Green Fine Tuning for Molecular Property Prediction

Emanuele Fontana

Abstract

This project focuses on the application of Green AI techniques for molecular property prediction. Following the CRISP-DM methodology, the work explores an adaptive Early Stopping mechanism based on Exponential Moving Average [1] (under review) for FineTuning Transformer-based models and Graph Neural Networks. The main goal is to analyze the trade-off between predictive accuracy and energy consumption with respect to a classic FineTuning Early Stopping method, showing that it is possible to achieve significant reductions in CO_2 emissions with minimal impact on performance.

Contents

1 Business Understanding **4**
- 1.1 Context and Motivation . . . 4
- 1.2 Business Objectives . . . 4
  - 1.2.1 Requirements . . . 4
  - 1.2.2 Expectations . . . 5
- 1.3 Assess situation . . . 5
- 1.4 Data Mining Goals . . . 6

2 Data Understanding **7**
- 2.1 SMILES Representation . . . 7
- 2.2 Dataset Description . . . 7
  - 2.2.1 BACE . . . 7
  - 2.2.2 BBBP . . . 8
  - 2.2.3 CEP . . . 8
  - 2.2.4 HIV . . . 8
  - 2.2.5 Malaria . . . 8
  - 2.2.6 Lipophilicity . . . 9
- 2.3 Data Exploration . . . 9
- 2.4 Data Distribution Analysis . . . 9
  - 2.4.1 BACE . . . 9
  - 2.4.2 BBBP . . . 10
  - 2.4.3 CEP . . . 10
  - 2.4.4 HIV . . . 11
  - 2.4.5 Malaria . . . 11
  - 2.4.6 Lipophilicity . . . 12

3 Data Preparation **13**
- 3.1 Preprocessing . . . 13
  - 3.1.1 SMILES Validation . . . 13
  - 3.1.2 Graph Generation . . . 14

4 Modelling **16**
- 4.1 Architectures . . . 16
  - 4.1.1 Sequence-based Models (Transformers) . . . 16
  - 4.1.2 Graph Neural Networks . . . 16
- 4.2 GFT Strategy . . . 16
  - 4.2.1 Formal Definition . . . 16
  - 4.2.2 Stopping Condition . . . 17
- 4.3 Fine tuning details and hyperparameters . . . 17
  - 4.3.1 Stopping criteria . . . 17
  - 4.3.2 Data Splitting . . . 17
- 4.4 Metrics . . . 18
- 4.5 Results . . . 18
  - 4.5.1 BACE Dataset (Classification) . . . 18
  - 4.5.2 HIV Dataset (Classification) . . . 19
  - 4.5.3 BBBP Dataset (Classification) . . . 19
  - 4.5.4 CEP Dataset (Regression) . . . 19
  - 4.5.5 Lipophilicity Dataset (Regression) . . . 19
  - 4.5.6 Malaria Dataset (Regression) . . . 19

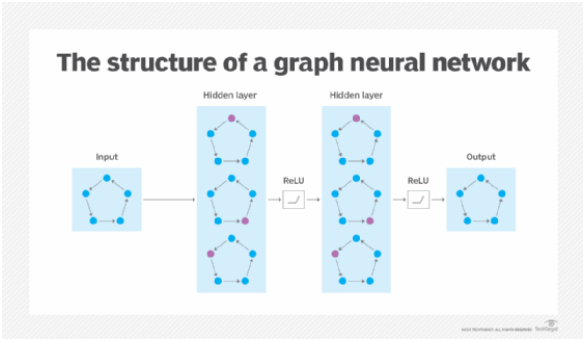
5 Evaluation **20**
5.1 BACE Dataset (Classification) 20
5.2 Dataset HIV (Classification) 21
5.3 Dataset BBBP (Classification) 21
5.4 CEP Dataset (Regression) 22
5.5 Lipophilicity Dataset (Regression) 23
5.6 Malaria Dataset (Regression) 24
5.7 Explainability Analysis 25
5.7.1 Motivation and Methodology 25
5.7.2 Feature Engineering 25
5.7.3 Key Findings 26
5.7.4 Implications for GFT Strategy 27

6 Conclusions **28**
6.1 Key Findings Across Datasets 28
6.2 Explainability Insights 28
6.3 Broader Impact 29

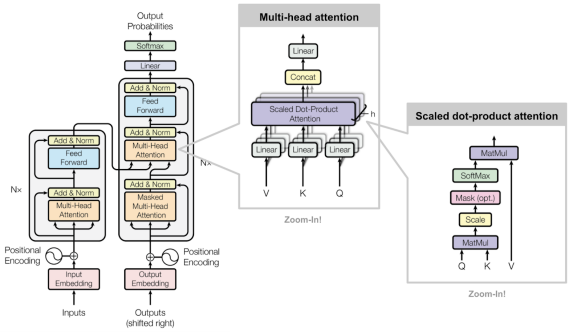
1 Business Understanding

1.1 Context and Motivation

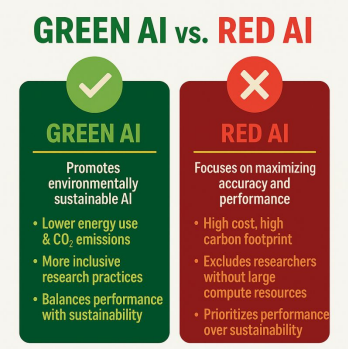
In recent years, the field of computational chemistry has seen a surge in the use of Deep Learning models, such as Transformers [2] and Graph Neural Networks (GNNs) [3], to accelerate drug discovery. However, the trend towards larger and more complex modelsoften referred to as "Red AI" [4]has led to a dramatic increase in computational costs and energy consumption. Following the **Green AI** paradigm [4], this project addresses the urgent need to make these processes sustainable. The motivation is twofold: ethical (reducing the carbon footprint) and economical (lowering the computational resources required for training), without abandoning the precision required for scientific discovery. Furthermore, this approach aligns with the **UN 2030 Agenda for Sustainable Development** [5], specifically contributing to **Goal 12 (Responsible Consumption and Production)** and **Goal 13 (Climate Action)**, by promoting energy-efficient innovation in the technological sector.



(a) Graph Neural Networks architecture



(b) Transformer architecture



(c) Green AI vs Red AI

(d) UN 2030 Agenda for Sustainable Development Goals

1.2 Business Objectives

1.2.1 Requirements

The primary business objective is to demonstrate that sustainable practices can be integrated into molecular property prediction pipelines without significant loss in performance. Specifically, the goals are:

- **Sustainability:** Drastically reduce the CO_2 equivalent emissions associated with the fine tuning phase of already existing pre-trained models.
- **Reliability:** Ensure that the "greener" models remain accurate enough to be useful for chemists and researchers in real-world screening scenarios.

1.2.2 Expectations

From a project management perspective, the experiment will be considered successful if: Based on previous similar approaches [6] the expected impact includes a reduction in CO_2 equivalent emissions between 20% and 50%, while maintaining predictive performance within a 10% margin of error compared to traditional fine-tuning methods.

1.3 Assess situation

Concerning models, the following architectures will be considered:

- **ChemBERTa:** A Transformer-based model pre-trained on a large corpus of molecular SMILES strings [7].
- **ChemBERTa2**: A bigger and improved version of ChemBERTa [8].
- **SELFormer**: A recent Transformer architecture specifically designed for molecular property prediction tasks [9].
- **SMILES_BERT**: A BERT model trained on a list of SMILES strings for various chemical tasks [10].
- **GraphMAE**: A GNN pre-trained using a masked autoencoder approach [11].

For energy consumption tracking CodeCarbon [12] will be used to monitor and log the energy usage and carbon emissions during the fine-tuning processes. CodeCarbon uses the following formula to estimate CO_2 emissions:

$$\text{CO}_2\text{eq} = \text{Energy Consumption (kWh)} \times \text{Carbon Intensity (g}\text{CO}_2\text{eq/kWh)} \tag{1}$$

where Carbon Intensity depends on the geographical location (in this case Apulia, Italy) and energy source mix of the data center where the computations are performed.
The project will use MoleculeNet [13], a collection of datasets for predicting molecular properties. In particular the following one will be used:

- **HIV**: A dataset used to predict if the HIV is active or inactive
- **BACE**: A dataset used to predict inhibitors of the human β-secretase 1 (BACE-1) enzyme (a target for Alzheimer's disease).
- **BBBP**: A dataset used to predict blood-brain barrier penetration (this barrier is crucial for drug delivery to the brain).
- **Lipophilicity**: A dataset used to predict the octanol/water distribution coefficient (logD at pH 7.4) of small molecules (a key property in drug design).
- **Malaria**: A dataset used to predict the ability of compounds to inhibit the growth of the malaria parasite *Plasmodium falciparum*.
- **CEP**: A dataset used to predict the efficiency of organic photovoltaic molecules (used in solar cells).

