Cross-Validation Methodology in Materials Science

Julien Brenneck[1], Alexander Dunn[2], Anubhav Jain[2]

[1]University of Massachusetts Amherst, [2]Lawrence Berkeley National Laboratory

Abstract

High-throughput Materials Science has turned towards statistical and machine learning methods in the effort to rapidly accelerate material characterization and discovery. While these methods promise to take advantage of the accumulating data, without effective quantification of performance and validation methodology there is no assurance that these models will generalize to new data. This work highlights potential pitfalls and best practices in the cross-validation of these models. In particular, it gives practical recommendations for cross-validation methodology.

Estimation of Generalization Error

When training a model $\hat{f}$ on training set T, we need to estimate how well it will do on a new set of data X, Y. The **generalization error** is defined as the expected loss L of the model over any new data, given training set T.

$$\mathrm{Err}_T = \mathbf{E}\left[L(Y, \hat{f}(X)) \mid T\right] \tag{1}$$

- **Test error** is average loss over a separate test sample.
- **Training error** is average loss over the training sample.

Training error does not estimate Err_T, and gives over-optimistic results, leading to over-fitting.

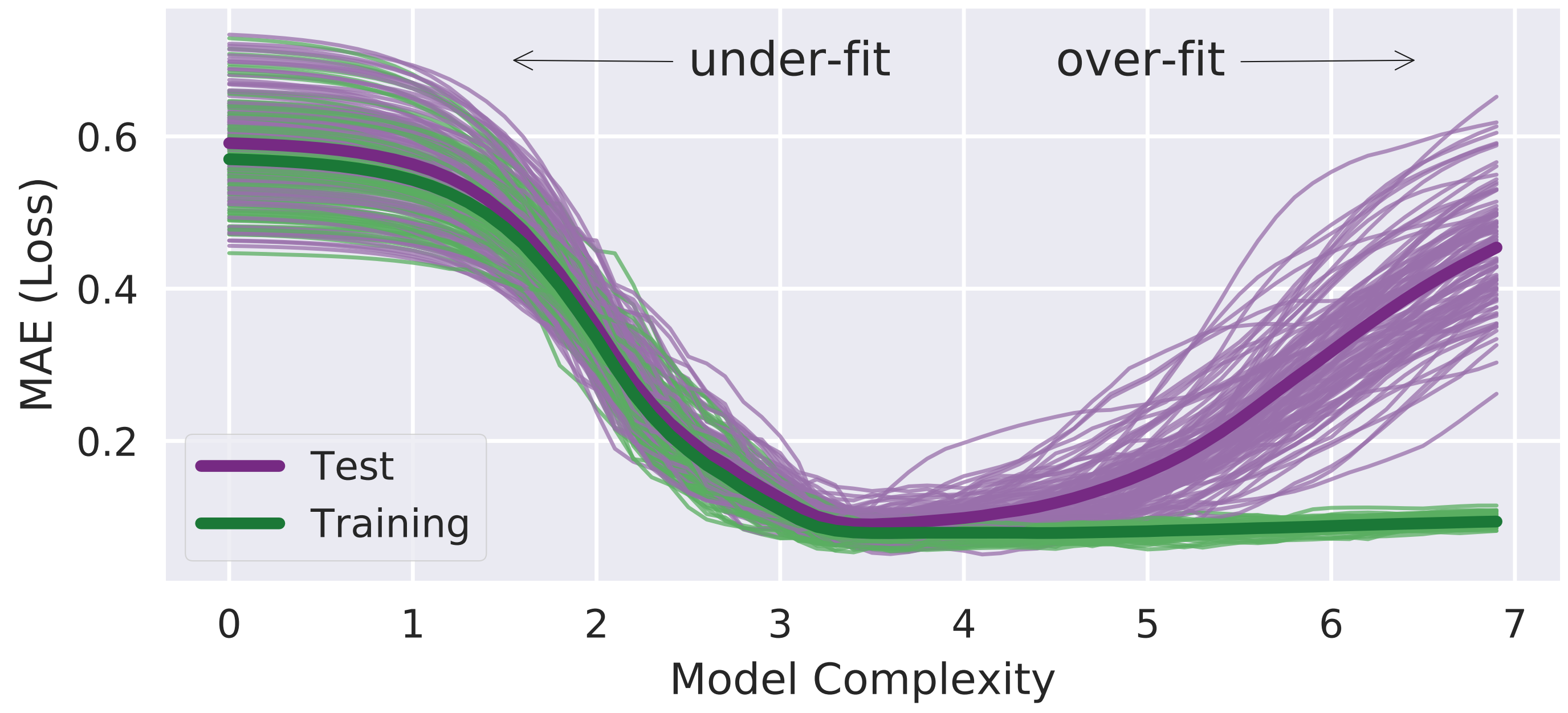


