SolvBERT for solvation free energy and solubility prediction: a demonstration of an NLP model for predicting the properties of molecular complexes

Jiahui Yu[1], Chengwei Zhang[2], Yingying Cheng[2], Yun-Fang Yang[2], Yuan-Bin She[2], Fengfan Liu[1], Weike Su[1], An Su*[2]

1. National Engineering Research Center for Process Development of Active Pharmaceutical Ingredients, Collaborative Innovation Center of Yangtze River Delta Region Green Pharmaceuticals, Zhejiang University of Technology, Hangzhou, 310014, P. R. China
2. College of Chemical Engineering, Zhejiang University of Technology, Hangzhou 310014, P. R. China

Corresponding author:
Prof. An Su
College of Chemical Engineering, Zhejiang University of Technology, Hangzhou
310014, P. R. China
Email: ansu@zjut.edu.cn

Supporting Information

1. Methods for measuring the solubility of out-of-sample solute-solvent combinations

The static gravimetric method was used to determine the solubility of solutes in pure solvents[1]. The detailed processes of the experiments are listed as follows:

(1) The mass of the solute is weighed on an analytical balance at about 1g.

(2) Use SolvBERT to predict the mass of solvent needed and calculate the approximate amount needed

(3) Add the solvent slowly, stirring while adding, and let stand to ensure that there are no obvious fine particles in the supernatant.

(4) Take about 2ml of the supernatant and transfer it to a Petri dish using a syringe with an organic membrane filter (to remove fine particles that cannot be separated by precipitation)

(5) Petri dishes containing the saturated solution were quickly weighed and then placed in a desiccator at 333.15 K for approximately 24 hours to evaporate all solvents.

(6) Remove the residual solid from the Petri dish after evaporation and weigh after cooling to room temperature. Remove from the desiccator, cool to room temperature and weigh. Weigh the process until the Petri dish with residual solids has no significant weight loss.

Table S1. Materials for measuring the solubility of out-of-sample solute-solvent combinations

Chemical name	Molecular formula	Molecular Weight(g/mol)	Mass fraction purity(%)	CAS registry number	Source
O-methyl-N-nitroisourea	$C_2H_5N_3O_3$	119.08	98	57538-27-9	Qingdao Dexin Chemical Co., Ltd.
3-Methyl-2-nitrobenzoic acid	$C_8H_7NO_4$	181.15	98	5437-38-7	Sahn Chemical Technology (Shanghai) Co., Ltd.
2-Amino-5-chloro-3-methylbenzoic acid	$C_8H_8ClNO_2$	185.61	97	20776-67-4	Accela ChemBio Co., Ltd.
1,4-naphthoquinone	$C_{10}H_6O_2$	158.15	97	130-15-4	Anhui Zesheng Technology Co., Ltd.
Anthracene	$C_{14}H_{10}$	178.23	98	120-12-7	Copyright Shanghai Aladdin Biochemical Technology Co., Ltd.
4-Chlorophtalic anhydride	$C_8H_3ClO_3$	182.56	98	118-45-6	Shanghai Eon Chemical Technology Co., Ltd.
Acetone	C_3H_8O	58.08	99.5	67-64-1	Shanghai Maclean Biochemical Technology Co., Ltd.
Ethanol	C_2H_6O	46.07	99.5	64-17-5	Shanghai Maclean Biochemical Technology Co., Ltd.
Methanol	CH_4O	32.042	99.5	67-56-1	Shanghai Lingfeng Chemical Reagent Co., Ltd.
Ethyl acetate	$C_4H_8O_2$	88.11	99.5	141-78-6	Shanghai Maclean Biochemical Technology Co., Ltd.
1-Propanol	C_3H_8O	60.1	99.5	71-23-8	Hangzhou Bongyi Chemical Co., Ltd.
Dichloromethane	CH_2Cl_2	84.933	99.5	75-09-2	Shanghai Maclean Biochemical Technology Co., Ltd

