

Screening High-Efficiency Deep Eutectic Solvents for Oxidative-Extractive Desulfurization via a Machine Learning Framework to Analyze Intermolecular Interactions

Changqing He ^{1,2,†}, Jiahao Xu ^{1,2,†}, Anupa Gunawardane ^{1,2}, Yang Xue ^{1,2}, Xueqian Liu ^{1,2}, Lin He ^{1,2,*} and Hong Sui ^{1,2}

¹ School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, China

² National Engineering Research Center of Distillation Technology, Tianjin 300072, China

* Correspondence: linhe@tju.edu.cn

† These authors contributed equally to this work.

S1 Mathematical Definitions

The mathematical definitions and their full names are shown in **Table S1**.

Table S1. The main mathematical definitions and full names.

Mathematical Definitions	Full Name	Mathematical Definitions	Full Name
ΔG	Solvation free energy	Prior_data	Absolute difference between molecular polarity descriptors of solute and solvent
Polarity_Descriptor_solute	Molecular polarity descriptor of the solute	Polarity_Descriptor_solvent	Molecular polarity descriptor of the solvent
RMSE	Root mean square error	R^2	Determination coefficient
y_i	Actual value	$\hat{y}_i$	Predicted value
$\bar{y}_i$	Mean value	n	Number of samples
D_{sub}	ΔG contribution of masked substructure	ND_{sub}	Normalized D_{sub}
ΔND_{sub}	ND_{sub} difference between target molecule and oil phase components	c_{before}	DBT concentration in oil phase before extraction
c_{after}	DBT concentration in oil phase after extraction	m_{initial}	Mass of initial DES
m_{recovery}	Mass of recovered DES	ρ	Bond critical point electron density
ΔE	Interaction energy	ΔE_{els}	Electrostatic energy
ΔE_{xrep}	Exchange repulsion energy	ΔE_{orb}	Orbital energy
ΔE_{disp}	Dispersion energy		

S2 DESs Library

Table S2 lists the SMILES strings for the hydrogen bond acceptor and donor components of the binary DES candidates.

Table S2. Hydrogen bond acceptor and donor components of the binary DES candidates.

Hydrogen Bond Acceptors	Hydrogen Bond Donors
[NH4+]	C(=O)(N)N
C[N+](C)(C)C	CNC(=O)N
CC[NH2+]CC	CNC(=O)NC
CC[N+](C)(CC)CC	CN(C)C(=O)N
CC[N+](CC)(CC)CC	CN(C)C(=O)N(C)C
CCCC[N+](CC)(CC)CC	C(=S)(N)N
CCC[N+](CCC)(CCC)CCC	CC(=O)N
CCCC[N+](CCCC)(CCCC)CCCC	c1ccc(cc1)C(=O)N
CCCCCC[N+](CCCCC)(CCCCC)CCCCC	C(=O)(C(F)(F)F)N
CCCCCCC[N+](CCCCCC)(CCCCCC)CCCCCCC	c1cc(ccc1C(=O)O)O
CCCCCCCC[N+](CCCCCCC)(CCCCCCC)CCCCCCCC	C(C(=O)O)C(CC(=O)O)C(=O)O)O
CCCCCCCC[N+](CCCCCCCC)(CCCCCCCC)CCCCCCCC	C(CO)N
CCCCCCCC[N+](C)(C)C	C(CO)N(CCO)CCO
CCCCCCCC[N+](C)(C)C	C(CCC(=O)O)CC(=O)O
CCCCCCCC[N+](C)(C)C	C(/C(=C/C(=O)O)/C(=O)O)C(=O)O
CCCCCCCC[N+](C)(C)C	c1ccc(cc1)C(=O)O

