Supporting information for

Data-Driven Insight into the Reductive Stability of Ion–Solvent Complexes in Lithium Battery Electrolytes

Yu-Chen Gao,[†,‡] Nan Yao,[†,‡] Xiang Chen,[†,*] Legeng Yu,[†] Rui Zhang,[₴] Qiang Zhang[†]

[†]Beijing Key Laboratory of Green Chemical Reaction Engineering and Technology, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China

[₴]Beijing Huairou Laboratory, Beijing 101400, China

I. Supporting Text

1. Methods

1.1. Molecular generation algorithm

According to the rules of graph theory, a molecule can be regarded as an undirected graph without self-loops, denoted as $G = <V, E>$. The atoms are considered as a set of vertices $V = \{v_1, v_2, ..., v_n\}$, and the bonds are treated as edges in the edge set $E = \{e_1, e_2, ..., e_m\}$. The atomic information is stored as attributes of the vertex set *V*, denoted as $A = \{a_1, a_2, ..., a_n\}$, and the bond information is represented as attributes of the edge set, denoted as $B = \{\mathrm{b}_1, b_2, ..., b_m\}$. The degree of a vertex refers to the weighted sum of the attributes of edges connected to the vertex. For example, the degree of the vertex V_i is denoted as $d_\mathrm{i} = b_{\mathrm{i},\,1} + b_{\mathrm{i},\,2} + ... + b_{\mathrm{i},\,\mathrm{p}}$, where p is the number of edges connected by V_i. Typically, as the hydrogen atom attributes are not considered in the molecular graph, the degree of carbon atoms is less than (unsaturated) or equal (saturated) to four, and that of oxygen atoms is less than (unsaturated) or equal (saturated) to two.

The implementation of the molecular generation algorithm was based on the Python code with the package Networkx[1] and the chemoinformatics toolkit RDKit.[2] Networkx was used for the creation, storage, update, and isomorphous judgment of the molecular graph. RDKit was used for the conversion from the graph data structure to Simplified Molecular Input Line Entry Specification (SMILES),[3] standardized output, and the screening of molecules. Networkx was utilized to construct initial base graphs, including formaldehyde and dimethyl ether, upon which vertices (atoms) and edges (bonds) are iteratively added in layers. The set of attributes for atoms that can be added included carbon and oxygen, and the set of attributes for edges that can be added included single and double bonds.

The algorithm for molecular graph generation can be seen in Table S1. In the process of generating molecules layer by layer according to the number of vertices (the number of heavy atoms), each round completed the tasks of vertex augmentation, edge augmentation, graph isomorphism judgment, and format conversion. Firstly, the round started with vertex augmentation, traversing each vertex in the graph. If the vertex was a carbon atom and its degree was less than four, a new node was added along with an associated edge. If the vertex was an oxygen atom and its degree was one, a new vertex with a carbon atom attribute was added, accompanied by a new edge with a single bond attribute. Secondly, edge augmentation was performed on the molecule that has been vertex augmented. All pairs of atoms in a molecular graph were iteratively traversed. If both atoms in the pair were unsaturated, a new edge would be added. This round ended when all pairs of atoms had been traversed and were saturated. To ensure the reliability of the molecules generated to some extent during the generation process, measures were taken to avoid forming peroxide bonds and distorted molecules (e.g., molecules with double bonds in ternary rings or bridgehead carbon atoms in small rings). Moreover, a graph isomorphism judgment algorithm was used to eliminate topologically identical molecules, ensuring the uniqueness of the generated molecular graph. The RDKit toolkit was utilized to convert the generated molecular graph into a standard SMILES format, facilitating the storage and reading of molecules. Because lithium metal anode can react with active hydrogen in the solvent, the RDKit toolkit was used to remove molecules with active hydrogen (inclusing alcohols, aldehydes, and acids). Ultimately, 1,399 molecules were generated as the initial database for potential electrolyte solvent molecules.

1.2. Clustering analysis and visualization

The feature representation of the 1399 molecules was conducted using the extended connectivity circular fingerprints (ECFPs)[4] from the DeepChem toolkit,[5] with each molecule represented as a 2048-dimensional vector. ECFPs computed the representation of a molecule by decomposing it into local neighborhoods and hashing these components into a bit vector of the specified size, which can effectively describe the structural characteristics of molecules. The generated molecular fingerprints were used as input for clustering using the t-distributed stochastic neighbor embedding (t-SNE)[6] algorithm in the scikit-learn[7] toolkit. Principal component analysis (PCA) was used to initialize the embedding of the molecules, with the perplexity setting to 8 and the random state setting to 0. The clustering results were colored according to the category of molecules, with linear carbonyl compounds corresponding to yellow, cyclic carbonyl compounds corresponding to red, linear ethers corresponding to green, and cyclic ethers corresponding to blue.