The hardware resources available for the experiments include:

- **CPU**: Intel(R) Core(TM) i714650HX
- **GPU**: NVIDIA GeForce RTX 4070 Laptop GPU with 8GB of VRAM
- **RAM**: 16 GB of SO-DIMM DDR5

1.4 Data Mining Goals

To achieve the business objectives outlined above, these objectives are translated into specific technical Data Mining goals. The core task involves supervised learning (both classification and regression) on molecular datasets. Dataset used for classification tasks are: HIV, BACE, BBB, while for regression tasks are: Lipophilicity, Malaria, CEP.

The technical objectives are:

- **Algorithm Implementation:** Develop and integrate the **Green FineTuning (GFT)** mechanism into the training loop of Transformers and GNNs.
- **Metric Evaluation:** Compare the GFT strategy against standard fine-tuning procedures using: Area under the Curve (ROC-AUC) for classification and Relative Squared Error (RSE) for regression tasks.
- **Trade-off Analysis:** Identify the optimal stopping point where the marginal gain in accuracy no longer justifies the marginal cost in emissions.
- **Explainability:** Analyze the results obtained and, using post-hoc explainability techniques, identify the factors that mainly contribute to sustainable fine-tuning

2 Data Understanding

2.1 SMILES Representation

The datasets utilized in this study primarily rely on the **SMILES** (Simplified Molecular Input Line Entry System) notation to represent chemical structures. SMILES is a line notation for encoding molecular structures using short ASCII strings, which allows chemical information to be easily stored and processed by computers.

In a SMILES string:

- **Atoms** are represented by their atomic symbols (e.g., `C` for Carbon, `N` for Nitrogen). Upper case letters usually indicate aliphatic (an "open chain" or a "simple ring") atoms, while lower case letters (e.g., `c`, `n`) denote aromatic (a "complex ring") atoms.
- **Bonds** are implied to be single unless specified otherwise. Double bonds are represented by `=`, triple bonds by `#`, and aromatic bonds are often implicit or denoted by colons.
- **Branching** is described using parentheses. For example, `CC(O)C` represents isopropanol.
- **Ring closures** are indicated by pairs of digits following the ring atoms. For instance, `C1CCCCC1` represents cyclohexane, where the two `1`s indicate the connection point of the ring.

This representation is crucial for machine learning tasks in chemistry (Cheminformatics) as it converts complex 3D topological graphs into a sequential 1D format that can be tokenized and processed by sequence-based models (such as Transformers) or reconstructed into graphs for Graph Neural Networks.

2.2 Dataset Description

For the experiments, standard datasets from the **MoleculeNet** suite were used, representative of various chemical-physical and biological properties. The selected datasets include BACE, BBBP, CEP, HIV, MALARIA and Lipophilicity, covering both classification and regression tasks.

2.2.1 BACE

Component	Description
mol	SMILES string representing the molecule
CID	Compound identifier (string)
Class	Binary label - categorical
pIC50	Numeric activity value - continue
Descriptors	Large set of chemical and topological descriptors (MW, AlogP, HBA, HBD, RB, Zagreb indices, Wiener index, connectivity indices, etc.).
Rows	1513 rows

Table 1: BACE raw CSV components

2.2.2 BBBP

Component	Description
num	Numeric progressive identifier - discrete
name	Compound name (string)
p_np	Binary label indicating blood-brain barrier permeability - categorical
smiles	SMILES string of the molecule
Rows	2050 rows

Table 2: BBBP raw CSV components

2.2.3 CEP

Component	Description
smiles	SMILES string of the organic molecule.
PCE	Power Conversion Efficiency - continue
Rows	29978 rows

Table 3: CEP raw CSV components

2.2.4 HIV

Component	Description
smiles	SMILES string representing the molecule.
activity	Descriptive activity label (e.g. CI = Confirmed Inactive, CA = Confirmed Active, CM = Confirmed Moderately Active) - categorical
HIV_active	Binary numeric label (CA and CM are considered as the same) - categorical
Rows	41126 rows

Table 4: HIV raw CSV components

2.2.5 Malaria

Component	Description
smiles	SMILES string of the compound
activity	Numeric activity value against malaria - continuous
Rows	9999 rows.

Table 5: Malaria raw CSV components

2.2.6 Lipophilicity

Component	Description
CMPD_CHEMBLID	ChEMBL compound identifier
exp	Experimental lipophilicity value continue
smiles	SMILES string of the molecule
Rows	4200 rows.

Table 6: Lipophilicity raw CSV components

2.3 Data Exploration

An initial exploratory data analysis (EDA) was conducted to understand the characteristics of each dataset. This included checking for missing values and visualization of the distributions. Luckily, all datasets were found to be clean, with no missing values.

2.4 Data Distribution Analysis

2.4.1 BACE

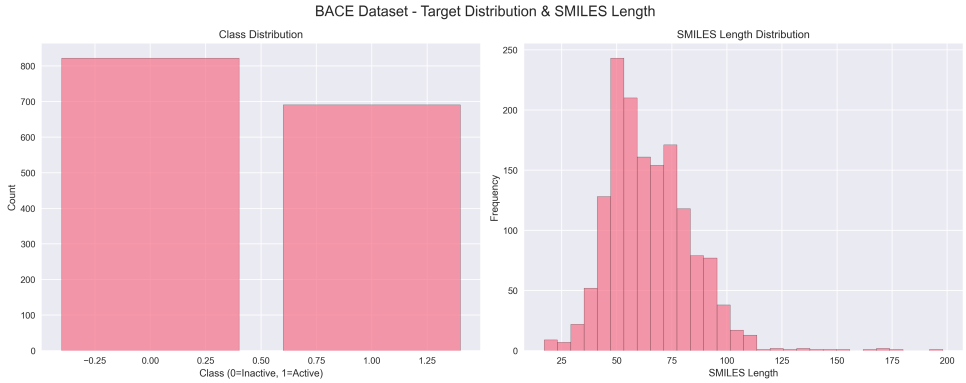


Figure 1: Data distribution analysis for BACE dataset

From the analysis it can be observed that classes are slightly imbalanced, with a higher number of inactive molecules compared to active ones. Regarding the lenght of the SMILES strings, most molecules have a length between 40 and 80 characters, with a few outliers having longer or shorter representations.

2.4.2 BBBP

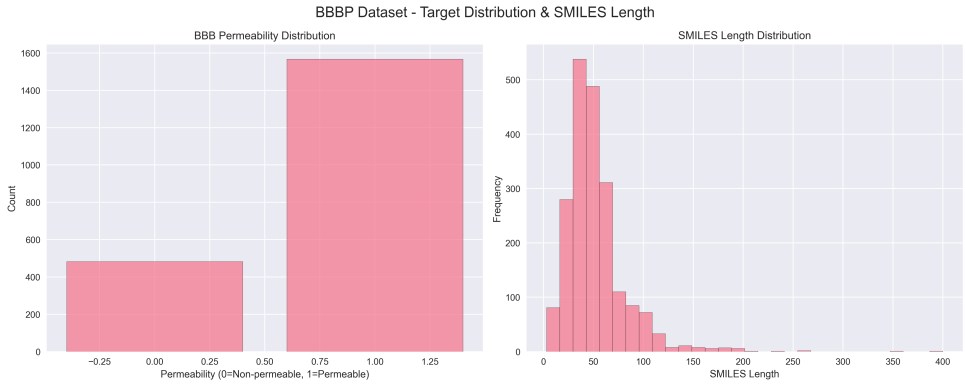


Figure 2: Data distribution analysis for BBBP dataset

From the analysis it can be observed that classes are strongly imbalanced, with a higher number of inactive molecules compared to active ones. Regarding the lenght of the SMILES strings, most molecules have a length between 40 and 60 characters, with a few outliers having longer or shorter representations.

