Non-Parametric Models for Yield Prediction in a Suzuki-Miyaura Coupling

b.benyahia@lboro.ac.uk

Abstract

The yield of a chemical reaction significantly influences efficiency, cost, safety, sustainability, and product quality in the chemical and pharmaceutical industries. Opportunities arise from automated flow chemistry platforms, generating real-time experimental data, for the application of machine learning methods, enhancing insights into reaction performance. In this study, non-parametric models are employed to predict Suzuki-Miyaura coupling reaction yields from 5760 automated flow-synthesis reactions. Non-parametric models, specifically, Gaussian Processes (GPs) with one-hot encoding are utilized, with a comparison of their performance against classic machine learning models, ensemble models, and a two-layer neural network. GPs, particularly those with Automatic Relevance Determination (ARD), demonstrate superiority over the evaluated machine learning models, revealing improvements in R² and root mean squared error metrics. The findings underscore the superiority of non-parametric models, especially GPs with ARD, over neural networks when using fractions of the dataset for training and predicting previously unseen reactions. Non-parametric models show potential for accurate yield predictions even with relatively small datasets, promising faster and more effective development of synthetic pathways in the chemical and pharmaceutical industries.

**Keywords**: Machine Learning, Gaussian Process, Flow Chemistry, Yield Prediction

* 1. Introduction

The use of algorithms and statistical models in chemical and pharmaceutical research has gained relevance, especially with the rise of High Throughput Experimentation (HTE) and continuous flow platforms. Automated platforms enhance reproducibility, efficiency, and real-time analytical data generation, leading to the rapid acquisition of standardized datasets (Bennett and Abolhasani, 2022). Experimentally obtained datasets, including negative outcomes, provide opportunities to develop realistic models for reaction outcome prediction, reaction condition prediction, and drug discovery (Angello et al., 2022). Model performance is evaluated with unseen test samples, making feature representation and model development active research areas (Pomberger et al., 2022). Chemical reaction yield, comparing obtained product to the theoretical maximum, is vital for optimizing reactions and resource use. The complexity stems from numerous reaction variables and challenges linked to the often small and varied data. This complexity escalates with descriptor and the choice of machine learning model, which directly impacts accuracy. Attaining high accuracy is essential in navigating this intricate process.

To address current challenges in predictive modelling, two benchmarks have been developed for machine learning in predicting reaction yields, namely the works done by Ahneman et al., (2018) on Buchwald-Hartwig amination and Perera et al., (2018) on Suzuki-Miyaura coupling. The Buchwald-Hartwig dataset was evaluated using various models, among which Random Forests (RFs) demonstrating the best performance. Descriptors' impact was explored through one-hot encoding and random features, affirming the models' capabilities (Chuang and Keiser, 2018). For the Suzuki-Miyaura dataset, a two-layer Neural Network, trained with one-hot encoding, predicted high yields with a subset of the data (Granda et al., 2018). Schwaller et al. (2021) used Natural Language Processing (NLP) models and SMILES representations to train a transformer model (BERT), outperforming RF models on the Buchwald-Hartwig Dataset and showing comparable results on Suzuki-Miyaura. Graph Neural Networks (GNNs) introduced a new dimension, with YieldGNN (Saebi et al., 2023) outperforming BERT and RF models. A GNN with permutation invariance processing (Kwon et al., 2022) outperformed both BERT and its data-augmented version, providing superior yield predictions along with uncertainty estimation.

Graph Neural Networks (GNNs) excel in capturing molecular relationships for superior predictive performance. However, there is an opportunity to enhance predictions further with additional machine learning (ML) models using non-structural descriptors, especially for small datasets like Suzuki-Miyaura. A promising approach is exploring GPs as non-parametric probabilistic models, known for modelling complex relationships and uncertainties, particularly in scenarios with limited data availability (Rasmussen, 2004; Stach et al., 2021).

