



## Application of Interpretable Group-embedded Graph Neural Networks for Pure Compound Properties

Aouichaoui, Adem R.N.; Fan, Fan; Abildskov, Jens; Sin, Gürkan

*Published in:*  
Computers and Chemical Engineering

*Link to article, DOI:*  
[10.1016/j.compchemeng.2023.108291](https://doi.org/10.1016/j.compchemeng.2023.108291)

*Publication date:*  
2023

*Document Version*  
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

*Citation (APA):*  
Aouichaoui, A. R. N., Fan, F., Abildskov, J., & Sin, G. (2023). Application of Interpretable Group-embedded Graph Neural Networks for Pure Compound Properties. *Computers and Chemical Engineering*, 176, Article 108291. <https://doi.org/10.1016/j.compchemeng.2023.108291>

---

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.



# Application of interpretable group-embedded graph neural networks for pure compound properties

Adem R.N. Aouichaoui, Fan Fan, Jens Abildskov, Gürkan Sin <sup>\*</sup>

*Process and Systems Engineering Center (PROSYS), Department of Chemical and Biochemical Engineering, Technical University of Denmark, Building 229, 2800 Kgs. Lyngby, Denmark*



## ARTICLE INFO

### Keywords:

Deep-learning  
Graph neural networks  
Group-contribution models  
Thermophysical properties  
Interpretability  
Pure compound properties

## ABSTRACT

The ability to evaluate pure compound properties of various molecular species is an important prerequisite for process simulation in general and in particular for computer-aided molecular design (CAMD). Current techniques rely on group-contribution (GC) methods, which suffer from many drawbacks mainly the absence of contributions for specific groups. To overcome this challenge, in this work, we extended the range of interpretable graph neural network (GNN) models for describing a wide range of pure component properties. The new model library contains 30 different properties ranging from thermophysical, safety-related, and environmental properties. All of these have been modeled with a suitable level of accuracy for compound screening purposes compared to current GC models used within CAMD applications. Moreover, the developed models have been subjected to a series of sanity checks using logical and thermodynamic constraints. Results show the importance of evaluating the model across a range of properties to establish their thermodynamic consistency.

## 1. Introduction

Molecules and chemical compounds are important for any chemical engineering application (Frenkel, 2011). An estimate of the theoretically available design space for organic compounds following the Lipinski rule of bioavailability is  $10^{60}$  (Reymond, 2015). For compounds up to 30 atoms, the estimate is approximately  $10^{24}$  (Reymond, 2015). Several studies have enumerated the possible compounds within a given design constraint by considering a molecule as a mathematical graph, where the atoms are replaced by graph nodes and chemical bonds by graph edges. The GDB17 is a generated database of organic compounds containing up to 17 atoms of Carbon (C), Nitrogen (N), Oxygen (O), Sulfur (S), and Halogens (Fluorine (F), Chlorine (Cl), Bromine (Br), Iodine (I)) that resides over 166 billion organic compounds (Ruddigkeit et al., 2012). The prospect of experimentally determining all relevant properties of such a chemical space is impractical for many research and engineering studies. Especially considering the diverse properties needed for various applications and domains covering ADME (absorption, distribution, metabolism, and excretion) properties, thermophysical properties, and environmental and safety-related properties, not to mention temperature and pressure sensitive properties and phase equilibria data. Commonly used predictive models are so-called quantitative

structure-property relationships (QSPR), which aim to establish a correlation between the molecular structure and the target property of interest (Katritzky et al., 2010). QSPR models take as input a numerical translation of the molecular structure referred to as molecular descriptor into a mathematical model that establishes the relation between the descriptor and the target property (Gasteiger, 2016; Katritzky et al., 2010). The nature of the descriptor and mathematical model are not universal and no widespread consensus has been established on how to best represent these. As such they widely differ depending on the field of application. Atomic coordinates (x-y-z coordinates) and calculated quantum mechanics-related properties are widely used for QSPR models for catalyst design (Parveen et al., 2019). Molecular fingerprints derived from algorithms detecting the presence of a set of predefined substructures such as the Morgan fingerprints and extended connectivity fingerprints (ECFP) are widely used for drug discovery (Liu and Zhou, 2008; Rogers and Hahn, 2010). For chemical engineering applications, QSPR models have largely been based on the concept of group-additivity models in the form of group-contribution (GC) models, where the molecule is described in terms of the occurrence of a set of predefined functional groups (Gani, 2019; Van Speybroeck et al., 2010). The popularity of this approach has given rise to a wide range of methods for describing the molecule in terms of groups (Tu, 1995; Constantinou and

\* Corresponding author.

E-mail address: [gsl@kt.dtu.dk](mailto:gsl@kt.dtu.dk) (G. Sin).

Gani, 1994; Hukkerikar et al., 2012b; Joback and Reid, 1987; Kliniewicz and Reid, 1984; Marrero and Gani, 2001; Meier, 2021a). The widespread use of such models is largely owing to their simplicity (linear additive models), transparent and clear aspect of interpretability (each group/fragment has a proper and fixed contribution towards the overall property) as well as their availability across a wide range of properties (Gani, 2019). One of the most detailed GC methods in terms of the defined groups and domain of applicability assumes a three-level hierarchical description of the molecule (Hukkerikar et al., 2012b). Each level provides a higher resolution and larger substructures of the molecule (220 first-order groups, 137 second-order groups, and 74 third-order groups). The GC model has been applied to 18 datasets related to the thermodynamic and physical properties of organic compounds covering liquid molar volumes, acentric factor, octanol-water partition coefficient, solubility parameters as well as melting/boiling points, critical point measurements, and enthalpy-related properties (energy of formation and energy of phase transition) (Hukkerikar et al., 2012b). A study further expanded on the already huge library of available properties to cover environmental-related properties (a total of 22 properties) (Hukkerikar et al., 2012a) and flammability-related properties (5 properties) (Frutiger et al., 2016a) making it the most widely applied GC method across properties. Establishing a wide repository of property models is necessary for in-silico screening purposes and plays an integral part in any computer-aided molecular design (CAMD) framework. Despite the advantages that GC methods present in terms of simplicity, interpretability, and a wide range of availability, they do come with a set of serious drawbacks. GC models are in essence linear additive models and as such, they struggle in capturing the non-linearity that some properties present. Potential proximity effects resulting from the relative positioning of groups to each other in a molecule are completely neglected. They are highly dependent on the data used for the regression in terms of determining the contribution of various groups as well as the absence of any established methods for determining the contribution of missing groups in the dataset, which results in a sparse contribution matrix. In an effort to remedy this, all data are usually used in the regression process, which constrains the ability to perform cross-validation (Hukkerikar et al., 2012b; Jhamb et al., 2020). As a consequence, the ability to generalize to unseen compounds is not well established. A final drawback resides in the definition of the groups, as not all are theoretically motivated or based on an understanding of the underlying phenomenon, which is especially true for the higher order groups (2<sup>nd</sup> and 3<sup>rd</sup> order groups) (Hukkerikar et al., 2013). While this ad-hoc definition for groups has improved the performance metrics of the models, it is often a misleading improvement as errors are “absorbed” by the defined group. A conclusion that is rooted in the fact that in many cases, the groups only feature once in a given dataset, and as such, the contribution of the group will take the value of the residual. Machine learning (ML) and artificial intelligence (AI) techniques have been applied to overcome some of the drawbacks presented by classical GC models, especially for capturing the potential non-linearity present in the datasets by replacing the linear model with Gaussian processes, neural networks, and support vector machines as well as employing cross-validation practices (Alshehri et al., 2022; Aouichaoui et al., 2021; Mondejar et al., 2017). However, the remaining shortcomings have largely been undealt with.

Recent advancements in the fields of ML and AI have the potential of solving some of the remaining drawbacks of GC methods and even replacing them completely. This can be done by performing feature extraction and construction using general and accessible molecular information. Graph neural networks (GNNs) are among the most used techniques for such a unified approach that combines the feature extraction process and the regression process (Hwang et al., 2020; Wieder et al., 2020; Zhou et al., 2018). A GNN takes as input a graph representation of the molecule where nodes correspond to atoms and edges correspond to the bonds linking the atoms. Each node and edge is associated with a feature vector containing information concerning the

atoms and bond e.g. atom type, number of hydrogen attached, and bond type. Through a series of arithmetic operations involving matrix multiplication, the graph representation is updated and converted into a numerical vector that is used as the molecular descriptor. These operations are denoted message-passing framework and comprise the mechanism that dictates the feature extraction process (Gilmer et al., 2017). This mechanism produces a molecular representation that is aware of the neighboring atoms and as such, it allows for taking proximity effects into account. An important advantage of the GNN models is that they are capable of learning a proper representation of the molecule from general atomic and bond features. In this way, they are less prone to dependency on the functional groups (representing specific atom/bond arrangements) defined in the group contribution methods. This could potentially allow the models to represent unseen molecules during training (Aouichaoui et al., 2023). In fact, recent years have witnessed a rapid increase in GNN models for molecular property prediction each developing and extending upon the existing models with the most recent developments from the field of deep learning (DL) (Wieder et al., 2020; Zhou et al., 2018). The first account of the GNN model was introduced by Davenaud et al. (Duvenaud et al., 2015) and was among the first models to learn specific molecular representation depending on the task. Gilmer et al. provided the first definition for the message-passing framework, which unifies many GNN models in the way they extract the molecular representation (Gilmer et al., 2017). Many studies involving new GNN models have been reported in the literature. For example, Coley et al. used attributed molecular graphs that use an extensive list of atom and bond features in order to produce a more informed molecular representation (Coley et al., 2017). The model was later altered to include bond-based message-passing rather than node-based message passing (Yang et al., 2019). A recent study combined the bond-based message-passing model with fixed descriptors obtained from various chemoinformatics software and applied it to a set of environmental properties (Zhang et al., 2022). Schweidtmann et al. developed a higher-order GNN model for fuel ignition properties (Schweidtmann et al., 2020). Xiong et al. combined recurrent neural networks and the attention mechanism as part of the mechanism generating the molecular representation and denoted the model attentive FP, achieving state-of-the-art performance compared to many existing models and providing an added benefit in the form of increased interpretability (Xiong et al., 2020). Zhang et al. proposed a three-level hierarchical feature extraction model (FraGAT) conducted on the node level of a molecular graph, a fragment level by enumerating all possible fragments of a molecular and a junction-tree level by combining the fragments into a new molecular graph (Zhang et al., 2021). Recently, a model inspired by the FraGAT introduced the groups as defined by GC models as fragments showcasing increased performance and faster computation time compared to the original model (Aouichaoui et al., 2023). Important to note that despite many of these studies and models achieving state-of-the-art performance in many applications, they –similar to GC models- come with a set of drawbacks. Depending on the data and complexity of the model, AI models are associated with long training times. This however can be remedied by using dedicated hardware in the form of graphical process units (GPUs). Another drawback is that AI-based models are high parametric models with very high degrees of freedom (DoF), which eventually result in parameter identifiability issues and unstable training procedures. Another issue is that despite gaining a lot of attention lately, the uncertainty quantification of such models is also not well established, with many methods available and no apparent conclusion on which method is best (Aouichaoui et al., 2022a; Hirschfeld et al., 2020; Scialia et al., 2020). Some of the applied methods include Monte-Carlo dropout, bootstrapping (either from the training data or from the errors), and deep ensemble, with the latter being one of the most widely used methods (Aouichaoui et al., 2022a; Hirschfeld et al., 2020; Scialia et al., 2020). A further drawback is the “black-box” nature of these models, a fact which might hinder their wider acceptance and applicability, especially in domain applications

**Table 1**

Atom and bond information used to featurize the molecular graph.

Feature	Explanation	Type	Size
<b>Atom features</b>			
Atom type	type of atom (C, N, O, S, F, Cl, Br, I, P)	One-hot encoding	9
No. of bonds	number of bonds attached to the atom (0, 1, 2, 3, 4, 5)	One-hot encoding	6
No. of H's	number of Hydrogen attached to the atom (0,1,2,3,4)	One-hot encoding	5
Valency	explicit valency (0,1,2,3,4,5)	One-hot encoding	6
Hybridization	hybridization (sp, sp <sup>2</sup> , sp <sup>3</sup> , sp <sup>3</sup> d, sp <sup>3</sup> d <sup>2</sup> )	One-hot encoding	5
Aromaticity	whether the atom is part of an aromatic system (0, 1)	binary	1
Chirality center	whether the atom is a center of chirality (0,1)	binary	1
Chirality type	type of chirality the atom is involved in (R, S)	One-hot encoding	2
Chirality tag	tag assigned to the chiral center ("unspecified", tetrahedral_CW, tetrahedral_CWW)	One-hot encoding	3
Formal Charge	charge assigned to individual atoms in a molecule	Integer	1
<b>Bond features</b>			
Bond type	bond type (single, double, triple, aromatic)	One-hot encoding	4
Conjugation	whether the bond is conjugated (0, 1)	Binary	1
Ring	whether the bond is part of a ring (0, 1)	Binary	1
Stereochemistry	Bond stereochemistry (none, any, Z/E, Cis/Trans)	One-hot encoding	6

that are traditionally based on a first principle understanding of the underlying tasks and phenomenon. Recent studies have made significant advancements in addressing the interpretability issue of deep-learning models by proposing several approaches to gain insight into the learning achieved by such models(Hasebe, 2021; Jiménez-Luna et al., 2021, 2020; Xiong et al., 2020; Zhang et al., 2021). In the context of GNN models, interpretability can be achieved by feature attribution, subgraph approaches, or the attention mechanism (Jiménez-Luna et al., 2021, 2020). The feature attribution approach consists of using the integrated gradient method in the backpropagation to produce an importance score for the feature elements in each node and edge (Jiménez-Luna et al., 2021). The method has previously been used for drug-related properties and showed that the method was able to highlight structural elements that are consistent with known pharmacophore motifs (Jiménez-Luna et al., 2021). The study does however also report some challenges with the method such as the presence of multicollinearity between features and instability in importance attribution for structurally similar compounds (Jiménez-Luna et al., 2021). However perhaps one major drawback is the fact that the insights are provided on an individual feature level and as such the insights are only provided at a micro-level. The attention mechanism can easily be integrated into the message-passing framework by applying a scoring function or weighting function to evaluate the importance of either the nodes or the bonds in the produced molecular descriptor generated through the readout function. This method has been used to highlight atoms and bonds with importance for water solubility of organic compounds (Xiong et al., 2020). However, in reality, it is rather functional groups and substructures of the molecule that influences the various properties of a molecule. As such, substructure (or subgraph) based approaches are more intuitive to chemists and the fundamental understanding of properties (Jiménez-Luna et al., 2020). A recent study integrated the concept of functional groups from the GC methods with GNNs to add an interpretable aspect to the GNN models (Aouichaoui et al., 2023). The interpretability consisted of using the attention mechanism to provide an importance ranking and color the fragments based on their importance. The models (AGC and GroupGAT) were tested in a range of properties to ensure the accuracy of the models is not compromised compared to state-of-the-art models and the insights provided were shown to be consistent with the insights gained from the classical GC models.

In this work, we extend and study a large collection of property models based on the AGC and GroupGAT models to rigorously test their performance on a wide range of different properties for different applications. The establishment of such a collection of property models is essential to construct a unified framework for the in-silico screening of organic compounds for different end uses. A further added value of this

work resides in a set of "sanity" checks that prove the models obtained abide by thermodynamic and logical constraints.

