



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/232552/>

Version: Accepted Version

Article:

Chen, B., Chen, H., Quan, Z. et al. (2025) SemaNet: Bridging words and numbers for predicting missing environmental data in life cycle assessment. *Environmental Science & Technology*, 59 (39). pp. 21131-21146. ISSN: 0013-936X

<https://doi.org/10.1021/acs.est.5c07557>

This document is the Accepted Manuscript version of a Published Work that appeared in final form in *Environmental Science & Technology*, copyright © American Chemical Society after peer review and technical editing by the publisher. To access the final edited and published work see <https://doi.org/10.1021/acs.est.5c07557>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

SemaNet: Bridging Words and Numbers for Predicting Missing Environmental Data in Life Cycle Assessment

Bin Chen,[†] Hong Chen,[‡] Zhishan Quan,[¶] Wei He,[§] Visakan Kadirkamanathan,[†]
Jose L. Casamayor,^{||} and Wei W. Xing^{*,[⊥]}

[†]*School of Electrical and Electronic Engineering, University of Sheffield, Sheffield, S1 4DT,
United Kingdom*

[‡]*City University of Hong Kong, Tat Chee Avenue, Hong Kong, 999077, China*

[¶]*College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen,
518054, China*

[§]*Department of Computer Science, University of Exeter, Exeter, EX4 4QF, United
Kingdom*

^{||}*Advanced Manufacturing Research Centre, University of Sheffield, Rotherham, S60 5TZ,
United Kingdom*

[⊥]*School of Mathematical and Physical Sciences, University of Sheffield, Sheffield, S3 7RH,
United Kingdom*

E-mail: w.xing@sheffield.ac.uk

Abstract

Life Cycle Assessment (LCA) is one of the most used methodologies for evaluating environmental impact, but its effective application is severely limited by missing data, an issue that existing methods for Life Cycle Inventory (LCI) data completion cannot

address effectively. This paper proposes a paradigm shift: rather than depending exclusively on numerical correlations, we leverage the extensive contextual information inherent in process descriptions via pre-trained language models, establishing a semantic bridge between qualitative descriptions and quantitative environmental flows. Our semantic-based neural network framework, SemaNet, achieves superior performance in predicting missing LCI values, surpassing existing state-of-the-art methods in various evaluation metrics. The results are significant: while existing approaches fail completely under high data sparsity, our method achieves high accuracy even with 100% missing numerical data while reducing computational requirements by 99% through the use of semantic filtering. This new method for LCI data completion significantly reduces the data collection efforts and time for LCA practitioners, making reliable and faster environmental impact assessment feasible, even when primary data does not exist, thus facilitating reliable sustainability assessment across industrial sectors.

keywords: Life Cycle Assessment, Semantic Learning, Neural Network Method, Data Efficiency, Life Cycle Inventory

Synopsis

Our framework enables reliable environmental prediction even without life cycle inventory data, improving the speed, cost, and feasibility of life cycle assessments across sectors.

1. Introduction

Life Cycle Assessment (LCA) has been recognised as one of the most established methodologies for thoroughly evaluating environmental impacts, offering essential insights across the entire product life cycle, from the extraction of raw materials to disposal at the end of life.¹ As the challenges of global sustainability become more intense, LCA has become indispensable for informed decision-making in diverse domains—guiding sustainable manufacturing prac-

tices, evaluating emerging low-carbon technologies, and optimizing environmental resource management.² For example, LCA facilitates a thorough exploration of the environmental trade-offs associated with renewable energy systems, revealing the hidden impacts in solar photovoltaic technology, which range from the extraction of materials to the production of panels and their eventual recycling.³

The foundation of a reliable LCA is the creation of a complete and accurate Life Cycle Inventory (LCI), which provides a comprehensive record of all the input and output flows across the entire lifecycle of a product or process.¹ However, constructing a comprehensive LCI presents a fundamental challenge: industrial process data often do not exist, and when they do, the collection of these data is very time-consuming and resource-intensive, requiring meticulous investigation and extensive laboratory testing.^{4,5} This challenge is further complicated by the prevalent practice of data withholding in industrial settings, where process data frequently constitutes significant intellectual property. While alternative approaches such as using proxy datasets and data extrapolation have been proposed,^{6,7} they bring with them significant uncertainty—studies have shown order-of-magnitude discrepancies between extrapolated estimates and direct measurements.^{8,9} This LCI data completion problem creates a contradictory scenario: while accurate, reliable and fast environmental impact assessments are becoming increasingly essential, they remain considerably constrained by the scarcity of available LCI data.

A possible way to address this challenge is to formulate the LCI data completion problem as a matrix completion task.¹⁰ In the matrix completion literature, traditional methods usually introduce mathematical and statistical techniques to predict missing values in structured datasets. For instance, Mazumder et al.¹¹ developed spectral regularisation algorithms that utilise low-rank structure via a soft-thresholded singular value decomposition, providing stable convergence for incomplete matrices. Soni et al. (2017) studied the noisy matrix completion problem, by using sparsity-regularized maximum likelihood estimation to recover matrices from different noise models. Jin et al.¹² introduced $\ell_{2,1}$ -norm regularised

sparse matrix factorisation techniques for robust data completion of matrices with noise or corrupted observations. Petrov et al.¹³ proposed a matrix completion framework that can model sparse measurement errors, employing a two-mask iterative update strategy alongside canonical-angle based subspace alignment to jointly recover the low-rank matrix and the sparse error component.

Matrix-oriented methods have inspired early attempts in the LCA domain. For instance, Liu et al.¹⁴ introduced a multiple imputation method to address high missing data in the steel plant’s inventory datasets, providing stable estimates even under non-linear and high-uncertainty conditions. Moreau et al.¹⁵ used kriging (a geostatistical interpolation method) to estimate absent LCI values in the hydroelectric power plant’s dataset. Hou et al.¹⁰ formulated the LCI gap-filling task as a similarity-based completion problem, leveraging structural regularities in unit process matrices and constructing LCI data completion strategies based on statistical similarity and data availability.

These early attempts are primarily based on traditional statistical or optimization frameworks, which rely on structured matrix representations and domain-specific assumptions. More recently, researchers in the artificial intelligence field have begun exploring machine learning techniques to forecast missing LCI data more flexibly and accurately.^{16–24} Significant contributions encompass the application of neural networks for predicting the effects of chemical production^{16,17} and the development of machine learning models aimed at estimating the ecotoxicity factors for a range of chemicals.¹⁹ Building on this trend, Zhao et al.²⁴ developed the extreme gradient boosting (XGBoost) algorithm to tackle the LCI data filling task, which achieves superior prediction performance with up to 20% of the data missing.

Despite these advances, a critical limitation remains: these approaches rely solely on numerical correlations among observed flows, entirely neglecting the important contextual information that is inherent in process descriptions. These descriptions contain technical specifications, geographical contexts, and operational conditions, which represent a significant amount of domain knowledge that has the potential to enhance prediction models. The

core idea is that process descriptions contain latent semantic patterns that can assist the prediction process for missing data even with few observed data, allowing for precise predictions even in situations of high data scarcity. While recent explorations like the Flamingo framework²⁵ have initiated the application of semantic information to LCI-related tasks such as automated environmental impact factor matching, they have yet to address the fundamental LCI data completion challenge of predicting missing unit process data, which remains a significant bottleneck in achieving reliable and fast LCAs at scale.

This paper presents a novel framework, SemaNet, which utilizes a semantic-based neural network to significantly improve the performance of predicting unit process data in LCA. Our approach bridges the link between textual process descriptions and numerical flow data, allowing for more precise predictions even in situations characterized by extreme data sparsity. The main contributions of this work are:

1. We introduce the first framework that bridges textual process descriptions with numerical flow data, enabling LCA prediction results of higher accuracy. By selecting only semantically relevant neighbors, our approach achieves superior results with just 1% of the data traditionally required.
2. We develop a unified neural architecture that jointly optimizes classification and regression, eliminating error propagation issues found in two-stage methods. This approach reduces computational time from hours to minutes compared to the state-of-the-art approaches.
3. Notably, we demonstrate data prediction capability in true zero-shot scenarios—where only process names are available with no flow data. This feature allows the LCA practitioners to perform preliminary life cycle assessments where other existing LCI completion methods completely fail, fundamentally transforming the potential for forecasting environmental data deficiencies.
4. We provide an open-source Ecoinvent data protocol that automatically extracts both

textual and numerical process data, with every step publicly inspectable. This guarantees that any researcher can rerun the same workflow on the same Ecoinvent release and obtain identical results, accelerating the adoption of semantically enhanced LCA methods.

The rest of this paper is organized as follows: Section 2 formulates the unit process data prediction problem and introduces our semantic reformulation; Section 3 provides our SemaNet framework and describes the experimental dataset; Section 4 presents comprehensive experimental results demonstrating our approach’s performance, and discusses key implications, limitations and avenues for future work.

2. Materials

This section establishes how unit process data is structured, examines traditional prediction approaches, and introduces our semantic reformulation.

2.1. LCA Data Structure and Matrix Formulation

LCI data is naturally represented as a matrix where columns correspond to unit processes and rows represent flows. Each unit process contains a structured name with: (1) a Reference Product Name, (2) a Geography Code, and (3) an Activity Name. Each flow represents either an intermediate exchange between processes or an elementary exchange with the environment. Table 1 illustrates this structure using electricity production as an example. The process “Electricity, high voltage//[AT] electricity production, hard coal” produces 1 kWh of electricity while consuming 0.386 kg of hard coal and emitting 0.804 kg of CO₂ and 0.0456 m³ of water.

Mathematically, we represent the LCI database as a matrix $\mathbf{Y} \in \mathbb{R}^{n \times m}$ with n processes and m flows. In practice, this matrix is extremely sparse—the Ecoinvent 3.11 database contains 26,034 processes and 29,271 flows with only approximately 0.22% non-zero entries.

This indicates that only about 0.22% of all possible process-flow combinations have a non-zero value, indicating an actual environmental exchange or flow between the corresponding process and flow, while the remaining 99.78% indicate no direct interaction. The sparsity reflects both the natural structure of industrial processes (each process uses only a small subset of all possible flows) and the widespread missing data problem in LCA.

2.2. Traditional Numerical Prediction Approaches

The state-of-the-art approach for predicting missing LCI data, developed by Zhao et al.,²⁴ formulates this as a matrix completion problem. For any process i , we define $\mathcal{K}_i$ as the set of known flow indices and $\mathcal{U}_i$ as the set of unknown flow indices that require prediction. The traditional prediction task employs a function:

$$\hat{y}_{ij} = f(\{y_{ik} \mid k \in \mathcal{K}_i\}, j), \quad \text{for each } j \in \mathcal{U}_i \quad (1)$$

This function learns patterns from known values to predict missing ones. While effective with limited missing data ($\leq 20\%$), this approach deteriorates rapidly as data sparsity increases and completely fails when data is 100% missing. The fundamental limitation is that numerical-only methods cannot recognize functional similarities between processes that share flow patterns based on their technical, geographical, or operational characteristics.

Table 1: Example Life Cycle Inventory Matrix (AT: Austria)

Flow Type	Electricity Production	Hard Coal Supply	...
	<i>Electricity, high voltage//[AT] electricity production, hard coal</i>	<i>Hard coal//[AT] market for hard coal</i>	
Electricity, high voltage (kWh)	1.000	0	...
Hard coal (kg)	0.386	1.000	...
CO ₂ emissions (kg)	0.804	0	...
Water emissions (m ³)	0.0456	0	...
...	...	...	...

