University of Nottingham
UK | CHINA | MALAYSIA

# Artificial Intelligence for Chemical Synthesis:

## Improving the Workflow of Medicinal Chemists using Computer-Aided Synthesis Planning

**Alexe L. Haywood, MSci**

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

July 2023

# Abstract

Machine learning techniques have numerous applications in modern drug discovery. Advances in computing power, machine learning algorithms and data availability have inspired renewed interest in artificial intelligence and automation in chemical synthesis. The field of Computer-Aided Synthesis Planning (CASP) aims to improve chemists' workflow by shortening the time required to synthesise compounds, giving them more time to analyse and design future experiments. In this thesis, we review contemporary CASP methodologies before developing machine learning models to predict reaction yield. State-of-the-art approaches to forward reaction prediction and retrosynthetic analysis tasks are outlined and compared using quantitative metrics.

Predicting reaction yield is a newer aspect of CASP that has received significantly less attention than forward reaction prediction and retrosynthetic planning. This is owing, in part, to a lack of curated reaction data reporting reaction yield. Using a combinatorial benchmark dataset generated using high throughput experimentation, we evaluate machine learning models to predict reaction yield. Our research focuses on linear, tree-based, and Support Vector Regression (SVR) machine-learning algorithms. Chemical reactivity regression tasks frequently use molecular descriptors based on time-consuming, computationally demanding quantum chemical calculations. Along with quantum chemical descriptors, we investigate a range of topological representations that are quicker to calculate and applicable to all molecules. SVR emerges as the most promising machine learning model across all molecular descriptors in a preliminary cross-validation test evaluating interpolation.

Rigorous out-of-sample tests are designed to reliably assess the extrapolation capabilities of the most promising SVR models. The performance of SVR models built on topological representations surpasses those constructed on quantum chemical descriptors. The top SVR models built on each descriptor are subjected to additional validation. A collection of previously unseen perspective chemical

reactions is compiled. Predictions are presented for synthetic assessment to validate and explore the extent of the generalisability of the top SVR models.

# Publications

1. **Alexe L. Haywood**, Joseph Redshaw, Magnus W. D. Hanson-Heine, Adam Taylor, Alex Brown, Andrew M. Mason, Thomas Gärtner, Jonathan D. Hirst, “Kernel Methods for Predicting Yields of Chemical Reactions", *J. Chem. Inf. Model.*, 2021, **62**(9), 2077–2092.

2. **Alexe L. Haywood**, Joseph Redshaw, Thomas Gärtner, Adam Taylor, Andrew M. Mason, Jonathan D. Hirst, “Machine Learning for Chemical Synthesis”, in *Machine Learning in Chemistry: The Impact of Artificial Intelligence*, ed. H. Cartwright, The Royal Society of Chemistry, 2020, ch. 7, pp. 169–194.

# Acknowledgements

First and foremost, I would like to express my deepest appreciation to Prof. Jonathan Hirst for being an excellent supervisor since the final year of my undergraduate degree in 2018. Thank you for your patience, feedback, and understanding throughout my doctoral studies. Your encouragement and support have helped me gain confidence in my abilities, from academic writing to presentations. I never imagined that throughout my PhD, you would enable me to build enough confidence to present, not only, to the School of Chemistry but also at national and international conferences.

I am also grateful to GSK for the opportunities it has given me. Thank you to my additional supervisors who provided knowledge, expertise, and guidance throughout my PhD: Thomas Gärtner, Alex Brown, Andy Mason, Adam Taylor, Stephen Pickett, Andrew Baxter, and Simon MacDonald. Thank you to Joe Redshaw for sharing the PhD experience with me and helping me understand (to me) difficult math. I also had the pleasure of collaborating with the Theme 1 Prosperity Partnership team. I thoroughly enjoyed working on this project with you and appreciated our discussions.

I would like to extend my sincere thanks to everyone in the computational chemistry department. Thank you to A46, A47, and A48 for being great friends and support group and providing me with some unforgettable memories. A special thank you to Steve Oatley, my personal hype man, who has greatly helped me in many aspects of my Master's and PhD. Thank you to Katherine Wickham, Adam Fouda, Joe Glover, Steve Mason, and Ross Amory for making demonstrating fun over the years. I could not have undertaken this journey without my chemistry girls, Abi Miller, Ellen Guest, Grace Belshaw, and Josh Baptiste. You girls (and Josh) are always there for me, and I appreciate all your support and inspiration.

It has not been an easy final couple of years, and I would not be writing this without the support of my family. Thank you to James, Mum, and Nigel for your

constant love and encouragement. Thank you to my Grandad for brewing me endless cups of tea and telling me the most inspiring stories of your life. Finally, I would like to thank my Dad, who always knew I would become a doctor long before I did. I hope you are both gazing down on me, Grandad and Dad, because I DID IT!

# Contents

# List of Figures

# List of Tables

# List of Abbreviations and Symbols

**%ee** Enantiomeric Excess.

$\log \mathbf{P}$ Octanol-Water Partition Coefficient.

$\chi^2$ Chi-squared.

$R^2$ Coefficient of Determination.

**0D** Zero-Dimensional.

**1D** One-Dimensional.

**2D** Two-Dimensional.

**3D** Three-Dimensional.

**ACS** American Chemical Society.

**ADME** Adsorption, Distribution, Metabolism and Excretion.

**AI** Artificial Intelligence.

**API** Active Pharmaceutical Ingredient.

**API** Application Programming Interface.

**AT** Augmented Transformer.

**BET** Bromodomain and Extra-Terminal.

**CART** Classification and Regression Trees.

**CAS** Chemical Abstracts Services.

**CASP** Computer-Aided Synthesis Planning.

**CASP** Critical Assessment of protein Structure Prediction.

**CNN** Convolutional Neural Network.

**DFT** Density Functional Theory.

**DMPK** Drug Metabolism and Pharmacokinetics.

**DMTA** Design-Make-Test-Analyse.

**DOE** Design of Experiment.

**EBM** Energy-Based Model.

**ECFP** Extended Connectivity Fingerprint.

**ELN** Electronic Laboratory Notebook.

**EPO** European Patent Office.

**FAERS** FDA Adverse Event Reporting System.

**FAIR** Findability, Accessibility, Interoperability, and Reusability.

**FCFP** Functional-Class Fingerprints.

**FDA** U.S. Food and Drug Administration.

**FMorgan** Feature Morgan.

**G2Gs** Graph-to-Graphs.

**GCMS** Gas Chromatography-Mass Spectrometry.

**GCN** Graph Convolutional Network.

**GLN** Graph Logic Network.

**GLP-1** Glucagon-Like Peptide-1.

**GNN** Graph Neural Network.

**GTA** Graph Truncated Attention.

**GTPN** Graph Transformation Policy Network.

**HOMO** Highest Occupied Molecular Orbital.

**HTE** High Throughput Experimentation.

**HTS** High Throughput Screening.

**ID** Identification.

**iLCT** International Linked Clinical Trials.

**ILP** Integer Linear Programming.

**InChI** IUPAC International Chemical Identifier.

**IND** Investigational New Drug Application.

**IP** Intellectual Property.

**IUPAC** International Union of Pure and Applied Chemistry.

**JWST** James Webb Space Telescope.

**LASSO** Least Absolute Shrinkage and Selection Operator.

**LCMS** Liquid Chromatography-Mass Spectrometry.

**LHASA** Logic and Heuristics Applied to Synthetic Analysis.

**LSRL** Least Square Regression Line.

**LSTM** Long Short-Term Memory.

**LUMO** Lowest Unoccupied Molecular Orbital.

**MACCS** Molecular ACCess Systems.

**MAE** Mean Absolute Error.

**MCTS** Monte Carlo Tree Search.

**MEGAN** Molecular Edit Graph Attention Network.

**MFFs** Multiple Fingerprint Features.

**MLP** Multi-Layer Perceptron.

**MSE** Mean Squared Error.

**MT** Molecular Transformer.

**NCI/CADD** Computer-Aided Drug Design Group of the National Cancer Institute.

**NERF** Non-autoregressive Electron Redistribution Framework.

**NeuralSym** Neural-Symbolic.

**NLP** Natural Language Processing.

**NMR** Nuclear Magnetic Resonance.

**NPPN** Node Pair Prediction Network.

**OLS** Ordinary Least Squares.

**ORD** Open Reaction Database.

**PAINS** Pan-Assay Interference Compounds.

**PN** Policy Network.

**QSAR** Quantitative Structure-Activity Relationship.

**QSPR** Quantitative Structure-Property Relationship.

**RDK** RDKit.

**RF** Random Forest.

**R-SMILES** Root-aligned SMILES.

**RBF** Gaussian Radial Basis Function.

**RDT** Reaction Decoder Tool.

**RetroXpert** Retrosynthesis eXpert.

**RMSE** Root Mean Squared Error.

**RNA** Ribonucleic acid.

**RNN** Recurrent Neural Network.

**RSC** Royal Society of Chemistry.

**RSS** Residual Sum of Squares.

**SAR** Structure-Activity Relationship.

**SCROP** Self-Corrected Retrosynthesis Predictor.

**SELECT** Safety, Environmental, Legal, Economics, Control and Throughput.

**Seq2Seq** Sequence-to-Sequence.

**SMARTS** SMILES Arbitrary Target Specification.

**SMILES** Simplified Molecular-Input Line-Entry System.

**SMIRKS** A Reaction Transform Language.

**SVM** Support Vector Machine.

**SVR** Support Vector Regression.

**USPTO** United States Patent and Trademark Office.

**WIPO** World Intellectual Property Organization.

**WL** Weisfeiler-Lehman.

**WLDN** Weisfeiler-Lehman Difference Network.

**WLN** Weisfeiler-Lehman Network.

# Chapter 1

# Artificial Intelligence for Drug Discovery

## 1.1 Drug Discovery and Development

The drug discovery process is initiated when there is a lack of effective medicines to treat or cure a disease or condition. Current treatments may be unsatisfactory, or few to no treatment options may be available. The discovery or improvement of medicines is not without challenges. The average time taken from the start of the drug discovery process to marketing a new drug is typically 12 to 15 years and costs above $2-3 billion.[4,5] Many steps are completed between the discovery and the approval of an Active Pharmaceutical Ingredient (API). The five primary stages (Figure 1.1) are early drug discovery, preclinical trials, clinical trials, U.S. Food and Drug Administration (FDA) review and approval, and FDA post-market safety monitoring.

### 1.1.1 Early Drug Discovery

Early drug discovery aims to develop drug candidates likely to succeed in preclinical and clinical trials. Drug candidates must be able to prevent or reverse the effects of a disease or condition. Drug design is an iterative process for detecting and examing new drug candidates (Figure 1.2). It begins with identifying a biological entity known as the target, which plays a significant role in a disease. Small organic molecules are designed to bind to a specific target. The desired therapeutic effect produced upon binding could result from activating or inhibiting a biological response. The ability of small molecules to bind selectively is important since binding to off-target molecules can cause side effects. The drug

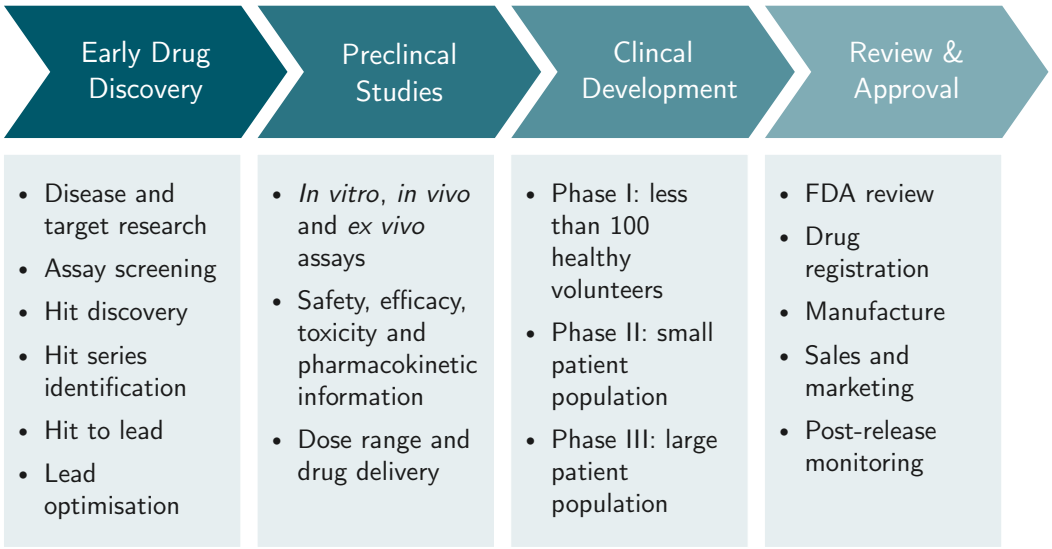



Figure 1.1: The drug discovery and development process.

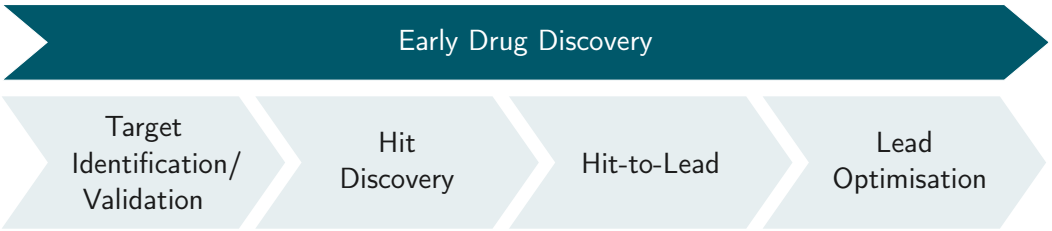



Figure 1.2: Early drug discovery workflow.

may act as a receptor agonist or antagonist or induce the opening or closing of an ion channel. Although small organic compounds have dominated treatments, biopharmaceuticals have also proven effective.[6] Peptides and therapeutic antibodies are examples of biopharmaceuticals. The focus of this thesis, on the other hand, will centre around small molecule drug design and development. Small molecules are screened against the biological target to identify compounds which bind to the target with a desired therapeutic effect. The active compounds are optimised to improve the selectivity, potency and physicochemical properties. Preclinical studies are conducted on the most promising drug candidates to establish the optimum dosage, method of administration, and toxicity. The main steps in early drug discovery are target identification and validation, hit discovery, hit series identification, hit-to-lead, and lead optimisation.

The initial focus is on the biological system and understanding the pathogenesis of the disease or disorder. Key biological targets in the pathogenesis are identified by searching through published literature and databases or via practical methods such as target deconvolution and discovery. The targets could include genes, proteins, or ribonucleic acid (RNA). Despite the nature of the biological entity, it needs to be disease-modifying and druggable. This means that the binding

of a drug molecule must cause a desirable biological response. For example, bromodomains are a potential druggable epigenetic target for treating diseases such as cancer, neurodegenerative disorders, inflammation, and obesity.[7] Small molecules are designed to inhibit protein interactions that can selectively regulate gene expression. Assuming all protein family members are equally druggable, a target can be considered druggable if drugs have successfully targeted other family members.[8] A more robust approach to predicting druggability is to use structural information about a target's binding site. Modelling tools can use this information to distinguish druggability within protein families,[9] such as the Bromodomain and Extra-Terminal (BET) protein family.[10] After identifying a druggable target, it is validated for suitability in drug development.

A hit molecule interacts with a drug target to produce a desired therapeutic effect, such as inhibition or activation. Occasionally hit molecules are discovered accidentally, but more frequently than not, hit molecules are identified by trial and error. High Throughput Screening (HTS) is an effective method for identifying hit molecules. Thousands to millions of small drug-like molecules are screened in various assays to identify those that interact with the therapeutic target and induce the desired biological response. HTS uses robotics and automation to test at the cellular, molecular, and biochemical levels. Utilising robotics, liquid handling devices, detectors, and data processing software allows large-scale compound libraries to be screened quickly. Other experimental screening techniques used alongside HTS include high-content, phenotypic, and fragment-based screening. Experimental screening is expensive in terms of money, labour, and time. Computational methods can reduce experimental screening while processing vast amounts of data. Virtual screening, for example, is a set of computational techniques that analyse small molecule libraries to identify potential hit candidates.

Experimental screening generates vast amounts of data that are difficult to manage. Numerous molecules with the necessary activity may be considered hits. Filters that remove frequent hitters, unwanted lipophilic compounds, or undesirable structural motifs may be applied. Compounds that include Pan-Assay Interference Compounds (PAINS)[11] are also excluded from further analysis. PAINS are substructures that frequently produce false positive results in high-throughput screening. Cheminformatics tools are introduced to sort compounds in a hit list. Computational clustering algorithms based on structural similarity scores categorise hits into series. A single average structure known as the cluster centroid represents each cluster. The clusters are ranked using confirmation experiments, and the top ones (hit series) are selected for further progression. Confirma-

tions include various assays performed closer to the physiological condition of the target to verify the hit molecules and measure efficacy values. Activity values, such as $IC_{50}$, can be calculated from dose-response curves. Synthetic tractability, feasibility, scale-up, associated costs, and patentability of the hit molecules are also evaluated. Alternatively, hits can be classified based on molecular scaffolds. Medicinal chemists regard structures with common scaffolds or cores to be comparable. This method of visualising series immediately demonstrates potential expansion around the shared core structure, which is beneficial in the lead optimisation phase. Small and simple hit molecules are favoured since they allow for increased molecular weight due to the inclusion of substituents to improve potency and selectivity.

After identifying a hit series, the molecules are explored and refined in the hit-to-lead phase. The central process that guides the exploration and exploitation of a hit series in hit-to-lead is the Design-Make-Test-Analyse (DMTA) cycle.[12] An efficient DMTA workflow requires medicinal chemistry, synthetic chemistry, computational chemistry and Drug Metabolism and Pharmacokinetics (DMPK) competence. Initially, the focus is on designing analogues of the hit series to test hypotheses (Design). Synthesising selected analogues (Make) for biochemical assays provides potency and physicochemical data (Test). Analysing the data verifies the specified hypotheses (Analyse). This insight aids the redesign process to improve properties. The DMTA cycle repeats until a molecule with desirable properties and potency is considered a lead molecule.

**Design.** Analogues of the hit series are designed in hit expansion using a combination of chemists' intuition and computational tools. Structure-Activity Relationship (SAR) analyses around the hit scaffolds aim to identify relationships between the chemical structure and biological activity or chemical properties. This approach formulates testable hypotheses with clear criteria for success. For example, whether the addition of a chemical group to the core structure improves a particular property or biological activity. The hypothesised SAR may have already been identified by the project experimentally or indicated computationally. Prediction tools such as Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR) models are an easy way to prioritise hypotheses before synthesis. These tools mathematically represent structure-activity and structure-property relationships, often using machine learning models. In QSAR, the chemical structure represented by molecular descriptors or physicochemical properties is related to biological activity. Whereas in QSPR, the chemical structure is related to a chemical property such as Octanol-Water Partition Coefficient ($\log P$), polar surface area, and Adsorption, Distri-

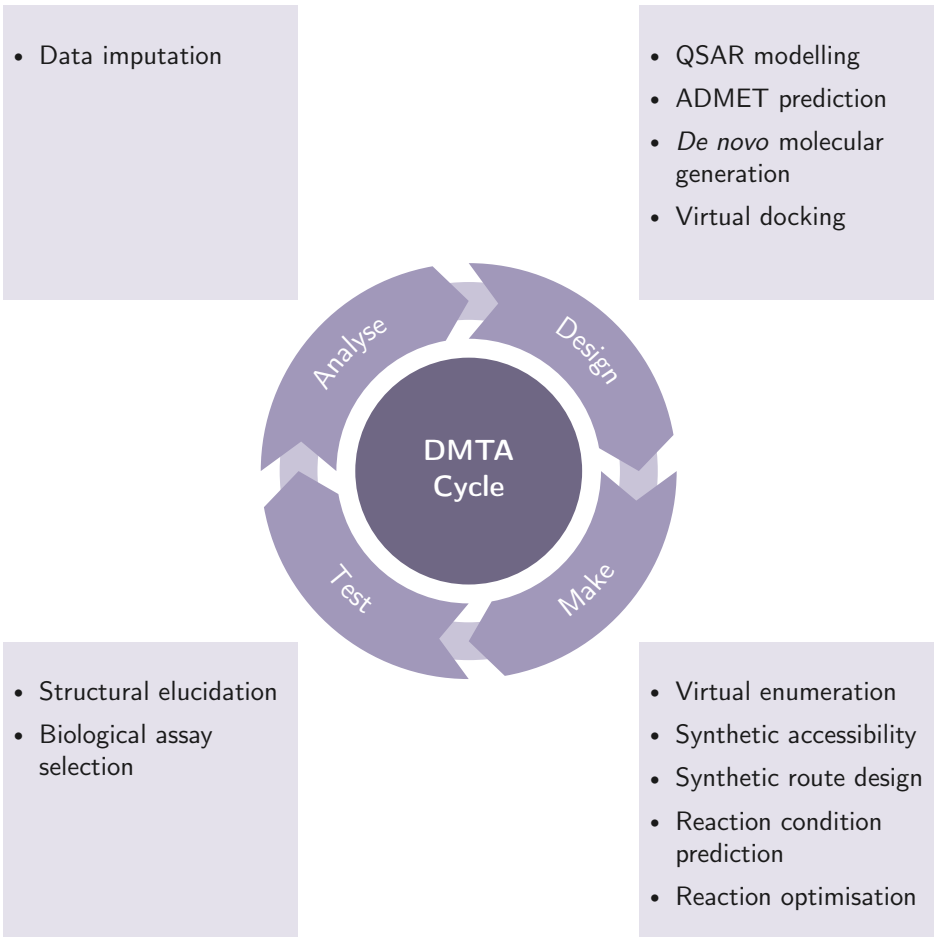



Figure 1.3: The Design-Make-Test-Analyse (DMTA) drug discovery cycle.

bution, Metabolism and Excretion (ADME) properties. ADME properties give a deeper understanding of a drug's pharmacological activity. Another practical computational tool is *de novo* drug design, which uses statistical models to generate drug-like compounds while prioritising the region of chemical space explored. Molecular docking programs also prioritise compounds by modelling the interaction between the small molecule and the biological target. Lipinski's rule of five is a rule of thumb to evaluate whether a compound is likely to be an orally active drug and proceed as a possible lead. Lipinski's rules are a collection of characteristics of small drug-like molecules that can operate as a filter to exclude or deprioritise compounds with poor physicochemical properties.[13] Lipinski's rule of five states that, in general, poor absorption or permeation is likely when a molecule exceeds two or more of the following properties: five hydrogen bond donors, ten hydrogen bond acceptors, a molecular mass of 500 daltons, and a $\log \mathrm{P}$ of five. While compounds with these molecular properties correlate well with high oral bioavailability in terms of solubility and permeability estimation, the values are considered guidelines. For example, Tinworth and Young suggest that molecular weight may not be as relevant as $\log \mathrm{P}$ and aromatic ring count in predicting the permeability and solubility of orally active drug molecules.[11]

**Make** Synthesising and biologically testing each compound in a laboratory requires a large amount of money, time, and resources. Computational tools can be utilised to reduce these limitations and the number of animal models by prioritising the compounds for wet lab experiments. Examples include virtual screening and Computer-Aided Synthesis Planning (CASP) tools. Virtual screening is an automated process for evaluating and filtering large libraries of compounds to a manageable amount for synthesis and testing. The libraries may be a predefined list of small molecules such as an in-house repository, public dataset, or commercial dataset. Alternatively, the libraries can be generated combinatorially from a set of pre-defined building blocks or in a more focused approach from a core scaffold with potential expansion points and a pre-defined list of substituents. This computational technique searches the virtual library to identify structures with improved biological activity. Ligand-based virtual screening uses information about the structure of ligands known to bind to the target without knowing the structure of the target. It is based on the assumption that molecules with similar structures have comparable properties, interactions with the target, and biological responses. Similarity searching, pharmacophore mapping, QSAR, and QSPR models are examples of ligand-based virtual screening techniques. Structure-based virtual screening requires knowledge of the structure of the biological target and its binding/active site. These approaches determine the ligands

that interact with the target by predicting binding affinity. Molecular docking is the most widely used structure-based virtual screening technique; other methods include molecular mechanics and molecular dynamics. Although CASP is currently in the early phases of development, it has the potential to improve the efficiency of the DMTA cycle. CASP tools include designing synthesis routes, predicting reaction conditions, and optimising chemical reactions. The tools aim to lighten the workload of synthetic chemists by providing synthesis recommendations. The most promising compounds are synthesised in the lab. Experimental methods such as combinatorial chemistry and high throughput chemistry may be employed to prepare a large number (tens to thousands) of compounds in a single process.

**Test** DMPK and physical chemistry assays on the synthesised compounds are performed. With an enhanced screening capability and a shift towards parallel testing, delivering *in vitro* data takes around ten working days.[12] Parallel testing results in information-rich data for each compound.

**Analyse** The assay results provide evidence to accept or reject the tested hypotheses. The prediction models implemented in the design stage are validated using the assay data. The assay data is analysed and converted into knowledge to aid redesign and propose new hypotheses. The effect of new proposed substructures can be observed via matched molecular pairs analysis against current benchmark compounds. An efficient DMPK cycle requires consistent quality reviews and tracking progression from concepts through synthesis and testing to results.[12]

The final stage of early drug discovery is lead optimisation. Lead optimisation aims to optimise or maintain the potency of lead compounds while balancing many other parameters. Toxicity, oral absorption, metabolic clearance *in vivo*, and activity in animal models must all be reviewed. Improving the selectivity against other biological targets is crucial since off-target interactions can lead to adverse effects. The DMTA cycle continues to explore analogues of one lead series and at least one backup series to improve properties. Computational property prediction models at this point may be sufficiently reliable to guide optimisation. Animal efficacy models are used to test drug safety and toxicity. A drug suitable for preclinical studies will exhibit high potency and good physicochemical and ADME properties while binding selectively to the target to cause the desired biological response.

### 1.1.2 Preclinical and Clinical Trials

Optimised lead compounds are thoroughly evaluated to provide sufficient evidence of safety, efficacy, toxicity, and pharmacokinetic information. *In vitro*, *in vivo* and *ex vivo* assays enable the evaluation of the drug candidates under conditions similar to those in living cells. *In vitro* studies are conducted in test tubes, *in vivo* studies on living animals, and *ex vivo* studies in cells or tissues of non-living animals. Before submitting an Investigational New Drug Application (IND) to progress to clinical trials, preclinical research must identify the following information. (i) Appropriate doses and the drug delivery system. Drug delivery may be targeted or controlled-release. Drug Delivery methods include oral, topical, membrane, intravenous, and inhalation. (ii) ADME pharmacokinetics properties to identify how the drug affects the body and if it interacts with healthy tissue. (iii) Possible side effects, adverse events, and the interaction with other treatments. (iv) Effects on gender, race or other ethnicity groups. (v) Scaled-up synthesis process to meet the sufficient qualities required in clinical trials.

If a drug candidate is successful in preclinical trials, it will progress to clinical trials to evaluate the drug in humans. Phase I of clinical trials involves fewer than 100 healthy volunteers. Human safety, pharmacokinetics, and side effects are investigated. Doses begin small and gradually increase if no risks are observed. Phase I trials aim to determine the best way to administer the drug while limiting toxicity and enhancing the therapeutic effect. Phase II clinical trials use a few hundred patients with the disease or condition, typically 100-500. The safety and efficacy of the drug are evaluated, optimum dose strength is determined, and adverse events are monitored. The data collected in the phase II study is used to optimise the design of the extensive phase III study. Phase III clinical trials use a few thousand patients with the disease or condition. The drug candidate is compared to existing treatments and a placebo. The drug efficacy is examined. Previously undetected long-term or rarer side effects in phases I or II are identified.

### 1.1.3 FDA Review and Post-Marketing Monitoring

Clinical trials establish the efficacy and safety of the drug. Before the drug is allowed to be sold commercially, the FDA must review and approve it. The FDA examines the results from the clinical trials to make a decision. Reasons for a drug to fail at this point include too toxic, insufficient efficacy, poor pharmacokinetics properties, poor bioavailability, or inadequate drug performance. If the drug is approved, the drug is launched on the market. The FDA continues to monitor the

drug post-market to ensure long-term safety, efficacy, and risks. Manufacturers, health professionals, and consumers are responsible for reporting issues to the FDA using the FDA Adverse Event Reporting System (FAERS).

### 1.1.4 Drug Repurposing

Drug repurposing is the investigation of pre-existing treatments with unanticipated effects as possible treatments for other diseases. It is an efficient approach since it decreases the number of steps required before clinical development, shortening the timeline and lowering the cost of drug discovery and development. Medications approved by the FDA that have shown beneficial effects in patients suffering from another disease can thus be considered a possible treatment option. New clinical trials are necessary to prove that the medication is effective for the novel purpose and does not cause side effects that people with the disease are susceptible to. A drug with repurposing potential may be discovered in preclinical/clinical research or by mining medical information databases.

The development of Parkinson's disease medication demonstrates a prime example of drug repurposing. Parkinson's disease is a progressive neurodegenerative disorder that affects dopamine levels in the brain. It is the second-most common neurological disorder, affecting more than 8 million people worldwide.[14] Dopamine-producing (dopaminergic) neurons are located in the substantia nigra structure of the basal ganglia, which is the part of the brain that controls movement. When people develop Parkinson's disease, these neurons become impaired or die. This results in lower dopamine production and hence causes movement problems. Symptoms are motor-related (e.g. shaking, tremors, muscle stiffness, and slowness) and nonmotor-related (e.g. cognitive impairment, mental health disorder and sleep disorders). The first approved drug repurposing treatment for Parkinson's disease was Amantadine. The FDA initially approved Amantadine for the treatment of the influenza virus. When patients with Parkinson's disease took this medication, it improved their symptoms. An International Linked Clinical Trials (iLCT) program for Parkinson's has been established to repurpose drugs to accelerate development.[15] A few FDA-approved drugs identified from this program are being considered as treatment options and are in clinical studies, including Ambroxol and Glucagon-Like Peptide-1 (GLP-1) agonists. Drug repurposing programs like this have a positive impact on the research community. It prompts the discovery and development of novel classes of drug candidates related to a repurposed biological target.[15]

## 1.2 Advances in Artificial Intelligence

Artificial Intelligence (AI) is the ability of a computer, computer-controlled robot, or software to perform a task that is typically associated with human intelligence. Typical intellectual human processes that machines could achieve include reasoning, discovering meaning, generalising, and learning from past experiences. Machines can process and extract relevant information from large amounts of data that are too large for humans to comprehend and interpret. AI is an object-achieving system that learns how to achieve a given goal by training on data. The more high-quality training data, the better the algorithm. For example, a machine can learn how to make a cake given only the ingredients and no recipe to follow. The output (response) of AI is dependent on the input variables. If you change the ingredients (input), the machine will make a different cake (output).

The computer pioneer Alan Turing completed significant early work on AI. In 1950, he proposed that in the future, machines such as digital computers would be capable of replicating human behaviour that was indistinguishable from a human being.[16] He devised the Turing test to determine such intelligence of machines. Turing described many concepts of AI in an unpublished report entitled "Intelligent Machinery". John McCarthy coined the term "Artificial Intelligence" in 1955 in his proposal for the first conference on AI at Dartmouth.

Growth in computing power, machine learning frameworks, data availability, and improved software and hardware has sparked enormous interest in the field of AI and particularly machine learning. Machine learning is a branch of AI which uses data-driven algorithms and analytics to build predictive models. Machine learning models are trained on a predefined dataset to learn patterns in the data without relying on rules. The training data can be labelled (supervised learning) or unlabelled (unsupervised learning). By learning from experience, a trained model can predict the output from a new set of inputs. Many non-linear algorithms are "black-box" methods, meaning it is difficult to determine the decision-making process behind the predictions. In the baking analogy, it is challenging to establish the recipe followed from the ingredients to a baked cake. The ability to comprehend why a machine learning model has made certain decisions or predictions is called interpretability. High interpretability is desirable for humans to trust the model and justify its use in daily life. Besides social acceptance, the ability to extract the additional knowledge captured by the model also enhances human understanding of the scientific topic.

Deep learning is a subcategory of machine learning that mimics brain neural

networks. Complex deep neural network algorithms are good at deciphering patterns and noise in large amounts of data. Well-known techniques include speech recognition, natural language processing, image recognition, and face recognition. They have many real-life applications, including virtual assistants, Identification (ID) validity, photo ID verification, and access control mechanisms in smartphone locks.

There have been several significant milestones that have advanced the capabilities of AI. In 1997, IBM's Deep Blue[17] supercomputer defeated a world-champion chess player. The expert system won by calculating every possible outcome, displaying the rapid evolution of computers. It took until 2016 for gaming capabilities to advance to the point where Deep Mind's AlphaGo[18] defeated the world's Go champion. AlphaGo is built on neural networks to analyse and learn while playing the game. A recent key milestone occurred in 2020 when Baidu's LinearFold[19] assisted in vaccine development during the early stages of the SARS-CoV-2 Covid-19 pandemic. LinearFold is an RNA folding algorithm. It predicted the secondary structure of the SARS-CoV-2 RNA sequence in under 30 seconds, 120 times faster than other methods.

AI is not limited to technological advances but also scientific advances, as demonstrated by LinearFold. Critical Assessment of protein Structure Prediction (CASP) is a forum that organises a biennial challenge for research groups to test protein-folding algorithms against experimental data not yet released to the public. Deep Mind's AlphaFold placed first in the 13$^{\text{th}}$ CASP (2018). AlphaFold is a deep learning algorithm that predicts the 3D structure of a protein from its amino acid sequence. A redesigned AlphaFold model demonstrated atomic accuracy in the 14$^{\text{th}}$ CASP (2020).[20,21] Applications of AI have extended to physics. Images taken by the James Webb Space Telescope (JWST) have recently been made public. Morpheus[22] is a deep learning algorithm that analyses data from the JWST to detect and classify galaxies in deep space. The neural networks in Morpheus are trained to classify every pixel in a JWST astronomical image and identify objects. In recent years, there has been a tremendous increase in the use of AI in chemistry.[23] Implementations of AI include predicting molecular properties, predicting bioactivities of new drugs, planning synthesis routes, optimising reaction conditions, and *de novo* drug design.[24–26]

AI, machine learning specifically, is a powerful tool with numerous applications. Compared to manually performing a repetitive task, automation through the use of AI is faster at decision-making, more efficient, available 24/7, and saves time. AI also improves accuracy and precision while reducing human error. Fewer errors save time and resources.

While AI has considerable scientific and technological benefits, AI costs money, time, and resources. Extensive initial investment is required to develop AI applications. Algorithms make predictions based on training data. If the training data is biased, the output will be discriminatory.

## 1.3 Computer-Aided Synthesis Planning

The application of AI and automation in chemical synthesis is an upcoming area to improve chemists' workflow. Reducing the timeline required to synthesise compounds allows more time for analysing and designing future experiments. Computer-Aided Synthesis Planning (CASP) tools exploit computational resources and mathematical algorithms to search through vast chemical and reaction search spaces. The tools intend to inform and inspire synthetic chemists while freeing time to focus on novel and complex problems, thereby improving productivity.

Chemical search engines, such as Reaxys and SciFinder$^{n}$, are examples of successful and well-integrated CASP tools. Their objective is to provide access to a wealth of knowledge derived from published literature, including journals, books, and patents. The tools search through a chemical database with millions of entries to find relevant chemical information and bioactivity data. Chemical compounds, reactions and properties can be retrieved by search engines, along with commercially available information. Credible citations supplement the recommended chemical data. The relevance and usefulness of information are difficult to capture. For example, chemical information highly relevant to a medicinal chemist may not be suitable for a process chemist. As a result, search solutions accommodate the need of the user through user-specific factors and preferences. Chemical search engines appeal to synthetic chemists due to the ease and quickness with which they can gather highly relevant and interpretable chemical information in a user-friendly and intuitive interface. Therefore, synthetic chemists can focus more time on conducting research and less time searching for relevant information.

Chemical search engines are a single type of CASP tool. Designing and optimising chemical reactions is usually based on synthetic chemists' knowledge, experience, and intuition. Increasing computational power and the establishment of reaction databases have led to advancements in data-driven decision-making tools. Reaction design and optimisation CASP tools have been developed to reduce the time and effort required to traverse through the vastness of chemical space via theoretically possible transformations. AI and machine learning algorithms have been

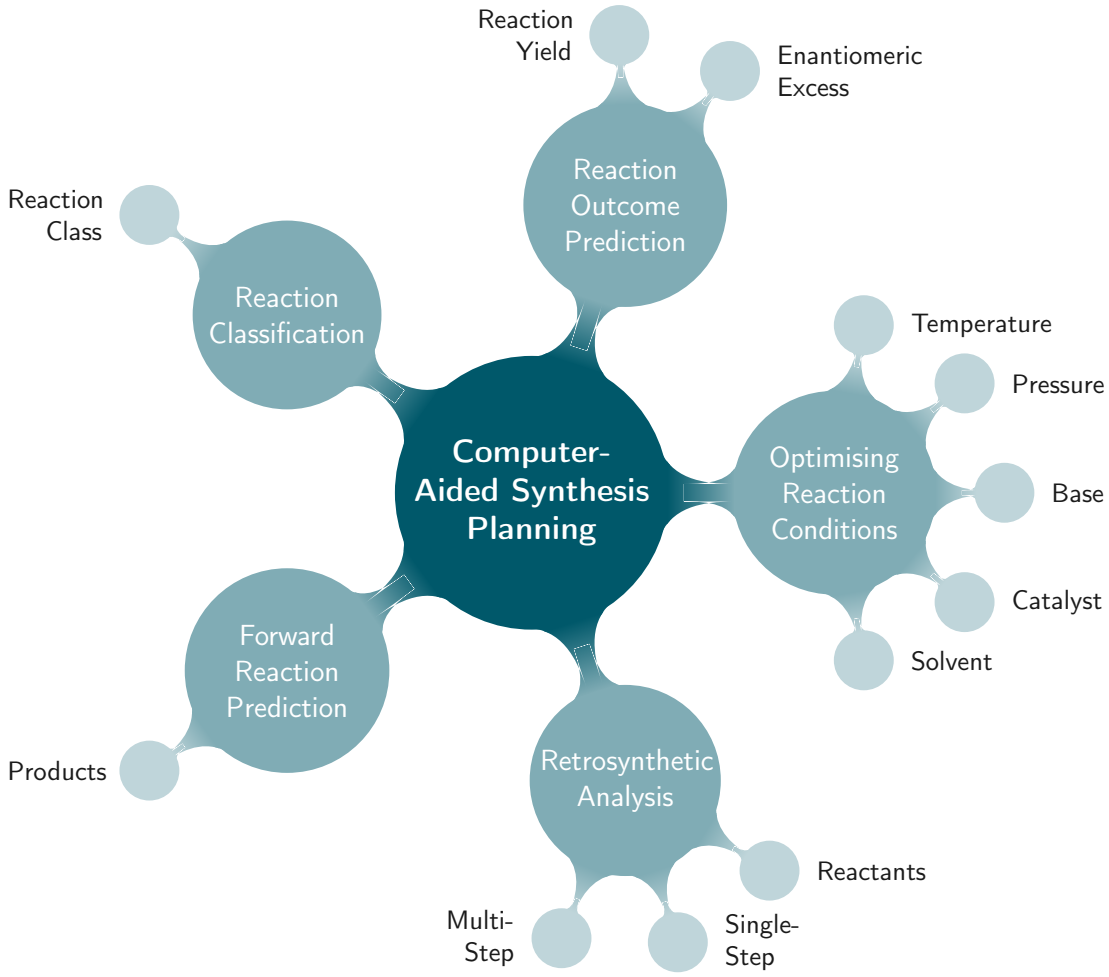



Figure 1.4: Overview of Computer-Aided Synthesis Planning (CASP) tools.

applied to predict chemical reactions and reaction outcomes, design retrosynthetic routes, classify reactions, and optimise reaction conditions (Figure 1.4).

The interest in CASP tools for forward reaction prediction and retrosynthetic route design has rapidly increased over the past seven years (Figure 1.5). Forward reaction prediction uses reactants, reagents, and conditions to predict the major product of a reaction. Retrosynthesis is the reverse of forward reaction prediction. Retrosynthetic route design involves the breakdown of a compound into precursors by disconnecting bonds or converting functional groups. Recursive breakdown occurs until the precursors are commercially available or in-house compounds or the pathway reaches a specified number of steps. Classification tools that identify the type of reaction occasionally supplement forward reaction prediction and retrosynthesis methods. CASP tools extend to the prediction of reaction outcomes and the optimisation of reaction conditions. Reaction outcomes include the reaction yield, Enantiomeric Excess (%ee), and selectivity. Reaction conditions include categorical variables such as catalysts and solvents and continuous variables such as temperature and pressure.

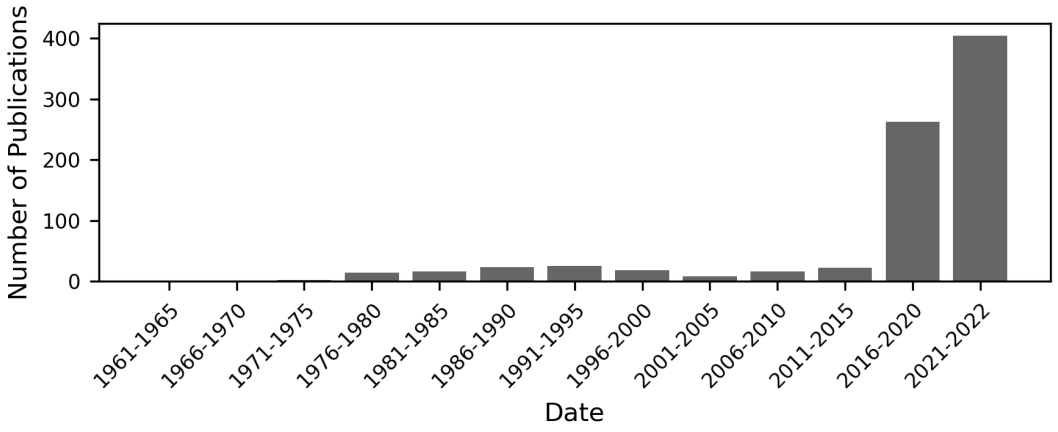



Figure 1.5: Number of publications about Computer-Aided Synthesis Planning (CASP) tools against time. Source, Google Scholar; date of access, 20/07/2023; publications contain at least one of the terms searched, 'computer-aided synthesis planning' or 'computer-assisted synthesis planning' or 'computer-aided retrosynthesis' or 'computer-assisted retrosynthesis' or 'computer-aided synthetic design' or 'computer-assisted synthetic design'.

CASP tools are continually improving and becoming readily accessible. Reaxys and SciFinder$^{n}$ have both integrated computer-aided retrosynthesis planning. Despite this, synthetic chemists remain wary of adopting CASP tools into their workflow. The benefits CASP tools can bring to synthetic chemists and ideal properties to increase integration into everyday decision-making are outlined below.

### 1.3.1 Integration of CASP

There is a role for CASP tools in drug discovery and development. CASP tools can determine synthetic pathways based on specific goals such as avoiding patent restrictions, complying with regulatory and environmental legislations, and achieving long-term sustainability. In drug discovery, the focus is on the ability to synthesise and functionalise scaffolds to a series of analogues for lead optimisation rather than on optimising reaction yield or selectivity. During hit identification and lead optimisation in drug discovery, chemists synthesise many compounds for assaying. Chemists must rapidly identify feasible synthetic routes. Ideally, when implementing de novo drug design tools, novel hits that are not synthetically accessible should be filtered out. Identifying synthetic routes to novel scaffolds which are not patented is crucial. Proposing reactions in a different context or inventing new chemical reactions would be beneficial. In drug development, the focus is on a few APIs. The aim is to identify optimal routes to meet Safety, Environmental, Legal, Economics, Control and Throughput (SELECT)

criteria.[27] The focus is on sustainability, atom economy, process mass intensity and material costs; while optimising reaction yield and selectivity and reducing the number of steps. Current efforts include green chemistry principles to suggest greener catalysts, reagents, and solvents while reducing energy consumption and waste.

Thus far, chemists are yet to implement these tools when planning synthetic routes to new molecules. A survey was completed in 2017 (13 chemists in two companies)[28] to determine what chemists expect from CASP tools. The survey revealed the following as the most desirable aspects to include. (i) A user-friendly interface that is easy to use, accessible, and has a shallow learning curve. (ii) Provide supportive literature examples with reasoning. (iii) A user-defined list of possible bonds to break. (iv) Lead the search to commercially available reactant precursors. (v) Recognise conflicting reactivity and propose protecting groups. (vi) Prioritise results based on user requirements. While aspects (i) to (v) can be implemented regardless of the chemist's area of expertise, prioritising the result is challenging. Chemists working in different parts of the chemical industry have distinct priorities when designing reactions. Their criteria for a suitable chemical reaction could include cost, greenness, reaction conditions (such as temperature or catalysts), selectivity, or the number of steps. CASP tools must be flexible and rank pathways based on criteria provided by the chemist. Providing a confidence score of predicted routes would also be valuable.

## 1.4 Scope and Objectives of Thesis

Advances in early drug discovery rely on the design and synthesis of novel molecules. Time, cost and efficiency pressures in the pharmaceutical industry are key drivers in accelerating drug design and development. Medicinal chemists are consequently biased toward robust reactions that are applicable to structurally diverse molecules and have a broad range of potential functionalisations, mild reaction conditions, and a reasonable time frame. The success of artificial intelligence and machine learning in other fields, such as image recognition and text processing, has sparked increased interest in their application to drug discovery.[29–31] This attention includes the design and synthesis of small molecules. The research in this thesis focuses on the development, evaluation, and growth of CASP to improve the workflow of medicinal chemists. Predicting reaction yield is an area of CASP reported significantly less in the literature. This thesis aims to explore machine learning models for the prediction of reaction yield. Building on the pioneering work of the Doyle group,[32–34] we investigate the applicability of structure-based

descriptors when building machine learning models to predict reaction yield. The specific objectives of each chapter are outlined below.

Chapter 2 describes the fundamentals of machine learning applied to chemistry-based research. An introduction to supervised learning is covered before delving into the mathematical details of individual regression techniques. We illustrate methods to represent the chemical structure of molecules, including machine-readable representations known as molecular descriptors. Methods to quantify the evaluation of the performance of machine learning models are also depicted.

Chapter 3 reviews distinguished CASP areas and compares existing state-of-the-art CASP tools. This chapter is an updated version of Machine Learning for Chemical Synthesis published in the Royal Society of Chemistry (RSC) book Machine Learning in Chemistry: The Impact of Artificial Intelligence.[35] Initially, an overview and a brief history of CASP is recounted. A discussion about current data sources and their potential limitations and progress follows. This literature review focuses on two types of CASP tools: forward reaction prediction and retrosynthetic analysis. The quantitative performance of current state-of-the-art methods on benchmark datasets is extracted from the literature and compared. This chapter concludes by emphasising existing issues and discussing the field's future direction.

Chapters 4 and 5 concentrate on the application of machine learning to predict reaction yield. These chapters constitute the bulk of my doctoral research, published in the American Chemical Society (ACS) Journal of Chemical Information and Modelling.[36]

Chapter 4 extends the work of Doyle *et al.* in developing and comparing the performance of machine learning models for the prediction of reaction yield.[32–34] This chapter evaluates several linear and non-linear machine learning algorithms. The aim of Chapter 4 is to examine whether the performance of models built on less computationally demanding, structure-based molecular descriptors is comparable to those built on the quantum chemical properties implemented by Doyle *et al.*

Chapter 5 implements a more rigorous evaluation technique to assess the performance of the Support Vector Regression (SVR) models identified as most promising in the preliminary evaluation conducted in Chapter 4. The aim of Chapter 5 is to construct a model to predict the reaction yield of unexplored Buchwald-Harwig reactions. We compare the performance of SVR models employing structure-based descriptors to models employing quantum chemical properties. The top SVR models built on each descriptor are subject to further external assessment.

Reaction yield values predicted by the top SVR models for the external test set are pending synthetic verification.

Chapter 6 summarises the outcomes and future direction of this research.

# Chapter 2

# Theory and Background

## 2.1 Introduction to Machine Learning

Artificial intelligence is the foundation of Computer-Aided Synthesis Planning (CASP) tools. It is a broad field regarding the ability of machines to imitate cognitive processes linked to human intelligence. Modern approaches to CASP tools incorporate machine learning, which is an application of artificial intelligence that enables machines to learn and improve through experience without explicit programming. Machine learning uses vast amounts of data, computer algorithms, and analytics to build predictive models. The three main types of machine learning are supervised, unsupervised, and reinforcement learning.

Supervised learning algorithms learn the relationship between input-output pairs. The input data is labelled, meaning the target variable is known. Regression and classification are the two types of problems supervised learning deals with. Regression algorithms are employed when the target data is continuous, and classification algorithms are employed when the target data is discrete. A supervised learning model aims to predict the output based on the input data.

Unsupervised learning algorithms identify underlying features and patterns in the input data without guidance. The input data is unlabelled, meaning there is no target variable. Clustering and dimensionality reduction are problems that fall under unsupervised learning. Clustering is the process of grouping data points and assigning segregated clusters such that similar data points lie in the same cluster. Dimensionality reduction is a technique to reduce the number of variables in the input data. An unsupervised learning model aims to interpret and organise the input data.

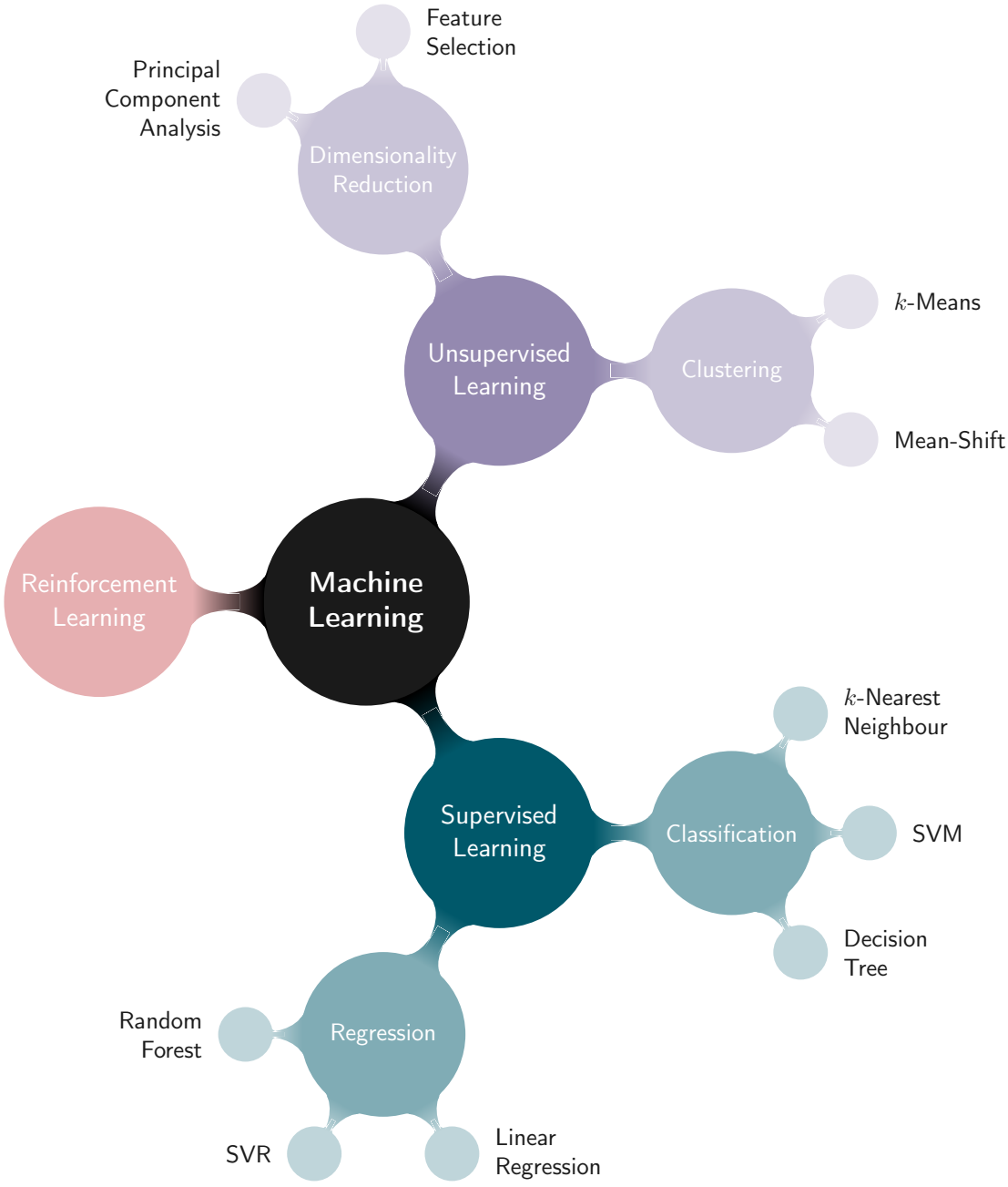



Figure 2.1: Three major types of machine learning: supervised, unsupervised, and reinforcement learning.

Reinforcement learning algorithms reward desired characteristics while penalising undesired ones. Through trial and error, an agent learns to perceive and interpret its environment using feedback from its decisions. These algorithms learn from outcomes to decide what action to take next. After each step, the algorithm receives feedback to determine whether the decision was beneficial. The objective is to maximise the rewards. Reinforcement learning encompasses exploitation or exploration, Markov's decision processes, and policy learning.

In the field of CASP, each of the three main types of machine learning has a role. Supervised learning can be employed to predict a predetermined variable, such as reaction yield using regression algorithms[32] or reaction type using classification algorithms.[37] Unsupervised learning can be used to cluster similar retrosynthesis pathways.[38] Reinforcement learning can be utilised to generate retrosynthesis pathways.[39]

The data used to build predictive models in this field contain chemical structures, often recorded in terms of their chemical composition and atomic configuration. These are known as molecular representations. Other notations, rather than encoding the exact structure of a compound, encode its physicochemical, structural, topological, or electronic structure. There are numerous types of molecular representation and descriptors. The most suitable one will depend on the task.

This chapter covers the theory and background of techniques used when applying machine learning to CASP. The work undertaken in this thesis focuses on developing a CASP tool to predict a predetermined value, the yield of chemical reactions. We first introduce the concept of supervised learning and how to build a predictive model in Section 2.2 before describing the regression algorithms in mathematical detail in Sections 2.3 and 2.4. We then focus on molecular representations, descriptors, and similarity measures commonly applied to cheminformatics in Sections 2.5 and 2.6. Finally, we provide details of performance evaluation metrics used to assess the generalisability of regression models in Section 2.7.

## 2.2 Supervised Machine Learning

Supervised machine learning algorithms model the relationship between independent variables and a dependent variable by maximising a performance criterion. Classification models are employed if the dependent variable is categorical, and regression models if continuous. Supervised machine learning methods are built on labelled data, known as training data, comprised of observed independent input values and corresponding dependent output. Training an algorithm deter-

mines a function (model) that describes the relationship between one dependent value (target) $y$ and two or more independent values (features) $\mathbf{x}$. A $p$-length vector of continuous or categorical values represents the features. Categories may be dummy-coded, whereby a number represents a single category. The trained model can then be used to predict the output values of novel inputs.

Parametric models assume the shape of the mathematical function, for example, linear. Making this assumption simplifies the optimisation of the function. The model uses training data to estimate the function's parameters. A disadvantage of parametric models is that the assumed form is not always the correct relationship of the data. If the presumed form is far from the unknown function, the performance of the model will be poor. Non-parametric models do not assume the function's form or shape. The function is estimated by fitting data points, allowing a broader range of functional forms. Compared to parametric models, non-parametric models require considerably more training examples to predict the function accurately. Detailed descriptions of machine learning algorithms can be found in *"The Elements of Statistical Learning"*, co-written by Hastie, Tibshirani and Friedman.[40]

### 2.2.1 Model Building

Initially, the prediction task is defined. In chemoinformatics, global models are designed for large datasets that cover broader chemical space, while local models are for small datasets with bias and narrow applicability range. The general process for building supervised machine learning models for a specified prediction task is illustrated in Figure 2.2.

#### Data Gathering

Machine learning algorithms require large amounts of training examples to perform well. When collecting the data, it is essential to consider its quantity, quality, coverage, diversity (bias), and reliability. The model's performance reflects these aspects. The model will overfit the training data if there is insufficient data. If the data is poor quality, the performance will be inaccurate. If the coverage is limited, the model will only perform well in a small region of feature space. If the data is biased and unreliable, the model will also be. Summary statistics give an overview of the data to capture its nature. For quantitative feature data, summary statistics include counts, mean, standard deviation, maximum data point, minimum data point, and percentile values. For categorical data, summary statistics include counts, the number of unique entries, the most frequent category and its value.

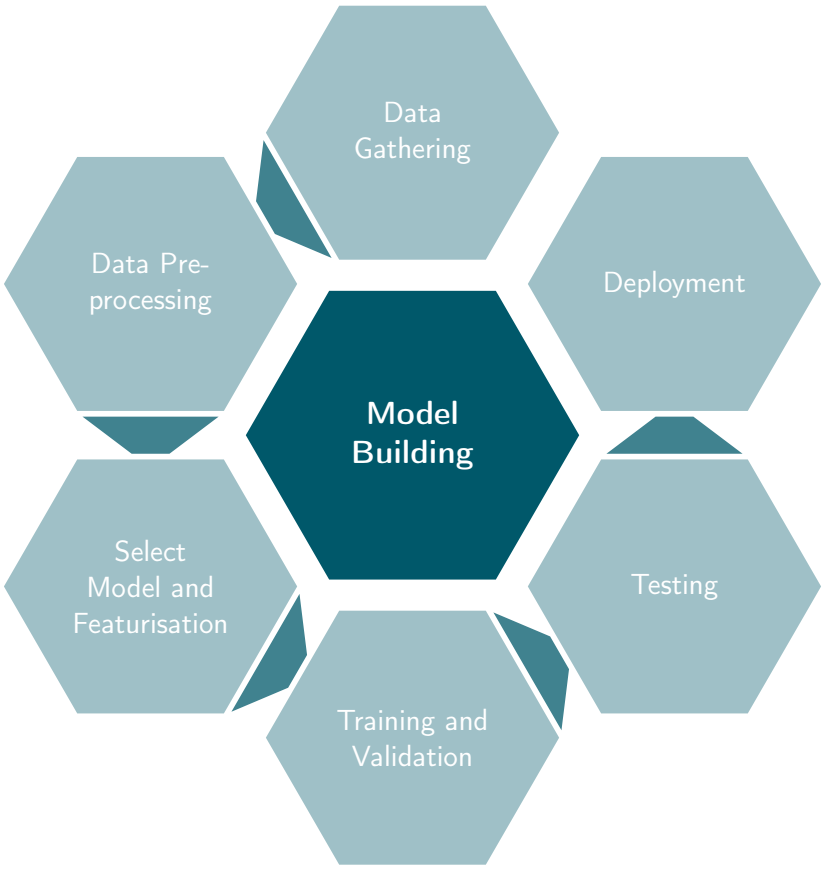



Figure 2.2: Overview of the model building process in supervised machine learning.

