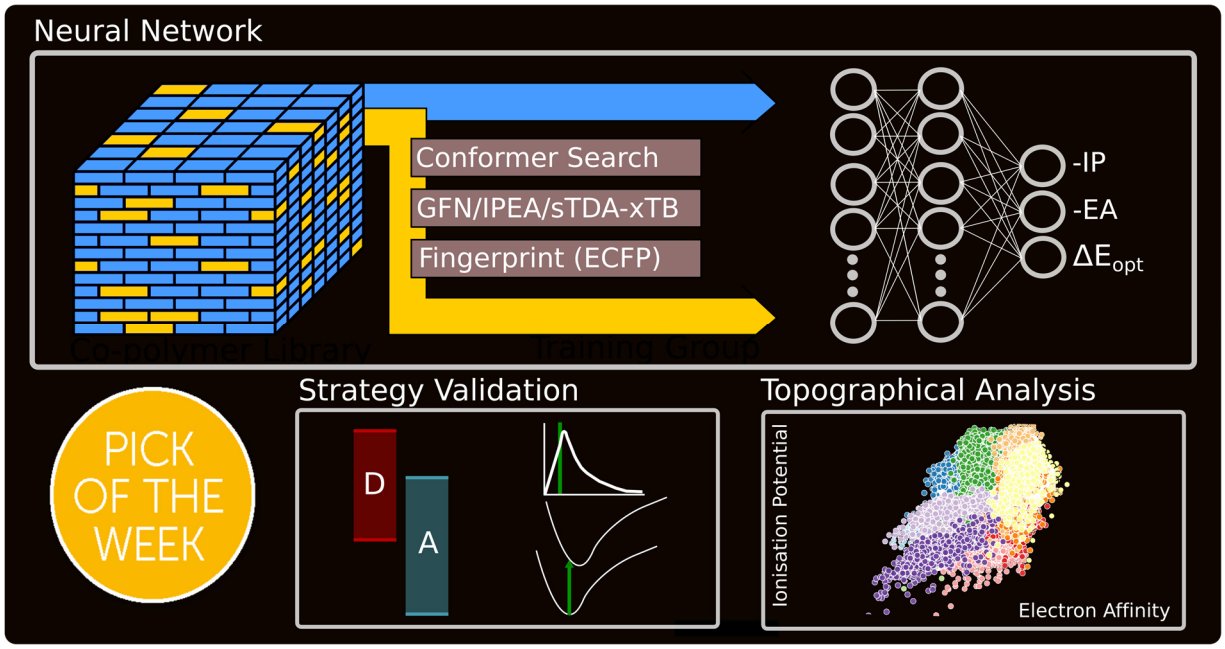
EMMC case study:

Mapping Binary Copolymer Property Space with Neural Networks

Interview of Dr Martijn Zwijnenburg, University College London

Writers: Alexandra Simperler and Gerhard Goldbeck



About Martijn Zwijnenburg

Dr Martijn Zwijnenburg (https://www.zwijnenburg-group.org/) is an associate professor at University College London. Dr Zwijnenburg's group works primarily on modelling the photo- and electrochemistry of materials, photocatalysis and understanding the structures of self-assembled and polymeric materials. He collaborates closely with several experimental groups to complement his theoretical results with experimental verification and thus, design better materials for important application such as renewable hydrogen production through photocatalytic water splitting. He is also interested in structure prediction methods, such as basin-hopping, that allow one to predict the lowest energy structures for a material (periodic solid, nanoparticle or polymer) given only its composition. He has published more than 80 articles in international journals and co-edited a book on the computational modelling of inorganic nanomaterials.

Essentials about data and modelling

During his PhD, Dr Zwijnenburg worked on understanding the structure-property relationships for zeolites [1] exploiting a database of hypothetical zeolites [2]. This may be seen as his first contact with data-driven modelling and was rather limited by only having structural data and energies and an arsenal of scripts to elucidate structure-activity relations. Also, the amount of available structures would often permit only a small data subspace to search for interesting relations. The advent of

more sophisticated semi-empirical methods permits an addition of electronic data such as ionisation potentials, electron affinities, excitation energies, band gaps and exciton binding energies. Using a combination of such semi-empirical methods and machine learning, Dr Zwijnenburg's group could calculate the properties of over 340,000 polymer structures based on commercially available monomers and on monomers suggested by their network of experimental collaborators. These data are available via a GitHub repository, https://github.com/ZwijnenburgGroup/2019-polymer-neural-network. The structures are stored as SMILES as fingerprints are too big. A Python module, pychemlp [3], for recreating Extended-Connectivity (Morgan) Fingerprints [4] is available. Besides GitHub, data are also provided with the electronic supplementary material of the respective publication and they are licensed as Creative Commons Attribution 3.0. The data are provided as .csv files as they are easy to use with Python. For this case, xtb (an extended tight-binding semi-empirical program package [5]) was used to provide relevant electronic properties.

