Predicting Coupled Cluster Amplitudes using Machine Learning

Petar N. Panayotov
Yusuf Hamied Department of Chemistry, University of Cambridge
Supervisor: Prof. Alex Thom

September 2025

Abstract

Coupled Cluster (CC) theory provides highly accurate predictions of electronic structure but suffers from steep computational scaling. This work explores the use of machine learning to predict single-excitation Coupled Cluster amplitudes (t_1), with the aim of reducing computational cost while retaining predictive accuracy. We describe the construction of localized molecular orbital features, preprocessing pipelines, and neural network architectures. We discuss preliminary results, challenges, possible improvements, and future work.

I. INTRODUCTION

The accurate description of the electronic structure is a central problem in computational chemistry. Among the many approximate methods developed, Coupled Cluster Theory (CC) has been established as the "gold standard" for quantum chemical calculations due to its systematic improvability and high accuracy for ground-state correlation energies. [1] In particular, CC singles and doubles (CCSD) and perturbative triples [CCSD(T)] methods have been widely adopted across quantum chemistry and materials science.

However, the accuracy of CC theory comes at a steep computational cost. For a system with n_o occupied and n_v virtual orbitals, solving the nonlinear CC amplitude equations scales as $\mathcal{O}(n_o^2 n_v^4)$. [2] This scaling quickly renders routine CC calculations infeasible for large molecules, especially when high accuracy is required for potential energy surfaces, reaction dynamics, or condensed-phase systems. Reducing the cost of solving for the excitation amplitudes without sacrificing predictive power is, therefore, an important challenge.

Machine Learning (ML) provides a promising route towards accelerating CC calculations. By training on data generated from conventional calculations, ML models can learn correlations between molecular features and CC amplitudes, enabling efficient prediction of amplitudes for new systems or geometries. In this way, ML can serve as a surrogate for the expensive solution of nonlinear equations, potentially reducing scaling while preserving accuracy.

In this work we are not only interested in predicting amplitudes at equilibrium geometries, but also across distorted and non-equilibrium geometries. To facilitate this, we adopt a minimal STO-3G basis combined with localized orbitals. The rationale is that localized orbitals allow the CC amplitudes to vary smoothly and locally as molecules' geometries are distorted,

thereby embedding "chemical intuition" directly into the machine learning features. In this way, the model is encouraged to learn transferable patterns that reflect underlying chemistry, rather than overfitting to specific geometries or orbital bases. A longer-term goal is to develop models that remain robust and transferable across different basis sets, extending their practical applicability.

In this project we focus on predicting the single excitation amplitudes t_1 of CC theory. Although the t_1 amplitudes are formally zero in canonical Hartree–Fock orbitals due to Brillouin's theorem, they become non-zero in localized orbital frameworks and distorted geometries. The ability to predict these amplitudes accurately is an essential step toward a general ML-assisted CC framework, as it provides insight into feature construction, preprocessing, and network design.

The aim of this work is therefore to develop and benchmark an ML model capable of predicting localized single excitation amplitudes from molecular orbital features. We present the theoretical background, describe the feature engineering and data preprocessing strategies, and discuss a neural network architecture tailored to the physical structure of the amplitudes. Finally, we provide preliminary results and highlight current challenges and directions for improvement.

II. BACKGROUND

A. Coupled Cluster Theory

Coupled Cluster Theory (CC) provides a systematic approach to include the two-electron correlation by exponential parameteterization of the wavefunction: [3]

$$|\Psi_{\mathrm{CC}}\rangle = e^{T} |\Phi_0\rangle \,, \tag{1}$$

where $|\Phi_0\rangle$ is the Hartree–Fock reference determinant and T is the cluster operator. The operator T is expressed as a sum of excitation operators,

$$T = T_1 + T_2 + T_3 + \cdots \,, \tag{2}$$

with T_1 generating single excitations, T_2 doubles, etc.

