

Supporting Information:
Boosting tree-assisted multitask deep learning
for small scientific datasets

Jian Jiang^{1,2}, Rui Wang², Menglun Wang², Kaifu Gao^{*2},
Duc Duy Nguyen² and Guo-Wei Wei^{2,3,4*}

¹*Research Center of Nonlinear Science, College of Mathematics
and Computer Science, Engineering Research Center of Hubei
Province for Clothing Information, Wuhan Textile University,
Wuhan, 430200, P R. China,*

²*Department of Mathematics, Michigan State University,
East Lansing, 48824, Michigan, USA*

³*Department of Electrical and Computer Engineering,
Michigan State University, East Lansing, 48824, Michigan, USA*

⁴*Department of Biochemistry and Molecular Biology, Michigan
State University, East Lansing, 48824, Michigan, USA*

S1. Evaluation metrics

Several different evaluation metrics, including Pearson correlation coefficient (R), root mean squared error (RMSE), mean absolute error (MAE), and Tanimoto coefficient ($S_{A,B}$) [1], are used to evaluate the performances of

*Corresponding author, Email: wei@math.msu.edu

Table S1. Essential introduction of four fingerprints.

fingerprint	Description	Number of features
Daylight	A path-based fingerprint contains 2048 bits and encodes all connectivity pathways in a given length through a molecule (Reference [71] in main text)	2048
MACCS	A substructure keys-based fingerprint with 166 structural keys based on SMARTS patterns (Reference [72] in main text)	166
Estate 1	A topological fingerprint based on electro-topological state indices, encoding the intrinsic electronic state of the atom as perturbed by the electronic influence of all other atoms in the molecule within the context of the topological character of the molecule. It means the number of times that each atom type is hit (Reference [73] in main text)	79
Estate 2	Similar to Estate 1, and it contains the sum of the estate indices for atoms of each type (Reference [73] in main text)	79

different models. These metrics are defined as follows:

$$R = \frac{\sum_{i=1}^N (\log \hat{X}_i - \overline{\log \hat{X}}) (\log X_i - \overline{\log X})}{\sqrt{\sum_{i=1}^N (\log \hat{X}_i - \overline{\log \hat{X}})^2} \sqrt{\sum_{i=1}^N (\log X_i - \overline{\log X})^2}} \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\log \hat{X}_i - \log X_i)^2} \quad (2)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\log \hat{X}_i - \log X_i| \quad (3)$$

and

$$S_{A,B} = \frac{\sum_{i=1}^N x_{iA} x_{iB}}{\sum_{i=1}^N (x_{iA})^2 + \sum_{i=1}^N (x_{iB})^2 - \sum_{i=1}^N x_{iA} x_{iB}}, \quad (4)$$

where \hat{X}_i and X_i stand for the predicted value and the experimented value for the i th molecule, respectively, $\overline{\log \hat{X}}$ and $\overline{\log X}$ are the average logarithm values of predicted and experimented result, respectively, N is the total number of molecules in the test set, and x_{iA} (x_{iB}) denotes the i th feature of molecule A (B).

Here R measures the linear relationship between experimental and predicted values of the test set, RMSE reflects the accuracy of the prediction. MAE is a measure of an average of the absolute difference between experimental and predicted results. The larger R , the smaller RMSE, or the smaller MAE, the more accurate is the prediction. Additionally, $S_{A,B} \in [0, 1]$ is used in the present work to calculate the degree of similarity between two molecule structures. A large average value of $S_{A,B}$ between two datasets means there is a high similarity between them.

S2. Datasets of toxicity prediction

The understanding of toxicity is very important to human health and environmental protection. Four quantitative toxicity datasets are used in the present work as follows. 1) LD₅₀ (oral rat LD50) is used to measure the amount of chemicals that can kill half of the rats when orally ingested. 2) IGC₅₀ (40h Tetrahymena pyriformis IGC₅₀) records 50% growth inhibitory concentration of Tetrahymena pyriformis organism after 40h. 3) LC₅₀ (96h fathead minnow LC₅₀) reports at the concentration of test chemicals in the water in milligrams per liter that results in 50% of fathead minnows to die after 96h. Finally, 4) LC₅₀-DM (48h Daphnia magna LC₅₀) denotes the concentration of test chemicals in the water in milligrams per liter that cause 50% Daphnia magna to die after 48h. The unit of toxicity reported in these four datasets is $-\log_{10}$ mol/L. All of these datasets are available from recent publications or public database [2, 3, 4]. The web site of database is <https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test>. Note that the size of these four datasets varies over a large range from 353 to 7413, which is a challenge for a model to keep the robust and consistent accuracy. We focus on the performance of the relatively small dataset, like LC₅₀-DM with a total size of 353 samples. The statistics of four datasets is given in Table S2.

