PCA of Atomic Descriptors and Similarity to Reference Structure

1. Per-Atom Descriptor Matrix

Let the per-atom descriptor matrix for a structure be:

$$\mathbf{D} = \begin{bmatrix} \mathbf{d}_1^\top \\ \mathbf{d}_2^\top \\ \vdots \\ \mathbf{d}_N^\top \end{bmatrix} \in \mathbb{R}^{N \times n}$$

where each row $\mathbf{d}_i \in \mathbb{R}^n$ is a descriptor for atom i, N is the number of atoms in the structure, and n is the dimensionality of each per-atom descriptor.

2. Principal Component Analysis (PCA)

We apply PCA to reduce the descriptor dimensionality from n to k, where $k \ll n$. The PCA projection matrix is:

$$\mathbf{P} \in \mathbb{R}^{n \times k}$$

Assuming column-wise centering of $\mathbf{D}$, the PCA-transformed matrix is:

$$\mathbf{Z} = \mathbf{D} \cdot \mathbf{P} \in \mathbb{R}^{N \times k}$$

Each row $\mathbf{z}_i \in \mathbb{R}^k$ represents the PCA-reduced descriptor of atom i.

3. Averaging Over Atoms

To obtain a single vector representing the entire structure, we compute the mean of the PCA-transformed descriptors across atoms:

$$\bar{\mathbf{z}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \in \mathbb{R}^k$$

4. Reference Structure

Repeat the same process for a reference structure, resulting in its mean PCA descriptor:

$$\bar{\mathbf{z}}_{\text{ref}} \in \mathbb{R}^k$$

5. Similarity Metrics

We can now compare the mean PCA vector $\bar{\mathbf{z}}$ of a given structure to the reference vector $\bar{\mathbf{z}}_{\text{ref}}$ using different similarity or distance metrics.

Cosine Similarity

$$s = \frac{\bar{\mathbf{z}}^\top \bar{\mathbf{z}}_{\text{ref}}}{\|\bar{\mathbf{z}}\| \cdot \|\bar{\mathbf{z}}_{\text{ref}}\|}$$

Hellinger Distance

$$s = \frac{1}{\sqrt{2}} \sqrt{\sum_{j=1}^{k} \left(\sqrt{\bar{z}_j} - \sqrt{\bar{z}_{\text{ref},j}} \right)^2}$$

6. Summary Flow

$$\mathbf{D} \in \mathbb{R}^{N \times n} \xrightarrow{\text{PCA on } n\text{-space}} \mathbf{Z} \in \mathbb{R}^{N \times k} \xrightarrow{\text{Mean over atoms}} \bar{\mathbf{z}} \in \mathbb{R}^{k} \xrightarrow{\text{Similarity to reference}} s \in \mathbb{R}$$