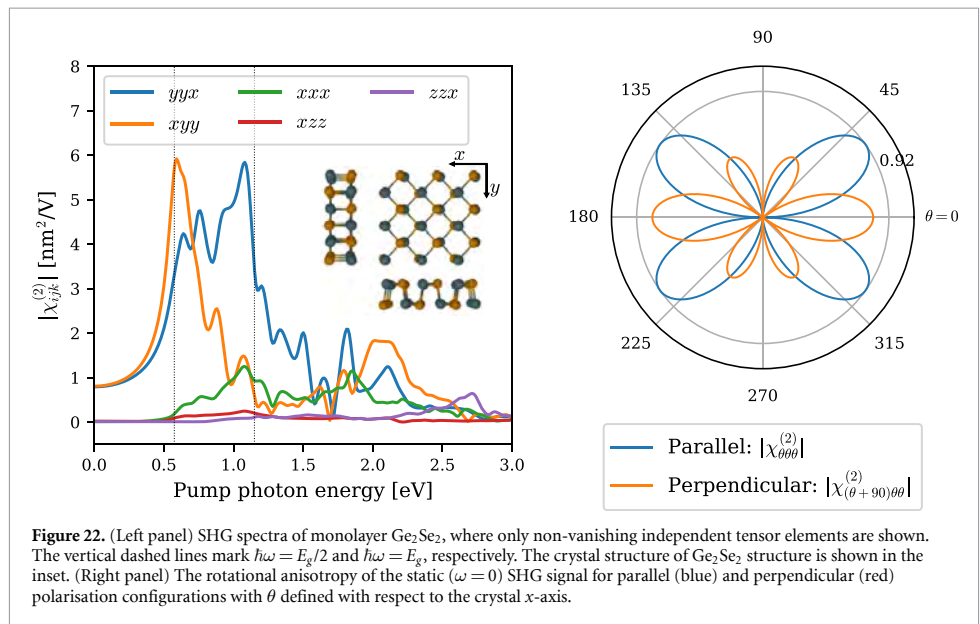
Figure 22. (Left panel) SHG spectra of monolayer Ge_2Se_2, where only non-vanishing independent tensor elements are shown. The vertical dashed lines mark $\hbar\omega = E_g/2$ and $\hbar\omega = E_g$, respectively. The crystal structure of Ge_2Se_2 structure is shown in the inset. (Right panel) The rotational 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 $\mathcal{F}(t) = \sum_i \mathcal{F}_i \mathbf{e}_i e^{-i\omega t}$+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 $\chi^{(2)}_{ijk}$,

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

where ε_0 denotes the vacuum permittivity. $\chi^{(2)}_{ijk}$ is a symmetric (due to intrinsic permutation symmetry i.e. $\chi^{(2)}_{ijk} = \chi^{(2)}_{ijk}$) rank-3 tensor with at most 18 independent elements. Furthermore, similar to the piezo-electric 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, $\chi^{(2)}_{xxx}$, $\chi^{(2)}_{xyy}$, $\chi^{(2)}_{xzz}$, $\chi^{(2)}_{yyx} = \chi^{(2)}_{yxy}$, and $\chi^{(2)}_{zzx} = \chi^{(2)}_{zxz}$, 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 m V^{-1}) is ill-defined for 2D materials (since the volume is ambiguous) [120]. Instead, the unambiguous *sheet* quadratic susceptibility (with SI units of m^2 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

Figure 23. Scatter plot (double log scale) of the static sheet quadratic susceptibility $|\chi_{ijk}^{(2)}|$ versus the static sheet linear susceptibility $|\chi_{ij}^{(1)}|$ for 67 TMDCs (with chemical formula MX_2 and space group 187). A few well known materials are highlighted.
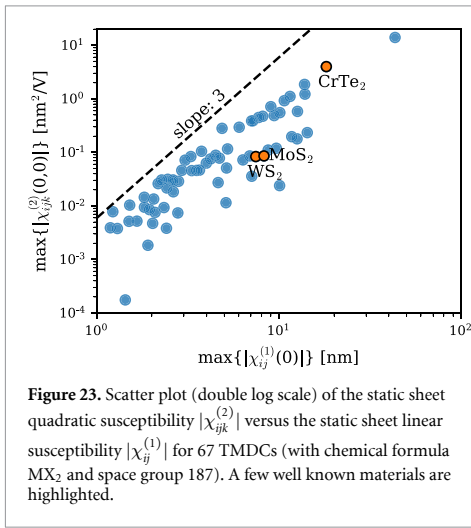
the linear and quadratic optical responses are functions of the transition dipole moments and transition energies. More interestingly, the strength of the quadratic response seems to a very good approximation to be given by a universal constant times the linear susceptibility to the power of three (ignoring polarisation indices), i.e.

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

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:
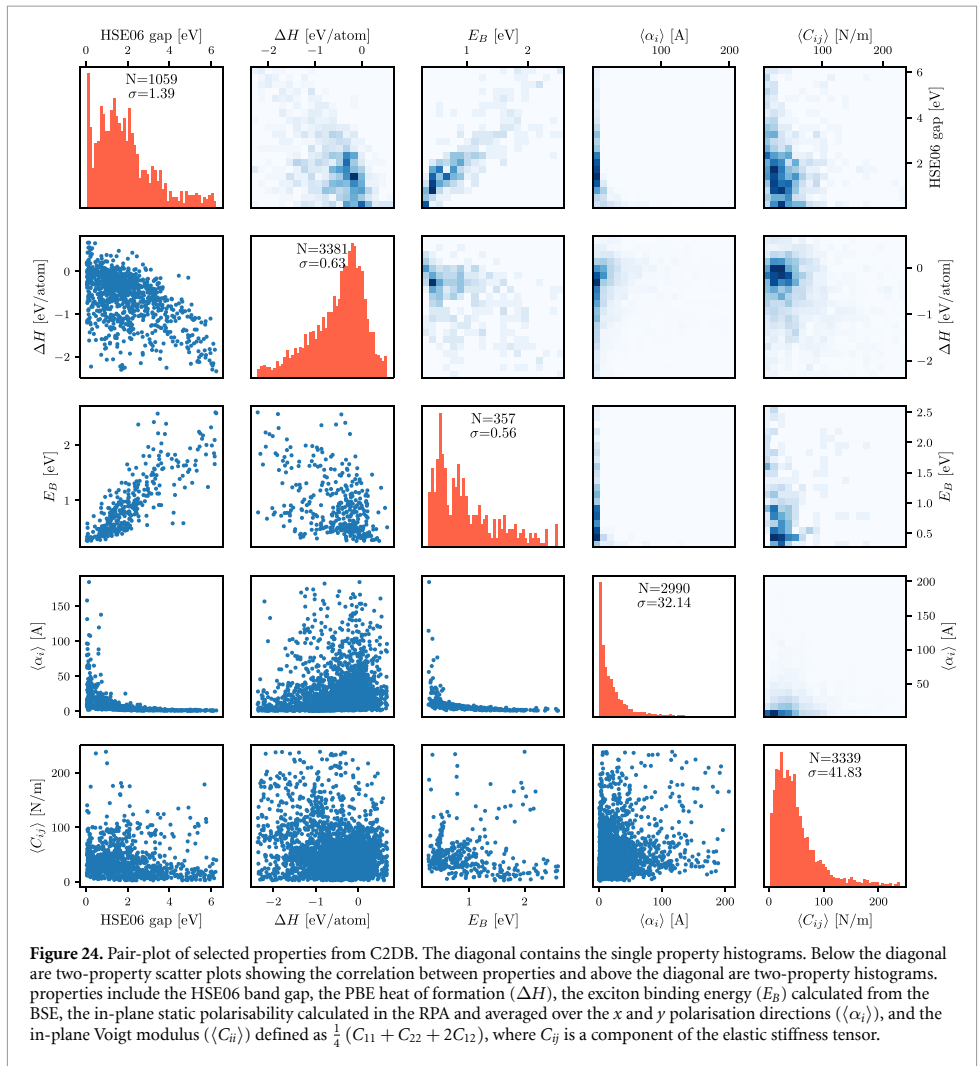
(a) *Complete:* The fingerprint should incorporate all the relevant input for the underlying problem, i.e. materials with different properties should have different fingerprints.

(b) *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.

(c) *Descriptive:* Materials with similar target values should have similar fingerprints.

(d) *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 build 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

Figure 24. Pair-plot of selected properties from C2DB. The diagonal contains the single property histograms. Below the diagonal are two-property scatter plots showing the correlation between properties and above the diagonal are two-property histograms. properties include the HSE06 band gap, the PBE heat of formation (ΔH), the exciton binding energy (E_B) calculated from the BSE, the in-plane static polarisability calculated in the RPA and averaged over the x and y polarisation directions ($\langle \alpha_i \rangle$), and the in-plane Voigt modulus ($\langle C_{ii} \rangle$) defined as $\frac{1}{4}\left(C_{11} + C_{22} + 2C_{12}\right)$, where C_{ij} is a component of the elastic stiffness tensor.

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 fingerprint[6] encodes the coupling between the PDOS at different atomic orbitals in both energy and real space. It is defined as:

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

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

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

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,

---

[6] Details will be published elsewhere.

Figure 25. Prediction scores (MAE normalised to standard deviation of property values) for the test sets of selected properties using a Gaussian process regression.

Figure 26. ML predicted HSE06 gap values vs. true values for Ewald, MBTR and PDOS fingerprints with MAE's for train and test set included. The PDOS is found to perform significantly better for the prediction of HSE06 gap.

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.

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] Haastrup 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] Bistritzer 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] Sohier 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, Klintenberg 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] Northup 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

9.5 Publication V: Atomic Simulation Recipes: A Python framework and library for automated workflows

M. GJERDING
T. SKOVHUS
A. RASMUSSEN
F. BERTOLDO
A. H. LARSEN
J. J. MORTENSEN
K. S. THYGESEN

Atomic Simulation Recipes – A Python framework and library for automated workflows

Morten Gjerding *, Thorbjørn Skovhus , Asbjørn Rasmussen , Fabian Bertoldo , Ask Hjorth Larsen , Jens Jørgen Mortensen , Kristian Sommer Thygesen

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

A B S T R A C T

The Atomic Simulation Recipes (ASR) is an open source Python framework for working with atomistic materials simulations in an efficient and sustainable way that is ideally suited for high-throughput projects. Central to ASR is the concept of a Recipe: a high-level Python script that performs a well defined simulation task robustly and accurately while keeping track of the data provenance. The ASR leverages the functionality of the Atomic Simulation Environment (ASE) to interface with external simulation codes and attain a high abstraction level. We provide a library of Recipes for common simulation tasks employing density functional theory and many-body perturbation schemes. These Recipes utilize the GPAW electronic structure code, but may be adapted to other simulation codes with an ASE interface. Being independent objects with automatic data provenance control, Recipes can be freely combined through Python scripting giving maximal freedom for users to build advanced workflows. ASR also implements a command line interface that can be used to run Recipes and inspect results. The ASR Migration module helps users maintain their data while the Database and App modules makes it possible to create local databases and present them as customized web pages.

1. Introduction

As computing power continues to increase and the era of exascale approaches, the development of software solutions capable of exploiting the immense computational resources becomes a key challenge for the scientific community. In the field of materials science, ab initio electronic structure (aiES) calculations are increasingly being conducted in a high-throughput fashion to screen thousands of materials for various applications [1–16] and to generate large reference data sets for training machine learning algorithms to predict fundamental materials properties [17–22] or design interatomic potentials [23–26]. The results from such aiES high-throughput calculations are often stored in open databases allowing the data to be efficiently shared and deployed beyond the original purpose [27–37].

While a few thousands of calculations can be managed manually, a paradigm in which data drives scientific discovery calls for dedicated workflow solutions that automatically submit and retrieve the calculations, store the results in organized data structures, and keep track of the origin, history and dependencies of all data, i.e. the data provenance. Ideally, the workflow should also attach explanatory descriptions to the data that allows them to be easily accessed, understood, and deployed – also by users with limited domain knowledge.

Materials scientists from the aiES community are employing a large and heterogeneous set of simulation codes based mainly on density functional theory (DFT) [38]. These codes differ substantially in the way they implement and solve the fundamental physical equations. This is due to the fact that different types of problems require different numerical approaches, e.g. high accuracy vs. large system sizes, periodic vs. finite vs. open boundary conditions, or ground state vs. excited state properties. In principle, the large pool of available aiES codes provides users with a great deal of flexibility and freedom to pick the code that best suits the problem at hand. In practice, however, the varying numerical implementations and the diverse and often rudimentary user interfaces make it challenging for users to switch between the different aiES codes leading to a significant "code barrier".

To some extent, a similar situation exists with respect to materials properties. Although aiES codes provide access to a rich variety of physical and chemical properties, individual researchers often focus on properties within a specific scientific domain. While this may be sufficient in many cases, several important contemporary problems

addressed by the aiES community are multi-physical in nature and require properties and insights from several domains. For example, evaluating the potential of a material as a photocatalyst involves an assessment of solar light absorption, charge transport, and chemical reactions at a solid–liquid interface. Calculating new types of properties for the first time is often a time-consuming process involving trial and error and the acquisition of technical, implementation-specific knowledge of no direct benefit for the user or the overall project aim. This situation may result in a "property barrier" that hampers researchers' exploitation of the full capacity of aiES codes.

In this paper, we introduce The Atomic Simulation Recipes (ASR) – a highly flexible Python framework for developing and working with computational materials workflows. The ASR reduces code and property barriers and makes it easy to perform high-throughput computations with advanced workflows while adhering to the FAIR Data Principles [39]. There are already some workflow solutions available in the field, some of the most prominent being AFlow [30], Fireworks [40], AiiDA [41], and Atomate [42]. However, these are either designed for one specific simulation code and/or constitute rather colossal integrated entities, the complexity of which could represent an entry barrier to some users. The ASR differs from the existing solutions in several important ways, and we expect it to appeal to a large crowd of computational researchers, e.g. those with Python experience who like to develop their own personalized (workflow) scripts and databases, less experienced users who prefer plug-and-play solutions, and those who wish to apply non-standard methodologies, e.g. compute GW band structures or Raman spectra, but feel they lack the expertise required for using standard low-level codes.

