Accelerated Discovery of High-Performance Polyamines for Solid-State Direct CO_2 Capture via Efficient Simulations and Bayesian Optimization

Junhe Chen[1*] **A N M Nafiz Abeer**[2*] **Alif Bin Abdul Qayyum**[2] **Zhihao Feng**[1]
Hyun-Myung Woo[3] **Byung-Jun Yoon**[2,4] **Seung Soon Jang**[1]

[1]Georgia Institute of Technology [2]Texas A&M University
[3]Incheon National University [4]Brookhaven National Laboratory

{junhechen, zfeng77}@gatech.edu
{nafiz.abeer, alifbinabdulqayyum, bjyoon}@tamu.edu
hmwoo@inu.ac.kr, seungsoon.jang@mse.gatech.edu

Abstract

Solid amine-based sorbents are a leading approach for direct air capture (DAC) of CO_2, owing to their energy efficiency and scalability. To enable data-driven discovery of improved sorbents, we developed a computational framework that integrates fragment-based polymer generation with Density Functional Theory (DFT), molecular dynamics (MD) relaxations, and grand canonical Monte Carlo (GCMC) sampling. This workflow provides accurate yet efficient estimates of CO_2 uptake while capturing key structure-property relationships across a diverse library of polymers assembled from well-characterized polyamines for DAC. Leveraging such adsorption data, we investigated the application of the Bayesian optimization (BO) strategy in accelerating the discovery process of high-performing polymer candidates with our developed simulation workflow. Computational experimental results demonstrated the sensitivity of this discovery process to the choice of molecular representation in the surrogate models of BO, especially in a small computational budget scenario, where polymer-specific pre-training provided an early advantage over models trained for general chemical space.

1 Introduction

Mitigating the adverse impact of rising atmospheric CO_2 [1, 2] necessitates the deployment of Carbon Capture and Storage (CCS) technologies, among which direct air capture (DAC) has emerged as a principal strategy. Because of the low atmospheric concentration of CO_2, considerable research has been devoted to the development of highly selective DAC sorbents. Conventional liquid-phase systems, although effective, are hindered by high thermal regeneration requirements[3, 4]. In contrast, solid adsorption [5–7] offers a more energy-efficient and scalable pathway. Within this domain, branched polyamines represent a particularly promising class [8, 9]. Their strong chemical affinity for CO_2, driven by the accessibility of amine functional groups, positions them as key candidates for next-generation sorbents capable of operating under ultradilute CO_2 conditions [10]. However, given the current stage of experimental work, further systematic studies will be important to clarify their behavior and guide the development of improved amine-based sorbents.

*Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Accelerated Materials Discovery (AI4Mat).

Machine learning-assisted design workflows have seen diverse applications in material discovery – ranging from inverse design paradigm [11] to self-driving laboratories (SDLs) [12] – enabling efficient exploration for materials meeting the design expectations. The field of polymer engineering is no exception to this; especially due to the high-impact application areas such as gas separation membranes, energy storage, fuel cells etc., there has been an increasing trend [13] to leverage machine learning techniques to accelerate the computational design workflow for polymer design. For instance, this includes efforts on molecular representation learning [14–16] for the reliable prediction of polymer characteristics (e.g., glass transition temperature, band gap etc.) as well as generative design approaches like [17] for polymer membranes in post-combustion carbon capture. Our work represents an effort toward incorporating the atomistic simulation, which is currently employed for gaining molecular-level insights behind the sorption mechanism [18], into the polymer sorbent design workflows for direct air capture of CO_2. Specifically, we developed an efficient simulation of the polymer's CO_2 adsorption mechanism, and generated adsorption data for a library of hypothetical (amine-based) polymers, enabling machine learning models of structure-property relationships required for accelerated identification of high-performing polymers. The key aspects of our work are as follows:

- We generated a dataset of 1,000 polymers and estimated their CO_2 adsorption capacity with our computationally efficient simulation – combining fast MD relaxations with grand canonical Monte Carlo (GCMC).
- We assessed the impact of molecular representation and surrogate models in Bayesian optimization for screening of polymers with high adsorption capacity.

2 Methodology

2.1 Dataset Generation

Our study began with nine well-characterized polyamines (e.g., polyethyleneimine (PEI), polypropyleneimine (PPI) etc.) commonly investigated for direct CO_2 capture due to high amine density and adsorption performance. These reference polymers were fragmented into chemically meaningful building blocks using BRICS[19] and RECAP[20] methods in RDKit[21], ensuring synthetically relevant bonds were cleaved while preserving functional group integrity. Building blocks were then recombined into a large ensemble of hypothetical linear polyamine structures, tailored to emulate realistic polymer backbones with varied chain lengths and terminal functionalities. All generated molecules underwent rigorous validity screening using RDKit. Improper sequences, structures or terminal atoms were discarded. Canonicalization and deduplication ensured that the final database comprised only chemically plausible and unique SMILES strings. This process resulted in a structurally diverse library poised for adsorption modeling.

