Analysis of Green AI Experiments on Molecular Transformers

Technical Report - CRISP-DM Methodology

Abstract

This report presents a detailed analysis of the application of Green AI techniques in the context of molecular property prediction. Following the CRISP-DM methodology, we explore the use of an adaptive Early Stopping mechanism (AER) on Transformer-based models and Graph Neural Networks. The objective is to evaluate the trade-off between predictive accuracy and energy consumption, demonstrating how significant reductions in CO_2 emissions can be achieved with minimal impact on performance.

Contents

1 Business Understanding

1.1 Context and Motivation

Training Deep Learning models, particularly in the field of computational chemistry and drug discovery, requires substantial computational resources. With the growing complexity of models (e.g., Transformers), the carbon footprint associated with their lifecycle has become a critical concern. The **Green AI** paradigm aims to make artificial intelligence more sustainable by optimizing energy efficiency without excessively compromising result quality.

1.2 Project Objectives

The main objective of this study is to evaluate the effectiveness of a custom early stopping mechanism, called **AER (Adaptive Accuracy-Emission Ratio)**, applied to molecular classification and regression tasks. Specific objectives include:

- Quantifying the reduction in CO_2eq emissions achievable through AER compared to classic training strategies.
- Analyzing the impact of AER on performance metrics (ROC-AUC for classification, RSE for regression).
- Identifying the optimal equilibrium point between accuracy and sustainability for different architectures (Transformers vs GNN).

2 Data Understanding

2.1 Dataset Description

For the experiments, standard datasets from the **MoleculeNet** suite were used, representative of various chemical-physical and biological properties. The data consists of SMILES (Simplified Molecular Input Line Entry System) strings and their corresponding target labels.

The analyzed datasets include:

- **BACE (Classification)**: Dataset containing beta-secretase 1 (BACE-1) inhibitors, a crucial target for Alzheimer's disease. The task is to predict whether a molecule inhibits the enzyme (binary).
- **BBBP (Classification)**: Blood-Brain Barrier Penetration dataset. Predicts whether a compound can penetrate the blood-brain barrier, essential for CNS drug development.
- **HIV (Classification)**: Dataset for predicting HIV replication inhibition, containing compounds tested for their ability to inhibit HIV viral replication.
- **CEP (Regression)**: Clean Energy Project. Contains data on the energy conversion efficiency of molecules for organic photovoltaics.
- **Lipophilicity (Regression)**: Measures the lipophilicity (LogD7.4) of small molecules, a fundamental property for drug absorption and distribution in the human body.
- **Malaria (Regression)**: Dataset focused on the efficacy of compounds against the malaria parasite.

3 Data Preparation

3.1 Preprocessing

Management and manipulation of chemical structures were performed using the **RDKit** library. SMILES strings were canonicalized and converted into graphical or sequential representations suitable for the models used.

3.2 Scaffold Splitting

A critical phase of data preparation is the division into training, validation, and test sets. To ensure a realistic evaluation of the model's generalization capability, **Scaffold Splitting** (80/10/10) was adopted instead of random splitting.

This technique groups molecules based on their Murcko scaffold (the central cyclic structure). Molecules with the same scaffold are assigned to the same set, simulating the real-world scenario of discovering new classes of chemical compounds structurally distinct from known ones.

3.3 Robustness Check

A control algorithm (*ensure_min_two_classes*) was implemented for classification tasks. In case the scaffold splitting produces a validation or test set with only one class (making ROC-AUC calculation impossible), the algorithm automatically rebalances the sets by taking specific examples from the training set.

4 Modeling

4.1 Architectures Used

Two model families were compared:

- **Transformers**: Models pre-trained on large corpora of SMILES strings, including **ChemBERTa**, **ChemBERTa-2**, **SELFormer**, and **SMILES-BERT**.
- **Graph Neural Networks (GNN)**: The **GraphMAE** (Graph Masked Autoencoder) model, which operates directly on the graph representation of the molecule.

