Discovery of Novel Reticular Materials for Carbon Dioxide Capture using GFlowNets

Flaviu Cipcigan
IBM Research - Europe
flaviu.cipcigan@ibm.com

Jonathan Booth
Science and Technology Facilities Council
jonathan.booth@stfc.ac.uk

Rodrigo Neumann Barros Ferreira
IBM Research
rneumann@br.ibm.com

Carine Ribeiro dos Santos
IBM Research
carineribeiro@ibm.com

Mathias Steiner
IBM Research
mathiast@br.ibm.com

Abstract

Artificial intelligence holds promise to improve materials discovery. GFlowNets are an emerging deep learning algorithm with many applications in AI-assisted discovery. Using GFlowNets, we generate porous reticular materials, such as metal organic frameworks and covalent organic frameworks, for applications in carbon dioxide capture. We introduce a new Python package (`matgfn`) to train and sample GFlowNets. We use `matgfn` to generate the `matgfn-rm` dataset of novel and diverse reticular materials with gravimetric surface area above 5000 m^2/g. We calculate single- and two-component gas adsorption isotherms for the top-100 candidates in `matgfn-rm`. These candidates are novel compared to the state-of-art ARC-MOF dataset and rank in the 90th percentile in terms of working capacity compared to the CoRE2019 dataset. We discover 15 hypothetical materials outperforming all materials in CoRE2019.

1 Introduction

Artificial intelligence (AI) holds promise to improve the scientific method [17, 1] and to accelerate scientific discovery. Applied to materials [1], AI unlocks vast search spaces and enables novel applications in pharmaceuticals [11, 10, 16, 9], batteries or carbon capture [24].

Reticular materials [34] such as Metal-Organic Frameworks (MOFs) and Covalent Organic Frameworks (COFs) are extended periodic structures connected via strong bonds [15]. They are synthesized by connecting building blocks known as secondary building units to form three-dimensional periodic structures [20]. By choosing the building blocks, the properties of a reticular material can be tuned to support many applications [34].

Reticular materials with high gravimetric surface area are particularly useful for applications in carbon capture, since carbon dioxide molecules adsorb at the internal surface area [14]. The larger the gravimetric surface area, the more gas molecules can be adsorbed per gram of material.

In this work, we use GFlowNets to generate reticular materials with high gravimetric surface area for applications in carbon capture. Our key contributions are:

1. The `matgfn` Phython library for training and sampling using GFlowNets.
2. A workflow using `matgfn` to generate reticular materials using secondary building units.

[1] Here, we conceptualise materials broadly to include molecules, proteins, crystals and complex materials.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

3. The `matgfn-rm` dataset of diverse and novel reticular materials with total internal surface area higher than 5000 m^2 g^{-1}. The top-100 reticular materials candidates are novel compared to the reference ARC-MOF dataset, rank in the 90th percentile in terms of simulated working capacity compared to the CoRE2019 dataset. We discover 15 hypothetical materials outperforming all materials in CoRE2019.

The code and dataset will be available at `http://github.com/flaviucipcigan/matgfn` and archived on Zenodo [8].

2 Background and related work

Generative Flow Networks GFlowNets [3, 4] are an emerging machine learning algorithm with many applications in AI-assisted materials discovery [19]. GFlowNets learn to generate composite objects $\underline{x}$ by sampling from an unnormalised distribution $p(\underline{x}) \propto R(\underline{x})$ where $R(\underline{x})$ is a user-specified positive reward function. A composite object $\underline{x}$ consists of symbols drawn from a vocabulary $\mathbb{V}$ and relationships between those symbols. For example, $\underline{x}$ can be a sequence $\underline{x} = [x_1, x_2, \ldots x_n]$ or a graph. The object $\underline{x}$ is built by through Markov Decision Process restricted to a directed acyclic graph. Transition probabilities $p(x_{i+1} \mid \underline{x})$ are approximated by a neural network called a flow model. GFlowNets need fewer evaluations of the reward function to generate samples with high reward, novelty and diversity when compared to alternatives such as Markov Chain Monte Carlo, Proximal Policy Optimisation or Bayesian Optimisation [3].

Building hypothetical reticular frameworks Trillions of hypothetical frameworks such as MOFs or COFs can be generated by placing secondary building units [20] into nodes and edges of a three dimensional topology [27]. A secondary building unit is an organic molecule or a coordination compound (a metal linked to organic atoms). A topology is a three dimensional arrangement of nodes and edges. Replacing nodes and edges with secondary building units results in a three dimensional point cloud of atoms connected by covalent or metal-organic bonds. We use the `pormake` secondary building units [22] and topology codes from the Reticular Chemistry Structure Resource [27]. Previously, deep autoencoders [35] and evolutionary methods [22] have been used to generate frameworks using this approach.

Reference datasets We use two reference datasets in this work. These datasets are not used for training models, but as comparison once training is done, as GFlowNet generates candidates using just a reward function. The CoRE2019 dataset [7] consists of 12,023 metal-organic frameworks with carbon dioxide uptake properties calculated by Moosavi *et al.* [25] using Grand Canonical Monte Carlo. ARC-MOF (reported in 2022) [6] is a collection of 279,610 MOFs from previous MOF datasets. It contains both experimental and hypothetical MOFs.

3 Generating reticular frameworks with GFlowNets

Python package We built a Python library called `matgfn` to train and sample GFlowNets. The library is built on top of PyTorch [29] and Gymnasium [31] and prioritises ease of use and code readability. We intend for `matgfn` to be a general Python package for generation of diverse types of materials from small molecules to framework materials. Architecturally, `matgfn` separates sampling, loss calculation, optimisation, and environment definition as modular Python classes. Each can be modified individually, to implement off-policy training or use improved losses, for example. We note similar architectural choices for `torchgfn`[21].

