Predicting exciton binding energies from ground-state properties

Malte Grunert,[*,†] Max Großmann,[*] and Erich Runge
Institute of Physics and Institute of Micro- and Nanotechnologies, Technische Universität Ilmenau, 98693 Ilmenau, Germany

(Received 19 February 2024; revised 16 July 2024; accepted 17 July 2024; published 22 August 2024)

We systematically investigate the correlations between exciton binding energies and various easily calculated ground-state properties ("predictors") that have been proposed to correlate with them. We do so for 51 bulk compounds, ranging from simple binary semiconductors to more exotic materials such as noble gas solids, perovskites, and layered systems and show that using only a few easily calculable parameters, a broadly applicable clustering of materials according to their exciton binding energies is possible. While no single one of the proposed parameters satisfactorily predicts the exciton binding energy, a combination of several parameters that enter the Wannier-Mott model or characterize the nature of the valence band and the ionicity of the compound allows a reliable automated clustering of exciton binding energies. The results allow an efficient estimation of the importance of excitonic effects without the need for expensive many-body perturbation theory calculations or other advanced numerically demanding *ab initio* approaches.

DOI: 10.1103/PhysRevB.110.075204

I. INTRODUCTION

The optical properties of many semiconductors and insulators are dominated by excitons, at least at low temperatures [1]. The inclusion of the effects of these quasiparticles is critical for the accurate description of important technologies such as solar cells [2], solar water splitting [3], artificial photosynthesis [4], attosecond physics [5], quantum computing [6], and other optical and electronic applications [7–9]. For many of these applications, the binding energy of the lowest bright exciton E_b is a crucial parameter, and its prediction is an important prerequisite for material classification, selection, and technology design.

Two well-known limiting cases/models of excitons are the Frenkel and the Wannier-Mott (WM) models [1]. In the former, excitons behave similarly to molecular excitations. They have a large binding energy of about 1000 meV, are highly localized, and are typically visualized as a pair of a filled and an empty molecular orbital. In the latter, excitons are pictured as an electron-hole pair forming a huge hydrogen-atom-like bound state delocalized over many unit cells with a small binding energy on the tens of meV scale. Depending on the exciton binding energy, computationally more (Bethe-Salpeter equation, BSE) or less (time-dependent density functional theory [10,11]) demanding methods are required to obtain accurate optical properties such as absorption spectra [12]. Therefore, the key to efficient calculations is to know the approximate exciton binding energy in advance. However, the accurate calculation of the excitonic binding energy is at least as expensive as the calculation of the optical properties themselves, if not more. We show that the results of cheap ground-state calculations already allow a broadly applicable clustering of materials according to their exciton binding energies.

The best-known approach to relate the exciton binding energy to ground-state properties is certainly the aforementioned Wannier-Mott (WM) model, which links the exciton binding energy E_b to the effective electron and hole masses m_e and m_h, respectively, and the static dielectric constant ε_r,

$$E_b^{\mathrm{WM}} = \frac{\mu_X}{\hbar^2}\left(\frac{e^2}{4\pi\varepsilon_0\varepsilon_r}\right)^2 \tag{1}$$

with ${\mu_X}^{-1} = {m_e}^{-1} + {m_h}^{-1}$. While it generally works well for the weakly bound excitons for which it was originally formulated, the definitions of both the effective masses and the dielectric function show certain ambiguities. The effective mass approximation holds only when the exciton wavefunction in reciprocal space is localized in a narrow **k**-point range and it needs significant modification in the presence of degenerate, nonparabolic, or anisotropic bands [13,14]. On the other hand, the static dielectric constant ε_r is made up of two components, the electronic ε_r^e and the ionic ε_r^i. Weakly bound excitons extended over various unit cells can be screened by the lattice, while strongly bound localized excitons cannot. While there are more sophisticated expressions for the resulting radius-dependent interpolation of the dielectric constant between ε_r^e and $\varepsilon_r^e + \varepsilon_r^i$ (see, e.g., Refs. [15,16]), commonly $\varepsilon_r^e + \varepsilon_r^i$ is used for weakly bound excitons (such as narrow-gap semiconductors), ε_r^e is used for strongly bound excitons, and an *ad hoc* intermediate value is used for intermediately strongly bound excitons (such as AlN) [1,17]. Furthermore,

[*]These authors contributed equally to this work.
[†]Contact author: malte.grunert@tu-ilmenau.de

the WM model breaks down for strongly bound excitons or more complex cases such as charge-transfer excitons or layered systems [18].

In recent years, there has been some research to expand or provide alternatives to the WM model. For example, Dvorak *et al.* [19] investigate weakly bound excitons in semiconductors and propose the colocalization of valence and conduction band states as a relevant additional parameter characterizing excitons. Jiang *et al.* [20] study 2D systems, where they find the universal relation $E_b = E_g/4$, which, however, only applies to 2D systems. The authors state explicitly that such simple relations do not exist for 3D systems [20].

In search of a more general way to predict exciton binding energies, or at least to categorize them as strongly or weakly bound, we investigate in the present paper the correlations between various ground-state properties and experimental binding energies for the lowest bright exciton for 51 different compounds across various classes of material systems, i.e., simple binary semiconductors and isolators, noble gas crystals, and more complex compounds such as α-Al_2O_3, As_2Se_3, $PbCl_2$, and inorganic and organic lead-halide perovskites.

II. CURATION OF THE EXPERIMENTAL DATASET

It is no easy feat to obtain accurate experimental exciton binding energies. This is because of two factors: First, for most materials there are very few, if any, published experiments, many of which are not performed with state-of-the-art methods. Second, for those materials for which various measurements are available, it is clear that a variety of factors can influence the obtained value of E_b. This is perhaps best illustrated by the case of gallium nitride, which was studied extensively in the 1990s. Here, strain, temperature, the growth method, and the growth substrate all influence the measured exciton binding energy (Fig. 5 in Ref. [21] summarizes some of the contemporary literature). Furthermore, the underlying model used in the analysis of the measurements can also play a significant role, see, e.g., Ref. [22] for the impact of different broadening expressions in the Elliot formula on the analysis of absorption spectra or the assumed form of the decay expression for temperature-dependent photoluminescence in the case of $MAPbI_3$. Finally, before the advent of modern experimental and theoretical tools the assignment of features in the optical properties to individual excitonic lines was far from unambiguous, see, e.g., the discussion in Ref. [23] for the case of alkaline earth sulfides. Fortunately, the reported values for a single material usually differ by only a few percent, while the range of reported E_b for all materials spans several orders of magnitude, from less than 1 meV to more than 1 eV. The largest range of exciton binding energies are found for the metal-halide perovskites $MAPbCl_3$ and $MAPbI_3$, for which values around 10–70 meV have been reported in the literature (see, e.g., Refs. [22,24]). For the current paper, this essentially means that by choosing an intermediate value of E_b for those compounds for which multiple values are reported, the results should not change qualitatively. On the other hand, no accurate quantitative predictions should be expected in those cases. In addition, we will later on divide the various materials into three categories (those with weakly, intermediately, and strongly bound excitons), which makes the results much more robust. The collected dataset is shown in Tables I–VI.

A few brief comments on the excitons contained in the dataset are probably useful and helpful: For almost all materials, the exciton energies used correspond to the energetically lowest dipole-allowed intrinsic exciton. For direct semiconductors this corresponds to a direct exciton, for indirect semiconductors to an indirect exciton. Exceptions to this are the IIa-VIb alkaline earth chalcogenides other than MgO and CaO, for many of which only direct excitons have been measured [25], although the assignment of excitonic peaks there was particularly difficult as mentioned above. According to the analysis of Lipari and Baldereschi on the energies of indirect excitons in realistic semiconductors [26], the difference between direct and indirect excitons lies mainly in slight differences in the correction terms to the simple WM model, which are not used in this paper. Furthermore, materials containing transition metals, lanthanides and actinides with partially filled d and f orbitals, such as the technologically important ZnO, CdTe, CuCl, MoS_2, and others, are intentionally ignored in this paper, since in these cases correlation effects can drastically complicate the electronic structure and exciton physics. However, we note that compounds with partially filled d or f bands can be easily identified and quantified as predictors by ground-state calculations. Thus, a generalization of the present paper would seem worthwhile if sufficient experimental data on excitons in materials with strong correlations were available. We are aware of about 20 high-quality measurements of the exciton binding energy in such materials, which is too small a sample to represent the rich and diverse physics of correlated electron systems.

