Exploring the Use of Transfer Learning for Property Prediction of Metal and Covalent Organic Frameworks

Submitted on: 7th May 2024
Word count: 9,702

Table of Contents

1. Abstract

The adoption of cleaner fuels to replace oil and gas is essential in reducing global greenhouse gas emissions and combatting climate change. Alternative fuels such as hydrogen and methane are gaseous and therefore challenging to store safely and efficiently. Metal organic frameworks (MOFs) and covalent organic frameworks (COFs) are two promising materials for gas storage applications, due to their light weight, large storage capacities and readily tuneable structures. Machine learning models can accelerate the discovery of novel MOFs and COFs by generating predictions of properties important in gas storage for large numbers of structures, faster and more cost-effectively than experiment or computational simulations. However, this machine learning-based discovery is hindered by the lack of available MOF and COF structure and property data. Here transfer learning, a machine learning technique, is used to overcome data scarcity in MOF thermal conductivity and COF gas adsorption data. The effectiveness of transfer learning is found to be limited by biases in the training data, and the complexity of the structure-property relationship to model. Though ineffective for MOF thermal conductivity prediction, transfer learning did improve the accuracy of COF gas adsorption predictions by up to 9.6%. This is significant, as it shows that more abundant MOF data can be leveraged for the discovery of optimised COFs. Overall, this project serves as an initial proof-of-concept study demonstrating the successes and challenges of machine learning and transfer learning in the discovery of novel MOFs and COFs optimised for gas storage applications.

Commented [JJ1]: Emissions of what? Be explicit. Also alternative fuels implies an alternative *to* something. Maybe, 'alternatives to X-based fuels are essential to...'

Commented [JJ2]: What is this?

Commented [JJ3R2]: (I get it but maybe say 'the relationship between molecular structure and thermophysical properties'. If the only thermophysical property you're using is conductivity, then just say 'conductivity'

Commented [JJ4]: Future work comes after this imo. But double check because idek if future work is mentioned in the abstract more generally

2. Introduction

2.1 Nanoporous Crystalline Materials

In 2022, global greenhouse gas emissions reached a record high, equivalent to 57 gigatons of CO_2.[1] This led the UN to call for an acceleration of green development initiatives to limit the impacts of climate change.[1] Greenhouse gas emissions can be lowered by replacing oil and gas with cleaner and more efficient alternative fuels. Common alternative fuels, such as hydrogen and methane, are gaseous and are thus challenging to store safely, limiting their wide-scale use. In an effort to reduce harmful emissions, novel materials must be developed that can effectively store gaseous fuels.

Nanoporous crystalline materials are promising candidate materials for gas storage applications.[2-4] Nanoporous materials are a class of stable, low-density materials containing permanent, well-defined pores between 1 to 100 nm wide.[5] They have demonstrated potential in a range of environmental applications, such as catalysis, sustainable batteries and gas storage.[6-11] The porosity of these materials endows them with high surface areas which enable them to store gases efficiently.

Two porous materials of growing interest for gas storage applications are metal organic frameworks (MOFs) and covalent organic frameworks (COFs).[12-14] Their structures are formed from molecular building blocks of nodes connected to organic linkers in an extended porous framework (Figure 1). The nodes in MOFs contain metal atoms, whereas the nodes in COFs are formed entirely from organic components.

Commented [JJ5]: Redundant unless you're specifically referring to the *most* severe, in which case, explicitly say 'most severe'.

Recommendation: cut 'severe'

Commented [JJ6]: Alternate != alternative.

Commented [JJ7]: Subscript?

Commented [JJ8]: Subscript? Not gonna repeat every time but you get the idea

Commented [JJ9]: V big logical gap. You need at least one paragraph before this explaining the following:

What does 'gas storage applications' mean? Give examples

Why do gas storage applications matter for lowering emissions?

Are the storage applications new? I.e., are you developing materials for a new product entirely, or just replacing the existing materials with new ones?

How does changing materials reduce emissions?

Commented [JJ10R9]: I can see you've given two examples in this sentence. You still need the paragraph I suggested in the original message

Commented [JJ11]: I've used efficiently because idk enough to add more detai. Idl 'efficiently' because it doesn't mean much with qualifiers - efficient in what sense?

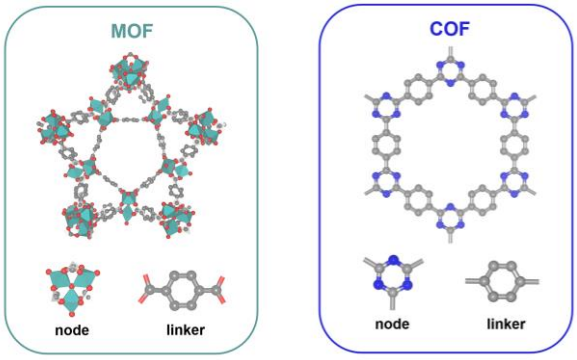


Figure 1. *Example MOF (MIL-101) and COF (CTF-1) substructures, which repeat to form extended frameworks.[15, 16] Both frameworks are assembled from node and linker molecular building blocks. Structures visualised using VESTA, with hydrogen atoms hidden.[17] Key: grey, carbon; red, oxygen; blue, nitrogen; green polyhedra, chromium.*

MOFs and COFs present three unique advantages over other porous materials. The first is their higher surface area and porosity, with MOF internal surface areas (~7000 $m^2 g^{-1}$) more than seven times greater than that of zeolites (<1000 $m^2 g^{-1}$).[18-20] Larger surface areas increase the number of accessible sites for gas adsorption, generally corresponding to improved gas storage capacity.[18, 21, 22]

Commented [JJ12]: One to nine written in full unless part of a measurement (e.g., 'seven times greater' vs 7 km)

Double digits written as Arabic numerals ('10 times greater')

The second advantage is the readily tuneable structure of MOFs and COFs, which enable adjustment of pore shape and size, and straightforward incorporation of functional groups – valuable in enhancing the selectivity of gas adsorption.[23]

Their final advantage is their remarkable structural diversity. A near infinite number of possible MOF/COF structures is possible, because MOFs and COFs possess an unrestricted selection of possible linkers and nodes, readily functionalised organic building blocks and a diverse range of potential topologies.

Accordingly, MOF/COF material space is enormous – the number of unique synthesised MOFs (˜100,000) exceeds the number of zeolites (˜300) by three orders of magnitude.[24, 25] This expansive design space presents a compelling opportunity to discover a novel structure optimised for gas storage.

2.2 Inverse Design

Finding a desired structure is challenging given the high structural diversity of MOFs and COFs.[26] Due to the near infinite number of possible materials, systematic testing of every potential material is impractical and often limited to known synthesisable materials. A better approach would be to narrow down the search to a few diverse structures with the most promising properties for a desired application.

Inverse design is an example of such an approach. It begins with a desired property, *i.e.* gas adsorption, and works backwards to identify potential materials exhibiting that property.[27] This property-to-structure approach allows for wider exploration of unknown materials, improving the likelihood of discovering a novel material with optimal properties.

Machine learning (ML) techniques have demonstrated success in identifying non-obvious structure-property relationships from data - a critical step towards inverse design.[27-29] In one study, ML improved a single MOF's gas adsorption by over 400%.[30] ML can also help to accelerate discovery by focusing experimental efforts on only the most high-performing materials.[31]

The following sections will provide background on the functional design of MOF and COF materials optimised for gas storage. Section 2.2.1 outlines the process of gas storage; Section 2.2.2 provides an overview of the ML methods used to identify materials optimised for gas storage.

2.2.1 Gas Adsorption and Thermal Conductivity

One of the main proposed gas storage applications of MOFs and COFs is focused on fuel storage in vehicles, which require the storage and release of gaseous fuels.[2, 32-34] At present, research is focused on discovering and developing a material viable for real-world applications. In particular, much work explores how to improve gas storage

Commented [JJ13]: Maybe:

'One of the main *proposed* applications of MOFs and COFs is fuel storage in vehicles, which requires the storage and release of gaseous fuels.'

capacity – the number of gas molecules that can be adsorbed or desorbed onto the surface of the material.

Gas adsorption in MOFs and COFs occurs mainly via physisorption, meaning that storage capacity scales with accessible surface area.[35, 36] One might hypothesise therefore, that storage capacity can be increased solely by increasing the size or number of pores, and hence the accessible surface, in the material.

However, bigger or more abundant pores correspond to lower charging rates, limiting applications in vehicular fuel storage.[37, 38] This is because highly porous materials are poor at dissipating the heat released upon adsorption. As the material warms during loading with gas, the adsorption process becomes less energetically favourable – reducing the maximum gas uptake. Thus, charging rates need to be restricted to minimise heating of the material and preserve good gas storage capacity.

To ensure practical charging rates, a material's thermal conductivity – it's ability to transfer heat – must be improved. Due to their high porosity, MOFs and COFs generally have low thermal conductivities; heat is transferred poorly across pores and is best transferred via vibrations through the lattice.[39]

However optimising structures for high thermal conductivities is not a straightforward task. Heat is transferred in MOFs and COFs by discrete vibrational energy carriers known as phonons.[40] Thermal conductivity is proportional to the mean free path travelled by phonons, *e.g.* the average distance a phonon can travel.[40] Phonon transport has a complex correlation to a material's structure, as it is affected by a large number of variables in ways that are still not fully understood.[40]

One study found that high thermal conductivity MOFs display relatively high densities and low porosity, however these properties were also found in low thermal conductivity MOFs.[41] Similarly, longer linkers have been found to correlate with higher porosity and thus lower thermal conductivity.[42] Thermal conductivity also decreases

Commented [JJ14]: Is it obvious what this means?

Commented [JJ15]: Idg how this affects storage *capacity*

Commented [JJ16]: Why did you italicise?

when atoms in the lattice have different masses, due to increased phonon scattering – impeding thermal transport.[43]

A practical gas storage material must be optimised for both gas adsorption and thermal conductivity. This will involve finding the right compromise – for example, a pore size large enough for sufficient gas storage capacity, but small enough for practical charging rates. Pore size is just one of many structural properties that must be investigated to develop a material optimised for use in gas storage devices.[32, 44] ML methods excel at solving such optimisation problems, with potential to model the complex interconnected relationship between a structure and its thermal conductivity and gas adsorption properties.[45]

2.2.2 Machine Learning for Property Prediction

An ML model can be built to predict a desired property for a given input structure.[46-49] A general workflow for the development of an ML model can be seen in Figure 2, consisting of data preparation, model selection and model generation. This section will discuss this workflow in depth, within the context of property prediction for MOFs and COFs. Here, 'training' refers to the process in which the algorithm is run, and 'training data' refers to the input data used to train the algorithm.

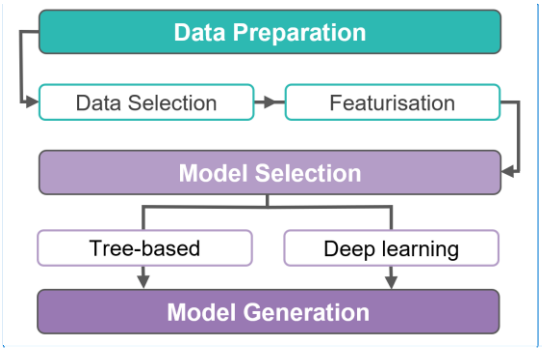


Figure 2. *A general machine learning workflow, as discussed in detail in Section 2.2.2.*

Commented [JJ17]: What does this mean?

Commented [JJ18]: No need to italicise. Also, e.g., in this context sounds informal

Commented [JJ19]: Maybe : 'that must be investigated' ?

Commented [JJ20]: Love this figure!

Data Preparation

An ML model makes predictions by mapping a relationship between inputs and outputs in a dataset. A model's performance is only as good as the dataset used to train it, since the dataset contains all the information that the model learns from.[50] The following section will outline a general method for the selection and preparation of input data.

Data Selection

The first step in ML model development is to choose a dataset of sufficient quality, quantity and diversity. Model performance scales with the number of data points (up to a limit) as the model has more examples to learn from. Diversity of both structural and property data is essential in ensuring that the model can make successful predictions on unseen data.

For property prediction, this dataset normally consists of input structural data labelled with the target property output. Structural data can be found in experimental or hypothetical databases. Hypothetical MOFs and COFs are generated computationally by combining a chosen set of nodes and linkers into different topologies.

Featurisation

Once a suitable dataset has been selected, it must be converted into a list of machine-readable features in a process called featurisation.[29,51] Property data is numerical and can be directly fed into the ML algorithm. However structural data is found in the form of files containing important structural information, including a list of coordinates for every atom in a unit cell. To be fed into an ML algorithm, this structural data must be converted into machine-readable structural features.

Commented [JJ21]: What's the difference between structural data and structural features?

I would explain v simply here.

Message me if unsure

The simplest structural features usually describe high-level properties, such as pore dimensions, atomic properties, topologies and extensive properties like density and

volume. More detailed descriptors describe the structure of node and linker subunits, normally represented as SMILES strings.[52] Among the most detailed features are crystal graphs, which translate the structure into a machine-readable graph, capturing detail of the entire 3D material.

The precision of features chosen should reflect the selected ML algorithm, with more sophisticated algorithms requiring more detailed descriptors. Another consideration is that more detailed features will increase the time and computational cost of training, but may not significantly improve model performance.

Model Selection

An ML model must be selected that is appropriate for both the input data and the structure-property relationship to be modelled. This subsection will provide a high-level overview of common ML algorithms relevant to MOF and COF property prediction.

Tree-Based Methods

Tree-based models use decision trees to generate predictions, and are among the simplest model architectures.[53] Decision trees work by iteratively splitting training data as to generate smaller groups of similar data (Figure 3).[54] This is based on the idea that structures with similar structural features will exhibit similar properties. Predictions are made by applying the splitting rules to new data.

Commented [JJ22]: Figure 3 doesn't show a random forest?

