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]

[1] Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

[2] Qingdao Institute for Theoretical and Computational Science, 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

Abstract

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.
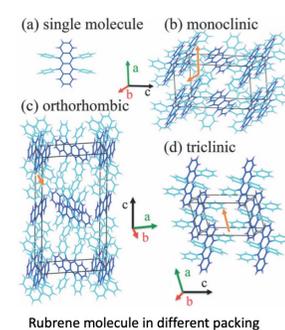
Introduction

Molecule 1 Molecule 2 Molecule 1 Molecule 2
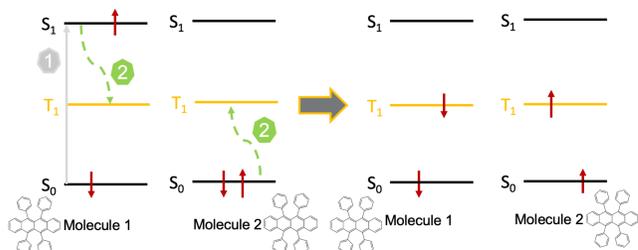
- Singlet fission is a spin-allowed process in which an organic chromophore in an excited singlet state shares its excitation energy with a neighboring ground-state chromophore and both are converted into triplet excited states. And it is thought to potentially break the Shockley Queisser limit of Solar Cells.
- 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.
- Primary criterion for SF: The thermodynamic driving force. The energy difference between the initial singlet state and final state of two triplets ($E_S - 2E_T$)
- PAH101 dataset: A set of 101 PAH crystal structures extracted from the Cambridge Structural Database (CSD). The systems in the PAH101 set represent diverse chemical families within the larger PAH class.
- In summary, to accelerate the computational discovery of potential materials for intermolecular SF in the solid state, we have used machine learning to generate models that are fast to evaluate and accurately predict the thermodynamic driving force.
- The SISSO machine-learning algorithm was used to generate models with a varying degree of complexity by combining physically motivated primary features, the most predictive models were selected by linear regression with cross validation.

Materials

Organic Molecular Crystal

(a) single molecule (b) monoclinic

(c) orthorhombic (d) triclinic

Rubrene molecule in different packing

Molecules (as motif) are packed with relatively weak intermolecular binding

Advantage to undergo SF:
- Molecular orbitals evolve into dispersed bands, excitations are easier to happen
- Excitations may be delocalized over many molecules, facilitate the exciton coupling

Wang, Xiaopeng, et al. "Effect of crystal packing on the excitonic properties of rubrene polymorphs." CrystEngComm18.38 (2016): 7353-7362.

Methodology

The primary features are calculated at DFT@PBE level. We used a system with 62 atoms per molecule as basic unit. The relative cost are multiples of basic units:

Feature Name	Description	Relative Cost
Gap^S	HOMO-LUMO gap	1
E_T^S	Triplet formation energy	3
DF^S	Driving force (DFT)	4
IP^S	Ionization potential	3
EA^S	Electron affinity	3
Gap^C	Crystal band gap	32
E_T^C	Triplet formation energy	92
DF^C	Driving force (DFT)	124
VB_{disp}^C	Valence-band dispersion	32
CB_{disp}^C	Conduction-band dispersion	32
H_{ab}	Transition matrix	88
$PolarTensor^S$	Polarization tensor	50
ϵ^C	Dielectric constant	130
$MolWt^S$	Molecular weight	0
ρ^C	Crystal density	0
$AtomNum^C$	Atom num in unit cell	0

Build Model: SISSO
(1) Feature Combination

$$\Phi_n \equiv \bigcup_{i=1}^{n} \hat{H}^{(m)}[\phi_1, \phi_2], \quad \forall \phi_1, \phi_2 \in \Phi_{i-1}$$

- Operation set: H = {$+, -, \times, \div, exp, log, (\square)^{-1,2,3}, \sqrt{\square}, \sqrt[3]{\square}, |\square|$}

(2) Linear Regression $M_{dim, Rung}$
Hyperparameters to restrict the model:
- Rung: Limit of number of feature combination (<=3)
- Dim (n): Dimension of final model (<=4)

(3) Leave-N-out Cross Validation (LCV)
The aim of this process is to prevent overfitting. We permutate N structures from the training set as unseen data and use them as validation sets to evaluate the model performance.