There are a large number of reported *ab initio* exciton binding energies, usually obtained by solving the BSE on the results of quasiparticle calculations [27]. However, we have chosen not to include any of them, as they are highly dependent on the level of theory and the approximations chosen, see for example Ref. [28] for a brief discussion on the impact of the static screening approximation on the BSE and Refs. [17,29] for a discussion on the inclusion of the lattice polarizability/electron-phonon interaction. We note that especially the latter is a rapidly developing field, see, e.g., Refs. [30,31] and the references contained within. Other relevant factors include the pseudopotential, the exchange-correlation functional, the lattice constant used, and many others. Furthermore, it is often not clear whether the reported exciton binding energy is fully converged, as it can converge extremely slowly with respect to the **k**-point grid density [28,32].

III. SELECTION OF PREDICTIVE GROUND-STATE PROPERTIES

We select various ground-state properties as possible predictors, taking into account both their computational complexity and their suggested correlation to the exciton binding energy.

The first group of properties comprises quantities related to the WM model [Eq. (1)], i.e., the electron and hole effective masses at the direct band gap m_e and m_h as well as the

electronic and ionic contributions ε_r^e and ε_r^i to the static dielectric constant.

Furthermore, we explore the connection between ionicity and exciton binding energies. As is well known, there is no clear definition of "ionicity" in solids [33]. As one possible measure, we choose the ionicity measure proposed by Abu-Farsakh and Qteish [34], which links ionicity to the mean distance between the centers of the maximally localized Wannier functions (MLWFs) of the valence bands and the atom centers [35]. It has the advantage that it can be obtained essentially parameter-free from ground-state wavefunctions. However, as discussed later, a certain ambiguity remains in the choice of the MLWFs.

Finally, we consider properties related to the localization of electrons and holes in the ground state. In general, one might expect more localized valence and conduction band states to correlate with smaller, more localized exciton states, which in turn would correlate with larger exciton binding energies. Specifically, we investigate as possible predictor the irreducible spread of the valence band MLWFs, defined below in Eq. (8). It is related to the band gap and thus indirectly to the electronic dielectric constant [36]. Here we note that the inverse relation between the band gap and the dielectric constant has recently been verified on a large material dataset [37]. Dvorak *et al.* [19] have already directly investigated the correlation between the irreducible spread of the valence band MLWFs and excitonic properties. However, they only presented results for seven simple materials.

Dvorak *et al.* also suggest the "colocalization" of valence and conduction states at the anion (in binary compounds) as a marker for increased exciton binding energies. We investigate this and related (generalized) quantities for a larger set of materials. In addition, we quantify the mean dispersion of the topmost valence band, again based on the assumption that excitons composed of less dispersive single-particle states are more localized and have higher exciton binding energies: The Coulomb attraction of electron and hole is larger the closer they are in real space, which in turn implies a larger range of wave vectors in reciprocal space. However, a larger range of wave vectors implies larger single-particle energies, unless the bands involved have little dispersion and their bandwidths are small.

To examine the correlations between these properties and the exciton binding energies, we perform density functional theory (DFT) calculations to evaluate these properties on the materials included in the dataset described above. We intentionally limit ourselves to easily calculated DFT quantities in order to make the prediction of exciton binding energies simple for future high-throughput projects. This also facilitates the application and testing of our predictors for future exciton binding energy measurements.

IV. COMPUTATIONAL DETAILS

For the compounds of the experimental dataset, computationally relaxed structure files were obtained from the Materials Project [38,39] and reduced to their primitive structure in the convention of Refs. [40,41] using PYMATGEN [42,43]. Two different DFT codes were used to evaluate the large number of different quantities in an efficient manner: In all cases, the PBE approximation [44] to the exchange-correlation functional was utilized. Unless otherwise indicated, calculations were carried out using the plane-waves code QUANTUM ESPRESSO [45,46] with the SG15 [47] optimized norm-conserving Vanderbilt pseudopotentials [48]. We used Γ-centered **k**-point grids with an even number of subdivisions defined through a structure-independent reciprocal density $\rho_{\mathbf{k}}$ as defined in PYMATGEN [49]. All calculations were converged with respect to the **k**-point grid and plane-wave cutoff until a convergence threshold of 1 kcal/mol per atom in the unit cell was reached. The starting value for the plane-wave cutoff was set to the recommended value of 60 Ry [47] and the **k**-point density to $\rho_{\mathbf{k}} = 1500\,\text{Å}^3$. Some calculations additionally required the calculation of MLWFs from the QUANTUM ESPRESSO results. For this, we used the WANNIER90 code [50]. The initial guesses for the MLWFs were defined through the selected columns of the density matrix (SCDM) algorithm [51,52] to improve the stability of the successive spread minimization. The components of the static dielectric constants were determined using a plane-wave basis and the projector augmented wave method [53] as implemented in the VASP code [54,55] utilizing the method by Gajdoš *et al.* [56]. For these calculations, we used **k**-point grids with a density of $\rho_{\mathbf{k}} = 3000\,\text{Å}^3$ and a plane-wave cutoff of 520 eV for all materials, as suggested by the Materials Project [38]. All calculations were carried out using an in-house, publicly available high-throughput framework [57]. Next, we introduce the operative definitions of the predictive properties.

Average effective mass. We define an inverse effective mass at a given **k**-point of interest as the average over the $N \geqslant 1$ degenerate or nearly degenerate Kohn-Sham bands $\epsilon_n(\mathbf{k})$ and over all Cartesian directions $i \in (x, y, z)$,

$$\left\langle \frac{1}{m_{e,h}(\mathbf{k})} \right\rangle = \frac{1}{N} \sum_n^N \frac{1}{3}\, tr\left\{ \left| \frac{1}{\hbar^2} \nabla_k^2 \epsilon_n(\mathbf{k}) \right| \right\} = \frac{1}{3N} \frac{1}{\hbar^2} \sum_n^N \sum_{i=1}^3 \left| \frac{\partial^2 \epsilon_n(\mathbf{k})}{\partial k_i^2} \right|. \tag{2}$$

To find the valence band maximum (VBM) and the conduction band minimum (CBM) and to evaluate the second derivatives with high numerical accuracy, a DFT calculation with twice the plane-wave energy cutoff and twice the number of subdivisions along each reciprocal axis in the **k**-point grid than required for electronic convergence was carried out. The **k**-vectors of the highest occupied and lowest unoccupied energy levels were interpreted as VBM and CBM. This may still slightly miss the VBM or CBM, if it is not located at a high symmetry point. The second derivatives were then calculated using second-order central finite differences around the identified **k**-points with the same increased plane-wave cutoff. The mass terms contributing to the sum in Eq. (2) correspond well with those given in [58]. The reduced mass of the exciton μ_X is then defined in analogy to the reduced mass of the WM model as the harmonic mean, which favors small effective masses,

$$\frac{1}{\mu_X(\mathbf{k}_{\text{VBM}}, \mathbf{k}_{\text{CBM}})} = \left\langle \frac{1}{m_e(\mathbf{k}_{\text{CBM}})} \right\rangle + \left\langle \frac{1}{m_h(\mathbf{k}_{\text{VBM}})} \right\rangle. \tag{3}$$

(Co)Localized states near the VBM/CBM. Inspired by Ref. [19] and similar ideas, we define measures quantifying the number of electrons localized at some ion species, separately for the VBM and the CBM. The number of, e.g., valence electrons per unit cell near the ion coordinate $\mathbf{R}_I$ with energy close to the VBM is given in terms of the local density of states ρ_{DOS} as

$$N_{v,\mathbf{R}_I} = \int_{|\mathbf{r}-\mathbf{R}_I|\leqslant R} \mathrm{d}\mathbf{r} \int_{E_{\mathrm{VBM}}-\delta}^{E_{\mathrm{VBM}}} \mathrm{d}E\ \rho_{\mathrm{DOS}}(E,\mathbf{r}). \tag{4}$$

We have checked that the results presented below depend only weakly on our somewhat arbitrary choice of the radius and the energy range, i.e., $R = 1$ Å and $\delta = 1$ eV (see the Supplemental Material, SM [59]). If the unit cell contains more than one ion of the same species, we sum over the number of equivalent ions: $N_{v,\mathrm{Species}} = \sum_{I\in\mathrm{Species}} N_{v,\mathbf{R}_I}$. Analogously, the quantities $N_{c,\mathbf{R}_I}$ and $N_{c,\mathrm{Species}}$ are also defined for the conduction band. Now, localization measures $N_{v|c}$ are defined as the fractions

$$N_{v|c} = \frac{N_{v|c,\mathrm{S}^\star}}{\sum_{\mathrm{Species}} N_{v|c,\mathrm{Species}}},$$
$$\mathrm{S}^\star = \arg\max_{\mathrm{Species}} \{N_{v,\mathrm{Species}}\}. \tag{5}$$

Note that following Ref. [19] Eq. (5) breaks the $c \leftrightarrow v$ symmetry and focuses on valence states, i.e., anionic sites.