The basic philosophy of ASR is to prioritize usability and simplicity over system perfection. More specifically, ASR is characterized by the following qualities:

- **Flexibility:** The Python scripting interface and high degree of modularity provide users with almost unlimited freedom for developing and deploying workflows.
- **Modularity:** The key components of ASR, namely the workflow development framework (ASR core), the Database and App modules, the task scheduler (MyQueue), and the simulation codes, are separate independent entities. Moreover, the Recipe library concept supports modular workflow designs and reuse of code.
- **Data locality:** Generated data is stored in a special folder named `.asr` where it can be accessed transparently via command line tools (similar to Git).
- **Compatibility:** For compatibility with external simulation codes, the ASR core is fully simulation code-independent while specific Recipe implementations communicate with simulation codes via the abstract ASE Calculator interface.
- **Minimalism and pragmatism:** ASR is based on simple solutions that work efficiently in practice. This makes ASR fast to learn, easy to use, and relatively uncomplicated to adapt to future demands.

At the core of ASR is the concept of a *Recipe*. In essence, a Recipe is a piece of code that can perform a certain simulation task (e.g. relax an atomic structure, calculate a Raman spectrum, or identify covalently bonded components of a material) while recording all relevant results and metadata. The use of Recipes makes it simple to run simulations from either Python or the command line. For example,

```
$ asr run "asr.bandstructure –atoms structure.json"
```

will calculate the electronic band structure of the material `structure.json`. Subsequently, the command

```
$ asr results asr.bandstructure
```

will produce a plot of the band structure. With two additional commands, the ASR results can be inspected in a web browser, see example in Fig. 6.

In practice, Recipes are implemented as Python modules building on the Atomic Simulation Environment (ASE) [43]. Recipes conform to certain naming and structured programming conventions, making them largely self-documenting and easy to read. To keep track of data provenance, Recipes utilize a caching mechanism that automatically logs all exchange of data with the user and other Recipes in a uniquely identifiable `Record` object. Not only does this guarantee the documentation and reproducibility of the results, it also allows ASR to determine whether a given Recipe task has already been performed (such that its result can be directly loaded and returned) and to detect if a Recipe task needs to be rerun because another piece of data in its dependency chain has changed. In addition, Recipes implement presentation and explanatory descriptions of their outputs and may also define a web panel for online presentation.

The Recipes of the current ASR library cover a variety of computational tasks and properties (see Table 1). Most of the 40+ available Recipes utilize DFT. However, some Recipes do not involve calls to a simulation code (e.g. symmetry analysis or construction of phase diagrams) while others employ beyond-DFT methodology (e.g. the GW method or the Bethe–Salpeter equation). These library Recipes can be used "out of the box" or modified to fit the user's need. New Recipes may be developed straightforwardly following the documentation and large body of available examples. Recipes can be combined into complex workflows using Python scripting for maximal flexibility and compatibility with ASE and other relevant Python libraries like PymatGen [44], Spglib [45] and Phonopy [46]. The Python workflows may be executed on supercomputers using the MyQueue [47] task scheduler front-end or other similar systems.

The ASR contains a number of tools for working with the ASE database module, which makes it easy to generate and maintain local materials databases. Relying on the Recipes' web panel implementations, these databases may be straightforwardly presented in a browser allowing for easy inspection, querying, and sharing of results on a local or public network. As an example of an ASR-driven database project we refer to the Computational 2D Materials Database (C2DB) [32,48].[1]

While the core of ASR, i.e. the Recipe concept and caching system, is fully simulation code-independent, most Recipe implementations of the current library contain calls to the specific aiES code GPAW [49]. At the moment, among the Recipes calling an external simulation code, only the asr.relax and asr.stiffness Recipes generalize to all ASE calculators, which support the calculation of stresses and atomic forces. We are currently working on a generalization of the ASE Calculator interface which will decouple Recipe implementations from simulation codes. In the future, many Recipes will therefore work with multiple simulation codes.

Another on-going effort is to generalize the organization of calculated results. For example results are currently presented mainly by material. This is practical for a database which primarily associates a number of properties with each material, but not for presenting sets of results parametrized over other variables than the material. These limitations will be removed over the next releases.

The rest of this paper is organised as follows: In Section 2 we provide a general overview of the main components of ASR. Section 3 zooms in on the central Recipe concept and its caching system while Section 4 gives an overview of the currently available Recipes. In Section 5, the Database and App modules are described. Section 6 gives a brief presentation of the Computational 2D Materials Database as an example of an ASR-driven high-throughput database project and provides a few concrete examples of Recipe implementations. Section 7 describes the different user interfaces supported by ASR while Sections 8 and 9 explain how ASR manages data migration and provenance, respectively. Sections 10 and 11 cover documentation and technical specifications.

---

[1] http://c2db.fysik.dtu.dk.

Table 1

List of Recipes currently implemented in the ASR library. Most of the Recipes depend explicitly on the GPAW electronic structure code. The Recipes are grouped under thematic headings and listed in alphabetic order.

Recipe name	Description
Atomic structure	
asr.database.duplicates	Remove duplicate structures from a database
asr.database.rmsd	Root mean square distance between structures
asr.dimensionality	Dimensionality of covalently bonded substructures of a material
asr.push	Push atoms along specific phonon mode
asr.relax	Relax atomic structure
asr.setup.defects	Generate native point defects
asr.setup.displacements	Generate structures with a single displaced atom
asr.setup.magnetize	Initialize atomic magnetic moments
asr.setup.reduce	Reduce supercell to primitive cell
asr.setup.symmetrize	Symmetrize an atomic structure
asr.structureinfo	Extract structural information
Thermodynamic properties	
asr.chc	Constrained convex hull stability analysis
asr.convex_hull	Convex hull stability analysis
asr.defectformation	Formation energy of neutral point defect
asr.fere	Fitted elemental reference energies
Mechanical properties	
asr.phonopy	Phonon band structure and dynamical stability
asr.piezoelectrictensor	Piezoelectric tensor
asr.stiffness	Stiffness tensor
Electronic properties	
asr.bader	Bader charge analysis
asr.bandstructure	Kohn–Sham band structure
asr.berry	Various band topology invariants
asr.borncharges	Born effective charge tensor
asr.deformationpotentials	Deformation potentials (only for 2D)
asr.dos	Density of states
asr.emasses	Effective masses
asr.fermisurface	Fermi surface
asr.formalpolarization	Formal polarization phase
asr.gs	Electronic ground state
asr.gw	G_0W_0 quasiparticle band structure
asr.hse	HSE06 band structure
asr.pdos	Orbital projected density of states
asr. projected_bandstructure	Orbital projected Kohn–Sham band structure
Magnetic properties	
asr.exchange	Magnetic exchange coupling
asr.magnetic_anisotropy	Magnetic anisotropy
asr.magstate	Determine magnetic state
Optical properties	
asr.bse	Optical absorption from Bethe–Salpeter Equation (BSE)
asr.infraredpolarizability	Infrared polarizability (caused by vibrations)
asr.plasmafrequency	Plasma frequency (from intraband transitions)
asr.polarizability	Optical polarizability (caused by electrons)
asr.raman	Raman spectrum (first-order)
asr.shg	Second harmonics generation
asr.shift	Shift current

Finally, Section 12 summarises the paper and presents our future perspectives for ASR.

2. Overview of ASR

Fig. 1 shows a schematic overview of the main components of the ASR and their mutual dependencies. An arrow from X to Y indicates a direct dependence of Y on X, e.g. via function calls (Y calls X). The ASR modules have been divided into the ASR core modules (Cache and Recipe) and the ASR user interfaces (command-line interface, Python, Task scheduler front-end, and Apps). In addition, the ASR Database and Data migration modules contain tools for working with databases and maintaining data, respectively.

Recipes implement specific, well defined materials simulation tasks

as Python modules building on the ASE [43] and other Python libraries. A Recipe integrates with a Cache module that keeps track of performed tasks and manages all relevant metadata. The Cache also allows the user to inspect the data generated by a Recipe via the ASR command line interface (CLI) or using Python. Likewise, the Recipes may be executed directly from the CLI or called via Python scripts, the latter giving maximal flexibility and compatibility with existing Python libraries. For the purpose of high-throughput computations, advanced Python workflows combining several Recipes may be constructed and executed remotely using task scheduling systems like MyQueue [47].

The ASR Cache and Recipe modules work on a folder/file basis. This locality of data makes the ASR highly transparent for the user. The ASR Database module contains functions for converting the ASR data stored in a tree of folders into an ASE database and vice versa. The ASR App module generates web pages for online presentation, browsing and searching of the databases generated by the ASR Database module. Finally, the Data migration module provides tools for transforming data (results or metadata) to ensure backward compatibility when Recipes are updated.

3. What is a Recipe?

A Recipe is a Python module implementing the Instructions needed to obtain a particular result, for example to relax an atomic structure, calculate an electronic band structure or a piezoelectric tensor. This section describes the structure and main components of a Recipe. A schematic overview of the Recipe concept is shown in Fig. 2.

3.1. Instruction

An Instruction is to be understood as a Python function wrapped in a caching layer provided by ASR, see Fig. 3. Whenever an Instruction is called, the caching layer intercepts the input arguments and asks the cache whether the result of the particular Instruction call already exists (cache hit) or whether there are no matching results (cache miss). If a matching result exists (because it was calculated previously), the caching layer skips the actual evaluation of the Instruction and simply reads and returns the previously calculated result. In the case of a cache miss, the Instruction is evaluated, after which the result is intercepted by the caching layer and stored together with the relevant metadata in a Record object. The precise content of the Record object and the conditions for a cache hit/miss are described in Section 3.3.

One of the great benefits of this design is its simplicity. Because the Instruction/caching layer is implemented as a simple wrapper around a Python function, usage of the caching functionality requires minimal additional knowledge. In practice, this means that working with ASR and implementing new ASR Recipes becomes really simple.

The caching system works on a per-folder basis (similar to Git): a cache is initialized by the user in a folder and any instruction evaluated within this folder or sub-folders will utilize this cache. This mimics the behaviour of the MyQueue task scheduler so as to maximize the synergy between these tools. In practice, the "one-cache-per-folder" system works well together with a "one-material-per-folder" structure. The latter is currently still a requirement for utilizing the Database functionalities described in Section 5. However, the caching system can work with several atomic structures in the same folder as the cache can distinguish ASR tasks performed on different atomic structures. Data written by ASR is encoded as JSON.

Any Instruction can be called directly by the user (from Python or the CLI), but special importance is given to the "main Instruction". The main Instruction usually provides the primary interface for the user to the Recipe and returns the final result of the Recipe. Other Instructions are called by the main Instruction and evaluated as needed. These may be Instructions implemented in the Recipe itself but may also be Instructions of separate Recipes. The main Instruction takes all input arguments required by the Recipe and uses them to call other Instructions.
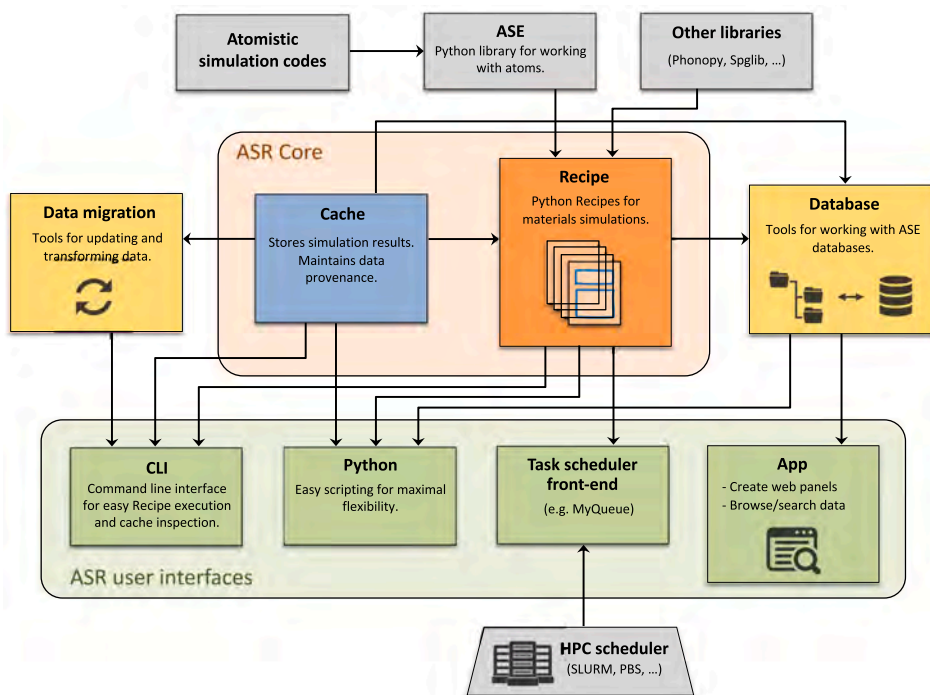
Fig. 1. Schematic overview of the main modules of the ASR and their interrelations. ASR consists of a Python library of Recipes for materials simulations and a caching system for recording of results and metadata. Recipes are envisioned to communicate with simulation codes via ASE interfaces, although most current Recipe implementations contain parts that are specific to the GPAW code. An arrow from X to Y means that Y calls X. The blue frames on the Instructions of the Recipe box symbolise a caching layer that records all data flow to/from the Recipes.
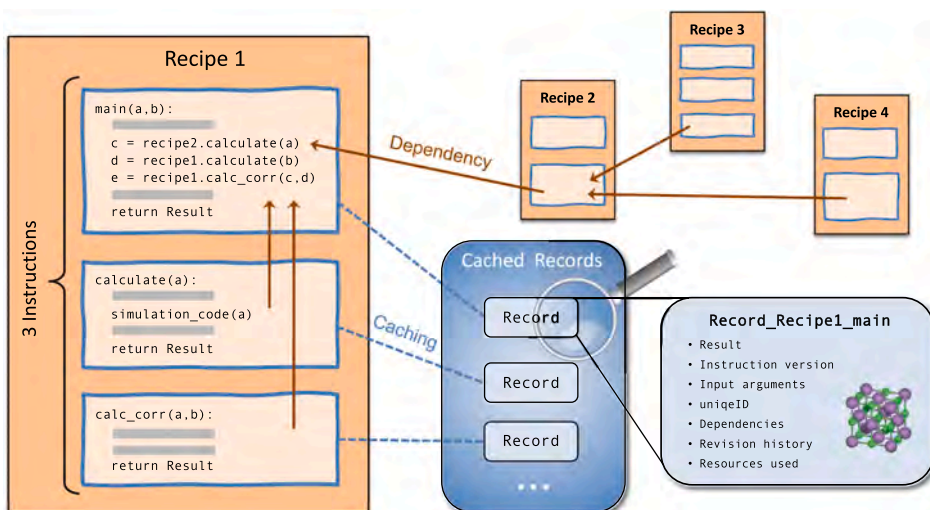
Fig. 2. A Recipe consists of a set of Instructions (see Fig. 3) implementing the computational steps needed to obtain a desired result. An Instruction may call other Instructions of the same, or separate, Recipes. An Instruction always returns a `Record` holding its result, normally represented as a `Result` data structure, together with the dependencies on other Instructions and all additional metadata required to trace back and reproduce the result.
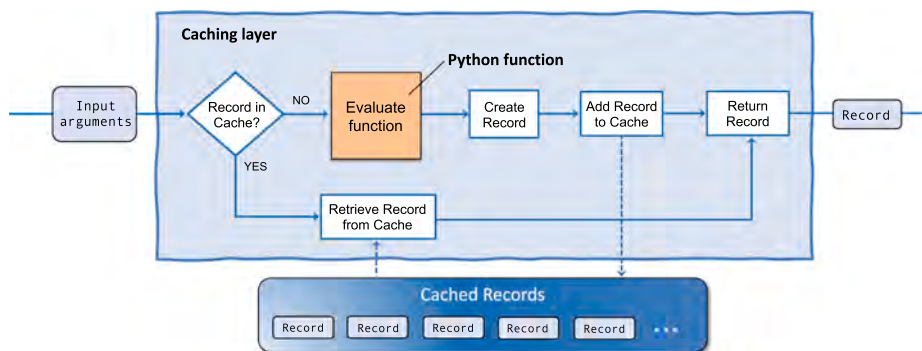
Fig. 3. An Instruction is a Python function (orange) wrapped in a caching layer (light blue). When the function is called with a set of input arguments, the caching layer consults the cache to check if a `Record` for that exact function call already exists. In case of a cache hit, the `Record` is read and returned. In case of a cache miss, the function is evaluated and the `Record` is stored before it is returned.