c1ccc(cc1)C(c1cccc1)(C(=O)O)O
C(C(=O)O)C(=O)O
C(C(C(=O)O)O)C(=O)O
C(=C\C(=O)O)\C(=O)O
C(=O)(C(=O)O)O
c1ccc(cc1)C(C(=O)O)O
c1ccc(cc1)CC(=O)O
c1ccc(cc1)CCC(=O)O
C(CC(=O)O)C(=O)O
c1ccc(cc1)/C=C/C(=O)O
C=C(CC(=O)O)C(=O)O
C(CCCC(=O)O)CCC(=O)O
C(C(=O)O)O
C(C(C(=O)O)O)C(=O)O)O
CC(=O)CCC(=O)O
c1ccc(c(c1)C(=O)O)O
C([C@H]([C@H]1C(=C(C(=O)O1)O)O)O)O
C(=O)(C(Cl)(Cl)Cl)O
C(C(=O)O)Cl
c1cc(c(cc1/C=C/C(=O)O)O)O
c1c(cc(c(c1O)O)O)C(=O)O
C(CC(=O)O)CC(=O)O
C(=C\C(=O)O)/C(=O)O
CC(C(=O)O)O
CCCCC(=O)O
C=CC(=O)O
CC(=C)C(=O)O
Cc1ccc(cc1)S(=O)(=O)O
c1cc(ccc1/C=C/C(=O)O)O
c1ccc2c(c1)C(=O)C(C2=O)(O)O
c1c(cc(c(c1C(=O)O)C(=O)O)C(=O)O
c1cc(c(c(c1)C(=O)O)C(=O)O)C(=O)O
CCCCC(=O)O
CC(C)(C)C(=O)O
C(F)(F)S(=O)(=O)O
c1coc(c1)CO
C([C@H](C([C@H](CO)O)O)O)O
c1ccc(c(c1)C(=O)O)C(=O)O
CCCC(=O)O
c1cc(cc(c1)C(=O)O)C(=O)O
CC(C)(C)CC(=O)O
C([C@H]([C@H]([C@H]([C@H](CO)O)O)O)O)
O
C([C@H]([C@H](CO)O)O)O
C([C@H]1[C@H]([C@H]([C@H]([C@H](O1)
O[C@H]([C@H]([C@H](CO)O)[C@H]([C@H](CO)O)
O)O)O)O)O
C(CO)O
CCCO
C(CO)CO
CC(CO)O
CCC(CO)O
CC(CCO)O
C(CCO)CO
CCCCCCCCCCCCCCCCCO
CCCCCCCCCCCCCCCCCO
CCCCCCCCCCCCCCCCCO
CC(C)(O)O
C([C@H]([C@H]([C@H]([C@H](CO)O)O)O)O)
O
c1cc(cc(c1)O)O
C(C(CO)O)O
C(COCCOCCO)O
c1ccc(cc1)O
c1ccc2c(c1)cccc2O
c1cc(ccc1O)Cl
Cc1ccc(cc1)O
Cc1cccc1O
COc1cccc1O
Cc1c(c(ccc1)O)C
CC1(CCC(c2c1ccc(c2)O)(C)C)C
C([C@H](C([C@H](CO)O)O)O)O
Cl[C@H]([C@H]([C@H]([C@H](C(O1)(CO)O)O)O)O)
C([C@H]1[C@H]([C@H]([C@H]([C@H](C(O1)O)O)O)
O)O)O)O

```

CCCCCCCCCCCCCCCC(=O)O
CCCCCCCC/C=C\CCCCCCCC(=O)O
C([C@@H]1[C@H]([C@@H]([C@H](C(O1)O)O)O)O)O
C1[C@H]([C@H]([C@@H](C(O1)(CO)O)O)O)O
C([C@@H]1[C@H]([C@@H]([C@H]([C@H](O1)O[C@]1[C@H]([C@@H]([C@H](O1)CO)O)O)CO)O)O)O)O

c1ccc2c(c1)ccc(=O)o2

CC(=O)c1c(cc(cc1)OC)O

CC1CCC(C(C1)O)C(C)C

CC1(C2CCC1(C(=O)C2)C)C

Cc1cc(c(cc1)C(C)C)O

c1ccc(cc1)OC(=O)c1cccc1O

c1ccc2c(c1)cccc2O

c1ccc2c(c1)C(=O)C(C2=O)(O)O
Cc1ccc(cc1)S(=O)(=O)O
C[S+](C)CC(=O)[O-]

C([C@@H]1[C@H]([C@@H]([C@H](C(O1)O)O)O)O)O
C1[C@H]([C@H]([C@@H](C(O1)O)O)O)O
C([C@@H]1[C@H]([C@@H]([C@H](O1)O[C@]1[C@H]([C@@H]([C@H](O1)CO)O)O)CO)O)O)O)O
C([C@@H]1[C@H]([C@@H]([C@H](C(O1)O)O)O)O)O
OC[C@@H]2[C@H]([C@@H]([C@H](C(O2)O)O)O)O)O)O)O
C1[C@H]([C@@H]2[C@H](O1)[C@H](CO2)O)O
C([C@@H]1[C@H]([C@@H]([C@H](C(O1)O)N)O)O)O)O
COc1c(ccc(c1)C=O)O
c1cnc[nH]1
C1CNC(=O)N1
c1c[nH]nc1
C/C=C(C)/C(=O)O
CCC(CO)(CO)CO
CCCCCCCCCCCC(=O)O
CCCCCCCCCCCCCCCC(=O)O
CCCCCCCCCCCCCCCC(=O)O
C[N+](C)(C)CC(=O)[O-]
C1C[C@H](NC1)C(=O)O
C(CC(=O)N)[C@@H](C(=O)O)N
C([C@@H](C(=O)O)N)S
C(C(=O)O)N
C(CC(=O)N[C@@H](CS)C(=O)NCC(=O)O)[C@@H](C(=O)O)N
C(C(F)(F)F)(C(F)(F)F)O
c1cc(c(c(c1)O)O)O
CN(C)C(=O)N
c1ncn[nH]1
C(CCCCCO)CCCCO
C1[C@H]([C@H]([C@@H](C(O1)(CO)O)O)O)O
CCCCCCCCCCCCCCC=CC(C(COC(=O)CCC(=O)O)CCOC)NC(=O)CCCCC)O
C(=O)(C(F)(F)F)N
CC(=O)O
CCCCCCCCCCCC(=O)O
C(CCCO)CCO
C1CCC(=O)NCC1
CNC(=O)NC
C([C@@H](C(=O)O)N)O
C[C@H]([C@@H](C(=O)O)N)O
CSCC[C@@H](C(=O)O)N
CCCCCCCCCCCCCO
C([C@@H](C(=O)O)N)C(=O)O
CCC(=O)O
C(CC(=O)O)[C@@H](C(=O)O)N
C(C[C@@H](C(=O)O)N)CN=C(N)N
C(COCCOCCOCCOCCO)O
C(CN)N
C(COCCOCCO)O
CCCCCCCC(=O)O
C(C(CO)(CO)CO)O
O
C(C(CO)O)O
C(C(=O)O)Cl
CC(CCO)O
CCCCO
C(CO)NCCO

