RSC Advances

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *RSC Adv.*, 2025, **15**, 8253

Many-body physics and machine learning enabled discovery of promising solar materials†

Tathagata Biswas, * Adway Gupta * and Arunima K. Singh*

In recent years, GW-BSE has been proven to be extremely successful in studying the quasiparticle (QP) bandstructures and excitonic effects in the optical properties of materials. However, the massive computational cost associated with such calculations restricts their applicability in high-throughput material discovery studies. Recently, we developed a Python workflow package, *py*GWBSE, to perform high-throughput GW-BSE simulations. In this work, using *py*GWBSE we create a database of various QP properties and excitonic properties of over 350 chemically and structurally diverse materials. Despite the relatively small size of the dataset, we obtain highly accurate supervised machine learning (ML) models *via* the dataset. The models predict the quasiparticle gap with an RMSE of 0.36 eV, exciton binding energies of materials with an RMSE of 0.29 eV, and classify materials as high or low excitonic binding energy materials with classification accuracy of 90%. We exemplify the application of these ML models in the discovery of 159 visible-light and 203 ultraviolet-light photoabsorber materials utilizing the Materials Project database.

Received 21st February 2025
Accepted 4th March 2025

DOI: 10.1039/d5ra01285f

rsc.li/rsc-advances

1 Introduction

Light-matter interaction is the fundamental physical phenomenon behind a wide variety of existing high-impact applications such as photovoltaics, photocatalysis, medical diagnostics, scientific instrumentation, imaging (*e.g.*, infrared imagers), and sensing (*e.g.*, for light detection) along with potential future quantum devices.[1–3] New opportunities to optimize the performance of these existing optical devices and to pave the road for emerging fields can be realized by the discovery and design of novel functional materials.

The Materials Genome Initiative (MGI)[4,5] was proposed in 2011 to enable the discovery, manufacturing, and deployment of advanced materials twice as fast and at a fraction of the cost compared to traditional methods. To achieve the MGI objectives one of the key strategies adopted was to harness the power of data and computational tools jointly with experimental investigations.[4,5] Since then a new paradigm for accelerated materials discovery has emerged by designing new compounds *in silico* using first-principles calculations and then performing experiments on the computationally designed candidates.[6–8] Several open-source databases have been developed to aid the accelerated material discovery goal such as the Materials Project,[9] Aflowlib,[10] C2DB,[11] ESP,[12] NoMaD,[13] OQMD[14] *etc.* The availability of such large data has opened up an emerging paradigm, the application of machine learning (ML) and other data science methods for material discovery, thus, making material discovery essentially a big-data problem.[15–20] ML accelerated material discovery has made a revolutionary impact in applications ranging from organic and solid-state LEDs, batteries, ferroelectric, high-κ dielectric, hydrogen storage, high-entropy alloys, and thermoplastics to shape memory alloys.[16,19,21]

The fundamental challenge of applying this well-established material discovery paradigm of combining first-principles computations and data science methods to applications where light–matter interaction is a key phenomenon is the unavailability of large datasets that are accurate enough in comparison to experimental observations. While some first-principles methods such as GW-BSE (Bethe–Salpeter equation) formalism for simulating excitonic effects can produce optical properties with sufficient accuracy they are computationally very expensive. Thus it is not surprising that the largest database of GW-BSE computed absorption spectra is that of ∼300 spectra of two-dimensional (2D) materials.[11,22]

The unavailability of large-scale data of first principles computed excited state properties can also be partially attributed to the lack of open-source computational tools to perform such high-throughput computations. Recently, the authors have developed *py*GWBSE,[23] a python workflow package that enables automated high-throughput GW-BSE simulations using VASP, one of the most widely used first-principles atomistic modeling software, and made it available through open-source licensing. While there have been recent efforts in the

---

Department of Physics, Arizona State University, Tempe, Arizona, 85281, USA. E-mail: arunimasingh@asu.edu

† Electronic supplementary information (ESI) available: The performance of various ML algorithms applied in this study, a detailed description of the features used in ML models, and a list of materials shortlisted for visible and UV light-based applications along with their ML-predicted properties. See DOI: **https://doi.org/10.1039/d5ra01285f**

development of similar workflow codes, they either suffer from issues like the absence of BSE capabilities[24] and database integration,[25] or use other *ab initio* modeling software for the excited-states calculations[26]

In this study, we demonstrate the applicability of ML models in predicting accurate excited state properties such as quasi-particle gap (QPG) and exciton binding energy (EBE). To accomplish this goal, we generated the largest database of first-principles computed QP and excitonic properties of bulk materials computed and curated using *py*GWBSE. This database contains static dielectric constants, effective masses, QP bandstructure, absorption spectra, and several other excited state properties such as EBE, integrated absorption coefficients, *etc.* of more than 350 bulk materials, and new materials are being added to the database continuously. We find that among the various ML regression algorithms, Random Forest Regression has the best performance for predicting QPG within an RMSE of 0.36 and EBE within an RMSE of 0.29 eV. We also find ML models that accurately classify materials to have integrated absorption coefficient (IAC),[2] anisotropy in absorption coefficient (AAC),[2] and excitonic binding energies suited for a good photoabsorber with an accuracy of ∼90%. Lastly, we apply the ML models developed in this work to identify promising materials that can absorb visible and ultraviolet (UV) radiation for photovoltaic or photocatalytic applications based on their QP and excitonic properties from a list of ∼7000 materials for which only the ground state properties are available in the Materials Project database.

