Journal of
Materials Chemistry A

ROYAL SOCIETY
OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *J. Mater. Chem. A*, 2021, **9**, 21175

Finely tuned inverse design of metal–organic frameworks with user-desired Xe/Kr selectivity†

Yunsung Lim, Junkil Park, Sangwon Lee and Jihan Kim *

Received 14th April 2021
Accepted 17th August 2021

DOI: 10.1039/d1ta03129e

rsc.li/materials-a

Inverse materials design entails providing desired properties as inputs and obtaining fine-tuned materials that fit the given criteria as outputs. Although this workflow would in principle lead to significant efficiency in materials design, it is difficult in practice to successfully implement a robust, accurate inverse design platform. In this work, we used a validated platform which integrates a genetic algorithm with machine learning to design user-desired metal–organic frameworks (MOFs) with the xenon/krypton separation being presented as a case study. Using our platform, we obtained two record-breaking MOFs that show significant improvement over the current record. Moreover, with facile modification in the cost function, we demonstrate that our platform can generate MOFs that are finely tuned to the specific desires of users across multiple properties and a range of property values.

Introduction

Advancement in the field of artificial intelligence has led to a significant shift in how scientists design and synthesize new nanomaterials.[1–3] In line with this progress, the development of material informatics has resulted in accumulation of big materials' data (in the form of both experimentally synthesized and hypothetical materials), which can be used as inputs to deep neural networks for materials design.[4–7] Moreover, with parallel efforts made in applying deep learning techniques to expedite and facilitate chemical synthesis,[8,9] the pace of materials design and deployment will accelerate significantly in the future.

Amongst many different materials, metal–organic frameworks (MOFs) are considered to be intriguing materials to deploy deep learning methods given the large number of available data and an even larger materials space that has yet to be explored. MOFs are comprised of tunable building blocks (*i.e.* metal nodes and organic linkers) that can connect in various different combinations, which has led to the synthesis of over hundreds of thousands of MOFs.[10] And many of these materials are considered to be promising candidates for various applications including gas separation, storage, catalysis, and sensors.[11–14] With that said, the exploration of the MOF space has barely scratched its surface since the facility of mixing and matching different inorganic and organic building blocks can lead to enormous numbers of MOFs that can in principle be synthesized. Although there have been many numbers of computational screening studies that have led to the discovery of optimal MOFs for different applications, most of these studies are confined to the pre-existing database of materials (*e.g.* experimentally synthesized CoRE MOFs[15]) and are limited to less than millions of structures.[16–21]

To this end, one efficient alternative way to search for optimal materials is *via* inverse design. Within this paradigm, materials are constructed with user-desired properties as part of the input (often in the form of objective function).[22–24] As such, theoretical materials space can be significantly reduced with less time spent in analyzing low performance materials during the search process compared to the conventional screening methodology. Especially, in the case of crystalline nanoporous materials, inverse design is potentially one of the most powerful ways to explore their unprecedented large materials space efficiently and as such, there have been few studies devoted to this topic. In particular, Kim *et al.* used a generative adversarial network (GAN) to design zeolites with user-desired methane heat of adsorption values.[25] Also, Yao *et al.* generated hypothetical MOFs (hMOFs) with high CO_2 separation performance within the chemical space with variational autoencoder (VAE).[26] Finally, Lee *et al.* combined a deep neural network (DNN) and genetic algorithm (GA) to discover hMOFs with record-breaking methane working capacity values for ANG applications.[27]

There have been some studies on other nanomaterials that focused on targeting specific values or conditions of functionalities within materials with a relatively simple structure (*e.g.* nanophotonic structures,[28,29] metal oxides,[30–32] alloys,[33] soft materials[34]). However, within the world of porous materials, most of the targeted materials involve using a global optimization algorithm to identify materials with the best properties for a given application.[26,27,35,36] However, in practical applications, it is often the case that once a material surpasses a certain

Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea. E-mail: jihankim@kaist.ac.kr

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d1ta03129e

(good-enough) threshold value for a given property, other metrics (such as materials cost and chemical/thermal stability) become more important.[37–39] From this point of view, a flexible tool that enables one to target fine-tuned user-desired properties (*e.g.* materials with selectivity higher than a certain value, materials with specific working capacity) would be invaluable for computational materials design. With these flexible objective functions in place, one can then consider other dimensions of properties amongst the pool of selected candidates that can facilitate the search for practical materials that can be deployed for industry purposes.

In this work, we revised the pre-developed workflow[27] to facilitate the finely tuned inverse design of MOFs. To demonstrate our capabilities, a test case study of xenon/krypton (Xe/Kr) separation from used nuclear fuel (UNF) off-gas has been chosen due to the importance of isolating pure xenon for various practical applications (*e.g.* X-ray, laser, medical industry, and analytical chemistry).[40–43] Also, separation of xenon from UNF off-gas by using adsorbents has been regarded as a promising alternative technique because of its less energy-intensive and cost effectiveness. The typical concentration of xenon and krypton is 400 ppm and 40 ppm within UNF off-gas and as such, we measured selective adsorbing capacity for xenon over krypton at very low pressure (*i.e.* infinite dilute condition). According to the results of the single column break through experiment from the previous work, the ratio of the Henry coefficient of xenon over krypton can be a suitable representative to measure selective adsorption of xenon from UNF off-gas.[44] Previously, Simon *et al.* screened 670 000 porous materials using a random forest algorithm to discover materials with the highest Xe/Kr selectivity ($S_{\mathrm{Xe/Kr}}$), with one of them (*i.e.* SBMOF-1) being successfully synthesized.[44,45] In this work, as shown in Fig. 1, we explored a much larger MOF space by integrating machine learning with a genetic algorithm and using a flexible cost function. By providing users with precise control on targeting MOFs with specific working capacity or selectivity, our work opens up a flexible means through which different objectives can be concurrently met, leading to a more personalized materials design. And it is our opinion that this type of flexibility will be important for users to generate materials for many applications in which multiple dimensions of properties (*e.g.* materials performance, cost, stability) are regarded as critical factors for consideration.

