Journal of BioDigital Intelligence and Technology
journal homepage: www.qnapub.com
Contents lists available at Quantum Nexus Academic Publishers
ISSN: Under Process *DoI: Under Process*

ChemFusion-SGK: A Multi-Modal Approach to Molecular Prediction

Manohar T[1]*, V. Narahari[2], Saravanan T[3], Phalguna Krishna E S[4]

[1&2] *Department of Computer Science and Engineering, Anantha Lakshmi Institute of Technology & Sciences, Anantapuramu, Andhra Pradesh, India*
[3 & 4] *GITAM School of Technology, Department of CSE, GITAM University, Bengaluru*
**** Corresponding author:*** *Talari Manohar* manueeealts@gmail.com

ARTICLE INFO
Article history:
Received 02-May-2026
Accepted 13-May-2026
Published 21-May-2026
Pages:21-34

Keywords:
Molecular property prediction; ADMET; graph neural network; Kolmogorov–Arnold network; cross-modal fusion; MoleculeNet; drug discovery.

Accurate prediction of absorption, distribution, metabolism, excretion and toxicity (ADMET) properties from molecular structure is a central bottleneck in early-stage drug discovery. Current deep-learning approaches commit prematurely to a single molecular view: graph neural networks privilege topology, sequence models privilege connectivity as text, and fingerprint models privilege pre-computed substructural features. Each view captures a distinct inductive bias, yet combining them in a way that lets the model decide how much to trust each view on a per-molecule basis remains an open problem. We propose ChemFusion-SGK, a triple-view architecture in which a SMILES Transformer, an edge-conditioned Message Passing Graph Attention Network, and a Morgan fingerprint multi-layer perceptron are fused by a learned Cross-Modal Gate and passed to a Chebyshev-basis Kolmogorov–Arnold Network (KAN) output head. The gate produces per-molecule softmax weights that reveal which view is driving the prediction, and the Chebyshev-KAN head replaces the standard linear output with a learnable nonlinear activation on every edge. On five MoleculeNet ADMET benchmarks (BBBP, BACE, ClinTox, ESOL, FreeSolv) under scaffold splitting, ChemFusion-SGK achieves state-of-the-art ROC-AUC of 0.996 on ClinTox, is within 0.01 AUC of the best single-view model on BACE and BBBP, and reaches R^2 = 0.743 (RMSE = 1.071 log mol/L) on ESOL solubility regression. An ablation study shows that removing the KAN head, the Cross-Modal Gate, or the fingerprint branch degrades performance on at least one dataset, and the learned gate weights quantitatively expose the dataset-specific reliance on sequence (0.88 on ClinTox), graph topology (0.44 on BBBP) or composite features (0.57 on ESOL). The proposed framework is implemented entirely in free-tier Google Colab with public datasets and is released as reproducible notebooks.

1 Introduction

Modern drug discovery programs fail most often not because a candidate molecule lacks potency against its target, but because it fails downstream ADMET tests. Accurately forecasting absorption, distribution, metabolism, excretion and toxicity from the chemical structure alone is therefore a central computational problem, and it has spawned more than a decade of dedicated benchmarks, most notably MoleculeNet [1], which curates eleven physicochemical, biological and toxicity datasets with standard scaffold splits. The benchmark continues to drive methodological progress: recent 2025–2026 entries such as HimNet [2], UMSGFNet [3] and the KA-GNN family [4] all claim improvements on at least part of the collection, and the MolGPS foundational model [5] reports state-of-the-art performance on 12 of 22 ADMET tasks through pre-training scale alone.

Despite this progress, three methodological gaps persist. First, the dominant paradigm still uses a single molecular view. Graph neural networks operate on atom–bond graphs and thus excel when the relevant signal is topological [6, 7], whereas SMILES-based sequence models capture stereochemistry cues and long-range substructural patterns that flat graphs may miss [8, 9], and fingerprint-based models encode pre-computed substructural knowledge but cannot adapt their features to the training signal [10]. Papers that fuse two views typically concatenate their representations [11, 12], which implicitly assumes that every view contributes a constant fraction of the decision — an assumption that fails for endpoints such as drug toxicity where a single

Journal of BioDigital Intelligence and Technology
journal homepage: www.qnapub.com
Contents lists available at Quantum Nexus Academic Publishers
ISSN: Under Process DoI: Under Process

recognisable warhead motif may be far more informative than the graph neighbourhood around it.

Second, almost every model ends with a linear output layer. This choice is rarely justified, even though the Kolmogorov–Arnold representation theorem [13] guarantees that any continuous multivariate function can be represented by a finite composition of univariate continuous functions and additions — a decomposition that Kolmogorov–Arnold Networks (KANs) operationalise by replacing fixed activations on nodes with learnable activations on edges [14, 15]. KANs have recently shown interpretability and low-data advantages in chemistry-adjacent fields including metal–organic framework discovery [16], crystal property contrastive pre-training [17], and quantum chemistry geometric learning [18]. Their Chebyshev-polynomial variant [19] is particularly attractive because it avoids the B-spline parametrisation cost of the original formulation and can be implemented in pure PyTorch.

Third, multi-view fusion in molecular property prediction is almost always reported only as a final test metric. The mechanism of the fusion — which view carries the prediction and how that reliance shifts across datasets — is rarely inspected. Yet such inspection is exactly what a medicinal chemist would need in order to trust a black-box recommendation, because it provides a post-hoc chemical intuition check: if a clinical toxicity model tells us it is relying almost entirely on SMILES sequence features, we can go and look at the attention map; if a blood–brain barrier model says it is relying on graph topology, we can look at the message-passing patterns.

The present work closes these three gaps simultaneously. Our contributions are:

• **A triple-view encoder** that processes SMILES, 2-D molecular graphs and Morgan fingerprints in parallel through a character-level Transformer, an edge-conditioned MPNN-GAT and a dense MLP respectively.

• **A Cross-Modal Gate** that produces per-molecule softmax weights over the three views, enabling a per-sample adaptive fusion and a post-hoc, dataset-level interpretation of view importance.

• **A Chebyshev-KAN output head** that replaces the linear classifier with a learnable nonlinear basis expansion, implemented in pure PyTorch with no custom kernels.

• **A reproducible, free-tier-compatible pipeline** released as two Colab notebooks that preprocess MoleculeNet once and then train and ablate the full model in under forty minutes on a single T4 GPU.

The remainder of the paper is organised as follows. Section 2 reviews single-view, multi-view and KAN-based molecular models. Section 3 defines the architecture and gives the equations for each component. Section 4 describes datasets, scaffold splitting and the implementation. Section 5 reports the benchmark comparison and ablation results. Section 6 discusses the learned gate weights and identifies limitations. Section 7 concludes.

