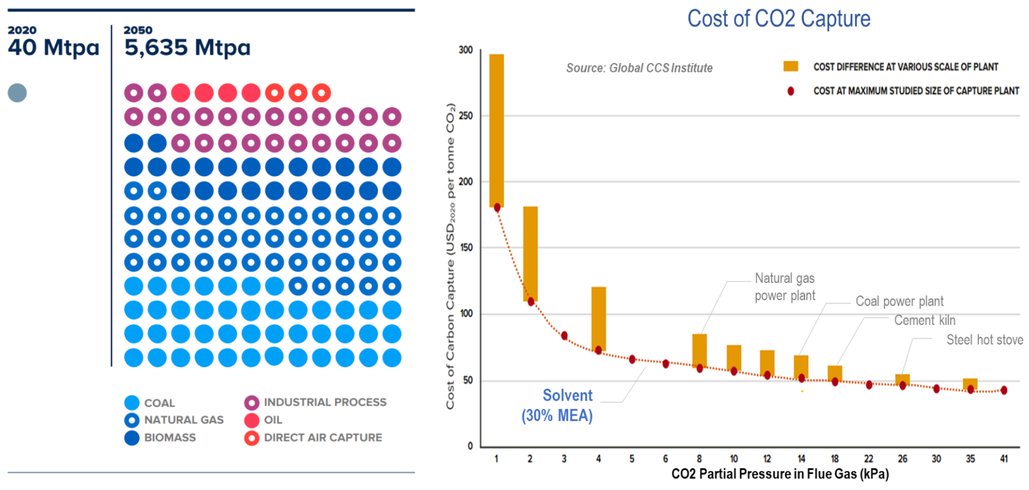
Graph Neural Network based Screening of Metal-Organic Frameworks for CO2 Capture

Zikri Bayraktar, Mengying Li, Shahnawaz Molla; *SLB*

ICLR International Conference On Learning Representations

Need for Carbon Dioxide Capture



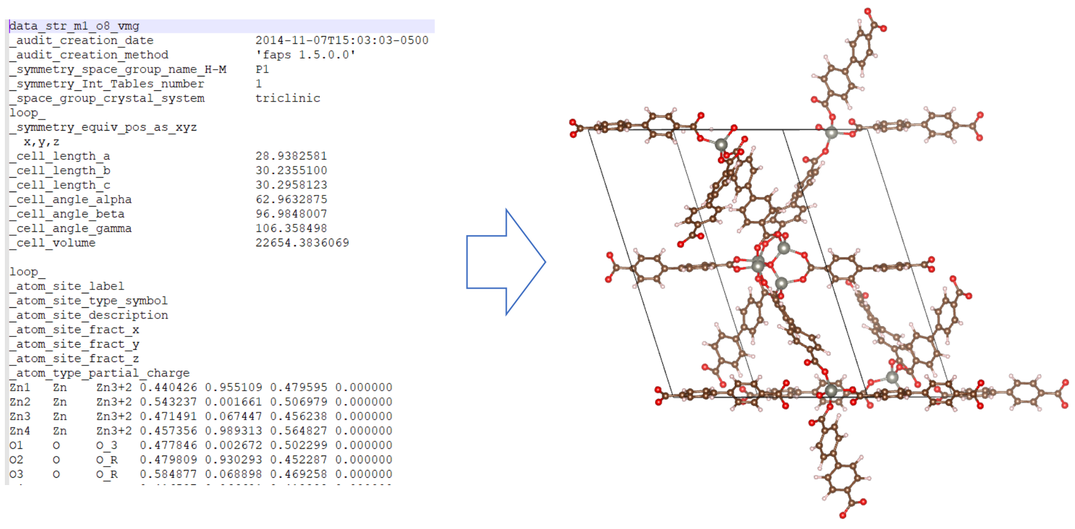
Carbon dioxide (CO_2) capture, utilization, and sequestration (CCUS) is expected to play an important role in the global campaign to limit global warming to 1.5 ◦C. However, to achieve this ambitious goal, we need to deploy CO_2 capture capabilities at gigaton capacity - approximately 150X increment by 2050.

Currently, CO_2 capture using chemical solvents is the most mature technology that has been proven to work at large-scale. However, these capture systems have large footprints and the energy demand to operate is very high. Studies conducted by U.S. National Energy Technology Laboratory suggest that implementing aqueous amine-based CO_2 capture can reduce the energy efficiency of a coal-fired power plant by 20-30%. The need for large thermal energy for solvent regeneration accounts for 20-25% of the operating cost even for advanced solvents.