Results and discussion

Record-breaking frameworks

Initially, 65 201 MOFs were generated from 1775 topologies and 953 building blocks, and their xenon and krypton Henry coefficients ($K_{\mathrm{H,Xe}}$ and $K_{\mathrm{H,Kr}}$) were computed *via* the Widom insertion method to obtain a large pool of training set data. At each genetic algorithm cycle, approximately 15 000 to 20 000 new MOFs were selected as best fit candidates and Monte Carlo (MC) simulations were conducted on these new structures. Then, they were appended to the pre-existing training set data. This cycle was repeated ten times to find optimal MOFs for Xe/Kr separation. Altogether, 245 618 MOFs were screened during this process and it can be seen that the procedure successfully removes low performing structures from one cycle to the next (Fig. 2, ESI Fig. S1 and S2†). Compared to the initial randomly generated MOF dataset (cycle 0; blue plot of Fig. 2a), subsequent cycles led to MOFs with higher selectivity. Furthermore, as can be seen in Fig. 2a, the selectivity value of the top 1% MOFs becomes larger as the cycle progresses (13, 23, 26, and 30 for

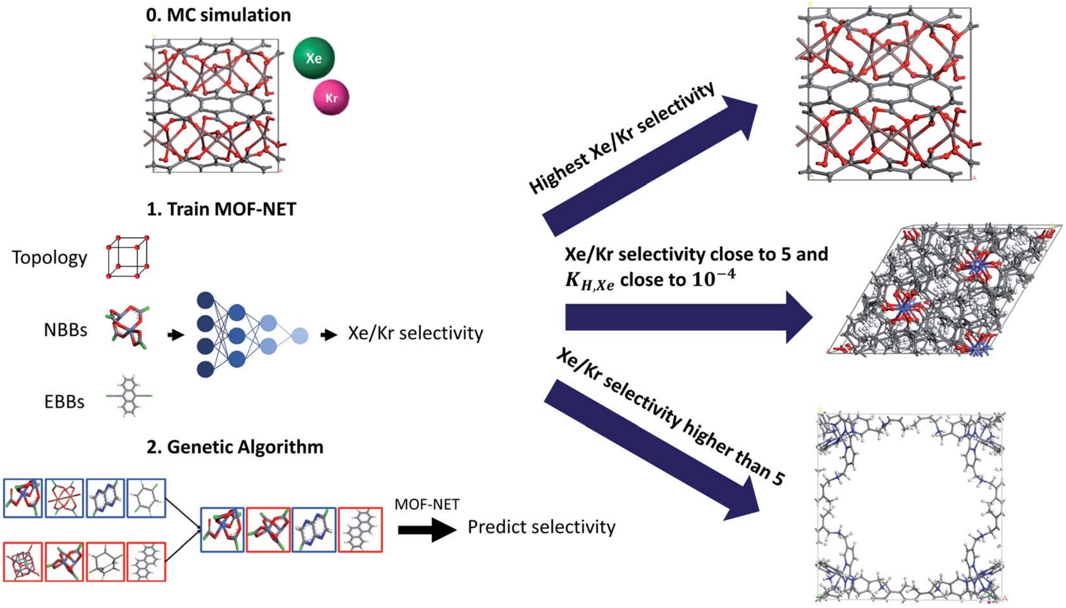
0. MC simulation
Xe
Kr
1. Train MOF-NET
Topology
NBBs
EBBs
Xe/Kr selectivity
2. Genetic Algorithm
MOF-NET
Predict selectivity
Highest Xe/Kr selectivity
Xe/Kr selectivity close to 5 and $K_{H,Xe}$ close to 10^{-4}
Xe/Kr selectivity higher than 5

Fig. 1 Overall schematics of our computational workflow. MOF-NET (machine learning model) was trained with results from the MC simulations. Trained MOF-NET weights were used to screen over the MOF space during the GA (genetic algorithm). The genetic algorithm worked as a flexible tool to generate frameworks with desired conditions.

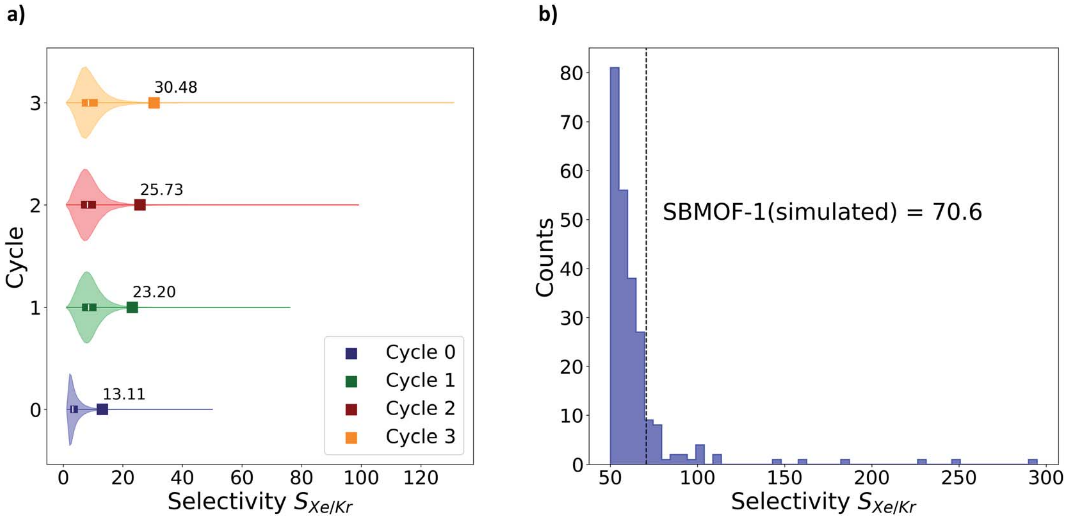

Fig. 2 Results of the consecutive genetic algorithm cycles to discover record-breaking MOFs. (a) Selectivity distribution of randomly generated MOFs for the initial training set (cycle 0) and frameworks generated from the 1st, 2nd, and 3rd cycle (large squares denote cut-off values for the top 1 percent MOFs; small rectangles denote the positions of the 1st quarter (25%) to 3rd quarter (75%) selectivity values). (b) Selectivity distribution of MOFs that have selectivity higher than 50 after ten cycles.

