Analisi degli Esperimenti di Green AI su Molecular Transformers

Report Tecnico

6 dicembre 2025

1 Metodologia e Gestione dei Dati

1.1 Preprocessing e Scaffold Splitting

La gestione dei dataset chimici è stata effettuata utilizzando la libreria RDKit. Una fase critica del processo è la suddivisione dei dati in set di training, validazione e test. Invece di una suddivisione casuale, è stato implementato lo **Scaffold Splitting** (80/10/10).

Questa tecnica raggruppa le molecole in base al loro scafold di Murcko (la struttura centrale ciclica). Le molecole con lo stesso scaffold vengono assegnate allo stesso set. Ciò garantisce una valutazione più realistica della capacità di generalizzazione del modello, simulando la scoperta di nuove strutture chimiche distinte da quelle note.

L'algoritmo implementato include inoltre un controllo di robustezza (*ensure_min_two_classes*): nel caso in cui la suddivisione per scaffold produca un set di validazione o test con una sola classe (rendendo impossibile il calcolo della ROC-AUC), l'algoritmo ribilancia i set prelevando esempi specifici dal training set.

1.2 Meccanismo di Early Stopping (AER)

Per ridurre l'impronta di carbonio, è stato implementato un callback personalizzato basato sul rapporto adattivo tra accuratezza ed emissioni (AER - Adaptive Accuracy-Emission Ratio). Il training viene interrotto quando il guadagno marginale in performance non giustifica più il costo energetico marginale:

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

Il training si arresta se $AER_t < \beta \cdot \text{EMA}(AER_{t-1})$, dove β è una soglia di tolleranza e EMA è la media mobile esponenziale.

2 Risultati Sperimentali

Di seguito vengono presentati i risultati per i quattro dataset analizzati. I modelli sono stati testati in configurazione "Classic" (fine-tuning completo) e "Early" (con arresto anticipato).

2.1 Dataset BACE (Classificazione)

Metriche utilizzate: ROC-AUC (maggiore è meglio).

Modello	AUC (Classic)	AUC (Early)	Δ Perf.	CO_2eq (kg) C.	CO_2eq (kg) E.	Δ Emiss.
ChemBERTa	0.961	0.993	+3.4%	1.03×10^{-3}	3.70×10^{-4}	**-63.9%**
ChemBERTa-2	0.967	0.974	+0.7%	3.06×10^{-4}	2.50×10^{-4}	**-18.3%**
SELFormer	0.947	0.993	+4.9%	1.85×10^{-3}	5.62×10^{-4}	**-69.6%**
SMILES-BERT	0.974	0.928	-4.7%	1.87×10^{-3}	8.72×10^{-4}	**-53.5%**

Tabella 1: Confronto su dataset BACE. In alcuni casi (ChemBERTa, SELFormer), la versione Early ha performato meglio, suggerendo che il training prolungato portava ad overfitting.

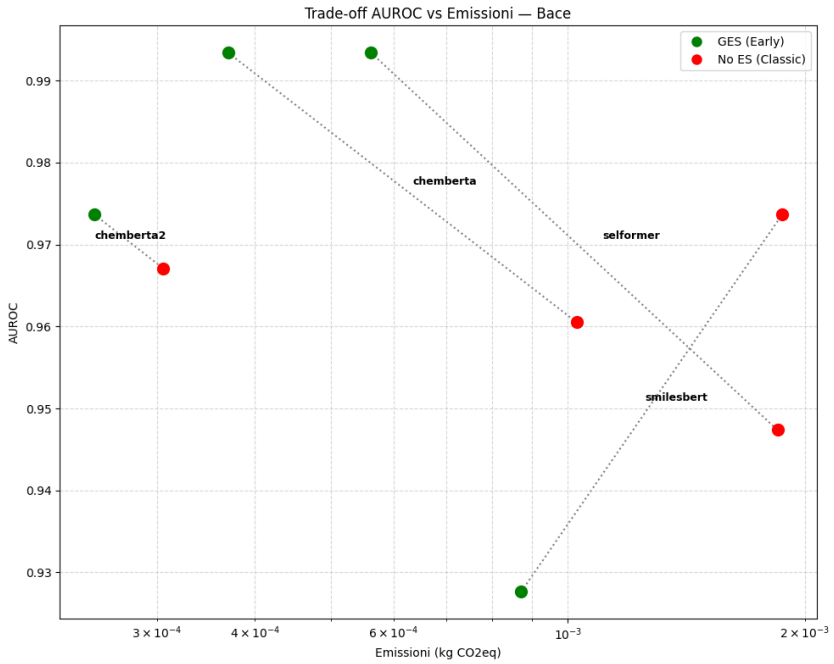

Figura 1: Curva ROC per ChemBERTa su BACE: confronto tra versione Classic ed Early Stopping

2.2 Dataset CEP (Regressione)

Metriche utilizzate: RSE (Relative Squared Error - minore è meglio).