Each valid polymer was converted into a 3D geometry, energy-minimized, and formatted for batch simulation in molecular dynamics (MD) and grand canonical Monte Carlo (GCMC) workflows (**Appendix** A.2). Fast MD was applied to relax polymer conformations and approximate packing behavior, while subsequent GCMC simulations in LAMMPS provided CO_2 uptake estimates under defined pressure and temperature conditions. Note that we refined the force field parameters (**Appendix** A.1) for accurate modeling of polymer-CO_2-H_2O interactions. The outputs of this computational pipeline (illustrated in Fig. 4) formed the foundational dataset of 1,000 polymers with estimated CO_2 adsorption capacity for downstream Bayesian optimization-based candidate screening.

2.2 Accelerating Discovery with Bayesian optimization

Our developed simulation pipeline for CO_2 adsorption capacity measurement (**Appendix** A.2) has an improved throughput via the combination of Fast MD and GCMC simulation. However, a naive approach to identifying the high-performing polymers by exhaustively simulating all polymer samples in a large pool leads to inefficient usage of computational resources. This work adopted the Bayesian optimization (BO) [22, 23] strategy to prioritize the polymer samples for simulation to accelerate the discovery process, i.e., to find the high-performing polymer samples within a fixed computational budget for simulations of CO_2 adsorption.

We initialize the BO-assisted discovery process with a small number (N_0) of polymer samples $x^{(i)}$ and their simulated adsorption capacity $y^{(i)} = f_{\text{sim}}(x^{(i)})$, where f_{sim} denotes the simulation process.

Our goal is to utilize such labeled data $\mathcal{D}_{\text{labeled}} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N_0}$ to identify the polymer sample with high adsorption from a pool of unlabeled data, $\mathcal{D}_{\text{unlabeled}} = \{(x^{(i)}\}_{i=1}^{N_{\text{pool}}}$ with N_{budget} number of simulations where $N_{\text{budget}} \ll N_{\text{pool}}$.

Using the labeled data $\mathcal{D}_{\text{labeled}}$, we first train a surrogate model $f_{\text{surrogate}} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ that predicts a distribution over the simulated CO_2 adsorption for the polymer sample $x \in \mathcal{X}$. We next obtain the predictive distribution for the samples in $\mathcal{D}_{\text{unlabeled}}$ using the trained surrogate model, and compute the utility of each unlabeled sample using an acquisition function. We perform the simulation for the unlabeled sample with the highest utility, and update our labeled data with this new sample and its observation, and remove the sample from the unlabeled data. This BO iteration – training of a surrogate with the labeled dataset, selection of an unlabeled sample to simulate, followed by updating the labeled and unlabeled dataset – is repeated for N_{budget} times. Algorithm 1 describes this iteration in detail. We have discussed the details of surrogate models and the acquisition functions in **Appendices** A.3 and A.4 respectively.

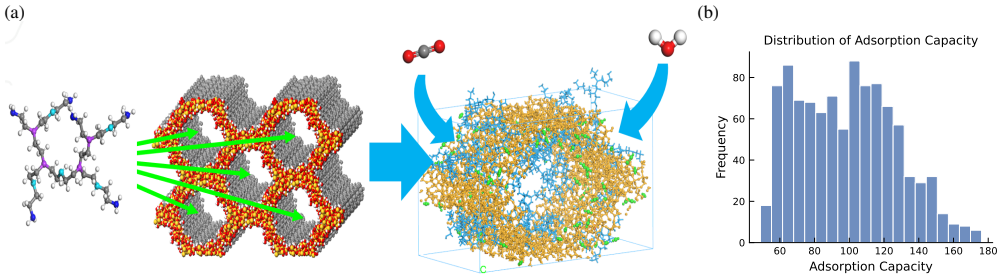

Figure 1: **(a) A model system of polyimine and a silica-based support material (MCM-41) loaded with CO_2 for GCMC and MD simulation. (b) Histogram of simulated adsorption capacity of** 1000 **polymer samples.**

3 Results and Discussion

3.1 Simulation Pipeline and Dataset

We refined the Lennard-Jones (LJ) parameters of CO_2-H_2O-amines interactions (**Appendix** A.1), resulting in accurate approximation of experimental energies (Fig. 5). We performed GCMC simulations (Fig. 1a, **Appendix** A.2) to estimate the CO_2 adsorption for polymer conformations generated by MD relaxation. Fig. 1b shows the histogram of the adsorption capacity of polymer samples.

3.2 Discovery of polymers with high CO_2 adsorption capacity

We investigated the Bayesian optimization-assisted accelerated discovery pipeline (Algorithm 1) with our generated dataset of 1000 polymer samples. Specifically, we analyzed the impact of different molecular representations in the surrogate models in different hypothetical scenarios of the initial labeled dataset. We considered 5 pre-trained models (polymer-specific: PolyNC [15], polyBERT [14] and general chemical space: MoLFormer [24], MolGen-large [25] and MiniMol [26]) and 3 choices