Having multiple Instructions in a Recipe is usually motivated by code reusability or reduction of resources. The former is relevant when another Recipe needs to perform an identical Instruction (see Section on *Dependencies*). The latter is relevant when the task can be divided into Instructions with different resource requirements, in which case the separation may save computational time or resources. In particular, this is useful if a recalculation of a subset of the generated data is required.

The *input arguments* of an Instruction comprise all the information required to specify its task. When calls to an external simulation code are involved, the input arguments include a code specification, the computational parameters like *k*-point density, basis set specification, or exchange–correlation (xc) functional, as well the atomic structure.

An Instruction carries a *version number* to facilitate data migrations, i. e. transformations of the values or organisation of data produced by the Instruction. This may be required for backward compatibility when Instructions are updated, see Section 8.

3.2. Dependencies

It often happens that an Instruction can benefit from the functionality implemented by other Instructions. An example is the main Instruction of the "band structure" Recipe which calls an Instruction of the "ground state" Recipe to compute the electron density that the band structure should be based on. The caching layer logs whenever an Instruction requests data from another Instruction and uses that information to build a list of data dependencies. The data dependency list is stored in the `Record` object making it possible to trace what other pieces of data were used in the construction of the current result.

Implementation of data-dependencies in Recipes requires no extra coding. Whenever an Instruction calls another Instruction, the caching layer will *automatically* intercept the call and (1) determine if there exists a matching Record (cache hit/miss); (2) log the data dependency by registering the unique IDs and revision UIDS (see Section 8) of any dependent Records.

3.3. The `Record` object

The `Record` object is the basic data unit of ASR. It stores the results of Instructions together with metadata documenting how the results were obtained, and is used by the cache system to identify already performed Instruction calls. The `Record` object contains the following information:

- `Result` object (see Section 3.4)
- Input arguments, if relevant including
  - Atomic structure

- Simulation code specification
- Computational parameters
- Instruction version (see Section 8)
- External codes versions
- Randomly generated unique ID
- Dependencies (see Section 3.2)
- Revision History (see Section 8)
- Execution time and resources (number of cores)

To identify a cache hit/miss when evaluating an Instruction, the caching layer searches the cache for `Records` with matching Instruction name, version, and input arguments. A cache hit is then defined as the existence of a matching `Record`. A recursive comparison is used to compare input arguments with those from existing `Records` within a small numerical tolerance for floating point numbers. Any later evaluations of the Instruction with identical arguments will result in a cache hit.

3.4. The `Result` object

To store and document the result produced by a Recipe, ASR offers a `Result` object that wraps the actual result data (stored as a Python dictionary) in a simple data structure that also contains specification of the result data types along with short explanatory descriptions of the data. In addition, the `Result` object may implement methods to present itself in different formats, see below. Using the `Result` object is optional, but in practice all Instructions that return more than a simple object or value utilizes a `Result` object for improved data documentation.

3.5. Presentation of results

The `Result` object may implement presentation options of the result data in various formats, for example text to terminal, figures, and web panels. The ASR Database and App modules draw on the Recipes' web panel implementation to create web pages for presenting, browsing, and distributing databases containing collected `Result` objects, see Sections 5 and 6. This provides an efficient way of inspecting and sharing data as it is generated, which is highly practical for projects involving multiple collaborators.

3.6. General principles for Recipe development

To maintain and exploit the modular structure of ASR, the development of new Recipes should follow a few general design principles. First, the task performed by a Recipe should be well defined and clearly

bounded to make it easy to use in different contexts. It should always be considered whether the Recipe could be split into smaller independent Recipes that could be useful individually. Additionally, it is encouraged that Recipes are designed/programmed so as to be as broadly applicable as possible, e.g. with respect to the type of material (structure dimensionality, chemical composition, magnetic/non-magnetic, metallic/insulating, etc.). Any information required to define the simulation task should be included in the input argument of the Recipe, i.e. hard coding of parameters should be avoided. This should be done to ensure a flexible use and enhance the data provenance (input arguments are stored in the `Records`). Recipes should employ conservative parameter settings as default to ensure that the results are numerically well converged independent of the application, e.g. material type. Finally, in order to keep ASR Recipes simple and easy-to-read, and in order to enhance the modularity, code-extensive functionalities should be separated out into ASE functions and called from ASR whenever it is possible and sensible, i.e. when the ASE function is useful in other contexts than the specific Recipe.

4. The Recipe library

The ASR currently provides more than 40 complete Recipes allowing users to perform a broad range of materials simulation tasks ranging from construction and analysis of crystal structures over DFT calculations of thermodynamic, mechanical, electronic, magnetic, and optical properties to many-body methods for evaluating response functions, quasiparticle band structures, and collective excitations. A non-exhaustive list of available Recipes is provided in Table 1. It should be stressed that the list constitutes a snapshot of the current state of the Recipe library, which is continuously expanding. For example, we are currently developing Recipes for creating and modeling layered van der Waals structures and point defects in semiconductors.

Most of the currently implemented Recipes rely specifically on the GPAW [49] electronic structure code. As previously mentioned, we are currently working on a generalisation of the ASE Calculator interface to make the Recipes – or a large portion of them – simulation code-independent. Until then, usage of ASR with other simulation codes than GPAW is possible by porting of existing Recipes or development of new ones. The amount of work involved will depend on the type of Recipe and the state of the ASE interface for the specific simulation code.

A few specific examples of Recipe implementations are given in Section 6 where we outline the main computational steps and the final output of the `asr.bandstructure` and `asr.emasses` Recipes, respectively.

5. The ASR Database and App modules

The ASR Database and web App modules make it possible to package, inspect, share, and present ASR-driven projects easily and efficiently. The main tools and opportunities provided by these modules are described in more detail below.

5.1. Database

The ASR Database module can be used to collect `Record` objects from a directory tree into an ASE database. This is achieved by the command `asr database fromtree`. The procedure assumes a "one-material-per-folder" structure, relying on the existence of an atomic structure file in each folder to select `Records` pertaining to that atomic structure. This assumption, which is not fundamental, was chosen as a practical solution given the data layout of the computational 2D materials database. Alternative, more general solutions are being explored. The Database module proceeds to collect atomic structure-`Record` data sets and assign them to a particular row of an ASE database. We shall refer to such a database as an ASR database. Once an ASR database has been collected, it is possible to define key–value-pairs and relate

property data to specific atomic structures.

The Database module also enables the reverse operation, that is, unpacking an existing ASR database to a directory tree containing `Record` objects. This is achieved by the command `asr database totree`. The function is useful when continuing a project, e.g. because existing data must be updated or new data must be added, for which the database is available but not the original directory tree. Moreover, the Database module provides tools for merging and splitting databases.

It is possible to collect a database for any number of materials/`Record` objects – even for a single material – and thereby take advantage of the App tools for presenting and inspecting results in a browser with no extra efforts. However, collecting databases is obviously most powerful in cases involving many materials/properties where the database makes it possible to search and filter the data via the defined key–value-pairs.

The easy installation of ASE through the standard PyPI Python package manager makes the ASE database format highly accessible. Furthermore, the portability of an ASE database (via several backends, e. g. SQLite, PostgreSQL, MariaDB and MySQL) enables easy packaging and distribution of data among different parties.

5.2. Web App

The ASE provides a flexible and easily extensible database web application making it possible to present and inspect the content of an ASE database in a browser. ASR leverages this ASE functionality to customize the web application layout and provide more sophisticated features such as the automatic generation of web panels, generation of figures, and documentation of the presented data by utilizing the web panel data structures encoded in the `Result` objects. Normally a Recipe generates one web panel. However, panels gathering data from several Recipes may be created. One example of the latter is the "Summary" panel of the C2DB web pages discussed in the next section. In this case, a number of Recipes write data to a web panel data structure named "Summary" in their `Result` object. This information is stored in the database when collected. When generating the C2DB web pages from the C2DB database, the App constructs all web panels that are defined in the data pertaining to a particular material. If several Recipes have written to the same web panel, the data will be combined in an order controlled by a priority keyword written together with the web panel data.

5.2.1. Adding information fields

To enhance the accessibility of the data, it is possible to add an explanatory description to specific data entries, i.e, key–value pairs and data files, of an ASR database. These descriptions will appear as text boxes when clicking a "?"-icon placed next to the data on the web panels, see Fig. 5. General information boxes for web panels are always generated by ASR. They contain a customised field that can be manually edited, e.g. providing a short explanation of the data presented in the panel and/or links to relevant literature, and an automatically generated field listing the ASR Recipes that have produced data for the web panel and the key input parameters for the calculations. An example of such an information box is shown in Fig. 5.

5.2.2. Linking rows of databases

ASR provides functionality to create links between rows of the same, or different, ASR databases. This allows the developer to connect relevant materials when designing web panels such that the end user can move swiftly between them when browsing databases. For example, the `asr.convex_hull` Recipe creates the convex hull phase diagram of a material using an ASR reference database of stable materials (originally from the OQMD [28]), and creates a table with links to all the materials on the phase diagram. Other examples, could be to link different defective versions of the same crystalline material or different isomers of the same material/molecule.

The links are defined in `links.json` files in the folders of the relevant materials. These files may be generated manually or automatically using the Recipe `asr.database.treelinks`. When collecting the database, ASR reads the `links.json` file for each folder and stores the information in the `Data` dictionary of the corresponding row. The Recipe `asr.database.crosslinks` then creates links between rows of the collected database and rows of other databases that are given as input to the Recipe. When generating the web panels, ASR uses this information to generate hyperlinks in `HTML` format and present them in the web application for each material.

6. High-throughput example: The C2DB

In this section we present an example of what can be accomplished by the ASR in the realm of data intensive high-throughput applications, showcase some examples of ASR-generated web panels, and discuss two specific Recipe implementations.

Historically, the ASR evolved in a symbiotic relationship with the Computational 2D Materials Database (C2DB) — an extensive database project organising various properties of more than 4000 two-dimensional (2D) materials [32,48].

The C2DB distinguishes itself from existing computational databases of bulk [28–30] and low-dimensional [50,15,51] materials by the large number of physical properties available. These include convex hull diagrams, stiffness tensors, phonons (at high-symmetry points), projected density of states, electronic band structures with spin–orbit effects, effective masses, band topology indices, work functions, Fermi surfaces, plasma frequencies, magnetic anisotropies, magnetic exchange couplings, Bader charges, Born charges, infrared polarisabilities, optical absorption spectra, Raman spectra, and second harmonics generation spectra. The use of beyond-DFT theories for excited state properties (GW band structures and BSE absorption for selected materials) and Berry-phase techniques for band topology and polarization quantities (spontaneous polarization, Born charges, piezoelectric tensors), are other unique features of the C2DB.

Building the first version of C2DB without a fully functioning workflow framework was a long and painstaking endeavour, but absolutely critical for the successful development of the ASR. Today, the entire C2DB project can be generated by a single (MyQueue) Python workflow script comprising a sequence of ASR Recipe calls and simple Python code for controlling and directing the workflow via statements like "`if band_gap > 0:`". Relying on the MyQueue task scheduler (see Section 7.3), generation of the C2DB is accomplished by the single command "`mq workflow c2db_workflow.py tree/*/*/*/`", which will submit the C2DB workflow in folders matching the pattern `tree/*/*/*/`. With the current C2DB workflow, this statement will launch up to 23 unique Instructions for each of the 4047 materials amounting to a total of 59822 individual aiES calculations (some Recipes like phonon and stiffness calculations launch multiple aiES calculations). When the current workflow is run with the GPAW code, about 258 calculations are unsuccessful (most often due to convergence errors in the self-consistency DFT cycle) corresponding to a success rate of 99.5%.

Apart from the data provenance control that ensures the documentation and reproducibility of the data, there are two aspects of the ASR that are particularly crucial for making high-throughput computations work efficiently in practice. First, the caching functionality ensures that Recipes which have already been performed are automatically skipped by ASR (unless something in the input for a Recipe has changed since it was last executed). This means that only a single workflow script needs to be maintained and submitted every time something has been changed, e.g., new materials have been added, the workflow script has been updated, it has been decided to rerun certain tasks with new parameters, or a Recipe has been modified. Such functionality is essential because running and maintaining high-throughput projects inevitably requires that subsets of calculations are repeated at different points in time. Secondly, the carefully designed and well tested Recipes minimise

the number of unsuccessful calculations and the risk of human errors.

