Supplementary Information

An accurate and interpretable deep learning model for yield prediction using hybrid molecular representations

Yingying Wang[abc], Xinyi Sun[abc], Yuanyuan Li*[abc], Li Wang*[abc] and Jinglai Zhang*[abc]

[a]*Henan Key Laboratory of Protection and Safety Energy Storage of Light Metal Materials, Henan University, Kaifeng, Henan 475004, China;*
[b]*Henan Province Engineering Research Center of Green Anticorrosion Technology for Magnesium Alloys, Henan University, Kaifeng, Henan 475004, China;*
[c]*College of Chemistry and Molecular Sciences, Henan University, Kaifeng, Henan 475004, China;*

Feature selection process

Prior to model construction, the dataset's molecular fingerprints undergo initial processing to reduce the dimension of the descriptors. A large number of redundant features with identical numerical distributions are removed from the molecular fingerprints, as these features do not contribute meaningfully to machine learning training models. As a result, the features of these descriptors have been considerably simplified.

Model input layers

In this study, three widely used molecular fingerprints, Molecular Access System (MACCS), Extended-Connectivity Fingerprint (ECFP), and Functional Connectivity Fingerprint (FCFP), are utilized as descriptors. The distinct molecular fingerprints are utilized as separate input layers rather than being integrated into a unified input layer, primarily for

*Corresponding authors
E-mail addresses: yylichem@henu.edu.cn (Y. Li), chemwangl@henu.edu.cn (L. Wang), zhangjinglai@henu.edu.cn (J. Zhang).

the following reasons:

Feature diversity and uniqueness: Distinct molecular fingerprints capture different aspects of molecular structure information. For instance, MACCS captures functional group information within molecules; ECFP reflects the local environment and connectivity of each atom within a specified radius; FCFP, based on ECFP, disregards distinctions between similar functional groups and represents a more generalized functional molecular fingerprint. Separate input layers better preserve each fingerprint's unique characteristics and prevent potential information overlap or masking that could arise if combined into a single layer.

Model flexibility and adaptability: The separate input layer design allows the network architecture to be tailored for each type of fingerprint. For example, in this study, hidden layers with varying numbers of nodes are established for each distinct input layer. This flexibility enables the model to more effectively adapt to different types of input data, thereby enhancing the overall performance.

Avoiding information confusion: Combining all fingerprints into a unified input layer may cause information confusion. For example, high-dimensional features of certain fingerprints might overshadow critical information from other fingerprints. By separating the input layers, such mutual interference can be avoided, allowing the model to learn the characteristics of each fingerprint more distinctly.

Reaction condition optimization process

The reaction conditions are optimized through deep learning model. The CO_2 pressure is always maintained at 0.1 MPa. The time range is set to [1 h, 2 h, 3 h, 4 h, 5 h], and the

temperature range is set to [50 ℃, 60 ℃, 70 ℃, 80 ℃, 90 ℃]. The catalyst dosage range is considered as [0.2 mol%, 0.4 mol%, 0.6 mol%, 0.8 mol%, 1.0 mol%, 1.2 mol%, 1.4 mol%]. The initial reaction conditions are 4 h, 90 ℃, 0.1 MPa, 0.8%. Not all combinations of reaction conditions are traversed, but the control variable method from experiments is employed to determine parameters one by one. With the aim of reducing reaction conditions as much as possible while ensuring a predicted yield close to 0.95, the reaction conditions are optimized by yield prediction through deep learning.

General applicability and potential limitations of the model

The deep learning approach presented in this study offers a novel and efficient alternative for predicting and designing IL catalysts for CO_2 cycloaddition reactions. While the current study focuses on imidazolium-based and pyrazolium-based ILs, the approach can be extended to other types of molecules and reactions with appropriate modifications. The use of hybrid fingerprint features allows the model to capture essential structural information for predicting catalytic performance. Additionally, the model's ability to handle high-dimensional data and complex relationships, combined with SHAP analysis and a two-step screening strategy, provides valuable insights and reduces experimental burden.

Despite promising results, the model's generalizability is limited by the dataset's focus on imidazolium-based and pyrazolium-based ILs. Extending the approach to other molecules or reactions will require more diverse training data. Additionally, the model's adaptability to more complex or less studied catalytic systems remains to be validated and may need further refinement. Future work should aim to expand the dataset, validate the model on additional

systems, and optimize computational efficiency to enhance its broader applicability.