Modello	RSE (Classic)	RSE (Early)	Δ Perf.	CO_2eq (kg) C.	CO_2eq (kg) E.	Δ Emiss.
ChemBERTa	0.347	0.311	+10.3%	1.60×10^{-2}	3.37×10^{-3}	**-78.9%**
ChemBERTa-2	0.279	0.345	-23.6%	2.07×10^{-3}	6.08×10^{-4}	**-70.6%**
SELFormer	0.328	0.393	-19.8%	3.16×10^{-2}	6.51×10^{-3}	**-79.4%**
SMILES-BERT	1.000	1.000	*Invar.*	3.25×10^{-2}	9.90×10^{-3}	**-69.5%**

Tabella 2: Confronto su dataset CEP. Nota: Per RSE, un valore più basso indica una performance migliore. Una Δ Perf negativa indica un peggioramento dell'errore (aumento del RSE).

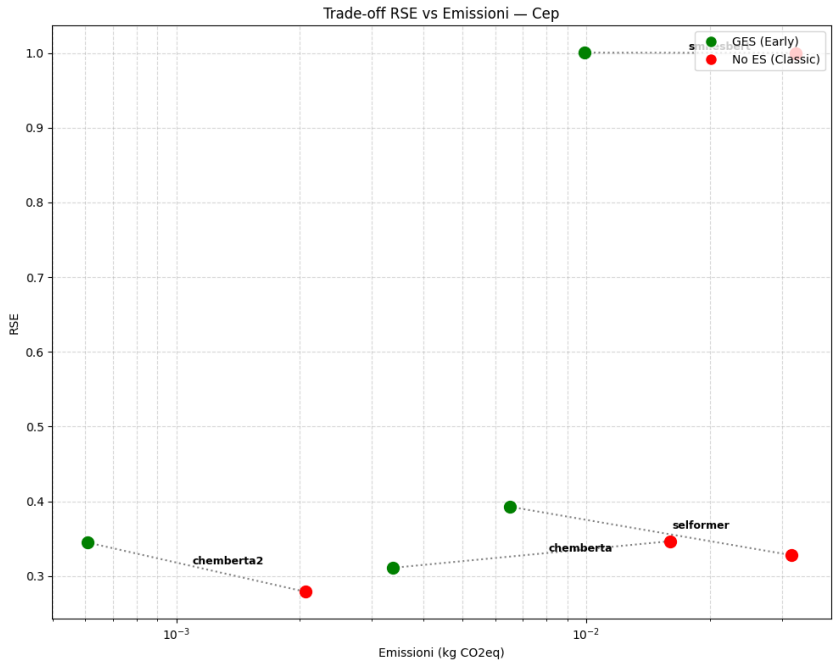

Figura 2: Andamento del RSE su CEP per ChemBERTa: confronto tra versione Classic ed Early Stopping

2.3 Dataset Lipophilicity (Regressione)

Metriche utilizzate: RSE (minore è meglio).

Modello	**RSE (Classic)**	**RSE (Early)**	**Δ Perf.**	**CO_2eq (kg) C.**	**CO_2eq (kg) E.**	**Δ Emiss.**
ChemBERTa	0.513	0.725	-41.3%	2.43×10^{-3}	6.48×10^{-4}	**-73.3%**
ChemBERTa-2	0.547	0.699	-27.8%	4.81×10^{-4}	2.86×10^{-4}	**-40.5%**
SELFormer	0.642	0.695	-8.3%	4.62×10^{-3}	1.11×10^{-3}	**-76.0%**
SMILES-BERT	0.692	0.752	-8.7%	4.73×10^{-3}	1.57×10^{-3}	**-66.7%**

Tabella 3: Confronto su Lipophilicity. Qui si nota il trade-off più marcato: grandi risparmi energetici corrispondono a una perdita significativa di precisione nel modello.