```

```

C(CO)N(CCO)CCO
c1ccccc1
C(CC#N)C#N
C1CCS(=O)(=O)C1
CC(CO)O
C(=O)O
C(CO)N
CCCCCCCCCO
CC(=O)NC
C(=O)N
CCCCC[C@H](C/C=C\CCCCCCCC(=O)O)O
CCCCCO
CCCCCCCCCO
CCCCCCCCCO
C1CCC(CC1)O
CC1CCC(C(C1)O)C(C)C
C(COCCOCCOCCO)O
C=CN
C(COCCO)O
c1cc(cnc1)C(=O)O
CCC(O)O
C([C@@H]1[C@@H]([C@@H]([C@H]([C@@H](
O1)O[C@@H]1[C@H](O[C@H]([C@@H]([C@H]1
O)O)O)CO)O)O)O)O
C[C@@H](C(=O)O)N
C1CC(=O)NC1=O
Cc1cc(c(c1)C(C)C)O
CC(C)Cc1ccc(cc1)C(C)C(=O)O
CC1(C2CCC1(C(=O)C2)CS(=O)(=O)O)C
CC1=CCC(CC1O)C(C)(C)O
CCCCCCCCC(CO)O
CC1(C2CCC1(C(=O)C2)C)C
C=CCCCCCCCC(=O)O
CCC(=O)N
[N+](=O)(O)[O-]
C1C[C@H](N(C1)CC(=O)[C@H]([C@@H]([C@@
H](CO)O)O)O)C(=O)O
CN(CCO)CCO
COc1cccc(c1)O
COc1ccc(cc1)O
C1C[C@@H]2[C@H]3CCCN4[C@H]3[C@@H](C
CC4)CN2C(=O)C1
CC(=O)C(=O)O
c1c([nH]cn1)C[C@@H](C(=O)O)N
CC1(C2CCC1(C(=O)C2)CS(=O)(=O)O)C
CC(C)(CO)C(C(=O)NCCCO)O
C([C@@H]1[C@@H]2[C@@H]([C@H]([C@H](O
1)O[C@@H]1[C@H](O[C@@H]([C@@H]([C@H]
1O)O)O[C@@H]1[C@H](O[C@@H]([C@@H]([C@
H]1O)O)O[C@@H]1[C@H](O[C@@H]([C@@H]
[C@@H]1O)O)O[C@@H]1[C@H](O[C@@H](O2
)[C@@H]([C@H]1O)O)CO)CO)CO)O)O)O
c1cc(cc1)O)N
CC(=CCC/C(=C/C(=O)[O-])/C)C
c1ccc2c(c1)C(=NS2(=O)=O)[O-]
C[N+](C)(C)CCO

