Molecular Energy Project Slides

Authors : Ayoub CHOUKRI , Axel OLOUGOUNA

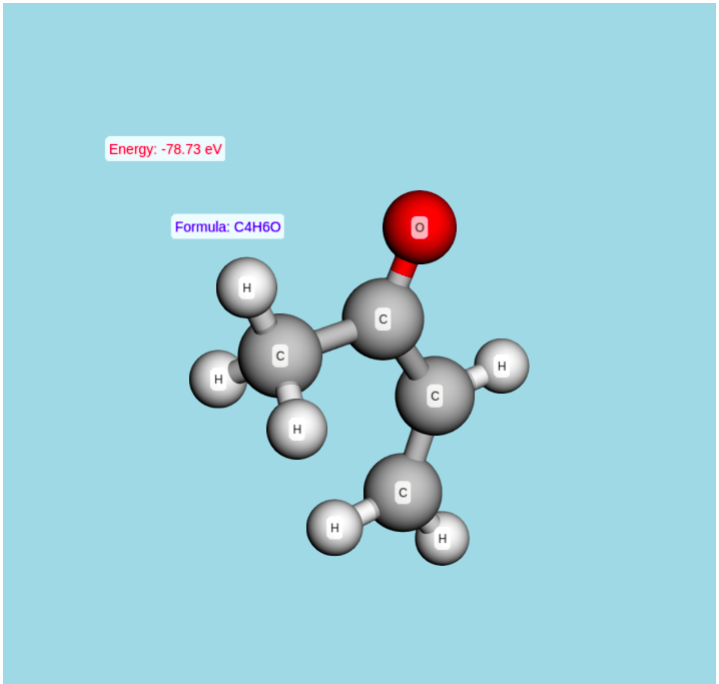
June 27, 2025

Outline

Table of Contents

QM7-X Dataset

- **Description**: 8300 small organic molecules with atomic positions and atomization energies.
- **Format**: `xyz` files with atomic symbols and 3D coordinates (x, y, z).
- **Preprocessing**: The dataset was preprocessed so that for each molecule we have:
  - a vector of atomic symbols, $\mathrm{S} = [S_1, S_2, \ldots, S_A]$,
  - the atomic positions $\mathrm{r} = [\mathrm{r}_1, \mathrm{r}_2, \ldots, \mathrm{r}_A]$,
  - and the atomization energy of the molecule E.
- **Example Visualization**:



Objectives and Approaches

- **Objective**: Predict atomization energy $E(\mathrm{r})$ of small organic molecules using 3D atomic positions.
  - **Translation invariance**:

$$E(\{\mathrm{r}_i + \mathrm{t}\}_{i=1}^{A}) = E(\{\mathrm{r}_i\}_{i=1}^{A}), \quad \forall \mathrm{t} \in \mathbb{R}^3$$

  - **Rotation invariance**:

$$E(\{R\mathrm{r}_i\}_{i=1}^{A}) = E(\{\mathrm{r}_i\}_{i=1}^{A}), \quad \forall R \in SO(3)$$

  - **Permutation invariance**:

$$E(\{\mathrm{r}_{\pi(i)}\}_{i=1}^{A}) = E(\{\mathrm{r}_i\}_{i=1}^{A}), \quad \forall \pi \in Sym(A)$$

- **Approaches**:
  - 3D Wavelet Scattering Transform
  - Transformer-based model with invariant features

Table of Contents

3D Wavelet Scattering: Main Idea

- **Concept**: Use 3D wavelet scattering to extract invariant features from atomic positions.
- **Invariance**: Features are invariant to translation, rotation, and permutation.
- **Process**:
  - Preprocess atomic data (nuclear charges, valence charges, scaled positions).
  - Construct electron density functions: $\rho_{\text{full}}, \rho_{\text{val}}, \rho_{\text{core}}$.
  - Compute scattering coefficients: zeroth-order, first-order, and second-order.
- **Feature Vector**: Concatenation of scattering coefficients for regression.

$$\mathrm{F} = \left(S_0(\rho_{\text{full}}), S_1(\rho_{\text{full}}), S_2(\rho_{\text{full}}),\ S_0(\rho_{\text{val}}), S_1(\rho_{\text{val}}), S_2(\rho_{\text{val}}),\ S_0(\rho_{\text{core}}), S_1(\rho_{\text{core}}), S_2(\rho_{\text{core}})\right)$$

To compute these invariant features, a data preprocessing step is required.

