Delta Machine Learning for Excited-State Property Correction Using Coulomb Matrix Descriptors

Daisi Williams
Under the academic guidance of Dr. Ezekiel Oyeniyi

Abstract

An accurate prediction of excited-state properties is essential in computational chemistry and materials discovery. High-level quantum chemical methods such as Coupled Cluster (CC2) provide reliable excitation energies but are computationally expensive, whereas semi-empirical methods like Density Functional Tight Binding (DFTB) are computationally efficient but less accurate. In this study, I used a delta-learning (Δ-ML) approach to systematically correct DFTB predictions toward CC2 accuracy using supervised machine learning models trained on Coulomb matrix molecular descriptors.

The correction function is formulated as learning the residual:

$$\Delta E = E^{\mathrm{CC2}} - E^{\mathrm{DFTB}},$$

which is then used to construct a corrected estimate

$$\widehat{E}^{\mathrm{CC2}} = E^{\mathrm{DFTB}} + \widehat{\Delta E}.$$

Neural networks and kernel-based methods were implemented and evaluated on a dataset of approximately 21,000 organic molecules. Significant improvement was observed for excitation energies, while oscillator strength corrections exhibited limited gains, consistent with the increased electronic complexity of transition intensities. This project demonstrates the effectiveness of residual learning for systematic physical model correction and highlights limitations of purely structural descriptors for electronic transition properties.

1 Problem Statement and Motivation

Excited-state properties, such as first excitation energy E_1 and oscillator strength f_1, play a central role in understanding optical absorption, photochemical processes, and materials design. However, the computational cost of high-accuracy quantum methods scales steeply with system size. Coupled Cluster theory at the CC2 level offers reliable excitation energies but is not practical for large-scale screening.

In contrast, DFTB provides computational efficiency but introduces systematic approximation errors. Rather than directly predicting CC2 quantities from molecular structure, this work leverages the hypothesis that systematic model errors are smoother and more learnable than absolute quantum observables.

Formally, given a molecular descriptor vector $\mathbf{x}$ derived from the Coulomb matrix representation, we define the residual learning objective:

$$\widehat{\Delta E_1} = f_\theta(\mathbf{x}),$$

where f_θ denotes a parameterized machine learning model (e.g., neural network, kernel ridge regression, or support vector regression). The corrected excitation energy is then given by

$$\widehat{E_1^{\mathrm{CC2}}} = E_1^{\mathrm{DFTB}} + \widehat{\Delta E_1}.$$

The model parameters θ are optimized by minimizing the empirical loss

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left| \Delta E_{1,i} - \widehat{\Delta E_{1,i}} \right|.$$

This formulation reflects a core principle of scientific machine learning: leveraging known physical approximations while learning only the systematic discrepancy between levels of theory.

2 Dataset and Feature Engineering

2.1 Dataset Description

The dataset consists of approximately 21,000 small organic molecules with excited-state properties computed at two different levels of quantum theory: Density Functional Tight Binding (DFTB) and Coupled Cluster at the CC2 level. For each molecule, the following quantities were available:

- First excitation energy: E_1^{DFTB}, E_1^{CC2}
- Second excitation energy: E_2^{DFTB}, E_2^{CC2}
- Oscillator strengths: f_1^{DFTB}, f_1^{CC2}, f_2^{DFTB}, f_2^{CC2}

The primary focus of this study was the correction of E_1 and f_1. Data preprocessing included removal of numerical outliers and verification of physically meaningful constraints (e.g., enforcing $f \geq 0$). The final dataset was split into training, validation, and test subsets using a 70% / 15% / 15% ratio, ensuring no data leakage between splits.

2.2 Coulomb Matrix Representation

Each molecule was encoded using a Coulomb matrix descriptor. For a molecule with nuclear charges $\{Z_i\}$ and interatomic distances $\{R_{ij}\}$, the Coulomb matrix $\mathbf{C}$ is defined as:

$$C_{ij} = \begin{cases} \frac{1}{2} Z_i^{2.4}, & \text{if } i = j, \\ \frac{Z_i Z_j}{R_{ij}}, & \text{if } i \neq j. \end{cases}$$

This representation provides a rotationally and translationally invariant encoding of molecular geometry and composition. To obtain a fixed-length descriptor across molecules of varying size, a reduced feature representation was constructed by selecting and ordering a subset of matrix elements, resulting in a 26-dimensional feature vector for each molecule.