2.3. Semantic Reformulation for Enhanced Prediction

To address these limitations, we propose integrating semantic information from process descriptions into the prediction function:

$$\hat{y}_{ij} = f^{\text{sem}}(\{y_{ik} \mid k \in \mathcal{K}_i\}, j, \mathbf{e}_i), \quad \text{for each } j \in \mathcal{U}_i, \quad (2)$$

where $\mathbf{e}_i$ represents semantic embeddings derived from the process i 's textual description.

This reformulation fundamentally transforms the prediction task in three critical aspects. First, it expands the information space by incorporating not just numerical correlations but also the contextual knowledge embedded in process descriptions. Second, it facilitates the identification of functional similarities among processes that share related descriptions (e.g., different photovoltaic production methods). Third and most importantly, it allows zero-shot prediction, i.e., generating estimates for processes with no observed flow data, relying solely on their textual descriptions.

The semantic embeddings effectively establish a connection between the qualitative domain knowledge encoded in process descriptions and the quantitative flow data. This approach improves both accuracy and interpretability, especially in situations of extreme data sparsity where methods relying solely on numerical data are inadequate.

The following section details our implementation of this semantic prediction approach, demonstrating its influence on LCI data prediction under different levels of data availability.

3. Methods

To implement the semantic-enhanced prediction idea outlined in Section 2, we developed a framework that bridges textual process descriptions with numerical flow data (as illustrated in Figure 1). The SemaNet framework addresses these challenges by recognizing that processes with semantic similarities frequently exhibit comparable flow patterns, despite

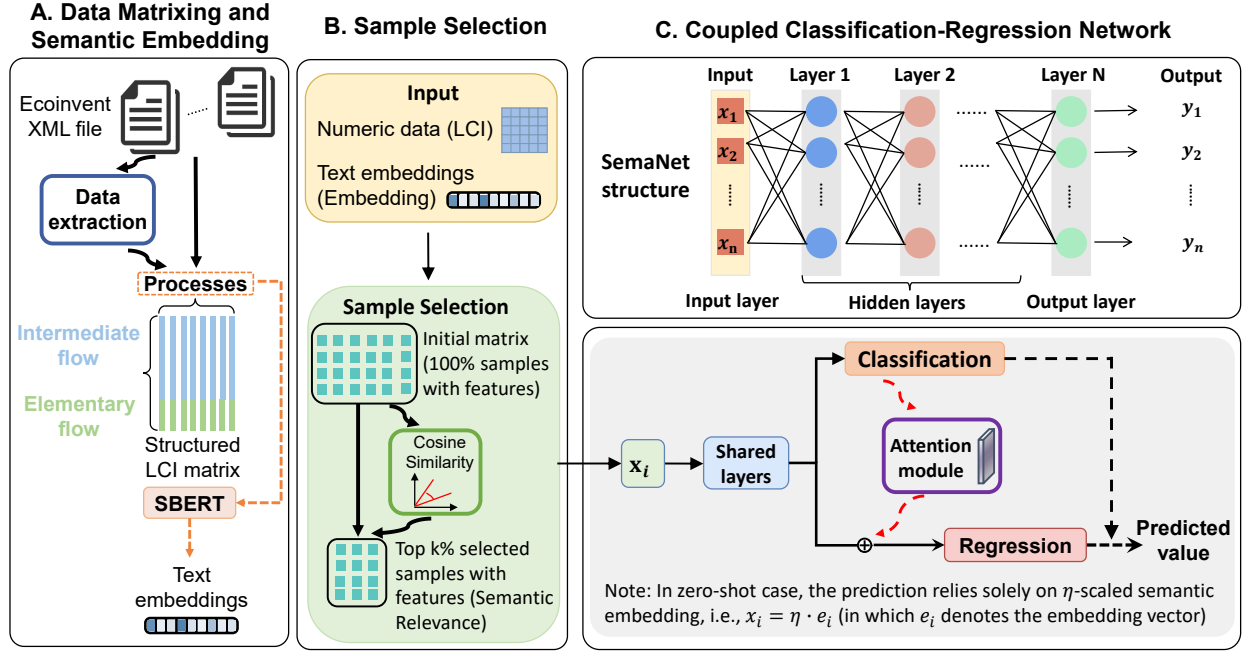


Figure 1: Architecture of the Semantic-based Neural Network (SemaNet) framework showing the process flow from semantic embedding through sample selection and feature fusion to coupled prediction.

variations in their numerical profiles. This insight allows us to make accurate predictions even in cases where conventional methods fail completely. The framework consists of four integrated components: semantic embedding of process descriptions, semantic-based sample selection, feature processing and fusion, and a coupled neural architecture that jointly handles classification and regression tasks.

3.1. Semantic Embedding of Process Descriptions

We employed the Sentence-Transformers (SBERT) framework²⁶ to convert textual process descriptions into dense vector representations. The SBERT framework offers a variety of pre-trained models with differing trade-offs in terms of semantic accuracy, computational cost, and embedding dimensionality¹.

¹A full list of available models can be found at: https://www.sbert.net/docs/sentence_transformer/pretrained_models.html

After preliminary evaluation, we selected ‘all-MiniLM-L6-v2’ as the sentence embedding model. Larger models like "all-mpnet-base-v2" have slightly better benchmark performance, but they produce 768-dimensional embeddings and use a lot more memory and time. Given that our framework needs to embed and compare more than 20 thousand unit process descriptions efficiently, these heavier models proved impractical in our context. In contrast, ‘all-MiniLM-L6-v2’ provides a favorable balance between speed and semantic quality. It generates compact 384-dimensional embeddings by default and contains only 22 million parameters, making it approximately five times faster than ‘all-mpnet-base-v2’ in practice. During our validation tests on LCI-related texts, this model consistently produced stable and meaningful semantic representations with minimal overhead.

For LCA processes with their structured naming convention (e.g., "electricity, high voltage//[AT] electricity production, hard coal"), we separately embedded the Reference Product, Geography, and Activity Name components. These three 384-dimensional vectors were then concatenated to form a 1152-dimensional semantic representation for each process.

This method effectively captures crucial technical and contextual relationships among processes. For example, when we analyzed photovoltaic processes from different regions using t-distributed stochastic neighbor embedding (t-SNE) visualization, processes with similar technical characteristics (like "electricity production, photovoltaic, 3kWp slanted-roof installation, multi-Si") clustered together despite geographical differences (CN-BJ versus CN-CQ). The average similarity among processes within semantic clusters was found to be high (e.g., ≥ 0.8010), in contrast to a lower between-cluster similarity (e.g., ≤ 0.4353). This quantitative evidence shows that our embeddings effectively reflect meaningful process relationships.

Employing sentence-level embedding offers significant advantages over word-level approaches in the analysis of LCA data. It preserves the contextual importance of specialized terms and reflects underlying operational similarities that might be missed in simplified representations. The understanding of semantical information is especially important when predicting flow data in situations where there is little to no observed data. In such cases,

the model must depend largely on textual information to identify relevant patterns.

3.2. Semantic-Based Sample Selection

One of the significant advancements of SemaNet is the careful selection of training data based on semantic similarity. Traditionally, the methodology in machine learning has relied on the use of all available data for training purposes. However, our findings suggest that this setting may hinder performance in predicting LCI data, specifically resulting in decreased prediction accuracy and increased computational complexity. This is because incorporating unrelated or semantically distant processes introduces irrelevant variations and noise, making it challenging for the model to identify meaningful and generalizable patterns relevant to the target LCI predictions.

To implement this selective training strategy, we leverage the semantic embedding vector (that is 1152 dimensions and has already been computed for each unit process description) to compute the similarity score between the target embedding and each candidate training sample. An intuitive way to obtain this similarity score is by calculating traditional distance metrics, such as Euclidean or Manhattan distances. However, these traditional distance metrics are normally ineffective in high-dimensional spaces, as most vectors in high-dimensional spaces have similar norms and closer distances.

Cosine similarity is a more advanced metric for the similarity score, as it focuses more on the angular difference between vectors rather than the magnitude difference. Benefiting from this characteristic, it is more semantically meaningful for comparing dense vector representations. Cosine similarity has become common with the rise of embedding models like Word2Vec²⁷ and GloVe,²⁸ where semantic relationships are encoded through vector direction rather than magnitude. It is also more robust to variation in phrasing, sentence length, and token composition—properties particularly relevant to process names in LCI databases.

Based on these considerations, we adopt cosine similarity between the target process

embedding and all training process embeddings:

$$\text{sim}(i, j) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\| \cdot \|\mathbf{e}_j\|} \quad (3)$$

to rank training samples.

Then, for each prediction task, we select only the top 1% most similar processes—approximately 290 out of 29,261 potential training examples. This selective approach contradicts the conventional wisdom that “more data is better” but delivers remarkable benefits. Training time drops from over 10 hours (XGBoost on full data) to just 1 minute on our approach, while simultaneously improving prediction accuracy across all missing scenarios.

This significant improvement arises from the focused learning environment created by semantic selection. Instead of attempting to learn general patterns across all industrial processes (that is an almost impossible task due to their diversity), our model concentrates on semantically similar processes that are expected to exhibit relevant flow characteristics. This method reflects the way humans would tackle the problem, emphasizing similar processes rather than attempting to identify universal patterns across entirely different industrial sectors.

3.3. Feature Processing and Fusion

The combination of semantic embeddings and numerical flow data poses a challenge, largely due to the distinct scales and features that each data type possesses. In this section, we developed a thoughtfully calibrated processing pipeline to facilitate the effective integration of these information sources.

For semantic embeddings, we apply a scaling transformation:

$$\mathbf{e}_i^{\text{scaled}} = \eta \cdot \mathbf{e}_i. \quad (4)$$

where the scaling η controls the influence of semantic information relative to numerical data.

At lower values, the model underutilizes semantic information, whereas at higher values, semantic information overshadows numerical patterns, leading to a decline in performance.

The LCI data presents unique challenges for normalization due to the wide range of magnitudes—flow values can range from 10^{-12} to 10^6 across different processes and flow types. To address this challenge, we implemented a two-step normalization process:

$$y_{ij}^{\text{row}} = \frac{y_{ij}}{\max(|\min(\mathbf{y}_i)|, \max(\mathbf{y}_i))} \quad \text{and} \quad y_{ij}^{\text{norm}} = \frac{y_{ij}^{\text{row}}}{\max(|\min(\mathbf{y}_{\cdot j})|, \max(\mathbf{y}_{\cdot j}))} \quad (5)$$

The initial step normalizes each process’s flows relative to its internal range, while the subsequent step normalizes each flow type across different processes. This approach preserves crucial relative relationships while also ensuring that all features fall within a range that is appropriate for training neural networks.

These processed features (i.e., scaled semantic embeddings and normalized flow values) are subsequently concatenated with the index of the flow that is being predicted to create the input vector:

$$\mathbf{x}_i = \text{concat}(\mathbf{e}_i^{\text{scaled}}, \{y_{ik}^{\text{norm}} \mid k \in \mathcal{K}_i\}, j). \quad (6)$$

This fusion approach allows the model to utilize both semantic information and numerical patterns simultaneously, effectively addressing the limitations inherent in approaches that rely solely on either aspect.

Note that, in the special case where no numeric data are available for the target process (i.e., a zero-shot scenario), the semantic embedding becomes the sole input feature for prediction. As illustrated in Figure 1, although no numerical LCI features are involved in this setting (i.e., $\mathbf{x}_i = \eta \cdot \mathbf{e}_i$), the value of η still influences the model’s behavior by adjusting the magnitude of the semantic signal entering the network. This modulation can subtly affect neuron activations and, consequently, the final output. Thus, η provides a meaningful control over how strongly the model responds to semantic information, even when no numeric data are present.

