Journal of Cheminformatics

RESEARCH Open Access

Multimodal graph fusion with statistically guided parsimonious descriptor selection for molecular property prediction

Yoonsuk Jang[1], Juyeon Lee[2], Keunhong Jeong[3*] and Jaeoh Kim[1*]

Abstract

Graph convolutional networks (GCN) are effective for learning molecular representations, but their reliance on local message passing and simple feature concatenation limits their ability to capture global physicochemical properties. We present KROnecker-product based multimodal fusion with Variable sElection for eXpressive molecular representation learning (KROVEX), a method that integrates graph embeddings with molecular descriptors through a Kronecker-product to explicitly model second-order interactions. Informative descriptors are identified using a two-stage procedure that combines iterative sure independence screening with Elastic Net regularization. The proposed approach was evaluated on two benchmark datasets (FreeSolv and ESOL) as well as two self-curated datasets with vapor pressure and aqueous solubility as the target property. Overall, our method outperformed not only GCN but also fusion-based baselines such as EGCN, D-MPNN, and BAN under both the random and scaffold split. More importantly, the fusion operates at the final embedding level, enabling consistent performance across different GNN backbones (e.g., GAT and GIN). KROVEX achieves state-of-the-art performance on vapor pressure prediction, establishing a new benchmark for this safety-critical property essential for environmental monitoring and industrial process design. Ablation studies further demonstrated that (1) statistically guided descriptor selection yields more informative features than predefined descriptors, and (2) Kronecker-product fusion provides greater improvements than simple concatenation as the number of descriptors increases. These results demonstrate that parsimonious descriptor selection combined with multimodal graph fusion enhances predictive performance and interpretability, providing a generalizable framework for molecular property prediction.

Scientific contribution

We present a novel multimodal fusion framework, KROVEX, which integrates graph embeddings with statistically selected molecular descriptors through Kronecker-product fusion. Grounded in the statistical philosophy of parsimony, our two-stage variable selection strategy not only enhances interpretability and generalization performance but also ensures compatibility with graph neural network architectures. Beyond benchmarks, KROVEX offers a principled and theoretically grounded alternative to conventional concatenation strategies in molecular machine learning.

*Correspondence:
Keunhong Jeong
doas1mind@gmail.com
Jaeoh Kim
jaeoh.k@inha.ac.kr
Full list of author information is available at the end of the article

Keywords Descriptor selection, Graph convolutional networks, Molecular property prediction, Multimodal fusion, Vapor pressure

Introduction

Graph convolutional networks (GCN) [1] have recently attracted considerable attention as a powerful approach that overcomes the limitations of traditional deep learning models restricted to Euclidean data. The key idea of GCN is to update node representations by aggregating information from neighbors through matrix operations. This allows GCN to handle irregular relationships and to learn efficiently on non-Euclidean structured data. This mechanism has led to widespread applications in material science [2, 3], social networks [4–6], and citation networks [7–9]. Notably, their application in chemistry has attracted researchers' extensive attention since molecular structures can be naturally represented as graphs. For example, in drug discovery [10, 11], molecular property prediction [12, 13], and protein interaction prediction [14].

Despite their success in chemistry, GCN struggles to fully represent complex physicochemical properties due to inherent limitations of message-passing architectures. Because of the shallow depth of GCN layers, global molecular structures are often overlooked [15–18]. Hierarchical structures of molecules are ignored by aggregation methods such as summation or averaging [19, 20]. Although several studies have attempted to address these issues—for example, FraGAT [21] encodes unified graph representation by extracting multi-scale representations from molecule graph fragments, EGCN [16] concatenates predefined descriptor with graph embeddings, CSGL [22] incorporates chemical synthesis routes to construct synthesis-aware molecular graphs, GROVER [23] employs large-scale self-supervised graph transformers to learn general-purpose molecular embeddings, and HieGT [24] models both atom-level and motif-level structures through a hierarchical graph transformer to capture local–global chemical patterns—these methods focus solely on enhancing graph-based representation learning or rely on simple concatenation fusion strategies.

Multimodal artificial intelligence (AI) is an expanding field that focuses on developing models capable of processing and integrating information from multiple sources [25]. By fusing heterogeneous inputs, multimodal AI seeks to create unified representations that capture richer insights than any single modality. In recent years, the integration of complementary information from different modalities, such as visual, textual, and audio, has made significant progress in areas such as emotion recognition, vision-language navigation, and human-computer interaction [26–28]. From a perspective of multimodal AI, the aforementioned limitations may be addressed by integrating information that cannot be captured by graph structures alone.

In molecular property prediction, however, multimodal approaches typically rely on simple vector concatenation [16, 29, 30]. This naive fusion fails to explicitly model interactions between modalities, limiting both the model's ability to capture complex relationships and its interpretability [31–33]. In contrast to simple concatenation, bilinear forms provide a principled way to capture cross-modal interactions that explain context-dependent behavior not recoverable by concatenation [34, 35]. Notably, such bilinear interactions can be equivalently represented as linear predictors on the Kronecker feature space, thereby enabling standard optimization and theoretical analysis while offering strictly greater expressive power. Beyond concatenation, Kronecker-product fusion can be viewed as a structured feature expansion equivalent to the degree-2 polynomial kernel restricted to cross-modality terms [36]. This perspective suggests that Kronecker-product fusion provides a theoretically grounded manner for modeling bilinear interactions between modalities, offering a clear link to classical polynomial kernel methods while avoiding redundant within-modality expansions. While Kronecker-product fusion enriches the representational capacity compared to concatenation [37, 38], it is equally important to ensure that such expressiveness does not lead to uncontrolled model complexity. To this end, a Rademacher complexity bound can be introduced, suggesting that the bilinear predictor class induced by Kronecker-product fusion admits generalization guarantees under norm constraints.

A variety of molecular descriptors are available, each providing distinct insights into molecular properties; however, Na et al. [16] integrated only three predefined descriptors into the model through simple concatenation. In contrast, the RDKit library [39] provides more than 200 descriptors, suggesting that incorporating a broader array could enhance the learning capabilities of GCN regarding molecular graph representations. However, employing all available descriptors can significantly complicate the model, leading to overfitting, multicollinearity, and high computational costs, which compromise predictive accuracy and generalizability. To avoid these issues while providing interpretability, a rational approach is required. Variable selection is a well-established statistical methodology, whose effect has been extensively documented in the literature [40–42]. Test-based selection is the most classical way of selecting variables in statistical models [43]. Since the introduction of LASSO [44], penalty-based methods have become a standard strategy in modern machine learning. By shrinking coefficients through penalty terms, they constrain model complexity and reduce variance at the cost of introducing a small bias, thereby improving the generalization performance of the model [45–47]. Screening-based methods have proven effective in high-dimensional settings. The earliest of these, Sure Independence Screening (SIS) [48], filters out irrelevant variables based on marginal correlations and has inspired several improved versions in subsequent studies [49, 50]. Crucially, selecting a compact subset of variables also induces an implicit low-rank effect on bilinear representation, thereby acting as a structural regularizer [51]. Guided by these statistical methodologies, informative descriptors can be extracted for multimodal fusion. Nevertheless, such approaches have rarely been explored in molecular property prediction.

Based on these considerations, we propose KROnecker-product based multimodal fusion with Variable sElection for eXpressive molecular representation learning (KROVEX), a method that integrates graph embeddings with molecular descriptors through a Kronecker-product. To the best of our knowledge, this is the first application of Kronecker-product fusion with variable selection for molecular property prediction. Notably, the fusion operates on the final embedding itself, suggesting that our approach is not tied to a specific graph encoder but can be adopted across different backbone architectures. Figure 1 illustrates overall workflow of KROVEX.

The main contributions of our study are as follows:

1. We introduce a novel multimodal fusion framework that integrates graph embeddings with statistically selected molecular descriptors through a Kronecker-product fusion.
2. We provide theoretical justification by interpreting Kronecker-product fusion as a polynomial kernel and establishing generalization guarantees via Rademacher complexity.
3. We demonstrate the effectiveness and generalizability of our model through comprehensive experiments on benchmark datasets and two self-curated datasets, under both the random and scaffold splits.

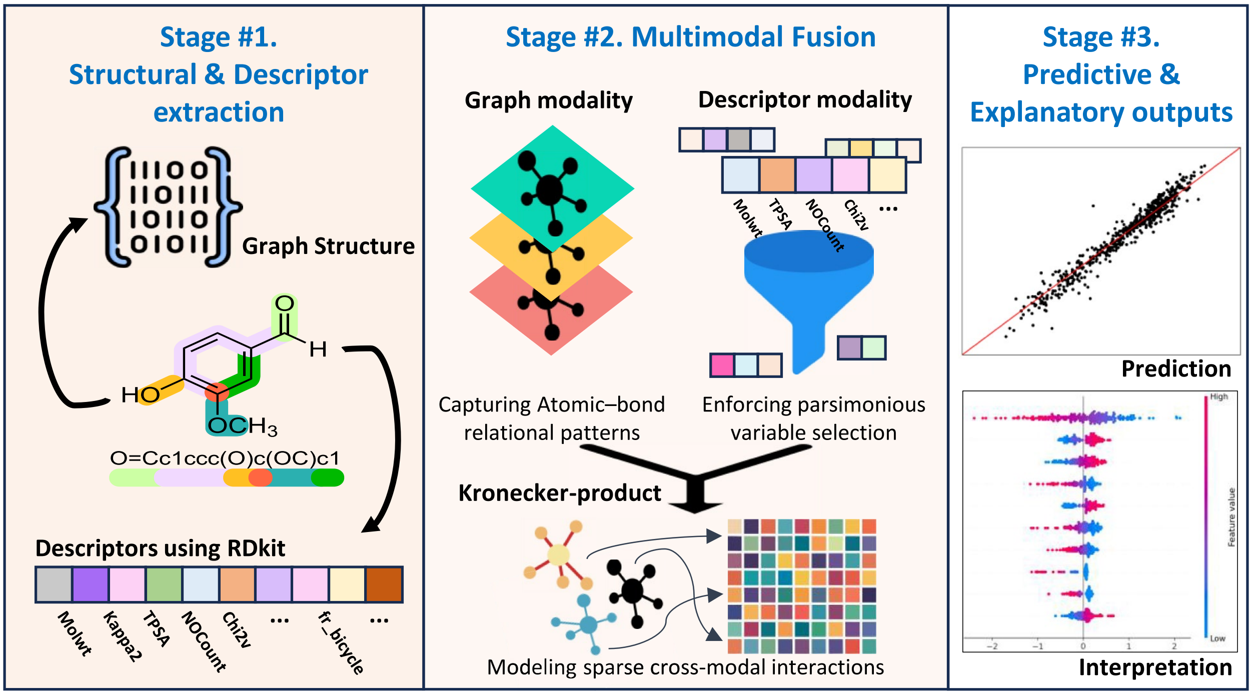


Fig. 1 Overall workflow of KROVEX. The model comprises three stages. Stage 1: Structural and Descriptor extraction from SMILES strings to construct molecular graphs and compute physicochemical descriptors. Stage 2: Multimodal Fusion between graph and descriptor modalities via Kronecker-product, enabling structured cross-modal interactions. Stage 3: Predictive and Explanatory outputs, where the model delivers accurate property predictions while providing chemical interpretation through descriptor contributions

The rest of this paper is organized as follows. "Methodology" section begins with the preliminaries for the proposed models and details our methodologies. "Experimental results" section reports the experimental results and the model's generalizability on two self-curated datasets. "Discussion" section discusses the implications and limitations of our findings. Finally, "Conclusions" section summarizes the key contributions of our study and offers concluding remarks.

Methodology

Let $d, k \in \mathbb{N}$. For $\mathbf{h} \in \mathbb{R}^d$ and $\mathbf{z} \in \mathbb{R}^k$, write the concatenation as $[\mathbf{h}; \mathbf{z}] \in \mathbb{R}^{d+k}$. For matrices U, V of the same size, we denote the Frobenius inner product by $\langle U, V \rangle_F := \mathrm{tr}(\mathrm{U}^\top \mathrm{V})$. Let $G = (V, E, X)$ be a molecular graph, where $V = \{v_i\}_{i=1}^N$ is the set of nodes (atoms) and $e_{ij} = (v_i, v_j) \in E$ is an edge (chemical bond) between nodes v_i and v_j with $|E| = M$. $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times D}$ is the node feature matrix. Let $A \in \{0,1\}^{N \times N}$ be the adjacency matrix, where $A_{ij} = 1$ if nodes v_i and v_j are connected, and $A_{ij} = 0$ otherwise. A graph convolutional layer updates node features as follows:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}(A + I)\tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}), \qquad (1)$$

where $H^{(l)}$ is the input node representation at layer l with $H^{(0)} = X$, and $\tilde{D}$ is the degree matrix with $\tilde{D}_{ii} = \sum_j (A + I)_{ij}$. $W^{(l)} \in \mathbb{R}^{D_l \times D_{l+1}}$ is the learnable weight matrix, and $\sigma(\cdot)$ is an activation function. For matrices $P \in \mathbb{R}^{m \times n}$, and $Q \in \mathbb{R}^{p \times q}$, their Kronecker-product $P \otimes Q \in \mathbb{R}^{mp \times nq}$ is defined as:

$$P \otimes Q \triangleq \begin{pmatrix} p_{11}Q & \cdots & p_{1n}Q \\ \vdots & \ddots & \vdots \\ p_{m1}Q & \cdots & p_{mn}Q \end{pmatrix}$$

where p_{ij} is the (i, j)-th element of P. Here, vectorization $\mathrm{vec}(\cdot)$ stacks columns, and for vectors u, v, we use the identity

$$\mathrm{vec}(\mathrm{uv}^\top) = \mathrm{v} \otimes \mathrm{u}.$$

Architecture

Figure 2 presents the detailed architecture of KROVEX, from graph construction to descriptor selection, multimodal fusion, and final prediction. It consists of five major components: molecular graph representation, GCN-based embedding extraction, descriptor selection, multimodal fusion, and prediction. While the input molecule is first represented as a graph processed through standard GCN layers to produce a graph embedding $\mathbf{h}_G$, the novelty of our approach lies in (1) statistically guided descriptor selection and (2) Kronecker-product based multimodal fusion. This enhanced GCN architecture

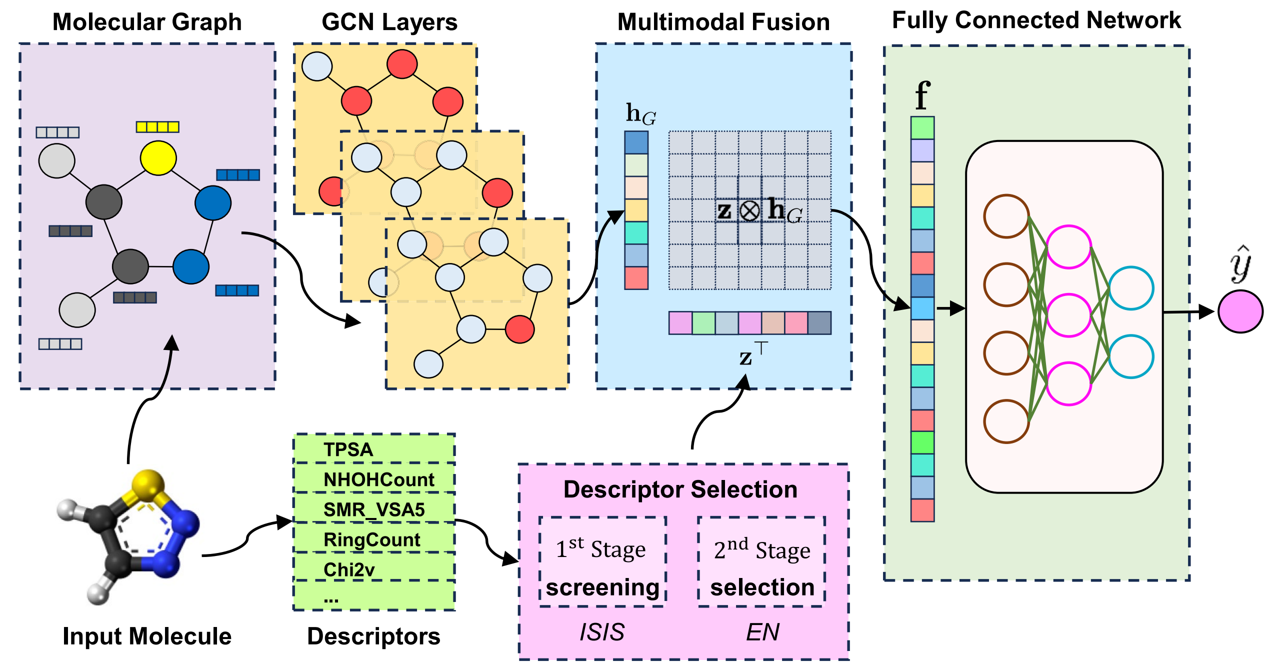


Fig. 2 The overall architecture of the proposed model. The model consists of five major components. (1) Molecular Graph Representation to encode atoms and bonds as a graph. (2) GCN-based Embedding Extraction to yield a graph embedding $\mathbf{h}_G$. (3) Descriptor Selection to extract an informative and parsimonious subset of descriptors $\mathbf{z}$. (4) Multimodal Fusion to integrate the graph embedding $\mathbf{h}_G$ with descriptors $\mathbf{z}$. (5) Prediction to yield the final prediction $\hat{y}$. ISIS refers to Iterative Sure Independence Screening; EN denotes Elastic Net regularization

explicitly models second-order feature interactions, leading to richer molecular representations.

While RDKit offers hundreds of molecular descriptors, naively including all of them risks overfitting, multicollinearity, and unnecessary complexity. We therefore employ a statistically principled variable selection procedure (e.g., Iterative Sure Independence Screening and Elastic Net) to extract an informative and parsimonious subset $\mathbf{z}$, improving interpretability and efficiency while implicitly regularizing the subsequent bilinear representation.

To integrate graph embeddings $\mathbf{h}_G$ with the selected descriptors $\mathbf{z}$, we construct a Kronecker feature map

$$\phi(\mathbf{h}_G, \mathbf{z}) := \mathrm{vec}(\mathbf{h}_G \mathbf{z}^\top) = \mathbf{z} \otimes \mathbf{h}_G, \tag{2}$$

which explicitly models cross-modal interactions. This operation is theoretically equivalent to a degree-2 polynomial kernel restricted to cross-modality terms, thereby enhancing expressiveness while avoiding redundant expansions within a single modality. The fused representation thus enables richer modeling of structure–descriptor relationships than simple concatenation, and it admits norm-based generalization guarantees via Rademacher complexity bounds. Additionally, because the fusion applies solely to the graph embedding $\mathbf{h}_G$ and forms multiplicative interactions with the descriptor $\mathbf{z}$, variations in the underlying message-passing scheme do not substantially change how the two information sources are combined.

The fused feature $\phi(\mathbf{h}_G, \mathbf{z})$ is then passed into a fully connected network to yield the final prediction $\hat{y}$. Here, the low-rank effect induced by descriptor selection acts as a structural regularizer, supporting both accuracy and generalization.

The algorithmic workflow is summarized in Algorithm 1.

Algorithm 1 The specific algorithm for multimodal graph fusion with descriptor selection through a Kronecker-product

```
Input: Molecular graph $G = (V, E, X)$
       Initial node embeddings $\mathbf{h}_v^{(0)} \leftarrow \mathbf{x}_v$
       Selected descriptor vector $\mathbf{z}$
Output: Predicted value $\hat{y}$ of molecular property
for $l \in \{0, 1, \dots, L-1\}$ do
    // Iterative message-passing
    $\mathbf{m}_v^{(l)} \leftarrow \mathrm{AGGREGATE}(\{\mathbf{h}_u^{(l)} : u \in \mathcal{N}(v)\})$
    $\mathbf{h}_v^{(l+1)} \leftarrow \mathrm{UPDATE}(\mathbf{h}_v^{(l)}, \mathbf{m}_v^{(l)})$
end for
$\mathbf{h}_G \leftarrow \mathrm{READOUT}(\{\mathbf{h}_v^{(L)} | v \in V\})$
// Kronecker-product fusion
$\mathbf{f} = \mathbf{z} \otimes \mathbf{h}_G$
$\hat{y} \leftarrow$ fully-connected network($\mathbf{f}$)
return $\hat{y}$
```

Descriptor selection

Let $Z \in \mathbb{R}^{N \times p_0}$ be the raw descriptor matrix, where Z_{ij} is the value of the j-th descriptor for the i-th molecule, and $p_0 = 209$ is the number of descriptors initially extracted using RDKit. Prior to selection, Z is preprocessed by removing descriptors with missing values or near-zero variance, and is standardized to zero mean and unit variance.

A two-stage selection procedure is then performed, combining screening-based and penalty-based methods.

In the first stage, Iterative Sure Independence Screening (ISIS) [48] is applied to filter out unimportant descriptors by iterating correlation learning and residual fitting. This step ensures that relevant descriptors are retained while reducing the dimensionality from p_0 to a smaller $p < p_0$. In the second stage, Elastic Net (EN) regularization [45] is employed to further remove irrelevant descriptors from the reduced $Z \in \mathbb{R}^{N \times p}$. EN combines the strengths of both LASSO (L_1) and Ridge (L_2) penalties such as sparsity-inducing and multicollinearity-handling. It solves the following optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - Z\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2, \tag{3}$$

where $\beta \in \mathbb{R}^p$ is the coefficient vector, $\mathbf{y} \in \mathbb{R}^N$ is the target vector, and tuning parameters λ_1 and λ_2 are selected via cross-validation. The L_1 and L_2 penalties are defined as:

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|, \quad \|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2.$$

When $\lambda_2 > 0$, the objective in (3) is strictly convex, yielding a unique minimizer under standard conditions; in practice this encourages sparsity while mitigating multicollinearity. By reducing the descriptor space from p_0 to $k \ll p_0$, the procedure induces an implicit low-rank effect on the subsequent bilinear representation. Further details are provided in "Multimodal fusion" section. Finally, we obtain a selected molecular descriptor vector $\mathbf{z} = [z_1, \dots, z_k]^\top \in \mathbb{R}^k$, where $k \leq p$ is the number of descriptors with non-zero coefficients after EN regularization.

Multimodal fusion

We construct a molecular graph G from SMILES. The node feature matrix $X \in \mathbb{R}^{N \times D}$ is defined, where each feature vector $\mathbf{x}_i \in \mathbb{R}^D$ contains basic atomic properties such as atomic weight, volume, radius, etc. Prior to forming the node feature matrix, these properties were standardized to zero mean and unit variance. Following Eq. (1), X is updated through two graph convolutional layers as:

$$H^{(1)} = \mathrm{ReLU}(\tilde{A}\mathrm{X}\mathrm{W}^{(0)})$$
$$H^{(2)} = \mathrm{ReLU}(\tilde{A}\mathrm{H}^{(1)}\mathrm{W}^{(1)}),$$

where $\tilde{A} = \tilde{D}^{-\frac{1}{2}}(A+I)\tilde{D}^{-\frac{1}{2}}$ is the normalized adjacency matrix. From the final layer $H^{(2)}$, we obtain the graph embedding $\mathbf{h}_G = [h_1, \ldots, h_d]^\top \in \mathbb{R}^d$ by averaging all node representations:

$$\mathbf{h}_G = \frac{1}{N}\sum_{i=1}^{N} \mathbf{h}_i, \tag{4}$$

where $\mathbf{h}_i = H^{(2)}_{i,:}$ denotes the representation of node v_i at the final GCN layer.

To integrate graph embedding (4) with molecular descriptors $\mathbf{z}$, concatenation can be employed; however, this fusion fails to explicitly model interactions between modalities (Proposition 1).

Proposition 1

(*Limitation of Concatenation in Expressing Cross-Interactions*) *Fix* $S \in \mathbb{R}^{d\times k}$ *and define the bilinear map*

$$f_S(\mathbf{h}_G, \mathbf{z}) := \langle S, \mathbf{h}_G\mathbf{z}^\top\rangle_F = \mathbf{h}_G^\top S\,\mathbf{z},\ (\mathbf{h}_G, \mathbf{z}) \in \mathbb{R}^d \times \mathbb{R}^k.$$

Consider the family of all affine maps on the concatenation,

$$\mathcal{G} = \left\{g(\mathbf{h}_G, \mathbf{z}) = \mathbf{a}^\top\mathbf{h}_G + \mathbf{b}^\top\mathbf{z} + c\right\},$$

where $\mathbf{a} \in \mathbb{R}^d$, $\mathbf{b} \in \mathbb{R}^k$, *and* $c \in \mathbb{R}$. *If* $S \neq 0$, *then there is no* $g \in \mathcal{G}$ *such that* $g(\mathbf{h}_G, \mathbf{z}) = f_S(\mathbf{h}_G, \mathbf{z})$ *for all* $(\mathbf{h}_G, \mathbf{z}) \in \mathbb{R}^d \times \mathbb{R}^k$. *Equivalently,* f_S *is not representable by any affine function of the concatenated features unless* $S = 0$.

Proposition 1 shows that simple concatenation cannot represent bilinear cross-interactions between modalities. Motivated by this limitation, we integrate the graph embedding with the molecular descriptor through a Kronecker-product:

$$\mathbf{f} = \mathbf{z} \otimes \mathbf{h}_G \in \mathbb{R}^{dk}, \tag{5}$$

producing a fused representation, which is passed into the fully connected network to yield the final prediction $\hat{y}$.

This Kronecker-product based approach captures second-order feature interactions between modalities, thereby providing a more expressive fusion mechanism for molecular property prediction. The following lemmas and theorems establish its theoretical foundation.

Lemma 1

(*Equivalence between Bilinear Forms and Kronecker Features*) *For any* $W \in \mathbb{R}^{d\times k}$ *and any* $(\mathbf{h}_G, \mathbf{z}) \in \mathbb{R}^d \times \mathbb{R}^k$,

$$\begin{aligned}\langle W, \mathbf{h}_G\mathbf{z}^\top\rangle_F &= \mathrm{tr}(\mathrm{W}^\top\mathbf{h}_\mathrm{G}\mathbf{z}^\top)\\ &= \mathbf{h}_G^\top W\,\mathbf{z}\\ &= \mathrm{vec}(\mathrm{W})^\top\mathrm{vec}(\mathbf{h}_\mathrm{G}\mathbf{z}^\top)\\ &= \mathrm{vec}(\mathrm{W})^\top(\mathbf{z} \otimes \mathbf{h}_\mathrm{G}).\end{aligned}$$

Hence a linear predictor on the Kronecker feature map (2) *is exactly a bilinear form in* $(\mathbf{h}_G, \mathbf{z})$, *with parameter vector* $\mathrm{vec}(\mathrm{W})$.

Lemma 1 demonstrates that every bilinear form $\mathbf{h}_G^\top W\mathbf{z}$ is equivalent to a linear predictor on the Kronecker feature map (2). This equivalence implies that second-order interactions between graph embeddings and molecular descriptors can be modeled within a standard linear framework, thereby enabling convex optimization and theoretical analysis while retaining strictly greater representational capacity than concatenation. Moreover, the Kronecker feature map induces a kernel equivalent to the degree-2 polynomial kernel restricted to cross-modal interactions (Lemma 2).

