CHE - 657

ENDTERM PROJECT

GROUP -16

Objective

The primary aim of this model is to predict the final yield of Dreher–Doyle dataset containing Buchwald–Hartwig C–N cross-coupling reactions by leveraging the SMILES representations of the four central molecular components: the **Ligand, Additive, Base, and Aryl halide** and **yield**. The prediction framework combines a **Neural Network** with a **ChemBERTa** transformer to learn chemical structure–yield relationships effectively.

Dataset Overview

We have a dataset containing 3955 rows and 5 columns - **Ligand, Additive, Base, and Aryl halide SMILES and yield data.** These reactions were carried at constant conditions

Temperature - 100 °C , Solvent - Dioxane/Toluene

	Ligand	Additive	Base	Aryl halide	Output
0	CC(C)C(C=C(C(C)C)C=C1C(C)C)=C1C2=C(P([C@@]3(C[...	CC1=CC(C)=NO1	CN(C)P(N(C)C)(N(C)C)=NP(N(C)C)(N(C)C)=NCC	ClC1=NC=CC=C1	70.410458
1	CC(C)C(C=C(C(C)C)C=C1C(C)C)=C1C2=C(P([C@@]3(C[...	O=C(OC)C1=CC=NO1	CN(C)P(N(C)C)(N(C)C)=NP(N(C)C)(N(C)C)=NCC	BrC1=NC=CC=C1	11.064457
2	CC(C)C(C=C(C(C)C)C=C1C(C)C)=C1C2=C(P(C3CCCCC3)...	O=C(OC)C1=CC=NO1	CN(C)P(N(C)C)(N(C)C)=NP(N(C)C)(N(C)C)=NCC	IC1=CC=C(CC)C=C1	10.223550

Innovation Aspect / Purpose

This model acts as a **high-speed computational screening tool.** Instead of chemists spending a lot of time running thousands of physical lab experiments to find the best reaction, this model predicts the yield for all those combinations in minutes. This allows researchers to **instantly identify** the most promising, **high-yield reactions** and focus their lab work only on the candidates most likely to succeed in industrial-scale production, **saving significant time, cost, and resources.**

Methodology

1.) Data Cleaning & Feature Engineering

- Load the Excel dataset containing 3955 rows and 5 columns .
- Remove invalid rows (missing data, etc).
- Combine the four component SMILES strings into a single "reaction SMILES" string.

2.) SMILES to Embeddings (ChemBERTa)

- Use the pre-trained ChemBERTa transformer to convert each "reaction SMILES" string into a 384-dimension numerical vector (embedding).

Methodology

3.) Train/Test Split

- Split the complete dataset (embeddings and yields) into a Training Set (80%) and a Test Set (20%).
- The Test Set is locked away for the final evaluation.

4.) Preprocessing Pipeline (PCA)

- Fit a StandardScaler and PCA (set to 95% variance) only on the Training Set.
- Transform both the Training Set and the Test Set using the fitted scaler and PCA.

Methodology

5.) Hyperparameter Tuning (Optuna)

- Define the Neural Network architecture (MLP Regressor).
- Use Optuna to automatically test hundreds of combinations (e.g., learning rate, layer size, dropout) to find the best-performing model.
- This tuning is done only on the Training Set.

Hyperparameter Search Space

```
# Define hyperparameter search space
lr = trial.suggest_float("lr", 1e-4, 1e-2, log=True)
batch_size = trial.suggest_categorical("batch_size", [32, 64, 128])
weight_decay = trial.suggest_float("weight_decay", 1e-5, 1e-3, log=True)
hidden_dim1 = trial.suggest_categorical("hidden_dim1", [256, 512, 768])
hidden_dim2 = trial.suggest_categorical("hidden_dim2", [128, 256, 512])
hidden_dim3 = trial.suggest_categorical("hidden_dim3", [64, 128, 256])
dropout1 = trial.suggest_float("dropout1", 0.1, 0.5)
dropout2 = trial.suggest_float("dropout2", 0.1, 0.4)
dropout3 = trial.suggest_float("dropout3", 0.0, 0.3)
```

Methodology

Using a neural network model configured with the hyperparameters optimized via Optuna.

The results of the Optuna hyperparameter tuning were

Best Hyperparameters:
- **Learning rate:** 0.00358
- **Batch size:** 128
- **Weight decay:** 4.08e-5
- **Hidden layer sizes:** 768, 128, 256
- **Dropout rates:** 0.106, 0.275, 0.223

Methodology

6.) K-Fold Cross-Validation

- To verify the best model's stability, perform a 5-fold cross-validation only on the Training Set.
- This gives a reliable average R2 score and confirms the model isn't just lucky.

