Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

DLG-Sol: Combining descriptor-based and language–3D models for aqueous solubility prediction

Hyunmo Goo
Hankuk University of Foreign Studies
Hyeongwoo Kong
gho3283@hufs.ac.kr

Hankuk University of Foreign Studies

Article

Keywords: aqueous solubility, molecular machine learning, molecular representations, XGBoost, three-dimensional message-passing neural network

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11122610/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

DLG-Sol: Combining descriptor-based and language–3D models for aqueous solubility prediction

Hyunmo Goo[1] and Hyeongwoo Kong[2*]

[1]Department of Chemistry, Hankuk University of Foreign Studies, 81 Oedae-ro, Mohyeon-eup, Cheoin-gu, Yongin-si, 17035, Gyeonggi-do, South Korea.

[2*]Division of Finance & AI, Hankuk University of Foreign Studies, 81 Oedae-ro, Mohyeon-eup, Cheoin-gu, Yongin-si, 17035, Gyeonggi-do, South Korea.

*Corresponding author(s). E-mail(s): gho3283@hufs.ac.kr;

Abstract

Molecular descriptors, molecular-language embeddings, and 3D graph embeddings have been used to predict aqueous solubility. However, it remains unclear how much combining models that use these inputs improves prediction over the individual components, or how the gains vary across datasets and molecular subgroups. We developed DLG-Sol by combining a Mordred-based XGBoost model with a neural model using ChemBERTa-derived language embeddings and graph embeddings from a distance-aware 3D message-passing neural network (3D-MPNN). Predictions are blended using a weight fixed across molecules within each evaluation fold or test split. Across six evaluations from four curated datasets, DLG-Sol had lower root-mean-squared error (RMSE) than either component. Paired-bootstrap 95% confidence intervals supported improvements in five evaluations against the descriptor model and four against the neural model. On a curated test set of 980 molecules, DLG-Sol achieved an RMSE of 0.6144, compared with 0.6571 for the released predictions of the original consensus graph neural network. Gains extended across molecular subgroups, although larger molecules still had higher prediction errors. Fixed-weight blending offered a simple, practical approach, as neither of the two evaluated blending schemes with molecule-specific weights provided a consistent improvement. Code and supporting materials for DLG-Sol are available at https://github.com/Tuesdayand/DLG-Sol/releases/tag/v1.0.5.

Keywords: aqueous solubility, molecular machine learning, molecular representations, XGBoost, three-dimensional message-passing neural network

1 Introduction

Aqueous solubility is an important physicochemical property that affects drug absorption, bioavailability, and formulation. Poor aqueous solubility can limit drug absorption, making solubility assessment important early in drug development [1, 2]. However, measuring solubility across large compound collections requires samples, analytical equipment, time, and resources [3, 4]. Compounds that have not yet been synthesized cannot be measured directly. Predicting aqueous solubility from molecular structure can therefore help prioritize compounds for experimental testing and reduce the burden of early-stage development.

Both empirical equations and physics-based calculations have been used to predict solubility. The General Solubility Equation (GSE) estimates aqueous solubility from melting point and the octanol–water partition coefficient [5]. Although straightforward to apply, its common reliance on measured melting points limits direct application to unsynthesized compounds [6]. The conductor-like screening model for real solvents (COSMO-RS) takes a different approach, combining molecular surface-charge information from quantum-chemical calculations with statistical thermodynamics to describe interactions in solution [7, 8]. Generating the required molecular surface-charge information involves electronic-structure calculations and, for flexible molecules, conformer sampling [4]. The availability of input properties and the computational effort required are therefore important considerations when screening large numbers of candidate compounds.

Machine learning offers an alternative by learning relationships between molecular structure and measured solubility. Predictions depend on both how a molecule is represented and how the model learns from that representation. Molecular descriptors summarize physicochemical and topological properties, whereas fingerprints encode substructures or local atomic environments [9, 10]. Neural networks can instead learn embeddings from inputs such as simplified molecular-input line-entry system (SMILES) strings or atom–bond graphs. For example, Ali et al. developed an XGBoost model using molecular descriptors [3], while Tang et al. used a self-attention-based message-passing neural network (SAMPN) to predict aqueous solubility from molecular graphs [11]. ChemBERTa-2 applied pretraining on SMILES strings to molecular property prediction [12]. These approaches use different molecular inputs and learning algorithms to obtain predictions.

A single representation and learning algorithm may not be equally effective for every molecule. As discussed in the review by Wigh et al., molecular representations simplify the underlying molecule, and their usefulness depends on whether they retain information relevant to the property being predicted [13]. Descriptors, strings, graphs, and three-dimensional structures provide molecular information in different forms. Different learning algorithms can also extract different predictive relationships from the same inputs [3]. These differences suggest that models may make complementary errors, providing a reason to investigate their combination. Such combinations can operate at two levels: features derived from different representations can be learned jointly, or predictions from separately trained models can be combined. Hybrid approaches thus encompass both the integration of molecular features and the integration of predictive models.

Recent studies illustrate both forms of integration. Molecular Property prediction with multimodal language and graph fusion (MolPROP) maps embeddings from a pretrained molecular-language model to atom nodes in a molecular graph, allowing language and graph information to be learned together [14]. The XG graph neural network (XGNN) combines an XGBoost model using descriptors and fingerprints with a directed message-passing neural network and uses a weighted average of their predictions to estimate aqueous solubility [15]. Ahmad et al. evaluated descriptors, fingerprints, and Mol2Vec embeddings and combined regression-model predictions through stacking [16]. These studies show that hybrid modelling involves more than increasing the number of input representations: it can also bring together models that process molecular information differently.

These developments support the use of combined representations and models for solubility prediction. An important remaining question is whether descriptor-based predictions provide additional value when a neural model already incorporates both molecular-language and three-dimensional information. Differences between the component models do not by themselves establish how much their combination improves on the stronger model. An improvement in overall performance also does not reveal how the benefit varies across evaluation datasets and molecular subgroups. Assessing this combination therefore requires direct comparisons against both components, estimates of the size and uncertainty of the gains, and an examination of the molecular groups that remain difficult to predict.

We developed the descriptor–language–geometry solubility model (DLG-Sol) to address this question. A developmental evaluation of molecular representations and predictor combinations guided the selection of a neural model that combines ChemBERTa-derived language embeddings with molecular embeddings from distance-aware three-dimensional message passing. We then combined its predictions with those of a separately trained Mordred–XGBoost model. DLG-Sol therefore integrates embeddings within the neural component and predictions between two distinct models. Across six evaluations from four datasets, we quantified improvements over each component and assessed their uncertainty using paired

analyses on the same molecules. Subgroup and residual analyses examined where gains occurred and where errors remained. Comparisons with selected external models and two molecule-specific weighting schemes provided further assessments of performance and the benefit of varying the blend weight. Together, these analyses quantified the predictive value of combining a language–3D neural model with a descriptor model and identified difficulties that persisted after combination.

