# UNiTE: Unitary N-body Tensor Equivariant Network with Applications to Quantum Chemistry

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Equivariant neural networks have been successful in incorporating various types of symmetries, but are mostly limited to vector representations of geometric objects. Despite the prevalence of higher-order tensors in various application domains, e.g. in quantum chemistry, equivariant neural networks for general tensors remain unexplored. Previous strategies for learning equivariant functions on tensors mostly rely on expensive tensor factorization which is not scalable when the dimensionality of the problem becomes large. In this work, we propose unitary $N$-body tensor equivariant neural network (UNiTE), an architecture for a general class of symmetric tensors called $N$-body tensors. The proposed neural network is equivariant with respect to the actions of a unitary group, such as the group of 3D rotations. Furthermore, it has a linear time complexity with respect to the number of non-zero elements in the tensor. We also introduce a normalization method, viz., Equivariant Normalization, to improve generalization of the neural network while preserving symmetry. When applied to quantum chemistry, UNiTE in combination with a physics-based molecular representation outperforms all state-of-the-art machine learning methods on multiple benchmarks. Finally, we show that UNiTE achieves a robust zero-shot generalization performance on diverse down stream chemistry tasks, while being three orders of magnitude faster than conventional numerical methods with competitive accuracy.

## 1 Introduction

Geometric deep learning is focused on building neural network models for geometric objects, and it needs to encode the symmetries present in the problem domain [1]. A geometric object is usually represented using a reference frame input to the neural network model. Symmetries are incorporated via the concept of *equivariance* defined as the property of being independent of the choice of reference frame.

One intuitive and common way to encode a geometric object is to represent it as the positions of a collection of points, i.e. a set of vectors. Examples include point clouds [2], grids [3] and meshes [4]. Many previous geometric learning methods, termed *equivariant neural networks*, have been designed by considering how the vectors transform under symmetry operations on the reference frames. These equivariant neural networks have successfully 'baked' symmetries into deep neural networks in various application domains, such as autonomous driving [5] and molecular design [6].

However, we identify two remaining challenges that are not addressed in prior works:

(a) Prior works mostly focus on vector (i.e. order-1 tensor) representations of a geometric entity, but constructing equivariant neural networks for general tensors is largely unexplored;

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

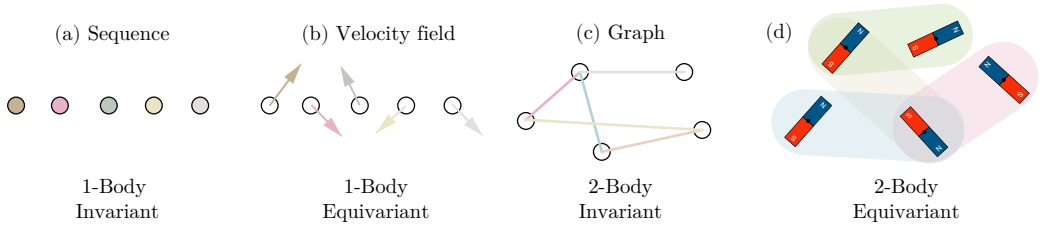


Figure 1: Examples of $N$-body tensors.

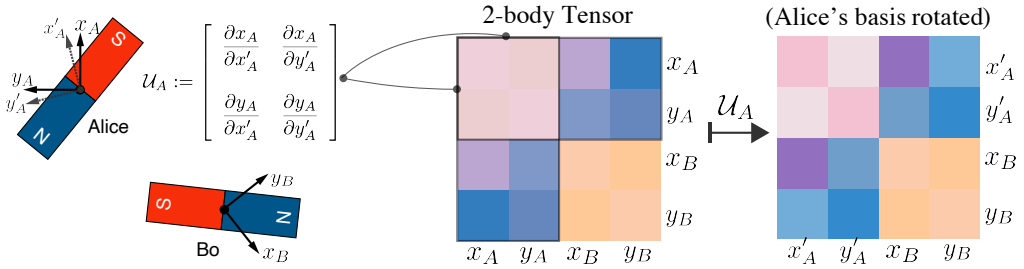


Figure 2: Illustrating an $N$-body tensor with $N = 2$. Imagine Alice and Bo are doing experiments with two bar magnets without knowing each other's reference frame. The magnetic interactions depend on both bar magnets' orientations and can be written as a 2-body tensor. When Alice make a rotation on her reference frame, sub-tensors containing index $A$ are transformed by a unitary matrix $\mathcal{U}_A$, giving rise to the 2-body tensor coefficients in the transformed basis. We design neural network to be equivariant to all such local basis transformations.

(b) They are often specific to one form of data (e.g., point clouds or grids) or one class of groups, while lacking a more generic framework.

Higher-order tensors are ubiquitous, and many famous problems in physical science, e.g. general relativity, are defined using tensors of order larger than one. However, it is challenging to directly apply equivariant neural networks for learning on those tensors. First, the order of learning targets of interest, mostly scalars (order-0) or vectors (order-1), usually differ from the order of input tensors and cannot be handled in those frameworks. Second, many equivariant neural networks rely on computing tensor products of vectors and their decompositions in the neural network building blocks [2, 7], which is not applicable to cases where the inputs are tensors. A few works designed for learning equivariant functions on tensors in specific domains rely on tensor factorization [8], which is not scalable when the order of tensors or the dimension of the physical spaces become large.

The absence of a more generic framework may limit their applicability when symmetries in the learning problem become more complicated. For example, in particle physics, each particle is linked with a different symmetry group [9]. Without a unified approach one has to exhaustively formulate equivariant neural networks for every instance of such systems.

In this work, we are interested in general tensors that encode *relations* among multiple geometric objects. We define a class of tensors $\mathbf{T}$, which we call *N-body tensors*, that can describe such $N$-object relations. Many forms of data can be interpreted as examples of $N$-body tensors (Figure 1). For example, sequences are scalar-valued features concatenated together to form a flattened array which is an order-1 tensor (e.g. word embeddings). They do not contain any positional information in physical space, therefore their values are trivially unchanged when rotating the reference frame. Thus sequences are classified as 1-body invariant tensors. Velocity fields can be understood as sequences with an extra directional information. When rotating the reference frame, the values recorded on $xyz$-components of each sub-vector changes equivariantly. Therefore velocity fields are interpreted as 1-body equivariant tensors. Graphs encode relational information between a set of nodes, defined

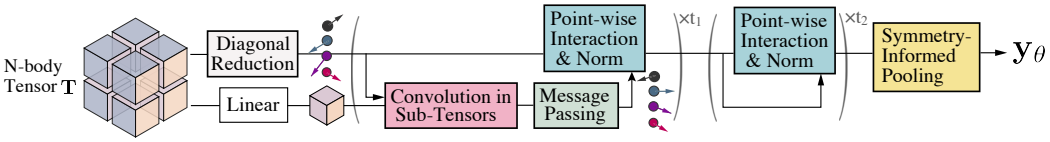


Figure 3: The UNiTE model architecture. UNiTE first initialize order-1 (vector) representations via a diagonal sub-tensor reduction layer (Section 4.2), then updates representations through convolution, message-passing (Section 4.1), and point-wise interaction blocks with Equivariant Normalization (Section 4.3). A symmetry-informed pooling is used at the end to readout the predictions $\mathbf{y}_\theta$.

by their adjacency matrices (order-2 tensors). Since the adjacency matrix of a graph only contains scalar values, graphs are also rotation-invariant and are 2-body invariant tensors.