In practice, truncations of T yield tractable approximations. The CCSD method retains T_1 and T_2, and CCSD (T) adds an aperturbative treatment of triple excitations. The nonlinear equations that determine the cluster amplitudes are obtained by projection of the similarity-transformed Schrödinger equation,

$$\langle \Phi_\mu | \, e^{-T} H e^{T} \, |\Phi_0\rangle = 0, \tag{3}$$

where $\{\Phi_\mu\}$ denote excited determinants. The formal computational scale of CCSD is $\mathcal{O}(n_o^2 n_v^4)$, where n_o and n_v are the numbers of virtual and occupied orbitals, respectively [2]. Although polynomial, this scaling is steep enough to prevent the routine use of CCSD for larger systems.

B. Single Excitation Amplitudes

For canonical Hartree–Fock orbitals, Brillouin's theorem implies that the single exciation amplitudes vanish,

$$t_i^a = 0 \quad \text{(canonical HF orbitals)}. \tag{4}$$

However, in localized orbital frameworks, especially in distorted geometries away from equilibrium, the t_1 amplitudes become nonzero and encode important electronic relaxation effects. Thus, these amplitudes provide a sensitive testbed for machine learning models.

In this work, the canonical t_1 amplitudes are obtained from CCSD calculations using PySCF. These amplitudes are then transformed for the localized orbitals via

$$\tilde{t}_{IA} = \left(U_{\mathrm{occ}}^{\dagger} \, t_1 \, U_{\mathrm{vir}} \right)_{IA} , \tag{5}$$

where U_{occ} and U_{vir} are the unitary rotation matrices that localize the occupied and virtual subspaces, respectively.

C. Orbital localization

To obtain chemically meaningful orbitals, both the occupied and virtual molecular orbitals are localized independently and then concatenated. Given canonical molecular orbitals

$$\psi_p = \sum_\mu \phi_\mu C_{\mu p}, \tag{6}$$

with coefficients $C_{\mu p}$ in the atomic orbital basis $\{\phi_\mu\}$, we separate the occupied and virtual subspaces:

$$\psi_i^{\mathrm{occ}} = \sum_\mu \phi_\mu C_{\mu i}, \tag{7}$$

$$\psi_a^{\mathrm{vir}} = \sum_\mu \phi_\mu C_{\mu a}. \tag{8}$$

Each subspace is then localized separately:

$$\tilde{\psi}_i^{\mathrm{occ}} = \sum_j \psi_j^{\mathrm{occ}} (U_{\mathrm{occ}})_{ji}, \tag{9}$$

$$\tilde{\psi}_a^{\mathrm{vir}} = \sum_b \psi_b^{\mathrm{vir}} (U_{\mathrm{vir}})_{ba}. \tag{10}$$

The two sets are concatenated to form the total localization matrix,

$$U = \begin{pmatrix} U_{\mathrm{occ}} & 0 \\ 0 & U_{\mathrm{vir}} \end{pmatrix}, \tag{11}$$

so that the localized orbital coefficients are $\tilde{C} = CU$

In this study, only the Pipek–Mezey localization procedure is applied to both occupied and virtual subspaces. [4] The localization is always initialized from the canonical molecular orbitals C provided by the Hartree–Fock calculation. However, because the iterative optimization in PySCF may occasionally converge to suboptimal local minima, we repeat the localization twenty times. Among these runs, the solution that maximizes the Pipek–Mezey objective function is selected, ensuring a consistent and maximally localized representation across different molecular geometries.

D. Motivation for Localized Orbital Features

A central goal of this study is to develop a machine learning framework that remains reliable across both equilibrium and distorted molecular geometries. To achieve this, the choice of orbital representation is critical. In canonical Hartree–Fock orbitals, the t_1 amplitudes vanish by Brillouin's theorem. Moreover, canonical orbitals themselves become highly delocalized when the geometry is distorted, leading to discontinuous and nonlocal variations that are challenging for a learning algorithm to capture.

In contrast, localized orbitals provide a representation in which the t_1 amplitudes vary smoothly and locally as molecular bonds are stretched or compressed. In this representation, the amplitudes associated with a given orbital are expected to change primarily in response to its local bonding environment, thereby embedding "chemical intuition " directly into the feature space. This is particularly important for training on non-equilibrium geometries, where chemical bonds may be progressively distorted.