Table S2. The summary of four datasets used in toxicity prediction.

Dataset	Total size	Train size	Test size	Max value	Min value
LD ₅₀	7413	5931	1482	7.201	0.291
IGC ₅₀	1792	1434	358	6.360	0.334
LC ₅₀	823	659	164	9.261	0.037
LC ₅₀ -DM	353	283	70	10.064	0.117

Table S3. The hyper-parameters in GBDT models for four datasets in toxicity prediction.

Dataset	n_estimators	max_depth	min_sample_split	learning_rate	subsample	max_feature
LD ₅₀	2000	7	3	0.01	0.3	sqrt
IGC ₅₀	10000	7	3	0.01	0.3	sqrt
LC ₅₀	2000	9	3	0.01	0.1	sqrt
LC ₅₀ -DM	2000	9	3	0.01	0.1	sqrt

S3. Partition coefficient (logP) and solvation free energy

The partition coefficient (P) is defined as the ratio of concentrations of a solute in the mixture of two immiscible solvents at equilibrium and is a very important quantity in pharmacology. The logarithm of this coefficient, logP, is one of the key parameters in drug design and discovery. Optimal logP with low molecular weight and low polar surface area play an important role in governing the kinetic and dynamic aspects of drug action [5]. The training set used for partition coefficient predictions was originally compiled by Cheng *et al.* [6], and in present work, it has 8199 compounds based on Hansch *et al.*'s compilation [7]. Additionally, the test set in our test contains 406 molecules and is named as FDA (Food and Drug Administration) [6].

Another dataset related to the drug discovery problem is the solvation free energy, which is collected by Wang *et al.* in present work for the purpose of testing their method named weighted solvent accessible surface area (WSAS) [8]. In this set, the total number of neutral molecules in the 2D SDF format is 383, which is divided into a training set and a test set having 289 and 94 molecules, respectively. Furthermore, the 3D Mol2 molecular structures are all generated by Schrödinger software except 11 molecules that are generated by Discovery Studio software and their IDs are 980, 69689, 95961, 6736, 1390,

6587, 7903, 12302, 7416, 6944, 7422, 11705, and 398. The summary of logP and solvation datasets is given in Table S4. To analyze these datasets, Estate 2, Estate 1, and MACCS fingerprints are used for comparison.

Table S4. The summary of logP and solvation datasets.

	Training set	Test set	Max value	Min value
logP	8199	406	7.57	-3.1
solvation	289	94	4.28	-11.96

S3.1 The performance of GBDT

The details of hyper-parameters used in the GBDT model are shown in Table S5. The performance of GBDT for two datasets with three different fingerprints is presented in Table S6. The current findings are the follows. 1) For the logP dataset, with a large size 8199, fingerprint Estate 2 achieves the best accuracy 0.893 for R^2 . After the process of consensus, the accuracy can be improved to 0.897 for R^2 . 2) For the solvation dataset, with a small size 289, MACCS fingerprint gives rise to the best performance with $R^2 = 0.954$, and the result of consensus is 0.921 for R^2 , which is better than that of logP dataset.

Table S5. The hyper-parameters in GBDT models for logP, logS and solvation datasets.

Dataset	n_estimators	max_depth	min_sample_split	learning_rate	subsample	max_feature
logP	20000	7	3	0.01	0.3	sqrt
logS	10000	7	3	0.01	0.3	sqrt
solvation	2000	9	3	0.01	0.1	sqrt

S3.2 The performance of multitask deep learning

To explore the effect of a large dataset, like logP, on the performance of a small dataset, like solvation, we put two datasets simultaneously as the input vector in the multitask deep learning model. The hyper-parameters for MDL are set as following: 1) the number of hidden layers is 7; 2) the number of neurons is 1000 for the first 4 hidden layers, and 100 for the next 3 hidden

Table S6. Comparison of prediction results R^2 , MAE and RMSE of GBDT model for logP and solvation datasets with Estate 2, Estate 1, MACCS fingerprints and the consensus.

	logP			solvation		
	R^2	MAE	RMSE	R^2	MAE	RMSE
Estate 2	0.893	0.331	0.649	0.882	0.697	1.015
Estate 1	0.870	0.381	0.701	0.881	0.677	1.019
MACCS	0.867	0.380	0.717	0.954	0.500	0.676
consensus	0.897	0.339	0.640	0.921	0.563	0.818

layers; 3) the number of epochs is 1000; 4) learning rate is 0.001 for the first 500 epochs and is 0.0001 for the rest of 500 epochs; 5) batch-size is 16; and 6) the optimizer is SGD (stochastic gradient descent) with the momentum value of 0.5.

