QH9: A Quantum Hamiltonian Prediction Benchmark for QM9 Molecules

Haiyang Yu[*]
Texas A&M University
College Station, TX 77843
haiyang@tamu.edu

Meng Liu[*]
Texas A&M University
College Station, TX 77843
mengliu@tamu.edu

Youzhi Luo
Texas A&M University
College Station, TX 77843
yzluo@tamu.edu

Alex Strasser
Texas A&M University
College Station, TX 77843
alexstrasser16410@tamu.edu

Xiaofeng Qian[†]
Texas A&M University
College Station, TX 77843
feng@tamu.edu

Xiaoning Qian[†]
Texas A&M University
College Station, TX 77843
xqian@ece.tamu.edu

Shuiwang Ji[†]
Texas A&M University
College Station, TX 77843
sji@tamu.edu

Abstract

Supervised machine learning approaches have been increasingly used in accelerating electronic structure prediction as surrogates of first-principle computational methods, such as density functional theory (DFT). While numerous quantum chemistry datasets focus on chemical properties and atomic forces, the ability to achieve accurate and efficient prediction of the Hamiltonian matrix is highly desired, as it is the most important and fundamental physical quantity that determines the quantum states of physical systems and chemical properties. In this work, we generate a new Quantum Hamiltonian dataset, named as QH9, to provide precise Hamiltonian matrices for 2,399 molecular dynamics trajectories and 130,831 stable molecular geometries, based on the QM9 dataset. By designing benchmark tasks with various molecules, we show that current machine learning models have the capacity to predict Hamiltonian matrices for arbitrary molecules. Both the QH9 dataset and the baseline models are provided to the community through an open-source benchmark, which can be highly valuable for developing machine learning methods and accelerating molecular and materials design for scientific and technological applications. Our benchmark is publicly available at https://github.com/divelab/AIRS/tree/main/OpenDFT/QHBench.

1 Introduction

Machine learning methods have shown great potential in accelerating computations in quantum chemistry tasks. For example, a variety of invariant geometric deep learning methods have been developed to encode pairwise distances and bond angles in molecular and materials systems [Schütt et al., 2018, Gasteiger et al., 2020, 2021, Liu et al., 2022, Wang et al., 2022] to accelerate the prediction of their chemical properties as data-driven surrogate approximations. To enhance the

---

[*]Equal contribution

[†]Equal senior contribution

prediction of vectorial properties, such as force fields, equivariant deep learning methods have been developed to capture permutation, translation, and rotation equivariance for equivariant property prediction [Satorras et al., 2021, Schütt et al., 2021, Thölke and Fabritiis, 2022, Thomas et al., 2018, Batzner et al., 2022, Fuchs et al., 2020, Liao and Smidt, 2023, Anderson et al., 2019, Brandstetter et al., 2022]. To support and facilitate the development of machine learning methods on quantum chemistry property prediction, many datasets have been generated to benchmark the respective tasks on molecular property prediction [Blum and Reymond, 2009, Ruddigkeit et al., 2012, Ramakrishnan et al., 2014, Wang et al., 2009, Nakata and Shimazaki, 2017], catalyst prediction [Chanussot et al., 2021, Tran et al., 2023], and force field prediction [Chmiela et al., 2017, 2023].

In addition to these quantum chemistry prediction tasks, the quantum Hamiltonian is another significant and fundamental physical property that determines the quantum states and various materials properties[Marzari and Vanderbilt, 1997, Souza et al., 2001, Qian et al., 2010, Marzari et al., 2012, Bai et al., 2022]. The quantum Hamiltonian can be calculated using Density Functional Theory (DFT) [Hohenberg and Kohn, 1964, Kohn and Sham, 1965] with a time complexity of $O(n^3 T)$, where n represents the number of electrons and T denotes the number of optimization steps required to achieve convergence. Given the high computational complexity of the DFT algorithms, accelerating such calculations for novel molecular and materials systems becomes a desirable but challenging task. To tackle this challenge, machine learning methods, such as quantum tensor networks [Li et al., 2022, Gong et al., 2023, Schütt et al., 2019, Yu et al., 2023, Unke et al., 2021], provide a highly promising approach for accelerating the DFT algorithms. These networks directly predict the final Hamiltonian matrix given the input 3D geometries, resulting in significant acceleration of calculations by orders of magnitude.

Unlike invariant chemical properties, Hamiltonian matrices obey intrinsic block-by-block matrix equivariance. This equivariance can be represented by the rotation Wigner D-Matrix, which may contain higher order rotations beyond 3D space. In order to make physically meaningful predictions, it is important to design quantum tensor network architectures that preserve this equivariance property. To perform systematic and in-depth study of this new task, there is a clear need to generate large-scale quantum tensor datasets and benchmarks. Currently, the only quantum Hamiltonian datasets include the MD17 [Schütt et al., 2019, Gastegger et al., 2020] and mixed MD17 [Yu et al., 2023] datasets, which consist of data for a single and four molecules, respectively.