3.4. Coupled Neural Network Architecture

A key limitation of the XGBoost method²⁴ is its dependence on a two-stage process: initially determining the presence of flows (zero versus non-zero), followed by the prediction of values for non-zero flows. This division leads to error propagation issues, as misclassifications occurring in the initial stage have a lasting impact on the regression outcomes in the subsequent stage. In contrast, our architecture addresses this limitation by jointly learning both tasks.

The proposed network is composed of three main components: a shared backbone, a classification branch, and a regression branch, which are all interconnected via an attention mechanism. The shared backbone consists of three fully connected layers, with each employing the Rectified Linear Unit (ReLU) activation function (to introduce the nonlinearities). This shared representation guarantees that both classification and regression tasks benefit from common feature learning.

For the classification branch, it utilizes the backbone’s output to estimate the probability of a non-zero flow:

$$\hat{p}_{ij} = \sigma(f_{\text{class}}(\mathbf{x}_i)), \quad (7)$$

where σ is the sigmoid activation function.

After obtaining this probability vector, rather than treating it as a separate preliminary step, we aim to let it play an active role in guiding the regression stage. Based on this idea, our framework feeds this classification output $\hat{p}_{ij}$ into a lightweight multi-head self-attention module:

$$\mathbf{a}_i = \text{Attention}(W\hat{p}_{ij}). \quad (8)$$

where W is a learned linear projection matrix that transforms the classification vector into a format suitable for attention processing. The transformed vector is then used as the query, key, and value in a single-token self-attention operation. This turns the attention mechanism into a channel-wise reweighting function that captures dependencies across flow classes. The

resulting attention vector $\mathbf{a}_i \in \mathbb{R}^{64}$ is concatenated with the shared feature representation $\mathbf{x}_i$ and passed into the regression branch:

$$\hat{v}_{ij} = f_{\text{reg}}(\text{concat}(\mathbf{x}_i, \mathbf{a}_i)), \quad (9)$$

The final prediction combines both outputs:

$$\hat{y}_{ij} = \hat{p}_{ij} \cdot \hat{v}_{ij} \quad (10)$$

This coupling allows feedback from the classification learning to influence the regression task and, conversely, allows classification to inform regression learning. When a flow is incorrectly categorized as zero but regression indicates a significant value, this inconsistency can guide parameter updates that reduce such errors in future iterations.

3.5. Model Training Optimization

Our approach involves optimizing the model through a loss function that effectively balances the objectives of classification and regression:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{class}} + \mathcal{L}_{\text{reg}}, \quad (11)$$

where $\mathcal{L}_{\text{class}}$ represents the binary cross-entropy loss for the classification task, $\mathcal{L}_{\text{reg}}$ denotes the mean squared error for the regression task, and λ controls their relative importance. To guarantee the balanced feedback for both tasks, we select $\lambda = 1.0$ in this paper.

We employ the Adam optimizer with an initial learning rate of 0.001 and incorporate a learning rate scheduler that decreases the rate by 50% when validation loss remains unchanged for 5 consecutive epochs. This adaptive approach allows rapid initial convergence while also providing fine-grained optimization as the training progresses. Furthermore, implementing early stopping with a patience of 20 epochs serves to mitigate overfitting while

also reducing unnecessary computational efforts.

A critical aspect of our training approach is the initialization strategy. In contrast to the method proposed by Zhao et al.²⁴ that requires hyperparameter tuning (that is computationally intensive) for each level of data sparsity, our approach demonstrates stability across different initialization seeds. This will be further illustrated in the later experiments.

3.6. Dataset and Implementation

We conducted our experiments using the Ecoinvent 3.11 database,²⁹ which is one of the most comprehensive and used LCI database available. After preprocessing, the dataset included 26,034 unique processes and 29,271 distinct flows, resulting in an extremely sparse matrix with around 0.22% non-zero entries.

A difficulty in using Ecoinvent data lies in its typical accessibility via proprietary LCA software (e.g., OpenLCA², SimaPro³), which limits the reproducibility and broad application of data-driven methods. To address this limitation, we developed an automated extraction pipeline specifically designed to process the native XML-format SPOLD files. This pipeline extracts not only the numerical flow data but also the textual process descriptions essential for semantic analysis.

Our pipeline adheres to a methodical framework: Initially, it associates activity IDs with their respective locations and names by traversing all SPOLD files. It then extracts activity descriptions, intermediate exchanges, and basic exchanges from each file, using activityLinkId references to rebuild the interrelations across processes. Upon cleansing and standardizing the data (eliminating erroneous entries, addressing duplicates, and normalizing units), it produces a structured matrix format appropriate for analysis. More details with pseudocode can be found in the supporting material.

All experiments were implemented in Python 3.8 using PyTorch for deep learning components and NVIDIA CUDA for GPU acceleration. The code was executed on a workstation

²The official website of the OpenLCA software: <https://www.openlca.org>

³The official website of the SimaPro software: <https://simapro.com>

with an Intel Core i9-14900KF CPU and NVIDIA RTX 4090 GPU. This configuration provided sufficient computational resources to effectively handle the challenging training scenarios. The entire implementation, including the data extraction pipeline and SemaNet code, will be made accessible to the public to guarantee reproducibility and facilitate adoption.

4. Results and Discussion

This section provides a detailed account of the experiments conducted to assess the effectiveness of our SemaNet framework. In more detail, we begin by validating the effectiveness of the semantic embedding approach, then evaluating the performance of classification and regression under different levels of data sparsity, and finally analyzing the framework’s computational efficiency and robustness. All the experiments were conducted utilizing the dataset and implementation outlined in Section 3.6, and all the evaluation metrics are summarized in the supporting material. Note that unless otherwise specified, all reported results are validated over test sets built from ten held-out unit processes and therefore reflect generalization performance rather than training behavior.

4.1. Semantic Embedding Validation

Prior to assessing the effectiveness of our predictions, we initially confirmed that our semantic embedding method effectively captures significant relationships among processes. We chose 30 processes from the Ecoinvent dataset and organized them into three groups of ten, each representing a unique domain.

- 1) Photovoltaic electricity generation processes varying by panel type (mono vs. polycrystalline silicon), installation method (slanted-roof vs. facade), and geographical location (e.g., CN-BJ, CN-CQ, CH, RoW).
- 2) Heat and power co-generation processes differing in capacity (50kW to 1MW), functionality (electricity-only vs. combined heat and power), and geography.

3) Land tenure and transformation processes involving agricultural maintenance, forest conversion, and transition to perennial crops across different Brazilian regions.

We generated embeddings for these processes using SBERT approach and applied t-SNE dimensionality reduction to visualize their relationships in a two-dimensional space, as shown in Figure 2.

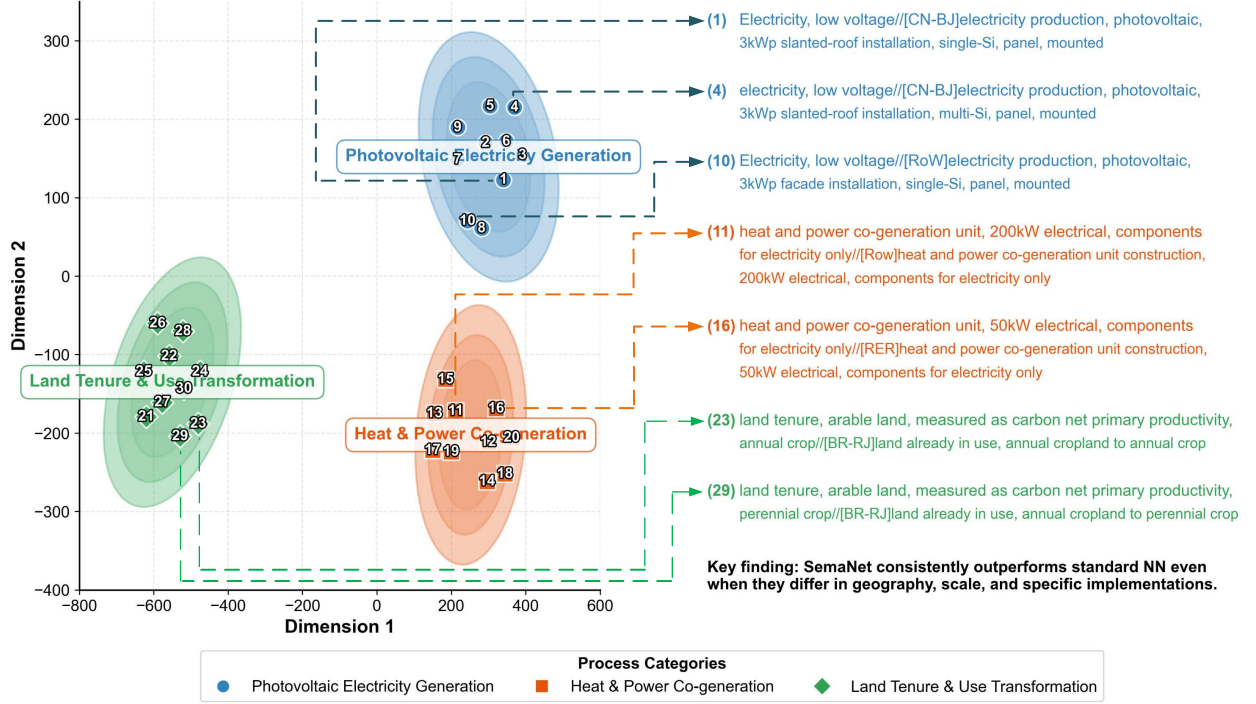


Figure 2: The t-SNE visualization of process embeddings showing clear clustering of semantically related processes. The three distinct clusters correspond to (1) photovoltaic electricity generation (blue), (2) heat and power co-generation (orange), and (3) land tenure and transformation (green). Representative process names for each cluster are shown on the right.

The visualization demonstrates clear clustering based on the type of process involved. Processes within each domain create unique clusters, as those with comparable technical characteristics tend to group together, even when separated by geographical boundaries. For example, the term “Electricity, low voltage//[CN-BJ] electricity production, photovoltaic, 3kWp slanted-roof installation, multi-Si” and the same process in [CN-CQ] are close to each other, indicating that the embedding effectively reflects their functional similarity, even though they originate from different locations. Interestingly, the visualization also reveals

that the electricity-related clusters (photovoltaic and co-generation) appear closer to each other than to the land transformation cluster. This configuration underscores their interdependence throughout energy systems.

The findings from quantitative analysis (which is summarized in Table 2) corroborated the visual observations made. The average cosine similarity between processes within the same cluster was 0.8010, while the average similarity between processes from different clusters was less than 0.4353. This distinct separation demonstrates that our embedding approach effectively captures the semantic relationships critical for recognizing similar processes.

Table 2: Average Cosine Similarity: Intra- vs. Inter-Cluster Processes

	Process (1)	Process (11)	Process (23)
Group 1 (Photovoltaic electricity generation)	0.8529	0.4353	0.2002
Group 2 (Heat and power co-generation)	0.3685	0.8010	0.2001
Group 3 (Land tenure and transformation)	0.1971	0.1673	0.8659

4.2. Effect of Top $k\%$ Semantic Selection on Accuracy and Runtime

Before formally evaluating SemaNet’s classification and regression ability, a preliminary step is to assess how much data is necessary for SemaNet’s training without affecting prediction accuracy. Consequently, we select one unit process out of the Ecoinvent dataset as the held-out test set and randomly mask 10% of the entries in the test column as targets to be predicted. Based on the idea of Top $k\%$ semantic selection, we form the training set by taking the top $k\%$ most semantically similar processes to the test process (from the cosine similarity of the rest 26,033 processes). For each k , SemaNet is trained for 1,000 iterations and after training, we evaluate on the masked entries. The accuracy of predictions is assessed using two common error metrics: Normalized Mean Absolute Error (NMAE) and Normalized Mean Squared Error (NMSE), which are shown in Figure 3.

