

Supplementary Material

Multi-Task Graph Convolutional Network Model with Improved Performance and Broad Applicability Domains for Identifying Aquatic Toxic Chemicals

Shengshe Ji, Wenjia Liu, Guijia Bi, Haobo Wang * and 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

* Correspondence: wanghaobodut@mail.dlut.edu.cn (H.W.); jwchen@dlut.edu.cn (J.C.)

S1. Curation of Aquatic Toxicity Data

Data Filtering and Taxonomic Verification

Data filtering was conducted based on standard exposure durations derived from Globally Harmonized System of Classification and Labelling of Chemicals (GHS) [9]. Fish acute toxicity records were restricted to 96-h measurements. Data for crustacean acute toxicity were limited to exposure periods within 48 or 24 h. Algal acute toxicity records were filtered to retain 72 or 96 h. Records with non-compliant exposure times were excluded.

Similarly, data filtering for chronic toxicity was conducted based on standard exposure durations. Fish chronic toxicity records were restricted to measurements of 28 or 32 days. Data for crustacean chronic toxicity were limited to exposure periods of 21 days. Algal chronic toxicity records were filtered to retain 72 h. Records with non-compliant exposure times were excluded.

To ensure that the compiled data are accurately mapped to valid biological entities, the species names of fish, crustaceans, and algae were initially retrieved and standardized using the National Center for Biotechnology Information (NCBI) Taxonomy database (<https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi> accessed on 23 November 2025). For species not identified in the NCBI database, domain-specific authoritative databases were subsequently employed for verification and correction. Specifically, fish taxonomic names were corrected based on FishBase (<https://fishbase.de/> accessed on 27 November 2025), and records with non-fish species were removed. Crustacean names were verified against the Global Biodiversity Information Facility (GBIF) (<https://www.gbif.org/> accessed on 31 November 2025), and records with non-crustacean taxa (e.g., aquatic insects such as *Baetis* sp.) were excluded. Algal nomenclature was standardized

using AlgaeBase (<https://www.algaebase.org/> accessed on 31 November 2025), with non-algal organisms discarded. Finally, when calculating the total number of species in the compiled datasets, unidentified species were included and treated as distinct species. For instance, records such as *Selenastrum* sp. were considered as independent species.

S2. Detailed Information on Dataset Processing

Before data discretization, all available experimental records for a given compound were retained rather than selecting a single representative value. Each record was first converted into a toxic or non-toxic label using the GHS-based thresholds described above. The final chemical label was then assigned according to the following rules. First, if all records for a compound yielded the same label, that label was directly assigned. Second, if inconsistent labels were obtained, a conservative consensus strategy was adopted. Under this rule, a compound was categorized as toxic if any single record indicated toxicity and as non-toxic only when all available records indicated non-toxicity. This conservative approach aligns with regulatory-oriented hazard screening, which prioritizes the avoidance of false negatives for potentially hazardous chemicals.



S3. Detailed Description of the Features Used in GCN

Node Features	Size	Description
atom_type	16	['B', 'C', 'N', 'O', 'F', 'Si', 'P', 'S', 'Cl', 'As', 'Se', 'Br', 'Te', 'I', 'At', 'others']
degree	7	Number of covalent bonds [0, 1, 2, 3, 4, 5, others] (one-hot)
formal_charge	1	Electrical charge (integer)
radical electrons	1	Radical electrons 1 number of radical electrons (integer)
hybridization	6	[sp, sp ² , sp ³ , sp ³ d, sp ³ d ² , other] (one-hot)
aromaticity	1	Whether the atom is part of an aromatic system [0/1] (one-hot)
Node Features	Size	Description
hydrogens	5	Number of connected hydrogens [0, 1, 2, 3, 4] (one-hot)
chirality	1	Whether the atom is chiral center [0/1] (one-hot)
chirality type	2	[R, S] (one-hot)
Edge Features	Size	Description
bond_type	4	[single bond, double bond, triple bond, aromatic bond] (one-hot)
conjugation	1	Whether the bond is conjugated [0/1] (one-hot)
ring	1	Whether the bond is in ring [0/1] (one-hot)
stereo	4	[StereoNone, StereoAny, StereoZ, StereoE] (one-hot)
atom pair	13	['CC'], ['CN'], ['NC'], ['ON'], ['NO'], ['CO'], ['OC'], ['CS'], ['SC'], ['SO'], ['OS'], ['NN'], ['SN'], ['NS'], ['CCl'], ['ClC'], ['CF'], ['FC'], ['CI'], ['IC'], ['CBr'], ['BrC'], ['others']

S4. Implementation Details of Single-Task Learning GCN Models

To ensure a fair comparison with the proposed multi-task graph convolutional network (MTL-GCN), the single-task learning GCN (STL-GCN) models were constructed using the same GCN algorithm but were trained independently for each of the six toxicity endpoints. Each STL model consists of a GCN-based feature generator for molecular representation learning and a subsequent multi-layer perceptron acting as the prediction head. To eliminate confounding factors, consistent data splitting strategies and training protocols were maintained, including the choice of optimizer, loss function, and early stopping mechanisms. Detailed configuration regarding the GCN and specific training hyperparameters are summarized in Table S10.

