SUPPORTING INFORMATION

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

Adroit T.N. Fajar[1,*], Guillaume Lambard[2], Md. Amirul Islam[3], Bidyut B. Saha[3], Zakiah D. Nurfajrin[4], Kevin Septioga[4]

[1]Center for Energy Systems Design (CESD), International Institute for Carbon-Neutral Energy Research (WPI-I2CNER), Kyushu University, 744 Motooka, Fukuoka 819-0395, Japan.
[2]Data-driven Materials Design Group, Center for Basic Research on Materials, National Institute for Materials Science, Namiki 1-1, Tsukuba 305-0044, Japan.
[3]International Institute for Carbon-Neutral Energy Research (WPI-I2CNER), Kyushu University, 744 Motooka, Fukuoka 819-0395, Japan.
[4]Department of Applied Chemistry, Graduate School of Engineering, Kyushu University, 744 Motooka, Fukuoka 819-0395, Japan.

[*]Corresponding Author. Email address: adroit@i2cner.kyushu-u.ac.jp

This supporting information file contains 14 pages, including additional descriptions of the methods (six notes), five figures, and one table.

Note S1

Data T0. Unlabeled data of ionic liquid (IL) structures was gathered from several data sources. All ILs were converted into the Simplified Molecular Input Line Entry System (SMILES) by concatenating the cation and anion components with a dot character. The concatenated SMILES strings were then transformed into their canonical forms using RDKit, and duplicate entries were removed. This process resulted in a dataset of 3,109 unique IL SMILES.

Data sources:

- Fan et al., 2024, Sci. Total Environ., 908, 168168.
- Chen et al., 2024, AIChE J., 70, e18392.
- Bakhtyari et al., 2023, Sci. Rep., 13, 12161.
- Li et al., 2023, Nat. Commun., 14, 2789.
- Liu et al., 2023, AIChE J., 69, e18182.
- Liu et al., 2023, J. Mol. Liq., 390, 122972.
- Boualem et al., 2022, J. Mol. Liq., 368, 120610.
- Chen et al., 2022, J. Mol. Liq., 350, 118546.
- Dhakal and Shah, 2022, Mol. Syst. Des. Eng., 7, 1344-1353.
- Duong et al., 2022, J. Chem. Phys., 156, 150-160.
- Cai et al., 2021, Desalination, 509, 115073.
- Carreira et al., 2021, Fluid Phase Equilib., 542, 113091.
- Nancarrow et al., 2021, Energy, 220, 119761.
- Makarov et al., 2021, J. Mol. Liq., 344, 117722.
- Lim et al., 2021, Sep. Purif. Technol., 258, 118019.
- Chen et al., 2021, Sep. Purif. Technol., 259, 118204.
- Li et al., 2021, Sep. Purif. Technol., 277, 119471.
- Shi et al., 2020, J. Mol. Liq., 304, 112756.
- Gras et al., 2020, ACS Sustain. Chem. Eng., 8, 15865-15874.
- Bui et al., 2020, Korean J. Chem. Eng., 37, 2262-2272.
- Tampucci et al., 2020, Pharmaceutics, 12, 1078.
- Kang et al., 2020, J. Hazard. Mater., 397, 122761.
- Kusumahastuti et al., 2019, Ecotoxicol. Environ. Saf., 172, 556-565.
- Parajó et al., 2019, Ecotoxicol. Environ. Saf., 184, 109580.
- Delgado-Mellado et al., 2019, SN Appl. Sci., 1, 1-9.
- Boudesocque et al., 2019, Sep. Purif. Technol., 210, 824-834.
- Shi, Jing, and Jia, 2016, J. Mol. Liq., 215, 640-646.
- Montalbán, Víllora, and Licence, 2018, Ecotoxicol. Environ. Saf., 150, 129-135.
- Biczak et al., 2018, Ecotoxicol. Environ. Saf., 155, 37-42.
- Diaz et al., 2018, Ecotoxicol. Environ. Saf., 162, 29-34.
- Ghanem et al., 2018, Chemosphere, 195, 21-28.
- Katsuta and Tamura, 2018, J. Solut. Chem., 47, 1293-1308.
- Sintra et al., 2017, Ecotoxicol. Environ. Saf., 143, 315-321.
- Shi, Jing, and Jia, 2017, Russ. J. Phys. Chem. A, 91, 692-696.
- Zarrougui et al., 2017, Sep. Purif. Technol., 175, 87-98.