In predicting reaction yields, GPs perform similarly to RFs regardless of the descriptors used in an HTE dataset of trimipramine (Pomberger et al., 2022). Simple one-hot encoding yields high-performing predictions, with slightly lower error for more complex descriptors. The capabilities of GP models, such as kernel selection and Automatic Relevance Determination (ARD), are yet to be explored in yield predicting models. This work assesses the predictive capabilities of GP models against classic models (k-Nearest Neighbour, Support Vector Machine (SVM)), ensemble models (RF, XGBoost), and a two-layer neural network (Granda et al., 2018) using Suzuki-Miyaura dataset with one-hot encoding. The GP model, with automatic relevance determination, outperforms other models. Evaluating the model on fractions of the dataset confirms its high performance and potential for handling relatively small datasets. Anticipated benefits include more accurate yield predictions, expediting synthetic pathway development in the chemical and pharmaceutical industries.

* 1. Methods
     1. Suzuki-Miyaura coupling dataset.

Perera and colleagues (2018) developed and automated flow system for reaction screening of a Suzuki-Miyaura coupling reaction involving the preparation of reaction segments at regular timer intervals and their analysis through high-resolution Liquid Chromatography-Mass Spectrometry (LC-MS). The process involved the injection of stock solution aliquots of the catalyst, reactants, ligands, and bases into a carrier solvent stream. This setup allowed the variation of a complete set of variables including 11 ligands, 7 bases, and 4 solvents, resulting in a total of 5760 reactions.

In the present work, a one-hot vector encoding of the set of variables was applied to explicitly evaluate the performance of machine learning models in the prediction of reaction yields. This type of encoding has been demonstrated to deliver high-quality predictions for reaction yields in previous studies on HTE-generated datasets. An overview of the reaction, along with the specific reactants, ligands, bases, solvents, and reaction conditions as well as an example of the vector encoding are detailed in Fig. 1.



**Fig. 1.** Suzuki-Miyaura coupling reaction overview with specifications of the set of variables along with a visual representation of one-hot vector encoding.

* + 1. Non-parametric Models

Gaussian Processes (GPs) are non-parametric models known for their flexibility and ability to capture complex non-linear relationships within data. GPs are defined by mean and covariance functions, as expressed in Eq. (1), allowing them to model the entire probability distribution over functions.

|  |  |
| --- | --- |
|  | (1) |

The mean is a function of ,which in practical applications takes to be the zero function, and the covariance expresses the expected covariance of at the points and . The covariance function involves hyper-parameters, subject to optimization, capturing the relationships of the encoding of the reactants according to the respective yield of the reactions.

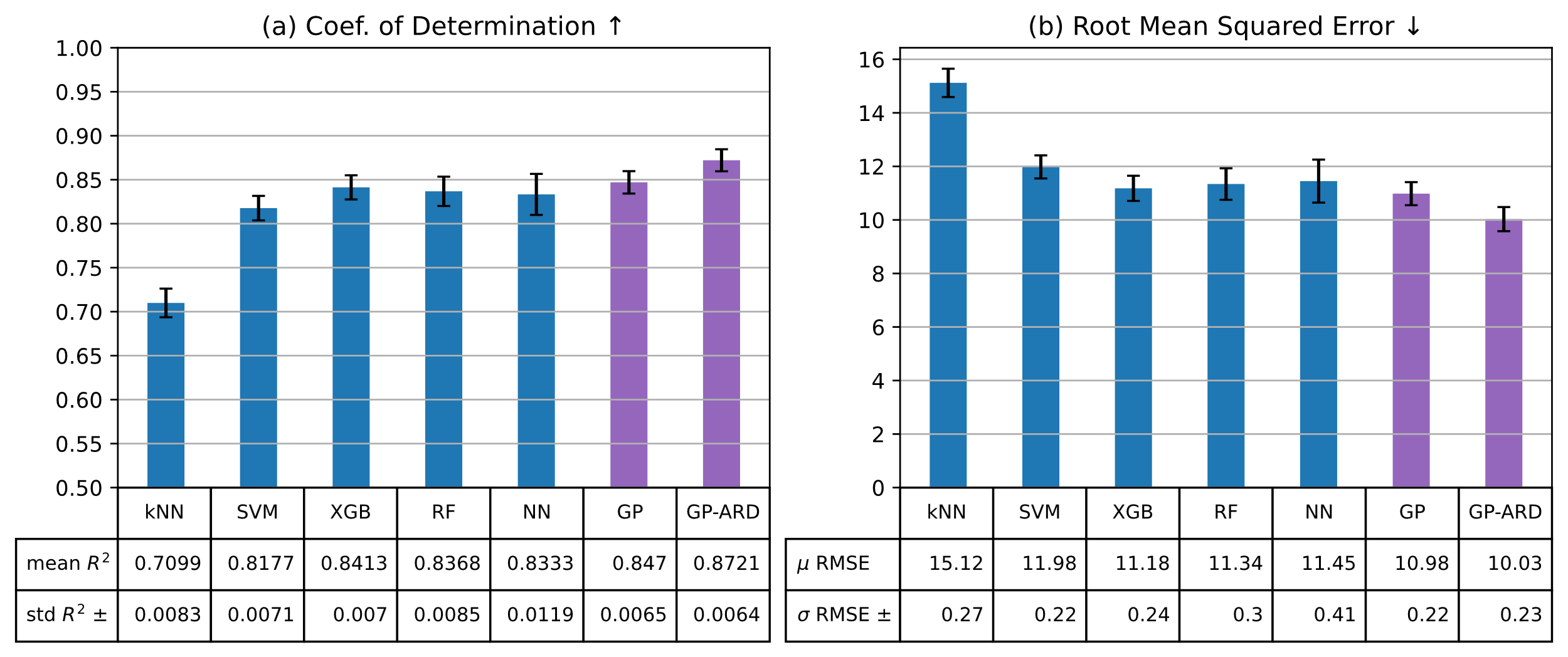
In this study, the evaluation of yield prediction performance led to the identification of the Matern 5/2 covariance with a single lengthscale hyperparameter as the initial effective choice. However, employing Automatic Relevance Determination (ARD), which assigns a lengthscale parameter to each input dimension, yielded an enhancement in predictive performance. The significance of these lengthscales lies in their role in defining the smoothness and oscillatory behaviour of the model. While a shared lengthscale in Matern 5/2 implies a uniform impact across dimensions, ARD's individual lengthscales enable customized adjustments, thereby influencing the contribution of each input dimension to the model's predictive accuracy.