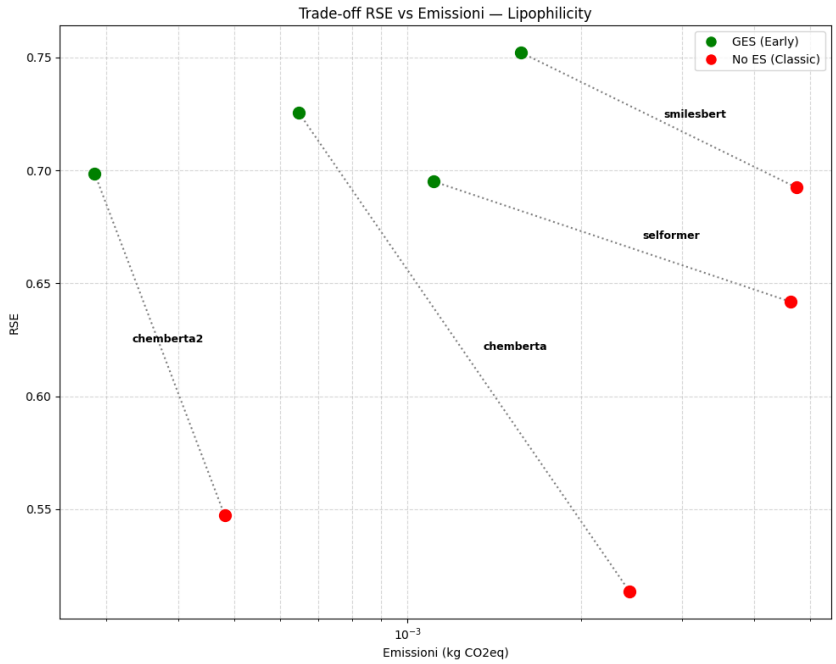

Figura 3: Andamento del RSE su Lipophilicity per ChemBERTa: confronto tra versione Classic ed Early Stopping

2.4 Dataset Malaria (Regressione)

Metriche utilizzate: RSE (minore è meglio).

Modello	**RSE (Classic)**	**RSE (Early)**	**Δ Perf.**	**CO_2eq (kg) C.**	**CO_2eq (kg) E.**	**Δ Emiss.**
ChemBERTa	0.926	0.931	-0.5%	5.49×10^{-3}	1.29×10^{-3}	**-76.4%**
ChemBERTa-2	0.995	0.870	+12.6%	8.40×10^{-4}	3.42×10^{-4}	**-59.3%**
SELFormer	1.068	0.991	+7.2%	1.07×10^{-2}	2.31×10^{-3}	**-78.3%**
SMILES-BERT	1.018	0.994	+2.4%	1.10×10^{-2}	3.45×10^{-3}	**-68.6%**

Tabella 4: Confronto su Malaria. In questo dataset, l'Early Stopping ha spesso migliorato o mantenuto invariate le performance (RSE più basso o simile), suggerendo che i modelli tendono a saturare o overfittare rapidamente su questi dati.

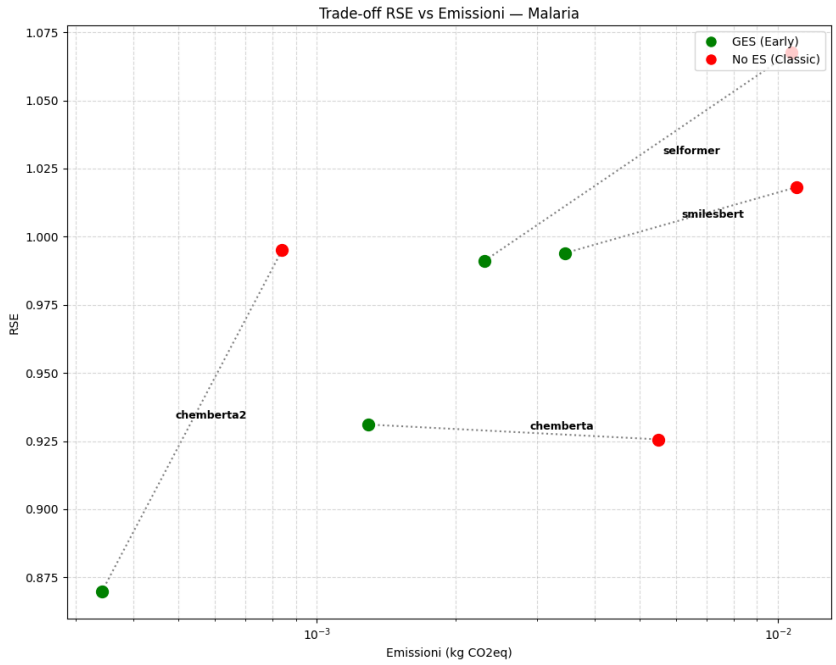

Figura 4: Andamento del RSE su Malaria per ChemBERTa: confronto tra versione Classic ed Early Stopping

3 Conclusioni

L'analisi dimostra che l'approccio *Green AI* tramite Early Stopping adattivo permette di ridurre drasticamente le emissioni di CO_2eq (spesso tra il 60% e l'80%).

- Nei task di **classificazione** (BACE), la riduzione delle emissioni avviene senza penalizzare l'accuratezza, anzi talvolta migliorandola grazie alla prevenzione dell'overfitting.
- Nei task di **regressione** (Lipophilicity, CEP), esiste un trade-off più tangibile: interrompere il training prematuramente può costare tra il 10% e il 40% in termini di errore (RSE), sebbene su dataset complessi come Malaria l'effetto sia mitigato.