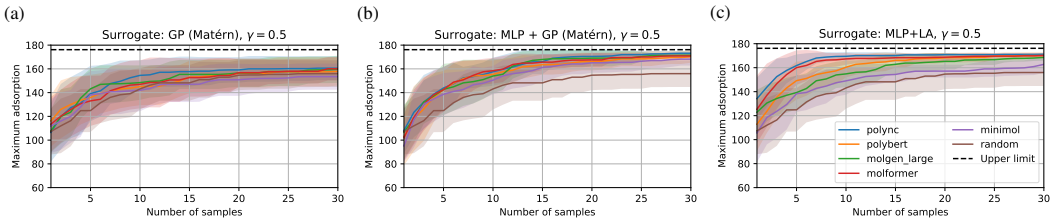

Figure 2: **Optimization traces for surrogate models: GP (a), GP-DKL (b), and MLP+LA (c) for** $\gamma = 0.5$**.** Each optimization traces show the progression of average (and ± 1 standard deviation over 20 trials) of the maximum adsorption capacity found by sequentially acquiring labels of unlabeled samples (in the x-axis). Along with 5 pre-trained models' embeddings, we have the random acquisition strategy as baseline. *Upper limit* denotes the maximum value in our dataset, $y_{\max}$.

of surrogate models: Gaussian process (GP with and without deep kernel learning) [27, 28] and Bayesian neural network (BNN) using Laplace approximation (LA) [29, 30]. For assessing the impact of initial $\mathcal{D}_{\text{labeled}}$, e.g., "*what if we start from samples with very low adsorption capacity?*", we randomly selected the initial set of labeled samples such that $y^{(i)} < \gamma y_{\text{max}}$ for $\gamma \in \{0.5, 0.3\}$ where y_{max} denotes the maximum adsorption capacity value in our dataset of 1000 samples. Each BO-assisted discovery process (repeated for 20 trials, with Thompson sampling [31] as acquisition function) starts with $N_0 = 10$ initial labeled data, and tries to find the sample with the best adsorption capacity from the remaining 990 unlabeled samples using $N_{\text{budget}} = 30$ simulations.

3.2.1 Impact of initial labeled data of moderate quality ($\gamma = 0.5$)

In Figs. 2a to 2c, we have showed the optimization result with $\gamma = 0.5$ for the GP, GP-DKL and MLP+LA-based surrogates with embedding from five pre-trained models. While the embedding from the PolyNC model showed slightly better discovery in early iterations (fewer than 15 query samples), there was overall minimal gain by using GP-based acquisition over the random strategy. With GP-DKL, we observed a significant improvement in acquiring high-performing polymer samples quickly compared to the random baseline. Furthermore, this surrogate showed lower sensitivity to embeddings from different pre-trained models, in contrast to the choice of MLP+LA-based surrogate (Fig. 2c). Specifically for the latter case, embeddings from PolyNC and MolFormer showed superior efficiency, i.e., acquiring samples with maximum adsorption capacity exceeding 160 (on average) within the first 5 queries of unlabeled polymer samples. However, the performance with embedding from MiniMol was marginally better than the random strategy. Since we have used same set of hyperparameters in training the surrogate (MLP+LA), such discrepancy may be attributed to the sensitivity of BNN-based surrogate model to its design choices [32].

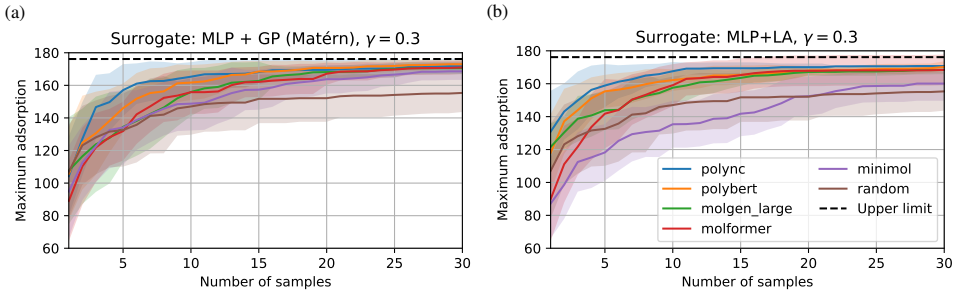

Figure 3: **Optimization traces for surrogate models: GP-DKL (a) and MLP+LA (b) for $\gamma = 0.3$.** The other settings of the experiments are same as in Fig. 2.

3.2.2 Impact of initial labeled data of poor quality ($\gamma = 0.3$)

With $\gamma = 0.3$, we effectively constrained the initial labeled dataset to contain samples with adsorption capacity lower than 53. In this scenario, both GP-DKL (Fig. 3a) and MLP+LA (Fig. 3b) based surrogates showed advantage of embeddings from polymer-specific pretrained models: PolyNC and polyBERT over other general chemical pre-trained models in accelerated discovery of high-quality samples. However, the differences in performance decreased as more labels were collected, except for the case with MiniMol in the MLP+LA-based surrogate, which showed worse performance than random acquisition in the first 20 queries. Nonetheless, this empirical findings highlight the importance of representation learning of polymer language models, specially in extremely unfavorable initialization. Furthermore, comparison with $\gamma = 0.5$ case indicates that the discovery rate became more sensitive to the choices of pre-trained models in the GP-DKL-based surrogate.

4 Conclusion

While MD simulation can effectively analyze the CO_2 adsoprtion mechanism, the associated computational cost constrains the use of MD simulation in the design process of amine-based polymer sorbents. In this work, we adopted a combination of fast MD with GCMC simulation as a rapid alternative to the lengthy MD simulation – providing improved throughput for estimating CO_2 adsorption capacity of the polymer. Our retrospective experiment with the adsoprtion data of 1000 linear polymer samples demonstrated that the BO strategy could further accelerate the discovery of high performing polymers

with limited number simulations. Importantly, it highlights the advantage of molecular representation from pre-trained polymer language models in quickly discovering polymers with high adsorption capacity, especially in cases where the initial labeled samples have low adsorption.

