Supplementary Information: 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;

Guide to the Supplementary Information

This document provides the methods and numerical detail supporting the component comparisons in the main manuscript. Tables S4–S5 give the exact primary effects and evaluation histories; Tables S12–S13 separate released-prediction comparisons from retrained implementations and historical context. Tables S18–S19 describe weighting sensitivity and model development, respectively. The analyses below explain how to interpret these different forms of evidence. Dataset and model definitions match those in the main text.

Developmental screening explains the choice of the neural component; the later exploratory grids assess alternative representations under a limited common budget. Neither is an independent, capacity-matched ablation of the final model. Supplementary Biogen and Ghanavati results are kept separate from the six primary evaluations.

Supplementary Tables (listed in the order they appear below):

- **Table S1:** dataset characteristics (Table S1).
- **Table S2:** molecular representations evaluated (Table S2).
- **Table S3:** 3D conformer-embedding failure counts (Table S3).
- **Table S4:** exact per-evaluation metrics and effects (Table S4).
- **Table S5:** provenance, overlap, and label-access qualifications for the main evaluations (Table S5).
- **Table S6:** ComPlat selection-and-tuning sensitivity results (Table S6).
- **Table S7:** Biogen and Ghanavati supplementary-panel metrics, effects, and evidence qualifications (Table S7).
- **Table S8:** final model hyperparameters (Table S8).
- **Table S9:** correction to the AqSolDBc descriptor pipeline and sensitivity analysis (Table S9).
- **Table S10:** software environments and random-seed records (Table S10).
- **Table S11:** blend coefficient α by evaluation and prediction unit (Table S11).
- **Table S12:** all 31 matched-row external-comparator comparisons, including six Bhattacharya–Roy interaction sensitivities (Table S12).
- **Table S13:** provenance and interpretation limits for external-comparator evidence, candidate-comparator audits, and contextual benchmarks (Table S13).

- **Table S14:** post-development exploratory-grid predictor search spaces (Table S14).
- **Table S15:** best single-representation result in each post-development exploratory screen (Table S15); complete grids are machine-readable Supplementary data.
- **Table S16:** post-development exploratory representation-pair screen, best-per-panel summary (Table S16); full 105-combination results are machine-readable Supplementary data.
- **Table S17:** the four evaluation–subgroup rows with fewer than 100 molecules in main-text Figure 2 (Table S17); all 96 DLG-Sol-versus-component chemical-subgroup effects, paired-bootstrap intervals, and sample sizes are provided as machine-readable Supplementary data.
- **Table S18:** fixed blend versus K=5 cluster-soft and PCA-linear gating, per-evaluation results supporting the main-text molecule-specific weighting results (Table S18).
- **Table S19:** developmental representation screening and neural-component selection summary (Table S19).
- **Table S20:** source-paper ComPlat and JCheM comparison contexts with RMSE and available MAE values (Table S20).

Supplementary Figures:

- **Supplementary Figure S1:** descriptor–neural residual correlations and large-error overlap across the six primary evaluations (Figure S1).

Supplementary methods and analyses (referenced by title):

- **TDC data partitions**.
- **ComPlat selection sensitivity**.
- **Biogen endpoint-shift evaluation**.
- **Retrospective Ghanavati panels**.
- **Descriptor preprocessing**.
- **3D-MPNN specification**.
- **External comparator assessment**.
- **Subgroup and residual analyses**.
- **Molecule-specific weighting**.
- **Developmental representation screening**.

Machine-readable Supplementary data files

The following UTF-8 comma-separated value files are supplied in `Additional_file_2.zip`. Content-based file names are used so that changes in Supplementary Table numbering do not alter the file identities.

- `main_evaluation_effects_12_rows.csv`: the 12 DLG-Sol-versus-component contrasts underlying main-text Table 2 and Supplementary Table S4, including RMSE, MAE, and 95% and 99% paired-bootstrap RMSE confidence intervals.
- `external_comparator_metrics_31_rows.csv`: all 31 rows in Supplementary Table S12, comprising the 24 implementation comparisons in main-text Table 4, the released Consensus GNN comparison described in the main-text external-model results, and six Bhattacharya–Roy interaction sensitivities. Values include RMSE, MAE, and 95% and 99% paired-bootstrap RMSE confidence intervals.
- `source_paper_benchmark_context.csv`: the historical point estimates and DLG-Sol rows reported in Supplementary Table S20, with metric availability and comparison-scope qualifications.
- `molecule_specific_weighting_14_rows.csv`: the 14 gate-versus-fixed-blend comparisons in Supplementary Table S18, including fixed and gate RMSE, paired differences, 95% confidence intervals, and bootstrap settings.
- `single_representation_grid_185_rows.csv`: the complete available single-representation grid summarized in Supplementary Table S15.
- `pair_representation_grid_315_rows.csv`: the complete 105-combination grid for each of the three OOF evaluations summarized in Supplementary Table S16.

- `chemical_subgroup_delta_rmse_96_rows.csv`: all 96 DLG-Sol-versus-component subgroup RMSE contrasts underlying main-text Figure 2 and Supplementary Table S17, with 95% paired-bootstrap confidence intervals and sample sizes.
- `chemical_subgroup_delta_mae_96_rows.csv`: the corresponding exploratory subgroup MAE contrasts.
- `seed_registry.csv`: a compact index of the principal model-fitting and bootstrap seed rules summarized in Supplementary Table S10. Complete unit-level rules and manifests are in the code archive.
- `descriptor_preprocessing_record.csv`: dataset-specific descriptor filtering, search-budget, and retained-feature records supporting the section "Descriptor preprocessing" and Tables S8–S9.
- `molprop_source_reproduction_summary.csv`: the source-domain MolPROP audit summarized in the section "External comparator assessment" and Supplementary Table S13. These audit results were not used as manuscript performance comparisons.
- `developmental_single_representation_screen_metrics.csv`: aggregate metrics from the seven-representation, five-predictor architecture-development screen summarized in the section "Developmental representation screening" and Supplementary Table S19.
- `developmental_pair_representation_screen_metrics.csv`: aggregate metrics from the 21-pair, five-predictor architecture-development screen summarized in the section "Developmental representation screening" and Supplementary Table S19.
- `developmental_focused_tuning_metrics.csv`: aggregate metrics from the focused tuning of developmentally prioritized representation–predictor combinations summarized in the section "Developmental representation screening" and Supplementary Table S19.

TDC data partitions

We retained the supplied scaffold train–test partition of the TDC ADMET Benchmark Group, comprising 7,985 training/validation rows and 1,997 test rows. Within the training/validation set, we created five OOF folds stratified by ten logS quantile bins, using seed 260722. Within each outer-training partition, 12.5% of the rows were reserved for validation using the same target bins and a seed equal to 260722 plus the zero-based fold index. These internal OOF and validation partitions were target-stratified, not scaffold splits. The supplied test rows were excluded from every internal partition used for model fitting or selection.

ComPlat selection sensitivity

The primary ComPlat OOF comparison combines two procedural differences that could favour DLG-Sol. Some DLG-Sol settings were selected using one global training/validation division before OOF fitting, whereas Ali XGB-125D retained its reported descriptor schema and XGBoost settings without a dataset-specific search budget comparable to the DLG-Sol descriptor branch. We therefore prespecified a 2×2 sensitivity analysis crossing the primary or fold-local DLG-Sol prediction with the reported-setting or investigator-tuned Ali XGB-125D prediction. The primary predictions, main-text Table 4 estimates, and inferential claims were locked before this analysis and were not replaced.

For fold-local DLG-Sol, each outer fold reselected the descriptor configuration from 36 candidates and selected the ChemBERTa and 3D-MPNN epoch from epochs 1–27 using only that outer fold's training population. The ChemBERTa *aug2* representation denotes fine-tuning with the canonical SMILES plus two independently randomized SMILES per training molecule. This representation, the ChemBERTa and 3D-MPNN architectures, fusion-head architecture, fusion-head seeds, and fusion-head training duration remained fixed under the prespecified development policy. Thus, this was fold-local reselection under the fixed architecture, not a fully nested redevelopment of all modelling decisions. For each held-out fold, α was estimated from the other four folds' cross-fitted component predictions and labels. The selected descriptor candidate, ChemBERTa epoch, 3D-MPNN epoch, Ali candidate, and α were 24/15/24/3/0.389505 in fold 0, 35/14/25/3/0.400621 in fold 1, 10/26/23/30/0.388266 in fold 2, 12/21/27/3/0.392920 in fold 3, and 35/18/27/30/0.376396 in fold 4.

For the tuned Ali sensitivity, every outer fold evaluated the same 36-candidate random-search budget used for the DLG-Sol descriptor branch plus the reported Ali setting as a 37th anchor. Selection

used training-side validation RMSE only. The reported anchor was not selected in any fold. The resulting model is an investigator-tuned sensitivity variant and is not presented as the model reported by Ali et al. Table S6 preserves this distinction.

The two procedural effects were detectable at the affected component level. Fold-local reselection increased the DLG-Sol descriptor RMSE by 0.006967 (95% CI [+0.002242, +0.011575]), while the change in the final DLG-Sol RMSE was 0.000654 (95% CI [−0.003622, +0.004792]) and was inconclusive. Matched-budget tuning reduced Ali RMSE by 0.009592 (95% CI [−0.013741, −0.005471]). In the symmetric comparison that allowed both corrections, fold-local DLG-Sol had RMSE 0.879868 and tuned Ali XGB-125D had RMSE 0.917690, giving a difference of −0.037822 (95% CI [−0.045872, −0.029676]). The paired 95% confidence intervals for all four prespecified external-comparator contrasts on ComPlat OOF continued to support lower RMSE for DLG-Sol when fold-local predictions replaced the primary predictions in this sensitivity analysis.

The epoch search had a fixed upper boundary of 27. The 3D-MPNN selected epoch 27 in folds 3 and 4, and ChemBERTa selected epoch 26 in fold 2. Performance beyond epoch 27 was not observed, so this analysis does not establish that its conclusions are invariant to a wider epoch search. We report these boundary selections without extending the search after inspecting the results.

On the ComPlat supplied test, the five fold-specific tuned-Ali refits were averaged before scoring. Tuned Ali had RMSE 0.557923, compared with 0.554278 for the reported setting. The fold-local OOF component predictions yielded an alternative training-side coefficient of $\alpha = 0.389547$, whereas the primary supplied-test coefficient was 0.352457. Both the primary DLG-Sol comparison with tuned Ali and the alternative-alpha comparison were inconclusive (Table S6). These results show that the tuning benefit observed in ComPlat OOF did not transfer to the supplied test. All confidence intervals used 10,000 paired molecular-row bootstrap resamples with seed 20260904.