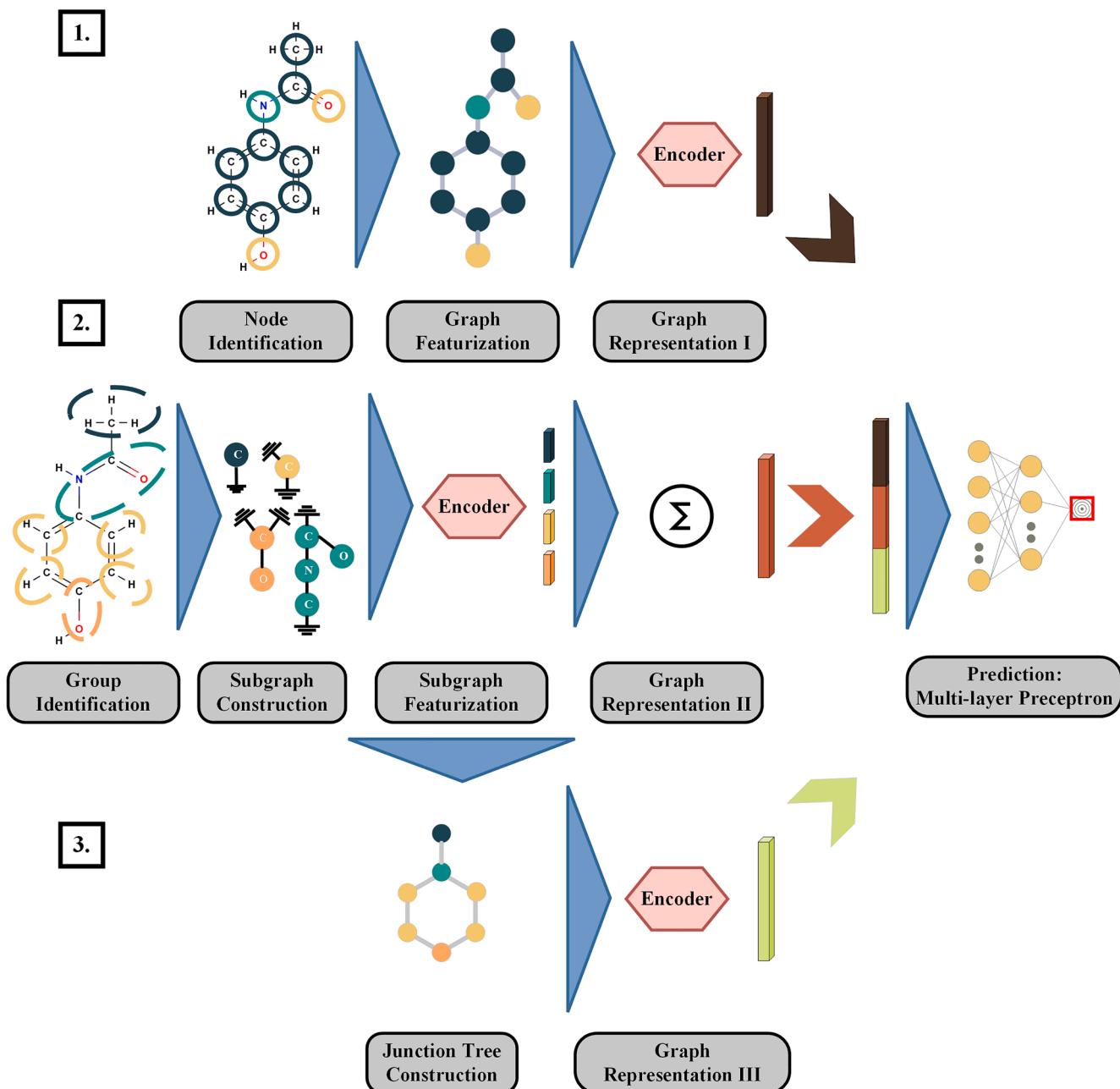
## 2. AGC and GroupGAT: interpretable graph neural networks using functional groups

In this section, the concept of GNNs for molecular property prediction is described. Furthermore, two GNNs that integrate functional groups into their structure are reintroduced along with the underlying mathematical schemes while highlighting the model hyperparameters.

### 2.1. Graph Neural Networks for molecular property prediction

GNNs are neural networks that operate on graph-structured data and are capable of learning a numerical representation of the input data and correlating the learned representation to the desired target output (either categorical or continuous variables) in an end-to-end learning framework (combining feature extraction and regression tasks). Since molecules can be described in terms of molecular graphs (similar to the ball and stick representation), they possess a suitable format for GNNs. The graph (or molecule) is made of a set of nodes or vertices (atoms) linked with edges (chemical bonds). Each node and edge is attributed (or featurized) with a set of information relating to the atom or bonds respectively. Table 1 showcases the atom and bond information used in this work. The feature extraction task is done through three steps: the message construction phase, the update phase, and the readout phase. These operations are done node-wise, where the node considered is denoted central node and the linked nodes are denoted neighboring nodes.

During the message construction phase, the neighboring atom and bonds are combined through a series of mathematical operations (denoted message function) to produce a representation that is combined with the currently hidden representation of the central node to alter its representation (update phase). As such, the latent representation of a node is influenced by the environment/neighborhood in which it is located. This is repeated "L" times corresponding to the number of message passing layers, after which the graph representation is converted into a vector representation through a readout function. The vector representation is then used as input to a multi-layer perceptron (MLP) (or a feedforward neural network) to correlate the obtained representation to the target value. Both operations (feature extraction and regression tasks) are combined through the backpropagation algorithm and as such, a task-specific representation (learned molecular representation or molecular descriptor) is produced.



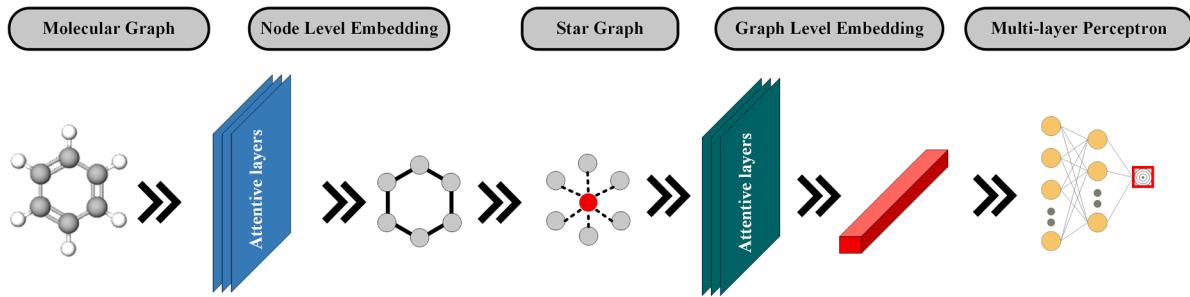
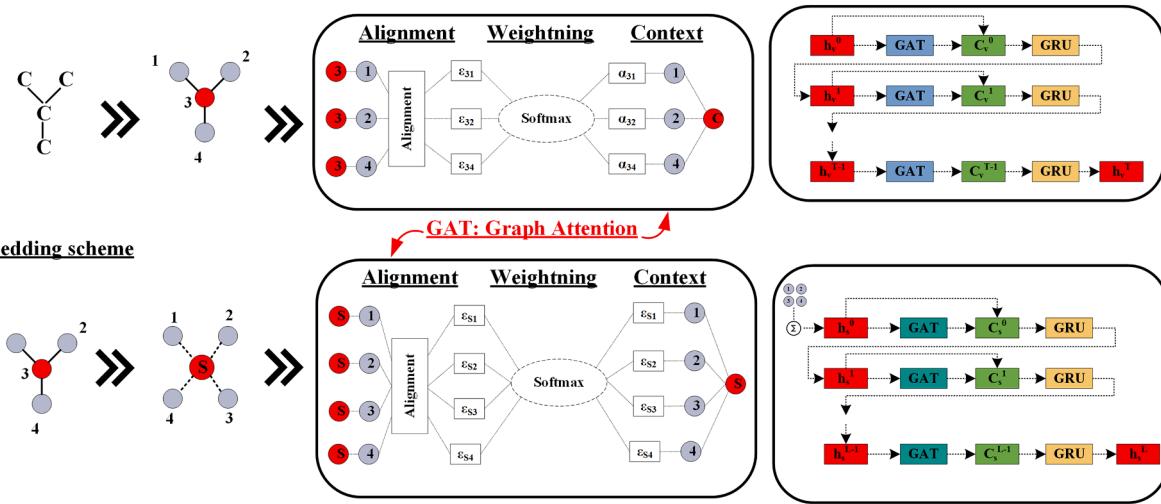
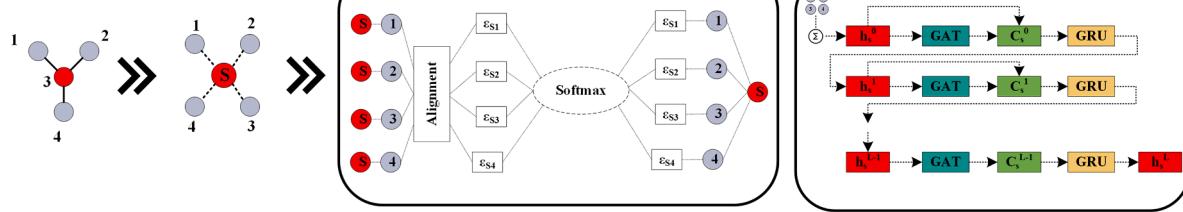
**Fig. 1.** Schematic of GroupGAT. The top branch (1) shows a node-based representation extraction using an encoder. Middle branch (2) shows the fragment-based representation using functional groups. The bottom branch (3) constructs a junction tree using the groups as nodes. All three representations are concatenated to produce the overall representation of the molecule. The representation is used as input to the regressor in the form of a multi-layer perceptron to produce a prediction.

## 2.2. GroupGAT: Group-Contribution Graph Attention Network

The GroupGAT model is inspired by the FraGAT model (Zhang et al., 2021). The model has previously been benchmarked against a wide range of message-passing neural network models including the FraGAT (Zhang et al., 2021), attentive FP (Xiong et al., 2020), and MPNN (Gilmer et al., 2017) and the D-MPNN (Yang et al., 2019). The study showed that the GroupGAT had increased performance and increased interpretability (Aouichaoui et al., 2023). The model is comprised of three branches that each produce a representation operating on different resolutions of the molecule, a similar approach to the three-level GC models (Hukkerikar et al., 2012b). The hierachal feature extraction is conducted on the following levels:

- 1 Complete molecular graph with nodes and edges as the center of the message-passing framework
- 2 Individual fragments or groups are considered individual graphs
- 3 A junction-tree model with groups (fragments) as the center of the message-passing framework

Each branch produces a representation that is concatenated to produce the overall molecular descriptor. An important difference between the FraGAT and the GroupGAT is that the groups (or fragments) defined for the GroupGAT are taken from the well-established groups defined in GC models (Hukkerikar et al., 2012b). This chemistry-informed fragmentation is inspired by chemists' understanding of the underlying mechanism in QSPR models (Meier, 2021a). GroupGAT only considers the first-order groups since they do not present any overlaps between

**A) Overall attentiveFP scheme****B) Node embedding scheme****C) Graph embedding scheme**

**Fig. 2.** Schematic overview of the attentive FP model: a complete overview of the attentive FP model (a), the node embedding scheme (b), and the graph embedding scheme (c)

substructures (Hukkerikar et al., 2012b; Marrero and Gani, 2001). A schematic of the overall structure of GroupGAT is shown in Fig. 1. For molecules that contain fragments not previously defined, the algorithm assigns the fragment as “*unknown*”. Though the fragment is considered unknown in the context of the defined groups, the model retains knowledge of the atoms and the arrangements associated with it and as such this does not pose a challenge for the model and does not limit its applicability to molecules that are not entirely fragmented in terms of the predefined set of groups. As such, the fragmentation can be considered a guide for the model to fragment the molecule as much as possible based on prior chemistry knowledge.

In essence, any GNN can be used to represent the encoder block. However, to retain the interpretability of the models, the attentive FP model is used (Xiong et al., 2020). The attentive FP model combines the graph attention mechanism with a unique readout function to produce the molecular representation. In the encoder, the representation is constructed over two steps: a node-level and a graph-level encoding process. The node-level embedding is very similar to the commonly used message-passing framework where the hidden node representation is repeatedly updated with information from neighboring atoms by stacking layers. The node hidden representation ( $h_v$ ) is initialized according to Eq. (1) using the initial atoms features ( $x_v^{atom}$ ) and adjustable weight matrix ( $W$ ) that is determined through backpropagation. The neighboring node hidden representations ( $h_w$ ) are initialized according to Eq. (2) using the initial atom ( $x_w^{atom}$ ) and edge features ( $x_{v,w}^{bond}$ ). The attention mechanism is summarized in three operations, all contributing to the construction of the message used to update the node latent representation: alignment (shown in Eq. (3)), weighting (shown in Eq. (4))

and context (shown in Eq. (5)). The purpose of the alignment is to produce a representation that combines the hidden representation ( $\epsilon_{v,w}$ ) of the central atom and the neighboring atoms through an adjustable weight matrix ( $W$ ) and a non-linear activation function (in this case a leaky ReLU). The weighting process aims to reduce the output of the alignment process into a weight coefficient ( $\alpha_{v,w}$ ) that sums up to 1 for all neighboring nodes using the softmax function. The coefficient is used as a weight for the importance/influence the hidden representation of the neighboring node should have when updating the hidden representation of the central atom as seen in the context step (Eq. (5)). The message produced through the context operation is then used to update the latent representation of the central node through a gated recurrent network (GRU) (as seen in Eq. (6)).

**Node-level embedding operations:**

*Center & Neighbor initialization:*

$$h_v^0 = \text{ReLU}(W \cdot x_v^{atom}) \text{ for } t = 0 \quad (1)$$

$$h_w^0 = \text{ReLU}\left(W \cdot [x_w^{atom}, x_{v,w}^{bond}]\right) \text{ for } t = 0 \quad (2)$$

**Alignment:**

$$\epsilon_{v,w}^{t+1} = \text{leakyReLU}(W \cdot [h_v^t, h_w^t]) \text{ for } t + 1 \leq T \quad (3)$$

**Weighting:**

$$\alpha_{v,w}^{t+1} = \text{softmax}\left(\epsilon_{v,w}^{t+1}\right) \text{ for } t + 1 \leq T \quad (4)$$

**Context:**

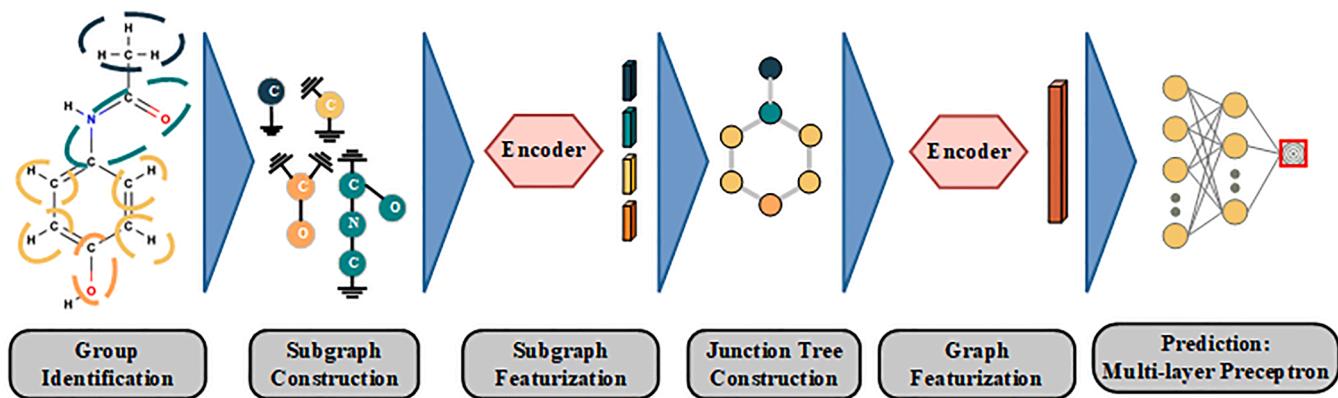


Fig. 3. Schematic of the attentive group-contrtribution model (AGC).

$$m_v^{t+1} = C_v^{t+1} = \text{elu} \left( \sum_{w \in \text{Neighbor}(v)} \alpha_{v,w} \cdot W \cdot h_w^t \right) \text{ for } t+1 \leq T \quad (5)$$

*Update:*

$$h_v^{t+1} = \text{GRU}(h_v^t, C_v^{t+1}) \text{ for } t+1 \leq T \quad (6)$$

After the final node embedding layer, a super node (also referred to as a virtual node) is constructed as an alternative representation of the complete molecule, with all the nodes of the graph converging towards it. This framework makes it possible to illustrate the importance each node has with regard to the overall graph (molecule) representation. The final hidden representation of the nodes is used to construct the hidden representation of the super node as shown in Eq. (7). The attention mechanism is applied again however it only considers one central node which is the super node (alignment in Eq. (8), the weighting in Eq. (9) and the context in (Eq. (10)). Important to note that the attention here considers all nodes towards the whole graph (represented as a super node herein) as such the importance of each atom towards the overall target property can be investigated through the weight coefficient produced. This aspect makes the attentive FP unique when using the attention mechanism for interpretability purposes. For the final time step ( $t+1=L$ ) the update can be considered as the readout step (see Eq. (11)). This representation is then used as an input to a multi-layer perceptron to regress and produce the target property.