Compared to the results of GBDT, we present the performance of MDL with logP and solvation datasets in Table S7, which suggests that with the help of large dataset logP, the accuracies of R^2 with Estate 2 and Estate 1 fingerprints of solvation are boosted by 0.9% and 3.2%, respectively. However, the accuracy of R^2 with MACCS fingerprint decreases by 5.6%. Through the method of consensus, the performance of solvation is improved by 0.2%, which is almost the same as that in Table S6. Opposite to the increase of accuracy of solvation dataset, for the large dataset, logP, except in the case of MACCS fingerprint, all accuracies of prediction with Estate 2 and Estate 1 fingerprints and the consensus method decrease by 4.4%, 1.1%, and 0.8%, respectively. Therefore, in this two-task deep learning of logP and solvation, the task with a small dataset does not benefit too much from the task with a large dataset. This can be explained from the similarity analysis given in Table S18.

S3.3 The performance of two BTAMDL models

As the number of neuron in the last hidden layer of the neural network in BTAMDL model is 100, the feature number of training data for GBDT in BTAMDL 1 is the same as 100. The details of hyper-parameters used in BTAMDL 1 for logP and solvation are in Table S8 and S9, respectively.

The performance of BTAMDL 1 is shown in Table S10. Through the

Table S7. Comparison of prediction results R^2 , MAE and RMSE of MDL for logP and solvation datasets with Estate 2, Estate 1 and MACCS fingerprints, and consensus method.

	logP			solvation		
	R^2	MAE	RMSE	R^2	MAE	RMSE
Estate 2	0.854	0.452	0.782	0.890	0.589	0.981
Estate 1	0.860	0.401	0.751	0.909	0.592	0.902
MACCS	0.884	0.364	0.678	0.901	0.639	0.956
consensus	0.890	0.351	0.666	0.923	0.516	0.829

Table S8. The hyper-parameters in BTAMDL 1 for logP dataset.

Dataset	n_estimators	max_depth	min_sample_split	learning_rate	subsample	max_feature
logP	10000	4	3	0.01	0.2	sqrt

Table S9. The hyper-parameters in BTAMDL 1 for solvation dataset.

Dataset	fingerprint	n_estimators	max_depth	min_sample_split	learning_rate	subsample	max_feature
solvation	Estate 2	2000	9	3	0.01	0.1	sqrt
	Estate 1	3000	2	7	0.004	0.2	sqrt
	MACCS	3600	2	5	0.003	0.2	sqrt

comparison between Table S10 and Table S7, we find that the accuracies R^2 of the consensus method are improved a little by 0.1% and 0.3% for logP and solvation datasets, respectively, which suggests that the algorithm of BTAMDL 1 can indeed further enhance the accuracy of prediction result of small and big datasets.

Next, we try to use BTAMDL 2 to further improve the performance of prediction, in which the feature number of inputs for GBDT is equal to the feature number of fingerprint plus that of activated outputs from MDL. So, the number of the new features in training data is increased from 79 to 179 for Estate 2 and Estate 1, and from 166 to 266 for MACCS, respectively. Table S11 gives the prediction results for logP and solvation datasets. It is demonstrated that with a comparison with BTAMDL 1 in Table S10, for logP, the accuracies R^2 for three fingerprints are decreased except the case with Estate 1, which results in a slight decrease 0.1% of accuracy with the method of consensus. However, for solvation, there is a slight increase of 0.5%

Table S10. Comparison of prediction results R^2 , MAE and RMSE with BTAMDL 1 for logP and solvation datasets with Estate 2, Estate 1, and MACCS fingerprints and consensus method.

	logP			solvation		
	R^2	MAE	RMSE	R^2	MAE	RMSE
Estate 2	0.855	0.449	0.779	0.894	0.616	1.002
Estate 1	0.858	0.402	0.757	0.913	0.582	0.886
MACCS	0.889	0.357	0.670	0.900	0.639	0.960
consensus	0.891	0.354	0.665	0.926	0.513	0.824

of accuracy with the method of consensus. Therefore, from these results, we can confirm that BTAMDL 2 may not further enhance the accuracy of prediction. This is due to the redundant features as analyzed below.