Results

The SF driving force in the PAH101 dataset are calculated by MBPT GW+BSE. A threshold of driving force is set to be -0.62 eV. The red structures are considered as SF candidates:
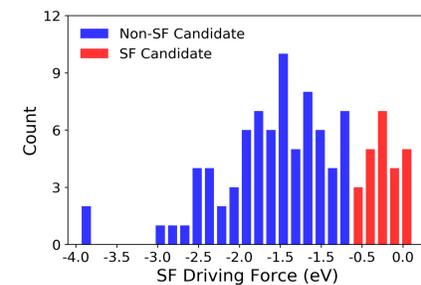
Cost Accuracy trade-off:
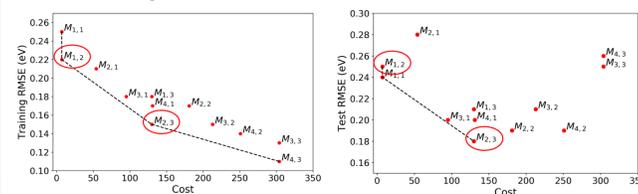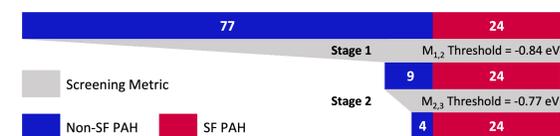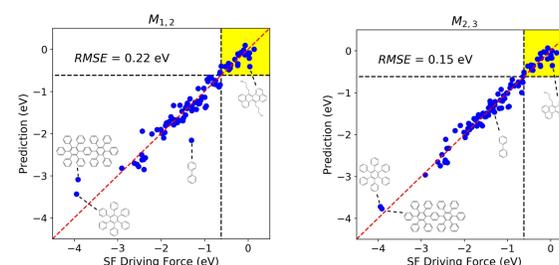

Hierarchical screening approach:
- Stage 1:

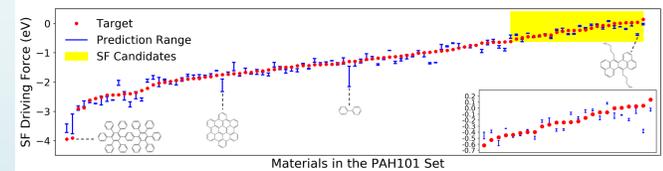$$M_{1,2} = 0.36 \times (Gap^S + EA^S) \times (DF^S \times \rho^C) + 0.33$$

- Stage 2:

$$M_{2,3} = -0.35 \times \frac{(E_T^C + EA^S) \times (E_T^S \times \rho^C)}{\log(AtomNum^C)/(AtomNum^C)^{\frac{1}{3}}} + 4.25 \times \frac{\log(\rho^C) \times (EA^S - CB_{disp}^C)}{EA^S/CB_{disp}^C - VB_{disp}^C/EA^S} + 0.61$$

Model performance:

The best model we have ($M_{2,3}$) is giving prediction with RMSE of 0.15 eV, screening out 73 non-SF structures and keep all 24 SF candidates

The range of predictions produced by Models M1,2 and M2,3 for the PAH101 dataset:

The materials are arranged in order of increasing SF driving force from left to right. The red dots indicate the GW+BSE@PBE SF driving force, 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 prediction range higher than 0.4 eV and the SF material with highest prediction error are shown.

Conclusion

- We have successfully used the SISSO machine-learning algorithm to find predictive models for excited state properties of molecular crystals using small amount of data.
- The accuracy of the SISSO-generated models demonstrated good performance with training RMSE below 0.2 eV, exceeded by far the accuracy baseline models based on DFT estimates of the single molecule and crystal SF driving force.
- The hierarchical screening approach downsized the 101-structure set to 28 but kept all 24 SF candidates.
- 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.

Acknowledgements