npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Finding predictive models for singlet fission by machine learning

Xingyu Liu[1], Xiaopeng Wang[2], Siyu Gao[1], Vincent Chang[1], Rithwik Tom[3], Maituo Yu[1], Luca M. Ghiringhelli[4] and Noa Marom[1,3,5]✉

Singlet fission (SF), the conversion of one singlet exciton into two triplet excitons, could significantly enhance solar cell efficiency. Molecular crystals that undergo SF are scarce. Computational exploration may accelerate the discovery of SF materials. However, many-body perturbation theory (MBPT) calculations of the excitonic properties of molecular crystals are impractical for large-scale materials screening. We use the sure-independence-screening-and-sparsifying-operator (SISSO) machine-learning algorithm to generate computationally efficient models that can predict the MBPT thermodynamic driving force for SF for a dataset of 101 polycyclic aromatic hydrocarbons (PAH101). SISSO generates models by iteratively combining physical primary features. The best models are selected by linear regression with cross-validation. The SISSO models successfully predict the SF driving force with errors below 0.2 eV. Based on the cost, accuracy, and classification performance of SISSO models, we propose a hierarchical materials screening workflow. Three potential SF candidates are found in the PAH101 set.

npj Computational Materials (2022)8:70 ; https://doi.org/10.1038/s41524-022-00758-y

INTRODUCTION

Singlet fission (SF) is the conversion of one photo-generated singlet-state exciton into two triplet-state excitons[1–11]. Intermolecular SF, where the triplet-state excitons are localized on different chromophores than the singlet exciton, occurs in molecular crystals[12–14]. SF may be utilized in solar cells to exploit the excess energy of high-energy photons and reduce the energy loss due to thermalization. Harvesting two charge carriers from one photon via SF could potentially increase the power conversion efficiency of solar cells beyond the Shockley–Queisser limit[15]. However, commercial SF-based solar cells have yet to be realized owing to the dearth of suitable materials[1,16]. Certain classes of molecular materials, such as oligoacenes, oligorylenes, and their derivatives, are experimentally known to undergo SF[17–26]. Although 200% quantum yield and ultra-fast SF have been observed experimentally[27,28], most of the known SF materials are not practical for use in commercial modules because they are chemically unstable and would degrade under operating conditions. It is therefore imperative to find new SF materials, possibly from different chemical families, in order to expand the available options. Computational exploration of the chemical space may significantly accelerate the discovery of candidates for SF in the solid state and guide experimental efforts in promising directions.

The primary criterion for SF to occur is the thermodynamic driving force. The energy difference between the initial singlet state and final state of two triplets ($E_S - 2E_T$) must be positive or at most slightly negative[1,3,29,30]. Organic molecular crystals that meet this requirement are rare, which explains why most known SF materials belong to restricted classes of molecules. Yet, most of the vast chemical space remains largely unexplored. Computationally efficient density functional theory (DFT) based on semi-local exchange-correlation functionals has been used extensively for high-throughput screening of materials[31–34]. However, DFT is a ground-state theory. Hence, it cannot directly describe the excited-state properties of chromophores that are of interest for SF. Time-dependent DFT (TDDFT) may be used to calculate the excitation energies of isolated molecules[35,36]. This relatively low-cost option has been adopted to screen molecules with up to 100 atoms in search of SF candidates[37]. However, SF-based solar cells utilize solid-state materials, i.e., molecular crystals[14], whose performance depends not only on the properties of the molecular constituents but also on crystal packing[38]. Therefore, it is desirable to screen molecular crystals, rather than isolated molecules, in search of potential SF materials. Many-body perturbation theory (MBPT) within the *GW* approximation paired with the Bethe–Salpeter equation (*GW* + BSE) is the state-of-the-art method for predicting the excitonic properties of organic molecular crystals with periodic boundary conditions[39–41]. Using this method, we have already identified several potential candidate materials for intermolecular SF in the solid state[16,29,42–45]. However, the high computational cost of *GW* + BSE calculations is prohibitive for large-scale screening of materials databases. Therefore, it is desirable to identify descriptors that are fast to evaluate and yield models that accurately predict *GW* + BSE results. To this end, machine-learning (ML) algorithms for feature selection may be used.

ML is increasingly employed in conjunction with first-principles simulations for materials discovery[46–55]. Typically, large datasets are required to train ML models, making data acquisition the computational bottleneck. The growing availability of datasets and repositories of DFT calculations[32,56–62] has facilitated the application of ML to the ground-state properties of materials. Applications of ML to excited-state properties are still relatively rare, owing to the high cost of data acquisition[63–66]. Incorporating physical and chemical knowledge may enable the construction of predictive ML models with small datasets.

Here, we employ the sure-independence-screening-and-sparsifying-operator (SISSO)[67] ML algorithm to identify low-cost

[1]Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [2]Qingdao Institute for Theoretical and Computational Sciences, Shandong University, Qingdao, Shandong 266237, People's Republic of China. [3]Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [4]NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society and Humboldt University, Berlin, Germany. [5]Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA 15213, USA. ✉email: nmarom@andrew.cmu.edu

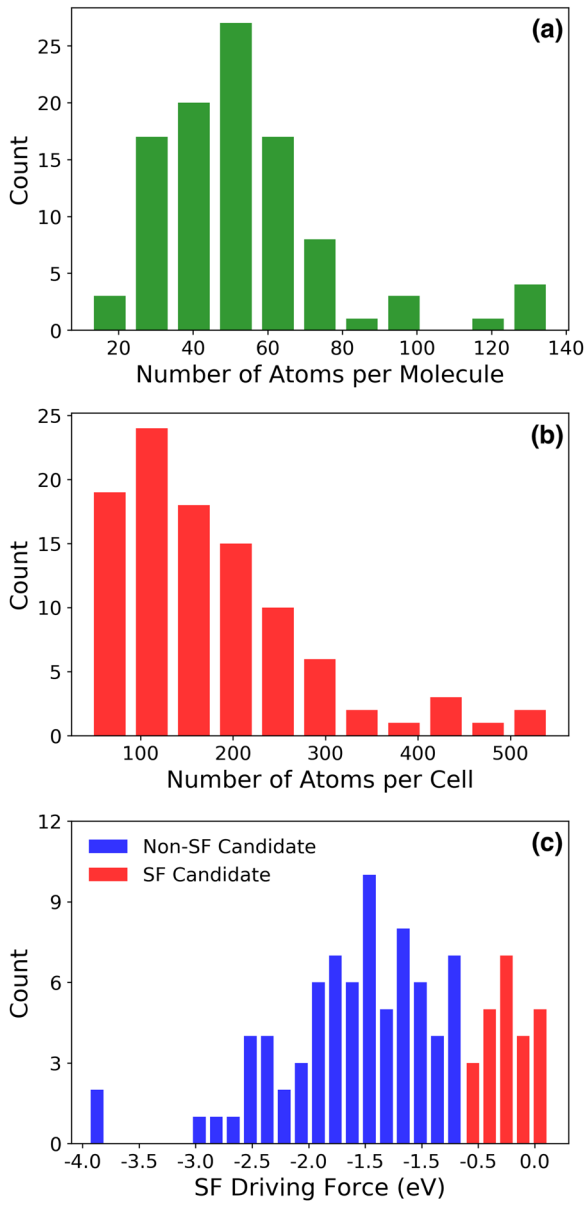

Fig. 1 Statistics of the PAH101 set. Distributions of **a** the number of atoms per molecule, **b** the number of atoms per unit cell, and **c** the GW+BSE@PBE SF driving force in the PAH101 dataset. SF candidates and non-SF candidates are colored in red and blue, respectively in panel **c**.

predictive models for SF. 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. SISSO essentially performs a computer experiment in which hypotheses are systematically generated and tested against reference data. Physical and chemical knowledge is leveraged in the choice of primary features and in the rules for combining them. An important advantage of SISSO is that it can work well with a relatively small amount of data. It has been demonstrated in several applications that SISSO can produce predictive models with as little as a few hundred[68–71], or even a few tens of training data points[72]. Moreover, SISSO-generated models are based on interpretable physical descriptors that may provide insight into which features are correlated with the target property[69–71].