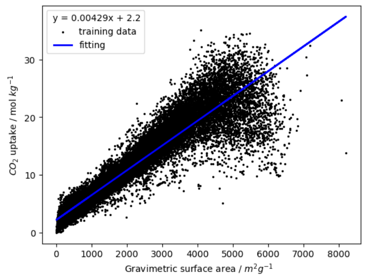

Figure 1: Regression of simulated high pressure CO_2 uptake to gravimetric surface area

Environment for reticular framework generation We configure a GFlowNet environment to build string sequences made out of text tokens. Those text tokens start with either an `N`, representing a node building block, or an `E`, representing an edge building block. For example, one of the potential generated sequences is ["N577", "N238", "N194", "E5", "E3", "E74"]. We use building blocks in the pormake database. The string sequences are transformed to a Crystallographic Information File (*.cif) by pormake to create a reticular framework. Not all strings create valid materials. Thus, during generation, building blocks were restricted such that (a) each topology had the correct number of nodes and edges, (b) the building blocks were placed in the correct order and (c) each slot had a compatible building block.

Reward We calculate the Gravimetric Surface Area GSA in m^2/g with Zeo++ [33] during the training loop of the GFlowNet. We configure Zeo++ with a probe radius of 1.525 Å and 2000 samples. The GFlowNet is given the following reward:

$$R(\underline{x}) = \mathcal{H}\left(GSA(\underline{x}) - C\right) * \exp\left(\frac{GSA(\underline{x}) - C}{C}\right) \tag{1}$$

where $\mathcal{H}(x)$ is the Heaviside step function $\mathcal{H}(x) = 0$ if $x < 0, 1$ if $x \geq 0$ and C is a cutoff. Zeo++ and pormake sometimes raise errors due to large distances between atoms. The reward is zero when an error occurred to encourage the GFlowNet to avoid materials with unrealistic bond lengths.

Relationship with CO_2 capture We demonstrate that the gravimetric surface area predicts CO_2 uptake by analysing approximately 30,000 MOFs from three databases: CoRE2019 [7], ARABG [2] and BW20K [5]. We performed univariate linear regression of CO_2 uptake at 16 bar using each of the geometric and chemical descriptors. The best performing descriptor was the gravimetric surface area with coefficient of determination is 0.88, RMSE is 2.41 mol kg^{-1} and Spearman's rank correlation coefficient of 0.97. Figure 1 shows the CO_2 uptake as a function of gravimetric surface area. We validated the regression using 50 rounds of 10-fold cross validation, with each cross-validation consisting of an 80-20 split between training and test data. The mean coefficient of determination is 0.88 ± 0.0002 and mean RMSE is 2.41 ± 0.022 mol/kg. The training and test values of coefficient of determination and RMSE are the same to two decimal places and the standard deviation of these metrics during cross validation are very small which shows that the correlation is robust and stable.

4 The matgfn-rm dataset

Training We trained a GFlowNet using Trajectory Balance loss [23] and an LSTM flow model. We use a learning rate of 5×10^{-3} for both the flow model and the partition function. We train for a maximum of 100,000 episodes and stop when the mean loss over 10,000 episodes is lower than 1.8. Eleven topologies were chosen: CDZ-E, CLD-E, EFT, FFC, TSG, TFF, ASC, DMG, DNQ, FSO, URJ. For each topology, two GFlowNets were trained, one with edges and one without. The performance is shown in Supplementary Information. Once the GFlowNets have been trained, they were sampled to generate matgfn-rm dataset of over 1 million hypothetical reticular frameworks.

Diversity analysis We compare the top-100 and top-100,000 candidates from matgfn-rm to the ARC-MOF dataset. For each CIF file, we compute the average minimum distance (AMD) descriptor [32] of length 100. This descriptor uniquely identifies crystal structures and is a continuous metric, meaning that the distance (measured using the Chebyshev metric) is zero for similar crystals. For visualisation, we perform dimensionality reduction to two dimensions using t-SNE implemented. We chose the implementation in openTSNE [30], use the Chebyshev distance metric and calculate nearest neighbours using nearest neighbour descent [13]. Figure 2 shows the result. The matgfn-rm materials are separated from most materials from ARC-MOF. Thus, the generated materials are novel compared to existing datasets. For a clearer virew of some of the overlapping regions, check Figure 18.

Simulated CO_2 capture performance In order to confirm the expectation of efficient CO_2 capture from an adsorption proxy (*i.e.*, the gravimetric surface area), we run Physics-based Grand Canonical Monte Carlo simulations for the top-100 generated materials in the matgfn-rm dataset [26, 28]. We simulated single-component adsorption isotherms for pure CO_2, from which we extract the CO_2 working capacity, and dual-component adsorption isotherms for dry flue gas (15% CO_2 and 85% N_2), from which we extract the CO_2/N_2 selectivity. All simulations were performed at 300 K, with pressures ranging from 0.15 to 16 bar. The working capacity was calculated as the difference in uptake of (single-component) CO_2 between 16 and 0.15 bar, while the selectivity was calculated as $S = {}^{(Q_{CO_2}/Q_{N_2})}\!/_{(f_{CO_2}/f_{N_2})}$, where Q_i is the uptake of species i at 0.15 bar and f_i is the concentration of species i in the input flue gas stream. Figure 3 shows the distribution of

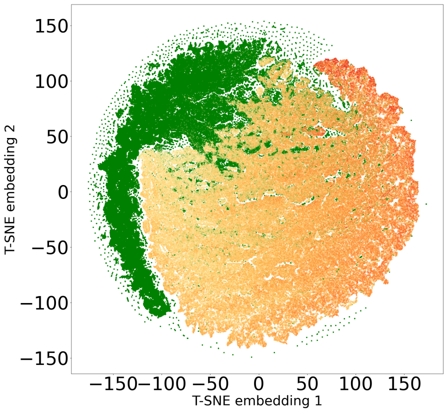

Figure 2: Two dimensional t-SNE embedding of the average minimum distance of ARC-MOF (green), and matgfn-rm (yellow to orange) materials. The colours in the matgfn-rm dataset are proportional to the reward, with light yellow signifying low reward and dark orange high reward.