2.4.3 CEP

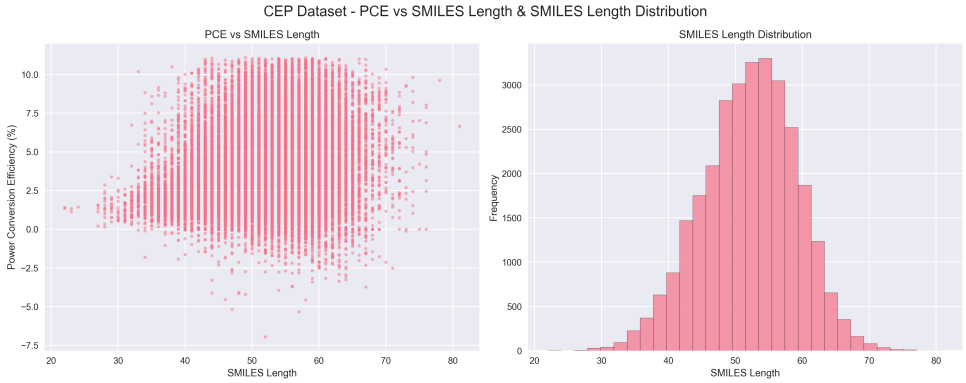


Figure 3: Data distribution analysis for CEP dataset

From the analysis it can be observed that PCE values are very close in the space with respect to the lenght of the SMILES strings with few outliers having lower PCE values. The lenght of the SMILES strings seems to follow a normal distribution

2.4.4 HIV

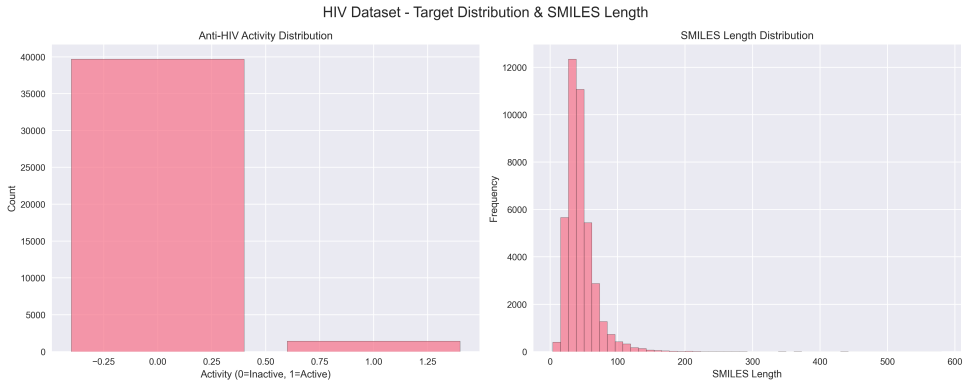


Figure 4: Data distribution analysis for HIV dataset

From the analysis it can be observed that classes are strongly imbalanced, with a higher number of inactive molecules compared to active ones. Regarding the lenght of the SMILES strings, it seems to follow a normal distribution with some outliers having longer representations.

2.4.5 Malaria

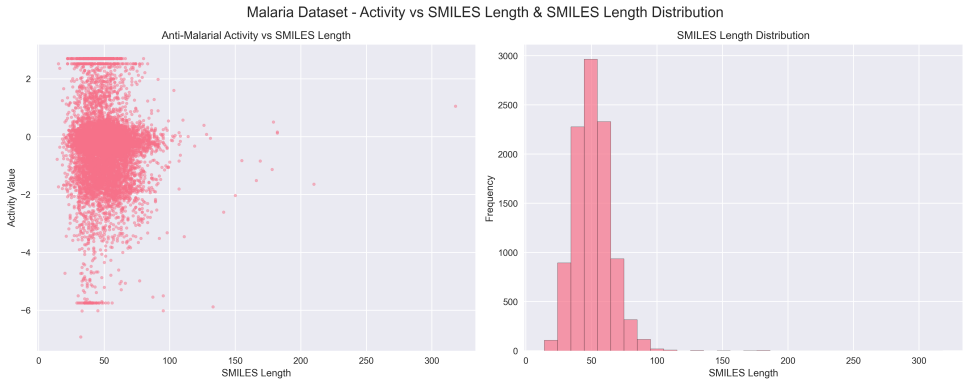


Figure 5: Data distribution analysis for Malaria dataset

From the analysis it can be observed that activity values are very close in the space with respect to the lenght of the SMILES strings with few outliers having lower values. The lenght of the SMILES strings seems to follow a normal distribution

2.4.6 Lipophilicity

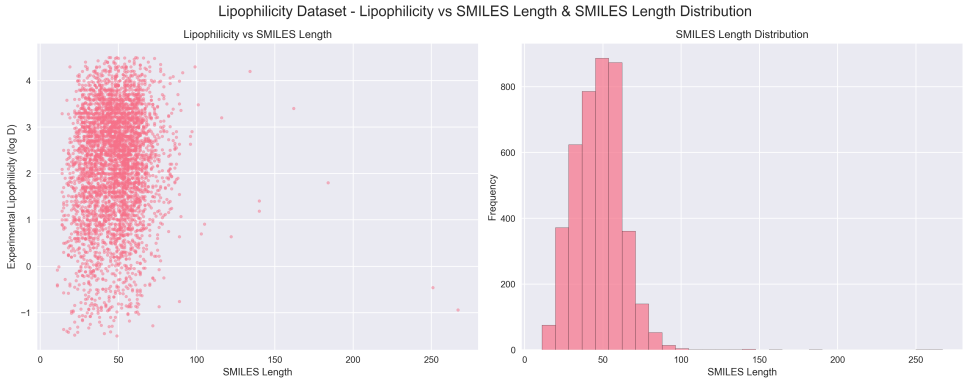


Figure 6: Data distribution analysis for Lipophilicity dataset

From the analysis it can be observed that exp values are very close in the space with respect to the lenght of the SMILES strings with few outliers having lower values. The lenght of the SMILES strings are concentrated between 20 and 80

3 Data Preparation

3.1 Preprocessing

For all datasets, two variants were prepared to accommodate different model architectures:

- **Transformer variant**: Only the SMILES string and the target class/label are retained. This format is suitable for sequence-based models, which process molecular representations as text sequences.
- **Graph variant**: Molecular graphs are generated from SMILES strings using the RDKit library. This representation captures the structural connectivity of atoms and bonds, making it ideal for graph neural networks like GraphMAE.

To ensure data quality, invalid SMILES strings were filtered out during preprocessing. A SMILES is considered valid if it can be successfully parsed into a molecular object by RDKit's `Chem.MolFromSmiles` function and passes sanitization checks (`Chem.SanitizeMol`). Invalid SMILES, which may represent malformed or chemically impossible structures, are excluded from the datasets to prevent errors in model training and evaluation.

3.1.1 SMILES Validation

Prior to graph generation, a rigorous validation pipeline is applied to ensure data integrity. Raw datasets often contain malformed strings or chemically impossible structures that can lead to runtime errors or degrade model performance.

The validation process relies on RDKit's internal chemistry engine and consists of two main stages:

1. **Parsing:** The raw SMILES string is parsed using the `Chem.MolFromSmiles` function. This step checks for syntax errors in the ASCII string (e.g., unclosed parentheses, invalid characters, or numbers indicating ring closures that do not match). If parsing fails, the function returns `None`, and the sample is discarded.
2. **Sanitization:** Successfully parsed molecules undergo a sanitization process via `Chem.SanitizeMol`. This function performs a series of chemical consistency checks, including:
    - **Valence Check:** Verifies that atoms do not exceed their maximum possible valence (e.g., a carbon atom with 5 bonds).
    - **Kekulization:** Converts aromatic rings into their alternating single-double bond representation (Kekulé form) to ensure structural correctness.
    - **Aromaticity Detection:** Identifies aromatic systems and ensures they obey Huckel's rule.

Only molecules that pass both parsing and sanitization are retained for the final dataset. This filtering step ensures that the resulting graphs represent physically valid chemical entities.

Dataset	Valid SMILES
BACE	1.513
BBBP	2.039
CEP	29.978
HIV	41.119
Malaria	9.999
Lipophilicity	4.200

Table 7: Summary of SMILES validation results across datasets.

3.1.2 Graph Generation

The transformation of 1D SMILES strings into graph structures is a critical step for enabling Geometric Deep Learning models. In this work, the graph generation process is performed using the **RDKit** library, which converts the raw string representation into a molecular graph object $G = (V, E)$.

In this graph representation:

- **Nodes (V)** represent the atoms in the molecule.
- **Edges (E)** represent the chemical bonds connecting them.

The process involves extracting specific chemical features for both atoms and bonds to create rich numerical representations (node features matrix X and edge attributes E_{attr}).

Node Featurization For each atom $v \in V$, a feature vector x_v is constructed using to capture essential chemical properties. The extracted features include:

- **Atomic Number:** Specifies the element type (e.g., C, N, O, F, etc.).
- **Chirality:** Describes the stereochemical configuration (e.g., unspecified, tetrahedral CW/CCW).
- **Degree:** The number of covalent bonds the atom forms.
- **Formal Charge:** The electrical charge assigned to the atom.
- **Num. Explicit Hs:** The number of hydrogen atoms explicitly attached.
- **Radical Electrons:** The number of unpaired electrons.
- **Hybridization:** The orbital hybridization state (sp, sp^2, sp^3, etc.).
- **Aromaticity:** A boolean flag indicating whether the atom is part of an aromatic ring system.
- **Explicit Valence:** The explicit valence of the atom.