4.2 Training Strategy: AER

To reduce the carbon footprint, a custom callback based on the adaptive ratio between accuracy and emissions was implemented. Training is interrupted when the marginal gain in performance no longer justifies the marginal energy cost:

$$AER_t = \frac{\%\Delta\text{Performance}_t}{\%\Delta\text{Emissions}_t} \tag{1}$$

Training stops if $AER_t < \beta \cdot \text{EMA}(AER_{t-1})$, where β is a tolerance threshold and EMA is the exponential moving average. This strategy ("Green-Early") was compared with a classic Early Stopping approach ("Classic") based on patience (patience=5).

4.3 Experimental Configuration

All experiments were conducted on a single NVIDIA RTX 4070 Mobile GPU, monitoring emissions through the `CodeCarbon` library.

- **Transformers**: 30 epochs, Batch Size 32, LR 1×10^{-4}, Warmup 5 epochs.

- **GraphMAE**: 100 epochs, Batch Size 32, LR 1×10^{-4}, Variable warmup (10, 25, 50 epochs).

5 Evaluation

This section presents the results obtained by comparing the "Classic" and "Green-Early" configurations.

5.1 BACE Dataset (Classification)

Metrics: ROC-AUC (higher is better).

Model	AUC (Classic)	AUC (Early)	Δ Perf.	CO_2eq (kg) C.	CO_2eq (kg) E.	Δ Emiss.
ChemBERTa	0.819	0.850	+3.8%	9.87×10^{-4}	5.68×10^{-4}	**-42.4%**
ChemBERTa-2	0.817	0.855	+4.6%	3.87×10^{-4}	2.70×10^{-4}	**-30.3%**
SELFormer	0.819	0.832	+1.6%	1.52×10^{-3}	1.18×10^{-3}	**-22.4%**
SMILES-BERT	0.789	0.844	+7.0%	2.71×10^{-3}	1.07×10^{-3}	**-60.5%**
GraphMAE (W10)	0.680	0.667	-1.9%	1.07×10^{-4}	7.18×10^{-5}	**-33.0%**
GraphMAE (W25)	0.711	0.700	-1.5%	1.91×10^{-4}	1.65×10^{-4}	**-13.6%**
GraphMAE (W50)	0.758	0.748	-1.3%	3.50×10^{-4}	4.17×10^{-4}	*+19.1%*

Table 1: Comparison on BACE dataset. In some cases (ChemBERTa, SMILES-BERT), the Early version performed better, suggesting that prolonged training led to overfitting.

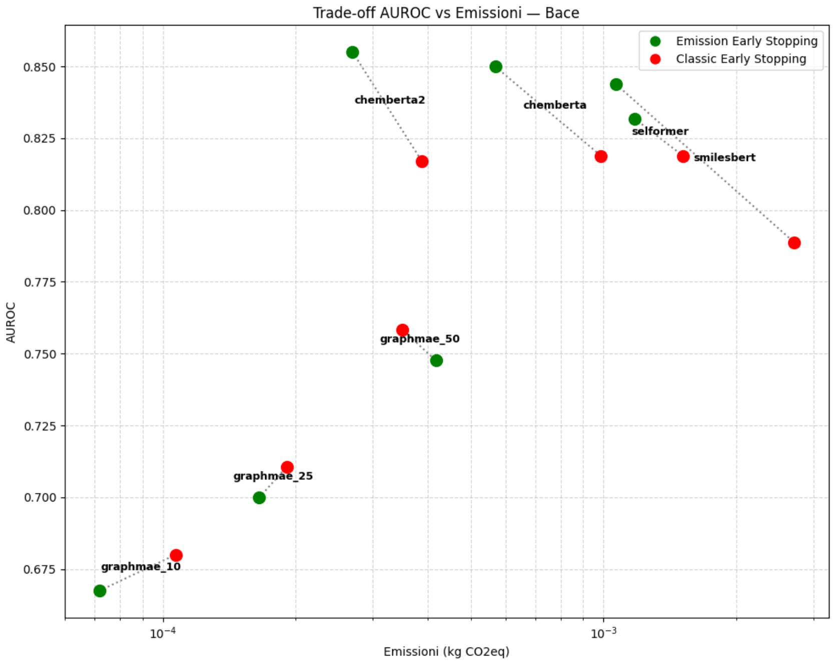

Figure 1: ROC curve for ChemBERTa on BACE: comparison between Classic and Early Stopping versions

5.2 Dataset HIV (Classification)

Metrics: ROC-AUC (higher is better).