Figure 1: Results of experiment comparing training error and test error as model complexity increases. This employed a Support Vector Machine regression on synthetic data, generated from a sinusoidal waveform with Gaussian noise.

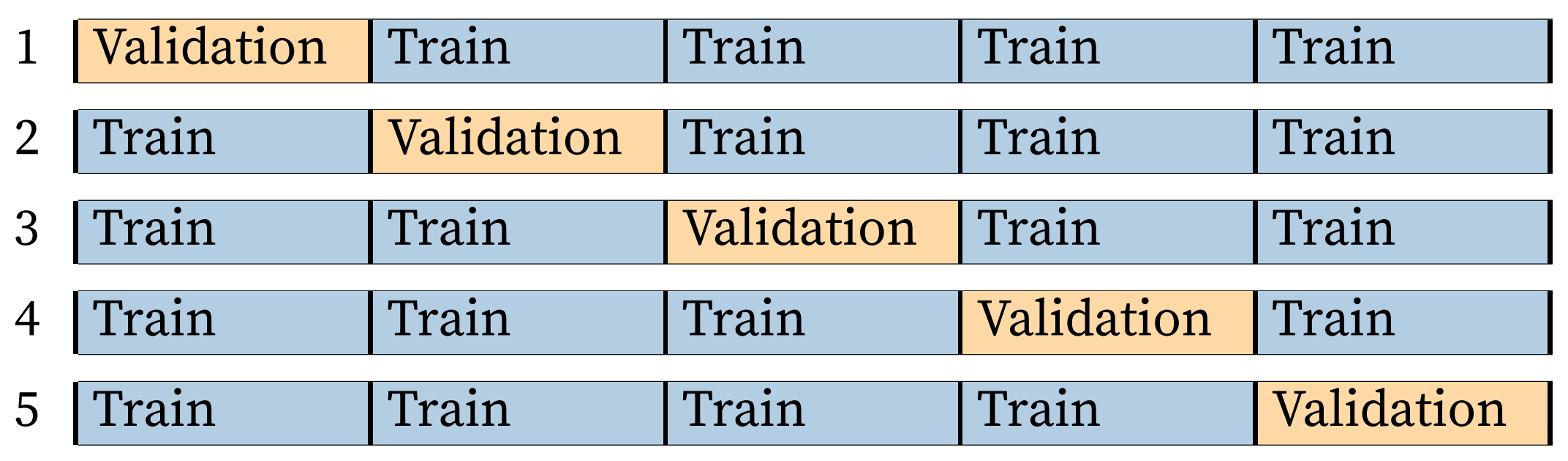
- Do not use training error to estimate model performance.
- Use hold-out test sets when possible.
- Use cross-validation, Bayesian models, the bootstrap, Bayesian information criterion, ect, for smaller data.

Materials Data

- Experimental data is small, often less than 1000 samples.
- Most common materials problems are trying to **extrapolate** from data instead of **interpolate**.
- Often very high number of features compared to size of the data, potentially introducing variance.

Cross-Validation

The **k-fold** cross-validation (CV) randomly samples the data into k distinct subsets of the same size.



1	Validation	Train	Train	Train	Train
2	Train	Validation	Train	Train	Train
3	Train	Train	Validation	Train	Train
4	Train	Train	Train	Validation	Train
5	Train	Train	Train	Train	Validation

Figure 2: 5-fold cross-validation. The data is randomly shuffled and then separated into 5 disjoint subsets of the same size, called folds. This gives 5 distinct train/validation splits.

- The k models are fit, each using a distinct validation set.
- Average of the k validation scores is taken.

Variance of the k-Fold Cross-Validation

How does the choice of k affect the result?