2 Results and discussions

As we intend to develop and apply ML models for predicting the QP and excitonic properties of a wide variety of materials, we first demonstrate the diversity of the materials within our GW-BSE computed dataset. In Fig. 1 we classify all the materials in the dataset according to their chemical composition, crystal systems, and the number of constituent atomic species. As one can see from Fig. 1(a), our materials dataset has a significant fraction of materials from all the seven crystal systems, except the triclinic phase, which constitutes less than 1% of our training set. Fig. 1(b) shows that our training set of materials consists of almost equal percentages of oxides, pnictides, halides, and chalcogenides. In addition, 26% of the materials in the training set don't fall in any of the aforementioned chemical groups. Given the diversity in the chemical compositions and the crystal symmetries of the materials in our dataset, we can expect that the ML model trained with the data can be used for a wide variety of materials. Moreover, we also look at the number of unique chemical species in the selected materials in Fig. 1(c) which shows that most of the materials in our dataset are binary (58%) and ternary (29%) compounds, which ensures that our ML model captures diverse chemical properties of multi-element compounds.

The diversity of our GW-BSE computed database is unique in terms of all the aforementioned characteristics. The only other database that hosts GW-BSE computed properties is that of Hasstrup *et al.*[11] which is limited to two-dimensional (2D) materials and therefore restrictive in terms of crystal systems and chemical compositions.

In the following two sections, Sections 2.1 and 2.2, we present the accuracy of various ML algorithms in predicting the QP and excitonic properties of materials. In particular, we focus on the QPG and the EBE of materials. We also discuss the most important features that were used in these models and their physical significance. In Section 2.3 we present ML models for classifying materials as high or low excitonic binding energy materials. Finally, Section 2.4 demonstrates how the ML models developed in this work allow the discovery of visible-light and UV-light photoabsorber materials by utilizing existing materials database without the need for any explicit GW-BSE simulations.

2.1 QP gap prediction

The bar plot in Fig. 2(a) shows the distribution of the QP gaps of the 314 materials' dataset. One can see that the materials have a broad range of QP gaps (0–15 eV).

For the ML prediction of the QPG, we tested four regression methods – the kernel ridge regression (KRR), the random forest regression (RF), the support vector machine (SVM), and the multi-layer perceptron (MLP)[27,28] methods. A 10-fold cross-validation was employed to ensure randomness in the training and test datasets. Note that while our dataset consists of GW-BSE calculations of nearly 350 materials about 35 of them were not included in the ML model development as they had unphysical values for some of the features considered in this study. The model obtained from the RF method performed best for the QP bandgap prediction with an R^2 score of 0.98 and

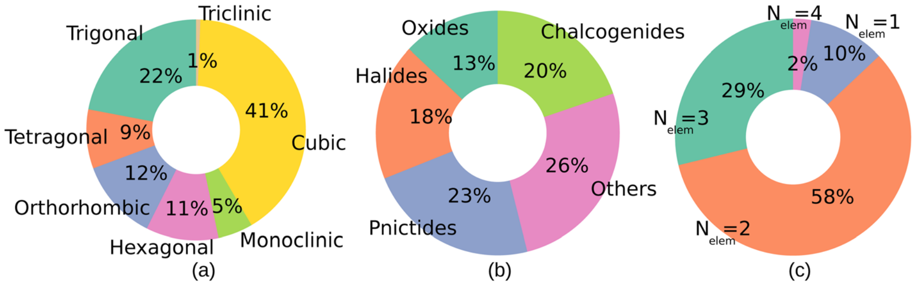
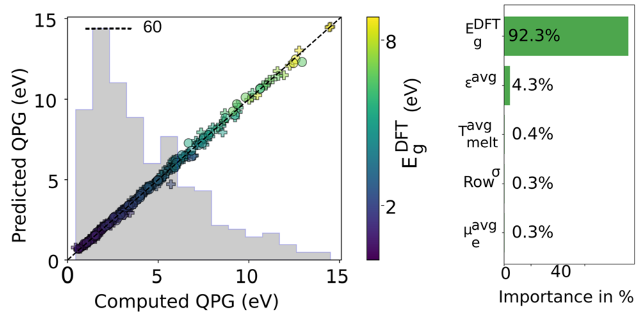


Fig. 1 A classification of the bulk materials' dataset considered in this work in terms of (a) the crystal system, (b) chemical composition, and (c) the number of constituent elements. The figure shows that the materials in the dataset are very diverse.



(a) (b)

Fig. 2 (a) Random forest regression model predicted QPG plotted against GW computed QPG. The training and test set data points are shown as + symbols and circle symbols, respectively. The distribution of the QPGs corresponding to the entire dataset is also shown in the background as a bar plot. To show the scale used for the histogram we have shown the highest value as reference in the figure. (b) The five most important features used to predict the QPG along with their percentage importance are shown as a bar plot. See Section II of the ESI† for a detailed description of the features used in this study. The training and test set data points in (a) are colored based on the DFT computed bandgaps, E_g^{DFT}, which is the most important feature in the prediction. The E_g^{DFT} is denoted by the color bar.

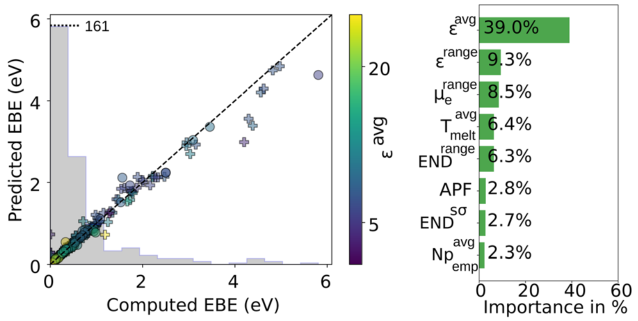
RMSE of only 0.36 eV. Fig. 2(a) compares the RF model predicted and GW computed QP gaps where the training set is shown with the '+' symbols and the test set *via* the circle symbols. The model obtained from the MLP also led to a very similar R^2 score as the one obtained from the RF method. Table S1 in the ESI† compares the performance of the models obtained from all four methods.