Biogen endpoint-shift evaluation

The fixed blend was scored on 2,153 Biogen public ADME compounds [1] with zero exact canonical-SMILES overlap against the audited scored sources. The fixed blend and XGB gave nearly identical RMSE, and their paired comparison was inconclusive. The fixed blend performed better than the neural model. RMSE and paired confidence intervals are reported in Table S7. Because this panel measures pH 6.8 CLND kinetic solubility, it was not pooled with the primary water-solubility evaluations. An exploratory analysis of five additional neural seed sets repeated the qualitative pattern but is not independent confirmation and did not change the primary model.

Retrospective Ghanavati panels

In a retrospective five-panel re-evaluation reconstructed from Ghanavati et al. [2], the fixed blend had lower point-estimate RMSE than both component models in all five panels. All five fixed-versus-XGB 95% confidence intervals excluded zero. The fixed-versus-neural interval excluded zero for All-Data but included zero for ESOL, AQUA, PHYS, and OCHEM. Panel-level estimates are reported in Table S7. The labels had been accessed previously. ESOL, AQUA, PHYS, and OCHEM are source-matched reconstructions, and OCHEM has a known random-row structural-overlap limitation. These results assess retrospective robustness but are not independent confirmation.

Descriptor preprocessing

Descriptor filtering was learned from training data and then applied to the rows being predicted. The procedures below explain the benchmark-specific choices behind main-text Figure 1A and Supplementary Tables S8–S9.

Each pipeline began with 1,613 raw two-dimensional Mordred descriptor columns. All descriptor filters were fitted on the corresponding training partition; scored rows did not determine descriptor retention or XGBoost selection. The preprocessing rules differed among the benchmark-specific pipelines, so we did not impose a uniform procedure retrospectively. In every pipeline, descriptor values were coerced to numeric form, columns containing any missing training value were removed, constant columns were removed, and columns with training variance at or below 10^{-12} were removed.

For AqSolDBc, the same filters were repeated inside every inner- and outer-training partition, followed by an absolute Pearson-correlation threshold of 0.95. The upper triangle of the correlation matrix was traversed in the original Mordred column order. Under this deterministic rule, the descriptor appearing later in the original column order was discarded from each highly correlated pair. Forty pre-specified XGBoost configurations were compared by pooled inner-OOF RMSE, and five seeded outer refits were averaged. The five outer folds retained 373, 353, 354, 352, and 356 descriptors.

For TDC, the fitting-partition filter used the same missingness, constant, and variance rules followed by an absolute-correlation threshold of 0.98, again dropping the later column in the original order. Twenty-four XGBoost candidates were selected on a training-only validation partition within each outer fold. The five folds retained 337, 341, 337, 340, and 338 descriptors.

For JCheM, each of the five supplied Train/Validation splits used the same rules and 0.98 threshold; 16 bounded-tree candidates were selected using that split's Validation rows, and the shared Test rows were not used for filtering or selection. The five fits retained 571, 575, 572, 574, and 615 descriptors.

For ComPlat, missingness, constant, and variance filtering were fitted on the training rows, but pairwise-correlation pruning was not applied because it was not part of the retained tree-model pipeline. Thirty-six candidate configurations were selected on the supplied training-side validation design. The five OOF fits retained 646, 650, 654, 650, and 650 descriptors, and the supplied-test refit retained 650. These retained-column counts describe the final benchmark-specific descriptor models, not the separate representation-grid baselines.

3D-MPNN specification

This section specifies the graph inputs and message-passing operations used to produce the 3D embeddings in main-text Figure 1. Generated conformers supply fixed distances; the network does not update coordinates.

Each node used 42 features. These comprised a 17-way atomic-number encoding (H, B, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Zn, Br, I, or other), a 7-way total-degree encoding (0–5 or other), a 6-way hybridization encoding (sp, sp^2, sp^3, sp^3d, sp^3d^2, or other), and 12 scalar features: formal charge, Gasteiger charge, total hydrogen count, aromatic and ring-membership flags, atomic mass divided by 200, and donor, acceptor, carboxyl, sulfonyl, phosphoryl, and quaternary-nitrogen flags. Each edge used 24 features: a bond-present flag, single-, double-, triple-, and aromatic-bond flags, conjugation and ring-membership flags, distance divided by 5 Å, and 16 Gaussian radial-basis values. The seven bond-related values were zero for noncovalent proximity edges.

The 19 molecule-level auxiliary features were formal charge of the original structure, formal charge of the retained fragment, absolute original charge, counts of positively and negatively charged atoms, topological polar surface area divided by 250, calculated logP divided by 10, hydrogen-bond donor count divided by 10, hydrogen-bond acceptor count divided by 20, rotatable-bond count divided by 30, ring count divided by 10, molecular weight divided by 1000, fragment count, largest-fragment heavy-atom ratio, counterion and conformer-success flags, and the minimum and mean conformer energies divided by 500 and their standard deviation divided by 100.

For an interatomic distance d, the 16 radial-basis values were

$$r_k(d) = \exp[-(d - c_k)^2/0.25],$$

where the centres c_k were evenly spaced from 0 to 5 Å. This encoding follows the general Gaussian radial-basis treatment of interatomic distances used in SchNet [3]; the 16 centres and width above are specific to this study. At message-passing layer ℓ, the message from sender j to receiver i, its sum at the receiver, and the residual receiver update were

$$m_{ij}^{(\ell)} = \phi_m\Big([h_i^{(\ell)} \,\|\, h_j^{(\ell)} \,\|\, e_{ij}]\Big), \qquad m_i^{(\ell)} = \sum_{j \in \mathcal{N}(i)} m_{ij}^{(\ell)},$$

$$h_i^{(\ell+1)} = \mathrm{SiLU}\Big(\mathrm{LayerNorm}\Big[h_i^{(\ell)} + \phi_u\Big([h_i^{(\ell)} \,\|\, m_i^{(\ell)}]\Big)\Big]\Big).$$

The message network mapped $2d + d_e$ inputs through a $2d$ hidden layer to d outputs, and the update network mapped $2d$ inputs through a $2d$ hidden layer to d outputs. Dropout was applied in both networks. Training-side validation selected two to four message-passing layers, hidden widths from 96 to 256, and whether to append the 19 auxiliary molecular features listed in Table S8.

External comparator assessment

External-comparator evidence was classified according to the provenance supported by the publicly available artefacts: authors' released predictions, same-split reproduction, schema-and-setting adaptation, independently retrained adaptation, or a historical point estimate. Authors' released row-level predictions were used whenever they were available on the scored rows. Otherwise, we assigned the evidence class according to the checkpoints, split definitions, training records, and code that could be recovered. When the public artefacts did not permit exact source-model reproduction, the resulting comparison was labelled as an adaptation and interpreted as the performance of that evaluated implementation. Retrained comparators used the same train/test partitions and the same full set of scored molecular rows as DLG-Sol, and split-specific preprocessing and fitting excluded the corresponding scored labels. RMSE differences used the paired 10,000-resample molecular-row bootstrap defined in the statistical-analysis subsection of the main manuscript. The same locked row-level predictions were resampled to obtain both 95% and 99% percentile intervals. The 25 primary external-model comparisons comprise 24 retrained-model contrasts across six evaluations and one released-prediction contrast; Table S12 additionally reports the six prespecified Bhattacharya–Roy interaction sensitivities. MolPROP was examined as a candidate comparator but excluded because the released code and artefacts did not allow us to link the available training and evaluation workflow to the reported source result. Implementation sources, environments, target-evaluation procedures, and permitted interpretations are summarized in Table S13. Prediction summaries and complete 95% and 99% intervals for the retained comparisons are supplied as machine-readable Supplementary data.

Independent reproduction and prospective generalization are related but distinct issues. Wigh et al. discuss reproducibility difficulties in emerging modelling methods, drawing on examples from generative modelling and reinforcement learning [4]. In aqueous-solubility modelling, Ali et al. reproduced a published model when potentially overlapping rows were retained but obtained materially weaker performance after molecular standardization exposed and removed 133 potential train–test duplicates [5]. This is evidence of sensitivity to preprocessing and overlap definitions, not failure to reproduce an unchanged protocol. Separately, Llompart et al. reported weak prospective performance of recently published solubility models on a newly curated dataset [6]; that result concerns external generalization rather than code-level reproduction.

XGNN — no public executable implementation. The XGNN article documented model configurations and training procedures in its Supplementary Materials [7], but we found no public executable implementation, checkpoint, or row-level predictions with which to conduct a code-faithful split-matched reconstruction. XGNN was therefore discussed as related prior work but not included in the evaluated comparator set.

PNNL GNN — independently retrained architecture adaptation. We first examined whether the reported source population and split could be identified from the method, repository, and archived data. The original study reported a 17,149-molecule population, whereas the available public SOMAS versions contained 11,692–11,696 rows and did not identify the original random split. The paper reported GNN RMSE 1.0722 on its fixed test set and 1.1156 ± 0.0378 in fivefold cross-validation, but the population and split mismatch precluded a like-for-like numerical reproduction test. The architecture and preprocessing could nevertheless be reconstructed and retrained on each benchmark-defined target split. We therefore treated the retrained models as independently trained architecture adaptations. They were not considered replications of the original study. The source result was classified as `NOT_IDENTIFIABLE` because the available data did not permit a like-for-like numerical reproduction test.

Ali XGB-125D — same-split reproduction on ComPlat and setting adaptations elsewhere. We retained the reported 125-descriptor schema and XGBoost settings and refitted the model within each benchmark-defined split. The ComPlat repository train and test rows were byte-identical to the

ComPlat files used in this study, so those results constitute a same-split reproduction of the released tabular procedure. They are not independent evidence because they reuse the authors' public rows and design. AqSolDBc, TDC, and JCheM are schema-and-setting adaptations because the original model was not reported for those datasets. Additional dataset-specific tuning comparable in breadth to DLG-Sol would have produced a new Ali-derived model instead of an evaluation of the reported settings. The results describe the fixed reported setting under each target protocol, not the maximum performance attainable after retuning.

Bhattacharya–Roy MLP–GNN — source value not reproduced. We obtained the public implementation and split and first identified the source paper's additive model as the primary architecture: its graph and descriptor outputs are summed without the optional multiplicative interaction term. We independently retrained that additive model using the public source split and retained the interaction-augmented implementation only as a prespecified sensitivity. The audit also examined a disjoint clean-split sensitivity to separate the public split's overlap issue from model fitting. For the validation-best source track, the reported versus observed mean/median MAE values were 1.06/1.07 versus 1.1342/1.1292 for the additive model and 1.10/1.08 versus 1.1573/1.1626 for the interaction-augmented model. The additive discrepancies were 7.0% for the mean and 5.5% for the median. On the disjoint clean-split sensitivity, validation-best mean/median MAE was 1.2590/1.2587 for the additive model and 1.2631/1.2719 for the interaction-augmented model; the corresponding mean/median RMSE values were 1.8568/1.7984 and 1.7978/1.8125. Under the prespecified criteria, source performance was `NOT_REPRODUCED` for either variant. Missing original transfer-learning checkpoints and complete training provenance prevented attribution of the residual difference to one implementation step. The six main target results therefore characterize independently retrained adaptations of the additive architecture. The interaction-augmented results remain a separate sensitivity in Tables S12 and S13.

