

# Green Fine Tuning for Molecular Property Prediction

Emanuele Fontana

# Outline

1 Business Understanding

2 Data Understanding

3 Data Preparation

4 Modeling

5 Evaluation

6 Conclusions

# Business Understanding

# Context and Motivation

- **Context:** Rise of Deep Learning in Drug Discovery (Transformers, GNNs).
- **Problem:** "Red AI" trend leading to high computational costs and energy consumption.
- **Solution:** **Green AI** paradigm - sustainable AI.
- **Alignment:** UN 2030 Agenda for Sustainable Development.
  - Goal 12: Responsible Consumption and Production.
  - Goal 13: Climate Action.

# Business Objectives

## Primary Goal

Demonstrate that sustainable practices can be integrated into molecular property prediction pipelines without significant loss in performance.

- **Sustainability:** Drastically reduce CO<sub>2</sub>eq emissions during fine-tuning.
- **Reliability:** Maintain predictive accuracy for real-world use.

## Expectations

- CO<sub>2</sub> reduction: 20% - 50%.
- Performance margin: Within 10% of traditional methods.

# Data Understanding

# Datasets (MoleculeNet)

## Classification Tasks:

- **HIV**: Predict if HIV is active or inactive.
- **BACE**: Predict inhibitors of BACE-1 enzyme (Alzheimer's).
- **BBBP**: Predict blood-brain barrier penetration.

## Regression Tasks:

- **Lipophilicity**: Predict octanol/water distribution coefficient.
- **Malaria**: Predict inhibition of malaria parasite growth.
- **CEP**: Predict efficiency of organic photovoltaic molecules.

# Dataset Details

Dataset	Task	Key Components	Rows
BACE	Classification	mol, CID, Class, pIC50	1,513
BBBP	Classification	num, name, p_np, smiles	2,050
HIV	Classification	smiles, activity	41,126
CEP	Regression	smiles, PCE	29,978
Malaria	Regression	smiles, activity	9,999
Lipophilicity	Regression	CMPD_CHEMBLID, exp, smiles	4,200

# Data Representation: SMILES

**SMILES** (Simplified Molecular Input Line Entry System):

- ASCII strings representing chemical structures.
- **Atoms:** C, N, O, etc. (Upper: aliphatic, Lower: aromatic).
- **Bonds:** Single (implicit), Double (=), Triple (#).
- **Branching:** Parentheses ().
- **Rings:** Numbers (e.g., C1CCCCC1).

Crucial for converting 3D structures into 1D sequences for Transformers or Graphs for GNNs.

# Data Preparation

# Preprocessing Pipeline

## ① SMILES Validation:

- **Parsing:** Check syntax errors using RDKit.
- **Sanitization:** Valence checks, Kekulization, Aromaticity detection.

## ② Data Variants:

- **Transformer Variant:** SMILES string + Target label.
- **Graph Variant:** Molecular graphs generated from SMILES.

# Graph Generation (for GNNs)

Conversion of SMILES to Graph  $G = (V, E)$  using RDKit.

**Node Features ( $V$ ):**

- Atomic Number
- Chirality
- Degree
- Formal Charge
- Hybridization
- Aromaticity

**Edge Features ( $E$ ):**

- Bond Type
- Stereochemistry
- Conjugation

Serialized as PyTorch Geometric Data objects (.pt files).

# Modeling

## Sequence-based (Transformers)

- Input: SMILES strings.
- Models:
  - ChemBERTa
  - ChemBERTa-2
  - SELFormer
  - SMILES-BERT

## Graph-based (GNNs)

- Input: Molecular Graphs.
- Model:
  - **GraphMAE**: Masked Autoencoder approach for self-supervised learning.