* + 1. Experimental Setup

To evaluate non-parametric models for yield prediction, GP models were compared with various machine learning (ML) models using the Suzuki-Miyaura coupling dataset. The data was split into 70% for training and 30% for testing. Hyperparameter tuning was conducted via grid search and five-fold cross-validation, optimized K-Nearest Neighbors, Support Vector Machine Regression, Random Forests, and XGBoost. Optimal hyperparameters were then used to train and compare the models. Simultaneously, a two-layer neural network was implemented with hyperparameters from a previous study (Granda et al., 2018). The network underwent a 200-epoch training regimen and it was evaluated using a validation dataset. The best-performing epoch was identified and retrained with the validation dataset. In parallel, GP regression models were deployed using the Matern 5/2 kernel with a unique lengthscale parameter for all input dimensions. Another GP model with Matern 5/2 kernel and Automatic Relevance Determination (ARD) was tested for predictive performance. This process was repeated 100 times with different random seeds for reproducibility. Additionally, to assess GP models on small datasets, GPs and Neural Networks were trained on fractions from 0.1 to 0.8 of the entire dataset and tested on the remaining data. This was performed across 100 random train/test splits for each fraction. Performance metrics included Root Mean Squared Error and R², evaluating prediction accuracy and how well a model explains variance in the dependent variable for unseen test data.