The Coulomb matrix captures nuclear charge interactions and interatomic distances, thereby encoding structural information relevant to electronic properties. However, it does not explicitly contain excited-state wavefunction information, which has implications for learning transition intensities, as discussed later.

2.3 Feature Scaling and Preprocessing

Prior to model training, features were standardized using a training-set-based normalization:

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j},$$

where μ_j and σ_j denote the mean and standard deviation of feature j computed exclusively on the training subset. This ensures that each feature dimension has zero mean and unit variance, which is particularly important for kernel methods and neural network optimization stability.

Target variables were constructed as either:

$$\Delta E_1 = E_1^{\mathrm{CC2}} - E_1^{\mathrm{DFTB}},$$

or

$$\Delta f_1 = f_1^{\mathrm{CC2}} - f_1^{\mathrm{DFTB}},$$

depending on the learning objective. In certain experiments involving oscillator strengths, a logarithmic transformation of the form $\log(1+f)$ was applied to mitigate heavy-tailed distributions and stabilize training.

This preprocessing pipeline ensures numerical stability, prevents information leakage, and aligns with best practices in scientific machine learning.

3 Models and Training Protocol

3.1 Learning Framework

Given molecular descriptors $\mathbf{x}_i \in \mathbb{R}^{26}$ and targets y_i (either ΔE_1 or Δf_1), the supervised learning task is to approximate a function

$$f_\theta : \mathbb{R}^{26} \to \mathbb{R}$$

parameterized by θ, such that

$$y_i \approx f_\theta(\mathbf{x}_i).$$

Model parameters were optimized by minimizing an empirical risk objective over the training dataset:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell\left(y_i, f_\theta(\mathbf{x}_i)\right),$$

where ℓ denotes a regression loss function. In practice, Mean Absolute Error (MAE) was used for evaluation due to its robustness to outliers and interpretability in physical units.

3.2 Kernel Ridge Regression

Kernel Ridge Regression (KRR) provides a non-parametric approach to regression by combining ridge regularization with the kernel trick. The optimization problem is given by:

$$\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{K}\boldsymbol{\alpha}\|_2^2 + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha},$$

where $\mathbf{K}$ is the kernel matrix defined by

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j),$$

and $\lambda > 0$ is a regularization parameter controlling model complexity.

Both radial basis function (RBF) and Laplacian kernels were explored:

$$k_{\text{RBF}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|_2^2\right),$$

$$k_{\text{Laplacian}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|_1\right).$$

Hyperparameters λ and γ were selected via cross-validation.

3.3 Support Vector Regression

Support Vector Regression (SVR) was employed to investigate margin-based regression performance. SVR minimizes an ϵ-insensitive loss function:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{N} \xi_i,$$

subject to

$$|y_i - (\mathbf{w}^\top \phi(\mathbf{x}_i) + b)| \leq \epsilon + \xi_i,$$

where $\phi(\cdot)$ denotes a kernel-induced feature mapping, C controls regularization strength, and ϵ defines the width of the insensitive zone. RBF kernels were primarily used.

3.4 Neural Network Architecture

A multilayer perceptron (MLP) was implemented to learn nonlinear mappings between Coulomb matrix features and target corrections. The architecture consisted of fully connected layers with ReLU activations:

$$\mathbf{h}^{(l+1)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)}\right),$$

where $\sigma(\cdot)$ denotes the ReLU activation function.

The final output layer produced a scalar regression output:

$$\hat{y} = \mathbf{W}^{(L)}\mathbf{h}^{(L-1)} + b^{(L)}.$$

Dropout regularization and early stopping were used to mitigate overfitting. Optimization was performed using the Adam optimizer with adaptive learning rates.

3.5 Evaluation Metrics

Model performance was assessed using:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Absolute error percentiles (P50, P90, P95, P99)
- Improvement factor relative to baseline

The improvement factor was defined as:

$$\text{Improvement} = \frac{\text{MAE}_{\text{baseline}}}{\text{MAE}_{\text{corrected}}}.$$

All reported results correspond to the held-out test set to ensure unbiased generalization estimates.