Consensus GNN — released predictions recovered, official-code retraining not treated as exact reproduction. On JCheM, the authors' released predictions were aligned directly to the same 980 rows and reproduced the reported RMSE of approximately 0.657. This is a direct released-prediction comparison and requires no reimplementation. In a separate evidence track, we independently retrained the public GraphConv architecture across all six evaluations using five seeded members per prediction unit, training-only SMILES augmentation, and an unweighted prediction mean. The adaptation retained the reported 64- and 128-unit graph-convolution layers, 256-unit dense layer, leaky-ReLU activation, batch normalization, dropout 0.1, L1 loss, and 130-epoch schedule. OOF members used only their outer-training partitions, and supplied-test members used the associated development populations. Exact augmentation commands, original random states, and the authors' complete checkpoint-selection procedure were unavailable. We therefore treated this model as an independently retrained official-code adaptation, not an exact reproduction. It used a locked Python 3.11.8 environment with TensorFlow/Keras 2.15.0, RDKit 2023.09.6, and DeepChem 2.7.2.dev20240326195015.

MolPROP — reproduction audit only. We evaluated the reported GATv2–ChemBERTa-2-77M-MLM architecture as an independently retrained official-code adaptation [8]. We first reconstructed the authors' ESOL scaffold split: the public train and holdout files contained 1,015 and 113 molecules, respectively, matched an independent DeepChem scaffold-split reconstruction, had no SMILES or scaffold overlap, and differed in labels only at floating-point precision. We then ran two independent 50-trial BOHB reconstructions. The first used the hyperparameter ranges reported in the paper, and the second used the different ranges implemented in the public graph-fusion code. Their mean ten-model holdout RMSEs were 1.0421 and 1.0067, respectively, compared with the published value of 0.777 ± 0.02, which the paper labels as RMSE. Extending the selected paper-space setting to 50 epochs did not improve the holdout result, so neither the alternative public-code search range nor the shorter HyperBand checkpoint explained the discrepancy.

Exact attribution was not possible because the public repository did not contain the manuscript ESOL checkpoints, selected ESOL hyperparameter record, Ray trial logs, or manuscript predictions needed to recover the reported run. Two additional code-level discrepancies were preserved and documented rather than silently repaired. First, the public holdout path averages one-molecule square-root MSE values at batch size one, which is mathematically MAE rather than dataset-level RMSE. The

published value remains described here as "reported as RMSE" because no released output links it to that evaluation script. Second, the public GATv2 optimizer construction consumes a shared parameter iterator while creating the graph group before testing membership for the ChemBERTa group; the latter group is consequently empty in the literal execution path even when the configuration states `freeze=false`. In our hash-pinned reconstruction, the two optimizer groups contained 22 and 0 parameter tensors and covered 3,023,873 parameters, leaving all 3,427,440 ChemBERTa parameters outside the optimizer. The omitted language parameters constituted 53% of the model's 6,451,313 parameters, so the language branch acted as a fixed feature extractor rather than a fine-tuned branch. The language learning rate and the number of frozen language layers—two of the five hyperparameters in the reported search space—therefore had no effect on the fitted parameters; all six examined combinations of those two settings produced optimizer groups with identical parameter counts and shapes.

Because learned language–graph fusion is a defining property of the reported MolPROP architecture, the released execution path did not train the language branch as described for MolPROP. We could not establish that this path implemented the reported language–graph fusion model. We do not know whether the published result used the public path or a corrected internal path. We therefore do not reinterpret the published value, assign its discrepancy to a single cause, or report the completed target-adaptation scores as a comparison with MolPROP. We retained the target-run artefacts for audit, but did not include their scores in the manuscript comparisons.

Across comparators, some adaptations retained reported hyperparameters and did not receive dataset-specific optimization comparable to DLG-Sol. Row-level bootstrap intervals quantify sampling uncertainty on the scored molecules but do not include retraining-seed uncertainty. The results apply to the implementations and adaptation procedures evaluated here, not to the best possible performance of each external model.

To make this study's procedures auditable and rerunnable, the public v1.0.5 reproducibility package records the source code, selected configurations, branch-specific software environments, random seeds, metric implementations, input-reconstruction recipes, and row-role and label-access validation checks. It also supplies machine-readable aggregate results. Raw third-party molecular rows, historical checkpoints, embeddings, predictions, and model weights are not redistributed. The release therefore supports protocol-level reconstruction and fresh training from lawfully obtained inputs rather than bitwise recovery of the historical fitted weights.

Subgroup and residual analyses

This analysis asks where the blend improves on its components and how their errors differ. Main-text Figure 2 reports relative RMSE changes; the accompanying CSV also reports absolute RMSEs, so improvement can be distinguished from remaining error.

We compared DLG-Sol directly with each component along seven structural-feature axes defined before inspecting the subgroup results. Molecular weight was divided into < 250, 250–< 400, and ≥ 400 Da; rotatable-bond count into 0, 1–4, and ≥ 5; and hydrogen-bond donor, hydrogen-bond acceptor, aromatic-ring, nitrogen-containing heterocycle, and carboxyl-group status into absent and present levels. Molecular weight and rotatable bonds were calculated with RDKit. Donor and acceptor status required at least one Lipinski donor or acceptor. Aromatic-ring status required a ring in which every atom was aromatic, nitrogen-containing heterocycle status required at least one ring nitrogen, and the carboxyl definition used the SMARTS pattern `[CX3](=O)[O;H1,-1]`.

The seven axes were analysed separately rather than crossed into a Cartesian partition, yielding 16 potentially overlapping subgroups.

For each subgroup in each evaluation, ΔRMSE was calculated as DLG-Sol minus component RMSE, so negative values favour DLG-Sol. Six evaluations therefore yielded 96 evaluation-by-subgroup rows, each with one XGB and one neural contrast. Confidence intervals used the paired bootstrap defined in the statistical-analysis subsection of the main manuscript. All rows were retained. The four evaluation–subgroup rows with fewer than 100 molecules are identified separately in Table S17. No multiplicity correction was applied.

For each component and evaluation, the residual was calculated as the prediction minus the measured log solubility. Paired component residuals were summarized using Pearson and Spearman

correlations. A molecule was assigned to a component's large-error subset when its absolute residual exceeded the sample standard deviation of that component's residuals within the evaluation. Supplementary Figure S1B reports descriptor-only, shared, and neural-only counts as fractions of all molecules in each evaluation. For a complementary summary, these three mutually exclusive categories were normalized within their union for each evaluation and then averaged with equal weight across the six evaluations. This gave 26.8% descriptor-only, 45.6% shared, and 27.6% neural-only membership. The two component-specific categories therefore totalled 54.4%. The overlap is a descriptive threshold summary and does not, by itself, demonstrate dependence beyond that represented by the residual correlations.

To examine whether either component had a stable advantage within a chemical subgroup, we compared the component RMSEs across the six evaluations. None of the 16 subgroup definitions had the same lower-RMSE component in all six. We then calculated the descriptor-minus-neural RMSE gap for each of the 96 evaluation–subgroup rows. The decomposition below concerns variation in these component performance gaps, not variation in individual molecular errors or measured solubility.

In a balanced, unweighted sum-of-squares decomposition of the gaps, evaluation identity accounted for 81.9% of the total variation, subgroup identity for 3.0%, and the remaining evaluation-by-subgroup variation for 15.1%. Each row received equal weight. These percentages describe point estimates, and no confidence intervals were calculated for them. The groups can overlap, and the subgroup analyses were not adjusted for multiple comparisons. The observed rankings do not prove that chemical specialization is absent or identify a causal representation mechanism.

Molecule-specific weighting

Table S18 reports the paired gate-versus-fixed-blend RMSE effects, sample sizes, and fixed-blend RMSE, and `molecule_specific_weighting_14_rows.csv` supplies the full-precision aggregate values. The main-text Methods subsection on molecule-specific weighting describes the fitting populations. The construction details below specify the two exploratory gates used in that comparison.

Both gates used ten structure-only RDKit variables: molecular weight, calculated logP, topological polar surface area, hydrogen-bond donor and acceptor counts, ring count, rotatable-bond count, fraction of sp^3 carbons, heavy-atom count, and formal charge. The variables were standardized on the fitting rows and projected by principal component analysis fitted on those same rows. For the cluster-soft gate, $K = 5$ and an eight-dimensional PCA representation were fixed before the comparison. K-means clusters were fitted on the development rows, and each cluster received the mean of the per-molecule convex coefficients that minimized squared error, shrunk toward the global mean with a pseudo-count of 200. For the PCA-linear gate, ridge regression mapped PCA coordinates to per-molecule convex coefficients that minimized squared error, and predictions were clipped to $[0, 1]$. Nested cross-validation on AqSolDBc development OOF selected eight principal components and a ridge penalty of 100 from prespecified candidate grids. These hyperparameters were transferred unchanged. For OOF evaluations, all transformations and coefficients were refitted without the held-out fold. Supplied-test gates were fitted only on the associated development predictions: ComPlat and TDC used their own OOF development rows, whereas JCheM and Biogen used AqSolDBc development OOF. Scored supplied-test labels were used only for final evaluation. These gates are project-specific tests of added routing complexity, not implementations of either cited mixture-of-experts architecture.

Across the six primary evaluations, the gates produced gains in TDC OOF but losses or inconclusive differences elsewhere, including losses for both gates on the ComPlat supplied test with paired 95% confidence intervals excluding zero. Both gates improved on the supplementary Biogen panel, with RMSE differences from the fixed blend of -0.0130 and -0.0119 (1.15% and 1.05% of the fixed-blend RMSE). The corresponding ComPlat supplied-test losses were 1.73% and 1.59%. Because Biogen had no gate-fitting partition and therefore used AqSolDBc development predictions, we cannot determine whether this isolated improvement reflects the endpoint shift or the use of gate parameters fitted on a different dataset. It is treated as hypothesis-generating rather than evidence of a generally superior routing rule. Within each evaluation, the gates used structure-only variables. The component rankings in the subgroup analysis changed across evaluations. This observation and the gate results do not exclude benefits from other weighting methods.

Developmental representation screening

The language–geometry component was selected through a staged architecture-development screen conducted before the final six-evaluation pipelines were fixed. The screen used 6,437 AqSolDBc training rows, 1,610 AqSolDBc validation rows, and the 2,251-row `OChemUnseen` file distributed with the curated AqSolDB source used by Llompart et al. [6]. The latter is called the *development panel* here to distinguish its role from independent validation. It is also distinct from the smaller Ghanavati `OCHEM` reconstruction described in the section "Retrospective Ghanavati panels".