Data Preprocessing

- **Nuclear Charges**: Map atomic symbols to atomic numbers Z_i.

$$\mathrm{Z} = [Z_1, Z_2, \ldots, Z_A], \quad Z_i \in \{1, 6, 7, 8, 16, 17\}$$

- **Valence Charges**: Estimate valence electrons based on Z_i.

$$v_i = \begin{cases} Z_i & \text{if } Z_i \leq 2 \\ Z_i - 2 & \text{if } 2 < Z_i \leq 10 \\ Z_i - 10 & \text{if } 10 < Z_i \leq 18 \end{cases}$$

$$\mathrm{v} = [v_1, v_2, \ldots, v_A]$$

- **Position Scaling**: Scale positions using minimum interatomic distance.

$$\mathrm{r}'_i = \mathrm{r}_i \cdot \frac{\delta}{d_{\text{min}}}, \quad \delta = \sigma\sqrt{-8\ln(\epsilon)}$$

$$R' = [\mathrm{r}'_1, \mathrm{r}'_2, \ldots, \mathrm{r}'_A]$$

- **Padding**: Uniform size with $A_{\text{max}} = 23$.

$$\mathrm{Z} = [Z_1, Z_2, \ldots, Z_{A_{\mathbf{max}}}], \quad \mathrm{v} = [v_1, v_2, \ldots, v_{A_{\mathbf{max}}}], \quad R' = [\mathrm{r}'_1, \mathrm{r}'_2, \ldots, \mathrm{r}'_{A_{\mathbf{max}}}]$$

Scattering Transform

- **Electron Density**:

$$\rho_{\text{full}}(\mathbf{r}) = \sum_{i=1}^{A} Z_i \exp\left(-\frac{|\mathbf{r} - \mathbf{r}_i'|^2}{2\sigma^2}\right)$$

$$\rho_{\text{val}}(\mathbf{r}) = \sum_{i=1}^{A} v_i \exp\left(-\frac{|\mathbf{r} - \mathbf{r}_i'|^2}{2\sigma^2}\right)$$

$$\rho_{\text{core}}(\mathbf{r}) = \rho_{\text{full}}(\mathbf{r}) - \rho_{\text{val}}(\mathbf{r}) = \sum_{i=1}^{A} (Z_i - v_i) \exp\left(-\frac{|\mathbf{r} - \mathbf{r}_i'|^2}{2\sigma^2}\right)$$

- **Scattering Coefficients**:

  For each $k \in \{\text{full}, \text{val}, \text{core}\}$, we compute $S_0(\rho_k)$, $S_1(\rho_k)$, and $S_2(\rho_k)$.

Scattering Coefficients: Formules

- **Zeroth-order**:

$$S_0(\rho) = \int_{\mathbb{R}^3} |\rho(\mathrm{r})|^p \, d\mathrm{r}$$

- **First-order**:

$$S_1(\rho; j) = \frac{1}{2L} \sum_{k=0}^{2L-1} \int_{\mathbb{R}^3} |\rho * \psi_{j,k}(\mathrm{r})|^p \, d\mathrm{r}$$

where $*$ denotes convolution, and $\psi_{j,k}$ is a 3D wavelet at scale j and rotation k.

- **Second-order**:

$$S_2(\rho; j_1, l_1, j_2, l_2) = \frac{1}{2L} \sum_{k=0}^{2L} \int_{\mathbb{R}^3} ||\rho * \psi_{j_1, l_1 - k}| * \psi_{j_2, l_2 - k}(\mathrm{r})| \, d\mathrm{r}$$

avec $j_2 \geq j_1 + 1$.

