Contents lists available at ScienceDirect

Artificial Intelligence Chemistry

journal homepage: www.journals.elsevier.com/artificial-intelligence-chemistry

Generating eco-friendly ionic liquids with enhanced CO_2 solubility using language models

Adroit T.N. Fajar [a,*], Guillaume Lambard [b], Md. Amirul Islam [c], Bidyut B. Saha [c], Zakiah D. Nurfajrin [d], Kevin Septioga [d]

[a] *Center for Energy Systems Design (CESD), International Institute for Carbon-Neutral Energy Research (WPI-I2CNER), Kyushu University, 744 Motooka, Fukuoka 819-0395, Japan*

[b] *Data-driven Materials Design Group, Center for Basic Research on Materials, National Institute for Materials Science, Namiki 1-1, Tsukuba 305-0044, Japan*

[c] *International Institute for Carbon-Neutral Energy Research (WPI-I2CNER), Kyushu University, 744 Motooka, Fukuoka 819-0395, Japan*

[d] *Department of Applied Chemistry, Graduate School of Engineering, Kyushu University, 744 Motooka, Fukuoka 819-0395, Japan*

ARTICLE INFO

Keywords:
Generative model
GPT
Prediction
SMILES
Decarbonization

ABSTRACT

This study presents a viable approach for designing eco-friendly ionic liquids (ILs) with enhanced CO_2 solubility using language models, specifically GPT-2 in conjunction with SMILES-X. The GPT-2 model was fine-tuned on a relatively small, unlabeled IL dataset and subsequently used to generate diverse IL structures. SMILES-X models, trained on IL datasets labeled with CO_2 solubility and eco-toxicity values, were employed to predict the properties of the generated ILs. Trends observed in the predicted IL properties were validated using density functional theory (DFT) and COSMO-RS calculations. The GPT-2 model was then fine-tuned iteratively, with the training data updated by including the top generated ILs from previous cycles. This iterative process led to a gradual improvement in the properties of the generated ILs. It was also observed, however, that continuously adding curated generated ILs to the training data eventually caused the model to produce correct but unrealistic IL structures. These findings highlight both the potential and limitations of language models in designing novel chemicals. Additionally, the CO_2 adsorption capacity of a surrogate IL was experimentally measured, demonstrating the potential of this approach in advancing decarbonization technologies.

1. Introduction

Climate change represents one of the most pressing global challenges, driven primarily by the unprecedented increase in atmospheric CO_2 levels. Since the Industrial Revolution, fossil fuel combustion has raised CO_2 concentrations from approximately 280 ppm in the pre-industrial era to over 424 ppm in November 2024 [1]. The Intergovernmental Panel on Climate Change has warned of severe consequences, including rising sea levels, extreme weather events, and disruptions to ecosystems [2]. To mitigate these impacts, there is an urgent need to transition from carbon-intensive energy systems to renewable energy sources and to develop technologies capable of capturing and sequestering excess CO_2 from the atmosphere [3].

Ionic liquids (ILs) are a promising material for advancing carbon capture technology, whether used directly in the liquid state, as additives to solid adsorbents, or in composite materials and membranes [4–6]. One of the primary advantages of ILs lies in their tunable physicochemical properties, which can be adjusted by altering the cation-anion combinations or incorporating specific functional groups. However, this potential has been underexplored, as evaluating numerous IL combinations through experiments or first-principles calculations is both costly and time-consuming. Most studies on ILs in CO_2 capture have only evaluated a limited number of individual or series of ILs [7], which restricts the full exploration of their potential in enhancing CO_2 capture technologies.

Data-driven approaches [8,9] and Generative AI have great potential in accelerating the exploration of ILs' chemical space. Generative deep learning models—such as variational autoencoder (VAE) and generative adversarial network (GAN)—have shown promising results in generating molecular structures [10–12], opening new avenues for inverse design. Recent reports highlight the application of generative models in exploring ILs for CO_2 capture. For example, Liu et al. employed a novel architecture combining a syntax-directed VAE, deep factorization machines, and gradient-based particle swarm optimization to smartly

[*] Corresponding author.
E-mail address: adroit@i2cner.kyushu-u.ac.jp (A.T.N. Fajar).

https://doi.org/10.1016/j.aichem.2025.100089
Received 30 January 2025; Received in revised form 7 May 2025; Accepted 21 May 2025
Available online 22 May 2025

design ILs for CO_2 capture [13]. Similarly, Chen et al. expanded this approach with a larger IL dataset to enhance exploration [14]. In another study, Lim proposed a combination of deep learning and quantitative structure-property relationship modeling to generate ILs optimized for CO_2 capture [15]. Despite their great potential, however, these approaches rely on training deep stack neural networks, mainly based on autoencoders, from the ground up. Considering that IL datasets are relatively small for generative models—for instance, the ILThermo Database (NIST) contains fewer than 3000 ILs [16]—training a model from scratch may not be the most effective strategy. Furthermore, most studies focus primarily on improving CO_2 solubility, overlooking a key challenge for industrial-scale implementation: IL toxicity. Addressing environmental toxicity of ILs is crucial [17,18], as it represents a major barrier to the widespread adoption of IL-based CO_2 capture technologies.

In this study, we present a practical approach for designing eco-friendly ionic liquids (ILs) with enhanced CO_2 solubility using language models. Rather than training a deep learning model from scratch, we fine-tuned an existing large language model (LLM)—the GPT-2 model [19]—using relatively small IL datasets, demonstrating its effectiveness for this task. GPT-2 was chosen for its efficiency in modeling sequential data, ease of fine-tuning, and accessibility through robust frameworks like Hugging Face [20]. While larger or specialized models (e.g., GPT-3/GPT-4, ChemBERTa) may offer enhanced performance, GPT-2 provides a balanced trade-off between computational cost and generative capability for targeted tasks. Additionally, we integrated the molecular characterization tool SMILES-X [21], also based on a language model, into the workflow to navigate the properties of the generated ILs. The potential of this approach was evaluated using established simulation methods (DFT and COSMO-RS) as well as experimental measurements.

2. Methods

2.1. Preparation of the datasets

Three IL datasets were collected from the literature, labeled as Data T0, Data T1, and Data T2 (see Note S1). Data T0 contains 3109 IL structures without associated property labels, Data T1 includes 564 IL structures with corresponding CO_2 solubility values, and Data T2 consists of 110 IL structures with corresponding EC_{50} values for eco-toxicity. All IL structures (concatenated cations and anions) were represented using the Simplified Molecular Input Line Entry System (SMILES) in their canonical forms. The cheminformatics tool RDKit [22] was employed to verify the accuracy and correctness of the SMILES strings, ensuring structural integrity before model training.