To further motivate towards the general $N$-body case, we can consider a toy model where many bar magnets are placed on a table (Figure 1d). The magnetic interaction between a pair of bar magnets (say, Alice and Bo) in fact cannot be simply expressed by the magnets' orientation vectors, but need to be written as a $2 \times 2$ matrix dependent on Alice's and Bo's reference frames. Values on the matrix are rotationally-equivariant because it columns and rows are transformed by a rotation matrix as Alice or Bo rotates their reference frames (Figure 2); hence, this is an example of 2-body equivariant tensor. Given $M$ bar magnets, the 2-body tensor representing the system depicted in Figure 1d is thus a $(2M \times 2M)$ square matrix stacked from such $2 \times 2$ matrices, which may be also viewed as a graph of those order-2 sub-tensors (matrices).

We propose UNiTE for the general case of $N$-body equivariant tensors where both $N$-object relations and physical-space information are present. **Our contributions are:**

- We present Unitary $N$-body Tensor Equivariant neural network (UNiTE), a novel architecture for $N$-body tensors of any $N$ in arbitrary-dimensional physical spaces. It is equivariant with respect to unitary transformations on the reference frames and tensor index permutations.
- The proposed approach realizes equivariance without requiring explicit tensor factorization operations. It is efficient, having a linear time complexity with respect to the number of non-zero elements in the tensor.
- Generalization and training stability of equivariant neural networks is improved using a simple but effective normalization scheme, Equivariant Normalization (EvNorm), proposed in this paper.

UNiTE has a modular architecture (Figure 3) that updates a set of latent order-1 (i.e. vectors) representations: $\mathbf{h}^{t=0} \mapsto \mathbf{h}^{t=1} \mapsto \cdots \mapsto \mathbf{h}^{t=t_f}$ and performs pooling at the end. The vector representations are defined on each geometric object $\mathbf{h}^t := [\mathbf{h}_1^t, \mathbf{h}_2^t, \cdots, \mathbf{h}_d^t]$.

The initial vector representations $\mathbf{h}^{t=0}$ are generated by decomposing diagonal sub-tensors of the $N$-body tensor into vectors without explicitly solving tensor factorization, based on a theoretical result in this work which is connected to the Wigner-Eckart Theorem in quantum physics [10]. We now give some intuitive explanations for our theoretical results: the diagonal sub-tensors can be viewed as isolated systems because they can be modeled as being infinitely apart and only interacting with an external field. The Wigner-Eckart Theorem describes rotational symmetry for such isolated systems, and we generalize it to $N$-body tensors in our work.

Each update step $\mathbf{h}^t \mapsto \mathbf{h}^{t+1}$ is composed of (a) Convolution, (b) Message passing, and (c) Point-wise interaction blocks. All such blocks are designed to be equivariant with respect to index permutations and both globally-applied and locally-applied reference-frame transformations, by (a) contracting the $N$-body tensor with vector representations $\mathbf{h}^t$ to make the contracted tensor dimensions invariant to reference-frame transformations, and (b) designing neural network layers on $\mathbf{h}^t$ that preserve its transformation rule under rotations and reflections. Thus, we have an end-to-end equivariant neural network.

In an update step $\mathbf{h}^t \mapsto \mathbf{h}^{t+1}$, each sub-tensor of the $N$-body tensor input $\mathbf{T}$ is first contracted with products of the vector representations $\mathbf{h}^t$. This tensor contraction operation can be interpreted as performing an $(N-1)$ dimensional convolution using the sub-tensors of $\mathbf{T}$ as convolution kernels, therefore we call it *convolution-in-sub-tensors* block. The convolution outputs from each sub-tensor

is an order-1 tensor (i.e. vectors); those convolution outputs are then passed into a *message-passing* block, which is analogous to a message-passing operation on edges in graph neural networks (GNNs) but are performed on edges connecting $N$ nodes. The outputs are then fed into a *point-wise interaction* block with the previous-step representation $\mathbf{h}^t$ to complete the update $\mathbf{h}^t \mapsto \mathbf{h}^{t+1}$. The point-wise interaction blocks are constructed as a stack of multi-layer perceptrons (MLPs), vector outer-products and skip connections. Within those blocks, a *matching layer* is used to 'glue' the basis of $\mathbf{h}^t$ with the basis used to define the tensor $\mathbf{T}$. Presence of the matching layer ensures equivariance is maintained.".

In addition, we propose a novel normalization layer, Equivariant Normalization (EvNorm). EvNorm normalizes the scales of vectors in the order-1 (i.e. vectors) representation, while recording those vectors' directional information to be recovered after normalization. Therefore it performs normalization without sacrificing model equivariance. EvNorm can be feasibly integrated within a point-wise interaction block through first applying EvNorm on the input vector representations, then using an MLP to the array of normalized vector scales, and finally multiplying the vector directions recorded by EvNorm to the MLP's output. In practice, we found using EvNorm within the point-wise interaction blocks greatly stabilizes training, improves model generalization, and eliminates the need for hand-tuning weight initializations and learning rates across different tasks.

In UNiTE, the only operation explicitly performed on the tensor input is the $(N-1)$ dimensional convolution, which has a linear time complexity with respect to the number of non-zero elements in the $N$-body tensor. The scalability of the architecture with respect to the number of physical interactions in the learning task is therefore ensured.

**Quantum chemistry with UNiTE.** One main motivation for us to develop UNiTE is to enable learning quantum chemistry properties based on 2-body and higher-order tensor representations of molecules. Such $N$-body tensor molecular representations are necessary to encode the interactions between electrons and atoms, since their motions are quantum-mechanical and are described by high-dimensional functions. Furthermore, the chemical properties predicted by the model need to satisfy their symmetry constraints. When applied to quantum chemistry, empirically we find UNiTE in combination with a quantum-mechanical-approximation-based molecular representation showing superior performance over all existing machine learning approaches for that domain which were only capable of utilizing molecular geometry inputs. On average, it outperforms the state-of-the-art methods by **150%** on QM9, **114%** on MD17 and **50-75%** on electron densities.

Beyond data efficiency on benchmarks in train-test split settings, a more practical and challenging aspect is whether the neural network can deduce down-stream properties of humans' interest once being trained on primitive physical quantities. This aspect is even more critical in the domain of quantum chemistry, where all chemical reactions are fundamentally related to the energies of molecules through the famous Schrödinger equation, but directly obtaining down-stream properties about chemical reactions by solving the Schrödinger equation is both theoretically and computationally prohibitive. In this work, we find UNiTE model pre-trained on energies of 236k molecules achieves robust performance on various practical, down stream chemistry tasks without any model fine-tuning. Even in such a zero-shot setting, it offers an accuracy similar or better than conventional numerical methods with up to **3-orders-of-magnitude** speedup.