#### Graph level embedding operations:

##### Graph initialization:

$$h_s^0 = \sum_{v \text{ in } G} h_v^T \text{ for } l=0 \quad (7)$$

##### Alignment:

$$e_{s,v}^{l+1} = \text{leakyReLU}(W \cdot [h_s^l, h_v^T]) \text{ for } l+1 \leq L \quad (8)$$

##### Weighting:

$$\alpha_{s,v}^{l+1} = \text{softmax}(e_{s,v}^{l+1}) \text{ for } l+1 \leq L \quad (9)$$

##### Context:

$$C_s^{l+1} = \text{elu} \left( \sum_{v \in \text{Neighbor}(s)} \alpha_{s,v} \cdot W \cdot h_v^l \text{ for } l+1 \leq L \right) \quad (10)$$

##### Update:

$$h_s^l = \text{GRU}(h_s^l, C_s^{l+1}) \text{ for } l+1 \leq L \quad (11)$$

A schematic of the attentive FP model can be seen in Fig. 2 as well as an inside view of the flow of information for the node embedding layers and the graph embedding layers.

The interpretability consists of showing the weight each fragment (or node) has when producing the final molecular representation. The

encoding is performed on all three hierarchical levels as shown in Fig. 1. As such, it is important to note that three distinct attention information are obtained from each branch. Therefore, the interpretability can be considered “partial” depending on the branch in focus. In order to establish “complete” interpretability within the attention mechanism framework, the attentive group-contrtribution (AGC) model is developed.

### 2.3. AGC: attentive group-contrtribution model

The AGC is a compact version of the GroupGAT and aims to provide “total” interpretability within the framework of the attention mechanism (Aouichaoui et al., 2023). The schematic of the AGC can be seen in Fig. 3. In brief, the top branch of the GroupGAT is discarded, while the latent representation produced from the middle branch (in the context of GroupGAT) is not used for the final representation of the graph but only to featurize the subgraphs in the junction-tree model. As such, the attentions obtained are for all fragments w.r.t. the overall graph. Similarly, the encoder block used is the attentive FP model.

The GroupGAT and AGC offer several advantages compared to classical GC models that are based on multi-linear regression. The models are only based on the first-order groups, which are rooted in the theoretical understanding of QSPRs. This is not always the case for the second and third-order groups, which have been defined due to convenience to improve the model predictions (Hukkerikar et al., 2013). Through the message-passing scheme, the GroupGAT and AGC account for possible proximity effects since the group representation is a function of the neighboring groups. The multi-layer perceptron allows for correlating highly non-linear behavior, while the classical GC are linear additivity models. Compared to other GNN models, the GroupGAT and AGC also provide more informative interpretability since the attention is performed based on functional groups rather than atoms. This could potentially prove useful in gaining insights into less understood properties.

## 3. Methods

The methods and steps used in this work are summarized in Fig. 4. The framework can overall be divided into the data processing step, model development, and model evaluation, each of which will be elaborated on in this section.

### 3.1. Data and preprocessing

The data needed for developing the model comprises of two elements, a molecular identifier in the form of SMILES (simplified molecular input line entry system) and a numerical target value. In this work, we investigate a total of 30 properties classified into three categories: thermophysical properties (18 properties), flammability/safety-related

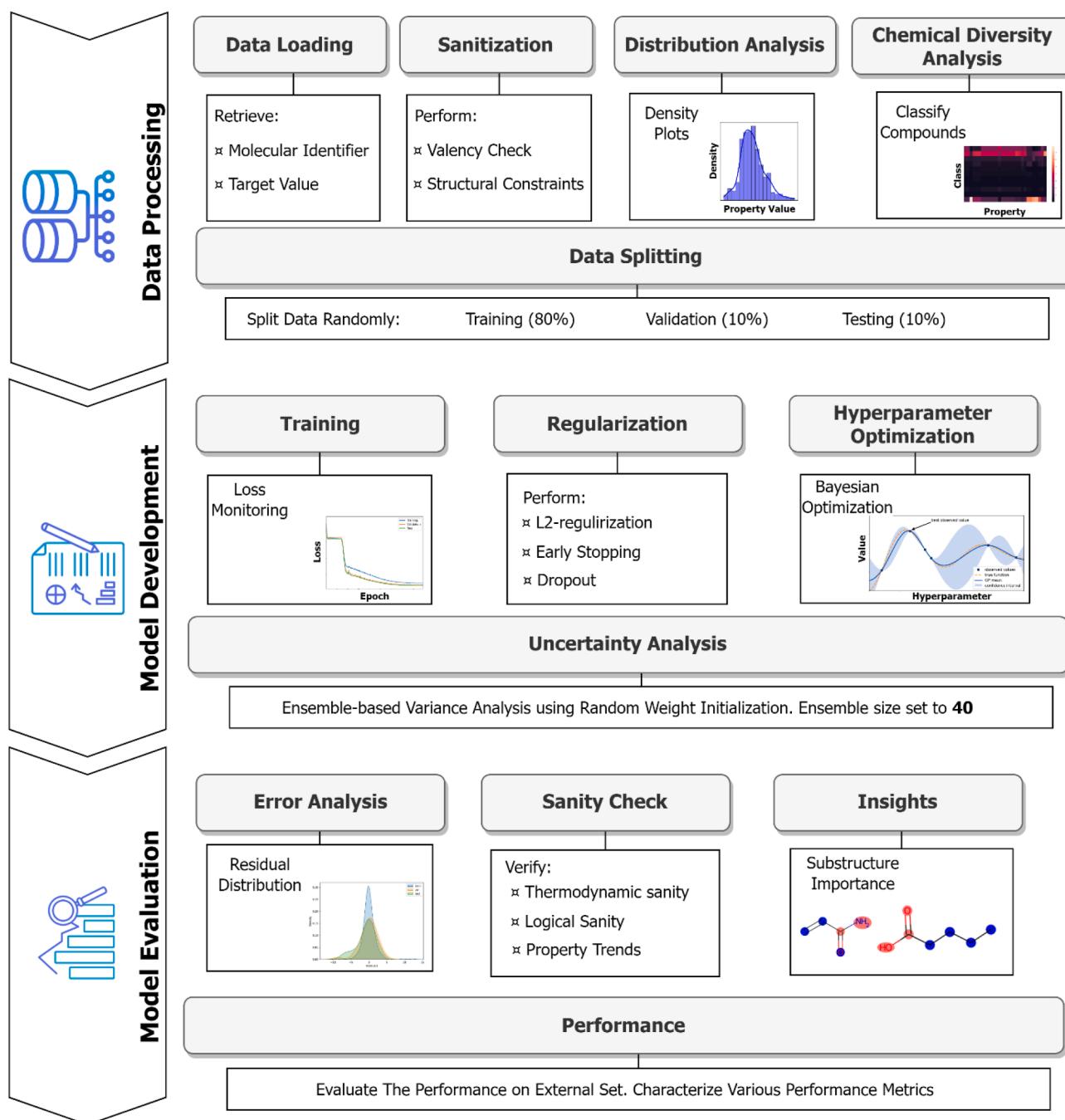


Fig. 4. Framework for developing AI-based property models.

properties (5 properties), and environmental properties (7 properties). The data collected mainly come from the Design Institute for Physical Properties under the American Institute of Chemical Engineers (AIChE DIPPR)(Rowley et al., 2019), the US environmental protection agency (EPA)(Mansouri et al., 2016) and other published experimental data (Alshehri et al., 2022; Nielsen et al., 2001). An overview of the property investigated, their definition, and the source from which they were obtained can be seen in Table 2. In this study, we opted for keeping the source of data homogenous in order to establish better benchmarks for future studies and as such, the dataset size could increase if we opted for more mixing of the source. These properties selected for analysis are of fundamental importance in many simulation studies focusing on product and process design studies such as the critical points are used as inputs to cubic equations of state, which is commonly used in many

commercial process simulators (Frutiger et al., 2017; Mondejar et al., 2019); the enthalpy of vaporization is used in energy balance calculations in process design (Cignitti et al., 2019), the flammability property is used in process safety studies (Frutiger et al., 2016a), the lethal dosages (LD50) and concentrations (LC50) are used to evaluate the human toxicity of the compounds (Enekvist et al., 2022; Karunanithi et al., 2006), to name a few examples.

The integrity of the compounds in terms of valency is inspected using the RDKit toolbox (Landrum, 2021). Furthermore, in this work, only organic compounds featuring either of the following atoms are considered: carbon (C), nitrogen (N), oxygen (O), sulfur (S), and halogens (fluorine (F), chlorine (Cl), bromine (Br), iodine (I) and phosphorous (P)). As such, the elemental composition of the compounds is included as a constraint any compound that does not fulfill these requirements is

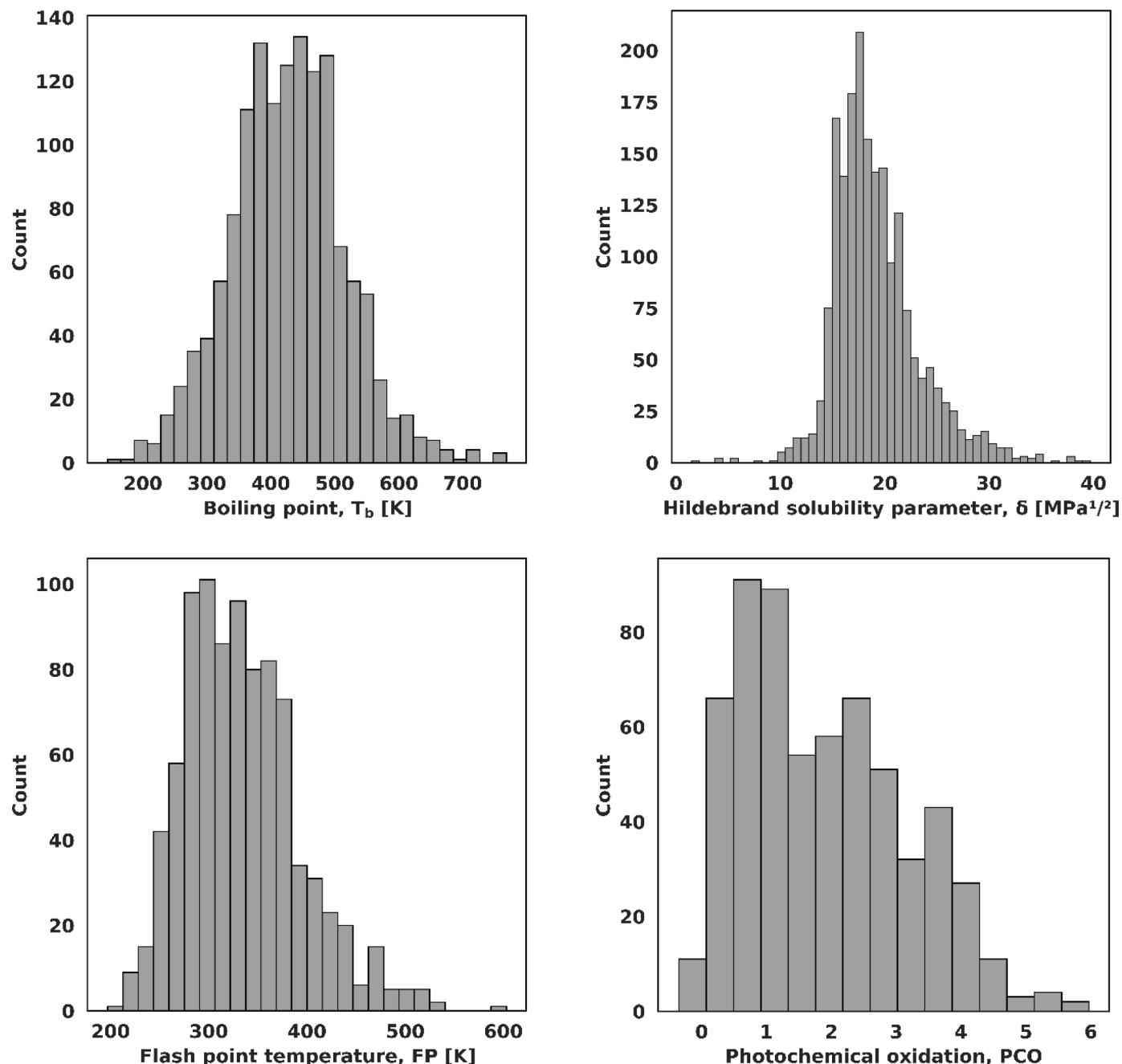
**Table 2**

Overview, definition, and source of the properties considered in this work.

Nr.	Property	Definition	Source
<b>Thermophysical properties</b>			
1	Normal boiling point ( $T_b$ )	The temperature at which a compound boils at 1 atm	Rowley et al., 2019
2	Normal melting point ( $T_m$ )	The temperature at which the melting of a compound occurs at 1 atm	Rowley et al., 2019
3	Octanol-water partition coefficient ( $\log K_{ow}$ )	The ratio of a chemical's concentration in the octanol phase to its concentration in the aqueous phase of a two-phase octanol/water system	Mansouri et al., 2016
4	Hansen solubility parameter - Dispersion ( $\delta_D$ )	The portion of the cohesive energy of a solvent attributed to dispersion	Nielsen et al., 2001
5	Hansen solubility parameter - Polar ( $\delta_P$ )	The portion of the cohesive energy of a solvent attributed to polarity	Nielsen et al., 2001
6	Hansen solubility parameter - H <sub>2</sub> -bond ( $\delta_H$ )	The portion of the cohesive energy of a solvent attributed to hydrogen bonding	Nielsen et al., 2001
7	Hildebrand solubility parameter ( $\delta$ )	A parameter that describes the total cohesive interactions that describe solubility based on the notion of "like dissolves like".	Rowley et al., 2019
8	Aqueous solubility (LogWs)	The amount of a chemical that will dissolve in liquid water and form a homogenous solution at a reference temperature	Mansouri et al., 2016
9	Acentric factor ( $\omega$ )	Conceptual number to describe the non-sphericity of molecules	Rowley et al., 2019
10	Critical temperature ( $T_c$ )	The temperature above which a gas cannot undergo liquefaction	Rowley et al., 2019
11	Critical pressure ( $P_c$ )	The minimum pressure for liquefaction of gas at the critical temperature	Rowley et al., 2019
12	Critical volume ( $V_c$ )	The Molar volume of a compound at the critical temperature and pressure	Rowley et al., 2019
13	Enthalpy of vaporization ( $H_{vap}$ )	The molar change in enthalpy associated with the isothermal transition from the liquid state to the vapor state of a compound at 298.15K	Rowley et al., 2019
14	Enthalpy of fusion ( $H_{fus}$ )	The molar change in enthalpy associated with the isothermal transition from the solid state to the liquid state of a compound at its melting point	Rowley et al., 2019
15	Enthalpy of formation ( $H_{for}$ )	The enthalpy change during the formation of the given substance in the ideal gas state at 298.15 K from elements in their standard states.	Rowley et al., 2019
16	Liquid molar volume ( $V_m$ )	The molar volume of a liquid at a reference temperature and pressure	Rowley et al., 2019
17	Ideal gas absolute entropy ( $S$ )	Absolute entropy of the ideal gas at 298.15 K and 1 bar.	Rowley et al., 2019
18	Refractive Index (RI)	The ratio of the speed of light in a vacuum to the speed of light in the substance. The incident light is the sodium D line (0.5896 microns).	Rowley et al., 2019
<b>Flammability and safety-related properties</b>			
19	Enthalpy of combustion ( $H_{com}$ )	The increase in enthalpy associated with the oxidation of a compound at 298.15 K and 1 atm to the products of the combustion process.	Rowley et al., 2019
20	Auto-ignition temperature (AiT)	The minimum temperature value for a compound to commence self-combustion in the air in the absence of an ignition source.	Rowley et al., 2019
21	Lower flammability limit (LFL)	The minimum concentration in the air will support the propagation of flames	Rowley et al., 2019
22	Upper flammability limit (UFL)	The maximum concentration in air that will support the propagation of flames	Rowley et al., 2019
23	Flashpoint (FP)	The lower temperature value, at which application of an ignition source causes the vapors of a compound in the air to ignite under specified conditions of the test (corrected to a pressure of 1 atm).	Rowley et al., 2019
<b>Environmental related properties</b>			
24	Bioconcentration factor (BCF)	The ratio of the chemical concentration in biota as a result of absorption via the respiratory surface to that in water at a steady state	Mansouri et al., 2016
25	Photochemical oxidation potential (PCO)	The result of reactions that take place between nitrogen oxides and volatile organic components exposed to UV radiation. It is expressed using a reference substance such as ethylene	Mansouri et al., 2016
26	Acid dissociation constant (pka)	A measure of the extent to which an acid dissociates in solution	Alshehri et al., 2022
27	Lethal dosage (LD <sub>50</sub> )	Amount of chemical (mass of chemical per body weight of the rat) that kills half of the rats through oral digestion	Alshehri et al., 2022
28	Lethal concentration (LC <sub>50</sub> )	Amount of chemical (in terms of liquid concentration in water) that kills half of the fathead Minnow in 96 hours.	Alshehri et al., 2022
29	Permissible exposure limit (OSHA-TWA)	A legal limit in the United States for exposure of an employee to a chemical substance or physical agent.	Alshehri et al., 2022
30	Biodegradability (BioD)	Quantification of the biodegradability of chemical compounds is described as the ratio of the biochemical oxygen demand (BOD) and chemical oxygen demand (COD)	Jhamb et al., 2020

