Transformation from Molecular Orbital Coefficients to Embedding Vectors: Detailed Mathematical Analysis

EGSMole Project

September 19, 2025

1 Overview

This document provides a detailed mathematical explanation of the transformation process from molecular orbital coefficients (MO coefficients) computed by PySCF to embedding vectors (MO embeddings) that can be used in E3NN (Equivariant Neural Networks).

2 Overall Transformation Framework

The transformation from molecular orbital coefficients $\mathbf{C} \in \mathbb{R}^{N_{\text{basis}} \times N_{\text{mo}}}$ to embedding vectors $\mathbf{E} \in \mathbb{R}^{N_{\text{mo}} \times N_{\text{atom}} \times N_{\text{features}}}$ is expressed by the following equation:

$$\mathbf{E} = \text{EmbedMOsOnAtoms}(\mathbf{C}, \mathbf{E}_{\text{energy}}, \mathbf{P}_{\text{atom}}, \mathbf{S}_{\text{irrep}}) \tag{1}$$

where:

- $\mathbf{C}$: Molecular orbital coefficient matrix ($N_{\text{basis}} \times N_{\text{mo}}$)
- $\mathbf{E}_{\text{energy}}$: Molecular orbital energies (N_{mo})
- $\mathbf{P}_{\text{atom}}$: Atom type embedding information
- $\mathbf{S}_{\text{irrep}}$: Irreducible representation strings for each atom

3 Step-by-Step Transformation Process

3.1 Step 1: d-orbital Correction

Correct the difference in d-orbital ordering between PySCF and E3NN:

$$\mathbf{C}_{\text{fixed}} = \mathbf{F}_D^T \mathbf{C} \tag{2}$$

where $\mathbf{F}_D$ is the d-orbital correction matrix.

3.2 Step 2: Transposition

Transpose the molecular orbital coefficients to enable independent processing of each molecular orbital:

$$\mathbf{C}_{\text{transposed}} = \mathbf{C}_{\text{fixed}}^{T} \in \mathbb{R}^{N_{\text{mo}} \times N_{\text{basis}}} \tag{3}$$

3.3 Step 3: Atom-wise Orbital Coefficient Extraction

For each atom a, extract the coefficients corresponding to its basis functions:

$$\mathbf{C}_a = \mathbf{C}_{\text{transposed}}[:, \text{slice}_a] \tag{4}$$

where slice_a is the index range of basis functions corresponding to atom a.

3.4 Step 4: Padding Process

Pad each atom's orbital coefficients to a unified dimension. Let $\mathbf{S}_a$ be the irreducible representation of atom a and $\mathbf{n}_{\text{target}}$ be the target number of orbitals:

$$\mathbf{n}_{\text{count}} = \text{CountIrreps}(\mathbf{S}_a) \tag{5}$$

$$\mathbf{n}_{\text{diff}} = \mathbf{n}_{\text{target}} - \mathbf{n}_{\text{count}} \tag{6}$$

$$\mathbf{S}_{\text{padded}} = \mathbf{S}_a + \text{GenerateDiffIrreps}(\mathbf{n}_{\text{diff}}) \tag{7}$$

The padded orbital coefficients are:

$$\mathbf{C}_{a,\text{padded}} = \text{Pad}(\mathbf{C}_a, \mathbf{n}_{\text{diff}}) \tag{8}$$

3.5 Step 5: Energy Information Addition

When molecular orbital energies are provided, add energy information to each atom's embedding vector:

$$\mathbf{E}_{a,\text{with\_energy}} = \text{Concat}(\mathbf{E}_{\text{energy}}, \mathbf{C}_{a,\text{padded}}) \tag{9}$$

where $\mathbf{E}_{\text{energy}}$ represents the energy values of each molecular orbital.

