Chemical Science

ROYAL SOCIETY OF CHEMISTRY

PERSPECTIVE

View Article Online
View Journal

Cite this: DOI: 10.1039/d4sc04064c

All publication charges for this article have been paid for by the Royal Society of Chemistry

Beyond chemical structures: lessons and guiding principles for the next generation of molecular databases

Timo Sommer, †[a] Cian Clarke †[a] and Max García-Melchor *[abc]

Databases of molecules and materials are indispensable for advancing chemical research, especially when enriched with electronic structure information from quantum chemistry methods like density functional theory. In this perspective, we review and analyze the current landscape of materials and molecular databases containing quantum chemical data. Our analysis reveals that the materials community has significantly benefited from data platforms such as the Materials Project, which seamlessly integrate chemical structures, electronic structure data, and open-source software. Conversely, quantum chemical data for molecular systems remains largely fragmented across individual datasets, lacking the comprehensive framework of a unified database. We distilled insights from these existing data resources into seven guiding principles termed QUANTUM, which build upon the foundational FAIR principles of data sharing (Findable, Accessible, Interoperable, and Reusable). These principles are aimed at advancing the development of molecular databases into robust, integrated data platforms. We conclude with an outlook on both short- and long-term objectives, guided by these QUANTUM principles, to foster future advancements in molecular quantum databases and enhance their utility for the research community.

Received 20th June 2024
Accepted 28th November 2024

DOI: 10.1039/d4sc04064c

rsc.li/chemical-science

1 Introduction

The dawn of the information age has profoundly transformed how research data is generated, stored, and disseminated. The advent of the World Wide Web in the late 1980s connected scientists like never before, fostering the expansion of chemical repositories such as the Cambridge Structural Database (CSD).[1,2] Originally established in 1965 as a compendium of published crystallographic data, the CSD has grown significantly since its

[a]School of Chemistry, CRANN and AMBER Research Centres, Trinity College Dublin, College Green, Dublin 2, Ireland. E-mail: maxgarcia@cicenergigune.com
[b]Center for Cooperative Research on Alternative Energy (CIC EnergiGUNE), Basque Research and Technology Alliance (BRTA), Alava Technology Park, Albert Einstein 48, 01510 Vitoria-Gasteiz, Spain
[c]IKERBASQUE, Basque Foundation for Science, Plaza de Euskadi 5, 48009 Bilbao, Spain
† Equal contribution.

Timo Sommer

Timo Sommer is a PhD candidate in computational chemistry under the supervision of Prof. Max García-Melchor at Trinity College Dublin, where he develops computational tools and datasets to screen transition metal complexes as catalysts for the oxygen evolution reaction. He earned his master's degree in Theoretical Physics from the Karlsruhe Institute of Technology, where he focused on data-driven methods to predict the critical temperature of superconductors.

Cian Clarke

Cian graduated from Trinity College Dublin in 2022 with a BA in Chemistry with Molecular Modelling. Soon after Cian joined the group of Prof. Max García-Melchor in Trinity College Dublin and under his supervision is currently pursuing a PhD in computational chemistry. The focus on Cian's research surrounds the development and in silico screening of novel water oxidation catalysts.

inception, now encompassing over 1.25 million curated entries. Similarly, resources like the Crystallographic Open Database (COD)[3–5] repository and the PubChem[6,7] database have enabled scientists to digitally catalogue and explore millions of unique molecules and materials. The emergence of data-driven platforms, notably the Materials Project,[8–11] has marked a significant evolution from traditional data resources to more sophisticated, interconnected platforms.

Quantum chemical (QC) methods, developed in the early 20th century, have empowered researchers to explore and predict the electronic structures of molecules and materials. Foundational approaches such as Hartree–Fock theory and density functional theory (DFT) paved the way for deeper insights into electronic and quantum effects. More advanced methods, including post–Hartree–Fock methods and time-dependent density functional theory (TD-DFT), have enhanced the analysis of electronic excitations and complex spectroscopic properties.[12,13] Additionally, computationally less expensive semi-empirical methods like xTB[14–16] and PM6/PM7 [17–19] have facilitated high-throughput screenings and the manipulation of chemical databases.[20,21] The utility of these databases can be greatly enhanced by the integration of QC data, broadening their applicability across various fields.

However, the accuracy of QC data is inherently dependent on the method and system being modelled. For instance, hybrid functionals in DFT, such as ωB97XD,[22] which include a percentage of Hartree–Fock exchange, are well-suited for reactivity studies involving systems with some electron correlation. Meanwhile, more accurate methods like coupled-cluster (CC) may be required for highly correlated systems. Additionally, the choice of basis set and the inclusion of relativistic effects are crucial considerations, particularly for systems containing heavy elements.[23,24] Thus, benchmarking QC methods against reliable experimental data or higher-level QC calculations is essential for validating predictions. Nevertheless, discrepancies can still arise due to incomplete theoretical models, such as the omission of solvent effects in reaction studies.[23] Furthermore, results obtained from QC calculations at different levels of theory are often not directly comparable, which highlights the need for standardized methodologies and cross-validation strategies.

Despite these challenges, recent advances underscore the potential of integrating QC data with large-scale databases. For example, users of the Materials Project have leveraged its QC data to identify efficient electrocatalysts for CO_2 reduction through active learning,[25] screen solid-state electrolytes for Li-ion batteries,[26] and develop interatomic potentials that accurately predict material properties.[27] Furthermore, specialized datasets like 2DMatpedia,[28,29] a collection of 2D materials, have enabled the development of advanced workflows, such as Gerber *et al.*'s work on predicting the properties of material interfaces.[30] Additionally, chemical data featuring electronic structure information is increasingly employed to train advanced machine learning (ML) algorithms to predict chemically relevant properties, including HOMO–LUMO gaps of molecules and semiconductor bandgaps.[31,32]

As chemical databases evolve, adherence to data management guidelines like the FAIR principles is becoming increasingly important.[33] These principles stipulate that data should be Findable, Accessible, Interoperable, and Reusable. For chemical structure databases, this means indexing entries with unique identifiers and ensuring that data such as molecular mass and formal charge is readily retrievable. Data should also be stored in universally accessible formats such as .xyz or .mol2 for molecular structures, .csv for tabular data, or .gml for graph representations. To promote reuse in subsequent studies, it is essential that data associated with each compound is diverse and abundant, underlying the practical benefits of these principles in modern research.

In this perspective, we review and analyze state-of-the-art QC materials and molecular databases, as well as various related datasets and repositories. These accounts are not intended to provide a holistic evaluation of each database but rather a targeted analysis to learn from their respective merits and limitations. Our review focuses on materials and molecular data resources that are open access, available for download, contain electronic structure information from QC calculations, and exclude macromolecules and reactions. Additionally, while acknowledging the many challenges of implementing and maintaining software and hardware for databases, our work focuses on discussing challenges of molecular and materials databases that are directly relevant to a chemistry audience.

Our analysis reveals that the materials community has benefited immensely from QC databases like the Materials Project, which provides geometric structures, electronic structure data, and associated software under a unified framework. In contrast, while the molecular community relies on several important structural databases and repositories of significant value, these resources would benefit from incorporating QC data and a comprehensive ecosystem of supporting software. Consequently, we propose seven guiding principles for a central molecular QC data platform to support research in the molecular community. These principles build upon the FAIR principles of data management and are collectively referred to as QUANTUM (Fig. 1). Thus, our work discusses key questions for the future development of molecular databases from a chemist's point of view.

Max García-Melchor

Dr Max García-Melchor is an Ikerbasque Research Professor at CIC EnergiGUNE, where he leads the Atomistic & Molecular Modelling for Catalysis group. His research leverages advanced computational methods and artificial intelligence to accelerate the discovery of catalytic systems for sustainable chemical and fuel production. With a PhD in Chemistry from the Universitat Autònoma de Barcelona and over 15 years of experience, he specializes in modelling (electro)catalytic reaction mechanisms and developing rational catalyst design approaches.

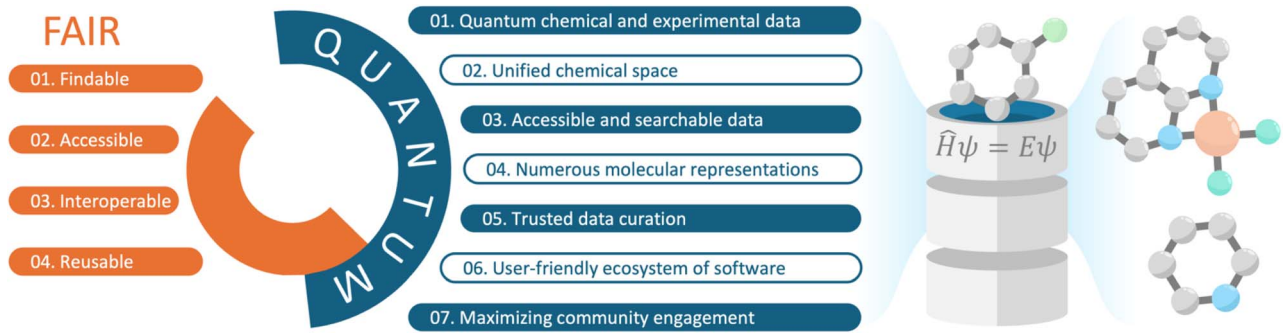

Fig. 1 Graphical summary of the proposed QUANTUM principles. The FAIR principles set the standard for scientific data management and sharing (left). We expand upon FAIR to include the QUANTUM principles (centre), which outline seven design guidelines for developing a QC platform for molecular systems (right).

2 Datasets, repositories and databases

For the purpose of this review, we categorize data resources into three primary groups: datasets, repositories, and databases (Fig. 2). It is important to note that these categories can sometimes overlap.

