Systems & Control Transactions

Original Research Article
Peer Reviewed Conference Proceeding
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

An End-to-End Pure Component Property Prediction Framework Based on a Hierarchical Molecular Fragmentation Method

Jianfeng Jiao[a] and Jie Li[a]*

[a] The University of Manchester, Department of Chemical Engineering, Manchester, UK
* Corresponding Author: jie.li-2@manchester.ac.uk

ABSTRACT

The accurate prediction of pure component properties has consistently been a critical issue in fields such as chemical engineering, biomedicine, and environmental science. In recent years, end-to-end deep learning methods have shown significant improvement over traditional machine learning approaches. This is due to their ability to automatically learn task-relevant representations from raw molecular data. In addition to accurate property prediction, researchers have increasingly focused on how specific fragment structures influence molecular properties. However, existing fragmentation methods based on predefined rules and group libraries struggle to capture novel molecular structures, which hampers the development of new materials and drugs. To address these challenges, this work proposes a hierarchical molecular fragmentation method. This method can automatically segment molecules into multiple fragments containing key functional groups. Then a three-branch graph attention network was constructed to achieve multi-level representation. Finally, a multi-layer perceptron is employed to establish the mapping relationship between molecular features and physical property values. Twenty datasets were used for validation, which can be grouped into four categories: Thermodynamic Properties, Pharmacokinetics, Toxicological Properties, and Industrial Safety. The results show that the best performance is achieved, with the average error reduced by 6.8% compared to existing research.

Keywords: Machine Learning, Artificial Intelligence, Algorithms, Multiscale Modelling, Property Prediction

INTRODUCTION

Accurate prediction of molecular physicochemical properties relies on effective molecular representations. Traditional descriptors like MACCS keys and Extended-Connectivity Fingerprints (ECFP) have been widely used to encode structural information into vectors [1]. However, these fixed-length representations suffer from limitations such as hash collisions, information loss in many-to-one mappings, and a lack of direct chemical interpretability. In parallel, Group-Contribution (GC) methods estimate properties by summing the contributions of predefined structural groups [2]. While recent integrations with machine learning have improved their nonlinear expression ability [3]. However, GC methods inherently ignore topological connectivity. Furthermore, their reliance on fixed libraries results in high-dimensional, sparse feature vectors that can lead to model overfitting and poor generalization to molecules containing novel substructures [4].

With the rise of deep learning, Graph Neural Networks (GNNs) have achieved success by treating molecules as graphs and updating atomic representations via message passing [5], meanwhile attention mechanisms have further improved accuracy [6]. However, chemists typically reason via specific structural motifs rather than treating the molecule as an indivisible whole. Hence, researchers have employed the predefined BRICS scheme to decompose molecules into fragments, subsequently treating each fragment as a subgraph and representing it using GNNs. Similarly, predefined group libraries have been used to fragment molecules, where each group is regarded as a subgraph. These approaches have been shown to effectively improve predictive accuracy,

because molecular properties are often governed by specific functional groups; for example, nitro, amino, and quinone groups often increase molecular toxicity, whereas carboxyl groups are widely recognised as detoxifying motifs. The resulting fragments can capture common functional groups and establish mappings to target properties via attention mechanisms. However, owing to the limited coverage of predefined rules, prediction accuracy often deteriorates when extrapolating to out-of-distribution (OOD) molecules. To address the aforementioned issues, Li et al. [7] proposed SMILES Pair Encoding (SPE) method, which learns frequent SMILES substrings from a large dataset, then segments molecules into tokens based on learned SMILES substrings. Benefiting from the large scale of the dataset, the learnable chemical space is substantially expanded, thereby alleviating the out-of-distribution (OOD) issue. However, SPE completely loses topological information. The model only knows that the molecule contains these three fragments. It does not know how these fragments are connected.

To address these challenges, based on a novel hierarchical fragmentation method, a molecular property prediction framework is proposed. Then, a three-branch GNNs was developed to enable representations at both the molecular-graph level and the fragment level. Twenty datasets were used for validation, which can be grouped into four categories: Thermodynamic Properties, Pharmacokinetics, Toxicological Properties, and Industrial Safety.