It is noteworthy that most of the previous studies of ML-based bandgap predictions have been limited to a particular material class such as studies of MXenes by Rajan *et al.*[29] or perovskites by Pilania *et al.*[30] In contrast, in this work, we have a dataset that includes materials without any such restrictions on chemical compositions or materials classes. Despite working with a more diverse set of materials, our RF model is capable of similar/better accuracy as the earlier studies.[29,30] In practical applications, we seldom have perfectly crystalline materials and we often see intrinsic defects (vacancies, Frenkel defects, Schottky defects, *etc.*) as well as extrinsic defects (substitutions, interstitials, inclusions, *etc.*) and often polycrystalline materials with planar defects like grain boundaries. These factors are known to modulate the bandgaps of materials 10–15% from the pure single-crystal materials that are typically studied through simulations. Thus, from a practical application point of view, an error in the bandgap of 0.3 eV for a material that has a bandgap of 2 eV or larger is accurate enough to develop and test these materials in the laboratory. However, for materials with a much smaller gap (closer to 0.5 eV) our method as well as experiments more advanced and rigorous studies are required to achieve desired bandgaps. Hence such materials need to be carefully evaluated computationally prior to experimental investigation.

Fig. 2(b) shows the most essential features for the QP gap predictions. The number of features used and their importance in the random forest algorithm is computed by calculating the Gini importance (see Methods Section 4.2 for more details). Fig. 2(b) shows that the DFT computed bandgap is the most important feature in the QP gap prediction with 92% importance. This is expected as the Kohn–Sham eigenvectors form the initial ansatz for the QP eigenvectors. DFT bandgaps are routinely used as a starting point in the determination of GW computed QP gaps. Furthermore, it is extremely promising that the macroscopic average dielectric constant (ε^{avg}) emerges as the second most important feature. While it is well known that the full frequency-dependent dielectric matrix in the plane wave basis $\varepsilon_q(G,G',\omega)$ is an important ingredient in the GW calculations, its determination is expensive and unsuitable for high-throughput computation. Here, by showing that the QPG gap can be predicted with a reasonable accuracy by using only the easily calculable static macroscopic dielectric constant $\varepsilon^{avg} = 1/\lim_{q\to 0}(\varepsilon_q(G = 0,G' = 0,\omega = 0)^{-1})$, our model emerges as an extremely useful tool in future high-throughput material discovery studies. Three of the 5 features shown in Fig. 2(b) have very low total importance, less than 1%. We examine the relevance of these features by comparing the RMSE values of RF models that include 1 to 9 of the most important features. An RF model obtained by including just the most important feature, *i.e.* the DFT gap, gives quite a large RMSE of 0.72 eV. Fig. S1 in the ESI† shows that the RMSE values for the QP gap prediction decrease from 0.44 eV to 0.36 eV when the number of features included in the model increases from 2 to 5, and thereafter it remains almost constant for up to 9 features. Thus, these three other features are also important for accurate QP gap predictions, despite their low, <1%, contribution to the total feature importance.

2.2 EBE prediction

Fig. 3(a) shows the distribution of the EBE of the materials in our dataset as a bar plot (shaded region in the background). We



(a) (b)

Fig. 3 (a) RF model predicted EBE plotted against GW-BSE computed EBE in the training (+ symbols) and test set (circle symbols). The distribution of EBEs corresponding to the entire dataset is also shown in the background as a bar plot. To show the scale used for the histogram we have shown the highest value as reference in the figure. (b) The eight most important features according to their Gini importance and their % importances are shown as a bar plot. See Section II of the ESI† for a detailed description of all the features used in this study. The training and test set data points in (a) are colored based on the average dielectric constant, ε^{avg}, which is the most important feature in the EBE prediction. The ε^{avg} value is denoted by the color bar.

note that the dataset consists of materials with EBE in a very wide range, up to 6 eV. However, the majority of the materials (over 80%) have EBE <1 eV. Fig. 3(a) also shows the RF model predicted EBE plotted against computed EBE for all the materials in our dataset. The RF model predicts the EBE of materials with an R^2 score of 0.86 and an RMSE of 0.29 eV (using 10-fold cross-validation). Table S1 in the ESI† compares the performance of the RF model with MLP, SVM, and KRR models which are all found to perform much worse than the RF model.

Fig. 3(b) shows the most important features and their % importance as computed using the Gini importance method. The most important features include properties like dielectric constants, effective masses of electrons and holes, and atomic packing fractions that are also considered in well-known physical theories of EBE. For instance, the average dielectric constant and the hole-effective mass features are also included in the Wannier–Mott (WM) model. In the WM model, the $\mathrm{EBE_{WM}} = \frac{\mu e^4}{2\hbar^2\varepsilon^2}$, where $\mu = \frac{\mu_e\mu_h}{\mu_e + \mu_h}$ is the reduced effective mass of an electron and a hole, e is the charge of an electron, $\hbar$ is the reduced Planck's constant and ε is the dielectric constant of the material. $\mathrm{EBE_{WM}}$, can thus be obtained from ground state properties without explicit BSE simulations.

In the case of Wannier–Mott (WM) excitons, the Coulomb attraction between e–h pairs is screened to a larger extent resulting in an exciton wavefunction that is spread over multiple unit cells and has low EBE. Since a majority of the materials in our dataset have a low (<1 eV) EBE, the Wannier–Mott model applies to them.

