EMMC case study:

Machine Learning for Organic Cage Property Prediction

Interview of Dr Kim Jelfs, Imperial College London

Writers: Alexandra Simperler and Gerhard Goldbeck

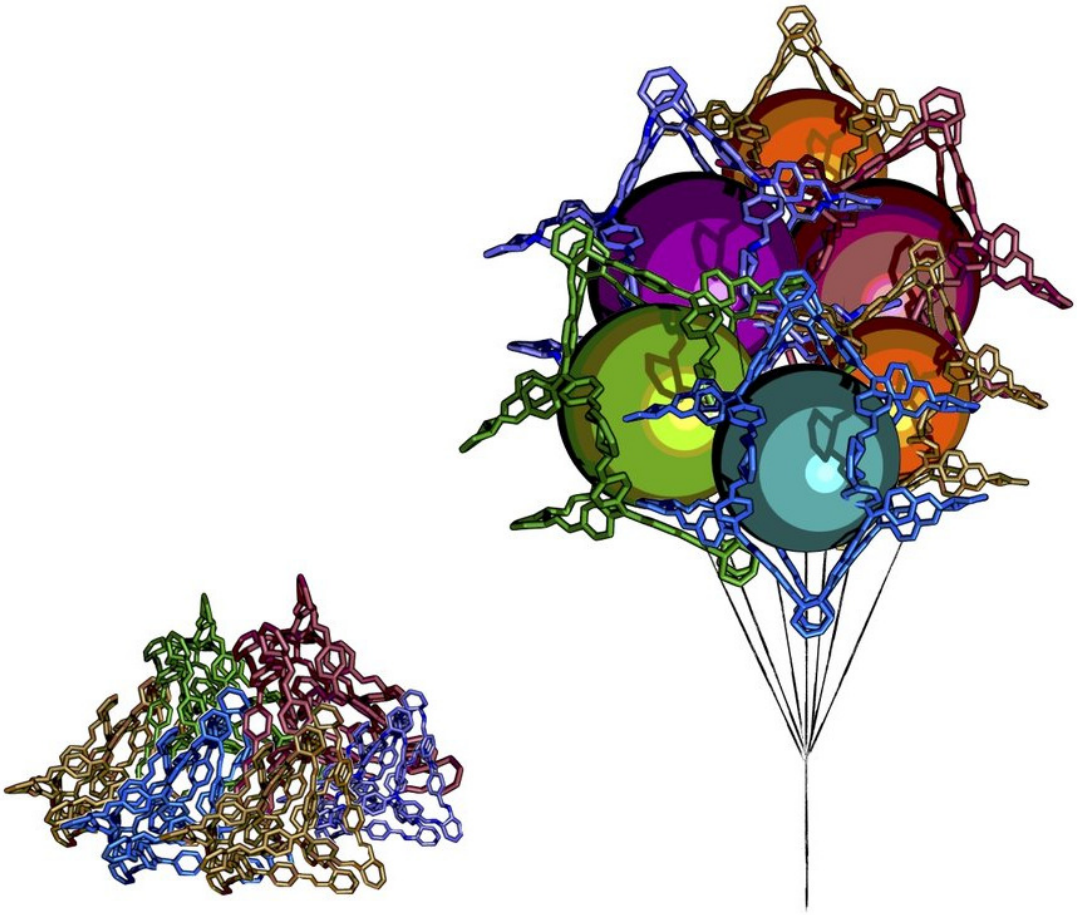


Figure by Dr Valentina Santolini

About Kim Jelfs

Dr Kim Jelfs (http://www.jelfs-group.org/) is a Senior Lecturer at Imperial College London. She and her group are focussing on computational approaches towards enabling functional molecular material discovery. Dr Jelfs is particularly interested in predicting these materials' assembly as individual units and how this then affects self-assembly and properties. She and her group are working on large scale computational screening of precursor libraries, and creating databases of viable, functional materials. Dr Jelfs is collaborating with experimentalists to allow synthetic realisation of her predictions.