2.2. Generative model

The generative model was developed by fine-tuning the GPT-2 model [19] (approximately 124 million parameters) from the Transformer library on Hugging Face [20] using Data T0. SMILES strings from Data T0 were tokenized using the GPT-2 tokenizer, with padding and truncation to standardize input lengths. The model was trained for up to 100 epochs, with early stopping employed to prevent overfitting. Key optimization techniques, including warmup steps, weight decay, and epoch-based evaluation, were applied to enhance model performance. After training, the model was evaluated on test data by monitoring test loss (see Note S2). Using the fine-tuned model, we targeted the generation of 10,000 unique and valid SMILES strings, with a maximum of 1 million attempts allowed to ensure quality and diversity. RDKit was used to validate each generated structure, filtering out invalid SMILES, unbalanced charges, radicals, and unchanged structures. This process ensured that only chemically valid ILs (labelled as Data G0) were retained for further analysis. Additional details about the GPT-2 model and how the new ILs were generated are provided in Note S2. The generative model was implemented using PyTorch [23]. The training typically took around 6–12 hours per cycle on four NVIDIA RTX A5000 GPUs, with later cycles taking slightly longer due to the growing size of the training data.

2.3. Prediction models

The prediction models were prepared using the molecular characterization tool SMILES-X [21], introduced by Lambard et al. CO_2 solubility (ML-1) and eco-toxicity (ML-2) predictors were trained on Data T1 and Data T2, respectively. The SMILES-X architecture was configured with three primary hyperparameters: the size of the embedding layer, and the number of units in both the LSTM and dense layers, with possible values of [8, 16, 32, 64, 128, 256, 512, 1024]. These hyperparameters were optimized using zero-cost geometry optimization to automatically identify the optimal architecture (Note S3). For further optimization, the batch size and learning rate (in powers of 10) were determined via Bayesian optimization, with search regions set to [8,16, 32,64] for batch size and [2.0, 2.5, 3.0, 3.1, 3.2, …, 4.0] for learning rate. The models were trained for up to 300 epochs, with generalization capabilities evaluated using k-fold cross-validation ($k = 3$ for both models). Each fold in the k-fold cross-validation was repeated three times using different random seeds (3 runs per fold). The reported performance metrics are averaged over these runs to ensure a reliable estimation of model generalization and to provide information on standard deviation. ML-1 took temperature and pressure as additional input features, while ML-2 only used the default [SMILES, property] inputs. The SMILES-X models were trained using a single NVIDIA RTX A5000 GPU. Each model required approximately 9 hours to complete the training. The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) [24] was then employed to rank the generated ILs based on ML-1 and ML-2 simultaneously (see Note S3).

2.4. DFT and COSMO-RS calculations

The COnductor-like Screening MOdel for Real Solvents (COSMO-RS) theory was employed to estimate key physical properties of selected representative ILs from Data G0. For each SMILES input (cation or anion), a conformer search was performed using the COSMOconfX package, following the default workflow for BP-TZVP-COSMO job templates. This workflow applies the Becke-Perdew (BP86) density functional theory (DFT) combined with the Triple Zeta Valence Polarized (TZVP) basis set, and outputs a *.*cosmo* file. All necessary DFT calculations during the conformer search were conducted using the TURBOMOLE package. Each cation or anion required approximately 3–6 hours on 8 CPU cores, depending on the number of atoms. The resulting *.*cosmo* files, which contain electronic density and surface charge information, were then used as inputs for COSMO-RS calculations. The COSMOtherm package was subsequently employed to calculate physical properties, including Henry's constant (K_H), Gibbs free energy of solvation (ΔG_{solv}), the water-octanol partition coefficient ($\log P$), and the activity coefficient ($\ln(\gamma)_{oct}$). COSMO-RS calculations are computationally inexpensive and took only a few seconds per IL once the COSMO files were prepared.

2.5. Iterative training and generation

The top 50 % of ILs from Data G0, based on the TOPSIS ranking, were added to the original training dataset (Data T0), resulting in an expanded training set of 5522 entries, labeled as Data T1. The GPT-2 model was then re-trained using Data T1, and the newly fine-tuned model was used to generate another set of ILs, labeled as Data G1. This iterative cycle was repeated up to 11 times, with the number of training ILs increasing in each cycle by incorporating the top half of the generated ILs from the previous cycle. The cumulative generated ILs were then evaluated using a combined score of CO_2 solubility and eco-toxicity (S-E score; see Note S4). Subsequently, a similarity search was conducted to identify commercially available ILs that match the top generated ILs (see Note S5). Additionally, the synthetic accessibility (SA) score was calculated for training and generated ILs (see Note S6). The distributions of SA scores were visualized using a kernel density estimate plot, creating a continuous density curve to facilitate easier trend identification.

2.6. Adsorption isotherm

The CO_2 adsorption-desorption behavior of a commercially available IL with the highest similarity (Tanimoto Index: 0.87) to one of the top 1000 generated ILs was investigated using a 3Flex analyzer (Micromeritics Instrument Corporation, USA). The IL sample, 1-(2-Hydroxyethyl)-3-methylimidazolium Bis(trifluoromethanesulfonyl) imide (>98.0 %), was purchased from Tokyo Chemical Industry Co., Ltd. and used without further purification. Prior to the analysis, the samples were degassed at 80°C for four hours under vacuum to remove contaminants. Approximately 40 mg of the sample was accurately measured and introduced into the analyzer's sample tubes, followed by an additional vacuum cycle. A precise quantity of CO_2 was introduced into the sample tube, and the pressure sensors continuously monitored and recorded the initial pressure, pressure decrease, and equilibrium points. The adsorption temperature was maintained at 0°C using a Julabo water circulator and an isothermal water bath. The investigated pressure range spanned from zero to 110 kPa. CO_2 density changes within the sample tube were determined using REFPROP software, based on the adsorption temperature, initial pressure, and final pressure.