cycle 0, 1, 2, and 3, respectively). For the subsequent cycles (cycle 4 to 10), dramatic enhancements in performance were not observed for newly generated MOFs because the procedure already covered the optimum locations within the materials space (see ESI Fig. S2†). Nevertheless, additional cycles were carried out to obtain a larger pool of candidate MOFs. After the 10th cycle, it was observed that 32 MOFs possessed higher selectivity than SBMOF-1, which holds the current record value (Fig. 2b).[44] From the selected candidate materials with high selectivity, MOFs with low adsorption properties (*i.e.* $K_{H,Xe} < 10^{-6}$ mol kg^{-1} Pa^{-1}) were excluded. However, Xe/Kr selectivity correlates well with $K_{H,Xe}$ (see ESI Fig. S3†),[44–46] so effectively only 10% of the MOFs were eliminated.

From the final 32 candidate materials, three more simulations were conducted to validate our result: (1) flexible molecular simulations, (2) RASPA simulations, and (3) polymorphic simulations. Then, flexible molecular simulations were conducted to observe the influence of framework flexibility on selectivity. Previously published studies have shown a relatively large discrepancy in the Xe/Kr selectivity values between the simulation and experimental data (see ESI Fig. S13†).[44] For example, the simulated Xe/Kr selectivity of SBMOF-1 is 71, but the experimental result is only 16.[44] One source of this discrepancy comes from fixing the MOF as rigid in most of these molecular simulations. As such, the flexible simulation of SBMOF-1 yielded a Xe/Kr selectivity value of 37, which is closer to the experimental value of 16.[47] Therefore, the selectivity values of the 32 candidate materials were recomputed using the "flexible snapshot" method. During the process, five MOFs were excluded due to a technical issue (*i.e.* overlapping atom pairs) regarding generating input files for molecular dynamics (MD) simulation when using the python library "lammps-interface".[48] At the end, 19 MOFs still showed higher selectivity compared to both flexible and rigid simulation results for SBMOF-1 (denoted by the red square point in Fig. 3). As shown in Fig. 3, flexible simulations yielded selectivity values that are in general smaller than that of rigid simulations which agrees with the trends found in previous studies.[47] Then, RASPA simulations were conducted to validate results from GPU based simulation for 19 promising candidates. 3 MOFs showed a significant discrepancy between results from the in-house GPU code and RASPA software due to the difference in pore blocking algorithms. While the in-house GPU code discerns inaccessible pores with an energy grid based flood fill algorithm, inaccessible pores were excluded by geometry based Monte Carlo integration

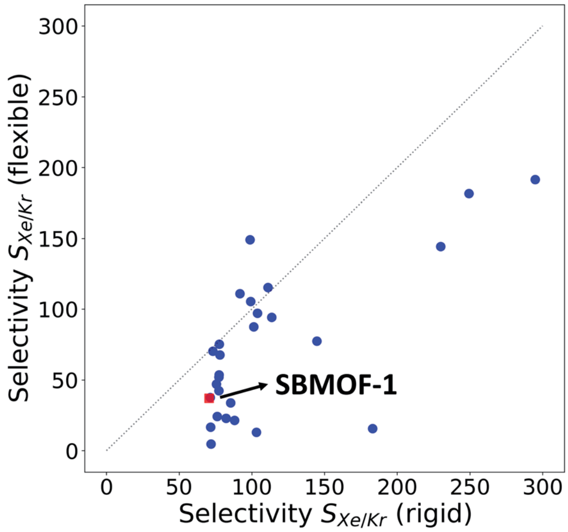

Fig. 3 Selectivity values for both flexible and rigid simulations for 27 candidate MOFs (blue circles) compared against the record holder of SBMOF-1 (red square).

during RASPA simulations. Therefore, we eliminate those 3 MOFs from our candidates.

Next, energy values were computed for 16 candidate materials to ascertain their experimental synthesizability. Given that the topology of these candidate materials might not be a suitable one for synthesis, polymorphic structures were generated for all of the materials and their energies were computed. Afterwards, two candidate MOFs with low energy values amongst their polymorphs were selected. Both of them respectively showed the third lowest energy among the ten polymorphs (see ESI Tables S2 and S3†). The two candidates are "htp+N5+N270+E0" ($S_{\mathrm{Xe/Kr}}$ (rigid): 100, $S_{\mathrm{Xe/Kr}}$ (flex): 97, $K_{\mathrm{H,Xe}}$: 1.6 × 10^{-1} mol kg^{-1} Pa^{-1}), and "htp+N5+N92+E0" ($S_{\mathrm{Xe/Kr}}$ (rigid): 78, $S_{\mathrm{Xe/Kr}}$ (flex): 75, $K_{\mathrm{H,Xe}}$: 5.8 × 10^{-2} mol kg^{-1} Pa^{-1}) (Fig. 4). The isosteric heat of adsorption (Q_{st}) was also obtained from the Widom insertion method to evaluate the strength of guest–host interactions for the final two candidates, which show that both candidate materials have higher Q_{st} for xenon than that for krypton (see Table S1 of ESI†). "htp+N5+N270+E0" and "htp+N5+N92+E0" consist of the gallium metal cluster (N270) and indium metal cluster (N92) respectively, but they share the same organic linker (BHC linker,[49] N5). Further details on the building blocks and the adsorption isotherm results for the two final candidates are shown in Section S2 of ESI.† Thus, the final candidates display significantly higher (roughly 1.5 times) theoretical selectivity than SBMOF-1.

