Graph Neural Networks on One- and Two-Body Integrals for Molecular Energy Prediction

Juan S. Carrasquilla-Gomez, and Rodrigo A. Vargas-Hernández
Department of Chemistry and Chemical Biology,
McMaster University, Hamilton, Canada
carrasqj@mcmaster.ca, vargashr@mcmaster.ca

Abstract

We present a graph neural network (GNN) framework for predicting molecular energies from molecular orbital graphs. Our approach leverages information from one- and two-electron integrals encoded as graph features, together with pooling strategies that map orbital-level predictions to molecular energies. The proposed approach was tested on 6 diatomic molecules, a total of 132 geometries, where the target molecular energy was computed using Full Configuration Interactions. Our results illustrate that a GNN architecture, even for this small dataset, is capable of learning the energy value given the molecular representation through the molecular orbitals.

1 Introduction

A central problem in quantum chemistry is the prediction of molecular ground-state energies, which requires solving the electronic structure problem (ESP). Approaches to the ESP range from mean-field methods, such as Hartree–Fock and density functional theory (DFT), to post–*ab initio* methods, where a molecular orbital (MO) basis is essential for capturing electronic correlation. Post–*ab initio* methods span a hierarchy of approximations, from perturbative approaches (e.g., Møller–Plesset perturbation theory) and coupled-cluster methods to Full Configuration Interaction (FCI), the ultimate reference model that provides numerically exact solutions. However, the exponential scaling of FCI with system size renders it intractable beyond small molecules. Molecular orbitals also play a central role in quantum algorithm–based approaches to the ESP; for example, the widely studied Variational Quantum Eigensolver (VQE) constructs its ansatz from MO representations.

Machine learning (ML) has become an important tool in molecular and materials simulations, as it can bypass explicit solutions of the electronic structure problem (ESP) and accelerate molecular modeling [1]. Applications range from materials discovery [2] to molecular property prediction and drug design [3]. A central challenge in these approaches is the choice of molecular representation, which strongly affects both predictive accuracy and generalization across chemical space [4]. Existing representations include string-based encodings such as SMILES, hand-crafted descriptors such as molecular fingerprints, and physics-inspired descriptors such as Coulomb matrices [5]. More recently, features derived from physico-chemical properties have also been employed.

Molecules can also be naturally expressed as graphs, where atoms are nodes and bonds are edges [6, 7]. This graph-based view preserves structural symmetries such as SE(3) equivariance [8] and has motivated the development of Graph Neural Networks (GNNs), which are now widely used for molecular property prediction [9–11].

Among physics-inspired options, atomic and molecular orbitals provide a compact basis derived from first principles and have demonstrated strong transferability in energy prediction tasks [12]. This "quantum chemistry–inspired" representation has been used in a plethora of supervised ML

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

models, from kernel methods to neural networks [12–15]. While some graph-based methods have incorporated physico-chemical descriptors [16] as node or edge features, a direct use of orbital representations within graph neural networks remains largely unexplored.

Some studies have explored combining orbital information with graph learning. OrbNet integrates molecular orbital features within a GNN to achieve accurate energy predictions [14], and similar ideas have been investigated by Chuiko et al. [17] using a Geminal-based feature representation. These results indicate that molecular orbitals provide a physically grounded feature space that can be incorporated with ML models. This leverages the molecular representation through one- and two-body integrals from Hartree-Fock orbitals as graph features for the regression of molecular energies.

2 Methodology

The electronic Hamiltonian in second quantization, for a set of N spin orbitals, is

$$\hat{H} = \sum_{pq}^{N} h_{pq}, a_p^\dagger a_q + \tfrac{1}{2} \sum_{pqrs}^{N} g_{pqrs}, a_p^\dagger a_q^\dagger a_r a_s, \tag{1}$$

where $a_p^\dagger$ (a_p) are the creation (annihilation) operators for spin orbital p, and h_{pq} and g_{pqrs} are the one- and two-electron integrals in the orbital basis. The electronic energy,

$$E_{\text{elec}} = \langle \Psi | \hat{H} | \Psi \rangle, \tag{2}$$

can be expressed in terms of the one- (Γ_{ik}) and two-particle ($D_{ij,kl}$) reduced density matrices (RDMs):

$$E_{\text{elec}} = \sum_{ik} \Gamma_{ik}, h_{ki} + \tfrac{1}{2} \sum_{ij,kl} D_{ij,kl}, g_{ijkl}. \tag{3}$$