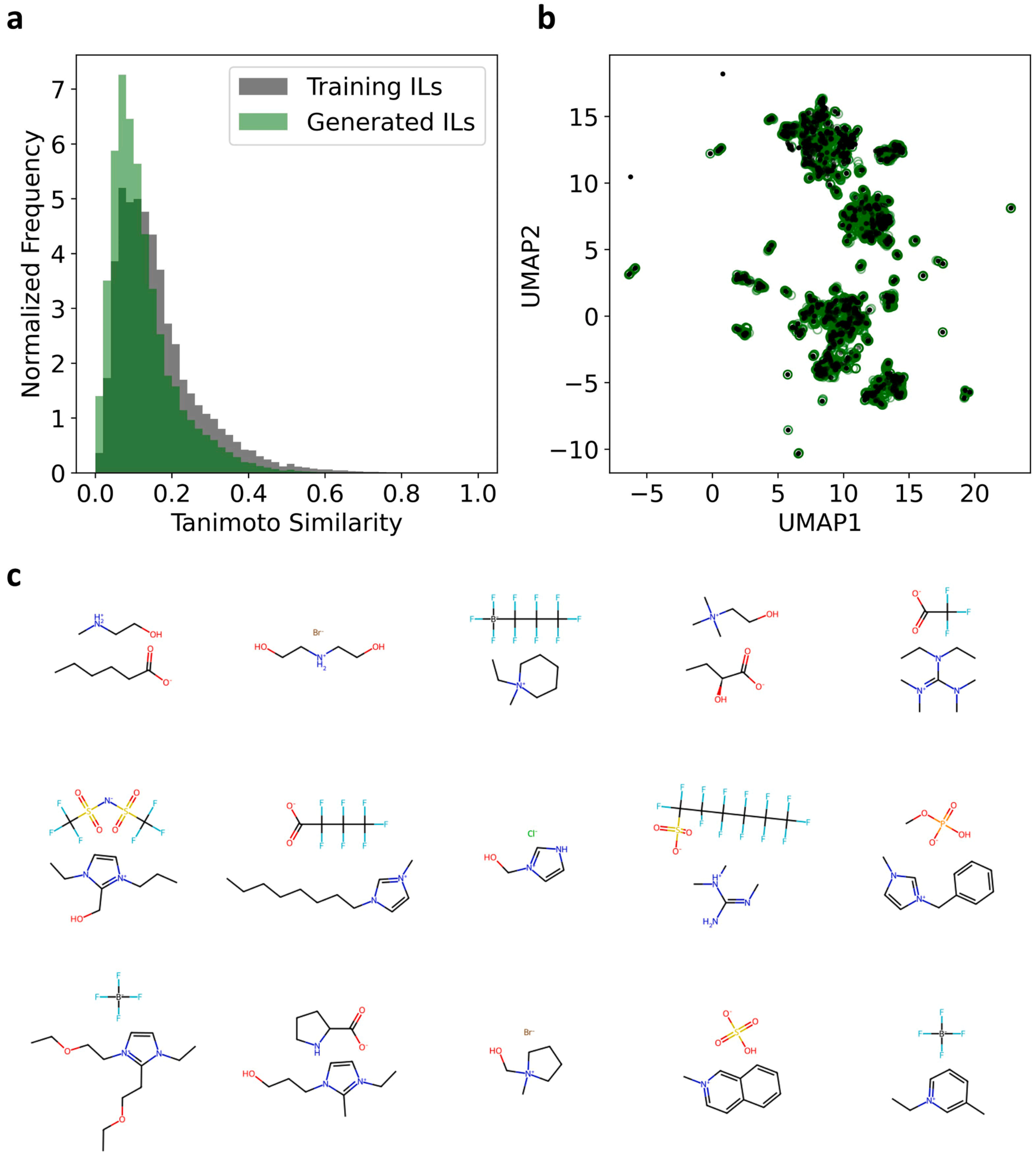

Fig. 1. Diversity profile of the generated ILs. (a) Tanimoto similarity histogram and (b) UMAP projections of both training and generated ILs. (c) Examples of several generated IL structures.

For comparison, the CO_2 adsorption–desorption behavior of a commercially available IL with a relatively low similarity score (Tanimoto Index: 0.74), namely 1-ethyl-3-methylimidazolium triflate (>98.0 %, Tokyo Chemical Industry Co., Ltd.), was also investigated.

3. Results and discussion

3.1. Generation of IL structures

Fine-tuning a pre-trained model, also known as transfer learning, is an effective approach for building task-specific models using relatively small datasets [25,26]. In this study, 3109 IL structures, represented by concatenated SMILES strings of the cation and anion, were used to fine-tune the GPT-2 model. Property labels were not required at this stage, as the training objective was solely to introduce IL grammar into the model. The learning curve (training vs. validation loss) during the fine-tuning process, along with the distribution of test losses, is shown in Figure S1. The average test loss was as low as 0.12, indicating that the model successfully recognized IL structure patterns. After training, the fine-tuned model generated a substantial number (4825) of unique, novel, and chemically valid IL structures (Data G0).

Fig. 1 shows the diversity profile of the generated ILs, along with examples of several structures. As shown in Fig. 1a, the Tanimoto similarity index [27] for both training and generated ILs is distributed at values close to 0, with average pairwise similarity as low as 0.16 and 0.12, respectively. This indicates that both training and generated ILs exhibit high sub-structural diversity. When projected onto a 2D plane using the uniform manifold approximation and projection (UMAP) [28] technique (Fig. 1b), the generated ILs also display a broad distribution that resembles, but does not replicate, the training data distribution. This demonstrates the effectiveness of a language model in molecular generation [29,30] and highlights the efficiency of fine-tuning an existing LLM for small and specialized datasets. Examples of the generated ILs are shown in Fig. 1c. At this stage, the generated ILs were random and not yet directed toward specific properties. Therefore, the molecular characterization tool SMILES-X, was then incorporated into the workflow to facilitate directed design.

3.2. Prediction of IL properties

The primary advantage of SMILES-X is that it performs well with small datasets and does not require molecular descriptors [21,31]. In this study, CO_2 solubility (ML-1) and IL eco-toxicity (EC_{50}; ML-2) predictors were trained using SMILES-X on datasets containing IL structures and their corresponding properties (Data T1 and Data T2, respectively). Fig. 2a,b shows the averaged cross-validation parity plots (measured vs. predicted) for ML-1 and ML-2. Parity plots for each run in each fold during k-fold cross-validation are shown in Figure S2. As seen in Fig. 2a, ML-1 exhibits a low averaged root mean square error (RMSE: 7.38 mmol/mol) and mean absolute error (MAE: 5.5 mmol/mol), with a coefficient of determination (R^2) of 0.72. Similarly, ML-2 shows reasonably low RMSE (0.70) and MAE (0.50), with an R^2 of 0.67 (Fig. 2b). Considering that these models were trained on small datasets without requiring molecular descriptors—typically derived from costly first-principle calculations [32,33]—these results are notably good.

Fig. 2c,d shows the distribution of predicted CO_2 solubility and EC_{50} values for both the training (Data T0) and generated (Data G0) ILs. Although the generated ILs are novel, their property distributions closely resemble those of the training ILs, which aligns with the structural distribution shown in the UMAP projections (Fig. 1b). This behavior is useful for directing the generated ILs toward specific properties. Notably, the objective of this study is to generate ILs with both