To provide a much larger and more realistic dataset, we generate a new quantum tensor dataset named QH9. This dataset contains Hamiltonian matrices for 130,831 stable molecular geometries and 2,399 molecular dynamic trajectories. In order to provide comprehensive studies for quantum tensor networks, we have designed four specific tasks. The first two tasks, QH-stable-iid and QH-stable-ood, aim to explore the performance of the networks in both in-distribution and out-of-distribution scenarios, specifically focusing on stable molecular geometries. The QH-dynamic-geo task follows the setting of the mixed MD17, containing the same molecule with different geometries in the training, validation, and test. On the other hand, the QH-dynamic-mol task splits the trajectories based on different molecules. Finally, we evaluate the transferrability of the trained models on molecules with larger sizes, thereby testing the models' ability to generalize beyond the training dataset. To demonstrate the quality of the predicted Hamilton matrix, we use four metrics. These metrics are based on the Mean Absolute Error (MAE) of the predicted Hamiltonian matrix $\boldsymbol{H}$, as well as the derived properties such as orbital energies ϵ and electronic wavefunction ψ. Furthermore, to evaluate the quality of the predicted Hamiltonian in accelerating DFT calculations, we calculate the DFT optimization ratio by taking the model predictions as DFT initialization.

2 Background and Related Works

2.1 Density Functional Theory (DFT)

Modeling the quantum states of physical systems is a central topic in computational quantum physics and chemistry. It aims to solve the Schrödinger equation [Schrödinger, 1926], which describes the electronic states shown as

$$\hat{H}\Psi\left(\boldsymbol{r}_1, \cdots, \boldsymbol{r}_n\right) = E\Psi\left(\boldsymbol{r}_1, \cdots, \boldsymbol{r}_n\right), \tag{1}$$

where $\Psi\left(\boldsymbol{r}_1, \cdots, \boldsymbol{r}_n\right)$ is the n-electronic wavefunctions and $\boldsymbol{r}$ is the 3D coordinates. Electronic eigenvalues and wavefunctions play an important role in calculating numerous crucial physical

properties, including the energy gap between the Highest Occupied Molecular Orbital (HOMO) and the Lowest Unoccupied Molecular Orbital(LUMO), *i.e.* the HOMO-LUMO gap, and charge density. However, due to the exponentially expanding input Hilbert space with the number of electrons, the computational cost to directly calculate many-electronic wavefunctions is extremely high. Therefore, various methods are proposed to approximate the solutions, such as the Hartree-Fock (HF) method [Szabo and Ostlund, 2012] that approximates the wavefunction itself, or density functional theory [Hohenberg and Kohn, 1964] that approximates the electron density. While the HF method scales with the number of electrons n as $O(n^4)$, DFT scales with $O(n^3)$ and therefore DFT is better suited for large-scale systems. DFT is based on the key discovery that the total energy and thus all ground-state properties of a system are uniquely determined by the ground-state electron density [Kohn and Sham, 1965].

Both of these approaches divide an n-electron system into a set of n non-interacting one-electron wavefunctions $\psi_i(\boldsymbol{r_i})$, also called molecular orbitals in molecular systems. These one-electron orbitals can then be approximated by a linear combination of basis functions $\phi_j(\boldsymbol{r})$ as $\psi_i(\boldsymbol{r}) = \sum_j C_{ij}\phi_j(\boldsymbol{r})$. The basis functions can be represented in analytical forms, such as Slater-type orbitals (STOs), Gaussian-type orbitals (GTOs), or plane waves, under numerical approximations for obtaining the coefficients matrix $\boldsymbol{C}$. With these approximations, the original Schrödinger Equation (1) for electrons can be transformed into a matrix form as

$$\boldsymbol{H}\boldsymbol{C}_i = \boldsymbol{\epsilon}_i\boldsymbol{S}\boldsymbol{C}_i, \tag{2}$$

where $\boldsymbol{H}$ is the Hamiltonian matrix, $\boldsymbol{S}$ is the overlap matrix, and $\boldsymbol{\epsilon}_i$ is the energy for the i-th orbital. The Hamiltonian matrix can be decomposed into the sum

$$\boldsymbol{H} = \boldsymbol{H}_{eN} + \boldsymbol{H}_{ee} + \boldsymbol{H}_{XC}, \tag{3}$$

which describes electron-ion interactions ($\boldsymbol{H}_{eN}$), electron-electron interactions ($\boldsymbol{H}_{ee}$, including kinetic energy and electron–electron Coulomb repulsion energy), and exchange-correlation energy ($\boldsymbol{H}_{XC}$). These matrices take the electron density $\boldsymbol{\rho}(\boldsymbol{r})$ as an input to evaluate the Hamiltonian matrix. The exchange-correlation energy functional used in this paper was B3LYP [Lee et al., 1988, Becke, 1993], which is a hybrid functional that includes both the exchange energy from the HF method as well as a correlation potential. We implement the GTO basis set Def2SVP [Weigend and Ahlrichs, 2005] in a post-HF method that incorporates aspects of DFT, namely, an exchange-correlation potential, in order to more accurately capture electron-electron interactions compared to the HF method, which uses a mean field approximation of electron density.

Equation (2) is satisfied for the final Hamiltonian matrix and its coefficient matrix once self-consistency is achieved using direct inversion in the iterative subspace (DIIS) [Pulay, 1980, 1982]. The equation is solved iteratively by building and solving the Hamiltonian and coefficient matrices, constructing an error vector based on a linear combination of energy differences in the previous steps, then diagonalizing and recalculating $\boldsymbol{H}$ until the error vector is below a convergence threshold.

