**Improving Prediction Performance on Solute Parameters Using Multi-Task Relational Graph Convolutional Networks with Explicit Hydrogens**

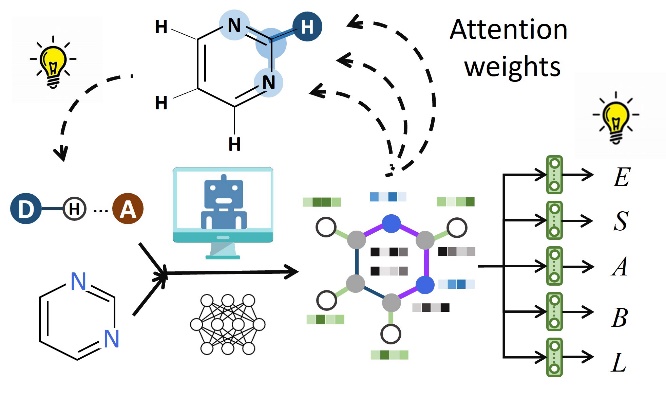
Zijun Xiao,† Minghua Zhu,†,‡ Jingwen Chen†,\*

† Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), Dalian Key Laboratory on Chemicals Risk Control and Pollution Prevention Technology, School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China

‡ Key Laboratory of Integrated Regulation and Resources Development of Shallow Lakes of Ministry of Education, College of Environment, Hohai University, Nanjing 210098, China

\* Corresponding Author: Jingwen Chen, e-mail: jwchen@dlut.edu.cn

**Table of Contents**



**◼ ABSTRACT**

This study proposed multi-task (MT) learning coupled with relational graph convolutional networks with attention weights (RGAN) and explicit hydrogens (abbreviated as MT-RGAN-H architecture) to construct prediction models on the solute parameters including excess molar refraction, dipolarity/polarizability, H-bond acidity (*A*), H-bond basicity (*B*), and logarithmic hexadecane−air partition coefficient. The resulting MT-RGAN-H model was proved to outperform single-task (ST) machine learning models, ST-RGAN models, MT-RGAN model, and previous models on the solute parameters. Importantly, the MT-RGAN-H model improved the prediction accuracy for *A* (5.6% increase) and *B* (4.3% increase) over previous models. The MT-RGAN-H architecture also enables learning from multiple interrelated endpoints, even from endpoints with limited labeled cases, overcoming challenges of data insufficiency in some tasks. By visualizing the attention weights, the MT-RGAN-H model was well interpreted. The predicted solute parameters were further employed to predict six physicochemical parameters of chemicals, achieving better prediction accuracy over the previous optimal models. Therefore, this study provides an integrative “end-to-end” prediction scheme for the solute parameters, laying a foundation for accurately predicting the environmental partition behavior of chemicals. The MT-RGAN-H architecture can be further applied to construct prediction models on other interrelated endpoints, supporting sound management of chemicals.

**◼ KEYWORDS**

*relational graph convolutional networks, multi-task learning, solute parameters, linear solvation energy relationship, attention mechanism, explicit hydrogen atoms*

**◼ SYNOPSIS**

A multi-task relational graph convolutional network model with explicit hydrogen atoms and attention weights was developed for accurately predicting solute parameters of chemicals.

1. **INTRODUCTION**

The solute parameters, *E*, *S*, *A*, *B*, *V*, and *L*, are basic constituents of linear solvation free energy relationships (LSERs) that have been widely used for accurately predicting various physicochemical parameters characterizing solvation, partition, and non-reactive toxicity of chemical pollutants.1,2 Herein, *E*, *S*, *A*, *B*, *V*, and *L* represent solute excess molar refraction [(cm3mol−1)/10], dipolarity/polarizability, H-bond acidity, H-bond basicity, McGowan molar volume [(cm3mol−1)/100], and logarithmic hexadecane−air partition coefficient, respectively.1,2

Nevertheless, there are only about 3800 compounds with experimentally determined values of the solute parameter,3 which is far less than the number of compounds used in the global market.4,5 The unavailability of the solute parameter values limits the practical application of the LSERs. Therefore, it is important to develop prediction models on the solute parameters, since experimental determination of the parameters is costly, laborious, and restricted by the availability of authentic chemical standards.

Although the quantum chemical methods, e.g. density functional theory (DFT), can be employed to calculate *E* and *V* values with high accuracy,6,7 and DFT with implicit solvent models can be employed to predict *L*;6,7 the calculation is time-consuming, with low-throughput, and depends on expertise in the theoretical calculation. The fragment contribution method has also been developed to estimate the solute parameter values.8,9 Although the fragment contribution method has easiness of calculation, it is prone to errors for compounds with complex steric structures or with multiple functional groups.10 Due to the involvement of multiple steps in these calculation methods, some of which involve commercial software, these methods are not convenient for programming and application. Therefore, it is preferable to develop integrative “end-to-end” models. “End-to-end” refers to training a single model to perform a task from raw input to final output, without any intermediate steps or feature engineering.10

Machine learning (ML) may serve as an alternative method for predicting the solute parameters with high-throughput and high prediction accuracy.11 Previous ML models on the solute parameters rely on manually selected molecular structural features, such as fragments, fingerprints, or descriptors, which are limited in their ability to fully characterize molecular structural information.11,12 Graph convolutional network (GCN) models based on molecular graph (MG) representation using atoms as nodes and bonds as edges, can automatically learn informative molecular features without manual intervention.13,14 Ulrich and Ebert15 developed GCN models to predict the solute parameters and achieved root mean squared error (*E*RMS) of 0.12 ~ 0.46 on a validation set of about 600 compounds.

However, in the previous MG-based chemical property prediction models, hydrogen atoms and the bonds connected to the hydrogen atoms (referred to as Q−H covalent bonds, where Q represents a non-hydrogen atom) were typically ignored.16,17 It is known that the hydrogen atoms and Q−H covalent bonds can significantly impact the stereochemistry, stability, and reactivity of molecules.18 The Q−H covalent bonds also play a crucial role in determining the overall properties of a molecule, including its acidity, electrophilicity, and hydrophobicity.19 Neglecting the hydrogen atoms in the MGs can lead to some molecules that are less distinguishable, such as benzene and hexane.20 Specially, GCN prediction of the parameters *A* and *B*, could be highly sensitive to the MG representation embedding hydrogen atoms and Q−H covalent bonds. Therefore, it is necessary to consider hydrogen atoms in the construction of GCN models for predicting the solute parameters.

As an extension of the conventional GCN, the relational graph convolutional network (RGCN) was proved to perform better in predicting the toxicity of compounds than the GCN.16 GCN is based on homogeneous graphs and regards all edges in the graph as the same type.16 However, molecular graphs are heterogeneous in nature, with multiple types of edges, including single bonds, double bonds, and various atom pairs. RGCN models can handle heterogeneous graphs and update the hidden states based on different edge types, enriching molecular representations,16,21 which may explain its better performance than GCN models. RGCN has achieved good performance in screening carcinogenic chemicals and predicting 50% lethal concentrations for *Daphnia magna*.16 It deserves investigation whether RGCN exhibits good performance in predicting the solute parameters.

It would be better for ML models to have interpretability. To enhance the interpretability of RGCN models, an attention mechanism was incorporated by assigning different attention weights to different atoms, thereby generating customized features for each prediction task.16 The attention weights allow models to automatically identify the most informative features.22 For example, graph attention network (GAT) which introduces the attention mechanism as a substitute for the statically normalized convolution used in the GCN, has been successfully applied to identify environmentally persistent, bio-accumulative, and toxic (PBT) chemicals.17 Therefore, RGCN models with the attention mechanism should be developed to predict the solute parameters.

Multi-task (MT) learning as a general learning paradigm has been shown to outperform single-task (ST) learning in many tasks including toxicity prediction.16 MT learning can utilize more data from different tasks than ST learning. With more data, MT learning can gain more robust and universal representations for multiple tasks, leading to better knowledge sharing among tasks, better performance of each task, and lower overfitting risk in each task.23 Therefore, MT models for predicting the solute parameters should be explored.

This study developed an MT-RGCN architecture based on MGs with explicit hydrogen atoms and with an attention layer to the network (hereinafter referred to as MT-RGAN-H). The resulting MT-RGAN-H model was proved to outperform ST-ML models, ST-RGAN models, MT-RGAN model, and previous prediction models on the solute parameters. This study also defined the applicability domains (ADs) for the MT-RGAN-H model, since ADs are necessary for regulatory acceptance of prediction models.17 The solute parameters predicted by the MT-RGAN-H model were further employed to develop LSER models for predicting selected physicochemical parameters characterizing environmental partition of compounds. The MT-RGAN-H model together with the LSER models may serve as versatile tools for high-throughput prediction of the environmental partition behavior of diverse organic compounds.