In addition, we adopt the minimal STO-3G basis as a practical starting point. Although quantitatively limited, this basis ensures that the localized orbitals remain compact and chemically interpretable, reducing redundancy in the feature set. The combination of Pipek–Mezey localization and a minimal basis yields features that are smooth, local, and robust across molecular geometries. This provides a natural inductive bias for the machine learning model, encouraging it to learn patterns that generalize beyond individual training points.

A further long-term motivation is the transferability of the model between basis sets. Because features are defined in terms of localized orbitals rather than atom- or geometry-specific descriptors, the approach is not inherently tied to the STO-3G basis. The expectation is that a model trained on

localized orbitals in one basis may retain predictive power when applied to localized in a different basis, thereby broadening applicability to more chemically realistic scenarios. [5]

As shown in Fig. 1, the localized t_1 amplitudes for CH_4 vary smoothly and systematically as a C–H bond is stretched or compressed. Here, the horizontal axis represents the length of the bond relative to the equilibrium value R/R_0. This behavior supports the hypothesis that localized orbitals provide a chemically meaningful feature space for machine learning models.

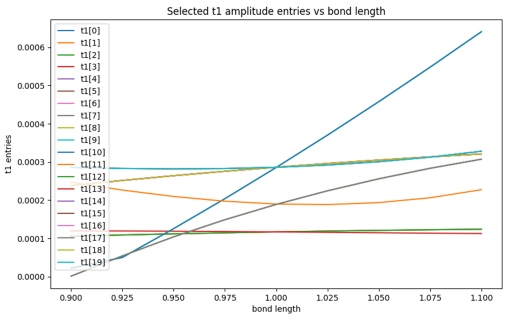

Figure 1: Variation of selected localized t_1 amplitudes for CH_4 as a function of one C–H bond length relative to its equilibrium value, R/R_0, in the STO-3G basis. Each curve corresponds to a distinct (i, a) orbital pair after Pipek–Mezey localization. The smooth, monotonic behavior of the amplitudes highlights the advantage of localized orbitals for encoding chemical intuition and facilitating machine learning models.

III. METHODS

A. Feature Construction

The predictive model relies on physically motivated descriptors extracted from localized molecular orbitals (LMOs). These features are designed to capture both local chemical environments and orbital-level information in a basis-independent manner. For each occupied–virtual orbital pair (i, a), we construct a feature vector $\mathbf{x}_{ia}$ as follows.

(i) Orbital populations. The electronic population of orbital $\tilde{\psi}_i$ on atom A is defined as

$$q_A = \sum_{\mu \in A} \sum_{\nu} \tilde{C}^*_{\mu i} \tilde{C}_{\nu i} S_{\mu\nu} = \sum_{\mu \in A} \left(\tilde{C}^\dagger S \tilde{C} \right)_{ii}, \tag{12}$$

where $S_{\mu\nu} = \langle \phi_\mu | \phi_\nu \rangle$ is the AO overlap matrix and $\tilde{C}$ are the coefficients of the LMOs. A threshold of $q_A > 0.15$ is used to determine which atoms the orbital is centered on. From this we extract categorical features encoding the atomic centers of each orbital, as well as the interatomic distances R_{AB} between centers.

(ii) Inverse interatomic distances. For orbitals localized on atoms A and B, we include descriptors of the form

$$d_{AB}^{-1} = \frac{1}{R_{AB}}, \tag{13}$$

for nearest-neighbor atoms associated with the orbital pair. These are standardized with masking, i.e. zero entries (padding) remain unscaled.

(iii) Angular momentum expectation. The expectation value of the z-component of the angular momentum operator for orbital $\tilde{\psi}_i$ is

$$\langle |L_z| \rangle_i = \sum_{\mu,\nu} \tilde{C}^*_{\mu i} \, |L_{\mu\nu}| \, \tilde{C}_{\nu i}, \tag{14}$$

where $|L| = U|\Lambda|U^\dagger$ is obtained by diagonalizing the AO angular momentum matrix L and taking absolute values of its eigenvalues Λ.

Because localized molecular orbitals (LMOs) may be centered on different numbers of atoms, we enforce a consistent alignment with respect to the z-axis prior to evaluation:

- For one-centered LMOs, the value of $\langle |L_z| \rangle$ is set to zero by convention, since no unique orientation exists.