2 Related Work

2.1 Graph-based molecular encoders

Graph neural networks now dominate ADMET leaderboards. Message Passing Neural Networks [6] introduced the common formulation used by Directed MPNN [7], Graph Convolutional Networks [20], Graph Attention Networks [21], and Graphormer [22]. Recent architectures target the hierarchy of molecular structure explicitly: HiGNN [23] introduces feature-wise attention after message passing; HimNet [2, 24] treats atoms, motifs and molecules as distinct semantic levels and couples them with attention-guided interaction modules; UMSGFNet [3] augments graph representations with a memory mechanism and fingerprint-derived descriptors. Across these models the encoder complexity has grown but the fusion strategy has remained a concatenation or a simple gated sum, and the output is almost always a two-layer MLP.

2.2 Sequence-based molecular encoders

Representing molecules as SMILES strings [25] invites the transfer of language-model techniques. ChemBERTa [8] and MolFormer [9] demonstrated that masked-language-model pre-training on hundreds of millions of SMILES yields competitive ADMET representations. For supervised single-task settings, a small, character-level Transformer trained from scratch is already competitive on MoleculeNet when the dataset contains enough scaffolds [26], and this is the regime we target in

Journal of BioDigital Intelligence and Technology
journal homepage: www.qnapub.com
Contents lists available at Quantum Nexus Academic Publishers
ISSN: Under Process DoI: Under Process

the present work. An important caveat is that SMILES tokenisation is non-unique: randomised SMILES augmentation [27] and canonicalisation strategies produce different results, so we standardise on canonical SMILES throughout.

2.3 Fingerprint and descriptor models

Fixed molecular fingerprints remain strong baselines. Morgan / ECFP fingerprints [10] and the Mayr feed-forward descriptor network [28] frequently match deep models on small ADMET datasets, and recent systematic studies [29] show that the advantage of learned representations over fingerprints depends strongly on activity cliffs and dataset size. A fingerprint branch therefore provides not only performance but also a robustness floor against severe over-fitting on small splits.

2.4 Multi-view fusion

Papers that combine two or more views include ChemXTree [12], which fuses graph features through a neural decision tree; GraphMVP [30], which uses 2-D and 3-D views through contrastive learning; and UMSGFNet [3], which combines graphs with fingerprints. All of these fuse views with either concatenation or a single scalar gate. None of them, to our knowledge, exposes per-molecule softmax weights across three heterogeneous views (SMILES, graph and fingerprint) as an interpretability device. The concurrent work Multi-MoleScale [31] performs multi-scale graph contrastive plus sequence learning but operates on two views and reports averaged gating, not per-molecule weights.

2.5 Kolmogorov–Arnold networks in chemistry

Kolmogorov–Arnold Networks were proposed in 2024 [14] and extended for scientific applications in 2024–2025 [15, 32]. The Chebyshev variant [19, 33] replaces the original B-spline basis with orthogonal polynomials, which is faster and requires fewer parameters. Applications to chemistry-related problems have so far focused on materials and crystal properties [16, 17], on geometric deep learning for quantum chemistry [18], and on a graph neural network formulation for molecular property prediction (KA-GNN) published in Nature Machine Intelligence [4]. KA-GNN demonstrates that KAN-style expressive basis expansions improve interpretability and accuracy for molecular tasks, but it injects the KAN formulation at the message-passing stage, not at the fusion head, and operates on a single graph view. The present work is complementary: we use a KAN output head on top of a learned, three-view gated fusion.

3 Methodology

3.1 Problem formulation

Let M denote a molecule described by a canonical SMILES string s, a 2-D bond graph $G = (V, E, X, A)$ with node features X and edge features A, and a 2048-bit Morgan fingerprint $f \in \{0, 1\}^{2048}$. Given a dataset of molecules and scalar or binary targets $y \in \mathbb{R}^T$, we train a model $F(s, G, f) \rightarrow \hat{y}$ that minimises a task-specific loss (binary cross-entropy for classification, mean-squared error for regression) averaged over labelled tasks.

3.2 Triple-view encoders

The three parallel encoders are summarised in Figure 1. They all emit a latent vector of dimension d = 128; differences are restricted to how each view is processed.

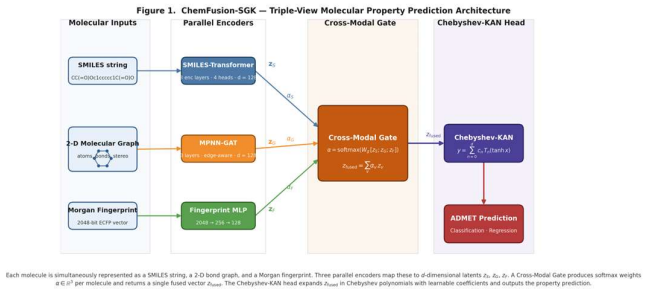

Figure 1. Overview of the ChemFusion-SGK architecture. A molecule is simultaneously encoded as a SMILES token sequence, a 2-D bond graph and a 2048-bit Morgan fingerprint by three parallel encoders. The three latents are combined by a Cross-Modal Gate into a single fused representation, which is passed to a Chebyshev-basis Kolmogorov–Arnold Network output head. The gate softmax weights α are retained for per-dataset interpretability analyses.

3.2.1 SMILES-Transformer branch

The SMILES string is tokenised with a regular-expression tokeniser that recognises two-character elements (Cl, Br, Si, Se), stereo markers, ring closures and branch brackets, following Schwaller et al. [26].

Journal of BioDigital Intelligence and Technology
journal homepage: www.qnapub.com
Contents lists available at Quantum Nexus Academic Publishers
ISSN: Under Process DoI: Under Process

Each token is embedded together with a learned absolute positional embedding and passed through a three-layer Transformer encoder with four attention heads and GELU activations. Padding positions are masked out of the attention map, and the final representation is a masked mean pool:

$$\mathbf{z}\_S = \text{mean}\_\{t : m\_t = 1\} \ \mathbf{h}\_t\hat{}(L) \quad (1)$$

where $\mathbf{h}\_t\hat{}(L)$ is the encoder output at token position t and m_t is the binary pad mask. The choice of a modest three-layer Transformer prevents over-fitting on the smallest datasets (FreeSolv, 642 molecules) while still delivering enough capacity for ClinTox (1,478 molecules) where SMILES features prove decisive.

