CHEME 6880 | Final Project – Supporting Information

Using supervised learning methods to predict Buchwald-Hartwig amination yields

from high throughput experimentation data

Nathan M. Lui

**Contents**

|  |  |  |
| --- | --- | --- |
| List of features | | 3 |
| List of models (w/ implementation) | | 4 |
| **Table S1** | Performance and optimized hyperparameters of models trained on DFT features using leave-one-molecule-out testing. | 5 |
| **Table S2** | Performance and optimized hyperparameters of models trained on one-hot encoded features using leave-one-molecule-out testing. | 7 |
| **Table S3** | Performance and optimized hyperparameters of models trained on DFT encoded features using randomly sampled testing. | 9 |
| **Figure S1** | Observed vs. predicted plot for the test set molecules of the best linear model. | 11 |
| **Figure S2** | Observed vs. predicted plot for the test set molecules of the best linear model and Żurański model. | 11 |
| **Figure S3** | Observed vs. predicted plot for the test set molecules of the best support vector regressor. | 12 |
| **Figure S4** | Observed vs. predicted plot for the test set molecules of the best support vector regressor and Żurański model. | 12 |
| **Figure S5** | Observed vs. predicted plot for the test set molecules of the best *k*-nearest neighbor regressor. | 13 |
| **Figure S6** | Observed vs. predicted plot for the test set molecules of the best *k*-nearest neighbor regressor and Żurański model. | 13 |
| **Figure S7** | Observed vs. predicted plot for the test set molecules of the best tree-based regressor. | 14 |
| **Figure S8** | Observed vs. predicted plot for the test set molecules of the best tree-based regressor and Żurański model. | 14 |
| **Figure S9** | Observed vs. predicted plot for the test set molecules of the best neural network. | 15 |
| **Figure S10** | Observed vs. predicted plot for the test set molecules of the best neural network and Żurański model. | 15 |
| Attached Files | | 16 |
| Supplementary References | | 17 |

**List of features**

Reproduced from Supplementary Information of Ahneman et al.1

\* refers to atom-specific descriptors

Additive Descriptors (n = 19)

EHOMO, ELUMO, Dipole Moment, Electronegativity, Hardness, Molecular Volume, Molecular

Weight, Ovality, Surface Area, \*C3 NMR Shift, \*C3 Electrostatic Charge, \*C4 NMR Shift, \*C4

Electrostatic Charge, \*C5 NMR Shift, \*C5 Electrostatic Charge, \*N1 Electrostatic Charge, \*O1

Electrostatic Charge, V1 Frequency, V1 Intensity

Aryl Halide Descriptors (n = 27)

EHOMO, ELUMO, Dipole Moment, Electronegativity, Hardness, Molecular Volume, Molecular

Weight, Ovality, Surface Area, \*C1 NMR Shift, \*C1 Electrostatic Charge, \*C2 NMR Shift, \*C2

Electrostatic Charge, \*C3 NMR Shift, \*C3 Electrostatic Charge, \*C4 NMR Shift, \*C4

Electrostatic Charge, \*H2 NMR Shift, \*H2 Electrostatic Charge, \*H3 NMR Shift, \*H3

Electrostatic Charge, V1 Frequency, V1 Intensity, V2 Frequency, V2 Intensity, V3 Frequency,

V3 Intensity

Base Descriptors (n = 10)

EHOMO, ELUMO, Dipole Moment, Electronegativity, Hardness, Molecular Volume, Molecular

Weight, Ovality, Surface Area, \*N1 Electrostatic Charge

Ligand Descriptors (n = 64)

Dipole Moment, \*C1 NMR Shift, \*C1 Electrostatic Charge, \*C2 NMR Shift, \*C2 Electrostatic

Charge, \*C3 NMR Shift, \*C3 Electrostatic Charge, \*C4 NMR Shift, \*C4 Electrostatic Charge,