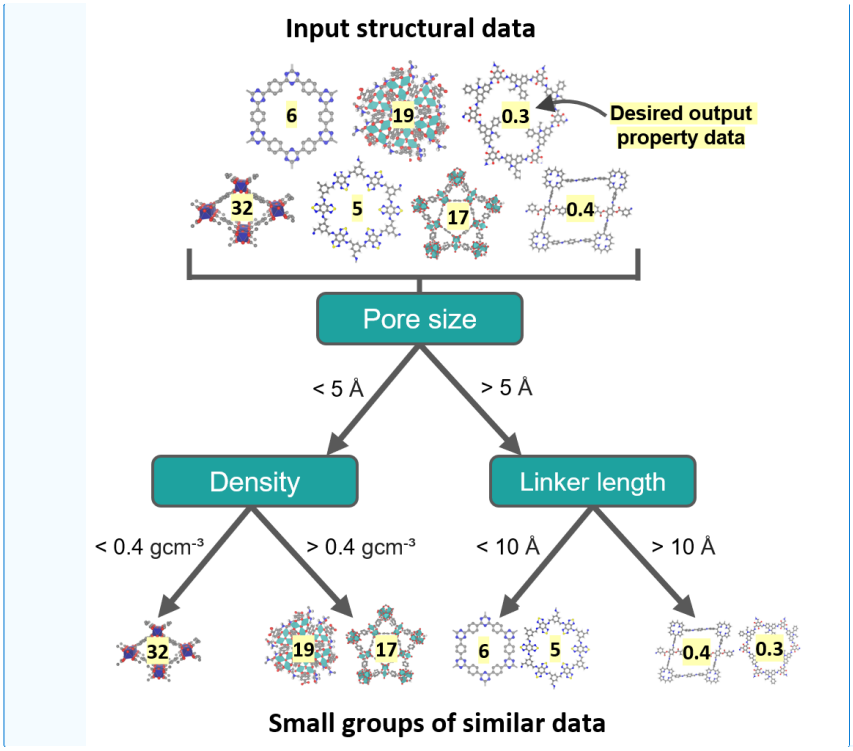


Figure 3. *An example of a simplified decision tree for MOF and COF property prediction. In this example, three structural features (pore size, density and linker length) are used to split the input structures into structurally similar sub-groups. During training, the algorithm decides on a splitting value that maximises the similarity of the desired property within each sub-group. For an unknown structure, predictions are made by following the relevant branches to the end, where the output value is the average of the desired property values of the sub-group.*

A single decision tree is highly sensitive to training data, resulting in overfitting – poor performance on unseen data. Random forest (RF) models mitigate this by combining the results from multiple decision trees, with each tree trained on a random subset of the input data.[55] RF models also handle outliers and high-dimensional data well.

XGBoost is another decision-tree based algorithm constructed from multiple gradient-boosted decision trees.[56] Gradient boosting involves running a gradient descent algorithm during training to minimise the prediction error for each tree. Consequently,

Commented [JJ23]: This figure is insanely good, v impressed

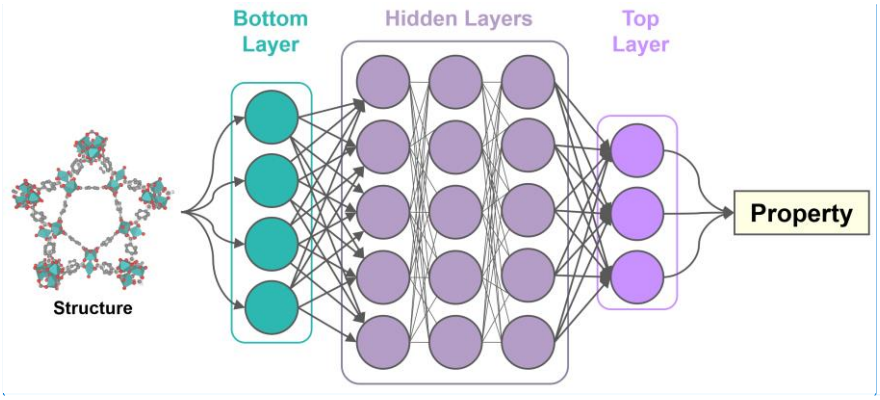
Commented [JJ24]: Is this a specific model or a type of model? Rn what you've written indicates it is a specific model, which I think is untrue

XGBoost tends to outperform RF and other common models in some prediction tasks.[57]

Commented [JJ25]: You can't say 'certain prediction tasks' without explaining what they are - don't need to get into specifics, just explain broadly what types of tasks they outperform RF and other common models in

Deep Learning Methods

Deep learning describes a subset of machine learning algorithms that contain neural network architectures.[58-60] An example neural network is shown in Figure 4.



Commented [JJ26]: *chef's kiss*, brillante!

Figure 4. *An example simplified representation of a neural network for MOF or COF property prediction. Each circle represents a node (a single processing unit). Nodes are arranged in vertical layers. Structural data is fed into the bottom layer of the neural network, through multiple hidden layers and finally through the top layer which generates a predicted property value for the input structure. Every node is connected to multiple other nodes in the layers above and below, forming a complex interconnected network.*

Neural networks are composed of layers of nodes. Each node is connected to several other nodes in the layer above and below, forming a complex interconnected network. Input data enters through the bottom layer and is processed by the nodes in each subsequent layer. An output is returned by the nodes in the top layer. Training a neural network involves optimising the weight each node applies to the data, as well as the threshold value, that if exceeded allows the node to pass its data to the layer above.

These multilayer neural networks are highly complex. Due to the large number of parameters in such a model, a significant amount of training data is required for effective training. Training data is often 'labelled' – meaning that the model receives the known output property data in addition to the input structural data. However,

Commented [JJ27]: Are you going to say that this is what 'deep' NNs are?

property data is often costly and time-consuming to obtain either experimentally or via physical simulations. This presents a challenge in generating labelled datasets large enough to effectively train a neural network.

To overcome this, neural networks can be trained on unlabelled data, *i.e.*, only the input structural data. Here the neural network learns unsupervised, independently learning patterns and trends in the structural data - leading to interesting insights. This process is often used for pretraining, allowing the model to develop an understanding of the underlying chemistry of the material, followed by fine-tuning for a specific property using a smaller set of labelled data.[61-65] This technique is a form of transfer learning, in which knowledge learned from one task is applied to a related task, improving performance especially in cases with little data available.

Two neural network models that have demonstrated good performance for material property prediction are crystal graph convolutional neural networks (CGCNNs) and Transformer models.

CGCNNs consist of two parts: crystal graphs and convolutional neural networks (CNNs).[66] As explained above, crystal graphs are machine-readable representations of an entire material structure. These crystal graphs are inputted into a CNN, which has been shown to generate predictions with accuracy comparable to more computational expensive quantum mechanical methods.[66]

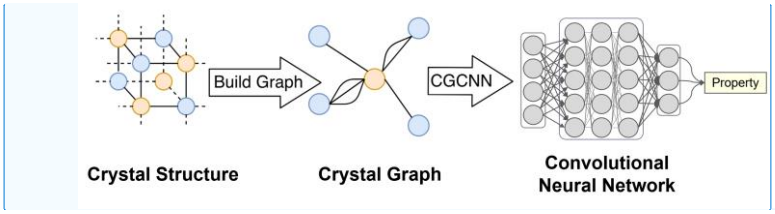


Figure 5. *A crystal graph convolutional neural network (CGCNN) converts a crystal structure into a machine-readable crystal graph. The crystal graph is input into a convolutional neural network which is trained to generate a property prediction. Figure adapted from reference [67].*

Commented [JJ28]: Why major? Also, if you do need a word here, maybe use 'main' or 'primary'. Sounds more apt than 'major'

Commented [JJ29]: This is so peng

Transformer models are used for natural language processing (NLP) and have demonstrated excellent performance in generative text prediction models like ChatGPT.[68-70] Unlike CGCNNs, Transformers only interpret text inputs, requiring MOF or COF structure files to be converted into a string of text which captures the important structural features. Despite having less detailed structural inputs, Transformers have demonstrated performance comparable to CGCNNs for MOF and COF property prediction tasks.[62, 63] This is beneficial, reducing the time and computational cost required by removing the need to generate an entire 3D structure for an entire dataset.

Commented [JJ30]: Why have you capitalised the T?

Commented [JJ31R30]: Is it really a proper noun?

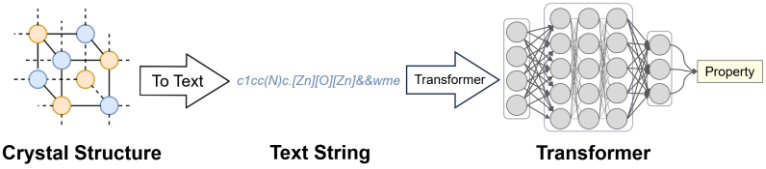


Figure 6. *A crystal structure can be converted into a representative text string, which is then input into a Transformer model which is trained to generate a property prediction. Adapted from reference* [67]

Commented [JJ32]: You need to write this in the standard style - 'Adapted from reference[68]'

Model Generation

Once the selected model has been trained on the dataset, its predictive performance must be evaluated. Evaluation consists of measuring the model performance on unseen data (data that the model has not been trained on). Cross-validation is a common method of model evaluation, involving splitting the dataset into training and test subsets. The model learns from the training data, and its performance is evaluated with the unseen test data. This assesses whether the model can replicate the underlying trend in the training data on unknown samples.

Three metrics are commonly used to quantify model performance. The first is mean absolute error (MAE), which measures the average magnitude of the error between known and predicted values (Equation 1). The second is the root mean square error (RMSE), a quadratic scoring function in which larger errors have increased contribution to the final average (Equation 2). The final metric is the coefficient of determination (R^2), which measures the overall fit of the model (Equation 3). The value

of R^2 lies between 0 and 1, where 1 is a perfect fit, and 0 describes a scenario in which the model does not explain any variation in the known property.

Commented [JJ33]: Don't think this is explained entirely correctly

$$MAE = \sum_{i=1}^{n} | P_i - O_i | \quad (1)$$

Where: P_i = predicted value, O_i = known value, n = number of datapoints

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \quad (2)$$

Where: P_i = predicted value, O_i = known value, n = number of datapoints

$$R^2 = 1 - \frac{\sum(P_i - \hat{P})^2}{\sum(P_i - \bar{P})^2} \quad (3)$$

Where: P_i = predicted value, $\hat{P}$ = value fitted to best fit line, $\bar{P}$ = average of predicted values

Initial model evaluation is often followed by iterative rounds of optimisation, such as the use of different datasets, feature representations or models to improve predictive power. One way to optimise is to tune the learning parameters of the ML algorithm, called hyperparameters.

Example hyperparameters include learning rate and model complexity, *i.e.* the number of trees in RF, or the number of layers in a neural network. Hyperparameter optimisation methods involve searching all possible hyperparameters for the combination that generates the highest predictive accuracy. This can be done exhaustively or randomly via grid and random search methods respectively.[71, 72]

3. Aims and Objectives

The discussion in the introduction can be summarised in three key points:

1) MOFs and COFs are materials that show promise for gas storage applications. Further optimisation of their properties – most especially their gas adsorption and thermal conductivity – is required before widespread use is possible.
2) Machine learning (ML) is a valuable tool to accelerate the discovery of novel MOFs and COFs with properties optimised for gas storage.
3) Transfer learning techniques enable accurate ML property predictions to be generated from even small datasets by borrowing knowledge learnt from larger pretrained models.

Commented [JJ34]: See my point at the beginning of the intro. The discussion of the different ML methods and how they work shouldn't be part of the intro imo

Cao *et al.* recently developed *MOFormer*, a Transformer model pretrained on over 400,000 MOFs and fine-tuned to predict MOF gas adsorption.[62] *MOFormer*'s large pretrained model has potential to be fine-tuned to predict other important properties (via transfer learning). This project will attempt to fine-tune *MOFormer* to predict two useful properties: MOF thermal conductivity and COF gas adsorption.

As far as the present author is aware, there are currently no published models for MOF thermal conductivity prediction in the literature. This is in part due to a lack of data; only a handful of MOFs' thermal conductivities have been determined experimentally.[73-77] Recently, Wilmer and colleagues calculated thermal conductivity for over 10,000 hypothetical MOFs via molecular dynamics simulations.[41] This moderately sized MOF Thermal Conductivity (MOF-TC) dataset is a promising starting point for a ML predictive model to discover high thermal conductivity MOFs optimised for gas storage.

COF research also suffers from limited data; less than 900 experimentally synthesised COFs have been reported in the literature.[78] COFs are of particular interest due to their superior stability, sustainability and lightness, which give them a practical advantage

over MOFs for vehicle fuel storage. Mercado *et al.* published a dataset of calculated maximum CH_4 gas adsorption for nearly 70,000 computer-generated hypothetical COFs.[79] This COF Gas Adsorption (COF-GA) dataset can hence be used to fine-tune *MOFormer* for COF gas adsorption.

This project will explore the effectiveness of transfer learning for generating useful predictions for MOF thermal conductivity and COF gas adsorption. The pretrained *MOFormer* model will act as a base model that will be fine-tuned to generate two predictive models for either property. The MOF-TC and COF-GA datasets will be used as input data for *MOFormer* to learn from.

Due to the lack of models for MOF thermal conductivity prediction in the literature, a model trained directly on the MOF-TC dataset will be developed. This will provide valuable insights into modelling MOF thermal conductivity, and also serve as a benchmark for how well transfer learning is able to improve predictions.