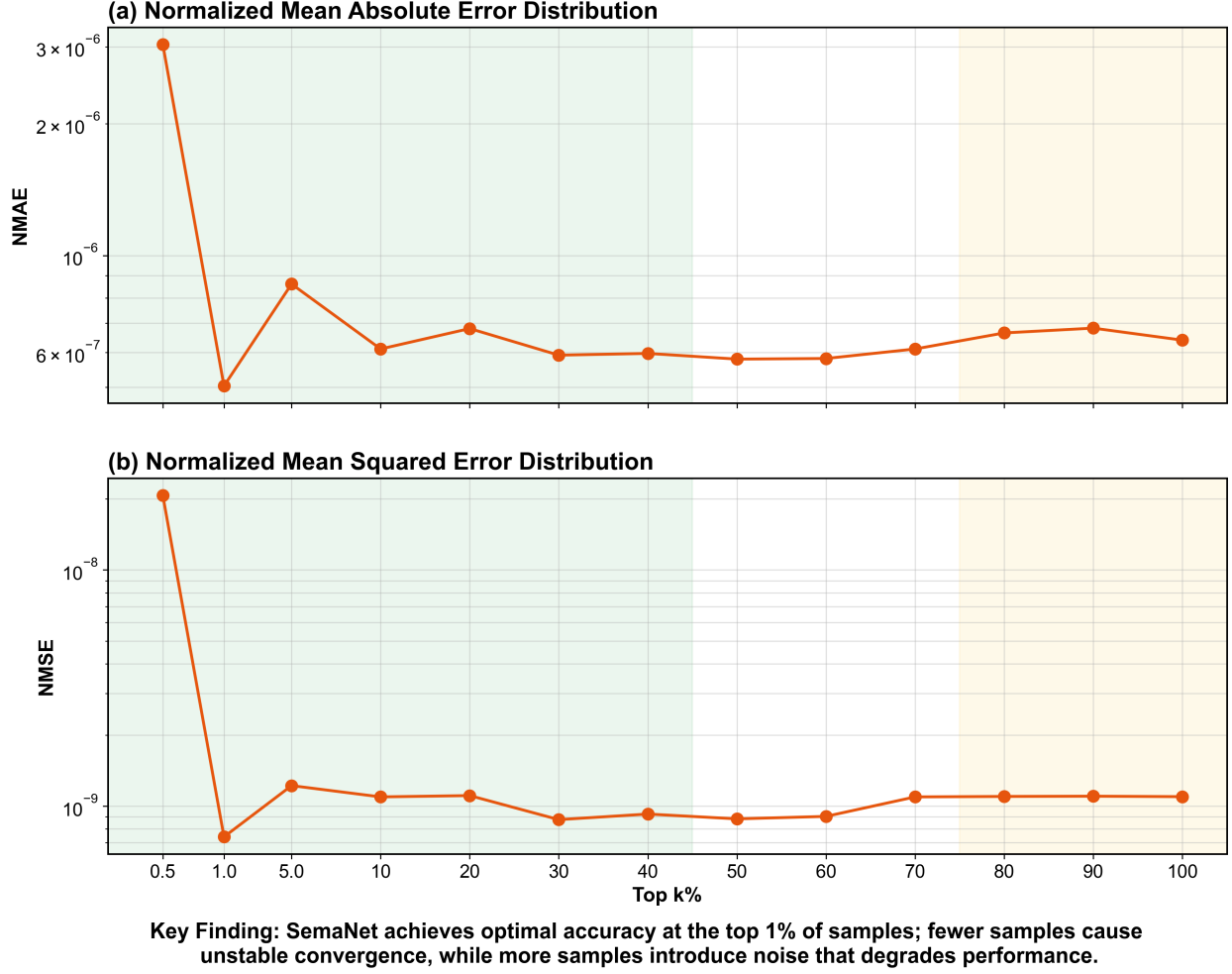
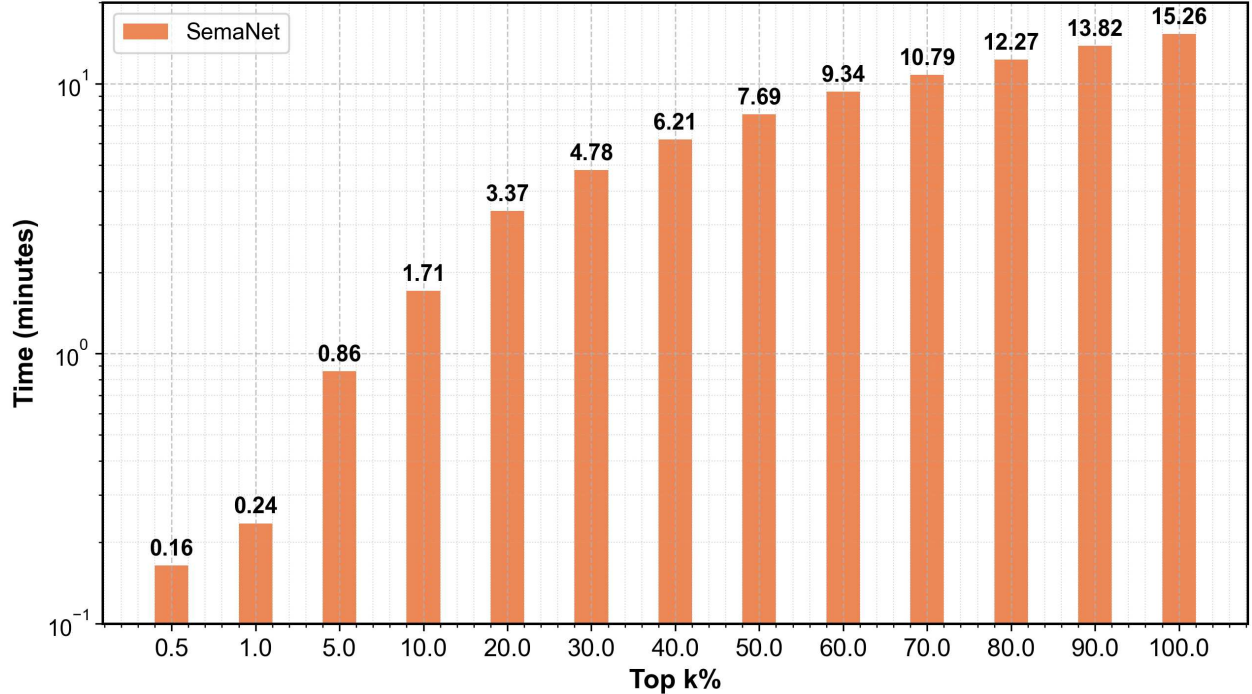


Figure 3: Regression errors (NMAE and NMSE) on the test set across top k% selection.

Figure 3 shows that both error metrics attain their minima at $k = 1\%$, i.e., when only $\approx 26,033 \times 0.01 \approx 260$ nearest neighbours are used for training. It can be observed that, using too few samples (when $k = 0.5\%$) yields unstable estimates with markedly higher errors, reflecting high variance due to insufficient coverage of relevant flow behaviours. On the other hand, enlarging the training set beyond the most relevant neighbours (e.g., $k \geq 5\%$) degrades accuracy: semantically distant processes act as distractors, injecting heterogeneity and noise that dilute the useful signal (“semantic drift”). This validates the intuition behind semantic filtering: the goal is not to maximise the amount of data, but to maximise the relevance of data. Besides the error evaluation, the training time of each $k\%$ is also reported in Figure 4.



Key Finding: Computation time increases with the number of selected samples, with top 1% offering an effective balance between accuracy and runtime.

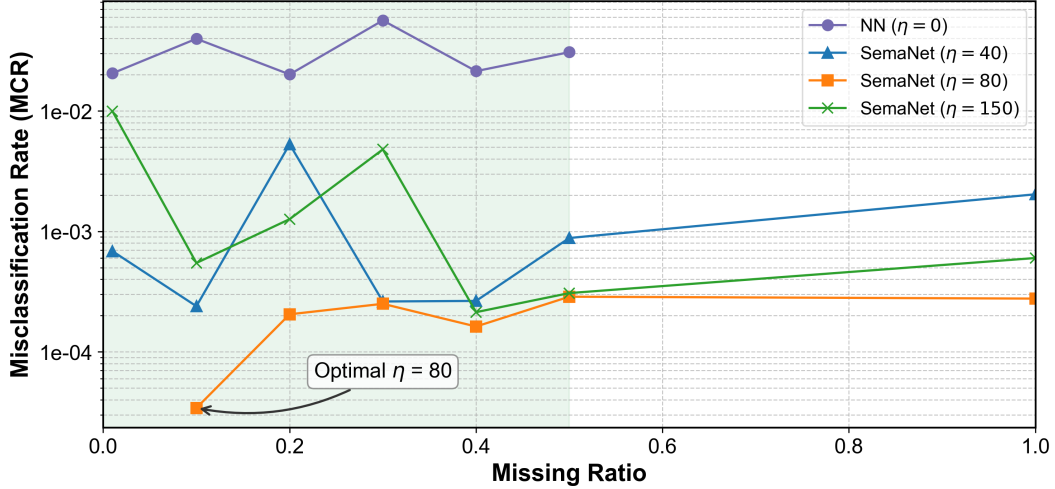
Figure 4: Computation time at different k% selection (evaluated on the test set).

Figure 4 shows that the runtime grows monotonically with k , as larger k increases the size of training set to affect computation time. Under the conventional full-data practice (i.e., $k = 100\%$), SemaNet takes 15.26 minutes in our setup, whereas using only the top 1% nearest neighbours finishes in 0.24 minutes—an $\approx 64\times$ speed-up. Coupled with Figure 3, we therefore fix $k = 1\%$ for the remainder of the study (Sections 4.3–4.8). This setting preserves the benefits of semantic relevance and enables rapid experimentation cycles, while avoiding the instability of extremely small neighbor sets and the negative transfer introduced by large, semantically heterogeneous training pools.

4.3. Prediction of Non-zero Flows as a Classification Task

A critical component of LCI prediction is accurately determining which flows exist (non-zero) versus which are absent (zero). We treat this as a classification task and evaluate the performance of our SemaNet model at different missing data ratios, from 10% to 100%.

Figure 5 illustrates the misclassification rate (MCR) of our SemaNet model at different missing ratios and semantic scaling factors (η).



Key Finding: SemaNet with $\eta = 80$ achieves up to 1000 $\times$ lower MCR compared to standard NN

Figure 5: Misclassification rate (MCR) on the test set at different data missing ratios for our SemaNet model with varying semantic scaling factors (η) compared to the baseline XGBoost approach. Lower values indicate better performance.

The findings indicate that the models that incorporate semantic information (i.e., $\eta > 0$) consistently outperform those relying solely on numerical data (i.e., $\eta = 0$) neural network (NN) across all missing ratios. For example, when 10% of the data is missing, our method utilizing $\eta = 80$ yields a MCR of 3.42e-05. This indicates a remarkable 99.9% reduction in classification errors in comparison to the numerical-only neural network, which has an MCR of 0.03977. Notably, our approach demonstrates remarkable effectiveness even with 100% missing data (zero-shot scenario), achieving an MCR of 2.77e-4. In this extreme situation, the numerical-only NN framework demonstrates a complete failure, as it is unable to generate any predictions at all.

