Environmental impacts prediction using graph neural networks on molecular graphs

Gao, Qinghe; Balhorn, Lukas Schulze; Laera, Alessandro; Meys, Raoul; Goßen, Jonas; Weber, Jana M.; Wernet, Gregor; Schweidtmann, Artur M.

DOI
10.1016/j.compchemeng.2025.109362

Publication date
2025

Document Version
Final published version

Published in
Computers & Chemical Engineering

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/cace

Environmental impacts prediction using graph neural networks on molecular graphs

Qinghe Gao [a], Lukas Schulze Balhorn [a], Alessandro Laera [a], Raoul Meys [b], Jonas Goßen [b], Jana M. Weber [c], Gregor Wernet [d], Artur M. Schweidtmann [a],*

[a] Process Intelligence Research Group, Department of Chemical Engineering, Delft University of Technology, Van der Maasweg 9, Delft 2629 HZ, The Netherlands
[b] Carbon Minds GmbH, Eupener Str. 165, Cologne 50933, Germany
[c] Pattern Recognition and Bioinformatics, Department of Intelligent Systems, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands
[d] Düsseldorf, Germany

ARTICLE INFO

Keywords:
Artificial intelligence
Graph convolutional neural networks
Life cycle assessment

ABSTRACT

The chemical industry needs to undergo a significant transformation towards more sustainable and circular production systems. To guide this transformation, estimating the environmental impacts of chemical production at early product screening or development stages is highly desirable. This study leverages the molecular structure of the process products with graph neural networks (GNNs) for early-stage environmental impact approximation of chemical processes. Specifically, we use end-to-end GNN models to predict fifteen environmental impact categories, utilizing a CarbonMinds dataset of 51,905 processes producing 791 molecules produced in 91 countries, augmented with country-specific energy mix data. Our analysis begins with a comparison of Quantitative Structure-Property Relationship (QSPR) and GNN models for the climate change impact category. Specifically, we develop three different GNN models: (i) GNN with only molecular structure, (ii) GNN with molecular structure and additional geographical features, and (iii) GNN with molecular structure and additional energy mix features. The results indicate that the three GNN models show an improvement over the QSPR models. Furthermore, benchmarking our GNN models against the existing literature in the climate change impact category reveals that our models perform comparably. We then extend our approach by developing both single- and multi-task GNN models to predict all fifteen impact categories. The findings indicate that multi-task learning can improve model performance in complex environmental impact predictions compared to single-task GNNs. Therefore, we recommend using a multi-task GNN for predicting multiple impact categories, with single-task models applied to fine-tune performance on underperforming categories. Although our proposed approach shows improvements over previous models, the prediction of environmental impacts solely based on molecular information remains a rough approximation.

1. Introduction

The chemical industry is currently in need of a significant paradigm shift towards more sustainable and circular processes. This transition requires reducing the environmental impact associated with existing chemical production processes. In this context, it is critical to recognize that modifications in the design of chemical products and processes at early technology readiness levels (TRLs) (Buchner et al., 2019) are most feasible and cost-effective. The role of computational prescreening techniques (Weber et al., 2021, 2022; Ulonska et al., 2016; Minten et al., 2024; Preuss et al., 2024; Blanco et al., 2024) is increasingly important, as these methods offer a systematic approach to assessing the environmental impact of chemical production processes and their viable alternatives during the early stage of development.

The state-of-the-art methodology for determining the environmental impact of processes and products is Life Cycle Assessment (LCA). LCA is an ISO-normed methodology for assessing the environmental impact of a product or process (Standard, 2006). The core principle of LCA is to trace the principal stages and processes in a product's lifecycle, from raw material extraction through manufacturing, usage, and recycling, to eventual disposal, pinpointing and calculating the environmental impacts at each stage. Because of its holistic nature, LCA requires detailed information about the processes and products across all phases of the life cycle inventory—often from cradle-to-gate

* Corresponding author.
E-mail address: a.schweidtmann@tudelft.nl (A.M. Schweidtmann).

https://doi.org/10.1016/j.compchemeng.2025.109362
Received 19 March 2025; Received in revised form 1 August 2025; Accepted 21 August 2025
Available online 28 August 2025

and beyond. Specifically, in this study, we focus exclusively on the cradle-to-gate phase. The significant data demands, spanning multiple levels of the supply chain of products, often present a major obstacle to conducting a thorough LCA. This challenge is particularly evident in the chemical industry and its supply chains, which are characterized by complex process networks with interdependent exchanges of energy, mass, and water, as well as the use of diverse intermediate products and feedstocks. Additionally, the high data demand for a full LCA study often hinders early-stage screening studies on multiple product and/or process alternatives. Therefore, to address these data gaps, LCA practitioners have deployed simplified approaches (Heidari et al., 2019; Nakamura and Nansai, 2016; Mattila, 2017; Yang et al., 2017), for example, streamlined LCA (Heidari et al., 2019), to overcome the data scarcity problem.

Streamlined LCA uses simplified models, databases, and assumptions to estimate impacts without extensive data collection and analysis as in full LCAs. The streamlined LCA methods consist of two parts: (1) the input data and (2) the analysis model. The data is usually collected from key aspects such as the molecular structure and it can be characterized by the technology readiness level as defined in Buchner et al. (2019). Several databases (Martínez-Rocamora et al., 2016) provide access to LCA impact category data, including Ecoinvent (Wernet et al., 2016) and the European Reference Life Cycle Database (ELCD) (Fazio et al., 2015). These resources offer valuable data for estimating environmental impacts across various categories. The analysis model maps the input data to output environment impact categories through a predictive model. Molecular structure models (MSMs) have been widely used in assessing chemical products (Kleinekorte et al., 2020; Parvatker and Eckelman, 2018). The rationale of MSMs is that the molecular structure encodes key information that has a direct impact on (i) the intricacy of their production process and (ii) their potential for danger and degradation at the end of life, which in turn affects aspects of their life cycle environmental impact. Unlike the prediction of conventional physicochemical properties (Alshehri et al., 2021; Chen et al., 2023), such as boiling point or solubility, predictions of LCIA impact scores aim to assess properties of an entire industrial supply chain, including aspects of production processes and specific process conditions. This complexity introduces greater variability and noise into the data, making accurate predictions more challenging. Various modeling approaches have been explored for predicting LCA impact categories within MSMs. These approaches can be broadly classified based on molecular representation methods and the techniques used to map these representations to corresponding impact category values. Molecular representations generally fall into three categories: molecular descriptors, molecular graphs, and molecular strings. Techniques for mapping these representations to impact values range from simple linear regressions to advanced machine learning models, including large language models, with training parameters varying from a few thousand to billions.

Quantitative structure–property relationship (QSPR) models (Hammett, 1935), utilizing molecular descriptors, as a type of MSMs have been successfully deployed to predict LCA impact categories in the previous literature. QSPRs characterize molecules with various structural, chemical, physical, and biological features, group contributions (Alshehri et al., 2021; Gani, 2019), referred to as molecular descriptors that are then mapped to a property of interest by a linear or nonlinear model, such as support vector machines, decision trees, random forests, and multilayer perceptrons (MLPs). For example, Wernet et al. (2008, 2009) first utilized both linear regression models and MLPs for molecular structures to predict several impact categories such as cumulative energy demand (CED), global warming potential (GWP), the biological and chemical oxygen demands (BOD and COD), the total organic carbon (TOC), the Ecoindicator'99(H/H): human health (HH), ecosystem quality (EQ), resources (R) and total (T) scores, achieving R^2 values from 0.41 to 0.69 across respective categories. Later, Song et al. (2017) also utilized molecular descriptors and MLPs to predict six impact categories, including CED, GWP, HH, EQ, T, and acidification (AC). The R^2 values ranged from 0.45 to 0.87. Sun et al. (2022) further utilized four common machine learning algorithms: support vector machines, decision trees, random forests, and MLPs to predict six impact categories, and the R^2 values ranged from 0.73 to 0.86. Additionally, Baxevanidis et al. (2021) deployed six different linear and nonlinear regression models to predict CED, GWP, and Ecoindicator'99. The R^2 values ranged from 0.08 to 0.26.

While predicting environmental impacts directly from molecular descriptors is an attractive approach, it presents challenges in achieving consistently accurate and generalizable models. Specifically, when using only product features such as molecular structure, these models most likely struggle to accurately predict aspects of the impact categories that depend on process-level decisions. This limitation arises because the models do not account for the production methods (e.g., derived from crude oil versus electrolyzers powered by green energy), which significantly influences environmental outcomes. To mitigate this problem, Calvo-Serrano et al. (2017, 2018b,a) added thermodynamic features with molecular descriptors as input, formulated into Mixed Integer Nonlinear Programming (MINLP) problems and achieve relative errors in the range 20%–44% for CED, GWP, COD, BOD, TOC, T, HH, EQ, and depletion of resources (Res). Also, Kleinekorte et al. (2019) additionally included process descriptors in MLPs, together with molecular descriptors (information commonly available at TRL 2 of process/product development), to predict seventeen impact categories with R^2 between 0 and 0.66. Karka et al. (2022) incorporated MLPs and decision trees to process molecular descriptors, process, and energy-related parameters, and predict twenty-three LCA metrics with R^2 between 0.50 and 0.88.

Recently, end-to-end learning, operating mainly on molecular graphs or molecular strings, has shown promising results in property prediction and molecule generation. In contrast to traditional QSPR models, the aim of end-to-end learning is to train a model to automatically learn from input data, such as molecular graphs or molecular strings, to final output predictions without the need for manual feature engineering or intermediate steps. Molecular strings with large language models (Sultan et al., 2025, 2024) for molecular property prediction are also a promising research direction. However, in this work, we mainly focus on the molecular graph perspective. In particular, graph neural networks (GNNs) have shown promising results in these tasks (Zhang et al., 2020; Buterez et al., 2024; Wang et al., 2024; Gao et al., 2024; Trivedi et al., 2024). GNNs can learn to predict properties from molecular graphs where nodes present as atoms and edges correspond to the bonds, which constitutes an end-to-end supervised learning setup without manual feature selection (Rittig et al., 2022). GNNs reached state-of-the-art accuracy in predicting molecular properties, for regression tasks, like electron affinity (Gao et al., 2024), excitation energies (Trivedi et al., 2024), and classification tasks, such as toxicity of molecules (Pope et al., 2018). In the context of environmental impact category prediction, Kleinekorte et al. (2023) recently utilized a GNNs-based encoder–decoder network for processing molecular graphs. In addition, they derived process descriptors from the stoichiometric reaction equation and combine molecular and process descriptors in a Gaussian process to predict the global warming impact (GWI). The model obtained an R^2 value of 0.53, demonstrating the potential of GNNs for learning molecular descriptors for environmental impact categories prediction tasks. Additionally, Zhang et al. (2024) proposed FineChem 2, which is based on a GNN and a transformer framework, to assess the product carbon footprints (PCF), the same as GWI and GWP, of chemicals.

With those advancements in utilizing GNNs in environmental impact prediction, there are still some remaining research gaps. Primarily, while GNNs have shown good results in predicting GWI (GWP or PCF), the results of other crucial impact categories like AC and HH still need to be explored. Furthermore, current GNN prediction models are made for individual impact categories separately, namely, single-task

GNN. However, multi-task GNNs have outperformed single-task GNNs on several molecular prediction tasks (Liu et al., 2022; Ramsundar et al., 2015; Chen et al., 2025; Hu et al., 2025). Multitask learning is a machine learning approach where a model is trained simultaneously on multiple related tasks. A shared representation is first learned across multiple tasks, capturing generalizable information. This shared representation serves as a foundation, upon which individual hidden layers are utilized to extract task-specific features. Consequently, a multi-task GNN has the potential to enhance predictive accuracy in the domain of environmental impact assessment by effectively integrating and learning from multiple related tasks. Furthermore, the current literature lacks a comparative analysis between GNNs and QSPR models. To guide future research effectively, particularly in the context of model selection, a comparison between these methodologies is necessary.

We propose to utilize GNNs to estimate the environmental impact categories. First, our contribution is applying GNN-based models to predict fifteen distinct midpoint environmental impact categories—a broader scope than most previous studies, which typically focus on less than three categories. Second, we utilize a comprehensive dataset in a total of 51,905 processes comprising 761 molecules, enriched with geographic and energy mix information from 91 different countries. Additionally, those impact scores are calculated and collected by the Environmental Footprint (EF) framework developed by Carbon Minds independently outside of this work, which is based on the Environmental Footprint v3.0 (European Commission, 2021) and is reported using a cradle-to-gate system boundary. Third, three different GNN models are proposed: 1. Molecular features only (GNN-M); 2. Molecular features and one-hot encoded country features (GNN-C); 3. Molecular features and energy mix features (GNN-E). Additionally, we establish a comparative framework by developing QSPR models using molecular descriptors from Song et al. (2017) as a benchmark to evaluate the performance of GNN models, with a specific focus on the climate change (CC) category because it is the most extensively studied and standardized impact category. CC values used in our study correspond to GWI — the total life-cycle climate impact per kg of product, expressed in kg CO_2-eq. Fourth, we develop single-task GNN-C and GNN-E to predict all fifteen impact categories separately. Additionally, to further improve prediction accuracy and data efficiency, we propose multi-task GNN-C and GNN-E. These models share message-passing layers across tasks while using separate MLP heads for each impact category, allowing shared learning across categories. Finally, we publish the trained models and code to support reusability and future research: https://github.com/process-intelligence-research/LCA_GNNs.git. Together, these contributions could provide data-driven tools for early-stage sustainability assessment.

2. Preliminaries

In this section, we briefly introduce the fundamentals of QSPR and GNN modeling.

