scientific reports

Check for updates

OPEN

Gas adsorption meets geometric deep learning: points, set and match

Antonios P. Sarikas[1], Konstantinos Gkagkas[2] & George E. Froudakis[1]✉

Thanks to their unique properties such as ultra high porosity and surface area, metal-organic frameworks (MOFs) are highly regarded materials for gas adsorption applications. However, their combinatorial nature results in a vast chemical space, precluding its exploration with traditional techniques. Recently, machine learning (ML) pipelines have been established as the go-to method for large scale screening by means of predictive models. These are typically built in a descriptor-based manner, meaning that the structure must be first coarse-grained into a 1D fingerprint before it is fed to the ML algorithm. As such, the latter can not fully exploit the 3D structural information, potentially resulting in a model of lower quality. In this work, we propose a descriptor-free framework called "AIdsorb", which can directly process raw structural information for predicting gas adsorption properties. To accomplish that, the structure is first treated as a point cloud and then passed to a deep learning algorithm suitable for point cloud analysis. As a proof of concept, AIdsorb is applied for predicting CO_2 uptake in MOFs, outperforming a conventional pipeline that uses geometric descriptors as input. Additionally, to evaluate the transferability of the proposed framework to different host-guest systems, CH_4 uptake in COFs is examined. Since AIdsorb bases its roots on raw structural information, its applicability extends to all fields of material science.

Metal-organic frameworks (MOFs), the first and most prominent "offspring" of reticular chemistry[1,2], are admittedly one if not, the most intriguing materials of the 21st century. Being essentially a combination of metal ions/clusters and organic linkers[3], MOFs equip researchers with a vast chemical playground for materials design, allowing them to tackle problems in a wide range of fields, spanning from gas storage and separation[4] to drug delivery[5]. Carbon capture is a prime example, where MOF-based sorbents have been deemed as green, low-cost and energy efficient solutions.

Owning to their unprecedented chemical and structural tunability, large databases of either experimental[6,7] or hypothetical[8–10] MOFs have already been developed, and more are expected to emerge in the coming years[11]. Searching for the most promising candidates across these catalogs is undoubtedly a non-trivial task. Obviously, experimentally synthesizing and characterizing each and every one of them is infeasible. Performance evaluation by means of molecular simulations provides a more efficient alternative, dramatically decreasing the time required to assess a single material. Nonetheless, exploring the MOFs space via brute-force computational screening is impractical, given its immensity. *How then we harness this materials space?*

In the era of big data, a subfield of artificial intelligence called *machine learning* (ML) comes to the rescue, enabling the efficient identification of promising materials through predictive models[12–15]. Building the latter amounts to training a (supervised) ML algorithm with a set of inputs and outputs. In ML jargon, inputs and outputs are known as *descriptors* and *labels*, respectively. Within our context, the descriptors provide a mathematical description of the structure, while the output is the property of interest.

The performance of ML models depends to a large extent on the way we select to mathematically describe a material. In other words, the amount of information that is "injected" into the descriptors can make the difference between a high-performing and a baseline model. Regarding gas adsorption in MOFs, various types of descriptors have been proposed with *geometric ones* being the first to be introduced[16] and widely used[17–20]. These descriptors typically include various textural properties such as void fraction and surface area, collectively summarizing the pore geometry of the framework. Chemical descriptors[21–24] are another type of MOF descriptors, aiming to capture the chemical character of the framework. For example, Fanourgakis et. al[25] introduced the number density of atom types, a standardized count (divided by the unit cell volume) of the atom types in the unit cell. Atom types provide information about the hybridization and connectivity of the MOF atoms, effectively

[1]Department of Chemistry, University of Crete, Voutes Campus, 70013 Heraklion, Crete, Greece. [2]Advanced Technology Division, Toyota Motor Europe NV/SA, Technical Center, Hoge Wei 33B, 1930 Zaventem, Belgium. ✉email: frudakis@uoc.gr

describing the chemistry of the framework. Energy-based descriptors[26–30], i.e. descriptors that take into account host-guest interactions, have also been developed. For instance, Bucior et al.[31] fingerprinted the potential energy landscape of H_2-MOF interactions through the construction of sorbate-sorbent energy histograms.