METHODOLOGY

Hierarchical molecular fragmentation method

The quality of molecular fragments depends on whether they contain functional groups that determine molecular properties or the core scaffolds that host these functional groups. Such motifs are typically highly cohesive in topology; they exhibit dense internal bonding and are connected to the rest of the molecule via only a small number of external bonds. In this subsection, we propose a novel hierarchical molecular fragmentation method. It systematically searches for highly cohesive topological substructures within molecules and enables explicit modelling of molecular fragments.

We treat each molecule as an undirected molecular graph $G = (V, E)$, where V are non-hydrogen nodes and $E$are bonds. Fragmentation is performed by a greedy-and-recursive procedure that repeatedly extracts one dominant local substructure and then applies the same procedure to the remaining subgraph.

Step 1: State initialization and candidate generation.

We first define the core C as a set of k directly connected atoms in graph G. We set a maximum core order k (in this work $k = 4$) and define $S_k(G)$ denote the collection of all such connected k -atom cores $S_k(G) = \{\, C \subseteq V \mid |C| = k, G[C] \text{ is connected} \,\}$.

For each core $C \in S_k(G)$, we define its environment as the core together with all atoms directly bonded to any atom in the core: $\Omega(C) = C \cup \{u \in V \mid \exists v \in C, (u, v) \in E\}$. Cores that yield the same environment are grouped, producing a candidate environment set $\mathcal{M}_{ca}$.

Step 2: Weight evaluation and greedy selection.

For each candidate environment $M \in \mathcal{M}_{ca}$, we assign a weight based on how many cores map to it across orders:

$$W(M) = \sum_{k=1}^{K} |\{C \in S_k(G) \mid \Omega(C) = M\}|$$

We then select the environment M^* with the largest weight as the dominant fragment, $M^* = \arg\max_{M \in \mathcal{M}_{ca}} W(M)$.

Step 3: State update and recursive splitting.

After selecting M^*, we remove M^* from the graph G to obtain the residual subgraph. At this stage, we examine whether the residual subgraph contains more than two non-hydrogen nodes. If so, the residual subgraph is treated as an intact unit, and Steps 1–2 are repeated to further fragment the residual subgraph until the stopping criterion is satisfied.

The final fragmentation results for the molecule O=C(N)Nc1ccc(C)cc1 are shown in **Fig. 1**. As can be seen, the proposed method successfully identifies the urea moiety (–NH–C(=O)–NH2) and isolates it as an individual fragment (blue), while the remaining part corresponds to a tolyl aromatic ring (red). For this molecule, the urea group governs its polarity and hydrogen-bonding capability and is therefore one of the most critical structural units for property characterization.

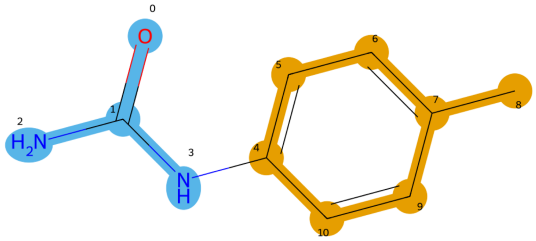

Figure 1. The fragmentation results of O=C(N)Nc1ccc(C)cc1 based on the proposed hierarchical molecular fragmentation method.

In addition, we fragmented the molecule using a predefined group-contribution (GC) library [8] and the BRICS rules [9], as shown in **Figs. 2–3**. Under the GC scheme, the phenyl ring is decomposed into multiple

aromatic carbon atoms, and the methyl group becomes disconnected from the ring, preventing the formation of an intact fragment. Under BRICS, the complete urea structure is not preserved; instead, it is split into a carbamoyl fragment (–NH2C(=O)–) and an –NH fragment.

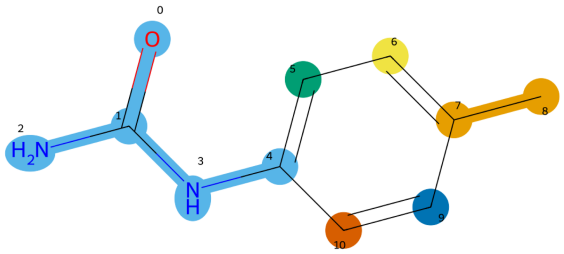

Figure 2. The fragmentation results of O=C(N)Nc1ccc(C)cc1 based on the group contribution method.

