Data Correlation

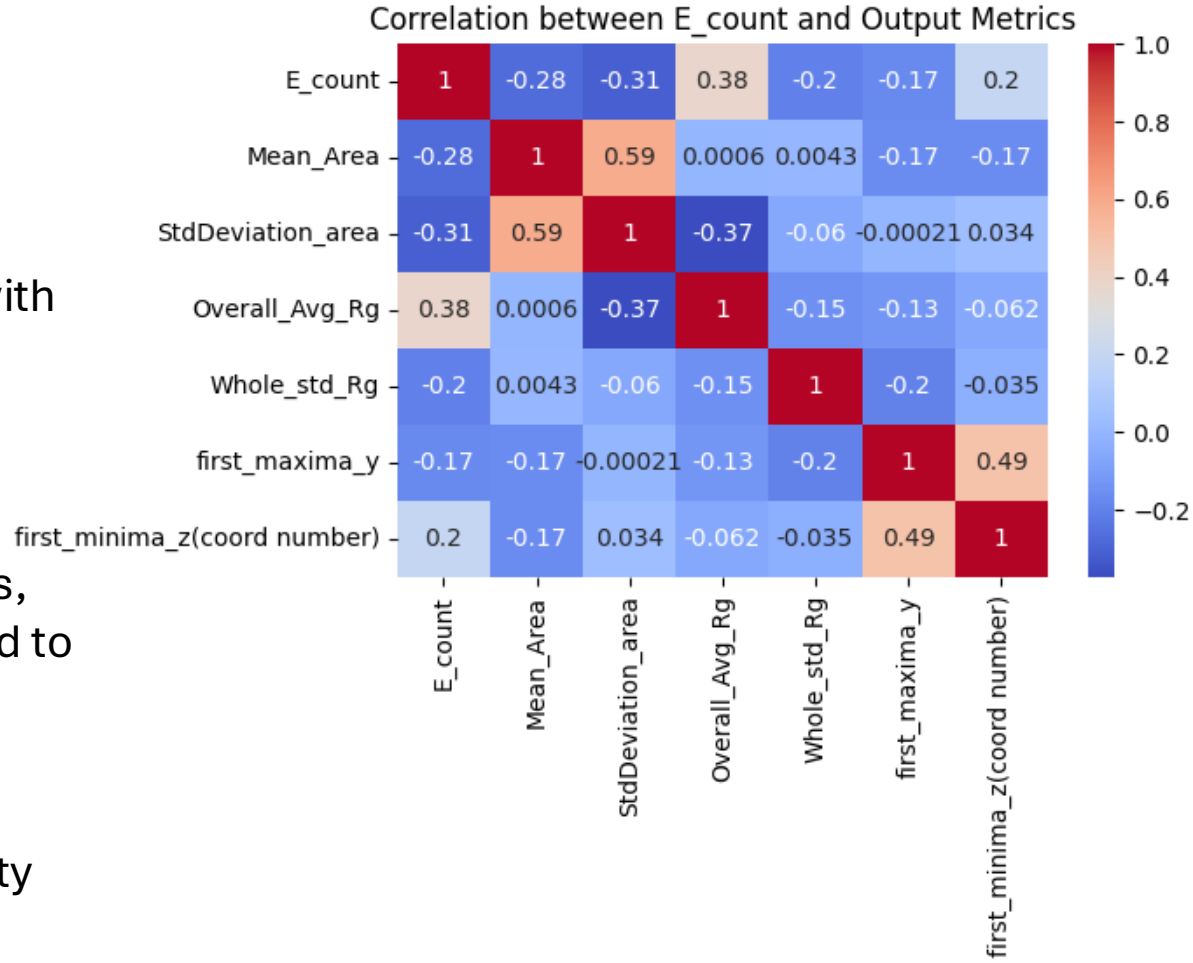
03/07/2025

Approach

- Collected data from the polymer files, and performed analysis.
- Modified the inputlist from string to a list in python with real values.
- Label Encoding for E and S values and normalizing the data to length of 20.
- Sum the values in the input vector and calculate the correlation between input and output metrics
- corr_agg = df[['S_count'] + output_cols].corr()

Research results

- At E = 1, S = -1
- `E_count` and `Overall_Avg_Rg` have a moderate positive correlation (**0.38**), indicating that polymers with more "E" labels tend to have higher average radius of gyration.
- `E_count` negatively correlates with `StdDeviation_area` (**-0.31**) and `Mean_Area` (**-0.28**). This suggests that as the count of "E" labels increases, both the mean and variability of the polymer area tend to decrease.
- `Mean_Area` and `StdDeviation_area` show a strong positive correlation (**0.59**), suggesting that higher average areas are associated with increased variability in area measurements.



Future work

- Analysis of the polymer simulations.
- Run 30 polymer simulations.
- Study the GNN based on a preexisting dataset of research papers.
- Used sequence charge decoration for E & S values.
- Modify the correlation values table, take the values with respect to the length of the input list.