# Green FineTuning (GFT) Strategy

**Core Idea:** Stop training when marginal performance gain doesn't justify energy cost.

## Instantaneous GFT Ratio

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

where:

- $\Delta P_t$  = improvement in validation metric at epoch  $t$
- $\Delta E_t$  = incremental CO<sub>2</sub>eq emitted during epoch  $t$

## Smoothed Trend (EMA)

To stabilize the volatile instantaneous efficiency:

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

where  $\alpha = 0.9$ . Initialization:  $S_0 = GFT_0$  (or 0).

# GFT Stopping Rule

## Stopping Condition

Training stops when instantaneous efficiency falls below the smoothed trend:

$$\text{Stop if } GFT_t < \beta \cdot S_t$$

where  $\beta = 0.2$  is the tolerance factor.

## Implementation Details

- $Perf_t$  is the validation metric:
  - ROC-AUC for classification tasks
  - Negative RMSE for regression tasks (to maintain "higher is better")
- Warmup period: 3 epochs (Transformers), variable for GraphMAE
- Energy tracking: CodeCarbon library

# Evaluation

# Results: Classification Tasks

- **BACE:**

- Transformers: +1.6% to +7.0% performance, -22% to -60% emissions.
- Early stopping prevents overfitting.

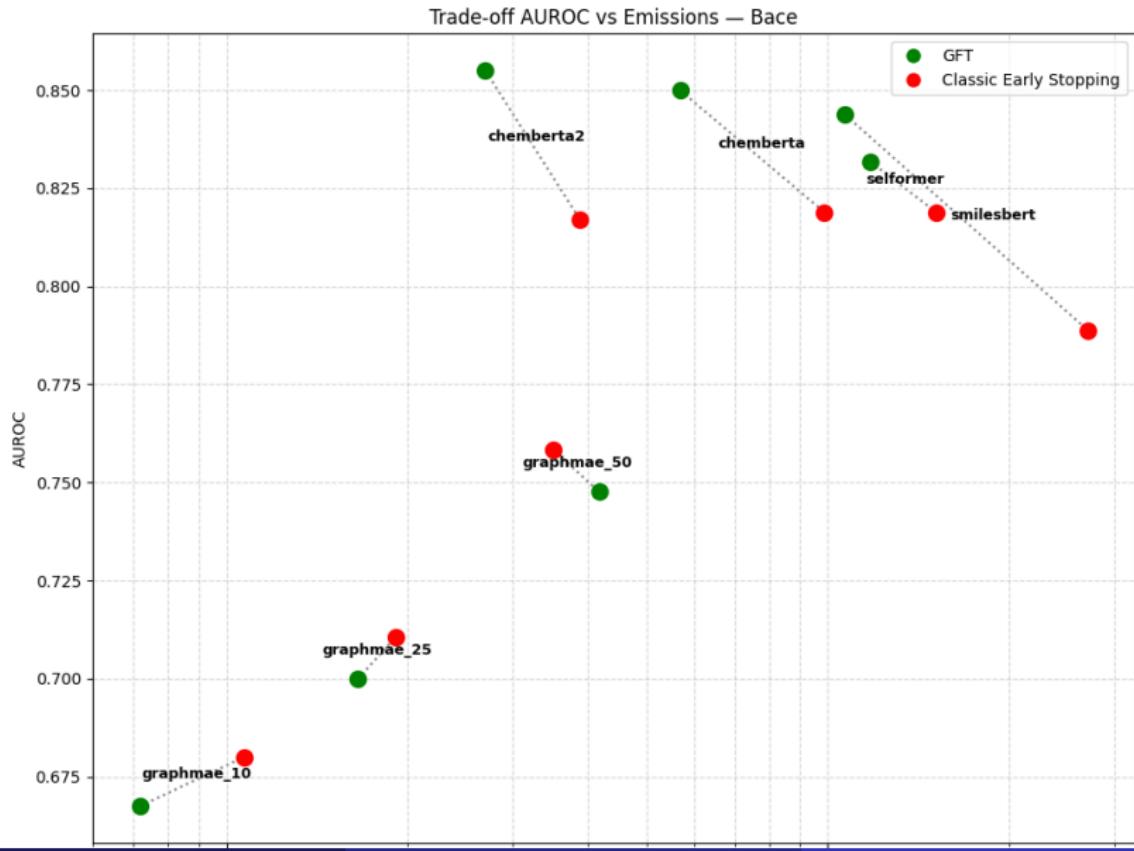
- **HIV:**

- Strong improvements (e.g., SELFormer +16.2% perf, -44% emissions).
- ChemBERTa-2: Stable perf, -60.6% emissions.