## 2 Related works

**Equivariant neural networks.** Equivariant neural networks were first introduced for homogeneous grid and point cloud data[2, 3, 7, 11], and have been generalized to symmetries on non-Euclidean manifolds[4, 12, 13]. While being powerful, those architectures are typically designed based on order-1 tensors in the geometry. In contrast, our goal is to develop equivariant neural network operating on a $N$-body tensors for geometric data that may entail higher-order relational information.

**Graph neural networks.** Graph neural networks (GNNs) are gaining popularity for learning on relational data [14, 15]. They are permutation-equivariant with respect to reordering the node indices in the graph. Recent works have extended graph neural networks to hyper-graphs [16, 17], as well as to equivariance under continuous symmetry transformations [11, 18] when the set of nodes in the graph correspond to point clouds in the Euclidean space. GNNs on point clouds primarily focused on the global 3D rotational symmetry, and achieve rotation-equivariance using either a set of harmonic

basis functions [2, 19] or the simpler standard basis [18, 20]. Some GNNs have also been developed with special interests in N-body physical simulation data or quantum chemistry [19–26].

**Equivariance for order-$N$ tensors.** To our best knowledge, closest to ours in spirit is a pair of recent works [8, 27] developed for a special class of order-2 $\mathrm{SU(n)}$ tensors in a high-energy physics problem. They proposed to apply learnable equivariant transformations on those order-2 tensors through eigen-decomposition, with a symmetrization algorithm eigenvalues to enforce permutation equivariance. But in a general order-$N$ setting, such an approach requires performing tensor factorization for each order-$N$ sub-tensor, which would introduce significant computational costs and can be intractable when $N$ or $n$ become large.

# 3 $N$-body tensors

We are interested in a class of tensors $\mathbf{T}$, for which each sub-tensor $\mathbf{T}(u_1, u_2, \cdots, u_N)$ describes relation among a collection of $N$ geometric objects defined in an $n$-dimensional physical space. For simplicity, we will introduce the tensors of interest using a special case based on point clouds embedded in the $n$-dimensional Euclidean space, associating a (possibly different) set of orthogonal basis with each point's neighbourhood. In this setting, our main focus is the change of the order-N tensor's coefficients when applying $n$-dimensional rotations and reflections to the local reference frames. Our proposed approach can be generalized to non-flat manifolds, harmonic basis and complex fields, and a general problem statement will be provided in Appendix A.2.

**Definition 1 ($N$-body tensor).** Let $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_d\}$ be $d$ points in $\mathbb{R}^n$ for each $u \in \{1, 2, \cdots, d\}$. For each point index $u$, we define an orthonormal basis (local reference frame) $\{\mathbf{e}_{u;v_u}\}$ centered at $\mathbf{x}_u$[1], and denote the space spanned by the basis as $V_u := \mathrm{span}(\{\mathbf{e}_{u;v_u}\}) \subseteq \mathbb{R}^n$. We consider a tensor $\hat{\mathbf{T}}$ defined via $N$-th direct products of the 'concatenated' basis $\{\mathbf{e}_{u;v_u}; (u, v_u)\}$:

$$\hat{\mathbf{T}} := \sum_{\vec{u},\vec{v}} T\big((u_1; v_1), (u_2; v_2), \cdots, (u_N; v_N)\big)\, \mathbf{e}_{u_1;v_1} \otimes \mathbf{e}_{u_2;v_2} \otimes \cdots \otimes \mathbf{e}_{u_N;v_N} \tag{1}$$

$\hat{\mathbf{T}}$ is a tensor of order-$N$ and is an element of $(\bigoplus_{u=1}^{d} V_u)^{\otimes N}$. We call its coefficients $\mathbf{T}$ an *N-body tensor* if $\mathbf{T}$ is invariant to global translations ($\forall\, \mathbf{x}_0 \in \mathbb{R}^n, \mathbf{T}[\mathbf{x}] = \mathbf{T}[\mathbf{x} + \mathbf{x}_0]$, and is symmetric:

$$T\big((u_1; v_1), (u_2; v_2), \cdots, (u_N; v_N)\big) = T\big((u_{\sigma_1}; v_{\sigma_1}), (u_{\sigma_2}; v_{\sigma_2}), \cdots, (u_{\sigma_N}; v_{\sigma_N})\big) \tag{2}$$

where $\sigma$ denotes arbitrary permutation on its dimensions $\{1, 2, \cdots, N\}$. Note that each sub-tensor, $\mathbf{T}_{\vec{u}}$, does not have to be symmetric. The shorthand notation $\vec{u} := (u_1, u_2, \cdots, u_N)$ indicates a subset of $N$ points in $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_d\}$ which then identifies a sub-tensor[2] $\mathbf{T}_{\vec{u}} := \mathbf{T}(u_1, u_2, \cdots, u_N)$ in the $N$-body tensor $\mathbf{T}$; $\vec{v} := (v_1, v_2, \cdots, v_N)$ index a coefficient $T_{\vec{u}}(v_1, v_2, \cdots, v_N) := T\big((u_1; v_1), (u_2; v_2), \cdots, (u_N; v_N)\big)$ in a sub-tensor $\mathbf{T}_{\vec{u}}$, where each index $v_j \in \{1, 2, \cdots, \dim(V_{u_j})\}$ for $j \in \{1, 2, \cdots, N\}$.

We aim to build neural networks $\hat{\mathcal{F}}_\theta : (\bigoplus_{u=1}^{d} V_u)^{\otimes N} \to \mathcal{Y}$ that map $\hat{\mathbf{T}}$ to order-1 tensor- or scalar-valued outputs $\mathbf{y} \in \mathcal{Y}$. While $\hat{\mathbf{T}}$ is independent of the choice of local reference frame $\mathbf{e}_u$, its coefficients $\mathbf{T}$ (i.e. the $N$-body tensor) vary when rotating or reflecting the basis $\mathbf{e}_u := \{\mathbf{e}_{u;v_u}; v_u\}$, i.e. acted by an element $\mathcal{U}_u \in \mathrm{O(n)}$. Therefore, the neural network $\hat{\mathcal{F}}_\theta$ should be constructed equivariant with respect to those reference frame transformations.

**Equivariance.** For a map $f : \mathcal{V} \to \mathcal{Y}$ and a group $G$, $f$ is said to be $G$-equivariant if for all $g \in G$ and $\mathbf{v} \in \mathcal{V}$, $g \cdot f(\mathbf{v}) = f(g \cdot \mathbf{v})$. In our case, the group $G$ is composed of (a) Unitary transformations $\mathcal{U}_u$ locally applied to basis: $\mathbf{e}_u \mapsto \mathcal{U}_u^\dagger \cdot \mathbf{e}_u$, which are rotations and reflections for $\mathbb{R}^n$. $\mathcal{U}_u$ induces transformations on tensor coefficients: $\mathbf{T}_{\vec{u}} \mapsto (\mathcal{U}_{u_1} \otimes \mathcal{U}_{u_2} \otimes \cdots \otimes \mathcal{U}_{u_N})\mathbf{T}_{\vec{u}}$, and an intuitive example

[1] We additionally allow for $\mathbf{0} \in \{\mathbf{e}_{u;v_u}\}$ to represent features in $\mathbf{T}$ that transform as scalars.