S5. Details on AD_{SAL} Characterization about Calculation

The applicability characterization (AD) method based on a structure-activity landscape (SAL) analysis, short for AD_{SAL} [64], was applied to characterize applicability domain (AD) in this study. In AD_{SAL}, ρ_s represents weighted molecular similarity density, and I_A describes weighted inconsistency of molecular activities. The method is based on the assumption that molecules with similar structures but conflicting activities can give rise to "activity cliffs" (ACs), resulting in local discontinuity in

structure-activity landscapes (SALs) [64,81]. The $w_{q,t}$ was calculated using an exponential function:

$$w_{q,t} = \exp \left\{ \frac{-\alpha \times (1 - S_{q,t})}{S_{q,t} + \varepsilon} \right\} \quad (S1)$$

where $S_{q,t}$ represents pairwise molecular similarity between compounds q and t ; α , a parameter, adjusts weighted contribution of the training compounds; and ε is a value used to prevent division by zero in the denominator. The $w_{q,t}$ and $S_{WD,t}$ were further used to calculate ρ_s and I_A .

The molecules located on the ACs can be identified by using weighted local discontinuity scores (S_{WD}). The S_{WD} for the training compound t ($S_{WD,t}$) was calculated as:

$$S_{WD,t} = \frac{\sum_{\{v \in T, t \neq v\}} w_{t,v} \cdot S_{t,v} \cdot |y_t - y_v|}{\sum_{\{v \in T, t \neq v\}} w_{t,v}} \quad (S2)$$

where the subscript t and v denote two training compounds, T represents the complete training set; $w_{t,v}$ signifies weighted molecular similarity between compound pairs t and v ; $S_{t,v}$ represents pairwise molecular similarity between compounds t and v ; y_t and y_v represent observed bioaccumulation parameter values for compounds t and v , respectively.

S6. Summary of Endpoints and Integrated Datasets

Endpoint	Number of Unique Chemicals	Class Distribution
Fish acute toxicity	4697	0 (520)/1 (4177)
Fish chronic toxicity	542	0 (155)/1 (387)
Crustacean acute toxicity	4285	0 (495)/1 (3790)
Crustacean chronic toxicity	1033	0 (397)/1 (636)
Algal acute toxicity	4107	0 (399)/1 (3708)
Algal chronic toxicity	2850	0 (1762)/1 (1088)

S7. Construction of Tanimoto Similarity Distributions

Pairwise structural diversity of the aquatic toxicity datasets was further characterized using Tanimoto similarity distributions. For each endpoint, the constructed dataset was compared with the corresponding reference dataset containing the largest number of chemicals. Morgan fingerprints were generated using RDKit (<http://www.rdkit.org> accessed on 3 December 2025), with a radius of 2 and a bit length of 2048. For each dataset, Tanimoto similarity was calculated between all unique molecular pairs within the dataset. The resulting similarity values were grouped into equal-width bins over the range from 0 to 1, and density distributions were plotted to compare the structural similarity patterns between the constructed dataset and the corresponding reference dataset. Meanwhile, a smoothed curve was overlaid on the distribution of the constructed dataset to present the overall trend more

clearly. This analysis was used to evaluate whether the constructed dataset still retained broad structural diversity relative to existing large reference datasets.

S8. Methods for Physicochemical Property Calculation

Physicochemical property distributions were further analyzed to evaluate whether the constructed datasets retained representative property variation relevant to aquatic toxicity modeling. Four descriptors relevant to aquatic toxicity were calculated, including logKow, molecular weight, APol, and topological polar surface area. The logKow values were calculated using the Crippen method implemented in RDKit (<http://www.rdkit.org> accessed on 4 December 2025). Molecular weight was calculated using the RDKit molecular descriptor function. Topological polar surface area was calculated using the RDKit TPSA algorithm. APol was approximated as the sum of atomic polarizability constants after explicit hydrogens were added to each molecule. For each descriptor, density histograms were generated using equal width bins to compare the distributions between the constructed dataset and the corresponding reference dataset. In addition, a Gaussian kernel density curve was overlaid on the constructed dataset to facilitate visualization of the overall trend. This analysis was used to assess whether the constructed datasets covered a broad physicochemical property space relevant to aquatic toxicity modeling.

