npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Predicting solid state material platforms for quantum technologies

Oliver Lerstøl Hebnes[1,2], Marianne Etzelmüller Bathen[3]✉, Øyvind Sigmundson Schøyen[2], Sebastian G. Winther-Larsen[2,4], Lasse Vines[5] and Morten Hjorth-Jensen[2,6]

Semiconductor materials provide a compelling platform for quantum technologies (QT). However, identifying promising material hosts among the plethora of candidates is a major challenge. Therefore, we have developed a framework for the automated discovery of semiconductor platforms for QT using material informatics and machine learning methods. Different approaches were implemented to label data for training the supervised machine learning (ML) algorithms logistic regression, decision trees, random forests and gradient boosting. We find that an empirical approach relying exclusively on findings from the literature yields a clear separation between predicted suitable and unsuitable candidates. In contrast to expectations from the literature focusing on band gap and ionic character as important properties for QT compatibility, the ML methods highlight features related to symmetry and crystal structure, including bond length, orientation and radial distribution, as influential when predicting a material as suitable for QT.

npj Computational Materials (2022)8:207 ; https://doi.org/10.1038/s41524-022-00888-3

INTRODUCTION

Quantum technologies (QT) based on solid state platforms have attracted a lot of attention during recent years. Among the promising applications that are already available we find important breakthroughs such as in vivo sensing of magnetic fields in cells[1], secure communication over large distances by separation of entangled photons[2] and, notably, various quantum information processing prototypes and architectures[3]. Quantum computers are in high demand to meet the increasing need for computing power to solve complex and high-dimensional scientific problems. Recent advances have highlighted the potential of quantum information processors to outperform state-of-the-art high-performance computing facilities. Indeed, Google's 53 qubit quantum computer based on superconducting electronics solved a computational problem that was beyond the capabilities of a 200000 core supercomputer[3]. Most recently, IBM announced its 127 qubit quantum processor[4]. Simultaneously, the concepts of entanglement and teleportation may eventually facilitate advanced quantum communication protocols such as quantum cryptography and the quantum internet[5], spurring further investigations into technologies based on quantum mechanics.

Several platforms are available for the development of quantum technologies, but the materials and fabrication technologies are less mature than those for, e.g., classical computers and sensors. An important concern in this context is that of scalability. For example, the best performing quantum computer prototypes available today rely on superconducting electronics that require millikelvin temperatures to operate, with the stability of interactions between qubits being an important issue. Instead, semiconductors are emerging as a promising alternative platform, offering competitive characteristics combined with the possibility of room temperature operation and mature and scalable material processing and fabrication.

Quantum technologies based on semiconductors rely on either defects or quantum dots where the latter can be of the self-assembled or nanostructured type[6]. Semiconductor defects can act as single-photon emitters or spin centers and are compatible with the three main QT categories of computing, communication and sensing[7]. These characteristics are most often found for the case of defects that introduce deep energy levels into the semiconductor band gap[8]. So-called deep level defects can trap charge carriers in localized states that are essentially isolated from the surroundings, making them highly suitable for QT due to, e.g., indistinguishable single-photon emission and long spin coherence times. The most well-known quantum compatible defect is the negatively charged nitrogen-vacancy (NV) center in diamond[9], but silicon carbide (SiC) and the various quantum emitters therein are strong contenders for quantum communication purposes especially due to the favorable emission wavelength region in the near infrared coupled with more mature material processing and fabrication (see, e.g., refs. [10–12]). However, semiconductor based QT is still in the early stages, and the issues left to address include identification of suitable host materials and candidate defects, and scalable and reproducible quantum device fabrication. Furthermore, a complete understanding of the requirements for a semiconductor material to manifest quantum compatible properties is lacking, and the selection of known quantum compatible host materials is slim[13,14].

The majority of discoveries of QT compatible characteristics in semiconductors has so far happened by serendipity, and there is an urgent need for a better and more systematic understanding of which material requirements must be met for QT compatible characteristics like single-photon emission and single spin control to manifest. In this context, a framework for dedicated materials search and analysis is needed.

The fourth science paradigm of big data driven science reveals the potential of targeted search for promising material systems in

[1]Sopra Steria, Information Technology and Services, N-4020 Stavanger, Norway. [2]Department of Physics and Center for Computing in Science Education, University of Oslo, N-0316 Oslo, Norway. [3]Advanced Power Semiconductor Laboratory, ETH Zürich, 8092 Zürich, Switzerland. [4]Menon Economics, N-0369 Oslo, Norway. [5]Department of Physics and Center for Materials Science and Nanotechnology, University of Oslo, N-0316 Oslo, Norway. [6]Department of Physics and Astronomy and Facility for Rare Isotope Beams, Michigan State University, East Lansing, MI 48824, USA. ✉email: bathen@aps.ee.ethz.ch

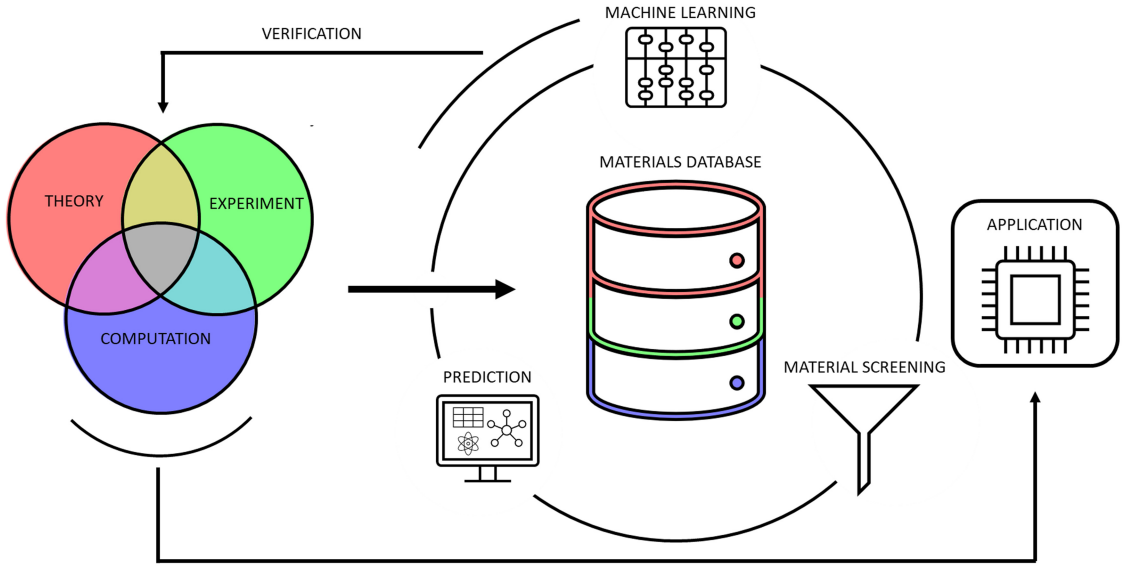

Fig. 1 Schematic of an example workflow in material informatics. Results from theory, experiment and computation are fed into material databases (arrow pointing to the center). A cycle involving material screening, machine learning and predictions leads to knowledge gain and ultimately applications in fields such as clean energy and quantum technology.

which to expect QT compatible properties. Rather than searching through a host of signals for those that match our criteria, we aim to *predict* which materials and signatures should be targeted for more detailed studies, following the framework illustrated in Fig. 1. This is made feasible by the availability of databases containing material properties for a wide range of different systems. In this work, the data in question are provided by bulk density functional theory (DFT) calculations to obtain the ground state properties of different elements and compounds. Combined with machine learning (ML) methods we provide a path towards precise classification of candidate materials. The inclusion of ML methods follows recent trends in applications of statistical learning, data science and machine learning for scientific discoveries, see for example refs. [15,16].

Herein, a framework is provided for the data mining and automated discovery of promising semiconductor hosts for QT using targeted database search and ML methods combined with knowledge from the field. Analyzing the output of the ML methods reveals that, given a suitable initial set of labeled materials for training and testing, it is possible to discern the physical mechanisms that govern a material's suitability for quantum applications. This framework does not distinguish between the specific mechanism giving rise to properties such as single-photon emission and long spin coherence times (e.g., semiconductor defects or quantum dots); instead, we attempt to target all materials that may accommodate the desired characteristics. The methodology developed herein can be modified for other material types and application areas provided that high quality databases containing relevant theoretical and/or experimental data is available.

The developed procedure relies on data extraction from different databases and the subsequent featurization of the data. An important aspect of the work is the database building and pertinent development of the datasets for the ML methods. Three different approaches to data mining were devised: (i) the criteria-based approach which is similar to that proposed by ref. [17], (ii) the extended criteria-based approach and (iii) the empirical approach. The two first data extraction protocols are based on broad material descriptors, leading to large sets of potentially suitable candidates[18,19]. The empirical approach, on the other hand, relies on including materials with experimentally proven advantageous characteristics in the training set and therefore yields a narrower set of possible candidates. The three resulting sets of labeled data were then analyzed with the four supervised ML methods logistic regression, decision trees, random forests and gradient boosting[19,20], yielding 47 predicted candidates that were common between all approaches and ML methods. Example materials among the predictions include $ZnGeP_2$, CdS, BP, BC_2N, GeC and InP. Focused theoretical and experimental studies are needed to verify the predictions that quantum properties may manifest as a result of defects or nanostructures in the above listed materials. Importantly, our findings also reveal which material properties are weighted by the ML methods upon predicting a material as suitable for QT applications, thereby opening up for new discoveries in the field of quantum technologies.

RESULTS

Information flow

The information stream in this work can be regarded as many interconnected modular parts. The initial step for gathering material data and building features is visualized by the outer flowchart in Fig. 2 (Jupyter notebooks containing the full workflow can be found at ref. [21]). We start by extracting all entries in the Materials Project (MP) database[22,23] that match a specific query. The MP database contains ground state properties of different materials that are computed using density functional theory. The DFT calculations in the database were performed using the Vienna ab initio simulation package (VASP)[24] and the Perdew–Burke–Ernzerhof (PBE)[25] exchange-correlation functional to calculate the electronic structure of the materials. We note that despite being immensely successful in describing a number of material properties, PBE is widely known for underestimating the band gap of semiconductors[26]. Therefore, not all properties predicted using the PBE functional are reliable, and the band gaps in particular cannot be trusted in absolute terms. Nonetheless, the functional is in wide use due to the combination of reasonable accuracy and high computational throughput, and is usually considered to be reliable for large-scale trends in semiconductor material properties.