```

S3 Fundamental Principles of Generating Substructures

Breaking of Retrosynthetically Interesting Chemical Substructures (**BRICS**) method is applied to generate substructures in this study, which is proposed by Degen et al. in 2008. It was initially designed for the field of medicinal chemistry with the aim of evaluating, generating, and combining potential structural fragments to form novel and effective chemical entities, thereby advancing drug design and retro-synthetic processes. **Figure S1** takes aspirin CC(=O)Oc1ccccc1C(=O)O as an example. The atom numbering in the figure is consistent with the atom order in the SMILES expression, excluding hydrogen atoms. Using the BRICS method, the broken bonds are [(1, 3), (3, 4), (9, 10)], and the obtained substructures are [1: (1, 0, 2), 3: (3), 4: (4, 5, 9, 6, 8, 7), 10: (10, 11, 12)]. Nowadays, the functionality of extracting substructures using the BRICS method has been integrated into the RDKit

library in Python.[4] The resulting substructures will be saved in dictionary format for ease of indexing and masking operations in the SME method.

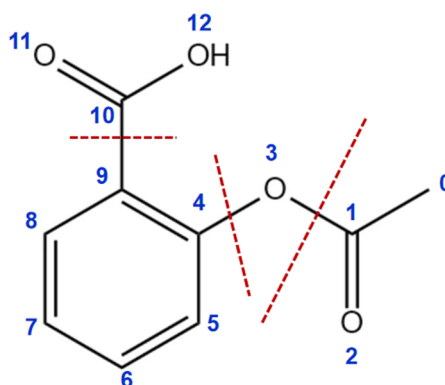


Figure S1. Substructures of aspirin generated via the BRICS method.

S3 Target Molecules: Various Oxidized Thiophenic Sulfur Compounds

Various oxidized thiophenic sulfur compounds serve as the research targets. **Figure S2** illustrates these molecules. The average ND_{sub} value for each molecular category acts as the evaluation index. This index quantifies the affinity of the model for the oxidized thiophenic sulfur system.

