Supplementary Information for

Informing geometric deep learning with electronic interactions to accelerate quantum chemistry

Zhuoran Qiao, Anders S. Christensen, Matthew Welborn, Frederick R. Manby, Anima Anandkumar, and Thomas F. Miller III

E-mail: anima@caltech.edu or tfm@caltech.edu

This PDF file includes:

- Supplementary text
- Figs. S1 to S4
- Tables S1 to S9
- SI References
Supporting Information Text

S1. Additional neural network details

A. Efficient GPU evaluation of spherical harmonics and Clebsch-Gordan coefficients. All O(3)-representation related operations are implemented through element-wise operations on arrays and gather-scatter operations, without the need of recursive computations that can be difficult to parallelize on GPUs at runtime. The real spherical harmonics (RSHs) are computed based on Equations 6.4.47-6.4.50 of (1), which reads:

\[ Y_{lm}(\mathbf{r}) = \frac{1}{\sqrt{4\pi}} \sqrt{\frac{(l-|m|)/2}{(l+|m|)/2}} \sum_{t=0}^{(l-|m|)/2} \frac{1}{2t+|m|} \sum_{u=0}^{(l+|m|)/2} C_{tuv}^{lm} \left( \frac{x}{|r|} \right)^{2t+|m|-2(u+v)} \left( \frac{y}{|r|} \right)^{2(u+v)} \left( \frac{z}{|r|} \right)^{2l-2|m|} \]  

[S1]

where \( \lfloor \cdot \rfloor \) is the floor function. The above scheme only requires computing element-wise powers of 3D coordinates and a linear combination of pre-tabulated coefficients. The Clebsch-Gordan (CG) coefficients are first tabulated using their explicit form with the transformation matrix between RSHs and CSHs (3).

B. Multiple input channels. UNiTE is naturally extended to inputs that possess extra feature dimensions, as in the case of AO features \( \mathbf{T} \) described in Section B the extra dimension equals the cardinality of selected QM operators. Those stacked features is processed by a learnable linear layer \( \mathbf{W}^{\text{in}} \) resulting in a fixed-size channel dimension. Each channel is then shared among a subset of convolution channels (indexed by \( i \)), instead of using one convolution kernel for all channels \( i \). For the numerical experiments of this work, \( \mathbf{T} \) are mixed into \( I \) input channels by \( \mathbf{W}^{\text{in}} \) and we assign a convolution channel to each input channel.

C. Restricted summands in Clebsch-Gordan coupling. For computational efficiency, in the Clebsch-gordan coupling Eq. (25) (i.e., Eq. (S43)) of a point-wise interaction block, we further restrict the angular momentum indices \( (l_1, l_2) \) within the range \( \{ (l_1, l_2) : l_1 + l_2 \leq l_{\text{max}}, 1 \leq l_1 \leq n, l_2 \leq n \} \) where \( l_{\text{max}} \) is the maximum angular momentum considered in the implementation.

D. Incorporating geometric information. Because the point cloud of atomic coordinates \( \mathbf{x} \) is available in addition to the atomic-orbital-based inputs \( \mathbf{T} \), we incorporated such geometric information through the following modified message-passing scheme to extend Eq. (17):

\[ (\mathbf{m}_l^i)_{lm} = \sum_{B \neq A} \bigoplus_{i,j} (\mathbf{m}_l^{i,j})_{AB} + Y_{lm}(\mathbf{x}_{AB}) (\mathbf{W}_i^l \cdot |\mathbf{m}_l^{i,j}||_{AB}_l) \cdot \alpha_{l}^{i,j} \]  

[S5]

where \( Y_{lm} \) denotes a spherical harmonics of degree \( l \) and order \( m \), \( \mathbf{x}_{AB} := \frac{\mathbf{x}_A - \mathbf{x}_B}{||\mathbf{x}_A - \mathbf{x}_B||} \) denotes the direction vector between atomic centers A and B, and \( \mathbf{W}_i^l \) are learnable linear functions.

E. Pooling layers. We define schemes for learning different classes of chemical properties with OrbNet-Equi without modifying the base UNiTE model architecture. We use \( A \) to denote an atom index, \( |A| \) to denote the total number of atoms in the molecule, \( z_A \in \mathbb{N}^+ \) to denote the atomic number of atom \( A \), and \( \mathbf{x}_A \in \mathbb{R}^3 \) to denote the atomic coordinate of atom \( A \).

E.1. Energetic properties. A representative target in this family is the molecular electronic energy \( E(x) \) (i.e., \( E_0 \) in the convention of QM9), which is rotation-invariant and proportional to the system size (i.e., extensive). The pooling operation is defined as:

\[ y_0 = \sum_A \mathbf{W}_0 \cdot ||\mathbf{h}_l^i|| + b_0^A \]  

[S6]

which is a direct summation over atom-wise contributions. \( \mathbf{W}_0 \) is a learnable linear layer and \( b_0^A \) are learnable biases for each atomic number \( z \). To account for nuclei contributions to molecular energies, we initialize \( b_0^A \) from a linear regression on the training labels with respect to \( \{ z_A \} \) to speed up training on those tasks. This scheme is employed for learning \( U_0, U, H, G, ZPVE \) and \( c_v \) on QM9, the energies part in MD17 and for the OrbNet-Equi/SDC21 model.
E.2. Dipole moment $\vec{\mu}$. The dipole moment $\vec{\mu}$ can be thought as a vector in $\mathbb{R}^3$. It is modelled as a combination of atomic charges $q_A$ and atomic dipoles $\vec{\mu}_A$, and the pooling operation is defined as

$$\vec{\mu}_q = \sum_A (\vec{R}_A \cdot q_A + \vec{\mu}_A)$$  \[S7\]

$$q_A = q_A' - \Delta q$$  \[S8\]

$$q_A' := W_{o,0} \cdot (h_{l}^{i})_{l=0,p=1} + b_{\mu_{A}}^0$$  \[S9\]

$$(\vec{\mu}_A)_m := W_{o,1} \cdot (h_{l}^{i})_{l=1,p=1,m} \text{ where } m \in \{x, y, z\}$$  \[S10\]

where $W_{o,0}$ and $W_{o,1}$ are learnable linear layers. Equation S8 ensures the translation invariance of the prediction through charge neutrality.

Note that OrbNet-Equi is trained by directly minimizing a loss function $\mathcal{L}(\vec{\mu}, \vec{\mu}_q)$ between the ground truth and the predicted molecular dipole moment vectors. For the published QM9 reference labels (4) only the dipole norm $\mu := ||\vec{\mu}||$ is available; we use the same pooling scheme to readout $\vec{\mu}_q$ but train on $\mathcal{L}(\mu, ||\vec{\mu}_q||)$ instead to allow for comparing to other methods in Table S1.

E.3. Polarizability $\alpha$. For isotropic polarizability $\alpha$, the pooling operation is defined as

$$\alpha = \sum_A (\alpha_A + \vec{R}_A \cdot \vec{\mu}_A)$$  \[S11\]

$$\alpha_A := W_{o,0} \cdot (h_{l}^{i})_{l=0,p=1} + b_{\alpha_{A}}^0$$  \[S12\]

$$\vec{\mu}_A = P_A^2 - \Delta P$$  \[S13\]

$$(P_A)_m := W_{o,1} \cdot (h_{l}^{i})_{l=1,p=1,m} \text{ where } m \in \{x, y, z\}$$  \[S14\]

E.4. Molecular orbital properties. For frontier molecular orbital energies, a global-attention based pooling is employed to produce intensive predictions:

$$a_A = \text{Softmax}(W_a \cdot ||h_{l}^{i}||) := \frac{W_a \cdot ||h_{l}^{i}||}{\sum_A W_a \cdot ||h_{l}^{i}||}$$  \[S15\]

$$y = \sum_A a_A \cdot (W_o \cdot ||h_{l}^{i}|| + b_{\mu_{A}}^0)$$  \[S16\]

where $W_a$ and $W_o$ are learnable linear layers and $b_{\mu_{A}}^0$ are learnable biases for each atomic number $z$. Similar to energy tasks, we initialize $b_{\mu_{A}}^0$ from a linear fitting on the targets to precondition training.

We take the difference between the predicted HOMO energies ($\epsilon_{\text{HOMO}}$) and LUMO energies ($\epsilon_{\text{LUMO}}$) as the HOMO-LUMO Gap ($\Delta \epsilon$) predictions.

E.5. Electronic spatial extent $\langle R^2 \rangle$. The pooling scheme for $\langle R^2 \rangle$ is defined as:

$$\langle R^2 \rangle_A = \sum_A ||\vec{R}_A - \vec{R}_0||^2 q_A + s_A$$  \[S17\]

$$\vec{R}_0 := \sum_A (\vec{R}_A \cdot q_A + \vec{\mu}_A)$$  \[S18\]

$$q_A := W_{o,0} \cdot (h_{l}^{i})_{l=0,p=1} + b_{\mu_{A}}^0$$  \[S19\]

$$(\vec{\mu}_A)_m := W_{o,1} \cdot (h_{l}^{i})_{l=1,p=1,m} \text{ where } m \in \{x, y, z\}$$  \[S20\]

$$s_A := W_{o,2} \cdot (h_{l}^{i})_{l=0,p=1}$$  \[S21\]

where $W_{o,0}$, $W_{o,1}$ and $W_{o,2}$ are learnable linear layers.

E.6. Electron densities $\rho(r)$. Both the ground truth and predicted electron densities $\rho(r)$ are represented as a superposition of atom-centered density fitting basis $\{\chi\}$,

$$\rho(r) = \sum_A \sum_l \sum_{m=-l}^{l} \sum_n n_{\text{max}}(n, l) d_{A}^{p|l|m} \chi_{A}^{n|m}(r)$$  \[S22\]
similar to the approach employed in (5); here we use the def2-TZVP-JKFIT density fitting basis for \( \{ \chi \} \). Computational details regarding obtaining the reference density coefficients \( \hat{d}_A^{nlm} \) are given in Section C, and the training loss function is defined in SI B.3. The pooling operation to predict \( \hat{\rho}(\vec{r}) \) from UNiTE is defined as

\[
\hat{d}_A^{nlm} := (W_u^A, \cdot \cdot (h_\sigma^A)_{l,p=1,m})_n
\]

where \( W_{A,j} \) are learnable weight matrices specific to each atomic number \( z \) and angular momentum index \( l \), and \( z_A \) denotes the atomic number of atom \( A \). This atom-centered expansion scheme compactly parameterizes the model-predicted density \( \hat{\rho}(\vec{r}) \). We stress that all UNiTE neural network parameters except for this density pooling layer Eq. (S23) are independent of the atomic numbers \( z \).