2 Methods

2.1 Datasets and evaluation designs

The four primary benchmarks were the curated version of AqSolDB [17] used in this study (AqSolDBc), the compiled solubility benchmark of Ali et al. (ComPlat), the Therapeutics Data Commons (TDC), and the curated water-solubility benchmark of Ulrich et al. (JCheM). Table 1 describes their endpoints and roles, together with two supplementary sources. Here, "aqueous solubility" covers related endpoints rather than one uniformly measured assay.

Each benchmark retained its recorded fold assignments or supplied train–test partitions. ComPlat and TDC each contributed both an out-of-fold (OOF) evaluation and a supplied-test evaluation. For TDC, we retained the ADMET Benchmark Group's supplied scaffold train–test partition and created five logS-stratified OOF folds within the training/validation set. AqSolDBc contributed an OOF evaluation and JCheM a supplied-test evaluation, giving six evaluations in total. In each evaluation, the models were scored on the same molecules. Scored labels were excluded from preprocessing, hyperparameter selection, coefficient estimation, and model fitting performed for that split. Some settings were selected earlier using overlapping development data and retained across outer folds (Section 2.2).

The evaluation populations also share source data. AqSolDBc and TDC derive from AqSolDB, as do 7,605 of the 9,800 records in the JCheM curation workflow [18]. Four of the six evaluations therefore trace to AqSolDB. Parts of the TDC and ComPlat test sets also overlap the AqSolDBc training population. We report the six evaluations separately, accounting for these shared origins rather than treating them as independent replications. Supplementary Tables S1–S3 describe the datasets, representations, and conformer-generation failures (Supplementary Information). Supplementary Tables S4–S5 give exact metrics, overlap counts, and label-access histories. The ComPlat analysis of fold-local DLG-Sol reselection and matched-budget Ali XGB-125D tuning is reported in Supplementary Information ("ComPlat selection sensitivity") and Supplementary Table S6.

The Biogen pH 6.8 chemiluminescent nitrogen detection (CLND) dataset [19] represents a different kinetic-solubility endpoint and was therefore treated as supplementary. The five panels reconstructed from Ghanavati et al. [20] were also treated as supplementary owing to their retrospective evaluation after prior label access. Their detailed roles and results are reported in Supplementary Table S7 and the associated Biogen and Ghanavati sections.

2.2 Developmental representation screening and neural-component selection

Model development began with a screen of seven molecular representations. These included Mordred descriptors and extended-connectivity fingerprints (ECFPs). The other five were embeddings generated by ChemBERTa, a three-dimensional message-passing neural network (3D-MPNN), an atom-level 2D graph neural network (GNN), a BRICS-subgroup GNN, and a Murcko-subgroup GNN. We evaluated each representation and all 21 representation pairs with five types of predictor models. These were XGBoost, tree ensembles (random forest or Extra Trees), regularized linear regression (ridge or Elastic Net), k-nearest-neighbour regression, and a multilayer perceptron (MLP).

The AqSolDBc development split comprised 6,437 training and 1,610 validation rows. Together, these form the 8,047-row population used in the subsequent AqSolDBc OOF evaluation. Hyperparameter trials were selected on the validation partition. An additional development panel was then used to rank the representation–predictor combinations.

On that panel, Mordred with XGBoost was the strongest single-representation candidate, with a root-mean-squared error (RMSE) of 1.276. Combining ChemBERTa embeddings and 3D-MPNN embeddings with an MLP predictor achieved the lowest RMSE across the initial single- and pair-representation

Table 1 Datasets, solubility endpoints, and roles in this study. Targets are $\log_{10}(S)$, with S in mol L^{-1}.

Dataset	Endpoint represented	Use in this study
AqSolDBc [21]	Curated thermodynamic solubility in water near pH 7 and approximately 300 K. The underlying AqSolDB sources are heterogeneous, so the dataset is not treated as a uniform intrinsic-solubility assay.	**Primary development OOF**. 8,047 molecules in five folds
ComPlat [3]	Training data compiled and curated by Ali et al. from four sources. The supplied test is the Huuskonen dataset [22]. Source assay conditions and intrinsic/apparent status are not uniformly resolved.	**Primary replication**. 17,937 OOF rows and 1,282 supplied-test rows
TDC [23]	AqSolDB-derived aqueous-solubility compilation with heterogeneous source conditions. It is not treated as uniformly intrinsic.	**Primary lineage-qualified replication**. Supplied scaffold train–test partition; 7,985 OOF rows and 1,997 supplied-test rows
JCheM [18]	Curated water solubility ($\log S_w$) for organic chemicals. The source defines water rather than intrinsic solubility.	**Primary supplied test and external-comparator comparison**. 980 rows
Biogen [19]	pH 6.8 CLND kinetic solubility, a different assay endpoint from the primary water-solubility compilations.	**Supplementary endpoint shift**. 2,153 molecules
Ghanavati reconstructions [20]	Five source-dependent aqueous-solubility panels: All-Data, ESOL, AQUA, PHYS, and OCHEM.	**Supplementary retrospective analysis**. Five test panels, $n = 224$–844

candidates (1.265), improving to 1.249 after focused tuning. This result guided selection of the language–geometry neural component. The development-panel labels thus contributed to architecture selection. These screening values document that selection process and are not directly comparable with the final-model evaluations in Table 2. Supplementary Information ("Developmental representation screening") and Supplementary Table S19 provide the screening design and ranked results, with complete grids in the machine-readable Supplementary data (Supplementary Data).

Separate post-development comparisons of single representations and representation pairs are summarized in Supplementary Tables S14–S16, with complete results in the machine-readable Supplementary data. These exploratory comparisons used a common low-budget search and were not capacity-matched ablations of the final neural model.

2.3 Model architecture

Figure 1 shows the DLG-Sol workflow from molecular input through the separately trained descriptor (panel A) and language–geometry neural (panel B) models to their fixed prediction-level combination (panel C). Both component models receive molecular structures represented by SMILES. Panel A produces the descriptor prediction, $\hat{y}_{\mathrm{XGB}}$, and panel B produces the neural prediction, $\hat{y}_{\mathrm{neural}}$.

Descriptor model (panel A).
Ali et al. found that expanding the molecular feature set did not always improve test performance [3]. Motivated by this finding, we filtered the descriptor set before training XGBoost. We calculated 1,613 raw two-dimensional Mordred descriptors from canonical SMILES [9]. Filtering based on the training data reduced this set to 337–654 descriptors, depending on the dataset and split. We then trained XGBoost on the retained descriptors using the squared-error objective [24]. Descriptor filtering and model selection used only the relevant training partition. The rows being scored did not determine which descriptors were retained or which XGBoost configuration was selected. The retained descriptors were supplied to the fitted XGBoost model to produce one descriptor prediction, $\hat{y}_{\mathrm{XGB}}$, per molecule. Dataset-specific filtering and configuration-search procedures are reported in Supplementary Information ("Descriptor preprocessing") and Supplementary Table S8. The correction to the AqSolDBc descriptor pipeline and the fold-specific tree counts are reported in Supplementary Table S9, and software environments and refit seeds are summarized in Supplementary Table S10.