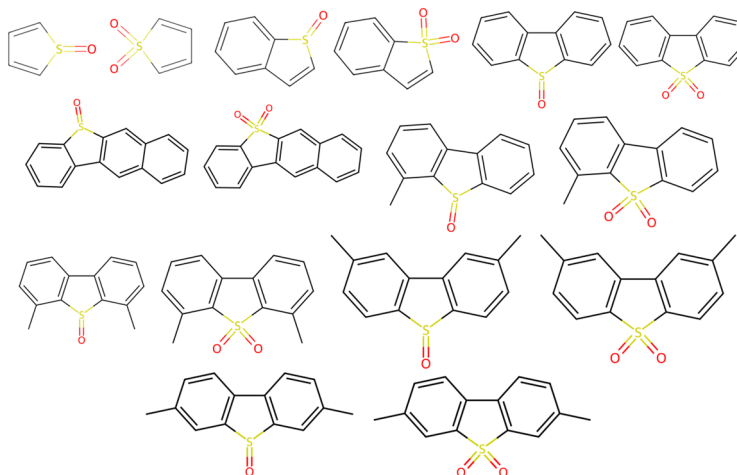


Figure S2. Target various oxidized thiophenic sulfur compounds.

S4 Test RMSE with the Introduction of Different Prior Data

This study investigates how different prior data regulate the predictive performance of the model (**Figure S3**). The experiments involve 30 iterations. The process sequentially incorporates three types of prior information for comparison. The information includes MPI (PDI-1), ESP interval distribution (PDI-2), and molecular surface area (PDI-3). Quantitative analysis assesses the RMSE of the test sets. The results show that molecular surface area acts as the primary factor among the prior variables. It provides the greatest contribution to decreasing the prediction error of the model and enhancing overall performance.

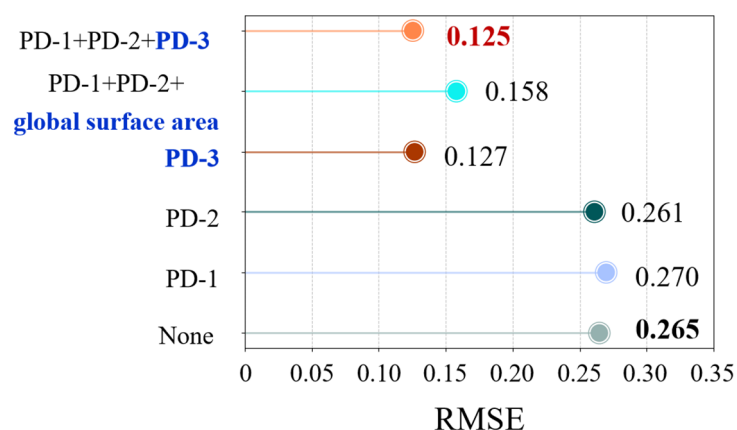


Figure S3. Prediction accuracy of PDI-DMPNN evaluated at 30 epochs varies with the introduction of different prior data types.

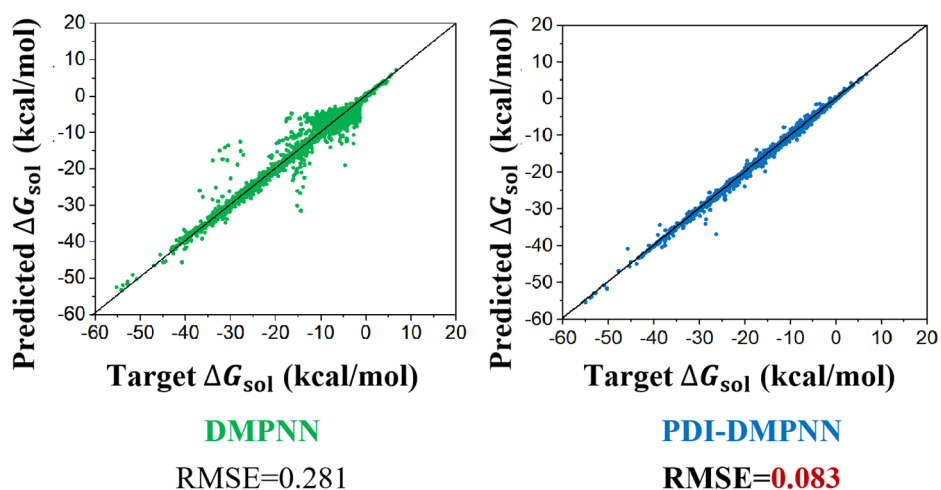


Figure S4. Comparative parity plots of DMPNN and PDI-DMPNN.

