MATDAT18: Materials and Data Science Hackathon

Team Composition (2 people max.)

Name	**Department**	**Institution**	**Email**
Noa Marom	MSE	Carnegie Mellon	nmarom@andrew.cmu.edu
Xingyu (Alfred) Liu	MSE	Carnegie Mellon	xingyul1@andrew.cmu.edu

Project Title

Finding Predictive Descriptors for Singlet Fission: Revealing Fundamental Physics in Data

Project Synopsis (approx. 100 words)

Singlet fission (SF) is the spontaneous conversion of one photogenerated singlet exciton into two triplet excitons. SF has the potential to significantly increase the efficiency of organic solar cells by harvesting two charge carriers from one photon. However, the realization of SF-based solar cells is hindered by dearth of suitable materials. Due to the high computational cost of excited-state quantum mechanical calculations, predictive descriptors that are fast to evaluate must be found in order to explore the chemical space in search of new SF materials.

Identified Data-Science Collaborative Need (approx. 100 words)

The project requires data science expertise in feature selection, machine learning, and multi-fidelity approaches. We would like to collaborate with data scientists on using feature selection methods to search for the lowest dimensional feature vectors that best predict the thermodynamic driving force for singlet fission. We would like to assess the performance of different feature selection algorithms with respect to the feature space size and possibly explore using multi-fidelity methods to evaluate features at varying degrees of accuracy and computational cost.

Data Origin and Access (*data must be available and sharable with data science teams* – please address: data source/origin, access privileges, sharing privileges)

All data is generated by us computationally and will be made available to the data science team if our project is selected. The data includes a reference set of singlet and triplet excitation energies calculated by many-body perturbation theory for ~100 molecular crystals and a feature space. Primary features are calculated by semi-local density functional theory (DFT). These primary features can be combined by mathematical operations (linear and non-linear) into secondary features to form larger feature spaces.

Project Description (approx. 1.5 pages, plus figures and references; please describe data size, form, dimensionality, uncertainties, number of examples, etc.)

The proposed research is related to the materials science topics of: Inverse data-driven materials design, Machine learning of quantitative structure property relationship (QSPR) models, Identifying descriptors of materials performance, and Materials discovery in large-scale databases. It requires data science expertise in feature selection, machine learning, and multi-fidelity approaches.

Singlet fission (SF) is the spontaneous conversion of one photogenerated singlet exciton into two triplet excitons. Intermolecular SF occurs in crystalline media, where it is mediated by coupling between chromophores in the excited state. Recently, there has been a surge of interest in SF thanks to its potential to significantly increase the efficiency of organic solar cells by harvesting two charge carriers from one photon. However, few materials are presently known to exhibit intermolecular SF with high efficiency, hindering the realization of solid-state SF solar cells. The chemical compound space of possible chromophores is infinitely vast and largely unexplored, as most research to date has focused on restricted classes of molecules, primarily acene derivatives. Exploring this configuration space by experimental means alone would be unfeasible. Therefore, we propose to develop a data driven approach to enable computational discovery of SF materials through computer simulations. This will enable a priori design of SF materials, leading to paradigm shifts in this emerging concept for renewable energy.

SF is a collective many-body quantum mechanical process, involving electrons and holes whose correlated wave-functions may extend over several molecules. The mechanism of SF and the factors governing its efficiency are only partially understood. This calls for a new paradigm for materials discovery in a domain with incomplete physical models. Obtaining data on properties associated with electronic excitations from simulations based on many-body perturbation theory is time consuming and obtaining data on SF efficiency from experimental

measurements is even more so. A central challenge is thus to identify descriptors that are fast to evaluate and correlate strongly with high SF efficiency.

We propose to conduct a computer experiment, in which thousands of hypotheses, represented by primary descriptors combined via mathematical operations, will be generated and evaluated in an unbiased way. Structural features and other ground state descriptors that are fast to evaluate and correlate strongly with SF efficiency will be revealed by applying ML algorithms for feature selection to a purpose-built first-principles dataset. The descriptors we identify will then be used to make experimentally verifiable predictions of new SF materials from previously unexplored chemical families.

Our target property is the thermodynamic driving force for SF, represented by the difference between the singlet energy and twice the triplet energy (E_S-$2E_T$). Reference values of these excitation energies may be calculated by using many-body perturbation theory within the GW approximation and the Bethe-Salpeter Equation (BSE). GW+BSE is currently the state-of-the-art method for calculating excitonic properties of periodic systems in the size range of few hundred atoms. These methods cannot be used for high-throughput screening of large datasets of materials due to their high computational cost. Our goal is to find low-cost features that are predictive of the target property. To this end, we have generated a reference dataset of GW+BSE calculations of the singlet and triplet energies of ~100 molecular crystals.

In addition, we have generated a feature space of descriptors that correspond to possibly relevant physical properties. The basis of the feature space is a set of physically motivated primary features, which may be relevant based on our current understanding of SF. To generate a space of ~10^3-10^4 features, primary features are combined through linear and non-linear operations. At the workshop, we would like to collaborate with data scientists on using feature selection methods to search for the lowest dimensional feature vectors that best predict the target property. We would like to assess the performance of different feature selection algorithms with respect to the feature space size and possibly explore using multi-fidelity methods to evaluate features at varying degrees of accuracy and computational cost.