2.1. Quantitative structure–property relationship

QSPR is a method used in chem- and bioinformatics to establish relationships between the molecular structure of compounds and their physical or chemical properties, consisting of two parts: (i) molecule description and (ii) property regression. In step (i), the structure of a molecule m is represented by selected molecular descriptors d_i. Mathematically, this can be described as $D : m \mapsto \mathbf{d}$ with $\mathbf{d} = [d_1, d_2, \ldots, d_n]^T$. In the step (ii), the function $F(\cdot)$ predicts the target property $\hat{p}$ based on the descriptors d_i (Katritzky et al., 2010) as $\hat{p} = F(\mathbf{d}) = F(D(m))$. The function $F(\cdot)$ can be linear or nonlinear (multivariate) regression methods, such as random forests, support vector machines, or MLPs. The choice of molecular descriptors depends on the molecules and properties of interest. Two frequently used descriptor types are structural groups and molecular fingerprints. Group contribution methods (GCMs) (Benson et al., 1969; Joback and Reid, 1987) decompose the molecular structure into predefined structural groups, e.g., >CH– (non-ring), or =CH– (ring), with the frequency of these groups serving as molecular descriptors. Alternatively, molecular fingerprints denote the molecules as a vector in which the molecular structure such as the count of substructures, similar to the GCM, or geometric distances between two atoms or structural groups (Todeschini and Consonni, 2000) are stored. A critical challenge in QSPR modeling is selecting appropriate descriptors for specific property prediction tasks, Cherkasov et al. (2014) as effective property prediction relies on the use of the most informative descriptors. However, informative descriptors are not always apparent prior to model development. Therefore, to mitigate this challenge, principal component analysis (PCA) has been utilized to extract informative features from molecular descriptors. PCA is a dimensionality reduction technique that transforms correlated variables into a smaller set of uncorrelated components while preserving as much variance as possible. Song et al. (2017) applied PCA to extract 60 features from a total of 3839 molecular descriptors, which were subsequently used to predict six impact categories. However, even with PCA, the initial selection of molecular descriptors remains a necessary first step, requiring manual effort. As a result, the development of predictive QSPR models often still relies on expert intuition (Katritzky et al., 2010; Cherkasov et al., 2014).

2.2. Graph neural networks

GNNs are a deep learning technique designed to learn properties directly from graph representations. Molecules are represented as graphs where atoms are denoted as nodes m, also referred to as vertices $v \in V$, and bonds are represented as edges $e_{vw} \in E$ connecting two nodes v and w. Importantly, a feature vector is typically assigned to each node and each edge, which includes specific information for the atom or bond, e.g., the atom type or the bond type. Mathematically, we denote the feature vector of a node v with $\mathbf{f}^v \in F^V$ and the feature vector of an edge e_{vw} with $\mathbf{f}^{e_{vw}} \in F^E$.

The framework of GNNs consists of two phases: The message passing phase and the readout phase as shown in Fig. 1. The message-passing phase is responsible for extracting structural information from the molecular graph through graph convolutions. Within the graph convolution process, each node exchanges information as messages that are passed along the edges to their neighboring nodes. These messages are based on the feature vectors of the corresponding neighboring nodes and edges. For example, at the beginning of graph convolution layer l, the initiated state of a node with the corresponding feature vector is denoted as $\mathbf{h}_v^0 = f^v$ with $l = 0$. Then, the molecular graph traverses the individual graph convolutional layers $l \in \{1, 2, \ldots, L\}$. Within a graph convolutional layer l, the hidden state of each node is updated based on the hidden states of the corresponding neighborhood nodes and the features of the associated edges. Thereby, with a total of L graph convolutional layers, each node receives direct information about an environment with a radius of L nodes. The updating process can be written as

$$\mathbf{m}_v^l = A_l\{M_l(\mathbf{h}_w^{l-1}, \mathbf{f}^{e_{vw}}) \mid w \in N(v)\} \quad (1)$$

$$\mathbf{h}_v^l = U_l(\mathbf{h}_v^{l-1}, \mathbf{m}_v^l) \quad (2)$$

The function M_l maps the *message* from one of its neighbor nodes w to node v, with $w \in N(v)$. The message is a function with two parts: (i) the hidden states of neighbor nodes w: $\mathbf{h}_w^{l-1}$, and (ii) the edge e_{vw} with the associated edge feature vector $\mathbf{f}^{e_{vw}}$. Furthermore, the aggregation function A_l aggregates the messages from all the neighbors of node v in $\mathbf{m}_v^l$. Then, the neighbor messages $\mathbf{m}_v^l$ are finally combined with the preceding hidden state of the node v itself, $\mathbf{h}_v^{l-1}$, in the update function U_l to update the hidden state of node v. Consequently, the hidden state of a node in layer l depends on its previous hidden state, the previous hidden states of its neighboring nodes, and the associated edges.

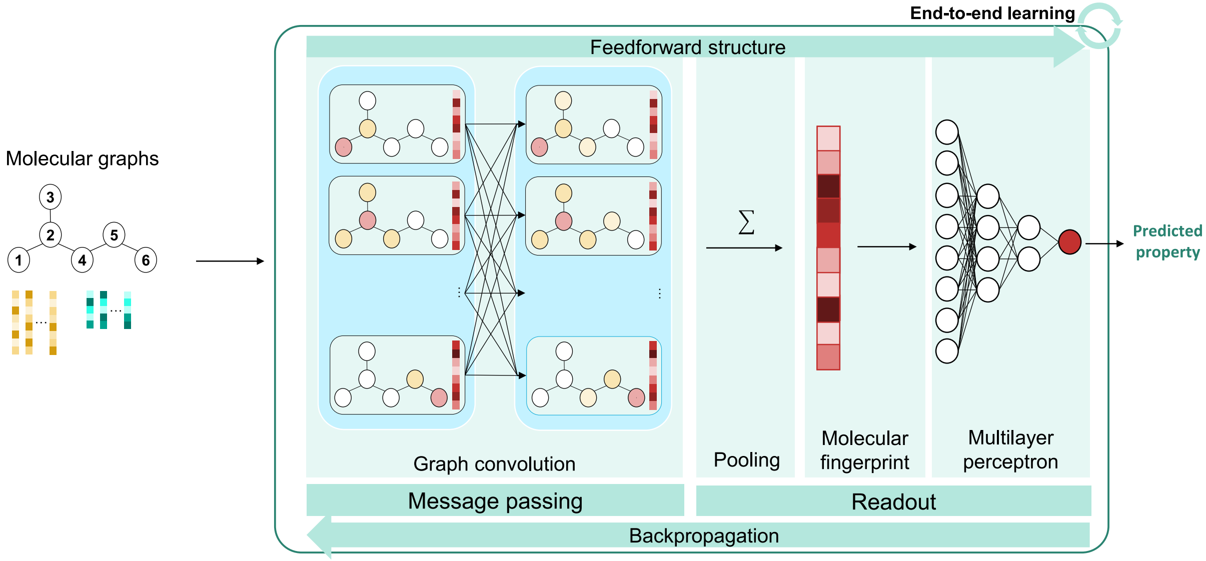


Fig. 1. Illustration of a GNN architecture for molecular property prediction (based on Schweidtmann et al., 2020).

With iterations of multiple graph convolutional layers L together, the information of all neighboring nodes within a distance of L nodes is passed to a node, thereby enabling learning of the so-called L-hop local environment of each node.

In the *readout phase*, a graph representation vector, the so-called molecular fingerprint vector $\mathbf{h}_G$ is aggregated from the learned structure information during message passing by the pooling function P such as $\mathbf{h}_G = P(\{\mathbf{h}_v^L \mid v \in V\})$. In particular, the pooling function P is commonly chosen as the mean, sum, or max function (Schweidtmann et al., 2023). Finally, the molecular fingerprint $\mathbf{h}_G$ is mapped to the property of interest by a regression model such as an MLP: $\hat{p} = \mathrm{MLP}(\mathbf{h}_G)$.

The functions in the message passing phase and pooling functions in the readout phase are both explicit and differentiable, which enables the training of the GNN with backpropagation in a supervised learning setup (Gilmer et al., 2017; Hamilton et al., 2017). This means that GNNs learn in an end-to-end manner, from the molecular graph to the property of interest instead of steps for the selection of informative molecular descriptors as is required in QSPR/QSAR modeling.

3. Datasets

The dataset comprises 51,905 processes producing 761 distinct molecules (represented by SMILES string (Weininger, 1988)) from 91 countries, where the target chemical is manufactured, which are incorporated as categorical features. Furthermore, each process contains 15 impact categories, as shown in Table 1. These impact results are process-specific, meaning the values for the same chemical product may vary depending on the production method and region. Acidification (AC) quantifies the potential of acidifying substances to disrupt ecosystems, expressed in mol H^+ equivalents. Climate change (CC) measures greenhouse gas emissions contributing to global warming in kg CO_2 equivalents.

Ecotoxicity, freshwater (ECO) assesses the toxic effects of chemicals on aquatic life, reported in comparative toxic units for ecosystems (CTUe). Energy resources, fossils (ER) reflects the depletion of non-renewable fossil fuels in megajoules. Eutrophication is divided into freshwater (EUF), marine (EUM), and terrestrial (EUT) categories, capturing nutrient enrichment impacts using phosphate (kg PO_4 eq), nitrogen (kg N eq), or mol N equivalents, respectively. Human toxicity (HT) estimates long-term effects of chemical exposure on human health in CTUh. Ionizing radiation (IR) represents human exposure to radioactive substances relative to uranium-235 (c kBq U-235 eq). Land use (LU) captures impacts on soil quality and ecosystem functionality. Material resources, metals/minerals (MR) quantifies the depletion of abiotic mineral and metal resources in kg Sb equivalents. Ozone depletion (OD) reflects the contribution to stratospheric ozone layer degradation in kg CFC-11 equivalents. Particulate matter formation (PMF) measures the increase in respiratory disease incidence from airborne particles. Photochemical ozone formation (POF) estimates the formation of ground-level ozone harmful to human health, reported in kg NMVOC equivalents. Lastly, water use (WU) accounts for water scarcity potential in m^3 water equivalents. These fully characterized scores require no further transformation and serve as direct prediction targets in our models. For additional details, we refer to the literature (Stellner et al., 2022; European Commission, 2021)

All the impact categories are calculated using the Environmental Footprint (EF) framework implemented into the Carbon Minds Database (V1.01, 2022) (Stellner et al., 2022). This framework is based on the Environmental Footprint v3.0 (EF v3.0) Life Cycle Impact Assessment (LCIA) method. EF v3.0 (European Commission, 2021) is a life cycle impact assessment method developed by the European Commission to support consistent, science-based environmental evaluation of products. It provides midpoint characterization factors across multiple impact categories and is aligned with the Product Environmental Footprint (PEF) framework for policy and industry applications. Unlike ReCiPe2016 (Huijbregts et al., 2016), EF v3.0 follows a single, consensus-based modeling approach without offering multiple perspectives. EF v3.0 is typically applied to chemical products using a cradle-to-gate system boundary, meaning it accounts for impacts from raw material extraction up to the point of product leaving the production site. It is important to mention that EF v3.0 includes 16 categories, due to the distinction between Human Toxicity of carcinogens and noncarcinogens. In this work, we count Human Toxicity as a single category. Furthermore, the collection and precalculation of the impact scores are done by Carbon Minds independently outside of this work.

In addition, Fig. 2 demonstrates the count of molecules in our dataset containing each functional group. Note that these are not counts of unique molecules, as a single molecule may contain multiple functional groups. Approximately 728 molecules are organic, and 33 molecules are non-organic. Specifically, the most frequent groups were alcohols (13%), aromatics (11%), and amines (10.5%), followed by halogenated compounds, ketones, and esters. This distribution reflects a broad spectrum of functionalized molecules relevant to industrial chemical processes. Approximately 9% of molecules could not be clearly assigned to any predefined group and were labeled as “Other”.

Table 1
Overview of fifteen selected impact categories. The corresponding impact scores and units are listed. Note that the term "potential" in Table 1 refers to the standard terminology of midpoint impact categories from the EF v3.0 LCIA method, which requires no further transformation.

Impact categories	Impact scores (unit)
Acidification (AC)	Accumulated exceedance (mol H+$_{eq}$)
Climate change (CC)	Global Warming Impact (kg CO$_{2\ eq}$)
Ecotoxicity, freshwater (ECO)	Comparative Toxic Unit for ecosystems (CTUe)
Energy resources, fossils (ER)	Abiotic depletion potential (MJ)
Eutrophication, freshwater (EUF)	Fraction of nutrients reaching freshwater end compartment (kg PO$_{4\ eq}$)
Eutrophication, marine (EUM)	Fraction of nutrients reaching marine end compartment (kg N$_{eq}$)
Eutrophication, terrestrial (EUT)	Accumulated exceedance (mol N$_{eq}$)
Human toxicity (HT)	Comparative Toxic Unit for humans (CTUh)
Ionizing radiation (IR)	Human exposure efficiency relative to U^{235}_{eq} (c kBq U^{235}_{eq})
Land use (LU)	Soil quality index (-)
Material resources, metals/minerals (MR)	Abiotic depletion potential (kg Sb$_{eq}$)
Ozone depletion (OD)	Ozone depletion potential (kg CFC-11$_{eq}$)
Particulate matter formation (PMF)	Impact on human health (Disease incidence)
Photochemical ozone formation, human health (POF)	Tropospheric ozone concentration increase (kg NMVOC$_{eq}$)
Water use (WU)	User deprivation potential (m^3 water$_{eq}$)