S9. Details on Method Chemical Clustering and Visualization

Firstly, Morgan fingerprints were generated for all curated molecules, using the open-source RDKit v2024.3.3 (<http://www.rdkit.org> accessed on 4 December 2025). The fingerprints were calculated with a radius of 2 and a fixed bit vector length of 1024 bits. To identify structurally chemical communities and assess the structural diversity of the dataset, chemical clustering was conducted using the Butina algorithm [97]. This unsupervised clustering approach is particularly advantageous for chemical datasets as it does not require a pre-specified number of clusters and guarantees that the centroid of each cluster is the most densely populated molecule. The clustering was based on pairwise Tanimoto similarity indices and similarity threshold was set to 0.87. To focus on significant chemical families and mitigate the influence of singleton outliers, only clusters containing at least two molecules were quantified as valid chemical communities.

To visually compare the chemical space coverage of the studied dataset against the reference database, principal component analysis (PCA) was performed. PCA reduces the dimensionality of the high-dimensional fingerprint vectors while preserving the maximum variance in the data. The molecules were projected into a two-dimensional space defined by the first two principal components (PC1 and PC2),

allowing for an assessment of the distribution and overlap of chemicals between the datasets.

S10. Calculation of Endpoint Label Consistency Rate

To evaluate the consistency of toxicity labels between different endpoints, a pairwise comparison was performed. For any given pair of endpoints (e.g., fish acute toxicity vs. fish chronic toxicity), shared compounds were identified based on their SMILES strings. The label consistency (C), is defined as the ratio of the number of compounds with identical toxicity labels between the two endpoints to the total number of shared compounds. The formula is calculated as follows:

$$C = \frac{n}{N} \times 100\% \quad (S3)$$

where C demotes label consistency (%); n represents the number of shared compounds where both endpoints assigned the same label (i.e., both 0 or both 1); N denotes the total number of shared compounds found in the intersection of the two endpoints.

S11. Processing and Screening of Chemicals Inventory

The constructed model was employed to predict aquatic toxicity for six endpoints across approximately 10^6 chemicals, which were curated from previous research [63]. After filtering the approximately one million initial chemicals with the DiPTox, 783,681 molecules remained. A total of $6 \times 783,681$ “end-to-end” predictions from the MGs to the toxicities took about 4 h on a computer equipped with a 3090 graphics processing unit. A total of 49,810 chemicals within AD were identified that simultaneously exhibit all six aquatic toxic effects and corresponding screening results are provided in Table S20.

S12. Details on Visualization of Inventory Chemicals with Six Endpoints Toxicity

First, generating Morgan fingerprint (radius = 2, nBits = 2048) similar to the above “9. Details on Method Chemical Clustering and Visualization”. Subsequently, principal component analysis (PCA) was applied to the fingerprint vectors to reduce the feature space. This preliminary linear reduction serves to mitigate noise and reduce computational redundancy while preserving the primary variance of the dataset. Following PCA, the reduced features were embedded into a two-dimensional manifold using the uniform manifold approximation and projection (UMAP) algorithm. To balance the preservation of local and global structures, the key parameters were set to $n_neighbors = 30$ and $min_dist = 0.1$. To identify distinct chemical communities within the projected space, k -means clustering was performed on the 2D embedded vectors. The number of clusters (k) was set to 16. Each cluster was assigned a unique alphabetical label (a–p) to facilitate subsequent analysis and visualization.