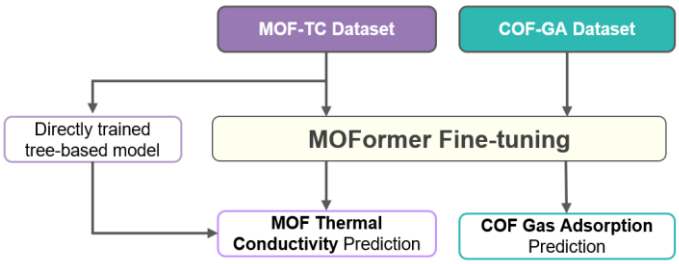


Figure 7. *Outline of project work. The MOF-TC and COF-GA datasets will be used to fine-tune MOFormer for MOF thermal conductivity and COF gas adsorption predictions respectively. An additional tree-based model will be trained directly on the MOF-TC dataset for MOF thermal conductivity prediction, serving as a benchmark for the fine-tuned MOFormer model.*

The aim of this project is to serve as an initial proof-of-concept study as to whether transfer learning could be beneficial in the discovery of novel MOF and COF materials for future green gas storage applications.

4. Computational Methods

All computations were run on Imperial College London's high-performance computing (HPC) resources. All scripts written for this project can be found on GitHub (https://github.com/jesrhea/TLforMaterialsDiscovery), including Python, Bash and YAML environment configuration files.

Commented [JJ35]: :oooooooooooooo

4.1 MOF Thermal Conductivity Predictive Models

4.1.1 MOF-TC Dataset

For all MOF thermal conductivity models, structure and thermal conductivity data for the MOF-TC dataset were taken from the repository generated by Islamov *et al.*[80] 10,194 MOFs were generated computationally via a combination of node and linkers. Structural data for the MOFs are stored in CIF file format. Thermal conductivity data is described in the *x*, *y* and *z* directions for each structure. The average thermal conductivity was calculated for each structure (Equation 4) and this average value used in training data for the subsequently generated models.

$$Average\ thermal\ conductivity = \frac{(k_x + k_y + k_z)}{3} \quad (4)$$

Where: k_i = thermal conductivity in the i direction.

4.1.2 Tree-Based Models

Data Featurisation

35 features were extracted from each MOF structure, using the Zeo++, pymatgen and mBUD software packages. A full list can be found in Table 1 in Section 5.1.15.1 .[81-83]

Zeo++ uses a graph representation of a structure's void space to calculate pore-related features. mBUD is used to generate node and linker sub-units, from which properties can be extracted. Other general properties, like topology, density and atomic composition were generated using pymatgen.

The program failed to generate features for some structures, meaning that the size of the featurised dataset (8,893 data points) is smaller than the original dataset (10,194 data points).

Model Generation

The scikit-learn and XGBoost packages were used for generating RF and XGBoost models respectively.[56, 84] All models were initialised with random_state = 0. Two hyperparameter optimisation methods were tested: random search and grid search.[71]

Stratified Sampling

Stratified sampling of the training dataset was achieved by using the train_test_split method from Scikit-learn, which ensures proportional sampling of user-defined outliers in the dataset.[84] Outliers were defined using z-score (Equation 5) that were greater than 2 or less than -2. This is illustrated in Figure 8 for a normally distributed dataset. For the MOF-TC dataset this translated to outliers with associated thermal conductivity values greater than 1.73 W m^{-1} K^{-1}.

Commented [JJ36]: ', from which properties can be extracted' ?

Commented [JJ37]: I think you need to explain how pymatgen generates properties. Not the technical details, but just whether it generates them at random or using some kind of data from the MOF structures

Commented [JJ38]: Are you saying that you had to remove the structures for which errors were present?

I'd split into:

The thing that generates features encountered errors for some structures.
Explain why there were errors if you know.
The features with errors in were removed.
Hence the featurised dataset is smaller than the original.

Commented [JJ39]: There is a hyphen between z and score so write as 'z-score'. Z doesn't need to be capitalised

Commented [JJ40R39]: Need to correct in figure

$$z = \frac{(x - \mu)}{\sigma} \quad (5)$$

Where: z = z-score, x = data value, μ = mean of dataset, σ = standard deviation of dataset.

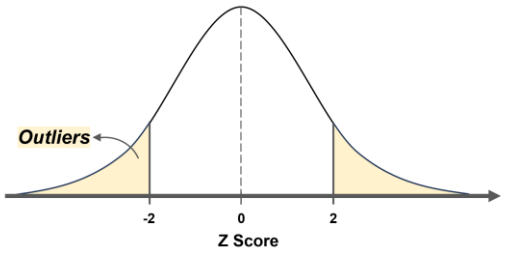


Figure 8. *Defining outliers using z-score for a normally distributed dataset.*

Feature Importance

The importance of each feature was calculated by measuring the decrease in predictive accuracy of the model upon assigning the feature a random value.

Commented [JJ41]: 'for' or 'of'?

Commented [JJ42]: I kind of get it but for clarity, I'd go step by step. In order to determine importance, each feature was assigned a random value and you found accuracy (accuracy of what?)

Did this X times, found mean decrease.

One with biggest mean decrease was most important, one with smallest...

Also idg why it's always a decrease in accuracy. Surely some will result in an increase?

4.1.3 Fine-tuning *MOFormer*

Generating MOFids

MOFormer generates predictions from MOFids, text-based representations of a MOF structure. MOFids for the structures in the MOF-TC dataset were kindly provided by a postdoctoral fellow in my research group. These MOFids required adjustment to be understood by *MOFormer*. The relevant scripts can be found on GitHub. MOFids were successfully generated for 9,458 structures from the MOF-TC dataset.

Model Generation

The pretrained *MOFormer* and CGCNN models, and scripts for training and testing were provided by the *MOFormer* authors.[62]

A block of code was written and added from line 118 to the *MOFormer* fine-tuning script (finetune_transformer.py), as shown in Figure 9. This allows the encoder layers of the Transformer to be frozen, leaving only the regression head to be trained. This code also provides a list of parameters and whether they are trainable to the standard output.

```
if self.config['freeze_base']:
    print('Freezing transformer encoder layers')
    for name, param in model.named_parameters():
        if not name.startswith('regressionHead'):
            param.requires_grad = False
summary(
    model=model,
    col_names=["num_params", "trainable"],
    col_width=20,
    row_settings=["var_names"]
)
```

Figure 9. *Block of code added to the finetune_transformer.py script (found in the published MOFormer repository). This freezes the encoder layers of the Transformer, so only the regression head is trained. It also outputs a list of model parameters and whether they are trainable.*

4.2 COF Gas Adsorption Prediction

COF structure and gas adsorption data generated by Mercado *et al.* were retrieved from the database on the Materials Cloud.[85] A representative subset of 10,000 COFs was generated from the over 69,000 COFs in the original database. The *MOFormer* model was fine-tuned for COF gas adsorption followed the same method as that used for fine-tuning for MOF thermal conductivity (Section 4.1.3).

Commented [JJ43]: What is the materials cloud?

Commented [JJ44]: Why is MOFormer italicised?

Commented [JJ45]: Maybe, 'The MOFormer was fine-tuned for COF gas adsorption following the same method as that used for fine-tuning MOF thermal conductivity [section reference showing where you explain fine-tuning for MOF thermal conductivity]'

Generating COFids

The hypothetical COF structures in the COF-GA database were generated by combining two linkers together using different bond types. The linker structures contain Br dummy atoms, that are replaced with the corresponding bonding atoms during COF structure formation. The file names for each structure contain information on the linker and bond type, and topology and catenation information.

COFids, analogous to MOFids, were generated for structures in the COF-GA to provide inputs for *MOFormer* fine-tuning. To generate COFids, a list of linker IUPAC names, (published by the COF-GA authors), were converted to SMILES using the RDKit Python package.[52, 86] Using RDKit, the dummy Br atoms were replaced by the corresponding bonding atoms for each linker. The final SMILES strings were then assembled into a COFid, together with the three-letter topology code and catenation number taken from the COF structure file name. A visualisation of COFid generation is shown in Figure 10.

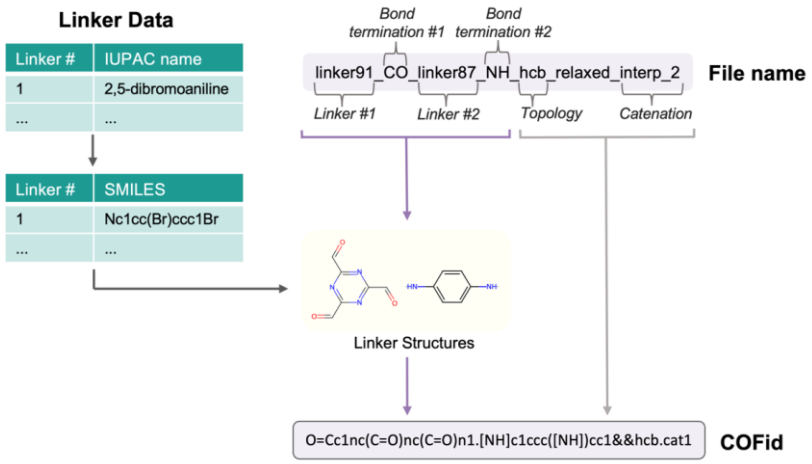


Figure 10. *Method for generating a COFid for a COF structure in the COF-GA dataset. Linker data is taken from reference [79], and lists the IUPAC name for each linker. The file name of each COF structure contains linker, bond, topology and catenation information. The RDKit software package is used to connect linkers with the corresponding bond termination atoms. A COFid is generated from SMILES strings of node and linker structures, combined with topology and catenation information.*

5. Results and Discussion

Commented [JJ46]: oooooooo

Commented [JJ47]: I have no idea what any of this means so can't help here sorry

Commented [JJ48]: Ensure caption is on same page before submitting

This section will discuss the results of fine-tuning *MOFormer* for both MOF thermal conductivity and COF gas adsorption using transfer learning. This will include a detailed discussion of *MOFormer* and the MOF-TC and COF-GA training datasets. Additionally, a tree-based model will be trained directly on the MOF-TC dataset, to provide insights on modelling thermal conductivity.

To assess the accuracy of models, MAE, RMSE and R^2 metrics introduced in Section 2.2.2**Error! Reference source not found.** will be used. It is important to note that MAE and RMSE can only be used to compare performance of models with the same dataset, as it is a non-normalised metric.

Commented [JJ49]: Add a reference to the section it was introduced in instead of this

5.1 Predicting MOF Thermal Conductivity

This section will discuss the results of tree-based models and a fine-tuned *MOFormer* model and evaluate their effectiveness in predicting MOF thermal conductivity. Particular attention will be given to the predictive performance for thermal conductivities between 1 to 3 W m^{-1} K^{-1}, which are necessary to achieve practical gas loading rates in gas storage devices.[87]

Commented [JJ50]: What's a loading rate?

As discussed in Section 2.2.1, the highly porous structure of MOFs impedes efficient thermal transport, meaning that MOFs typically have very low thermal conductivities. This is reflected in the MOF-TC database, where 97% of structures have thermal conductivities less than 1 W m^{-1} K^{-1} (Figure 11).[41] This results in a lack of datapoints representing high thermal-conductivity structures (between 1 to 3 W m^{-1} K^{-1}), making generating predictions for high thermal conductivity MOFs challenging.

Commented [JJ51]: Figure 6 doesn't show this?

Commented [JJ52R51]: I think you mean Figure 11

The following experiments aim to evaluate the effectiveness of transfer learning to overcome this lack-of-data problem. Section 5.1.1 will present the results of a tree-based model trained directly on the MOF-TC dataset, which will be compared to the

MOFormer model fine-tuned for MOF thermal conductivity presented in Section 5.1.2. The results of both experiments will be discussed in further detail in Section 5.1.3.

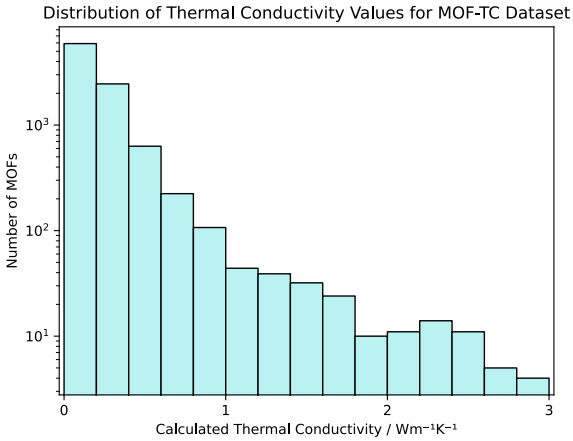


Figure 11. *Distribution of MOFs from the MOF-TC dataset with thermal conductivity values below 3 W m^{-1} K^{-1}. The dataset is highly skewed towards low thermal conductivity MOFs (< 1 W m^{-1} K^{-1}). There are only 194 MOF structures with desired thermal conductivity values (between 1 and 3 W m^{-1} K^{-1}) for practical gas storage devices.*

5.1.1 Direct Training with Tree-Based Models

To begin, tree-based models were used to model the MOF-TC dataset. Given the present lack of predictive models for MOF thermal conductivity in the literature, it is hoped this experiment will provide novel insights into modelling thermal conductivity. These directly trained models will also provide a useful benchmark to assess the performance of the *MOFormer*-based model against.

Commented [JJ53]: What insights? Are you just trying to say that there's a lack of predictive models and that your work will begin to fill the gap?

Before model training, the MOF-TC structural data were converted into machine-readable features. A total of 35 features known to affect thermal conductivity were generated, including surface area, topology, atomic composition and node and linker properties.[41-43, 88] A full list is shown in Table 1.

Commented [JJ54]: Maybe add a reference to the section where you explained why this was necessary

Table 1. *A list of structural features extracted from the MOF-TC dataset.*

Commented [JJ55]: Table captions should be above the table. Different to figure captions

Pore-related properties	Geometric properties	Compositional properties
• Pore limiting diameter. • Largest capacity diameter. • Global cavity diameter. • Probe-occupiable void fraction. • Surface area accessible to a spherical probe.	• Unit cell parameters (a, b, c, α, β, γ). • Number of vertices and edges. • Topology. • Volume. • Density. • Dimensionality.	• Number of carbon, hydrogen, nitrogen and oxygen atoms per unit cell. • Chemical formula. • Node/linker length. • Node/linker/metal mass. • Node/linker SMILES structure. • Node connectivity.