[2] For example, if there are $d = 5$ points defined in the 3-dimensional Euclidean space $\mathbb{R}^3$ and each point is associated with a standard basis $(x, y, z)$, then for the example of N=4, there are $5^4$ sub-tensors and each sub-tensor $\mathbf{T}_{(u_1,u_2,u_3,u_4)}$ contains $3^4 = 81$ elements with indices spanning from $xxxx$ to $zzzz$. In total, $\mathbf{T}$ contains $(5 \times 3)^4$ coefficients. The coefficients of $\mathbf{T}$ are in general complex-valued as formally discussed in Definition A1, but are real-valued for the special case introduced in Definition 1 .

for infinitesimal basis rotations in $N = 2, n = 2$ is shown in Figure 2; (b) Tensor index permutations: $(\vec{u}, \vec{v}) \mapsto \sigma(\vec{u}, \vec{v})$; (c) Global translations: $\mathbf{x} \mapsto \mathbf{x} + \mathbf{x}_0$. For conciseness, we borrow the term $G$-equivariance to say $\hat{\mathcal{F}}_\theta$ is equivariant to all the symmetry transformations listed above.

To reiterate, **our goal** is to propose neural networks $\hat{\mathcal{F}}_\theta$ that are not only $G$-equivariant, but also scalable for general values of $N$ and $n$, and efficient for implementation and training in practice.

# 4 Unitary $N$-body tensor equivariant neural network (UNiTE)

We propose our Unitary $N$-body Tensor Equivariant (UNiTE) neural network architecture (Figure 3). Given an input $N$-body tensor, UNiTE first performs an efficient tensor order reduction on diagonal sub-tensors to generate a set of representations $\mathbf{h}^{t=0}$. Then it updates the representations with $t_1$ stacks of convolution, message-passing, and point-wise interaction blocks, followed by $t_2$ stacks of point-wise self interaction blocks. Finally, a pooling operation is applied to the final representations $\mathbf{h}^{t=t_f}$ at $t_f = t_1 + t_2$ to readout the predictions.

## 4.1 Convolutions and message passing on $\mathbf{T}$

**Convolution in sub-tensors.** In an update step $\mathbf{h}^t \mapsto \mathbf{h}^{t+1}$, sub-tensors of $\mathbf{T}$ are contracted with products of the order-1 (i.e. vectors) representations $\mathbf{h}^t$:

$$(\mathbf{m}_{\vec{u}}^t)_{v_1}^i = \sum_{v_2,\cdots,v_N} T_{\vec{u}}(v_1, v_2, \cdots, v_N) \prod_{j=2}^{N} \left(\rho_{u_j}(\mathbf{h}_{u_j}^t)\right)_{v_j}^i \tag{3}$$

which can be viewed as a $(N-1)$-dimensional convolution operation between each sub-tensor $\mathbf{T}_{\vec{u}}$ (as convolution kernels) and the signal $\prod_{j=2}^{N} \mathbf{h}_{u_j}^t$ in the $i$-th convolution channel. This $(N-1)$-dimensional convolution gives an order-1 tensor output $\mathbf{m}_{\vec{u}}^t$ for each sub-tensor index $\vec{u}$. $\rho_u$ is called a *matching layer* at index $u$, which will be defined later.

**Message-passing on convolution outputs.** $\mathbf{m}_{\vec{u}}^t$ are then aggregated into each index $u$ by summing over the indices $u_2, u_3, \cdots, u_N$, analogous to a 'message-passing' between nodes and edges in common realizations of graph neural networks [15]. We define the following equivariant message passing scheme on the (hyper-)graph defined by the the set of indices $\vec{u}$ where $\mathbf{T}_{\vec{u}}$ is non-zero:

$$\tilde{\mathbf{m}}_{u_1}^t = \sum_{u_2,u_3,\cdots,u_N} \bigoplus_i (\mathbf{m}_{\vec{u}}^t)^i \cdot \alpha_{\vec{u}}^t \tag{4}$$

$$\mathbf{h}_{u_1}^{t+1} = \phi\big(\mathbf{h}_{u_1}^t, \rho_{u_1}^\dagger(\tilde{\mathbf{m}}_{u_1}^t)\big) \tag{5}$$

where $\alpha_{\vec{u}}^t$ are reference-frame-invariant, scalar-valued weights for improving the network capacity, and their practical parameterizations are discussed in Appendix A.8. In Equation 5, the aggregated equivariant messages $\tilde{\mathbf{m}}_u^t$ are interacted with $\mathbf{h}_u^t$ through an point-wise interaction block $\phi(\cdot,\cdot)$ to complete the update $\mathbf{h}_u^t \mapsto \mathbf{h}_u^{t+1}$, which we elaborate in Section 4.3. Equations 3-5 form the backbone of UNiTE, which can be shown to satisfy $G$-equivariance (see Appendix A.2 for proofs).

## 4.2 Embedding through diagonal sub-tensor reduction

Here we take a step backward and discuss the construction of initial order-1 representations, $\mathbf{h}^{t=0}$. We note that in general those order-1 embeddings can be *uniquely* extracted from the diagonal sub-tensors of the $N$-body tensor $\mathbf{T}$ based on group representation theory. We use the short-hand notation $\mathbf{T}_u$ to denote the $u$-th diagonal sub-tensors of $\mathbf{T}$, i.e. $\{\mathbf{T}_u\} := \{\mathbf{T}_{\vec{u}}; u_1 = u_2 = \cdots = u_N = u\}$.

**Theorem 1 (Diagonal sub-tensors of an $N$-body tensor are reducible)** *(Informal) $\mathbf{T}_u$ can be uniquely expressed as a linear combination of order-1 tensors (i.e. vectors). Such a linear combination coefficients do not depend on $\mathbf{T}$. Those order-1 tensors transform under the irreducible representations of the group $\mathrm{O(n)}$ when rotating or reflecting the reference frames.*

The above theorem is a consequence of group representation properties of $\mathrm{O(n)}$ and realizing that $\mathbf{T}_u$ has a symmetric tensor decomposition which allows for identifying a special case of the Schur–Weyl duality. See Appendix A.2 for a formal statement and proof. Based upon Theorem 1, we can obtain a practically useful result for generating the desired order-1 embeddings $\mathbf{h}_u^{t=0}$ from $\mathbf{T}_u$:

**Lemma 1** *Let $(\boldsymbol{\pi}_l)_m$ denote the $m$-th basis component for the $l$-th irreducible representation of $\mathrm{O(n)}$. For each allowed $l, m$, there exist $K_{l,m} \times n^N$ $\mathbf{T}$-independent scalar coefficients $Q^{\vec{v}}_{klm}$ parameterizing a linear transformation $\psi$ that performs $\mathbf{T}_{\vec{u}} \mapsto \mathbf{h}_u$, if $u_1 = u_2 = \cdots = u_N = u$:*

$$\big(\psi(\mathbf{T}_u)\big)_{klm} := \sum_{\vec{v}} T_u(v_1, v_2, \cdots v_N)\, Q^{\vec{v}}_{klm} \quad \text{for} \quad k \in \{1, 2, \cdots, K_l\} \tag{6}$$