1. **MATERIALS AND METHODS**
   1. **Data Collection and Curation.** Datasets consisting of 7974 compounds were retrieved from the UFZ-LSER Database.3 The UFZ-LSER Database contains 3885 compounds for which the values of *E*, *S*, *A*, *B*,and *L* were derived from different experimental studies reported in different literature. Another source of data in the UFZ-LSER Database was the Absolv Database9 which comprises solute parameter values of 7881 compounds, for which sources or methods to derive the solute parameters were not specified.

The data were preprocessed with Python 3.8.24 The simplified molecular input line entry system (SMILES) codes were converted into canonical SMILES codes and duplicate molecules were removed. The mixture and inorganic molecules were removed since they failed to be converted into MGs. Average values of the solute parameters were adopted when multiple values were reported for a compound.

The preprocessing resulted in 3838 compounds for which the solute parameter values have specified experimental sources in the literature. Among the 3838 compounds, the number of compounds with *E*, *S*, *A*, *B*,and *L* values is 3407, 3258, 3074, 2703, and 3308, respectively. Each dataset was randomly divided into training sets and validation sets with a ratio of 4:1. The training and validation sets on the five parameters were then combined as the training and validation sets for the MT model. Each training set was randomly and equally divided into 10 folds, with 9 folds used as internal training sets and 1 fold used as an internal test set. This process was repeated iteratively until each fold had been used once as the internal test set, which is termed 10-fold cross validation.

The preprocessing on the data from the Absolv Database9 led to 7834 compounds, among which the number of compounds with the *E*, *S*, *A*, *B*,and *L* values, after excluding the compounds covered in the aforementioned datasets of the 3838 compounds, are 4534, 4161, 4814, 4419, and 3611, respectively. These datasets were employed as supplementary external validation sets. The data curation process was illustrated in Figure S1 in Supporting Information (SI).

Datasets on selected physicochemical properties were also constructed based on previous studies,25-30 including liquid vapor pressure (*P*LV, mmHg), water solubility (*S*W, mol/L), Henry’s law constants [*K*H, (atm·m3)/mol]*,* *n*-octanol/air partition coefficients(*K*OA), *n*-octanol/water partition coefficients(*K*OW), and soil organic carbon normalized adsorption coefficients (*K*OC, L/kg). After the pre-processing, the datasets on *P*LV, *S*W, *K*H*, K*OA, *K*OW, and *K*OC,contain 5850, 9433, 2137, 588, 17243, and 837 compounds with the corresponding empirical values, respectively. The solute parameters predicted by the MT-RGAN-H model were adopted to develop LSER models on the physicochemical properties, and the LSER models were compared with those based on the solute parameters from other sources.

* 1. **Construction of MT-RGAN-H Model.** One MT-RGAN-H model, one MT-RGAN model, five ST-RGAN-H models, and five ST-RGAN models were constructed for comparison. All the RGAN models have four modules according to their functions: input module, general feature extractor (GFE), customized feature extractor (CFE), and predictor.The framework of the MT-RGAN-H model is shown in Figure 1(a). The GFE contains two RGCN layers and the CFE contains one layer with attention weights and three fully connected layers (FCLs). This combination was proved to be optimal in previous studies.16 The input module and GFE were consistent for the MT and ST models, while the CFE and predictor varied depending on the tasks. PyTorch31 was adopted to build the RGAN models.

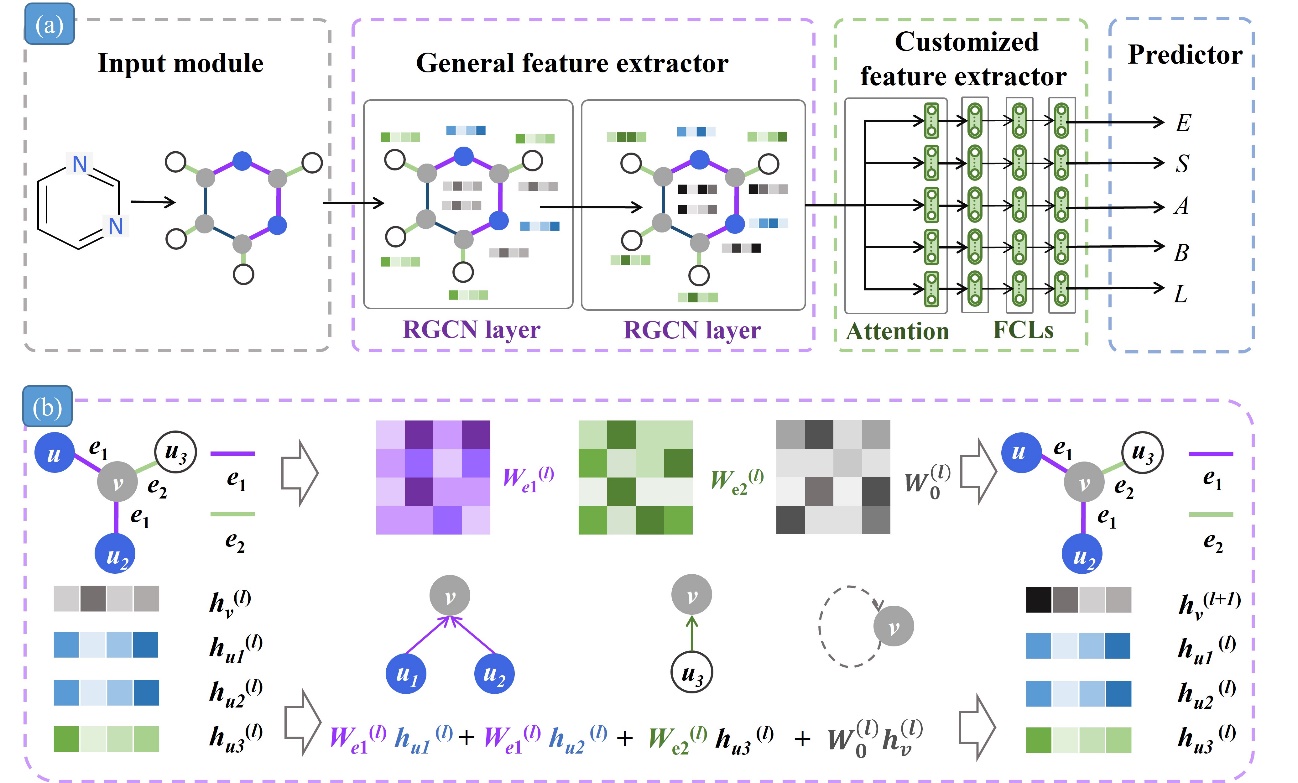


Figure 1. (a) Overview of the MT-RGAN-H architecture; (b) A schematic diagram illustrating propagation in a part of molecular structure with 4 nodes, where *v* (in gray) represents a center atom; *h*(*l*) and *h*(*l*+1) show the node embedding before and after the message passing step, respectively; the neighboring nodes are labeled as *u* (in blue and white); *e* represents edge (relation), and is marked with purple and green to distinguish different relations; *W* represents the weight matrix.