To train SISSO, we compile a purpose-built dataset of $GW + \mathrm{BSE}$ calculations of the SF driving force, $E_S - 2E_T$, of 101 molecular crystals of polycyclic aromatic hydrocarbons (PAHs). Most known SF materials are PAHs, in particular acenes, rylenes, and their derivatives[17–26]. However, PAHs, broadly defined as compounds comprising carbon and hydrogen atoms and containing multiple aromatic rings, encompass a multitude of chemical families, which have not been explored in the context of SF. To maximize the chances of discovering new classes of SF materials, we have selected a set of PAH crystals, representing diverse chemical families. For the same set of materials, 16 physically motivated primary features are calculated. Because the properties of molecular crystals depend on both the single-molecule properties and the crystal packing in the solid state, the primary features include both single-molecule and crystal features. SISSO produces several predictive models with varying degrees of complexity. The most accurate models generated yield a training set root-mean-square error (RMSE) below 0.2 eV, which is on par with $GW + \mathrm{BSE}$. Moreover, the best-performing models have a near-perfect classification accuracy for determining whether or not a given material is a promising SF candidate. Based on considerations of the model accuracy vs. the computational cost of primary feature evaluation, a hierarchical screening approach is proposed to narrow down the candidate pool. The variance between the predictions of different SISSO-generated models may be used as a measure of uncertainty. Based on the SISSO-generated models, three potential SF candidates are identified: 9-(4-biphenyl) cyclopenta[*a*]phenalene (BCPP), tetrabenzo[*de,hi,op,st*]pentacene (TBPT), and 5,6–11,12-diphenylenenaphthacene (DPNP). These compounds belong to chemical families of PAHs that have not been previously explored in the context of SF.

RESULTS AND DISCUSSION

The PAH101 dataset

Because most known SF materials are PAHs, we focus on this class of compounds to maximize the chances of discovery. In addition, restricting the chemical space means that ML models trained on small data are more likely to succeed in producing accurate predictions. A set of 101 PAH crystal structures was extracted from the Cambridge Structural Database (CSD)[73]. The systems in the PAH101 set represent diverse chemical families within the larger PAH class. The chromophore size in the PAH101 set ranges from 12 to 136 atoms and the crystal unit cell size ranges from 44 to 544 atoms, as shown in Fig. 1a–b . Figure 1c shows the SF driving force distribution obtained for the PAH101 set with $GW + \mathrm{BSE}$, based on the Perdew-Burke-Ernzerhof (PBE)[74] DFT functional, denoted as $GW + \mathrm{BSE@PBE}$. We note that $GW + \mathrm{BSE@PBE}$ systematically underestimates the thermodynamic driving force for SF. This is partly owing to the underlying approximations and partly because additional effects, such as electron-phonon coupling[75], entropic effects[30], and kinetics are not considered. Therefore, we assess prospective SF candidates based on their predicted SF driving force relative to the known SF materials pentacene, tetracene, and rubrene[16,29,42–44]. Pentacene has been observed to undergo rapid SF with a 200% triplet yield[27,28]. SF in tetracene is slightly endoergic[20,76]. Rubrene is known to undergo both SF and the reverse process of triplet-triplet annihilation (TTA), where two triplet excitons are converted into one singlet exciton[77–80]. Therefore, we consider the $GW + \mathrm{BSE@PBE}$ SF driving force of rubrene, −0.62 eV, which is even lower than that of tetracene, as the lower limit for viable SF candidates. Indeed, the SF driving force of anthracene, a well-known TTA material[81,82], is below that of rubrene. Thus, even if renormalization of the exciton energies due to phonons were considered, which may tilt the energy

balance in favor of SF in some cases[75], materials with a $GW+$BSE@PBE SF driving force below that of rubrene would still be unlikely to exhibit SF. The PAH101 set contains materials with a broad range of SF driving force in order for the SISSO-generated models to be able to distinguish between materials that are likely and those that are unlikely to undergo SF.

Primary features

The primary features are a collection of descriptors that may be physically relevant to the target property[62,83], in this case, the SF driving force. The excitonic properties of molecular crystals depend on the single-molecule properties as well as the crystal packing[25,38,42,84–91]. Therefore, we consider single-molecule descriptors, denoted by an "S" superscript, and crystal descriptors, denoted by a "C" superscript, as primary features. For computational efficiency, the primary features are calculated at the DFT@PBE level, as described in the Methods section. For single-molecule features, we consider properties that could be correlated with the excitation energies of the chromophore, including the DFT HOMO-LUMO gap (Gap^S), ionization potential (IP^S), electron affinity (EA^S), triplet-state formation energy (E_T^S), and the trace of the polarization tensor ($\mathrm{PolarTensor}^S$). The IP^S and EA^S are calculated based on DFT total energy differences between neutral and charged species. Similarly, the triplet-state formation energy is obtained from the DFT total energy difference between the triplet-state and singlet-state systems. $\mathrm{PolarTensor}^S$ is calculated using the PBE exchange-correlation functional coupled with the many-body dispersion (MBD) method (PBE + MBD)[92]. In addition, we consider a DFT-based estimation of the thermodynamic driving force for SF, where the singlet excitation energy is approximated by the HOMO-LUMO gap and the triplet excitation energy is approximated by the triplet-state formation energy: $\mathrm{DF}^S = \mathrm{Gap}^S - 2E_T^S$.

Crystal features include the DFT bandgap (Gap^C), the triplet-state formation energy (E_T^C), as well as the DFT estimate of the SF thermodynamic driving force, $\mathrm{DF}^C = \mathrm{Gap}^C - 2E_T^C$. In addition, we consider features that reflect the effect of crystal packing and the strength of coupling between neighboring molecules. The fundamental gap of a crystal is narrower than that of a single molecule owing to the combined effect of band dispersion and polarization[43]. Therefore, the crystal features include the valence-band dispersion ($\mathrm{VB}^C_{\mathrm{disp}}$) and conduction-band dispersion ($\mathrm{CB}^C_{\mathrm{disp}}$)[93], as well as the dielectric constant (ϵ^C) as descriptors of the screening effect in a crystal. ϵ^C is calculated using the Clausius–Mossotti relation, with the static polarizability obtained from PBE + MBD[43,94]. Because the intermolecular SF process involves charge/energy transfer between neighboring chromophores, we also consider a descriptor of the intermolecular electronic coupling, the transition matrix element, $H_{ab} = \langle \mathbf{\Phi}_a | \hat{\mathbf{H}} | \mathbf{\Phi}_b \rangle$, between the initial state $\mathbf{\Phi}_a$ of molecule a, and the final state $\mathbf{\Phi}_b$ of molecule b, where $\hat{\mathbf{H}}$ is the Hamiltonian. For hole transport, molecule a is positively charged and molecule b is neutral. The states $\mathbf{\Phi}_a$and $\mathbf{\Phi}_b$ represent the corresponding HOMO. H_{ab} is calculated within the frozen orbital DFT approach[95–97]. Different dimers extracted from the same molecular crystal result in different values of H_{ab}. Hence, we use the average of the three highest H_{ab} values to represent the intermolecular coupling strength in a given crystal. Finally, we consider chemical descriptors, including the molecular weight MolWt^S, the crystal density ρ^C, and the number of atoms in the unit cell $\mathrm{AtomNum}^C$. A full list of the primary features and their descriptions is provided in Supplementary Table 1.

To evaluate the relative computational cost of calculating different primary features, we used a representative system with 62 atoms per molecule and a total of 248 atoms (four molecules) in the unit cell. The CPU time spent on one single-molecule DFT@PBE calculation is considered the basic unit of computational cost. The computational cost of calculating each primary feature is expressed as multiples of that basic unit. For features whose evaluation requires multiple DFT calculations (for example, EA^S requires two DFT@PBE calculations for the neutral and anion) the computational cost of all calculations is summed up. Descriptors such as ρ^C do not require any calculations and therefore have a cost of zero. A full list of the primary features with their relative computational cost is provided in Supplementary Table 2.

Model generation with SISSO

The SISSO training was performed with the SISSO package available at the SISSO GitHub Repository:[67] https://github.com/rouyang2017/SISSO. SISSO can generate a huge feature space with billions (or even trillions) of elements by iteratively combining the primary features using linear and nonlinear elementary mathematical operations[67]. To avoid generating unphysical features, addition and subtraction are allowed only for primary features with the same units. Two key parameters of SISSO are the model dimension and feature rung, which is the number of iterations used to build combined features. Here, the maximal rung (Rung) was set to 3 and the maximal dimension (Dim) was set to 4. These values are found to be sufficient to identify the optimal model complexity, as shown below. The resulting models are denoted as $M_{\mathrm{Dim,\ Rung}}$. The operator set $H = \left\{+, -, \times, \div, \exp, \log, ()^{-1}, ()^2, ()^3, \sqrt{}, \sqrt[3]{}, |\cdot|\right\}$ was used for feature construction. The maximum complexity, i.e., the maximum number of operators in one combined feature, was set to 10. With these settings, a total of 584, 5×10^5, and 5×10^{11} features were generated with Rung = 1, 2, and 3, respectively.