\*C5 NMR Shift, \*C5 Electrostatic Charge, \*C6 NMR Shift, \*C6 Electrostatic Charge, \*C7

NMR Shift, \*C7 Electrostatic Charge, \*C8 NMR Shift, \*C8 Electrostatic Charge, \*C9 NMR

Shift, \*C9 Electrostatic Charge, \*C10 NMR Shift, \*C10 Electrostatic Charge, \*C11 NMR Shift,

\*C11 Electrostatic Charge, \*C12 NMR Shift, \*C12 Electrostatic Charge, \*C13 NMR Shift,

\*C13 Electrostatic Charge, \*C14 NMR Shift, \*C14 Electrostatic Charge, \*C15 NMR Shift,

\*C15 Electrostatic Charge, \*C16 NMR Shift, \*C16 Electrostatic Charge, \*C17 NMR Shift,

\*C17 Electrostatic Charge, \*H11 NMR Shift, \*H11 Electrostatic Charge, \*H3 NMR Shift, \*H3

Electrostatic Charge, \*H4 NMR Shift, \*H4 Electrostatic Charge, \*H9 NMR Shift, \*H9 Electrostatic Charge, \*P1 Electrostatic Charge, V1 Frequency, V1 Intensity, V2 Frequency, V2

Intensity, V3 Frequency, V3 Intensity, V4 Frequency, V4 Intensity, V5 Frequency, V5 Intensity,

V6 Frequency, V6 Intensity, V7 Frequency, V7 Intensity, V8 Frequency, V8 Intensity,

V9 Frequency, V9 Intensity, V10 Frequency, V10 Intensity

**List of models examined (w/ implementation)**

All models were trained on the original data as well as normalized data

* Linear models:
  + Ordinary Least Squares (LinearRegerssion()))
  + Ridge Regression (RidgeCV())
  + LASSO Regression (LassoCV())
  + Elastic Net (ElasticNetCV())
  + Stochastic Gradient Descent Regression (SGDRegressor())
  + Generalized Linear Model – GLM (TweedieRegressor())
* *k*-nearest Neighbors (KneighborRegressor())
* Support Vector Regression (SVR())
* Tree-based models:
  + Decision Tree Regression (DecisionTreeRegressor())
  + Random Forest Regression (RandomForestRegressor())
  + Gradient Boosted Regression Trees – GBRT (GradienBoosingRegressor())
  + Adaboost Regression Tree (AdaboostRegressor())
  + XGBoosted Regression Tree (xgb.XGBRegressor())
* Neural Network Regression (MLPRegressor())

**Table S1** | Performance and optimized hyperparameters of models trained on DFT features using leave-one-molecule-out testing.