- Rantamäki et al., 2017, Sci. Rep., 7, 46673.
- Panigrahi et al., 2016, Sep. Purif. Technol., 171, 263-269.
- Montalbán et al., 2016, Chemosphere, 155, 405-414.
- Papaiconomou et al., 2016, ChemistrySelect, 1, 3892-3900.
- Chen et al., 2015, ACS Sustain. Chem. Eng., 3, 3167-3174.
- Costa et al., 2015, J. Hazard. Mater., 284, 136-142.
- Ghanem et al., 2015, J. Mol. Liq., 212, 352-359.
- Hernández-Fernández et al., 2015, Ecotoxicol. Environ. Saf., 116, 29-33.
- Rout and Binnemans, 2014, Ind. Eng. Chem. Res., 53, 6500-6508.
- Ventura et al., 2014, Ecotoxicol. Environ. Saf., 102, 48-54.
- Das and Roy, 2014, Chemosphere, 104, 170-176.
- Onghena et al., 2014, Dalton Trans., 43, 11566-11578.
- Peric et al., 2013, J. Hazard. Mater., 261, 99-105.
- Izadiyan et al., 2013, Ecotoxicol. Environ. Saf., 87, 42-48.
- Viboud et al., 2012, J. Hazard. Mater., 215, 40-48.
- Hossain et al., 2011, Chemosphere, 85, 990-994.
- Alvarez-Guerra and Irabien, 2011, Green Chem., 13, 1507-1516.
- Luis, Garea, and Irabien, 2010, J. Mol. Liq., 152, 28-33.
- Samorì et al., 2007, Environ. Toxicol. Chem., 26, 2379-2382.
- Couling et al., 2006, Green Chem., 8, 82-90.
- Ranke et al., 2004, Ecotoxicol. Environ. Saf., 58, 396-404.

Data T1. Data on IL structures labeled with corresponding CO_2 solubility (mmol/mol) values at specific temperatures and pressures were collected from various sources. From this raw dataset, only ILs with CO_2 solubility values measured near ambient temperature (258–323 K) and pressure (40–200 kPa) were included. This selection process resulted in a dataset of 564 IL entries.

Data sources:

- Liu, Tianxiong, et al., 2023, AIChE Journal, 69.10, e18182.
- Liu, Zongyang, et al, 2023, Journal of Molecular Liquids, 391, 123308.
- Song, Zhen, et al., 2020, Chemical Engineering Science, 223, 115752.

Data T2. Data on IL structures labeled with their corresponding eco-toxicity levels (EC_{50} values in μM) was collected from our previous study, comprising 110 entries. Further details on data collection are available at DOI: 10.1021/acssuschemeng.2c03480

- Fajar et al., 2022, ACS Sustainable Chem. Eng., 10, 12698.

Note S2

Test loss. The test loss used here is cross-entropy loss (also known as negative log-likelihood loss), calculated between the model's predicted probability distribution and the actual distribution of the target tokens. For each token in the test dataset, the model predicts a probability distribution over the vocabulary, and the loss measures the deviation of these predictions from the actual tokens, as described in Equation (*S*1). After fine-tuning the GPT-2 model on the IL dataset, the model exhibited a test loss of 0.12, which is very low. This suggests that the model's predicted probability distributions are closely aligned with the actual distributions of the target tokens in the test dataset, meaning it assigns higher probabilities to the correct next tokens.

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} logP_{model(w_i|w_{<1})} \qquad (S1)$$

Where:

- N is the total number of tokens.
- w_i is the i-th token.
- $w_{<1}$ are the preceding tokens.
- $P_{model(w_i|w_{<1})}$ is the probability the model assigns to the correct next token.