Irrespective of their type, descriptors inevitably introduce the following problems into a ML pipeline.

1. *Need to be designed*, a process which requires a significant amount of human effort and domain knowledge.
2. *Require calculation*, adding an extra computational overhead to the pipeline and as such, slowing down the deployment of the model for large scale screening.
3. More importantly, they *may lead to significant information loss* and hence decrease model's performance, as a 3D object, the structure, is coarse-grained into a 1D (or 2D) fingerprint.In reticular chemistry and of course chemistry, "every Angstrom matters", meaning that *the ML algorithm should ideally be aware of the exact arrangement of atoms in 3D space*. To put it differently, if our aim is to model the underlying structure-property relationship, *why provide the algorithm with a description of the structure and not the structure itself?*

In this work we present "AIdsorb" (Fig. 1), a *descriptor-free* framework that can *directly consume raw structural information to predict gas adsorption properties*. To achieve this, we:

1. Treat the structure as a point cloud
2. Choose a suitable algorithm for learning on point cloudsA *point cloud*, being essentially a *set of 3D points and associated information*, provides a natural and lossless way to mathematically represent a structure. In our case, the 3D points correspond to *atomic positions*, while the associated information corresponds to *atomic numbers and optionally extra chemical information*. We refer to such a point cloud as "molecular point cloud". Extra information added to each point can be in the form of atomic properties, such as the electronegativity and ionization energy of the atom, or properties summarizing the local environment of the atom, e.g. the average electronegativity of the first coordination sphere. In this work only atomic properties were considered, namely electronegativity, van der Waals radius and dipole polarizability, which are collectively denoted as $\mathcal{F}$. More details can be found on Section 1 of Supplementary Information (SI).

With regards to the choice of ML algorithm, we turn our attention to geometric deep learning (DL)[32], the branch of DL that deals with unstructured data, such as graphs and point clouds. For this study, the algorithm of choice is a lightweight version of PointNet[33] (see Fig. 1 and SI for more details), a simple yet robust DL architecture for point cloud processing.

Although DL algorithms are notorious for being data hungry, training such as algorithms nowadays and expecting them to generalize well—at least within the field of reticular chemistry—should be reasonable, given the vast amount of data currently available[34]. Thanks to the ability of DL algorithms to perform automatic

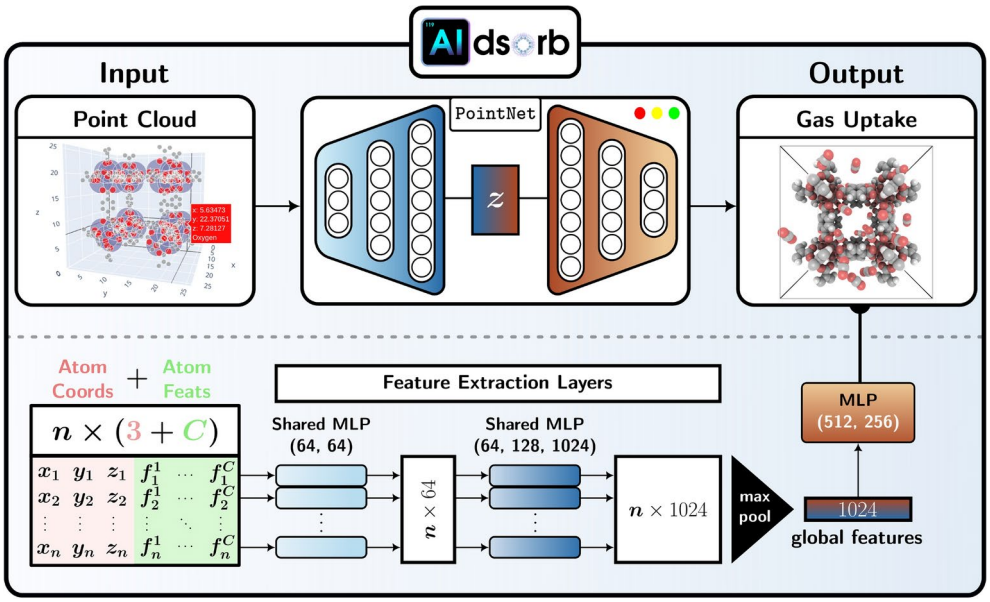

Fig. 1. (Top) Generalized framework to predict gas adsorption properties by using a molecular point cloud as input. (Bottom) The lightweight version of PointNet architecture used in this work. Each point in the cloud is processed identically and independently by the feature extraction layers (shared MLPs). After passing these layers, each point has been transformed from being represented with $3 + C$ features into 1024 features. Then, a max pooling layer aggregates the per-point features into a global "signature" for the molecular point cloud. The latter is fed into a MLP that generates predictions for the property(ies) interest (hereon gas uptake). Numbers in parentheses are layer sizes. MLP, multi-layer perceptron.