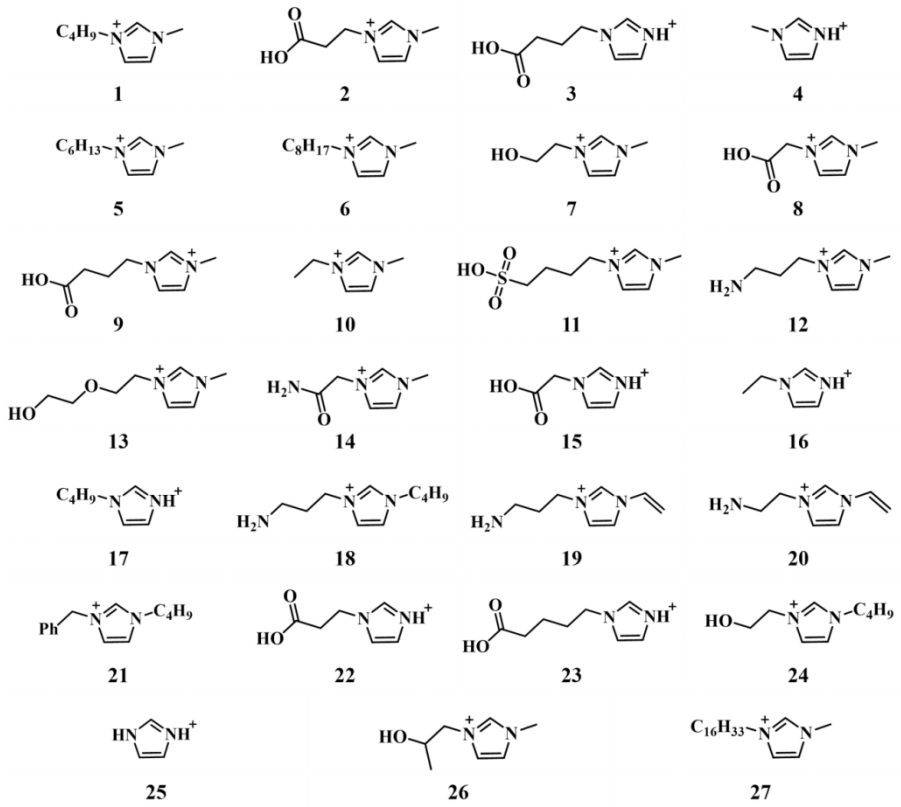

Fig. S1. 27 types of imidazolium-based cation structures in the dataset.

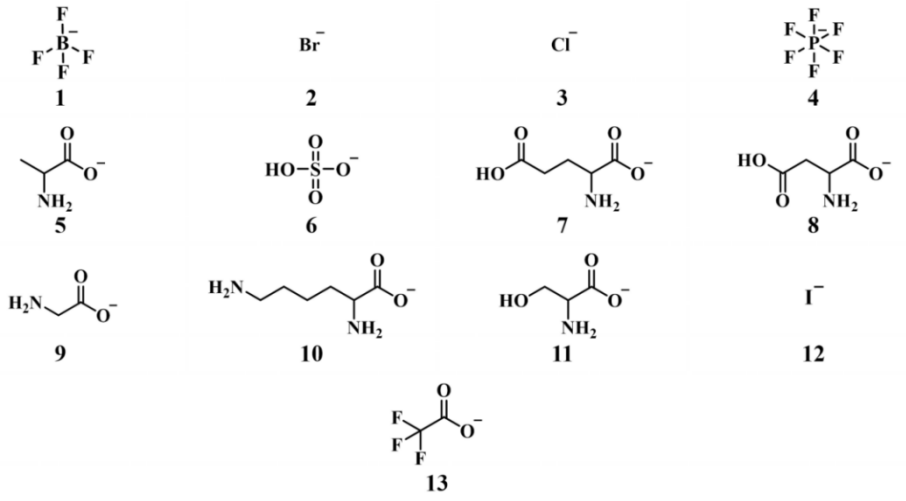

Fig. S2. 13 types of anion structures associated with imidazolium-based ionic liquids in the dataset.