Since the adsorption estimates require validation by long-timescale MD simulations for high-confidence CO_2 uptakes, one can adopt a multi-fidelity Bayesian optimization approach [33] to discover high-quality polymers by leveraging both types of simulations. Exploration of the chemical space, e.g. branched and cross-linked architectures beyond linear polymers with de-novo design approaches, is another direction for future efforts. Specifically, this will also allow a more thorough investigation into the utility of different molecular representations in the polymer discovery process.

References

[1] Budget, g. c. fossil co2 emissions at record high in 2023 url = `https://globalcarbonbudget.org/fossil-co2-emissions-at-record-high-in-2023/.`, note = accessed: 2025-08-13 .

[2] Global carbon atlas url = `https://globalcarbonatlas.org/`, note = accessed: 2025-08-13 .

[3] H. Ahn, M. Luberti, Z. Liu, and S. Brandani. Process configuration studies of the amine capture process for coal-fired power plants. *International Journal of Greenhouse Gas Control*, 16:29–40, 2013.

[4] K. Goto, K. Yogo, and T. Higashii. A review of efficiency penalty in a coal-fired power plant with post-combustion co2 capture. *Applied Energy*, 111:710–720, 2013.

[5] R. V. Siriwardane, M.-S. Shen, E. P. Fisher, and J. A. Poston. Adsorption of co2 on molecular sieves and activated carbon. *Energy & Fuels*, 15(2):279–284, 2001.

[6] H. P. Huang, Y. Shi, W. Li, and S. G. Chang. Dual alkali approaches for the capture and separation of co2. *Energy & Fuels*, 15(2):263–268, 2001.

[7] K. I. Kim, R. Lawler, H. J. Moon, P. Narayanan, M. A. Sakwa-Novak, C. W. Jones, and S. S. Jang. Distribution and transport of co2 in hydrated hyperbranched poly(ethylenimine) membranes: A molecular dynamics simulation approach. *ACS Omega*, 6(4):3390–3398, 2021. PMID: 33553957.

[8] A. Samanta, A. Zhao, G. K. H. Shimizu, P. Sarkar, and R. Gupta. Post-combustion co2 capture using solid sorbents: A review. *Industrial & Engineering Chemistry Research*, 51(4):1438–1463, 2012.

[9] P. Sharma, S. Chakrabarty, S. Roy, and R. Kumar. Molecular view of co2 capture by polyethylenimine: Role of structural and dynamical heterogeneity. *Langmuir*, 34(17):5138–5148, 2018. PMID: 29641903.

[10] E. S. Sanz-Pérez, C. R. Murdock, S. A. Didas, and C. W. Jones. Direct capture of co2 from ambient air. *Chemical Reviews*, 116(19):11840–11876, 2016. PMID: 27560307.

[11] B. Sanchez-Lengeling and A. Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[12] G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson, M. Skreta, N. Yoshikawa, S. Corapi, G. D. Akkoc, F. Strieth-Kalthoff, M. Seifrid, and A. Aspuru-Guzik. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732, 2024. PMID: 39137296.

[13] H. Tran, R. Gurnani, C. Kim, G. Pilania, H.-K. Kwon, R. P. Lively, and R. Ramprasad. Design of functional and sustainable polymers assisted by artificial intelligence. *Nature Reviews Materials*, 9(12):866–886, 2024.

[14] C. Kuenneth and R. Ramprasad. polybert: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nature communications*, 14(1):4099, 2023.

[15] H. Qiu, L. Liu, X. Qiu, X. Dai, X. Ji, and Z.-Y. Sun. Polync: a natural and chemical language model for the prediction of unified polymer properties. *Chemical Science*, 15(2):534–544, 2024.

[16] F. Wang, W. Guo, M. Cheng, S. Yuan, H. Xu, and Z. Gao. Mmpolymer: A multimodal multitask pretraining framework for polymer property prediction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2336–2346, 2024.

[17] R. Giro, H. Hsu, A. Kishimoto, T. Hama, R. F. Neumann, B. Luan, S. Takeda, L. Hamada, and M. B. Steiner. Ai powered, automated discovery of polymer membranes for carbon capture. *npj Computational Materials*, 9(1):133, 2023.

[18] M. Robertson, J. Qian, and Z. Qiang. Polymer sorbent design for the direct air capture of co2. *ACS Applied Polymer Materials*, 6(23):14169–14189, 2024.

[19] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, and M. Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3(10):1503–1507, 2008.

[20] X. Q. Lewell, D. B. Judd, S. P. Watson, and M. M. Hann. Recap retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of chemical information and computer sciences*, 38(3):511–522, 1998.

[21] G. Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

[22] P. I. Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[23] A. Kristiadi, F. Strieth-Kalthoff, M. Skreta, P. Poupart, A. Aspuru-Guzik, and G. Pleiss. A sober look at llms for material discovery: are they actually good for bayesian optimization over molecules? In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[24] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

[25] Y. Fang, N. Zhang, Z. Chen, L. Guo, X. Fan, and H. Chen. Domain-agnostic molecular generation with chemical feedback. In *International Conference on Learning Representations*, 2024.