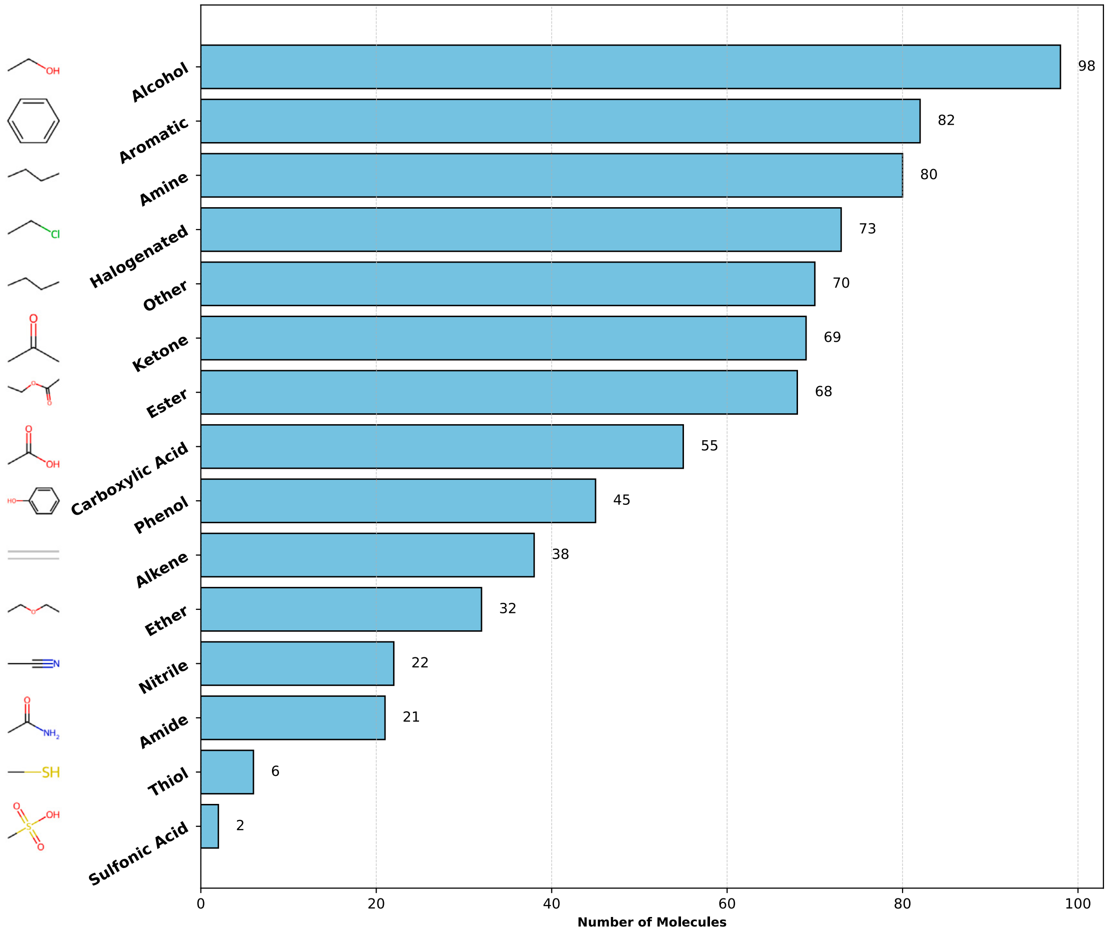


Fig. 2. Functional group counts of the molecules in the dataset. Notably, one molecule could potentially contain more than one functional groups.

Molecular features with categorical country features offer a broad overview, however, they fall short of capturing the important contributions of different energy sources and the dynamic shifts in policy and technology that significantly influence the environmental impact categories. To address this limitation and enrich the predictive capability of our model, we integrate energy mix data that details the specific contributions of various energy sources to carbon emissions. This data is collected from the International Energy Agency (IEA) open-source version for 49 countries. The rest of the countries are assigned to broader regions (e.g., Africa, Middle East). Moreover, these data encompasses industry consumption energy data across eight product flows: Coal, peat, and oil shale; Crude, NGL, and feedstocks; Oil products; Natural gas; Nuclear; Renewables and waste; Electricity and heat. The detailed information is shown in Table 8.

4. Methods

In this section, we introduce the development of the GNN model. First, we describe the data representation approaches for molecules, countries, and energy mix features in Section 4.1. Next, we present the GNN and QSPR architectures in Section 4.2. Finally, the model extensions for single/multi-task learning are described.

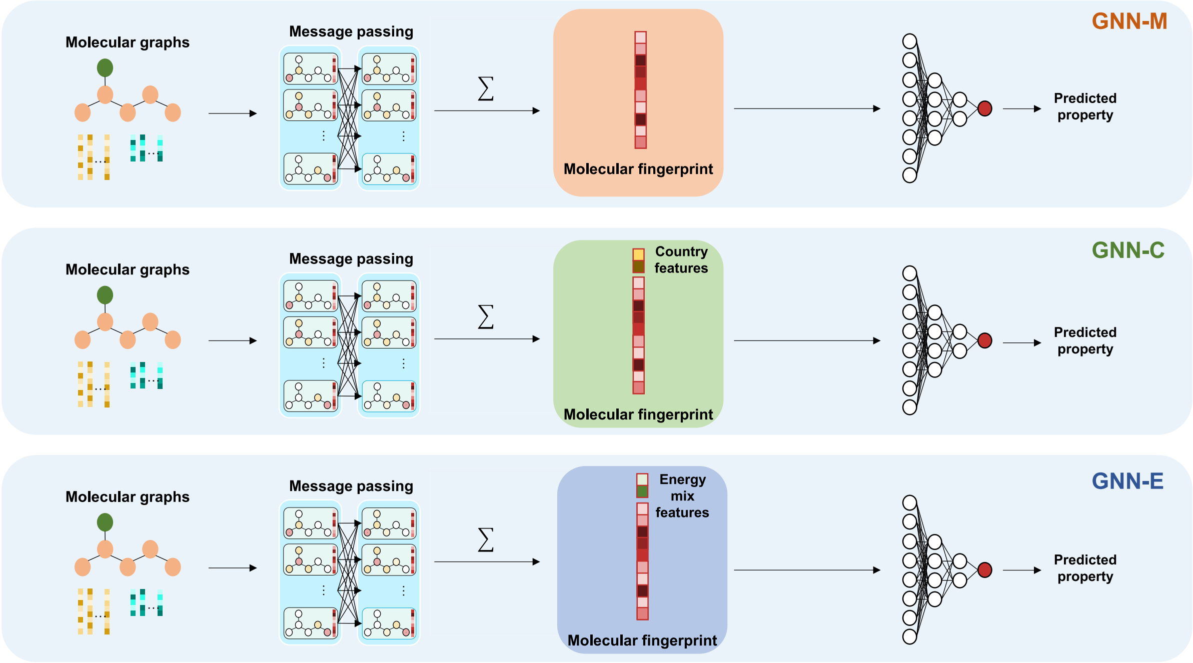


Fig. 3. Illustration of three GNN architectures: GNN-M, GNN-C and GNN-E.

Table 2
Atom features for initial node feature vector (Gilmer et al., 2017; Schweidtmann et al., 2020; Rittig et al., 2023) in GNN models. Features are typically encoded as one-hot vectors.

Feature	Description	Dimension
Atom type	Atom type (e.g., C, H, O) ordered by atomic number	19
Is in ring	Whether the atom is part of a ring	1
Is aromatic	Whether the atom is part of an aromatic ring	1
Number of neighbors	Number of bonded atoms	6
Hybridization	Atom hybridization (e.g., s, sp, sp2...)	5

4.1. Data representation

Data representation is critical since it directly affects the training process for ML models. In QSPR, molecules are denoted as molecular descriptors. Here, each molecule is described by 56 molecular descriptors as proposed in the earlier work by Song et al. (2017). These descriptors are automatically generated by the software *Dragon 7* (Mauri et al., 2006). The list of descriptors can be found in Song et al. (2017) and contains common descriptors such as molecular weight, number of aromatic rings, number of functional groups, and number of halogen atoms. We represent each product molecule with a vector d of the length 56, where each entry corresponds to a descriptor with continuous or binary values. Notably, those descriptors are only utilized for QSPR models training.

In GNNs, molecules can be naturally represented by molecular graphs where atoms are denoted as nodes and edges are represented as edges. Each node and edge has one associated feature vector, which is outlined in Tables 2 and 3. These features are local graph-based representations and are fundamentally different from the global, descriptor-based features used in the QSPR model. For example, the QSPR descriptors include molecular weight, percentage of N atoms and fragment counts — each computed from the entire molecular graph — whereas the GNN models operate on atomic and bonding features such as atom type, hybridization state, and bond type, which are used in the message passing process. All the features are represented as a one-hot encoder in the corresponding dimension for this feature, where a single entry with value one at the index corresponds to the value of the feature. Specifically, molecules are given as SMILES and we convert SMILES strings into molecular graphs through RDkit (Landrum et al., 2024).

Table 3
Bond features for initial edge feature vector (Gilmer et al., 2017; Schweidtmann et al., 2020; Rittig et al., 2023) in GNN models. Features are typically encoded as one-hot vectors.

Feature	Description	Dimension
Bond type	Bond type as in single, double, triple or aromatic	4
Is conjugated	Whether the bond is conjugated	1
Is in ring	Whether the bond is in a ring	1

4.2. Model architectures

We first set up a QSPR model to do the comparative study with GNN-based models. The QSPR model is adapted from Song et al. (2017), which contains an $\mathbf{MLP}_{\text{QSPR}}$ with two fully connected layers with 16 hidden dimensions and ReLu activation function. Taking molecular descriptors $\mathbf{d}$ as input, the QSPR model calculates the target property $\hat{\mathbf{p}}$ by Eq. (3). The QSPR model parameters are listed in Table 10.

$$\hat{\mathbf{p}} = \mathbf{MLP}_{\text{QSPR}}(\mathbf{d}) \tag{3}$$

Fig. 3 shows three GNN models that we propose: 1. Molecular features only (GNN-M); 2. Molecular features and one-hot encoded country features (GNN-C); 3. Molecular features and energy mix features (GNN-E). All three GNN-based models deploy the same model architecture for molecular feature representation learning. We first convert the molecules into attributed molecular graphs that serve as input to the GNN. In the message-passing phase, gated recurrent units (GRUs) (Cho et al., 2014) are used to explore local atomic environments

Table 4
Overview of model inputs, outputs, and implementation frameworks.

Model	Input	Output
QSPR model	Molecular descriptors (Dragon 7)	Climate change (CC) score
GNN-M	Molecular graphs (RDKit)	Climate change (CC) score
Single-task GNN-C/GNN-E	Molecular graphs + country or energy-mix vectors	One impact category score
Multi-task GNN-C/GNN-E	Molecular graphs + country or energy-mix vectors	15 impact category scores

within the molecular graphs, which has shown promising results in molecular property prediction (Withnall et al., 2020; Meng et al., 2019; Schweidtmann et al., 2020). Specifically, the hidden state in layer l is updated by Eq. (4):

$$\mathbf{h}_v^l = \mathbf{GRU}(\mathbf{h}_v^{l-1}, \sigma(\theta_v \cdot \mathbf{h}_v^{l-1} + \mathbf{m}_v^l)) \tag{4}$$

where $\mathbf{m}_v^l$ is given by Eq. (5).

$$\mathbf{m}_v^l = \sum_{w \in N(v)} \mathbf{MLP}_{\theta_e}(\mathbf{f}^{e_{vw}}) \cdot \mathbf{h}_w^{l-1} \tag{5}$$

The edge features $\mathbf{f}^{e_{vw}}$ are first mapped by a **MLP** into a parameter matrix θ_e, which is further multiplied with the hidden states of neighborhoods of node v, $\mathbf{h}_w^{l-1}$ where $w \in N(v)$, to get message $\mathbf{m}_v^l$. The message $\mathbf{m}_v^l$ is then added to hidden state $\mathbf{h}_v^{l-1}$ multiplied with a parameter matrix θ_v. The result is passed into an activation function, exponential linear unit (ELU), and then together with hidden state $\mathbf{h}_v^{l-1}$ to get the final updated hidden state $\mathbf{h}_v^l$ after **GRU**. Furthermore, after l layers of graph convolution, the molecular fingerprint $\mathbf{h}_G$ is calculated by Eq. (6).

$$\mathbf{h}_G = \sum_{v \in V} \mathbf{h}_v \tag{6}$$

where the pooling function is the sum. To note that, the initial hidden states $\mathbf{h}_v^0$ are passed into a shallow **MLP** with **ReLu** activation to ensure the uniform dimension.

All three GNN-based models use the same message passing layers to process molecular graphs but differ in how they incorporate additional features for property prediction, which are all implemented by the Pytorch-Geometric package. The detailed model parameters can be found in Tables 11, 13 and 12 for GNN-M, GNN-C, and GNN-E in the Appendix section, respectively. It is important to mention that the concatenation of learnable molecular embedding is similar to traditional approaches that use molecular descriptors alongside thermodynamic or process descriptors to predict impact scores, as demonstrated in several previous studies (Karka et al., 2022; Kleinekorte et al., 2019; Calvo-Serrano et al., 2018b). The key distinction in our work is that the molecular representation is now learned dynamically via GNNs rather than being manually engineered.

- **GNN-M**: The molecular fingerprint $\mathbf{h}_G$ is directly fed into an MLP to predict the target property:

$$\hat{\mathbf{p}} = \mathbf{MLP}_{\text{GNN-M}}(\mathbf{h}_G) \tag{7}$$

- **GNN-C**: A one-hot encoded country feature $\mathbf{F}_c$ (91-dimensional) is concatenated with the molecular fingerprint before training:

$$\hat{\mathbf{p}} = \mathbf{MLP}_{\text{GNN-C}}\left[\mathbf{Cat}(\mathbf{h}_G, \mathbf{F}_c)\right] \tag{8}$$

- **GNN-E**: The model extends GNN-C by also incorporating a seven-dimensional energy mix feature $\mathbf{F}_e$:

$$\hat{\mathbf{p}} = \mathbf{MLP}_{\text{GNN-E}}\left[\mathbf{Cat}(\mathbf{h}_G, \mathbf{F}_e)\right] \tag{9}$$

4.3. Training approach

We began by training a QSPR model and three GNN-based models (GNN-M, GNN-C, and GNN-E) for the CC impact category. The QSPR model takes molecular descriptors (listed in Table 9) as input and was evaluated using 10-fold cross-validation. Training hyperparameters for this model are provided in Table 10.

The GNN-M model was trained under the same cross-validation process, using molecular graphs as input. Details of its architecture and training settings are summarized in Table 11. For the GNN-C and GNN-E models, the molecular graph is learned through message passing to produce a molecular fingerprint. In GNN-C, this fingerprint is concatenated with a one-hot-encoded geographical feature vector (dimension = 91), while in GNN-E, it is concatenated with an energy mix feature vector (dimension = 7). Both are passed through multilayer perceptrons (MLPs) for CC prediction. Architectural and training configurations for these models are shown in Tables 12 and 13.

