Green Fine Tuning for Molecular Property Prediction

Emanuele Fontana

Outline

Business Understanding

Context: Green AI vs Red AI

Red AI Issues

- Maximizing accuracy regardless of cost.
- Massive models (Transformers, Large GNNs).
- **High CO_2 emissions & Energy cost.**

Green AI Solution

- Efficiency as a core metric.
- Sustainable fine-tuning cycles.
- **UN Agenda 2030:** Goals 12 & 13.

Business Objectives

Goal: Integrate sustainable practices without losing predictive accuracy.

- **Sustainability:** Reduce CO_2eq emissions by **20% – 50%**.
- **Reliability:** Maintain performance within a **10% margin** of error.
- **Trade-off Analysis:** Identify the trade-off between energy and accuracy.

Data Understanding

Dataset Description (MoleculeNet)

Dataset	Task	Key Features / Targets	Rows
BACE	Class.	`mol`, `Class` (Binary), `pIC50`	1,513
BBBP	Class.	`smiles`, `p_np` (Permeability)	2,050
HIV	Class.	`smiles`, `HIV_active` (0/1)	41,126
CEP	Regr.	`smiles`, `PCE` (Efficiency)	29,978
Malaria	Regr.	`smiles`, `activity`	9,999
Lipophilicity	Regr.	`exp` (LogD), `smiles`	4,200

Data Representation: SMILES

SMILES (Simplified Molecular Input Line Entry System):

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

Data Preparation

Preprocessing Pipeline: RDKit

Pipeline implemented:

1. Parsing & Validation

- `Chem.MolFromSmiles`: Syntax check.
- `Chem.SanitizeMol`: Chemical consistency check.
  - Valence Check, Kekulization, Aromaticity Detection.

2. Output Formats

- **Transformers:** Raw SMILES strings + Labels.
- **GNNs:** Graph Objects $G = (V, E)$ serialized as `.pt` files.

Graph Generation Details

Transformation of SMILES into Graph features for GraphMAE.

Node Features (V)

Captured for each atom:

- Atomic Number
- Chirality
- Degree
- Formal Charge
- Hybridization
- Aromaticity
- Num. Explicit Hs

Edge Features (E)

Captured for each bond:

- Bond Type (Single, Double, Triple, Aromatic)
- Stereochemistry
- Conjugation status

Modeling & Strategy

Architectures

Sequence-based (Transformers):

- **ChemBERTa** & **ChemBERTa-2**: BERT-like models trained on chemical corpora.
- **SELFormer** & **SMILES-BERT**: Specialized architectures.

Graph-based (GNNs):

- **GraphMAE**: Masked Autoencoder.
- Tested with variable warmup epochs (W10, W25, W50) to analyze stability.

Green Fine Tuning (GFT) Algorithm

Objective: Stop when marginal performance gain $<$ marginal energy cost.

Instantaneous Ratio

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

Smoothing & Stopping Condition

We use Exponential Moving Average (EMA) with $\alpha = 0.9$:

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

Stop Rule: If $GFT_t < \beta \cdot S_t$ (with $\beta = 0.2$).