Datasets are collections of data typically generated and presented by a single set of authors in a publication resulting from a specific research project. Datasets are often formatted as .csv files for tabular data or .json files for more complex data structures, and are commonly uploaded to online portals like Figshare[34] or GitHub.[35] Due to their specific nature, new datasets emerge frequently, reflecting ongoing advancements in research. In this review, we highlight a selection of notable materials and molecular datasets to illustrate their diversity and utility.

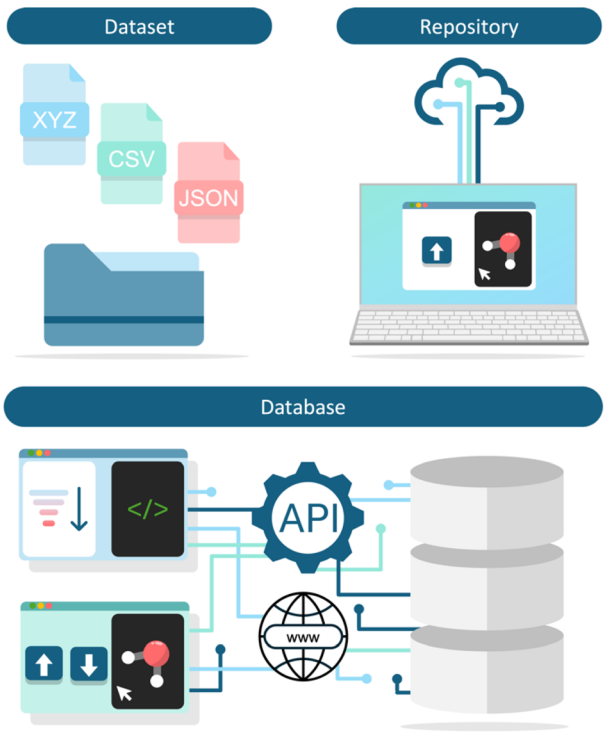

Fig. 2 Overview of different categories of data resources. Datasets comprise data typically formatted as individual .xyz, .csv or .json files, repositories facilitate the online upload and cataloguing of data, and databases allow users to access entries *via* online web interfaces and support advanced querying and connectivity *via* an API.

Repositories allow users to upload material and molecular structure information to an online portal, sharing their results with the broader scientific community. Entries in repositories are typically indexed with a unique identifier, which aids in ensuring traceability and reproducibility in scientific research. Each entry in a repository usually represents one user submission, not one molecule. For instance, ioChem-BD is a web-based repository for chemical structures derived from QC calculations and has many entries where the same chemical structure was calculated with different QC methods.[36,37] While repositories can offer advanced features similar to those found in databases, the wide variety of user submissions can lead to less consistent entries. Another subtype of repository, referred to as dataset repository, does not contain individual molecules or materials but discrete datasets uploaded by various users, for example the Computational Materials Repository.[38,39]

Databases generally differ from datasets and repositories by providing enhanced functionalities that facilitate searching, filtering, and querying entries through user-friendly interfaces (*e.g.* websites), while also being curated and regularly updated. In contrast to repositories, entries in a database usually represent one chemical structure and all data connected to the structure is contained in one entry, such as in the PubChem Compounds database. Databases typically support an application programming interface (API), allowing integration with programming languages such as Python, which fosters a robust ecosystem of software and functionalities for data manipulation and processing. For example, the Materials Project can be easily accessed *via* the Materials Project API.[40] By augmenting data with systems that adhere to the FAIR principles (Findable, Accessible, Interoperable, Reusable), databases significantly increase the impact and utility of their data for the research community. However, developing and maintaining a comprehensive database is often more challenging than creating standalone datasets due to the need for continuous curation and enhancement. In addition to the term database, we will

occasionally use the term platform to emphasize a particularly extensive and well-developed database which contains many different functionalities.

3 Materials data resources

In Table 1, we summarize four major general computational materials databases: AFLOW,[41–43] OQMD,[44–46] the Materials Project, and JARVIS-DFT.[47–49] These databases are centralized, housing large amounts of internally curated data computed predominantly using consistent DFT methods to increase comparability between different entries.

The AFLOW and OQMD stand out for their significantly large sizes, with 3.5M and 1.2M structures, respectively. Many of these are derived from the ICSD,[80,81] a commercial database containing 299k inorganic crystal structures. AFLOW and OQMD further expand their collections by incorporating hypothetical materials, generated by substituting elements in existing structural prototypes, thus extending beyond experimentally confirmed structures.

The JARVIS-DFT database, with 76k structures, distinguishes itself with a diverse range of 3D, 2D, 1D, and 0D materials. This diversity makes it a versatile resource for a broad spectrum of research needs. Moreover, JARVIS-DFT is integrated within the

Table 1 Material databases, datasets, repositories, and dataset repositories that contain QC data. The 'Size' column indicates the number of entries in each data resource. The 'Source' column specifies the origin of the structures

Name	Size	Method	Source	Content
Material databases				
AFLOW[41–43]	3.5M	DFT	ICSD, Pauling File, prototypes	Inorganic bulk materials
OQMD[44–46]	1.2M	DFT	ICSD & prototypes	Inorganic bulk materials
Materials Project[8–11]	1.0M	DFT	ICSD & others	153k bulk materials (main data), and 222k organic molecules, 4k battery materials, 25k battery electrolytes, 20k MOFs, 560k catalyst surfaces, and 41k synthesis recipes
JARVIS-DFT[47–49]	76k	DFT	MP, ICSD, AFLOW, OQMD, COD	3D, 2D, 1D and 0D materials at varying levels of DFT theory
Organic Materials DB[50,51]	41k	DFT	COD	Organic and organometallic materials
Material datasets				
OC20 [52]	1.3M	DFT	MP	Surfaces with N,C,O-containing adsorbates
ARC-MOF[53]	280k	DFT	Multiple papers	MOFs
InterMatch[30]	199k	DFT	MP	Interfaces of materials
Schmidt *et al.*[54]	175k	DFT	MP & others	Chemically diverse bulks
Bare *et al.*[55]	67k	DFT	ABO_3 prototype	ABO_3 perovskite bulks
OC22 [56]	62k	DFT	MP	Surfaces of oxide materials, coverages, and adsorbates
QMOF[57]	20k	DFT	CSD	MOFs
ECD-cubic[58]	17k	DFT	MP	Cubic bulks
2DMatpedia[28,29]	6.4k	DFT	MP	2D materials
Emery & Wolverton[59]	5.3k	DFT	ABO_3 prototype	ABO_3 perovskite bulks
C2DB[60–62]	4.0k	DFT	Prototypes	2D materials
C1DB[63]	820	DFT	ICSD, COD & prototypes	1D materials
Choudhary *et al.*[64]	430	DFT	MP	2D materials
CURATED COFs[65]	308	DFT	Materials Cloud	COFs
Material repositories				
NOMAD[66–69]	12M	DFT & others	Submissions, MP, OQMD, AFLOW, and others	9M bulk crystals, 75k surfaces; 5k 2D, 33k 1D materials, 2.8M organic and inorganic molecules
ioChem-BD[36,37]	356k	DFT	Submissions	38k materials and 318k molecules, chemically diverse
Catalysis-Hub[70,71]	132k	DFT	Submissions	Structures, reaction energies, and barriers for surface reactions, including various tools
Material dataset repositories				
Materials Data Facility[72–74]	>650 sets	Mixed	Mixed	Datasets from publications
MPContribs[75]	45 sets	Mixed	Mixed	Community contributions to MP
Computational Materials Repository[38,39]	31 sets	Mixed	Mixed	Datasets from publications
Materials Cloud[76,77]	17 sets	Mixed	Mixed	Datasets from publications
MatBench[78,79]	13 sets	Mixed	Mixed	Datasets for benchmarking ML algorithms, hosted by MP

JARVIS infrastructure, which includes a force-field database (JARVIS-FF) and ML tools (JARVIS-ML), offering a suite of resources for computational materials science.

The Materials Project database is particularly notable for its extensive and widely used ecosystem of data, functionalities, and Python tools, all integrated into a unified framework. Launched in 2011 as part of the Materials Genome Initiative,[8,10] the Materials Project features a set of 153k bulk materials as its main data resource but has since expanded to include 222k organic molecules, 4k battery materials, 25k battery electrolytes, 20k metal–organic frameworks (MOFs), 560k catalyst surfaces, and 41k synthesis recipes.[9] The Materials Project prioritizes consistency between QC calculations, initially employing only two different DFT methods: PBE+*U* for transition metal oxides and sulfides, and PBE for all other systems.[10] The Materials Project also offers numerous utilities to support research, such as tools for generating phase stability diagrams and Pourbaix diagrams. It has released multiple open-source Python packages like Pymatgen,[82] Atomate,[83] FireWorks,[84] and Custodian.[82] Additionally, community initiatives such as MPContribs,[75] which allows users to contribute their data to existing entries, and MP-Complete,[85] which facilitates submission and voting on new structures, have fostered a collaborative research environment.

In addition to these databases, Table 1 displays three repositories of materials QC data: NOMAD,[66–69] ioChem-BD and Catalysis-Hub. The ioChem-BD contains 38k submissions of QC calculations for materials and 318k submissions for molecules, some of which correspond to identical chemical structures, while Catalysis-Hub also hosts data on surface reactions and provides tools for analysis. The NOMAD, established in 2015, allows uploads from any user employing supported computational chemistry codes and incorporates substantial data from AFLOW, OQMD, and the Materials Project. Adhering firmly to the FAIR principles, NOMAD ensures all data is universally accessible. At present it features 9M bulk materials, 5k 2D materials, 33k 1D materials, 75k surfaces, and a recent addition of 2.8M organic and inorganic molecules. The extensive coverage of NOMAD spans a large chemical space and includes data calculated with a variety of computational codes and methods. To navigate this vast database, the NOMAD website provides advanced tools to query and filter by chemical space, computational QC code, QC methods, applications, or data origin.