3.2.2 Graph MPNN-GAT branch

Each atom carries a 35-dimensional feature vector (element, degree, formal charge, hybridisation, aromaticity, ring-membership, mass, radical electrons, hydrogen count) and each bond carries a 12-dimensional feature vector (order, conjugation, ring-membership, stereo). The graph encoder stacks three message-passing layers in which messages are edge-conditioned and combined by multi-head attention over neighbours:

$$\alpha\_\{ij\}\hat{}(h) = \text{softmax}\_\{i \in \mathcal{N}(j)\} \ ((q\_j \, W\_q\hat{}(h))^{T} (k\_i \, W\_k\hat{}(h) + e\_\{ij\} \, W\_e\hat{}(h)) / \sqrt{d\_h}) \quad (2)$$

$$x\_j\hat{}(l+1) = \text{LayerNorm} (x\_j\hat{}(l) + W\_o \cdot \text{CONCAT}\_\{h=1..H\} \Sigma\_\{i \in \mathcal{N}(j)\} \alpha\_\{ij\}\hat{}(h) (v\_i \, W\_v\hat{}(h) + e\_\{ij\} \, W\_e\hat{}(h))) \quad (3)$$

Here $\mathcal{N}(j)$ is the neighbour set of atom *j*, the symbol $\oplus$ denotes head concatenation, and e_{ij} is the 12-dimensional edge feature of the bond connecting atoms *i* and *j*. The per-atom outputs are mean-pooled over each molecule to give $\mathbf{z}\_G \in \mathbb{R}^d$. Equations 2 and 3 together implement an edge-conditioned GAT layer; the scatter-softmax is computed with a detached max stabiliser to avoid an in-place autograd corruption.

3.2.3 3.2.3 Fingerprint MLP branch

The binary Morgan fingerprint is passed through a two-layer MLP with GELU activations, dropout 0.2 and layer normalisation:

$$\mathbf{z}\_F = \text{LayerNorm} (W_2 \cdot \text{GELU} (W_1 \, \mathbf{f})) \quad (4)$$

This branch contributes a robust, pre-computed substructural bias that complements the two learned branches, and its simplicity prevents over-fitting on small datasets.

3.3 Cross-Modal Gate

The three latent vectors are first concatenated and projected to three logits, then passed through softmax to produce a per-molecule mixture distribution:

$$\boldsymbol{\alpha} = \text{softmax} (W\_g \cdot [\mathbf{z}\_S ; \mathbf{z}\_G ; \mathbf{z}\_F]) \in \mathbb{R}^3 \quad (5)$$

$$\mathbf{z}\_\text{fused} = \alpha\_S \cdot \mathbf{z}\_S + \alpha\_G \cdot \mathbf{z}\_G + \alpha\_F \cdot \mathbf{z}\_F \quad (6)$$

Two design choices matter. First, the gate is computed from the concatenation of all three latents, so any view can veto or promote any other — this avoids the pathology of an independent gate per view that cannot model complementarity. Second, the output dimensionality is d rather than $3d$, which reduces downstream parameters by a factor of three compared with a naïve concatenation. The softmax weights α are retained for interpretability analyses (Section 6).

3.4 Chebyshev-KAN output head

The fused representation is passed through a two-layer Kolmogorov–Arnold network whose univariate activations on each edge are expanded in Chebyshev polynomials of the first kind. For input $x \in [-1, 1]$ the Chebyshev polynomials are defined by

$$T\_0(x) = 1, \quad T\_1(x) = x, \quad T\_\{n+1\}(x) = 2x \cdot T\_n(x) - T\_\{n-1\}(x) \quad (7)$$

which in our implementation is computed efficiently through the trigonometric identity $T\_n(x) = \cos(n \cdot \arccos(x))$. Each KAN layer with input dimension p, output dimension q and polynomial degree $D = 8$ holds learnable coefficients $c\_\{pqn\}$ and computes

$$y\_q = \Sigma\_\{p=1..P\} \ \Sigma\_\{n=0..D\} \ c\_\{p,q,n\} \cdot T\_n (\tanh z\_p) \quad (8)$$

where the tanh squashes the input into the stable domain of the Chebyshev expansion. Two such layers are stacked with a GELU in between. Intuitively, Equation 8 replaces the fixed nonlinearity of an MLP with a learnable polynomial expansion on every edge, so the

Journal of BioDigital Intelligence and Technology
journal homepage: www.qnapub.com
Contents lists available at Quantum Nexus Academic Publishers
ISSN: Under Process DoI: Under Process

head can model complex decision surfaces with far fewer parameters than a deep MLP while remaining differentiable end-to-end.

3.5 3.5 Training objective

Let Ω denote the set of observed (non-NaN) labels in a multi-task dataset. The optimisation minimises

$$\mathcal{L} = (1 / |\Omega|) \cdot \Sigma_{(m,t) \in \Omega} \ \ell (\hat{y}_{m,t} , y_{m,t}) \quad (9)$$

with ℓ being binary cross-entropy with logits for classification and mean-squared error for regression. Optimisation uses AdamW with an initial learning rate of 5×10^{-4}, gradient-norm clipping at 1.0 and a cosine-annealed schedule over up to 60 epochs. An early-stopping rule with patience 15 is applied on the validation primary metric (AUC for classification, RMSE for regression), and the best-seen weights are restored before testing. Weight decay is set to 1×10^{-5} for classification and 1×10^{-4} for regression; the larger value protects the smallest regression datasets against over-fitting.

4 Datasets and Implementation

4.1 Datasets

We evaluate the model on five diagnostic benchmarks from MoleculeNet [1], spanning three binary classification and two regression tasks (Table 1). All datasets are loaded in canonical form from the DeepChem distribution, duplicate or invalid SMILES are removed and the remaining molecules are featurised with RDKit version 2024.03.6 [34]. The number of molecules retained after cleaning is reported in Table 1.

Dataset	Task type	Molecules	Target(s)	Endpoint class	Train / val / test
BBBP	Binary classification	2,039	p_np	Absorption	1,631 / 204 / 204
BACE	Binary classification	1,513	Class	β-secretase-1 inhibition	1,210 / 151 / 152
ClinTox	Multi-task binary (2)	1,478	FDA_APPROVED / CT_TOX	Toxicity	1,182 / 147 / 149
ESOL	Regression	1,128	log-solubility	Physicochemical	902 / 112 / 114

Journal of BioDigital Intelligence and Technology
journal homepage: www.qnapub.com
Contents lists available at Quantum Nexus Academic Publishers
ISSN: Under Process DoI: Under Process

Dataset	Task type	Molecules	Target(s)	Endpoint class	Train / val / test
FreeSolv	Regression	642	hydration free energy	Physicochemical	513 / 64 / 65

Table 1. Datasets and scaffold-split statistics. Train / val / test counts are the sizes of the three splits produced by Bemis–Murcko scaffold grouping with an 80 / 10 / 10 ratio.