Model	AUC (Classic)	AUC (Early)	∆ Perf.	CO_2eq (kg) C.	CO_2eq (kg) E.	∆ Emiss.
ChemBERTa	0.628	0.721	+14.8%	1.97×10^{-2}	2.40×10^{-2}	*+21.8%*
ChemBERTa-2	0.784	0.783	-0.1%	4.31×10^{-3}	1.70×10^{-3}	**-60.6%**
SELFormer	0.559	0.650	+16.3%	3.86×10^{-2}	2.16×10^{-2}	**-44.0%**
SMILES-BERT	0.473	0.552	+16.7%	3.98×10^{-2}	2.66×10^{-2}	**-33.2%**
GraphMAE (W10)	0.717	0.708	-1.3%	1.66×10^{-3}	1.37×10^{-3}	**-17.5%**
GraphMAE (W25)	0.749	0.741	-1.1%	3.93×10^{-3}	3.43×10^{-3}	**-12.7%**
GraphMAE (W50)	0.762	0.764	+0.3%	7.33×10^{-3}	6.24×10^{-3}	**-14.9%**

Table 2: Comparison on HIV dataset. Transformers show positive performance gains with early stopping, while GNNs maintain stable performance with energy savings.

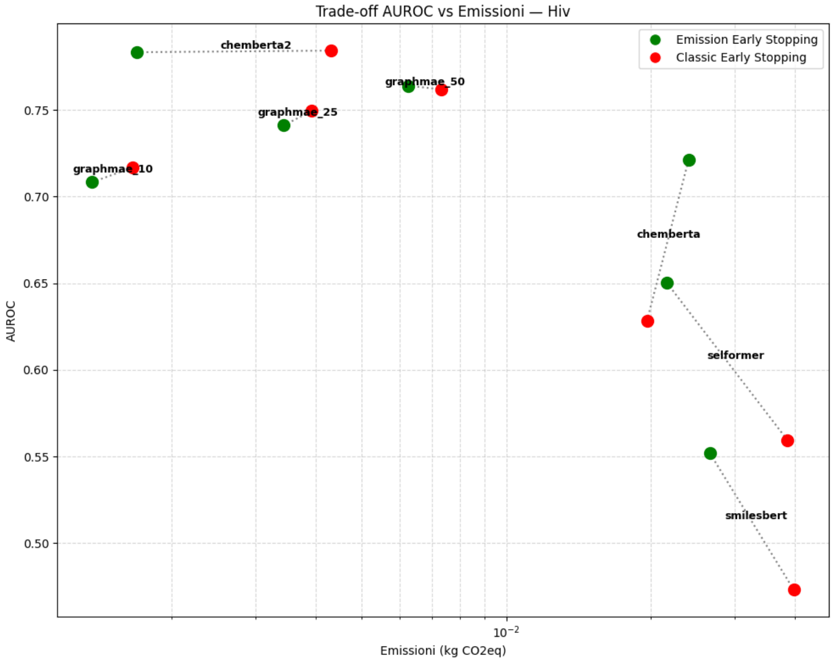

Figure 2: ROC curve for ChemBERTa on HIV: comparison between Classic and Early Stopping

5.3 Dataset BBBP (Classification)

Metrics: ROC-AUC (higher is better).

Model	AUC (Classic)	AUC (Early)	∆ Perf.	CO_2eq (kg) C.	CO_2eq (kg) E.	∆ Emiss.
ChemBERTa	0.708	0.625	-11.7%	3.58×10^{-4}	2.95×10^{-4}	**-17.6%**
ChemBERTa-2	0.771	0.667	-13.5%	2.03×10^{-4}	2.11×10^{-4}	*+3.9%*
SELFormer	0.604	0.646	+6.9%	3.52×10^{-4}	3.09×10^{-4}	**-12.2%**
SMILES-BERT	0.542	0.438	-19.2%	4.10×10^{-4}	3.07×10^{-4}	**-25.1%**
GraphMAE (W10)	0.455	0.424	-6.8%	9.72×10^{-6}	5.37×10^{-6}	**-44.8%**
GraphMAE (W25)	0.485	0.485	*Unchanged*	1.59×10^{-5}	1.18×10^{-5}	**-25.8%**
GraphMAE (W50)	0.576	0.515	-10.6%	2.75×10^{-5}	2.31×10^{-5}	**-16.0%**

Table 3: Comparison on BBBP dataset. BBBP shows variable behavior: some models maintain stability while others show trade-offs between performance and emissions.

