PAH101: A GW+BSE Dataset of 101 Polycyclic Aromatic Hydrocarbon (PAH) Molecular Crystals

Siyu Gao‡[1], Xingyu Liu‡[1], Yiqun Luo[2], Xiaopeng Wang[3], Kaiji Zhao[1], Vincent Chang[1], Bohdan Schatschneider[4], and Noa Marom[1,2,5,*]

[1]Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[2]Department of Physics, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[3]School of Foundational Education, University of Health and Rehabilitation Sciences, Qingdao 266113, China
[4]Department of Chemistry and Biochemistry, California State Polytechnic University at Pomona, Pomona, CA, 91768, USA
[5]Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[*]corresponding author(s): Noa Marom (nmarom@andrew.cmu.edu)

ABSTRACT

The excited-state properties of molecular crystals are important for applications in organic electronic devices. The GW approximation and Bethe-Salpeter equation (GW+BSE) is the state-of-the-art method for calculating the excited-state properties of crystalline solids with periodic boundary conditions. We present the PAH101 dataset of GW+BSE calculations for 101 molecular crystals of polycyclic aromatic hydrocarbons (PAHs) with up to ∼500 atoms in the unit cell. The data records include the GW quasiparticle band structure, the fundamental band gap, the static dielectric constant, the first singlet exciton energy (optical gap), the first triplet exciton energy, the dielectric function, and optical absorption spectra for light polarized along the three lattice vectors. We envision the dataset being used to (i) identify correlations between DFT and GW+BSE quantities, (ii) discover materials with desired electronic/ optical properties in the dataset itself, and (iii) train machine learned models to help in materials discovery efforts. We provide examples to illustrate these three use cases.

Background & Summary

Computational materials design and discovery requires exploring the infinitely vast chemical space using quantum mechanical methods that can reliably predict the electronic and optical properties of candidate materials. The computational cost of quantum mechanical simulations increases rapidly with the method accuracy and system size. This limits the scope of simulations that can be performed within a reasonable time in terms of the number of systems explored, their size, the accuracy of the predicted properties, and the types of phenomena that can be investigated.[1–7]

Density functional theory (DFT) is the workhorse of first-principles simulations.[8] DFT relies on approximate exchange-correlation functionals to describe the many-body quantum mechanical interactions between electrons. Computationally efficient semi-local functionals have been used extensively for high-throughput materials screening.[9–18] However, DFT is a ground-state theory, therefore it is inherently unable to describe excited-state properties of interest, such as fundamental band gaps, singlet and triplet excitation energies, optical gaps (*i.e.*, the first singlet excitation energy), and optical absorption spectra. The excited states of isolated molecules may be calculated relatively efficiently with time dependent DFT (TDDFT).[19–23] The excited states of crystalline systems may be calculated using Green's function based many-body perturbation theory (MBPT) within the GW approximation and Bethe-Salpeter equation (BSE)[24–28], which lends itself more easily than TDDFT to periodic implementations. Unfortunately, the high computational cost of GW+BSE simulations makes it unfeasible to use these methods for large scale materials exploration.

Machine learning (ML) may accelerate computational materials discovery by bypassing the need to perform expensive first-principles simulations.[29–39] To this end, statistical models are constructed based on training data to make predictions for new data points. Training ML models, especially deep neural networks (DNN), typically requires huge datasets. Therefore, data acquisition is often the bottleneck of applying ML to computational materials discovery. With the supercomputing resources available nowadays, acquiring DFT training data with semi-local functionals is relatively fast. This has led to the proliferation of DFT datasets.[29,40–47] As a result, ML models have been trained predominantly on semi-local DFT data, which limits their applicability to structural and ground state properties. Owing to the high computational cost of GW+BSE, such datasets are scarce and the amount of data they contain is relatively small compared to DFT datasets.[46,48,49] We note that the GW datasets

‡ These authors contributed equally to this work.

cited here comprise small isolated molecules, which are considerably faster to calculate than periodic molecular crystals with hundreds of atoms in the unit cell. Recently, ML has been applied to predict the *GW* quasiparticle energies of small molecules.[50]

It is challenging to construct transferable ML models based on "small data". This has limited the applicability of ML to excited state properties of molecular crystals. Emerging approaches to ML with small data include multi-fidelity approaches. These methods combine a small amount of high-fidelity data with a large amount of low-fidelity data, which, although not as accurate, is sufficiently correlated with the high-fidelity data for statistical inference.[51–60] Recently, high-quality results have been achieved by fine-tuning a pre-trained DNN model with small datasets or combining feature selection with DNN.[61,62] Other approaches involve using low-fidelity features, selected based on physical/chemical knowledge, to construct surrogate models that are predictive of high-fidelity data. One such approach is the sure-independence-screening-and-sparsifying-operator (SISSO)[63,64] ML algorithm. The input of SISSO is a set of primary features, which are physical descriptors that could be correlated with the target property. SISSO generates a huge feature space by iteratively combining the primary features using linear and nonlinear algebraic operations. Subsequently, linear regression is performed to identify the most predictive models. Physical and chemical knowledge is leveraged in the choice of primary features and in the rules for combining them. SISSO has been demonstrated to work well with a relatively small amount of data for several different types of materials systems and properties.[13,65–83]

One application that requires predicting the excited-state properties of molecular crystals is singlet fission (SF), the conversion of one singlet exciton into two triplet excitons.[84–87] The efficiency of solar cells can be boosted by augmenting traditional absorbers with SF materials.[88–90] The SF material can convert photons with energies high above the traditional absorber's band gap into two charge carriers instead of losing their excess energy to heat. Currently, few classes of materials are known to undergo intermolecular SF in the solid state, and insufficient stability under operating conditions precludes their utilization in commercial modules.[84,85,91,92] Therefore, there is a need for computational discovery of new SF materials. The primary criterion for a material to undergo SF is the thermodynamic driving force, *i.e.,* the difference between the singlet exciton energy and twice the triplet exciton energy, $E_S - 2E_T$, which can be calculated using GW+BSE.[93–97]

Recently, we have used SISSO to find models based on low-cost DFT properties that can reliably predict the GW+BSE SF driving force.[98] SISSO generated several models that predicted the GW+BSE SF driving force with errors below 0.2 eV. Based on considerations of accuracy and computational cost, two SISSO models were selected to build a two-step hierarchical classifier for screening promising candidates for SF. To train SISSO, we generated a dataset of *GW*+BSE calculations of the SF driving force of 101 molecular crystals of polycyclic aromatic hydrocarbons (PAHs). PAHs are compounds comprising carbon and hydrogen atoms and containing multiple aromatic rings. Most SF materials are PAHs. In addition to SF, PAHs and their functionalized derivatives have versatile applications in organic electronic devices.[99–109] To form the PAH101 set, crystal structures of unsubstituted PAHs (containing only C and H atoms) were extracted from the Cambridge Structural Database (CSD)[110]. The PAH101 set contains several sub-classes including acenes, rylenes, zethrenes, as well as various compounds that do not belong to any particular family. As shown in Figure 1, the PAH101 set contains molecules ranging in size from 12 atoms in benzene (CSD Reference: BENZEN) to 136 atoms in two pyrene-stabilized acenes 9,11,13,22,24,26-Hexaphenyltetrabenzo[de,rs,wx,k_1l_1]nonacene (CSD Reference: KECLAH), 9,11,13,14,15,16,18,20-Octaphenyldibenzo[de,c_1d_1]heptacene (CSD Reference: TAYSUJ), and a phenylated pentacene 1,2,3,4,6,8,9,10,11,13-Decaphenylpentacene (CSD Reference: VEBJAO). The crystal size in the PAH101 set ranges from 44 atoms in the unit cell for biphenyl (CSD Reference: BIPHEN) to 544 atoms in 1,2,3,4,6,8,9,10,11,13-Decaphenylpentacene (CSD Reference: VEBJAO).

The PAH101 dataset contains GW+BSE results for the electronic and optical properties of molecular crystals, as well as the DFT-level SISSO primary features used in Ref.[98]. We envision this dataset being used for computational discovery of crystalline organic semiconductors and chromophores with desired properties for applications in various organic electronic devices. For example, the dataset contains information on optical gaps and absorption spectra, which could be used to search for chromophores that absorb light in a certain energy range. In addition, the dataset contains singlet and triplet excitation energies, which can be used to evaluate candidate chromophores for triplet-triplet annihilation (TTA) and thermally activated delayed fluorescence (TADF). TTA chromophores can be used for harvesting photons with energies below the absorption threshold of a solar cell by up-conversion of two low-energy triplet excitons into one singlet exciton that can be absorbed.[111,112] TADF chromophores can be used to enhance the efficiency of OLEDs by converting electrically generated non-radiative triplet excitons into radiative singlet excitons.[23,113,114] The dataset also contains GW band structures. The band dispersion is related to transport in crystalline organic semiconductors, which affects the performance of organic electronic devices such as field effect transistors (OFETs).[115] Furthermore, this dataset can be used as a resource for comparing and benchmarking the performance of various electronic structure methods for calculating the electronic and optical properties of molecular crystals. Finally, this dataset can be used to augment other datasets, e.g., DFT datasets for molecular crystals or TDDFT datasets for isolated molecules to train multi-fidelity ML models for predicting various electronic and optical properties of molecular crystals. In summary, because the PAH101 is a unique set of GW+BSE data for molecular crystals, we expect it to be a resource of great usefulness to the computational community.

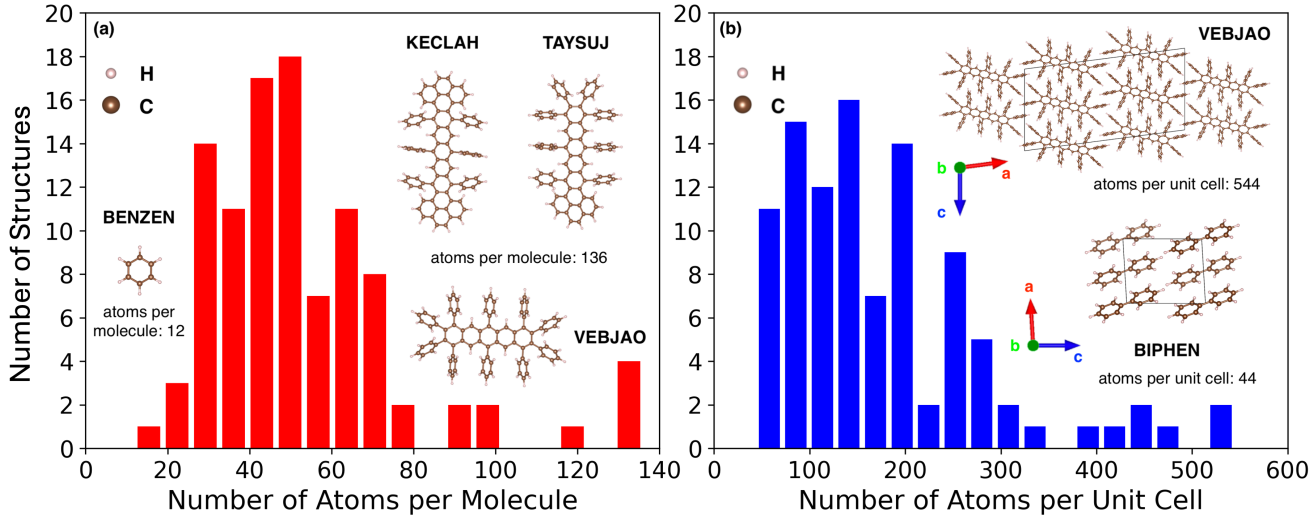

Figure 1. Histograms of the number of atoms (a) in a single molecule and (b) in a crystal unit cell for the materials in the PAH101 set. Also shown are illustrations of the molecular structures of benzene (BENZEN), 9,11,13,22,24,26-Hexaphenyltetrabenzo[de, rs, wx, k_1l_1]nonacene (KECLAH), 9,11,13,14,15,16,18,20-Octaphenyldibenzo[de, c_1d_1]heptacene (TAYSUJ), 1,2,3,4,6,8,9,10,11,13-Decaphenylpentacene (VEBJAO), and the crystal structures of 1,2,3,4,6,8,9,10,11,13-Decaphenylpentacene (VEBJAO) and Biphenyl (BIPHEN).

Methods

Hydrogen Addition

The starting geometries of the 101 molecular crystals were extracted from the Cambridge Structural Database (CSD).[110] The CSD reference codes for each material are available in the data records. Some of the CIF files in the CSD are missing the hydrogen atom positions, which cannot be determined by X-ray diffraction. To provide an approximate position for each missing H atom, we have developed the Hydrogen Append (HAppend) code, available in the GitHub repo: https://github.com/BLABABA/HAppend. HAppend is written in Python and uses RDKit[116] and Pymatgen[117]. The workflow of HAppend is illustrated in Figure 2 using BEANTR as an example. All H atoms were removed from the CIF file for the purpose of demonstration. HAppend does not use the symmetry information provided in the CIF file. In step (1) the unit cell is replicated to build a super-cell so that any molecular fragments inside the unit cell can be completed. In step (2) all the complete molecules and molecular fragments are identified. Subsequently, any broken fragments at the supercell boundary (colored in blue in Figure 2) are removed. In step (3) all the complete molecules are extracted. Only two molecules are shown in Figure 2 for demonstration purposes. Step (4) is identifying the missing hydrogen sites and appending H atoms to each molecule. A detailed schematic of step (4) is shown in the bottom row of Figure 2. In step 4a the missing hydrogen sites are identified by checking the type of hybridization of each carbon atom against the number of valence electrons participating in covalent bonds. In this example, all C atoms in the aromatic rings have sp^2 hybridization. In step 4b H atoms are attached to atoms with unpaired valence electrons. The bond length and angle are determined based on the bonded neighbors and hybridization type. In this example, given that the C atom is sp^2 hybridized, the two H-C-C angles should be about 120°. This process is performed for all atoms in the BEANTR molecule and the completed molecule is obtained after step 4c. Step (5) reconstructs the complete super-cell with appended H atoms. Step (6) reduces the super-cell back to the original unit cell with all the coordinates for the missing H atoms now known. Finally, sanity checks are performed to verify that the structure is correct. The structure is checked against the expected chemical formula (if provided in the CIF file from the CSD). In addition, RDKit is used to repeat step 4b and confirm that the explicit valence matches with the type of hybridization for each atom. If the sanity check fails, the user may have to attach H atoms manually. HAppend is not limited to PAHs and may be used to add missing H atoms to other types of organic molecules.