discarded. The remaining data undergoes a set of statistical analyses in order to characterize the dataset. Summary statistics of the data can be seen in the supplementary materials. In this work, the data split is done randomly as we believe this allows fair benchmarking of model performance across diverse data sets/property space. While this might result in compound features being absent in one of the splits or structurally similar compounds being present in all splits, this can be addressed by repeating the random data splitting several times (Hwangbo et al., 2021). On the other hand, this random data splitting

does not require the definition of heuristics and rules to determine the species in each split, which might ultimately also introduce a certain bias. In the end, each data splitting method may have its use depending on the end application of the model itself: random splitting (with sufficient repetition) allows obtaining a range of model performance achievable, while heuristics/knowledge-based data splitting allows the development of a more customized model with high accuracy for certain application range of molecules. Further analysis of this aspect is however beyond the scope of this contribution. The data splitting is done at



**Fig. 5.** Property data distributions for a selected subset of properties from each class: the normal boiling point ( $T_b$ ), the Hildebrand total solubility parameter ( $\delta$ ), the flash point temperature (FP), and the photochemical oxidation (PCO).

random with 80% of data allocated for training and the remaining evenly split (10%) between validation and testing. Various splits were tested and the split resulting in the lowest validation loss was chosen. During this process, the performance metric on the test set was not calculated and not considered for determining the split. The target values are further scaled using a z-score based using the mean and the variance of the data allocated to the training to prevent data leakage. The domain of applicability of any QSPR model can be characterized through a wide range of techniques, most of which, are based on statistics and heuristics such as the William plots (Netzeva et al., 2005), principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP). In this work, we adopt a more chemically intuitive definition of the domain of applicability where we conduct a chemical diversity

analysis by classifying compounds into hydrocarbons, oxygenated, nitrogenated, chlorinated, fluorinated, brominated, iodinated, sulfonated, phosphorous-containing or silicon-containing. In case a compound can pertain to more than one group, it will be considered multifunctional. The distribution of a selected subset of the property dataset can be seen in Fig. 5. The remaining can be seen in the supplementary material and an overview of the classes present in the various dataset can be seen in Fig. 6.

### 3.2. Model development and systematic training

The models selected are the AGC and GroupGAT models described earlier and were implemented using the Deep Graph Library (DGL) using the Pytorch deep learning framework (Zheng et al., 2021). For the

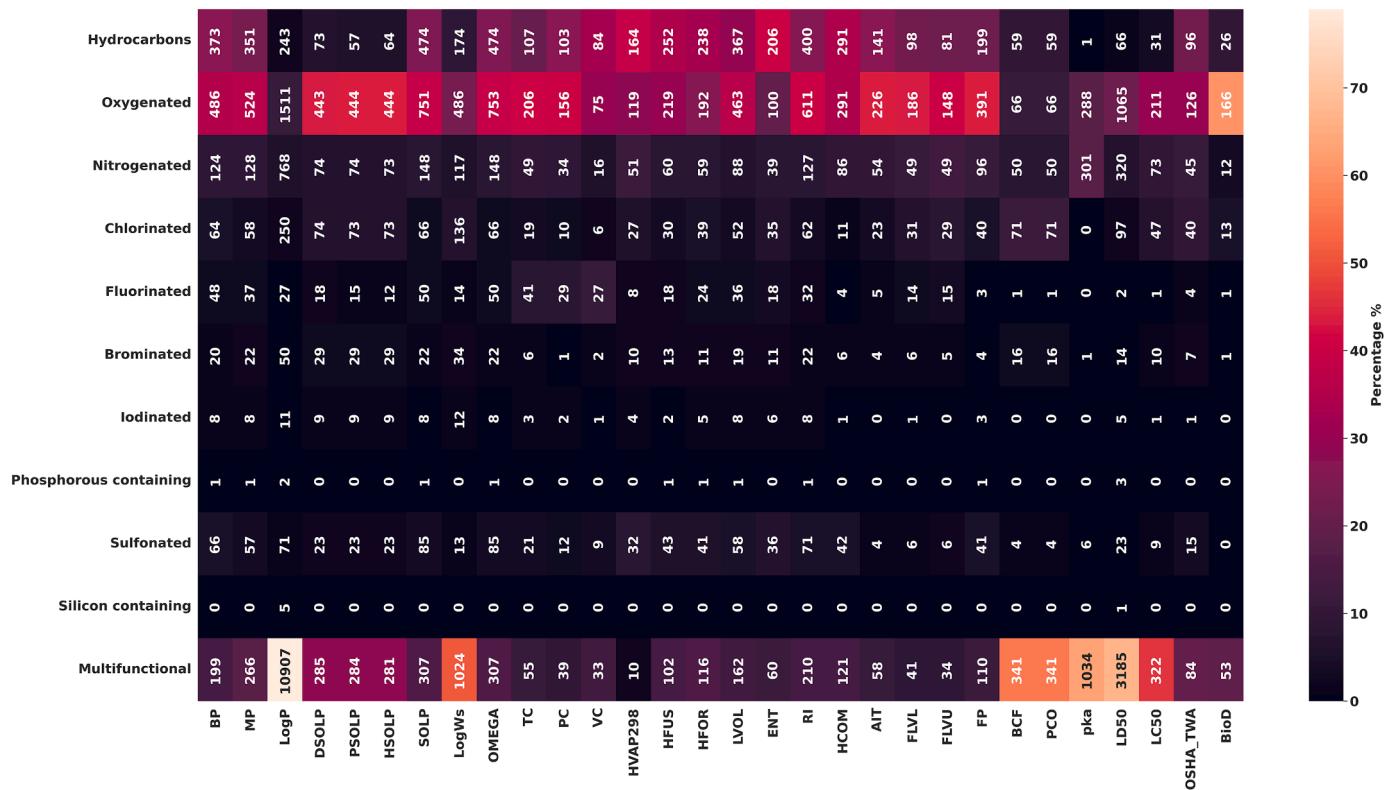


Fig. 6. Domain of applicability of the models based on chemical diversity analysis.

**Table 3**  
Hyperparameter design space for the AGC and GroupGAT.

Parameter	Range
<b>Global</b>	
Initial learning rate	[10 <sup>-5</sup> , 10 <sup>-1</sup> ]
Learning rate reduction factor	[0.4, 0.99]
Weight decay	[10 <sup>-6</sup> , 10 <sup>-1</sup> ]
Patience	30
Max number of epochs	300
<b>GroupGAT</b>	
Hidden dimension	B1, B2, B3: integer([32, 256])
Node embedding layers (T)	B1, B3: integer([1, 4]) / B2: integer([1,3])
Graph embedding layers (L)	B1, B3: integer([1, 4]) / B2: integer([1,2])
Representation dropout	B1, B2, B3, final: [0, 0.4]
MLP layers	Integer([1,4])
<b>AGC</b>	
Hidden dimension	integer([32, 256])
Node embedding layers (T)	B1: integer([1, 2]) / B2: integer([1, 4])
Graph embedding layers (L)	B1: integer([1, 2]) / B2: integer([1, 4])
Representation dropout	[0, 0.25]
MLP dropout	[0, 0.4]
MLP layers	Integer([1,4])

The ReLU activation function is used for the MLP for all properties and models. The number of hidden dimensions of a layer in the MLP corresponds to half of the previous layer.

B1, B2, and B3 correspond to the various hierachal branches in the models. The hyperparameters are not shared between each branch and thus the optimization is done for each branch individually.

training, no batching was used as the size of the dataset did not provide any issues in estimating the true gradient (batch size was set to be equal to the training size). The model training was done by the ADAM optimizer over a maximum of 500 epochs with the mean absolute error as the objective function (Kingma and Ba, 2015). In order to avoid overfitting, a series of regularization measures have been implemented such as weight decay (L2-regularization), early stopping, and dropout. The

hyperparameters of the models are determined using Bayesian optimization as implemented in the pyGPGO (Jiménez and Ginebra, 2017) toolbox. The hyperparameter design space consists in these cases of the hyperparameters of the encoder (in this case the attentive FP model) as well as some training-specific parameters as shown in Table 3. A main difference in the training procedure presented here is that the hyperparameters of each encoder block are independent (can vary between branches). This allows for a more comprehensive exploration of the hyperparameter space compared to a previous study (Aouichaoui et al., 2023).

The variance in the model prediction which can be regarded as the uncertainty is quantified using an ensemble of 40 independently trained models with different weight initialization. The choice of ensemble size was determined through a maximum-to-sum plot. Maximum to sum plot is essentially an empirical analysis of the convergence of the moments of a distribution (Cirillo and Taleb, 2020). In this case, we focused on the distribution of mean absolute error (MAE) of ensemble models as a function of N being the number of ensemble models in the sample space. This method has previously been employed in a similar study (Aouichaoui et al., 2023). At convergence, the ratio ( $R_n^p$ ) of the cumulative max ( $M_n^p$ ) over the cumulative sum ( $S_n^p$ ) of a performance metric (e.g. MAE) with a  $p$  order (up to 4) moment tends towards zero for  $n$  evaluations as shown in Eq.(12).

$$R_n^p = \frac{M_n^p}{S_n^p} \rightarrow 0 \text{ as } n \rightarrow \infty \quad (12)$$

Important to note is that the data split also affects the variance in the model performance, which can be more dominant for properties with limited size. This is however not considered in the main part of this work, where the main focus was to study the feasibility of GNN models for a wide range of properties. For future works, indeed this data splitting method can be addressed as part of the hyperparameter training to reduce any potential variance in the model performance -as shown from initial tests presented in the supplementary material this could be a

**Table 4**  
Logical and thermodynamic consistency checks.

Nr.	Properties	Condition	Reference
1	T <sub>c</sub> , T <sub>b</sub>	T <sub>c</sub> > T <sub>b</sub>	logical
2	T <sub>b</sub> , T <sub>m</sub>	T <sub>b</sub> > T <sub>m</sub>	logical
3	δ <sub>D</sub> , δ <sub>P</sub> , δ <sub>H</sub> , δ	δ = $\sqrt{\delta_D^2 + \delta_P^2 + \delta_H^2}$	Undavalli et al., 2021
4	AIT, FP	AIT > FP	logical
5	UFL, LFL	UFL > LFL	logical

relevant factor for some properties. An early version of the code is made available at: <https://github.com/gsi-lab/GC-GNN>.

### 3.3. Model evaluation

The model performance across the three data folds (training, validation, and testing) is characterized by calculating a set of performance metrics consisting of the root mean squared error (RMSE), mean absolute error (MAE), the correlation coefficient ( $R^2$ ), mean absolute percentage error (MAPE) and median absolute percentage error (MDAPE). The MDAPE is especially relevant for properties that exhibit a value in the vicinity of 0. Consider a true value of 0.1 and a potential prediction of 0.5. the MAPE would be 500% while the median will consider this as an outlier (thanks to the use of rank statistics) while still preserving the general shape of the distribution of errors. While the MDAPE can be considered the unitless performance metric which allows comparing model performance across diverse properties with different units, the MAE metric is the unit-rooted error metric and useful for specific engineering applications. These metrics are evaluated for each member of the ensemble and the mean performance value and the standard deviation in terms of percentage of the mean are reported. The errors are inspected using residual distribution plots. Some of the studied properties are interrelated in terms of thermodynamic and logical understanding. Furthermore, there are a set of established trends for some properties with increasing carbon number. It is important to investigate

these in order to establish the sanity of the models developed. [Table 4](#) provides an overview of the checks conducted.

In addition to the various sanity checks conducted, the developed models are capable of providing insights into the contribution (in terms of attention weights) of groups to the final hidden representation of the molecule. The interpretability obtained for GroupGAT and AGC are instance based and consists of inspecting and visualizing the attention weights in the graph embedding stage where the molecule is considered a virtual node with all the groups (subgraphs) linked to it. As such, the attention visualization is conducted at the encoder of the third branch for the GroupGAT and the second encoder for AGC. The attention weights are thus visualized through a color code, where the intensity of the color corresponds to a higher attention weight.