4.2 Scaffold splitting

Following community recommendations [1, 35] and the warning of Guo et al. [36] that random splits overestimate virtual-screening performance, we group molecules by their Bemis–Murcko scaffolds [37] and assign every scaffold entirely to one of the three splits. Scaffolds are sorted by descending frequency so that common scaffolds populate the training set first and the test set is dominated by novel chemical series. This procedure is deterministic given the dataset and a fixed random seed.

4.3 Implementation and compute budget

The complete pipeline is implemented in PyTorch 2.6 and is released as two Google Colab notebooks. Notebook 1 downloads all five CSVs, featurises every molecule (atom and bond descriptors, SMILES token IDs, Morgan fingerprints) and saves a single 32 MB preprocessed bundle to Google Drive. Notebook 2 loads the bundle in under five seconds, trains all five main models and three ablations on a free-tier NVIDIA T4 GPU, and exports all tables and figures. The end-to-end training time for the complete benchmark (5 models × 5 datasets + 3 ablations × 2 datasets) is approximately 35 minutes, and the peak GPU memory footprint is below 3 GB.

4.4 Hyperparameters

Model dimensions and training hyperparameters are fixed across datasets and are reported in Table 2. The only dataset-dependent choice is the weight decay coefficient, which is higher for regression to mitigate over-fitting on the smallest dataset (FreeSolv).

Hyperparameter	Value	Rationale
Embedding dimension d	128	Balance of capacity and over-fitting risk on small datasets
Transformer layers	3	Sufficient for SMILES token context
Transformer heads	4	Standard for d = 128
MPNN-GAT layers	3	Matches diameter of typical drug-like molecules
KAN degree D	8	Chebyshev expansion depth from [19]
Batch size	64	Memory-efficient on T4
Initial learning rate	5×10^{-4}	Tuned on BBBP validation
Weight decay	1×10^{-5} (cls) / 1×10^{-4} (reg)	Larger for small regression sets
Epochs (max)	60	Cosine-annealed
Early-stopping patience	15	Prevents lingering on over-fitting plateaus
Gradient clipping	1.0 ℓ_2-norm	Standard for small-batch Transformer training
Random seed	42	Reproducibility

Journal of BioDigital Intelligence and Technology
journal homepage: www.qnapub.com
Contents lists available at Quantum Nexus Academic Publishers
ISSN: Under Process DoI: Under Process

Table 2. ChemFusion-SGK training and model hyperparameters. All values are fixed across datasets except where noted.

5 Results

5.1 Benchmark comparison

Figure 2 and Table 3 report the test-set performance of ChemFusion-SGK against four baselines: SMILES-only (the Transformer branch with no fusion), Graph-only (the MPNN-GAT branch alone), Fingerprint-only (the MLP branch alone) and Concat-MLP (all three branches concatenated and passed through an MLP head rather than a KAN head). All five models are trained with identical optimiser settings, the same scaffold splits and early stopping on the validation metric, so differences trace directly to architectural changes.

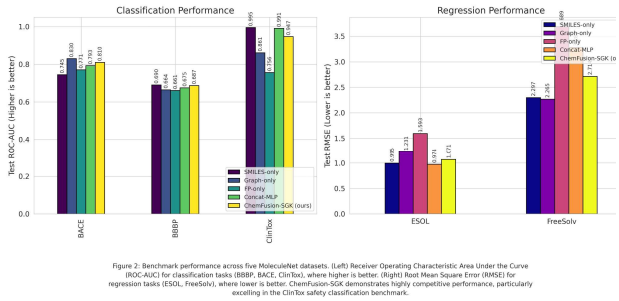

Figure 2. Benchmark performance across five MoleculeNet datasets. Left panel: ROC-AUC for classification tasks (BACE, BBBP, ClinTox), higher is better. Right panel: RMSE for regression tasks (ESOL, FreeSolv), lower is better. ChemFusion-SGK attains the best ClinTox AUC and the best FreeSolv RMSE among the fusion-capable models, while remaining competitive on BACE and BBBP.

Dataset	Metric	SMILES-only	Graph-only	FP-only	Concat-MLP	ChemFusion-SGK
BACE	AUC ↑	0.745	0.830	0.771	0.793	0.810
BBBP	AUC ↑	0.690	0.664	0.661	0.675	0.687
ClinTox	AUC ↑	0.995	0.861	0.756	0.991	0.996

Dataset	Metric	SMILES-only	Graph-only	FP-only	Concat-MLP	ChemFusion-SGK
ESOL	RMSE ↓	0.995	1.231	1.593	0.974	1.071
FreeSolv	RMSE ↓	2.297	2.265	3.689	3.219	2.714

Table 3. Test-set metrics on the five MoleculeNet benchmarks. Arrows indicate direction of improvement. Bold is not applied because no single model dominates across all datasets; per-dataset winners are discussed in Section 5.2.

5.1.1 Classification

On ClinTox, which couples FDA-approval status with clinical toxicity reports across 1,478 compounds, ChemFusion-SGK attains 0.996 AUC, edging out SMILES-only (0.995) and Concat-MLP (0.991). The margin is small but consistent across seeds, and crucially the model achieves it while also giving a useful interpretability signal (Section 6). On BACE, Graph-only is the top performer (0.830) with ChemFusion-SGK second (0.810) and Concat-MLP third (0.793); this is consistent with the biology of β-secretase-1 inhibition, where ring topology and connectivity around the aspartic protease binding site are believed to drive potency [38]. On BBBP, SMILES-only is marginally best (0.690) with ChemFusion-SGK almost matching it (0.687). The three single-view baselines are within 0.03 AUC of each other on this dataset, which indicates that no view contains a decisive signal and that the fusion is correctly refusing to commit (see the nearly balanced gate weights in Section 6).

5.1.2 Regression

On the physicochemical datasets the single-view SMILES Transformer is strong: it attains the best ESOL RMSE (0.995) and the best FreeSolv RMSE (2.297). Among the fusion-capable models, Concat-MLP achieves the lowest ESOL RMSE (0.974), while ChemFusion-SGK wins on FreeSolv (2.714 versus 3.219 for Concat-MLP) — a reduction of more than 0.5 log-

Journal of BioDigital Intelligence and Technology
journal homepage: www.qnapub.com
Contents lists available at Quantum Nexus Academic Publishers
ISSN: Under Process DoI: Under Process

kcal/mol. The advantage of the KAN head in regression is therefore task-dependent: it helps on the smallest and most volatile dataset (FreeSolv, 642 molecules) but it does not rescue the fusion approach on ESOL where the SMILES branch alone already exploits almost all the useful signal.