[26] K. Klaser, B. Banaszewski, S. Maddrell-Mander, C. McLean, L. Müller, A. Parviz, S. Huang, and A. W. Fitzgibbon. Minimol: A parameter-efficient foundation model for molecular learning. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*, 2024.

[27] C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[28] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Deep kernel learning. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 370–378, Cadiz, Spain, 09–11 May 2016. PMLR.

[29] D. J. C. Mackay. *Bayesian methods for adaptive models*. California Institute of Technology, 1992.

[30] E. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer, and P. Hennig. Laplace redux-effortless bayesian deep learning. *Advances in neural information processing systems*, 34:20089–20103, 2021.

[31] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[32] T. Cinquin, S. Lo, F. Strieth-Kalthoff, A. Aspuru-Guzik, G. Pleiss, R. Bamler, T. G. Rudner, V. Fortuin, and A. Kristiadi. What actually matters for materials discovery: Pitfalls and recommendations in bayesian optimization. In *AI for Accelerated Materials Design-ICLR 2025*, 2025.

[33] J. Song, Y. Chen, and Y. Yue. A general framework for multi-fidelity bayesian optimization with gaussian processes. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3158–3167. PMLR, 16–18 Apr 2019.

[34] S. L. Mayo, B. D. Olafson, and W. A. Goddard. Dreiding: a generic force field for molecular simulations. *The Journal of Physical Chemistry*, 94(26):8897–8909, 1990.

[35] M. Levitt, M. Hirshberg, R. Sharon, K. E. Laidig, and V. Daggett. Calibration and testing of a water model for simulation of the molecular dynamics of proteins and nucleic acids in solution. *The Journal of Physical Chemistry B*, 101(25):5051–5061, 1997.

[36] X. Shen, H. Du, R. H. Mullins, and R. R. Kommalapati. Polyethylenimine applications in carbon dioxide capture and separation: From theoretical study to experimental work. *Energy Technology*, 5(6):822–833, 2017.

[37] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[38] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 1015–1022, Madison, WI, USA, 2010. Omnipress.

[39] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

A Appendix

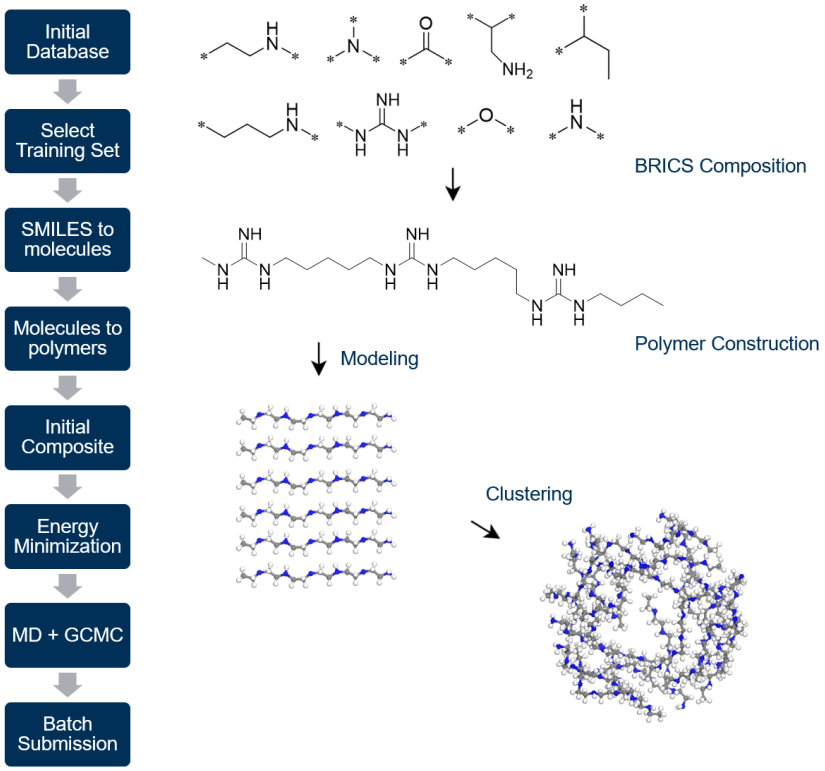

Figure 4: **Computational pipeline for polymer sorbent screening.** Molecular fragments are assembled into polymers via BRICS composition, deposited into porous matrices, and equilibrated. The resulting composites undergo molecular dynamics (MD) and grand canonical Monte Carlo (GCMC) simulations, enabling large-scale automated screening for CO_2 capture.

A.1 Force Fields for Polyamine-CO_2-H_2O Interactions

In this study, the DREIDING force field [34] and the F3C force field [35] were employed to model HB-PEI, HB-PPI, CO_2, and water. The DREIDING force field has been extensively applied in prior simulation studies across a wide range of materials, with results showing strong consistency with experimental data [36]. The functional form of the DREIDING potential is expressed as

$$E_{\text{total}} = E_{\text{vdW}} + E_{\text{Q}} + E_{\text{bond}} + E_{\text{angle}} + E_{\text{torsion}} + E_{\text{inversion}}, \tag{4}$$

where E_{total}, E_{vdW}, E_{Q}, E_{bond}, E_{angle}, E_{torsion}, and $E_{\text{inversion}}$ correspond to the total, van der Waals, electrostatic, bond stretching, angle bending, torsional, and inversion energies, respectively. Electrostatic contributions, E_{Q}, were determined from Mulliken population analysis.