**F. Time complexity.** The asymptotic time complexity of UNiTE model inference is \( \mathcal{O}(BNI) \), where \( B \) is the number of non-zero elements in \( T \), and \( I \) denotes the number of convolution channels in a convolution block Eq. (15). This implies UNiTE scales as \( \mathcal{O}(N(nd)^N) \) if the input is dense, but can achieve a lower time complexity for sparse inputs, e.g., when long-range cutoffs are applied. We note that in each convolution block Eq. (15) the summand \( T_{x,\sigma} \cdot \prod_{i=2}^N ( \rho_{w_j}(h_i^x) )_{i} \neq 0 \) only if the tensor coefficient \( T_{x,\sigma} \neq 0 \); therefore Eq. (15) can be exactly evaluated using \( ((N-1)BI) \) arithmetic operations. In each message passing block Eq. (17) the number of arithmetic operations scales as \( \mathcal{O}(B'I) \) where \( B' \) is the number of indices \( \vec{u} \) such that \( m^I_\vec{u} \neq 0 \), and \( B' \leq B \). The embedding block \( \phi \) and the point-wise interaction block \( \psi \) has \( \mathcal{O}(d) \) time complexities since they act on each point independently and do not contribute to the asymptotic time complexity.

**S2. Theoretical results**

We formally introduce the problem of interest, restate the definitions of the building blocks of UNiTE (Methods A) using more formal notations, and prove the theoretical results claimed in this work. We first generalize the input data domain to a generic class of tensors beyond quantum chemistry quantities; for brevity we call such inputs \( N \)-body tensors.

**A. \( N \)-body tensors (informal).** We are interested in a class of tensors \( T \), for which each sub-tensor \( T(u_1,u_2,\cdots,u_N) \) describes relation among a collection of \( N \) geometric objects defined in an \( n \)-dimensional physical space. For simplicity, we will first 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.

**Definition S1 (\( N \)-body tensor).** Let \( \{x_1,x_2,\cdots,x_N\} \) be \( d \) points in \( \mathbb{R}^d \) for each \( u \in \{1,2,\cdots,d\} \). For each point index \( u \), we define an orthonormal basis (local reference frame) \( \{e_{u,v_u}\} \) centered at \( x_u \), and denote the space spanned by the basis as \( V_u := \text{span} \{e_{u,v_u}\} \subseteq \mathbb{R}^d \). We consider a tensor \( \hat{T} \) defined via \( N \)-th direct products of the ‘concatenated’ basis \( \{e_{u,v_u}; (u,v_u)\} \):

\[
\hat{T} := \sum_{\sigma} T((u_1;v_1),(u_2;v_2),\cdots,(u_N;v_N)) \ e_{u_1,v_1} \otimes e_{u_2,v_2} \otimes \cdots \otimes e_{u_N,v_N}
\]

\( \hat{T} \) is a tensor of order-\( N \) and is an element of \( (\mathbb{R}^d)^\otimes_N \). We call its coefficients \( T \) an \( N \)-body tensor if \( T \) is invariant to global translations \( \forall x_0 \in \mathbb{R}^n, T[x] = T[x + x_0] \), and is symmetric:

\[
T((u_1;v_1),(u_2;v_2),\cdots,(u_N;v_N)) = T((u_1;v_1),(u_2;v_2),\cdots,(u_N;v_N))
\]

where \( \sigma \) denotes arbitrary permutation on its dimensions \( \{1,2,\cdots,N\} \). Note that each sub-tensor, \( T_\sigma \), 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 \( \{x_1,x_2,\cdots,x_N\} \) which then identifies a sub-tensor \( T_{\vec{u}} \) in the \( N \)-body tensor \( T \); \( \vec{v} := (v_1,v_2,\cdots,v_N) \) index a coefficient \( T_\vec{v}(v_1,v_2,\cdots,v_N) := T((u_1;v_1),(u_2;v_2),\cdots,(u_N;v_N)) \) in a sub-tensor \( T_{\vec{u}} \), where each index \( v_j \in \{1,2,\cdots,\text{dim}(V_{u_j})\} \) for \( j \in \{1,2,\cdots,N\} \).

(a) Sequence (b) Velocity field (c) Graph (d) AO features

1-Body Invariant 1-Body Equivariant 2-Body Invariant 2-Body Equivariant

*We additionally allow for \( \vec{v} \in \{e_{u,v_u}\} \) to represent features in \( T \) that transform as scalars.*
We aim to build neural networks $\mathcal{F}_\theta : (\bigoplus_{u=1}^d V_u)^\otimes N \to \mathcal{Y}$ that map $\tilde{T}$ to order-1 tensor- or scalar-valued outputs $y \in \mathcal{Y}$.

While $\tilde{T}$ is independent of the choice of local reference frame $e_u$, its coefficients $T$ (i.e. the N-body tensor) vary when rotating or reflecting the basis $e_u := \{e_{u,v_u} : v_u\}$, i.e. acted by an element $U_u \in O(n)$. Therefore, the neural network $\mathcal{F}_\theta$ should be constructed equivariant with respect to those reference frame transformations.

### B. 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 $v \in \mathcal{V}$, $\varphi(g) \cdot f(v) = f(\varphi'(g) \cdot v)$ where $\varphi(g)$ and $\varphi'(g)$ are the group representation of element $g$ on $\mathcal{V}$ and $\mathcal{Y}$, respectively. In our case, the group $G$ is composed of (a) Unitary transformations $U_u$ locally applied to basis: $e_u \mapsto U_u e_u$, which are rotations and reflections for $\mathbb{R}^n$. $U_u$ induces transformations on tensor coefficients: $T_\mathcal{V} \mapsto (U_{u_1} \otimes U_{u_2} \otimes \cdots \otimes U_{u_n}) T_\mathcal{V}$, and an intuitive example for infinitesimal basis rotations in $N = 2$, $n = 2$ is shown in Figure S2; (b) Tensor index permutations: $(\bar{\tilde{u}}, \bar{\tilde{v}}) \mapsto \sigma(\bar{\tilde{u}}, \bar{\tilde{v}})$; (c) Global translations: $x \mapsto x + x_0$. For conciseness, we borrow the term $G$-equivariance to say $\mathcal{F}_\theta$ is equivariant to all the symmetry transformations listed above.

### C. N-body tensors

Here we generalize the definition of N-body tensors to the basis of irreducible group representations instead of a Cartesian basis. The atomic orbital features discussed in the main text fall into this class, since the angular parts of atomic orbitals (i.e., spherical harmonics $Y_{lm}$) form the basis of the irreducible representations of group $SO(3)$.

#### Definition S2

Let $G_1, G_2, \cdots, G_d$ denote unitary groups where $G_u \subset U(n)$ are closed subgroups of $U(n)$ for each $u \in \{1, 2, \ldots, d\}$. We denote $G := G_1 \times G_2 \times \cdots \times G_d$. Let $(\pi_L, \mathcal{V}_L)$ denote an irreducible unitary representation of $U(n)$ labelled by $L$. For each $u \in \{0, 1, 2, \ldots, d\}$, we assume there is a finite-dimensional Banach space $V_u \simeq \bigoplus_{k \in \mathbb{N}} \mathcal{V}_L^{\otimes K_L}$, where $K_L \in \mathbb{N}$ is the multiplicity of $\mathcal{V}_L$ (e.g. the number of feature channels associated with representation index $L$), with basis $\{\pi_{L,M} u\}$ such that $\operatorname{span}\{\pi_{L,M} u : k, L, M\} = V_u$ for each $u \in \{1, 2, \ldots, d\}$ and $k \in \{1, 2, \ldots, K_L\}$, and span $\{\pi_{L,M} u : M\}$ $\simeq \mathcal{V}_L$ for each $u, L$.

We denote $V := \bigoplus_u V_u$, and index notation $v := (k, L, M)$. For a tensor $T \in \mathcal{V}^{\otimes N}$, we call the coefficients $\mathbf{T}$ of $\mathbf{T}$ in the N-th direct products of basis $\{\pi_{L,M} u : L, M, u\}$ an N-body tensor, if $\mathbf{T} = \sigma(\mathbf{T})$ for any permutation $\sigma \in \operatorname{Sym}(N)$ (i.e. permutation invariant).

#### Note

Note that the vector spaces $V_u$ do not need to be embed in the same space $\mathbb{R}^n$ as in the special case from Definition S1, but can be originated from general 'parameterizations' $u \mapsto V_u$, e.g., coordinate charts on a manifold.

#### Corollary S1

If $V_u = \mathbb{C}^n$, $G_u = U(n)$ and $\pi_{L,M,u} = \mathbf{e}_M$ where $\{\mathbf{e}_M\}$ is a standard basis of $\mathbb{C}^n$, then $\mathbf{T}$ is an N-body tensor if $\mathbf{T}$ is permutation invariant.

#### Proof

For $V_u = \mathbb{C}^n$, $\pi : G_u \to U(\mathbb{C}^n)$ is a fundamental representation of $U(n)$. Since the fundamental representations of a Lie group are irreducible, it follows that $\{\mathbf{e}_M\}$ is a basis of an irreducible representation of $U(n)$, and $\mathbf{T}$ is an N-body tensor.

Similarly, when $V_u = \mathbb{R}^n$ and $G_u = O(n) \subset U(n)$, $\mathbf{T}$ is an N-body tensor if $\tilde{T}$ is permutation invariant. Then we can recover the special case based on point clouds in $\mathbb{R}^n$ in Definition S1.

Procedures for constructing complete bases for irreducible representations of $U(n)$ with explicit forms are established (6). A special case is $G_u = SO(3)$, for which a common construction of a complete set of $\{\pi_{L,M} u\}$ is using the spherical harmonics $\pi_{L,M,u} := Y_{lm}$; this is an example that polynomials $Y_{lm}$ can be constructed as a basis of square-integrable functions on the 2-sphere $L^2(S^2)$ and consequently as a basis of the irreducible representations $(\pi_L, \mathcal{V}_L)$ for all $L (7)$.

1 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 $\sum_{u=1}^{n} u_{u_1,u_2,u_3,u_4}$ contains $3^6 = 81$ elements with indices spanning from $x_{1234}$ to $x_{5678}$. In total, T contains $(5 \times 3)^4$ coefficients. The coefficients of T are in general complex-valued as formally discussed in Definition S2, but are real-valued for the special case introduced in Definition S1.
D. Decomposition of diagonals $T_u$. We consider the algebraic structure of the diagonal sub-tensors $T_u$, which can be understood from tensor products of irreducible representations.