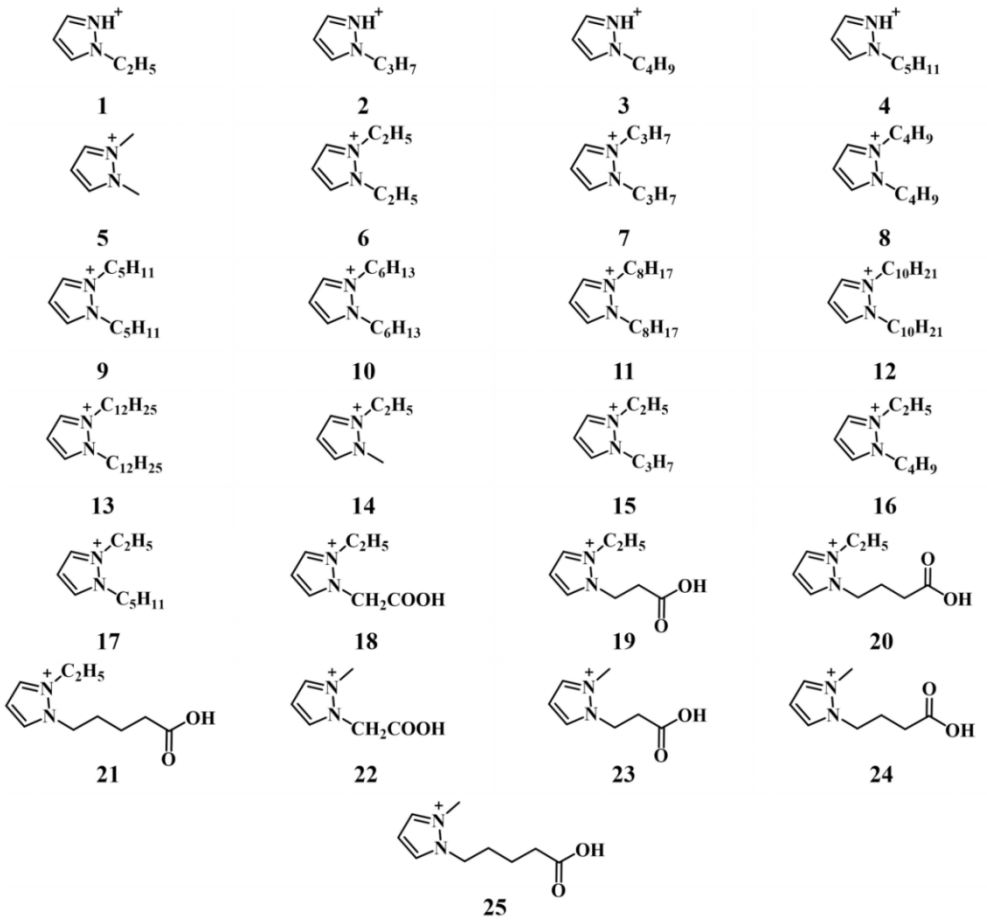

Fig. S3. 25 types of pyrazolium-based cation structures in the dataset.

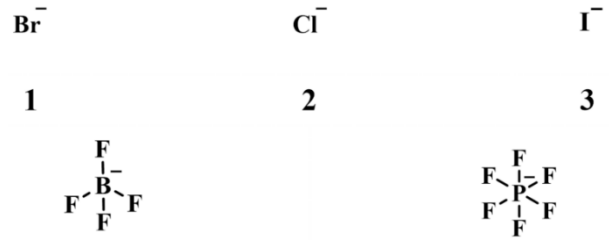

Fig. S4. 5 types of anion structures associated with pyrazolium-based ionic liquids in the dataset.