1.3. Density functional theory (DFT) calculation

The DFT calculations were carried out in Gaussian (G16) program[8] with Becke's three-parameter hybrid method using the Lee–Yang–Parr correlation functional (B3LYP)[9] at 6-311+G(d,p) level. The solvation effect was considered with the universal solvation model of SMD.[10] Frequency analysis was conducted to ensure the ground state of molecular structures. Based on the optimized structures, the highest occupied molecular orbital (HOMO), the lowest unoccupied molecular orbital (LUMO) energy level, and corresponding atomic orbital components were analyzed using the Natural Bond Orbitals (NBO) theory.[11]

The formation energy (E_{f}) of a solvent is defined as follows (taking ternary compound $\mathrm{C_xH_yO_z}$ as an example):

$$E_{\mathrm{f}} = \frac{E(\mathrm{C_xH_yO_z}) - xE(\mathrm{C}) - yE(\mathrm{H}) - zE(\mathrm{O})}{x + y + z} \#(2)$$

where $E(\mathrm{C_xH_yO_z})$ represents the total energy of the ternary compound. $E(\mathrm{C})$, $E(\mathrm{H})$, and $E(\mathrm{O})$ represents the total energy difference between a CH_4 molecule and two H_2 molecules, half of the total energy of an H_2 molecule, and half of the total energy of an O_2 molecule, respectively.

The binding energy (E_{b}) between a Li^+ and a solvent is defined as follows:

$$E_{\mathrm{b}} = E_{\mathrm{cluster}} - (E_{\mathrm{Li^+}} + E_{\mathrm{solvent}}) \#(3)$$

where E_{cluster}, $E_{\mathrm{Li^+}}$, and E_{solvent} represent the total energy of the Li^+–solvent cluster, Li^+, and solvent, respectively.

1.4. Machine learning methods

Ten machine learning algorithms were utilized to predict the LUMO energy levels of 1399 coordinated molecules. All machine learning algorithm implementations come from the scikit-learn[7] package, including linear regression (LR), decision tree regression (DT), support vector machines regression (SVM), k-nearest neighbor regression (KNN), random forest regression (RF), adaptive boosting regression

(AdaBoost), gradient boosting regression (GB), bagging regression (Bagging), extremely randomized trees regression (Extra Trees), and multi-layer perceptron regression (MLP). The configurations of these models are shown in the Table S3. Four-fold cross-validation (CV) is used for model selection (Table S4, Table S6, and Table S7). The hold-out method is used to evaluate the optimal model.

Shapley additive explanation (SHAP) was combined with the trained model to improve the interpretability and observe the influence of each molecular feature on the LUMO energy level of coordinated solvents. The analysis was based on the SHAP package.

II. Supporting Figures

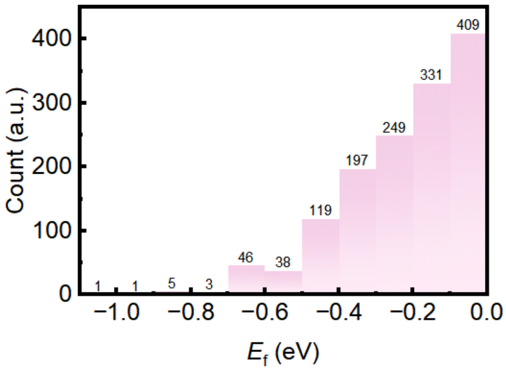

Figure S1. Histogram of formation energy (E_f) of the ion–solvent complexes.