2.2 Group Equivariance and Equivariant Matrices

In many quantum chemistry problems, the molecular property to be predicted (e.g., energy and force) is internally invariant or equivariant to transformations in SE(3) group, including rotations and translations. Formally, for an n-atom molecule whose 3D atom coordinates are $\boldsymbol{r}_1, ..., \boldsymbol{r}_n$, any transformation in SE(3) group can be described as changing the 3D atom coordinates to $\boldsymbol{R}\boldsymbol{r}_1 + \boldsymbol{t}, ..., \boldsymbol{R}\boldsymbol{r}_n + \boldsymbol{t}$. Here, the translation vector $\boldsymbol{t} \in \mathbb{R}^3$ is an arbitrary 3D vector, and the rotation matrix $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ satisfies that $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}, |\boldsymbol{R}| = 1$. Let $\boldsymbol{f}(\cdot)$ map the 3D atom coordinates to an $(2\ell + 1)$-dimensional prediction target vector, we say $\boldsymbol{f}$ is order-ℓ SE(3)-equivariant if

$$\boldsymbol{f}(\boldsymbol{R}\boldsymbol{r}_1 + \boldsymbol{t}, ..., \boldsymbol{R}\boldsymbol{r}_n + \boldsymbol{t}) = D^\ell(\boldsymbol{R})\boldsymbol{f}(\boldsymbol{r}_1, ..., \boldsymbol{r}_n) \tag{4}$$

holds for any rotation matrix $\boldsymbol{R}$ and translation vector $\boldsymbol{t}$, where $D^\ell(\boldsymbol{R}) \in \mathbb{C}^{(2\ell+1)\times(2\ell+1)}$ is the order-ℓ Wigner-D matrix of $\boldsymbol{R}$. To accurately predict SE(3)-equivariant properties, an effective approach is to develop neural network models that are designed to maintain the same equivariance relations between inputs and outputs as in Equation (4). Recently, many studies have proposed SE(3)-equivariant neural network architectures by using SE(3)-invariant feature encoding [Schütt et al., 2018, Gasteiger et al., 2020, 2021, Liu et al., 2022], tensor product operations [Thomas et al., 2018, Brandstetter et al., 2022, Liao and Smidt, 2023], or atomic cluster expansion framework [Batatia et al., 2022, Drautz, 2019, Dusson et al., 2022, Kovács et al., 2021].

Different from vector-like molecular properties, the Hamiltonian matrix $\boldsymbol{H}$ has a much more complicated SE(3) equivariance pattern that is associated with the intrinsic angular momentum of the atomic orbital pairs. In computational quantum chemistry algorithms such as DFT, the Hamiltonian matrix $\boldsymbol{H}$ can be used to represent the interactions between these atomic orbitals, and the block $\boldsymbol{H}_{ij}$ in Hamiltonian matrix represents the interactions between the atomic orbitals i in atom a_i with angular momentum ℓ_i and atomic orbitals j in atom a_j with angular momentum ℓ_j, and the shape of this block $\boldsymbol{H}_{ij}$ is $(2\ell_i + 1) \times (2\ell_j + 1)$. Usually, the atomic orbitals are arranged sequentially for the orbitals in the same atom and with the same angular momentum. For example, $\boldsymbol{H}_{ij}$ can be located within the s_i-th to $(s_i + 2\ell_i)$-th row, and the s_j-th to $(s_j + 2\ell_j)$-th column of Hamiltonian matrix $\boldsymbol{H}$. Specifically, its SE(3) equivariance can be described as

$$\boldsymbol{H}_{ij}\left(\boldsymbol{\rho}(\boldsymbol{R}\boldsymbol{r} + \boldsymbol{t})\right) = D^{\ell_i}(\boldsymbol{R})\boldsymbol{H}_{ij}\left(\boldsymbol{\rho}(\boldsymbol{r})\right)D^{\ell_j}(\boldsymbol{R}), \tag{5}$$

where $\boldsymbol{\rho}(\boldsymbol{r})$ is the electronic density at positioin $\boldsymbol{r}$ and Hamiltonian matrix $\boldsymbol{H}$ is a function of the electronic density $\boldsymbol{\rho}(\boldsymbol{r})$ in the DFT algorithm. In other words, the SE(3) equivariance of different submatrices in $\boldsymbol{H}$ has different mathematical forms, which is much more complicated than the SE(3) equivariance of vector-like molecular properties. Hence, it is much more challenging to develop SE(3)-equivariant neural network architectures for the prediction of Hamiltonian matrices. Nowadays, only a few studies [Li et al., 2022, Gong et al., 2023, Yu et al., 2023, Unke et al., 2021] have made initial exploration in this direction.

2.3 Datasets for Quantum Chemistry

To facilitate the usage of machine learning models to predict chemistry properties and accelerate simulations, numerous quantum chemistry datasets have been built to provide extensive and faithful data. Here, we introduce several existing datasets that have been constructed for different tasks respectively, including molecular property prediction, catalyst modeling, molecular force field prediction, and molecular Hamiltonian matrix prediction. For molecular property prediction, the QM7 [Blum and Reymond, 2009] dataset was initially constructed using 7,165 molecules from the organic molecule database GDB-13 [Blum and Reymond, 2009], with each selected molecule having no more than 7 heavy atoms. The primary purpose of creating the QM7 dataset is to provide atomization energies as the target molecular property. Then QM9 [Ramakrishnan et al., 2014, Schütt et al., 2018] was built based on GDB-17 [Ruddigkeit et al., 2012] to provide 134k stable small organic molecules with no more than 9 heavy atoms in each molecule. Moreover, it provides 13 different important quantum chemistry properties, including HOMO and LUMO energies. Based on the molecules from PubQChem [Wang et al., 2009, 2017, Kim et al., 2019, 2021, 2023], PubQChemQC [Nakata and Shimazaki, 2017] provides 3M ground-state molecular structures as well as the HOMO-LUMO gap and excitation energies for 2M molecules. In addition to the molecular property datasets, OC20 [Chanussot et al., 2021] and OC22 [Tran et al., 2023] were developed to provide the data of interactions of catalysts on material surfaces. They provide the geometries of the initial structures to predict the final structures or energies as well as the relaxation trajectories with energy and atomic forces. For the molecular force field prediction datasets, MD17 [Chmiela et al., 2017] and MD22 [Chmiela et al., 2023] contain atomic forces for molecular and supramolecular trajectories respectively as valuable datasets to develop machine learning methods. The last category is the Hamiltonian matrices datasets. MD17 [Schütt et al., 2019, Gastegger et al., 2020] provides the Hamiltonian matrices for single molecular dynamic trajectories to study the Hamiltonian matrices for molecules with various geometries. Building upon this dataset, mixed MD17 [Yu et al., 2023] combines four molecular trajectories in the MD17 to study Hamiltonian matrix prediction tasks with multiple molecules. Alongside the increasing interest in Hamiltonian matrix prediction, there is a growing need for datasets that include Hamiltonian matrices with a greater number of molecules to facilitate the subsequent studies.

