Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting Information. General QSPR Protocol for Atomic/Inter-atomic Properties Predictions: Fragments based Graph Convolutional Neural Network (F-GCN)

Peng Gao,[†] Jie Zhang,[*,‡,¶] Hongbo Qiu,[§] and Shuaifei Zhao[‖]

†*School of Chemistry and Molecular Bioscience, University of Wollongong, NSW 2500, Australia*

‡*Centre of Chemistry and Chemical Biology, Bioland Laboratory (Guangzhou Regenerative Medicine and Health-Guangdong Laboratory), Guangzhou 53000, China*

¶*School of Chemical Engineering, East China University of Science and Technology, Shanghai 200237, China*

§*Department of Chemical Engineering, Monash University, Clayton, VIC 3800, Australia*

‖*Institute for Frontier Materials (IFM), Deakin University, Perth, WA, Australia*

E-mail: j.chang@ecust.edu.cn

Encoding of the molecular fragments

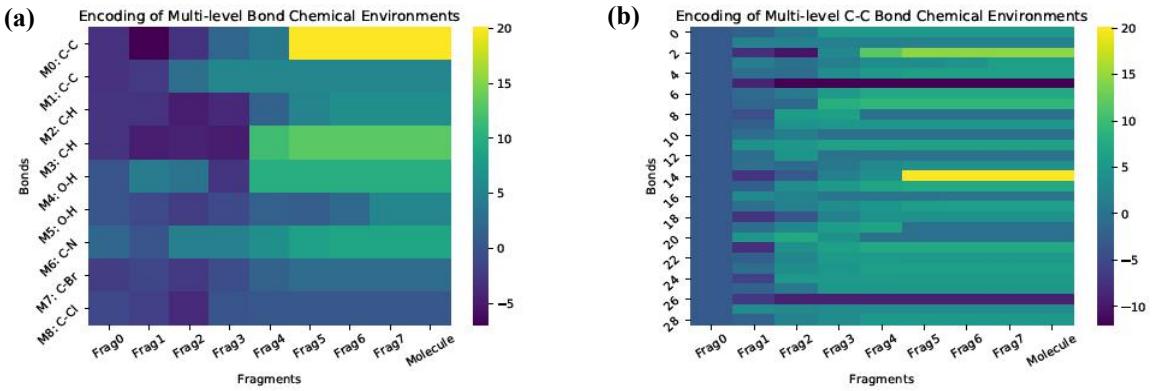

Figure S. 1. (a). Encoding of various kinds of chemical bonds among multi-level fragments by F-GCN; (b). Encoding of C−C bonds among multi-level fragments by F-GCN.

Table 1: The list of the applied molecules for Figure 1a.

Molecule (expressed in SMILES)	Bonding type	Bonding site 1	Bonding site 2
CC(C)(C)c2cccc(c1cccc(C(C)(C)C)c1O)c2O	C−C	8	9
CCCC(C)CC	C−C	2	3
C2=C(c1ccccc1)Cc3ccccc23	C−H	8	N.A
CC(C)(C)CC3c1ccccc1c2ccccc23	C−H	5	N.A
CC1(C)C=C(Cl)C(C)(C)N1O	O−H	10	N.A
CC(=O)c1ccc(O)cc1	O−H	7	N.A
CN(C)Cc1ccccc1	C−N	0	1
BrCc1ccccc1	C−Br	0	1
FC(F)(F)c1ccccc1Cl	C−Cl	9	10

[a)] The experimental BDEs were obtained from iBond 2.0 databank;[1] the SMILES formulas were processed by RDKit package.[2]

Table 2: The list of the applied molecules for Figure 1b.