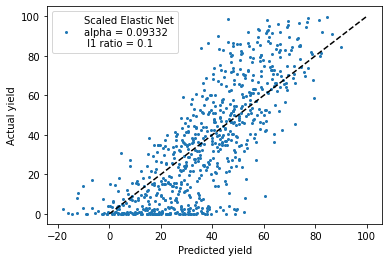
|  |  |  |  |
| --- | --- | --- | --- |
| Model | Optimized Model Hyper/parameters | Test RMSE | Test *R*2 |
| OLS |  | 164612.23 | -32733729.15 |
| Scaled OLS |  | 5.8 e11 | -4.1e20 |
| Ridge | α = 0.0910 | 17.26430 | 0.63994 |
| Scaled Ridge | α = 244.20 | 17.86942 | 0.61426 |
| LASSO | α = 39.49 | 19.43688 | 0.54362 |
| Scaled LASSO | α = 0.215 | 18.18718 | 0.60042 |
| Elastic Net | α = 39.49  ℓ1 ratio = 1.0 | 19.43688 | 0.54362 |
| Scaled Elastic Net | α = 0.093  ℓ1 ratio = 0.1 | 17.90920 | 0.61254 |
| SGD Regression | α = 1000  learning rate: adaptive  using early stopping | 1.3e13 | -2.1e23 |
| Scaled SGD Regression | α = 0.1  learning rate: adaptive  using early stopping | 17.94914 | 0.61081 |
| GLM | Poisson regression  α = 1 | 28.84749 | -0.00528 |
| Scaled GLM | Gaussian regression  α = 0.2 | 18.23403 | 0.59836 |
| Żurański GLM2 | Gaussian regression  α = 0.1 | 17.92750 | 0.61175 |
| *k*-nn | *k* = 3  distance metric: ℓ1-norm  distance-weighted | 15.43473 | 0.71221 |
| Scaled *k*-nn | *k* = 5  distance metric: ℓ1-norm  distance-weighted | 12.25547 | 0.81856 |
| Żurański *k*-nn2 | *k* = 3  distance metric: ℓ2-norm | 13.24342 | 0.78813 |
| SVR | C = 1000 | 18.27959 | 0.59635 |
| Scaled SVR | C = 100 | 13.63759 | 0.77533 |
| Żurański SVR2 | C = 0.5  γ = 0.007 | 22.25789 | 0.40153 |
| Decision Tree | minimum samples per leaf = 300 | 24.25820 | 0.28913 |
| Scaled Decision Tree | minimum samples per leaf = 300 | 24.25820 | 0.28913 |
| Random Forest | trees = 300  maximum features = 7 | 13.17499 | 0.79031 |
| Żurański Random Forest2 | trees = 20 | 19.29864 | 0.55009 |
| Gradient Boosted Regression Trees | trees = 200 | 16.11411 | 0.68632 |
| Adaboost Regression Trees | learning rate = 0.1  trees = 100  maximum tree depth = 3 | 18.80018 | 0.57303 |
| XGBoost Regression Trees | learning rate = 0.3  number of trees = 20  maximum tree depth = 3  gamma = 0 | 17.14594 | 0.64486 |
| Żurański XGBoost Regression Trees2 | learning rate = 0.3  number of trees = 15  maximum tree depth = 6  gamma = 0 | 17.37165 | 0.63545 |
| Neural Network | α = 0.00001  hidden layer nodes: 100 | 20.97158 | 0.46871 |
| Scaled Neural Network | α = 0.0001  hidden layer nodes: 6 | 15.22099 | 0.72013 |
| Żurański Neural Network2 | α = 0.0001  hidden layer nodes: 4 | 16.75460 | 0.66089 |

**Table S2** | Performance and optimized hyperparameters of models trained on one-hot encoded features using leave-one-molecule-out testing.