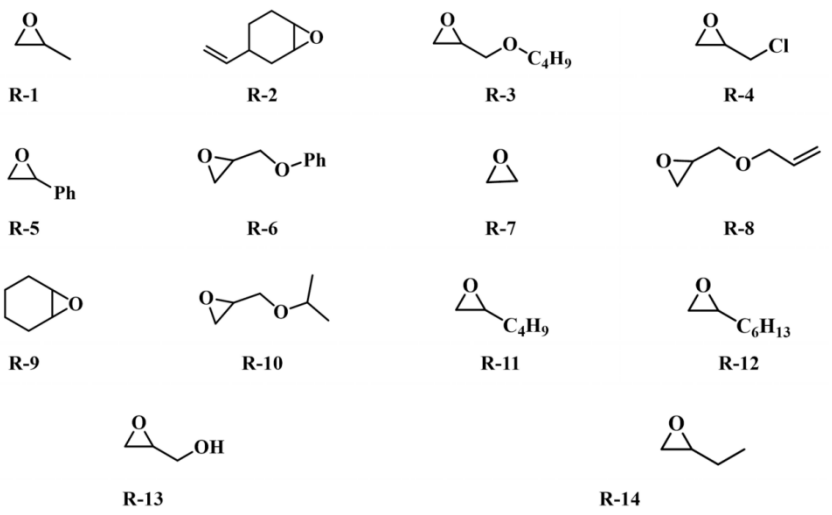

Fig. S5. 14 types of reactant structures in the dataset.

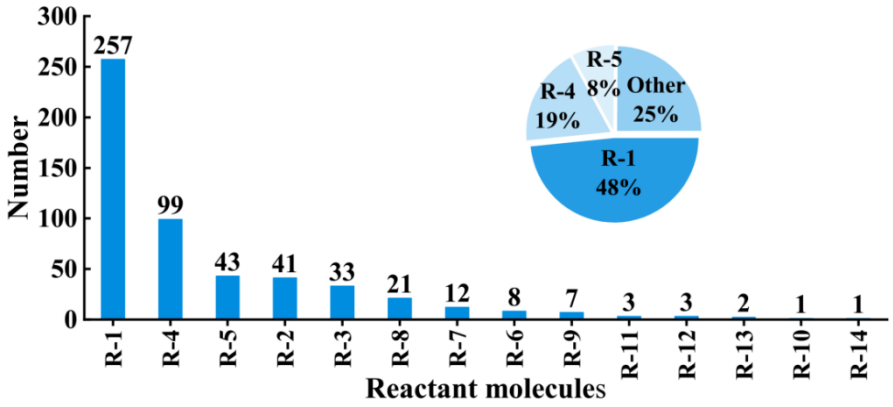

Fig. S6. Number of data points for each of the 14 types of reactants in the dataset.

Table S1

Parameters of DNN models.

Model	Random seed	Learning rate	Epochs
M	28	0.006	310
E	29	0.005	374
F	6	0.004	486
ME	3	0.007	715
MF	25	0.007	390
EF	2	0.007	262
MEF	3	0.007	385

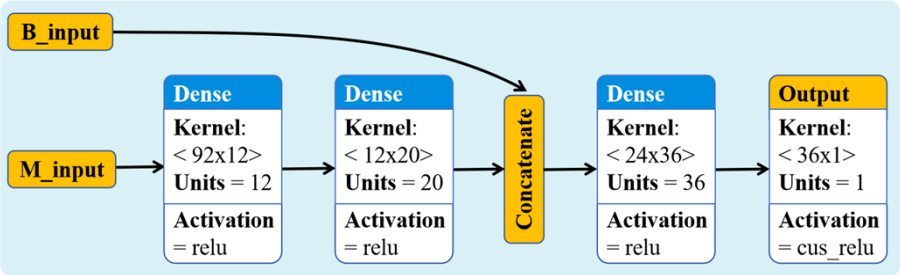

Fig. S7. DNN model based on M feature.

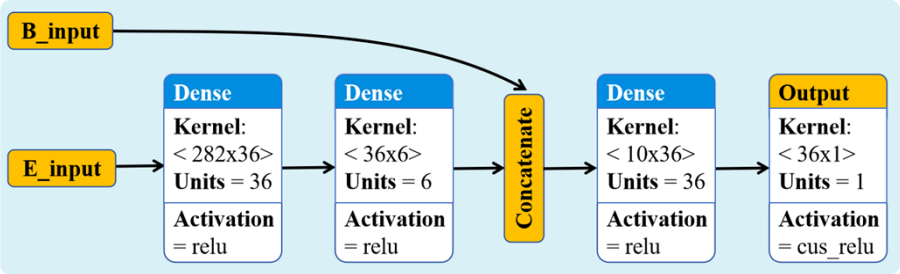

Fig. S8. DNN model based on E feature.