It can also be concluded from the figure that the scaling factor η plays a crucial role in balancing semantic and numerical information. The improvement in performance as η increases from 0 to 80 highlights the importance of semantic information. However, further increasing η to 200 leads to a decline in classification performance. This indicates that an overemphasis on semantic information (at the expense of numerical patterns) may not yield

beneficial results.

In order to conduct a comprehensive evaluation that extends beyond the binary MCR, we assessed our method using additional classification metrics: Precision, Recall, F1-Score, G-Mean, and Matthews Correlation Coefficient (MCC). Figure 6 illustrates radar plots that compare SemaNet ($\eta = 80$) to the numerical-only NN ($\eta = 0$) across these metrics at different missing ratios.

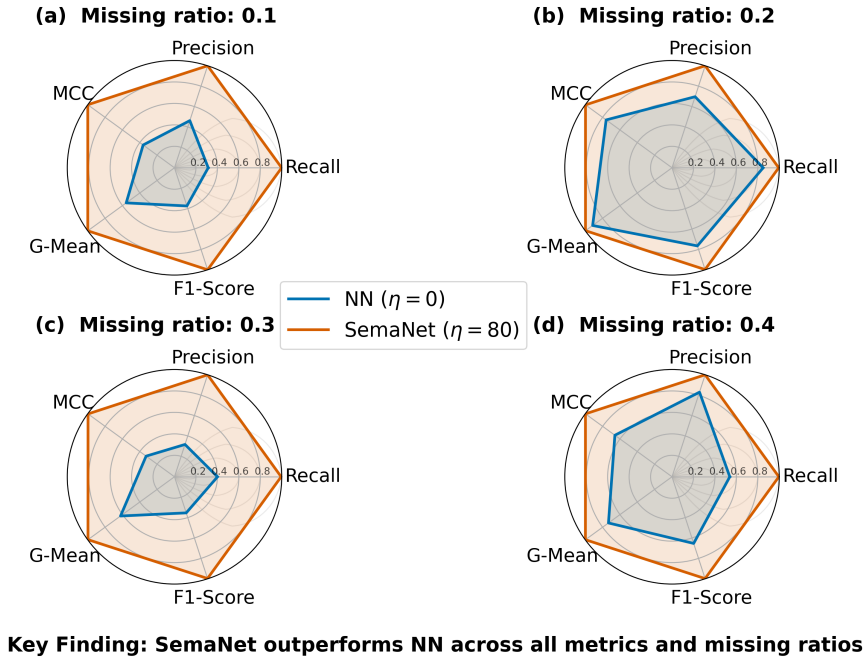
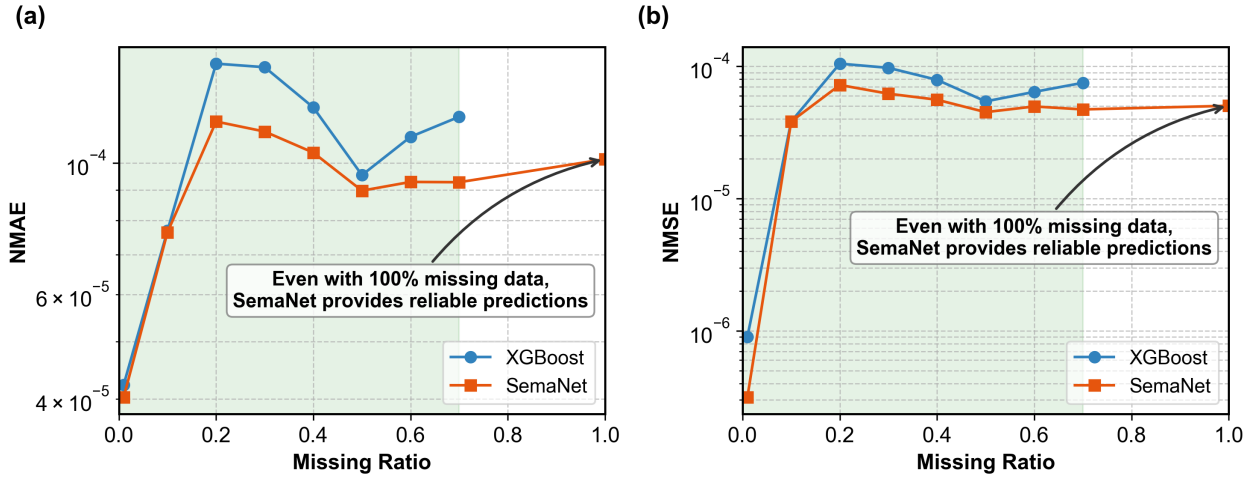


Figure 6: Radar plots comparing test set classification performance across multiple metrics at different missing ratios. The SemaNet model with semantic information (orange) consistently outperforms the numerical-only model (blue) across all metrics and sparsity conditions.

The radar plots demonstrate that semantic embeddings improve all aspects of classification performance, extending beyond just MCR. The enhancement in Recall is especially significant, as it demonstrates the model’s ability to accurately recognize true non-zero flows. This metric is particularly important in LCA applications, as neglecting process flows which contribute to environmental impacts substantially could lead to considerable underestimation of environmental impacts. These findings demonstrate that our semantic approach fundamentally enhances classification performance, facilitating reliable flow identification under scenarios where traditional methods encounter difficulties or are entirely ineffective.

4.4. Qualitative Predictions of Non-zero Flow

While classification determines which flows exist, regression predicts their actual values, which is a critical task for accurate environmental impact assessment. The accuracy of predictions is also assessed using NMAE and NMSE. Figure 7 presents a comparison of the regression performance of our SemaNet approach against the state-of-the-art method (i.e., XGBoost) across different missing ratios.



Key Finding: SemaNet yields consistently lower errors than XGBoost in terms of both NMAE and NMSE.

Figure 7: Regression errors (NMAE and NMSE) on the test set at different missing ratios for SemaNet (orange) versus XGBoost (blue). Lower values indicate better performance. Note that XGBoost cannot make predictions at 100% missing data.

The figure shows that our SemaNet framework outperforms XGBoost in all cases of missing data ratios, which highlights the effectiveness of semantic embeddings in revealing the relationships and preserving prediction accuracy. Note that XGBoost cannot create estimates when all numerical data is completely absent, while SemaNet still provides reasonable flow forecasts solely on text embeddings. This demonstrates the robustness of our strategy, as it can still yield predictions despite the scarcity of data, which is desirable in practice.

Figure 8 shows the convergence of MCR for SemaNet and XGBoost at a 10% missing ratio. It demonstrates that SemaNet begins reducing MCR as early as epoch 30, while XGBoost does not exhibit any sign of convergence until epoch 80. It can also be seen from the figure that SemaNet reaches a stable low-error plateau at epoch 70 and no more improvement

from that until epoch 90. This phenomenon demonstrates that although the model is trained jointly on classification and regression objectives, the classification head tends to stabilize first, resulting in a temporary plateau in MCR while regression error continues to decrease. After epoch 90, SemaNet continues to improve its classification behaviour and eventually achieves a final MCR of 7×10^{-5} , which is better than XGBoost’s 1×10^{-4} .

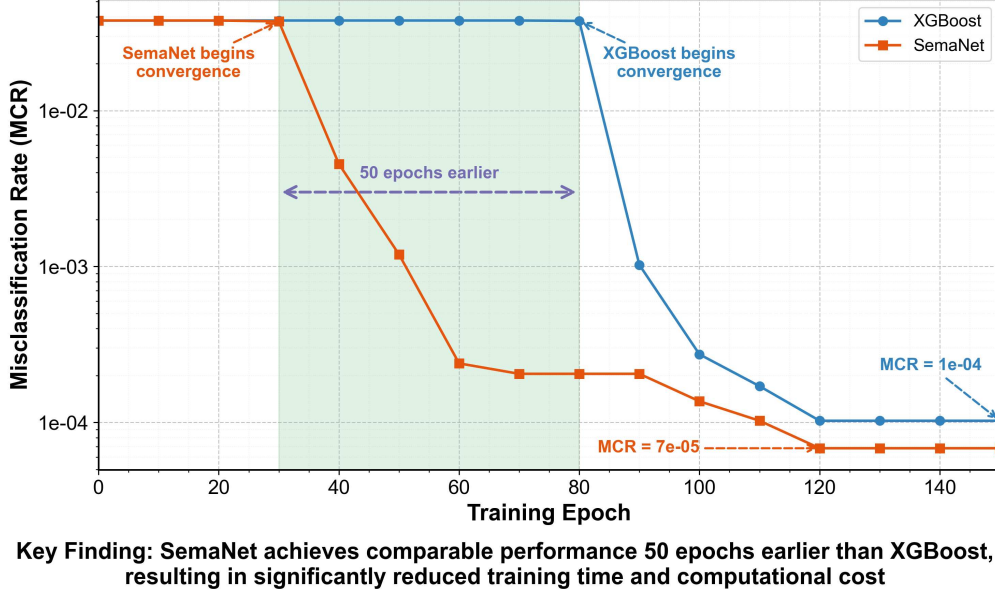


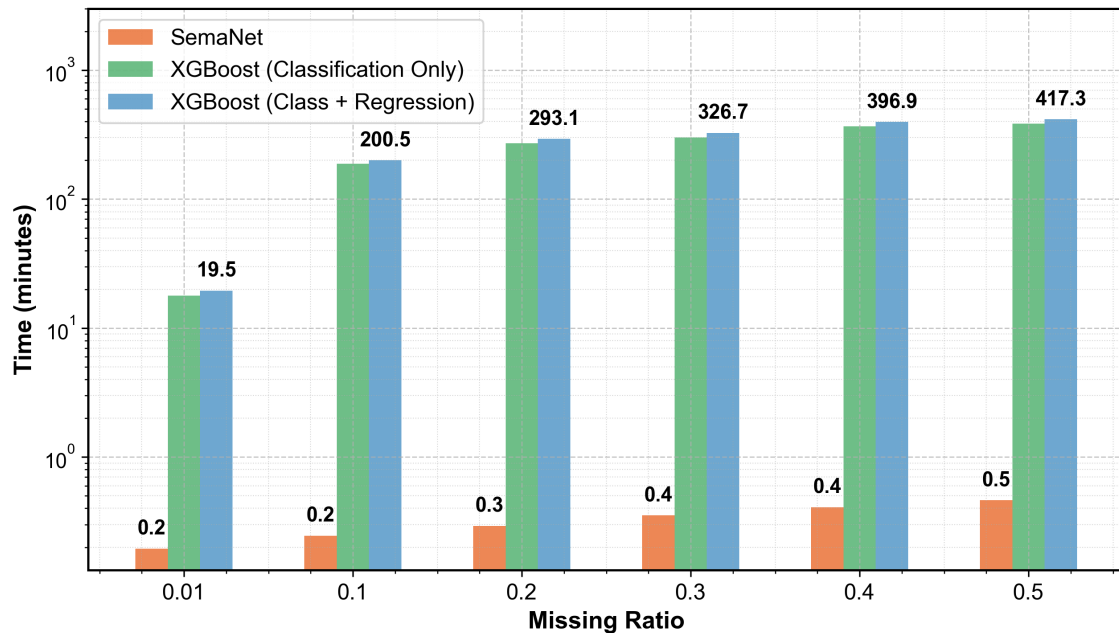
Figure 8: Convergence of MCR for SemaNet versus XGBoost at 10% missing data. SemaNet achieves lower error rates and faster convergence dynamics.

The performance advantage of SemaNet over XGBoost can be attributed to three main factors: (1) the integration of semantic information that captures functional similarities among processes, (2) a selective training strategy that emphasizes on the most relevant processes, and (3) the coupled neural architecture that prevents error propagation between classification and regression tasks. These advancements work together to notably enhance prediction accuracy, particularly in challenging high-sparsity situations.