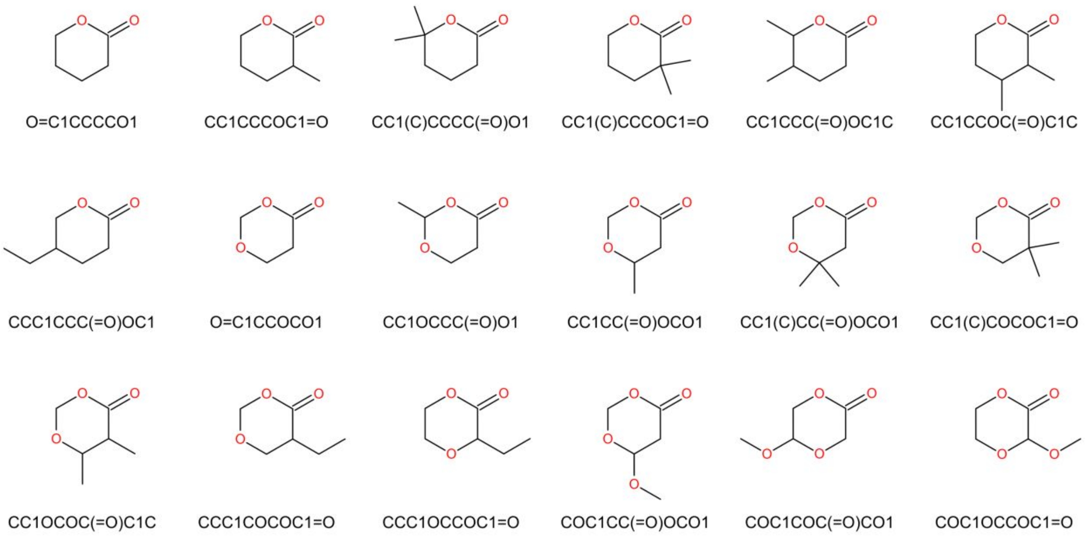

Figure S2. The structural formulas and corresponding SMILES representations of the molecules, in which the LUMO energy levels increase after coordinating with a Li-ion.

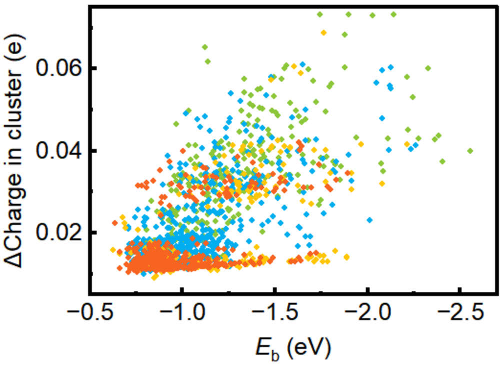

Figure S3. The correlation between the charge transfer during the formation of the ion–solvent complexes and binding energy (E_b). Each point represents a molecule. The linear carbonyl compounds, cyclic carbonyl compounds, linear ethers, and cyclic ethers are marked with yellow, red, green, and blue, respectively.

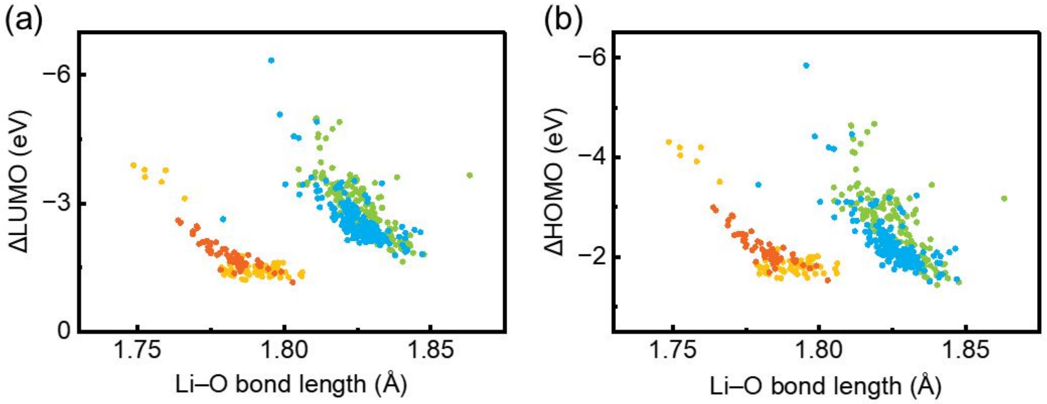

Figure S4. Effects of Li–O bond length on the LUMO and HOMO energy level change for molecules containing only one oxygen atom. The correlation between the (a) LUMO and (b) HOMO energy level change and Li–O bond length. Each point represents a molecule. The linear carbonyl compounds, cyclic carbonyl compounds, linear ethers, and cyclic ethers are marked with yellow, red, green, and blue, respectively.