First we note that for a sub-tensor $T_u \in V_{u_1} \otimes V_{u_2} \otimes \cdots \otimes V_{u_N}$, the action of $g \in G$ is given by

$$g \cdot T_u = (\pi(g_{u_1}) \otimes \pi(g_{u_2}) \otimes \cdots \otimes \pi(g_{u_N}))T_u$$  \[S26\]

for diagonal sub-tensors $T_u$, this reduces to the action of a diagonal sub-group

$$g \cdot T_u = (\pi(g_u) \otimes \pi(g_u) \otimes \cdots \otimes \pi(g_u))T_u$$  \[S27\]

which forms a representation of $G_u \in U(n)$ on $V_u^{\otimes N}$. According to the isomorphism $V_u \cong \bigoplus_L (\mathcal{V}^L)^{\otimes K_L}$ in Definition S2 we have $\pi(g_u) \cdot \mathbf{v} = \bigoplus_L U_{g_u}^{L_1} \cdot \mathbf{v}_L$ for $\mathbf{v} \in V_u$ where $\mathbf{v}_L \in \mathcal{V}^L$, more explicitly

$$g \cdot T_u(k, \bar{\ell}) = (U_{g_u}^{L_1} \otimes U_{g_u}^{L_2} \otimes \cdots \otimes U_{g_u}^{L_N})T_u(k, \bar{\ell})$$  \[S28\]

where we have used the shorthand notation $T_u(k, \bar{\ell}) := T_u((k_1, L_1), (k_2, L_2), \ldots, (k_N, L_N))$ and $U_{g_u}^L$ denotes the unitary matrix representation of $g_u \in U(n)$ on $\mathcal{V}^L$ expressed in the basis $\{\pi_{L,M;u}:\{M\}\}$, on the vector space $\mathcal{V}^L$ for the irreducible representation labelled by $L$. Therefore $T_u(k, \bar{\ell}) \in \mathcal{V}^{L_1} \otimes \mathcal{V}^{L_2} \otimes \cdots \otimes \mathcal{V}^{L_N}$ is the representation space of an $N$-fold tensor product representations of $U(n)$. We note the following theorem for the decomposition of $T_u(k, \bar{\ell})$:

**Theorem S1** (Theorem 2.1 and Lemma 2.2 of (8)). The representation of $U(n)$ on the direct product of $\mathcal{V}^{L_1}, \mathcal{V}^{L_2}, \ldots, \mathcal{V}^{L_N}$ decomposes into direct sum of irreducible representations:

$$\mathcal{V}^{L_1} \otimes \mathcal{V}^{L_2} \otimes \cdots \otimes \mathcal{V}^{L_N} \cong \bigoplus_L \sum_{\mu} \mathcal{V}^{L;\mu}$$  \[S29\]

and

$$\sum_L \mu(L_1, L_2, \ldots, L_N; L) \dim(\mathcal{V}^L) = \prod_{u=1}^N \dim(\mathcal{V}^{L_u})$$  \[S30\]

where $\mu(L_1, L_2, \ldots, L_N; L)$ is the multiplicity of $L$ denoting the number of replicas of $\mathcal{V}^L$ being present in the decomposition of $\mathcal{V}^{L_1} \otimes \mathcal{V}^{L_2} \otimes \cdots \otimes \mathcal{V}^{L_N}$.

Note that we have abstracted the labelling details for $U(n)$ irreducible representations into the index $L$. See (8) for proof and details on representation labelling. We now state the following result for generating order-1 representations (Materials and methods, A.1):

**Corollary S2.** There exists an invertible linear map $\psi: V_u^{\otimes N} \rightarrow V_u^e := \bigoplus \mathcal{V}^{L;\mu(L,V_u)}$ where $\mu(L,V_u) \in \mathbb{N}$, such that for any $T_u$, $L$ and $\nu \in \{1, 2, \ldots, \mu(L,V_u)\}$, $\psi(g_u \cdot T_u)_{\nu,L} = U_{g_u}^L \psi(T_u)_{\nu,L}$ if $\mu(L,V_u) > 0$.

**Proof.** First note that each block $T_u(k, \bar{\ell})$ of $T_u$ is an element of $\mathcal{V}^{L_1} \otimes \mathcal{V}^{L_2} \otimes \cdots \otimes \mathcal{V}^{L_N}$ up to an isomorphism. Eq. (S30) in Theorem S1 states there is an invertible linear map $\psi^e: \mathcal{V}^{L_1} \otimes \mathcal{V}^{L_2} \otimes \cdots \otimes \mathcal{V}^{L_N} \rightarrow \bigoplus_L (\mathcal{V}^L)^{\otimes \mu(L_1, L_2, \ldots, L_N; L)}$, such that $\pi(g_u) = (\psi^e)^{-1} \circ \pi(g_u) \circ \psi^e$ for any $g_u \in G_u$, where $\pi: G_u \rightarrow U(\mathcal{V}^{L_1} \otimes \mathcal{V}^{L_2} \otimes \cdots \otimes \mathcal{V}^{L_N})$ and $\pi: G_u \rightarrow U(\bigoplus_L (\mathcal{V}^L)^{\otimes \mu(L_1, L_2, \ldots, L_N; L)})$ are representations of $G_u$. Note that $\pi$ is defined as a direct sum of irreducible representations of $U(n)$, i.e. $\pi(g_u) \psi^e(T_u(k, \bar{\ell})) := \bigoplus_{\nu,L} \mu(L_1, L_2, \ldots, L_N; L)$ directly satisfies $\psi(g_u \cdot T_u)_{\nu,L} = \psi(\pi(g_u)T_u)_{\nu,L} = U_{g_u}^L \psi(T_u)_{\nu,L}$ for $\nu \in \{1, 2, \ldots, \mu(L,V_u)\}$. Since each $\psi^e_L$ is finite-dimensional and invertible, it follows that the finite direct sum $\psi$ is invertible.

For Hermitian tensors, we conjecture the same result for SU(2), O(2) and O(3) as each irreducible representation is isomorphic to its complex conjugate.

We then formally restate the proposition in Methods A.1 which was originally given for orthogonal representations of O(3) (i.e., the real spherical harmonics):

**Corollary S3.** For each $L$ where $\mu(L,V_u) > 0$, there exist $n_L \times \dim(V_u)^N$ $T$-independent coefficients $Q^e_{\nu,L,M}$ parameterizing the linear transformation $\psi$ that performs $T_u \mapsto h_u := \psi(T_u)$, if $u_1 = u_2 = \cdots = u_N = u$:

$$\psi(T_u)_{\nu,L,M} := \sum_{v} T_u(v_1, v_2, \cdots, v_N)Q^e_{\nu,L,M}$$  \[S31\]

such that the linear map $\psi$ is injective, $\sum_L n_L \leq \dim(V_u)^N$, and for each $g_u \in G_u$:

$$\psi(g_u \cdot T_u)_{\nu,L} = U_{g_u}^L \psi(T_u)_{\nu,L}$$  \[S32\]
Proof. According to Definition S2, a complete basis of $V^\otimes N$ is given by \( \{ \pi_{L_1,M_{1,u}} \otimes \pi_{L_2,M_{2,u}} \otimes \cdots \otimes \pi_{L_N,M_{N,u}}; (k, \tilde{L}, \tilde{M}) \} \) and a complete basis of \((V_u^*)_L \) is \( \{ \pi_{L,M,u}^*; (k, \tilde{L}, \tilde{M}) \} \). Note that \( V_u \) and \((V_u^*)_L \) are both finite dimensional. Therefore an example of \( \mathbf{Q}_L \) is the \( \dim(V_u^*)_L \times \mu(L; V_u) \) matrix representation of the bijective map \( \psi \) in the two basis, which proves the existence. \( \square \)

Note that Corollary S3 does not guarantee the resulting order-1 representations \( h_u := \psi(T_u) \) (i.e. vectors in \( V_u^* \)) to be invariant under permutations \( \sigma \), as the ordering of \( \{ \nu \} \) may change under \( T \mapsto \sigma(T) \). Hence, the symmetric condition on \( T \) is important to achieve permutation equivariance for the decomposition \( V^\otimes N \rightarrow V_u^* \); we note that \( T_u \) has a symmetric tensor factorization and is an element of \( \text{Sym}^N(V_u) \), then algebraically the existence of a permutation-invariant decomposition is ensured by the Schur-Weyl duality (9) giving the fact that all representations in the decomposition of \( \text{Sym}^N(V_u) \) commute with the symmetric group \( S_N \). With the matrix representation \( \mathbf{Q} \) in Eq. (S31), clearly for any \( \sigma \), \( \psi(\sigma(T_u)) = \sigma(T_u) \cdot \mathbf{Q} = T_u \cdot \mathbf{Q} = \psi(T_u) \).

For general asymmetric \( N \)-body tensors, we expect the realization of permutation equivariance to be sophisticated and may be achieved through tracking the Schur functors from the decomposition of \( V^\otimes N \rightarrow V_u^* \), which is considered out of scope of the current work. Additionally, the upper bound \( \sum_L n_L \leq \dim(V_u^*)_L \) is in practice often not saturated and the contraction Eq. (S31) can be simplified. For example, when \( N > 2 \) it suffices to perform permutation-invariant decomposition on symmetric \( T_u \) recursively through Clebsch-Gordan coefficients \( \mathbf{C} \) which has the following property:

\[
\mathbf{C}^\nu_{L_1,L_2} := (U^1_{g_{u_1}} \otimes U^2_{g_{u_2}}) \cdot (\mathbf{C}^\nu_{L_1,L_2}) = U^L_{g_{u}} \quad \text{for} \quad \nu \in \{ 1, 2, \ldots, \mu(L_1, L_2; L) \}
\]

i.e., \( \mathbf{C} \) parameterizes the isomorphism \( \psi_L \) of Theorem S2 for \( N = 2 \), \( \tilde{L} = (L_1, L_2) \). Then \( \psi \) can be constructed with the procedure \( (V_u^*)_\otimes N \rightarrow V_u^* \otimes (V_u^*)_L \otimes \cdots \otimes (V_u^*)_L \rightarrow V_u^* \) without explicit order-\( N+1 \) tensor contractions, where each reduction step can be parameterized using \( \mathbf{C} \).