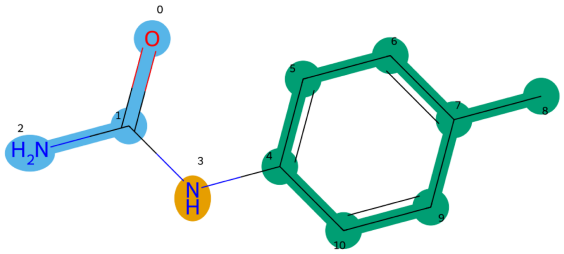

Figure 3. The fragmentation results of O=C(N)Nc1ccc(C)cc1 based on the BRICS method.

HierAttnGNN

The HierAttnGNN contains three branches. The first branch learns representations at the whole molecule level. The second branch encodes molecular fragments. The third branch is a junction tree that represents the connectivity relationships between fragments, as shown in **Fig. 4.**

Molecule level representation

This branch learns a global embedding for the complete molecular graph. Given the molecular graph defined above, let x_v and $x_{e(v,u)}$ denote the node and edge features, as shown in **Table 1-2** respectively.

Table 1 Atom information used to featurize the molecular graph

No.	Feature Name	Description	Encoding	Dim
1	Atom Symbol	Atom type (C, N, O, S, F, P, Cl, Br, I)	One-hot	9
2	Atom Degree	Number of covalent bonds (0–5)	One-hot	6
3	Total Num Hs	Total number of hydrogen atoms (0–4)	One-hot	5
4	Implicit Valence	Implicit valence electrons (0–5)	One-hot	6
5	Hybridization	Orbital hybridization type (SP, SP2, SP3, SP3D, SP3D2)	One-hot	5
6	Is Aromatic	True (1) / False (0)	Boolean (Numeric)	1
7	Chirality Possible	Potential for chirality (RDKit property)	Boolean (Numeric)	1
8	Chirality Type (CIP Code)	CIP code classification (R, S)	One-hot	2
9	Formal Charge	Electrical charge of the atom	Direct Numeric	1
Total				36

Table 2 Bond information used to featurize the molecular graph

No.	Feature Name	Description	Encoding	Dim
1	Is Single Bond	Single, Double, Triple, or Aromatic	Boolean (0/1)	1
2	Is Double Bond	Conjugation status of the bond	Boolean (0/1)	1
3	Is Triple Bond	Checks if the bond type is Triple	Boolean (0/1)	1
4	Is Aromatic Bond	Checks if the bond type is Aromatic	Boolean (0/1)	1
5	Is Conjugated	Checks if the bond is conjugated	Boolean (0/1)	1
6	Is In Ring	Ring membership status	Boolean (0/1)	1
Total				6

We first project the atom and bond feature vectors into a shared hidden space, yielding the initial hidden states of atoms and bonds, $h_v^{(0)}$ and $h_{e(v,u)}^{(0)}$, as in Eqs. (1)–(2).