The conditions for the initial MP query are that the materials must derive from the Inorganic Crystal Structure Database (ICSD) and have a band gap wider than 0.1 eV to exclude metallic compounds. The ICSD is the world's largest database for completely identified inorganic crystal structures[27]. In a parallel step, entries that are

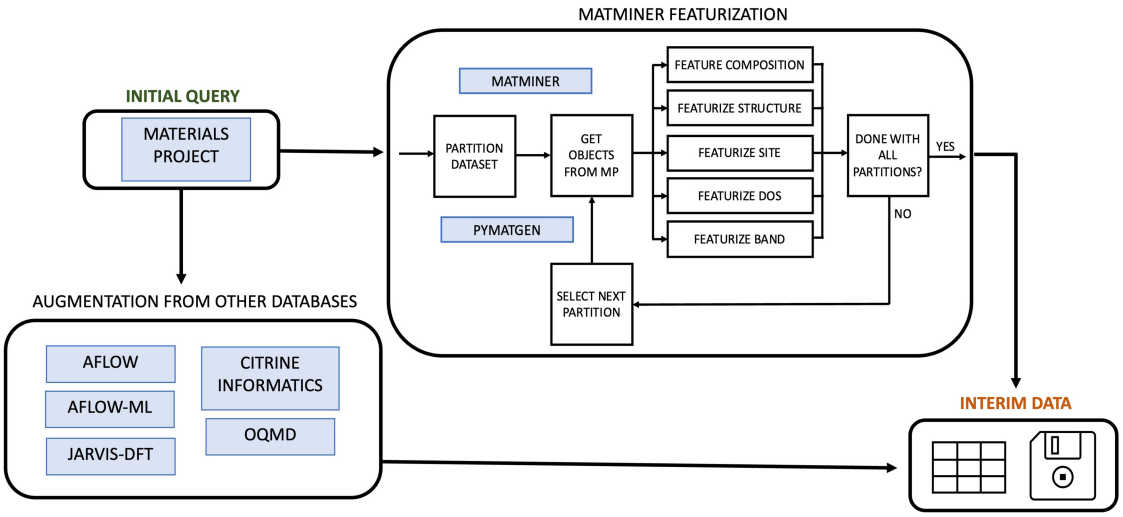

Fig. 2 Featurization workflow. The project workflow starting from an initial Materials Project (MP) query, and ending with a featurized dataset with entries from several other databases. To limit the memory and computational usage, the data is partitioned into smaller subsets where the respective Pymatgen objects (Pymatgen is a robust, open-source Python library for materials analysis[69]) are obtained through a query to be used in the following featurization steps. This process is repeated iteratively until all the data has been featurized. *DOS* refers to density of states and *band* to the electronic band structure.

deemed similar to the entries from the initial query are extracted from the following databases: the Open Quantum Materials Database (OQMD)[28,29], JARVIS-DFT[30], AFLOW[31–33], AFLOW-ML[34] and the Citrination platform[35]. The results of these steps are combined into a dataset for further analysis.

After material extraction tools from the open-source library Matminer[36] were applied to generate thousands of features from the data. We will refer to this process as featurization. A schematic visualization of the featurization process in Matminer is shown in Fig. 2 and focuses on a material's composition, structure, atomic sites, density of states and band structure. The 39 featurizers (each generates several features) selected as material descriptors in this work are described in the Supplementary Information at ref. [37]. The selection of features was kept rather wide to avoid making *ex ante* assumptions on which features best describe a solid state material platform for QT.

The constructed dataset encompasses compounds formed by a plethora of combinations of surfaces, interfaces, nanostructures, compositions and structures, but this complexity is not necessarily reflected in the material descriptors. Furthermore, the extracted data was obtained from high-throughput density functional theory calculations. Indeed, there are possible errors associated with every step, starting from an initial calculation, adding of data in the database, featurization and preprocessing the data, data mining and labeling, and finally training a model and making a prediction. Unfortunately, if an error occurs in the first part of the process, it will be carried along and get increasingly harder to detect. Thus, the output strongly depends on the quality of data in the employed databases.

Data mining

The complete dataset consists of 25,000 materials. A subset of these materials is labeled into either suitable or unsuitable candidates for QT while the remainder will stay unlabeled. The labeled data is then grouped into training and test sets for the ML methods. The ML methods are trained using the training set and then evaluated on the test set. Finally, the ML methods are applied to the unlabeled data for which predictions of QT suitability are made.

Several challenges accompany the labeling process. Although QT compatible properties are becoming increasingly well studied for the case of, e.g., nitrogen-vacancy centers in diamond, single-photon emitters in silicon carbide and quantum dot (QD) structures[6,9,12], relatively few candidate materials are known to be suitable[13,14]. An additional consideration is that the physical mechanisms promoting favorable properties are not fully understood. Conversely, defining materials as unsuitable candidates for QT is in many ways equally difficult, as the mechanisms preventing quantum compatible characteristics to manifest are not known either. The strategy for selection of unsuitable candidates is, thus, that the negation of the criteria used to select suitable candidates should give unsuitable candidates. A side-effect of this selection is that the criteria for unsuitable candidates can become equally, or more, restrictive compared to the criteria for suitable candidates, resulting in skewed datasets with fewer unsuitable than suitable candidates. This method for selecting unsuitable candidates has shortcomings but will serve as a starting point and demonstration for how the labeling procedure could be improved. Below three separate procedures for labeling materials as suitable or unsuitable candidates for QT are described.

The first approach to labeling a selection of materials in the dataset is based on the criteria proposed by ref. [17]. They suggest a data mining process consisting of four stages by systematically evaluating the suitability of host materials taken from the Materials Project. Note that the data mining process in ref. [17] alone was intended for further experimental studies and not necessarily for a targeted machine learning search. Nonetheless, we adapt the selection criteria from ref. [17] for the present purposes. In the framework of the *criteria-based approach*, suitable candidates are labeled according to the following steps:

1. Include materials that;
   - contain elements with more than 50% natural abundance of spin zero isotopes,
   - crystallize in non-polar space groups, are calculated to be nonmagnetic, and are present in the ICSD.
2. Pragmatically remove toxic, radioactive and otherwise "difficult" materials;
   - exclude Th, U, Cd and Hg because they are radioactive and/or toxic in the most stable forms,
   - exclude any rare-earth metals (because of the difficulty

Table 1. Known quantum compatible materials and defect candidates.

Material	Band gap (eV)	Defect candidates	References
Diamond	5.5	N_CV_C, Si_CV_C, Ge_CV_C	71–76
SiC	2.2–3.3	V_{Si}, $V_{Si}V_C$, $C_{Si}V_C$, N_CV_{Si}	8,11,77–81
Si	1.1	P, G, unidentified	60,82,83
(2D) *h*-BN	6.0	Unidentified defects	84–86
(2D) MoS_2, WSe_2, WS_2	<2.5	Bound excitons	38
ZnO	3.4	Unidentified defects	87
ZnS (zincblende)	3.6	Unidentified defects	88
GaAs	1.4	Quantum dots	89
GaN	3.4	Quantum dots, unidentified defects	90,91
AlN	6.0	Unidentified defects	92

The materials have demonstrated quantum compatible characteristics such as single-photon emission and coherent spin manipulation. The subscript denotes lattice site and *V* refers to a vacancy.

of obtaining pure materials free of isotopes with nuclear spin) and noble gases (due to the lack of stable solid phases),
- exclude transition metal elements with unpaired electrons like Fe and Ni because of their paramagnetism; Ru and Os are also excluded because they only exist in the dataset as complex cluster structures.

3. Include only materials with a band gap larger than 0.5 eV calculated using DFT and the PBE functional. The value of 0.5 eV was chosen to match that typically predicted for silicon by PBE-level DFT calculations.
4. Ensure that the energy above hull is less than 0.2 eV per atom.

The inclusion criteria are based on ref. [8] and targeted primarily at semiconductors that can host deep level defects with spin qubit capabilities. In this context, long spin coherence times are needed, necessitating an environment that can be depleted of nuclear spins and permanent magnetism. Moreover, non-polar materials are assumed to be preferable to obtain sharp and indistinguishable single-photon emission from defects. Transition metal elements are eliminated if they have unpaired electrons because the presence of permanent electric dipole moments may have a detrimental impact on the optical coherence of defect emission. Finally, the energy above hull requirement ensures that the selected compounds are thermodynamically stable. Note that larger cells are sometimes needed to verify antiferromagnetic ordering so the criteria mainly target ferromagnetic ordering under the labels magnetic/non-magnetic.

Next, unsuitable candidates are labeled according to the reverse requirements of the above; as materials in the ICSD that crystallize in polar space groups, are calculated to be magnetic and have a band gap larger than 0.1 eV in the MP database (to exclude metals but include lower-band gap semiconductors). Crucially, only materials that satisfy *all* of the reverse requirements are labeled as unsuitable. In other words, unlabeled materials may satisfy some of the selection criteria, but not all. The resulting set of labeled materials contains 1581 materials where 35% are labeled as unsuitable and 65% as suitable for QT applications, evidencing that the criteria for labeling unsuitable candidates are more restrictive than those for the suitable ones. The remaining ~ 23500 unlabeled materials may still contain suitable candidates for QT, motivating for the use of ML methods to classify the materials and deduce trends in material properties.

Considering the materials present in the labeled dataset of the criteria-based approach more closely, diamond is classified as a suitable candidate in good agreement with experimental observations. Carbon in two-dimensional graphite-like structures are marked as suitable as well, along with all structures of silicon and one entry of silicon carbide. Note that this is the 3C polytype of SiC, meaning that the most well-established quantum compatible SiC polytype, 4H-SiC, was not classified in the labeling process and is found in the unlabeled dataset. Among other relevant candidates we find that ZnS, ZnSe, ZnO and ZnTe were all labeled as suitable in the criteria-based approach.

Next, the criteria-based approach was adjusted to expand the data labeling process beyond practical considerations. The second approach is therefore named *the extended criteria-based approach* and involves removing stage two from the approach above. Moreover, certain additional elements that have shown promising properties but were initially excluded due to the lack of spin zero isotopes are also included.

The following steps constitute the process of labeling suitable candidates in the extended criteria-based approach:

1. Include materials that;
   - contain elements where more than half have a natural abundance of spin zero isotopes, including Al, P, Ga, As, B and N,
   - crystallize in non-polar space groups, are calculated to be nonmagnetic and are present in the ICSD.
2. Only keep materials that have a band gap larger than 1.5 eV in the MP database. The higher band gap requirement (as compared to the criteria-based approach) is included here to avoid labeling an unfeasibly large number of materials.
3. Ensure that the energy above hull is less than 0.2 eV per atom.

For unsuitable candidates, the same strategy as for the criteria-based approach was implemented. The result is an unbalanced set of labeled materials that is 78% larger than for the criteria-based approach and having ~75% of the materials found in the suitable group.

The findings from both the data mining and machine learning procedures for the extended criteria-based approach did not differ substantially from those obtained using the criteria-based approach. This is attributed to the similarities in the selection processes. Therefore, detailed discussion on the findings from the extended criteria-based approach can be found in the Supplementary Information at ref. [37]. Qualitative conclusions drawn for the criteria-based approach herein also hold for the extended criteria-based approach, indicating that the removal of so-called practical considerations did not have a significant impact on the results. Nonetheless, predictions from the extended criteria-based approach will be employed below to contrast with and filter the findings from the empirical approach.

In the *empirical approach*, we apply knowledge from the field (see for instance refs. [11,13,14,38] for an overview) to guide the search for promising material hosts. In other words, the labeled data contains candidates where quantum compatible properties have been either experimentally demonstrated or theoretically predicted, including the materials suggested in ref. [8] as promising deep level defect hosts. Materials where single-photon emission and spin manipulation have been observed but attributed to excitonic effects or quantum dots (formed by self-assembly or lithographic structuring) rather than being defect related were also included.