We then extended GNN-C and GNN-E to cover all 15 impact categories. These models were trained in both single-task and multi-task settings. In single-task training, a separate model is trained independently for each target property—this is the approach used for the initial CC models—resulting in different fingerprints for the same molecule across tasks. In contrast, multi-task learning employs a shared model where all tasks utilize the same message passing layers for molecular representation, while separate MLPs are used for each prediction target. This allows the model to learn generalizable features across tasks while still capturing task-specific signals in the readout layers.

Multi-task learning has been shown to improve generalization and reduce overfitting by encouraging shared feature learning across tasks (Ruder, 2017). Prior work has demonstrated its advantages in molecular property prediction (Burkardt et al., 2021; Dahl et al., 2014; Ramsundar et al., 2015). In our case, we apply a multi-task strategy to jointly predict 15 LCA impact categories using a shared GNN backbone and 15 parallel MLP heads.

All models were trained on a dataset of 761 molecules, with 51,905 total processes across countries and impact categories, using 10-fold cross-validation. The detailed input–output structure is listed in Table 4. Hyperparameters and architectural configurations for single-task GNN-C and GNN-E models are listed in Tables 12 and 13, respectively. Learning rates were tuned individually for each category and are reported in Table 16. Multi-task model configurations for both GNN-C and GNN-E are detailed in Tables 14 and 15.

To avoid overfitting and improve generalization, we apply several standard regularization techniques during model training. Specifically, we employ early stopping, which stops training when performance on the validation set ceases to improve for a fixed number of epochs, preventing overfitting to the training data. Additionally, we apply a learning rate decay strategy, which gradually reduces the learning rate over time to allow finer convergence in later training stages. All corresponding training hyperparameters are summarized in the Appendix tables. Importantly, folds are constructed based on the number of unique molecules, rather than the total number of samples. For the QSPR and GNN-M models, where only molecular descriptors or molecular graphs are used as input, each molecule corresponds to a single data point (e.g., one climate change impact value). In these cases, each fold consists of 80% of the molecules for training, 10% for validation, and 10% for testing. This ensures that each molecule is seen only once across the folds. In contrast, the GNN-C and GNN-E models incorporate additional contextual information—namely, geographical and energy mix features. Here, each molecule may be associated with multiple samples due to its presence across different countries. However, not all chemicals have complete data for all 91 countries. As a result, when folds are split based on unique molecules, the number

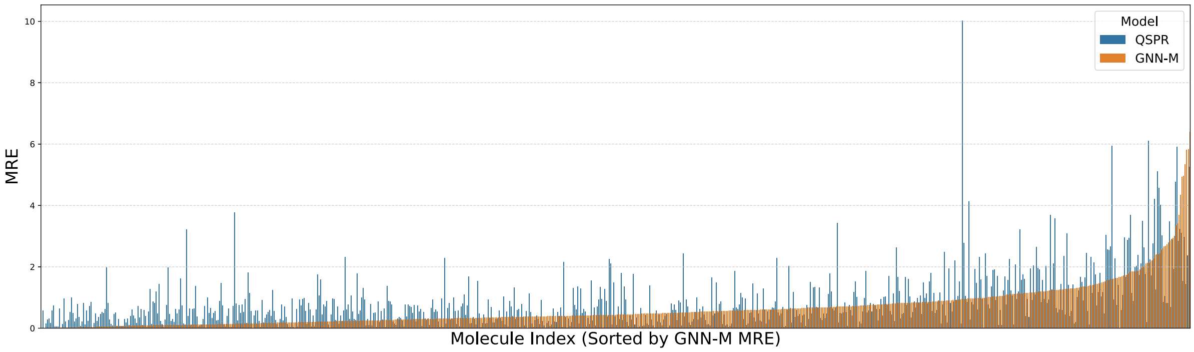
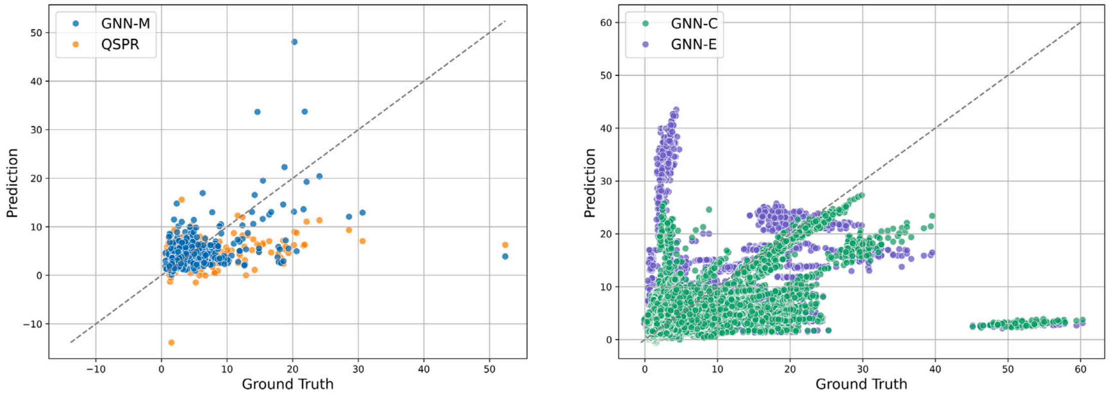


Fig. 4. Per-molecule Mean Relative Error (MRE) comparison between the GNN-M and QSPR models. Molecules are sorted by GNN-M MRE to highlight performance differences across the dataset.



(a) Parity plot of QSPR and GNN-M models

(b) Parity plot of GNN-C and GNN-E models

Fig. 5. Parity plots of QSPR, GNN-M, GNN-C, and GNN-E for predicting CC category.

of training, validation, and test samples may vary across folds. The reason for splitting by unique molecules rather than by sample is to evaluate the model's ability to generalize to entirely unseen chemicals. This approach reflects a more realistic deployment scenario where the model is applied to new molecules not encountered during training.

4.4. Evaluation metrics

To assess model performance, we use two standard regression metrics: the mean relative error (MRE) and the coefficient of determination (R^2).

The MRE is defined as:

$$\mathrm{MRE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \tag{10}$$

where y_i and $\hat{y}_i$ denote the ground truth and predicted values, respectively, and n is the number of samples. MRE provides a relative measure of prediction error, normalized by the true values, and is robust to scale differences across categories.

The R^2 score is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{11}$$

where $\bar{y}$ is the mean of the ground truth values. R^2 quantifies the proportion of variance in the target variable that is explained by the model. An R^2 value of 1 indicates perfect prediction, while a value of 0 means the model performs no better than predicting the mean.

Notably, R^2 can be negative when the model performs worse than the baseline prediction $\hat{y}_i = \bar{y}$. A negative R^2 suggests that the model fails to capture the underlying trend in the data and instead increases the prediction error relative to a naive mean-based approach. Note that throughout this manuscript, we use the coefficient of determination (R^2) rather than correlation coefficients (Pearson or Spearman), which are also commonly reported in the literature.

5. Results and discussion

In this section, we first compare the QSPR models with GNNs for the CC impact category. Next, we discuss the results of single- and multi-task GNN with additional country and energy mix features.

5.1. Comparing QSPR and GNN for the climate change impact category

We benchmark our proposed models on the CC impact category, also as referred to GWP or GWI in the literature, because this impact category is widely studied in the literature. Table 5 shows results of the MRE and R^2 for this benchmark.

First, we compare the QSPR models to three GNN models. One observation is that GNN-M outperforms the QSPR model as MRE decreases by 6% and R^2 increases by 46.9%. This indicates that in this case, the end-to-end learning manner enables GNN-based models to automatically extract effective features from molecular graphs. Fig. 5(a) further verifies that the GNN-M (blue) demonstrates improved predictive accuracy compared to the QSPR model (orange). GNN-M predictions are more tightly clustered along the parity line. Fig. 4 compares the MRE of GNN-M and QSPR models across 761 molecules. The x-axis shows molecule indices sorted by GNN-M's error from lowest to highest, while the y-axis shows the MRE for each molecule. The plot illustrates how prediction accuracy varies per molecule and highlights that GNN-M generally achieves lower errors than QSPR. Notably, the highest

Table 5
Results of QSPR and GNN models, as well as previous literature, for CC impact categories. The assessment metrics utilized include MRE and R^2 on the test dataset. Each metric represents the mean value ± standard deviation of the best models.

Model	Model type	Dataset (# molecules)	Additional features	MRE /%	R^2
QSPR	MLPs	761	–	108 ± 6	0.17 ± 0.03
GNN-M	GNN	761	–	101 ± 30	0.32 ± 0.09
GNN-C	GNN	761	✓	72 ± 5	0.39 ± 0.04
GNN-E	GNN	761	✓	77 ± 2	0.35 ± 0.02
Literature					
Wernet et al. (2008)	MLPs & linear models	103	–	–	0.37 ± 0.36
Wernet et al. (2009)	MLPs	338	–	58.2	0.41 ± 0.23
Song et al. (2017)	MLPs	166	–	50	0.48[a]
Calvo-Serrano et al. (2017, 2018b)	MINLPs	83	✓	41.82	0.55[b]
Calvo-Serrano et al. (2018a)	MINLPs	90	✓	30	0.55[b]
Kleinekorte et al. (2019)	MLPs	63	✓	–	0.30[a]
Baxevanidis et al. (2021)	Six linear/nonlinear models	214	–	53	0.26
Karka et al. (2022)	MLPs & classification tree	–	✓	–	0.72 ± 0.14
Kleinekorte et al. (2023)	GNN	166	✓	–	0.53
Zhang et al. (2024)	GNN & Transformer	547	–	38.6	-

[a] R^2 Pearson

[b] R^2 Spearman

prediction errors from GNN-M are primarily associated with sulfur-containing compounds. This is likely because sulfur-related chemicals often require detailed process-specific information — such as oxidation states, reaction pathways, or treatment technologies — which cannot be captured by molecular descriptors alone. To further compare the predictions of sulfur-related chemicals remains challenging, as, to the best of our knowledge, no existing studies report model predictions for specific individual molecules in this category. Moreover, the incorporation of the country and energy mix features improved both R^2 and RMSE performance, as depicted by GNN-C and GNN-E results. For example, compared with GNN-M, the R^2 increases by 21.8% and 9%, and RMSE drops by 28.7% and 23.8% for GNN-C and GNN-E, respectively. Additionally, both GNN-C and GNN-E models have lower standard deviations than GNN-M, which indicates that the results are likely more reproducible and reliable. Furthermore, the GNN-C model depicts slightly better performance than the GNN-E model in terms of MRE and R^2. One plausible reason is that the preprocessing of the energy mix data requires several simplifications due to the lack of information from several countries. Those simplifications may account for the high variance and less accurate predictions than GNN-E.

When inspecting Fig. 5(b), both GNN-C and GNN-E models are able to capture the overall trend. However, substantial deviations from the parity line are observed across both models. The GNN-E model (purple) tends to overestimate the impact for low- to mid-range values, while the GNN-C model (green) behaves more conservatively, but underestimates in most cases. Notably, the sulfur dichloride process has a true CC impact greater than 45 in the plot. Both models significantly underestimate this value, predicting well below 10. This discrepancy arises because the majority of sulfur-containing compounds in the training data have much lower CC values—typically below 2. For instance, disulfur dichloride (S_2Cl_2), which is structurally similar to sulfur dichloride (SCl_2), consistently shows CC values around 1.5 across various countries. As a result, the models generalize a learned pattern that “sulfur-containing compounds” are associated with low CC, and fail to account for outliers like sulfur dichloride. This large gap is explained by differences in industrial production processes. Disulfur dichloride is synthesized directly from molten sulfur and chlorine gas in a relatively simple and energy-efficient process. In contrast, sulfur dichloride is produced by further chlorinating disulfur dichloride, requiring additional energy, materials, and process complexity (Ulonska et al., 2016). This is a limit of molecular-based predictions when process context (e.g., feedstock origin, energy use, reaction severity) is crucial.

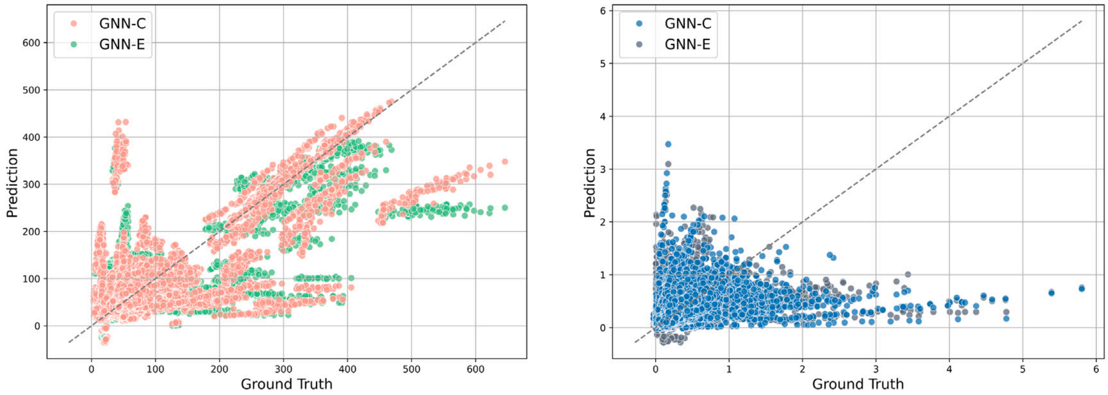
At first glance, the results in Table 5 show that our GNN-C and GNN-E models yield R^2 values that are comparable with, but in most cases lower than, those reported in the literature. Specifically, the GNN-C model achieves an R^2 of 0.39, which is comparable with the R^2 values of 0.37 and 0.36 reported by Wernet et al. (2009) and Kleinekorte et al. (2019), respectively. Furthermore, when compared with Karka et al. (2022), GNN-C and GNN-E models demonstrate a substantially lower R^2, implying a potential need for integrating additional process information to enhance prediction accuracy.