Lemma 2

(*Equivalence between Kronecker Features and Degree-*2 *Cross-Polynomial Kernel*) *Recall the Kronecker feature map*

$$\phi(\mathbf{h}_G, \mathbf{z}) := \mathrm{vec}(\mathbf{h}_G\mathbf{z}^\top) = \mathbf{z} \otimes \mathbf{h}_\mathrm{G} \ \in \mathbb{R}^{\mathrm{dk}}.$$

Then the induced kernel $K_\times\big((\mathbf{h}_G, \mathbf{z}), (\mathbf{h}'_G, \mathbf{z}')\big) := \langle\phi(\mathbf{h}_G, \mathbf{z}), \phi(\mathbf{h}'_G, \mathbf{z}')\rangle$ *satisfies*

$$K_\times\big((\mathbf{h}_G, \mathbf{z}), (\mathbf{h}'_G, \mathbf{z}')\big) = \langle\mathbf{h}_G, \mathbf{h}'_G\rangle\,\langle\mathbf{z}, \mathbf{z}'\rangle.$$

Moreover, letting $x := [\mathbf{h}_G; \mathbf{z}] \in \mathbb{R}^{d+k}$ *and* $x' := [\mathbf{h}'_G; \mathbf{z}']$ *and denoting the homogeneous degree-*2 *polynomial kernel by* $K_2(u, u') = (u^\top u')^2$, *we have the identity*

$$\begin{aligned}&\langle\mathbf{h}_G, \mathbf{h}'_G\rangle\,\langle\mathbf{z}, \mathbf{z}'\rangle\\ &\quad= \tfrac{1}{2}\Big\{K_2(x, x') - K_2(\mathbf{h}_G, \mathbf{h}'_G) - K_2(\mathbf{z}, \mathbf{z}')\Big\}.\end{aligned}$$

Hence $K_\times$ *coincides with the degree-*2 *polynomial kernel on the concatenated input restricted to cross-modality monomials* $\{h_i z_j\}$, *i.e., it excludes within-modality terms such as* $h_i h'_{i'}$ *and* $z_j z'_{j'}$.

Table 1 Predictive performance comparison of our model versus competing models for FreeSolv datasets, demonstrating the effectiveness of descriptor selection and Kronecker-product fusion

# of Descriptors[1]	GCN [1]	EGCN [16]	EGCN [16] with DS[2]	D-MPNN [58]	BAN [59]	KROVEX (ours)
0	8.020 ± 0.517	–	–	1.341 ± 0.086	–	–
	2.007 ± 0.050	–	–	0.773 ± 0.021	–	–
	15.186 ± 2.114	–	–	5.227 ± 0.883	–	–
	2.840 ± 0.186	–	–	1.558 ± 0.096	–	–
1		6.238 ± 0.384	–	–	–	–
	–	1.619 ± 0.039	–	–	–	–
	–	16.030 ± 3.210	–	–	–	–
	–	2.846 ± 0.287	–	–	–	–
2		6.360 ± 0.333	–	–	–	–
	–	1.745 ± 0.034	–	–	–	–
	–	14.580 ± 3.436	–	–	–	–
	–	2.535 ± 0.271	–	–	–	–
3		5.713 ± 0.360	5.070 ± 0.210	–	–	3.498 ± 0.178
	–	1.595 ± 0.026	1.458 ± 0.032	–	–	1.340 ± 0.026
	–	14.624 ± 3.077	11.808 ± 2.027	–	–	9.277 ± 1.413
	–	2.649 ± 0.279	2.802 ± 0.185	–	–	2.527 ± 0.205
5		–	2.002 ± 0.124	–	–	1.401 ± 0.088
	–	–	1.007 ± 0.023	–	–	0.827 ± 0.023
	–	–	5.544 ± 1.109	–	–	3.704 ± 0.501
	–	–	1.714 ± 0.144	–	–	1.490 ± 0.098
7		–	1.664 ± 0.068	–	–	1.310 ± 0.073
	–	–	0.963 ± 0.020	–	–	0.827 ± 0.023
	–	–	3.914 ± 0.512	–	–	7.927 ± 4.344
	–	–	1.556 ± 0.103	–	–	1.637 ± 0.191
10		–	1.516 ± 0.061	–	–	1.230 ± 0.076
	–	–	0.909 ± 0.020	–	–	0.765 ± 0.015
	–	–	2.903 ± 0.275	–	–	4.963 ± 2.027
	–	–	1.287 ± 0.061	–	–	1.474 ± 0.183
20		–	1.262 ± 0.076	–	–	1.107 ± 0.068
	–	–	0.777 ± 0.024	–	–	0.626 ± 0.020
	–	–	3.280 ± 0.496	–	–	2.774 ± 0.449
	–	–	1.285 ± 0.096	–	–	1.255 ± 0.132
50		–	1.120 ± 0.069	–	–	**0.973 ± 0.072**
	–	–	0.640 ± 0.018	–	–	**0.597 ± 0.014**
	–	–	3.884 ± 0.919	–	–	**2.606 ± 0.427**
	–	–	1.292 ± 0.154	–	–	1.141 ± 0.076
ALL[3]		–	–	1.034 ± 0.133	1.025 ± 0.082	–
	–	–	–	0.620 ± 0.024	0.599 ± 0.015	–
	–	–	–	4.446 ± 0.421	3.246 ± 0.458	–
	–	–	–	1.481 ± 0.113	**1.126 ± 0.111**	–

KROVEX showed consistently strong performance on most metrics

The first two values represent MSE and MAE under the random split, while the last two values represent MSE and MAE under the scaffold split. All experiments were repeated 15 times, and the results are presented as the means ± standard error for each metric. The best and second-best performances for each metric are marked bold and underlined

[1] Indicates the number of descriptors integrated into the model

[2] Indicates descriptor selection

[3] Denotes the complete descriptor set for each model (200 for D-MPNN and 208 for BAN)

Table 2 Predictive performance comparison of our model versus competing models for ESOL datasets, demonstrating the effectiveness of descriptor selection and Kronecker-product fusion

# of Descriptors[1]	GCN [1]	EGCN [16]	EGCN [16] with DS[2]	D-MPNN [58]	BAN [59]	KROVEX (ours)
0	3.317 ± 0.080	–	–	0.438 ± 0.011	–	–
	1.475 ± 0.015	–	–	0.469 ± 0.018	–	–
	5.426 ± 1.019	–	–	0.742 ± 0.032	–	–
	1.747 ± 0.112	–	–	**0.620 ± 0.017**	–	–
1		2.015 ± 0.077	–	–	–	–
	–	1.092 ± 0.013	–	–	–	–
	–	4.115 ± 1.085	–	–	–	–
	–	1.331 ± 0.083	–	–	–	–
2		0.960 ± 0.027	–	–	–	–
	–	0.774 ± 0.021	–	–	–	–
	–	1.564 ± 0.131	–	–	–	–
	–	0.966 ± 0.026	–	–	–	–
3		0.918 ± 0.028	0.686 ± 0.030	–	–	0.637 ± 0.019
	–	0.761 ± 0.013	0.607 ± 0.009	–	–	0.590 ± 0.012
	–	1.512 ± 0.150	1.085 ± 0.168	–	–	1.034 ± 0.111
	–	0.937 ± 0.032	0.733 ± 0.036	–	–	0.745 ± 0.037
5		–	0.606 ± 0.024	–	–	0.592 ± 0.024
	–	–	0.580 ± 0.008	–	–	0.567 ± 0.012
	–	–	0.868 ± 0.088	–	–	0.984 ± 0.090
	–	–	0.702 ± 0.032	–	–	0.717 ± 0.030
7		–	0.554 ± 0.012	–	–	0.492 ± 0.015
	–	–	0.536 ± 0.008	–	–	0.520 ± 0.010
	–	–	0.829 ± 0.092	–	–	1.148 ± 0.204
	–	–	0.670 ± 0.024	–	–	0.687 ± 0.022
10		–	0.569 ± 0.016	–	–	0.460 ± 0.011
	–	–	0.538 ± 0.007	–	–	0.498 ± 0.006
	–	–	0.839 ± 0.050	–	–	0.964 ± 0.090
	–	–	0.676 ± 0.022	–	–	0.740 ± 0.032
20		–	0.470 ± 0.010	–	–	0.453 ± 0.023
	–	–	0.508 ± 0.007	–	–	0.484 ± 0.007
	–	–	0.794 ± 0.069	–	–	0.874 ± 0.076
	–	–	0.640 ± 0.023	–	–	0.693 ± 0.022
63		–	0.477 ± 0.015	–	–	**0.423 ± 0.022**
	–	–	0.479 ± 0.008	–	–	**0.469 ± 0.013**
	–	–	0.785 ± 0.068	–	–	**0.730 ± 0.054**
	–	–	0.659 ± 0.022	–	–	0.628 ± 0.019
ALL[3]		–	–	0.432 ± 0.032	0.441 ± 0.010	–
	–	–	–	0.479 ± 0.018	0.485 ± 0.005	–
	–	–	–	0.795 ± 0.017	0.775 ± 0.053	–
	–	–	–	0.637 ± 0.018	0.651 ± 0.024	–

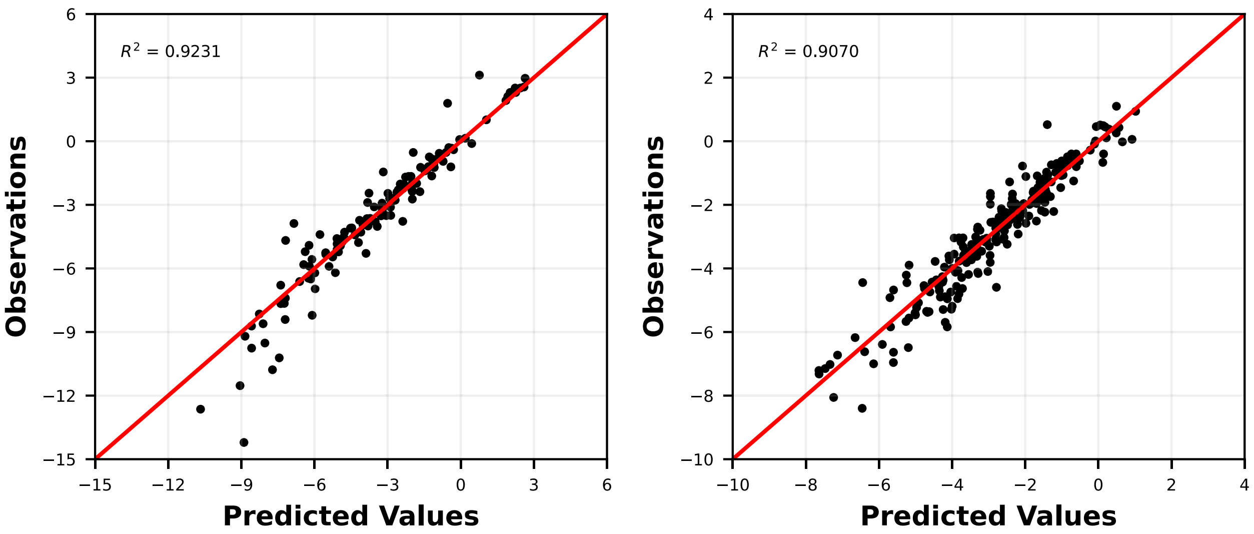
KROVEX outperformed competitors with the best predictive performance

The first two values represent MSE and MAE under the random split, while the last two values represent MSE and MAE under the scaffold split. All experiments were repeated 15 times, and the results are presented as the means ± standard error for each metric. The best and second-best performances for each metric are marked bold and underlined

[1] Indicates the number of descriptors integrated into the model

[2] Indicates descriptor selection

[3] Denotes the complete descriptor set for each model (200 for D-MPNN and 208 for BAN.)



(a) Freesolv

(b) ESOL

Fig. 3 Scatter plots of true versus predicted values for the two test datasets

Lemma 2 demonstrates that Kronecker-product fusion can be interpreted as modeling precisely the bilinear cross-terms between $\mathbf{h}_G$ and $\mathbf{z}$, while excluding within-modality quadratic terms. This characterization not only clarifies the representational advantage over simple concatenation but also connects our approach to the classical polynomial kernel framework, thereby providing a principled mathematical foundation for the proposed model. While Kronecker-product fusion enriches expressiveness, model complexity must be controlled. To this end, we introduce the Rademacher complexity (6) and establish a Rademacher bound (Theorem 1). We first denote the Euclidean inner product by $\langle u, v\rangle$. For $W \in \mathbb{R}^{d\times k}$, we write $\|W\|_F$ and $\|W\|_*$ for the Frobenius and nuclear norms, respectively. Given samples $\{(\mathbf{h}_{G_i}, \mathbf{z}_i)\}_{i=1}^{n}$ with $\mathbf{h}_{G_i} \in \mathbb{R}^d$ and $\mathbf{z}_i \in \mathbb{R}^k$, we assume boundedness $\|\mathbf{h}_{G_i}\| \le B_h$ and $\|\mathbf{z}_i\| \le B_z$ almost surely. Given a function class $\mathcal{F}$ on pairs $(\mathbf{h}_G, \mathbf{z})$ and a sample $S = \{(\mathbf{h}_{G_i}, \mathbf{z}_i)\}_{i=1}^{n}$, its empirical Rademacher complexity is

$$\widehat{\mathfrak{R}}_n(\mathcal{F}; S) := \mathbb{E}_{\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(\mathbf{h}_{G_i}, \mathbf{z}_i)\right], \tag{6}$$

where $\sigma_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{-1, +1\}$.

Theorem 1

(*Rademacher bound for Frobenius-bounded bilinear class*) *Let*

$$\mathcal{F}_\Lambda := \left\{ (\mathbf{h}_G, \mathbf{z}) \mapsto \langle W, \mathbf{h}_G \mathbf{z}^\top \rangle_F :\ \|W\|_F \le \Lambda \right\}.$$

Then, for any sample S with $\|\mathbf{h}_{G_i}\| \le B_h, \|\mathbf{z}_i\| \le B_z$,

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_\Lambda; S) \le \frac{\Lambda\, B_h\, B_z}{\sqrt{n}}.$$

Corollary 1

(*Generalization via Rademacher complexity*) *Let $\ell : \mathbb{R}\times\mathcal{Y} \to [0,1]$ be L-Lipschitz in its first argument (e.g., MAE with $L = 1$). For any $f \in \mathcal{F}_\Lambda$ and any $\delta \in (0,1)$, with probability at least $1-\delta$ over an i.i.d. sample of size n,*

$$\begin{aligned}\mathbb{E}[\ell(f(\mathbf{h}_G, \mathbf{z}), y)] \ &\le \frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{h}_{G_i}, \mathbf{z}_i), y_i) \\ &\quad + 2L\widehat{\mathfrak{R}}_n(\mathcal{F}_\Lambda; S) \\ &\quad + 3\sqrt{\frac{\log(2/\delta)}{2n}}.\end{aligned}$$

Combining with Theorem 1 yields an $\mathcal{O}\big(\Lambda B_h B_z/\sqrt{n}\big)$ excess-risk term.

Theorem 1 indicates that, despite the enhanced expressive power gained from Kronecker-product fusion, the associated bilinear hypothesis class retains favorable generalization properties under Frobenius or nuclear norm constraints. In other words, the statistical reliability of the model is preserved while achieving richer feature interactions. As discussed in "Descriptor selection" section, the descriptor selection stage effectively reduces dimensionality from p_0 to a much smaller $k \ll p_0$. Although our model does not enforce an explicit low-rank factorization on the parameter matrix, this dimensionality reduction restricts the Kronecker-product embedding (5) to a lower-dimensional subspace, implicitly reducing the effective rank of the bilinear representation $\mathbf{h}_G^\top W \mathbf{z}$. This acts as a structural regularizer akin to low-rank constraints. To theoretically support this perspective, Theorem 2 establishes Rademacher complexity bounds under explicit low-rank or nuclear-norm constraints on to W. Consider a rank–r factorization $W = UV^\top$ with $U \in \mathbb{R}^{d\times r}$, $V \in \mathbb{R}^{k\times r}$, and columns $U = [\mathbf{u}_1, \ldots, \mathbf{u}_r]$, $V = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$. The matrix factorization satisfies the standard norm inequalities:

$$\begin{aligned} \|W\|_F &\le \|U\|_F \, \|V\|_F, \\ \|W\|_* &\le \tfrac{1}{2}\big(\|U\|_F^2 + \|V\|_F^2\big). \end{aligned} \tag{7}$$

The first bound follows from submultiplicativity, since $\|UV^\top\|_F \le \|U\|_F \, \|V^\top\|_2 \le \|U\|_F \|V\|_F$. The second bound is the standard factorization characterization of the trace norm: for any factorization $W = UV^\top$, the nuclear norm satisfies $\|W\|_* \le \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2)$, with equality attained at an optimal factorization.

Theorem 2

(*Rademacher bounds under low-rank factorization*) *Suppose $W = UV^\top$ with $\|U\|_F \le A$ and $\|V\|_F \le B$. Then $f(\mathbf{h}_G, \mathbf{z}) = \langle W, \mathbf{h}_G \mathbf{z}^\top \rangle_F$ belongs to $\mathcal{F}_\Lambda$ with $\Lambda \le AB$, and therefore*

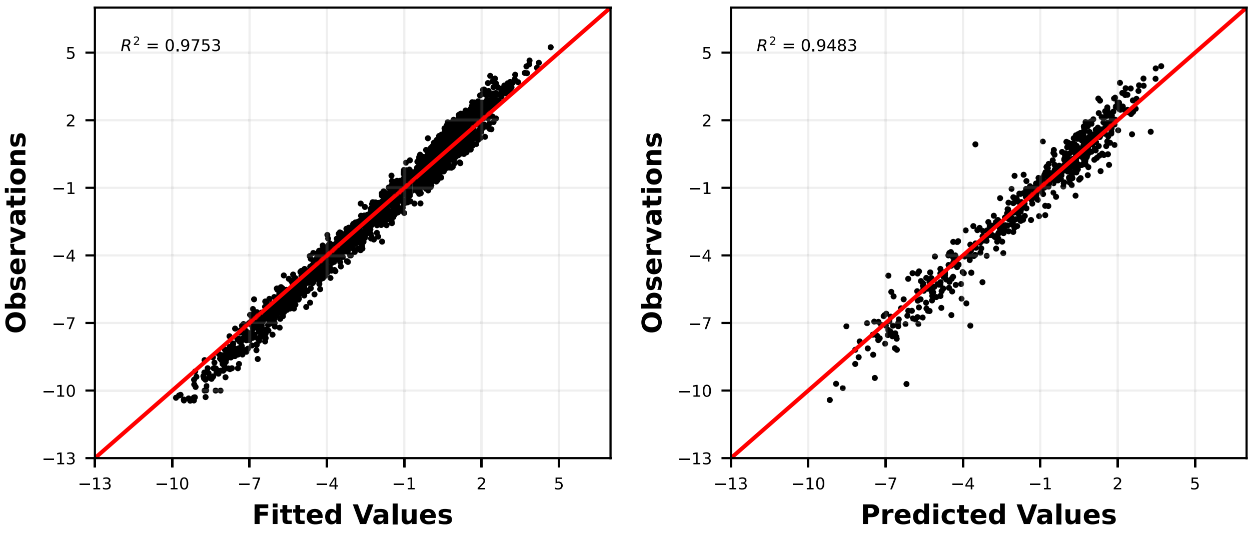
$$\widehat{\mathfrak{R}}_n \;\le\; \frac{AB\, B_h\, B_z}{\sqrt{n}}.$$

Alternatively, if $\|W\|_* \le \tau$, then with the nuclear-norm ball $\mathcal{F}_\tau := \{(\mathbf{h}_G, \mathbf{z}) \mapsto \langle W, \mathbf{h}_G \mathbf{z}^\top \rangle_F : \|W\|_* \le \tau\}$,

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_\tau; S) \;\le\; \frac{\tau\, B_h\, B_z}{\sqrt{n}}.$$

Consequently, Theorem 2 bridges the gap between practical dimensionality reduction and explicit low-rank factorization, ensuring both generalization reliability and computational efficiency. The proofs of all the above are provided in "Appendix A".

Here, we present the computational complexity of our model. Let N be the number of nodes in a molecular graph, D the hidden feature dimension, p_0 the number of input descriptors, d the dimension of graph embedding, and h the hidden width of the descriptor encoder. A standard GCN with L-layer graph encoder requires time complexity of $\mathcal{O}(L(|E|D + ND^2))$ [52], and the descriptor encoder (a lightweight MLP) has a cost of $\mathcal{O}(p_0 h)$. Since the fusion module activates only a statistically selected subset of $k \ll p_0$ descriptor dimensions, it incurs a compact interaction with $\mathcal{O}(dk)$, which is



(a) Training dataset

(b) Test dataset

Fig. 4 Scatter plots of fitted and predicted vapor pressure values for the training and test datasets

effectively linear in the descriptor dimension when k is small. The overall time complexity therefore becomes $\mathcal{O}_{Time} = \mathcal{O}\big(L(|E|D + ND^2) + p_0h + dk\big)$. We next analyze the space complexity. The space complexity of GCN training algorithms with an L-layer is known as $\mathcal{O}(L(ND + D^2))$ [52]. The descriptor encoder (a lightweight MLP) requires $\mathcal{O}(p_0h)$ parameters. The fusion stores only a dk-dimensional representation, i.e., $\mathcal{O}(dk)$ with $k \ll p_0$. Consequently, the total space complexity becomes $\mathcal{O}_{Space} = \mathcal{O}\big(L(ND + D^2) + p_0h + dk\big)$, which remains dominated by the underlying GCN encoder.

Experimental results

Four datasets were used in the experiments: two publicly available molecular datasets and two self-curated datasets. The details of these datasets are provided in "Physical chemistry properties–Physicochemical properties(solubility)" sections. To rigorously evaluate model performance, both the random data split and scaffold split were employed. Compared to random split, scaffold split partitions the data based on molecular scaffold structures, providing more challenges for prediction. Under both split settings, five-fold cross validation with two loss functions—mean squared error (MSE) and mean absolute error (MAE)—was employed. Each model was trained for 300 epochs per fold, and the average MSE and MAE were used as evaluation metrics. To ensure statistical significance of the results, all experiments were repeated 15 times and the mean and standard error of each metric were reported. In addition to the GCN backbone, we also conducted supplementary experiments with GAT [53] and GIN [54] backbone to assess whether our method retain its performance advantages across different encoder architectures. All experiments were conducted on a desktop with an Intel Core i7 CPU, 32GB RAM, and a GeForce RTX 4060Ti GPU. The following Python libraries and frameworks were used: ISIS [50] for variable screening, DGL for graph modeling, RDKit [39] for molecular structure processing, and mendeleev for element property retrieval.

Physical chemistry properties

Molecular machine learning tasks can be categorized according to molecular property [55], and in this study we used two representative datasets that fall under the category of physical chemistry tasks. The first dataset, FreeSolv (Free Solvation) [56], contains computational and experimental hydration free energies for 642 neutral molecules in water. The computational values are derived from alchemical free energy using molecular dynamics simulations; however, we consider only the experimentally obtained hydration free energies. The second dataset, ESOL (Estimated SOLubility) [57], contains water solubility data for 2,874 compounds. In this study, we used 1,128 compounds curated for machine learning tasks by Wu et al. [55].

The target values of the two datasets—hydration free energy (FreeSolv) and aqueous solubility (ESOL)—exhibit approximately normal-like distributions (see Appendix Fig. 6). To explore the relationship between molecular descriptors and target properties, we divided the target values into three tertiles (low, mid, and high). A total of 209 molecular descriptors were initially calculated, of which 177 and 183 were retained for the two datasets, respectively (see "Descriptor selection" section on preprocessing), and subsequently reduced to two principal components via Principal Component Analysis (PCA). Appendix Fig. 9 shows three natural clusters, corresponding to the low, mid, and high target groups. This implies that molecular descriptors can be useful in predicting target properties. The summary statistics of 209 descriptors are provided in Appendix Tables 16 and 17.

To validate the effect of the proposed method, we performed the following ablation experiments. First, we evaluated the effect of descriptor selection by gradually integrating the selected descriptors into the model. Before integration, the descriptors were sorted in descending order of the absolute values of their regression coefficients. Second, we compared two fusion strategies—concatenation and Kronecker-product—to examine how different integration methods affect predictive performance. These ablation experiments allowed us to validate effect of descriptor selection and Kronecker-product fusion. For comparative evaluation, we included the following baseline models in our experiments: GCN [1], and fusion-based models such as EGCN [16], D-MPNN [58], and BAN [59]. Tables 1 and 2 summarizes the results. GCN always underperformed compared to a model that integrates at least one descriptor through concatenation. This indicates that molecular descriptors provide useful information for molecular property prediction.

Examining the effect of descriptor selection, we found that concatenating the top three descriptors with graph embedding generally outperformed the EGCN baseline that concatenates three predefined descriptors. For example, on FreeSolv under random split settings, the selected descriptors achieved an MSE of 5.070, compared to 5.713 for EGCN (Table 1). This provides clear evidence of the effect of descriptor selection, generating enhanced graph representations. Moreover, performance was consistently improved as the number of descriptors increased. Although an exception was observed under the scaffold split, this result does not alter the overall trend. These findings indicate that descriptor selection effectively guides the model to identify the most informative features, thereby enriching graph representations and promoting convergence toward lower prediction error.

Next, we examined the effect of Kronecker-product. Compared to concatenation, integrating the top three descriptors via Kronecker-product consistently outperformed EGCN. For example, on FreeSolv under random split, our model achieved an MSE of 3.498, compared to 5.713 for EGCN. Moreover, when using the same descriptors, the Kronecker-product significantly outperformed concatenation in most settings. On ESOL, although concatenation performed better under scaffold split, Kronecker-product fusion achieved the best performance when all selected descriptors were integrated. Except for this case, our findings consistently support that Kronecker-product fusion provides superior performance, suggesting its effectiveness in modeling second-order interactions between descriptors and graph embeddings.

To further benchmark against other fusion strategies, we compared our model with BAN [59]. Although BAN achieved performance comparable to ours, it required substantially higher computational costs. The average training time per epoch was 0.80 s/epoch for BAN, versus 0.66 s/epoch for ours. Furthermore, because BAN applies bilinear attention over all descriptor without selection, it consumed markedly more GPU memory (1.17 MB) than our model (0.32 MB).

Consequently, integrating all automatically selected descriptors—50 from FreeSolv and 63 from ESOL—through a Kronecker-product achieved the lowest prediction errors (with minor exceptions). Figure 3 displays the predicted values for the test sets, where 20% of the data were randomly allocated for testing. The coefficient of determination (R^2) reached approximately 0.92 for FreeSolv and 0.91 for ESOL.

To elucidate the physicochemical rationale behind descriptor selection, we conducted Shapley Additive exPlanations (SHAP) analysis for both concatenation and Kronecker-product fusion, revealing that statistically selected descriptors align with established chemical principles. Here, percentages in parentheses indicate the contribution of each descriptor within the selected set (Appendix Table 13). For FreeSolv (Fig. 13b), *SlogP_VSA2* dominates (13.15%), capturing the hydrophobic effect central to solvation thermodynamics. Hydrogen bonding descriptors *NHOHCount* (12.40%) and *NumHDonors* (6.40%) rank highly, reflecting enthalpic contributions from water-solute interactions. This hierarchy matches Abraham's solvation parameters, confirming mechanistic learning. Concatenation (Fig. 13a) shows dispersed, less interpretable patterns. For ESOL (Fig. 13d), *MolLogP* leads (14.57%), consistent with Yalkowsky's General Solubility Equation where lipophilicity inversely correlates with solubility. *MaxEStateIndex* and *FpDensityMorgan2* capture electronic and structural complexity effects on crystal packing. The concentrated SHAP distributions indicate decisive feature utilization, contrasting with concatenation's uncertain patterns (Fig. 13c). These SHAP analysis highlights the interpretability benefits of Kronecker-product fusion and its superiority over concatenation in capturing chemically grounded structure–property interactions.