Table 1 contains an overview of known semiconductor materials with demonstrated quantum compatible characteristics. The table forms the basis for picking suitable candidates for the

empirical approach. The properties being studied arise from mechanisms related to, e.g., point defects, bound excitons, and both self-assembled and lithographically structured quantum dots and nanostructures such as 2D materials. Quantum emission signatures have been assigned to specific defects in both diamond and SiC, but for most other materials, secure identification of the responsible defects or structure related mechanism is still lacking.

The strategy for picking suitable candidates in the empirical approach is;

1. Select candidate materials that match the formulas in Table 1, or the formulas ZnSe, AlP, GaP, AlAs, ZnTe, CdS[8] and SiGe[39], as these materials have been predicted to behave as suitable quantum hosts based on favorable properties such as a wide band gap and low spin-orbit coupling.
2. Ensure that the candidates are present in the ICSD.
3. Perform a manual screening for appropriate crystallographic structures.

After the first stage of picking candidates we are left with a list of 202 matching formulas which includes 12 entries that have a band gap of less than 0.4 eV. These 12 entries are calculated to be thermodynamically unstable in terms of the energy above hull, and will decompose into other materials in the list—incidentally, the resulting structures all have calculated band gaps that are substantially larger than 0.5 eV. All of the 202 structures were included in the labeled dataset apart from C (mp-568410) which is a metal according to AFLOW-ML.

Entries matching the formulas C, SiC, BN, MoS_2, WSe_2 and WS_2 were manually screened to see whether they have a matching structure to the respective candidates discussed earlier and summarized in Table 1. For carbon, three-dimensional diamond-like structures as explicitly stated in the column tags from the MP were admitted. Additionally, we find several two-dimensional structures of carbon with a large band gap (>1.5 eV) among the data. These were added as suitable candidates. Complex structures (e.g., C_{28}, C_{48} and C_{60}) were moved to the test set in our machine learning studies. For SiC we admitted all entries which included the 2H, 3C, 4H, 6H and 15R polytypes. Concerning BN, MoS_2, WSe_2 and WS_2, only two-dimensional structures were admitted. Other non-matching structures were moved to the test set to see whether or not they are predicted as suitable by the ML methods applied in a later stage.

The materials AlP, GaP, AlAs, ZnTe and CdS were manually screened for tetrahedrally coordinated structures, and have been included since ref. [8] identified them as potentially promising candidates due to suitable material properties. Note that only tetrahedrally coordinated structures of the given formulas were labeled as suitable after imposing a band gap restriction of 0.5 eV.

Following the three screening steps in the empirical approach, a total of 187 entries were labeled as suitable candidates for the training set. Notably, only elementary (unary) and binary materials are labeled as suitable in the empirical approach. To complete the dataset, 400 materials were added and labeled as unsuitable. These were picked at random from the pool of unsuitable candidates from the two previous approaches, in addition to those that were marked as unsuitable during the manual screening process.

Note that there is some potential for inherent bias in the dataset, in part due to experimental work being limited by the availability and cost of materials and processing. Moreover, the discovery of the quantum compatible properties of the NV center in diamond naturally led early searchers to comparable materials such as silicon carbide.

The labeled data for the criteria-based and empirical approaches is visualized in Fig. 3 as parallel coordinate plots for selected features informed by the criteria proposed by ref. [8]. Parallel coordinate schemes[40,41] represent a multi-dimensional data tuple as one polyline crossing a parallel axis. The selected features are found on the *x*-axis, while the *y*-axis shows the value of the data. Thus, parallel coordinate plots can turn complex many-dimensional data into a compact two-dimensional representation. Due to possible data cluttering, the figure visualizes a

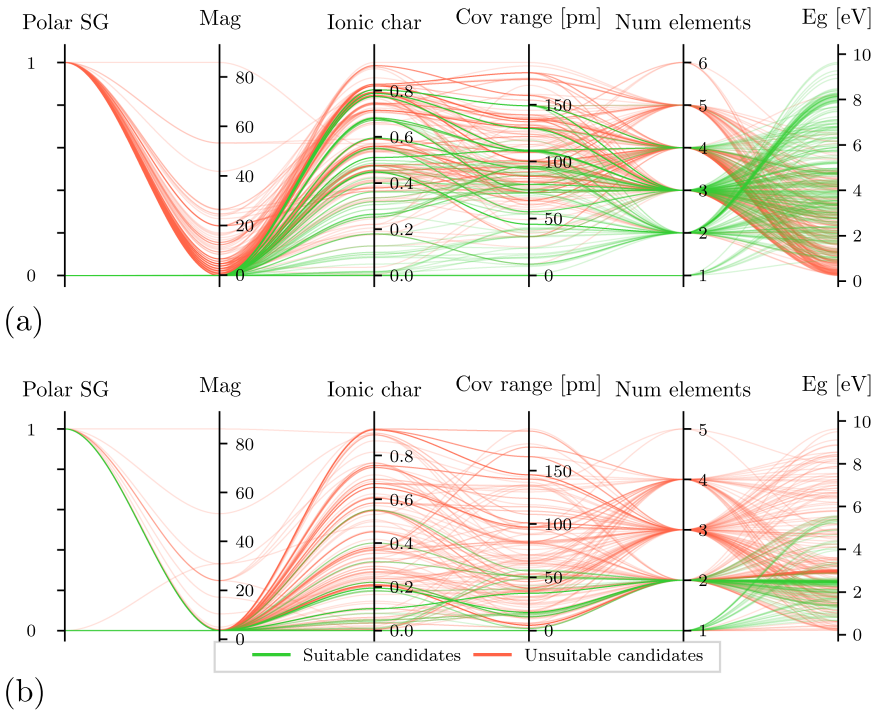

Fig. 3 Parallel coordinate plots. Parallel coordinate plots for the (**a**) criteria-based and (**b**) empirical approaches. To limit data cluttering up to 250 entries for each class were randomly collected. The axes show total magnetization (mag), space group (SG), ionic character (ionic char), covalent range (cov range) as calculated from elemental properties, number of elements (num elements) and energy gap (Eg) as extracted from the MP database.

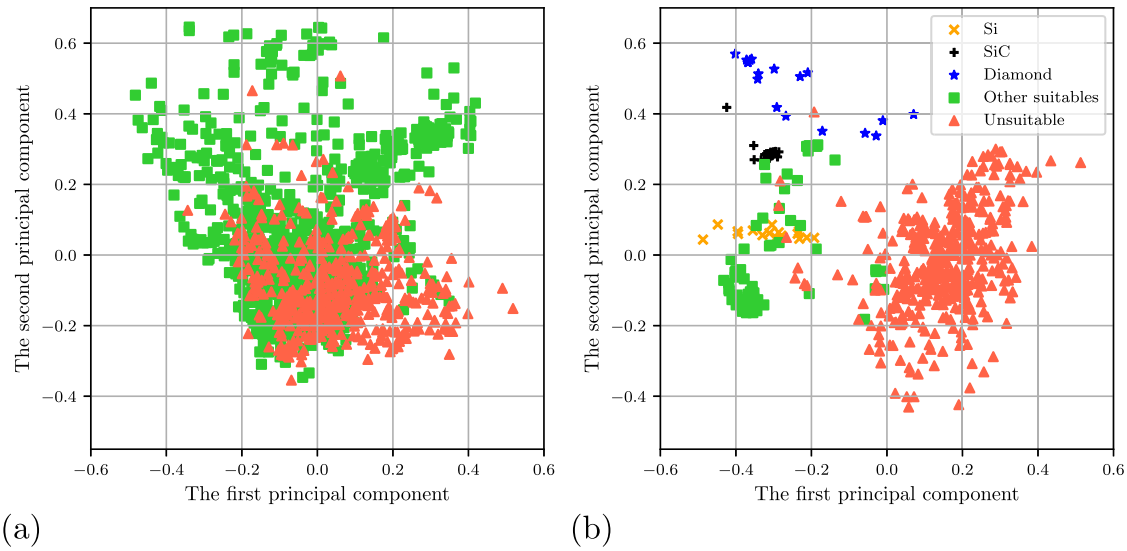

Fig. 4 Two-dimensional scatter plots. Scatter plots for the (**a**) criteria-based and (**b**) empirical approaches. The eigenvectors corresponding to the two largest eigenvalues of the covariance-matrix were identified, that is, the two most important principal components of the initial data from the Materials Project query. Then, the labeled datasets were transformed and visualized as 2D scatter plots. Green squares display suitable candidates, along with the black (SiC), blue (diamond) and yellow (Si) symbols for the empirical approach, and red triangles represent unsuitable candidates.

random sample of each class (suitable or unsuitable) with an upper limit of 250 per class with transparent lines. The green and red polylines represent suitable and unsuitable candidates, respectively.

Several differences are observed between the criteria-based approach and the empirical approach to data mining. While neither case passes materials that exhibit magnetization through as suitable, suitable candidates can crystallize in both polar and non-polar space groups in the empirical approach. Moreover, the ranges of covalent radius and maximum ionic character span a substantially smaller parameter space for the empirical approach than for the criteria-based approach (see Fig. 3). Overall, there seems to be greater overlap in the distributions in material properties for suitable and unsuitable candidates in the case of the criteria-based approach as compared to the empirical one.

To reduce the dimensionality of the labeled data a principal component analysis (PCA) was performed[42]. In its standard form, PCA relates the variance of the features with the eigenvalues of the covariance matrix[19,20,42]. We identify the two largest eigenvalues of the covariance matrix[19] of the complete initial dataset from the MP database, and transform the three labeled datasets according to the corresponding two eigenvectors. The result of this procedure is displayed in the scatter plots of Fig. 4. Note that some minor differences between the approaches may occur due to the process of removing the mean and scaling to unit variance. Red triangles represent unsuitable candidates while otherwise colored symbols (green, blue, black and yellow) represent suitable candidates. Due to the complexity of reducing the large amount of features down to only two, suitable and unsuitable candidates for the criteria-based approach are largely overlapping. The logic behind categorizing materials in two classes (suitable and unsuitable) appears to not translate into a distinct separation, at least not in the representation of Fig. 4 for the criteria-based approach. Hence, using this approach for predicting QT material hosts could prove challenging for any model that would try to glean a clear-cut boundary between materials that are and are not suitable for QT. We therefore expect that the criteria-based approach could need supplementary dimensions for further distinguishing between the materials in the two categories.

In the case of the empirical approach, on the other hand, a clear trend can be discerned where the upper left part of Fig. 4 is dominated by suitable candidates while the unsuitable ones are similarly restricted to the lower right corner, albeit with some exceptions. Interestingly, we observe that different configurations of the famously quantum compatible materials of diamond (blue), silicon carbide (black) and silicon (yellow) are grouped together but each material class is separated in its own region. This strongly indicates that the empirical approach may be capable of separating materials based on their underlying properties, emphasizing the importance of having a logical framework for the data mining process.

Machine learning and principal component analysis

Four well-tested ML algorithms were applied to the labeled data to classify specific materials as candidate systems for QT, yielding multiple sets of predicted candidates. The ML methods employed herein are logistic regression, decision trees, and the ensemble methods random forests and gradient boosting[18–20]. Principal component analysis was again performed, but now on the three training sets separately (as opposed to the case of Fig. 4 where all panels are based on the same model). Next, in the evaluation of the ML methods we apply a 5×5 stratified cross-validation[19] when searching for the optimal hyperparameter combinations. Note that all four ML methods have high evaluation metrics for the optimal hyperparameters. Further details on the evaluation of the ML methods and the results from the principal component analyses for all four ML methods are shown in the Supplementary Information at ref. [37]. The data labeling approaches define sets of labeled data of varying sizes, where the smallest is from the empirical approach. For small datasets, it is proven to be beneficial to repeat the cross-validation analysis[19], as this is a method that allows us to measure the stability of the predictions against perturbations (i.e., few different entries) in the training data[43].