With the input module, the SMILES codes were transformed into MGs using the Deep Graph Library.32 The MGs were encoded by a set of atoms (*v* ϵ *V*g) and bonds (*e* ϵ *E*g), where *v* and *e* represent each atom or bond, respectively. The features of the MGs used in the RGAN model are listed in Table 1. Unlike the previously constructed GNN models,17 the RGAN-H introduced hydrogen atoms as atom features and new atom pairs as bond features (also relation types). The MGs with explicit hydrogen atoms have more features than the conventional ones: hydrogen atom nodes and five hydrogen-related atom pairs (['CH', 'HC'], ['OH', 'HO'], ['NH', 'HN'], ['PH', 'HP'], ['SH', 'HS']). All the atom and bond features were calculated by the RDKit toolkit.33

**Table 1.** Atom and bond features of the molecular graphs used in the RGAN models.

|  |  |  |
| --- | --- | --- |
| **Atom feature** | **Size** | **Description** |
| Atom symbol | 18/19*a* | ['H'*a*, 'B', 'C', 'N', 'O', 'F', 'Si', 'P', 'S', 'Cl', 'As', 'Se', 'Br', 'Te', 'I', 'At', 'Fe', 'Hg', 'other'] |
| Degree | 7 | Number of covalent bonds [0, 1, 2, 3, 4, 5, others] |
| Formal charge | 1 | Electrical charge |
| Radical electrons | 1 | Number of radical electrons |
| Hybridization | 6 | [sp, sp2, sp3, sp3d, sp3d2, other] |
| Aromaticity | 1 | Whether the atom is part of an aromatic system |
| Hydrogens | 5 | Number of connected hydrogen atoms [0, 1, 2, 3, 4] |
| Chirality | 1 | Whether the atom is a chiral center [0/1] |
| Chirality type | 2 | [R, S] |
| **Bond Feature** | **Size** | **Description** |
| Bond Type | 4 | [single, double, triple, aromatic] |
| Conjugation | 1 | Whether the bond is in the conjugated [0/1] |
| Ring | 1 | Whether the bond is in the ring [0/1] |
| Stereo | 4 | Stereo configuration of a bond [StereoNone, StereoAny, StereoZ, StereoE] |
| Atom Pair | 13/18*a* | ['CC'], ['CN', 'NC'], ['ON', 'NO'], ['CO', 'OC'], ['CS', 'SC'], ['SO', 'OS'], ['NN'], ['SN', 'NS'], ['CCl', 'ClC'], ['CF', 'FC'], ['CI', 'IC'], ['CBr', 'BrC'], ['CH',' HC']*a*, ['OH', 'HO']*a*, ['NH', 'HN']*a*, ['PH', 'HP']*a*, ['SH', 'HS']*a*, ['others'] |
| *a* The features were only applied in the RGAN models based on MGs with explicit hydrogen atoms. | | |

The GFE consisted of two RGCN layers and adopted a feature-sharing architecture in the MT models. Propagation for each node *ν* in the RGCN layerswas calculated via:34

(1)

where *hν*(*l*+1) is the state vector of node *ν* after *l* + 1 iterations*;* *e* stands for different types of edges (chemical bonds); *E*g denotes the set of all edge types centered on node *v*; *u* stands for the neighbor node connecting to node *ν*; *Nve* denotes the set of the neighbor nodes of node *ν* under the relation *e*; *We*(*l*) represents weight matrix for node *u* with edge *e* (ϵ *E*g) after *l* iterations; *hu*(*l*)and *hv*(*l*) denotes state vectors of neighbor node *u* and node *v* in *l* iterations, respectively; *W*0(*l*) represents weight matrix for node *v* in *l* iterations; ReLU is a linear rectification function commonly used in neural networks.34

Figure 1(b) illustrates the propagation of node *v* in a RGCN layer. The RGCN layer explicitly incorporates the bond information (e.g., *e*1 and *e*2, under the relation *e* ϵ *E*g) by training the multiple weight matrices (e.g., *We*1 and *We*2), in contrast to some algorithms such as GCN and GAT that train only one weight matrix.34 After the iterations by eq. (1), the state vector *hv*(*l*+1) was obtained, which was also referred to as the general feature of node *v* that was to be further utilized for extracting customized features.

The CFE assigns different attention weights to different atoms (nodes) for a specific task. The attention weight on node *ν* (*ω*ν) was calculated as follows:22

*ω*ν =sigmoid (*W*·*h*ν + *B*) (2)

where sigmoid is an activation function that limits the attention weight *ω*ν between 0 and 1;22 *W* and *B* represent trainable matrices with *B* representing bias*.* The customized feature of node *ν* was calculated as the dot product of *ω*ν and *h*v*.* To enable the CFE to learn from the features of the GFE, and to customize the features for specific tasks, FCLs were added.

A loss function as indicated in eq. (3) was used in the predictor to measure differences between the predicted and experimental values, and to determine whether the model has converged.35

(3)

where  represents predicted the value of molecule *m* from task *t*, *yt,m* represents the experimental value of molecule *m* corresponding with task *t*, *M*represents the number of molecules in the training sets, and *T* represents the number of the tasks. It was the purpose of the training process to optimize a set of hyperparameters by minimizing the *Loss* value.

With all the model parameters optimized “end-to-end”, the optimal hyperparameters of the RGAN models were obtained as: number of neurons of each RGCN layer is 256; number of RGCN layer is 2; number of neurons of the layer with attention weights is 256, number of neurons of each FCL is 256; number of FCLs is 3; dropout rate of each RGCN layer is 0.2; dropout rate of each FCL is 0.2; optimizer is Adam35 with an initial learning rate of 10-3 and a weight decay of 10-5; batch size is 256; number of epochs is 300, and early stopping is used to avoid overfitting with a maximum epoch to 50.

* 1. **Construction of ST-ML models.** For comparison, a total of 36 ST-ML models for each solute parameter were built based on 4 groups of molecular descriptors or 5 fingerprints, coupled with four machine learning algorithms.

The 4 groups of molecular descriptors include those calculated by Mordred,36 Gaussian,37 Dragon (Ver. 06),38 and the coalition of Gaussian and Dragon (CGD). The Mordred descriptors were calculated by the Mordred module36 in Python from the SMILES codes. The Gaussian (09 program with revision D.01) was adopted to calculate the quantum chemical descriptors with the method of B3LYP/6-31G(d) (The LANL2DZ basis set for I atom).37,39,40 The descriptors were Z-score standardized using the Standard Scaler function in the Scikit-learn module.41 The descriptors with zero variance or with a standard deviation below 0.001 were removed.

The 5 groups of fingerprints include extended-connectivity fingerprints with a diameter of 4 bonds (ECFP4) and 6 bonds (ECFP6),42 Avalon fingerprints (AFP),43 Molecular ACCess System (MACCS),44 and RDKit fingerprints (RDKFP).33 They were generated using the RDKit toolkit.33 Dimensions of the descriptors and fingerprints (Table S1), and values of the Gaussian and Dragon descriptors are provided in SI.

The four ML algorithms are random forest (RF),45,46 support vector machines (SVM),47 gradient boosting regression (GBR),48 and artificial neural networks (ANN),49 which were executed by the Scikit-learn module and TensorFlow module in Python.41 The hyperparameter optimization on the ST-ML models was based on the Bayesian optimization using HyperOpt package50,51 and UltraOpt package.52

* 1. **Model Performance Evaluation.** The coefficient of determination on the training set (*R*2tra) and root mean square error on the training set (*E*RMS-tra) were employed to evaluate model fitting performance.53,54 A 10-fold cross validation coefficient of determination (*R*2cv)was used as a criterion to select the best hyperparameter combinations.53,54 Prediction performance of the models was evaluated by coefficient of determination on the validation sets (*R*2val), and the external validation sets (*R*2ext); root mean square error on the validation sets (*E*RMS-val), and the external validation sets (*E*RMS*-*ext).
  2. **AD Characterization.** A state-of-the-art AD characterization method abbreviated as ADSAL{*ρ*s*,q* ≥ *ρ*s,T, *I*A,*q* ≤ *I*A,T} was adopted, where the subscript SAL is abbreviated from the term structure-activity landscapes defined as variations in activity/property with molecular similarity distributions of compounds;55,56 *ρ*s*,q* and *I*A,*q* stand for weighted molecular similarity density of a query compound represented by the subscript *q* in the chemical space of the training set, and weighted inconsistency in molecular activities, respectively; *ρ*s,Tand *I*A,T are the corresponding thresholds. *ρ*s and *I*A were found to be key metrics for characterizing dataset modelability.56 The ADSAL has been proven to outperform the others in previous studies.55

*ρ*s,*q* was calculated as:

(4)

where the subscripts *q* and *t* represent a query compound and a training compound; *T*c represents the training set of compounds. *ρ*s,*q* ranges from zero to infinity. *wq,t*  represents weighted similarity for the pair of compounds *q* and *t*.

As *wq,t* based on the exponential function performed better than that on the piecewise function,55 *wq,t* was calculated as:

(5)

where *S*M,*q*,*t* stands for pairwise molecular similarity (*S*M) between compounds *q* and *t*; *a* is a parameter that modulates the weighted contribution of the training compounds; *ε* is an infinitesimal quantity to ensure that the denominator is not zero.

*I*A,*q* was calculated as:

(6)

where *S*WD,*t* stands for weighted local discontinuity scores (*S*WD) for the training compound *t*. The range of *I*A values is [0,1]. Details on the AD characterization are provided in the SI.