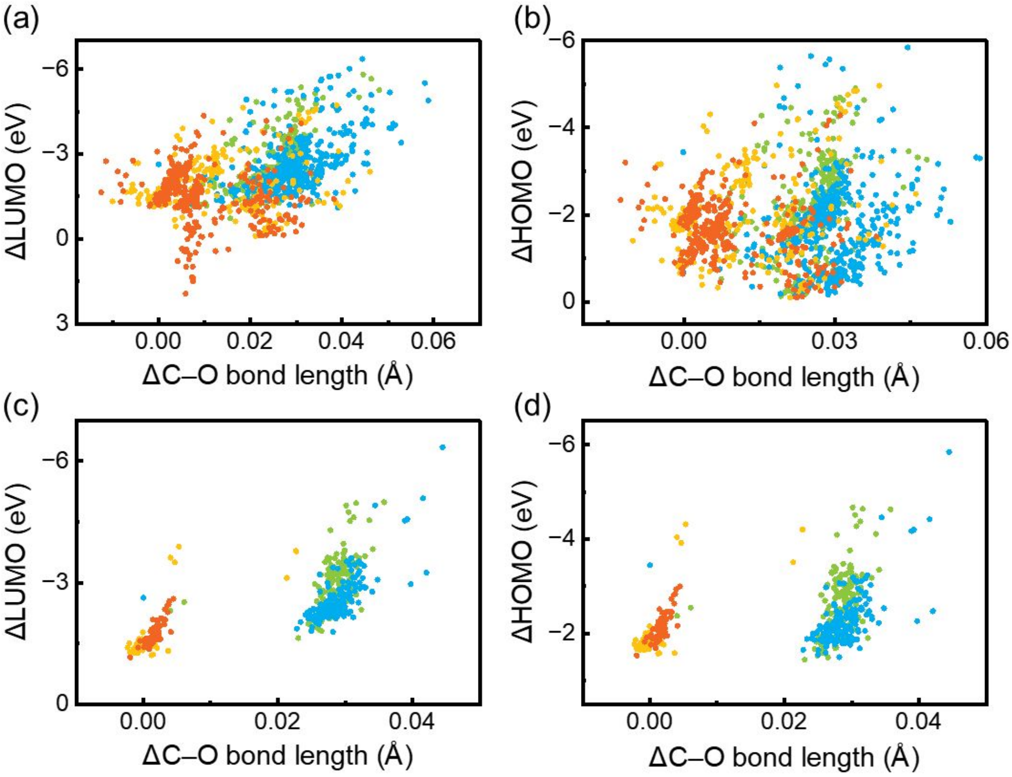

Figure S5. Effects of C–O bond length on the LUMO and HOMO energy level change. The correlation between the (a) LUMO and (b) HOMO energy level change and C–O bond length. The correlation between the (c) LUMO and (d) HOMO energy level change and C–O bond length, for molecules containing only one oxygen atom. Each point represents a molecule. The linear carbonyl compounds, cyclic carbonyl compounds, linear ethers, and cyclic ethers are marked with yellow, red, green, and blue, respectively.

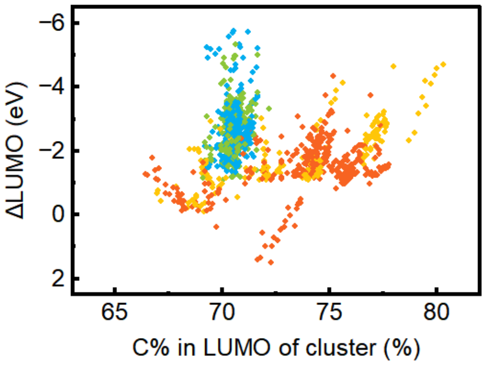

Figure S6. The correlation between the LUMO energy level change and the ratio of carbon 2p orbitals in the LUMOs. Each point represents a molecule. The linear carbonyl compounds, cyclic carbonyl compounds, linear ethers, and cyclic ethers are marked with yellow, red, green, and blue, respectively.