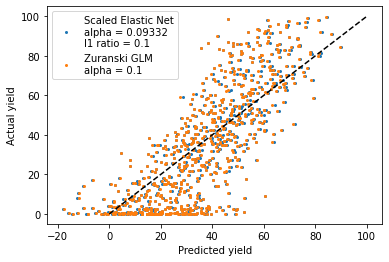
|  |  |  |  |
| --- | --- | --- | --- |
| Model | Optimized Model Hyper/parameters | Test RMSE | Test *R*2 |
| OLS |  | 5.4e12 | -3.6e22 |
| Scaled OLS |  | 5.8e14 | -4.1e26 |
| Ridge | α = 2.0236 | 17.28109 | 0.63924 |
| Scaled Ridge | α = 33.932 | 17.28669 | 0.63901 |
| LASSO | α = 0.0033 | 17.15413 | 0.64453 |
| Scaled LASSO | α = 0.00834 | 17.15078 | 0.64466 |
| Elastic Net | α = 0.0033  ℓ1 ratio = 1.0 | 17.15413 | 0.64453 |
| Scaled Elastic Net | α = 0.00834  ℓ1 ratio = 1.0 | 17.15078 | 0.64466 |
| SGD Regression | α = 0.001  learning rate: adaptive  using early stopping | 17.40642 | 0.63399 |
| Scaled SGD Regression | α = 0.1  learning rate: adaptive  using early stopping | 17.19814 | 0.64270 |
| GLM | Poisson regression  α = 1 | 18.71425 | 0.57886 |
| Scaled GLM | Gaussian regression  α = 0.2 | 16.78881 | 0.65950 |
| Żurański GLM2 | Gaussian regression  α = 0.1 | 17.64454 | 0.62391 |
| *k*-nn | *k* = 5  distance metric: ℓ2-norm  distance-weighted | 15.73572 | 0.70088 |
| Scaled *k*-nn | *k* = 5  distance metric: ℓ2-norm  distance-weighted | 15.20846 | 0.72059 |
| Żurański *k*-nn2 | *k* = 3  distance metric: ℓ2-norm | 15.48280 | 0.71042 |
| SVR | C = 100 | 14.59498 | 0.74268 |
| Scaled SVR | C = 100  γ = 0.01 | 14.19613 | 0.75655 |
| Żurański SVR2 | C = 0.5  γ = 0.007 | 22.58015 | 0.38408 |
| Decision Tree | minimum samples per leaf = 10 | 15.94972 | 0.69269 |
| Scaled Decision Tree | minimum samples per leaf = 10 | 15.94972 | 0.69269 |
| Random Forest | trees = 50  maximum features = 11 | 14.36821 | 0.75061 |
| Żurański Random Forest2 | trees = 20 | 16.72889 | 0.66193 |
| Gradient Boosted Regression Trees | learning rate = 1.0  trees = 700 | 15.65603 | 0.70390 |
| Adaboost Regression Trees | learning rate = 0.1  trees = 100  maximum tree depth = 5 | 19.81363 | 0.52576 |
| XGBoost Regression Trees | learning rate = 0.3  number of trees = 40  maximum tree depth = 6  gamma = 0 | 14.81151 | 0.73499 |
| Żurański XGBoost Regression Trees2 | learning rate = 0.3  number of trees = 15  maximum tree depth = 6  gamma = 0 | 15.93111 | 0.69341 |
| Scaled Neural Network | α = 0.0001  hidden layer nodes: 6 | 14.33062 | 0.75191 |
| Żurański Neural Network2 | α = 0.0001  hidden layer nodes: 4 | 14.77841 | 0.73617 |

**Table S3** | Performance and optimized hyperparameters of models trained on DFT encoded features using a randomly sampled test set.