One interesting point is that there are similarities between the final candidate materials and MIL-116. MIL-116 consists of two carboxylate arms, mellitate linker and octahedral metal cluster (Al, Ga and In).[50] However, this particular MOF was not generated because the functionalized mellitate linker was not included in our database. Thus, we should keep in mind that the unexpected structures can be generated as main clusters during synthesis of the candidates.

User-desired frameworks with specific values of properties

Most of the prior studies regarding inverse design tend to target generating materials with optimal properties. However, it is interesting to note whether our workflow can yield materials with more general user-desired property values, as the precise target of material properties might be very important when it comes to specific applications. As such, the cost function of our genetic algorithm was adjusted to target generation of MOFs with two user-desired selectivity values of 5 and 10 (see Section S3-1, ESI†). Compared to the randomly generated MOFs (up to 63 916 MOFs from the initial training set), there is a marked shift in the distributions of the 893 ($S_{\mathrm{Xe/Kr}} = 5$) and 996 ($S_{\mathrm{Xe/Kr}} = 10$) generated user-desired MOFs (Fig. 5a). It can be seen that the peaks of the distributions align closely with the user-desired target of selectivity values of 5 (turquoise, Fig. 5a) and 10 (magenta, Fig. 5a). It has to be pointed out that the targeting selectivity value of 5 leads to a more accurate distribution compared to that of 10 as the prediction capability of the machine learning model seems to decrease with higher selectivity values due to the ever-decreasing number of samples in higher selectivity regions.

Next, the cost function was expanded to incorporate a different property (*i.e.* $K_{\mathrm{H,Xe}}$). Similar to selectivity, it is possible to generate frameworks with user-desired $K_{\mathrm{H,Xe}}$ values if a machine learning model can make accurate predictions on these and the cost function is appropriately adjusted (see Section S3-1, ESI†). Fig. 5b shows the distribution of $K_{\mathrm{H,Xe}}$ for randomly generated MOFs as well as those of the user-desired targets of 10^{-5} mol kg^{-1} Pa^{-1}, 10^{-4} mol kg^{-1} Pa^{-1}, 10^{-3} mol kg^{-1} Pa^{-1}, and 10^{-2} mol kg^{-1} Pa^{-1}. 994, 954, 842, and 835 MOFs were respectively generated for the aforementioned user-desired $K_{\mathrm{H,Xe}}$. As can be seen, the peaks of the distributions follow the target well for 10^{-5} mol kg^{-1} Pa^{-1} and 10^{-4} mol kg^{-1} Pa^{-1}, but not for higher K_{H} values. This is similar to what we

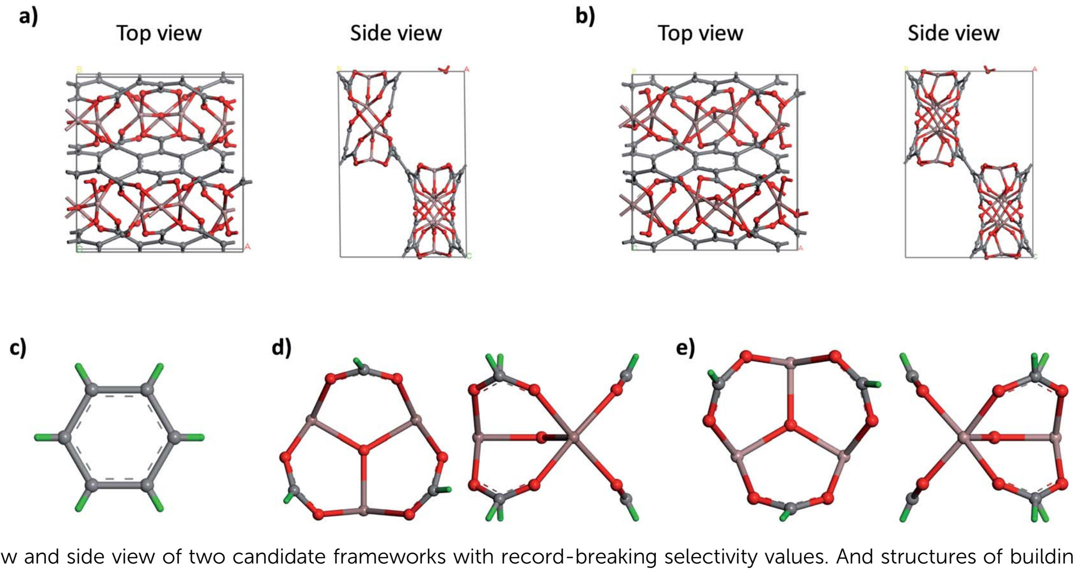

Fig. 4 Top view and side view of two candidate frameworks with record-breaking selectivity values. And structures of building blocks which were used to construct top 2 frameworks. (a) htp+N5+N270+E0. (b) htp+N5+N92+E0. (c) N5. (d) N92. (e) N270. For (c), (d), and (e), connection points are colored as light green for discrimination.