Figure 5 visualizes different parameters for the most important principal components ranked in descending order by (a) the explained variance for the Ferrenti (upper panel) and empirical (lower panel) approach, and (b) the gradient boosting coefficients for the corresponding approaches. This differs from Fig. 4, where the two most important eigenvectors for all approaches that

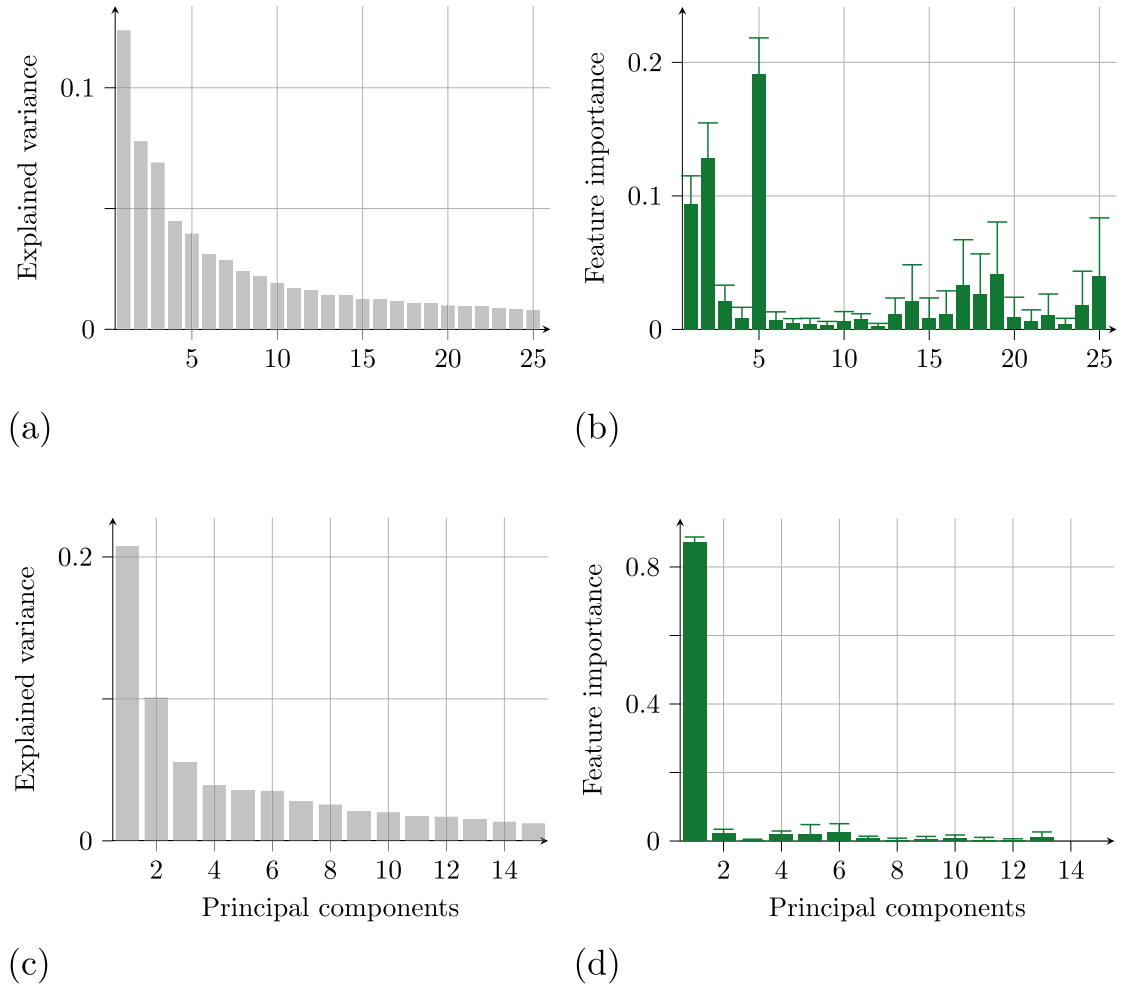

Fig. 5 Explained variance and feature importance. Visualization of different parameters for the 25 most principal components ranked in descending order by the explained variance for the (**a**) criteria-based and (**c**) empirical approaches, and (**b**) and (**d**) the mean gradient boosting coefficients during 5×5 cross-validation with the standard deviation (s.d.) for the corresponding approaches as error bars. Note that in (**a**) and (**b**) only the results of the PCA analysis are visualized, while (**c**) and (**d**) contains results of the PCA reduced training sets using gradient boosting[19,70]. The latter shows that for the empirical approach most of the physics is represented by a few features. The results in (**b**) and (**d**) are similar to those obtained with logistic regression, decision trees and random forests.

originate from the same covariance matrix are shown. For the criteria-based approach (Fig. 5a) to reach the 95% accumulated explained variance, a total of 144 principal components must be included. In the case of the empirical approach and in contrast to the criteria-based approach, decision trees and random forests exhibit the best performance for just a few principal components, and experience a considerable degree of overfitting when involving more principal components. Gradient boosting also experiences the best performance for just a few principal components.

Figure 5(b) reveals that the most important feature for the gradient boosting method in the criteria-based approach is the fifth principal component (fifth largest eigenvalue). The trend is maintained for all four ML methods (see the Supplementary Information at ref. [37]). By selecting the highest values in this eigenvector we find that the corresponding features originate from the DFT computed band gap of the elemental solids among the elements in the compound as calculated by the Materials Agnostic Platform for Informatics and Exploration (MagPie)[44]. The second most important principal component for the criteria-based approach exhibits significant contributions from the covalent radius, the ionic character and the packing efficiency among the elements in the composition. The data originate from elemental calculations from MagPie and are aggregated as either minimum, mean, standard deviation, or maximum. While the first principal component encompasses the largest explained variance, it does not provide a clear and specific information on which features it represents.

The empirical approach differs in many aspects from the criteria-based approach. Firstly, the number of principal components necessary to obtain 95% variance is reduced to 103 from 144. Thus, the variance of the training set is described with fewer principal components, indicating a simpler model. Secondly, the first principal component is by far the most important feature for all ML methods in the empirical approach, as visualized in the lower panel of Fig. 5b for gradient boosting. Similar conclusions are reached when using logistic regression, decisions trees and random forests as classification methods (see the Supplementary Information at ref. [37]). The distinct importance of the first principal component partly explains why we experience a high accuracy for only a single feature. The first principal component's corresponding features is a complex combination of several material properties, but we find that it includes bond orientational parameters, coordination numbers, and the radial distribution function of a compound's crystal system. The standard deviation of the radial distribution function appears multiple times in the list

of features and is of particular importance. Thirdly, the empirical approach differs in how much explained variance is retained by the first component, which is 21%, compared to 14% for the criteria-based approach. We find the difference striking considering that the approaches share the same ultimate goal, but where the training sets apparently exhibit large and significant variations.

Intriguingly, the principal components that are deemed important by the approaches differ substantially. The criteria-based approach places particular value on the material's band gap and the ionic or covalent character of the bonding. Indeed, precisely these features were used as guidance in the data labeling process. Thus, the ML methods seem to perpetuate the criteria imposed in the criteria-based approach at least to some extent. They may, however, still be recognizing other patterns than those originally intended in the data selection process. For the empirical approach, on the other hand, the selection was not guided in terms of specific properties. Here, the ML methods appear to be informed by other characteristics than band gap and ionic character that are more related to symmetry and crystal structure, based on the repeated appearance of bond length, orientation and radial distribution in the first principal component. In the empirical approach, the ML methods are recognizing more complex mechanisms in the crystal structure and bonding as common trends among materials that are suitable for QT applications.

Predicting suitable material hosts for quantum technology

After training and validating the ML algorithms on the labeled datasets, the ML methods were applied to unlabeled data to obtain predictions for suitable QT host materials. The number of candidates predicted by each of the four ML methods logistic regression, decision trees, random forests and gradient boosting is visualized in the Supplementary Information at ref. [37]. An overview of the number of predicted suitable candidates as a function of a given ML algorithm's confidence is summarized in Table 2 for the criteria-based, extended criteria-based and empirical approaches. Note that the criteria-based approaches are employed as a filter on the predictions of the empirical one.

From the predicted dataset of 23,623 materials all four ML methods agree on a total of 6804 suitable candidates in the criteria-based approach, however, many of these materials are predicted with an incidence similar to that of a coin flip. Raising the minimum confidence cut-off for a prediction to, e.g., 0.75 instead of 0.5, yields only 1784 suitable candidates that the ML algorithms agree on. The ML methods admit almost all materials with a chemical formula matching the known suitable candidates (see Table 1) that were present in the labeled data. This can allow materials with complex structures (e.g., nanostructures or 2D materials) to be labeled as suitable candidates. Notably, the ML methods do not maintain the band gap restriction from the training set definition, where all materials with band gaps lower than 0.5 eV were eliminated. This trend is not carried over to the predicted data. Indeed, many entries with band gaps lower than 0.5 eV are predicted as suitable candidates by all four ML methods employed herein—despite the principal component analysis revealing that the band gap is an important feature for the machine learning classification in the criteria-based approach. This indicates that the ML methods are not exactly recognizing the initial selection criteria, instead finding other patterns in the dataset. On the other hand, due to the known underestimation of band gaps by the PBE functional, the band gaps could in reality be larger for many of the materials.

Table 2. Number of predicted suitable materials vs. ML model confidence.

Approach	0.50	0.75	0.85
Criteria-based approach	6804	1784	258
Extended criteria-based approach	9227	4735	2001
Empirical approach	214	66	9
All approaches together	47	6	0

Number of predicted suitable candidates as a function of a given ML method's confidence when the four methods in an approach agree. A threshold value of 0.5 represents the confidence of a coin flip while 1.0 is a fully confident prediction.

The ML methods predict materials as suitable that are not expected according to, e.g., ref. [8]. Indeed, NaCl is predicted as a suitable candidate to minimum confidences of 0.83 and 0.61 for two different configurations, despite the strong electrostatic interactions between Na and Cl and the ionic character of their bonding. Note that NaCl was excluded from being labeled as both unsuitable and suitable in the criteria-based approach and was therefore found in the unlabeled dataset. Conversely, 4H-SiC, which was unlabeled in the selection process, is predicted as a suitable candidate by all four ML methods in the criteria-based approach. For the empirical approach, NaCl was included in the labeled data in the training set as an unsuitable candidate, while 4*H*-SiC was labeled as suitable in the initial selection process.

The ML methods that were trained on the data extracted in the empirical approach predict substantially fewer candidates as compared to the criteria-based approach. A total of 842, 1197, 543 and 596 materials are classified as suitable candidates by logistic regression, decision trees, random forests and gradient boosting, respectively. All the four ML methods agree on a total of 214 suitable candidates to 0.5 confidence. Note that 51 of these have a band gap of 0.5 eV or smaller. Increasing the threshold to 0.75 or 0.85 yields 66 or 9 predicted suitable candidate materials in the empirical approach, respectively (see Table 2).