3 Datasets, Tasks, Methods, and Metrics

3.1 Datasets

Dataset Generation. For the QH9 dataset, we use open-source software PySCF [Sun et al., 2018, 2020] to conduct computational quantum chemistry calculations. In the QH9, there are two sub datasets. The first one is the QH-stable dataset containing Hamiltonian matrices for 130,831 molecules with their geometries. The stable molecular geometries come from a subset of dataset from QM9

(a) The histogram of the molecule size for QH9-static dataset.

(b) Number of heavy atom percentage in QH9-static dataset.

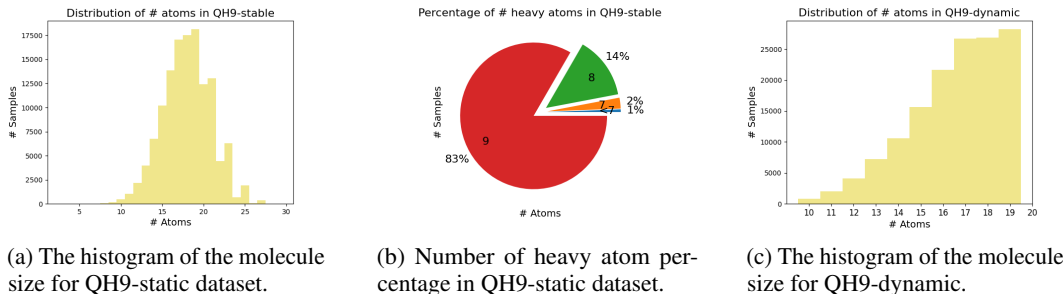(c) The histogram of the molecule size for QH9-dynamic.

Figure 1: The dataset statistics on QH9-stable and QH9-dynamic, including molecule size distribution and percentage of molecules with different number of heavy atoms.

Table 1: The statistics of our defined four tasks.

Task	# Total geometries	# Molecules	# Training/validation/testing geometries
QH-stable-iid	$130,831$	$130,831$	$104,664/13,083/13,084$
QH-stable-ood	$130,831$	$130,831$	$104,001/17,495/9,335$
QH-dynamic-geo	$143,940$	$2,399$	$119,950/11,995/11,995$
QH-dynamic-mol	$143,940$	$2,399$	$115,140/14,340/14,460$

dataset which is widely used in molecular property prediction tasks in Schütt et al. [2018], Gasteiger et al. [2020], Liu et al. [2022], Wang et al. [2022]. The second one is the QH-dynamic dataset, it has molecular trajectories for $2,399$ molecules and each trajectory contains 60 geometries. To obtain the accurate Hamiltonian matrices for this dataset, we set the hyper-parameters of the DFT algorithms to a tight level. Specifically, we set the grid density level to 3 to calculate accurate electronic density, and the SCF convergence condition is set to SCF tolerance of 10^{-13} and gradient threshold of 3.16×10^{-5} to ensure the the final states achieve tight convergence. For the density functional, we select the B3LYP exchange-correlation functional to conduct DFT calculations, and GTO orbital basis Def2SVP is selected to approximate the electronic wavefunctions. To accelerate and achieve the convergence of SCF algorithm, we use DIIS algorithm with consideration of the 8 previous steps. For the QH9-dynamic dataset, molecular dynamics simulations are conducted under the microcanonical ensemble, where the number of particles, volume, and energy remain constant (NVE). The temperature is set to 300K, and the time step for recording the molecular trajectory is set to 2.419×10^{-3} fs.

Dataset Statistics The statistical data, including the number of molecules and geometries for QH9-stable and QH9-dynamic, is presented in Table 1. These molecules consist of no more than 9 heavy atoms and are composed of four specific heavy atoms: carbon (C), nitrogen (N), oxygen (O), and fluorine (F). The distribution of molecule size for QH9-stable and QH9-dynamic is shown in Figure 1a and Figure 1c. Meanwhile, the percentage of molecules in QH9-stable with different the number of heavy atoms is shown in 1b.

3.2 Tasks

To comprehensively evaluate the quantum Hamiltonian prediction performance, we define the following tasks based on the obtained stable and dynamic geometries in the QH9 dataset.

QH-stable-iid. We first randomly divide the obtained stable geometries in QH9 into three subsets, including 80% for training, 10% for validation, and 10% for testing. This serves as the basic evaluation task for predicting quantum Hamiltonian matrices.