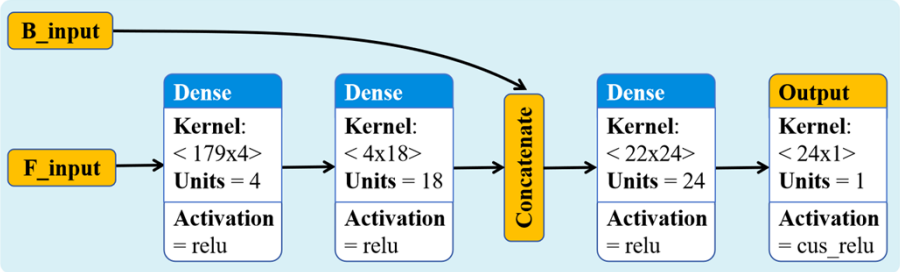

Fig. S9. DNN model based on F feature.

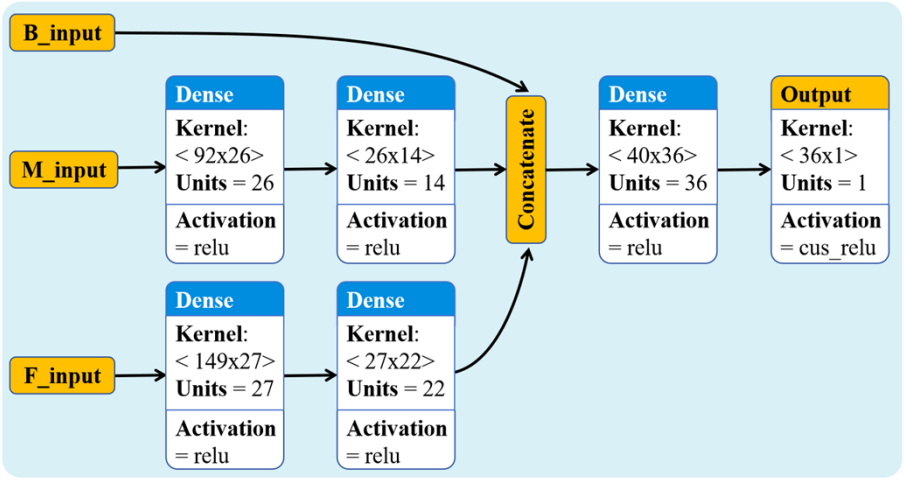

Fig. S10. DNN model based on MF feature.

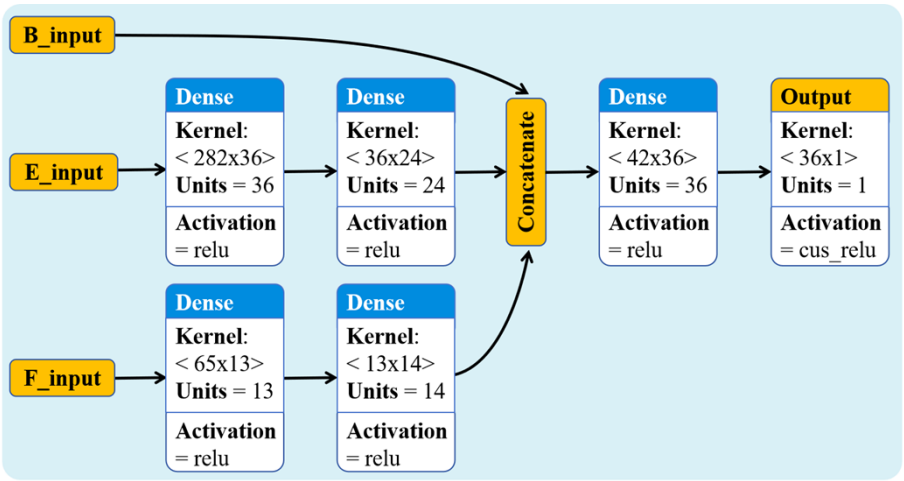

Fig. S11. DNN model based on EF feature.