Consider the 9 materials that were classified as suitable to a confidence of 0.85 or higher by all four ML methods in the empirical approach; BN, CdSe (2 structures), BC_2N (2 structures), InAs, CuI (2 structures), and $ZnCd_3Se_4$. The nine materials (considering different crystal structures) each belong to one of the four crystal systems cubic, hexagonal, tetragonal and orthorombic. Figure 6 visualizes the four different crystal systems while Table 3 lists important material properties of the relevant materials as reported in the MP database. Interestingly, all nine materials appear to be four-fold coordinated, and the first two lattice vectors (a and b) are identical in all cases whereas c may differ. There appears to be a high degree of symmetry present for all materials, but no elemental and perfectly symmetric semiconductors were found in the list.

Turning to each of the individual material predictions, two compositions of CdSe (the cubic mp-2691 and the hexagonal mp-1070) are predicted as suitable, possibly as a consequence of the similar compound CdS being labeled as a suitable candidate in the training set. The two compounds of CdSe share similar properties with calculated band gaps in the MP database of 0.5 eV and 0.6 eV, respectively. The compound BN (mp-1639) was also predicted as suitable possibly due to a different BN crystal structure being present in the training set and labeled as a suitable candidate.

In the case of BC_2N we find two compositions with the same chemical formula; the orthorhombically coordinated (mp-629458) and the tetragonally structured (mp-1008523) BC_2N. The former takes on a polar space group while the latter does not. The band gaps are listed as 1.9 eV and 1.7 eV, respectively, in the MP database. Both structures exhibit a strong covalent character and have been studied for applications as nanostructures for electronic devices[45], hydrogen storage[46] and superhard materials[47,48]. Interestingly, the diamond-like structure of BC_3N was recently predicted as a promising spin qubit material host[49]. By creating a boron vacancy one can in theory obtain a defect center with

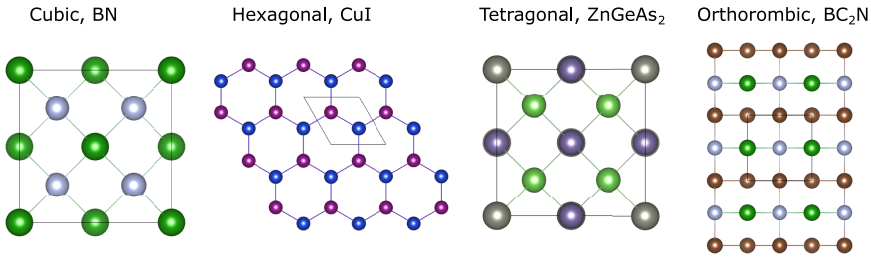

Fig. 6 Crystal structures. Example illustrations of the crystal systems that the 9 materials predicted by the empirical approach to 0.85 confidence, and the 6 materials predicted by all approaches to 0.75 confidence, belong to. Four different symmetry categories are observed; cubic, hexagonal, tetragonal and orthorombic. The viewpoints for all materials are down along the *c*-axis.

Table 3. Material properties from the MP database.

Approach	Material	Crystal structure	MP code	Density (g cm^{-3})	Band gap (eV)	a, b, c (Å)	α, β, γ (°)
Empirical approach to 0.85 confidence	CdSe	Hexagonal	mp-1070	5.3	0.6	4.4, 4.4, 7.2	90, 90, 120
	CuI	Hexagonal	mp-569346	5.8	1.2	4.3, 4.3, 7.0	90, 90, 120
	CuI	Cubic	mp-22895	5.8	1.2	4.3, 4.3, 4.3	60, 60, 60
	CdSe	Cubic	mp-2691	5.3	0.5	4.4, 4.4, 4.4	60, 60, 60
	BN	Cubic	mp-1639	3.5	4.6	2.6, 2.6, 2.6	60, 60, 60
	InAs	Cubic	mp-20305	5.3	0.3	4.4, 4.4, 4.4	60, 60, 60
	$ZnCd_3Se_4$	Cubic	mp-1078597	5.3	1.7	6.1, 6.1, 6.1	90, 90, 90
	BC_2N	Tetragonal	mp-1008523	3.3	1.6	2.6, 2.6, 3.7	90, 90, 90
	BC_2N	Orthorombic	mp-629458	3.4	1.8	2.5, 2.6, 3.6	90, 90, 90
All approaches to 0.75 confidence	$CdSnP_2$	Tetragonal	mp-5213	4.6	0.7	7.2, 7.2, 7.2	131.1, 131.1, 71.7
	GeSe	Cubic	mp-10759	5.5	0.4	4.0, 4.0, 4.0	60, 60, 60
	InP	Cubic	mp-20351	4.6	0.5	4.2, 4.2, 4.2	60, 60, 60
	InP	Hexagonal	mp-966800	4.6	0.5	4.2, 4.2, 6.9	90, 90, 120
	SiSn	Cubic	mp-1009813	4.4	0.4	4.3, 4.3, 4.3	60, 60, 60
	$ZnGeAs_2$	Tetragonal	mp-4008	5.2	0.6	7.0, 7.0, 7.0	131.4, 131.4, 71.2

Material properties for the 9 materials predicted by the empirical approach (>0.85 confidence) and the 6 materials predicted by all approaches (>0.75 confidence).

similar properties to those found for the NV center in diamond. Whether this is also possible for BC_2N remains to be seen. Note that BC_3N was not represented in the MP database at the time of data extraction and is therefore not included in our dataset.

The compounds InAs (cubic, mp-20305), CuI (cubic, mp-22895 and hexagonal, mp-569346) and $ZnCd_3Se_4$ (cubic, mp-1078597) are listed in the MP database with band gaps of 0.3 eV, 1.2 eV, 1.2 eV and 1.7 eV, respectively. To the best of our knowledge, $ZnCd_3Se_4$ has yet to be synthesized, while self-assembled InAs quantum dots are exciting possible materials to use in quantum technology[50]. CuI has recently been synthesized as single-crystal epitaxial films and was shown to exhibit remarkable optoelectronic properties[51]. Interestingly, the material exhibits a large ionic character and shares few similarities with known QT compatible hosts such as, e.g., Si, SiC and diamond. The prediction of CuI in two configurations thus indicates that ionic character alone is not an obstacle for a material to be quantum compatible. It is unknown at this time whether any potentially favorable properties of CuI would originate from deep level defects or nanostructuring (e.g., quantum dots).

Out of the nine predicted materials (threshold of 0.85) we note that both instances of CuI are stated in the MP database to decompose into trigonal CuI, which was not present in our predictions. Hexagonal CdSe decomposes into cubic CdSe (in the list of 9), the cubic BN decomposes to hexagonal BN (mp-984, not in the predictions), and $ZnCd_3Se_4$ decomposes to ZnSe + CdSe (in the list). Finally, both compounds of the BC_2N structure are listed to decompose into hexagonal BN + C. However, synthesis may still be possible depending on the growth conditions.

Lowering the cut-off requirement from 0.85 to 0.75 for all ML methods results in 66 candidate materials for the empirical approach. The full list of these 66 candidates is displayed in the Supplementary Information at ref. [37]. In addition to the nine materials discussed above and some elemental and binary semiconductors, the list of 66 predicted suitable candidates now also includes ternary compounds of the formula ABC_2. For the ABC_2 structures, the elements Ga, Cd, In and Zn can occupy the A-site, Cu, Sn, Ag and Ge take the B-site, while S, Se, Te, P or As may reside at the C site. Most of the predicted compounds include at least one toxic element with one exception: $ZnGeP_2$ (mp-4524) in a chalcopyrite-like tetragonal crystal structure with an indirect band gap of 1.2 eV[52] as reported in the MP database. In comparison, the experimentally reported band gap is somewhat larger at 2.0 eV[53]. $ZnGeP_2$ crystallizes in a non-polar space group, possesses no magnetic moment, exhibits covalent bonding and has been reported as an excellent mid-IR transparent crystal material that is suitable for nonlinear optical applications[52]. Importantly, it is possible to integrate sources of photon quantum states based on nonlinear optics with $ZnGeP_2$[54]. $ZnGeP_2$ is therefore identified as an eligible candidate material for QT, but it remains to be seen whether the candidate can facilitate, e.g., the isolated deep energy levels often associated with defects exhibiting quantum compatible properties, or instead be a candidate for nanostructure based QT.

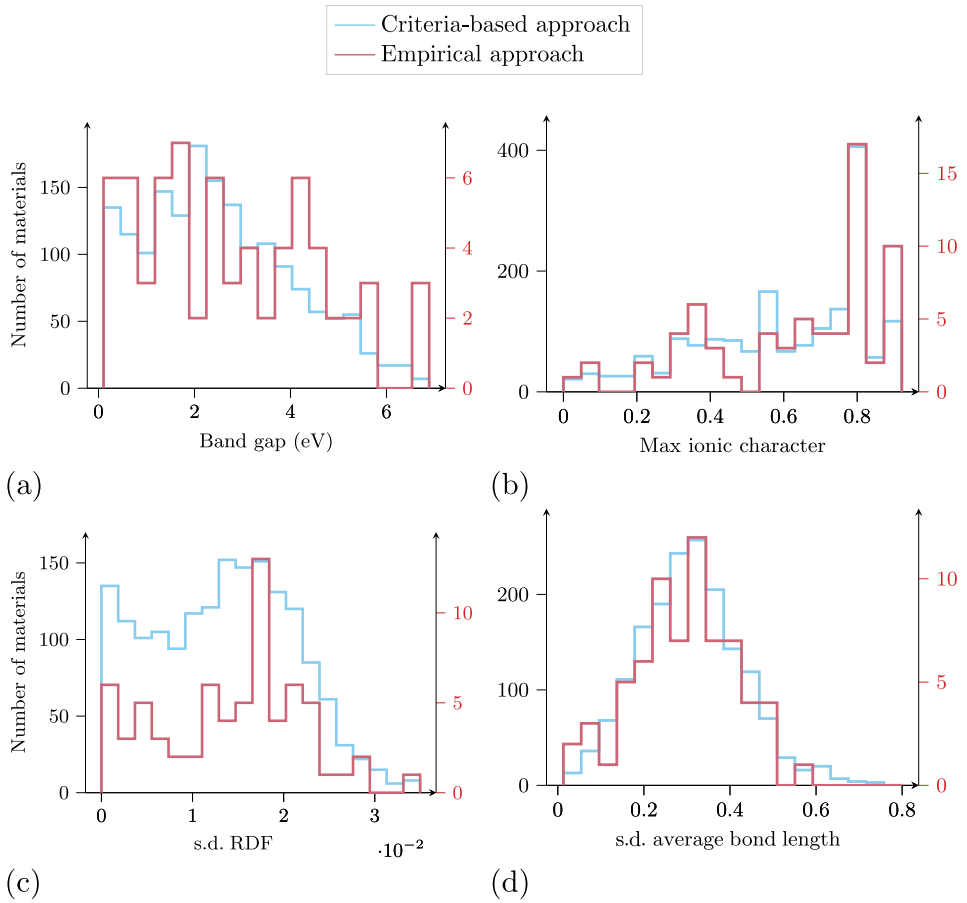

Fig. 7 Material distributions for different properties. Histograms of predicted suitable materials as a function of the (**a**) band gap, (**b**) maximum ionic character, (**c**) standard deviation of the radial distribution function (RDF) and (**d**) standard deviation of the average bond length. The total number of predicted materials is 6804 for the criteria-based approach, and 214 for the empirical approach. The criteria-based approach refers to the left *y*-axis and the empirical approach to the right.