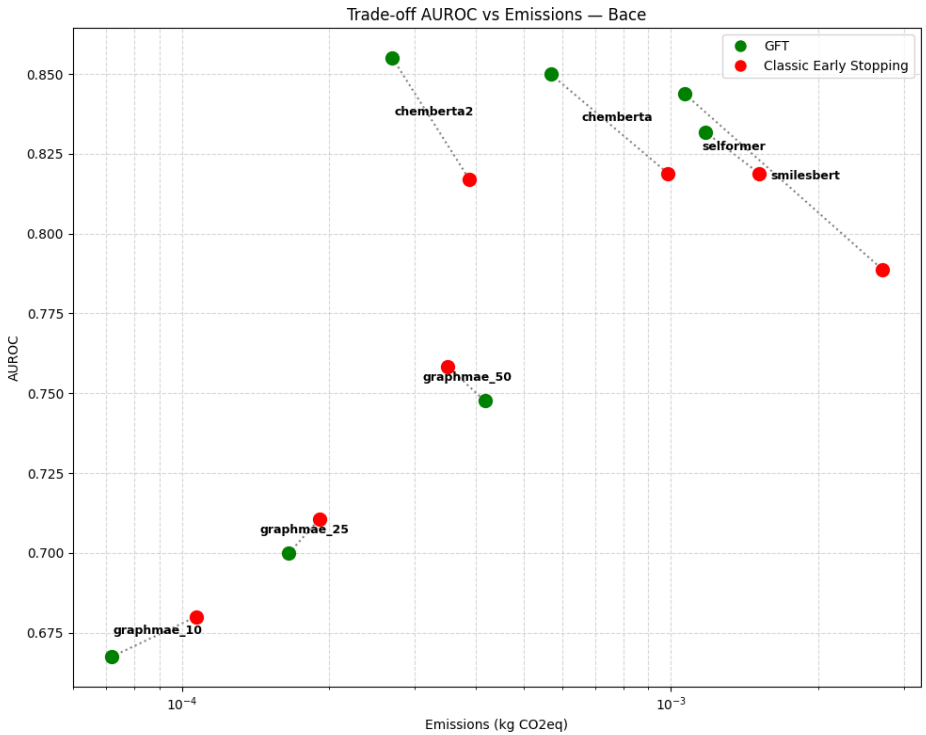
Note: Energy tracked via CodeCarbon (Hardware: RTX 4070, Intel i7).

Results: Trade-off Analysis

BACE Dataset (Classification)

Results:

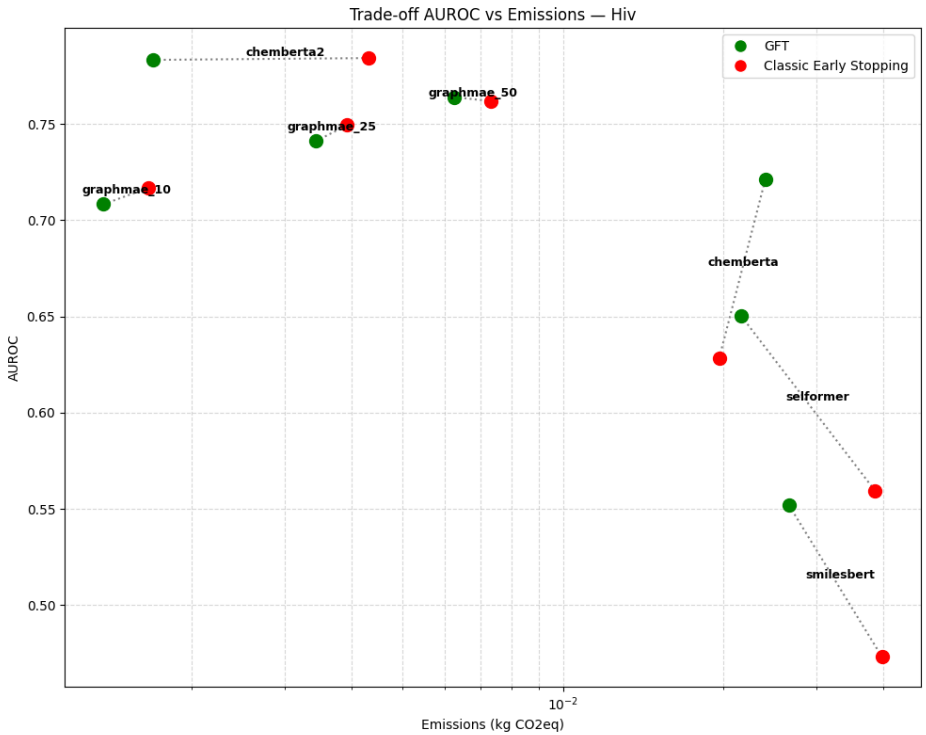
- **SMILES-BERT:** +7.0% Perf, -60.5% Emissions.
- Early stopping effectively prevents overfitting.