This results in a node feature matrix $X \in \mathbb{R}^{|V| \times F_{node}}$, where $|V|$ is the number of atoms and F_{node} is the dimensionality of the atom feature vector.

Edge Featurization For each bond $e_{ij} \in E$ connecting atom i and atom j, a feature vector is created to describe the bond properties:

- **Bond Type:** Single, Double, Triple, or Aromatic.
- **Stereochemistry:** Stereo configuration (e.g., None, Z, E, Any).
- **Conjugation:** Boolean flag indicating if the bond is conjugated.

The connectivity of the graph is stored in coordinate format (COO), creating an edge index tensor necessary for message-passing operations in GNNs.

Graph Construction and Serialization Once the feature extraction is complete, the individual components are converted into PyTorch tensors and aggregated into a unified structure. Specifically:

- The node features matrix X is converted to a tensor of shape $[|V|, F_{node}]$.
- The connectivity list is converted to an `edge_index` tensor of shape $[2, |E|]$ (LongTensor).
- The edge attributes are converted to a tensor of shape $[|E|, F_{edge}]$.
- The target label y is attached as a tensor (e.g., float for regression tasks or long for classification).

These tensors are encapsulated into a single PyTorch Geometric `Data` object, which acts as a container for the entire graph instance. Finally, the processed dataset is serialized and saved to disk as a binary file with the `.pt` extension using `torch.save`. This format allows for efficient loading during the training phase, enabling the `DataLoader` to dynamically batch multiple graphs into a single large disconnected graph for parallel processing on the GPU.

4 Modelling

This section details the deep learning architectures employed for molecular property prediction and introduces the proposed adaptive fine tuning strategy designed to optimize the trade-off between performance and environmental impact.

4.1 Architectures

Molecular representation learning is approached through two distinct paradigms: sequence-based modelling (treating molecules as SMILES strings) and graph-based modelling (treating molecules as molecular graphs).

4.1.1 Sequence-based Models (Transformers)

In this paradigm, molecules are represented as SMILES strings. The Transformer architecture [2] is leveraged, specifically models pre-trained on large chemical corpora via Masked Language Modeling (MLM). The specific models evaluated are:

- **ChemBERTa**
- **ChemBERTa-2**
- **SELFormer**
- **SMILES-BERT**

4.1.2 Graph Neural Networks

Alternatively, molecules are naturally represented as graphs $G = (V, E)$, where atoms constitute the nodes V and chemical bonds the edges E. To this end, **GraphMAE** is employed. Unlike standard GNNs, GraphMAE focuses on self-supervised learning by masking a portion of the input graph (nodes or edges) and reconstructing it, forcing the model to learn robust structural representations.

4.2 GFT Strategy

To address the environmental concerns of fine tuning deep learning models, a custom callback mechanism is introduced: the Green FineTuning. Standard fine tuning procedures typically rely on "patience" based solely on validation loss. In contrast, this approach integrates energy consumption directly into the stopping criterion.

4.2.1 Formal Definition

The core idea is to interrupt fine tuning when the marginal gain in predictive performance no longer justifies the marginal energy cost. We define the instantaneous GFT ratio (GFT_t) at epoch t as:

$$GFT_t = \frac{\Delta P_t}{\Delta E_t} = \frac{Perf_t - Perf_{t-1}}{Emission_t - Emission_{t-1}} \tag{2}$$

where ΔP_t represents the percentage improvement in the validation metric and ΔE_t is the incremental CO_2eq produced during the epoch.

4.2.2 Stopping Condition

Since raw metrics can be volatile, we employ an Exponential Moving Average (EMA) to smooth the GFT signal, ensuring stability in the decision process. The smoothed value S_t is updated as follows:

$$S_t = \alpha \cdot GFT_t + (1 - \alpha) \cdot S_{t-1} \tag{3}$$

where $\alpha \in [0, 1]$ is the smoothing factor.

4.3 Fine tuning details and hyperparameters

Parameter	Value	Description
Seed	42	seed fixed for reproducibility
Alpha (α)	0.9	EMA smoothing factor for GFT
Beta (β)	0.2	GFT threshold factor
Warmup epochs	3 (Transformers), variable (GraphMAE)	Initial epochs before applying early stopping
Patience	5	Patience for classic early stopping
Optimizer	Adam	Optimization algorithm
Learning rate	10^{-4}	Step size for optimizer
Batch size	32	Number of samples per gradient update
Epochs	30 (Transformers), 100 (GraphMAE)	Maximum number of training epochs

Table 8: Fine tuning hyperparameters and configuration.

4.3.1 Stopping criteria

The fine tuning process leverages two distinct stopping mechanisms depending on the chosen strategy.

Classic Strategy The baseline approach uses a traditional early stopping protocol based exclusively on predictive performance. This strategy monitors the validation metric and halts fine tuning if the model fails to improve upon the best recorded performance for a consecutive number of epochs specified by the *patience* parameter. In this mode, energy consumption is tracked for comparison purposes but does not influence the stopping decision.

Green-Early Strategy (GFT) The proposed adaptive strategy dynamically evaluates the trade-off between performance gain and energy cost at each evaluation step. Instead of relying on a fixed patience, the algorithm computes the instantaneous efficiency ratio GFT_t, defined as the quotient between the relative improvement in validation performance (ΔP) and the relative increase in cumulative emissions (ΔE). To filter out noise and volatility inherent in training metrics, this instantaneous ratio is smoothed using an EMA, denoted as S_t. The stopping decision is governed by a tolerance threshold β. The training halts when the current instantaneous efficiency (GFT_t) falls below a fraction β of the historical smoothed trend (S_t):

$$GFT_t < \beta \cdot S_t$$

4.3.2 Data Splitting

The dataset was partitioned into fine tuning (80%), validation (10%), and test (10%) sets using a scaffold splitting strategy. Unlike random splitting, this method segregates molecules based on their two-dimensional structural frameworks (Bemis-Murcko scaffolds) computed using RDKit, thereby testing the model's ability to generalize to unseen chemical spaces. Scaffolds were identified and sorted by cluster size in descending order to ensure that the most representative structures were prioritized for the training set.

To address the potential issue of class imbalancewhere a scaffold-based split might result in validation or test sets lacking positive or negative examples a custom balancing algorithm was applied. This procedure verifies the presence of all class labels in the target subsets. If a subset (validation or test) contains fewer than two classes (in classification tasks), the algorithm searches the training set for scaffolds containing the missing label. To minimize the impact on dataset distribution, the algorithm prioritizes the smallest available scaffolds in the training set, moving them to the target set to restore class diversity while preserving structural separation. The training set is used to fine-tune the models, the validation set guides early stopping decisions, and the test set provides an unbiased evaluation of the final model performance.

4.4 Metrics

For classification tasks the ROC-AUC metric

$$\text{ROC-AUC} = \frac{1}{2}\sum_{i=1}^{n}(X_i - X_{i-1})(Y_i + Y_{i-1}) \tag{4}$$

where (X_i, Y_i) are the points on the ROC curve, is used for both validation and test evaluations. In this case higher values indicate better performance. For regression tasks two different metrics were used:

- Relative Mean Squared Error (RMSE) metric

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

  where y_i are the true values and $\hat{y}_i$ are the predicted values, is used for validation evaluations.

- Relative Squared Error (RSE) metric

$$\text{RSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

  where y_i are the true values, $\hat{y}_i$ are the predicted values, and $\bar{y}$ is the mean of the true values, is used for test evaluations.

In both regression cases lower values indicate better performance.