After feature generation, linear regression is performed to yield the model prediction (each model is the scalar product of the SISSO-identified descriptor with the vector of fitted coefficients, via linear regression) and the models are ranked based on their prediction performance. Optimal subspaces are selected from the huge feature space by sure-independence screening (SIS). The number of features saved after SIS is set to 500. On each such subspace, the sparse solution is determined by l_0 normalization (the sparsifying-operator, SO). To assess the optimal model complexity (i.e., Rung and Dim), leave-N-out cross-validation (LCV) is performed, i.e., the performance of the trained models is assessed on unseen data. N data points are held out as an unseen validation set and the remaining data points are used for model training. This process is repeated several times. Here, we use $N = 10$. In the LCV practice, data points are typically randomly assigned to the validation set. Here, rather than the model with the smallest overall prediction error, we are interested in a regression model with higher prediction accuracy at the high SF driving force range in order to identify promising SF candidates with high confidence. Hence, a modified LCV scheme is used, which prioritizes the selection of PAHs with a higher SF driving force than rubrene for the validation set. The selection probability of materials with $E_S - 2E_T \geq -0.62\mathrm{eV}$ is boosted by a factor of 10 compared to other PAHs. For each combination of Rung and Dim, 40 rounds of LCV are performed. In each round, the model with the lowest RMSE for the validation set is selected. Finally, the model that yields the lowest RMSE of the 40 models for the combined training and validation data is selected as $M_{\mathrm{Dim,Rung}}$. We note that the regression coefficients may have units, such that the overall units of the resulting models are eV. A subset of 10 PAH crystals of different sizes with a range of SF driving force values are completely left out of the SISSO training to serve as the test set of unseen data. The SISSO training is performed using the remaining 91 crystals. As a baseline for assessing the performance of SISSO-generated models we use our human-generated models, the DFT estimates of the single molecule and crystal SF driving force DF^S and DF^c.

Because each SISSO-generated model comprises different primary features, each model has a different computational cost.

Here, the computational cost for each model is evaluated by summing over the costs of all the primary features included in the model. The cost of features that appear in the model more than once is counted only once because no additional calculation is required. As mentioned above, SISSO is adopted to train regression models, i.e., predicting the SF driving force, by minimizing the prediction error. However, the same model can also be assessed (without retraining) as a classification model if the two classes of interest are SF vs. non-SF materials. To this end, the materials are classified based on the value of the SF driving force with a threshold of −0.62 eV, corresponding to the SF driving force of rubrene, as explained above. True positive and true negative are defined here based on whether the ML model is in agreement with the GW+BSE reference data regarding whether or not the SF driving force of a certain material is above or below −0.62 eV. The classification performance of each model is assessed based on sensitivity, specificity, and accuracy. Sensitivity is the fraction of correctly identified SF candidates, defined as the number of true positives (TP) divided by the total number of positive labels, which includes true positives and false negatives (FN), TP/(TP + FN). Conversely, specificity is the fraction of correctly identified non-SF candidates, defined as the number of true negatives (TN) divided by the total number of negative labels, which includes true negatives and false positives (FP), TN/(TN + FP). Accuracy measures the overall fraction of correct classifications, which is given by the sum of true positives and true negatives divided by the sum of all labels, (TP + TF)/(TP + FN + TN + FP). The classification performance of all SISSO-generated models for the test set and training set are reported in Supplementary Tables 3, 4, respectively.

Model selection and performance evaluation

Table 1 summarizes the training set and test set RMSE of the best models produced by SISSO with each combination of Dim and Rung. The training set comprises all data used for training, including both the training and validation data in all cross-validations, and the test set comprises the ten data points unseen by the SISSO training process. The formulas of all models are provided in the Supplementary Notes and some models are selected for further discussion in the main text. Overall, all the SISSO-generated models have a higher prediction accuracy than the baseline models DF^S and DF^C. Both SISSO-generated models and the baseline DFT estimation model perform significantly better than the mean value. The prediction error is expected to decrease with the model complexity, i.e., with increasing Rung and Dim, until a saturation point is reached, beyond which the accuracy deteriorates because of overfitting. For Rung = 1 models, the training set RMSE decreases monotonically with increasing model dimension. The test set RMSE, however, peaks at $M_{2,1}$ with both three and four-dimensional models achieving lower RMSE. The better performance of higher dimensional models indicates that the SISSO training does not saturate at $M_{2,1}$. Rather, some PAHs may be more sensitive to the descriptors included in $M_{2,1}$. The improvements from three dimensions to four dimensions for both the training and test sets are marginal, suggesting that the model complexity has saturated. For Rung = 2 models, the training RMSE decreases with increasing dimension, whereas the test RMSE increases slightly for $M_{3,2}$. The slightly worse performance of $M_{3,2}$ for the test set, compared to $M_{2,2}$ and $M_{4,2}$, is negligible, suggesting the model complexity is saturating but the optimum is not reached. For models with Rung = 3, the training RMSE decreases monotonically with the increase in model dimension. However, for the test set, the model performance deteriorates significantly from two dimensions to three and four dimensions. This suggests that Dim = 2 is the saturation point. Similarly, increasing the Rung for models with the same dimension improves the accuracy until an optimum is reached. In general, at fixed Dim, the test RMSE shows a minimum at Rung = 3 for one and two-dimensional models, Rung = 2 for three and four-dimensional ones. The overall lowest test RMSE of 0.18 eV is achieved with $M_{2,3}$, suggesting that this model has the optimal complexity. We note that most of the features included in the low-complexity models are single-molecule properties. These results imply that the SF driving force is heavily dependent on the molecular characteristics. However, because the PAH101 set only contains four sets of polymorphs (rubrene, perylene, diindeno [1,2,3-cd:1′,2′,3′-lm]perylene, and p-quaterphenyl), the effect of crystal packing may be underrepresented.

Table 1. The computational cost and prediction accuracy, represented by the RMSE for the training set and test set, of SISSO-generated models.

Model	Cost	Training RMSE (eV)	Test RMSE (eV)
Mean	0	0.85	0.85
DF^S	4	0.51	0.50
DF^C	124	0.77	0.69
$\text{포}_{1,1}$	7	0.25	0.24
$\text{포}_{2,1}$	54	0.21	0.28
$\text{포}_{3,1}$	95	0.18	0.20
$\text{포}_{4,1}$	131	0.17	0.20
$\text{포}_{1,2}$	7	0.22	0.25
$\text{포}_{2,2}$	181	0.17	0.19
$\text{포}_{3,2}$	213	0.15	0.21
$\text{포}_{4,2}$	251	0.14	0.19
$\text{포}_{1,3}$	130	0.18	0.21
$\text{포}_{2,3}$	130	0.15	0.18
$\text{포}_{3,3}$	304	0.13	0.25
$\text{포}_{4,3}$	304	0.11	0.26

The mean value, and the human-generated models, DF^S and DF^C, are shown as a baseline for comparison. SISSO models are denoted as $M_{\text{Dim, Rung}}$.

To decide which model(s) to use for materials screening, we consider the computational cost in addition to the model accuracy. The relative computational cost of SISSO-generated models is given in Table 1. Figure 2 shows a Pareto chart, in which the model accuracy, represented by the validation and test set RMSE, is plotted against its relative computational cost. The validation set RMSE is calculated using the corresponding train/validation split that produces the final SISSO model. A Pareto chart based on the training RMSE is provided in Supplementary Fig. 3, which leads to similar conclusions. More complex models tend to have a higher computational cost because they require evaluating more primary features. However, some primary features have a higher computational cost than others. In general, crystal features cost more than single-molecule features. Therefore, models with similar complexity may have a different computational cost depending on the specific features they contain. It is worth noting that a *GW* + BSE@PBE calculation for a mid-sized molecular crystal with 180 atoms per unit cell may consume more than 10^6 CPU hours, which is higher than the computational cost of all the primary features by a factor of 10^4. Both $M_{1,1}$ and $M_{1,2}$ are on the Pareto front. However, $M_{1,2}$ yields a lower validation RMSE with the same computational cost. Hence, $M_{1,1}$ is not considered further for materials screening. $M_{2,3}$ and $M_{4,3}$ are on the validation set Pareto front. $M_{2,3}$ is also on the test set Pareto front. The test set RMSE for $M_{4,3}$ suggests this model may overfit the training data. Therefore, $M_{2,3}$ is selected as a second-level screening model after $M_{1,2}$.

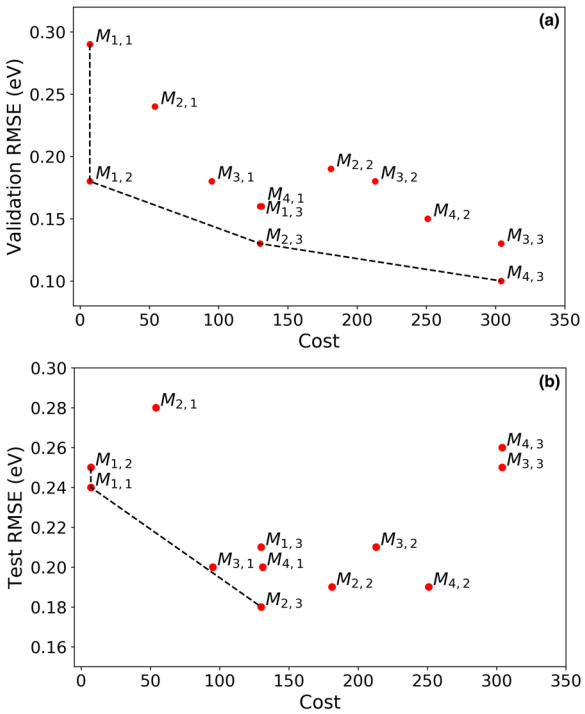

Fig. 2 Pareto charts. Pareto Chart of the accuracy, represented by the **a** training set validation RMSE and **b** unseen test set RMSE, vs. the relative computational cost of SISSO-generated models. The dashed line indicates the Pareto front.