The coefficients are concatenated to form the feature vector used for regression.

Coulomb Matrix

- **Construction**: Symmetric matrix encoding electrostatic interactions.

$$C_{ij} = \begin{cases} 0.5 Z_i^{2.4} & \text{if } i = j \\ \frac{Z_i Z_j}{|\mathbf{r}_i' - \mathbf{r}_j'|} & \text{if } i \neq j \end{cases}$$

- **Invariance**: Sorted eigenvectors ensure translational, rotational, and permutational invariance.
- **Padding**: Eigenvectors padded to $A_{\text{max}} = 23$.
- **Feature Fusion**: The eigenvectors are concatenated with the scattering features to form the input vector for the regression model.

Ridge Regression

- **Model**: Linear regression with L2 regularization.

$$\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta^T \mathrm{x}_i \right)^2 + \alpha \|\beta\|_2^2$$

- **Cross-Validation**: 15-fold CV to select optimal α.
- **Results**:
  - Optimal $\alpha = 0.008498$, Mean MSE ≈ 0.024213.
  - Kaggle score: 0.125, indicating strong performance.

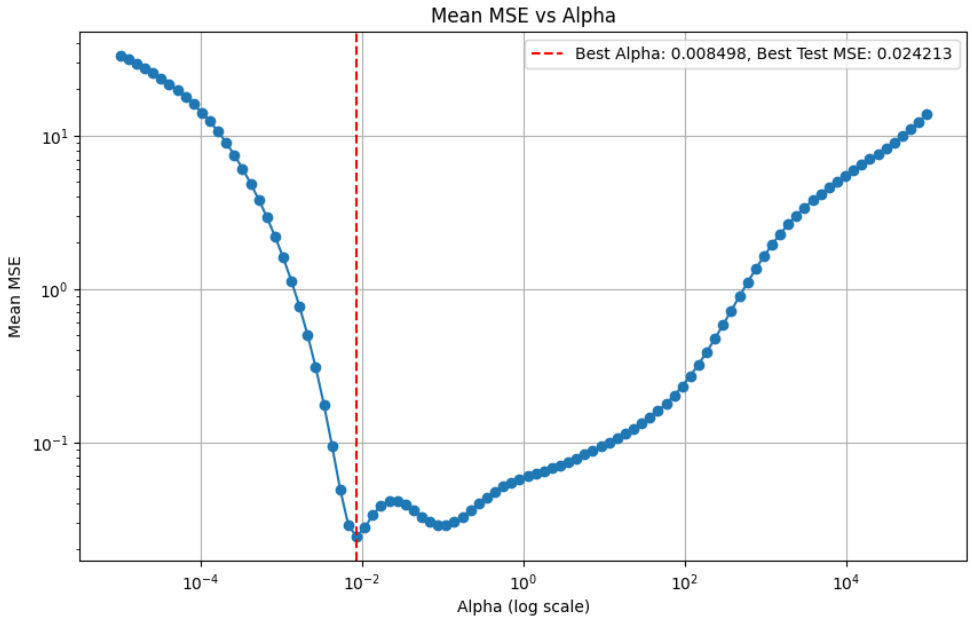


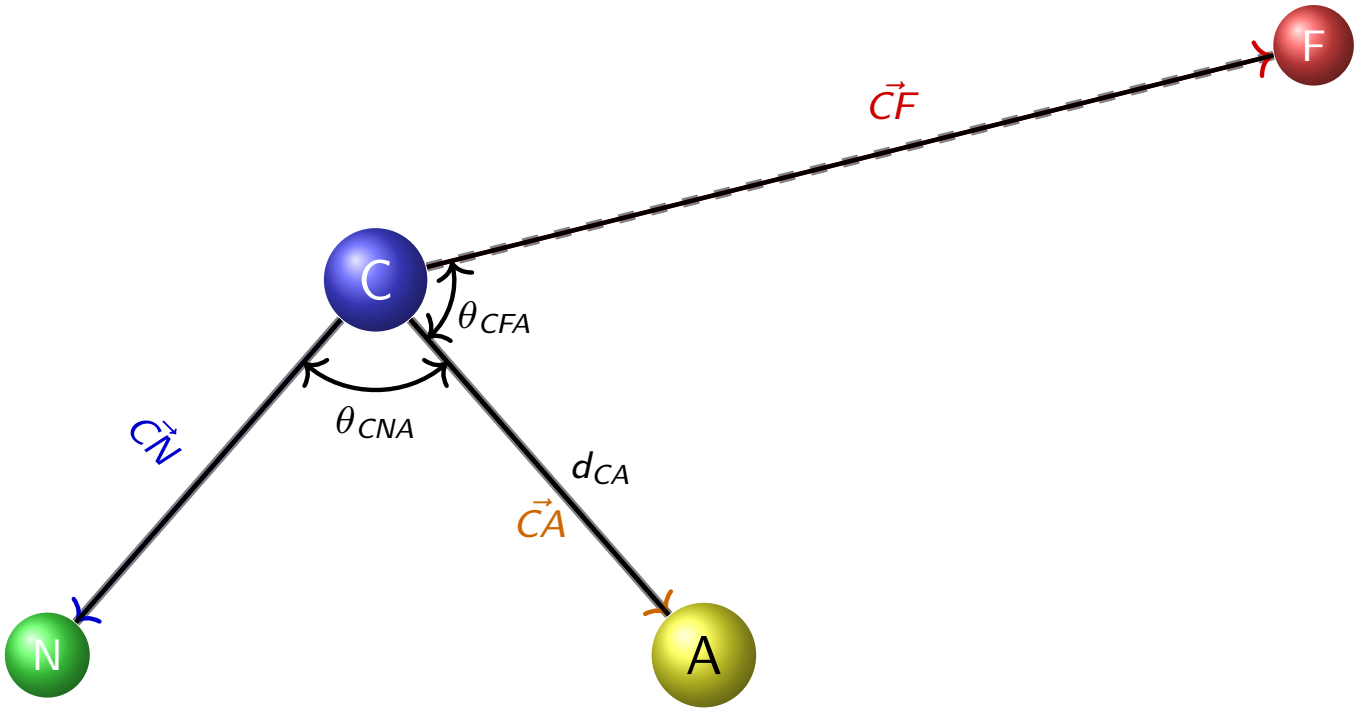
Figure: Mean CV MSE vs. α.

Table of Contents

Transformers: Main Idea

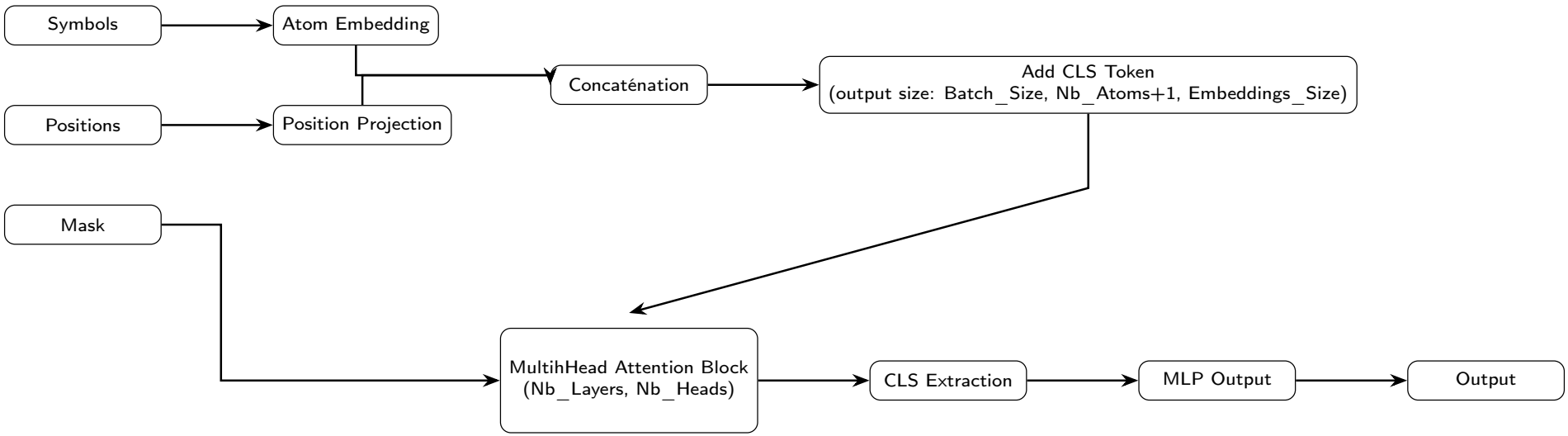
- **Concept**: Use Transformer to model variable-length molecular data (QM7-X dataset).
- **Input**:
  - Atom types (e.g., H, C, N, O, F) and 3D positions.
  - Learned embeddings capturing atomic types and geometric relations (distances, angles).
- **Key Advantage**:
  - Handles variable-size molecules via self-attention.
  - Ensures permutational invariance: model output independent of atom order.
- **Data Transformation**:
  - Compute invariant features to respect translation, rotation, and permutation symmetries.
- **Goal**: Predict molecular energies by learning atomic configuration contributions.

