Supporting information for "BonDNet: a graph neural network for the prediction of bond dissociation energies for charged molecules"

Mingjian Wen, Samuel M. Blau, Evan Walter Clark Spotte-Smith, Shyam Dwaraknath, and Kristin A. Persson*

E-mail: kapersson@lbl.gov

Datasets

A summary of the PubChem, ZINC, and BDNCM bond dissociation energy (BDE) datasets is presented in Table S1. Different energies are adopted in the three datasets: the PubChem BDE dataset uses the gas-phase enthalpy at 298 K and 1 atm, the ZINC BDE dataset uses the electronic structure energy, and the BDNCM BDE dataset uses the Gibbs free energy at 298.15 K. We choose to use the Gibbs free energy for the BDNCM dataset because both enthalpic and entropic contributions are critical to accurately capturing reaction thermodynamics of charged and moderately sized molecules in solvent at ambient temperature and pressure. The PubChem and ZINC BDE datasets are discussed in detail in Refs. 1 and 2, respectively, and we refer the readers to them for more information. Here, we discuss how the new BDNCM BDE dataset is constructed.

The BDNCM dataset was constructed with an initial target application of predicting the early formation of the solid electrolyte interphase (SEI) formation in lithium-ion battery systems. We first collected a set of principal molecules that are used in or thought to be formed in lithium-ion battery electrolytes as reported in the literature. These include cyclic carbonates (e.g. ethylene carbonate), linear carbonates (e.g. ethyl methyl carbonate), their lithium coordination complexes and fluorinated derivatives, lithiated dicarbonates (e.g. lithium ethylene dicarbonate), water, and more. We then fragmented these molecules by systematically breaking all of their bonds to generate a set of fragments, and this process was iteratively applied to the fragments, fragments of fragments, etc. in order to generate all possible unique fragments. Each principle molecule and each unique fragment, defined by the molecular connectivity, was then assigned charges -1, 0, and +1 to generate a preliminary set. We optimized the geometry of each molecule in the preliminary set using density functional theory; subsequently, we performed a vibrational frequency analysis to calculate the Gibbs free energy. After filtering the optimized structures again for uniqueness, defined now by the molecular connectivity and the charge, we obtained 8518 unique molecules made up of C, H, O, F, and Li, and these molecules yield 64312 homolytic and heterolytic bond breaking reactions. The calculations of the Gibbs free energy were performed using Q-Chem 5.2.2[3] at the ωB97X-V[4] level theory with the def2-TZVPPD basis set.[5] The implicit SMD model[6] was chosen to approximate the thermodynamic properties of the solvent.

Table S1: Summary of the BDNCM, PubChem, and ZINC BDE datasets

	BDNCM	PubChem	ZINC
Energy type	Gibbs free energy	enthalpy	electronic structure energy
Chemical species	C, H, O, F, Li	C, H, O, N	C, H, O, N, S
# unique molecules	8518	249374	4343
# bond-breaking reactions	64312	290644	16626
Molecular charge	-1, 0, 1	0	0
Bond dissociation type	hteterolytic, homolytic	homolytic	homolytic
Breaking ring bond	yes	no	no

Input features

The input atom, bond, and global features for the BonDNet model are presented in Tables S2 and S3. The features are chosen based on the characteristics of the datasets. The BDNCM

BDE dataset contains charged molecules coordinated with metal ions, so “charge” is chosen as a global feature and “coordinate” is chosen as a bond feature. This complex dataset also contains organic and inorganic species, closed-shell and radical molecules, making it difficult to determine the “valence”, “aromatic”, “hybridization”, “acceptor”, and “donor” information of atoms and the “conjugated” and “bond type” information of bonds. Therefore, these are not included as input features. The PubChem and ZINC BDE datasets only have neutral organic molecules, so neither the “charge” global feature nor the “coordiante” bond feature is necessary. However, the features ignored by the BDNCM BDE dataset are well suited for them and thus included.