To accurately capture the specific interactions between amine–CO_2, amine–H_2O, and CO_2–H_2O pairs, interaction energy profiles as a function of distance were first obtained using DFT at the B3LYP-D3/6-31G** level of theory (Fig. 5). From these calculations, Lennard-Jones (LJ) potential parameters (D

Algorithm 1 Bayesian Optimization for Accelerated Discovery

Require: Simulator for estimating CO_2 adsorption capacity $f_{\text{sim}}(x)$, initial dataset $\mathcal{D}^0_{\text{labeled}} = \mathcal{D}_{\text{labeled}} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N_0}$, initial pool $\mathcal{D}^0_{\text{unlabeled}} = \mathcal{D}_{\text{unlabeled}} = \{(x^{(i)}\}_{i=1}^{N_{\text{pool}}}$, acquisition function $\alpha(x)$, computational budget N_{budget}

1: **for** $k = 1$ to N_{budget} **do**

2: Train a surrogate model $f^k_{\text{surrogate}}$ with labeled dataset, $\mathcal{D}^{k-1}_{\text{labeled}}$

3: Select sample to simulate next:

$$x^{(k)} = \underset{x \in \mathcal{D}^{k-1}_{\text{unlabeled}}}{\arg\max}\, \alpha(x \mid f^k_{\text{surrogate}}) \tag{1}$$

4: Simulate for adsorption capacity: $y^{(k)} = f_{\text{sim}}(x^{(k)})$

5: Update datasets:

$$\mathcal{D}^k_{\text{labeled}} = \mathcal{D}^{k-1}_{\text{labeled}} \cup \{(x^{(k)}, y^{(k)})\} \tag{2}$$

$$\mathcal{D}^k_{\text{unlabeled}} = \mathcal{D}^{k-1}_{\text{unlabeled}} \setminus \{(x^{(k)})\} \tag{3}$$

6: **end for**

7: **Return:** $x^* = \arg\max_{(x^{(i)}, y^{(i)}) \in \mathcal{D}^{N_{\text{budget}}}_{\text{labeled}}} y^{(i)}$

and r_0) in Eq. (5) were optimized to describe the off-diagonal van der Waals interactions:

$$E_{\text{off-diagonal}}(r) = D\left[\left(\frac{r_0}{r}\right)^{12} - 2\left(\frac{r_0}{r}\right)^{6}\right], \tag{5}$$

where D is the potential well depth and r_0 is the equilibrium separation distance.

The binding energy between species A and B was calculated as

$$E_{\text{binding}} = E_{(A+B)} - E_A - E_B, \tag{6}$$

where $E_{(A+B)}$ denotes the total energy of the optimized molecular pair, and E_A and E_B are the corresponding isolated species. This formulation enables direct comparison of DFT-derived interaction energies with those predicted by conventional mixing rules. As illustrated in Fig. 5, the interaction curves based on the optimized LJ parameters show excellent agreement with DFT results, in contrast to the less accurate predictions of the geometric-mean mixing rule. Notably, the mean square error (MSE) was reduced by approximately 98% with the newly optimized parameters compared to those obtained from the mixing rule, demonstrating the robustness of this refinement. Consequently, the optimized LJ parameters were adopted for all subsequent MD simulations in this study.

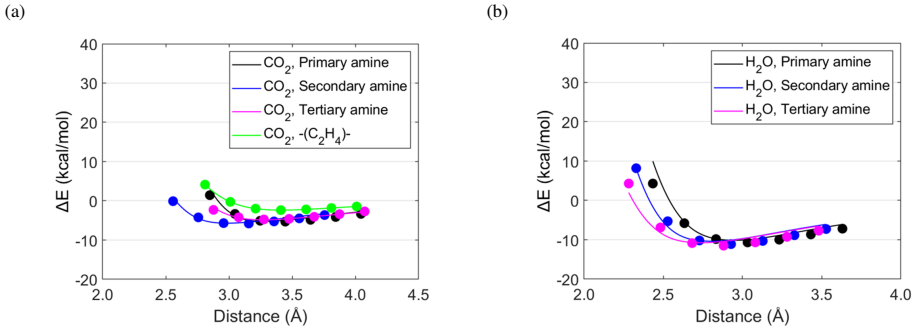

Figure 5: **DFT-calculated interaction energy curves for amine–CO_2 and amine–H_2O pairs.** (a) Binding energy profiles of CO_2 with primary, secondary, and tertiary amines, as well as with the –(C_2H_4)– linker group. (b) Binding energy profiles of H_2O with primary, secondary, and tertiary amines. Data points represent DFT calculations (B3LYP-D3/6-31G**), while lines correspond to Lennard-Jones (LJ) potential fits.

It is noted that i) among the binding energies of CO_2 with amines (-5.31 kcal/mol, -5.78 kcal/mol, and -4.75 kcal/mol, for primary, secondary, and tertiary amines, respectively), the CO_2-secondary amine pair has the strongest binding energy while the CO_2-tertiary amine pair has the weakest binding energy; ii) the binding energies of the CO_2-amine pairs are weaker than those of the H_2O-amine pairs (-10.67 kcal/mol, -11.10 kcal/mol, and -11.50 kcal/mol for primary, secondary, and tertiary amines, respectively).