The product N_vN_c of both values is a measure of colocalization, i.e., it estimates in a mean-field sense the chance to find electrons and holes near the same atomic site. Thus, a large N_vN_c product should be a predictor for strong, Frenkel-like excitons. We calculate the spatially resolved integrated density of states (i.e., the number of states per volume) using the QUANTUM ESPRESSO postprocessing code pp.x. Our definition of colocalization differs from that of Dvorak *et al.* in Ref. [19] by minor differences in the range of integration and, more importantly, by the normalization, i.e., the denominator in Eq. (5). The latter allows to go beyond the binary compounds considered by Dvorak *et al.* and to handle more complex compounds on equal footing.

Dielectric constants. The electronic component of the static dielectric constant ε_r^e was calculated using density functional perturbation theory. The method is described in detail in Ref. [56]. The ionic part ε_r^i was determined from the dynamical matrix, i.e., the Hessian matrix of the total energy with respect to the ionic positions, calculated using a finite difference approach.

Wannierization. Various quantities are obtained from MLWFs [50], so it is warranted to discuss the Wannierization procedure separately. In general, Wannier functions $w_{n\mathbf{R}}(\mathbf{r}) = w_{n\mathbf{0}}(\mathbf{r}-\mathbf{R})$ can be defined for a group of J bands with Bloch states $|\psi_{n\mathbf{k}}\rangle$ via

$$|w_{n\mathbf{R}}\rangle = \frac{V}{(2\pi)^3}\int_{\mathrm{BZ}} \mathrm{d}\mathbf{k} e^{-i\mathbf{k}\cdot\mathbf{R}} \sum_{m=1}^{J} U_{mn}(\mathbf{k})|\psi_{m\mathbf{k}}\rangle \tag{6}$$

with the volume of the unit cell V and an arbitrary $\mathbf{k}$-dependent unitary $J\times J$ matrix $U_{mn}(\mathbf{k})$. The freedom in choosing U corresponds to the gauge freedom to choose the phase factor of the Bloch states. The Wannier functions as defined above are not necessarily localized. The MLWFs are obtained by minimizing the total spread

$$\Omega[U(.)] = \sum_{n=1}^{J}[\langle w_{n\mathbf{0}}|r^2|w_{n\mathbf{0}}\rangle - |\langle w_{n\mathbf{0}}|\mathbf{r}|w_{n\mathbf{0}}\rangle|^2] \tag{7}$$

with respect to U. Numerically, the MLWFs are obtained using WANNIER90 [50]. In this paper, we only cover properties related to the valence bands, accordingly we only Wannierize the topmost valence band manifold. As a starting point for the Wannierization, the SCDM algorithm is used [51,52]. We remark that the spread defined in Eq. (7) can be decomposed into two positive parts, the irreducible spread Ω_I,

$$\Omega_I = \sum_{n=1}^{J}\left[\langle w_{n\mathbf{0}}|r^2|w_{n\mathbf{0}}\rangle - \sum_{m\mathbf{R}}|\langle w_{m\mathbf{R}}|\mathbf{r}|w_{n\mathbf{0}}\rangle|^2\right], \tag{8}$$

which is independent of the choice of $U_{mn}(\mathbf{k})$, and the reducible spread, i.e., the remainder. The normalized irreducible spread is related to the direct band gap [36],

$$\frac{\Omega_I}{J} \leqslant \frac{3\hbar^2}{2m_eE_g} \tag{9}$$

and thus indirectly to the static dielectric constant [37].

Ionicity. Following the ionicity measure proposed by Abu-Farsakh and Qteish [34], we investigate the mean distance of the Wannier centers from their respective closest ions. The distance of the center of the nth MLWF from the closest ion is

$$\Xi_n = \min_{I\in\mathrm{Ions}} |\langle w_{n\mathbf{0}}|\mathbf{r}|w_{n\mathbf{0}}\rangle - \mathbf{R}_\mathrm{I}|^2. \tag{10}$$

The mean distance of the MLWF centers to the nearest ions of a group of J' bands is

$$\Xi = \frac{1}{J'}\sum_{n=1}^{J'}\Xi_n. \tag{11}$$

Reference [34] does not specify (a) which J Bloch states to include in the Wannierization and (b) which J' Wannier functions to include in Eq. (11). In an all-electron code [60], one could simply use all filled states. In a pseudopotential code like the codes used in this paper, this, however, introduces a dependence on the number of included (semi)core states, as discussed later. We opted to use $J' = J$ and to include only the top-most valence band manifold in both the Wannierization and the construction of the mean. The resulting value is then normalized by the average bond length, which is calculated by averaging over all unique minimum distances between unequal lattice sites ($i \neq j$) in a supercell. Fully covalent compounds thus have a value of $\Xi = 0.5$, corresponding to bond-centered MLWFs, while ionic compounds or compounds such as noble gas solids have $\Xi = 0$, corresponding to ion-centered MLWFs. We observed rather small Ξ_n values with valence manifold MLWFs looking, e.g., like slightly deformed p orbitals on the anion already for some bonds with partially covalent character, such as MgO, cf. discussion of Fig. 1(c) below.

Dispersion of the topmost valence band. As a measure of the mean dispersion of a band, we originally used the mean

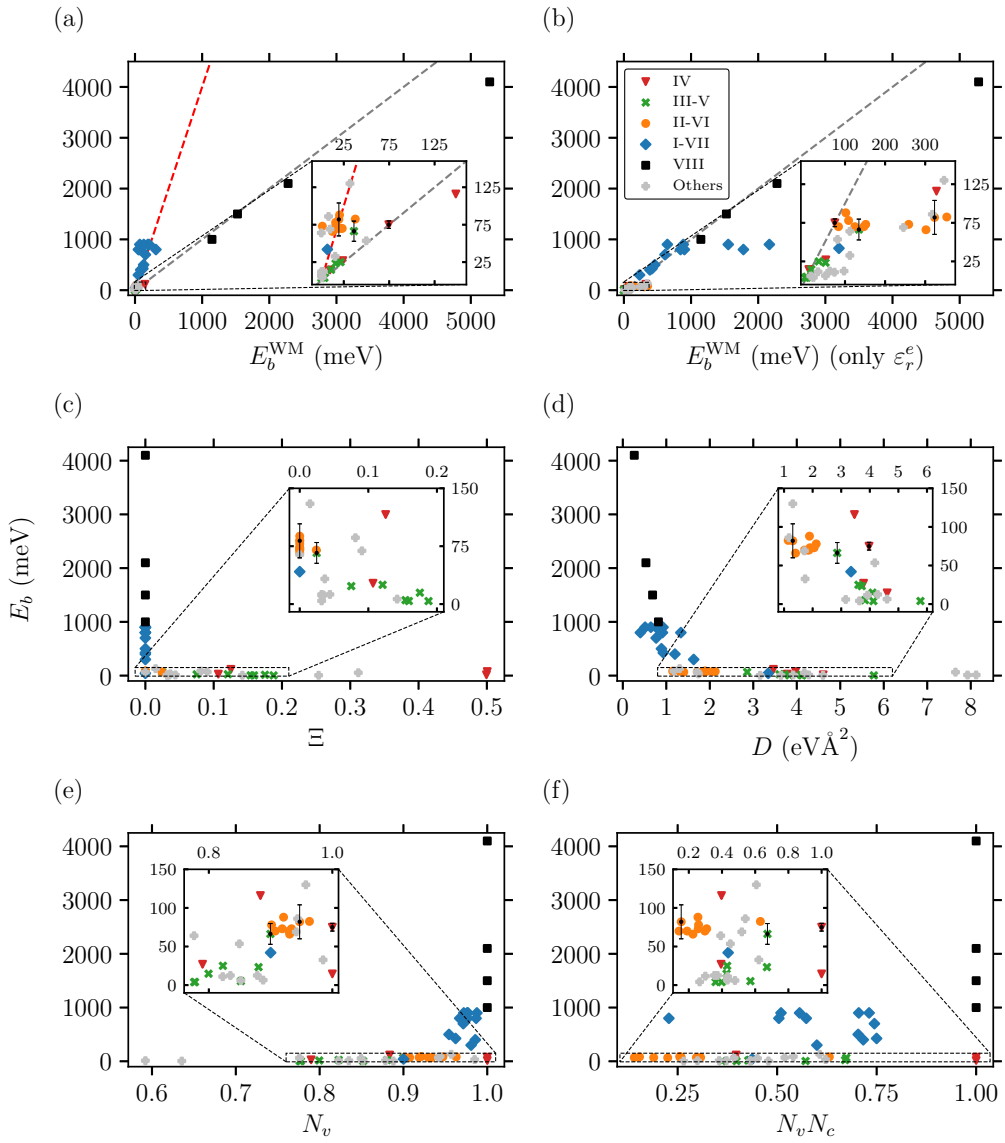

FIG. 1. Correlations between various ground-state properties on the x axis and the experimentally measured exciton binding energy E_b on the y axis, with the legend shown in panel (b). All graphs follow the same coloring scheme, with colors and markers corresponding to material classes. Error bars are included in the inset for materials with multiple available measurements whose standard deviation is larger than 5 meV (see Appendix). (a) Prediction of the WM model. The dashed-red and black lines correspond to $E_b = 4E_b^{\mathrm{WM}}$ and $E_b = E_b^{\mathrm{WM}}$, respectively. (b) Prediction of the WM model including only the electronic contribution to the dielectric function. (c) Ionicity defined through the mean distance of the MLWF centers. (d) Scaled dispersion of the topmost valence band. (e) Share of localized states near anion close to the VBM. (f) Share of colocalized states near anion close to the VBM and the CBM.