In order to evaluate the model performance across the PAH101 training set, in particular for materials in the region of interest for SF, Fig. 3 shows correlation plots between the model prediction and the reference values of the SF driving force obtained with *GW* + BSE@PBE. A correlation plot for the baseline human-generated model, DF^S, is also shown for comparison. The correlation plots for the training set and test set for all SISSO-generated models are provided in Supplementary Figs. 1, 2. As shown in Fig. 3a, DF^S systematically underestimates the SF driving force. The SISSO-generated models are overall more predictive than the baseline human-generated model. For the models on the Pareto front, the training set RMSE gradually decreases with the model complexity. A few systems, whose molecular structures are shown in Fig. 3, consistently appear as outliers across models. The majority of the outliers comprise benzene rings connected by a single covalent bond, whereas most of the systems in the PAH101 set are conjugated aromatic compounds, in which interconnected rings share extended π-orbitals. Hence, the lower prediction accuracy for these systems may be attributed to their somewhat different chemistry. Because most of these outliers are not in the region of interest for SF, they are not a cause for concern. One outlier in the SF candidate range is the zethrene derivative 7,14-Di-*n*-butyldibenzo[*de,mn*]naphthacene (CSD reference code KAGGIK)[98]. Its SF driving force is significantly underestimated by most SISSO-generated models (except for $M_{4,3}$). Because such errors are not observed for other zethrene derivatives, we attribute this to the long alkyl side chains of KAGGIK, which make it chemically distinct from most other chromophores in the PAH101 set.

Hierarchical screening workflow

We propose a hierarchical screening approach based on different SISSO-generated models with increasing cost and accuracy to gradually narrow down the candidate pool. To select models for hierarchical screening we also consider their classification performance, shown in Table 2. Correct classification of candidate materials is important in order for the promising SF candidates to proceed to the next step of screening and the non-promising candidates to be discarded. If a false positive occurs, a material is misclassified as promising, in which case it proceeds to screening with more accurate models and may be discarded subsequently. However, if a false negative occurs, a material is misclassified as non-promising and discarded, which results in the loss of a promising candidate. Therefore, screening thresholds should be set to avoid false negatives and tolerate a small number of false positives. The hierarchical screening workflow is illustrated in Fig. 4 for the PAH101 set. The first stage of screening is performed with the low-cost model $M_{1,2}$:

$$M_{1,2} = 0.36 \times \left(\mathrm{Gap}^S - \mathrm{EA}^S\right) \times \left(\mathrm{DF}^S \times \rho^C\right) + 0.33 \quad (1)$$

$M_{1,2}$ only requires three DFT calculations for a single molecule and the crystal density, which requires no calculations, and yields an RMSE of 0.22 eV. As shown in Table 2, similar to the other SISSO-generated models, $M_{1,2}$ yields 100% sensitivity for the training set. However, one of the three additional SF candidates in the test set is not correctly classified, resulting in a sensitivity of 0.67. Both the training set and test set produce almost 100% specificity, implying high confidence in the classification of non-SF candidates. In order to correctly classify all SF materials, the selection threshold is adjusted by subtracting the model RMSE of 0.22 eV from the true positive threshold of −0.62 eV, to give a threshold of −0.84 eV. With this threshold, all 24 SF candidates in the PAH101 set and nine non-promising materials pass the first stage of screening. Thus, model $M_{1,2}$ already eliminates the vast majority of non-SF materials in the dataset.

As shown in Figure 2a, $M_{2,3}$ yields a significantly higher accuracy at a computational cost that is about 20 times higher than that of $M_{1,2}$. Equation 2 shows the features included in the model:

$$M_{2,3} = -0.35 \times \frac{\left(E_T^C + \mathrm{EA}^S\right) \times \left(E_T^S \times \rho^C\right)}{\log\left(\mathrm{AtomNum}^C\right) / \left(\mathrm{AtomNum}^C\right)^{\frac{1}{3}}} + 4.25 \times \frac{\log\left(\rho^C\right) \times \left(\mathrm{EA}^S - \mathrm{CB}^C_{\mathrm{disp}}\right)}{\mathrm{EA}^S/\mathrm{CB}^C_{\mathrm{disp}} - \mathrm{VB}^C_{\mathrm{disp}}/\mathrm{EA}^S} + 0.61 \quad (2)$$

The only single-molecule features included in $M_{2,3}$ are the electron affinity EA^S and triplet-state formation energy, E_T^S. The remaining features are crystal features, including the crystal density ρ^C, the number of atoms in the unit cell, the conduction band and valence band dispersion, $\mathrm{CB}^C_{\mathrm{disp}}, \mathrm{VB}^C_{\mathrm{disp}}$, and the triplet-state formation energy, E_T^S. $M_{2,3}$ achieves almost 100% classification accuracy for the training set. In addition, $M_{2,3}$ yields 100% on all three metrics of sensitivity, specificity, and accuracy for the test set. Based on its performance, $M_{2,3}$ is selected for the second stage of screening with a selection threshold of $-0.62-0.15 = -0.77$ eV, where 0.15 eV is the training set RMSE. We note that some materials admitted by the threshold of −0.77 eV could turn out to be promising for SF if renormalization of the exciton energies due to phonons is considered in post-processing[75]. At the second level of screening, all 24 SF candidates in the PAH101 set and four non-promising materials pass, filtering out almost half of the non-promising candidates from the first stage. Owing to the high computational cost of *GW* + BSE calculations, every non-promising material filtered out may save 10^5–10^6 CPU hours.

The variance between the predictions of different models for a given material may be used as a measure of uncertainty. Figure 5 shows the range of predictions produced by the two models selected for the hierarchical screening workflow, $M_{1,2}$ and $M_{2,3}$ for all the materials in the PAH101 set, arranged in order of increasing SF driving force from left to right. For almost 90% of the PAH101 set, the predictions of the two models are within 0.2 eV of each other. Most of the materials for which the predictions of the two models significantly diverge are outside of the promising region for SF. As shown in Fig. 5, the three materials with high