We additionally conducted experiments using GAT and GIN as alternative backbones. The corresponding results are provided in the "Appendix C", which show consistent gains over the baseline models.

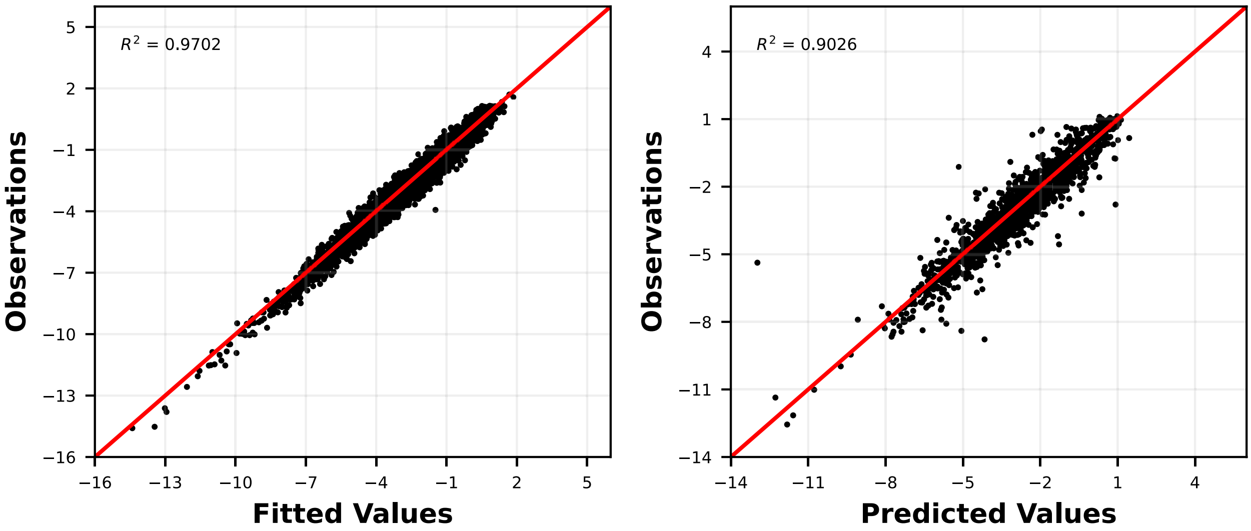
Physicochemical properties (vapor pressure)

We further conducted experiments on a self-curated gas dataset consisting of 3152 compounds represented as SMILES strings. This dataset was compiled from diverse sources including in-house experiments, public databases, and published literature. The target property was the log-transformed vapor pressure.

Vapor pressure is a critical physicochemical property, as it reflects volatility and toxicity, the potential risk of airborne dispersion, and its relevance to environmental regulation. Despite its importance, vapor pressure has rarely been explored in complicated deep learning. In this study, we provide a systematic analysis of vapor pressure using multimodal graph fusion with descriptor selection.

The distribution of vapor pressure values is shown in Appendix Fig. 7a. It is slightly left-skewed, indicating that the mean is smaller than the median due to a few low values. Examination of the boxplot revealed 20 data points lying beyond the whiskers, corresponding to potential outliers. Based on the interquartile range, these outliers were removed, resulting in a final dataset of 3,132 compounds (Appendix Fig. 8). A total of 209 descriptors were initially calculated from SMILES, of which 190 were retained for exploratory analysis prior to descriptor selection (see "Descriptor selection" section).

To assess whether molecular descriptors provide meaningful information for predicting vapor pressure, we applied PCA to the 190 preprocessed descriptors. This dimensionality reduction preserved most of the variance in the first few principal components. The vapor pressure values were divided into three tertiles (low, mid, and high) and these labels were assigned to the PCA-reduced descriptors. As shown in Appendix Fig. 10, three distinct clusters emerged, indicating that the molecular descriptors contain informative features related to vapor pressure. Subsequently, 23 descriptors were selected through the descriptor selection



(a) Training dataset

(b) Test dataset

Fig. 5 Scatter plots of fitted and predicted solubility values for the training and test datasets

procedure. The correlation heatmap of these descriptors is displayed in Appendix Fig. 11, where blue indicates positive correlations and red indicates negative correlations. For instance, *NumAromaticRings* is strongly positively correlated with *BertzCT* ($r = 0.84$) and strongly negatively correlated with *HallKierAlpha* ($r = -0.65$). Summary statistics of the initial 209 descriptors are provided in Appendix Table 18.

Following the ablation experiments as described in “Physical chemistry properties” section, we numerically validate the effects of descriptor selection and the Kronecker-product fusion. These ablation experiments further support the validity of our method and establish its generalizability. The results are summarized in Table 3. As expected, GCN exhibited the lowest performance (MSE = 6.668 under random split), in contrast, integrating even a single descriptor into GCN significantly improved performance (MSE = 4.378), suggesting that chemically informative descriptors provide more substantial gains in representation learning than message-passing mechanisms alone.

To evaluate the effect of descriptor selection, we concatenated the top three statistically selected descriptors with graph embeddings. This yielded a marked improvement over EGCN baseline. For example, the model with selected descriptors achieved an MSE of 1.192 compared to 1.464 for EGCN. These findings demonstrate that statistical descriptor selection is more effective than manual selection in terms of predictive accuracy and representation learning. Furthermore, as the number of integrated descriptors increased, predictive performance improved steadily. This indicates that statistically guided descriptor selection promotes convergence to an optimal solution, thereby enhancing both efficiency and interpretability.

We next examined the effect of Kronecker-product fusion. When integrating the top three descriptors, our model achieved an MSE of 1.079, outperforming the baseline value of 1.464. Under identical descriptor settings, Kronecker-product fusion consistently outperformed concatenation. Moreover, as the number of descriptors increased, Kronecker-product fusion provided faster and more stable performance gains, underscoring its superior ability to capture cross-modal interactions between descriptors and graph embeddings. Integrating all 23 selected descriptors via the Kronecker-product yielded the lowest prediction error, with an MSE of 0.444 compared to 1.464 for EGCN. Although BAN achieved a lower MAE of 0.484 under random split, this minor gain comes at the cost of higher training time and memory usage. This result validates the effectiveness and efficiency of combining descriptor selection with Kronecker fusion for complex molecular property prediction.

Figure 4 presents predicted versus true values for both training and test sets, with 20% of the data randomly allocated for testing. While test set predictions were more dispersed around the identity line than those of the training set, no overfitting was observed. The R^2 reached approximately 0.98 for the training set and 0.95 for the test set.

To clarify the chemical factors contributing to vapor pressure predictions, we conducted SHAP analysis

Table 3 Predictive performance comparison of our model versus competing models for the self-curated vapor pressure datasets, demonstrating the effectiveness of descriptor selection and Kronecker-product fusion

# of Descriptors[1]	GCN [1]	EGCN [16]	EGCN [16] with DS[2]	D-MPNN [58]	BAN [59]	KROVEX (ours)
0	6.668 ± 0.142	–	–	0.566 ± 0.031	–	–
	1.966 ± 0.022	–	–	0.537 ± 0.007	–	–
	10.680 ± 0.972	–	–	0.992 ± 0.032	–	–
	2.639 ± 0.150	–	–	0.720 ± 0.016	–	–
1		4.378 ± 0.116	–	–	–	–
	–	1.541 ± 0.021	–	–	–	–
	–	5.569 ± 0.435	–	–	–	–
	–	1.801 ± 0.073	–	–	–	–
2		1.653 ± 0.047	–	–	–	–
	–	0.955 ± 0.012	–	–	–	–
	–	1.994 ± 0.105	–	–	–	–
	–	1.060 ± 0.027	–	–	–	–
3		1.464 ± 0.034	1.192 ± 0.026	–	–	1.079 ± 0.032
	–	0.879 ± 0.008	0.814 ± 0.008	–	–	0.782 ± 0.010
	–	1.781 ± 0.088	1.621 ± 0.066	–	–	1.769 ± 0.167
	–	1.002 ± 0.021	0.988 ± 0.027	–	–	1.000 ± 0.025
5		–	0.946 ± 0.027	–	–	0.818 ± 0.017
	–	–	0.710 ± 0.007	–	–	0.678 ± 0.008
	–	–	1.247 ± 0.055	–	–	1.139 ± 0.145
	–	–	0.825 ± 0.021	–	–	0.887 ± 0.044
7		–	0.858 ± 0.017	–	–	0.657 ± 0.016
	–	–	0.687 ± 0.007	–	–	0.593 ± 0.009
	–	–	1.232 ± 0.096	–	–	1.201 ± 0.162
	–	–	0.792 ± 0.015	–	–	0.812 ± 0.041
10		–	0.700 ± 0.014	–	–	0.528 ± 0.011
	–	–	0.613 ± 0.007	–	–	0.540 ± 0.006
	–	–	1.130 ± 0.084	–	–	0.924 ± 0.103
	–	–	0.753 ± 0.020	–	–	0.753 ± 0.047
20		–	0.598 ± 0.015	–	–	0.447 ± 0.009
	–	–	0.563 ± 0.006	–	–	0.497 ± 0.006
	–	–	0.793 ± 0.042	–	–	0.789 ± 0.057
	–	–	0.669 ± 0.018	–	–	0.699 ± 0.030
23		–	0.592 ± 0.019	–	–	**0.444 ± 0.009**
	–	–	0.560 ± 0.008	–	–	0.491 ± 0.006
	–	–	**0.762 ± 0.032**	–	–	0.770 ± 0.053
	–	–	0.658 ± 0.015	–	–	**0.621 ± 0.037**
ALL[3]		–	–	0.538 ± 0.022	0.486 ± 0.022	–
	–	–	–	0.503 ± 0.007	**0.484 ± 0.007**	–
	–	–	–	0.963 ± 0.033	0.879 ± 0.046	–
	–	–	–	0.737 ± 0.014	0.682 ± 0.021	–

ROVEX showed consistently strong performance on most metrics

The first two values represent MSE and MAE under the random split, while the last two values represent MSE and MAE under the scaffold split. All experiments were repeated 15 times, and the results are presented as the means ± standard error for each metric. The best and second-best performances for each metric are marked bold and underlined

[1] Indicates the number of descriptors integrated into the model

[2] Indicates Descriptor Selection

[3] Denotes the complete descriptor set for each model (200 for D-MPNN and 196 for BAN.)

Table 4 Predictive performance comparison of our model versus competing models for the self-curated solubility datasets, demonstrating the effectiveness of descriptor selection and Kronecker-product fusion

# of Descriptors[1]	GCN [1]	EGCN [16]	EGCN [16] with DS[2]	D-MPNN [58]	BAN [59]	KROVEX (ours)
0	3.018 ± 0.049	–	–	0.597 ± 0.024	–	–
	1.348 ± 0.010	–	–	0.578 ± 0.006	–	–
	3.370 ± 0.173	–	–	0.672 ± 0.016	–	–
	1.437 ± 0.033	–	–	0.650 ± 0.011	–	–
1		2.263 ± 0.036	–	–	–	–
	–	1.128 ± 0.006	–	–	–	–
	–	2.504 ± 0.099	–	–	–	–
	–	1.206 ± 0.022	–	–	–	–
2		1.321 ± 0.018	–	–	–	–
	–	0.874 ± 0.007	–	–	–	–
	–	1.536 ± 0.055	–	–	–	–
	–	0.934 ± 0.014	–	–	–	–
3		1.263 ± 0.018	1.086 ± 0.019	–	–	1.052 ± 0.021
	–	0.852 ± 0.006	0.771 ± 0.007	–	–	0.720 ± 0.008
	–	1.499 ± 0.054	1.261 ± 0.045	–	–	1.256 ± 0.042
	–	0.924 ± 0.011	0.841 ± 0.010	–	–	0.818 ± 0.013
5		–	1.132 ± 0.030	–	–	0.911 ± 0.027
	–	–	0.769 ± 0.009	–	–	0.697 ± 0.012
	–	–	1.313 ± 0.055	–	–	1.201 ± 0.045
	–	–	0.830 ± 0.011	–	–	0.811 ± 0.012
7		–	1.054 ± 0.020	–	–	0.785 ± 0.023
	–	–	0.743 ± 0.007	–	–	0.624 ± 0.006
	–	–	1.216 ± 0.038	–	–	1.152 ± 0.073
	–	–	0.810 ± 0.012	–	–	0.746 ± 0.009
10		–	0.958 ± 0.020	–	–	0.676 ± 0.020
	–	–	0.722 ± 0.006	–	–	0.590 ± 0.005
	–	–	1.123 ± 0.030	–	–	0.974 ± 0.045
	–	–	0.772 ± 0.009	–	–	0.712 ± 0.011
20		–	0.751 ± 0.021	–	–	0.463 ± 0.014
	–	–	0.601 ± 0.005	–	–	0.446 ± 0.004
	–	–	0.968 ± 0.107	–	–	0.670 ± 0.029
	–	–	0.670 ± 0.013	–	–	**0.566 ± 0.008**
30		–	0.756 ± 0.002	–	–	**0.448 ± 0.015**
	–	–	0.585 ± 0.006	–	–	**0.441 ± 0.004**
	–	–	0.902 ± 0.045	–	–	**0.666 ± 0.031**
	–	–	0.651 ± 0.007	–	–	0.573 ± 0.010
ALL[3]		–	–	0.586 ± 0.025	0.509 ± 0.014	–
	–	–	–	0.529 ± 0.005	0.511 ± 0.002	–
	–	–	–	0.680 ± 0.014	0.693 ± 0.019	–
	–	–	–	0.593 ± 0.007	0.624 ± 0.008	–

ROVEX outperformed competitors with the best predictive performance

The first two values represent MSE and MAE under the random split, while the last two values represent MSE and MAE under the scaffold split. All experiments were repeated 15 times, and the results are presented as the means ± standard error for each metric. The best and second-best performances for each metric are marked bold and underlined

[1] Indicates the number of descriptors integrated into the model

[2] Indicates descriptor selection

[3] Denotes the complete descriptor set for each model (200 for D-MPNN and 196 for BAN)

for both concatenation and Kronecker-product fusion (Fig. 14a, b and Appendix Table 14). *TPSA* dominates (11.84%) as polar surface area determines intermolecular forces governing volatility. *MolMR* (10.44%) captures size-polarizability effects, while *Kappa1* (9.37%) reflects branching's impact on molecular contact. The model correctly identifies *fr_halogen*'s competing effects on volatility. In contrast, concatenation (Fig. 14a) fails to prioritize these chemically relevant features. These results confirm that Kronecker fusion not only improves predictive accuracy but also captures chemically grounded structure–property relationships that simple concatenation fails to reveal.

We further evaluated our model using GAT and GIN as alternative backbones. Tables 7 and 11 summarize the results, showing consistent improvements over the baseline models.

Physicochemical properties (solubility)

We evaluated our model on a self-curated aqueous solubility dataset consisting of 8,789 compounds represented as SMILES strings. This dataset was constructed from public sources and curated laboratory reports, and includes solubility measurements not present in ESOL. The purpose of this experiment was to evaluate whether the performance gains observed in ESOL extend to structurally and chemically diverse solubility regimes.

The experimental procedure, including descriptor selection, model training, and evaluation was identical to that used for other datasets. Consistent with the other datasets, our model achieved strong predictive performance on solubility. Table 4 summarizes the predictive performance, and the Fig. 5 shows predicted versus true values for both training and test sets, with 20% of the data randomly allocated for testing. The model achieved an R^2 of approximately 0.97 on the training and 0.90 on the test set.

The comprehensive SHAP analysis was conducted to validate the chemical relevance (Fig. 14c, d, and Appendix Table 14). Like the other prediction model and SHAP analysis results, the solubility dataset shows similar patterns where Kronecker fusion identifies chemically intuitive descriptors. *MolMR* (12.64%) captures molecular volume and polarizability crucial for solvation cavity formation. *PEOE_VSA7* (9.96%) quantifies charge-weighted surface area for electrostatic interactions, while *fr_halogen* (7.75%) captures halogen-specific effects. These selections comprehensively cover the hydrophobic-hydrophilic balance fundamental to dissolution.

Across all datasets, the SHAP analysis demonstrates Kronecker fusion's advantages: chemical selectivity over mathematical abstractions, concentrated contribution patterns indicating confident predictions, and explicit multiplicative interaction modeling through $h_i z_j$ terms. These interactions capture context-dependent effects fundamental to chemistry—how a hydroxyl group's contribution varies with molecular size—which concatenation cannot model. The descriptor selections validate against established models (COSMO-RS, Delaney's equation, EPA's SPARC), demonstrating that KROVEX learns genuine structure–property relationships suitable for mechanistic understanding.

We additionally conducted experiments using GAT and GIN as alternative backbones. The corresponding results are provided in the Appendix Tables 8 and 12, where the findings were consistent with those from the other datasets.

Discussion

This study demonstrates that statistically guided descriptor selection combined with Kronecker-product fusion yields consistent and substantial improvements in molecular property prediction across physicochemical benchmarks (FreeSolv, ESOL), a safety-critical vapor-pressure, and a self-curated solubility dataset (Tables 1, 2, 3, 4, Figs. 3, 4, 5). Two findings are most notable. First, integrating even a small number of statistically selected descriptors significantly outperforms message-passing–only baselines, showing that curated physicochemical attributes complement structural features captured by GCN. Second, as the number of descriptors increases, Kronecker-product fusion consistently surpasses simple concatenation, highlighting the importance of explicitly modeling cross-modal interactions rather than relying on feature accumulation. Beyond predictive gains, these findings have methodological implications for multimodal learning in cheminformatics. While concatenation remains the most common fusion strategy, our results indicate that explicitly modeling second-order interactions in a capacity-controlled manner can improve both accuracy and interpretability. For applications where safety, regulation, and scientific insight are critical, this structured approach is preferable to opaque feature expansions or very deep networks.

The ablation results, together with the analysis in "Multimodal fusion" section, clarify the effectiveness of KROVEX. The ISIS→Elastic Net pipeline identifies a compact, informative subset of descriptors while suppressing redundancy and multicollinearity. Empirically, performance improves monotonically (with minor fluctuations) as additional selected descriptors are introduced, and selected triplets already surpass prior fixed choices. Methodologically, shrinking from $p_0 = 209$ to $k \ll p_0$ induces an implicit low-rank effect on the subsequent bilinear layer, providing structural regularization without explicitly constraining rank [51]. Concatenation cannot express multiplicative cross-modal interactions with an affine head (Proposition 1), whereas Kronecker features implement precisely those bilinear terms (Lemma 1). Equivalently, KROVEX realizes a degree-2 cross-polynomial kernel on $[\mathbf{h}_G; \mathbf{z}]$ that excludes within-modality quadratics (Lemma 2). This exclusion focuses model capacity on chemically plausible interactions (e.g., how a structural motif's effect is modulated by polarity or surface area) and avoids unnecessary quadratic self-expansions that inflate variance. The bilinear hypothesis class admits $\mathcal{O}(\Lambda B_h B_z/\sqrt{n})$ Rademacher bounds under Frobenius or nuclear-norm constraints (Theorems 1, 2), explaining why the added expressiveness of Kronecker fusion does not yield overfitting in our experiments. In practice, descriptor selection further reduces the effective dimension of the Kronecker space (dk), complementing norm-based regularization to stabilize training. KROVEX preserves a direct mapping from selected descriptors to interaction terms $h_i z_j$, enabling *post hoc* attribution that is more transparent than deep, purely learned stacks. Two levels of interpretation arise naturally: (1) *main-effect* contributions of individual descriptors (e.g., topological polar surface area, aromatic proportion) and (2) *interaction* contributions via prominent $h_i z_j$ terms, which reveal where in the learned structural subspace a descriptor is most influential. For safety-relevant endpoints such as vapor pressure, such structured interpretability facilitates model auditing and hypothesis generation (e.g., identifying structural regions where volatility is especially sensitive to polarity or size).

FreeSolv and ESOL are the most widely adopted benchmarks in molecular property prediction. Wu et al. [55]'s MoleculeNet systematically compared graph neural networks (GCN, GAT, etc.) on these datasets, emphasizing the necessity of descriptor supplementation. Subsequently, Na et al. [16] reported performance improvements over baselines by simply concatenating a few predefined descriptors, though this approach failed to adequately capture descriptor diversity and graph interactions. These pioneering studies established FreeSolv and ESOL as internationally recognized evaluation standards while demonstrating that descriptor utilization is crucial for model improvement. Recently, various approaches including KA-GNN [60], Graph Structure Learning-based models [61], and quantized GNN [62] have achieved performance improvements on FreeSolv and ESOL. However, these methods often rely on indiscriminate descriptor usage or simple concatenation strategies. To overcome these limitations, KROVEX selects statistically significant descriptors and models second-order descriptor-graph embedding interactions through a Kronecker-product fusion. As a result, KROVEX reduced the baseline MSE from 5.713 to 0.973 on FreeSolv and from 0.918 to 0.423 on ESOL, substantially outperforming existing methods. This demonstrates that KROVEX provides superior generalization performance and interpretability compared to simple concatenation.

Vapor pressure is a critical property for understanding volatile hazardous gas behavior and is essential for environmental regulation and process safety design. Traditional approaches have relied on physicochemical models like the Antoine equation or COSMO-RS, as well as QSPR-based regression. However, these methods suffer from high computational costs or arbitrary descriptor selection, limiting their generalization performance for large-scale compound prediction. To address these challenges, recent deep learning models such as GRAPPA [63], GC2NN [64], D-MPNN [65], and Super Learner Ensembles [66] have been proposed, reporting improved performance. Unlike these approaches, KROVEX selects only statistically significant descriptors and explicitly learns second-order interactions between descriptors and graph embeddings through a Kronecker-product fusion. Consequently, on our self-curated gas dataset, KROVEX achieved a significant reduction in prediction error (MSE = 0.444) compared to baseline GCN/EGCN, surpassing simple physical models and conventional QSPR approaches. The exceptional R^2 of 0.95 achieved on the test set surpasses typical QSPR models for this property. This performance gain likely stems from KROVEX's ability to model multiplicative interactions between structural features and global descriptors—precisely the type of relationships that govern phase transitions. This demonstrates that KROVEX provides a novel methodology that simultaneously achieves generalization capability and interpretability for vapor pressure prediction.

To situate our findings relative to existing graph–descriptor fusion methods, we compared KROVEX with Chemprop (D-MPNN), and Bilinear Attention. Across four datasets, KROVEX consistently achieved lower prediction errors under both random and scaffold splits (with minor exception). ESOL was the only dataset for which the two models (ours, and Chemprop) exhibited similar performance. This is consistent with previous studies indicating that ESOL is a small-scale dataset with low noise and limited structural diversity, often leading to comparable results across different molecular models. Notably, while Chemprop adopts a descriptor-fusion strategy, our experiments indicate that such indiscriminate descriptor usage does not lead to significant performance gains and may unnecessarily complicate the model. These observations suggest that the proposed sparsity-aware descriptor selection and Kronecker-product fusion provide performance gains beyond those obtained through conventional descriptor-concatenation architectures. The comparison results are provided in Appendix Table 15.

Bilinear Attention—a widely used and powerful feature-interaction mechanism in multimodal learning—requires a learned $d \times p_0$ projection matrix or multiple attention heads, resulting in a computational complexity on the order of $\mathcal{O}(dp_0)$ (or $\mathcal{O}((d + p_0)r)$ for low-rank-factorizations with $r \ll \min(d, p_0)$). While BAN achieved performance comparable to ours, this came at the cost of higher training time and memory usage, underscoring computational efficiency of KROVEX.

To validate that these gains do not depend on a particular graph encoder, we carried out additional experiments using two commonly used GNN encoders: GAT, GIN. The results showed that KROVEX maintains its performance advantage across all these backbones, with consistent gains over all baseline models (Tables 5, 6, 7, 8, 9, 10, 11, 12). This is largely due to how the Kronecker-product fusion layer operates. The fusion is applied to the final graph embedding $\mathbf{h}_G$ produced by the encoder and forms multiplicative interactions $h_i z_j$ between graph-derived features and molecular descriptors. Since this computation depends mainly on the embedding itself, differences in the underlying message-passing scheme (linear aggregation, attention weighting, or WL-based aggregation) do not substantially change how the two information sources are combined.

In light of these results, the consistent performance gains across four datasets suggest that KROVEX generalizes to tasks where targets reflect both local structural patterns and global physicochemical attributes. The stable behavior across GNN encoders further indicates that the multiplicative interaction modeling captures relevant structure–property relations regardless of the chosen backbone.

For practitioners, we recommend: (1) starting from a broad, standardized descriptor pool; (2) applying ISIS for initial screening and Elastic Net for refinement; (3) using Kronecker fusion with Frobenius or nuclear-norm regularization while monitoring dk; and (4) auditing learned $h_i z_j$ terms to generate mechanistic insights and detect spurious associations.

Despite its strong empirical performance, KROVEX has several concrete limitations. First, the evaluation was conducted on four datasets, primarily covering small- and medium-scale molecular prediction tasks; broader validation on larger and more chemically diverse benchmarks is needed to fully assess generalizability. Second, the framework currently relies solely on 2D molecular descriptors, without incorporating 3D conformational or quantum-chemical information, which may limit predictive fidelity for stereochemically sensitive or highly flexible molecules. Third, although KROVEX performs well under scaffold splits, the present modeling strategy does not explicitly encode structural or conformational constraints, suggesting that more chemically informed or geometry-aware extensions could provide additional robustness. Fourth, descriptor selection is performed as an external preprocessing step rather than being jointly optimized with the GNN encoder, limiting the benefits of fully end-to-end multimodal learning. Finally, while the Kronecker fusion module is more parameter-efficient than standard bilinear mechanisms, it still increases embedding dimensionality and may introduce computational overhead when applied to very large descriptor spaces.

To address these limitations, several directions for future work are envisioned. Integrating descriptor selection into an end-to-end training pipeline would allow feature relevance to be optimized jointly with the graph encoder, mitigating the disconnect introduced by external preprocessing. Exploring higher-order fusion architectures with explicit rank constraints may counteract the dimensional growth inherent to Kronecker embeddings while preserving expressive interactions. In parallel, incorporating uncertainty quantification will be essential for safety-critical molecular applications, particularly where increased feature dimensionality may exacerbate reliability issues. From a chemical modeling perspective, extending the framework from 2D to conformer-aware or quantum-informed descriptors would better capture structure-dependent properties, and evaluating KROVEX on larger, more diverse datasets–including multi-condition prediction tasks–would provide stronger evidence of robustness under realistic deployment scenarios. Collectively, these extensions would broaden the scope of KROVEX while preserving its core strengths in interpretability and theoretical grounding, reinforcing the principle that statistically guided parsimony combined with structured multimodal fusion offers a compelling alternative to conventional concatenation strategies in molecular machine learning.