*such that the linear map $\psi$ is injective, $\sum_{l,m} K_{l,m} \leq n^N$, and for each $g_u \in \mathrm{O(n)}$:*

$$\psi\big(g_u(\mathbf{T}_u)\big)_{kl} = \pi_l(g_u) \cdot \big(\psi(\mathbf{T}_u)\big)_{kl} \tag{7}$$

**Wigner-Eckart layer $\psi$.** Lemma 1 implies that we can use a fixed set of at most $n^{2N}$ coefficients $\mathbf{Q}$ to uniquely map the diagonal sub-tensors of $\mathbf{T}$ to order-1 $G$-equivariant embeddings $\mathbf{h}_u^{t=0} := \psi(\mathbf{T}_u)$, without solving tensor factorization on $\mathbf{T}_u$. $\mathbf{Q}$ can be shown to further decompose into products of $\mathrm{O(n)}$ Clebsch-Gordan coefficents which can simplify the contraction (6), and in practice $\mathbf{Q}$ can be numerically tabulated using integrals of radial functions and spherical harmonics (see Appendix A.2, A.5). Lemma 1 can be regarded as a generalization of the Wigner-Eckart theorem originally defined for $(N = 2, n = 3)$, therefore we also refer to $\psi$ defined in (6) as a *Wigner-Eckart layer*.

### 4.3 Interacting and normalizing representations with equivariance

**Equivariant Normalization (EvNorm).** We propose a normalization scheme on order-1 tensors to improve generalization while preserving $G$-equivariance. Given an order-1 tensor $\mathbf{x}$, we define $\mathrm{EvNorm} : \mathbf{x} \mapsto (\bar{\mathbf{x}}, \hat{\mathbf{x}})$ where $\bar{\mathbf{x}}$ and $\hat{\mathbf{x}}$ are given by

$$\bar{x}_{kl} := \frac{||\mathbf{x}_{kl}|| - \mu^x_{kl}}{\sigma^x_{kl}} \quad \text{and} \quad \hat{x}_{klm} := \frac{x_{klm}}{||\mathbf{x}_{kl}|| + 1/\beta_{kl} + \epsilon} \tag{8}$$

where $k$ index a feature channel, $\mu^x_{kl}$ and $\sigma^x_{kl}$ are mean and variance estimates of the invariant content $||\mathbf{x}||$; they can be obtained from either batch or layer statistics as in normalization schemes developed for scalar neural networks [28, 29]; $\beta_{kl}$ are positive, learnable scalars controlling the fraction of vector scale information from $\mathbf{x}$ to be retained in $\hat{\mathbf{x}}$, and $\epsilon$ is a numerical stability factor. The proposed EvNorm operation (8) decouples the order-1 tensor $\mathbf{x}$ to the normalized scalar-valued tensor $\bar{\mathbf{x}}$ suitable for being transformed by an MLP, and a 'pure-direction' tensor $\hat{\mathbf{x}}$ that can be later multiplied to the MLP-transformed normalized invariant content to finish updating $\mathbf{x}$. Note that in (8), zero is always a fixed point of the map $\mathbf{x} \mapsto \hat{\mathbf{x}}$ and the vector directions information $\mathbf{x}$ is always preserved.

**Point-wise interaction block $\phi$.** We propose a *point-wise interaction block* ($\phi$ in (5)) as a modular component to construct $\hat{\mathcal{F}}_\theta$, which equivariantly update $\mathbf{h}_u^{t+1} = \phi(\mathbf{h}_u^t, \mathbf{g}_u)$ by coupling another order-1 (i.e. vector) tensor $\mathbf{g}_u$ (e.g. $\tilde{\mathbf{m}}_u^t$ in (5), or $\mathbf{h}_u^t$ itself) with $\mathbf{h}_u^t$, and performing normalizations:

$$\mathbf{f}_u^t = \mathrm{MLP}_1(\bar{\mathbf{h}}_A^t) \odot \hat{\mathbf{h}}_u^t \quad \text{where} \quad (\bar{\mathbf{h}}_u^t, \hat{\mathbf{h}}_u^t) = \mathrm{EvNorm}(\mathbf{h}_u^t) \tag{9}$$

$$(\mathbf{q}_u)_{klm} = (\mathbf{g}_u)_{klm} + \sum_{l_1, l_2} \sum_{m_1, m_2} (\mathbf{f}_u^t)_{kl_1m_1} (\mathbf{g}_u)_{kl_2m_2}\, C^{lm}_{l_1m_1;l_2m_2} \tag{10}$$

$$\mathbf{h}_u^{t+1} = \mathbf{h}_u^t + \mathrm{MLP}_2(\bar{\mathbf{q}}_u) \odot \hat{\mathbf{q}}_u \quad \text{where} \quad (\bar{\mathbf{q}}_u, \hat{\mathbf{q}}_u) = \mathrm{EvNorm}(\mathbf{q}_u) \tag{11}$$

where $C^{lm}_{l_1m_1;l_2m_2}$ are Clebsch-Gordan coefficients known for relating vectors to their tensor products, and $\mathrm{MLP}_1$ and $\mathrm{MLP}_2$ denote multi-layer perceptrons. See Appendix A.2 for the proof on $G$-equivariance.

**Symmetry-informed pooling.** Once the representations $\mathbf{h}_u^t$ are updated to the last step $\mathbf{h}_u^{t_f}$, a pooling operation $\{\mathbf{h}_u^{t_f}\} \mapsto \mathbf{y}_\theta$ can be employed to readout the target prediction. Due to equivariance, we can flexibly address the the symmetry prior of the learning task by designing pooling schemes without modifying the model architecture. For quantum-chemistry tasks, we define a class of pooling operations detailed in Appendix A.7; for example, a molecule's dipole vector can be predicted as $\vec{\mu} = \sum_u (\vec{r}_u \cdot q_u + \vec{\mu}_u)$ where $\vec{r}_u$ is the atom $u$'s position, and atomic charges $q_u$ and atomic dipoles $\vec{\mu}_u$ can be respectively predicted using scalar ($l = 0$) and vector ($l = 1$) components of $\mathbf{h}_u^{t_f}$.