The first stage evaluated seven individual representations—Mordred, ECFP, ChemBERTa, the 3D-MPNN, atom-level 2D GNN, BRICS-subgroup GNN, and Murcko-subgroup GNN—with five downstream predictor families. The second stage evaluated all 21 unordered representation pairs with the same five predictor families. Preprocessing was fitted on the AqSolDBc training rows. Within each representation–predictor combination, the hyperparameter trial was selected by AqSolDBc validation RMSE. Candidate families for focused tuning were then prioritized partly from their observed development-panel rankings, including the overall and large-error-subset rankings. The development-panel labels therefore influenced architecture-family selection even though they did not select the hyperparameter trial within a combination.

Mordred with XGBoost was the strongest single-representation configuration by development-panel RMSE (1.2755; 1.276 after rounding), with corresponding validation RMSE 0.8386. ChemBERTa and the 3D-MPNN concatenated and fitted with an MLP achieved the lowest development-panel RMSE among the initial single- and pair-representation candidates (1.2654; 1.265 after rounding), with validation RMSE 0.8124. Focused tuning of prioritized pair–predictor families reduced the ChemBERTa–3D-MPNN–MLP development-panel RMSE to 1.2492 (1.249 after rounding), with validation RMSE 0.8038. These results motivated the subsequent development of the final language–geometry neural component.

These values document architecture selection rather than independent evidence for final-model performance. The development-panel rankings were inspected before focused tuning. The resulting values are therefore not independent validation, and the screen was not capacity-matched to the subsequently optimized DLG-Sol components. A retrospective overlap audit found that 680 of the 2,251 development-panel rows shared canonical SMILES with at least one final primary evaluation; the per-evaluation counts were 33 for AqSolDBc OOF, 529 for ComPlat OOF, 214 for TDC OOF, 69 for the TDC supplied test, 44 for the JCheM supplied test, and 65 for the ComPlat supplied test. These per-evaluation counts overlap and therefore do not sum to the union count. At the InChIKey level, the union comprised 800 of the 2,219 distinct InChIKeys in the development panel. This overlap further precludes interpreting the panel as independent validation.

The later grids in Tables S14–S16 were rerun after model development under the final representation policy and a common low-budget predictor protocol. They provide descriptive breadth across alternative representations but are not the original selection experiment and do not estimate the maximum attainable performance of those representations. Aggregate configuration-level results from all three developmental stages are supplied as machine-readable Supplementary data.

Supplementary Table S1 Dataset distributions of measured logS, molecular weight (MW; g/mol), and ring count. Molecular weight and ring count were calculated from canonical SMILES using RDKit.

Panel	n	logS mean	logS SD	MW mean (median)	MW SD	Rings mean (median)
AqSolDBc development OOF	8,047	−2.85	2.31	251.7 (225.2)	139.4	1.54 (1)
ComPlat OOF	17,937	−3.01	2.20	280.1 (260.3)	127.5	1.82 (2)
TDC OOF	7,985	−2.76	2.37	255.9 (218.3)	171.3	1.30 (1)
TDC official Test	1,997	−3.40	2.30	309.8 (273.3)	223.6	2.36 (2)
JCheM official Test	980	−2.97	2.04	245.8 (222.2)	122.3	1.54 (1)
ComPlat official Test	1,282	−2.74	2.04	199.2 (178.2)	94.3	1.37 (1)

Supplementary Table S2 Molecular representations evaluated in this study. ChemBERTa and the explicit-polar-H, distance-aware 3D-MPNN form the neural component of the fixed blend. Mordred descriptors form the descriptor component. The remaining four representations are used only in the single- and pair-representation grids.

Representation	**Definition**	**Generation procedure**
Mordred descriptors	Physicochemical and topological indices, including atom counts, connectivity, and charge-distribution terms. Filtered subsets, rather than one fixed full schema, were used.	Canonical SMILES → RDKit molecular object → Mordred calculation → fitting-partition column filtering.
ECFP4-like (Morgan fingerprint)	A 2,048-bit binary vector encoding the presence of circular substructures around each atom.	RDKit Morgan fingerprint with radius 2; standard atom invariants; feature invariants and chirality not explicitly enabled.
ChemBERTa (SMILES)	A 768-dimensional embedding from the pretrained ChemBERTa model named in Methods, fine-tuned end to end and mean-pooled over non-padding tokens.	Maximum length 256; canonical plus two randomized training SMILES; canonical SMILES at inference; learning rate 2×10^{-5}.
Explicit-polar-H distance-aware 3D-MPNN	Message-passing representation of 3D conformers using scalar atom states and distance-expanded edge attributes. Polar hydrogens bonded to N or O are explicit nodes.	ETKDGv3 generates up to five conformers; MMFF or UFF optimization; up to three lowest-energy conformers retained; covalent and 4.5 Å proximity edges; conformer embeddings averaged by molecule. Failed embeddings use the label-free fallback in Table S3.
Atom-level 2D GNN	Message-passing graph neural network embedding over the full 2D molecular graph (atoms as nodes, bonds as edges), without 3D coordinates.	2D graph constructed from canonical SMILES; standard message-passing GNN.
BRICS-subgroup GNN	Graph neural network embedding over a coarsened graph whose nodes are BRICS bond-decomposition fragments rather than individual atoms, encoding functional-group-scale structure.	BRICS fragmentation of the 2D graph, then subgroup-level message passing.
Murcko-subgroup GNN	Graph neural network embedding over a coarsened graph whose nodes are Murcko-scaffold ring/linker fragments, encoding scaffold-scale structure.	Murcko-scaffold decomposition of the 2D graph, then subgroup-level message passing.

Supplementary Table S3 3D conformer-embedding fallback counts and rates in the final benchmark-aligned dataset caches. TDC, JCheM, and ComPlat rows summarize their complete cached dataset populations rather than individual train/test partitions.

Dataset cache	**n**	**Embedding failures**	**Failure rate (%)**
AqSolDBc	8,047	12	0.149
TDC	9,982	31	0.311
JCheM	9,798	18	0.184
ComPlat	19,219	7	0.036
Biogen + Cui-62 (external cache)	2,215	0	0.000

Supplementary Table S4 Exact RMSE and MAE results for the main-text evaluations. Within each metric, values are ordered as DLG-Sol / descriptor / neural. Δdescriptor and Δneural are DLG-Sol-minus-component RMSE; negative values favour DLG-Sol. Displayed intervals are 95% paired-bootstrap RMSE CIs; corresponding 99% intervals used for the main-text significance tiers and numerical MAE contrasts are supplied as machine-readable Supplementary data.

Evaluation/split	**Endpoint**	**n**	**DLG-Sol / descriptor / neural**	**Paired RMSE effect [95% CI]**
AqSolDBc development OOF	curated aqueous logS	8,047	RMSE: 0.8070 / 0.8506 / 0.8567; MAE: 0.5431 / 0.5762 / 0.5865	Δdescriptor −0.0436 [−0.0531, −0.0343]; Δneural −0.0497 [−0.0584, −0.0410]
ComPlat OOF	ComPlat logS	17,937	RMSE: 0.8792 / 0.9017 / 0.9532; MAE: 0.5539 / 0.5769 / 0.6107	Δdescriptor −0.0225 [−0.0285, −0.0165]; Δneural −0.0740 [−0.0847, −0.0637]
TDC OOF	TDC AqSolDB logS	7,985	RMSE: 0.9863 / 1.0654 / 1.0607; MAE: 0.6414 / 0.6909 / 0.6929	Δdescriptor −0.0791 [−0.0962, −0.0621]; Δneural −0.0743 [−0.0910, −0.0578]
TDC official Test	TDC AqSolDB logS	1,997	RMSE: 1.0015 / 1.0908 / 1.0049; MAE: 0.7091 / 0.7733 / 0.7178	Δdescriptor −0.0893 [−0.1095, −0.0691]; Δneural −0.0033 [−0.0212, +0.0140]
TDC official Test, five seed retrainings	TDC AqSolDB logS	1,997×5	RMSE: 0.9996–1.0111 / 1.0881–1.0995 / 0.9989–1.0289; MAE: 0.7090–0.7155 / 0.7731–0.7795 / 0.7135–0.7312	All five DLG-Sol-minus-descriptor RMSE CIs exclude zero; DLG-Sol-minus-neural point differences range from −0.0218 to +0.0007
JCheM shared official Test	curated water solubility logSw	980	RMSE: 0.6144 / 0.6551 / 0.6247; MAE: 0.4193 / 0.4496 / 0.4288	Δdescriptor −0.0407 [−0.0539, −0.0273]; Δneural −0.0103 [−0.0249, +0.0033]
ComPlat official Test	ComPlat logS	1,282	RMSE: 0.5599 / 0.5615 / 0.6723; MAE: 0.4092 / 0.4110 / 0.4989	Δdescriptor −0.0016 [−0.0122, +0.0089]; Δneural −0.1123 [−0.1306, −0.0941]

Supplementary Table S5 Provenance, overlap, and evidentiary role of the main-text evaluations. Where numerical overlap counts are given, canonical-SMILES and full-InChIKey matches generated with RDKit 2023.09.6 are reported separately against the 8,047-row AqSolDBc population. These qualifications prevent the six evaluations from being interpreted as six independent external validations.

Evaluation/split	Model-training relation	Overlap and label-access qualification	Evidentiary role
AqSolDBc development OOF	Fold-local descriptor filtering and nested XGB selection; five-seed outer refit; design-fixed $\alpha = 0.5$; held-out descriptor values and labels excluded from the XGB pipeline	Internal curated master population. The neural architecture and schedule had prior Train/Val development-label access and were not reselected in a complete nested loop.	Primary development OOF evidence
ComPlat OOF	Final fold fitting and fold-specific coefficient fitting excluded the held-out-fold labels. Upstream settings used one global Train/Val preselection and were not fully nested.	Separate training distribution, but not a frozen external transfer test; 3,898/17,937 rows overlap AqSolDBc by canonical SMILES and 4,153/17,937 by full InChIKey. Labels were previously accessed under OOF evaluation.	Replication across a different curated benchmark
TDC OOF	Supplied train_val OOF	AqSolDB-derived lineage; 4,240/7,985 rows overlap AqSolDBc by canonical SMILES and 4,442/7,985 by full InChIKey. Labels were previously accessed under OOF evaluation.	Lineage-qualified replication
TDC official Test	Supplied train_val training; official labels excluded from split-specific selection	All 1,997 rows were canonicalizable and InChIKey-valid; 1,095 overlap AqSolDBc by canonical SMILES and 1,147 by full InChIKey. Public labels had been accessed previously and were not first-open sealed.	Supplied-test evidence
TDC official Test, five seed retrainings	Same frozen code, data, and splits; complete pipeline separately retrained with distinct seed offsets	Same AqSolDB-lineage audit as above. Each replication emitted no-label predictions before final scoring.	Training-seed robustness, not independent replication
JCheM shared official Test	Supplied Train/Val fitting; common Test excluded from split-specific selection	AqSolDB-derived curation lineage; 476/980 rows overlap AqSolDBc by canonical SMILES and 505/980 by full InChIKey. Authors' released predictions are available on the same rows, whose labels had been accessed previously as a benchmark.	Supplied-test and fair same-row comparison
ComPlat official Test	ComPlat Train and aug2 OOF coefficient; historical full-train XGB inference	Canonical-SMILES overlaps comprise 750 rows with AqSolDBc Train and 170 with Validation; full-InChIKey overlaps comprise 777 and 173 rows, respectively. Official labels were excluded from reported model and coefficient fitting, but had been accessed previously as a benchmark.	Inconclusive descriptor comparison; resolved neural comparison; conditional literature context

Supplementary Table S6 ComPlat selection-and-tuning sensitivity analysis. Panel A crosses the primary or fold-local DLG-Sol prediction with the reported-setting or investigator-tuned Ali XGB-125D prediction on the 17,937 OOF rows. Panel B isolates changes induced by fold-local DLG-Sol reselection or matched-budget Ali tuning. Panel C reports secondary comparisons on the 1,282-row supplied test. Differences are left-model minus right-model RMSE, so a negative value favours the left model. Confidence intervals are percentile 95% paired-bootstrap intervals from 10,000 molecular-row resamples with seed 20260904. These sensitivity variants do not replace the primary predictions.