In Fig. 4 we compare the RMSE accuracy of both the WM model and the ML model as a function of the EBE. This RMSE as a function of EBE has been calculated by considering only materials with EBE in a 1 eV window around a certain EBE value. Our results show that the ML model has a much lower RMSE than the WM model. Furthermore, while consistently poorer than the ML model, the WM works comparatively well at low EBE but fails dramatically in the high EBE region. This is not surprising, since the materials that have very high EBE in the range of 4–5 eV are expected to exhibit Frenkel or Charge Transfer (CT) type excitons. CT excitons are more localized with very strong Coulomb attraction between e–h pairs and therefore have high EBE. The EBE of a CT exciton is given by the expression, $\mathrm{EBE_{CT}} = \frac{e^2}{4\pi\varepsilon r_{\mathrm{CT}}}$ where r_{CT} is the separation between the electron and hole of an exciton or radius of exciton wavefunction. Unlike the WM model, this model can not be used to predict the EBE of solid-state materials using ground-state DFT computed properties, as r_{CT} can not be computed without solving the BSE. CT excitons are usually localized in a length scale of the order of the size of a unit cell of materials and are also expected to have smaller r_{CT} for materials with tighter packing efficiency. Thus one can assume that $r_{\mathrm{CT}} = fV^{-1/3}$, where f is a dimensionless proportionality constant and V is the volume of the unit cell, allowing the estimation $\mathrm{EBE_{CT}}$ from ground state properties without the need of BSE simulations.

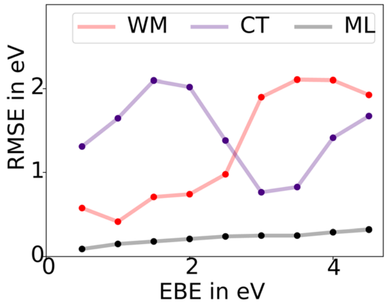


Fig. 4 The RMSE accuracy of both Wannier–Mott (WM) and charge transfer/Frenkel (CT) models as well the ML model as a function of the EBE. The RMSE as a function of EBE has been calculated by considering only materials with EBE in a 1 eV window around a certain EBE value. Our results show ML model performs better in any energy window than the WM or CT model. WM model works fairly well in the case of low EBE but fails in the high EBE region. The CT model is less accurate than the WM model in the low EBE region but works better in the high EBE region.

In Fig. 4 we present the RMSE accuracy of EBE obtained from the CT model ($f = 0.5$) as a function of the EBE. In comparison to the ML model, the CT model is consistently poorer with high RMSE values. However, as expected, it performs better than the WM model in the high EBE region.

Overall, the ML model performs much better in any energy window in comparison to the WM or CT model. We think this superior predicting capability comes from the inclusion of additional material properties not present in the WM model such as packing fraction and range of electron and hole effective masses and dielectric constants. By including such attributes our ML model is capturing the physics of not only the low EBE excitons but also the higher EBE regime where the CT model is more perhaps applicable than the WM model. Therefore, one can in principle build a more general empirical model for excitons based on the properties revealed by our ML model.

2.3 ML classification of exciton binding energies and absorption-related properties

In the previous two sections, we discussed the capabilities of our ML model in predicting accurate QPG and EBE of a wide variety of materials by solving a regression problem. This undoubtedly has a lot of potential applications in identifying materials with specific bandgap requirements such as for ultra-wide bandgap semiconductors for power electronics applications or optical devices in visible or UV light applications. However, for tractable BSE calculations with reasonable k-grids, the computed EBEs have an error of ∼100 meV for low EBE values, which is subsequently extended in our ML model. At the same time, the EBE values calculated from the BSE method present a strong correlation with experimental EBEs.[31,32] Thus, from a material designing/discovery point of view, an equally useful exercise would be to identify materials with desirable optical/excitonic properties such as having a low EBE, which is a classification problem. In this section, we apply classification algorithms to classify materials based on their excitonic properties.

Among excitonic properties, low EBEs are preferred in applications where free e–h pairs are desired for example in photocatalytic materials.[2] In addition, two other parameters derived from BSE obtained absorption spectra—the integrated absorption coefficient (IAC) in the solar wavelength range of interest and anisotropy in absorption coefficient (AAC)—are useful to quantify the potential of a material for solar-energy absorption, for example in photovoltaics and photocatalysts. The methods section describes the calculation of IAC and AAC from the frequency-dependent absorption spectra. In a previous study, we have established that low EBE materials are those that have EBE smaller than 0.2 eV, high (visible/UV)-light IAC materials have an IAC larger than $>10.5 \times 10^4$ cm^{-1} eV, and high AAC materials have AAC ≥ 0.8.

Fig. 5 shows the results of classification obtained by the RF method in the form of confusion matrixes. Fig. 5(a) shows the confusion matrix for classifying materials as low EBE. Fig. 5(b) shows the matrix for classifying materials as high UV-light IAC and Fig. 5(c) for classifying materials as high AAC materials. For the EBE classification, only 21 of the 305 materials were classified incorrectly resulting in a high classification accuracy of 93%. The classification models for the IAC and AAC resulted in an accuracy of 94% and 83% for the 355 materials in the dataset, respectively. Additionally, we employed the AdaBoost, stochastic gradient descent (SGD), and MLP[27,28] methods for the classification. However, the RF performed best among the four methods. ESI Table S2† shows the comparison between the four methods.