- For two-centered LMOs, the bond vector is rotated to align with the z-axis.
- For three-centered LMOs, the plane defined by the three atoms is aligned such that its normal vector coincides with the z-axis.

This procedure ensures that the computed values of $\langle|L_z|\rangle$ are comparable across geometries and molecules, avoiding discontinuities due to arbitrary orbital orientations.

(iv) Orbital energies. For each localized orbital $\tilde{\psi}_i$, we compute the expectation value of the Fock operator:

$$\tilde{\epsilon}_i = \langle\tilde{\psi}_i|\hat{F}|\tilde{\psi}_i\rangle = \sum_p |U_{pi}|^2\,\epsilon_p, \tag{15}$$

which is a weighted average of the canonical MO energies ϵ_p with weights from the unitary transformation U.

(v) Orbital energy gap. For each occupied–virtual pair (i, a) we also compute the orbital energy gap.

$$\Delta\epsilon_{ia} = \epsilon_a - \epsilon_i. \tag{16}$$

The motivation for including $\Delta\epsilon_{ia}$ is rooted in many-body perturbation theory. In MBPT, double-excitation amplitudes scale as the inverse of the orbital energy denominators, while single excitations vanish in the canonical basis by Brillouin's theorem. Nevertheless, once orbitals are localized or distorted away from equilibrium, the t_1 amplitudes become nonzero, and it is reasonable to assume that they follow a similar inverse-gap trend. In coupled cluster calculations, MBPT amplitudes are commonly used as initial guesses, and the inverse-gap dependence remains a strong physical heuristic. Thus, the gap carries direct information about the expected scale of the t_1 amplitudes. In the present workflow, raw $\Delta\epsilon_{ia}$ is computed at this stage, while its nonlinear transforms and integration into the network architecture are described in Sections III.B and III.C.

(vi) Amplitude targets. The canonical CCSD t_1 amplitudes are expressed in the localized orbital basis by applying the occupied and virtual orbital rotations:

$$\tilde{t}_{IA} = \left(U_{\text{occ}}^{\dagger}\, t_1\, U_{\text{vir}}\right)_{IA}, \tag{17}$$

where U_{occ} and U_{vir} are the unitary matrices from Pipek–Mezey localization of the occupied and virtual subspaces. These localized amplitudes serve as the supervised targets for training.

Final feature vectors. The descriptors above yield fixed-length feature vectors $\mathbf{x}_i$ and $\mathbf{x}_a$ for each occupied and virtual orbital, respectively. These form the core inputs to the machine learning model. Additional physics-informed descriptors, most notably the orbital energy gap $\Delta\epsilon_{ia}$ and its nonlinear transforms, are introduced in Section III.C as part of the neural network architecture, since their inclusion is directly motivated by the structure of many-body perturbation theory.

B. Data Preprocessing

The raw orbital descriptors vary substantially in magnitude and distribution across molecules. To ensure numerical stability and comparability between training and test data, all features are preprocessed according to the following protocol.

(i) Standardization of orbital energies. Occupied and virtual orbital energies are standardized by z-score normalization,

$$\epsilon' = \frac{\epsilon - \mu_\epsilon}{\sigma_\epsilon}, \tag{18}$$

where μ_ϵ and σ_ϵ denote the mean and standard deviation computed from the training set only. This centers the distribution at zero with unit variance, preventing data leakage into validation and test sets.

(ii) Masked standardization of distance features. Inverse interatomic distances may include zero entries, corresponding to padding for non-existent atom pairs. These entries are preserved as zero, while nonzero values are standardized,

$$d' = \begin{cases} \dfrac{d - \mu_d}{\sigma_d}, & d \neq 0, \\ 0, & d = 0, \end{cases} \tag{19}$$

where μ_d and σ_d are computed only over nonzero distances.