4.5. Computational Efficiency

Beyond accuracy improvements, our SemaNet framework also delivers substantial computational advantages. To demonstrate this efficiency improvement, we compare their compu-

tation time (with 150 training epochs) at missing data ratios varying from 0.01 to 0.5 (as shown in Figure 9). Leveraging semantic-based filtering of the top 1% of data and highly parallelized GPU tensor operations, our SemaNet methodology (orange bars) reliably achieves training completion within one minute across all evaluated situations.



Key Finding: SemaNet demonstrates a significant computational advantage over XGBoost.

Figure 9: Computation time at different data missing ratio (evaluated on the test set).

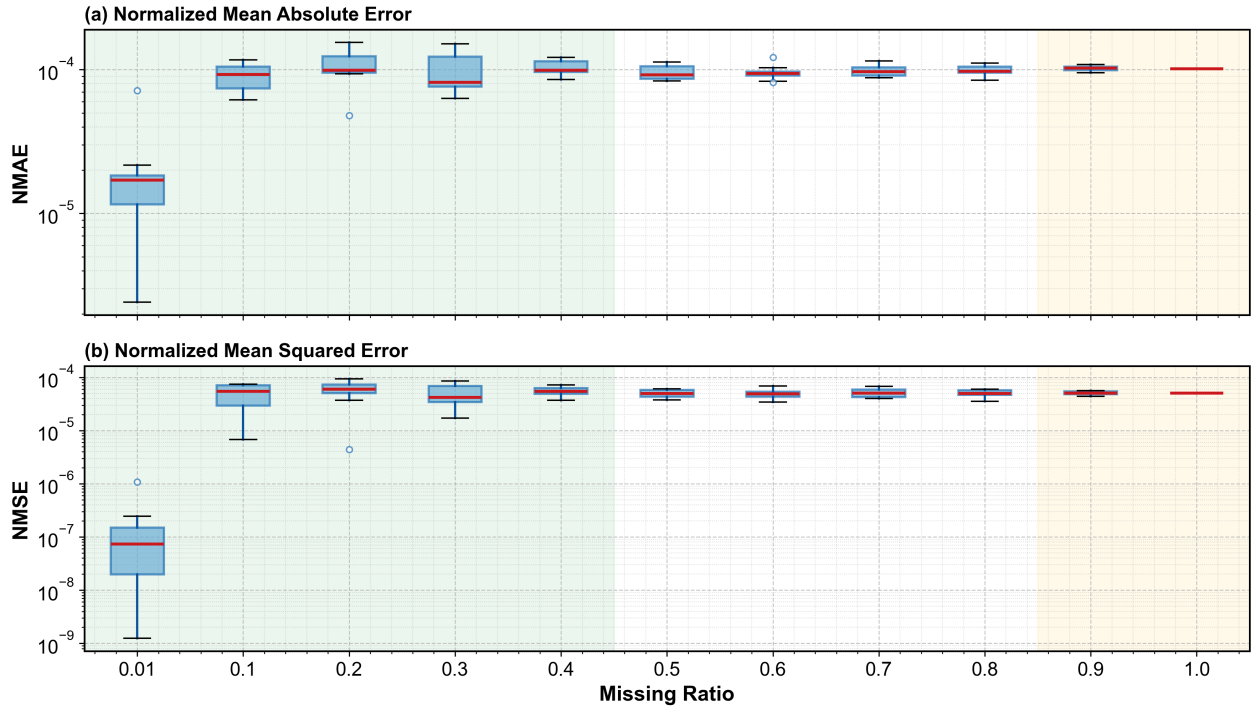
In contrast, the classification-only variant of XGBoost (green bars) and its dual-stage pipeline (blue bars) require significantly more time—sometimes exceeding hundreds of minutes (from 19.5 minutes at 1% missing data to 417.3 minutes—around 7 hours at 50% missing data). Even though the regression stage in XGBoost can save time by skipping zero flows for prediction, it still has to handle non-zero data, which increases computing time.

In summary, the proposed SemaNet framework improved performance by utilizing semantic filtering to decrease the size of training data, simultaneously in both classification and regression performance, which has significant practical implications. In practice, LCA practitioners often need to assess various scenarios or revise LCA results as new industry data is updated. With XGBoost, each new analysis could require overnight processing, which significantly restricts the ability to engage in iterative exploration. Our SemaNet framework

enables rapid assessment cycles, yielding results in minutes instead of hours. Therefore, it expands the opportunities available in practical LCA applications.

4.6. Robustness Analysis

To ensure that our results represent real improvements instead of being influenced by particular data setups, we undertook a comprehensive robustness analysis through multiple experimental runs. We executed ten independent experiments with randomized flow orders and initialization seeds to assess consistency and reliability. Figure 10 illustrates boxplots of NMAE and NMSE across these independent experiments.



Key finding: Error variance decreases at high missing ratios as predictions rely more on semantic information

Figure 10: Distribution of regression errors on the test set across ten experimental runs with different random initializations. Boxplots show median, quartiles, and range at each missing ratio, demonstrating consistent performance despite randomization.

In all 10 experiments, the proposed method can attain high prediction accuracy at low missing ratios and regularly escalates with increasing missing ratios. Remarkably, even with 100% data absence, it still generates significant estimates—highlighting the model’s capacity

to depend exclusively on semantic embeddings in the absence of numerical features. The uniformity observed across many random configurations demonstrates the robustness of the approach, guaranteeing its application to incomplete LCI datasets.

An interesting pattern emerges when analyzing the boxplots corresponding to various missing ratios. As the missing ratio increases, the variability in error distributions tends to diminish, leading to more compact boxplots at elevated levels of sparsity ($\geq 50\%$). This counterintuitive finding indicates that when numerical data becomes extremely sparse, the model increasingly depends on semantic information, which offers a consistent foundation for predictions despite the randomization of numerical inputs.

The robustness analysis provides additional evidence that our semantic approach fundamentally alters the prediction challenge rather than simply offering minor enhancements. By establishing a reliable foundation based on process semantics, we develop a robust prediction framework that consistently performs well, even in the face of significant data constraints and diverse experimental conditions.

4.7. Validation on Structurally Diverse and Non-Standardized LCI Dataset Beyond Ecoinvent

To further assess the generalizability of our semantic-based framework beyond Ecoinvent, we evaluated SemaNet on an independent LCI dataset, the Tiangong LCA database⁴. Unlike Ecoinvent, which is highly standardized and rigorously structured, Tiangong records open-access process data from 70 different sectors in China, including manufacturing, agriculture, energy, and waste treatment.

Due to the diversity in different sectors, the Tiangong dataset exhibits a broader spectrum of real-world data characteristics: process descriptions may be inconsistently structured, flow completeness varies significantly, and terminology is often heterogeneous. These properties resemble the practical conditions encountered when LCA practitioners rely on self-reported

⁴The official website of Tiangong database: <https://www.tiangong.earth/zh/data>

data, literature sources, or early-stage product development inputs (settings where process descriptors may be vague, incomplete, or inconsistent). In this regard, Tiangong serves as a practical proxy for the real-world challenges of incomplete LCI data with variation in either flows or descriptors compared to Ecoinvent.

Following the same workflow as described in Section 3, we applied SemaNet to this unstructured dataset and obtained the results as shown in Figures 11 and 12. From Figure 11, the regression accuracy of SemaNet outperforms the XGBoost method across different missing ratios. Most impressively, even with 100% missing data (i.e., only process descriptions are available to the model), SemaNet can generate usable preliminary predictions solely based on semantic information. In contrast, XGBoost is unable to operate (i.e., make any meaningful predictions) in this scenario due to the absence of numerical features.

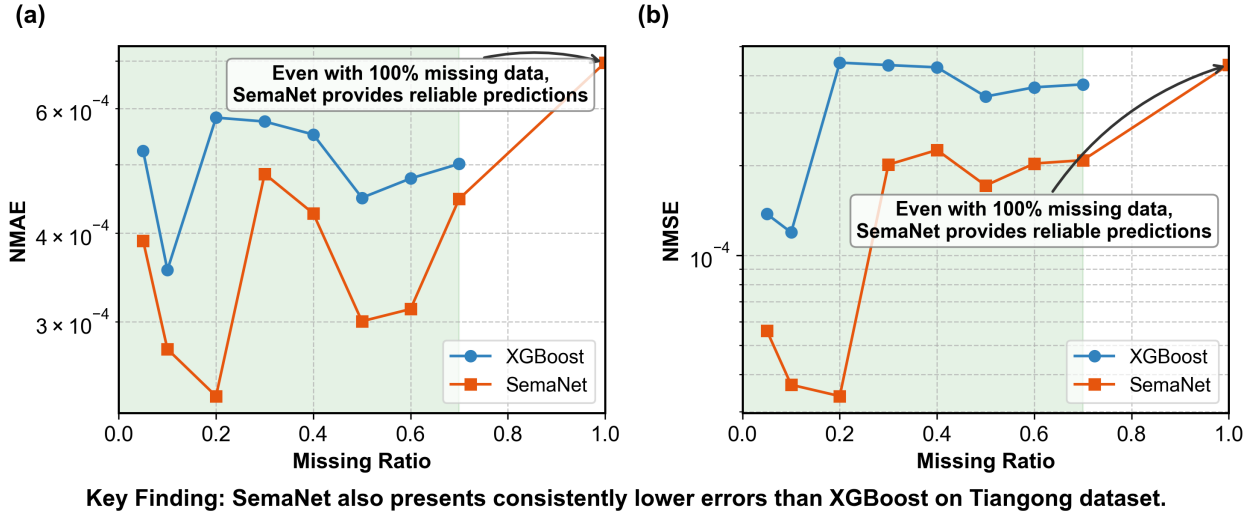
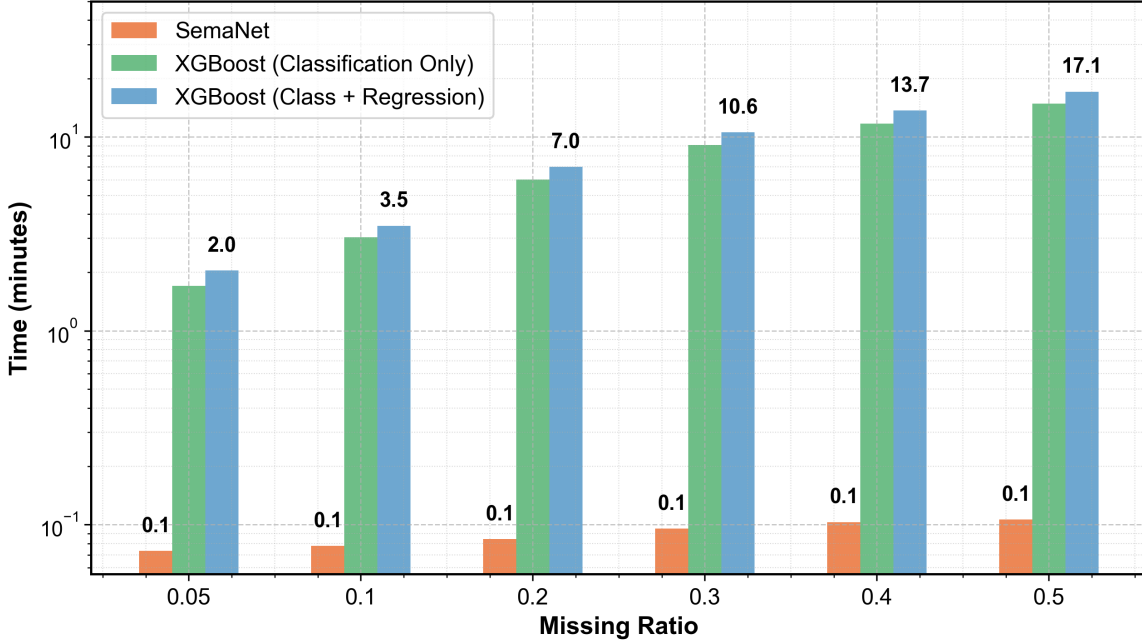


Figure 11: Regression errors (NMAE and NMSE) on the test set at different missing ratios for SemaNet (orange) versus XGBoost (blue). Lower values indicate better performance. Note that XGBoost cannot make predictions at 100% missing data.