GPT-2 model. GPT-2 is a generative language model based on the transformer architecture, which uses stacks of decoder blocks consisting of masked multi-head self-attention layers, position-wise feed-forward networks, and layer normalization. Each decoder block processes the input sequence in parallel, leveraging self-attention to learn dependencies between tokens—here, the atomic and structural symbols in SMILES strings. The core of GPT-2's ability to model sequences lies in the self-attention mechanism, as described in Equation (*S*2), where Q, K, and V are the query, key, and value matrices derived from the input embeddings, and d_k is the dimensionality of the keys. This allows the model to weigh the relevance of each token with respect to others in the sequence. Masked self-attention ensures that the model can only attend to past tokens during generation, preserving causality.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (S2)$$

In this study, SMILES strings representing ILs were tokenized and used as input for fine-tuning the pretrained GPT-2 model. The model learned the syntactic patterns and chemical substructures common in the training data. Once fine-tuned, the model was used to generate new ILs by providing an initial token (here, [PAD]), after which the model sampled one token at a time until a complete SMILES string was produced. This generation process is autoregressive, meaning that each new token is generated based on the previously generated ones. The resulting SMILES were validated using RDKit to ensure chemical correctness—invalid, duplicate, and syntactically incorrect molecules were discarded, and only unique, valid IL structures were retained for further analysis.

Note S3

Zero-cost geometry optimization. To efficiently identify the best-performing architecture for the SMILES-X prediction models, we employed a zero-cost geometry optimization approach. This method allows the selection of optimal hyperparameter combinations—such as embedding size, number of LSTM units, and number of dense layer units—without requiring full training for each candidate architecture. Instead, a lightweight evaluation (e.g., initial loss or proxy score from a small batch or early training step) is used to approximate model performance. This significantly reduces computational cost while effectively guiding the search toward promising architectures.

TOPSIS method. In this study, the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) was employed to rank the generated ILs based on two criteria: CO_2 solubility (ML-1) and IL eco-toxicity (ML-2). The first step involved normalizing the criteria using z-score standardization to remove the influence of scale, as shown in Equation (*S*3); where x_{ij} is the original value, μ_j is the mean, and σ_j is the standard deviation of criterion j. Next, a weighted normalized decision matrix was constructed with equal weights (w_j, 0.5 for both criteria), given by Equation (*S*4). The ideal (best) and negative-ideal (worst) solutions were then identified by selecting the maximum and minimum values, respectively, of the weighted normalized values for each criterion. Distances to the ideal, D_i^+, and negative-ideal, D_i^-, solutions were calculated using the Euclidean distance formula, as shown in Equations (*S*5) and (*S*6). Finally, the relative closeness to the ideal solution for each IL was determined using Equation (*S*7). This relative closeness score, C_i, was used to rank the ILs, with higher values indicating closer proximity to the ideal solution.

$$x_{ij}^{norm} = \frac{x_{ij} - \mu_j}{\sigma_j} \qquad (S3)$$

$$x_{ij}^{weighted} = x_{ij}^{norm} \cdot w_j \qquad (S4)$$

$$D_i^+ = \sqrt{\sum_{j=1}^{n} \left(x_{ij}^{weighted} - x_j^{ideal} \right)^2} \qquad (S5)$$

$$D_i^- = \sqrt{\sum_{j=1}^{n} \left(x_{ij}^{weighted} - x_j^{negative\ ideal} \right)^2} \qquad (S6)$$

$$C_i = \frac{D_i^+}{D_i^+ + D_i^-} \qquad (S7)$$

Note S4

S-E score. The combined score for CO_2 solubility and eco-toxicity (S-E score) was calculated as the normalized dot product of the predicted CO_2 solubility (S_{CO_2}) and predicted $logEC_{50}$ values, as described in Equation (*S*8). It is important to note that the S-E score and TOPSIS ranking serve different but complementary purposes in evaluating ILs. The S-E score provides a straightforward combined assessment of solubility and toxicity, while TOPSIS enables a more nuanced prioritization by applying specific selection weights to each property.

$$SE = \frac{S_{CO_2} \cdot logEC_{50}}{1000} \qquad (S8)$$

Note S5

Similarity search. To identify commercially available ILs with structural similarity to the top generated ILs, a similarity search was conducted using the Tanimoto index. Two datasets were prepared: the top 1,000 ILs (ranked by S-E score) from the cumulative generated ILs up to cycle 4 (Dataset 1) and a list of 337 commercially available ILs (sourced from the iolitec, Merck, and TCI product catalogs). SMILES representations of ILs from Dataset 1 and Dataset 2 were converted into molecular fingerprints using Morgan fingerprints with a radius of 2 and a 2048-bit vector length. Pairwise Tanimoto similarities between fingerprints from the two datasets were then computed. The Tanimoto similarity, defined in Equation (*S*9), quantifies structural overlap between two fingerprint bit vectors, A and B, yielding a score between 0 (no similarity) and 1 (identical). A threshold of 0.7 was set to identify high-similarity pairs, and all pairs exceeding this threshold were stored for further analysis.