5.4 CEP Dataset (Regression)

Metrics: RSE (Relative Squared Error - lower is better).

Model	RSE (Classic)	RSE (Early)	Δ Perf.	CO_2eq (kg) C.	CO_2eq (kg) E.	Δ Emiss.
ChemBERTa	0.306	0.289	+5.5%	1.93×10^{-2}	1.13×10^{-2}	**-41.2%**
ChemBERTa-2	0.337	0.280	+16.9%	3.52×10^{-3}	1.11×10^{-3}	**-68.5%**
SELFormer	0.567	0.280	+50.6%	4.76×10^{-2}	1.92×10^{-2}	**-59.7%**
SMILES-BERT	1.009	0.980	+2.9%	5.22×10^{-2}	1.65×10^{-2}	**-68.4%**
GraphMAE (W10)	0.411	0.525	-27.7%	9.64×10^{-3}	9.72×10^{-4}	**-89.9%**
GraphMAE (W25)	0.423	0.481	-13.7%	8.73×10^{-3}	2.51×10^{-3}	**-71.2%**
GraphMAE (W50)	0.417	0.454	-8.9%	9.53×10^{-3}	4.81×10^{-3}	**-49.5%**

Table 4: Comparison on CEP dataset.

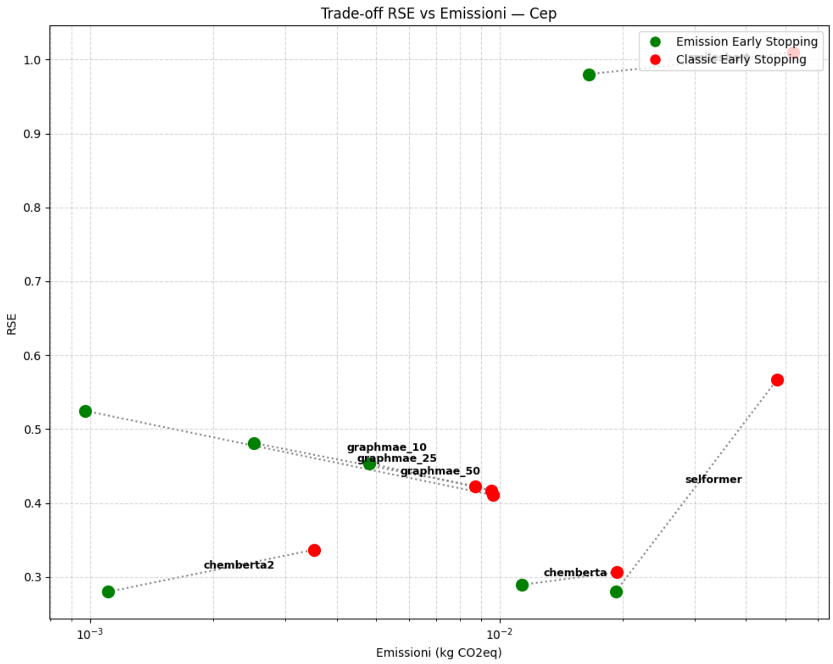

Figure 3: RSE trend on CEP for ChemBERTa: comparison between Classic and Early Stopping versions

5.5 Lipophilicity Dataset (Regression)

Metrics: RSE (lower is better).