S5 Environmental Assessment of Component-2 Candidates

For the environmental assessment, substances with acute toxicity, ecological hazards, or flammability were strictly excluded. Substances with irritancy were conditionally retained.

Table S3. Environmental assessment of Component-2 candidates.

SMILES	Chemical Name	Acute Toxicity	Ecological Hazard	Flammability	Irritancy
<chem>CC(=O)N</chem>	Acetamide	√	√	×	×
<chem>C(=O)(N)N</chem>	Urea	×	×	×	×
<chem>C(CO)N</chem>	Ethanolamine	×	×	×	√
<chem>C(=N)(N)N</chem>	Guanidine	√	√	√	√
<chem>CCN</chem>	Ethylamine	×	×	√	√
<chem>CC(CCO)O</chem>	1,3-Butanediol	×	×	√	×
<chem>CCCCCN</chem>	Butylamine	√	√	√	√
<chem>CC(C(=O)O)O</chem>	L-Lactic Acid	×	×	×	×
<chem>C1CC(N)C(=O)O</chem>	Azetidine-2-carboxylic acid	×	×	×	√
<chem>C(CO)NCCO</chem>	Diethanolamine	√	√	×	√
<chem>C(C(=O)O)C(CC(=O)O)C(=O)O</chem>	Citric Acid	×	×	×	√

S6 HPLC Test Results of the Control Experiments

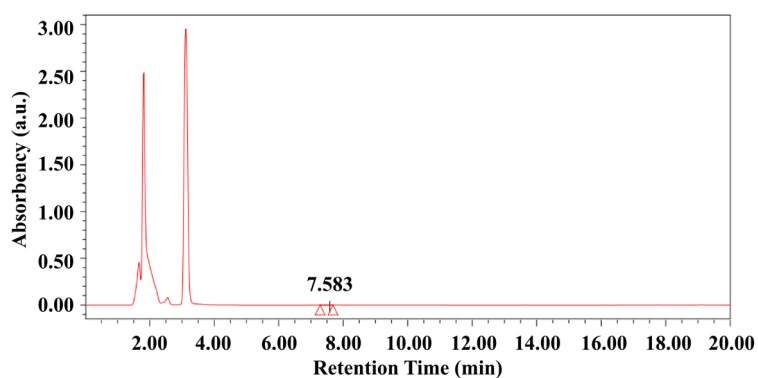


Figure S5. HPLC detected the spectrum of DBT in DES-2 phase after OEDS.

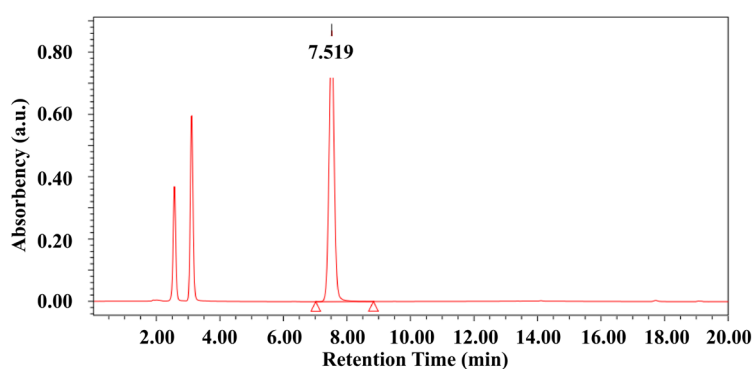


Figure S6. Oil phase HPLC spectrum after OEDS using glycerol.