S13. Visualization of the Chemical Space Distribution

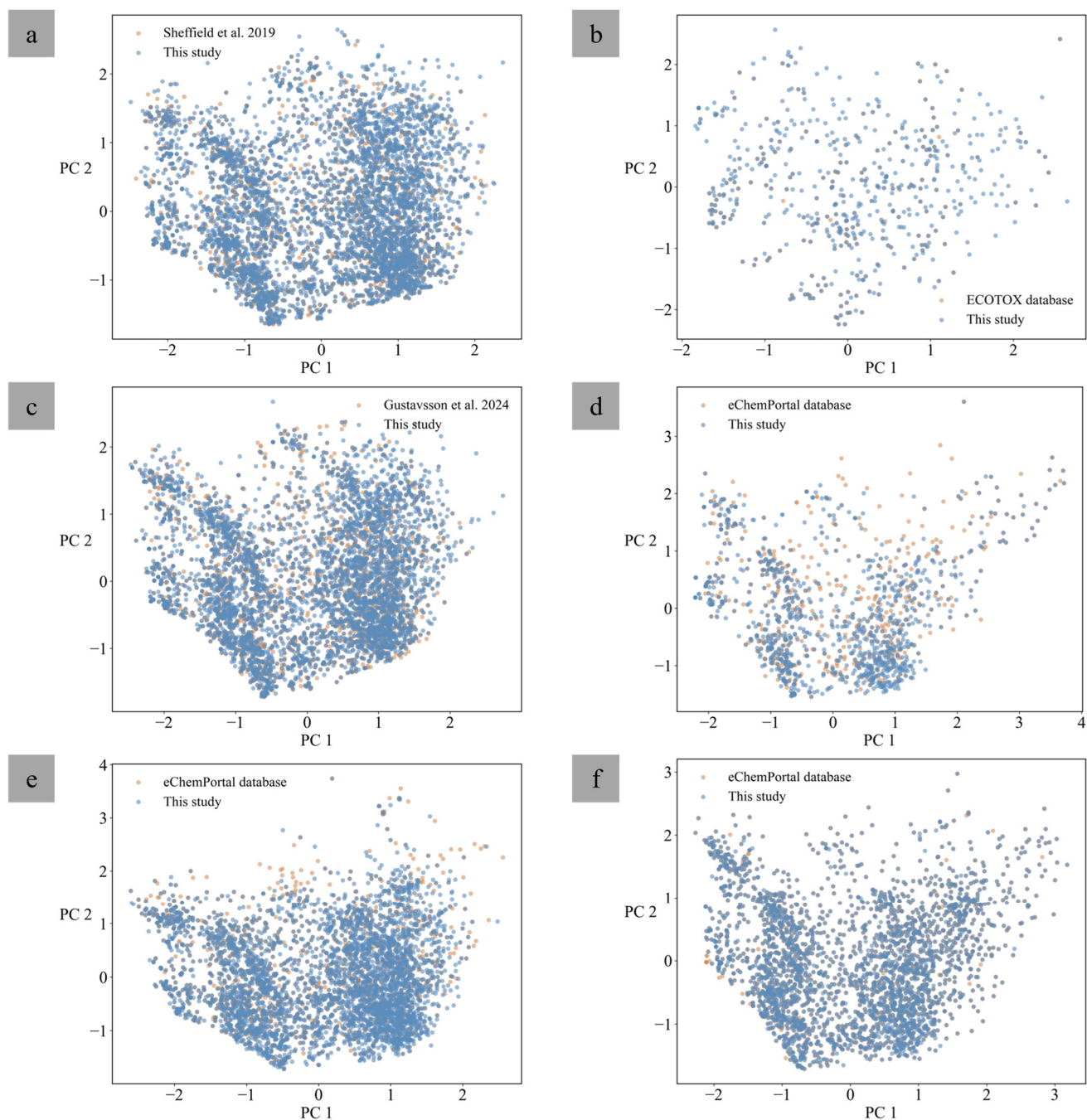


Figure S1. Visualization of the chemical space distribution for six aquatic toxicity endpoints through principal component analysis, including (a) fish acute toxicity [27], (b) fish chronic toxicity [71], (c) crustacean acute toxicity [26], (d) crustacean chronic toxicity [72], (e) algal acute toxicity [72] and (f) algal chronic toxicity [72]. In each plot, molecules from the current study are represented by blue points, while those from prior studies or databases are shown in orange.