1. **RESULTS AND DISCUSSION**
   1. **Performance of the MT-RGAN-H Model.** The performance metrics for the RGAN models and the optimal ST-ML models are listed in Table 2. Generally, the *R*2val (0.819 ~ 0.983) and *R*2ext (0.817 ~ 0.945) values for the ST-RGAN-H models, are larger than those for the ST-RGAN models (0.775 ~ 0.981, and 0.781 ~ 0.928, respectively). Therefore, the ST-RGAN-H models generally outperform the ST-RGAN models on the solute parameters.

**Table 2.** Comparison of the RGAN models with the optimal ST-ML models.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Architecture** | ***n*tra** | ***n*val** | ***n*ext** | ***R*2tra** | ***E*RMS-tra** | ***R*2cv** | ***R*2val** | ***E*RMS-val** | ***R*2ext** | ***E*RMS-ext** |
| *E* | ST-RGAN | 2726 | 681 | 4534 | 0.986 | 0.100 | 0.972 | 0.975 | 0.130 | 0.805 | 0.373 |
| ST-RGAN-H | 0.989 | 0.091 | 0.976 | **0.978** | **0.081** | **0.886** | **0.286** |
| MT-RGAN | 0.985 | 0.100 | 0.974 | 0.973 | 0.083 | 0.873 | 0.306 |
| MT-RGAN-H | 0.983 | 0.109 | 0.977 | 0.974 | 0.132 | 0.882 | 0.289 |
| GBR+CGD | 0.999 | 0.008 | 0.978 | 0.976 | 0.123 | 0.827 | 0.324 |
| *S* | ST-RGAN | 2656 | 633 | 4161 | 0.949 | 0.150 | 0.850 | 0.917 | 0.188 | 0.836 | 0.354 |
| ST-RGAN-H | 0.958 | 0.136 | 0.858 | **0.923** | **0.183** | 0.839 | 0.345 |
| MT-RGAN | 0.959 | 0.142 | 0.848 | 0.914 | 0.188 | 0.831 | 0.371 |
| MT-RGAN-H | 0.943 | 0.169 | 0.859 | 0.914 | 0.187 | **0.855** | **0.342** |
| GBR+CGD | 0.999 | 0.025 | 0.903 | 0.908 | 0.192 | 0.824 | 0.383 |
| *A* | ST-RGAN | 2470 | 604 | 4814 | 0.943 | 0.064 | 0.905 | 0.775 | 0.137 | 0.781 | 0.187 |
| ST-RGAN-H | 0.938 | 0.068 | 0.910 | 0.819 | 0.122 | 0.817 | 0.173 |
| MT-RGAN | 0.952 | 0.058 | 0.908 | 0.785 | 0.137 | 0.784 | 0.184 |
| MT-RGAN-H | 0.922 | 0.075 | 0.921 | **0.843** | 0.113 | **0.828** | **0.167** |
| GBR+Mordred | 0.998 | 0.013 | 0.899 | 0.810 | **0.111** | 0.780 | 0.189 |
| *B* | ST-RGAN | 2163 | 540 | 4419 | 0.972 | 0.097 | 0.922 | 0.956 | 0.120 | 0.928 | 0.178 |
| ST-RGAN-H | 0.973 | 0.090 | 0.932 | 0.957 | 0.117 | 0.937 | 0.186 |
| MT-RGAN | 0.975 | 0.087 | 0.925 | 0.954 | 0.121 | 0.933 | 0.168 |
| MT-RGAN-H | 0.965 | 0.104 | 0.938 | **0.959** | **0.115** | **0.944** | **0.154** |
| SVR+CGD | 0.968 | 0.094 | 0.944 | 0.944 | 0.125 | 0.920 | 0.182 |
| *L* | ST-RGAN | 2647 | 662 | 3611 | 0.991 | 0.288 | 0.977 | 0.981 | 0.406 | 0.903 | 1.334 |
| ST-RGAN-H | 0.985 | 0.368 | 0.980 | **0.983** | **0.394** | 0.945 | 1.081 |
| MT-RGAN | 0.991 | 0.275 | 0.977 | 0.979 | 0.415 | 0.897 | 1.464 |
| MT-RGAN-H | 0.982 | 0.392 | 0.978 | 0.980 | 0.404 | **0.947** | **1.054** |
| SVR+CGD | 0.991 | 0.276 | 0.977 | 0.982 | 0.396 | 0.920 | 1.281 |
| *n*tra, *n*val, and *n*ext: number of the compounds in the training, validation, and external validation sets, respectively; *R*2tra, *R*2val and *R*2ext: coefficient of determination for the training, validation, and external validation sets, respectively; *R*2CV: 10-fold cross validation coefficient of determination; *E*RMS-tra, *E*RMS-val and *E*RMS-ext:root mean square error for the training, validation and external validation sets, respectively. The optimal ST-ML models are expressed as “algorithm+ descriptor/fingerprint”, e.g., SVR+CGD. | | | | | | | | | | | |

Specifically, for parameter *A*, the ST-RGAN-H model outperforms the ST-RGAN model by > 5.7% on *R*2val. A previous study also found that models based on SMILES with hydrogens outperformed the hydrogen-free models by more than 4.0% on *R*2.20 Hydrogen atoms were usually ignored in most reported molecular property prediction models.20 In the ST-RGAN-H model, the MGs with hydrogen atoms can facilitate determining the numbers of the chemical bonds attached to non-hydrogen atoms and can provide more information than the MGs without hydrogen atoms.57 The model results have indeed confirmed this, as the performance of the model with explicit hydrogen atoms is superior on both the validation set and external validation set.

Hydrogen bonding is crucial for biological systems.58 In molecular initiating events of toxicity pathways, the initial interactions between xenobiotics and biomacromolecules are mainly influenced by hydrogen bondings.59 Non-reactive toxicity of chemicals is also associated with their ability to form hydrogen bonds with biomacromolecules.60 Therefore, the MGs with explicit hydrogen atoms can also be applied in predicting the binding affinities and the non-reactive toxicities.

It can be seen that compared with the ST-RGAN models, the MT-RGAN model increased *R*2val on *A* (1.3% increases), and increased *R*2ext on *E* (8.4%), *A* (0.4%), and *B* (0.5%). The MT learning could leverage information from the multiple correlated solute parameters, making it easier to achieve better performance than the ST models. However, the *R*2ext values of the MT-RGAN model on the parameters *S* and *L* are slightly lower than those from the ST-RGAN models. The basic assumption of MT learning is that the tasks are associated with each other, and thus the shared information among different tasks can lead to better learning performance if all the tasks are learned jointly, compared with that of ST learning.61 If there is no or weak relevance among the tasks, the performance of the MT models may be weakened.61

Figure S2 shows that 225 labeled compounds exclusively belong to the training set of *L* (8.5% of the total). The exclusively labeled compounds in the training sets of *E*, *S*, *A*, and *B* account for 2.6%, 2.0%, 0.3%, and 0.3% of the total, respectively. Therefore, the overlap between the training set of *L* and the other parameters is the lowest, followed by the parameters *E* and *S*. The low overlaps may reduce the relevance among the solute parameters. As a result, the MT-RGAN model on *L*, *E*, and *S* did not outperform the ST-RGAN models greatly.

The MT-RGAN-H model (*R*2ext = 0.828 ~ 0.947) significantly outperformed the ST-RGAN-H models on *S* (0.2% increases)*,* *A* (1.3%), *B* (0.7%), and *L* (0.2%), and slightly underperformed the ST-RGAN-H models on *E* (*R*2ext with 0.5% decrease).

Previous studies indicated that an endpoint with a large training set tends to have low benefits from the knowledge-sharing strategy.62,63 Since the large training sets contain ample data to generate valuable hidden representations, the need for additional information from the sharing strategy is reduced.62,63 Additional information from the other endpoints can be redundant for the endpoint with a large training set. This redundancy can introduce "noise" that may even degrade the overall performance of the model on that endpoint. As *E* has the largest training set among all the solute parameters, it is not surprising that the MT-RGAN-H model performs slightly worse than the ST-RGAN-H model on *E*.

The MT model can be trained on multiple inter-related endpoints, even endpoints with limited labeled compounds.16 The feature makes it possible for the MT model to address the issue of data scarcity in the environmental science and engineering field. For instance, the MT-RGAN-H model can be trained for predicting multiple partition coefficients with significant differences in data volumes, such as the log*K*OW (with over 17,000 values)25,27 and log*K*OA (less than 1,000)25,30. Similarly, this model can also be used to simultaneously predict acute toxicity endpoints to multiple organisms. Undoubtedly, this will provide vital data support for chemical management.