Table S2: Input atom, bond, and global features for the BDNCM dataset

Feature type	Feature name	Description
Atom	atom type	chemical specie of an atom (one-hot)
	degree	number of bonds an atom forms (integer)
	# hydrogens	number of hydrogens connected to an atom (integer)
	ring status	whether an atom is in a ring (binary)
	ring size	number of atoms in the ring (3–7), “null” if the atom is not in a ring (one-hot or null)
Bond	ring status	whether a bond is in a ring (binary)
	ring size	number of atoms in the ring (3–7), “null” if the bond is not in a ring (one-hot or null)
	coordinate	whether it is a coordinate bond (binary)
Global	# atoms	number of atoms in a molecule (integer)
	# bonds	number of bonds in a molecule (integer)
	weight	weight of a molecule (integer)
	charge	total charge $(-1, 0, 1)$ of a molecule (one-hot)

Table S3: Input atom, bond, and global features for the PubChem and ZINC BDE datasets

Feature type	Feature name	Description
Atom	atom type	chemical specie of an atom (one-hot)
	degree	number of bonds an atom forms (integer)
	# hydrogens	number of hydrogens connected to an atom (integer)
	ring status	whether an atom is in a ring (binary)
	ring size	number of atoms in the ring (3–7), “null” if the atom is not in a ring (one-hot or null)
	valence	valence of an atom (integer)
	aromatic	whether an atom forms aromatic bond (binary)
	hybridization	s, sp^1, sp^2, or sp^3 (one-hot or null)
	acceptor	whether an atom accepts electrons (binary)
	donor	whether an atom donates electrons (binary)
Bond	ring status	whether a bond is in a ring (binary)
	ring size	number of atoms in the ring (3–7), “null” if the bond is not in a ring (one-hot or null)
	conjugated	whether it is a conjugated bond (binary)
	bond type	single, double, triple, or aromatic (one-hot or null)
Global	# atoms	number of atoms in a molecule (integer)
	# bonds	number of bonds in a molecule (integer)
	weight	weight of a molecule (integer)

BonDNet hyperparameters

As discussed in the main text, the hyperparameters determining the structure of BonDNet are selected based on the model performance on the validation set. We conducted a grid search over six hyperparameters and their optimal values that result in the the smallest mean absolute error (MAE) on the validation set are presented in Table S4. Some hyperparameters need more explanation. (1) “graph-to-graph module hidden layer size” denotes the size of the weights and biases in the two-layer fully connected neural networks (FCNNs). More specifically, it is the number of columns of the weight matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ and the length of the bias vectors $\mathbf{b}_1$ and $\mathbf{b}_2$ in the feature update functions $\phi_1, \phi_2, \ldots, \phi_9$ in Eqs. (4) (5) and (7) in the main text. (2) “# hidden layer in graph-to-property module” denotes the number of hidden layers in the FCNN used to map the concatenated reaction feature (Eq. (9) in the main text) to the output BDE. (3) “graph-to-property module hidden layer sizes” denotes the corresponding sizes of the hidden layers in the FCNN.

Table S4: Hyperparameters of BonDNet for the three datasets

	BDNCM	PubChem	ZINC
input feature embedding size	24	24	24
# graph-to-graph modules	3	3	3
graph-to-graph module hidden layer size	192	256	64
graph-to-graph module dropout ratio	0	0.1	0.1
# hidden layers in graph-to-property module	2	3	2
graph-to-property module hidden layer sizes	384, 192	512, 256, 128	128, 64

Reactant-only model hyperparameters

The reactant-only model applies multiple graph-to-graph modules and then map the features of the breaking bond in the last graph-to-graph module to the BDE using an FCNN. The hyperamarameters are determined in the same way as BonDNet, and the optimal values are presented in Table S5. Note that the listed hyperparameters are the optimal values when the whole PubChem BDE dataset is used. The optimal values vary when training on a subset of the PubChem BDE dataset. This is also the case for BonDNet when training on a subset.

Table S5: Hyperparameters of the reactant-only model for the PubChem BDE dataset

input feature embedding size	24
# graph-to-graph modules	4
graph-to-graph module hidden layer size	256
graph-to-graph module dropout ratio	0.1
# hidden layer in FCNN	3
FCNN hidden layer sizes	256, 128, 64

Reactions with large test error

The reactions with the 10 largest prediction errors for the BDNCM test set are given in Fig. S1. They can be broadly categorized into two groups: (1) reactions that are underrepresented in the dataset, including d), f), and j); and (2) reactions that are more complex than one-bond dissociation, including a), b), c), e), g), h), and i). Below, we provide a more detailed analysis as to why BonDNet yields large errors for them.

The dataset has 9 molecules with CCCO double rings as in the reactant of d) and 151 molecules with a CCCO single ring as in the product. However, there is only one reaction that breaks a bond in CCCO double rings and form a CCCO single ring (that is the reaction shown in d)). It is expected that a machine learning model such as BonDNet cannot give correct predictions for such underrepresented data. This is also the case for reaction f). For j), although the dataset has BDEs for more than 500 $F \rightarrow Li^+$ bonds, most of them are Li^+ coordinated to large molecules containing F, very different from the reaction in j). Actually, this reaction is very unlikely to exist since it is extremely difficult to obtain a +1 charged F_2 due to the large electronegativiy of F.