Table 1 also lists various materials datasets that cover specific areas of chemical space not extensively detailed in the major databases, such as surfaces, interfaces, MOFs, covalent organic frameworks (COFs), and 1D or 2D materials. Moreover, dataset repositories such as the Materials Data Facility,[72–74] MPContribs,[75] Computational Materials Repository,[38,39] Materials Cloud,[76,77] and MatBench[78,79] compile individual materials datasets, facilitating broader access to diverse data.

From Table 1 we can also observe that 7 out of the 14 materials datasets have been generated using and manipulating structures from the Materials Project (MP). The remaining datasets include hypothetical structures or materials from distinct chemical spaces not present in the Materials Project at the time of publication, such as MOFs or COFs. This underscores the significant impact of the Materials Project as a trusted resource, frequently used for downstream research projects. The Materials Project's ecosystem of functionalities and Python packages supports these projects, promoting widespread community engagement.

Overall, the Materials Project exemplifies the concept of a QC platform, a comprehensive database that integrates structures, electronic structure information, software, and community contributions. This concept is central to our perspective, highlighting the substantial benefits the Materials Project provides to the materials community. By promoting a robust ecosystem where data is consistently curated, easily accessible, and actively contributed to by researchers worldwide, the Materials Project not only serves as a vital resource but also accelerates scientific breakthroughs and innovation in materials science.

4 Molecular data resources

The molecular research community benefits from several important databases and repositories that strongly support data sharing and collaboration. These resources provide comprehensive structural data for each entry but typically lack QC information. Table 2 presents a selection of the most prominent

Table 2 Prominent molecular databases and repositories without QC data. All of them contain 3D structural information

Name	Size	Content
Molecular databases		
HugeMDB[86]	1.7B	Conformers of molecules from PubChem
ZINC20 [87,88]	230M	Commercially available compounds
ChemSpider[89,90]	129M	Chemically diverse molecules
PubChem[6,7]	118M	Chemically diverse molecules
ChemDB[91,92]	5.0M	Small commercially available molecules
ChEMBL[93,94]	2.0M	Bioactive molecules
[a]DrugBank[95,96]	500k	Pharmaceuticals
COCONUT[97,98]	400k	Natural products
Molecular repositories		
[a]CSD[1,2]	1.0M	Small and medium sized organic and inorganic crystallized molecules
COD[3–5]	514k	Crystal structures of organic, inorganic, organometallic compounds and minerals, excluding biopolymers

[a] Not fully open access.

molecular structure databases and repositories that do not include QC data. Several of these, like the PubChem database and the CSD repository, are widely used resources in the molecular community, supporting various applications that require molecular structures. However, the absence of electronic structure information limits their broader utility, especially in data-driven applications.

To address this limitation, Nakata and Shimazaki created the PubChemQC dataset by computing QC properties for 94% of all molecules present in the PubChem database as of August 2016.[21,99,100] While this effort added significant value, the dataset remains separate from the PubChem database and does not integrate with its search and API functionalities. This separation restricts users, especially in fields like organic photovoltaics, from querying PubChem for molecules with specific HOMO–LUMO gaps.

Table 3 provides an overview of molecular databases, datasets, and repositories that include electronic structure data. While there are multiple comprehensive datasets for monometallic transition metal complexes (TMCs) like the tmQMg[133,134] and datasets of extracted ligands,[136,143–145] data for other classes of inorganic molecules are less commonly provided. Among datasets containing both organic and inorganic molecules, the PubChemQC dataset covers the largest chemical space by far. Other datasets are either small in scale or contain a large number of data points for a small number of species, such as the DES370K.[106] Additionally, these datasets are predominantly focused on organic molecules, with fewer entries for inorganic compounds. Other significant sources of electronic structure data including both organic and inorganic molecules are the two repositories ioChem-BD and NOMAD. While the ioChem-BD features 318k user-submitted QC calculations for chemically diverse molecules, the NOMAD contains the largest number of entries among all molecular data resources, featuring 2.8M organic and inorganic molecules. However, despite these large numbers, the decentralized nature of the ioChem-BD and the NOMAD and the diversity of their entries introduce challenges, such as susceptibility to human errors and inconsistencies, which can complicate downstream research.

For organic molecules, significant efforts have been made to generate extensive datasets with QC information. One of the pioneering examples is the QM9 dataset, which includes DFT properties for all 134k enumerated molecules with up to nine heavy atoms within the chemical space of C, H, O, N, and F.[123,146] Other datasets provide electronic structure data for various molecular conformers, non-equilibrium geometries, and open-shell molecules.[111,115–117,120]

Despite the generation of substantial electronic structure data for predominantly organic molecules, this valuable information largely remains outside the framework of a comprehensive database. The Clean Energy Project Database (CEPDB)[101,102] contains 2.3M organic photovoltaic candidates while the Organic Crystals in Electronic and Light-Oriented Technologies (OCELOT)[103,104] database contains 56k crystalline organic semiconductors, making both large but specialized databases. Currently, the Materials Project is the only major general database that includes molecules with enriched QC properties.[101,102] Initially focused on materials, the Materials Project has since begun expanding to include molecules. It currently contains 222k organic molecules, with plans to include inorganic molecules in the future.[11] However, the Materials Project and its ecosystem remain primarily oriented towards materials, affecting its adoption by the molecular research community.

Despite the inclusion of both structural and QC information in the Materials Project database and the NOMAD repository, neither resource is optimized for molecular applications. Widely used molecular repositories such as the CSD and COD still lack electronic structure information. This gap underscores a critical need for a dedicated molecular QC platform, which could significantly enhance research capabilities in fields ranging from pharmaceuticals to organic electronics.

5 Guiding principles for a unified molecular quantum database

Analyzing and comparing the existing materials and molecular databases summarized in Tables 1–3 reveals a significant disparity between the two research communities. The materials community benefits immensely from the Materials Project, a robust QC platform that integrates extensive data, advanced functionalities, and active community engagement. In stark contrast, the molecular community lacks an equivalent comprehensive platform. This gap is further emphasized by the recent expansions of the Materials Project database and the NOMAD repository to incorporate molecular systems, even though both remain primarily focused on materials.

Despite our initial classification of dataset, database, repository, and dataset repository, these distinctions are not always well-defined, especially between a database and a repository. For example, the NOMAD is considered a repository because it collects QC data from many different sources, but it also incorporates data from databases like the Materials Project and features an advanced user interface. While the Materials Project is classified as a database due to its mostly centralized data generation, it also functions as a repository by collecting experimental and computational community data *via* MPContribs.[147] Therefore, a key consideration for developing molecular QC databases is what balance of in-house data generation, curation, and user contribution is novel and needed in the molecular community. In this view, while there are already two major QC repositories for molecular data, the ioChem-BD and the NOMAD, the Materials Project is the only general QC database containing molecular structures. However, these are only a recent addition and are currently limited to organic molecules. Thus, there is a significant opportunity within the molecular community for a QC database encompassing not only organic but also inorganic chemistries.

A general QC molecular database would be well-positioned to evolve into a large platform, similar to the Materials Project, but specifically optimized for molecular structures. This platform could support both experimental and QC user-

Table 3 Molecular databases, datasets, repositories and dataset repositories that contain QC data. The table is divided into six categories, describing the type of data resource (database, dataset, repository, dataset repository) and the chemical space covered (organic, organic and inorganic, transition metal complexes). An '-sp' in the 'Method' column denotes single-point calculations, often preceded by a geometry relaxation using a less computationally intensive method, such as xTB. Computational methods mentioned: semi-empirical (xTB and PM6/PM7), Hartree–Fock, DFT, TD-DFT, Gaussian-4 theory using second-order Møller–Plesset perturbation theory (G4MP2), complete active space self-consistent field (CASSCF), and coupled-cluster (CC)