* 1. Results
     1. Model Comparison

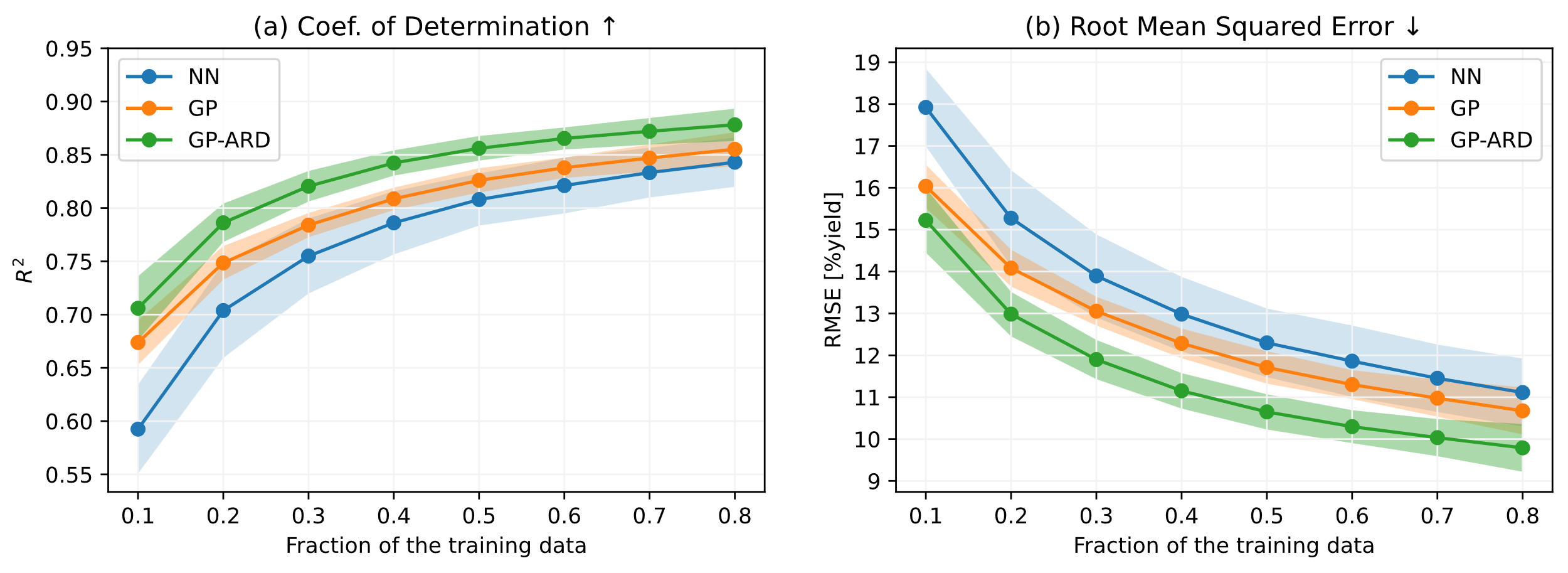
The best-performing models from each of the 100 70/30 train/test splits were compared. As shown in Fig. 2a, k-Nearest Neighbour demonstrated the poorest predictive performance among all the tested models, while SVM, Neural Networks, and Ensemble models (Random Forest and XGBoost) exhibited similar performance, with their R2 and RMSE falling within the confidence interval of the Neural Network; Ensemble models performed better. Gaussian Processes with Matern 5/2 showed slightly superior performance to XGBoost. Notably, GP with Automatic Relevance Determination (ARD) demonstrated the best performance, with its 95% confidence interval values surpassing the best possible performance of its counterparts. Additionally, GP models exhibited less variability in performance. Regarding Root Mean Squared Error (Fig. 2b), the results aligned with the R² metric, where GP with ARD displayed the lowest error, indicating superior predictive accuracy.



**Fig. 2.** Predictive performance metrics on test data for the evaluated models. Including the mean and standard deviation across 100 random 70/30 train/test splits for each model. Values and 95% confidence intervals are shown for: A) Coefficient of determination, and B) RMSE.

* + 1. Predictive Performance Fractions of the Dataset

The performance of predicting reaction yields was assessed using dataset fractions ranging from 0.1 to 0.8, with the remaining portion used for predictions. Both configurations of GPs were compared with a two-layer neural network, each fraction trained 100 times with different random seeds. Fig. 3 illustrates a performance trend, showing that GP models exhibit significantly better performance when trained with 10% of the dataset. GPs consistently outperform neural networks, as evident in the depicted confidence intervals for each model. Notably, GP with Automatic Relevance Determination (ARD) consistently outperforms both models. It's worth highlighting that GP with ARD, trained with 0.4 of the dataset, achieves slightly better performance than training the neural network and GP model with 0.8 of the dataset.



**Fig. 3.** Predictive performance of training dataset fractions for Neural Network, GP, and GP with ARD models. The shaded regions around each line depict the 95% confidence interval based on 100 tests for: A) Coefficient of determination, and B) Root mean squared error.