Procedures for computing \( \mathbf{C} \) in general are established (10, 11). For the main results reported in this work \( O(3) \simeq \text{SO}(3) \times \mathbb{Z}_2 \) is considered, where \( \mu(L_1, L_2; L) \leq 1 \) and the basis of an irreducible representation \( \pi_{L,M} \) can be written as \( \pi_{L,M} := [l, m, p] \) where \( p \in \{1, -1\} \) and \( m \in \{-l, -l+1, \ldots, -1, 1, \ldots, l\} \). \( \mu(l, m, p) \) can be thought as a spherical harmonic \( Y_{lm} \) but may additionally flips sign under point reflections \( \tilde{p} \) depending on the parity index \( p: \tilde{p} \| l, m, p \| = p \cdot (-1)^{l-l+1} \) \( l, m, p \) where \( \forall x \in \mathbb{R}^3, \tilde{p}(x) = -x \).

Clebsch-Gordan coefficients \( \mathbf{C} \) for \( O(3) \) is given by:

\[
\mathbf{C}^\nu_{I_1I_2} := \mathbf{C}^\nu_{I_1I_2} \quad \text{for} \quad \nu = \{ 1, 2, \ldots, \mu(L_1, L_2; L) \}
\]

where \( \mathbf{C}^\nu_{I_1I_2} \) are \( \text{SO(3)} \) Clebsch-Gordan coefficients. For \( N = 2 \), the problem reduces to using Clebsch-Gordan coefficients to decompose \( T_u \) as a combination of matrix representations of \( \text{spherical tensor operators} \) which are linear operators transforming under irreducible representation of \( \text{SO}(3) \) based on the the Wigner-Eckart Theorem (see (2) for formal derivations). Remarkably, a recent work (12) discussed connections of a class of neural networks to the Wigner-Eckart Theorem in the context of operators in spherical CNNs, which also provides a thorough review on this topic.

Both \( V_u \) and \( V_u^* \) are defined as direct sums of the representation spaces \( \mathbb{V}^L \) of irreducible representations of \( G_u \), but each \( L \) may be associated with a different multiplicity \( K_L \) or \( K^\nu_L \) (e.g. different numbers of feature channels). We also allow for the case that the definition basis \( \{ \mathbf{e}_{u,v} \} \) for the \( N \)-body tensor \( T \) differ from \( \{ \pi_{u,v,L,M} \} \) by a known linear transformation \( \mathbf{D}_u \) such that \( \mathbf{e}_{u,v} := \sum_{L,M} (D_u)_v^M \pi_{u,v,L,M} \) or \( (D_u)_v^M := \langle \cdot, \cdot \rangle \) denotes an Hermitian inner product, and we additionally define if \( K_L = 0 \), \( \langle \mathbf{e}_{u,v}, \pi_{u,v,L,M} \rangle := 0 \). We then give a natural extension to Definition S2:

\begin{definition}
We extend the basis in Definition S2 for \( N \)-body tensors to \( \{ \mathbf{e}_{u,v} \} \) where \( \text{span}(\{ \mathbf{e}_{u,v}; v \}) = V_u^* \), if
\end{definition}

\[
\mathbf{D}_u \cdot \pi^2(g_u) = \pi^1(g_u) \cdot \mathbf{D}_u \quad \forall g_u \in G_u
\]

where \( \pi^1 \) and \( \pi^2 \) are matrix representations of \( g_u \) on \( V_u \subset V_u^* \) in basis \( \{ \mathbf{e}_{u,v} \} \) and in basis \( \{ \pi_{u,v,L,M} \} \). Note that \( \pi^2(g_u) \cdot v = U^L_{g_u} \cdot v \) for \( v \in \mathbb{V}^L \).

\section{Generalized neural network building blocks}

We clarify that in all the sections below \( n \) refers to a feature channel index within an irreducible representation group labelled by \( L \), which should not be confused with \( \dim(V_u) \). More explicitly, we note \( n \in \{ 1, 2, \ldots, N^L \} \) where \( N^L \) is the number of vectors in the order-1 tensor \( h_u \) that transforms under the \( L \)-th irreducible representation \( G_u \) (i.e. the multiplicity of \( L \) in \( h_u^u \)). \( M \in \{ 1, 2, \ldots, \dim(\mathbb{V}^L) \} \) indicates the \( M \)-th component of a vector in the representation space of the \( L \)-th irreducible representation of \( G_u \), corresponding to a basis vector \( \pi_{u,v,L,M} \). We also denote the total number of feature channels in \( h \) as \( N^h := \sum L N^L \)

For a simple example, if the features in the order-1 representation \( h^u \) are specified by \( L \in \{ 0, 1, 2 \} \), \( N^{h^0}_L = 8 \), \( N^{h^1}_L = 4 \), \( \dim(\mathbb{V}^L=0) = 1 \), and \( \dim(\mathbb{V}^L=1) = 5 \), then \( N^h = 8 + 4 + 12 = 24 \) and \( h^u \) is stored as an array with \( \sum L N^h \cdot \dim(\mathbb{V}^L) = (8 \times 1 + 4 \times 5) = 28 \) elements.

We reiterate that \( \vec{u} := (u_1, u_2, \ldots, u_N) \) is a sub-tensor index (location of a sub-tensor in the \( N \)-body tensor \( T \)) and \( \vec{v} := (v_1, v_2, \ldots, v_N) \) is an element index in a sub-tensor \( T_{\vec{u}} \).

7 of 16
Convolution and message passing. We generalize the definition of a block convolution module Eq. (15) to order-\(N\) and complex numbers:

\[
\begin{align*}
\mathbf{m}_2^i & = \sum_{v_2, \ldots, v_N} T_d(v_1, v_2, \ldots, v_N) \prod_{j=2}^N (\rho_{u_j} (\mathbf{h}_u^i))^{i}_{v_j} \\
\mathbf{m}_1^i & = \sum_{u_2, u_3, \ldots, u_N} \bigoplus_{i,j} (\mathbf{m}_2^i)^{i} \cdot \alpha_{u_j}^{i,j}
\end{align*}
\]

Message passing modules Eq. (17)-Eq. (27) is generalized to order \(N\):

\[
\begin{align*}
\mathbf{m}_1^{t+1} & = \phi (\mathbf{h}_u^t, \rho_{u_1} (\mathbf{m}_u^i)) \\
\mathbf{h}_u^{t+1} & = \phi (\mathbf{h}_u^t, \rho_{u_1} (\mathbf{m}_u^i))
\end{align*}
\]

EvNorm. We write the EvNorm operation Eq. (22) as EvNorm : \( \mathbf{h} \mapsto (\mathbf{h}, \mathbf{\tilde{h}}) \) where

\[
\begin{align*}
\tilde{h}_{nL} & := \frac{|h_{nL}|-\mu_{nL}^2}{\sigma_{nL}^2} \quad \text{and} \quad \hat{h}_{nLM} := \frac{x_{nLM}}{|h_{nL}|-1/\beta_{nL} + \epsilon}
\end{align*}
\]

Point-wise interaction \( \phi \). We adapt the notations and explicitly expand Eq. (24)-Eq. (26) for clarity. The operations within a point-wise interaction block \( \mathbf{h}_u^{t+1} = \phi (\mathbf{h}_u^t, \mathbf{g}_u) \) are defined as:

\[
\begin{align*}
(b_i^{t+1})_{nLM} & = (\text{MLP}_2 (b_i^t))_{nLM} \text{ where } (\mathbf{h}_u^t, \mathbf{h}_u) = \text{EvNorm}(\mathbf{h}_u^t) \\
(q_i^{t+1})_{nLM} & = (\mathbf{g}_u)_{nLM} + \sum_{l_1, l_2, M_1, M_2} \sum_{l_1, l_2, M_1, M_2} (b_i^{t+1})_{nL1M1M2} \cdot C^{\nu(n)}_{l_1M1, l_2M2} \\
(h_i^{t+1})_{nLM} & = (\mathbf{h}_u^t)_{nLM} + (\text{MLP}_2 (q_i^t))_{nLM} \text{ where } (q_i^t, q_i) = \text{EvNorm}(q_i)
\end{align*}
\]

where \( \nu : \mathbb{N}^+ \to \mathbb{N}^+ \) assigns an output multiplicity index \( \nu \) to a group of feature channels \( n \).

For the special example of \( O(3) \) where the output multiplicity \( \nu(L_1, L_2; L) \leq 1 \) (see Theorem S1 for definitions), we can restrict \( \nu(n) \equiv 1 \) for all values of \( n \), and Eq. (S41) can be rewritten as

\[
(q_i_{nLM})_{nLM} = (\mathbf{g}_u)_{nLM} + \sum_{l_1, l_2, M_1, M_2} \sum_{l_1, l_2, M_1, M_2} (b_i^{t+1})_{nL1M1M2} \cdot C^{\nu(n)}_{l_1M1, l_2M2} \\
\]

which is based on the construction of \( C^{\nu(n)}_{l_1M1, l_2M2} \) in Eq. (S34). The above form recovers Eq. (25).

Matching layers. Based on Definition S3, we can define generalized matching layers \( \rho_u \) and \( \rho^*_u \) as

\[
\begin{align*}
(\rho_u (\mathbf{h}_u^i))^{i}_{v} & = \sum_{L,M} (\mathbf{W}_L^i) \cdot (\mathbf{h}_u^{i,M}) \cdot (\mathbf{e}_{u,v}, \mathbf{\pi}_{u,L,M}) \\
(\rho^*_u (\mathbf{m}_u^i))^{i}_{LM} & = \sum_{v} (\mathbf{W}_L^i) \cdot (\mathbf{m}_u^{i,v}) \cdot (\mathbf{\pi}_{u,L,M}, \mathbf{e}_{u,v})
\end{align*}
\]

where \( \mathbf{W}_L^i \) are learnable \((1 \times N_L^i)\) matrices; \( \mathbf{W}_L^i \) are learnable \((N_L^i \times (N_i^1))\) matrices where \( N_i^1 \) denotes the number of convolution channels (number of allowed \( i \) in Eq. (15)).

F. G-equivariance. With main results from Corollary S2 and Corollary S3 and basic linear algebra, the equivariance of UNITE can be straightforwardly proven. G-equivariance of the Diagonal Reduction layer \( \psi \) is stated in Corollary S3, and it suffices to prove the equivariance for other building blocks.