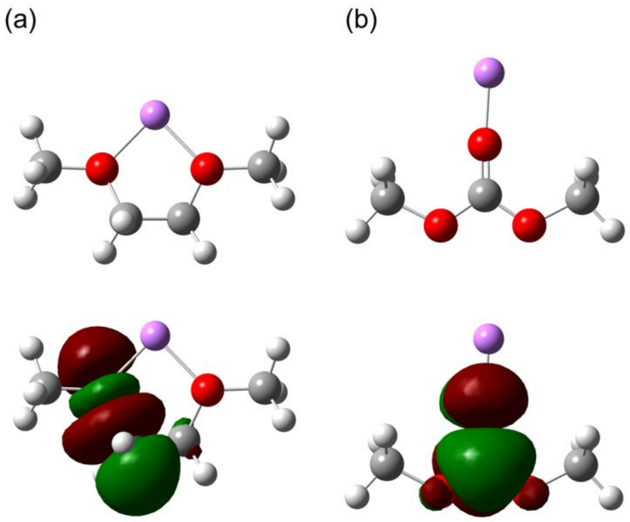

Figure S7. The optimized geometric structures and visual LUMOs of ion–solvent complexes. (a) Li^+–DME. (b) Li^+–DMC. The hydrogen, carbon, oxygen, and lithium atoms are marked with white, gray, red, and purple, respectively. The red and green regions of LUMOs represent the positive and negative parts of orbitals, respectively. Isovalue: 0.02.

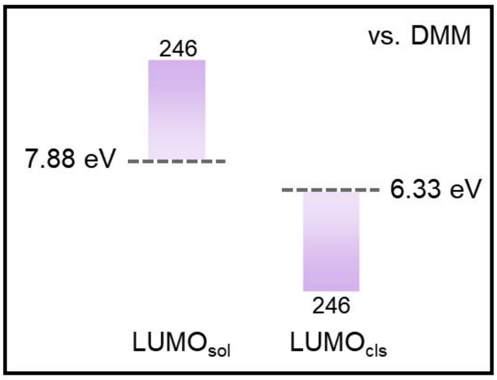

Figure S8. The LUMO energy level changes after the formation of ion–solvent complexes structures based on dimethoxymethane (DMM). There are 246 molecules whose LUMO energy levels are higher than DMM in pure solvent, but they are lower than DMM after the forming of ion–solvent complexes.

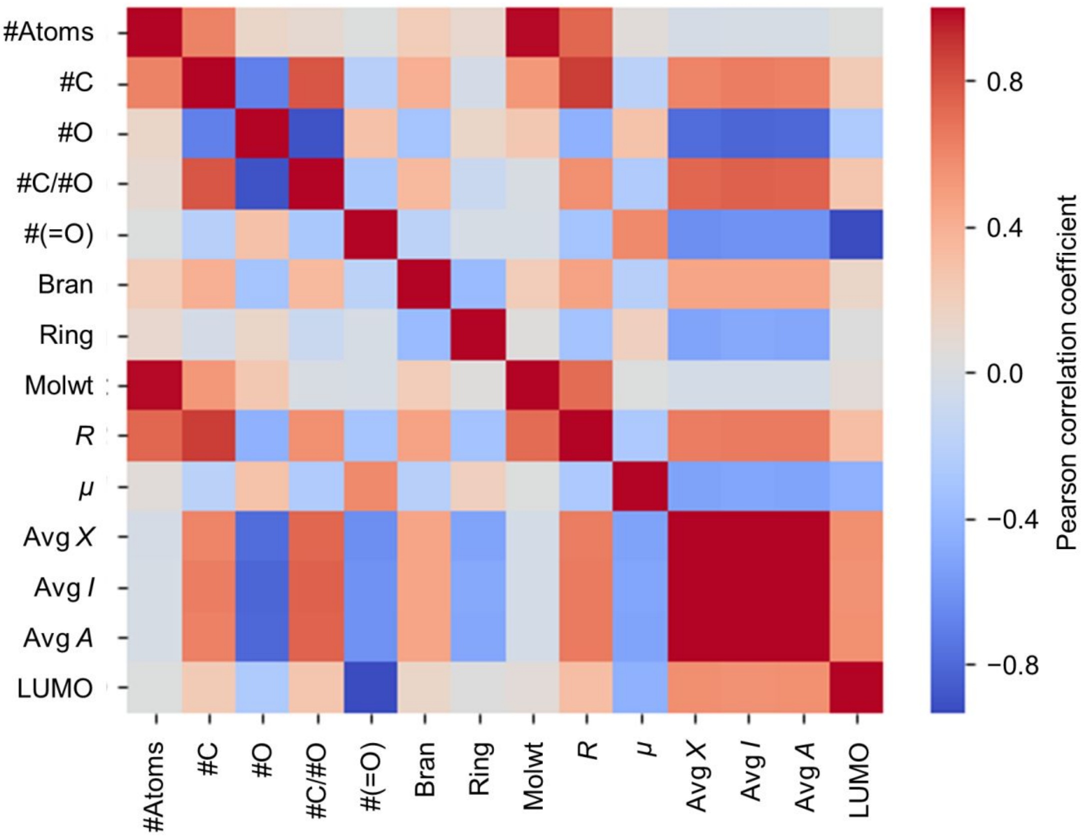

Figure S9. Pearson correlation analysis heat map of molecular features. Colorbar: Pearson correlation coefficient. Red indicates a positive correlation, while blue indicates a negative correlation. The meanings of features are shown in Table S2.