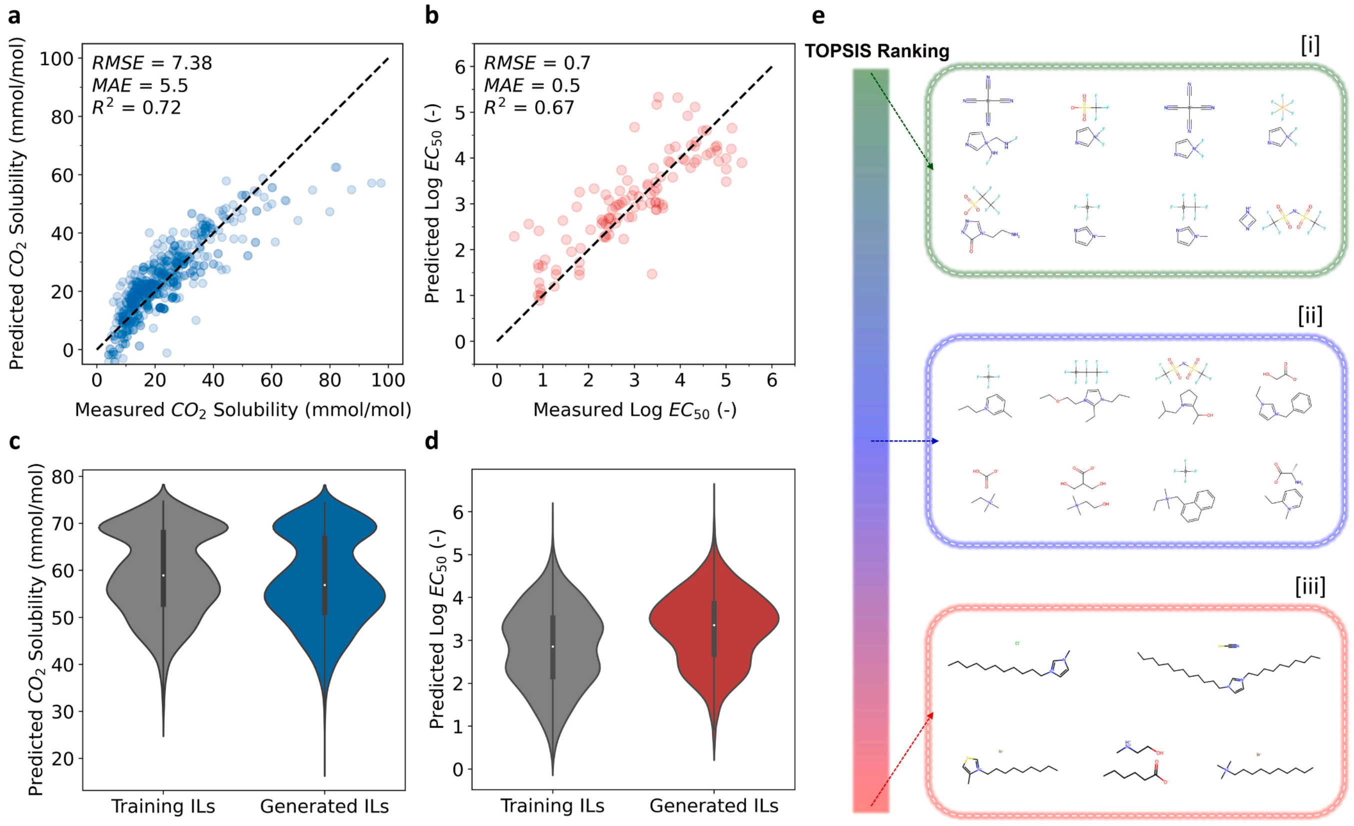

Fig. 2. Characterizing IL properties with prediction models. Averaged cross-validation parity plots of (a) ML-1 (CO_2 solubility predictor) and (b) ML-2 (IL eco-toxicity predictor, EC_{50} values were measured in μM). Distribution of (c) predicted CO_2 solubility and (d) predicted EC_{50} values for the training (Data T0) and generated (Data G0) ILs. (e) Examples of IL structures ranked at the [i] top, [ii] middle, and [iii] bottom based on ML-1 and ML-2 using the TOPSIS method.

high CO_2 solubility and low eco-toxicity, requiring simultaneous consideration of ML-1 and ML-2. To achieve this, a multicriteria decision analysis method—the TOPSIS method—was applied to rank the generated ILs. Fig. 2e shows examples of the generated ILs ranked at the top, middle, and bottom of the TOPSIS scale. ILs selected from the top rank (indicated by the green-blue section in the schematic scale bar) are more likely to exhibit higher CO_2 solubility and lower eco-toxicity compared to the rest of the list.

3.3. Insight from simulations

To gain further insights from the TOPSIS ranking of the generated ILs, we performed DFT and COSMO-RS calculations. Four physicochemical properties were calculated for five ILs from each of the top (1, 11, 21, 31, 41), middle (2393, 2403, 2413, 2423, 2433), and bottom (4785, 4795, 4805, 4815, 4825) ranks. The properties included Henry's constant (K_H), Gibbs free energy of solvation (ΔG_{solv}), the water-octanol partition coefficient of the cation ($\log P^+$), and the activity coefficient towards n-octanol ($\ln(\gamma)_{oct}$). The COSMO views of these 15 representative IL structures are shown in Figure S3. K_H and ΔG_{solv} are related to IL$-CO_2$ interactions, while $\log P^+$ and $\ln(\gamma)_{oct}$ provide insight into the ILs' hydrophilicity/lipophilicity, which may indicate IL$-$lipid interactions (represented by octanol as a lipid surrogate) in cell membranes and, by extension, potential toxicity levels (see Table 1).

As shown in Table 1, ILs in the top rank generally exhibit lower K_H values with negative ΔG_{solv}, indicating stronger and spontaneous interactions with CO_2. The top-ranked ILs also display lower $\log P^+$ and positive $\ln(\gamma)_{oct}$ values, suggesting they are less lipophilic and may have minimal interaction with lipid structures, such as cell membranes, thus potentially reducing toxicity risk. Although ILs in the middle rank also exhibit favorable interactions with CO_2 (low K_H and negative ΔG_{solv}), their toxicity risk tends to increase, as indicated by higher $\log P^+$ and lower $\ln(\gamma)_{oct}$ values. In the bottom rank, K_H values generally increase slightly, though ΔG_{solv} remains negative, maintaining favorable CO_2 interactions. However, changes in toxicity indicators are more pronounced, with further increases in $\log P^+$ and decreases in $\ln(\gamma)_{oct}$, which may elevate toxicity risk due to a stronger likelihood of interaction with lipid-like structures that could lead to cell membrane disruption. This analysis supports the validity of the TOPSIS ranking and suggests it can be effectively used to identify ILs with desirable properties.

3.4. Navigating IL generation

To improve the properties of the generated ILs, GPT-2 was fine-tuned iteratively, assisted by the SMILES-X models (see Methods section for details). Fig. 3a shows the number of training entries, generated ILs, and cumulative generated ILs through cycle 10. Interestingly, while the training dataset steadily increased over the cycles (from cycle 0: 3109 to cycle 10: 23,764 entries), the number of chemically valid generated ILs did not grow proportionally. The number of generated ILs gradually decreased from cycle 0 to cycle 3 (G0: 4825; G1: 4294; G2: 3240; G3: 3093), then fluctuated in subsequent cycles, ultimately increasing by cycle 10. Despite this irregularity, the cumulative count of generated ILs (excluding duplicates) continued to grow across cycles, indicating that the self-generated synthetic data were quantitatively useful for expanding the molecular search space.