derivative $\tilde{D}$ as

$$\tilde{D} = \frac{1}{3}\sum_{i=1}^{3}\frac{1}{N_{\mathbf{k}}}\sum_{j=1}^{N_{\mathbf{k}}}\left|\frac{\partial E_n(\mathbf{k}_j)}{\partial k_i}\right|. \tag{12}$$

The gradient of the topmost valence band at each of the $N_{\mathbf{k}}$ $\mathbf{k}$-points, i.e., $\nabla_k E_n(\mathbf{k}_j)$, was obtained using WANNIER90 [50]. However, while $\tilde{D}$ was found to generally have good predictive power, materials with highly anisotropic primitive cells were found to be significant outliers, as shown within the SM [59]. Empirically, we found a correction of these outliers by evaluating a "scaled" mean dispersion D defined as

$$D = \frac{1}{3}\sum_{i=1}^{3}\frac{1}{N_{\mathbf{k}}}\sum_{j=1}^{N_{\mathbf{k}}}\left|\frac{\partial E_n(\mathbf{k}_j)}{\partial k_i}\right| \cdot L_i \tag{13}$$

where L_i is the largest value of the ith Cartesian component of the unit cell vectors. One could argue heuristically that this normalization corrects for bands with a low gradient along the Cartesian axis i caused by an elongated unit cell in that direction, e.g., where $L_z \gg L_x, L_y$. We are aware of the arbitrariness of this choice, since the values of D and $\tilde{D}$ obviously depend on the choice of unit cell and even the coordinate system: They change when a different or larger unit cell is used. They also change when the coordinate

system is rotated, because the individual components and the 1-norm $||\nabla_{\mathbf{k}}\epsilon||_1 = \sum_i |\partial_i\epsilon|$ enters instead of the rotational invariant 2-norm $||\nabla_{\mathbf{k}}\epsilon||_2 = (\sum_i |\partial_i\epsilon|^2)^{1/2}$. To mitigate this arbitrariness, we use a fixed method for obtaining standardized primitive unit cells, previously used for high-throughput electron-phonon coupling calculations for conventional superconductors [40,41]. As an example, we show as Fig. S7 within the SM [59] the correlations between Eq. (12) evaluated using either the L1- or the L2-norm. Results for the predictive power of this scaled dispersion D of the uppermost valence band are presented below. The analogously defined dispersion of the lowest conduction band yielded no significant predictive power (not shown). Note that the gradients obtained via WANNIER90 [50] define "bands" by enumerating sorted eigenvalues. Details on the superior performance of the less plausible D over $\tilde{D}$ are given within the SM [59].

The fact that plausible, but coordinate-system-dependent quantities depending on the choice of the unit cell such as $\tilde{D}$ and D have predictive power and that the latter shows superior results (i) highlights our still very incomplete understanding of exciton binding energies and (ii) poses the interesting, very general question how to deal in ML and ML-related research with predictors showing significant predictive power but—at least at present—insufficient physical justification. A positive attitude to the latter question is the view that today's "missing physical plausibility" reflects some deeper connection to be discovered and understanding to be obtained in the future.

V. RESULTS

A. Correlations

The correlations between some selected properties and E_b are shown in Fig. 1. The remaining correlations are shown in the SM [59]. It is evident that most correlations between the shown properties and E_b are significant in the sense that the data points suggest a monotonous trend and/or a natural clustering. It seems worthwhile to briefly discuss each correlation to see whether it agrees with the underlying *a priori* intuition. In each case, there will be outliers. As always in machine learning or high-throughput studies, outliers are a mixed blessing: On the one hand, they show that the presented observation, trend, or classification is not always valid. On the other hand, they suggest that some unusual physical phenomenon makes this sample special and, thus, interesting.

Wannier-Mott prediction. At first glance, Fig. 1(a) seems to indicate a roughly linear correlation $E_b \approx 4E_b^{\mathrm{WM}}$ except for the noble gas solids, where $E_b \approx E_b^{\mathrm{WM}}$ seems to hold. As the simplified isotropic Wannier-Mott model used in this paper cannot fully incorporate the diversity of the covered materials, a simple linear correlation is not to be expected. Indeed, the inset of Fig. 1(a) zooming in on the low-energy range shows two groups: For excitons with low binding energy (below and around 25 meV), the WM model using the calculated full static dielectric constant predicts E_b often quantitatively well. The general underestimation of E_b is most likely caused by the general underestimation of band gaps in DFT, which in turn leads to an overestimation of the dielectric constant. For excitons with an intermediate binding energy in the range 50–150 meV, the quantitative prediction is significantly worse, while a rough qualitative trend $E_b \approx 4E_b^{\mathrm{WM}}$ can be observed. As we are more interested in the predictive power of quantities than in *a priori* arguments, we also looked at the correlation of E_b with the WM model evaluated with only the electronic contribution to the dielectric constant, see Fig. 1(b). We do so because it has been suggested repeatedly that the ionic part of the screening contributes less for excitons with increased binding energy (see, e.g., Refs. [61,62] for the case of MgO; the theoretical underpinning is described in Refs. [15,16]). A general overestimation of the binding energy is observed.

Ionicity. Figure 1(c), which shows the binding energy plotted against the ionicity measure Ξ of Eq. (11), confirms the qualitative intuition: More ionic compounds with smaller Ξ generally have larger exciton binding energies. However, Ξ evaluated only on the valence band manifold collapses to a numerical zero already for many compounds which one would consider "intermediately ionic" like the binary II-VI compounds. The cause of this is evident when looking at the MLWFs of, e.g., MgO: The valence band manifold MWLFs consists of three p orbitals centered directly on the oxygen atom. Evaluating Eq. (10) for orbitals which are anti-/symmetric with respect to some atom yields zero. One can remedy this by including additional valence band manifolds, e.g., lower-lying s orbitals of the anion (which for tetrahedrally coordinated compounds typically yields sp^3-looking MLWFs). However, such an *ad hoc* inclusion is extremely difficult to automate and to generalize properly for more complex compounds. Simply including all valence bands introduces a dependence on the pseudopotentials employed in the DFT calculation, as the value of Ξ would then depend on how many (semi)core states are included.

Band dispersion. The valence band dispersion D shown in Fig. 1(d) also exhibits the expected correlation, with less dispersed, i.e., more localized, valence band states generally corresponding to larger exciton binding energies. We have not analyzed the equivalent property for the conduction band in detail, but a cursory analysis showed—in view of the explicit μ_X dependence of the WM model somehow surprisingly—that the conduction band dispersion has considerably less predictive power regarding the exciton binding energy. On second thought, this is not so surprising since in semiconductors band gap, electron mass(es), hole mass(es), and the electronic contribution to the dielectric constant are strongly correlated with each other. For example, in $\mathbf{k}\cdot\mathbf{p}$ theory, these quantities are all determined by the momentum matrix element and, thus, are all intimately related with each other.

Localization and colocalization. As shown in Fig. 1(e), in general the exciton binding energy E_b increases with an increased proportion of localized states near the anion N_v evaluated via Eq. (5). This agrees well with the earlier observation that more ionic compounds possess more strongly bound excitons. The approach, however, fails to account for compounds with a single species, where N_v trivially collapses to 1. Analyzing the corresponding property N_c for the conduction bands shows hardly any predictive power (see the SM [59] including Ref. [63]).