Proof of G-equivariance for the convolution block Eq. (S36). For any \( g \in G \):

\[
\begin{align*}
\sum_{v_2, \ldots, v_N} (g \cdot T_d(v_1, v_2, \ldots, v_N)) \prod_{j=2}^N (\rho_{u_j} (g \cdot \mathbf{h}_u^i)^{i}_{v_j}) \\
= \sum_{v_2, \ldots, v_N} ((\bigotimes_{a} \pi^a (g_{a})) \cdot T_d(v_1, v_2, \ldots, v_N)) \prod_{j=2}^N (\rho_{u_j} (\pi^a (g_{a})) \cdot \mathbf{h}_u^{i}_{v_j}) \\
= \sum_{v_2, \ldots, v_N} ((\bigotimes_{a} \pi^a (g_{a}) \cdot T_d(v_1, v_2, \ldots, v_N)) \prod_{j=2}^N (\sum_{L,M} (\mathbf{D}_{a_j})^{L,M}_{v_j} \cdot (\mathbf{W}_{L}^{i} \cdot (\mathbf{e}_{a_j,v} \cdot \mathbf{\pi}_{a_j,L,M}))^{i}_{v_j} \\
= \sum_{v_2, \ldots, v_N} ((\bigotimes_{a} \pi^a (g_{a}) \cdot T_d(v_1, v_2, \ldots, v_N)) \prod_{j=2}^N (\sum_{L,M} (\mathbf{W}_{L}^{i} \cdot (\mathbf{D}_{a_j})^{L,M}_{v_j} \cdot (\mathbf{e}_{a_j,v} \cdot \mathbf{\pi}_{a_j,L,M}))^{i}_{v_j}
\]
\[
\begin{align*}
&\sum_{v_2, \ldots, v_N} \left( \bigotimes_d \pi^1(g_{u_j}) \cdot T_u \right) (v_1, v_2, \ldots, v_N) \prod_{j=2}^N \left( \sum_{L, M} \left( W^L_L \cdot (\pi^1(g_{u_j}) \cdot D^M_{v_j} \cdot h^L_{u_j}) \right) v_j \right)^i \\
&= \sum_{v_2, \ldots, v_N} \left( \bigotimes_d \pi^1(g_{u_j}) \cdot T_u \right) (v_1, v_2, \ldots, v_N) \prod_{j=2}^N \left( (\pi^1(g_{u_j}))^* \cdot (\rho_u(h^L_{u_j}))^* \right) v_j \\
&= \sum_{v_2, \ldots, v_N} \sum_{v'_1, v'_2, \ldots, v'_N} \left( (\pi^1(g_{u_j}))_{v_j, v'_j} \cdot T_u (v'_1, v'_2, \ldots, v'_N) \right) \prod_{j=2}^N \left( (\pi^1(g_{u_j}))^* \cdot (\rho_u(h^L_{u_j}))^* \right)_{v'_j} \\
&= \sum_{v'_1} \prod_{j=2}^N \left( (\pi^1(g_{u_1}))_{v_1, v'_1} T_u (v'_1, v'_2, \ldots, v'_N) \right) \prod_{j=2}^N \left( (\rho_u(h^L_{u_j}))^* \right)_{v'_j} \\
&= \sum_{v'_1} \prod_{j=2}^N \left( \pi^1(g_{u_1})_{v_1, v'_1} \cdot (m^L_{v_1})^i \right) \\
&= \pi^1(g_{u_1}) \cdot (m^L_{v_1})^i = (g \cdot (m^L_{v_1}))^i \\
\end{align*}
\]

Proof of G-equivariance for the message passing block Eq. (S37)-Eq. (S38). From the invariance condition \( g \cdot \alpha^L_{u} = \alpha^L_{u} \), clearly

\[
\sum_{u_2, u_3, \ldots, u_N} \bigoplus_{i,j} (g \cdot m^L_{u})^i \cdot \alpha^L_{u} = \sum_{u_2, u_3, \ldots, u_N} \bigoplus_{i,j} (\pi^1(g_{u_1}) \cdot (m^L_{u})^i) \cdot \alpha^L_{u} = \pi^1(g_{u_1}) \cdot \sum_{u_2, u_3, \ldots, u_N} \bigoplus_{i,j} (m^L_{u})^i \cdot \alpha^L_{u} = \pi^1(g_{u_1}) \cdot m^L_{u_1} = g \cdot m^L_{u_1}
\]

Proof of G-equivariance for EvNorm Eq. (S39). Note that the vector norm \( \|x_{nL}\| \) is invariant to unitary transformations \( x_{nL} \mapsto U^L_{g_u} \cdot x_{nL} \). Then \( (g \cdot x) = \|x_{nL}\|/\|x_{nL}\| = x \), and \( (g \cdot x_{nL}) = \|x_{nL}\|/\|x_{nL}\| = \pi^2(g_{u}) \cdot x_{nL} = g \cdot x_{nL} \).

Proof of G-equivariance for the point-wise interaction block Eq. (S40)-Eq. (S42). Equivariances for Eq. (S40) and Eq. (S42) are direct consequences of the equivariance of EvNorm \( (g \cdot x) = x \) and \( (g \cdot x_{nL}) = g \cdot x_{nL} \), if \( g \cdot x_{nL} = \pi^2(g_{u}) \cdot x_{nL} \equiv U^L_{g_u} \cdot x_{nL} \).

Then it suffices to prove \( g \cdot (q_{u})_{nL} = U^L_{g_u} \cdot (q_{u})_{nL} \), which is ensured by Eq. (S33):

\[
(g_{u} \cdot g_{u})_{nL,M} + \sum_{L_1, L_2} (g_{u} \cdot f_{L_1, L_2})_{nL,M,1} (g_{u} \cdot g_{u})_{nL,M,2} C^{\nu(n),LM}_{L_1,M_1, L_2,M_2} = (U^L_{g_u} \cdot g_{u})_{nL,M} + \sum_{L_1, L_2} (U^L_{g_u} \cdot f_{L_1, L_2})_{nL,M,1} (U^L_{g_u} \cdot g_{u})_{nL,M,2} C^{\nu(n),LM}_{L_1,M_1, L_2,M_2}
\]

For permutation equivariance, it suffices to realize \( \sigma(T) \equiv T \) due to the symmetric condition in Definition S2 so Eq. (S36) is invariant under \( \sigma \), the permutation invariance of \( \psi \) (see Equation S31), and the actions of \( \sigma \) on network layers in \( \phi \) defined for a single dimension \{\(u, v\}\} are trivial (since \( \sigma(u) \equiv u \)).
### S3. Supplementary numerical results

#### Table S1. Test mean absolute errors (MAEs) on QM9 for atomistic deep learning models and OrbNet-Equi trained on 110k samples. OrbNet-Equi results on first 8 tasks are obtained by training on the residuals between the DFT reference labels and the tight-binding QM model estimations (delta-learning), because the tight-binding results for these targets can be directly obtained from the single-point calculation that featurizes the molecule. Results on the last 4 targets are obtained through directly training on the target properties (direct-learning).

| Target | Unit | SchNet | Cormorant | DimeNet++ | PaiNN | SphereNet | OrbNet-Equi |
|--------|------|--------|-----------|-----------|-------|-----------|-------------|
| $\mu$  | mD   | 33     | 38        | 29.7      | 12    | 26.9      | 6.3±0.2     |
| $\epsilon_{\text{HOMO}}$ | meV | 41     | 32.9      | 24.6      | 27.6  | 23.6      | 9.9±0.02    |
| $\epsilon_{\text{LUMO}}$ | meV | 34     | 38        | 19.5      | 20.4  | 18.9      | 12.7±0.3    |
| $\Delta \epsilon$ | meV | 63     | 38        | 32.6      | 45.7  | 32.3      | 17.3±0.3    |
| $U_0$  | meV  | 14     | 22        | 6.3       | 5.9   | 6.3       | 3.5±0.1     |
| $U$    | meV  | 19     | 21        | 6.3       | 5.8   | 7.3       | 3.5±0.1     |
| $H$    | meV  | 14     | 21        | 6.5       | 6.0   | 6.4       | 3.5±0.1     |
| $G$    | meV  | 14     | 20        | 7.6       | 7.4   | 8.0       | 5.2±0.1     |
| $\alpha_{a}$ | | 0.235  | 0.085     | 0.044     | 0.045 | 0.047     | 0.036±0.002 |
| $\langle R^2 \rangle$ | | 0.073  | 0.961     | 0.331     | 0.066 | 0.292     | 0.030±0.001 |
| ZPVE  | meV  | 1.7    | 2.0       | 1.2       | 1.3   | 1.1       | 1.11±0.04   |
| $c_v$  | $\frac{\text{meV}}{\text{m}^2}$ | 0.033  | 0.026     | 0.023     | 0.024 | 0.022     | 0.022±0.001 |

| std. MAE | %     | 1.76   | 1.44      | 0.98      | 1.01  | 0.94      | 0.47        |
| log. MAE | -     | -5.2   | -5.0      | -5.7      | -5.8  | -5.7      | -6.4        |

#### Table S2. Test MAEs on the rMD17 dataset in terms of energies (in meV) and forces (in meV/Å) for models trained on 1000 sample geometries for each molecular system. For OrbNet-Equi, both direct learning and delta learning results are reported.

| Molecule       | FCHL19 (13) | NequP ($l = 3$) (14) | OrbNet-Equi (direct learning) | OrbNet-Equi (delta learning) |
|----------------|-------------|----------------------|------------------------------|------------------------------|
| Aspirin        | Energy      | 6.2                  | 2.3                          | 2.4                          | 1.8                          |
|                | Forces      | 20.9                 | 8.5                          | 7.6                          | 6.1                          |
| Azobenzene     | Energy      | 2.8                  | 0.7                          | 1.1                          | 0.63                         |
|                | Forces      | 10.8                 | 3.6                          | 4.2                          | 2.7                          |
| Ethanol        | Energy      | 0.9                  | 0.4                          | 0.62                         | 0.42                         |
|                | Forces      | 6.2                  | 3.4                          | 3.7                          | 2.6                          |
| Malonaldehyde  | Energy      | 1.5                  | 0.8                          | 1.2                          | 0.80                         |
|                | Forces      | 10.2                 | 5.2                          | 7.1                          | 4.6                          |
| Naphthalene    | Energy      | 1.2                  | 0.2                          | 0.46                         | 0.27                         |
|                | Forces      | 6.5                  | 1.2                          | 2.6                          | 1.5                          |
| Paracetamol    | Energy      | 2.9                  | 1.4                          | 1.9                          | 1.2                          |
|                | Forces      | 12.2                 | 6.9                          | 7.1                          | 4.5                          |
| Salicylic Acid | Energy      | 1.8                  | 0.7                          | 0.73                         | 0.52                         |
|                | Forces      | 9.5                  | 4.0                          | 3.8                          | 2.9                          |
| Toluene        | Energy      | 1.6                  | 0.3                          | 0.45                         | 0.27                         |
|                | Forces      | 8.8                  | 1.6                          | 2.5                          | 1.6                          |
| Uracil         | Energy      | 0.4                  | 0.4                          | 0.58                         | 0.35                         |
|                | Forces      | 4.2                  | 3.2                          | 3.8                          | 2.4                          |
| Benzene        | Energy      | 0.3                  | 0.04                         | 0.07                         | 0.02                         |
|                | Forces      | 2.6                  | 0.3                          | 0.73                         | 0.27                         |

### A. The QM9 dataset

We provide the QM9 MAEs on all 12 target properties as reported in Table S1. The standardized MAE and standardized log MAE in Table S1 are computed following the Appendix C of (20). Uncertainties for test MAEs are obtained by statistical bootstrapping with sample size 5000 and 100 iterations.