(iii) Gap feature transforms. From the raw orbital energy gap $\Delta\epsilon_{ia}$ we construct a physics-informed feature vector motivated by many-body perturbation theory:

$$\phi(\Delta\epsilon_{ia}) = \Big[\, 1,\ \ln\Big(\Delta\epsilon_{ia}\Big),\ (\Delta\epsilon_{ia})^{-1} \Big]. \tag{20}$$

This choice reflects the approximate inverse dependence of single-excitation amplitudes on orbital energy denominators. The transformed components are standardized by z scores, while the constant feature 1 is retained as is. The raw data $\Delta\epsilon_{ia}$ is not passed directly to the network.

(iv) Other features. The expectation values of $|L_z|$ are retained without scaling, since their values are already bounded and chemically interpretable. The features of the atomic population are treated as categorical integers, with zero reserved for padding.

(v) Consistency across dataset splits. All preprocessing parameters (means and variances) are fixed from the training data and subsequently applied to the validation data. This guarantees consistency across splits and prevents information leakage.

C. Neural Network Architecture

The full network is illustrated in Fig. 2. Its design reflects both the structure of the orbital excitation problem and inductive biases from many-body perturbation theory. In particular, we enforce symmetry between occupied and virtual orbitals by using two parallel encoders, incorporate gap-dependent features explicitly, and factorize the output into sign and magnitude channels.

Orbital encoders. Occupied and virtual orbitals are processed separately by embedding and multilayer perceptrons (MLPs). For an occupied orbital i with feature vector x_i, the encoder is

$$\mathbf{h}_i^{(1)} = \rho\Big(x_i W_i^{(1)} + \mathbf{b}_i^{(1)}\Big), \tag{21}$$

$$\mathbf{h}_i^{(2)} = \rho\Big(\mathbf{h}_i^{(1)} W_i^{(2)} + \mathbf{b}_i^{(2)}\Big), \tag{22}$$

$$\mathbf{h}_i \ = \rho\Big(\mathbf{h}_i^{(2)} W_i^{(3)} + \mathbf{b}_i^{(3)}\Big), \tag{23}$$

where $W_i^{(\ell)}$ and $\mathbf{b}_i^{(\ell)}$ are learnable weights and biases. The same structure (with independent parameters) produces $\mathbf{h}_a$ from virtual features x_a. The nonlinearity is

$$\rho(x) = \ln\Big(0.5 + 0.5e^{x}\Big), \tag{24}$$

chosen for smoothness, non-saturation, and compatibility with wide feature scales.

Gap feature injection. From the orbital gap $\Delta\epsilon_{ia}$ we construct

$$\phi(\Delta\epsilon_{ia}) = \Big[\, 1,\ \ln(\Delta\epsilon_{ia}),\ (\Delta\epsilon_{ia})^{-1} \Big], \tag{25}$$

which is standardized as described in Section B.The inverse term reflects the approximate $1/\Delta\epsilon$ scaling of amplitudes in many-body perturbation theory, while the logarithmic term provides a more slowly varying feature that is consistent with the network's regression of log-magnitudes. Including both allows the model to exploit the known physical dependence of amplitudes on orbital gaps while retaining flexibility in how this dependence is expressed. The combined feature vector is

$$\boldsymbol{\xi}_{ia} = [\, \mathbf{h}_i \,;\, \mathbf{h}_a \,;\, \phi(\Delta\epsilon_{ia}) \,]. \tag{26}$$

Interaction term. Pairwise interactions between occupied and virtual spaces are captured by

$$g_{ia} = \mathbf{h}_i M \mathbf{h}_a^\top + \mathbf{u} \cdot (\mathbf{h}_i \odot \mathbf{h}_a) + \mathbf{w}_g \cdot \boldsymbol{\xi}_{ia} + c_g. \quad (27)$$

with $M \in \mathbb{R}^{d\times d}$, vectors $\mathbf{u}$, $\mathbf{w}_g$, and bias c_g as learnable parameters. Here $\odot$ denotes elementwise multiplication. This bilinear–linear form encodes both multiplicative couplings and additive contributions.