Comparison (left vs right)	**Left RMSE**	**Right RMSE**	**RMSE difference [95% CI]**	**Inference**
A. ComPlat OOF 2×2 *comparison*				
Primary DLG-Sol vs reported-setting Ali	0.879214	0.927282	-0.048068 $[-0.055718, -0.040223]$	resolved
Fold-local DLG-Sol vs reported-setting Ali	0.879868	0.927282	-0.047414 $[-0.055593, -0.039183]$	resolved
Primary DLG-Sol vs investigator-tuned Ali	0.879214	0.917690	-0.038476 $[-0.046500, -0.030386]$	resolved
Fold-local DLG-Sol vs investigator-tuned Ali	0.879868	0.917690	-0.037822 $[-0.045872, -0.029676]$	resolved
B. Change induced by each sensitivity procedure				
Fold-local vs primary descriptor	0.908712	0.901746	$+0.006967$ $[+0.002242, +0.011575]$	resolved
Fold-local vs primary neural network	0.949282	0.953223	-0.003941 $[-0.012137, +0.004378]$	inconclusive
Fold-local vs primary DLG-Sol	0.879868	0.879214	$+0.000654$ $[-0.003622, +0.004792]$	inconclusive
Investigator-tuned vs reported-setting Ali	0.917690	0.927282	-0.009592 $[-0.013741, -0.005471]$	resolved
C. ComPlat supplied-test secondary comparisons				
Primary DLG-Sol vs investigator-tuned Ali	0.559936	0.557923	$+0.002013$ $[-0.012222, +0.016858]$	inconclusive
Alternative-alpha DLG-Sol vs investigator-tuned Ali	0.562584	0.557923	$+0.004661$ $[-0.009941, +0.020083]$	inconclusive

Supplementary Table S7 Supplementary Biogen and Ghanavati evaluations. These panels are not pooled with the six primary evaluations. Each fixed-blend prediction is compared with both final component models on identical molecular rows. Effects are differences between the original-sample RMSEs; brackets show paired-bootstrap 95% confidence intervals.

Panel/split	Endpoint	n	Reported RMSE; paired effect [95% CI]	Model-training relation	Exact-overlap audit	Development status & label-access	Manuscript role
Biogen public ADME structure-only panel	pH 6.8 CLND kinetic solubility	2,153	Fixed / XGB / neural: 1.1309 / 1.1317 / 1.1951; fixed−XGB −0.0008 [−0.0189, +0.0167]; fixed−neural −0.0642 [−0.0746, −0.0536]	deployment model fit on AqSolDBc Train 6,437; alpha fixed on Val 1,610 before Biogen label access	zero exact canonical-SMILES overlap against 21 audited scored sources	primary no-label handoff, then one planned primary scoring; later exploratory seed script had one documented additional direct source-label access before I/O correction	Supplementary assay-shift stress test; not pooled
Ghanavati All-Data re-evaluation	aqueous solubility, author-workbook-exact split	789	Fixed / XGB / neural: 0.5658 / 0.5955 / 0.5977; fixed−XGB −0.0298 [−0.0432, −0.0161]; fixed−neural −0.0319 [−0.0512, −0.0130]	train-side OOF alpha; final aug2 heads refit on the author-workbook split	not asserted as zero-overlap external evidence	previously accessed development-adjacent labels	Supplementary retrospective robustness
Ghanavati ESOL re-evaluation	aqueous solubility, source-matched reconstruction	224	Fixed / XGB / neural: 0.5879 / 0.6547 / 0.5911; fixed−XGB −0.0668 [−0.1094, −0.0229]; fixed−neural −0.0033 [−0.0624, +0.0569]	as above	source-matched, not author-file-exact; no independent-overlap claim	previously accessed development-adjacent labels	Supplementary retrospective robustness
Ghanavati AQUA re-evaluation	aqueous solubility, source-matched reconstruction	262	Fixed / XGB / neural: 0.5658 / 0.5967 / 0.5916; fixed−XGB −0.0309 [−0.0528, −0.0105]; fixed−neural −0.0258 [−0.0541, +0.0021]	as above	source-matched, not author-file-exact; no independent-overlap claim	previously accessed development-adjacent labels	Supplementary retrospective robustness
Ghanavati PHYS re-evaluation	aqueous solubility, source-matched reconstruction	401	Fixed / XGB / neural: 0.5150 / 0.5668 / 0.5428; fixed−XGB −0.0518 [−0.0829, −0.0219]; fixed−neural −0.0278 [−0.0549, +0.0008]	as above	source-matched, not author-file-exact; no independent-overlap claim	previously accessed development-adjacent labels	Supplementary retrospective robustness
Ghanavati OCHEM re-evaluation	aqueous solubility, source-matched random-row reconstruction	844	Fixed / XGB / neural: 0.5457 / 0.5951 / 0.5506; fixed−XGB −0.0494 [−0.0716, −0.0293]; fixed−neural −0.0050 [−0.0248, +0.0160]	as above	known 176 canonical-structure train/test overlaps; no final aug2 group-disjoint rerun	previously accessed development-adjacent labels	Supplementary retrospective robustness; not chemical-generalization evidence

Note: Negative effects favour the fixed blend. "Previously accessed" describes label and provenance history, not a failure of split-specific test-label exclusion. The archived JCheM-subset and Cui-62 historical analyses are preserved outside the submission PDF because they do not evaluate the final model as independent evidence.

Supplementary Table S8: Final descriptor–neural model hyperparameters.

Component	Specification
Descriptor branch	2D Mordred descriptors and XGBoost squared-error regression. TDC, ComPlat, and JCheM use their benchmark-specific fitting-partition filters and validation procedures. AqSolDBc repeats missingness, constant, variance ($\leq 10^{-12}$), and pairwise-correlation (> 0.95) filtering within each inner- and outer-training partition. For correlated pairs, descriptors are traversed in original column order: the earlier retained descriptor is kept and the later descriptor is removed. Forty pre-specified configurations are compared by pooled inner-OOF RMSE; early stopping uses 120 rounds, the final tree count is the median inner best iteration plus one, and five seeded outer fits are averaged.
SMILES representation	A single-task encoder initialized from seyonec/ChemBERTa-zinc-base-v1; all encoder layers fine-tuned for log-solubility prediction; learning rate 2×10^{-5}; canonical SMILES plus two randomized SMILES strings per training molecule; pretrained GELU activations retained; 768-dimensional masked mean pooling; maximum length 256; canonical SMILES only at inference. Embeddings were held fixed during fusion-head training.
3D graph construction	Largest organic fragment; up to five ETKDGv3 conformers; MMFF or UFF optimization; top three by energy; heavy atoms and N/O-bound hydrogens as nodes; covalent and 4.5 Å proximity edges
3D node and edge attributes	42 node features and 24 edge features encoding atom identity and state, bond status, scaled interatomic distance, and 16 Gaussian radial-basis values. The complete feature contract is given in the section "3D-MPNN specification".
3D message passing	Scalar message passing with summed incoming messages, SiLU, residual node updates, layer normalization, and global mean pooling. The search space comprised 2–4 layers, hidden widths of 96, 128, 192, or 256, dropout of 0.05, 0.10, 0.15, 0.25, or 0.35, batch sizes of 64, 96, or 128, learning rates from 2×10^{-4} to 3×10^{-3}, weight decay from 10^{-7} to 5×10^{-4}, and graph-only or graph-plus-auxiliary representations. Selection used training-side validation RMSE.
Auxiliary 3D features	19 molecule-level features covering charge state, hydrogen bonding, topology, fragment state, conformer success, and conformer-energy summaries. The complete list is given in the section "3D-MPNN specification", and inclusion was selected using training-side validation.
Conformer aggregation	Arithmetic mean of the retained conformer representations for each molecule
Fusion preprocessing	Training-partition-only modality-wise transformation followed by direct concatenation. ChemBERTa and 3D blocks were standardized separately. AqSolDBc, ComPlat, and TDC additionally applied dimension-wise clipping at percentiles 0.1 and 99.9, estimated from their fitting rows; the JCheM pipeline did not apply this clipping. The final neural model used neither PCA projection nor blockwise L_2 normalization.
Fusion regressors	Four separately trained residual MLP heads receive the same concatenated input. Each head uses a 384-unit input projection followed by two residual blocks and one linear output. Each block applies layer normalization, a 384-to-384 linear transformation, ReLU, dropout 0.45, a second 384-to-384 linear transformation, skip addition, and ReLU. Training uses MSE, AdamW with learning rate 5×10^{-4} and weight decay 5×10^{-4}, batch size 128, and gradient-norm clipping at 5.0.
Fusion-head mapping	Code indices 0–3 use base seeds 260901, 260902, 260903, and 260904 and fixed epoch counts 29, 82, 40, and 21, respectively. The epoch counts were frozen as the per-seed median best epochs from the AqSolDBc five-fold development procedure. Fold-based pipelines add $100000\times$ the fold index to each base seed; the full-train ComPlat supplied-test refit uses the base seeds.
Neural prediction	Unweighted arithmetic mean of the four residual-head predictions. The heads share the same ChemBERTa and 3D representations and the same architecture; they differ in initialization, minibatch order, and fixed training duration.
Blend	$(1-\alpha)\hat{y}_{XGB} + \alpha\hat{y}_{neural}$. AqSolDBc uses $\alpha = 0.5$, fixed by design, in every fold. ComPlat and TDC OOF fit one coefficient per held-out fold from the other four folds' OOF predictions. JCheM fits one coefficient per supplied split from that split's validation predictions. TDC and ComPlat supplied tests each use one evaluation-wide coefficient fitted from complete training-side OOF predictions.