Structural Relaxation

Full unit cell relaxation was performed with either CASTEP[118] or FHI-aims[119,120] (which code was used is reported in the data records). The Perdew, Burke, and Ernzerhof (PBE) exchange-correlation functional[121] was used with the Tkatchenko-Scheffler (TS) pairwise dispersion method[122]. For relaxations performed with CASTEP, norm-conserving pseudopotentials were utilized

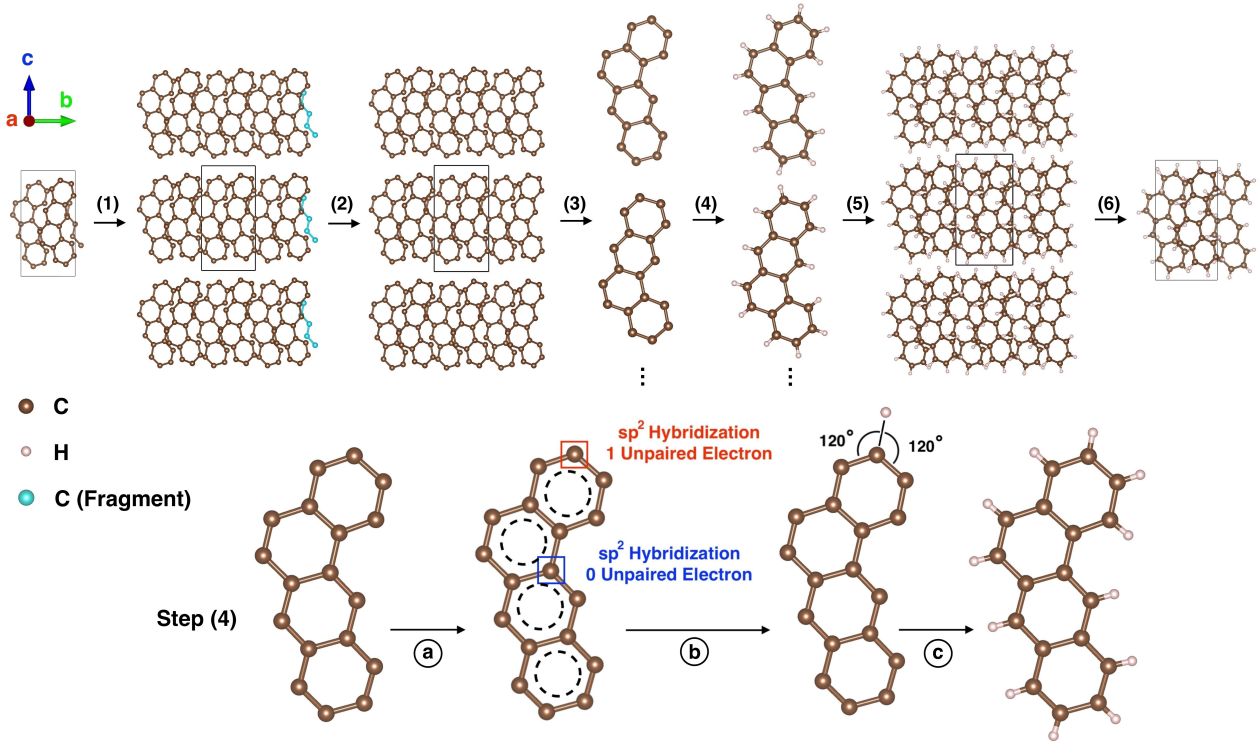

Figure 2. Schematic illustration of the workflow of adding missing H atoms with HAppend, demonstrated for the BEANTR crystal. The top row shows the steps of (1) super-cell construction, (2) removal of broken molecular fragments (colored in blue), (3) extraction of molecules, (4) addition of H atoms to all molecules, (5) reconstruction of the supercell with the H atoms attached to all molecules, and (6) reduction of the supercell to a single unit cell. For Steps (3) and (4) only two molecules are shown for clarity. The bottom row presents a detailed view of the hydrogen addition step: (a) identification of missing H sites, (b) calculation of approximate H atom positions, and (c) attachment of H atoms to the molecule.

for carbon and hydrogen. The plane-wave basis set cutoff was 750 eV. A Monkhorst-Pack k-grid with a spacing of 0.07 $Å^{-1}$ was adopted. The convergence thresholds for total energy, maximum force, maximum stress, and maximum displacement were 5×10^{-6} eV/atom, 0.01 eV/Å, 0.02 GPa, and 5×10^{-4} $Å^{-1}$, respectively. For structures relaxed with FHI-aims, the tight numerical settings and tier-2 basis sets were used. The fully relaxed crystal structures and the molecular geometries extracted from them are provided in the data records. The GW+BSE calculations were performed for the fully relaxed crystal structures.

DFT Features

The data records include the DFT primary features used for SISSO in Ref.[98]. The DFT features of molecules and crystals were calculated using FHI-aims[119,120]. From considerations of computational efficiency, the DFT primary features were calculated with locally-optimized geometries. The crystal structures were relaxed with the lattice vectors fixed at the experimental values and the single molecule properties were calculated using molecules extracted from these locally-optimized crystal structures. All primary features were calculated with the PBE functional.[121] using the tight numerical setting and tier-2 basis sets of FHI-aims.[119]

Mean-Field Wave Function Calculation

The Quantum ESPRESSO package[123] was used to compute the DFT eigenvectors and eigenvalues, which served as the starting point for non-self-consistent *GW*+BSE calculations, using the PBE functional. Norm-conserving pseudopotentials were chosen in order to take advantage of the simplification of matrix elements in *GW*+BSE calculations.[124] Troullier–Martins norm-conserving pseudopotentials were generated using FHI98PP-converted with fhi2upf.x v.5.0.2 from Abinit Project. The kinetic energy cutoff was 50 Ry. The k-point grids used for each material are reported in the data records.

GW+BSE Calculations

The BerkeleyGW package[124] was used to perform *GW*+BSE calculations. From considerations of computational cost, non-self-consistent G_0W_0 calculation were performed. 550 unoccupied states were included in the dielectric function and the

self-energy operator evaluations. The static remainder correction was applied to accelerate the convergence. The screened and bare Coulomb interaction cutoffs were 10 Ry and 40, respectively. The Bethe-Salpeter equation was solved within the Tamm-Dancoff approximation (TDA) with 24 valence bands and 24 conduction bands included. The fine k-point grid wave-functions were generated using a fine k-point grid twice as dense as the coarse k-point grid. The coarse and fine k-point grid settings for each material are reported in the data records.

Data Records

The PAH101 dataset is available via the NOvel MAterials Discovery (NOMAD) repository[125] and can be accessed at DOI: 10.17172/NOMAD/2024.12.05-1. The data are provided in YAML (.yaml) format. Each file is named as *CSD-REFERENCE*.archive.yaml, where *CSD-REFERENCE* is the CSD reference code for each structure. The data structure for each material record is described in Table 1. The top level sections are *struct_id*, *geometry*, *dft*, and *gwbse*. The *struct_id* section contains the CSD reference code. The *geometry* section provides the fully relaxed crystal structure and the single molecule geometry extracted from it. The *dft* section contains all the SISSO primary features used in Ref.[98]. The *gwbse* section provides quasi-particle (QP) and excitonic properties for the PAH101 crystals, including the fundamental gap, quasiparticle band structure, the static dielectric constant, the first singlet exciton energy (optical gap), the first triplet exciton energy, the full dielectric function, and optical absorption spectra for light polarized along the three lattice vectors. The GW static dielectric constant is not available for some of the materials in the dataset because some data that was not needed for Ref.[98] was not preserved.

Technical Validation

Crystal Structures

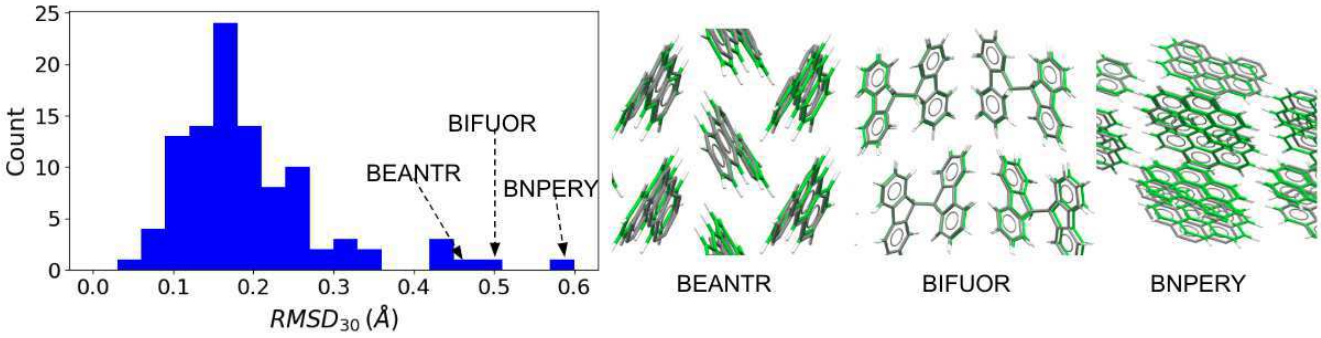

Figure 3. Histogram of the $RMSD_{30}$ of crystal structures relaxed with PBE+TS compared to the experimental structures from the CSD for the PAH101 set. The similarity overlay plots generated by Mercury are shown for BNPERY, BIFUOR, and BEANTR with the experimental structures colored in gray,and the relaxed structures colored in green.

To verify the results of full unit cell relaxation with PBE+TS, the root-mean square distance (RMSD) between the relaxed structures and the experimental structures was calculated. We used the COMPACK[129] molecular overlay method, implemented as the Crystal Packing Similarity tool, in Mercury 2023.2.0[130]. COMPACK overlays clusters of molecules taken from each crystal, within given distance and angle tolerances, and minimizes the RMSD between atoms, typically excluding hydrogen. The output of COMPACK is the number of molecules that could be overlaid and the RMSD. COMPACK comparisons were performed with a cluster of 30 molecules and distance and angle tolerances of 35% and 35°. H atoms were not included. These were the settings used for structure comparison in the 7^{th} crystal structure prediction blind test[131,132]. Figure 3 shows a histogram of the RMSD obtained for the PAH101 set. For the majority of the structures in the dataset the RMSD is below 0.3 Å. The three structures with the largest RMSDs are BNPERY, BIFUOR, and BEANTR. All three are monoclinic structures with larger than average deviations in their b lattice parameter and β angle. For instance, the relaxed b parameter of BEANTR is 6.00 Å, compared to 6.50 Å in the experimental structure, a deviation of 7.7 %. The relaxed structure of BNPERY has a β angle of 92.2°, compared to 98.5°in the experimental structure. Some differences between structures relaxed by DFT at 0K and structures experimentally characterized at room temperature are to be expected.[133] Overall, the performance of PBE+TS is within the community accepted standards of agreement with experiment, as established in the crystal structure prediction blind tests.[131,132] It is possible that performing relaxations with the more accurate many-body dispersion (MBD)[128] method would reduce the RMSD.

struct_id	the CSD reference code for this structure		
geometry	*relaxed_crystal*	the DFT-relaxed crystal structure saved in Pymatgen Structure format	
	molecule	the single molecule geometry extracted from *relaxed_crystal*, saved in Pymatgen Molecule format	
	chemical_formula	chemical formula of the single molecule	
	relax_code	code used to perform crystal structure relaxation	
dft	*gap_s*	the single molecule gap, calculated based on the energy difference between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO)	
	Et_s	the single molecule triplet formation energy, calculated based on the total energy difference between the ground-state and triplet-state molecule	
	DF_s	the single molecule DFT estimate for the SF driving force, calculated by taking the difference between *gap_s* and twice *Et_s*	
	IP_s	the single molecule ionization potential (IP), calculated based on the total energy difference between a cation and neutral molecule	
	EA_s	the single molecule electron affinity (EA), calculated based on the total energy difference between an anion and neutral molecule	
	bandgap	the crystal band gap	
	Et	the crystal triplet formation energy, calculated based on the total energy difference between the ground-state and triplet-state crystal	
	DF	the crystal DFT estimate for the SF driving force, calculated by taking the difference between *bandgap* and twice *Et*	
	VBdisp	the valence band dispersion, *i.e.*, the energy range of the HOMO-derived band	
	CBdisp	the conduction band dispersion, *i.e.*, the energy range of the LUMO-derived band	
	h_{ab}	the transfer integral, calculated with fragment orbital DFT[126]	
	polarization	the trace of the polarization tensor for a single molecule, calculated with DFT using the PBE functional and the range-separated self-consistently screened version of many-body dispersion (MBD@rsSCS) method[127,128]	
	epsilon_mbd	the dielectric constant calculated with PBE+MBD@rsSCS	
	weight_s	the molecular weight in atomic mass units (amu)	
	density	the crystal density in amu Å^{-3}	
	eigenvalues	the eigenvalues for the single molecules, data stored as $n \times 4$ matrix, whose columns are: State, Occupation, Eigenvalue [Ha], Eigenvalue [eV]	
	kgrid	the k-grid settings for the calculation of crystal primary features	
gwbse	*absorption*	*a*	Optical absorption spectrum for light polarized along the a, b, and c crystal axes. Each absorption data record contains four columns: energy (eV), the imaginary and real parts of the dielectric function ε_2 and ε_1, and the normalized joint density of states.
		b	
		c	
	bandstructure	*kpoints*	the high-symmetry k-point path used to calculate the *GW* band structure
		val	the values of band structure, saved as $n \times 8$ matrix, whose columns are: spin, band index, k-point coordinate x, k-point coordinate y, k-point coordinate z, mean-field energy, quasi-particle energy, difference between mean-field and quasi-particle energy
	bse_Es	the singlet exciton energy (optical gap) calculated with BSE	
	bse_Et	the triplet exciton energy calculated with BSE	
	bse_DF	the SF driving force for a crystal, $bse\_Es - 2 \times bse\_Et$	
	kgrid_coarse	the k-grid used for coarse grid wave-function calculation	
	kgrid_fine	the k-grid used for fine grid wave-function calculation	
	fundamental_gap	the fundamental gap calculated with *GW*	
	bse_Es_bind	the singlet-state exciton binding energy	
	bse_Et_bind	the triplet-state exciton binding energy	
	epsilon_gw	the dielectric constant calculated with *GW*; N.A. entered if not available	

Table 1. Data records: Description of the data structure of the PAH101 set with explanations of all entries.