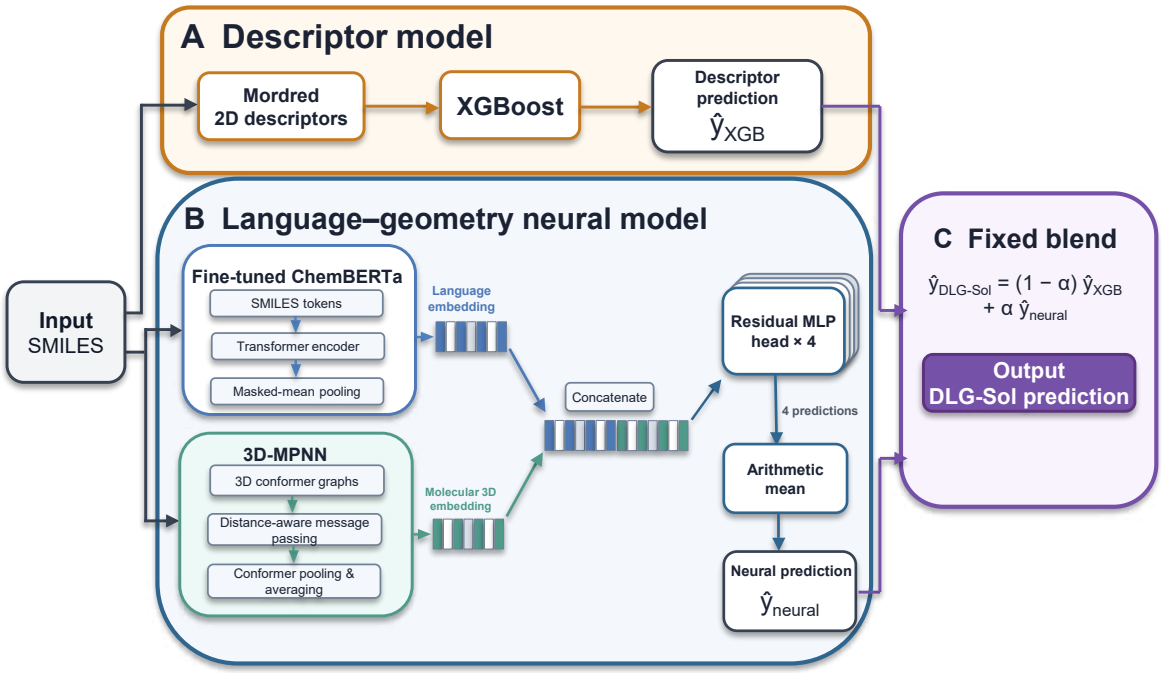


Fig. 1 Architecture of DLG-Sol. Panel A calculates Mordred descriptors and applies XGBoost (XGB) to obtain the descriptor prediction, $\hat{y}_{\mathrm{XGB}}$. In panel B, masked-mean pooling of ChemBERTa token states produces a molecular-language embedding, while node pooling within each retained conformer followed by averaging across conformers produces a molecular 3D embedding. The preprocessed embeddings are concatenated and supplied to four separately trained residual multilayer perceptron (MLP) heads with identical architecture. The arithmetic mean of their four predictions defines $\hat{y}_{\mathrm{neural}}$. Panel C combines the final predictions from the two components using a coefficient fixed before scoring and applied unchanged to every molecule in the corresponding held-out fold or test split.

ChemBERTa model (panel B).

The ChemBERTa model in panel B used a single-task encoder initialized from the pretrained ChemBERTa zinc-base v1 model [25]. Tokenized SMILES were passed through the fine-tuned transformer, which retained its pretrained Gaussian error linear unit activations [26]. Masked-mean pooling over the non-padding final-layer token states then produced one 768-dimensional molecular-language embedding per molecule. During fine-tuning, each molecule was represented by its canonical SMILES string and two independently randomized SMILES strings [27], and all encoder layers were optimized only for log-solubility prediction. Inference used canonical SMILES only. Fine-tuning settings are summarized in Supplementary Table S8.

3D-MPNN (panel B).

Input preparation for the 3D-MPNN model in panel B began with conformer generation. For each molecule, we retained the largest organic fragment and generated up to five conformers with experimental-torsion knowledge distance geometry version 3 (ETKDGv3) [28, 29]. Conformers were optimized with the Merck molecular force field when parameters were available and otherwise with the universal force field [30, 31]. Up to three lowest-energy conformers were retained as separate graphs. Graph nodes represented heavy atoms and hydrogens bonded to nitrogen or oxygen. We refer to the latter as explicit polar hydrogens. Each graph included covalent bonds and non-bonded atom pairs within 4.5 Å as edges. Each edge was represented by bond features, interatomic distance, and 16 Gaussian radial-basis values. We used the general form of Gaussian distance encoding used in SchNet, with study-specific choices for the number, centres, and width of the basis functions [32].

Each retained conformer graph was processed by the distance-aware 3D message-passing neural network (3D-MPNN), which followed the general message-passing framework [33]. Each message was computed from the receiver-node state, sender-node state, and edge attributes, and incoming messages were summed. A residual update, layer normalization, and a sigmoid-weighted linear unit (SiLU)

activation then updated each node representation [34]. Global mean pooling over nodes produced one representation per conformer. Averaging the retained conformer representations then produced one molecular 3D embedding. When conformer generation or graph embedding failed for a molecule, the missing 3D representation was replaced with a label-free mean embedding calculated from the relevant fitting partition. Fewer than 0.4% of records in each final benchmark-specific processed dataset required this fallback (Supplementary Table S3). The encoder uses scalar node features and fixed interatomic distances without updating atomic coordinates. A high-level definition of the 3D representation is provided in Supplementary Table S2, and feature dimensions and searched configuration ranges are summarized in Supplementary Table S8. Exact node, edge, and auxiliary molecular features, together with the message equations, radial-basis parameters, and layer dimensions, are given in Supplementary Information ("3D-MPNN specification").

Embedding concatenation and residual MLP prediction (panel B).
The molecular-language and molecular 3D embeddings were processed separately using statistics from the relevant fitting partition and then directly concatenated, as shown at the centre of panel B. Each benchmark pipeline standardized both modalities. The AqSolDBc, ComPlat, and TDC pipelines additionally winsorized each embedding dimension at percentiles 0.1 and 99.9, estimated from the training partition, whereas the JCheM pipeline did not. No principal-component projection or L_2 normalization was used in the final neural model.