S14. Visualizations of Tanimoto Similarity Distribution

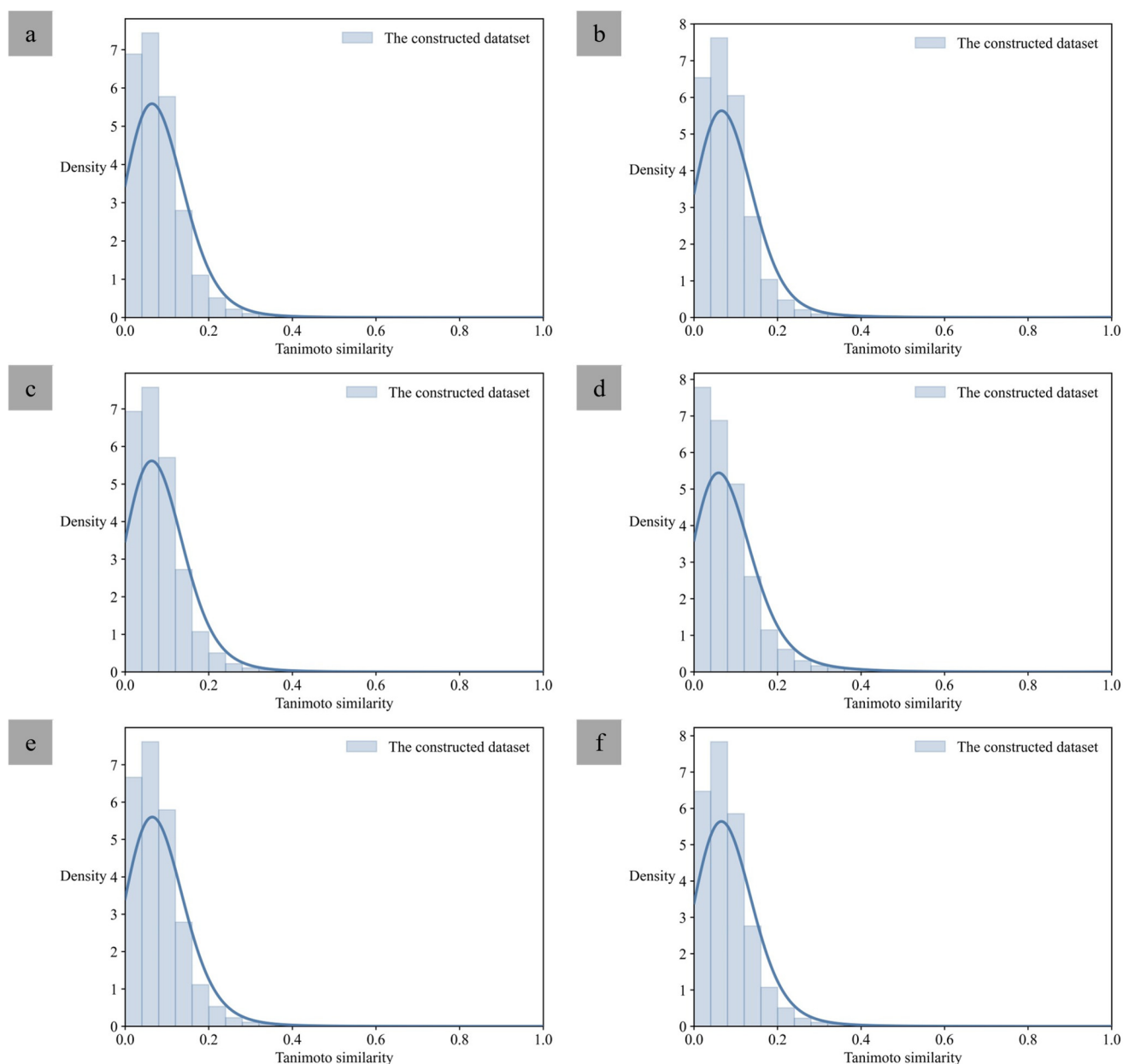
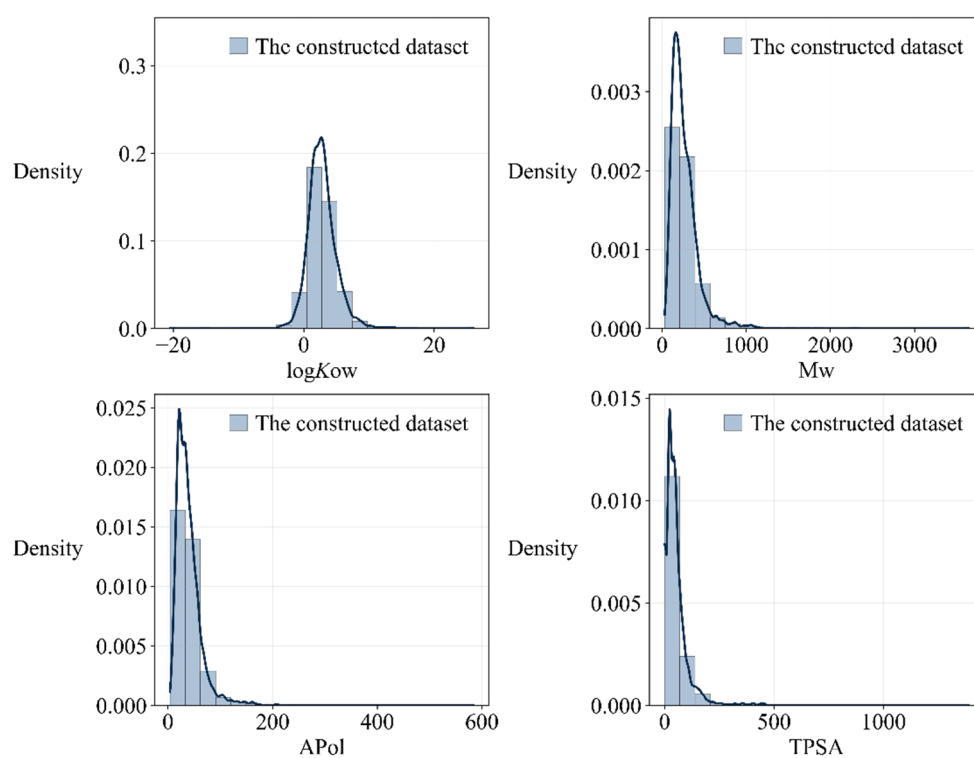