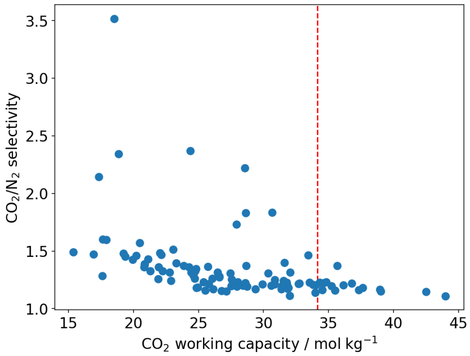

Figure 3: Simulated CO_2 working capacity and CO_2 / N_2 selectivity for the top-100 matgfn-rm materials. The (red) dashed line represents the highest working capacity found in the CoRE2019 dataset, which is surpassed by 15 of the top-100 matgfn-rm materials.

absolute (working capacity) and relative (selectivity) capture metrics for the top-100 matgfn-rm materials. All top-100 materials are (modestly) more selective towards CO_2 than N_2 and exhibit very high CO_2 working capacities, corresponding to the 90th percentile of the experimentally-realised CoRE2019 dataset [25]. Fifteen of the top-100 matgfn-rm materials have working capacities that are higher than all materials found in the CoRE2019 dataset. In particular, we highlight in Figure 4 the covalent organic framework 005-ffc-10217 that achieved the highest CO_2 working capacity of approximately 44 mol/kg.

Relaxation and validity check Due to the hypothetical nature of the generated MOFs, the crystalline structures are not guaranteed to be perfect. We therefore used the mofchecker library [18] to perform basic consistency checks on the generated CIFs. According to mofchecker, all of the top-100 matgfn-rm are porous (metal-) organic materials. However, due to the hypothetical interatomic distances sometimes being larger (or shorter) than the typical bond lengths, some atoms are flagged as either over- or under-coordinated. In order to obtain a more realistic structure, we performed atomic coordinate and unit cell relaxation using the CHGNet [12] interatomic potential. Relaxing the structures solves most of the structural problems, with 98% presenting neither atomic overlaps nor over-coordination of C, N and H atoms, respectively. In particular, for the high-performing 005-ffc-10217 structure, relaxation led to a 23% reduction in the unit cell volume, bringing the CO_2 working capacity down to 37.5 mol/kg, which is still larger than those found in the CoRE2019 dataset. The relaxed pore size of 005-ffc-10217 is approximately 87 Å. Structural relaxation changes the average minimum

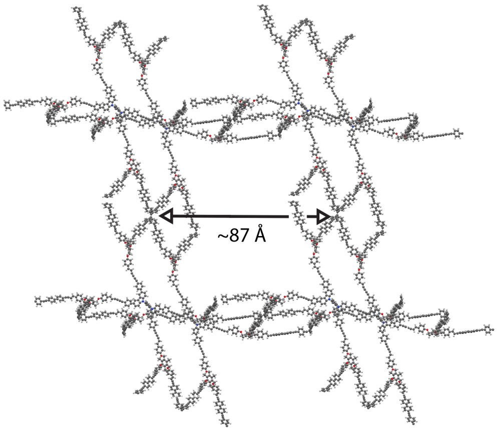

Figure 4: A render of the relaxed structure of 005-ffc-10217, the highest performing structure in the matgfn-rm dataset. We show here the $2 \times 2 \times 2$ supercell.

distance descriptors by a small amount and thus the diversity analysis still holds. See Figure 18 for an illustration of the effect of relaxation on the t-SNE embeddings.

5 Conclusion

In summary, we built a workflow using GFlowNets to generate diverse and novel reticular frameworks with gravimetric surface area greater than 5000 m^2/g. As a key result, the top-100 candidates of the resulting `matgfn-rm` dataset have working capacities in the top 90^{th} percentile of CoRE2019 reference dataset. Moreover, 15 of the top-100 `matgfn-rm` materials have working capacities that are higher than all materials found in the CoRE2019 dataset. Further tests are underway to confirm the stability and synthesizability of the materials generated in our study. Nevertheless, our results clearly demonstrate the potential of GFlowNets for materials discovery in carbon capture applications.

6 Contributions

Flaviu Cipcigan: **Project**: Conceptualisation, Project Administration. **Paper**: Writing - coordination, Writing – original draft, Writing – review & editing. **Software**: Main author of `matgfn`, contributor to MOF application code, **Results**: Methodology, Experiments, Formal Analysis, Validation, diversity analysis of `matgfn-rm`

Jonathan Booth: **Project**: Project Administration, trained GFlowNets on MOF topologies **Paper**: Writing – MOF result section, Writing – review & editing. **Software**: created reward function and other necessary code for MOF application of GFlownets **Results**: Methodology, Experiments, Formal Analysis, Validation

Rodrigo Neumann Barros Ferreira: **Paper**: Writing – original draft, Writing – review & editing. **Results**: Validation and simulations of `matgfn-rm` dataset.

Carine Ribeiro dos Santos: **Results**: Validation and simulations of `matgfn-rm` dataset.

Mathias Steiner: **Project**: Project Administration. **Paper**: Writing – original draft, Writing – review & editing.

7 Acknowledgements

We thank Dan Cunnington and Won Kyung Lee for valuable comments on the manuscript. Flaviu Cipcigan and Jonathan Booth were supported by the Hartree National Centre for Digital Innovation, a collaboration between STFC and IBM.

References

[1] Ankit Agrawal and Alok N. Choudhary. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Materials*, 4:053208, 2016.

[2] R. Anderson, E. Argueta, A. Biong, and D. Gomez-Gualdron. Role of pore chemistry and topology in the CO2 capture capabilities of MOFs: from molecular simulation to machine learning. *Chem. Mater.*, 2018.

[3] Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation. *CoRR*, abs/2106.04399, 2021.

[4] Yoshua Bengio, Tristan Deleu, Edward J. Hu, Salem Lahlou, Mo Tiwari, and Emmanuel Bengio. GFlowNet Foundations. *ArXiv*, abs/2111.09266, 2021.

[5] Peter G. Boyd, Arunraj Chidambaram, Enrique García-Díez, Christopher P. Ireland, Thomas D. Daff, Richard Bounds, Andrzej Gładysiak, Pascal Schouwink, Seyed Mohamad Moosavi, M. Mercedes Maroto-Valer, Jeffrey A. Reimer, Jorge A. R. Navarro, Tom K. Woo, Susana Garcia, Kyriakos C. Stylianou, and Berend Smit. Data-driven design of metal–organic frameworks for wet flue gas CO_2 capture. *Nature*, 576(7786):253–256, Dec 2019.