Conclusions

This study presented a novel multimodal fusion framework for molecular property prediction that integrates graph embeddings with molecular descriptors via a Kronecker-product. The framework combines statistically guided descriptor selection with Kronecker-product-based multimodal fusion, providing both interpretability and enhanced representational capacity. Across benchmark datasets and self-curated datasets, KROVEX consistently outperformed baseline models, attaining an R^2 of approximately 0.95 on the vapor-pressure prediction test set, demonstrating strong performance for this safety-critical property. Ablation studies confirmed that statistically guided descriptor selection yields more informative features than predefined descriptors, and that Kronecker-product fusion offers greater improvements than simple concatenation. From both optimization and theoretical perspectives, Rademacher-complexity bounds demonstrate that enhanced representational capacity can be achieved without uncontrolled model complexity; furthermore, descriptor selection induces an implicit low-rank effect that confines the Kronecker embedding to a tractable subspace. In addition, empirical evaluations across multiple GNN backbones–including GCN, GAT, and GIN–show that KROVEX maintains consistent performance gains, indicating that the proposed fusion mechanism is robust to architectural variations in graph encoders. Compared with bilinear fusion techniques, KROVEX achieves similar or superior predictive accuracy while avoiding the parameter growth and overfitting tendencies inherent to high-dimensional bilinear operations. These findings highlight the practical advantages of Kronecker-based fusion, particularly in data-limited molecular learning settings.

While the current implementation relies on 2D molecular descriptors and independent feature-selection stages, which may limit its applicability to conformationally flexible molecules, these design choices balance interpretability with computational efficiency. KROVEX thus provides a theoretically grounded and practically effective framework that delivers improved predictive performance while retaining interpretability, offering a promising avenue for molecular property prediction in cheminformatics and a foundation for future extensions to broader multimodal learning tasks in chemistry and materials science.

Appendix A: Proofs

This section contains the proofs for all propositions, lemmas, and theorems stated in the main text.

A.1: Proposition 1

Proof

Suppose, toward a contradiction, that there exist $\mathbf{a}, \mathbf{b}, c$ *with* $\mathbf{a}^\top \mathbf{h}_G + \mathbf{b}^\top \mathbf{z} + c \equiv \mathbf{h}_G^\top S\,\mathbf{z}$ *for all* $(\mathbf{h}_G, \mathbf{z})$*. Fix indices* $i \in \{1, \dots, d\}$ *and* $j \in \{1, \dots, k\}$ *and set* $\mathbf{h}_G = t\,\mathbf{e}_i$, $\mathbf{z} = s\,\mathbf{e}_j$ *for arbitrary* $s, t \in \mathbb{R}$ *(where* $\{\mathbf{e}_i\}$ *are standard basis vectors). Then*

$$a_i\,t \;+\; b_j\,s \;+\; c \;=\; S_{ij}\,ts \qquad \text{for all } s, t \in \mathbb{R}.$$

Setting $s = 0$ gives $a_i\,t + c = 0$ for all t, so $a_i = 0$ and $c = 0$. Setting $t = 0$ then yields $b_j\,s = 0$ for all s, hence $b_j = 0$. With $a_i = b_j = c = 0$, the identity reduces to $0 = S_{ij}\,ts$ for all s, t, implying $S_{ij} = 0$. Since (i, j) were arbitrary, $S = 0$, contradicting the assumption. Therefore no such g exists when $S \neq 0$. □

A.2: Lemma 1

Proof

The equalities $\langle W, \mathbf{h}_G\mathbf{z}^\top\rangle_F = \mathrm{tr}(\mathrm{W}^\top \mathbf{h}_{\mathrm{G}}\mathbf{z}^\top) = \mathbf{h}_{\mathrm{G}}^\top \mathrm{W}\,\mathbf{z}$ *follow from properties of the trace and Frobenius inner product. The identity* $\mathrm{vec}(\mathbf{h}_{\mathrm{G}}\mathbf{z}^\top) = \mathbf{z} \otimes \mathbf{h}_{\mathrm{G}}$ *is standard, and* $\langle A, B\rangle_F = \mathrm{vec}(\mathrm{A})^\top \mathrm{vec}(\mathrm{B})$ *yields the remaining equalities.* □

A.3: Lemma 2

Recall the identity $\mathrm{vec}(\mathrm{u}\mathrm{v}^\top) = \mathrm{v} \otimes \mathrm{u}$, and the Kronecker-product property $(\mathbf{a} \otimes \mathbf{b})^\top (\mathbf{c} \otimes \mathbf{d}) = \langle \mathbf{a}, \mathbf{c}\rangle\,\langle \mathbf{b}, \mathbf{d}\rangle$.

Proof

By $\phi(\mathbf{h}_G, \mathbf{z}) = \mathbf{z} \otimes \mathbf{h}_G$ *and the basic Kronecker identity,*

$$\begin{aligned}\langle \phi(\mathbf{h}_G, \mathbf{z}), \phi(\mathbf{h}_G', \mathbf{z}')\rangle &= (\mathbf{z} \otimes \mathbf{h}_G)^\top (\mathbf{z}' \otimes \mathbf{h}_G') \\ &= \langle \mathbf{z}, \mathbf{z}'\rangle\,\langle \mathbf{h}_G, \mathbf{h}_G'\rangle,\end{aligned}$$

which proves the first claim. For the second identity, note that

$$\begin{aligned}K_2(x, x') = (x^\top x')^2 &= (\mathbf{h}_G^\top \mathbf{h}_G' + \mathbf{z}^\top \mathbf{z}')^2 \\ &= (\mathbf{h}_G^\top \mathbf{h}_G')^2 + (\mathbf{z}^\top \mathbf{z}')^2 + 2\,\langle \mathbf{h}_G, \mathbf{h}_G'\rangle\,\langle \mathbf{z}, \mathbf{z}'\rangle.\end{aligned}$$

Rearranging yields $\langle \mathbf{h}_G, \mathbf{h}_G'\rangle\,\langle \mathbf{z}, \mathbf{z}'\rangle = \frac{1}{2}\{K_2(x, x') - (\mathbf{h}_G^\top \mathbf{h}_G')^2 - (\mathbf{z}^\top \mathbf{z}')^2\}$. □

Remark

Equivalently, $K_\times$ *is the tensor-product kernel of the two linear kernels* $k_h(\mathbf{h}_G, \mathbf{h}_G') = \langle \mathbf{h}_G, \mathbf{h}_G'\rangle$ *and* $k_z(\mathbf{z}, \mathbf{z}') = \langle \mathbf{z}, \mathbf{z}'\rangle$*, i.e.,* $K_\times = k_h \otimes k_z$*.*

A.4: Theorem 1

Proof

By definition,

$$\begin{aligned}\widehat{\mathfrak{R}}_n &= \mathbb{E}_\sigma\Big[\sup_{\|W\|_F \le \Lambda} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \langle W, \mathbf{h}_{G_i}\mathbf{z}_i^\top\rangle_F\Big] \\ &= \frac{\Lambda}{n}\,\mathbb{E}_\sigma \big\|\, \textstyle\sum_{i=1}^{n} \sigma_i\,\mathbf{h}_{G_i}\mathbf{z}_i^\top \big\|_F,\end{aligned}$$

where the last equality uses Cauchy–Schwarz in the Frobenius inner product. By Jensen and the orthogonality of Rademacher signs,

$$\begin{aligned}\mathbb{E}_\sigma \big\|\textstyle\sum_{i=1}^{n} \sigma_i \mathbf{h}_{G_i}\mathbf{z}_i^\top\big\|_F &\le \Big(\mathbb{E}_\sigma \big\|\textstyle\sum_{i=1}^{n} \sigma_i \mathbf{h}_{G_i}\mathbf{z}_i^\top\big\|_F^2\Big)^{1/2} \\ &= \Big(\textstyle\sum_{i=1}^{n} \|\mathbf{h}_{G_i}\mathbf{z}_i^\top\|_F^2\Big)^{1/2}.\end{aligned}$$

Since $\|\mathbf{h}_{G_i}\mathbf{z}_i^\top\|_F = \|\mathbf{h}_{G_i}\|\,\|\mathbf{z}_i\|$, we obtain $\big(\sum_i \|\mathbf{h}_{G_i}\|^2\|\mathbf{z}_i\|^2\big)^{1/2} \le \sqrt{n}\,B_h\,B_z$, hence the claimed bound. □

A.5: Theorem 2

Proof
The first claim follows by (7) and Theorem 1 with $\Lambda \le AB$. For the nuclear-norm class, by duality $\sup_{\|W\|_* \le \tau}\langle W, M\rangle_F = \tau\|M\|_{\mathrm{op}}$ for any matrix M. Hence

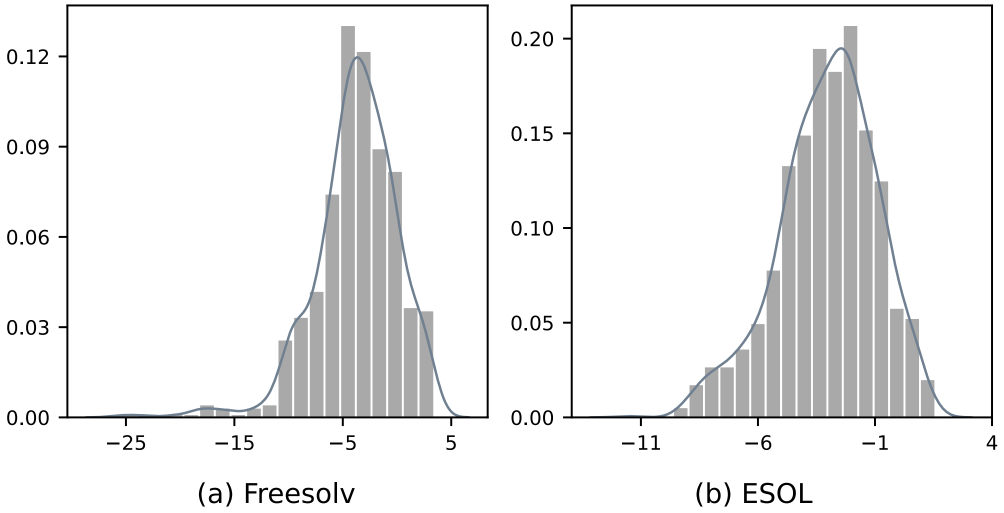
$$\begin{aligned}\widehat{\mathfrak{R}}_n &= \frac{1}{n}\,\mathbb{E}_{\boldsymbol{\sigma}} \sup_{\|W\|_* \le \tau} \Big\langle W, \sum_{i=1}^{n} \sigma_i \mathbf{h}_{G_i}\mathbf{z}_i^\top \Big\rangle_F \\ &\le \frac{\tau}{n}\,\mathbb{E}_{\boldsymbol{\sigma}} \Big\| \sum_{i=1}^{n} \sigma_i \mathbf{h}_{G_i}\mathbf{z}_i^\top \Big\|_{\mathrm{op}}.\end{aligned}$$

Using $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_F$ and the same second-moment calculation as in Theorem 1 gives the stated bound. □

Appendix B: Figures

B.1: Histograms

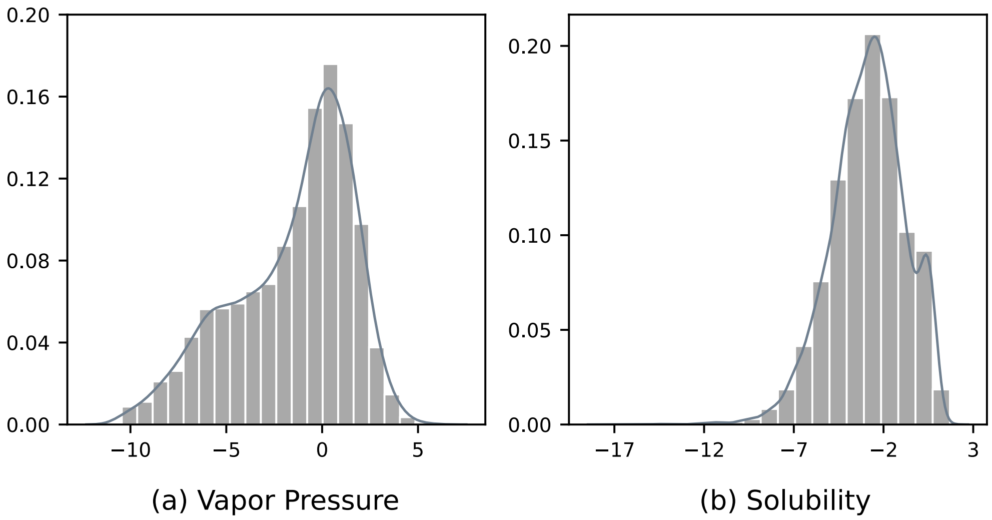
See Figs. 6 and 7.



(a) Freesolv

(b) ESOL

Fig. 6 Distributions of the target property values: the left panel represents the plot for FreeSolv dataset, and the right panel represents the plot for ESOL dataset



(a) Vapor Pressure

(b) Solubility

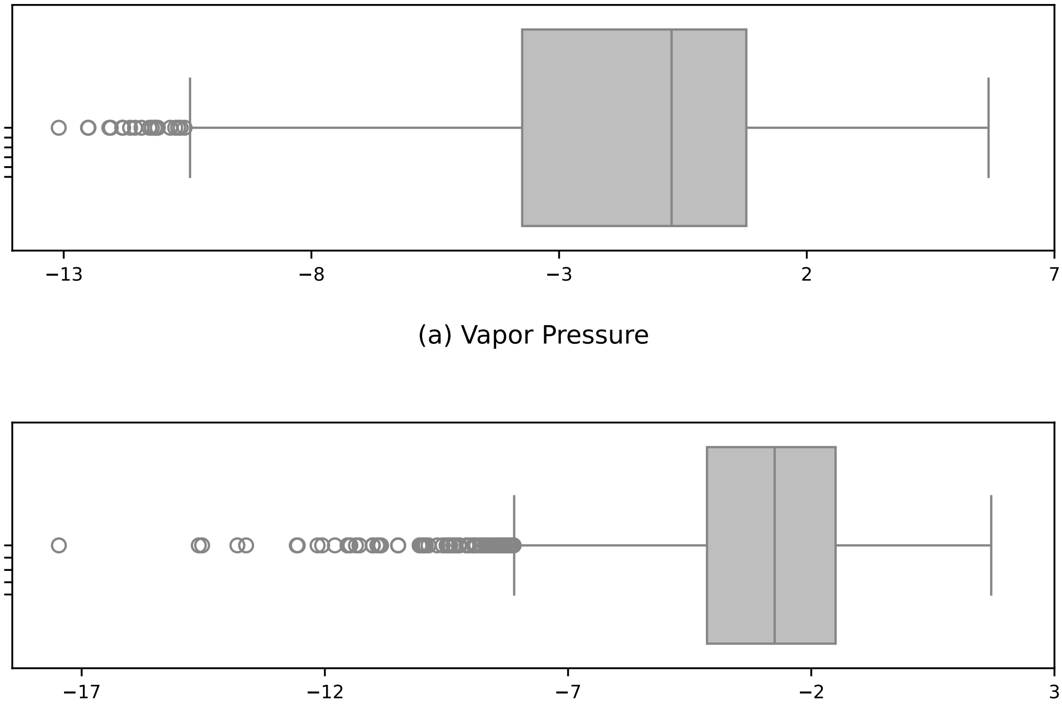
Fig. 7 Distribution of the target property values: the left panel represents the plot for self-curated gas dataset, and the right panel represents the plot for the self-curated solubility dataset

B.2: Boxplot

See Fig. 8.

B.3: PCA results

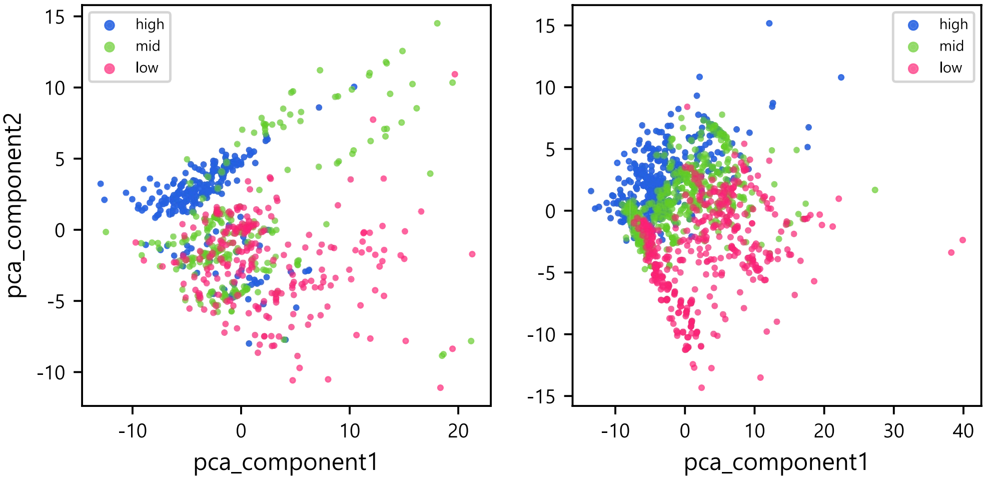
See Figs. 9 and 10.



(a) Vapor Pressure

(b) Solubility

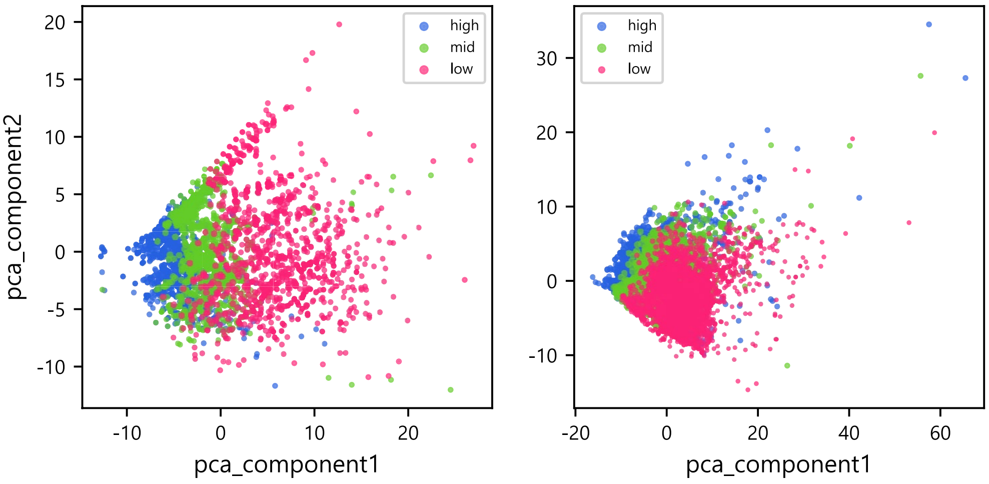
Fig. 8 Boxplot of the log-transformed vapor pressure for the self-curated gas dataset, and of solubility for the self-curated solubility dataset



(a) Freesolv

(b) ESOL

Fig. 9 Two-dimensional PCA results of molecular descriptors, colored by target tertiles. The left panel shows the results for FreeSolv dataset, and the right panel shows the results for ESOL dataset



(a) Vapor Pressure (b) Solubility

Fig. 10 Two-dimensional PCA results of molecular descriptors, colored by target tertiles. The left panel shows the results for gas dataset, and the right panel shows the results for solubility dataset

B.4: Heatmap

See Figs. 11 and 12.

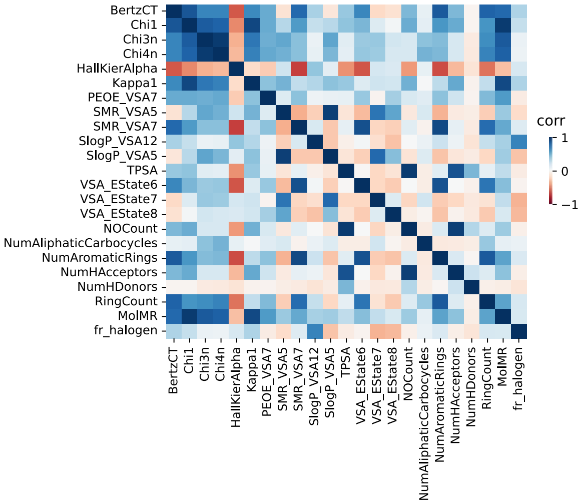


Fig. 11 Heatmap of correlation coefficients between 23 selected descriptors for gas dataset

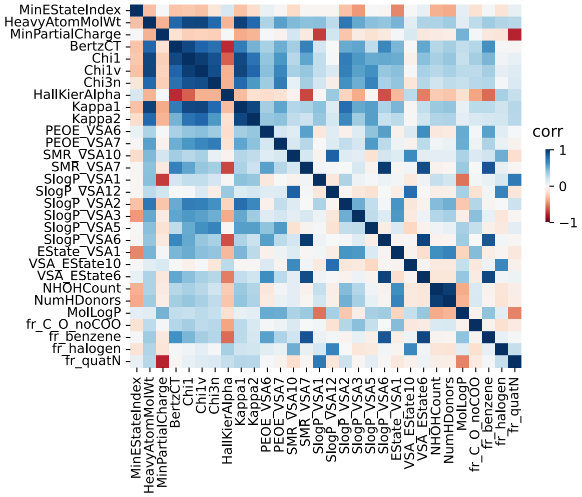


Fig. 12 Heatmap of correlation coefficients between 30 selected descriptors for solubility dataset

Appendix C: Ablation experiments With alternative graph network backbones

C.1: Graph attention network

See Tables 5, 6, 7 and 8.

Table 5 Predictive performance comparison on the FreeSolv dataset using GAT as the backbone

# of Descriptors[1]	GAT [53]	EGAT[2]	EGAT with DS[3]	GAT_{KROVEX}
0	6.097 ± 0.329	–	–	–
	1.867 ± 0.026	–	–	–
	13.501 ± 1.772	–	–	–
	2.792 ± 0.170	–	–	–
1		4.545 ± 0.321	–	–
	–	1.496 ± 0.039	–	–
	–	14.708 ± 2.839	–	–
	–	2.644 ± 0.281	–	–
2		4.772 ± 0.197	–	–
	–	1.651 ± 0.021	–	–
	–	14.389 ± 3.304	–	–
	–	2.504 ± 0.271	–	–
3		4.309 ± 0.258	3.312 ± 0.115	3.140 ± 0.131
	–	1.512 ± 0.035	1.591 ± 0.032	1.330 ± 0.025
	–	14.562 ± 3.046	10.1107 ± 1.799	8.218 ± 0.964
	–	2.524 ± 0.263	2.659 ± 0.217	2.448 ± 0.196
5		–	1.725 ± 0.136	1.417 ± 0.095
	–	–	0.932 ± 0.027	0.817 ± 0.024
	–	–	5.062 ± 1.084	3.898 ± 0.498
	–	–	1.511 ± 0.173	1.509 ± 0.117
7		–	1.522 ± 0.067	1.291 ± 0.083
	–	–	0.934 ± 0.017	0.799 ± 0.018
	–	–	3.650 ± 0.627	3.701 ± 0.464
	–	–	1.413 ± 0.111	1.542 ± 0.157
10		–	1.436 ± 0.063	1.139 ± 0.066
	–	–	0.885 ± 0.023	0.714 ± 0.020
	–	–	2.992 ± 0.418	3.866 ± 0.937
	–	–	1.317 ± 0.088	1.516 ± 0.231
20		–	1.315 ± 0.067	1.161 ± 0.074
	–	–	0.747 ± 0.023	0.631 ± 0.018
	–	–	3.337 ± 0.638	**2.566 ± 0.362**
	–	–	1.279 ± 0.091	**1.074 ± 0.065**
50		–	1.120 ± 0.067	**1.042 ± 0.0735**
	–	–	0.638 ± 0.022	**0.594 ± 0.014**
	–	–	4.209 ± 1.052	2.874 ± 0.428
	–	–	1.277 ± 0.183	1.211 ± 0.117

The results show that KROVEX maintains its performance advantage, demonstrating the robustness and generalizability across different backbone architectures

The first two values represent MSE and MAE under the random split, while the last two values represent MSE and MAE under the scaffold split. All experiments were repeated 15 times, and the results are presented as the means ± standard error for each metric. The best and second-best performances for each metric are marked bold and underlined

[1] Indicates the number of descriptors integrated into the model

[2] Refers to a GAT model that integrates predefined descriptors via concatenation

[3] Indicates descriptor selection

Table 6 Predictive performance comparison on the ESOL dataset using GAT as the backbone

# of Descriptors[1]	GAT [53]	EGAT[2]	EGAT with DS[3]	GAT_{KROVEX}
0 2.975 ± 0.128	–	–	–	
	1.401 ± 0.021	–	–	–
	4.287 ± 0.450	–	–	–
	1.659 ± 0.071	–	–	–
1		1.915 ± 0.078	–	–
	–	1.079 ± 0.019	–	–
	–	2.675 ± 0.324	–	–
	–	1.283 ± 0.073	–	–
2		0.792 ± 0.036	–	–
	–	0.682 ± 0.011	–	–
	–	1.039 ± 0.073	–	–
	–	0.804 ± 0.021	–	–
3		0.793 ± 0.031	0.632 ± 0.020	0.617 ± 0.022
	–	0.679 ± 0.012	0.582 ± 0.008	0.607 ± 0.012
	–	1.094 ± 0.076	1.046 ± 0.122	1.145 ± 0.139
	–	0.840 ± 0.032	0.728 ± 0.032	0.767 ± 0.026
5		–	0.548 ± 0.016	0.591 ± 0.020
	–	–	0.546 ± 0.008	0.585 ± 0.012
	–	–	0.882 ± 0.104	1.017 ± 0.091
	–	–	0.670 ± 0.030	0.748 ± 0.037
7		–	0.522 ± 0.014	0.487 ± 0.016
	–	–	0.525 ± 0.008	0.506 ± 0.008
	–	–	0.771 ± 0.070	1.010 ± 0.100
	–	–	0.658 ± 0.018	0.712 ± 0.025
10		–	0.525 ± 0.012	0.479 ± 0.014
	–	–	0.538 ± 0.010	0.502 ± 0.008
	–	–	0.886 ± 0.093	0.993 ± 0.098
	–	–	0.666 ± 0.025	0.728 ± 0.036
20		–	0.439 ± 0.010	0.445 ± 0.017
	–	–	0.487 ± 0.007	0.488 ± 0.008
	–	–	0.842 ± 0.087	0.899 ± 0.053
	–	–	**0.626 ± 0.019**	0.711 ± 0.645
63		–	0.460 ± 0.016	**0.403 ± 0.015**
	–	–	0.485 ± 0.006	**0.445 ± 0.006**
	–	–	**0.681 ± 0.046**	0.707 ± 0.045
	–	–	0.643 ± 0.021	0.645 ± 0.020

The results show that KROVEX maintains its performance advantage, demonstrating the robustness and generalizability across different backbone architectures

The first two values represent MSE and MAE under the random split, while the last two values represent MSE and MAE under the scaffold split. All experiments were repeated 15 times, and the results are presented as the means ± standard error for each metric. The best and second-best performances for each metric are marked bold and underlined

[1] Indicates the number of descriptors integrated into the model

[2] Refers to a GAT model that integrates predefined descriptors via concatenation

[3] Indicates descriptor selection

Table 7 Predictive performance comparison on the self-curated vapor pressure dataset using GAT as the backbone