QH-stable-ood. We further split the stable geometries in QH9 by molecular size based on the number of constituting atoms. The training set consists of molecules with 3 to 20 atoms, maintaining a similar number of training samples as in the QH-stable-iid split. The validation set includes molecules with 21 to 22 atoms, while the testing set has molecules with 23 to 29 atoms. This task allows for an evaluation of the model's generalization ability under an out-of-distribution training setup.

5

QH-dynamic-geo. For this split and the following QH-dynamic-mol split, there are $2,399$ molecular dynamics trajectories, while each trajectory includes 60 geometries. In QH-dynamic-geo, the split is performed geometry-wise. Specifically, for each molecule, 60 geometries are randomly divided into 50 for training, 5 for validation, and 5 for testing. Here, the molecules in the test set are visible during training but the geometric structures are different from training structures.

QH-dynamic-mol. In this QH-dynamic-mol split, the $2,399$ molecules are divided into training, validation, and testing subsets in a ratio of $0.8/0.1/0.1$. Importantly, different from the above QH-dynamic-geo setup, all 60 geometries corresponding to a specific molecule are grouped together and assigned to the same subset. This setup introduces a more challenging task than QH-dynamic-geo since the geometries in the testing set correspond to different molecules as those in training.

3.3 Methods

To predict the quantum Hamiltonian matrix, several quantum tensor networks have been proposed. SchNorb [Schütt et al., 2019] uses pairwise distance and direction as the input geometric information to predict the final Hamiltonian matrix. However, SchNorb lacks the ability to ensure matrix equivariance and relies on data augmentation techniques to encourage equivariance. Another network, DeepH [Li et al., 2022], uses invariant local coordinate systems and a global coordinate system to handle the equivariance challenge. It uses the geometric features within these invariant local coordinate systems to predict invariant Hamiltonian matrix blocks. Next, as a post-processing step, DeepH applies a rotation using Wigner D-Matrix to transform the Hamiltonian matrix blocks from local coordinate systems back to the global coordinate system. Currently, DeepH is applied on predicting Hamiltonian matrices for materials. PhiSNet [Unke et al., 2021] uses an equivariant model architecture that inherently guarantees matrix equivariance. However, current implementation of PhiSNet is limited to supporting single molecule. This limitation arises from the design of the matrix prediction module in PhiSNet, which is designed to predict matrices for the same molecules with fixed matrix size. Therefore, equivariant quantum tensor network QHNet [Yu et al., 2023] is selected as the main baseline method in the QH9 benchmark currently. QHNet has an extendable expansion module that is built upon intermediate full orbital matrices, enabling its capability to effectively handle different molecules. This flexibility allows QHNet to accommodate various molecules in the QH9 benchmark.

3.4 Metrics

To evaluate the quality of the predicted Hamiltonian matrix, we adopt several metrics that are used to measure both approximation accuracy and computational efficiency.

MAE on Hamiltonian matrix H. This metric calculates the Mean Absolute Error (MAE) between the predicted Hamiltonian matrix and the ground-truth labels from DFT calculation. Each Hamiltonian matrix consists of diagonal blocks and non-diagonal blocks, representing the interactions within individual atoms and the interactions between pairs of atoms, respectively. When the atom pair is distant, the values in the Hamiltonian matrix blocks are typically close to zero. Consequently, as the molecules increase in size, the proportion of distant atom pairs also increases, causing the overall mean value of the Hamiltonian matrix to decrease. Hence, in the subsequent experiments, we compare the MAEs of the diagonal and non-diagonal blocks separately as well as the total MAE on the Hamiltonian matrix.

MAE on occupied orbital energies ϵ. Orbital energy, which includes the Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO) energies, is a highly significant chemical property. It can be determined by diagonalizing the Hamiltonian matrix using Equation 2. Hence, this metric can serve as a measure to reflect the quality of the predicted Hamiltonian matrix in accurately deducing the desired property. Specifically, it calculates the MAE on all the occupied molecular orbital energies ϵ derived from the predicted and the ground-truth Hamiltonian matrix.

Cosine similarity of orbital coefficients ψ. Electronic wavefunctions can describe the quantum states of molecular systems and are used to derive a range of chemical properties. In order to measure the similarity between the ground-truth wavefunctions and the predicted wavefunctions, we calculate the cosine similarity of the coefficients for the occupied molecular orbitals ψ. The corresponding

Table 2: The overall performance on the testing set on the defined four tasks. The unit for the Hamiltonian **H** and eigenenergies ϵ is Hartree denoted by E_h.

| Dataset | Model | **H** $[10^{-6}E_h]\downarrow$ | | | $\epsilon\ [10^{-6}E_h]\downarrow$ | $\psi\ [10^{-2}]\uparrow$ |
		diagonal	non-diagonal	all		
QH-stable-iid	QHNet	130.99	83.95	87.24	2055.53	94.46
QH-stable-ood	QHNet	135.63	80.26	83.22	893.91	92.50
QH-dynamic-geo	QHNet	80.91	75.53	75.89	475.14	96.67
QH-dynamic-mol	QHNet	299.49	112.43	126.69	2208.69	91.27

coefficients **C** are derived from the predicted and ground-truth Hamiltonian matrix shown in Equation 2.

Optimization step ratio. Besides the metrics assessing molecular properties, this optimization step ratio metric is introduced to measure the quality of the predicted Hamiltonian matrix in accelerating DFT calculation. Specifically, it calculates the ratio of the number of optimization steps between initializing with the predicted Hamiltonian matrix and using a random initialization. When the Hamiltonian matrix is accurately predicted, the Self-Consistent Field (SCF) algorithm is close to the convergence condition, leading to a significant reduction in the number of optimization steps.