**Matching layer $\rho_u$ .** One subtlety that must be addressed for the convolution-message-passing layers (3-5) is that the basis $\mathbf{e}_u$ for the local reference frame at point $u$ may differ from the underlying basis for the vector representations $\mathbf{h}_u^t$. However, we can construct matching layers $\rho_u$ and $\rho_u^\dagger$ to connect the representations defined on the two sets of basis:

$$\left(\rho_u(\mathbf{h}_u^t)\right)_v^i = \sum_{l,m} \mathbf{W}_l^i \cdot (\mathbf{h}_u^t)_{lm} \cdot \langle \mathbf{e}_{u;v}, (\boldsymbol{\pi}_l)_m \rangle \tag{12}$$

$$\left(\rho_u^\dagger(\tilde{\mathbf{m}}_u^t)\right)_{lm} = \sum_{v} \mathbf{W}_l^\dagger \cdot (\tilde{\mathbf{m}}_u^t)_v \cdot \langle (\boldsymbol{\pi}_l)_m, \mathbf{e}_{u;v} \rangle \tag{13}$$

where $\mathbf{W}_l, \mathbf{W}_l^\dagger$ are learnable linear functions, and a careful treatment for the inner product $\langle \mathbf{e}_u, \boldsymbol{\pi} \rangle$ will be discussed in Appendix A.8. For a simple example that $\mathbf{e}_u$ is the standard basis of $\mathbb{R}^3$ and there is only one convolution channel $i$, $\rho_u^\dagger$ is (up to a constant) given by $\left(\rho_u^\dagger(\mathbf{v})\right)_{l=1,p=1,m} = Y_{l=1,m}(\mathbf{v})$ and $\left(\rho_u^\dagger(\mathbf{v})\right)_{l\neq 1,p,m} \equiv 0$, where $\mathbf{v} \in \mathbb{R}^3$ and $Y_{lm}$ is the spherical harmonic of degree $l$ and order $m$.

**Time complexity.** The asymptotic time complexity of UNiTE is $\mathcal{O}(BNI)$, where $B$ is the number of indices $(\vec{u}, \vec{v})$ such that $T_{\vec{u},\vec{v}} \neq 0$ (i.e., the number of non-zero elements in $\mathbf{T}$), and $I$ denotes the number of convolution channels in a convolution block (3). This implies UNiTE scales as $\mathcal{O}(N(nd)^N)$ if the $N$-body tensor $\mathbf{T}$ is dense, but can achieve a lower time complexity if $\mathbf{T}$ is sparse.

*Proof. (Informal) In each convolution block* (3) *the summand* $T_{\vec{u},\vec{v}} \cdot \prod_{j=2}^N \left(\rho_{u_j}(\mathbf{h}_{u_j}^t)\right)_{v_j}^i \neq 0$ *only if the tensor coefficient* $T_{\vec{u},\vec{v}} \neq 0$*; therefore* (3) *can be exactly evaluated using* $((N-1)BI)$ *arithmetic operations. It can be shown that in each message passing block* (4) *the number of arithmetic operations scales as* $\mathcal{O}(B'I)$ *where* $B'$ *is the number of indices* $\vec{u}$ *such that* $\mathbf{m}_{\vec{u}}^t \neq 0$*, and* $B' \leq B$*. It can be shown that the embedding block* $\phi$ *and the point-wise interaction block* $\psi$ *has* $\mathcal{O}(d)$ *time complexities and do not contribute to the asymptotic time complexity.*

See Appendix A.2 for the full proof. Several more technical aspects, including the case of multiple input channels and efficient implementations are discussed in Appendix A.6 and A.8.

# 5 Experimental results

**Problem statement.** Quantum chemistry studies interactions among atoms and electrons in a molecular system. We aim to learn quantum chemistry properties through utilizing 2-body tensor representations for molecules originated from physical approximations on such electron-atom interactions, in this work using a semi-empirical quantum mechanics model. A scientific background and the semi-empirical tight-binding quantum chemistry approach employed to generate the tensor-based molecular representation are provided in Appendix A.3-A.4. We note that compared to methods that were developed for solely utilizing molecular geometry inputs, constructing this physic-based molecular representation incurs a moderate computational overhead; model inference time breakdowns and discussions are provided in Appendix A.1. All the ground-truth labels of the benchmarks were generated based on a conventional numerical method, i.e. Density Function Theory (DFT).

We first explore UNiTE's performance on learning quantum-chemical properties on several open-source machine learning datasets. Then, using a UNiTE model trained only on energies, we perform zero-shot generalization to a variety of quantum chemical benchmarks, with comparison to physics-based and ML models. We use the **same** set of model hyperparameters for obtaining all experimental results. See Appendix A.9 for hyperparameter and training details.

**QM9.** The QM9 dataset [30] contains 134k small organic molecules with up to 9 heavy (CNOF) atoms in their equilibrium geometries, with scalar-valued chemical properties computed by DFT. QM9 is widely used to benchmark deep learning methods [19–22, 26, 31]. Following previous works, we use 110000 random samples as the training set and another 10831 samples as the test set. As shown in Table 1, we observe state-of-the-art performance on all 12 targets with a 150% average decrease of MAE relative to the second best model. We also perform experiments on two representative targets, energy $U_0$ and dipole vector $\vec{\mu}$, for which a plethora of task-specific ML models has previously been developed [32–37]; UNiTE outperforms all those methods across all sizes of training data. See Appendix A.11 for reference calculation details of $\vec{\mu}$.

Table 1: Experiments on QM9. (Left) Prediction MAEs on multiple targets for models trained on 110k samples. The best results on each task are marked in bold and the second-bests are indicated by underline. The results on first 8 tasks are obtained with a preconditioning strategy based on the physical approximation used to construct the molecular representation (see Appendix A.7). UNiTE achieves state-of-the-art on all 12 targets, outperforming the second-best (SphereNet) by 150% on average. (Right) Comparing UNiTE to task-specific models and deep learning methods for (a) energy $U_0$ and (b) dipole moment vector $\vec{\mu}$ at different training data sizes. Results obtained with the preconditioning strategy (black) and obtained without preconditioning (brown) are both included.

| Target | Unit | SchNet | PhysNet | Cormorant | DimeNet++ | PaiNN | SphereNet | **UNiTE** |
|---|---|---|---|---|---|---|---|---|
| $\mu$ | mD | 33 | 53 | 38 | 29.7 | <u>12</u> | 26.9 | **6.3** |
| $\epsilon_{\text{HOMO}}$ | meV | 41 | 34 | 32.9 | 24.6 | 27.6 | <u>23.6</u> | **9.9** |
| $\epsilon_{\text{LUMO}}$ | meV | 34 | 24.7 | 38 | 19.5 | 20.4 | <u>18.9</u> | **12.7** |
| $\Delta\epsilon$ | meV | 63 | 42.5 | 38 | 32.6 | 45.7 | <u>32.3</u> | **17.3** |
| $U_0$ | meV | 14 | 8.2 | 22 | 6.3 | <u>5.9</u> | 6.3 | **3.5** |
| $U$ | meV | 19 | 8.3 | 21 | 6.3 | <u>5.8</u> | 7.3 | **3.5** |
| $H$ | meV | 14 | 8.4 | 21 | 6.5 | <u>6.0</u> | 6.4 | **3.5** |
| $G$ | meV | 14 | 9.4 | 20 | 7.6 | <u>7.4</u> | 8.0 | **5.2** |
| $\alpha$ | $a_0^3$ | 0.235 | 0.062 | 0.085 | <u>0.044</u> | 0.045 | 0.047 | **0.036** |
| $\langle R^2 \rangle$ | $a_0^2$ | 0.073 | 0.765 | 0.961 | 0.331 | <u>0.066</u> | 0.292 | **0.030** |
| ZPVE | meV | 1.7 | 1.4 | 2.0 | 1.2 | 1.3 | **1.1** | **1.1** |
| $c_v$ | $\frac{\text{cal}}{\text{molK}}$ | 0.033 | 0.028 | 0.026 | 0.023 | 0.024 | **0.022** | **0.022** |
| std. MAE | % | 1.76 | 1.37 | 1.44 | 0.98 | 1.01 | 0.94 | **0.47** |
| log. MAE | - | -5.2 | -5.4 | -5.0 | -5.7 | -5.8 | -5.7 | **-6.4** |