However, it is important to note that our models exhibit higher MRE compared to previous studies, with MRE values ranging from 72% to 77%, whereas earlier works report values between 30% and 58.2%. These differences can be partially attributed to the distinct methodologies and datasets employed. Our study leverages a significantly larger dataset of 761 molecules, which is substantially greater than the 63 to 338 molecules used in previous studies. While this expanded dataset enhances the potential for more generalizable models, it also increases the complexity of the prediction task, potentially leading to the observed higher MRE values. It is important to note that our dataset is based on commercially available and transparent data sources, ensuring broader applicability. However, the increased dataset size also introduces greater structural and physicochemical variability, which inherently makes direct comparisons with smaller datasets more challenging.

Additionally, a direct comparison of R^2 values across studies is complicated by the use of different evaluation metrics and model configurations. For instance, some studies used Pearson (Song et al., 2017; Kleinekorte et al., 2019) or Spearman (Calvo-Serrano et al., 2018a,b) R^2 metrics, while we employed the coefficient of determination. This variability in evaluation methods further complicates the comparison. However, the standard deviation of R^2 in our models is significantly lower than in previous studies, suggesting better reproducibility and model stability. Overall, despite these challenges, our results suggest that GNN-C and GNN-E models are promising tools for environmental impact prediction due to their improved performance in R^2 and RMSE over baseline models like GNN-M, as well as their lower standard deviations, indicating better reproducibility. Their ability to incorporate complex features, such as country and energy mix, enhances their applicability across different impact categories, potentially making them valuable for further development in environmental modeling.

5.2. Single-task model GNN

Table 6 shows the performance of the single-task GNN-C and GNN-E models across fifteen impact categories in terms of MRE and R^2. As



(a) Parity plot of single-task GNN-C and GNN-E models for predicting ER category (b) Parity plot of single-task GNN-C and GNN-E models for predicting IR category

Fig. 6. Parity plots of single-task GNN-C and GNN-E for predicting ER and IR category, respectively.

Section 5.1 discussed, here we omit training the QSPR and GNN-M because GNN-C and GNN-E depict more promising results.

First, although both GNN-C and GNN-E models demonstrate some ability to predict the majority of impact categories, the overall performance is limited. For example, nine out of fifteen R^2 and ten out of fifteen R^2 are non-negative for GNN-C and GNN-E, respectively. Additionally, neither GNN-C nor GNN-E consistently outperforms the other across all categories. Specifically, GNN-E shows higher R^2 values in ten out of fifteen categories than GNN-C, suggesting that GNN-E may better explain the variability in the data for those categories. This could be important in cases where the focus is on understanding overall trends and patterns, such as exploratory analyses or when dealing with noisy data. On the other hand, GNN-C exhibits a lower MRE in twelve out of fifteen categories than GNN-E, which may indicate that GNN-C provides more accurate predictions with smaller average errors. This might be more relevant when precise predictions are needed, such as in applications where the magnitude of errors has a significant impact. As a result, the choice of model should be guided by the specific requirements of the impact category under study. For example, a model with a higher R^2 may be preferable for explaining variability, while a model with lower MRE might be chosen when accuracy and precision are critical

Moreover, both models show better performance in predicting CC, ER, EUF, and WU categories than others. This indicates that the additional country and energy mix features can account for geographical, regulatory, and energy-related differences across countries and provide models with more context for predicting those environmental impacts. For example, GNN-E has superior performance on water-related impact categories such as EUF and WU. One potential reason is that energy mix features provide related information about energy consumption patterns and their associated water footprints across different regions. Energy production and consumption significantly influence water usage, with various energy sources requiring differing amounts of water for extraction, processing, and generation, thereby giving more accurate predictions. However, it is important to note that the models exhibit higher prediction errors in certain categories, such as LU and IR. This is expected, as these impact categories are less directly influenced by molecular properties and are instead driven by broader land management practices and nuclear-related processes, which are not explicitly captured in our input features.

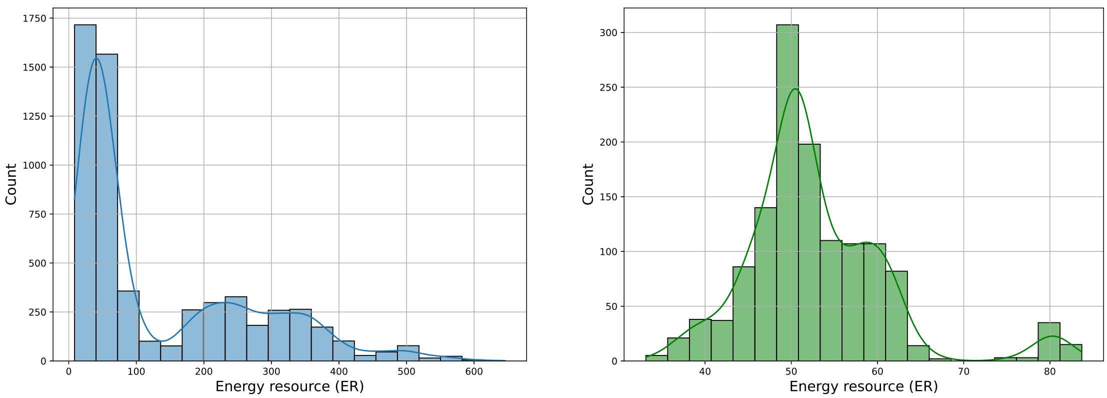
Furthermore, we selected the ER category and IR category to discuss the performance through Fig. 6. In Fig. 6(a), both models demonstrate reasonable performance for the ER category, with the majority of predictions aligning with the parity line. However, we observe that one region (top left) is significantly overestimated, and one region (middle right) is underestimated. When inspecting the corresponding chemicals, all significant prediction errors—both underestimations (e.g., carbon disulfide, 1,3,4-thiadiazolidine-2,5-dithione, and tetramethyl thiuram disulfide) and overestimations (e.g., thioformamide)—originated from sulfur-containing compounds. As shown in Fig. 7(a), sulfur-related chemicals exhibit a highly skewed and widespread ER distribution, ranging from very low (< 10) to extremely high values (> 600). In contrast, the ER values for alkene-containing molecules (Fig. 7(b))) form a narrower (10–90) and more symmetric distribution centered around moderate values. The wide variance in ER values for sulfur-related compounds likely limits the model's generalizability. This may stem from the diverse nature of sulfur production processes, which differ substantially in energy intensity depending on feedstocks, technologies, and purification steps. These results highlight the need to consider intra-family heterogeneity and suggest that incorporating process descriptors may improve prediction for such chemically diverse families. On the other hand, Fig. 6(b) reveals that both models significantly underperform on the IR category. Predictions for both GNN-C and GNN-E are heavily clustered at low values and fail to reproduce the tail of the ground truth distribution.

In conclusion, Table 6 suggests that neither model consistently outperforms the other across all categories. The choice between GNN-C and GNN-E may depend on the specific impact category of interest. Therefore, we further deploy multi-task GNN to investigate the causes of high variance in model predictions, assess the robustness of the models, and explore potential improvements in model architecture or training data.

5.3. Multi-task GNN

We further investigate multi-task GNN-C and GNN-E models and the MRE and R^2 across fifteen impact categories (see results in Table 7). Furthermore, Fig. 9 provides a clear comparison of the performance between single-task and multi-task GNN models.

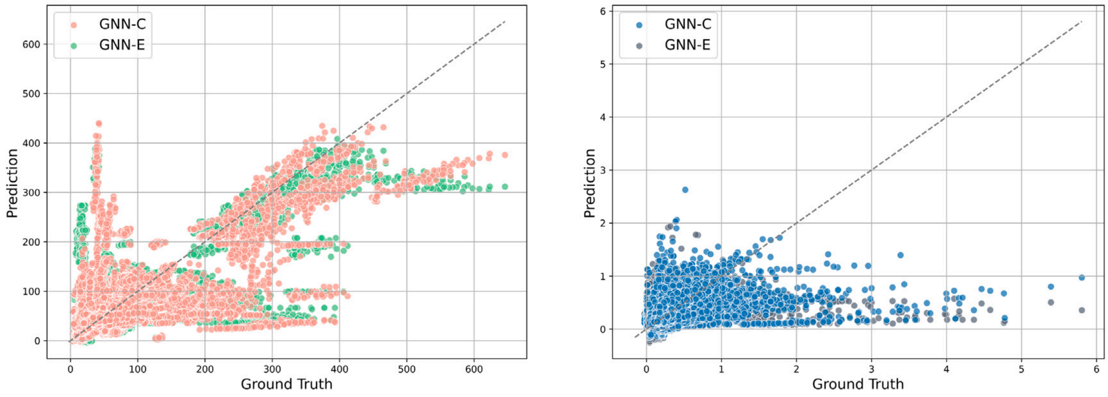
First, we observe a consistent improvement in MRE across nearly all impact categories when compared to the single-task GNN-C and GNN-E as Fig. 9(a) shows. For example, the MRE of fourteen out of fifteen categories for multi-task GNN-C decreases by ranging from 8% (PMF category) to 73% (OD category) compared with single-task GNN-C. The same trend echoes in multi-task GNN-E. The MRE of thirteen out of fifteen categories decreases by ranging from 16% (CC category) to 79% (IR category) compared with single-task counterparts. This enhancement verifies the hypothesis that multi-task GNN models learn a shared representation that captures underlying patterns common to all tasks, thereby building a more robust and generalizable model framework. This shared learning facilitates the transfer of knowledge



(a) Data distribution of Sulfur-related family for ER category

(b) Data distribution of alkenes-related family for ER category

Fig. 7. Data distribution of Sulfur-related family and alkenes-related family for ER category.



(a) Parity plot of multi-task GNN-C and GNN-E models for predicting ER category

(b) Parity plot of multi-task GNN-C and GNN-E models for predicting IR category

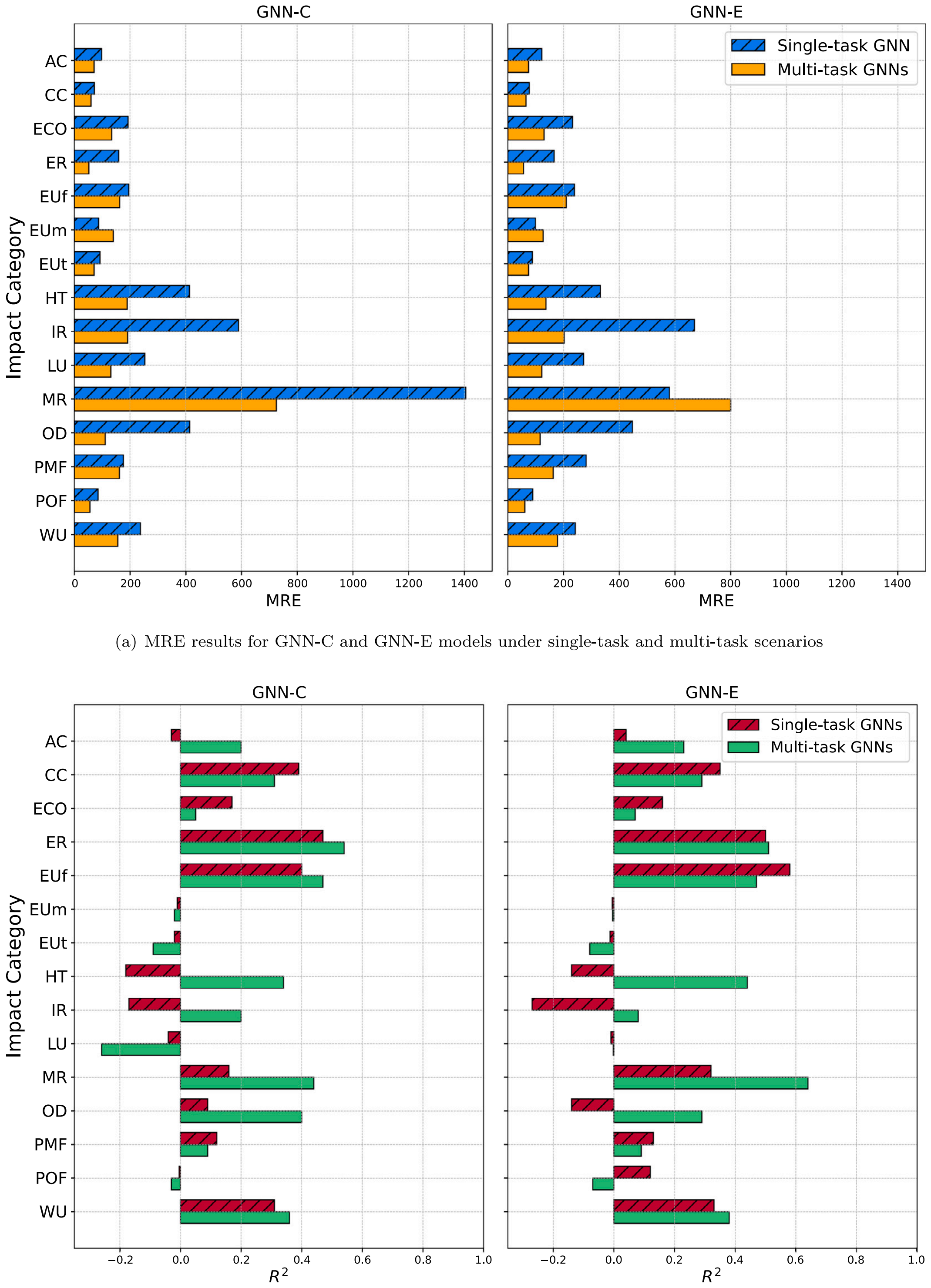
Fig. 8. Parity plots of multi-task GNN-C and GNN-E for predicting ER and IR category, respectively.

Table 6
Results of single-task GNN-C and GNN-E models evaluated across fifteen categories. The assessment metrics utilized include MRE and R^2. Each metric represents the mean value ± standard deviation obtained from three independent runs.