Output heads. Two heads predict sign and log-magnitude:

$$\hat{s}_{ia} = \tanh(\alpha_s g_{ia} + \mathbf{v}_s \cdot \boldsymbol{\xi}_{ia} + b_s)\,, \quad (28)$$

$$\widehat{\log|t_{ia}|} = \alpha_m g_{ia} + \mathbf{v}_m \cdot \boldsymbol{\xi}_{ia} + b_m, \quad (29)$$

with $\alpha_s, \alpha_m \in \mathbb{R}$ scalars and $\mathbf{v}_s, \mathbf{v}_m$ weight vectors. The final amplitude prediction is

$$\hat{t}_{ia} = \hat{s}_{ia}\, \exp\Big(\widehat{\log|t_{ia}|}\Big) / \Delta\epsilon_{ia}. \quad (30)$$

Loss and optimization. Training is based on a composite objective that balances accurate amplitude regression, sign consistency, weight regularization, and physics-inspired monotonicity constraints. Given reference amplitudes t_{ia} and predictions $(\hat{t}_{ia}, \hat{s}_{ia})$ on a mini-batch $\mathcal{B}$, the loss reads

$$\mathcal{J} = \frac{1}{|\mathcal{B}|} \sum_{(i,a)\in\mathcal{B}} \underbrace{\mathcal{L}_{\text{Huber}}(t_{ia}, \hat{t}_{ia})}_{\text{amplitude fit}} + \lambda_{\text{sign}} \underbrace{\mathcal{L}_{\text{sign}}(t_{ia}, \hat{s}_{ia})}_{\text{sign fit}} + \lambda_\ell \|\Theta\|^2 + \lambda_{\text{mono}}\, \mathcal{R}_{\text{gap}}. \quad (31)$$

The first term is a Huber loss, combining squared error near the origin with linear growth on large deviations, thereby reducing sensitivity to outliers. Explicitly,

$$\mathcal{L}_{\text{Huber}}(t, \hat{t}) = \begin{cases} \frac{1}{2}(t-\hat{t})^2, & |t-\hat{t}| \le \delta, \\ \delta\,(|t-\hat{t}| - \frac{1}{2}\delta), & |t-\hat{t}| > \delta, \end{cases} \quad (32)$$

where δ is the transition point between quadratic and linear regimes.

The second term supervises the predicted sign $\hat{s}_{ia}$ against the ground-truth $\text{sign}(t_{ia})$, weighted by λ_{sign}. With $p = (1+\hat{s}_{ia})/2$ as the predicted sign probability and $y_{\text{sign}} = (1+\text{sign}(t_{ia}))/2$ as the binary label, the loss is the binary cross-entropy

$$\mathcal{L}_{\text{sign}}(t, \hat{s}) = -\Big[y_{\text{sign}} \log p + (1 - y_{\text{sign}}) \log(1-p) \Big]. \quad (33)$$

The third term is standard ℓ_2 regularization on all network parameters Θ, weighted by λ_ℓ.

The final term $\mathcal{R}_{\text{gap}}$ penalizes violations of the monotonic dependence of amplitudes on the orbital gap $\Delta\epsilon$, weighted by λ_{mono}. In practice, this is implemented via finite differences:

$$\mathcal{R}_{\text{gap}} = \max\left(0,\, \frac{|\,\hat{t}_{ia}(\Delta\epsilon+\eta)|\, - |\,\hat{t}_{ia}(\Delta\epsilon)|}{\eta}\right), \quad (34)$$

where η is a small step size. This encourages $|\hat{t}_{ia}|$ to decrease as $\Delta\epsilon$ increases, consistent with the physical inverse-gap trend.

Optimization is performed with Adam, with learning rate tuned on the validation set.

Design rationale. Each architectural component was chosen to reflect the structure of the excitation problem and to embed prior physical knowledge where possible. The use of twin encoders allows occupied and virtual orbitals to be processed separately, ensuring that their representations are learned independently while maintaining a symmetric treatment of the two spaces. The explicit inclusion of the gap feature vector $\phi(\Delta\epsilon)$ incorporates the inverse-gap dependence suggested by many-body perturbation theory, reducing the need for the network to learn this trend from scratch. The bilinear interaction term g_{ia} provides a mechanism for capturing pairwise couplings between orbitals that would be missed by

simple concatenation of embeddings, thus improving expressivity. Finally, factorizing the output into sign and magnitude channels mirrors the perturbative structure of the amplitudes: the overall scale depends primarily on the inverse gap, while the sign arises from interference between orbital contributions. This separation was intended to stabilize training and to improve interpretability of the learned model.