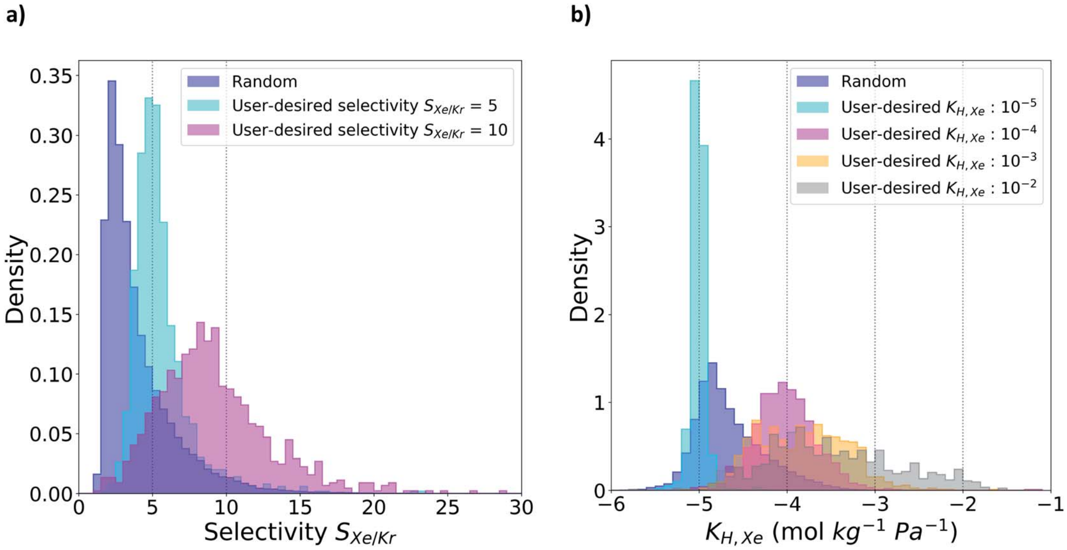

Fig. 5 Distributions of randomly generated MOFs *versus* user-desired generated MOFs. The axis of $K_{H,Xe}$ was converted to log scale for convenience. (a) Density *versus* selectivity. (b) Density *versus* $K_{H,Xe}$.

observed in the case for selectivity and has to do with a dearth of structures in these extreme value regions. With that said, it can be seen from the tails of the distributions that targeting higher values of K_H leads to more MOFs with higher K_H being generated.

Finally, the two aforementioned properties (*i.e.* selectivity and K_H) were combined in the cost function for simultaneous targeting. Specifically, in our cost function, the values of the two properties were standardized to provide equal weights to them (see Section S3-2, ESI†). Considering previous results, "$S_{Xe/Kr} = 10$, $K_{H,Xe} = 10^{-4}$ mol kg^{-1} Pa^{-1}", "$S_{Xe/Kr} = 5$, $K_{H,Xe} = 10^{-4}$ mol kg^{-1} Pa^{-1}", and "$S_{Xe/Kr} = 5$, $K_{H,Xe} = 10^{-5}$ mol kg^{-1} Pa^{-1}" were selected as target conditions. For each case, 1,816, 986, and 984 MOFs were respectively generated. Fig. 6 indicates that the cost function can be adjusted to produce MOFs that simultaneously possess user-desired properties along both property dimensions. Specifically, while only 4.56%, 1.32%, and 0.85% of randomly generated MOFs are within ±30% of target values of the properties, these proportions increase to 24.56%, 12.17%, and 15.04% respectively for MOFs with user-desired cost functions (see ESI Fig. S7–9†).

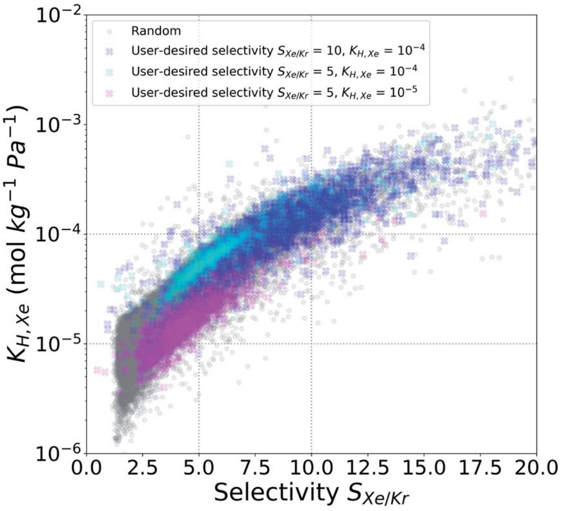

Fig. 6 Performance of the revised genetic algorithm to optimize frameworks with two properties (selectivity and $K_{H,Xe}$). The opaqueness of colors of a specific area increases with the frequency of frameworks in the area.

User-desired frameworks with minimum selectivity

It is often the case in real-life applications that there is a certain threshold property value that the material should exceed and anything beyond that becomes less important. When the material is "good enough" across one property dimension, then other factors such as materials cost or chemical/thermal stability can become more important. To demonstrate this capacity, our cost function was adjusted to discover materials which satisfy an arbitrary condition from the materials space. Specifically, we contrived a new cost function that can take on a minimum selectivity value (see Section S3-3, ESI†) of 5, 10, 15, and 20. Overall, 956, 694, 646, and 736 user-desired MOFs were respectively generated for the aforementioned sets. For each of these cases, the performances for the random set of MOFs (magenta), targeted selectivity set (*i.e.* cost function from the previous section, turquoise), and the minimum selectivity set (blue) were compared. It can be seen from Fig. 7 that 84% and 49.6% of MOFs extracted from the new cost function of $S_{Xe/Kr} > 5$ and $S_{Xe/Kr} > 10$ indeed have higher selectivity than 5 and 10, respectively. These values are much higher than that of both the targeted cost function and randomly generated MOF sets. However, this performance reduces for higher selectivity values of 15 and 20 (see Fig. 7 and ESI Fig. S10†). This degradation of performance can be attributed to the low prediction accuracy of the machine learning model in the high selectivity regime and