[6] Jake Burner, Jun Luo, Andrew White, Adam Mirmiran, Ohmin Kwon, Peter G. Boyd, Stephen Maley, Marco Gibaldi, Scott Simrod, Victoria Ogden, and Tom K. Woo. ARC–MOF: A diverse database of metal-organic frameworks with DFT-derived partial atomic charges and descriptors for machine learning. *Chemistry of Materials*, 35(3):900–916, January 2023.

[7] Yongchul G. Chung, Emmanuel Haldoupis, Benjamin J. Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D. Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S. Camp, Ben Slater, J. Ilja Siepmann, David S. Sholl, and Randall Q. Snurr. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: CoRE MOF 2019. *Journal of Chemical & Engineering Data*, 64(12):5985–5998, November 2019.

[8] Flaviu Cipcigan. Zenodo archive for flaviucipcigan/matgfn, December 2023. https://doi.org/10.5281/zenodo.10246465.

[9] Flaviu Cipcigan, Paul Smith, Jason Crain, Anders Hogner, Leonardo De Maria, Antonio Llinas, and Ekaterina Ratkova. Membrane permeability in cyclic peptides is modulated by core conformations. *Journal of Chemical Information and Modeling*, 61(1):263–269, December 2020.

[10] Daniel Crusius, Jason R. Schnell, Flaviu Cipcigan, and Philip C. Biggin. MacroConf – dataset &amp workflows to assess cyclic peptide solution structures. *Digital Discovery*, 2(4):1163–1177, 2023.

[11] Payel Das, Tom Sercu, Kahini Wadhawan, Inkit Padhi, Sebastian Gehrmann, Flaviu Cipcigan, Vijil Chenthamarakshan, Hendrik Strobelt, Cicero dos Santos, Pin-Yu Chen, Yi Yan Yang, Jeremy P. K. Tan, James Hedrick, Jason Crain, and Aleksandra Mojsilovic. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering*, 5(6):613–623, March 2021.

[12] Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, pages 1–11, 2023.

[13] Wei Dong, Charikar Moses, and Kai Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 577–586, New York, NY, USA, 2011. Association for Computing Machinery.

[14] Omar K. Farha, Ibrahim Eryazici, Nak Cheon Jeong, Brad G. Hauser, Christopher E. Wilmer, Amy A. Sarjeant, Randall Q. Snurr, SonBinh T. Nguyen, A. Özgür Yazaydın, and Joseph T. Hupp. Metal–organic framework materials with ultrahigh surface areas: Is the sky the limit? *Journal of the American Chemical Society*, 134(36):15016–15021, August 2012.

[15] Ralph Freund, Stefano Canossa, Seth M. Cohen, Wei Yan, Hexiang Deng, Vincent Guillerm, Mohamed Eddaoudi, David G. Madden, David Fairen-jimenez, Hao Lyu, Lauren K. Macreadie, Zhe Ji, Yuanyuan Zhang, Bo Wang, Frederik Haase, Christof Wöll, Orysia Zaremba, Jacopo Andreo, Stefan Wuttke, and Christian S. Diercks. 25 years of Reticular Chemistry. *Angewandte Chemie*, 2021.

[16] Katharine Hammond, Flaviu Cipcigan, Kareem Al Nahas, Valeria Losasso, Helen Lewis, Jehangir Cama, Fausto Martelli, Patrick W. Simcock, Marcus Fletcher, Jascindra Ravi, Phillip J. Stansfeld, Stefano Pagliara, Bart W. Hoogenboom, Ulrich F. Keyser, Mark S. P. Sansom, Jason Crain, and Maxim G. Ryadnov. Switching cytolytic nanopores into antimicrobial fractal ruptures by a single side chain mutation. *ACS Nano*, 15(6):9679–9689, April 2021.

[17] Tony Hey, Stewart Tansley, and Kristin Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[18] Kevin Maik. Jablonka. Mofchecker 1.0.0. *https://github.com/kjappelbaum/mofchecker*, 2023.

[19] Moksh Jain, Tristan Deleu, Jason S. Hartford, Cheng-Hao Liu, Alex Hernández-García, and Yoshua Bengio. GFlowNets for AI-Driven Scientific Discovery. *ArXiv*, abs/2302.00615, 2023.

[20] Markus J. Kalmutzki, Nikita Hanikel, and Omar M. Yaghi. Secondary building units as the turning point in the development of the reticular chemistry of MOFs. *Science Advances*, 4(10), October 2018.

[21] Salem Lahlou, Joseph D. Viviano, Victor Schmidt, and Yoshua Bengio. `torchgfn`: A PyTorch GFlowNet library, 2023.

[22] Sangwon Lee, Baekjun Kim, Hyun Cho, Hooseung Lee, Sarah Yunmi Lee, Eun Seon Cho, and Jihan Kim. Computational screening of trillions of metal–organic frameworks for high-performance methane storage. *ACS Applied Materials & Interfaces*, 13(20):23647–23654, May 2021.

[23] Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in GFlownets. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[24] James L McDonagh, Benjamin H Wunsch, Stamatia Zavitsanou, Alexander Harrison, Bruce Elmegreen, Stacey Gifford, Theodore van Kessel, and Flaviu Cipcigan. Machine guided discovery of novel carbon capture solvents. *arXiv preprint arXiv:2303.14223*, 2023.

[25] Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G. Boyd, Yongjin Lee, Berend Smit, and Heather J. Kulik. Understanding the diversity of the metal-organic framework ecosystem. *Nature Communications*, 11(1):4068, Aug 2020.

[26] Rodrigo Neumann Barros Ferreira, Breanndan O Conchuir, Tonia Elengikal, Binquan Luan, Ricardo Luis Ohta, Felipe Lopes Oliveira, Ashish Mhadeshwar, Jayashree Kalyanaraman, Anantha Sundaram, Joseph Falkowski, et al. Cloud-Based, High-Throughput, End-To-End Computational Screening of Solid Sorbent Materials for Carbon Capture. In *Proceedings of the 16th Greenhouse Gas Control Technologies Conference*, 2022.

