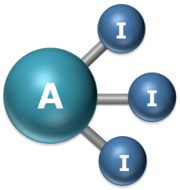

Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Optimising Flatland: Inverse design of desalination membranes
Interim Report
Project Dates: 01/05/2020 - 15/12/2020
The University of Sheffield

Dr J. Grant Hill and Adam N. Hill
University of Sheffield

Report Date: 30/09/2020

AI3SD-Project-Series:Report5_Hill_Interim

Optimising Flatland: Inverse design of desalination membranes
AI3SD-Project-Series:Report5_Hill_Interim
Report Date: 30/09/2020
DOI: 10.5258/SOTON/P0040

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Project Details

Title	Optimising Flatland: Inverse design of desalination membranes
Funding reference	AI3SD-FundingCall2_008
Lead Institution	University of Sheffield
Project Dates	01/05/2020 - 15/12/2020
Website	N/A
Keywords	Machine learning; Generative models; Inverse design

2 Project Team

2.1 Principal Investigator

Name and Title	Dr J. Grant Hill
Employer name / University Department Name	University of Sheffield / Department of Chemistry
Work Email	grant.hill@sheffield.ac.uk
Website Link (if available)	http://www.grant-hill.group.shef.ac.uk

2.2 Co-Investigators

Name and Title	Prof. Patrick W. Fowler
Employer name / University Department Name	University of Sheffield / Department of Chemistry
Work Email	p.w.fowler@sheffield.ac.uk
Website Link (if available)	https://www.sheffield.ac.uk/chemistry/people/academic/patrick-w-fowler

Name and Title	Dr Jonathan A. Foster
Employer name / University Department Name	University of Sheffield / Department of Chemistry
Work Email	jona.foster@sheffield.ac.uk
Website Link (if available)	https://foster.group.shef.ac.uk

Name and Title	Dr Peyman Z. Moghadam
Employer name / University Department Name	University of Sheffield / Department of Chemical and Biological Engineering
Work Email	p.moghadam@sheffield.ac.uk
Website Link (if available)	https://www.sheffield.ac.uk/cbe/people/academic-staff/peyman-z-moghadam

Name and Title	Dr Kim Jelfs
Employer name / University Department Name	Imperial College London / Department of Chemistry
Work Email	k.jelfs@imperial.ac.uk
Website Link (if available)	http://www.jelfs-group.org

2.3 Researchers & Collaborators

Adam N. Hill has been seconded full-time from his PhD studies to work on this project. Although the current project is unrelated to his PhD research, Adam has experience of writing scientific code and the application of machine learning in chemistry.

3 Publicity Summary

The aim of this project is to establish an integrated discovery pipeline for developing new materials for water desalination. This will be achieved by creating a large dataset of ultrathin membrane materials and then applying artificial intelligence (AI) design paradigms to predict the ideal materials for this application. The best candidate materials will then undergo experimental evaluation.

4 Executive Summary

The aim of this project is to use AI generative design to develop new 2D materials for desalination, with an eventual goal of integrated discovery where experimental results are used to refine AI predictions. To ensure this feasibility study has realistic projected outcomes, it is focused on generating a data set via automated computational methods, using a generative model in the inverse design of 2D materials for desalination, and in an initial experimental validation. If the generative model predicts materials that prove to be effective for desalination, this would make a strong case for further investment from bodies such as EPSRC towards an automated, integrated discovery pipeline for 2D materials.

5 Aims and Objectives

Over 1.2 billion people around the world currently lack access to safe drinking water, a number the World Water Council estimate will rise to 3.9 billion in coming decades as a result of continued industrialisation and climate change. Three quarters of the Earth's surface is covered in water, but current desalination technology is prohibitively expensive and energy intensive. Two-dimensional materials, in particular metal-organic nanosheets, have shown massive potential for creating a new generation of ultrathin membranes with exceptionally high flux rates and

selectivity for desalination. The overall goal of the research project is to discover metal-organic nanosheets for desalination using a combination of AI methods and experimental validation. This is sub-divided into the following aims:

- Use generative AI methods to inverse design MONs with ideal properties for desalination and exfoliation.
- Produce a sustainable software pipeline/framework that can be adapted for the inverse design of other 2D materials.
- Establish an effective collaboration across the interdisciplinary research team.
- Obtain preliminary experimental results that will act as a springboard for larger research proposals targeting closed-loop 2D materials discovery.

6 Methodology

6.1 Scientific Methodology

To train AI methods a large dataset is typically required to achieve reasonable results. However, for 2D materials there is insufficient experimental data for training a generative model – by their very nature 2D nanomaterials are poorly crystalline and hence difficult to characterise. In this work we have adopted a "synthetic data" approach, in which a computer algorithm is used to construct the dataset. After defining a large number of relevant building blocks that include organic linkers, secondary building units and topological nets, the construction of the nanosheets from these blocks is automated using the Topologically Based Crystal Constructor (ToBaCCo) code,[1,2] which was originally designed for 3D metal-organic frameworks. An interactive example of this data generation pipeline is demonstrated through a Jupyter notebook. The resulting large set of hypothetical nanosheets can then be evaluated in terms of relevant properties such as pore size, before the data set is used to train generative AI methods.

6.2 AI Methodology

We plan to train both a Generative Adversarial Network (GAN) and a Variational Autoencoder (VAE) to compare the performance of both generative networks. A GAN consists of two parts, a generator and a discriminator, which are trained concurrently, with the generator trying to achieve realism, and the discriminator trying to perceive realism. Playing off each other, the two networks recurrently improve each other. A VAE operates by learning the underlying probability distribution of our dataset, and then making predictions based upon that distribution. In both cases, the expectation is that new nanosheets that reside in different areas of material space will be predicted. To facilitate this, an inverse design paradigm will be used, where the correct properties for desalination are defined and the generative model is biased towards predicting new nanosheets with these properties.