4 Results and Interpretation

4.1 Excitation Energy Prediction (E_1)

The Δ-learning approach substantially improved excitation energy prediction relative to the DFTB baseline. On the held-out test set, the Mean Absolute Error (MAE) decreased from 0.07435 Hartree (baseline) to 0.04173 Hartree after correction, corresponding to an improvement factor of approximately 1.78×.

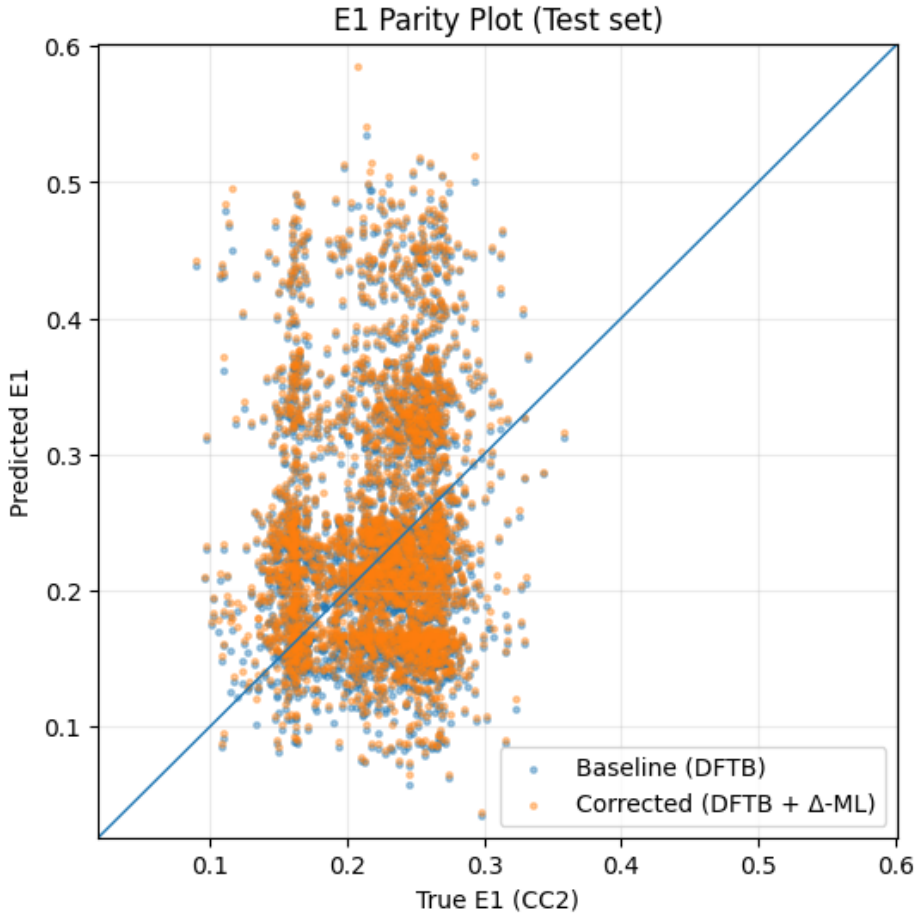

Figure 1: Graph 1: Parity plot for E_1 (baseline and corrected) against CC2 reference values.

Interpretation (Graph 1 – Parity Plot): The baseline predictions exhibit systematic deviation from the identity line, reflecting structured approximation error in DFTB. After Δ-learning

Table 1: Test-set performance for first excitation energy E_1 (DFTB $\rightarrow$ CC2). The corrected prediction uses $\hat{E}_1^{CC2} = E_1^{DFTB} + \widehat{\Delta E_1}$.

Model	**MAE**	**RMSE**	**Improvement (MAE)**
Baseline (DFTB)	0.0744	0.0970	1.0000
Δ-ML Corrected	0.0417	0.0534	1.7810