Table 1
Physicochemical properties related to IL$-CO_2$ interactions (K_H and ΔG_{solv}) and IL-octanol interactions (used as a simplified model of lipid-like structures; $\log P^+$ and $\ln(\gamma)_{oct}$) calculated using COSMO-RS theory.

IL index[#]	K_H (bar)	ΔG_{solv} (kJ·mol^{-1})	$\log P^+$	$\ln(\gamma)_{oct}$
1	164.64	−0.23	0.26	5.06
11	181.50	−0.28	0.07	5.37
21	170.15	−0.05	0.71	7.98
31	190.62	0.26	1.18	7.41
41	210.50	−0.28	−0.05	5.37
2393	186.59	−0.01	3.77	1.68
2403	160.53	−0.35	2.49	−0.20
2413	170.05	−0.96	2.07	−4.34
2423	198.04	−0.04	5.34	0.70
2433	238.58	0.08	5.61	0.88
4785	159.35	−0.21	6.42	−3.54
4795	167.05	−0.29	4.59	−0.61
4805	284.01	0.17	3.47	0.44
4815	344.27	−0.01	2.78	2.72
4825	237.78	−0.14	8.05	0.79

[#] IL index in the TOPSIS ranking.

It is important to note, however, that LLMs may collapse when trained on recursively generated data, as recently reported by Shumailov et al. [34]. This phenomenon was also observed in the context of chemical structure generation. As shown in Fig. 3b, high-quality ILs were generated up to cycle 4, after which the model appeared to collapse at cycle 5, producing N-based structures. These ILs exhibited valid chemical structures but were unrealistic and likely challenging to synthesize. On one hand, this behavior suggests that the model learned the importance of nitrogen atoms in CO_2 capture, likely due to the presence of Lewis base sites that interact with CO_2 to form carbamate ions or carbamic acid [35,36]. On the other hand, the iterative training process led to an overemphasis on nitrogen, ultimately limiting the model's performance. While LLMs hold great potential for advancing molecular design, users should be mindful of these limitations. Accordingly, only the cumulative generated ILs up to cycle 4 (Data G0–G4) were considered for further analysis.

Fig. 3c shows the combined CO_2 solubility and eco-toxicity (S-E) scores of the original training ILs (Data T0) and the generated ILs through cycle 4 (Data G0–G4). As observed, the maximum S-E scores improved incrementally from Data T0 to Data G4 (T0: 0.26; G0: 0.27; G1: 0.28; G2: 0.27; G3: 0.28; G4: 0.28), demonstrating the effectiveness of the iterative process in enhancing the properties of the generated ILs. Interestingly, the maximum S-E scores continued to increase through cycle 10 (see Figure S4), despite the generation of N-based structures. This highlights that, when used with caution, an iteratively fine-tuned LLM can be a practical and effective tool for exploring the vast chemical space of ILs.

3.5. Potential implementations

Borrowing a common practice from modern drug discovery workflows [37,38], a similarity search was conducted to identify commercially available ILs with structures similar to those of the top generated ILs. Table S1 lists commercially available ILs with similarity scores ≥ 0.7 to the top 1000 generated ILs through cycle 4 (Data G0–G4). Notably, none of the commercially available ILs achieved a similarity score ≥ 0.9, indicating that the model successfully explored previously uncharted chemical space, particularly in the context of the target properties. To further evaluate this finding, we assessed the CO_2 adsorption capacity of the commercial IL with the highest similarity score (0.87), namely 1-(2-Hydroxyethyl)-3-methylimidazolium Bis(trifluoromethanesulfonyl) imide, as a surrogate to the generated IL. A structure comparison between the generated and commercial ILs is shown in Fig. 4a.

As shown in Fig. 4b, the selected commercial IL exhibited excellent CO_2 adsorption capacity, achieving approximately 70 mmol/mol at 110 kPa (predicted as 66.74 mmol/mol by ML-1). Notably, this value is comparable to the highest CO_2 solubility region in the original training data (see Fig. 2a,c). For comparison, the CO_2 adsorption capacity of a commercially available IL with a relatively low similarity score (0.74) was also measured, showing a reduced capacity of approximately 36 mmol/mol at 110 kPa (see Figure S5). This finding demonstrates how a combination of generative models and similarity search can identify existing chemicals with unrecognized potential for novel applications, analogous to drug repurposing in the pharmaceutical domain [39]. To further investigate how the IL interacts with CO_2, we analyzed the