HIV Dataset (Classification)

Results:

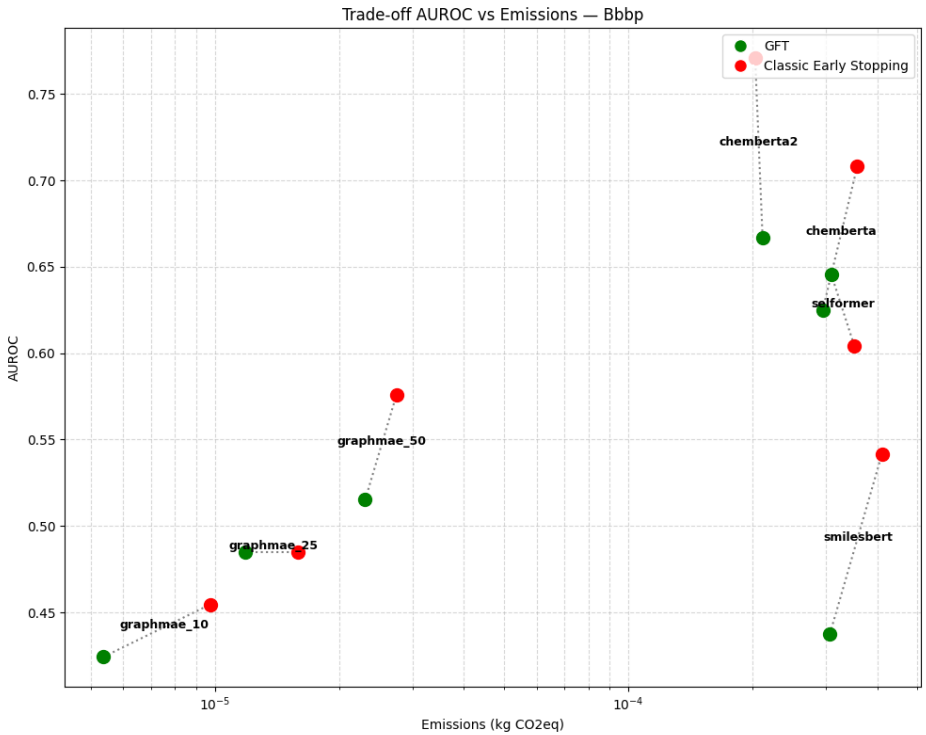
- **SELFormer:** +16.2% Perf, -44.0% Emissions.
- **ChemBERTa-2:** -60.6% Emissions with stable performance.



BBBP Dataset (Classification)

Results:

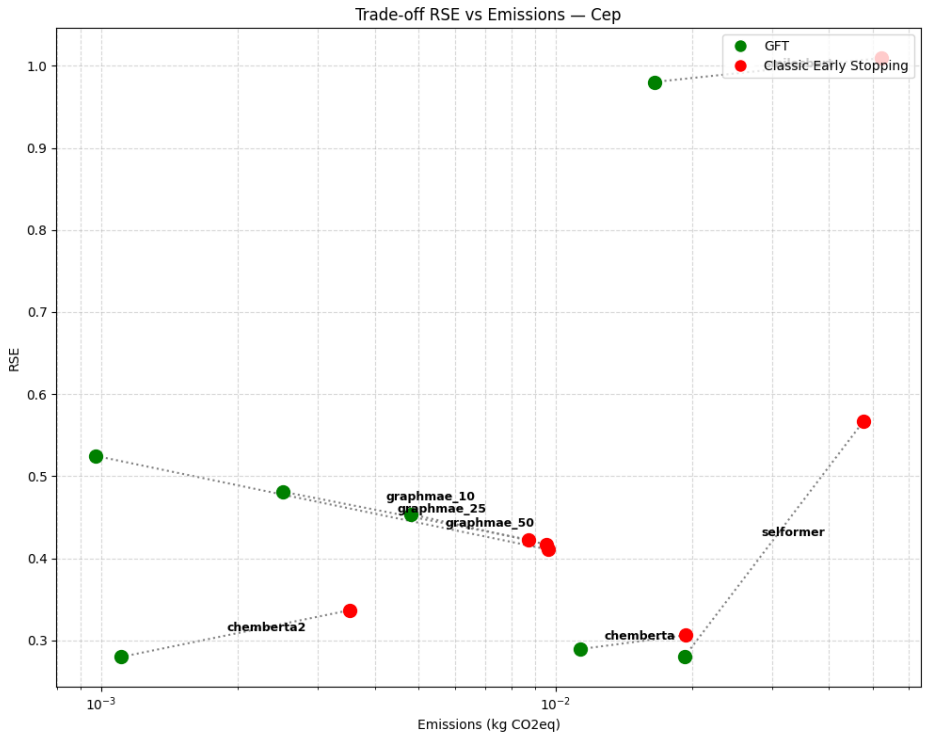
- Mixed results. Some models degrade (e.g., ChemBERTa).
- Highlighted importance of architecture selection.