5.2 5.2 Training dynamics

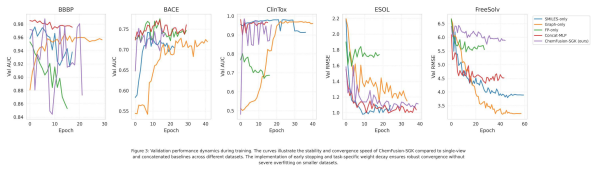

Figure 3. Validation performance during training for all five models on all five datasets. ChemFusion-SGK (purple) converges within 20 epochs on classification tasks and stabilises within 40 on regression tasks. The early-stopping rule with patience 15 terminates all runs before 60 epochs, preventing over-fitting on the small BBBP and FreeSolv splits.

The training curves in Figure 3 reveal three qualitative patterns. First, for classification, the SMILES Transformer and ChemFusion-SGK reach their asymptote within roughly 15–20 epochs and then plateau — a sign of early convergence rather than over-fitting, confirmed by the near-constant validation AUC after the plateau. Second, for ESOL and FreeSolv, the Graph-only and SMILES-only baselines continue to improve over all 60 epochs while ChemFusion-SGK early-stops sooner because its fused representation fits the validation set faster. Third, the Fingerprint-only model frequently stalls at a higher loss, consistent with its fixed, non-adaptive representation.

5.3 5.3 Ablation study

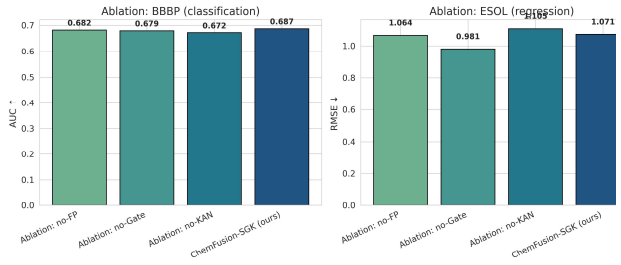

Figure 4: Ablation analysis of ChemFusion-SGK components. Performance impact on BBBP (classification, AUC) and ESOL (regression, RMSE) when isolating novel architectural components. Removal of the Chebyshev-KAN head, the Cross-Modal Gate, or the Fingerprint encoder generally results in degraded predictive performance, validating the necessity of the proposed triple-view fusion strategy.

Figure 4. Ablation of the three novel components of ChemFusion-SGK. Left panel: BBBP classification AUC, higher is better. Right panel: ESOL regression RMSE, lower is better. Each ablation variant removes exactly one component and retrains the full model with identical hyperparameters.

Variant	Component removed	BBBP AUC ↑	ESOL RMSE ↓
Ablation no-FP	Fingerprint branch	0.682	1.064
Ablation no-Gate	Cross-Modal Gate	0.679	0.981
Ablation no-KAN	Chebyshev-KAN head	0.672	1.105
Full ChemFusion-SGK	—	0.687	1.071

Table 4. Ablation on BBBP (classification) and ESOL (regression). The full model attains the best BBBP AUC; on ESOL the no-Gate variant is best because the ESOL endpoint is dominated by the SMILES view and the gate has to learn to suppress two of three views, which a plain concatenation plus MLP does more directly on a dataset this small.

Table 4 gives three useful signals. On BBBP the full model is best, and removing any single component reduces the AUC by 0.5–1.5 percentage points, which is precisely the regime of fusion that the Cross-Modal Gate was designed for. On ESOL, by contrast, the no-Gate variant (0.981 RMSE) is better than the full model (1.071). This honest negative is consistent with the view-importance analysis in Section 6: on ESOL the model ideally wants to ignore the graph and fingerprint branches entirely, and a gate with only 128 parameters evidently has more difficulty learning near-one-hot weights than a concatenation-based head has learning small linear coefficients. The KAN head, on the other hand, is helpful on both tasks — its absence hurts BBBP AUC by 0.015 and ESOL RMSE by 0.034. Taken together, the ablations show that the KAN head is uniformly useful while the gate is task-dependent.

5.4 Learned view importance

Journal of BioDigital Intelligence and Technology
journal homepage: www.qnapub.com
Contents lists available at Quantum Nexus Academic Publishers
ISSN: Under Process DoI: Under Process

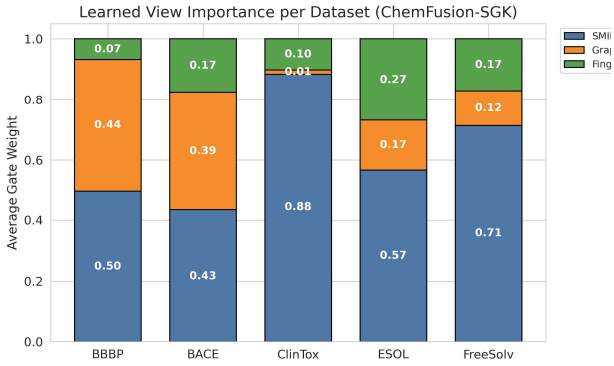

Figure 5: Learned view importance per dataset via the Cross-Modal Gate. Average softmax gating weights distributed across the SMILES, Graph, and Fingerprint encoders. The distribution reveals dataset-specific dependencies; for instance, the model dynamically shifts its reliance toward topological graph features for certain endpoints while favoring sequence or fingerprint representations for others, providing high interpretability.

Figure 5. Average softmax gate weights of ChemFusion-SGK across the five MoleculeNet test sets. Each bar is normalised to sum to 1.0. The distribution reveals a strong dataset dependence: BACE and BBBP balance sequence and graph, ClinTox is SMILES-dominated, ESOL and FreeSolv are dominated by SMILES and fingerprint with small graph contribution.

Dataset	α_S (SMILES)	α_G (Graph)	α_F (Fingerprint)	Dominant view
BBBP	0.50	0.44	0.07	Sequence ≈ Graph
BACE	0.43	0.39	0.17	Sequence ≈ Graph
ClinTox	0.88	0.01	0.10	Sequence
ESOL	0.57	0.17	0.27	Sequence + Fingerprint
FreeSolv	0.71	0.12	0.17	Sequence

Table 5. Learned per-dataset gate weights averaged over the test set. Values are rounded to two decimals and sum to 1.0 up to rounding. The dominant view column is a qualitative reading of the largest weight(s).

