**Data Analysis & Feature Engineering**

Our dataset consists of **metal-organic frameworks (MOFs)** obtained from the **QMOF database**, where **DFT simulations** were conducted for **~10k–20k MOFs** at four different accuracy levels: *PBE, HLE17, HSE06, and HSE06*\*. Each property in the database is reported at all four DFT levels, and our target property is the **band gap**.

**Key Insights from Exploratory Data Analysis (EDA)**

• Band gap distributions vary significantly across DFT levels. The **HSE06\*** and **HSE06** levels yielded **higher and more consistent band gaps**, making them the **preferred targets for ML**.

• Shell character (open vs. closed) impacts band gap predictions at lower DFT levels, highlighting the **inadequacy of PBE and HLE17** for some materials.

• **System size** did **not** introduce **noticeable bias**—smaller MOFs (used at higher DFT levels) showed similar structural and band gap distributions as larger MOFs.

**Feature Engineering & Dimensionality Reduction**

We explored four major feature sets:

• **Chemical composition-based**: stoich\_45, stoich\_120

• **Structure-sensitive**: sine\_coulomb\_matrix, orbital\_field\_matrix

All feature sets were analyzed using PCA to retain ≥90% variance. The sine Coulomb matrix was transformed from eigenvalue lists into a density representation, improving performance.

Predictive evaluations (via PCA + linear regression) showed that:

• **stoich\_45 and sine Coulomb matrix density features performed best**

• Combining stoich\_45 with stoich\_120 didn’t add value

• Analysis was limited to linear relationships—nonlinear dependencies were not captured at this stage

We refined stoich\_45 using:

• **Recursive Feature Elimination (RFE)**

• **Random Forest Feature Importance**

• **Lasso (L1) Regularization**

The **intersection of all three methods** gave us a reduced **23-feature set**, while PCA reduction yielded an alternative **8-feature set**.

We narrowed down to three optimal feature sets:

1. **PCA-reduced stoich\_45 (8 features)**

**2. Feature-selected stoich\_45 (23 features)**

**3. Transformed sine Coulomb matrix density representation**

**(+ an optional combined version)**