Post–Hartree–Fock (post-HF) methodologies such as configuration interaction (CI), coupled cluster (CC), Møller–Plesset perturbation theory (MP2), and complete active space self-consistent field (CASSCF) all rely on a set of molecular orbitals (MOs) to construct h_{pq} and g_{pqrs}, typically starting from the Hartree–Fock reference. This motivates a general formalism of the form

$$E_{\text{elec}} \approx f_\theta(\{\psi_i\}_i^M), \tag{4}$$

where $\{\psi_i\}_i^M$ denotes the set of MOs and f_θ is a learnable function (e.g., a neural network, kernel method, symbolic regression, etc.). This idea has been explored previously using atomic orbital (AO) representations [12–14]. AO-based features are directly connected to the computation of forces via $\nabla_{\mathbf{R}i} E_{\text{elec}}$, but obtaining localized MOs often requires additional projection procedures such as Pipek–Mezey or Boys localization.

2.1 Graph representation of MOs

We employ molecular orbitals (MOs) as the molecular representation by constructing a graph object $\mathcal{G} = (V, E)$ from the Fock ($\mathbf{F}$), Coulomb ($\mathbf{J}$), and exchange ($\mathbf{K}$) matrices. In this representation, the nodes (V) correspond to the diagonal elements of these matrices, while the edges (E) are given by the off-diagonal elements. Since the Fock matrix $\mathbf{F}$ is diagonal in the MO basis, it only contributes node features and does not provide edge information. To reduce the number of edges, we prune connections between two MOs whenever $\max\big(|J_{ij}|, |K_{ij}|\big) < \epsilon$, where ϵ is a small cutoff [15]; here set to 10^{-10}. This criterion assumes that orbitals with negligible interaction need not be connected and also improves the efficiency of the graph convolution steps. An example of this construction is shown in Fig. 1 for a toy molecule with two molecular orbitals (e.g., H_2). All matrices were computed using PySCF and stored to avoid recomputation during training. The matrix calculation takes on average 0.15 s per geometry.

2.2 Graph neural network for energy prediction

The molecular energy is predicted in a regression setting using the MO-based graph representation. Node features are initialized as $\mathbf{x}_i = [F_{ii}, J_{ii}, K_{ii}]$, and embedded into a latent space of dimension

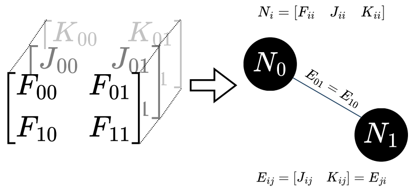

Figure 1: Matrix to graph parametrization for a pair of molecular orbitals.

64 through an MLP. The message-passing stage consists of three graph convolution layers. The first layer is a `NNConv` [18], which incorporates edge features to update the node embeddings. Subsequent layers refine the representation through nonlinear transformations and neighborhood aggregation. For graph-level readout, we evaluated both mean and sum pooling, finding that the latter yielded more accurate results. Finally, the pooled graph embedding is passed through a two-layer multilayer perceptron (MLP) to produce the energy prediction. We considered two variations of the same GNN: one trained individually for each system (GNN_i) and one trained jointly across all systems (GNN_G).

Both general architectures were selected after an extensive search over different layer depths and convolutional combinations. Given the dataset size, we restricted the model to a maximum of four layers after the initial NNConv layer, which incorporates edge features. The neural network used to map edge features within NNConv was varied from a simple linear augmenter to MLPs. The best performance was achieved with an MLP of dimensions $[3, 32, hid\_dim, hid\_dim]$ and LeakyReLU activations. The latent space dimension (hid_dim) was varied between 8 and 128 (in powers of 2). Both 64 and 128 dimensions gave comparable results, with 64 ultimately selected due to its lower computational cost. Batch normalization after activation layers was also tested, but consistently led to larger training and test errors, so it was discarded. Finally, motivated by universal machine learning interatomic potentials, where the energy scale also depends on system size or mass, we adopted a `global_add_pooling` operation to aggregate node embeddings.

2.3 Data and training procedure

We used the PennyLane [19] `Molecules` dataset, available at `https://pennylane.ai/datasets/collection/qchem`, which reports seven distinct diatomic molecules: CO (12 geometries), HF (42), Li_2 (12), LiH (42), N_2 (12), and O_2 (12). For all systems, the reference ground-state energies were computed using Full Configuration Interaction (FCI) with the STO-3G basis set. For both GNN_G and GNN_i, the dataset was split at the geometry level, with 80% of the geometries from each molecule used for training and the remaining 20% held out for testing. We used Adam with a learning rate of 6×10^{-4} and a batch size of 8. The GNN_i models were trained for 1,000 epochs, while GNN_G was trained for 2,000 epochs. All models were run on a GPU Nvidia RTX 4050.