$$h_v^{(0)} = \mathrm{ReLU}(\mathbf{W}_{node} x_v + b_n) \quad (1)$$

$$h_{e(v,u)}^{(0)} = \mathbf{W}_{edge} x_{e(v,u)} + b_e \quad (2)$$

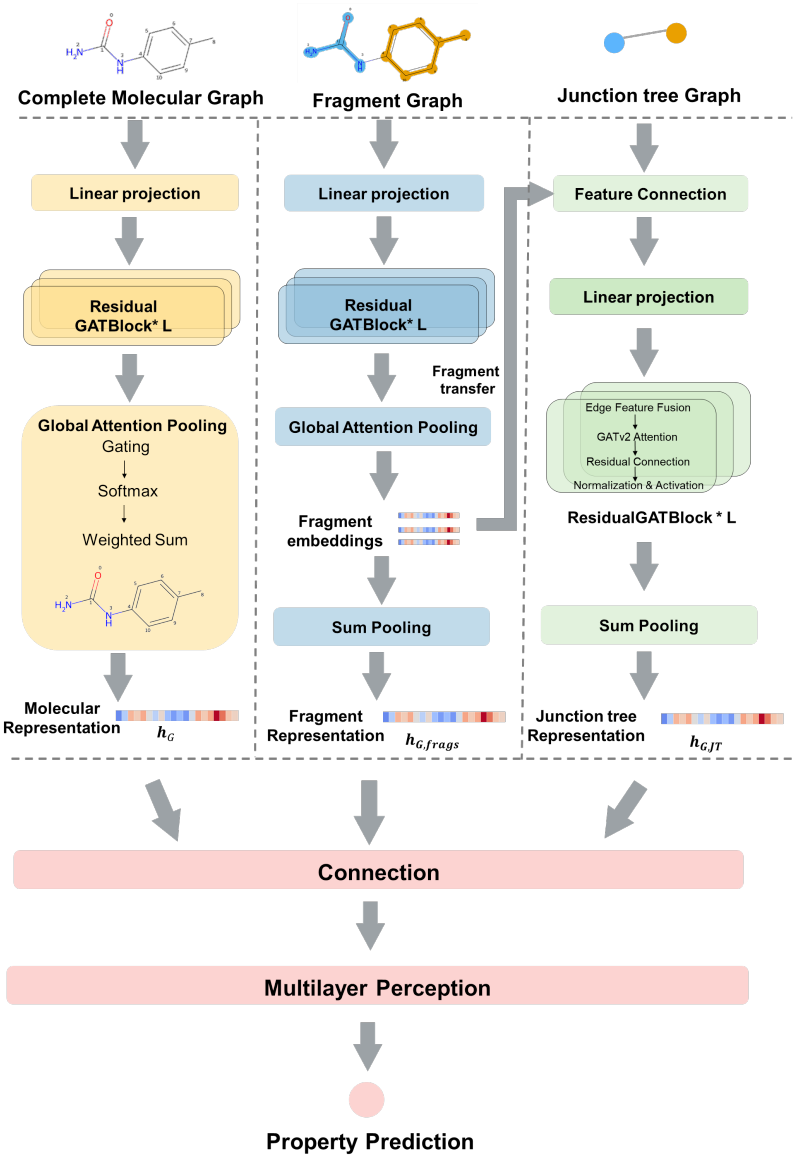

Figure 4. The architecture of the proposed HierAttnGNN framework, consisting of molecule-level, fragment-level, and junction-tree branches

We then stack L ResidualGATBlock modules to update node representations. Each block is built on a GATv2 layer, which performs neighbourhood aggregation with learnable attention. At layer l, edge features incident to node v are first fused into the node state, so that subsequent attention can depend on both node and bond information, as in Eq. (3).

$$h'_v = h_v^{(l)} + \sum_{u \in \mathcal{N}(v)} h_{e(v,u)}^{(l)} \tag{3}$$

Next, for each neighbour u of v, GATv2 computes an unnormalised attention score a_{vu} using a dynamic attention function (Eq. (4)), which is more expressive than the original GAT and can capture more complex structural dependencies [10]. The scores are normalised with a softmax to obtain the importance weights α_{vu} (Eq. (5)).

$$a_{vu} = w_{attn}^T \bullet \mathrm{LeakyReLU}(\mathbf{W}_L h'_v + \mathbf{W}_R h'_u) \tag{4}$$

$$\alpha_{vu} = \mathrm{softmax}_u(a_{vu}) = \frac{\exp(a_{vu})}{\sum_{k \in \mathcal{N}(v) \cup \{v\}} \exp(a_{vk})} \tag{5}$$

where, $\mathbf{W}_L$ and $\mathbf{W}_R$ are learnable weight matrices. Specifically, $W_L, W_R \in \mathbb{R}^{d_h \times d_h}$ are learnable linear transformation matrices. The vector $w_{\mathrm{attn}} \in \mathbb{R}^{d_h}$ is a learnable attention weight vector.

Using α_{vu}, the model forms a weighted sum of neighbour features to update each node. A multi-head

mechanism is applied for stability and capacity, and the outputs of K heads are concatenated, as shown in Eq. (6).

$$h_v^{(l+1)} = \|_{k=1}^{K} \sum_{u \in \mathcal{N}(v) \cup \{v\}} \alpha_{vu}^{(k)} \mathbf{W}^{(k)} h_u' \quad (6)$$

where, $\|$ denotes concatenation.