Name	Size	Method	Source	Content
Organic molecular databases				
CEPDB[101,102]	2.3M	DFT	Enumerated	Organic compounds for photovoltaics
Materials Project[8–11]	1.0M	DFT	ICSD & others	153k bulk materials (main data), and 222k organic molecules, 4k battery materials, 25k battery electrolytes, 20k MOFs, 560k catalyst surfaces, 41k synthesis recipes
OCELOT[103,104]	56k	DFT	CSD, community	Crystalline organic semiconductors
Organic + inorganic molecular datasets				
PubChemQC[21,99,100]	86M	PM6 + DFT-sp	PubChem	Organic and organometallic molecules containing first-row transition metals
SPICE[105]	1.1M	DFT	Literature, PubChem, DES370K	Conformations of small molecules, dimers, dipeptide, and solvated amino acids
DES370K[106]	370K	DFT + CC-sp	Literature	370k data points of dimer interactions of 392 mostly organic molecules
Alexandria library[107]	2.7k	DFT	PubChem, ChemSpider	Mostly organic molecules
CCCBDB[108]	2.2k	DFT	Literature	Gas-phase atoms and small molecules
QuestDB[109,110]	>500	CC & others	Literature	Vertical excitation energies for small- and medium-sized molecules
Organic molecular datasets				
GEOM[111]	37M	xTB	AICures, QM9	37M conformers of 450k organic molecules
Transition1x[112]	10M	DFT-sp	Grambow *et al.*[113]	Molecular configurations along the potential energy surface of 11 961 reactions
ANI-1x[114]	5.0M	DFT	GDB11, ChEMBL, generated	Small molecules
QM7-X[115]	4.2M	DFT	QM7	Equilibrium and non-equilibrium structures of small organic molecules
QMugs[116]	2.0M	xTB + DFT-sp	ChEMBL	2M conformers of 665K biologically relevant organic molecules
WS22 [117]	1.2M	DFT	Literature	1.2M data points of equilibrium and non-equilibrium geometries of 10 species
VQ24 [118]	836k	DFT & xTB	Generated	Enumerated molecules with up to 5 heavy atoms from C, N, O, F, Si, P, S, Cl, Br
Frag20 [119]	566k	DFT	ZINC, PubChem	Small organic molecules from ZINC and PubChem
ANI-1ccx[114]	500k	DFT + CC-sp	ANI-1x	Subset of ANI-1x recomputed with CC-sp
John *et al.*[120]	240k	DFT	PubChem	Open- and closed-shell small organic molecules
QM-symex[121,122]	173k	DFT & TD-DFT	Generated	Includes point group and excited states of small molecules
QM9 [123]	134k	DFT	GDB-17	Small organic molecules with up to 9 heavy atoms
Kim *et al.*[124]	134k	G4MP2	QM9	Refinement of QM9
Narayanan *et al.*[125]	133k	G4MP2	QM9	Refinement of QM9
FORMED[126]	117k	xTB, DFT-sp & TD-DFT	CSD	Organic molecules from the CSD
OE62 [127]	62k	DFT	CSD	Organic molecules from the CSD
MQMspin[128]	13k	DFT & CASSCF	QM9	Small organic carbene molecules
HOPV15 [129]	6.0k	DFT	Literature	6k conformers of 353 p-type molecules for organic photovoltaics + exp. data
VERDE Materials DB[130,131]	1.8k	DFT	Generated	Light-responsive π-conjugated organic molecules
HAB79 [132]	921	DFT & CASSCF	Literature	Benchmark dataset for DFT
Transition metal complex (TMC) datasets				
tmQM[133]	80k	xTB + DFT-sp	CSD	Monometallic TMCs
tmQMg[134]	60k	DFT	tmQM	Subset of tmQM with full DFT and graphs from natural bond orders
SC1MC-2022 [135]	7.0k	Hartree–Fock	Generated	TMCs assembled from ligands

Table 3 (*Contd.*)

Name	Size	Method	Source	Content
OHLDB[136]	1.4k	DFT	Enumerated	Homoleptic TMCs
divTMC[137]	855	DFT	CSD	Octahedral TMCs assembled from monodentate ligands
16OSTM10 [138]	160	DFT	CSD	Open-shell TMCs for conformer benchmark
ROST61 [139]	61	CC	Literature	Open-shell TMCs for DFT functional benchmark
MOR41 [140]	41	CC	Literature	Closed-shell TMCs for DFT functional benchmark
Organic + inorganic molecular repositories				
NOMAD[66–69]	12M	DFT & others	Submissions, MP, OQMD, AFLOW, and others	9M bulks, 75k surfaces; 5k 2D, 33k 1D materials, 2.8M organic and inorganic molecules
ioChem-BD[36,37]	356k	DFT mixed	Submissions	38k materials and 318k molecules, chemically diverse
Organic + inorganic molecular dataset repositories				
QCarchive[141,142]	47 sets	Mixed	Mixed	Datasets from publications

contributions in the form of analytical spectra such as ultraviolet-visible (UV-Vis) and X-ray diffraction (XRD), as well as QC input and output files. The unification of different chemical systems and the integration of computational and experimental data are central to making data more Findable, Accessible, Interoperable, and Reusable (FAIR). By collecting data in a widely recognized platform, it becomes more visible to researchers across various disciplines and is more likely to be repurposed for different applications. For example, the bulk structures in the Materials Project have been used not only for screening bulk properties, but also as a source for generating surface slabs,[52,56] interfaces,[30] and 2D materials.[28,29,64]

The unification of data within a single platform becomes particularly impactful in the context of ML applications, where large and diverse datasets are essential for training robust models. Notably, ML methods such as transfer learning, multi-task learning, and multi-fidelity learning can leverage heterogeneous data to optimize performance predictions for specific targets. For example, Yamada *et al.* employed transfer and multi-task learning to predict the experimental heat capacity at constant pressure (C_P) for 58 polymers. They pre-trained their model on small organic molecules from the QM9 dataset, utilizing QC calculated heat capacities at constant volume (C_V) rather than experimental C_P values, reducing the mean absolute error (MAE) of predicting the polymeric C_P by 35%.[148] Similarly, Moore *et al.* combined QC and experimental data in a transfer learning framework to predict the experimental HOMO–LUMO gap of 26 commercially available polymer donors, achieving a 72% reduction in root mean squared error compared to DFT predictions.[149]

The potential of ML is further enhanced by multi-fidelity learning, where data of varying reliability, such as calculations performed at multiple levels of theory, is integrated. For instance, Chen *et al.* used multi-fidelity learning to improve predictions of experimental material band gaps by augmenting experimental datasets with QC data derived from the Materials Project at three different levels of DFT theory reducing the MAE by 22%.[150] In each of these studies, a critical yet time-intensive step was the collection and curation of data from multiple sources. A centralized, unified database would have streamlined this process significantly, highlighting the transformative potential of such platforms for accelerating data-driven discoveries.

Despite the potential benefits of unifying data on a single platform, several challenges must be addressed. A significant hurdle is how to incorporate data from different computational and experimental sources in a way that is most useful for users. The Materials Project facilitates this by enabling data annotations *via* MPContribs,[147] while the PubChem handles this issue by identifying new submissions based on their chemical structure and, when possible, linking them to existing entries.[151]

Another challenge in integrating computational and experimental properties involve semantic issues, where properties with similar names may refer to subtly different concepts. For instance, experimental overpotentials in electrocatalysis are referenced to a specific current density,[152] whereas theoretical overpotentials calculated using QC methods are not. These differences need to be clarified for users and can complicate data exchange through standardized, logic-based language (ontologies) such as the PubChemRDF project, which uses ontologies like CHEMINF[153] to express the PubChem knowledge in a consistent and machine-understandable format.[151]

In addition to studying the chemical properties of individual molecules, a major area of interest in chemistry is the interaction between species in chemical reactions, which can be modelled using QC calculations. For instance, the Gibbs energy of H adsorption (ΔG_H) is a QC-derived reaction descriptor for the hydrogen evolution reaction that allows the prediction of catalytic performance. However, such values are not intrinsic to a single molecule and often depend on the properties of multiple molecular structures. Similarly, reaction parameters such as temperature, pressure, reactant concentration, and solvent depend on the conditions of the reaction, not just the individual molecules. Consequently, reactions require different organizational structures, such as those provided by the Open Reaction Database[154] or the Catalysis-Hub repository for surface reactions.[70,71]

Consequently, our review and evaluation of a diverse range of molecular and material data resources have led us to identify seven key principles crucial for establishing a unified molecular QC database. These principles, which we refer to as the QUANTUM principles, are illustrated in Fig. 3. Designed to build upon the foundational FAIR principles, the QUANTUM principles address the unique needs and challenges in realizing a QC platform for the molecular community. While some of these principles are already partially implemented in existing molecular databases, others highlight critical areas requiring further development and innovation.

5.1. Quantum chemical and experimental data

The integration of QC and experimental data into a unified molecular database presents both opportunities and challenges. Ideally, a comprehensive database would include a wide range of experimentally measured properties for each molecule, such as nuclear magnetic resonance (NMR), infrared (IR), and UV-Vis spectroscopic data, and XRD analyses, as well as physical properties like melting point, hardness, and even color. However, obtaining such data consistently across a broad chemical space is challenging. For example, difficulties in crystallization can hinder XRD analysis.[155] Conversely, QC calculations can be applied to a much broader range of systems, offering valuable insights into the electronic structure of molecules. For instance, Kneiding *et al.* computed properties such as HOMO–LUMO gaps, polarizability, dipole moments, and Gibbs energies for 60k transition metal complexes using a variety of DFT methods.[134] The inclusion of QC data in a database is therefore intended to complement experimental data by filling gaps and providing theoretical insights that can enhance our understanding of molecular properties and reactivity.

However, care must be taken when using and creating QC data to ensure that it is appropriate for the corresponding chemical system and balances both speed and accuracy. It can be more beneficial to focus on fewer, high-quality data points at suitable levels of theory than to amass data with methods that may not be well-suited for the intended purpose. On the other hand, ML techniques can leverage data from computationally inexpensive but less precise QC methods and improve their reliability and speed by incorporating either more accurate QC data or experimental data during training.[148–150] These methods, such as multi-fidelity learning, can dramatically enhance the predictive power of models, even when relying on less accurate or incomplete datasets.