### Data Pre-processing

The collected data is cleaned and converted into a machine-readable format. Cleaning data entails detecting errors and removing unwanted, missing, and duplicate data. If the input variables are already in a numerical format, they can be implemented in their raw form or standardised if necessary. The data will not be in a machine-readable format in many circumstances. Numerous types of descriptors can represent the input variables. For cheminformatic-based problems, molecular descriptors can represent one-, two-, or three-dimensional structures of molecules.

Once the data is in an appropriate format, it is usually divided into training, validation, and test sets. The model is optimised using the training set. The validation set is then used to detect overfitting, select hyperparameter values, and compare models. Finally, the test set evaluates the performance of the model and should be representative of the application range. When data is scarce, the validation set may be bypassed. In this case, the hyperparameter values would be optimised using the training set.

Standard methods to split data include randomised, time-based, and leave-one-class-out splits. Taking a random split of the data is not always the best approach. In many circumstances, there is an uneven distribution of data. Random splitting on data with naturally clustered similar examples gives a too-optimistic view of performance. A time-split of the data is preferred whenever possible because it simulates how the model will be used in real life, with older training data than test data. Time splits work best with extensive datasets consisting of millions of examples. With fewer data, a time split can result in quite different distributions of data between training, validation, and test sets. Leave-one-class-out is a frequent solution to limited data. The features of the training set are initially grouped. Data points in one of the groups are held out of the training set and used as the test set.

### Select Model and Featurisation

Methodologies for selecting molecular descriptors or machine learning algorithms are yet to be standardised. Often it is necessary to exhaustively compare various algorithms with various molecular descriptors to identify the most suitable for the problem specified. Simpler, well-performing models should not be overlooked as they may provide better predictions than complex models prone to overfitting. Benchmarking the algorithms and descriptors is desirable. Baseline models may be pre-defined or generated during model development.

### Training and Validation

Machine learning frameworks are available to provide model implementation, such as scikit-learn for the Python programming language. During training, the model learns to identify patterns in the training data by determining the model's parameters. The model uses the training examples to adjust the parameters gradually.

Validation is an initial evaluation of the trained model. The quality of the model is checked against the validation set. With a sufficient dataset size, the validation set is a subset of the dataset generated in data splitting. The trained model predicts the validation set before calculating the performance statistics. If the dataset is not large enough to afford a validation set, validation can be achieved using cross-validation of the training set. Cross-validation is a technique to partition the training set into iterations of training and test sets. A popular technique called $k$-fold cross-validation divides the training data into $k$ equal groups. For each iteration of training and testing, a different group is the test data, and the remaining $k-1$ groups are the training data. The average performance statistics

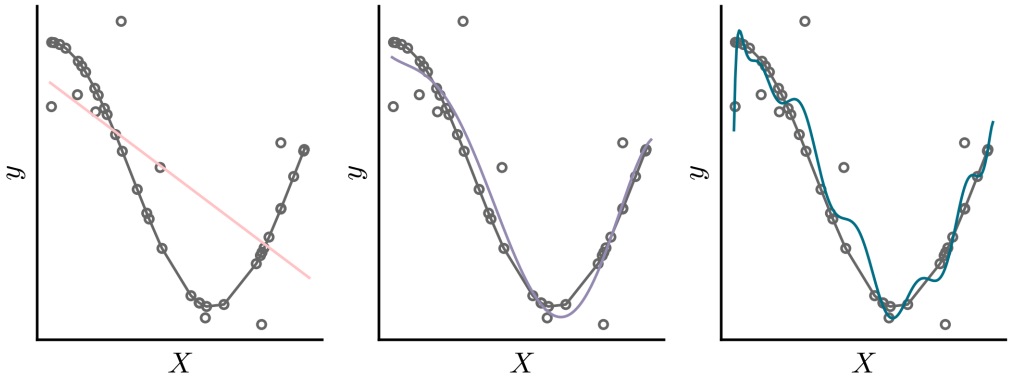



Figure 2.3: Overview of model fit. Light pink line, underfitting; purple, balanced; dark teal, underfitting.

of the $k$ models are calculated.

There is a fine balance between underfitting and overfitting the parameters in the model (Figure 2.3). A model is underfitting the training data if it performs poorly when the training data is used as the test set. Underfitting is when the model has not captured the complexity of the training data. An underfitted model does not capture the relationship between the input and target variables. The performance statistics from the test data can determine if the model is overfitting the training data. If the model performs well on the training data but poorly on the test data, the model is overfitting. Overfitting is when the model fits the training data too closely. The overfitted model is then unable to predict unseen examples accurately.

Validation, or cross-validation, is also used to determine hyperparameter values. Hyperparameters are parameters that are not optimised during training. Although the default values are helpful, tuning the hyperparameters could further improve the model's performance.

### Testing

The trained model is evaluated on the test set generated in data splitting to determine the performance on previously unseen data. The choice of performance metric may impact the quality of the model. Therefore, multiple metrics that fit the application should be evaluated.

The superior model and featurisation identified from testing may be subject to external testing to determine how well the model performs under real-life conditions. This additional testing uses an external test set, which was not part of the initially gathered data.

### Deployment

Once satisfied with the performance, the model can be deployed.[25] Deployment involves registering the model and generating documentation and guidelines for use. The model is integrated into existing tools and workflow with accessibility for non-data scientists where required. Model ownership and accountability, monitoring, and maintenance ensure the model is updated with newly available training data and any issues are fixed.

## 2.3 Parametric Regression Algorithms

Parametric algorithms assume the function of the mathematical model that relates the features with the targets. In this section, the regression algorithms discussed assume the relationship is linear. The models define the relationship between the target $y_i$ and the features $\mathbf{x_i}$, , given a dataset $\{y_i, x_{i1}, \ldots, x_{ip}\}_i^n = 1$ of $n$ statistical units (i.e., $n$ training samples), using the function

$$\begin{aligned} y_i &= w_0 + w_1 x_{i1} + \ldots + w_p x_{ip} + \epsilon_i \\ &= w_0 + \sum_{j=1}^{p} x_{ij} w_i + \epsilon \\ &= \mathbf{x}_i^\intercal \mathbf{w} + \epsilon_i \end{aligned}$$

for $i = 1, \ldots, n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_i p)$ is a $p$-length vector of features, $\mathbf{w} = (w_1, \ldots, w_p)$ is a $p$ length vector of coefficients, $w_0$ is the intercept and $\epsilon_i$ is the error term. For a dataset of $n$ samples, the equations for all samples can be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon \tag{2.1}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\intercal \\ \vdots \\ \mathbf{x}_n^\intercal \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$\mathbf{y}$ is the vector of observed target values $y_i$ $(i = 1, \ldots, n)$. $\mathbf{X}$ is a matrix of $n$ rows of $p$-feature vectors $\mathbf{x}_i$ $(i = 1, \ldots, n)$ where $x_{i0} = 1$. $\mathbf{w}$ is a $p + 1$ length vector of coefficients. $\epsilon$ is a vector of the error terms $\epsilon_i$ $(i = 1, \ldots, n)$. Regression coefficients are estimated in the training of the linear regression algorithm to fit a given dataset. From these estimates, predicted target values can be calculated using

$$\begin{aligned} \hat{y} &= \hat{w}_0 + \sum_{j=1}^{p} x_{ij} w_i \\ &= \mathbf{x}^\intercal \hat{\mathbf{w}}. \end{aligned}$$

### 2.3.1 Multiple Linear Regression

#### Ordinary Least Squares

Ordinary Least Squares (OLS) regression aims to identify the line of best fit with the lowest error, known as the Least Square Regression Line (LSRL). The OLS model estimates the parameters of the linear function by minimising the Residual Sum of Squares (RSS). Residuals are the differences between the observed targets $y$ and targets predicted by the linear function of features $\hat{y}$. The residual of the $i^{\text{th}}$ training point is given by $\epsilon_i = y_i - \hat{y}_i$. Figure 2.4 illustrates of squared residuals for a simple linear regression model. The equation for calculating the sum of

squared residual is given below.

$$\begin{aligned}\text{RSS} &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left( y_i - w_0 - \sum_{j=1}^{p} x_{ij} w_j \right)^2 \\ &= \sum_{i=1}^{n} (y_i - \mathbf{x_i}^\intercal \mathbf{w})^2\end{aligned}$$

The equation for calculating RSS can be rewritten in the $l_p$-norm notation, $\|\mathbf{x}\|_p = (\sum_i |x_i|^p)^{1/p}$, as shown in (Equation 2.2). The estimated multiple least squares regression coefficients $\mathbf{w}$ that minimise Equation 2.2 can be used to predict the target of novel input features.

$$\begin{aligned}\text{RSS} &= (\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2)^2 \\ &= \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2\end{aligned} \tag{2.2}$$

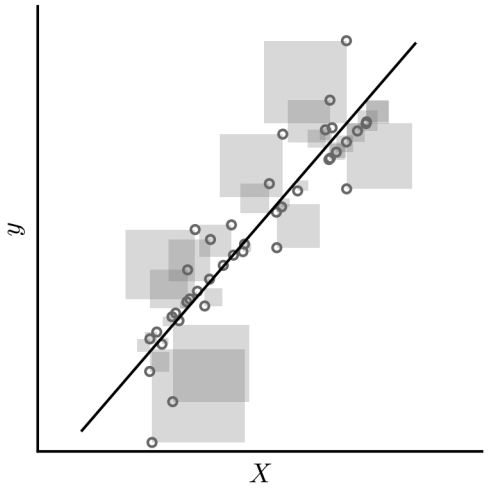



Figure 2.4: Illustration of squared residuals that are minimised in linear regression. Scatter points, observed data; line, least square regression line $\mathbf{y} = \mathbf{w}\mathbf{X}$; translucent squares, squared residuals.

The OLS method is easy to implement, efficient to compute, and straightforward to interpret. While OLS performs well when the training data is linearly separable, it has several limitations. The method is sensitive to outliers. The presence

of these extreme target values in the training data can cause drastic effects on the resulting trained model due to the minimisation of the sum of squared errors. The magnitude of the residuals will be squared during the sum of squares calculation, meaning the model will try to reduce these extensive residual errors in training. Extreme errors alter the least squared solution the most, significantly affecting the final model.

The OLS method assumes there is no relationship between the independent variables. When features are correlated, changing one independent variable causes changes in the others. A linear regression model trained on collinear data produces an unstable model. It can be challenging to differentiate between the independent effects of the collinear features with the target. Multiple solutions are considered acceptable, resulting in less precise coefficient estimation with higher uncertainty. The unstable model becomes harder to interpret and may cause overfitting of the data, resulting in poor performance on an external test set.

Another issue arises when there are too many independent variables. As the number of features approaches or exceeds the number of training examples, the model overfits the data by fitting both the underlying structure of the data and noise. Using an extensive number of features requires plenty of training examples to distinguish between the features correlated with the output and those merely correlated by chance. Including redundant features will impair the model and should be avoided.

### LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) is a regularisation technique used to overcome overfitting in linear regression and improve generalisability. This method employs shrinkage, which reduces the regression coefficients towards zero. An $l_1$ regularisation term is added as a constraint to the residual sum of least squares calculation, equal to the $l_1$-norm of the coefficient vector. The optimisation problem solved by LASSO is the minimisation of

$$\begin{aligned} \hat{w}^{\text{lasso}} &= \underset{w}{\operatorname{argmin}} \left\{ \text{RSS} + \alpha \left( l_1\text{-norm} \right) \right\} \\ &= \underset{w}{\operatorname{argmin}} \left\{ \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \alpha \|\mathbf{w}\|_1 \right\} \end{aligned} \tag{2.3}$$

where RSS is the sum of squared residuals, $\|\mathbf{w}\|_1$ is the $l_1$-norm and $\alpha \geq 0$ is a non-negative pre-defined regularisation parameter. The value of $\alpha$ is tuned as a hyperparameter. The $l_1$-norm is defined as the sum of the absolute value of the

magnitude of the components in a vector, as shown below.

$$\|\mathbf{w}\|_1 = \sum_{j=1}^{p} |w_j|$$

LASSO essentially translates each coefficient by a factor of $\alpha$ and truncates it at zero. The intensity of the $l_1$ constraint is determined by the constant $\alpha$, where $\alpha = 0$ is the OLS optimisation problem. As $\alpha$ increases, a larger proportion of the coefficients will equal zero and be eliminated from the model. If $\alpha$ is sufficiently large enough, all coefficients will equal zero, resulting in a null model.

The enhanced performance of the LASSO method over the OLS method is attributable to the trade-off between the variance and the bias of the LASSO method. Variance is the amount a function would change when using a different training set. If minor changes to the training set cause large changes to the estimated function, the model has high variance and is considered unstable or flexible. Bias is the error in the predicted targets compared to the training examples due to oversimplifying the complex relationship between the features and the target. Bias is usually a result of assuming the relationship is linear. If the predictions significantly differ from the actual training values, for example, when predicting a non-linear relationship with a linear model, the resulting model will have a high bias.

As $\alpha$ increases, the model variance decreases and becomes more stable, although at the expense of a slight bias increase. The reduction in model complexity helps overcome the overfitting of the training data and mitigate multicollinearity present in the features. The sparse models generated by LASSO regression have fewer non-zero coefficients and hence are reliant on a smaller number of features. This variable selection results in more interpretable models. The LASSO method becomes limited if multicollinearity, where one feature correlates with multiple other features, is present in the training data. The LASSO model tends to focus on one of the features of the group rather than considering all of them.

### Ridge Regression

Ridge regression is a regularisation technique based on shrinkage, similar to LASSO. The coefficients are proportionally shrunk toward zero. Generally, the resulting coefficients are all non-zero. Therefore, sparse models, reliant on fewer features, cannot be attained with the ridge regression method. An $l_2$ regularisation term is added to the OLS optimisation problem. In ridge regression, coeffi-

cients are estimated by minimising a penalised residual sum of squares

$$\begin{aligned}\hat{w}^{\text{ridge}} &= \underset{w}{\operatorname{argmin}} \left\{\text{RSS} + \alpha \left(l_2\text{-norm}\right)^2\right\} \\ &= \underset{w}{\operatorname{argmin}} \left\{\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \alpha\|\mathbf{w}\|_2^2\right\}\end{aligned} \tag{2.4}$$

where RSS is the sum of squared residuals, and $\alpha \geq 0$ is a non-negative pre-defined regularisation parameter. The value of $\alpha$ is tuned as a hyperparameter. The square of the $l_2$-norm, $\|\mathbf{w}\|_2^2$, is equal to the sum of the square of the magnitude of the components in the coefficient vector (see below).

$$\begin{aligned}\|\mathbf{w}\|_2^2 &= \left(\left(\sum_{j=1}^{p} |w_j|^2\right)^{1/2}\right)^2 \\ &= \sum_{j=1}^{p} |w_j|^2\end{aligned}$$

The penalty is not applied to the intercept $w_0$. The intercept is essentially the mean of the training target values with the features centred to zero.

When $\alpha$ is zero, the optimisation problem is the same as OLS, and the $l_2$ penalty term has no effect. As $\alpha$ increases ($\alpha \to \infty$), the influence of the $l_2$ penalty increases. The resulting decrease in the coefficient estimates toward zero during optimisation reduces model flexibility, resulting in lower variance but higher bias. As a result, a minor change in the training data should not cause a significant change in the regression coefficients. Reducing the complexity of the model reduces data overfitting and improves generalisability until a tipping point where the trade-off between variance and bias begins to cause underfitting. The ridge regression method is advantageous when the number of features $p$ is close to, or larger than, the number of training examples $n$. In this case, the OLS method will be very flexible, whereas the ridge method can perform well by trading off small increases in bias for large decreases in variance.

### Elastic-Net

The elastic-net method adds $l_1$ and $l_2$ regularisation techniques to the OLS optimisation equation.

$$\begin{aligned}\hat{w}^{\text{elastic net}} &= \underset{w}{\operatorname{argmin}} \left\{\text{RSS} + \alpha_1 \left(l_1\text{-norm}\right) + \alpha_2 \left(l_2\text{-norm}\right)^2\right\} \\ &= \underset{w}{\operatorname{argmin}} \left\{\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \alpha_1\|\mathbf{w}\|_1 + \alpha_2\|\mathbf{w}\|_2^2\right\}\end{aligned} \tag{2.5}$$

The $\alpha_1$ and $\alpha_2$ regularisation parameters control the intensity of the $l_1$ and $l_2$ constraints, respectively. When $\alpha_1 = 0$, the optimisation problem is ridge regression; when $\alpha_2 = 0$, the optimisation problem is LASSO regression. The optimisation equation can be rewritten to depend on a single regularisation parameter $\alpha$. The value of $\alpha$ is tuned as a hyperparameter. A ratio parameter describes the ratio between the $\alpha_1$ and $\alpha_2$ parameters, known as the $l_1$-ratio.

$$\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \alpha\,(l_1\text{-ratio})\,\|\mathbf{w}\|_1 + \alpha\,(1 - l_1\text{-ratio})\,\|\mathbf{w}\|_2^2 \tag{2.6}$$

When the $l_1$-ratio $= 0$, the optimisation problem is the same as ridge regression and when the $l_1$-ratio $= 1$, it is the same as LASSO regression. Intermittent values of the $l_1$-ratio determine a trade-off between the $l_1$-norm and the squared $l_2$-norm of the coefficient vector.

Implementing the $l_1$ regularisation technique allows the elastic-net model to be sparse and coefficients to be reduced to zero for redundant features. Adding the $l_2$ penalty overcomes limitations in the LASSO method when multiple features are correlated. The elastic-net method, therefore, performs variable selection and regularisation simultaneously, which is advantageous when the number of features is larger than the number of training examples.

### 2.3.2 Linear Support Vector Regression

In epsilon-Support Vector Regression ($\epsilon$-SVR), the aim is to find a function

$$\mathbf{y} = \mathbf{w}\mathbf{X} + w_0$$

that describes the relationship between the observed target values $\mathbf{y}$ and observed features $\mathbf{X}$ by fitting the error within a certain threshold $\epsilon$, where $\mathbf{w} = (w_1, \ldots, w_p)$ is a vector of coefficients and $w_0$ is the intercept. This linear function is known as the hyperplane and is shown in Figure 2.5 as a solid line. The observed variables should deviate by a maximum distance $\epsilon$ from the hyperplane. This defines an area known as the $\epsilon$-insensitive tube or $\epsilon$-tube, which sets a margin for the observed variables. The boundary of the $\epsilon$-tube is defined as

$$\mathbf{y} = \mathbf{w}\mathbf{X} + w_0 + \epsilon$$
$$\mathbf{y} = \mathbf{w}\mathbf{X} + w_0 - \epsilon$$

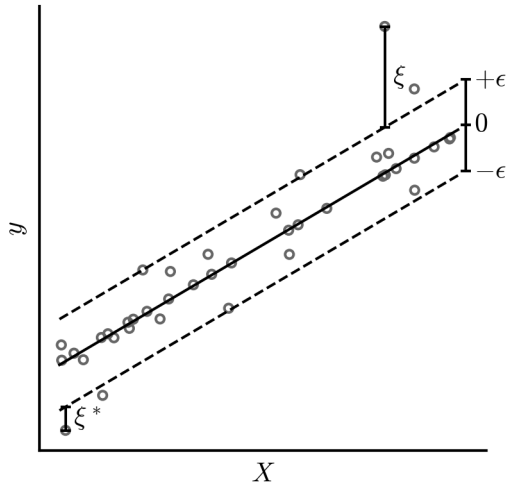



Figure 2.5: Illustration of linear support vector regression. Scatter points, observed data; solid line, hyperplane $\mathbf{y} = \mathbf{wX} + w_0$; dotted lines, boundary lines $\mathbf{y} = \mathbf{wX} + w_0 + \epsilon$ and $\mathbf{y} = \mathbf{wX} + w_0 - \epsilon$.

and can be observed in Figure 2.5 as two dashed boundary lines. The hyperplane, therefore, satisfies the constraints below.

$$\mathbf{y} - \mathbf{wX} - w_0 \leq \epsilon$$
$$\mathbf{y} - \mathbf{wX} - w_0 \geq -\epsilon$$

The linear hyperplane is determined by a convex optimisation problem to minimise the norm value

$$J(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\intercal\mathbf{w} = \frac{1}{2}\|\mathbf{w}\|^2$$

subject to all residuals being less than $\epsilon$ as shown in Equation (2.7).

$$\min_{\mathbf{w}} J(\mathbf{w}) = \min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 \tag{2.7}$$

It may be possible that no function would satisfy these constraints, and functions that are found may overfit the observed data. Slack variables ($\xi$ and $\xi^*$) are introduced to allow for errors larger than $\epsilon$. Slack variables are defined from the boundary lines (Figure 2.5) and, therefore, can only be greater than or equal to zero:

$$\xi \geq 0$$
$$\xi^* \geq 0.$$

If the error is above the $\epsilon$-tube it is denoted as $\xi$ and has the format:

$$y_i - \mathbf{x}_i^\intercal \mathbf{w} - w_0 \leq \epsilon + \xi_i.$$

If the data point is below the $\epsilon$-tube it is denoted as $\xi^*$ and has the format:

$$y_i - \mathbf{x}_i^\intercal \mathbf{w} - w_0 \geq -\epsilon - \xi_i^*.$$

Any data points within the $\epsilon$-tube are considered to have an error of zero. The data points $x_i$ that fall outside of the $\epsilon$-tube are called support vectors and are used to define the hyperplane by contributing to the objective function,[41] also known as the primal formula (Equation 2.8).

$$J(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2.8}$$

Equation 2.9 describes the optimisation problem solved during the training of the SVR algorithm. The hyperparameter $C$ can be tuned to determine the toleration of points outside of $\epsilon$; as $C$ increases, the tolerance increases. This parameter helps to prevent overfitting (regularisation) and is considered a trade-off between the flatness of the hyperplane and the tolerance of deviations larger than $\epsilon$.

$$\begin{aligned} \min_{\mathbf{w},\xi,\xi^*} \quad & J(\mathbf{w}) \\ \min_{\mathbf{w},\xi,\xi^*} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{subject to} \quad & y_i - \mathbf{w}^T x_i \;\leq\; \epsilon \;+\; \xi_i \\ & \mathbf{w}^T x_i - y_i \;\leq\; \epsilon \;+\; \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \\ & C > 0 \end{aligned} \tag{2.9}$$

The primal optimisation problem is simpler to solve computationally in its dual formulation. A Lagrange function $L(\alpha)$ can be constructed from the primal objective function (Equation 2.8) by introducing positive Lagrange multipliers ($\alpha$ and $\alpha^*$) for each feature:

$$\begin{aligned} L(\alpha) = \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i + \alpha_i^*)(\alpha_j + \alpha_j^*)(\mathbf{x_i}\mathbf{x_j}) \\ + \epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{n} y_i(\alpha_i + \alpha_i^*) \end{aligned}$$

The construction of the Lagrange function is detailed in Smola and Schölkopf.[41] The dual optimisation problem is thus given by Equation 2.10.

$$\begin{aligned} \min_{\alpha,\alpha^*} \quad & L(\alpha) \\ \min_{\alpha,\alpha^*} \quad & \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i+\alpha_i^*)(\alpha_j+\alpha_j^*)(\mathbf{x_i}\mathbf{x_j}) \\ & +\epsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)-\sum_{i=1}^{n}y_i(\alpha_i+\alpha_i^*) \\ \text{subject to} \quad & \sum_{i=1}^{n}(\alpha_i+\alpha_i^*)=0 \\ & 0\le\alpha_i\le C \\ & 0\le\alpha_i^*\le C \end{aligned} \tag{2.10}$$

The parameter $\mathbf{w}$ can be written as a linear combination of the observed training features $\mathbf{x}_i$:

$$\mathbf{w}=\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)\mathbf{x}_i$$

The equation of the hyperplane can be rewritten as

$$y=\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)(\mathbf{x}_i\mathbf{x})+w_0.$$

Lagrange multipliers ($\alpha$ and $\alpha^*$) are set to zero for data points inside the $\epsilon$-tube. Data points outside the $\epsilon$-tube have positive Lagrange multipliers, known as support vectors.

## 2.4 Non-Parametric Regression Algorithms

The shape of the mathematical function assumed in parametric models is not always the same as the actual relationship between a set of features and targets. Non-parametric models do not assume the form of the function and hence can fit a broader range of forms.

### 2.4.1 Support Vector Regression

The relationship between the observed targets and features is not always linear and, therefore, cannot always be described by a linear model (Figure 2.6). In non-linear SVR, a kernel function maps the input data to a higher-dimensional feature space where linear regression is performed. A non-linear kernel function

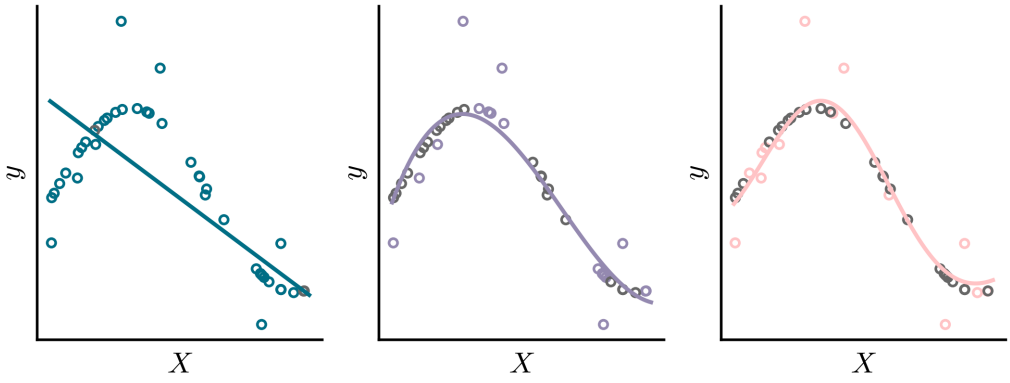



Figure 2.6: Illustration of support vector regression implementing the linear (dark teal), polynomial (purple) and RBF (light pink) kernels, on a one dimensional dataset with 40 data points. Scatter points, observed data; coloured scatter points, support vectors; solid line, hyperplane $y = \sum_{i=1}^{n}(\alpha_i + \alpha_i^*)K(\mathbf{x}_i, \mathbf{x}) + w_0$.

is defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \theta(\mathbf{x}_i), \theta(\mathbf{x}_j) \rangle$$

where $\theta(\mathbf{x})$ maps $\mathbf{x}$ to high-dimensional space. Examples of non-linear kernel functions include polynomial, Gaussian Radial Basis Function (RBF), and sigmoid functions. The sigmoid equation is not a valid kernel but has been applied successfully; see Schölkopf[42] for further details. The equations of these kernel functions between two data points ($\mathbf{x}_i$ and $\mathbf{x}_j$) are shown in Table 2.1.

Table 2.1: Kernel equations on two data points $\mathbf{x}_i$ and $\mathbf{x}_j$

| Kernel Name | Equation, $k(\mathbf{x}_i, \mathbf{x}_j)$ |
|---|---|
| Linear | $\mathbf{x}_i^\intercal \mathbf{x}_j$ |
| Polynomial | $(\gamma_p(\mathbf{x}_i^\intercal \mathbf{x}_j) + c_p)^d$ |
| RBF | $\exp\left(-\gamma_r \lVert \mathbf{x}_i - \mathbf{x}_j \rVert^2\right)$ |
| Sigmoid | $\tanh\left(\gamma_s(\mathbf{x}_i^\intercal \mathbf{x}_j) + c_s\right)$ |

The non-linear kernel function $K(\mathbf{x_i}, \mathbf{x_j})$ replaces the dot product ($\mathbf{x}_i\mathbf{x}$) in the

dual formula (Equation 2.11).

$$\begin{aligned} \min_{\alpha,\alpha^*} \quad & \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i+\alpha_i^*)(\alpha_j+\alpha_j^*)K(\mathbf{x_i},\mathbf{x_j}) \\ & +\epsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)-\sum_{i=1}^{n}y_i(\alpha_i+\alpha_i^*) \\ \text{subject to} \quad & \sum_{i=1}^{n}(\alpha_i+\alpha_i^*)=0 \\ & 0\le\alpha_i\le C \\ & 0\le\alpha_i^*\le C \end{aligned} \tag{2.11}$$

The equation of the hyperplane can be rewritten as

$$y=\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)K(\mathbf{x}_i,\mathbf{x})+w_0.$$

### 2.4.2 $k$-Nearest Neighbours

The aim of $k$-nearest neighbours is to find the nearest $k$ training points to a new prediction point and calculate the average target value. The nearest $k$ training points are identified by calculating the distance between the features of the new point and the features of the training points. The distance between the features of a training point $\mathbf{x}_i$ and a test point $\mathbf{x}_j$ can be calculated by metrics such as Euclidean distance, Manhattan distance or Minkowski distance (Equation 2.12a to 2.12c).

$$\text{Euclidean distance} \qquad \sqrt{(\mathbf{x}_i-\mathbf{x}_j)^2}=\sqrt{\mathbf{x}_i^2-2\mathbf{x}_i\mathbf{x}_j+\mathbf{x}_j^2} \tag{2.12a}$$

$$\text{Manhattan distance} \qquad |\mathbf{x}_i-\mathbf{x}_j| \tag{2.12b}$$

$$\text{Minkowski distance} \qquad \left[\sum(w|\mathbf{x}_i-\mathbf{x}_j|^p)\right]^{1/p} \tag{2.12c}$$

The distances are sorted, and the closest $k$ points are selected ($N_0$). The predicted target value of the new point $x_0$ is the average of these selected points, as shown below.

$$\hat{f}\,(x_0)=\hat{y}_0=\frac{1}{k}\sum_{x_i\in N_0}y_i$$

For example, in Figure 2.7, the predicted value represented by the black cross is calculated from the average of the $k$ closest points ($k=1,5,10$), highlighted by coloured scatter points. As $k$ increases, the predicted function becomes smoother and less flexible, resulting in a lower variance but higher bias. The value of $k$ is

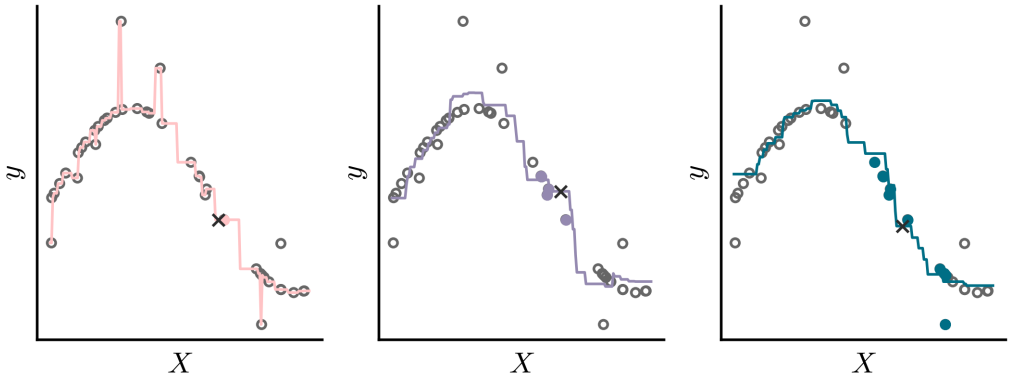



Figure 2.7: Illustration of $k$-nearest neighbours implementing $k = 1$ (light pink), $k = 5$ (purple) and $k = 10$ (dark teal), on a one-dimensional dataset with 40 data points. Scatter points, observed data; black cross, predicted data point; coloured scatter points, nearest training points to the predicted data point; solid line, predicted relationship between the features and the target.

tuned as a hyperparameter.

The computation of the distances is typically completed using one of the following algorithms: brute force, $k$-dimensional ($k$D) tree[43] or ball tree. The brute force algorithm calculates distances between the new point and all training points. For large datasets, this method can become infeasible.

The $k$D tree algorithm encodes implicit information to reduce the number of distances that require computing. The $k$D tree algorithm partitions the training data into a binary tree structure along each dimension $k$. An example dataset $\mathbf{x} = \begin{pmatrix} x_1 & x_2 \end{pmatrix}$ with two dimensions $x_1$ and $x_2$ ($k = 2$) is shown in Figure 2.8a. The corresponding $k$D binary tree is shown in Figure 2.8b. The levels of the binary tree are known as discriminators and range from zero to $k - 1$. At each level of the binary tree, the discriminator alternates between each dimension, $x_1$ and $x_2$ in the example shown. The nodes in the binary tree represent a split through the data along the dimension determined by the level. The root node corresponds to the split of the data along the $x_1$ axis at point $l_1$. Data points with $x_1 < l_1$ will be on the left (subtree to the left of the root), and data points with $x_1 > l_1$ will be on the right (subtree to the right of the root). Each side of $l_1$ is split along the $x_2$ dimension, represented by the two child nodes labelled $l_2$ and $l_3$. The leaf nodes $R_1$ to $R_4$ correspond to the final subsets of the training data. The $k$D tree is traversed to determine the nearest $k$ points by locating the boundary box that the new point is in. Distances are calculated between the new point and training points in this boundary box and training points in neighbouring boundary boxes. From these distances, the closest $k$ training points are identified. The $k$D tree algorithm can become infeasible for higher dimensional data.

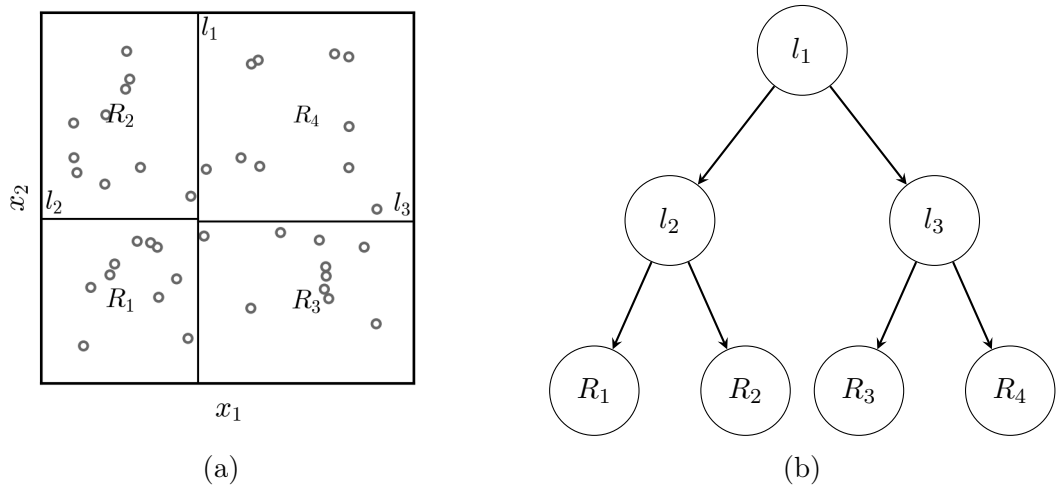



Figure 2.8: Illustration of the $k$D algorithm. $l_1$, Split 1 along $x_1$ dimension; $l_2$ and $l_3$, Split 2 and 3 along $x_2$ dimension; $R_1$ to $R_4$, final subsets of the dataset. (a) Plot of two-dimensional (2D) data $(x_1, x_2)$ with 40 data points. Scatter points, 2D data; solid lines, $k$D tree nodes. (b) Binary tree structure. $l_1$, root node; $l_2$ and $l_3$, inner nodes; $R_1$ to $R_4$, leaf nodes.

The ball tree algorithm is efficient for organising data points in high dimensional space. It constructs a binary tree, similar to the $k$D tree (Figure 2.8), except the data is partitioned into $n$-dimensional hyperspheres rather than boxes. Initially, the data is split into two hyperspheres where any $n$-dimensional point will belong to only one of these spheres, even if these spheres intersect one another. Each data point belongs to the hypersphere with the closest centroid. The hyperspheres can each be divided in two again to create sub-hyperspheres. The distance between the data points and the centroid of the sub-hyperspheres will determine which sub-hypersphere the point belongs. This splitting process is repeated until a certain depth is reached. For a new point, the tree is explored until a leaf node is reached. The distances between the new point and the points inside this hypersphere are calculated, and the closed $k$ points are identified.

The dimensionality of the feature space hinders the performance of $k$-nearest neighbours. A finite number of training data in high dimensional feature space leads to sparse data, known as the curse of dimensionality. Increasing the dimensionality of the features causes the closest distance between two points to approach the average distance of all points. The model performance decreases as it will only be slightly better than taking the average target value of the dataset.

### 2.4.3 Decision Trees

Tree-based models predict target values using binary trees where the feature space is split based on specific rules into non-overlapping regions. The algorithm is known as the Classification and Regression Trees (CART) algorithm.[44] The root node represents the entire dataset and is split into two or more sub-nodes following a rule. A divided node is the parent node, and the resulting sub-nodes are the child nodes. The feature space $\mathbf{X}$ is recursively partitioned to form a tree structure where the intermediate subsets are internal nodes and the final $M$ subsets of the dataset are leaf nodes $\mathbf{R} = (R_1, \ldots, R_M)$. Training the decision tree model defines the rules for splitting at each non-leaf node. At each leaf node $m$, the subset of the training data $R_m$ is represented by a constant $c_m$, corresponding to the predicted target value. The decision tree model is given by Equation 2.13. Depending on the loss function used to determine the splits of the data, $c_m$ is either the mean or median of the training data. The indicator function $I()$ returns one if true and zero otherwise.

$$f(\mathbf{x}) = \sum_{m=1}^{M} c_m I\left(\mathbf{x} \in R_m\right) \tag{2.13}$$

The decision tree recursively splits the feature space so that each child node has similar observed target values. This greedy approach is called recursive binary splitting. In the training of the algorithm, the splitting rules are determined. Candidate splits are explored for each division of feature space, and their quality is evaluated to determine the best division. A candidate division is denoted as

$$\theta = (j, t_m)$$

where $j$ is the feature of the candidate and $t_m$ is the threshold value for node $m$. Given data $Q_m$ at node $m$ with $n_m$ samples, the resulting child nodes (subsets) are denoted as $Q_m^{left}(\theta)$ and $Q_m^{right}(\theta)$. The left child node will contain data with features lower than or equal to the threshold value

$$Q_m^{left}(\theta) = \{(x, y) | x_j \leq t_m\}$$

and the right child node the remaining data

$$\begin{aligned} Q_m^{right}(\theta) &= Q_m \setminus Q_m^{left}(\theta) \\ &= \{(x, y) | x_j > t_m\}. \end{aligned}$$

A loss function $H(Q_m)$ is used to assess the quality of the candidate split $G(Q_m, \theta)$ of node $m$.

$$G(Q_m, \theta) = \frac{n_m^{left}}{n_m} H(Q_m^{left}(\theta)) + \frac{n_m^{right}}{n_m} H(Q_m^{right}(\theta))$$

Minimise the loss function

$$\theta^* = \text{argmin}_\theta G(Q_m, \theta)$$

to ensure data points in each child node have similar observed target values while the two child nodes are as different as possible. For a regression task, this loss function identifies the extent the predictions deviate from the observed target values and can be the Mean Squared Error (MSE), Poisson deviance or Mean Absolute Error (MAE). The MSE and Poisson loss functions set the predicted values of the leaf nodes $c_m$ to the mean $\bar{y}_m$ of the training values and can be calculated by

$$\begin{aligned} \text{MSE} \qquad & H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} (y - \bar{y}_m)^2 \\ \text{Poisson Deviation} \qquad & H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} \left( y \log \frac{y}{y_m} - y - \bar{y}_m \right)^2 \\ \text{where} \qquad & \bar{y}_m = \frac{1}{n_m} \sum_{y \in Q_m} y \end{aligned}$$

where $n_m$ is the number of training samples in node $m$, $y$ are the observed target values in the data $Q_m$ and $\bar{y}_m$ is the predicted target value given by the mean of the data $Q_m$. The MAE loss function sets the predicted values of the leaf nodes to the median median$(y)_m$ and is calculated by

$$\begin{aligned} \text{MAE} \qquad & H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} |(y - \text{median}(y)_m| \\ \text{where} \qquad & \text{median}(y)_m = \underset{y \in Q_m}{\text{median}}(y)_m \end{aligned}$$

The search and split process is repeated to build the decision tree until a stop criterion is met. Possible stop criteria include minimum training examples in a leaf node or a maximum tree depth.

Decision trees have several advantages over other regression algorithms. They are simple to understand and easy to interpret. The algorithm mirrors human decision-making and can be displayed graphically to aid understanding and interpretability. Decision trees are beneficial when features contain categorical data since they do not require dummy coding.

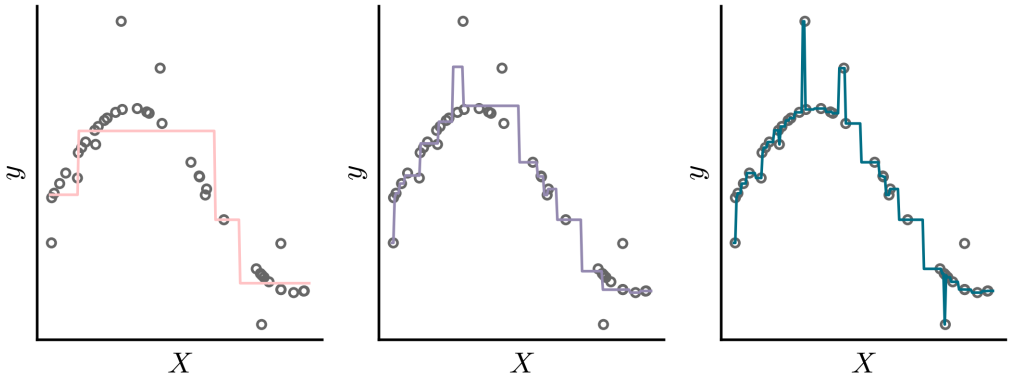



Figure 2.9: Illustration of decision tree on a one-dimensional dataset with 40 data points with a maximum depth of 2 (light pink), 5 (purple) and no maximum depth (dark teal). Scatter points, observed data; solid line, predicted relationship between the features and the target.

Despite many advantages, decision trees suffer from high variance. Minor changes in the training data cause extensive changes to the decision tree via different splits, changing the predictions. This leads to overfitting of training data and results in poor predictions on external test sets. The functions relating the features to the targets generated by decision trees are not smooth (Figure 2.9), which can hinder the performance.

### Cost-Complexity Pruning

Large decision trees are often too complex, resulting in overfitting of the training data. Cost-complexity pruning, also known as weakest link pruning, reduces the size of a decision tree by sequentially removing internal nodes. Not all subtrees are explored. The internal nodes that give the smallest increase in the loss function are removed. For a large decision tree $T_0$, a subtree $T \subset T_0$ is obtained by minimising the cost-complexity function defined as

$$C_\alpha(T) = \sum_{m=1}^{|T|} n_m H(Q_m) + \alpha|T|$$

where $n_m$ is the number of examples in node $m$, $H(Q_m)$ is the loss function, and $\alpha \geq 0$ is a positive tuning parameter.

The value for $\alpha$ determines a trade-off between the decision tree's complexity (size) and the fit of the training data. When alpha is zero ($\alpha = 0$), the tree is not pruned, resulting in the large initial decision tree $T_0$. As $\alpha$ increases, more internal nodes are removed. This decreases the number of terminal leaf nodes and the tree size, resulting in lower model variance and improved interpretability

at the expense of a slight increase in bias.

### 2.4.4 Ensemble Methods

Ensemble methods combine multiple models trained on the same dataset. Ensemble methods based on decision trees include bagging, random forests, and gradient boosting. The bagging and random forest models use the bootstrap method to obtain subsets of the training data. In the bootstrap method, samples are randomly selected from the training data, with sample replacement. This is repeated $B$ times to generate $B$ subsets with a sample size equal to that of the original training set. Sampling is performed with replacement meaning the same training point can be in multiple bootstrapped subsets.

#### Bagging

In bagging, separate decision trees are built on different subsets of the training data, and the results are averaged. Splitting the training data into $B$ different training subsets is called bootstrapping. A separate decision tree is trained on each of the $B$ training subsets. For a new prediction point $x_0$, the predicted target value is calculated as the average prediction over the $B$ decision trees (Equation 2.14).

$$\hat{f}_{bag}(x_0) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x_0) \tag{2.14}$$

If a feature in the training data has a strong correlation with the targets, the feature will be selected as a splitting criterion in many of the $B$ decision trees. This causes the trees to be highly correlated and with high variance. The resultant flexible model is unstable and susceptible to minor changes in the training data, which causes overfitting.

#### Random Forests

Random forests are an alternative to bagging, which reduce the overfitting of training data. Random forests consist of multiple de-correlated decision trees built on bootstrapped training data. A random forest tree is built for each subset of the bootstrapped training data. A random-forest tree differs from a decision tree by how the features are split. A random subset of features $m$ is selected from the total $p$ features. The data is then split along the dimension of the best-split point in the $m$ features. The predicted target value of a new point $x_0$ is calculated as the average prediction over the $B$ random-forest trees (Equation 2.15).

$$\hat{f}_{rf}(x_0) = \frac{1}{B}\sum_{b=1}^{B} T_b(x_0) \tag{2.15}$$

Selecting only a subset of features for training the random-forest trees results in a decorrelation of the $B$ random-forest trees. If a single feature has a strong correlation with the targets, the random selection of features reduces the number of times it is selected as a splitting criterion because it may not be present in the $m$ features. Decorrelating the decision trees reduces the variance with a modest increase in bias and loss of interpretability. As a result, the model is more reliable, less flexible, and less susceptible to overfitting. Reducing $m$ reduces the correlation between the trees in the ensemble meaning small values of $m$ may be beneficial for datasets containing a large number of features.

### Gradient Boosting

Gradient boosting differs from bagging and random forest as decision trees are grown sequentially rather than bootstrapping the dataset and building individual trees. Each tree is built on the information generated by the previous tree. The decision trees in gradient boosting are built on residuals, the difference between the observed and predicted target values, rather than on the target values themselves.

The gradient boosting algorithm has two requirements, a labelled training set $\{(x_i, y_i)\}_{i=1}^{n}$ and a differentiable loss function $L(y_i, f(x_i))$. The most common loss function used for gradient boosting in a regression setting is the squared error

$$\begin{aligned} L(y_i, f(x)) &= \frac{1}{2}\left(\text{observed} - \text{predicted}\right)^2 \\ &= \frac{1}{2}\left(y_i - f(x_i)\right)^2. \end{aligned}$$

The derivative of this loss function is given by

$$\begin{aligned} \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} &= -\frac{2}{2}\left(y_i - f(x_i)\right) \\ &= -\left(y_i - f(x_i)\right). \end{aligned}$$

The boosting model is initialised with a constant value $f(x_0) = \text{argmin}_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$ where $L(y_i, \gamma)$ is the loss function, $y_i$ is the $i^{\text{th}}$ observed value and $\gamma$ is the predicted value. The value for $\gamma$ that minimises this loss function is determined by

equating the derivative of the loss function to zero.

$$\begin{aligned}\sum_{i=1}^{n} \frac{\partial L(y_i, \gamma)}{\partial \gamma} &= \sum_{i=1}^{n} \frac{\partial \left(\frac{1}{2}\left(y_i - \gamma\right)^2\right)}{\partial \gamma} \\ &= -\sum_{i=1}^{n} \left(y_i - \gamma\right) \\ &= 0\end{aligned}$$

This equation is solved to find $\gamma$ which equates to the mean target value $\frac{1}{n}\sum_{i=1}^{n} y_i$. The initial constant model $f_0(x_i)$ is a single leaf node tree equal to the mean target value.

Once initialised, the algorithm builds consecutive decision trees up to a predefined maximum number of trees $M$. For each tree $m$, pseudo residuals are calculated for the training points $i = 1, \ldots, n$ by the negative gradient.

$$\begin{aligned}r_{i,m} &= -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \\ &= -\frac{\partial \left(\frac{1}{2}\left(y_i - f_{m-1}(x_i)\right)^2\right)}{\partial f_{m-1}(x_i)} \\ &= \left(y_i - f_{m-1}(x_i)\right)\end{aligned}$$

For the first decision tree $m = 1$, the initial function $f_0(x_i)$ is used to calculate the predicted value $f_{m-1}(x_i)$. A regression tree is fitted to predict the pseudo residuals $r_{i,m}$ (not target values). The leaf nodes of the regression tree define terminal regions $R_{j,m}$ for $j = 1, \ldots, J_m$. For each region $R_{j,m}$, the data in the region is represented by the constant value $\gamma_{j,m}$, calculated as the minimisation of

$$\begin{aligned}\gamma_{j,m} &= \text{argmin}_{\gamma} \sum_{x_i \in R_{i,j}} L(y_i, f_{m-1}(x_i) + \gamma) \\ &= \text{argmin}_{\gamma} \sum_{x_i \in R_{i,j}} \frac{1}{2}\left(y_i - \left(f_{m-1}(x_i) + \gamma\right)\right)^2.\end{aligned}$$

This is equal to the average residual values of the training samples contained in the region $R_{j,m}$. The predicted target value $f_m(x)$ of each training sample is then calculated using

$$f_m(x) = f_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{j,m} I(x \in R_{j,m})$$

where $f_{m-1}(x)$ is the predicted target value calculated using the previous model,

$\nu$ is a learning rate between zero and one, and $\sum_{j=1}^{J_m} \gamma_{j,m} I(x \in R_{j,m})$ is the sum of the predicted residuals. The learning rate determines the effect each tree has on the final prediction of the target value to prevent overfitting of the training set. This decreases the variance with a slight increase in the bias. The predicted target value of a new point $x_0$ is calculated using the function from the final $M^{\text{th}}$ decision tree (Equation 2.16).

$$f_M(x) = f_{M-1}(x) + \nu \sum_{j=1}^{J_M} \gamma_{j,M} I(x \in R_{j,M}) \tag{2.16}$$

**Neural Networks**

Neural networks form the basis of deep learning, a subcategory of machine learning. The anatomy of the human brain inspires deep learning algorithms. Neurons are highly interconnected human brain cells that transmit electrical signals to each other to process information. Artificial neural networks consist of interconnected layers of artificial neurons, known as nodes, that process and analyse data to solve mathematical problems.

The simple architecture of an artificial neural network has an input layer, one or more hidden layers, and an output layer. The input layer receives the data and processes, analyses, or categorises it before passing it on to the next layer. The hidden layer takes the data from the previous layer (input or hidden) and further processes it before passing it on to the next layer (hidden or output). The output layer then predicts the final forecast. This final layer may consist of one or more nodes. Binary classification, for example, has a single node with a value of zero or one indicating no or yes, whereas multi-class classification has several output nodes, one for each class. While a simple architecture may have only one hidden layer, a deep neural network architecture may have numerous hidden layers with millions of interconnected neurons.

The nodes in each layer are connected to the following layer. These connections are assigned an associated weight value to demonstrate the importance of the variables. A higher weight indicates a larger contribution to the final prediction and vice versa. The inputs (activations) to each node are multiplied by their connection weight to compute a weighted sum,

$$a_0 w_0 + a_1 w_1 + \ldots + a_n w_n$$

where $a$ is the activation of a single node in the current layer, $\{a_0, a_1, \ldots, a_n\}$ are the activations of the previous layer, and $\{w_0, w_1, \ldots, w_n\}$ are the weights

connecting the activations of the previous layer with the node in the current layer. A bias value is added to the weighted sum.

$$a_0 w_0 + a_1 w_1 + \ldots + a_n w_n + b$$

The bias value, $b$, accounts for inactivity, in other words, how high the weighted sum needs to be before the neuron is meaningfully active. An activation function is applied to the resulting value, which acts as a threshold to determine whether or not the neuron is activated.

$$\sigma(a_0 w_0 + a_1 w_1 + \ldots + a_n w_n + b)$$

Activation functions may return a range between 0 and 1, for example, the sigmoid function.

$$S(a) = \frac{1}{1 + e^{-a}}$$

The activation function most commonly employed in deep learning is the rectified linear unit (ReLU) function. It returns 0 if it receives any negative input and itself for any positive value.

$$\text{ReLU}(a) = \max(0, a)$$

The activation of a neuron measures how positive the relevant weighted sum is. A node's activation can be expressed in matrix form. See Equation 2.17, where $l$ represents the current layer, $l-1$ represents the previous layer, $w$ represents the weights, $a$ represents the activation values, and $b$ represents the biases. Each neuron is a function that takes the preceding layer's outputs and returns a positive number.

$$\begin{aligned} a_0^{(l)} &= \sigma(w_{0,0} a_0^{(l-1)} + w_{0,1} a_1^{(l-1)} + \ldots + w_{0,n} a_n^{(l-1)} + b_0) \\ a^{(l)} &= \sigma(W a^{(l-1)} + b) \end{aligned} \tag{2.17}$$

Training a neural network entails iteratively performing the forward and backward propagation cycle. Initially, the weights and biases are randomly assigned. The data flows through the network from the input layer through the hidden layers to the output layer. The activation of nodes governs the flow of data. If a node is activated, its output becomes the input to the next layer in the network. If a node is not activated, it inhibits the data passing to the next layer. This process is termed forward propagation. The final prediction is the neuron in the

output layer with the highest value. The values assigned to the output layer are analogous to probability scores.

The observed output is known during training and is compared to the predicted output to determine an error value. The magnitude of the error is transferred back through the neural network. This process is termed backward propagation. The weights are adjusted based on the error information. The forward and backward propagation cycle is performed iteratively for multiple input-output pairs.

Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), sequence-to-sequence, and transformer are examples of neural network architectures.

## 2.5 Molecular Descriptors

There are many ways of representing a molecule. Names, codes, topology, and properties are all examples of molecular representations. For example, the name of the painkiller paracetamol is a shortened form of its chemical name ***par***a-**acet**yl-**am**ino-phen**ol**. The trade names, Panadol and Tylenol, are examples of alternate common names for paracetamol, as is the abbreviation APAP from the alternative chemical name [*N*-]**a**cetyl-***p****ara***-**a**mino**p**henol, and the International Union of Pure and Applied Chemistry (IUPAC) nomenclature *N*-(4-hydroxyphenyl)acetamide. Figure 2.10 depicts the structure of paracetamol, while Table 2.2 shows common representations.

Machine learning for chemistry requires encoding chemical structures in a machine-readable format that an algorithm can process. Todeschini and Consonni define a molecular descriptor formally as *"the final result of a logical and mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardised experiment"*.[45] Descriptors are determined experimentally or theoretically derived from a symbolic representation of a molecule. They encode a molecule's structural, physicochemical, electronic, or topological nature. The choice of descriptor is critical as it can affect an algorithm's performance. It is consequently essential

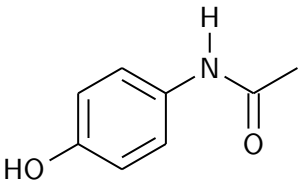



Figure 2.10: The chemical structure of paracetamol.

Table 2.2: A list of commonly accepted representations for the chemical paracetamol.

| Representation Name | Representation of Paracetamol |
|---|---|
| Name | Paracetamol |
| Trade Names | Panadol, Tylenol |
| Other Names | *para*-aceyl-amino-phenol, [*N*-]acetyl-*para*-aminophenol, APAP |
| IUPAC Name | *N*-(4-hydroxyphenyl)acetamide |
| CAS Registry Number | 103-90-2 |
| ChEMBL ID | CHEMBL112 |
| Canonical SMILES | CC(=O)NC1=CC=C(C=C1)O |
| InChI | 1S/C8H9NO2/c1-6(10)9-7-2-4-8(11)5-3-7/h2-5,11H,1H3,(H,9,10) |
| InChIKey | RZVAJINKPMORJF-UHFFFAOYSA-N |

to evaluate several representations. A descriptor must be able to discriminate between molecules, have values that change gradually with modest structural changes, and be interpretable. It must also capture the necessary information for the specified problem, i.e. correlate well with the target variable. It is beneficial for descriptors to obey physical invariants,[46] meaning the descriptor of a molecule should be independent of distinct characteristics of the molecule's representation. These characteristics may include numbering, labelling, reference frame, translation, rotation, or molecular conformations.

The perspective of a molecule can characterise molecular descriptors: global, local, or field. Descriptors based on the global perspective of a molecule consider the whole structure, for example, volume, surface area, dipole moment, and molecular graph. Local descriptors only contain information about atoms, bonds, or fragments within a molecule, including atomic charges, bond polarizabilities, and molecular fingerprints. The final perspective of a molecule is the molecular fields surrounding the molecule, such as the electrostatic potential.

The dimensionality of the structural representation can also classify molecular descriptors. Zero-Dimensional (0D) descriptors include constitutional descriptors and atom counts. One-Dimensional (1D) descriptors consider a list of structural fragments. Two-Dimensional (2D) descriptors take into account the connectivity in a molecule. Three-Dimensional (3D) descriptors encode the geometry of a molecule. Each class of descriptor is described in further detail below.

### 2.5.1 Zero-Dimensional and One-Dimensional Descriptors

0D and 1D descriptors do not contain information about a molecule's chemical structure, geometry, or atom connectivity. These representations consist of occurrence frequencies, bulk properties, and molecular properties. 0D descriptors are constitutional descriptors derived from the molecular formula of a molecule. Examples include molecular weight, atom counts, and bond counts. 1D descriptors are obtained from structural fragments, including the number of hydrogen bond donors, hydrogen bond acceptors and rings, and functional group counts. Table 2.3 shows 0D and 1D descriptors of paracetamol. Low-dimensional descriptors are easily obtained and quick to calculate. They are frequently combined with other chemical descriptors of the same or higher dimensions since they provide minimal information about the molecule.