The feature set used for IAC and AAC classification was selected following a similar strategy employed for the QP gap and EBE prediction described earlier. In the ESI Fig. S2 (IAC) and S3 (AAC)† we have shown the most important features along with their % importance. In the case of IAC prediction mean dielectric constant (67.1%) and DFT computed bandgap (4.6%) emerge as the two most important features. The emergence of these two properties as the most important features can be understood from the fact that for a high absorption in the visible spectrum, 1.7–3.5 eV, one needs to have a material with a QP gap in the same range and also needs to have significant absorption coefficient ($\propto \varepsilon(\omega)$) in that energy range. As we have seen in Fig. 2 for the QP gap prediction mean dielectric constant and DFT computed bandgap are the two most important features, it is not surprising that they are also equally important for the IAC predictions. Moreover, the high importance of the mean dielectric constant also signifies a high degree of correlation between the static dielectric constant of a material and frequency-dependent dielectric function. In the prediction of anisotropy in visible light absorption (AAC) we find that the range of dielectric constant ($\max\{\varepsilon_x, \varepsilon_y, \varepsilon_z\} - \min\{\varepsilon_x, \varepsilon_y, \varepsilon_z\}$) is the most important feature (58.8%). All the other important features in the prediction of AAC have importance <5%. Therefore, one can identify a material with a high degree of anisotropy in visible light absorption by looking at the anisotropy in the static dielectric constant, which once again highlights the importance of static dielectric constants in the excitonic properties. The Materials Project (MP)[9] database currently holds ∼150 000 materials, but only ∼7000 (4.7%) of them have computed static dielectric constants. We believe that the static dielectric constant using DFT is quite inexpensive to calculate but is a crucial parameter to understand material applicability for a wide variety of electronic and optoelectronic applications and therefore it would be useful to compute and curate it for more materials in existing materials databases.

2.4 Screening materials from MP database

The high accuracy (90%) of our ML classifiers enables their application to a larger material set to screen for materials suitable for photoabsorption-based applications such as photovoltaics, and photocatalysis.

We perform such screening on ∼7083 materials that have static dielectric constants and DFT computed band structure available in the MP database. We find that out of 7000 materials, only 159 passed the criterion of low EBE, high IAC, and high AAC in the visible-light region, 1.7–3.5 eV. Regarding the UV-light absorption, 3.5–4.2 eV, we found 237 materials passed the aforementioned criteria.

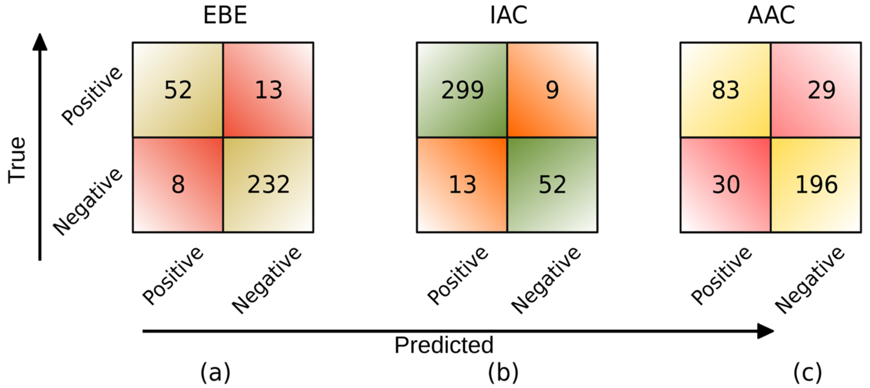


Fig. 5 Confusion matrix illustrating the number of true positives, (TP), true negatives (TN), false negatives (FN), and false negatives (FN), for predicting if materials possess a low exciton binding energy (EBE), <0.2 eV, high integrated absorption coefficient (IAC), $>10.5 \times 10^4$ cm^{-1} eV, and anisotropy in absorption coefficient, >0.8. All the incorrect classifications (TPs and TNs) are shown in different shades of red and correct classifications (FPs and FNs) are shown in shades of green and yellow.

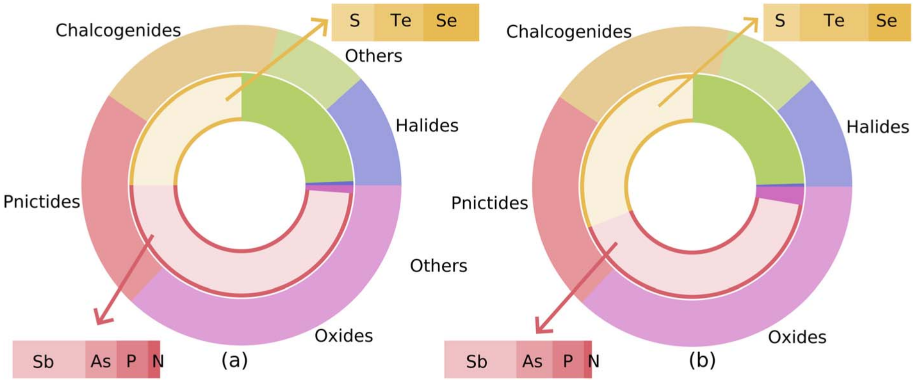


Fig. 6 Materials obtained from the MP database and selected for visible (a) and UV light (b) applications through the ML-based screening process are categorized into different chemical compositions. The outer rings represent the fraction of the starting set of ~7000 materials belonging to different chemical groups whereas the inner rings represent the same for selected (159 for visible and 203 for UV) materials. In the case of visible light absorption, we see significantly fewer oxides pass through the screening process, whereas, pnictides and chalcogenides comprise a major fraction of the selected materials.