Initial Results

Random forest (RF) and XGBoost models were trained and tested on the structural feature data. Grid search and random search hyperparameter optimisation methods were also used for each model (Table S1).[71] The performance metrics of the best performing RF and XGBoost models are shown in **Error! Reference source not found.**.

Commented [JJ56]: Have you explained what these methods are? If not, add references to sources that explain them

Table 2. Results for the best performing random forest and XGBoost models directly trained on the MOF-TC database.

Commented [JJ57]: Move above table

Model	**MAE** / W m^{-1} K^{-1}	**RMSE** / W m^{-1} K^{-1}	**R^2**
Random Forest	0.11	0.71	0.35
XGBoost	0.15	0.68	0.39

Both models have similarly poor overall performance. The R^2 values show that less than half of the variance of thermal conductivity is predicted by both models. The errors for both MAE and RMSE are also poor. The most accurate MAE value is 0.11 W m^{-1} K^{-1}, which is poor considering that 25% of structures have thermal conductivities below 0.1 W m^{-1} K^{-1}. (The margin of error is itself bigger than the thermal conductivity value of 25% of structures).

Commented [JJ58]: Are you saying that the margin of error is itself bigger than the conductivities themselves? Maybe say so explicitly

The model predictions have been illustrated in Figure 12. Predicted thermal conductivity is plotted against calculated thermal conductivity for the unseen test

dataset. For a perfect predictive model, all lines would fall on the line of equality (y=x, shown in grey on the inset plot). Points above the line are overpredictions, and points below it are underpredictions.

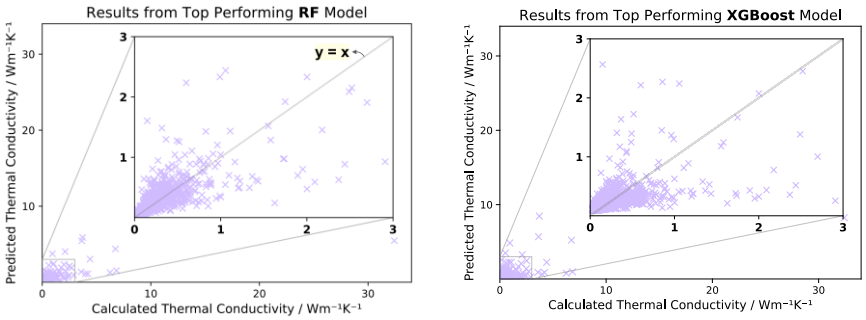


Figure 12. *Results of predicted and calculated thermal conductivity for the best performing RF and XGBoost models.*

For both graphs in Figure 12, the most immediate observation is the high concentration of low thermal conductivity structures, reflecting the skewed distribution of the MOF-TC dataset. Both the RF and XGBoost models have highly dispersed predictions for values higher than 1 W m^{-1} K^{-1}. This decrease in model accuracy with increasing thermal conductivity is to be expected. As there are fewer data points with high thermal conductivities to train on, it is harder to extract a robust trend for these points and hence their prediction accuracy decreases.

Model accuracy decreases significantly from the training data to the test data (Table 3) - a sign the model is overfitting to the training data. (A textbook example of overfitting can be seen in Figure 13.)

Instead of learning the underlying trend between structure and thermal conductivity during training, the models have learnt to replicate the distribution of the training data, thus leading to poor performance on the unseen test data. The model may also be overfitting to the more abundant low thermal conductivities, which could be a further contributor to poorer predictions of high thermal conductivity values.[89]

Commented [JJ59]: Y=x is called the line of equality

Commented [JJ60]: Rather than saying 'highly dispersed', maybe refer directly to 'spread', which is the statistical property you're alluding to here

Commented [JJ61]: Where's the training data and where is the test data in the figure?

Commented [JJ62]: Isn't this the same thing as what you're saying in the previous sentence when you talk about replicating the distribution of the training data?

Table 3. *RMSE performance for best performing random forest and XGBoost models on training and test sets. Performance degrades from training to test sets as a result of overfitting.*

Model	RMSE / W m^{-1} K^{-1}	
	Training Set	**Test Set**
Random Forest	0.20	0.71
XGBoost	0.28	0.68

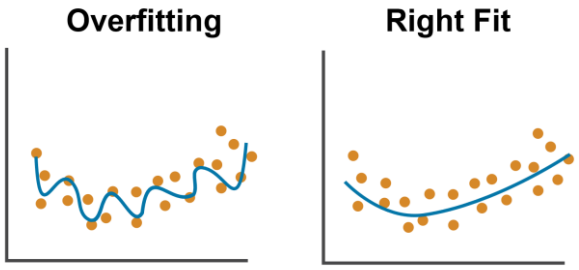


Figure 13. *Textbook examples of overfitting and what a good fit looks like. From reference [90].*

Commented [JJ63]: Should be 13(a) and 13(b), not left and right. Also, don't place a table within a figure. Create a separate table with associated caption, and a separate figure

Commented [JJ64R63]: When you say 'the same dataset' at the end of your caption here, it sounds like you're talking about the dataset you used in this report.

Commented [JJ65R63]: Also, I don't think you're referencing the image from ref 91 correctly. Consult Imperial's guide on referencing and check how to reference images properly.

Commented [JJ66]: Big gap

Tackling Data Imbalance

Both models perform poorly when predicting high thermal conductivity MOFs between 1 to 3 W m^{-1} K^{-1}. This is likely a consequence of the lack of high thermal conductivity structures in the training data. More specifically, the *proportion* of the dataset consisting of high thermal conductivity structures is very small. This means that during training, the model does not have enough examples to learn from to generate accurate predictions.

One way to combat this data imbalance is stratified sampling, which ensures that there is a proportional number of high thermal conductivity structures in the training and test set. For example, if the training set comprises of 80% of the entire dataset, then it should include 80% of high thermal conductivity structures.

Commented [JJ67]: Don't get this - I think there's a mistake. Are you trying to say that if the training set has 80% high thermal conductivity structures, then the test dataset should also have 80% high TC structures?

An RF model was trained on a stratified dataset and the results are shown in **Error! Reference source not found.** below. Though the MAE is similar to the non-stratified model, there is an improvement in RMSE and R^2 values due to stratification, indicating that there are fewer extremely inaccurate predictions. However, when plotted, the spread of high thermal conductivity values remains large (Figure 14). This suggests that despite correcting the *proportion* of high thermal conductivity MOFs, there are still too *few* high thermal conductivity structures to generate an accurate prediction. Similar results are found when training an XGBoost model on stratified data (Table S2).

Commented [JJ68]: Referring to spread as I mentioned earlier

Commented [JJ69]: Be more clear what 'their' is referring to. Also don't use 'numbers' because it sounds like you might be talking about their actual values

Commented [JJ70]: Why have you called the section at the end 'Supporting Info'? Shouldn't it be called 'Appendix'

Table 4. *Results for random forest models with and without data stratification. Stratification results in improved performance in terms of RMSE and R^2, but little improvement in MAE.*

Commented [JJ71]: Table caption above

Model	**MAE** / W m^{-1} K^{-1}	**RMSE** / W m^{-1} K^{-1}	**R^2**
Random Forest	0.11	0.71	0.35
*Random Forest with **stratified** data*	0.10	0.39	0.70

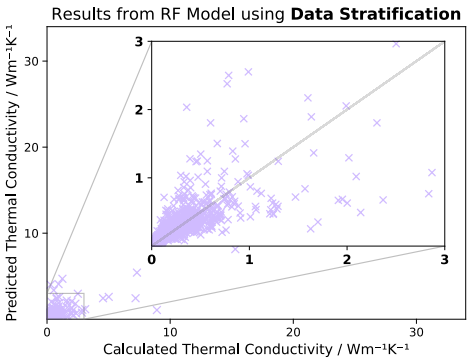


Figure 14. *Plotted results of calculated and predicted thermal conductivity of a random forest model using data stratification. High thermal conductivity predictions, between 1-3 W m^{-1} K^{-1}, remain inaccurate.*

Commented [JJ72]: You need to cite the figure in the body of the text. In the paragraph above this ideally. Reading that paragraph I didn't realise you had actually included the figure in this report.

Also as a general rule for reports, if you include a figure, you have to cite it in the body of the report

The improvement as a result of stratification is likely due to a decrease in overfitting to low thermal conductivity values. This is evidenced by the similar training scores for the RF models with and without stratification, as shown in Table 5. This further

Commented [JJ73]: Don't get this. I think it stems from my confusion around your explanation of stratified sampling, which needs to be reworded

supports the theory that there is simply insufficient data to develop accurate predictions for high thermal conductivity MOFs from the selected structural features.

Commented [JJ74]: Isn't this just repetition from earlier?

__Table 5.__ RMSE performance for random forest models with and without data stratification. Both models display similar accuracy on the training set, but the model using stratification has a higher accuracy on the test set – indicating reduced overfitting.

Commented [JJ75]: Table caption above

Model	RMSE / W m^{-1} K^{-1}	
	Training Set	**Test Set**
Random Forest	0.20	0.71
*Random Forest with **stratified** data*	0.23	0.39

Physical Interpretation of Results

Ideally, the predictive model would learn the *physical* relationship between a MOF structure and its thermal conductivity. This would increase the applicability of the model and may provide novel insights into the physical drivers of the structure-property relationship.

Commented [JJ76]: Reliability has a statistical meaning unrelated to what you're saying here. Also regardless of this, I don't think it's the right word to convey what you're trying to say. Be more explicit

In this section, the parameters of the trained tree-based models for MOF thermal conductivity will be investigated to obtain a physical interpretation. This will also help to identify whether their poor predictive performance is solely due to insufficient data, or whether there is also a lack of understanding of the underlying structure thermal conductivity relationship.

Commented [JJ77]: Understanding is not the right word because you're implying the model itself has some kind of insight/intelligence. It doesn't - you are gleaning the insight

Physical insights can be determined from the relative importance of each input structural feature to the accuracy of the model. The model's feature importances can then be compared with important features established in the literature.

As discussed previously in Section 2.2.1, thermal conductivity is correlated to density and porosity. When analysing the MOF-TC dataset, Wilmer and colleagues found that

density and structural features relating to porosity, such as pore size, play a big role in determining thermal conductivity.[41] Similarly, linker length and mass mismatch between node and linkers are also expected to appear as significant contributors to thermal conductivity in the ML model.

Commented [JJ78]: Why?

The top 10 most important features for the previously presented RF model are shown below (Figure 15). Notably, the top two most important features have a much larger effect on model accuracy than the others.

Largest capacity diameter (LCD), the most important calculated feature, describes the largest diameter sphere that can move through a structure's pores. This is physically related to thermal conductivity, as it is linked to porosity. However, it is surprising that this is the most important feature; within the literature, density is considered to have a better correlation with thermal conductivity.

Another surprising observation is that the number of carbons per unit cell is ranked third most important. It is unclear why the number of carbons is more important than the number of oxygens, nitrogens or hydrogens, which were also input features. A possible reason is that the number of carbons is likely more closely correlated to density, but then one would expect density to be of similar, if not greater importance.

Commented [JJ79]: You've written the whole report in the third person so don't switch to first person now

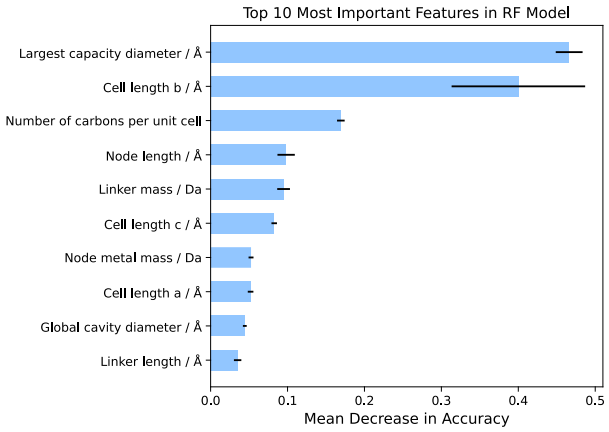


Figure 15. *Top 10 most important features in the random forest model. The importance of each feature is determined by the mean decrease in predictive accuracy when the feature is assigned a random value. A larger decrease in accuracy corresponds to greater importance of that feature to the model. Error bars show the standard deviation of each feature importance.*

Commented [JJ80]: ? I think I mentioned that this was unclear earlier in the report too. If you explain what you mean at that earlier point in the report, you can add a section reference here back to that earlier point - so that the reader can refer back if they are confused

Perhaps the most unexpected result is that *cell length b* is ranked as the second most important feature. There is no physical relationship between thermal conductivity and this particular cell length. *Cell length b* is also very similar to *cell lengths a* and *c*, which are ranked eighth and sixth respectively, seemingly arbitrarily.

Commented [JJ81]: Have you measured the correlation/found the correlation in the literature? Can't keep using the statistical term 'correlation' unless you actually have experimental evidence to back whatever assertion it is you're making with that word

Commented [JJ82R81]: Btw, however you word this, you need to find an external reference to back the assertion you're making

From these feature importances, it can be concluded that the RF model does not rely on physically understood relationships between MOF structure and thermal conductivity to generate its predictions. This may be a cause of the poor predictive performance of the tree-based models, in addition to the lack of data. This is not unexpected, as thermal conductivity is a challenging property to predict; phonon transport depends on a large number of complex variables which are still not fully understood, especially for highly porous extended MOF structures.[40]

To summarise, the directly trained tree-based models outlined above cannot overcome the scarcity of data or the challenges in modelling thermal conductivity, and hence perform poorly in predicting the desired high thermal conductivity MOFs.

5.1.2 Fine-tuning *MOFormer* for Thermal Conductivity Prediction

This section will discuss fine-tuning the pretrained *MOFormer* model for MOF thermal conductivity using the MOF-TC dataset. It is hoped that a more sophisticated neural-network based model combined with chemical intuition developed from pretraining on 400,000 MOF structures will improve predictive accuracy compared to the tree-based models discussed in the previous section.