feature extraction from raw data, combining a DL architecture with molecular point clouds introduces a versatile paradigm for data-driven material science, bypassing the obstacles of manually crafted descriptors.

The ability of the proposed approach to directly process raw structural information, allows it to be applied to any host-guest system for modeling any property of interest. As a proof of concept, AIdsorb is applied on MOFs for predicting CO_2 uptake. Additionally, we showcase its transferability by examining CH_4 uptake on COFs. In both cases, the suggested pipeline is compared with conventional ones that use geometric descriptors as input.

Results and discussion

In order to evaluate our pipeline's performance, PointNet is trained (see Section 3.3 of SI for training details) and tested on the University of Ottawa database[9], for predicting CO_2 uptake at 298K and 0.15 bar. For the sake of comparison, a conventional model is built with the random forest (RF) algorithm[35], serving as our baseline. For both pipelines the same random subsets of 291 984 and 24 331 materials were used as training and test sets, respectively. As it can be seen from the parity plots of Fig. 2, the PointNet model achieves a R^2 value of 0.897, outperforming the conventional one, which shows a R^2 value of 0.753. This performance gap of approximately 20% highlights the importance of preserving and not coarse-graining raw structural information.

Furthermore, the transferability of the approach is demonstrated by applying AIdsorb to the COFs database created by Mercado et. al[36], for predicting CH_4 uptake at 298K and 5.8 bar. In this case, PointNet is trained and tested with a random subset of 59 363 and 6984 materials, respectively. As shown in Fig. S2, the predictions of the resulting model are in great agreement with the ground truth values ($R^2 = 0.966$). Again, the PointNet model performs better than the conventional one ($R^2 = 0.946$), which was trained and tested on the same data as the former. It should be noted that the performance gap in this case is less pronounced compared to the CO_2 case. This should be attributed to the fact that geometric descriptors are sufficient when modeling gases with negligible electrostatic interactions, such as CH_4[16,37] or X_e[18]. That is, the coarse-grained structural information that geometric descriptors encode suffices to accurately predict CH_4 uptake but is not enough when predicting CO_2 uptake.

To understand whether the addition of chemical information into point clouds affects the predictive accuracy of PointNet, the latter was trained with different types of point clouds for both MOFs-CO_2 and COFs-CH_4 cases. Specifically, PointNet was trained with point clouds containing information about:

1. coordinates only
2. coordinates and atomic numbers
3. coordinates, atomic numbers and atomic propertiesAs can be seen from Table S3, the performance of PointNet systematically improves when information about the atomic number is incorporated into the point cloud ($xyz + Z$). However, when atomic properties are added to the point cloud ($xyz + Z + \mathcal{F}$) no significant improvements (or none at all) are observed.

This may be attributed to the limitation of the PointNet architecture to combine the individual atomic properties and extract useful local features—features describing the local chemical environment around each atom—since it process each point independently (see Fig. 1). A straightforward approach to bypass this limitation is to enrich the point cloud with features that encode local chemical information for an atom, such as atom types[25]. Alternatively, instead of adding local features manually, one can replace PointNet with an architecture capable of extracting local features[38].