3 Results and Discussion

The performance of both models, GNN_i and GNN_G, is presented in Fig. 2 and summarized in Table 1. Figure 2 shows the potential energy surfaces (PES) for all six molecules, where the symbols $\times$ denote test data outside the training set. Both models successfully capture the relationship between interatomic distance and molecular energy using the information provided by the molecular orbitals. For H_2, however, the normalization procedure led to a form of feature "degeneracy," where, across the sampled range of interatomic distances, the MO graph was characterized by only two predominant values of $\boldsymbol{F}$, $\boldsymbol{J}$, and $\boldsymbol{K}$, reason why results for H_2 are not reported. We aim to further investigate other normalization approaches for $\boldsymbol{F}$, $\boldsymbol{J}$, and $\boldsymbol{K}$ compatible with different chemical systems.

The noisier performance of GNN_i may stem from the smaller dataset used for its training, which constrains its ability to accurately represent the energy as a function of the integrals. In contrast, GNN_G is trained on the full dataset, allowing it to capture deeper relationships between the molecular orbitals and the energy, extending the knowledge from similar molecules to others with alike properties.

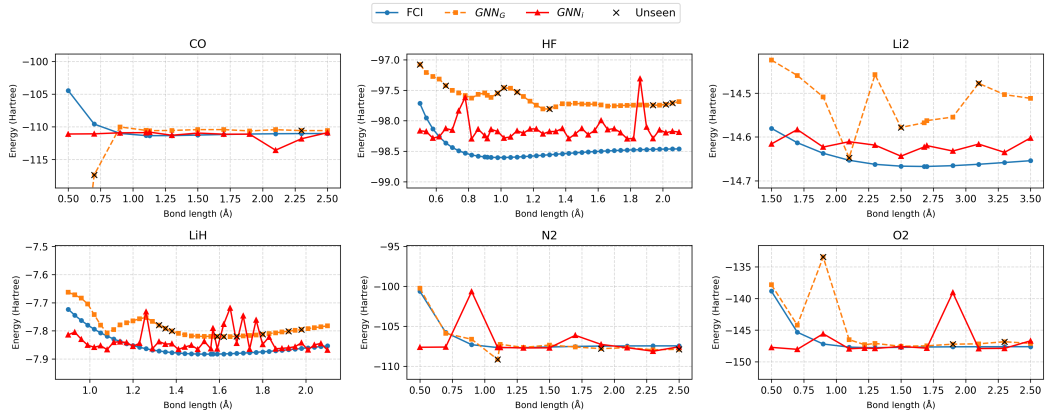

Figure 2: Predicted PES for diatomics. Blue curves represent the FCI energy, and orange and red the energy predicted with GNN_G and GNN_i, respectively. Symbols × are the test data; 80/20 split.

To evaluate performance, we computed the mean absolute error (MAE) between the predicted and FCI reference energies. Given the small dataset size, the data point with the largest error was excluded from the calculation to avoid disproportionate bias. Our average error per molecule is ~ 0.5, while high, illustrating that GNN_G architecture with additional data can be used to predict molecular energies.

Finally, one of the main motivations for using a GNN on molecular orbitals (MOs) is its ability to predict properties directly at the graph level. An important application is the Variational Quantum Eigensolver (VQE), a leading algorithm for near-term quantum chemistry that combines classical optimization with quantum state preparation to approximate molecular ground-state energies [20, 21]. In VQE, the molecular energy is obtained as the expectation value of the Hamiltonian, $\mathrm{E}_{\mathrm{VQE}}(\boldsymbol{\theta}) = \langle\psi(\boldsymbol{\theta})|H|\psi(\boldsymbol{\theta})\rangle$ with $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$ denoting the circuit parameters, H the molecular Hamiltonian, and $|\psi(\boldsymbol{\theta})\rangle$ the ansatz wavefunction. In practice, many ansätze are inspired by classical quantum chemistry methods such as coupled cluster (CC), where $\boldsymbol{\theta}$ corresponds to single and double excitation amplitudes. These excitations can be regarded as graph-level observables, making them natural targets for prediction with a GNN-based approach.

As a proof of concept, we conducted a secondary experiment using the learned embeddings from GNN_G. Specifically, we employed the node embeddings obtained after the first convolution layer and applied permutationally invariant operations to construct inputs for an MLP with two hidden layers of shape [4×64, 4×256, 4×256] and [4×256, 4×256, 1]. This model was trained for 1,000 epochs to predict the single-excitation parameter θ_{pr} between orbitals p and r, using the values provided in the PennyLane [19] `Molecules` dataset through its circuit "wires." The resulting model achieved a cosine similarity of **0.4514**, demonstrating the feasibility of leveraging GNN-based MO representations for downstream quantum computing tasks such as excitation prediction.