Reaction a) looks like a one-bond-breaking reaction, but a closer examination reveals that the reactant and the product have very different electronic structures: the radical is on an oxygen atom in the reactant, whereas it is on a carbon atom in the product. Thus, the reference BDE calculated from DFT contains a substantial contribution from the rearrangement of the electron density. This cannot be explained by one-bond dissociation that BonDNet is designed to model. The other reactions in group (2) are all complex reactions. We expect the reactant of b) to be a planar molecule, but it is actually not (probably due to insufficient geometry optimization), and thus the reference BDE contains some conformational energy. For c) and g), breaking a bond leads to the change of a neighboring single bond to a double bond. For e), both the two electrons forming the $O-O$ bond in the reactant move to the $C=O$ bond in the product. For h), the $O \rightarrow Li^+$ coordinate bond in the reactant results in strain in the neighboring bonds in the ring, which are released upon bond dissociation. Therefore, the reference BDE contains strain energy. Finally, for i), breaking the bond results in the change of the $O-C=C$ conjugated system in the reactant to the $O=C-C$ conjugated system in the product. The current form of BonDNet is designed to predict BDEs based on graph representation of molecules, and it make predictions only

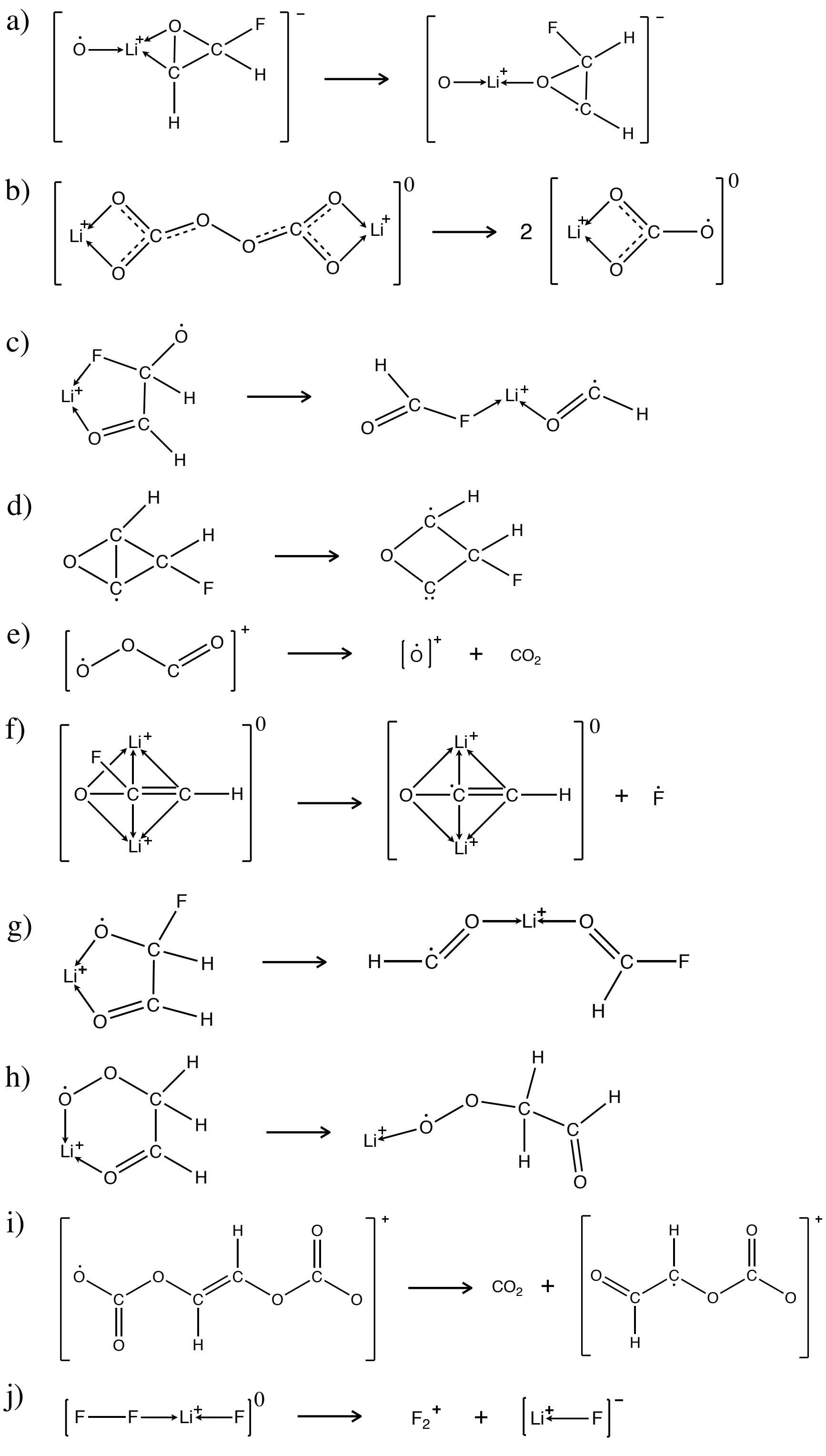

Figure S1: Reactions with large prediction error for the BDNCM test set.

using molecular connectivity information and simple atom, bond, and global features. The complex reactions discussed above are beyond BonDNet's scope, and future extension can be made to deal with these complex reactions.

Prediction for large drug-like molecules

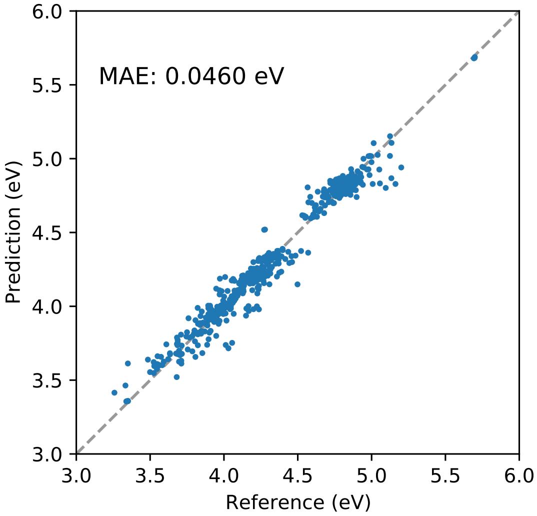

Figure S2: BDEs predicted by BonDNet versus DFT reference values for a set of drug-like molecules much larger than the molecules in the training set. The predictions are made using the BonDNet model trained on the PubChem BDE Dataset, and drug-like molecules and the DFT reference energies are from Ref. 1. BonDNet achieves a mean absolute error (MAE) of 0.0460 eV for the drug-like molecules, and ALFABET achieves an MAE of 0.0494 eV.[1]