GW+BSE Convergence

The results of GW+BSE calculations with the BerkeleyGW code are sensitive to the convergence of several parameters[49,134,135]. Because of the large number of calculations performed for the PAH101 set, we have chosen parameters that provide a balance between accuracy and computational cost. The convergence of the settings used for the PAH101 dataset has been demonstrated previously for selected systems[93,98]. Figure 4 shows the convergence with respect to coarse k-point grid and the number of empty bands used in the *GW* step for representative materials. The number of k-points is inversely proportional to the unit cell

size. Benzene has the smallest unit cell in the PAH101 set and therefore requires a relatively large number of k-points. 9,9'-bifluorenyl (CSD reference code BIFUOR) represents a system of intermediate size. For both materials, increasing the number of k-points beyond the chosen settings leads to a change of less than 0.001 eV in the GW band gap. For the representative materials fluoranthene (CSD reference code FLUANT02) and 6-phenylpentacene (CSD reference code VEBKAP), increasing the number of empty bands beyond 550 leads to a change of less than 0.02 eV in the GW band gap. For fluoranthene, increasing the number of valence and conduction bands used for the BSE step beyond 24 leads to a change of less than 0.06 eV in the optical gap. The settings used for the PAH101 set are sufficiently robust for "production" calculations. Notably, in the time that passed since the PAH101 set was generated, there have been advances in streamlining the convergence of MBPT calculations.[136–139] These have focused primarily on inorganic crystals with a few atoms in the unit cell. A workflow that converges the settings for each system individually would be too expensive for systems of the size of the PAH101 set. If a certain material is of particular interest, then more detailed calculations may be pursued with ultra-converged settings and/or more accurate methods than G_0W_0@PBE.

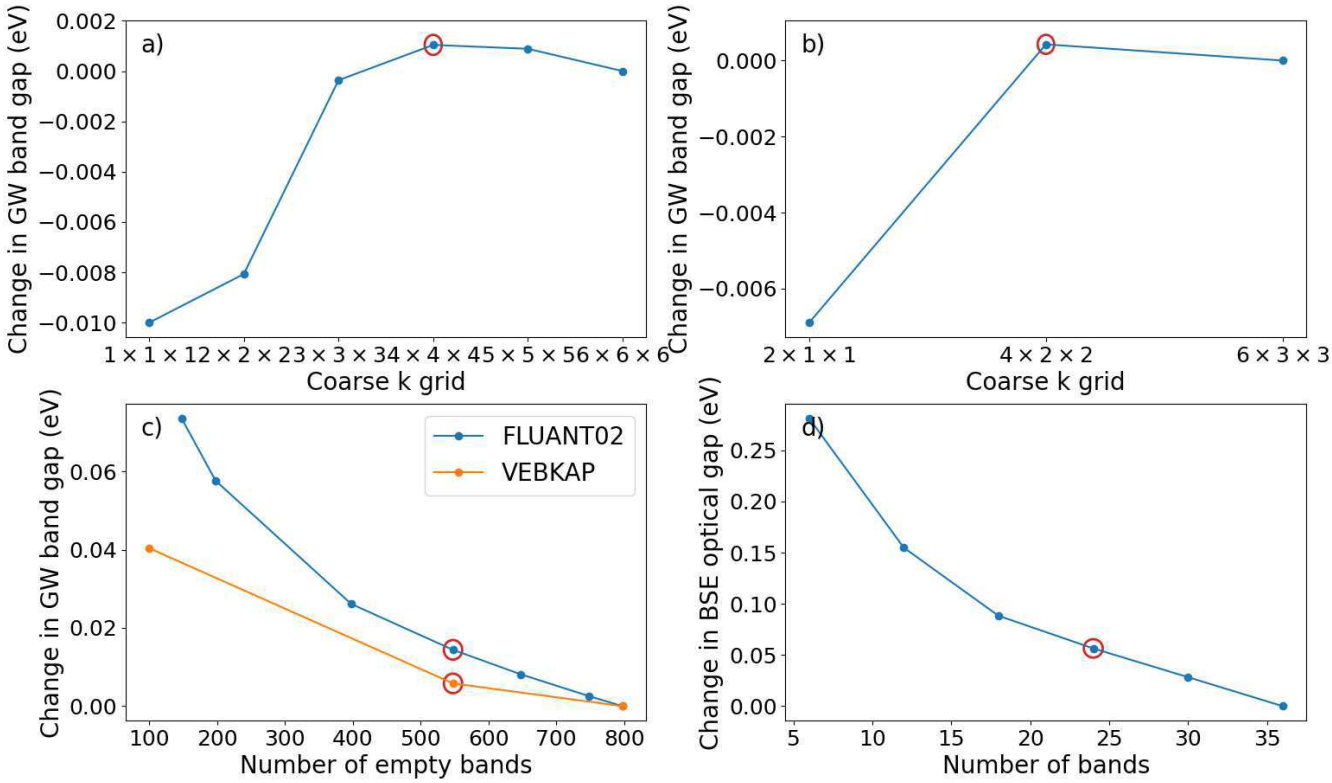

Figure 4. Convergence of *GW*+BSE calculations for representative materials. Change in the *GW* band gap as a function of the coarse *k*-point grid for (a) benzene and (b) 9,9'-bifluorenyl (CSD reference code BIFUOR). (c) Change in the *GW* band gap as a function of the number of empty bands for fluoranthene (CSD reference code FLUANT02) and 6-phenylpentacene (CSD reference code VEBKAP). (d) Change in the BSE optical gap as a function of the number of fine bands for fluoranthene. The chosen settings are circled in red.

Optical Absorption

The GW+BSE approach has been benchmarked extensively for isolated molecules, for which high-level quantum chemistry reference data can be calculated.[49,140–144] For molecular crystals no benchmark studies are available, owing to the difficulty of obtaining reference data for large systems with periodic boundary conditions. Therefore, we are only able to validate the results of GW+BSE by comparison to experiments. Table 2 shows a comparison of the GW+BSE optical gaps (singlet exciton energies) to experimental values and GW+BSE values reported by others, where available. The GW+BSE values reported here are within 0.2 eV or less of the values reported by others. The results of GW+BSE calculations can differ because of differences in the implementation and convergence settings, as discussed extensively in Ref.[49]. Because the absorption edge is not abrupt, Tauc plots are typically used to extract the optical gap from absorption spectra.[145–148] This can lead to some uncertainty in the experimental values. Here, if multiple experimental values are found for the same material, they are within

0.1 eV or less of each other in most cases. For the entries marked with *, we used the Tauc method to extract the optical gap from the experimental data because no value for the optical gap was reported in the paper. For the entries marked with **, there is a larger uncertainty in the optical gaps extracted using the Tauc method because the absorption edge does not decay to zero. In most cases, the GW+BSE optical gaps are within 0.2 eV or less from experimental values.

Table 2. Optical gaps obtained using *GW*+BSE (E_g^{GW+BSE}) compared with experimental values (E_g^{Exp}) and *GW*+BSE values reported by others (E_g^{GW+BSE} in Ref), where available. Entries marked with * were extracted by us from absorption spectra using the Tauc method. Entries marked with ** have an absorption spectrum that is non-zero in the low-energy region, leading to a larger uncertainty in the optical gaps extracted using the Tauc method.

CSD Ref. Code	Compound Name	E_g^{GW+BSE} (eV)	E_g^{Exp} (eV)	E_g^{GW+BSE} in Ref (eV)
BENZEN	Benzene	4.83	4.69-4.8[149]	5.0[150]
ANTCEN	Anthracene	3.22	3.16[151]	3.3[150]
TETCEN01	Tetracene	2.24	2.38[152,153]	2.4[150]
PENCEN	Pentacene	1.72	1.8-1.85[154–156]	1.7-1.8[150,157,158]
ZZZDKE01	Hexacene	1.17	1.37*-1.4[159–161]	1.0[150]
QQQCIG04	Rubrene (Orthorhombic)	2.28	2.32[162]	
QQQCIG13	Rubrene (Monoclinic)	2.62	2.36[163]	
QQQCIG14	Rubrene (Triclinic)	2.30	2.31[163]	
PERLEN05	Perylene (SHB)	2.61	2.58*[164,165]	
PERLEN07	Perylene (HB)	2.45	2.49*[164,165]	
POBPIG	Diindeno[1,2,3-cd:1',2',3'-lm]perylene	2.21	2.25[166]	
QUATER10	Quaterrylene	1.33	1.48-1.60[167–169]	
CORONE01	Coronene	2.96	2.9-2.92*[170,171]	
HBZCOR	Hexabenzo(bc,ef,hi,kl,no,qr)coronene	2.70	2.80[172,173]	
BEANTR	1,2-Benzanthracene	3.27	3.14[174]	
BIPHEN	Biphenyl	3.41	4.1-4.18[175–178]	
CRYSEN01	Chrysene	3.66	3.6**[179]	
TERPHE02	p-Terphenyl	4.17	3.9**[179]	
BNPERY	1,12-Benzoperylene	2.80	2.4-2.5*[180]	
KUBVUY	10,10'-Diphenyl-9,9'-bianthryl	3.23	2.9*[181]	
KUBWAF01	9,9'-Bianthracenyl	3.05	2.7-2.8*[182]	

The GW+BSE absorption spectra are validated by comparison to thin film experimental data for representative materials.[172,179,183] For an anisotropic crystal the absorbance depends on the polarization direction of the incident light. Most absorption experiments are performed on polycrystalline samples and even in experiments performed on single crystals the crystallographic orientation of the sample with respect to the polarization of the incident light is often unknown. This introduces some uncertainties in the comparison with experiments. We calculate the absorbance for light polarized along the *a*, *b*, and *c* lattice vectors and normalize the maximum of the total absorbance to one. The results are shown in Figure 5. For 1,2-benzanthracene (BEANTR), coronene (CORONE01), and hexabenzo(bc,ef,hi,kl,no,qr)coronene (HBZCOR) the agreement of the GW+BSE spectra with experiment is very good. For chrysene (CRYSEN01), p-terphenyl (TERPHE02), and triphenylene (TRIPHE12) the agreement is more qualitative.

In addition to the unknown direction of the polarization with respect to the crystal axes, there are other factors, both on the experimental side and on the theoretical side that can contribute to discrepancies. In ref.[179] the crystal structure of the films is not reported. The crystal structures used in our calculations are the common forms of p-terphenyl and chrysene, but both materials have other polymorphs reported in the CSD (for triphenylene all CSD entries appear to be the same structure but we cannot rule out the appearance of a different thin film polymorph). In polycrystals there can be contributions from grain boundaries (in samples comprising very small crystallites, which is not the case here, there can be surface contributions as well). Furthermore, we do not consider vibrational contributions in our simulations. Sources of errors in GW+BSE calculations include the DFT exchange-correlation functional used for the mean-field starting point, numerical convergence of various settings (k-point grids, number of empty states used in the GW step, the number of bands used in the BSE step), the non-self-consistency in the GW step, the plasmon pole approximation used in the GW step,[49] the Tamm–Dancoff approximation used in the BSE step,[25,150,184–188] and the static approximation for *W* used in the BSE step.[184,189] See also Ref.[190] for additional discussion. The significance of different sources of errors can be material dependent. In the future, it would be desirable to rigorously

assess the contributions of different sources of errors in GW+BSE by comparison to high-level theories or well-controlled experiments (performed on single crystals with well-defined polarization) for a diverse benchmark set of molecular crystals.