6.1. Recipe and web page examples

Below we present a few examples of output generated by the ASR-C2DB workflow (for a full impression we refer the reader to the C2DB website).

6.1.1. Search page

Fig. 4 shows the C2DB search page, which consists of a search/filtering section followed by a list of the database rows presented by a selected number of key–value pairs. Clicking one of the highlighted key names once (twice) will sort the rows in increasing (decreasing) order of that key. Which keys should be shown by default can be customized, but the user can always add extra keys via the "Add column" button. By default, the search page generated by the ASR App module will contain only the search field in the upper section, but additional fields or buttons may be added for easy filtering according to the most relevant parameters.

6.1.2. "Summary" panel

Fig. 5 shows the C2DB web page for monolayer MoS_2. All the web panels produced by the various Recipes of the workflow are seen, but only the "Summary" panel is unfolded. This panel is designed to provide an overview of the most basic properties of the material, and gathers data from the `Result` objects generated by the following Recipes: `asr.gs`, `asr.gw`, `asr.hse`, `asr.phonons`, `asr.magstate`, `asr.stiffness`, `asr.convex_hull` and `asr.structureinfo`.

Fig. 5 also shows the information box of the "Effective masses" web panel. It contains a short explanation of the effective mass tensor and how it is evaluated by the Recipe as well as a link to a relevant paper. The automatically generated part shows that the panel contains data generated by the `asr.emasses` Recipe. The two fields at the top of the page "Download raw data" and "Browse raw data" provide access to the entire data set comprised by all `Result` objects of the specific material entry of the database.

6.1.3. "Band structure" Recipe

As another example, Fig. 6 shows the "Electronic band structure" panel for monolayer CrW_3S_8 as calculated and presented by the Recipe `asr.bandstructure`. The band structure is calculated with the PBE xc-functional including spin–orbit interactions. The out-of-plane spin projections of the states is shown by the color code. The main computational steps carried out by this Recipe are:

- Perform a self-consistent ground state calculation (by calling the `calculate` Instruction of the ground state Recipe `asr.gs`) to obtain a converged electron density.
- Determine crystal symmetries and corresponding band path (uses ASE functionalities).
- Calculate the Kohn–Sham eigenvalues along the band path. For magnetic materials, this step calls the Recipe `asr.magnetic_anisotropy` to obtain the magnetic easy axis for evaluating spin projections.
- Call the main Instruction of the ground state Recipe to get the Fermi level (in 3D) or the vacuum level (in < 3D) for use as zero-point energy for the band structure.

In addition to these computational steps, the main Instruction of the Recipe formats two figures to present the band structure itself and the Brillouin zone with the band path and the positions of the valence band maximum (VBM) and conduction band minimum (CBM). Note that the position of the VBM and CBM, as well as a number of other properties like the band gap and band edge energies (not shown), are determined by the Recipe `asr.gs`, which is called by `asr.bandstructure`.

Fig. 4. The search page of C2DB with the first few rows of the database shown below. The default web page generated by ASR includes only the top most search field, but the panel can be customized by additional fields and buttons for more convenient data filtering.

6.1.4. "Effective masses" Recipe

Fig. 7 shows a screenshot of the "Effective masses" panel for mono-layer CrW_3S_8 generated by the Recipe `asr.emasses`. The effective mass tensor is calculated with the PBE xc-functional including spin–orbit interactions. The color code represents the spin projections along the *z*-axis. In addition to the effective masses themselves, the Recipe evaluates a "band parabolicity" parameter defined as the mean absolute relative error (MARE) between the parabolic fit and the true bands in an energy range of 25 meV. The main computational steps carried out by this Recipe involve three subsequent *k*-point grid refinements; specifically:

- Perform a self-consistent ground state calculation on a uniform *k*-point grid (by calling the `calculate` Instruction of the Recipe `asr.gs`) to obtain a converged electron density as well as Kohn–Sham band energies.
- Locate the preliminary positions of the VBM and CBM and calculate band energies on a higher-density *k*-point grid around the VBM and CBM to locate the VBM and CBM positions with higher accuracy.
- Define final high-density *k*-point grids in the vicinity of the VBM and CBM points, and calculate band energies.
- Locate VBM and CBM and fit bands by second-order polynomial using band energies in an energy range of 1 meV from the band extremum.
- Calculate band structures for the web panel and evaluate the "parabolicity parameter".

It should be noted that even though effective mass calculations appear to be a simple task, it is surprisingly tricky to design a scheme that performs efficiently, robustly, and accurately across all types of band structures including flat bands, highly dispersive bands, highly anisotropic bands, and bands exhibiting complex spin–orbit effects like Rashba splittings.

6.1.5. General comments

In contrast to the "Summary" panel, which has been customized for the C2DB project (that is, the web panel sections of the relevant Recipes have been appropriately adjusted), the "Electronic band structure" and "Effective masses" panels are the default web panels produced by the `asr.bandstructure` and `asr.emasses` Recipes, respectively.

The examples given here concern two-dimensional (2D) materials. However, the Recipes `asr.bandstructure` and `asr.emasses` (like all other Recipes of the current ASR library) apply also to 1D and 3D materials, as well as 0D where it is meaningful. As mentioned in Section 3.6, this kind of generality should always be strived for when designing Recipes. Achieving this may be straightforward or more involved depending on the Recipe. The Recipe for the stiffness tensor represents an easy case, where the dimensionality merely dictates the number of axes along which the material must be strained. The Recipe for the band structure is more involved in this regard, as the determination of the band path requires separate treatments in 2D and 3D as does the determination of the spin projection axis (in 2D the out-of-plane direction is a natural choice while in 3D the magnetic easy axis is more appropriate).

7. User interfaces

The ASR can be used via four different interfaces, c.f. Fig. 1: A command line interface (CLI), a Python interface, a task scheduling front-end, and an app-based interface. Below we describe each interface in more detail.

7.1. The CLI

The CLI provides convenient commands for easy interaction with ASR via the `cache` and `run` subcommands. The `cache` subcommand allows inspection of the `Records` stored in the cache, in particular their
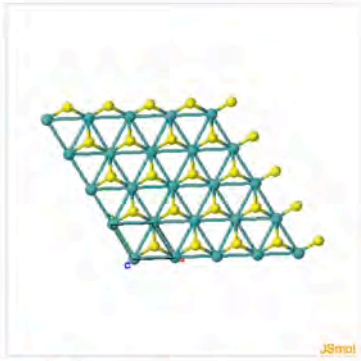
Fig. 5. Screenshot of the web page for monolayer MoS_2 from the C2DB project (only the "Summary" panel is unfolded). The panel presents data from the Result objects generated by the following Recipes: `asr.gs`, `asr.gw`, `asr.hse`, `asr.phonons`, `asr.magstate`, `asr.stiffness`, `asr.convex_hull`, `asr.structureinfo`.
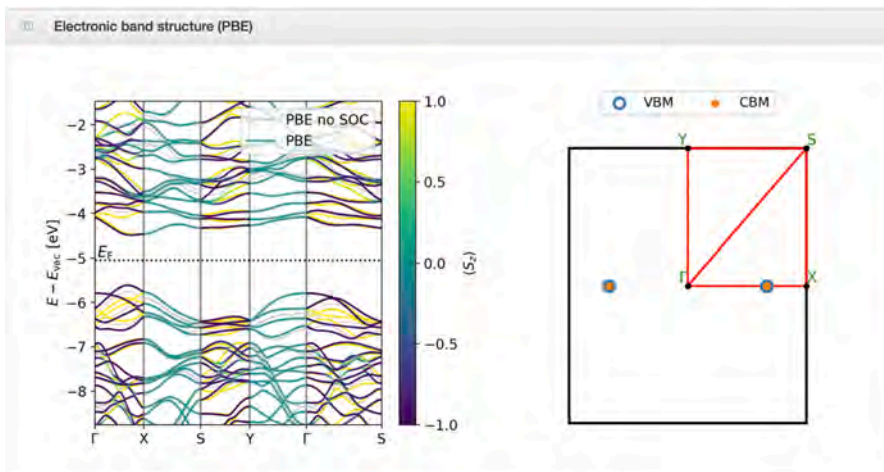
Fig. 6. Screenshot of the "Band structure" panel for monolayer CrW_3S_8 from the C2DB project. The web panel contains data computed by the `asr.band-structure` Recipe.
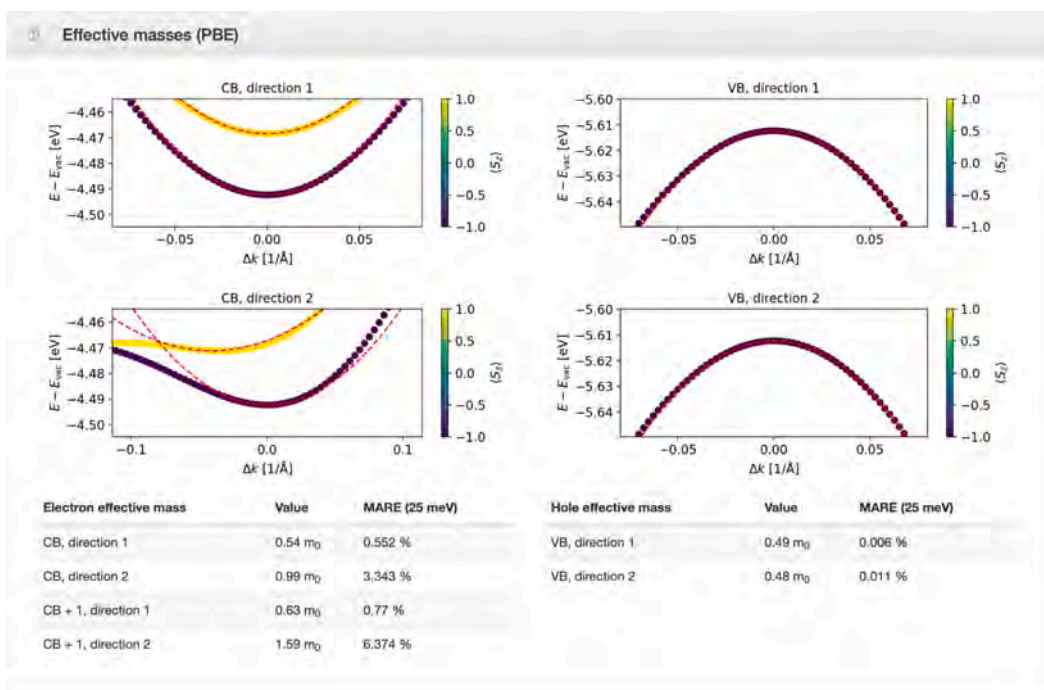

Electron effective mass	Value	MARE (25 meV)	Hole effective mass	Value	MARE (25 meV)
CB, direction 1	0.54 m_0	0.552 %	VB, direction 1	0.49 m_0	0.006 %
CB, direction 2	0.99 m_0	3.343 %	VB, direction 2	0.48 m_0	0.011 %
CB + 1, direction 1	0.63 m_0	0.77 %			
CB + 1, direction 2	1.59 m_0	6.374 %			

Fig. 7. Screenshot of the "Effective masses" panel for monolayer CrW_3S_8 from the C2DB project. The panel contains data computed by the `asr.emasses` Recipe.

Result data. For example, `$ asr cache ls name=asr.gs` will list all Records produced by the "ground state" Recipe. The `run` subcommand can be used to execute Recipes directly from the command line. For example, `$ asr run asr.gs` will run the *ground state* Recipe.

7.2. Python interface

The Python scripting interface allows inspection of Records and execution of Recipes directly from Python. This makes it possible to implement more complex logic and integrate directly with ASE and any other tools in the user's Python toolkit.

7.3. MyQueue interface

For high-throughput computations, ASR can be used in combination with a workflow manager that can handle the interaction with the scheduler of the supercomputer, such as Fireworks [40] or MyQueue [47]. The latter is a personal, decentralized, and lightweight front-end

for schedulers (currently supporting SLURM, PBS, and LSF), which has been co-designed with ASR. MyQueue has a command line interface, which allows for submission of thousands of jobs in one command and provides easy-to-use tools for generating an overview of the status of jobs ('done', 'queued', 'failed' etc.). It also has a Python interface that can be used to define workflows. A Python script defines a dependency tree of *tasks* that MyQueue will submit without user involvement. The dependencies take the form: "if task X is done then submit task Y". MyQueue works directly with folders and files, which makes it transparent and easy to use. Together ASR and MyQueue provide a powerful and extremely flexible toolkit for high-throughput materials computations.

Individual Instructions of the Recipes may be defined as separate MyQueue tasks, such that computational resources can be specifically dedicated each Instruction ensuring a flexible and efficient execution of any workflow. It is, however, not a requirement to specify resources on a per Instruction basis, in which case the resources specified for the main Instruction will apply to all Instructions of the Recipe.

7.4. App interface

The App interface is a web-based read-only interface that allows the user to present and inspect the data stored in an ASE database on a local or public network. Distributing the data on a local network is convenient for larger projects and/or projects involving several users, as it allows for easy sharing and monitoring of the data as the project evolves. Once a project is finalized, the App may be used as a platform to present the data to the world via web pages. The data presentation used by the App is defined in the `Result` object of the Recipes.

8. Data maintenance

It sometimes happens that a Recipe, or one of its Instructions, has to be updated, e.g. because a bug has been detected or it has been found appropriate to store additional metadata. Such updates may imply that previously generated `Records` are no longer consistent with the current implementation of the Recipe. Depending on the nature of the change made to the Recipe, it may be possible to update the `Record` objects

without rerunning the Recipe (data migration) or it may be necessary to rerun the entire Recipe or some of its Instructions (data regeneration).

To support the migration of data, ASR implements a simple versioning system for Instructions. An Instruction is associated with a integer version number which is stored in the `Record` and identifies the version of the Instruction at the time of creation. When an Instruction is changed, its version number may be increased by the developer. Since the caching layer matches the current Instruction version number against `Records` in the cache (see Section 3.3), older `Records` would no longer yield cache hits and are then said to be invalidated.