### B. MD17 and rMD17 datasets

The MD17 dataset (21) contains energy and force labels from molecular dynamics trajectories of small organic molecules, and is used to benchmark ML methods for modelling a single instance of a molecular potential energy surface. Recently the revised-MD17 (rMD17) dataset (13) was reported with improved label fidelity. For both the MD17 and

---

1 With respect to $\omega$B97X-D3/def2-TZVP. Note that ANI-2x is trained on a different DFT theory and the number is provided for reference only.
Table S3. OrbNet-Equi test force MAEs (in kcal/mol/Å) on the original MD17 dataset using 1000 training geometries.

| Molecule         | OrbNet-Equi (direct learning) | OrbNet-Equi (delta learning) |
|------------------|-------------------------------|-----------------------------|
| Aspirin          | 0.156                         | 0.118                       |
| Ethanol          | 0.092                         | 0.069                       |
| Malonaldehyde    | 0.159                         | 0.128                       |
| Naphthalene      | 0.064                         | 0.048                       |
| Salicylic Acid   | 0.097                         | 0.067                       |
| Toluene          | 0.072                         | 0.057                       |
| Uracil           | 0.098                         | 0.072                       |

Table S4. OrbNet-Equi inference time breakdowns (mean/std in milliseconds) for the calculation of energy and forces on the Hutchison dataset (15).

| Feature generation | NN inference | NN back propagation | Nuclear gradients calculation |
|--------------------|--------------|---------------------|-------------------------------|
| 85.8 ± 40.1        | 181 ± 83     | 273 ± 73            | 33.2 ± 1.8                    |

Table S5. Summary statistics of representative semi-empirical quantum mechanics (GFN-xTB and GFN2-xTB), machine learning (ANI-2x), density functional theory (B97-3c) methods and OrbNet-Equi/SDC21 on down-stream tasks. See section E regarding results on geometry optimization tasks.

| Task                        | Dataset       | Metric                  | GFN-xTB      | GFN2-xTB     | ANI-2x       | B97-3c       | OrbNet-Equi/SDC21 |
|-----------------------------|---------------|-------------------------|--------------|--------------|--------------|--------------|-------------------|
| Conformer ordering          | Hutchison     | Med. $R^2$ / DLPNO-CCSD(T) | 0.62±0.04    | 0.64±0.04    | 0.63±0.06    | 0.90±0.01    | 0.87±0.02         |
| Conformer ordering          | Hutchison     | Med. $R^2$ / B97X-D3/def2-TZVP | 0.64±0.04    | 0.69±0.04    | 0.68±0.04    | 0.97±0.01    | 0.96±0.01         |
| Torsion profiles            | TorsionNet    | MAE‡ (kcal/mol)          | 0.948±0.017  | 0.731±0.013  | 0.893±0.017  | 0.284±0.006  | 0.173±0.003       |
| Geometry optimization       | ROT34         | Avg. RMSD (Å)            | 0.227±0.087  | 0.210±0.072  | -            | 0.063±0.013    | 0.045±0.005       |
| Geometry optimization       | MOONF         | Avg. RMSD (Å)            | 0.899±0.106  | 0.603±0.064  | -            | 0.511±0.072    | 0.227±0.042       |

Table S6. MAEs (in kcal/mol) of binding energy predictions on the S66x10 (19) dataset, computed for different inter-molecular distances in $r_e$ unit with $r_e$ being the equilibrium inter-molecular distance. B97X-D3/def2-TZVP binding energy results are used as the reference level of theory. Standard errors of the mean are reported in the parentheses.

| Distance ($r_e$) | GFN-xTB        | GFN2-xTB       | ANI-2x         | B97-3c         | OrbNet-Equi/SDC21 |
|------------------|----------------|---------------|----------------|----------------|-------------------|
| 0.7              | 6.758(2.192)   | 6.887(2.219)  | 2.323(0.596)   | 1.785(0.603)   | 1.644(0.465)      |
| 0.8              | 2.622(0.790)   | 2.859(0.879)  | 1.143(0.243)   | 0.975(0.245)   | 0.924(0.283)      |
| 0.9              | 1.408(0.195)   | 1.301(0.271)  | 1.010(0.160)   | 0.592(0.103)   | 0.539(0.151)      |
| 0.95             | 1.436(0.169)   | 1.208(0.180)  | 0.975(0.159)   | 0.501(0.083)   | 0.494(0.112)      |
| 1.0              | 1.555(0.173)   | 1.192(0.177)  | 0.968(0.148)   | 0.443(0.067)   | 0.350(0.088)      |
| 1.05             | 1.596(0.174)   | 1.196(0.184)  | 0.950(0.146)   | 0.375(0.052)   | 0.309(0.087)      |
| 1.1              | 1.577(0.175)   | 1.180(0.184)  | 0.940(0.160)   | 0.304(0.043)   | 0.332(0.094)      |
| 1.25             | 1.269(0.169)   | 0.976(0.171)  | 0.964(0.170)   | 0.134(0.021)   | 0.443(0.073)      |
| 1.5              | 0.827(0.153)   | 0.576(0.121)  | 0.850(0.136)   | 0.061(0.016)   | 0.469(0.068)      |
| 2.0              | 0.334(0.089)   | 0.166(0.037)  | 0.713(0.207)   | 0.029(0.007)   | 0.282(0.074)      |

As shown in Table S2 and Table S3, OrbNet-Equi with direct learning achieves competitive accuracy when compared against the best results reported by kernel methods (13, 23) and graph neural networks (GNNs) (14, 20, 24). Additional performance gains are observed when the models are trained with the delta learning strategy, resulting in the lowest test MAEs for both energy and forces on most of the test molecules. We note that MD17 represents an ideal case scenario where abundant high-level reference calculations are available on the potential energy surface of a single chemical system and the configurations of interest only spans a thermally-accessible energy scale. Despite the highly-interpolative nature of this learning task, OrbNet-Equi results still matches the accuracy that can be achieved with state-of-the-art neural network potentials for which system-dependent optimizations are often employed.
Table S7. Subset-averaged WTMAD-2 (WTMAD-2 = \( \frac{1}{N} \sum_{i=1}^{N} \text{WTMA}_i \), see Methods C) on the GMTKN55 collection of benchmarks, reported for all potential energy methods considered in this study against the CCSD(T):CBS reference values. Standard errors of the mean are reported in the parentheses. For cases in which no reaction within a subset is supported by a method, results are marked as “-“.

| Group          | Subset | GPN1-xTB | GPN2-xTB | ANI-2x | B97-3c | \( \omega \) | B97x3D | OrbNet-Equi/SDC21 | OrbNet-Equi/SDC21 (filtered) |
|----------------|--------|----------|----------|--------|--------|-----------|--------|------------------|-----------------------------|
| React. barriers|        |          |          |        |        |           |        |                  |                             |
| BH76           | 64.2(6.6) | 59.9(6.6) | 68.9(55.2) | 21.0(1.7) | 6.9(0.6) | 57.0(6.0) | 35.7(6.7) |
| BHPERI         | 25.4(2.1) | 27.9(2.3) | 65.7(9.6) | 12.5(0.8) | 7.8(0.9) | 10.5(2.4) | 10.5(2.4) |
| BHDIV10        | 10.5(3.7) | 50.3(9.3) | 132.2(20) | 7.6(1.0) | 2.2(0.7) | 1.3(0.3) | 1.3(0.3)  |
| BSR36          | 8.2(0.7)  | 9.7(1.6)  | 31.0(3.3) | 6.7(0.5) | 15.3(1.6) | 24.1(2.4) | 24.1(2.4) |
| CDIE20         | 28.6(4.8) | 25.3(4.1) | 50.9(9.5) | 27.8(3.1) | 10.1(2.3) | 16.0(3.6) | 16.0(3.6) |
| IOS34          | 24.6(3.7) | 26.9(4.2) | 50.5(9.8) | 7.3(1.4) | 4.7(0.6) | 7.3(3.3) | 7.3(3.3)  |
| ISOL24         | 28.3(4.4) | 30.3(4.7) | 18.9(4.1) | 13.5(2.8) | 7.1(1.2) | 7.6(1.3) | 7.6(1.3)  |
| C60ISO         | 4.6(1.0)  | 3.4(0.9)  | 26.0(2.9) | 3.6(1.1) | 7.2(1.2) | 2.3(0.4) | 2.3(0.4)  |
| PArEl          | 55.8(13.2) | 72.0(18.5) | -        | 22.1(6.2) | 8.2(2.0) | 43.9(3.9) | 43.9(3.9) |
| Inter. mol. NCI|        |          |          |        |        |           |        |                  |                             |
| RG18           | 31.8(7.1) | 11.0(3.1) | 11.8(3.0) | 11.1(1.8) | 53.6(13.1) | -       |               |
| ADIM6          | 17.2(1.8) | 19.5(4.2) | 5.8(1.2)  | 8.9(2.1) | 6.2(2.2) | 4.5(1.2) | 4.5(1.2)  |
| S22            | 10.4(1.7) | 5.9(0.9)  | 11.7(2.8) | 2.2(0.4) | 2.8(0.5) | 4.1(0.6) | 4.1(0.6)  |
| S66            | 11.2(0.8) | 7.6(0.6)  | 11.5(1.2) | 3.4(0.4) | 5.4(0.4) | 5.1(0.5) | 5.1(0.5)  |
| HEAVY28        | 30.0(9.6) | 27.8(5.0) | -        | 36.8(4.0) | 12.2(2.2) | 54.3(8.8) | -          |
| WATER27        | 5.2(0.7)  | 2.1(0.3)  | 33.4(8.1) | 6.6(0.9) | 10.0(1.6) | 12.0(1.5) | 12.0(1.5) |
| CARBHB12       | 6.3(1.4)  | 16.9(6.7) | 58.3(17.1) | 19.5(4.3) | 7.8(1.2) | 19.7(3.9) | 19.7(3.9) |
| PNICO23        | 31.0(6.7) | 14.7(2.7) | 251.8(7.3) | 21.8(2.6) | 5.0(0.7) | 39.1(9.5) | 39.1(9.5) |
| HAL59          | 16.6(2.6) | 15.8(1.7) | 74.4(4.7) | 20.1(2.9) | 4.2(0.4) | 35.3(6.0) | 35.3(6.0) |
| AHB21          | 11.8(2.6) | 7.5(1.2)  | -        | 8.3(1.2) | 8.6(1.2) | 18.6(3.7) | 18.6(3.7) |
| CHB6           | 8.4(3.9)  | 11.5(2.3) | -        | 2.9(1.2) | 2.8(0.9) | 24.0(10.7) | 24.0(10.7) |
| IL16           | 3.0(0.6)  | 2.2(0.3)  | -        | 1.2(0.3) | 1.1(0.2) | 2.5(0.4) | 2.5(0.4)  |
| Intra. mol. NCI|        |          |          |        |        |           |        |                  |                             |
| IDISP          | 26.1(14.4) | 27.1(17.0) | 82.5(45.7) | 15.6(5.3) | 11.3(3.6) | 19.7(8.9) | 19.7(8.9) |
| ICONF          | 45.7(13.5) | 28.3(4.8) | 37.3(27.2) | 6.6(1.5) | 5.9(1.4) | 29.8(10.3) | 29.8(10.3) |
| ACONF          | 20.5(3.4) | 6.0(1.3)  | 7.3(1.1)  | 4.8(1.1) | 2.7(0.3) | 1.6(0.4) | 1.6(0.4)  |
| Amino20x4      | 26.0(2.1) | 22.2(2.2) | 23.2(2.2) | 7.6(0.7) | 6.1(0.5) | 7.0(0.6) | 7.0(0.6)  |
| PCONF21        | 76.0(14.1) | 61.6(10.1) | 77.2(18.5) | 29.0(5.4) | 11.7(2.0) | 17.8(2.5) | 17.8(2.5) |
| MCONF          | 16.5(1.2) | 19.7(1.5) | 9.4(0.9)  | 3.8(0.4) | 5.5(0.4) | 5.4(0.5) | 5.4(0.5)  |
| SCONF          | 30.9(12.0) | 20.3(6.2) | 30.5(7.0) | 9.5(2.1) | 3.7(1.2) | 6.6(1.3) | 6.6(1.3)  |
| UP023          | 12.3(1.6) | 28.9(2.9) | -        | 5.0(0.8) | 9.4(1.0) | 10.7(1.5) | 10.7(1.5) |
| BUT14DIOL      | 19.4(1.6) | 25.4(1.8) | 28.8(1.7) | 8.4(0.5) | 8.3(0.4) | 11.9(0.5) | 11.9(0.5) |
C. Electron-density-derived electric moment predictions. Based on the predicted charge densities from models described in Section (Main text, Accurate modelling for electron densities), we have computed the electric moment (dipole $\mu$, quadrupole $\Theta$, and octupole $\Omega$) deviations relative to the reference DFT theory ($\omega$B97X-D3/Def-TZVP) for 128 random samples from the QM9 and the BFDb-SSI test sets, respectively. To further guide the interpretation of results, we also compare the predicted multipole errors to the deviations between two density functionals, B3LYP/Def2-TZVP and $\omega$B97X-D3/Def-TZVP.