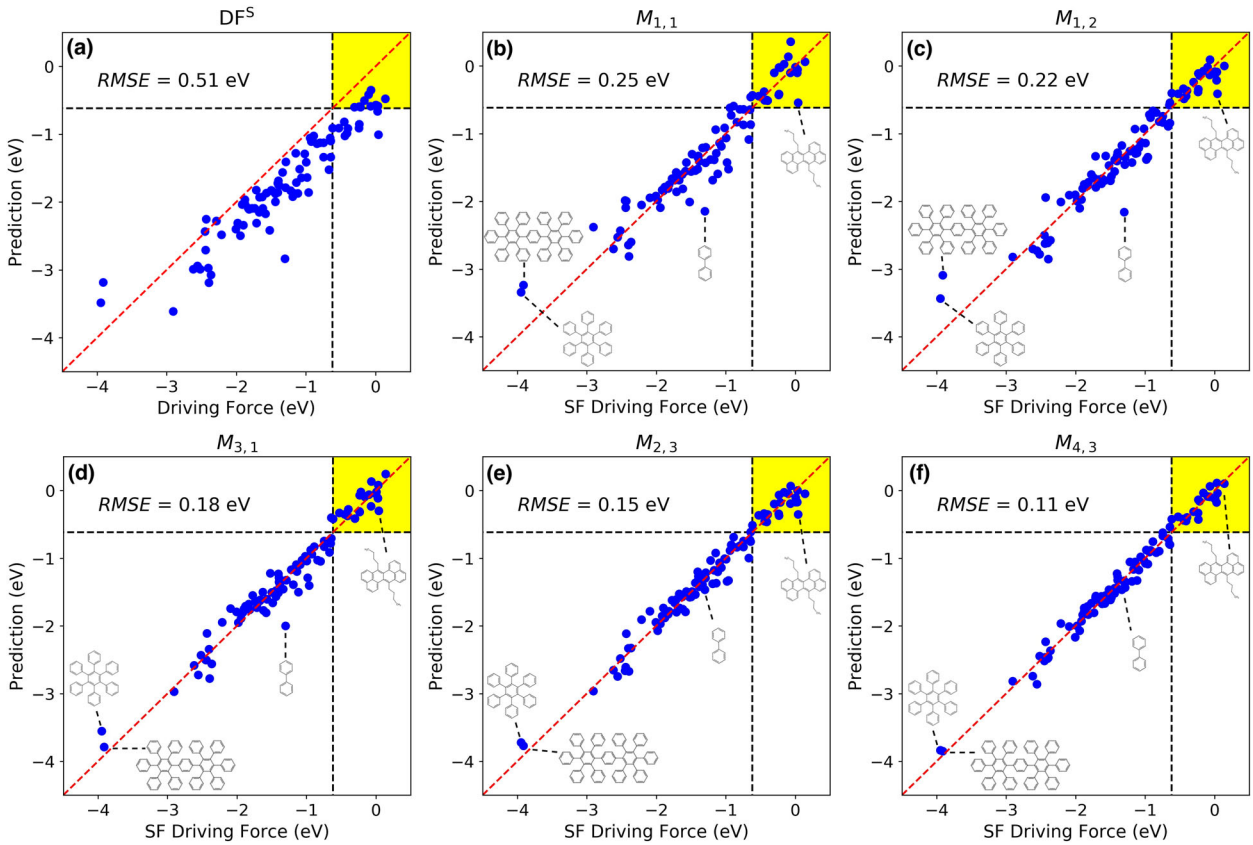

Fig. 3 Correlation plots of selected models. Model prediction as a function of the *GW* + BSE SF driving force for **a** the baseline model, DF^S, and the five models on or close to the Pareto front: **b** $M_{1,1}$, **c** $M_{1,2}$, **d** $M_{3,1}$, **e** $M_{2,3}$, and **f** $M_{4,3}$. The training set RMSE and molecular structures of four outliers are also shown. The true positive region for SF candidates is colored in yellow.

Table 2. Classification performance of the SISSO-generated models in terms of sensitivity, specificity, and accuracy with respect to the SF driving force threshold of −0.62 eV.

Model	Sensitivity		Specificity		Accuracy	
	Training	Test	Training	Test	Training	Test
DF^S	0.53	0.0	1.0	1.0	0.89	0.7
$M_{1,1}$	1.0	0.67	0.96	1.0	0.97	0.9
$M_{1,2}$	1.0	0.67	0.97	1.0	0.98	0.9
$M_{2,3}$	1.0	1.0	0.99	1.0	0.99	1.0
$M_{4,3}$	1.0	0.67	0.99	1.0	0.99	0.9

The DF^S baseline model is shown for comparison.

prediction uncertainty in the non-SF candidate region are molecules with singly-bonded benzene rings and a graphene nanoflake. Both classes are rare in the PAH101 set, leading to a high uncertainty between different models due to insufficient training data. In the SF candidate region, no significant uncertainty is observed. The improved model performance in the region of interest for SF may be attributed to the preferential selection of materials from this region for the LCV validation set. One material, the zethrene derivative 7,14-Di-*n*-butyldibenzo[*de,mn*]naphthacene (CSD reference code KAGGIK) has a relatively high prediction error. KAGGIK is a zethrene derivative with two long alkyl side groups, making it chemically distinct from most of the PAH101 set. Most of the materials with high prediction variance are the same outliers, for which the models with lower complexity have high prediction errors in Fig. 3. Within a hierarchical screening workflow, materials for which the predictions of different models significantly diverge may be selected for *GW*+BSE calculations even if they are not promising candidates for SF for the purpose of model refinement.

Promising SF candidates

Further analysis is performed, using GW+BSE, for the materials that are consistently classified as promising by the selected SISSO-generated models. For most of the promising SF candidates in the PAH101 set, including pentacene, tetracene, rubrene, quaterrylene, phenylated acenes, pyrene-fused acenes, and zethrene derivatives, detailed analyses have been published elsewhere[16,29,42–44]. Three additional promising SF candidates discovered among the materials studied here are BCPP, TBPT, and DPNP. Their crystal structures, reported in refs. [99–101], are visualized in Fig. 6. These compounds belong to chemical families of PAHs not previously explored in the context of SF. BCPP and DPNP are non-alternant PAHs containing five-membered rings fused with six-membered rings. TBPT is somewhat reminiscent of a rylene. In Fig. 7 BCPP, TBPT, and DPNP are compared to the known SF materials tetracene, rubrene, diphenyltetracene (DPT), and diphenylpentacene (DPP) with respect to a two-dimensional descriptor for SF performance[16,29,43,44]. The primary descriptor is the SF driving force, plotted on the x-axis. A high driving force indicates that a material is likely to undergo SF at a high rate. However, an overly high driving force would lead to energy losses in solar energy conversion. Therefore, a driving force between tetracene and pentacene is considered optimal.

The secondary descriptor, displayed on the *y*-axis, is the degree of charge transfer character (%CT) of the singlet exciton wave

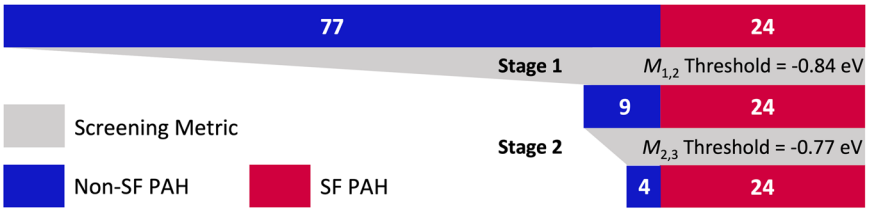

Fig. 4 Hierarchical screening workflow. Schematic of the hierarchical screening workflow based on models $M_{1\ 2}$, and $M_{2,3}$, applied to the PAH101 set.

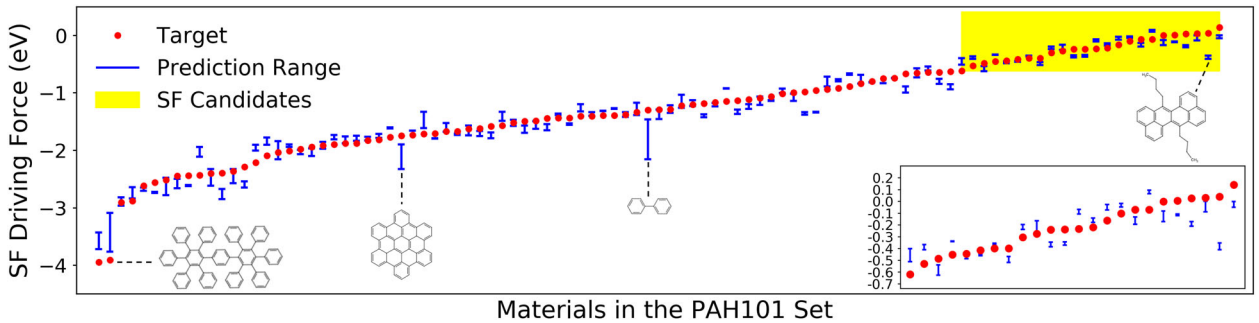

Fig. 5 Uncertainty analysis. The range of predictions produced by models $M_{1,2}$ and $M_{2,3}$ for the PAH101 set. The materials are arranged in order of increasing GW+BSE SF driving force from left to right. The red dots indicate the *GW* + BSE@PBE SF driving force, and the blue error bars represent the prediction range of the two SISSO models. The region of promising SF candidates is highlighted in yellow and magnified in the inset. Molecular structures of non-SF materials with a prediction range higher than 0.4 eV and the SF material with the highest prediction error are also shown.

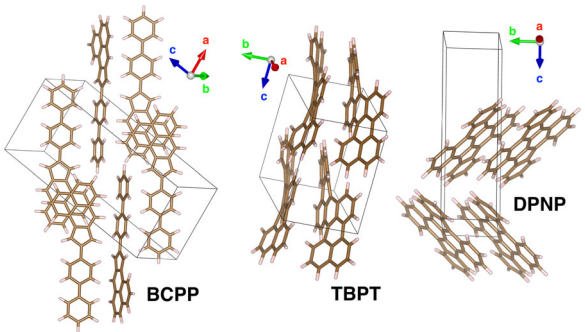

Fig. 6 Crystal structures of SF candidates. The crystal structures of BCPP, TBPT, and DPNP. The carbon and hydrogen atoms are colored in brown and white, respectively.