The concatenated embedding was supplied to four separately trained residual multilayer perceptron (MLP) heads with identical architecture. Each head produced one scalar prediction. The heads used rectified linear unit activations [35] and differed in initialization, minibatch order, and fixed training duration. They were optimized with mean-squared error and AdamW [36]. The arithmetic mean of their four predictions defined the final neural prediction, $\hat{y}_{\text{neural}}$. Exact head layouts, seeds, fixed epoch counts, and optimization settings are reported in Supplementary Table S8.

Fixed prediction-level blending (panel C).
Panel C combines the final component predictions $\hat{y}_{\text{XGB}}$ and $\hat{y}_{\text{neural}}$ as

$$\hat{y}_{\text{DLG-Sol}} = (1-\alpha)\hat{y}_{\text{XGB}} + \alpha\hat{y}_{\text{neural}},$$

where α is the weight assigned to the neural prediction. The component models were trained separately and were not joined by end-to-end optimization.

We used equal weights for AqSolDBc and estimated the weights for the other evaluations from development predictions and labels. Each weight was fixed before scoring and shared by all molecules in the corresponding fold or test split. Supplementary Table S11 gives the fitting procedures and coefficient values.

For JCheM, we first blended the component predictions using the validation-fitted coefficient for each of the five supplied splits. We then averaged the five blended predictions for each of the 980 shared test molecules before calculating RMSE and mean absolute error (MAE).

2.4 External comparator models

We used authors' released predictions whenever they were available for the molecules being evaluated. Other comparisons used models fitted from public code or reported procedures on the same splits as DLG-Sol. Supplementary Table S13 distinguishes released predictions, same-split reproductions, adaptations of reported descriptor schemas and settings, independently retrained adaptations, and historical point estimates. Paired analyses required predictions for the same molecules. Source-paper point estimates without those predictions are presented separately in Table 3 and Supplementary Table S20.

Four comparator models were evaluated across the six primary evaluations. The Pacific Northwest National Laboratory (PNNL) graph neural network (GNN) [37], additive Bhattacharya–Roy MLP–GNN [38], and Consensus GNN [18] were independently retrained. Ali XGB-125D used its reported descriptor schema and settings [3]. Its ComPlat evaluations reproduced the released procedure on the same splits, while the other evaluations adapted that procedure to new datasets. For Consensus GNN,

the authors' released JCheM predictions provided an additional comparison alongside the retrained official-code adaptation. The Bhattacharya–Roy extension with a multiplicative interaction term was evaluated as a supplementary sensitivity analysis.

Public artefacts did not permit exact source-model reproduction for several comparators. Those results describe the evaluated adaptations. Some adaptations retained the reported hyperparameters without a dataset-specific search comparable to DLG-Sol's, so the comparisons do not estimate each source model's best attainable performance. Supplementary Information ("External comparator assessment") and Supplementary Table S13 document the available artefacts, reconstruction attempts, and tuning conditions.

XGNN [15] was not included in the comparator set because no public executable implementation or row-level predictions were available for a split-matched evaluation.

MolPROP [14] was also excluded because the released artefacts did not link the available training and evaluation workflow to the published result. Its audit is reported in Supplementary Information ("External comparator assessment") and Supplementary Table S13.

2.5 Molecule-specific weighting

To test whether the blend benefited from weights that vary between molecules, we fitted two gates to the same component predictions. Such weighting follows the general mixture-of-experts approach [39], also explored for solubility prediction in ALipSol [40]. The first, a cluster-soft gate, used soft membership in five clusters to assign blend weights. The second, a PCA-linear gate, mapped principal component analysis (PCA) coordinates to a weight clipped to $[0, 1]$. Both gates used ten structure-derived variables. Their construction is detailed in Supplementary Information ("Molecule-specific weighting").

Gate hyperparameters were selected on AqSolDBc development OOF predictions and transferred unchanged. Transformations and coefficients were fitted using the development data available to each evaluation. For OOF evaluations, fitting excluded the held-out fold. For supplied tests, ComPlat and TDC used their own development OOF predictions, while JCheM and Biogen used AqSolDBc development OOF predictions. Scored test labels were excluded from gate fitting.

2.6 Model evaluation, statistical analysis, and implementation

Root-mean-squared error (RMSE) is the primary metric, and mean absolute error (MAE) is secondary:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \qquad \mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,$$

where y_i and $\hat{y}_i$ are the measured and predicted log-solubility values for molecule i.

For comparisons on the same molecular rows, we report the RMSE difference between DLG-Sol and its comparator. Percentile 95% and 99% confidence intervals were obtained from 10,000 molecular-row bootstrap resamples that retained each row's paired predictions [41]. An interval excluding zero indicates a detectable difference at the stated confidence level. An interval containing zero is treated as inconclusive rather than as evidence of equivalence.

Chemical-subgroup analyses used seven structural-feature axes defined before inspecting the subgroup results. These described molecular weight, rotatable bonds, hydrogen-bond donors and acceptors, aromatic rings, nitrogen-containing heterocycles, and carboxyl groups. The axes defined 16 potentially overlapping subgroups, analysed separately in each evaluation. Supplementary Information ("Subgroup and residual analyses") gives the definitions and the residual-correlation and large-error-overlap analyses. Evaluation–subgroup rows with fewer than 100 molecules are listed in Supplementary Table S17, with complete subgroup effects, confidence intervals, and sample sizes in the machine-readable Supplementary data. These exploratory subgroup analyses were not adjusted for multiple comparisons. The bootstrap intervals quantify molecular-row sampling uncertainty conditional on the fitted predictions; they do not include variability from repeating model development or training.

DLG-Sol was implemented in Python 3.9.23 with RDKit 2023.09.6, Mordred 1.2.0, scikit-learn 1.6.1, XGBoost 2.1.1, PyTorch 2.7.1, PyTorch Geometric 2.6.1, and Transformers 4.50.3. Neural training used

NVIDIA RTX 4090 graphics processing units (GPUs). Descriptor and tree-model calculations used an Intel Xeon w7-3465X processor. Software environments and random-seed records are summarized in Supplementary Table S10. Complete split-, model-, and contrast-level seed records are supplied with the machine-readable Supplementary data and code archive.

2.7 Code availability

Source code, configurations, software environments, and workflows are available in the DLG-Sol v1.0.5 release at https://github.com/Tuesdayand/DLG-Sol/releases/tag/v1.0.5. The original DLG-Sol code is distributed under the MIT License. The release supports fresh training, prediction, fixed blending, and scoring from user-supplied benchmark inputs verified by file hashes. It does not include historical fitted weights or third-party pretrained weights and code; instructions for obtaining the required third-party resources are provided in the repository.