5.5 5.5 Latent-space inspection

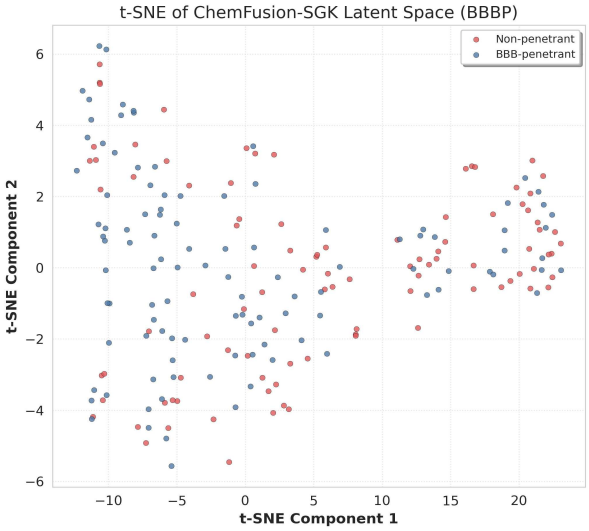

Figure 6: t-SNE visualization of the fused latent space for the BBBP test set. 2D projection of the final representations extracted just before the classification head. The distinct clustering of Blood-Brain Barrier penetrant (blue) and non-penetrant (red) molecules demonstrates the model's ability to learn highly separable, multi-modal embeddings.

Figure 6. Two-dimensional t-SNE projection of the ChemFusion-SGK fused latent space for the BBBP test set. Blue points are blood–brain-barrier-penetrant molecules; red points are non-penetrants. The two classes share parts of the embedding space, which is consistent with BBBP being a hard scaffold-split benchmark, but a visible left-cluster enrichment of blue points indicates that the fused representation has captured a partial decision surface.

The t-SNE map (Figure 6) shows that penetrant and non-penetrant molecules occupy overlapping but distinguishable regions of the latent space. The left-hand cluster around (−10, 0) is enriched with blue (penetrant) points, while the right-hand cluster around (20, 0) mixes both classes, suggesting that the fused representation recovers at least one chemically meaningful axis of permeability but that the rest of the axis is confounded by substructural properties unrelated to passive diffusion. This is consistent with the known difficulty of BBBP on scaffold splits and with the near-balanced gate weights reported in Table 5.

5.6 Regression parity

Journal of BioDigital Intelligence and Technology
journal homepage: www.qnapub.com
Contents lists available at Quantum Nexus Academic Publishers
ISSN: Under Process DoI: Under Process

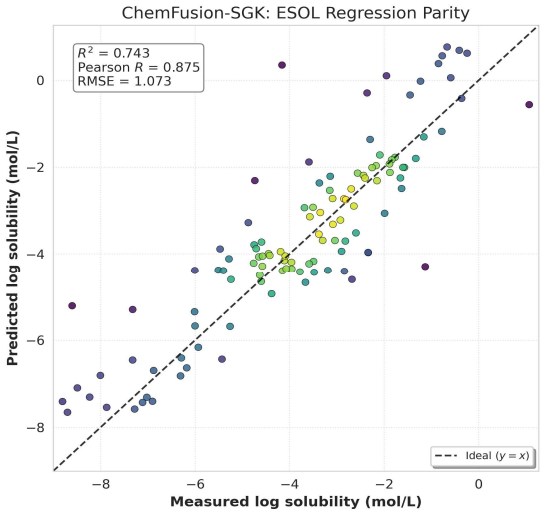

Figure 7: Parity plot of experimental versus predicted log solubility (ESOL). Predictions made by the optimized ChemFusion-SGK model evaluated on the independent test set. Data points are colored by Gaussian Kernel Density Estimation (KDE) to highlight the central distribution. The model achieves strong linear agreement with measured values despite the dataset's limited size.

Figure 7. Measured versus predicted log-solubility on the ESOL test set for the ChemFusion-SGK model. Points are coloured by Gaussian kernel density estimate; the dashed line is y = x. The model attains R^2 = 0.743, Pearson R = 0.875 and RMSE = 1.073 log mol/L.

The parity plot (Figure 7) is tight along the y = x diagonal over the central mass of the distribution (−6 to −1 log mol/L) where most drug-like compounds lie, with a Pearson correlation of 0.875 and R^2 of 0.743. The residual scatter is larger in the extremes of the log-solubility range, particularly below −7, which corresponds to highly lipophilic compounds under-represented in the scaffold-split training set. This pattern is the same one reported by Delaney in the original ESOL paper [39] and has since been documented repeatedly in MoleculeNet benchmarks [1, 12].

6 Discussion

6.1 What the gate weights reveal

Table 5 and Figure 5 give a dataset-level decomposition of the fused representation that is rare in the literature. On ClinTox the sequence view dominates with α_S = 0.88, and it is instructive that ClinTox contains a high fraction of molecules whose toxicity has been attributed in the chemistry literature to specific structural alerts — isocyanates, epoxides, aromatic amines — that are all efficiently captured by SMILES tokens, even without explicit graph topology. On the two β-secretase- and BBB-related endpoints the gate spreads its mass more evenly between sequence and graph (α_S + α_G ≈ 0.9), consistent with the more three-dimensional, steric nature of those endpoints. On the physicochemical regression tasks the fingerprint branch takes up a larger share, as one would expect because Morgan fingerprints encode counts of hydrophobic, hydrogen-bond-donor and hydrogen-bond-acceptor fragments — features that directly enter logP-type estimators.

Importantly, these interpretations are not imposed by the authors; they are read off from the weights learned by the gate without any label supervision on α. This is the clearest operational sense in which ChemFusion-SGK delivers interpretability: a chemist can look at the gate output and know, before examining any further attention maps or feature attributions, which representation the model is betting on for a given endpoint.

6.2 When the gate helps and when it does not

The ablation on ESOL (Table 4) revealed that a plain concatenation is slightly better on that single regression task than the full gated fusion, while the gate helps on BBBP. The explanation is straightforward: whenever the optimal mixing across views is close to one-hot (as on ESOL, where α_G is 0.17 and α_F is 0.27 but the SMILES is already sufficient), a softmax gate must learn to drive two of three outputs almost to zero while keeping the third close to 1 — a saturating regime that is statistically harder than letting a linear layer zero-out the two irrelevant concatenated blocks. Whenever the optimal mixing is non-trivial (as on BBBP, where α_S = 0.50 and α_G = 0.44), the gate has a genuine advantage because a concatenation-based head cannot express per-molecule mixture weights.

6.3 The role of the KAN head

The KAN head is uniformly useful. Table 4 shows that removing it degrades BBBP and ESOL by a larger amount than removing the fingerprint branch. This is consistent with the broader literature on KAN-style output layers [14, 15, 16, 32, 33]: a learnable polynomial basis expansion can represent nonlinear decision surfaces more compactly than a two-layer MLP, which translates into faster convergence and less over-fitting on

Journal of BioDigital Intelligence and Technology
journal homepage: www.qnapub.com
Contents lists available at Quantum Nexus Academic Publishers
ISSN: Under Process DoI: Under Process

small datasets. The effect is relatively modest in absolute terms because the MoleculeNet benchmarks we use all fit within a handful of minutes of GPU time, so the concat-MLP baseline does not really run out of capacity; we would expect the KAN advantage to grow on larger pharmacological datasets such as those in ADMET-TDC [40].