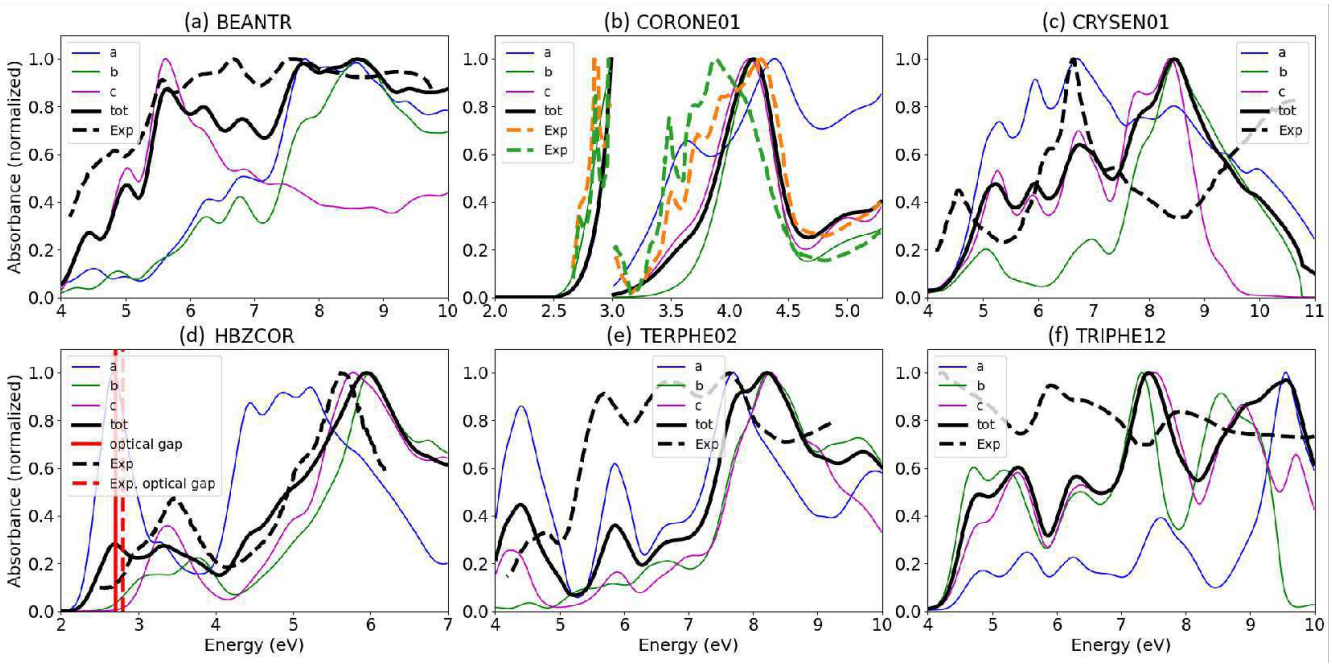

Figure 5. Absorption spectra calculated using *GW*+BSE compared with thin film experiments[172,179,183] for (a) 1,2-benzanthracene (BEANTR), (b) coronene (CORONE01) with the region around the absorption edge magnified for clarity, (c) chrysene (CRYSEN01), (d) hexabenzo(bc,ef,hi,kl,no,qr)coronene (HBZCOR), (e) p-terphenyl (TERPHE02), and (f) triphenylene (TRIPHE12).

Usage Notes

The PAH101 set is the currently the largest trove available of *GW*+BSE data for molecular crystals. As such, it offers unique opportunities to (i) learn about correlations between DFT and *GW*+BSE values of various properties, (ii) discover materials with desired electronic/ optical properties in the dataset itself, and (iii) train machine learning models to help in materials discovery efforts. Here, we provide examples for these three use cases.

Reliability of DFT Models

In materials discovery workflows it is desirable to use models that are fast to evaluate for preliminary screening of a large number of candidates. Semi-local DFT has been used extensively for this purpose. However, such models must be sufficiently reliable to at least capture the correct trends. Here, we perform statistical analysis across our dataset to examine whether selected DFT models are sufficiently predictive of *GW*+BSE quantities. The PAH101 dataset may similarly serve as a resource for researchers interested in comparing the results of other DFT and TDDFT models to *GW*+BSE.

Figure 6 shows correlation plots between selected properties calculated by DFT with the PBE functional and *GW*+BSE@PBE. In Panel (a) single-molecule and crystal DFT quantities are compared to the GW+BSE crystal optical gap. The fundamental gap of a molecule corresponds to the difference between the ionization potential (IP) and electron affinity (EA). The fundamental gap of a molecular crystal (calculated by *GW*) is typically significantly narrower than the single molecule fundamental gap because of screening and band dispersion in the crystal.[97] The optical gap of a molecular crystal is narrower than the fundamental gap because of the exciton binding energy.[191] The IP and EA calculated based on on DFT total energy differences are better estimates than the Kohn-Sham eigenvalues of the HOMO and LUMO. However, it has been shown that the molecular fundamental gaps obtained from PBE IP-EA have errors of 0.89 eV on average compared to reference data.[192] As expected, the molecular PBE IP-EA values significantly overestimate the *GW*+BSE optical gaps of the corresponding molecular crystal. Although there is correlation with the overall trend of the *GW*+BSE optical gaps, the spread of the PBE IP-EA values is too large to be considered as a reliable predictor.

It is well known that molecular HOMO-LUMO gaps and crystal band gaps are significantly underestimated by (semi-)local functionals such as PBE, owing to the self-interaction error (SIE).[193] For the PAH101 set, both the PBE single molecule

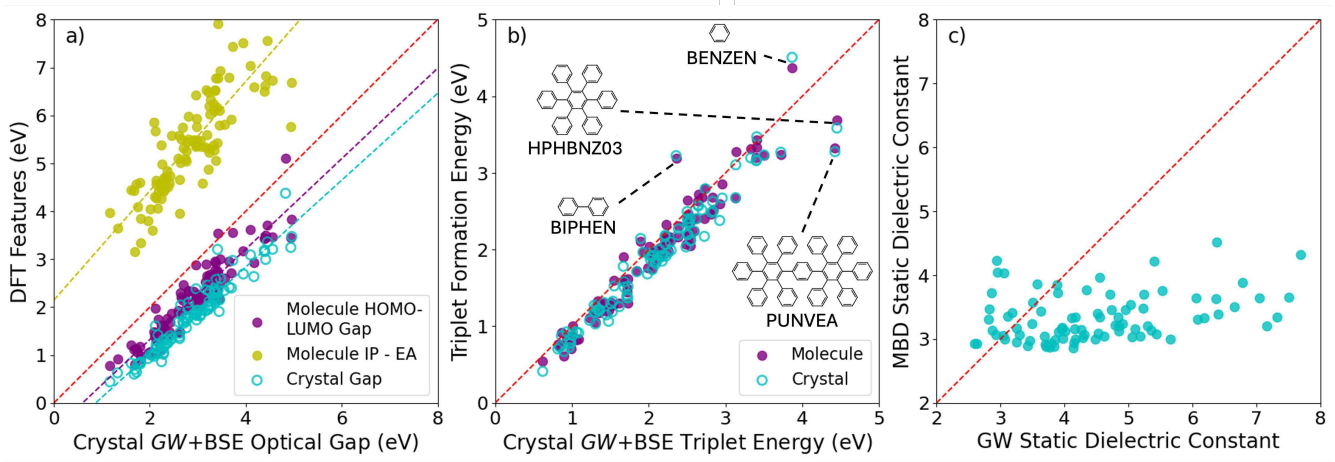

Figure 6. Correlations between DFT and *GW*+BSE across the PAH101 set for selected properties: (a) DFT molecular IP-EA, molecular HOMO-LUMO gaps, and crystal band gaps compared to *GW*+BSE optical gaps. (b) DFT triplet formation energy of the molecule and crystal compared to the *GW*+BSE triplet exciton energy. Molecular structures of some outliers are also shown. (c) DFT dielectric constant calculated by using the MBD polarizability in the Clausius-Mossotti equation compared with the *GW* static dielectric constant.

HOMO-LUMO gap and crystal gap systematically underestimate but correlate well with the *GW*+PBE optical gaps. Based on this, these values may be sufficiently reliable for rough preliminary screening based on relative trends among materials. The single molecule PBE HOMO-LUMO gap is particularly attractive for this purpose because it is very fast to evaluate. Furthermore, there are large datasets of single molecule[46] and crystal PBE gap[47] that can be mined. We note, however, the effect of SIE is material-dependent.[143,194] Compounds whose HOMO and/or LUMO are highly localized may be affected more severely than PAHs, whose frontier molecular orbitals are typically delocalized over the aromatic system. Therefore, it would be prudent to reevaluate the reliability of DFT-PBE molecular and crystal gaps for more diverse data sets.

In Panel (b) the single molecule and crystal DFT triplet formation energies are compared to the *GW*+BSE triplet excitation energies. Overall, the single molecule and crystal DFT values are quite close to each other and to the *GW*+BSE triplet exciton energies, with MAEs of 0.20 eV and 0.23 eV, respectively and R^2 values of 0.89 and 0.86, respectively. The reasons for this agreement need to be investigated further (we are not aware of any benchmark studies of DFT triplet formation energies). The four most significant outliers, whose molecular structures are shown, are: biphenyl (BIPHEN), benzene (BENZEN), 2',2''',3',3''',5',5''',6',6'''-octaphenyl-p–quinquephenyl (PUNVEA), and hexaphenylbenzene (HPHBNZ03). These compounds are characterized by phenyl rings connected by single C-C bonds, whereas the majority of compounds in the PAH101 set are characterized by extended aromatic systems. Our results indicate that DFT triplet formation energies are fairly reliable as lower-cost descriptors for preliminary screening. However, based on the nature of the outliers, it would be prudent to validate these findings for more diverse materials.

Panel (c) shows a comparison between the static dielectric constant calculated by PBE+MBD and by *GW*. The *GW* value corresponds to the dielectric function value at 0 frequency and 0 wave-vector, $\varepsilon(\omega = 0, q = 0)$. The DFT value is obtained by using the MBD polarizability in the Clausius-Mossotti relation, as described in Ref.[97]. The comparison reveals that the DFT values are narrowly distributed around 3 and, in general, do not correlate with the *GW* values. For some materials the values obtained from PBE+MBD may fortuitously agree with experimental and/or *GW* values;[195] however, even with the self-consistent screening approach used in the MBD method,[127,128] DFT does not capture the many-body physics contained in the *GW* dielectric function. This demonstrates that it is important to consider larger sets of materials to assess the reliability of methods.

Materials Discovery

The electronic and optical properties of most of the materials in the PAH101 set have not been thoroughly investigated experimentally. Some of the quantities calculated here, such as triplet excitation energies, are difficult to probe experimentally and require highly specialized techniques and facilities. Therefore, although the PAH101 set is relatively small, it is possible that some useful materials would be found in it. Here, we provide examples for some of the electronic and optical properties relevant for organic electronic devices that can be extracted from the dataset. The dataset can be searched for materials with a

particular property or combination of properties. As demonstrated below, the dataset may provide insights on structure-property relations and expose gaps in our understanding of the properties of molecular crystals that call for further investigation.

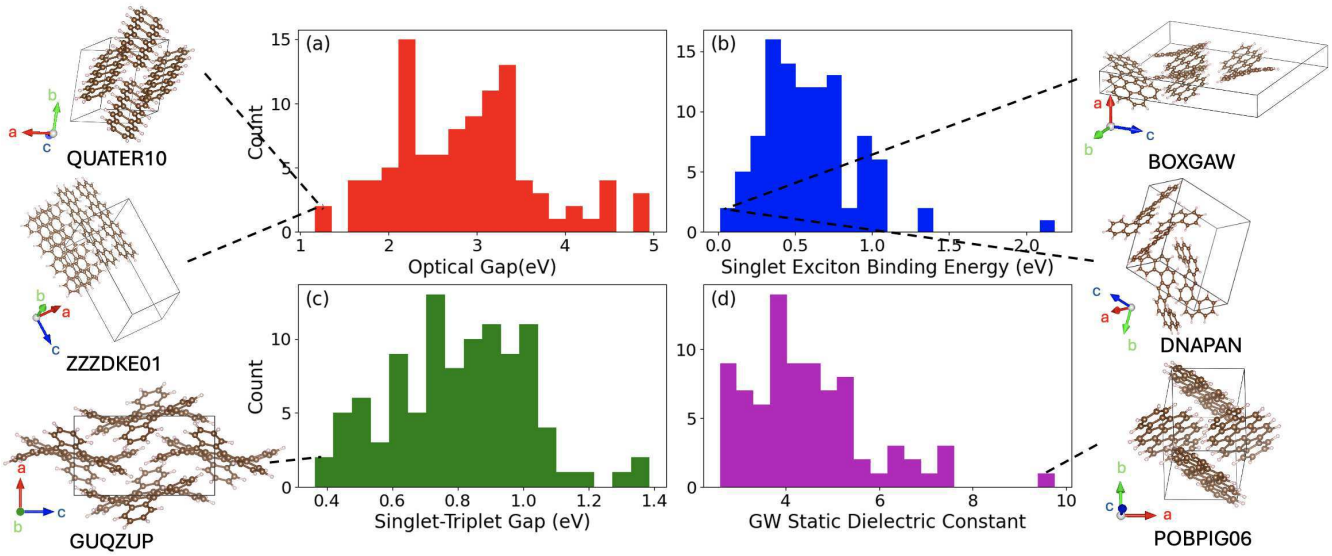

Figure 7. Distributions of (a) the singlet exciton eneries, which correspond to the optical gaps, (b) the singlet exciton binding energies, (c) the singlet-triplet gaps, and (d) the *GW* static dielectric constant across the PAH101 dataset. Some crystal structures are also shown.