Model	RSE (Classic)	RSE (Early)	Δ Perf.	CO_2eq (kg) C.	CO_2eq (kg) E.	Δ Emiss.
ChemBERTa	0.525	0.601	-14.5%	4.90×10^{-3}	1.34×10^{-3}	**-72.7%**
ChemBERTa-2	0.452	0.452	*Unchanged*	8.20×10^{-4}	3.40×10^{-4}	**-58.5%**
SELFormer	0.655	0.649	+0.9%	3.30×10^{-3}	2.85×10^{-3}	**-13.6%**
SMILES-BERT	0.655	0.725	-10.7%	4.75×10^{-3}	2.49×10^{-3}	**-47.6%**
GraphMAE (W10)	0.608	0.968	-59.2%	1.31×10^{-3}	1.36×10^{-4}	**-89.6%**
GraphMAE (W25)	0.608	0.830	-36.5%	1.31×10^{-3}	3.33×10^{-4}	**-74.6%**
GraphMAE (W50)	0.608	0.685	-12.7%	1.32×10^{-3}	6.53×10^{-4}	**-50.5%**

Table 5: Comparison on Lipophilicity. Here the most pronounced trade-off is observed: large energy savings correspond to a significant loss in model precision.

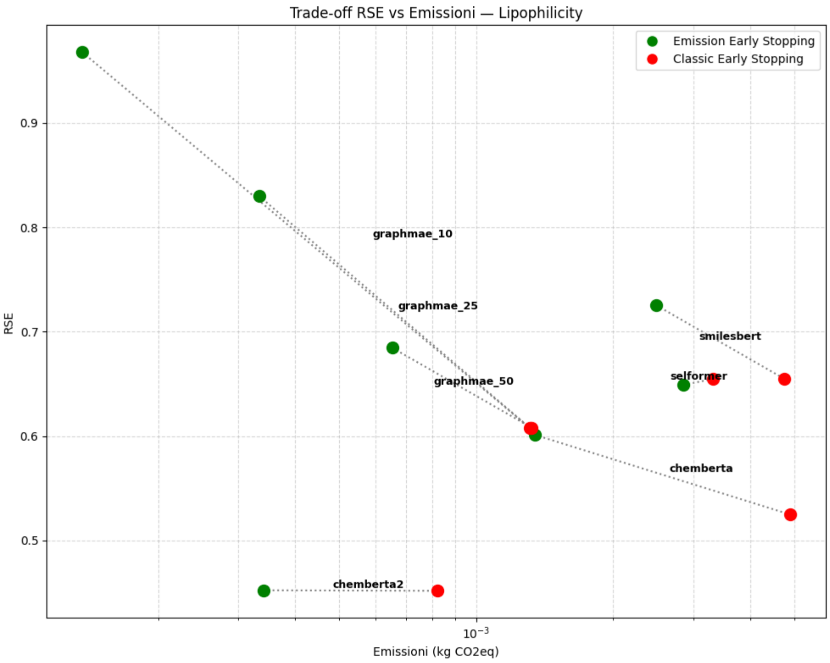

Figure 4: RSE trend on Lipophilicity for ChemBERTa: comparison between Classic and Early Stopping versions

5.6 Malaria Dataset (Regression)

Metrics: RSE (lower is better).

Model	**RSE (Classic)**	**RSE (Early)**	**△ Perf.**	**CO_2eq (kg) C.**	**CO_2eq (kg) E.**	**△ Emiss.**
ChemBERTa	1.058	1.149	-8.6%	4.47×10^{-3}	5.00×10^{-3}	*+11.9%*
ChemBERTa-2	0.878	0.849	+3.3%	6.74×10^{-4}	5.15×10^{-4}	**-23.6%**
SELFormer	1.164	1.096	+5.8%	1.07×10^{-2}	1.18×10^{-2}	*+10.0%*
SMILES-BERT	0.991	0.991	*Unchanged*	1.53×10^{-2}	5.61×10^{-3}	**-63.4%**
GraphMAE (W10)	0.915	0.933	-2.0%	8.80×10^{-4}	5.45×10^{-4}	**-38.1%**
GraphMAE (W25)	0.980	0.953	+2.8%	1.66×10^{-3}	1.37×10^{-3}	**-17.4%**
GraphMAE (W50)	1.104	1.057	+4.3%	2.96×10^{-3}	2.73×10^{-3}	**-7.8%**

Table 6: Comparison on Malaria dataset.