Molecule	CO	HF	Li_2	LiH	N_2	O_2
MAE	1.3399	0.8448	0.1207	0.0659	0.3065	0.6331

Table 1: Testing performance of the GNN_G (All measures in Hartree).

4 Summary

In this work, we introduced a GNN-based framework for energy prediction, demonstrating how orbital graph representations can capture the essential physics of electronic structure. Although the dataset was small, our results indicate that GNN-based architectures can be developed for molecular orbitals. In future work, it would be valuable to explore more sophisticated preprocessing of the one- and two-body integrals. For instance, recent studies have proposed sign-preserving logarithmic transformations that separate positive and negative entries, log-scale them, and standardize independently before concatenation [15]. Such approaches stabilize learning across the wide dynamic range of integral values while retaining sign information, and could further improve the robustness and generalization of orbital-based graph models.

Acknowledgment

JSCG acknowledges support from the Mitacs GRI program. RAVH acknowledges the support from the Digital Research Alliance of Canada and NSERC Discovery Grant No. RGPIN-2024-06594.

References

[1] Abdulrahman Aldossary, Jorge Arturo Campos-Gonzalez-Angulo, Sergio Pablo-García, Shi Xuan Leong, Ella Miray Rajaonson, Luca Thiede, Gary Tom, Andrew Wang, Davide Avagliano, and Alán Aspuru-Guzik. In silico chemical experiments in the age of ai: From quantum chemistry to machine learning and back. *Advanced Materials*, 36(30):2402369, 2024. doi: https://doi.org/10.1002/adma.202402369. URL `https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/adma.202402369`.

[2] Shunning Li, Yuanji Liu, Dong Chen, Yi Jiang, Zhiwei Nie, and Feng Pan. Encoding the atomic structure for machine learning in materials science. *WIREs Computational Molecular Science*, 12(1):e1558, 2022. doi: https://doi.org/10.1002/wcms.1558. URL `https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1558`.

[3] Guy Durant, Fergus Boyles, Kristian Birchall, and Charlotte M. Deane. The future of machine learning for small-molecule drug discovery will be driven by data. *Nature Computational Science*, 4(10):735–743, Oct 2024. ISSN 2662-8457. doi: 10.1038/s43588-024-00699-0. URL `https://doi.org/10.1038/s43588-024-00699-0`.

[4] Yonatan Harnik and Anat Milo. A focus on molecular representation learning for the prediction of chemical properties. *Chem Sci*, 15(14):5052–5055, March 2024.

[5] Grier M. Jones, Brittany Story, Vasileios Maroulas, and Konstantinos D. Vogiatzis. *Molecular Representations for Machine Learning*. American Chemical Society, Washington, DC, USA, 2023. doi: 10.1021/acsinfocus.7e7006. URL `https://pubs.acs.org/doi/abs/10.1021/acsinfocus.7e7006`.

[6] Maria Boulougouri, Pierre Vandergheynst, and Daniel Probst. Molecular set representation learning. *Nature Machine Intelligence*, 6(7):754–763, Jul 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00856-0. URL `https://doi.org/10.1038/s42256-024-00856-0`.

[7] Jay Morgan, Adeline Paiement, and Christian Klinke. Domain-informed graph neural networks: a quantum chemistry case study, 2022. URL `https://arxiv.org/abs/2208.11934`.

[8] Patrick Reiser, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, and Pascal Friederich. Graph neural networks for materials science and chemistry. *Communications Materials*, 3(1):93, Nov 2022. ISSN 2662-4443. doi: 10.1038/s43246-022-00315-6. URL `https://doi.org/10.1038/s43246-022-00315-6`.

[9] William L. Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications, 2018. URL `https://arxiv.org/abs/1709.05584`.

[10] Oliver Wieder, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, and Thierry Langer. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37:1–12, 2020. ISSN 1740-6749. doi: https://doi.org/10.1016/j.ddtec.2020.11.009. URL `https://www.sciencedirect.com/science/article/pii/S1740674920300305`.

[11] Zhenxing Wu, Jike Wang, Hongyan Du, Dejun Jiang, Yu Kang, Dan Li, Peichen Pan, Yafeng Deng, Dongsheng Cao, Chang-Yu Hsieh, and Tingjun Hou. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nature Communications*, 14(1):2585, May 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38192-3. URL `https://doi.org/10.1038/s41467-023-38192-3`.