To mitigate over-smoothing and to stabilise optimisation, each block combines residual connections with batch normalisation, giving the complete update in Eq. (7).

$$h_{v,\text{out}}^{(l+1)} = \text{BatchNorm}(\text{ELU}(h_v^{(l+1)}) + h_v^{(l)}) \quad (7)$$

After L GATv2 layers, we obtain the final node embeddings $\{h_v^{(L)}\}$. A Global Attention Pooling readout is then applied to produce the graph-level representation h_G, as in Eq. (8), where W_{gate} learns the gating (importance) score of each node and $h_G$is taken as the final molecule-level representation.

$$h_{\mathcal{G}} = \sum_{v \in V} \left(\frac{\exp(\mathbf{W}_{gate} h_v^{(L)})}{\sum_{u \in V} \exp(\mathbf{W}_{gate} h_u^{(L)})} \right) h_v^{(L)} \quad (8)$$

where, $h_v^{(L)}$ is the final feature vector of node $v$after L GNN layers. $\mathbf{W}_{gate}$ is a learnable linear transformation used to compute the gating or importance score of each node. The aggregated vector $h_{\mathcal{G}}$ is taken as the final representation of the molecular graph.

Fragment level representation

The second branch encodes molecular fragments produced by the hierarchical fragmentation. Each fragment is treated as an independent subgraph and is processed using the same representation pipeline as the whole-molecule branch, i.e., feature initialization, GATv2-based message passing, and Global Attention Pooling. This yields a fragment embedding $h_{\mathcal{G}_{frag}}$ that summarizes the corresponding fragment subgraph.

To aggregate fragment-level information back to the parent molecule, we apply sum pooling over all fragment embeddings that originate from the same molecular graph G. The resulting vector $h_{G,frags}$ is defined as:

$$h_{\mathcal{G},frags} = \sum_{g_{frag} \in \mathcal{F}(\mathcal{G})} h_{\mathcal{G}_{frag}} \quad (9)$$

where, $F(\mathcal{G})$ denotes the set of fragments derived from the original molecular graph $\mathcal{G}$.

Junction tree

In the third branch, we explicitly model fragment connectivity by constructing a junction tree, where each molecular fragment is treated as a super node and bonds between fragments are treated as edges. Message passing on this coarse grained graph enables the model to capture spatial and chemical interaction patterns among fragments. The initial feature vector of each junction tree node, $x_{v,jt\text{-}aug}$, is formed by concatenating two parts: (i) static fragment descriptors and (ii) a dynamic fragment embedding. In particular, the scalar W_i incorporates the fragment weight, allowing the network to recognise more stable structural cores and assign them higher importance during message passing, as shown in Eq. (10).

$$x_{v,jt-aug} = [c_i;\ W_i; h_{\mathcal{G},frag}] \quad (10)$$

where, $c_i$is the number of atoms contained in fragment i, W_i is the statistical weight of this fragment, and $h_{\mathcal{G},\text{frag}}$ is the graph level embedding of this fragment obtained from the second branch.

Next, the augmented feature vector is passed through a linear projection layer and mapped into the hidden space of this branch, yielding the initial hidden state of each junction tree node, as shown in Eq. (11).

$$h_{v,jt}^{(0)} = \mathbf{W}_{jt} x_{v,jt-aug} + b_{jt} \quad (11)$$

As in the first two branches, node features on the junction tree graph are then updated by L ResidualGATBlock modules. Information is propagated between nodes along edges, and the updated hidden states $h_{v,jt}^{(L)}$ represent each fragment in the context of its fragment neighbourhood. Finally, to obtain the junction tree view representation of the whole molecule, we aggregate the final hidden states of all junction tree nodes using sum pooling, as shown in Eq. (12).

$$h_{\mathcal{G},jt} = \sum_{v \in V_{JT}(\mathcal{G})} h_{v,jt}^{(L)} \quad (12)$$

where, $V_{JT}(\mathcal{G})$ denotes the set of nodes in the junction tree corresponding to molecule $\mathcal{G}$.

DATASETS AND MODEL TRAINING

This work uses 20 datasets released by Alshehri et al. [3, 11], covering four property categories: thermophysical, pharmacokinetic, toxicological, and industrial safety properties, summarised in **Table 3**. Thermophysical property prediction is directly relevant to process design and operation; for example, the normal boiling point T_b is a basic design parameter that affects equipment selection and operating conditions, while the critical constants (T_c, P_c) determine the critical point and are widely used in equations of state and phase-equilibrium calculations, including supercritical-fluid applications. Accurate prediction of safety- and environment-related properties is likewise essential for chemical safety management and environmental risk evaluation. As a representative example, auto-ignition temperature is important for fire and explosion prevention and risk assessment, and reliable prediction of solubility is critical for toxicological evaluation, drug discovery, and human-health risk assessment.

