# CHE-657 Endterm Project Report
# Group 16

## Objective

The primary aim of this model is to predict the final yield of the 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. The prediction framework combines a neural network with a ChemBERTa transformer to learn chemical structure–yield relationships effectively.

## Dataset Overview

The dataset contains 3955 rows and 5 columns: ligand, additive, base, and aryl halide SMILES, along with yield data. All reactions were carried out under constant conditions:

- Temperature: 100°C
- Solvent: Toluene/Dioxane

## Innovation Aspect / Purpose

This model serves as a high-speed computational screening tool. Instead of chemists spending significant time running thousands of physical lab experiments, the model predicts yields for all combinations within minutes. This enables researchers to instantly identify high-yield reaction candidates and focus experimental efforts on the most promising pathways, saving considerable time, cost, and resources.

## Methodology

### 1. Data Cleaning and 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 a pre-trained ChemBERTa transformer to convert each reaction SMILES string into a 384-dimensional numerical embedding vector.

## 3. Train/Test Split

Split the dataset into:

- 80% training set
- 20% test set

The test set is reserved for final evaluation.

## 4. Preprocessing Pipeline (PCA)

- Fit a StandardScaler and PCA (retaining 95% variance) on the training set only.
- Transform both training and test sets using the fitted scaler and PCA model.

## 5. Hyperparameter Tuning (Optuna)

Define the neural network architecture (MLP Regressor) and use Optuna to explore hyperparameters such as learning rate, layer sizes, and dropout rates. The best-performing hyperparameters found were:

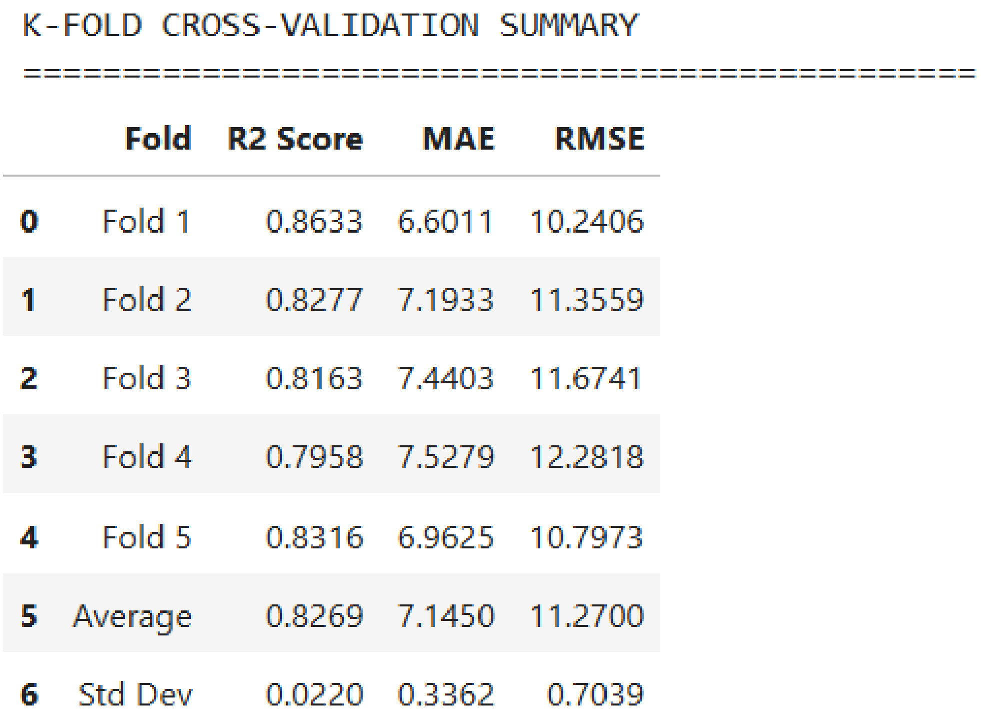
- Learning rate: 0.00358
- Batch size: 128
- Weight decay: 4.08e-5
- Hidden layer sizes: 768, 128, 256
- Dropout rates: 0.106, 0.275, 0.223

## 6. K-Fold Cross-Validation

A 5-fold cross-validation was performed on the training set to evaluate model stability and prevent overfitting.