[12] Matthew Welborn, Lixue Cheng, and Thomas F. III Miller. Transferability in machine learning for electronic structure via the molecular orbital basis. *Journal of Chemical Theory and Computation*, 14(9):4772–4779, 2018. doi: 10.1021/acs.jctc.8b00636. URL `https://doi.org/10.1021/acs.jctc.8b00636`. PMID: 30040892.

[13] Lixue Cheng, Jiace Sun, J. Emiliano Deustua, Vignesh C. Bhethanabotla, and III Miller, Thomas F. Molecular-orbital-based machine learning for open-shell and multi-reference systems with kernel addition gaussian process regression. *The Journal of Chemical Physics*, 157(15): 154105, 10 2022. ISSN 0021-9606. doi: 10.1063/5.0110886. URL `https://doi.org/10.1063/5.0110886`.

[14] Zhuoran Qiao, Matthew Welborn, Animashree Anandkumar, Frederick R. Manby, and III Miller, Thomas F. Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *The Journal of Chemical Physics*, 153(12):124111, 09 2020. ISSN 0021-9606. doi: 10.1063/5.0021955. URL `https://doi.org/10.1063/5.0021955`.

[15] Christian Venturella, Jiachen Li, Christopher Hillenbrand, Ximena Leyva Peralta, Jessica Liu, and Tianyu Zhu. Unified deep learning framework for many-body quantum chemistry via green's functions. *Nature Computational Science*, 5(6):502–513, Jun 2025. ISSN 2662-8457. doi: 10.1038/s43588-025-00810-z. URL `https://doi.org/10.1038/s43588-025-00810-z`.

[16] Daniil A. Boiko, Thiago Reschützegger, Benjamin Sanchez-Lengeling, Samuel M. Blau, and Gabe Gomes. Advancing molecular machine learning representations with stereoelectronics-infused molecular graphs. *Nature Machine Intelligence*, 7(5):771–781, May 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01031-9. URL `https://doi.org/10.1038/s42256-025-01031-9`.

[17] Valerii Chuiko and Paul W. Ayers. Predicting energy of the quantum system from one- and two-electron integrals using deep learning, 2025. URL `https://arxiv.org/abs/2504.03849`.

[18] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry, 2017. URL `https://arxiv.org/abs/1704.01212`.

[19] Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M. Sohaib Alam, Guillermo Alonso-Linaje, B. AkashNarayanan, Ali Asadi, Juan Miguel Arrazola, Utkarsh Azad, Sam Banning, Carsten Blank, Thomas R Bromley, Benjamin A. Cordier, Jack Ceroni, Alain Delgado, Olivia Di Matteo, Amintor Dusko, Tanya Garg, Diego Guala, Anthony Hayes, Ryan Hill, Aroosa Ijaz, Theodor Isacsson, David Ittah, Soran Jahangiri, Prateek Jain, Edward Jiang, Ankit Khandelwal, Korbinian Kottmann, Robert A. Lang, Christina Lee, Thomas Loke, Angus Lowe, Keri McKiernan, Johannes Jakob Meyer, J. A. Montañez-Barrera, Romain Moyard, Zeyue Niu, Lee James O'Riordan, Steven Oud, Ashish Panigrahi, Chae-Yeun Park, Daniel Polatajko, Nicolás Quesada, Chase Roberts, Nahum Sá, Isidor Schoch, Borun Shi, Shuli Shu, Sukin Sim, Arshpreet Singh, Ingrid Strandberg, Jay Soni, Antal Száva, Slimane Thabet, Rodrigo A. Vargas-Hernández, Trevor Vincent, Nicola Vitucci, Maurice Weber, David Wierichs, Roeland Wiersema, Moritz Willmann, Vincent Wong, Shaoming Zhang, and Nathan Killoran. Pennylane: Automatic differentiation of hybrid quantum-classical computations, 2022. URL `https://arxiv.org/abs/1811.04968`.

[20] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, 5(1), July 2014. ISSN 2041-1723. doi: 10.1038/ncomms5213. URL `http://dx.doi.org/10.1038/ncomms5213`.

[21] Jules Tilly, Hongxiang Chen, Shuxiang Cao, Dario Picozzi, Kanav Setia, Ying Li, Edward Grant, Leonard Wossnig, Ivan Rungger, George H. Booth, and Jonathan Tennyson. The variational quantum eigensolver: A review of methods and best practices. *Physics Reports*, 986:1–128, November 2022. ISSN 0370-1573. doi: 10.1016/j.physrep.2022.08.003. URL `http://dx.doi.org/10.1016/j.physrep.2022.08.003`.