Dvorak *et al.* [19] evaluate the number of colocalized valence and conduction states, but they do not go into much detail regarding the exact definition. Somehow surprisingly, the analysis using our colocalization measure N_vN_c in Fig. 1(f)

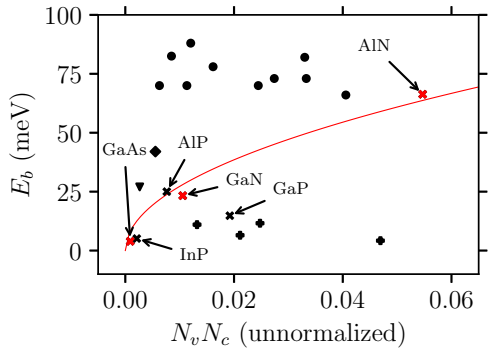

FIG. 2. Unnormalized number of colocalized states near the anion N_vN_c vs the experimental exciton binding energy E_b for binary compounds. The different markers correspond to different material classes as indicated in the legend of Fig. 1(b). Red markers indicate the materials analyzed by both us and Dvorak *et al.* [19]. The red line is the numerical fitting of $E_b = \alpha\sqrt{N_vN_c}$ through all red markers, as suggested in Ref. [19]. The data point for NaCl is off the scale of the plot to the top right, shown in the SM [59]. As it does not agree with the overall trend, it was ignored in the square-root fit.

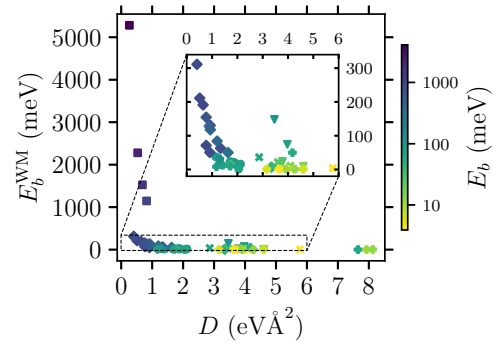

FIG. 3. Wannier-Mott prediction E_b^{WM} over the mean dispersion of the valence band D with color-coded experimental energies E_b. The different markers correspond to different material classes as indicated in the legend of Fig. 1(b).

shows a significantly worse predictive power than N_v alone. As shown in Fig. 2, the unnormalized number of colocalized states near the anion, i.e., ignoring the denominators in Eq. (5), recovers the square-root trend shown by Dvorak *et al.* [19] for the materials analyzed by both them and us, but does not generalize to the other binary materials.

As shown in the SM [59], we also do reproduce the results of Ref. [19] regarding a roughly linear relation between the electron localization length (expressed through the irreducible spread of the MLWFs [36]) and the electronic part of the dielectric constant. The results of Souza *et al.* [36] would, as discussed above, predict an inverse proportionality between E_G and Ω_I/J and thus a roughly square-root dependence between ε_e and Ω_I/J. Due to the inability of GGA functionals to describe small bandgap materials (large Ω_I/J) and with the spread we observe between different compounds, it is based on the presently available data not possible to decide which relation would better fit to our results.

B. Two-predictor classifications

Data analysis based on individual predictors, as described in the previous subsection, ideally leads to bijective correlations, or at least makes it possible to separate different data clusters by straight lines or simple boundaries. Using just one predicting property, this is evidently not possible with the selected predictors. Even though strong correlations were found, there were several outliers for each property. Fortunately, the outliers tend to be different for different traits. Therefore, the next logical step is to combine two or more properties as predictors. In this paper, we do not go beyond combinations of two properties, as initial attempts to go beyond more than two parameters using support vector regression or random forests led to significant overfitting.

We found empirically that, at least for our dataset, the most useful of the two-properties-to-E_b correlations is the combination of the scaled dispersion of the topmost valence band D and the energy E_b^{WM} of the WM model. This is shown in Fig. 3, where the color indicates the experimental binding energy: The exciton binding energy E_B clearly increases as one moves to the upper left region.

There is, apart from a few outliers, a roughly linear relation between N_v and D (shown as Figs. S3(b) and S4(b) in the SM [59]). Therefore, it is not surprising that a similarly high predictive power can be achieved using E_b^{WM} in combination with the proportion N_v of localized valence states on the anion (shown as Figs. S3(a) and S4(a) in the SM [59]). Using N_v instead of D avoids the construction of MLWFs, but introduces two additional parameters in the form of the radius and the energy range over which the integration of the spatially resolved density of states is carried out [see Eq. (4)]. More two-property correlations are shown within the SM [59].

We proceed with a discussion of the relation between D, the WM model and E_b. The trends seen in Fig. 3 can be used to cluster the exciton binding energies into three generally well-separated groups from 0 to 50 meV, 50 to 150 meV, and above 150 meV, see top panel of Fig. 4. The only clear outliers are SnO_2 and $CsPbCl_3$. Both materials are also outliers in the combination of the WM model E_b^{WM} and the share of localized valence states N_v, as shown within the SM [59]. We note that their exciton binding energies of 32.7 meV and 64 meV, respectively, are not too far from the somewhat arbitrary cluster limit of 50 meV, and that only one primary source was available for each material.

The predictive power does not degrade too much if the dielectric constant instead of the WM model is used. This choice avoids all ambiguities related to the calculation of the reduced exciton mass μ_X. The correlation of the two properties and the resulting clustering is shown in the middle panel of Fig. 4 using the DFT-calculated full dielectric constant. Corresponding figures using only the ionic or electronic contribution to the dielectric function are shown within the SM [59].

Finally, even using only the indirect bandgap instead of the WM model or the dielectric constant still results in significant predictive power (bottom panel of Fig. 4).

The results generally confirm the common belief that weak screening (corresponding to a large WM prediction, a small dielectric constant, and a large gap) is necessary for strongly bound excitons. In addition, the results tentatively suggest

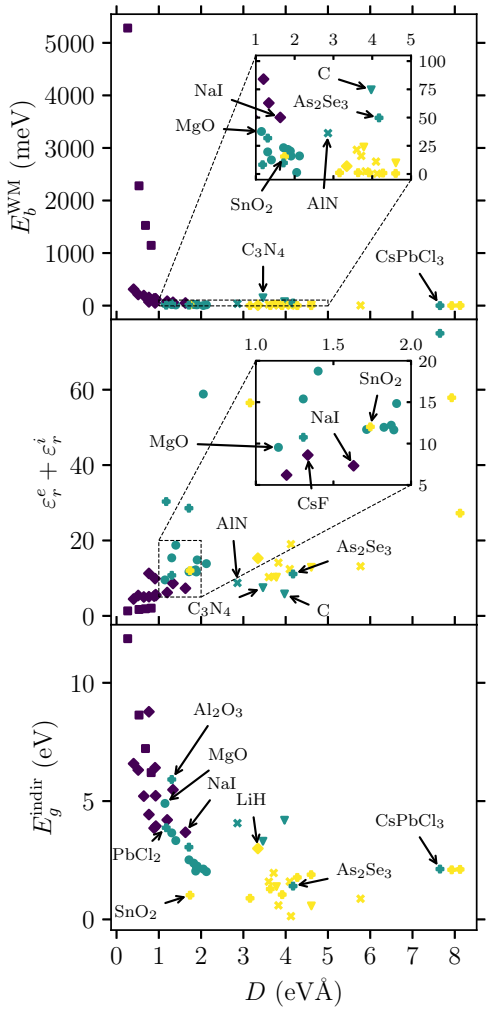

FIG. 4. Wannier-Mott prediction E_b^{WM}, full static dielectric constant $\varepsilon_r^e + \varepsilon_r^i$ and indirect band gap E_g^{indir} over the mean dispersion of the valence band D with the color determined from clustering E_b (yellow: $E_b < 50$ meV, turquoise: 50 meV $< E_b <$ 150 meV, purple: 150 meV $< E_b$). Interesting materials on the edges between clusters and outliers are annotated with arrows. The different markers correspond to different material classes as indicated in the legend of Fig. 1(b).

that especially weakly bound excitons require not only strong screening but also delocalized valence states.

VI. CONCLUSION

In summary, we have presented a broadly applicable method of clustering materials according to their exciton binding energies using only ground-state properties. We have carried out ground-state DFT calculations for 51 compounds from a large variety of material systems, and based on these, evaluated a diverse set of ground-state properties expected to correlate with the binding energy of the lowest bright exciton. We were able to confirm in most cases the proposed correlations, while also confirming the existence of several outliers. In addition to the WM model, properties more directly related to localization and ionicity also have significant predictive power. We believe that more refined predictors especially for the ionicity, for example using definitions based on Born effective charges [64], are a promising direction for future studies.