Fig. 6 examines the chemical compositions of the materials in the starting set of ~7000 materials (outer rings) as well as ones that emerged as suitable for visible-light applications, Fig. 6(a), and UV-light, Fig. 6(b), respectively, inner rings. We find that most of the visible-light materials are either pnictides or chalcogenides. Note that almost half of the starting set of materials, 47%, are oxides but only 2 oxides pass through the screening. This is not surprising since oxides mostly have larger bandgaps than the visible light region. More oxides, six, are found in the screened materials for the UV light region. Most of the materials for UV absorption also belong to pnictides or chalcogenides. Furthermore, we find that almost half of the selected chalcogenides are tellurides. In the ESI Tables S6 and S7† we list all the screened materials for both visible and UV absorption respectively along with their ML-predicted QP gap and EBE values. In Tables S6 and S7,† we have also reported whether a screened material has been already synthesized before and has an ICSD ID and the computed value of energy above hull. We find that the majority of these materials 193 out of 234 materials have been already synthesized. Moreover, we find that 168 out of 193 previously synthesized materials have computed energy above the hull value of 0 eV. An examination of the screened materials shows that several of these materials have been studied in the context of photoabsorption-related applications, for example, GeTe,[33] AlSb,[34] SnSe,[35] *etc.* Thus it is likely that the other screened materials can be promising novel materials for photoabsorption-related applications.

3 Conclusion

In this study, we report the largest dataset of excited state properties of bulk materials calculated using the state-of-the-art GW-BSE formalism. This database has been created using the open-source Python workflow package *py*GWBSE and made publicly available through the website **https://hydrogen.cmd.lab.asu.edu/gwbse-data**. Using this dataset we have developed ML models that can predict the QP gap and EBE of a diverse set of materials with excellent accuracy. Despite having a limited dataset size of ~350 materials, the ML models predict the QP gap with an accuracy of 0.36 eV (RMSE) and EBE with an accuracy of 0.29 eV. Moreover, we demonstrated that using the ML models developed on this dataset we can screen materials based on excited state properties such as EBE, IAC, and AAC by using only ground state DFT computed properties. Using our ML models we utilize the relevant existing DFT computed date in the MP database, available for 7000 materials, and screen 159 materials for visible and 203 materials for UV light-based applications. This work presents a robust framework to utilize the underlying potential of the large datasets of ground-state properties available in open-source materials databases, allowing us to obtain excited state properties without the need for explicit simulations.

4 Computational methods

4.1 GW-BSE calculations

Quasi-particle (QP) energies, the energy required to add or remove an electron from an interacting many-electron system is not a ground state property of the system and therefore can't be computed accurately using DFT. These QP energies can be computed correctly at a reasonable computational cost using many-body perturbation theory within GW approximation. Within GW approximation one computes the self-energy as the product of one-particle Green's function (G) and the screened Coulomb interaction (W).[36,37] The QP energies are evaluated as a correction to the KS eigenvalues using first-order perturbation by assuming the difference between exchange–correlation potential (V_{xc}) and self-energy is sufficiently small. QP energies, and thus QPGs, of all the materials in this study were computed using the one-shot G_0W_0 method *via* the high-throughput workflow code *py*GWBSE.[23] *py*GWBSE workflows allow high-throughput first principles atomistic simulations *via* the VASP[38] package by high-throughput automated input file generation, submission to supercomputing platforms, analysis of post-simulation data, and storage of metadata and data in a MongoDB database. We use the PBE54 psuedopotentials as employed in VASP to determine the

Kohn–Sham eigenvalues and eigenvectors. A plane wave cutoff of 500 eV and a *k*-grid with reciprocal density 200 $Å^{-3}$ was used for all the GW-BSE calculations. A fixed value of 100 eV for the screened Coulomb energy cutoff and 80 for the number of frequency grid points was used for all the GW calculations since these parameters don't display a significant dependence on the material of choice.[23] The number of unoccupied bands for the GW simulations was selected such that the QPGs converged within <0.1 eV. The macroscopic dielectric constants reported here were calculated using density functional perturbation theory following the formalism originally introduced by Baroni *et al.*[39] and later implemented for projector-augmented wave methodology by Gajdoš *et al.*[40] Once the QP energies were obtained, the absorption spectra, $\varepsilon(\omega)$, and EBE were computed by solving the Bethe–Salpeter equation (BSE). BSE is a two-particle equation that explicitly includes the electron–hole interactions or the excitonic effects within the Tamn–Dancoff approximation.[41,42] The number of valence (v) and conduction (c) bands included in the BSE calculation are selected such that the vertical v → c transitions of energy less than 3 eV were obtained. From a BSE solution, the optical gap is the lowest energy excitation. The EBEs are the difference between the optical gaps and the direct QP gaps.

To quantify the fraction of incident light that can be absorbed by a material in a desired frequency range, we can compute the integrated absorption coefficient, IAC. IAC is obtained by integrating the BSE computed frequency-dependent absorption coefficient, α_{int}.[23] In the case of light polarization along x axis,

$$\alpha_{\text{int}}{}^{x} = \int_{\omega_{\min}}^{\omega_{\max}} \alpha^{x}(\omega)\mathrm{d}\omega$$

$$\text{where } \alpha^{x}(\omega) = \frac{2\pi\left(\left|\sqrt{\left(\varepsilon_1{}^{x}(\omega)\right)^2 + \left(\varepsilon_2{}^{x}(\omega)\right)^2}\right| - \left(\varepsilon_1{}^{x}(\omega)\right)^{1/2}\right)}{\lambda} \qquad (1)$$

where $\varepsilon_1(\omega)$ is the real and $\varepsilon_2(\omega)$ the imaginary part of the dielectric function. ω and λ are the frequency and wavelength of incident radiation. The $\varepsilon_2(\omega)$ is obtained by solving BSE and $\varepsilon_1(\omega)$ is computed using the Kramers–Kronig relation.

To assess whether a material has a preference for absorbing light of certain polarization we calculate the anisotropy in absorption coefficient (AAC), $\alpha_{\text{int}}^{\text{aniso}}$. $\alpha_{\text{int}}^{\text{aniso}}$ is defined as the ratio of min($\alpha_{\text{int}}{}^{x}$,$\alpha_{\text{int}}{}^{y}$,$\alpha_{\text{int}}{}^{z}$) and max($\alpha_{\text{int}}{}^{x}$,$\alpha_{\text{int}}{}^{y}$,$\alpha_{\text{int}}{}^{z}$).