## 4. Results

### 4.1. Model performance

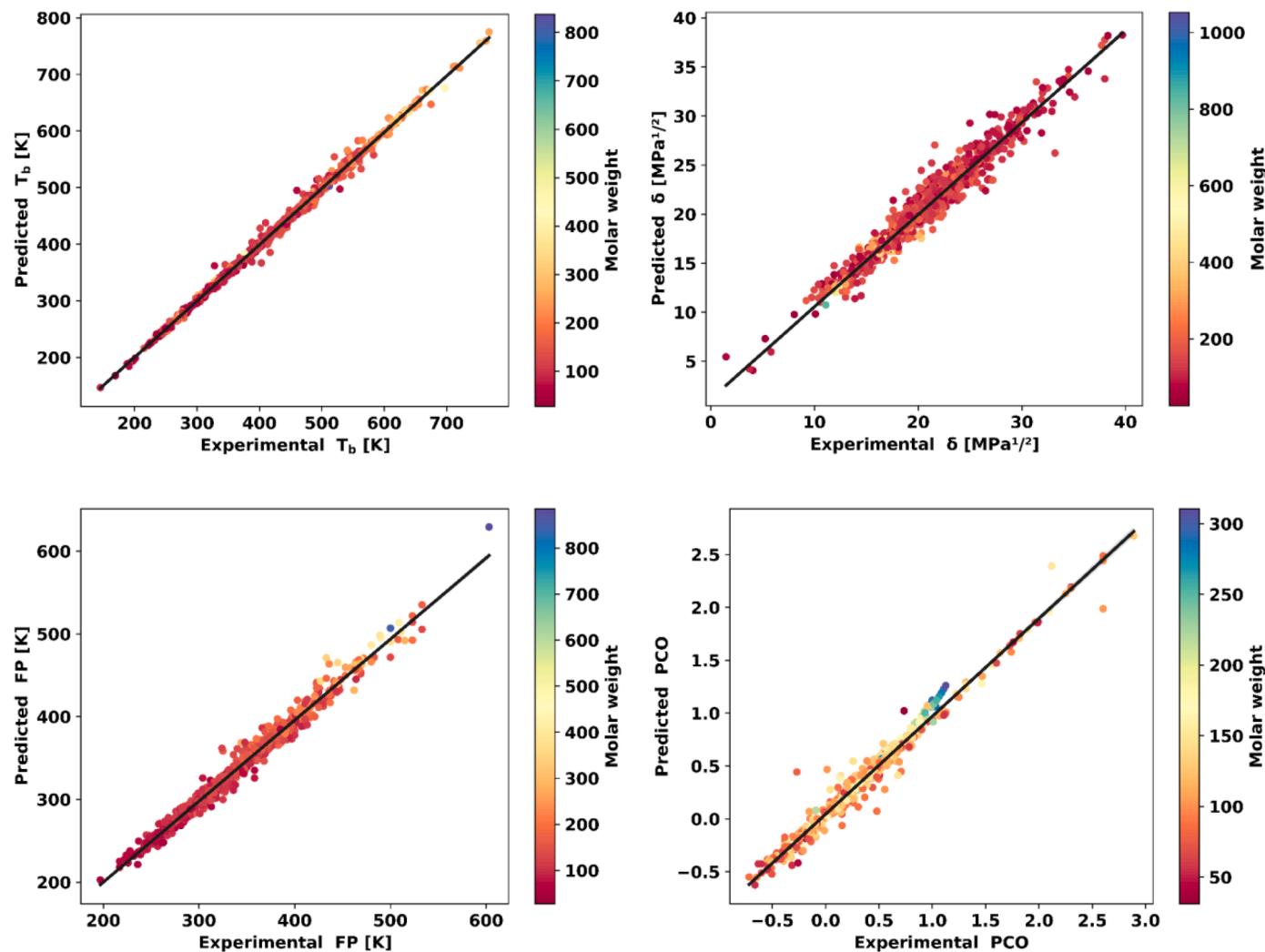
The model performance is characterized through a set of metrics and is evaluated by the model's ability to describe the complete available design space (using all data) and its ability to generalize to unseen compounds (using the test set). The shown metrics here are the  $R^2$ , MAE, and the MDAPE. The obtained results for GroupGAT can be seen in [Table 5](#), while those for AGC can be seen in the supplementary material. The uncertainty of the model performance is indicated in terms of the standard deviations reported as a percentage of the reported metrics mean. A selection parity plot can be seen in [Fig. 7](#). The remainder of the parity plots as well as the error distribution plots can be seen in the supplementary material.

#### 4.1.1. Thermophysical properties

The GroupGAT model successfully models almost all accounts of the 18 thermophysical properties considered in this study, with MDAPE below 10% for 15 of these and the remaining below 15% considering the complete dataset. Considering the test set exclusively, the MDAPE is

**Table 5**  
Performance overview of the GroupGAT on the test set and the complete dataset.

Property	No. data	$R^2$		MAE		MDAPE %	
		Test	Total	Test	Total	Test	Total
T <sub>b</sub>	1,389	0.99 $\pm$ 0.12%	0.99 $\pm$ 0.12%	6.00 $\pm$ 7.97%	3.68 $\pm$ 17.40%	0.97 $\pm$ 14.81%	0.61 $\pm$ 23.40%
T <sub>m</sub>	1,452	0.86 $\pm$ 1.64%	0.92 $\pm$ 1.20%	24.29 $\pm$ 4.96%	18.63 $\pm$ 7.67%	7.24 $\pm$ 7.81%	5.62 $\pm$ 8.71%
log K <sub>ow</sub>	13,845	0.94 $\pm$ 0.35%	0.97 $\pm$ 0.55%	0.29 $\pm$ 3.41%	0.20 $\pm$ 11.43%	10.94 $\pm$ 5.08%	7.65 $\pm$ 12.35%
δ <sub>D</sub>	1,028	0.88 $\pm$ 0.85%	0.92 $\pm$ 0.73%	0.46 $\pm$ 2.92%	0.37 $\pm$ 4.39%	1.95 $\pm$ 7.33%	1.58 $\pm$ 5.44%
δ <sub>P</sub>	1,008	0.70 $\pm$ 2.22%	0.82 $\pm$ 1.87%	1.53 $\pm$ 2.40%	1.29 $\pm$ 4.37%	17.77 $\pm$ 5.53%	14.71 $\pm$ 4.66%
δ <sub>H</sub>	1,008	0.87 $\pm$ 1.24%	0.90 $\pm$ 0.73%	1.22 $\pm$ 3.91%	1.07 $\pm$ 3.63%	12.95 $\pm$ 7.72%	10.91 $\pm$ 4.84%
δ	1,912	0.89 $\pm$ 1.67%	0.94 $\pm$ 1.63%	0.77 $\pm$ 6.28%	0.64 $\pm$ 14.57%	2.54 $\pm$ 12.45%	2.28 $\pm$ 15.34%
LogWs	2,010	0.95 $\pm$ 0.20%	0.96 $\pm$ 0.48%	0.42 $\pm$ 2.42%	0.31 $\pm$ 6.88%	15.67 $\pm$ 7.45%	9.96 $\pm$ 8.15%
ω	1,914	0.95 $\pm$ 0.55%	0.98 $\pm$ 0.33%	0.05 $\pm$ 4.77%	0.03 $\pm$ 11.72%	6.11 $\pm$ 10.41%	3.64 $\pm$ 17.27%
T <sub>c</sub>	507	0.99 $\pm$ 0.27%	0.99 $\pm$ 0.09%	8.70 $\pm$ 8.74%	4.73 $\pm$ 25.51%	0.91 $\pm$ 16.86%	0.55 $\pm$ 37.78%
P <sub>c</sub>	386	0.99 $\pm$ 0.26%	0.99 $\pm$ 0.27%	0.13 $\pm$ 11.81%	0.11 $\pm$ 13.34%	3.19 $\pm$ 16.97%	2.52 $\pm$ 16.16%
V <sub>c</sub>	253	0.99 $\pm$ 0.22%	0.99 $\pm$ 0.05%	0.01 $\pm$ 12.33%	0.01 $\pm$ 9.57%	2.39 $\pm$ 19.71%	1.31 $\pm$ 12.94%
H <sub>vap</sub>	425	0.99 $\pm$ 0.34%	0.98 $\pm$ 0.39%	1.12 $\pm$ 13.05%	1.02 $\pm$ 15.75%	2.03 $\pm$ 20.56%	1.68 $\pm$ 18.21%
H <sub>fus</sub>	740	0.97 $\pm$ 0.56%	0.95 $\pm$ 1.19%	3.55 $\pm$ 6.22%	2.48 $\pm$ 16.07%	19.17 $\pm$ 10.91%	13.12 $\pm$ 17.95%
H <sub>f</sub>	726	0.98 $\pm$ 0.28%	0.99 $\pm$ 0.06%	19.27 $\pm$ 7.91%	11.36 $\pm$ 15.89%	5.70 $\pm$ 20.72%	4.07 $\pm$ 22.31%
V <sub>m</sub>	1,254	0.99 $\pm$ 0.02%	0.99 $\pm$ 0.04%	0.00 $\pm$ 12.14%	0.00 $\pm$ 13.99%	1.11 $\pm$ 18.52%	1.14 $\pm$ 18.06%
S	511	0.99 $\pm$ 0.09%	0.95 $\pm$ 0.19%	0.06 $\pm$ 0.11%	0.04 $\pm$ 11.60%	1.06 $\pm$ 16.38%	0.90 $\pm$ 14.38%
RI	1,544	0.89 $\pm$ 5.59%	0.96 $\pm$ 0.90%	0.01 $\pm$ 8.95%	0.01 $\pm$ 16.77%	0.39 $\pm$ 7.76%	0.26 $\pm$ 18.36%
H <sub>comb</sub>	853	0.99 $\pm$ 0.02%	0.99 $\pm$ 0.02%	45.76 $\pm$ 13.06%	38.52 $\pm$ 17.18%	0.83 $\pm$ 22.82%	0.74 $\pm$ 24.61%
AiT	515	0.76 $\pm$ 5.15%	0.83 $\pm$ 6.63%	41.36 $\pm$ 9.41%	29.70 $\pm$ 27.53%	4.13 $\pm$ 16.91%	3.10 $\pm$ 33.61%
LFL	432	0.94 $\pm$ 3.31%	0.98 $\pm$ 0.58%	0.27 $\pm$ 15.45%	0.17 $\pm$ 13.16%	8.16 $\pm$ 19.12%	6.36 $\pm$ 17.77%
UFL	367	0.64 $\pm$ 8.86%	0.88 $\pm$ 3.03%	2.02 $\pm$ 14.03%	1.66 $\pm$ 33.68%	12.54 $\pm$ 27.26%	9.42 $\pm$ 48.35%
FP	888	0.97 $\pm$ 0.48%	0.99 $\pm$ 0.30%	6.83 $\pm$ 7.84%	3.75 $\pm$ 18.72%	1.45 $\pm$ 11.61%	0.81 $\pm$ 21.91%
BCF	608	0.86 $\pm$ 1.54%	0.93 $\pm$ 0.75%	0.36 $\pm$ 4.33%	0.25 $\pm$ 8.25%	19.12 $\pm$ 10.03%	11.23 $\pm$ 10.34%
PCO	608	0.78 $\pm$ 5.98%	0.95 $\pm$ 1.12%	0.11 $\pm$ 10.13%	0.06 $\pm$ 28.17%	15.55 $\pm$ 23.35%	10.74 $\pm$ 33.91%
pka	1,631	0.82 $\pm$ 2.14%	0.92 $\pm$ 1.61%	1.01 $\pm$ 4.29%	0.68 $\pm$ 10.22%	11.42 $\pm$ 9.74%	8.24 $\pm$ 12.65%
LD <sub>50</sub>	4,781	0.62 $\pm$ 1.92%	0.77 $\pm$ 3.00%	0.35 $\pm$ 7.75%	0.27 $\pm$ 5.61%	12.40 $\pm$ 3.66%	9.55 $\pm$ 6.67%
LC <sub>50</sub>	705	0.78 $\pm$ 2.04%	0.86 $\pm$ 2.17%	0.55 $\pm$ 4.96%	0.44 $\pm$ 9.54%	10.89 $\pm$ 11.27%	8.55 $\pm$ 11.43%
PEI	418	0.84 $\pm$ 6.47%	0.84 $\pm$ 9.12%	0.44 $\pm$ 19.72%	0.44 $\pm$ 25.95%	11.22 $\pm$ 32.07%	10.05 $\pm$ 29.74%
BioD	232	0.59 $\pm$ 28.31%	0.74 $\pm$ 19.87%	0.15 $\pm$ 22.64%	0.12 $\pm$ 30.25%	54.97 $\pm$ 49.22%	32.83 $\pm$ 23.37%



**Fig. 7.** Parity plot for a selected subset of properties from each class: the normal boiling point ( $T_b$ ), the Hildebrand total solubility parameter ( $\delta$ ), the flash point temperature (FP), and the photochemical oxidation (PCO). Molar weight is used as the color code.

below 10% for 13 out of the 18 accounts with the remaining being below 20%. In general, the performance of the AGC model (available in the supplementary materials) is slightly worse than the GroupGAT. This could indicate the benefits derived from the hierarchical representation learning that the GroupGAT model offers. Notably are the good performance achieved for the critical temperature, the boiling point, absolute entropy, and the refractive index with MDAPE being below 1%. Solubility-related properties such as the Hansen and Hildebrand parameters are proving to be challenging especially when it comes to unseen data. This is not a surprising observation as previous models also failed to describe such properties (Hukkerikar et al., 2012b). This is due to the complex thermodynamics and structural information that is involved e.g. the symmetry present in the molecule, the polarity, and the electronegativity.

A literature survey was conducted to identify models with a similar purpose to assess the performance of the models developed for the thermophysical properties with an emphasis on GC and GNN models (shown in Table 6). Notably, for 9 properties there are no preexisting GNN models, while for 4 of the properties GNN models have been developed as part of our previous work. This highlights that many properties of interest in the chemical engineering practice are largely overlooked when developing GNN models. The developed GroupGAT model showcases the best performance of any GNN model across all 18 thermophysical properties. Especially noteworthy are the great improvements achieved for the normal melting point (12 K in terms of MAE

to  $T_m$  compared to previously published models (Coley et al., 2017; Sivaraman et al., 2020)). The same goes for any other machine-learning or deep-learning model developed for any of the properties except for the combination of groups and Gaussian process regression (GC+GPR) (Alshehri et al., 2022). It is important to note that the cross-validation metrics for these models are not presented and no regularization techniques were reported. Considering the extremely small error metrics and confidence bounds that the models exhibit, it could indicate the presence of overfitting.

Nonetheless, compared to the GC models based on the Marrero-Gani groups, the developed models exhibit great improvements when considering the normal boiling point, solubility-related properties, critical properties, and the enthalpy of fusion. For the remaining properties, the models perform similarly except for the enthalpy of formation. This is because new groups were defined to reduce the error exhibited by the original based on the same groups (Hukkerikar et al., 2013, 2012b). The model based on the Constantinou-Gani group definitions is commonly used in computer-aided molecular design and in-silico screening of compounds due to their compatibility with the groups defined in the UNIFAC model for liquid activity coefficients (Cignitti et al., 2019; Constantinou and Gani, 1994; Hansen et al., 1991). As can be seen in Table 6, all developed models perform similarly in many accounts although the models in this work are trained on much larger data. This indicates that the currently developed models provide an accuracy level that is well comparable with the currently used

**Table 6**

Comparison of the developed property models for thermophysical properties with existing models.