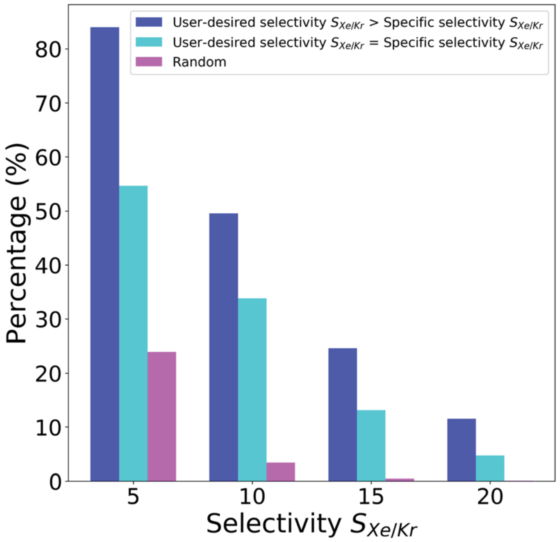

Fig. 7 Percentage of frameworks with higher values than the specified selectivity values of 5, 10, 15, and 20. Blue histograms denote the results of the new cost function. Turquoise histograms denote the results of the cost function to generate frameworks with specific selectivity (*i.e.* cost function from the previous section). Magenta histograms denote the results of randomly generated frameworks (initial training set).

scarcity of frameworks with high selectivity within the MOF space. Nonetheless, this demonstrates the power of our workflow that can adjust the cost function to cater towards the various demands from inverse design.

Conclusion

In this work, we integrated a genetic algorithm and a deep neural network to discover MOF candidates with record-breaking Xe/Kr selectivity and demonstrated that our tool can be used to efficiently explore the vast MOF materials space. Specifically, two MOF candidates possessed significantly higher selectivity values (both in flexible and rigid simulations) compared to SBMOF-1. But beyond just targeting MOFs with the best material properties, we successfully expanded our platform to incorporate fine-tuned targeting of user-desired properties. As such, one can simply modify the cost function across multiple properties and provide significant benefits in materials design that allow users to consider several conditions simultaneously. We believe that this type of design strategy will become more popularized and we surmise that our workflow/methodology will work seamlessly with other porous materials and other applications as well.

Materials and methods

Construction of hypothetical MOFs (hMOFs)

Both the top-down and the bottom-up approaches have been used in the past to construct hypothetical MOFs.[19,51] In this work, we used our in-house developed top-down MOF generator called, PORMAKE,[27] to construct hypothetical MOFs. By using the topology-based MOF construction method, a high degree of structural diversity can be guaranteed.

The database of PORMAKE consists of three components: (1) node building block (NBB), (2) edge building block (EBB), and (3) topology. NBBs are clusters that have more than two connection points, while EBBs are clusters with only two connection points (where the connection point is defined as an abstract point within the chemical structure that can be connected to another building block with the conjunction between two connection points treated as a single bond). Topologies contain information regarding the spatial positions of each of the NBBs (with the EBBs connecting each of the NBBs). The details of PORMAKE can be found in Lee *et al.*[27]

In this work, 719 NBBs, 234 EBBs, and 1775 topologies were used as components, and materials with very large unit cells (*i.e.* cell lengths > 60 Angstrom or # of atoms/unit cell > 1500) were disregarded due to high computational costs. The NBBs and EBBs were taken from ToBaCCo and CoRE MOF databases,[15,51] and the topologies were taken from the RCSR database.[52]

Integration of machine learning and a genetic algorithm

A machine learning model was created to predict specific properties and to reduce the computational cost of running a large number of molecular simulations. We used the machine learning model, MOF-NET,[27] which can predict Xe/Kr selectivity values directly from the MOFs that are represented as "topology + NBBs + EBBs" in the neural network (*e.g.* one of the well-known MOFs, HKUST-1, is represented as "tbo+N10+N409+E0"). Each topology and building block are mapped into their unique integral indices, which is very similar to word tokenization in natural language processing (NLP). Topologies, NBBs, and EBBs have different sets of indices, respectively. MOF-NET adopts the concept of the word embedding technique, Word2Vec,[53] to analyze relationships among topologies and building blocks. Due to the embedding technique, the architecture of MOF-NET can be elaborately designed to consider the effects of not only individual building blocks (NBBs and EBBs) but also interactions between the building blocks.

As an optimization algorithm for inverse design, we used the Multi-Species Genetic Algorithm (MSGA) developed by Lee *et al.*[27] For each of the 1775 topologies, independently parallel simulations were conducted using the NBBs and the EBBs (regarded as chromosomes) as inputs to the MSGA. Similar to other genetic algorithm implementations, crossover occurs between two chromosomes (*i.e.* NBBs and EBBs) while mutations of these chromosomes are possible. The cost function of the MSGA can be changed to find both the MOFs with the highest Xe/Kr selectivity value as well as user-desired Xe/Kr selectivity values and xenon adsorption property values (see Section S3, ESI†).

As shown in Fig. 1, the computational workflow can be divided into two parts. The first part consists of the machine learning model that takes in MOFs as inputs (in the form of NBBs, EBBs, and topologies) and Xe/Kr selectivity as outputs. For the training set purposes, selectivity values of the generated

MOFs were obtained using Monte Carlo (MC) simulations. The second part consists of running the MSGA and generating new hypothetical MOFs with user-desired conditions. The cost function of MSGA was changed to generate optimal MOFs. The trained weights of MOF-NET were used to screen the expansive MOF space efficiently during the MSGA.

Molecular simulation details

In this work, Xe/Kr selectivity under dilute conditions was selected as the target property. Given the dilute condition, Xe/Kr selectivity was calculated as the ratio of the xenon and krypton Henry coefficients (eqn (1)). We calculated the Henry coefficients by using the Widom insertion method *via* in-house GPU simulation. Parameters for the force field were taken from Boato *et al.* and the universal force field (UFF) was used to represent the MOF atoms (see Section S6 of ESI†).[54,55] The Lorentz–Berthelot mixing rule was applied to calculate parameters between different atomic species. All of the simulations were conducted at $T = 298$ K and $P = 1$ bar.

$$\text{Xe/Kr selectivity} = \frac{\text{Xenon Henry coefficient }(K_{\text{H,Xe}})}{\text{Krypton Henry coefficient }(K_{\text{H,Kr}})} \quad (1)$$