Table 2.3: Zero- and one-dimensional molecular descriptors of paracetamol[3]

| Descriptor Name | Value |
|---|---|
| Empirical Formula | $C_8H_9NO_2$ |
| Molecular Weight | 151.16 g $mol^{-1}$ |
| Hydrogen Bond Donor Count | 2 |
| Hydrogen Bond Acceptor Count | 2 |
| Rotational Bond Count | 1 |
| Heavy Atom Count | 11 |
| Formal Charge | 0 |

### 2.5.2 Two-Dimensional Descriptors

2D descriptors encode a molecule's topology (i.e. connectivity) based on structural fragments, atom connectivity, and topological indices. A chemical structure is the spatial arrangement of atoms and bonds derived from the mathematical field of graph theory.[47] It is also known as a molecular graph. The molecular graph of paracetamol is depicted in Figure 2.11.

A molecular graph $G = (V, E)$ consists of nodes $V$ corresponding to the atoms and edges $E$ corresponding to the bonds.[48] The vertices $v_i$ and $v_j$ ($v_i \in V$ and $v_j \in V$) of an edge $\{v_i, v_j\} \in E$ are considered connected and are known as the endpoints. The edge that connects two vertices is incident with them, the vertex is incident with the connected edge, and the two vertices connected by the edge are said to be adjacent.

Molecular graphs are undirected since the edges do not have a direction, coloured since the nodes are assigned a discrete label, and weighted since the edges are assigned a number (Figure 2.11). The nodes are coloured according to the atom

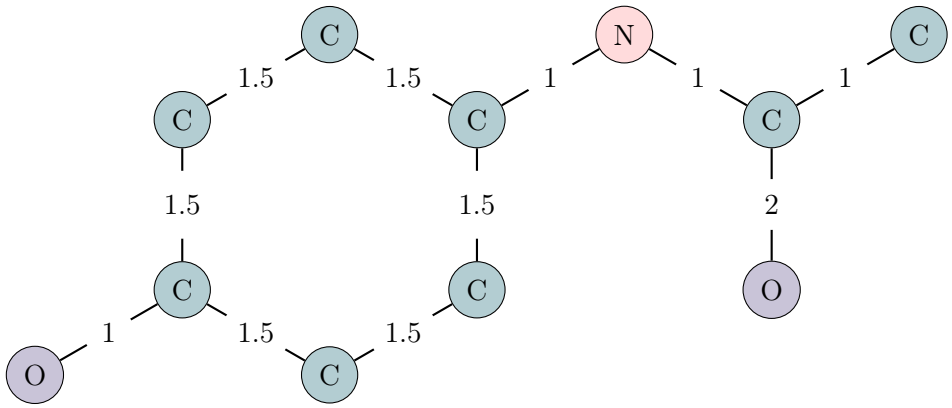



Figure 2.11: The molecular graph of paracetamol.

type, such as carbon (C), nitrogen (N), and oxygen (O). The edges can be weighted based on bond order, such as single (1), double (2), triple (3), and aromatic (1.5). Hydrogen atoms are commonly omitted from molecular graphs when the edges are weighted since they are represented implicitly through valences. Molecular graphs are fast to calculate. They are also easy to interpret as molecules with identical molecular graphs correspond to the language of organic chemists.

Topological descriptors are derived from the molecular graph representation. The molecular graph representation can be directly used as a descriptor or converted to numerical values or vectors such as molecular fingerprints and graph-based kernels. An alternative 2D representation is line notations that describe a molecule as a sequence of characters.

### Molecular Fingerprints

Molecular fingerprints are topological descriptors derived from the molecular graph representation. Fingerprints are vectors composed of binary digits denoting the presence or absence of structural features, which are quick and easy to calculate. There are two main classes of molecular fingerprints, structure-key and hash-key fingerprints, also recognised as knowledge-based and information-based descriptors.

**Structure-Key Fingerprints.** Structure-key fingerprints encode a molecule as a fixed-length bitstring based on a list of predefined substructures or fragments (subgraphs). Each bit corresponds to the presence or absence of a subgraph in the predefined list. If the subgraph is present, the bit is set to one; otherwise, it is set to zero (Figure 2.12). Structure keys cannot encode information about the entire molecular structure. As structure keys are based on an explicit dictionary, they cannot encode features outside the predefined list. They have limited applicability

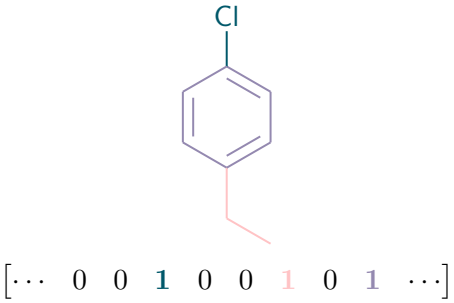



Figure 2.12: Illustration of a structure-key fingerprint. Subgraphs and corresponding bits in the fingerprint are colour-coded.

as a result. The Molecular ACCess Systems (MACCS)[49] keys are an example of structure-key fingerprints frequently used in drug discovery. MACCS keys consist of 166 bits corresponding to 166 public structural keys based on SMILES Arbitrary Target Specification (SMARTS) strings.

**Hash-Key Fingerprints** Hash-key fingerprints define the connectivity of molecules without using a predefined list of subgraphs. They intend to encode the structural information of the entire molecule, akin to the mathematical graph-vector transform discussed later. Substructure enumeration techniques encode linear or circular subgraphs. Figure 2.13 and Algorithm 2.1 outline the general hashing algorithm for generating a hash-key fingerprint. Initially, subgraphs within a molecule are identified by following linear or circular paths along bonds up to a specified size. These subgraphs are then converted to numeric values using a hash function which indicates the bit positions in the fingerprint. The optimum bit length of hash-key fingerprints can vary depending on the type of fingerprint or the specific problem and hence should be tuned as a hyperparameter. As the hash function converts the molecule into a number within a fixed range, bit collisions are likely. A bit collision is when two different subgraphs are converted to the same number and, therefore, the same bit. Hashed fingerprints are irreversible, meaning determining the molecule from the fingerprint is impossible. Examples of hash-key fingerprints include the RDKit path-based fingerprints[50] and Morgan circular fingerprints.[51]

**RDKit Fingerprints** The RDKit fingerprints are based on the Daylight fingerprint.[52] Subgraphs within a molecule are generated by following topological paths starting from each atom up to a predefined path length (number of bonds). Figure 2.15 illustrates this process for the molecule 4-ethylbenzyl chloride, with the maximum path length set to three bonds (Figure 2.14). A hash is generated for each subgraph using the bond order and atom types of the individual bonds. The hash is then used as the seed to a random generator which generates $n$ ran-

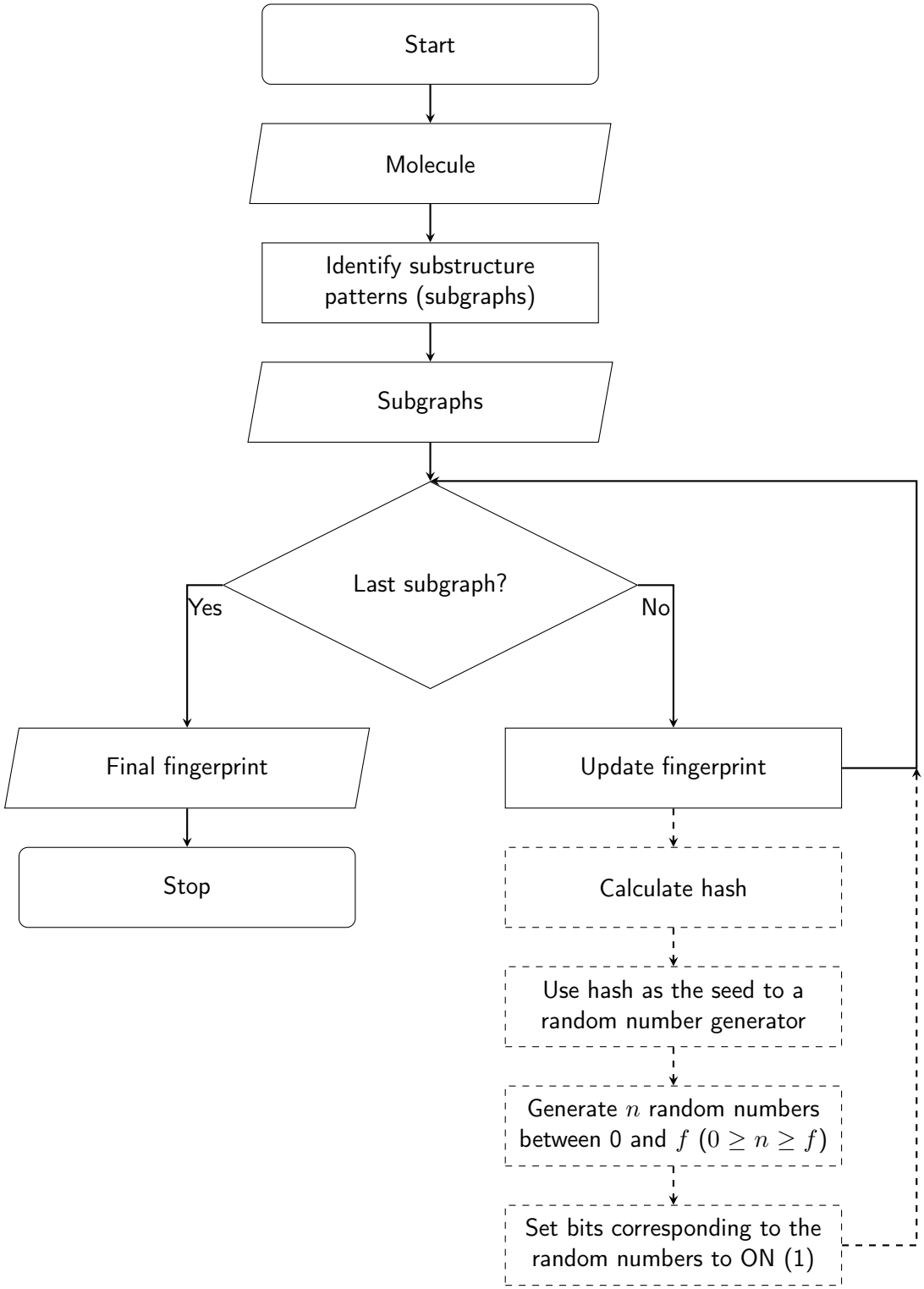



Figure 2.13: Flow diagram of the general hashing algorithm for generating a hash-key fingerprint.

**Algorithm 2.1** General hashing algorithm for generating a hash-key fingerprint

```
function GENERATEFINGERPRINT(molecule, bitLength, subgraphSize)
    fingerprint = INITIALISEFINGERPRINT(bitLength)
    subgraphs = GETSUBGRAPHS(subgraphSize)
    for subgraph in subgraphs do
        seed = HASH(subgraph)
        indicies = RANDOM(seed, start = 0, end = bitLength)
        for index is indicies do
            fingerprint[index] = 1
    return fingerprint
```

dom numbers between zero and the length of the fingerprint. The indices of the bits that correspond to the random numbers are set to ON (1).

**Morgan Fingerprints** Circular fingerprints consider the circular environment surrounding each atom rather than linear paths.[53] The Morgan fingerprint is similar to the Extended Connectivity Fingerprint (ECFP).[53] The Morgan fingerprint defines the circular radius, whereas ECFP defines the diameter. The algorithm initially assigns an identifier to each non-hydrogen atom in the molecule. The identifier is based on the Daylight atomic invariants:[54] the number of adjacent non-hydrogen atoms; valence minus the number of hydrogens; the atomic number; the atomic mass; the atomic charge; the number of attached hydrogens; and additionally, whether the atom is in at least one ring. Feature Morgan fingerprints are a variant similar to the Functional-Class Fingerprints (FCFP), which differ in the assignment of the atom identifier. Each atom is assigned a code based on its role: hydrogen-bond acceptor; hydrogen-bond donor; aromatic; halogen; basic; and acidic. The next step of the algorithm iteratively updates each identifier to include the identifier and bond order of neighbouring atoms. The iteration process repeats up to a predefined radius. The value for the radius commonly implemented ranges between one and three. Figure 2.16 illustrates this iteration process on the 4-ethylbenzyl chloride for atom 1. The iteration process is performed for all atoms, meaning the final identifiers contain partial implicit information about other areas of the molecule (Figure 2.17). Lastly, the identifiers are folded into the fixed length of the bit vector using a hash function, where bit collisions may occur.

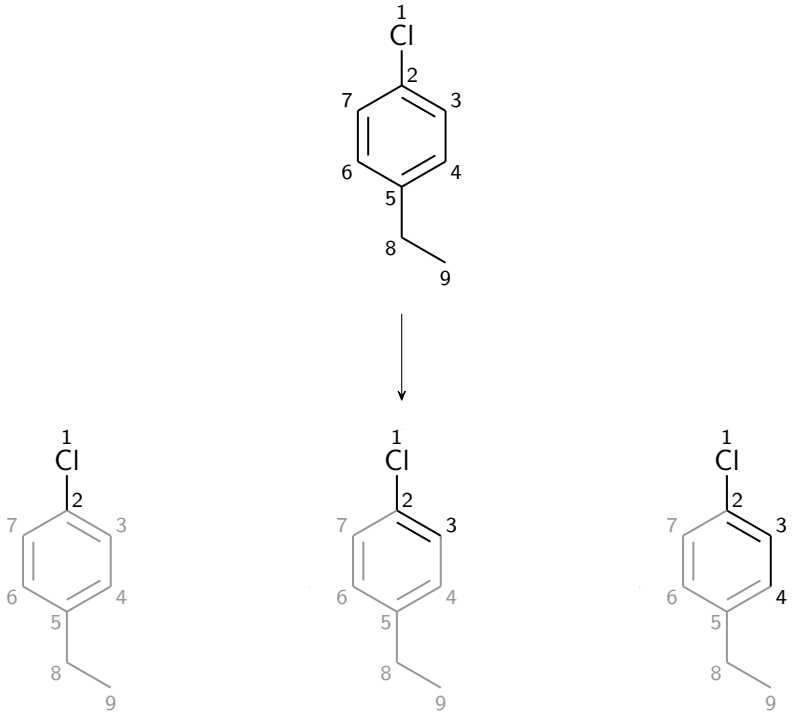



Figure 2.14: Illustration of the path identification process in RDKit fingerprints for atom **1** in the orginal molecule, 4-ethylbenzyl chloride. Paths are extended up to a maximum path length of three bonds.

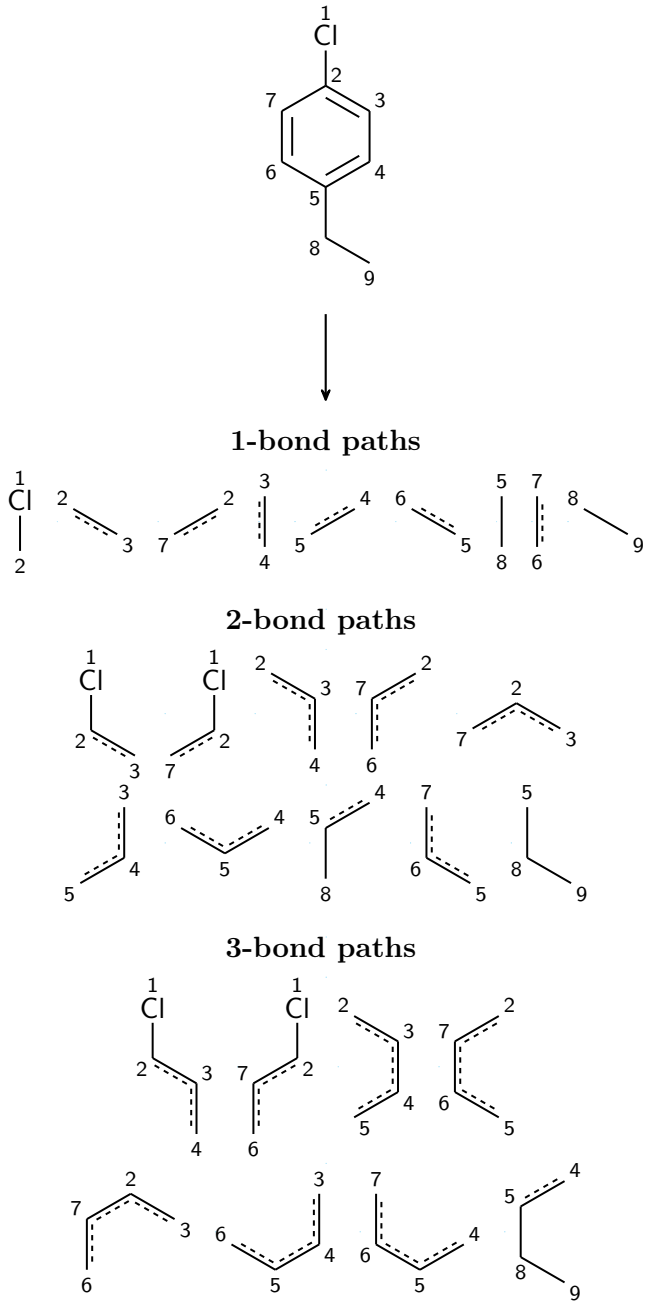



Figure 2.15: Illustration of subgraph identification using the RDKit topological fingerprint. For each atom in the orginal molecule, 4-ethylbenzyl chloride, paths are extended up to a maximum path length of three bonds.

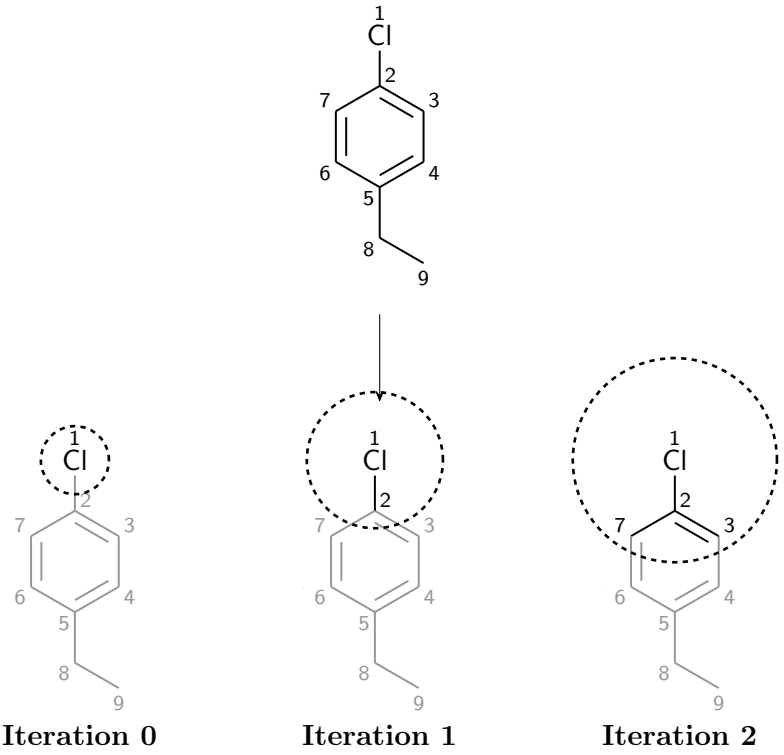



Figure 2.16: Illustration of the iterative updating of an atom identifier. The example uses atom **1** in 4-ethylbenzyl chloride. At iteration 0, the initial identifier for atom **1** only contains information about the chlorine atom and its bonds. After the first iteration, the identifier includes information about the immediate neighbours of atom **1**. By the second iteration, information regarding the meta-carbon atoms on the phenyl ring is incorporated into the atom **1** identifier. The iteration process will terminate after a predetermined number of iterations. Performing more iterations increases the circular environment surrounding the atoms in the identifier.

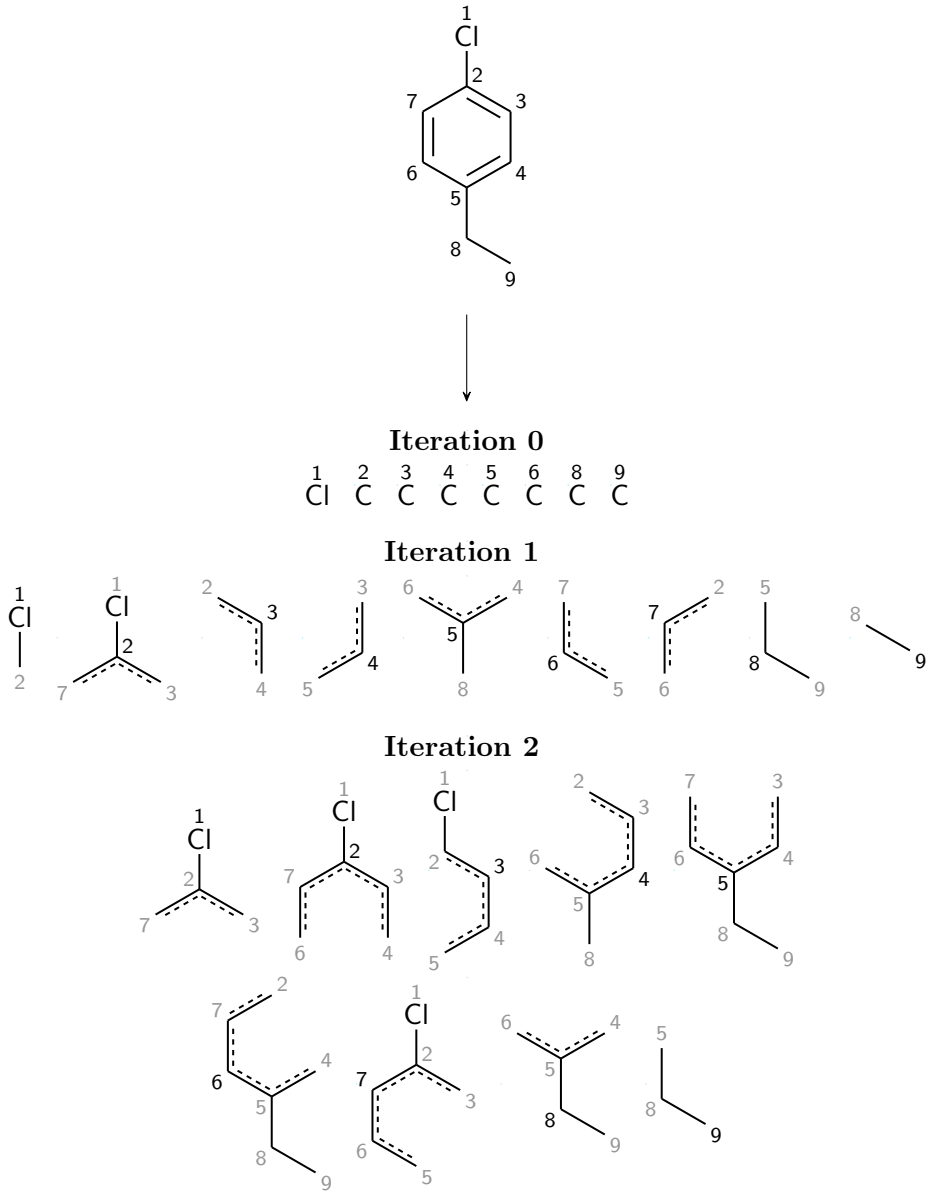



Figure 2.17: Illustration of the identification of subgraphs in the Morgan circular fingerprint with a radius up to two bonds.

### Weisfeiler-Lehman Graph Kernel

The Weisfeiler-Lehman (WL) subtree graph kernel[55] is based on the concepts used in the Weisfeiler-Lehman test of isomorphism. If two graphs are structurally identical, i.e., the mapping between the nodes is equivalent, they are said to be isomorphic. An example of two isomorphic graphs is the cis and trans isomers of an alkene molecule. The nodes are connected by the same edges regardless of the isomerism (Figure 2.18). When a graph matches a subgraph in another graph, this is known as subgraph isomorphism.

Figure 2.19 outlines the algorithm for the WL isomorphism test between two molecular graphs, $G$ and $G'$. The molecular graphs are initialised by mapping the atoms present in both graphs to numbers. When using the subtree kernel as the base kernel, the node labels are iteratively updated to include information about the circular neighbourhood of the atoms, similar to the Morgan circular fingerprint. The number of iterations $h$ is defined prior to the calculation and requires tuning as a hyperparameter. The algorithm iteratively updates the node labels in four main steps: multiset-label determination, sorting each multiset, label compression, and relabelling.

**Multiset-Label Determination.** A multiset-label $M_i(v)$ is assigned to each node $v$ of the molecular graphs. This is determined by identifying the direct (one-bond distance) neighbours $u$ in neighbourhood $\mathcal{N}(\sqsubseteq)$ to the node $v$.

$$M_i(v) = \{l_{i-1}(u) | u \in \mathcal{N}(v)\}$$

**Sorting Each Multiset.** The elements in the multiset $M_i(v)$ are sorted in ascending order to ensure all identical labels are compressed to the same number. The sorted multiset-label $M_i(v)$ is converted into a string. The current node label $l_{i-1}(v)$ is added as a prefix to the sorted multiset-label string with the format

$$s_i(v) = (l_i(v), M_i(v))$$

where $i$ is the iteration number and $v$ is the node.

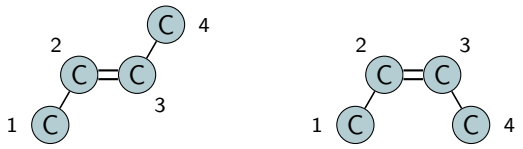



Figure 2.18: An example of two isomorphic graphs, *trans*-but-2-ene and *cis*-but-2-ene.

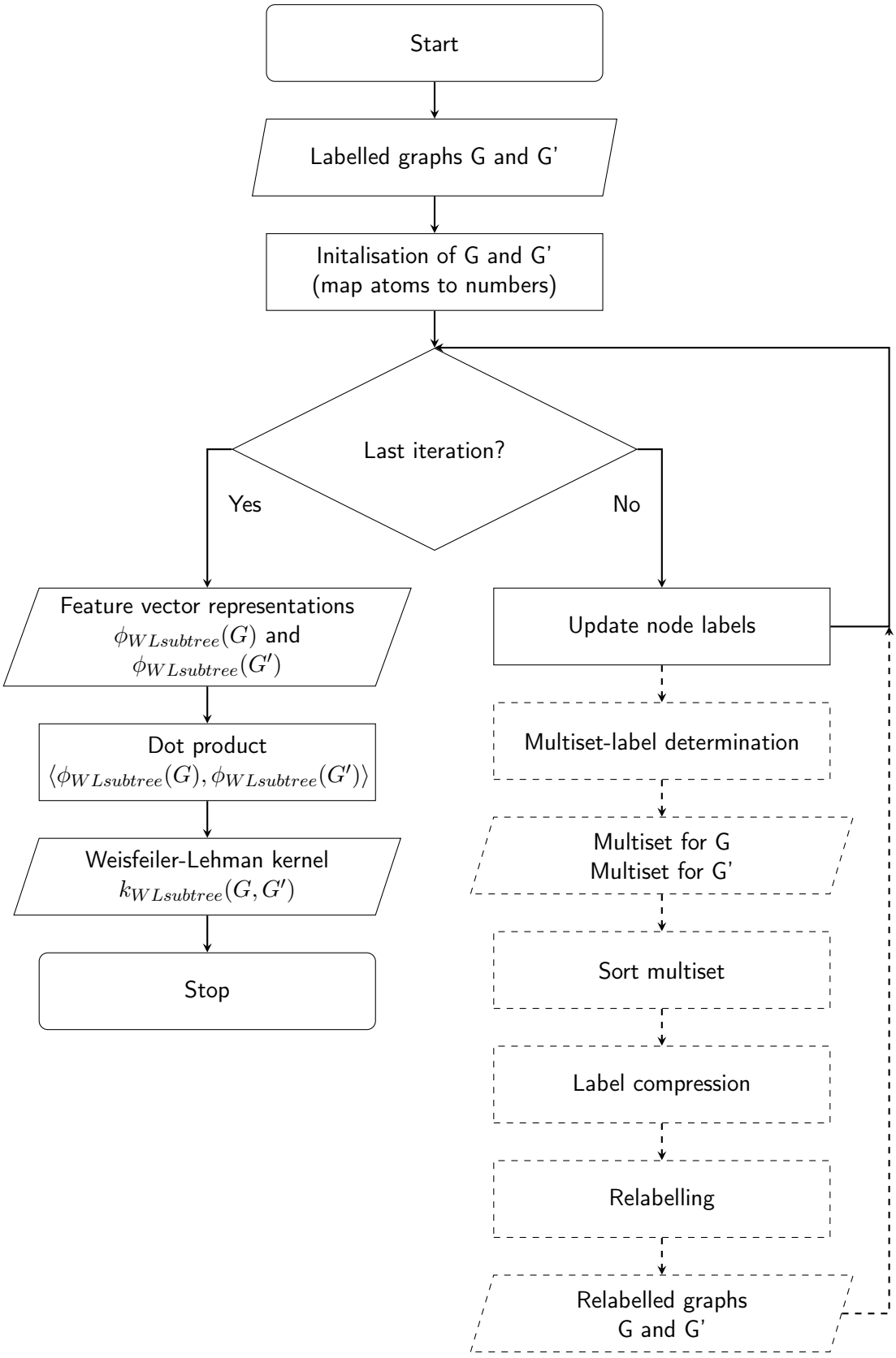



Figure 2.19: Schematic of the Weisfeiler-Lehman algorithm.

**Label Compression.** The multiset-label strings $s_i(v)$ for all nodes $v$ in both molecules $G$ and $G'$ are sorted in ascending order. A function $f : \Sigma^* \to \Sigma$ is used to map these strings $s_i(v)$ to new compressed labels $C_i(v)$.

**Relabelling.** The nodes $l_i(v)$ in the two molecules $G$ and $G'$ are relabelled using these new compressed labels:

$$l_i(v) = f(s_i(v)) = C_i(v).$$

Any two nodes with the same multiset-label strings $s_i(v) = s_i(w)$ will have the same new compressed label $f(s_i(v)) = f(s_i(v)) = C_i(v) = C_i(w)$.

Once the $h$ iterations have been completed, the molecular graphs $G$ and $G'$ are converted into feature vector representations $\phi(G)$ and $\phi(G')$. The vector $\phi$ is a concatenation of the counts $c_i$ of the original node labels $l_0(v)$ and the counts $c_i$ of the compressed node labels $l_i(v) = C_i(v)$ in the two graphs. Mathematically, the feature vectors are defined as

$$\phi(G) = \big(c_0(G, \sigma_{01}), \ldots, c_0(G, \sigma_{0|\sum_0|}), \ldots, c_h(G, \sigma_{h1}), \ldots, c_h(G, \sigma_{h|\sum_h|})\big)$$
$$\phi(G') = \big(c_0(G', \sigma_{01}), \ldots, c_0(G', \sigma_{0|\sum_0|}), \ldots, c_h(G', \sigma_{h1}), \ldots, c_h(G', \sigma_{h|\sum_h|})\big)$$

where $\sum_i \subseteq \sum$ is the set of numbers that occur at least once as node labels in $G$ or $G'$ by the end of the $i^{th}$ iteration, $\sum_0$ is the set of original node labels of $G$ or $G'$, assuming all $\sum_i$ are pairwise disjoint and every $\sum_i = \sigma_{i1}, \ldots, \sigma_{i|\sum_i|}$ is ordered and a map $c_i : \{G, G'\} \times \sum_i \to \mathbb{N}$ such that $c_i(G, \sigma_{ij})$ is the number of occurrences of the number $\sigma_{ij}$ in the graph $G$.

The dot product of the two vectors is calculated to give the WL subtree kernel between the two molecules as defined in Equation 2.18.

$$k_{WLsubtree}(G, G') = \langle \phi_{WLsubtree}(G), \phi_{WLsubtree}(G') \rangle \tag{2.18}$$

Figure 2.20 illustrates a single iteration of the WL kernel between two molecules, 4-ethylbenzyl chloride and 4-methoxybenzyl bromide.

The WL kernel is calculated for a pair of molecules. This kernel notation can be used directly as an input to an SVR model or a neural network. While it is not the norm to use the kernel notation as an input to other machine learning algorithms, a mathematical transformation can extract the features from the kernel notation.

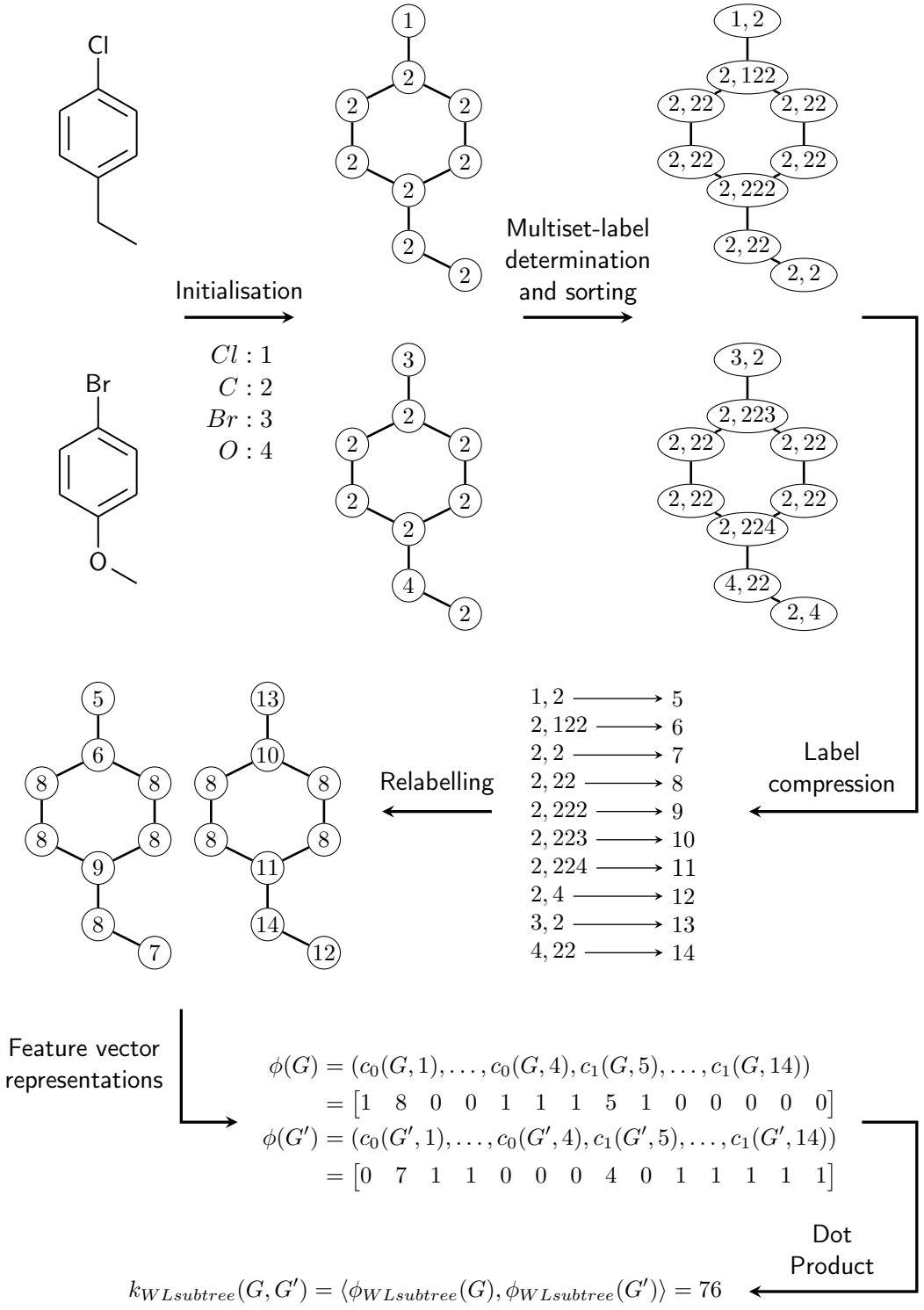



Figure 2.20: Illustration of the calculation of the Weisfeiler-Lehman kernel between two molecules (4-ethylbenzyl chloride and 4-methoxybenzyl bromide) represented by molecular graphs. The following (atom: number) mapping is used: $Cl : 1$, $C : 2$, $Br : 3$ and $O : 4$. The multiset label is determined. For example, the carbon atom (2) in the phenyl ring in 4-methoxybenzyl bromide connected to a chlorine atom (1) and two carbon atoms (2) has a multiset-label of 122. After sorting the multiset and adding the prefix, the final multiset-label string for the aforementioned carbon atom is $(2, 122)$. After label compression and relabelling, the final label of the carbon for the first iteration is **6**.

### Line Notation

Line notations are a condensed representation of the chemical structure. They contain connectivity information analogous to topological descriptors but overlook information such as protonation states and geometry. Examples include Simplified Molecular-Input Line-Entry System (SMILES) and IUPAC International Chemical Identifier (InChI).

The most widely used line notation is SMILES and its extensions, SMILES Arbitrary Target Specification (SMARTS) and A Reaction Transform Language (SMIRKS). Atomic symbols in square brackets represent the atoms. The organic subset of elements, which includes B, C, N, O, P, S, F, Cl, Br, and I, do not require brackets. With a few exceptions, hydrogen atoms are implicit. The SMILES string of ethanol ($CH_3CH_2OH$), for example, is CCO. A charged atom requires a square bracket, explicit hydrogen atoms, the number of charges if greater than one, and the sign of the charge. A plus symbol represents a positive charge, and a subtraction symbol a negative. For example, the SMILES string of the hydroxide anion ($HO^-$) is [OH−], whereas, for the hydronium cation ($H_3O^+$), it is [OH3+].

Symbols represent the bonds depending on the type: a single bond is '-', double is '=', triple is '#', quadruple is '$', and aromatic is ':'. Examples are illustrated in the SMILES string of carbon dioxide ($CO_2$) O=C=O and carbon monoxide (CO) [C−]#[O+]. Many bonds are usually implicit. These include single bonds, bonds between aliphatic atoms assumed to be single, and aromatic bonds. The '.' character represents a non-bond whereby two parts interact non-covalently, such as sodium chloride is [Na+].[Cl−].

Rings are broken at an arbitrary point, and ring closure labels are written on each end to show connectivity between the non-adjacent atoms. For example, the SMILES string of the aliphatic ring cyclohexane is C1CCCCC1. For a second ring closure, the label would be **2**, as in the bicyclic compound Decalin C1CCC2CCCCC2C1. The SMILES string of an aromatic ring has three forms. The Kekulé form has alternating single and double bonds; for benzene, the string would be C1=CC=CC=C1. The aromatic character form; C1:C:C:C:C:C1. The most common form is writing the aromatic atoms in lowercase letters and omitting the bond characters, such as 'b', 'c', 'n', 'o', 'p', and 's', for B, C, N, O, P, and S, respectively. For benzene, the SMILES string would be c1ccccc1. A hydrogen atom bonded to an aromatic nitrogen must be explicitly represented, as in pyrrole [nH]1cccc1.

Branching is denoted with parentheses. For example, the SMILES string of

(a) (b)


Figure 2.21: The L-enantiomer (a) and D-enantiomer (b) of the amino acid alanine.

ethanol can also be written as C(O)C, indicating that the carbon atom is attached to an oxygen atom and another carbon atom.

Although SMILES string may define stereochemistry, it is not essential. Symbols represent different types of stereochemistry: '\' and '/' define *cis* and *trans* isomers, whereas '@' and '@@' define the tetrahedral configuration. The characters '\' and '/' indicate the direction of the single bonds adjacent to the double bond in cis and trans isomers. For example, *cis*-but-2-ene is C/C=C/C, whereas *trans*-but-2-ene is C/C=\C. The '@' character implies a clockwise tetrahedral carbon, whereas '@@' implies anti-clockwise. A stereocentre vital to life is in amino acids (except for glycine), the building blocks of proteins. Figure 2.21 illustrates the two enantiomers of alanine, L and D, where L-enantiomers are essential for proteins. The SMILES string for L-alanine is N[C@@H](C)C(=O)O, and for D-alanine is N[C@H](C)C(=O)O.

SMARTS strings specify subgraphs within a molecule. They are particularly relevant for substructure searching, i.e., identifying a subgraph in a molecular graph. While nearly all SMILES strings are valid SMARTS patterns, this is not true in reverse. SMARTS patterns introduce additional logical operators and special symbols to enable more generalised structures. For example, [C, N] means the atom can be an aliphatic carbon or aliphatic nitrogen. The symbol '∼' matches any bond. The documentation provides a complete list of examples.[56]

The SMIRKS string describes a chemical reaction. The SMILES strings of the reactants, reagents, and products are separated by a '>' symbol. Multiple molecules can populate each field by delineating with a dot (.)

A single molecule has multiple valid SMILES strings. The string depends on the starting atom, the path around the molecule, and branching. Several canonicalization techniques have been devised. It is crucial to ensure that a molecule is represented by a single SMILES string when developing machine learning models.

### 2.5.3 Three-Dimensional Descriptors

Three-Dimensional (3D) descriptors are derived from the 3D conformation of a molecule. They range from geometrical representations of the 3D molecular structure to properties calculated from the 3D structure. While 3D descriptors provide additional structural information they can be time-consuming to calculate.

3D topology descriptors define the connectivity of atoms in 3D space. The 3D representation of molecular graphs is one example. The node or edge labels can encode atomic coordinates, bond angles, and chirality. The 3D molecular graph is suitable for a single static depiction of a molecule but inadequate if the atoms rearrange over time, such as tautomers.

Space-filling models, commonly known as CPK models, can be used to calculate molecular properties. Spheres represent the atoms in these models. The centre of the sphere is at the nucleus of the atom, and the radius of the sphere is proportional to the atom's Van der Waals radius. Properties calculated from the CPK model include area, volume, polar and non-polar surface area can all be calculated from the CPK model.

More advanced calculations, such as quantum chemical, can provide information about molecular orbitals, vibrational modes, and molecular structure. Example quantum chemical calculations include Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO) energies, vibrational frequencies and intensities, and Nuclear Magnetic Resonance (NMR) chemical shifts.

## 2.6 Molecular Similarity

The similar property principle in drug discovery states that two structurally similar ligands tend to have comparable properties and reactivity.[57] These properties may be physicochemical, biological affinities, or ADME. The quantification of molecular similarity in terms of topology is beneficial in exploring similar regions of molecular space likely to exhibit similar characteristics. Similarity coefficients have applications in cheminformatics, including property prediction, molecular generation, and synthesis route design.

Similarity coefficients are calculated on molecular descriptors. In similarity calculations, the chemical structures are generally represented using binary molecular fingerprints. Examples of similarity coefficients include the Tanimoto coefficient, Euclidean distance, and the Dice coefficient. Table 2.4 lists the equations for calculating similarity coefficients between two molecules, A and B, represented

by fingerprints, $F_A$ and $F_B$; where $a$ is the number of ON bits in fingerprint $F_A$, $b$ is the number of ON bits in fingerprint $F_B$ and $c$ is the number of shared ON bits in $F_A$ and $F_B$. The resulting quantitative values are typically normalised on a scale of zero (no similarity) to one (identical).

Table 2.4: Equations for calculating similarity coefficients between two molecules represented by molecular fingerprints

| Similarity Measure | Equation |
|---|---|
| Tanimoto coefficient | $T(F_A, F_B) = \frac{c}{a+b-c}$ |
| Dice coefficient | $D(F_A, F_B) = \frac{2c}{a+b}$ |
| Euclidean distance | $D(F_A, F_B) = \sqrt{a+b-2c}$ |

The Tanimoto coefficient[58,59] is widely accepted in cheminformatics and medicinal chemistry as a molecular similarity measure. It is worth noting that, in practice, different fingerprint descriptors may give vastly different Tanimoto scores. Molecular fingerprints may differ in size and type. While Morgan fingerprints are relatively sparse, topological-path fingerprints are more dense.

## 2.7 Performance Evaluation

The performances of the regression models can be evaluated using the Coefficient of Determination ($R^2$) and Root Mean Squared Error (RMSE) for data points outside of the training set.

### 2.7.1 Coefficient of Determination

The coefficient of determination, denoted as $R^2$, is a measure of how well the model replicates the observed targets. It calculates the proportion of variability within the predicted targes that can be explained by the regression model. The mathematical representation is given in Equation 2.19, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$; $\hat{y}_i$ is the predicted value of the $i$-th sample; $y_i$ is the corresponding observed (experimental) value; and $n$ is the total number of samples. The residual sum of squares, $SS_{res}$ is the discrepancy between the observed and predicted target values, signifying the variability of the target data explained by the model. The total sum of squares, $SS_{tot}$ is proportional to the variance of the target data.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SS_{res}}{SS_{tot}} \tag{2.19}$$

The value for the upper bound of $R^2$ is one, which indicates a better goodness-of-fit. The $R^2$ value can be negative if the mean of the data is a better fit to the observed values than the predicted values, i.e., $SS_{res} > SS_{tot}$.

### 2.7.2 Root Mean Squared Error

The RMSE measures how far the predicted values are from the observed target values. Mean Squared Error (MSE) calculates the mean of the squared residuals, where the residuals are the differences between the observed and predicted target values. The RMSE is the root of the MSE as shown in Equation 2.20.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{SS_{res}}{n}} = \sqrt{MSE} \tag{2.20}$$

As the predicted target values get closer to the observed values, the residual errors are reduced, lowering the RMSE. The value for RMSE ranges from zero to positive infinity and has the same units as the target variable. A value of zero means the predicted values are equal to the observed values, a lower value indicates a better fit and lower errors.

# Chapter 3

# Artificial Intelligence for Chemical Synthesis

## 3.1 Introduction

The synthesis of new molecules is essential for progress in the pharmaceutical industry and academia. Examples include medicinal and process chemistry in the development of pharmaceuticals. The prediction and development of synthetic routes are crucial in exploring reaction space to identify more appealing or novel syntheses. Anticipating how molecules react (forward reaction prediction) and how molecules can be synthesised (retrosynthesis) currently relies on synthetic chemists' scientific intuition, expertise and experience. Continuous improvements in computational resources and mathematical algorithms have accelerated the development of tools to help chemists explore reaction space and increase synthesis success rates. These are known as Computer-Aided Synthesis Planning (CASP) tools. In this chapter, we introduce basic approaches to CASP and review established benchmark chemical datasets employed in this field. We discuss recently developed applications for forward reaction prediction and retrosynthetic analysis. In particular, the quantitative performance of state-of-the-art tools on benchmark datasets is extracted from the literature and compared.

CASP tools are challenging to develop due to the high dimensionality of chemical and reaction search space. Artificial Intelligence (AI) has shaped the progression of CASP tools with a recent focus on data-driven machine-learning approaches. The overall aim of CASP tools is to reduce the timelines of chemical syntheses by aiding the everyday decision-making of synthetic chemists. Areas that would benefit from such tools include drug design, novel route discovery and the design

of biologically active compounds.

Retrosynthesis is the process of synthetic planning that begins with the product and works backwards to determine the starting materials.[60] Retrosynthetic analysis is the logical, problem-solving technique applied when planning synthetic routes to attain a target molecule. E. J. Corey formalised the retrosynthetic procedure.[61] During the 1960s and 1970s, retrosynthetic analysis was developed and demonstrated in practice.[60,62,63] A target molecule is transformed into simpler precursors by the imaginary disconnection of a bond, the reverse of a chemical reaction. The disconnection is chosen based on recognising key substructures or "retrons" present in the target molecule that are the products of known, reliable reactions. The two synthons resulting from the disconnection are generalised ionic or neutral fragments corresponding to idealised reagents. The synthetic equivalent of a synthon is a reagent that acts as the synthon. Reagents and reaction conditions are often not considered during retrosynthetic analysis. For a single synthon, there may be multiple reasonable synthetic equivalents; additional research would be required to determine the best choice. Each derived precursor becomes the target molecule for further analysis. This retrosynthetic procedure is repeated until simple structures, commercially available compounds, or in-house compounds are reached.

A synthesis tree is a directed acyclic graph of alternative retrosynthetic routes for a single target molecule. The root node represents the target molecule, the internal nodes are the intermediate structures, and the leaf nodes are the starting materials. The edges connecting the nodes refer to potential chemical pathways. Analysing the graph edges can determine the feasibility and efficiency of the synthetic routes. Retrosynthetic analysis requires the knowledge, experience and intuition of synthetic chemists. The textbook, The Logic of Chemical Synthesis,[64] summarises the principles of retrosynthesis. Corey's pioneering work was recognised and honoured with a Nobel Prize in 1990 for "the development of the theory and methodology of organic synthesis".[61]

Automating retrosynthetic analysis would significantly reduce the time taken to plan, improve efficiency, and reduce the costs of chemical syntheses. In the 1960s, Corey first researched and demonstrated using computers to plan synthetic routes by retrosynthetic methodology and AI.[60,62] Early pioneering work in this field primarily led to developing template-based approaches, including Corey's Logic and Heuristics Applied to Synthetic Analysis (LHASA) program.[65–67] Template-based methods rely on a knowledge base of reaction templates. The framework of template-based methods consists of encoding chemical reactions in a machine-readable format, selecting and prioritising the most suitable reaction templates,

applying templates to the target molecule in retrosynthetic analysis, and pruning and ranking the resultant precursors. The history of template-based retrosynthetic analysis tools has been reviewed thoroughly.[28,62,66–74]

Historically, knowledge bases of reaction templates for template-based approaches were manually curated. This labour-intensive process requires significant time and effort to keep up to date as novel syntheses are identified. Advancements in computing power, storage capacity, data availability, reaction databases and data-driven algorithms have provoked renewed interest in CASP in the past decade. Contemporary retrosynthesis methodologies frequently incorporate machine learning techniques, a subfield of AI. Models are built on training data to learn relationships and make predictions. Machine learning algorithms not only contribute to the automated extraction of reaction rules from pre-compiled reaction databases but also to the ranking of the reaction templates and the scoring of reaction precursors.

Deep learning is a subfield of machine learning with algorithms inspired by the human brain. Speech recognition, image recognition, and natural language processing are examples of various fields which have successfully implemented deep learning techniques. Deep learning is also prevalent in contemporary CASP strategies. Graph neural networks (GNN) process data represented as graphs and have been implemented to process molecular graphs.[75–78] Natural Language Processing (NLP) techniques have governed template-free strategies to CASP by directly translating reactants to products, or vice versa.[79] These techniques include sequence-to-sequence and Transformer models.[80–82] Detailed perspectives of contemporary CASP strategies can be found in[24,26,28,73,74,83–88].

CASP techniques extend to forward reaction prediction and reaction condition optimisation. Forward reaction prediction is the prediction of products given the reactants, reagents and a set of reaction conditions. Experiments predominantly used to identify reaction outcomes are expensive, time-consuming and require experienced chemists. It would thus be beneficial for computational tools to identify the major product and any side products and validate retrosynthetic predictions. Optimising reaction conditions, such as catalysts and solvents, is also essential in synthesis planning. Changing a set of reaction conditions, even slightly, could result in the formation of a different major product or a failed reaction. Integrating CASP with high-throughput screening and robotic equipment holds much promise for the future of reaction optimisation.

In the past decade, CASP tools have been revolutionised by the availability of big data, the establishment of reaction databases and the advancement of data-

driven techniques. This chapter reviews contemporary CASP strategies, primarily focusing on retrosynthetic analysis and forward reaction prediction. Initially, the fundamentals of CASP are described. We then evaluate the two dominant approaches to retrosynthetic analysis and forward reaction prediction: template-based and template-free. The advantages and limitations of each approach are discussed in a detailed comparison of methods. Lastly, the future outlook and potential challenges in this field are outlined.

## 3.2 The Nature of the Chemical Data

CASP tools are based on historical chemical reaction data, which captures how an experiment was performed and its outcome. The number of reactions in the literature grows exponentially, doubling every 10-15 years.[89,90] Approximately 3000-5000 novel reaction classes with distinct mechanisms are published every year.[91] A chemical reaction chemically transforms one or more reactants to products under specific conditions. Chemists use chemical equations as a symbolic representation of chemical reactions. Reactants are on the left of the arrow, products on the right, reagents above, and operating conditions below (Figure 3.1).

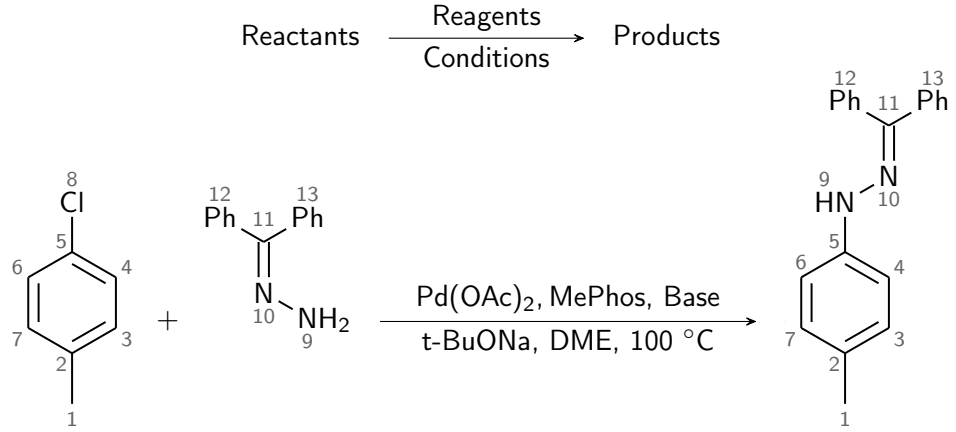



Figure 3.1: An example of an atom-mapped reaction. Reaction taken from reference[1].

The reactant(s) undergo a change in connectivity at the reaction centre during a chemical reaction and contribute to the atoms in the product(s). Atom-to-atom mapping aligns the atoms in the reactant species to the product species. It is a valuable technique to ensure the conservation of atoms and aid the identification of the reaction centre (Figure 3.1). Depending on the reaction type, multiple products may form. These can be classified as primary products, by-products or side-products. Primary products are the desired products of a chemical reaction,

while by-products are produced directly from the desired chemical reaction; both appear as part of the fully balanced chemical equation.[92] Side-products refer to undesired products of a competitive pathway or further degradation of the primary products.[92] These reduce the experimental reaction yield of the primary products. Varying reaction conditions could potentially suppress the formation of side products. The environmental conditions, such as reagents and physical variables, under which a reaction is optimal are defined in the chemical equation. Reagents participate in the chemical reaction but are not consumed and do not contribute to the atoms in the product. Examples of reagents include catalysts and solvents. Physical variables include temperature and pressure. Along with the chemical equation, chemists also record supplementary data such as experimental procedures, reaction class, experimental reaction yield, and Enantiomeric Excess (%ee).

CASP methods typically employ supervised learning methods to learn patterns of chemical reactivity from chemical reaction data. Training such machine learning algorithms requires large amounts of labelled data. The input-output pairs could be product-reactants for retrosynthesis or vice versa for reaction prediction. Chemical reaction data can be manually curated by expert chemists or automatically extracted from data sources. In the era of big data, current methods primarily use reaction knowledge mining from data sources, including in-house Electronic Laboratory Notebook (ELN), literature, patents, and commercially available reaction databases. Chemical reaction data are noisy and can contain duplicate and even erroneous reactions. The extracted reaction data is cleaned, filtered, and converted into a machine-readable representation.[93,94] The lack of a conventional approach for data extraction has resulted in different types of information, managed in different ways, used as the input to the machine learning methods. This section discusses molecular representations, data sources, and the progression of data sources.

### 3.2.1 Molecular Representations

For a machine learning algorithm to learn from chemical structures, they must be represented in a way that an algorithm can process. Molecular descriptors are based on structural, physiochemical, electronic, or topological properties. The type of molecular descriptor used can affect the performance of the machine learning model. Molecular descriptors commonly implemented in synthesis planning tools are based on the chemical structure of molecules; examples include molecular fingerprints, molecular graphs, and Simplified Molecular-Input Line-Entry System (SMILES)[95] strings. The theory of these molecular descriptors

is detailed in Chapter 2, Section 2.5. Structure-based descriptors are relatively quick and easy to calculate. Section 3.3 discusses how molecular descriptors are employed in various approaches to CASP.

### 3.2.2 Data Sources

Chemical reaction data is stored in various formats, including patents, journal articles, and ELNs. Despite the vast amount of data sources, few curated databases record organic syntheses in a structured, standardised format and are freely and publicly available. Current CASP tools are developed using licensed commercial data, proprietary in-house ELN data or the open-access United States Patent and Trademark Office (USPTO) dataset.[96]

#### Commercial Database Systems

Commercial database systems provide access to chemical reaction data reported in scientific journal publications and patents. Reaxys[97] from Elsevier and SciFinder$^{n}$[98] from Chemical Abstracts Services (CAS) are chemical search engines designed to retrieve chemical information and data from published literature. These repositories contain millions of chemical compounds, reactions, and properties with appropriate citations (Table 3.1). Pistachio[99,100] from NextMove Software is a reaction database and search system containing millions of reactions extracted from patent data (USPTO, EPO and WIPO). Subject to a license agreement, chemists can search through vast regions of chemical and reaction space to fill knowledge gaps.

Table 3.1: Commercial Database Systems

| Database | Reference | Number of Reactions (Million) |
|---|---|---|
| Reaxys | 97,101 | 49 |
| SciFinder$^{n}$ | 98,102 | 150 |
| Pistachio | 99 | 9 |

Chemical search engines are indispensable in practically every synthetic chemist's workflow. Analysing relevant syntheses and the reactivity of analogous structures aids decision-making in synthetic chemistry. While analysing chemical information is a manual process that requires the intellect and time of expert synthetic chemists, chemical search engines enable vast amounts of data to be instantly filtered and prioritised. These tools facilitate a more focused and directed approach to conducting synthesis research on a shorter timescale.

Licenses can be purchased to extract large quantities of chemical reaction data

from commercial databases for data analysis or the training of machine learning models. Typically, the chemical information provided is not recorded in a standardised format. Details, such as the structures of reactants, reagents, products, and experimental procedures, are left as unstructured text in the original document.[103] Research groups have extracted chemical reactions from Reaxys and built models for forward reaction prediction,[104,105] retrosynthetic planning[105–107] and the prediction of reaction conditions.[108] Segler and Waller used the chemical reaction data extracted from Reaxys to develop a multi-step retrosynthesis application using neural networks and symbolic AI.[104–106] The Bishop group use the Reaxys data in a different approach to multi-step retrosynthetic planning based on reinforcement learning.[107] Pistachio has also been utilised in CASP tools.[81,109] The Lee group utilised the non-public patent data from Pistachio to generate a time-split test set in the development of a machine translation model for forward reaction prediction.[81]

Although commercial databases provide access to millions of chemical reactions, the data is biased toward high-yielding reactions and requires a commercial license. CASP tools developed and validated using commercial data may be based on biased machine learning models with limited comparability to open-source applications.