One of the key properties for device applications is the optical gap, whose distribution in the dataset is shown in Figure 7a. The PAH101 set contains materials with a wide range of optical gaps. Crystalline quaterrylene (QUATER10) and hexacene (ZZZDKE01) have the smallest optical gaps of 1.33 eV and 1.17 eV, respectively. If a material is sought with an optical gap of up to about 5 eV it may be found in the dataset. Absorption spectra for light polarized along the three crystal axes are also provided in the dataset (see Table 1), such that materials can be sought with broad absorption and/or absorption peaks in certain energy ranges.

The singlet exciton binding energy, whose distribution is shown in Figure 7b, corresponds to the difference between the *GW* fundamental gap and the optical gap. This is the energy required to split photogenerated excitons into free charge carriers in organic solar cells. In most organic materials the exciton binding energy is significant compared to inorganic materials because the dielectric screening of charges is not as strong. However, some materials in the PAH101 set have low exciton binding energies (in parentheses), including: anthra(2,1,9,8-hijkl)benzo(de)naphtho(2,1,8,7-stuv)pentacene (BOXGAW; 0.013 eV), dinaphtho(1,2-a:1',2'-h)anthracene (DNAPAN; 0.071 eV), tetrabenzo(de,no,st,c1d1)heptacene (TBZHCE; 0.130 eV), benzo[lm]chryseno[1,12,11,10-opqrab]perylene (YUNYAJ; 0.165 eV), and hexabenzo(bc,ef,hi,kl,no,qr)coronene (HBZCOR; 0.169 eV). All of these compounds are characterized by very extended and/or elongated π systems, which likely lead to an already low molecular exciton binding energy (not calculated here), further reduced by dielectric screening in the solid form. Triplet exciton binding energies are also provided in the dataset (see Table 1). They are typically significantly higher than singlet exciton binding energies.

Another property of interest for device applications is the singlet-triplet gap, *i.e.*, the energy difference between the lowest singlet excited state and the lowest triplet excited state, both of which are included in the PAH101 dataset (see Table 1). The singlet-triplet gap is a key property for organic light emitting diodes (OLEDs). Most of the electrically generated excitons in OLEDs are triplet excitons, which cannot decay radiatively to the ground state. In thermally activated delayed fluorescence (TADF) chromophores, a small singlet-triplet gap enables reverse intersystem crossing (RISC) from the lowest triplet excited state to the lowest singlet excited state, which subsequently decays to the ground state, emitting a photon.[196–198] Figure 7c shows the distribution of singlet-triplet gaps in the PAH101 dataset. Small singlet-triplet gaps are rare among this class of materials. The materials with lowest singlet-triplet gaps (in parentheses) are: trinaphtho[1,2,3,4-fgh:1',2',3',4'-pqr:1",2",3",4"-za_1_b_1_]trinaphthylene (GUQZUP; 0.36 eV), 9,18-diphenyltetrabenz(a,c,h,j)anthracene (FACPEE; 0.38 eV), acenaphtho[3,2,1,8-fghij]tetrabenzo[a,c,m,o]picene (VUFHUA; 0.435 eV), benzo(1,2,3-bc:4,5,6-b',c')dicoronene (YOFCUR; 0.44 eV), and 2-(naphthalen-2-yl)azulene (PUJQIV; 0.45 eV). Even the lowest singlet-triplet gaps in the PAH101 set would be considered marginal or too high for TADF. However, examining these materials may reveal new classes of chromophores that could be interesting for further investigation and fine-tuning by chemical modification. Charge transfer (CT) excitations

between spatially separated HOMO and LUMO states are considered key to achieving small singlet-triplet gaps in TADF chromophores.[113,197] With the exception of PUJQIV, the materials with the smallest singlet-triplet gaps in the PAH101 set bear no resemblance to the donor-acceptor compounds typically used for TADF. Rather, they are large PAHs with extended π systems. FACPEE, VUFHUA, and YOFCUR have segments that could lead to CT-like intramolecular excitations. GUQZUP (shown in Figure 7c) can be described as a graphene flake with no obvious segments. The twisted conformation it adopts in the crystal structure may contribute to orbital localization and CT-like excitations. The effect of crystal packing and intermolecular vs. intramolecular CT excitations on singlet-triplet gaps is also not well-understood and should be further investigated in relation to TADF in crystalline materials.[199,200]

Figure 7d shows the distribution of the *GW* static dielectric constant in the PAH101 dataset. There is a prevalent perception in the organic electronic community that all organic solids have a similar dielectric constant of about 3 (we have not been able to trace the origin of this perception to a particular paper). The DFT values obtained for the PAH101 set (see Figure 6c) may confirm this perception, but the *GW* values tell a different story. Several materials in the dataset have *GW* static dielectric constants (in parentheses) that are significantly higher than 3, including: diindeno[1,2,3-cd:1',2',3'-lm]perylene (POBPIG06; 9.75), benzo[lm]chryseno[1,12,11,10-opqrab]perylene (YUNYAJ; 7.51), hexacene (ZZZDKE01; 7.41), indeno(7,7a,1,2,3-lmno)-1,12-ethenochrysene (SURTAA; 7.33), and tetrabenzo[a,d,j,m]coronene (SETTES; 7.05). These are compounds with extended and/or elongated π systems, which are probably highly polarizable (the molecular polarizability is not calculated here). The crystal packing probably also contributes significantly to the dielectric screening. Most of the research on organic materials with high dielectric constants has been on polymers for applications in bulk heterojunction organic solar cells (*e.g.*,[201]), which are very different for the materials in the PAH101 set. This calls for further investigation of the dielectric behavior of molecular crystals. We note that the full dielectric function, which contains information on the frequency dependence and anisotropy, is available in the dataset (in the absorption entry, see Table 1).

Machine Learning

To demonstrate how the PAH101 dataset can be reused to train ML models for other purposes than SF, we use SISSO to find predictive models for the *GW* fundamental band gap, whose values are also provided in the dataset. The dataset can be used in a similar manner to train ML models other than SISSO to predict any of the quantities included in the dataset. In addition, it can be used to supplement larger lower-fidelity datasets to train multi-fidelity models.

SISSO models were trained following the same procedure used in Ref.[98]. The same primary features were used (also provided in the PAH101 dataset), with the exception of *DF_s* and *DF_c*, because the DFT estimate for the SF driving force is not a physically meaningful descriptor in relation to the fundamental band gap. The same 10 structures as in Ref.[98] were withheld as an unseen test set and the remaining 91 structures were used for model training. Features were constructed with a maximum rung (the number of times primary features are combined) of 3 and a maximum dimension (Dim) of 4. Features were combined using the operator set $H = \{+, -, \times, \div, \exp, \log, ()^{-1}, ()^2, ()^3, \sqrt{}, \sqrt[3]{}, |\cdot|\}$. The maximum complexity, i.e., the maximum number of operators in one combined feature, was set to 10. A total of 5×10^2, 4×10^5, and 6×10^{10} features were generated by SISSO with a rung of 1, 2, and 3, respectively.

After feature generation, SISSO performs linear regression to yield the model prediction, where each model is the scalar product of the SISSO-generated feature with a vector of fitted coefficients. Then, the models are ranked according to their prediction performance. Sure independence screening (SIS) is used to select optimal subspaces from the huge feature space. The number of features saved after SIS was set to 20. SISSO then uses ℓ_0-norm minimization as a sparsifying operator (SO) to determine the sparse solution for each such subspace. For each combination of dimension and rung, 40 rounds of leave-10-out cross validation (LCV) were performed. In each round, 10 data points (out of the 91 points used for model training) were randomly selected and held out as an unseen validation set. The model with the lowest RMSE for the validation set was selected in each round. Finally, the model with the lowest root mean square error (RMSE) for the combined LCV training and validation data was selected out of the 40 models. This model is denoted as $M_{\mathrm{Dim,Rung}}$. A full account of the SISSO models is provided in the SI.

The computational cost of SISSO-generated models varies depending on the number and type of primary features they contain. The cost of each model was evaluated by summing over the costs of all the primary features included in it. The cost of features that appear in the model more than once was counted only once. The computer time required to calculate the single molecule PBE gap, Gap^S, was assigned a value of 1 cost unit and the cost of other features is tabulated in the SI as multiples of that unit. The cost of all the primary features has been updated from the values given in Ref.[98] to account for new developments in the latest version of FHI-aims. In particular, the MBD calculation has become significantly more efficient than in older versions of the code. The cost was averaged over the 10 structures in the validation set, rather than picking one system of average size, as in Ref.[98]. Figure 8a shows a Pareto chart of the accuracy vs. the computational cost of the SISSO models considered here. The "train" RMSE is calculated for the training set of 91 structures. The "test" RMSE is calculated for the 10

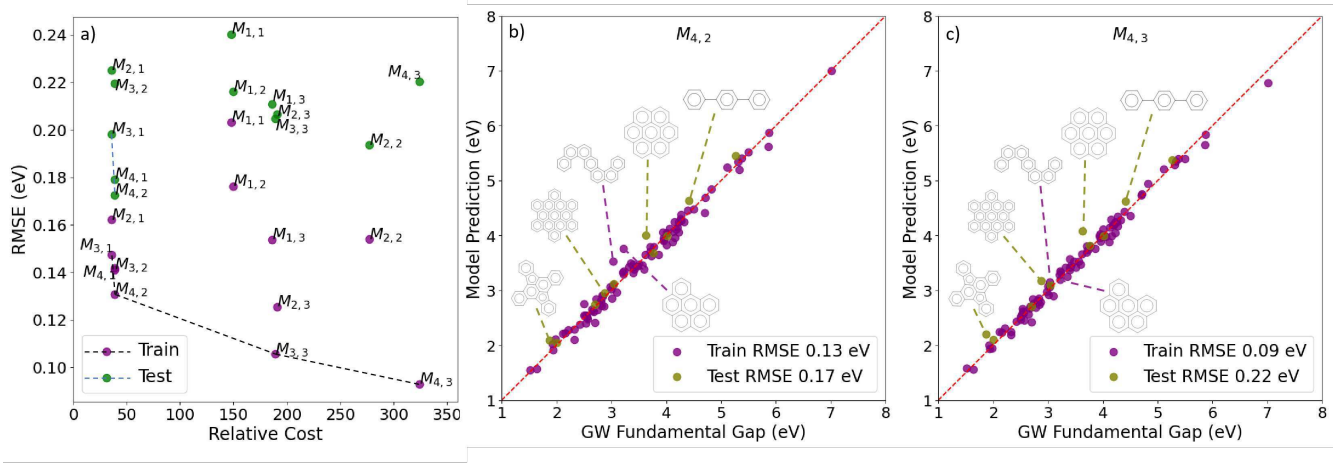

Figure 8. Performance of SISSO-generated models for predicting the *GW* fundamental band gaps of molecular crystals: (a) Pareto chart of the accuracy vs. the computational cost of SISSO-generated models. The "train" accuracy corresponds to the RMSE obtained for the LCV validation set during training and the "test" accuracy corresponds to the withheld set of 10 materials not included in the training. The dashed lines indicate the Pareto front. Model prediction as a function of the *GW* fundamental band gap for (b) $M_{4,2}$ and (c) $M_{4,3}$. Molecular structures of some of the outliers are also shown.

withheld materials, which were excluded from the LCV. The best balance of cost and accuracy is provided by the $M_{4,2}$ model:

$$M_{4,2} = 0.90 \times \frac{E_{\mathrm{T}}^{\mathrm{S}} \times \mathrm{Gap}^{\mathrm{C}}}{\mathrm{Gap}^{\mathrm{S}} \times \rho^{\mathrm{C}}} - 0.063 \times \frac{\ln\left(\mathrm{CB}_{\mathrm{disp}}^{\mathrm{C}}\right) \times \mathrm{AtomNum}^{\mathrm{C}}}{\mathrm{MolWt}^{\mathrm{S}}} + 197 \times \frac{\left(\mathrm{CB}_{\mathrm{disp}}^{\mathrm{C}}\right)^3}{\mathrm{EA}^{\mathrm{S}} \times \mathrm{MolWt}^{\mathrm{S}}} + 0.035 \times \frac{\mathrm{EA}^{\mathrm{S}}}{\mathrm{Gap}^{\mathrm{S}} \times \ln\left(\rho^{\mathrm{C}}\right)} + 1.67 \quad (1)$$

The $M_{3,3}$ and $M_{4,3}$ models, whose computational cost is considerably higher, have a better accuracy for the training set. However, their RMSE increases significantly for the unseen test set, which is indicative of over-fitting. This is also seen in the correlation plots in Figure 8b,c. Interestingly, SISSO does not produce any models that can predict the crystal fundamental gap based only on single molecule features (the equations of all models are provided in the SI).

Code availability

- The *HAppend* code for adding missing hydrogen atoms to molecular crystal structures is available in the GitHub Repository HAppend, together with scripts for making band structure and absorption plots.
- Scripts for calculating the SISSO primary features and for processing SISSO results are available in the GitHub repository MLfeat_FHI-aims.
- The BerkeleyGW code for performing *GW*+BSE calculations[124] is available at the BerkeleyGW website.
- The FHI-aims code,[119] used to perform some relaxations and calculate DFT features, is available at the FHI-aims website.
- The Quantum ESPRESSO code,[123] used to calculate the mean-field wave functions for subsequent *GW*+BSE calculations, is available at the Quantum ESPRESSO website.
- The SISSO code,[63] used to perform sure independent screening and sparsifying operator model training, is available at the GitHub Repository SISSO

References

1. Louie, S. G., Chan, Y.-H., da Jornada, F. H., Li, Z. & Qiu, D. Y. Discovering and understanding materials through computation. *Nat. Mater.* **20**, 728–735 (2021).