$$S = \frac{A \cap B}{A \cup B} \qquad (S9)$$

Note S6

SA score. The synthetic accessibility (SA) score is a computational metric used to estimate the ease of synthesizing a molecule based on its structural complexity and fragment contributions. It integrates knowledge of molecular fragments from large chemical databases and penalizes structural features associated with synthetic difficulty, such as rare substructures, high molecular complexity, and the presence of stereocenters. The score is calculated as defined in Equation ($S10$), where c represents the molecular complexity factor derived from properties such as the number of atoms, rings, and bonds; f is the average fragment contribution determined by the frequency of molecular fragments in a chemical database; and a accounts for stereochemical complexity. The SA score ranges from approximately 1 (highly accessible, easy to synthesize) to 10 (low accessibility, difficult to synthesize). This approach provides a practical estimate for prioritizing molecules in cheminformatics workflows, particularly in drug discovery and material design.

$$SA = c + f - a \qquad (S10)$$

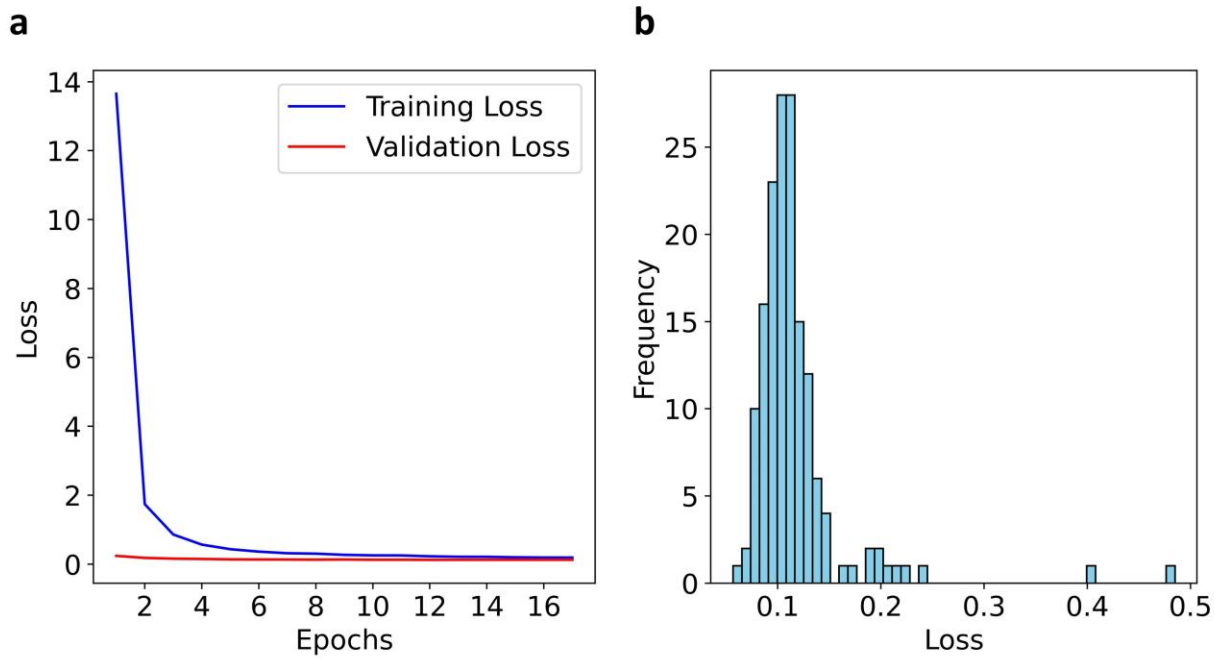

Figure S1. (a) The learning curve (training vs. validation loss) during model fine-tuning and (b) the distribution of test losses.