### Electronic Laboratory Notebooks

An ELN is a software tool that replaces traditional paper laboratory notebooks. Chemists document experimental procedures, experimental data, and supplementary notes in laboratory notebooks. A laboratory notebook is a legal document that can act as evidence in legal matters (e.g., patent disputes) to protect Intellectual Property (IP).[110] Capturing and storing vital experimental research in a digital format benefits the user, organisation, external collaborators, and the broader scientific community. ELNs enable researchers to access, search, share and backup experimental documentation. Digitally capturing experiments facilitates long-term data storage, reduced data misplacement or loss risk, enhanced experimental record availability, IP protection, collaboration, and open science.[110–112] Standardising experimental records improves the reproducibility of experiments and simplifies the curation of chemical data. Developing ELNs to store data in a machine-readable format improves interoperability and allows integration with third-party tools.[112,113] If permitted, the data can also be exported as an external dataset to support open science.

ELNs store valuable information about chemical reactions including reagents, experimental conditions, reaction yields, and various spectra. ELNs help facilitate

the compilation of high quality, reliable, and reproducible data.[110] Data is often recorded for all chemical reactions performed regardless of their success. A data source containing low-yielding or failed chemical reactions, or both, is advantageous for training machine learning algorithms. Although ELNs are prominent in industrial research, uptake is limited in academia.[110,112,114,115] ELN data generated in industry are not usually accessible to the public, nor in commercial databases, due to confidentiality reasons. Within the pharmaceutical industry and through various collaborations with academia, ELNs have proved a valuable data source in developing CASP tools, such as reaction prediction,[116] retrosynthetic analysis,[94,116] and the prediction of reaction conditions.[117]

### Open-Source Patent Data

In 2017, D. Lowe released a large dataset of machine-readable chemical reactions to the public.[96] The dataset contained approximately 1.8 million organic chemical reactions from US patents and applications published between 1976 and September 2016. Text mining was used to extract the following experimental details from the USPTO: structures of reactants, products and reagents, reaction conditions, reaction yield, synthesis steps, and patent source.[96,118–120] A complete list of the details extracted can be found in Appendix A. The open-source USPTO 1976-2016 dataset is a subset of the reactions in the Pistachio database, which includes chemical reactions extracted from the USPTO dating from 1976 to May 2018.

The chemical reactions in the USPTO 1976-2016 dataset are encoded as Reaction SMILES,[56] where each molecule in a chemical reaction are expressed as SMILES strings. As an example, the USPTO 1976-2016 dataset contains the Reaction SMILES shown in Figure 3.2.[96] The reactions in the dataset have been atom-mapped, meaning corresponding atoms in the reactants and products are labelled accordingly in the Reaction SMILES string.

The USPTO 1976-2016 dataset is a common open-source dataset used in developing CASP tools. It contains prevalent chemical reactions used in medicinal chemistry. The raw data is often preprocessed as it contains frequent duplicate chemical reactions due to similar text in multiple patents and numerous cases of incorrect atom mapping.[96] A handful of public benchmarking datasets have been derived from the raw data and implemented in several research groups (Table 3.2). In descending order of dataset size, prevalent subsets of the USPTO 1976-2016 dataset in the literature include USPTO-FULL,[75,121] USPTO-MIT,[122,123] and USPTO-50K.[124]

(a)

Reactants > Reagents > Products

(b)

[CH2:1]([O:3][C:4](=[O:9])[CH2:5][C:6](=[NH:8])[NH2:7])[CH3:2] . Cl[CH2:11][CH:12]=O
>
CCOC(C)=O
>
[CH2:1]([O:3][C:4]([C:5]1[CH:12]=[CH:11][NH:8][C:6]=1[NH2:7])=[O:9])[CH3:2]

(c)

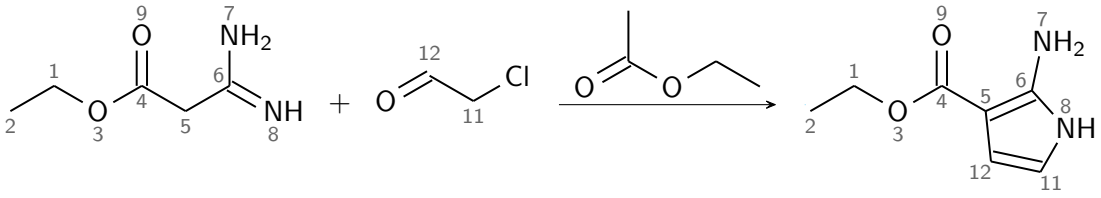



Figure 3.2: Patent US09447100B2, paragraph 0548, from the USPTO 1976-2016 dataset. (a) Generic reaction smiles. (b) Reaction smiles of patent US09447100B2. (c) Atom-mapping of patent US09447100B2.

Table 3.2: USPTO Benchmarking Datasets

| Dataset | Reference | Number of Reactions | Split | | |
|---|---|---|---|---|---|
| | | | Train | Validation | Test |
| USPTO 1976-2016 | Lowe Figshare[96] | 1,808,937 | - | - | - |
| USPTO-FULL | GLN Repo[121] | 1,013,118 | 810,496 | 101,311 | 101,311 |
| USPTO-MIT | WLN Repo[123] | 479,035 | 409,035 | 30,000 | 40,000 |
| USPTO-50K (Liu) | Seq2Seq Repo[125] | 50,037 | 40,029 | 5,004 | 5,004 |
| USPTO-50K (Coley) | GLN Repo[126] | 50,016 | 40,008 | 5,001 | 5,007 |

The USPTO-FULL dataset contains approximately one million cleaned, unique single-step reactions.[75] The reactions with multiple products were duplicated into multiple reactions producing a single product to generate a dataset of entirely single-step reactions. Chemical reactions that were duplicates or had incorrect atom mappings were removed. Dai *et al.* divided the USPTO-FULL dataset into 80%:10%:10% training/validation/test sets.[75]

The USPTO-MIT dataset was derived from the USPTO 1978-2016 after removing duplicate and erroneous reactions. The dataset contains approximately 480,000 fully atom-mapped single- and multi-step reactions without stereochemical information, split into 400,000 training, 30,000 validation, and 40,000 test reactions. The USPTO-MIT is not as diverse as the other datasets, including USPTO-50K.

The first benchmarking dataset derived from open-source patent data was published in 2016 before the full USPTO 1976-2016 dataset was released. As a result,

the USPTO-50K dataset only contains patent data up to 2015. The atom mappings between reactants and products were updated to ensure high accuracy, duplicate reactions were removed, and reaction types were assigned. A total of 50,000 reactions were randomly selected. The reactions were classified into ten distinct reaction types (Table 3.3) which cover common reactions in the medicinal chemist's toolkit.[124] Subsequent studies by two separate research groups, Liu *et al.*[125,127] and Coley *et al.*,[75,126,128] have further processed the USPTO-50K dataset. Both groups generate single-step reactions by splitting the reactions with multiple products into multiple single-product reactions. Any chemical reactions with trivial products, such as inorganic ions and solvent molecules, were removed. The resultant 50,000 reactions were split into 80:10:10 percent training/validation/test sets. USPTO-50K is considered the standard benchmarking dataset in the analysis of CASP tools. Multiple research groups employ this benchmark for model comparison.

Table 3.3: Ten Reaction Classes in the USPTO-50K Dataset[124]

| Reaction Class | Reaction Name | Size (%) |
|---|---|---|
| 1 | Heteroatom alkylation and arylation | 30.3 |
| 2 | Acylation and related processes | 23.8 |
| 3 | C-C bond formation | 11.3 |
| 4 | Heterocycle formation | 1.8 |
| 5 | Protections | 1.3 |
| 6 | Deprotections | 16.5 |
| 7 | Reductions | 9.2 |
| 8 | Oxidations | 1.6 |
| 9 | Functional group interconversions | 3.7 |
| 10 | Functional group addition | 0.5 |

### 3.2.3 Progression of Data Sources

There are various sources of chemical reaction data, including published literature, patents, and ELN data. Despite the variety of data sources, there are underlying issues when using the data to build data-driven models. The chemical reaction data are often biased, not stored in a machine-readable format, and not publicly available.

Reaction data published in journals and patents are biased toward high-yielding reactions.[129,130] Low-yielding and failed reactions are reactions with unreacted starting materials or unexpected products. Examples of these reactions are uncommon in the literature, which is particularly problematic when implementing supervised machine learning models. Counterexamples are required in training to prevent model bias and optimise generalisability. One way to overcome the ab-

sence of failed data is to generate artificial negative examples. Reaction templates are applied in the forward direction to the reactants of reported reactions. This generates chemically plausible "wrong/false" products, for example with incorrect regioselectivity.[104,106,131] Alternatively, negative examples can be generated by shuffling the pairs of products and corresponding reactions.[106]

Although significant scientific knowledge is available in the published literature, it is not always free to access or straightforward to export. Open-source publishing and reproducibility of data have been subject to discussion.[132,133] In patents, journal articles and supporting information, published reaction data are often recorded in a hard-to-parse PDF format. The structural drawings of chemical structures are challenging to translate into a machine-readable format. Text mining is currently required to retrieve reaction data from text. Providing reaction data in a readily machine-readable format is vital for enhanced scientific growth. It would help to improve reproducibility and to expand on previously published work. Establishing guidelines enforced by journals could improve publishing practices, such as uploading reaction information to a structured repository, ensuring the data is machine-readable, and disclosing source code.

ELNs are commonly used in the pharmaceutical industry. Reactions frequently used by medicinal chemists restrict the reaction space covered in ELNs. ELNs may not contain all reaction information. Material costs and knowledge transfer protocols for scale-up, purification and crystallisation may be stored elsewhere.[134] ELN data generated in the pharmaceutical industry is frequently inaccessible as data confidentiality limits access, collaboration and sharing with third parties.

Multi-step syntheses are difficult to record in ELNs. As a result, ELNs favour single-step reactions. Reaction data from High Throughput Experimentation (HTE) and Design of Experiment (DOE) screening are similarly challenging to record in ELNs. Screening techniques generate hundreds to thousands of data points at a quick pace. If the ELN software does not support recording high throughput data, each reaction must be manually recorded as an individual entry in the ELN.

ELNs require continuous upgrading to ensure user needs are satisfied, improve the infrastructure and interoperability, and streamline data extraction for subsequent reuse. The entry fields are a mix of free-text, restrictive-text, and multiple-choice options. Increasing the number of specific entry fields increases rigidity and ease of exportation to generate machine-readable datasets; at the expense of adaptability.

While ELNs and commercial databases are useful data sources, they are not considered further in this thesis due to the lack of availability for benchmarking software. Benchmarks are required to compare the performance of CASP tools, including the reaction data underlying the tool and assessment criteria. Multi-step retrosynthetic planning also requires a standard for the database of commercially available building blocks. A benchmark dataset must be open source to negate financial costs and accessibility issues. The current standard benchmarking reaction dataset is the USPTO-50K dataset. This dataset should be considered with caution as it may contain prophetic examples.[135] Prophetic examples are anticipated experimental methods and results which are yet to be proven. They are acceptable in US patents, provided they are not in the past tense. Models based on prophetic examples may predict synthetic routes based on undetermined reactions.

Progress towards a centralised public repository of chemical reaction data is a priority. A centralised repository would accelerate the development and growth of downstream applications involving CASP tools. Any data repository should adhere to the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles for data management.[136] Both humans and machines must be able to find the data. The location should be apparent for quick and easy retrieval. Providing the data in a machine-readable format and consistent representation is essential for automated data retrieval. The data should be open-source whenever possible, which is particularly important when comparing CASP tools. Data processing, analysis, and storage require interoperability. The infrastructure should support integration with third-party data, applications and workflows. For example, all reaction data should be able to be recorded regardless of the reaction setup, such as bench reactions, automated high throughput screening, and flow chemistry. The fundamental goal is to ensure that the data is reusable. Well-described data is required to facilitate easy replication. Maintaining high-quality data minimises the amount of time to clean and remove noise. The FAIR principles for data management extend to metadata (information about the data) and software infrastructure (access to the data).

The Open Reaction Database (ORD) is one endeavour towards a centralised repository.[103] It is an open-access schema and infrastructure for improving access and encouraging the sharing of chemical reaction data. Structured, public and freely available data are provided in a user-friendly interface for viewing and downloading.

## 3.3 Approaches to Computer-Aided Synthesis Planning

The following sections focus on two primary types of synthesis planning tools that are inverse processes: forward reaction prediction and retrosynthetic analysis. Forward reaction prediction aims to estimate the likely products of a reaction using prior knowledge of reactant precursors. Single-step retrosynthesis begins with a target molecule and proceeds backwards, predicting potential bond disconnections to acquire simpler reactant precursors.

Forward reaction prediction is less complex than single-step retrosynthesis as the input (reactants) contains all reacting functional groups. This limits the number of possible reaction types and thus reduces the number of possible outputs.

There is no single correct answer in single-step retrosynthesis. The reacting functional groups in the output (reactants) are absent from the input (products). For example, consider the Buchwald-Hartwig reaction shown in Figure 3.3. The halide leaving group is not present in the amine product. For a single bond disconnection, multiple reaction types may be plausible. As demonstrated in Figure 3.3, the carbon-nitrogen bond in the target could be synthesised via a Buchwald-Hartwig or Chan-Lam reaction. Unless the reaction type is assigned and provided, a large number of precursors may be feasible from a single bond disconnection. In larger molecules, there may be several disconnection sites, resulting in a wider pool of possible precursors.

Template-based and template-free frameworks are two dominant approaches to forward reaction prediction and single-step retrosynthesis. These approaches are defined and outlined in this section.

### 3.3.1 Template-Based Framework

Traditionally, computer programs for retrosynthetic analysis and forward reaction planning were based on reaction templates. The terms "*reaction template*" and "*reaction rule*" are commonly used interchangeably in the literature and refer to encoding a chemical transformation in a machine-readable format. The term "*reaction template*" is used herein. Reaction templates can be defined in either the forward (reactants to products) or reverse (products to reactants) direction.

Figure 3.4 depicts the template-based framework for single-step retrosynthesis and forward reaction planning. The three primary components are a pre-defined library of reaction templates, a template application engine, and a scoring func-

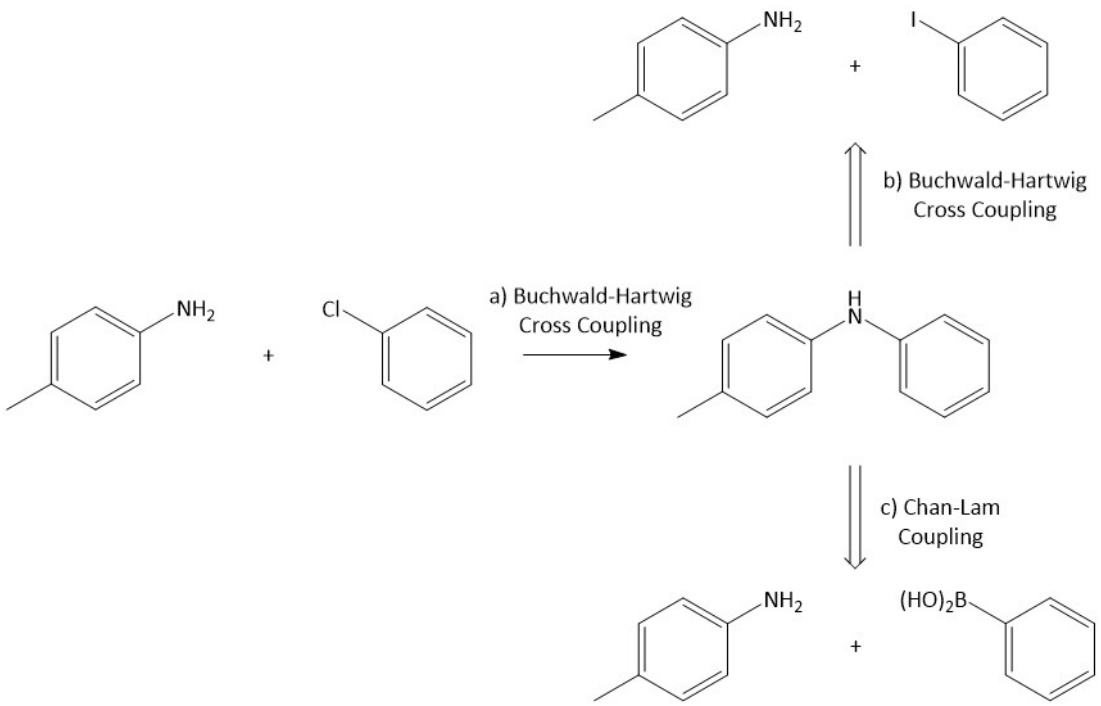



Figure 3.3: An example of forward synthesis planning (a) and retrosynthesis route design (b) and (c).

tion. The library of reaction templates contains possible disconnections in retrosynthesis or possible reactions in forward reaction planning. Reaction templates cannot be applied to all molecules. For example, to apply a template of a carbon-nitrogen bond disconnection, the target molecule must contain a carbon-nitrogen single bond. The template application engine must determine if the reaction template is applicable before enumerating the pathway in the forward or reverse direction. The scoring function can be used to score and rank the feasibility of the generated candidates[131] or more commonly the reaction templates.[75,76,105,106,128,137–140]

### Defining Reaction Templates

A reaction template encodes the changes in atom connectivity during a chemical reaction.[141,142] Not all atoms and bonds are included in a reaction template. The chemical transformation is generalised to enable the reaction template to be applied to overlapping sets of molecules.[143] Minimal reaction templates only encode the reaction centre. The reaction centre is the change in atoms, bonds and bond orders during bond formation and breaking. Reaction templates are frequently extended beyond the reaction centre to include neighbouring atoms that may influence the chemical reaction. The neighbouring atoms could be admissible substituents, incompatible groups, or have physical-organic effects. Such effects include electron densities, steric bulk, and molecular strain. Complex templates

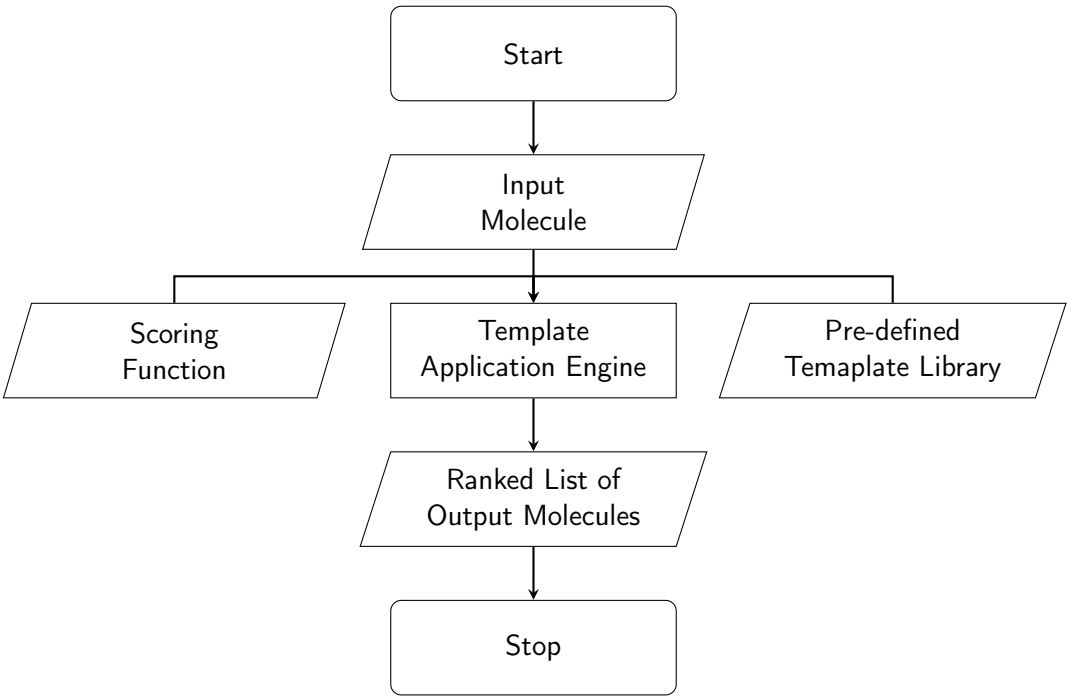



Figure 3.4: Template-based framework for single-step retrosynthesis and forward reaction planning.

may contain reactivity conflicts, protection requirements, stereoselectivity,[137] or regioselectivity.[91,144] Templates may not include additional information due to data availability, inconsistency, and difficulty encoding.[94,131] To minimise the reaction search space, reaction conditions are frequently omitted from templates and solved as an independent problem.

The performance of template-based models is dependent on the number of reaction templates and the size of the templates. The number of reaction templates defines the size of the reaction search space. There is an inevitable trade-off between generalisation and specificity when defining the size of the reaction templates. Increasing the number of atoms in the template increases the specificity and accuracy of the transform while reducing the number of molecules in the template is applicable (generalisability). Specific templates cover a smaller amount of overlapping chemical reactions. As a result, the number of templates, the computational cost, and the time required to encode chemical reactions increases. If the templates contain too few atoms neighbouring the reaction centre, they become overly generic, and crucial information about the indirect effects of neighbouring atoms is neglected. A model based on this limited information cannot perceive potential reactivity conflicts, which may result in incorrect template application and inaccurate predictions. If too many neighbouring atoms are included in the template, they are overly specific with a limited scope, which results in poor generalisability.

Reaction templates are written in a user-friendly syntax, usually specific to the developers. The SMILES Arbitrary Target Specification (SMARTS) language, an extension of the SMILES line notation, and descriptors derived from molecular graphs are examples of commonly used machine-readable formats. Early programs manually hand-coded reaction templates.[65–67,145,146] This method suffered from three fundamental issues. (1) The intuition of expert synthetic chemists with years of experience was required to encode each reaction. (2) Manual coding was a labour-intensive task that could not be scaled up to include an adequate number of chemistries. Template libraries were typically incomplete and covered a limited scope of reaction classes.[65,67] (3) The complexity of reactions made codification problematic. Chemical reactions are not straightforward due to reactivity conflicts and dependence on reaction conditions.

The well-known program Synthia (formerly Chematica), commercialised by Sigma-Aldrich, incorporates the largest database of hand-coded reaction templates.[91,144,147] The database took over ten years to curate. It contains a total of 75,000 hand-coded reaction templates. To hand-code a chemical reaction, the mechanism is initially studied and understood. The core of the transformation is coded as reaction SMARTS. Functional groups that need to be protected, groups that are always incompatible in the reaction, typical reaction conditions, representative literature sources, and other additional information are also included in the reaction template.[91] Quality control procedures reduced human error in the encoding process. These procedures included scripts for syntax checking, testing template applicability, peer-review cross-checking, and final verification before input into the reaction template database.[91] As novel chemistries are discovered and old chemistries refined, the database of hand-coded reaction templates is continually updated. Synthia has demonstrated that hand-coding is feasible and manageable if sufficient time is invested. Synthia is a successful implementation of hand-coded reaction templates in the prediction of retrosynthetic pathways.[91,144,147]

Automated template extraction from published reaction data is more efficient and time-saving. The reaction centre and neighbouring atoms are extracted algorithmically from an atom-mapped SMILES pattern. Heuristics are then used to incorporate groups known to influence the reaction. Atom-mapped reactions are required to identify the atoms and bonds that change during the chemical transformation.[94,106,137,141,148–151] Tools, such as the Reaction Decoder Tool (RDT), are available to calculate atom map indices if they are unknown.[149,150,152] The reaction centre is identified by iterating around the molecule, recording changes in atom environments. The shell- or radius-based approach is commonly used to incor-

porate neighbouring atoms up to a pre-specified number of bonds. Radius-based template extraction is implemented in RDChiral,[141,153] Monte Carlo Tree Search (MCTS),[106] ARChem (Route Designer),[148] and InfoChem's CLASSIFY.[154]

Reaction templates generated from algorithmic extraction are less robust than manually hand-coded templates. Extracted templates may be chemically inaccurate, contain duplicate and nonexclusive templates, and neglect long-distance effects. Canonicalization is the process of producing a unique reaction template for a chemical transformation, which is essential in deduplication. Poorly canonicalized templates describe the same chemical transformation on the same set of molecules.[143] Nonexclusive templates describe the same chemical transformation but include different or no special groups on overlapping sets of molecules.[143] Poorly canonicalized and nonexclusive templates add unnecessary noise to the library of reaction templates. RDChiral is a template extraction and application algorithm that canonicalizes the extracted reaction templates while handling stereochemical information.[141,153] Not all template extraction techniques handle stereochemical information correctly. Automated extraction does not account for long-distance effects. Atoms or groups that are more than four atoms from the reaction centre are typically not encoded in the reaction template. The accuracy of automated template extraction has been discussed thoroughly.[91,143,155] The advantages of automated template extraction are speed and scalability. Extracted template libraries are easier to maintain compared to hand-coded libraries. Algorithmic template extraction has been utilised in forward reaction prediction[105] and retrosynthesis planning.[75,76,94,105–107,128,137–140]

### Template Selection

There may be numerous suitable reaction templates for a single target molecule or reactants in the case of reaction prediction. For large template libraries, applying every template is computationally expensive and often intractable. Various algorithms have been developed to identify the appropriate templates while accounting for chemical context, trade-offs, and reactivity. Template selection can be considered a multi-class classification problem where each template in the template library is a class. A machine learning model learns how to classify reactions into applicable templates from the library. The template selection models are restricted to predicting templates in the template library. While the models can interpolate known reactions encoded in the template library to a novel target, they cannot extrapolate to reactions not present in the library or novel chemistries.

One approach to selecting appropriate reaction templates to apply to the input

molecule is to train a neural network to predict the probability of relevance for each reaction template in the library.[75,105,139] NeuralSym (Neural-Symbolic) is an example of this approach, which uses molecular fingerprints to represent the molecules. Using machine learning to determine template applicability can be biased towards popular reaction classes. When there are too few examples of a reaction class in the dataset, it is unclear which substituents are admissible or conflicting. Fortunato *et al.* built on the NeuralSym framework, employing a pre-train and data augmentation strategy to reduce bias toward well-represented templates and thereby extending the scope of the training set.[139] Although this strategy enhanced the performance of rare templates, the model remained limited to interpolation. The Graph Logic Network (GLN) adopts a similar framework with a molecular graph representation to improve interpretability.[75]

An alternative approach is a similarity-based template application implemented in Retrosim[128] and exclusively used in retrosynthesis. Rather than defining a library of reaction templates, Retrosim generates them on demand. Reaction templates are extracted from the most similar products in the training set to the target molecule. The reaction templates are applied to the target to generate reactant candidates. The candidates are scored and ranked using a combination of reactant and product similarities.

The template selection approaches discussed above have relied on the global features of the target molecules. Chen *et al.* proposed a graph-based framework based on locally derived reaction templates.[76] The applicability of the local templates is evaluated using atom or bond features at each reaction centre. As a result, the proposed LocalRetro model focuses more on local information.

### Advantages and Limitations

Template and similarity-based approaches are interpretable as they align with how synthetic chemists think. The proposed pathways can be traced to the underlying data of successfully performed reactions and the reasoning behind the decision-making. Despite this, encoding chemical reactions is a bottleneck. The constant maintenance of template libraries is time-consuming and can be challenging when including new chemistries. Although hand-coding requires care to eliminate human errors, the quality is higher than extracted templates. In automated template extraction, reactions frequently have incomplete or erroneous atom mapping, resulting in duplicate and nonexclusive templates. Extracted templates focus on the local environment surrounding the reaction centre, potentially overlooking non-local influential groups. The advantage of automating template extraction is scalability. Manual encoding is laborious and requires the knowledge

of synthetic chemists. Template-based approaches cannot accurately predict reactions outside of the template library.[105] Rather than discover novel chemistries, template-based methods aim to assist synthetic chemists with routine synthesis tasks.

### 3.3.2 Template-Free Framework

Template-free approaches are fully data-driven, do not rely on predefined reaction templates, and do not require chemical knowledge. Molecular graphs or SMILES strings are typically used to represent the input compound(s). Template-free approaches are further classified as semi-template generation or machine translation methods. Semi-template generation is a two-step method aligned with template-based approaches. Machine translation is a single-step method that is comparable to language translation.

#### Semi-Template-Based Methods

In semi-template-based methods, the output (reactants or products) is predicted by generating intermediates or synthons from input molecules represented by molecular graphs. The two-step procedure is similar to template-based approaches. Rather than selecting reaction templates in the first step, a graph neural network or transformer[156] is used to identify the reaction centre to generate synthons. Convolutional, graph attention and message-passing neural networks are examples of graph neural networks that have been implemented. In the second step, the synthons are completed to produce the output using either a graph generative,[157] transformer,[156,158] or subgraph selection model.[78] Semi-template-based methods have been developed for forward reaction prediction[122,159] and retrosynthetic analysis.[77,78,156–158,160]

#### Machine Translation Methods

Machine translation is a subfield of NLP that focuses on translating text from one language to another.[79] When applied to synthetic chemistry, machine translation methods are trained end-to-end to learn the syntax of chemical reactions. The molecules can be represented by text notation, such as SMILES strings. In forward reaction prediction, the reactant SMILES strings are directly transformed into product SMILES strings, and vice versa for retrosynthesis.

Neural sequence-to-sequence models are based on an encoder-decoder architecture composed of two recurrent neural networks and an attention mechanism.[80,127,161] Each token in the SMILES string is considered sequentially and is assumed to

be related to its neighbour. Sequence-to-sequence models learn the local environments in the SMILES strings and their influence on neighbouring tokens. Two tokens far apart in a SMILES string could be topologically near in the equivalent molecular graph. As a result, these models may have difficulty identifying long-range token relationships. Sequence-to-sequence models have been implemented in reaction prediction[80,161] and retrosynthesis tasks.[127]

The transformer architecture has surpassed the neural sequence-to-sequence architecture. Transformer models are based on a fully attention-based encoder-decoder architecture.[162] Due to the absence of recurrent neural networks, transformer models can correlate individual tokens in the SMILES strings regardless of their location and hence capture long-range interactions. Transformer models have been successfully applied to forward reaction prediction[81,116,163] and retrosynthesis.[116,163–169]

**Advantages and Limitations**

Template-free methods have some advantages over template-based methods. They do not require atom mapping, atom features or any chemical knowledge. As template-free methods are not dependent on a library of reaction templates, they are scalable to large datasets at a fraction of the computational cost and can generalise to novel chemistries.

Semi-template-based methods require developing and training two independent models, one for synthon generation and the other for synthon completion. Semi-template-based methods, like template-based methods, are analogous to a synthetic chemist's thought process. As a result, the predictions generated by these methods are easy to interpret. These methods, however, rely implicitly on predetermined reaction templates and atom-mapped data.

Machine translation approaches are based on an implicit representation of the global environment of molecules. A single model is trained end-to-end to transform the input into the output. The interpretability of machine translation methods is limited. Attention weights of the atoms (i.e. the SMILES tokens) can be used to identify which aspects of the input species influence the output species. Although the attention mechanism enables the models to be interpreted, it is not intuitive to synthetic chemists. Machine translation approaches suffer from SMILES invalidity, a lack of diversity, and chemically implausible predictions. As machine translation methods do not directly learn the terminology of the SMILES notation, they cannot be guaranteed to predict valid SMILES. If a predicted SMILES string is grammatically incorrect, the molecule is consid-

ered invalid even if the grammatically correct version is viable or the ground truth.

### 3.3.3 Performance Evaluation

The performance of synthesis planning models is a balance between accuracy, reliability, and adaptability. Top-$n$ exact match accuracy is the most widely used metric for evaluating overall performance. It is the percentage of test set molecules whose ground truth was ranked within the top-$n$ predictions. A higher top-$n$ accuracy indicates a better model performance. The top-1, top-3, top-5, and top-10 accuracies are commonly reported.

The reliability of a model's predictions could be estimated to improve its interpretability. The probability that a prediction is correct can be used to calculate a confidence score. This degree of uncertainty can then be used to determine if a prediction is incorrect.[81,116] Template-free models can suffer from syntax errors in their predictions. This is common in machine translation methods. Top-$n$ invalid rate is the percentage of chemically invalid predictions in the top-$n$ predictions. A higher top-$n$ invalid rate indicates more syntax errors.

The adaptability of synthesis planning models is crucial. While the accuracy of a model on a small amount of data may be high, the accuracy may decay as the amount of data increases.[160] The accuracy may also vary when trained and then tested on different regions of chemical space. For example, when a model is trained on patent data and used to predict synthetic routes in ELNs.[116] Accuracy, reliability, and adaptability are valuable to evaluate in synthesis planning.

## 3.4 Retrosynthetic Analysis

Retrosynthesis is a fundamental task in synthesis planning in drug research and development. The goal of retrosynthetic analysis on a novel target molecule is to discover the complete synthetic pathway, a succession of single-step transformations. The common painkiller paracetamol, for example, can be synthesised in two steps. The initial step is a reduction of para-nitrophenol, followed by the acetylation of para-aminophenol with acetic acid or acetic anhydride.

Single-step computational models can be applied to the target recursively until the precursors are readily available or a termination criterion is satisfied. More accurate single-step predictions would improve the success rate of multi-step approaches. The reaction search space grows exponentially as the number of reaction steps increases. Due to the enormous potential search space, the com-

putational cost is substantial. Efficient and effective computer-aided approaches are required to navigate the search space.

### 3.4.1 Single-Step Retrosynthesis

Single-step retrosynthesis is the simplest form of retrosynthesis, which aims to break down the target into potential reactant precursors given a target molecule. For the single-step retrosynthesis problem, several template-based, semi-template-based, and machine translation methods have been developed.

This section compares the top-$n$ accuracy of several single-step retrosynthesis applications on the USPTO-50K benchmark dataset. All top-$n$ accuracy values are taken directly from the literature. The template-based models are Retrosim, Neuralsym, GLN, and LocalRetro. The semi-template-based models are G2Gs, RetroXpert, GraphRetro, RetroPrime, MEGAN, R-SMILES, and Graph2Edits. The machine-translation-based models are Seq2Seq, Augmented Transformer, SCORP, GTA, Dual-TF, Tied Transformer, Graph2SMILES, and R-SMILES. All methods are outlined below.

**Retrosim.** Retrosim[128] is a similarity-based approach in which disconnections are made strategically based on similarity to known reactions. Reaction precedents are retrieved from a knowledge base based on product similarity. Reaction templates are extracted from reaction precedents and applied to the target, generating candidate precursors. Reactant similarity is used to score the candidate precursors. The overall similarity is calculated from the reactant and product similarity scores and used to rank the candidate precursors.

**Neuralsym.** The hybrid Neural-Symbolic (NeuralSym)[105] model uses neural networks to prioritise reaction templates before application. NeuralSym learns the named reaction used to synthesise the target by multiclass classification for template selection.

**GLN.** The conditional Graph Logic Network (GLN)[75] is built on GNNs to learn when to apply reaction templates. GLN uses graph embeddings to model the conditional joint probability of rules and reactants.

**LocalRetro.** LocalRetro[76] focuses on local reaction templates involving atom and bond edits. A GNN accounts for the local reactivity and a global attention mechanism accounts for the nonlocal effects of the chemical reaction. The pre-

dicted local templates are scored, ranked, and applied to the target molecule to obtain the final ranked reactants.

**G2Gs.** The Graph-to-Graphs (G2Gs)[77] method converts a target molecular graph to a set of reactant molecular graphs. The reaction centre of the target is identified using a GNN and used to split the target into synthons. Graph translation is used to convert the synthons into the final reactant graphs.

**RetroXpert.** The Retrosynthesis eXpert (RetroXpert)[158,170] initially uses a GNN to identify the reaction centre of the target and generate synthon molecular graphs. The synthon molecular graphs are converted to SMILES strings. A Transformer-based[162] sequence-to-sequence model generates reactant SMILES strings from the synthon SMILES strings.

**GraphRetro.** GraphRetro[78] uses a GNN to predict a series of graph edits, converting a target to synthons. To expand the synthons to reactants, leaving groups are attached from predefined chemical rules.

**RetroPrime.** RetroPrime[156] consists of two Transformer models: one that converts the product into synthons and another that converts the synthons to reactants. SMILES strings represent the molecules.

**MEGAN.** Molecular Edit Graph Attention Network (MEGAN)[171] initially generates reactants by performing a sequence of graph edits, i.e. bond changes, to the target. A graph attention network modifies the target sequentially by generating intermediate substrates until it terminates and gives the predicted reactants.

**R-SMILES.** The Root-aligned SMILES (R-SMILES)[172] specifies a tightly aligned one-to-one mapping between the product and reactants SMILES. R-SMILES uses the same starting atom (root) of the SMILES string of the reactants and products, decreasing the edit distance. The model is a Transformer with data augmentation fine-tuned on reaction data after being pre-trained on unlabelled data.

**Graph2Edits.** Graph2Edits[160] uses a GNN to predict graph edits sequentially, generating intermediates. To complete the reactants, a leaving group is attached to the intermediate. Although comparable to MEGAN, the Graph2Edits method has a simpler graph-to-edits network architecture.

**Seq2Seq.** The Sequence-to-Sequence (Seq2Seq)[80] model is a machine translation method which maps the product SMILES string to reactant SMILES strings. Two RNNs and an attention mechanism comprise the model. The Long Short-Term Memory (LSTM) network, an RNN, is used as the encoder and the decoder.

**MT** The Molecular Transformer (MT)[81] is a machine translation method based on the attention Transformer architecture.[162] The product SMILES string is converted to reactant SMILES strings.

**AT.** The Augmented Transformer (AT)[82] uses data augmentation of the SMILES string to decrease overfitting and improve the accuracy of the neural network Transformer architecture.

**SCROP.** The Self-Corrected Retrosynthesis Predictor (SCROP)[166] has a Transformer framework with an additional neural network-based syntax corrector. The syntax corrector reduces the number of invalid SMILES strings.

**GTA.** The Graph Truncated Attention (GTA)[173] adds molecular graph information into the attention layers of a transformer model.

**Dual-TF.** Dual-TF[174] is an Energy-Based Model (EBM) framework that combines graph- and sequence-based models with various energy functions. A dual EBM variant is constructed based on the agreement between forward and backward reaction prediction.

**Tied-TF.** Tied-TF[165] couples two transformers with latent modelling. One transformer is for retrosynthesis prediction, while the other is for forward reaction prediction.

**R-SMILES** The Root-aligned SMILES (R-SMILES)[172] specifies a tightly aligned one-to-one mapping between the reactant and product SMILES. R-SMILES uses the same starting atom (root) of the SMILES string of the reactants and products, decreasing the edit distance. The model is a Transformer with data augmentation fine-tuned on reaction data after being pre-trained on unlabelled data.

**Graph2SMILES.** The Graph2SMILES[175] model has a graph-to-sequence architecture which combines a GNN encoder with a Transformer decoder. The

product is represented by molecular graphs and translated to reactant SMILES strings, eliminating the need for SMILES augmentation.

The benchmark dataset for single-step retrosynthesis is the USPTO-50K dataset consisting of approximately 50,000 reactions across ten reaction classes. The USPTO-50K is a high-quality dataset with accurate atom mappings between the reactants and products. Single-step retrosynthesis was performed in two settings: with and without prior knowledge of the reaction class. Tables 3.4 and 3.5 show the top-n exact match accuracy results of the single-step retrosynthesis models on the USPTO-50K benchmark. The accuracy results are taken directly from the references. The training, validation, and test sets may differ slightly between models. Zhong et al. provide a comprehensive review and detailed quantitative analysis of single-step retrosynthesis methods on several publicly available datasets.[88]

Table 3.4: Top-$n$ Accuracy of the Single-Step Retrosynthesis Models on the USPTO-50K Dataset with Reaction Class Unknown*

| Method | Model | Reference | Accuracies (%) | | | |
|---|---|---|---|---|---|---|
| | | | Top-1 | Top-3 | Top-5 | Top-10 |
| Template-based | Retrosim | [128] | 37.3 | 54.7 | 63.3 | 74.1 |
| | Neuralsym | [105] | 44.4 | 65.3 | 72.4 | 78.9 |
| | GLN | [75] | 52.5 | 69.0 | 75.6 | 83.7 |
| | LocalRetro | [76] | **53.4** | **77.5** | **85.9** | **92.4** |
| Semi-template | G2Gs | [77] | 48.9 | 67.6 | 72.5 | 75.5 |
| | RetroXpert | [158,170] | 50.4 | 61.1 | 62.3 | 63.4 |
| | GraphRetro | [78] | 53.7 | 68.3 | 72.2 | 75.5 |
| | RetroPrime | [156] | 51.4 | 70.8 | 74.0 | 76.1 |
| | MEGAN | [171] | 48.1 | 70.7 | 78.4 | 86.1 |
| | *R-SMILES* | [172] | 49.1 | 68.4 | 75.8 | 82.2 |
| | Graph2Edits | [160] | **55.1** | **77.3** | **83.4** | **89.4** |
| Machine-translation | Seq2Seq | [127] | 37.4 | 52.4 | 57.0 | 61.7 |
| | MT | [116] | 43.8 | 60.5 | - | - |
| | AT | [82] | 53.5 | - | 81.0 | 85.7 |
| | SCROP | [166] | 43.7 | 60.0 | 65.2 | 68.7 |
| | GTA | [173] | 51.1 | 67.6 | 74.8 | 81.6 |
| | Dual-TF | [174] | 53.6 | 70.7 | 74.6 | 77.0 |
| | Tied-TF | [165] | 47.1 | 67.2 | 73.5 | 78.5 |
| | R-SMILES | [172] | **56.3** | **79.2** | **86.2** | **91.0** |
| | Graph2SMILES | [175] | 52.9 | 66.5 | 70.0 | 72.9 |

LocalRetro surpasses the other template-based models with and without the reaction class predefined except for the top-1 accuracy with the reaction class predefined. The top-1 accuracy of the LocalRetro without reaction class known is 53.4%. For the top-3, top-5 and top-10 accuracy, LocalRetro outperforms

Table 3.5: Top-$n$ Accuracy of the Single-Step Retrosynthesis Models on the USPTO-50K Dataset with Reaction Class Known*

| Method | Model | Reference | Accuracies (%) | | | |
|---|---|---|---|---|---|---|
| | | | Top-1 | Top-3 | Top-5 | Top-10 |
| Template-based | Retrosim | [128] | 52.9 | 73.8 | 81.2 | 88.1 |
| | Neuralsym | [105] | 44.4 | 65.3 | 72.4 | 78.9 |
| | GLN | [75] | **64.2** | 79.1 | 85.2 | 90.0 |
| | LocalRetro | [76] | 63.9 | **86.8** | **92.4** | **96.3** |
| Semi-template | G2Gs | [77] | 61.0 | 81.3 | 86.0 | 88.7 |
| | RetroXpert | [158,170] | 62.1 | 75.8 | 78.5 | 80.9 |
| | GraphRetro | [78] | 63.9 | 81.5 | 85.2 | 88.1 |
| | RetroPrime | [156] | 64.8 | 81.6 | 85.0 | 86.9 |
| | MEGAN | [171] | 60.7 | 82.0 | 87.5 | 91.6 |
| | *R-SMILES* | [172] | - | - | - | - |
| | Graph2Edits | [160] | **67.1** | **87.5** | **91.5** | **93.8** |
| Machine-translation | Seq2Seq | [127] | - | - | - | - |
| | MT | [116] | - | - | - | - |
| | AT | [82] | - | - | - | - |
| | SCROP | [166] | 59.0 | 74.8 | 78.1 | 81.1 |
| | GTA | [173] | - | - | - | - |
| | Dual-TF | [174] | 65.7 | 81.9 | 84.7 | 85.9 |
| | Tied-TF | [165] | - | - | - | - |
| | R-SMILES | [172] | - | - | - | - |
| | Graph2SMILES | [175] | - | - | - | - |

the other template-based models with at least 6.3% and 8.5% margins with and without reaction class predefined, respectively.

Graph2Edits achieves a top-1 accuracy of 55.1% and 67.1% when the reaction class is unknown and known, respectively. Graph2Edits reaches state-of-the-art performance for semi-template-based methods, outperforming the other models by at least a 3.3% margin for larger $n$ values ($n = 3, 5, 10$) when the reaction class is unknown. Although Graph2Edits surpasses template-based LocalRetro in top-1 accuracy, it falls short in larger $n$ values.

The top-$n$ accuracy of the models, when the reaction class is known, will not be discussed as only two models reported this information. When the reaction class is unknown, R-SMILES outperforms the current best product-to-reactant machine translation methods by 2.7%, 8.5%, 5.2%, and 5.3% in top-1, top-3, top-5, and top-10 accuracy. R-SMILES is the state-of-the-art in single-step retrosynthesis, with a top-1 accuracy of 56.3%, which is 2.9% and 1.2% higher than the top template-based and semi-template-based methods.

A drawback to the product-to-reactant machine translation methods is the pre-

diction of invalid SMILES strings. The standard transformer architecture is prone to predicting incorrect SMILES strings as it does not learn the grammar of the SMILES notation. Although predicted SMILES strings may be grammatically incorrect, the grammatically correct version could be the ground truth. The SCROP model incorporates a syntax corrector, which reduces the SMILES invalidity rate but does not significantly improve the accuracy.[166] Data augmentation is now employed to prevent SMILES invalidity at an additional computational cost.[82]

### 3.4.2 Multi-Step Retrosynthesis

A target molecule is broken down recursively until the precursors are commercially available in multi-step retrosynthesis. In the pharmaceutical industry, the average synthetic route is 8.1 steps.[176] This vast reaction search space makes efficient searching and planning of multi-step syntheses challenging. The multi-step retrosynthesis task can be represented as a synthesis tree or directed acyclic graph, with the target molecule at the root. A branching factor and depth restrict the synthesis tree. The branching factor specifies the number of possible steps from a particular molecule, while the depth is the maximum number of steps before termination. An objective function guides the search and repeats until termination. Pathway termination will occur if the potential precursors are commercially available or a predefined depth constraint is reached. For retrosynthesis, the branching factor is usually high and the depth low.

The multi-step planning framework contains three phases: selection, expansion, and update. A selection policy identifies the most promising nodes to expand, i.e., the most promising molecules to propose reactants for. Selection can be based on heuristics or a node value function. An expansion policy applies a pre-trained single-step retrosynthesis model to the selected nodes. The relevant values along the path are then updated.

The number of potential synthetic routes grows exponentially as the number of steps in a pathway increases. Multi-step retrosynthesis tools require intelligent algorithms to identify the most promising branches to avoid exhaustive calculations of all combinations. Unfortunately, choosing the optimum predicted reaction at each step may not result in the most efficient pathway. Until the total cost of the synthetic route is calculated, the effect of each decision is unknown. The cost could be related to the price of starting materials, number of steps, waste generated, greenness, ease of product purification, sustainability, safety hazards, or environmental hazards.

*Top-$n$ accuracy values are taken directly from the corresponding reference. Highest top-$n$ accuracies are highlighted in **bold**.

There are limited assessment criteria for multi-step retrosynthesis. A double-blind AB test comparison of proposed and reported routes is valuable but time-consuming and laborious. Quantitative assessment metrics include the success rate of different iterations and the number of iterations, reaction nodes, and molecule nodes. PaRoutes is a benchmarking framework for comparing multi-step retrosynthesis methods.[177] It consists of two sets of patent-extracted routes, a list of available compounds, and a collection of reactions for training single-step models. It evaluates performance based on route quality, search speed, and route diversity. PaRoutes offers an unbiased assessment and comprehension of subtle differences in state-of-the-art approaches. Although the framework is currently limited to template-based methods, work on including template-free methods is ongoing.

### Monte Carlo Tree Search

Inspired by game AI, the well-known approach published by Segler and Waller[106] combines the Neural-Symbolic (NeuralSym)[105] single-step model with Monte Carlo Tree Search (MCTS).[178] Three neural networks incorporated into the MCTS algorithm comprise the 3N-MCTS model. The neural networks are the expansion policy, in-scope filter and rollout phase. MCTS is a heuristic search algorithm that selects the best-unexpanded node, expands it with a template-based single-step model, filters the results, evaluates the new molecules in a rollout phase, and updates the scores along the pathway. Based on current position values, the selection policy chooses the most promising node, balancing exploitation (highest-scoring nodes) and exploration (unvisited nodes).[179] The section policy employed is a version of the one used in AlphaGO.[18] The expansion policy, a single-step, template-based retrosynthesis model, expands the selected node. An in-scope filter classifies the reactions as feasible or not to remove any unfeasible chemical reactions. The rollout phase then evaluates the new nodes (precursors) by iteratively applying a comparable lightweight single-step model. Based on the difficulty of the synthesis, reward values are assigned to molecules and utilised to update the tree position values. While the 3N-MCTS model can account for stereochemistry, it does not consider the prediction of reaction conditions in pathway generation. ASKCOS,[138] AiZynthFinder,[140] and AutoSynRoute[180] adopt the MCTS technique in their implementations. AutoSynRoute combines MCTS with a template-free Transformer model.

### Commercial Software

Most commercial software are closed-source meaning the algorithms, reaction data, and reaction templates are not publicly available. Due to the limited information, the focus will be on real-life applications of the methods. Commercially available synthesis planning software includes InfoChem's ICSYNTH,[137] Wiley's ChemPlanner™ (formerly ARChem Route Designer)[148] that has integrated into CAS' SciFinder$^n$ and Merck KGaA's Synthia™ (formerly Chematica).[91,144,147,181–184] These applications are based on reaction templates. Synthia™ contains a template library of over 100,000 reaction templates hand-coded by expert synthetic chemists. The templates encode possible incompatibility groups, protecting groups, and reaction conditions. ICSYNTH and ChemPlanner™ automatically extracted templates from reaction databases, consisting of the reaction centre and neighbouring atom/groups. ICSYNTH also provides tools to generate template libraries from in-house data, which can be used alone or in conjunction with the supplied libraries.

Scoring functions based on user-defined criteria, evaluate each synthetic step and direct the search to commercially available substrates. Synthia™ and ICSYNTH filter the predictions to remove any unwanted structures. Synthia™ defines additional heuristics to penalise non-selective reactions, strained intermediates, and unlikely structural motifs. ICSYNTH compares the predicted structure to a predefined list of prohibited structures. ChemPlanner™ uses a similar method to eliminate functional group incompatibilities. Synthia™ and ChemPlanner™ account for regiochemistry and electronic effects, while Synthia™ can also handle stereochemistry and steric effects. Each program generates synthesis trees that include references to supporting literature and the cost of starting materials. Commercial multi-step synthesis planning software may have good validation scores, but chemists will remain sceptical until sufficient evidence supports success. The applicability of the software to drug-like targets and bioactive molecules has been the primary objective.

InfoChem, in collaboration with AstraZeneca, has evaluated the performance of ICSYNTH to act as an idea generator.[137] The ability of the program to predict routes to therapeutic targets, present in AstraZeneca's commercial drug projects or the literature, was assessed. Proposed pathways were compared to the literature, brainstormed proposals, and the in-house experience of chemists working on the projects. Without prior knowledge of the brainstorm proposals, ICSYNTH rediscovered known chemistries, returned brainstorm suggestions and provided new unreported synthetic routes. Due to its unbiased nature, it also identified an unconventional transformation that led to a non-intuitive solution to a prob-

lem.[137] ICSYNTH can generate ideas to complement synthetic chemists.

ChemPlanner™ experimentally validated proposed synthetic pathways to medically relevant targets.[185] These pathways were compared to routes developed by synthetic chemists. ChemPlanner™ identified almost all routes designed by synthetic chemists. Not all identified routes were considered the best due to the scoring functions within the program. ChemPlanner™ also proposed alternative paths with fewer steps and lower costs. ChemPlanner™ can assist synthetic chemists in designing synthetic pathways, saving time and money.

Synthia™ uses network theory, high-power computing, AI and expert chemical knowledge to design retrosynthetic pathways.[144] It can construct novel synthesis routes to medicinally and industrially relevant targets. The program found synthetic pathways to eight commercial bioactive substances and natural products.[144] These targets were selected as previous attempts at synthesis were low yielding, not scalable, or failed. Synthetic chemists experimentally verified the predicted pathways. Synthia™ designed routes that increased reaction yield while saving time and money. It also successfully predicted a pathway to a target without known synthetic paths. An extension of Synthia™ enables one to avoid patented routes.[91] When constructing pathways to novel molecules, it is critical not to violate existing patented routes. A bond preservation approach identifies and prevents the disconnection of bonds essential to patents. To validate this technique, three commercial drugs, Linezolid, Sitagliptin, and Panobinostat, were evaluated by Synthia™. Pathways predicted with and without the patent bond constraint. Without constraints, Synthia™ proposed similar routes to the patents. However, by selecting the bonds that must not be broken, the program could navigate around them and identify alternative routes. Synthia recently demonstrated passing a Turing-like test, where synthetic routes designed by Synthia were indistinguishable from those designed by synthetic chemists.[184]

## 3.5 Forward Reaction Prediction

Synthetic pathways predicted by chemists or machines are not guaranteed to be experimentally feasible. Synthetic chemists assess the feasibility of synthetic routes by reviewing the literature for similar transformations. Following this, the chemical reactions are performed experimentally. This process requires experienced synthetic chemists and substantial time and money. Forward reaction prediction is an alternative technique for determining the plausibility of proposed chemical reactions. The aim is to predict the major product of a chemical reaction given the reactants, reagents, and occasionally reaction conditions.

This technique can also identify selectivity patterns and potentially harmful or difficult-to-separate side products or impurities.

In principle, there is only one correct answer in forward reaction prediction. In practice, however, minor variations in reaction conditions, such as solvent and temperature, can affect the major product and reaction yield.

This section compares the top-$n$ accuracy of several forward reaction prediction applications on the USPTO-MIT benchmark dataset. All top-$n$ accuracy values are taken directly from the literature. The current state-of-the-art in forward reaction prediction is template-free methods. The approaches compared in this section are the semi-template-based methods: WLDN, WLDN5, GTPN, Symbolic, MEGAN; and the machine-translation methods: Seq2Seq, Molecular Transformer, Augmented Transformer, Chemformer, R-SMILES. These reaction prediction models are outlined below.

**WLDN.** The Weisfeiler-Lehman Difference Network (WLDN)[122] is a semi-template-based method which predicts a series of graph edits. A Weisfeiler-Lehman Network (WLN) first identifies the reaction centre as pairwise atom interactions. Enumerating all feasible bond configurations between atoms in the reaction centre generates product candidates. A WLDN ranks the prospective products.

**WLDN5.** WLDN5[159] improves on the WLDN model by combining the reaction centre prediction and candidate ranking into a single task.

**GTPN.** The Graph Transformation Policy Network (GTPN)[186] is a semi-template-based method that uses reinforcement learning to determine the optimal sequence of bond changes to transform the reactants into products. A Graph Neural Network (GNN) is used to model the reactants. A Node Pair Prediction Network (NPPN) predicts a single change in connectivity. A Policy Network (PN) generates an intermediate graph as an input for the following step until it terminates. The final graph generated is the predicted product.

**Symbolic.** The Symbolic[187] method is a semi-template-based method which integrates deep neural networks with probabilistic and symbolic inference. A Graph Convolutional Network (GCN) predicts the likelihood of changes in connectivity which then govern a probability distribution over potential products. Integer Linear Programming (ILP) infers the most probable product candidate from the probability distribution. In ILP, heuristic constraints ensure that the products are chemically valid.

**MEGAN.** The Molecular Edit Graph Attention Network (MEGAN)[171] is a semi-template-based method. The model generates products by performing a sequence of graph edits, i.e. bond changes, to the reactants. A graph attention network modifies the reactants sequentially by generating intermediate substrates until it terminates and gives the predicted products.

**Seq2Seq** The Sequence-to-Sequence (Seq2Seq)[80] model is a machine translation method which maps reactant SMILES strings to product SMILES strings. Two RNNs and an attention mechanism comprise the model. The Long Short-Term Memory (LSTM) network, an RNN, is used as the encoder and the decoder.

**MT** The Molecular Transformer (MT)[81] is a machine translation method based on the attention Transformer architecture.[162] The reactant SMILES string are converted to product SMILES strings.

**AT** The Augmented Transformer (AT)[82] uses data augmentation of the SMILES string to decrease overfitting and improve the accuracy of the neural network Transformer architecture.

**Chemformer** The Chemformer[164] model is a pre-train-fine-tune Transformer-based model that uses transfer learning to improve convergence and accuracy. The model was pre-trained on a large dataset of unlabelled SMILES before being fine-tuned on the forward reaction prediction task.

**R-SMILES** The Root-aligned SMILES (R-SMILES)[172] specifies a tightly aligned one-to-one mapping between the reactant and product SMILES. R-SMILES uses the same starting atom (root) of the SMILES string of the reactants and products, decreasing the edit distance. The model is a Transformer with data augmentation fine-tuned on reaction data after being pre-trained on unlabelled data.

**Graph2SMILES** The Graph2SMILES[175] model has a graph-to-sequence architecture which combines a GNN encoder with a Transformer decoder. The reactants are represented by molecular graphs and translated to product SMILES strings, eliminating the need for SMILES augmentation.

**NERF** The Non-autoregressive Electron Redistribution Framework (NERF)[188] models the electron flow in reactants. Molecular graphs represent the reactants and products. A GNN encodes the reactant graphs. A decoder models the

electron movement probabilities of bond breaking and bond formation for each atom in the reactants. The electron movement probabilities are converted to bond changes, generating the products.

The popular benchmark dataset for reaction prediction is the USPTO-MIT dataset, composed of approximately 480,000 reactions.[122] Forward reaction prediction was performed in two settings: "separated" and "mixed". In USPTO-MIT-Separated, a separator token separates the reactants and reagents. In USPTO-MIT-Mixed, the reactants and reagents are not separated. Forward reaction prediction on the mixed dataset is more challenging than on the separated dataset since the model must distinguish between reactants and reagents.

Tables 3.6 and 3.7 illustrate the predictive accuracy of the reaction prediction models on the USPTO-MIT benchmark dataset. Not all models, notably semi-template models, are tested on the mixed dataset. When evaluating the models on both datasets, removing the distinction between reactants and reagents before evaluation reduces the accuracy. The transformer-based models tend to have higher accuracies than the semi-template methods. Implementing a fully attention-based Transformer model over the Seq2Seq method improves the accuracy by 10.1% to 7.8% from top-1 to top-5. Including data augmentation and pre-training further enhances the performance of the Transformer model by at least 1.6%.