4.2 Machine learning

All the ML models were created using the Scikit-Learn ML library.[43] Over 150 features were used for training the ML models for this work. Section II in the ESI† describes the features considered in this work. The features include DFT-computed material properties that are available in the MP database including the DFT bandgaps and the dielectric constants. In order to include the effective masses as features, we computed the effective masses from the bandstructures available in the MP database by employing the sumo[44] code and the MP Application Programming Interface (API).

The most important features were determined by computing the Gini Importance[45] method. The Gini Importance of each feature is calculated as the decrease in node impurity weighted by the probability of reaching that node. Furthermore, for the prediction of QPG and EBE, we chose the minimum number of features that were needed to obtain an RMSE value converged within 0.01 eV.

Abbreviations

QP	Quasiparticle
EBE	Exciton binding energy
DFT	Density functional theory

Data availability

The GW-BSE calculations performed to create the dataset were performed using publicly available *py*GWBSE code (**https://github.com/cmdlab/pyGWBSE**). The GW-BSE dataset will be made available through the website (**https://hydrogen.cmd.lab.asu.edu/gwbse-data**) upon publication. The codes used to develop ML models and apply them to the Materials Project dataset have been made publicly available (**https://github.com/cmdlab/ML-GWBSE**). The materials shortlisted in this study for visible and UV lightbased applications along with their ML-predicted properties has been reported in the ESI.†

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgements

This research was supported by the U.S. Department of Energy (DOE), Office of Science, Basic Energy Sciences (BES), under Award Number DE-SC0024184 (physics-guided and machine learning enabled discovery of solar materials), ULTRA, an Energy Frontier Research Center funded by the U.S. Department of Energy (DOE), Office of Science, Basic Energy Sciences (BES), under Award # DE-SC0021230 (GW-BSE high-throughput simulations for accurate prediction of properties of ultra-wide band gap materials), Arizona State University start-up funds, and by the National Science Foundation (NSF) under Award Number 2235447 (*ab initio* simulations of semiconducting materials for electronic applications). The authors acknowledge the San Diego Supercomputer Center under the NSF-XSEDE and NSF-ACCESS Award No. DMR150006, and the Research Computing at Arizona State University for providing HPC resources. This research used resources from the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

1 N. J. Borys, C. Argyropoulos and L. Ye, Light and matter interactions: Recent advances in materials, theory,

fabrication, and characterization, *APL Mater.*, 2022, **10**(6), 060401.

2 T. Biswas and A. K. Singh, Excitonic effects in absorption spectra of carbon dioxide reduction photocatalysts, *npj Comput. Mater.*, 2021, **7**(1), 1–10.

3 M. R. Hasan, E. S. Arinze, A. K. Singh, *et al.*, An antimony selenide molecular ink for flexible broadband photodetectors, *Adv. Electron. Mater.*, 2016, **2**(9), 1600182.

4 J. J. De Pablo, B. Jones, C. L. Kovacs, V. Ozolins and A. P. Ramirez, The materials genome initiative, the interplay of experiment, theory and computation, *Curr. Opin. Solid State Mater. Sci.*, 2014, **18**(2), 99–117.

5 Materials Genome Initiatve, https://www.mgi.gov/, accessed: 2023-07-13.

6 A. Shinde, S. K. Suram, Q. Yan, *et al.*, Discovery of manganese-based solar fuel photoanodes via integration of electronic structure calculations, pourbaix stability modeling, and high-throughput experiments, *ACS Energy Lett.*, 2017, **2**(10), 2307–2312.

7 L. Zhou, A. Shinde, J. H. Montoya, *et al.*, Rutile alloys in the Mn–Sb–O system stabilize Mn3+ to enable oxygen evolution in strong acid, *ACS Catal.*, 2018, **8**(12), 10938–10948.

8 A. K. Singh, J. H. Montoya, J. M. Gregoire and K. A. Persson, Robust and Synthesizable Photocatalysts for CO2 Reduction: A Data-Driven Materials Discovery, *Nat. Commun.*, 2019, **10**(1), 443.

9 A. Jain, S. P. Ong, G. Hautier, *et al.*, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.*, 2013, **1**(1), 011002.

10 S. Curtarolo, W. Setyawan, S. Wang, *et al.*, AFLOWLIB. ORG: A distributed materials properties repository from high-throughput ab initio calculations, *Comput. Mater. Sci.*, 2012, **58**, 227–235.

11 S. Haastrup, M. Strange, M. Pandey, *et al.*, The Computational 2D Materials Database: high-throughput modeling and discovery of atomically thin crystals, *2D Materials*, 2018, **5**(4), 042002.

12 C. Ortiz, O. Eriksson and M. Klintenberg, Data mining and accelerated electronic structure theory as a tool in the search for new functional materials, *Comput. Mater. Sci.*, 2009, **44**(4), 1042–1049.

13 The NoMaD Repository, http://nomad-repository.eu/cms/, accessed: 2023-02-22.

14 S. Kirklin, J. E. Saal, B. Meredig, *et al.*, The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies, *npj Comput. Mater.*, 2015, **1**(1), 1–15.

15 Z. Yao, Y. Lum, A. Johnston, *et al.*, Machine learning for a sustainable energy future, *Nat. Rev. Mater.*, 2023, **8**(3), 202–215.

16 J. E. Saal, A. O. Oliynyk and B. Meredig, Machine learning in materials discovery: Confirmed predictions and their underlying approaches, *Annu. Rev. Mater. Res.*, 2020, **50**, 49–69.