Property	Reference	Type	No. data	Metric
T <sub>b</sub> (K)	Current work Alshehri et al., 2022 Aouichaoui et al., 2022b Hukkerikar et al., 2012b Qu et al., 2022 Constantinou and Gani, 1994*	GNN GC+GP GC+MLR GC+MLR GNN GC+MLR	1,389 5,276 3,510 3,510 3,850 285	MAE: 3.7 MAE: 3.9 MAE: 4.7 MAE: 6.2 MAE: 6.0 MAE: 7.71 <sup>a</sup> - 5.35 <sup>b</sup>
T <sub>m</sub> (K)	Current work Alshehri et al., 2022 Aouichaoui et al., 2022b Wytttenbach et al., 2020 Aouichaoui et al., 2023 Constantinou and Gani, 1994* Hukkerikar et al., 2012b Coley et al., 2017 Sivaraman et al., 2020	GNN GC+GPR GC+MLR QSPR GNN GC+MLR GC+MLR GNN GNN	1,452 9,249 5,183 1,800 3,035 312 5,183 3,019 33,408	MAE: 18.7 MAE: 5.3 MAE: 14.5 MAE: 16.70 MAE: 17.4 MAE: 17.40 <sup>a</sup> - 14.03 <sup>b</sup> MAE: 16.0 MAE: 26.2 MAE: 29.3
Log K <sub>ow</sub>	Current work Alshehri et al., 2022 Stefanis et al., 2004)* Zhang et al., 2022 US EPA, 2023 Aouichaoui et al., 2022b Hukkerikar et al., 2012a Tang et al., 2020 Coley et al., 2017 Xiong et al., 2020	GNN GC+GPR GC+MLR GNN QSPR GC+MLR GC+MLR GNN GNN GNN	13,845 12,193 422 10,668 10,668 12,193 12,193 4,200 282 4,200	MAE: 0.20 MAE: 0.04 MAE: 0.23 <sup>a</sup> - 0.18 <sup>b</sup> MAE: 0.23 MAE: 0.31 MAE: 0.39 MAE: 0.48 MAE: 0.41 MAE: 0.45 MAE: 0.58
δ <sub>D</sub> (MPa)	Current work Alshehri et al., 2022 Aouichaoui et al., 2022b Stefanis and Panayiotou, 2008* Hukkerikar et al., 2012b Sanchez-Lengeling et al., 2019	GNN GC+GPR GC+MLR GC+MLR GC+MLR QSPR+GPR	1,028 1,073 1,037 344 1,037 193	MAE: 0.37 MAE: 0.04 MAE: 0.37 MAE: 0.44 <sup>a</sup> - 0.41 <sup>b</sup> MAE: 0.60 MAE: 0.68
δ <sub>P</sub> (MPa)	Current work Alshehri et al., 2022 Stefanis and Panayiotou, 2008* Aouichaoui et al., 2022b Hukkerikar et al., 2012b Sanchez-Lengeling et al., 2019	GNN GC+GPR GC+MLR GC+MLR GC+MLR QSPR+GPR	1,008 1,017 350 1,017 1,017 193	MAE: 1.29 MAE: 0.12 MAE: 1.10 <sup>a</sup> - 0.86 <sup>b</sup> MAE: 1.16 MAE: 1.81 MAE: 1.93
δ <sub>H</sub> (MPa)	Current work Alshehri et al., 2022 Aouichaoui et al., 2022b Stefanis and Panayiotou, 2008* Hukkerikar et al., 2012b Sanchez-Lengeling et al., 2019	GNN GC+GPR GC+MLR GC+MLR GC+MLR QSPR+GPR	1,008 1,051 1,016 375 1,016 193	MAE: 1.07 MAE: 0.07 MAE: 0.83 MAE: 0.88 <sup>a</sup> - 0.80 <sup>b</sup> MAE: 1.28 MAE: 1.57
δ (MPa)	Current work Alshehri et al., 2022 Aouichaoui et al., 2022b Stefanis et al., 2004)* Hukkerikar et al., 2012b	GNN GC+GPR GC+MLR GC+MLR GC+MLR	1,912 1,384 1,384 1017 1,384	MAE: 0.64 MAE: 0.11 MAE: 0.72 MAE: 0.99 <sup>a</sup> - 0.90 <sup>b</sup> MAE: 1.08
Log W <sub>s</sub>	Current work Alshehri et al., 2022 Aouichaoui et al., 2023 Coley et al., 2017 Tang et al., 2020 Xiong et al., 2020 Hukkerikar et al., 2012a	GNN GC+GPR GNN GNN GNN GNN GC+MLR	2,010 2,565 1,128 1,116 1,311 1,128 4,681	MAE: 0.31 MAE: 0.13 MAE: 0.32 MAE: 0.40 MAE: 0.45 MAE: 0.51 MAE: 0.98
ω	Current work Alshehri et al., 2022 Aouichaoui et al., 2022b Hukkerikar et al., 2012b Constantinou et al., 1995*	GNN GC+GPR GC+MLR GC+MLR GC+MLR	1,914 1,723 1,723 1,723 181	MAE: 0.03 MAE: 0.01 MAE: 0.03 MAE: 0.05 MAE: 0.16 <sup>a</sup> - 0.01 <sup>b</sup>
T <sub>c</sub> (K)	Current work Alshehri et al., 2022 Aouichaoui et al., 2022b Hukkerikar et al., 2012b Constantinou and Gani, 1994)* Aouichaoui et al., 2022b Su et al., 2019	GNN GC+GPR GC+MLR GC+MLR GC+MLR GNN DNN	507 776 858 858 285 491 1,792	MAE: 4.73 MAE: 2.30 MAE: 4.57 MAE: 7.72 MAE: 9.12 <sup>a</sup> - 4.85 <sup>b</sup> MAE: 17.00 MAE: 23.00
P <sub>c</sub> (MPa)	Current work Alshehri et al., 2022 Aouichaoui et al., 2022b Aouichaoui et al., 2022a Constantinou and Gani, 1994* Hukkerikar et al., 2012b Su et al., 2019	GNN GC+GPR GC+MLR GNN GC+MLR GC+MLR DNN	386 774 852 371 269 852 1,726	MAE: 0.11 MAE: 0.04 MAE: 0.08 MAE: 0.13 MAE: 0.14 <sup>a</sup> - 0.11 <sup>b</sup> MAE: 0.14 MAE: 0.18

(continued on next page)

**Table 6 (continued)**

Property	Reference	Type	No. data	Metric	
$V_c$ ( $\text{m}^3 \text{ kmol}^{-1}$ )	Current work	GNN	253	MAE: 0.001	
	Alshehri et al., 2022	GC+GPR	773	MAE: 0.004	
	Aouichaoui et al., 2022a	GC+MLR	797	MAE: 0.006	
	Hukkerikar et al., 2012b	GC+MLR	797	MAE: 0.007	
	Su et al., 2019	DNN	1801	MAE: 0.007	
	Aouichaoui et al., 2022b	GNN	250	MAE: 0.008	
	Constantinou and Gani, 1994*	GC+MLR	251	MAE: 0.008 <sup>a</sup> - 0.006 <sup>b</sup>	
	Current work	GNN	425	MAE: 1.02	
	Alshehri et al., 2022	GC+GPR	425	MAE: 0.45	
	Aouichaoui et al., 2022a	GC+MLR	705	MAE: 0.71	
$H_{\text{vap}}^{298}$ ( $\text{kJ mol}^{-1}$ )	Hukkerikar et al., 2012b	GC+MLR	705	MAE: 1.29	
	Constantinou and Gani, 1994*	GC+MLR	225	MAE: 1.40 <sup>a</sup> - 1.11 <sup>b</sup>	
	Current work	GNN	740	MAE: 2.48	
	Alshehri et al., 2022	GC+GPR	749	MAE: 0.36	
	Aouichaoui et al., 2022a	GC+MLR	764	MAE: 1.72	
	Aouichaoui et al., 2023	GNN	730	MAE: 2.61	
	Hukkerikar et al., 2012b	GC+MLR	764	MAE: 5.03	
	Wytenbach et al., 2020	QSPR	2332	MAE: 5.20	
	Current work	GNN	726	MAE: 11.36	
	Hukkerikar et al., 2013	GC+MLR	861	MAE: 1.75	
$H_f$ ( $\text{kJ mol}^{-1}$ )	Aouichaoui et al., 2022a	GC+MLR	882	MAE: 2.92	
	Meier, 2022, 2021a, 2021b	GC+MLR	458	MAE: 1.43	
	Constantinou and Gani, 1994*	GC+MLR	373	MAE: 5.45 <sup>a</sup> - 3.71 <sup>b</sup>	
	Zheng et al., 2022	DNN	534	MAE: < 5.70	
	Alshehri et al., 2022	GC+GPR	1,059	MAE: 7.60	
	Aouichaoui et al., 2023	GNN	741	MAE: 15.50	
	Trinh et al., 2022	QSPR+SVR	1276	MAE: 16.93	
	Current work	GNN	1,254	MAE: 0.001	
	Alshehri et al., 2022	GC+GPR	1,059	MAE: 0.001	
	Aouichaoui et al., 2022a	GC+MLR	1,056	MAE: 0.001	
$V_m$ ( $\text{m}^3 \text{ kmol}^{-1}$ )	Constantinou et al., 1995*	GC+MLR	312	MAE: 0.001 <sup>a</sup> - 0.001 <sup>b</sup>	
	Hukkerikar et al., 2012b	GC+MLR	1,056	MAE: 0.003	
	Current work	GNN	511	MAE: 0.011	
	Trinh et al., 2022	QSPR+SVR	1276	MAE: 0.023	
	Benson, 1999	GC+MLR	14	MAE: 0.002	
	RI	Current work	GNN	1,544	MAPE: 0.26
	Cai et al., 2017	GC+MLR	106	MAPE: 0.73	
	Naef and Acree, 2022	GC+MLR	5988	MAPE: 0.76	
	Gharagheizi et al., 2014	GC+MLR	11918	MAPE: 0.83	

GNN: Graph neural Networks, GC: group-contribution, MLR: multi-linear regression, GPR: Gaussian process regression, SVR: support vector regression, DNN: Deep neural networks, QSPR: Quantitative structure-property relation, \*Constantinou-Gani groups based

<sup>a</sup> first-order groups only

<sup>b</sup> first and second-order groups

CAMD-related property models if not better. Furthermore, the GC-based models are trained on the entirety of the data available with no cross-validation performed which limits the models' ability to extrapolate and is dependent on the presence of the defined group in the data to assign a contribution to it.

#### 4.1.2. Flammability and safety-related properties

Among the flammability and safety-related properties, the GroupGAT model attains good performance considering the enthalpy of combustion, the flash point as well as the lower flammability limit. The results obtained for the enthalpy of combustion are comparable to those obtained using higher-order GC models (Frutiger et al., 2016b). However, important to note, is that in the current study no outlier deletion was conducted, and all data were retained for model development. Considering the MDAPE, the obtained models for the UFL, AIT, as well as FP, exceed those obtained using GC methods with outlier deletion (Frutiger et al., 2016a). However, it does remain a challenge to successfully model both the AIT and the UFL, especially since the former exhibits different limiting values for the homologous series which explains the complicated expressions developed for that purpose in GC methods (Frutiger et al., 2016a; Hukkerikar et al., 2012b). Another factor one needs to consider when modeling the AIT is the fact that measurement of this property experimentally is challenging and not trivial as the AIT is very close to the thermal decomposition temperature

of many compounds. Despite having more data, the GroupGAT outperforms the usual FP model used for CAMD screening (Frutiger et al., 2016a; Stefanis et al., 2004) and as such offers a better alternative for doing so. Similarly, the observation made in the previous analysis concerning the GC model combined with GPR is also valid for the safety-related properties (Alshehri et al., 2022). Considering the other AI or ML-based methods, the currently developed models based on GroupGAT vastly outperform these (see Table 7).

#### 4.1.3. Environmental related properties

Despite, significant improvements compared to GC-based models, the environmental-related properties remain a challenge for QSPR modeling (see Table 8). Compared to GC models, improvements have been achieved on all accounts, where in many cases the MAE has been halved (Hukkerikar et al., 2012a). It is however also important to highlight the fact that the model does struggle with some properties, especially the biodegradability dataset for which the MDAPE is 32%. In general, the biodegradability data is a complex property to model as several factors affect this measurement. Namely, the data is collected by performing a biochemical oxygen demand (BOD) test that relies on the microorganism's ability used in the test to degrade and decompose the said chemical compound under aerobic conditions monitored over several days (Metcalf & Eddy et al., 2014). Therefore, it is not surprising that high modeling accuracy is challenging to achieve due to potential

**Table 7**

Comparison of the developed property models for Flammability and safety-related properties with existing models.

Property	Reference	Type	No. data	Metric
$H_{\text{com}}$ (kJ mol <sup>-1</sup> )	Current work	GNN	853	MAE: 38.52
	Frutiger et al., 2016b	GC+MLR	794	MAE: 13.09
	Aouichaoui et al., 2023	GNN	847	MAE: 32.80
	Park et al., 2021	QSPR+MLR+NLI	1,850	MAE: 60.60
	Cao et al., 2009	GC+ANN	1,496	MAE: 155.40
	Current work	GNN	515	MAE: 29.70
	Alshehri et al., 2022	GC+GPR	571	MAE: 3.75
	Frutiger et al., 2016a	GC+MLR	513	MAE: 12.33
	Hukkerikar et al., 2012b	GC+MLR	570	MAE: 13.51
	Yang et al., 2021	DNN	480	MAE: 35.57
$\text{LFL}$ (vol-%)	Current work	GNN	432	MAE: 0.17
	Aouichaoui et al., 2021	GC+GPR	443	MAE: 0.12
	Aouichaoui et al., 2021	GC+DNN	443	MAE: 0.20
	Park et al., 2021	QSPR+MLR+NLI	1,733	MAE: 0.21
	Frutiger et al., 2016a	GC+MLR	443	MAE: 0.24
	Yang et al., 2021	DNN	443	MAE: 0.31
	Aouichaoui et al., 2021	GC+MLR	443	MAE: 0.39
	Gharagheizi, 2008	QSPR+MLR	1056	MAE: 7.68
	Current work	GNN	367	MAE: 1.66
	Yuan et al., 2019	QSPR + RF	79	MAE: 0.06
$\text{UFL}$ (vol-%)	Frutiger et al., 2016a	GC+MLR	351	MAE: 0.99
	Yang et al., 2021	DNN	329	MAE: 1.60
	Park et al., 2021	QSPR +MLR+NLI	1,711	MAE: 2.44
	Current work	GNN	888	MAE: 3.75
	Alshehri et al., 2022	GC+GPR	512	MAE: 1.10
	Frutiger et al., 2016a	GC+MLR	927	MAE: 6.77
	Park et al., 2021	QSPR +MLR+NLI	1,741	MAE: 7.31
	Yang et al., 2021	DNN	1176	MAE: 9.40
	Wen et al., 2022	GNN	1651	MAE: 9.60
	Stefanis et al., 2004*	GC+MLR	418	MAE: 11.9 <sup>a</sup> - 10.7 <sup>b</sup>
	Sun et al., 2020	GNN	10,575	MAE: 17.8

NLI: non-linear interactions

sources of variability in the measurement technique itself. However, the model developed for biodegradability in this work compares well against the GC model developed in the literature (Jhamb et al., 2020).

As a final remark, the developed GNN-based methods present in many cases the first attempt for modeling a wide range of properties and their performance exceeds in many cases currently available models based on either classical approaches such as GC models or AI/ML-based methods. However, it is important to highlight the fact that almost none of the models were trained on the same data, and as such, it makes a direct one-to-one comparison difficult to make. As such, we suggest the establishment of a standardized set of data for each property to better benchmark developed models in the literature. Such data is widely available for various quantum mechanics properties and ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties related to drug discovery in the form of e.g. Tox21 and PDBbind (Wu et al., 2018). Furthermore, it is important to consult the chemical species present for the various properties as well as the variance produced for a specific prediction in order to assess the reliability and applicability of the models for specific types of compounds. Absent features or compound types will ultimately result in high variance when assessing the uncertainty of the prediction provided by the ensemble (Ryu et al., 2019).

#### 4.2. Interpretability

We showcase one example of how the integration of functional groups and the attention mechanism provides an aspect of interpretability to the GNN model. Fig. 8 shows the attention visualization of paracetamol w.r.t to the flashpoint property. The attention weights attribute each fragment with a weight that sums to unity. The larger the attributed weight the more important the fragment is. This is visualized using a red color code where a more intense shade of red corresponds to a higher attention weight. In the shown case, the “a-C-NHCO” group has the largest attention weight followed by “aC-OH”, “a-CH” and “-CH3”

respectively. This order of importance is consistent with the contribution coefficient of the groups published in the literature (Frutiger et al., 2016a; Hukkerikar et al., 2012b). The same visualization is also obtained for the normal boiling point which also corresponds to the order of contributions published in the literature. The proposed interpretability assigns different importance for the same groups (e.g. a-CH) depending on their neighboring groups which underlies the proximity effects that the method is capable of capturing.

Important to note is that interpretability is local and instance based and as such there is no guarantee that a given group will retain the importance from one compound to another or across properties. As such, the interpretability must be evaluated on a case-by-case approach. Furthermore, we suspect that the insights are highly dependent on the goodness of fit of the model. This is because the attention coefficient is small (between 0 and 1) and as such could be very sensitive to the overall and specific performance of the model towards a given type of species in the data. It is also important to take into account the fact of collinearity, identifiability as well as the uncertainty present in the group contributions determined in the literature. As such, these could also provide a misleading conclusion (Frutiger et al., 2016b). While some counterfactual examples have previously been produced in this regard to highlight the need to assess the effect of the model accuracy and weight initialization on the interpretability produced, other techniques can be used to generate new examples where the model interpretability fails (Aouichaoui et al., 2023; Wellawatte et al., 2022).

##### 4.2.1. Sanity checks

A series of thermodynamic and logical checks have been identified as presented in the method section. The aim here is to test the validity of the models in following the thermodynamic and logical trends expected for the properties and the compounds. The results are depicted in Fig. 9.

For this, all properties of interest were evaluated on all compounds (1920 compounds in total) which had a SMILES identifier in the DIPPR database (Rowley et al., 2019). For the first consistency check, the ratio

**Table 8**

Comparison of the developed property models for environmental-related properties with existing models.