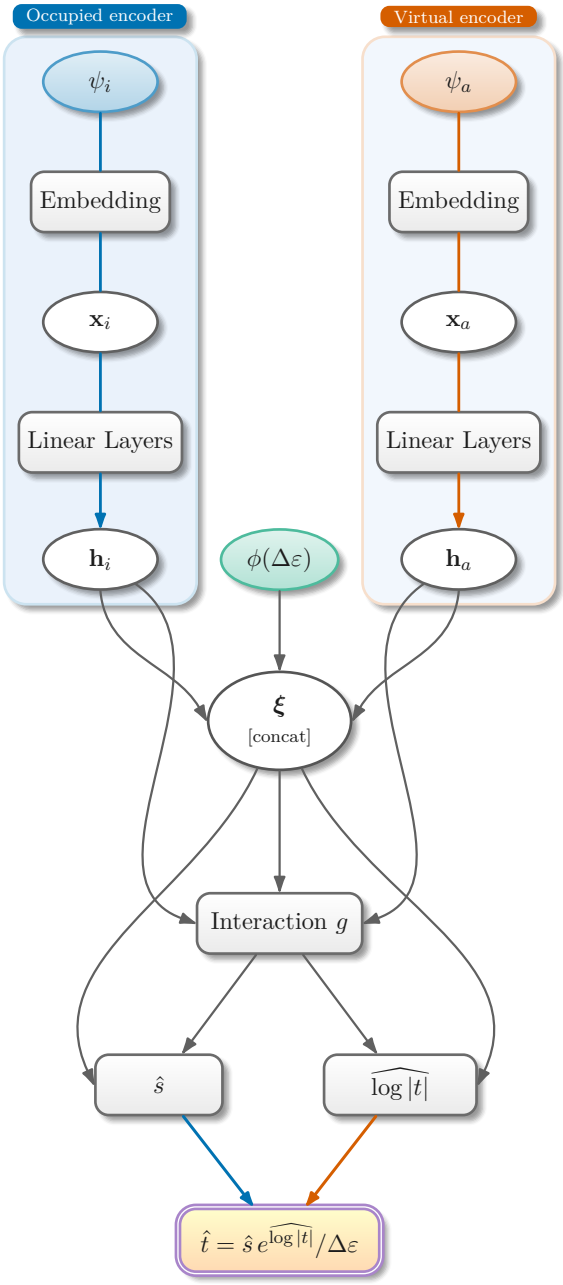

Figure 2: Neural network architecture. Occupied (blue) and virtual (vermillion) encoders produce $\mathbf{h}_i$ and $\mathbf{h}_a$; gap features $\phi(\Delta\varepsilon)$ join the concatenation $\boldsymbol{\xi}$, which interacts via g to yield sign and magnitude heads combined into $\hat{t}$.

Training protocol. All models were trained using Adam optimizer with learning rates in the range 10^{-4}–10^{-3} and batch sizes of 64. Training proceeded for up to 200 epochs, with early stopping based on validation loss. The Huber loss parameter was fixed at $\delta = 1.0$. The regularization weights were set as $\lambda_{\text{sign}} = 1.0$, $\lambda_\ell = 10^{-4}$, and $\lambda_{\text{mono}} = 0.1$, unless otherwise stated. Multiple hyperparameter settings were explored, but predictive accuracy remained limited, as discussed in Section IV.

IV. RESULTS AND DISCUSSION

A. Neural network training results

As shown previously in Fig. 1, localized t_1 amplitudes vary smoothly with molecular distortions, providing a promising target for supervised learning. Therefore, we trained the neural network described in Section III to predict these amplitudes from orbital features.

In practice, the model failed to achieve meaningful predictive accuracy across a relatively large dataset of molecules and geometries. Although training loss decreased slightly during optimization, validation loss plateaued at a high value and did not improve across epochs. On the held-out test sets, the mean absolute error remained well above chemical accuracy thresholds.