In summary, the MT-RGAN-H architecture has potential applications in constructing “end-to-end” prediction models on multiple physicochemical properties, environmental behavior, and toxicological parameters of chemicals that are inter-related.

* 1. **Comparison with ST-ML Models.** Based on the different fingerprints, descriptors, and algorithms, 36 ST-ML models for each solute parameter were constructed for comparison with the MT-RGAN-H model. The performance metrics for the ST-ML models are listed in Table S2. It can be seen that the ST-ML models based on the GBR algorithm and CGD descriptors generally perform the best.

Among the ST-ML models, the models based on the GBR algorithm usually performed the best. Previous studies also found the GBR algorithm exhibited the optimal performance for constructing models on screening persistent, mobile, and toxic chemicals64 and toxicity prediction65. This can be attributed to the fact that the GBR is an ensemble learning method that combines the strengths of multiple individual models.48 Although the models based on the GBR algorithm are typically the best in the ST-ML models, their performance on the validation and external validation sets is inferior to that of the MT-RGAN-H model.

The fingerprints did not perform as well as the descriptors when using the same ML algorithm. The fingerprints are binary values (1 or 0) that provide a simplified representation of molecules, focusing on local constitutional information.42 Consequently, the fingerprints have limitations in conveying detailed information about molecular structures.66 In contrast, the molecular descriptors provide a more comprehensive representation of molecular structures. They encode an amount of information from the molecules, encompassing a wide range of features, including simple features like atom count and complex features like electrostatic potential.67

However, both the fingerprints and the descriptors may have limitations in representing molecules.67 The graph-based models can automatically generate and select "molecular features" without manual intervention, simplifying modeling and achieving high performance.66 The MT-RGAN-H model exhibits stronger predictive power (*R*2val = 0.845 ~ 0.983, *R*2ext = 0.828 ~ 0.947) than the optimal ST-ML models (*R*2val = 0.810 ~ 0.982, *R*2ext = 0.780 ~ 0.920). Wang et al.17 also found that MG-based models outperformed descriptor-based models for screening PBT chemicals.

For the external validation set with over 4800 chemicals from the Absolv Database9, the *R*2 values of the MT-RGAN-H model were improved by 2.4% to 6.7%, and the *E*RMS values were decreased by 11.9% to 21.5%, relative to the ST-ML model. This indicates that the generalization ability of the MT-RGAN-H model is superior to that of the ST-ML models.

The four optimal ST-ML models rely on CGD descriptors that were calculated by DFT optimization on the molecular structures. In a server node with 16 cores and 24GB memory, the DFT optimization of a single molecule takes 2 minutes to 10 hours. In contrast, training and evaluating the MT-RGAN-H model with the same server node on the training sets only took 67 minutes, much shorter than the time needed for calculating molecular descriptors for the ST-ML models. If the MT-RGAN-H model is trained and evaluated on a computer equipped with an RTX 3090 graphic process unit, it only takes 56 minutes. Notably, the MT-RGAN-H model can simultaneously predict the five solute parameters, while each ST-ML model can only predict one. Running the ST-ML models multiple times increases time and resource consumption. Thus, using the MT-RGAN-H model for predicting the solute parameters is more computationally efficient.

The ST-ML models not only involve an extended computation period but also demand professional knowledge in quantum chemistry.6 The end-to-end models that rely only on open source codes are easier to be integrated into websites or software than the DFT calculations that may rely on commercial software. Therefore, the MT-RGAN-H model was selected as the best one for further evaluation.

**3.3. Comparison with Previous models.** Table 3 lists the performance metrics of the MT-RGAN-H models that were re-trained on the same datasets used in previous studies.6,11,15 Xiao et al.6 developed the “DUT models” for predicting *S*, *A*, and *B*, based on multiple linear regression with CGD descriptors. Zhang et al.11 employed deep neural networks coupled with 2D descriptors to construct ST models called “PaDEL-DNN” on *E*, *S*, *A*,and *B*, based on a training set of 1978 compounds. Ulrich et al.15 employed GCN to construct ST models called "taut-ST" based on rigorously filtered datasets for the parameters *E*, *S*, *A*, *B*,and *L.*

**Table 3.** Comparison on the different models for estimating the solute parameters.

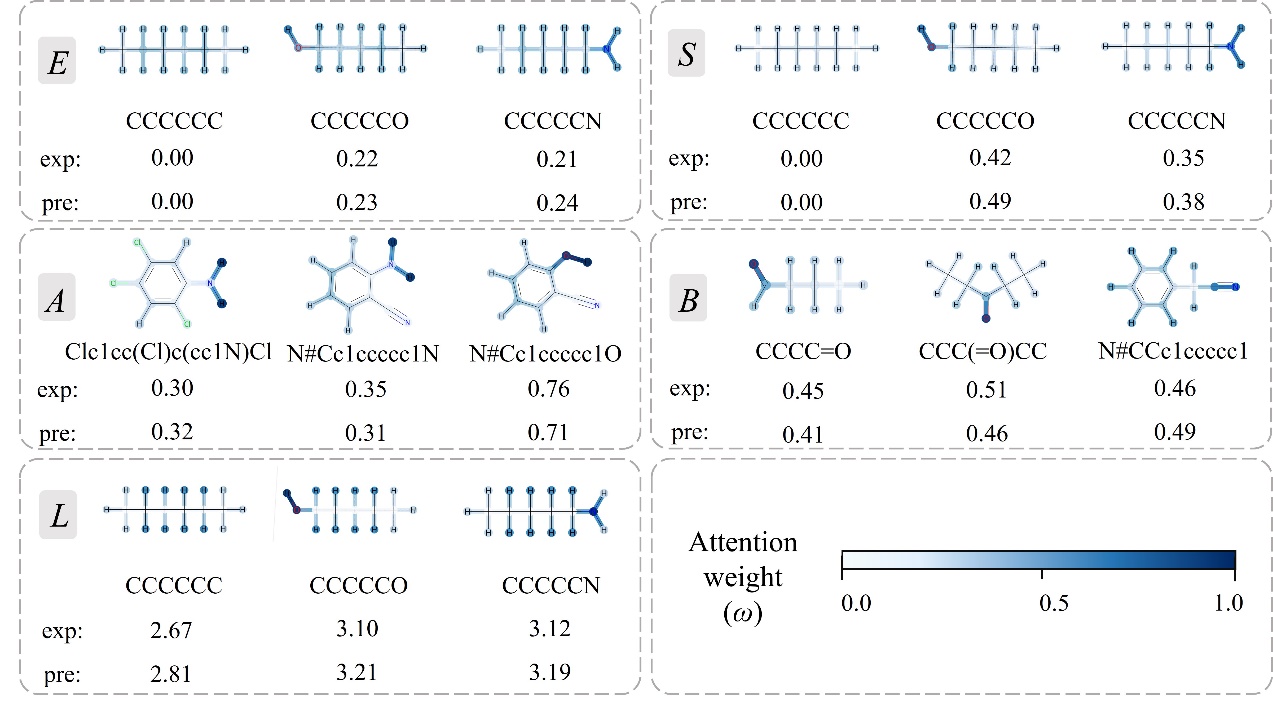
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Solute parameters** | **Models** | ***n*tra** | ***n*val** | ***R2*tra** | ***E*RMS*-*tra** | ***R*2val** | ***E*RMS*-*val** |
| *S* | MT-RGAN-H | 2603 | 645 | **0.94** | **0.17** | **0.92** | **0.18** |
| DUT6 | 0.90 | 0.20 | **0.92** | 0.19 |
| *A* | MT-RGAN-H | 2401 | 590 | **0.92** | **0.07** | **0.92** | **0.07** |
| DUT6 | **0.92** | **0.07** | 0.91 | 0.08 |
| *B* | MT-RGAN-H | 2127 | 531 | **0.97** | **0.10** | **0.96** | **0.10** |
| DUT6 | 0.93 | 0.13 | 0.94 | 0.14 |
| *E* | MT-RGAN-H | 1582 | 396 | **0.99** | **0.09** | 0.97 | 0.13 |
| PaDEL-DNN11 | 0.98 | 0.11 | **0.98** | **0.10** |
| *S* | MT-RGAN-H | 1582 | 396 | **0.95** | **0.13** | **0.87** | **0.19** |
| PaDEL-DNN11 | 0.88 | 0.21 | **0.87** | 0.20 |
| *A* | MT-RGAN-H | 1582 | 396 | **0.92** | **0.07** | **0.88** | **0.07** |
| PaDEL-DNN11 | 0.87 | 0.10 | 0.86 | 0.08 |
| *B* | MT-RGAN-H | 1582 | 396 | **0.96** | **0.08** | **0.95** | **0.10** |
| PaDEL-DNN11 | 0.95 | 0.11 | 0.93 | 0.11 |
| *E* | MT-RGAN-H | 4568 | 635 | **0.99** | **0.10** | **0.99** | **0.10** |
| taut-ST15 | **0.99** | **0.10** | 0.97 | 0.12 |
| *S* | MT-RGAN-H | 4582 | 636 | 0.95 | 0.15 | **0.93** | **0.19** |
| taut-ST15 | **0.96** | **0.11** | 0.80 | 0.22 |
| *A* | MT-RGAN-H | 4580 | 635 | **0.93** | **0.08** | **0.93** | **0.09** |
| taut-ST15 | 0.92 | 0.10 | 0.88 | 0.11 |
| *B* | MT-RGAN-H | 4396 | 607 | **0.97** | 0.14 | **0.97** | **0.10** |
| taut-ST15 | **0.97** | **0.13** | 0.93 | 0.14 |
| *L* | MT-RGAN-H | 4011 | 549 | **0.99** | **0.29** | **0.99** | **0.33** |
| taut-ST15 | 0.99 | 0.30 | 0.98 | 0.42 |
| *n*tra and *n*val: number of the compounds in the training and validation sets, respectively; *R*2tra and *R*2val: coefficient of determination for the training and validation sets, respectively; *E*RMS-tra and *E*RMS-val:root mean square error for the training and validation sets, respectively. References (6, 11, 15) for the previous models are marked. | | | | | | | |