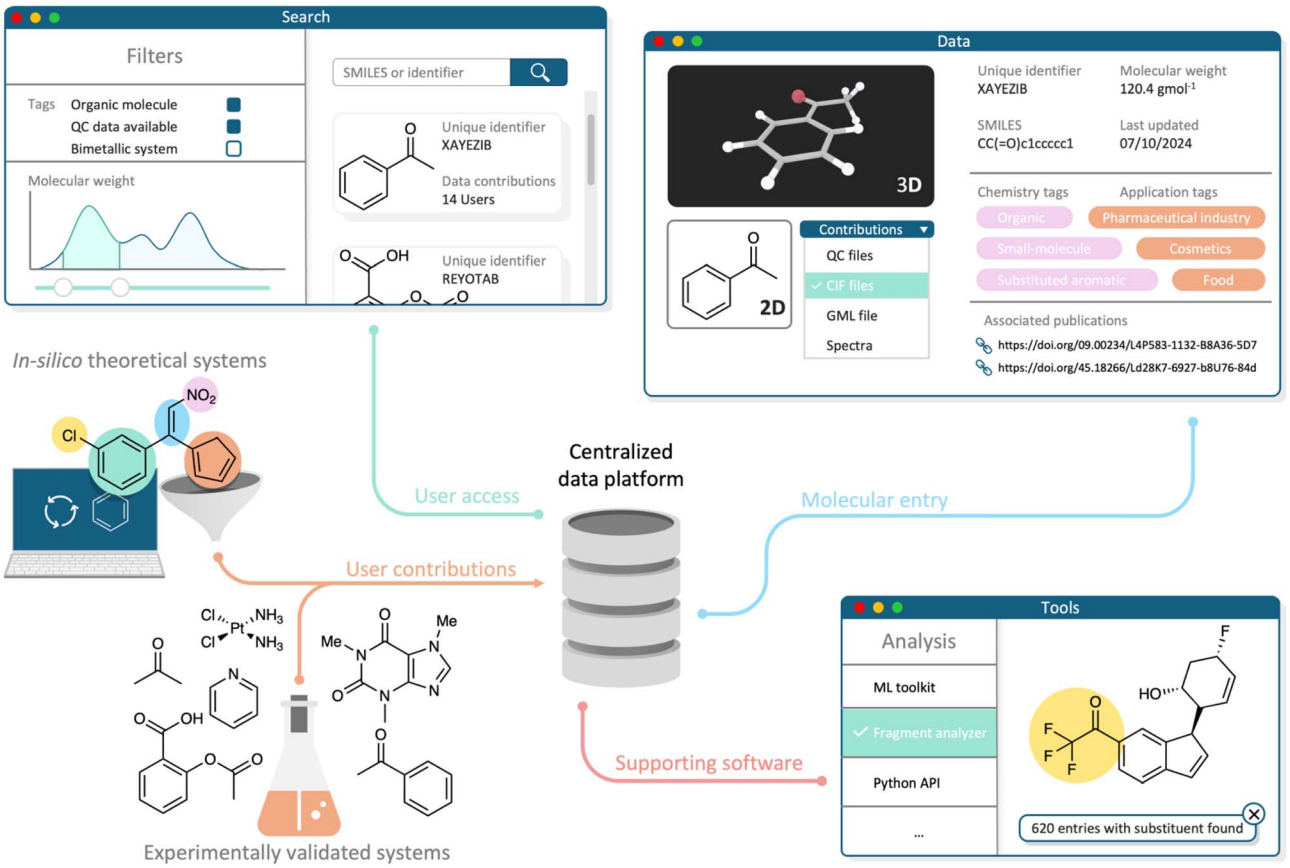

Fig. 3 Schematic overview of a molecular QC platform adhering to the QUANTUM principles. The platform allows users to contribute organic and inorganic molecular structures, either experimentally validated or derived from theoretical studies. A web interface enables users to search and query molecules based on diverse properties or tags. Each entry encompasses multiple attributes, including structural details, QC data, and spectral information. The platform also offers integrated software tools for advanced data queries and analyses.

5.2. Unified chemical space

A comprehensive molecular database would benefit from covering a wide chemical space, including both organic and inorganic molecules, while recognizing that macromolecules may require special considerations. This enables researchers to explore a diverse array of molecular chemistries, including organometallics, TMCs, main-group organic chemistry, as well as molecules used in medicinal chemistry, catalysis, agrochemicals, and beyond, all while using the same database infrastructure. In addition to benefiting data-driven methods such as ML, unifying chemical systems in a single platform enables the reuse of data across various fields of chemistry. For example, the development of cisplatin illustrates how a compound initially observed for inhibiting the cell division of *Escherichia coli* in electrochemical experiments eventually became a widely used chemotherapy drug.[156,157]

Beyond experimentally validated structures, a robust QC platform should also accommodate hypothetical structures generated through various methodologies, such as bottom-up workflows, scaffold diversification inspired by experiments, and generative ML techniques.[158] For example, molSimplify[159] offers a bottom-up approach by assembling monometallic transition metal complexes from a predefined set of ligands. Similarly, Jin *et al.* developed a generative ML model that incrementally constructs organic molecules by predicting substructure connections, enabling the exploration of new chemical spaces.[160]

Evaluating the synthetic feasibility of hypothetical structures is a key challenge, as it involves factors such as byproduct formation, yield, and ease of characterization.[158] Computational tools like MegaSyn address this by assessing synthetic viability of organic molecules, using methodologies that evaluate the relative abundance of synthetically accessible molecular fragments within a given compound.[161] To the same end, the DART platform allows the generation of bottom-up molecular datasets by assembling novel TMCs from ligands in the CSD with established synthetic precedents, aiming to maximize their synthetic viability.[162] These tools help prioritize structures that are more likely to be experimentally realizable, thus streamlining efforts in synthesis and validation.

Nonetheless, hypothetical structures remain valuable even when synthetic feasibility is uncertain. Such systems, especially those with QC data, can serve as training datasets for ML models or as input for high-throughput screenings. By integrating diverse experimental and theoretical molecules from various domains into a unified platform, the QC database can facilitate interdisciplinary innovation, providing access to an expansive and interconnected chemical space.

5.3. Accessible and searchable data

To support public research, the molecular QC platform would benefit from being open access with a modern web interface that facilitates querying and filtering of target molecules. This should include simple descriptors like empirical formula or molecular weight, as well as more complex properties like the HOMO–LUMO gap or sub-structure searches using SMARTS.[163] An API should also be available for programmatic batch access to support data-driven applications and extensive computational analyses.

5.4. Numerous molecular representations

To capture the complexity of chemical structures, the database should support multiple molecular representations that complement each other. This includes 3D structures from experimental XRD and optimized 3D structures from QC calculations. Critical structural details of the 2D molecular graph, such as connectivity and bond orders, commonly represented by SMILES[164] strings, should also be included.

QC calculations also enable the addition of quantitative information such as atomic charges and spins. If necessary, computationally derived bonds and bond orders can also be assigned using methods such as natural bond orbital analysis,[165] as was done for the tmQMg dataset.[134] This data can be useful for example in ML applications as molecular features.

To represent molecular structures numerically, various methods are employed, depending on the desired application. For instance, 3D molecular structures can be encoded into fixed-sized vectors using Smooth Overlap of Atomic Positions (SOAP) features,[166] while 2D molecular graph representations can be expressed either as a fixed-size vector using autocorrelation[167,168] or molecular fingerprints,[169] or they can be used to directly train graph neural networks.[170] Notably, 2D molecular graphs can incorporate geometric properties such as bond distances and QC-derived properties such as atomic charges. However, these fixed-size vector and graph features are typically not stored in databases due to their computational efficiency and dependence on user-defined hyperparameters. Instead, they are often generated on-the-fly using Python packages such as DScribe,[171] RDKit,[172] and molSimplify.[159] This approach ensures flexibility and adaptability, allowing users to tailor the representations to specific tasks or datasets.

In addition to including 3D coordinates and 2D graph representations of a molecule, it can also be beneficial to include data corresponding to the conformational space of a compound. For instance, Eastman *et al.*[105] emphasized the importance of broad conformational datasets, not limited to only the lowest energy conformers, for training ML potentials. They developed the SPICE dataset, which includes 1.1M conformers and trained a set of ML potentials applicable to a broad region of chemical space.

To effectively collate this data, each entry should also have a unique identifier assigned, as SMILES alone is not always sufficient for defining molecules, especially when capturing different conformations of the same molecule. The database should also enable smart data relations between entries, such as identifying isomers or clustering similar molecules. Additionally, tagging molecules with specific applications (*e.g.* organic photovoltaics), as is done in the NOMAD and ioChem-BD, and linking them to related publication DOIs, like in the CSD, could significantly boost research efficiency and breakthroughs.

5.5. Trusted data curation

Ensuring that the molecular QC database is a trusted community resource requires regular curation and updates. Integrating community data consistently within the database framework is essential to maintain its reliability. Both the Materials Project

and the PubChem provide valuable examples of strongly curated databases managing the inclusion of community data. This can also be supported by automated validation and normalization procedures as described for the PubChem.[151]

Especially for QC data, inclusion and curation becomes particularly important due to the large range of QC methods and different requirements for different chemical systems. Thus, a QC molecular platform needs to adopt a consistent framework to accept, process, and display data contributions from the community. The implementation and realization must be considered by the developers of the database, considering the target audience, technical details, and available funding, and cannot be imposed, but can develop over time.

5.6. User-friendly ecosystem of software

Offering user-friendly software and functionalities is essential to create an accessible QC platform. For example, the widely used Python package Pymatgen provides API access to the Materials Project and various tools for analyzing and manipulating materials and molecules. A robust ecosystem of web apps and open-source software enhances the database's utility and promotes community contributions to software, reinforcing the database's status within the community.

5.7. Maximizing community engagement

The ultimate value of a QC platform lies in its frequent use by the scientific community. The Materials Project's most relevant accomplishment is not just the diversity of its data but its status as a trusted and widely used resource. This status was achieved by integrating structural and electronic structure data with extensive open-source software, which mobilized the community to further contribute to data and software, forming a positive feedback loop. To cultivate a similar status, a molecular QC platform needs to engage with the community to meet their needs, incentivize contributions to open-source software, and facilitate the incorporation of data from downstream projects by other researchers.