3 Results

3.1 Performance relative to the component models

DLG-Sol had lower RMSE than either component in all six evaluations (Table 2). Paired 95% confidence intervals supported the improvements over both components in each OOF evaluation. On the supplied tests, improvements were supported over the descriptor model for TDC and JCheM, and over the neural model for ComPlat. Overall, five of the six descriptor comparisons and four of the six neural comparisons supported an improvement.

The gain over the stronger component was smaller on the supplied tests. On the ComPlat test, the descriptor model had RMSE 0.562, whereas the neural model had RMSE 0.672. Combining them reduced RMSE to 0.560. On the TDC test, the neural model was stronger, and blending reduced RMSE from 1.005 to 1.002. The paired intervals for these two differences included zero, as did the interval for the JCheM comparison with the neural model.

Table 2 RMSE for DLG-Sol and its components. ΔRMSE is DLG-Sol minus the indicated component, so negative values favour DLG-Sol. The blend coefficient was fixed across molecules within each fitted fold, split, or evaluation model. A dagger (†) indicates that the paired-bootstrap 99% confidence interval for the indicated difference lies entirely below zero. Exact differences and intervals are reported in Supplementary Table S4 and the machine-readable Supplementary data.

Evaluation	n	DLG-Sol	Descriptor (XGB)	ΔRMSE DLG-Sol − XGB	Neural network (NN)	ΔRMSE DLG-Sol − NN
OOF						
AqSolDBc	8,047	**0.807**	0.851	−0.044†	0.857	−0.050†
ComPlat	17,937	**0.879**	0.902	−0.023†	0.953	−0.074†
TDC	7,985	**0.986**	1.065	−0.079†	1.061	−0.074†
Supplied test						
TDC	1,997	**1.002**	1.091	−0.089†	1.005	−0.003
JCheM	980	**0.614**	0.655	−0.041†	0.625	−0.010
ComPlat	1,282	**0.560**	0.562	−0.002	0.672	−0.112†

3.2 Gains and remaining errors across molecular subgroups

The blend gains extended across chemical subgroups defined by molecular size, flexibility, hydrogen bonding, and functional groups (Figure 2). DLG-Sol had lower point-estimate RMSE than both components in 81 of the 96 evaluation-by-subgroup rows, including all 16 subgroups in each OOF evaluation. None of the remaining subgroup comparisons showed higher RMSE for DLG-Sol supported by a paired 95% confidence interval. The improvement was therefore broadly distributed across the groups examined.

Larger molecules remained more difficult to predict. DLG-Sol RMSE increased from the lowest to the highest molecular-weight group in all six evaluations. For AqSolDBc OOF, RMSE was 0.6692 below 250 Da (n = 4,708), 0.8713 from 250 to below 400 Da (n = 2,564), and 1.2372 at or above 400 Da

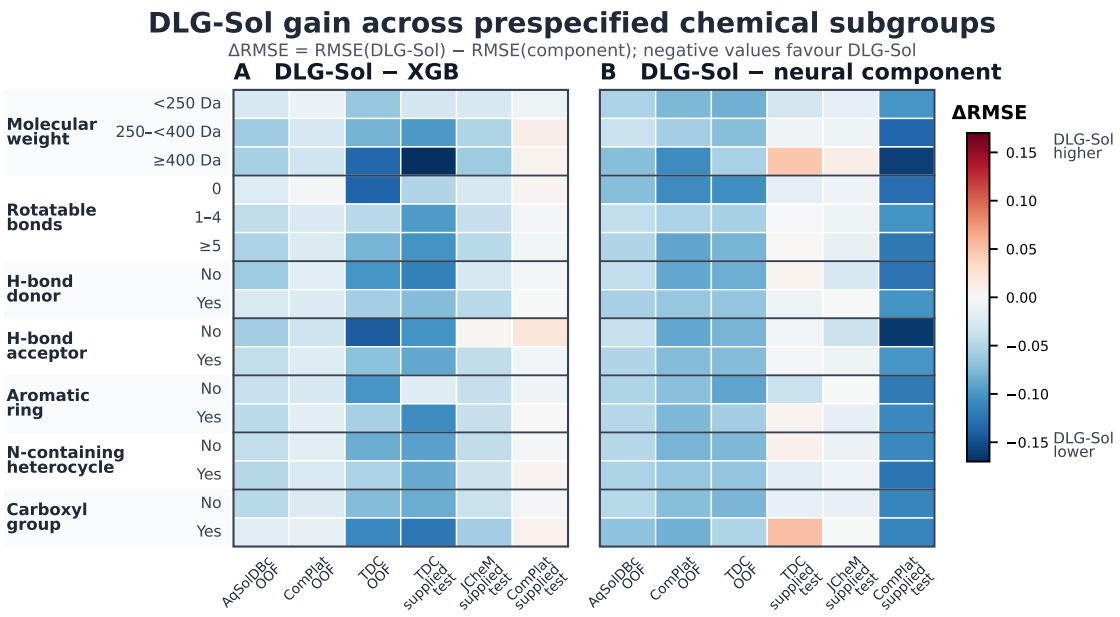


Fig. 2 Distribution of blend gain across chemical subgroups. Each cell shows ΔRMSE, calculated as DLG-Sol minus the indicated component, for one evaluation and one subgroup. Blue values favour DLG-Sol. Negative values in both panels indicate a lower point-estimate RMSE than both components. Seven structural-feature axes define 16 potentially overlapping subgroups, analysed separately. Six evaluations give 96 evaluation-by-subgroup rows, with two contrasts per row. Subgroups with fewer than 100 molecules are listed in Supplementary Table S17. Exact effects, paired-bootstrap confidence intervals, and sample sizes are provided as machine-readable Supplementary data. The analysis is exploratory and was not adjusted for multiple comparisons.

($n = 775$). The corresponding descriptor RMSEs were 0.6988, 0.9319, and 1.2937; neural RMSEs were 0.7224, 0.9095, and 1.3092. The blend improved on both components in each bin, while absolute errors remained highest for the largest molecules.

Within the largest-molecule subgroup, the neural model had lower point-estimate RMSE than DLG-Sol on the TDC and JCheM tests, whereas the descriptor model had lower RMSE on the ComPlat test. The JCheM and ComPlat test sets contained 86 and 41 molecules with molecular weights of at least 400 Da, respectively. Absolute subgroup RMSEs, sample sizes, and all paired effects are supplied in the machine-readable Supplementary data.

The component with lower RMSE differed across evaluations. None of the 16 subgroups consistently favoured the same component across all six evaluations. At the molecular level, the component residuals were correlated but not identical, with mean Pearson $r = 0.787$ and mean Spearman $\rho = 0.719$ across evaluations. Supplementary Figure S1 shows the residual correlations and overlap between large-error subsets.

3.3 Performance relative to external comparator models

We evaluated DLG-Sol on the ComPlat test set from Ali et al. and the JCheM test set from Ulrich et al. and compared its performance with the models reported in those studies [3, 18]. Table 3 retains the published comparisons and adds the DLG-Sol results on the corresponding benchmark test sets.