Solvent-based capture systems are also expensive to build due to the large footprint of the packed bed towers used for CO_2 absorption and desorption. As a result, new separation technologies using either solid sorbents, membranes, or cryogenics are being developed to address the new material needs to improve the CO_2 capture capacity, selectivity over other gases, and to reduce the regeneration energy demand. There is urgency to develop energy-efficient and compact capture processes for widespread adoption of CC in the industry to reduce global warming.

Metal-Organic Frameworks for CO2 Capture

A variety of solid sorbent-based materials are being investigated for their CO_2 absorption capacity, CO_2 selectivity, stability, and regeneration capacity to efficiently remove CO_2 from point sources. Specifically, metal-organic frameworks (MOFs), a class of highly porous crystalline materials, have come across as promising candidates due to their exceedingly high CO_2 adsorption capacities that are inherently tunable through their unique structural features and more importantly, their scalability for industrial applications. MOFs are built through transition metal clusters and organic linkers into structural building units (SBUs) that are further extended into well-defined, uniformed structures with high surface area and structural stability. Those highly tunable SBUs in MOFs offer exceptional adaptability in isoreticular manipulation as well as their controllable functionalities in CO_2 adsorption through their theoretical, infinite combinations among metal ions, organic linkers, functional group, and optimization of pores spacing.

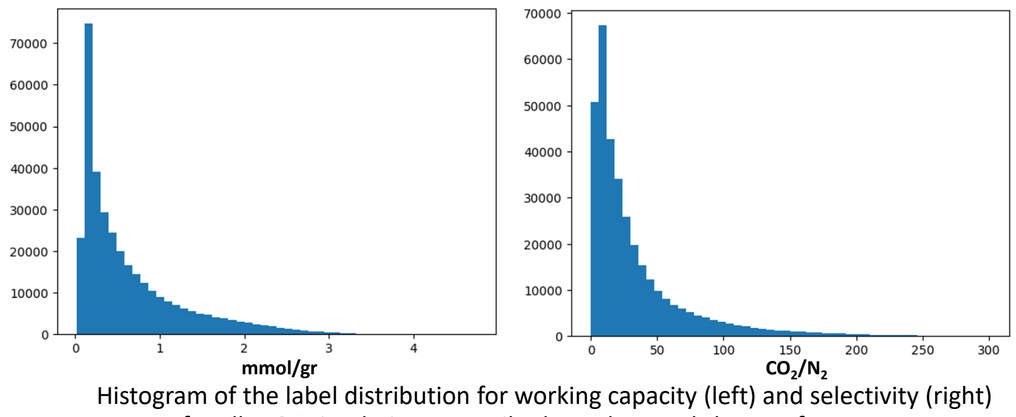


```
data_str_m1_o8_vmg
_audit_creation_date                2014-11-07T15:03:03-0500
_audit_creation_method              'faps 1.5.0.0'
_symmetry_space_group_name_H-M      P1
_symmetry_Int_Tables_number         1
_space_group_crystal_system         triclinic
loop_
_symmetry_equiv_pos_as_xyz
  x,y,z
_cell_length_a                      28.9382581
_cell_length_b                      30.2355100
_cell_length_c                      30.2958123
_cell_angle_alpha                   62.9632875
_cell_angle_beta                    96.9848007
_cell_angle_gamma                   106.358498
_cell_volume                        22654.3836069

loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_description
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_type_partial_charge
Zn1   Zn    Zn3+2  0.440426  0.955109  0.479595  0.000000
Zn2   Zn    Zn3+2  0.543237  0.001661  0.506979  0.000000
Zn3   Zn    Zn3+2  0.471491  0.067447  0.456238  0.000000
Zn4   Zn    Zn3+2  0.457356  0.989313  0.564827  0.000000
O1    O     O_3    0.477846  0.002672  0.502299  0.000000
O2    O     O_R    0.479809  0.930293  0.452287  0.000000
O3    O     O_R    0.584877  0.068898  0.469258  0.000000
```