Table S11. Comparison of prediction results with BTAMDL 2 for logP and solvation datasets with three fingerprints and the consensus method.

	logP			solvation		
	R^2	MAE	RMSE	R^2	MAE	RMSE
Estate 2	0.851	0.471	0.792	0.894	0.602	0.964
Estate 1	0.859	0.416	0.757	0.913	0.580	0.880
MACCS	0.881	0.392	0.694	0.904	0.638	0.937
consensus	0.889	0.361	0.669	0.928	0.514	0.808

Figure S1 shows the relationship between the number of features and the prediction accuracy based on the importance order of the feature for logP and solvation datasets. In Figure S1 (a), for logP dataset with fingerprints Estate 2 and Estate 1, the value of R^2 increases quickly as the number of features increases. The maximal value is reached with 11 features (see inset), which is just 6.1% of the total of 179 features. For MACCS, there is little fluctuation in R^2 with the increase of the number of features, compared to that of the other two fingerprints. The maximum value of R^2 is reached with around 21 features (see inset), which is just 7.9% of the total 266 features. These results suggest that most of the features in the above training data are redundant, which gives the reason why the performance of BTAMDL 2 is not improved in Table S11. In Figure S1 (b), for solvation dataset, similar behavior of the accuracy R^2 with the number of features is found, that is,

the value of R^2 sharply increases at a small number of features. Additionally, a small number of features is needed for R^2 to become steady, compared to the logP dataset in Figure S1 (a). They are only 2.8% and 1.7% of the total features for Estate 1 and Estate 2, respectively (see inset). In particular, for MACCS, even one feature could make the model robust and accurate (see inset). Hence, in this case, redundant features do not enhance the accuracy of the model but make the model time-consuming.

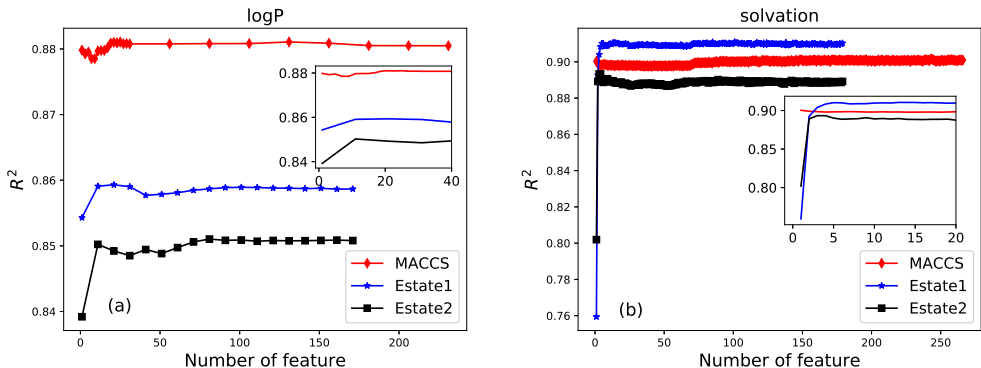


Figure S1. Tendency of the prediction accuracy R^2 along with number of features for three different fingerprints, MACCS, Estate 1, and Estate 2 of logP and solvation datasets.

S3.4 Comparison with other methods

A detailed comparison with other logP prediction methods is shown in Table S12, where the methods of GBDT-ESTD⁺-2-AD and MT-ESTD-1 are based on 3D descriptors [5]. Since GBDT-ESTD⁺-2-AD model contains some molecules from the NIH dataset in its training set, its performance is better than those of our models. As pointed out in the literature [9], ALOGPS has a similar problem: its training set includes all of the compounds in the FDA test set for the PHYSPROP database [10]. Therefore, compared with methods with a clear separation between the training set and test set, the present methods performed the best.

Table S12. Comparison of logP predictions between our models (green part) with other different methods (pink part) on the FDA test set.