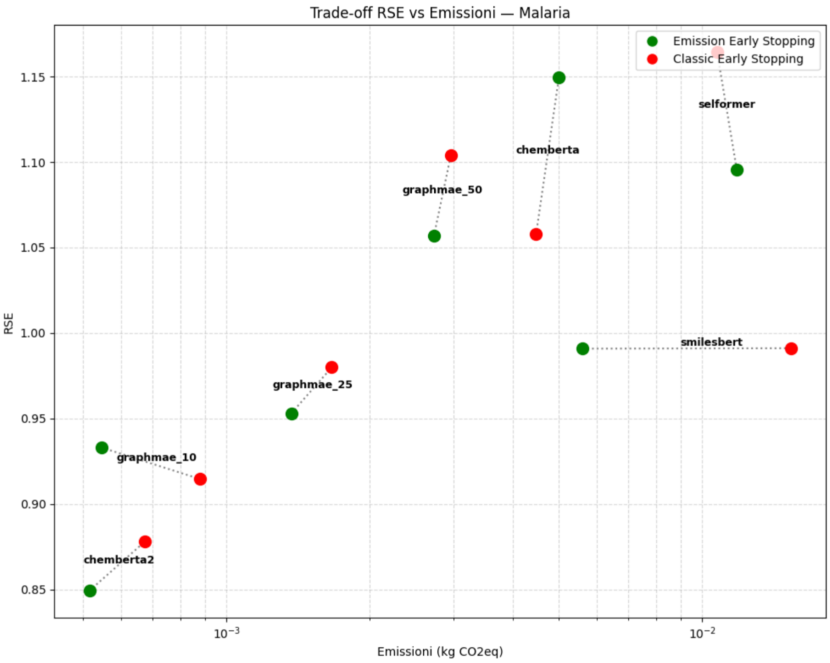

Figure 5: RSE trend on Malaria for ChemBERTa: comparison between Classic and Early Stopping versions

5.7 Warmup Sensitivity Analysis (GraphMAE)

For the GraphMAE model, a specific analysis was conducted by varying the number of warmup epochs (10, 25, 50) to evaluate the impact on Early Stopping stability.

- **Complex Datasets (CEP, Lipophilicity)**: A longer warmup (50 epochs) proved essential. With only 10 warmup epochs, the model tended to stop too early, resulting in much higher RSE (e.g., Lipophilicity: RSE 0.96 with W10 vs 0.68 with W50).
- **Malaria Dataset**: Conversely, prolonged warmup worsened performance (RSE 1.05 with W50 vs 0.93 with W10), suggesting that for this dataset the model quickly reaches peak performance and further epochs lead to overfitting.
- **Trade-off**: Using a 50-epoch warmup ensures more robust performance closer to fine-tuning with classic early stopping, while maintaining significant energy savings (about 50%).

5.8 Explainability Analysis

5.8.1 Motivation and Methodology

Understanding which factors most influence model emissions and performance is crucial for optimizing Green AI strategies. To address this, we conducted a comprehensive explainability analysis using **Partial Correlation**, a statistical technique that isolates the unique contribution of each feature while controlling for all other variables.

Unlike simple correlation, partial correlation removes confounding effects, revealing the true independent relationship between each feature and the target outcomes (emissions and performance). This approach is particularly valuable when dealing with multicollinearity among features, which is common in machine learning experiments where multiple factors interact.

5.8.2 Feature Engineering

For each experiment, we extracted and engineered a comprehensive set of features:

- **Model Architecture Features**: Number of parameters, hidden size, number of layers, model family (Transformer vs GNN)
- **Training 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)
- **Training Strategy**: Binary indicator for early stopping usage

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

5.8.3 Key Findings

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

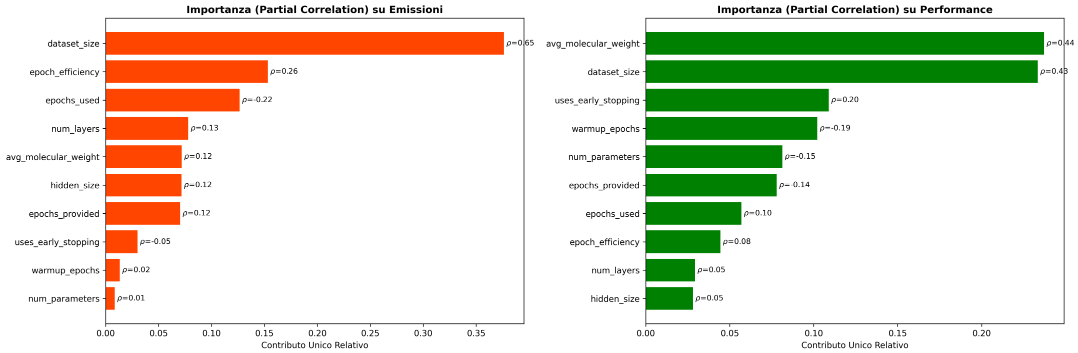

Figure 6: Feature importance for emissions (left) and performance (right) based on partial correlation analysis. The values shown are normalized relative contributions, with ρ indicating the direction and strength of the partial correlation.