Data Transformation



- **Features**: $\|\vec{CA}\|$, θ_{CAF}, θ_{CAN} (signed angles).
- **Invariance**: This ensures Translation and rotation invariant.

Transformer Architecture



- **Components**: Atom embedding, position projection, CLS token, multi-head attention, MLP output.
- **Hyperparameters**: Embedding size = 1024, 30 heads, 1 attention block.
- **Number Of Parameters**: 163000000 parameters.

Permutation Invariance

- **Definition**: Model predictions are unchanged when the order of input atoms is permuted.
- **Implementation**:
  - Atoms treated as independent tokens, no positional encoding.
  - Atoms sorted by distance from the center atom to ensure consistent input structure.
- **Impact**: Predictions depend only on relative geometric information, not on arbitrary atom ordering.

Translation Invariance

- **Definition**: Model predictions remain consistent under spatial translations of the molecule.
- **Mathematical Proof**:
  - Distance $\|\vec{CA}\| = \sqrt{(x_A - x_C)^2 + (y_A - y_C)^2 + (z_A - z_C)^2}$ is invariant as translation terms cancel.
  - Angles θ_{CAF}, θ_{CAN} depend on vectors unchanged by translation ($\vec{CA}' = \vec{CA}$).
  - Nearest and furthest atoms remain the same as distances $\|\vec{Ci}\|$ are preserved.
- **Impact**: Model is invariant to translations, ensuring consistent predictions regardless of molecular position in space.

Rotation Invariance

- **Definition**: Model predictions are invariant to rotations of the molecule around any axis.
- **Mathematical Proof**:
  - Norm invariance: $\|\vec{CA}'\| = \|R\vec{CA}\| = \|\vec{CA}\|$, since R is an orthogonal matrix ($R^T R = I$).
  - Dot product invariance: For rotated vectors $\vec{CA}' = R\vec{CA}$, $\vec{CF}' = R\vec{CF}$,

$$\vec{CA}' \cdot \vec{CF}' = (R\vec{CA})^T (R\vec{CF}) = \vec{CA}^T R^T R\vec{CF} = \vec{CA}^T \vec{CF} = \vec{CA} \cdot \vec{CF}$$

  - Angles θ_{CAF}, θ_{CAN} preserved since $\cos(\theta'_{CAF}) = \cos(\theta_{CAF})$.
  - Nearest and furthest atoms remain consistent as norms $\|\vec{CA}\|$ and $\|\vec{Ci}\|$ are invariant.
- **Impact**: Model ensures consistent predictions regardless of molecular orientation.