Method	R^2
GBDT-ESTD ⁺ -2-AD (2D+3D) [5]	0.935
MT-ESTD-1 [5]	0.920
ALOGPS (2D) [6]	0.908
GBDT consensus	0.897
BTAMD1 1	0.891
MDL consensus	0.890
BTAMD1 2	0.889
XLOGP3 (2D) [6]	0.872
XLOGP3-AA (2D) [6]	0.847
CLOGP (2D) [6]	0.838
TOPKAT (2D) [6]	0.815
ALOGP98 (2D) [6]	0.800
KowWIN (2D) [6]	0.771
HINT (2D) [6]	0.491

S4. Aqueous solubility (logS) and Solvation

In this section, we use aqueous solubility (logS) dataset, instead of logP dataset, as one of the test data in the prediction, aiming at finding out the influence of different data size on the improvement of the performance of small dataset. Like the case, we also want to identify the reason behind this improvement, if possible.

In drug discovery and other pharmaceutical fields, aqueous solubility, denoted by its logarithm value logS, is very important for excluding molecules with undesirable water solubility on early stages since the solubility greatly influences many processes, namely, absorption, distribution, metabolism, and elimination [11, 12]. There are several well-defined aqueous solubility datasets used in prediction models [13]. In the present work, we test our models on a relatively small dataset where has 1290 samples in the train set [14] and 21 samples in the test set [9]. The summary of logS and solvation datasets is given in Table S13. In these datasets test, the Estate 2, Estate 1, and MACCS fingerprints are used for comparison.

Table S13. The summary of logS and solvation datasets used.

	Training set	Test set	Max value	Min value
logS	1290	21	0.39	-8.08
solvation	289	94	4.28	-11.96

S4.1 The performance of GBDT

The details of the hyper-parameters used in the GBDT model are shown in Table S5. The performance of GBDT for two datasets with three different fingerprints is presented in Table S14. It is seen that: 1) for the logS dataset, with a size of 1290, MACCS fingerprint can achieve the accuracy 0.896 for R^2 . However, the consensus accuracy is 0.886; 2) for the solvation dataset, with a size of 289, MACCS fingerprint obtains the best performance of $R^2 = 0.954$, and the result of consensus is 0.921 for R^2 , which is better than that of logS dataset.

Table S14. Comparison of prediction results R^2 , MAE and RMSE of GBDT model for logS and solvation datasets with Estate 2, Estate 1, MACCS fingerprints and the method of consensus.

	logS			solvation		
	R^2	MAE	RMSE	R^2	MAE	RMSE
Estate 2	0.804	0.646	0.900	0.882	0.697	1.015
Estate 1	0.870	0.550	0.792	0.881	0.677	1.019
MACCS	0.896	0.608	0.710	0.954	0.500	0.676
consensus	0.886	0.570	0.732	0.921	0.563	0.818

S4.2 The performance of multitask deep learning

To improve the performance of small dataset, solvation, a two-task MDL model with logS and solvation datasets is constructed. As the size of logS dataset is smaller than that of logP dataset, some parameters in neural networks are changed as follows: 1) one hidden layer; 2) the number of neurons in hidden layer is 900; 3) number of epoch is 1900; 4) learning rate is 0.001; 5) batch-size is 2; 6) the optimizer is SGD (stochastic gradient descent) with a momentum value of 0.2. We show the performance of the prediction of the MDL model in Table S15. It is seen that for the small set, solvation, all

values of R^2 are improved in three fingerprints and the consensus method, by 5.7%, 7.0%, 0.6%, and 4.6%, respectively, compared to those with GBDT model in Table S14. For the large set, logS, though all values of R^2 with three fingerprints are dropped. However, the R^2 of the consensus method increases 0.6% and is also better than that in Table S14. Therefore, with MDL, logS and solvation predictions are benefited mutually. The prediction with a small data size can be significantly improved by a large data size.

Table S15. Comparison of prediction results R^2 , MAE and RMSE of MDL for logS and solvation datasets with Estate 2, Estate 1, MACCS fingerprints and consensus method.

	logS			solvation		
	R^2	MAE	RMSE	R^2	MAE	RMSE
Estate 2	0.713	0.971	1.167	0.932	0.517	0.774
Estate 1	0.845	0.643	0.847	0.943	0.460	0.713
MACCS	0.839	0.616	0.817	0.960	0.434	0.604
consensus	0.891	0.590	0.723	0.963	0.407	0.569

S4.3 Comparison with other methods

Table S16 presents a comparison between our models and other methods on the logS dataset. Our models outperform all other state-of-the-art 2D and 3D methods. In Table S17, the comparison is given between different methods on the solvation dataset. Our models outperform all other methods. Note that FFT [15] and WSAS [2] did not give R^2 results.

Table S16. Comparison of logS predictions between our models (green part) with other different methods (pink part).