In order to reduce the computational cost, most of the Henry coefficient calculations were held by an in-house GPU code with the blocking algorithm.[56] For longer simulations for the few selective high-performance MOFs, the RASPA software was used.[57] During RASPA simulations, Zeo++ software was used to identify inaccessible pores and block them.[58] The Forcite module of Material studio[59] was used to conduct geometric optimization for the MOFs prior to the MC simulations. For further structural analysis for high-performance MOFs, Zeo++ software was used.[58]

Also, it has been demonstrated in the past that more reliable results for Xe/Kr selectivity can be obtained by considering the MOFs to be flexible (especially for frameworks with high selectivity).[47] To incorporate flexibility, the "flexible snapshots" method, which was devised by Gee *et al.*, was used to obtain accurate selectivity values for high-performing frameworks.[60,61] This method entails taking snapshots from a trajectory of molecular dynamics (MD) simulations and running MC simulations on each of the snapshots. The LAMMPS software package[62] was used to generate the trajectory and the MD simulation was conducted with the UFF-fix-metal (UFF-FM) force field within the NVT ensemble. The structures were equilibrated for 30 ps and the production run lasted for 30 ps. Six snapshots (for every 5 ps) were obtained and the resulting averaged Henry coefficient was calculated to account for the framework flexibility.

Polymorph finder

Synthesizability is one major issue for hypothetical MOFs and one key aspect for synthesizability is the generation of polymorphic structures. In this context, polymorphic MOFs refer to structures that possess the same chemical compositions with different topologies.[63] Among many polymorphic structures, the structure with the lowest energy is likely to be synthesized experimentally.[64]

And as such, we developed an algorithm to generate all polymorphic MOF structures of a given reference framework with the idea that structures with high Xe/Kr selectivity values but relatively high energy would be disregarded at the end. The algorithm consists of two parts: identifying connectivity (connection information of node building blocks) and comparing the ratio of combinations of building blocks (NBB-EBB-NBB) in a framework. Every polymorphic structure should have the same connectivity distribution and the ratio of combinations of building blocks (see Section S5 of ESI for details†). After the generation, the energy for each polymorph was computed by the Forcite module in Material Studio[59] *via* geometric optimization. During geometric optimization, the force of every atom in structure is calculated from the potential energy expression, allowing the atoms to move until their overall forces become smaller than defined convergence tolerance. Thus, the structure with a minimum potential energy surface can be obtained after the geometric optimization. Lastly, the computed energy was divided by the number of metal atoms for proper normalization.

Author contributions

Y. L. and J. K. wrote the manuscript. Y. L. conducted overall computational studies involving machine learning, genetic algorithm, and molecular simulation. Y. L. and J. P. developed a polymorph finder algorithm and analyzed the synthesizability of hypothetical structures. J. K. formulated the project. All authors contributed to exchanging ideas and discussed on this work.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT, & Future Planning under grant no. 2021R1A2C2003583.

Notes and references

1 K. M. Jablonka, D. Ongari, S. M. Moosavi and B. Smit, *Chem. Rev.*, 2020, **120**, 8066–8129.

2 S. Chong, S. Lee, B. Kim and J. Kim, *Coord. Chem. Rev.*, 2020, **423**, 213487.

3 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.

4 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.

5 L. Sarkisov and J. Kim, *Chem. Eng. Sci.*, 2015, **121**, 322–330.

6 K. Takahashi and Y. Tanaka, *Dalton Trans.*, 2016, **45**, 10497–10499.

7 R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi and C. Kim, *npj Comput. Mater.*, 2017, **3**, 54.

8 E. Kim, K. Huang, S. Jegelka and E. Olivetti, *npj Comput. Mater.*, 2017, **3**, 53.

9 C. W. Coley, L. Rogers, W. H. Green and K. F. Jensen, *J. Chem. Inf. Model.*, 2018, **58**, 252–261.

10 O. M. Yaghi, M. O'Keeffe, N. W. Ockwig, H. K. Chae, M. Eddaoudi and J. Kim, *Nature*, 2003, **423**, 705–714.

11 J.-R. Li, R. J. Kuppler and H.-C. Zhou, *Chem. Soc. Rev.*, 2009, **38**, 1477–1504.

12 R. E. Morris and P. S. Wheatley, *Angew. Chem., Int. Ed.*, 2008, **47**, 4966–4981.

13 J. Lee, O. K. Farha, J. Roberts, K. A. Scheidt, S. T. Nguyen and J. T. Hupp, *Chem. Soc. Rev.*, 2009, **38**, 1450–1459.

14 L. E. Kreno, K. Leong, O. K. Farha, M. Allendorf, R. P. Van Duyne and J. T. Hupp, *Chem. Rev.*, 2012, **112**, 1105–1125.

15 Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp, B. Slater, J. I. Siepmann, D. S. Sholl and R. Q. Snurr, *J. Chem. Eng. Data*, 2019, **64**, 5985–5998.

16 G. Avci, I. Erucar and S. Keskin, *ACS Appl. Mater. Interfaces*, 2020, **12**, 41567–41579.

17 T. D. Burns, K. N. Pai, S. G. Subraveti, S. P. Collins, M. Krykunov, A. Rajendran and T. K. Woo, *Environ. Sci. Technol.*, 2020, **54**, 4536–4544.

18 Z. Qiao, Q. Xu and J. Jiang, *J. Mater. Chem. A*, 2018, **6**, 18898–18905.

19 C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp and R. Q. Snurr, *Nat. Chem.*, 2012, **4**, 83–89.

20 Y. G. Chung, D. A. Gómez-Gualdrón, P. Li, K. T. Leperi, P. Deria, H. Zhang, N. A. Vermeulen, J. F. Stoddart, F. You, J. T. Hupp, O. K. Farha and R. Q. Snurr, *Sci. Adv.*, 2016, **2**, e1600909.