Figure S2. Pairwise Tanimoto similarity distributions for the constructed datasets across six aquatic toxicity endpoints. Histograms show the density distributions of pairwise Tanimoto similarity values within each dataset. The blue smoothed fitted curve is overlaid on the histogram of the constructed dataset to facilitate visualization of the overall trend. (a–f) correspond to fish acute toxicity, fish chronic toxicity, crustacean acute toxicity, crustacean chronic toxicity, algal acute toxicity, and algal chronic toxicity, respectively.

S15. Visualization of Physiochemical Property Distribution

(A)



(B)

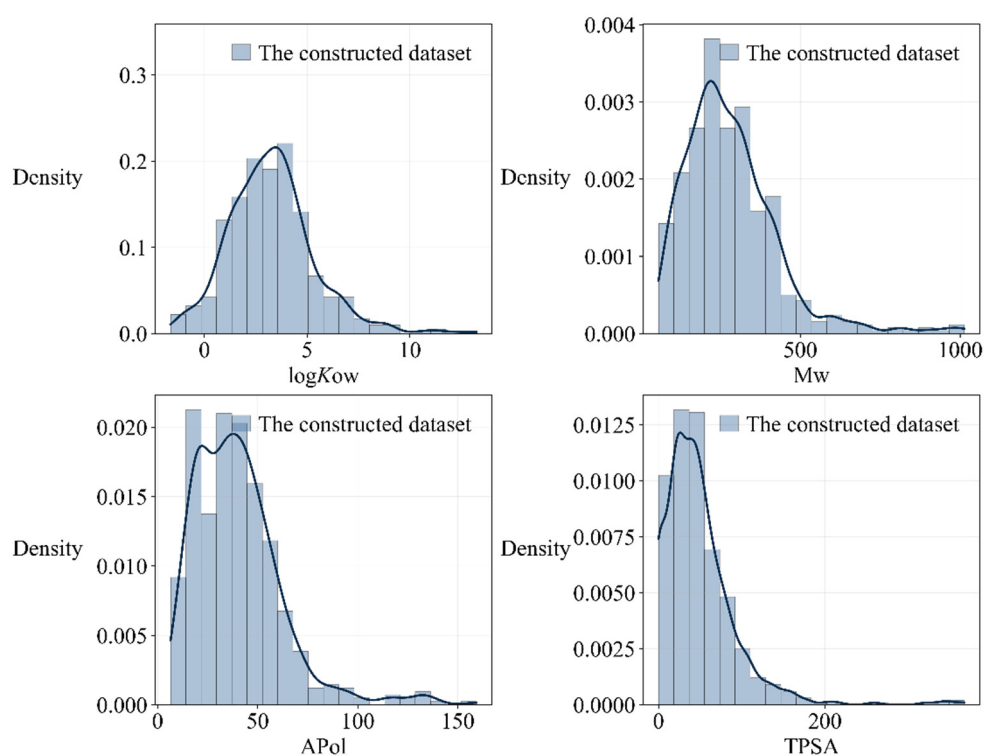
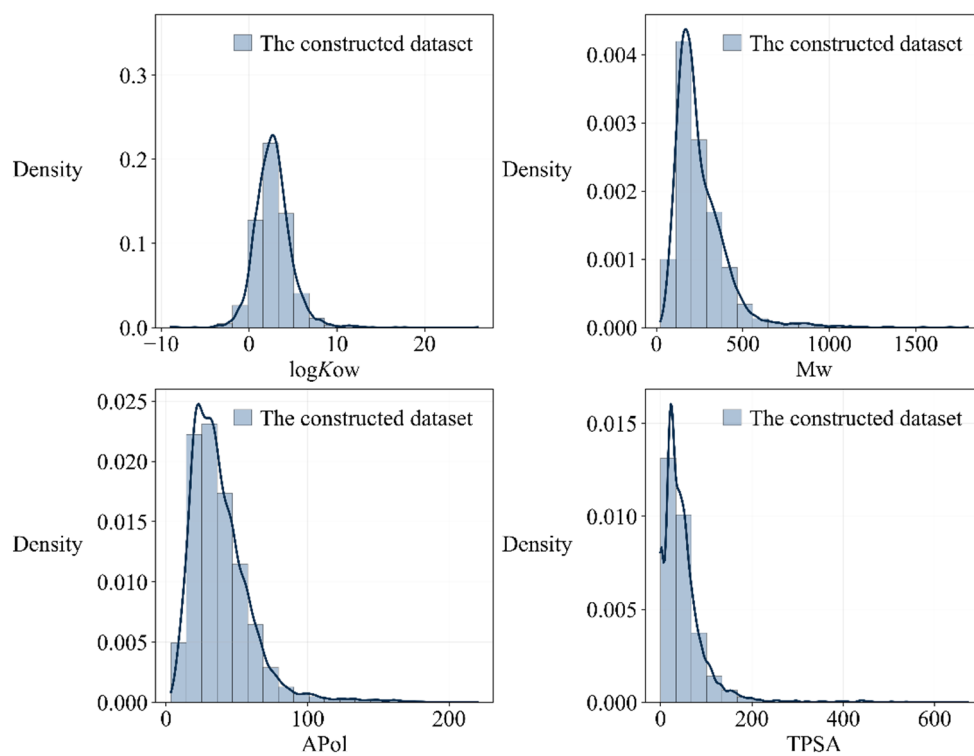