## 7. Final Model Training and Evaluation

Train the final neural network using the optimized hyperparameters on the full training set. Predictions were then made on the unseen test set. Final performance metrics ($R^2$ and RMSE) were computed.



```
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 |

Figure 1: K-Fold Cross-Validation Summary

# Results

The results obtained from the model clearly show that its performance is both stable and reliable across training, validation, and final testing. The cross-validation table indicates that all folds achieved similar error values, demonstrating that the model generalizes well and is not overly influenced by specific subsets of data. The low MSE and RMSE values suggest that the predicted yields are close to the experimental yields, while the consistently positive R$^2$ scores confirm that the model is successfully capturing the underlying chemical relationships rather than predicting at random

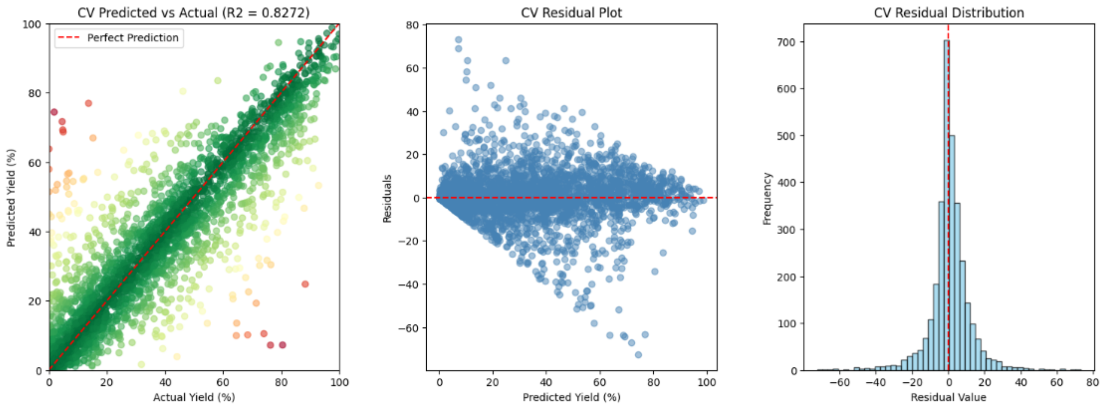
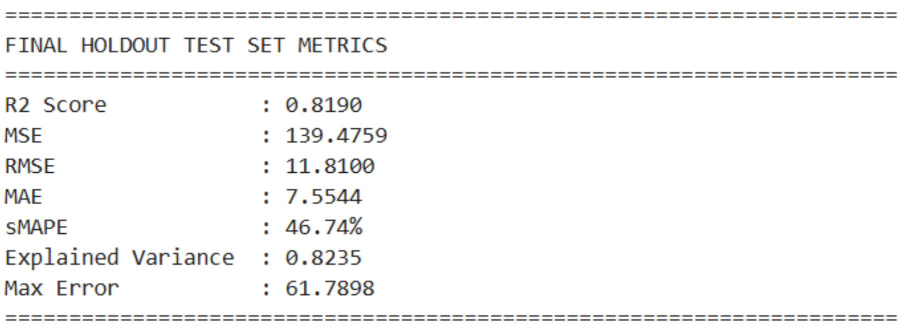


Figure 2: Cross-Validation Analysis Plots



```
======================================================================
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
======================================================================
```

Figure 3: Test Set Results

# Conclusion

Based on the data set we found the trends which gives high yields :

**Ligands (bulky biaryl phosphines,OMe or bulky aryl rings)**

• Ligands that are bulky and strongly electron-rich tend to give higher yields as it accelerates oxidative addition of aryl iodides/bromides and reductive elimination to form the C–N bond.

• The Steric bulk promotes fast reductive elimination as it suppress side reactions to keep the catalytic cycle fast increasing the yield.

• Bulky biaryl-phosphine ligands improves oxidative addition and help stabilize Pd(0).

**Bases (substituted amines, amidine-type bases)**

• Strong bases(alkoxides, amides, bulky organic bases) deprotonate the amines and fasten C–N bond formation.

• Bulky, non-coordinating bases reduce unwanted Pd binding that slows down the catalytic cycle.

• Very weak or strongly coordinating bases leads to slower turnover and lower yields.

**Aryl halides (I and Br, fluoro, carbonyl, or alkoxy groups)**

• Aryl iodides and bromides are highly reactive and result to high yields.

• Electron-withdrawing groups(–CN, –F, –CO, –CF3) strongly favor oxidative addition and gives higher yields.

• Electron-rich aryl halides(–OC, –CC substituents) will still react efficiently with strong, due to strong ligation.

**Additives (neutral heterocyclic carbonyl/nitrogen compounds)**

• Mildly coordinating additives(N–O, C=O, or heterocyclic scaffolds), which provide weak coordination stabilizes Pd intermediates.

• Additives suppress Pd black formation and help maintaining a high concentration of active Pd(0).

• Additives resemble mild halide proton shuttles, improving oxidative addition and amine activation.

• Additives help tuning coordination, solubility, and reaction rates.

The study demonstrates that the optimized neural network model can successfully learn meaningful structure–yield relationships and generalize to unseen reactions. The model enables high-speed virtual screening of thousands of reaction combinations, significantly reducing the experimental load on chemists. This work highlights how machine learning can accelerate chemical discovery and improve efficiency in synthetic chemistry workflows.

# References

https://chem.libretexts.org/Bookshelves/Inorganic_Chemistry/Supplemental_Modules_and_Websites_(Inorganic_Chemistry)/Catalysis/Catalyst_Examples/Buchwald-Hartwig_Amination

https://www.organic-chemistry.org/namedreactions/buchwald-hartwig-reaction.shtm