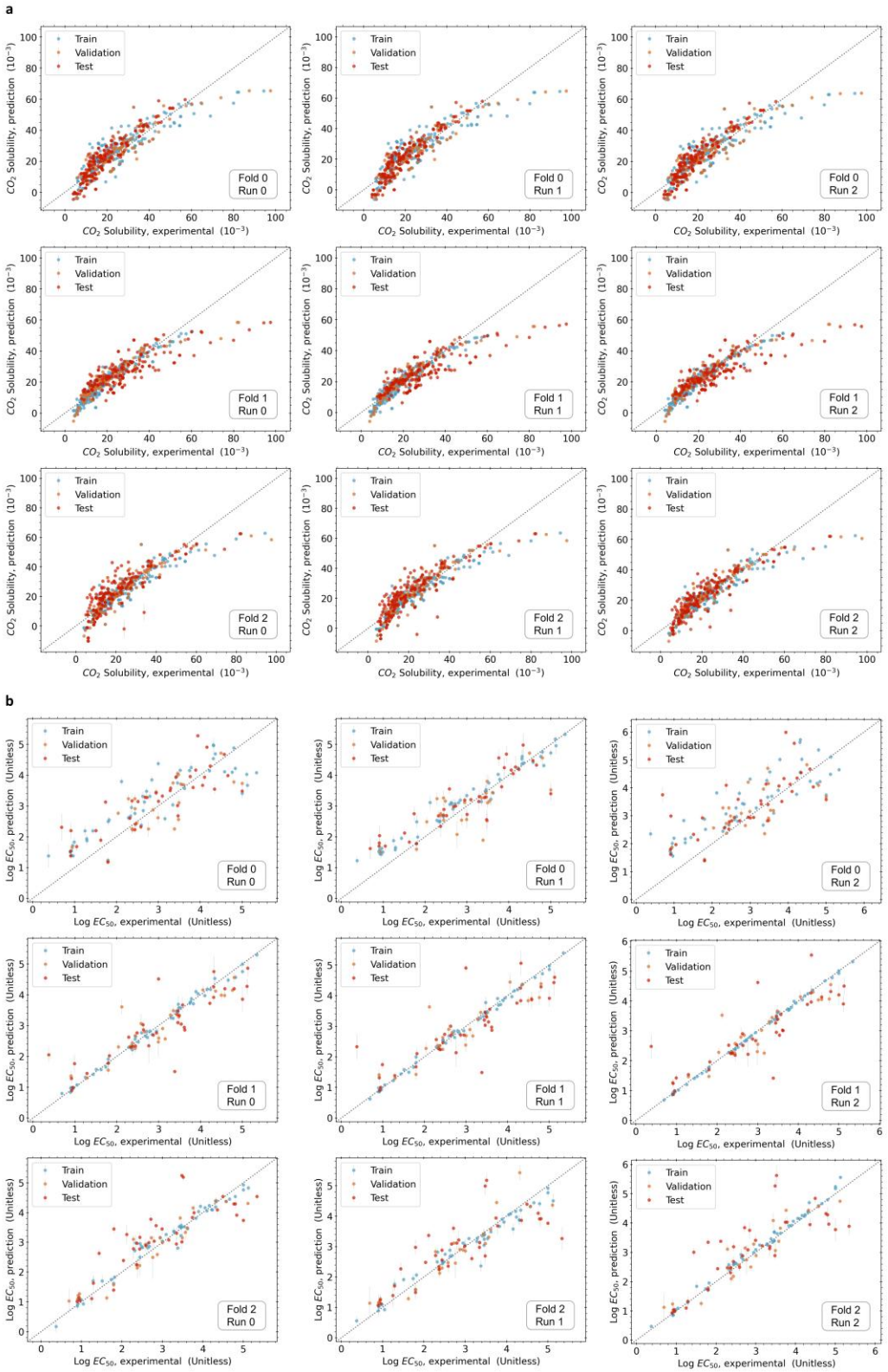

Figure S2. Parity plots for each run in each fold during the training of SMILES-X with (a) Data T1 (CO_2 solubility) and (b) Data T2 (IL eco-toxicity; EC_{50}).