Dr Jelfs was awarded the Harrison-Meldola Memorial Prize in 2018 for her outstanding work as a scientist.

Essentials about data and modelling

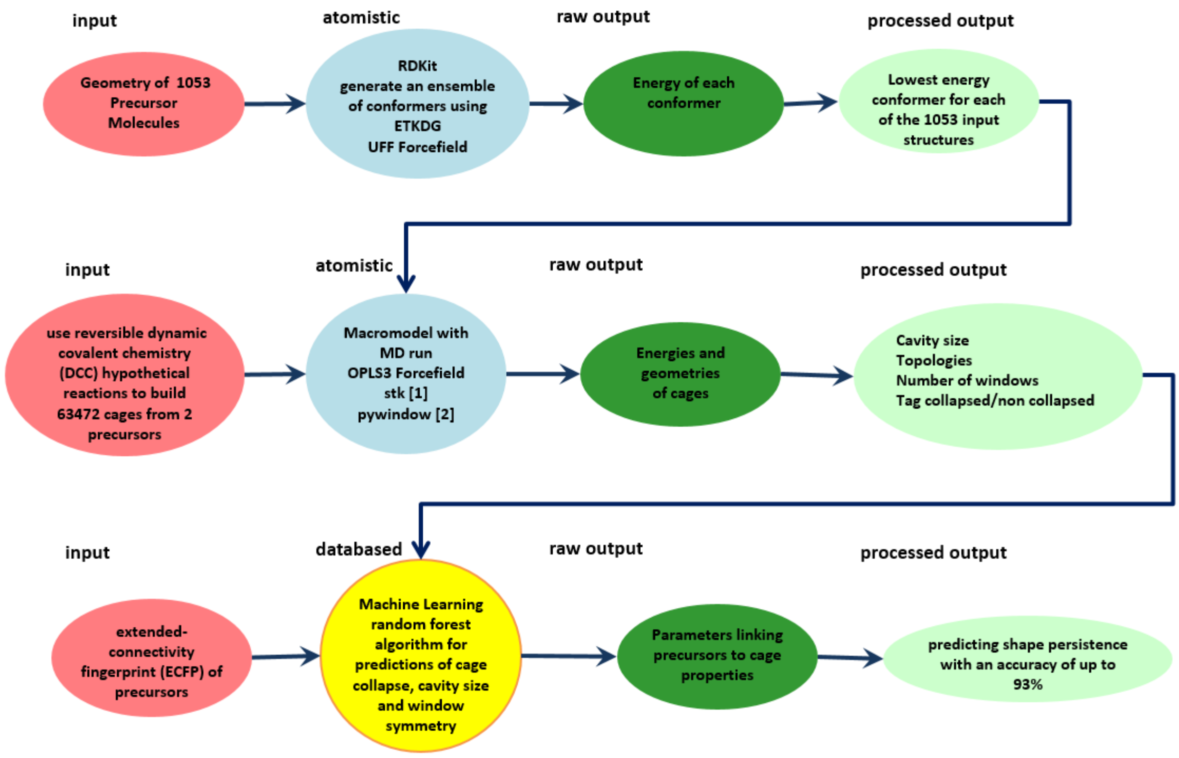
Dr Jelfs is very interested in the prediction of porous organic cages as they are an interesting material for applications such as encapsulation, molecular separations and catalysis. The compounds are lacking an extended network of covalent bonds in the solid state which makes them quite flexible. This flexibility is an advantage if one would like to actively change the porosity of a compound. However, this can turn into a disadvantage as a certain shape persistence is required to prevent these materials from collapsing. An experimentalist could spend around a year to create such porous organic cages and test them. Thus, a computational tool to predict shape persistence starting from an individual molecular unit could enable experimentalists to test in silico before starting their synthesis. Dr Jelfs wanted to develop a tool for experimentalists to use themselves without having to purchase materials modelling software and learn how to perform extensive computation. It was also pertinent to create a tool that is rather fast, hence machine learning was the way to go. The corresponding source code is available at https://github.com/lukasturcani/cage_prediction under MIT license.