Figure S3. Cont.

(C)



(D)

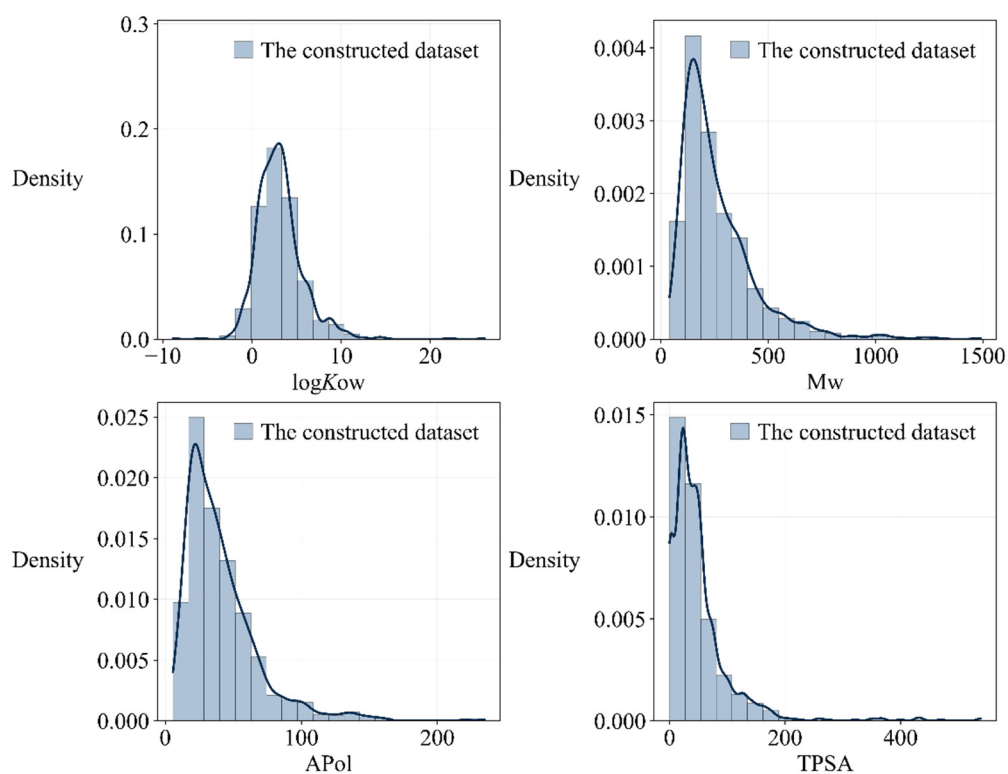
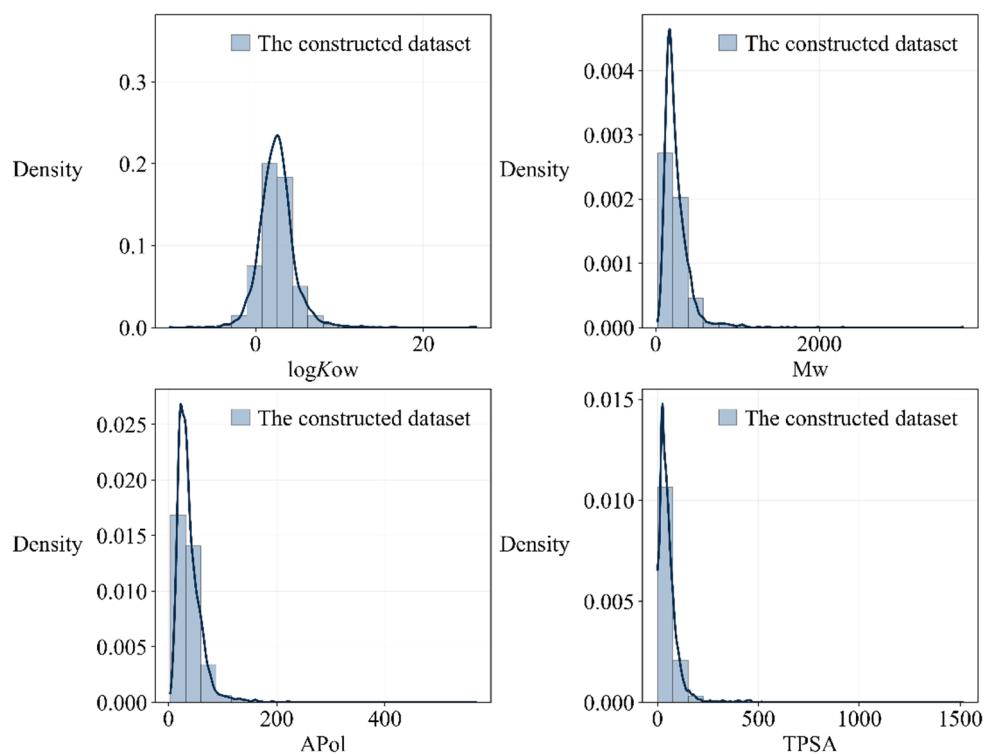