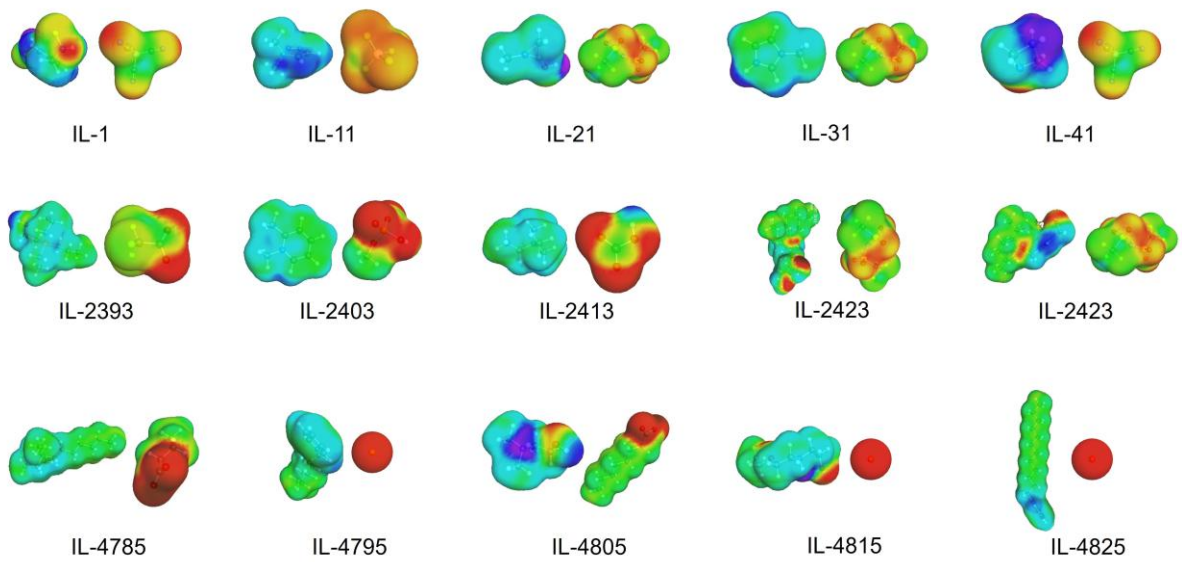

Figure S3. COSMO-RS visualizations (COSMO views) of 15 representative IL structures listed in Table 1, illustrating the spatial distribution of electronic charge for ILs ranked at the top, middle, and bottom positions in the TOPSIS ranking.

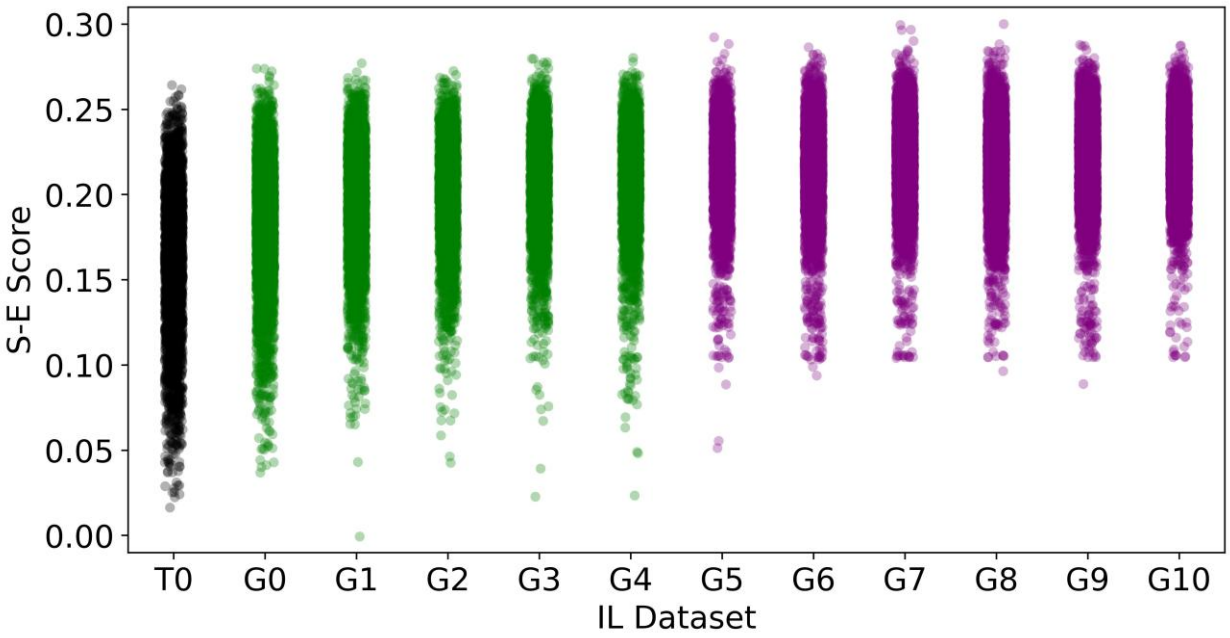

Figure S4. Combined CO_2 solubility and eco-toxicity (S-E) scores of the original training ILs (Data T0) and generated ILs through cycle 10 (Data G0–G10).