It can be seen from Table 3 that the re-trained MT-RGAN-H models performed better (*R*2val: 0.87 ~ 0.99) than the previous models (*R*2val: 0.77 ~ 0.98) on the same datasets. It deserves mentioning that the MT-RGAN-H model achieves good prediction accuracy for the parameters that are highly sensitive to hydrogen atoms, including the *A* (*R*2val > 0.92) and *B* (*R*2val > 0.95), which was not achievable in the previous models.6,11,15 This further proves that the combination of MT learning and MGs with explicit hydrogens can improve model prediction performance.

It is worth mentioning that Ulrich et al.15 also developed an MT model called “taut-MT” with a dataset containing the five solute parameters (*E*, *S*, *A*, *B*,and *L*).15 The taut-MT model was based on a parameter-sharing mechanism that shares the same hidden layers of the network among the five solute parameters.15 This type of network structure requires the training set to have no missing label values.15 The MT-RGAN-H model overcomes the issue of incomplete label values in the dataset by constructing an additional mark for each instance to indicate the positions of the missing labels,16 as can be seen from Figure S3. This allows the MT-RGAN-H model to skip the missing labels during training and evaluation.16

In the training set of the MT-RGAN-H model, only 704 compounds have all five solute parameter values, while 3769 compounds have missing values for at least one solute parameter (see Figure S2). If the MT-RGAN-H model is trained only on the 704 compounds, it clearly has limited coverage of the chemical space and may limit the AD of the model. However, obtaining complete experimental data for all the solute parameters is challenging. Therefore, the MT-RGAN-H model maximized the utilization of the training sets and can effectively address the challenges posed by the data limitations.

**3.4. Model interpretation.** The MT-RGAN-H architecture assigns varying attention weights (*ω*) to different substructures within the molecules. Figure 2 showcases a subset of molecular structures along with their corresponding *ω*. Similar figures for the remaining molecules are presented at SI.



**Figure 2.** Selected molecular structures along with their corresponding attention weights (*ω*) for the five solute parameters, where “exp” stands for the experimental value, and “pre” stands for the predicted values.

For *E*, Figure 2 showcases a comparison among hexane, 1-pentanol, and 1-pentylamine. The MT-RGAN-H model assigns high *ω* to the atoms that increase molecular polarizability, such as O and N atoms. *E* characterizes intermolecular dispersion interactions, which increase with the polarizability.2 Therefore, focusing on the atoms that increase molecular polarizability is reasonable. For *S*, the MT-RGAN-H model assigned high *ω* to polar functional groups such as -OH and -NH2, which is reasonable since *S* describes interactions related to the surface polarity of the molecules.1

For *A*, fragments such as -NH2 and -OH typically have high *ω* values, since the groups are typically hydrogen-bond donors.2 For *B*, fragments such as -C≡N and -C=O typically have high *ω*, since the groups are typical hydrogen-bond acceptors.2

For *L*, the MT-RGAN-H model assigns high *ω* on atoms or groups that increase molecular polarizability since *L* describes cavity formation and dispersion interactions.2 Unlike *E* for which the *ω* focuses only on specific functional groups, *ω* of *L* pays attention to the overall structure of the molecules. Therefore, the MT-RGAN-H model can identify customized features for the different solute parameters and exhibit human-understandable and interpretable hints.

**3.5. AD Characterization.** As can be seen from Figure S4, increasing *ρ*s,T or decreasing *I*A,T leads to a decrease in the number of compounds in the validation sets within the AD (*N*val) and an increase in the *R*2val values. Therefore, setting more stringent thresholds for the ADSAL resulted in fewer compounds kept within the AD, coupled with improved model performance. This represents the inherent nature of a robust AD, highlighting the effectiveness of the ADSAL methodology on the MT-RGAN-H model.

Various *ρ*s,T and *I*A,T thresholds were defined to evaluate their effects on the *N*val and *R*2val values. Optimal ADSAL was proposed to enhance the *R*2val values significantly and maximize the *N*val values jointly. The *ρ*s,T values of the optimal ADSAL on the five solute parameters are 2.5, 1.5, 3.5, 2.5, and 3.5, respectively; and the corresponding *I*A,T values are 0.80, 0.50, 0.70, 0.70, and 0.90. With the optimal ADSAL, the *R*2val values for *E*, *S*, *A*, *B*,and *L* were improved up to 0.981, 0.920, 0.896, 0.968, and 0.980, respectively; with the corresponding *N*val values being 542, 576, 410, 405, and 656. By applying the optimal ADSAL to the external validation sets, the *R*2ext values were also improved (the maximum improvement is 2.2%). Therefore, the MT-RGAN-H model coupled with the optimal ADSAL can be utilized to predict the solute parameters for various compounds.

**3.6. Model application.** A total of six physicochemical parameters (log*P*LV, log*S*W, log*K*H, log*K*OA log*K*OW, and log*K*OC) were selected to further evaluate the accuracy of the solute parameters predicted by the MT-RGAN-H model. Table 4 lists the performance metric of the six parameters predicted by the different LSER equations, where the solute parameters were predicted by either the MT-RGAN-H model or the ST-taut models.

**Table 4.** Comparison of the LSER models constructed with the different sets of solute parameter values.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Indicators** | **log*P*LV** | **log*S*W** | **log*K*H** | **log*K*H** | **log*K*OA** | **log*K*OA** | **log*K*OW** | **log*K*OW** | **log*K*OC** |
| MT-RGAN-H | *r* | 0.725 | **0.781** | **0.843** | **0.857** | **0.971** | **0.971** | **0.829** | **0.817** | **0.860** |
| *E*RMS | 3.659 | **2.174** | **2.071** | **1.677** | **0.873** | **0.852** | **1.369** | **1.302** | 0.715 |
| ST-taut | *r* | **0.732** | 0.746 | 0.804 | 0.831 | 0.953 | 0.952 | 0.723 | 0.736 | 0.850 |
| *E*RMS | **3.610** | 2.263 | 2.098 | 2.407 | 1.268 | 1.215 | 1.862 | 1.508 | **0.666** |
| *r*: Pearson correlation coefficient between the predicted and experimental values; *E*RMS: root mean square error. | | | | | | | | | | |

As can be seen from Table 4, the LSER models with the solute parameters predicted by the MT-RGAN-H model generally have higher Pearson correlation coefficients (*r*)between the predicted and experimental values,and lower *E*RMS than the LSERs with the solute parameters predicted by the ST-taut models. Given the typically high prediction performance of the models on *E* (*R*2val > 0.95, *E*RMS < 0.15) and *L* (*R*2val > 0.95, *E*RM*S* < 0.50),11,15 accurate prediction of *S*, *A*, and *B* becomes crucial for improving the prediction of the LSERs. Notably, the MT-RGAN-H model improved the prediction accuracy on *S* (16.3% increase), *A* (5.6%), and *B* (4.3%); over the ST-taut models. It was the improvement that promoted the prediction performance of the LSERs based on the MT-RGAN-H derived solute parameters. Therefore, the MT-RGAN-H derived solute parameters have the potential to be employed for constructing LSER models on other physicochemical, environmental partition behavior parameters, as well as non-reactive toxicity of chemicals.