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 training.
2. **Epoch Efficiency** ($\rho = 0.26$): Positive correlation indicates that longer training (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 training 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 Green AI approach.

4. **Warmup Epochs** ($\rho = -0.19$): Negative correlation indicates that excessive warmup may delay convergence or lead to suboptimal training 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.

5.8.4 Added Variable Plots (Partial Regression Plots)

To visualize these relationships while controlling for confounding variables, we created added variable plots showing the residual relationships:

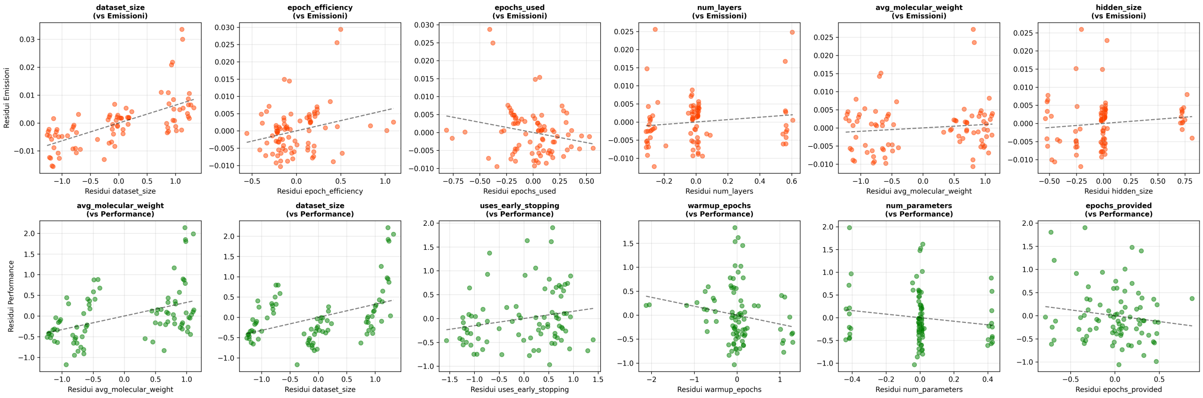

Figure 7: Added variable plots showing residual relationships between top features and target variables after controlling for all other features. Top row shows relationships with emissions; bottom row shows relationships with performance. The trend lines indicate the partial correlation direction.

These plots reveal the *unique* contribution of each feature after removing the influence of all other variables. The scatter pattern confirms:

- Strong linear residual relationships for dataset_size with both emissions and performance
- Moderate positive relationship between early stopping and performance (bottom row, middle panel)
- Complex interactions between epoch efficiency and emissions when other factors are controlled

5.8.5 Implications for Green AI 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 training 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 Deployment and Conclusions

The analysis conducted demonstrates that the *Green AI* 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 Green AI 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 Green AI 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

6.3 Practical Recommendations

Based on our comprehensive analysis, we recommend the following Green AI practices:

1. **Implement Adaptive Early Stopping**: The AER 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.

6.4 Broader Impact

In conclusion, adopting metrics like AER and implementing explainability-driven optimization represents a promising strategy for making artificial intelligence more sustainable. This is especially relevant in research and development contexts where training numerous models is frequent, such as:

- Drug discovery campaigns requiring screening of thousands of molecular candidates
- Hyperparameter optimization and neural architecture search
- Continuous model retraining in production environments
- Educational and experimental settings where compute budgets are limited

By demonstrating that Green AI 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.

Future work should explore extending these techniques to larger language models, investigating the interaction between early stopping and other efficiency techniques (quantization, pruning), and developing automated systems for per-dataset adaptive training strategies based on real-time emission and performance monitoring.