* 1. Discussion

The evaluation of GP models for reaction yield prediction indicates that optimizing a single lengthscale parameter for all input dimensions achieves comparable or superior performance to ensemble models, two-layer neural networks, and transformer-based models reported in previous studies. Optimizing lengthscales for each input dimension significantly impacts the model's ability to discern individual dimensions' influence on predictive accuracy, resulting in substantial performance improvement against all tested models, comparable to GNN-based models known for superior structural information encoding (Saebi et al., 2023). Gaussian processes show adaptability for modelling intricate relationships without a specific functional form, pivotal in datasets with complex patterns or challenging scenarios for traditional ML models. ARD kernels enhance adaptability by automatically adjusting feature relevance, prioritizing those significantly contributing to yield prediction. The non-parametric nature of GPs with ARD proves beneficial in scenarios with varying feature importance, and automatic feature selection within ARD kernels prevents overfitting. This study unveils GPs' intriguing capability to achieve high performance even with small datasets (10% of the total), attributed to their non-parametric adaptability. Despite the reported slightly lower prediction errors with additional descriptors in ML models, the limitation of one-hot encoding lies in its inability to capture physical or chemical information. This study explicitly focuses on the impact of non-parametric models on yield predictions, achieving high performance even with limited dataset fractions. In reaction optimization, accuracy in yield prediction becomes crucial for informed decision-making, anticipating faster and more effective synthetic pathway development in the chemical and pharmaceutical industries.

* 1. Conclusion

In conclusion, the integration of advanced algorithms and statistical models in chemical and pharmaceutical research, driven by high-throughput experimentation and continuous flow platforms, has yielded significant progress. Benchmark studies on real-time analytical data showcase the potential of innovative approaches, such as Natural Language Processing models and Graph Neural Networks, surpassing traditional models in predicting reaction yields. Furthermore, exploring predictive modelling challenges highlights the efficacy of Gaussian Processes, especially in enhancing accuracy for datasets like Suzuki-Miyaura, even with small sizes. The use of GPs with automatic relevance determination demonstrates superior performance compared to classic and ensemble models, holding promise for advancing accurate reaction yield predictions and expediting synthetic pathway development in the chemical and pharmaceutical industries. Future work will involve evaluating non-parametric models with additional descriptors and exploring different datasets for yield prediction in subsequent reactions.

**Acknowledgments**

This research has received funding from the UKRI (reference number 10038378) as part of the European Union – Health and Digital Executive Agency (HADEA) under the call HORIZON-HLTH-2021-IND-07 – grant agreement No 101057430.

References

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

N.H. Angello, V. Rathore, W. Beker, A. Wołos, E.R. Jira, R. Roszak, et al., 2022, Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling. Science, 378, 399–405.

J.A. Bennett, M. Abolhasani, 2022, Autonomous chemical science and engineering enabled by self-driving laboratories. Curr Opin Chem Eng, 36, 100831.

K.V. Chuang, M.J. Keiser, 2018, Comment on “Predicting reaction performance in C–N cross-coupling using machine learning.” Science, 362, 589–604.

J.M. Granda, L. Donina, V. Dragone, D-L. Long, L. Cronin, 2018, Controlling an organic synthesis robot with machine learning to search for new reactivity. Nature, 559, 377–381.

Y. Kwon, D. Lee, Y.S. Choi, S. Kang, 2022, Uncertainty-aware prediction of chemical reaction yields with graph neural networks. J Cheminform, 14, 2.

D. Perera, J.W. Tucker, S. Brahmbhatt, C.J. Helal, A. Chong, W. Farrell, et al., 2018, A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. Science, 359, 429–434.

A. Pomberger, A.A. Pedrina McCarthy, A. Khan, S. Sung, C.J. Taylor, M.J. Gaunt, et al., 2022, The effect of chemical representation on active machine learning towards closed-loop optimization. React Chem Eng, 7, 1368–1379.

C.E. Rasmussen, 2004, Gaussian Processes in machine learning. Lect Notes Comput Sci, 3176, 63–71.

M. Saebi, B. Nan, J.E. Herr, J. Wahlers, Z. Guo, A.M. Zurański, et al., 2023, On the use of real-world datasets for reaction yield prediction. Chem Sci, 14, 4997–5005.

P. Schwaller, A.C. Vaucher, T. Laino, J.-L. Reymond, 2021, Prediction of chemical reaction yields using deep learning. Mach Learn Sci Technol, 2, 015016.

E. Stach, B. DeCost, A.G. Kusne, J. Hattrick-Simpers, K.A. Brown, K.G. Reyes, et al., 2021, Autonomous experimentation systems for materials development: A community perspective. Matter, 4, 2702–2726.