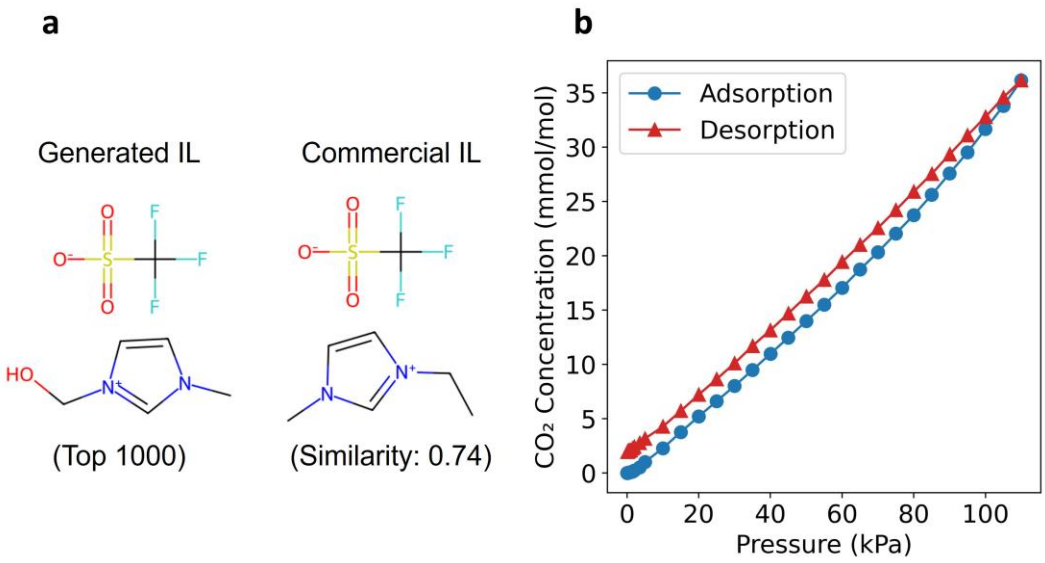

Figure S5. (a) Structural comparison of a generated IL and a commercial IL with 74% similarity, namely 1-ethyl-3-methylimidazolium triflate. (b) Experimentally measured CO_2 adsorption-desorption behavior of the commercial IL with a relatively low similarity score.

Table S1. Commercially available ILs with similarity scores ≥ 0.7 to the top 1,000 generated ILs through cycle 4.