Molecule (expressed in SMILES)	Bonding type	Bonding site 1	Bonding site 2
C/C=C	C−C	2	3
ClC(Br)C(Cl)Br	C−C	1	3
c4ccc(C(c1ccccc1)(c2ccccc2)c3ccccc3)cc4	C−C	4	5
C=Cc1ccc(CC)cc1	C−C	6	7
C=CCCc1ccccc1	C−C	2	3
FC(F)(F)C(F)(F)C(F)(F)F	C−C	1	4
NCC1CCCCC1	C−C	1	2
CN(C)Cc1ccccc1	C−C	3	4
CCCCC(C)(C)C	C−C	3	4
C=CCC(C)C	C−C	2	3
CCCCC(C)C	C−C	2	3
CNCC(=O)O	C−C	2	3
CCC(C)CO	C−C	2	4
CC(=O)c1ccccc1	C−C	0	1
CC(C)(C)c2cccc(c1cccc(C(C)(C)C)c1O)c2O	C−C	8	9
CC(C)Cc1ccccc1	C−C	1	3
COC(=O)C(F)(F)F	C−C	2	4
CCCCCCC=O	C−C	5	6
c2ccc(c1ccccc1)cc2	C−C	3	4
CCCCCCCCC	C−C	3	4
NCc1cnccn1	C−C	1	2
CC(C)(CN)c1ccccc1	C−C	1	5
CCCC(C)CC	C−C	2	3
CCCCCCC=O	C−C	4	5
C=CC(C)(C)C	C−C	1	2
C=CCC(C)C	C−C	1	2
F/C(F)=C(F)(F)=C(F)	C−C	3	5
COC(=O)CCl	C−C	2	4
C=CCCCC	C−C	1	2

[a)] The experimental BDE data were obtained from iBond 2.0 databank;[1] the SMILES formulas were processed by RDKit package.[2]

The architecture of F-GCN

Our development was based on graph convolutional neural network (GCN). First, starting from the target bond, a series of molecular fragments were generated, as depicted in the Fig. 1 of the main text. The two atoms connected by the target bond make up the first-level fragment. The second level fragment consists of the first-level fragment and the atoms which share covalent bonds with the atoms in the first level. The following fragments are generated following this manner. There are up to 9 levels of fragments until the whole molecule is included. Then these fragments will be encoded by separate GCNs, and a list of bits can be obtained, indicating the fingerprints of fragmentary graphs. And all the fingerprinted fragments contain the original bond and its chemical environment at different levels.

The fragments were solved by RDkit that is a powerful tool for information extraction. The bond type and bond order were recorded for refined learning. At the same time, the atomic descriptors of the two bonded atoms were also included, which compose a 60-dimension vector. The selected atomic descriptors include minimum ring size, CrippenlogP, atomic number, hybridization state, aromatic, Gasteiger charge, number of connected hydrogen, proton donor, proton acceptor, element symbol, CrippenMR, TPSA, LaASA and etc. The selected inter-atomic descriptors include bond type, bond length and etc.

In the readout stage, all the mentioned information can be combined to predict the bond dissociation energy. The molecular and fragmentary graphs were generated within the modified TencentAlchemyDataset. The architecture of F-GCN was designed with reference to SchNet. The neural network was realized within the DGL framework and Pytorch 1.5 with CUDA 10.1. The model was trained on a Linux machine running Ubuntu 20.04 with a NVIDIA 1080 TI graphical processing unit.

To train the model, ADAM optimizer with a initial learning rate of 0.001 was applied. The learning rate was scheduled with ReduceLROnPlateau, with parameters: gamma=0.9, patience=10, min(lr)=0.0001, eps=1e-8.