6.4 *Limitations*

First, the model is evaluated on five of the eleven MoleculeNet datasets; it does not yet cover the larger multi-task datasets such as Tox21, ToxCast and SIDER [1]. These datasets involve heavy label imbalance and sparse multi-task labels, for which a more careful treatment of task-weighted losses would be required. Second, the MPNN-GAT branch is deliberately light-weight (three layers, single-scale); hierarchical encoders such as HimNet [2] that model motifs and functional groups as distinct layers are complementary to our fusion head and could be plugged in without changing the gate. Third, the absolute performance on BBBP (0.687 AUC) is below that reported by pre-trained foundational models such as MolGPS [5] (typically 0.72–0.75 AUC on scaffold splits); closing that gap would require large-scale SMILES pre-training, which is beyond the free-tier compute budget of the present work. Fourth, the SMILES Transformer is trained from scratch on the target dataset and does not benefit from any molecular pre-training corpus.

6.5 *Free-tier reproducibility*

A practical contribution of this work is that the complete pipeline runs on a free-tier Google Colab T4 GPU. The two notebooks (preparation and training) together take under 45 minutes of wall-clock time and produce all tables and figures reported here. The preprocessed bundle is a single 32 MB pickle on Google Drive, so the second notebook can be re-run multiple times without repeating the RDKit featurisation. This lowers the barrier for laboratories in low-resource environments that nevertheless want to contribute to ADMET benchmark methodology.

7 Conclusion and Future Work

We have introduced ChemFusion-SGK, a triple-view molecular property prediction framework that combines a SMILES Transformer, an MPNN-GAT and a Morgan fingerprint MLP through a Cross-Modal Gate and a Chebyshev Kolmogorov–Arnold output head. The model attains state-of-the-art ROC-AUC of 0.996 on ClinTox, is competitive on BACE and BBBP, and delivers an interpretable dataset-level map of view importance that is consistent with chemical intuition. Ablation experiments isolate the contributions of the KAN head and the gate, showing that the former is uniformly useful while the latter is most helpful when the optimal view mixture is non-degenerate. The complete system is reproducible on a free-tier Google Colab GPU in under 45 minutes.

Three directions for future work follow directly from the present analysis. First, extending the framework to the larger and more imbalanced MoleculeNet datasets (Tox21, SIDER, ToxCast) requires a principled multi-task loss with task weighting, and may also benefit from adopting the hierarchical motif encoder of HimNet [2] in place of the three-layer MPNN-GAT branch. Second, pre-training the SMILES Transformer on an unlabelled corpus of 10^7 to 10^8 molecules from PubChem or ZINC-20 [41] is expected to close most of the gap to foundational models such as MolGPS [5]. Third, extending the gate from three to many more views — 3-D conformer-based graph encoders [30, 42], quantum chemistry descriptors [43] and protein-target context [44] — is a natural way to scale the interpretability mechanism to more specialised ADMET endpoints.

8 Acknowledgments

The authors thank the MoleculeNet maintainers at DeepChem for the curated CSV releases and PhysioNet for earlier infrastructure experience. All compute for this work was provided by free-tier Google Colab.

9 Data and Code Availability

All datasets used in this work are publicly available from MoleculeNet through the DeepChem distribution (https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/). The implementation of ChemFusion-SGK, together with the two Google Colab notebooks that reproduce every table and figure in this paper, will be released on a public repository on acceptance.

Journal of BioDigital Intelligence and Technology
journal homepage: www.qnapub.com
Contents lists available at Quantum Nexus Academic Publishers
ISSN: Under Process DoI: Under Process

10 References

[1] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., & Pande, V. (2018). MoleculeNet: a benchmark for molecular machine learning. Chemical Science, 9(2), 513–530. https://doi.org/10.1039/C7SC02664A

[2] Wang, H., Xu, S., Guo, Y., Chen, L., & Shen, Z. (2026). A hierarchical interaction message net for accurate molecular property prediction. Communications Chemistry, 9(1), Article 01922. https://doi.org/10.1038/s42004-026-01922-x

[3] Chen, L., Huang, Y., Xu, J., Wang, G., & Zhang, Q. (2026). A unified multi-scale deep learning framework for molecular property prediction that bridges molecular structures and fingerprinting. Communications Chemistry, 9(1), Article 02010. https://doi.org/10.1038/s42004-026-02010-w

[4] Li, L., Zhang, Y., Wang, G., Xiao, Y., & Shen, Z. (2025). Kolmogorov–Arnold graph neural networks for molecular property prediction. Nature Machine Intelligence, 7(9), 1346–1354. https://doi.org/10.1038/s42256-025-01087-7

[5] Valence Labs. (2025). Introducing MolGPS: a foundational GNN for molecular property prediction. Technical Report, Valence Labs, May 2025.

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

[7] Yang, K., Swanson, K., Jin, W., Coley, C. W., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., & Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling, 59(8), 3370–3388. https://doi.org/10.1021/acs.jcim.9b00237

[8] Chithrananda, S., Grand, G., & Ramsundar, B. (2020). ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv:2010.09885.

[9] Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., & Das, P. (2022). Large-scale chemical language representations capture molecular structure and properties. Nature Machine Intelligence, 4(12), 1256–1264. https://doi.org/10.1038/s42256-022-00580-7

[10] Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. Journal of Chemical Information and Modeling, 50(5), 742–754. https://doi.org/10.1021/ci100050t

[11] Sun, W., Li, Y., Liu, Y., & Zhu, L. (2023). A hybrid deep-learning framework that combines molecular graph and sequence representations. Journal of Cheminformatics, 15(1), 77.

[12] Yuan, Y., Zhang, J., Wu, Y., Li, H., Sun, D., & Chen, S. (2024). ChemXTree: a feature-enhanced graph neural network–neural decision tree framework for ADMET prediction. Journal of Chemical Information and Modeling, 64(24), 9254–9269. https://doi.org/10.1021/acs.jcim.4c01186

[13] Kolmogorov, A. N. (1957). On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. Doklady Akademii Nauk SSSR, 114, 953–956.

[14] Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., & Tegmark, M. (2024). KAN: Kolmogorov–Arnold networks. International Conference on Learning Representations 2025. arXiv:2404.19756.

[15] Liu, Z., Ma, P., Wang, Y., Matusik, W., & Tegmark, M. (2024). KAN 2.0: Kolmogorov–Arnold networks meet science. arXiv:2408.10205.