4.5 Results

4.5.1 BACE Dataset (Classification)

Model	AUROC (Classic)	AUROC (Early)	Δ Perf.	Ep. (C.)	Ep. (E.)	Emis. C.	Emis. E.	Δ Emiss.
ChemBERTa	0.819	0.850	+3.8%	10	5	9.87×10^{-4}	5.68×10^{-4}	-42.4%
ChemBERTa-2	0.817	0.855	+4.6%	21	5	3.87×10^{-4}	2.70×10^{-4}	-30.3%
SELFormer	0.819	0.832	+1.6%	8	6	1.52×10^{-3}	1.18×10^{-3}	-22.4%
SMILES-BERT	0.789	0.844	+7.0%	15	5	2.71×10^{-3}	1.07×10^{-3}	-60.5%
GraphMAE (W10)	0.680	0.667	-1.9%	14	10	1.07×10^{-4}	7.18×10^{-5}	-33.0%
GraphMAE (W25)	0.711	0.700	-1.5%	29	25	1.91×10^{-4}	1.65×10^{-4}	-13.4%
GraphMAE (W50)	0.758	0.748	-1.4%	54	50	3.50×10^{-4}	4.17×10^{-4}	*+19.0%*

Table 9: Comparison on Bace dataset

4.5.2 HIV Dataset (Classification)

Model	AUROC (Classic)	AUROC (Early)	Δ Perf.	Ep. (C.)	Ep. (E.)	Emis. C.	Emis. E.	Δ Emiss.
ChemBERTa	0.628	0.721	+14.8%	9	11	1.97×10^{-2}	2.40×10^{-2}	*+22.1%*
ChemBERTa-2	0.784	0.783	-0.1%	14	5	4.31×10^{-3}	1.70×10^{-3}	-60.6%
SELFormer	0.559	0.650	+16.2%	9	5	3.86×10^{-2}	2.16×10^{-2}	-44.0%
SMILES-BERT	0.473	0.552	+16.6%	9	6	3.98×10^{-2}	2.66×10^{-2}	-33.1%
GraphMAE (W10)	0.717	0.708	-1.2%	14	10	1.66×10^{-3}	1.37×10^{-3}	-17.7%
GraphMAE (W25)	0.749	0.741	-1.1%	29	25	3.93×10^{-3}	3.43×10^{-3}	-12.8%
GraphMAE (W50)	0.762	0.764	+0.3%	54	50	7.33×10^{-3}	6.24×10^{-3}	-14.8%

Table 10: Comparison on Hiv dataset

4.5.3 BBBP Dataset (Classification)

Model	AUROC (Classic)	AUROC (Early)	Δ Perf.	Ep. (C.)	Ep. (E.)	Emis. C.	Emis. E.	Δ Emiss.
ChemBERTa	0.708	0.625	-11.8%	9	5	3.58×10^{-4}	2.95×10^{-4}	-17.5%
ChemBERTa-2	0.771	0.667	-13.5%	13	5	2.03×10^{-4}	2.11×10^{-4}	*+4.2%*
SELFormer	0.604	0.646	+6.9%	9	5	3.52×10^{-4}	3.09×10^{-4}	-12.1%
SMILES-BERT	0.542	0.438	-19.2%	9	5	4.10×10^{-4}	3.07×10^{-4}	-25.2%
GraphMAE (W10)	0.455	0.424	-6.7%	14	10	9.72×10^{-6}	5.37×10^{-6}	-44.7%
GraphMAE (W25)	0.485	0.485	*Unchanged*	29	25	1.59×10^{-5}	1.18×10^{-5}	-25.6%
GraphMAE (W50)	0.576	0.515	-10.5%	54	50	2.75×10^{-5}	2.31×10^{-5}	-16.0%

Table 11: Comparison on Bbbp dataset

4.5.4 CEP Dataset (Regression)

Model	RSE (Classic)	RSE (Early)	Δ Perf.	Ep. (C.)	Ep. (E.)	Emis. C.	Emis. E.	Δ Emiss.
ChemBERTa	0.306	0.289	+5.5%	12	7	1.93×10^{-2}	1.13×10^{-2}	-41.2%
ChemBERTa-2	0.337	0.280	+16.9%	18	5	3.52×10^{-3}	1.11×10^{-3}	-68.6%
SELFormer	0.567	0.280	+50.6%	15	6	4.76×10^{-2}	1.92×10^{-2}	-59.7%
SMILES-BERT	1.009	0.980	+2.9%	16	5	5.22×10^{-2}	1.65×10^{-2}	-68.4%
GraphMAE (W10)	0.411	0.525	-27.8%	100	10	9.64×10^{-3}	9.72×10^{-4}	-89.9%
GraphMAE (W25)	0.423	0.481	-13.9%	91	26	8.73×10^{-3}	2.51×10^{-3}	-71.3%
GraphMAE (W50)	0.417	0.454	-8.8%	99	50	9.53×10^{-3}	4.81×10^{-3}	-49.6%

Table 12: Comparison on Cep dataset

4.5.5 Lipophilicity Dataset (Regression)

Model	RSE (Classic)	RSE (Early)	Δ Perf.	Ep. (C.)	Ep. (E.)	Emis. C.	Emis. E.	Δ Emiss.
ChemBERTa	0.525	0.601	-14.6%	21	5	4.90×10^{-3}	1.34×10^{-3}	-72.6%
ChemBERTa-2	0.452	0.452	-0.1%	23	5	8.20×10^{-4}	3.40×10^{-4}	-58.5%
SELFormer	0.655	0.649	+0.9%	7	6	3.30×10^{-3}	2.85×10^{-3}	-13.7%
SMILES-BERT	0.655	0.725	-10.7%	10	5	4.75×10^{-3}	2.49×10^{-3}	-47.6%
GraphMAE (W10)	0.608	0.968	-59.3%	100	10	1.31×10^{-3}	1.36×10^{-4}	-89.6%
GraphMAE (W25)	0.608	0.830	-36.6%	100	25	1.31×10^{-3}	3.33×10^{-4}	-74.7%
GraphMAE (W50)	0.608	0.685	-12.6%	100	50	1.32×10^{-3}	6.53×10^{-4}	-50.6%

Table 13: Comparison on Lipophilicity dataset

4.5.6 Malaria Dataset (Regression)

Model	RSE (Classic)	RSE (Early)	Δ Perf.	Ep. (C.)	Ep. (E.)	Emis. C.	Emis. E.	Δ Emiss.
ChemBERTa	1.058	1.149	-8.7%	8	9	4.47×10^{-3}	5.00×10^{-3}	*+11.7%*
ChemBERTa-2	0.878	0.849	+3.3%	7	5	6.74×10^{-4}	5.15×10^{-4}	-23.6%
SELFormer	1.164	1.096	+5.9%	10	11	1.07×10^{-2}	1.18×10^{-2}	*+10.0%*
SMILES-BERT	0.991	0.991	*Unchanged*	14	5	1.53×10^{-2}	5.61×10^{-3}	-63.5%
GraphMAE (W10)	0.915	0.933	-2.0%	16	10	8.80×10^{-4}	5.45×10^{-4}	-38.1%
GraphMAE (W25)	0.980	0.953	+2.7%	29	25	1.66×10^{-3}	1.37×10^{-3}	-17.8%
GraphMAE (W50)	1.104	1.057	+4.3%	54	50	2.96×10^{-3}	2.73×10^{-3}	-7.9%

Table 14: Comparison on Malaria dataset

5 Evaluation

This section analyzes the experimental results obtained by applying the GFT strategy across various datasets and model architectures with respect to the Business Objectives defined earlier. The evaluation focuses on visualizing the trade-off between CO_2 equivalent emissions and predictive performance, as well as understanding the factors influencing these outcomes through explainability techniques.

5.1 BACE Dataset (Classification)

The BACE dataset demonstrates highly favorable results for the GFT, particularly with Transformer-based models. All four Transformer architectures (ChemBERTa, ChemBERTa-2, SELFormer, SMILES-BERT) show performance improvements ranging from +1.6% to +7.0% when using early stopping, with SMILES-BERT achieving the most substantial gain of +7.0%. This suggests that early stopping effectively prevents overfitting on this dataset.

Emission reductions are consistently strong across Transformer models, ranging from -22.4% to -60.5%, with SMILES-BERT achieving the highest emission reduction of -60.5% while simultaneously delivering the best performance improvement. This represents an ideal GFT scenario where environmental benefits align with enhanced model performance.

GraphMAE models show more mixed results, with slight performance degradations (-1.3% to -1.9%) but still achieving emission reductions in most configurations. Notably, GraphMAE W50 shows an emission increase of +19.0%, indicating that longer warmup periods may not be beneficial for this dataset complexity.

This suggests that Transformer architectures pre-trained on molecular corpora tends to overfit after few epochs on BACE (in fact GFT used less epochs than classic fine tuning while achieving better results). Surprisingly, GFT with 50 epochs warmup for GraphMAE leads to an higher emission than classic fine tuning, probably because the emission is "justified" by a better performance, but in reality the model is overfitting too, finding the best trade-off in the middle (25 epochs warmup).