To facilitate the migration of invalidated `Records`, it is possible to specify `Migrations` that can be associated with an Instruction and thereby provide a way to bring old `Records` up to date. In practice, a `Migration` bundles a `Record` transformation function, a unique migration ID and a human readable description of the effect of the migration, see Fig. 8. In general, a `Record` transformation function induces a change to a `Record`. For example, this could be to convert a `Record` of version n to a later version $n+1$ without rerunning the Instruction, but in general the effect of the transformation could be anything. Use of transformation functions is typically possible when the update involves changes to metadata and/or data restructuring while the actual result of the Instruction is unchanged.

When a `Migration` is applied to a `Record`, a `Revision` object is produced. A `Revision` contains a randomly generated UID, the UID of the applied `Migration`, an explanatory description of the changes made to the `Record`, and an automatically generated list of the `Record` entries that were changed, added or deleted. The auto-generated list of changes is constructed by comparing the `Record` returned by the transformation function to the input `Record`.

Upon migration of a `Record`, a revision history is updated by the latest `Revision` and stored in the migrated `Record`. The revision history can be inspected by users to learn which revisions, if any, have previously been applied to a given `Record`.

A `Selector` is used to identify the `Records` to be migrated, e.g. based on the Instruction name and version number. The `Selector` is bundled together with a `Migration` into a `MigrationSelector`, which can determine whether a particular `Record` matches the selection criteria of the `Selector`. To migrate a `Record`, ASR searches
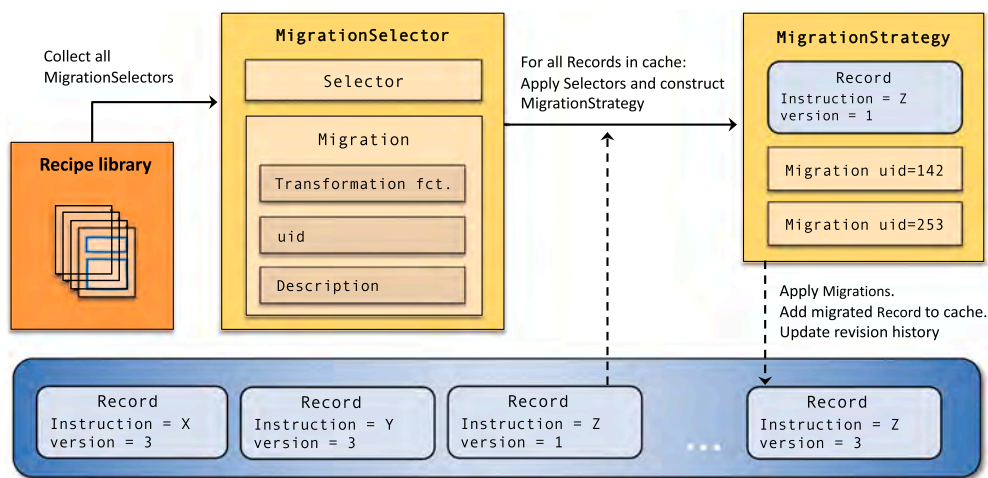

Fig. 8. To support data maintenance, ASR provides migration tools for bringing `Records` up-to-date with the latest version of the Instructions that produced them thereby avoiding recalculations whenever possible. The ASR migration procedure consists of the main steps: (1) Collect all `MigrationSelectors` from all available `Instructions`. (2) Select the migratable `Records` of the cache. (3) Determine a migration strategy (an ordered list of `Migrations`) for each migratable `Record`. (4) Apply transformation functions to migrate the `Records` and add them to the cache. (5) Update the revision history by a `Revision` object that documents the effect of the migration.

through all Recipes to collect their `MigrationSelectors` (if they have any) and apply them to the `Record` to find a "migration strategy", i.e., which `Migrations` to be applied and in which order. The migration strategy is then encoded in a `MigrationStrategy`, which couples a particular `Record` to an ordered list of `Migrations`. The particular `MigrationStrategy` can then be applied to the Cache to execute the migration of the associated `Record`.

ASR provides a simple CLI, via `asr cache migrate`, to analyse existing `Records` in the cache, and identify migratable `Records`. Whenever the `asr` version used in a given project is upgraded, a project participant should identify migratable `Records`, migrate them and then rerun the project workflow. Up-to-date `Records` will then be taken directly from the cache, whereas the Instructions with invalidated `Records` and no associated `Migration`, i.e. `Records` that the developer cannot migrate directly to the newest version, will be rerun.

In order to minimize the computational cost of bringing data up-to-date with ASR, developers are strongly advised to supply `Migrations` with their Recipe updates whenever possible.

To provide the best conditions for the long term deployment of ASR-generated data, the `asr` version of important projects should be upgraded regularly and the project workflow rerun. Obviously, this action may induce changes in the data. Whether this is acceptable or not is ultimately a strategic decision. However, for dynamic data projects, a regular version upgrade not only ensures that the data is of the highest quality, it also makes it easier for other parties to deploy the data because existing results (`Records`) can be reused directly with the newest version of ASR without having to rerun Recipes to bring the data up-to-date.

9. Data provenance

Simply stated, data provenance is the documentation of the circumstances under which a piece of data came into existence. This includes how the data originally was constructed, how the data has changed over time (also known as data-lineage) and a documentation of relevant system specifications such as architecture, operating system, important system packages, executables etc. If data provenance is handled perfectly, then data will in principle be reproducible, i.e. given access to exactly the same systems and software, any piece of data can be reproduced. In a scientific context, where reproducibility is key, data provenance is naturally very important.

In ASR, the basic unit of data is the `Record` object, which connects the result of an Instruction with various pieces of contextual metadata, see Section 3.3. Taken together, the metadata tell the story of how the original `Record` came into existence (Instruction name/version and input arguments), which other `Records` were implicitly used for the construction of this `Record` (Dependencies), what external package versions were used, and how the `Record` has transformed over time (Revision history). For simplicity, since it would be outside the scope of ASR, system information is not stored with the `Record`, which, in our experience, is not practically relevant for the purposes of ASR. As such, we characterize ASR as practically, but not perfectly, data provenant.

10. Documentation

ASR itself is documented on Read the Docs. The data is documented through the `Record` and `Result` objects, see previous Section on data provenance.

11. Technical specifications

Some technical specifications are listed in Table 2. ASR can be installed via pip using the command `pip install asr`.

ASR requires or is normally used with the following software:

• Python libraries: ASE, numpy, matplotlib, plotly, flask, click

Table 2
Technical specifications.

Source code	https://gitlab.com/asr-dev/asr
Releases	https://pypi.org/project/asr/
License	GNU GPLv3 or newer (free software)
Documentation	https://asr.readthedocs.io/en/latest/

• Computational and workflow software: GPAW or other ASE codes, MyQueue (SLURM/PBS/LSF)
• Optional extras: spglib, phonopy, and pymatgen (for Recipes); jinja, mysql or other ASE database backends

For community support see https://asr.readthedocs.io/en/latest/src/contact.html.

12. Summary and outlook

This article has introduced The Atomic Simulation Recipes (ASR) as an open source Python framework for developing materials simulation workflows and managing the data they produce.

To facilitate the transition to a paradigm of data-intensive science, ASR was designed to support the development of materials simulation workflows that operate in accordance with the FAIR data principles, by providing tools and concepts that are general enough that they do not restrict the user whilst being concrete enough to make a real difference. The ASR achieves this through the notion of a Recipe: a general Python script that performs a well defined simulation task and is wrapped in a caching layer that logs all relevant metadata without involving the user. This construction places essentially no restrictions on the developer's freedom to design and control the workflow, but resolves the critical and complex issue of keeping track of the data provenance. We stress that the core of ASR, i.e. the Recipe concept and the caching system, is fully simulation code independent. In particular, it is not tied to materials simulations and could potentially be useful in other areas of computational science.

Beyond the built-in data documentation, there are many benefits of using standardized, well tested, and well documented Recipes. For example, it saves time and promotes a more sustainable scripting culture by reducing the need for individual researchers to write and maintain their own personal scripts (which can be hard for other to read and are often lost when the developer leaves the group). Furthermore, it reduces the risk of human errors and lowers the barrier for researchers to undertake simulation tasks with which they have little prior experience.

The fact that Recipes are independent units with own data provenance control implies that they can be freely combined to create advanced workflows using Python scripting for maximal flexibility. Such workflows can be executed on supercomputers using a workflow management software that supports a Python interface. To this end, we have developed the MyQueue [47] task manager that works as a front-end to the most common schedulers (currently SLURM, PBS, and LSF). While MyQueue will resubmit jobs that have timed out or crashed due to lack of memory, code-related failures must be handled manually. In the future, ASR should integrate more closely with MyQueue to permit that errors from the simulation codes are automatically analysed and reacted upon. Along the same lines, an automated estimation of the HPC resources (time/memory/nodes) required by individual tasks could limit the number of failed jobs and improve the utilization of resources.

The current Recipe library already covers a wide range of materials simulation tasks and more are continuously being added. Of special importance are Recipes for advanced beyond-DFT calculations where the benefits in terms of a lowered user barrier, improved data quality, and increased utilization of computing resources, are particularly large. The Recipe concept should also be advantageous for implementation of machine learning methods that could integrate with ASR databases and "standard" Recipes to make for more intelligent and computationally

efficient workflows.

The ASR makes extensive use of the Atomic Simulation Environment (ASE) as a toolkit to process atomistic calculations. In particular, ASE is used as a front-end for ASR to communicate with external simulation codes. This has the clear advantage that ASR can become decoupled from the simulation codes. This decoupling is currently not in place, and the majority of the existing Recipe implementations contain code parts that are specific to the GPAW electronic structure code. To make ASR fully simulation code-independent, the ASE Calculator interfaces must be further generalized. This includes extensions of the interfaces to access outputs of calculations as well as a systematic mechanism to control multi-step tasks. The adaptation of this interface to multiple codes will eventually require a community effort that we hope many code developers will take part in. Until then, Recipes must to some extent be code specific.

CRediT authorship contribution statement

Morten Gjerding: Conceptualization, Methodology, Software, Validation, Writing - original draft, Writing - review & editing, Visualization. **Thorbjørn Skovhus:** Conceptualization, Software, Validation, Writing - review & editing, Visualization. **Asbjørn Rasmussen:** Conceptualization, Software, Validation, Writing - review & editing. **Fabian Bertoldo:** Conceptualization, Software, Validation, Writing - review & editing. **Ask Hjorth Larsen:** Conceptualization, Software, Validation, Writing - review & editing. **Jens Jørgen Mortensen:** Conceptualization, Software, Validation, Writing - review & editing. **Kristian Sommer Thygesen:** Conceptualization, Validation, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition, Project administration.

Declaration of Competing Interest

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

Acknowledgments

References

[1] J. Greeley, T.F. Jaramillo, J. Bonde, I. Chorkendorff, J.K. Nørskov, Nat. Mater. 5 (2006) 909–913.
[2] G.K. Madsen, J. Am. Chem. Soc. 128 (2006) 12140–12146.
[3] S. Curtarolo, et al., Nat. Mater. 12 (2013) 191–201.
[4] S. Kirklin, B. Meredig, C. Wolverton, Adv. Energy Mater. 3 (2013) 252–262.
[5] K.B. Ørnsø, J.M. Garcia-Lastra, K.S. Thygesen, Phys. Chem. Chem. Phys. 15 (2013) 19478–19486.
[6] Z. Zhang, et al., ACS Omega 4 (2019) 7822–7828.
[7] W. Chen, et al., J. Mater. Chem. C 4 (2016) 4414–4426.
[8] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A.M. Brockway, A. Aspuru-Guzik, J. Phys. Chem. Lett. 2 (2011) 2241–2251.
[9] S. Bhattacharya, G.K. Madsen, Phys. Rev. B Condens. Matter 92 (2015), 085205.
[10] I.E. Castelli, et al., Energy Environ. Sci. 5 (2012) 5814–5819.
[11] G. Hautier, A. Miglio, G. Ceder, G.M. Rignanese, X. Gonze, Nat. Commun. 4 (2013) 1–7.
[12] L. Yu, A. Zunger, Phys. Rev. Lett. 108 (2012), 068701.
[13] K. Kuhar, M. Pandey, K.S. Thygesen, K.W. Jacobsen, ACS Energy Lett. 3 (2018) 436–446.
[14] M. Aykol, S. Kim, V.I. Hegde, D. Snydacker, Z. Lu, S. Hao, S. Kirklin, D. Morgan, C. Wolverton, Nat. Commun. 7 (2016) 1–12.
[15] N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I.E. Castelli, A. Cepellotti, G. Pizzi, et al., Nat. Nanotechnol. 13 (2018) 246–252.
[16] L.Q. Chen, L.D. Chen, S.V. Kalinin, G. Klimeck, S.K. Kumar, J. Neugebauer, I. Terasaki, NPJ Comput. Mater. 1 (2015) 1–2.
[17] M. Rupp, A. Tkatchenko, K.R. Müller, O.A. Von Lilienfeld, Phys. Rev. Lett. 108 (2012), 058301.
[18] J. Lee, A. Seko, K. Shitara, K. Nakayama, I. Tanaka, Phys. Rev. B 93 (2016), 115104.
[19] T. Xie, J.C. Grossman, Phys. Rev. Lett. 120 (2018), 145301.
[20] L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M. Scheffler, Phys. Rev. Lett. 114 (2015) 105503.
[21] P.B. Jørgensen, E.G. del Río, M.N. Schmidt, K.W. Jacobsen, Phys. Rev. B 100 (2019) 104114.
[22] K. Ghosh, A. Stuke, M. Todorović, P.B. Jørgensen, M.N. Schmidt, A. Vehtari, P. Rinke, Adv. Sci. 6 (2019) 1801367.
[23] V.L. Deringer, G. Csányi, Phys. Rev. B 95 (2017), 094203.
[24] S. Lorenz, A. Groß, M. Scheffler, Chem. Phys. Lett. 395 (2004) 210–215.
[25] J. Behler, M. Parrinello, Phys. Rev. Lett. 98 (2007), 146401.
[26] N. Artrith, A. Urban, Comput. Mater. Sci. 114 (2016) 135–150.
[27] K.S. Thygesen, K.W. Jacobsen, Science 354 (2016) 180–181.
[28] J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, JOM 65 (2013) 1501–1509.
[29] A. Jain, et al., APL Mater. 1 (2013), 011002.
[30] S. Curtarolo, W. Setyawan, G.L. Hart, M. Jahnatek, R.V. Chepulskii, R.H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, et al., Comput. Mater. Sci. 58 (2012) 218–226.
[31] C. Draxl, M. Scheffler, J. Phys.: Mater. 2 (2019), 036001.
[32] S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P.S. Schmidt, N.F. Hinsche, M.N. Gjerding, D. Torelli, P.M. Larsen, A.C. Riis-Jensen, et al., 2D Mater 5 (2018) 042002.
[33] S.S. Borysov, R.M. Geilhufe, A.V. Balatsky, PloS One 12 (2017), e0171501.
[34] K.T. Winther, M.J. Hoffmann, J.R. Boes, O. Mamun, M. Bajdich, T. Bligaard, Sci. Data 6 (2019) 1–10.
[35] L. Talirz, S. Kumbhar, E. Passaro, A.V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S.P. Huber, S. Zoupanos, et al., Sci. Data 7 (2020) 1–12.
[36] R. Armiento, Mach. Learn. Meets Quant. Phys. (2020) 377–395.
[37] L. Himanen, A. Geurts, A.S. Foster, P. Rinke, Adv. Sci. 6 (2019) 1900808.
[38] W. Kohn, L.J. Sham, Phys. Rev. 140 (1965) A1133.
[39] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, Santos L.B. da Silva, P.E. Bourne, et al., Sci. Data 3 (2016) 1–9.
[40] A. Jain, et al., Concurr. Comput. 27 (2015) 5037–5059.
[41] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, B. Kozinsky, Comput. Mater. Sci. 111 (2016) 218–230.
[42] K. Mathew, J.H. Montoya, A. Faghaninia, S. Dwarakanath, M. Aykol, H. Tang, I.h. Chu, T. Smidt, B. Bocklund, M. Horton, et al., Comput. Mater. Sci. 139 (2017) 140–152.
[43] A.H. Larsen, J.J. Mortensen, J. Blomqvist, I.E. Castelli, R. Christensen, M. Dułak, J. Friis, M.N. Groves, B. Hammer, C. Hargus, et al., J. Phys.: Condens. Mat. 29 (2017), 273002.
[44] S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K.A. Persson, G. Ceder, Comput. Mater. Sci. 68 (2013) 314–319.
[45] A. Togo, I. Tanaka, 2018 arXiv preprint arXiv:1808.01590.
[46] A. Togo, L. Chaput, I. Tanaka, Phys. Rev. B 91 (2015), 094306.
[47] J. Mortensen, M. Gjerding, K. Thygesen, J. Open Sour. Softw. 5 (2020) 1844.
[48] M. Gjerding, et al., 2D Mater 8 (2021), 044002.
[49] J. Enkovaara, C. Rostgaard, J.J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. Hansen, et al., J. Phys.: Condens. Mat. 22 (2010), 253202.
[50] M. Ashton, J. Paul, S.B. Sinnott, R.G. Hennig, Phys. Rev. Lett. 118 (2017), 106101.
[51] J. Zhou, L. Shen, M.D. Costa, K.A. Persson, S.P. Ong, P. Huck, Y. Lu, X. Ma, Y. Chen, H. Tang, et al., Sci. Data 6 (2019) 1–10.