As shown in Figure S3 and Figure S4, although no explicit DFT moments information was incorporated into model training, the OrbNet-Equi predicted density still yields dipole and higher moment errors comparable to the disagreement among modern density functionals (i.e., B3LYP v.s. $\omega$B97X-D3).

![Electric dipole moment deviations](image1)

**Fig. S3.** Electric dipole moment deviations $\varepsilon_\mu = \|\vec{\mu}_{\text{pred}} - \vec{\mu}_{\text{ref}}\|$ derived from the OrbNet-Equi charge density model are plotted against the (left) $L^1$ density error $\varepsilon_\rho$ and (right) dipole moment deviations of B3LYP/Def2-TZVP relative to the $\omega$B97X-D3/Def-TZVP reference, over randomly-sampled molecules from the (red) QM9 and (blue) BFDb-SSI test dataset. The grey dashed line indicates $x = y$. Average $\varepsilon_\mu$ of OrbNet-Equi density-derived dipole predictions: 0.185 ± 0.018 Debye (QM9), 0.152 ± 0.018 Debye (BFDb-SSI). Average $\varepsilon_\mu$ of B3LYP/Def2-TZVP density-derived dipole: 0.075 ± 0.007 Debye (QM9), 0.071 ± 0.015 Debye (BFDb-SSI).

![Electric quadrupole moment deviations](image2)

**Fig. S4.** Electric quadruple moment deviations $\varepsilon_\Theta = \|\Theta_{\text{pred}} - \Theta_{\text{ref}}\|$ and octupole moment errors $\varepsilon_\Omega = \|\Omega_{\text{pred}} - \Omega_{\text{ref}}\|$ are plotted against their counterparts between B3LYP/Def2-TZVP and $\omega$B97X-D3/Def-TZVP calculations. $\|\|$ denotes a tensor Frobenius norm. Average $\varepsilon_\Theta$ of OrbNet-Equi density-derived quadrupole: 0.856 ± 0.075 a.u. (QM9), 0.532 ± 0.050 Debye (BFDb-SSI); Average $\varepsilon_\Omega$ of OrbNet-Equi density-derived octupole: 3.68 ± 0.76 a.u. (QM9), 5.32 ± 0.53 Debye (BFDb-SSI).

All electric moments are defined in the Cartesian convention (e.g., $\Theta_{\alpha\beta} := \sum_A z_A R_{A,\alpha} R_{A,\beta} - \int \rho(\vec{r}) r_\alpha r_\beta d\vec{r}$) and are computed using numerical integration over the predicted electron densities with a grid spacing of $[\Delta x, \Delta y, \Delta z] = [0.1, 0.1, 0.1]$ Bohr and grid boundaries set to 5 Bohr from the outermost atom of the molecular system. The convergence of numerical integration is confirmed by comparing the numerically-computed dipole deviations between B3LYP and $\omega$B97X-D3 functionals to the exact electronic-integral-based results. Note that presented molecular quadrupoles and octupoles are origin-dependent due to...
non-vanishing dipole moments, and the molecular principal axes of inertia are used for all numerical calculations.

While a lower average dipole prediction error on QM9 was achieved by directly training OrbNet-Equi on dipole labels (Figure 3b), we note that the charge density model provides finer-scale information due to the availability of detailed charge distributions. Moreover, we anticipate that by introducing analytic electric moment terms to the electron density loss function, both the description on electrostatics (e.g., multipoles and electrostatic potential) and the density prediction accuracy can be further improved. We consider that as a promising future direction.

D. Inference timings. Wall-clock timing results for evaluating the pretrained OrbNet-Equi/SDC21 model are reported on the Hutchison dataset (15) which represents the distribution of realistic drug-like molecules. All timing results are obtained using 16 cores of an Intel Xeon Gold 6130 CPU. We note that due to the use of a Lagrangian formalism we previously developed (25) to efficiently compute the analytic nuclear gradients in which the operator derivatives with respect to atomic coordinates are not explicitly evaluated, for such medium-sized organic molecules the overhead for computing energy and forces in addition to only computing the energies is still dominated by neural network back propagation.

E. Downstream benchmarks. Table S5-S7 provide summary statistics of method performances on downstream main-group quantum chemistry benchmarks considered in this study.

Additional computational details for geometry optimization experiments are provided as follows. The symmetry-corrected root-mean-square-deviations (RMSDs) are computed between the test geometries and the reference DFT (ωB97X-D3/def2-TZVP) geometries following a Hungarian algorithm (26) to account for equivalent atoms. OrbNet-Equi/SDC21 and GFN-xTB results are obtained using Entos Qcore version 1.1.0 with the L-BFGS algorithm with tight thresholds (energy change after iteration < 1E-6 a.u., gradient RMS < 4.5E-4 a.u., max element of gradient < 3E-4 a.u., optimization step RMS < 1.8E-3 a.u., and max element of optimization step < 1.2E-3 a.u.). GFN2-xTB results are obtained with the XTB (27) package with default settings. ANI-2x energy calculations are performed with the TorchANI (28) package and geometry optimizations are performed with the geomTRIC optimizer (29) with default settings. Using this software setting, ANI-2x optimizations are found unable to converge on 21 out of 52 conformers on MCONF and 1 out of 12 conformers on ROT34, and the average errors are reported as “–”. On the subsets that ANI-2x geometry optimizations converged, the average RMSD for ANI-2x is 0.154±0.038 Å on ROT34 and 0.324±0.026 Å on MCONF versus the reference geometries. The histograms and kernel density estimations displayed in Figure 5d are computed on the subsets where all methods successfully converged, that is, 11 conformers from ROT34 and 31 conformers from MCONF.

S4. Hyperparameters and training details

A. Model hyperparameters. We use the same set of model hyperparameters to obtain all numerical experiment results on open benchmarks reported in this work. The hyperparameters are summarized in Table S8 and Table S9. The hyperparameters for the OrbNet-Equi/SDC21 model is locally optimized based on a 4219-sample validation set from the training data distribution, with the additional difference of using LayerNorm (30) for mean μ and variance σ estimates, and ϵ = 0.5 in all EvNorm layers. This choice is made to improve model robustness when applied to extrapolative molecular geometries.

Table S8. The model hyperparameters for OrbNet-Equi used for benchmarking studies.

| Symbol | Meaning | Defined in | Value(s) |
|--------|---------|------------|----------|
| N      | Total number of feature channels in h^l | Methods. A, Eq. (12) | 256 |
| N_{lp} | Number of feature channels for each (l, p) in h^l | Methods. A, Eq. (12) | See Table S9 |
| t_1    | Number of convolution-message-passing update steps | Methods. A | 4 |
| t_2    | Number of post-update point-wise interaction modules | Methods. A | 4 |
| I      | Number of convolution channels i | Methods. A, Eq. (15) | 8 |
| J      | Number of attention heads j | Methods. A, Eq. (17) | 8 |
| d^{MLP} | Depth of MLPs | MLPs in Eq. (24)-Eq. (26) | 2 |
| Activation function | MLPs in Eq. (24)-Eq. (26) | Swish (31) |
| N^ξ    | Number of Radial basis functions ξ | Methods. A, Eq. (21) | 16 |
| Estimation scheme for (μ, σ) for EvNorm, t < t_1 | Methods. A, Eq. (22) | BatchNorm (32) |
| Estimation scheme for (μ, σ) for EvNorm, t ≥ t_1 | Methods. A, Eq. (22) | LayerNorm (30) |
| Initialization of β_{nlp} in EvNorm layers | Methods. A, Eq. (22) | Uniform (0.5, 1.5) |
| ϵ      | Stability factor ϵ in EvNorm layers | Methods. A, Eq. (22) | 0.1 |
| Total number of parameters | - | 2.1M |
Table S9. The number of feature channels $N_{lp}$ for each representation group $(l, p)$ of $h^l$ used in this work across all values of $l$. Note that $l_{\text{max}} = 4$.

| $N_{lp}$ | $l = 0$ | $l = 1$ | $l = 2$ | $l = 3$ | $l = 4$ |
|----------|---------|---------|---------|---------|---------|
| $p = +1$ | 128     | 48      | 24      | 12      | 6       |
| $p = -1$  | 24      | 8       | 4       | 2       | 0       |

B. Training. For all training setups we use the Adam optimizer (33) with maximal learning rate $5 \times 10^{-4}$ and parameters $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-4}$. The loss function denoted as $\mathcal{L}$ below refers to a SmoothL1Loss function (34). Batch sizes and the total number of epochs are adjusted for different benchmarks to account for their vastly different training set sizes, as detailed below. No additional regularization such as weight decay or early stopping is employed.