17 Y. Liu, T. Zhao, W. Ju and S. Shi, Materials discovery and design using machine learning, *J Materiomics.*, 2017, **3**(3), 159–177.

18 E. O. Pyzer-Knapp, J. W. Pitera, P. W. Staar, *et al.*, Accelerating materials discovery using artificial intelligence, high performance computing and robotics, *npj Comput. Mater.*, 2022, **8**(1), 84.

19 A. K. Singh, R. Gorelik and T. Biswas, Data-Driven Discovery of Robust Materials for Photocatalytic Energy Conversion, *Annu. Rev. Condens. Matter Phys.*, 2023, **14**(1), 237–259.

20 S. B. Torrisi, A. K. Singh, J. H. Montoya, T. Biswas and K. A. Persson, Two-dimensional forms of robust CO2 reduction photocatalysts, *npj 2D Mater. Appl.*, 2020, **4**(1), 24.

21 Z. Rao, P. Y. Tung, R. Xie, *et al.*, Machine learning–enabled high-entropy alloy discovery, *Science*, 2022, **378**(6615), 78–85.

22 A. Rasmussen, T. Deilmann and K. S. Thygesen, Towards fully automated GW band structure calculations: What we can learn from 60.000 self-energy evaluations, *npj Comput. Mater.*, 2021, **7**(1), 22.

23 T. Biswas and A. K. Singh, py GWBSE: a high throughput workflow package for GW-BSE calculations, *npj Comput. Mater.*, 2023, **9**(1), 22.

24 X. Gonze, B. Amadon, G. Antonius, *et al.*, The ABINIT project: Impact, environment and recent developments, *Comput. Phys. Commun.*, 2020, **248**, 107042.

25 J. Deslippe, G. Samsonidze, D. A. Strubbe, M. Jain, M. L. Cohen and S. G. Louie, BerkeleyGW: A massively parallel computer package for the calculation of the quasiparticle and optical properties of materials and nanostructures, *Comput. Phys. Commun.*, 2012, **183**(6), 1269–1289.

26 M. Bonacci, J. Qiao, N. Spallanzani, *et al.*, Towards high-throughput many-body perturbation theory: efficient algorithms and automated workflows, *npj Comput. Mater.*, 2023, **9**(1), 74.

27 G. Bonaccorso, *Machine Learning Algorithms*, Packt Publishing Ltd, 2017.

28 M. Mohammed, M. B. Khan and E. B. M. Bashier, *Machine Learning: Algorithms and Applications*, Crc Press, 2016.

29 A. C. Rajan, A. Mishra, S. Satsangi, *et al.*, Machine-learning-assisted accurate band gap predictions of functionalized mxene, *Chem. Mater.*, 2018, **30**, 4031–4038, DOI: 10.1021/acs.chemmater.8b00686.

30 G. Pilania, A. Mannodi-Kanakkithodi, B. Uberuaga, R. Ramprasad, J. Gubernatis and T. Lookman, Machine learning bandgaps of double perovskites, *Sci. Rep.*, 2016, **6**(1), 19375.

31 J. Sun, J. Yang and C. A. Ullrich, Low-cost alternatives to the Bethe-Salpeter equation: Towards simple hybrid functionals for excitonic effects in solids, *Phys. Rev. Res.*, 2020, **2**(1), 013091.

32 A. Molina-Sánchez, D. Sangalli, K. Hummer, A. Marini and L. Wirtz, Effect of spin-orbit interaction on the optical spectra of single-layer, double-layer, and bulk MoS 2, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2013, **88**(4), 045412.

33 M. Noman, M. Abden and M. Islam, *Germanium Telluride Absorber Layer, a Proposal for Low Illumination Photovoltaic Application Using AMPS 1D*, IEEE, 2018, pp. 1–5.

34 R. Dhakal, Y. Huh, D. Galipeau and X. Yan, AlSb compound semiconductor as absorber layer in thin film solar cells, *Solar Cells—New Aspects and Solutions*, 2011.

35 V. R. Minnam Reddy, S. Gedi, B. Pejjai and C. Park, Perspectives on SnSe-based thin film solar cells: a comprehensive review, *J. Mater. Sci.: Mater. Electron.*, 2016, **27**, 5491–5508.

36 M. S. Hybertsen and S. G. Louie, Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 1986, **34**(8), 5390.

37 G. Onida, L. Reining and A. Rubio, Electronic excitations: density-functional versus many-body Green's-function approaches, *Rev. Mod. Phys.*, 2002, **74**(2), 601.

38 M. Shishkin and G. Kresse, Implementation and performance of the frequency-dependent G W method within the PAW framework, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2006, **74**(3), 035101.

39 S. Baroni and R. Resta, Ab initio calculation of the macroscopic dielectric constant in silicon, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 1986, **33**(10), 7017.

40 M. Gajdoš, K. Hummer, G. Kresse, J. Furthmüller and F. Bechstedt, Linear optical properties in the projector-augmented wave methodology, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2006, **73**(4), 045112.

41 M. Rohlfing and S. G. Louie, Electron-hole excitations and optical spectra from first principles, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2000, **62**(8), 4927.

42 T. Sander, E. Maggio and G. Kresse, Beyond the Tamm-Dancoff approximation for extended systems using exact diagonalization, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2015, **92**(4), 045209.

43 O. Kramer and O. Kramer, Scikit-learn, *Machine Learning for Evolution Strategies*, 2016, pp. 45–53.

44 A. Ganose, A. Jackson and D. Scanlon, sumo: Command-line tools for plotting and analysis of periodic *ab initio* calculations, *J. Open Source Softw.*, 2018, **3**(28), 717.

45 L. Breiman, *Classification and Regression Trees*, Routledge, 2017.