[27] Michael O'Keeffe, Maxim A. Peskov, Stuart J. Ramsden, and Omar M. Yaghi. The reticular chemistry structure resource (RCSR) database of, and symbols for, crystal nets. *Accounts of Chemical Research*, 41(12):1782–1789, October 2008.

[28] Felipe Lopes Oliveira, Conor Cleeton, Rodrigo Neumann Barros Ferreira, Binquan Luan, Amir H Farmahini, Lev Sarkisov, and Mathias Steiner. CRAFTED: An exploratory database of simulated adsorption isotherms of metal-organic frameworks. *Scientific Data*, 10(1):230, 2023.

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019.

[30] Pavlin G. Poličar, Martin Stražar, and Blaž Zupan. openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. *bioRxiv*, 2019.

[31] Mark Towers, Jordan K. Terry, Ariel Kwiatkowski, John U. Balis, Gianluca de Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Arjun KG, Markus Krimmel, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Andrew Tan Jin Shen, and Omar G. Younis. Gymnasium, March 2023.

[32] Daniel Widdowson, Marco M. Mosca, Angeles Pulido, Andrew I. Cooper, and Vitaliy Kurlin. Average minimum distances of periodic point sets – foundational invariants for mapping periodic crystals. *MATCH Communications in Mathematical and in Computer Chemistry*, 87(3):529–559, December 2021.

[33] Thomas F. Willems, Chris H. Rycroft, Michaeel Kazi, Juan C. Meza, and Maciej Haranczyk. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials*, 149(1):134–141, 2012.

[34] Omar M Yaghi, Markus J Kalmutzki, and Christian S Diercks. *Introduction to reticular chemistry: metal-organic frameworks and covalent organic frameworks*. John Wiley & Sons, 2019.

[35] Zhenpeng Yao, Benjamín Sánchez-Lengeling, N. Scott Bobbitt, Benjamin J. Bucior, Sai Govind Hari Kumar, Sean P. Collins, Thomas Burns, Tom K. Woo, Omar K. Farha, Randall Q. Snurr, and Alán Aspuru-Guzik. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence*, 3(1):76–86, January 2021.

Appendix A: Training details of GFlowNets on all MOF topologies

Figure 5 shows the trajectory balance losses for training a GFlowNet on the ASC topology without edges while figure 6 shows the logZ. All other training runs on other topologies showed similar behaviour.

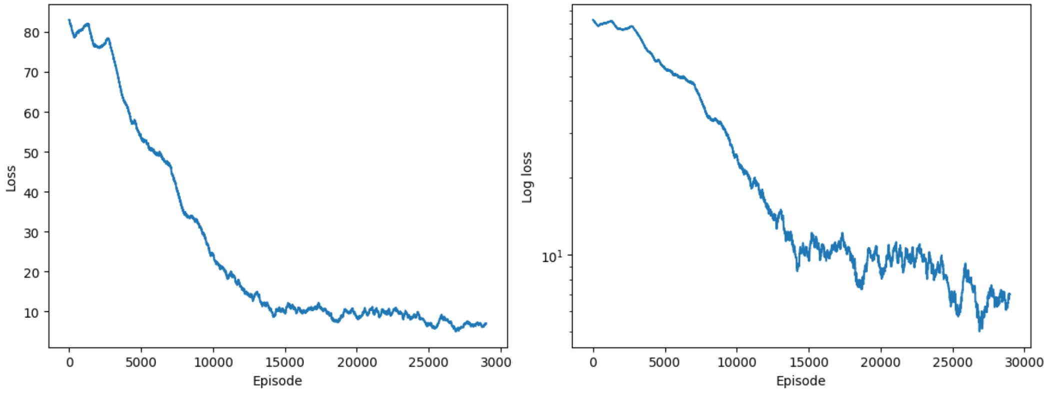

Figure 5: trajectory balance losses for training a GFlowNet on the ASC topology without edges. Losses are smoothed with a 1,000 episode window moving average due to the discovery of a high performing MOF causing a one-episode long spike in the loss.

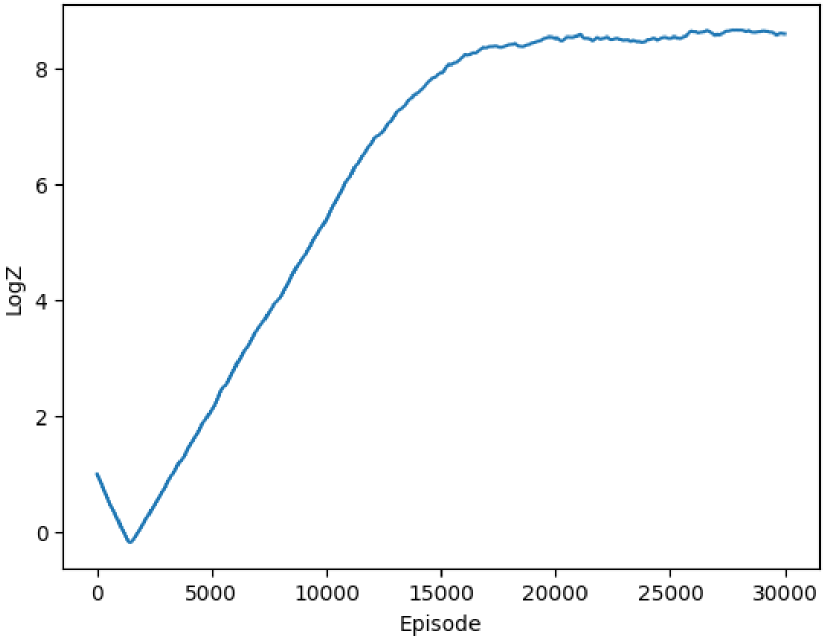

Figure 6: logZ during training for the ASC topology without edges.

The figures below show the performance of the GFlowNet vs random sampling for all eleven topologies with and without edges.