6 Conclusions and outlook

In this perspective, we have reviewed and analyzed the current landscape of materials and molecular databases, datasets, repositories, and dataset repositories, with a particular focus on those incorporating electronic structure properties from QC calculations. Our analysis highlights the considerable benefits that the materials community has gained from robust QC databases like the Materials Project. This platform seamlessly integrates structural data with consistently calculated electronic structure information and supports a vibrant ecosystem of open-source software, driving downstream research and fostering significant community contributions in data and software development. The success of the Materials Project exemplifies the concept and the potential of a well-integrated QC platform.

In contrast, the molecular community, while leveraging several widely used structural databases and repositories, does not benefit from a dedicated platform that includes both electronic structure information and a comprehensive ecosystem of supporting software. To bridge this gap, we propose the seven QUANTUM principles aimed at developing a unified molecular QC platform. These principles draw inspiration from the diverse databases, datasets, repositories, and dataset repositories reviewed herein. Although our focus is on enhancing molecular databases, the QUANTUM principles also offer valuable insights for advancing existing materials databases. They provide a strategic roadmap for researchers in both the molecular and materials communities to collaborate on improving current databases and identifying critical strategies for future developments.

Significant molecular data resources like the PubChem database and the CSD repository already align with several of the QUANTUM principles. However, the most pressing short-term development we have identified is the integration of electronic structure data from QC calculations into these molecular structural databases. The name QUANTUM is therefore not only intended as an acronym but also as a reflection of the urgency of this particular principle. Meanwhile, platforms like the Materials Project and NOMAD, traditionally focused on materials, are beginning to expand their scope to include molecular data, signaling a major shift towards integrating molecular systems into QC platforms.

Looking ahead, we anticipate significant mid-term progress to emerge from the development of associated software that supports and facilitates community contributions of molecular data. In the long term, we envision the establishment of a unified database that fully adheres to all seven QUANTUM principles, serving as the central QC platform for molecular research. This platform would host a vast array of molecular structures, QC calculations, and experimental properties, underpinned by a comprehensive ecosystem of software and functionalities. It would include a subset of highly curated structures with consistent QC calculations while also acting as a repository for users to submit experimental and computational data.

Once established, we foresee that such a QC platform will revolutionize the field of molecular discovery, mirroring the transformative impact that the Materials Project has had on materials research. We therefore urge the research community to unify their efforts and collaborate in establishing a molecular QC platform that will drive future advancements and innovation in chemistry.

Data availability

Data sharing is not applicable to this manuscript as no datasets were generated or analysed in this perspective.

Author contributions

All authors contributed to the initial conceptualization and the outline of the paper. T. S. and C. C. reviewed and analyzed the existing materials and molecular databases and co-wrote the first draft of the manuscript. M. G.-M. supervised the process in

all stages and reviewed and edited the first draft. All authors contributed to the final version of the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors are very grateful for the financial support provided by the Science Foundation Ireland (SFI-20/FFP-P/8740).

References

1 Cambridge Structural Database, **https://www.ccdc.cam.ac.uk/**, accessed 9 May 2024.
2 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallogr., Sect. B:Struct. Sci.*, 2016, **72**, 171–179.
3 Crystallography Open Database, **https://www.crystallography.net/cod/**, accessed 9 May 2024.
4 S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck and A. Le Bail, *J. Appl. Crystallogr.*, 2009, **42**, 726–729.
5 S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moeck, R. T. Downs and A. Le Bail, *Nucleic Acids Res.*, 2012, **40**, D420–D427.
6 PubChem, **https://pubchem.ncbi.nlm.nih.gov/**, accessed 9 May 2024.
7 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E. E. Bolton, *Nucleic Acids Res.*, 2023, **51**, D1373–D1380.
8 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.
9 Materials Project, **https://next-gen.materialsproject.org/**, accessed 8 May 2024.
10 A. Jain, J. Montoya, S. Dwaraknath, N. E. R. Zimmermann, J. Dagdelen, M. Horton, P. Huck, D. Winston, S. Cholia, S. P. Ong and K. Persson, in *Handbook of Materials Modeling: Methods: Theory and Modeling*, ed. W. Andreoni and S. Yip, Springer International Publishing, Cham, 2020, pp. 1751–1784.
11 E. W. Clark Spotte-Smith, O. Archer Cohen, S. M. Blau, J. M. Munro, R. Yang, R. D. Guha, H. D. Patel, S. Vijay, P. Huck, R. Kingsbury, M. K. Horton and K. A. Persson, *Digital Discovery*, 2023, **2**, 1862–1882.
12 A. Chrostowska and C. Darrigan, in *Organosilicon Compounds*, ed. V. Y. Lee, Academic Press, 2017, pp. 115–166.
13 A. Perera, Y. C. Park and R. J. Bartlett, in *Comprehensive Computational Chemistry*, ed. M. Yáñez and R. J. Boyd, Elsevier, Oxford, 1st edn, 2024, pp. 18–46.
14 S. Grimme, C. Bannwarth and P. Shushkov, *J. Chem. Theory Comput.*, 2017, **13**, 1989–2009.
15 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.
16 C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher and S. Grimme, *Wiley Interdiscip. Rev.:Comput. Mol. Sci.*, 2021, **11**, e1493.
17 J. J. P. Stewart, *J. Mol. Model.*, 2007, **13**, 1173–1213.
18 J. J. P. Stewart, *J. Mol. Model.*, 2013, **19**, 1–32.
19 W. Thiel, *Wiley Interdiscip. Rev.:Comput. Mol. Sci.*, 2014, **4**, 145–157.
20 H. Neugebauer, B. Bädorf, S. Ehlert, A. Hansen and S. Grimme, *J. Comput. Chem.*, 2023, **44**, 2120–2129.
21 M. Nakata, T. Shimazaki, M. Hashimoto and T. Maeda, *J. Chem. Inf. Model.*, 2020, **60**, 5891–5899.
22 J.-D. Chai and M. Head-Gordon, *J. Chem. Phys.*, 2008, **128**, 084106.
23 M. Bursch, J.-M. Mewes, A. Hansen and S. Grimme, *Angew. Chem., Int. Ed.*, 2022, **61**, e202205735.
24 P. J. Hay and W. R. Wadt, *J. Chem. Phys.*, 1985, **82**, 270–283.
25 M. Zhong, K. Tran, Y. Min, C. Wang, Z. Wang, C.-T. Dinh, P. De Luna, Z. Yu, A. S. Rasouli, P. Brodersen, S. Sun, O. Voznyy, C.-S. Tan, M. Askerka, F. Che, M. Liu, A. Seifitokaldani, Y. Pang, S.-C. Lo, A. Ip, Z. Ulissi and E. H. Sargent, *Nature*, 2020, **581**, 178–183.
26 K. Jun, Y. Sun, Y. Xiao, Y. Zeng, R. Kim, H. Kim, L. J. Miara, D. Im, Y. Wang and G. Ceder, *Nat. Mater.*, 2022, **21**, 924–931.
27 C. Chen and S. P. Ong, *Nat. Comput. Sci.*, 2022, **2**, 718–728.
28 J. Zhou, L. Shen, M. D. Costa, K. A. Persson, S. P. Ong, P. Huck, Y. Lu, X. Ma, Y. Chen, H. Tang and Y. P. Feng, *Sci. Data*, 2019, **6**, 86.
29 2D Materials Encyclopedia, **http://www.2dmatpedia.org/**, accessed 8 May 2024.
30 E. Gerber, S. B. Torrisi, S. Shabani, E. Seewald, J. Pack, J. E. Hoffman, C. R. Dean, A. N. Pasupathy and E.-A. Kim, *Nat. Commun.*, 2023, **14**, 7921.
31 F. Zheng, Z. Zhu, J. Lu, Y. Yan, H. Jiang and Q. Sun, *Chem. Phys. Lett.*, 2023, **814**, 140358.
32 F. Dinic, I. Neporozhnii and O. Voznyy, *Comput. Mater. Sci.*, 2024, **231**, 112580.
33 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, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, *Sci. Data*, 2016, **3**, 160018.
34 figshare, **https://figshare.com/**, accessed 13 June 2024.
35 GitHub, **https://github.com**, accessed 13 June 2024.
36 ioChem-BD, **https://www.iochem-bd.org/**, accessed 9 May 2024.
37 M. Álvarez-Moreno, C. de Graaf, N. López, F. Maseras, J. M. Poblet and C. Bo, *J. Chem. Inf. Model.*, 2015, **55**, 95–103.