Each structure from the MOF-TC dataset was converted into a MOFid – a text-based description of a MOF structure used as an input to *MOFormer*.[62, 91] MOFids describe topology, catenation and node and linker structural information. In *MOFormer*, the MOFid is first passed through multiple encoder layers. In a process called embedding, these encoders learn how to best represent the MOFid so as to optimise model performance. Next the regression head learns to model the relationship between this learned embedding and the desired property (thermal conductivity). This process is illustrated in Figure 16.

Commented [JJ83]: Is this a widely understood word?

Commented [JJ84]: Add a reference to figure 16 in the body of the report

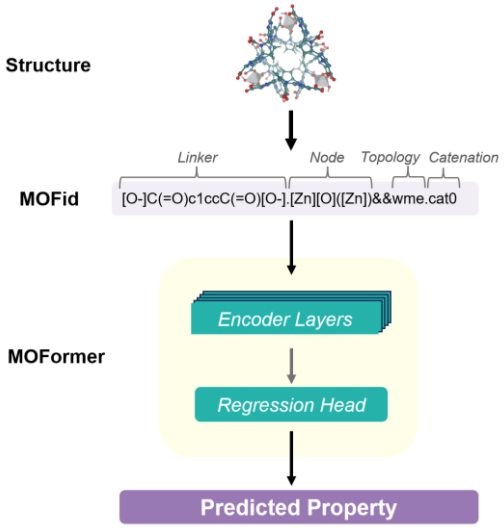


Figure 16. *MOFormer workflow. The MOF structure is converted into a MOFid. MOFid is inputted into MOFormer, where it is fed through multiple encoder layers and a regression head which then generates a property prediction.*

MOFid is a greatly simplified representation of an actual MOF structure. To compensate for this, pretraining of *MOFormer* occurs alongside a CGCNN, which takes the entire MOF structure as an input. Pretraining focuses on maximising the similarity of the embedded MOFids with the embedded MOF structures from the CGCNN. This step improves the accuracy of *MOFormer* and the CGCNN in predicting MOF gas adsorption.

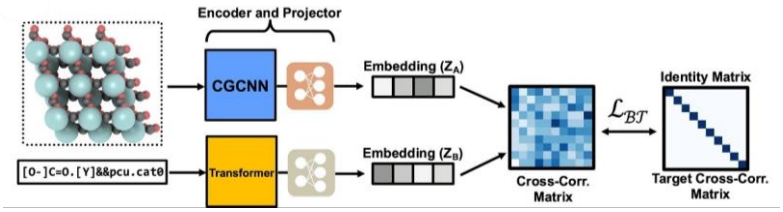


Figure 17. *MOFormer pretraining alongside a CGCNN. Pretraining aims to maximise the similarity between the CGCNN and Transformer embedding of the same MOF structure. Figure taken from reference* [62].

For thermal conductivity prediction, *MOFormer* will be initialised with the weights from the pretrained model before being retrained on the labelled MOF-TC dataset. *MOFormer* will also be initialised with random weights and trained on MOF-TC from scratch to determine the effectiveness of pretraining in improving model accuracy.

Similarly, the CGCNN model will also be trained using pretrained weights and from scratch. This will provide a further transfer learning experiment, as well as insights into how a different model and MOF structural input will affect model predictions.

Model Results

The performance of *MOFormer* models – trained from the pretrained weights and also separately from scratch (initialised with random weights) – is shown in Table 6. Pretraining improved model prediction as measured by MAE by 26%. This indicates that the information gained during pretraining improves MOF thermal conductivity prediction.

Table 6. *Results for MOFormer model initialised with pretrained weights and random weights (from scratch). Pretraining improves performance compared to the scratch model. Compared to the directly trained random forest model, pretraining performs slightly worse.*

Commented [JJ85]: Table caption above

Model	**MAE** / W m^{-1} K^{-1}	**RMSE** / W m^{-1} K^{-1}	R^2
MOFormer *trained from* ***scratch***	0.19	0.77	0.02

***MOFormer** fine-tuned **pretrained** model*	0.14	0.66	0.27
*Directly trained **RF** model (Discussed in Section 4.1.2)*	0.11	0.71	0.35

Commented [JJ86]: Maybe add a reference to the section where this came from

However, the 26% improvement observed above is misleading. When compared to the best directly trained tree-based model (discussed in Section 5.1.1), pretraining shows no significant improvement in model performance. This suggests that the improvement in accuracy between the *MOFormer* model trained from scratch and from the pretrained model is primarily due to the poor performance of the former, rather than the strong performance of the latter. This is confirmed when the results are plotted, as in Figure 18.

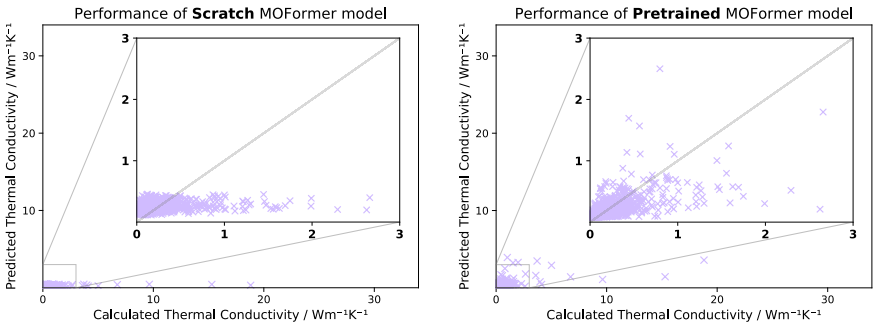


Figure 18. *Results for MOFormer model trained from scratch and from the pretrained model.*

From Figure 18. Results for MOFormer model trained from scratch and from the pretrained model.Figure 18 it is clear that the model trained from scratch is highly inaccurate, with predictions restricted between 0.1 and 0.5 W m^{-1} K^{-1}, forming a horizontal band. This may indicate that the model trained from scratch is overfitting to the training data, and only predicting the most abundant thermal conductivities found in the MOF-TC dataset (0.2 - 0.4 W m^{-1} K^{-1}, see Figure 11). This is not unexpected; deep learning methods are prone to overfitting due to their high number of tuneable parameters.[92]

Commented [JJ87]: Isn't it pretty bad in this range too?

Commented [JJ88]: Can you give it a different name? 'scratch' doesn't really work, it should be 'from scratch' if you're going along these lines. But it still sounds weird

The pretrained model displays a better correlation than the model trained from scratch, but the predictions are still dispersed – especially for high thermal conductivities (1 - 3 W m^{-1} K^{-1}). Predictions for high thermal conductivities tend to be smaller than the true values, with most points below the y=x line. This is indicative of overfitting to the more abundant low thermal conductivity values (< 1 W m^{-1} K^{-1}), a result of the imbalanced training set.

Commented [JJ89]: From scratch/something else

Commented [JJ90]: Wdym? Also stop saying 'with respect to'

Fixing the Cut-Off

A closer look at the pretrained model's low thermal conductivity predictions reveals a cut-off at approximately 0.1 W m^{-1} K^{-1}, below which there are no predicted values (Figure 19). This is unexpected given that25% of the data is below 0.1 W m^{-1} K^{-1}. Hence, it suggests a problem not with the data, but with the *MOFormer* model itself.

Commented [JJ91]: Label the cut-off line on the figure

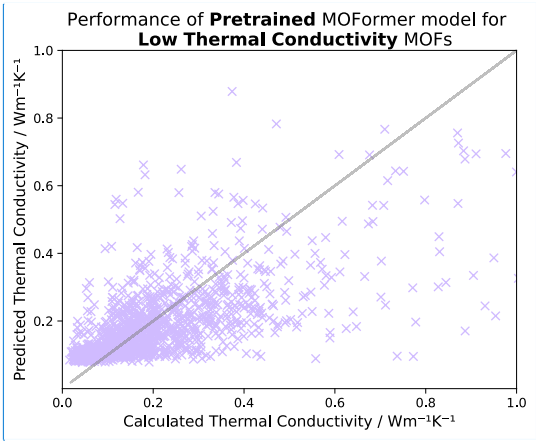


Figure 19. *Performance of the pretrained MOFormer model for low thermal conductivity (<1 W m^{-1} K^{-1}) data. There is a cut-off that can be seen at ~0.1 W m^{-1} K^{-1}.*

Commented [JJ92]: Put figure 19 after the place where you mention it in the text, not before

Two approaches were used in an attempt to remove this cut-off, with the aim to improve the accuracy of predictions for both low and high thermal conductivity values.

The first method involves the length of MOFids. As all MOFids must be the same length to be read by *MOFormer*, MOFids with more than 512 characters are truncated. These truncated MOFids may cause some inaccuracies in the model, leading to the peculiar cut-off. It is therefore worth investigating whether fine-tuning the model using non-truncated MOFs alone improves performance.

Commented [JJ93]: I went with 'non-truncated' in the end because I saw you used it in figures below. But it should be 'untruncated'

The second approach freezes the encoder layers of *MOFormer*, so the layers preserve the pretrained weights. Only the regression head responsible for predictions is retrained. This allows the embedding learned during pretraining on 400,000 structures alongside a CGCNN, to be retained. This learned embedding has likely developed a more comprehensive understanding of the relationship between a 3D MOF structure and MOFid. Hence preserving this learned embedding may help improve model prediction.

Commented [JJ94]: 'developed an understanding' makes it sound like the algorithm is actually learning in the way a human would. Reword

The ability of these two approaches to improve the prediction of low thermal conductivity MOFs can be assessed from the results plotted in Figure 20. Looking at the two graphs, it is clear that both methods have succeeded in removing the cut-off at ~0.1 W m^{-1} K^{-1}.

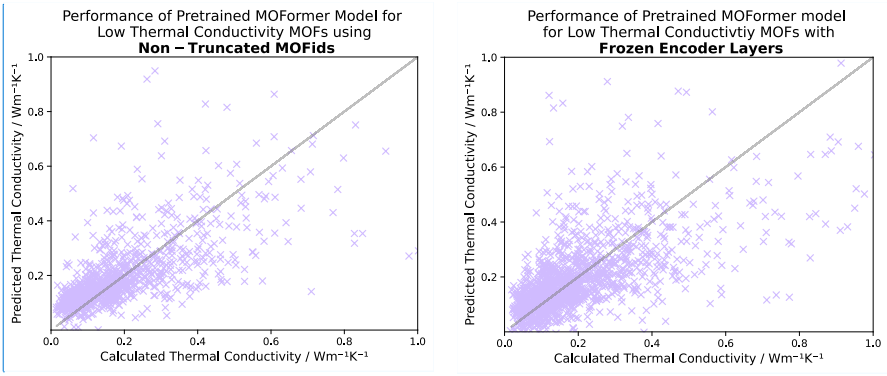


***Figure 20**. Results for low thermal conductivity data (<1 W m^{-1} K^{-1}) for pretrained MOFormer models using only non-truncated MOFid as inputs, and freezing the MOFormer encoder layers.*

Commented [JJ95]: I would use a and b throughout the report for figures like this and refer to a or b specifically in the body of the text as needed. But if you cba leave it

Commented [JJ96]: Place figure after first reference in body of report

For the model using only non-truncated MOFids, prediction accuracy in terms of RMSE and R^2 decrease compared to the original pretrained model (Table 7). This decrease

in performance is likely due to a decrease in the size of the training data by 30%, as a result of removing overly long MOFids.

Table 7. Results of pretrained MOFormer models, with non-truncated MOFids and encoder layers frozen.

MOFormer Model	**MAE** / W m^{-1} K^{-1}	**RMSE** / W m^{-1} K^{-1}	R^2
*Model fine-tuned from **pretrained** model*	0.14	0.66	0.27
*Pretrained model trained on **non-truncated MOFids***	0.19	0.98	0.36
*Pretrained model trained with **frozen encoder layers***	0.15	0.60	0.41

In contrast, the fine-tuned model where the encoder layers were frozen displays a slight improvement in accuracy in terms of RMSE and R^2 compared to the original pretrained method (Table 7). This may be because the embeddings learned during *MOFormer* pretraining have a more comprehensive understanding of how MOFid relates to a 3D MOF structure.

Despite improving performance for low thermal conductivities, Figure 21 shows that the two methods remain poor at predicting high thermal conductivity values. This indicates that the *MOFormer* models still struggle to make up for the scarcity of data for high thermal conductivity MOFs.

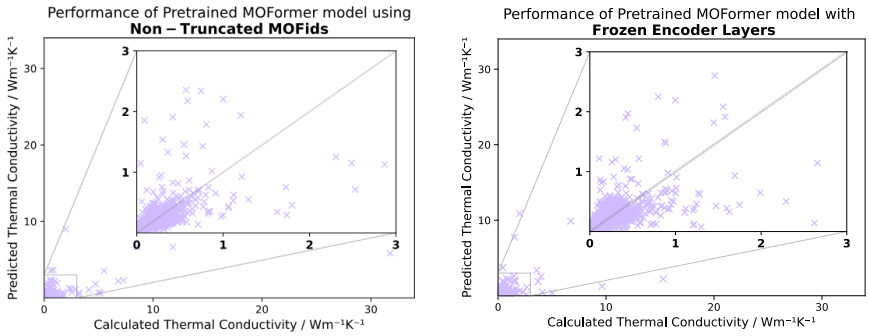


Figure 21. Results for pretrained MOFormer models using only non-truncated MOFid as inputs, and freezing the MOFormer encoder layers.

Commented [JJ97]: Table caption above

Commented [JJ98]: Replace - they don't 'understand' anything

Commented [JJ99]: Place figure after first reference in body of report

CGCNN Results

Shown below are the results from training the CGCNN model from scratch and initialising using pretrained weights. **Error! Reference source not found.** shows that retraining had little effect on MAE, but decreased model accuracy as measured by RMSE and R^2. This supports the results from the previous *MOFormer* models, where it is shown that pretraining has little effect on improving model accuracy. Similar to *MOFormer* models, visualisation of the results shows highly dispersed and inaccurate predictions for high thermal conductivity structures (Figure 22).