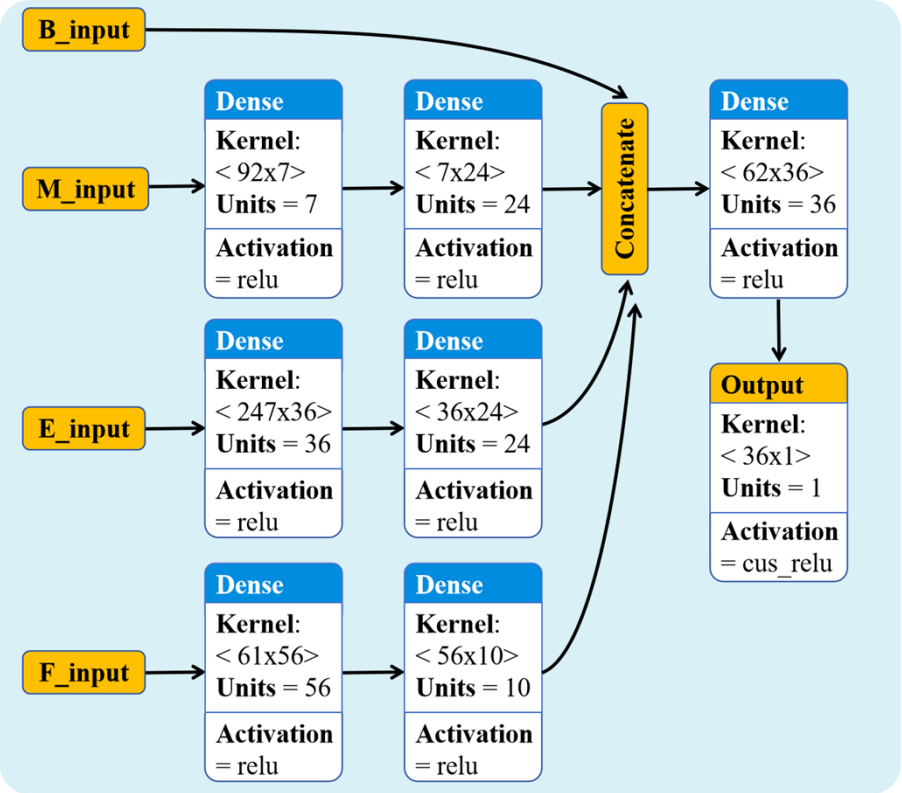

Fig. S12. DNN model based on MEF feature.

Table S2

Parameters of traditional algorithms.

Algorithms	M
DT	random_state=35, max_depth=24
RF	max_depth=17, random_state=16, max_features=84, n_estimators=11
GBR	learning_rate=0.396, random_state=7, n_estimators=200
XGB	random_state=0, max_depth=2, n_estimators=81
SVM	Kernel='rbf', C=470, epsion=1e-06, gamma=0.01

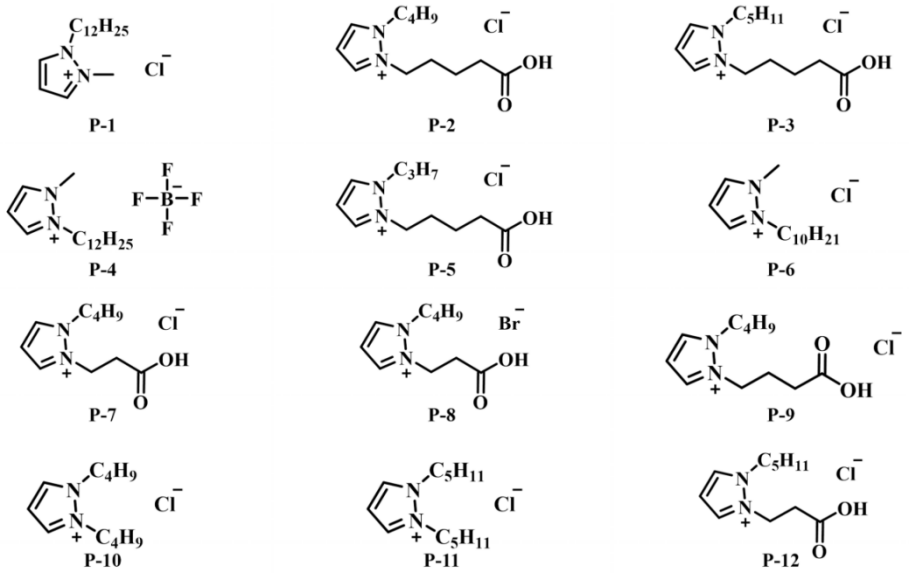

Fig. S13. The structures of screened pyrazolium-based ionic liquids.