On ComPlat, DLG-Sol achieved RMSE 0.5599, below the values reported for the ANN, MLR, and SVM models in the source paper. Ali et al. reported RMSE 0.55 for XGB-298. Row-level predictions for XGB-298 were unavailable to us, so we could not assess the statistical significance of the difference between the two models.

On JCheM, DLG-Sol had the lowest RMSE (0.6144) among the models listed for that test set. RMSE was 0.6571 for the original authors' released Consensus GNN predictions, and the paired 95% confidence interval on the same 980 test molecules supported the improvement with DLG-Sol (Supplementary

Table S12). Evaluation sample counts and training conditions differed across the other published models, so we report their values separately from this same-molecule statistical comparison.

Table 3 Source-paper benchmark comparisons with DLG-Sol added.

Benchmark context	Model or method	Source	n	RMSE
A. Ali et al. ComPlat/Huuskonen comparison				
ComPlat test	ANN	Huuskonen et al. [42]	1,294	0.71
	ANN	Tetko et al. [43]	1,291	0.62
	ANN	Yan and Gasteiger [44]	1,294	0.59
	MLR	Delaney [45]	1,290	0.87
	MLR	Hou et al. [46]	1,294	0.63
	SVM	Schroeter et al. [47]	1,290	0.60
	MLR	Ali et al. (2012) [48]	1,290	0.94
	MLR	Daina et al. [49]	1,290	1.15
	ANN	Bjerrum and Sattarov [50]	1,297	0.65
	Consensus	Sorkun and Koelman [51], overlap-affected	1,290	(0.53)
	Consensus	Ali et al. overlap-corrected rework [3]	1,290	0.73
	XGB-298	Ali et al. (source study) [3]	1,282	0.55
	DLG-Sol	This work	1,282	0.5599
B. Ulrich et al. JCheM comparison				
JCheM test	Consensus GNN	Ulrich et al. (released predictions) [18]	980	0.6571
	EPI Suite [52]	Ulrich et al. comparison [18]	934	1.386
	OCHEM [53]	Ulrich et al. comparison [18]	980	0.703†
	ACD GALAS [54]	Ulrich et al. comparison [18]	980	0.845
	DLG-Sol	This work	980	**0.6144**

Note: Panels A and B present benchmark values compiled from Table 3 of Ali et al. [3] and Table 3 of Ulrich et al. [18], respectively, alongside DLG-Sol results from this study. Consensus GNN RMSE was recomputed from released predictions (original report: 0.657). Bold marks the lowest JCheM point estimate; paired inference is available only against released Consensus GNN predictions. Parentheses flag the overlap-affected Sorkun result. The OCHEM value (†) was flagged for training–test overlap. MAE values and comparison details are in Supplementary Table S20.

Separately, we retrained four published models on the same data splits as DLG-Sol and compared them across the six evaluations (Table 4). The Consensus GNN in this comparison was independently retrained by us and is distinct from the original authors' released predictions described above. Its JCheM test RMSE was 0.8095.

Against Ali XGB-125D, paired intervals supported lower RMSE for DLG-Sol in four evaluations. Ali XGB-125D had lower point estimates on TDC OOF and the ComPlat supplied test, but the 95% confidence intervals for both differences included zero. A separate ComPlat OOF analysis reselected DLG-Sol settings within each fold and gave the Ali model a matched tuning budget. DLG-Sol had lower RMSE in that comparison (0.8799 versus 0.9177), with the paired interval supporting the improvement (Supplementary Table S6).

DLG-Sol had lower RMSE than each of the three retrained neural models—PNNL GNN, the additive Bhattacharya–Roy MLP–GNN, and Consensus GNN—in all six evaluations. The paired 95% intervals supported these differences. Implementation and tuning details are given in Section 2.4 and Supplementary Table S13. Complete RMSE, MAE, and paired intervals, including the Bhattacharya–Roy interaction sensitivity, are reported in Supplementary Table S12.

Table 4 Absolute matched-row RMSE for DLG-Sol and retained external-comparator implementations across six evaluations. Bold marks the lowest point estimate. A dagger (†) on a DLG-Sol value indicates that each of the four DLG-Sol-minus-comparator differences in that column has a paired-bootstrap 99% confidence interval entirely below zero; unmarked values can still have individual contrasts whose intervals exclude zero. PNNL, Bhattacharya–Roy, and Consensus GNN are independently retrained adaptations. Their provenance and interpretation limits are documented in Supplementary Table S13. The authors' released Consensus predictions are reported separately in Table 3 and Supplementary Tables S12 and S20. Complete RMSE and MAE values and intervals are given in Supplementary Table S12 and the machine-readable Supplementary data.

Model and evidence class	AqSolDBc OOF	ComPlat OOF	TDC OOF	TDC test	JCheM test	ComPlat test
DLG-Sol (this work)	**0.807**†	**0.879**†	0.986	**1.002**†	**0.614**†	0.560
PNNL GNN (independently retrained) [37]	0.908	1.032	1.141	1.105	0.700	0.616
Ali XGB-125D (split-adapted, ComPlat same-split) [3]	0.846	0.927	**0.980**	1.037	0.658	**0.554**
Bhattacharya–Roy additive MLP–GNN (independently retrained) [38]	1.064	1.162	1.241	1.111	0.718	0.627
Consensus GNN (independently retrained official-code adaptation) [18]	1.154	1.331	1.685	1.800	0.810	0.824

3.4 Comparison of fixed-weight and molecule-specific blending

Neither molecule-specific gate improved consistently on the fixed blend across the primary evaluations (Supplementary Table S18). Both gates reduced RMSE on TDC OOF and increased it on the ComPlat supplied test, with paired 95% confidence intervals supporting these changes. Elsewhere, differences were inconclusive or favoured the fixed blend.

Both gates improved on the supplementary Biogen panel, reducing RMSE by 0.0130 and 0.0119. Biogen measures a different kinetic-solubility endpoint and is shown separately. The gate comparisons support the fixed coefficient as a practical choice for the primary evaluations, while the Biogen result motivates further testing when the endpoint changes. Gate-fitting procedures are described in Section 2.5 and Supplementary Information ("Molecule-specific weighting").

4 Discussion

Adding descriptor-model predictions improved a neural model that already used molecular-language and 3D graph embeddings. DLG-Sol had lower RMSE than either component in all six evaluations. Paired confidence intervals supported improvements over both components in each of the three OOF evaluations. Among DLG-Sol and its two component models, DLG-Sol had the lowest RMSE on the ComPlat, TDC, and JCheM test sets. On each test set, however, the additional gain over the better-performing of XGBoost and the neural model was small and statistically inconclusive. The benefit of combining the models was therefore clearer in the OOF evaluations.