Impact	Single-task GNN-C		Single-task GNN-E	
category	MRE/%	R^2	MRE/%	R^2
AC	**98 ± 8**	−0.03 ± 0.02	122 ± 14	**0.04 ± 0.08**
CC	**72 ± 5**	**0.39 ± 0.04**	77 ± 2	0.35 ± 0.02
ECO	**193 ± 16**	**0.17 ± 0.08**	232 ± 30	0.16 ± 0.09
ER	**159 ± 2**	0.47 ± 0.04	166 ± 10	**0.50 ± 0.05**
EUF	**195 ± 48**	0.40 ± 0.16	239 ± 15	**0.58 ± 0.06**
EUM	87 ± 8	−0.01 ± 0.01	99 ± 5	**0.01 ± 0.01**
EUT	92 ± 6	−0.02 ± 0.03	**88 ± 3**	**−0.01 ± 0.01**
HT	413 ± 71	−0.18 ± 0.03	**332 ± 247**	**−0.14 ± 0.2**
IR	**589 ± 32**	**−0.17 ± 0.11**	670 ± 20	−0.27 ± 0.08
LU	**253 ± 10**	−0.04 ± 0.02	272 ± 36	**−0.01 ± 0.03**
MR	1405 ± 1312	0.16 ± 0.35	**580 ± 109**	**0.32 ± 0.37**
OD	**414 ± 22**	**0.09 ± 0.09**	447 ± 11	−0.14 ± 0.06
PMF	**176 ± 4**	0.12 ± 0.04	281 ± 28	**0.13 ± 0.03**
POF	**85 ± 4**	0.004 ± 0.022	89 ± 6	**0.12 ± 0.06**
WU	**237 ± 42**	**0.36 ± 0.19**	242 ± 12	0.33 ± 0.10

Table 7
Results of multi-task GNN-C and GNN-E models evaluated across fifteen categories. The assessment metrics utilized include MAE, MRE, and R^2. Each metric represents the mean value ± standard deviation obtained from three independent runs.

Impact	Multi-task GNN-C		Multi-task GNN-E	
category	MRE/%	R^2	MRE/%	R^2
AC	71 ± 1	0.20 ± 0.09	74 ± 1	**0.23 ± 0.04**
CC	60 ± 2	**0.31 ± 0.04**	65 ± 3	0.29 ± 0.03
ECO	134 ± 9	0.05 ± 0.09	130 ± 8	**0.07 ± 0.02**
ER	52 ± 1	**0.54 ± 0.03**	56 ± 1	0.51 ± 0.02
EUF	163 ± 18	0.47 ± 0.10	210 ± 27	**0.47 ± 0.02**
EUM	140 ± 27	−0.02 ± 0.00	127 ± 4	**−0.005 ± 0.006**
EUT	71 ± 2	−0.09 ± 0.07	74 ± 6	**−0.08 ± 0.09**
HT	190 ± 22	0.34 ± 0.24	137 ± 16	**0.44 ± 0.02**
IR	191 ± 6	**0.20 ± 0.01**	202 ± 3	0.08 ± 0.00
LU	131 ± 7	−0.26 ± 0.45	122 ± 11	**−0.002 ± 0.037**
MR	725 ± 219	0.44 ± 0.46	800 ± 189	**0.64 ± 0.11**
OD	111 ± 3	**0.40 ± 0.05**	116 ± 0	0.29 ± 0.02
PMF	162 ± 7	0.09 ± 0.04	163 ± 4	**0.09 ± 0.03**
POF	56 ± 1	**−0.03 ± 0.00**	61 ± 3	−0.07 ± 0.01
WU	156 ± 15	0.36 ± 0.19	178 ± 7	**0.38 ± 0.05**



(a) MRE results for GNN-C and GNN-E models under single-task and multi-task scenarios

(b) R^2 results for GNN-C and GNN-E models under single-task and multi-task scenarios

Fig. 9. MRE and results for GNN-C and GNN-E models under single-task and multi-task scenarios.

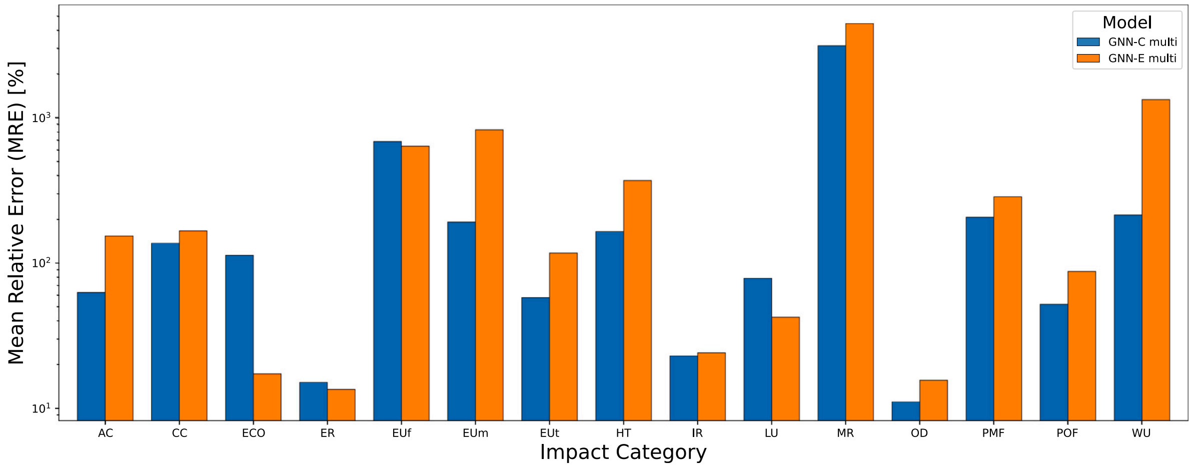


Fig. 10. Mean relative error (MRE, log scale) for propylene predicted with the multitask GNN-C (blue) and GNN-E (orange) models across 15 impact categories. For each category, the bar represents the MRE averaged over all country contexts, while the GNN-E bar is averaged over the corresponding national energy-mix scenarios.

between tasks, with insights or patterns learned in one task potentially benefiting others, particularly if they are interrelated.

For multi-task GNN models, eleven out of fifteen R^2 are non-negative for both GNN-C and GNN-E models. When compared with their single-task counterparts, the R^2 presents a mixed landscape with some categories showing improvements and others slight declines as Fig. 9(b) depicts. For instance, for multi-task GNN-C, in the categories of AC, HT, and IR, the performance has been improved by 115%, 152%, and 185%. While for categories ECO and POF, the R^2 significantly decreases by 216% and 112%. Additionally, for multi-task GNN-E, we do observe significant improvements in HT (131%), IR (456%), and LU (327%) categories, but for the ECO, EUM, and POF categories, the R^2 decreases by 132%, 224% and 260%, respectively, compared with single-task GNN-E. Notably, for both GNN-C and GNN-E, the multi-task models significantly improve the R^2 in HT and IR categories but decline in ECO and POF categories. This discrepancy could potentially be attributed to the nature of the tasks and their underlying data structures. HT and IR categories share similarities in how environmental exposures impact human health, allowing the multi-task GNN to leverage shared representations and improve predictive performance. In contrast, ECO and POF involve complex, distinct mechanisms—ecological interactions in aquatic environments for ECO, and atmospheric chemistry for POF—that do not easily align with the patterns learned from other tasks. Consequently, the multi-task model may struggle to generalize across these divergent categories, leading to a worse performance. These findings highlight the importance of task similarity and data alignment in multi-task learning frameworks, suggesting that while multi-task GNNs can enhance performance for related tasks, they may require careful design or even task-specific models when dealing with highly specialized or unrelated properties.

We further examine the parity plots for ER and IR categories for multi-task models. Compared to the single-task results, both models exhibit slightly improved alignment with the parity line, particularly in the mid-range values of the ER category (Fig. 8(a)). The overall trend is supported by the performance metrics: the corresponding improvements in R^2 and reductions in MRE (as shown in Table 7) confirm that generalization and error mitigation have indeed improved. For the IR category (Fig. 8(b)), although predictions still remain challenging, both models show improved dispersion compared to their single-task counterparts.

For the CC category, multi-task GNN models perform worse than single-task models. This may be because multi-task learning prioritizes general features, reducing the ability of the model to capture CC-specific patterns. Therefore, a single-task GNN is recommended for predicting CC category. Additionally, to cross-compare the multi-task results of GNN-C and GNN-E, both models display similar performances in terms of MRE and R^2. GNN-E marginally outperforms in eight impact categories for R^2 and maintains overall smaller standard deviations compared to GNN-C, indicating more reproducible results. On the other hand, GNN-C shows better MRE in ten out of fifteen impact categories than GNN-E.

Additionally, Fig. 10 shows the mean relative error (MRE, log scale) for propylene predicted by the multitask GNN-C and GNN-E models across the fifteen EF midpoint impact categories. Errors span more than three orders of magnitude: the lowest MREs occur for ER, OD, and IR (all below roughly the 20% range), whereas MR and WU exhibit extremely large errors (>10^3%), with EUf and EUm also high. Intermediate error levels (tens to low hundreds of percent) are observed for AC, CC, EUt, LU, PMF, and POF. Categories such as ER, OD, and IR are largely governed by upstream energy-conversion or halocarbon-related emission factors that scale well with the elemental composition of propylene and the national energy mix. They also show comparatively low inter-database variability, which helps the model keep MREs below 20% (Kim et al., 2023; Alyaseri and Zhou, 2019). Conversely, MR and WU require detailed information on ore grades, mining routes, and regional water scarcity indices, all of which vary widely across countries and databases (Kim et al., 2023; Alyaseri and Zhou, 2019). Because such process- and location-specific flows are absent from the molecular graph, the model under-predicts or over-predicts by orders of magnitude. Comparing the two models, adding energy-mix information (GNN-E, orange) substantially improves ECO and modestly reduces LU (and slightly ER/IR), but increases error relative to the country-only variant (GNN-C, blue) in categories such as AC, EUm, HT, MR, PMF, POF, and WU. Taken together, the figure highlights that molecular-based multitask models are most useful for energy- or chemistry-dominated categories, whereas process-intensive categories will still require additional life-cycle inventory data for reliable screening. This illustrative case also underscores the limitations of predictive models in practical applications—particularly in later stages of process development, where accurate life-cycle assessments depend on detailed process designs and simulation studies beyond molecular structure and national energy mix.

Overall, we offer the following recommendations for future work in early-stage prediction of life cycle impact categories. First, the choice of impact category is critical. When using only molecular features in combination with geographical or energy mix data, categories such as land use (LU) may be less suitable due to their weak correlation with molecular structure. We recommend excluding such categories from purely structure-based modeling efforts. Second, it is important to inspect the data distribution, particularly with respect to chemical diversity

Table 8
Overview of the energy mix data from IEA.

Product	Description
Coal, peat and oil shale	Coal includes all coal, both primary (including hard coal and lignite) and derived fuels (including patent fuel, coke oven coke, gas coke, BKB, gas works gas, coke oven gas, blast furnace gas and other recovered gases). Peat (including peat products) and oil shale are also included in this figure where applicable.
Crude, NGL and feedstocks	Crude oil comprises crude oil, natural gas liquids, refinery feedstocks and additives as well as other hydrocarbons (including emulsified oils, synthetic crude oil, mineral oils extracted from bituminous minerals such as oil shale, bituminous sand, etc. and oils from coal and gas liquefaction).
Oil products	Oil products comprise refinery gas, ethane, LPG, aviation gasoline, motor gasoline, jet fuels, kerosene, gas/diesel oil, fuel oil, naphtha, white spirit, lubricants, bitumen, paraffin waxes, petroleum coke and other oil products.
Natural gas	Natural gas includes both ‘associated’ and ‘non-associated’ gas as well as colliery gas (excluding natural gas liquids).
Nuclear	Nuclear shows the primary heat equivalent of the electricity produced by a nuclear power plant with an average thermal efficiency of 33%.
Renewables and waste	Renewables and waste comprises hydro, geothermal, solar, wind and tide/wave/ocean energy and the use of these energy forms for electricity and heat generation, as well as solid biofuels, liquid biofuels, biogases, industrial waste and municipal waste.
Electricity	Electricity shows final consumption and trade in electricity, which is accounted at the same heat value as electricity in final consumption (that is, 1 GWh = 0.0036 PJ).
Heat	Heat shows the disposition of heat produced for sale. The large majority of the heat included in this column results from the combustion of fuels although some small amounts are produced from electrically powered heat pumps and boilers. Any heat extracted from ambient air by heat pumps is shown as production.
Total	Total equals the total of products included in the dataset.

and class imbalance. If a small subset of molecules dominates the dataset or if certain classes are underrepresented, data augmentation or rebalancing strategies (Wieder et al., 2020) could be considered to improve model generalization in future work. Finally, for applications involving multiple impact categories, we suggest initially developing a multi-task GNN model using molecular and energy mix or geographical features. This approach enables shared representation learning across tasks. If performance is suboptimal for certain categories, those specific targets may benefit from additional fine-tuning using task-specific (single-task) GNNs. These strategies can help improve the robustness and interpretability of predictive models in the context of sustainable process and product design.

6. Conclusion

We propose an end-to-end GNN-based approach for predicting fifteen distinct environmental impact categories using a comprehensive dataset from CarbonMind. This dataset includes 51,905 processes producing 791 unique molecules across 91 countries. We further integrate energy mix data from the IEA, corresponding to the same countries, to enrich our dataset. Our analysis begins by comparing the performance of QSPR and GNN models specifically for the climate change impact category, revealing that GNN models outperform the QSPR model. Furthermore, benchmarking our GNN models against existing literature for the climate change category demonstrates that our models achieve comparable performance. Expanding our approach, we developed both single- and multi-task GNN-C and GNN-E models to predict all fifteen impact categories. The results suggest that multi-task learning can enhance model performance in complex environmental impact predictions compared to single-task GNNs, considering the MRE. In summary, we recommend selecting impact categories based on their relevance to molecular structure, and excluding those with weak correlation — such as land use (LU) — when relying solely on molecular and contextual features. It is also important to examine the data distribution and address class imbalance to improve training performance. For multi-category prediction tasks, we suggest first training a multi-task GNN with molecular and geographic or energy mix features. If certain categories show poor predictive performance, dedicated single-task models can then be developed for those specific targets. Overall, the predictive performance of our computational models for impact categories remains limited. In scenarios where detailed process data is unavailable or inaccessible — such as early-stage process design, supply chain assessments, or cases involving proprietary or confidential information — the models can provide useful approximations to support preliminary decision-making.