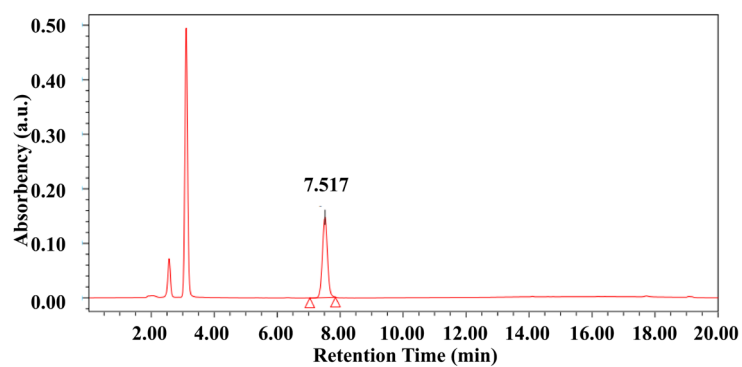


Figure S7. Oil phase HPLC spectrum after OEDS using citric acid.

S7 HPLC Test Results of the Control Experiments

Figures S8 and S9 were the standard spectra of glycerol and citric acid in the National Institute of Standards and Technology database.

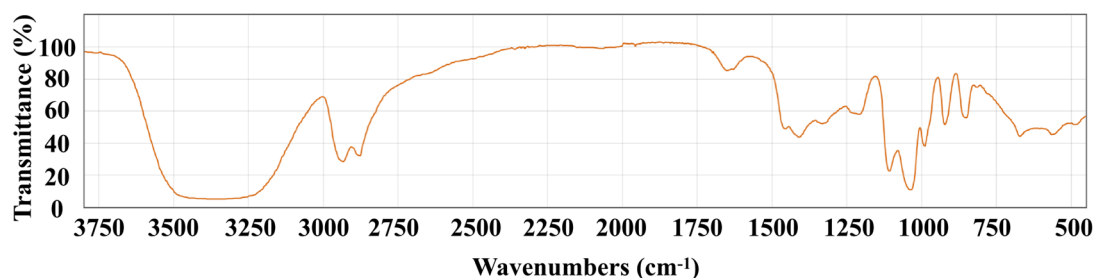


Figure S8. Standard FTIR spectrum of glycerol.

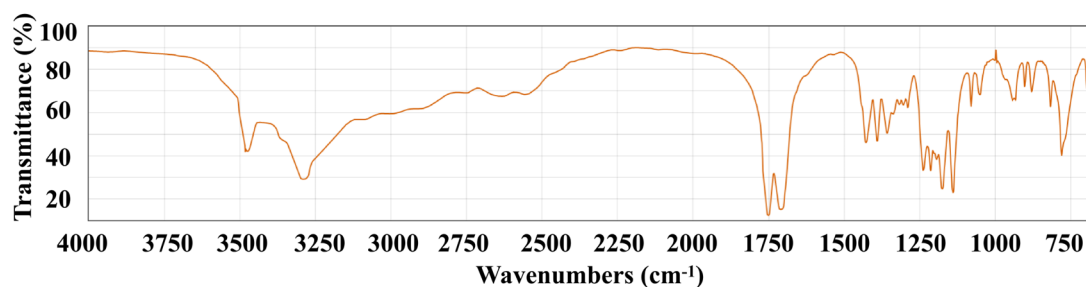


Figure S9. Standard FTIR spectrum of citric acid.

S8 Supplementary characterization of DES-2

Figure S10 gave the variation of the thermal transition temperature (melting point or glass transition temperature) with the molar ratio of glycerol to citric acid. When the molar ratio was 1:1, 2:1, and 3:1, the T_g was -26.8, -41.6, and -53.4 °C, respectively. This indicated that the T_g gradually decreased with the increase of glycerol ratio. When the glycerol content was higher, the molecular mobility of the system was stronger, and it was more difficult to vitrify. T_g needed to be low enough, in order to ensure that DES remained a stable liquid under operating conditions. T_g of the 1:1 composition was higher (-26.8 °C), and its macro eutectic degree was low, and the low temperature fluidity was insufficient. Therefore, 2:1-3:1 was more suitable as the composition range of the DES system.