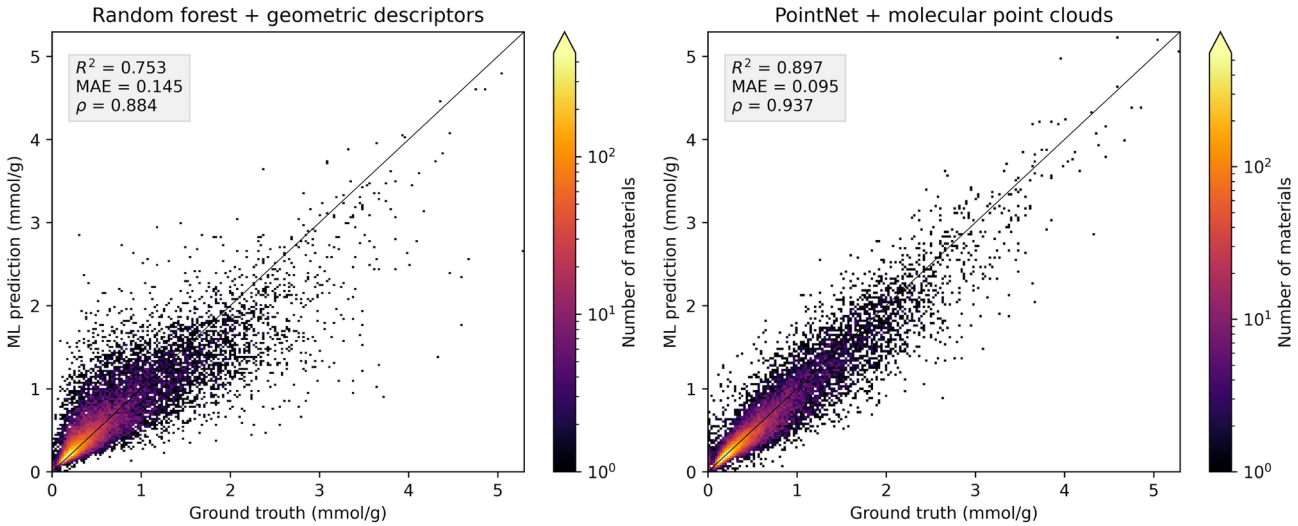

Fig. 2. Parity plots for conventional model (left) and PointNet model (right) regarding CO_2 uptake in MOFs. All metrics were measured on the test set. R^2, coefficient of determination (unitless); MAE, mean absolute error (capacity units); ρ, Spearman's rho (unitless).

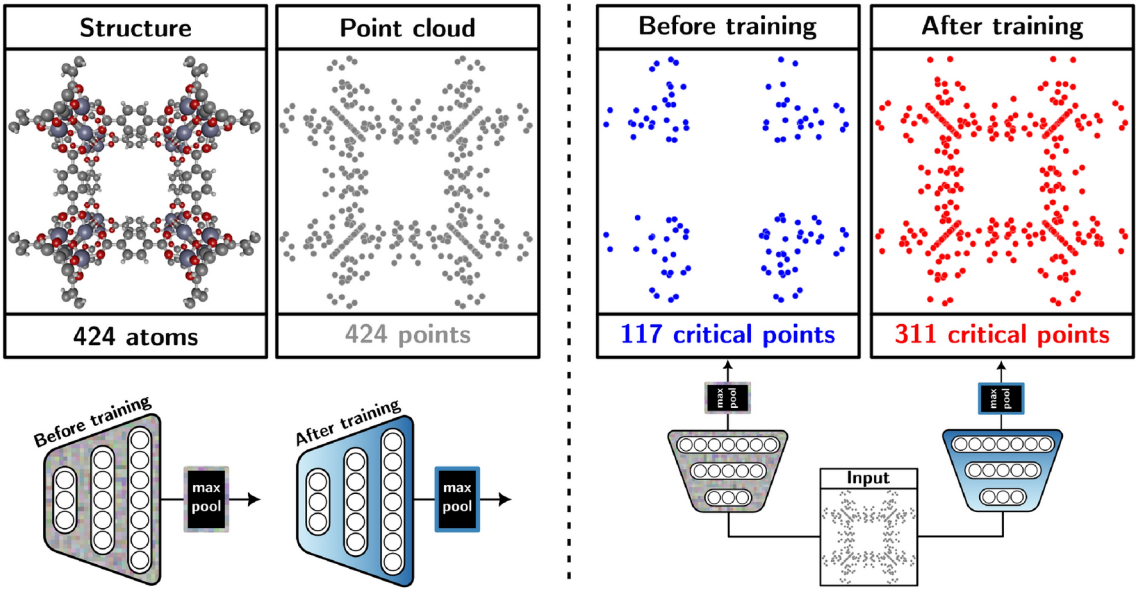

Fig. 3. Critical points of IRMOF-1. (Blue) The critical points are obtained by passing the point cloud of IRMOF-1 through PointNet before training starts, i.e. when PointNet has its parameters randomly initialized. (Red) The critical points are obtained by passing the point cloud of IRMOF-1 through PointNet after training ends, i.e. when PointNet has its parameters optimized. For the sake of clarity, atomic information is omitted from the point cloud drawings.