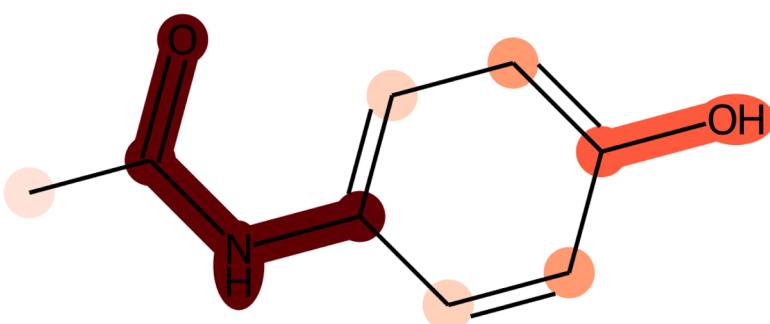
Property	Reference	Type	No. data	Metric
BCF	Current work	GNN	608	MAE: 0.25
	Alshehri et al., 2022	GC+GPR	589	MAE: 0.05
	Medina et al., 2021	GNN	473	MAE: 0.44
	Zhao et al., 2008	QSPR	473	MAE: 0.45
	Hukkerikar et al., 2012a	GC+MLR	662	MAE: 0.47
PCO	Current work	GNN	608	MAE: 0.06
	Alshehri et al., 2022	GC+GPR	606	MAE: 0.02
	Hukkerikar et al., 2012a	GC+MLR	639	MAE: 0.13
pKa	Current work	GNN	1631	MAE: 0.68
	Alshehri et al., 2022	GC+GPR	1634	MAE: 0.16
	Zhou et al., 2018	GC+ANN	1622	MAE: 0.17
	Mayr et al., 2022	GNN	208	MAE: 0.71
	Zhou et al., 2018	GC+MLR	1622	MAE: 1.18
LD50 (mol l <sup>-1</sup> )	Current work	GNN	4781	MAE: 0.27
	Alshehri et al., 2022	GC+GPR	4904	MAE: 0.03
	Hukkerikar et al., 2012a	GC+MLR	5995	MAE: 0.35
LC50 (mol l <sup>-1</sup> )	Karim et al., 2021	Ensemble	7380	MAE: 0.40
	Current work	GNN	705	MAE: 0.44
	Alshehri et al., 2022	GC+GPR	705	MAE: 0.06
	Martin and Young, 2001	GC+MLR	397	MAE: 0.26
	Hukkerikar et al., 2012a	GC+MLR	809	MAE: 0.48
OSHA_TWA (mol m <sup>-3</sup> )	Karim et al., 2021	Ensemble	823	MAE: 0.48
	Current work	GNN	418	MAE: 0.44
	Alshehri et al., 2022	GC+GPR	422	MAE: 0.06
BioD	Hukkerikar et al., 2012a	GC+MLR	425	MAE: 0.44
	Current work	GNN	232	R <sup>2</sup> : 0.74
	Jhamb et al., 2020	GC+MLR	232	R <sup>2</sup> : 0.69

of the critical temperature to the normal boiling point was calculated. Logically, this should be above unity. The sanity check attained over a 99.99% success rate. In fact, this check holds for all compounds except cyanogen. The reason behind this faulty prediction is not clear as no experimental evaluation of the critical point was found for the compound. For the second consistency check, the ratio of the normal boiling point to the melting point achieved values above unity for 99.99% of the cases. The condition did not hold for six compounds (carbamide, melamine, hexamethylenetetramine, adamantane, diamantane, and carbamoyl chloride). Most of these compounds undergo sublimation and such could be a reason for their faulty outcome in the consistency check. The third consistency check, the ratio of the auto-ignition temperature to the flash point, showcased a success rate of 98.75%, where 24 compounds did not adhere to the logical constraint set. This could be explained by the performance of the auto-ignition temperature model (MAE of 41K for the test set), which could be significant in some of the cases as this only showcases the median and as such for some extreme cases could be much larger. Another reason could be the fact that the experimental determination of these properties is rather challenging as they are

associated with a degree of safety precautions in addition to the existence of various techniques to evaluate these and as such the consistency of the available experimental data could be brought to question (See DIPPR manual) (Rowley et al., 2019). For 96% of the cases, the constraint for the flammability limits is valid. For the compounds that do not follow the logical constraints, the LFL and UFL values are very close to each other, and as such the need for accuracy increases to uphold the constraint. For the last sanity check, we also provide two independent ways of estimating the total solubility parameter (Hildebrand solubility parameter), where one is provided directly through prediction and the other through the prediction of the individual components (the Hansen solubility parameters). As can be seen in Fig. 10, there is a mismatch in some predictions, although the trend is qualitatively present. The correlation coefficient obtained through this is app. 0.7, which also showcases the discrepancies between the two calculation methods. One obvious reason is the discrepancies obtained for each of the models (Hildebrand and the three Hansen solubility parameters as seen in Table 5, especially for the test set).

An important insight to be reported and gained through the model development process is the significant impact of the activation function used. This will have a huge impact on the thermodynamic sanity of the models developed. An example would be the use of leakyReLU, which is not suitable for properties that should not exhibit negative values such as the critical pressure (pressure is always positive) and the flammability limits (percentages). Despite setting the slope of the negative linear part of the leakReLU to a very small number, for cases with limiting values, this might still result in a negative value. As such, the recommendation is to use ReLU for properties such as critical pressure, solubility parameters, and volumes. However, the use of ReLU for the final layer might result in “dead neurons” due to the negative gradients flowing through the models. This renders the weight initialization more crucial for the convergence of the training. For the LFL and UFL since they are limited between 0%-100% and a potential linear function would cause the percentage to exceed the upper limit, the recommendation is to use the sigmoid activation function. This breach of logical constraints is down to the fact the GNN models are not physics-informed compared to classical QSPR models such as the GC models where the functional transformation of the properties provides an aspect of logical constraints. As such, we suggest that it is necessary to test the limits of the developed models despite performance metrics showing excellent performance. Despite the reported interpretability w.r.t. the important functional groups in a given compound for a specific property being consistent with chemistry knowledge, this does not necessarily mean that the models are sane and thermodynamically consistent internally or across properties. Training model independently does not force the properties to be mutually consistent thermodynamically or logically. As such, multi-task learning could potentially prove to be a possible direction for future

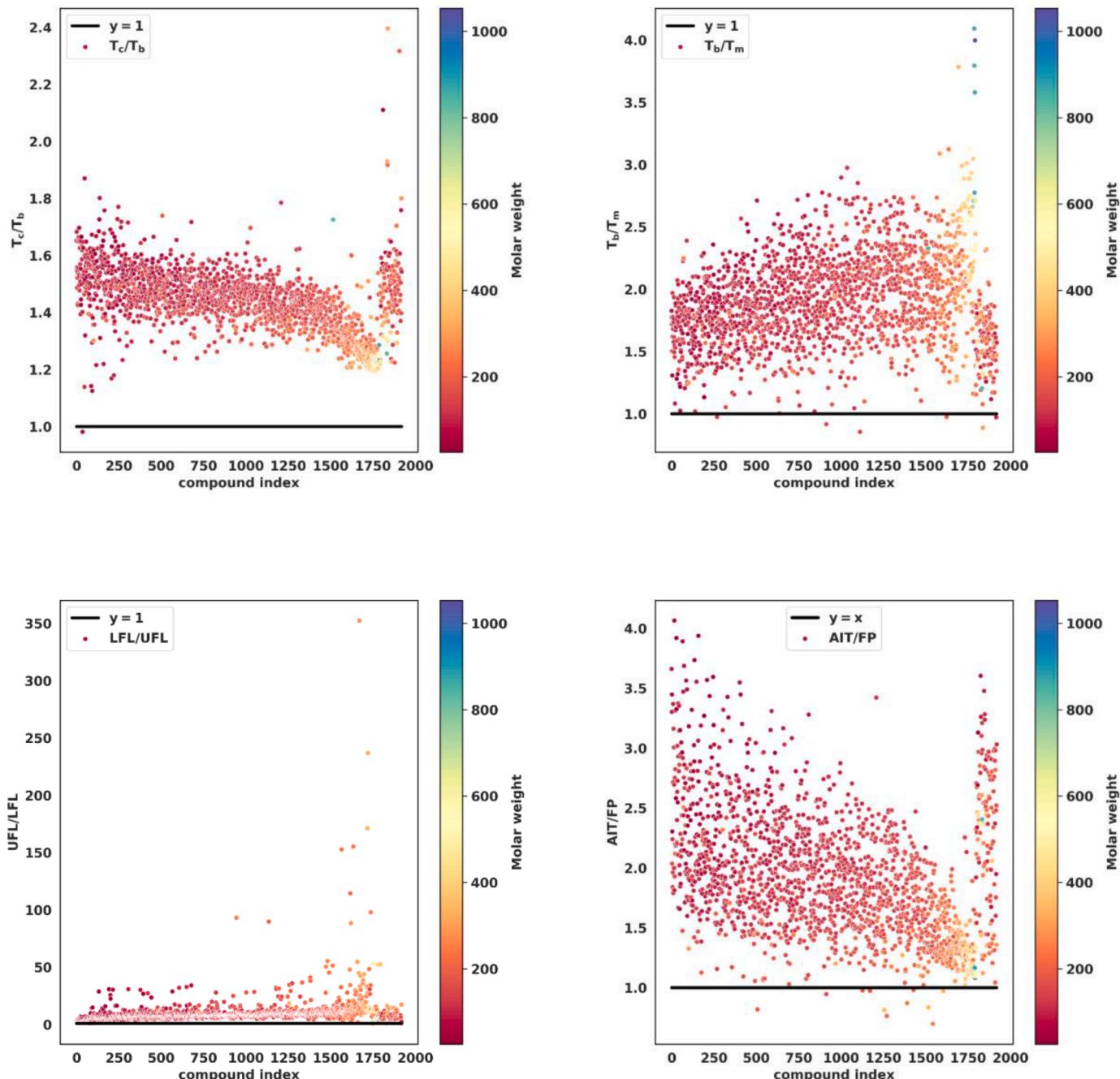
### Attention visualization



**Fig. 8.** Attention visualization (property: flashpoint, compound: paracetamol). A darker shade of red indicates higher attention weight.  
<sup>1</sup> refers to the coefficients published in (Hukkerikar et al., 2012b).

### Group significance

GC <sup>1</sup>	GroupGAT
aC-NHCO	aC-NHCO
aC-C-OH	aC-C-OH
a-CH	a-CH
-CH3	-CH3



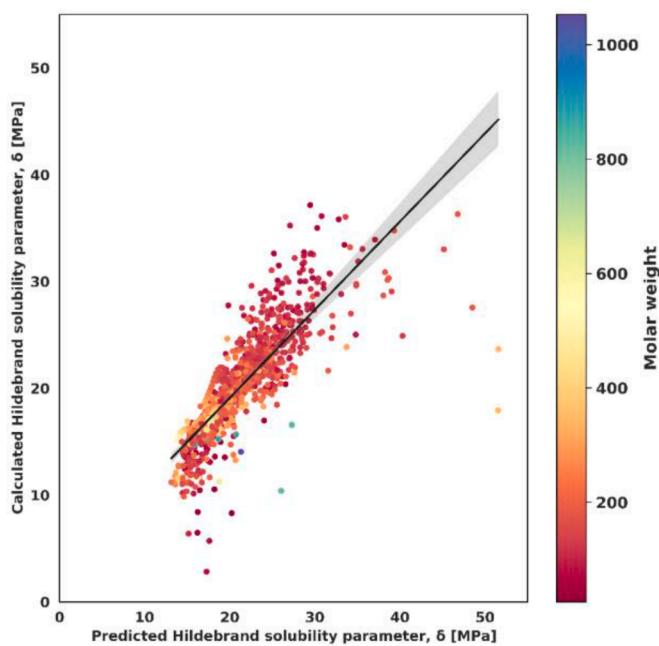
**Fig. 9.** Sanity plots: (a) the ratio of the critical temperature to the boiling point, (b) the ratio of the boiling point to the melting point, (c) the ratio of the upper and lower flammability limits, (d) the ratio of the auto-ignition temperature to the flash point temperature. Molar weight is used as the color code.

research to insure this. Multi-task learning is capable of leveraging internal correlation between target endpoints and as such could learn some logical or thermodynamic constraints such as AIT being larger than FP.

## 5. Conclusion

GNNs are versatile models that are capable of modeling a wide range of properties. In this work, we extended the range of applicability of two interpretable graph neural network models that integrates the concept of functional groups to provide an aspect of interpretability to the models to cover a total of 30 properties. The results showcase this fact by providing models with accuracy suitable for their intended purpose

which is the enable computer-aided screening based on pure compound properties. However, some properties remain a challenge, especially those related to solubility and the toxicity of a compound. The integration of functional groups showcased insights that are consistent with those obtained from GC models and might provide further insight into important substructures impacting the endpoint prediction. Despite good performance, future QSPR developments must test the limits of their models by performing logical and thermodynamic sanity checks on a benchmark set of compounds. Through this study, we showcased that cross-property validation might not hold despite each property model showcasing good performance individually and that the choice of activation function can help guide the physical and thermodynamic consistency of the models.



**Fig. 10.** parity plot showcasing the discrepancies between calculated (through Hansen parameters) Hildebrand parameters and directly predicted Hildebrand parameters.

#### CRediT authorship contribution statement

**Adem R.N. Aouichaoui:** Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Fan Fan:** Methodology, Software, Validation, Formal analysis, Writing – review & editing, Visualization. **Jens Abildskov:** Writing – review & editing, Supervision, Funding acquisition. **Gürkan Sin:** Writing – review & editing, Supervision, Funding acquisition.

#### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Data availability

The authors do not have permission to share data.

#### Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.compchemeng.2023.108291](https://doi.org/10.1016/j.compchemeng.2023.108291).