38 CMR—Computational Materials Repository, https://cmr.fysik.dtu.dk/, accessed 8 May 2024.
39 D. D. Landis, J. S. Hummelshøj, S. Nestorov, J. Greeley, M. Dułak, T. Bligaard, J. K. Nørskov and K. W. Jacobsen, *Comput. Sci. Eng.*, 2012, **14**, 51–57.
40 The Materials Project API, https://next-gen.materialsproject.org/api, accessed 14 October 2024.
41 Aflow – Automatic FLOW for Materials Discovery, https://www.aflowlib.org/, accessed 8 May 2024.
42 S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko and D. Morgan, *Comput. Mater. Sci.*, 2012, **58**, 218–226.
43 M. Esters, C. Oses, S. Divilov, H. Eckert, R. Friedrich, D. Hicks, M. J. Mehl, F. Rose, A. Smolyanyuk, A. Calzolari, X. Campilongo, C. Toher and S. Curtarolo, *Comput. Mater. Sci.*, 2023, **216**, 111808.
44 OQMD, https://oqmd.org/, accessed 8 May 2024.
45 J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, *JOM*, 2013, **65**, 1501–1509.
46 J. Shen, S. D. Griesemer, A. Gopakumar, B. Baldassarri, J. E. Saal, M. Aykol, V. I. Hegde and C. Wolverton, *JPhys Mater.*, 2022, **5**, 031001.
47 NIST-JARVIS, https://jarvis.nist.gov/, accessed 8 May 2024.
48 K. Choudhary, K. F. Garrity, A. C. E. Reid, B. DeCost, A. J. Biacchi, A. R. Hight Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, A. Davydov, J. Jiang, R. Pachter, G. Cheon, E. Reed, A. Agrawal, X. Qian, V. Sharma, H. Zhuang, S. V. Kalinin, B. G. Sumpter, G. Pilania, P. Acar, S. Mandal, K. Haule, D. Vanderbilt, K. Rabe and F. Tavazza, *npj Comput. Mater.*, 2020, **6**, 173.
49 D. Wines, R. Gurunathan, K. F. Garrity, B. DeCost, A. J. Biacchi, F. Tavazza and K. Choudhary, *Appl. Phys. Rev.*, 2023, **10**, 041302.
50 Organic Materials Database, https://omdb.mathub.io/, accessed 8 May 2024.
51 S. S. Borysov, R. M. Geilhufe and A. V. Balatsky, *PLoS One*, 2017, **12**, e0171501.
52 L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, A. Palizhati, A. Sriram, B. Wood, J. Yoon, D. Parikh, C. L. Zitnick and Z. Ulissi, *ACS Catal.*, 2021, **11**, 6059–6072.
53 J. Burner, J. Luo, A. White, A. Mirmiran, O. Kwon, P. G. Boyd, S. Maley, M. Gibaldi, S. Simrod, V. Ogden and T. K. Woo, *Chem. Mater.*, 2023, **35**, 900–916.
54 J. Schmidt, H.-C. Wang, T. F. T. Cerqueira, S. Botti and M. A. L. Marques, *Sci. Data*, 2022, **9**, 64.
55 Z. J. L. Bare, R. J. Morelock and C. B. Musgrave, *Sci. Data*, 2023, **10**, 244.
56 R. Tran, J. Lan, M. Shuaibi, B. M. Wood, S. Goyal, A. Das, J. Heras-Domingo, A. Kolluru, A. Rizvi, N. Shoghi, A. Sriram, F. Therrien, J. Abed, O. Voznyy, E. H. Sargent, Z. Ulissi and C. L. Zitnick, *ACS Catal.*, 2023, **13**, 3066–3084.
57 A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein and R. Q. Snurr, *Matter*, 2021, **4**, 1578–1597.
58 F. Q. Wang, K. Choudhary, Y. Liu, J. Hu and M. Hu, *Sci. Data*, 2022, **9**, 59.
59 A. A. Emery and C. Wolverton, *Sci. Data*, 2017, **4**, 170153.
60 C2DB, https://c2db.fysik.dtu.dk/, accessed 9 May 2024.
61 S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, J. Gath, K. W. Jacobsen, J. J. Mortensen, T. Olsen and K. S. Thygesen, *2D Mater.*, 2018, **5**, 042002.
62 M. N. Gjerding, A. Taghizadeh, A. Rasmussen, S. Ali, F. Bertoldo, T. Deilmann, N. R. Knøsgaard, M. Kruse, A. H. Larsen, S. Manti, T. G. Pedersen, U. Petralanda, T. Skovhus, M. K. Svendsen, J. J. Mortensen, T. Olsen and K. S. Thygesen, *2D Mater.*, 2021, **8**, 044002.
63 H. Moustafa, P. M. Larsen, M. N. Gjerding, J. J. Mortensen, K. S. Thygesen and K. W. Jacobsen, *Phys. Rev. Mater.*, 2022, **6**, 064202.
64 K. Choudhary, I. Kalish, R. Beams and F. Tavazza, *Sci. Rep.*, 2017, **7**, 5179.
65 D. Ongari, A. V. Yakutovich, L. Talirz and B. Smit, *ACS Cent. Sci.*, 2019, **5**, 1663–1675.
66 NOMAD, https://nomad-lab.eu/nomad-lab/, accessed 8 May 2024.
67 C. Draxl and M. Scheffler, *MRS Bull.*, 2018, **43**, 676–682.
68 C. Draxl and M. Scheffler, *JPhys Mater.*, 2019, **2**, 036001.
69 L. Sbailò, Á. Fekete, L. M. Ghiringhelli and M. Scheffler, *npj Comput. Mater.*, 2022, **8**, 1–7.
70 Catalysis-Hub, https://www.catalysis-hub.org/, accessed 8 May 2024.
71 K. T. Winther, M. J. Hoffmann, J. R. Boes, O. Mamun, M. Bajdich and T. Bligaard, *Sci. Data*, 2019, **6**, 75.
72 The Materials Data Facility (MDF), https://materialsdatafacility.org/, accessed 8 May 2024.
73 B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke and I. Foster, *JOM*, 2016, **68**, 2045–2052.
74 B. Blaiszik, L. Ward, M. Schwarting, J. Gaff, R. Chard, D. Pike, K. Chard and I. Foster, *MRS Commun.*, 2019, **9**, 1125–1133.
75 Materials Project, MPContribs Explorer, https://next-gen.materialsproject.org/contribs, accessed 8 May 2024.
76 The Materials Cloud, https://www.materialscloud.org/home, accessed 8 May 2024.
77 L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, C. S. Adorf, C. W. Andersen, O. Schütt, C. A. Pignedoli, D. Passerone, J. VandeVondele, T. C. Schulthess, B. Smit, G. Pizzi and N. Marzari, *Sci. Data*, 2020, **7**, 299.
78 MatBench, https://matbench.materialsproject.org/, accessed 8 May 2024.
79 A. Dunn, Q. Wang, A. Ganose, D. Dopp and A. Jain, *npj Comput. Mater.*, 2020, **6**, 138.
80 ICSD, https://icsd.products.fiz-karlsruhe.de/, accessed 13 May 2024.
81 D. Zagorac, H. Müller, S. Ruehl, J. Zagorac and S. Rehme, *J. Appl. Crystallogr.*, 2019, **52**, 918–925.

82 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Comput. Mater. Sci.*, 2013, **68**, 314–319.
83 K. Mathew, J. H. Montoya, A. Faghaninia, S. Dwarakanath, M. Aykol, H. Tang, I. Chu, T. Smidt, B. Bocklund, M. Horton, J. Dagdelen, B. Wood, Z.-K. Liu, J. Neaton, S. P. Ong, K. Persson and A. Jain, *Comput. Mater. Sci.*, 2017, **139**, 140–152.
84 A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter and K. A. Persson, *Concurr. Comput. Pract. Exp.*, 2015, **27**, 5037–5059.
85 MP-Complete, **https://sciencegateways.org/resources/mp-complete**, accessed 6 October 2024.
86 Huge MDB, **https://www.multi-d.com/**, accessed 9 May 2024.
87 ZINC20, **https://zinc.docking.org/**, accessed 9 May 2024.
88 J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield and R. A. Sayle, *J. Chem. Inf. Model.*, 2020, **60**, 6065–6073.
89 ChemSpider, **https://www.chemspider.com/**, accessed 9 May 2024.
90 H. E. Pence and A. Williams, *J. Chem. Educ.*, 2010, **87**, 1123–1124.
91 ChemDB, **https://cdb.ics.uci.edu/**, accessed 9 May 2024.
92 J. H. Chen, E. Linstead, S. J. Swamidass, D. Wang and P. Baldi, *Bioinformatics*, 2007, **23**, 2348–2351.
93 ChEMBL Database, **https://www.ebi.ac.uk/chembl/**, accessed 9 May 2024.
94 B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, M. P. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A. P. Bento, M. F. Adasme, P. Monecke, G. A. Landrum and A. R. Leach, *Nucleic Acids Res.*, 2024, **52**, D1180–D1192.
95 DrugBank, **https://go.drugbank.com/**, accessed 9 May 2024.
96 D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, *Nucleic Acids Res.*, 2006, **34**, D668–D672.
97 COCONUT: Natural Products Online, **https://coconut.naturalproducts.net/**, accessed 9 May 2024.
98 M. Sorokina, P. Merseburger, K. Rajan, M. A. Yirik and C. Steinbeck, *J. Cheminf.*, 2021, **13**, 2.
99 M. Nakata and T. Maeda, *J. Chem. Inf. Model.*, 2023, **63**, 5734–5754.
100 M. Nakata and T. Shimazaki, *J. Chem. Inf. Model.*, 2017, **57**, 1300–1308.
101 CEPDB, **https://www.molecularspace.org/**, accessed 8 May 2024.
102 J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.
103 OCELOT – Organic Crystals in Electronic and Light-Oriented Technologies, **https://oscar.as.uky.edu/**, accessed 2 October 2024.
104 Q. Ai, V. Bhat, S. M. Ryno, K. Jarolimek, P. Sornberger, A. Smith, M. M. Haley, J. E. Anthony and C. Risko, *J. Chem. Phys.*, 2021, **154**, 174705.
105 P. Eastman, P. K. Behara, D. L. Dotson, R. Galvelis, J. E. Herr, J. T. Horton, Y. Mao, J. D. Chodera, B. P. Pritchard, Y. Wang, G. De Fabritiis and T. E. Markland, *Sci. Data*, 2023, **10**, 11.
106 A. G. Donchev, A. G. Taube, E. Decolvenaere, C. Hargus, R. T. McGibbon, K.-H. Law, B. A. Gregersen, J.-L. Li, K. Palmo, K. Siva, M. Bergdorf, J. L. Klepeis and D. E. Shaw, *Sci. Data*, 2021, **8**, 55.
107 M. M. Ghahremanpour, P. J. van Maaren and D. van der Spoel, *Sci. Data*, 2018, **5**, 180062.
108 NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101, **http://cccbdb.nist.gov/**, accessed 8 May 2024.
109 QUEST: A Database of Highly-Accurate Excitation Energies, **https://lcpq.github.io/QUESTDB_website/**, accessed 8 May 2024.
110 M. Véril, A. Scemama, M. Caffarel, F. Lipparini, M. Boggio-Pasqua, D. Jacquemin and P.-F. Loos, *Wiley Interdiscip. Rev.:Comput. Mol. Sci.*, 2021, **11**, e1517.
111 S. Axelrod and R. Gómez-Bombarelli, *Sci. Data*, 2022, **9**, 185.
112 M. Schreiner, A. Bhowmik, T. Vegge, J. Busk and O. Winther, *Sci. Data*, 2022, **9**, 779.
113 C. A. Grambow, L. Pattanaik and W. H. Green, *Sci. Data*, 2020, **7**, 137.
114 J. S. Smith, R. Zubatyuk, B. Nebgen, N. Lubbers, K. Barros, A. E. Roitberg, O. Isayev and S. Tretiak, *Sci. Data*, 2020, **7**, 134.
115 J. Hoja, L. Medrano Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio Jr and A. Tkatchenko, *Sci. Data*, 2021, **8**, 43.
116 C. Isert, K. Atz, J. Jiménez-Luna and G. Schneider, *Sci. Data*, 2022, **9**, 273.
117 M. Pinheiro Jr, S. Zhang, P. O. Dral and M. Barbatti, *Sci. Data*, 2023, **10**, 95.
118 D. Khan, A. Benali, S. Y. H. Kim, G. F. von Rudorff and O. A. von Lilienfeld, *arXiv*, 2024, preprint, arXiv:2405.05961, DOI: **10.48550/arXiv.2405.05961**.
119 J. Lu, S. Xia, J. Lu and Y. Zhang, *J. Chem. Inf. Model.*, 2021, **61**, 1095–1104.
120 P. C. S. John, Y. Guan, Y. Kim, B. D. Etz, S. Kim and R. S. Paton, *Sci. Data*, 2020, **7**, 244.
121 J. Liang, Y. Xu, R. Liu and X. Zhu, *Sci. Data*, 2019, **6**, 213.
122 J. Liang, S. Ye, T. Dai, Z. Zha, Y. Gao and X. Zhu, *Sci. Data*, 2020, **7**, 400.
123 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *Sci. Data*, 2014, **1**, 140022.
124 H. Kim, J. Y. Park and S. Choi, *Sci. Data*, 2019, **6**, 109.
125 B. Narayanan, P. C. Redfern, R. S. Assary and L. A. Curtiss, *Chem. Sci.*, 2019, **10**, 7449–7455.
126 J. T. Blaskovits, R. Laplaza, S. Vela and C. Corminboeuf, *Adv. Mater.*, 2024, **36**, 2305602.
127 A. Stuke, C. Kunkel, D. Golze, M. Todorović, J. T. Margraf, K. Reuter, P. Rinke and H. Oberhofer, *Sci. Data*, 2020, **7**, 58.