2. Luo, S., Li, T., Wang, X., Faizan, M. & Zhang, L. High-throughput computational materials screening and discovery of optoelectronic semiconductors. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **11**, e1489 (2021).

3. Marzari, N., Ferretti, A. & Wolverton, C. Electronic-structure methods for materials design. *Nat. materials* **20**, 736–749 (2021).

4. Stein, H. S. & Gregoire, J. M. Progress and prospects for accelerating materials science with automated and autonomous workflows. *Chem. science* **10**, 9640–9649 (2019).

5. Szczypiński, F. T., Bennett, S. & Jelfs, K. E. Can we predict materials that can be synthesised? *Chem. Sci.* **12**, 830–840 (2021).

6. Jain, A., Shin, Y. & Persson, K. A. Computational predictions of energy materials using density functional theory. *Nat. Rev. Mater.* **1**, 1–13 (2016).

7. Pyzer-Knapp, E. O., Suh, C., Gómez-Bombarelli, R., Aguilera-Iparraguirre, J. & Aspuru-Guzik, A. What is high-throughput virtual screening? a perspective from organic materials discovery. *Annu. Rev. Mater. Res.* **45**, 195–216 (2015).

8. Teale, A. M. *et al.* Dft exchange: sharing perspectives on the workhorse of quantum chemistry and materials science. *Phys. chemistry chemical physics* **24**, 28700–28781 (2022).

9. Allen, A. E. & Tkatchenko, A. Machine learning of material properties: Predictive and interpretable multilinear models. *Sci. advances* **8**, eabm7185 (2022).

10. Fiedler, L., Shah, K., Bussmann, M. & Cangi, A. Deep dive into machine learning density functional theory for materials science and chemistry. *Phys. Rev. Mater.* **6**, 040301 (2022).

11. Foppa, L., Purcell, T. A., Levchenko, S. V., Scheffler, M. & Ghiringhelli, L. M. Hierarchical symbolic regression for identifying key physical parameters correlated with bulk properties of perovskites. *Phys. Rev. Lett.* **129**, 055301 (2022).

12. Hoock, B., Rigamonti, S. & Draxl, C. Advancing descriptor search in materials science: feature engineering and selection strategies. *New J. Phys.* **24**, 113049 (2022).

13. Peng, J. *et al.* Human-and machine-centred designs of molecules and materials for sustainability and decarbonization. *Nat. Rev. Mater.* **7**, 991–1009 (2022).

14. Doan, H. A., Wang, X. & Snurr, R. Q. Computational screening of supported metal oxide nanoclusters for methane activation: Insights into homolytic versus heterolytic c–h bond dissociation. *The J. Phys. Chem. Lett.* **14**, 5018–5024 (2023).

15. Jain, A., Voznyy, O. & Sargent, E. H. High-throughput screening of lead-free perovskite-like materials for optoelectronic applications. *The J. Phys. Chem. C* **121**, 7183–7187 (2017).

16. Diao, X. *et al.* High-throughput screening of stable and efficient double inorganic halide perovskite materials by dft. *Sci. Reports* **12**, 12633 (2022).

17. Broberg, D. *et al.* High-throughput calculations of charged point defect properties with semi-local density functional theory—performance benchmarks for materials screening applications. *npj Comput. Mater.* **9**, 72 (2023).

18. Brunin, G., Ricci, F., Ha, V.-A., Rignanese, G.-M. & Hautier, G. Transparent conducting materials discovery using high-throughput computing. *npj Comput. Mater.* **5**, 63 (2019).

19. Casida, M. E. Time-dependent density-functional theory for molecules and molecular solids. *J. Mol. Struct. THEOCHEM* **914**, 3–18 (2009).

20. Adamo, C. & Jacquemin, D. The calculations of excited-state properties with time-dependent density functional theory. *Chem. Soc. Rev.* **42**, 845–856 (2013).

21. Laurent, A. D. & Jacquemin, D. Td-dft benchmarks: a review. *Int. J. Quantum Chem.* **113**, 2019–2039 (2013).

22. Herbert, J. M. Density-functional theory for electronic excited states. In *Theoretical and computational photochemistry*, 69–118 (Elsevier, 2023).

23. Wang, X., Gao, S., Zhao, M. & Marom, N. Benchmarking time-dependent density functional theory for singlet excited states of thermally activated delayed fluorescence chromophores. *Phys. Rev. Res.* **4**, 033147 (2022).

24. Rohlfing, M. & Louie, S. G. Electron-hole excitations and optical spectra from first principles. *Phys. Rev. B* **62**, 4927–4944, 10.1103/PhysRevB.62.4927 (2000).

25. Sharifzadeh, S. Many-body perturbation theory for understanding optical excitations in organic molecules and solids. *J. Physics: Condens. Matter* **30**, 153002 (2018).

26. Blase, X., Duchemin, I. & Jacquemin, D. The bethe–salpeter equation in chemistry: relations with td-dft, applications and challenges. *Chem. Soc. Rev.* **47**, 1022–1043 (2018).

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

28. Blase, X., Duchemin, I., Jacquemin, D. & Loos, P.-F. The bethe–salpeter equation formalism: From physics to chemistry. *The J. Phys. Chem. Lett.* **11**, 7371–7382 (2020).

29. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom* **65**, 1501–1509 (2013).

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

31. Sorkun, M. C., Astruc, S., Koelman, J. V. A. & Er, S. An artificial intelligence-aided virtual screening recipe for two-dimensional materials discovery. *npj Comput. Mater.* **6**, 106 (2020).

32. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).

33. Deng, B. *et al.* Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* **5**, 1031–1041 (2023).

34. Brockherde, F. *et al.* Bypassing the kohn-sham equations with machine learning. *Nat. communications* **8**, 872 (2017).

35. Schleder, G. R., Padilha, A. C., Acosta, C. M., Costa, M. & Fazzio, A. From dft to machine learning: recent approaches to materials science–a review. *J. Physics: Mater.* **2**, 032001 (2019).

36. Li, H. *et al.* Deep-learning density functional theory hamiltonian for efficient ab initio electronic-structure calculation. *Nat. Comput. Sci.* **2**, 367–377 (2022).

37. Huang, B., von Rudorff, G. F. & von Lilienfeld, O. A. The central role of density functional theory in the ai age. *Science* **381**, 170–175 (2023).

38. Choudhary, K. *et al.* Recent advances and applications of deep learning methods in materials science. *npj Comput. Mater.* **8**, 59 (2022).

39. Bhat, V., Ganapathysubramanian, B. & Risko, C. Rapid estimation of the intermolecular electronic couplings and charge-carrier mobilities of crystalline molecular organic semiconductors through a machine learning pipeline. *The J. Phys. Chem. Lett.* **15**, 7206–7213, 10.1021/acs.jpclett.4c01309 (2024). PMID: 38973725.

40. Jain, A. *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials* **1** (2013).

41. Curtarolo, S. *et al.* Aflow: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).

42. Choudhary, K. *et al.* The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj computational materials* **6**, 173 (2020).

43. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. data* **1**, 1–7 (2014).

44. Schreiner, M., Bhowmik, A., Vegge, T., Busk, J. & Winther, O. Transition1x-a dataset for building generalizable reactive machine learning potentials. *Sci. Data* **9**, 779 (2022).

45. O'Mara, J., Meredig, B. & Michel, K. Materials data infrastructure: a case study of the citrination platform to examine data import, storage, and access. *Jom* **68**, 2031–2034 (2016).

46. Stuke, A. *et al.* Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Sci. data* **7**, 58 (2020).

47. Olsthoorn, B., Geilhufe, R. M., Borysov, S. S. & Balatsky, A. V. Band gap prediction for large organic crystal structures with machine learning. *Adv. Quantum Technol.* **2**, 1900023, https://doi.org/10.1002/qute.201900023 (2019). https://onlinelibrary.wiley.com/doi/pdf/10.1002/qute.201900023.

48. Fediai, A., Reiser, P., Peña, J., Friederich, P. & Wenzel, W. Accurate gw frontier orbital energies of 134 kilo molecules. *Sci. Data* **10**, 10.1038/s41597-023-02486-4 (2023).

49. van Setten, M. J. *et al.* Gw 100: Benchmarking g 0 w 0 for molecular systems. *J. chemical theory computation* **11**, 5665–5687 (2015).

50. Venturella, C., Hillenbrand, C., Li, J. & Zhu, T. Machine learning many-body green's functions for molecular excitation spectra. *J. Chem. Theory Comput.* **20**, 143–154, 10.1021/acs.jctc.3c01146 (2024). PMID: 38150268, https://doi.org/10.1021/acs.jctc.3c01146.

51. Fare, C., Fenner, P., Benatan, M., Varsi, A. & Pyzer-Knapp, E. O. A multi-fidelity machine learning approach to high throughput materials screening. *npj Comput. Mater.* **8**, 257 (2022).

52. Palizhati, A. *et al.* Agents for sequential learning using multiple-fidelity data. *Sci. reports* **12**, 4694 (2022).

53. Chen, C., Zuo, Y., Ye, W., Li, X. & Ong, S. P. Learning properties of ordered and disordered materials from multi-fidelity data. *Nat. Comput. Sci.* **1**, 46–53 (2021).

54. Batra, R. & Sankaranarayanan, S. Machine learning for multi-fidelity scale bridging and dynamical simulations of materials. *J. Physics: Mater.* **3**, 031002 (2020).

55. Liu, D. & Wang, Y. Multi-fidelity physics-constrained neural network and its application in materials modeling. *J. Mech. Des.* **141**, 121403 (2019).

56. Pilania, G., Gubernatis, J. E. & Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **129**, 156–163 (2017).

57. Islam, M., Thakur, M. S. H., Mojumder, S. & Hasan, M. N. Extraction of material properties through multi-fidelity deep learning from molecular dynamics simulation. *Comput. Mater. Sci.* **188**, 110187 (2021).

58. Yang, J., Manganaris, P. & Mannodi-Kanakkithodi, A. Discovering novel halide perovskite alloys using multi-fidelity machine learning and genetic algorithm. *The J. Chem. Phys.* **160** (2024).

59. Liu, X., De Breuck, P.-P., Wang, L. & Rignanese, G.-M. A simple denoising approach to exploit multi-fidelity data for machine learning materials properties. *npj Comput. Mater.* **8**, 233 (2022).

60. Greenman, K. P., Green, W. H. & Gómez-Bombarelli, R. Multi-fidelity prediction of molecular optical peaks with deep learning. *Chem. science* **13**, 1152–1162 (2022).

61. Feng, S., Zhou, H. & Dong, H. Using deep neural network with small dataset to predict material defects. *Mater. & Des.* **162**, 300–310 (2019).

62. De Breuck, P.-P., Hautier, G. & Rignanese, G.-M. Materials property prediction for limited datasets enabled by feature selection and joint learning with modnet. *npj Comput. Mater.* **7**, 83 (2021).

63. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M. & Ghiringhelli, L. M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.* **2**, 083802, 10.1103/PhysRevMaterials.2.083802 (2018). 1710.03319.

64. Purcell, T. A. R., Scheffler, M. & Ghiringhelli, L. M. Recent advances in the SISSO method and their implementation in the SISSO++ code. *The J. Chem. Phys.* **159**, 114110, 10.1063/5.0156620 (2023). https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/5.0156620/18931308/114110_1_5.0156620.pdf.

65. Cao, G. *et al.* Artificial intelligence for high-throughput discovery of topological insulators: The example of alloyed tetradymites. *Phys. Rev. Mater.* **4**, 034204, 10.1103/PhysRevMaterials.4.034204 (2020).

66. Bartel, C. J. *et al.* New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **5**, eaav0693 (2019).

67. Andersen, M., Levchenko, S. V., Scheffler, M. & Reuter, K. Beyond Scaling Relations for the Description of Catalytic Materials. *ACS Catal.* **9**, 2752–2759, 10.1021/acscatal.8b04478 (2019). 1902.07495.

68. Bartel, C. J. *et al.* Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry. *Nat. Commun.* **9**, 4168, 10.1038/s41467-018-06682-4 (2018). 1805.08155.

69. Foppa, L. *et al.* Materials genes of heterogeneous catalysis from clean experiments and artificial intelligence. *MRS Bull.* **46**, 1016–1026, 10.1557/s43577-021-00165-6 (2021). 2102.08269.

70. Hart, G. L., Mueller, T., Toher, C. & Curtarolo, S. Machine learning for alloys. *Nat. Rev. Mater.* **6**, 730–755 (2021).

71. Luo, Y., Li, M., Yuan, H., Liu, H. & Fang, Y. Predicting lattice thermal conductivity via machine learning: A mini review. *npj Comput. Mater.* **9**, 4 (2023).

72. Song, Z. *et al.* Distilling universal activity descriptors for perovskite catalysts from multiple data sources via multi-task symbolic regression. *Mater. Horizons* **10**, 1651–1660 (2023).

73. Han, Z.-K. *et al.* Single-atom alloy catalysts designed by first-principles calculations and artificial intelligence. *Nat. communications* **12**, 1833 (2021).

74. Hoffmann, N., Cerqueira, T. F., Schmidt, J. & Marques, M. A. Superconductivity in antiperovskites. *NPJ Comput. Mater.* **8**, 150 (2022).

75. Guo, Z., Hu, S., Han, Z.-K. & Ouyang, R. Improving symbolic regression for predicting materials properties with iterative variable selection. *J. Chem. Theory Comput.* **18**, 4945–4951 (2022).