Training and Results

- **Training**: Adam optimizer, learning rate 3×10^{-5}, 450 epochs.
- **Results**: Kaggle score = 0.306, indicating good generalization.
- **Loss Evolution**:

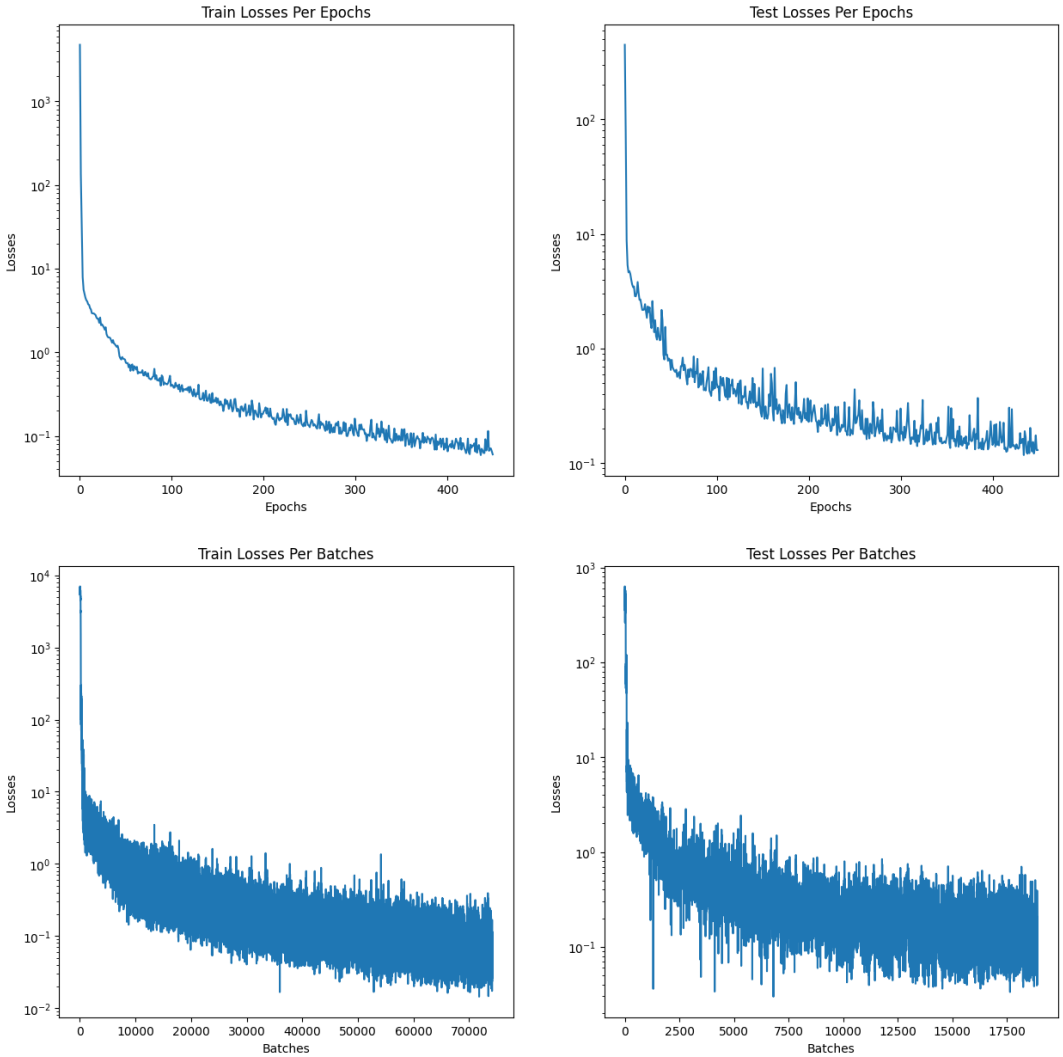


Table of Contents

Conclusion

- **Summary**:
  - 3D Wavelet Scattering: High-quality invariant features, Kaggle score 0.125.
  - Transformer: Invariant geometric features, Kaggle score 0.311.
- **Key Insights**: Both approaches ensure invariance, with Transformers excelling in modeling complex interactions.
- **Future Work**: Combine GNN and Transformers, explore advanced architectures (Path Advanced Graph Neural Networks).
- **Reference**: arXiv:1905.12712