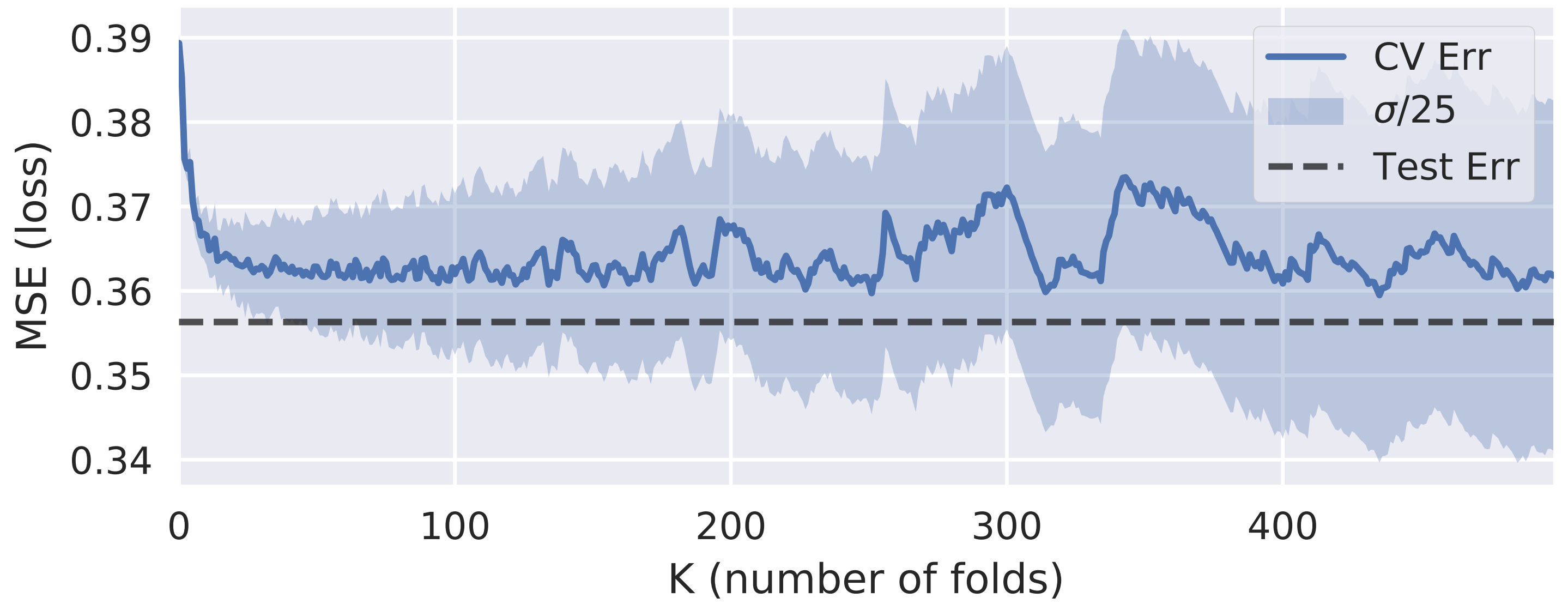


Figure 3: Computational experiment showing the sample mean and standard deviation of the k-fold scores for increasing values of k. This employed a Support Vector Regression (SVR) using a Radial Basis Function (RBF) kernel on synthetic data, with Mean Squared Error (MSE) loss.

This highlights part of the bias-variance trade-off.

- Small k gives higher bias, lower variance in CV estimate.
- Larger k gives lower bias, higher variance in CV estimate.
- Larger k much more computationally expensive!
- In general, **k=5 or k=10** is a practical compromise.

When $k = N$ we get **Leave One Out CV** (LOOCV). For some models LOOCV has low variance and can be computed efficiently, but in general can have high variance and is computationally expensive.

There is **no unbiased estimator** for k-fold CV variance.

- Bias can be same order of magnitude as total variance.
- Conservative estimates of CV variance difficult in practice.
- Robust comparisons of model performance difficult to make, as confidence intervals require variance estimates.

Repeated k-fold reduces variance of the CV estimate, without affecting bias. Recommended when computationally feasible. Repeats k-fold CV with new random folds and averages the scores.

Model Selection for Materials Data

The process of model selection, choosing a model and hyper-parameters, can heavily bias estimated model performance without proper methodology.