4 Experiments

Setup. To assess how deep learning approaches perform on the proposed dataset, we conduct experiments on the four designed tasks, as described in Section 3.2. To be more specific, we evaluate the performance of QHNet [Yu et al., 2023], a recently proposed SE(3)-equivariant network specifically designed for efficient and accurate quantum Hamiltonian matrix prediction. QHNet is known for its effectiveness and efficiency in handling the task at hand, making it a suitable testing method for our benchmark evaluation. For quantitative evaluation, we use the metrics as introduced in Section 3.4. Our implementation is based on PyTorch [Paszke et al., 2019], PyG [Fey and Lenssen, 2019], and e3nn [Geiger et al., 2022]. We train models on either (1) a single 48GB Nvidia GeForce RTX A6000 GPU and Intel Xeon Silver 4214R CPU, or (2) a single Nvidia A100 GPU and Intel Xeon Gold 6258R CPU.

Following the model setup in QHNet, in all implemented models, we employ five node-wise interaction layers to aggregate messages from neighboring nodes to update the node irreducible representations. We train all models with a total training step of either $210,000$ or $260,000$ using a batch size of 32. To expedite the convergence of model training, following the QHNet setup, we implement a learning rate scheduler. The scheduler gradually increases the learning rate from 0 to a maximum value of 5×10^{-4} over the first $1,000$ warm-up steps. Subsequently, the scheduler linearly reduces the learning rate, ensuring it reaches 1×10^{-7} by the final step.

Overall performance. We first evaluate the overall performance of the model on the four defined tasks by demonstrating its accuracy of the predicted Hamiltonian matrices on the testing set. As summarized in Table 2, the employed QHNet models can achieve a reasonably low MAE in predicting the Hamiltonian matrices on all proposed tasks. For reference, QHNet can achieve an MAE of $83.12\times10^{-6}E_h$ on the mixed MD17 dataset, which has a similar setup to our QH-dynamic-geo setup. In addition to MAE on Hamiltonian matrices, the trained models also achieve low errors on the predicted occupied orbital energies and orbital coefficients. This aligns with the prior reported work that QHNet is effective to predict the Hamiltonian matrices for multiple molecules [Yu et al., 2023]. Notably, compared to the existing Hamiltonian matrix datasets, such as MD17 [Chmiela et al., 2017] and mixed MD17 [Yu et al., 2023], our proposed tasks involve predicting Hamiltonian matrices for significantly more molecules. Overall, we anticipate that the proposed new datasets and corresponding tasks can serve as more challenging and realistic testbeds for future research in Hamiltonian matrix prediction.

Investigation on out-of-distribution generalization. Since we maintain a similar number of training samples for QH-stable-iid and QH-stable-ood, it is feasible to compare the performance of these two settings to investigate the out-of-distribution challenge in predicting Hamiltonian matrices. It

Table 3: The performance of in-distribution (ID) training and out-of-distribution (OOD) training on the constructed evaluation set for the OOD investigation.

| Training schema | Models | $\mathbf{H}\ [10^{-6}E_h] \downarrow$ | | | $\epsilon\ [10^{-6}E_h] \downarrow$ | $\psi\ [10^{-2}] \uparrow$ |
		diagonal	non-diagonal	all		
ID	QHNet	91.87	59.69	61.42	663.05	94.28
OOD	QHNet	135.58	80.07	83.04	827.65	92.77

is worth noting that we cannot directly compare the performance on their respective test sets, as reported in Table 2, to demonstrate the out-of-distribution generalizability challenge. This is because the molecules in the QH-stable-ood test set have a larger number of atoms on average than those in QH-stable-iid. As explained in Section 3.4, molecules with larger size typically have more distant atom pairs, thus leading to a lower overall mean value of the Hamiltonian matrix. Hence, numerical results on molecules with different sizes are not directly comparable.

To examine the presence of the out-of-distribution issue in the Hamiltonian prediction task, we adopt an alternative evaluation strategy. To be specific, we assess models that have been trained respectively on the QH-stable-iid and QH-stable-ood training sets, employing the same set of samples for evaluation in each instance. Specifically, we use the intersecting set of the QH-stable-iid and QH-stable-ood testing sets as our evaluation dataset. Clearly, the samples contained within this evaluation set are previously unseen during the training phase of either model, thereby maintaining the integrity of the assessment. The evaluation set contains 923 molecules with 23 to 29 atoms. Under this experimental setup, the primary challenge faced by the model trained on the QH-stable-ood training set stems from the novelty of molecular sizes during the evaluation phase. On the other hand, the model trained on the QH-stable-iid training set benefits from having been exposed to such molecular sizes during training. We denote that these two models are trained under out-of-distribution (OOD) and in-distribution (ID) training schema respectively in Table 3. By comparing the performance on the identical evaluation set, it becomes apparent that the model employing the ID training schema outperforms its OOD-trained counterpart, across all metrics including Hamiltonian MAE and predicted orbital energies and coefficients. Such a performance gap demonstrates that the out-of-distribution issue in molecular size is actually a valid concern particularly when extending trained models to molecular sizes not encountered during training.