76. Ma, B. *et al.* An interpretable machine learning strategy for pursuing high piezoelectric coefficients in (k0.5na0.5)nbo3-based ceramics. *npj Comput. Mater.* **9**, 229, 10.1038/s41524-023-01187-1 (2023).

77. Mou, L.-H., Han, T., Smith, P. E. S., Sharman, E. & Jiang, J. Machine learning descriptors for data-driven catalysis study. *Adv. Sci.* **10**, 2301020, https://doi.org/10.1002/advs.202301020 (2023). https://onlinelibrary.wiley.com/doi/pdf/10.1002/advs.202301020.

78. Ren, C., Li, Q., Ling, C. & Wang, J. Mechanism-guided design of photocatalysts for co2 reduction toward multicarbon products. *J. Am. Chem. Soc.* **145**, 28276–28283, 10.1021/jacs.3c11972 (2023). PMID: 38095164, https://doi.org/10.1021/jacs.3c11972.

79. Oh, S.-H., Yoo, S.-H. & Jang, W. Small dataset machine-learning approach for efficient design space exploration: engineering znte-based high-entropy alloys for water splitting. *npj Comput. Mater.* **10**, 166, 10.1038/s41524-024-01341-3 (2024).

80. Khatua, R., Das, B. & Mondal, A. Physics-informed machine learning with data-driven equations for predicting organic solar cell performance. *ACS Appl. Mater. & Interfaces* 10.1021/acsami.4c10868, 10.1021/acsami.4c10868 (2024). PMID: 39388716, https://doi.org/10.1021/acsami.4c10868.

81. Tian, S., Zhou, K., Yin, W. & Liu, Y. Machine learning enables the discovery of 2d invar and anti-invar monolayers. *Nat. Commun.* **15**, 6977, 10.1038/s41467-024-51379-6 (2024).

82. Jacobs, R., Liu, J., Abernathy, H. & Morgan, D. Machine learning design of perovskite catalytic properties. *Adv. Energy Mater.* **14**, 2303684, https://doi.org/10.1002/aenm.202303684 (2024). https://onlinelibrary.wiley.com/doi/pdf/10.1002/aenm.202303684.

83. Wang, H., Ouyang, R., Chen, W. & Pasquarello, A. High-quality data enabling universality of band gap descriptor and discovery of photovoltaic perovskites. *J. Am. Chem. Soc.* **146**, 17636–17645, 10.1021/jacs.4c03507 (2024). PMID: 38698551, https://doi.org/10.1021/jacs.4c03507.

84. Smith, M. B. & Michl, J. Singlet fission. *Chem. reviews* **110**, 6891–6936 (2010).

85. Smith, M. B. & Michl, J. Recent advances in singlet fission. *Annu. review physical chemistry* **64**, 361–386 (2013).

86. Monahan, N. & Zhu, X.-Y. Charge transfer–mediated singlet fission. *Annu. review physical chemistry* **66**, 601–618 (2015).

87. Lee, J. *et al.* Singlet exciton fission photovoltaics. *Accounts chemical research* **46**, 1300–1311 (2013).

88. Xia, J. *et al.* Singlet fission: progress and prospects in solar cells. *Adv. Mater.* **29**, 1601652 (2017).

89. Pazos-Outón, L. M. *et al.* A silicon–singlet fission tandem solar cell exceeding 100% external quantum efficiency with high spectral stability. *ACS energy letters* **2**, 476–480 (2017).

90. Daiber, B., van den Hoven, K., Futscher, M. H. & Ehrler, B. Realistic efficiency limits for singlet-fission silicon solar cells. *ACS energy letters* **6**, 2800–2808 (2021).

91. Fallon, K. J. *et al.* Exploiting excited-state aromaticity to design highly stable singlet fission materials. *J. Am. Chem. Soc.* **141**, 13867–13876 (2019).

92. Rao, A. & Friend, R. H. Harnessing singlet exciton fission to break the shockley–queisser limit. *Nat. reviews materials* **2**, 1–12 (2017).

93. Wang, X., Garcia, T., Monaco, S., Schatschneider, B. & Marom, N. Effect of crystal packing on the excitonic properties of rubrene polymorphs. *CrystEngComm* **18**, 7353–7362 (2016).

94. Wang, X., Liu, X., Cook, C., Schatschneider, B. & Marom, N. On the possibility of singlet fission in crystalline quaterrylene. *The J. chemical physics* **148** (2018).

95. Liu, X. *et al.* Pyrene-stabilized acenes as intermolecular singlet fission candidates: importance of exciton wave-function convergence. *J. Physics: Condens. Matter* **32**, 184001 (2020).

96. Liu, X., Tom, R., Gao, S. & Marom, N. Assessing zethrene derivatives as singlet fission candidates based on multiple descriptors. *The J. Phys. Chem. C* **124**, 26134–26143 (2020).

97. Wang, X. *et al.* Phenylated acene derivatives as candidates for intermolecular singlet fission. *The J. Phys. Chem. C* **123**, 5890–5899 (2019).

98. Liu, X. *et al.* Finding predictive models for singlet fission by machine learning. *npj Comput. Mater.* **8**, 70 (2022).

99. Wu, J., Pisula, W. & Müllen, K. Graphenes as potential material for electronics. *Chem. reviews* **107**, 718–747 (2007).

100. Anthony, J. E. Functionalized acenes and heteroacenes for organic electronics. *Chem. reviews* **106**, 5028–5048 (2006).

101. Hou, J., Inganäs, O., Friend, R. H. & Gao, F. Organic solar cells based on non-fullerene acceptors. *Nat. materials* **17**, 119–128 (2018).

102. Congreve, D. N. *et al.* External quantum efficiency above 100% in a singlet-exciton-fission–based organic photovoltaic cell. *Science* **340**, 334–337 (2013).

103. Weiss, L. R. *et al.* Strongly exchange-coupled triplet pairs in an organic semiconductor. *Nat. Phys.* **13**, 176–181 (2017).

104. Katz, H. E. & Huang, J. Thin-film organic electronic devices. *Annu. Rev. Mater. Res.* **39**, 71–92 (2009).

105. Brédas, J.-L., Norton, J. E., Cornil, J. & Coropceanu, V. Molecular understanding of organic solar cells: the challenges. *Accounts chemical research* **42**, 1691–1699 (2009).

106. Anthony, J. E. The larger acenes: Versatile organic semiconductors. *Angewandte Chemie Int. Ed.* **47**, 452–483, https://doi.org/10.1002/anie.200604045 (2008). https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.200604045.

107. Wang, C., Dong, H., Hu, W., Liu, Y. & Zhu, D. Semiconducting -conjugated systems in field-effect transistors: A material odyssey of organic electronics. *Chem. Rev.* **112**, 2208–2267, 10.1021/cr100380z (2012). PMID: 22111507, https://doi.org/10.1021/cr100380z.

108. Mei, J., Diao, Y., Appleton, A. L., Fang, L. & Bao, Z. Integrated materials design of organic semiconductors for field-effect transistors. *J. Am. Chem. Soc.* **135**, 6724–6746, 10.1021/ja400881n (2013). PMID: 23557391, https://doi.org/10.1021/ja400881n.

109. Khasbaatar, A. *et al.* From solution to thin film: Molecular assembly of -conjugated systems and impact on (opto)electronic properties. *Chem. Rev.* **123**, 8395–8487, 10.1021/acs.chemrev.2c00905 (2023). PMID: 37273196, https://doi.org/10.1021/acs.chemrev.2c00905.

110. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The cambridge structural database. *Acta Crystallogr. Sect. B: Struct. Sci. Cryst. Eng. Mater.* **72**, 171–179 (2016).

111. Wang, X., Tom, R., Liu, X., Congreve, D. N. & Marom, N. An energetics perspective on why there are so few triplet–triplet annihilation emitters. *J. Mater. Chem. C* **8**, 10816–10824 (2020).

112. Wang, X. & Marom, N. An energetics assessment of benzo [a] tetracene and benzo [a] pyrene as triplet–triplet annihilation emitters. *Mol. Syst. Des. & Eng.* **7**, 889–898 (2022).

113. Chen, X.-K., Kim, D. & Brédas, J.-L. Thermally activated delayed fluorescence (tadf) path toward efficient electroluminescence in purely organic materials: molecular level insight. *Accounts Chem. Res.* **51**, 2215–2224 (2018).

114. Wang, X., Wang, A., Zhao, M. & Marom, N. Inverted lowest singlet and triplet excitation energy ordering of graphitic carbon nitride flakes. *The J. Phys. Chem. Lett.* **14**, 10910–10919 (2023).

115. Coropceanu, V. *et al.* Charge transport in organic semiconductors. *Chem. reviews* **107**, 926–952 (2007).

116. Landrum, G. *et al.* Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **8**, 31 (2013).

117. Ong, S. P. *et al.* Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).

118. Clark, S. J. *et al.* First principles methods using castep. *Zeitschrift für kristallographie-crystalline materials* **220**, 567–570 (2005).

119. Blum, V. *et al.* Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **180**, 2175–2196 (2009).

120. Havu, V., Blum, V., Havu, P. & Scheffler, M. Efficient o (n) integration for all-electron electronic structure calculation using numeric basis functions. *J. Comput. Phys.* **228**, 8367–8379 (2009).

121. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. review letters* **77**, 3865 (1996).

122. Tkatchenko, A. & Scheffler, M. Accurate molecular van der waals interactions from ground-state electron density and free-atom reference data. *Phys. review letters* **102**, 073005 (2009).

123. Giannozzi, P. *et al.* Quantum espresso: a modular and open-source software project for quantum simulations of materials. *J. physics: Condens. matter* **21**, 395502 (2009).

124. Deslippe, J. *et al.* Berkeleygw: A massively parallel computer package for the calculation of the quasiparticle and optical properties of materials and nanostructures. *Comput. Phys. Commun.* **183**, 1269–1289 (2012).

125. Scheidgen, M. *et al.* Nomad: A distributed web-based platform for managing materials science research data. *J. Open Source Softw.* **8**, 5388, 10.21105/joss.05388 (2023).

126. Schober, C., Reuter, K. & Oberhofer, H. Critical analysis of fragment-orbital dft schemes for the calculation of electronic coupling values. *The J. Chem. Phys.* **144**, 054103, 10.1063/1.4940920 (2016). https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/1.4940920/14800371/054103_1_online.pdf.

127. Ambrosetti, A., Reilly, A. M., DiStasio, R. A. & Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. *The J. chemical physics* **140** (2014).

128. Tkatchenko, A., DiStasio, R. A., Car, R. & Scheffler, M. Accurate and efficient method for many-body van der waals interactions. *Phys. Rev. Lett.* **108**, 236402, 10.1103/PhysRevLett.108.236402 (2012).

129. Chisholm, J. A. & Motherwell, S. Compack: a program for identifying crystal structure similarity using distances. *J. applied crystallography* **38**, 228–231 (2005).

130. Macrae, C. F. *et al.* Mercury 4.0: From visualization to analysis, design and prediction. *J. applied crystallography* **53**, 226–235 (2020).

131. Hunnisett, L. M. *et al.* The seventh blind test of crystal structure prediction: structure generation methods. *Acta Crystallogr. Sect. B* **80**, 517–547, 10.1107/S2052520624007492 (2024).

132. Hunnisett, L. M. *et al.* The seventh blind test of crystal structure prediction: structure ranking methods. *Acta Crystallogr. Sect. B* **80**, 548–574, 10.1107/S2052520624008679 (2024).

133. Schatschneider, B., Monaco, S., Liang, J.-J. & Tkatchenko, A. High-throughput investigation of the geometry and electronic structures of gas-phase and crystalline polycyclic aromatic hydrocarbons. *The J. Phys. Chem. C* **118**, 19964–19974, 10.1021/jp5064462 (2014). https://doi.org/10.1021/jp5064462.

134. Sharifzadeh, S., Tamblyn, I., Doak, P., Darancet, P. T. & Neaton, J. B. Quantitative molecular orbital energies within a G_0W_0 approximation. *Eur. Phys. J. B* **85**, 323, 10.1140/epjb/e2012-30206-0 (2012). 1204.0509.

135. Filip, M. R., Qiu, D. Y., Del Ben, M. & Neaton, J. B. Screening of excitons by organic cations in quasi-two-dimensional organic–inorganic lead-halide perovskites. *Nano letters* **22**, 4870–4878 (2022).

136. Biswas, T. & Singh, A. pygwbse: a high throughput workflow package for gw-bse calculations. *npj Comput. Mater.* **9**, 22, 10.1038/s41524-023-00976-y (2023).

137. Bonacci, M. *et al.* Towards high-throughput many-body perturbation theory: efficient algorithms and automated workflows. *npj Comput. Mater.* **9**, 74, https://doi.org/10.1038/s41524-023-01027-2 (2023).

138. Rasmussen, A., Deilmann, T. & Thygesen, K. Towards fully automated gw band structure calculations: What we can learn from 60.000 self-energy evaluations. *npj Comput. Mater.* **7**, 10.1038/s41524-020-00480-7 (2021).

139. Großmann, M., Grunert, M. & Runge, E. A robust, simple, and efficient convergence workflow for gw calculations. *npj Comput. Mater.* **10**, 135, 10.1038/s41524-024-01311-9 (2024).

140. Jacquemin, D., Duchemin, I. & Blase, X. Benchmarking the bethe–salpeter formalism on a standard organic molecular set. *J. Chem. Theory Comput.* **11**, 3290–3304 (2015).

141. Jacquemin, D., Duchemin, I., Blondel, A. & Blase, X. Benchmark of bethe-salpeter for triplet excited-states. *J. Chem. Theory Comput.* **13**, 767–783 (2017).