In order to get some insights into model's internals, we visualize the *critical points* of IRMOF-1's point cloud (424 points) after passing it through the PointNet model trained on the MOFs dataset (see Section 3.3 of SI for training details). These are the points that contribute to the pooled global feature and hence, these are *the only points that can contribute to model's output*. In other words, PointNet takes into account only these points and discards every other point. As can be seen from Fig. 3, when the point cloud is passed through a randomly initialized PointNet, the critical points (117 points, shown in blue) don't properly account for the geometry of the framework. In contrast, when the point cloud is passed through the trained PointNet, the critical points (311 point, shown in red) effectively summarize the skeleton of the framework, similar to the original application in computer vision where the critical points summarize the skeleton of the object[33]. That is, the output of PointNet is dictated by a set of points that effectively capture the geometry of the material.

In conclusion, we demonstrated that AIdsorb can yield accurate predictions regarding gas adsorption in porous materials by just using a molecular point cloud as input. As the choice of ML algorithm strongly affects the quality of the resulting model, coupling AIdsorb with a more refined architecture[38,39] can further improve its performance. Additional enhancements on data efficiency or performance can be achieved by employing improved training schemes, such as self-supervised pre-training[40,41] and auxiliary learning[42]. Finally, since molecular point clouds can essentially represent any chemical system, the presented approach can be extended and applied beyond the realm of reticular chemistry, for elucidating any structure-property relationship.

Data & Code Availability

Point clouds used in this work are available from the corresponding author on reasonable request. Labels (gas uptake values) and geometrics descriptors are publicly available in: https://archive.materialscloud.org/record/2018.0016/v3 (MOFs) and https://archive.materialscloud.org/record/2018.0003/v2 (COFs). The source code of "AIdsorb" package is available at: https://github.com/frudakis-research-group/aidsorb. The version used in this work is specified by the tag "pointnet_paper". The package can be installed from PyPI: https://pypi.org/project/aidsorb/. Documentation for the package is available at: https://aidsorb.readthedocs.io/en/stable/.

Received: 24 July 2024; Accepted: 14 October 2024
Published online: 09 November 2024

References

1. Yaghi, O. M. Emergence of metal-organic frameworks. In *Introduction to Reticular Chemistry*. Chap. 1, 1–27. https://doi.org/10.1002/9783527821099.ch1 (Wiley, 2019).
2. Yaghi, O. M. The reticular chemist. *Nano Lett.* **20**(12), 8432–8434. https://doi.org/10.1021/acs.nanolett.0c04327 (2020).
3. Yaghi, O. M. Reticular chemistry in all dimensions. *ACS Cent. Sci.* **5**(8), 1295–1300. https://doi.org/10.1021/acscentsci.9b00750 (2019).
4. Li, B. et al. Porous metal-organic frameworks for gas storage and separation: what, how, and why?. *J. Phys. Chem. Lett.* **5**(20), 3468–3479. https://doi.org/10.1021/z501586e (2014).
5. Lawson, H. D., Walton, S. P. & Chan, C. Metal-organic frameworks for drug delivery: a design perspective. *ACS Appl. Mater. Interfaces* **13**(6), 7004–7020. https://doi.org/10.1021/acsami.1c01089 (2021).