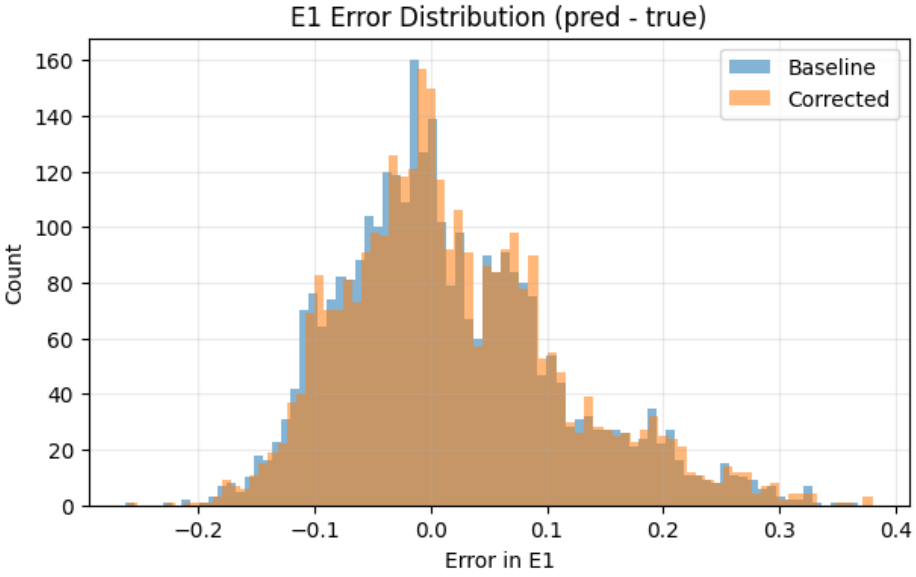

Figure 2: Graph 2: Error distribution for E_1 (prediction minus reference).

correction, predictions align more closely with the identity line, indicating that the residual model successfully captures systematic discrepancies between levels of theory.

Interpretation (Graph 2 – Error Distribution): The baseline error distribution is broader and shifted relative to zero. After correction, the distribution narrows and centers closer to zero, indicating reduced bias and variance. This confirms that the learned residual reduces both systematic and random components of prediction error.

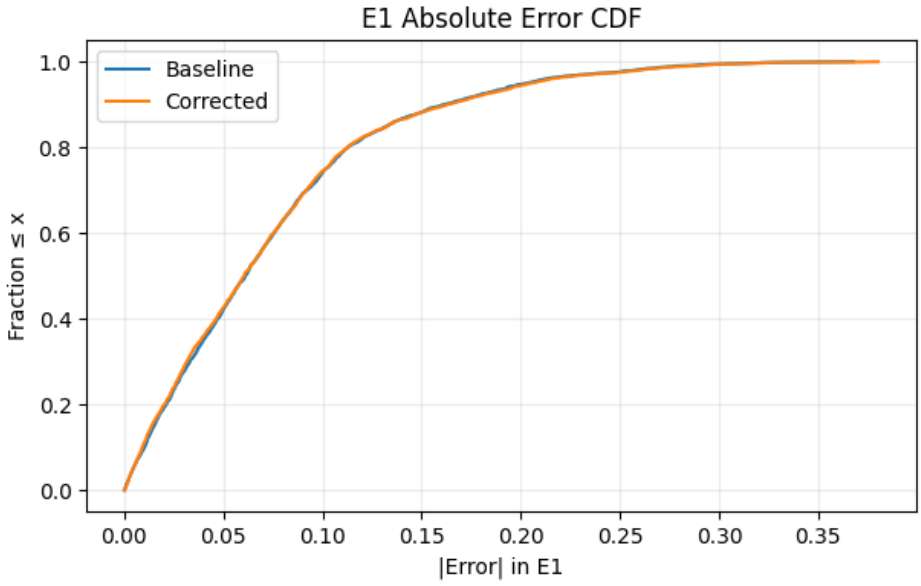

Figure 3: Graph 3: Absolute error distribution (or CDF) for E_1 baseline and corrected predictions.

Interpretation (Graph 3 – Absolute Error): The corrected model shifts the absolute error distribution toward lower values across the entire percentile range. Improvements are observed not only in mean error but also in higher percentiles (P90, P95), demonstrating consistent generalization rather than isolated improvements.

4.2 Oscillator Strength Prediction (f_1)

In contrast to excitation energy, oscillator strength exhibited more limited improvement under Δ-learning. This reflects the increased physical complexity of transition intensities, which depend on transition dipole moments and excited-state wavefunctions.

Table 2: Test-set performance for oscillator strength f_1 (DFTB → CC2). The corrected prediction uses $\hat{f}_1^{CC2} = f_1^{DFTB} + \widehat{\Delta f_1}$.