The improvements were not confined to a particular molecular subgroup. DLG-Sol outperformed both components in most evaluation-by-subgroup comparisons. Across all six evaluations, however, DLG-Sol had higher RMSE in the higher-molecular-weight groups than in the lower-molecular-weight groups. Thus, although blending improved overall performance, predicting solubility remained more difficult for larger molecules. Which component had lower subgroup error varied across evaluations, and neither of the two evaluated blending schemes with molecule-specific weights consistently outperformed fixed-weight blending. These results support fixed-weight blending as a simple, practical option for combining the two component models evaluated here.

The first limitation is that the evaluations are not fully independent. Several evaluations share data sources, and some development data used for model design overlap with later evaluation sets. Labels for the molecules being evaluated were excluded from model fitting within each split. However, the full development process, including representation and model selection, was not repeated independently for every evaluation. The six evaluations therefore should not be interpreted as six fully independent validations.

The second limitation concerns what can be concluded from the model comparisons. Representation screening was used to select the neural component, rather than to isolate the contributions of language and 3D information using models of matched capacity. The observed gains therefore demonstrate the benefit of the selected model combination, rather than the independent contribution of each information source. Implementations and tuning budgets also differed across external models. These comparisons show performance under the evaluated conditions, but do not establish how the models would rank if each were fully optimized. Prospective evaluation on independent compounds not used in model development is needed to determine whether the observed gains extend to new molecules.

5 Conclusions

Combining descriptor predictions with those of a language–geometry neural model reduced point-estimate RMSE relative to either component in all six evaluations. The gains extended across molecular subgroups, although some supplied-test improvements were small and uncertain and larger molecules remained difficult to predict. These results support combining descriptor-model predictions with those of a language–3D neural model for aqueous-solubility prediction. Independent evaluation under clearly defined assay conditions is the next step in assessing the wider applicability of this combination.

Supplementary Information. PDF containing supplementary methods and analyses, Supplementary Tables S1–S20, and Supplementary Figure S1.

Supplementary Data. ZIP archive. CSV files containing the numerical results, confidence intervals, model configurations, and reproducibility records underlying the article and Supplementary Information.

Data Availability

The numerical results, paired confidence intervals, model configurations, and reproducibility records generated in this study are provided in the Supplementary Data ZIP archive and the DLG-Sol v1.0.5 release at https://github.com/Tuesdayand/DLG-Sol/releases/tag/v1.0.5.

The third-party datasets are available from their original sources. AqSolDBc and the OChem development data described by Llompart et al. [21] are deposited in Recherche Data Gouv (version 2.0, https://doi.org/10.57745/CZVZIA) [55]. The ComPlat training and Huuskonen test data used by Ali et al. [3] are available at https://github.com/ComPlat/water-solubility-prediction [56]. The TDC data [23] are available as `Solubility_AqSolDB` in the ADMET Benchmark Group at https://tdcommons.ai/benchmark/admet_group/overview/ [57]. The curated dataset and released Consensus GNN predictions of Ulrich et al. [18] are available at https://github.com/nadinulrich/log_Sw_prediction [58].

For the supplementary analyses, the Biogen solubility measurements of Fang et al. [19] are available in `ADME_public_set_3521.csv` at https://github.com/molecularinformatics/Computational-ADME [59]. The source datasets for the Ghanavati reconstructions [20] are available at https://github.com/amingh1995/Aqueous-Solubility-Prediction [60].

Source versions, input-file hashes, preprocessing instructions, and access or redistribution conditions are documented in the DLG-Sol repository. Third-party source rows are not redistributed in our release. For TDC, exact input reconstruction requires the recorded runtime-export files; a current download may differ slightly in numerical precision and is not guaranteed to match their hashes.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00354727). This work was also supported by the Hankuk University of Foreign Studies Research Fund (Of 2026).

Acknowledgements

The authors thank Hojun Kim for reviewing the manuscript for English grammar and usage.

References

[1] Amidon, G. L., Lennernäs, H., Shah, V. P. & Crison, J. R. A theoretical basis for a biopharmaceutic drug classification: The correlation of in vitro drug product dissolution and in vivo bioavailability. *Pharm. Res.* **12**, 413–420 (1995).

[2] Di, L., Fish, P. V. & Mano, T. Bridging solubility between drug discovery and development. *Drug Discov. Today* **17**, 486–495 (2012).

[3] Ali, M. *et al.* Advancing aqueous solubility prediction: A machine learning approach for organic compounds using a curated data set. *J. Chem. Inf. Model.* **65**, 8426–8434 (2025).

[4] Mac Fhionnlaoich, N. *et al.* A hybrid approach to aqueous solubility prediction using COSMO-RS and machine learning. *Chem. Eng. Res. Des.* **209**, 67–71 (2024).

[5] Ran, Y. & Yalkowsky, S. H. Prediction of drug solubility by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **41**, 354–357 (2001).

[6] Ali, J., Camilleri, P., Brown, M. B., Hutt, A. J. & Kirton, S. B. Revisiting the general solubility equation: In silico prediction of aqueous solubility incorporating the effect of topographical polar surface area. *J. Chem. Inf. Model.* **52**, 420–428 (2012).

[7] Klamt, A. Conductor-like screening model for real solvents: A new approach to the quantitative calculation of solvation phenomena. *J. Phys. Chem.* **99**, 2224–2235 (1995).

[8] Klamt, A., Eckert, F., Hornig, M., Beck, M. E. & Bürger, T. Prediction of aqueous solubility of drugs and pesticides with COSMO-RS. *J. Comput. Chem.* **23**, 275–281 (2002).

[9] Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: A molecular descriptor calculator. *J. Cheminform.* **10**, 4 (2018).

[10] Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

[11] Tang, B. *et al.* A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *J. Cheminform.* **12**, 15 (2020).

[12] Ahmad, W., Simon, E., Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa-2: Towards chemical foundation models. Preprint at https://arxiv.org/abs/2209.01712 (2022).

[13] Wigh, D. S., Goodman, J. M. & Lapkin, A. A. A review of molecular representation in the age of machine learning. *WIREs Comput. Mol. Sci.* **12**, e1603 (2022).

[14] Rollins, Z. A., Cheng, A. C. & Metwally, E. MolPROP: Molecular Property prediction with multimodal language and graph fusion. *J. Cheminform.* **16**, 56 (2024).

[15] Pan, B., Li, S., Hou, X. & Wang, X. XGNN: A chemometric dual-tower model for predicting aqueous solubility. *J. Mol. Graph. Model.* **145**, 109370 (2026).

[16] Ahmad, W. *et al.* Integrating molecular descriptors and fingerprints via meta-learning for aqueous solubility prediction. *Sci. Rep.* (2026). Article in press. https://doi.org/10.1038/s41598-026-65722-y.

[17] Sorkun, M. C., Khetan, A. & Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Sci. Data* **6**, 143 (2019).