Supplementary Table S9 Correction to the AqSolDBc descriptor pipeline. All blend rows use the same neural predictions and $\alpha = 0.5$. The legacy condition is retained only to quantify the effect of replacing the global development schema. The nested condition is the primary analysis reported in the manuscript. Δ is equal-blend minus its corresponding XGB RMSE. Fold-specific retained-descriptor and tree counts are listed below the table.

Descriptor condition	**XGB RMSE**	**Equal-blend RMSE**	**Δ [95% CI]**
Legacy global schema and development-selected configuration	0.8589	0.8114	-0.0475 [-0.0568, -0.0382]
Fold-local schema with frozen legacy configuration	0.8513	0.8072	-0.0440 [-0.0540, -0.0344]
Fold-local schema and nested XGB selection (primary)	0.8506	0.8070	-0.0436 [-0.0532, -0.0342]

The nested and fold-local-schema-only XGB estimates differed by -0.0007 RMSE (95% CI $[-0.0036, +0.0023]$). The substantive correction therefore came from fold-local descriptor construction. The interval did not support an additional benefit from retuning. For correlation pruning, the absolute Pearson matrix was traversed through its upper triangle in the original Mordred column order. When a pair exceeded 0.95, the descriptor occurring later in that order was removed. This rule is deterministic and label-free, although column-order-dependent. The five nested outer folds retained 373, 353, 354, 352, and 356 descriptors and used 3,456, 3,332, 3,064, 2,601, and 3,198 trees, respectively.

Supplementary Table S10 Software environments and key random-seed records. A single seed cannot summarize the complete workflow because model members, folds, trials, and statistical contrasts used separate deterministic offsets. Complete unit-level seeds and package locks are provided in the machine-readable Supplementary data and code archive.

Workflow	Environment or implementation record	Seed record
DLG-Sol and representation baselines	Python 3.9.23; 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; Transformers 4.50.3	Fold-, trial-, encoder-, graph-, and fusion-member seeds are stored in the corresponding configuration and manifest files.
AqSolDBc repaired descriptor model	Nested filtering and XGBoost selection described in the section "Descriptor preprocessing" and Table S9	Five outer-refit seeds: 260717, 260727, 260737, 260740, and 260747; paired-bootstrap seed 20260821.
TDC, ComPlat, and JCheM descriptor models	Benchmark-specific scripts and artefacts described in the section "Descriptor preprocessing"	Base seeds 260722 (TDC), 260722 (ComPlat), and 260725 (JCheM), with deterministic fold and trial offsets documented in output manifests.
Complementarity analyses	Paired molecular-row bootstrap with 10,000 resamples per component contrast	Base seed 20260823; XGB contrast seed $20260823+1000p+10s$ and neural contrast seed one greater, where p and s index evaluation and subgroup.
External-comparator comparisons excluding MolPROP	Comparator-specific locked environments; Consensus-GNN used Python 3.11.8, TensorFlow/Keras 2.15.0, RDKit 2023.09.6, and DeepChem 2.7.2.dev20240326195015	Paired molecular-row bootstrap used 10,000 resamples with base seed 20260822; training-unit seeds are stored in the completion manifests.
MolPROP reproduction audit	Compatibility environment with Python 3.8.12, PyTorch 1.12.1/CUDA 11.3, PyTorch Geometric 2.3.1, Transformers 4.26.0, Ray 2.1.0, DeepChem 2.7.1, and RDKit 2023.3.3; the environment could not be resolved exactly because the file pinned an unavailable dependency that was not used by the workflow	Source reconstruction used seed 123 and retained all trial and checkpoint hashes; target-run seed and scoring records were retained for audit, but their performance results were not used in the manuscript.
ComPlat selection-and-tuning sensitivity	Fold-local descriptor, ChemBERTa, 3D-MPNN, residual-MLP, and Ali XGB-125D training in locked environments; 25 production units with prediction hashes written before scoring	Inner-split seed $260928 + f$ for outer fold f; 36 descriptor and 36 random Ali candidates per fold, with the reported Ali setting included as an additional anchor; paired-bootstrap seed 20260904.

Supplementary Table S11 Blend coefficient α by evaluation and prediction unit. α is the neural-component weight in $(1-\alpha)\hat{y}_{XGB} + \alpha\hat{y}_{neural}$. AqSolDBc uses an equal weight fixed by design; its fitted coefficients are reported only as a sensitivity analysis.

Evaluation	**Coefficient unit and fitting data**	α
AqSolDBc development OOF	Same design-fixed value in every outer fold; no label-based coefficient fit	0.5000 in folds 0–4. Fitted sensitivity only: 0.4668, 0.4991, 0.4808, 0.4845, 0.4856
ComPlat OOF	Fold-specific; other four folds' OOF component predictions and labels	0.3578, 0.3472, 0.3484, 0.3589, 0.3499
TDC OOF	Fold-specific; other four folds' OOF component predictions and labels	0.5107, 0.5310, 0.5139, 0.4750, 0.5096
TDC supplied test	Evaluation-wide; complete TDC training-side OOF predictions and labels	0.5080
JCheM supplied test	Split-specific; corresponding validation predictions and labels. The five split-specific blended predictions were averaged for each shared test molecule before scoring	0.5042, 0.4575, 0.5181, 0.4462, 0.5755
ComPlat supplied test	Evaluation-wide; complete ComPlat training-side OOF predictions and labels	0.3525
Biogen supplementary panel	Deployment-wide; AqSolDBc validation predictions and labels	0.6284

Supplementary Table S12: Thirty-one matched-row comparisons with external comparator models. PNNL, Ali, Bhattacharya–Roy, and one Consensus-GNN evidence track were retrained or adapted to the benchmark-defined evaluation splits; a separate Consensus-GNN comparison uses the authors' released JCheM predictions. The table contains the 25 comparisons summarized in the main text and six prespecified Bhattacharya–Roy interaction sensitivities. Absolute performance columns give RMSE / MAE. Negative Δ favours DLG-Sol; displayed effects and 95% paired-bootstrap CIs are for RMSE and use 10,000 molecular-row resamples. Corresponding 99% RMSE intervals and numerical MAE contrasts are supplied as machine-readable Supplementary data.

Comparator model	Evaluation	n	DLG-Sol RMSE / MAE	Comparator RMSE / MAE	Δ (DLG-Sol−comparator) [95% CI]
PNNL GNN [9]	AqSolDBc development OOF	8,047	0.8070 / 0.5431	0.9077 / 0.6461	−0.1007 [−0.1150, −0.0864]
PNNL GNN	ComPlat OOF	17,937	0.8792 / 0.5539	1.0319 / 0.7081	−0.1527 [−0.1668, −0.1386]
PNNL GNN	TDC OOF	7,985	0.9863 / 0.6414	1.1411 / 0.7673	−0.1548 [−0.1798, −0.1305]
PNNL GNN	TDC supplied test	1,997	1.0015 / 0.7091	1.1047 / 0.7821	−0.1031 [−0.1313, −0.0751]
PNNL GNN	JCheM supplied test	980	0.6144 / 0.4193	0.6995 / 0.4989	−0.0852 [−0.1155, −0.0566]
PNNL GNN	ComPlat supplied test	1,282	0.5599 / 0.4092	0.6157 / 0.4709	−0.0557 [−0.0796, −0.0315]
Ali XGB-125D† [5]	AqSolDBc development OOF	8,047	0.8070 / 0.5431	0.8457 / 0.5719	−0.0386 [−0.0510, −0.0265]
Ali XGB-125D†	ComPlat OOF	17,937	0.8792 / 0.5539	0.9273 / 0.5974	−0.0481 [−0.0558, −0.0402]
Ali XGB-125D†	TDC OOF	7,985	0.9863 / 0.6414	0.9797 / 0.6336	+0.0066 [−0.0149, +0.0287]
Ali XGB-125D†	TDC supplied test	1,997	1.0015 / 0.7091	1.0374 / 0.7519	−0.0358 [−0.0633, −0.0089]
Ali XGB-125D†	JCheM supplied test	980	0.6144 / 0.4193	0.6583 / 0.4478	−0.0439 [−0.0654, −0.0235]
Ali XGB-125D†	ComPlat supplied test	1,282	0.5599 / 0.4092	0.5543 / 0.4082	+0.0057 [−0.0093, +0.0206]
Bhattacharya–Roy additive MLP-GNN‡ [10]	AqSolDBc development OOF	8,047	0.8070 / 0.5431	1.0639 / 0.7581	−0.2569 [−0.2795, −0.2353]
Bhattacharya–Roy additive MLP-GNN‡	ComPlat OOF	17,937	0.8792 / 0.5539	1.1624 / 0.8362	−0.2832 [−0.3011, −0.2657]
Bhattacharya–Roy additive MLP-GNN‡	TDC OOF	7,985	0.9863 / 0.6414	1.2415 / 0.8592	−0.2551 [−0.2872, −0.2231]

Continued on next page

Supplementary Table S12: Continued.

Comparator model	Evaluation	n	DLG-Sol RMSE / MAE	Comparator RMSE / MAE	Δ (DLG-Sol−comparator) [95% CI]
Bhattacharya–Roy additive MLP-GNN‡	TDC supplied test	1,997	1.0015 / 0.7091	1.1114 / 0.8120	−0.1098 [−0.1415, −0.0787]
Bhattacharya–Roy additive MLP-GNN‡	JCheM supplied test	980	0.6144 / 0.4193	0.7179 / 0.5136	−0.1035 [−0.1292, −0.0785]
Bhattacharya–Roy additive MLP-GNN‡	ComPlat supplied test	1,282	0.5599 / 0.4092	0.6271 / 0.4819	−0.0672 [−0.0959, −0.0381]
Bhattacharya–Roy interaction-augmented MLP-GNN§	AqSolDBc development OOF	8,047	0.8070 / 0.5431	1.0265 / 0.7482	−0.2195 [−0.2398, −0.1995]
Bhattacharya–Roy interaction-augmented MLP-GNN§	ComPlat OOF	17,937	0.8792 / 0.5539	1.1806 / 0.8360	−0.3014 [−0.3288, −0.2791]
Bhattacharya–Roy interaction-augmented MLP-GNN§	TDC OOF	7,985	0.9863 / 0.6414	1.2298 / 0.8390	−0.2434 [−0.2833, −0.2060]
Bhattacharya–Roy interaction-augmented MLP-GNN§	TDC supplied test	1,997	1.0015 / 0.7091	1.0678 / 0.7816	−0.0662 [−0.0978, −0.0343]
Bhattacharya–Roy interaction-augmented MLP-GNN§	JCheM supplied test	980	0.6144 / 0.4193	0.7210 / 0.5263	−0.1067 [−0.1362, −0.0777]
Bhattacharya–Roy interaction-augmented MLP-GNN§	ComPlat supplied test	1,282	0.5599 / 0.4092	0.6642 / 0.5188	−0.1042 [−0.1334, −0.0748]
Ulrich et al. Consensus GNN¶ [11]	JCheM supplied test	980	0.6144 / 0.4193	0.6571 / 0.4745	−0.0427 [−0.0755, −0.0082]
Ulrich et al. Consensus GNN (retrained adaptation)‖	AqSolDBc development OOF	8,047	0.8070 / 0.5431	1.1539 / 0.7703	−0.3469 [−0.3993, −0.2989]
Ulrich et al. Consensus GNN (retrained adaptation)‖	ComPlat OOF	17,937	0.8792 / 0.5539	1.3306 / 0.9624	−0.4514 [−0.4733, −0.4296]
Ulrich et al. Consensus GNN (retrained adaptation)‖	TDC OOF	7,985	0.9863 / 0.6414	1.6854 / 1.0810	−0.6991 [−0.8580, −0.5731]

Continued on next page

Supplementary Table S12: Continued.