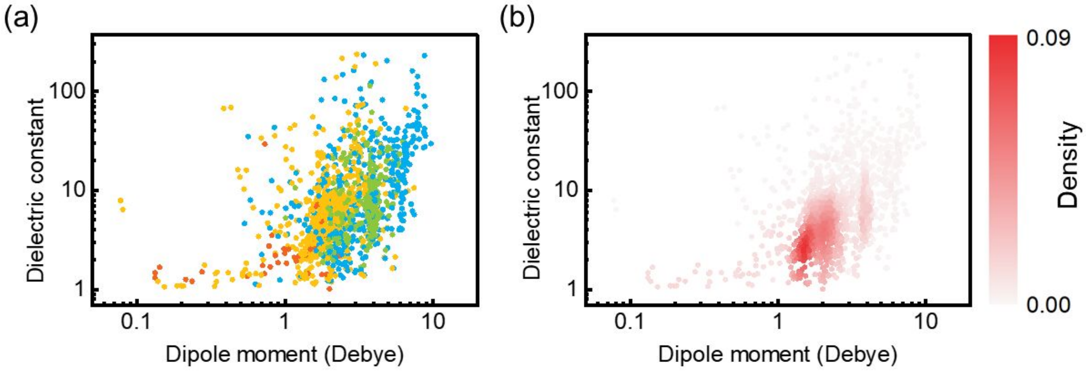

Figure S10. The correlation between dipole moment and dielectric constant of solvents. (a) Each point represents a molecule. The linear carbonyl compounds, cyclic carbonyl compounds, linear ethers, and cyclic ethers are marked with yellow, red, green, and blue, respectively. (b) Each point is colored by point density. Colorbar: The density of points. Darker colors represent more points.

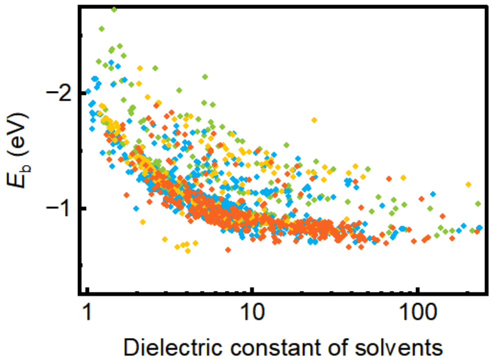

Figure S11. The correlation between the dielectric constant of solvents and binding energy. Each point represents a molecule. The linear carbonyl compounds, cyclic carbonyl compounds, linear ethers, and cyclic ethers are marked with yellow, red, green, and blue, respectively.