Chemformer outperforms the other models in top-1 accuracy, establishing a state-of-the-art top-1 accuracy of 92.8% on the separated dataset and 91.3% on the mixed dataset. The top-1 accuracy of R-SMILES is marginally ($\leq 0.5\%$) lower than Chemformer. Except for this, R-SMILES obtains better top-2 and top-5 accuracy results.

Table 3.6: Top-$n$ Accuracy of the Reaction Prediction Models on the USPTO-MIT-Separated Dataset[†]

| Method | Model | Reference | Accuracies (%) | | | |
|---|---|---|---|---|---|---|
| | | | Top-1 | Top-2 | Top-3 | Top-5 |
| Semi-template | WLDN | [122] | 79.6 | - | 87.7 | 89.2 |
| | WLDN5 | [159] | 85.6 | 90.5 | 92.8 | 93.4 |
| | GTPN | [186] | 83.2 | - | 86.0 | 86.5 |
| | Symbolic | [187] | 90.4 | 93.2 | 94.1 | 95.0 |
| | MEGAN | [171] | 89.3 | 92.7 | 94.4 | 95.6 |
| | NERF | [188] | 90.7 | 92.3 | 93.3 | 93.7 |
| Machine-translation | Seq2Seq | [80] | 80.3 | 84.7 | 86.2 | 87.5 |
| | MT | [81] | 90.4 | 93.7 | **94.6** | 95.3 |
| | AT | [82] | 92.0 | 95.4 | - | 97.0 |
| | R-SMILES | [172] | 92.3 | **95.8** | - | **97.5** |
| | Chemformer | [164] | **92.8** | - | - | 94.9 |
| | Graph2SMILES | [175] | - | - | - | - |

Table 3.7: Top-$n$ Accuracy of the Reaction Prediction Models on the USPTO-MIT-Mixed Dataset[†]

| Method | Model | Reference | Accuracies (%) | | | |
|---|---|---|---|---|---|---|
| | | | Top-1 | Top-2 | Top-3 | Top-5 |
| Semi-template | WLDN | [122] | - | - | - | - |
| | WLDN5 | [159] | - | - | - | - |
| | GTPN | [186] | - | - | - | - |
| | Symbolic | [187] | - | - | - | - |
| | MEGAN | [171] | 86.3 | 90.3 | 92.4 | 94.0 |
| | NERF | [188] | - | - | - | - |
| Machine-translation | Seq2Seq | [80] | - | - | - | - |
| | MT | [81] | 88.6 | 92.4 | 93.5 | 94.9 |
| | AT | [82] | 90.6 | 94.4 | - | 96.1 |
| | R-SMILES | [172] | 91.0 | **95.0** | - | **96.8** |
| | Chemformer | [164] | **91.3** | - | - | 93.7 |
| | Graph2SMILES | [175] | 90.3 | - | **94.0** | 94.8 |

[†]Top-$n$ accuracy values are taken directly from the corresponding reference. Highest top-$n$ accuracies are highlighted in **bold**.

## 3.6 Future Outlook and Potential Challenges

Knowing whether a compound is synthesisable before executing chemical reactions is beneficial. CASP tools propose synthetic routes, ideally with high confidence. Contemporary approaches to retrosynthetic analysis and forward reaction prediction are classified as template-based or template-free. Recently the number of template-free models has surged due to higher coverage, scalability and diversity. Automated extraction of data and the ability to learn from historical reaction data has resulted in contemporary AI approaches outperforming traditional expert-based systems in terms of cost and efficiency. There are numerous benefits to incorporating CASP into synthetic chemists' daily routines. These include shortening synthesis planning timelines, reducing the number of steps in a pathway, lowering costs, and offering alternative, unconventional routes.

When designing paths to novel compounds, chemists are biased towards robust reactions.[189] In CASP programs, AI and machine learning reduce human bias in predictions, potentially expanding the toolbox of medicinal chemists. Despite their advantages and recent success, CASP tools are not routine. As the inclusion of reaction conditions increases computational complexity, proposed methods typically omit reaction conditions. Few attempts have integrated reaction conditions such as reagents, catalysts, solvents, and temperature.[39,190,191] Finding the shortest path drives multi-step retrosynthesis, which can be problematic. The models cannot replicate the literature route as they fail to account for protection and deprotection strategies. Underlying chemical reaction data, evaluation methodologies, and interpretability also limit the tools.

Chemical reaction data is a significant limitation of current methods. Existing databases are still insufficient in terms of volume and diversity. High-quality, reproducible reaction data is critical for advancing CASP tools. Benchmark datasets derived from the USPTO patent data suffer from prophetic examples, inaccuracies in atom mapping, noisy stereochemical data, and inconsistencies. Open-source benchmarks are required to ensure fair model comparison. The large benchmark datasets should incorporate counter-examples to enable algorithms to perform optimally. There is a push for open sharing of machine-readable reaction data, open-source code, and changes to publication standards to reflect this.

Rigorous validation methods to assess model generalisability are required to ensure that reported results are not misleading.[151,190] While top-$n$ accuracy is justified for reaction prediction, its use in single-step retrosynthesis is misleading. The top-$n$ accuracy examines whether a model can predict the ground truth. Compared to forward reaction prediction, retrosynthesis rarely originates from a

single set of precursors. Depending on the functional groups present in the target molecule, multiple disconnection sites may exist, leading to numerous valid pathways. Top-$n$ accuracy cannot determine whether alternative routes are feasible. Feasibility is determined experimentally. However, the cost, reaction yield and diversity are frequently overlooked in model development.

Poor model interpretability has created suspicion among synthetic chemists. There are four main aspects to interpretability: transparency, justification, informativeness, and uncertainty estimation.[192] These aspects enable synthetic chemists to comprehend why the model generated the predictions. They also assist computational chemists in determining the causes of poor model performance and guide model improvement. Template-based models rely on templates for interpretability, whereas semi-template models rely on synthons. Neither method of interpretation explains why a model made a particular prediction. Many CASP methods employ "black-box" algorithms, such as neural networks, non-linear support vector machines, or random forests. While "black-box" algorithms have high predictive power, their decisions and predictions are unexplainable. Current acceptance and confidence of "black-box" approaches rely on sufficient justification by chemical literature or experimentally determined results. Attributing the relationship between properties and structural fragments has improved the interpretability of molecular prediction models.[193–195] Similar methods implemented in CASP could improve interpretability.

CASP programs aim to assist synthetic chemists in the decision-making process for designing chemical pathways. We project CASP tools to become increasingly common in modern laboratories imminently. Synthesis planning models should improve when high-quality reaction data becomes available and algorithms advance. The future of automated reaction optimisation is the integration of robotics and CASP. These units would increase productivity in synthesis laboratories without replacing bench chemists. Early and strong engagement with synthetic chemists will facilitate more rapid development of CASP with broader acceptance and uptake from the community.

In this chapter, we reviewed retrosynthesis and forward reaction prediction tasks. We discussed reaction data sources and their progression. Contemporary state-of-the-art approaches to CASP were classified, outlined, and compared on benchmark patent data. Finally, we highlighted potential challenges and the outlook of this field.

# Chapter 4

# Machine Learning for Predicting Yields of Chemical Reactions

## 4.1 Introduction

The availability of large reaction datasets and high-performance computing have been key in the development of computer-aided chemistry.[35] For example, in molecular design,[196] retrosynthetic planning tools,[106,116,166,197,198] reaction prediction[80,116,159], and the optimisation of reaction conditions.[108,199,200] Whilst the prediction of biological activities and molecular properties using Quantitative Structure-Activity Relationship (QSAR) or Quantitative Structure-Property Relationship (QSPR) models have been well-studied,[29,201] reactivity prediction, has been explored much less. This is largely due to a lack of appropriately curated data, for example, on reaction yield and Enantiomeric Excess (%ee). Performing a large number of experimental reactions is expensive, time-consuming, resource-consuming and requires synthetic chemists. High-throughput chemistry, along with batch and flow systems, have recently opened opportunities to generate reaction data for use in machine learning.[32,202,203] In Chapter 4 and Chapter 5, we focus on developing machine learning models to predict reaction yield using a high-throughput reaction dataset. This research aims to determine whether models built using structure-based descriptors have comparable performance metrics to those constructed using calculated properties. This chapter discusses the pioneering work in predicting reaction yield undertaken by the Doyle group[32–34] before outlining our approach and detailing preliminary work.

A dataset consisting of chemical structures or reactions must be converted to a machine-readable format before it is presented to a machine learning algorithm.

Molecular descriptors are based on the structural, physiochemical, electronic, or topological nature of molecules. Quantum chemical descriptors are common for the prediction of chemical reactivity.[32,204–206] They have also been used to build kernel-based QSAR and QSPR models, employing the Gaussian Radial Basis Function (RBF) kernel.[207–209] Site-specific, atomic properties including Nuclear Magnetic Resonance (NMR) shifts, vibrational frequencies, vibrational intensities and partial atomic charges have been used, along with global descriptors such as Highest Occupied Molecular Orbital (HOMO) energies, Lowest Unoccupied Molecular Orbital (LUMO) energies, dipole moment and polar surface area. Three-dimensional steric descriptors have been included in models of catalyst selectivity to improve predictions, by capturing important conformational information.[205,206] Quantum chemical descriptors are typically calculated using Density Functional Theory (DFT), which can be computationally demanding. Therefore, quantum chemical descriptors may not always be appropriate for large datasets, particularly if the dataset contains large molecules. Site-specific descriptors calculated for a reaction dataset require shared structural features for each reaction component.[32,204,205] For example, when calculating the carbon NMR shift for a specific carbon atom in a reaction component, that atom is a requirement across the dataset. If a reaction dataset has a large variety of molecules for a single reaction component, there may only be a few key shared atoms. In this scenario, alternative representations are required, such as structure-based descriptors that can be calculated for all molecules.

Molecular fingerprints represent the two-dimensional topology of a molecule. Examples include Molecular ACCess Systems (MACCS) Keys[49], Morgan circular fingerprints[53], and RDKit (RDK) fingerprints[50]. They are fast and easy to calculate, making them a popular choice for representing molecules. They are established in machine learning for virtual screening[210] and have emerged in the prediction of reaction conditions.[108,199] Sandfort *et al.*[211] have shown that two-dimensional, structure-based molecular fingerprints can achieve similar accuracy to quantum chemical descriptors in the prediction of chemical reactivity. Reactions were represented by a concatenation of Multiple Fingerprint Features (MFFs) and were used to build random forest models to predict reaction yields and %ee.[211] Fingerprints have also been utilised in kernel-based QSAR/QSPR relationship models, using the Tanimoto or RBF kernel.[212–214]

Labelled molecular graphs are another two-dimensional representation that depict the connectivity of a set of nodes, labelled with atom type, by a set of edges that are labelled by the bond order. From herein, we refer to labelled molecular graphs as molecular graphs. The global molecular structure is considered, in contrast to

the local environments in fingerprints. The kernel trick can be applied to molecular graphs to build machine learning models based on kernel methods, including Support Vector Machine (SVM).[215] Kriege *et al.*.[216] give a detailed overview of graph kernels and provide guidelines to aid researchers in the identification of successful kernels for different applications. The Weisfeiler-Lehman (WL)[55] algorithm is well-established in the field of Computer-Aided Synthesis Planning (CASP). The WL algorithm has been embedded in a neural network[122,217] and applied to the prediction of chemical reactivity.[122,159] Molecular graphs have been used in combination with deep learning to generate graph convolutional network models for reaction prediction,[159] retrosynthetic route design[197] and the prediction of reaction conditions.[218]

The prediction of reaction yields and enantiomeric excess are multidimensional problems as reaction outcomes depend on multiple reaction parameters, including both categorical and continuous variables. Minor changes in the reaction conditions such as catalyst(s), reagent(s), solvent(s), as well as temperature and pressure, can result in radically different reaction outcomes or possibly failed reactions. Even with expert synthetic chemists' chemical intuition and experience, chemical reactivity and reaction outcomes can be challenging to anticipate. High-throughput experimentation enables the screening of multiple discrete reaction variables (catalysts, reagents, solvents) on a nanomolar scale.[219,220] A matrix of parallel reactions is performed on a plate at the desired temperature and pressure, with the same reaction time. The samples in each well are analysed using Liquid Chromatography-Mass Spectrometry (LCMS) or Gas Chromatography-Mass Spectrometry (GCMS). There are challenges associated with such high throughput chemistry. These include the handling of very small volumes of liquid and evaporative solvent loss due to the use of volatile organics and solubility. The technique has proved useful for the optimisation of reaction conditions, as well as the discovery of new chemical reactivity in the pharmaceutical industry and academia.[219,220] It is also a lower-cost alternative for generating reaction data with which to build machine learning models.[32,206,211,221]

In this study, we investigate the use of structure-based descriptors in developing machine learning models to predict reaction yield. Structure-based descriptors are derived from the topology of molecules. They are simple and quick to compute and are applicable to any molecule. Quantum chemical descriptors are atomic, molecular, and vibrational properties calculated using DFT. In contrast, they can be computationally demanding and are not applicable to every molecule. The site-specific descriptors require key shared atoms across the dataset. Models that employ structure-based descriptors can predict a broader spectrum of reactants,

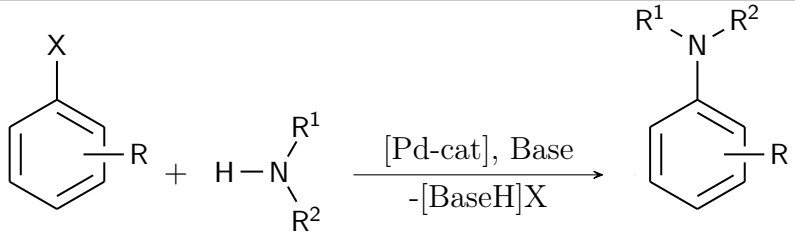
and the descriptors are generated on a much quicker timescale. A comparison of these approaches is demonstrated herein. We discuss the pioneering work of Doyle *et al.*,[32,34] who developed a random forest model constructed using quantum chemical descriptors to predict reaction yield. This pioneering work was facilitated by the generation of a high-throughput dataset which included reaction yield. The reactions in the dataset were Buchwald-Hartwig reactions, which we describe thoroughly, including mechanistic detail and the role of the catalyst. We outline a broad overview of our approach to predicting the yield of chemical reactions. Our preliminary work covers two main objectives. Firstly, we determine the optimum parameters for the structure-based descriptors. Secondly, we compare machine learning algorithms and identify which is the most promising to investigate further using more rigorous testing.

In 2018, Doyle *et al.*[32] reported an open-source combinatorial dataset which included reaction yields. The dataset contained approximately 4600 Buchwald-Hartwig amination reactions. We describe the Buchwald-Hartwig reaction, including mechanistic detail and the role of the catalyst. In the pioneering work by the Doyle group, the reactions in this dataset were represented by chemical properties and used to build machine learning models to predict reaction yield.[32]

### 4.1.1 Buchwald-Hartwig Amination Reaction

The Buchwald-Hartwig reaction (Scheme 4.1) is a well-established methodology for the formation of $sp^2$ carbon-nitrogen bonds. This type of palladium catalysed C-N cross-coupling of amines and aryl halides has attracted particular attention, due to its wide application in the pharmaceutical and agrochemical industries.[222–224]

**Scheme 4.1** General reaction scheme of the Buchwald-Hartwig amination reaction.

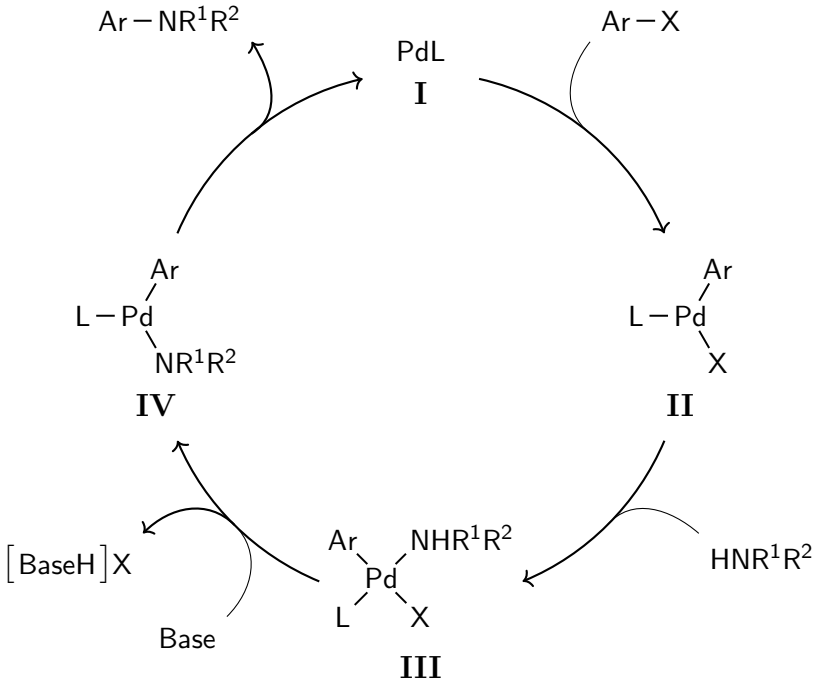



Palladium catalysed cross-coupling reactions between aryl bromines and aminostannanes were pioneered by Migita *et al.*[225] in 1983. The limitations of forming aryl amines using aminostannanes include narrow substrate scope and the use of toxic aminostannanes. These limitations were overcome by Buchwald and

Hartwig, who in 1995 reported a tin-free palladium catalysed coupling of aryl bromides and amines.[226,227] This initial approach, along with the subsequent work by the two research groups on the methodology and mechanistic understanding, led to the Buchwald-Hartwig amination reaction (Scheme 4.1) which has broadened the substrate scope and improved scalability.

The mechanism of the Buchwald-Hartwig cross-coupling reaction is well understood and the general scheme is shown in Scheme 4.2.[1] The palladium catalyst $Pd^0$ (species **I**) is initially inserted into the aryl halide by oxidative addition, forming species **II**. The amine coordinates to the $Pd^{II}$ via ligand exchange (species **III**). The resulting increase in acidity allows for the deprotonation of the amine by a hindered base, to form a palladium amine complex (species **IV**). The reductive amination of this complex yields the aryl amine product and the regenerated palladium catalyst (species **I**). A β-hydride elimination side reaction can compete with reductive amination to give a hydrodehalogenated arene and an imine by-product.

**Scheme 4.2** General reaction mechanism of the Buchwald-Hartwig amination reaction




Understanding the mechanistic detail of the Buchwald-Hartwig reaction has led to the development of multiple catalytic systems which have improved the reactivity and scope. Early catalytic systems based on monodentate ligands, such as P(*o*-tolyl)$_3$, required high temperatures and had limited substrate scope. The use of biphosphine ligands improved the substrate scope to include the coupling of primary amines, aryl chlorides and aryl triflates. The success of coupling primary

amines is partially owing to the preference of the reductive elimination reaction over the competitive $\beta$-hydride elimination. This is due to the chelating effect of the bidentate ligands to the palladium atom. The application of sterically hindered biaryl phosphine ligands improved the substrate scope further to include additional aryl pseudohalides (sulfonates and mesylates), heteroaryl halides and a larger range of aryl chlorides. A number of reactions containing the sterically hindered monodentate ligands can be performed under milder conditions. The commercially available CyPF-$^t$Bu bidentate ligand and biaryl phosphine ligands BrettPhos and RuPhos are considered standard.[228] Buchwald has published a guide to aid the selection of reaction conditions and dialkylbiaryl phosphines ligands.[229]

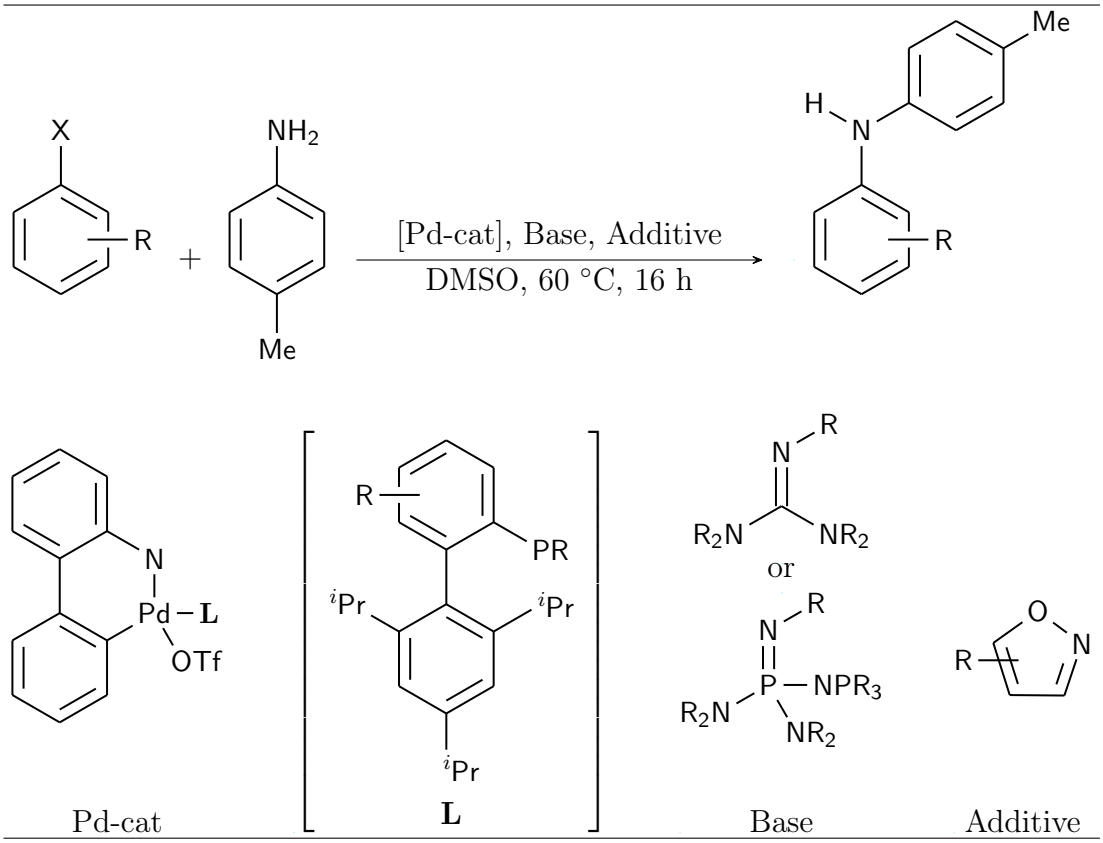
### 4.1.2 Pioneering Work on the Prediction of Reaction Yield

An open-source combinatorial dataset, including reaction yields, was reported by Doyle *et al.*[32] The experiments were performed on three 1536-well high-throughput plates with the use of the Mosquito robot. The dataset contains a set of Buchwald-Hartwig amination reactions between 4-methylanaline and 15 aryl/heteroaryl halides (Scheme 4.3). The reactions varied in three hindered bases and four monophosphine catalyst ligands.

Aniline products are important building blocks for the synthesis of small drug-like molecules.[230] This key transformation can however be limited if the substrates contain a five-membered ring with a heteroatom-heteroatom bond. Despite the drug-like characteristics of such heterocycles, for example, isoxazoles, they are not common in approved pharmaceuticals.[230] Doyle *et al.* assessed the effect of compounds containing isoxazole heterocycles on the reaction performance. Glorius developed an approach to identify catalysis-inhibiting sub-structures by deliberately adding representative fragments to the catalytic mixture.[231] Using this methodology, Doyle *et al.* added a selection of potentially inhibitory isoxazole additives to the Buchwald-Hartwig reactions. This allowed assessment of the additive's effect on the reaction performance, without the need to synthesise and isolate isoxazole (or other) containing aryl halides as a prior step to performing the coupling reactions. A total of 23 isoxazole additives were investigated.

All possible combinations of the 15 aryl halides, four ligands, three bases, 23 additives, aryl halide control and additive control, formed a total of 4608 reactions. Doyle et al. used this data to build machine-learning models to predict reaction yield. The reactions were represented using molecular, atomic and vibrational

**Scheme 4.3** Buchwald-Hartwig reactions performed by Doyle *et al.*, including generic structures of the palladium catalyst, base and additive. Full list of structures can be found in (Figure B.1 to B.4).




calculated properties. These quantum chemical descriptors were calculated using density functional theory. A variety of linear and non-linear regression models were evaluated with a 70-30% train-test split of the data. The random forest model exhibited the best performance metrics in predicting reaction yield, Coefficient of Determination ($R^2$) equal to 0.92 and Root Mean Squared Error (RMSE) equal to 7.8%.

Datasets with combinatorial structure have an intrinsic pattern (i.e., the presence or absence of molecules) which can lead to large variations in the performance of a model, depending on the train-test split of the data.[232] By splitting the data randomly, the reaction components in the test reactions will also be present in different training reactions. This type of in-sample test, where descriptors of molecules in the test reactions are already observed in training, can result in an unreliable representation of model generalisability. Models may fit the pattern of the data, rather than the relationship between chemically meaningful descriptors and the observed data. These models would therefore struggle when extrapolating to unseen chemical entities.

One-hot encodings[33] can be used as a baseline descriptor to validate model perfor-

mance and reveal potential patterns within the training data that may be fitted by models built on chemically meaningful descriptors. The one-hot encoding of a reaction simply denotes the presence or absence of each molecule as a vector and encapsulates no information beyond this. For the same random 70-30% train-test split of the Buchwald-Hartwig data, Chuang and Keiser showed models built on one-hot encodings exhibited near identical performance to the models built on quantum chemical descriptors ($R^2 = 0.90$ and RMSE = 8.6%).[33]

The 70-30% train-test split of the data is an example of a hold-out cross-validation test. In hold-out cross-validation, the data is only split once, and hence only a single model is evaluated. Often this is considered a bottleneck as the performance metrics will be dependent on the data points that reside in the training and test sets. This causes high variance in performance as the performance may differ if another division was made. The variance can be reduced by using $k$-fold cross-validation instead. This involves splitting the data into $k$ subsets and building a separate model using each subset as the test set and the remaining subsets as the training data. The average performance of the models is calculated and used for model evaluation. By training and testing on multiple splits of the data, $k$-fold cross-validation gives a more stable and reliable indication of performance. Although cross-validation may overestimate model generalisability, it is a useful technique for the selection of appropriate models or the optimisation of model and descriptor parameters.

In this chapter, preliminary $k$-fold cross-validation on the Doyle *et al.* dataset is performed. A selection of linear, tree-based and Support Vector Regression (SVR) machine learning models are built on quantum chemical descriptors and two types of structure-based descriptors: molecular fingerprints and molecular graphs. Structure-based descriptors are applicable to a wider range of molecules and are less computationally demanding than quantum chemical descriptors. The performances of the machine learning models are used to identify optimum parameters of the molecular descriptors and select appropriate models for the task of predicting reaction yield.

## 4.2 Computational Methods

### 4.2.1 Dataset

The data used in this study were 4608 single-step reactions reported by Doyle *et al.*[32] This open-access dataset contains the reactants, products, reaction conditions and yields of a single reaction class, the Buchwald-Hartwig amination

reaction (Scheme 4.3). The reactions varied in 23 isoxazole additives, 15 aryl halides, three bases and four Buchwald ligands (Figure B.1 to B.4). The data was generated using ultra-high-throughput experimentation in three 1536-well plates, giving a full matrix of reaction components including controls. Once the control reactions and reactions containing additive seven were removed, a total of 3955 reactions remained. Additive seven was removed as quantum chemical descriptors could not be calculated;[32] see Section 4.2.2 for details. The names of the aryl halide, additive, base and ligand in each reaction were converted to Simplified Molecular-Input Line-Entry System (SMILES) strings.[95] This was completed using the Computer-Aided Drug Design Group of the National Cancer Institute (NCI/CADD) Chemical Identifier Resolver Application Programming Interface (API)[233] except for a few unrecognised names, which were drawn and converted to SMILES strings in ChemDraw.

### 4.2.2 Molecular Descriptors and Preprocessing

A total of five molecular descriptors were evaluated in this study (Table 4.1). The focus of this work was to assess the generalisability of machine learning models built on structure-based descriptors. Three varieties were considered for their simplicity, ease of calculation and broad applicability. These were concatenated molecular fingerprints, Tanimoto kernel descriptors derived from molecular fingerprints and WL kernel descriptors derived from molecular graphs. Quantum chemical descriptors were also considered to allow comparison with the pioneering work by Doyle *et al.*[32] One-hot encodings were used as a baseline descriptor due to the combinatorial nature of the dataset. Detailed descriptions of the descriptors can be found in Chapter 2, Section 2.5.

Table 4.1: Format and Notation of the Descriptors for a Single Reaction

| Descriptor | Additive | Aryl Halide | Base | Ligand |
|---|---|---|---|---|
| One-hot Encodings | $[O_{A_1} \cdots O_{A_n}$ | $O_{H_1} \cdots O_{H_n}$ | $O_{B_1} \cdots O_{B_n}$ | $O_{L_1} \cdots O_{L_n}]$ |
| Quantum Chemical | $[D_1^A \cdots D_{19}^A$ | $D_1^H \cdots D_{27}^H$ | $D_1^B \cdots D_{10}^B$ | $D_1^L \cdots D_{64}^L]$ |
| Fingerprints | $[\cdots 0\ 1 \cdots$ | $\cdots 0\ 1 \cdots$ | $\cdots 0\ 1 \cdots$ | $\cdots 0\ 1 \cdots]$ |
| Tanimoto Kernel | $[\ k_{T_A}\ ]$ | $[\ k_{T_H}\ ]$ | $[\ k_{T_B}\ ]$ | $[\ k_{T_L}\ ]$ |
| WL Kernel | $[\ k_{WL_A}\ ]$ | $[\ k_{WL_H}\ ]$ | $[\ k_{WL_B}\ ]$ | $[\ k_{WL_L}\ ]$ |

#### One-hot Encodings

One-hot encodings of chemical reactions are binary vectors that denote the presence (1) or absence (0) of each molecule in the training reactions. This is shown in Table 4.1, where $A_n$, $H_n$, $B_n$ and $L_n$ are the number of additives, aryl halides, bases and ligands present in the training reactions. One-hot encodings represent

the reactions without using chemically meaningful information and therefore by construction are not able to generalise to unseen chemical entities. Building machine learning models on one-hot encodings can reveal underlying patterns in combinatorial datasets and should be used as a validation method.

### Quantum Chemical Descriptors

A combination of calculated molecular, atomic and vibrational properties for the additive ($D^A$), aryl halide ($D^H$), base ($D^B$) and ligand ($D^L$) formed a set of quantum chemical descriptors for each reaction (Table 4.1). The Spartan '14 interface for the Q-Chem quantum chemical software package[234,235] was used to calculate 120 descriptors per reaction using the density functional B3LYP with the 6-31G(d) basis set.[236,237] A full list of the descriptors consisting of 19 additive, 27 aryl halide, 10 base and 64 ligand descriptors can be found in Appendix B, Section B.2. These quantum chemical descriptors for the dataset were calculated by Doyle *et al.*[32]. The descriptors were standardised by centring the data to have zero mean and scaling to unit variance. The mean and standard deviation were calculated on the training set and used to standardise both the training and test sets.

The molecular descriptors included molecular volume, surface area, ovality, molecular weight, $E_{HOMO}$, $E_{LUMO}$, electronegativity, hardness and dipole moment. The atomic descriptors, NMR shifts and electrostatic charge, were calculated for shared atoms in each reaction component (Figure 4.1). The shared atoms were predetermined by Doyle *et al*. The common molecular vibrational modes across the set of molecules for each reagent class were identified. This was accomplished by comparing the similarity of the molecular vibrations. For each vibrational mode, the rotated atomic movement data of the predefined shared atoms was extracted. The atomic movement data was multiplied by the atom's atomic mass to obtain weighted atomic movement data. The Pearson correlation coefficient was calculated using the weighted atomic movement data between two molecular vibrations, each from a different molecule. A correlation matrix of Pearson coefficients for every molecular vibration for a pair of molecules was constructed. Using the correlation matrix, Pearson coefficient values that were above 0.5, as well as the highest value in the row and column, were considered shared molecular vibrations if the vibrational frequency was above 500cm$^{-1}$. The vibrational frequencies and infrared transition intensities were calculated for the common modes.

The atomic and vibrational descriptors cannot be calculated for additive seven. In this molecule, the *C4 labelled atom is a nitrogen atom and is not shared with

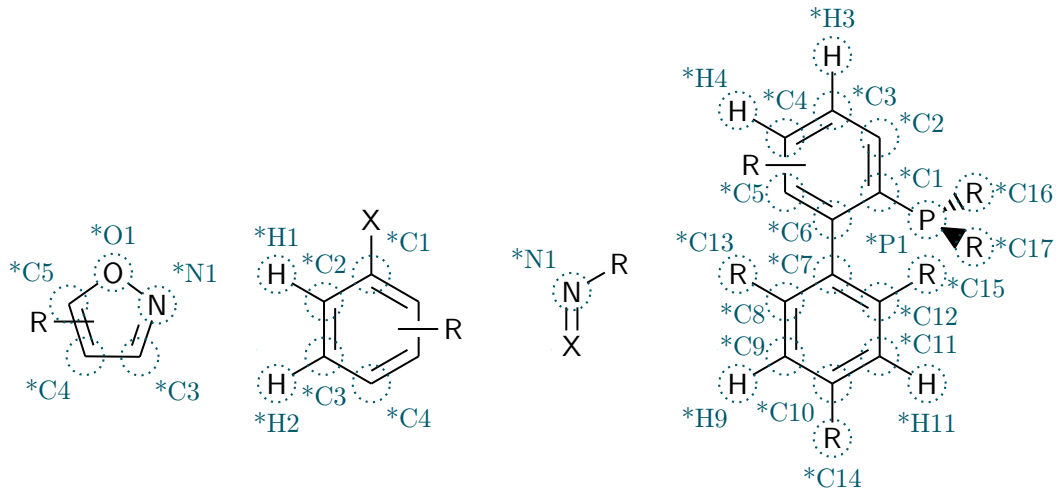



Figure 4.1: Shared atoms for each reaction component.

the other additives. Reactions containing additive seven were therefore removed from the dataset.

Large discrepancies in a few quantum chemical descriptors of 1-bromo-4-methoxybenzene were identified. The energies of the HOMO (−6.00eV) and LUMO (−0.35eV) reported by Doyle *et al.* were more negative than the other aryl halides (Table 4.2). The values of the electronegativity and hardness were both calculated from the HOMO and LUMO energies using the equations below.

$$\text{Electronegativity (eV)} = \frac{-(\ \text{E}_{HOMO}\ +\ \text{E}_{LUMO}\ )}{2}$$

$$\text{Hardness (eV)} = \frac{-(\ \text{E}_{HOMO}\ -\ \text{E}_{LUMO}\ )}{2}$$

The discrepancies therefore extended to these descriptors. The quantum chemical descriptors were recalculated for 1-bromo-4-methoxybenzene and gave HOMO (−0.22eV) and LUMO (−0.01eV) energies closer in value to the other aryl halides (Table 4.2). These results replaced the values reported by Doyle *et al.* in model development.

Table 4.2: Discrepancies in the Quantum Chemical Descriptors of 1-bromo-4--methoxybenzene

| Descriptor | Other Aryl Halides | 1-bromo-4-methoxybenzene | |
|---|---|---|---|
| | | Doyle et al.[32] | This Work |
| $\text{E}_{HOMO}$ (eV) | −0.2648 to −0.2176 | −6.0000 | −0.2204 |
| $\text{E}_{LUMO}$ (eV) | −0.0429 to −0.0104 | −0.3500 | −0.0128 |
| Electronegativity (eV) | 0.12 to 0.15 | 3.18 | 0.12 |
| Hardness (eV) | 0.10 to 0.11 | 2.82 | 0.10 |

### Molecular Fingerprints

Three types of molecular fingerprints were implemented using the RDKit package: MACCS Keys,[49] RDK fingerprints[50] and Morgan circular fingerprints.[53] Fingerprints are hashes, binary bit vectors, of a specified length. The bit length of the MACCS fingerprint is 167-bit, where the first bit is always zero and the remaining bits correspond to 166 public MACCS Keys. The first bit is zero to allow for the original numbering of the MACCS keys (1-166), due to zero indexing in Python. Bit lengths from 32 to 2048 were explored for the topological fingerprints: RDK fingerprints and Morgan circular fingerprints. The predefined path length (number of bonds) used in the RDK fingerprint was seven. Morgan and Feature Morgan (FMorgan) fingerprints were investigated with radii up to three.

**Concatenated Fingerprints** The fingerprint of the aryl halide, additive, base and ligand in each reaction was calculated. The fingerprints of the reaction components were used to generate fingerprint descriptors, by concatenating to form a single reaction fingerprint (Table 4.1).

**Tanimoto Kernel Descriptors** Tanimoto similarity scores were calculated between the fingerprints of molecules within the same reaction class, as implemented in RDKit. For two molecules in a single reaction class represented by molecular fingerprints ($F_{m_1}$ and $F_{m_2}$), the Tanimoto similarity[58,59] is defined as

$$k_T\left(F_{m_1}, F_{m_2}\right) = \frac{c}{a+b-c} \tag{4.1}$$

where $a$ and $b$ are the number of bits set in fingerprints $F_{m_1}$ and $F_{m_2}$, and $c$ is the number of bits set in common in $F_{m_1}$ and $F_{m_2}$. Although slight changes in the structure of small molecules can lead to substantial changes in the Tanimoto similarity, it is a very well-established measure and thus appropriate for us to consider. To calculate the Tanimoto kernel between two reactions ($R_x$, $R_{x'}$), the Hadamard product of the reaction component kernels was taken. This is shown in Equation 4.2, where $A_i$, $H_i$, $B_i$ and $L_i$ are the additive, aryl halide, base and ligand in reaction $i$.

$$k(R_x, R_{x'}) = k\left(A_x, A_{x'}\right) k\left(H_x, H_{x'}\right) k\left(B_x, B_{x'}\right) k\left(L_x, L_{x'}\right) \tag{4.2}$$

The training kernel is a symmetrical matrix generated using this method between all pairs of training reactions. The test kernel matrix is generated by calculating the Tanimoto kernel between the test reactions and training reactions. For a

single reaction ($R_x$) in the training or test set, the Tanimoto kernel (Table 4.1) is in the format

$$K_{R_x} = k_{A_x} k_{H_x} k_{L_x} k_{B_x} \tag{4.3}$$

where the Tanimoto kernel of each reaction component is

$$k_{T_{M_x}} = [k(M_x, M_1) \cdots k(M_x, M_n)]$$

$M$ is the reaction component class and $n$ is the number of training reactions.

#### Molecular Graphs

A molecular graph represents the topology of a molecule by a set of labelled nodes corresponding to the atoms, connected by a set of labelled edges corresponding to the bonds. From the SMILES string of each molecule in the dataset, the atomic symbol, the index of each atom, the bond order, the index of each bond and the adjacency matrix were obtained using RDKit.[50] This information was parsed to a module within GraKel to generate the molecular graph representation.[238]

**WL Kernel Descriptors** WL subtree graph kernels[55] were calculated for each reaction component using GraKel. The number of iterations, hyperparameter $h$ (also referred to as the WL depth), from two to ten were explored. The Hadamard product of reaction component kernels was calculated to give the WL reaction kernel as shown in Equation 4.2. The training and test kernel matrices were also generated using the same method as the Tanimoto kernel descriptors. For a single reaction, $R_x$, the format of the WL kernel (Table 4.1) is shown in Equation 4.3.

### 4.2.3 Machine Learning Models

Machine learning models relating descriptors to reaction yield were developed and implemented using scikit-learn.[239] A variety of linear, tree-based and support vector regression models were evaluated. Detailed descriptions of the machine learning algorithms can be found in Chapter 2, Section 2.3 and 2.4. The quantum chemical descriptors, concatenated molecular fingerprints and one-hot encodings can be used directly as an input to the machine learning models. The Tanimoto and WL kernel descriptors are in matrix form. These can be used directly as a precomputed kernel for the SVR models. The kernel-based descriptors cannot be used directly as an input to the linear and tree-based models. The features of each reaction must be extracted from the kernel matrices using a mathematical function for the linear and tree-based models. These features can then be given

to the models as an input.

The linear models explored were linear regression, Least Absolute Shrinkage and Selection Operator (LASSO), ridge, elastic net and Bayesian ridge. These assume a linear relationship between the molecular descriptors and the target (i.e., reaction yield). The tree-based models include decision tree, gradient boosting and random forest. The latter two are ensemble methods that build multiple decision trees. These models learn a series of rules from data features to predict the target values. SVR uses a kernel function to map input data to a higher dimensional feature space where regression is performed linearly. The kernel functions explored were linear, polynomial, RBF and sigmoid. All four kernels were applied to the quantum chemical, concatenated molecular fingerprint and one-hot encoding feature vectors, with the hyperparameters set to the values in Table 4.3, using scikit-learn. Although Kriege *et al.*[216] suggest there is little benefit in the combination of the WL kernel with non-linear kernels, we explored the WL kernel in combination with linear and non-linear kernels for completeness. Both kernel-based descriptors were used as a precomputed kernel as well as with the polynomial, RBF and sigmoid kernels applied to the individual entries of the kernel descriptor matrix. The hyperparameters of the kernels were tuned over the values in Table 4.3.

Table 4.3: Hyperparameters of the Kernel Functions

| Kernel | Hyperparameter | Condition | Descriptor | Values |
|---|---|---|---|---|
| Polynomial | $\gamma$ | $> 0$ | Non-kernel | $1.0/n_{features}$ |
| | | | Kernel | 1, 10, 100, 1000 |
| | $c$ | $\geq 0$ | Non-kernel | 1 |
| | | | Kernel | 1 |
| | $d$ | $> 0$ | Non-kernel | 3 |
| | | | Kernel | 3 |
| RBF | $\gamma$ | $> 0$ | Non-kernel | $1.0/n_{features}$ |
| | | | Kernel | 1, 10, 100, 1000 |
| Sigmoid | $\gamma$ | $> 0$ | Non-kernel | $1.0/n_{features}$ |
| | | | Kernel | 1, 10, 100, 1000 |
| | $c$ | $\geq 0$ | Non-kernel | 1 |
| | | | Kernel | 1 |

### 4.2.4 Model Building and Evaluation

A preliminary five-fold cross-validation test was performed on the dataset, using the following models: linear regression, LASSO, ridge, elastic net, Bayesian ridge, decision tree, gradient boosting, random forest and SVR (with linear, polynomial, RBF, sigmoid and precomputed kernels). The dataset was shuffled and split into

five groups. In turn, the models were trained on four-folds of the data and validated on the fifth. The average performance statistics of the five test sets were calculated for each machine learning model and used to identify the best combination of hyperparameters in Table 4.4 and 4.3. The optimum values for the bit length of the molecular fingerprints and the WL depth of the WL graph kernel were identified. The linear, tree-based and SVR models were compared to identify the best-performing model for further evaluation.

Table 4.4: Hyperparameter Grid

| Machine Learning Algorithm | Hyperparameter | Values |
|---|---|---|
| Support Vector Regression | C | 1, 10, 100, 1000 |
| | epsilon | 1, 5, 10 |
| Linear Regression | fit_intercept | True, False |
| Lasso | alpha | 1, $1\times10^{-1}$, $1\times10^{-2}$, $1\times10^{-3}$, $1\times10^{-4}$ |
| Ridge | alpha | 1, $1\times10^{-1}$, $1\times10^{-2}$, $1\times10^{-3}$, $1\times10^{-4}$ |
| Elastic Net | alpha | 0.01, 0.1, 0.2, 0.5 |
| Bayesian Ridge | alpha_1 | $1\times10^{-4}$, $1\times10^{-6}$, $1\times10^{-8}$ |
| | alpha_2 | $1\times10^{-4}$, $1\times10^{-6}$, $1\times10^{-8}$ |
| | lambda_1 | $1\times10^{-4}$, $1\times10^{-6}$, $1\times10^{-8}$ |
| | lambda_2 | $1\times10^{-4}$, $1\times10^{-6}$, $1\times10^{-8}$ |
| Decision Tree | N/A | N/A |
| Gradient Boosting | n_estimators | 250, 500, 750, 1000 |
| | learning_rate | 0.05, 0.1, 0.15, 0.2 |
| Random Forest | n_estimators | 250, 500, 750, 1000 |

The performances of the regression models were evaluated by $R^2$ and RMSE for data points outside of the training set. All analysis was performed using scikit-learn. Machine learning models built on one-hot encodings were used as a baseline, for comparison.

## 4.3 Results and Discussion

### 4.3.1 Parameter Optimisation of the Descriptors

Cross-validation was used to determine optimum parameters for the molecular descriptors. This included the bit length and radii (where applicable) of the molecular fingerprints, as well as the WL depth of the WL kernel descriptors.

#### Descriptors Derived from Molecular Fingerprints

The bit length of the Morgan, FMorgan and RDK molecular fingerprints can affect the performance of the models built on concatenated fingerprints and Tanimoto kernel descriptors. The bit length of the MACCS keys is not variable and

therefore was not considered in parameter optimisation. As bit length increases, there is a larger capacity for storing additional information about the molecule. It also decreases the darkness of the fingerprint, meaning a lower proportion of bits are set to ON (1) which decreases the probability of bit collisions. Cross-validation was used to identify the optimum bit length; values of 32, 64, 128, 256, 512, 1024 and 2048 were evaluated.

Generally, the bit length of the Morgan and FMorgan fingerprints, with radii of one to three, had little effect on the performance of the models (Figure 4.2 and 4.3; see Appendix B, Section B.3.1 for numeric details). The models built on FMorgan fingerprints had consistently poor predictive performance in cross-validation, with $R^2 \leq 0.51$ and RMSE$\geq$ 19.3%. Morgan fingerprints differ from FMorgan fingerprints in the initial encoding of the atoms. The Morgan fingerprints encode each atom's properties and connectivity, whereas FMorgan fingerprints generalise this information into roles of the atom, for example, whether the atom is a hydrogen-bond acceptor or donor, aromatic or a halogen. The FMorgan fingerprints cannot distinguish between halide atoms and therefore models built on these fingerprints would not recognise the lower reactivity of the aryl chlorides, nor the higher reactivity of the aryl iodides. Models built on descriptors derived from the FMorgan fingerprints were not able to capture the correlation between the molecules in the reactions and the reaction yield, thus were omitted from further analysis.

The predictive performance of the models built on the RDK fingerprint generally improved with bit length up to 512 (Figure 4.2 and 4.3). The models built on RDK fingerprints were the only models majorly affected by the bit length and hence determined the optimum bit length of the molecular fingerprints. Henceforth, a bit length of 512 was used in further testing of the models built on descriptors derived from molecular fingerprints.

The radius of the Morgan fingerprint defines the radius of the circular neighbourhood for each atom and hence the size of the subgraphs that are encoded in the fingerprint, see Chapter 2, Section 2.5 for more details. Radii from one to three were explored. The larger the radius, the more information about the molecule is encoded as the subgraphs will be larger. The Morgan fingerprints with a radius higher than one will contain all the fingerprint bits of lower radii. This also increases the fingerprint darkness and consequently increases the number of bit collisions. Variations in the radius caused minor effects on the performance of the models built on Morgan fingerprints with a bit length of 512 (Figure 4.4). Encoding more information in the Morgan fingerprints did not notably increase model performance, as a result only the smallest radius of neighbouring atoms

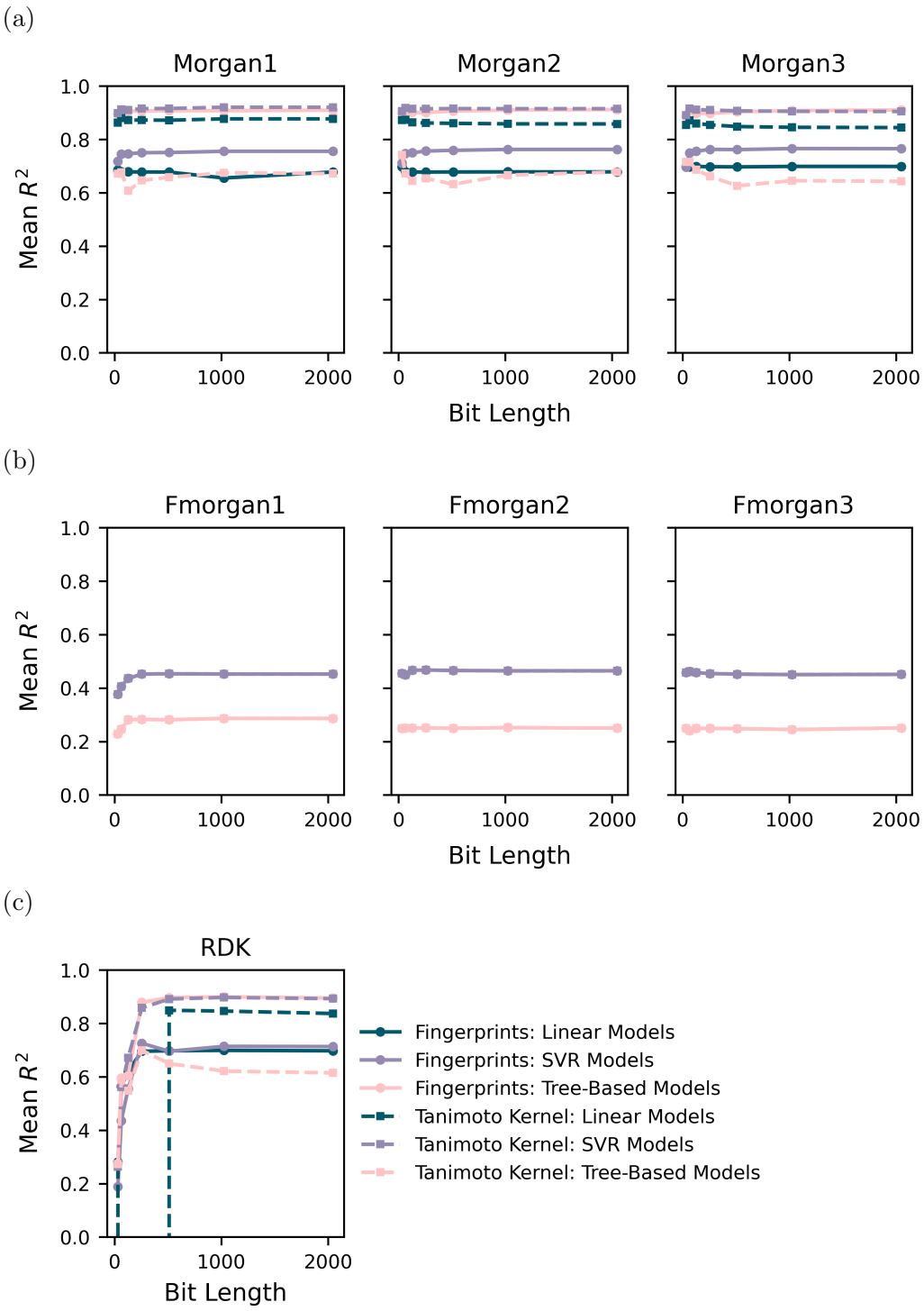



Figure 4.2: Average cross-validated Coefficient of Determination ($R^2$) of the linear models (dark blue), Support Vector Regression (SVR) models (purple) and tree-based models (pale pink) against the bit length of the molecular fingerprints: (a) Morgan, (b) Feature Morgan (FMorgan) and (c) RDK. Solid line, models built on concatenated fingerprints; dashed line, models built on Tanimoto kernel descriptors. See Appendix B, Section B.3.1 for numeric details.

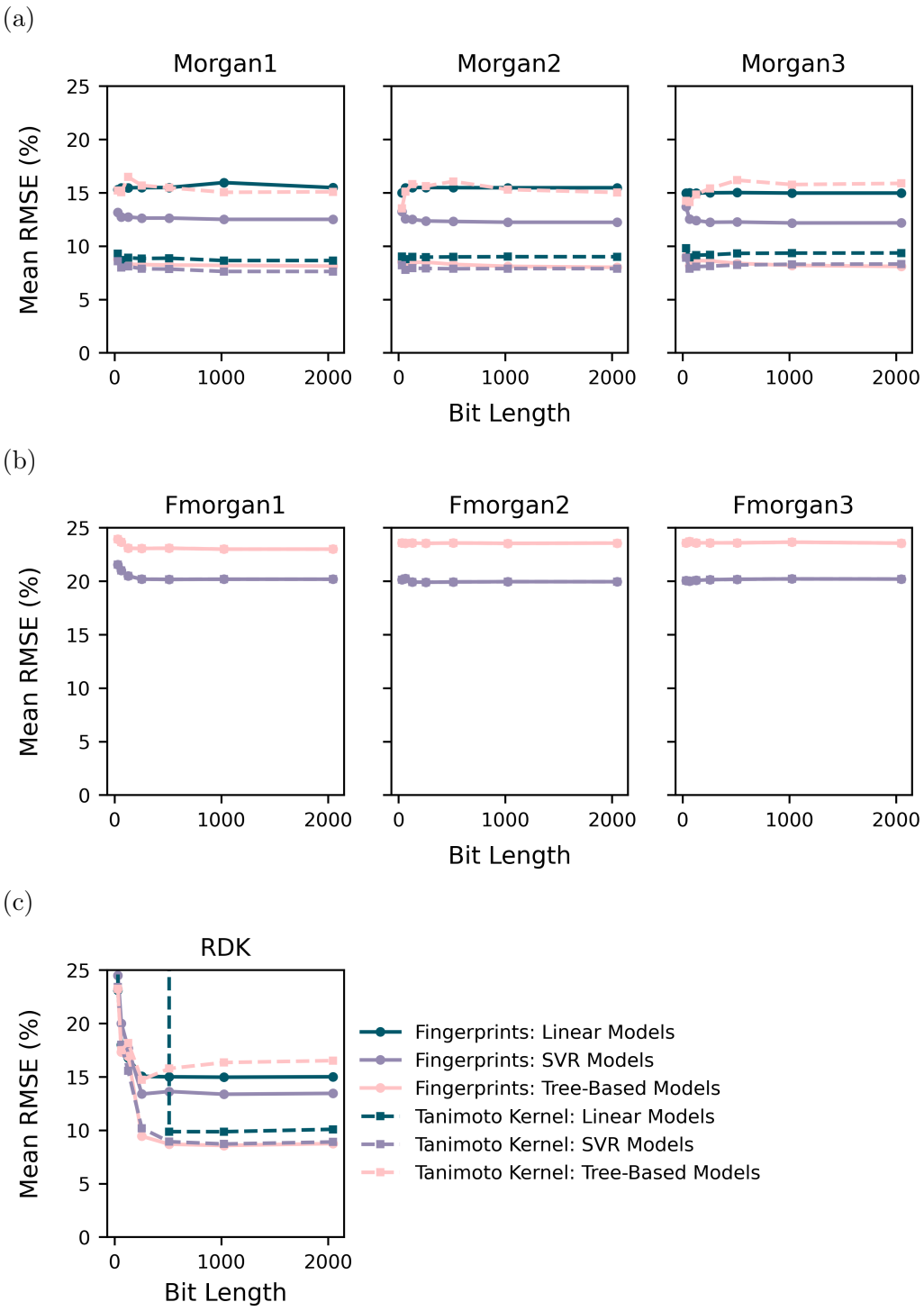



Figure 4.3: Average cross-validated Root Mean Squared Error (RMSE) of the linear models (dark blue), Support Vector Regression (SVR) models (purple) and tree-based models (pale pink) against the bit length of the molecular fingerprints: (a) Morgan, (b) Feature Morgan (FMorgan) and (c) RDK. Solid line, models built on concatenated fingerprints; dashed line, models built on Tanimoto kernel descriptors. see Appendix B, Section B.3.1 for numeric details.

(Morgan1) was considered for further analysis.

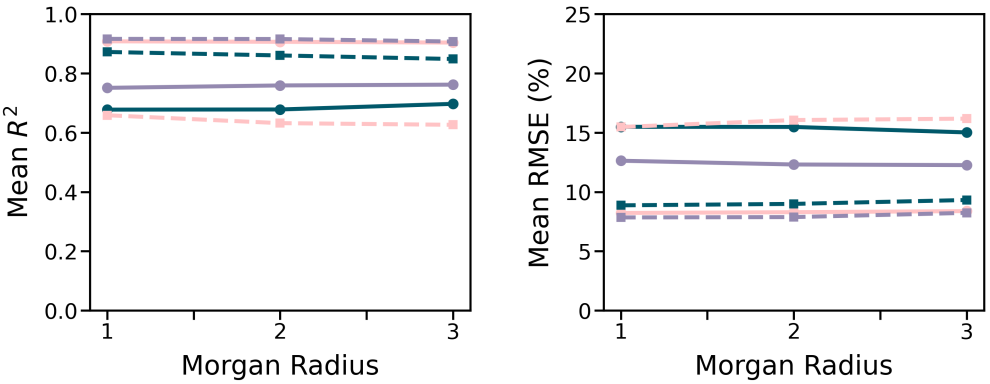



Figure 4.4: Average cross-validated performance of the linear models (dark blue), Support Vector Regression (SVR) models (purple) and tree-based models (pale pink) against the radius of the Morgan fingerprints with the bit length set to 512. Solid line, models built on concatenated fingerprints; dashed line, models built on Tanimoto kernel descriptors.

### Descriptors Derived from Molecular Graphs

The number of iterations $h$, also known as the WL depth, determines the circular neighbourhoods of the atoms which are used to generate the WL kernel descriptors; see Chapter 2, Section 2.5 for more details. The larger the WL depth, the more information about the atoms' connectivity is encoded. Cross-validation was used to evaluate the influence of the WL depth on the performance of the models.

The performance of the linear, tree-based and SVR models built on WL kernel descriptors, with WL depths varying from two to ten, are shown in Figure 4.5. Numeric details of the individual model performances can be found in Appendix B, Section B.3.1. The average performance of the linear models with a WL depth of two is poor. The resulting negative $R^2$ value is predominantly due to the linear regression model. Although the other linear models performed reasonably well with $R^2 \geq 0.63$ and RMSE $\leq 16.7\%$, the performance was lower than higher WL depths. Increasing the WL depth above two had little to no effect on the average performance of the linear models, which remained at approximately an $R^2$ of 0.87 and a RMSE of 9.1%. The tree-based models exhibited a lower average performance compared to the linear (with WL depth less than two) and the SVR models. The average performance of the tree-based models ranged from 0.72 to 0.76 for the $R^2$ and 14.1 to 13.3% for the RMSE. The SVR models had the highest performance at each WL depth. As WL depth increased to five, the model performance also increased to $R^2 = 0.90$ and RMSE = 8.5%. The aver-

age performance of the linear, tree-based and SVR models built on WL kernel descriptors only marginally improved with WL depth greater than five. The WL depth was therefore set to five for further analysis.