B.1. QM9. For QM9 tasks we optimize the model using the loss $\mathcal{L}(y, y_\theta)$ for each target $y$. We use a batch size of 64 when the training sample size $> 1000$, and a batch size of 16 for smaller training sizes. We employ a learning rate schedule of first performing linear warmup for 100 epochs to the maximal learning rate followed by a cosine learning rate annealing (35) for 200 epochs. Models are trained on a single Nvidia Tesla V100-SXM2-32GB GPU, taking around 36 hours for training runs with 110k training samples.

B.2. MD17. For MD17, we optimize the model by simultaneously training on energies $E(R)$ and forces $F(R)$, using the following loss function:

$$\mathcal{L}_{E+F}(E, F; \theta) := c_E \cdot \mathcal{L}(E(R); E_\theta(R)) + c_F \cdot \frac{1}{3[A]} \sum_{\sigma} \sum_{m \in \{x, y, z\}} \mathcal{L}(\frac{\partial E_\theta(R)}{\partial R_{A,m}} - F_{A,m}(R))$$  \[S46\]

Following previous works (14, 20, 36), we set $c_E = 1$, $c_F = 1000 \text{Å}$ for training on the rMD17 labels, and $c_E = 0$, $c_F = 1000 \text{Å}$ for training on the original MD17 labels. For each molecular system, we use the 1000 geometries of the ‘train 01’ subset given by (13) for training and the 1000 geometries of the ‘test 01’ subset for testing. We use a batch size of 8, and train the model on a single Nvidia Tesla V100-SXM2-32GB GPU for 1500 epochs using a step decay learning rate schedule, taking around 36 hours for a single training run.

For proof-of-principle purposes, the gradients of tight-binding features with respect to atomic coordinates are obtained using finite difference with a 5-point stencil for each degree of freedom. The grid spacing between the stencil points are set to 0.01 Bohr. We note that in principle, this cost of evaluating and storing feature gradients can be avoided if the electronic structure method is implemented with back-propagation and the model can be trained end-to-end on both energy and force labels.

B.3. Electron densities. For electron density, we train the models on the analytic $L^2$ density loss following (37):

$$\mathcal{L}_\rho(\hat{\rho}, \hat{\rho}) := \int \| \hat{\rho}(\vec{r}) - \hat{\rho}(\vec{r}) \|^2 d\vec{r} = (\mathbf{d} - \hat{\mathbf{d}})^T \mathbf{S}_\rho (\mathbf{d} - \hat{\mathbf{d}})$$  \[S47\]

where the density coefficients $\mathbf{d} := \bigoplus_{A,u,l,m} d_{A,u,l,m}^{\text{fit}} \{ \chi \}$. A sparse COO format is used for $\mathbf{S}_\rho$ to efficiently compute $\mathcal{L}_\rho$ during batched training. We use a batch size of 64 and a cosine annealing learning schedule for training; the models are trained on a single Nvidia Tesla V100-SXM2-32GB GPU for 2000 epochs on the BFDb-SS1 dataset taking 10 hours, and for 500 epochs on the QM9 dataset taking 120 hours.

B.4. The OrbNet-Equi/SDC21 model. The training dataset (see Methods C) contains different geometries $b_\eta$ for each molecule $\eta$ in the dataset. We train on a loss function following (38):

$$\mathcal{L}_{\text{Gr}}(E(\eta, b_\eta), E_\theta(\eta, b_\eta)) := \mathcal{L}(E(\eta, b_\eta), E_\theta(\eta, b_\eta))$$  \[S48\]

where $\hat{b}_\eta$ is a geometry randomly sampled from all the geometries $\{ b_\eta \}$ of each molecule $\eta$ within each mini-batch during training. We use $c_{\text{Gr}} = 10$ in this work. We train the model for 125 epochs on a Nvidia Tesla V100-SXM2-32GB GPU using a batch size of 64 and a cosine annealing learning rate schedule taking 64 hours.

References

1. T Helgaker, P Jorgensen, J Olsen, Molecular electronic-structure theory, (John Wiley & Sons), (2014).
2. JJ Sakurai, ED Commins, Modern quantum mechanics, revised edition (1995).
3. MA Blanco, M Flórez, M Bermejo, Evaluation of the rotation matrices in the basis of real spherical harmonics. J. Mol. Struct. THEOCHEM 419, 19–27 (1997).
4. R Ramakrishnan, PO Dral, M Rupp, OA Von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules. Sci. Data 1, 1–7 (2014).
5. A Fabrizio, A Grisafi, B Meyer, M Ceriotti, C Corminboeuf, Electron density learning of non-covalent systems. *Chem. science* **10**, 9424–9432 (2019).

6. Al Molev, Gelfand-tsetlin bases for classical lie algebras. *arXiv preprint math/0211289* (2002).

7. BC Hall, Quantum theory for mathematicians. (Springer) Vol. 267, (2013).

8. W Klink, T Ton-That, Multiplicity, invariants, and tensor product decompositions of compact groups. *J. Math. Phys.* **37**, 4668–4685 (1996).

9. W Fulton, J Harris, *Representation theory: a first course*. (Springer Science & Business Media) Vol. 129, (2013).

10. A Alex, M Kalus, A Huckebleyber, J von Delft, A numerical algorithm for the explicit calculation of su (n) and sl (n, c) clesch–gordan coefficients. *J. Math. Phys.* **52**, 023507 (2011).

11. S Gilske, W Klink, T Ton-That, Algorithms for computing u (n) clesch gordan coefficients. *Acta Appl. Math.* **95**, 51–72 (2007).

12. L Lang, M Weiler, A wigner-ekart theorem for group equivariant convolution kernels. *arXiv preprint arXiv:2010.10952* (2020).

13. AS Christensen, OA von Lilienfeld, On the role of gradients for machine learning of molecular energies and forces. *Mach. Learn. Sci. Technol.* **1**, 045018 (2020).

14. S Batzner, et al., Se (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *arXiv preprint arXiv:2104.03164* (2021).

15. D Fölmsbee, G Hutchison, Assessing conformer energies using electronic structure and machine learning methods. *Int. J. Quantum Chem.* p. e26381 (2020).

16. BK Rai, et al., Torsionnet: A deep neural network to rapidly predict small-molecule torsional energy profiles with the accuracy of quantum mechanics. *J. Chem. Inf. Model.* **62**, 785–800 (2020) PMID: 35119861.

17. T Risthaus, M Steinmetz, S Grimme, Implementation of nuclear gradients of range-separated hybrid density functionals and benchmarking on rotational constants for organic molecules. *J. Comput. Chem.* **35**, 1509–1516 (2014).

18. UR Fogueri, S Kozuch, A Karton, JM Martin, The melatonin conformer space: Benchmark and assessment of wave function and df methods for a paradigmatic biological and pharmacological molecule. *The J. Phys. Chem. A* **117**, 2269–2277 (2013).

19. DG Smith, LA Burns, K Patkowski, CD Sherrill, Revised damping parameters for the d3 dispersion correction to density functional theory. *The journal physical chemistry letters* **7**, 2197–2203 (2016).

20. J Klicpera, J Groß, S Günnewann, Directional message passing for molecular graphs in *International Conference on Learning Representations (ICLR)*. (2020).

21. S Chmiela, et al., Machine learning of accurate energy-conserving molecular force fields. *Sci. advances* **3**, e1603015 (2017).

22. P Izmailov, D Podoprikhin, T Garipov, D Vetrov, AG Wilson, Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407* (2018).

23. S Chmiela, HE Sauceda, KR Müller, A Tkatchenko, Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. communications* **9**, 1–10 (2018).

24. KT Schütt, OT Unke, M Gastegger, Equivariant message passing for the prediction of tensorial properties and molecular spectra. *arXiv preprint arXiv:2102.03150* (2021).

25. Z Qiao, et al., Multi-task learning for electronic structure to predict and explore molecular potential energy surfaces. *arXiv preprint arXiv:2011.02680* (2020).

26. WJ Allen, RC Rizzo, Implementation of the hungarian algorithm to account for ligand symmetry and similarity in structure-based design. *J. chemical information modeling* **54**, 518–529 (2014).

27. Semiempirical Extended Tight-Binding Program Package, (2020, accessed July 14, 2020) [https://github.com/grimme-lab/xtb](https://github.com/grimme-lab/xtb).

28. X Gao, F Ramezanghorbani, O Isayev, JS Smith, AE Roitberg, Torchani: A free and open source pytorch-based deep learning implementation of the ani neural network potentials. *J. chemical information modeling* **60**, 3408–3415 (2020).

29. LP Wang, C Song, Geometry optimization made simple with translation and rotation coordinates. *The J. chemical physics* **144**, 214108 (2016).

30. JL Ba, JR Kiros, GE Hinton, Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

31. P Ramachandran, B Zoph, QV Le, Searching for activation functions. *arXiv preprint arXiv:1710.05941* (2017).

32. S Ioffe, C Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift in *International Conference on Machine Learning*. pp. 448–456 (2015).

33. DP Kingma, J Ba, Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

34. R Girshick, Fast r-cnn in *Proceedings of the IEEE international conference on computer vision*. pp. 1440–1448 (2015).

35. I Loshchilov, F Hutter, Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

36. K Schütt, et al., Schnet: A continuous-filter convolutional neural network for modeling quantum interactions in *Advances in neural information processing systems*. pp. 991–1001 (2017).

37. A Grisafi, et al., Transferable machine-learning model of the electron density. *ACS Cent. Sci.* **5**, 57–64 (2019).

38. Z Qiao, M Welborn, A Anandkumar, FR Manby, TF Miller III, Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *The J. Chem. Phys.* **153**, 124111 (2020).