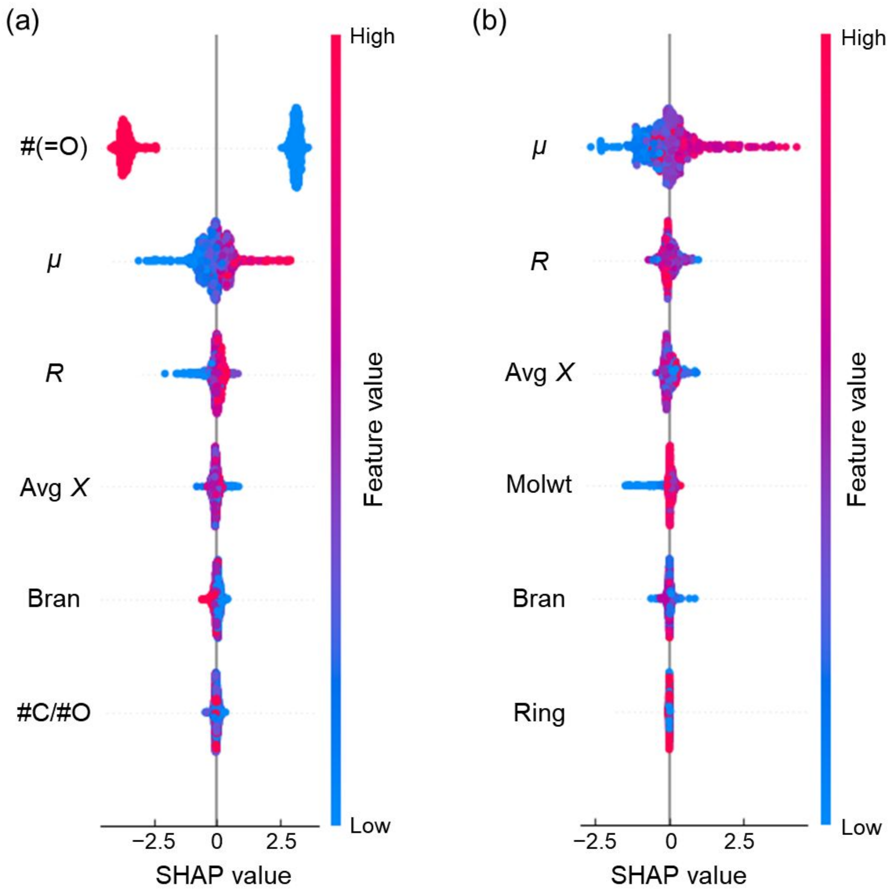

Figure S12. Interpretable machine learning for predicting LUMO energy levels of ion–solvent complexes. (a) Shapley feature ranking for all 1399 molecules. (b) Shapley feature ranking for all carbonyl molecules. SHAP value means the contribution of the sample point to the model performance.

Ⅲ. Supporting Tables

Table S1. Molecular generation algorithm.

Molecular generation algorithm

```
Input: $N$, $|Base.\text{vertices}| = n$ , Empty list $L_1$, $L_2$
Output: $G$
if $N = n$
    initial $Base$
else
    $L_1 \leftarrow$ Read file($N - 1$); $L_1 = [g_1, g_2, \ldots, g_i]$
    while $|L| > 0$ do
        while $|g_i.\text{vertices}| > N$ do
            $G_i \leftarrow$ Add vertices($g_i$)
            $L_2 \leftarrow$ Isomorphism($G_i$)
        end while
    end while
Read file($L_2$)
$G \leftarrow$ GraphToSmiles($L_2$)
```

Table S2. Alternative molecular descriptors.

Descriptor	Symbol
The number of atoms	#Atoms
The number of carbon atoms	#C
The number of oxygen atoms	#O
The ratio of carbon atoms to oxygen atoms	#C/#O
The number of carbonyl oxygens	#(=O)
The number of branches	Bran
The number of rings	Ring
Molecular weight	Molwt
Molecular radius	*R*
Molecular dipole moment	μ
Average electronegativity	Avg *X*
Average ionization energy	Avg *I*
Average electron affinity	Avg *A*

Table S3. Ten machine learning algorithms and configurations.

Algorithm	Function	Configuration
1 LR	LinearRegression	Fit intercept: True Positive: False
2 DT	DecisionTreeRegressor	Criterion: Squared error Max depth: None Min samples leaf: 1
3 SVM	SVR	Kernel: Rbf Degree: 3
4 KNN	KNeighborsRegressor	N neighbors: 5 Weights: Uniform
5 RF	RandomForestRegressor	N estimators: 100 Criterion: Squared error Max depth: None Min samples leaf: 1
6 AdaBoost	AdaBoostRegressor	N estimators: 50 Learning rate: 1.0 Loss: linear
7 GB	GradientBoostingRegressor	N estimators: 100 Learning rate: 0.1 Loss: Squared error
8 Bagging	BaggingRegressor	N estimators: 100 Max features: 1.0
9 Extra Trees	ExtraTreeRegressor	Criterion: Squared error Max depth: None Min samples leaf: 1
10 MLP	MLPRegressor	Hidden layer sizes: (100,) Activation: Relu Solver: Adam Alpha: 0.0001 Learning rate: Constant Learning rate init: 0.001

Table S4. Four-fold cross-validation (CV) among 1399 molecules for model selection. The table presents the values of CV using mean absolute error (MAE) as the metric and their average value.