About the Case Study

The case is based on the paper "Mapping binary copolymer property space with neural networks" L. Wilbraham, R.S. Sprick, K.E. Jelfs, M.A. Zwijnenburg, Chem. Sci., 2019,10, 4973-4984. DOI: 10.1039/C8SC05710A

Dr Zwijnenburg and Dr Liam Wilbraham, a very talented postdoc working in his group that has since moved to the University of Glasgow, applied machine learning techniques to the optoelectronic property landscape of conjugated organic copolymers and could test very diverse monomer compositions. The trained neural network can link the properties of the constituent monomers of a copolymer to the properties of the copolymer itself. The analysis of the search space revealed that for a copolymer to have a significantly smaller optical gap than its related homopolymers, the potentials of these should be substantially offset and arranged in a staggered fashion. Thus, this work proposes conditions that greatly enhance the likelihood of its experimental realisation. They also identified promising monomers which target specific regions of the property space, which is relevant to a variety of applications, such as organic photovoltaics, light emitting diodes, and thermoelectrics.

For this particular case, which were your objectives?

We wanted to work with a large data set to learn about trends and get a feeling for what optoelectronic properties of copolymers are possible, and to understand the range of properties that can be realised. The size of the data set was also important in order for our analysis to not be biased.

For this Case study did you create and/or apply a data-based model

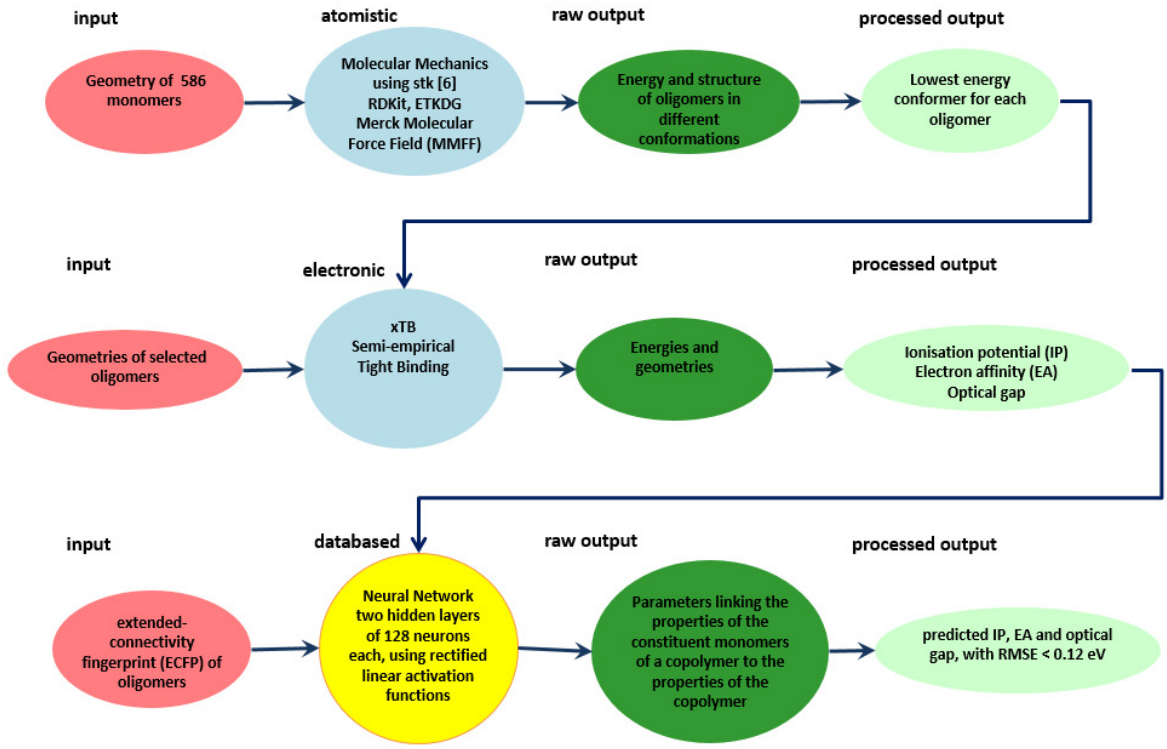
We used Machine Learning to create a data-based model. We extract general features of our property space that would otherwise be obscured in smaller datasets. We could identify simple models that effectively relate the properties of these copolymers to the homopolymers of their constituent monomers, and thus, challenge common ideas behind copolymer design.

How did data play a key role in problem solving?