Model	MAE (a.u.)	Improvement (MAE)
Baseline (DFTB)	0.0327	1.0000
Δ-ML Corrected	0.0315	1.0380

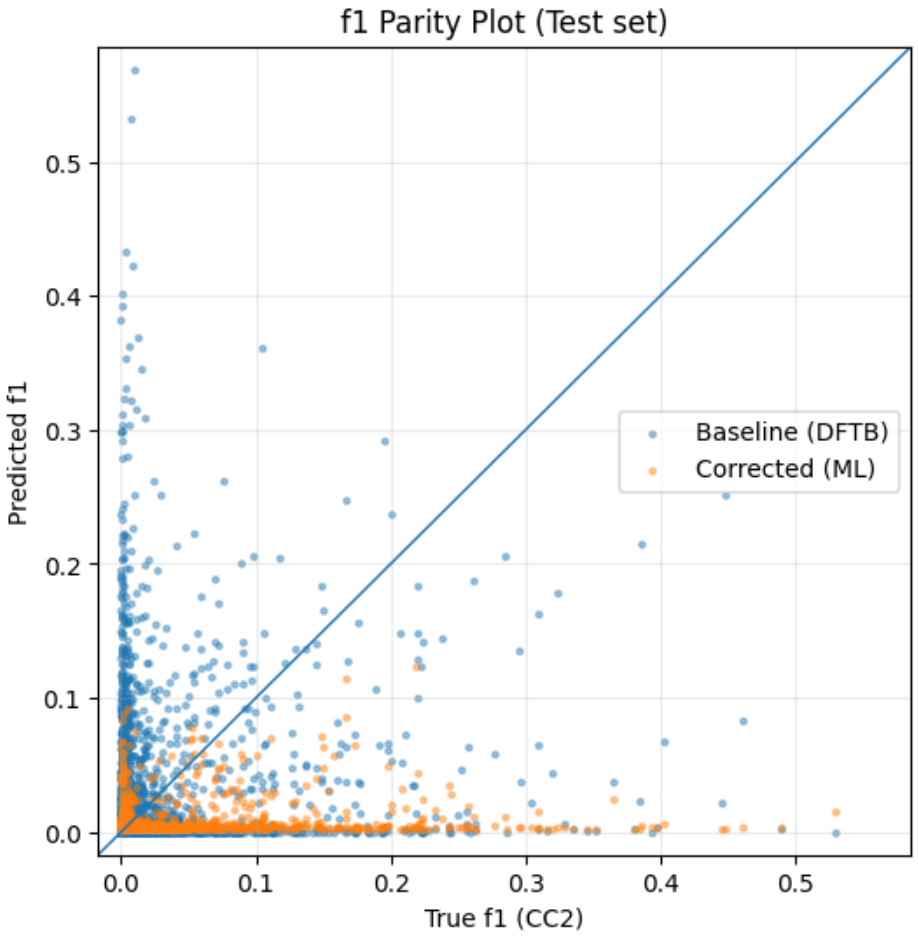

Figure 4: Graph 4: Parity plot for f_1 baseline and corrected predictions.

Interpretation (Graph 4 – Parity Plot): The oscillator strength baseline shows significant dispersion from the identity line. While machine learning reduces average deviation, large outliers remain. This suggests that geometric descriptors alone may not fully encode the electronic information governing transition intensities.