Among the list of 66 we also highlight the predictions of interesting materials like Ge, sharing many characteristics with Si and C in addition to the periodic column number, GeC, BP and InP. Much like SiC, device design based on Ge can take advantage of the mature large-scale fabrication of silicon due to the material's comparable properties. Similar considerations could be made for the case of GeC. Data from the MP database suggests that the cubic compound GeC (mp-1002164) is a covalently bonded semiconductor having a band gap of 1.8 eV. Consequently, with SiC being widely known as a highly suitable host material for QT compatible defects, we encourage further research on GeC due to its similarities with SiC. Next, BP (mp-1479 and mp-1008559) is present in the predictions in both the cubic and hexagonal structures, with indirect band gaps calculated at 1.5 eV and 1.1 eV, respectively. Both configurations of BP are nonmagnetic, non-toxic and BP has been synthesized with a potential for large-scale production[55]. Lastly, InP (mp-966800) in the hexagonal structure is reported in the Materials Project to have a direct band gap of 0.5 eV and is considered as one of the most promising candidates to compete with Cd- or Pb- based QDs for, e.g., display and lighting applications[56,57]. The possibility of using InP-based quantum dots for QT applications should therefore be considered.

Comparing to the work of ref. [17], they suggest a list of 541 viable hosts after the data mining procedure. Among these, only a single material is present in the list of 66 candidates predicted by the four ML methods in the empirical approach: the nontoxic compound MgSe (mp-10760) which crystallizes in the rock-salt structure, is expected to have a 2.0 eV band gap and decomposes to a similar MgSe configuration. MgSe is notable for its available spin-zero isotopes in accordance with the criteria set by ref. [8]. We note that these properties may favor defects acting as spin centers with qubit potential and MgSe is thus identified as an interesting host material in this regard.

The number of materials predicted by the empirical approach is restricted enough to enable close scrutiny of the various suggested candidates. The same cannot be said for the criteria-based approach, however, as seen from Table 2. Manual verification through a literature survey will often not be possible, and perfecting automatic data mining and analysis is therefore an important goal of material informatics[58].

Despite notable differences, as mentioned above, there is overlap of predicted materials between the approaches. All approaches (including the extended criteria-based approach) and their corresponding ML methods agree on a total of 47 potential candidates to 0.5 confidence (see the Supplementary Information at ref. [37] for the full list). Several interesting materials that were also discussed above for the predictions by the empirical approach are present among these 47, including BP, CdSe, GeC, InP and Ge. However, certain materials that cannot be classified as semiconductors are also included, such as P, I, N_2 and H_2. We note that none of these were included by the empirical approach when the threshold was set to 0.75 or above.

Importantly, there are 6 materials that all approaches (criteria-based, extended criteria-based and empirical) and ML methods agree on above a 0.75 threshold. These are $ZnGeAs_2$ (tetragonal mp-4008), $CdSnP_2$ (tetragonal mp-5213), GeSe (cubic mp-10759),

InP (cubic mp-20351), InP (hexagonal mp-966800) and SiSn (cubic mp-1009813). Here, we can distinguish three groupings in crystal structure: cubic, hexagonal and tetragonal (illustrated in Fig. 6). The relevant material properties are summarized in the lower part of Table 3, with comparable trends to those for the 9 materials predicted by the empirical approach in terms of, e.g., crystal structure. The six-fold coordination of GeSe is an outlier as compared to the four-fold coordination of the other materials. We highlight these six compounds, along with the nine predicted by the empirical approach to 0.85 confidence, as particularly interesting for future in-depth theoretical and experimental studies.

Taking a closer look at the reasoning behind the choices made by the different ML methods during the classification process, we can start to identify important driving forces for manifestation of quantum compatible properties in semiconductors. The analysis of the principal components extracted from the ML methods revealed that the most important principal component for the criteria-based approach encompasses features related to the band gap and chemical environment. This means that the band gap criterion imposed in the training set selection is at least somewhat satisfied. The criteria-based approach does not entirely reproduce the logic of the initial selection process, however, as several low band gap (<0.5 eV) materials were highlighted as suitable by the ML methods. For the empirical approach, on the other hand, band gap related features were not recognized as important in the dominant principal component.

Figure 7a displays the number of predicted suitable materials as a function of band gap (from the MP database), and reveals that both approaches predicted a substantial amount of materials with a low band gap (below 0.5 eV). Moreover, the distribution of materials that were identified as suitable is rather broad with the criteria-based approach exhibiting a peak around a 2.5 eV band gap. Coincidentally, the band gap of, e.g., 4*H*-SiC is usually computed at around 2.5 eV using the PBE functional. The empirical approach exhibits a more scattered data distribution.

The second most important principal component in the criteria-based approach encompasses properties such as the ionic character, covalent radius and maximum packing efficiency (see the Supplementary Information at ref. [37] for the predicted material distributions of the latter two features). Intriguingly, as shown in Fig. 7b, the predicted candidates distribute over a broad range of ionic characters for both the criteria-based and empirical approaches—even peaking at a relatively high ionic character of 0.8. We note that this may be a result of the distribution of the initial dataset of 25,000 materials with a maximum around a similar value (not shown). The minor peak in the empirical approach's predicted materials around 0.4 ionic character is not present for the criteria-based approach nor the overall data distribution. For reference, all SiC entries in the dataset have maximum ionic characters of ~0.1. Furthermore, the covalent radii of the materials (see the Supplementary Information) exhibit two distinct peaks in the data distribution. The trend of two data peaks is repeated for the maximum packing efficiency but is much more prominent for the empirical approach. This indicates that the material density, or in other words the bond length, is an important parameter for QT suitability.

The ML methods in the empirical approach consistently identified the first principal component as the predominant one. Identifying the single most important feature in this principal component proved challenging as it is the combined impact of several features that matters. Here, the standard deviation of the radial distribution function (RDF) has a particularly strong impact since it appears four times in different forms in the top ten list over dominating features. One configuration of the standard deviation of the RDF is demonstrated in Fig. 7c, with two others being included in the Supplementary Information at ref. [37]. Intriguingly, the standard deviation of the RDF exhibits substantial discrepancies between the criteria-based and the empirical approach. For the empirical approach, there is a sharp peak in the preferred value for the standard deviation of the RDF, while the criteria-based approach displays an even distribution across a broader range. These observations emphasize the importance of symmetry related material properties for QT suitability.

We interpret the standard deviation of the RDF such that zero standard deviation in the RDF means that there is no variation in the radial symmetry throughout the material. Similarly, zero standard deviation in the average bond length would mean that all bonds are identical throughout the crystal. Intriguingly, the peaks in the distributions of the predicted materials are not found for perfectly symmetric materials with identical bond lengths; instead, some variation in the bond and wavefunction distributions is found to be optimal for a material to be suitable for QT. Note that the maxima in Fig. 7c seem to appear for moderate standard deviations in the RDF, indicating that a certain degree of symmetry is necessary for a material to act as a suitable QT host. The exact degree and type of symmetry is still open to debate and merits further study. Similar symmetry related features such as the standard deviation of the average bond length (see Fig. 7d), the site fingerprint of the chemical environment and the bond orientation (see the Supplementary Information at ref. [37]) are also influential in guiding the ML algorithms upon classifying suitable and unsuitable materials for QT in the empirical approach.

Interestingly, none of the nine materials predicted by the empirical approach above a 0.85 threshold, nor of the six materials predicted by all approaches to 0.75 confidence, are elemental. At most three different elements are present, but the emphasis is clearly on binary compounds. This resonates with the observations from the principal component analysis; an optimal degree of crystalline order likely exists for a material to manifest QT compatible properties, but some variations throughout the crystal in, e.g., bond length and symmetry are needed. This is in contrast to the expectations that guided the formulation of the test and training sets in the criteria-based approach. Where most previous works have highlighted features such as band gap, polarity and ionic character as vital for a semiconductor to manifest quantum compatible features, our results reveal that local variations in the crystal structure related to symmetry and bond angles, length and orientation should be considered equally and may even be more important.

DISCUSSION

Considering the trends in machine learning selectivity in light of the specific defect centers we know to be quantum compatible reveals fascinating characteristics. None of the known quantum emitters or spin centers seem to appear in completely uniform systems. For example, in high symmetry materials like diamond and silicon, QT compatible defects are not intrinsic; neither the silicon vacancy in silicon nor the carbon vacancy in diamond, for example, have exhibited single-photon emission or controllable spin coherence. Instead, quantum effects often appear after impurities are introduced, as evidenced by the phosphorous and carbon impurities in silicon[59,60] and the nitrogen-vacancy, germanium-vacancy, tin-vacancy and lead-vacancy centers in diamond[61]. For a binary system like silicon carbide, on the other hand, intrinsic defects like the silicon vacancy and the divacancy are appropriate for our goals. While these trends have yet to be verified on a grander scale, our findings provide strong indications that local variations in the crystal structure are paramount for QT compatible properties to manifest in a semiconductor host.

It should be noted that performing machine learning on a dataset derived using preconceived notions for which material properties are important may reproduce several of the initial selection criteria. Nonetheless, the criteria-based approach was included in the present work to highlight expectations from the

literature and contrast them with the findings from the empirical one. Additionally, ML methods are often capable of recognizing other patterns than those intended for the data, opening up the possibility that also the criteria-based approach could yield new insights. Finally, the criteria-based approaches are employed as a filter on the empirical approach, to provide a better tuned list of candidates for future experimental studies.

To summarize, we have developed data extraction tools and strategies for data mining and labeling to enable the automated search and analysis of host materials for QT. The clear separation between suitable and unsuitable candidates after data labeling, along with the smaller number of principal components needed for obtaining optimal performance of the ML algorithms, indicate that the empirical approach based on findings from the literature is highly suitable for performing this type of guided search through materials databases. The principal component analyses of the ML methods' performance imply that the criteria-based approach with its strong focus on band gap and bonding character when assessing a material's quantum compatibility to some extent reproduces the specifications of the data labeling process, as expected. Valuable insight is gained from comparing the important features of the Ferrenti with those for the empirical approach which highlights the importance of symmetry related properties in the bond orientations and wavefunctions over expected features related to band gap and bonding. As such, the expected properties are not able to capture the full physics of the problem. The contrast between the criteria-based and empirical approaches reveals that the problem of QT compatibility is more complex than being related to band gap and bonding character alone. Based on our findings we propose that the manifestation of quantum effects in semiconductors is related to the crystal structure symmetry and bonding.