7.) Final Model Training & Evaluation

- Train one final Neural Network using the best hyperparameters on the entire Training Set.
- Make final predictions on the unseen Test Set.
- Report the R2 and RMSE scores from this final test as the true model performance.

Results

1.) K - FOLD CROSS VALIDATION SUMMARY

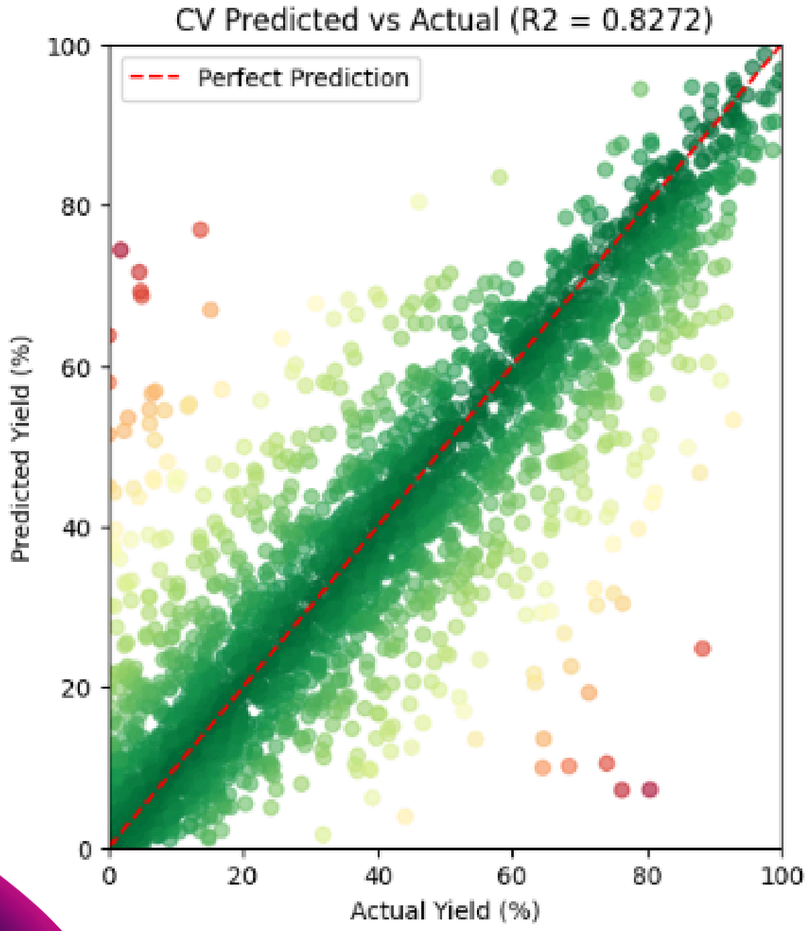
K-FOLD CROSS-VALIDATION SUMMARY

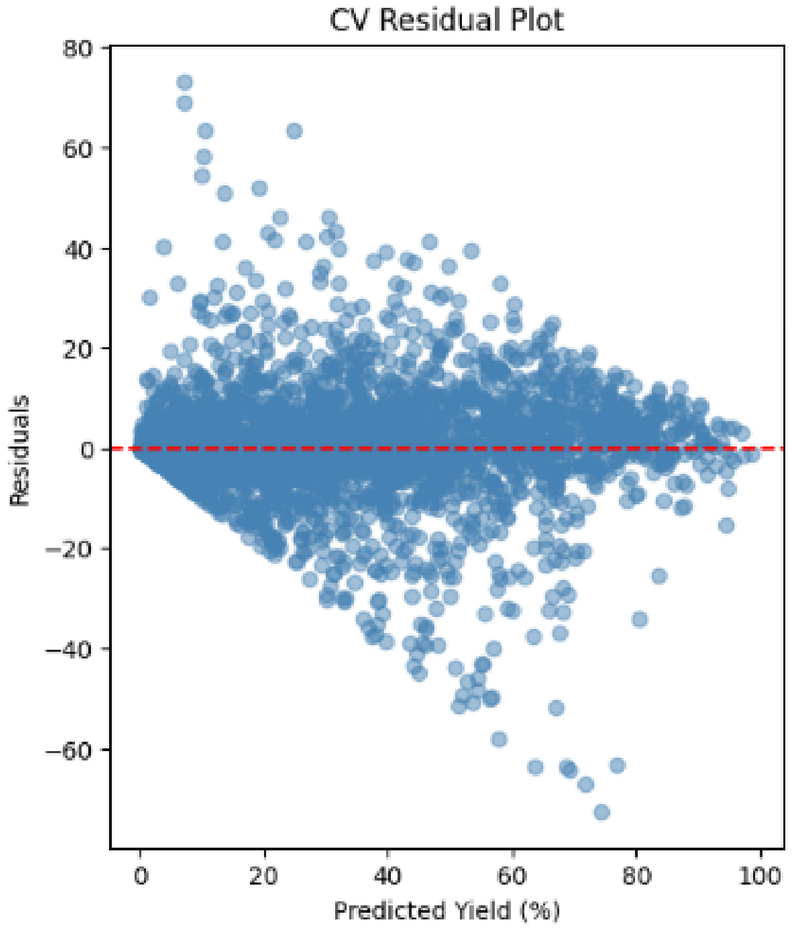
==========

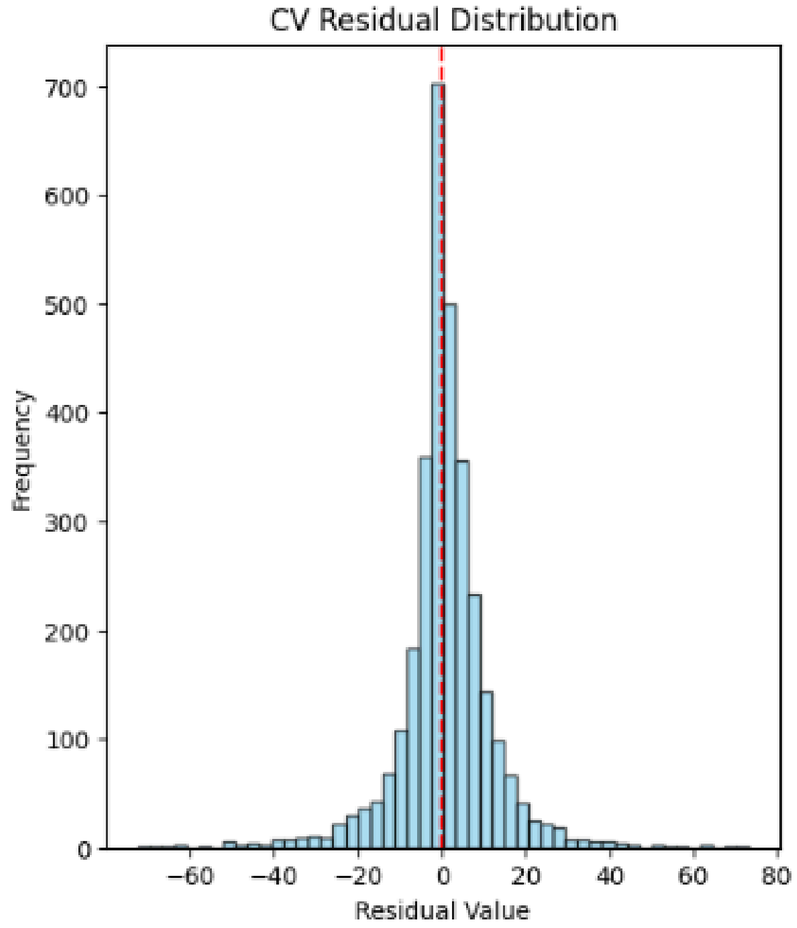
	Fold	R2 Score	MAE	RMSE
0	Fold 1	0.8633	6.6011	10.2406
1	Fold 2	0.8277	7.1933	11.3559
2	Fold 3	0.8163	7.4403	11.6741
3	Fold 4	0.7958	7.5279	12.2818
4	Fold 5	0.8316	6.9625	10.7973
5	Average	0.8269	7.1450	11.2700
6	Std Dev	0.0220	0.3362	0.7039

Results

2.) CROSS VALIDATION ANALYSIS PLOTS







Results

3.) TEST SET RESULTS

```
============================================================
FINAL HOLDOUT TEST SET METRICS
============================================================
R2 Score                : 0.8190
MSE                     : 139.4759
RMSE                    : 11.8100
MAE                     : 7.5544
sMAPE                   : 46.74%
Explained Variance      : 0.8235
Max Error               : 61.7898
============================================================
```

Conclusion

To conclude, our work shows that an optimized neural network, the model is able to learn meaningful structure–yield relationships and generalize to unseen reactions.

The model can screen thousands of reaction combinations within minutes, allowing chemists to focus their experimental effort only on high-potential reactions, reducing both time and resource consumption. Ultimately, demonstrating how machine learning can work hand-in-hand with synthetic chemistry to accelerate discovery.