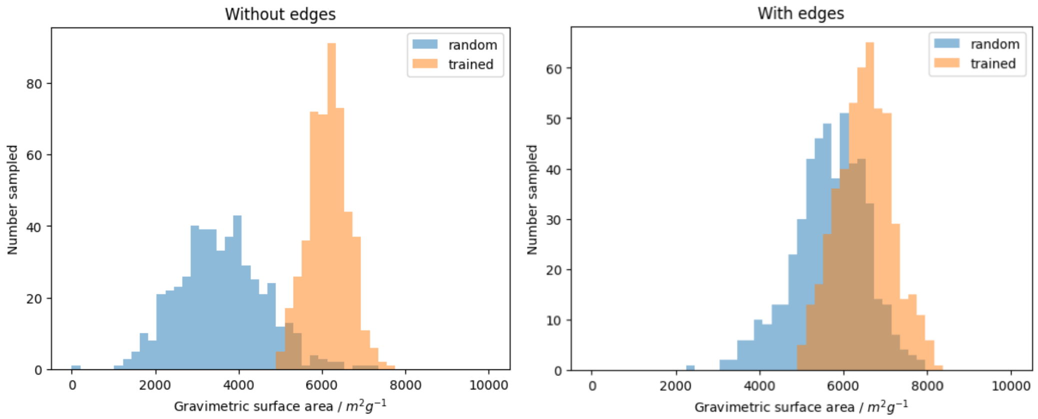

Figure 7: Performance of the GFlowNet trained on the CDZ-E topology.

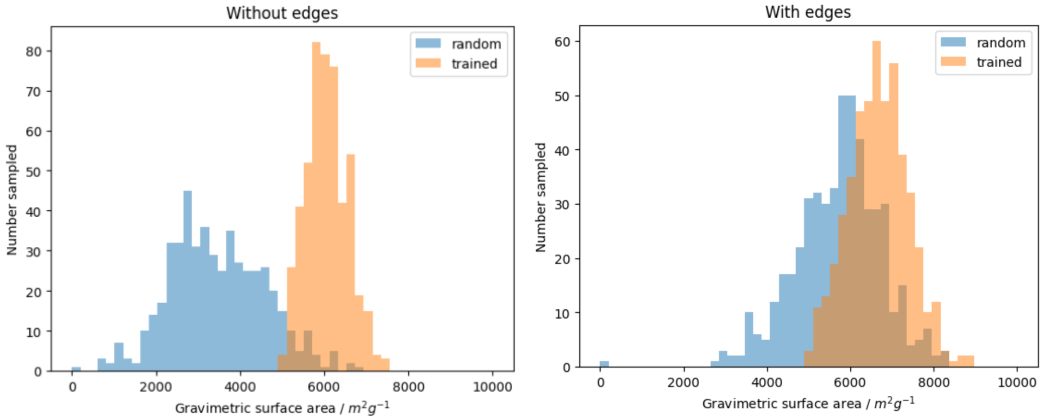

Figure 8: Performance of the GFlowNet trained on the CDL-E topology.

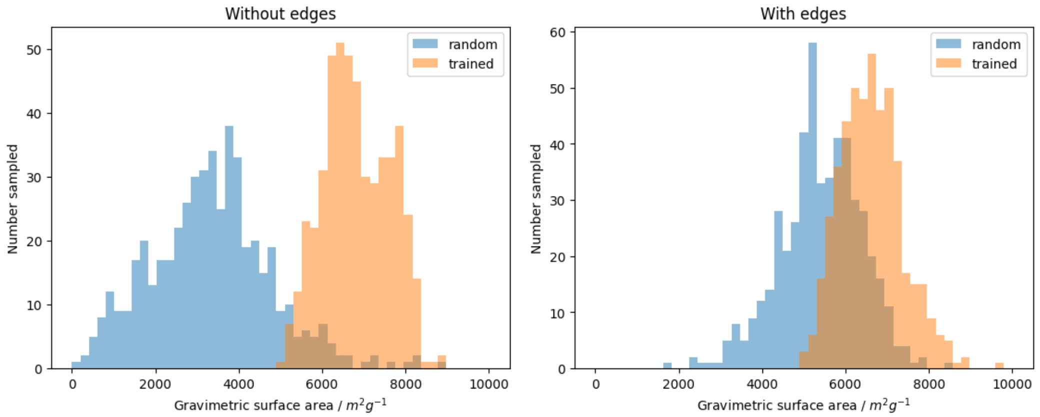

Figure 9: Performance of the GFlowNet trained on the EFT topology.

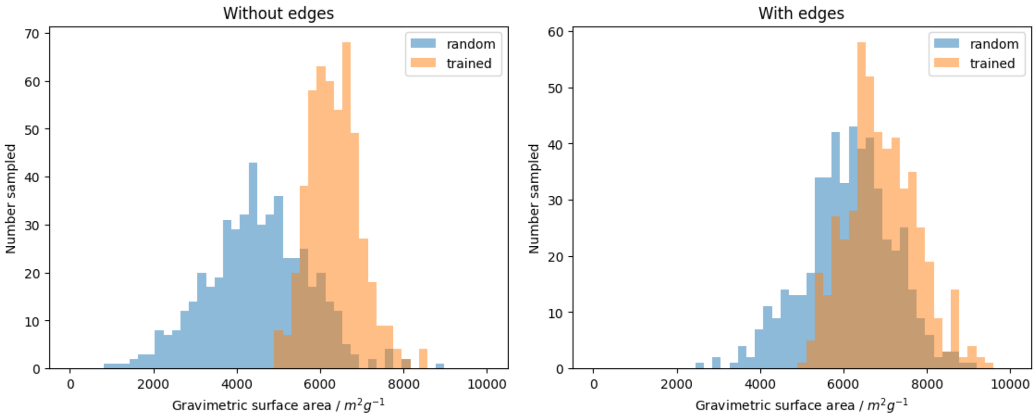

Figure 10: Performance of the GFlowNet trained on the FFC topology.