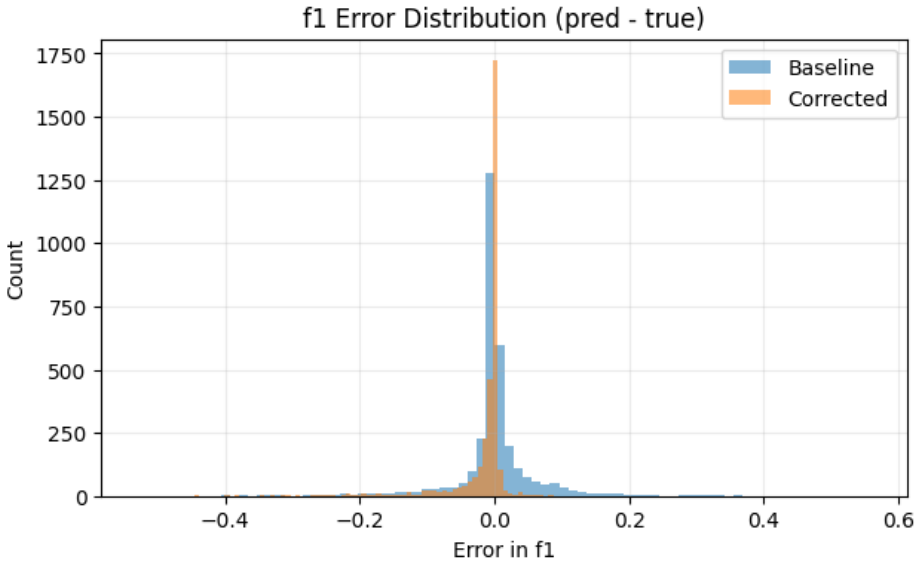

Figure 5: Graph 5: Error distribution for oscillator strength f_1.

Interpretation (Graph 5 – Error Distribution): Although the corrected model reduces central dispersion, heavy-tailed behavior persists. The error distribution remains skewed, indicating that certain transitions are intrinsically difficult to predict from structure alone.

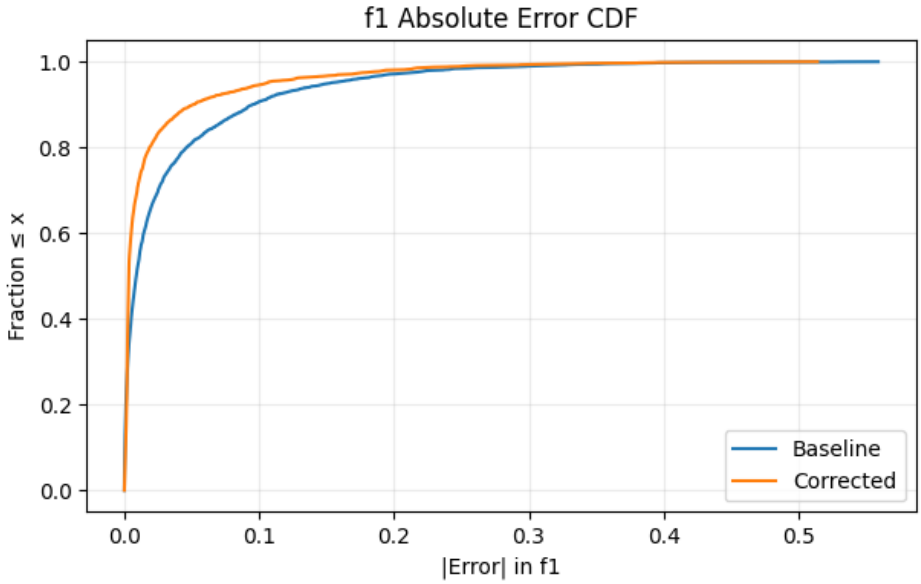

Figure 6: Graph 6: Absolute error distribution (or CDF) for oscillator strength.

Interpretation (Graph 6 – Absolute Error): Median error improves moderately; however, high-percentile errors (P95, P99) remain substantial. This confirms that oscillator strength prediction is more sensitive to electronic structure details not explicitly encoded in Coulomb matrix descriptors.