[18] Ulrich, N., Voigt, K., Kudria, A., Böhme, A. & Ebert, R.-U. Prediction of the water solubility by a graph convolutional-based neural network on a highly curated dataset. *J. Cheminform.* **17**, 55 (2025).

[19] Fang, C. *et al.* Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. *J. Chem. Inf. Model.* **63**, 3263–3274 (2023).

[20] Ghanavati, M. A., Ahmadi, S. & Rohani, S. A machine learning approach for the prediction of aqueous solubility of pharmaceuticals: a comparative model and dataset analysis. *Digit. Discov.* **3**, 2085–2104 (2024).

[21] Llompart, P. *et al.* Will we ever be able to accurately predict solubility? *Sci. Data* **11**, 303 (2024).

[22] Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **40**, 773–777 (2000).

[23] Huang, K. *et al. Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery and development. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS 2021)* (2021).

[24] Chen, T. & Guestrin, C. *XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).

[25] Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. Preprint at https://arxiv.org/abs/2010.09885 (2020).

[26] Hendrycks, D. & Gimpel, K. Gaussian error linear units (GELUs). Preprint at https://arxiv.org/abs/1606.08415 (2016).

[27] Bjerrum, E. J. SMILES enumeration as data augmentation for neural network modeling of molecules. Preprint at https://arxiv.org/abs/1703.07076 (2017).

[28] Riniker, S. & Landrum, G. A. Better informed distance geometry: Using what we know to improve conformation generation. *J. Chem. Inf. Model.* **55**, 2562–2574 (2015).

[29] Wang, S., Witek, J., Landrum, G. A. & Riniker, S. Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *J. Chem. Inf. Model.* **60**, 2044–2058 (2020).

[30] Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17**, 490–519 (1996).

[31] Rappé, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A. & Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).

[32] Schütt, K. T. *et al. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. Advances in Neural Information Processing Systems 30* (2017).

[33] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. *Neural message passing for quantum chemistry. Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of *Proceedings of Machine Learning Research*, 1263–1272 (2017).

[34] Elfwing, S., Uchibe, E. & Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Netw.* **107**, 3–11 (2018).

[35] Nair, V. & Hinton, G. E. *Rectified linear units improve restricted Boltzmann machines. Proceedings of the 27th International Conference on Machine Learning*, 807–814 (2010).

[36] Loshchilov, I. & Hutter, F. *Decoupled weight decay regularization. International Conference on Learning Representations* (2019).

[37] Panapitiya, G. *et al.* Evaluation of deep learning architectures for aqueous solubility prediction. *ACS Omega* **7**, 15695–15710 (2022).

[38] Bhattacharya, S. & Roy, A. An additive MLP-GNN framework for characterizing chemical and structural contributions to aqueous solubility. Preprint at https://arxiv.org/abs/2607.02212 (2026).

[39] Jacobs, R. A., Jordan, M. I., Nowlan, S. J. & Hinton, G. E. Adaptive mixtures of local experts. *Neural Comput.* **3**, 79–87 (1991).

[40] Wu, J. *et al.* ALipSol: An attention-driven mixture-of-experts model for lipophilicity and solubility prediction. *J. Chem. Inf. Model.* **62**, 5975–5987 (2022).

[41] Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **7**, 1–26 (1979).

[42] Huuskonen, J., Rantanen, J. & Livingstone, D. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur. J. Med. Chem.* **35**, 1081–1088 (2000).

[43] Tetko, I. V., Tanchuk, V. Y., Kasheva, T. N. & Villa, A. E. P. Estimation of aqueous solubility of chemical compounds using E-State indices. *J. Chem. Inf. Comput. Sci.* **41**, 1488–1493 (2001).

[44] Yan, A. & Gasteiger, J. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J. Chem. Inf. Comput. Sci.* **43**, 429–434 (2003).

[45] Delaney, J. S. ESOL: Estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **44**, 1000–1005 (2004).

[46] 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.* **44**, 266–275 (2004).

[47] Schroeter, T. S. *et al.* Estimating the domain of applicability for machine learning QSAR models: A study on aqueous solubility of drug discovery molecules. *J. Comput.-Aided Mol. Des.* **21**, 485–498 (2007).

[48] Ali, J., Camilleri, P., Brown, M. B., Hutt, A. J. & Kirton, S. B. In silico prediction of aqueous solubility using simple QSPR models: the importance of phenol and phenol-like moieties. *J. Chem. Inf. Model.* **52**, 2950–2957 (2012).

[49] Daina, A., Michielin, O. & Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **7**, 42717 (2017).

[50] Bjerrum, E. J. & Sattarov, B. Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. *Biomolecules* **8**, 131 (2018).

[51] Sorkun, M. C., Koelman, J. M. V. A. & Er, S. Pushing the limits of solubility prediction via quality-oriented data selection. *iScience* **24**, 101961 (2021).

[52] US Environmental Protection Agency. EPI Suite: Estimation program interface. https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface. Accessed 22 September 2026.

[53] OCHEM. Online chemical modeling environment. https://ochem.eu. Accessed 22 September 2026.

[54] Advanced Chemistry Development, Inc. ACD/Percepta, 2020 release, GALAS. https://www.acdlabs.com (2020).

[55] Llompart, P. *et al.* Données de réplication pour : Towards the improvement of thermodynamic solubility prediction – a review. *Recherche Data Gouv*, version 2.0. https://doi.org/10.57745/CZVZIA (2023).

[56] ComPlat. water-solubility-prediction: dataset and code. *GitHub*. https://github.com/ComPlat/water-solubility-prediction/tree/d4feda24b4bcb9efee9953605ba56af0e52dba4d (2025).

[57] Therapeutics Data Commons. ADMET Benchmark Group: Solubility_AqSolDB dataset. https://tdcommons.ai/benchmark/admet_group/overview/. Accessed 22 September 2026.

[58] Ulrich, N., Voigt, K., Kudria, A., Böhme, A. & Ebert, R.-U. log_Sw_prediction: curated dataset, splits, and predictions. *GitHub*. https://github.com/nadinulrich/log_Sw_prediction/tree/3da09dc96acde6c5c51d784307559067864ca779 (2025).

[59] Biogen. Computational-ADME: public dataset and code. *GitHub*. https://github.com/molecularinformatics/Computational-ADME. Accessed 22 September 2026.

[60] Ghanavati, M. A., Ahmadi, S. & Rohani, S. Aqueous-Solubility-Prediction: datasets and code. *GitHub*. https://github.com/amingh1995/Aqueous-Solubility-Prediction (2024).

Author contributions

Hyunmo Goo conceptualized the study, developed the methodology and software, performed validation, formal analysis, investigation, data curation, and visualization, and wrote the original draft. Hyeongwoo Kong contributed to validation, provided resources, supervised and administered the project, acquired funding, and reviewed and edited the manuscript.

Additional Information

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DLGSolSupplementaryData.zip
- DLGSolSupplementaryInformation.pdf