Figure S3. *Cont.*

(E)



(F)

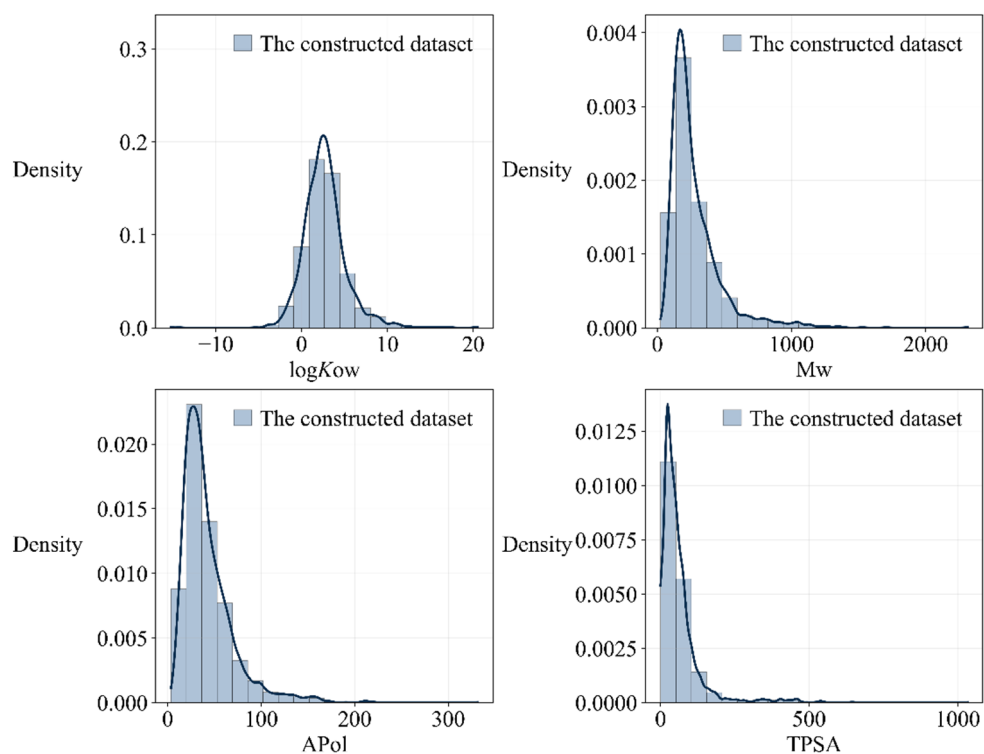


Figure S3. Physicochemical property distributions of descriptors relevant to aquatic toxicity for the constructed datasets across six endpoints. Histograms show the density distributions of logKow, molecular weight, the sum of atomic polarizabilities (APol), and topological polar surface area (TPSA) for each endpoint. The blue smoothed fitted curve was overlaid on the constructed dataset to facilitate visualization of the overall distribution trend. (A–F) correspond to fish acute toxicity, fish chronic toxicity, crustacean acute toxicity, crustacean chronic toxicity, algal acute toxicity, and algal chronic toxicity, respectively.