142. Forster, A. & Visscher, L. Gw100: A slater-type orbital perspective. *J. Chem. Theory Comput.* **17**, 5080–5097 (2021).

143. Knight, J. W. *et al.* Accurate ionization potentials and electron affinities of acceptor molecules iii: a benchmark of gw methods. *J. chemical theory computation* **12**, 615–626 (2016).

144. Bruneval, F., Hamed, S. M. & Neaton, J. B. A systematic benchmark of the ab initio bethe-salpeter equation approach for low-lying optical excitations of small organic molecules. *The J. Chem. Phys.* **142** (2015).

145. Tauc, J. Optical properties and electronic structure of amorphous ge and si. *Mater. research bulletin* **3**, 37–46 (1968).

146. Viezbicke, B. D., Patel, S., Davis, B. E. & Birnie III, D. P. Evaluation of the tauc method for optical absorption edge determination: Zno thin films as a model system. *physica status solidi (b)* **252**, 1700–1710 (2015).

147. Makuła, P., Pacia, M. & Macyk, W. How to correctly determine the band gap energy of modified semiconductor photocatalysts based on uv–vis spectra. *The J. Phys. Chem. Lett.* **9**, 6814–6817, 10.1021/acs.jpclett.8b02892 (2018). https://doi.org/10.1021/acs.jpclett.8b02892.

148. Klein, J. *et al.* Limitations of the tauc plot method. *Adv. Funct. Mater.* **33**, 2304523, https://doi.org/10.1002/adfm.202304523 (2023). https://onlinelibrary.wiley.com/doi/pdf/10.1002/adfm.202304523.

149. Schnepp, O. Electronic spectra of molecular crystals. *Annu. Rev. Phys. Chem.* **14**, 35–60 (1963).

150. Rangel, T. *et al.* Structural and excited-state properties of oligoacene crystals from first principles. *Phys. Rev. B* **93**, 115206 (2016).

151. Ahn, T.-S. *et al.* Experimental and theoretical study of temperature dependent exciton delocalization and relaxation in anthracene thin films. *The J. chemical physics* **128** (2008).

152. Lim, S.-H., Bjorklund, T. G., Spano, F. C. & Bardeen, C. J. Exciton delocalization and superradiance in tetracene thin films and nanoaggregates. *Phys. review letters* **92**, 107402 (2004).

153. Bree, A. & Lyons, L. 998. photo-and semi-conductance of organic crystals. part vi. effect of oxygen on the surface photo-current and some photochemical properties of solid anthracene. *J. Chem. Soc. (Resumed)* 5179–5186 (1960).

154. Park, S., Kim, S., Kim, J., Whang, C. & Im, S. Optical and luminescence characteristics of thermally evaporated pentacene films on si. *Appl. physics letters* **80**, 2872–2874 (2002).

155. Faltermeier, D., Gompf, B., Dressel, M., Tripathi, A. K. & Pflaum, J. Optical properties of pentacene thin films and single crystals. *Phys. Rev. B* **74**, 125416 (2006).

156. Jentzsch, T., Juepner, H., Brzezinka, K.-W. & Lau, A. Efficiency of optical second harmonic generation from pentacene films of different morphology and structure. *Thin solid films* **315**, 273–280 (1998).

157. Sharifzadeh, S., Biller, A., Kronik, L. & Neaton, J. B. Quasiparticle and optical spectroscopy of the organic semiconductors pentacene and ptcda from first principles. *Phys. Rev. B* **85**, 125307 (2012).

158. Tiago, M. L., Northrup, J. E. & Louie, S. G. Ab initio calculation of the electronic and optical properties of solid pentacene. *Phys. Rev. B* **67**, 115212 (2003).

159. Sun, D. *et al.* Anisotropic singlet fission in single crystalline hexacene. *Iscience* **19**, 1079–1089 (2019).

160. Watanabe, M. *et al.* The synthesis, crystal structure and charge-transport properties of hexacene. *Nat. chemistry* **4**, 574–578 (2012).

161. Busby, E. *et al.* Multiphonon relaxation slows singlet fission in crystalline hexacene. *J. Am. Chem. Soc.* **136**, 10654–10660 (2014).

162. Najafov, H., Lee, B., Zhou, Q., Feldman, L. C. & Podzorov, V. Observation of long-range exciton diffusion in highly ordered organic semiconductors. *Nat. materials* **9**, 938–943 (2010).

163. Huang, L. *et al.* Rubrene micro-crystals from solution routes: their crystallography, morphology and optical properties. *J. Mater. Chem.* **20**, 159–166 (2010).

164. Tanaka, J. The electronic spectra of aromatic molecular crystals. ii. the crystal structure and spectra of perylene. *Bull. Chem. Soc. Jpn.* **36**, 1237–1249 (1963).

165. Mulder, B. Photoconductivity spectra of stable and metastable single-crystals of perylene. *Recueil des Travaux Chimiques des Pays-Bas* **84**, 713–728 (1965).

166. Kurrle, D. & Pflaum, J. Exciton diffusion length in the organic semiconductor diindenoperylene. *Appl. Phys. Lett.* **92** (2008).

167. Maruyama, Y., Iwaki, T., Kajiwara, T., Shirotani, I. & Inokuchi, H. Molecular orientation and absorption spectra of quaterrylene evaporated film. *Bull. Chem. Soc. Jpn.* **43**, 1259–1261 (1970).

168. Maruyama, Y., Inokuchi, H. & Harada, Y. Electronic properties of quaterrylene, c40h20. *Bull. Chem. Soc. Jpn.* **36**, 1193–1198 (1963).

169. Fuchs, M. & Scheffler, M. Ab initio pseudopotentials for electronic structure calculations of poly-atomic systems using density-functional theory. *Comput. Phys. Commun.* **119**, 67–98 (1999).

170. Nijegorodov, N., Mabbs, R. & Downey, W. Evolution of absorption, fluorescence, laser and chemical properties in the series of compounds perylene, benzo (ghi) perylene and coronene. *Spectrochimica Acta Part A: Mol. Biomol. Spectrosc.* **57**, 2673–2685 (2001).

171. Xiao, J. *et al.* Preparation, characterization, and photoswitching/light-emitting behaviors of coronene nanowires. *J. Mater. Chem.* **21**, 1423–1427 (2011).

172. Proehl, H. *et al.* Comparison of ultraviolet photoelectron spectroscopy and scanning tunneling spectroscopy measurements on highly ordered ultrathin films of hexa-peri-hexabenzocoronene on au (111). *Phys. Rev. B* **63**, 205409 (2001).

173. Schatschneider, B., Monaco, S., Liang, J.-J. & Tkatchenko, A. High-throughput investigation of the geometry and electronic structures of gas-phase and crystalline polycyclic aromatic hydrocarbons. *The J. Phys. Chem. C* **118**, 19964–19974 (2014).

174. Brodin, M. & Soskin, M. Investigation of the absorption spectrum of a single crystal of 1.2-benzanthracene in the region of lowest electronic transitions. *OPTIKA I SPEKTROSKOPIYA* **6**, 600–604 (1959).

175. Ramasesha, S., Albert, I. & Sinha, B. Optical and magnetic properties of the exact ppp states of biphenyl. *Mol. Phys.* **72**, 537–547 (1991).

176. Coffman, R. & McClure, D. S. The electronic spectra of crystalline toluene, dibenzyl, diphenylmethane, and biphenyl in the near ultraviolet. *Can. J. Chem.* **36**, 48–58 (1958).

177. Puschnig, P. *et al.* Pressure studies on the intermolecular interactions in biphenyl. *Synth. metals* **116**, 327–331 (2001).

178. Gondo, Y. Electronic structure and spectra of biphenyl and its related compound. *The J. Chem. Phys.* **41**, 3928–3938 (1964).

179. Hino, S., Veszprémi, T., Ohno, K., Inokuchi, H. & Seki, K. Absorption spectra of volatile aromatic hydrocarbon films in the vacuum ultraviolet region. *Chem. Phys.* **71**, 135–144 (1982).

180. Mukherjee, B. & Ganguly, S. Anisotropy of the electronic spectra of a single crystal of 1, 12-benzperylene (c22h12). *Proc. Phys. Soc.* **83**, 93 (1964).

181. Pu, Y.-J. *et al.* Absence of delayed fluorescence and triplet–triplet annihilation in organic light emitting diodes with spatially orthogonal bianthracenes. *J. Mater. Chem. C* **7**, 2541–2547 (2019).

182. Manna, B., Nandi, A. & Chandrakumar, K. Comparative study of exciton dynamics in 9, 9-bianthracene nanoaggregates and thin films: Observation of singlet–singlet annihilation-mediated triplet exciton formation. *The J. Phys. Chem. C* **126**, 10762–10771 (2022).

183. Tanaka, J. The electronic spectra of pyrene, chrysene, azulene, coronene and tetracene crystals. *Bull. Chem. Soc. Jpn.* **38**, 86–102 (1965).

184. Puschnig, P., Meisenbichler, C. & Draxl, C. Excited state properties of organic semiconductors: breakdown of the tamm-dancoff approximation. *arXiv preprint arXiv:1306.3790* (2013).

185. Lettmann, T. & Rohlfing, M. Electronic excitations of polythiophene within many-body perturbation theory with and without the tamm–dancoff approximation. *J. Chem. Theory Comput.* **15**, 4547–4554 (2019).

186. Ma, Y., Rohlfing, M. & Molteni, C. Excited states of biological chromophores studied using many-body perturbation theory: Effects of resonant-antiresonant coupling and dynamical screening. *Phys. Rev. B* **80**, 241405, 10.1103/PhysRevB.80.241405 (2009).

187. Lettmann, T. & Rohlfing, M. Finite-momentum excitons in rubrene single crystals. *Phys. Rev. B* **104**, 115427, 10.1103/PhysRevB.104.115427 (2021).

188. Ambrosch-Draxl, C., Nabok, D., Puschnig, P. & Meisenbichler, C. The role of polymorphism in organic thin films: oligoacenes investigated from first principles. *New J. Phys.* **11**, 125010 (2009).

189. Zhang, X., Leveillee, J. A. & Schleife, A. Effect of dynamical screening in the bethe-salpeter framework: Excitons in crystalline naphthalene. *arXiv preprint arXiv:2302.07948* (2023).

190. Wang, X. *et al.* Computational discovery of intermolecular singlet fission materials using many-body perturbation theory. *The J. Phys. Chem. C* **0**, null, 10.1021/acs.jpcc.4c01340 (0). https://doi.org/10.1021/acs.jpcc.4c01340.

191. Sharifzadeh, S., Biller, A., Kronik, L. & Neaton, J. B. Quasiparticle and optical spectroscopy of the organic semiconductors pentacene and ptcda from first principles. *Phys. Rev. B* **85**, 125307, 10.1103/PhysRevB.85.125307 (2012).

192. Gallandi, L., Marom, N., Rinke, P. & Körzdörfer, T. Accurate ionization potentials and electron affinities of acceptor molecules ii: Non-empirically tuned long-range corrected hybrid functionals. *J. Chem. Theory Comput.* **12**, 605–614, 10.1021/acs.jctc.5b00873 (2016). PMID: 26731340, https://doi.org/10.1021/acs.jctc.5b00873.

193. Golze, D., Dvorak, M. & Rinke, P. The gw compendium: A practical guide to theoretical photoemission spectroscopy. *Front. chemistry* **7**, 377 (2019).

194. Marom, N. *et al.* Benchmark of g w methods for azabenzenes. *Phys. Rev. B* **86**, 245127 (2012).

195. Schatschneider, B. *et al.* Electrodynamic response and stability of molecular crystals. *Phys. Rev. B* **87**, 060104, 10.1103/PhysRevB.87.060104 (2013).

196. Parker, C. & Hatchard, C. Triplet-singlet emission in fluid solutions. phosphorescence of eosin. *Transactions Faraday Soc.* **57**, 1894–1904 (1961).

197. Endo, A. *et al.* Efficient up-conversion of triplet excitons into a singlet state and its application for organic light emitting diodes. *Appl. Phys. Lett.* **98** (2011).

198. Yang, Z. *et al.* Recent advances in organic thermally activated delayed fluorescence materials. *Chem. Soc. Rev.* **46**, 915–1016 (2017).

199. Cai, X. *et al.* Purely organic crystals exhibit bright thermally activated delayed fluorescence. *Angewandte Chemie Int. Ed.* **58**, 13522–13531 (2019).

200. Zhan, L. *et al.* A simple organic molecule realizing simultaneous tadf, rtp, aie, and mechanoluminescence: understanding the mechanism behind the multifunctional emitter. *Angewandte Chemie* **131**, 17815–17819 (2019).

201. Brebels, J., Manca, J. V., Lutsen, L., Vanderzande, D. & Maes, W. High dielectric constant conjugated materials for organic photovoltaics. *J. Mater. Chem. A* **5**, 24037–24050, 10.1039/C7TA06808E (2017).

Acknowledgements

Work at CMU was supported by the National Science Foundation (NSF) Designing Materials to Revolutionize and Engineer our Future (DMREF) program under award DMR-2323749. This research used resources of the Argonne Leadership Computing Facility (ALCF), which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357 and of the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy, under Contract DE-AC02-05CH11231.

Author contributions statement

S. G., X. L., Y. L., X. W., K. Z., and V. C. performed the calculations and curated the data. S. G. and Y. L. performed additional analysis and validation. B.S. provided the structures relaxed with CASTEP. N. M., S. G., X. L., Y. L., and X. W. wrote the manuscript. S. G. and X. L. contributed equally to this work. N. M. conceived and led the project.

Competing interests

The authors declare no competing interests.