CEP Dataset (Regression)

Results:

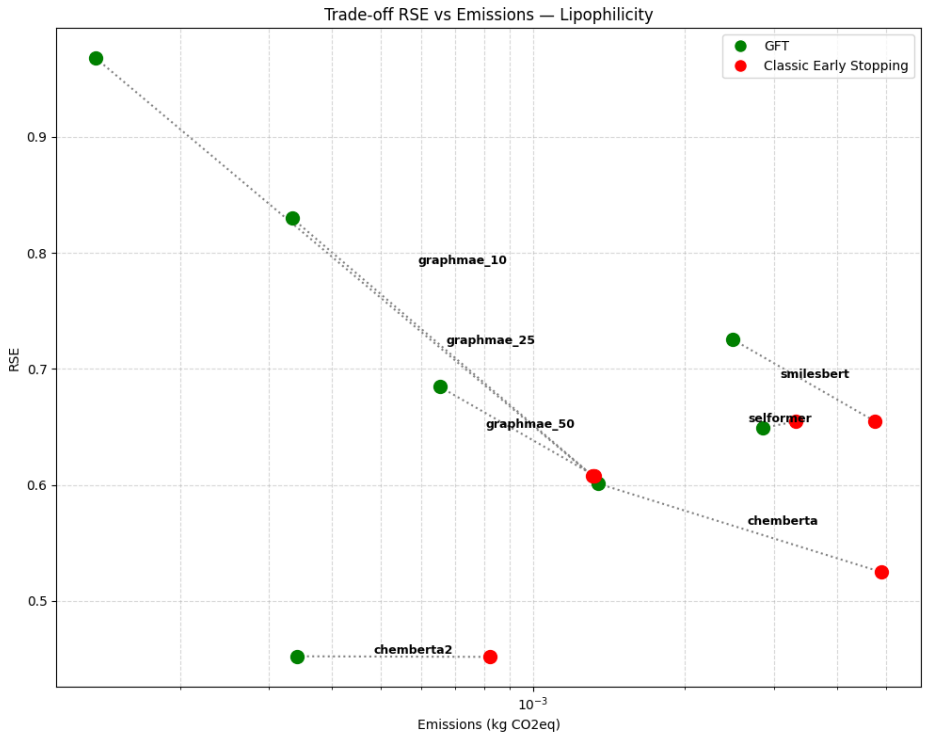
- **SELFormer:** Massive win (+50.6% Perf, -59.7% Emissions).
- GraphMAE shows strong trade-off sensitivity to warmup.



Lipophilicity Dataset (Regression)

Challenge:

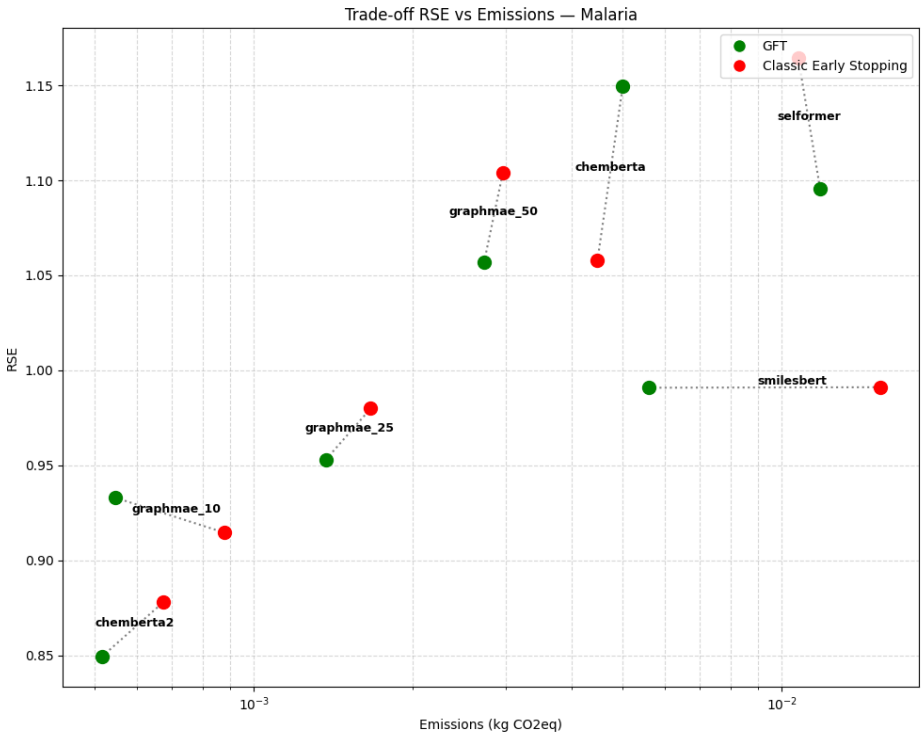
- Most difficult dataset.
- Significant trade-offs: GraphMAE W10 saved 89% energy but lost 59% accuracy.
- Requires longer training.