A key outcome of the DFT-based parameterization is the capability to reproduce the observations from experimental and computational studies that amine-H_2O interactions are stronger than amine-CO_2 interactions, validating the suitability of the newly developed force field for investigating CO_2 capture in both HB-PEI and HB-PPI.

A.2 CO_2 Adsorption Estimation via Improved Throughput Simulation

For each polymer candidate in our dataset, we evaluated its CO_2 adsorption properties using a two-stage simulation workflow that combined fast molecular dynamics (MD) relaxation and grand canonical Monte Carlo (GCMC) sampling. The overall objective was to obtain accurate yet computationally efficient predictions of CO_2 capacity under specified capture conditions.

Fast MD simulations were first used to relax polymer configurations and approximate their packing density prior to adsorption calculations. Initial geometries were energy-minimized and equilibrated in the NVT ensemble at T = 298 K for 500 ps, followed by NPT ensemble simulations at P = 1 bar for 500 ps to achieve realistic bulk densities. This approach ensured that chain conformations, void spaces, and local environments were well-represented before introducing sorbate molecules. The relaxed polymer configurations served as input for subsequent GCMC adsorption simulations.

GCMC simulations were carried out using the LAMMPS / RASPA framework to directly estimate the equilibrium uptake of CO_2 at low partial pressures relevant to DAC. All simulations were performed at T = 298 K, with a CO_2 partial pressure of 0.4 mbar to mimic atmospheric capture conditions. The DREIDING + F3C force field parameters described in **Appendix** A.1 were used for polymer-CO_2-H_2O interactions. Each GCMC run consisted of 1 million trial moves, with an insertion-to-deletion ratio of 1:1 and displacement/rotation trials included to enhance sampling efficiency. The calculated adsorption capacity was expressed in mmol g^{-1}, averaged over the production phase of the simulation after equilibration. These results formed the quantitative performance labels for downstream machine learning and Bayesian optimization workflows.

A.3 Molecular Representation of Polymer and Surrogate Model

Molecular Embeddings: For building the surrogate model to predict the adsorption capacity of a polymer sample x, we have used the representation $\mathbf{z}$ encoded by the pre-trained models for molecule space as a numerical representation of sample x. In our work, we considered both polymer-specific and general molecular pre-trained models for encoding the polymer samples. The two polymer-specific pretrained models: PolyNC [15] and polyBERT [14], encode the PSMILES strings of polymer x into the embeddings of their corresponding chemical tokens. We obtained the representation $\mathbf{z}$ by taking the average over these chemical token embeddings, excluding the special tokens. For representation from the general molecular space, we computed the embedding $\mathbf{z}$ from pre-trained MoLFormer [24], MolGen-large [25], and MiniMol [26]. For these models, the "*" symbols in PSMILES (indicating the connection points in monomer) were removed, and the resulting monomer was encoded to an embedding $\mathbf{z}$. Note that MoLFormer and MolGen-large are chemical language models for SMILES and SELFIES representation respectively, and the embedding $\mathbf{z}$ is the token-wise average embedding of the monomer. For MiniMol, which is a pre-trained graph neural network (GNN), $\mathbf{z}$ is the max-pool aggregation of the node embeddings of the molecular graph of the monomer. Fig. 6 shows the UMAP [37] visualization of the 1000 polymer samples – with simulated CO_2 adsorption capacity – based on the embedding $\mathbf{z}$ from these pretrained models.

Surrogate Models: In our work, we have used Gaussian Process (GP) [27], GP with deep kernel learning (GP-DKL) [28], and Bayesian neural network with Laplace approximation (LA) [29, 30] as surrogate models mapping embedding $\mathbf{z}$ to the predictive distribution of the adsorption capacity of polymer samples. For GP, the prediction for sample x is $f(\mathbf{z}) \sim \mathcal{GP}(m, k)$ where $m : \mathcal{Z} \to \mathbb{R}$ is the mean function, and $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is the kernel function (e.g., Matérn-$5/2$ kernel used in our