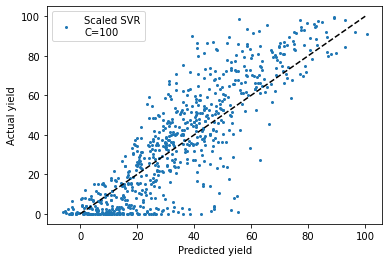
|  |  |  |  |
| --- | --- | --- | --- |
| Model | Optimized Model Hyper/parameters | Test RMSE | Test *R*2 |
| OLS |  | 15.17728 | 0.69440 |
| Scaled OLS |  | 15.15040 | 0.69548 |
| Ridge | α = 0.001 | 15.21968 | 0.69269 |
| Scaled Ridge | α = 0.001 | 15.17068 | 0.69466 |
| LASSO | α = 6.85239 | 17.17224 | 0.60878 |
| Scaled LASSO | α = 0.00948 | 15.40995 | 0.68496 |
| Elastic Net | α = 6.85239  ℓ1 ratio = 1.0 | 17.17224 | 0.60878 |
| Scaled Elastic Net | α = 0.00948  ℓ1 ratio = 1.0 | 15.40995 | 0.68496 |
| SGD Regression | α = 0.001  learning rate: adaptive | 1.4e13 | -2.9e23 |
| Scaled SGD Regression | α = 0.1  penalty: ℓ1  learning rate: adaptive | 15.99681 | 0.66050 |
| GLM | Poisson regression  α = 1 | 16.99227 | 0.61694 |
| Scaled GLM | Poisson regression  α = 0 | 13.05295 | 0.77396 |
| Żurański GLM2 | Gaussian regression  α = 0.1 | 16.43260 | 0.64176 |
| *k*-nn | *k* = 5  distance metric: ℓ1-norm  distance-weighted | 16.07998 | 0.65697 |
| Scaled *k*-nn | *k* = 5  distance metric: ℓ1-norm  distance-weighted | 15.57180 | 0.68970 |
| Żurański *k*-nn2 | *k* = 3  distance metric: ℓ2-norm | 17.91193 | 0.57435 |
| SVR | C = 1000 | 17.33799 | 0.60119 |
| Scaled SVR | C = 100  γ = 0.01 | 11.58464 | 0.82195 |
| Żurański SVR2 | C = 0.5  γ = 0.007 | 20.50922 | 0.44196 |
| Decision Tree | minimum samples per leaf = 1 | 9.56427 | 0.87864 |
| Scaled Decision Tree | minimum samples per leaf = 1 | 9.56427 | 0.87864 |
| Random Forest | trees = 50 | 7.61278 | 0.92311 |
| Żurański Random Forest2 | trees = 20 | 7.49203 | 0.92553 |
| Gradient Boosted Regression Trees | learning rate = 1.0  trees = 700 | 8.39844 | 0.90642 |
| Adaboost Regression Trees | learning rate = 0.1  trees = 10  maximum tree depth = 3 | 12.81256 | 0.78221 |
| XGBoost Regression Trees | learning rate = 0.3  number of trees = 40  maximum tree depth = 9  gamma = 0 | 0.82505 | 0.99910 |
| Żurański XGBoost Regression Trees2 | learning rate = 0.3  number of trees = 15  maximum tree depth = 6  gamma = 0 | 6.57220 | 0.94287 |



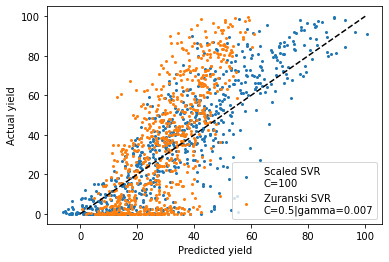
**Figure S1** | Observed vs. predicted plot for the test set molecules of the best linear model (elastic net regression, α = 0.09332, ℓ1 ratio = 0.1, *R*2 = 0.61).



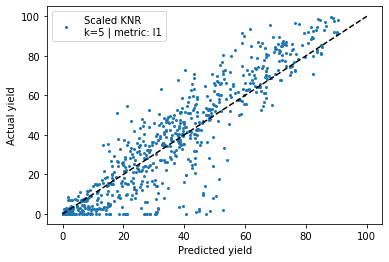
**Figure S2** | Observed vs. predicted plot for the test set molecules of the best linear model (blue, elastic net regression, α = 0.09332, ℓ1 ratio = 0.1, *R*2 = 0.61) and Żurański model (orange, generalized linear model, α = 0.1, *R*2 = 0.61).



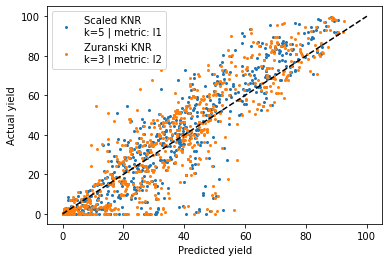
**Figure S3** | Observed vs. predicted plot for the test set molecules of the best support vector regressor (C = 100, *R*2 = 0.78).



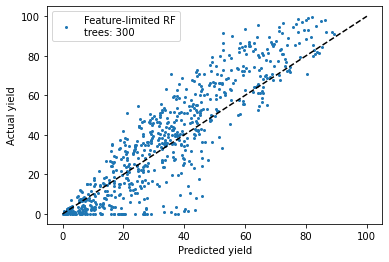
**Figure S4** | Observed vs. predicted plot for the test set molecules of the best support vector regressor (blue, C = 100, *R*2 = 0.78) and Żurański model (orange, C = 0.5, γ = 0.007, *R*2 = 0.40).