Table 3. Dataset categories and descriptions

Classification	Property	Sample number	Description
Thermophysical properties	T_b	5276	Normal boiling point (K)
	T_m	9248	Normal melting point (K)
	T_c	776	Critical temperature (K)
	P_c	774	Critical pressure (bar)
	V_c	773	Critical volume (cc/mol)
	H_v	425	Enthalpy of vaporization at 298 K (kJ/mol)
	H_{fus}	748	Normal enthalpy of fusion (kJ/mol)
	H_f	1059	Standard enthalpy of formation (kJ/mol)
	G_f	756	Gibbs energy of formation at 298 K (kJ/mol)
	L_{mv}	1055	Liquid molar volume at 298 K (cc/mol)
	HSolP	1368	Hildebrandt solubility parameter at 298 K) (MPa1/2)
Pharmacokinetics properties	BCF	589	Bioconcentration factor
	LogP	12184	Octanol–water partition coefficient
	$LogW_s$	2563	Aqueous solubility (mol/L)
	PCO	606	Photochemical oxidation potential
	pKa	1632	Acid dissociation constant
Toxicological Properties	$LC50_{(FM)}$	705	Fathead Minnow 96-H LC50 (mol/L)
	LD50	4780	Toxicity (oral rat) (mol/kg)
Industrial Safety	OSHA-TWA	422	Permissible exposure limit (mol/m3)
	AiT	570	Auto ignition temperature (K)

A six-fold cross-validation is used for dataset splitting. Specifically, 10% of the samples are first held out as an independent test set. This test set is strictly excluded from model training, hyperparameter tuning, and validation, and is used only once for the final evaluation of the selected model. The remaining 90% of the data are then randomly and evenly divided into six folds. For each run, five folds are used for training and the remaining fold is used for validation, and the procedure is repeated six times such that every fold serves as the validation set exactly once. To prevent data leakage, data standardisation is performed only after the test set has been separated.

Training, evaluation metrics and hyperparameters

Model parameters were optimised using AdamW with decoupled weight decay, which applies weight decay independently from the gradient-based Adam update and thus provides more effective regularisation than coupling L2 penalties into the gradient. The training objective was to minimise the mean squared error (MSE). To improve training stability and avoid gradient explosion, we applied gradient clipping: after backpropagatio, the global L2 norm of all parameter gradients $\| \mathbf{g} \|_2$was computed, and when it exceeded a threshold g_{clip}, gradients were rescaled. Following common practice, we set $g_{\text{clip}} = 1$. To mitigate overfitting, two complementary strategies were used. First, early stopping was adopted by monitoring the validation root mean squared error (RMSE); if no improvement was observed for 25 consecutive epochs, training was terminated and the checkpoint with the lowest validation RMSE was retained. Second, a learning-rate scheduler was employed: when validation performance plateaued, the learning rate was multiplied by 0.5 until reaching a minimum of 1×10^{-5}, enabling finer optimisation in later training stages.

Generalisation performance was evaluated using

RMSE and the coefficient of determination R^2. Under six-fold cross-validation, model selection and reporting followed a consistent protocol: within each fold, the checkpoint achieving the lowest validation RMSE was saved as the best model for that fold. After all six folds, the six resulting models were evaluated on the independent held-out test set, and the reported test metrics were computed as the average across these six models rather than from a single run.

Because different fragmentation schemes produce fragments with substantially different sizes and topological complexity, a single fixed hyperparameter configuration can unfairly constrain some models and obscure their attainable performance. We therefore applied Bayesian hyperparameter optimisation using Optuna [12]. A unified search space was defined, and for each GNNs variant constructed under a specific fragmentation method, 50 Optuna trials were executed independently. To avoid information leakage, the dataset splits used during hyperparameter search were generated with random seeds independent of those used in final evaluation. Validation RMSE was used as the optimisation objective; the best hyperparameter set was then fixed and used for the subsequent six-fold cross-validation. The complete search space and ranges are reported in **Table 4**.