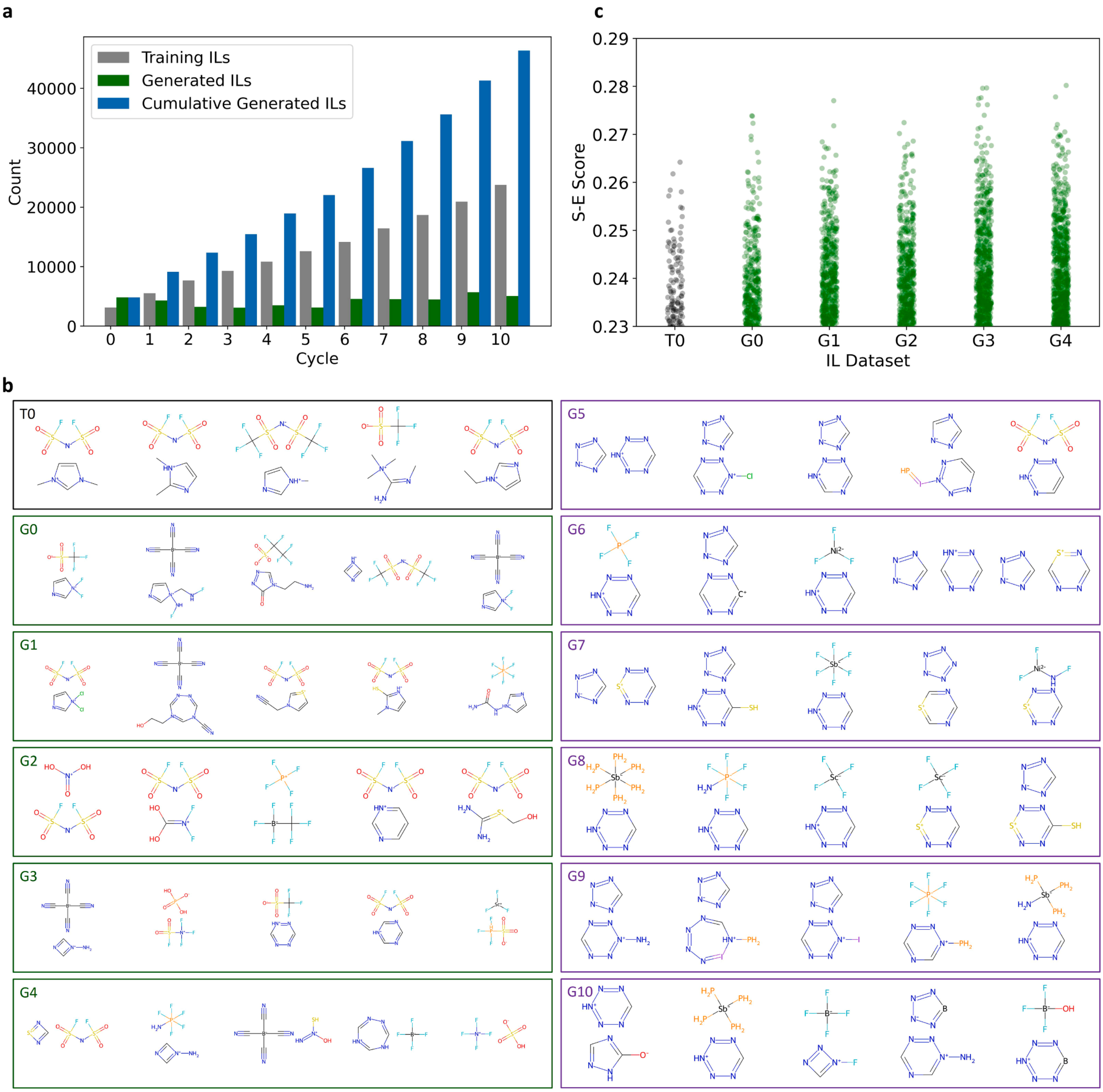

Fig. 3. Improvement in properties of the generated ILs. (a) Number of training, generated, and cumulative generated ILs throughout iterative training and generation cycles. (b) Structures of the top 5 ILs from the original training dataset (Data T0) and generated ILs through cycle 10 (Data G0–G10). (c) Combined CO_2 solubility and eco-toxicity (S-E) scores of the original training ILs (Data T0) and generated ILs through cycle 4 (Data G0–G4).

σ-profile of the cation, anion, and CO_2 (Fig. 4c). The cation contributes moderate dipole and hydrogen-bonding interactions with CO_2 (peaks around −0.01 e/Å), while the anion provides significant van der Waals (peaks around 0 e/Å^2) and dipole interactions due to its sharp polar regions (peaks around 0.01 e/Å^2). The complementary interactions between the ionic liquid components and CO_2 suggest a synergistic effect, with the anion playing a dominant role in enhancing CO_2 solubility.

Nevertheless, it is important to note that the tested IL was not one of the ILs directly suggested by our models. Evaluating the generated ILs will require novel chemical synthesis, which is a complex and non-trivial task that will be addressed in future work. As an initial insight, the synthetic accessibility (SA) score, developed by Ertl and Schuffenhauer [40], was used to assess the feasibility of synthesizing the generated ILs. Fig. 4d–f show the distribution of SA scores for Data T0, Data G0–G4, and Data G5–G10, respectively. The SA score ranges from 1 to 10, where lower scores indicate higher synthetic accessibility (easier synthesis), and higher scores represent lower accessibility (more challenging synthesis). The SA score distribution for ILs in Data T0 (Fig. 4d) is similar to that of Data G0–G4 (Fig. 4e), suggesting that synthesizing the generated ILs is as feasible as synthesizing ILs already reported in the literature. Furthermore, like the training ILs, the majority of the generated ILs have SA scores below 5, indicating they are “relatively easy” to synthesize based on fragment contributions and complexity penalties. In contrast, the SA scores of ILs in Data G5–G10 shift to a higher range (>5), confirming that the quality of the generated ILs was compromised after cycle 5. Overall, this analysis highlights the potential of the generated