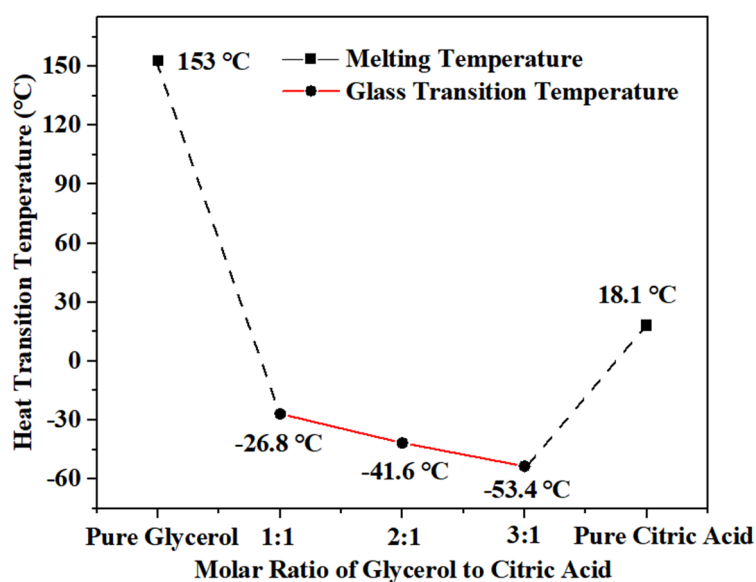


Figure S10. Change of thermal transition temperature with the molar ratio of glycerol to citric acid.

Figure S11 was the TGA curve of DES-2. The total weight loss of the sample was 93.19 %. Among them, the weight loss in the range of 189.5-232.7 °C was 28.16 %, mainly from water removal and partial

decarboxylation of citric acid. The weight loss of 41.40 % in the range of 272.0-318.0 °C was attributed to the volatilization or pyrolysis of low molecular weight components. The weight loss of 29.46 % in the range of 383.9-416.0 °C was related to the pyrolysis and carbonization of the hydrogen bond cross-linked structure with higher thermal stability.

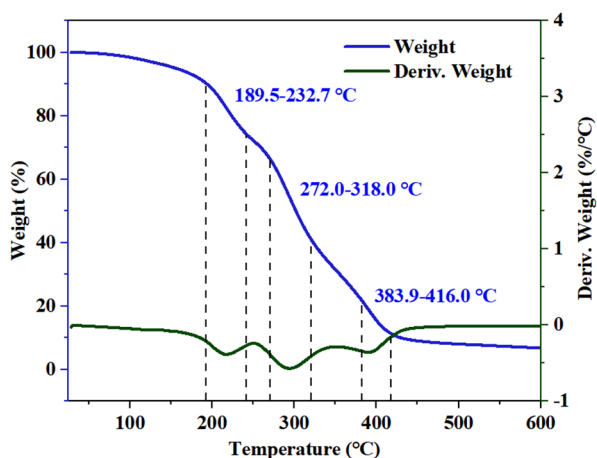


Figure S11. TGA curve of DES-2 (glycerol to citric acid molar ratio of 2:1).

S9 Hydrogen-bonding Energies between DBT and DES-2 Components

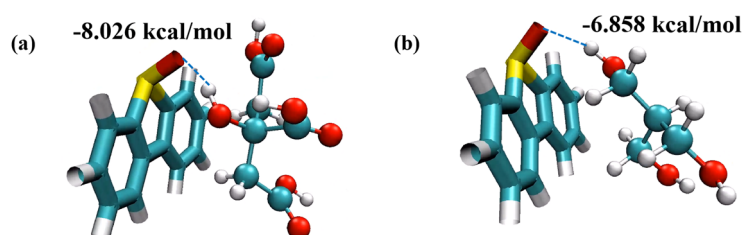


Figure S12. (a) Hydrogen-bonding energy in the extraction of DBTO by citric acid. (b) Hydrogen-bonding energy in the extraction of DBTO by glycerol.