Generated ILs	Commercial ILs	Similarity	Product Name
CnIcc[n+](CCO)c1.O=S(=O)(F)[N-]S(=O)(=O)C(F)(F)F	CnIcc[n+](CCO)c1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.87	1-(2-Hydroxyethyl)-3-methylimidazolium bis(trifluoromethylsulfonyl)imide
CCnIcc[n+](CC)c1.O=S(=O)(F)[N-]S(=O)(=O)C(F)(F)F	CCnIcc[n+](CC)c1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.86	1,3-Diethylimidazolium bis(trifluoromethylsulfonyl)imide
CC[n+]1ccn(C)c1.O=S(=O)(F)[N-]S(=O)(=O)C(F)(F)F	CC[n+]1ccn(C)c1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.86	1-Ethyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide
CCnIcc[n+](CC)c1.NS(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	CCnIcc[n+](CC)c1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.84	1,3-Diethylimidazolium bis(trifluoromethylsulfonyl)imide
CnIcc[n+](CO)c1.O=S(=O)([N-]S(=O)(=O)C(F)(F)F)C(F)(F)F	CnIcc[n+](CCO)c1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.79	1-(2-Hydroxyethyl)-3-methylimidazolium bis(trifluoromethylsulfonyl)imide
CC[NH+]1C=CN=C1.O=S(=O)([N-]S(=O)(=O)C(F)(F)F)C(F)F	CC[NH+]1C=CN=C1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.79	1-Ethylimidazolium bis(trifluoromethylsulfonyl)imide
CC[n+]1ccn(C)c1.O=S(=O)(F)[N-]S(=O)(=O)C(F)(F)F	CC[n+]1ccn(C)c1.[N-](S(=O)(=O)F)S(=O)(=O)F	0.78	1-Ethyl-3-methylimidazolium bis(fluorosulfonyl)imide
CnIcc[n+](CP)c1.O=S(=O)([N-]S(=O)(=O)C(F)(F)F)C(F)(F)F	CC[n+]1ccn(C)c1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.76	1-Ethyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide
CnIcc[n+](CS)c1.O=S(=O)([N-]S(=O)(=O)C(F)(F)F)C(F)(F)F	CC[n+]1ccn(C)c1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.76	1-Ethyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide
CnIcc[n+](CO)c1.O=S(=O)([N-]S(=O)(=O)C(F)(F)F)C(F)(F)F	CC[n+]1ccn(C)c1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.76	1-Ethyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide
CnIcc[n+](CO)c1.O=S(=O)([O-])C(F)(F)F	CC[n+]1ccn(C)c1.[O-][S](=O)(=O)C(F)(F)F	0.74	1-Ethyl-3-methylimidazolium triflate
CnIcc[n+](CO)c1.O=S(=O)(F)[N-]S(=O)(=O)F	CC[n+]1ccn(C)c1.[N-](S(=O)(=O)F)S(=O)(=O)F	0.74	1-Ethyl-3-methylimidazolium bis(fluorosulfonyl)imide
CC[n+]1ccn(C)c1.O=S(=O)(F)[N-]S(=O)(=O)C(F)(F)F	CC[n+]1ccn(C)c1.C(C(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(C(F)(F)F)(F)F)(F)(F)F	0.73	1-Ethyl-3-methylimidazolium bis(pentafluoroethylsulfonyl)imide
NC[NH+]1C=CN=C1.O=S(=O)([N-]S(=O)(=O)C(F)(F)F)C(F)(F)F	CC[NH+]1C=CN=C1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.72	1-Ethylimidazolium bis(trifluoromethylsulfonyl)imide
O=S(=O)([N-]S(=O)(=O)C(F)(F)F)C(F)(F)F.OC[NH+]1C=CN=C1	CC[NH+]1C=CN=C1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.72	1-Ethylimidazolium bis(trifluoromethylsulfonyl)imide
CC[NH+]1C=CN=C1.O=S(=O)([O-])C(F)(F)F	CC[NH+]1C=CN=C1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.71	1-Ethylimidazolium bis(trifluoromethylsulfonyl)imide
CCnIcc[n+](F)c1.O=S(=O)([N-]S(=O)(=O)C(F)(F)F)C(F)(F)F	CCnIcc[n+](CC)c1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.71	1,3-Diethylimidazolium bis(trifluoromethylsulfonyl)imide
CC1=NC=C[NH+]1C.O=S(=O)([O-])C(F)(F)F	C[NH+]1C=CN=C1C.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.71	1,2-Dimethylimidazolium bis(trifluoromethylsulfonyl)imide
C[n+]1ccn(CO)c1.O=S(=O)([O-])C(F)(F)F	CCCnIcc[n+](C)c1.[O-][S](=O)(=O)C(F)(F)F	0.70	1-Methyl-3-propylimidazolium trifluoromethanesulfonate
CnIcc[n+](CP)c1.O=S(=O)([N-]S(=O)(=O)C(F)(F)F)C(F)(F)F	CnIcc[n+](CCO)c1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.70	1-(2-Hydroxyethyl)-3-methylimidazolium bis(trifluoromethylsulfonyl)imide
CnIcc[n+](CS)c1.O=S(=O)([N-]S(=O)(=O)C(F)(F)F)C(F)(F)F	CnIcc[n+](CCO)c1.C(F)(F)(F)S(=O)(=O)[N-]S(=O)(=O)C(F)(F)F	0.70	1-(2-Hydroxyethyl)-3-methylimidazolium bis(trifluoromethylsulfonyl)imide
C[n+]1ccn(CCC#N)c1.O=S(=O)([O-])C(F)(F)F	CCCnIcc[n+](C)c1.[O-][S](=O)(=O)C(F)(F)F	0.70	1-Methyl-3-propylimidazolium trifluoromethanesulfonate