6. Moghadam, P. Z. et al. Development of a Cambridge structural database subset: a collection of metal-organic frameworks for past, present, and future. *Chem. Mater.* **29**(7), 2618–2625. https://doi.org/10.1021/acs.chemmater.7b00441 (2017).
7. Chung, Y. G. et al. Advances, updates, and analytics for the computation-ready, experimental metal-organic framework database: CoRE MOF 2019. *J. Chem. Eng. Data* **64**(12), 5985–5998. https://doi.org/10.1021/acs.jced.9b00835 (2019).
8. Wilmer, C. E. et al. Large-scale screening of hypothetical metal-organic frameworks. *Nat. Chem.* **4**(2), 83–9 (2011).
9. Boyd, P. G. et al. Data-driven design of metal-organic frameworks for wet flue gas CO_2 capture. *Nature* **576**(7786), 253–256. https://doi.org/10.1038/s41586-019-1798-7 (2019).
10. Rosen, A. S. et al. Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery. *Matter* **4**(5), 1578–1597. https://doi.org/10.1016/j.matt.2021.02.015 (2021).
11. Lee, S. et al. Computational screening of trillions of metal-organic frameworks for high-performance methane storage. *ACS Appl. Mater. Interfaces* **13**(20), 23647–23654. https://doi.org/10.1021/acsami.1c02471 (2021).
12. Fanourgakis, G. S. et al. Fast screening of large databases for top performing nanomaterials using a self-consistent, machine learning based approach. *J. Phys. Chem. C* **124**(36), 19639–19648. https://doi.org/10.1021/acs.jpcc.0c05491 (2020).
13. Choudhary, K. et al. Graph neural network predictions of metal organic framework CO2 adsorption properties. *Comput. Mater. Sci.* **210**, 111388. https://doi.org/10.1016/j.commatsci.2022.111388 (2022).
14. Kang, Y. et al. A multi-modal pre-training transformer for universal transfer learning in metal-organic frameworks. *Nat. Mach. Intell.* **5**(3), 309–318. https://doi.org/10.1038/s42256-023-00628-2 (2023).
15. Sarikas, A. P., Gkagkas, K. & Froudakis, G. E. Gas adsorption meets deep learning: voxelizing the potential energy surface of metal-organic frameworks. *Sci. Rep.* https://doi.org/10.1038/s41598-023-50309-8 (2024).
16. Fernandez, M. et al. Large-scale quantitative structure-property relationship (QSPR) analysis of methane storage in metal-organic frameworks. *J. Phys. Chem. C* **117**(15), 7681–7689. https://doi.org/10.1021/jp4006422 (2013).
17. Zhang, X. et al. Machine learning prediction on properties of nanoporous materials utilizing pore geometry barcodes. *J. Chem. Inf. Model.* **59**(11), 4636–4644. https://doi.org/10.1021/acs.jcim.9b00623 (2019).
18. Liang, H. et al. XGBoost: an optimal machine learning model with just structural features to discover MOF adsorbents of Xe/Kr. *ACS Omega* **6**(13), 9066–9076. https://doi.org/10.1021/acsomega.1c00100 (2021).
19. Ahmed, A. & Siegel, D. J. Predicting hydrogen storage in MOFs via machine learning. *Patterns* **2**(7), 100291. https://doi.org/10.1016/j.patter.2021.100291 (2021).
20. Ren, E. & Coudert, F.-X. Prediction of the diffusion coefficient through machine learning based on transition-state theory descriptors. *J. Phys. Chem. C* **128**(16), 6917–6926. https://doi.org/10.1021/acs.jpcc.4c00631 (2024).
21. Fernandez, M. et al. Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO_2 capture. *J. Phys. Chem. Lett.* **5**(17), 3056–3060. https://doi.org/10.1021/jz501331m (2014).
22. Borboudakis, G. et al. Chemically intuited, large-scale screening of MOFs by machine learning techniques. *npj Comput. Mater.* **3**, 1–7 (2017).
23. Pardakhti, M. et al. Machine learning using combined structural and chemical descriptors for prediction of methane adsorption performance of metal organic frameworks (MOFs). *ACS Comb. Sci.* **19**(10), 640–645. https://doi.org/10.1021/acscombsci.7b00056 (2017).
24. Burner, J. et al. High-performing deep learning regression models for predicting low-pressure CO2 adsorption properties of metal-organic frameworks. *J. Phys. Chem. C* **124**(51), 27996–28005. https://doi.org/10.1021/acs.jpcc.0c06334 (2020).