Table 4 The detailed search space and ranges of all hyperparameters

Hyperparameter	Search space and ranges
Learning Rate	$[1\times10^{-4}, 1\times10^{-2}]$
Hidden Dimension	{128, 256, 512}
Attention Heads	{4, 8}
Layers	[2, 11]
Dropout	[0.0, 0.5]
Weight Decay	$[1\times10^{-6}, 1\times10^{-3}]$

RESULTS

The final predictive results are reported in **Table 5-8**. The results of the GP-WP model were taken directly from the literature [13]. The BRICS-GNN, SPE-GNN and GC-GNN models were built under different fragmentation methods. For each dataset, molecular fragments were generated for every molecule, a three-branch architecture was constructed, and property prediction was performed. HiGNN [14] and AttentiveFP [6] model were evaluated using publicly available molecular encoding pipelines, and public implementations were used for prediction and hyperparameter optimisation. To further validate the effectiveness of incorporating fragment weight features in the junction tree branch, an ablation model, Hier-AttnGNN-Ablation, was also constructed. All results are reported as the average performance on the test set over six-fold cross validation.

Table 5 Predictive performance of different models across Pharmacokinetics property datasets

	BCF	LogP	$LogW_s$	PCO	pKa
HierAttnGNN	0.72	0.39	0.76	0.14	1.67
HierAttnGNN-Ablation	0.91	0.403	0.77	0.15	1.76
GPR-WP	0.97	0.408	0.79	0.16	1.78
BRICS-GNN	0.91	0.403	0.79	0.18	1.72
SPE-GNN	0.9	0.43	0.92	0.21	1.95
GC-GNN	0.73	0.4	0.77	0.18	1.87
HiGNN	0.79	0.412	0.78	0.18	1.77
Attention FP	0.74	0.405	0.77	0.16	1.89

Table 6 Predictive performance of different models across Toxicological and Industrial Safety property datasets

	AiT	$LC50_{(FM)}$	LD50	OSHA-TWA
HierAttnGNN	76.09	0.67	0.43	0.66
Hier-AttnGNN-Ablation	78.3	0.703	0.45	0.74
GPR-WP	94.64	0.76	0.43	1.3
BRICS-GNN	89.97	0.96	0.44	0.78
SPE-GNN	83.47	0.89	0.47	1.08
GC-GNN	81.22	0.78	0.46	0.88
HiGNN	81.44	0.69	0.46	0.87
Attention FP	83.01	0.68	0.45	0.77

Table 7 Predictive performance of different models across Thermophysical property datasets (Part 1)

	T_b	V_c	T_c	P_c	G_f
Hier-AttnGNN	19.19	27.31	21.06	2.43	22.45
Hier-AttnGNN-Ablation	19.97	29.57	25.76	3.15	25.39
GPR-WP	25.39	33.43	86.8	2.77	44.37
BRICS-GNN	23.06	32.73	27.5	2.53	32.24
SPE-GNN	27.86	75.12	43.24	4.02	36.11
GC-GNN	19.4	29.88	21.37	2.47	24.66
HiGNN	19.27	34.62	23.21	2.84	24.06
Attention FP	20.77	30.44	22.62	2.99	26.14

Table 8 Predictive performance of different models across Thermophysical property datasets (Part 2)

	H_f	H_{fus}	HSolP	H_v	L_{mv}
Hier-AttnGNN	20.95	3.74	1.66	4.82	0.0068
Hier-AttnGNN-Ablation	24.25	3.88	1.68	4.95	0.0103
GPR-WP	37.42	4.099	2.31	5.72	0.007
BRICS-GNN	28.56	4.44	1.78	5.97	0.019
SPE-GNN	45.72	4.53	1.72	5.95	0.02
GC-GNN	24.9	5.3	1.71	5.01	0.0105
HiGNN	27.83	4.04	1.7	5.18	0.007
Attention FP	26.69	3.76	1.72	4.98	0.0074

HierAttnGNN achieved the best performance across all 20 prediction tasks, with an average RMSE of 12.44, representing a 6.8% reduction relative to GC-GNN at 13.36. Among the models sharing the same three-branch architecture, SPE-GNN consistently produces the highest errors, confirming that discarding topological connectivity between fragments severely impairs predictive accuracy. BRICS-GNN generally outperforms SPE-GNN but suffers notable degradation where its predefined rules fail to preserve intact functional groups, as evidenced by RMSE values of 0.96 on $LC50_{(FM)}$ and 89.97 on AiT, compared with 0.67 and 76.09 for HierAttnGNN. GC-GNN performs comparably to HierAttnGNN on classical thermophysical properties where its predefined library provides adequate coverage, yet it yields the worst RMSE of 5.3 on H_{fus}, illustrating how fine-grained fragmentation of out-of-library substructures leads to sparse feature vectors and poor generalisation. AttentiveFP achieves competitive results on properties governed by local atomic environments, attaining RMSE values of 3.76 on H_{fus} and 0.68 on $LC50_{(FM)}$, but falls behind on tasks requiring multi-level structural reasoning, such as Hf and pKa. HiGNN ranks among the top three on many datasets but inherits the functional-group integrity limitations of BRICS, leading to higher errors on toxicological and structurally diverse thermophysical datasets.