Malaria Dataset (Regression)

Results:

- **ChemBERTa-2:** +3.3% Perf, -23.6% Emissions.
- GraphMAE benefits significantly from longer warmup (W50).

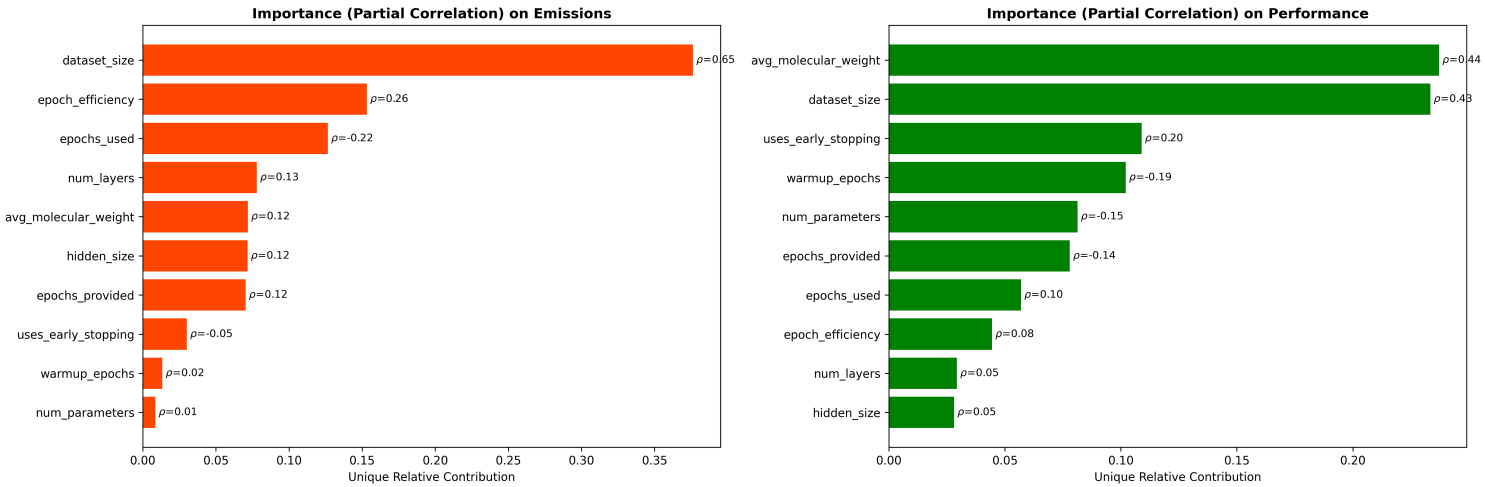


Explainability & Conclusions

What drives Emissions?

Partial Correlation Analysis (ρ):

1. **Dataset Size ($\rho = 0.65$):** The dominant factor.
2. **Epoch Efficiency ($\rho = 0.26$):** Ratio of epochs used.
3. **Model Architecture:** Secondary impact compared to data size.



Final Recommendations

1. Adopt GFT

Achieved **60-80% emission reduction** in best cases (Classification) often with performance gains.

2. Data-Centric Approach

Since Dataset Size is the main emission driver, prioritize **Data Pruning** and Quality over Model Size.

3. Adaptive Strategies

- **Classification:** Transformers are highly robust to Early Stopping.
- **Regression/GNNs:** Require careful Warmup tuning (start with 50 epochs for complex tasks).

Thank You!