Comparator model	Evaluation	n	DLG-Sol RMSE / MAE	Comparator RMSE / MAE	Δ (DLG-Sol−comparator) [95% CI]
Ulrich et al. Consensus GNN (retrained adaptation)‖	TDC supplied test	1,997	1.0015 / 0.7091	1.7998 / 1.1040	−0.7983 [−1.1198, −0.5087]
Ulrich et al. Consensus GNN (retrained adaptation)‖	JCheM supplied test	980	0.6144 / 0.4193	0.8095 / 0.5826	−0.1952 [−0.2510, −0.1414]
Ulrich et al. Consensus GNN (retrained adaptation)‖	ComPlat supplied test	1,282	0.5599 / 0.4092	0.8242 / 0.6291	−0.2642 [−0.3046, −0.2227]

†The reported 125-descriptor schema and model settings were adapted to each benchmark-defined split; the ComPlat rows additionally constitute a same-split reproduction. ‡The source-aligned additive Bhattacharya–Roy architecture was independently retrained from scratch; this does not reproduce the paper's original AqSolDB/BigSolDB2 transfer-learning result. §The optional multiplicative-interaction extension was retained as a prespecified sensitivity analysis and was not selected after scoring. ¶Authors' released predictions on the identical JCheM rows; no model reimplementation. ‖Official GraphConv architecture and paper-described augmentation independently retrained on the benchmark-defined splits; exact augmentation commands, random states, and the authors' complete checkpoint-selection procedure were unavailable, so this is not an exact reproduction. Scored labels were excluded from comparator fitting and final prediction generation, but this is not a first-ever label-access claim.

Supplementary Table S13: Provenance, source-domain status, and permitted interpretation for external-comparator evidence. Matched row counts and paired confidence intervals are reported in Table S12.

Comparison	Evidence type	Code or prediction source	Target-evaluation training and tuning	Source-domain status	Permitted interpretation
JCheM Consensus GNN on shared test	Authors' released predictions	Authors' released JCheM prediction file	No reimplementation; aligned to the same 980 test rows	Released score recovered (RMSE 0.6571)	Direct paired released-prediction comparison
Consensus GNN, six evaluations	Independently retrained official-code adaptation	Public GraphConv implementation and reported augmentation	Five seeded members per unit; fitting-side augmentation; reported architecture and schedule retained	Not treated as exact reproduction	Split-matched adaptation comparison; not the authors' released checkpoints
PNNL GNN, six evaluations	Independently retrained architecture adaptation	Reported method and authors' public repository	Architecture retrained on every benchmark-defined split; no source-domain checkpoint selected	`NOT_IDENTIFIABLE`: source paper reported RMSE 1.0722 on its fixed test and 1.1156 ± 0.0378 in fivefold cross-validation, but the 17,149 reported rows and original split could not be recovered from the 11,692–11,696 public rows	Across the six target evaluations, RMSEs ranged from 0.6157 to 1.1411; these are split-matched architecture comparisons, not source-domain level checks or replications of the original study population
Ali XGB-125D, AqSolDBc/ TDC/JCheM	Schema-and-setting adaptation	Reported 125-descriptor schema and settings	Schema and settings refitted within each benchmark-defined split; no comparable dataset-specific retuning	Not applicable outside source split	Paired implementation comparison, not replication of the original ComPlat study
Ali XGB-125D, ComPlat OOF/test	Same-split reproduction	Authors' repository rows, schema, and settings	The ComPlat rows used here are byte-identical to the repository files	Same-split reproduction	Paired but not independent evidence

Continued on next page

Supplementary Table S13: Continued.

Comparison	Evidence type	Code or prediction source	Target-evaluation training and tuning	Source-domain status	Permitted interpretation
Bhattacharya–Roy additive model, six evaluations	Independently retrained architecture adaptation; source-aligned main comparator	Reported additive architecture, public data, and public split	Architecture retrained on every benchmark-defined split; reported settings retained	`NOT_REPRODUCED`: source-domain mean/median MAE 1.1342/1.1292 versus 1.06/1.07 reported, 7.0%/5.5% higher, respectively	Split-matched adaptation, not the original transfer-learning result
Bhattacharya–Roy interaction-augmented model, six evaluations	Prespecified architecture sensitivity	Optional multiplicative-interaction extension from the same implementation source	Retained independently of outcome; not used to select the main comparator	`NOT_REPRODUCED`: 1.1573/1.1626 versus 1.10/1.08 reported	Prespecified sensitivity only
MolPROP candidate comparator	Reproduction audit only	Public MolPROP revision `7eea0a7`; reconstructed ESOL split; hash-pinned ChemBERTa snapshot	Two 50-trial source BOHB reconstructions; target adaptation completed before the optimizer defect was quantified, then excluded from comparison	`NOT_REPRODUCED`: mean ten-model ESOL holdout RMSE 1.0421 (paper space) and 1.0067 (code space) versus 0.777 ± 0.02 reported as RMSE; manuscript checkpoint, selected ESOL configuration, and logs unavailable; released optimizer omitted the language branch	Not used as a comparator; reproduction audit only

Continued on next page

Supplementary Table S13: Continued.

Comparison	Evidence type	Code or prediction source	Target-evaluation training and tuning	Source-domain status	Permitted interpretation
XGNN candidate comparator	No executable evaluation artefact	Published article and Supplementary Materials only; no public executable implementation, checkpoint, or row-level predictions identified	Not evaluated	Exact source-model reproduction or a split-matched evaluation was not identifiable or achievable from the publicly available artefacts	Related prior work only; no numerical performance comparison
ComPlat XGB-298	Conditional context	Reported point estimate	Same official-test context; no paired adaptation claim	Not assessed here	Numerical context only; no superiority claim
ComPlat historical comparison rows	Historical point estimates	Ali et al. Table 3	No reimplementation; row counts and standardization differ across studies	Not assessed here	Source-paper context only; not a uniformly paired comparison
JCheM EPI Suite, OCHEM, and ACD GALAS	Historical point estimates	Ulrich et al. Table 3	No reimplementation; row-level predictions unavailable to us; EPI Suite covers 934 rather than 980 rows	Not assessed here	Source-paper context only; no paired inference

Supplementary Table S14 Post-development exploratory-grid predictor search spaces. Each panel×representation×predictor combination drew 20 random trials per outer fold; selection used inner-validation RMSE only. These low-budget searches were not capacity-matched to the optimized DLG-Sol components and therefore describe the screen rather than the maximum attainable performance of each representation. Parameter names used in the code are shown in parentheses.

Predictor	**Parameter**	**Candidate values or distribution**
XGBoost	Number of trees (`n_estimators`)	(40, 50, 60)
	Maximum tree depth (`max_depth`)	(2, 3, 4, 5)
	Learning rate (`learning_rate`)	(0.02, 0.04, 0.07, 0.10)
	Row subsampling (`subsample`)	(0.70, 0.85, 1.00)
	Column subsampling (`colsample_bytree`)	(0.50, 0.75, 1.00)
	Minimum child weight (`min_child_weight`)	(1, 3, 6)
	L_2 regularization (`reg_lambda`)	(0.5, 1, 2)
Random forest / extra-trees	Ensemble type (`kind`)	(extra-trees or random forest)
	Number of trees (`n_estimators`)	(50, 75, 100)
	Maximum depth (`max_depth`)	(unrestricted, 10, 18)
	Minimum leaf size (`min_samples_leaf`)	(1, 2, 4)
	Feature fraction (`max_features`)	(0.40, 0.70, 1.00)
Ridge	Regularization strength (`alpha`)	(0.01, 0.1, 1, 10, 100, 300)
k-nearest-neighbours	Number of neighbours (`n_neighbors`)	(3, 5, 10, 20, 40, 80)
	Neighbour weighting (`weights`)	(uniform or distance)
	Minkowski order (`p`)	(1 or 2); Tanimoto geometry replaced p for ECFP
MLP	Hidden layers (`hidden_layer_sizes`)	(one or two layers; 64, 128, or 256 units per layer)
	L_2 regularization (`alpha`)	$(10^u;\ u \sim \mathrm{Uniform}(-6, -3))$
	Initial learning rate (`learning_rate_init`)	$(10^u;\ u \sim \mathrm{Uniform}(-4, -2.5))$

Supplementary Table S15 Best single-representation result in each post-development exploratory screen. The first five evaluations included 35 configurations (seven representations × five predictor families); the ComPlat supplied test included 10 configurations (Mordred and ECFP × five predictor families). Values are point-estimate RMSE. Predictor budgets were not matched to the optimized DLG-Sol components, so these results do not support a final-model performance comparison. Complete configuration-level results are provided as machine-readable Supplementary data.