4.3 Neural Network Training Behavior

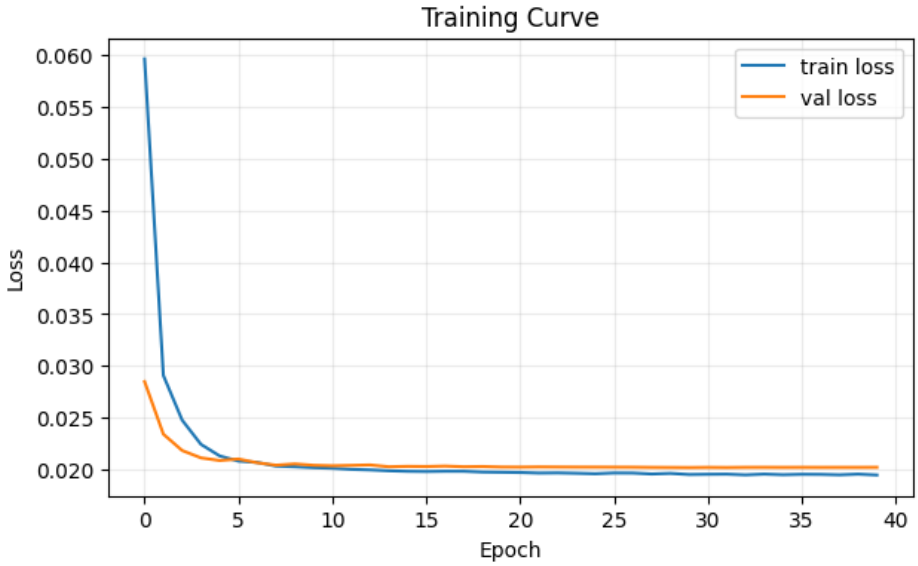

Figure 7: Graph 7: Training and validation loss curves for neural network model.

Interpretation (Graph 7 Training Curve): The training and validation loss curves demonstrate stable convergence without significant divergence, indicating controlled model capacity and effective regularization. Early stopping prevented overfitting while maintaining strong generalization performance.

4.4 Overall Interpretation

These results highlight a key scientific insight: residual learning is highly effective when systematic model errors are smooth functions of molecular structure, as observed for excitation energies. However, oscillator strengths depend more strongly on electronic wavefunction characteristics, limiting the predictive power of purely geometric descriptors.

Relation to Prior Work

This work builds on the Δ-learning framework introduced by Ramakrishnan et al. [1], who showed that learning the residual between two levels of quantum theory can be more effective than predicting absolute properties directly. In their study on ground-state molecular properties, residual learning significantly reduced prediction errors when correcting lower-level calculations toward higher-level references.

Here, the same idea was applied to excited-state properties from the QM8 dataset. By learning the residual $\Delta E_1 = E_1^{CC2} - E_1^{DFTB}$ from Coulomb matrix descriptors, the Mean Absolute Error was reduced from 0.07435 Hartree to 0.04173 Hartree on the test set, corresponding to a 1.78× improvement over the DFTB baseline.

While the improvement is more modest than what was reported for ground-state properties, it confirms that Δ-learning remains effective for excitation energies. In contrast, oscillator strength showed only limited improvement ($\sim$ 1.04×), highlighting the greater electronic complexity of transition intensities and the limitations of purely structure-based descriptors.

Overall, this study extends the Δ-learning approach to excited-state observables while also revealing where its effectiveness begins to diminish.

5 Discussion, Reflection, and Contribution

This project highlights both the potential and the limits of machine learning in scientific modeling. The results show that Δ-learning is very effective for excitation energy correction. By learning only the systematic difference between DFTB and CC2, the model significantly reduced prediction error. This supports an important idea in scientific machine learning: when physical approximations already exist, learning the residual can generalize better than predicting absolute quantities from scratch.

Oscillator strength prediction, however, was more challenging. Although direct learning improved upon the DFTB baseline, Δ-learning produced only marginal gains. This is physically reasonable. Oscillator strengths depend strongly on electronic wavefunctions and transition dipole moments, while the Coulomb matrix encodes only structural information. The persistent heavy-tailed errors suggest that geometry alone cannot fully describe transition intensities.

An important insight from this study is that predictive performance must be interpreted in light of underlying physics. Some properties are naturally more learnable than others given the same descriptor. The contrast between excitation energies and oscillator strengths illustrates how descriptor expressivity constrains model accuracy.

This project was conducted under the academic guidance of Dr. Ezekiel Oyeniyi, who provided research direction and conceptual feedback. I independently implemented the computational pipeline, including data preprocessing, model training (KRR, SVR, and neural networks), hyperparameter tuning, evaluation, and visualization.

All code is publicly available at:

`https://github.com/daisiwilliams20-tech/ml-corrections-dftb-excitations`

Overall, this work reflects my interest in combining physics and machine learning to improve computational efficiency while maintaining scientific rigor.

References

[1] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, *Machine Learning of Quantum Chemical Properties*, Physical Review Letters **114**, 143002 (2015).

[2] QM8 Dataset, Quantum Machine Learning Benchmark Dataset, available at: `https://deepchemdata.s3-us-west-1.amazonaws.com/datasets/qm8.csv`