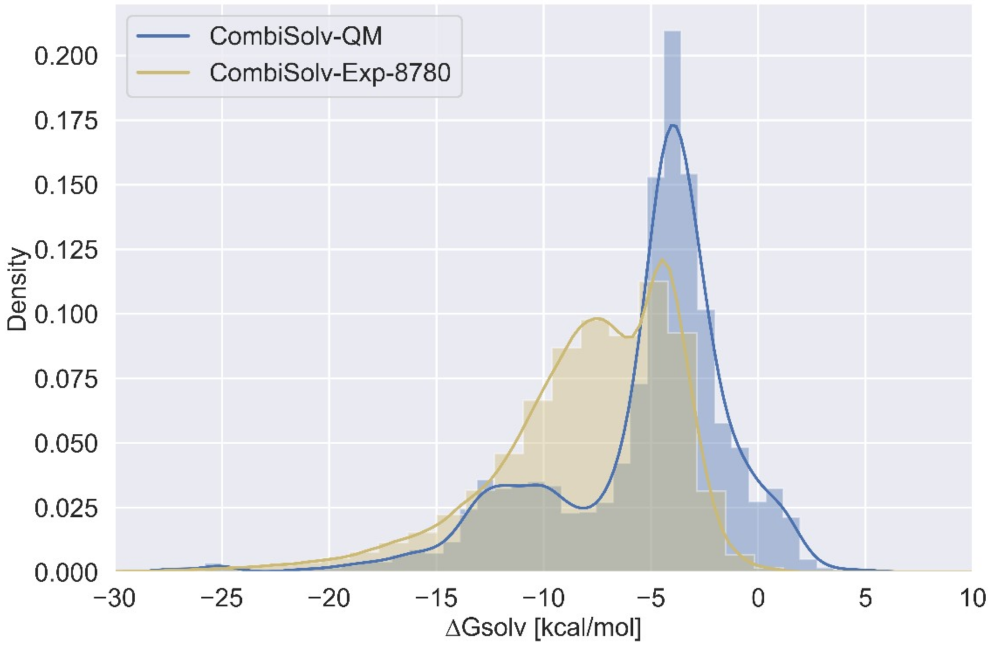

Figure S1. The distribution of ΔG_{solv} for the CombiSolv-QM and CombiSolv-Exp-8780 databases.

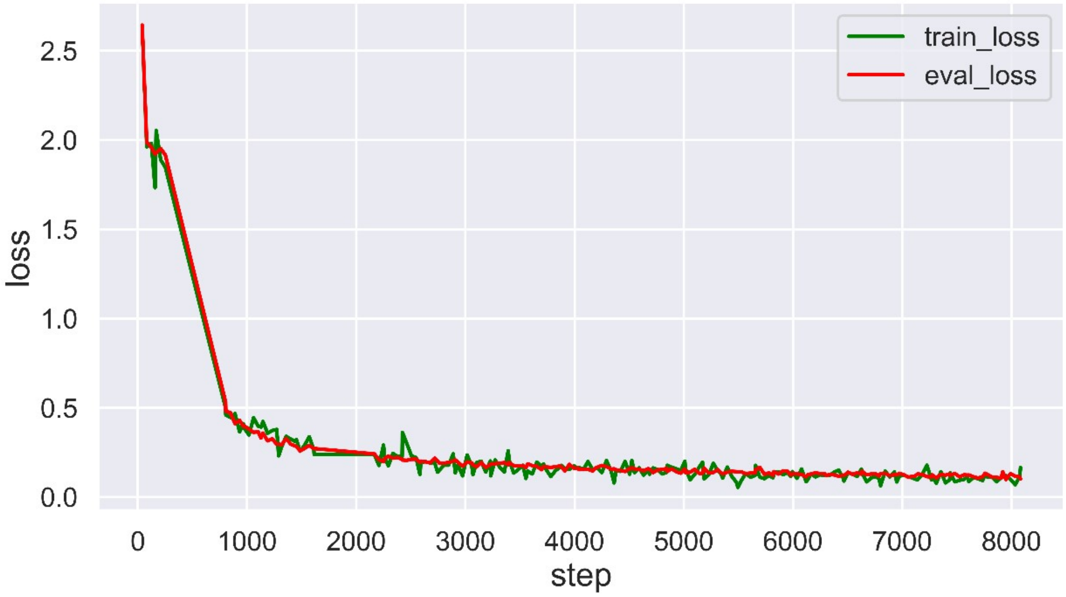

Figure S2. The learning curves for the training of SolvBERT model

Table S2. Hyperparameter optimization for SolvBERT

number	learning_rate	hidden_dropout_rate	R2	RMSE	MAE
1	0.0004	0.1	0.966	0.78	0.50
2	0.0005	0.4	0.951	0.94	0.55
3	0.00009	0.1	0.97	0.77	0.46
5	0.0001	0.2	0.97	0.77	0.46
6	0.00007	0.1	0.978	0.79	0.48
7	0.0001	0.1	0.949	1.00	0.61
8	0.00008	0.2	0.974	0.71	0.49
9	0.00002	0.4	0.974	0.71	0.49
10	0.00008	0.4	0.981	0.60	0.37

Table S3. Hyperparameter optimization for GCN

number	dropout	lr	r2
1	0.2	0.002	0.84
2	0.8	0.0018	0.59
3	0.3	0.0007	0.74
4	0.3	0.002	0.87
5	0.3	0.0002	0.76
6	0.4	0.00002	0.52
7	0.2	0.0016	0.84
8	0.1	0.0004	0.80
9	0.2	0.004	0.89
10	0.1	0.004	0.92

References

1. Huang, H.; Qiu, J.; He, H.; Guo, Y.; Liu, H.; Hu, S.; Han, J.; Zhao, Y.; Wang, P., Solubility Behavior and Polymorphism of N-Acetyl-dl-methionine in 16 Individual Solvents from 283.15 to 323.15 K. *Journal of Chemical & Engineering Data* **2021,** *66* (5), 2182-2191.