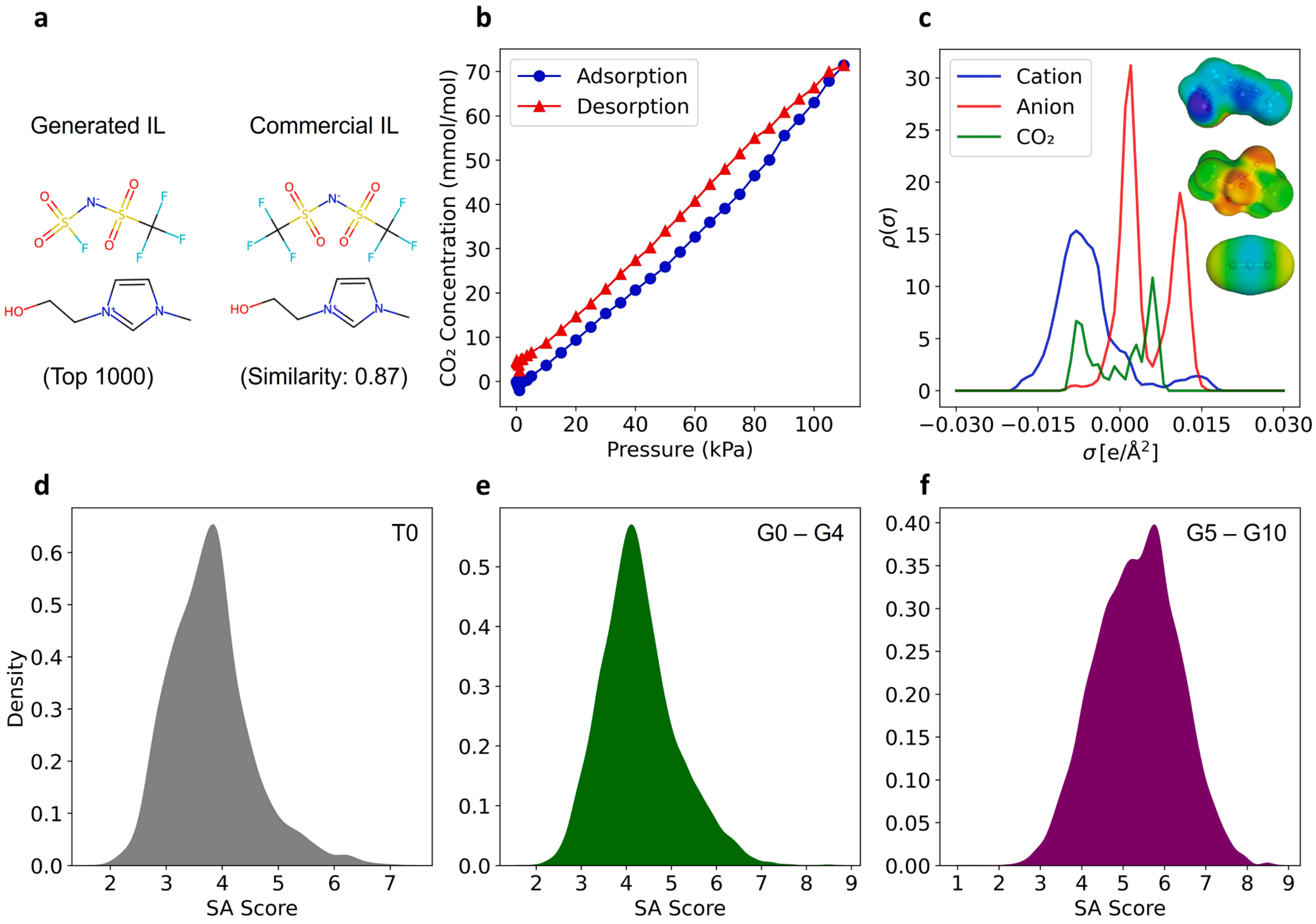

Fig. 4. Potential implementation of the findings. (a) Structural comparison of a generated IL and a commercial IL with 87 % similarity. (b) Experimentally measured CO_2 adsorption-desorption behavior of the selected commercial IL. (c) σ-profile of the commercial IL and CO_2, illustrating possible complementary interactions. (d–f) Synthetic accessibility (SA) scores of the original training ILs (Data T0), generated ILs through cycle 4 (Data G0–G4), and generated ILs after cycle 4 (Data G5–G10), respectively.

ILs (up to cycle 4) for real-world applications. Nevertheless, it should be noted that while the SA score can provide initial insight into synthesis feasibility, practical challenges in IL synthesis (e.g., precursor availability and reaction pathways) are not accounted for by this metric. Future work should address these challenges by incorporating more comprehensive synthesis knowledge into the models.

4. Conclusions

This study demonstrates the effectiveness of using language models, specifically GPT-2 combined with SMILES-X, for designing novel ILs with targeted properties. The fine-tuned GPT-2 model successfully recognized IL patterns (test loss: 0.12) and generated structurally diverse ILs (Tanimoto index: 0.12). The SMILES-X models provided reasonably accurate predictions for CO_2 solubility (MAE: 5.5 mmol/mol) and IL eco-toxicity (MAE: 0.5) using a relatively small dataset without requiring molecular descriptors. The TOPSIS method was employed to rank the generated ILs based on two SMILES-X models simultaneously, and the observed trends aligned well with DFT and COSMO-RS calculations. Iterative fine-tuning of GPT-2 with the training data augmented by generated ILs curated by SMILES-X models significantly expanded the chemical search space, although this process carries the risk of model collapse after a certain point. Nevertheless, when used cautiously, LLMs can be powerful tools for exploring the vast chemical space of ILs. None of the commercially available ILs exhibited structural similarity greater than 90 % to the generated ILs. However, the tested IL with 87 % similarity demonstrated reasonably high CO2 solubility (≈70 mmol/mol at 110 kPa). Comprehensive assessment of the generated ILs will require novel chemical synthesis, and the low SA score distribution (<5) indicates the feasibility of such synthesis. Future work will focus on developing a comprehensive synthesis protocol for novel ILs designed by generative models, considering reaction pathways, precursor availability, and cost analysis. Additionally, future work will also explore the use of both smaller, chemically specialized models (e.g., MolGPT) to reduce computational cost and overfitting, as well as larger-scale models (e.g., LLaMA-2) to assess whether more expressive architectures offer improved performance in IL generation tasks. These findings will then be integrated into laboratory assessments to advance decarbonization technologies.

CRediT authorship contribution statement

Adroit T.N. Fajar: Writing – original draft, Resources, Methodology, Investigation, Funding acquisition, Conceptualization. **Guillaume Lambard:** Writing – review & editing, Methodology, Conceptualization. **Md. Amirul Islam:** Writing – review & editing, Investigation. **Bidyut B. Saha:** Resources. **Zakiah D. Nurfajrin:** Data curation. **Kevin Septioga:** Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Center for Energy Systems Design (CESD), International Institute for Carbon-Neutral Energy Research (WPI-I2CNER), which was established by the World Premier International Research Center Initiative (WPI), MEXT, Japan. Additional support was provided by (1) Kyushu University Education and Research Support Program for Mathematical and Data Science and (2) ROIS NII Open Collaborative Research (251FP-22667).

Code availability

All codes used in this study have been deposited at: https://github.com/adroitfajar/ionic-gen.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.aichem.2025.100089.

Data availability

All datasets used (Data T0–T2) and generated (Data G0–G10) in this study have been deposited at: https://doi.org/10.6084/m9.figshare.28159445.

References

[1] Carbon Dioxide, ⟨Https://Climate.Nasa.Gov/Vital-Signs/Carbon-Dioxide/?Intent= 121⟩ (2024).

[2] Climate Change 2021: The Physical Science Basis, ⟨Https://Www.Ipcc.Ch/Report/Ar6/Wg1/⟩ (2021).

[3] Net Zero by 2050: A Roadmap for the Global Energy Sector, ⟨Https://Www.Iea.Org/Reports/Net-Zero-by-2050⟩ (2021).

[4] S. Zeng, X. Zhang, L. Bai, X. Zhang, H. Wang, J. Wang, D. Bao, M. Li, X. Liu, S. Zhang, Ionic-liquid-based CO2 capture systems: structure, interaction and process, Chem. Rev. 117 (2017) 9625–9673, https://doi.org/10.1021/acs.chemrev.7b00072.

[5] F. Weisshar, A. Gau, J. Hack, N. Maeda, D.M. Meier, Toward carbon dioxide capture from the atmosphere: lowering the regeneration temperature of polyethylenimine-based adsorbents by ionic liquid, Energy Fuels 35 (2021) 9059–9062, https://doi.org/10.1021/acs.energyfuels.1c00392.

[6] A.V. Bhaskar Reddy, M. Moniruzzaman, M.A. Bustam, M. Goto, B.B. Saha, C. Janiak, Ionic liquid polymer materials with tunable nanopores controlled by surfactant aggregates: a novel approach for CO2capture, J. Mater. Chem. A Mater. 8 (2020) 15034–15041, https://doi.org/10.1039/c9ta13077b.

[7] W. Faisal Elmobarak, F. Almomani, M. Tawalbeh, A. Al-Othman, R. Martis, K. Rasool, Current status of CO2 capture with ionic liquids: development and progress, Fuel 344 (2023), https://doi.org/10.1016/j.fuel.2023.128102.

[8] X. Zhang, J. Wang, Z. Song, T. Zhou, Data-driven ionic liquid design for CO2Capture: molecular structure optimization and DFT verification, Ind. Eng. Chem. Res 60 (2021) 9992–10000, https://doi.org/10.1021/acs.iecr.1c01384.

[9] Z. Song, H. Shi, X. Zhang, T. Zhou, Prediction of CO2 solubility in ionic liquids using machine learning methods, Chem. Eng. Sci. 223 (2020), https://doi.org/10.1016/j.ces.2020.115752.

[10] B. Sanchez-Lengeling, A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, n.d. ⟨https://www.science.org⟩.