Machine learning in general requires a large amount of data and what is more, can handle them. This enabled us to go beyond a small subset of data and we could find trends and also verify if trends are really there.

What methodologies have been applied?

We used stk, the supramolecular toolkit [6], to construct polymer models, xtb to calculate their properties, RDKit to calculate the fingerprints, TensorFlow to implement the neural network in, while the whole workflow was implemented in Python.



What were the expected improvements by adding data to your modelling?

In materials sciences, we have for a long time been able to create structural data, as you can see from the example of Mike Tracey and his database of Hypothetical Zeolites [2]. We, however, were limited at the time to the subspace of hypothetical zeolites we could study and worked with the limited properties we could calculate, such as structural parameters and energies obtained with atomistic models. By using semi-empirical methods and machine learning we can now calculate many more properties that are unattainable from atomistic calculations, e.g. materials' optoelectronics properties, and even in principle consider chemistry, i.e. bond breaking and formation.

For this particular case, did you have to invest a lot of work to make the data usable?

We had to get monomer data from different sources such as chemical catalogues, and asked our experimental network, what was missing, what should be considered. It took some work to merge and collect these data. Liam, had to investigate and learn about best practises to make this data ready for machine learning. These best practices comprise how to generate data for machine learning, i.e. the fingerprints, and how to choose training- and test-sets for the actual training of our neural network.

For this particular case, what did you do with the data w.r.t. data-science?

We generated a very large amount of data using the machine-learning model we trained, which we subsequently analysed in terms of sub-sets and common features for particular regions of property space.

For this particular case, what did you do with the data w.r.t. materials applications?

Besides using the data, we generated to understand the optoelectronic properties of conjugated polymers and test data-driven models that predict these properties, we also identified monomers that commonly occur in polymers for certain regions of property space, associated with applications such as organic photovoltaics and organic light emitting diodes.

For this particular case, what was the quantitative value of combining data with materials modelling?

Machine learning requires some time to make it work, but if compared to earlier approaches of attempting to find structure activity relations it takes less person time. This reduced amount of time spent on learning speeds up the research. The combination of us tackling a hot topic (i.e. machine learning) and then be able to answer questions that otherwise could not have been answered did lead to publications with a higher impact.

What investments were made during the project?

We needed some time to get data together and permit Liam to learn new skills. From the perspective of infrastructure, we did not need investments as we could work on a normal computer with decent memory, but we did not need specialised equipment with GPUs, for example.

What sort of obstacles or barriers (if any) did you have to overcome to use data driven modelling?

The barriers are very low as machine learning is becoming a contender in the materials science world. I needed a little convincing from Liam to go ahead. As this field is so new, there are many ideas floating around but one has to follow the right ones to make them tangible. Thus, we had to manage how to go about this project sensibly.

Did using data improve your competitiveness/innovation power?

Yes. These days, experimental groups have adopted materials modelling skills and can solve problems we would have aided them years ago in collaborations. Thus, we materials modellers can go ahead and concentrate on questions nobody could answer so far. This made us look beyond a subsystem of data to larger search spaces where we can learn more global things about materials relations. We also get exposure to new circles within sciences and we can exchange new ideas, and hence, become more innovative.

What would you need the community to provide to enable data-driven materials modelling?

We would like to make data accessible to the experimentalist community. This would need a follow-up grant to make this possible. So far, we can share our data quite easily with skilled expert modellers. I see as an ideal funding setup to enable a scientist to create data, perform data-based modelling on them and then find ways to make data and data-based models accessible to the non-expert. Another point I would like to raise is that we need dedicated repositories with an infrastructure that can not only host our data and data-based models, but also maintain them and keep them running.

References

[1] “Towards understanding the thermodynamic viability of Zeolites and related frameworks through a simple topological model” M.A. Zwijnenburg, S.T. Bromley, M.D. Foster, R.G. Bell, O. Delgado-Friedrichs, J.C. Jansen, T. Maschmeyer; Chem. Mater. 2004, 16, 3809. https://pubs.acs.org/doi/abs/10.1021/cm049256k

[2] MD Foster and MMJ Treacy, A Database of Hypothetical Zeolite Structures: http://www.hypotheticalzeolites.net

[3] https://github.com/ZwijnenburgGroup/pychemlp

[4] “Extended-Connectivity Fingerprints.” D. Rogers, M. Hahn, M. J. Chem. Inf. Model. 2010, 50, 742–754.

[5] https://www.chemie.uni-bonn.de/pctc/mulliken-center/software/xtb/xtb

[6] “STK: A Python Toolkit for Supramolecular Assembly.” L. Turcani, E. Berardo, K.E. Jelfs, J. Comp. Chem 39 (2018) 1931-1942 https://onlinelibrary.wiley.com/doi/10.1002/jcc.25377