The corresponding database comprises structures of precursor cores (i.e. molecules suitable to build organic cages) each with locations of functional groups marked and these structures are stored in SMILES representation. Cages were built and added to the database which is hosted by Imperial College London at https://data.hpc.imperial.ac.uk/resolve/?doi=4618. This university repository clearly states the author and a Creative Commons license is in place.

About the Case Study

The case is based on Dr Jelfs' paper "Machine Learning for Organic Cage Property Prediction", L. Turcani, R. L. Greenaway, K. E. Jelfs, Chem. Mater. 31 (2019) 714-727 https://doi.org/10.1021/acs.chemmater.8b03572

This work was chiefly conducted by Lukas Turcani, where Dr. Rebecca Greenaway provided insight on the synthesis of cages and developed the library of precursors used for the study. The publication describes how they use machine learning to predict shape persistence and cavity size in porous organic cages. Therefore, they created the largest computational database of these molecules comprising 63,472 cages, formed through a range of reaction chemistries and in multiple topologies. The idea was to use this database and identify features which lead to the formation of shape persistent cages. They developed machine learning models capable of predicting shape persistence with an accuracy of up to 93%, thereby reducing the time taken to predict this property to milliseconds, and removing the need for specialist software.


input
atomistic
raw output
processed output
Geometry of 1053 Precursor Molecules
RDKit generate an ensemble of conformers using ETKDG UFF Forcefield
Energy of each conformer
Lowest energy conformer for each of the 1053 input structures
input
atomistic
raw output
processed output
use reversible dynamic covalent chemistry (DCC) hypothetical reactions to build 63472 cages from 2 precursors
Macromodel with MD run OPLS3 Forcefield stk [1] pywindow [2]
Energies and geometries of cages
Cavity size
Topologies
Number of windows
Tag collapsed/non collapsed
input
databased
raw output
processed output
extended-connectivity fingerprint (ECFP) of precursors
Machine Learning random forest algorithm for predictions of cage collapse, cavity size and window symmetry
Parameters linking precursors to cage properties
predicting shape persistence with an accuracy of up to 93%

For this particular case, which were your objectives?

Our aim was to provide a tool to experimentalists that can guide them to the right molecular precursors to synthesise a promising porous organic cage. The tool should be easy to use and allow a quick insight into whether their selected precursors are likely to form a shape persistent cage or not. Thus, the experimentalists are less likely to spend time on a material that is less promising.

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

We developed a large computational cage database consisting of over 60,000 cage molecules and then trained random forest models for the prediction of cage collapse, cavity size and window symmetry. Thus, we created our own data-based model.

How did data play a key role in problem solving?

We want to start from a set of molecules and predict whether a cage is stable or collapses. Already small changes in a precursor can have a large impact on a cage and make it collapse. Thus, a large amount of data has to be provided to enable ML to catch the finer details.

What methodologies have been applied?

We have used Molecular Dynamics, Machine Learning and python.