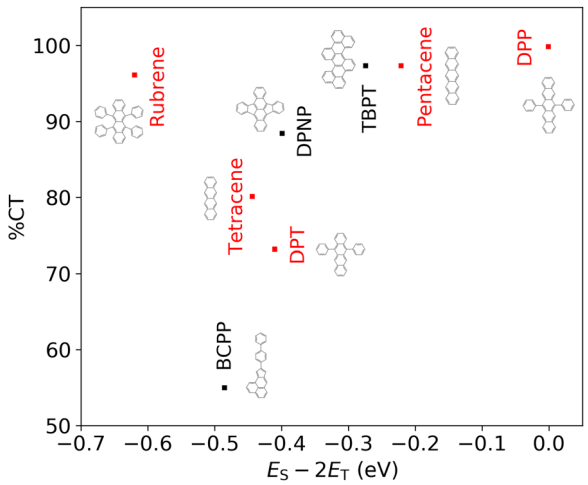

Fig. 7 2D descriptor of SF performance. BCPP, TBPT, and DPNP compared to known SF materials, colored in red, with respect to a two-dimensional descriptor calculated with GW + BSE. The thermodynamic driving force for SF ($E_S - 2E_T$) is displayed on the x-axis and the singlet exciton charge transfer character (%CT) is displayed on the y-axis.

function. This descriptor is motivated by the growing body of experimental evidence for the involvement of an intermediate charge transfer state in the SF process[4,102–105]. A singlet exciton with a high degree of charge transfer character, i.e., with the hole and the electron probability distributions centered on different molecules, is thought to be favorable for SF[4,21,106–108]. The SF driving force of BCPP is comparable to tetracene but its %CT is significantly lower. Considering the relatively slow fission rate in crystalline tetracene[109–111], slow SF could be observed in the BCPP crystal. DPNP has a comparable SF driving force to that of DPT and a much higher %CT of almost 90%. TBPT has a slightly lower SF driving force than pentacene and a comparable %CT. Based on this, DPNP and TBPT may undergo faster SF than tetracene with a smaller energy loss than pentacene.

In summary, to accelerate the computational discovery of potential materials for intermolecular singlet fission in the solid state, we have used machine learning to generate models that are fast to evaluate and accurately predict the thermodynamic driving force, which is the primary criterion for singlet fission to occur. To this end, a dataset of *GW* + BSE calculations of the SF driving force of 101 polycyclic aromatic hydrocarbons (PAH101) was compiled. The SISSO machine-learning algorithm was used to generate models with a varying degree of complexity by combining physically motivated primary features. Subsequently, the most predictive models were selected by linear regression with cross-validation.

Several SISSO-generated models demonstrated good prediction performance with a training set RMSE below 0.2 eV. The accuracy of the SISSO-generated models exceeded by far the accuracy of human generated baseline models based on DFT estimates of the single molecule and crystal SF driving force. The few outliers, most of which were outside the region of interest for SF, were

somewhat chemically different than most chromophores in the PAH101 set. Based on considerations of cost, accuracy, and classification performance we have proposed a hierarchical screening workflow comprising two SISSO-generated models with increasing cost and accuracy. Thresholds were set based on model RMSE to allow a small number of false positives while ensuring that no viable SF candidates were missed. All 24 promising SF candidates in the PAH101 set successfully passed through the workflow with only four false positives. In a materials screening scenario, *GW* + BSE calculations would be performed only for the materials that pass all stages of the SISSO-based screening. In addition, we have proposed using the variance in the predictions of different SISSO-generated models for a given material as a measure of uncertainty. A large variance in the SISSO model predictions for a certain material may indicate that it should be selected for *GW* + BSE calculations, even if it is not a promising SF candidate, for the purpose of model retraining and refinement.

Finally, three potentially promising SF materials that have not been reported previously were discovered in the PAH101 set: BCPP, TBPT, and DPNP. For these materials, further analysis was performed using *GW* + BSE. They were compared to known SF materials with respect to a two-dimensional descriptor based on the thermodynamic driving force and the singlet exciton charge transfer character. BCPP was found to have a thermodynamic driving force comparable to tetracene but a significantly lower CT character, indicating that it may undergo slow singlet fission. TBPT and DPNP were found to have a thermodynamic driving force between tetracene and pentacene and a high degree of singlet exciton CT character. This indicates that they may undergo faster SF than tetracene with a smaller energy loss (higher energy efficiency) than in pentacene. BCPP, TBPT, and DPNP belong to chemical families that have not been studied in the context of SF to date. This may help steer experimental efforts in new directions.

Thus, we have successfully used the SISSO machine-learning algorithm to find predictive models for excited-state properties of molecular crystals, whose computational cost is sufficiently low to enable large-scale screening in search of SF materials. In the future, we will use the SISSO-generated models to screen materials datasets. We note that the present models are not expected to perform well for materials that are significantly chemically different than PAHs because that would be an extrapolation. However, there are many additional PAHs in the CSD and PAH structures that continue to be solved and added at an increasing rate with the advent of 3D electron diffraction (e.g., ref. [45]). As additional data are acquired the SISSO-generated models may be retrained and refined for more chemically diverse systems. A similar approach may be used for other materials discovery efforts where properties of interest are expensive to compute or measure, making training data scarce.

METHODS

Primary feature calculation

Crystal features were evaluated for a locally-optimized geometry with the unit cell lattice vectors fixed at their experimental values. Single-molecule features were evaluated for molecules extracted from these locally-optimized crystal structures. The primary features were calculated using the FHI-aims package[112,113] with the PBE functional, tight numerical settings, and tier-2 basis sets[112]. Details of the k-point grid settings for each crystal are provided in the Supplementary Information.

SF driving force calculation

The SF driving force of crystals was calculated after full unit cell relaxation. The Quantum ESPRESSO[114] package was used to generate the mean-field eigenvalues and eigenfunctions using the PBE exchange-correlation functional with Troullier–Martins norm-conserving pseudopotentials[115]. The wave functions were generated using a kinetic energy cutoff of 50 Ry. The BerkeleyGW package[116] was used to conduct many-body perturbation theory (MBPT) calculations within the *GW* approximation and to solve the Bethe–Salpeter equation (BSE). About 550 unoccupied bands were included in the calculation of the *GW* dielectric function and self-energy operator. The static remainder correction was applied to accelerate the convergence with respect to the number of unoccupied states[117]. Twenty-four valence bands and 24 conduction bands were included in the calculation of the BSE kernel. The Tamm–Dancoff approximation (TDA) was applied when solving the BSE[116]. The coarse and fine k-point grid settings for each crystal are provided in the Supplementary Discussions.

DATA AVAILABILITY

The data are available in the Supplementary Information.

CODE AVAILABILITY

The SISSO Fortran code is available at GitHub Repository: https://github.com/rouyang2017/SISSO.

Received: 24 November 2021; Accepted: 22 March 2022;
Published online: 19 April 2022

REFERENCES

1. Smith, M. B. & Michl, J. Singlet fission. *Chem. Rev.* **110**, 6891–6936 (2010).
2. Casanova, D. Theoretical modeling of singlet fission. *Chem. Rev.* **118**, 7164–7207 (2018).
3. Rao, A. & Friend, R. H. Harnessing singlet exciton fission to break the Shockley-Queisser limit. *Nat. Rev. Mater.* **2**, 17063 (2017).
4. Monahan, N. & Zhu, X. Y. Charge transfer-mediated singlet fission. *Annu. Rev. Phys. Chem.* **66**, 601–618 (2015).
5. Smith, M. B. & Michl, J. Recent advances in singlet fission. *Annu. Rev. Phys. Chem.* **64**, 361–386 (2013).
6. Minami, T. & Nakano, M. Diradical character view of singlet fission. *J. Phys. Chem. Lett.* **3**, 145–150 (2012).
7. Lee, J. et al. Singlet exciton fission photovoltaics. *Acc. Chem. Res.* **46**, 1300–1311 (2013).
8. Ito, S., Nagami, T. & Nakano, M. Molecular design for efficient singlet fission. *J. Photochem. Photobiol. C.* **34**, 85–120 (2018).
9. Felter, K. M. & Grozema, F. C. Singlet fission in crystalline organic materials: recent insights and future directions. *J. Phys. Chem. Lett.* **10**, 7208–7214 (2019).
10. Walker, B. J., Musser, A. J., Beljonne, D. & Friend, R. H. Singlet exciton fission in solution. *Nat. Chem.* **5**, 1019–1024 (2013).
11. Xia, J. et al. Singlet fission: progress and prospects in solar cells. *Adv. Mater.* **29**, 1601652 (2017).
12. Congreve, D. N. et al. External quantum efficiency above 100% in a singlet-exciton-fission-based organic photovoltaic cell. *Science* **340**, 334–337 (2013).
13. Ehrler, B., Wilson, M. W., Rao, A., Friend, R. H. & Greenham, N. C. Singlet exciton fission-sensitized infrared quantum dot solar cells. *Nano Lett.* **12**, 1053–1057 (2012).
14. Ehrler, B. et al. In situ measurement of exciton energy in hybrid singlet-fission solar cells. *Nat. Commun.* **3**, 1019 (2012).
15. Hanna, M. C. & Nozik, A. J. Solar conversion efficiency of photovoltaic and photoelectrolysis cells with carrier multiplication absorbers. *J. Appl. Phys.* **100**, 074510 (2006).
16. Liu, X. et al. Pyrene-stabilized acenes as intermolecular singlet fission candidates: importance of exciton wave-function convergence. *J. Phys. Condens. Matter.* **32**, 184001 (2020).
17. Hummer, K., Puschnig, P. & Ambrosch-Draxl, C. Lowest optical excitations in molecular crystals: bound excitons versus free electron-hole pairs in anthracene. *Phys. Rev. Lett.* **92**, 147402 (2004).
18. Hummer, K. & Ambrosch-Draxl, C. Oligoacene exciton binding energies: their dependence on molecular size. *Phys. Rev. B* **71**, 081202 (2005).
19. Zimmerman, P. M., Bell, F., Casanova, D. & Head-Gordon, M. Mechanism for singlet fission in pentacene and tetracene: from single exciton to two triplets. *J. Am. Chem. Soc.* **133**, 19944–19952 (2011).
20. Rangel, T. et al. Structural and excited-state properties of oligoacene crystals from first principles. *Phys. Rev. B* **93**, 115206 (2016).
21. Sharifzadeh, S. et al. Relating the physical structure and optoelectronic function of crystalline TIPS-pentacene. *Adv. Funct. Mater.* **25**, 2038–2046 (2015).
22. Minami, T., Ito, S. & Nakano, M. Theoretical study of singlet fission in oligoacenes. *J. Phys. Chem. Lett.* **3**, 2719–2723 (2012).