- At this interim stage no generative AI component has been implemented, but this is the sole focus of the second part of this project.
- Initially both a GAN and VAE will be trained on our dataset to ascertain which looks most promising, then the better performing generative method will be optimised.
- A unique string-based identifier, known as a MOFid,[3] has been generated for each entry in the dataset. It is intended that this string will also be used as representation within the generative networks.
- The generative networks will be implemented in Python, using sustainable software practices and existing machine learning / AI frameworks.

7 Interim Results

We have generated a large database of hypothetical, computationally predicted 2D nanosheet structures based on a set of predetermined building blocks: topographies, secondary building units (SBUs), and organic linkers. The process for generating the nanosheets is outlined below, and a demonstration is provided in a supplemental Jupyter Notebook.

- A total of 194 2D topography files were extracted from the full topography database contained within ToBaCCo. These topologies had been extracted from the Reticular Chemistry Structure Resource (RCSR)[4]
- We have identified 63 SBUs suitable for the creation of 2D materials from both the 131 units identified in MOF chemistry by Tranchemontagne et al.[5], and the SBUs provided with ToBaCCo. We have created these SBUs in the Avogadro molecule editor and automated their conversion into the crystallographic information file (CIF) file format used throughout our pipeline.
- Using existing expertise within the Foster group 10 unique and common organic linker backbones were identified that can be used in 2D nanosheet generation.
- Using RDKit and code developed in this project, the selected organic linker backbones were then mutated with 32 different functional groups to generate a large database of 30,976 varied organic linkers, which were also converted into CIF format.
- Once the sets of all three necessary building blocks were collected, ToBaCCo was used to automate the construction of 2D nanosheets, with a CIF generated for every permutation of the building blocks.
- Each resulting nanosheet was then optimised using molecular mechanics and the universal force-field (UFF)[6] through Open Babel,[7,8] and the pore size calculated using Zeo++.[9]
- The program MOFid[3] was used to generate a unique identifier for each nanosheet constructed.

Flexibility in terms of future use and development has been a key element in the design of this data set, and its mostly automated generation. For example, extending the current database with nanosheets containing any new SBUs discovered, or including additional organic linker backbones that provide promising experimental results would be straightforward. The current choices of topology, SBU and linker building blocks were motivated by relatively constrained mutations of experimentally verified structures, with the aim of generating nanosheets that are feasible to synthesise. Greater variation within this mutation process, or extension to 3D MOF generation, would require only minor modifications to specific parts of the pipeline.

8 Outputs

- A new computational 2D nanosheet database that includes the 3D structure and connectivity (in CIF format), along with a unique identifier (MOFid) and calculated pore sizes. After further refinement, the initial release of this database will be made freely available via ORDA, the University of Sheffield's online research data hub, providing an open and citeable dataset to the community.
- The software pipeline for generation of the dataset will be made available via GitHub, for hosted version control.

9 Progress Summary

- The project start was delayed due to COVID-19, with the researcher beginning work on the project on 15th June 2020. AI^3 Science Discovery Network+ have agreed to a short no-cost extension to cover these delays. This project is running through a period of local lockdowns and varied government guidelines on home working, as such meetings that were intended to be face to face have happened online instead.
- The project has three planned activities: generating the dataset; training a generative model; initial experimental validation. Activity one has been completed, and a new 2D nanosheet database has been generated.
- The database was constructed with particular care to provide a working pipeline of programs that allow for quick and easy nanosheet generation, should extension of the dataset prove desirable.

10 Next Steps

- Construct the two generative networks selected for further investigation (GAN and VAE).
- Train both networks and analyse their performance to determine the most promising.
- Take best-performing generative network forward, continuing to optimise hyperparameters and improve predictions.
- Use the now trained and optimised generative method to produce a hypothetical "best" water desalination membrane based on pore size, and interlayer strengths.
- Synthesis and characterisation of this membrane by the experimental team.

11 References

References

[1] Y. J. Colón, D. A. Gómez-Gualdrón and R. Q. Snurr, *Cryst. Growth Des.*, 2017, **17**, 5801–5810.

[2] R. Anderson and D. A. Gómez-Gualdrón, *Cryst. Eng. Comm.*, 2019, **21**, 1653–1665.

[3] B. J. Bucior, A. S. Rosen, M. Haranczyk, Z. Yao, M. E. Ziebel, O. K. Farha, J. T. Hupp, J. I. Siepmann, A. Aspuru-Guzik and R. Q. Snurr, *Cryst. Growth Des.*, 2019, **19**, 6682–6697.

[4] M. O'Keeffe, M. A. Peskov, S. J. Ramsden and O. M. Yaghi, *Acc. Chem. Res.*, 2008, **41**, 1782–1789.

[5] D. J. Tranchemontagne, J. L. Mendoza-Cortés, M. O'Keeffe and O. M. Yaghi, *Chem. Soc. Rev.*, 2009, **38**, 1257–1283.

[6] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard and W. M. Skiff, *J. Am. Chem. Soc.*, 1992, **114**, 10024–10035.

[7] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminformatics*, 2011, **3**, 33.

[8] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *Open Babel*, https://openbabel.org/, 2020.

[9] T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza and M. Haranczyk, *Micropor. Mesopor. Mat.*, 2012, **149**, 134–141.

12 Data & Software Links

Software/database developed in this work:

- Jupyter Notebook with example of database structure

Existing programs used in this work:

- HDF5 Database
- MOFid
- OpenBabel
- RDKit
- ToBaCCo 3.0
- Zeo++