CRediT authorship contribution statement

Qinghe Gao: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lukas Schulze Balhorn:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Alessandro Laera:** Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Raoul Meys:** Writing – review & editing, Resources, Project administration, Methodology, Data curation, Conceptualization. **Jonas Goßen:** Writing – review & editing, Validation, Software, Investigation, Formal analysis, Data curation. **Jana M. Weber:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Conceptualization. **Gregor Wernet:** Writing – review & editing, Validation, Methodology, Formal analysis. **Artur M. Schweidtmann:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT 4o in order to correct the grammar and polish the sentences. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Declaration of competing interest

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

Acknowledgments

This work was supported by the Netherlands Organisation for Scientific Research (NWO) under the Veni grant.

Appendix

See Tables 8–16.

Table 9
56 molecular descriptors from Song et al. (2017) used in QSPR models.

Descriptors abbreviation	Descriptors full name	Descriptor category
MW	Molecular weight	Constitutional indices
AMW	Average molecular weight	Constitutional indices
nBM	Number of multiple bonds	Constitutional indices
RBN	Number of rotatable bonds	Constitutional indices
nF	Number of Fluorine atoms	Constitutional indices
N%	Percentage of N atoms	Constitutional indices
O%	Percentage of O atoms	Constitutional indices
D/Dtr05	Distance/detour ring index of order 5	Ring descriptors
D/Dtr10	Distance/detour ring index of order 10	Ring descriptors
MAXDP	Maximal electrotopological positive variation	Topological indices
Psi_i_A	Intrinsic state pseudoconnectivity index - type S average	Topological indices
Yindex	Balaban Y index	Information indices
CIC4	Complementary Information Content index (neighborhood symmetry of 4-order)	Information indices
CIC5	Complementary Information Content index (neighborhood symmetry of 5-order)	Information indices
VR1_D/Dt	Randic-like eigenvector-based index from distance/detour matrix	2D matrix-based descriptors
SpDiam_B(m)	spectral diameter from Burden matrix weighted by mass	2D matrix-based descriptors
ATSC2m	Centred Broto-Moreau autocorrelation of lag 2 weighted by mass	2D autocorrelations
ATSC1p	Centred Broto-Moreau autocorrelation of lag 1 weighted by polarizability	2D autocorrelations
GATS6m	Geary autocorrelation of lag 6 weighted by mass	2D autocorrelations
GATS7s	Geary autocorrelation of lag 7 weighted by I-state	2D autocorrelations
P_VSA_LogP_1	P_VSA-like on LogP, bin 1	P_VSA-like descriptors
P_VSA_LogP_2	P_VSA-like on LogP, bin 2	P_VSA-like descriptors
P_VSA_LogP_8	P_VSA-like on LogP, bin 8	P_VSA-like descriptors
P_VSA_MR_3	P_VSA-like on Molar Refractivity, bin 3	P_VSA-like descriptors
T(N..Cl)	Sum of topological distances between N..Cl	2D Atom Pairs
T(O..F)	Sum of topological distances between O..F	2D Atom Pairs
T(O..Cl)	Sum of topological distances between O..Cl	2D Atom Pairs
T(F..Cl)	Sum of topological distances between F..Cl	2D Atom Pairs
F03[C-O]	Frequency of C - O at topological distance 3	2D Atom Pairs
F03[C-Cl]	Frequency of C - Cl at topological distance 3	2D Atom Pairs
MLOGP2	Squared Moriguchi octanol-water partition coeff. ($\log P^2$)	Molecular properties

Table 10
QSPR model configuration and training hyperparameters.

Model component	Configuration/Parameters
Input features	52 descriptors
Hidden layers	Linear(52, 16) → Linear(16, 16)
Output layer	Linear(16, 1)
Activation function	ReLU
Total trainable parameters	1137
Output	Predicted climate change category
Optimizer	Adam
Learning rate	5e−6
Batch size	20
Decreasing factor (LR scheduler)	0.9
Learning rate decay patience	10 epochs
Training epochs	800
Early stop patience	15 epochs

Table 11
GNN-M model configuration and training hyperparameters.

Model component	Configuration/Parameters
Input features	Molecular graphs
Input layer	Linear(32, 64) (2112 params)
Edge network (NN)	Linear(13, 128) + Linear(128, 4096) (530,176 params)
Graph convolution	NNConv (parameters in NN)
Recurrent layer	GRU(64, 64) (24,768 params)
Fully connected head	Linear(64, 256) → Linear(256, 128) → Linear(128, 64) → Linear(64, 1) (Total: 57,857 params)
Output	Predicted climate change category
Total trainable parameters	**614,913**
Optimizer	Adam
Learning rate	0.0001
Batch size	20
Decreasing factor (LR scheduler)	0.9
Learning rate decay patience	10 epochs
Training epochs	500
Early stop patience	15 epochs
Activation functions	ReLU (NN), ELU (GNN layers)
Number of message passing steps	3
Pooling method	Add
Weight initialization	Normal(0, 0.1)

Table 12
Single-task GNN-E model configuration and training hyperparameters.

Model component	Configuration/Parameters
Input features	Molecular graphs and energy-mix features
Input layer	Linear(32, 64) (2112 params)
Edge network (NN)	Linear(13, 128) + Linear(128, 4096) (530,176 params)
Graph convolution	NNConv (parameters in NN)
Recurrent layer	GRU(64, 64) (24,768 params)
Fully connected head	Linear(71, 256) → Linear(256, 128) → Linear(128, 64) → Linear(64, 1) (Total: 59,649 params)
Output	Predicted single impact category
Total trainable parameters	**616,705**
Optimizer	Adam
Batch size	20
Decreasing factor (LR scheduler)	0.9
Learning rate decay patience	10 epochs
Training epochs	500
Early stop patience	15 epochs
Activation functions	ReLU (NN), ELU (GNN layers)
Number of message passing steps	3
Pooling method	Add
Weight initialization	Normal(0, 0.1)

Table 13
Single-task GNN-C model configuration and training hyperparameters.

Model component	Configuration/Parameters
Input features	Molecular graphs and country categorical features
Input layer	Linear(32, 64) (2112 params)
Edge network (NN)	Linear(13, 128) + Linear(128, 4096) (530,176 params)
Graph convolution	NNConv (parameters in NN)
Recurrent layer	GRU(64, 64) (24,768 params)
Fully connected head	Linear(155, 256) → Linear(256, 128) → Linear(128, 64) → Linear(64, 1) (Total: 81,153 params)
Output	Predicted single impact category
Total trainable parameters	**638,209**
Optimizer	Adam
Batch size	20
Decreasing factor (LR scheduler)	0.9
Learning rate decay patience	10 epochs
Training epochs	500
Early stop patience	15 epochs
Activation functions	ReLU (NN), ELU (GNN layers)
Number of message passing steps	3
Pooling method	Add
Weight initialization	Normal(0, 0.1)

Table 14
Multi-task GNN-C model configuration and training hyperparameters.

Model component	Configuration/Parameters
Input features	Molecular graphs, and country categorical features
Input layer	Linear(32, 64) (2112 params)
Edge network (NN)	Linear(13, 128) + Linear(128, 4096) (530,176 params)
Graph convolution	NNConv (parameters in NN)
Recurrent layer	GRU(64, 64) (24,768 params)
Multi-task output head	15 MLPs with shared structure: Linear(155, 256) → Linear(256, 128) → Linear(128, 64) → Linear(64, 1) (Total: 1,217,295 params)
Output	Predicted all 15 impact categories
Total trainable parameters	**1,774,351**
Optimizer	Adam
Learning rate	0.001
Batch size	20
Decreasing factor (LR scheduler)	0.9
Learning rate decay patience	10 epochs
Training epochs	500
Early stop patience	15 epochs
Activation functions	ReLU (edge net), ELU (GNN + MLPs)
Number of message passing steps	3
Pooling method	Add
Weight initialization	Normal(0, 0.1)

Table 15
Multi-task GNN-E model configuration and training hyperparameters.

Model component	Configuration/Parameters
Input features	Molecular graphs and energy-mix features
Input layer	Linear(32, 64) (2112 params)
Edge network (NN)	Linear(13, 128) + Linear(128, 4096) (530,176 params)
Graph convolution	NNConv (parameters in NN)
Recurrent layer	GRU(64, 64) (24,768 params)
Multi-task output head	15 MLPs with shared structure: Linear(71, 256) → Linear(256, 128) → Linear(128, 64) → Linear(64, 1) (Total: 894,735 params)
Output	Predicted all 15 impact categories
Total trainable parameters	**1,451,791**
Optimizer	Adam
Learning rate	0.001
Batch size	20
Decreasing factor (LR scheduler)	0.9
Learning rate decay patience	10 epochs
Training epochs	500
Early stop patience	15 epochs
Activation functions	ReLU (edge net), ELU (GNN + MLPs)
Number of message passing steps	3
Pooling method	Add
Weight initialization	Normal(0, 0.1)

Table 16
Category-specific learning rates for single-task GNN-C and GNN-E models.

Impact category	GNN-C	GNN-E
AC	1.00E–03	5.00E–04
CC	5.00E–05	5.00E–04
ECO	5.00E–04	1.00E–04
ER	5.00E–04	1.00E–04
EUf	1.00E–03	1.00E–03
EUm	5.00E–04	5.00E–04
EUt	5.00E–04	5.00E–04
HT	1.00E–03	1.00E–03
IR	1.00E–04	1.00E–03
LU	1.00E–04	1.00E–04
MR	1.00E–04	1.00E–03
OD	5.00E–04	5.00E–04
PMF	5.00E–04	5.00E–05
POF	5.00E–05	5.00E–05
WU	5.00E–04	5.00E–04

Data availability

The data that has been used is confidential.

References

Alshehri, A.S., Tula, A.K., You, F., Gani, R., 2021. Next generation pure component property estimation models: With and without machine learning techniques. AIChE J. 68 (6), http://dx.doi.org/10.1002/aic.17469.

Alyaseri, I., Zhou, J., 2019. Handling uncertainties inherited in life cycle inventory and life cycle impact assessment method for improved life cycle assessment of wastewater sludge treatment. Heliyon 5 (11), e02793. http://dx.doi.org/10.1016/j.heliyon.2019.e02793.

Baxevanidis, P., Papadokonstantakis, S., Kokossis, A., Marcoulaki, E., 2021. Group contribution-based LCA models to enable screening for environmentally benign novel chemicals in CAMD applications. AIChE J. 68 (3), http://dx.doi.org/10.1002/aic.17544.

Benson, S.W., Cruickshank, F.R., Golden, D.M., Haugen, G.R., O'Neal, H.E., Rodgers, A.S., Shaw, R., Walsh, R., 1969. Additivity rules for the estimation of thermochemical properties. Chem. Rev. 69 (3), 279–324. http://dx.doi.org/10.1021/cr60259a002.

Blanco, C., Pauliks, N., Donati, F., Engberg, N., Weber, J., 2024. Machine learning to support prospective life cycle assessment of emerging chemical technologies. Curr. Opin. Green Sustain. Chem. 50, 100979. http://dx.doi.org/10.1016/j.cogsc.2024.100979.

Buchner, G.A., Stepputat, K.J., Zimmermann, A.W., Schomäcker, R., 2019. Specifying technology readiness levels for the chemical industry. Ind. Eng. Chem. Res. 58 (17), 6957–6969. http://dx.doi.org/10.1021/acs.iecr.8b05693.

Burkardt, P., Fleischmann, M., Wegmann, T., Braun, M., Knöll, J., Schumacher, L., vom Lehn, F., Lehrheuer, B., Meinke, M., Pitsch, H., Kneer, R., Schröder, W., Pischinger, S., 2021. On the use of active pre-chambers and bio-hybrid fuels in internal combustion engines. In: Engines and Fuels for Future Transport. Springer Singapore, pp. 205–231. http://dx.doi.org/10.1007/978-981-16-8717-4_9.

Buterez, D., Janet, J.P., Kiddle, S.J., Oglic, D., Lió, P., 2024. Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. Nat. Commun. 15 (1), http://dx.doi.org/10.1038/s41467-024-45566-8.

Calvo-Serrano, R., González-Miquel, M., Guillén-Gosálbez, G., 2018a. Integrating COSMO-based σ-profiles with molecular and thermodynamic attributes to predict the life cycle environmental impact of chemicals. ACS Sustain. Chem. Eng. 7 (3), 3575–3583. http://dx.doi.org/10.1021/acssuschemeng.8b06032.

Calvo-Serrano, R., González-Miquel, M., Papadokonstantakis, S., Guillén Gosálbez, G., 2017. Cradle-to-gate environmental impact prediction from chemical attributes using mixed-integer programming. In: 27th European Symposium on Computer Aided Process Engineering. Elsevier, pp. 1999–2004. http://dx.doi.org/10.1016/b978-0-444-63965-3.50335-4.

Calvo-Serrano, R., González-Miquel, M., Papadokonstantakis, S., Guillén-Gosálbez, G., 2018b. Predicting the cradle-to-gate environmental impact of chemicals from molecular descriptors and thermodynamic properties via mixed-integer programming. Comput. Chem. Eng. 108, 179–193. http://dx.doi.org/10.1016/j.compchemeng.2017.09.010.

Chen, G., Song, Z., Qi, Z., Sundmacher, K., 2023. A scalable and integrated machine learning framework for molecular properties prediction. AIChE J. 69 (10), http://dx.doi.org/10.1002/aic.18185.

Chen, X., Zhang, Z.-J., Hong, X., Ackermann, L., 2025. Integrating a multitask graph neural network with DFT calculations for site-selectivity prediction of arenes and mechanistic knowledge generation. Nat. Synth. http://dx.doi.org/10.1038/s44160-025-00770-2.

Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todeschini, R., Consonni, V., Kuz'min, V.E., Cramer, R., Benigni, R., Yang, C., Rathman, J., Terfloth, L., Gasteiger, J., Richard, A., Tropsha, A., 2014. QSAR modeling: Where have you been? Where are you going to? J. Med. Chem. 57 (12), 4977–5010. http://dx.doi.org/10.1021/jm4004285.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. http://dx.doi.org/10.48550/ARXIV.1406.1078.