The combination of different properties can remove most of the outliers and, while still not providing a full quantitative prediction, allows for a highly reliable categorization. Such a computationally inexpensive categorization paves the way for efficient computation of optical properties, e.g., in the context of high-throughput studies. An easy-to-use workflow for evaluating all predictors as defined above is provided. If more high-quality experimental data or theoretical data were available, one could look at higher correlations involving more properties using for example, support vector machines, kernel ridge regression, random forests, or other small data machine learning methods [65].

The data and workflows supporting the results of this paper are publicly available at [66]. The workflow is designed to handle materials whose structure is available on the Materials Project [38] or as a CIF file. Possible future updates will be made available via GitHub [57].

ACKNOWLEDGMENTS

We thank the staff and especially Mr. Henning Schwanbeck of the Zentrum für Mikro- und Nanotechnologien and of the Compute Center of the Technische Universität Ilmenau for providing an excellent research environment. Additionally we would also like to thank Miguel A. L. Marques, Bochum, Germany, for inspiring discussions and the provision of the automated symmetry detection aiding the QUANTUM ESPRESSO workflows. This work is supported by the Deutsche Forschungsgemeinschaft DFG (Grant No. 537033066).

APPENDIX: EXCITON BINDING ENERGY DATASET

In this Appendix, Tables I–VI, we present the curated dataset containing experimental exciton binding energies for 51 bulk compounds, split into meaningful material groupings. If more than one plausible reference for the exciton binding energy was found for a given material, we used the average value in our analysis. The Materials Project [38] ID of the given material is given in the ID column.

TABLE I. Dataset for group IV materials.

Group	Material	ID	E_b (meV)	Reference
IV	Diamond	mp-66	70	[67]
			80	[68]
	Si	mp-149	14.3	[69]
			14.7	[70]
	C_3N_4	mp-1985	116	[71]
	3C-SiC	mp-8062	27	[72]

TABLE II. Dataset for group III-V materials.

Group	Material	ID	E_b (meV)	Reference
III-V	AlN	mp-661	48	[73]
			71	[74]
			80	[75]
			20	[76]
			20	[77]
			20	[78]
			20.4	[79]
			21	[80]
	GaN	mp-804	25	[81]
			25	[21]
			25.3	[82]
			26	[83]
			26.4	[84]
			26.7	[85]
	AlP	mp-1550	25	[86]
	GaP	mp-2490	10	[87]
			19.5	[88]
	GaAs	mp-2534	3.1-4.0	[89]
			4.2	[90]
	InP	mp-20351	5.1	[91]
	$AlGaAs_2$	mp-1228891	4	[92]

TABLE III. Dataset for group II-VI materials.

Group	Material	ID	E_b (meV)	Reference
II-VI	SrS	mp-1087	70	[25]
	BaSe	mp-1253	78	[25]
	MgO	mp-1265	80	[62]
			85	[61]
	BaO	mp-1342	73	[25]
	CaSe	mp-1415	70	[25]
	BaS	mp-1500	73	[25]
	CaS	mp-1672	70	[25]
	SrO	mp-2472	66	[25]
	CaO	mp-2605	60	[25]
			104	[61]
	SrSe	mp-2758	88	[25]

TABLE IV. Dataset for group I-VII materials.

Group	Material	ID	E_b (meV)	Reference
I-VII	KF	mp-463	900	[93]
	NaF	mp-682	800	[93]
	LiF	mp-1138	900	[93]
	CsF	mp-1784	800	[93]
	RbF	mp-2064	800	[93]
	NaCl	mp-22862	900	[93]
	KI	mp-22898	400	[1]
			450	[94]
	RbI	mp-22903	500	[94]
	NaBr	mp-22916	400	[93]
	KCl	mp-23193	900	[93]
	KBr	mp-23251	700	[93]
	NaI	mp-23268	300	[1]
	LiH	mp-23703	42	[95]

TABLE V. Dataset for the rare gas solids.

Group	Material	ID	E_b (meV)	Reference
VIII	Ne	mp-111	4100	[93]
	Ar	mp-23155	2100	[93]
	Kr	mp-612118	1500	[93]
	Xe	mp-611517	1000	[93]

TABLE VI. Dataset for all materials not included in previous tables.

Group	Material	ID	E_b (meV)	Reference
Others	SnO_2	mp-856	32.7	[96]
	As_2Se_3	mp-909	50	[97]
			57	[98]
	α-Al_2O_3	mp-1143	130	[99]
	Sb_2Se_3	mp-2160	6	[100]
	InBr	mp-22870	11.6	[101]
	TlBr	mp-22875	6.5	[102]
	$CsPbCl_3$	mp-23037	64	[103]
	TlCl	mp-23167	11	[102]
	InI	mp-23202	4.2	[101]
	$PbCl_2$	mp-23291	86	[104]
	$PbBr_2$	mp-28077	69	[104]
	$MaPbBr_3$	mp-732337	10-15	[24]
	$MaPbI_3$	mp-995214	10-15	[24]

[1] M. Fox, *Optical Properties of Solids*, 2nd ed., Oxford Master Series in Physics (Oxford University Press, London, 2010).

[2] M. A. Green, A. Ho-Baillie, and H. J. Snaith, The emergence of perovskite solar cells, Nat. Photon. **8**, 506 (2014).

[3] Y. Li, Y.-L. Li, B. Sa, and R. Ahuja, Review of two-dimensional materials for photocatalytic water splitting from a theoretical perspective, Catal. Sci. Technol. **7**, 545 (2017).

[4] Y. Shi, G. Zhan, H. Li, X. Wang, X. Liu, L. Shi, K. Wei, C. Ling, Z. Li, H. Wang *et al.*, Simultaneous manipulation of bulk excitons and surface defects for ultrastable and highly selective CO_2 photoreduction, Adv. Mater. **33**, 2100143 (2021).

[5] F. Krausz and M. Ivanov, Attosecond physics, Rev. Mod. Phys. **81**, 163 (2009).

[6] A. Kavokin, T. C. H. Liew, C. Schneider, P. G. Lagoudakis, S. Klembt, and S. Hoefling, Polariton condensates for classical and quantum computing, Nat. Rev. Phys. **4**, 435 (2022).

[7] R. D. Schaller and V. I. Klimov, High efficiency carrier multiplication in PbSe nanocrystals: Implications for solar energy conversion, Phys. Rev. Lett. **92**, 186601 (2004).

[8] A. Khan, K. Balakrishnan, and T. Katona, Ultraviolet light-emitting diodes based on group three nitrides, Nat. Photon. **2**, 77 (2008).

[9] C. Dang, J. Lee, C. Breen, J. S. Steckel, S. Coe-Sullivan, and A. Nurmikko, Red, green and blue lasing enabled by single-exciton gain in colloidal quantum dot films, Nat. Nanotechnol. **7**, 335 (2012).

[10] E. Runge and E. K. U. Gross, Density-functional theory for time-dependent systems, Phys. Rev. Lett. **52**, 997 (1984).

[11] E. K. U. Gross and W. Kohn, Local density-functional theory of frequency-dependent linear response, Phys. Rev. Lett. **55**, 2850 (1985).

[12] Y.-M. Byun, J. Sun, and C. A. Ullrich, Time-dependent density-functional theory for periodic solids: Assessment of excitonic exchange-correlation kernels, Electron. Struct. **2**, 023002 (2020).

[13] A. Siarkos, E. Runge, and R. Zimmermann, Center-of-mass properties of the exciton in quantum wells, Phys. Rev. B **61**, 10854 (2000).

[14] A. Siarkos and E. Runge, Quantum-wire exciton dispersion in a multiband real-space scheme, Phys. Rev. B **61**, 16854 (2000).

[15] S. D. Mahanti and C. M. Varma, Effective electron-hole interactions in polar semiconductors, Phys. Rev. B **6**, 2209 (1972).

[16] R. S. Knox, *Theory of Excitons* (Academic Press, San Diego, CA, 1963).

[17] F. Bechstedt, K. Seino, P. H. Hahn, and W. G. Schmidt, Quasiparticle bands and optical spectra of highly ionic crystals: AlN and NaCl, Phys. Rev. B **72**, 245114 (2005).

[18] E. Baldini, L. Chiodo, A. Dominguez, M. Palummo, S. Moser, M. Yazdi-Rizi, G. Auböck, B. Mallett, H. Berger, A. Magrez *et al.*, Strongly bound excitons in anatase TiO_2 single crystals and nanoparticles, Nat. Commun. **8**, 13 (2017).

[19] M. Dvorak, S.-H. Wei, and Z. Wu, Origin of the variation of exciton binding energy in semiconductors, Phys. Rev. Lett. **110**, 016402 (2013).

[20] Z. Jiang, Z. Liu, Y. Li, and W. Duan, Scaling universality between band gap and exciton binding energy of two-dimensional semiconductors, Phys. Rev. Lett. **118**, 266401 (2017).