# of Descriptors[1]	GAT [53]	EGAT[2]	EGAT with DS[3]	GAT_{KROVEX}
0	5.658 ± 0.117	–	–	–
	1.781 ± 0.017	–	–	–
	8.462 ± 0.715	–	–	–
	2.307 ± 0.115	–	–	–
1		4.189 ± 0.082	–	–
	–	1.498 ± 0.011	–	–
	–	5.359 ± 0.433	–	–
	–	1.811 ± 0.077	–	–
2		1.507 ± 0.033	–	–
	–	0.928 ± 0.026	–	–
	–	2.000 ± 0.124	–	–
	–	1.050 ± 0.027	–	–
3		1.391 ± 0.028	1.134 ± 0.025	1.074 ± 0.024
	–	0.873 ± 0.005	0.788 ± 0.007	0.785 ± 0.001
	–	1.898 ± 0.151	1.579 ± 0.093	1.697 ± 0.090
	–	1.051 ± 0.032	0.977 ± 0.033	1.030 ± 0.031
5		–	0.909 ± 0.028	0.774 ± 0.002
	–	–	0.686 ± 0.006	0.654 ± 0.007
	–	–	1.237 ± 0.082	1.266 ± 0.121
	–	–	0.831 ± 0.018	0.873 ± 0.039
7		–	0.795 ± 0.018	0.661 ± 0.013
	–	–	0.650 ± 0.007	0.585 ± 0.008
	–	–	1.091 ± 0.087	1.251 ± 0.166
	–	–	0.782 ± 0.022	0.801 ± 0.033
10		–	0.706 ± 0.019	0.504 ± 0.011
	–	–	0.618 ± 0.007	0.536 ± 0.005
	–	–	0.945 ± 0.052	0.940 ± 0.098
	–	–	0.739 ± 0.021	0.744 ± 0.044
20		–	0.634 ± 0.020	0.449 ± 0.001
	–	–	0.561 ± 0.005	0.498 ± 0.006
	–	–	0.771 ± 0.049	0.808 ± 0.047
	–	–	0.668 ± 0.017	**0.648 ± 0.020**
23		–	0.593 ± 0.016	**0.447 ± 0.011**
	–	–	0.551 ± 0.006	**0.485 ± 0.006**
	–	–	0.764 ± 0.043	**0.712 ± 0.059**
	–	–	0.660 ± 0.015	0.650 ± 0.026

The results show that KROVEX maintains its performance advantage, demonstrating the robustness and generalizability across different backbone architectures

The first two values represent MSE and MAE under the random split, while the last two values represent MSE and MAE under the scaffold split. All experiments were repeated 15 times, and the results are presented as the means ± standard error for each metric. The best and second-best performances for each metric are marked bold and underlined

[1] Indicates the number of descriptors integrated into the model

[2] Refers to a GAT model that integrates predefined descriptors via concatenation

[3] Indicates descriptor selection

Table 8 Predictive performance comparison on the self-curated solubility dataset using GAT as the backbone

# of Descriptors[1]	GAT [53]	EGAT[2]	EGAT with DS[3]	GAT$_{KROVEX}$
0	2.918 ± 0.049	–	–	–
	1.347 ± 0.009	–	–	–
	3.197 ± 0.146	–	–	–
	1.414 ± 0.037	–	–	–
1		2.236 ± 0.036	–	–
	–	1.120 ± 0.008	–	–
	–	2.437 ± 0.103	–	–
	–	1.170 ± 0.018	–	–
2		1.300 ± 0.043	–	–
	–	0.849 ± 0.014	–	–
	–	1.458 ± 0.039	–	–
	–	0.906 ± 0.014	–	–
3		1.300 ± 0.037	1.106 ± 0.022	1.054 ± 0.026
	–	0.832 ± 0.013	0.762 ± 0.011	0.726 ± 0.007
	–	1.509 ± 0.047	1.360 ± 0.095	1.332 ± 0.082
	–	0.917 ± 0.016	0.814 ± 0.008	0.822 ± 0.013
5		–	1.050 ± 0.023	0.950 ± 0.036
	–	–	0.756 ± 0.011	0.693 ± 0.008
	–	–	1.231 ± 0.030	1.351 ± 0.129
	–	–	0.819 ± 0.012	0.796 ± 0.010
7		–	1.041 ± 0.026	0.750 ± 0.027
	–	–	0.726 ± 0.007	0.610 ± 0.006
	–	–	1.305 ± 0.076	1.060 ± 0.045
	–	–	0.803 ± 0.010	0.750 ± 0.013
10		–	0.958 ± 0.022	0.692 ± 0.022
	–	–	0.699 ± 0.007	0.577 ± 0.005
	–	–	1.130 ± 0.041	0.993 ± 0.046
	–	–	0.762 ± 0.011	0.705 ± 0.010
20		–	0.744 ± 0.022	0.444 ± 0.016
	–	–	0.601 ± 0.007	0.438 ± 0.002
	–	–	0.978 ± 0.128	0.677 ± 0.029
	–	–	0.666 ± 0.012	**0.575 ± 0.011**
30		–	0.781 ± 0.030	**0.444 ± 0.012**
	–	–	0.598 ± 0.007	**0.435 ± 0.004**
	–	–	0.895 ± 0.033	**0.672 ± 0.030**
	–	–	0.673 ± 0.011	0.580 ± 0.011

The results show that KROVEX maintains its performance advantage, demonstrating the robustness and generalizability across different backbone architectures

The first two values represent MSE and MAE under the random split, while the last two values represent MSE and MAE under the scaffold split. All experiments were repeated 15 times, and the results are presented as the means ± standard error for each metric. The best and second-best performances for each metric are marked bold and underlined

[1] Indicates the number of descriptors integrated into the model

[2] Refers to a GAT model that integrates predefined descriptors via concatenation

[3] Indicates descriptor selection

C.2: Graph isomorphism network

See Tables 9, 10, 11 and 12.

Table 9 Predictive performance comparison on the FreeSolv dataset using GIN as the backbone

# of Descriptors[1]	GIN [54]	EGIN[2]	EGIN with DS[3]	GIN$_{KROVEX}$
0	2.542 ± 0.167	–	–	–
	1.116 ± 0.023	–	–	–
	6.193 ± 0.970	–	–	–
	1.742 ± 0.133	–	–	–
1		3.015 ± 0.421	–	–
	–	1.119 ± 0.034	–	–
	–	7.235 ± 2.166	–	–
	–	1.777 ± 0.112	–	–
2		3.053 ± 0.671	–	–
	–	1.068 ± 0.037	–	–
	–	5.712 ± 1.001	–	–
	–	1.748 ± 0.143	–	–
3		2.737 ± 0.430	1.653 ± 0.211	2.131 ± 0.097
	–	1.045 ± 0.032	0.955 ± 0.046	0.961 ± 0.028
	–	6.305 ± 1.192	6.253 ± 1.489	5.392 ± 0.704
	–	1.725 ± 0.115	1.542 ± 0.116	1.868 ± 0.150
5		–	1.427 ± 0.079	1.181 ± 0.070
	–	–	0.838 ± 0.030	0.723 ± 0.018
	–	–	5.180 ± 1.078	3.286 ± 0.387
	–	–	1.417 ± 0.112	1.345 ± 0.107
7		–	1.371 ± 0.104	1.196 ± 0.079
	–	–	0.783 ± 0.023	0.697 ± 0.018
	–	–	5.135 ± 1.227	4.227 ± 0.792
	–	–	1.352 ± 0.120	1.684 ± 0.340
10		–	1.488 ± 0.124	1.719 ± 0.244
	–	–	0.736 ± 0.021	0.888 ± 0.087
	–	–	5.316 ± 1.406	5.283 ± 2.498
	–	–	1.345 ± 0.120	1.396 ± 0.128
20		–	1.208 ± 0.100	1.253 ± 0.092
	–	–	0.670 ± 0.021	0.734 ± 0.046
	–	–	3.867 ± 0.636	3.296 ± 0.544
	–	–	1.435 ± 0.173	1.210 ± 0.095
50		–	1.406 ± 0.191	**0.968 ± 0.065**
	–	–	0.631 ± 0.022	**0.579 ± 0.017**
	–	–	4.843 ± 1.039	**2.916 ± 0.473**
	–	–	1.330 ± 0.158	**1.192 ± 0.066**

The results show that KROVEX maintains its performance advantage, demonstrating the robustness and generalizability across different backbone architectures

The first two values represent MSE and MAE under the random split, while the last two values represent MSE and MAE under the scaffold split. All experiments were repeated 15 times, and the results are presented as the means ± standard error for each metric. The best and second-best performances for each metric are marked bold and underlined

[1] Indicates the number of descriptors integrated into the model

[2] Refers to a GIN model that integrates predefined descriptors via concatenation

[3] Indicates descriptor selection

Table 10 Predictive performance comparison on the ESOL dataset using GIN as the backbone

# of Descriptors[1]	GIN [54]	EGIN[2]	EGIN with DS[3]	GIN_{KROVEX}
0	1.056 ± 0.052	–	–	–
	0.778 ± 0.017	–	–	–
	3.172 ± 0.702	–	–	–
	1.197 ± 0.076	–	–	–
1		1.001 ± 0.063	–	–
	–	0.745 ± 0.013	–	–
	–	3.986 ± 1.423	–	–
	–	1.170 ± 0.088	–	–
2		0.574 ± 0.018	–	–
	–	0.586 ± 0.013	–	–
	–	0.930 ± 0.101	–	–
	–	0.754 ± 0.034	–	–
3		0.585 ± 0.039	0.508 ± 0.024	0.524 ± 0.028
	–	0.561 ± 0.008	0.522 ± 0.011	0.547 ± 0.013
	–	1.040 ± 0.151	1.507 ± 0.417	0.976 ± 0.100
	–	0.789 ± 0.056	0.731 ± 0.044	0.741 ± 0.044
5		–	0.479 ± 0.028	0.487 ± 0.020
	–	–	0.512 ± 0.019	0.536 ± 0.015
	–	–	0.884 ± 0.111	0.888 ± 0.073
	–	–	0.721 ± 0.025	0.713 ± 0.030
7		–	0.479 ± 0.022	0.432 ± 0.010
	–	–	0.501 ± 0.012	0.482 ± 0.005
	–	–	0.887 ± 0.068	1.020 ± 0.167
	–	–	0.681 ± 0.027	0.680 ± 0.024
10		–	0.475 ± 0.029	0.448 ± 0.014
	–	–	0.501 ± 0.012	0.484 ± 0.005
	–	–	1.033 ± 0.267	0.920 ± 0.072
	–	–	0.661 ± 0.035	0.716 ± 0.030
20		–	0.435 ± 0.016	0.433 ± 0.018
	–	–	0.471 ± 0.009	0.471 ± 0.007
	–	–	1.076 ± 0.234	0.863 ± 0.057
	–	–	0.670 ± 0.032	0.698 ± 0.023
63		–	0.479 ± 0.035	**0.416 ± 0.016**
	–	–	0.482 ± 0.009	**0.454 ± 0.006**
	–	–	1.416 ± 0.271	**0.749 ± 0.055**
	–	–	0.752 ± 0.050	**0.652 ± 0.022**

The results show that KROVEX maintains its performance advantage, demonstrating the robustness and generalizability across different backbone architectures

The first two values represent MSE and MAE under the random split, while the last two values represent MSE and MAE under the scaffold split. All experiments were repeated 15 times, and the results are presented as the means ± standard error for each metric. The best and second-best performances for each metric are marked bold and underlined

[1] Indicates the number of descriptors integrated into the model

[2] Refers to a GIN model that integrates predefined descriptors via concatenation

[3] Indicates descriptor selection

Table 11 Predictive performance comparison on the self-curated vapor pressure dataset using GIN as the backbone

# of Descriptors[1]	GIN [54]	EGIN[2]	EGIN with DS[3]	GIN_{KROVEX}
0	2.257 ± 0.104	–	–	–
	1.109 ± 0.020	–	–	–
	5.306 ± 0.496	–	–	–
	1.730 ± 0.054	–	–	–
1		1.971 ± 0.087	–	–
	–	1.004 ± 0.018	–	–
	–	6.625 ± 1.608	–	–
	–	1.635 ± 0.114	–	–
2		0.851 ± 0.021	–	–
	–	0.657 ± 0.007	–	–
	–	1.504 ± 0.164	–	–
	–	0.868 ± 0.024	–	–
3		0.871 ± 0.021	0.843 ± 0.065	0.755 ± 0.022
	–	0.667 ± 0.013	0.630 ± 0.015	0.632 ± 0.004
	–	1.425 ± 0.146	1.474 ± 0.251	1.255 ± 0.125
	–	0.863 ± 0.030	0.873 ± 0.059	0.836 ± 0.030
5		–	0.781 ± 0.033	0.682 ± 0.023
	–	–	0.629 ± 0.025	0.603 ± 0.009
	–	–	1.235 ± 0.098	1.229 ± 0.175
	–	–	0.792 ± 0.021	0.872 ± 0.057
7		–	0.749 ± 0.029	0.554 ± 0.012
	–	–	0.607 ± 0.010	0.548 ± 0.005
	–	–	1.123 ± 0.081	1.050 ± 0.126
	–	–	0.778 ± 0.027	0.775 ± 0.041
10		–	0.664 ± 0.027	0.507 ± 0.013
	–	–	0.570 ± 0.009	0.529 ± 0.005
	–	–	1.031 ± 0.083	0.897 ± 0.084
	–	–	0.720 ± 0.025	0.777 ± 0.051
20		–	0.622 ± 0.028	0.460 ± 0.017
	–	–	0.541 ± 0.005	0.491 ± 0.006
	–	–	0.933 ± 0.068	**0.766 ± 0.049**
	–	–	0.694 ± 0.028	0.715 ± 0.031
23		–	0.656 ± 0.036	**0.449 ± 0.015**
	–	–	0.553 ± 0.012	**0.483 ± 0.005**
	–	–	1.101 ± 0.104	0.813 ± 0.048
	–	–	0.709 ± 0.024	**0.674 ± 0.037**

The results show that KROVEX maintains its performance advantage, demonstrating the robustness and generalizability across different backbone architectures

The first two values represent MSE and MAE under the random split, while the last two values represent MSE and MAE under the scaffold split. All experiments were repeated 15 times, and the results are presented as the means ± standard error for each metric. The best and second-best performances for each metric are marked bold and underlined

[1] Indicates the number of descriptors integrated into the model

[2] Refers to a GIN model that integrates predefined descriptors via concatenation

[3] Indicates descriptor selection

Table 12 Predictive performance comparison on the self-curated solubility dataset using GIN as the backbone

# of Descriptors[1]	GIN [54]	EGIN[2]	EGIN with DS[3]	GIN$_{KROVEX}$
0	1.569 ± 0.067	–	–	–
	0.867 ± 0.008	–	–	–
	2.378 ± 0.196	–	–	–
	1.127 ± 0.042	–	–	–
1		1.208 ± 0.041	–	–
	–	0.775 ± 0.009	–	–
	–	2.326 ± 0.506	–	–
	–	1.011 ± 0.066	–	–
2		0.783 ± 0.041	–	–
	–	0.604 ± 0.583	–	–
	–	0.946 ± 0.051	–	–
	–	0.698 ± 0.017	–	–
3		0.681 ± 0.023	0.790 ± 0.033	0.793 ± 0.034
	–	0.583 ± 0.007	0.632 ± 0.031	0.641 ± 0.011
	–	0.963 ± 0.085	1.060 ± 0.076	1.129 ± 0.126
	–	0.701 ± 0.022	0.742 ± 0.033	0.726 ± 0.014
5		–	0.853 ± 0.053	1.206 ± 0.437
	–	–	0.650 ± 0.034	0.584 ± 0.008
	–	–	1.140 ± 0.094	1.173 ± 0.139
	–	–	0.793 ± 0.038	0.715 ± 0.015
7		–	0.778 ± 0.029	0.623 ± 0.023
	–	–	0.632 ± 0.022	0.546 ± 0.006
	–	–	1.018 ± 0.072	0.941 ± 0.062
	–	–	0.722 ± 0.022	0.679 ± 0.021
10		–	0.780 ± 0.050	0.603 ± 0.020
	–	–	0.589 ± 0.013	0.524 ± 0.007
	–	–	0.952 ± 0.066	0.886 ± 0.055
	–	–	0.780 ± 0.064	0.656 ± 0.014
20		–	0.574 ± 0.018	0.455 ± 0.017
	–	–	0.550 ± 0.012	**0.424 ± 0.003**
	–	–	0.788 ± 0.048	0.676 ± 0.040
	–	–	0.624 ± 0.021	0.569 ± 0.010
30		–	0.571 ± 0.015	**0.432 ± 0.012**
	–	–	0.541 ± 0.011	0.427 ± 0.003
	–	–	0.749 ± 0.038	**0.674 ± 0.044**
	–	–	0.619 ± 0.011	**0.564 ± 0.011**

The results show that KROVEX maintains its performance advantage, demonstrating the robustness and generalizability across different backbone architectures

The first two values represent MSE and MAE under the random split, while the last two values represent MSE and MAE under the scaffold split. All experiments were repeated 15 times, and the results are presented as the means ± standard error for each metric. The best and second-best performances for each metric are marked bold and underlined

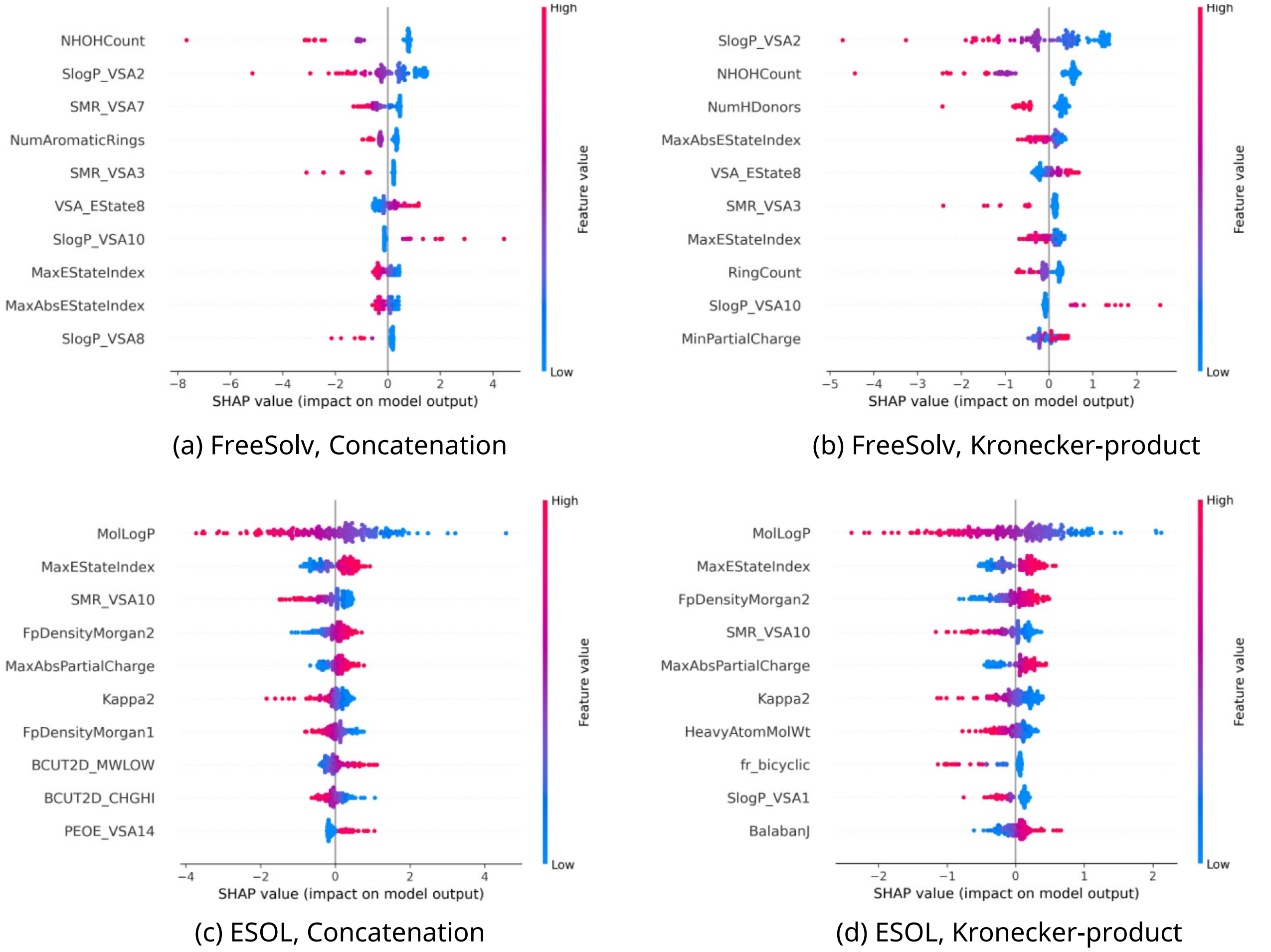
[1] Indicates the number of descriptors integrated into the model

[2] Refers to a GIN model that integrates predefined descriptors via concatenation

[3] Indicates descriptor selection

Appendix D: SHAP analysis

See Figs. 13, 14 and Tables 13, 14.



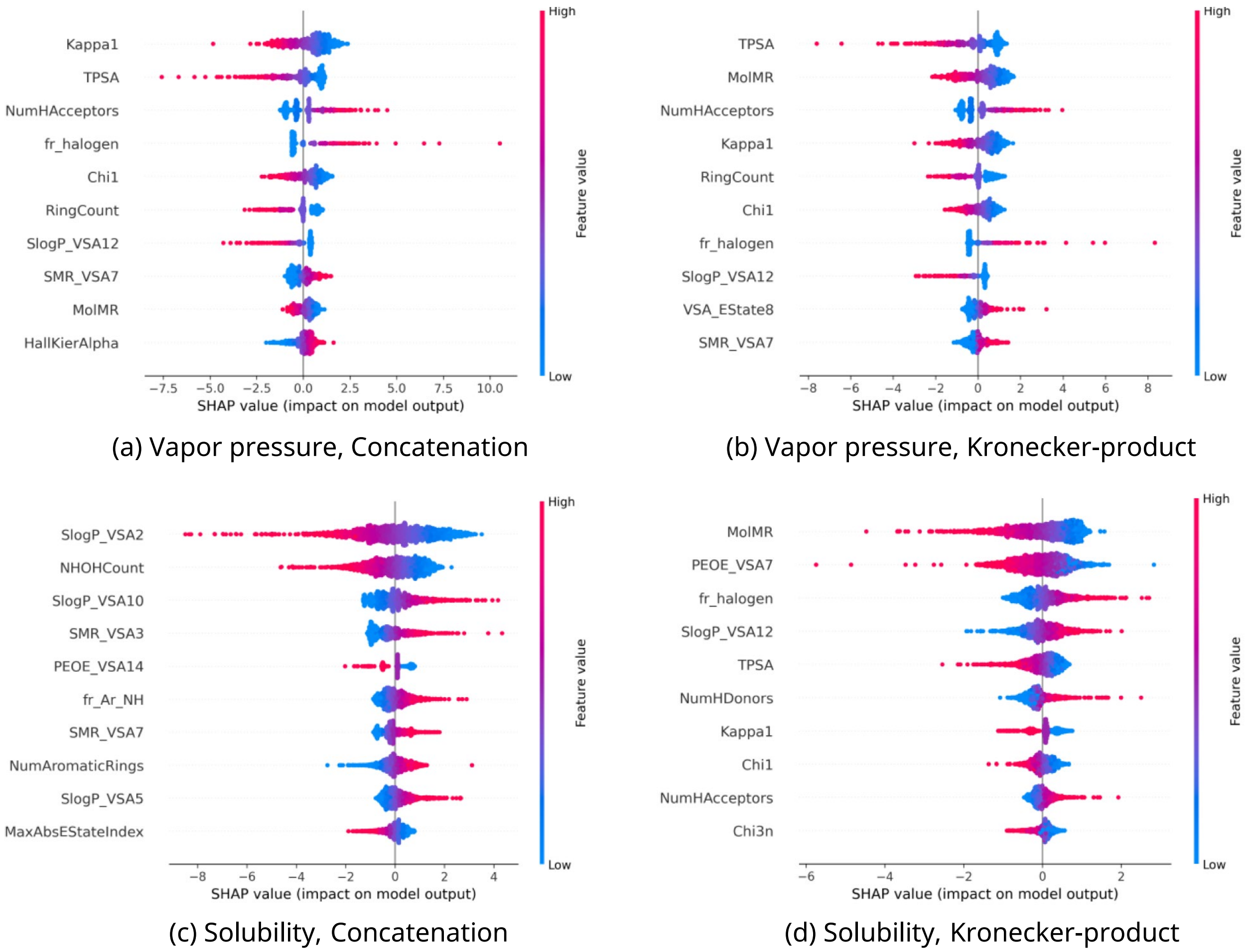
(a) FreeSolv, Concatenation

(b) FreeSolv, Kronecker-product

(c) ESOL, Concatenation

(d) ESOL, Kronecker-product

Fig. 13 SHAP analysis comparing descriptor importance between concatenation and Kronecker-product fusion across Two molecular properties. The y-axis indicates top 10 contributing descriptors for **a**, **b** FreeSolv, and **c**, **d** ESOL. SHAP values (x-axis) indicate feature impact on predictions, with colors representing feature magnitude (pink: high, blue: low). Kronecker fusion (right panels) demonstrates more focused, chemically meaningful descriptor selection compared to concatenation (left panels)



(a) Vapor pressure, Concatenation

(b) Vapor pressure, Kronecker-product

(c) Solubility, Concatenation

(d) Solubility, Kronecker-product

Fig. 14 SHAP analysis comparing descriptor importance between concatenation and Kronecker-product fusion across Two molecular properties. The y-axis indicates top 10 contributing descriptors for **a**, **b** Self-curated vapor pressure, and **c**, **d** Self-curated solubility datasets. SHAP values (x-axis) indicate feature impact on predictions, with colors representing feature magnitude (pink: high, blue: low). Kronecker fusion (right panels) demonstrates more focused, chemically meaningful descriptor selection compared to concatenation (left panels)

Table 13 Mean absolute contributions of top 10 contributing descriptors between concatenation and Kronecker-product fusion for Freesolv and ESOL datasets

FreeSolv			
Concatenation		**Kronecker-product**	
Descriptor	**Contribution**	**Descriptor**	**Contribution**
NHOHCount	14.35%	SlogP_VSAP2	13.15%
SlogP_VSA2	10.77%	NOHOCount	12.40%
SMR_VSA7	6.48%	NumHDonors	6.40%
NumAromaticRings	4.81%	MaxAbsEStateIndex	3.75%
SMR_VSA3	4.73%	VSA_EState8	3.61%
VSA_EState8	4.70%	SMR_VSA3	3.59%
SlogP_VSA10	3.88%	MaxEStateIndex	3.53%
MaxEStateIndex	3.78%	RingCount	3.46%
MaxAbsEStateIndex	3.50%	SlogP_VSA10	3.21%
SlogP_VSA8	3.30%	MinPartialCharge	3.20%
ESOL			
Concatenation		Kronecker-product	
Descriptor	Contribution	Descriptor	Contribution
MolLogP	15.77%	MolLogP	14.57%
MaxEStateIndex	6.45%	MaxEStateIndex	5.87%
SMR_VSA10	5.14%	FpDensityMorgan2	5.10%
FpDensityMorgan2	3.80%	SMR_VSA10	4.89%
MaxAbsPartialCharge	3.55%	MaxAbsPartialCharge	4.72%
Kappa2	3.38%	Kappa2	4.60%
FpDensityMorgan1	3.34%	HeavyAtomMolWt	3.69%
BCUT2D_MWLOW	3.08%	fr_bicyclic	3.57%
BCUT2D_CHGHI	2.98%	SlogP_VSA1	3.51%
PEOE_VSA14	2.95%	BalabanJ	3.40%

Table 14 Mean absolute contributions of top 10 contributing descriptors between concatenation and Kronecker-product fusion for self-curated vapor pressure and solubility datasets

Vapor pressure			
Concatenation		**Kronecker-product**	
Descriptor	**Contribution**	**Descriptor**	**Contribution**
Kappa1	11.20%	TPSA	11.84%
TPSA	10.90%	MolMR	10.44%
NumHAcceptors	9.71%	NumHAcceptors	9.48%
fr_halogen	8.32%	Kappa1	9.37%
Chi1	8.32%	RingCount	7.14%
RingCount	7.21%	Chi1	7.12%
SlogP_VSA12	6.39%	fr_halogen	6.94%
SMR_VSA7	5.69%	SlogP_VSA12	5.86%
MolMR	4.93%	VSA_EState8	4.42%
HallKierAlpha	4.59%	SMR_VSA7	4.18%
Solubility			
Concatenation		Kronecker-product	
Descriptor	Contribution	Descriptor	Contribution
SlogP_VSA2	14.39%	MolMR	12.64%
NHOHCount	9.32%	PEOE_VSA7	9.96%
SlogP_VSA10	7.02%	fr_halogen	7.75%
SMR_VSA3	6.53%	SlogP_VSA12	7.04%
PEOE_VSA14	4.65%	TPSA	5.91%
fr_Ar_NH	4.48%	NumHDonors	5.23%
SMR_VSA7	4.48%	Kappa1	5.08%
NumAromaticRings	4.31%	Chi1	3.78%
SlogP_VSA5	4.03%	NumHAcceptors	3.25%
MaxAbsEStateIndex	3.84%	Chi3n	3.16%

Appendix E: Comparison

See Table 15.