Geometry-wise *vs.* molecule-wise generalization. We further explore geometry-wise and molecule-wise generalizability by analyzing the difficulty differences between the QH-dynamic-geo and QH-dynamic-mol tasks. We consider the results in Table 2 for these two tasks to be comparable given that both models are trained with a similar number of geometry structures. We note that the model in the QH-dynamic-geo task demonstrates numerically better test performance than the model in the QH-dynamic-mol task. This is consistent with our intention when designing the tasks. Specifically, in QH-dynamic-geo, although the geometric structures in the test set are different, the molecules themselves are not entirely novel to the model due to the exposure during the training phase. In comparison, the QH-dynamic-mol task presents a more challenging and demanding scenario. In particular, the test set geometries in QH-dynamic-mol correspond to entirely different molecules than those seen during training. This task requires the model to generalize from its learned patterns to the unseen molecular structures. To summarize, both tasks serve as valuable testbeds in evaluating the model's generalization ability, and our analysis shows that QH-dynamic-mol task, which requires extrapolating to entirely new molecular structures, is notably more challenging and demanding.

Accelerating the DFT calculation. We further measure the quality of the predicted Hamiltonian matrix by evaluating its ability in accelerating the DFT calculation. As introduced in Section 3.4, we compute the ratio of optimization steps required when initializing with the predicted Hamiltonian matrix as compared to a random initialization. In this experiment, following our data collection process, we use PySCF [Sun et al., 2018] to perform the DFT calculation

Table 4: The performance of accelerating the DFT calculation.

Dataset	Models	Optimization step ratio $\downarrow$
QH-stable-iid	QHNet	0.718
QH-stable-ood	QHNet	0.723
QH-dynamic-geo	QHNet	0.674
QH-dynamic-mol	QHNet	0.733

with using B3LYP exchange-correlation functional and def2SVP basis set. We select DIIS as the SCF algorithm for the DFT calculation and set a grid density level of 3 to ensure an accurate DFT calculation. For each dataset, we compute the average optimization step ratio for 50 randomly selected molecules. As shown in Table 4, when initializing from the predicted Hamiltonian matrices given by QHNet, it requires fewer optimization steps to reach the converged Hamiltonian matrix, which indicates that the predicted Hamiltonian matrix is close to the convergence condition. This experimental result demonstrates that machine learning approaches are helpful in accelerating the DFT calculation.

5 Conclusion

We are interested in accelerating computation of quantum Hamiltonian matrices, which fundamentally determine the quantum states of physical systems and chemical properties. While various invariant and equivariant deep learning methods have been developed recently, current quantum Hamiltonian datasets consist of Hamiltonian matrices of molecular dynamic trajectories for only a single and four molecules, respectively. To significantly expand the size and variety of such datasets, we generate a much larger dataset based on the QM9 molecules. Our dataset provides precise Hamiltonian matrices for 130,831 stable molecular geometries and 2,399 molecular dynamics trajectories with 60 geometries in each trajectory. Extensive and carefully designed experiments are conducted to demonstrate the quality of our generated data.

Acknowledgements

References

Brandon Anderson, Truong Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. *Advances in neural information processing systems*, 32, 2019.

Hexin Bai, Peng Chu, Jeng-Yuan Tsai, Nathan Wilson, Xiaofeng Qian, Qimin Yan, and Haibin Ling. Graph neural network for Hamiltonian-based material property prediction. *Neural Computing and Applications*, 34:4625–4632, 2022.

Ilyes Batatia, David Peter Kovacs, Gregor N. C. Simm, Christoph Ortner, and Gabor Csanyi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=YPpSngE-ZU.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, 2022.

Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 04 1993. ISSN 0021-9606. doi: 10.1063/1.464913. URL https://doi.org/10.1063/1.464913.

Lorenz C Blum and Jean-Louis Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database gdb-13. *Journal of the American Chemical Society*, 131(25): 8732–8733, 2009.

Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J Bekkers, and Max Welling. Geometric and physical quantities improve e(3) equivariant message passing. In *International Conference on Learning Representations*, 2022.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021.

Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5):e1603015, 2017.

Stefan Chmiela, Valentin Vassilev-Galindo, Oliver T Unke, Adil Kabylda, Huziel E Sauceda, Alexandre Tkatchenko, and Klaus-Robert Müller. Accurate global machine learning force fields for molecules with hundreds of atoms. *Science Advances*, 9(2):eadf0873, 2023.

Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, 99(1):014104, 2019.

Genevieve Dusson, Markus Bachmayr, Gábor Csányi, Ralf Drautz, Simon Etter, Cas van der Oord, and Christoph Ortner. Atomic cluster expansion: Completeness, efficiency and stability. *Journal of Computational Physics*, 454:110946, 2022.

Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv Preprint, arXiv:1903.02428*, 2019.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3D roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33:1970–1981, 2020.

Michael Gastegger, Adam McSloy, M Luya, Kristof T Schütt, and Reinhard J Maurer. A deep neural network for molecular wave functions in quasi-atomic minimal basis representation. *The Journal of Chemical Physics*, 153(4):044123, 2020.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. GemNet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34: 6790–6802, 2021.

Mario Geiger, Tess Smidt, Alby M., Benjamin Kurt Miller, Wouter Boomsma, Bradley Dice, Kostiantyn Lapchevskyi, Maurice Weiler, Michał Tyszkiewicz, Simon Batzner, Dylan Madisetti, Martin Uhrin, Jes Frellsen, Nuri Jung, Sophia Sanborn, Mingjian Wen, Josh Rackers, Marcel Rød, and Michael Bailey. Euclidean neural networks: e3nn, April 2022. URL https://doi.org/10.5281/zenodo.6459381.

Xiaoxun Gong, He Li, Nianlong Zou, Runzhang Xu, Wenhui Duan, and Yong Xu. General framework for E(3)-equivariant neural network representation of density functional theory Hamiltonian. *Nature Communications*, 14(1):2848, 2023.