[21] A. Alemu, B. Gil, M. Julier, and S. Nakamura, Optical properties of wurtzite GaN epilayers grown on *a*-plane sapphire, Phys. Rev. B **57**, 3761 (1998).

[22] D. M. Niedzwiedzki, H. Zhou, and P. Biswas, Exciton binding energy of $MAPbI_3$ thin film elucidated via analysis and modeling of perovskite absorption and photoluminescence properties using various methodologies, J. Phys. Chem. C **126**, 1046 (2022).

[23] P. Ghosh and B. Ray, Luminescence in alkaline earth sulphides, Prog. Cryst. Growth Charact. Mater. **25**, 1 (1992).

[24] M. Baranowski and P. Plochocka, Excitons in metal-halide perovskites, Adv. Energy Mater. **10**, 1903659 (2020).

[25] Y. Kaneko and T. Koda, New developments in IIa–VIb (alkaline-earth chalcogenide) binary semiconductors, J. Cryst. Growth **86**, 72 (1988).

[26] N. O. Lipari and A. Baldereschi, Energy levels of indirect excitons in semiconductors with degenerate bands, Phys. Rev. B **3**, 2497 (1971).

[27] R. M. Martin, L. Reining, and D. M. Ceperley, *Interacting Electrons* (Cambridge University Press, Cambridge, 2016).

[28] F. Fuchs, C. Rödl, A. Schleife, and F. Bechstedt, Efficient $\mathcal{O}(N^2)$ approach to solve the Bethe-Salpeter equation for excitonic bound states, Phys. Rev. B **78**, 085103 (2008).

[29] M. R. Filip, J. B. Haber, and J. B. Neaton, Phonon screening of excitons in semiconductors: Halide perovskites and beyond, Phys. Rev. Lett. **127**, 067401 (2021).

[30] L. Adamska and P. Umari, Bethe-Salpeter equation approach with electron-phonon coupling for exciton binding energies, Phys. Rev. B **103**, 075201 (2021).

[31] A. M. Alvertis, J. B. Haber, Z. Li, C. J. N. Coveney, S. G. Louie, M. R. Filip, and J. B. Neaton, Phonon screening and dissociation of excitons at finite temperatures from first principles, Proc. Natl. Acad. Sci. USA **121**, e2403434121 (2024).

[32] A. M. Alvertis, A. Champagne, M. Del Ben, F. H. da Jornada, D. Y. Qiu, M. R. Filip, and J. B. Neaton, Importance of nonuniform Brillouin zone sampling for *ab initio* Bethe-Salpeter equation calculations of exciton binding energies in crystalline solids, Phys. Rev. B **108**, 235117 (2023).

[33] C. R. A. Catlow and A. M. Stoneham, Ionicity in solids, J. Phys. C **16**, 4321 (1983).

[34] H. Abu-Farsakh and A. Qteish, Ionicity scale based on the centers of maximally localized Wannier functions, Phys. Rev. B **75**, 085201 (2007).

[35] N. Marzari, A. A. Mostofi, J. R. Yates, I. Souza, and D. Vanderbilt, Maximally localized Wannier functions: Theory and applications, Rev. Mod. Phys. **84**, 1419 (2012).

[36] I. Souza, T. Wilkens, and R. M. Martin, Polarization and localization in insulators: Generating function approach, Phys. Rev. B **62**, 1666 (2000).

[37] P. J. M. A. Carriço, M. Ferreira, T. F. T. Cerqueira, F. Nogueira, and P. Borlido, High-refractive-index materials screening from machine learning and *ab initio* methods, Phys. Rev. Mater. **8**, 015201 (2024).

[38] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, Commentary: The materials project: A materials genome approach to accelerating materials innovation, APL Mater. **1**, 011002 (2013).

[39] S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, and K. A. Persson, The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles, Comput. Mater. Sci. **97**, 209 (2015).

[40] T. F. T. Cerqueira, A. Sanna, and M. A. L. Marques, Sampling the materials space for conventional superconducting compounds, Adv. Mater. **36**, 2307085 (2023).

[41] K. Gao, W. Cui, J. Shi, A. P. Durajski, J. Hao, S. Botti, M. A. L. Marques, and Y. Li, Prediction of high-T_c superconductivity in ternary actinium beryllium hydrides at low pressure, Phys. Rev. B **109**, 014501 (2024).

[42] A. Togo and I. Tanaka, `Spglib`: A software library for crystal symmetry search, arXiv:1808.01590.

[43] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, Comput. Mater. Sci. **68**, 314 (2013).

[44] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. **77**, 3865 (1996).

[45] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo *et al.*, QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials, J. Phys.: Condens. Matter **21**, 395502 (2009).

[46] P. Giannozzi Jr, O. Andreussi, T. Brumme, O. Bunau, M. B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni *et al.*, Advanced capabilities for materials modelling with QUANTUM ESPRESSO, J. Phys.: Condens. Matter **29**, 465901 (2017).

[47] M. Schlipf and F. Gygi, Optimization algorithm for the generation of ONCV pseudopotentials, Comput. Phys. Commun. **196**, 36 (2015).

[48] D. R. Hamann, Optimized norm-conserving Vanderbilt pseudopotentials, Phys. Rev. B **88**, 085117 (2013).

[49] A. Jain, G. Hautier, C. J. Moore, S. Ping Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, A high-throughput infrastructure for density functional theory calculations, Comput. Mater. Sci. **50**, 2295 (2011).

[50] G. Pizzi, V. Vitale, R. Arita, S. Blügel, F. Freimuth, G. Géranton, M. Gibertini, D. Gresch, C. Johnson, T. Koretsune *et al.*, WANNIER90 as a community code: New features and applications, J. Phys.: Condens. Matter **32**, 165902 (2020).

[51] V. Vitale, G. Pizzi, A. Marrazzo, J. R. Yates, N. Marzari, and A. A. Mostofi, Automated high-throughput Wannierisation, npj Comput. Mater. **6**, 66 (2020).

[52] A. Damle, L. Lin, and L. Ying, Compressed representation of Kohn-Sham orbitals via selected columns of the density matrix, J. Chem. Theory Comput. **11**, 1463 (2015).

[53] P. E. Blöchl, O. Jepsen, and O. K. Andersen, Improved tetrahedron method for Brillouin-zone integrations, Phys. Rev. B **49**, 16223 (1994).

[54] G. Kresse and J. Furthmüller, Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set, Phys. Rev. B **54**, 11169 (1996).

[55] G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B **59**, 1758 (1999).

[56] M. Gajdoš, K. Hummer, G. Kresse, J. Furthmüller, and F. Bechstedt, Linear optical properties in the projector-augmented wave methodology, Phys. Rev. B **73**, 045112 (2006).

[57] M. Grunert and M. Großmann, Github: Predicting exciton binding energies from ground state properties, http://github.com/magr4826/ExcitonBindingPrediction (2024).

[58] J. L. Janssen, Y. Gillet, S. Poncé, A. Martin, M. Torrent, and X. Gonze, Precise effective masses from density functional perturbation theory, Phys. Rev. B **93**, 205147 (2016).

[59] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevB.110.075204 for correlations and classifications with one or two predictors not shown in the main text and details on the dependence of the predictors in the main text on their parameters.

[60] S. Tillack, A. Gulans, and C. Draxl, Maximally localized Wannier functions within the (L)APW+LO method, Phys. Rev. B **101**, 235102 (2020).

[61] R. C. Whited and W. C. Walker, Exciton spectra of CaO and MgO, Phys. Rev. Lett. **22**, 1428 (1969).

[62] D. M. Roessler and W. C. Walker, Electronic spectrum and ultraviolet optical properties of crystalline MgO, Phys. Rev. **159**, 733 (1967).

[63] P. K. de Boer and R. A. de Groot, The origin of the conduction band in table salt, Am. J. Phys. **67**, 443 (1999).

[64] G. Wellenhofer, K. Karch, P. Pavone, U. Rössler, and D. Strauch, Pressure dependence of static and dynamic ionicity of SiC polytypes, Phys. Rev. B **53**, 6071 (1996).

[65] P. Xu, X. Ji, M. Li, and W. Lu, Small data machine learning in materials science, npj Comput. Mater. **9**, 42 (2023).

[66] M. Grunert and M. Großmann, Zenodo: Predicting exciton binding energies from ground state properties v1.2, https://zenodo.org/records/11235900 (2024).

[67] C. D. Clark, P. J. Dean, and P. V. Harris, Intrinsic edge absorption in diamond, Proc. R. Soc. London A **277**, 312 (1964).