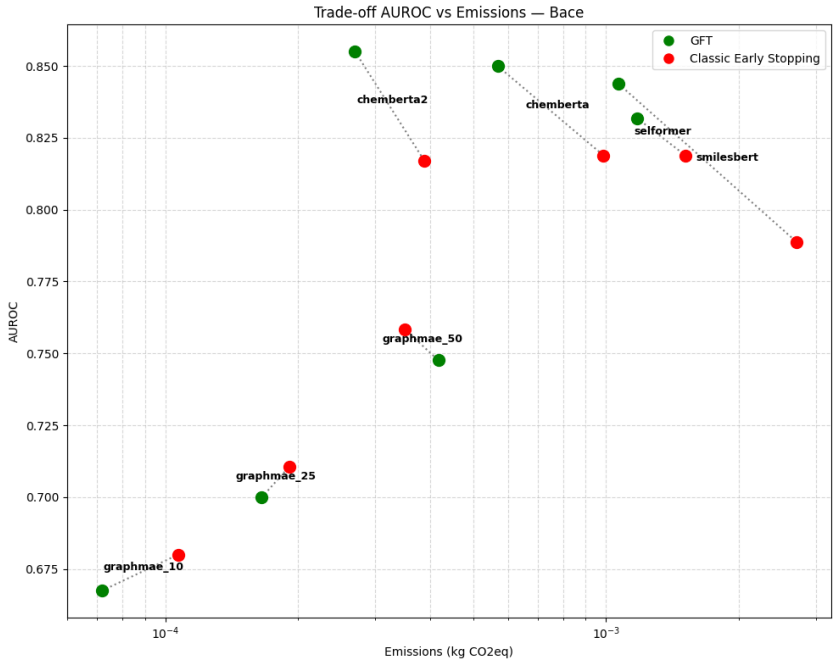


Figure 7: Trade-off comparison on BACE

5.2 Dataset HIV (Classification)

The HIV dataset presents particularly striking results for GFT, with some of the most dramatic performance improvements observed across all datasets. ChemBERTa shows exceptional improvement of +14.8%, though this comes at the cost of a +22.1% increase in emissions, representing one of the few cases where performance gains outweigh emission savings.

SELFormer and SMILES-BERT demonstrate remarkable consistency with improvements of +16.2% and +16.6% respectively, while achieving substantial emission reductions of -44.0% and -33.1%. ChemBERTa-2 maintains stable performance (-0.1%) while delivering significant emission savings of -60.6%.

GraphMAE models show minimal performance impact (-1.1% to +0.3%) with consistent emission reductions across all warmup configurations. The W50 configuration even achieves a slight performance improvement (+0.3%) while maintaining emission savings of -14.8%, suggesting that after 50 epochs the model starts to overfit.

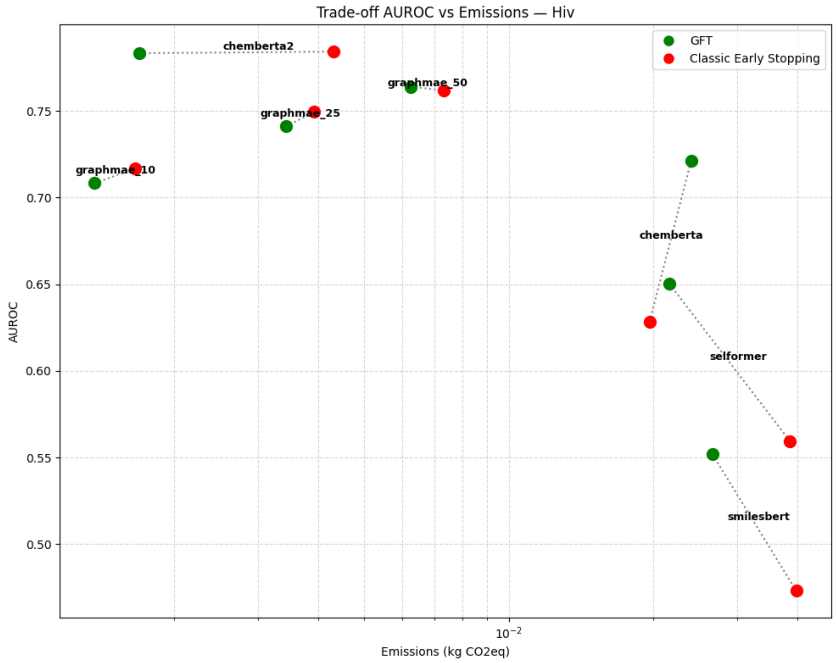


Figure 8: Trade-off comparison on HIV

5.3 Dataset BBBP (Classification)

The BBBP dataset reveals more challenging trade-offs, highlighting the importance of model selection for GFT strategies. The results show significant variability across different architectures, with some models experiencing substantial performance degradations.

Notably, ChemBERTa and SMILES-BERT show significant performance drops of -11.8% and -19.2% respectively, indicating that early stopping may be premature for these models on this dataset. However, SELFormer demonstrates resilience with a +6.9% improvement while achieving -12.1% emission reduction, suggesting that certain architectures are inherently more compatible with early stopping strategies.

ChemBERTa-2 shows the most concerning pattern with both performance degradation (-13.5%) and a slight emission increase (+4.2%), indicating poor optimization dynamics. GraphMAE models show moderate performance impacts (-6.7% to unchanged) while consistently achieving emission reductions ranging from -16.0% to -44.7%.

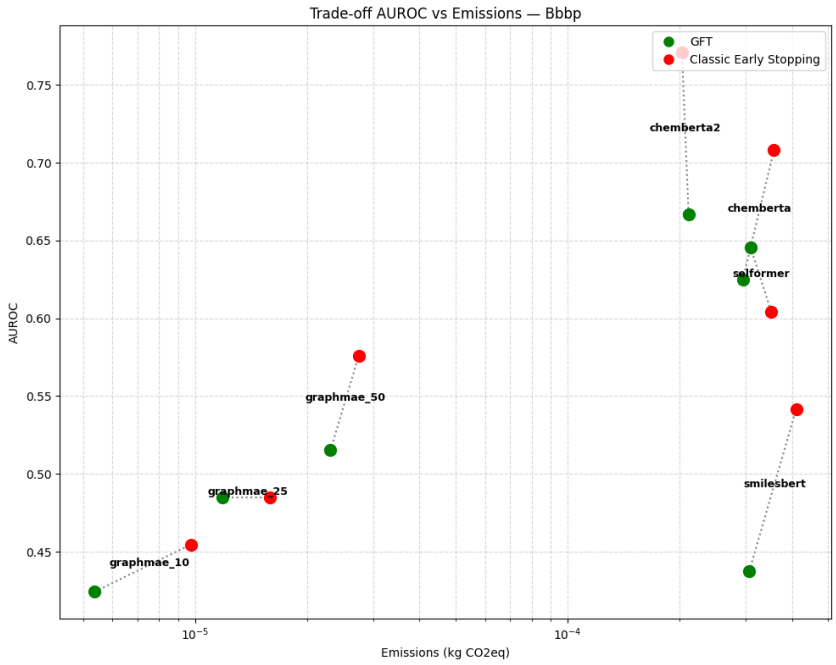


Figure 9: Trade-off comparison on BBBP

5.4 CEP Dataset (Regression)

The CEP dataset showcases some of the most impressive GFT results, particularly demonstrating that regression tasks can benefit substantially from early stopping when properly configured. SELFormer achieves an exceptional +50.6% performance improvement while reducing emissions by -59.7%, representing one of the strongest win-win scenarios observed across all experiments.

ChemBERTa-2 also delivers very good results with +16.9% performance improvement and -68.6% emission reduction, while ChemBERTa shows modest but positive gains (+5.5% performance, -41.2% emissions). SMILES-BERT maintains stable performance (+2.9%) with substantial emission savings (-68.4%).

GraphMAE models present a clear trade-off pattern where shorter warmup periods (W10) result in larger performance degradations (-27.8%) but achieve the most dramatic emission reductions (-89.9%). Longer warmup periods (W50) moderate the performance impact (-8.8%) while still achieving significant emission savings (-49.6%). This demonstrates the critical importance of warmup tuning for regression tasks with graph-based models.