Table 8. Results for CGCNN models trained from scratch and from a pretrained model.

Commented [JJ100]: Table caption above

CGCNN Model	**MAE** / W m^{-1} K^{-1}	**RMSE** / W m^{-1} K^{-1}	R^2
*Model trained from **scratch***	0.15	0.80	0.30
*Model trained from **pretrained** model*	0.14	0.89	0.13

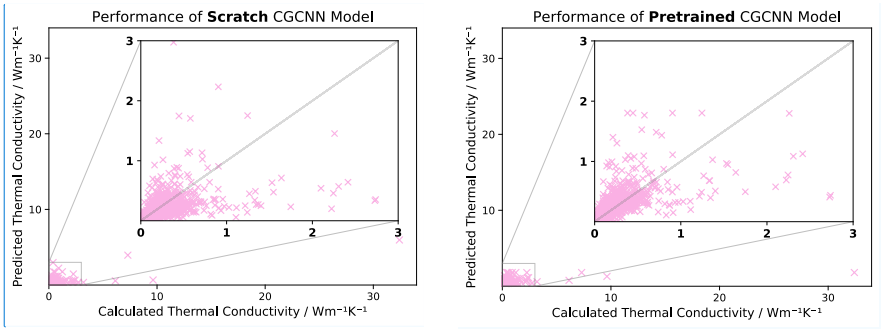


Figure 22. Calculated and predicted values for CGCCN models trained from scratch and from a pretrained model.

Commented [JJ101]: Reference figure in body

It is interesting to observe that CGCNN models are less accurate than the *MOFormer* models, despite learning from more detailed inputs - crystal graphs that represent the entire MOF structure. One reason for this might be that crystal graphs are too detailed, making it difficult to extract the structural features important to thermal conductivity. This extraction may be easier with MOFids, that represent only the key features of each MOF structure.

5.1.3 Outlook for MOF Thermal Conductivity Prediction

The results in this section show that transfer learning had little effect in improving MOF thermal conductivity predictions. All models performed poorly in predicting the high thermal conductivity values that are desired for gas storage applications.

Commented [JJ102]: Are they predicting the conductivity or the MOFs themselves? Because this bit I'm highlighting implies the latter

The most apparent reason for this poor performance is the sparsity of high thermal conductivity data. MOFs tend to have low thermal conductivities, in part due to their high porosity, resulting in less high thermal conductivity data. Transfer learning fails to overcome this scarce-data problem, with the best pretrained model generating results of similar accuracy to a much simpler tree-based model. One might suggest the generation of larger datasets to resolve the data scarcity issue. But this approach presents its own problems It is highly time-consuming and resource intensive to obtain simulated conductivity values; the 10,000 structures in the MOF-TC database would require 80,000 calculations.

Another potential reason for the inaccurate predictions of MOF thermal conductivity is the fundamental difficulty in modelling the structure-thermal conductivity

relationship for a MOF. Even for the much simpler case of monoatomic 2D graphene, thermal conductivity can vary by up to 74% for structures with the same porosity depending on the specific pore distribution on the graphene sheet.[93] The complexity in MOF thermal conductivity prediction was demonstrated by the relative feature importances in the RF model, which were seemingly arbitrary and did not reflect any known physical relationships (Section 4.1.2). This inherent difficulty may also be why there are no studies in the literature putting forth an ML model for MOF thermal conductivity prediction.

It was anticipated that the pretrained *MOFormer* model would develop enough chemical intuition to be able to overcome both the limited data and the difficulty in modelling thermal conductivity (Section **Error! Reference source not found.**). One reason this has failed may be the lack of structural overlap between the MOF-TC dataset and the dataset *MOFormer* was pretrained on - only six structures were common to both datasets. It would be interesting to see whether performance would improve with pretraining on structures more similar to those in the MOF-TC database.

A further possible reason for the lack of improvement when fine-tuning *MOFormer* for thermal conductivity is that it was developed for the prediction of static properties - band gap and gas adsorption. Attempting to fine-tune for thermal conductivity, a more complicated transport property, is perhaps too far from *MOFormer*'s original purpose for transfer learning to be successful in this case.

One way to improve thermal conductivity predictions would be to change the structural representation of MOFs. Since thermal conductivity is a dynamic property that describes phonon transport, conventional structural features may not be the best representation of MOFs. Cheng *et al.* reported that there was no direct correlation between common structural properties and thermal conductivity for the 18 MOFs they studied.[94] They instead created new features describing the orientation and distribution of phonon transport pathways, that were found to have a positive correlation with thermal conductivity.[94] This presents a potential route for further work

Commented [JJ103]: Add reference to relevant section

Commented [JJ104]: Where? Reference the section

Commented [JJ105]: Is this a proposal for further work?

Commented [JJ106]: What you're implying here is that transfer learning is the reason for the failure. As if the process of using MOFormer for prediction for thermal conductivity is itself transfer learning.

Message me if unsure what I'm getting at here

Commented [JJ107]: 18 MOFs *they* studied? If so, include the word 'they'. Otherwise unclear whether it's 18 MOFs you studied

in developing more accurate predictive models using features more relevant to thermal conductivity.

A final challenge of modelling thermal conductivity is that there is no simple way of confirming whether predictions are accurate, due to difficulties in making experimental measurements. Calculated thermal conductivities of hypothetical MOFs do not consider attributes such as defects, packing disorder, grain boundaries or adsorbate effects. These attributes could cause a discrepancy between experimental and simulated thermal conductivity values.[76,95] Though less challenging to obtain than experimental measurements, simulating thermal conductivities is highly time-consuming and resource intensive.

Commented [JJ108]: Idg this sentence. Split into two

Commented [JJ109]: Repetition? I think you mentioned the 80,000 calculations earlier in the report

5.2 Predicting COF Gas Adsorption

In this section, *MOFormer* will be fine-tuned for COF gas adsorption prediction. As in the previous section, both a pretrained model and a model trained from scratch will be produced to evaluate the benefit of pretraining on model prediction.

Commented [JJ110]: Reword - see my previous comments on this

The relationship between a COF structure and its gas adsorption is more straightforward, so it should be easier to develop a predictive model for than thermal conductivity. The main question to be answered by these experiments is how well a model pretrained on MOFs can be used for COF property prediction. Here, the success of transfer learning depends on the structural similarity of MOFs and COFs as understood by the machine.

Commented [JJ111]: This doesn't make sense

10,000 structures and their gas adsorption properties from the COF-GA database will be used for model training and testing. Each structure has two gas adsorption values, one at a high loading pressure (65 bar) and one at a low discharging pressure (5.8 bar).[96] Here, we address each case separately, Section 5.2.1 will discuss predictions at high pressure and Section 5.2.2 will discuss predictions at low pressure.

5.2.1 High Pressure Results

The results for the *MOFormer* models for COF gas adsorption at high pressure (65 bar) are shown below in **Error! Reference source not found.**. Both models show promising results, with average R^2 values of 0.9, meaning 90% of the variance in gas adsorption can be explained by either model. This suggests that the models have a comprehensive understanding of how a COF structure relates to gas adsorption. This is further supported by the fact that accurate predictions are obtained for higher gas adsorptions (> 60 mol kg^{-1}), despite with the fact there are fewer datapoints in this range (Figure 23).

Commented [JJ112]: See previous comments on use of the word 'understanding'

Table 9. Results for MOFormer model for COF gas adsorption prediction at 65 bar. Models are trained from scratch, from a pretrained model with and without freezing the encoder layers.

Commented [JJ113]: Table caption above table

MOFormer Model	**MAE** / mol kg^{-1}	**RMSE** / mol kg^{-1}	R^2
Model trained from ***scratch***	2.18	3.47	0.89
Model trained from ***pretrained*** *model*	2.05	3.30	0.91
Model trained from ***pretrained*** *model with* ***encoder layers frozen***	4.82	6.58	0.63

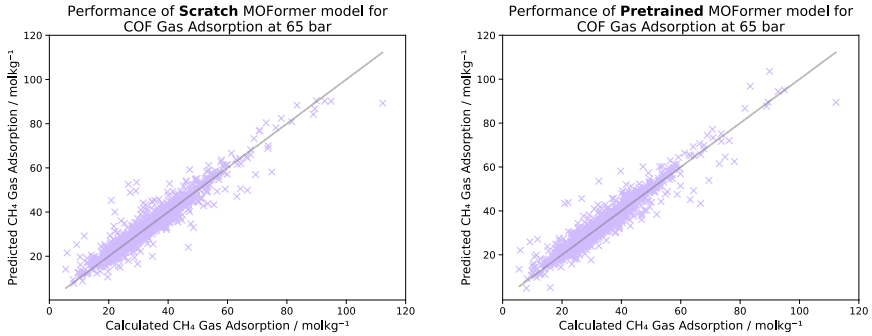


Figure 23. *Results for MOFormer model for COF gas adsorption prediction at 65 bar. The models were trained from scratch and from a pretrained model.*

The pretrained model has better predictive power than the model trained from scratch, with an improvement of 5.8% as measured by MAE. This is an improvement on the average accuracy of fine-tuning for MOF gas adsorption (4.9%) as published in the original *MOFormer* paper.[62] This is significant as it demonstrates that MOFs and COFs are sufficiently structurally similar that transfer learning on a pretrained MOF model for COF property prediction is effective.

Commented [JJ114]: Add reference to the paper and where in the paper this comes from

These results also show that COFids, COF equivalents of MOFids, can be used as an accurate representation of COF structures. This is confirmed when the encoder layers

are frozen and the regression head of *MOFormer* alone is tuned. Freezing the encoder layers forces the model to interpret COFids in the same way it would MOFids, resulting in a 2.4-fold increase in prediction error in terms of MAE.

Commented [JJ115]: I don't get this

5.2.2 Low Pressure Results

The results for the *MOFormer* models using pretrained weights and trained from scratch for COF gas adsorption at low pressure (5.8 bar) are shown in Table 10 and Figure 24. These results are significantly worse than the predictive models for gas adsorption at high pressures. However, as was the case for high-pressure models, pretraining is beneficial and improves model performance by 9.6% as measured by MAE.

Table 10. *Results for MOFormer model for COF gas adsorption prediction at 5.8 bar. The models are trained from scratch and from a pretrained model.*

Commented [JJ116]: Caption above table and also need to reference table before first appearance

MOFormer Model	**MAE** / mol kg^{-1}	**RMSE** / mol kg^{-1}	R^2
Model trained from ***scratch***	0.32	0.54	0.54
Model trained from ***pretrained*** *model*	0.29	0.48	0.63

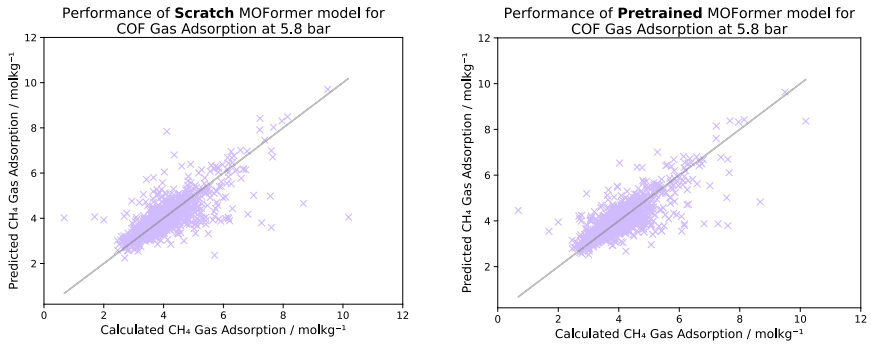


Figure 24. *Results for MOFormer model for COF gas adsorption prediction at 5.8 bar. The models are trained from scratch and from a pretrained model.*

The relatively poor performance of low-pressure predictions is likely due to the difference in dataset distributions. From the distributions of the two datasets shown in Figure 25, it is clear that the distribution of COFs at high adsorption pressures is more uniform, whereas for low pressures there is a more significant skew around 4 mol kg^{-1}. This might be the reason for the inaccurate predictions at low pressures. Techniques such as normalisation or oversampling, which add or remove points to create a more uniform distribution of data, could be used to potentially improve predictions.[97, 98]

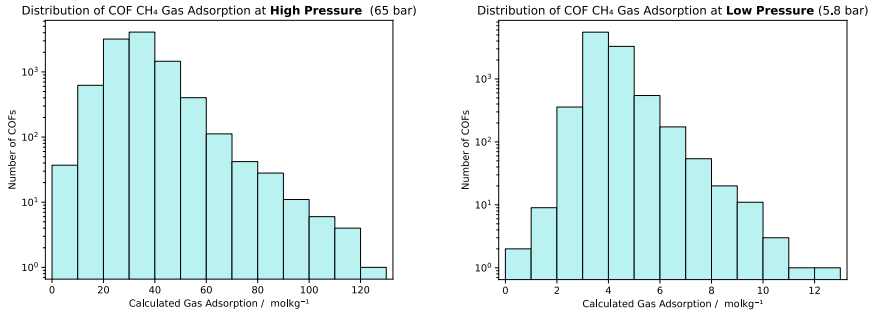


Figure 25. *COF distribution for COF-GA dataset at 5.8 and 65 bar respectively. The 65 bar dataset is more uniformly distributed than the 5.8 bar dataset.*

For both high and low adsorption pressures, the models still achieved strong predictive performance for COF gas adsorption. Both models also benefit from pretraining from the *MOFormer* model, showing transfer learning is successful in this case. The strong performance of the models also confirm that MOF to COF transfer learning is feasible, which presents promising opportunities for COF material discovery. The far larger volume of public MOF data can be leveraged to predict COFs with optimised properties for practical gas storage applications.

For both high and low pressures, the CGCNN models were less accurate than the corresponding *MOFormer* models (Table S3). This may be due to the more complex crystal graph inputs of the CGCNN, increasing the difficulty of extracting only the most important structure-property relationships.