Bibliography

[1] T. Bell, *The ancient history of copper*, edited by ThoughtCo, https://www.thoughtco.com/copper-history-pt-i-2340112, accesed on 08.04.2022.

[2] C. D. A. Inc., editor, *The beginnings of bronze*, https://www.copper.org/education/history/60centuries/raw_material/thebeginnings.html, accesed on 08.04.2022.

[3] C. E. N. Bromehead, *Practical geology in ancient britain, Pt. 2: the metals* (Proc. Geol. Ass., 1947).

[4] G. W. Pickard, "Means for receiving intelligence communicated by electric waves," U.S. Patent **836** (1906).

[5] W. Shockley, "The theory of p-n junctions in semiconductors and p-n junction transistors," Bell Syst. Tech. J. **28**, 435 (1949).

[6] J. S. Park, S. Kim, Z. Xie, and A. Walsh, "Point defect engineering in thin-film solar cells," Nat. Rev. Mater. **3**, 194 (2018).

[7] J. P. Dowling and G. J. Milburn, "Quantum technology: the second quantum revolution," Philos. Trans. Royal Soc. A **361**, 1655 (2003).

[8] P. Michler, *Quantum dots for quantum information technologies*, Vol. 1 (Springer, 2017).

[9] C. Monroe, "Quantum information processing with atoms and photons," Nature **416**, 238 (2002).

[10] E. Lucero, M. Hofheinz, M. Ansmann, R. C. Bialczak, N. Katz, M. Neeley, A. O'Connell, H. Wang, A. Cleland, and J. M. Martinis, "High-fidelity gates in a single Josephson qubit," Phys. Rev. Lett. **100**, 247001 (2008).

[11] C. E. Dreyer, A. Alkauskas, J. L. Lyons, A. Janotti, and C. G. Van de Walle, "First-principles calculations of point defects for quantum technologies," Annu. Rev. Mater. Sci. **48**, 1 (2018).

[12] J. J. Pla, K. Y. Tan, J. P. Dehollain, W. H. Lim, J. J. Morton, D. N. Jamieson, A. S. Dzurak, and A. Morello, "A single-atom electron spin qubit in silicon," Nature **489**, 541 (2012).

[13] J. Weber, W. Koehl, J. Varley, A. Janotti, B. Buckley, C. Van de Walle, and D. D. Awschalom, "Quantum computing with defects," PNAS **107**, 8513 (2010).

[14] I. Aharonovich, S. Castelletto, D. Simpson, C.-H. Su, A. Greentree, and S. Prawer, "Diamond-based single-photon emitters," Rep. Prog. Phys. **74**, 076501 (2011).

[15] V. Dobrovitski, G. Fuchs, A. Falk, C. Santori, and D. Awschalom, "Quantum control over single spins in diamond," Annu. Rev. Condens. Matter Phys. **4**, 23 (2013).

[16] M. W. Doherty, N. B. Manson, P. Delaney, F. Jelezko, J. Wrachtrup, and L. C. Hollenberg, "The nitrogen-vacancy colour centre in diamond," Phys. Rep. **528**, 1 (2013).

[17] G. Davies and M. Hamer, "Optical studies of the 1.945 eV vibronic band in diamond," Proc. Math. Phys. Eng. Sci. **348**, 285 (1976).

[18] A. Gali, M. Fyta, and E. Kaxiras, "*ab initio* supercell calculations on nitrogen-vacancy center in diamond: Electronic structure and hyperfine tensors," Phys. Rev. B **77**, 155206 (2008).

[19] J. Weber, W. Koehl, J. Varley, A. Janotti, B. Buckley, C. Van de Walle, and D. Awschalom, "Defects in SiC for quantum computing," Journal of Applied Physics **109**, 102417 (2011).

[20] T. Hornos, A. Gali, and B. G. Svensson, "Large-scale electronic structure calculations of vacancies in 4H-SiC using the Heyd-Scuseria-Ernzerhof screened hybrid density functional," in Materials science forum, Vol. 679 (Trans Tech Publ, 2011), pages 261–264.

[21] E. Janzén, A. Gali, P. Carlsson, A. Gällström, B. Magnusson, and N. T. Son, "The silicon vacancy in SiC," Physica B: Condensed Matter **404**, 4354 (2009).

[22] J. Davidsson, V. Ivády, R. Armiento, N. Son, A. Gali, and I. A. Abrikosov, "First principles predictions of magneto-optical data for semiconductor point defect identification: the case of divacancy defects in 4H-SiC," New J. Phys. **20**, 023035 (2018).

[23] T. T. Tran, C. Elbadawi, D. Totonjian, C. J. Lobo, G. Grosso, H. Moon, D. R. Englund, M. J. Ford, I. Aharonovich, and M. Toth, "Robust multicolor single photon emission from point defects in hexagonal boron nitride," ACS Nano **10**, 7331 (2016).

[24] G. Grosso, H. Moon, B. Lienhard, S. Ali, D. K. Efetov, M. M. Furchi, P. Jarillo-Herrero, M. J. Ford, I. Aharonovich, and D. Englund, "Tunable and high-purity room temperature single-photon emission from atomic defects in hexagonal boron nitride," Nat. Commun. **8**, 1 (2017).

[25] K. Barthelmi, J. Klein, A. Hötger, L. Sigl, F. Sigger, E. Mitterreiter, S. Rey, S. Gyger, M. Lorke, M. Florian, et al., "Atomistic defects as single-photon emitters in atomically thin MoS_2," Appl. Phys. Lett. **117**, 070501 (2020).

[26] A. Jahan, M. Y. Ismail, S. Sapuan, and F. Mustapha, "Material screening and choosing methods - a review," Mater. Des. **31**, 696 (2010).

[27] O. Cf, "Transforming our world: the 2030 agenda for sustainable development," United Nations: New York, NY, USA (2015).

[28] K. S. Novoselov, A. K. Geim, S. V. Morozov, D.-e. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, "Electric field effect in atomically thin carbon films," Science **306**, 666 (2004).

[29] K. S. Novoselov, D. Jiang, F. Schedin, T. Booth, V. Khotkevich, S. Morozov, and A. K. Geim, "Two-dimensional atomic crystals," Proc. Natl. Acad. Sci. U.S.A. **102**, 10451 (2005).

[30] K. Novoselov, o. A. Mishchenko, o. A. Carvalho, and A. Castro Neto, "2D materials and van der Waals heterostructures," Science **353**, aac9439 (2016).

[31] G. R. Bhimanapati, Z. Lin, V. Meunier, Y. Jung, J. Cha, S. Das, D. Xiao, Y. Son, M. S. Strano, V. R. Cooper, et al., "Recent advances in two-dimensional materials beyond graphene," ACS Nano **9**, 11509 (2015).

[32] M. P. Levendorf, C.-J. Kim, L. Brown, P. Y. Huang, R. W. Havener, D. A. Muller, and J. Park, "Graphene and boron nitride lateral heterostructures for atomically thin circuitry," Nature **488**, 627 (2012).

[33] J. Jiang, T. Xu, J. Lu, L. Sun, and Z. Ni, "Defect engineering in 2D materials: precise manipulation and improved functionalities," Research **2019** (2019).

[34] Z. Lin, B. R. Carvalho, E. Kahn, R. Lv, R. Rao, H. Terrones, M. A. Pimenta, and M. Terrones, "Defect engineering of two-dimensional transition metal dichalcogenides," 2D Mater. **3**, 022002 (2016).

[35] Q. Liang, Q. Zhang, X. Zhao, M. Liu, and A. T. Wee, "Defect engineering of two-dimensional transition-metal dichalcogenides: applications, challenges, and opportunities," ACS Nano **15**, 2165 (2021).

[36] K. M. McCreary, A. G. Swartz, W. Han, J. Fabian, and R. K. Kawakami, "Magnetic moment formation in graphene detected by scattering of pure spin currents," Phys. Rev. Lett. **109**, 186604 (2012).

[37] H. González-Herrero, J. M. Gómez-Rodríguez, P. Mallet, M. Moaied, J. J. Palacios, C. Salgado, M. M. Ugeda, J.-Y. Veuillen, F. Yndurain, and I. Brihuega, "Atomic-scale control of graphene magnetism by using hydrogen atoms," Science **352**, 437 (2016).

[38] R. Friend and A. Yoffe, "Electronic properties of intercalation complexes of the transition metal dichalcogenides," Adv. Phys. **36**, 1 (1987).

[39] J. Wan, S. D. Lacey, J. Dai, W. Bao, M. S. Fuhrer, and L. Hu, "Tuning two-dimensional nanomaterials by intercalation: materials, properties and applications," Chem. Soc. Rev. **45**, 6742 (2016).

[40] X. Zhao, P. Song, C. Wang, A. C. Riis-Jensen, W. Fu, Y. Deng, D. Wan, L. Kang, S. Ning, J. Dan, et al., "Engineering covalently bonded 2D layered materials by self-intercalation," Nature **581**, 171 (2020).

[41] J. R. Schaibley, H. Yu, G. Clark, P. Rivera, J. S. Ross, K. L. Seyler, W. Yao, and X. Xu, "Valleytronics in 2D materials," Nat. Rev. Mater. **1**, 1 (2016).

[42] K. F. Mak and J. Shan, "Photonics and optoelectronics of 2D semiconductor transition metal dichalcogenides," Nat. Photonics **10**, 216 (2016).

[43] X. Liu and M. C. Hersam, "2D materials for quantum information science," Nat. Rev. Mater. **4**, 669 (2019).

[44] M. Toth and I. Aharonovich, "Single photon sources in atomically thin materials," Annu. Rev. of Phys. Chem. **70**, 123 (2019).

[45] J. F. Janak, "Proof that $\frac{\partial E}{\partial n_i} = \varepsilon$ in density-functional theory," Phys. Rev. B **18**, 7165 (1978).

[46] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, et al., "The Atomic Simulation Environment—a Python library for working with atoms," J. Condens. Matter Phys. **29**, 273002 (2017).

[47] E. Schrödinger, "An undulatory theory of the mechanics of atoms and molecules," Phys. Rev. **28**, 1049 (1926).

[48] K. Capelle, "A bird's-eye view of density-functional theory," Braz. J. Phys. **36**, 1318 (2006).

[49] M. Born and R. Oppenheimer, "Zur Quantentheorie der Molekeln," Ann. Phys. **389**, 457 (1927).

[50] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," Phys. Rev. **136**, B864 (1964).

[51] C. Fiolhais, F. Nogueira, and M. A. Marques, *A primer in density functional theory*, Vol. 620 (Springer Science & Business Media, 2003).

[52] M. Levy, "Electron densities in search of Hamiltonians," Phys. Rev. A **26**, 1200 (1982).

[53] R. M. Dreizler and J. da Providência, *Density functional methods in physics*, Vol. 123 (Springer Science & Business Media, 2013).

[54] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," Phys. Rev. **140**, A1133 (1965).

[55] W. Koch and M. C. Holthausen, *A chemist's guide to density functional theory* (John Wiley & Sons, 2015).

[56] R. G. Parr, "Density functional theory of atoms and molecules," in *Horizons of quantum chemistry* (Springer, 1980), pages 5–15.

[57] D. M. Ceperley and B. J. Alder, "Ground state of the electron gas by a stochastic method," Phys. Rev. Lett. **45**, 566 (1980).