The findings presented here firmly establish that material informatics is a viable and important route to new discoveries in important fields. Our focus has been on predicting new candidate materials to host single-photon emitters and spin centers for quantum technology applications, but the developed framework is suitable for other fields as well. Two possible paths are suggested to further exploit the findings presented herein. One aspect is the pursuit of experimental verification of QT compatible effects in the materials predicted as suitable by machine learning. Another, perhaps even more important, route is to use the features and trends identified during the data mining and prediction processes to understand the distinct material characteristics that enable quantum effects to manifest, opening thereby up for new discoveries in the field of quantum technologies.

METHODS

Databases

The Materials Project[22,23] is an open-source project containing ground state properties of materials calculated using density functional theory (DFT) as implemented in the Vienna Ab initio Simulation Package (VASP)[24]. The Perdew-Burke-Ernzerhof[25] (PBE) functional is used to calculate band structures, while for transition metals, a $+U$ correction is applied to correct for correlation effects[62]. The project is known as the initiator of materials genomics and offers a variety of calculated properties for over one hundred thousand inorganic crystalline materials, with frequent updates and extensions. Data extraction from Materials Project was performed in December of 2020 for the criteria-based and extended criteria-based approaches, and in March 2021 for the empirical approach. Therefore, the initial dataset for the two former approaches includes 77 more materials than that for the empirical approach due to erroneous entries that have been removed from the Materials Project database.

The Open Quantum Materials Database (OQMD)[28,29] contains thermodynamic and structural properties of more than 600,000 materials. The calculations are performed with the VASP software and the electron exchange and correlation are described with the PBE functional. The $+U$ extension is included for several calculations considering specific elements[63]. Data extraction from OQMD was done in February of 2021.

JARVIS-DFT[30] is an open-source database based on the VASP software and consists of roughly 40,000 three-dimensional materials using the vdW-DF-OptB88 (OPT) functional[64,65]. Structures included in the database are originally taken from the Materials Project[22,23], and then re-optimized using the OPT functional. Finally, the combination of the OPT and modified Becke-Johnson (mBJ) functionals[66] is used to obtain a representative band gap for each structure[67]. Data extraction from JARVIS-DFT was done in January of 2021, were we utilized the version made available on 2021-04-30 (see ref. [30]).

The AFLOW[31–33] repository is an automatic software framework for the calculations of a wide range of inorganic material properties. They utilize the PBE functional (with the $+U$ correction for certain cases) within VASP to relax and optimize all structures from the ICSD. Data extraction was performed in the period of January to February of 2021.

AFLOW-ML[34] is an application programming interface (API) that uses machine learning to predict thermo-mechanical and electronic properties based on the chemical composition and atomic structure alone, which are denoted as *fragment descriptors*. Initially, the API decides whether a given material is a metal or an insulator, where the latter is confirmed with an additional regression method to predict the band gap. The accuracy is validated by a five-fold cross-validation process for each ML method, where they report a 93% prediction success of their initial binary classification method. In this work we utilized the Property Labeled Material Fragments (PLMF) openly available at their website (see ref. [34]). We extract the crystallographic information files (CIF) for the crystals from Materials Project, use the CIF files as input to AFLOW-ML, which then returns an anticipated band gap. This process was executed during January of 2021.

Note that from the initial criterion, we define that a material is required to have an ICSD identifier (ID) in Materials Project. This ID is included in the AFLOW, AFLOW-ML, JARVIS-DFT and OQMD databases as well. If two different databases contain different values for the same property, such as the band gap, we added both of them as columns (features) to our data to avoid any data cluttering. An additional analysis was performed to uncover any differences between the space groups reported in the Materials Project and the other databases, yielding an average match of 97%. We note that the small deviation might arise due to errors in either database, and is not necessarily reflected in the remaining features of the data. For the experimental values from Citrine, we could only verify the chemical formula since the experimental data is lacking information regarding the structure (e.g., space group, symmetry) of the material.

The data regarding a material's magnetic character is extracted from the Materials Projects database. Indeed, the majority of these calculations are based on the primitive cell, however, Materials Project performs an initial relaxation of cell and lattice parameters using a $1000 \cdot (\text{number of atoms in the cell})^{-1}\mathbf{k}$-point mesh to ensure that all properties calculated are representative of the idealized unit cell for each respective structure. As a result, we can find, e.g., Fe labeled as ferromagnetic and NiO as antiferromagnetic in our dataset. Furthermore, MP contains structures of varying sizes for the same material. We have included all structures and materials from the Materials Project, but we have not checked which materials are represented as a larger cell in our data. We have thereby not verified whether antiferromagnetic ordering has been investigated for all cases. This is an improvement that could be made to our method in a future study.

Material informatics

Matminer[36] is an open source toolkit for material analysis written in Python. Matminer provides modules to extract information from a wide variety of databases. Additionally, they provide the tools to construct possibly thousands of features from calculations based on a material's composition, structure and electronic properties from DFT calculations, and have frameworks for visualization and automatic machine learning. To apply Matminer's featurization tools, we extend an existing implementation by ref. [68], which was used to generate a supervised machine learning framework called the MODnet. The implementation by ref. [68] provides featurization for a material's composition, structure and atomic sites. However, Matminer also provides featurization tools for a material's density of states (DOS) and band structure. Therefore, we extend their implementation to facilitate such featurizations. The features selected for featurization herein are summarized in the Supplementary Information at ref. [37].

Pymatgen, a robust and open-source Python library for material analysis[69], was also employed to extract and generate features for several of the databases mentioned above.

Machine learning

Machine learning represents the science of giving computers the ability to learn without being explicitly programmed. The idea is that generic algorithms exist which can be used to find patterns in a broad class of datasets without having to write code specifically for each problem. The algorithm builds its own logic based on the data.

The approaches to machine learning are many, but are often split into two main categories: supervised and unsupervised. In supervised learning we know the answer to a problem, and let the computer deduce the logic behind it. On the other hand, unsupervised learning is a method for identifying patterns and relationships in datasets without any prior knowledge of the system. Many researchers also operate with a third category, namely reinforcement learning. This is a paradigm of learning inspired by behavioral psychology, where learning is achieved by trial-and-error, solely from rewards and punishment. In this work our focus is on supervised learning only with labeled data for classification problems.

In this work we have applied four well-known and tested ML methods for classification problems, these are (see for example[18,19] for discussions and applications):

1. Logistic regression,
2. Decision trees,
3. Random forests,
4. Gradient boosting.

Logistic regression[19] is a simple and frequently used method for binary and multi-category classification problems. In addition to logistic regression, we have also applied and tested the predictions made by decision trees and ensemble methods like random forests and gradient boosting, the latter through the application of the computationally efficient XGBoost library[70]. Gradient boosting and random forests use decision trees as weak learners and improve their predictability. For random forests this is implemented through a collection of randomized decision trees where a subset of the features in the datasets are selected randomly when building a decision tree. Boosting methods like gradient boosting use decision trees as weak learners and improve upon these by an iterative process that involves the estimation of the gradients of the cost/loss function[19]. Pure decision trees can easily lead to overfitting of the data under study, leading to an ML method that exhibits a high variance. Ensemble methods like random forests and gradient boosting on the other hand tend to soften the overfitting problem, resulting in both a small bias and a reduced variance of the employed method, see for example refs. [18,19] for an in-depth discussion of the bias-variance trade-off in machine learning. Gradient boosting implemented through the XGBoost library[70] is widely used by data scientists to achieve state-of-the-art results on many machine learning challenges.

Our findings are corroborated by the fact that all four ML methods predict a small set of the unlabeled materials as suitable, while agreeing on a large part of these materials. The methods we have chosen are all well tested, with random forests and gradient boosting methods tending to outperform the others, resulting normally in a small bias and variance[18–20].

DATA AVAILABILITY

The datasets that support the findings of this study are available online at ref. [21].

CODE AVAILABILITY

The codes employed to develop the machine learning results are available online at ref. [21].

Received: 6 April 2022; Accepted: 1 September 2022;
Published online: 28 September 2022

REFERENCES

1. Le Sage, D. et al. Optical magnetic imaging of living cells. *Nature* **496**, 486–489 (2013).
2. Ursin, R. et al. Entanglement-based quantum communication over 144 km. *Nat. Phys.* **3**, 481–486 (2007).
3. Arute, F. et al. Quantum supremacy using a programmable superconducting processor. *Nature* **574**, 505–510 (2019).
4. Chow, J., Dial, O. & Gambetta, J. IBM quantum breaks the 100-qubit processor barrier. https://research.ibm.com/blog/127-qubit-quantum-processor-eagle (2021).
5. Kimble, H. J. The quantum internet. *Nature* **453**, 1023–1030 (2008).
6. Aharonovich, I., Englund, D. & Toth, M. Solid-state single-photon emitters. *Nat. Photon.* **10**, 631–641 (2016).
7. Awschalom, D. D., Hanson, R., Wrachtrup, J. & Zhou, B. B. Quantum technologies with optically interfaced solid-state spins. *Nat. Photon.* **12**, 516–527 (2018).
8. Weber, J. R. et al. Quantum computing with defects. *Proc. Natl. Acad. Sci. USA* **107**, 8513–8518 (2010).
9. Doherty, M. W. et al. The nitrogen-vacancy colour centre in diamond. *Phys. Rep.* **528**, 1–45 (2013).
10. Castelletto, S., Rosa, L. & Johnson, B. C. Silicon carbide for novel quantum technology devices. In *Advanced Silicon Carbide Devices and Processing* (InTech, 2015).
11. Son, N. T. et al. Developing silicon carbide for quantum spintronics. *Appl. Phys. Lett.* **116**, 190501 (2020).
12. Bathen, M. E. & Vines, L. Manipulating single-photon emission from point defects in diamond and silicon carbide. *Adv. Quantum Technol.* **4**, 2100003 (2021).
13. Atatüre, M., Englund, D., Vamivakas, N., Lee, S.-Y. & Wrachtrup, J. Material platforms for spin-based photonic quantum technologies. *Nat. Rev. Mater.* **3**, 38–51 (2018).
14. Zhang, G., Cheng, Y., Chou, J.-P. & Gali, A. Material platforms for defect qubits and single-photon emitters. *Appl. Phys. Rev.* **7**, 031308 (2020).
15. Deiana, A. M. et al. Applications and techniques for fast machine learning in science. *Front. Big Data* **5**, 787421 (2022).
16. Carleo, G. et al. Machine learning and the physical sciences. *Rev. Mod. Phys.* **91**, 045002 (2019).
17. Ferrenti, A. M., de Leon, N. P., Thompson, J. D. & Cava, R. J. Identifying candidate hosts for quantum defects via data mining. *npj Comput. Mater.* **6**, 1 (2020).
18. Mehta, P. et al. A high-bias, low-variance introduction to machine learning for physicists. *Phys. Rep.* **810**, 1–124 (2019).
19. Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning: Data Mining, Inference and Prediction (Springer Verlag, Berlin, 2009), 2nd edn.
20. Murphy, K. Machine learning : a probabilistic perspective (MIT Press, 2012).
21. Hebnes, O. L. Predicting solid-state qubit material hosts. https://zenodo.org/record/6345880 (2022).
22. Jain, A. et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).