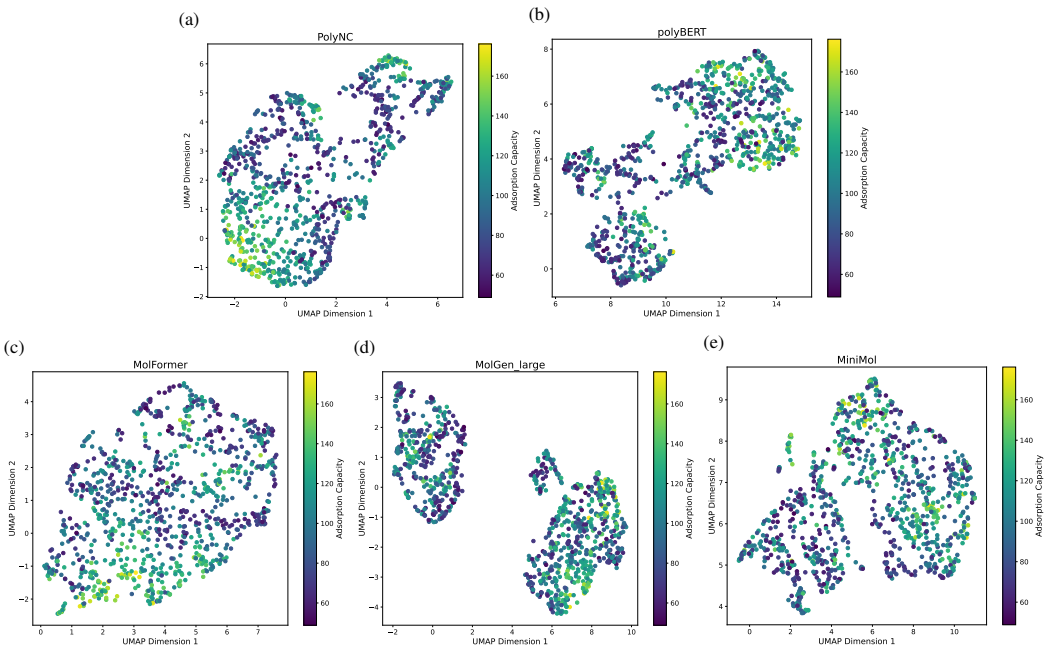

Figure 6: **UMAP visualization of 1000 polymer samples based on their representation from polymer-specific pre-trained models: PolyNC, polyBERT and general chemical pre-trained models: MolFormer, MolGen_large, MiniMol**

work) defined on the embedding $\mathbf{z}$ of the sample. In GP-DKL, the kernel $k(\mathbf{z}, \mathbf{z}')$ is replaced with $k(\phi(\mathbf{z}), \phi(\mathbf{z}'))$ where ϕ represents a non-linear mapping with learnable parameters which is jointly trained with the base GP. We have used a multilayer perceptron (MLP) network with one hidden layer and one output layer – both with 50 neurons and ReLU activation as the non-linear function ϕ in our experiment with GP-DKL. For LA, we have used the linearized Laplace approximation from [30], where the output of a neural network $f_{\boldsymbol{\theta}}(\mathbf{z})$ is linearized near a known model weights in the parameter space. Specifically, we first trained a deterministic network $f_{\boldsymbol{\theta}^*} : \mathcal{Z} \to \mathbb{R}$ ($\boldsymbol{\theta}^*$ being trained parameters); followed by optimization of prior precision (over neural network's weights) and observation noise via marginal likelihood maximization. The predictive distribution for a sample x is approximated by $\mathcal{N}(f_{\boldsymbol{\theta}^*}(\mathbf{z}), \mathbf{J}_{\boldsymbol{\theta}^*}(\mathbf{z})\boldsymbol{\Sigma}\mathbf{J}_{\boldsymbol{\theta}^*}(\mathbf{z})^{\mathrm{T}})$, where the $\mathbf{J}_{\boldsymbol{\theta}^*}(\mathbf{z}) = \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\mathbf{z})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ is the Jacobian evaluated at model weights $\boldsymbol{\theta}^*$, and $\boldsymbol{\Sigma} = \left(-\nabla^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{D})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right)^{-1}$ is the inverse of the Hessian. Note, $\mathcal{D}$ in this expression of $\boldsymbol{\Sigma}$ denotes the training data used for training the model $f_{\boldsymbol{\theta}^*}$, which is an MLP network with two hidden layers (each having 50 neurons and ReLU activation) in our experiments.

A.4 Acquisition Function

The acquisition function $\alpha(x|f_{\text{surrogate}})$ defines the utility of performing simulation for measuring CO_2 adsorption capacity of sample x based on the surrogate model $f_{\text{surrogate}}$ learned with the labeled data. The choice of the acquisition strategy dictates the trade-off in terms of the "exploration vs exploitation" goal of the Bayesian optimization. For example, the greedy acquisition strategy prioritizes sample with the highest predictive mean, i.e., exploiting the knowledge of the trained surrogate, which may be overconfident given the limited number of labeled data samples. In contrast, the random acquisition is purely exploration-based without leveraging the current observation data. Eqs. (7) to (10) shows the utility of sample x under greedy, upper confidence bound (UCB) [38], Thompson sampling [31](used

in our work), and expected improvement (EI) [39] acquisition functions, respectively.

$$\alpha_{\text{greedy}}(x|f_{\text{surrogate}}) = \mu_x \tag{7}$$

$$\alpha_{\text{UCB}}(x|f_{\text{surrogate}}) = \mu_x + \beta\sigma_x \quad \text{with } \beta > 0 \tag{8}$$

$$\alpha_{\text{Thompson}}(x|f_{\text{surrogate}}) = \tilde{y} \quad \text{where} \quad \tilde{y} \sim \mathcal{N}(\mu_x, \sigma_x^2) \tag{9}$$

$$\alpha_{\text{EI}}(x|f_{\text{surrogate}}) = \mathbb{E}\left[\max(\tilde{y} - y^*, 0)\right] \text{ where } \tilde{y} \sim \mathcal{N}(\mu_x, \sigma_x^2) \tag{10}$$

Here μ_x and σ_x^2 are the predictive mean and variance from the surrogate model for polymer x. y^* denotes the best objective values within the current set of samples with measured objectives, i.e., the training samples for the surrogate.