25. Fanourgakis, G. S. et al. A universal machine learning algorithm for large-scale screening of materials. *J. Am. Chem. Soc.* **142**(8), 3814–3822. https://doi.org/10.1021/jacs.9b11084 (2020).
26. Simon, C. M. et al. What are the best materials to separate a xenon/krypton mixture?. *Chem. Mater.* **27**(12), 4459–4475. https://doi.org/10.1021/acs.chemmater.5b01475 (2015).
27. Fanourgakis, G. S. et al. A generic machine learning algorithm for the prediction of gas adsorption in nanoporous materials. *J. Phys. Chem. C* **124**(13), 7117–7126. https://doi.org/10.1021/acs.jpcc.9b10766 (2020).
28. Orhan, I. B. et al. Accelerating the prediction of CO2 capture at low partial pressures in metal-organic frameworks using new machine learning descriptors. *Commun. Chem.* https://doi.org/10.1038/s42004-023-01009-x (2023).
29. Shi, K. et al. Two-dimensional energy histograms as features for machine learning to predict adsorption in diverse nanoporous materials. *J. Chem. Theory Comput.* https://doi.org/10.1021/acs.jctc.2c00798 (2023).
30. Deng, Z. & Sarkisov, L. Engineering machine learning features to predict adsorption of carbon dioxide and nitrogen in metal-organic frameworks. *J. Phys. Chem. C* https://doi.org/10.1021/acs.jpcc.4c01692 (2024).
31. Bucior, B. J. et al. Energy-based descriptors to rapidly predict hydrogen storage in metal-organic frameworks. *Mol. Syst. Des. Eng.* **4**, 162–174. https://doi.org/10.1039/C8ME00050F (2019).
32. Bronstein, M. M. et al. *Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges*. https://doi.org/10.48550/ARXIV.2104.13478 (2021).
33. Qi, C. R. et al. *PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation*. https://doi.org/10.48550/ARXIV.1612.00593 (2016).
34. Bobbitt, N. S. et al. MOFX-DB: an online database of computational adsorption data for nanoporous materials. *J. Chem. Eng. Data* **68**(2), 483–498. https://doi.org/10.1021/acs.jced.2c00583 (2023).
35. Breiman, L. *In Machine Learning* **45**(1), 5–32. https://doi.org/10.1023/a:1010933404324 (2001).
36. Mercado, R. et al. In silico design of 2D and 3D covalent organic frameworks for methane storage applications. *Chem. Mater.* **30.15**, 5069–5086. https://doi.org/10.1021/acs.chemmater.8b01425 (2018).
37. Suyetin, M. The application of machine learning for predicting the methane uptake and working capacity of MOFs. *Faraday Discuss.* **231**, 224–234. https://doi.org/10.1039/d1fd00011j (2021).
38. Qi, C. R. et al. *PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space*. https://doi.org/10.48550/ARXIV.1706.02413 (2017).
39. Wang, Y. et al. *Dynamic Graph CNN for Learning on Point Clouds*. https://doi.org/10.48550/ARXIV.1801.07829 (2018).
40. Cao, Z. et al. MOFormer: self-supervised transformer model for metal-organic framework property prediction. *J. Am. Chem. Soc.* **145.5**, 2958–2967. https://doi.org/10.1021/jacs.2c11420 (2023).
41. Wang, J. et al. A comprehensive transformer-based approach for high-accuracy gas adsorption predictions in metal-organic frameworks. *Nat. Commun.* https://doi.org/10.1038/s41467-024-46276-x (2024).
42. Cui, J. et al. Direct prediction of gas adsorption via spatial atom interaction learning. *Nat. Commun.* https://doi.org/10.1038/s41467-023-42863-6 (2023).

Acknowledgements

This research has been co-financed by Toyota Motor Europe NV/SA. Authors would like to acknowledge financial support from European Union: Horizon Europe (project MOST-H_2; Grant agreement no. 101058547). The Research implemented was supported by the University of Crete Research Committee funds (Project Code Number 3650).

Author contributions

Conceptualization, A.P.S. and G.E.F.; Methodology and software development, A.P.S.; Investigation, A.P.S.; Resources, K.G. and G.E.F.; Writing-original draft preparation, A.P.S.; Writing-review and editing, K.G. and G.E.F.; Supervision, G.E.F. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-76319-8.

Correspondence and requests for materials should be addressed to G.E.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.