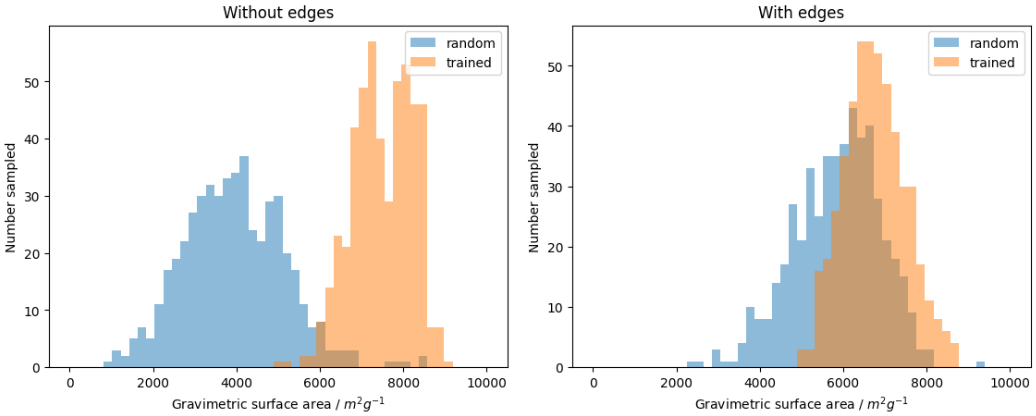

Figure 11: Performance of the GFlowNet trained on the TSG topology.

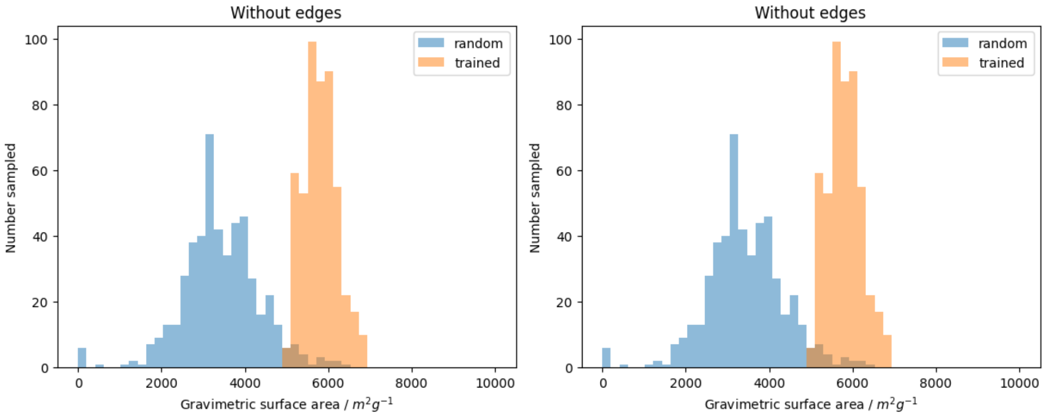

Figure 12: Performance of the GFlowNet trained on the TFF topology.

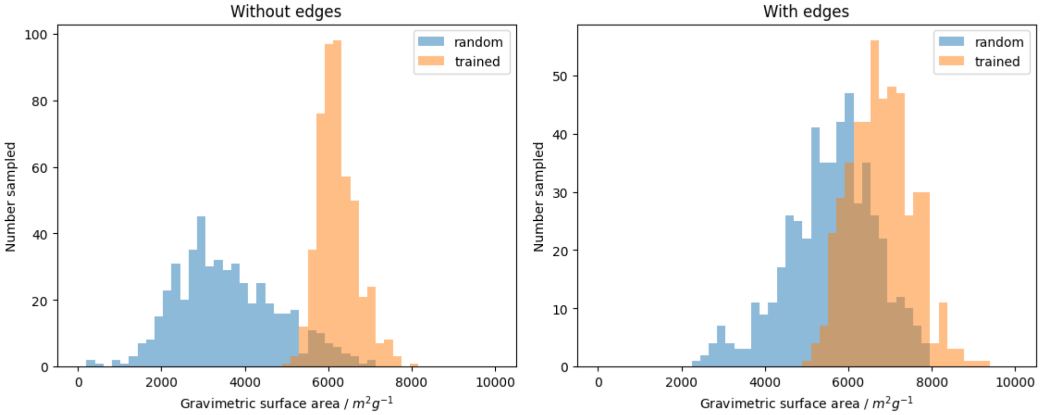

Figure 13: Performance of the GFlowNet trained on the ASC topology.

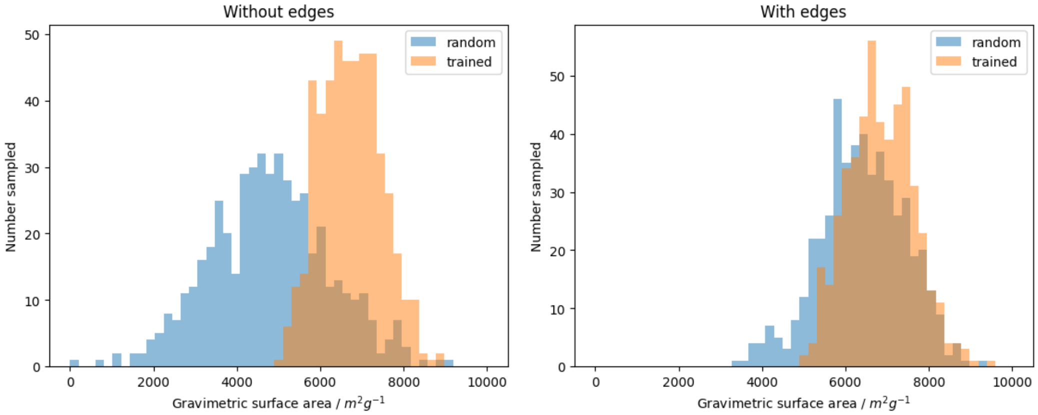

Figure 14: Performance of the GFlowNet trained on the DMG topology.

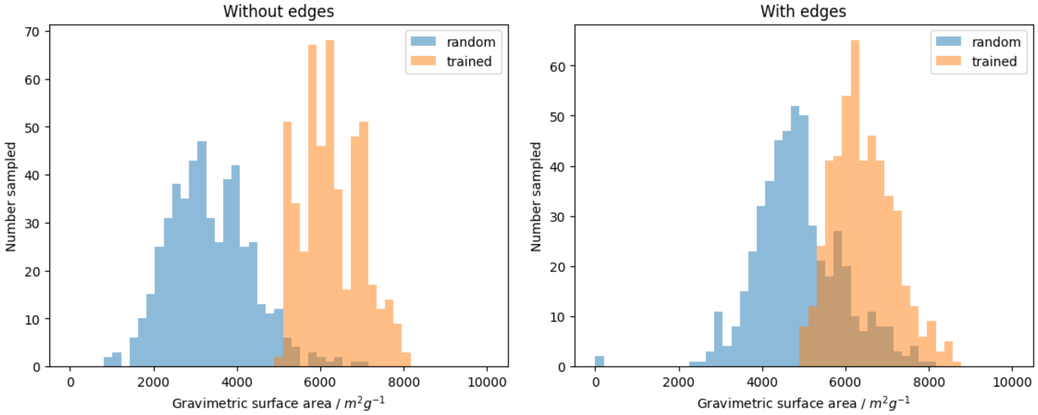

Figure 15: Performance of the GFlowNet trained on the DNQ topology.