Table 15 Comparison between Chemprop (D-MPNN) and KROVEX on four molecular property datasets

Dataset	Model	R-MSE	R-MAE	S-MSE	S-MAE
Freesolv	Chemprop	1.034 ± 0.133	0.620 ± 0.024	4.446 ± 0.421	1.481 ± 0.113
	KROVEX	0.973 ± 0.072	0.597 ± 0.014	2.606 ± 0.427	1.141 ± 0.076
ESOL	Chemprop	0.432 ± 0.032	0.479 ± 0.018	0.795 ± 0.017	0.637 ± 0.018
	KROVEX	0.423 ± 0.022	0.469 ± 0.013	0.730 ± 0.054	0.628 ± 0.019
Vapor Pressure	Chemprop	0.538 ± 0.022	0.503 ± 0.007	0.963 ± 0.033	0.737 ± 0.014
	KROVEX	0.444 ± 0.009	0.491 ± 0.006	0.770 ± 0.053	0.621 ± 0.037
Solubility	Chemprop	0.586 ± 0.025	0.529 ± 0.005	0.680 ± 0.014	0.593 ± 0.007
	KROVEX	0.448 ± 0.015	0.441 ± 0.004	0.666 ± 0.031	0.573 ± 0.010

Columns use the following abbreviations: R-MSE = Random-split MSE, R-MAE = Random-split MAE, S-MSE = Scaffold-split MSE, S-MAE = Scaffold-split MAE

Appendix F: Summary statistics

F.1: FreeSolv

See Tables 16.

Table 16 Summary statistics of hydration free energy and 209 molecular descriptors

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
Hydration free energy (target)	−3.80	3.85	−25.47	−5.73	−3.53	−1.21	3.43
MaxEStateIndex	6.91	3.36	0.00	3.83	6.16	10.05	13.97
MinEStateIndex	0.11	1.36	−5.97	−0.25	0.36	0.95	4.00
MaxAbsEStateIndex	6.91	3.36	0.00	3.83	6.16	10.05	13.97
MinAbsEStateIndex	0.67	0.62	0.00	0.20	0.54	1.04	5.97
qed	0.49	0.10	0.24	0.43	0.49	0.55	0.89
MolWt	138.95	72.70	16.04	94.12	120.88	159.70	498.66
HeavyAtomMolWt	129.51	72.60	12.01	84.08	112.09	150.09	498.66
ExactMolWt	138.59	72.25	16.03	94.00	120.57	159.60	493.69
NumValenceElectrons	49.00	21.46	8.00	36.00	44.00	58.00	132.00
NumRadicalElectrons	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MaxPartialCharge	0.13	0.14	−0.34	0.01	0.10	0.25	0.56
MinPartialCharge	−0.28	0.15	−0.51	−0.40	−0.30	−0.10	−0.05
MaxAbsPartialCharge	0.29	0.15	0.05	0.12	0.32	0.41	0.56
MinAbsPartialCharge	0.13	0.11	0.00	0.04	0.09	0.21	0.46
FpDensityMorgan1	1.22	0.37	0.23	1.00	1.20	1.44	2.00
FpDensityMorgan2	1.68	0.44	0.36	1.39	1.70	2.00	2.67
FpDensityMorgan3	1.95	0.50	0.50	1.57	2.00	2.33	2.94
BCUT2DMWHI	24.70	19.92	12.01	15.01	16.47	32.24	126.92
BCUT2DMWLOW	10.43	0.97	9.73	10.14	10.27	10.51	32.07
BCUT2DCHGHI	1.90	0.32	−0.34	1.79	1.91	2.07	3.01
BCUT2DCHGLO	−1.91	0.25	−2.49	−2.06	−1.94	−1.84	−0.08
BCUT2DLOGPHI	2.00	0.33	−0.48	1.87	2.02	2.18	2.84
BCUT2DLOGPLOW	−1.86	0.32	−2.82	−2.04	−1.87	−1.71	0.65
BCUT2DMRHI	5.83	1.67	2.13	4.75	5.65	6.29	14.20
BCUT2DMRLOW	0.56	0.77	−0.79	−0.11	0.39	1.10	7.59
BalabanJ	2.75	0.58	0.00	2.40	2.83	3.05	4.81
BertzCT	154.29	167.36	0.00	29.37	89.26	205.79	707.58
Chi0	6.71	3.00	0.00	4.83	5.98	7.81	18.07
Chi0n	5.28	2.21	0.00	3.82	5.01	6.42	13.18
Chi0v	5.84	2.62	0.00	4.25	5.31	6.71	17.34
Chi1	4.12	2.00	0.00	2.81	3.80	4.91	11.15
Chi1n	2.85	1.33	0.00	1.98	2.68	3.58	7.77
Chi1v	3.25	1.67	0.00	2.30	2.93	3.82	14.48
Chi2n	1.94	1.08	0.00	1.22	1.79	2.56	6.31
Chi2v	2.40	1.66	0.00	1.48	2.06	2.85	15.93
Chi3n	1.11	0.82	0.00	0.51	1.01	1.53	5.92
Chi3v	1.44	1.44	0.00	0.68	1.14	1.72	13.57
Chi4n	0.65	0.61	0.00	0.19	0.55	0.91	5.30
Chi4v	0.85	1.15	0.00	0.22	0.61	1.00	13.71
HallKierAlpha	−0.48	0.74	−2.62	−0.98	−0.49	−0.04	2.06
Ipc	2451.55	13171.85	0.00	19.09	64.05	203.57	229928.38
Kappa1	7.05	3.17	0.00	5.05	6.30	8.47	20.84
Kappa2	3.38	2.18	−27.04	2.16	2.96	4.38	11.41
Kappa3	32.88	448.17	−104.04	1.26	2.16	4.00	9507.96
LabuteASA	56.43	26.48	7.50	39.18	50.97	64.40	175.52
PEOEVSA1	3.62	4.30	0.00	0.00	4.74	5.11	30.64
PEOEVSA10	1.76	3.49	0.00	0.00	0.00	0.00	24.42
PEOEVSA11	1.12	3.91	0.00	0.00	0.00	0.00	23.00

Table 16 (continued)

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
PEOEVSA12	0.45	1.85	0.00	0.00	0.00	0.00	17.18
PEOEVSA13	0.64	2.05	0.00	0.00	0.00	0.00	11.81
PEOEVSA14	1.30	3.07	0.00	0.00	0.00	0.00	23.69
PEOEVSA2	2.08	3.88	0.00	0.00	0.00	4.79	20.23
PEOEVSA3	0.44	1.64	0.00	0.00	0.00	0.00	14.38
PEOEVSA4	1.11	3.86	0.00	0.00	0.00	0.00	35.12
PEOEVSA5	1.61	5.36	0.00	0.00	0.00	0.00	46.40
PEOEVSA6	19.56	18.32	0.00	5.70	17.51	30.33	116.01
PEOEVSA7	12.28	10.25	0.00	5.92	12.13	18.55	61.18
PEOEVSA8	5.80	7.23	0.00	0.00	5.02	10.71	39.45
PEOEVSA9	3.85	6.69	0.00	0.00	0.00	6.33	50.23
SMRVSA1	5.17	6.35	0.00	0.00	4.79	9.52	35.12
SMRVSA10	10.44	16.27	0.00	0.00	5.78	11.94	116.01
SMRVSA2	0.10	0.71	0.00	0.00	0.00	0.00	5.26
SMRVSA3	0.93	2.67	0.00	0.00	0.00	0.00	18.69
SMRVSA4	1.00	2.96	0.00	0.00	0.00	0.00	23.67
SMRVSA5	14.96	14.49	0.00	0.00	13.34	26.19	65.21
SMRVSA6	4.79	7.15	0.00	0.00	0.00	6.61	40.27
SMRVSA7	16.47	18.73	0.00	0.00	10.11	30.33	78.37
SMRVSA8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SMRVSA9	1.74	4.38	0.00	0.00	0.00	0.00	23.00
SlogPVSA1	1.30	2.99	0.00	0.00	0.00	0.00	17.07
SlogPVSA10	1.38	4.50	0.00	0.00	0.00	0.00	35.92
SlogPVSA11	1.21	3.86	0.00	0.00	0.00	0.00	23.00
SlogPVSA12	7.48	16.13	0.00	0.00	0.00	11.60	116.01
SlogPVSA2	9.43	9.72	0.00	0.00	6.29	12.71	68.27
SlogPVSA3	3.03	5.33	0.00	0.00	0.00	4.79	41.71
SlogPVSA4	2.84	5.17	0.00	0.00	0.00	5.92	26.15
SlogPVSA5	14.56	14.24	0.00	0.00	12.24	26.19	65.21
SlogPVSA6	12.19	14.81	0.00	0.00	0.00	24.27	67.24
SlogPVSA7	1.42	5.52	0.00	0.00	0.00	0.00	50.23
SlogPVSA8	0.77	3.06	0.00	0.00	0.00	0.00	32.32
SlogPVSA9	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TPSA	20.40	22.64	0.00	0.00	17.07	26.30	123.66
EStateVSA1	2.41	6.02	0.00	0.00	0.00	0.00	44.34
EStateVSA10	3.04	5.46	0.00	0.00	0.00	4.79	35.12
EStateVSA11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EStateVSA2	3.59	7.17	0.00	0.00	0.00	5.81	63.18
EStateVSA3	4.37	7.13	0.00	0.00	0.00	6.09	43.09
EStateVSA4	5.19	6.86	0.00	0.00	0.00	7.02	38.52
EStateVSA5	7.72	10.49	0.00	0.00	4.90	12.13	51.37
EStateVSA6	5.08	8.66	0.00	0.00	0.00	6.92	55.45
EStateVSA7	6.78	10.67	0.00	0.00	0.00	12.13	60.66
EStateVSA8	9.37	12.73	0.00	0.00	4.98	13.50	73.60
EStateVSA9	8.06	16.10	0.00	0.00	0.00	5.73	116.01
VSAEState1	3.86	9.75	−3.70	0.00	0.00	4.46	92.88
VSAEState10	3.05	7.80	0.00	0.00	0.00	1.85	60.86
VSAEState2	4.32	6.90	−2.78	0.00	0.00	9.64	35.91
VSAEState3	2.95	5.51	−4.23	0.00	0.00	5.69	52.18

Table 16 (continued)

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
VSAEState4	0.65	1.70	−4.51	0.00	0.00	0.82	13.36
VSAEState5	0.12	1.20	−23.88	0.00	0.00	0.20	8.00
VSAEState6	3.06	4.62	−0.43	0.00	0.00	6.09	21.86
VSAEState7	1.41	2.85	−9.96	0.00	0.00	2.59	13.50
VSAEState8	2.35	2.57	−3.51	0.00	2.00	4.05	11.47
VSAEState9	0.29	1.03	−3.81	0.00	0.00	0.00	6.92
FractionCSP3	0.57	0.40	0.00	0.14	0.67	1.00	1.00
HeavyAtomCount	8.72	4.19	1.00	6.00	8.00	10.00	24.00
NHOHCount	0.40	0.75	0.00	0.00	0.00	1.00	6.00
NOCount	1.40	1.59	0.00	0.00	1.00	2.00	9.00
NumAliphaticCarbocycles	0.09	0.35	0.00	0.00	0.00	0.00	4.00
NumAliphaticHeterocycles	0.06	0.23	0.00	0.00	0.00	0.00	1.00
NumAliphaticRings	0.14	0.43	0.00	0.00	0.00	0.00	5.00
NumAromaticCarbocycles	0.44	0.68	0.00	0.00	0.00	1.00	4.00
NumAromaticHeterocycles	0.08	0.28	0.00	0.00	0.00	0.00	2.00
NumAromaticRings	0.52	0.69	0.00	0.00	0.00	1.00	4.00
NumHAcceptors	1.31	1.44	0.00	0.00	1.00	2.00	8.00
NumHDonors	0.34	0.63	0.00	0.00	0.00	1.00	6.00
NumHeteroatoms	2.21	2.17	0.00	1.00	2.00	3.00	11.00
NumRotatableBonds	1.63	1.97	0.00	0.00	1.00	3.00	12.00
NumSaturatedCarbocycles	0.05	0.24	0.00	0.00	0.00	0.00	3.00
NumSaturatedHeterocycles	0.03	0.18	0.00	0.00	0.00	0.00	1.00
NumSaturatedRings	0.08	0.32	0.00	0.00	0.00	0.00	4.00
RingCount	0.67	0.81	0.00	0.00	1.00	1.00	5.00
MolLogP	1.93	1.49	−3.59	1.16	1.78	2.58	9.89
MolMR	36.58	16.61	5.02	25.61	33.61	42.47	101.98
frAlCOO	0.02	0.12	0.00	0.00	0.00	0.00	1.00
frAlOH	0.10	0.45	0.00	0.00	0.00	0.00	6.00
frAlOHnoTert	0.10	0.44	0.00	0.00	0.00	0.00	6.00
frArN	0.03	0.19	0.00	0.00	0.00	0.00	2.00
frArCOO	0.01	0.08	0.00	0.00	0.00	0.00	1.00
frArN	0.13	0.48	0.00	0.00	0.00	0.00	4.00
frArNH	0.03	0.24	0.00	0.00	0.00	0.00	3.00
frArOH	0.08	0.27	0.00	0.00	0.00	0.00	1.00
frCOO	0.02	0.15	0.00	0.00	0.00	0.00	1.00
frCOO2	0.02	0.15	0.00	0.00	0.00	0.00	1.00
frCO	0.26	0.50	0.00	0.00	0.00	0.00	3.00
frCOnoCOO	0.24	0.48	0.00	0.00	0.00	0.00	3.00
frCS	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frHOCCN	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frImine	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frNH0	0.23	0.59	0.00	0.00	0.00	0.00	4.00
frNH1	0.08	0.33	0.00	0.00	0.00	0.00	3.00
frNH2	0.06	0.25	0.00	0.00	0.00	0.00	2.00
frNO	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frNdealkylation1	0.01	0.10	0.00	0.00	0.00	0.00	2.00
frNdealkylation2	0.00	0.06	0.00	0.00	0.00	0.00	1.00
frNhpyrrole	0.03	0.24	0.00	0.00	0.00	0.00	3.00
frSH	0.01	0.09	0.00	0.00	0.00	0.00	1.00

Table 16 (continued)

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
fraldehyde	0.04	0.19	0.00	0.00	0.00	0.00	1.00
fralkylcarbamate	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fralkylhalide	0.36	1.07	0.00	0.00	0.00	0.00	8.00
frallylicoxid	0.13	0.54	0.00	0.00	0.00	0.00	6.00
framide	0.04	0.23	0.00	0.00	0.00	0.00	2.00
framidine	0.00	0.00	0.00	0.00	0.00	0.00	0.00
franiline	0.07	0.29	0.00	0.00	0.00	0.00	2.00
frarylmethyl	0.17	0.49	0.00	0.00	0.00	0.00	3.00
frazide	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frazo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frbarbitur	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frbenzene	0.44	0.68	0.00	0.00	0.00	1.00	4.00
frbenzodiazepine	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frbicyclic	0.15	0.71	0.00	0.00	0.00	0.00	12.00
frdiazo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frdihydropyridine	0.00	0.00	0.00	0.00	0.00	0.00	0.00
froxazole	0.00	0.00	0.00	0.00	0.00	0.00	0.00
froxime	0.00	0.04	0.00	0.00	0.00	0.00	1.00
frparahydroxylation	0.09	0.38	0.00	0.00	0.00	0.00	4.00
frphenol	0.08	0.27	0.00	0.00	0.00	0.00	1.00
frphenolnoOrthoHbond	0.07	0.26	0.00	0.00	0.00	0.00	1.00
frphosacid	0.00	0.07	0.00	0.00	0.00	0.00	1.00
frphosester	0.00	0.07	0.00	0.00	0.00	0.00	1.00
frpiperdine	0.00	0.06	0.00	0.00	0.00	0.00	1.00
frpiperzine	0.00	0.07	0.00	0.00	0.00	0.00	1.00
frpriamide	0.01	0.08	0.00	0.00	0.00	0.00	1.00
frprisulfonamd	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frpyridine	0.04	0.19	0.00	0.00	0.00	0.00	1.00
frquatN	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frsulfide	0.02	0.14	0.00	0.00	0.00	0.00	1.00
frsulfonamd	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frsulfone	0.00	0.07	0.00	0.00	0.00	0.00	1.00
frtermacetylene	0.01	0.10	0.00	0.00	0.00	0.00	1.00
frtetrazole	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frthiazole	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frthiocyan	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frthiophene	0.00	0.06	0.00	0.00	0.00	0.00	1.00
frunbrchalkane	0.29	0.87	0.00	0.00	0.00	0.00	6.00
frurea	0.00	0.04	0.00	0.00	0.00	0.00	1.00

STD indicates the standard deviation, measuring the spread of a variable around its mean. MIN and MAX indicate the minimum and maximum values observed for each variable, respectively. Quartiles divide the data into four equal parts when sorted in ascending order: Q1 represents the 25th percentile, Q2 the median (50th percentile), and Q3 the 75th percentile

F.2: ESOL

See Table 17.

Table 17 Summary statistics of Solubility and 209 molecular descriptors

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
Solubility (target)	−3.05	2.10	−11.60	−4.32	−2.86	−1.60	1.58
MaxEStateIndex	8.24	3.71	0.00	5.05	9.19	11.48	17.26
MinEStateIndex	−0.18	1.29	−5.57	−0.61	−0.05	0.74	4.00
MaxAbsEStateIndex	8.24	3.71	0.00	5.05	9.19	11.48	17.26
MinAbsEStateIndex	0.49	0.49	0.00	0.11	0.31	0.79	4.00
qed	0.55	0.15	0.15	0.45	0.53	0.65	0.93
MolWt	203.94	102.74	16.04	121.18	182.18	270.37	780.95
HeavyAtomMolWt	191.49	99.32	12.01	112.09	172.10	256.66	716.44
ExactMolWt	203.46	102.40	16.03	121.09	182.08	270.16	780.43
NumValenceElectrons	72.06	36.15	8.00	44.00	64.00	94.00	312.00
NumRadicalElectrons	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MaxPartialCharge	0.17	0.14	−0.12	0.05	0.16	0.30	0.65
MinPartialCharge	−0.30	0.15	−0.75	−0.41	−0.34	−0.12	−0.04
MaxAbsPartialCharge	0.31	0.15	0.04	0.14	0.35	0.43	0.75
MinAbsPartialCharge	0.17	0.12	0.00	0.05	0.15	0.28	0.44
FpDensityMorgan1	1.13	0.36	0.19	0.90	1.14	1.35	2.00
FpDensityMorgan2	1.66	0.44	0.31	1.39	1.73	2.00	2.67
FpDensityMorgan3	2.07	0.54	0.42	1.73	2.15	2.47	3.23
BCUT2DMWHI	25.66	19.01	12.01	16.16	16.53	35.50	126.92
BCUT2DMWLOW	10.20	0.39	9.41	10.02	10.16	10.33	12.01
BCUT2DCHGHI	2.07	0.31	−0.08	1.89	2.06	2.23	3.32
BCUT2DCHGLO	−2.05	0.25	−2.81	−2.19	−2.04	−1.92	−0.08
BCUT2DLOGPHI	2.17	0.28	0.14	2.03	2.18	2.34	3.13
BCUT2DLOGPLOW	−2.02	0.34	−3.03	−2.26	−2.00	−1.80	0.14
BCUT2DMRHI	6.18	1.51	2.50	5.36	5.91	6.41	14.20
BCUT2DMRLOW	0.48	0.70	−0.80	−0.12	0.25	1.00	3.31
BalabanJ	2.68	0.57	0.00	2.30	2.72	3.03	4.81
BertzCT	330.45	275.39	0.00	78.99	275.71	513.73	1552.14
Chi0	9.88	4.85	0.00	5.98	8.67	13.10	39.19
Chi0n	7.75	3.88	0.00	4.91	7.08	9.93	32.90
Chi0v	8.48	4.04	0.00	5.36	7.51	11.10	32.90
Chi1	6.30	3.28	0.00	3.79	5.61	8.42	26.01
Chi1n	4.37	2.43	0.00	2.62	3.95	5.56	20.94
Chi1v	4.90	2.61	0.00	2.95	4.26	6.49	20.94
Chi2n	3.24	2.18	0.00	1.76	2.85	4.21	18.82
Chi2v	3.89	2.54	0.00	2.09	3.26	4.93	18.82
Chi3n	2.18	1.91	0.00	0.99	1.71	2.81	15.95
Chi3v	2.68	2.28	0.00	1.17	2.04	3.47	17.29
Chi4n	1.47	1.59	0.00	0.52	1.02	1.86	12.78
Chi4v	1.81	1.92	0.00	0.64	1.24	2.41	16.03
HallKierAlpha	−0.92	1.01	−4.68	−1.59	−0.92	−0.11	3.48
Ipc	2.17×10^{9}	5.24×10^{10}	0.00	64.05	505.76	10105.88	1.46×10^{12}
Kappa1	9.99	4.95	0.00	6.16	8.64	12.97	40.51
Kappa2	4.36	2.48	0.00	2.57	3.85	5.60	25.00
Kappa3	4.09	33.72	−27.04	1.39	2.27	3.74	1128.96
LabuteASA	83.33	40.44	8.74	51.72	73.14	111.84	323.32
PEOEVSA1	5.43	7.43	0.00	0.00	4.74	9.47	79.86
PEOEVSA10	3.73	6.75	0.00	0.00	0.00	5.75	73.75
PEOEVSA11	1.70	4.10	0.00	0.00	0.00	0.00	35.58

Table 17 (continued)

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
PEOEVSA12	1.64	3.83	0.00	0.00	0.00	0.00	20.41
PEOEVSA13	1.01	2.83	0.00	0.00	0.00	0.00	34.83
PEOEVSA14	1.87	3.48	0.00	0.00	0.00	5.63	23.63
PEOEVSA2	3.57	5.83	0.00	0.00	0.00	4.79	40.45
PEOEVSA3	2.01	4.03	0.00	0.00	0.00	4.39	28.11
PEOEVSA4	0.99	3.35	0.00	0.00	0.00	0.00	22.95
PEOEVSA5	1.61	6.43	0.00	0.00	0.00	0.00	92.81
PEOEVSA6	25.23	21.57	0.00	11.60	20.27	36.40	167.95
PEOEVSA7	20.51	18.10	0.00	6.42	16.38	28.52	113.22
PEOEVSA8	7.91	8.71	0.00	0.00	6.07	12.39	61.84
PEOEVSA9	5.57	7.86	0.00	0.00	0.00	9.85	54.43
SMRVSA1	7.34	9.05	0.00	0.00	4.92	9.90	79.86
SMRVSA10	16.80	18.06	0.00	0.00	11.60	23.32	139.21
SMRVSA2	0.12	0.84	0.00	0.00	0.00	0.00	10.52
SMRVSA3	2.55	4.87	0.00	0.00	0.00	4.90	29.53
SMRVSA4	2.82	6.84	0.00	0.00	0.00	0.00	39.92
SMRVSA5	19.74	22.69	0.00	0.00	13.66	31.50	197.18
SMRVSA6	5.39	8.02	0.00	0.00	0.00	7.11	64.75
SMRVSA7	25.68	23.38	0.00	0.00	25.31	38.42	149.19
SMRVSA8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SMRVSA9	2.31	4.72	0.00	0.00	0.00	0.00	40.07
SlogPVSA1	2.90	4.58	0.00	0.00	0.00	5.32	24.84
SlogPVSA10	2.59	5.09	0.00	0.00	0.00	4.79	35.92
SlogPVSA11	1.34	3.50	0.00	0.00	0.00	0.00	28.75
SlogPVSA12	9.27	17.72	0.00	0.00	0.00	11.60	139.21
SlogPVSA2	14.09	16.20	0.00	4.43	11.21	19.28	168.11
SlogPVSA3	4.54	6.89	0.00	0.00	0.00	6.72	41.71
SlogPVSA4	4.81	8.36	0.00	0.00	0.00	6.92	47.44
SlogPVSA5	18.45	20.61	0.00	0.00	13.09	27.69	167.95
SlogPVSA6	20.43	19.95	0.00	0.00	18.20	30.33	126.68
SlogPVSA7	1.81	5.48	0.00	0.00	0.00	0.00	50.23
SlogPVSA8	2.53	6.97	0.00	0.00	0.00	0.00	64.63
SlogPVSA9	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TPSA	34.87	35.38	0.00	0.00	26.30	55.44	268.68
EStateVSA1	5.83	12.25	0.00	0.00	0.00	6.09	111.94
EStateVSA10	5.19	7.30	0.00	0.00	4.39	9.59	56.17
EStateVSA11	0.04	0.45	0.00	0.00	0.00	0.00	8.78
EStateVSA2	6.20	8.84	0.00	0.00	4.62	10.83	61.35
EStateVSA3	7.21	8.84	0.00	0.00	5.56	12.01	60.28
EStateVSA4	7.81	10.15	0.00	0.00	5.57	11.84	64.21
EStateVSA5	9.97	13.91	0.00	0.00	6.07	13.89	154.10
EStateVSA6	7.51	11.10	0.00	0.00	0.00	12.14	62.63
EStateVSA7	10.48	15.00	0.00	0.00	0.00	18.20	97.07
EStateVSA8	12.42	15.69	0.00	0.00	6.92	16.18	84.93
EStateVSA9	10.11	17.11	0.00	0.00	4.74	11.60	139.21
VSAEState1	4.21	9.72	−3.84	0.00	0.00	4.82	96.82
VSAEState10	3.88	8.59	0.00	0.00	0.00	5.19	79.99
VSAEState2	8.52	10.88	−14.32	0.00	3.28	12.94	67.99
VSAEState3	5.43	9.50	−9.32	0.00	2.13	8.59	108.91

Table 17 (continued)