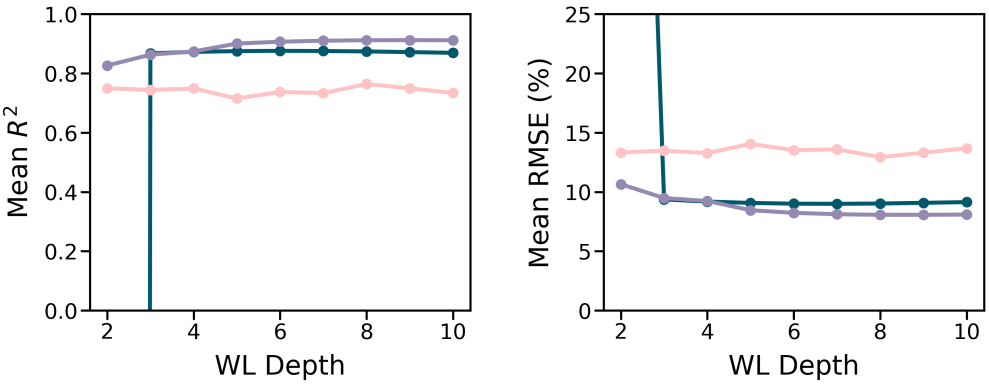



Figure 4.5: Cross-validated performance of the linear models (dark blue), Support Vector Regression (SVR) models (purple) and tree-based models (pale pink) against the Weisfeiler-Lehman (WL) depth of the WL kernel descriptors.

### 4.3.2 Cross-Validation Performance

The average five-fold cross-validation performance of the linear, tree-based and SVR models built on one-hot encodings, quantum chemical, concatenated fingerprints, Tanimoto kernel and WL kernel descriptors are shown in Table 4.5 to 4.7. The concatenated fingerprints are denoted as “Fingerprints: fingerprint type" and the Tanimoto descriptors as “Tanimoto: fingerprint type".

The performance of the linear models averaged over the descriptors ranged from 0.65 to 0.79 for the $R^2$ and 16.1 to 11.8% for the RMSE. On average, adding either the LASSO or ridge regularisation parameters did not improve the performance of the linear regression models (average $R^2 = 0.78$ and RMSE = 12.2%). Including both regularisation terms in the elastic net method deteriorated the performance by $-0.13$ for the $R^2$ and $+3.9\%$ for the RMSE. The Bayesian ridge method also did not improve the performance of the linear regression model.

For the tree-based models, the decision trees had average performance metrics across the descriptors of 0.67 for the $R^2$ and 14.9% for the RMSE. The decision trees were outperformed by the ensemble models. The gradient boosting and random forest models had an average $R^2$, RMSE scores across the descriptors of 0.86, 10.0% and 0.84, 10.4%, respectively. These ensemble methods combine multiple decision trees trained over the same dataset. This decreases the variance of the predictions with a slight increase in model bias, resulting in a less flexible model that is less prone to overfitting.

Table 4.5: Average Cross-Validated Performance of the Linear Models

| ML Algorithm | Linear Model | | | | | Mean |
|---|---|---|---|---|---|---|
| | Linear Regression | Lasso | Ridge | Elastic Net | Bayesian Ridge | |
| Mean $R^2$: | | | | | | |
| One-hot | 0.69 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| Quantum Chemical | 0.70 | 0.69 | 0.70 | 0.68 | 0.70 | 0.69 |
| WL | 0.92 | 0.93 | 0.93 | 0.67 | 0.93 | 0.87 |
| Fingerprints: Morgan1 | 0.67 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 |
| Fingerprints: RDK | 0.69 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| Fingerprints: MACCS | 0.62 | 0.64 | 0.64 | 0.64 | 0.64 | 0.63 |
| Tanimoto: Morgan1 | 0.93 | 0.94 | 0.94 | 0.61 | 0.94 | 0.87 |
| Tanimoto: RDK | 0.91 | 0.92 | 0.92 | 0.59 | 0.92 | 0.85 |
| Tanimoto: MACCS | 0.88 | 0.91 | 0.92 | 0.62 | 0.90 | 0.84 |
| Mean | 0.78 | 0.79 | 0.79 | 0.65 | 0.79 | |
| Mean RMSE (%): | | | | | | |
| One-hot | 15.2 | 15.0 | 15.0 | 15.0 | 15.0 | 15.0 |
| Quantum Chemical | 15.0 | 15.1 | 15.0 | 15.4 | 15.1 | 15.1 |
| WL | 7.6 | 7.4 | 7.3 | 15.7 | 7.3 | 9.1 |
| Fingerprints: Morgan1 | 15.6 | 15.4 | 15.4 | 15.5 | 15.5 | 15.5 |
| Fingerprints: RDK | 15.2 | 15.0 | 15.0 | 15.0 | 15.0 | 15.0 |
| Fingerprints: MACCS | 16.8 | 16.4 | 16.4 | 16.4 | 16.4 | 16.5 |
| Tanimoto: Morgan1 | 7.0 | 6.8 | 6.8 | 17.0 | 6.7 | 8.9 |
| Tanimoto: RDK | 8.2 | 7.9 | 7.8 | 17.6 | 7.8 | 9.9 |
| Tanimoto: MACCS | 9.4 | 8.2 | 8.0 | 16.9 | 8.6 | 10.2 |
| Mean | 12.2 | 11.9 | 11.8 | 16.1 | 11.9 | |

The linear SVR models had an average performance across the descriptors of $R^2$ equal to 0.68 and RMSE equal to 15.4%. For the non-linear SVR models, the descriptors were converted to higher dimensional feature space using a kernel function. Although the sigmoid function has been successfully used as a valid kernel, for this regression task it performed worse than the linear SVR model, with an average $R^2$ of 0.59 and RMSE of 16.6%. The SVR models implementing the polynomial and RBF kernels had a better predictive performance, with a respective average $R^2$ of 0.91 and 0.90 and RMSE of 8.3 and 8.4%. The SVR models built on the precomputed kernel descriptors (WL and Tanimoto) also performed well, with $R^2$ from 0.90 to 0.94 and RMSE 8.4 to 6.8%. Due to the moderate performance of the linear SVR model and the SVR model implementing the sigmoid kernel, these models were not considered in further analysis.

The one-hot encodings, quantum chemical descriptors and concatenated fingerprints performed better in combination with the tree-based models (average $R^2$: 0.87 to 0.91, RMSE: 9.7 to 8.2%). The predictive performance of the quantum chemical random forest model exhibited similar performance to the 70-30%

Table 4.6: Average Cross-Validated Performance of the Tree-Based Models

| ML Algorithm | Tree-Based Model | | | Mean |
|---|---|---|---|---|
| | Decision Tree | Gradient Boosting | Random Forest | |
| Mean $R^2$: | | | | |
| One-hot | 0.81 | 0.90 | 0.90 | 0.87 |
| Quantum Chemical | 0.87 | 0.92 | 0.93 | 0.91 |
| WL | 0.51 | 0.86 | 0.78 | 0.72 |
| Fingerprints: Morgan1 | 0.88 | 0.91 | 0.93 | 0.91 |
| Fingerprints: RDK | 0.86 | 0.90 | 0.93 | 0.90 |
| Fingerprints: MACCS | 0.87 | 0.89 | 0.92 | 0.89 |
| Tanimoto: Morgan1 | 0.43 | 0.81 | 0.73 | 0.66 |
| Tanimoto: RDK | 0.44 | 0.79 | 0.72 | 0.65 |
| Tanimoto: MACCS | 0.32 | 0.77 | 0.71 | 0.60 |
| Mean | 0.67 | 0.86 | 0.84 | |
| Mean RMSE (%): | | | | |
| One-hot | 11.8 | 8.7 | 8.7 | 9.7 |
| Quantum Chemical | 9.8 | 7.8 | 7.2 | 8.3 |
| WL | 19.1 | 10.2 | 12.8 | 14.1 |
| Fingerprints: Morgan1 | 9.4 | 8.2 | 7.1 | 8.2 |
| Fingerprints: RDK | 10.2 | 8.5 | 7.4 | 8.7 |
| Fingerprints: MACCS | 9.8 | 9.1 | 7.6 | 8.8 |
| Tanimoto: Morgan1 | 20.6 | 11.8 | 14.0 | 15.5 |
| Tanimoto: RDK | 20.4 | 12.4 | 14.4 | 15.8 |
| Tanimoto: MACCS | 22.5 | 13.1 | 14.7 | 16.8 |
| Mean | 14.9 | 10.0 | 10.4 | |

hold-out test performed by Doyle et al. In comparison, the kernel-based descriptors (WL and Tanimoto) performed better in combination with the SVR models (average $R^2$: 0.85 to 0.92, RMSE: 10.1 to 7.8%). There was little to no improvement in the performance of the SVR models built on the kernel-based descriptors implementing additional kernels. Nevertheless, these models were considered in further analysis for completeness and to allow direct comparison with other descriptors.

The performances of the models built on one-hot encodings were comparable to those built on chemically meaningful descriptors. The split of the dataset in $k$-fold cross-validation is random; analogous to the 70-30% hold-out split performed by Doyle et al. Due to the combinatorial nature of the data, each of the five cross-validation test sets likely contained the same molecules but in different reactions. This enabled the models to learn the reactivity of the individual molecules and resulted in an unreliable superior performance. The results of the cross-validation test were to determine which models and descriptors were unsuitable for this specific regression task of predicting reaction yield. Should a model perform

Table 4.7: Average Cross-Validated Performance of the Support Vector Regression (SVR) Models

| ML Algorithm | SVR Kernel | | | | | Mean |
|---|---|---|---|---|---|---|
| | Linear | Polynomial | RBF | Sigmoid | Pre-computed | |
| Mean $R^2$: | | | | | | |
| One-hot | 0.70 | 0.90 | 0.91 | 0.59 | | 0.77 |
| Quantum Chemical | 0.70 | 0.90 | 0.90 | 0.43 | | 0.73 |
| WL | | 0.93 | 0.92 | 0.83 | 0.92 | 0.90 |
| Fingerprints: Morgan1 | 0.68 | 0.91 | 0.92 | 0.49 | | 0.75 |
| Fingerprints: RDK | 0.70 | 0.91 | 0.92 | 0.26 | | 0.70 |
| Fingerprints: MACCS | 0.64 | 0.87 | 0.86 | 0.31 | | 0.67 |
| Tanimoto: Morgan1 | | 0.93 | 0.91 | 0.89 | 0.94 | 0.92 |
| Tanimoto: RDK | | 0.91 | 0.89 | 0.85 | 0.92 | 0.89 |
| Tanimoto: MACCS | | 0.91 | 0.90 | 0.69 | 0.90 | 0.85 |
| Mean | 0.68 | 0.91 | 0.90 | 0.59 | | |
| Mean RMSE (%): | | | | | | |
| One-hot | 15.0 | 8.5 | 8.1 | 17.5 | | 12.3 |
| Quantum Chemical | 15.1 | 8.8 | 8.6 | 20.5 | | 13.2 |
| WL | | 7.4 | 7.7 | 11.2 | 7.5 | 8.5 |
| Fingerprints: Morgan1 | 15.5 | 8.0 | 7.5 | 19.5 | | 12.6 |
| Fingerprints: RDK | 15.0 | 8.1 | 7.9 | 23.5 | | 13.6 |
| Fingerprints: MACCS | 16.5 | 9.9 | 10.0 | 22.7 | | 14.8 |
| Tanimoto: Morgan1 | | 7.3 | 8.0 | 9.2 | 6.8 | 7.8 |
| Tanimoto: RDK | | 8.4 | 8.9 | 10.5 | 7.9 | 8.9 |
| Tanimoto: MACCS | | 8.3 | 8.7 | 15.1 | 8.4 | 10.1 |
| Mean | 15.4 | 8.3 | 8.4 | 16.6 | | |

poorly in cross-validation, it is also likely to perform poorly when predicting on unseen data.

## 4.4 Conclusions

The preliminary evaluation of the machine learning models used to predict the yield of a set of Buchwald-Hartwig amination reactions was accomplished with five-fold cross-validation. Linear, tree-based and SVR models were built on one-hot encodings, quantum chemical descriptors, concatenated molecular fingerprints, Tanimoto kernel descriptors and WL kernel descriptors. The performance of the models was compared using the $R^2$ and RMSE performance metrics.

Optimum parameters of the structure-based descriptors were identified. Bit lengths of the Morgan, FMorgan and RDK molecular fingerprints, varying from 32 to 2048 were evaluated. Altering the bit length had the most effect on the models built using RDK fingerprint, where these models showed little to no im-

provement with bit lengths above 512 bits. The predictions of the models built on FMorgan fingerprints had a poor correlation to the experimental values and were therefore omitted from further analyses. The radius of the Morgan fingerprints had a very minor effect on the performance of the models. Further analyses of the models built on concatenated fingerprints and Tanimoto kernel descriptors were performed with a bit length of 512 and Morgan fingerprints with a radius of one. The WL depth of the WL kernel descriptors was assessed, with depths from two to ten. The models built on WL kernel descriptors with a depth of five were found to have the optimum performance. A value of five was therefore used in additional testing.

The preliminary cross-validation work uncovered that, in general, non-linear models (SVR and tree-based) often outperformed the linear models in predicting reaction yield. The average cross-validated performance across all descriptors gave average $R^2$ values ranging from 0.65 to 0.79 for the five linear models, 0.67 to 0.86 for the three tree-based models, and 0.59-0.91 for the three non-linear SVR models. This trend is also observed in the performance of linear versus non-linear SVR models. The SVR models implementing the polynomial, RBF and precomputed kernels showed better performance statistics, with $R^2$ metrics ranging from 0.86 (the RBF kernel applied to the MACCS fingerprints) to 0.94 (the precomputed Morgan1 Tanimoto kernel). The linear SVR models showed lower performance statistics with an average $R^2$ value of 0.68 across all descriptors. The sigmoid kernel also performed poorly, with an average $R^2$ of 0.59 across all descriptors. Overall, the SVR models, implementing the RBF and Sigmoid kernels, showed marginally better performance metrics on average across all descriptors ($R^2$ around 0.90) compared to the random forest model ($R^2$ of 0.84).

The non-linear models built on structure-based descriptors showed comparable results to those built on quantum chemical calculations. The Doyle group's random forest model based on quantum chemical descriptors obtained an $R^2$ score of 0.93. By comparison, the $R^2$ values for the random forest models constructed on concatenated fingerprints ranged from 0.92 to 0.93. The random forest models built using kernel-based descriptors exhibited slightly worse $R^2$ values, ranging from 0.71 to 0.78. The similar performance metrics between the structure-based and quantum chemical models extend to SVR. The SVR models with the polynomial and RBF kernels applied to the quantum chemical descriptors had an $R^2$ of 0.90. The structure-based SVR models with the polynomial, RBF, and precomputed kernels had $R^2$ ranging from 0.86 to 0.94. Based on these preliminary cross-validation results, structure-based descriptors should be suitable for this regression task.

We have demonstrated the suitability of structure-based descriptors in predicting reaction yield. Structure-based descriptors may be a viable alternative to quantum chemical properties due to their relative quickness to generate. We have established the ideal parameters for the structure-based descriptors. We examined machine learning algorithms using a preliminary cross-validation test to select non-linear SVR as the most promising for additional exploration. Due to the moderate performance of the SVR models implementing the linear SVR sigmoid kernels, these models were not considered further. Chapter 5 subjects these kernel methods to additional testing and external validation to investigate the limits to the generalisability of these models.

# Chapter 5

# Kernel Methods for Predicting Yields of Chemical Reactions

## 5.1 Introduction

Molecular descriptors employed in regression tasks related to chemical reactivity have often been based on time-consuming, computationally demanding quantum chemical calculations, usually Density Functional Theory (DFT). The Doyle group pioneered the prediction of reaction yield by developing a random forest model built on quantum chemical descriptors.[32–34] While calculated properties based on shared atoms are common for representing reaction components in a single reaction class, it limits the domain of applicability. Structure-based descriptors derived from the molecular graph structure, such as fingerprints and graph kernels, are quicker to calculate and applicable to any molecule. Several structure-based random forests and deep learning models have been reported for the regression task of predicting reaction yield.[37,211,240,241]

In the previous chapter (Chapter 4), preliminary cross-validation was performed using the Buchwald-Hartwig combinatorial dataset. We evaluated numerous molecular descriptors and machine learning algorithms. The molecular descriptors included quantum chemical, molecular fingerprints, and kernel-based descriptors. The machine learning algorithms included linear, tree-based, and SVR methods. The non-linear SVR models provided the best overall performance statistics across all descriptors. SVRs have also been employed successfully in a variety of other regression tasks in chemistry, such as predicting bioactivity, toxicity-related properties, and physicochemical properties.[29,212–214,242]

The pioneering work published by Doyle *et al.* focused on models built on calcu-

lated properties.[32] It was reported that the random forest method outperformed SVR in an initial 70-30% hold-out test. The only SVR model evaluated was linear SVR, which is not able to describe non-linear relationships between features and target values. In our preliminary cross-validation work, we investigated several non-linear kernels. The performance of the models was compared using the Coefficient of Determination ($R^2$) and Root Mean Squared Error (RMSE) performance metrics. The linear SVR built on quantum chemical descriptors had a predictive performance ($R^2 = 0.70$ and RMSE = 15.1%) much lower than the non-linear SVR implementing the polynomial and Gaussian Radial Basis Function (RBF) kernels, which both had an $R^2$ of 0.90 and RMSE of 8.3 and 8.4%, respectively. These values are relatively close to the random forest model with $R^2$ of 0.93 and RMSE of 7.2%.

The SVR algorithm has proven to be appropriate for regression tasks related to Quantitative Structure-Activity Relationship (QSAR) and cheminformatics.[243] SVR also demonstrates encouraging preliminary results for predicting reaction yield, as described in Chapter 4. In this chapter, we study the application of kernel approaches for predicting the yield of combinatorial reaction data. Compared to previously used quantum chemical calculations, structure-based descriptors offer speed, alignment with the language of synthetic chemists, and applicability to every molecule. This chapter aims to develop SVR models employing structure-based descriptors and compare their performance to models employing quantum chemical properties. The best-performing SVR model for each descriptor is subject to external validation. We establish an external validation procedure to illustrate a feasible real-life implementation of the models.

### 5.1.1 Pioneering Work on the Prediction of Reaction Yield

Cross-validation analyses on the Buchwald-Hartwig dataset are insufficient to gain a reliable understanding of model generalisability to unseen reactions. This is owing to the combinatorial nature of the data. By providing the model with training data that contains molecules in the test reactions, the models can learn the reactivity of the individual molecules. Although chemically meaningful descriptors are provided, the models perform no better than those built on one-hot encodings. A more appropriate assessment of model generalisability is to test the models with molecules not present in the training set.[232] In this type of out-of-sample test, a set of reactions containing specific molecules (one or more reaction components) not present in the training set are withheld from model training and used to assess the predictive ability of the trained model.

In the original work by Doyle *et al.*, out-of-sample test sets were designed by splitting the reactions along the high-throughput plates, where each plate contained a separate set of additives.[32] The random forest model built on quantum chemical descriptors was trained using the reactions on Plate **1** and Plate **2**, then tested using Plate **3**. In a technical comment, Chuang and Keiser identified that alternative splits of the plates resulted in much lower performance, suggesting the random forest model built on quantum chemical descriptors was limited.[33] Designing out-of-sample tests without considering the distribution of chemical reactivity covered in the training set was unreliable. The reactions on Plate **2** contained the most inhibitory additives, diminishing the reaction yields (0-10%). As a result, when failed reactions were not present in training, the models overpredicted reaction yield. The reactions on Plate **3** contained more high-yielding reactions ($> 80\%$). Therefore, when Plate **3** was used as the test set, the models underpredicted the higher-yielding reactions. Reactions that cover a broad range of chemical space and observed variables must be used in model training and addressed in the test-set design.

The out-of-sample test sets were redesigned in a technical response using activity ranking.[34] The focus remained on the ability of the models to predict reactions containing unseen additives. The mean yield of the reactions containing each additive was ranked from lowest to highest. All training sets included the highest- and lowest-yielding additives. Test sets were constructed by taking every fourth molecule from the remaining additives. Repeating this three more times created four test sets in total. Using activity ranking to design test sets ensured the model was trained on a wide range of reaction yields. The quantum chemical random forest model showed good generalisability across the additive dimension, with a mean $R^2$ of 0.69 and RMSE of 14.9% in the additive ranked test. Doyle *et al.* did not perform out-of-sample tests along any other dimension.[34]

The experiments performed by Doyle *et al.* aimed to assess additive effects on the reaction yield of the Buchwald-Hartwig amination. From evaluating the relative importance of the calculated properties used to construct the random forest model, Doyle *et al.* hypothesised that an oxidative additive of the isoxazole to the palladium catalyst side could act as a competitive side reaction and result in lower-yielding reactions. Additional experiments supported their hypothesis that electrophilic isoxazole additives undergo a competitive side reaction. The N-O oxidative addition of the electrophilic isoxazoles to palladium ($Pd^0$) resulted in a lower yield of the aniline products in the Buchwald-Hartwig reactions.

### 5.1.2 Our Aims

Using machine learning methods for predicting reaction yield from combinatorial data has real-life applications, for example, pharmaceutical discovery and development. High-throughput experimentation is a practical workflow for chemists working on a specific problem with increased time pressures.[244] In lead optimisation, multiple analogues of hit molecules are synthesised to improve the potency, target selectivity, toxicity, physiochemical and Adsorption, Distribution, Metabolism and Excretion (ADME) properties. Synthesising various analogues of a hit molecule involves making slight changes to the reactants. In the case of the Buchwald-Hartwig reaction, this would be the aryl halide or amine. Since only a single amine was considered in the Buchwald-Hartwig dataset, the amine dimension of chemical space cannot be explored without performing additional reactions. However, it is possible to examine the model's generalisability along the aryl halide dimension.

The pioneering work of the Doyle group concentrated on developing a model to evaluate additive effects on the yield of Buchwald-Hartwig amination reactions.[32] We focus on constructing a model to predict the reaction yield of unexplored Buchwald-Harwig reactions; this covers variations in the aryl halide reactant, catalyst ligand, and base. To assess the viability of a model capable of predicting reaction yield when altering the reaction components, we perform rigorous testing using the combinatorial dataset reported by Doyle *et al.*

Initially, we emphasise the importance of test set design, not just when evaluating model performance in the additive dimension of the dataset but also in the aryl halide dimension. We then explore the extent of model generalisability when the training dataset contains only a few examples of a reaction component; this is the case for the base and catalyst ligand in the Buchwald-Hartwig combinatorial dataset. We design leave-one-base-out and leave-one-ligand-out tests, which are analogous to leave-one-out tests.

We conduct out-of-sample tests using activity ranking for a more reliable evaluation of model performance along the dataset's additive and aryl halide dimensions. An additive ranked test allows for a direct comparison of our kernel models to the work of Doyle *et al.*[34] We also investigate an aryl halide ranked test to determine which models are appropriate for external examination against reactions unfamiliar to the model.

If a molecule not observed by the model differs significantly from the training data, the model will struggle to predict accurately. We explore the domain of applicability of the models to determine the scope of reaction space the models can

predict. To ascertain the domain of applicability, we evaluate the performance of the models with respect to similarity to training. The similarity score quantifies how similar a reaction in the test set is to the training set. We complete this analysis for both activity-ranked tests.

Lastly, we design an external examination procedure which imitates real-life circumstances, working with combinatorial reaction data. The aim is to investigate if employing yield prediction models trained on combinatorial data is viable in a medicinal setting, such as when synthesising analogues of potential drug molecules. We also study the extent of model generalisability by considering reactions with high and low degrees of similarity to the training data. Although we have yet to conduct any experiments, we report and compare the yields of the reactions in the proposed combinatorial dataset.

## 5.2 Methodology

### 5.2.1 Dataset

The combinatorial dataset reported by Doyle *et al.*,[32] consisting of 4608 Buchwald-Hartwig reactions, was used in this study. The structure of the Buchwald-Hartwig dataset and the cleaning process is described in Chapter 4, Section 4.2.1.

A set of prospective combinatorial reactions was compiled to externally validate the top performing SVR models. The proposed reactions follow the same combinatorial framework as the Doyle dataset, differing in the aryl halide, base, catalyst ligand, and additive, with the same shared atoms for each reaction component (Figure 4.1). All possible combinations of 59 aryl halides, three bases, four catalyst ligands and two additives, formed a total of 1416 proposed reactions. The aryl halides cover *ortho-*, *meta-* and *para-* substituents, with a range of electron withdrawing and electron donating groups (Figure B.5 to B.7). Five of the aryl halides are present in the Buchwald-Hartwig dataset and will be used as benchmarks. The base DBU and catalyst ligand BrettPhos were selected along with the two higher-yielding bases and ligands from the Buchwald-Hartwig dataset: MTBD, BTMG, *t*-BuXPhos and *t*-BuBrettPhos (Figure B.8 and B.9). To investigate whether the reactions of the *ortho*-substituted halopyridines are proceeding via an alternative reaction pathway, the prospective reactions will also be performed without a catalyst. As the aim of these reactions is to assess model generalisability, with particular interest along the aryl halide dimension, the reactions will be carried out with and without a single high-yielding isoxa-

zole additive: 3-methylisoxazole (Figure B.10). The proposed reactions will be performed experimentally using high-throughput chemistry to identify reaction yields.

### 5.2.2 Molecular Descriptors and Preprocessing

The molecular descriptors explored were one-hot encodings, quantum chemical descriptors, concatenated molecular fingerprints, Tanimoto kernel descriptors and Weisfeiler-Lehman (WL) kernel descriptors. These descriptors represented the reactions in the Buchwald-Hartwig dataset and prospective reactions. The methodology of the descriptors, including the generation process, is detailed in Chapter 4, Section 4.2.2.

The quantum chemical descriptors, consisting of molecular, atomic and vibrational properties, were calculated for the prospective reactions using the same methodology as Doyle *et al.*[32] The calculations were submitted to Spartan and the features were manually extracted from the resulting text files.* During the calculation of the atomic descriptors, an issue was found in the calculation of the Nuclear Magnetic Resonance (NMR) shift for the aryl iodides. The issue is detailed in Appendix B, Section B.5. The aryl iodides were included in the initial model development for consistency with the methodology used by Doyle *et al.*, but due to the ambiguity in the calculations, were not included in the yield predictions of the prospective reactions. The five aryl iodides removed from Doyle's dataset resulted in 1320 excluded reactions, leaving 3288 training reactions in the external examination.

Five-fold cross-validation on the Buchwald-Hartwig dataset revealed optimal parameters of the molecular fingerprints and WL kernel descriptors. The optimum bit length of the Morgan and RDK fingerprints was 512-bits. The optimum WL depth of the WL graph kernel was five. These values were therefore used in the out-of-sample and validation tests.

#### Encoding Missing Molecules

The descriptors must account for the missing molecules included in the prospective reactions. For the quantum chemical descriptors, concatenated molecular fingerprints and one-hot encodings, the bits corresponding to the missing molecules were set to zero. For the kernel-based descriptors, the missing molecules were incorporated in the calculation of the kernel of each reaction component. For

*The quantum chemical descriptors for the prospective reactions were calculated by Magnus W. D. Hanson-Heine.

example, the kernel between two molecules ($m_1$ and $m_2$) is defined below.

$$k'(m_1, m_2) = \begin{cases} k(m_1, m_2) + 1, & \text{if } m_1 \text{ and } m_2 \text{ are both present} \\ 2, & \text{if } m_1 \text{ and } m_2 \text{ are both missing} \\ 1, & \text{otherwise} \end{cases}$$

If both molecules were present the kernel of the two molecules is the original kernel plus one, if both molecules were missing the kernel equals two, otherwise the kernel is equal to one. This method is only applied when the training or test data includes missing molecules.

### 5.2.3 Model Building and Evaluation

Machine learning models relating descriptors to reaction yield were developed using the SVR method as implemented in scikit-learn.[239] Mathematical details of the SVR algorithm can be found in Chapter 2, Section 2.4. The linear and sigmoid kernel functions were not explored in this chapter due to the moderate performance in cross-validation. SVR models implementing the polynomial, RBF and pre-computed (where applicable) kernel functions were evaluated.

The hyperparameters of the SVR models ($\epsilon$ and $C$) were optimised in scikit-learn by performing an exhaustive grid-search over the specified parameter grid (Table 4.4 and 4.3). This was accomplished by performing five-fold cross-validation on the training set. For each train-test split of the data, the training set was shuffled and split into five groups. In turn, each of the five groups was used to test a model trained on the remaining four groups. The average performance statistics were calculated and compared to identify the best combination of hyperparameters. The grid search cross-validated and training set performances are reported for each model in the out-of-sample tests and the prospective SVR models in Appendix B, Section B.7 and B.8. The best combination of hyperparameters was used to build the SVR model on the training set to predict the yield of the test set. The hyperparameters of the prospective SVR models are reported in Appendix B, Section B.9.

The performances of the models were evaluated by $R^2$ and RMSE using data points outside of the training set. All analysis was performed using scikit-learn.

### 5.2.4 Test Set Design

Out-of-sample test sets were designed to assess model generalisability to unseen molecules along each reaction component (additive, aryl halide, base and ligand). The models were tested on a specific set of molecules that were withheld from model training.

#### Without Activity Ranking

Out-of-sample test sets designed without activity ranking were investigated. The Buchwald-Hartwig dataset was split based on the additives on each plate, as performed by Chuang and Keiser[33]. This split of the data is referred to as the *additive: plate* out-of-sample test (Table 5.1).

Table 5.1: Additives in the Test Sets Split by High Throughput Plate Number

| Additive | Low Yielding | | | | | | | | | | | | | | | | | High Yielding | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 13 | 10 | 11 | 14 | 16 | 18 | 9 | 8 | 15 | 2 | 1 | 21 | 22 | 23 | 17 | 20 | 12 | 4 | 5 | 6 | 3 | 19 |
| Plate 1 | | | | | | | | | | 2 | 1 | | | | | | | 4 | 5 | 6 | 3 | |
| Plate 2 | 13 | 10 | 11 | 14 | | | 9 | 8 | 15 | | 1 | | | | | | 12 | | | | | |
| Plate 3 | | | | | 16 | 18 | | | | | | 21 | 22 | 23 | 17 | 20 | | | | | | 19 |

The dataset was also split along the aryl halide dimension in two different ways (Table 5.2). The first split was based on the ring type of the aryl halide, either phenyl or pyridyl, referred to as the *aryl halide: ring type* test. The second, based on the halide present in the aryl halide, is referred to as *aryl halide: halide type* test.

Table 5.2: Aryl Halides in the Test Sets Split by Ring Type and Halide

| Aryl Halide | Low Yielding | | | | | | | | | | | | | | | High Yielding |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4 | 7 | 1 | 13 | 5 | 2 | 6 | 3 | 14 | 10 | 8 | 15 | 11 | 9 | 12 | |
| Phenyl | 4 | 7 | 1 | | 5 | 2 | 6 | 3 | | | 8 | | | 9 | | |
| Pyridyl | | | | 13 | | | | | 14 | 10 | | 15 | 11 | | 12 | |
| Cl | 4 | 7 | 1 | 13 | | | | | | 10 | | | | | | |
| Br | | | | | 5 | 2 | | | 14 | | 8 | | 11 | | | |
| I | | | | | | | 6 | 3 | | | | 15 | | 9 | 12 | |

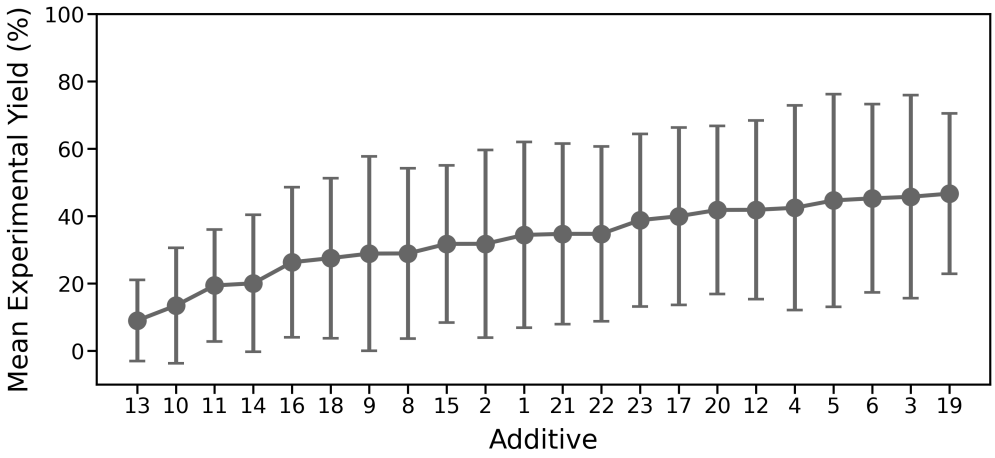
Due to the small number of bases (three) and catalyst ligands (four) in the dataset, two leave-one-molecule-out tests were performed. In the first test, the dataset was split into three test sets based on the base used in the reactions, herein called the *leave-one-base-out*. For the second test, the dataset was split into four test sets based on the ligand used in the reactions, herein called *leave-one-ligand-out*. In turn, each test set was withheld from model training.

### With Activity Ranking

The generalisability of the models along the additive and aryl halide dimensions was assessed using activity-ranked tests. Activity ranking was used to ensure the models were trained on a range of reaction yields.[34] The reactions excluded controls and reactions containing additive **7**, as these were not included in model development. Initially, the mean yields of the reactions containing each molecule within the two reaction components were ranked from lowest to highest (Figure 5.1). The molecules with the highest and lowest mean yields were included in all training sets. Test sets were constructed from the remaining molecules by taking every $n^{th}$ molecule, where $n$ is four for the additive ranked test (Table 5.3) and three for the aryl halide ranked test (Table 5.4).

The order of the mean yields of the additives calculated in this work differs slightly from Doyle et al.[34] There are two instances where the mean reaction yields of two additives are within 0.1%. The interchange of the additives in the ranked order resulted in slightly different additive ranked test sets (Table 5.3). The different test sets caused very minor differences in the performance of the quantum chemical random forest model, with a mean $R^2$ of 0.68 and mean RMSE 15.3% in this work, compared to a mean $R^2$ of 0.69 and mean RMSE of 14.9% reported by Doyle *et al.*[34]

(a)




(b)

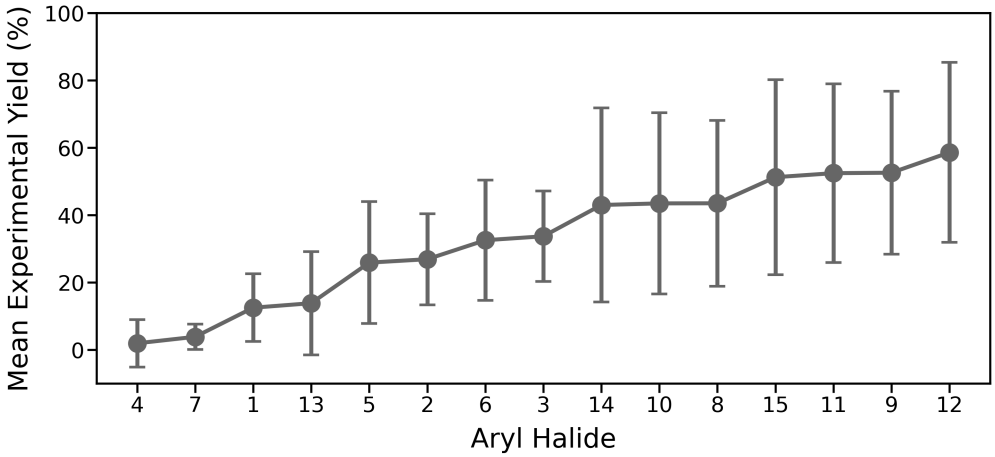



Figure 5.1: Mean experimental yield of the reactions containing each (a) additive and (b) aryl halide. Errorbars; standard deviation.

Table 5.3: Additive Ranked Test Sets[†]

This Work

| Additive | Low Yielding | | | | | | | | | | | | | | | | | | | | | High Yielding |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 13 | 10 | 11 | 14 | 16 | 18 | 9 | 8 | 15 | 2 | 1 | 21 | 22 | 23 | 17 | 20 | 12 | 4 | 5 | 6 | 3 | 19 |
| Set 1 | | 10 | | | | 18 | | | | 2 | | | | 23 | | | | 4 | | | | |
| Set 2 | | | 11 | | | | 9 | | | | 1 | | | | 17 | | | | 5 | | | |
| Set 3 | | | | 14 | | | | 8 | | | | 21 | | | | 20 | | | | 6 | | |
| Set 4 | | | | | 16 | | | | 15 | | | | 22 | | | | 12 | | | | 3 | |

Doyle *et al.*[34]

| Additive | Low Yielding | | | | | | | | | | | | | | | | | | | | | High Yielding |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 13 | 10 | 11 | 14 | 16 | 18 | 9 | 8 | 15 | 2 | 1 | 21 | 22 | 23 | 17 | 20 | 12 | 4 | 5 | 6 | 3 | 19 |
| Set 1 | | 10 | | | | 18 | | | | **15** | | | | 23 | | | | 4 | | | | |
| Set 2 | | | 11 | | | | 9 | | | | 1 | | | | 17 | | | | 5 | | | |
| Set 3 | | | | 14 | | | | 8 | | | | 21 | | | | **12** | | | | 6 | | |
| Set 4 | | | | | 16 | | | | **2** | | | | 22 | | | | **20** | | | | 3 | |

[†]Differences in ranking order are highlighted in **bold**.

Table 5.4: Aryl Halide Ranked Test Sets

| Aryl Halide | Low Yielding | | | | | | | | | | | | | High Yielding |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4 | 7 | 1 | 13 | 5 | 2 | 6 | 3 | 14 | 10 | 8 | 15 | 11 | 9 | 12 |
| Set 1 | | 7 | | | 5 | | | 3 | | | 8 | | | 9 | |
| Set 2 | | | 1 | | | 2 | | | 14 | | | 15 | | | |
| Set 3 | | | | 13 | | | 6 | | | 10 | | | 11 | | |

## 5.3 Results and Discussion

### 5.3.1 Diversity of the Buchwald-Hartwig Dataset

The reactions in the Buchwald-Hartwig dataset cover a range of yields (Figure 5.2). The majority of reactions are low yielding (0 to 10%) due to the use of inhibitory additives and the lower reactivity of the aryl chlorides. The lowest proportion of reactions are high yielding (90 to 100%). When assessing the generalisability of a model, it is important to ensure an even spread of chemical reactivity is included in both the training and the test sets. If specific ranges of chemical reactivity are excluded from training, the resulting models can be biased in its predictions and result in the under or overprediction of the test reactions.

Chuang and Keiser have shown that splitting the Buchwald-Hartwig dataset by high-throughput plate (where all inhibitory additives were present on a single plate) leads to an inaccurate estimation of model performance.[33] This was due

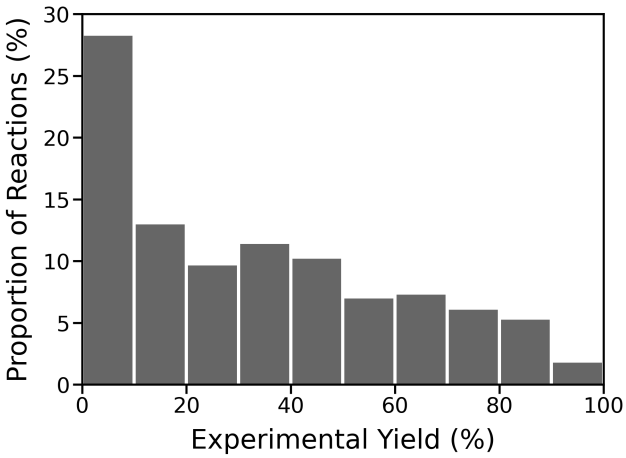



Figure 5.2: Distribution of experimental yields, excluding control reactions and reactions containing 5-phenyl-1,2,4-oxadiazole (additive **7**), corresponding to 3955 data points.

to the underrepresentation of inhibitory additives which led to an uneven cover of reaction yields in the training set (Figure B.13a). The underrepresentation of less reactive molecules in the training set is prevalent in all out-of-sample tests designed without activity ranking (aryl halide halide type, aryl halide aryl type, leave-one-base-out and leave-one-ligand-out). The aryl chlorides are less reactive than the bromides and the iodides (Figure B.13c), the pyridyl halides are lower yielding compared to the phenyl halides (Figure B.13b) and the reactions containing the XPhos ligand gave yields less than 70% (Figure B.13e). When these less reactive molecules are used as the test set, it is likely that these lower yielding reactions will be overpredicted by the models. Splitting data using activity ranking ensures the models are trained and tested on similar distributions of reaction yields (Figure B.14).

It is important to assess whether the reactions in the test set are within the domain of applicability. The similarity of the test reactions to the training reactions was evaluated. The pairwise Tanimoto score calculated using the Morgan2 fingerprint was used as the similarity metric. For each reaction component in a single test reaction, the similarity was calculated between the test molecule and all training molecules in the same reaction component class. The dot product of the reaction component similarity vectors was calculated. The maximum product was taken as the maximum similarity to training value. The maximum similarity to training score ranges from zero, which indicates not similar, to one, which indicates an identical reaction. For the additive and aryl halide ranked tests, the maximum similarity to training ranged from 0.30 to 0.65 and from 0.30 to 0.60, respectively (Figure 5.3). The models are expected to predict instances with

low maximum similarity scores less accurately than those with high maximum similarity scores.

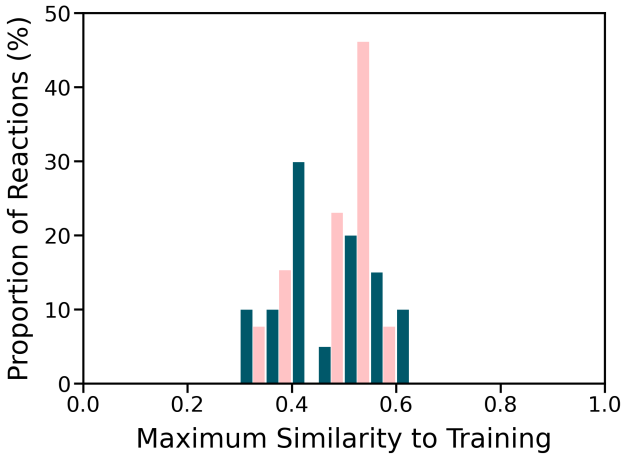



Figure 5.3: Distributions of maximum similarity to training for the additive ranked test sets (dark blue bars) and aryl halide ranked test sets (pale pink bars). Maximum similarity to training was calculated using the maximum product of pairwise Tanimoto scores, with the Morgan2 fingerprint, of the reaction components.

If all combinations of the additives, aryl halides, bases and ligands are in the dataset (this is not always the case as reactions with missing yield data were removed), the maximum similarity to training is dependent upon the unseen molecules in the test sets (i.e. the additives in the additive ranked test and the aryl halides in the aryl halide ranked test). For example, if the reaction $R_1 = (A_1, H_1, B_1, L_1)$ is in the training set and the reaction $R_2 = (A_2, H_1, B_1, L_1)$ is in the test set (where $A_n$ is the $n^{th}$ additive, $H_n$ is the $n^{th}$ aryl halide, $B_n$ is the $n^{th}$ base and $L_n$ is the $n^{th}$ ligand), then the similarity score would only be dependent on the additives in the reactions as shown in Equation 5.1. The maximum similarity to training scores of the additives and aryl halides for both activity-ranked tests can be found in Table B.10.

$$\begin{aligned} T(R_1, R_2) &= T(A_1, A_2)\, T(H_1, H_1)\, T(B_1, B_1)\, T(L_1, L_1) \\ &= T(A_1, A_2) \cdot 1 \cdot 1 \cdot 1 \\ &= T(A_1, A_2) \end{aligned} \tag{5.1}$$

### 5.3.2 Prediction of Reaction Yield

The generalisability of the machine learning models was evaluated using out-of-sample tests, designed without and with activity ranking. The models were built

on quantum chemical descriptors, concatenated molecular fingerprints, Tanimoto kernel descriptors and WL kernel descriptors. The performances of the SVR and baseline models in the out-of-sample tests are shown in Table 5.5, 5.6 and 5.7. The performance metrics are reported as the average over the test sets for the specified split of the data. The grid search cross-validated (on the training set), training set and test set performances of the models for the individual test sets can be found in Appendix B, Section B.7 and B.8. The performances of the models built on one-hot encodings are reported to assess whether the models were fitting any underlying combinatorial structure in the training reactions.

### Without Activity Ranking

To demonstrate the importance of test set design, out-of-sample tests were designed without considering the distribution of reaction yields covered in the training and test sets. The following out-of-sample tests were designed without activity ranking: additive plate split (Table 5.1), aryl halide ring and halide splits (Table 5.2), leave-one-base-out and leave-one-ligand-out.

In the additive plate split, the models built on structure-based descriptors had similar performances ($0.50 < R^2 < 0.53$, $17.9\% > \text{RMSE} > 17.4\%$) to the quantum chemical random forest model reported by Doyle et al.[32–34]. The structure-based models also overpredicted the yield of the reactions containing the inhibitory additives on Plate **2**. The models in the aryl halide splits based on ring type and halide had low performances. The $R^2$ for the aryl chloride test set in the halide split was negative for all models, due to the overprediction of the less reactive aryl chlorides.

The average model performances in the leave-one-base-out and leave-one-ligand-out tests were modest (Table 5.5). The SVR model built on the MACCS fingerprints with the polynomial kernel applied was the only model to outperform ($R^2 = 0.57$) the one-hot encodings model ($R^2 = 0.53$) in the leave-one-base-out test. For the leave-one-ligand-out test, all models had a negative $R^2$ for the XPhos test set due to the uneven representation of low yields in the training set, which resulted in the overprediction of the reaction yields. The quantum chemical model had a poor performance across all ligand test sets (Table B.25). The SVR model built on the Tanimoto kernel descriptors with the polynomial kernel applied ($R^2 = 0.48$) outperformed the other models ($R^2 \leq 0.32$) in the leave-one-ligand-out test.

The lower average performances of these tests (Table 5.5) in comparison to the activity ranked tests (Table 5.6 and 5.7) underscore the importance of test set

Table 5.5: Mean Performance Statistics for the Top Reaction Yield Prediction Models Built Using the Support Vector Regression (SVR) Algorithm and Baseline Random Forest Models Without Activity Ranking‡

| Descriptor | Kernel | $R^2$ | | RMSE (%) | |
|---|---|---|---|---|---|
| Additive: Plate | | | | | |
| One-hot Encodings | Polynomial | 0.47 | (0.28) | 18.8 | (3.9) |
| Quantum Chemical | RBF | 0.17 | (0.31) | 23.5 | (2.7) |
| Fingerprints: Morgan1 | RBF | 0.51 | (0.35) | 17.6 | (5.3) |
| Tanimoto: Morgan1 | Polynomial | 0.53 | (0.30) | 17.4 | (4.6) |
| WL | Precomputed | 0.50 | (0.32) | 17.9 | (4.5) |
| *Quantum Chemical Random Forest* | | *0.54* | *(0.34)* | *16.9* | *(5.2)* |
| *One-hot Random Forest* | | *0.41* | *(0.47)* | *18.9* | *(6.0)* |
| Aryl Halide: Ring Type | | | | | |
| One-hot Encodings | Polynomial | -0.21 | (0.29) | 28.6 | (1.7) |
| Quantum Chemical | RBF | -0.68 | (0.82) | 34.4 | (14.5) |
| Fingerprints: MACCS | RBF | 0.34 | (0.17) | 21.6 | (6.7) |
| Tanimoto: MACCS | Precomputed | 0.21 | (0.22) | 23.5 | (7.5) |
| WL | Precomputed | -0.04 | (0.22) | 26.4 | (1.9) |
| *Quantum Chemical Random Forest* | | *-0.36* | *(0.06)* | *30.8* | *(6.3)* |
| *One-hot Random Forest* | | *-0.52* | *(1.0)* | *30.4* | *(5.8)* |
| Aryl Halide: Halide Type | | | | | |
| One-hot Encodings | Polynomial | -0.47 | (1.11) | 26.9 | (7.7) |
| Quantum Chemical | RBF | -0.95 | (0.92) | 32.3 | (6.8) |
| Fingerprints: MACCS | RBF | -0.29 | (1.15) | 24.5 | (9.3) |
| Tanimoto: MACCS | Polynomial | -0.20 | (1.01) | 23.9 | (8.5) |
| WL | Precomputed | -0.27 | (1.08) | 24.3 | (9.3) |
| *Quantum Chemical Random Forest* | | *-0.05* | *(0.62)* | *23.2* | *(6.9)* |
| *One-hot Random Forest* | | *-0.72* | *(2.11)* | *26.3* | *(14.7)* |
| Leave-One-Base-Out | | | | | |
| One-hot Encodings | RBF | 0.53 | (0.25) | 17.7 | (4.9) |
| Quantum Chemical | RBF | -0.30 | (0.39) | 30.1 | (6.9) |
| Fingerprints: MACCS | Polynomial | 0.57 | (0.17) | 17.0 | (3.3) |
| Tanimoto: MACCS | Precomputed | 0.45 | (0.24) | 19.7 | (6.6) |
| WL | Precomputed | 0.52 | (0.21) | 18.2 | (5.6) |
| *Quantum Chemical Random Forest* | | *0.54* | *(0.23)* | *17.5* | *(4.8)* |
| *One-hot Random Forest* | | *0.53* | *(0.26)* | *17.7* | *(5.0)* |
| Leave-One-Ligand-Out | | | | | |
| One-hot Encodings | Polynomial | 0.32 | (0.54) | 18.9 | (2.1) |
| Quantum Chemical | RBF | -0.13 | (0.23) | 27.1 | (6.2) |
| Fingerprint: Morgan1 | RBF | 0.30 | (0.90) | 16.6 | (6.4) |
| Tanimoto: MACCS | Polynomial | 0.48 | (0.62) | 14.3 | (7.1) |
| WL | RBF | 0.42 | (0.77) | 14.9 | (6.0) |
| *Quantum Chemical Random Forest* | | *0.32* | *(0.95)* | *15.6* | *(7.9)* |
| *One-hot Random Forest* | | *0.36* | *(1.02)* | *14.1* | *(8.1)* |

‡$R^2$ and RMSE statistics are reported in the format "mean (standard deviation)" for the specified test sets. Performance statistics for the individual test sets can be found in Table B.13, B.16, B.19, B.22, B.25. Baseline random forest models are in *italics*.

design.[33,34] These splits of the data give a low, misrepresentative estimate of model performance, due to the uneven distribution of reaction yield across the test sets (Figure B.13).

### With Activity Ranking

The performance of the yield prediction models built on quantum chemical descriptors, concatenated fingerprints, Tanimoto kernel descriptors and WL kernel descriptors for the additive and aryl halide ranked tests are shown in Table 5.6 and 5.7, Figure 5.4. The random forest model built on quantum chemical descriptors from Doyle *et al.*[34] was included for comparison. The performance of the SVR and random forest models built on one-hot encodings are reported to assess whether the models were fitting any underlying combinatorial structure in the training reactions.

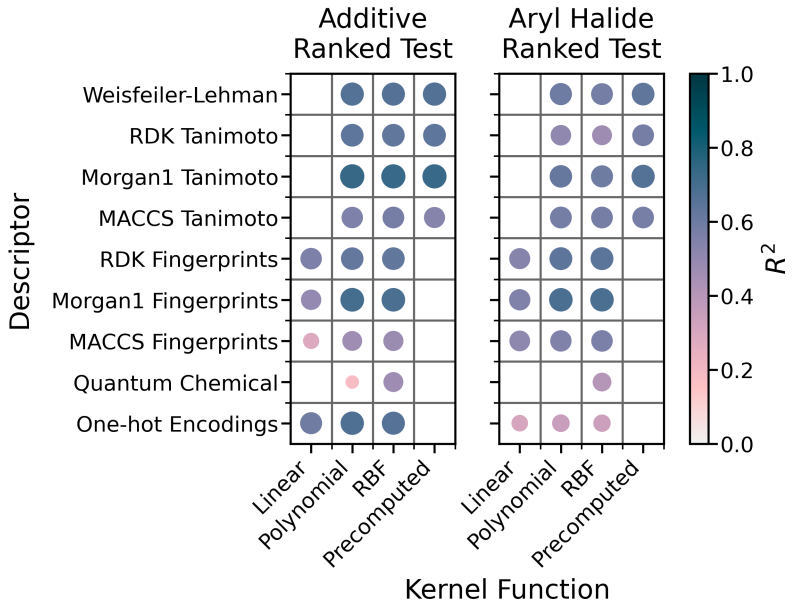



Figure 5.4: Coefficient of Determination ($R^2$) performance comparison of the Support Vector Regression (SVR) models built on one-hot encodings, quantum chemical descriptors, concatenated fingerprints, Tanimoto kernel descriptors and Weisfeiler-Lehman (WL) kernel descriptors with a range of kernels, in the activity ranked tests. Marker size is proportional to $R^2$. Numeric values can be found in Table 5.6 and 5.7.

A few trends in the performance of the algorithms, kernels and descriptors were present in both the additive and aryl halide ranked tests. The SVR models built on one-hot encodings had a better predictive performance than the random forest models built on the same one-hot encodings. Random forest and methods based on decision trees may not handle well the sparsity that one-hot encoding introduces into the dataset. This therefore set a higher baseline for the SVR

Table 5.6: Mean Performance Statistics for the Reaction Yield Prediction Models Built Using the Support Vector Regression (SVR) Algorithm and Baseline Random Forest Models in the Additive Ranked Tests[§]

| Descriptor | Kernel | $R^2$ | | RMSE (%) | |
|---|---|---|---|---|---|
| One-hot Encodings | Linear | 0.59 | (0.05) | 17.4 | (1.5) |
| | Polynomial | **0.68** | (0.05) | **15.4** | (1.5) |
| | RBF | 0.66 | (0.06) | 15.9 | (1.8) |
| Quantum Chemical | Linear | -0.56 | (1.23) | 32.2 | (12.7) |
| | Polynomial | 0.18 | (0.39) | 24.0 | (6.1) |
| | RBF | **0.47** | (0.16) | **19.6** | (3.4) |
| Fingerprints: MACCS | Linear | 0.29 | (0.18) | 22.8 | (2.4) |
| | Polynomial | 0.47 | (0.18) | 19.6 | (3.2) |
| | RBF | 0.48 | (0.15) | 19.4 | (2.7) |
| Fingerprints: Morgan1 | Linear | 0.50 | (0.13) | 19.2 | (2.7) |
| | Polynomial | **0.70** | (0.13) | **14.7** | (3.3) |
| | RBF | 0.69 | (0.14) | 14.9 | (3.6) |
| Fingerprints: RDK | Linear | 0.56 | (0.06) | 18.0 | (1.3) |
| | Polynomial | 0.62 | (0.09) | 16.6 | (1.8) |
| | RBF | 0.63 | (0.07) | 16.5 | (1.6) |
| Tanimoto: MACCS | Polynomial | 0.56 | (0.17) | 17.8 | (3.4) |
| | RBF | 0.58 | (0.16) | 17.5 | (3.4) |
| | Precomputed | 0.54 | (0.17) | 18.3 | (3.3) |
| Tanimoto: Morgan1 | Polynomial | 0.74 | (0.11) | 13.8 | (3.1) |
| | RBF | 0.73 | (0.10) | 13.9 | (2.9) |
| | Precomputed | **0.73** | (0.13) | **13.8** | (3.5) |
| Tanimoto: RDK | Polynomial | 0.64 | (0.05) | 16.4 | (1.3) |
| | RBF | 0.63 | (0.05) | 16.5 | (1.3) |
| | Precomputed | 0.64 | (0.05) | 16.3 | (1.3) |
| WL | Polynomial | 0.67 | (0.17) | 15.4 | (4.0) |
| | RBF | 0.66 | (0.16) | 15.6 | (3.8) |
| | Precomputed | **0.67** | (0.18) | **15.3** | (4.2) |
| Baseline Random Forest Models: | | | | | |
| One-hot | | 0.59 | (0.11) | 17.4 | (2.8) |
| Quantum Chemical | | 0.68 | (0.11) | 15.3 | (3.0) |

[§] $R^2$ and RMSE statistics are reported in the format "mean (standard deviation)" for the specified test. Performance statistics for the individual test sets can be found in Table B.28. For each type of descriptor, the models with the best performance are highlighted in **bold**.

models (additive split: $R^2 < 0.68$, RMSE $> 15.4\%$; aryl halide split: $R^2 < 0.35$, RMSE $> 20.9\%$) than random forest, for model comparison. The one-hot encoding models, in the aryl halide ranked test, have a much lower performance than in the additive ranked test. This could be due to the aryl halide present in the reaction, generally having a larger effect on the reaction yield than the additive (Figure 5.1), base or ligand (Figure B.13) present. There are only four additives that are considered reaction poisons (additives **1**, **4**, **7** and **13**) and hence have a large effect on the reaction yield. One-hot encoding models tend to fit the intrinsic pattern in the combinatorial training data (i.e. the presence/absence of

Table 5.7: Mean Performance Statistics for the Reaction Yield Prediction Models Built Using the Support Vector Regression (SVR) Algorithm and Baseline Random Forest Models in the Aryl Halide Ranked Tests[¶]

| Descriptor | Kernel | $R^2$ | | RMSE (%) | |
|---|---|---|---|---|---|
| One-hot Encodings | Linear | 0.31 | (0.05) | 21.6 | (0.5) |
| | Polynomial | **0.35** | (0.04) | **20.9** | (0.5) |
| | RBF | 0.34 | (0.09) | 21.0 | (0.9) |
| Quantum Chemical | Linear | -505.21 | (875.86) | 336.4 | (549.2) |
| | Polynomial | -4.34 | (7.85) | 47.3 | (41.0) |
| | RBF | **0.41** | (0.14) | **19.9** | (2.6) |
| Fingerprints: MACCS | Linear | 0.52 | (0.01) | 18.1 | (0.8) |
| | Polynomial | 0.55 | (0.17) | 17.2 | (3.7) |
| | RBF | 0.56 | (0.16) | 17.1 | (3.5) |
| Fingerprints: Morgan1 | Linear | 0.55 | (0.05) | 17.5 | (1.5) |
| | Polynomial | **0.69** | (0.05) | **14.6** | (1.7) |
| | RBF | 0.68 | (0.05) | 14.6 | (1.6) |
| Fingerprints: RDK | Linear | 0.54 | (0.05) | 17.7 | (1.3) |
| | Polynomial | 0.64 | (0.11) | 15.4 | (2.4) |
| | Sigmoid | 0.09 | (0.06) | 24.9 | (1.6) |
| Tanimoto: MACCS | Polynomial | 0.58 | (0.05) | 16.8 | (1.2) |
| | RBF | 0.58 | (0.04) | 16.9 | (0.7) |
| | Precomputed | 0.57 | (0.13) | 16.9 | (2.9) |
| Tanimoto: Morgan1 | Polynomial | 0.62 | (0.06) | 16.0 | (1.1) |
| | RBF | 0.59 | (0.06) | 16.5 | (1.1) |
| | Precomputed | **0.67** | (0.06) | **15.0** | (1.4) |
| Tanimoto: RDK | Polynomial | 0.50 | (0.13) | 18.2 | (2.3) |
| | RBF | 0.47 | (0.13) | 18.8 | (2.1) |
| | Precomputed | 0.57 | (0.13) | 16.8 | (2.4) |
| WL | Polynomial | 0.60 | (0.05) | 16.4 | (0.9) |
| | RBF | 0.58 | (0.05) | 16.8 | (0.8) |
| | Precomputed | **0.63** | (0.06) | **15.7** | (1.2) |
| Baseline Random Forest Models: | | | | | |
| One-hot | | -0.04 | (0.33) | 26.2 | (3.3) |
| Quantum Chemical | | 0.20 | (0.17) | 23.2 | (1.5) |

[¶] $R^2$ and RMSE statistics are reported in the format "mean (standard deviation)" for the specified test. Performance statistics for the individual test sets can be found in Table B.31. For each type of descriptor, the models with the best performance are highlighted in **bold**.

each molecule). In the additive ranked test, the models learn the reactivity of the aryl halides, bases and ligands in training and are able to predict the yield of reactions in the test set to a relatively high level. However, in the aryl halide ranked test, the models struggle to extrapolate to unseen aryl halides as they have a larger effect on the reaction yield than the additives, bases and ligands that were fitted in training. This is supported by the following observation. In the aryl halide ranked test, the predicted yields (made by the one-hot encoding model) of the reactions containing the four inhibitory additives, which have a clear effect on lowering the reaction yield, are closer to experimental values than

most of the other additives. If the molecules in the test set have a clear effect on the reaction yield and are also observed in training, the model can learn the reactivity of these molecules and appear to extrapolate well.