- Feature selection can introduce bias, with more features making the effect worse, a concern for materials data.
- Feature selection needs to be a part of model selection.

An experiment was conducted using materials data to show the susceptibility to over-fitting in model selection.

Model	Train Err	10-fold CV Err	Nested CV Err	Test Err
KRR	10.17	18.26	21.41	21.61
GBTR	3.37	14.37	16.38	20.90
GPR	0.00	32.63	33.96	35.45

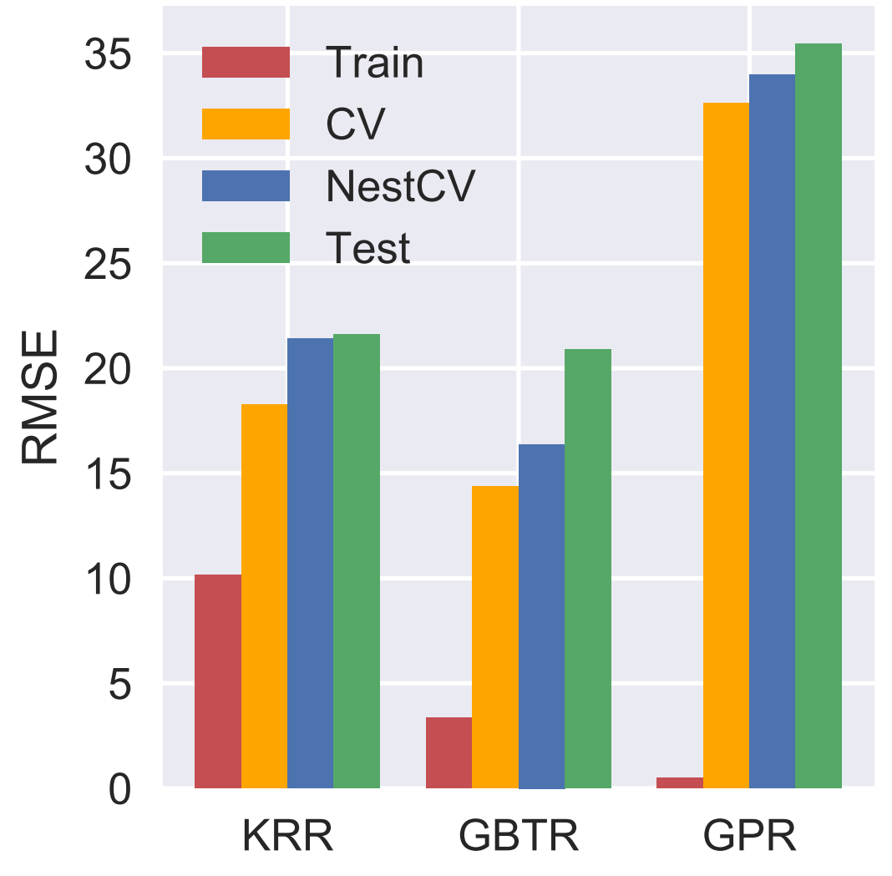


Table 1: Comparing prediction errors of an over-fit model. Model selection CV estimates can seriously underestimate true test error. Predicting bulk modulus based on 140 compositional and structural features, from matminer's ElementProperty and density featurizers, with a Kernel Ridge Regression (KRR), Gradient Boosted Tree Regression (GBTR), and a Gaussian Process Regression (GPR). Error is RMSE in GPa.

- Use a hold out test set for model selection.
- Consider nested CV or Bayesian methodology for small data.
- Over-fitting in model selection is a problem in practice, and can cause over or under fitting in the final model.
- Feature selection can cause bias if not part of cross-validation.

Bias is less important than variance in model selection (assuming a constant bias) as the "best" model will still be chosen.

Conclusion

- Use a validation set or k-fold CV with $k = 5$ or $k = 10$ depending on size of data, and avoid LOOCV (in general).
- Use repeated k-fold CV to increase confidence in CV estimate.
- Model selection and feature selection need to be internal to validation scheme to avoid over-fitting.
- A good methodology is using a test / train split, then using CV for model selection on the training data, and measuring final model performance on the test data.

Acknowledgements

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program. Additional funding was provided by the U.S. Department of Energy, Office of Basic Energy Sciences, Early Career Research Program.