3.6 Step 6: Atom Type Information Addition

Add one-hot encoding of atom type to each atom's embedding vector:

$$\mathbf{A}_{a,\text{one\_hot}} = \text{OneHot}(\text{AtomType}(a), N_{\text{max\_atoms}}) \tag{10}$$

The final embedding vector is:

$$\mathbf{E}_{a,\text{final}} = \text{Concat}(\mathbf{A}_{a,\text{one\_hot}}, \mathbf{E}_{a,\text{with\_energy}}) \tag{11}$$

4 Dimension Calculation

4.1 Orbital Contribution Dimension

The orbital contribution dimension for each atom is determined by the dimension of the irreducible representation after padding:

$$d_{\text{orbital}} = \sum_{l=0}^{l_{\max}} n_l \cdot (2l+1) \tag{12}$$

where:

- n_l: Number of orbitals with angular momentum l
- $2l+1$: Degeneracy of orbitals with angular momentum l

4.2 Energy Dimension

When molecular orbital energies are provided:

$$d_{\text{energy}} = N_{\text{mo}} \tag{13}$$

4.3 Atom Type Dimension

One-hot encoding of atom types:

$$d_{\text{atom\_type}} = N_{\text{max\_atoms}} \tag{14}$$

4.4 Total Dimension

The dimension of the final embedding vector is:

$$N_{\text{features}} = d_{\text{atom\_type}} + d_{\text{orbital}} + d_{\text{energy}} \tag{15}$$

5 Concrete Example: Methane Molecule (CH_4)

5.1 Input Data

- Molecular orbital coefficients: $\mathbf{C} \in \mathbb{R}^{34\times34}$
- Molecular orbital energies: $\mathbf{E}_{\text{energy}} \in \mathbb{R}^{34}$
- Basis set: def2-svp

5.2 Irreducible Representations for Each Atom

$$\text{Carbon atom}: \quad \mathbf{S}_0 = \text{1x0e+1x0e+1x0e+1x1o+1x1o+1x2e} \tag{16}$$

$$\text{Hydrogen atoms}: \quad \mathbf{S}_i = \text{1x0e+1x0e+1x1o} \quad (i = 1, 2, 3, 4) \tag{17}$$

5.3 Padding Process

Target number of orbitals: $\mathbf{n}_{\text{target}} = [3, 2, 1]$ (s, p, d orbitals)

$$\text{Carbon atom}: \quad \mathbf{n}_{\text{count}} = [3, 2, 1], \quad \mathbf{n}_{\text{diff}} = [0, 0, 0] \tag{18}$$

$$\text{Hydrogen atoms}: \quad \mathbf{n}_{\text{count}} = [2, 1, 0], \quad \mathbf{n}_{\text{diff}} = [1, 1, 1] \tag{19}$$

5.4 Dimension Calculation

$$d_{\text{orbital}} = 3 \times 1 + 2 \times 3 + 1 \times 5 = 14 \tag{20}$$

$$d_{\text{energy}} = 34 \tag{21}$$

$$d_{\text{atom\_type}} = 0 \quad \text{(when energies are included)} \tag{22}$$

$$N_{\text{features}} = 0 + 14 + 34 = 48 \tag{23}$$

5.5 Final Output

$$\mathbf{E} \in \mathbb{R}^{34 \times 5 \times 48} \tag{24}$$

6 Mathematical Properties

6.1 Rotation Invariance

The embedding vectors possess rotation-invariant properties with respect to molecular rotations:

$$\mathbf{E}(R\mathbf{M}) = \mathbf{E}(\mathbf{M}) \tag{25}$$

where R is a rotation matrix and $\mathbf{M}$ is the molecular geometry.

6.2 Permutation Invariance

Invariant with respect to permutation of identical atoms:

$$\mathbf{E}(\sigma(\mathbf{M})) = \sigma(\mathbf{E}(\mathbf{M})) \tag{26}$$

where σ represents permutation of identical atoms.

7 Implementation Considerations

7.1 Memory Efficiency

For large molecules, the embedding vector size becomes significant, making memory-efficient implementation crucial.

7.2 Numerical Stability

In the padding process, it is necessary to ensure that zero-padding is numerically stable.

7.3 Parallel Processing

Since each atom's processing is independent, parallelization is possible.

8 Conclusion

The transformation from molecular orbital coefficients to embedding vectors is a crucial process that converts quantum chemistry calculation results into a format suitable for machine learning. This transformation enables efficient processing of molecular electronic structure information in E3NN.

Each stage of the transformation process is mathematically well-defined, properly converting molecular orbital information into embedding vectors while preserving rotation invariance and permutation invariance.