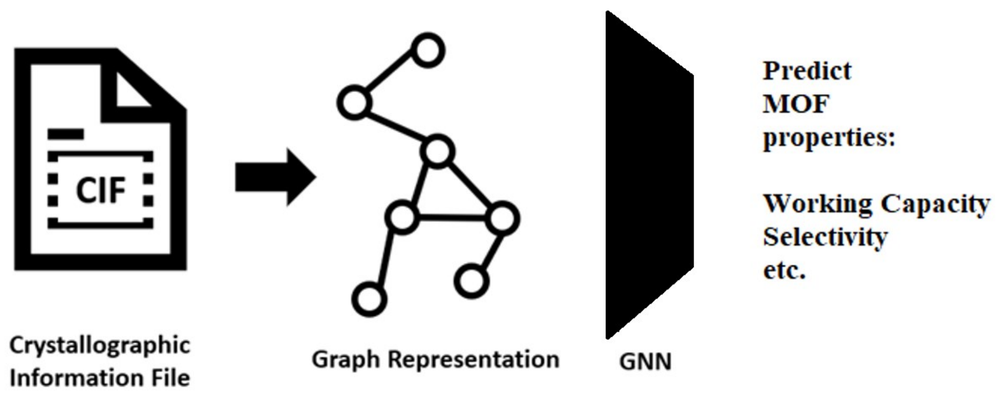
For post-combustion CO_2 capture processes, there is an urgent need to understand the role of impurities, water in particular, the variability in the concentrations of impurities in the feed stream and more importantly, their impact on the stability of the sorbents as well as the energy requirement for the regeneration processes. However, to-date, only a small number of MOFs have been studied experimentally to meet target-specific requirements, which warrants the necessity for a faster and reliable way to design and screen MOFs for post-combustion capture processes.

Past ML work on MOFs screening was driven by the subject matter expert's (SME) time-consuming analysis of the properties through feature engineering with geometric, chemical, and energy-based descriptors. Graph Neural Network, on the other hand, can be used to represent complex crystal structures of MOF using their crystallographic information, including but not limited to their unit cell values, space groups, identities, and position of the atoms, as well as properties of each bond.



Histogram of the label distribution for working capacity (left) and selectivity (right) for all MOF simulations. Heavily skewed towards low performance.