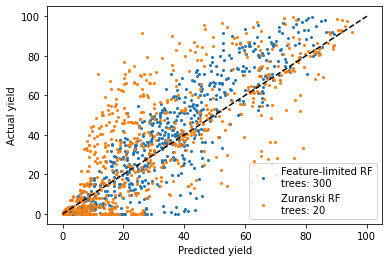
**Figure S5** | Observed vs. predicted plot for the test set molecules of the best *k*-nearest neighbor regressor (*k* = 5, distance metric: ℓ1, distance weighted, *R*2 = 0.82).



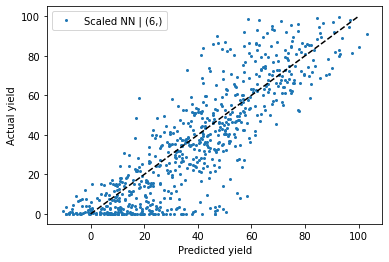
**Figure S6** | Observed vs. predicted plot for the test set molecules of the best *k*-nearest neighbor regressor (blue, *k* = 5, distance metric: ℓ1, distance weighted, *R*2 = 0.82) and Żurański model (orange, *k* = 3, distance metric: ℓ2, unweighted, *R*2 = 0.79).



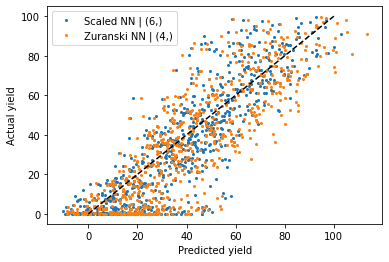
**Figure S7** | Observed vs. predicted plot for the test set molecules of the best tree-based regressor (random forest, trees = 300, feature-limited, *R*2 = 0.79).



**Figure S8** | Observed vs. predicted plot for the test set molecules of the best tree-based regressor (orange random forest, trees = 300, feature-limited, *R*2 = 0.79) and Żurański model (random forest, trees = 20, *R*2 = 0.55).



**Figure S9** | Observed vs. predicted plots for the test set molecules of the best neural network (6 hidden nodes in a single layer, *R*2 = 0.72).



**Figure S10** | Observed vs. predicted plot for the test set molecules of the best neural network (blue, 6 hidden nodes in a single layer, *R*2 = 0.72) and Żurański model (orange, 4 hidden nodes in a single layer, *R*2 = 0.66).

**Attached files**

dataProcessing.ipynb

Python notebook for pre-processing of data, conversion from smiles to features, and pickling of final datasets/indices

analysis-lomocv.ipynb

Python notebook for training/testing models on the DFT featurized dataset using the leave-one-molecule-out CV/split method.

DFT-LOMOCV.txt

Model parameters and training/testing errors for models trained on the DFT featurized dataset using the leave-one-molecule-out CV/split method.

analysis-randomCV.ipynb

Python notebook for training/testing models on the DFT featurized dataset using the random CV/split method.

DFT-RandomCV.txt

Model parameters and training/testing errors for models trained on the DFT featurized dataset using the random CV/split method.

analysis-onehot-lomocv.ipynb

Model parameters and training/testing errors for models trained on the one-hot encoded dataset using the leave-one-molecule-out CV/split method.

OneHot-LOMOCV.txt

Model parameters and training/testing errors for models trained on the one-hot encoded dataset using the leave-one-molecule-out CV/split method.

analysis-onehot-rCV.ipynb

Model parameters and training/testing errors for models trained on the one-hot encoded dataset using the random CV/split method.

OneHot-RandomCV.txt

Model parameters and training/testing errors for models trained on the one-hot encoded dataset using the random CV/split method.

**Supplementary References**

1. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science (80-. ).* **360**, 186–190 (2018).

2. Żurański, A. M., Martinez Alvarado, J. I., Shields, B. J. & Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **54**, 1856–1865 (2021).