- **BBBP:**

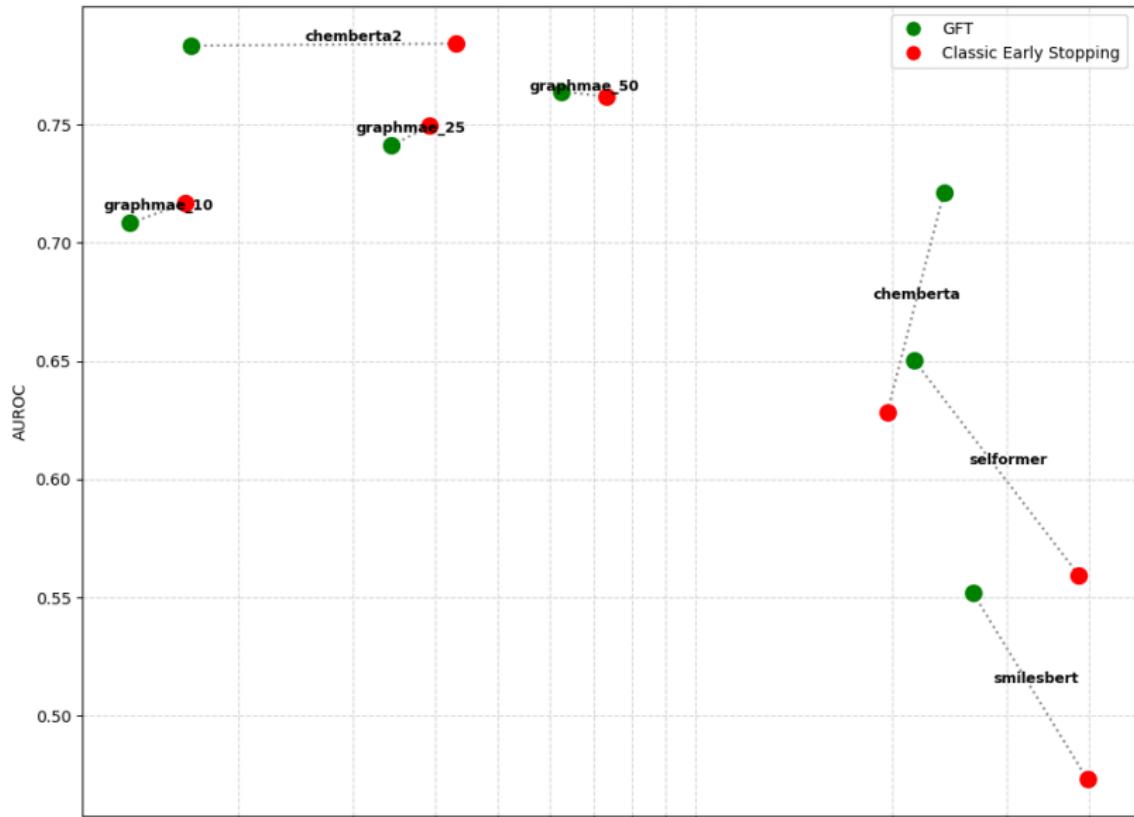
- Mixed results. Some models degrade (ChemBERTa -11.8%).
- SELFormer resilient (+6.9% perf).