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

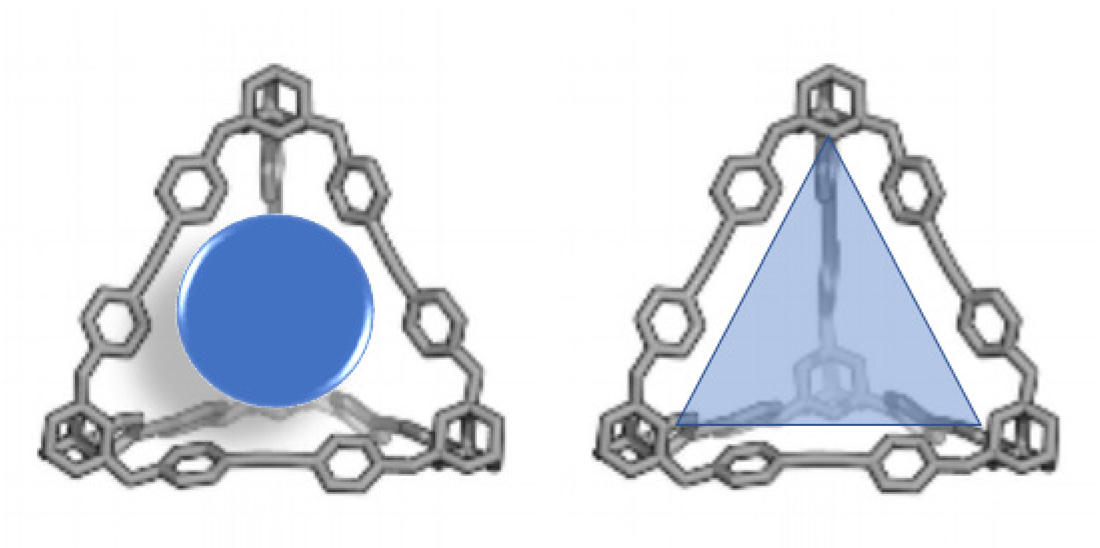
Traditionally atomistic methods would be used to predict properties of cages. However, modelling a large number of molecules is still time-consuming when using atomistic models and intractable with electronic models, particularly as these cages are large, typically with more than 100 atoms. By using machine learning techniques and data we could combat excessive computational times.

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

To properly use our machine learning we had to classify our cages as either "collapsed" or "not collapsed". We invested a lot of work to get our classification right.

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

We have created a database of 63,472 cages formed through a range of reaction chemistries and in multiple topologies. Comprised in the database are the structural data of the cages (geometries, bond order, nature of functional groups, …), cavity size (marked by a blue sphere) and window size (marked by a blue triangle).



We use our in-house developed *pywindow* tool [2] to calculate the window size.

All data-based models rely only on the molecular graphs of cage precursors, i.e. we analyse the cage precursors in our database to determine which structural features are conducive to the formation of shape persistent cages. The experimentalist has then to provide their precursor of choice, and from its structural features they can predict if their cage is likely to be stable or not.

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

Our tool permits the combination of two precursor species. We can use our organic cage predictor tool to predict properties of hypothetical cages and to enable finding novel cages.

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

Our tool will save person time and speed up research for both the modeller and the experimentalist. As modellers we learned how to engage with data and machine learning and explore new structures of interest faster. We can spend computation time on promising porous structures rather than on uninteresting, collapsed cages. The experimentalists can save time in the lab as our tool enables them to select precursors for porous materials virtually before going to the lab.

What investments were made during the project?

We needed a PhD scholarship for Lukas Turcani and he taught himself about machine learning which he achieved easily within the scope of his PhD. We could work with our usual commercial licenses and hardware access, so there were no investments on that level.

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

One expected obstacle was certainly the data generation, which was needed to make machine learning a success.

Did using data improve your competitiveness/innovation power?

My peers are often looking into individual systems; we however, with our tool can provide a faster and wider insight into many systems. Thus, data enabled us to become more competitive and innovative.

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

I feel that projects like NOMAD are better set up for inorganic materials and I have to think if and how this could work for our organic systems. Our university does provide repositories and servers, but these require maintenance which is provided only to a certain extent. Data driven materials modelling could profit from national or international repositories that understand the versatility of materials and go beyond just being a pure hosting service. Especially the provision of tools like our organic cage property predictor require working (chemistry) libraries and other infrastructure that need to be maintained. Thus, such infrastructure requires experts who can provide these additional services.

References

[1] “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

[2] "pywindow: Automated Structural Analysis of Molecular Pores", M. Miklitz and K. E. Jelfs, J. Chem. Inf. Model. 58 (2018) 2387-2391 https://doi.org/10.1021/acs.jcim.8b00490