Figure 12 demonstrates SemaNet’s exceptional scalability and computational efficiency. The semantic-based sample filtering mechanism constrains training time to under ten seconds across all evaluated missing ratios, enabling rapid iterative deployment in industrial environments. This computational advantage proves critical for real-time applications requiring frequent model updates—evaluating new suppliers, adapting to evolving product specifica-

tions, or responding to dynamic design constraints during early development phases.



Key Finding: SemaNet demonstrates computational advantage over XGBoost on TianGong dataset.

Figure 12: Computation time at different data missing ratio (evaluated on the test set).

4.8. Quantifying Architectural Gains via Module-swap Ablations with XGBoost

To substantiate the semantic gains of the proposed SemaNet framework and evaluate whether integrating XGBoost can benefit the design, we extend the experiment in the last Section. Four approaches are compared: (i) SemaNet (semantics in classification and regression); (ii) XGBoost(non-semantic baseline); (iii) SemaNet + XGBoost (SemaNet for classification, XGBoost for regression); and (iv) XGBoost + SemaNet (XGBoost for classification, SemaNet for regression). We first evaluate their predictive performance using the error metrics NMAE and NMSE across different missing ratios.

Figure 13 shows that SemaNet attains the lowest NMAE/NMSE across the entire range of missing ratios and, critically, continues to produce meaningful predictions in the zero-shot case (100% missing). This robustness follows from applying semantic information in both

modules: the classifier uses embeddings to gate plausible non-zero flows, and the regressor maps the same embeddings to values, yielding a coherent semantic prior even when no numerics are available. Among other models, SemaNet + XGBoost ranks second overall. The gain arises because semantic classification exploits the sparsity of LCI matrices to identify a compact, high-precision set of candidate non-zero flows, so the subsequent tree-based regression operates on a well-filtered target set.

By contrast, XGBoost + SemaNet performs competitively only at low sparsity (missing ratio < 0.2): with sufficient numerical evidence, the non-semantic classifier can still pass a reasonably accurate mask to the semantic regressor, consistent with prior observations on tree-based models under modest missingness.²⁴ As the missing ratio increases, classification without semantic cues deteriorates and the propagated mask becomes the accuracy bottleneck, which explains the rise of its error curves. Nevertheless, both hybrids improve upon the fully non-semantic XGBoost baseline across all missing ratios, indicating that injecting semantics into either classification or regression already confers a tangible benefit. While at 100% missing (zero-shot), any pipeline that contains a non-semantic module fails in predicting numerical values: a non-semantic classifier has no features to identify non-zero flows, and a non-semantic regressor has no usable inputs to map to values.

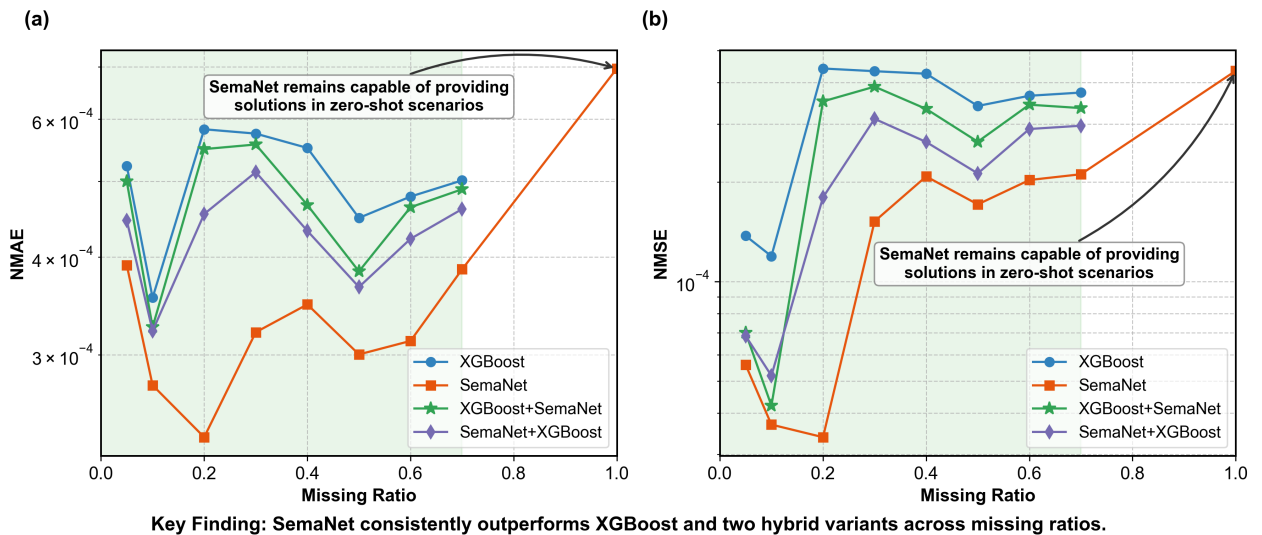
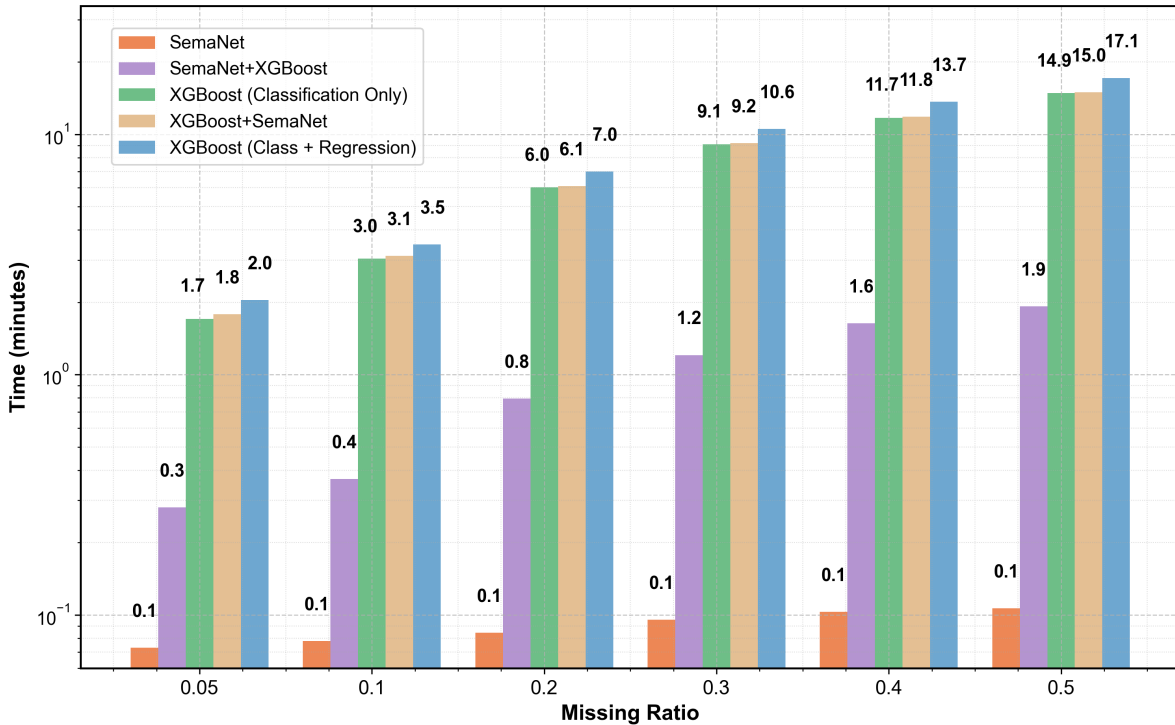


Figure 13: Error comparison (NMAE and NMSE) for SemaNet, XGBoost, and two hybrids across missing ratios.

Besides the prediction error, we also record the training time of each approach, as shown in Figure 14. SemaNet is the fastest, and its curve is nearly flat as the missing ratio increases, reflecting that the semantic pipeline is intrinsically lightweight under the fixed Top $k\%$ setting and is only weakly affected by how many entries are masked. SemaNet + XGBoost ranks second: semantic classification first identifies a small set of candidate non-zero flows in the sparse matrix, so the subsequent tree-based regression is executed on a markedly reduced subset; runtime grows modestly with higher missing ratios as more targets require value estimation. XGBoost + SemaNet is slower because the front-end classifier lacks semantic cues and must operate on high-dimensional, mostly-zero numerics, which increases computational burden before the semantic regressor is applied. The slowest pipeline is the fully non-semantic baseline XGBoost(classification + regression), which lacks semantic filtering to reduce the candidate set and therefore runs both stages over nearly all entries.



Key Finding: SemaNet consistently outperforms standard XGBoost and its integrated variants in computational efficiency.

Figure 14: Runtime comparison for SemaNet, XGBoost, and the hybrids at different missing ratios.

4.9. Practical Implications for LCA

SemaNet addresses a fundamental challenge in industrial LCA: generating reliable environmental data when proprietary constraints or data scarcity prevent conventional assessment approaches. The framework synthesizes missing LCI data through semantic inference, enabling comprehensive environmental assessments with 60-80% less input data while maintaining prediction accuracy.

This data efficiency translates into quantifiable cost reductions. Traditional LCA assessments incur substantial expenses through data acquisition (\$20,000-\$40,000), expert consultation (\$15,000-\$30,000), database licensing (\$5,000-\$15,000), and extended project timelines that delay decision-making (\$10,000-\$15,000 in opportunity costs). SemaNet’s semantic inference reduces data collection requirements by 60-80%, eliminating most field studies and primary data gathering. Expert consultation shifts from extensive data interpretation to framework validation, reducing associated costs by 70%. Timeline compression from 6-8 months to 6-8 weeks further minimizes opportunity costs and project overhead. These combined efficiencies yield total assessment costs of \$10,000-\$20,000, representing a 70-80% reduction from conventional \$50,000-\$100,000 evaluations. These improvements lower implementation barriers across industrial sectors, particularly benefiting small and medium enterprises that traditionally lack resources for comprehensive environmental assessments. Manufacturing organizations can now integrate LCA into standard product development cycles rather than treating environmental assessment as a specialized, external process.

SemaNet’s computational efficiency enables two complementary deployment strategies. Expert users can integrate the framework as an interactive module within design workflows, generating real-time predictions from textual specifications. Alternatively, practitioners can employ SemaNet offline to create semantically enriched LCI databases for standard plug-and-play deployment. While current implementation prioritizes the interactive mode, integration with mainstream LCA platforms represents a natural extension for broader adoption.

4.10. Toward Trustworthy and Scalable AI-Based LCI Prediction

Ensuring the robustness and scientific legitimacy of the AI-generated data is critically important before applying SemaNet in practical LCI data prediction. This section outlines some concerns and discusses how they can be addressed or mitigated.

4.10.1. Uncertainty Estimation in AI-Generated LCI Values

Uncertainty estimation is a fundamental requirement for the scientific use of any AI-generated environmental data. Unlike conventional LCI datasets such as Ecoinvent, which incorporate uncertainty through pedigree matrices and statistical distributions, most machine learning-based methods (including our own method SemaNet) typically produce point estimates without an explicit quantification of prediction confidence.