[58] S. H. Vosko, L. Wilk, and M. Nusair, "Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis," Can. J. Phys. **58**, 1200 (1980).

[59] J. Perdew, E. McMullen, and A. Zunger, "Density-functional theory of the correlation energy in atoms and ions: a simple analytic model and a challenge," Phys. Rev. A **23**, 2785 (1981).

[60] J. P. Perdew and Y. Wang, "Accurate and simple analytic representation of the electron-gas correlation energy," Phys. Rev. B **45**, 13244 (1992).

[61] R. O. Jones and O. Gunnarsson, "The density functional formalism, its applications and prospects," Rev. Mod. Phys. **61**, 689 (1989).

[62]   J. P. Perdew, "Density functional theory and the band gap problem," Int. J. Quantum Chem. **28**, 497 (1985).

[63]   J. P. Perdew, "Density-functional approximation for the correlation energy of the inhomogeneous electron gas," Phys. Rev. B **33**, 8822 (1986).

[64]   J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865 (1996).

[65]   A. D. Becke, "A new mixing of Hartree-Fock and local density-functional theories," J. Chem. Phys. **98**, 1372 (1993).

[66]   J. P. Perdew, M. Ernzerhof, and K. Burke, "Rationale for mixing exact exchange with density functional approximations," J. Chem. Phys. **105**, 9982 (1996).

[67]   C. Adamo and V. Barone, "Toward reliable density functional methods without adjustable parameters: the PBE0 model," J. Chem. Phys. **110**, 6158 (1999).

[68]   J. Heyd, G. E. Scuseria, and M. Ernzerhof, "Hybrid functionals based on a screened coulomb potential," J. Chem. Phys. **118**, 8207 (2003).

[69]   J. Paier, M. Marsman, K. Hummer, G. Kresse, I. C. Gerber, and J. G. Ángyán, "Screened hybrid density functionals applied to solids," J. Chem. Phys. **124**, 154709 (2006).

[70]   J. P. Perdew, R. G. Parr, M. Levy, and J. L. Balduz Jr, "Density-functional theory for fractional particle number: derivative discontinuities of the energy," Phys. Rev. Lett. **49**, 1691 (1982).

[71]   H. Eschrig, *The fundamentals of density functional theory*, Vol. 2 (Springer, 2003).

[72]   M. Cococcioni and S. De Gironcoli, "Linear response approach to the calculation of the effective interaction parameters in the LDA+U method," Physical Review B **71**, 035105 (2005).

[73]   J. C. Slater and J. H. Wood, "Statistical exchange and the total energy of a crystal," Int. J. Quantum Chem. **5**, 3 (1970).

[74]   A. R. Williams, R. A. de Groot, and C. B. Sommers, "Generalization of Slater's transition state concept," J. Chem. Phys. **63**, 628 (1975).

[75]   J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. Hansen, et al., "Electronic structure calculations with GPAW: a real-space implementation of the projector augmented-wave method," J. Condens. Matter Phys. **22**, 253202 (2010).

[76]   G. Kresse and J. Furthmüller, "Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set," Phys. Rev. B **54**, 11169 (1996).

[77]   C. Freysoldt, B. Grabowski, T. Hickel, J. Neugebauer, G. Kresse, A. Janotti, and C. G. Van de Walle, "First-principles calculations for point defects in solids," Rev. Mod. Phys. **86**, 253 (2014).

[78]   S. G. Louie, M. Schlüter, J. R. Chelikowsky, and M. L. Cohen, "Self-consistent electronic states for reconstructed Si vacancy models," Phys. Rev. B **13**, 1654 (1976).

[79]   S. Zhang and J. E. Northrup, "Chemical potential dependence of defect forma-
       tion energies in GaAs: Application to Ga self-diffusion," Phys. Rev. Lett. **67**,
       2339 (1991).

[80]   C. G. Van de Walle, D. Laks, G. Neumark, and S. Pantelides, "First-principles
       calculations of solubilities and doping limits: Li, Na, and N in ZnSe," Phys. Rev.
       B **47**, 9425 (1993).

[81]   G. Makov and M. Payne, "Periodic boundary conditions in *ab initio* calcula-
       tions," Phys. Rev. B **51**, 4014 (1995).

[82]   M. Leslie and N. Gillan, "The energy and elastic dipole tensor of defects in ionic
       crystals calculated by the supercell method," J. Phys. C: Solid State Phys. **18**,
       973 (1985).

[83]   A. Walsh, "Correcting the corrections for charged defects in crystals," npj Com-
       put. Mater. **7**, 1 (2021).

[84]   S. Lany and A. Zunger, "Assessment of correction methods for the band-gap
       problem and for finite-size effects in supercell defect calculations: case studies
       for ZnO and GaAs," Phys. Rev. B **78**, 235104 (2008).

[85]   J. Xiao, K. Yang, D. Guo, T. Shen, H.-X. Deng, S.-S. Li, J.-W. Luo, and S.-H.
       Wei, "Realistic dimension-independent approach for charged-defect calculations
       in semiconductors," Phys. Rev. B **101**, 165306 (2020).

[86]   M. C. da Silva, M. Lorke, B. Aradi, M. F. Tabriz, T. Frauenheim, A. Rubio, D.
       Rocca, and P. Deák, "Self-consistent potential correction for charged periodic
       systems," Phys. Rev. Lett. **126**, 076401 (2021).

[87]   C. Freysoldt, J. Neugebauer, and C. G. Van de Walle, "Fully *ab initio* finite-
       size corrections for charged-defect supercell calculations," Phys. Rev. Lett. **102**,
       016402 (2009).

[88]   H.-P. Komsa, T. T. Rantala, and A. Pasquarello, "Finite-size supercell correc-
       tion schemes for charged defect calculations," Phys. Rev. B **86**, 045112 (2012).

[89]   H.-P. Komsa, N. Berseneva, A. V. Krasheninnikov, and R. M. Nieminen, "Charged
       point defects in the flatland: Accurate formation energy calculations in two-
       dimensional materials," Phys. Rev. X **4**, 031044 (2014).

[90]   C. G. Van de Walle and J. Neugebauer, "First-principles calculations for defects
       and impurities: Applications to III-nitrides," J. Appl. Phys. **95**, 3851 (2004).

[91]   K. Reuter and M. Scheffler, "Composition, structure, and stability of RuO_2
       (110) as a function of oxygen pressure," Phys. Rev. B **65**, 035406 (2001).

[92]   S. Sanna, T. Frauenheim, and U. Gerstmann, "Validity of the Slater-Janak
       transition-state model within the LDA+U approach," Phys. Rev. B **78**, 085201
       (2008).

[93]   Y. Li, S. Sanna, and W. G. Schmidt, "Modeling intrinsic defects in $LiNbO_3$
       within the Slater-Janak transition state model," J. Chem. Phys. **140**, 234113
       (2014).

[94]   M. Pandey, F. A. Rasmussen, K. Kuhar, T. Olsen, K. W. Jacobsen, and K. S. Thygesen, "Defect-tolerant monolayer transition metal dichalcogenides," Nano Lett. **16**, 2234 (2016).

[95]   C. Göransson, W. Olovsson, and I. A. Abrikosov, "Numerical investigation of the validity of the Slater-Janak transition-state model in metallic systems," Phys. Rev. B **72**, 134203 (2005).

[96]   P. Mori-Sánchez, A. J. Cohen, and W. Yang, "Localization and delocalization errors in density functional theory and implications for band-gap prediction," Phys. Rev. Lett. **100**, 146401 (2008).

[97]   J. Buckeridge, "Equilibrium point defect and charge carrier concentrations in a material determined through calculation of the self-consistent Fermi energy," Comput. Phys. Commun. **244**, 329 (2019).

[98]   I. Aharonovich, D. Englund, and M. Toth, "Solid-state single-photon emitters," Nat. Photonics **10**, 631 (2016).

[99]   A. Alkauskas, J. L. Lyons, D. Steiauf, and C. G. Van de Walle, "First-principles calculations of luminescence spectrum line shapes for defects in semiconductors: the example of GaN and ZnO," Phys. Rev. Lett. **109**, 267401 (2012).

[100] A. M. Stoneham, *Theory of defects in solids: electronic structure of defects in insulators and semiconductors* (Oxford University Press, 2001).

[101] A. Alkauskas, B. B. Buckley, D. D. Awschalom, and C. G. Van de Walle, "First-principles theory of the luminescence lineshape for the triplet transition in diamond NV centres," New J. Phys. **16**, 073026 (2014).

[102] A. Alkauskas, M. D. McCluskey, and C. G. Van de Walle, "Tutorial: Defects in semiconductors—combining experiment and theory," J. Appl. Phys. **119**, 181101 (2016).

[103] S. Manti, "Effects of lattice imperfections on the optical and electronic properties of two-dimensional materials," PhD thesis (Technial University of Denmark, 2021).

[104] J. N. Eckstein and J. Levy, "Materials issues for quantum computation," MRS Bull. **38**, 783 (2013).

[105] B. Gardas, J. Dziarmaga, W. H. Zurek, and M. Zwolak, "Defects in quantum computers," Sci. Rep. **8**, 1 (2018).

[106] F. J. Heremans, C. G. Yale, and D. D. Awschalom, "Control of spin defects in wide-bandgap semiconductors for quantum technologies," IEEE Proc. **104**, 2009 (2016).

[107] H. Seo, H. Ma, M. Govoni, and G. Galli, "Designing defect-based qubit candidates in wide-gap binary semiconductors for solid-state quantum technologies," Phys. Rev. Mat. **1**, 075002 (2017).

[108] W. F. Koehl, B. B. Buckley, F. J. Heremans, G. Calusine, and D. D. Awschalom, "Room temperature coherent control of defect spin qubits in silicon carbide," Nature **479**, 84 (2011).

[109] L. Childress, M. Gurudev Dutt, J. Taylor, A. Zibrov, F. Jelezko, J. Wrachtrup, P. Hemmer, and M. Lukin, "Coherent dynamics of coupled electron and nuclear spin qubits in diamond," Science **314**, 281 (2006).

[110] Y.-T. Huang, S. R. Kavanagh, D. O. Scanlon, A. Walsh, and R. L. Hoye, "Perovskite-inspired materials for photovoltaics and beyond—from design to devices," Nanotechnology **32**, 132004 (2021).

[111] G. Ceder, Y.-M. Chiang, D. Sadoway, M. Aydinol, Y.-I. Jang, and B. Huang, "Identification of cathode materials for lithium batteries guided by first-principles calculations," Nature **392**, 694 (1998).

[112] S. Wang, Z. Wang, W. Setyawan, N. Mingo, and S. Curtarolo, "Assessing the thermoelectric properties of sintered compounds *via* high-throughput *ab-initio* calculations," Phys. Rev. X **1**, 021012 (2011).

[113] K. Yang, W. Setyawan, S. Wang, M. Buongiorno Nardelli, and S. Curtarolo, "A search model for topological insulators with high-throughput robustness descriptors," Nat. Mater. **11**, 614 (2012).

[114] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design," Nat. Mater. **12**, 191 (2013).

[115] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, and B. Kozinsky, "AiiDA: automated interactive infrastructure and database for computational science," Comp. Mat. Sci. **111**, 218 (2016).

[116] S. P. Huber, S. Zoupanos, M. Uhrin, L. Talirz, L. Kahle, R. Häuselmann, D. Gresch, T. Müller, A. V. Yakutovich, C. W. Andersen, et al., "Aiida 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance," Sci. Data **7**, 1 (2020).

[117] A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, et al., "FireWorks: a dynamic workflow system designed for high-throughput applications," Concurr. Comput. **27**, 5037 (2015).

[118] S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, et al., "AFLOW: an automatic framework for high-throughput materials discovery," Comp. Mat. Sci. **58**, 218 (2012).

[119] R. Armiento et al., *The high-throughput toolkit (httk)*, http://httk.openmaterialsdb.se/, 2019.

[120] R. Armiento, *In machine learning meets quantum physics 377–395*, 2020.

[121] S. Maffioletti and R. Murri, "GC3Pie: a Python framework for high-throughput computing," in Proceedings of the egi community forum (Citeseer, 2012).

[122] D. Broberg, B. Medasani, N. E. Zimmermann, G. Yu, A. Canning, M. Haranczyk, M. Asta, and G. Hautier, "PyCDT: a Python toolkit for modeling point defects in semiconductors and insulators," Comput. Phys. Commun. **226**, 165 (2018).

[123] E. Péan, J. Vidal, S. Jobic, and C. Latouche, "Presentation of the PyDEF post-treatment Python software to compute publishable charts for defect energy formation," Chem. Phys. Lett. **671**, 124 (2017).

[124] J. Davidsson, V. Ivády, R. Armiento, and I. A. Abrikosov, "ADAQ: automatic workflows for magneto-optical properties of point defects in semiconductors," Comput. Phys. Commun. **269**, 108091 (2021).

[125] A. Goyal, P. Gorai, H. Peng, S. Lany, and V. Stevanović, "A computational framework for automation of point defect calculations," Comput. Mater. Sci. **130**, 1 (2017).

[126] K. Yim, J. Lee, D. Lee, M. Lee, E. Cho, H. S. Lee, H.-H. Nahm, and S. Han, "Property database for single-element doping in ZnO obtained by automated first-principles calculations," Sci. Rep. **7**, 1 (2017).

[127] J. J. Mortensen, M. Gjerding, and K. S. Thygesen, "MyQueue: task and work-flow scheduling system," J. Open Source Softw. **5**, 1844 (2020).

[128] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, et al., "Commentary: The Materials Project: a materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).

[129] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: the Open Quantum Materials Database (OQMD)," JOM **65**, 1501 (2013).

[130] S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, et al., "The Computational 2D Materials Database: high-throughput modeling and discovery of atomically thin crystals," 2D Mater. **5**, 042002 (2018).

[131] N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I. E. Castelli, A. Cepellotti, G. Pizzi, et al., "Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds," Nat. Nanotechnol. **13**, 246 (2018).

[132] P. E. Blöchl, "First-principles calculations of defects in oxygen-deficient silica exposed to hydrogen," Phys. Rev. B **62**, 6158 (2000).

[133] O. V. Yazyev, I. Tavernelli, L. Helm, and U. Röthlisberger, "Core spin-polarization correction in pseudopotential-based electronic structure calculations," Phys. Rev. B **71**, 115110 (2005).