GNN-based Screening of MOFs

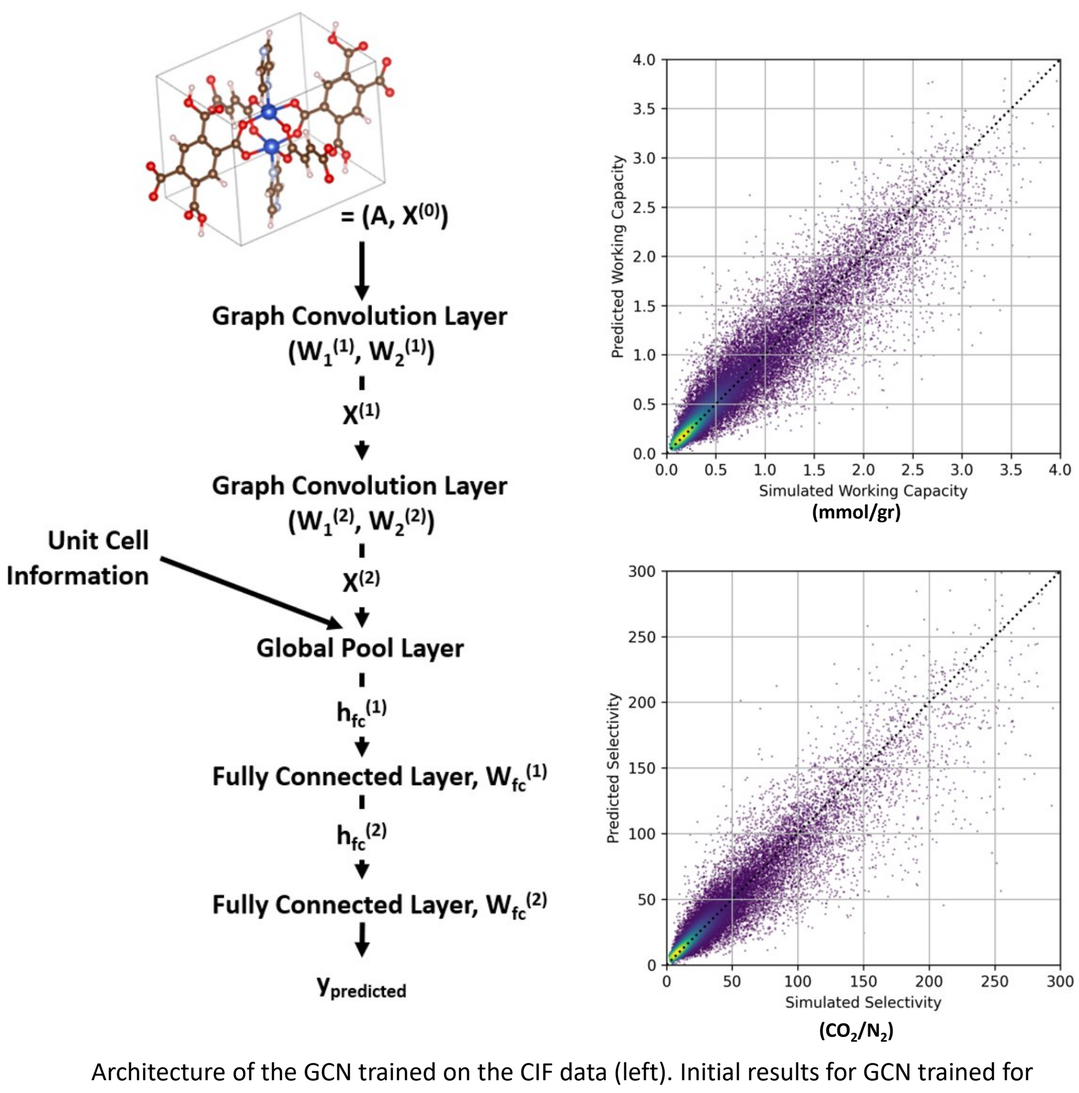


Various types of data in many aspects of our lives can be represented by graphs and there are solutions already implemented using Graph Neural Networks (GNN). Recommender systems in online shopping sites, traffic prediction in mobile phone applications, weather forecasting and our beloved social-networks, all use different types of GNNs in their core. GNNs are also utilized in science for physics simulations, molecule classification or molecule design, drug discovery, protein discovery and many other tasks.

Our proposed method utilizes the CIF file content and encode it in graph representation to train GNN models to predict material performance as illustrated in Figure above. Material performances are computed through Grand Canonical Monte Carlo (GCMC) method. Unlike simple molecules, a single MOF crystal can make up to a thousand of bonds, which brings a challenge for any ML method to handle such nonuniform dataset. In our preliminary results, we have demonstrated GCNs coded in PyG can easily handle the training of such complex data structures as undirected graphs in a regression problem of predicting MOF properties such as the CO_2 working capacity or selectivity. As expected from the problem definition, i.e., screening large number of potential candidates, most of the MOF crystals are low performing. This can be seen by the label distribution for working capacity and selectivity in Figure at the bottom of previous column. This creates a challenge as the predicted data is skewed for this regression task.

Preliminary Results

In this work, we implemented and trained end-to-end GNN models directly from the existing CIF content and used the corresponding GCMC data to predict their CO_2 working capacity or CO_2/N_2 selectivity as depicted in the illustration above. We utilized the CIF files for 340,000 MOFs where over 250,000 of the files were missing their bond distances. Using the unit cell dimensions and atom site fractional coordinates, we calculated the bond distances to use them as one of the node features. Along with the bond distances, one-hot encoded bond types constitute the edge attributes for each MOF. Adjacency matrix which represents which atoms are connected through bonds (i.e., graph edges) was encoded as edge lists. Node attributes include the atomic weight of each atom and one-hot encoded atom types. Crystal unit cell dimensions and angles are also utilized along with the one-hot encoded crystal type information as shown on the left in Figure 3. Exploiting the information provided in each CIF files, one can eliminate the need for hand-picked geometric or chemical descriptors and reduce the SME's workload. Initially, we focused on training graph convolution network (GCN) models and achieved R^2 score ranging 0.87 to 0.89, easily. A sample test result is shown on the right side of Figure below. In comparison to our preliminary results, prior ML work using only SME-designed geometric descriptors alone reached R^2 score up to 0.75, whereas custom designed atomic property-weighted radial distribution function (AP-RDF) features can push the ML model R^2 score to a range of 0.83 and 0.94, respectively. Similarly, it was shown that CNN based end-to-end screening can achieve R^2 score up to 0.91.



Architecture of the GCN trained on the CIF data (left). Initial results for GCN trained for working capacity (top right) and selectivity (bottom right).