MATDAT18: Materials and Data Science Hackathon

Team Composition (2 people max.)

Name	Department	Institution	Email
Stanislav Borysov	Department of Management Engineering	DTU	stabo@dtu.dk
Bart Olsthoorn	Condensed Matter, Statistical and Biological Physics	NORDITA	bartol@kth.se

Project Title

Computational discovery of novel organic metals and narrow-gap semiconductors with generative models

Project Synopsis (approx. 100 words)

The main goal of this project is to predict novel stable crystal structures using the recent advances in generative modeling from the machine learning field, bypassing computationally demanding first-principles optimization calculations. Particular focus is made on prediction of organic crystals with metallic and semiconducting properties. Finding such materials in the organic space, which is mainly populated by wide-gap insulators, is a highly non-trivial task. At the same time, these functional materials offer promising technological applications in electronics and offer a basis for environmentally friendly technologies. The electronic structure data from the Organic Materials Database (OMDB) [https://omdb.diracmaterials.org/] will be available for training.

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

Prediction of novel stable crystal structures and compounds is an extremely demanding task in terms of computational resources. For example, finding stable configuration for a single compound itself involves many rounds of solving first-principles equations, not mentioning exploring new compounds. To accelerate this search, we propose to employ recent striking advances in generative modelling from the machine learning field. Particularly, we aim to focus on variational autoencoders (VAE) and generative adversarial networks (GAN) which are capable of generating new high-quality samples from the estimated distribution of the data. We anticipate that an approach based on generative methods initiates a novel direction in predicting stable crystal structures with specific functional properties that stimulates further computational and experimental investigation.

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

1. Approximately 25,000 crystal structures (in the form of CIF files) and calculated band gaps (CSV file) from the Organic Materials Database [https://omdb.diracmaterials.org/] will be fully available for the data science teams and public open access.
2. Additional organic crystal structures (more than 300,000) can be obtained from the open-access Crystallography Open Database (COD) [http://www.crystallography.net/cod/].

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

Organic narrow-gap semiconductors and metals offer promising technological applications in organic-based electronics, such as flexible OLED based displays or photovoltaics. However, less than 5% of randomly chosen organic crystals exhibit the necessary electronic properties for such specific applications. Therefore, finding materials candidates is a highly non-trivial task, often relying on time-consuming first-

principles calculations. In 2016, the Nordic Institute for Theoretical Physics (Nordita) launched an Organic Materials Database (OMDB) [https://omdb.diracmaterials.org/] hosting electronic structure properties of about 25,000 organic crystals [1]. Of all the materials contained in the OMDB, about 1000 of those are organic semiconductors (or metals) with a band gap between 0 and 1 eV. The goal of this project is to identify more novel crystal structures to expand this set of 1000 materials. The OMDB data provided will consist of CIF files, containing positions of atoms and symmetry properties of crystals, and a list of band gaps (a single float number for each material) for these materials calculated using density functional theory (DFT). Additionally, more than 300,000 previously synthesized organic crystal structures in the Crystallography Open Database (COD) [2] are available as examples of stable crystal structures. All crystal structure files are stored in the standard CIF text file format (around 10 kB each) that is straightforward to load with for example Pymatgen (Python package).

The search space of all possible spatial arrangements of atoms that make up stable crystal structures is enormous. Currently, in order to verify the stability of a structure, first-principles calculations are performed on a trial-and-error basis. As a result, the design of novel materials with certain key functionalities is of high computational demand. In contrast, developing a statistical intuition of the search space allows for a quicker identification of similar materials as well as the characterization of novel materials. The main goal of this project is to apply current exciting progress in the machine learning field to initiate a new approach for functional materials design.

Nowadays, we witness striking progress in generative models, which are designed to generate new, unobserved data points governed by the same distribution as data. Such generative models as a Variational AutoEncoder (VAE) [3] and Generative Adversarial Networks (GANs) [4] have been successfully applied in many research areas including physics (such models are implemented in most of the modern deep learning frameworks). However, the most vivid examples are usually given by the ability of GANs to generate photo-realistic images. The first step of this project is to use these machine learning methods to predict novel stable crystal structures, bypassing time-consuming first-principles calculations. Next, the output of the models will be conditioned on the band gap size (key functionality) using the narrow-band organic crystal structures from the OMDB. Combining these target examples from the OMDB and generative models will open a path towards the search for novel organic materials with desired properties.

Before applying machine learning methods to the crystal structures, a suitable representation of the data (i.e. atom positions within the unit cell) has to be chosen. Choosing a suitable representation of an arbitrary crystal structure is a known non-trivial task. The description should ideally be unique and not depend on the arbitrary choice of the coordinate system. In the past 10 years, different approaches have been proposed, so several representations can be considered for this project (some useful methods for crystal structures are listed in [4]). Alternatively, inventing a new representation scheme for crystal structures that leverages symmetry information (i.e. the space group) would be extremely valuable.

As a choice for generative models, there are several options that can be investigated, such as GANs and VAEs. The main idea behind these methods is to represent data in some linear latent space of random variables using a sequence of non-linear transformations, for example, an artificial neural network. We intend to approach the problem of finding stable organic semiconductors and metals by first training a GAN on all the organic crystal structures in the COD and afterwards finding similar materials in the output to the target “semiconducting” from the OMDB. For this approach, the large amount of crystal structures in the COD helps to find stable structures and the gap information from the OMDB can be used for the identification of desired properties.

[1] “Organic Materials Database: an open-access online database for data mining”, S.S. Borysov, R.M. Geilhufe, A.V. Balatsky, PLoS ONE 12(2), e0171501 (2017)
[2] “Crystallography Open Database”, http://www.crystallography.net/cod/
[3] "Tutorial on Variational Autoencoders", C. Doersch, arXiv preprint 1606.05908 (2016)
[4] "Generative Adversarial Nets", I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, In *Advances in neural information processing systems*, p. 2672 (2014)
[5] "How to represent crystal structures for machine learning: Towards fast prediction of electronic properties", K.T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K.R. Müller, E.K.U. Gross, Phys. Rev. B 89 (2014)