Dahl, G.E., Jaitly, N., Salakhutdinov, R., 2014. Multi-task neural networks for QSAR predictions. http://dx.doi.org/10.48550/ARXIV.1406.1231.

European Commission, 2021. Commission recommendation (EU) 2021/2279 of 15 december 2021 on the use of the environmental footprint methods to measure and communicate the life cycle environmental performance of products and organisations. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32021H2279; Official Journal of the European Union, L 471, 30.12.2021, pp. 1–396.

Fazio, S., Garraín, D., Mathieux, F., De la Rúa, C., Recchioni, M., Lechón, Y., 2015. Method applied to the background analysis of energy data to be considered for the European reference life cycle database (ELCD). SpringerPlus 4 (1), http://dx.doi.org/10.1186/s40064-015-0914-x.

Gani, R., 2019. Group contribution-based property estimation methods: Advances and perspectives. Curr. Opin. Chem. Eng. 23, 184–196. http://dx.doi.org/10.1016/j.coche.2019.04.007.

Gao, Q., Dukker, T., Schweidtmann, A.M., Weber, J.M., 2024. Self-supervised graph neural networks for polymer property prediction. Mol. Syst. Des. Eng. 9 (11), 1130–1143. http://dx.doi.org/10.1039/d4me00088a.

Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E., 2017. Neural message passing for quantum chemistry. http://dx.doi.org/10.48550/ARXIV.1704.01212.

Hamilton, W., Ying, Z., Leskovec, J., 2017. Inductive representation learning on large graphs. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc..

Hammett, L.P., 1935. Some relations between reaction rates and equilibrium constants.. Chem. Rev. 17 (1), 125–136. http://dx.doi.org/10.1021/cr60056a010.

Heidari, M., Mathis, D., Blanchet, P., Amor, B., 2019. Streamlined life cycle assessment of an innovative bio-based material in construction: A case study of a phase change material panel. Forests 10 (2), 160. http://dx.doi.org/10.3390/f10020160.

Hu, F., Chen, D., Liu, Q., Wu, S., 2025. Improving multi-task GNNs for molecular property prediction via missing label imputation. Mach. Intell. Res. 22 (1), 131–144. http://dx.doi.org/10.1007/s11633-023-1443-7.

Huijbregts, M.A.J., Steinmann, Z.J.N., Elshout, P.M.F., Stam, G., Verones, F., Vieira, M., Zijp, M., Hollander, A., van Zelm, R., 2016. ReCiPe2016: A harmonised life cycle impact assessment method at midpoint and endpoint level. Int. J. Life Cycle Assess. 22 (2), 138–147. http://dx.doi.org/10.1007/s11367-016-1246-y.

Joback, K.G., Reid, R.C., 1987. Estimation of pure-component properties from group-contributions. Chem. Eng. Commun. 57 (1–6), 233–243. http://dx.doi.org/10.1080/00986448708960487.

Karka, P., Papadokonstantakis, S., Kokossis, A., 2022. Digitizing sustainable process development: From ex-post to ex-ante LCA using machine-learning to evaluate bio-based process technologies ahead of detailed design. Chem. Eng. Sci. 250, 117339. http://dx.doi.org/10.1016/j.ces.2021.117339.

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 (10), 5714–5789. http://dx.doi.org/10.1021/cr900238d.

Kim, T., Benavides, P.T., Kneifel, J.D., Beers, K.L., Hawkins, T.R., 2023. Cross-database comparisons on the greenhouse gas emissions, water consumption, and fossil-fuel use of plastic resin production and their post-use phase impacts. Resour. Conserv. Recycl. 198, 107168. http://dx.doi.org/10.1016/j.resconrec.2023.107168.

Kleinekorte, J., Fleitmann, L., Bachmann, M., Kätelhön, A., Barbosa-Póvoa, A., von der Assen, N., Bardow, A., 2020. Life cycle assessment for the design of chemical processes, products, and supply chains. Annu. Rev. Chem. Biomol. Eng. 11 (1), 203–233. http://dx.doi.org/10.1146/annurev-chembioeng-011520-075844.

Kleinekorte, J., Kleppich, J., Fleitmann, L., Beckert, V., Blodau, L., Bardow, A., 2023. Appropriate life cycle assessment: A process-specific, predictive impact assessment method for emerging chemical processes. ACS Sustain. Chem. Eng. 11 (25), 9303–9319. http://dx.doi.org/10.1021/acssuschemeng.2c07682.

Kleinekorte, J., Kröger, L., Leonhard, K., Bardow, A., 2019. A neural network-based framework to predict process-specific environmental impacts. In: 29th European Symposium on Computer Aided Process Engineering. Elsevier, pp. 1447–1452. http://dx.doi.org/10.1016/b978-0-12-818634-3.50242-3.

Landrum, G., Tosco, P., Kelley, B., Ric, Cosgrove, D., Sriniker, Vianello, R., Gedeck, Schneider, N., Jones, G., Kawashima, E., Nealschneider, D., Dalke, A., Cole, B., Swain, M., Turk, S., Savelev, A., Vaucher, A., Wójcikowski, M., Take, I., Scalfani, V.F., Probst, D., Ujihara, K., Walker, R., Godin, G., Pahl, A., Lehtivarjo, J., Berenger, F., Strets123, Biggs, J.D., 2024. rdkit/rdkit: 2023_09_6 (Q3 2023) release. http://dx.doi.org/10.5281/zenodo.591637.

Liu, R., Liu, Y., Duan, J., Hou, F., Wang, L., Zhang, X., Li, G., 2022. Ensemble learning directed classification and regression of hydrocarbon fuels. Fuel 324, 124520. http://dx.doi.org/10.1016/j.fuel.2022.124520.

Martínez-Rocamora, A., Solís-Guzmán, J., Marrero, M., 2016. LCA databases focused on construction materials: A review. Renew. Sustain. Energy Rev. 58, 565–573. http://dx.doi.org/10.1016/j.rser.2015.12.243.

Mattila, T.J., 2017. Use of input–output analysis in LCA. In: Life Cycle Assessment. Springer International Publishing, pp. 349–372. http://dx.doi.org/10.1007/978-3-319-56475-3_14.

Mauri, A., Consonni, V., Pavan, M., Todeschini, R., et al., 2006. Dragon software: An easy approach to molecular descriptor calculations. Match 56 (2), 237–248.

Meng, M., Wei, Z., Li, Z., Jiang, M., Bian, Y., 2019. Property prediction of molecules in graph convolutional neural network expansion. In: 2019 IEEE 10th International Conference on Software Engineering and Service Science. ICSESS, IEEE, http://dx.doi.org/10.1109/icsess47205.2019.9040723.

Minten, H., Vandegehuchte, B.D., Jaumard, B., Meys, R., Reinert, C., Bardow, A., 2024. Early-stage impact assessment tool (ESTIMATe) for the life cycle assessment of CO2-based chemicals. Green Chem. 26 (15), 8728–8743. http://dx.doi.org/10.1039/d4gc00964a.

Nakamura, S., Nansai, K., 2016. Input–output and hybrid LCA. In: Special Types of Life Cycle Assessment. Springer Netherlands, pp. 219–291. http://dx.doi.org/10.1007/978-94-017-7610-3_6.

Parvatker, A.G., Eckelman, M.J., 2018. Comparative evaluation of chemical life cycle inventory generation methods and implications for life cycle assessment results. ACS Sustain. Chem. Eng. 7 (1), 350–367. http://dx.doi.org/10.1021/acssuschemeng.8b03656.

Pope, P., Kolouri, S., Rostrami, M., Martin, C., Hoffmann, H., 2018. Discovering molecular functional groups using graph convolutional neural networks. http://dx.doi.org/10.48550/ARXIV.1812.00265.

Preuss, N., Alshehri, A.S., You, F., 2024. Large language models for life cycle assessments: Opportunities, challenges, and risks. J. Clean. Prod. 466, 142824. http://dx.doi.org/10.1016/j.jclepro.2024.142824.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., Pande, V., 2015. Massively multitask networks for drug discovery. http://dx.doi.org/10.48550/ARXIV.1502.02072.

Rittig, J.G., Ben Hicham, K., Schweidtmann, A.M., Dahmen, M., Mitsos, A., 2023. Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids. Comput. Chem. Eng. 171, 108153. http://dx.doi.org/10.1016/j.compchemeng.2023.108153.

Rittig, J.G., Gao, Q., Dahmen, M., Mitsos, A., Schweidtmann, A.M., 2022. Graph neural networks for the prediction of molecular structure-property relationships. http://dx.doi.org/10.48550/ARXIV.2208.04852.

Ruder, S., 2017. An overview of multi-task learning in deep neural networks. http://dx.doi.org/10.48550/ARXIV.1706.05098.

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 (9), 11395–11407. http://dx.doi.org/10.1021/acs.energyfuels.0c01533.

Schweidtmann, A.M., Rittig, J.G., Weber, J.M., Grohe, M., Dahmen, M., Leonhard, K., Mitsos, A., 2023. Physical pooling functions in graph neural networks for molecular property prediction. Comput. Chem. Eng. 172, 108202. http://dx.doi.org/10.1016/j.compchemeng.2023.108202.

Song, R., Keller, A.A., Suh, S., 2017. Rapid life-cycle impact screening using artificial neural networks. Environ. Sci. Technol. 51 (18), 10777–10785. http://dx.doi.org/10.1021/acs.est.7b02862.

Standard, I., 2006. Environmental Management-Life Cycle Assessment-Requirements and Guidelines. ISO, London.

Stellner, L., Kätehön, A., Vögler, O., Hermanns, R., Suh, S., Bardow, A., Meys, R., 2022. Methodology cm.chemicals. Carbon Minds GmbH, Cologne.

Sultan, A., Rausch-Dupont, M., Khan, S., Kalinina, O., Volkamer, A., Klakow, D., 2025. Transformers for molecular property prediction: Domain adaptation efficiently improves performance. http://dx.doi.org/10.48550/ARXIV.2503.03360.

Sultan, A., Sieg, J., Mathea, M., Volkamer, A., 2024. Transformers for molecular property prediction: Lessons learned from the past five years. http://dx.doi.org/10.48550/ARXIV.2404.03969.

Sun, Y., Wang, X., Ren, N., Liu, Y., You, S., 2022. Improved machine learning models by data processing for predicting life-cycle environmental impacts of chemicals. Environ. Sci. Technol. 57 (8), 3434–3444. http://dx.doi.org/10.1021/acs.est.2c04945.

Todeschini, R., Consonni, V., 2000. Handbook of molecular descriptors. Wiley, http://dx.doi.org/10.1002/9783527613106,

Trivedi, D., Patrikar, K., Mondal, A., 2024. Graph-based networks for accurate prediction of ground and excited state molecular properties from minimal features. Mol. Syst. Des. Eng. 9 (12), 1275–1284. http://dx.doi.org/10.1039/d4me00113c.

Ulonska, K., Skiborowski, M., Mitsos, A., Viell, J., 2016. Early-stage evaluation of biorefinery processing pathways using process network flux analysis. AIChE J. 62 (9), 3096–3108. http://dx.doi.org/10.1002/aic.15305.

Wang, H., Zhang, A., Zhong, Y., Tang, J., Zhang, K., Li, P., 2024. Chain-aware graph neural networks for molecular property prediction. In: Wren, J. (Ed.), Bioinformatics 40 (10), http://dx.doi.org/10.1093/bioinformatics/btae574.

Weber, J.M., Guo, Z., Lapkin, A.A., 2022. Discovering circular process solutions through automated reaction network optimization. ACS Eng. Au 2 (4), 333–349. http://dx.doi.org/10.1021/acsengineeringau.2c00002.

Weber, J.M., Guo, Z., Zhang, C., Schweidtmann, A.M., Lapkin, A.A., 2021. Chemical data intelligence for sustainable chemistry. Chem. Soc. Rev. 50 (21), 12013–12036. http://dx.doi.org/10.1039/d1cs00477h.

Weininger, D., 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 28 (1), 31–36. http://dx.doi.org/10.1021/ci00057a005.

Wernet, G., Bauer, C., Steubing, B., Reinhard, J., Moreno-Ruiz, E., Weidema, B., 2016. The ecoinvent database version 3 (part I): Overview and methodology. Int. J. Life Cycle Assess. 21 (9), 1218–1230. http://dx.doi.org/10.1007/s11367-016-1087-8.

Wernet, G., Hellweg, S., Fischer, U., Papadokonstantakis, S., Hungerbühler, K., 2008. Molecular-structure-based models of chemical inventories using neural networks. Environ. Sci. Technol. 42 (17), 6717–6722. http://dx.doi.org/10.1021/es7022362.

Wernet, G., Papadokonstantakis, S., Hellweg, S., Hungerbühler, K., 2009. Bridging data gaps in environmental assessments: Modeling impacts of fine and basic chemical production. Green Chem. 11 (11), 1826. http://dx.doi.org/10.1039/b905558d.

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. http://dx.doi.org/10.1016/j.ddtec.2020.11.009.

Withnall, M., Lindelöf, E., Engkvist, O., Chen, H., 2020. Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction. J. Cheminformatics 12 (1), http://dx.doi.org/10.1186/s13321-019-0407-y.

Yang, Y., Heijungs, R., Brandão, M., 2017. Hybrid life cycle assessment (LCA) does not necessarily yield more accurate results than process-based LCA. J. Clean. Prod. 150, 237–242. http://dx.doi.org/10.1016/j.jclepro.2017.03.006.

Zhang, S., Liu, Y., Xie, L., 2020. Molecular mechanics-driven graph neural network with multiplex graph for molecular structures. http://dx.doi.org/10.48550/ARXIV.2011.07457.

Zhang, D., Wang, Z., Oberschelp, C., Bradford, E., Hellweg, S., 2024. Enhanced deep-learning model for carbon footprints of chemicals. ACS Sustain. Chem. Eng. 12 (7), 2700–2708. http://dx.doi.org/10.1021/acssuschemeng.3c07038.