Algorithm	CV-MAE1 (eV)	CV-MAE2 (eV)	CV-MAE3 (eV)	CV-MAE4 (eV)	Avg-MAE (eV)
1 LR	0.732	0.565	0.730	1.166	0.798
2 DT	0.813	0.728	0.980	1.151	0.918
3 SVM	1.082	0.628	1.249	1.046	1.001
4 KNN	0.738	0.622	0.860	1.020	0.810
5 RF	0.639	0.563	0.769	0.931	0.726
6 AdaBoost	0.669	0.608	0.945	1.020	0.811
7 GB	0.624	0.548	0.769	0.962	0.726
8 Bagging	0.641	0.596	0.778	0.970	0.746
9 Extra Trees	0.731	0.755	0.956	1.236	0.919
10 MLP	0.677	0.567	0.940	1.002	0.796

Table S5. Feature importance ranking among 1399 molecules.

Importance	Tree-based	Permutation	SHAP
1	#(=O)	#(=O)	#(=O)
2	μ	μ	μ
3	R	R	R
4	Avg X	Avg X	Avg X
5	#C/#O	#C/#O	Bran
6	Bran	Bran	#C/#O

Table S6. Four-fold cross-validation (CV) among 771 ether molecules for model selection. The table presents the values of CV using mean absolute error (MAE) as the metric and their average value.

Algorithm	CV-MAE1 (eV)	CV-MAE2 (eV)	CV-MAE3 (eV)	CV-MAE4 (eV)	Avg-MAE (eV)
1 LR	0.620	0.586	0.553	0.613	0.593
2 DT	0.748	0.718	0.697	0.794	0.739
3 SVM	0.611	0.547	0.541	0.582	0.570
4 KNN	0.628	0.582	0.539	0.588	0.584
5 RF	0.589	0.561	0.545	0.584	0.570
6 AdaBoost	0.630	0.630	0.547	0.639	0.612
7 GB	0.612	0.543	0.541	0.583	0.570
8 Bagging	0.578	0.571	0.581	0.610	0.585
9 Extra Trees	0.761	0.732	0.722	0.783	0.749
10 MLP	0.799	0.666	0.669	0.708	0.711

Table S7. Four-fold cross-validation (CV) among 628 carbonyl compound molecules for model selection. The table presents the values of CV using mean absolute error (MAE) as the metric and their average value.

Algorithm	CV-MAE1 (eV)	CV-MAE2 (eV)	CV-MAE3 (eV)	CV-MAE4 (eV)	Avg-MAE (eV)
1 LR	0.762	0.883	1.210	1.029	0.971
2 DT	0.720	0.929	1.216	0.989	0.964
3 SVM	0.639	0.744	1.028	0.840	0.812
4 KNN	0.639	0.817	1.087	1.179	0.931
5 RF	0.537	0.728	0.987	0.908	0.790
6 AdaBoost	0.807	0.962	1.131	1.130	1.007
7 GB	0.588	0.712	1.056	0.962	0.829
8 Bagging	0.556	0.758	1.017	0.906	0.809
9 Extra Trees	0.757	0.900	1.280	1.050	0.997
10 MLP	0.615	0.838	1.080	1.033	0.891

Table S8. Feature importance ranking among 771 ether molecules.

Importance	Tree-based	Permutation	SHAP
1	μ	μ	μ
2	R	R	R
3	Molwt	Bran	Bran
4	Avg X	Molwt	Molwt
5	Bran	Avg X	Avg X
6	Ring	Ring	Ring

Table S9. Feature importance ranking among 628 carbonyl compound molecules.

Importance	Tree-based	Permutation	SHAP
1	μ	μ	μ
2	R	R	R
3	Molwt	Avg X	Avg X
4	Avg X	Molwt	Molwt
5	Bran	Bran	Bran
6	Ring	Ring	Ring

Reference

(1) Hagberg, A. A.; Schult, D. A.; Swart, P. Exploring Network Structure, Dynamics, and Function Using Networkx. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Pasadena, CA USA, 2008; pp 11–15.

(2) Rdkit: Open-Source Cheminformatics. https://www.rdkit.org/.

(3) Weininger, D. Smiles, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comp. Sci.* **1988**, *28*, 31–36.

(4) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(5) Ramsundar, B. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019.

(6) Laurens, V. D. M.; Hinton, G. Visualizing Data Using T-Sne. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

(7) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(8) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe,

D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16 Rev. C.01. Wallingford, CT, 2016.

(9) Becke, A. D. Density - Functional Thermochemistry. Iii. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(10) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(11) Reed, A. E.; Curtiss, L. A.; Weinhold, F. Intermolecular Interactions from a Natural Bond Orbital, Donor-Acceptor Viewpoint. *Chem. Rev.* **1988**, *88*, 899–926.