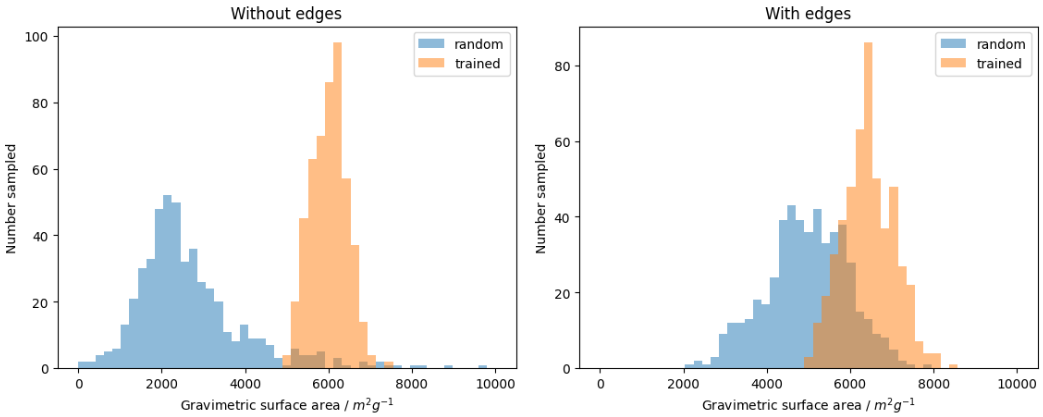

Figure 16: Performance of the GFlowNet trained on the FSO topology.

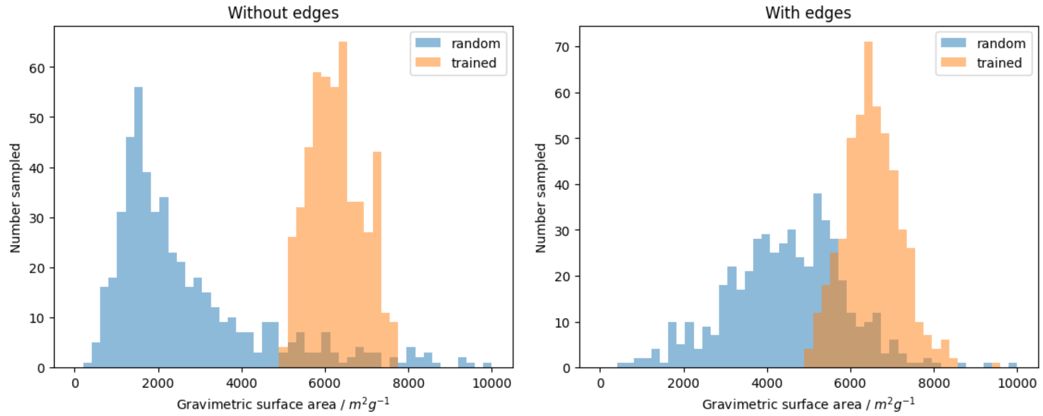

Figure 17: Performance of the GFlowNet trained on the URJ topology.

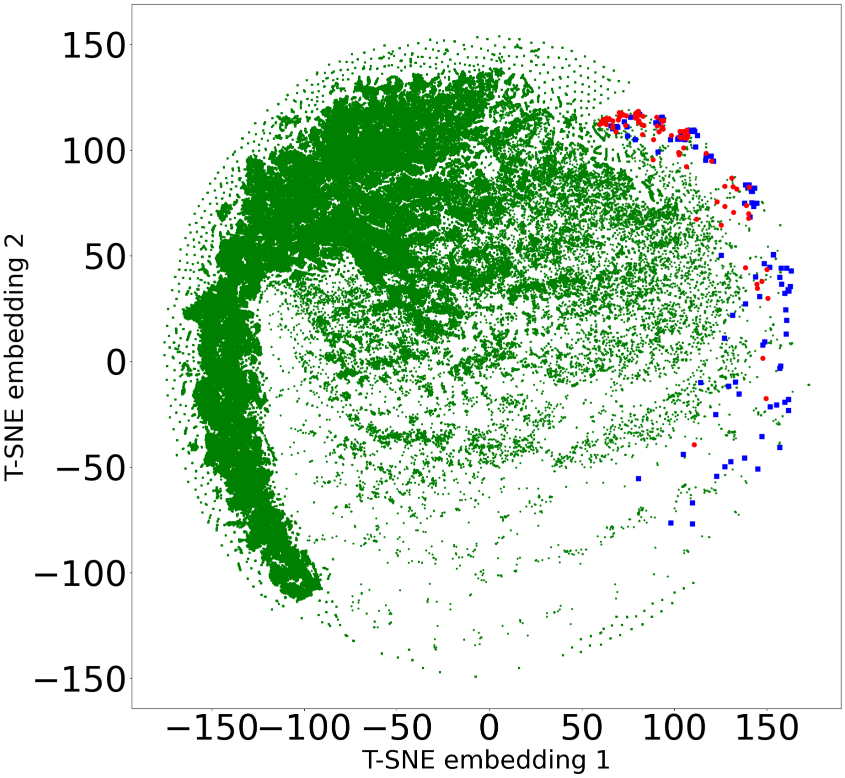

Figure 18: Two dimensional t-SNE embedding of the average minimum distance descriptor of ARC-MOF (green) and the top-100 `matgfn-rm` structures. The red circles are the unrelaxed top-100 structures. The blue squares are the relaxed top-100 structures, with two structures missing due to structural relaxation errors.