# Trade-off: BACE

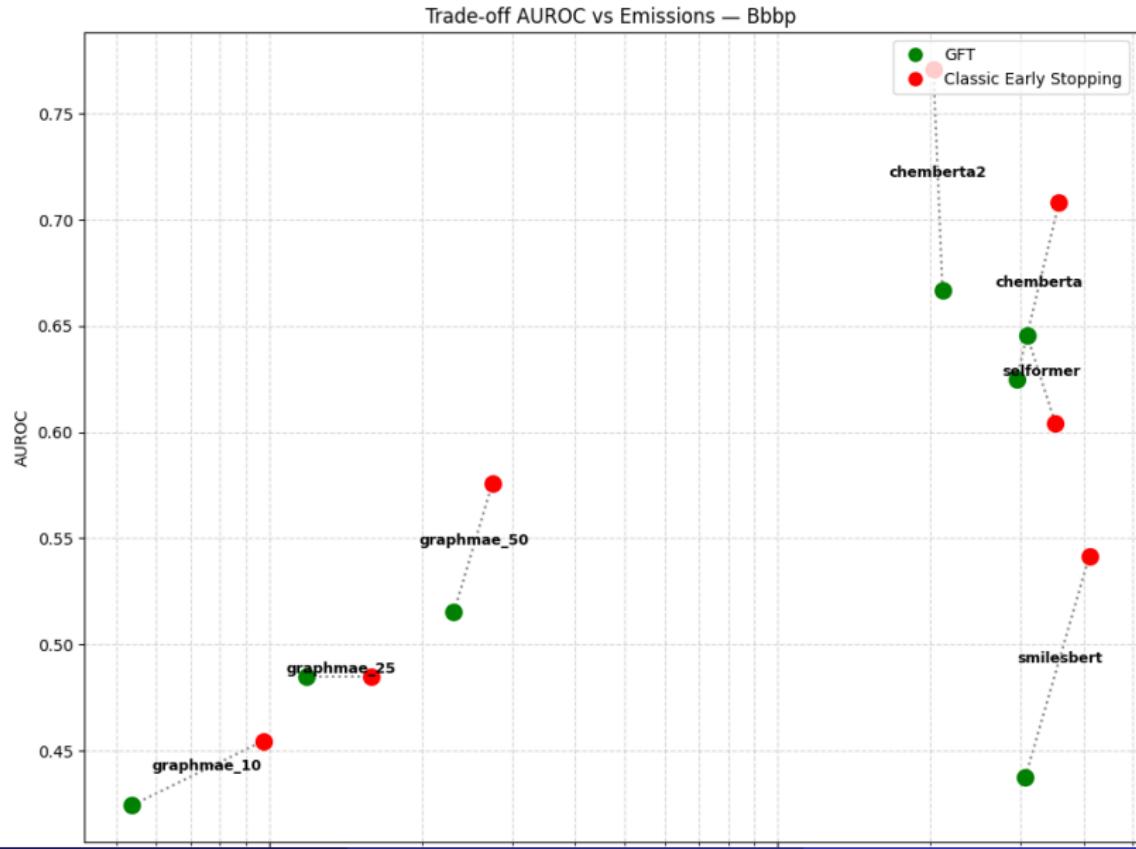


# Trade-off: HIV

Trade-off AUROC vs Emissions — Hiv



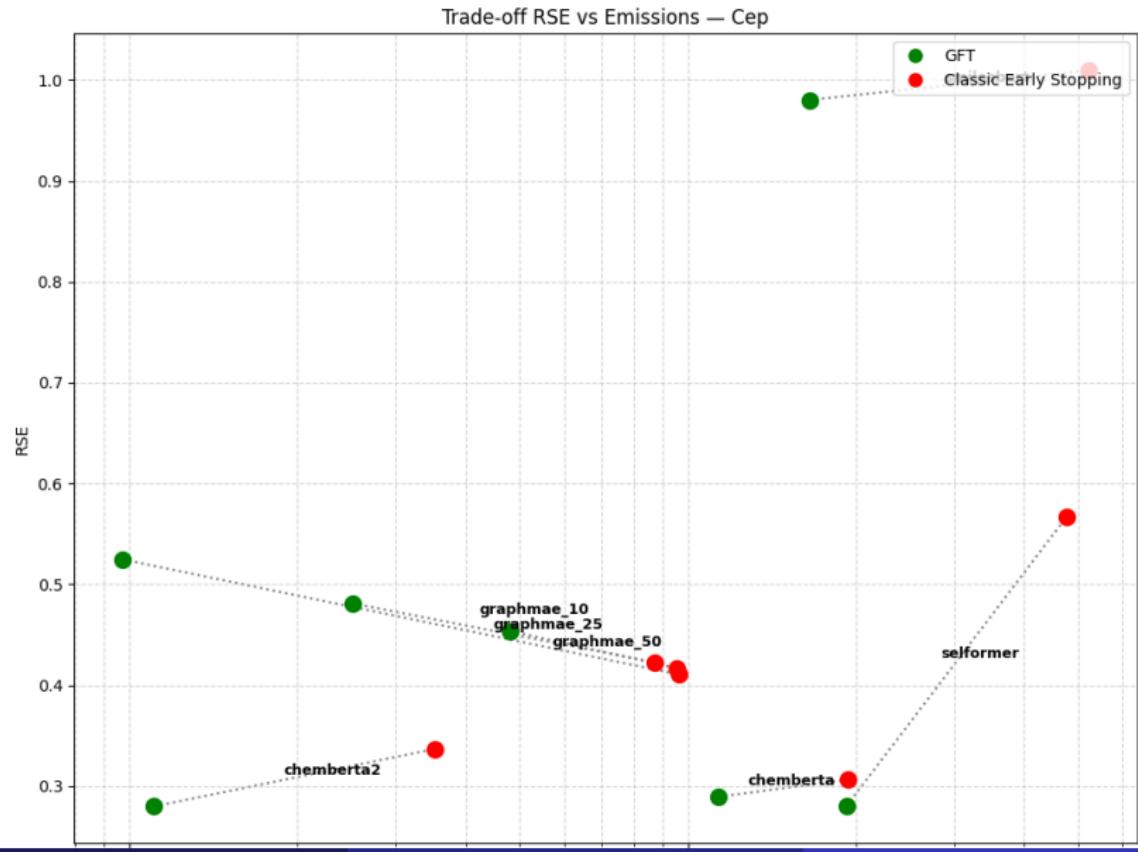
# Trade-off: BBBP



# Results: Regression