23. Renaud, N., Sherratt, P. A. & Ratner, M. A. Mapping the relation between stacking geometries and singlet fission yield in a class of organic crystals. *J. Phys. Chem. Lett.* **4**, 1065–1069 (2013).
24. Eaton, S. W. et al. Singlet exciton fission in polycrystalline thin films of a slip-stacked perylenediimide. *J. Am. Chem. Soc.* **135**, 14701–14712 (2013).
25. Eaton, S. W. et al. Singlet exciton fission in thin films of tert-butyl-substituted terrylenes. *J. Phys. Chem. A.* **119**, 4151–4161 (2015).
26. Budden, P. J. et al. Singlet exciton fission in a modified acene with improved stability and high photoluminescence yield. *Nat. Commun.* **12**, 1527 (2021).
27. Jundt, C. et al. Exciton dynamics in pentacene thin films studied by pump-probe spectroscopy. *Chem. Phys. Lett*. **241**, 84–88 (1995).
28. Wilson, M. W. et al. Ultrafast dynamics of exciton fission in polycrystalline pentacene. *J. Am. Chem. Soc.* **133**, 11830–11833 (2011).
29. Wang, X., Liu, X., Cook, C., Schatschneider, B. & Marom, N. On the possibility of singlet fission in crystalline quaterrylene. *J. Chem. Phys.* **148**, 184101 (2018).
30. Chan, W. L., Ligges, M. & Zhu, X. Y. The energy barrier in singlet fission can be overcome through coherent coupling and entropic gain. *Nat. Chem.* **4**, 840–845 (2012).
31. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
32. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
33. 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).
34. Olivares-Amaya, R. et al. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy Environ. Sci.* **4**, 4849–4861 (2011).
35. Jacquemin, D., Wathelet, V., Perpète, E. A. & Adamo, C. Extensive TD-DFT benchmark: singlet-excited states of organic molecules. *J. Chem. Theory Comput.* **5**, 2420–2435 (2009).
36. Laurent, A. D. & Jacquemin, D. TD-DFT benchmarks: a review. *Int. J. Quantum Chem.* **113**, 2019–2039 (2013).
37. Padula, D., Omar, Ö. H., Nematiaram, T. & Troisi, A. Singlet fission molecules among known compounds: finding a few needles in a haystack. *Energy Environ. Sci.* **12**, 2412–2416 (2019).
38. Ryerson, J. L. et al. Two thin film polymorphs of the singlet fission compound 1,3-diphenylisobenzofuran. *J. Phys. Chem. C.* **118**, 12121–12132 (2014).
39. Sharifzadeh, S. Many-body perturbation theory for understanding optical excitations in organic molecules and solids. *J. Phys.: Condens. Matter* **30**, 153002 (2018).
40. Marom, N. Accurate description of the electronic structure of organic semiconductors by *GW* methods. *J. Phys. Condens. Matter* **29**, 103003 (2017).
41. 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).
42. 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).
43. Wang, X. et al. Phenylated acene derivatives as candidates for intermolecular singlet fission. *J. Phys. Chem. C.* **123**, 5890–5899 (2019).
44. Liu, X., Tom, R., Gao, S. & Marom, N. Assessing zethrene derivatives as singlet fission candidates based on multiple descriptors. *J. Phys. Chem. C.* **124**, 26134–26143 (2020).
45. Hall, C. L. et al. 3D electron diffraction structure determination of terrylene, a promising candidate for intermolecular singlet fission. *ChemPhysChem* **22**, 1631–1637 (2021).
46. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
47. Gubernatis, J. E. & Lookman, T. Machine learning in materials design and discovery: examples from the present and suggestions for the future. *Phys. Rev. Mater.* **2**, 120301 (2018).
48. Goldsmith, B. R., Esterhuizen, J., Liu, J. X., Bartel, C. J. & Sutton, C. Machine learning for heterogeneous catalyst design and discovery. *AIChE J.* **64**, 2311–2323 (2018).
49. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-driven materials science: status, challenges, and perspectives. *Adv. Sci.* **6**, 1900808 (2019).
50. Ong, S. P. Accelerating materials science with high-throughput computations and machine learning. *Comput. Mater. Sci.* **161**, 143–150 (2019).
51. Mueller, T., Kusne, A. G. & Ramprasad, R. Machine learning in materials science: recent progress and emerging applications. *Rev. Comput. Chem.* **29**, 186–273 (2016).
52. Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.* **115**, 1058–1073 (2015).
53. Janet, J. P. et al. Designing in the face of uncertainty: exploiting electronic structure and machine learning models for discovery in inorganic chemistry. *Inorg. Chem.* **58**, 10592–10606 (2019).
54. Haghighatlari, M. et al. ChemML: a machine learning and informatics program package for the analysis, mining, and modeling of chemical and materials data. *Comput. Mol. Sci.* **10**, e1458 (2020).
55. Kim, J., Kang, D., Kim, S. & Jang, H. W. Catalyze materials science with machine learning. *ACS Mater. Lett.* **3**, 1151–1171 (2021).
56. Curtarolo, S. et al. AFLOW: an automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
57. Nakata, M. & Shimazaki, T. PubChemQC project: a large-scale first-principles electronic structure database for data-driven chemistry. *J. Chem. Inf. Model.* **57**, 1300–1308 (2017).
58. Hachmann, J. et al. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry - the Harvard Clean Energy Project. *Energy Environ. Sci.* **7**, 698–704 (2014).
59. Kirklin, S. et al. The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
60. Olsthoorn, B., Matthias Geilhufe, R., Borysov, S. S. & Balatsky, A. V. Band gap prediction for large organic crystal structures with machine learning. *Adv. Quantum Technol.* **2**, 1900023 (2019).
61. Stuke, A. et al. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Sci. Data* **7**, 58 (2020).
62. Ghiringhelli, L. M. et al. Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats. *npj Comput. Mater.* **3**, 46 (2017).
63. Zheng, C. et al. Automated generation and ensemble-learned matching of X-ray absorption spectra. *npj Comput. Mater.* **4**, 12 (2018).
64. Timoshenko, J. et al. Neural network approach for characterizing structural transformations by X-ray absorption fine structure spectroscopy. *Phys. Rev. Lett.* **120**, 225502 (2018).
65. Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
66. Paruzzo, F. M. et al. Chemical shifts in molecular solids by machine learning. *Nat. Commun*. **9**, 4501 (2018).
67. 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 (2018).
68. Cao, G. et al. Artificial intelligence for high-throughput discovery of topological insulators: the example of alloyed tetradymites. *Phys. Rev. Mater.* **4**, 034204 (2020).
69. Bartel, C. J. et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **5**, eaav0693 (2019).
70. Andersen, M., Levchenko, S. V., Scheffler, M. & Reuter, K. Beyond scaling relations for the description of catalytic materials. *ACS Catal.* **9**, 2752–2759 (2019).
71. Bartel, C. J. et al. Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry. *Nat. Commun.* **9**, 4168 (2018).
72. Foppa, L. et al. Materials genes of heterogeneous catalysis from clean experiments and artificial intelligence. *MRS Bull.* **46**, 1016–1026 (2021).
73. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge structural database. *Acta Cryst.* **72**, 171–179 (2016).
74. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
75. Alvertis, A. M. et al. Impact of exciton delocalization on exciton-vibration interactions in organic semiconductors. *Phys. Rev. B* **102**, 081122 (2020).
76. Thorsmølle, V. K. et al. Morphology effectively controls singlet-triplet exciton relaxation and charge transport in organic semiconductors. *Phys. Rev. Lett.* **102**, 017401 (2009).
77. Schulze, T. F. & Schmidt, T. W. Photochemical upconversion: present status and prospects for its application to solar energy conversion. *Energy Environ. Sci.* **8**, 103–125 (2015).
78. Cheng, Y. Y. et al. Kinetic analysis of photochemical upconversion by triplet-triplet annihilation: beyond any spin statistical limit. *J. Phys. Chem. Lett.* **1**, 1795–1799 (2010).
79. Wolf, E. A., Finton, D. M., Zoutenbier, V. & Biaggio, I. Quantum beats of a multiexciton state in rubrene single crystals. *Appl. Phys. Lett.* **112**, 083301 (2018).
80. Ma, L. et al. Singlet fission in rubrene single crystal: direct observation by femtosecond pump-probe spectroscopy. *Phys. Chem. Chem. Phys.* **14**, 8307–8312 (2012).
81. Simon, Y. C. & Weder, C. Low-power photon upconversion through triplet-triplet annihilation in polymers. *J. Mater. Chem.* **22**, 20817–20830 (2012).
82. Singh-Rachford, T. N. & Castellano, F. N. Photon upconversion based on sensitized triplet-triplet annihilation. *Coord. Chem. Rev.* **254**, 2560–2573 (2010).

83. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Mater.* **114**, 105503 (2015).
84. Arias, D. H., Ryerson, J. L., Cook, J. D., Damrauer, H. & Johnson, J. C. Polymorphism influences singlet fission rates in tetracene thin films. *Chem. Sci.* **7**, 1185–1191 (2016).
85. Bhattacharyya, K. & Datta, A. Polymorphism controlled singlet fission in TIPS-anthracene: role of stacking orientation. *J. Phys. Chem. C.* **121**, 1412–1420 (2017).
86. Wang, L., Olivier, Y., Prezhdo, O. V. & Beljonne, D. Maximizing singlet fission by intermolecular packing. *J. Phys. Chem. Lett.* **5**, 3345–3353 (2014).
87. Armstrong, Z. T., Kunz, M. B., Jones, A. C. & Zanni, M. T. Thermal annealing of singlet fission microcrystals reveals the benefits of charge transfer couplings and slip-stacked packing. *J. Phys. Chem. C.* **124**, 15123–15131 (2020).
88. Dillon, R. J., Piland, G. B. & Bardeen, C. J. Different rates of singlet fission in monoclinic versus orthorhombic crystal forms of diphenylhexatriene. *J. Am. Chem. Soc.* **135**, 17278–17281 (2013).
89. Buchanan, E. A. et al. Molecular packing and singlet fission: the parent and three fluorinated 1,3-diphenylisobenzofurans. *J. Phys. Chem. Lett.* **10**, 1947–1953 (2019).
90. Feng, X., Kolomeisky, A. B. & Krylov, A. I. Dissecting the effect of morphology on the rates of singlet fission: Insights from theory. *J. Phys. Chem. C.* **118**, 19608–19617 (2014).
91. Sutton, C., Tummala, N. R., Beljonne, D. & Brédas, J. L. Singlet fission in rubrene derivatives: impact of molecular packing. *Chem. Mater.* **29**, 2777–2787 (2017).
92. 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 (2012).
93. Hammouri, M. et al. High-throughput pressure-dependent density functional theory investigation of herringbone polycyclic aromatic hydrocarbons: part 2. Pressure-dependent electronic properties. *J. Phys. Chem. C.* **122**, 2838–2844 (2018).
94. Marom, N., Körzdörfer, T., Ren, X., Tkatchenko, A. & Chelikowsky, J. R. Size effects in the interface level alignment of dye-sensitized TiO_2 clusters. *J. Phys. Chem. Lett.* **5**, 2395–2401 (2014).
95. Kunkel, C., Schober, C., Margraf, J. T., Reuter, K. & Oberhofer, H. Finding the right bricks for molecular legos: a data mining approach to organic semiconductor design. *Chem. Mater.* **31**, 969–978 (2019).
96. Yu, M. et al. Anomalous pressure dependence of the electronic properties of molecular crystals explained by changes in intermolecular electronic coupling. *Synth. Met.* **253**, 9–19 (2019).
97. Schober, C., Reuter, K. & Oberhofer, H. Critical analysis of fragment-orbital DFT schemes for the calculation of electronic coupling values. *J. Chem. Phys.* **144**, 054103 (2016).
98. Wu, T.-C. et al. Synthesis, structure, and photophysical properties of dibenzo[*de,mn*]naphthacenes. *Angew. Chem. Int. Ed. Engl.* **122**, 7213–7216 (2010).
99. Shea, K. M., Lee, K. L. & Danheiser, R. L. Synthesis and properties of 9-alkyl- and 9-arylcyclopenta[*a*]phenalenes. *Org. Lett.* **2**, 2353–2356 (2000).
100. Izuoka, A., Wakui, K., Fukuda, T., Sato, N. & Sugawara, T. Refined molecular structure of tetrabenzo[*de,hi,op,st*]pentacene. *Acta Cryst.* **48**, 900–902 (1992).
101. Bennett, A. & Hanson, A. W. The structure of diphenylene naphthacene. *Acta Cryst.* **6**, 736–739 (1953).
102. Kim, V. O. et al. Singlet exciton fission via an intermolecular charge transfer state in coevaporated pentacene-perfluoropentacene thin films. *J. Chem. Phys.* **151**, 164706 (2019).
103. Miyata, K., Conrad-Burton, F. S., Geyer, F. L. & Zhu, X. Y. Triplet pair states in singlet fission. *Chem. Rev.* **119**, 4261–4292 (2019).
104. Margulies, E. A. et al. Direct observation of a charge-transfer state preceding high-yield singlet fission in terrylenediimide thin films. *J. Am. Chem. Soc.* **139**, 663–671 (2017).
105. Chan, W. L. et al. The quantum coherent mechanism for singlet fission: experiment and theory. *Acc. Chem. Res.* **46**, 1321–1329 (2013).
106. Sharifzadeh, S., Darancet, P., Kronik, L. & Neaton, J. B. Low-energy charge-transfer excitons in organic solids from first-principles: the case of pentacene. *J. Phys. Chem. Lett.* **4**, 2197–2201 (2013).
107. Broch, K. et al. Robust singlet fission in pentacene thin films with tuned charge transfer interactions. *Nat. Commun.* **9**, 954 (2018).
108. Hart, S. M., Silva, W. R. & Frontiera, R. R. Femtosecond stimulated Raman evidence for charge-transfer character in pentacene singlet fission. *Chem. Sci.* **9**, 1242–1250 (2018).
109. Burdett, J. J., Müller, A. M., Gosztola, D. & Bardeen, C. J. Excited state dynamics in solid and monomeric tetracene: the roles of superradiance and exciton fission. *J. Chem. Phys.* **133**, 144506 (2010).
110. Burdett, J. J. & Bardeen, C. J. The dynamics of singlet fission in crystalline tetracene and covalent analogs. *Acc. Chem. Res.* **46**, 1312–1320 (2013).
111. Wilson, M. W. B. et al. Temperature-independent singlet exciton fission in tetracene. *J. Am. Chem. Soc.* **135**, 16680–16688 (2013).
112. Blum, V. et al. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **180**, 2175–2196 (2009).
113. Havu, V., Blum, V., Havu, P. & Scheffler, M. Efficient O(N) integration for all-electron electronic structure calculation using numeric basis functions. *J. Chem. Phys.* **228**, 8367–8379 (2009).
114. Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).
115. Troullier, N. & Martins, J. L. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B* **43**, 1993–2006 (1991).
116. 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).
117. Deslippe, J., Samsonidze, G., Jain, M., Cohen, M. L. & Louie, S. G. Coulomb-hole summations and energies for *GW* calculations with limited number of empty orbitals: a modified static remainder approach. *Phys. Rev. B* **87**, 165124 (2013).

ACKNOWLEDGEMENTS

We thank Dr. Runhai Ouyang from Shanghai University for his support in training SISSO and interpreting the results. We thank Dr. Volker Blum and Dr. Yi Yao from Duke University, and Dr. William Paul Huhn from Argonne National Laboratory for their support on DFT calculations with FHI-aims. Work at CMU was supported by the National Science Foundation (NSF) Division of Materials Research through grant DMR-2021803. 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 US Department of Energy, under Contract DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS

X.L. performed part of the calculations and the SISSO training, collected, and analyzed the data. X.W., S.G., V.C., R.T., and M.Y. performed part of the calculations. L.M.G. advised on data analysis and results interpretation. N.M. led the project. All authors contributed to writing the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-022-00758-y.

Correspondence and requests for materials should be addressed to Noa Marom.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022