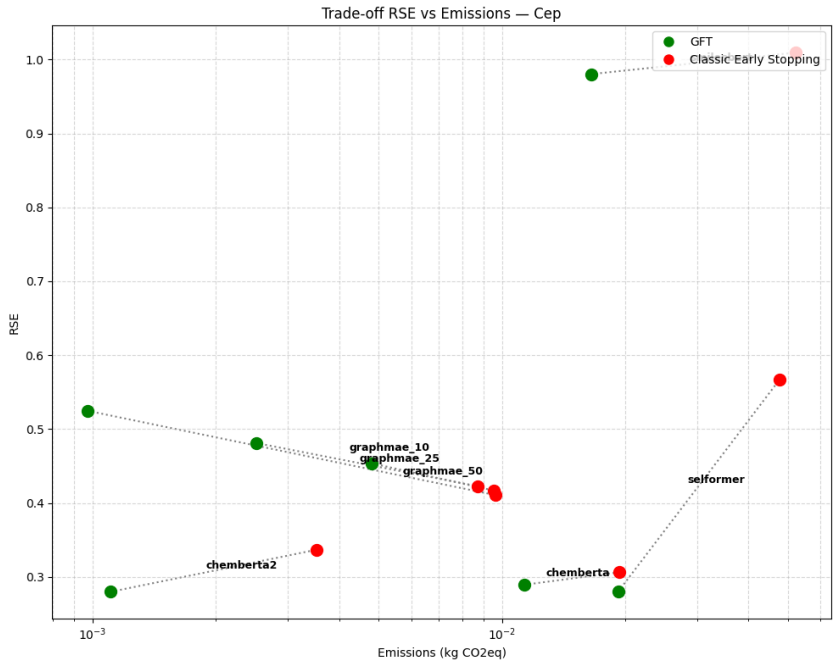


Figure 10: Trade-off comparison on CEP

5.5 Lipophilicity Dataset (Regression)

Metrics: RSE (lower is better).

The Lipophilicity dataset presents the most challenging scenario for GFT implementation, with most models experiencing performance degradations when early stopping is applied. This suggests that lipophilicity prediction requires extended fine tuning periods to achieve optimal performance.

SELFormer stands out as the only consistently positive case with a modest +0.9% improvement and -13.7% emission reduction. ChemBERTa-2 maintains nearly identical performance (-0.1%) while achieving substantial emission savings (-58.5%), making it a viable GFT option.

The most concerning results come from GraphMAE models, particularly with shorter warmup periods. GraphMAE W10 shows a dramatic -59.3% performance degradation, despite achieving -89.6% emission reduction. Even GraphMAE W50 shows significant performance loss (-12.6%) with -50.6% emission savings.

ChemBERTa and SMILES-BERT show moderate performance impacts (-10.7% to -14.6%) while achieving substantial emission reductions (-47.6% to -72.6%). These results highlight that for complex regression tasks like lipophilicity prediction, the trade-off between performance and emissions requires careful consideration, and extended fine tuning may be necessary for optimal results.

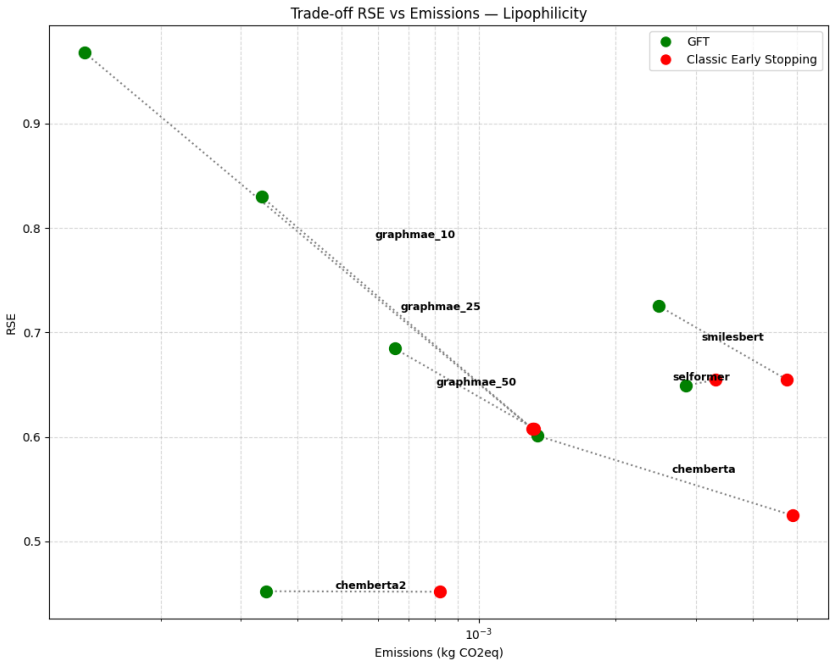


Figure 11: Trade-off comparison on Lipophilicity

5.6 Malaria Dataset (Regression)

Metrics: RSE (lower is better).

The Malaria dataset demonstrates moderate success for GFT strategies with several models showing positive performance impacts alongside emission reductions. ChemBERTa-2 delivers consistent improvement (+3.3%) with emission savings (-23.6%), while SELFormer shows +5.9% performance improvement, though with a slight emission increase (+10.0%).

SMILES-BERT maintains unchanged performance while achieving substantial emission reduction (-63.5%), representing an ideal efficiency scenario. ChemBERTa shows the only clear negative trade-off with -8.7% performance degradation and +11.7% emission increase, suggesting poor optimization dynamics for this model-dataset combination.

GraphMAE models demonstrate interesting patterns across warmup configurations. W10 shows slight performance degradation (-2.0%) with good emission savings (-38.1%), while W25 and W50 both achieve performance improvements (+2.7% and +4.3% respectively) with emission reductions (-17.8% and -7.9%). This suggests that longer warmup periods are beneficial for the Malaria dataset, allowing GraphMAE to find better optimization paths while maintaining efficiency gains.

Overall, the Malaria dataset shows that with proper model selection and configuration, regression tasks can achieve both performance and efficiency improvements.

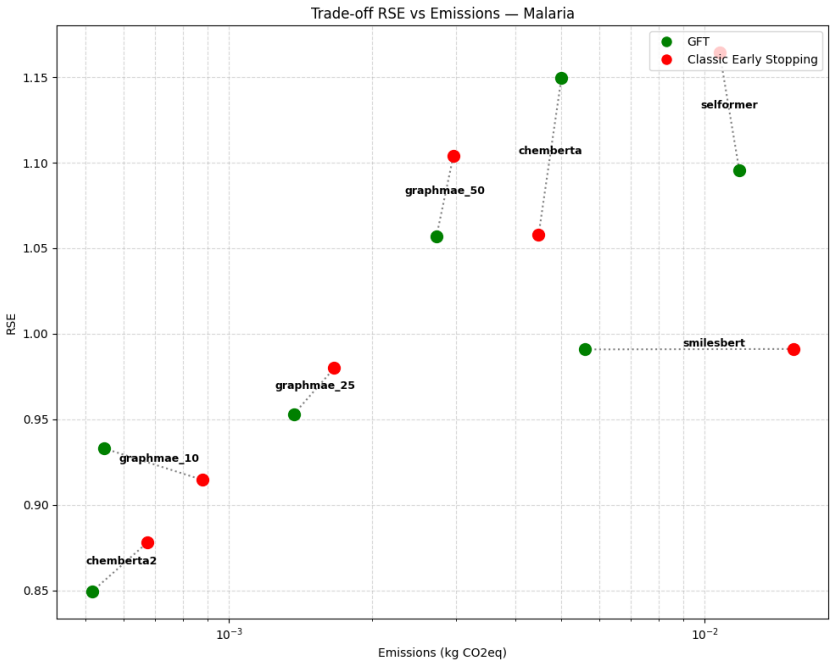


Figure 12: Trade-off comparison on Malaria

5.7 Explainability Analysis

5.7.1 Motivation and Methodology

Understanding which factors most influence model emissions and performance is crucial for optimizing GFT. To address this,a comprehensive explainability analysis using **Partial Correlation**, a statistical technique that isolates the unique contribution of each feature while controlling for all other variables, was conducted. This approach is particularly valuable when dealing with multicollinearity among features, which is common in machine learning experiments where multiple factors interact.

5.7.2 Feature Engineering

For each experiment, a comprehensive set of features was engineered to capture relevant aspects of model architecture, fine tuning configuration, dataset characteristics, and fine tuning strategy. The features included:

- **Model Architecture Features**: Number of parameters, hidden size, number of layers, model family (Transformer vs GNN)
- **Fine tuning Configuration**: Epochs provided, epochs used, warmup epochs, epoch efficiency (ratio of epochs used to epochs provided)
- **Dataset Characteristics**: Dataset size, average molecular weight, task type (classification vs regression)
- **Fine tuning Strategy**: Binary indicator for early stopping usage

All features were standardized using z-score normalization before analysis to ensure fair comparison across different scales.