[134] J. Davidsson, "Color centers in semiconductors for quantum applications - a high-throughput search of point defects in SiC," PhD thesis (Linköping University, 2021).

[135] C. F. Powell, J. H. Oxley, J. M. Blocher, and J. Klerer, "Vapor deposition," J. Electrochem. Soc. **113**, 266C (1966).

[136] D. H. Lowndes, D. Geohegan, A. Puretzky, D. Norton, and C. Rouleau, "Synthesis of novel thin-film materials by pulsed laser deposition," Science **273**, 898 (1996).

[137] D. Dijkkamp, T. Venkatesan, X. Wu, S. Shaheen, N. Jisrawi, Y. Min-Lee, W. McLean, and M. Croft, "Preparation of Y-Ba-Cu oxide superconductor thin films using pulsed laser evaporation from high T_c bulk material," Applied Physics Letters **51**, 619 (1987).

[138] D. B. Chrisey, G. K. Hubler, et al., "Pulsed laser deposition of thin films," (1994).

[139] J. Madsen, P. Liu, J. B. Wagner, T. W. Hansen, and J. Schiøtz, "Accuracy of surface strain measurements from transmission electron microscopy images of nanoparticles," Adv. Struct. Chem. Imaging. **3**, 1 (2017).

[140] C. T. Koch, *Determination of core structure periodicity and point defect density along dislocations* (Arizona State University, 2002).

[141] W. Zhou, X. Zou, S. Najmaei, Z. Liu, Y. Shi, J. Kong, J. Lou, P. M. Ajayan, B. I. Yakobson, and J.-C. Idrobo, "Intrinsic structural defects in monolayer molybdenum disulfide," Nano Lett. **13**, 2615 (2013).

[142] J. Hong, Z. Hu, M. Probert, K. Li, D. Lv, X. Yang, L. Gu, N. Mao, Q. Feng, L. Xie, et al., "Exploring atomic defects in molybdenum disulphide monolayers," Nat. Commun. **6**, 1 (2015).

[143] A. M. Van Der Zande, P. Y. Huang, D. A. Chenet, T. C. Berkelbach, Y. You, G.-H. Lee, T. F. Heinz, D. R. Reichman, D. A. Muller, and J. C. Hone, "Grains and grain boundaries in highly crystalline monolayer molybdenum disulphide," Nat. Mater. **12**, 554 (2013).

[144] H.-P. Komsa and A. V. Krasheninnikov, "Native defects in bulk and monolayer MoS_2 from first principles," Phys. Rev. B **91**, 125304 (2015).

[145] G. Levi, A. V. Ivanov, and H. Jónsson, "Variational calculations of excited states via direct optimization of the orbitals in DFT," Faraday Discuss. **224**, 448 (2020).

[146] E. Fortunato, P. Barquinha, and R. Martins, "Oxide semiconductor thin-film transistors: a review of recent advances," Adv. Mater. **24**, 2945 (2012).

[147] R. Kormath Madam Raghupathy, H. Wiebeler, T. D. Kühne, C. Felser, and H. Mirhosseini, "Database screening of ternary chalcogenides for p-type transparent conductors," Chem. Mater. **30**, 6794 (2018).

[148] J. Xu, J.-B. Liu, J. Wang, B.-X. Liu, and B. Huang, "Prediction of novel p-type transparent conductors in layered double perovskites: a first-principles study," Adv. Funct. Mater. **28**, 1800332 (2018).

[149] A. Bhatia, G. Hautier, T. Nilgianskul, A. Miglio, J. Sun, H. J. Kim, K. H. Kim, S. Chen, G.-M. Rignanese, X. Gonze, et al., "High-mobility bismuth-based transparent p-type oxide from high-throughput material screening," Chem. Mater. **28**, 30 (2016).

[150] S.-C. Lu and J.-P. Leburton, "Electronic structures of defects and magnetic impurities in MoS_2 monolayers," Nanoscale Res. Lett. **9**, 1 (2014).

[151] N. Onofrio, D. Guzman, and A. Strachan, "Novel doping alternatives for single-layer transition metal dichalcogenides," J. Appl. Phys. **122**, 185102 (2017).

[152] A. N. Andriotis and M. Menon, "Tunable magnetic properties of transition metal doped MoS_2," Phys. Rev. B **90**, 125304 (2014).

[153] P. Rastogi, S. Kumar, S. Bhowmick, A. Agarwal, and Y. S. Chauhan, "Doping strategies for monolayer MoS_2 via surface adsorption: a systematic study," J. Phys. Chem. C **118**, 30309 (2014).

[154] Z. Fu, B. Yang, N. Zhang, Z. Lu, Z. Yang, and D. Ma, "Tuning the physical and chemical properties of 2D InSe with interstitial boron doping: a first-principles study," J. Phys. Chem. C **121**, 28312 (2017).

[155] P. M. Coelho, H.-P. Komsa, K. Lasek, V. Kalappattil, J. Karthikeyan, M.-H. Phan, A. V. Krasheninnikov, and M. Batzill, "Room-temperature ferromagnetism in $MoTe_2$ by post-growth incorporation of vanadium impurities," Adv. Electron. Mater. **5**, 1900044 (2019).

[156] J. Karthikeyan, H.-P. Komsa, M. Batzill, and A. V. Krasheninnikov, "Which transition metal atoms can be embedded into two-dimensional molybdenum dichalcogenides and add magnetism?" Nano Lett. **19**, 4581 (2019).

[157] O. Lehtinen, N. Vats, G. Algara-Siller, P. Knyrim, and U. Kaiser, "Implantation and atomic-scale investigation of self-interstitials in graphene," Nano Lett. **15**, 235 (2015).

[158] G. Kresse and J. Furthmüller, "Efficiency of *ab-initio* total energy calculations for metals and semiconductors using a plane-wave basis set," Comput. Mat. Sci. **6**, 15 (1996).

[159] S. Manti, M. K. Svendsen, N. R. Knøsgaard, P. M. Lyngby, and K. S. Thygesen, "Predicting and machine learning structural instabilities in 2D materials," arXiv preprint arXiv:2201.08091 (2022).

[160] A. Rasmussen, T. Deilmann, and K. S. Thygesen, "Towards fully automated GW band structure calculations: What we can learn from 60.000 self-energy evaluations," npj Comput. Mater. **7**, 1 (2021).

[161] T. Olsen, E. Andersen, T. Okugawa, D. Torelli, T. Deilmann, and K. S. Thygesen, "Discovering two-dimensional topological insulators from high-throughput computations," Phys. Rev. Mater. **3**, 024005 (2019).

[162] D. Torelli, K. S. Thygesen, and T. Olsen, "High throughput computational screening for 2D ferromagnetic materials: the critical role of anisotropy and local correlations," 2D Mate. **6**, 045018 (2019).

[163] D. Torelli and T. Olsen, "First principles heisenberg models of 2D magnetic materials: the importance of quantum corrections to the exchange coupling," J. Condens. Matter Phys. **32**, 335802 (2020).

[164] A. Taghizadeh, U. Leffers, T. G. Pedersen, and K. S. Thygesen, "A library of *ab initio* Raman spectra for automated identification of 2D materials," Nat. Commun. **11**, 1 (2020).

[165] A. Taghizadeh, K. S. Thygesen, and T. G. Pedersen, "Two-dimensional materials with giant optical nonlinearities near the theoretical upper limit," ACS Nano **15**, 7155 (2021).

[166]   A. C. Riis-Jensen, T. Deilmann, T. Olsen, and K. S. Thygesen, "Classifying the electronic and optical properties of Janus monolayers," ACS Nano **13**, 13354 (2019).

[167]   D. Torelli, H. Moustafa, K. W. Jacobsen, and T. Olsen, "High-throughput computational screening for two-dimensional magnetic materials based on experimental databases of three-dimensional compounds," npj Comput. Mater. **6**, 1 (2020).

[168]   P. M. Larsen, M. Pandey, M. Strange, and K. W. Jacobsen, "Definition of a scoring parameter to identify low-dimensional materials components," Phys. Rev. Mater. **3**, 034003 (2019).

[169]   N. R. Knøsgaard and K. S. Thygesen, "Representing individual electronic states for machine learning GW band structures of 2D materials," Nat. Commun. **13**, 1 (2022).

[170]   M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., "The FAIR Guiding Principles for scientific data management and stewardship," Sci. Data **3**, 1 (2016).

[171]   C. Goble, S. Cohen-Boulakia, S. Soiland-Reyes, D. Garijo, Y. Gil, M. R. Crusoe, K. Peters, and D. Schober, "FAIR computational workflows," Data Intell. **2**, 108 (2020).

[172]   J. L. Lyons and C. G. Van de Walle, "Computationally predicted energies and properties of defects in GaN," npj Comput. Mater. **3**, 1 (2017).

[173]   J. P. Perdew and K. Schmidt, "Jacob's ladder of density functional approximations for the exchange-correlation energy," in AIP Conference Proceedings, Vol. 577, 1 (American Institute of Physics, 2001), pages 1–20.

[174]   C. G. Van de Walle and A. Janotti, "Advances in electronic structure methods for defects and impurities in solids," pss (b) **248**, 19 (2011).

[175]   X. Zhang, M. E. Turiansky, and C. G. Van de Walle, "Correctly assessing defect tolerance in halide perovskites," J. Phys. Chem. C **124**, 6022 (2020).

[176]   H.-P. Komsa and A. Pasquarello, "Assessing the accuracy of hybrid functionals in the determination of defect levels: application to the As antisite in GaAs," Phys. Rev. B **84**, 075207 (2011).

[177]   P. Rinke, A. Janotti, M. Scheffler, and C. G. Van de Walle, "Defect formation energies without the band-gap problem: combining density-functional theory and the GW approach for the silicon self-interstitial," Phys. Rev. Lett. **102**, 026402 (2009).

[178]   W. Chen and A. Pasquarello, "Accuracy of GW for calculating defect energy levels in solids," Phys. Rev. B **96**, 020101 (2017).

[179]   M. Hedström, A. Schindlmayr, G. Schwarz, and M. Scheffler, "Quasiparticle corrections to the electronic properties of anion vacancies at GaAs (110) and InP (110)," Phys. Rev. Lett. **97**, 226401 (2006).

[180] Y. Ma, M. Rohlfing, and A. Gali, "Excited states of the negatively charged nitrogen-vacancy color center in diamond," Phys. Rev. B **81**, 041204 (2010).

[181] M. Bockstedte, F. Schütz, T. Garratt, V. Ivády, and A. Gali, "*ab initio* description of highly correlated states in defects for realizing quantum bits," npj Quantum Mater. **3**, 1 (2018).

[182] D. Zgid, E. Gull, and G. K.-L. Chan, "Truncated configuration interaction expansions as solvers for correlated quantum impurity models and dynamical mean-field theory," Phys. Rev. B **86**, 165128 (2012).

[183] P. Delaney, J. C. Greer, and J. A. Larsson, "Spin-polarization mechanisms of the nitrogen-vacancy center in diamond," Nano Lett. **10**, 610 (2010).

[184] L. Hedin, "New method for calculating the one-particle Green's function with application to the electron-gas problem," Phys. Rev. **139**, A796 (1965).

[185] A. Ruban, H. L. Skriver, and J. K. Nørskov, "Surface segregation energies in transition-metal alloys," Phys. Rev. B **59**, 15990 (1999).

[186] N. C. Frey, D. Akinwande, D. Jariwala, and V. B. Shenoy, "Machine learning-enabled design of point defects in 2D materials for quantum and neuromorphic information processing," ACS Nano **14**, 13406 (2020).

[187] J. B. Varley, A. Samanta, and V. Lordi, "Descriptor-based approach for the prediction of cation vacancy formation energies and transition levels," J. Phys. Chem. Lett. **8**, 5059 (2017).

[188] S. A. Miller, M. Dylla, S. Anand, K. Gordiz, G. J. Snyder, and E. S. Toberer, "Empirical modeling of dopability in diamond-like semiconductors," npj Comp. Mater. **4**, 1 (2018).

[189] V. Sharma, P. Kumar, P. Dev, and G. Pilania, "Machine learning substitutional defect formation energies in ABO_3 perovskites," J. Appl. Phys. **128**, 034902 (2020).

The effect of disorder in materials always had a fundamental impact on society and technology. For example, a large crack in our car window negatively impacts its structural stability and we are forced to replace it with a new and perfect, *i.e.* pristine, window. But disorder can also be the cause for desirable property changes: the beauty of colored glass in church windows can only be achieved by 'messing up' pristine glass and adding salts, minerals, and other imperfections to the mix.

Characterizing materials on the microscopic level requires describing some of the smallest entities of any solid state system, namely atoms and electrons. In the same way as macroscopic disruptions like cracks affect a system's physical behaviour, imperfections on the atomic level, *i.e.* manipulation/replacement/removal of single atoms in the unperturbed crystal, can alter a host material's pristine properties. Even though those disruptions, *i.e.* defects, on the microscopic scale are so small that they cannot be seen by the human eye, they can still have a major impact on the pristine material itself, for example by negatively affecting the efficiencies of photovoltaic materials. On the contrary, certain manifestations of defects in semiconducting structures can act as realisations of qubits, the quantum mechanical counterpart of regular bits in conventional computers, which have the potential to fundamentally revolutionize the way we solve difficult problems on a computer. Therefore, understanding the fundamental processes involved in the formation of defects in solid state crystals, as well as their impact on a host material's physical behaviour is of utmost importance and technological relevance.

In this thesis, we use computational methods to systematically study and simulate defects in crystalline materials. By understanding small disruptions to the pristine material one inevitably begins to understand the unperturbed system better and, thereby, paves the way towards defect-mediated applications for various technological challenges, *e.g.* the transition towards quantum technologies. In particular, we look at materials within the realm of two-dimensional crystals with thicknesses of only few nanometers, a length scale 100,000 smaller than the width of a hair. With our simulations we (i) help in identifying single defect species experimental samples, (ii) compute thermodynamic, electronic, magneto-optical properties for hundreds of defect systems in experimentally known and theoretically predicted atomically thin materials, and (iii) study adsorption and absorption processes of doped 2D monolayers.

Technical University of Denmark
DTU Physics
Department of Physics

Fysikvej 311
2800 Kongens Lyngby, Denmark
Phone: +45 4525 3344
info@fysik.dtu.dk

www.fysik.dtu.dk

DTU