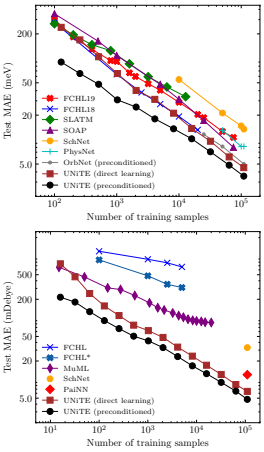


Table 2: Benchmarking UNiTE against representative semi-empirical quantum mechanics (SEQM), machine learning (ML), and density functional theory (DFT) methods on down-steam tasks.

| Task | Benchmark | Metric | *SEQM* GFN2-xTB | *ML* ANI-2x | *DFT* B97-3c | (Ours) UNiTE |
|---|---|---|---|---|---|---|
| Speed[3] | [43] | Relative time-to-solution↓ | ~1 | ~1 | ~200 | ~1 |
| Data efficiency | - | Training dataset size ↓ | - | 8.9M | - | 236K |
| Drug chemistry coverage | [43] | Sample coverage rate ↑ | 100% | 81% | 100% | 100% |
| General chemistry coverage | [44] | Subset coverage rate ↑ | 100% | 36% | 100% | 67% |
| Conformer ordering | [43] | $R^2$[DLPNO] ↑ | 0.63±0.04 | 0.63±0.06 | **0.90±0.01** | **0.89±0.02** |
| Torsion profiles | [45] | MAE (kcal/mol) ↓ | 0.73±0.01 | 0.90±0.01 | 0.29±0.00 | **0.17±0.00** |
| Reaction energies | [44] | WTMAD-2 ↓ | 36.1±3.8 | 19.2±10.1 | **14.6±1.6** | **14.6±4.3** |
| Intra-molecular interactions | [44] | WTMAD-2 ↓ | 25.1±2.4 | 29.6±5.8 | **8.6±0.9** | 10.3±1.7 |
| Geometry optimizations | [46]<br>[47] | RMSD (Å) ↓ | 0.21±0.08<br>0.60±0.06 | -<br>- | **0.06±0.01**<br>0.51±0.07 | **0.06±0.01**<br>**0.18±0.02** |

**MD17.** The MD17 dataset [38] contains energy and force labels from molecular dynamics trajectories of eight small organic molecules, and is used to benchmark ML methods for modelling a single instance of a molecular potential energy surface. We train UNiTE on energies and forces of 1000 geometries of each molecule and test on another 1000 geometries of the same molecule, using reported dataset splits and revised labels [39] (see Appendix A.9 for details). As shown in Table A1, UNiTE achieves over 110% average improvements on both energies and forces, when compared to kernel methods [39, 40] and graph neural networks [20, 24, 25].

**Electron density.** We next focus on the more challenging task of predicting the electron density of molecules $\rho(\vec{r}) : \mathbb{R}^3 \to \mathbb{R}$ which plays an essential role in the formulation of quantum chemistry. Equivariance enables UNiTE to efficiently learn $\rho(\vec{r})$ in a compact spherical expansion basis (see Appendix A.7). Compared to two baselines [41, 42] developed for learning $\rho(\vec{r})$, UNiTE achieves 50-75% reduction in mean $L^1$ density error. Remarkably, UNiTE is also more efficient at training compared to SA-GPR [41] which has a cubic training time complexity, and at inference compared to DeepDFT [42] which requires evaluating part of the neural network at each grid point $\vec{r}$.

**Down stream chemistry tasks.** To evaluate UNiTE's performance as a black-box quantum chemistry method, we train a UNiTE model on the DFT energies of 236k samples with broad chemical

[3] All based on CPU timings. While neural networks are better parallelized on GPUs, no GPU-based implementations are available for GFN2-xTB and B97-3c and we report timings with the same hardware setup.

space coverage, non-equilibrium geometries, and include charged systems (see Appendix A.10). Without any model fine-tuning, we directly apply it to down-stream tasks commonly used to benchmark quantum-chemistry simulation methods (detailed in Appendix A.12). In this zero-shot setting, our pretrained model achieves accuracy similar or better than a popular DFT method [48] while being around 200x faster and is significantly better than representative semi-empirical quantum mechanics methods [49] or machine learning methods [50] which offer comparable speeds.

# 6 Discussion and conclusions

We propose UNiTE, a neural network framework for learning general $N$-body tensors. It shows superior performance when applied to quantum chemistry, achieving up to three orders of magnitude speedup compared to DFT on down stream chemical tasks. A limitation of our approach is the model cannot be applied to several edge cases such as all diagonal sub-tensors are zeros, therefore is not universal; however we expect such examples to be rare in physical sciences. A natural future direction is to extend the formalism to general asymmetric order-$N$ tensors through representation-theoretic techniques. Given its demonstrated performance on practical tasks, we anticipate UNiTE to be useful in scientific applications in chemistry, quantum physics and other domains.

# Broader Impacts

Our contribution is a neural network framework with primary focus on tensor-based problems in physical sciences. Data used in this study do not contain human-related or offensive content. Although we do not foresee any direct negative societal impacts associated with human-related objects, we note that training the model may produce carbon emissons that should be taken into consideration for large-scale application scenarios. However, we also anticipate our approach to reduce carbon footprint in long terms by replacing more compute-intensive solvers. We expect our approach to be practically beneficial to the society when applied to tasks such as drug discovery and the study of elementary particles.

# References


[1] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

[2] 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.

[3] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

[4] Pim de Haan, Maurice Weiler, Taco Cohen, and Max Welling. Gauge equivariant mesh cnns: Anisotropic convolutions on geometric graphs. *arXiv preprint arXiv:2003.05425*, 2020.

[5] Robin Walters, Jinxi Li, and Rose Yu. Trajectory prediction using equivariant continuous convolution. *arXiv preprint arXiv:2010.11344*, 2020.

[6] Gregor NC Simm, Robert Pinsler, Gábor Csányi, and José Miguel Hernández-Lobato. Symmetry-aware actor-critic for 3d molecular design. *arXiv preprint arXiv:2011.12747*, 2020.

[7] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. *arXiv preprint arXiv:1807.02547*, 2018.

[8] Denis Boyda, Gurtej Kanwar, Sébastien Racanière, Danilo Jimenez Rezende, Michael S Albergo, Kyle Cranmer, Daniel C Hackett, and Phiala E Shanahan. Sampling using su (n) gauge equivariant flows. *Physical Review D*, 103(7):074504, 2021.

[9] Ta-Pei Cheng and Ling-Fong Li. *Gauge theory of elementary particle physics*. Oxford university press, 1994.