Method	R^2
MDL consensus	0.891
GBDT consensus	0.886
MT-ESTD ⁺ -1 (3D) [5]	0.884
Drug-LOGS (2D) [9]	0.884
Klopman MLR (2D) [14]	0.846

Table S17. Comparison of solvation predictions between our models (green part) with other different methods (pink part). Reference [36] is in main text.

Method	R^2	MAE
MDL consensus	0.963	0.407
GBDT consensus	0.921	0.563
$\text{EIC}_{\text{E},3.5,0.3;\text{E},2.5,1.3}^{\text{HH}}$ [36]	0.920	0.558
Consensus ^H [36]	0.920	0.567
FFT [15]	NA	0.570
$\text{EIC}_{\text{E},3.5,0.3}^{\text{H}}$ [36]	0.904	0.575
$\text{EIC}_{\text{L},3,1.3}^{\text{H}}$ [36]	0.906	0.592
$\text{EIC}_{\text{L},3,1.3;\text{L},6.5,0.3}^{\text{HH}}$ [36]	0.907	0.608
WSAS [2]	NA	0.660

S4.4 The similarity analysis between datasets

From the prediction results of MDL in subsection S3.2 and S4.2, it is seen that the transfer learning enhancement of a small dataset from a larger dataset is not always proportional to the size of the larger dataset. The logP data (8199) is much larger than that of logS data (1290). But solvation prediction R^2 is enhanced more from the logS dataset than from the logP dataset. This behavior can be understood from the similarity analysis. As shown in Table S18, the similarity between logP and solvation datasets averaged over three different fingerprints is 0.609, while that between logS and solvation datasets is 0.645. This similarity analysis explains why in the transfer learning, the logS dataset can provide more enhancement to the solvation prediction than the logP dataset. Therefore, one can conclude that the similarity between the datasets in multitask learning is essential to the smaller dataset’s prediction improvement.

References

- [1] Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **2015**, *7*, 1-13.

Table S18. The similarity of two large datasets, logP and logS, with small dataset, solvation. The number in the bracket is the size of the training set.

fingerprint	logP (8199)	logS (1290)
Estate 2	0.776	0.847
Estate 1	0.745	0.814
MACCS	0.306	0.276
Increment of R^2 for solvation	0.2%	3.9%

- [2] Martin, T. User’s guide for test (version 4.2) (toxicity estimation software tool): A program to estimate toxicity from molecular structure **2016**.
- [3] Akers, K. S.; Sinks, G. D.; Schultz, T. W. Structure-toxicity relationships for selected halogenated aliphatic chemicals. *Environ. Toxicol. Pharmacol.* **1999**, *7*, 33-39.
- [4] Zhu, H.; Tropsha, A.; Fourches, D.; Vamek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, IV. Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* **2008**, *48*, 766-784.
- [5] Wu, K.; Zhao, Z.; Wang, R.; Wei, G. W. TopP-S: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. *J. Comput. Chem.* **2018**, *39*, 1444-1454.
- [6] Cheng, T. J.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of octanol-water partition coefficients by guiding an additive model with knowledge. *J. Chem. Inf. Model.* **2007**, *47*, 2140-2148.
- [7] Hansch, C.; Leo, A.; Livingstone, D. J. Exploring QSAR Fundamentals and Applications in Chemistry and Biology. Pesticide Biochemistry and Physiology **1996**.
- [8] Wang, J.; Wang, W.; Huo, S.; Lee, M.; Kollman, P. A. Solvation model based on weighted solvent accessible surface area. *J. Phys. Chem. B*, **2002**, *B*, 5055-5067.

- [9] Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266-275.
- [10] Howard, P.; Meylan, W. Physical/chemical property database (physprop) **1999**.
- [11] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **2001**, *1*, 3-26.
- [12] Di, L.; Kerns, E. H. Biological assay challenges from compound solubility: strategies for bioassay optimization. *Drug Discovery Today.* **2006**, *11*, 446-451.
- [13] Wang, J.; Krudy, G.; Hou, T.; Zhang, W.; Holland, G.; Xu, X. Development of reliable aqueous solubility models and their application in druglike analysis. *J. Chem. Inf. Model.* **2007**, *47*, 1395-1404.
- [14] Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474-482.
- [15] Wang, B.; Wang, C.; Wu, K.; Wei, G. W. Breaking the polar-nonpolar division in solvation free energy prediction. *J. Comput. Chem.* **2018**, *39*, 217-233.