Prediction error distribution

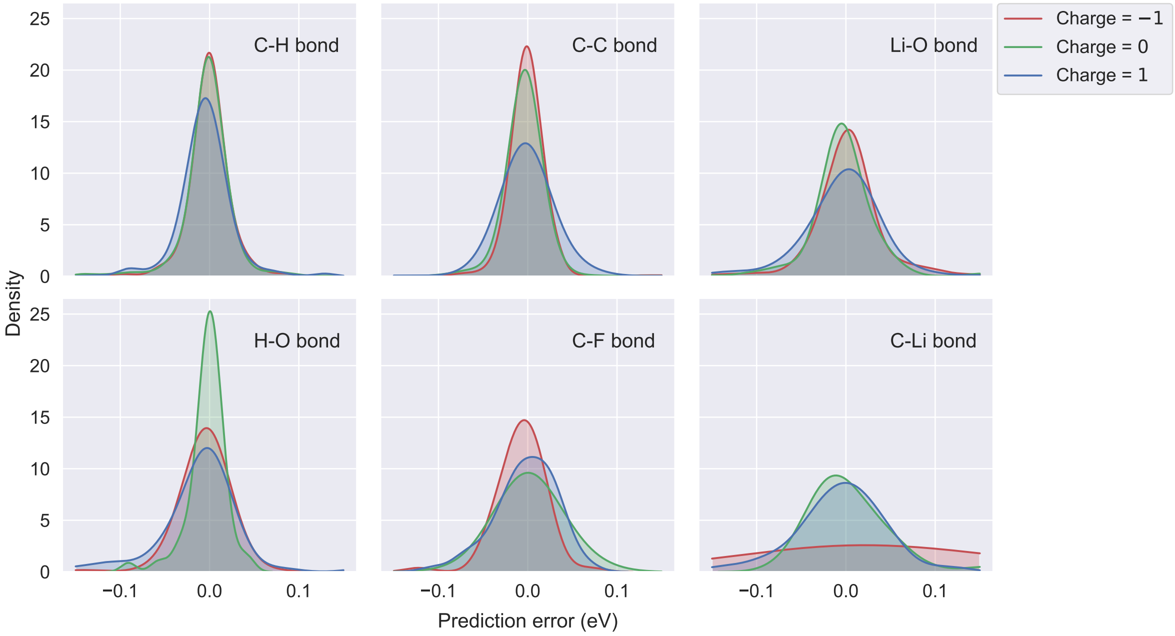

Figure S3: Distribution of BonDNet prediction error by reactant charge for the BDNCM dataset.

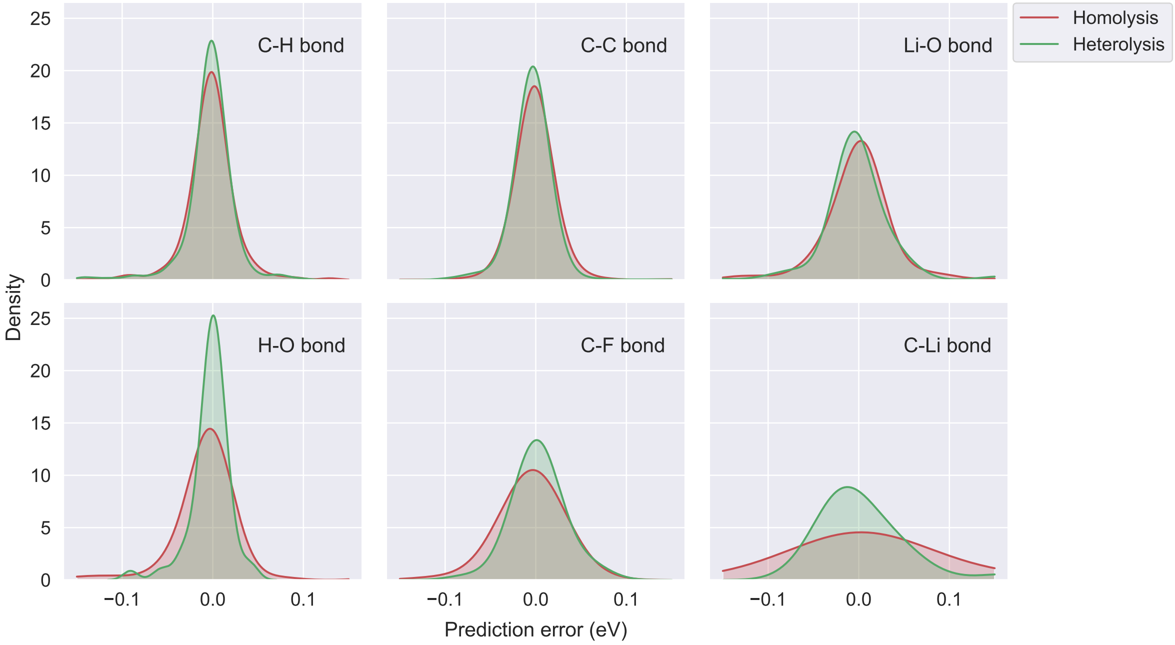

Figure S4: Distribution of BonDNet prediction error by bond dissociation type for the BDNCM dataset.

References

(1) John, P. C. S.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S. Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nature Communications* **2020**, *11*, 2328.

(2) Qu, X.; Latino, D. A.; Aires-de Sousa, J. A big data approach to the ultra-fast prediction of DFT-calculated bond energies. *Journal of cheminformatics* **2013**, *5*, 34.

(3) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X., et al. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Molecular Physics* **2015**, *113*, 184–215.

(4) Mardirossian, N.; Head-Gordon, M. ωB97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Physical Chemistry Chemical Physics* **2014**, *16*, 9904–9924.

(5) Hellweg, A.; Rappoport, D. Development of new auxiliary basis functions of the Karlsruhe segmented contracted basis sets including diffuse basis functions (def2-SVPD, def2-TZVPPD, and def2-QVPPD) for RI-MP2 and RI-CC calculations. *Physical Chemistry Chemical Physics* **2015**, *17*, 1010–1017.

(6) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. Self-consistent reaction field model for aqueous and nonaqueous solutions based on accurate polarized partial charges. *Journal of Chemical Theory and Computation* **2007**, *3*, 2011–2033.