Table S3

The 12 screened pyrazolium-based ionic liquids.

Structure	Time (h)	T (℃)	P (MPa)	Amount (mol%)	Predicted yield
P-1	7	90	0.1	1	1.00
P-2	7	90	0.1	1	0.96
P-3	7	90	0.1	1	0.91
P-4	7	90	0.1	1	0.87
P-5	7	90	0.1	1	0.87
P-6	7	90	0.1	1	0.85
P-7	7	90	0.1	1	0.83
P-8	7	90	0.1	1	0.83
P-9	7	90	0.1	1	0.83
P-10	7	90	0.1	1	0.82
P-11	7	90	0.1	1	0.82
P-12	7	90	0.1	1	0.80

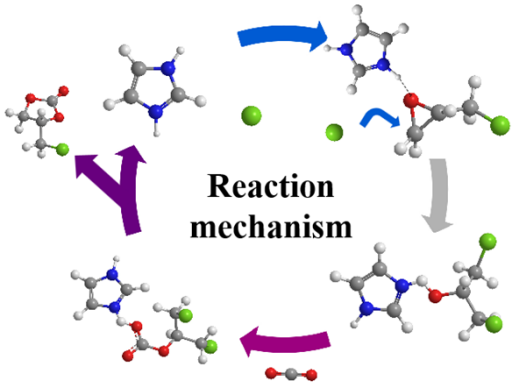

Fig. S14. The mechanism of the CO_2 cycloaddition reaction catalyzed by ionic liquids.

Table S4

The energy barriers, optimized reaction conditions, and corresponding yields of 19 imidazole-based ionic liquids.

ILs	Time (h)	T (°C)	P (MPa)	Amount (mol%)	Predicted yield	Energy barriers (kcal mol^{-1})
M-1	4	80	0.1	1.2	0.96	16.0
M-2	4	80	0.1	1.0	0.98	17.2
M-3	4	70	0.1	1.0	0.96	18.8
M-4	4	80	0.1	1.0	0.98	19.8
M-5	4	60	0.1	1.0	0.98	20.2
M-6	4	80	0.1	1.0	0.96	20.3
M-7	4	50	0.1	1.2	0.99	20.4
M-8	4	80	0.1	1.0	0.95	20.4
M-9	4	50	0.1	1.2	0.97	20.7
M-10	4	90	0.1	0.8	0.94	20.8
M-11	4	70	0.1	1.0	0.95	21.1
M-12	4	90	0.1	1.0	0.95	21.8
M-13	4	90	0.1	0.8	0.95	22.4
M-14	4	60	0.1	0.8	0.97	22.6
M-15	4	90	0.1	1.0	0.94	23.1
M-16	4	80	0.1	1.2	0.95	23.3

M-17	4	90	0.1	0.8	0.95	29.4
M-18	4	80	0.1	1.0	0.95	34.8
M-19	4	70	0.1	1.2	0.97	40.5

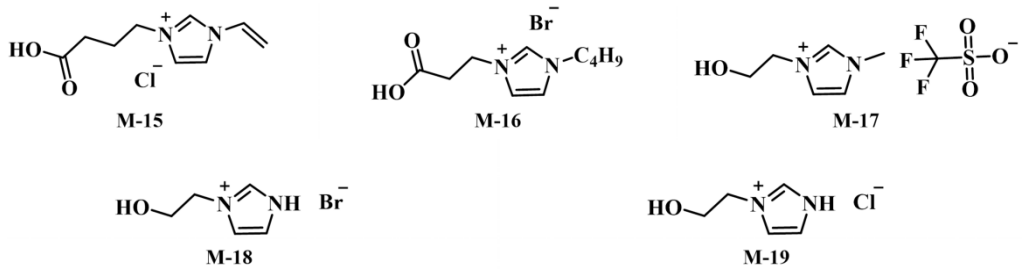

Fig. S15. The structures of 5 imidazole-based ionic liquids with energy barriers more than 23.0 kcal mol^{-1}.