CONCLUSION

A hierarchical molecular fragmentation method is proposed in this work. Without relying on predefined rules or a static vocabulary, the proposed method can automatically partition a molecule into fragments that contain key functional groups. On this basis, a three-branch GNNs is constructed to learn hierarchical representations from the atom level to the fragment level and further to the junction tree level. Evaluation on 20 benchmark datasets shows that HierAttnGNN significantly outperforms existing mainstream model. The average RMSE is reduced by 6.8% relative to the second-best model.

ACKNOWLEDGEMENTS

Authors appreciate financial support from China Scholarship Council (202406440073)

AUTHOR IDENTIFIERS

Author ORCIDs:
Jianfeng Jiao: 0009-0009-8246-3861
Jie Li: 0000-0001-5196-2136

REFERENCES

1. Rogers D, Hahn M. Extended-connectivity fingerprints. J. Chem. Inf. Model. 50:742-754 (2010). https://doi.org/10.1021/ci100050t
2. Gani R. Group contribution-based property estimation methods: advances and perspectives. Current Opinion in Chemical Engineering 23:184-196 (2019). https://doi.org/10.1016/j.coche.2019.04.007
3. Alshehri AS, Tula AK, You F, Gani R. Next generation pure component property estimation models: with and without machine learning techniques. AIChE Journal 68: (2021). https://doi.org/10.1002/aic.17469
4. Aouichaoui ARN, Fan F, Abildskov J, Sin G. Application of interpretable group-embedded graph neural networks for pure compound properties. Computers & Chemical Engineering 176:108291 (2023). https://doi.org/10.1016/j.compchemeng.2023.108291
5. Gilmer J, Schoenholz S S, Riley P F, Vinyals O, Dahl G E. Neural message passing for quantum chemistry. in *International conference on machine learning* 1263–1272 (2017). https://doi.org/proceedings.mlr.press/v70/gilmer17a
6. Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, Li Z, Luo X, Chen K, Jiang H, Zheng M. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J. Med. Chem. 63:8749-8760 (2019). https://doi.org/10.1021/acs.jmedchem.9b00959
7. Li X, Fourches D. SMILES pair encoding: a data-driven substructure tokenization algorithm for deep learning. J. Chem. Inf. Model. 61:1560-1569 (2021). https://doi.org/10.1021/acs.jcim.0c01127
8. Hukkerikar AS, Sarup B, Ten Kate A, Abildskov J,

Sin G, Gani R. Group-contribution+ (GC+) based estimation of properties of pure components: improved property estimation and uncertainty analysis. Fluid Phase Equilibria 321:25-43 (2012). https://doi.org/10.1016/j.fluid.2012.02.010
9. Wang J, Wang Y. Brics-based generation and ai-assisted screening of ionic liquids with mechanistic insights into lithium transport in electrolytes. J. Chem. Inf. Model. 65:10961-10976 (2025). https://doi.org/10.1021/acs.jcim.5c01824
10. Brody S, Alon U, Yahav E. How attentive are graph attention networks? *arXiv Prepr. arXiv2105.14491* (2021) https://doi.org/10.48550/arXiv.2105.14491
11. Jiao J, Gao X, Li J. Pure component property estimation framework using explainable machine learning methods. Chinese Journal of Chemical Engineering 84:158-178 (2025). https://doi.org/10.1016/j.cjche.2025.05.011
12. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining :2623-2631 (2019). https://doi.org/10.1145/3292500.3330701
13. Cao X, Gong M, Tula A, Chen X, Gani R, Venkatasubramanian V. An improved machine learning model for pure component property estimation. Engineering 39:61-73 (2024). https://doi.org/10.1016/j.eng.2023.08.024
14. Zhu W, Zhang Y, Zhao D, Xu J, Wang L. Hignn: a hierarchical informative graph neural network for molecular property prediction equipped with feature-wise attention. J. Chem. Inf. Model. 63:43-55 (2022). https://doi.org/10.1021/acs.jcim.2c01099