5.7.3 Key Findings

Factors Influencing Emissions The partial correlation analysis reveals the following hierarchy of importance for predicting CO_2 emissions:

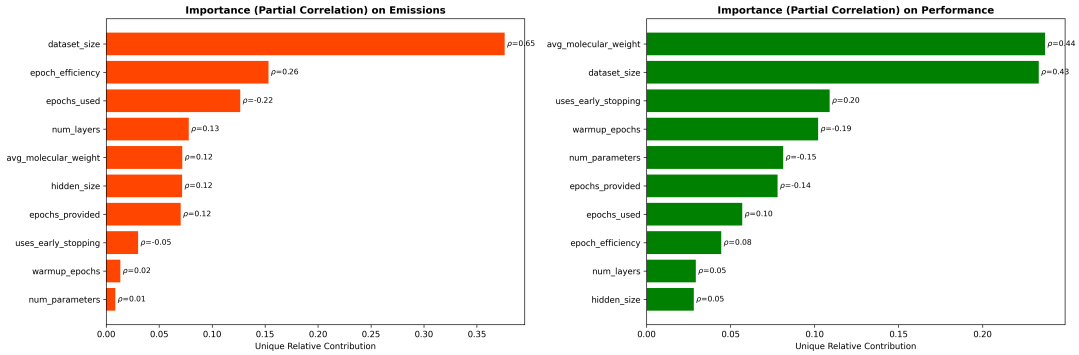


Figure 13: Feature importance for emissions (left) and performance (right) based on partial correlation analysis. The values shown are normalized relative contributions, with ρ indicating the direction and strength of the partial correlation.

Top emission drivers:

1. **Dataset Size** ($\rho = 0.65$): Largest contributor to emissions. Larger datasets require more computational resources per epoch and often necessitate longer fine tuning.
2. **Epoch Efficiency** ($\rho = 0.26$): Positive correlation indicates that longer fine tuning (higher epoch usage ratio) directly increases emissions.
3. **Epochs Used** ($\rho = -0.22$): Interestingly shows negative correlation when controlling for other factors, suggesting interaction effects with dataset size and efficiency.
4. **Model Architecture** (num_layers, hidden_size): Moderate influence ($\rho \approx 0.12 - 0.13$), indicating that while model size matters, it's less critical than dataset size and fine tuning duration.

Notably, **early stopping usage** shows minimal direct impact ($\rho = -0.05$), suggesting its emission reduction benefits are mediated primarily through reducing epochs used rather than any inherent efficiency gain.

Factors Influencing Performance The performance analysis reveals a different set of priorities:

Top performance drivers:

1. **Average Molecular Weight** ($\rho = 0.44$): Strongest predictor, likely reflecting dataset complexity and information content.
2. **Dataset Size** ($\rho = 0.43$): Critical for model generalization, consistent with standard machine learning principles.
3. **Early Stopping Usage** ($\rho = 0.20$): Positive contribution suggests early stopping helps prevent overfitting, validating the GFT approach.
4. **Warmup Epochs** ($\rho = -0.19$): Negative correlation indicates that excessive warmup may delay convergence or lead to suboptimal fine tuning dynamics for some tasks.
5. **Model Capacity** (num_parameters, $\rho = -0.15$): Surprisingly negative, potentially indicating overfitting with larger models or diminishing returns beyond a certain model size for these molecular tasks.

5.7.4 Implications for GFT Strategy

The explainability analysis provides actionable insights:

1. **Prioritize Dataset Optimization**: Since dataset size is the dominant factor for emissions, techniques like data pruning, active learning, or efficient data sampling could yield substantial emission reductions.

2. **Early Stopping is Effective**: The positive partial correlation with performance validates that early stopping not only reduces emissions but can actually improve generalization by preventing overfitting.

3. **Model Size is Secondary**: The relatively small contribution of model architecture parameters suggests that emission reduction efforts should focus on fine tuning efficiency rather than always choosing smaller models, which might compromise performance disproportionately.

4. **Warmup Period Tuning**: The complex relationship between warmup epochs and performance across different datasets (as seen in the GraphMAE analysis) underscores the need for dataset-specific hyperparameter optimization.

6 Conclusions

The analysis conducted demonstrates that the *GFT* approach through adaptive Early Stopping permits drastic reductions in CO_2eq emissions (often between 60% and 80%) while maintaining competitive performance.

6.1 Key Findings Across Datasets

- In **classification** tasks (BACE, BBBP, HIV), emission reduction often occurs without penalizing accuracy, sometimes even improving it through overfitting prevention. Notably:
  - BACE showed consistent improvements with early stopping across most transformer models (+3.8% to +7.0% for successful cases)
  - BBBP demonstrated robust performance maintenance with ChemBERTa and ChemBERTa-2 models
  - HIV exhibited strong results particularly with ChemBERTa-2 and SMILES-BERT
- In **regression** tasks (Lipophilicity, CEP, Malaria), a more tangible trade-off exists: stopping training prematurely can cost between 10% and 40% in terms of error (RSE), although on complex datasets like Malaria the effect is mitigated. Key observations:
  - CEP showed exceptional results with SELFormer (+50.6% improvement with early stopping)
  - Lipophilicity demonstrated the most pronounced trade-off, requiring careful balance between efficiency and accuracy
  - Malaria showed mixed results, with warmup configuration playing a critical role

6.2 Explainability Insights

The partial correlation analysis revealed crucial insights for optimizing GFT strategies:

- **Dataset size** emerges as the dominant factor affecting emissions, suggesting that data efficiency techniques should be a primary focus for emission reduction
- **Early stopping** positively correlates with performance while reducing emissions, validating its effectiveness as a GFT strategy
- **Model architecture** has surprisingly limited direct impact on emissions compared to training configuration, indicating that epoch optimization is more critical than model size reduction
- **Warmup epochs** show complex, dataset-dependent relationships with performance, emphasizing the need for adaptive hyperparameter tuning

Based on our comprehensive analysis, the following best practices are recommended:

1. **Implement GFT**: The GFT mechanism consistently provides substantial emission reductions (30-80%) with minimal or even positive performance impacts, especially for classification tasks.
2. **Optimize Dataset Usage**: Given that dataset size is the primary emission driver, invest in data quality over quantity. Techniques such as active learning, data pruning, and curriculum learning should be prioritized.

3. **Task-Specific Warmup Tuning**: For regression tasks, particularly with GNN architectures, carefully tune warmup epochs based on dataset complexity. Start with longer warmup periods (50 epochs) for complex datasets and reduce for simpler ones.

4. **Consider Transformer Models for Classification**: Transformer-based molecular models (ChemBERTa, ChemBERTa-2) consistently show better resilience to early stopping in classification tasks, making them preferred choices when emission reduction is a priority.

5. **Monitor Partial Correlations**: Implement feature monitoring during training to identify which factors are most influencing emissions and performance in real-time, enabling dynamic optimization.

6.3 Broader Impact

In conclusion, this work highlights the significant potential of GFT strategies, particularly adaptive Early Stopping, to reduce the environmental footprint of molecular model fine-tuning without sacrificing performance. The findings advocate for a paradigm shift in model training practices, emphasizing efficiency and sustainability. By demonstrating that GFT is not merely about sacrifice but about intelligent optimization that can maintain or even enhance model quality, this work contributes to making sustainable AI practices more accessible and appealing to practitioners. The explainability framework provided enables researchers to make informed decisions about where to focus their optimization efforts for maximum environmental benefit with minimal performance cost.

References

[1] Giuseppe Spillo, Allegra De Filippo, Vincenzo Monopoli, Cataldo Musto, Michela Milano, and Giovanni Semeraro. *While My RecSys Gently Weeps*: Energy-efficient early stopping with exponential weighted moving average for green recommender systems. *ACM Transactions on Information Systems, Under Review*, 2025.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[3] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[4] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020. Available also as arXiv:1907.10597.

[5] United Nations. Transforming our world: the 2030 agenda for sustainable development, 2015. Resolution adopted by the General Assembly on 25 September 2015.

[6] Giuseppe Spillo, Allegra De Filippo, Emanuele Fontana, Michela Milano, and Giovanni Semeraro. Training green and sustainable recommendation models: Introducing carbon footprint data into early stopping criteria. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '25, page 341346, New York, NY, USA, 2025. Association for Computing Machinery.

[7] Chemberta-zinc-base-v1. `https://huggingface.co/seyonec/ChemBERTa-zinc-base-v1`.

[8] Chemberta-2. `https://huggingface.co/DeepChem/ChemBERTa-77M-MLM`.

[9] Atakan Yüksel, Erva Ulusoy, Atabey Ünlü, Gamze Deniz, and Tunca Doan. Selformer: Molecular representation learning via selfies language models, 2023.

[10] Smiles-bert. `https://huggingface.co/JuIm/SMILES_BERT`.

[11] Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 594–604, 2022.

[12] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

[13] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.