Tasks

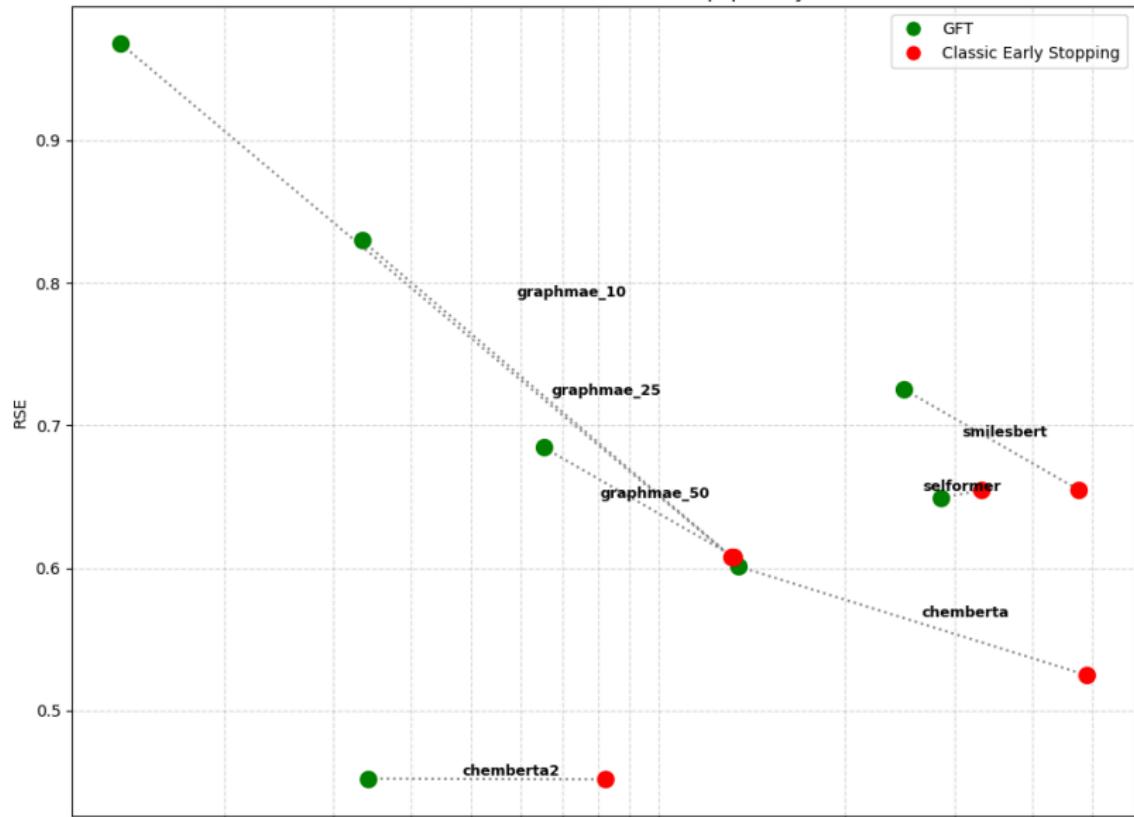
- **CEP:**
  - Excellent results. SELFormer: +50.6% perf, -59.7% emissions.
  - GraphMAE: Trade-off between warmup and performance.
- **Lipophilicity:**
  - Most challenging. Many models degrade.
  - Requires extended fine-tuning.
- **Malaria:**
  - Moderate success. ChemBERTa-2: +3.3% perf, -23.6% emissions.
  - GraphMAE benefits from longer warmup.