[10] Jun John Sakurai and Eugene D Commins. Modern quantum mechanics, revised edition, 1995.

[11] Fabian B Fuchs, Daniel E Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *arXiv preprint arXiv:2006.10503*, 2020.

[12] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2019.

[13] Danilo Jimenez Rezende, George Papamakarios, Sébastien Racaniere, Michael Albergo, Gurtej Kanwar, Phiala Shanahan, and Kyle Cranmer. Normalizing flows on tori and spheres. In *International Conference on Machine Learning*, pages 8083–8092. PMLR, 2020.

[14] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=SJU4ayYgl`.

[15] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272, 2017.

[16] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3558–3565, 2019.

[17] Song Bai, Feihu Zhang, and Philip HS Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021.

[18] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. *arXiv preprint arXiv:2102.09844*, 2021.

[19] Brandon Anderson, Truong Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14537–14546. Curran Associates, Inc., 2019. URL `http://papers.nips.cc/paper/9596-cormorant-covariant-molecular-neural-networks.pdf`.

[20] Kristof T Schütt, Oliver T Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. *arXiv preprint arXiv:2102.03150*, 2021.

[21] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in neural information processing systems*, pages 991–1001, 2017.

[22] Oliver T Unke and Markus Meuwly. Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.*, 15(6):3678–3693, 2019.

[23] Ziteng Liu, Liqiang Lin, Qingqing Jia, Zheng Cheng, Yanyan Jiang, Yanwen Guo, and Jing Ma. Transferable multi-level attention neural network for accurate prediction of quantum chemistry properties via multi-task learning. *ChemRxiv*, 12588170:v1, 2020.

[24] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations (ICLR)*, 2020.

[25] Simon Batzner, Tess E Smidt, Lixin Sun, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, and Boris Kozinsky. Se (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *arXiv preprint arXiv:2101.03164*, 2021.

[26] Yi Liu, Limei Wang, Meng Liu, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3d graph networks. *arXiv preprint arXiv:2102.05013*, 2021.

[27] Gurtej Kanwar, Michael S Albergo, Denis Boyda, Kyle Cranmer, Daniel C Hackett, Sébastien Racaniere, Danilo Jimenez Rezende, and Phiala E Shanahan. Equivariant flow-based sampling for lattice gauge theory. *Physical Review Letters*, 125(12):121601, 2020.

[28] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[29] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

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

[31] Johannes Klicpera, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020.

[32] Bing Huang and O. Anatole von Lilienfeld. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nature Chemistry*, 12(10):945–951, September 2020. doi: 10.1038/s41557-020-0527-z. URL `https://doi.org/10.1038/s41557-020-0527-z`.

[33] Felix A Faber, Anders S Christensen, Bing Huang, and O Anatole von Lilienfeld. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.*, 148(24):241717, 2018. doi: 10.1063/1.5020710.

[34] Albert P Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R Kermode, Gábor Csányi, and Michele Ceriotti. Machine learning unifies the modeling of materials and molecules. *Science advances*, 3(12):e1701816, 2017.

[35] Zhuoran Qiao, Matthew Welborn, Animashree Anandkumar, Frederick R Manby, and Thomas F Miller III. Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *The Journal of Chemical Physics*, 153(12):124111, 2020.

[36] Anders S Christensen, Felix A Faber, and O Anatole von Lilienfeld. Operators in quantum machine learning: Response properties in chemical space. *The Journal of chemical physics*, 150(6):064105, 2019.

[37] Max Veit, David M Wilkins, Yang Yang, Robert A DiStasio Jr, and Michele Ceriotti. Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles. *The Journal of Chemical Physics*, 153(2):024113, 2020.

[38] 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.

[39] Anders S Christensen and O Anatole von Lilienfeld. On the role of gradients for machine learning of molecular energies and forces. *Machine Learning: Science and Technology*, 1(4): 045018, 2020.

[40] Stefan Chmiela, Huziel E Sauceda, Klaus-Robert Müller, and Alexandre Tkatchenko. Towards exact molecular dynamics simulations with machine-learned force fields. *Nature communications*, 9(1):1–10, 2018.

[41] Alberto Fabrizio, Andrea Grisafi, Benjamin Meyer, Michele Ceriotti, and Clemence Corminboeuf. Electron density learning of non-covalent systems. *Chemical science*, 10(41):9424–9432, 2019.

[42] Peter Bjørn Jørgensen and Arghya Bhowmik. Deepdft: Neural message passing network for accurate charge density prediction. *arXiv preprint arXiv:2011.03346*, 2020.

[43] Dakota Folmsbee and Geoffrey Hutchison. Assessing conformer energies using electronic structure and machine learning methods. *Int. J. Quantum Chem.*, page e26381, 2020. doi: 10.1002/qua.26381.

[44] Lars Goerigk, Andreas Hansen, Christoph Bauer, Stephan Ehrlich, Asim Najibi, and Stefan Grimme. A look at the density functional theory zoo with the advanced gmtkn55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Physical Chemistry Chemical Physics*, 19(48):32184–32215, 2017.

[45] Brajesh Rai, Vishnu Sresht, Qingyi Yang, Rayomond J Unwalla, Meihua Tu, Alan M Mathiowetz, and Gregory A Bakken. Torsionnet: A deep neural network to rapidly predict small molecule torsion energy profiles with the accuracy of quantum mechanics. 2020.

[46] Tobias Risthaus, Marc Steinmetz, and Stefan Grimme. Implementation of nuclear gradients of range-separated hybrid density functionals and benchmarking on rotational constants for organic molecules. *Journal of Computational Chemistry*, 35(20):1509–1516, 2014.

[47] Uma R Fogueri, Sebastian Kozuch, Amir Karton, and Jan ML Martin. The melatonin conformer space: Benchmark and assessment of wave function and dft methods for a paradigmatic biological and pharmacological molecule. *The Journal of Physical Chemistry A*, 117(10): 2269–2277, 2013.

[48] Jan Gerit Brandenburg, Christoph Bannwarth, Andreas Hansen, and Stefan Grimme. B97-3c: A revised low-cost variant of the b97-d density functional method. *The Journal of chemical physics*, 148(6):064104, 2018.

[49] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. GFN2-xTB — an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.*, 15(3): 1652–1671, 2019.

[50] Christian Devereux, Justin S Smith, Kate K Davis, Kipton Barros, Roman Zubatyuk, Olexandr Isayev, and Adrian E Roitberg. Extending the applicability of the ani deep learning molecular potential to sulfur and halogens. *Journal of Chemical Theory and Computation*, 16(7):4192–4202, 2020.

# Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Confidence interval estimates for results in Table 1 is provided in Appendix A.1. The methods used to compute errorbars for results in Table 1 and Table 2 are detailed in Appendix A.13.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes] We mention that existing datasets used in this work are open source in Section 5, and list their licenses in the supplemental material.

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] As mentioned in the broader impact statement, the quantum chemistry datasets used in this work do not curate data from human sources.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] As mentioned in the broader impact statement, the quantum chemistry datasets used in this work do not contain such information.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not conduct research with human subjects.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]