The quantum chemical descriptors do not have a linear relationship to the reaction yield, as the linear SVR model predictions show no statistical correlation. The extremely poor performance of the linear SVR model in the aryl halide ranked test is mainly due to the poor predictions of reactions containing 1-bromo-4-methoxybenzene. This aryl bromide has a high $\nu_1$ frequency (1630cm$^{-1}$) in comparison to the other aryl halides (699 to 745cm$^{-1}$). Therefore, when the quantum descriptors were scaled, this $\nu_1$ frequency had an anomalously high value of 59 (usually expect values between $\pm 3$). It is likely that this partially led to predictions of reactions containing 1-bromo-4-methoxybenzene in the range $-2100$ to $-2181$% for the linear SVR model and the prediction of the constant value (15.9.0%) for the quantum chemical SVR model with the RBF kernel applied. Non-linear kernels (polynomial and RBF) were considered, to transform the input data into higher dimensional feature space, where regression could be performed linearly. For the one-hot, quantum chemical and concatenated fingerprint descriptors, the performance of the SVR models implementing the polynomial and RBF kernels were better than linear SVR. The application of non-linear kernels to the WL and Tanimoto kernel descriptors did not substantially improve the performance of the SVR models and therefore are not considered nor discussed further. The SVR algorithm performs better with the structure-based descriptors compared to the quantum chemical descriptors. It is encouraging that the Morgan fingerprints capture enough chemical information that they outperform the quantum chemical descriptors which were adopted by Doyle *et al.*[32] The best combinations of descriptors and kernel functions were the same for both activity-ranked tests (Table 5.8). The top performing SVR model of each descriptor are henceforth referred to as $P$(One-hot), $R$(Quantum), $P$(Fingerprints), Tanimoto and WL.

Table 5.8: Top Performing Support Vector Regression (SVR) Model for each Descriptor in the Activity Ranked Tests

| Descriptor | Kernel Function | SVR Model Name |
|---|---|---|
| One-Hot Encodings | Polynomial | $P$(One-hot) |
| Quantum Chemical | RBF | $R$(Quantum) |
| Fingerprints: Morgan1 | Polynomial | $P$(Fingerprints) |
| Tanimoto: Morgan1 | Precomputed | Tanimoto |
| WL | Precomputed | WL |

In the additive ranked test, the performance of the top SVR model for each descriptor ranged from 0.47 to 0.73 for the $R^2$ and 19.6 to 13.8% for the RMSE

(Table 5.6). According to a Chi-squared ($\chi^2$) test, the top two highest performing models $P$(Fingerprints) and Tanimoto are not statistically, significantly different. The $R^2$, RMSE performance of the respective $P$(Fingerprints) and Tanimoto models was 0.70, 14.7% and 0.73, 13.8%. Under the null hypothesis that the distributions of the residual yield are the same, the $p$-value was calculated as 0.06 (Figure 5.5a, Table 5.9). The $P$(One-hot), $R$(Quantum) and WL models are significantly different with $p$-values less than $10^{-7}$. The SVR models built on one-hot encodings and quantum chemical descriptors have shorter, broader peaks in the distribution of residual yields, meaning larger associated errors (Figure 5.5a). The random forest algorithm learns more from the quantum chemical descriptors ($R^2 = 0.68$, RMSE = 15.3%) than the SVR algorithm ($R^2 \leq 0.47$, RMSE $\geq$ 19.6%).

Table 5.9: Pairwise Chi-squared ($\chi^2$) Results Calculated on the Distributions of Residual Yield Between the Top Performing Support Vector Regression (SVR) models for each descriptor and Random Forest (RF) Baseline

| Model A | Model B | Ranked Test | |
|---|---|---|---|
| | | Additive | Aryl halide |
| $P$(One-hot) | $P$(One-hot) | 1 | 1 |
| | $R$(Quantum) | $3\times10^{-28}$ | $3\times10^{-66}$ |
| | $P$(Fingerprints) | $1\times10^{-23}$ | $3\times10^{-115}$ |
| | Tanimoto | $2\times10^{-36}$ | $2\times10^{-85}$ |
| | WL | $2\times10^{-18}$ | $1\times10^{-69}$ |
| | Quantum RF | $1\times10^{-61}$ | $4\times10^{-44}$ |
| $R$(Quantum) | $R$(Quantum) | 1 | 1 |
| | $P$(Fingerprints) | $1\times10^{-35}$ | $8\times10^{-70}$ |
| | Tanimoto | $8\times10^{-50}$ | $3\times10^{-38}$ |
| | WL | $4\times10^{-24}$ | $1\times10^{-54}$ |
| | Quantum RF | $6\times10^{-34}$ | $2\times10^{-47}$ |
| $P$(Fingerprints) | $P$(Fingerprints) | 1 | 1 |
| | Tanimoto | $6\times10^{-2}$ | $6\times10^{-14}$ |
| | WL | $8\times10^{-7}$ | $5\times10^{-25}$ |
| | Quantum RF | $1\times10^{-18}$ | $9\times10^{-88}$ |
| Tanimoto | Tanimoto | 1 | 1 |
| | WL | $9\times10^{-10}$ | $2\times10^{-14}$ |
| | Quantum RF | $3\times10^{-15}$ | $2\times10^{-88}$ |
| WL | WL | 1 | 1 |
| | Quantum RF | $3\times10^{-14}$ | $3\times10^{-67}$ |
| Quantum RF | Quantum RF | 1 | 1 |

Model performances along the aryl halide dimension were significantly lower than along the additive dimension for the baseline and quantum chemical models (Table 5.7, Figure 5.4). Models built on structure-based descriptors had a similar performance to those in the additive ranked test. There is a large difference in performance between the structure-based descriptors, with an $R^2$ of 0.63 to 0.69,

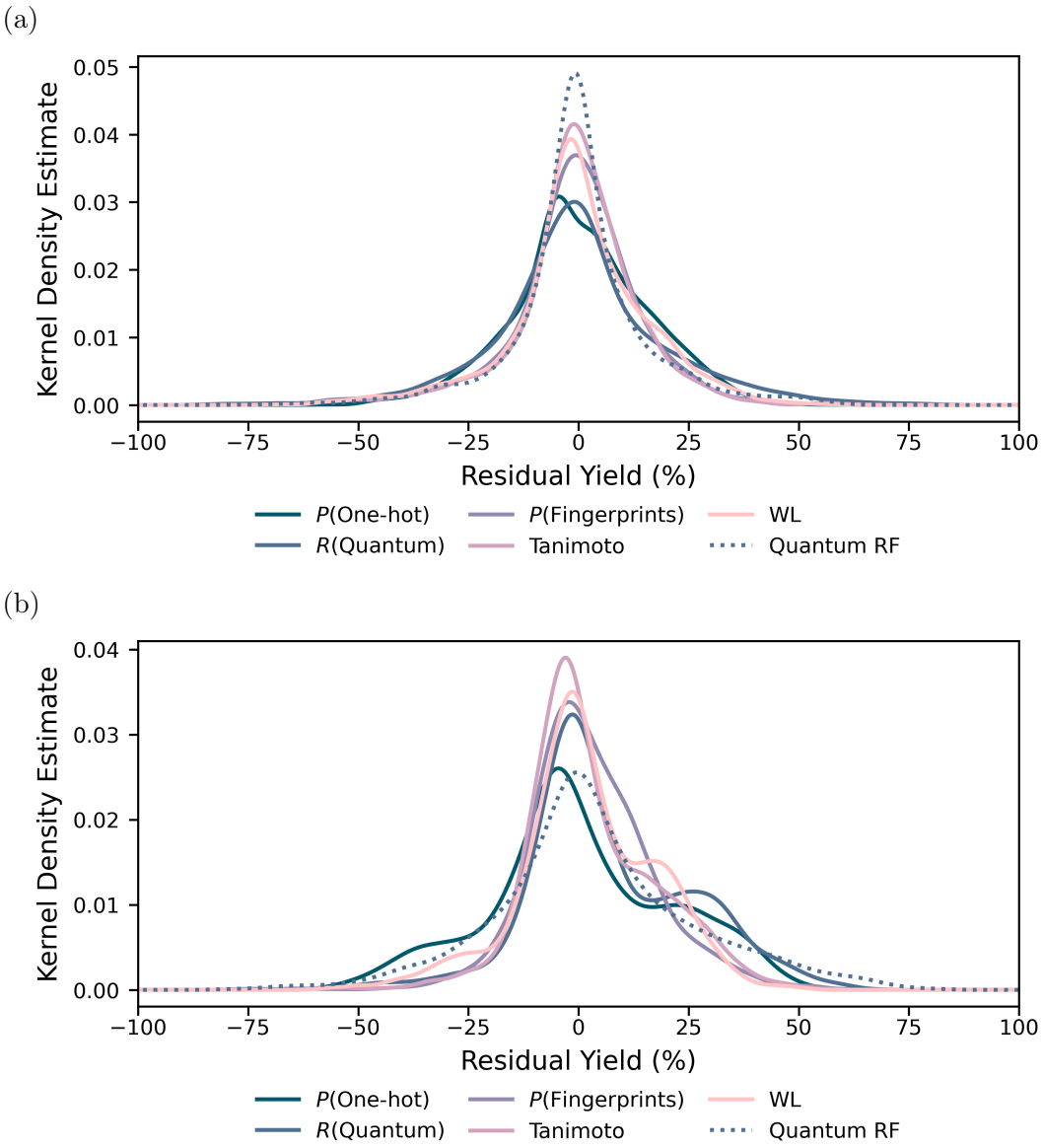
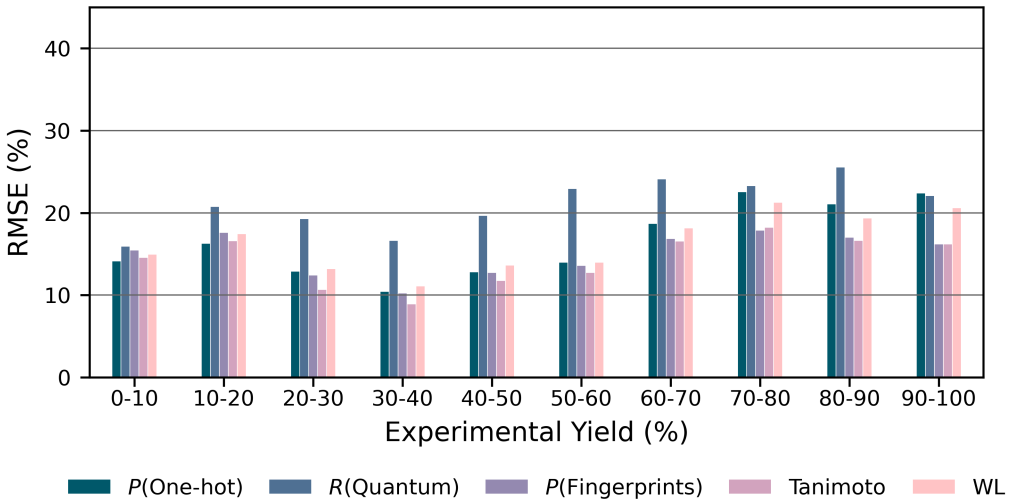



Figure 5.5: Distributions of residual yield for the (a) additive ranked test and (b) aryl halide ranked test. Residual yield was calculated as observed yield minus predicted yield. Numeric details for the test of statistical significance can be found in Table 5.9.

compared to the $R$(Quantum) and $P$(One-hot) models with $R^2$ of 0.41 and 0.35, respectively. The low performance of the quantum chemical and one-hot encoding models suggests that they may only be fitting the intrinsic pattern in the training set and therefore, struggle to extrapolate to the unseen aryl halides. In general, there is an even distribution of residual yield centred around 0% for the top SVR model per descriptor (Figure 5.5b). All models however tend to underpredict the reaction yields of the unseen aryl halides as shown by the smaller, secondary peaks (between 12.5 to 37.5%) in the distribution of residual yield (Figure 5.5b). This is partially due to the under-representation of higher reaction yields (Fig-

ure 5.2), resulting in poorer model performances (Figure 5.6b). This issue is also observed in the additive ranked test to a lesser extent (Figure 5.6a).

(a)




(b)

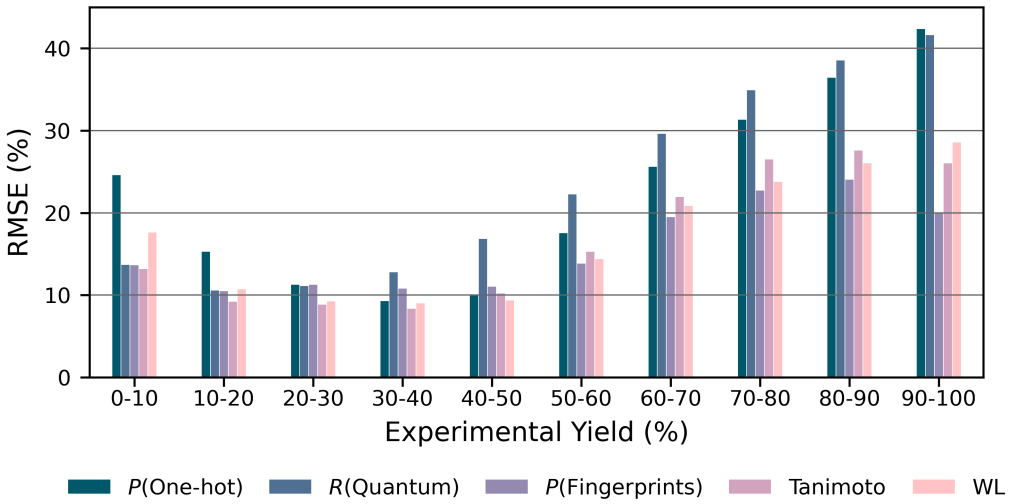



Figure 5.6: Root Mean Squared Error (RMSE) performance against the experimental yield for the (a) additive ranked test and (b) aryl halide ranked test.

### 5.3.3 Domain of Applicability

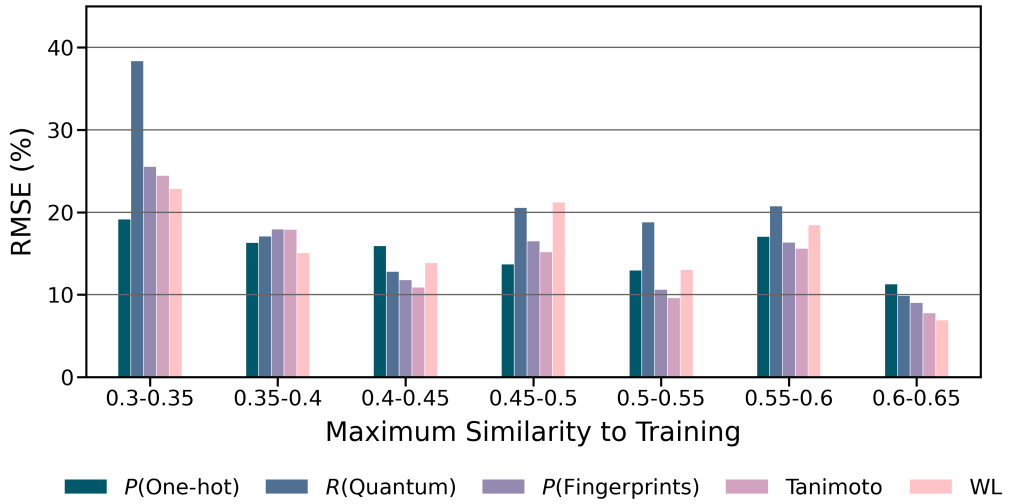
Assessing model performance with respect to maximum similarity to training reactions helps to identify molecules that may be outside the domain of applicability. Maximum similarity to training is defined as the maximum product of pairwise Tanimoto scores (between molecules in the training and test sets) of the reaction components.

In the additive ranked test, the models performed poorly for reactions in the lowest maximum similarity to training interval, 0.30 to 0.35 (Figure 5.7a). These

reactions contain the additives: benzo[*c*]isoxazole (additive **10**) and benzo[*d*]-isoxazole (additive **15**). The performance of the models, considering the additives individually, are generally good for additive **15** (Figure B.15d) but very poor for additive **10** (Figure B.15a). The models overpredict the yield of reactions containing the inhibitory additive **10** and result in negative $R^2$ and high RMSE scores. These reactions may therefore be outside the domain of applicability. Generally, the performance of the models improves with maximum similarity to training (Figure 5.7a), as expected. The models have a high RMSE ($> 15\%$) for the reactions in the maximum similarity to training intervals 0.35 to 0.40 (additive **1** and **14**) and 0.55 to 0.60 (additive **4**, **6** and **9**). This is mainly due to the underprediction of high-yielding reactions, which is a result of the under-representation of higher reaction yields (Figure 5.2 and 5.6a). The structure-based SVR models demonstrate good prediction statistics for reactions with maximum similarity to training greater than 0.35.

For the aryl halide ranked test, there is a slight improvement in the performance statistics as maximum similarity to training increases (Figure 5.7b). The higher-yielding ($> 50\%$) reactions containing ethyl-substituted aryl halides (0.30 to 0.40), 2-halopyridines and 3-iodopyridine (0.45 to 0.50) are underpredicted by the models (Figure B.16), due to the under-representation of higher reaction yields (Figure 5.2 and 5.7b). Reactions containing the trifluoromethyl and methoxy substituted aryl halides, as well as the remaining 3-halopyridines (0.50 to 0.55), are generally predicted well by the models. It is important to consider the $R^2$ and RMSE together when assessing goodness of fit.[245] This is demonstrated in the model performance of reactions containing 1-chloro-4-ethylbenzene (aryl halide **7**) and 1-chloro-4-(trifluoromethyl)benzene (aryl halide **1**). These reactions are low yielding and therefore only cover a small range of reaction yields. While this leads to low $R^2$ scores across all models (Figure B.16), the RMSE scores are good ($\leq 15\%$) for at least half of the models.

(a)




(b)

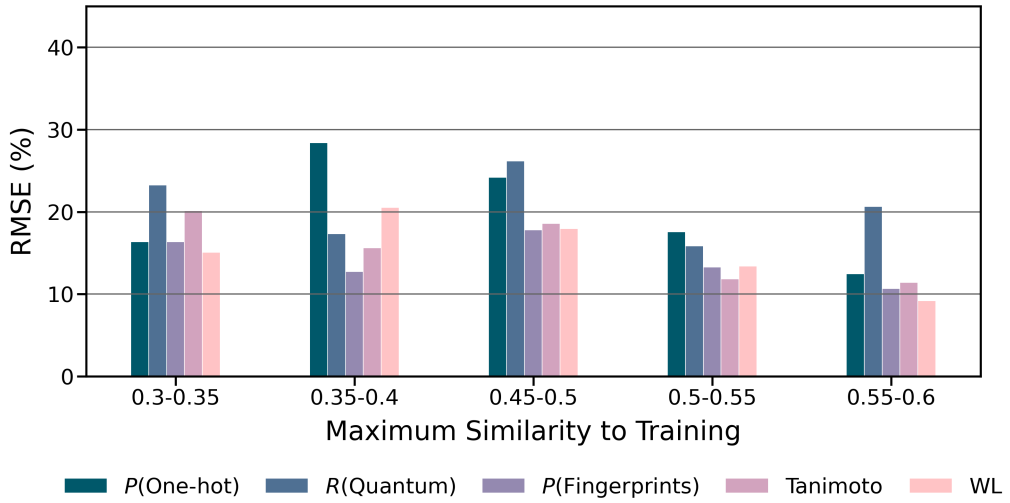



Figure 5.7: Root Mean Squared Error (RMSE) performance against maximum similarity to training for (a) the additive ranked test and (b) the aryl halide ranked test.

### 5.3.4 Predictions of Prospective Reactions

A set of combinatorial reactions was compiled to validate the generalisability of the SVR models, with particular interest along the aryl halide dimension. Although no experiments have been completed, comparing trends in the predictions of reaction yields between models is beneficial. Here, we present predicted yields of the proposed reactions prior to experimentation. The SVR model with the best predictive performance for each descriptor in the aryl halide ranked test was employed: $R$(Quantum), $P$(Fingerprints), Tanimoto, WL and the $P$(One-hot) baseline. The aryl halides in the prospective reactions cover a range of maximum similarity to training scores between 0.15 to 0.60 (Figure 5.8). This excludes the

five aryl halides that are present in the Doyle *et al.* training set, where the maximum similarity to training was 1.00. In the aryl halide ranked test, the models predicted the yield of reactions containing the aryl halide with the lowest maximum similarity to training score (0.30 to 0.35) reasonably well (Figure 5.7b). The models may, however, struggle to extrapolate to the aryl halides in the prospective reactions with maximum similarity scores lower than 0.30 (over half of the unseen aryl halides). These models in the base and ligand leave-one-out tests generally showed comparable correlation to the best kernel-descriptor combinations in these tests (Table B.22 and B.25). The poor performance of the quantum chemical model in these tests indicates that the model is limited and may be unable to extrapolate to unseen bases and ligands.

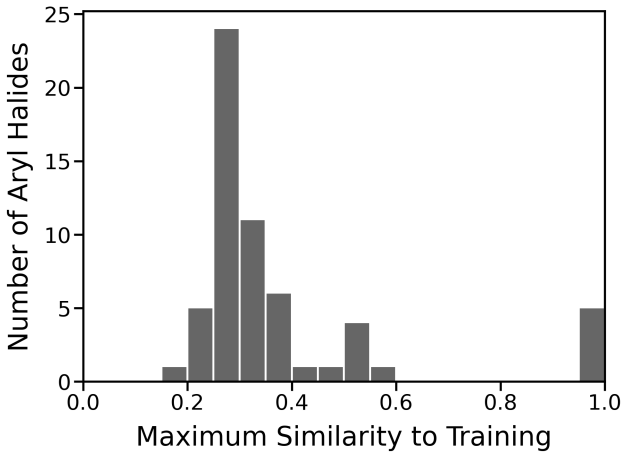



Figure 5.8: Distributions of maximum similarity to training for all prospective reactions. Maximum similarity to training was calculated using the maximum pairwise Tanimoto scores (using the Morgan2 fingerprint) of the aryl halides in the training and test set.

Two tests were designed to investigate the predictive ability of the SVR models identified as the top descriptor-kernel combinations in the aryl halide ranked test. The first test considered all 1416 proposed reactions for the comparison of the structure-based descriptors and one-hot encodings. These descriptors can be applied to any molecule and are quick and easy to calculate. In this test, the Fingerprints: Morgan1-Polynomial, Tanimoto: Morgan1-Precomputed, WL-Precomputed and the One-hot-Polynomial models were trained on the Doyle *et al.* dataset, including additive control reactions (i.e. no additive present); a total of 4135 reactions. The second test only considered a subset of the proposed reactions to compare the quantum chemical descriptors with the structure-based descriptors. The quantum chemical descriptors have a limited application range as they require predefined, key shared atoms to be present for each reaction component. The subset excluded any molecules where quantum chemical descriptors

could not be calculated; this included aryl iodides (see the Supporting Information for further details). This prospective test set contained a total of 882 reactions, a combination of 49 aryl halides, two additives, three bases and three ligands. The SVR models were trained on a subset of 2757 reactions from the Doyle *et al.* dataset, including additive control reactions. The predicted yields of each reaction, calculated in both tests, are shown in Figure B.19 and B.20.

The models built on chemically meaningful descriptors predicted the yield of test reactions that are also present in the training reactions accurately, with $R^2 \geq 0.98$ and RMSE $\leq 4.4\%$ (Figure B.17 and B.18). In both tests, the one-hot encodings model predicted an arbitrary number irrespective of the aryl halide present in the reaction. The predictions were primarily dependent on the type of base and ligand in the reaction (Figure B.22 and B.24). The reaction containing the base DBU and the reactions performed without a catalyst ligand contribute the most towards the broad peak at approximately 35% in the distribution of predicted yield for the subset of proposed reactions (Figure 5.9). The base MTBD and catalyst ligand t-BuXPhos have a broader range of higher yields contributing to the peak around 50%. The same trend was observed when all prospective reactions were considered (Figure B.23 and B.25). There is minimal difference in the distributions of the predicted yield of the reactions containing each base for the chemically meaningful SVR models. The baseline one-hot encodings model was unable to extrapolate to unseen aryl halides from fitting the underlying pattern in the training data. Therefore, it is anticipated that the models built on quantum chemical and structure-based descriptors were learning from chemically meaningful information.

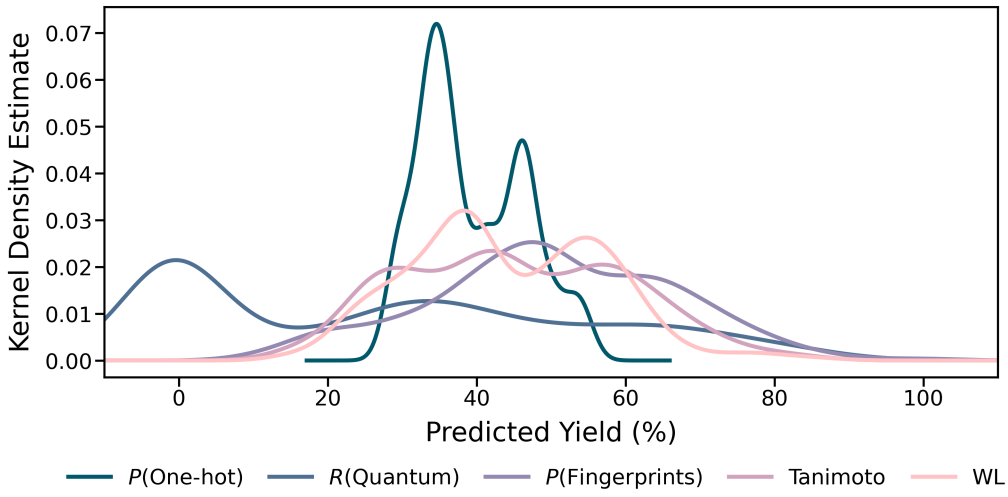



Figure 5.9: Distributions of predicted reaction yield for the subset (882) of validation reactions.

Reactions performed without a catalyst were included in the prospective reactions to evaluate the following synthetic hypothesis; the reactions containing *ortho*-substituted halopyridines are proceeding via an alternative reaction pathway, leading to higher reaction yields. No examples of these reactions were provided in training and therefore, may be beyond the limits of the models. The Quantum SVR model predicted the yield of reactions without the presence of a catalyst to be a negative arbitrary number (-0.64%), irrespective of the aryl halide or additive present in the reactions. Predictions of this negative number suggest these reactions are outside the domain of applicability for the Quantum SVR model. These predictions largely contributed to the distinct peak in the distribution of predicted yield around 0% (Figure 5.9). The structure-based models predicted a smaller range of yields for reactions performed without a catalyst ($\sim$ 30%) compared to the reactions containing a catalyst ($\gtrsim$ 60%, Figure B.24 and B.25). This could indicate a potential limitation in the ability of the structure-based models to predict reactions without a catalyst. The chemically meaningful models predicted similar trends in the reactivity of the catalyst ligands (Figure B.24 and B.25), following the order: BrettPhos (where applicable) < no catalyst < *t*-BuBrettPhos < *t*-BuXPhos.

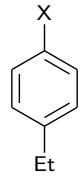
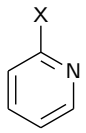
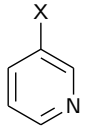
The prospective reactions were designed to validate the applicability of the SVR models to unseen aryl halides that are not present in the training set. The models built on chemically meaningful descriptors predicted higher yields for reactions containing aryl bromides and aryl iodides (where applicable) compared to reactions containing aryl chlorides (Figure B.26 and B.27). Using the reactions containing the *ortho*-halo-substituted isopropylbenzene and *para*-halo-substituted methylpyridazine molecules as examples, there is an increase in mean predicted yield from the chloride to bromide to iodide (Table 5.10). This trend is plausible, as it follows the trend in the training reactions (Figure B.13c). Comparing the mean yield of reactions containing 1-chloro-4-isopropylbenzene ($\sim$ 30% to 45%) with a similar alkyl-substituted aryl halide used in training (1-chloro-4-ethylbenzene, $\sim$ 4%), suggests the models may have overpredicted these reaction yields. Aryl halides with substituents at the ortho position are sterically hindered which could potentially lower the reactivity. As there are no reactions containing *ortho*-substituted aryl halides in the training set, it is possible that the predictions were influenced by the higher yielding *ortho*-substituted pyridines (Table 5.10). The pyridazine molecules contain a nitrogen atom at both the ortho and meta positions. It is interesting that the structure-based models again appear to make predictions based on the higher-yielding *ortho*-substituted pyridines, whereas the quantum chemical model predicts reactivity closer to the lower-yielding *meta*-

substituted pyridines (Table 5.10).

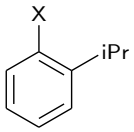
Despite the similar trends between the quantum chemical model and the structure-based models, the predictions are only slightly correlated (Pearson correlation coefficient of $< 0.67$, Figure B.28). The structure-based models are expected to be more robust than the quantum chemical models for extrapolating to unseen chemical entities. The predictions of the structure-based models are well correlated and have a Pearson correlation coefficient of $> 0.83$ (Figure B.29 and B.30).

Table 5.10: Mean Experiment Yields of Aryl Halides in the Training Set (Top) and Mean Predicted Yields of Aryl Halides in the Prospective Reactions (Bottom)‖

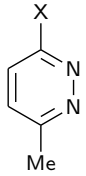
**Mean Experimental Yields (%) of Aryl Halides in the Training Set**

| | Cl | Br | I |
|---|---|---|---|
|  | 3.9 (3.8) | 43.5 (24.6) | 52.6 (24.2) |
|  | 44.1 (26.8) | 53.3 (26.5) | 59.3 (26.6) |
|  | 14.9 (16.2) | 43.9 (29.1) | 52.3 (29.0) |

**Mean Predicted Yields (%) of Aryl Halides in the Prospective Reactions**



| | Cl | | Br | | I | |
|---|---|---|---|---|---|---|
| | Subset | All | Subset | All | Subset | All |
| $P$(One-hot) | 42.4 (6.7) | 47.0 (7.7) | 42.4 (6.7) | 47.0 (7.7) | - | 47.0 (7.7) |
| $R$(Quantum) | 29.6 (4.4) | - | 73.3 (4.7) | - | - | - |
| $P$(Fingerprints) | 36.0 (4.2) | 33.4 (8.6) | 67.0 (4.1) | 55.3 (10.6) | - | 63.2 (8.9) |
| Tanimoto | 35.0 (3.7) | 34.9 (7.0) | 56.1 (3.8) | 53.1 (9.3) | - | 59.4 (8.0) |
| WL | 41.8 (3.8) | 42.8 (8.2) | 55.4 (3.7) | 55.0 (10.3) | - | 56.6 (9.0) |



| | Cl | | Br | | I | |
|---|---|---|---|---|---|---|
| | Subset | All | Subset | All | Subset | All |
| $P$(One-hot) | 42.4 (6.7) | 47.0 (7.7) | 42.4 (6.7) | 47.0 (7.7) | - | 47.0 (7.7) |
| $R$(Quantum) | -4.9 (3.6) | - | 13.7 (2.2) | - | - | - |
| $P$(Fingerprints) | 45.3 (4.4) | 39.8 (12.4) | 61.2 (4.1) | 57.2 (9.7) | - | 63.2 (7.3) |
| Tanimoto | 41.0 (3.7) | 39.0 (8.6) | 58.3 (3.8) | 55.0 (9.1) | - | 60.7 (7.6) |
| WL | 49.8 (4.2) | 48.9 (10.2) | 57.3 (4.0) | 56.2 (10.5) | - | 57.6 (9.6) |

‖Experimental and predicted yields are reported in the format "mean (standard deviation)". Reactions performed without a ligand were excluded.

## 5.4 Conclusions

Anticipating reaction yield is a complex task which requires experimental verification, resources, and time. A tool which provides confidence that a reaction will produce sufficient yield would be valuable to synthetic chemists. In this chapter, we developed machine learning models to predict the yield of unexplored Buchwald-Hartwig reactions. The performance of SVR models was assessed along each reaction component of the Buchwald-Hartwig combinatorial dataset. We demonstrated that topological descriptors are a suitable input to an SVR model to predict yields of combinatorial reaction data.

Poorer model performance in the tests designed without activity ranking highlighted the importance of accounting for the distribution of reaction yields in the training and test sets. Only a few examples of bases and catalyst ligands are present in the combinatorial dataset. In leave-one-base-out and leave-one-ligand-out experiments, we investigated how well the SVR model could predict unknown bases and ligands given only a few instances in training. The moderate performances of the SVR models in both leave-one-out tests suggest that the models might benefit from training on a broader range of bases and ligands.

Out-of-sample tests were constructed using activity ranking. These tests provided a more reliable evaluation of the ability of the models to predict the yield of reactions containing unseen additives and aryl halides. The SVR models built on structure-based descriptors were closely compared to those built on quantum chemical calculations. The SVR models built on structure-based descriptors demonstrated good prediction statistics in each test. Specifically, those employing molecular fingerprints surpassed the models employing other descriptors consistently, demonstrating the robustness of molecular fingerprints. In the aryl halide ranked test, the models built on the Morgan 1 fingerprint achieved an $R^2$of 0.69 and an RMSE of 14.6%, significantly better than the quantum chemical random forest model, which achieved an $R^2$ of 0.20 and RMSE of 23.2%. The applicability, ease and quickness of calculating molecular fingerprints make them particularly attractive (Table 5.11).

The aryl halide ranked test revealed the optimum kernel applied to each descriptor type in the SVR model: quantum chemical calculations, RBF; concatenated fingerprints, polynomial; Tanimoto and WL kernel descriptors, no additional kernel. The SVR models with these descriptor-kernel combinations were subject to an external examination against unfamiliar reactions. We designed the external examination procedure to imitate how medicinal chemists may use yield prediction models on combinatorial data. Buchwald-Hartwig reactions were carefully

Table 5.11: Comparison of the Molecular Descriptors used in this Study[d]

| Descriptor | Speed | Applicability to molecules | Generalisability | |
|---|---|---|---|---|
| | | | Additive | Aryl Halide |
| Quantum Chemical | + | Subset | + | + |
| Fingerprints | +++ | All | ++++ | ++++ |
| Tanimoto Kernel | ++ | All | ++++ | ++++ |
| WL Kernel | ++ | All | +++ | +++ |
| One-hot Encodings | ++++ | All | ++ | ++ |

[d] Speed and generalisability are ranked from poor (+) to good (++++). The ranking of generalisability refers to the performance of the top SVR model for each descriptor.

curated with the assistance of medicinal chemists to propose a set that varied in their similarity to the training reactions, allowing us to study the limits of the models' generalisability.

Despite not performing the high-throughput experiments, we reported and compared the predicted yields of the proposed Buchwald-Harwig reactions. We observed similar trends in the reactivity of the molecules along each reaction component throughout the chemically meaningful SVR models. The reaction yields predicted by the structure-based models were reasonably correlated. We anticipate, based on the performances of the models in the preceding tests and the analysis of the predicted yields, that the structure-based models may extrapolate better than the quantum chemical models. The SVR models used to predict the prospective Buchwald-Hartwig reactions and instructions on their use are available on GitHub (`https://github.com/alexehaywood/yield_prediction`). The chemically meaningful models should be used with caution since they may struggle to extrapolate to unseen aryl halides with a similarity to training score of less than 0.30 and reactions without a catalyst. It may be beneficial to account for the scope of applicability before utilising the models to offer confidence in the model's predictions.

The experimental yields of the proposed reactions will be determined using high-throughput experimentation. Analysis of the errors in the yield predictions for the external test set is required to assess the scope and limits of the SVR models. In the future, it would be interesting to explore the transferability of the structure-based SVR models to different reaction types or alternative regression-related problems. The approach presented could also be applied to larger datasets when they become publicly available.

This chapter demonstrates the applicability of computationally less demanding structure-based descriptors in predicting reaction yield. The SVR models learnt from a relatively small (a few thousand instances) combinatorial dataset, prov-

ing their use in facilitating chemical synthesis and optimising reaction conditions.

# Chapter 6

# Concluding Remarks

Machine learning algorithms are a valuable prediction tool in drug discovery and development. Synthetic chemists can benefit from Computer-Aided Synthesis Planning (CASP) tools. They enhance workflows and productivity while shortening timelines. Knowing which reactions are likely to succeed minimises the number of experiments undertaken, lowering the number of chemicals used and the cost of the experiments. The incorporation of Safety, Environmental, Legal, Economics, Control and Throughput (SELECT) criteria into CASP tools aids in the development of greener and more sustainable reactions. This thesis reviewed contemporary approaches to CASP before focusing specifically on predicting reaction yield.

Contemporary approaches to forward reaction prediction and retrosynthesis can be divided into template-based and template-free methods. Template-free methods focus on selecting templates from a predefined library and applying them to the input structures to generate a ranked list of chemical reactions. Although this approach is interpretable, its applicability is limited by the library of reaction templates. These models cannot extrapolate to chemistries outside the library or discover or predict new chemistries. Template-free methods consist of semi-template-based and machine translation techniques. Both techniques take advantage of machine learning algorithms and do not require a predefined library of templates. Semi-template-based strategies follow the same input-to-synthon-to-ouput workflow as the template-based methods, enabling interpretability. They differ in how the templates are defined. Semi-template-based approaches generate templates on demand from a database. Machine translation methods follow a direct input-to-product workflow. Although they implement "black-box" algorithms resulting in limited interpretability, they are state-of-the-art in both

forward reaction prediction and retrosynthetic analysis.

Research into developing tools for the reaction yield prediction task is emerging. The pioneering work of the Doyle group focused on a random forest model built on quantum chemical descriptors.[32] In Chapter 4, regression algorithms for predicting reaction yield were evaluated, including linear, tree-based, and Support Vector Regression (SVR) methods. The regression algorithms were built on structure-based and quantum chemical descriptors. Preliminary cross-validation tests demonstrated that the performance of the non-linear SVR algorithms, implementing the polynomial and Gaussian Radial Basis Function (RBF) kernels, was comparable to that of the quantum chemical random forest model. The SVR-polynomial and SVR-RBF models built on quantum chemical descriptors both had Coefficient of Determination ($R^2$) values of 0.90 and Root Mean Squared Error (RMSE) values of 8.3% and 8.4%, respectively, compared to Doyle's random forest model, which had an $R^2$ value of 0.93 and RMSE value of 7.2%.

The pioneering work by the Doyle group was criticised for misrepresenting model performance in an out-of-sample test.[33] When designing the training and test sets, the distribution of reaction yield was not considered. As a result, the reported performance metrics were not a reliable representation of generalisability. In a technical response, the Doyle group redesigned the test sets using activity ranking along the additive dimension.[34] Chapter 5 investigated the generalisability of SVR models built on structure-based and quantum chemical descriptors using more rigorous testing. Out-of-sample tests were designed using activity ranking along each dimension of the Buchwald-Hartwig dataset. SVR models built on structure-based descriptors demonstrated good performance statistics and outperformed the SVR and random forest models built on quantum chemical descriptors. The top-performing SVR models from the aryl halide ranked test were subjected to further examination. The top descriptor-kernel combinations were quantum chemical-RBF, Morgan1 fingerprints-polynomial, Morgan1 Tanimoto kernel, and WL kernel. The Morgan1 fingerprints-polynomial SVR model performed best with an $R^2$ value of 0.69 and an RMSE value of 14.6%; this was significantly better than Doyle's quantum chemical random forest model, which had an $R^2$ value of 0.20 and an RMSE value of 23.2%. Prospective Buchwald-Hartwig reactions were compiled to examine the applicability of the SVR models to unseen molecules, with a particular interest in aryl halides. The predicted reaction yields for this set of reactions were presented and compared prior to experimentation.

The work undertaken in this thesis has contributed to the early development of tools for predicting reaction yield. We have demonstrated that limited combi-

natorial data, commonly generated in medicinal chemistry, can be used to build local models. Medicinal chemistry focuses on the design and synthesis of analogous compounds. With further research, localised yield prediction models can be developed and deployed in the workflow of medicinal chemists. Knowing whether a chemical reaction will fail or have a high yield prior to experimentation would be incredibly beneficial. We have also shown for this particular yield prediction problem that structure-based descriptors incorporate enough information to outperform quantum chemical descriptors. The applicability, ease, and quickness of generating molecular fingerprints make them significantly more appealing than calculating quantum chemical properties.

Predicting reaction yield is a relatively new area of CASP and has been explored much less than forward reaction prediction and retrosynthetic planning. This is partially due to the lack of curated reaction data which reports reaction yield. The current public benchmarking dataset is the high throughput combinatorial dataset published by the Doyle group, which only contains around 4,000 Buchwald- Hartwig reactions. Using this dataset to develop yield prediction tools restricts them to a single reaction class and a relatively low amount of data. Although deep learning has shown considerable success in many fields, including chemistry and CASP, the scarcity of data presents issues with implementing these models. Deep learning models are prone to overfitting and require many more training examples than are currently available to the public. Therefore, this work focused on simpler non-linear models with proven success in Quantitative Structure-Activity Relationship (QSAR) evaluation.

Immediate future work on this project would entail performing the prospective reactions experimentally using high throughput experimentation. The experimental reaction yields can be used to validate and determine the limits of the top SVR models. The data would also be valuable to the community for further development of reaction yield prediction tools. Integrating yield prediction tools into synthesis planning workflows is a broader vision for the future of the field. One approach may be to develop a foundation for localised models, refine the model on a specific reaction class, and update the training data to integrate reaction yield data as it is generated from high throughput experiments. Another approach is to initially focus on curating a larger dataset including many reaction classes. A large dataset of numerous reactions allows a global model to be developed. Integrating a global model into computer-aided retrosynthesis software would provide additional quantitative suggestions along with the synthetic route. This study provides foundation work and shows the potential to develop such localised and global tools. Higher-quality datasets and focused research into

reaction-specific descriptors are likely to progress the area of yield prediction. There is potential to improve the accuracy and reliability of predicting the yield of chemical reactions to reduce the timeline of chemical syntheses.

# Bibliography


[1] R. Dorel, C. P. Grugel and A. M. Haydl, *Angew. Chem. Int. Ed.*, 2019, **58**, 17118–17129.

[2] D. T. Ahneman and A. G. Doyle, *Github: rxnpredict*, `https://github.com/doylelab/rxnpredict`, (accessed July 2023).

[3] N. C. for Biotechnology Information, *PubChem Compound Summary for CID 1983, Acetaminophen*, 2023, `https://pubchem.ncbi.nlm.nih.gov/compound/Acetaminophen`, (accessed July 2023).

[4] J. Hughes, S. Rees, S. Kalindjian and K. Philpott, *Br. J. Pharmacol.*, 2011, **162**, 1239–1249.

[5] J. A. DiMasi, H. G. Grabowski and R. W. Hansen, *J. Health Econ.*, 2016, **47**, 20–33.

[6] M. Kesik-Brodacka, *Biotechnol. Appl. Biochem.*, 2018, **65**, 306–322.

[7] O. Gilan, I. Rioja, K. Knezevic, M. J. Bell, M. M. Yeung, N. R. Harker, E. Y. N. Lam, C. wa Chung, P. Bamborough, M. Petretich, M. Urh, S. J. Atkinson, A. K. Bassil, E. J. Roberts, D. Vassiliadis, M. L. Burr, A. G. S. Preston, C. Wellaway, T. Werner, J. R. Gray, A.-M. Michon, T. Gobbetti, V. Kumar, P. E. Soden, A. Haynes, J. Vappiani, D. F. Tough, S. Taylor, S.-J. Dawson, M. Bantscheff, M. Lindon, G. Drewes, E. H. Demont, D. L. Daniels, P. Grandi, R. K. Prinjha and M. A. Dawson, *Science*, 2020, **368**, 387–394.

[8] A. L. Hopkins and C. R. Groom, *Nat. Rev. Drug Discov.*, 2002, **1**, 727–730.

[9] A. C. Cheng, R. G. Coleman, K. T. Smyth, Q. Cao, P. Soulard, D. R. Caffrey, A. C. Salzberg and E. S. Huang, *Nat. Biotechnol.*, 2007, **25**, 71–75.

[10] L. R. Vidler, N. Brown, S. Knapp and S. Hoelder, *J. Med. Chem.*, 2012, **55**, 7346–7359.

[11] C. P. Tinworth and R. J. Young, *J. Med. Chem.*, 2020, **63**, 10091–10108.

[12] A. T. Plowright, C. Johnstone, J. Kihlberg, J. Pettersson, G. Robb and R. A. Thompson, *Drug Discov. Today*, 2012, **17**, 56–62.

[13] C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Deliv. Rev.*, 2001, **46**, 3–26.

[14] M. J. Armstrong and M. S. Okun, *JAMA*, 2020, **323**, 548.

[15] S. R. W. Stott, R. K. Wyse and P. Brundin, *Front. Neurosci.*, 2021, **15**, 232.

[16] A. M. Turing, *Mind*, 1950, **LIX**, 433–460.

[17] M. Campbell, A. Hoane and F. hsiung Hsu, *Artif. Intell.*, 2002, **134**, 57–83.

[18] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel and D. Hassabis, *Nature*, 2016, **529**, 484–489.

[19] L. Huang, H. Zhang, D. Deng, K. Zhao, K. Liu, D. A. Hendrix and D. H. Mathews, *Bioinformatics*, 2019, **35**, i295–i304.

[20] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.

[21] K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. A. A. Kohl, A. Potapenko, A. J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reiman, S. Petersen, A. W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper and D. Hassabis, *Nature*, 2021, **596**, 590–596.

[22] R. Hausen and B. E. Robertson, *Astrophys. J. Suppl. Ser.*, 2020, **248**, 20.

[23] Z. J. Baum, X. Yu, P. Y. Ayala, Y. Zhao, S. P. Watkins and Q. Zhou, *J. Chem. Inf. Model.*, 2021, **61**, 3197–3212.

[24] W. Hu, Y. Liu, X. Chen, W. Chai, H. Chen, H. Wang and G. Wang, *IEEE Trans. Artif. Intell.*, 2023, 1–21.

[25] A. Volkamer, S. Riniker, E. Nittinger, J. Lanini, F. Grisoni, E. Evertsson, R. Rodríguez-Pérez and N. Schneider, *Artificial Intelligence in the Life Sciences*, 2023, **3**, 100056.

[26] Z. Tu, T. Stuyver and C. W. Coley, *Chem. Sci.*, 2023, **14**, 226–244.

[27] M. Butters, D. Catterick, A. Craig, A. Curzons, D. Dale, A. Gillmore, S. P. Green, I. Marziano, J.-P. Sherlock and W. White, *Chem. Rev.*, 2006, **106**, 3002–3027.

[28] O. Engkvist, P.-O. Norrby, N. Selmi, Y. hong Lam, Z. Peng, E. C. Sherer, W. Amberg, T. Erhard and L. A. Smyth, *Drug Discov. Today*, 2018, **23**, 1203–1218.

[29] J. B. O. Mitchell, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2014, **4**, 468–481.

[30] E. J. Griffen, A. G. Dossetter, A. G. Leach and S. Montague, *Drug Discov. Today*, 2018, **23**, 1373–1384.

[31] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer and S. Zhao, *Nat. Rev. Drug Discov.*, 2019, **18**, 463–477.

[32] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.

[33] K. V. Chuang and M. J. Keiser, *Science*, 2018, **362**, eaat8603.

[34] J. G. Estrada, D. T. Ahneman, R. P. Sheridan, S. D. Dreher and A. G. Doyle, *Science*, 2018, **362**, eaat8763.

[35] A. L. Haywood, J. Redshaw, T. Gärtner, A. Taylor, A. M. Mason and J. D. Hirst, in *Machine Learning in Chemistry: The Impact of Artificial Intelligence*, ed. H. Cartwright, The Royal Society of Chemistry, 2020, ch. 7, pp. 169–194.

[36] A. L. Haywood, J. Redshaw, M. W. D. Hanson-Heine, A. Taylor, A. Brown, A. M. Mason, T. Gärtner and J. D. Hirst, *J. Chem. Inf. Model.*, 2022, **62**, 2077–2092.

[37] D. Probst, P. Schwaller and J.-L. Reymond, *Digital Discovery*, 2022, **1**, 91–97.

[38] Y. Mo, Y. Guan, P. Verma, J. Guo, M. E. Fortunato, Z. Lu, C. W. Coley and K. F. Jensen, *Chemical Science*, 2021, **12**, 1469–1478.

[39] X. Wang, Y. Qian, H. Gao, C. Coley, Y. Mo, R. Barzilay and K. F. Jensen, *Chem. Sci.*, 2020, **11**, 10959–10972.

[40] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer New York, 2nd edn., 2009.

[41] A. J. Smola and B. Schölkopf, *Stat. Comput.*, 2004, **14**, 199–222.

[42] B. Schölkopf, *Ph.D. thesis*, TU Berlin, 1997.

[43] J. L. Bentley, *Commun. ACM*, 1975, **18**, 509–517.

[44] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification And Regression Trees*, Routledge, 1984, pp. 1–358.

[45] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley, 2009, vol. 2, pp. 1–252.

[46] B. J. Braams and J. M. Bowman, *Int. Rev. Phys. Chem.*, 2009, **28**, 577–606.

[47] N. L. Biggs, E. K. Lloyd and R. J. Wilson, *Graph theory 1736-1936*, Oxford University Press, 1986.

[48] R. Diestel, *Graph Theory*, Springer Berlin, Heidelberg, 2017, vol. 173.

[49] J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1273–1280.

[50] G. A. Landrum, *RDKit: Open-Source Cheminformatics*, `http://www.rdkit.org`, (accessed July 2023).

[51] H. L. Morgan, *Journal of Chemical Documentation*, 1965, **5**, 107–113.

[52] *Daylight Theory: Fingerprints*, `https://www.daylight.com/dayhtml/doc/theory/theory.finger.html`, (accessed July 2023).

[53] D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.

[54] D. Weininger, A. Weininger and J. L. Weininger, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 97–101.

[55] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn and K. M. Borgwardt, *J. Mach. Learn. Res.*, 2011, **12**, 2539–2561.

[56] D. C. I. Systems, *Daylight Theory Manual, Chapter 5: A Reaction Transform Language*, `https://www.daylight.com/dayhtml/doc/theory/index.pdf`, (accessed July 2023).

[57] M. A. Johnson and G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, John Wiley and Sons, 1990.

[58] D. Bajusz, A. Rácz and K. Héberger, *J. Cheminformatics*, 2015, **7**, 20.

[59] D. J. Rogers and T. T. Tanimoto, *Science*, 1960, **132**, 1115–1118.

[60] E. J. Corey, *Pure Appl. Chem.*, 1967, **14**, 19–38.

[61] E. J. Corey, *Angew. Chem. Int. Ed.*, 1991, **30**, 455–465.

[62] E. J. Corey and W. T. Wipke, *Science*, 1969, **166**, 178–192.

[63] E. J. Corey, *Quarterly Reviews, Chemical Society*, 1971, **25**, 455.

[64] E. J. Corey and X.-M. Cheng, *The Logic of Chemical Synthesis*, John Wiley and Sons, 1989.

[65] E. J. Corey, W. T. Wipke, R. D. Cramer and W. J. Howe, *J. Am. Chem. Soc.*, 1972, **94**, 421–430.

[66] D. A. Pensak and E. J. Corey, in *Computer-Assisted Organic Synthesis*, ed. W. T. Wipke and W. J. Howe, American Chemical Society, Washingtin, D. C., USA, 1977, vol. 61, ch. 1, pp. 1–32.

[67] E. J. Corey, A. K. Long and S. D. Rubenstein, *Science*, 1985, **228**, 408–418.

[68] W.-D. Ihlenfeldt and J. Gasteiger, *Angew. Chem. Int. Ed.*, 1996, **34**, 2613–2633.

[69] M. H. Todd, *Chem. Soc. Rev.*, 2005, **34**, 247.

[70] A. Cook, A. P. Johnson, J. Law, M. Mirzazadeh, O. Ravitz and A. Simon, *WIREs Computational Molecular Science*, 2012, **2**, 79–107.

[71] O. Ravitz, *Drug Discov. Today Technol.*, 2013, **10**, e443–e449.

[72] W. A. Warr, *Mol. Inform.*, 2014, **33**, 469–476.

[73] I. I. Baskin, T. I. Madzhidov, I. S. Antipin and A. A. Varnek, *Russian Chem. Rev.*, 2017, **86**, 1127–1156.

[74] Z. Wang, W. Zhang and B. Liu, *Chinese J. Chem.*, 2021, **39**, 3127–3143.

[75] H. Dai, C. Li, C. W. Coley, B. Dai and L. Song, Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2019, pp. 8872–8882.

[76] S. Chen and Y. Jung, *JACS Au*, 2021, **1**, 1612–1620.

[77] C. Shi, M. Xu, H. Guo, M. Zhang and J. Tang, Proceedings of the 37th International Conference on Machine Learning, 2020, pp. 8818–8827.

[78] V. R. Somnath, C. Bunne, C. W. Coley, A. Krause and R. Barzilay, Proceedings of the 35th International Conference on Neural Information Processing Systems, 2021.

[79] S. Yang, Y. Wang and X. Chu, 2020, arXiv:2002.07526.

[80] P. Schwaller, T. Gaudin, D. Lányi, C. Bekas and T. Laino, *Chem. Sci.*, 2018, **9**, 6091–6098.

[81] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.

[82] I. V. Tetko, P. Karpov, R. V. Deursen and G. Godin, *Nat. Commun.*, 2020, **11**, 5575.

[83] C. W. Coley, W. H. Green and K. F. Jensen, *Acc. Chem. Res.*, 2018, **51**, 1281–1289.

[84] A. F. de Almeida, R. Moreira and T. Rodrigues, *Nat. Rev. Chem.*, 2019, **3**, 589–604.

[85] T. J. Struble, J. C. Alvarez, S. P. Brown, M. Chytil, J. Cisar, R. L. Desjarlais, O. Engkvist, S. A. Frank, D. R. Greve, D. J. Griffin, X. Hou, J. W. Johannes, C. Kreatsoulas, B. Lahue, M. Mathea, G. Mogk, C. A. Nicolaou, A. D. Palmer, D. J. Price, R. I. Robinson, S. Salentin, L. Xing, T. Jaakkola, W. H. Green, R. Barzilay, C. W. Coley and K. F. Jensen, *J. Med. Chem.*, 2020, **63**, 8667–8682.

[86] Z. Wang, W. Zhao, G. Hao and B. Song, *Org. Chem. Front.*, 2021, **8**, 812–824.

[87] Y. Sun and N. V. Sahinidis, *Curr. Opin. Chem. Eng.*, 2022, **35**, 100721.

[88] Z. Zhong, J. Song, Z. Feng, T. Liu, L. Jia, S. Yao, T. Hou and M. Song, 2023, arXiv:2301.05864.

[89] M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell and B. A. Grzybowski, *Angew. Chem. Int. Ed.*, 2005, **44**, 7263–7269.

[90] K. J. M. Bishop, R. Klajn and B. A. Grzybowski, *Angew. Chem. Int. Ed.*, 2006, **45**, 5348–5354.

[91] K. Molga, P. Dittwald and B. A. Grzybowski, *Chem*, 2019, **5**, 460–473.

[92] W. Watson, *Org. Process Res. Dev.*, 2012, **16**, 1877–1877.

[93] M. C. Swain and J. M. Cole, *J. Chem. Inf. Model.*, 2016, **56**, 1894–1904.

[94] C. D. Christ, M. Zentgraf and J. M. Kriegl, *J. Chem. Inf. Model.*, 2012, **52**, 1745–1756.

[95] D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.

[96] D. M. Lowe, *Chemical reactions from US patents (1976-Sep2016)*, 2017, `https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873`, (accessed July 2023).

[97] *Reaxys*, `https://www.reaxys.com`, (accessed July 2023).