21 Y. Bao, R. L. Martin, C. M. Simon, M. Haranczyk, B. Smit and M. W. Deem, *J. Phys. Chem. C*, 2015, **119**, 186–195.

22 B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360–365.

23 A. Zunger, *Nat. Rev. Chem.*, 2018, **2**, 0121.

24 J. Noh, G. H. Gu, S. Kim and Y. Jung, *Chem. Sci.*, 2020, **11**, 4871–4881.

25 B. Kim, S. Lee and J. Kim, *Sci. Adv.*, 2020, **6**, eaax9324.

26 Z. Yao, B. Sánchez-Lengeling, N. S. Bobbitt, B. J. Bucior, S. G. H. Kumar, S. P. Collins, T. Burns, T. K. Woo, O. K. Farha, R. Q. Snurr and A. Aspuru-Guzik, *Nat. Mach. Intell.*, 2021, **3**, 76–86.

27 S. Lee, B. Kim, H. Cho, H. Lee, S. Y. Lee, E. S. Cho and J. Kim, *ACS Appl. Mater. Interfaces*, 2021, **13**, 23647–23654.

28 D. Liu, Y. Tan, E. Khoram and Z. Yu, *ACS Photonics*, 2018, **5**, 1365–1369.

29 S. So, J. Mun and J. Rho, *ACS Appl. Mater. Interfaces*, 2019, **11**, 24264–24268.

30 H. J. Xiang, B. Huang, E. Kan, S.-H. Wei and X. G. Gong, *Phys. Rev. Lett.*, 2013, **110**, 118702.

31 Y.-Y. Zhang, W. Gao, S. Chen, H. Xiang and X.-G. Gong, *Comput. Mater. Sci.*, 2015, **98**, 51–55.

32 R. Dong, Y. Dan, X. Li and J. Hu, *Comput. Mater. Sci.*, 2020, 110166, DOI: 10.1016/j.commatsci.2020.

33 M. Mahfouf, M. Jamei and D. A. Linkens, *Mater. Manuf. Processes*, 2005, **20**, 553–567.

34 A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman and R. Ramprasad, *Sci. Rep.*, 2016, **6**, 20952.

35 M. Zhou, A. Vassallo and J. Wu, *J. Membr. Sci.*, 2020, **598**, 117675.

36 S. Lee, B. Kim and J. Kim, *J. Mater. Chem. A*, 2019, **7**, 2709–2716.

37 I. G. Tapeinos, A. Miaris, P. Mitschang and N. D. Alexopoulos, *Compos. Sci. Technol.*, 2012, **72**, 774–787.

38 S. LeBlanc, S. K. Yee, M. L. Scullin, C. Dames and K. E. Goodson, *Renewable Sustainable Energy Rev.*, 2014, **32**, 313–327.

39 M. Ding, X. Cai and H.-L. Jiang, *Chem. Sci.*, 2019, **10**, 10209–10230.

40 J. B. West and J. Morton, *At. Data Nucl. Data Tables*, 1978, **22**, 103–107.

41 E. R. Ault, R. S. Bradford Jr and M. L. Bhaumik, *Appl. Phys. Lett.*, 1975, **27**, 413–415.

42 C. Lynch, J. Baum, R. Tenbrinck and R. B. Weiskopf, *Anesthesiology*, 2000, **92**, 865–870.

43 S. French and M. Novotny, *Anal. Chem.*, 1986, **58**, 164–166.

44 D. Banerjee, C. M. Simon, A. M. Plonka, R. K. Motkuri, J. Liu, X. Chen, B. Smit, J. B. Parise, M. Haranczyk and P. K. Thallapally, *Nat. Commun.*, 2016, **7**, 11831.

45 C. M. Simon, R. Mercado, S. K. Schnell, B. Smit and M. Haranczyk, *Chem. Mater.*, 2015, **27**, 4459–4475.

46 D. Banerjee, C. M. Simon, S. K. Elsaidi, M. Haranczyk and P. K. Thallapally, *Chem*, 2018, **4**, 466–494.

47 M. Witman, S. Ling, S. Jawahery, P. G. Boyd, M. Haranczyk, B. Slater and B. Smit, *J. Am. Chem. Soc.*, 2017, **139**, 5547–5557.

48 P. G. Boyd, S. M. Moosavi, M. Witman and B. Smit, *J. Phys. Chem. Lett.*, 2017, **8**, 357–363.

49 K. M. L. Taylor, A. Jin and W. Lin, *Angew. Chem., Int. Ed.*, 2008, **47**, 7722.

50 C. Volkringer, T. Loiseau, N. Guillou, G. Férey, D. Popov, M. Burghammer and C. Riekel, *Solid State Sci.*, 2013, **26**, 38–44.

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

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

53 M. Tomas, C. Kai, C. Greg and D. Jeffrey, 2013, arXiv:1301.3781.

54 G. Boato and G. Casanova, *Physica*, 1961, **27**, 571–589.

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

56 J. Kim, R. L. Martin, O. Rübel, M. Haranczyk and B. Smit, *J. Chem. Theory Comput.*, 2012, **8**, 1684–1693.

57 D. Dubbeldam, S. Calero, D. E. Ellis and R. Q. Snurr, *Mol. Simul.*, 2016, **42**, 81–101.

58 T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza and M. Haranczyk, *Microporous Mesoporous Mater.*, 2012, **149**, 134–141.

59 BIOVIA, *Material Studio, 2019*, Dassault Systèmes, San Diego, 2019.

60 J. A. Gee and D. S. Sholl, *J. Phys. Chem. C*, 2016, **120**, 370–376.

61 M. Agrawal and D. S. Sholl, *ACS Appl. Mater. Interfaces*, 2019, **11**, 31060–31068.

62 S. Plimpton, *J. Comput. Phys.*, 1995, **117**, 1–19.

63 D. Aulakh, J. R. Varghese and M. Wriedt, *Inorg. Chem.*, 2015, **54**, 8679–8684.

64 R. Anderson and D. A. Gómez-Gualdrón, *Chem. Mater.*, 2020, **32**, 8106–8119.