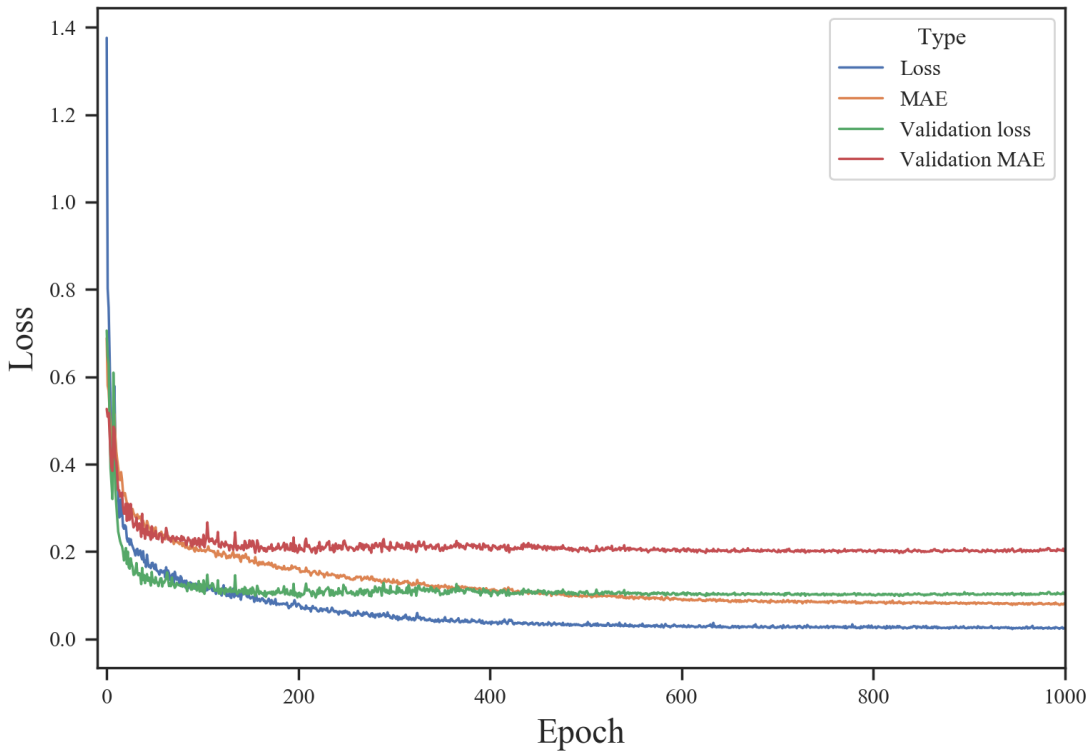

Figure S. 2. The plot of averaging loss of F-GCN with epoch.

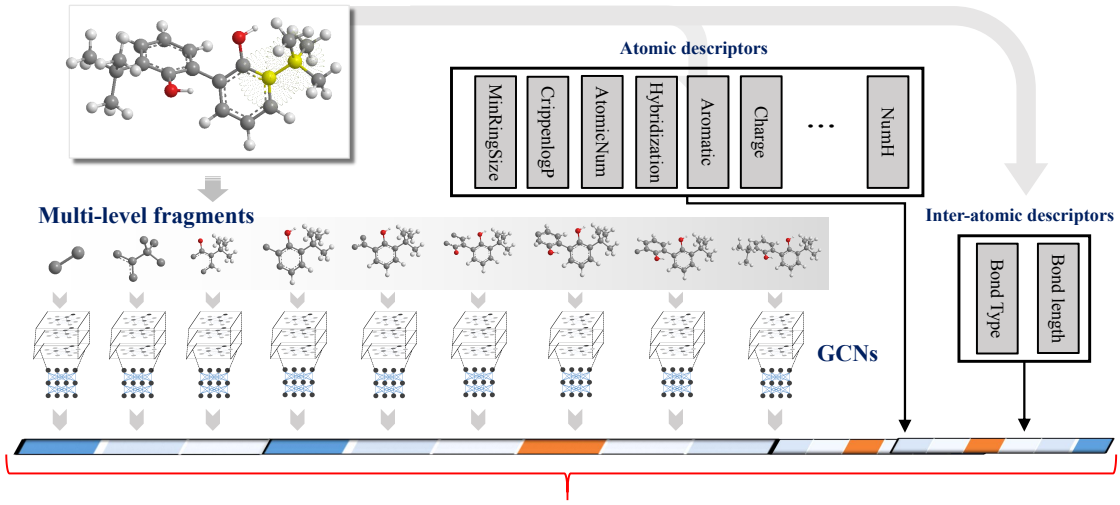

Figure S. 3. The general description of the F-GCN's workflow, designed for chemical properties predictions.

References

(1) Internet Bond-energy Databank (pKa and BDE)—iBonD Home Page. 2020; http://ibond.nankai.edu.cn/.

(2) Landrum, G. A. RDKit: Open-source cheminformatics software. 2018; http://www.rdkit.org.