P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964. doi: 10.1103/PhysRev.136.B864. URL https://link.aps.org/doi/10.1103/PhysRev.136.B864.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research*, 47(D1):D1102–D1109, 2019.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research*, 49(D1):D1388–D1395, 2021.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. PubChem 2023 update. *Nucleic Acids Research*, 51 (D1):D1373–D1380, 2023.

W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965. doi: 10.1103/PhysRev.140.A1133. URL https://link.aps.org/doi/10.1103/PhysRev.140.A1133.

Dávid Péter Kovács, Cas van der Oord, Jiri Kucera, Alice EA Allen, Daniel J Cole, Christoph Ortner, and Gábor Csányi. Linear Atomic Cluster Expansion Force Fields for Organic Molecules: Beyond RMSE. *Journal of Chemical Theory and Computation*, 17(12):7696–7711, 2021.

Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785–789, Jan 1988. doi: 10.1103/PhysRevB.37.785. URL https://link.aps.org/doi/10.1103/PhysRevB.37.785.

He Li, Zun Wang, Nianlong Zou, Meng Ye, Runzhang Xu, Xiaoxun Gong, Wenhui Duan, and Yong Xu. Deep-learning density functional theory hamiltonian for efficient ab initio electronic-structure calculation. *Nature Computational Science*, 2(6):367–377, 2022.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KwmPfARgOTD.

Yi Liu, Limei Wang, Meng Liu, Yuchao Lin, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3D molecular graphs. In *International Conference on Learning Representations*, 2022.

Nicola Marzari and David Vanderbilt. Maximally localized generalized Wannier functions for composite energy bands. *Phys. Rev. B*, 56:12847–12865, 1997.

Nicola Marzari, Arash A. Mostofi, Jonathan R. Yates, Ivo Souza, and David Vanderbilt. Maximally localized Wannier functions: Theory and applications. *Rev. Mod. Phys.*, 84:1419–1475, 2012.

Maho Nakata and Tomomi Shimazaki. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *Journal of Chemical Information and Modeling*, 57(6):1300–1308, 2017.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Péter Pulay. Convergence acceleration of iterative sequences. the case of scf iteration. *Chemical Physics Letters*, 73(2):393–398, 1980. ISSN 0009-2614. doi: https://doi.org/10.1016/0009-2614(80)80396-4. URL https://www.sciencedirect.com/science/article/pii/0009261480803964.

Péter Pulay. Improved scf convergence acceleration. *Journal of Computational Chemistry*, 3(4): 556–560, 1982. doi: https://doi.org/10.1002/jcc.540030413. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.540030413.

Xiaofeng Qian, Ju Li, and Sidney Yip. Calculating phase-coherent quantum transport in nanoelectronics with *ab initio* quasiatomic orbital basis set. *Phys. Rev. B*, 82:195442, 2010.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):1–7, 2014.

Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning*, pages 9323–9332. PMLR, 2021.

Erwin Schrödinger. An undulatory theory of the mechanics of atoms and molecules. *Physical Review*, 28(6):1049, 1926.

Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021.

Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. SchNet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

Kristof T Schütt, Michael Gastegger, Alexandre Tkatchenko, K-R Müller, and Reinhard J Maurer. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature Communications*, 10(1):5024, 2019.

Ivo Souza, Nicola Marzari, and David Vanderbilt. Maximally localized Wannier functions for entangled energy bands. *Phys. Rev. B*, 65:035109, 2001.

Qiming Sun, Timothy C Berkelbach, Nick S Blunt, George H Booth, Sheng Guo, Zhendong Li, Junzi Liu, James D McClain, Elvira R Sayfutyarova, Sandeep Sharma, et al. PySCF: the Python-based simulations of chemistry framework. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(1):e1340, 2018. doi: https://doi.org/10.1002/wcms.1340.

Qiming Sun, Xing Zhang, Samragni Banerjee, Peng Bao, Marc Barbry, Nick S Blunt, Nikolay A Bogdanov, George H Booth, Jia Chen, Zhi-Hao Cui, et al. Recent developments in the PySCF program package. *The Journal of Chemical Physics*, 153(2), 2020. doi: https://doi.org/10.1063/5.0006074.

Attila Szabo and Neil S Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Courier Corporation, 2012.

Philipp Thölke and Gianni De Fabritiis. Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations*, 2022.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3D point clouds. *arXiv Preprint, arXiv:1802.08219*, 2018.

Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, et al. The open catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, 2023.

Oliver Unke, Mihail Bogojeski, Michael Gastegger, Mario Geiger, Tess Smidt, and Klaus-Robert Müller. SE(3)-equivariant prediction of molecular wavefunctions and electronic densities. *Advances in Neural Information Processing Systems*, 34:14434–14447, 2021.

Limei Wang, Yi Liu, Yuchao Lin, Haoran Liu, and Shuiwang Ji. ComENet: Towards complete and efficient message passing for 3D molecular graphs. In *The 36th Annual Conference on Neural Information Processing Systems*, pages 650–664, 2022.

Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, 37(suppl_2):W623–W633, 2009.

Yanli Wang, Stephen H Bryant, Tiejun Cheng, Jiyao Wang, Asta Gindulyte, Benjamin A Shoemaker, Paul A Thiessen, Siqian He, and Jian Zhang. PubChem bioassay: 2017 update. *Nucleic Acids Research*, 45(D1):D955–D963, 2017.

Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 2005.

Haiyang Yu, Zhao Xu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Efficient and equivariant graph networks for predicting quantum Hamiltonian. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.