To address this limitation, we outline several promising directions for future work. First, approximate Bayesian techniques (e.g., Monte Carlo analysis and deep ensembles) could be incorporated into the model to yield confidence intervals over predictions. Second, Gaussian Process Regression (GPR) offers a principled alternative for delivering both point estimates and uncertainty bounds, which has been illustrated on environmental inventory modeling.²² Third, our very recent conference paper³⁰ has explored neural process architectures that enable active learning and uncertainty-aware predictions, however, it has not incorporated the semantic information in the framework.

While these approaches remain outside the scope of the present study, it is a highly promising future direction to transit from point-based prediction to uncertainty-aware modeling. Our future work will consider the integration of these approaches to further enhance the reliability, transparency, and applicability of AI-generated LCI data.

4.10.2. Data Quality Transfer and Representativeness

When using AI to generate LCI data, one of the most important things is to keep the quality of the data and its ability to represent the context. In our current SemaNet framework, it

captures contextual signals implicitly through a semantic embedding scheme. As described in Section 3.1, each unit process name is decomposed into three components (i.e., Reference Product, Geography, and Activity Name), which are separately encoded using a pre-trained sentence transformer. The resulting vectors are combined into a 1152-dimensional representation that keeps structure-aware semantic alignment.

This design enables SemaNet to focus more on samples that are relevant to the meaning and context when it trains and makes predictions. For example, two processes that use similar technologies but come from different regions are likely to be separated in the embedding space, preserving geographic specificity in the inferred values. While this serves as an effective proxy, it is still an approximation of true metadata transfer.

To improve representational fidelity, the future extensions of SemaNet may explicitly incorporate auxiliary metadata (e.g., reliability scores, regional weights and temporal tags) into the sample selection framework. This would allow the model to better reflect not only numerical similarity, but also data quality dimensions as defined by International Reference Life Cycle Data System (ILCD)³¹ and ISO 14044 standards.³² This would further align the proposed framework with established LCA data quality principles.

4.10.3. Model Reliability for Novel Technologies

Another critical problem of machine learning-based prediction frameworks, including SemaNet, lies in their ability to generalize to novel or emerging technologies. These kinds of technologies often diverge significantly from the historical data used for training. Even though the text looks similar, they may not share meaningful underlying characteristics. As a result, relying on existing analogues can lead to unjustified extrapolations and overconfident predictions.

To mitigate these risks, a possible solution is a semantic novelty detection mechanism that measures the distance between the target process and all training samples in the embedding space. If this similarity is less than a certain level, the prediction is marked as low-confidence,

which warns users that the extrapolation may not be reliable. The recent advances in large language models (LLMs) provide an opportunity to go further. LLMs can evaluate semantic similarity across deeper dimensions (e.g., technological functionality, environmental relevance, and contextual operation) offering more meaningful insight into whether a process meaningfully aligns with historical precedents.

To guarantee the transparency and methodological rigor, it is recommended that future AI-generated LCI datasets should adhere to the ILCD data creation protocol.³¹ This protocol provides a structured way for documenting assumptions, indicating uncertainties, and ensuring representativeness across technical, temporal, and geographic dimensions. At the same time, expert validation remains essential. In accordance with ISO 14044,³² any AI-generated data destined for public databases, particularly for emerging technologies, must undergo expert evaluation to determine its plausibility, completeness, and conformity with physical and technological limitations.

Finally, future extensions may combine semantic data-driven inference with mechanistic models or expert-informed priors. These hybrid methods could provide increased robustness in low-data regimes and ensure that AI-generated inventories remain scientifically credible and practically applicable.

4.10.4. Practical Constraints and Deployment Challenges

Beyond the aforementioned methodological concerns, several broader limitations remain that influence the scalability and generalizability of SemaNet in practical deployment environments. First, the current framework supports only English-language descriptions, limiting its applicability in multilingual industrial settings. Second, SemaNet requires on-demand prediction for each process rather than generating reusable gap-filled databases, posing usability challenges for non-technical LCA practitioners.

To address these concerns, future research may pursue universal semantic models capable of cross-domain prediction without domain-specific retraining. This goal presents several

technical obstacles: industrial vocabulary remains highly heterogeneous, current datasets lack coverage across manufacturing sectors, and existing neural models struggle to generalize environmental knowledge. Achieving this vision requires developing multilingual training corpora, curating structured ontologies, and architecting transformer-based models specifically tailored to industrial LCI tasks.

Acknowledgement

This research was supported by UK Engineering and Physical Sciences Research Council KE-IAA 193664.

Supporting Information Available

The following files are available free of charge.

- Additional information on the automated data extraction protocol, the formulation of evaluation metrics, and the names of thirty unit process examples. (PDF)

References

- (1) Hellweg, S.; Milà i Canals, L. Emerging approaches, challenges and opportunities in life cycle assessment. *Science* **2014**, *344*, 1109–1113.
- (2) Hauschild, M. Z.; Rosenbaum, R. K.; Olsen, S. I. *Life Cycle Assessment–Theory and Practice*; Springer, 2018.
- (3) Virtuani, A.; Block, A. B.; Wyrsh, N.; Ballif, C. The carbon intensity of integrated photovoltaics. *Joule* **2023**, *7*, 2511–2536.
- (4) Guinée, J. B.; Heijungs, R.; Huppes, G.; Zamagni, A.; Masoni, P.; Buonamici, R.;

- Ekvall, T.; Rydberg, T. Life cycle assessment: past, present, and future. *Environmental Science & Technology* **2011**, *45*, 90–96.
- (5) Curran, M. A. Life cycle assessment: a review of the methodology and its application to sustainability. *Current Opinion in Chemical Engineering* **2013**, *2*, 273–277.
- (6) Canals, L. M. i.; Azapagic, A.; Doka, G.; Jefferies, D.; King, H.; Mutel, C.; Nemecek, T.; Roches, A.; Sim, S.; Stichnothe, H. Approaches for addressing life cycle assessment data gaps for bio-based products. *Journal of Industrial Ecology* **2011**, *15*, 707–725.
- (7) Pascual-González, J.; Pozo, C.; Guillén-Gosálbez, G.; Jiménez-Esteller, L. Combined use of MILP and multi-linear regression to simplify LCA studies. *Computers & Chemical Engineering* **2015**, *82*, 34–43.
- (8) Reap, J.; Roman, F.; Duncan, S.; Bras, B. A survey of unresolved problems in life cycle assessment—Part 2: Impact assessment and interpretation. *The International Journal of Life Cycle Assessment* **2008**, *13*, 374–388.
- (9) Liu, T.; Wang, Q.; Su, B. A review of carbon labeling: standards, implementation, and impact. *Renewable and Sustainable Energy Reviews* **2016**, *53*, 68–79.
- (10) Hou, P.; Cai, J.; Qu, S.; Xu, M. Estimating missing unit process data in life cycle assessment using a similarity-based approach. *Environmental Science & Technology* **2018**, *52*, 5259–5267.
- (11) Mazumder, R.; Hastie, T.; Tibshirani, R. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research* **2010**, *11*, 2287–2322.
- (12) Jin, X.; Miao, J.; Wang, Q.; Geng, G.; Huang, K. Sparse matrix factorization with $L_{2,1}$ norm for matrix completion. *Pattern Recognition* **2022**, *127*, 108655.

- (13) Petrov, S.; Zamarashkin, N. Matrix completion with sparse measurement errors. *Calcolo* **2023**, *60*, 9.
- (14) Liu, Y.; Gong, X. Z.; Wang, Z. H.; Liu, W.; Nie, Z.-R. Multiple imputation for missing data in life cycle inventory. *Materials Science Forum*. 2009; pp 21–27.
- (15) Moreau, V.; Bage, G.; Marcotte, D.; Samson, R. Statistical estimation of missing data in life cycle inventory: an application to hydroelectric power plants. *Journal of Cleaner Production* **2012**, *37*, 335–341.
- (16) Wernet, G.; Hellweg, S.; Fischer, U.; Papadokonstantakis, S.; Hungerbühler, K. Molecular-structure-based models of chemical inventories using neural networks. *Environmental Science & Technology* **2008**, *42*, 6717–6722.
- (17) Song, R.; Keller, A. A.; Suh, S. Rapid life-cycle impact screening using artificial neural networks. *Environmental Science & Technology* **2017**, *51*, 10777–10785.
- (18) Meron, N.; Blass, V.; Thoma, G. Selection of the most appropriate life cycle inventory dataset: new selection proxy methodology and case study application. *The International Journal of Life Cycle Assessment* **2020**, *25*, 771–783.
- (19) Hou, P.; Jolliet, O.; Zhu, J.; Xu, M. Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. *Environment International* **2020**, *135*, 105393.
- (20) Liao, M.; Kelley, S.; Yao, Y. Generating energy and greenhouse gas inventory data of activated carbon production using machine learning and kinetic based process simulation. *ACS Sustainable Chemistry & Engineering* **2019**, *8*, 1252–1261.
- (21) Ghoroghi, A.; Rezgui, Y.; Petri, I.; Beach, T. Advances in application of machine learning to life cycle assessment: a literature review. *The International Journal of Life Cycle Assessment* **2022**, *27*, 433–456.

- (22) Dai, T.; Jordaan, S. M.; Wemhoff, A. P. Gaussian process regression as a replicable, streamlined approach to inventory and uncertainty analysis in life cycle assessment. *Environmental Science & Technology* **2022**, *56*, 3821–3829.
- (23) Khadem, S. A.; Bensebaa, F.; Pelletier, N. Optimized feed-forward neural networks to address CO₂-equivalent emissions data gaps—Application to emissions prediction for unit processes of fuel life cycle inventories for Canadian provinces. *Journal of Cleaner Production* **2022**, *332*, 130053.
- (24) Zhao, B.; Shuai, C.; Hou, P.; Qu, S.; Xu, M. Estimation of unit process data for life cycle assessment using a decision tree-based approach. *Environmental Science & Technology* **2021**, *55*, 8439–8446.
- (25) Balaji, B.; Vunnava, V. S. G.; Domingo, N.; Gupta, S.; Gupta, H.; Guest, G.; Srinivasan, A. Flamingo: environmental impact factor matching for life cycle assessment with zero-shot machine learning. *ACM Journal on Computing and Sustainable Societies* **2023**, *1*, 1–23.
- (26) Reimers, N.; Gurevych, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019; pp 3982–3992.
- (27) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. *Efficient estimation of word representations in vector space*; arXiv preprint arXiv:1301.3781, 2013; (accessed 2025-09-01).
- (28) Pennington, J.; Socher, R.; Manning, C. D. Glove: global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014; pp 1532–1543.
- (29) Wernet, G.; Bauer, C.; Steubing, B.; Reinhard, J.; Moreno-Ruiz, E.; Weidema, B. The ecoinvent database version 3 (part I): Overview and methodology. *The International Journal of Life Cycle Assessment* **2016**, *21*, 1218–1230.

- (30) Xing, W. W.; Chen, H.; Chen, Z.; Quan, Z.; Laratte, B.; Walsh, M.; Pu, J.; Casamayor, J. L. Adaptive LCI data completion: integrating neural processes and active learning for enhanced life cycle assessment. *Procedia CIRP* **2025**, *135*, 136–141.
- (31) EC-JCR *International reference life cycle data system (ILCD) handbook: general guide for life cycle assessment—detailed guidance*; Publications Office of the European Union, 2010.
- (32) ISO *ISO 14040—Environmental management—Life cycle assessment—Principles and framework*; International Organization for Standardization: Geneva, Switzerland, 2006.