# Trade-off: CEP

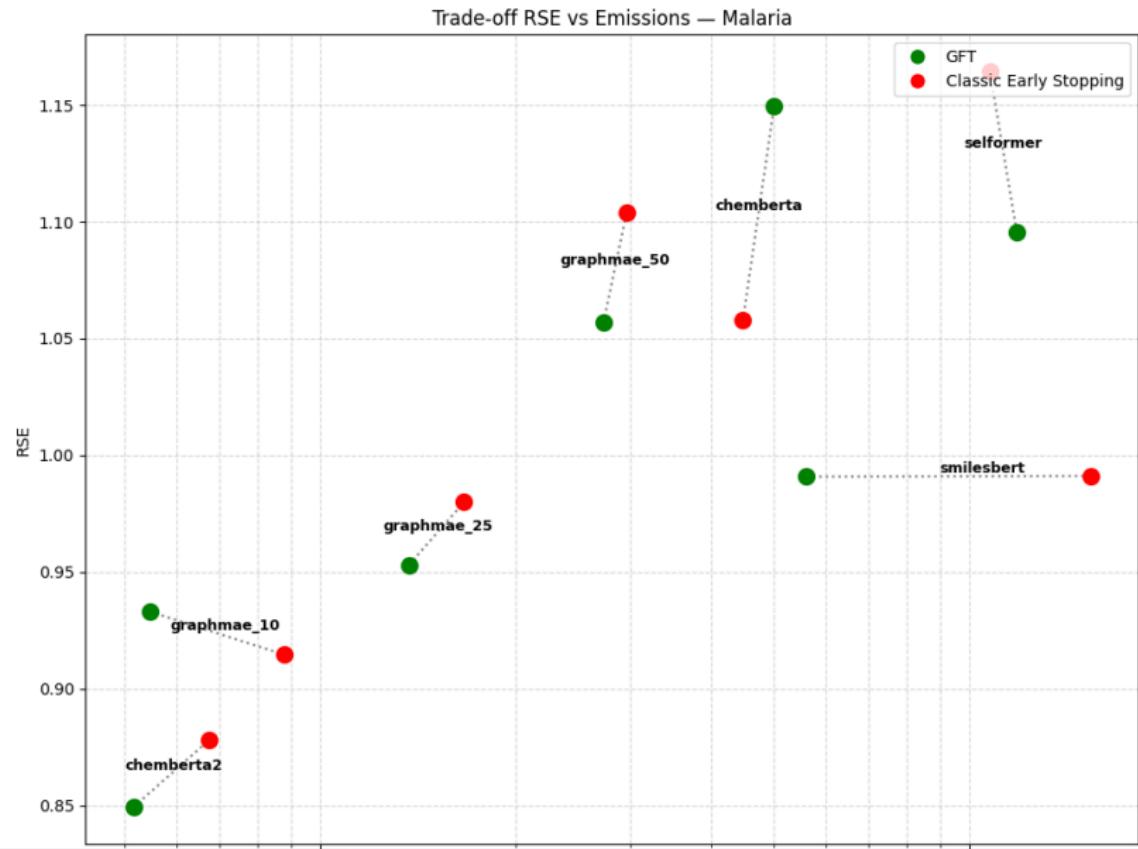


# Trade-off: Lipophilicity

Trade-off RSE vs Emissions — Lipophilicity



# Trade-off: Malaria

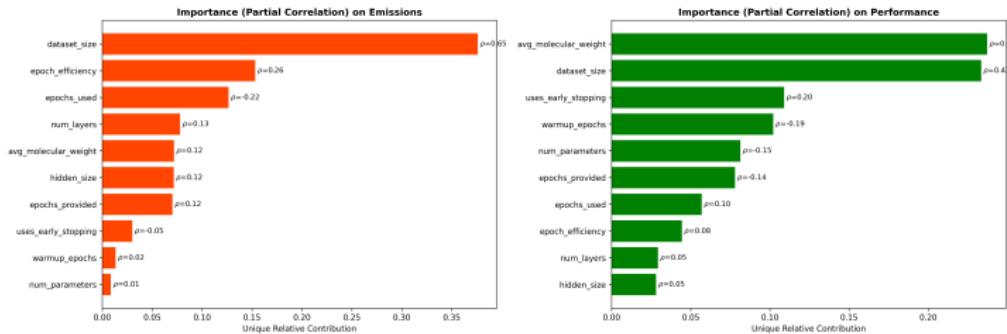


# Explainability Analysis

- **Motivation:** understand which factors drive emissions and performance to improve GFT.
- Use **Partial Correlation** to isolate each feature's unique contribution while controlling for others.
- Valuable when features are collinear across experiments.

# Factors Influencing Emissions

- **Top drivers:** Dataset Size ( $\rho = 0.65$ ), Epoch Efficiency ( $\rho = 0.26$ ).
- **Interactions:** Epochs Used shows interaction effects with dataset size and efficiency.
- **Model architecture:** moderate influence (num\_layers, hidden\_size,  $\rho \approx 0.12$ ).
- **Early stopping:** minimal direct effect ( $\rho = -0.05$ ) its benefit is mediated by fewer epochs used.



# Conclusions

# Key Findings

- **GFT Effectiveness:** Drastic emission reductions (60%-80%) often possible with maintained or improved accuracy.
- **Classification vs Regression:**
  - Classification: Often "win-win" (prevents overfitting).
  - Regression: Clearer trade-off, requires careful tuning.
- **Explainability Insights:**
  - **Dataset Size:** Dominant factor for emissions.
  - **Early Stopping:** Positively correlates with performance.

# Recommendations

- ① **Implement GFT:** Standardize adaptive early stopping.
- ② **Optimize Dataset Usage:** Focus on data quality over quantity.
- ③ **Task-Specific Tuning:** Adjust warmup epochs based on task complexity (especially for Regression/GNNs).
- ④ **Model Selection:** Transformers show high resilience in classification tasks.

**Thank You!**