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
VSAEState4	0.84	2.11	−7.26	0.00	0.00	1.35	19.28
VSAEState5	0.17	1.04	−9.68	0.00	0.00	0.36	5.90
VSAEState6	5.06	6.45	−2.64	0.00	2.13	8.05	32.11
VSAEState7	1.77	3.78	−21.77	0.00	0.00	2.70	35.39
VSAEState8	2.52	2.87	−5.45	0.00	1.97	4.35	13.27
VSAEState9	0.24	1.34	−9.42	0.00	0.00	0.00	11.47
FractionCSP3	0.44	0.37	0.00	0.10	0.36	0.80	1.00
HeavyAtomCount	13.29	6.88	1.00	8.00	12.00	18.00	55.00
NHOHCount	0.77	1.21	0.00	0.00	0.00	1.00	11.00
NOCount	2.43	2.42	0.00	0.00	2.00	4.00	16.00
NumAliphaticCarbocycles	0.29	0.91	0.00	0.00	0.00	0.00	6.00
NumAliphaticHeterocycles	0.17	0.45	0.00	0.00	0.00	0.00	4.00
NumAliphaticRings	0.46	1.05	0.00	0.00	0.00	0.00	8.00
NumAromaticCarbocycles	0.75	0.94	0.00	0.00	1.00	1.00	7.00
NumAromaticHeterocycles	0.19	0.47	0.00	0.00	0.00	0.00	2.00
NumAromaticRings	0.93	1.01	0.00	0.00	1.00	2.00	7.00
NumHAcceptors	2.11	2.15	0.00	0.00	2.00	3.00	16.00
NumHDonors	0.70	1.09	0.00	0.00	0.00	1.00	11.00
NumHeteroatoms	3.35	2.79	0.00	1.00	3.00	5.00	16.00
NumRotatableBonds	2.18	2.64	0.00	0.00	1.00	3.00	23.00
NumSaturatedCarbocycles	0.20	0.71	0.00	0.00	0.00	0.00	6.00
NumSaturatedHeterocycles	0.11	0.37	0.00	0.00	0.00	0.00	3.00
NumSaturatedRings	0.31	0.84	0.00	0.00	0.00	0.00	7.00
RingCount	1.39	1.32	0.00	0.00	1.00	2.00	8.00
MolLogP	2.45	1.85	−7.57	1.41	2.34	3.40	10.39
MolMR	54.02	25.28	6.73	34.42	48.52	72.26	192.61
frAlCOO	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frAlOH	0.25	0.87	0.00	0.00	0.00	0.00	11.00
frAlOHnoTert	0.21	0.82	0.00	0.00	0.00	0.00	11.00
frArN	0.04	0.23	0.00	0.00	0.00	0.00	3.00
frArCOO	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frArN	0.30	0.85	0.00	0.00	0.00	0.00	6.00
frArNH	0.04	0.25	0.00	0.00	0.00	0.00	4.00
frArOH	0.10	0.37	0.00	0.00	0.00	0.00	5.00
frCOO	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frCOO2	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frCO	0.53	0.86	0.00	0.00	0.00	1.00	6.00
frCOnoCOO	0.53	0.86	0.00	0.00	0.00	1.00	6.00
frCS	0.01	0.12	0.00	0.00	0.00	0.00	2.00
frHOCCN	0.00	0.03	0.00	0.00	0.00	0.00	1.00
frImine	0.01	0.12	0.00	0.00	0.00	0.00	2.00
frNH0	0.51	1.00	0.00	0.00	0.00	1.00	6.00
frNH1	0.27	0.61	0.00	0.00	0.00	0.00	4.00
frNH2	0.08	0.30	0.00	0.00	0.00	0.00	3.00
frNO	0.01	0.12	0.00	0.00	0.00	0.00	2.00
frNdealkylation1	0.01	0.10	0.00	0.00	0.00	0.00	1.00
frNdealkylation2	0.00	0.04	0.00	0.00	0.00	0.00	1.00
frNhpyrrole	0.04	0.25	0.00	0.00	0.00	0.00	4.00
frSH	0.00	0.07	0.00	0.00	0.00	0.00	1.00

Table 17 (continued)

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
fraldehyde	0.02	0.12	0.00	0.00	0.00	0.00	1.00
fralkylcarbamate	0.00	0.03	0.00	0.00	0.00	0.00	1.00
fralkylhalide	0.26	0.96	0.00	0.00	0.00	0.00	12.00
frallylicoxid	0.22	0.73	0.00	0.00	0.00	0.00	8.00
framide	0.38	0.94	0.00	0.00	0.00	0.00	8.00
framidine	0.00	0.00	0.00	0.00	0.00	0.00	0.00
franiline	0.22	0.58	0.00	0.00	0.00	0.00	4.00
frarylmethyl	0.18	0.53	0.00	0.00	0.00	0.00	4.00
frazide	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frazo	0.00	0.03	0.00	0.00	0.00	0.00	1.00
frbarbitur	0.03	0.18	0.00	0.00	0.00	0.00	2.00
frbenzene	0.75	0.94	0.00	0.00	1.00	1.00	7.00
frbenzodiazepine	0.01	0.07	0.00	0.00	0.00	0.00	1.00
frbicyclic	0.58	1.40	0.00	0.00	0.00	0.00	12.00
frdiazo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frdihydropyridine	0.00	0.03	0.00	0.00	0.00	0.00	1.00
frepoxide	0.00	0.05	0.00	0.00	0.00	0.00	1.00
frester	0.10	0.34	0.00	0.00	0.00	0.00	3.00
frether	0.30	0.76	0.00	0.00	0.00	0.00	9.00
frfuran	0.01	0.08	0.00	0.00	0.00	0.00	1.00
frguanido	0.00	0.05	0.00	0.00	0.00	0.00	1.00
frhalogen	0.74	1.54	0.00	0.00	0.00	1.00	12.00
frhdrzine	0.01	0.10	0.00	0.00	0.00	0.00	1.00
frhdrzone	0.00	0.03	0.00	0.00	0.00	0.00	1.00
frimidazole	0.01	0.12	0.00	0.00	0.00	0.00	1.00
frimide	0.09	0.40	0.00	0.00	0.00	0.00	4.00
frisocyan	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frisothiocyan	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frketone	0.11	0.40	0.00	0.00	0.00	0.00	3.00
frketoneTopliss	0.07	0.27	0.00	0.00	0.00	0.00	2.00
frlactam	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frlactone	0.01	0.08	0.00	0.00	0.00	0.00	1.00
frmethoxy	0.07	0.30	0.00	0.00	0.00	0.00	3.00
frmorpholine	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frnitrile	0.02	0.16	0.00	0.00	0.00	0.00	2.00
frnitro	0.07	0.31	0.00	0.00	0.00	0.00	3.00
frnitroarom	0.06	0.30	0.00	0.00	0.00	0.00	3.00
frnitroaromnonortho	0.03	0.19	0.00	0.00	0.00	0.00	3.00
frnitroso	0.00	0.00	0.00	0.00	0.00	0.00	0.00
froxazole	0.00	0.05	0.00	0.00	0.00	0.00	1.00
froxime	0.00	0.06	0.00	0.00	0.00	0.00	1.00
frparahydroxylation	0.13	0.42	0.00	0.00	0.00	0.00	3.00
frphenol	0.09	0.35	0.00	0.00	0.00	0.00	4.00
frphenolnoOrthoHbond	0.09	0.35	0.00	0.00	0.00	0.00	4.00
frphosacid	0.01	0.07	0.00	0.00	0.00	0.00	1.00
frphosester	0.01	0.07	0.00	0.00	0.00	0.00	1.00
frpiperdine	0.01	0.08	0.00	0.00	0.00	0.00	1.00
frpiperzine	0.00	0.03	0.00	0.00	0.00	0.00	1.00
frpriamide	0.01	0.13	0.00	0.00	0.00	0.00	2.00

Table 17 (continued)

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
frprisulfonamd	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frpyridine	0.05	0.26	0.00	0.00	0.00	0.00	2.00
frquatN	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frsulfide	0.03	0.17	0.00	0.00	0.00	0.00	1.00
frsulfonamd	0.02	0.17	0.00	0.00	0.00	0.00	2.00
frsulfone	0.00	0.06	0.00	0.00	0.00	0.00	1.00
frtermacetylene	0.01	0.11	0.00	0.00	0.00	0.00	1.00
frtetrazole	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frthiazole	0.01	0.08	0.00	0.00	0.00	0.00	1.00
frthiocyan	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frthiophene	0.00	0.04	0.00	0.00	0.00	0.00	1.00
frunbrchalkane	0.30	1.41	0.00	0.00	0.00	0.00	21.00
frurea	0.07	0.27	0.00	0.00	0.00	0.00	2.00

STD indicates the standard deviation, measuring the spread of a variable around its mean. MIN and MAX indicate the minimum and maximum values observed for each variable, respectively. Quartiles divide the data into four equal parts when sorted in ascending order: Q1 represents the 25th percentile, Q2 the median (50th percentile), and Q3 the 75th percentile

F.3: Self-curated gas

See Table 18.

Table 18 Summary statistics of Vapor Pressure and 209 molecular descriptors

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
Vapor pressure (target)	−1.53	3.09	−10.45	−3.69	−0.70	0.79	5.67
MaxEStateIndex	7.09	3.61	0.00	3.66	6.55	10.37	14.27
MinEStateIndex	−0.14	1.54	−9.81	−0.51	0.29	0.87	3.50
MaxAbsEStateIndex	7.09	3.61	0.00	3.66	6.55	10.37	14.27
MinAbsEStateIndex	0.65	0.66	0.00	0.19	0.48	1.00	7.66
qed	0.51	0.12	0.09	0.43	0.50	0.58	0.94
MolWt	172.91	92.31	16.04	110.10	144.16	218.04	943.17
HeavyAtomMolWt	160.60	90.25	12.01	96.09	132.10	204.19	943.17
ExactMolWt	172.50	91.89	16.03	109.97	143.95	217.97	933.18
NumValenceElectrons	61.59	29.09	7.00	42.00	54.00	76.00	242.00
NumRadicalElectrons	0.00	0.05	0.00	0.00	0.00	0.00	2.00
MaxPartialCharge	0.14	0.16	−0.41	0.00	0.12	0.29	1.00
MinPartialCharge	−0.28	0.16	−1.00	−0.42	−0.30	−0.10	0.00
MaxAbsPartialCharge	0.29	0.16	0.00	0.10	0.33	0.45	1.00
MinAbsPartialCharge	0.15	0.13	0.00	0.04	0.12	0.26	0.54
FpDensityMorgan1	1.15	0.39	0.15	0.88	1.15	1.40	2.00
FpDensityMorgan2	1.63	0.47	0.24	1.31	1.67	2.00	2.67
FpDensityMorgan3	1.95	0.53	0.31	1.57	2.00	2.36	3.15
BCUT2DMWHI	24.84	17.54	12.01	14.64	16.50	32.92	126.92
BCUT2DMWLOW	10.36	1.97	9.41	10.06	10.20	10.38	79.90
BCUT2DCHGHI	2.00	0.28	−0.41	1.87	1.99	2.17	3.32
BCUT2DCHGLO	−2.00	0.23	−2.81	−2.13	−2.01	−1.91	0.00
BCUT2DLOGPHI	2.11	0.29	−0.48	1.98	2.12	2.28	3.13
BCUT2DLOGPLOW	−1.94	0.29	−3.03	−2.12	−1.95	−1.80	0.85
BCUT2DMRHI	5.97	1.51	0.82	4.87	5.70	6.43	14.20
BCUT2DMRLOW	0.47	0.67	−1.11	−0.01	0.33	0.77	8.93
BalabanJ	2.84	0.64	0.00	2.43	2.85	3.15	8.21
BertzCT	211.61	213.48	0.00	52.74	136.72	311.80	1219.17
Chi0	8.37	3.92	0.00	5.70	7.40	10.49	33.58
Chi0n	6.73	3.17	0.00	4.57	6.13	8.25	24.86
Chi0v	7.43	3.60	0.00	5.01	6.51	9.16	32.86
Chi1	5.18	2.61	0.00	3.31	4.46	6.79	17.64
Chi1n	3.72	1.94	0.00	2.41	3.43	4.68	16.41
Chi1v	4.35	2.67	0.00	2.71	3.68	5.38	35.57
Chi2n	2.68	1.59	0.00	1.60	2.43	3.47	13.34
Chi2v	3.46	2.85	0.00	1.89	2.82	4.32	44.30
Chi3n	1.59	1.11	0.00	0.77	1.42	2.21	7.71
Chi3v	2.19	2.25	0.00	0.96	1.65	2.84	35.39
Chi4n	0.97	0.84	0.00	0.35	0.78	1.40	6.34
Chi4v	1.42	2.39	0.00	0.42	0.94	1.81	47.12
HallKierAlpha	−0.55	0.82	−4.01	−1.00	−0.48	−0.04	3.48
Ipc	1.30×10^{5}	2.86×10^{6}	0.00	34.40	130.65	1589.72	1.34×10^{8}
Kappa1	9.04	4.46	0.00	6.02	7.79	11.21	38.07
Kappa2	4.45	3.20	−27.04	2.51	3.59	5.63	32.00
Kappa3	14.90	255.23	−104.04	1.58	2.81	4.87	9507.96
LabuteASA	70.41	33.75	6.10	46.24	60.65	88.39	213.51
PEOEVSA1	3.80	4.60	0.00	0.00	4.52	5.11	32.92
PEOEVSA10	2.14	4.22	0.00	0.00	0.00	5.02	38.99
PEOEVSA11	1.14	3.79	0.00	0.00	0.00	0.00	34.63

Table 18 (continued)

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
PEOEVSA12	0.68	2.54	0.00	0.00	0.00	0.00	34.90
PEOEVSA13	0.51	1.95	0.00	0.00	0.00	0.00	20.35
PEOEVSA14	2.31	5.72	0.00	0.00	0.00	0.00	82.86
PEOEVSA2	1.96	3.83	0.00	0.00	0.00	4.79	30.34
PEOEVSA3	1.16	3.17	0.00	0.00	0.00	0.00	40.46
PEOEVSA4	1.55	5.64	0.00	0.00	0.00	0.00	118.54
PEOEVSA5	1.63	5.62	0.00	0.00	0.00	0.00	92.81
PEOEVSA6	26.03	23.86	0.00	6.58	23.20	38.97	212.89
PEOEVSA7	16.45	13.93	0.00	6.42	12.84	24.27	159.30
PEOEVSA8	5.97	7.92	0.00	0.00	4.47	11.09	55.85
PEOEVSA9	4.41	7.13	0.00	0.00	0.00	6.61	52.86
SMRVSA1	6.29	8.15	0.00	0.00	4.79	9.59	118.54
SMRVSA10	12.78	17.39	0.00	0.00	5.97	18.53	159.30
SMRVSA2	0.13	0.96	0.00	0.00	0.00	0.00	15.79
SMRVSA3	1.07	3.11	0.00	0.00	0.00	0.00	24.69
SMRVSA4	1.86	4.15	0.00	0.00	0.00	0.00	29.59
SMRVSA5	21.99	23.29	0.00	4.84	13.85	33.30	212.89
SMRVSA6	5.02	7.94	0.00	0.00	0.00	6.67	56.38
SMRVSA7	18.75	21.35	0.00	0.00	12.15	34.89	109.19
SMRVSA8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SMRVSA9	1.84	4.41	0.00	0.00	0.00	0.00	29.72
SlogPVSA1	1.42	3.23	0.00	0.00	0.00	0.00	34.66
SlogPVSA10	1.86	6.13	0.00	0.00	0.00	0.00	118.54
SlogPVSA11	1.19	3.52	0.00	0.00	0.00	0.00	23.52
SlogPVSA12	8.32	15.87	0.00	0.00	0.00	11.60	159.30
SlogPVSA2	10.52	11.59	0.00	0.00	6.54	15.48	77.10
SlogPVSA3	4.13	6.55	0.00	0.00	0.00	6.42	46.02
SlogPVSA4	3.63	6.36	0.00	0.00	0.00	5.92	47.44
SlogPVSA5	20.80	22.86	0.00	0.00	13.85	32.34	212.89
SlogPVSA6	14.69	17.78	0.00	0.00	11.65	24.27	109.19
SlogPVSA7	1.85	7.28	0.00	0.00	0.00	0.00	117.85
SlogPVSA8	1.33	4.72	0.00	0.00	0.00	0.00	53.86
SlogPVSA9	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TPSA	21.70	24.27	0.00	0.00	17.07	35.53	209.48
EStateVSA1	3.49	7.75	0.00	0.00	0.00	5.60	77.10
EStateVSA10	3.46	6.76	0.00	0.00	0.00	4.79	118.54
EStateVSA11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EStateVSA2	3.92	6.92	0.00	0.00	0.00	5.97	61.35
EStateVSA3	5.41	8.03	0.00	0.00	0.00	6.61	50.73
EStateVSA4	6.38	7.99	0.00	0.00	5.56	11.38	64.21
EStateVSA5	11.78	18.40	0.00	0.00	5.69	16.69	199.05
EStateVSA6	5.22	9.53	0.00	0.00	0.00	6.92	72.80
EStateVSA7	7.84	12.79	0.00	0.00	0.00	12.18	91.00
EStateVSA8	13.55	16.42	0.00	0.00	6.92	20.77	159.30
EStateVSA9	8.68	15.10	0.00	0.00	0.00	10.25	139.21
VSAEState1	5.93	15.99	−3.84	0.00	0.00	5.09	349.32
VSAEState10	3.25	7.27	−4.38	0.00	0.00	3.48	79.99
VSAEState2	4.58	7.30	−14.32	0.00	0.00	9.96	56.15
VSAEState3	2.83	4.99	−9.32	0.00	0.00	4.18	40.21

Table 18 (continued)

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
VSAEState4	0.73	1.75	−7.91	0.00	0.00	1.09	15.41
VSAEState5	0.08	2.66	−55.43	0.00	0.00	0.49	10.46
VSAEState6	3.59	5.61	−29.43	0.00	0.00	6.68	38.97
VSAEState7	2.27	4.76	−25.31	0.00	0.47	3.32	45.89
VSAEState8	3.27	3.73	−5.88	0.00	2.25	5.14	38.30
VSAEState9	0.13	1.38	−20.28	0.00	0.00	0.00	11.43
FractionCSP3	0.57	0.38	0.00	0.20	0.61	1.00	1.00
HeavyAtomCount	10.98	5.46	1.00	7.00	10.00	14.00	40.00
NHOHCount	0.35	0.72	0.00	0.00	0.00	0.00	7.00
NOCount	1.57	1.83	0.00	0.00	1.00	2.00	16.00
NumAliphaticCarbocycles	0.12	0.44	0.00	0.00	0.00	0.00	6.00
NumAliphaticHeterocycles	0.07	0.28	0.00	0.00	0.00	0.00	4.00
NumAliphaticRings	0.19	0.53	0.00	0.00	0.00	0.00	6.00
NumAromaticCarbocycles	0.52	0.78	0.00	0.00	0.00	1.00	6.00
NumAromaticHeterocycles	0.09	0.30	0.00	0.00	0.00	0.00	2.00
NumAromaticRings	0.61	0.85	0.00	0.00	0.00	1.00	6.00
NumHAcceptors	1.52	1.74	0.00	0.00	1.00	2.00	12.00
NumHDonors	0.31	0.59	0.00	0.00	0.00	1.00	5.00
NumHeteroatoms	2.60	2.72	0.00	1.00	2.00	4.00	28.00
NumRotatableBonds	2.67	3.24	0.00	0.00	2.00	4.00	30.00
NumSaturatedCarbocycles	0.09	0.37	0.00	0.00	0.00	0.00	6.00
NumSaturatedHeterocycles	0.04	0.24	0.00	0.00	0.00	0.00	4.00
NumSaturatedRings	0.13	0.44	0.00	0.00	0.00	0.00	6.00
RingCount	0.80	0.97	0.00	0.00	1.00	1.00	6.00
MolLogP	2.59	1.75	−4.66	1.51	2.44	3.42	13.12
MolMR	46.22	22.14	2.50	30.70	40.64	59.25	159.18
frAlCOO	0.03	0.19	0.00	0.00	0.00	0.00	2.00
frAlOH	0.10	0.34	0.00	0.00	0.00	0.00	4.00
frAlOHnoTert	0.08	0.32	0.00	0.00	0.00	0.00	4.00
frArN	0.02	0.17	0.00	0.00	0.00	0.00	3.00
frArCOO	0.01	0.09	0.00	0.00	0.00	0.00	2.00
frArN	0.13	0.53	0.00	0.00	0.00	0.00	5.00
frArNH	0.01	0.08	0.00	0.00	0.00	0.00	2.00
frArOH	0.05	0.21	0.00	0.00	0.00	0.00	2.00
frCOO	0.04	0.21	0.00	0.00	0.00	0.00	2.00
frCOO2	0.04	0.21	0.00	0.00	0.00	0.00	2.00
frCO	0.31	0.58	0.00	0.00	0.00	1.00	3.00
frCOnoCOO	0.28	0.56	0.00	0.00	0.00	0.00	3.00
frCS	0.01	0.09	0.00	0.00	0.00	0.00	2.00
frHOCCN	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frImine	0.01	0.11	0.00	0.00	0.00	0.00	3.00
frNH0	0.29	0.73	0.00	0.00	0.00	0.00	6.00
frNH1	0.08	0.32	0.00	0.00	0.00	0.00	4.00
frNH2	0.04	0.24	0.00	0.00	0.00	0.00	3.00
frNO	0.00	0.09	0.00	0.00	0.00	0.00	2.00
frNdealkylation1	0.01	0.08	0.00	0.00	0.00	0.00	1.00
frNdealkylation2	0.00	0.06	0.00	0.00	0.00	0.00	1.00
frNhpyrrole	0.01	0.08	0.00	0.00	0.00	0.00	2.00
frSH	0.01	0.11	0.00	0.00	0.00	0.00	2.00

Table 18 (continued)

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
fraldehyde	0.02	0.16	0.00	0.00	0.00	0.00	2.00
fralkylcarbamate	0.00	0.04	0.00	0.00	0.00	0.00	2.00
fralkylhalide	0.38	1.38	0.00	0.00	0.00	0.00	27.00
frallylicoxid	0.22	0.71	0.00	0.00	0.00	0.00	8.00
framide	0.07	0.34	0.00	0.00	0.00	0.00	5.00
framidine	0.00	0.04	0.00	0.00	0.00	0.00	2.00
franiline	0.07	0.32	0.00	0.00	0.00	0.00	4.00
frarylmethyl	0.17	0.50	0.00	0.00	0.00	0.00	4.00
frazide	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frazo	0.00	0.06	0.00	0.00	0.00	0.00	1.00
frbarbitur	0.00	0.02	0.00	0.00	0.00	0.00	1.00
frbenzene	0.52	0.78	0.00	0.00	0.00	1.00	6.00
frbenzodiazepine	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frbicyclic	0.18	0.70	0.00	0.00	0.00	0.00	12.00
frdiazo	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frdihydropyridine	0.00	0.00	0.00	0.00	0.00	0.00	0.00
froxazole	0.00	0.03	0.00	0.00	0.00	0.00	1.00
froxime	0.01	0.09	0.00	0.00	0.00	0.00	2.00
frparahydroxylation	0.08	0.34	0.00	0.00	0.00	0.00	4.00
frphenol	0.04	0.21	0.00	0.00	0.00	0.00	2.00
frphenolnoOrthoHbond	0.04	0.21	0.00	0.00	0.00	0.00	2.00
frphosacid	0.01	0.11	0.00	0.00	0.00	0.00	2.00
frphosester	0.01	0.11	0.00	0.00	0.00	0.00	2.00
frpiperdine	0.00	0.08	0.00	0.00	0.00	0.00	3.00
frpiperzine	0.00	0.06	0.00	0.00	0.00	0.00	3.00
frpriamide	0.00	0.08	0.00	0.00	0.00	0.00	2.00
frprisulfonamd	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frpyridine	0.02	0.14	0.00	0.00	0.00	0.00	2.00
frquatN	0.00	0.02	0.00	0.00	0.00	0.00	1.00
frsulfide	0.03	0.17	0.00	0.00	0.00	0.00	2.00
frsulfonamd	0.00	0.05	0.00	0.00	0.00	0.00	1.00
frsulfone	0.00	0.06	0.00	0.00	0.00	0.00	1.00
frtermacetylene	0.01	0.11	0.00	0.00	0.00	0.00	2.00
frtetrazole	0.00	0.02	0.00	0.00	0.00	0.00	1.00
frthiazole	0.00	0.06	0.00	0.00	0.00	0.00	1.00
frthiocyan	0.00	0.02	0.00	0.00	0.00	0.00	1.00
frthiophene	0.01	0.09	0.00	0.00	0.00	0.00	1.00
frunbrchalkane	0.67	2.28	0.00	0.00	0.00	0.00	28.00
frurea	0.01	0.10	0.00	0.00	0.00	0.00	2.00

STD indicates the standard deviation, measuring the spread of a variable around its mean. MIN and MAX indicate the minimum and maximum values observed for each variable, respectively. Quartiles divide the data into four equal parts when sorted in ascending order: Q1 represents the 25th percentile, Q2 the median (50th percentile), and Q3 the 75th percentile

F.4: Self-curated solubility

See Table 19.