128 M. Schwilk, D. N. Tahchieva and O. A. von Lilienfeld, *arXiv*, 2020, preprint, arXiv:2004.10600, DOI: **10.48550/arXiv.2004.10600**.
129 S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann and A. Aspuru-Guzik, *Sci. Data*, 2016, **3**, 160086.
130 Verdematerials DB, **https://www.verdematerialsdb.com/**, accessed 9 May 2024.
131 B. G. Abreha, S. Agarwal, I. Foster, B. Blaiszik and S. A. Lopez, *J. Phys. Chem. Lett.*, 2019, **10**, 6835–6841.
132 O. G. Ziogos, A. Kubas, Z. Futera, W. Xie, M. Elstner and J. Blumberger, *J. Chem. Phys.*, 2021, **155**, 234115.
133 D. Balcells and B. B. Skjelstad, *J. Chem. Inf. Model.*, 2020, **60**, 6135–6146.
134 H. Kneiding, R. Lukin, L. Lang, S. Reine, T. B. Pedersen, R. D. Bin and D. Balcells, *Digital Discovery*, 2023, **2**, 618–633.
135 P. Golub, P. Beran, A. Antalik and J. Brabec, *arXiv*, 2023, preprint, arXiv:2101.06090, DOI: **10.48550/arXiv.2101.06090**.
136 S. Gugler, J. Paul Janet and H. J. Kulik, *Mol. Syst. Des. Eng.*, 2020, **5**, 139–152.
137 C. Duan, A. J. Ladera, J. C.-L. Liu, M. G. Taylor, I. R. Ariyarathna and H. J. Kulik, *J. Chem. Theory Comput.*, 2022, **18**, 4836–4845.
138 A. A. Otlyotov, A. D. Moshchenkov, L. Cavallo and Y. Minenkov, *Phys. Chem. Chem. Phys.*, 2022, **24**, 17314–17322.
139 L. R. Maurer, M. Bursch, S. Grimme and A. Hansen, *J. Chem. Theory Comput.*, 2021, **17**, 6134–6151.
140 S. Dohm, A. Hansen, M. Steinmetz, S. Grimme and M. P. Checinski, *J. Chem. Theory Comput.*, 2018, **14**, 2596–2608.
141 The MolSSI QCArchive, **https://qcarchive.molssi.org/**, accessed 2 October 2024.
142 D. G. A. Smith, D. Altarawy, L. A. Burns, M. Welborn, L. N. Naden, L. Ward, S. Ellis, B. P. Pritchard and T. D. Crawford, *Wiley Interdiscip. Rev.:Comput. Mol. Sci.*, 2021, **11**, e1491.
143 T. Gensch, G. dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario, M. S. Sigman and A. Aspuru-Guzik, *J. Am. Chem. Soc.*, 2022, **144**, 1205–1217.
144 S.-S. Chen, Z. Meyer, B. Jensen, A. Kraus, A. Lambert and D. H. Ess, *J. Chem. Inf. Model.*, 2023, **63**, 7412–7422.
145 H. Kneiding, A. Nova and D. Balcells, *Nat. Comput. Sci.*, 2024, **4**, 263–273.
146 L. Ruddigkeit, R. van Deursen, L. C. Blum and J.-L. Reymond, *J. Chem. Inf. Model.*, 2012, **52**, 2864–2875.
147 Materials Project, MPContribs Documentation, **https://docs.materialsproject.org/services/mpcontribs**, accessed 10 October 2024.
148 H. Yamada, C. Liu, S. Wu, Y. Koyama, S. Ju, J. Shiomi, J. Morikawa and R. Yoshida, *ACS Cent. Sci.*, 2019, **5**, 1717–1730.
149 G. J. Moore, O. Bardagot and N. Banerji, *Adv. Theory Simul.*, 2022, **5**, 2100511.
150 C. Chen, Y. Zuo, W. Ye, X. Li and S. P. Ong, *Nat. Comput. Sci.*, 2021, **1**, 46–53.
151 G. Fu, C. Batchelor, M. Dumontier, J. Hastings, E. Willighagen and E. Bolton, *J. Cheminf.*, 2015, **7**, 34.
152 A. M. Appel and M. L. Helm, *ACS Catal.*, 2014, **4**, 630–633.
153 J. Hastings, L. Chepelev, E. Willighagen, N. Adams, C. Steinbeck and M. Dumontier, *PLoS One*, 2011, **6**, e25513.
154 S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen and C. W. Coley, *J. Am. Chem. Soc.*, 2021, **143**, 18820–18826.
155 H. Li, Y. Li, J. Jiao and C. Lin, *Results Chem.*, 2023, **5**, 100859.
156 S. Dasari and P. B. Tchounwou, *Eur. J. Pharmacol.*, 2014, **740**, 364–378.
157 B. Rosenberg, L. Van Camp and T. Krigas, *Nature*, 1965, **205**, 698–699.
158 C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay and K. F. Jensen, *Wiley Interdiscip. Rev.:Comput. Mol. Sci.*, 2022, **12**, e1608.
159 E. I. Ioannidis, T. Z. H. Gani and H. J. Kulik, *J. Comput. Chem.*, 2016, **37**, 2106–2117.
160 W. Jin, R. Barzilay and T. Jaakkola, in *Artificial Intelligence in Drug Discovery*, ed. N. Brown, The Royal Society of Chemistry, 2020, pp. 228–249.
161 F. Urbina, C. T. Lowden, J. C. Culberson and S. Ekins, *ACS Omega*, 2022, **7**, 18699–18713.
162 C. Clarke, T. Sommer, F. Kleuker and M. García-Melchor, *ChemRxiv*, 2024, preprint, DOI: **10.26434/chemrxiv-2024-tljj9**.
163 SMARTS – A Language for Describing Molecular Patterns, **https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html**, accessed 21 March 2024.
164 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
165 E. D. Glendening, C. R. Landis and F. Weinhold, *Wiley Interdiscip. Rev.:Comput. Mol. Sci.*, 2012, **2**, 1–42.
166 A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2013, **87**, 184115.
167 J. P. Janet and H. J. Kulik, *J. Phys. Chem. A*, 2017, **121**, 8939–8954.
168 L. Morán-González, J. E. Betten, H. Kneiding and D. Balcells, *ChemRxiv*, 2024, preprint, DOI: **10.26434/chemrxiv-2023-5wbkr-v2**.
169 D. Boldini, D. Ballabio, V. Consonni, R. Todeschini, F. Grisoni and S. A. Sieber, *J. Cheminf.*, 2024, **16**, 35.
170 P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer and P. Friederich, *Commun. Mater.*, 2022, **3**, 1–18.
171 L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, *Comput. Phys. Commun.*, 2020, **247**, 106949.
172 RDKit: Open-source cheminformatics, **https://www.rdkit.org/**, accessed 14 October 2024.