**4. Conclusions**

This study developed an MT-RGAN-H model to predict the five solute parameters, which overcomes limitations in data scarcity and exhibits good prediction performance. Compared with previous models, the MT-RGAN-H model performed prominently in predicting parameters susceptible to hydrogen atoms, including H-bond acidity (5.6% increases) and H-bond basicity (4.3% increases). Therefore, the RGAN-H strategy holds great potential for accurately predicting molecular properties (such as the acid dissociation constant p*K*a) or activities related to hydrogen atoms and Q-H bonds. Meanwhile, the solute parameters predicted by the MT-RGAN-H model were further used to predict six physicochemical parameters of more than 17,000 chemicals based on the LSERs, achieving better or comparable prediction accuracy compared to previous optimal models. Therefore, the MT-RGAN-H can be further applied for supporting sound management of chemicals.

**◼ ASSOCIATED CONTENT**

**Supporting Information**

The Supporting Information is available free of charge at \*\*\*\*\*\*\*\*\*.

A Microsoft Word file (.docx) and a Microsoft Excel file (.xlsx) which give details on the supplementary figures and tables mentioned in the text.

Source codes of the MT-RGAN-H model, curated data on the solute parameters, Gaussian and Dragon descriptors, and molecular structures along with their corresponding attention weights (*ω*) for the five solute parameters were uploaded to the GitHub (https://github.com/Azukix/MT-RGAN-H/).

**◼ AUTHOR INFORMATION**

**Corresponding Author**

**Jingwen Chen -** *Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), Dalian Key Laboratory on Chemicals Risk Control and Pollution Prevention Technology, School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China;*

http://orcid.org/0000-0002-5756-3336; E-mail: [jwchen@dlut.edu.cn](mailto:jwchen@dlut.edu.cn)

**Authors**

**Zijun Xiao** － *Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), Dalian Key Laboratory on Chemicals Risk Control and Pollution Prevention Technology, School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China*;Email: xzj960729@mail.dlut.edu.cn

**Minghua Zhu** － *Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), Dalian Key Laboratory on Chemicals Risk Control and Pollution Prevention Technology, School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China*; *Key Laboratory of Integrated Regulation and Resources Development of Shallow Lakes of Ministry of Education, College of Environment, Hohai University, Nanjing 210098, China;* orcid.org/0009-0007-3724-5182; Email: mhzhu@hhu.edu.cn

Notes

The authors declare no competing financial interest.

**◼ ACKNOWLEDGMENTS**

This study was supported by the National Natural Science Foundation of China (22136001, 22206022) and the National Key R&D Program of China (2022YFC3902100).

**◼ REFERENCES**

(1) Endo, S.; Goss, K. Applications of polyparameter linear free energy relationships in environmental chemistry. *Environ. Sci. Technol.*, **2014**, *48*(21), 12477-12491.

(2) Abraham, M. H.; Ibrahim, A.; Zissimos, A. M. Determination of sets of solute descriptors from chromatographic measurements. *J. Chromatogr. A*, **2004**, *1037*(1-2), 29-47.

(3) Ulrich, N.; Endo, S.; Brown, T. N.; Watanabe, N.; Bronner, G.; Abraham, M. H.; Goss, K. U. LSER database ver. 3.2.1, **2017**. http://www.ufz.de/lserd. (accessed 2023.10.20).

(4) Allen, D. T.; Gathergood, N.; Licence, P.; Subramaniam, B. Expectations for manuscripts contributing to the field of solvents in *ACS Sustainable Chemistry & Engineering*. *ACS Sustain. Chem. Eng.*, **2020**, *8*(39), 14627-14629.

(5) Wang, Z.; Walker, G. W.; Muir, D. C. G.; Nagatani-Yoshida, K. Toward a global understanding of chemical pollution: a first comprehensive analysis of national and regional chemical inventories. *Environ. Sci. Technol.*, **2020**, *54*(5), 2575-2584.

(6) Xiao, Z. J.; Chen, J. W.; Wang, Y.; Wang, Z. Y. In silico package models for deriving values of solute parameters in linear solvation energy relationships. *SAR QSAR Environ. Res.*, **2023**, *34*(1), 21–37.

(7) Liang, Y.; Xiong, R.; Sandler, S. I.; Di Toro, D. M. Quantum chemically estimated Abraham solute parameters using multiple solvent-water partition coefficients and molecular polarizability. *Environ. Sci. Technol.*, **2017**, *51*(17), 9887-9898.

(8) Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. Estimation of molecular linear free energy relation descriptors using a group contribution approach. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*(5), 835-845.

(9) ACD/Labs. ACD/Percepta, **2012**.

(10) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. QSAR without borders. *Chem. Soc. Rev.*, **2020**, *49*(11), 3525-3564.

(11) Zhang, K.; Zhang, H. Predicting solute descriptors for organic chemicals by a deep neural network (DNN) using basic chemical structures and a surrogate metric. *Environ. Sci. Technol.*, **2022**, *56*(3), 2054-2064.

(12) Chung, Y.; Vermeire, F. H.; Wu, H.; Walker, P. J.; Abraham, M. H.; Green, W. H. Group contribution and machine learning approaches to predict Abraham solute parameters, solvation free energy, and solvation enthalpy. *J. Chem. Inf. Model.*, **2022**, *62*(3), 433-446.

(13) Sun, M.; Zhao, S.; Gilvary, C.; Elemento, O.; Zhou, J.; Wang, F. Graph convolutional networks for computational drug development and discovery. *Brief. Bioinform.*, **2020**, *21*(3), 919-935.

(14) Zhou, Y.; Wang, Z.; Huang, Z.; Li, W.; Chen, Y.; Yu, X.; Tang, Y.; Liu, G. In silico prediction of ocular toxicity of compounds using explainable machine learning and deep learning approaches. *J. Appl. Toxicol.*, **2024**, *44*, 892-907.

(15) Ulrich, N.; Ebert, A. Can deep learning algorithms enhance the prediction of solute descriptors for linear solvation energy relationship approaches? *Fluid Phase Equilib.*, **2022**, *555*, 113349.

(16) Wu, Z.; Jiang, D.; Wang, J.; Hsieh, C.; Cao, D.; Hou, T. Mining toxicity information from large amounts of toxicity data. *J. Med. Chem.*, **2021**, *64*(10), 6924–6936.

(17) Wang, H.; Wang, Z.; Chen, J.; Liu, W. Graph attention network model with defined applicability domains for screening PBT chemicals. *Environ. Sci. Technol.*, **2022**, *56*(10), 6774-6785.

(18) Civis, S.; Lamanec, M.; Špirko, V.; Kubista, J.; Špet Ko, M.; Hobza, P. Hydrogen bonding with hydridic hydrogen-experimental low-temperature IR and computational study: is a revised definition of hydrogen bonding appropriate? *J. Am. Chem. Soc.*, **2023**, *145*(15), 8550-8559.

(19) van der Lubbe, S. C. C.; Fonseca Guerra, C. The nature of hydrogen bonds: a delineation of the role of different energy components on hydrogen bond strengths and lengths. *Chem.-Asian J.*, **2019**, *14*, 2760-2769.

(20) Zhang, X.; Wu, C.; Yang, Z.; Wu, Z.; Yi, J.; Hsieh, C.; Hou, T.; Cao, D. MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction. *Brief. Bioinform.*, **2021**, *22*(6), 1-14.

(21) Munikoti, S.; Agarwal, D.; Das, L.; etc. Challenges and opportunities in deep reinforcement learning with graph neural networks: a comprehensive review of algorithms and applications. 8 Nov 2022.https://arxiv.org/abs/2206.07922. (accessed 2023-08-23)

(22) Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing*, **2021**, *452*, 48-62.

(23) Zhang, Y.; Yang, Q. An overview of multi-task learning. *Natl. Sci. Rev.*, **2018**, *5*(1), 30-43.

(24) Python, W. Python. *Python Releases for Windows*, **2021**, *24*.

(25) Card, M. L.; Gomez-Alvarez, V.; Lee, W. H.; Lynch, D. G.; Orentas, N. S.; Lee, M. T.; Wong, E. M.; Boethling, R. S. History of EPI Suite and future perspectives on chemical property estimation in us toxic substances control act new chemical risk assessments. *Environ. Sci.-Process. Impacts*, **2017**, *19*(3), 203-212.