It is important to note that if the training set were restricted to a single molecule such as CH_4 and its distorted geometries, the network would likely succeed, since the mapping between features and amplitudes in that case is smooth and local (see Fig. 1). The present work, however, sought transferability across molecules, and it is precisely in this more challenging regime that the model fails to generalize.

B. Analysis of difficulties

The limited performance of the present neural network can be attributed to several factors, both on the data side and in the model design.

Dataset limitations. First, the dataset, while including a range of molecules, remains too small and too close to equilibrium geometries. As discussed in Section II, our goal is to predict amplitudes reliably across distorted geometries, yet the training set contains relatively few non-equilibrium structures. A larger dataset with systematic sampling of stretched, bent, and otherwise distorted geometries will be essential to expose the model to the full range of amplitude variation.

Local versus global orbital interactions. Second, the present model considers only pairwise interactions between a given occupied orbital ϕ_i and virtual orbital ϕ_a when predicting the corresponding amplitude t_{ia}. This ignores the fact that in correlated wavefunctions, excitations are coupled across many orbitals simultaneously. In chemical terms, orbitals interact not in isolation but as part of a collective many-body environment. A model that predicts each t_{ia} independently is therefore fundamentally limited.

V. CONCLUSIONS AND OUTLOOK

In this work we explored a physics-informed machine learning approach to predicting localized single-excitation amplitudes from coupled cluster theory. A pipeline was developed to construct orbital-level descriptors, preprocess features, and train a neural network architecture with explicit inverse-gap scaling and sign–magnitude factorization. While the model encoded meaningful inductive biases, it failed to achieve predictive accuracy across a diverse dataset of molecules and geometries. This outcome highlights both the promise and the limitations of orbital-based machine learning in its current form.

Future progress will require advances in two key directions. First, dataset size and diversity must be increased substantially, with systematic inclusion of non-equilibrium geometries. Second, the model architecture must move beyond local pairwise mappings. A promising avenue is to provide the network with the full matrix of orbital descriptors for each molecule,

and to predict all t_1 amplitudes jointly. This calls for architectures capable of capturing many-body interactions, with transformers as a natural candidate due to their attention-based representation of orbital couplings. Finally, feature sets should be enriched with physics-informed descriptors such as two-electron Coulomb and exchange integrals, which directly encode electron–electron interactions. Together, such datasets and architectures could allow machine learning models to achieve true transferability across molecules and basis sets, moving closer to a practical acceleration of correlated wavefunction methods.

ACKNOWLEDGEMENTS

The author wishes to thank Prof. Alex Thom for supervision, guidance, and many insightful discussions throughout the course of this project. The author is also grateful to the members of the Yusuf Hamied Department of Chemistry for providing a stimulating research environment. Computational resources were provided by the University of Cambridge.

Financial support for this work was generously provided by the Friends of Peterhouse Fund, established through donations from Petreans, whose assistance made it possible to undertake this summer research project.

References

[1] Rodney J. Bartlett and Monika Musiał. Coupled-cluster theory in quantum chemistry. *Reviews of Modern Physics*, 79(1):291–352, 2007.

[2] Isaiah Shavitt and Rodney J. Bartlett. *Many-Body Methods in Chemistry and Physics: MBPT and Coupled-Cluster Theory*. Cambridge University Press, 2009. See Chapter 9.5, p. 291 for the CCSD scaling $\mathcal{O}(n_o^2 n_v^4)$.

[3] Jiří Čížek. On the correlation problem in atomic and molecular systems. calculation of wavefunction components in ursell-type expansion using quantum field theoretical methods. *Journal of Chemical Physics*, 45:4256–4266, 1966.

[4] János Pipek and Paul G. Mezey. A fast intrinsic localization procedure applicable for ab initio and semiempirical linear combination of atomic orbital wave functions. *Journal of Chemical Physics*, 90(9):4916–4926, 1989.

[5] Matthew Welborn, Lixue Cheng, and Thomas F. Miller III. Transferability in machine learning for electronic structure via the molecular orbital basis. *Journal of Chemical Theory and Computation*, 14(9):4772–4779, 2018.