[68] P. J. Dean, E. C. Lightowlers, and D. R. Wight, Intrinsic and extrinsic recombination radiation from natural and synthetic aluminum-doped diamond, Phys. Rev. **140**, A352 (1965).

[69] T. Nishino, M. Takeda, and Y. Hamakawa, Analysis of derivative spectrum of indirect exciton absorption in silicon, Solid State Commun. **12**, 1137 (1973).

[70] K. L. Shaklee and R. E. Nahory, Valley-orbit splitting of free excitons? The absorption edge of Si, Phys. Rev. Lett. **24**, 942 (1970).

[71] Y. Shi, J. Li, C. Mao, S. Liu, X. Wang, X. Liu, S. Zhao, X. Liu, Y. Huang, and L. Zhang, van Der Waals gap-rich BiOCl atomic layers realizing efficient, pure-water CO_2-to-CO photocatalysis, Nat. Commun. **12**, 5923 (2021).

[72] R. Humphreys, D. Bimberg, and W. Choyke, Wavelength modulated absorption in SiC, Solid State Commun. **39**, 163 (1981).

[73] E. Silveira, J. A. Freitas, O. J. Glembocki, G. A. Slack, and L. J. Schowalter, Excitonic structure of bulk AlN from optical reflectivity and cathodoluminescence measurements, Phys. Rev. B **71**, 041201(R) (2005).

[74] L. Chen, B. J. Skromme, R. F. Dalmau, R. Schlesser, Z. Sitar, C. Chen, W. Sun, J. Yang, M. A. Khan, M. L. Nakarmi *et al.*,

Band-edge exciton states in AlN single crystals and epitaxial layers, Appl. Phys. Lett. **85**, 4334 (2004).
[75] J. Li, K. B. Nam, M. L. Nakarmi, J. Y. Lin, H. X. Jiang, P. Carrier, and S.-H. Wei, Band structure and fundamental optical transitions in wurtzite AlN, Appl. Phys. Lett. **83**, 5163 (2003).
[76] M. Smith, G. D. Chen, J. Z. Li, J. Y. Lin, H. X. Jiang, A. Salvador, W. K. Kim, O. Aktas, A. Botchkarev, and H. Morkoç, Excitonic recombination in GaN grown by molecular beam epitaxy, Appl. Phys. Lett. **67**, 3387 (1995).
[77] D. C. Reynolds, D. C. Look, W. Kim, O. Aktas, A. Botchkarev, A. Salvador, H. Morkoç, and D. N. Talwar, Ground and excited state exciton spectra from GaN grown by molecular-beam epitaxy, J. Appl. Phys. **80**, 594 (1996).
[78] M. Tchounkeu, O. Briot, B. Gil, J. P. Alexis, and R.-L. Aulombard, Optical properties of GaN epilayers on sapphire, J. Appl. Phys. **80**, 5352 (1996).
[79] J. F. Muth, J. H. Lee, I. K. Shmagin, R. M. Kolbas, H. C. Casey, B. P. Keller, U. K. Mishra, and S. P. DenBaars, Absorption coefficient, energy gap, exciton binding energy, and recombination lifetime of GaN obtained from transmission measurements, Appl. Phys. Lett. **71**, 2572 (1997).
[80] W. Shan, B. D. Little, A. J. Fischer, J. J. Song, B. Goldenberg, W. G. Perry, M. D. Bremser, and R. F. Davis, Binding energy for the intrinsic excitons in wurtzite GaN, Phys. Rev. B **54**, 16369 (1996).
[81] M. Leroux, B. Beaumont, N. Grandjean, C. Golivet, P. Gibart, J. Massies, J. Leymarie, A. Vasson, and A. Vasson, Comparative optical characterization of GaN grown by metal-organic vapor phase epitaxy, gas source molecular beam epitaxy and halide vapor phase epitaxy, Mater. Sci. Eng. B **43**, 237 (1997).
[82] S. Chichibu, A. Shikanai, T. Azuhata, T. Sota, A. Kuramata, K. Horino, and S. Nakamura, Effects of biaxial strain on exciton resonance energies of hexagonal GaN heteroepitaxial layers, Appl. Phys. Lett. **68**, 3766 (1996).
[83] A. Shikanai, T. Azuhata, T. Sota, S. Chichibu, A. Kuramata, K. Horino, and S. Nakamura, Biaxial strain dependence of exciton resonance energies in wurtzite GaN, J. Appl. Phys. **81**, 417 (1997).
[84] B. Skromme, Optical and magneto-optical characterization of heteroepitaxial gallium nitride, Mater. Sci. Eng., B **50**, 117 (1997).
[85] D. Volm, K. Oettinger, T. Streibl, D. Kovalev, M. Ben-Chorin, J. Diener, B. K. Meyer, J. Majewski, L. Eckey, A. Hoffmann, H. Amano, I. Akasaki, K. Hiramatsu, and T. Detchprohm, Exciton fine structure in undoped GaN epitaxial films, Phys. Rev. B **53**, 16543 (1996).
[86] M. Lorenz, R. Chicotka, G. Pettit, and P. Dean, The fundamental absorption edge of AlAs and AlP, Solid State Commun. **8**, 693 (1970).
[87] P. J. Dean and D. G. Thomas, Intrinsic absorption-edge spectrum of gallium phosphide, Phys. Rev. **150**, 690 (1966).
[88] R. G. Humphreys, U. Rössler, and M. Cardona, Indirect exciton fine structure in GaP and the effect of uniaxial stress, Phys. Rev. B **18**, 5590 (1978).
[89] A. R. Goi, A. Cantarero, K. Syassen, and M. Cardona, Effect of pressure on the low-temperature exciton absorption in GaAs, Phys. Rev. B **41**, 10111 (1990).
[90] D. D. Sell, Resolved free-exciton transitions in the optical-absorption spectrum of GaAs, Phys. Rev. B **6**, 3750 (1972).
[91] H. Mathieu, Y. Chen, J. Camassel, J. Allegre, and D. S. Robertson, Excitons and polaritons in InP, Phys. Rev. B **32**, 4042 (1985).
[92] B. Monemar, K. K. Shih, and G. D. Pettit, Some optical properties of the Al_xGa_{1-x} As alloys system, J. Appl. Phys. **47**, 2604 (1976).
[93] F. Bechstedt, *Many-Body Approach to Electronic Excitations: Concepts and Applications* (Springer, Berlin, Heidelberg, 2015).
[94] D. M. Roessler and W. C. Walker, Exciton structure in the ultraviolet spectra of KI and RbI, J. Optic. Soc. Am. **57**, 677 (1967).
[95] V. Plekhanov, Comparative study of isotope and chemical effects on the exciton states in LiH crystals, Prog. Solid State Chem. **29**, 71 (2001).
[96] K. Reimann and M. Steube, Experimental determination of the electronic band structure of SnO_2, Solid State Commun. **105**, 649 (1998).
[97] J. Ristein and G. Weiser, Quenching by electric fields of the luminescence of As_2Se_3 single crystals, Solid State Commun. **57**, 639 (1986).
[98] R. S. Sussmann, T. M. Searle, and I. G. Austin, The excitonic gaps and Urbach tail in crystalline As_2Se_3: Absorption and electroabsorption studies, Philos. Mag. B **44**, 665 (1981).
[99] R. H. French, D. J. Jones, and S. Loughin, Interband electronic structure of α-alumina up to 2167 K, J. Am. Ceram. Soc. **77**, 412 (1994).
[100] J. Krustok, R. Kondrotas, R. Nedzinskas, K. Timmo, R. Kaupmees, V. Mikli, and M. Grossberg, Identification of excitons and biexcitons in Sb_2Se_3 under high photoluminescence excitation density, Adv. Opt. Mater. **9**, 2100107 (2021).
[101] N. Ohno, M. Yoshida, K. Nakamura, J. Nakahara, and K. Kobayashi, Magneto-reflectance of excitons in indium halides, J. Phys. Soc. Jpn. **53**, 1548 (1984).
[102] R. Z. Bachrach and F. C. Brown, Exciton-optical properties of TlBr and TlCl, Phys. Rev. B **1**, 818 (1970).
[103] M. Baranowski, P. Plochocka, R. Su, L. Legrand, T. Barisien, F. Bernardot, Q. Xiong, C. Testelin, and M. Chamarro, Exciton binding energy and effective mass of $CsPbCl_3$: A magneto-optical study, Photon. Res. **8**, A50 (2020).
[104] M. Fujita, M. Itoh, Y. Bokumoto, H. Nakagawa, D. L. Alov, and M. Kitaura, Optical spectra and electronic structures of lead halides, Phys. Rev. B **61**, 15731 (2000).