(26) Diderich, R. Tools for category formation and read-across: overview of the OECD (Q)SAR application toolbox. *In Silico Toxicology: Principles and Applications*, **2010**, 390-412.

(27) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA models for predicting physicochemical properties and environmental fate endpoints. *J. Cheminformatics*, **2018**, *10*(10), 1-19.

(28) Dimitrov, S. D.; Diderich, R.; Sobanski, T.; Pavlov, T. S.; Chankov, G. V.; Chapkanov, A. S.; Karakolev, Y. H.; Temelkov, S. G.; Vasilev, R. A.; Gerova, K. D. QSAR Toolbox-workflow and major functionalities. *SAR QSAR Environ. Res.*, **2016**, *27*(3), 203-219.

(29) Fu, Z.; Chen, J.; Li, X.; Wang, Y.; Yu, H. Comparison of prediction methods for octanol-air partition coefficients of diverse organic compounds. *Chemosphere*, **2016**, *148*, 118-125.

(30) Baskaran, S.; Lei, Y. D.; Wania, F. A database of experimentally derived and estimated octanol-air partition ratios (*K*OA). *J. Phys. Chem. Ref. Data*, **2021**, *50*(4), 043101.

(31) Ketkar, N.; Moolayil, J.; Ketkar, N.; Moolayil, J. Introduction to pytorch. *Deep Learning with Python: Learn Best Practices of Deep Learning Models with Pytorch*, **2021**, 27-91.

(32) Wang, M. Y. In *Deep graph library: towards efficient and scalable deep learning on graphs*, ICLR workshop on representation learning on graphs and manifolds, **2019**.

(33) Landrum, G. RDKit: a software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, **2013**, *8*.

(34) Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; van den Berg, R.; Titov, I.; Welling, M. In *Modeling relational data with graph convolutional networks*, Cham, **2018**; Gangemi, A.; Navigli, R.; Vidal, M.; Hitzler, P.; Troncy, R.; Hollink, L.; Tordai, A.; Alam, M., Ed. Springer International Publishing: Cham; pp 593-607.

(35) Kingma, D. P.; Ba, J. Adam: a method for stochastic optimization. 30 Jan 2017. https://arxiv.org/abs/1412.6980. (accessed 2023-08-23)

(36) Moriwaki, H.; Tian, Y.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminformatics*, **2018**, *10*(4), 1–14.

(37) Frisch, M.; Trucks, G.; Schlegel, H. B.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V. Gaussian 09, **2013**.

(38) Srl, T. Dragon (software for molecular descriptor calculation) ver. 6.0., **2010**.

(39) Wadt, W. R.; Hay, P. J. Ab initio effective core potentials for molecular calculations. Potentials for main group elements Na to Bi. *The Journal of Chemical Physics*, **1985**, *82*(1), 284-298.

(40) Petersson, G. A.; Al Laham, M. A. A complete basis set model chemistry. II. Open‐shell systems and the total energies of the first‐row atoms. *The Journal of Chemical Physics*, **1991**, *94*(9), 6081-6090.

(41) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: machine learning in python. *The Journal of Machine Learning Research*, **2011**, *12*, 2825-2830.

(42) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, **2010**, *50*(5), 742–754.

(43) Glodek, A.; Horrigan, S.; Wachtel, K.; Katz, S.; Glanowski, S.; Soppet, D.; Liu, D.; Hise, D.; Castaneda, J.; Ebner, R. Avalon compound profiling toolkit: use of gene transcription biomarker signatures for drug discovery. *Cancer Res.*, **2008**, *68*(9), 796.

(44) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of mdl keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, **2002**, *42*(6), 1273-1280.

(45) Breiman, L. Random forests. *Mach. Learn.*, **2001**, *45*, 5-32.

(46) Yan, X.; Sedykh, A.; Wang, W.; Zhao, X.; Yan, B.; Zhu, H. In silico profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale*, **2019**, *11*(17), 8352-8362.

(47) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.*, **1995**, *20*, 273-297.

(48) Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, **2001**, 1189-1232.

(49) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature*, **1986**, *323*(6088), 533-536.

(50) Bergstra, J.; Komer, B.; Eliasmith, C.; Yamins, D.; Cox, D. D. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, **2015**, *8*(1), 14008.

(51) Bergstra, J.; Yamins, D.; Cox, D. D. Hyperopt: distributed asynchronous hyper-parameter optimization. *Astrophysics Source Code Library*, **2022**, 2205-2208.

(52) TQCAI; Auto-flow. Auto-flow/ultraopt: v0.1.0. In Zenodo: 2021.

(53) Zhao, C. Y.; Zhang, R. S.; Zhang, H. X.; Xue, C. X.; Liu, H. X.; Liu, M. C.; Hu, Z. D.; Fan, B. T. QSAR study of natural, synthetic and environmental endocrine disrupting compounds for binding to the androgen receptor. *SAR QSAR Environ. Res.*, **2005**, *16*(4), 349-367.

(54) Roy, K.; Ambure, P.; Kar, S.; Ojha, P. K. Is it possible to improve the quality of predictions from an "intelligent" use of multiple QSAR/QSPR/QSTR models? *J. Chemom.*, **2018**, *32*(1), e2992.

(55) Liu, W.; Wang, Z.; Chen, J.; Tang, W.; Wang, H. Machine learning model for screening thyroid stimulating hormone receptor agonists based on updated datasets and improved applicability domain metrics. *Chem. Res. Toxicol.*, **2023**, *36*(6), 947-958.

(56) Wang, H.; Liu, W.; Chen, J.; Wang, Z. Applicability domains based on molecular graph contrastive learning enable graph attention network models to accurately predict 15 environmental end points. *Environ. Sci. Technol.*, **2023**, *57*(44), 16906-16917.

(57) Kumagai, T.; Shiotari, A.; Okuyama, H.; Hatta, S.; Aruga, T.; Hamada, I.; Frederiksen, T.; Ueba, H. H-atom relay reactions in real space. *Nat. Mater.*, **2012**, *11*(2), 167-172.

(58) Tan, H.; Wang, X.; Hong, H.; Benfenati, E.; Giesy, J. P.; Gini, G. C.; Kusko, R.; Zhang, X.; Yu, H.; Shi, W. Structures of endocrine-disrupting chemicals determine binding to and activation of the estrogen receptor α and androgen receptor. *Environ. Sci. Technol.*, **2020**, *54*(18), 11424-11433.

(59) Madden, J. C.; Enoch, S. J.; Paini, A.; Cronin, M. T. A review of *in silico* tools as alternatives to animal testing: principles, resources and applications. *Alternatives to Laboratory Animals*, **2020**, *48*(4), 146-172.

(60) Wang, S.; Yan, L. C.; Zheng, S. S.; Li, T. T.; Fan, L. Y.; Huang, T.; Li, C.; Zhao, Y. H. Toxicity of some prevalent organic chemicals to tadpoles and comparison with toxicity to fish based on mode of toxic action. *Ecotoxicol. Environ. Saf.*, **2019**, *167*, 138-145.

(61) Zhang, W.; Deng, L.; Wu, D. Overcoming negative transfer: a survey. 9 Aug 2021. https://arxiv.org/abs/2009.00909v4. (accessed 2023-08-23)

(62) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-task neural networks for QSAR predictions. 4 Jun 2014. https://arxiv.org/abs/1406.1231. (accessed 2023-08-23)

(63) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep neural nets as a method for quantitative structure–activity relationships. *J. Chem. Inf. Model.*, **2015**, *55*(2), 263-274.

(64) Zhao, Q.; Yu, Y.; Gao, Y.; Shen, L.; Cui, S.; Gou, Y.; Zhang, C.; Zhuang, S.; Jiang, G. Machine learning-based models with high accuracy and broad applicability domains for screening PMT/vPvM substances. *Environ. Sci. Technol.*, **2022**, *56*(24), 17880-17889.

(65) Jiang, J.; Wang, R.; Wei, G. GGL-TOX: geometric graph learning for toxicity prediction. *J. Chem. Inf. Model.*, **2021**, *61*(4), 1691-1700.

(66) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided. Mol. Des.*, **2016**, *30*(8), 595-608.

(67) Jeong, J.; Choi, J. Artificial intelligence-based toxicity prediction of environmental chemicals: future directions for chemical management applications. *Environ. Sci. Technol.*, **2022**, *56*(12), 7532-7543.