23. Jain, A. et al. The materials project: Accelerating materials design through theory-driven data and tools. In *Handbook of Materials Modeling* (Springer International Publishing, 2018).
24. Kresse, G. & Furthmüller, J. Efficient iterative schemes fo rab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
25. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
26. Freysoldt, C. et al. First-principles calculations for point defects in solids. *Rev. Mod. Phys.* **86**, 253–305 (2014).
27. Zagorac, D., Müller, H., Ruehl, S., Zagorac, J. & Rehme, S. Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. *J. Appl. Crystallogr.* **52**, 918–925 (2019).
28. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *JOM* **65**, 1501–1509 (2013).
29. Kirklin, S. et al. The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
30. Choudhary, K. et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput. Mater.* **6**, 173 (2020).
31. Curtarolo, S. et al. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
32. Curtarolo, S. et al. AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
33. Calderon, C. E. et al. The AFLOW standard for high-throughput materials science calculations. *Comput. Mater. Sci.* **108**, 233–238 (2015).
34. Isayev, O. et al. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **8**, 1 (2017).
35. O'Mara, J., Meredig, B. & Michel, K. Materials data infrastructure: A case study of the citrination platform to examine data import, storage, and access. *JOM* **68**, 2031–2034 (2016).
36. Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
37. Hebnes, O. L. et al. Supplementary Information available at https://github.com/mhjensen/PredictingSolidStateQubitCandidates (2022).
38. Toth, M. & Aharonovich, I. Single photon sources in atomically thin materials. *Annu. Rev. Phys. Chem.* **70**, 123–142 (2019).
39. Hardy, W. J. et al. Single and double hole quantum dots in strained Ge/SiGe quantum wells. *Nanotechnology* **30**, 215202 (2019).
40. Inselberg, A. & Dimsdale, B. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In Proceedings of the First *IEEE* Conference on Visualization: Visualization 90 (IEEE Comput. Soc. Press, 1990).
41. Inselberg, A. The plane with parallel coordinates. *Vis. Comput.* **1**, 69–91 (1985).
42. Jolliffe, I. T. Principal Component Analysis (Springer, Berlin, 2002).
43. Beleites, C. & Salzer, R. Assessing and improving the stability of chemometric models in small sample size situations. *Anal. Bioanal. Chem.* **390**, 1261–1271 (2008).
44. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 1–7 (2016).
45. Gao, Y. et al. Superhard sp2-sp3 hybridized bc2n: A 3d crystal with 1d and 2d alternate metallicity. *J. Appl. Phys.* **121**, 225103 (2017).
46. Cai, Y., Xiong, J., Liu, Y. & Xu, X. Electronic structure and chemical hydrogen storage of a porous sp^3 tetragonal BC_2N compound. *J. Alloys Compd.* **724**, 229–233 (2017).
47. Li, H., Xiao, X., Tie, J. & Lu, J. Electronic and magnetic properties of bare armchair BC_2N nanoribbons. *J. Magn. Magn. Mater.* **426**, 641–645 (2017).
48. Jiang, C.-L., Zeng, W., Liu, F.-S., Tang, B. & Liu, Q.-J. The shape type of bonds and the direction of phonons in orthorhombic BC_2N from first-principles calculations. *J. Phys. Chem. Solids.* **140**, 109349 (2020).
49. Wang, D., Liu, L. & Zhuang, H. L. Spin qubit based on the nitrogen-vacancy center analog in a diamond-like compound C_3BN. *J. Appl. Phys.* **130**, 225702 (2021).
50. Liu, J. et al. Single self-assembled InAs/GaAs quantum dots in photonic nanostructures: The role of nanofabrication. *Phys. Rev. Appl.* **9**, 064019 (2018).
51. Ahn, D. et al. Intrinsically p-type cuprous iodide semiconductor for hybrid light-emitting diodes. *Sci. Rep.* **10**, 3995 (2020).
52. Zhang, S. R., Xie, L. H., Ouyang, S. D., Chen, X. W. & Song, K. H. Electronic structure, chemical bonding and optical properties of the nonlinear optical crystal $ZnGeP_2$ by first-principles calculations. *Phys. Scr.* **91**, 015801 (2015).
53. Xing, G. C., Bachmann, K. J., Posthill, J. B. & Timmons, M. L. $ZnGeP_2$: A wide bandgap chalcopyrite structure semiconductor for nonlinear optical applications. *MRS Proc.* **162**, 615 (1989).
54. Caspani, L. et al. Integrated sources of photon quantum states based on nonlinear optics. *Light Sci. Appl.* **6**, e17100–e17100 (2017).
55. Mukhanov, V. A., Vrel, D., Sokolov, P. S., Le Godec, Y. & Solozhenko, V. L. Ultra-fast mechanochemical synthesis of boron phosphides, BP and $B_{12}P_2$. *Dalton Trans.* **45**, 10122–10126 (2016).
56. Zhang, H. et al. High-brightness blue InP quantum dot-based electroluminescent devices: The role of shell thickness. *J. Phys. Chem.* **11**, 960–967 (2020).
57. Won, Y.-H. et al. Highly efficient and stable InP/ZnSe/ZnS quantum dot light-emitting diodes. *Nature* **575**, 634–638 (2019).
58. Rickman, J., Lookman, T. & Kalinin, S. Materials informatics: From the atomic-level to the continuum. *Acta Mater.* **168**, 473–510 (2019).
59. He, Y. et al. A two-qubit gate between phosphorus donor electrons in silicon. *Nature* **571**, 371–375 (2019).
60. Redjem, W. et al. Single artificial atoms in silicon emitting at telecom wavelengths. *Nat. Electron.* **3**, 738–743 (2020).
61. Thiering, G. & Gali, A.Color centers in diamond for quantum applications. In *Diamond for Quantum Applications*, 1–36 (Elsevier, 2020).
62. Wang, L., Maxisch, T. & Ceder, G. Oxidation energies of transition metal oxides within the GGA + U framework. *Phys. Rev. B* **73**, 195107 (2006).
63. Stevanović, V., Lany, S., Zhang, X. & Zunger, A. Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies. *Phys. Rev. B* **85**, 115104 (2012).
64. Thonhauser, T. et al. Van der Waals density functional: Self-consistent potential and the nature of the van der Waals bond. *Phys. Rev. B* **76**, 125112 (2007).
65. Klimeš, J., Bowler, D. R. & Michaelides, A. Van der Waals density functionals applied to solids. *Phys. Rev. B* **83**, 195131 (2011).
66. Tran, F. & Blaha, P. Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential. *Phys. Rev. Lett.* **102**, 226401 (2009).
67. Choudhary, K. et al. Computational screening of high-performance optoelectronic materials using OptB88vdw and TB-mBJ formalisms. *Sci. Data* **5**, 180082 (2018).
68. Breuck, P.-P. D., Evans, M. L. & Rignanese, G.-M. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODNet. *J. Phys. Condens. Matter* **33**, 404002 (2021).
69. Ong, S. P. et al. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
70. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, 785-794 (Association for Computing Machinery, New York, NY, USA*, 2016).
71. Taylor, J. M. et al. High-sensitivity diamond magnetometer with nanoscale resolution. *Nat. Phys.* **4**, 810–816 (2008).
72. Balasubramanian, G. et al. Ultralong spin coherence time in isotopically engineered diamond. *Nat. Mater.* **8**, 383–387 (2009).
73. Barclay, P. E., Fu, K.-M. C., Santori, C., Faraon, A. & Beausoleil, R. G. Hybrid nanocavity resonant enhancement of color center emission in diamond. *Phys. Rev. X* **1**, 011007 (2011).
74. Gordon, L. et al. Quantum computing with defects. *MRS Bull.* **38**, 802–807 (2013).
75. Rogers, L. J. et al. All-optical initialization, readout, and coherent preparation of single silicon-vacancy spins in diamond. *Phys. Rev. Lett.* **113**, 263602 (2014).
76. Bhaskar, M. K. et al. Quantum nonlinear optics with a germanium-vacancy color center in a nanoscale diamond waveguide. *Phys. Rev. Lett.* **118**, 223603 (2017).
77. Widmann, M. et al. Coherent control of single spins in silicon carbide at room temperature. *Nat. Mater.* **14**, 164–168 (2014).
78. Christle, D. J. et al. Isolated electron spins in silicon carbide with millisecond coherence times. *Nat. Mater.* **14**, 160–163 (2015).
79. Castelletto, S. et al. A silicon carbide room-temperature single-photon source. *Nat. Mater.* **13**, 151–156 (2014).
80. Zargaleh, S. A. et al. Nitrogen vacancy center in cubic silicon carbide: A promising qubit in the 1.5μm spectral range for photonic quantum networks. *Phys. Rev. B* **98**, 165203 (2018).
81. Falk, A. L. et al. Polytype control of spin qubits in silicon carbide. *Nat. Commun* **4**, 1819 (2013).
82. Muhonen, J. T. et al. Storing quantum information for 30 seconds in a nanoelectronic device. *Nat. Nanotechnol.* **9**, 986–991 (2014).
83. Durand, A. et al. Broad diversity of near-infrared single-photon emitters in silicon. *Phys. Rev. Lett.* **126**, 083602 (2021).
84. Tran, T. T. et al. Robust multicolor single photon emission from point defects in hexagonal boron nitride. *ACS Nano* **10**, 7331–7338 (2016).
85. 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).
86. Hayee, F. et al. Revealing multiple classes of stable quantum emitters in hexagonal boron nitride with correlated optical and electron microscopy. *Nat. Mater.* **19**, 534–539 (2020).
87. Morfa, A. J. et al. Single-photon emission and quantum characterization of zinc oxide defects. *Nano Lett.* **12**, 949–954 (2012).
88. Stewart, C. et al. Quantum emission from localized defects in zinc sulfide. *Opt. Lett.* **44**, 4873 (2019).
89. Bluhm, H. et al. Dephasing time of GaAs electron-spin qubits coupled to a nuclear bath exceeding 200 μs. *Nat. Phys.* **7**, 109–113 (2010).
90. Roux, F. L. et al. Temperature dependence of the single photon emission from interface-fluctuation GaN quantum dots. *Sci. Rep.* **7**, 16107 (2017).

91. Berhane, A. M. et al. Photophysics of GaN single-photon emitters in the visible spectral range. *Phys. Rev. B* **97**, 165202 (2018).
92. Xue, Y. et al. Single-photon emission from point defects in aluminum nitride films. *J. Phys. Chem.* **11**, 2689–2694 (2020).

ACKNOWLEDGEMENTS

The work of L.V. and M.E.B. was supported by the Research Council of Norway and the University of Oslo through the frontier research projects FUNDAMeNT (no. 251131) and QuTe (no. 325573). The work of M.E.B. was supported by an ETH Zurich Postdoctoral Fellowship. The work of M.H.J. was supported by the U.S. Department of Energy, Office of Science, office of Nuclear Physics under grant No. DE-SC0021152 and U.S. National Science Foundation Grants Nos. PHY-1404159 and PHY-2013047. The work of SGWL and øSS was supported by the Norwegian Directorate for International Cooperation and Quality Enhancement in Higher Education (DIKU) which supports the Center for Computing in Science Education (CCSE).

AUTHOR CONTRIBUTIONS

M.E.B., L.V. and M.H.J. conceived the main theme of the project. O.H. developed the programs and performed the bulk of the work while M.E.B. lead the writing process. All authors have contributed to the writing of the paper and the discussion and analysis of the data.

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/s41524-022-00888-3.

Correspondence and requests for materials should be addressed to Marianne Etzelmüller Bathen.

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.