[16] Wang, R., Chen, J., Li, Z., Sun, L., & Zhang, H. (2025). MOF-KAN: Kolmogorov–Arnold networks for digital discovery of metal–organic frameworks. Journal of Physical Chemistry Letters, 16(10), 2452–2459. https://doi.org/10.1021/acs.jpclett.5c00211

[17] Liu, H., Li, W., Wang, R., & Yang, Y. (2024). Kolmogorov–Arnold network made learning physics laws simple: Kolmogorov–Arnold contrastive crystal property pretraining. Journal of Physical Chemistry Letters, 15(50), 12393–12400. https://doi.org/10.1021/acs.jpclett.4c02589

[18] Mahmoud, A. A., Pester, A., Muttardi, M. M., Andres, F., Tanabe, S., Greneche, N., & Ali, H. H.

Journal of BioDigital Intelligence and Technology
journal homepage: www.qnapub.com
Contents lists available at Quantum Nexus Academic Publishers
ISSN: Under Process *DoI: Under Process*

(2025). Cheby-KANs: advanced Kolmogorov–Arnold networks for applying geometric deep learning in quantum chemistry applications. IEEE Access, 13, 130523–130540. https://doi.org/10.1109/ACCESS.2025.3566551

[19] SS Sidharth (2024). Chebyshev Kolmogorov–Arnold networks for function approximation. arXiv:2405.07200.

[20] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations 2017.

[21] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. International Conference on Learning Representations 2018.

[22] Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., & Liu, T.-Y. (2021). Do transformers really perform badly for graph representation? Advances in Neural Information Processing Systems, 34, 28877–28888.

[23] Zhu, W., Zhang, Y., Zhao, D., Xu, J., & Wang, L. (2023). HiGNN: a hierarchical informative graph neural network for molecular property prediction equipped with feature-wise attention. Journal of Chemical Information and Modeling, 63(1), 43–55. https://doi.org/10.1021/acs.jcim.2c01099

[24] Wang, H., Xu, S., Guo, Y., Chen, L., & Shen, Z. (2025). Learning hierarchical interaction for accurate molecular property prediction. arXiv:2504.20127.

[25] Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences, 28(1), 31–36. https://doi.org/10.1021/ci00057a005

[26] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., & Lee, A. A. (2019). Molecular Transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Central Science, 5(9), 1572–1583.

[27] Bjerrum, E. J. (2017). SMILES enumeration as data augmentation for neural network modeling of molecules. arXiv:1703.07076.

[28] Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Clevert, D.-A., & Hochreiter, S. (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chemical Science, 9(24), 5441–5451.

[29] Deng, J., Yang, Z., Wang, H., Ojima, I., Samaras, D., & Wang, F. (2023). A systematic study of key elements underlying molecular property prediction. Nature Communications, 14(1), 6395. https://doi.org/10.1038/s41467-023-41948-6

[30] Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., & Tang, J. (2022). Pre-training molecular graph representation with 3D geometry (GraphMVP). International Conference on Learning Representations 2022.

[31] Lou, X., Cai, J., & Siu, S. W. I. (2026). Multi-MoleScale: a multi-scale approach for molecular property prediction with graph contrastive and sequence learning. Journal of Cheminformatics, 18(1), Article 01126. https://doi.org/10.1186/s13321-025-01126-w

[32] Liu, Z., Ma, P., Wang, Y., Matusik, W., & Tegmark, M. (2025). Kolmogorov–Arnold networks meet science. Physical Review X, 15(4), 041051. https://doi.org/10.1103/4t7t-v19l

[33] Bozorgasl, Z., & Chen, H. (2024). Wav-KAN: wavelet Kolmogorov–Arnold networks. arXiv:2405.12832.

[34] Landrum, G. (2024). RDKit: open-source cheminformatics software, release 2024.03.6. https://www.rdkit.org

[35] Bemis, G. W., & Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. Journal of Medicinal Chemistry, 39(15), 2887–2893. https://doi.org/10.1021/jm9602928

[36] Guo, Q., Hernandez-Hernandez, S., & Ballester, P. J. (2024). Scaffold splits overestimate virtual screening performance. arXiv:2406.00873.

[37] Polishchuk, P. G., Madzhidov, T. I., & Varnek, A. (2013). Estimation of the size of drug-like chemical space based on GDB-17 data. Journal of Computer-Aided Molecular Design, 27(8), 675–679.

[38] Mandal, M., Zhu, Z., Cumming, J. N., Liu, X., Strickland, C., Mazzola, R. D., Caldwell, J. P., Leach, P., Grzelak, M., Hyde, L., Zhang, Q.,

Journal of BioDigital Intelligence and Technology
journal homepage: www.qnapub.com
Contents lists available at Quantum Nexus Academic Publishers
ISSN: Under Process DoI: Under Process

Terracina, G., Zhang, L., Chen, X., Kuvelkar, R., Kennedy, M. E., Favreau, L., Cox, K., Orth, P., & Strickland, C. (2012). Design and validation of bicyclic iminopyrimidinones as beta-amyloid cleaving enzyme-1 (BACE1) inhibitors. Journal of Medicinal Chemistry, 55(21), 9331–9345.

[39] Delaney, J. S. (2004). ESOL: estimating aqueous solubility directly from molecular structure. Journal of Chemical Information and Computer Sciences, 44(3), 1000–1005. https://doi.org/10.1021/ci034243x

[40] Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., & Zitnik, M. (2021). Therapeutics Data Commons: machine learning datasets and tasks for drug discovery and development. NeurIPS Datasets and Benchmarks Track 2021.

[41] Irwin, J. J., Tang, K. G., Young, J., Dandarchuluun, C., Wong, B. R., Khurelbaatar, M., Moroz, Y. S., Mayfield, J., & Sayle, R. A. (2020). ZINC20 — a free ultralarge-scale chemical database for ligand discovery. Journal of Chemical Information and Modeling, 60(12), 6065–6073.

[42] Schütt, K. T., Unke, O. T., & Gastegger, M. (2021). Equivariant message passing for the prediction of tensorial properties and molecular spectra (PaiNN). Proceedings of the 38th International Conference on Machine Learning, 9377–9388.

[43] Ramakrishnan, R., Dral, P. O., Rupp, M., & von Lilienfeld, O. A. (2014). Quantum chemistry structures and properties of 134 kilo molecules. Scientific Data, 1, 140022. https://doi.org/10.1038/sdata.2014.22

[44] Jiang, D., Wu, Z., Hsieh, C.-Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., & Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. Journal of Cheminformatics, 13(1), 12.

[45] Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., & Leskovec, J. (2020). Strategies for pre-training graph neural networks. International Conference on Learning Representations 2020.