#### References

- Alshehri, A.S., Tula, A.K., You, F., Gani, R., 2022. Next generation pure component property estimation models: with and without machine learning techniques. *AIChE J.* 68.
- Aouichaoui, A.R.N., Al, R., Abildskov, J., Sin, G., 2021. Comparison of group-contribution and machine learning-based property prediction models with uncertainty quantification. In: Computer Aided Chemical Engineering. Elsevier Masson SAS, pp. 755–760.
- Aouichaoui, A.R.N., Fan, F., Mansouri, S.S., Abildskov, J., Sin, G., 2023. Combining group-contribution concept and graph neural networks toward interpretable molecular property models. *J. Chem. Inf. Model.* 63, 725–744.
- Aouichaoui, A.R.N., Mansouri, S.S., Abildskov, J., Sin, G., 2022a. Uncertainty estimation in deep learning-based property models: graph neural networks applied to the critical properties. *AIChE J.* 68.
- Aouichaoui, A.R.N., Mansouri, S.S., Abildskov, J., Sin, G., 2022b. Application of outlier treatment towards improved property prediction models. In: 32nd European Symposium on Computer Aided Process Engineering. Elsevier Masson SAS, pp. 1357–1362.
- Benson, S.W., 1999. New methods for estimating the heats of formation, heat capacities, and entropies of liquids and gases. *J. Phys. Chem. A* 103, 11481–11485.
- Cai, C., Marsh, A., Zhang, Y.H., Reid, J.P., 2017. Group contribution approach to predict the refractive index of pure organic components in ambient organic aerosol. *Environ. Sci. Technol.* 51, 9683–9690.
- Cao, H.Y., Jiang, J.C., Pan, Y., Wang, R., Cui, Y., 2009. Prediction of the net heat of combustion of organic compounds based on atom-type electrotopological state indices. *J. Loss Prev. Process Ind.* 22, 222–227.
- Tu, Chein-Hsiun, 1995. Group-contribution estimation of critical temperature with only chemical structure. *Chem. Eng. Sci.* 50, 3515–3520.
- Cignitti, S., Rodriguez-Donis, I., Abildskov, J., You, X., Shcherbakova, N., Gerbaud, V., 2019. CAMD for entrainer screening of extractive distillation process based on new thermodynamic criteria. *Chem. Eng. Res. Des.* 147, 721–733.
- Cirillo, P., Taleb, N.N., 2020. Tail risk of contagious diseases. *Nat. Phys.* 16, 606–613.
- Coley, C.W., Barzilay, R., Green, W.H., Jaakkola, T.S., Jensen, K.F., 2017. Convolutional embedding of attributed molecular graphs for physical property prediction. *J. Chem. Inf. Model.* 57, 1757–1772.
- Constantinou, L., Gani, R., 1994. New group contribution method for estimating properties of pure compounds. *AIChE J.* 40, 1697–1710.
- Constantinou, L., Gani, R., O'Connell, J.P., 1995. Estimation of the acentric factor and the liquid molar volume at 298 K using a new group contribution method. *Fluid Phase Equilib.* 103, 11–22.
- Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., Adams, R.P., 2015. Convolutional networks on graphs for learning molecular fingerprints. *Adv. Neural Inf. Process. Syst.* 2224–2232, 2015-Janua.
- Enekvist, M., Liang, X., Zhang, X., Dam-Johansen, K., Kontogeorgis, G.M., 2022. Computer-aided design and solvent selection for organic paint and coating formulations. *Prog. Org. Coat.* 162, 106568.
- Frenkel, M., 2011. Thermophysical and thermochemical properties on-demand for chemical process and product design. *Comput. Chem. Eng.* 35, 393–402.
- Frutiger, J., Bell, I., O'Connell, J.P., Kroenlein, K., Abildskov, J., Sin, G., 2017. Uncertainty assessment of equations of state with application to an organic Rankine cycle. *Mol. Phys.* 115, 1225–1244.
- Frutiger, J., Marcarie, C., Abildskov, J., Sin, G., 2016a. Group-contribution based property estimation and uncertainty analysis for flammability-related properties. *J. Hazard. Mater.* 318, 783–793.
- Frutiger, J., Marcarie, C., Abildskov, J., Sin, G., 2016b. A comprehensive methodology for development, parameter estimation, and uncertainty analysis of group contribution based property models—an application to the heat of combustion. *J. Chem. Eng. Data* 61, 602–613.
- Gani, R., 2019. Group contribution-based property estimation methods: advances and perspectives. *Curr. Opin. Chem. Eng.* 23, 184–196.
- Gasteiger, J., 2016. Chemoinformatics: achievements and challenges, a personal view. *Molecules* 21, 151.
- Gharagheizi, F., 2008. Quantitative structure–property relationship for prediction of the lower flammability limit of pure compounds. *Energy Fuels* 22, 3037–3039.
- Gharagheizi, F., Ilani-Kashkouli, P., Kamari, A., Mohammadi, A.H., Ramjugernath, D., 2014. Group contribution model for the prediction of refractive indices of organic compounds. *J. Chem. Eng. Data* 59, 1930–1943.
- Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E., 2017. Neural message passing for quantum chemistry. In: 34th International Conference on Machine Learning. ICML, pp. 2053–2070, 2017.
- Hansen, H.K., Rasmussen, P., Fredenslund, A., Schiller, M., Gmehling, J., 1991. Vapor-liquid equilibria by UNIFAC group contribution. 5. Revision and extension. *Ind. Eng. Chem. Res.* 30, 2352–2355.
- Hasebe, T., 2021. Knowledge-embedded message-passing neural networks: improving molecular property prediction with human knowledge. *ACS Omega* 6, 27955–27967.
- Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R., Coley, C.W., 2020. Uncertainty quantification using neural networks for molecular property prediction. *J. Chem. Inf. Model.* 60, 3770–3780.
- Hukkerikar, A.S., Kalakul, S., Sarup, B., Young, D.M., Sin, G., Gani, R., 2012a. Estimation of environment-related properties of chemicals for design of sustainable processes: development of group-contribution+ (GC+) property models and uncertainty analysis. *J. Chem. Inf. Model.* 52, 2823–2839.
- Hukkerikar, A.S., Meier, R.J., Sin, G., Gani, R., 2013. A method to estimate the enthalpy of formation of organic compounds with chemical accuracy. *Fluid Phase Equilib.* 348, 23–32.
- Hukkerikar, A.S., Sarup, B., Ten Kate, A., Abildskov, J., Sin, G., Gani, R., 2012b. Group-contribution+ (GC+) based estimation of properties of pure components: Improved property estimation and uncertainty analysis. *Fluid Phase Equilib.* 321, 25–43.
- Hwang, D., Yang, S., Kwon, Y., Lee, K.H., Lee, G., Jo, H., Yoon, S., Ryu, S., 2020. Comprehensive study on molecular supervised learning with graph neural networks. *J. Chem. Inf. Model.* 60, 5936–5945.
- Hwangbo, S., Al, R., Chen, X., Sin, G., 2021. Integrated model for understanding N2O emissions from wastewater treatment plants: a deep learning approach. *Environ. Sci. Technol.* 55, 2143–2151.
- Jham, S., Hospital, I., Liang, X., Pilloud, F., Piccione, P.M., Kontogeorgis, G.M., 2020. Group contribution method to estimate the biodegradability of organic compounds. *Ind. Eng. Chem. Res.* 59, 20916–20928.
- Jiménez, J., Ginebra, J., 2017. pyGPGO: bayesian optimization for python. *J. Open Source Software* 2, 431–433.

- Jiménez-Luna, J., Grisoni, F., Schneider, G., 2020. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* 2, 573–584.
- Jiménez-Luna, J., Skalic, M., Weskamp, N., Schneider, G., 2021. Coloring molecules with explainable artificial intelligence for preclinical relevance assessment. *J. Chem. Inf. Model.* 61, 1083–1094.
- Joback, K.G., Reid, R.C., 1987. Estimation of pure-component properties from group-contributions. *Chem. Eng. Commun.* 57, 233–243.
- Karim, A., Riahi, V., Mishra, A., Newton, M.A.H., Dehzangi, A., Balle, T., Sattar, A., 2021. Quantitative toxicity prediction via meta ensembling of multitask deep learning models. *ACS Omega* 6, 12306–12317.
- Karunanithi, A.T., Achenie, L.E.K., Gani, R., 2006. A computer-aided molecular design framework for crystallization solvent design. *Chem. Eng. Sci.* 61, 1247–1260.
- Katritzky, A.R., Kuanar, M., Slavov, S., Hall, C.D., Karelson, M., Kahn, I., Dobchev, D.A., 2010. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chem. Rev.* 110, 5714–5789.
- Kingma, D.P., Ba, J.L., 2015. Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, pp. 1–15.
- Klincewicz, K.M., Reid, R.C., 1984. Estimation of critical properties with group contribution methods. *AIChE J.* 30, 137–142.
- Landrum, G., 2021. RDKit: Open-source cheminformatics.
- Liu, R., Zhou, D., 2008. Using molecular fingerprint as descriptors in the QSPR study of lipophilicity. *J. Chem. Inf. Model.* 48, 542–549.
- Mansouri, K., Grulke, C.M., Richard, A.M., Judson, R.S., Williams, A.J., 2016. An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling. *SAR QSAR Environ. Res.* 27, 911–937.
- Marrero, J., Gani, R., 2001. Group-contribution based estimation of pure component properties. *Fluid Phase Equilif.* 183–184, 183–208.
- Martin, T.M., Young, D.M., 2001. Prediction of the acute toxicity (96-h LC50) of organic compounds to the fathead minnow (*Pimephales promelas*) using a group contribution method. *Chem. Res. Toxicol.* 14, 1378–1385.
- Mayr, F., Wieder, M., Wieder, O., Langer, T., 2022. Improving small molecule pKa prediction using transfer learning with graph neural networks. *Front. Chem.* 10.
- Medina, E.I.S., Linke, S., Sundmacher, K., 2021. Prediction of Bioconcentration Factors (BCF) using Graph Neural Networks. *Computer Aided Chemical Engineering*. Elsevier B.V., pp. 991–997.
- Meier, R.J., 2022. Group contribution revisited: the enthalpy of formation of organic compounds with “chemical accuracy” part III. *Appl. Chem.* 2, 213–228.
- Meier, R.J., 2021a. Group contribution revisited: the enthalpy of formation of organic compounds with “chemical accuracy”. *Chem. Eng.* 5, 24.
- Meier, R.J., 2021b. Group contribution revisited: the enthalpy of formation of organic compounds with “chemical accuracy. Part II. *Appl. Chem.* 1, 111–129.
- Metcalf & Eddy, 2014. *Wastewater engineering: treatment and resource recovery*. McGraw Hill Education.
- Mondejar, M.E., Cignitti, S., Abildskov, J., Woodley, J.M., Haglind, F., 2017. Prediction of properties of new halogenated olefins using two group contribution approaches. *Fluid. Phase Equilib.* 433, 79–96.
- Mondejar, M.E., Frutiger, J., Cignitti, S., Abildskov, J., Sin, G., Woodley, J.M., Haglind, F., 2019. Uncertainty in the prediction of the thermophysical behavior of new halogenated working fluids. *Fluid. Phase Equilib.* 485, 220–233.
- Naef, R., Acree, W.E., 2022. Revision and extension of a generally applicable group additivity method for the calculation of the refractivity and polarizability of organic molecules at 298.15 K. *Liquids* 2, 327–377.
- Netzeva, T.I., Worth, A.P., Aldenbergh, T., Benigni, R., Cronin, M.T.D., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D.W., Schultz, T.W., Stanton, D.T., Van De Sandt, J.J.M., Tong, W., Veith, G., Yang, C., 2005. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. *ATLA Altern. Lab. Anim.* 33, 155–173.
- Nielsen, T.L., Abildskov, J., Harper, P.M., Papaenconomou, I., Gani, R., 2001. The CAPEC database. *J. Chem. Eng. Data* 46, 1041–1044.
- Park, S., Bailey, J.P., Pasman, H.J., Wang, Q., El-Halwagi, M.M., 2021. Fast, easy-to-use, machine learning-developed models of prediction of flash point, heat of combustion, and lower and upper flammability limits for inherently safer design. *Comput. Chem. Eng.* 155, 107524.
- Parveen, R., Cundari, T.R., Younker, J.M., Rodriguez, G., McCullough, L., 2019. DFT and QSAR studies of ethylene polymerization by zirconocene catalysts. *ACS Catal.* 9, 9339–9349.
- Qu, C., Kearsley, A.J., Schneider, B.I., Keyrouz, W., Allison, T.C., 2022. Graph convolutional neural network applied to the prediction of normal boiling point. *J. Mol. Graph. Model.* 112.
- Reymond, J.L., 2015. The chemical space project. *Acc. Chem. Res.* 48, 722–730.
- Rogers, D., Hahn, M., 2010. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754.
- Rowley, R.I., Wilding, W.V., Oscarson, J.L., Giles, N.F., 2019. DIPPR data compilation of pure chemical properties.
- Ruddigkeit, L., Van Deursen, R., Blum, L.C., Reymond, J.L., 2012. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* 52, 2864–2875.
- Ryu, S., Kwon, Y., Kim, W.Y., 2019. A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification. *Chem. Sci.* 10, 8438–8446.
- Sanchez-Lengeling, B., Roch, L.M., Perea, J.D., Langner, S., Brabec, C.J., Aspuru-Guzik, A., 2019. A bayesian approach to predict solubility parameters. *Adv. Theory Simul.* 2.
- Scalia, G., Grambow, C.A., Pernici, B., Li, Y.P., Green, W.H., 2020. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *J. Chem. Inf. Model.* 60, 2697–2717.
- Schweidtmann, A.M., Rittig, J.G., König, A., Grohe, M., Mitsos, A., Dahmen, M., 2020. Graph neural networks for prediction of fuel ignition quality. *Energy Fuels* 34, 11395–11407.
- Sivaraman, G., Jackson, N.E., Sanchez-Lengeling, B., Vázquez-Mayagoitia, Á., Aspuru-Guzik, A., Vishwanath, V., de Pablo, J.J., 2020. A machine learning workflow for molecular analysis: application to melting points. *Mach. Learn. Sci. Technol.* 1, 025015.
- Stefanis, E., Constantinou, L., Panayiotou, C., 2004. A group-contribution method for predicting pure component properties of biochemical and safety interest. *Ind. Eng. Chem. Res.* 43, 6253–6261.
- Stefanis, E., Panayiotou, C., 2008. Prediction of hansen solubility parameters with a new group-contribution method. *Int. J. Thermophys.* 29, 568–585.
- Su, Y., Wang, Z., Jin, S., Shen, W., Ren, J., Eden, M.R., 2019. An architecture of deep learning in QSPR modeling for the prediction of critical properties using molecular signatures. *AIChE J.* 65, 1–11.
- Sun, X., Krakauer, N.J., Politowicz, A., Chen, W.T., Li, Q., Li, Z., Shao, X., Sunaryo, A., Shen, M., Wang, J., Morgan, D., 2020. Assessing graph-based deep learning models for predicting flash point. *Mol. Inform.* 39.
- Tang, B., Kramer, S.T., Fang, M., Qiu, Y., Wu, Z., Xu, D., 2020. A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *J. Cheminform.* 12, 15.
- Trinh, C., Meimaroglou, D., Lasala, S., Herbinet, O., 2022. Machine Learning for the prediction of the thermochemical properties (enthalpy and entropy of formation) of a molecule from its molecular descriptors. pp. 1471–1476.
- Undavalli, V.K., Ling, C., Khandelwal, B., 2021. Impact of alternative fuels and properties on elastomer compatibility. *Aviation Fuels*. Elsevier, pp. 113–132.
- US EPA, 2023. *Estimation Programs Interface SuiteTM for Microsoft® Windows*.
- Van Speybroeck, V., Gani, R., Meier, R.J., 2010. The calculation of thermodynamic properties of molecules. *Chem. Soc. Rev.* 39, 1764–1779.
- Wellawatte, G.P., Seshadri, A., White, A.D., 2022. Model agnostic generation of counterfactual explanations for molecules. *Chem. Sci.* 13, 3697–3705.
- Wen, H., Su, Y., Wang, Z., Jin, S., Ren, J., Shen, W., Eden, M., 2022. A systematic modeling methodology of deep neural network-based structure-property relationship for rapid and reliable prediction on flashpoints. *AIChE J.* 68.
- Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., Langer, T., 2020. A compact review of molecular property prediction with graph neural networks. *Drug Discov. Today Technol.* 37, 1–12.
- Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S., Leswing, K., Pande, V., 2018. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* 9, 513–530.
- Wyttnebach, N., Niederquell, A., Kuentz, M., 2020. Machine estimation of drug melting properties and influence on solubility prediction. *Mol. Pharm.* 17, 2660–2671.
- Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., Zheng, M., 2020. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* 63, 8749–8760.
- Yang, A., Su, Y., Wang, Z., Jin, S., Ren, J., Zhang, X., Shen, W., Clark, J.H., 2021. A multi-task deep learning neural network for predicting flammability-related properties from molecular structures. *Green Chem.* 23, 4451–4465.
- Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., Barzilay, R., 2019. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* 59, 3370–3388.
- Yuan, S., Jiao, Z., Quddus, N., Kwon, J.S.-I., Mashuga, C.V., 2019. Developing quantitative structure–property relationship models to predict the upper flammability limit using machine learning. *Ind. Eng. Chem. Res.* 58, 3531–3537.
- Zhang, J., Wang, Q., Su, Y., Jin, S., Ren, J., Eden, M., Shen, W., 2022. An accurate and interpretable deep learning model for environmental properties prediction using hybrid molecular representations. *AIChE J.* 68.
- Zhang, Z., Guan, J., Zhou, S., 2021. FraGAT: a fragment-oriented multi-scale graph attention model for molecular property prediction. *Bioinformatics* 37, 2981–2987.
- Zhao, C., Boriani, E., Chana, A., Roncaglioni, A., Benfenati, E., 2008. A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). *Chemosphere* 73, 1701–1707.
- Zheng, D., Wang, M., Gan, Q., Song, X., Zhang, Z., Karypis, G., 2021. Scalable graph neural networks with deep graph library. In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM, pp. 1141–1142.
- Zheng, P., Yang, W., Wu, W., Isayev, O., Dral, P.O., 2022. Toward chemical accuracy in predicting enthalpies of formation with general-purpose data-driven methods. *J. Phys. Chem. Lett.* 13, 3479–3491.
- Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M., 2018. Graph neural networks: a review of methods and applications 1–22.
- Zhou, T., Jhamb, S., Liang, X., Sundmacher, K., Gani, R., 2018. Prediction of acid dissociation constants of organic compounds using group contribution methods. *Chem. Eng. Sci.* 183, 95–105.