Commented [JJ117]: I wouldn't say that it's more uniform just because the actual distribution is nowhere near uniform. I know what you're trying to get at, but using the word uniform here is not good practice

Commented [JJ118]: What is 'it'?

Commented [JJ119]: Ooo nice sentence!

Commented [JD120R119]: Jamie wuz here

6. Conclusion

MOFormer was fine-tuned using transfer learning for MOF thermal conductivity and COF gas adsorption with differing success. For COF gas adsorption, transfer learning improved predictions by an average of 7.7% as measured by MAE. However, it had little effect on MOF thermal conductivity predictions, showing little improvement compared to tree-based models. Looking at the *MOFormer* models trained from scratch for both properties reveals a similar trend: accurate predictions of COF gas adsorption and inaccurate predictions of MOF thermal conductivity.

It is likely that pretraining is not effective in improving predictions in the latter case due to the inherent challenges in predicting MOF thermal conductivity, demonstrated by the poor accuracy of models trained directly on the MOF-TC dataset. There are two driving factors for this. The first is the MOF-TC dataset is severely imbalanced, resulting in overfitting to low thermal conductivity data and generally poor predictions for the desired high thermal conductivity MOFs. The second is that thermal conductivity is a complex transport property with non-direct correlation to structural features. In contrast, gas adsorption is governed by physisorption which is strongly correlated with structural features such as surface area.[36]

Further work is therefore required before a model for predicting MOFs with optimised thermal conductivity for gas storage applications is developed. Two potential avenues that could be explored are pretraining on more similar MOF structures, or developing bespoke features for thermal conductivity prediction..[94]

In contrast, the *MOFormer* model fine-tuned for COF gas adsorption is successful in demonstrating that a MOF- based model can be repurposed for COF property prediction. This is significant in showing transfer learning can be utilised to develop accurate models using the small amount of available COF data by leveraging the greater volume of available MOF data.

The success of COF gas adsorption models also demonstrates that COFs can be represented by a text string (MOFid), allowing COF structures to be screened faster and at less computational cost compared to generating entire 3D structures. There is also an improvement in accuracy in using MOFid for all *MOFormer* models compared to the CGCNN. This may be because it is easier to extract important features from structures represented by MOFids compared to more complicated crystal graph representations.

Commented [JJ121]: idgi

MOFs and COFs represent a growing field of sustainable materials that can contribute to green gas storage applications, important in our efforts to combat climate change. This project presented initial findings which hope to contribute to the development of ML-based models to accelerate the discovery of novel optimised materials.

Commented [JJ122]: Too opinionated in the context of the conclusion section

6.1 Future Work

There are multiple opportunities for future work to further develop this project.

Commented [JJ123]: Didn't you already discuss further work in the third paragraph of the conclusion?

In order to discover a MOF or COF material suitable for gas storage in vehicles, there are many other properties, such as pore size, thermal stability and enthalpy of adsorption, to be optimised aside from thermal conductivity and gas adsorption capacity.[8] Developing a model that can predict all relevant material properties will require more structure and property data and more research into how the properties interplay.

Commented [JJ124]: What type of data?

This report has shown that transfer learning can be used to successfully retrain a MOF model for COF gas adsorption. Future work could focus on other COF properties for which predictions could be improved when fine-tuning an established MOF model.

Commented [JJ125]: Reword - opinion

The models developed in this project have used computer-generated structures with simulated property values. To find a material viable for practical use, computational work must be conducted in tandem with experimental research. Experiments must be

conducted in the lab to determine the true values of material properties, and the results must be compared with predictions arising from computational models.

On the computational side, future work could focus on assessing the synthesisability of a hypothetical structure, to guide experimental efforts to only the most promising materials. This could be done via a predictive model, or by calculating properties such as free energy to determine experimental feasibility.[61, 99]

Commented [JJ126]: Is this what you're trying to say?

Commented [JJ127]: Isn't this essentially what you said in the intro?

Commented [JJ128]: Idg what free energy has to do with this but might make sense to a chemist?

7. References

1. Programme, U. N. E. *Emissions Gap Report 2023: Broken Record – Temperatures hit new highs, yet world fails to cut emissions (again)*.
2. Ma, S.; Zhou, H.-C., Gas storage in porous metal–organic frameworks for clean energy applications. *Chemical Communications* **2010,** *46* (1), 44-53.
3. Kumar, K. V.; Preuss, K.; Titirici, M.-M.; Rodríguez-Reinoso, F., Nanoporous Materials for the Onboard Storage of Natural Gas. *Chemical Reviews* **2017,** *117* (3), 1796-1825.
4. Knerelman, E. I.; Karozina, Y. A.; Shunina, I. G.; Sedov, I. V., Highly Porous Materials as Potential Components of Natural Gas Storage Systems: Part 1 (A Review). *Petroleum Chemistry* **2022,** *62* (6), 561-582.
5. Mishra, R.; Militky, J.; Venkataraman, M., 7 - Nanoporous materials. In *Nanotechnology in Textiles*, Mishra, R.; Militky, J., Eds. Woodhead Publishing: 2019; pp 311-353.
6. Rios, G.; Centi, G.; Kanellopoulos, N. K., *Nanoporous materials for energy and the environment / edited by Gilbert Rios, Gabriele Centi, Nick Kanellopoulos*. Pan Stanford Pub.: Singapore, 2012.
7. Zeng, H.; Qu, X.; Xu, D.; Luo, Y., Porous Adsorption Materials for Carbon Dioxide Capture in Industrial Flue Gas. *Frontiers in Chemistry* **2022,** *10*.
8. Morris, R. E.; Wheatley, P. S., Gas Storage in Nanoporous Materials. *Angewandte Chemie International Edition* **2008,** *47* (27), 4966-4981.
9. Zhu, T.; Han, Y.; Liu, S.; Yuan, B.; Liu, Y.; Ma, H., Porous Materials Confining Single Atoms for Catalysis. *Frontiers in Chemistry* **2021,** *9*.
10. Liu, Y.; Zhai, Y.; Xia, Y.; Li, W.; Zhao, D., Recent Progress of Porous Materials in Lithium-Metal Batteries. *Small Structures* **2021,** *2* (5), 2000118.
11. Yang, Q.; Liu, Q.; Ling, W.; Dai, H.; Chen, H.; Liu, J.; Qiu, Y.; Zhong, L., Porous Electrode Materials for Zn-Ion Batteries: From Fabrication and Electrochemical Application. *Batteries* **2022,** *8* (11), 223.
12. Suh, M. P.; Park, H. J.; Prasad, T. K.; Lim, D.-W., Hydrogen Storage in Metal–Organic Frameworks. *Chemical Reviews* **2012,** *112* (2), 782-835.
13. Ozdemir, J.; Mosleh, I.; Abolhassani, M.; Greenlee, L. F.; Beitle, R. R.; Beyzavi, M. H., Covalent Organic Frameworks for the Capture, Fixation, or Reduction of CO2. *Frontiers in Energy Research* **2019,** *7*.
14. Li, H.; Li, L.; Lin, R.-B.; Zhou, W.; Zhang, Z.; Xiang, S.; Chen, B., Porous metal-organic frameworks for gas storage and separation: Status and challenges. *EnergyChem* **2019,** *1* (1), 100006.
15. Kuhn, P.; Antonietti, M.; Thomas, A., Porous, Covalent Triazine-Based Frameworks Prepared by Ionothermal Synthesis. *Angewandte Chemie International Edition* **2008,** *47* (18), 3450-3453.
16. Lebedev, O. I.; Millange, F.; Serre, C.; Van Tendeloo, G.; Férey, G., First Direct Imaging of Giant Pores of the Metal−Organic Framework MIL-101. *Chemistry of Materials* **2005,** *17* (26), 6525-6527.
17. Momma, K.; Izumi, F., VESTA: a three-dimensional visualization system for electronic and structural analysis. *Journal of Applied Crystallography* **2008,** *41* (3), 653-658.

18. Farha, O. K.; Eryazici, I.; Jeong, N. C.; Hauser, B. G.; Wilmer, C. E.; Sarjeant, A. A.; Snurr, R. Q.; Nguyen, S. T.; Yazaydın, A. Ö.; Hupp, J. T., Metal–Organic Framework Materials with Ultrahigh Surface Areas: Is the Sky the Limit? *Journal of the American Chemical Society* **2012,** *134* (36), 15016-15021.
19. Galarneau, A.; Mehlhorn, D.; Guenneau, F.; Coasne, B.; Villemot, F.; Minoux, D.; Aquino, C.; Dath, J.-P., Specific Surface Area Determination for Microporous/Mesoporous Materials: The Case of Mesoporous FAU-Y Zeolites. *Langmuir* **2018,** *34* (47), 14134-14142.
20. Bae, Y.-S.; Yazaydın, A. Ö.; Snurr, R. Q., Evaluation of the BET Method for Determining Surface Areas of MOFs and Zeolites that Contain Ultra-Micropores. *Langmuir* **2010,** *26* (8), 5475-5483.
21. Papaefstathiou, G. S.; MacGillivray, L. R., Inverted metal–organic frameworks: solid-state hosts with modular functionality. *Coordination Chemistry Reviews* **2003,** *246* (1), 169-184.
22. Hönicke, I. M.; Senkovska, I.; Bon, V.; Baburin, I. A.; Bönisch, N.; Raschke, S.; Evans, J. D.; Kaskel, S., Balancing Mechanical Stability and Ultrahigh Porosity in Crystalline Framework Materials. *Angewandte Chemie International Edition* **2018,** *57* (42), 13780-13783.
23. Li, J.-R.; Kuppler, R. J.; Zhou, H.-C., Selective gas adsorption and separation in metal–organic frameworks. *Chemical Society Reviews* **2009,** *38* (5), 1477-1504.
24. Li, A.; Bueno-Perez, R.; Wiggin, S.; Fairen-Jimenez, D., Enabling efficient exploration of metal–organic frameworks in the Cambridge Structural Database. *CrystEngComm* **2020,** *22* (43), 7152-7161.
25. Ch. Baerlocher, D. B., Bernd Marler and L.B. McCusker, Database of Zeolite Structures.
26. Mroz, A. M.; Posligua, V.; Tarzia, A.; Wolpert, E. H.; Jelfs, K. E., Into the Unknown: How Computation Can Help Explore Uncharted Material Space. *Journal of the American Chemical Society* **2022,** *144* (41), 18730-18743.
27. Sanchez-Lengeling, B.; Aspuru-Guzik, A., Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018,** *361* (6400), 360-365.
28. Wang, J.; Wang, Y.; Chen, Y., Inverse Design of Materials by Machine Learning. *Materials* **2022,** *15* (5), 1811.
29. Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B., Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chemical Reviews* **2020,** *120* (16), 8066-8129.
30. Collins, S. P.; Daff, T. D.; Piotrkowski, S. S.; Woo, T. K., Materials design by evolutionary optimization of functional groups in metal-organic frameworks. *Science Advances* **2016,** *2* (11), e1600954.
31. Liu, Y.; Zhao, T.; Ju, W.; Shi, S., Materials discovery and design using machine learning. *Journal of Materiomics* **2017,** *3* (3), 159-177.
32. Mason, J. A.; Veenstra, M.; Long, J. R., Evaluating metal–organic frameworks for natural gas storage. *Chemical Science* **2014,** *5* (1), 32-51.
33. Pezzella, G.; Bhatt, P. M.; AlHaji, A.; Ramirez, A.; Grande, C. A.; Gascon, J.; Eddaoudi, M.; Sarathy, S. M., Onboard capture and storage system using metal-organic frameworks for reduced carbon dioxide emissions from vehicles. *Cell Reports Physical Science* **2023,** *4* (7), 101467.
34. Zhu, L.; Zhang, Y.-B., Crystallization of Covalent Organic Frameworks for Gas Storage Applications. *Molecules* **2017,** *22* (7), 1149.

35. Zhao, D.; Yuan, D.; Zhou, H.-C., The current status of hydrogen storage in metal–organic frameworks. *Energy & Environmental Science* **2008,** *1* (2), 222-235.
36. Kuppler, R. J.; Timmons, D. J.; Fang, Q.-R.; Li, J.-R.; Makal, T. A.; Young, M. D.; Yuan, D.; Zhao, D.; Zhuang, W.; Zhou, H.-C., Potential applications of metal-organic frameworks. *Coordination Chemistry Reviews* **2009,** *253* (23), 3042-3066.
37. Wieme, J.; Vandenbrande, S.; Lamaire, A.; Kapil, V.; Vanduyfhuys, L.; Van Speybroeck, V., Thermal Engineering of Metal–Organic Frameworks for Adsorption Applications: A Molecular Simulation Perspective. *ACS Applied Materials & Interfaces* **2019,** *11* (42), 38697-38707.
38. Makal, T. A.; Li, J.-R.; Lu, W.; Zhou, H.-C., Methane storage in advanced porous materials. *Chemical Society Reviews* **2012,** *41* (23), 7761-7779.
39. Freitas, S. K. S.; Borges, R. S.; Merlini, C.; Barra, G. M. O.; Esteves, P. M., Thermal Conductivity of Covalent Organic Frameworks as a Function of Their Pore Size. *The Journal of Physical Chemistry C* **2017,** *121* (48), 27247-27252.
40. Nomura, M.; Shiomi, J.; Shiga, T.; Anufriev, R., Thermal phonon engineering by tailored nanostructures. *Japanese Journal of Applied Physics* **2018,** *57* (8), 080101.
41. Islamov, M.; Babaei, H.; Anderson, R.; Sezginel, K. B.; Long, J. R.; McGaughey, A. J. H.; Gomez-Gualdron, D. A.; Wilmer, C. E., High-throughput screening of hypothetical metal-organic frameworks for thermal conductivity. *npj Computational Materials* **2023,** *9* (1), 11.
42. Wieser, S.; Kamencek, T.; Schmid, R.; Bedoya-Martínez, N.; Zojer, E., Exploring the Impact of the Linker Length on Heat Transport in Metal–Organic Frameworks. *Nanomaterials* **2022,** *12* (13), 2142.
43. Wieser, S.; Kamencek, T.; Dürholt, J. P.; Schmid, R.; Bedoya-Martínez, N.; Zojer, E., Identifying the Bottleneck for Heat Transport in Metal–Organic Frameworks. *Advanced Theory and Simulations* **2021,** *4* (1), 2000211.
44. Ahmed, A.; Seth, S.; Purewal, J.; Wong-Foy, A. G.; Veenstra, M.; Matzger, A. J.; Siegel, D. J., Exceptional hydrogen storage achieved by screening nearly half a million metal-organic frameworks. *Nature Communications* **2019,** *10* (1), 1568.
45. Agrawal, A.; Choudhary, A., Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Materials* **2016,** *4* (5).
46. Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C., Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials* **2017,** *3* (1), 54.
47. Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A., Machine learning for molecular and materials science. *Nature* **2018,** *559* (7715), 547-555.
48. Rodrigues, J. F.; Florea, L.; de Oliveira, M. C. F.; Diamond, D.; Oliveira, O. N., Big data and machine learning for materials science. *Discover Materials* **2021,** *1* (1), 12.
49. Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L., Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* **2019,** *5* (1), 83.
50. Halevy, A.; Norvig, P.; Pereira, F., The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems* **2009,** *24* (2), 8-12.
51. Lin, J.; Liu, Z.; Guo, Y.; Wang, S.; Tao, Z.; Xue, X.; Li, R.; Feng, S.; Wang, L.; Liu, J.; Gao, H.; Wang, G.; Su, Y., Machine learning accelerates the investigation of targeted MOFs: Performance prediction, rational design and intelligent synthesis. *Nano Today* **2023,** *49*, 101802.

52. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988,** *28* (1), 31-36.
53. Quinlan, J. R., Simplifying decision trees. *International Journal of Man-Machine Studies* **1987,** *27* (3), 221-234.
54. Quinlan, J. R., Induction of decision trees. *Machine Learning* **1986,** *1* (1), 81-106.
55. Breiman, L., Random Forests. *Machine Learning* **2001,** *45* (1), 5-32.
56. Chen, T.; Guestrin, C., XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery: San Francisco, California, USA, 2016; pp 785–794.
57. Liang, H.; Jiang, K.; Yan, T.-A.; Chen, G.-H., XGBoost: An Optimal Machine Learning Model with Just Structural Features to Discover MOF Adsorbents of Xe/Kr. *ACS Omega* **2021,** *6* (13), 9066-9076.
58. Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J. L.; Holm, E.; Ong, S. P.; Wolverton, C., Recent advances and applications of deep learning methods in materials science. *npj Computational Materials* **2022,** *8* (1), 59.
59. Agrawal, A.; Choudhary, A., Deep materials informatics: Applications of deep learning in materials science. *MRS Communications* **2019,** *9* (3), 779-792.
60. LeCun, Y.; Bengio, Y.; Hinton, G., Deep learning. *Nature* **2015,** *521* (7553), 436-444.
61. Wei, X.; Lu, Z.; Ai, Y.; Shen, L.; Wei, M.; Wang, X., Implementing and understanding the unsupervised transfer learning in metal organic framework toward methane adsorption from hypothetical to experimental data. *Separation and Purification Technology* **2024,** *330*, 125291.
62. Cao, Z.; Magar, R.; Wang, Y.; Barati Farimani, A., MOFormer: Self-Supervised Transformer Model for Metal–Organic Framework Property Prediction. *Journal of the American Chemical Society* **2023,** *145* (5), 2958-2967.
63. Kang, Y.; Park, H.; Smit, B.; Kim, J., A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. *Nature Machine Intelligence* **2023,** *5* (3), 309-318.
64. Cooper, G. M.; Colón, Y. J., Metal–organic framework clustering through the lens of transfer learning. *Molecular Systems Design & Engineering* **2023,** *8* (8), 1049-1059.
65. Wang, J.; Liu, J.; Wang, H.; Zhou, M.; Ke, G.; Zhang, L.; Wu, J.; Gao, Z.; Lu, D., A comprehensive transformer-based approach for high-accuracy gas adsorption predictions in metal-organic frameworks. *Nature Communications* **2024,** *15* (1), 1904.
66. Xie, T.; Grossman, J. C., Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters* **2018,** *120* (14), 145301.
67. Sanyal, S.; Balachandran, J.; Yadati, N.; Kumar, A.; Rajagopalan, P.; Sanyal, S.; Talukdar, P. P., MT-CGCNN: Integrating Crystal Graph Convolutional Neural Network with Multitask Learning for Material Property Prediction. *ArXiv* **2018,** *abs/1811.05660*.
68. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; Polosukhin, I., Attention is All you Need. Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Eds. 2017; Vol. 30.
69. Ray, P. P., ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* **2023,** *3*, 121-154.

70. OpenAI, Introducing ChatGPT. 2022.
71. Bergstra, J.; Bengio, Y., Random search for hyper-parameter optimization. *Journal of machine learning research* **2012,** *13* (2).
72. Hutter, F.; Hoos, H. H.; Leyton-Brown, K. In *Sequential Model-Based Optimization for General Algorithm Configuration*, Learning and Intelligent Optimization, Berlin, Heidelberg, 2011//; Coello, C. A. C., Ed. Springer Berlin Heidelberg: Berlin, Heidelberg, 2011; pp 507-523.
73. Cui, B.; Audu, C. O.; Liao, Y.; Nguyen, S. T.; Farha, O. K.; Hupp, J. T.; Grayson, M., Thermal Conductivity of ZIF-8 Thin-Film under Ambient Gas Pressure. *ACS Applied Materials & Interfaces* **2017,** *9* (34), 28139-28143.
74. Gunatilleke, W. D. C. B.; Wei, K.; Niu, Z.; Wojtas, L.; Nolas, G.; Ma, S., Thermal conductivity of a perovskite-type metal–organic framework crystal. *Dalton Transactions* **2017,** *46* (39), 13342-13344.
75. Huang, B. L.; Ni, Z.; Millward, A.; McGaughey, A. J. H.; Uher, C.; Kaviany, M.; Yaghi, O., Thermal conductivity of a metal-organic framework (MOF-5): Part II. Measurement. *International Journal of Heat and Mass Transfer* **2007,** *50* (3), 405-411.
76. Babaei, H.; DeCoster, M. E.; Jeong, M.; Hassan, Z. M.; Islamoglu, T.; Baumgart, H.; McGaughey, A. J. H.; Redel, E.; Farha, O. K.; Hopkins, P. E.; Malen, J. A.; Wilmer, C. E., Observation of reduced thermal conductivity in a metal-organic framework due to the presence of adsorbates. *Nature Communications* **2020,** *11* (1), 4010.
77. Huang, J.; Xia, X.; Hu, X.; Li, S.; Liu, K., A general method for measuring the thermal conductivity of MOF crystals. *International Journal of Heat and Mass Transfer* **2019,** *138*, 11-16.
78. Daniele Ongari, M. J. P., Aliaksandr V. Yakutovich, Leopold Talirz, Berend Smit, Building a consistent and reproducible database for adsorption evaluation in Covalent-Organic Frameworks. Materials Cloud Archive, 2023.
79. Mercado, R.; Fu, R.-S.; Yakutovich, A. V.; Talirz, L.; Haranczyk, M.; Smit, B., In Silico Design of 2D and 3D Covalent Organic Frameworks for Methane Storage Applications. *Chemistry of Materials* **2018,** *30* (15), 5069-5086.
80. Meiirbek Islamov, H. B., Ryther Anderson, Kutay B. Sezginel, Jeffrey R. Long, Alan J. H. McGaughey, Diego A. Gomez-Gualdron & Christopher E. Wilmer, thermal-transport-MOFs. GitHub, 2022.
81. Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M., Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials* **2012,** *149* (1), 134-141.
82. Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G., Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **2013,** *68*, 314-319.
83. Halder, P.; Prerna; Singh, J. K., Building Unit Extractor for Metal–Organic Frameworks. *Journal of Chemical Information and Modeling* **2021,** *61* (12), 5827-5840.
84. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E., Scikit-learn: Machine Learning in P ython. *Journal of Machine Learning Research* **2011,** *12*, 2825--2830.
85. Rocio Mercado, R.-S. F., Aliaksandr V. Yakutovich, Leopold Talirz, Maciej Haranczyk, Berend Smit, In Silico Design of 2D and 3D Covalent Organic Frameworks for Methane Storage Applications. 2018.0003/v3 ed.; Materials Cloud Archive, 2018.

86. RDKit: Open-source cheminformatics.
87. Broom, D. P.; Webb, C. J.; Hurst, K. E.; Parilla, P. A.; Gennett, T.; Brown, C. M.; Zacharia, R.; Tylianakis, E.; Klontzas, E.; Froudakis, G. E.; Steriotis, T. A.; Trikalitis, P. N.; Anton, D. L.; Hardy, B.; Tamburello, D.; Corgnale, C.; van Hassel, B. A.; Cossement, D.; Chahine, R.; Hirscher, M., Outlook and challenges for hydrogen storage in nanoporous materials. *Applied Physics A* **2016,** *122* (3), 151.
88. Pardakhti, M.; Moharreri, E.; Wanik, D.; Suib, S. L.; Srivastava, R., Machine Learning Using Combined Structural and Chemical Descriptors for Prediction of Methane Adsorption Performance of Metal Organic Frameworks (MOFs). *ACS Combinatorial Science* **2017,** *19* (10), 640-645.
89. Chang, C.-Y.; Hsu, M.-T.; Esposito, E. X.; Tseng, Y. J., Oversampling to Overcome Overfitting: Exploring the Relationship between Data Set Composition, Molecular Descriptors, and Predictive Modeling Methods. *Journal of Chemical Information and Modeling* **2013,** *53* (4), 958-971.
90. MathWorks Overfitting. https://www.mathworks.com/discovery/overfitting.html.
91. Bucior, B. J.; Rosen, A. S.; Haranczyk, M.; Yao, Z.; Ziebel, M. E.; Farha, O. K.; Hupp, J. T.; Siepmann, J. I.; Aspuru-Guzik, A.; Snurr, R. Q., Identification Schemes for Metal–Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis. *Crystal Growth & Design* **2019,** *19* (11), 6682-6697.
92. Bartlett, P. L.; Montanari, A.; Rakhlin, A., Deep learning: a statistical viewpoint. *Acta Numerica* **2021,** *30*, 87-201.
93. Wan, J.; Jiang, J.-W.; Park, H. S., Machine learning-based design of porous graphene with low thermal conductivity. *Carbon* **2020,** *157*, 262-269.
94. Cheng, R.; Li, W.; Wei, W.; Huang, J.; Li, S., Molecular Insights into the Correlation between Microstructure and Thermal Conductivity of Zeolitic Imidazolate Frameworks. *ACS Applied Materials & Interfaces* **2021,** *13* (12), 14141-14149.
95. Smith, D. S.; Alzina, A.; Bourret, J.; Nait-Ali, B.; Pennec, F.; Tessier-Doyen, N.; Otsu, K.; Matsubara, H.; Elser, P.; Gonzenbach, U. T., Thermal conductivity of porous materials. *International Journal of Materials Research* **2013,** *28* (17), 2260-2272.
96. ARPA-E Methane Opportunities for Vehicular Energy. https://arpa-e.energy.gov/technologies/programs/move (accessed 15/04/24).
97. Shelke, M. M. S.; Deshmukh, D. P. R.; Shandilya, V. K. In *A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique*, 2017.
98. Jeni, L. A.; Cohn, J. F.; Torre, F. D. L. In *Facing Imbalanced Data--Recommendations for the Use of Performance Metrics*, 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, 2-5 Sept. 2013; 2013; pp 245-251.
99. Anderson, R.; Gómez-Gualdrón, D. A., Large-Scale Free Energy Calculations on a Computational Metal–Organic Frameworks Database: Toward Synthetic Likelihood Predictions. *Chemistry of Materials* **2020,** *32* (19), 8106-8119.

8. Supporting Information

Table S1. *Results for random forest and XGBoost models trained on the MOF-TC dataset with different hyperparameter optimisation. The best performing models for random forest and XGBoost that are discussed in Section 5.1.1 are shown in bold.*

Model	Number of trees	Hyperparameter optimisation method	**MAE** / W m^{-1} K^{-1}	**RMSE** / W m^{-1} K^{-1}	R^2
Random Forest	**128**	**None**	**0.11**	**0.71**	**0.35**
Random Forest	185	Random search	0.14	0.79	0.19
Random Forest	185	Grid search	0.11	0.71	0.35
XGBoost	128	None	0.16	0.72	0.32
XGBoost	185	Random search	0.15	0.71	0.35
XGBoost	**185**	**Grid search**	**0.15**	**0.68**	**0.39**

Table S2. *Results for XGBoost model trained on the MOF-TC dataset with and without stratified sampling. Stratifed sampling yields a slight improvement in model predictions.*

Model	**MAE** / W m^{-1} K^{-1}	**RMSE** / W m^{-1} K^{-1}	R^2
XGBoost	0.15	0.68	0.39
XGBoost with ***stratified*** *data*	0.14	0.42	0.65

Table S3. *The CGCCN model results trained on the COF-GA dataset for high and low pressures, trained with random weights (from scratch) and from pretrained weights. All predictions generated by the CGCNN are less accurate than those generated by MOFormer when trained on the same data. Similar to MOFormer, for these CGCNN models pretraining does reduce prediction errors.*

Dataset	**CGCNN** Model	**MAE** / mol kg^{-1}	**RMSE** / mol kg^{-1}	R^2
High pressure (65 bar)	*Trained from* ***scratch***	5.05	6.95	0.56
High pressure (65 bar)	*Trained from* ***pretrained*** *model*	4.92	6.73	0.59
Low pressure (5.8 bar)	*Trained from* ***scratch***	0.46	0.67	0.27
Low pressure (5.8 bar)	*Trained from* ***pretrained*** *model*	0.45	0.66	0.30