[11] R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams, A. Aspuru-Guzik, Automatic chemical design using a data-driven continuous representation of molecules, ACS Cent. Sci. 4 (2018) 268–276, https://doi.org/10.1021/acscentsci.7b00572.

[12] A. Maziarka, J. Pocha, K. Kaczmarczyk, T. Rataj, M. Danel, Warchoł, Mol-CycleGAN: a generative model for molecular optimization, J. Chemin.-. 12 (2020), https://doi.org/10.1186/s13321-019-0404-1.

[13] X. Liu, J. Chu, S. Huang, A. Li, S. Wang, M. He, Machine learning-based design of ionic liquids at the atomic scale for highly efficient CO2 capture, ACS Sustain Chem. Eng. 11 (2023) 8978–8987, https://doi.org/10.1021/acssuschemeng.3c01191.

[14] X. Chen, G. Chen, K. Xie, J. Cheng, J. Chen, Z. Song, Z. Qi, Exploring the chemical space of ionic liquids for CO2 dissolution through generative machine learning models, Green. Chem. Eng. (2024), https://doi.org/10.1016/j.gce.2024.06.005.

[15] H. Lim, Extension of scoring-assisted generative exploration for ionic liquids (SAGE-IL) and its application to ionic liquid design for CO2 capture, Mater. Today Adv. 24 (2024), https://doi.org/10.1016/j.mtadv.2024.100529.

[16] Q. Dong, C.D. Muzny, A. Kazakov, V. Diky, J.W. Magee, J.A. Widegren, R. D. Chirico, K.N. Marsh, M. Frenkel, ILThermo: a free-access web database for thermodynamic properties of ionic liquids, J. Chem. Eng. Data 52 (2007) 1151–1159, https://doi.org/10.1021/je700171f.

[17] R.M. Cuéllar-Franca, P. García-Gutiérrez, S.F.R. Taylor, C. Hardacre, A. Azapagic, A novel methodology for assessing the environmental sustainability of ionic liquids used for CO2 capture, Faraday Discuss. 192 (2016) 283–301, https://doi.org/10.1039/c6fd00054a.

[18] Z. Wang, Z. Song, T. Zhou, Machine learning for ionic liquid toxicity prediction, Processes 9 (2021) 1–10, https://doi.org/10.3390/pr9010065.

[19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, Open. Blog 1 (2019) 1–9.

[20] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. Von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T.Le Scao, S. Gugger, M. Drame, Q. Lhoest, A.M. Rush, Transformers: State-of-the-Art Natural Language Processing, n.d. ⟨https://github.com/huggingface/⟩.

[21] G. Lambard, E. Gracheva, SMILES-X: autonomous molecular compounds characterization for small datasets without descriptors, Mach. Learn Sci. Technol. 1 (2020), https://doi.org/10.1088/2632-2153/ab57f3.

[22] RDKit: Open-source cheminformatics. ⟨https://www.rdkit.org⟩, (n.d.).

[23] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, 2019: pp. 8024–8035. ⟨https://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf⟩.

[24] C.-L. Hwang, K. Yoon, Multiple Attribute Decision Making: Methods and Applications, Springer, Berlin, Heidelberg, 1981, https://doi.org/10.1007/978-3-642-48318-9.

[25] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016. ⟨https://www.deeplearningbook.org⟩.

[26] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems, Curran Associates Inc., 2012: pp. 1097–1105.

[27] T.T. Tanimoto, Elementary Mathematical Theory of Classification and Prediction, 1958.

[28] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv Preprint ArXiv:1802.03426 (2018). ⟨https://arxiv.org/abs/1802.03426⟩.

[29] D. Flam-Shepherd, K. Zhu, A. Aspuru-Guzik, Language models can learn complex molecular distributions, Nat. Commun. 13 (2022), https://doi.org/10.1038/s41467-022-30839-x.

[30] W. Feng, L. Wang, Z. Lin, Y. Zhu, H. Wang, J. Dong, R. Bai, H. Wang, J. Zhou, W. Peng, B. Huang, W. Zhou, Generation of 3D molecules in pockets via a language model, Nat. Mach. Intell. 6 (2024) 62–73, https://doi.org/10.1038/s42256-023-00775-6.

[31] E. Gracheva, G. Lambard, S. Samitsu, K. Sodeyama, A. Nakata, Prediction of the coefficient of linear thermal expansion for the amorphous homopolymers based on chemical structure using machine learning, Sci. Technol. Adv. Mater.: Methods 1 (2021) 213–224, https://doi.org/10.1080/27660400.2021.1993729.

[32] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, Proceedings of the 34th International Conference on Machine Learning (ICML) 70 (2017) 1263–1272. ⟨https://arxiv.org/abs/1704.01212⟩.

[33] A.T.N. Fajar, A.D. Hartono, R.M. Moshikur, M. Goto, Ionic liquids curated by machine learning for metal extraction, ACS Sustain Chem. Eng. 10 (2022) 12698–12705, https://doi.org/10.1021/acssuschemeng.2c03480.

[34] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, Y. Gal, AI models collapse when trained on recursively generated data, Nature 631 (2024) 755–759, https://doi.org/10.1038/s41586-024-07566-y.

[35] F. Weisshar, A. Gau, J. Hack, N. Maeda, D.M. Meier, Toward carbon dioxide capture from the atmosphere: lowering the regeneration temperature of polyethylenimine-based adsorbents by ionic liquid, Energy Fuels 35 (2021) 9059–9062, https://doi.org/10.1021/acs.energyfuels.1c00392.

[36] J. Hack, S. Frazzetto, L. Evers, N. Maeda, D.M. Meier, Branched versus linear structure: lowering the CO2 desorption temperature of polyethylenimine-functionalized silica adsorbents, Energies 15 (2022), https://doi.org/10.3390/en15031075.

[37] C.Y. Jia, J.Y. Li, G.F. Hao, G.F. Yang, A drug-likeness toolbox facilitates ADMET study in drug discovery, Drug Discov. Today 25 (2020) 248–258, https://doi.org/10.1016/j.drudis.2019.10.014.

[38] J. Lyu, S. Wang, T.E. Balius, I. Singh, A. Levit, Y.S. Moroz, M.J. O'Meara, T. Che, E. Algaa, K. Tolmachova, A.A. Tolmachev, B.K. Shoichet, B.L. Roth, J.J. Irwin, Ultra-large library docking for discovering new chemotypes, Nature 566 (2019) 224–229, https://doi.org/10.1038/s41586-019-0917-9.

[39] S. Pushpakom, F. Iorio, P.A. Eyers, K.J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, A. Norris, P. Sanseau, D. Cavalla, M. Pirmohamed, Drug repurposing: progress, challenges and recommendations, Nat. Rev. Drug Discov. 18 (2018) 41–58, https://doi.org/10.1038/nrd.2018.168.

[40] P. Ertl, A. Schuffenhauer, Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions, J. Chemin.-. 1 (2009), https://doi.org/10.1186/1758-2946-1-8.