Table 19 Summary statistics of Solubility and 209 molecular descriptors

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
Solubility (target)	−2.87	2.06	−17.47	−4.14	−2.75	−1.50	1.70
MaxEStateIndex	9.51	3.26	0.00	6.26	10.61	11.92	17.81
MinEStateIndex	−0.64	1.27	−6.88	−0.92	−0.35	0.00	4.00
MaxAbsEStateIndex	9.51	3.26	0.00	6.26	10.61	11.92	17.81
MinAbsEStateIndex	0.23	0.34	0.00	0.00	0.10	0.31	4.50
qed	0.58	0.19	0.02	0.46	0.60	0.73	0.94
MolWt	284.14	128.69	16.04	194.22	268.32	350.26	1583.58
HeavyAtomMolWt	265.18	120.79	12.01	182.11	250.19	326.55	1475.73
ExactMolWt	283.64	128.48	16.03	194.08	268.12	349.22	1582.65
NumValenceElectrons	102.21	46.31	2.00	70.00	98.00	126.00	618.00
NumRadicalElectrons	0.00	0.06	0.00	0.00	0.00	0.00	3.00
MaxPartialCharge	0.25	0.20	−0.41	0.12	0.24	0.33	3.00
MinPartialCharge	−0.46	0.21	−1.00	−0.49	−0.44	−0.34	1.00
MaxAbsPartialCharge	0.48	0.24	0.00	0.35	0.45	0.50	3.00
MinAbsPartialCharge	0.23	0.14	0.00	0.12	0.24	0.32	1.00
FpDensityMorgan1	1.15	0.29	0.19	1.00	1.17	1.33	2.00
FpDensityMorgan2	1.75	0.38	0.29	1.54	1.81	2.00	2.75
FpDensityMorgan3	2.26	0.49	0.33	2.00	2.35	2.59	3.45
BCUT2DMWHI	30.84	23.15	12.01	16.41	19.41	35.45	126.93
BCUT2DMWLOW	10.09	0.31	0.01	9.95	10.10	10.21	14.01
BCUT2DCHGHI	2.19	0.22	−0.37	2.06	2.17	2.31	3.30
BCUT2DCHGLO	−2.18	0.20	−2.80	−2.32	−2.18	−2.05	−0.08
BCUT2DLOGPHI	2.23	0.20	−0.48	2.13	2.24	2.35	3.09
BCUT2DLOGPLOW	−2.32	0.35	−3.21	−2.47	−2.30	−2.11	0.14
BCUT2DMRHI	6.33	1.38	2.13	5.72	5.93	6.33	14.28
BCUT2DMRLOW	0.06	0.53	−2.00	−0.14	0.03	0.28	3.31
BalabanJ	1.77	1.28	−0.00	0.00	2.11	2.75	6.05
BertzCT	473.44	312.92	0.00	259.62	427.36	644.54	4074.50
Chi0	13.44	5.97	0.00	9.42	12.82	16.23	81.48
Chi0n	11.00	5.03	0.00	7.43	10.42	13.72	62.08
Chi0v	11.87	5.29	0.00	8.03	11.25	14.85	62.97
Chi1	8.76	4.01	0.00	5.95	8.33	10.78	50.88
Chi1n	6.26	3.08	0.00	4.05	5.84	7.96	35.30
Chi1v	6.71	3.18	0.00	4.37	6.33	8.52	38.53
Chi2n	4.74	2.62	0.00	2.87	4.31	6.13	28.79
Chi2v	5.25	2.86	0.00	3.25	4.85	6.76	82.66
Chi3n	3.28	2.17	0.00	1.75	2.85	4.34	21.07
Chi3v	3.65	2.25	0.00	2.01	3.26	4.84	21.60
Chi4n	2.27	1.78	0.00	1.05	1.83	3.02	16.52
Chi4v	2.55	1.86	0.00	1.22	2.17	3.43	19.15
HallKierAlpha	−1.25	1.07	−9.20	−1.85	−1.32	−0.58	12.66
Ipc	4.89×10^{25}	4.58×10^{27}	0.00	723.09	9095.77	1.38×10^{5}	4.29×10^{29}
Kappa1	14.87	7.19	0.00	9.89	13.90	18.47	91.71
Kappa2	6.77	3.93	0.00	4.09	6.02	8.60	62.81
Kappa3	4.93	20.85	−27.04	2.18	3.40	5.34	1128.96
LabuteASA	116.63	51.36	8.74	79.76	111.36	144.36	627.65
PEOEVSA1	12.78	12.61	0.00	5.11	9.90	15.74	165.89
PEOEVSA10	5.80	8.21	−0.06	0.00	5.41	6.61	121.83
PEOEVSA11	2.54	5.28	0.00	0.00	0.00	5.11	72.77

Table 19 (continued)

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
PEOEVSA12	1.78	3.73	0.00	0.00	0.00	0.00	53.39
PEOEVSA13	1.12	2.84	0.00	0.00	0.00	0.00	35.44
PEOEVSA14	3.95	9.62	0.00	0.00	0.00	5.97	354.69
PEOEVSA2	4.61	5.59	0.00	0.00	4.79	5.84	57.84
PEOEVSA3	2.73	4.35	0.00	0.00	0.00	4.79	33.67
PEOEVSA4	1.21	3.63	0.00	0.00	0.00	0.00	52.68
PEOEVSA5	3.46	7.45	0.00	0.00	0.00	0.00	92.81
PEOEVSA6	24.00	21.26	0.00	6.42	19.41	36.03	251.92
PEOEVSA7	28.90	21.60	0.00	12.84	24.35	38.92	161.93
PEOEVSA8	13.58	12.55	0.00	5.56	11.64	19.26	114.14
PEOEVSA9	9.88	11.66	0.00	0.00	6.42	13.21	117.88
SMRVSA1	14.76	16.13	0.00	4.79	9.90	19.37	467.94
SMRVSA10	16.94	14.47	0.00	5.97	12.41	23.56	170.67
SMRVSA2	0.14	1.00	0.00	0.00	0.00	0.00	15.79
SMRVSA3	4.47	5.13	0.00	0.00	4.90	5.32	61.80
SMRVSA4	3.48	7.11	0.00	0.00	0.00	5.73	94.71
SMRVSA5	25.61	25.66	−0.06	6.92	19.77	37.89	296.55
SMRVSA6	15.46	15.18	0.00	0.00	13.09	24.95	137.90
SMRVSA7	32.50	25.03	0.00	12.13	29.83	48.03	180.54
SMRVSA8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SMRVSA9	2.98	5.40	0.00	0.00	0.00	5.75	68.51
SlogPVSA1	8.41	11.88	0.00	0.00	5.32	10.63	413.41
SlogPVSA10	2.64	4.76	0.00	0.00	0.00	5.69	73.64
SlogPVSA11	2.21	4.47	0.00	0.00	0.00	5.75	46.00
SlogPVSA12	7.79	12.75	0.00	0.00	0.00	12.22	135.55
SlogPVSA2	30.18	23.76	−0.06	13.09	25.54	40.36	370.27
SlogPVSA3	7.04	7.89	0.00	0.00	4.79	10.21	94.49
SlogPVSA4	4.25	7.58	0.00	0.00	0.00	6.92	71.02
SlogPVSA5	24.23	21.70	0.00	6.92	19.41	34.12	290.45
SlogPVSA6	26.40	21.32	0.00	11.11	24.27	41.14	170.18
SlogPVSA7	0.80	2.79	0.00	0.00	0.00	0.00	35.16
SlogPVSA8	2.41	5.71	0.00	0.00	0.00	0.00	65.34
SlogPVSA9	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TPSA	53.98	47.27	0.00	24.06	46.17	69.92	896.66
EStateVSA1	7.52	15.13	0.00	0.00	0.00	10.02	250.73
EStateVSA10	6.79	7.98	0.00	0.00	4.79	9.59	109.41
EStateVSA11	0.02	0.37	0.00	0.00	0.00	0.00	14.38
EStateVSA2	13.08	13.52	0.00	5.41	11.66	18.45	354.69
EStateVSA3	9.64	10.44	0.00	0.00	6.42	13.21	117.88
EStateVSA4	15.69	15.14	−0.06	4.57	12.04	24.95	104.72
EStateVSA5	14.05	19.02	0.00	0.00	6.92	19.26	192.62
EStateVSA6	8.34	11.66	0.00	0.00	0.00	12.99	124.63
EStateVSA7	13.96	16.80	0.00	0.00	6.92	24.27	139.53
EStateVSA8	18.33	20.56	0.00	4.98	11.05	25.67	221.16
EStateVSA9	8.93	10.62	0.00	0.00	5.11	11.69	136.33
VSAEState1	6.87	11.77	−3.70	0.00	2.95	7.96	152.95
VSAEState10	1.84	4.33	−5.10	0.00	0.00	1.49	65.67
VSAEState2	11.77	12.06	−3.76	1.35	10.75	16.59	164.67
VSAEState3	7.57	11.62	−9.32	0.00	3.20	10.42	153.27

Table 19 (continued)

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
VSAEState4	2.10	3.12	−9.03	0.00	1.14	3.16	38.76
VSAEState5	0.12	1.76	−34.01	−0.22	0.00	0.65	12.58
VSAEState6	6.90	6.81	−9.66	0.00	6.07	10.57	44.64
VSAEState7	2.30	4.25	−36.18	0.00	1.16	3.81	62.70
VSAEState8	3.47	3.72	−6.81	0.29	2.49	5.21	32.12
VSAEState9	0.76	2.42	−40.11	0.00	0.00	1.63	17.51
FractionCSP3	0.47	0.30	0.00	0.25	0.43	0.67	1.00
HeavyAtomCount	18.69	8.49	1.00	13.00	18.00	23.00	109.00
NHOHCount	1.47	1.90	0.00	0.00	1.00	2.00	38.00
NOCount	3.90	2.87	0.00	2.00	3.00	5.00	50.00
NumAliphaticCarbocycles	0.28	0.83	0.00	0.00	0.00	0.00	7.00
NumAliphaticHeterocycles	0.47	0.78	0.00	0.00	0.00	1.00	8.00
NumAliphaticRings	0.75	1.16	0.00	0.00	0.00	1.00	10.00
NumAromaticCarbocycles	0.90	0.85	0.00	0.00	1.00	1.00	7.00
NumAromaticHeterocycles	0.25	0.52	0.00	0.00	0.00	0.00	6.00
NumAromaticRings	1.15	0.95	0.00	0.00	1.00	2.00	7.00
NumHAcceptors	3.24	2.46	0.00	2.00	3.00	4.00	36.00
NumHDonors	1.23	1.52	0.00	0.00	1.00	2.00	30.00
NumHeteroatoms	4.85	3.05	0.00	3.00	4.00	6.00	53.00
NumRotatableBonds	4.20	3.47	0.00	2.00	4.00	6.00	47.00
NumSaturatedCarbocycles	0.19	0.69	0.00	0.00	0.00	0.00	6.00
NumSaturatedHeterocycles	0.34	0.67	0.00	0.00	0.00	1.00	6.00
NumSaturatedRings	0.53	0.98	0.00	0.00	0.00	1.00	10.00
RingCount	1.90	1.41	0.00	1.00	2.00	3.00	16.00
MolLogP	1.84	2.43	−46.67	0.78	2.01	3.33	16.43
MolMR	73.98	32.35	0.00	50.30	70.79	92.32	370.22
frAlCOO	0.11	0.44	0.00	0.00	0.00	0.00	6.00
frAlOH	0.31	0.91	0.00	0.00	0.00	0.00	14.00
frAlOHnoTert	0.26	0.83	0.00	0.00	0.00	0.00	13.00
frArN	0.06	0.28	0.00	0.00	0.00	0.00	3.00
frArCOO	0.02	0.17	0.00	0.00	0.00	0.00	3.00
frArN	0.33	0.81	0.00	0.00	0.00	0.00	8.00
frArNH	0.05	0.23	0.00	0.00	0.00	0.00	3.00
frArOH	0.11	0.40	0.00	0.00	0.00	0.00	6.00
frCOO	0.14	0.46	0.00	0.00	0.00	0.00	6.00
frCOO2	0.14	0.47	0.00	0.00	0.00	0.00	6.00
frCO	0.71	0.94	0.00	0.00	0.00	1.00	12.00
frCOnoCOO	0.59	0.85	0.00	0.00	0.00	1.00	12.00
frCS	0.01	0.13	0.00	0.00	0.00	0.00	2.00
frHOCCN	0.01	0.12	0.00	0.00	0.00	0.00	2.00
frImine	0.03	0.21	0.00	0.00	0.00	0.00	4.00
frNH0	0.93	1.10	0.00	0.00	1.00	1.00	9.00
frNH1	0.41	0.66	0.00	0.00	0.00	1.00	8.00
frNH2	0.21	0.52	0.00	0.00	0.00	0.00	8.00
frNO	0.02	0.18	0.00	0.00	0.00	0.00	6.00
frNdealkylation1	0.15	0.40	0.00	0.00	0.00	0.00	4.00
frNdealkylation2	0.10	0.32	0.00	0.00	0.00	0.00	4.00
frNhpyrrole	0.05	0.23	0.00	0.00	0.00	0.00	3.00
frSH	0.01	0.10	0.00	0.00	0.00	0.00	2.00

Table 19 (continued)

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
fraldehyde	0.01	0.09	0.00	0.00	0.00	0.00	2.00
fralkylcarbamate	0.01	0.12	0.00	0.00	0.00	0.00	2.00
fralkylhalide	0.15	0.70	0.00	0.00	0.00	0.00	12.00
frallylicoxid	0.14	0.67	0.00	0.00	0.00	0.00	12.00
framide	0.36	0.78	0.00	0.00	0.00	0.00	10.00
framidine	0.02	0.16	0.00	0.00	0.00	0.00	2.00
franiline	0.25	0.57	0.00	0.00	0.00	0.00	6.00
frarylmethyl	0.19	0.50	0.00	0.00	0.00	0.00	6.00
frazide	0.00	0.01	0.00	0.00	0.00	0.00	1.00
frazo	0.00	0.04	0.00	0.00	0.00	0.00	1.00
frbarbitur	0.01	0.11	0.00	0.00	0.00	0.00	2.00
frbenzene	0.90	0.85	0.00	0.00	1.00	1.00	7.00
frbenzodiazepine	0.00	0.06	0.00	0.00	0.00	0.00	1.00
frbicyclic	0.58	1.21	0.00	0.00	0.00	1.00	15.00
frdiazo	0.00	0.02	0.00	0.00	0.00	0.00	2.00
frdihydropyridine	0.00	0.03	0.00	0.00	0.00	0.00	1.00
froxazole	0.01	0.11	0.00	0.00	0.00	0.00	2.00
froster	0.13	0.41	0.00	0.00	0.00	0.00	6.00
frother	0.58	1.00	0.00	0.00	0.00	1.00	11.00
frfuran	0.01	0.12	0.00	0.00	0.00	0.00	2.00
frguanido	0.01	0.13	0.00	0.00	0.00	0.00	4.00
frhalogen	0.67	1.09	0.00	0.00	0.00	1.00	12.00
frhdrzine	0.01	0.13	0.00	0.00	0.00	0.00	3.00
frhdrzone	0.01	0.11	0.00	0.00	0.00	0.00	2.00
frimidazole	0.02	0.16	0.00	0.00	0.00	0.00	2.00
frimide	0.04	0.25	0.00	0.00	0.00	0.00	4.00
frisocyan	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frisothiocyan	0.00	0.03	0.00	0.00	0.00	0.00	2.00
frketone	0.11	0.38	0.00	0.00	0.00	0.00	4.00
frketoneTopliss	0.08	0.31	0.00	0.00	0.00	0.00	4.00
frlactam	0.01	0.07	0.00	0.00	0.00	0.00	2.00
frlactone	0.01	0.13	0.00	0.00	0.00	0.00	6.00
frmethoxy	0.15	0.52	0.00	0.00	0.00	0.00	6.00
frmorpholine	0.02	0.15	0.00	0.00	0.00	0.00	2.00
frnitrile	0.02	0.15	0.00	0.00	0.00	0.00	3.00
frnitro	0.05	0.24	0.00	0.00	0.00	0.00	3.00
frnitroarom	0.03	0.18	0.00	0.00	0.00	0.00	3.00
frnitroaromnonortho	0.02	0.14	0.00	0.00	0.00	0.00	3.00
frnitroso	0.01	0.09	0.00	0.00	0.00	0.00	2.00
froxazole	0.01	0.11	0.00	0.00	0.00	0.00	1.00
froxime	0.00	0.06	0.00	0.00	0.00	0.00	2.00
frparahydroxylation	0.18	0.48	0.00	0.00	0.00	0.00	4.00
frphenol	0.10	0.39	0.00	0.00	0.00	0.00	6.00
frphenolnoOrthoHbond	0.09	0.38	0.00	0.00	0.00	0.00	6.00
frphosacid	0.01	0.13	0.00	0.00	0.00	0.00	6.00
frphosester	0.01	0.10	0.00	0.00	0.00	0.00	6.00
frpiperdine	0.11	0.40	0.00	0.00	0.00	0.00	6.00
frpiperzine	0.05	0.22	0.00	0.00	0.00	0.00	4.00
frpriamide	0.05	0.26	0.00	0.00	0.00	0.00	6.00

Table 19 (continued)

Descriptor	Mean	STD	Min	Q1	Q2	Q3	Max
frprisulfonamd	0.00	0.00	0.00	0.00	0.00	0.00	0.00
frpyridine	0.08	0.29	0.00	0.00	0.00	0.00	6.00
frquatN	0.12	0.39	0.00	0.00	0.00	0.00	2.00
frsulfide	0.05	0.26	0.00	0.00	0.00	0.00	2.00
frsulfonamd	0.02	0.17	0.00	0.00	0.00	0.00	2.00
frsulfone	0.00	0.07	0.00	0.00	0.00	0.00	2.00
frtermacetylene	0.01	0.10	0.00	0.00	0.00	0.00	2.00
frtetrazole	0.00	0.04	0.00	0.00	0.00	0.00	1.00
frthiazole	0.01	0.10	0.00	0.00	0.00	0.00	2.00
frthiocyan	0.00	0.04	0.00	0.00	0.00	0.00	1.00
frthiophene	0.01	0.10	0.00	0.00	0.00	0.00	2.00
frunbrchalkane	0.48	1.74	0.00	0.00	0.00	0.00	33.00
frurea	0.03	0.19	0.00	0.00	0.00	0.00	4.00

STD indicates the standard deviation, measuring the spread of a variable around its mean. MIN and MAX indicate the minimum and maximum values observed for each variable, respectively. Quartiles divide the data into four equal parts when sorted in ascending order: Q1 represents the 25th percentile, Q2 the median (50th percentile), and Q3 the 75th percentile

Abbreviations

GCN	Graph convolutional networks
GNN	Graph neural networks
GAT	Graph attention networks
GIN	Graph isomorphism networks
AI	Aritifical intelligence
ISIS	Iterative sure independence screening
EN	Elastic net
MSE	Mean squared error
MAE	Mean absolute error
FreeSolv	Free solvation
ESOL	Estimated solubility
PCA	Principal component analysis
R^2	Coefficient of determination
SHAP	Shapley additive explanations

Acknowledgements

Not applicable.

Author contributions

Y.J conducted data analysis, experimentation, and drafted the manuscript. J.L collected, and provided the gas dataset. K.J provided feedback with chemistry knowledge. J.K designed, supervised this study from a statistical and machine learning perspective, and provided critical feedback on the manuscript. All authors contributed to discussions, read and approved the final manuscript.

Funding

This study was supported by Tech. Challenge for Future Program Policing Program through the Korea Institutes of Police Technology(KIPoT) funded by the Korean National Police Agency* (RS-2024-0048545). This study was also supported by the INHA UNIVERSITY Research Grant.

Availability of data and materials

All source codes and benchmark datasets that support the findings of this study are available on GitHub at https://github.com/gdsjayjang/KROVEX. The self-curated data are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Statistics and Data Science, Inha University, 100, Inha-ro, Michuhol-gu, Incheon 22212, Republic of Korea. [2]Department of Biomedical Engineering, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 0284, Republic of Korea. [3]Department of Physics and Chemistry, Korea Military Academy, 574, Hwarang-ro, Nowon-gu, Seoul 01805, Republic of Korea.

Received: 4 October 2025 Accepted: 11 December 2025
Published online: 11 January 2026

References

1. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv:1609.02907
2. Xie T, Grossman JC (2018) Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett 120(14):145301
3. Karamad M, Magar R, Shi Y, Siahrostami S, Gates ID, Barati Farimani A (2020) Orbital graph convolutional neural network for material property prediction. Phys Rev Mater 4(9):093801
4. Vijayan R, Mohler G (2018) Forecasting retweet count during elections using graph convolution neural networks. In: 2018 IEEE 5th international conference on data science and advanced analytics (DSAA). IEEE, pp 256–262
5. Gao L, Wang H, Zhang Z, Zhuang H, Zhou B (2022) HetInf: social influence prediction with heterogeneous graph neural network. Front Phys 9:787185
6. Liu H, Wei J, Xu T (2023) Community detection based on community perspective and graph convolutional network. Expert Syst Appl 231:120748
7. Jeong C, Jang S, Park E, Choi S (2020) A context-aware citation recommendation model with Bert and graph convolutional networks. Scientometrics 124:1907–1922
8. Lin G, Wang J, Liao K, Zhao F, Chen W (2020) Structure fusion based on graph convolutional networks for node classification in citation networks. Electronics 9(3):432

9. Cummings D, Nassar M (2020) Structured citation trend prediction using graph neural networks. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 3897–3901
10. Li J, Cai D, He X (2017) Learning graph-level representation for drug discovery. arXiv:1709.03741
11. Li X, Yan X, Gu Q, Zhou H, Wu D, Xu J (2019) Deepchemstable: chemical stability prediction with an attention-based graph convolution network. J Chem Inf Model 59(3):1044–1049
12. Shang C, Liu Q, Chen K-S, Sun J, Lu J, Yi J, Bi J (2018) Edge attention-based multi-relational graph convolutional networks. arXiv:1802.04944
13. Liu K, Sun X, Jia L, Ma J, Xing H, Wu J, Gao H, Sun Y, Boulnois F, Fan J (2019) Chemi-Net: a molecular graph convolutional network for accurate drug property prediction. Int J Mol Sci 20(14):3389
14. Fout A, Byrd J, Shariat B, Ben-Hur A (2017) Protein interface prediction using graph convolutional networks. In: Advances in neural information processing systems, vol 30
15. Li Q, Han Z, Wu X-M (2018) Deeper insights into graph convolutional networks for semi-supervised learning. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
16. Na GS, Kim HW, Chang H (2020) Costless performance improvement in machine learning for graph-based molecular analysis. J Chem Inf Model 60(3):1137–1145
17. Ji Z, Shi R, Lu J, Li F, Yang Y (2022) ReLMole: molecular representation learning based on two-level graph similarities. J Chem Inf Model 62(22):5361–5372
18. Zhou J, Li S, Huang L, Xiong H, Wang F, Xu T, Xiong H, Dou D (2020) Distance-aware molecule graph attention network for drug-target binding affinity prediction. arXiv:2012.09624
19. Ying Z, You J, Morris C, Ren X, Hamilton W, Leskovec J (2018) Hierarchical graph representation learning with differentiable pooling. In: Advances in neural information processing systems, vol 31
20. Pei H, Wei B, Chang KC-C, Lei Y, Yang B (2020) Geom-GCN: geometric graph convolutional networks. arXiv:2002.05287
21. Zhang Z, Guan J, Zhou S (2021) FraGAT: a fragment-oriented multi-scale graph attention model for molecular property prediction. Bioinformatics 37(18):2981–2987
22. Li A, Casiraghi E, Rousu J (2025) CSGL: chemical synthesis graph learning for molecule representation. Bioinformatics 41(7):355
23. Rong Y, Bian Y, Xu T, Xie W, Wei Y, Huang W, Huang J (2020) Self-supervised graph transformer on large-scale molecular data. Adv Neural Inf Process Syst 33:12559–12571
24. Wang Z, Liu Y, Hu W et al (2024) Molecular representation learning via hierarchical graph transformer. APSIPA Trans Signal Inf Process 14(2):e107
25. Xu P, Zhu X, Clifton DA (2023) Multimodal learning with transformers: a survey. IEEE Trans Pattern Anal Mach Intell 45(10):12113–12132
26. Huang F, Canny JF, Nichols J (2019) Swire: sketch-based user interface retrieval. In: Proceedings of the 2019 CHI conference on human factors in computing systems
27. Lv F, Chen X, Huang Y, Duan L, Lin G (2021) Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2554–2562
28. Chen C, Al-Halah Z, Grauman K (2021) Semantic audio-visual navigation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15516–15525
29. Wang X, Li Z, Jiang M, Wang S, Zhang S, Wei Z (2019) Molecule property prediction based on spatial graph embedding. J Chem Inf Model 59(9):3817–3828
30. Wu J, Su Y, Yang A, Ren J, Xiang Y (2023) An improved multi-modal representation-learning model based on fusion networks for property prediction in drug discovery. Comput Biol Med 165:107452
31. He X, Du X, Wang X, Tian F, Tang J, Chua T-S (2018) Outer product-based neural collaborative filtering. arXiv:1808.03912
32. Liu Z, Zhou B, Chu D, Sun Y, Meng L (2024) Modality translation-based multimodal sentiment analysis under uncertain missing modalities. Inf Fusion 101:101973
33. Luvembe AM, Li W, Li S, Xu G, Wu X, Liu F (2025) An adaptive auto fusion with hierarchical attention for multimodal fake news detection. Expert Syst Appl 285:127930
34. Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv:1606.01847
35. Ben-Younes H, Cadene R, Cord M, Thome N (2017) Mutan: multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2612–2620
36. Gao Y, Beijbom O, Zhang N, Darrell T (2016) Compact bilinear pooling. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 317–326
37. Chen RJ, Lu MY, Wang J, Williamson DF, Rodig SJ, Lindeman NI, Mahmood F (2020) Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. IEEE Trans Med Imaging 41(4):757–770
38. Li C, Hou Y, Li W, Ding Z, Wang P (2024) DFN: a deep fusion network for flexible single and multi-modal action recognition. Expert Syst Appl 245:123145
39. Landrum G et al (2013) RDKit: a software suite for cheminformatics, computational chemistry, and predictive modeling. Greg Landrum 8(31.10):5281
40. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3(Mar):1157–1182
41. Lima E, Davies P, Kaler J, Lovatt F, Green M (2020) Variable selection for inferential models with relatively high-dimensional data: between method heterogeneity and covariate stability as adjuncts to robust selection. Sci Rep 10(1):8002
42. Chamlal H, Benzmane A, Ouaderhman T (2024) Elastic net-based high dimensional data selection for regression. Expert Syst Appl 244:122958
43. Desboulets LDD (2018) A review on variable selection in regression analysis. Econometrics 6(4):45
44. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 58(1):267–288
45. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol 67(2):301–320
46. Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101(476):1418–1429
47. Liu X, Molstad AJ, Chi EC (2023) A convex-nonconvex strategy for grouped variable selection. Electron J Stat 17(2):2912–2961
48. Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol 70(5):849–911
49. Fan J, Samworth R, Wu Y (2009) Ultrahigh dimensional feature selection: beyond the linear model. J Mach Learn Res 10:2013–2038
50. Saldana DF, Feng Y (2018) SIS: an R package for sure independence screening in ultrahigh-dimensional statistical models. J Stat Softw 83:1–25
51. Mahoney MW, Drineas P (2009) CUR matrix decompositions for improved data analysis. Proc Natl Acad Sci 106(3):697–702
52. Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS (2020) A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst 32(1):4–24
53. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. arXiv:1710.10903
54. Xu K, Hu W, Leskovec J, Jegelka S (2018) How powerful are graph neural networks? arXiv:1810.00826
55. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V (2018) MoleculeNet: a benchmark for molecular machine learning. Chem Sci 9(2):513–530
56. Mobley DL, Guthrie JP (2014) FreeSolv: a database of experimental and calculated hydration free energies, with input files. J Comput Aided Mol Des 28(7):711–720
57. Delaney JS (2004) ESOL: estimating aqueous solubility directly from molecular structure. J Chem Inf Comput Sci 44(3):1000–1005
58. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M et al (2019) Analyzing learned molecular representations for property prediction. J Chem Inf Model 59(8):3370–3388
59. Kim J-H, Jun J, Zhang B-T (2018) Bilinear attention networks. In: Advances in neural information processing systems, vol 31
60. Li L, Zhang Y, Wang G, Xia K (2024) KA-GNN: Kolmogorov–Arnold graph neural networks for molecular property prediction. arXiv:2410.11323
61. Zhao B, Xu W, Guan J, Zhou S (2024) Molecular property prediction based on graph structure learning. Bioinformatics 40(5):304

62. Rasool A, Ul Rahman J, Uwitije R (2025) Enhancing molecular property prediction with quantized GNN models. J Cheminform 17(1):81
63. Hoffmann M, Hasse H, Jirasek F (2025) GRAPPA—a hybrid graph neural network for predicting pure component vapor pressures. Chem Eng J Adv 22:100750
64. Krüger M, Galeazzo T, Eremets I, Schmidt B, Pöschl U, Shiraiwa M, Berkemeier T (2025) Improved vapor pressure predictions using group contribution-assisted graph convolutional neural networks (GC2NN). EGUsphere 2025:1–22
65. Lin Y-H, Liang H-H, Lin S-T, Li Y-P (2024) Advancing vapor pressure prediction: a machine learning approach with directed message passing neural networks. J Taiwan Inst Chem Eng 105926
66. Hyun Nam J, Lee S, Jo S, Kim J, Lee J, Koo J, Lee B, Jeong K, Yu D (2025) Improving vapor pressure prediction through integration of multiple molecular representations: a super learner approach. J Chemom 39(2):70003

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.