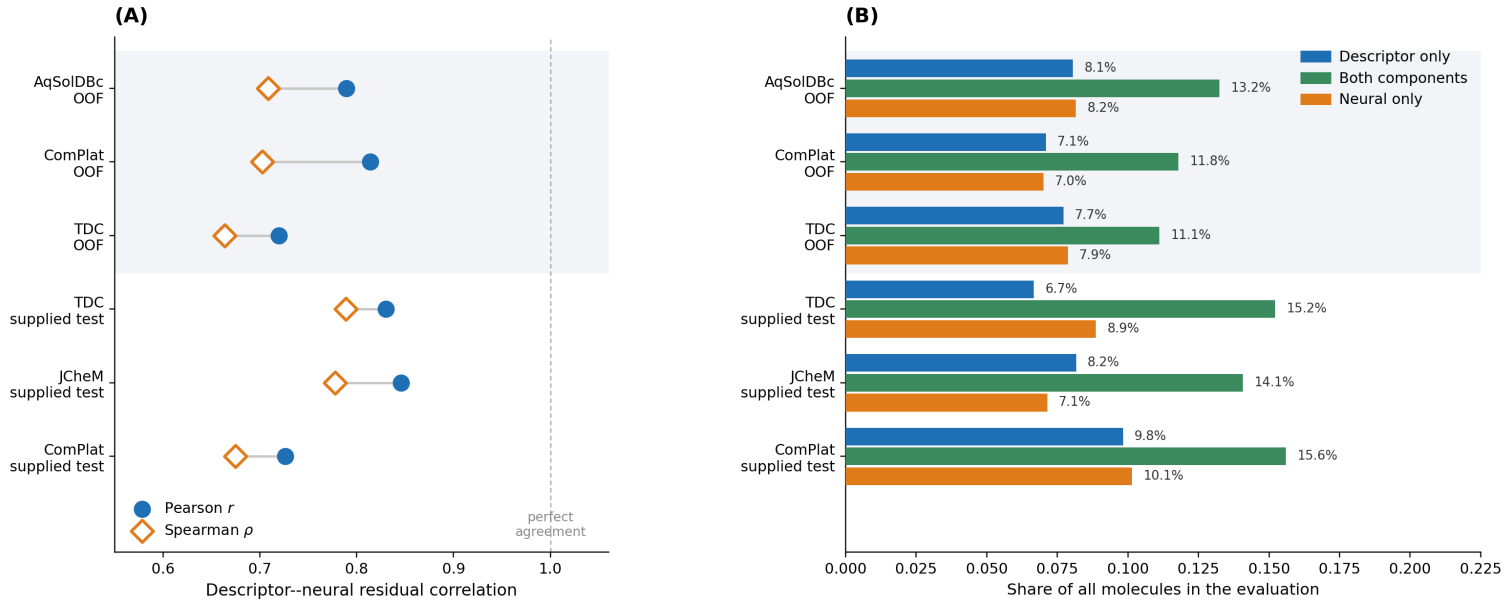
Evaluation	n	**Candidates**	**Best RMSE**	**Best single-representation configuration**
AqSolDBc development OOF	8,047	35	0.8790	Mordred + random forest
ComPlat OOF	17,937	35	0.9504	Mordred + random forest
TDC OOF	7,985	35	1.0646	BRICS-subgroup GNN + random forest
TDC supplied test	1,997	35	1.0836	ChemBERTa + ridge regression
JCheM supplied test	980	35	0.6594	Mordred + multilayer perceptron
ComPlat supplied test	1,282	10	0.5723	Mordred + random forest

Supplementary Table S16 Post-development exploratory representation-pair screen: the combination with the lowest point-estimate RMSE in each primary OOF panel among 105 evaluated combinations (21 unordered representation pairs × five predictor families). Predictor budgets were not capacity-matched to the optimized DLG-Sol components. Full configuration-level results are in the machine-readable Supplementary data.

Panel	**Best pair RMSE**	**Best pair combination**
AqSolDBc development OOF	0.8476	ChemBERTa + Atom-level 2D GNN, random forest
TDC OOF	0.9963	Atom-level 2D GNN + BRICS-subgroup GNN, random forest
ComPlat OOF	0.9473	Mordred + ECFP, random forest

Supplementary Table S17 Evaluation–subgroup rows in main-text Figure 2 with fewer than 100 molecules. Both component contrasts in a row were calculated on the same molecules.

Evaluation	**Chemical axis**	**Subgroup**	n
TDC supplied Test	H-bond acceptor	No	98
JCheM supplied Test	Molecular weight	$\geq$ 400 Da	86
JCheM supplied Test	H-bond acceptor	No	87
ComPlat supplied Test	Molecular weight	$\geq$ 400 Da	41



Supplementary Figure S1 Descriptor–neural residual correlations and large-error overlap. (A) Pearson and Spearman correlations of paired residuals. (B) Fractions of all molecules assigned to descriptor-only, shared, or neural-only large-error subsets. Within each evaluation, a molecule entered a component's subset when its absolute residual exceeded the sample standard deviation of that component's residuals. Molecules outside the three displayed groups did not exceed this threshold for either component. Shaded rows denote OOF evaluations.

Supplementary Table S18 Fixed blend versus K=5 cluster-soft and PCA-linear gating using the same component-model predictions. The main text summarizes their direction and consistency across evaluations. Δ is (gating − fixed) RMSE, so negative values favour gating. Gate hyperparameters were selected on AqSolDBc development OOF only. Transformations and coefficients followed the panel-specific fitting rules in the section "Molecule-specific weighting". Full-precision values are supplied in `molecule_specific_weighting_14_rows.csv`.

Panel	**n**	**Fixed**	**Δ K=5 (95% CI)**	**Δ PCA (95% CI)**
AqSolDBc development OOF	8,047	0.8070	+0.0002 ([−0.0002,+0.0005])	−0.0004 ([−0.0013,+0.0004])
ComPlat OOF	17,937	0.8792	+0.0027 ([+0.0004,+0.0050])	+0.0018 ([−0.0008,+0.0046])
ComPlat official Test	1,282	0.5599	+0.0097 ([+0.0063,+0.0131])	+0.0089 ([+0.0057,+0.0120])
TDC OOF	7,985	0.9863	−0.0008 ([−0.0016,−0.0001])	−0.0033 ([−0.0055,−0.0010])
TDC official Test	1,997	1.0015	+0.0014 ([+0.0009,+0.0021])	−0.0008 ([−0.0023,+0.0005])
JCheM official Test	980	0.6144	+0.0005 ([−0.0002,+0.0011])	−0.0006 ([−0.0019,+0.0006])
Biogen (Supplementary panel)	2,153	1.1309	−0.0130 ([−0.0172,−0.0088])	−0.0119 ([−0.0158,−0.0081])

Supplementary Table S19 Developmental representation screening and neural-component selection. The five configurations with the lowest development-panel RMSE are shown for each stage. Within each configuration, hyperparameter trials were selected on the AqSolDBc validation partition. The rankings on the 2,251-row `OChemUnseen` development panel were then used to prioritize families for focused tuning. Bold marks the lowest development-panel RMSE within a stage. These records document architecture selection. They are not independent validation or capacity-matched comparisons with the final DLG-Sol model. Complete results are supplied as machine-readable Supplementary data.

Stage	Rank	Representation(s)	Predictor	Validation RMSE	Development RMSE
Single-representation screen (35)	1	Mordred	XGBoost	0.8386	**1.2755**
	2	Mordred	Extra Trees	0.8997	1.2892
	3	ChemBERTa	Random forest	0.8766	1.2977
	4	ChemBERTa	MLP	0.8671	1.3015
	5	Mordred	*k*-nearest neighbours	1.0673	1.3186
Pair-representation screen (105)	1	ChemBERTa + 3D-MPNN	MLP	0.8124	**1.2654**
	2	ChemBERTa + BRICS-GNN	Extra Trees	0.8440	1.2716
	3	ChemBERTa + Murcko-GNN	Extra Trees	0.8471	1.2760
	4	Mordred + ECFP	Extra Trees	0.8649	1.2892
	5	ChemBERTa + 3D-MPNN	XGBoost	0.8216	1.2897
Focused pair tuning (12)	1	ChemBERTa + 3D-MPNN	MLP	0.8038	**1.2492**
	2	ChemBERTa + BRICS-GNN	XGBoost	0.8539	1.2756
	3	ChemBERTa + 3D-MPNN	XGBoost	0.8184	1.2853
	4	ChemBERTa + Murcko-GNN	XGBoost	0.8530	1.2889
	5	Mordred + ECFP	XGBoost	0.8287	1.2943

Supplementary Table S20 Source-paper ComPlat and JCheM benchmark contexts with RMSE and available MAE values. Panels A and B reconstruct the model membership and source point estimates of Table 3 in Ali et al. [5] and Table 3 in Ulrich et al. [11], respectively, and add only the DLG-Sol row from this study. A — means that MAE was not reported in the source paper and could not be computed from released row-level predictions. For Consensus GNN, RMSE and MAE were recomputed from the authors' released predictions on the same 980 JCheM rows; the source paper reported RMSE 0.657 but not MAE. Ali et al. reported MAE 0.68 and RMSE 0.59 for the Yan and Gasteiger row (‡). Because MAE cannot exceed RMSE when both metrics are calculated from the same residuals under their standard definitions, these values are retained only as transcribed source-paper context and are not treated as a mutually consistent metric pair. Parentheses retain Ali et al.'s notation for the overlap-affected Sorkun result. Ulrich et al. flagged the OCHEM value (†) because test-set chemicals were present in that model's training set. Historical rows are contextual rather than uniformly paired: row counts, standardization, and training populations differ. Only DLG-Sol versus the released Consensus GNN predictions supports the same-row paired inference reported in the main text and Supplementary Table S12.

Benchmark context	Model or method	Source or evidence	n	RMSE	MAE
A. Ali et al. ComPlat/Huuskonen comparison					
ComPlat test	ANN	Huuskonen et al. [12]; source-paper point estimate	1,294	0.71	—
	ANN	Tetko et al. [13]; source-paper point estimate	1,291	0.62	—
	ANN	Yan and Gasteiger [14]; source-paper point estimate	1,294	0.59	0.68‡
	MLR	Delaney [15]; source-paper point estimate	1,290	0.87	0.68
	MLR	Hou et al. [16]; source-paper point estimate	1,294	0.63	0.52
	SVM	Schroeter et al. [17]; source-paper point estimate	1,290	0.60	0.43
	MLR	Ali et al. (2012) [18]; source-paper point estimate	1,290	0.94	0.72
	MLR	Daina et al. [19]; source-paper point estimate	1,290	1.15	0.93
	ANN	Bjerrum and Sattarov [20]; source-paper point estimate	1,297	0.65	—
	Consensus	Sorkun and Koelman [21]; overlap-affected source-paper point estimate	1,290	(0.53)	(0.39)
	Consensus	Ali et al. overlap-corrected rework [5]	1,290	0.73	0.54
	XGB-298	Ali et al. source-study model [5]	1,282	0.55	0.40
	DLG-Sol	This work; scored on the 1,282 standardized rows	1,282	0.5599	0.4092
B. Ulrich et al. JCheM comparison					
JCheM test	Consensus GNN	Ulrich et al. released predictions [11]; RMSE and MAE recomputed	980	0.6571	0.4745
	EPI Suite [22]	Source-paper point estimate; 46 rows without predictions	934	1.386	—
	OCHEM [23]	Source-paper point estimate; source authors noted test-set chemicals in the model training set	980	0.703†	—
	ACD GALAS [24]	Source-paper point estimate	980	0.845	—
	DLG-Sol	This work; same 980 rows as the released Consensus predictions	980	0.6144	0.4193

References

[1] 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).

[2] 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).

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

[4] 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).

[5] 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).

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

[7] 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).

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

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

[10] 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).

[11] 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).

[12] 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).

[13] 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).

[14] 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).

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

[16] 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).

[17] 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).

[18] 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).

[19] 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).

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

[21] 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).

[22] 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.

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

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