[98] *CAS SciFinder$^n$ - Chemical Compound Database*, `https://www.cas.org/solutions/cas-scifinder-discovery-platform/cas-scifinder-n`, (accessed July 2023).

[99] J. Mayfield, I. Lagerstedt and I. Lagerstedt, *NextMove Software Pistachio*, 2023, `https://www.nextmovesoftware.com/pistachio.html`, (accessed July 2023).

[100] J. Mayfield, I. Lagerstedt and R. Sayle, *Pistachio "Fantastic reactions and how to use them"*, 2021, `https://nextmovesoftware.com/talks/Mayfield_Pistachio_NIHReactions_202105.pdf`, (accessed July 2023).

[101] *Reaxys Fact Sheet*, 2019, `https://www.elsevier.com/__data/assets/pdf_file/0005/91616/Reaxys-Fact-Sheet-2019-web.pdf`, (accessed July 2023).

[102] *CAS Reactions*, `https://www.cas.org/cas-data/cas-reactions`, (accessed July 2023).

[103] S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen and C. W. Coley, *J. Am. Chem. Soc.*, 2021, **143**, 18820–18826.

[104] M. H. S. Segler and M. P. Waller, *Chem. Eur. J.*, 2017, **23**, 6118–6128.

[105] M. H. S. Segler and M. P. Waller, *Chem. Eur. J.*, 2017, **23**, 5966–5971.

[106] M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.

[107] J. S. Schreck, C. W. Coley and K. J. M. Bishop, *ACS Cent. Sci.*, 2019, **5**, 970–981.

[108] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2018, **4**, 1465–1476.

[109] A. Toniato, P. Schwaller, A. Cardinale, J. Geluykens and T. Laino, *Nat. Mach. Intell.*, 2021, **3**, 485–494.

[110] S. G. Higgins, A. A. Nogiwa-Valdez and M. M. Stevens, *Nat. Protoc.*, 2022, **17**, 179–189.

[111] C. L. Bird, C. Willoughby and J. G. Frey, *Chem. Soc. Rev.*, 2013, **42**, 8157–8175.

[112] S. Kanza, C. Willoughby, N. Gibbins, R. Whitby, J. G. Frey, J. Erjavec, K. Zupančič, M. Hren and K. Kovač, *J. Cheminformatics*, 2017, **9**, 31.

[113] S. J. Coles, J. G. Frey, C. L. Bird, R. J. Whitby and A. E. Day, *J. Cheminformatics*, 2013, **5**, 52.

[114] N. H. Goddard, R. Macneil and J. Ritchie, *Autom. Exp.*, 2009, **1**, 4.

[115] F. Rudolphi and L. J. Goossen, *J. Chem. Inf. Model.*, 2012, **52**, 293–301.

[116] A. A. Lee, Q. Yang, V. Sresht, P. Bolgar, X. Hou, J. L. Klug-McLeod and C. R. Butler, *Chem. Commun.*, 2019, **55**, 12152–12155.

[117] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73–76.

[118] D. M. Lowe, *Ph.D. thesis*, University of Cambridge, 2012.

[119] D. M. Lowe and R. A. Sayle, *J. Cheminformatics*, 2015, **7**, S5.

[120] N. Schneider, D. M. Lowe, R. A. Sayle, M. A. Tarselli and G. A. Landrum, *J. Med. Chem.*, 2016, **59**, 4385–4402.

[121] H. Dai, C. Li, C. Coley, B. Dai and L. Song, *USPTO-FULL*, `https://www.dropbox.com/sh/6ideflxcakrak10/AAB6bLHH32CvtGTjRsXTCL02a/uspto_multi?dl=0&subfolder_nav_tracking=1`, (accessed July 2023).

[122] W. Jin, C. W. Coley, R. Barzilay and T. Jaakkola, Proceedings of the 31st Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2017, pp. 2604–2613.

[123] W. Jin, C. W. Coley, R. Barzilay and T. Jaakkola, *USPTO-MIT*, `https://github.com/wengong-jin/nips17-rexgen/blob/master/USPTO/data.zip`, (accessed July 2023).

[124] N. Schneider, N. Stiefl and G. A. Landrum, *J. Chem. Inf. Model.*, 2016, **56**, 2336–2346.

[125] B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. L. Nguyen, S. Ho, J. Sloane, P. Wender and V. Pande, *USPTO-50K-Liu*, 2017, `https://github.com/pandegroup/reaction_prediction_seq2seq/tree/master/processed_data`, (accessed July 2023).

[126] H. Dai, C. Li, C. W. Coley, B. Dai and L. Song, *USPTO-50K-Coley*, 2019, `https://www.dropbox.com/sh/6ideflxcakrak10/AADN-TNZnuGjvwZYiLk7zvwra/schneider50k?dl=0&subfolder_nav_tracking=1`, (accessed July 2023).

[127] B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. L. Nguyen, S. Ho, J. Sloane, P. Wender and V. Pande, *ACS Cent. Sci.*, 2017, **3**, 1103–1113.

[128] C. W. Coley, L. Rogers, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 1237–1245.

[129] X. Jia, A. Lynch, Y. Huang, M. Danielson, I. Lang'at, A. Milder, A. E. Ruby, H. Wang, S. A. Friedler, A. J. Norquist and J. Schrier, *Nature*, 2019, **573**, 251–255.

[130] R.-R. Griffiths, P. Schwaller and A. A. Lee, 2021, arXiv:2105.02637.

[131] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434–443.

[132] W. P. Walters, *J. Chem. Inf. Model.*, 2013, **53**, 1529–1530.

[133] R. D. Clark, *J. Cheminformatics*, 2019, **11**, 62.

[134] A. Thakkar, S. Johansson, K. Jorner, D. Buttar, J.-L. Reymond and O. Engkvist, *React. Chem. Eng.*, 2021, **6**, 27–51.

[135] J. Freilich and L. L. Ouellette, *Science*, 2019, **364**, 1036–1037.

[136] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, *Sci. Data*, 2016, **3**, 160018.

[137] A. Bøgevig, H.-J. Federsel, F. Huerta, M. G. Hutchings, H. Kraut,

T. Langer, P. Löw, C. Oppawsky, T. Rein and H. Saller, *Org. Process Res. Dev.*, 2015, **19**, 357–368.

[138] C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison and K. F. Jensen, *Science*, 2019, **365**, eaax1566.

[139] M. E. Fortunato, C. W. Coley, B. C. Barnes and K. F. Jensen, *J. Chem. Inf. Model.*, 2020, **60**, 3398–3407.

[140] S. Genheden, A. Thakkar, V. Chadimová, J. L. Reymond, O. Engkvist and E. Bjerrum, *J. Cheminformatics*, 2020, **12**, 1–9.

[141] C. W. Coley, W. H. Green and K. F. Jensen, *J. Chem. Inf. Model.*, 2019, **59**, 2529–2537.

[142] W. L. Chen, D. Z. Chen and K. T. Taylor, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2013, **3**, 560–593.

[143] E. Heid, J. Liu, A. Aude and W. H. Green, *J. Chem. Inf. Model.*, 2022, **62**, 16–26.

[144] T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. Trice and B. A. Grzybowski, *Chem*, 2018, **4**, 522–532.

[145] W. L. Jorgensen, E. R. Laird, A. J. Gushurst, J. M. Fleischer, S. A. Gothe, H. E. Helson, G. D. Paderes and S. Sinclair, *Pure Appl. Chem.*, 1990, **62**, 1921–1932.

[146] H. Satoh and K. Funatsu, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 34–44.

[147] S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk and B. A. Grzybowski, *Angew. Chem. Int. Ed.*, 2016, **55**, 5904–5937.

[148] J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade and H. Y. Ando, *J. Chem. Inf. Model.*, 2009, **49**, 593–602.

[149] S. A. Rahman, G. Torrance, L. Baldacci, S. M. Cuesta, F. Fenninger, N. Gopal, S. Choudhary, J. W. May, G. L. Holliday, C. Steinbeck and J. M. Thornton, *Bioinformatics*, 2016, **32**, 2065–2066.

[150] P. P. Plehiers, G. B. Marin, C. V. Stevens and K. M. V. Geem, *J. Cheminformatics*, 2018, **10**, 11.

[151] A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist and E. J. Bjerrum, *Chem. Sci.*, 2020, **11**, 154–168.

[152] W. Jaworski, S. Szymkuć, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Rydzewski, A. Gambin and B. A. Grzybowski, *Nat. Commun.*, 2019, **10**, 1434.

[153] C. Coley, *RDChiral*, `https://github.com/connorcoley/rdchiral`, (accessed July 2023).

[154] H. Kraut, J. Eiblmaier, G. Grethe, P. Löw, H. Matuszczyk and H. Saller, *J. Chem. Inf. Model.*, 2013, **53**, 2884–2895.

[155] A. Thakkar and J.-L. Reymond, *Chimia*, 2022, **76**, 294.

[156] X. Wang, Y. Li, J. Qiu, G. Chen, H. Liu, B. Liao, C.-Y. Hsieh and X. Yao, *J. Chem. Eng.*, 2021, **420**, 129845.

[157] Z. Chen, O. R. Ayinde, J. R. Fuchs, H. Sun and X. Ning, *Commun. Chem.*, 2023, **6**, 102.

[158] C. Yan, Q. Ding, P. Zhao, S. Zheng, J. Yang, Y. Yu and J. Huang, Proceedings of the 34th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2020, pp. 11248–11258.

[159] C. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay and K. F. Jensen, *Chem. Sci.*, 2019, **10**, 370–377.

[160] W. Zhong, Z. Yang and C. Y.-C. Chen, *Nat. Commun.*, 2023, **14**, 3009.

[161] J. Nam and J. Kim, 2016, arXiv:1612.09529.

[162] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2017.

[163] Z. Guo, S. Wu, M. Ohno and R. Yoshida, *J. Chem. Inf. Model.*, 2020, **60**, 4474–4486.

[164] R. Irwin, S. Dimitriadis, J. He and E. J. Bjerrum, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 015022.

[165] E. Kim, D. Lee, Y. Kwon, M. S. Park and Y.-S. Choi, *J. Chem. Inf. Model.*, 2021, **61**, 123–133.

[166] S. Zheng, J. Rao, Z. Zhang, J. Xu and Y. Yang, *J. Chem. Inf. Model.*, 2020, **60**, 47–55.

[167] K. Lin, Y. Xu, J. Pei and L. Lai, 2019, arXiv:1906.02308.

[168] P. Karpov, G. Godin and I. V. Tetko, Proceedings of the Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions, Cham, 2019, pp. 817–830.

[169] B. Chen, T. Shen, S. Jaakkola and R. Barzilay, 2019, arXiv:1910.09688.

[170] *RetroXpert*, `https://github.com/uta-smile/RetroXpert`, (accessed July 2023).

[171] M. Sacha, M. Błaz, P. Byrski, P. Dąbrowski-Tuman, M. Chromin, R. Loska, P. Włodarczyk-Pruszyn and S. Jastrzębski, *J. Chem. Inf. Model.*, 2021, **61**, 59.

[172] Z. Zhong, J. Song, Z. Feng, T. Liu, L. Jia, S. Yao, M. Wu, T. Hou and M. Song, *Chem. Sci.*, 2022, **13**, 9023–9034.

[173] S.-W. Seo, Y. Y. Song, J. Y. Yang, S. Bae, H. Lee, J. Shin, S. J. Hwang and E. Yang, Proceedings of the 35th AAAI Conference on Artificial Intelligence, Washington, DC, USA, 2021, pp. 531–539.

[174] R. Sun, H. Dai, L. Li, S. Kearnes and B. Dai, Proceedings of the 35th conference on Neural Information Processing Systems, Red Hook, NY, USA, 2021, pp. 10186–10194.

[175] Z. Tu and C. W. Coley, *J. Chem. Inf. Model.*, 2022, **62**, 3503–3513.

[176] J. S. Carey, D. Laffan, C. Thomson and M. T. Williams, *Org. Biomol. Chem.*, 2006, **4**, 2337.

[177] S. Genheden and E. Bjerrum, *Digital Discovery*, 2022, **1**, 527–539.

[178] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis and S. Colton, *IEEE Trans. Comput. Intell. AI Games*, 2012, **4**, 1–43.

[179] A. Kishimoto, B. Buesser, B. Chen and A. B. Eaton, Proceedings of the 33rd conference on Neural Information Processing Systems, Red Hook, NY, USA, 2019.

[180] K. Lin, Y. Xu, J. Pei and L. Lai, *Chem. Sci.*, 2020, **11**, 3355–3364.

[181] M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz,

P. E. Fuller, B. A. Grzybowski and K. J. M. Bishop, *Angew. Chem. Int. Ed.*, 2012, **51**, 7928–7932.

[182] T. Badowski, K. Molga and B. A. Grzybowski, *Chem. Sci.*, 2019, **10**, 4640–4651.

[183] K. Molga, P. Dittwald and B. A. Grzybowski, *Chem. Sci.*, 2019, **10**, 9219–9232.

[184] B. Mikulak-Klucznik, P. Gołębiowska, A. A. Bayly, O. Popik, T. Klucznik, S. Szymkuć, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, T. Badowski, K. A. Scheidt, K. Molga, J. Mlynarski, M. Mrksich and B. A. Grzybowski, *Nature*, 2020, **588**, 83–88.

[185] S.-A. Stark, R. Neudert, R. Threlfall, R. Neudert and R. Threlfall, *Wiley ChemPlanner predicts experimentally verified synthesis routes in medicinal chemistry*, `https://www.chemanager-online.com/en/whitepaper/wiley-chemplanner-predicts-experimentally-verified-synthesis-routes-medicinal-chemistry`, (accessed July 2023).

[186] K. Do, T. Tran and S. Venkatesh, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, NY, USA, 2019, pp. 750–760.

[187] W. W. Qian, N. T. Russell, C. L. W. Simons, Y. Luo, M. D. Burke and J. Peng, *Integrating Deep Neural Networks and Symbolic Inference for Organic Reactivity Prediction*, 2020, `https://chemrxiv.org/engage/chemrxiv/article-details/60c7476dbb8c1a4fad3daa77`, (accessed July 2023).

[188] H. Bi, H. Wang, C. Shi, C. Coley, J. Tang and H. Guo, Proceedings of the 38th International Conference on Machine Learning, 2021, pp. 904–913.

[189] J. Boström, D. G. Brown, R. J. Young and G. M. Keserü, *Nat. Rev. Drug Discov.*, 2018, **17**, 709–727.

[190] P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano and T. Laino, *Chem. Sci.*, 2020, **11**, 3316–3325.

[191] A. C. Vaucher, P. Schwaller, J. Geluykens, V. H. Nair, A. Iuliano and T. Laino, *Nat. Commun.*, 2021, **12**, 2573.

[192] J. Jiménez-Luna, F. Grisoni and G. Schneider, *Nat. Mach. Intell.*, 2020, **2**, 573–584.

[193] Y. Xu, J. Pei and L. Lai, *J. Chem. Inf. Model.*, 2017, **57**, 2672–2685.

[194] Z. Wu, D. Jiang, J. Wang, C.-Y. Hsieh, D. Cao and T. Hou, *J. Med. Chem.*, 2021, **64**, 6924–6936.

[195] L. Jia, Z. Feng, H. Zhang, J. Song, Z. Zhong, S. Yao and M. Song, *Adv. Intell. Syst.*, 2022, **4**, 2200104.

[196] D. C. Elton, Z. Boukouvalas, M. D. Fuge and P. W. Chung, *Mol. Syst. Des. Eng.*, 2019, **4**, 828–849.

[197] S. Ishida, K. Terayama, R. Kojima, K. Takasu and Y. Okuno, *J. Chem. Inf. Model.*, 2019, **59**, 5026–5033.

[198] E. P. Gajewska, S. Szymkuć, P. Dittwald, M. Startek, O. Popik, J. Mlynarski and B. A. Grzybowski, *Chem*, 2020, **6**, 280–293.

[199] E. Walker, J. Kammeraad, J. Goetz, M. T. Robo, A. Tewari and P. M. Zimmerman, *J. Chem. Inf. Model.*, 2019, **59**, 3645–3654.

[200] M. R. Maser, A. Y. Cui, S. Ryou, T. J. DeLano, Y. Yue and S. E. Reisman, *J. Chem. Inf. Model.*, 2021, acs.jcim.0c01234.

[201] Y.-C. Lo, S. E. Rensi, W. Torng and R. B. Altman, *Drug Discov. Today*, 2018, **23**, 1538–1546.

[202] R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. S. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer and G. S. Sittampalam, *Nat. Rev. Drug Discov.*, 2011, **10**, 188–195.

[203] J. M. Granda, L. Donina, V. Dragone, D.-L. Long and L. Cronin, *Nature*, 2018, **559**, 377–381.

[204] S. Singh, M. Pareek, A. Changotra, S. Banerjee, B. Bhaskararao, P. Balamurugan and R. B. Sunoj, *Proceedings of the National Academy of Sciences U. S. A.*, 2020, **117**, 1339–1345.

[205] J. P. Reid and M. S. Sigman, *Nature*, 2019, **571**, 343–348.

[206] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**, eaau5631.

[207] L. D. Hughes, D. S. Palmer, F. Nigsch and J. B. O. Mitchell, *J. Chem. Inf. Model.*, 2008, **48**, 220–232.

[208] A. P. Harding, D. C. Wedge and P. L. A. Popelier, *J. Chem. Inf. Model.*, 2009, **49**, 1914–1924.

[209] N. Dong, W. cong Lu, N. yi Chrn, Y. cheng Zhu and K. xian Chen, *Acta Pharmacol. Sin.*, 2005, **26**, 107–112.

[210] J. L. Melville, E. K. Burke and J. D. Hirst, *Comb. Chem. High Throughput Screen.*, 2009, **12**, 332–343.

[211] F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks and F. Glorius, *Chem*, 2020, **6**, 1379–1390.

[212] T. Miyao, K. Funatsu and J. Bajorath, *J. Chem. Inf. Model.*, 2019, **59**, 983–992.

[213] Y. Lu, S. Anand, W. Shirley, P. Gedeck, B. P. Kelley, S. Skolnik, S. Rodde, M. Nguyen, M. Lindvall and W. Jia, *J. Chem. Inf. Model.*, 2019, **59**, 4706–4719.

[214] T. Miyao and K. Funatsu, *J. Chem. Inf. Model.*, 2019, **59**, 2626–2641.

[215] T. Gärtner, P. Flach and S. Wrobel, Learning Theory and Kernel Machines, Berlin, Heidelberg, 2003, pp. 129–143.

[216] N. M. Kriege, F. D. Johansson and C. Morris, *Appl. Netw. Sci.*, 2020, **5**, 6.

[217] T. Lei, W. Jin, R. Barzilay and T. Jaakkola, Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 3181–3190.

[218] G. Marcou, J. A. de Sousa, D. A. R. S. Latino, A. de Luca, D. Horvath, V. Rietsch and A. Varnek, *J. Chem. Inf. Model.*, 2015, **55**, 239–250.

[219] K. D. Collins, T. Gensch and F. Glorius, *Nat. Chem.*, 2014, **6**, 859–871.

[220] A. B. Santanilla, E. L. Regalado, T. Pereira, M. Shevlin, K. Bateman, L.-C. Campeau, J. Schneeweis, S. Berritt, Z.-C. Shi, P. Nantermet, Y. Liu, R. Helmy, C. J. Welch, P. Vachal, I. W. Davies, T. Cernak and S. D. Dreher, *Science*, 2015, **347**, 49–53.

[221] J. J. Henle, A. F. Zahrt, B. T. Rose, W. T. Darrow, Y. Wang and S. E. Denmark, *J. Am. Chem. Soc.*, 2020, **142**, 11578–11592.

[222] C. Torborg and M. Beller, *Adv. Synth. Catal.*, 2009, **351**, 3027–3043.

[223] J. Magano and J. R. Dunetz, *Chem. Rev.*, 2011, **111**, 2177–2250.

[224] P. Ruiz-Castillo and S. L. Buchwald, *Chem. Rev.*, 2016, **116**, 12564–12649.

[225] M. Kosugi, M. Kameyama and T. Migita, *Chem. Lett.*, 1983, **12**, 927–928.

[226] A. S. Guram, R. A. Rennels and S. L. Buchwald, *Angew. Chem. Int. Ed.*, 1995, **34**, 1348–1350.

[227] J. Louie and J. F. Hartwig, *Tetrahedron Lett.*, 1995, **36**, 3609–3612.

[228] D. Maiti, B. P. Fors, J. L. Henderson, Y. Nakamura and S. L. Buchwald, *Chem. Sci.*, 2011, **2**, 57–68.

[229] D. S. Surry and S. L. Buchwald, *Chem. Sci.*, 2011, **2**, 27–50.

[230] E. Vitaku, D. T. Smith and J. T. Njardarson, *J. Med. Chem.*, 2014, **57**, 10257–10274.

[231] K. D. Collins and F. Glorius, *Acc. Chem. Res.*, 2015, **48**, 619–627.

[232] A. F. Zahrt, J. J. Henle and S. E. Denmark, *ACS Comb. Sci.*, 2020, **22**, 586–591.

[233] N. C. I. C.-A. D. D. N. group (2009-2020), *Chemical Identifier Resolver*, 2020, `https://cactus.nci.nih.gov/chemical/structure`, (accessed July 2023).

[234] Y. Shao, L. F. Molnar, Y. Jung, J. Kussmann, C. Ochsenfeld, S. T. Brown, A. T. Gilbert, L. V. Slipchenko, S. V. Levchenko, D. P. O'Neill, R. A. D. Jr, R. C. Lochan, T. Wang, G. J. Beran, N. A. Besley, J. M. Herbert, C. Y. Lin, T. V. Voorhis, S. H. Chien, A. Sodt, R. P. Steele, V. A. Rassolov, P. E. Maslen, P. P. Korambath, R. D. Adamson, B. Austin, J. Baker, E. F. C. Byrd, H. Dachsel, R. J. Doerksen, A. Dreuw, B. D. Dunietz, A. D. Dutoi, T. R. Furlani, S. R. Gwaltney, A. Heyden, S. Hirata, C.-P. Hsu, G. Kedziora, R. Z. Khalliulin, P. Klunzinger, A. M. Lee, M. S. Lee, W. Liang, I. Lotan, N. Nair, B. Peters, E. I. Proynov, P. A. Pieniazek, Y. M. Rhee, J. Ritchie, E. Rosta, C. D. Sherrill, A. C. Simmonett, J. E. Subotnik, H. L. W. III, W. Zhang, A. T. Bell, A. K. Chakraborty, D. M. Chipman, F. J. Keil, A. Warshel, W. J. Hehre, H. F. S. III, J. Kong, A. I. Krylov, P. M. W. Gill and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2006, **8**, 3172–3191.

[235] Y. Shao, Z. Gan, E. Epifanovsky, A. T. B. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kuś, A. Landau, J. Liu, E. I. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele, E. J. Sundstrom, H. L. Woodcock, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. O. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C.-M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Diedenhofen, R. A. DiStasio, H. Do, A. D. Dutoi, R. G. Edgar, S. Fatehi, L. Fusti-Molnar, A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. D. Hanson-Heine, P. H. P. Harbach, A. W. Hauser, E. G.

Hohenstein, Z. C. Holden, T.-C. Jagau, H. Ji, B. Kaduk, K. Khistyaev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowalczyk, C. M. Krauter, K. U. Lao, A. D. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Luenser, P. Manohar, S. F. Manzer, S.-P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, E. Neuscamman, C. M. Oana, R. Olivares-Amaya, D. P. O'Neill, J. A. Parkhill, T. M. Perrine, R. Peverati, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, S. M. Sharada, S. Sharma, D. W. Small, A. Sodt, T. Stein, D. Stück, Y.-C. Su, A. J. W. Thom, T. Tsuchimochi, V. Vanovschi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, J. Yang, S. Yeganeh, S. R. Yost, Z.-Q. You, I. Y. Zhang, X. Zhang, Y. Zhao, B. R. Brooks, G. K. Chan, D. M. Chipman, C. J. Cramer, W. A. Goddard, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. Schaefer, M. W. Schmidt, C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xu, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J.-D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C.-P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. V. Voorhis, J. M. Herbert, A. I. Krylov, P. M. W. Gill and M. Head-Gordon, *Mol. Phys.*, 2015, **113**, 184–215.

[236] A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.

[237] P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623–11627.

[238] G. Siglidis, G. Nikolentzos, S. Limnios, C. Giatsidis, K. Skianis and M. Vazirgianis, *J. Mach. Learn. Res.*, 2020, **21**, 1–5.

[239] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

[240] Y. Kwon, D. Lee, Y.-S. Choi and S. Kang, *J. Cheminformatics*, 2022, **14**, 2.

[241] A. Sato, T. Miyao and K. Funatsu, *Mol. Inform.*, 2022, **41**, 2100156.

[242] O. Ivanciuc, in *Reviews in Computational Chemistry*, ed. K. B. Lipkowitz and T. R. Cundari, John Wiley & Sons, Ltd, Volume 23 edn., 2007, ch. 6, pp. 291–400.

[243] K. Hasegawa and K. Funatsu, *Curr. Comput.-Aided Drug Des.*, 2010, **6**, 24–36.

[244] S. M. Mennen, C. Alhambra, C. L. Allen, M. Barberis, S. Berritt, T. A. Brandt, A. D. Campbell, J. Castañón, A. H. Cherney, M. Christensen, D. B. Damon, J. E. de Diego, S. García-Cerrada, P. García-Losada, R. Haro, J. Janey, D. C. Leitch, L. Li, F. Liu, P. C. Lobben, D. W. C. MacMillan, J. Magano, E. McInturff, S. Monfette, R. J. Post, D. Schultz, B. J. Sitter, J. M. Stevens, I. I. Strambeanu, J. Twilton, K. Wang and M. A. Zajac, *Org. Process Res. Dev.*, 2019, **23**, 1213–1242.

[245] D. L. J. Alexander, A. Tropsha and D. A. Winkler, *J. Chem. Inf. Model.*, 2015, **55**, 1316–1322.

# Appendix A

# Open-Source Patent Data

Text mining was used to extract experimental details from the United States Patent and Trademark Office (USPTO).[96,118–120] The full dataset can be downloaded from figshare. The downloadables files and file descriptions can be found in Table A.1. The reaction SMILES (*.rsmi) files contain the reaction SMILES, patent number, paragraph number, year, text-mined yield, and calculated yield. The 'text-mined' yield is the yield reported in the experimental section. The 'calculated' yield is the yield calculated from the quantities of the reagents used and the product obtained. The Chemical Markup Language (*.cml) files contain the source, text, reaction SMILES, reactant list, product list, spectator list, and reaction action list. The source includes citations such as paragraph number and patent (application) ID. The text is the complete experimental method. The reaction SMILES is the line notation of the reaction, including reactants, reagents, and products. The reactant list includes the name, SMILES, and InChI of the reactants. The product list includes name, SMILES, InChI, text-mined yield, calculated yield, appearance, and state of the products. The spectator list includes solvents and quantities used. The reaction action list is ordered steps on how to perform the synthesis; each step includes the action performed, components acted on, conditions, time, and temperature.

Table A.1: Downloadable Files and File Descriptions of the USPTO 1976-2016 Dataset

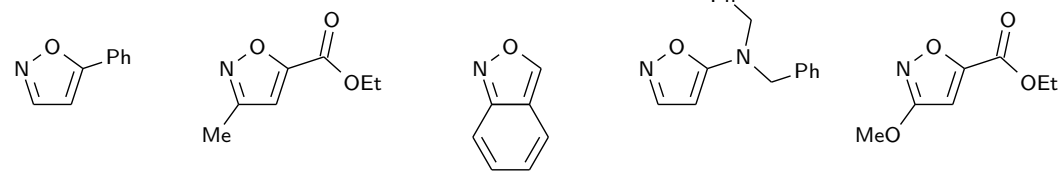
| File Name | Description |
|---|---|
| `2001_Sep2016_USPTOapplications_cml.7z` | Annoted patent applications from 2001-2016 in XML format. |
| `1976_Sep2016_USPTOgrants_cml.7z` | Annotated patent grants from 1976-2016 in XML format. |
| `1976_Sep2016_USPTOgrants_smiles.7z` | Reaction SMILES for patent grants from 1976-2016 in a Reaction SMILES file (.*rsmi). |
| `2001_Sep2016_USPTOapplications_smiles.7z` | Reaction SMILES for patent applications from 2001-2016 in a Reaction SMILES file (.*rsmi). |
| `xml_xsd.zip` | Schema Definition for XML files. |

# Appendix B

# Predicting Yields of Chemical Reactions

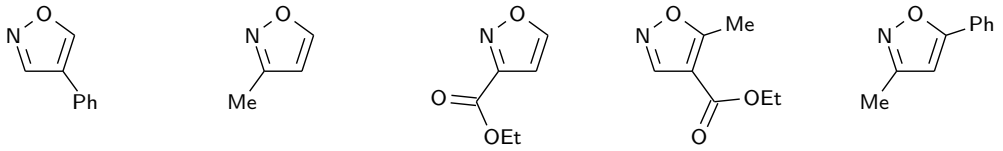
## B.1 Buchwald-Hartwig Dataset

Set 1:



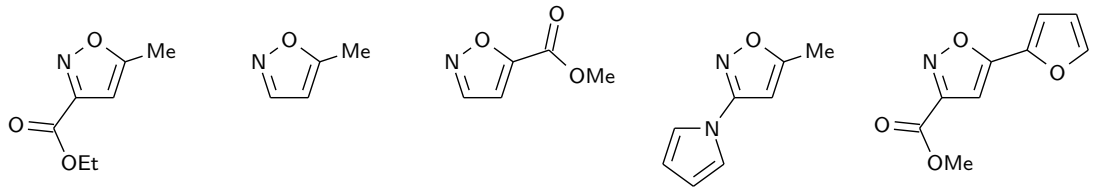

**2** **4** **10** **18** **23**

Set 2:



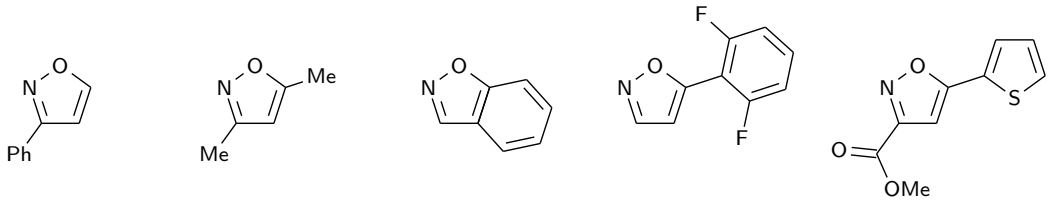

**1** **5** **9** **11** **17**

Set 3:



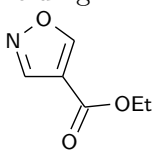

**6** **8** **14** **20** **21**

Set 4:



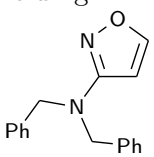

**3** **12** **15** **18** **22**

Highest Yielding:



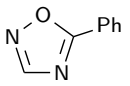

**13**

Lowest Yielding:




**19**

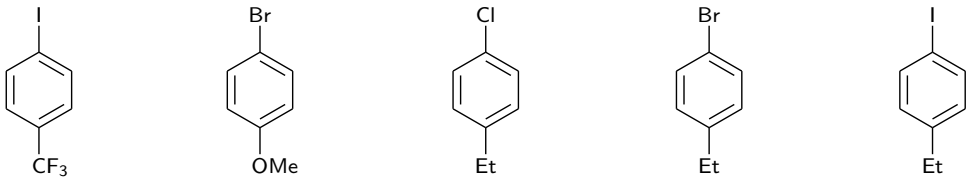
Miscellaneous:




**7**

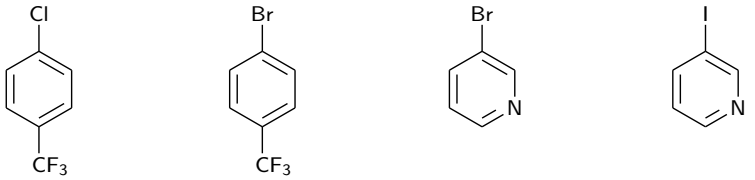
Figure B.1: Additives in the Buchwald-Hartwig dataset, compiled by Doyle *et al.*[32] Sets one to four are the additive ranked test sets. Additive **13** and **19** are the highest and lowest-yielding additives, respectively. Reactions containing additive **7** were removed from the Buchwald-Hartwig dataset as quantum chemical descriptors cannot be calculated.

Set 1:



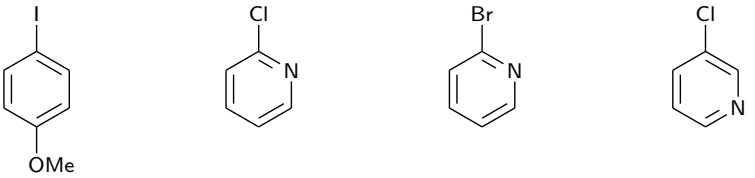

**3** **5** **7** **8** **9**

Set 2:



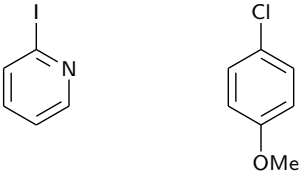

**1** **2** **14** **15**

Set 3:



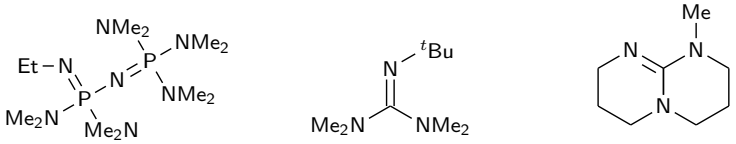

**6** **10** **11** **13**

Highest Yielding: Lowest Yielding:




**12** **4**

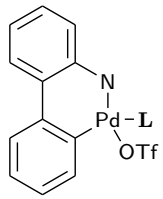
Figure B.2: Aryl halides in the Buchwald-Hartwig dataset, compiled by Doyle *et al.*[32] Sets one to three are the aryl halide ranked test sets. Aryl Halide **12** and **4** are the highest and lowest yielding aryl halides, respectively.



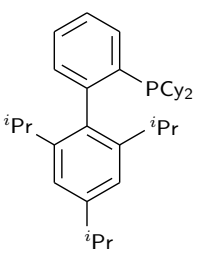

**$P_2$Et** **BTMG** **MTBD**

Figure B.3: Bases in the Buchwald-Hartwig dataset, compiled by Doyle *et al.*[32]

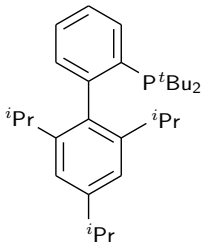
**Pd Catalyst:**



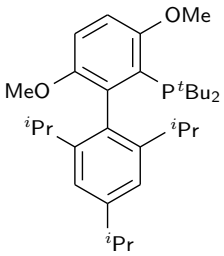

**L:**



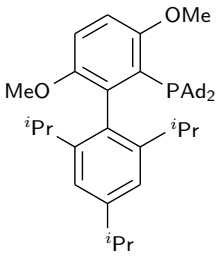

**XPhos**




**t-BuXPhos**




**t-BuBrettPhos**




**AdBrettPhos**

Figure B.4: Catalyst ligands in the Buchwald-Hartwig dataset, compiled by Doyle *et al.*[32]

## B.2 Quantum Chemical Descriptors

**Additive Descriptors** ($n = 19$) $E_{HOMO}$, $E_{LUMO}$, 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, $\nu_1$ Frequency, $\nu_1$ Intensity.

**Aryl Halide Descriptors** ($n = 27$) $E_{HOMO}$, $E_{LUMO}$, 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, $\nu_1$ Frequency, $\nu_1$ Intensity, $\nu_2$ Frequency, $\nu_2$ Intensity, $\nu_3$ Frequency, and $\nu_3$ Intensity.

**Base Descriptors** ($n = 10$) $E_{HOMO}$, $E_{LUMO}$, 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 S24 Electrostatic Charge, *P1 Electrostatic Charge, $\nu_1$ Frequency, $\nu_1$ Intensity, $\nu_2$ Frequency, $\nu_2$ Intensity, $\nu_3$ Frequency, $\nu_3$ Intensity, $\nu_4$ Frequency, $\nu_4$ Intensity, $\nu_5$ Frequency, $\nu_5$ Intensity, $\nu_6$ Frequency, $\nu_6$ Intensity, $\nu_7$ Frequency, $\nu_7$ Intensity, $\nu_8$ Frequency, $\nu_8$ Intensity, $\nu_9$ Frequency, $\nu_9$ Intensity, $\nu_10$ Frequency, $\nu_10$ Intensity

# B.3 Cross-Validation

## B.3.1 Parameter Optimisation of the Descriptors

### Descriptors Derived from Molecular Fingerprints

Table B.1: Average Cross-Validated Coefficient of Determination of the Tuned Linear Models Built on Molecular Fingerprints and Tanimoto Kernel Descriptors with Bit Lengths from 32 to 2048

| Descriptor | Bit Length | Linear Models | | | | | Mean |
|---|---|---|---|---|---|---|---|
| | | LR | Lasso | Ridge | Elastic Net | BR | |
| Fingerprints | | | | | | | |
| Morgan1 | 32 | 0.67 | 0.69 | 0.69 | 0.69 | 0.69 | 0.69 |
| | 64 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 |
| | 128 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 |
| | 256 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 |
| | 512 | 0.67 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 |
| | 1024 | 0.56 | 0.68 | 0.68 | 0.68 | 0.68 | 0.66 |
| | 2048 | 0.67 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 |
| Morgan2 | 32 | 0.70 | 0.70 | 0.70 | 0.69 | 0.70 | 0.70 |
| | 64 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 |
| | 128 | 0.67 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 |
| | 256 | 0.67 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 |
| | 512 | 0.67 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 |
| | 1024 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 |
| | 2048 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 | 0.68 |
| Morgan3 | 32 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| | 64 | 0.69 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| | 128 | 0.69 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| | 256 | 0.69 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| | 512 | 0.69 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| | 1024 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| | 2048 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| FMorgan1 | 32 | <-1.00 | 0.41 | 0.41 | 0.36 | 0.41 | <-1.00 |
| | 64 | <-1.00 | 0.45 | 0.44 | 0.37 | 0.44 | <-1.00 |
| | 128 | <-1.00 | 0.48 | 0.48 | 0.39 | 0.48 | <-1.00 |
| | 256 | <-1.00 | 0.49 | 0.49 | 0.41 | 0.49 | <-1.00 |
| | 512 | <-1.00 | 0.49 | 0.49 | 0.41 | 0.49 | <-1.00 |
| | 1024 | <-1.00 | 0.49 | 0.49 | 0.41 | 0.49 | <-1.00 |
| | 2048 | <-1.00 | 0.49 | 0.49 | 0.41 | 0.49 | <-1.00 |
| FMorgan2 | 32 | <-1.00 | 0.49 | 0.48 | 0.40 | 0.48 | <-1.00 |
| | 64 | <-1.00 | 0.49 | 0.47 | 0.39 | 0.47 | <-1.00 |
| | 128 | <-1.00 | 0.49 | 0.47 | 0.41 | 0.48 | <-1.00 |
| | 256 | <-1.00 | 0.50 | 0.46 | 0.42 | 0.48 | <-1.00 |
| | 512 | <-1.00 | 0.50 | 0.45 | 0.42 | 0.48 | <-1.00 |
| | 1024 | <-1.00 | 0.50 | 0.45 | 0.41 | 0.47 | <-1.00 |
| | 2048 | <-1.00 | 0.50 | 0.45 | 0.41 | 0.47 | <-1.00 |
| FMorgan3 | 32 | <-1.00 | 0.49 | 0.48 | 0.42 | 0.48 | <-1.00 |
| | 64 | <-1.00 | 0.49 | 0.46 | 0.40 | 0.47 | <-1.00 |
| | 128 | <-1.00 | 0.49 | 0.45 | 0.40 | 0.47 | <-1.00 |
| | 256 | <-1.00 | 0.49 | 0.43 | 0.40 | 0.46 | <-1.00 |
| | 512 | <-1.00 | 0.50 | 0.43 | 0.39 | 0.46 | <-1.00 |
| | 1024 | <-1.00 | 0.49 | 0.43 | 0.39 | 0.46 | <-1.00 |
| | 2048 | <-1.00 | 0.49 | 0.43 | 0.39 | 0.46 | <-1.00 |
| RDK | 32 | 0.28 | 0.28 | 0.28 | 0.28 | 0.28 | 0.28 |
| | 64 | 0.56 | 0.57 | 0.57 | 0.56 | 0.57 | 0.56 |
| | 128 | 0.62 | 0.62 | 0.62 | 0.62 | 0.62 | 0.62 |
| | 256 | 0.69 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| | 512 | 0.69 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| | 1024 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |

Table B.1: Average Cross-Validated Coefficient of Determination of the Tuned Linear Models Built on Molecular Fingerprints and Tanimoto Kernel Descriptors with Bit Lengths from 32 to 2048 (Continued)

| Descriptor | Bit Length | Linear Models | | | | | Mean |
|---|---|---|---|---|---|---|---|
| | | LR | Lasso | Ridge | Elastic Net | BR | |
| | 2048 | 0.69 | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| Tanimoto | | | | | | | |
| Morgan1 | 32 | 0.92 | 0.93 | 0.93 | 0.61 | 0.93 | 0.86 |
| | 64 | 0.93 | 0.94 | 0.94 | 0.65 | 0.94 | 0.88 |
| | 128 | 0.93 | 0.94 | 0.94 | 0.62 | 0.94 | 0.87 |
| | 256 | 0.93 | 0.94 | 0.94 | 0.61 | 0.94 | 0.87 |
| | 512 | 0.93 | 0.94 | 0.94 | 0.61 | 0.94 | 0.87 |
| | 1024 | 0.94 | 0.94 | 0.94 | 0.62 | 0.94 | 0.88 |
| | 2048 | 0.94 | 0.94 | 0.94 | 0.62 | 0.94 | 0.88 |
| Morgan2 | 32 | 0.92 | 0.93 | 0.93 | 0.64 | 0.93 | 0.87 |
| | 64 | 0.94 | 0.94 | 0.94 | 0.61 | 0.94 | 0.87 |
| | 128 | 0.94 | 0.94 | 0.94 | 0.56 | 0.94 | 0.86 |
| | 256 | 0.94 | 0.94 | 0.94 | 0.54 | 0.94 | 0.86 |
| | 512 | 0.94 | 0.94 | 0.94 | 0.53 | 0.94 | 0.86 |
| | 1024 | 0.94 | 0.94 | 0.94 | 0.52 | 0.94 | 0.86 |
| | 2048 | 0.94 | 0.94 | 0.94 | 0.52 | 0.94 | 0.86 |
| Morgan3 | 32 | 0.90 | 0.91 | 0.92 | 0.62 | 0.92 | 0.85 |
| | 64 | 0.93 | 0.94 | 0.94 | 0.62 | 0.93 | 0.87 |
| | 128 | 0.94 | 0.94 | 0.94 | 0.55 | 0.94 | 0.86 |
| | 256 | 0.94 | 0.94 | 0.94 | 0.52 | 0.94 | 0.85 |
| | 512 | 0.94 | 0.94 | 0.94 | 0.49 | 0.94 | 0.85 |
| | 1024 | 0.94 | 0.94 | 0.94 | 0.47 | 0.94 | 0.85 |
| | 2048 | 0.94 | 0.94 | 0.94 | 0.46 | 0.94 | 0.84 |
| FMorgan1 | 32 | <-1.00 | 0.41 | 0.41 | 0.36 | 0.41 | <-1.00 |
| | 64 | <-1.00 | 0.45 | 0.44 | 0.37 | 0.44 | <-1.00 |
| | 128 | <-1.00 | 0.48 | 0.48 | 0.39 | 0.48 | <-1.00 |
| | 256 | <-1.00 | 0.49 | 0.49 | 0.41 | 0.49 | <-1.00 |
| | 512 | <-1.00 | 0.49 | 0.49 | 0.41 | 0.49 | <-1.00 |
| | 1024 | <-1.00 | 0.49 | 0.49 | 0.41 | 0.49 | <-1.00 |
| | 2048 | <-1.00 | 0.49 | 0.49 | 0.41 | 0.49 | <-1.00 |
| FMorgan2 | 32 | <-1.00 | 0.49 | 0.48 | 0.40 | 0.48 | <-1.00 |
| | 64 | <-1.00 | 0.49 | 0.47 | 0.39 | 0.47 | <-1.00 |
| | 128 | <-1.00 | 0.49 | 0.47 | 0.41 | 0.48 | <-1.00 |
| | 256 | <-1.00 | 0.50 | 0.46 | 0.42 | 0.48 | <-1.00 |
| | 512 | <-1.00 | 0.50 | 0.45 | 0.42 | 0.48 | <-1.00 |
| | 1024 | <-1.00 | 0.50 | 0.45 | 0.41 | 0.47 | <-1.00 |
| | 2048 | <-1.00 | 0.50 | 0.45 | 0.41 | 0.47 | <-1.00 |
| FMorgan3 | 32 | <-1.00 | 0.49 | 0.48 | 0.42 | 0.48 | <-1.00 |
| | 64 | <-1.00 | 0.49 | 0.46 | 0.40 | 0.47 | <-1.00 |
| | 128 | <-1.00 | 0.49 | 0.45 | 0.40 | 0.47 | <-1.00 |
| | 256 | <-1.00 | 0.49 | 0.43 | 0.40 | 0.46 | <-1.00 |
| | 512 | <-1.00 | 0.50 | 0.43 | 0.39 | 0.46 | <-1.00 |
| | 1024 | <-1.00 | 0.49 | 0.43 | 0.39 | 0.46 | <-1.00 |
| | 2048 | <-1.00 | 0.49 | 0.43 | 0.39 | 0.46 | <-1.00 |
| RDK | 32 | 0.27 | 0.28 | 0.28 | 0.23 | 0.28 | 0.27 |
| | 64 | <-1.00 | 0.61 | 0.61 | 0.47 | 0.61 | <-1.00 |
| | 128 | <-1.00 | 0.72 | 0.73 | 0.56 | 0.73 | <-1.00 |
| | 256 | <-1.00 | 0.89 | 0.89 | 0.59 | 0.89 | <-1.00 |
| | 512 | 0.91 | 0.92 | 0.92 | 0.59 | 0.92 | 0.85 |
| | 1024 | 0.91 | 0.92 | 0.92 | 0.56 | 0.92 | 0.85 |
| | 2048 | 0.91 | 0.92 | 0.92 | 0.52 | 0.92 | 0.84 |

Table B.2: Average Cross-Validated RMSE of the Tuned Linear Models Built on Molecular Fingerprints and Tanimoto Kernel Descriptors with Bit Lengths from 32 to 2048

| Descriptor | Bit Length | Linear Models | | | | | Mean |
|---|---|---|---|---|---|---|---|
| | | Linear Regression | Lasso | Ridge | Elastic Net | Bayesian Ridge | |
| Fingerprints | | | | | | | |
| Morgan1 | 32 | 15.7 | 15.2 | 15.2 | 15.3 | 15.2 | 15.3 |
| | 64 | 15.5 | 15.5 | 15.5 | 15.5 | 15.5 | 15.5 |
| | 128 | 15.5 | 15.4 | 15.4 | 15.5 | 15.5 | 15.5 |
| | 256 | 15.5 | 15.4 | 15.4 | 15.5 | 15.5 | 15.5 |
| | 512 | 15.6 | 15.4 | 15.4 | 15.5 | 15.5 | 15.5 |
| | 1024 | 17.9 | 15.4 | 15.4 | 15.5 | 15.5 | 16.0 |
| | 2048 | 15.6 | 15.4 | 15.4 | 15.5 | 15.5 | 15.5 |
| Morgan2 | 32 | 15.0 | 15.0 | 15.0 | 15.1 | 15.0 | 15.0 |
| | 64 | 15.5 | 15.4 | 15.4 | 15.4 | 15.4 | 15.5 |
| | 128 | 15.6 | 15.5 | 15.5 | 15.5 | 15.5 | 15.5 |
| | 256 | 15.6 | 15.5 | 15.5 | 15.5 | 15.5 | 15.5 |
| | 512 | 15.6 | 15.4 | 15.5 | 15.5 | 15.5 | 15.5 |
| | 1024 | 15.5 | 15.5 | 15.5 | 15.5 | 15.5 | 15.5 |
| | 2048 | 15.5 | 15.5 | 15.5 | 15.5 | 15.5 | 15.5 |
| Morgan3 | 32 | 15.0 | 15.0 | 15.0 | 15.0 | 15.0 | 15.0 |
| | 64 | 15.2 | 15.0 | 15.0 | 15.0 | 15.0 | 15.0 |
| | 128 | 15.1 | 15.0 | 15.0 | 15.0 | 15.0 | 15.0 |
| | 256 | 15.2 | 15.0 | 15.0 | 15.0 | 15.0 | 15.0 |
| | 512 | 15.3 | 15.0 | 15.0 | 15.0 | 15.0 | 15.0 |
| | 1024 | 15.0 | 14.9 | 15.0 | 15.0 | 15.0 | 15.0 |
| | 2048 | 15.1 | 15.0 | 15.0 | 15.0 | 15.0 | 15.0 |
| FMorgan1 | 32 | >100.0 | 20.9 | 20.9 | 21.9 | 20.9 | >100.0 |
| | 64 | >100.0 | 20.2 | 20.3 | 21.6 | 20.3 | >100.0 |
| | 128 | >100.0 | 19.6 | 19.6 | 21.3 | 19.6 | >100.0 |
| | 256 | >100.0 | 19.5 | 19.6 | 21.0 | 19.5 | >100.0 |
| | 512 | >100.0 | 19.5 | 19.6 | 21.0 | 19.5 | >100.0 |
| | 1024 | >100.0 | 19.5 | 19.6 | 21.0 | 19.5 | >100.0 |
| | 2048 | >100.0 | 19.5 | 19.6 | 21.0 | 19.5 | >100.0 |
| FMorgan2 | 32 | >100.0 | 19.5 | 19.6 | 21.1 | 19.6 | >100.0 |
| | 64 | >100.0 | 19.6 | 19.8 | 21.3 | 19.8 | >100.0 |
| | 128 | >100.0 | 19.4 | 19.9 | 20.9 | 19.7 | >100.0 |
| | 256 | >100.0 | 19.3 | 20.1 | 20.7 | 19.7 | >100.0 |
| | 512 | >100.0 | 19.3 | 20.1 | 20.8 | 19.7 | >100.0 |
| | 1024 | >100.0 | 19.3 | 20.2 | 20.9 | 19.8 | >100.0 |
| | 2048 | >100.0 | 19.3 | 20.2 | 20.9 | 19.8 | >100.0 |
| FMorgan3 | 32 | >100.0 | 19.5 | 19.6 | 20.9 | 19.6 | >100.0 |
| | 64 | >100.0 | 19.5 | 20.1 | 21.2 | 19.8 | >100.0 |
| | 128 | >100.0 | 19.4 | 20.3 | 21.1 | 19.9 | >100.0 |
| | 256 | >100.0 | 19.4 | 20.5 | 21.2 | 20.0 | >100.0 |
| | 512 | >100.0 | 19.4 | 20.6 | 21.3 | 20.1 | >100.0 |
| | 1024 | >100.0 | 19.4 | 20.6 | 21.3 | 20.1 | >100.0 |
| | 2048 | >100.0 | 19.4 | 20.6 | 21.3 | 20.1 | >100.0 |
| RDK | 32 | 23.1 | 23.1 | 23.1 | 23.1 | 23.1 | 23.1 |
| | 64 | 18.1 | 18.0 | 18.0 | 18.0 | 18.0 | 18.0 |
| | 128 | 16.9 | 16.8 | 16.8 | 16.8 | 16.8 | 16.8 |
| | 256 | 15.2 | 15.0 | 15.0 | 15.0 | 15.0 | 15.0 |
| | 512 | 15.2 | 15.0 | 15.0 | 15.0 | 15.0 | 15.0 |
| | 1024 | 15.0 | 15.0 | 15.0 | 15.0 | 15.0 | 15.0 |
| | 2048 | 15.2 | 15.0 | 15.0 | 15.0 | 15.0 | 15.0 |
| Tanimoto | | | | | | | |
| Morgan1 | 32 | 7.8 | 7.2 | 7.1 | 17.1 | 7.1 | 9.3 |
| | 64 | 7.1 | 6.8 | 6.8 | 16.0 | 6.7 | 8.7 |
| | 128 | 7.2 | 6.9 | 6.9 | 16.8 | 6.8 | 8.9 |
| | 256 | 7.0 | 6.8 | 6.8 | 17.0 | 6.7 | 8.8 |
| | 512 | 7.0 | 6.8 | 6.8 | 17.0 | 6.7 | 8.9 |
| | 1024 | 6.8 | 6.5 | 6.6 | 16.8 | 6.5 | 8.6 |

Table B.2: Average Cross-Validated RMSE of the Tuned Linear Models Built on Molecular Fingerprints and Tanimoto Kernel Descriptors with Bit Lengths from 32 to 2048 (Continued)

| Descriptor | Bit Length | Linear Models | | | | | Mean |
|---|---|---|---|---|---|---|---|
| | | Linear Regression | Lasso | Ridge | Elastic Net | Bayesian Ridge | |
| | 2048 | 6.8 | 6.5 | 6.6 | 16.8 | 6.5 | 8.6 |
| Morgan2 | 32 | 7.5 | 7.1 | 7.0 | 16.5 | 7.0 | 9.0 |
| | 64 | 6.8 | 6.7 | 6.7 | 17.0 | 6.7 | 8.8 |
| | 128 | 6.8 | 6.7 | 6.7 | 18.1 | 6.8 | 9.0 |
| | 256 | 6.6 | 6.6 | 6.6 | 18.5 | 6.6 | 9.0 |
| | 512 | 6.6 | 6.6 | 6.6 | 18.6 | 6.6 | 9.0 |
| | 1024 | 6.5 | 6.6 | 6.5 | 18.9 | 6.5 | 9.0 |
| | 2048 | 6.5 | 6.5 | 6.5 | 18.9 | 6.5 | 9.0 |
| Morgan3 | 32 | 8.6 | 8.0 | 7.9 | 16.7 | 7.7 | 9.8 |
| | 64 | 7.0 | 6.9 | 6.9 | 16.9 | 7.0 | 8.9 |
| | 128 | 6.9 | 6.9 | 6.9 | 18.2 | 6.9 | 9.1 |
| | 256 | 6.7 | 6.8 | 6.7 | 19.0 | 6.7 | 9.2 |
| | 512 | 6.7 | 6.8 | 6.7 | 19.5 | 6.7 | 9.3 |
| | 1024 | 6.7 | 6.7 | 6.7 | 19.9 | 6.7 | 9.3 |
| | 2048 | 6.7 | 6.7 | 6.7 | 20.0 | 6.7 | 9.4 |
| FMorgan1 | 32 | >100.0 | 20.9 | 20.9 | 21.9 | 20.9 | >100.0 |
| | 64 | >100.0 | 20.2 | 20.3 | 21.6 | 20.3 | >100.0 |
| | 128 | >100.0 | 19.6 | 19.6 | 21.3 | 19.6 | >100.0 |
| | 256 | >100.0 | 19.5 | 19.6 | 21.0 | 19.5 | >100.0 |
| | 512 | >100.0 | 19.5 | 19.6 | 21.0 | 19.5 | >100.0 |
| | 1024 | >100.0 | 19.5 | 19.6 | 21.0 | 19.5 | >100.0 |
| | 2048 | >100.0 | 19.5 | 19.6 | 21.0 | 19.5 | >100.0 |
| FMorgan2 | 32 | >100.0 | 19.5 | 19.6 | 21.1 | 19.6 | >100.0 |
| | 64 | >100.0 | 19.6 | 19.8 | 21.3 | 19.8 | >100.0 |
| | 128 | >100.0 | 19.4 | 19.9 | 20.9 | 19.7 | >100.0 |
| | 256 | >100.0 | 19.3 | 20.1 | 20.7 | 19.7 | >100.0 |
| | 512 | >100.0 | 19.3 | 20.1 | 20.8 | 19.7 | >100.0 |
| | 1024 | >100.0 | 19.3 | 20.2 | 20.9 | 19.8 | >100.0 |
| | 2048 | >100.0 | 19.3 | 20.2 | 20.9 | 19.8 | >100.0 |
| FMorgan3 | 32 | >100.0 | 19.5 | 19.6 | 20.9 | 19.6 | >100.0 |
| | 64 | >100.0 | 19.5 | 20.1 | 21.2 | 19.8 | >100.0 |
| | 128 | >100.0 | 19.4 | 20.3 | 21.1 | 19.9 | >100.0 |
| | 256 | >100.0 | 19.4 | 20.5 | 21.2 | 20.0 | >100.0 |
| | 512 | >100.0 | 19.4 | 20.6 | 21.3 | 20.1 | >100.0 |
| | 1024 | >100.0 | 19.4 | 20.6 | 21.3 | 20.1 | >100.0 |
| | 2048 | >100.0 | 19.4 | 20.6 | 21.3 | 20.1 | >100.0 |
| RDK | 32 | 23.4 | 23.1 | 23.1 | 24.0 | 23.1 | 23.3 |
| | 64 | >100.0 | 16.9 | 17.0 | 19.9 | 17.0 | >100.0 |
| | 128 | >100.0 | 14.4 | 14.3 | 18.2 | 14.3 | >100.0 |
| | 256 | >100.0 | 9.2 | 9.0 | 17.4 | 9.0 | >100.0 |
| | 512 | 8.2 | 7.9 | 7.8 | 17.6 | 7.8 | 9.9 |
| | 1024 | 8.0 | 7.8 | 7.8 | 18.1 | 7.7 | 9.9 |
| | 2048 | 8.0 | 7.9 | 7.8 | 18.9 | 7.8 | 10.1 |

Table B.3: Average Cross-Validated Coefficient of Determination of the Tuned SVR Models Built on Molecular Fingerprints and Tanimoto Kernel Descriptors with Bit Lengths from 32 to 2048

| Descriptor | Bit Length | SVR Kernel | | | | | Mean |
|---|---|---|---|---|---|---|---|
| | | Linear | Polynomial | RBF | Sigmoid | Precomputed | |
| Fingerprints | | | | | | | |
| Morgan1 | 32 | 0.69 | 0.92 | 0.93 | 0.34 | | 0.72 |
| | 64 | 0.68 | 0.91 | 0.92 | 0.47 | | 0.75 |
| | 128 | 0.68 | 0.91 | 0.92 | 0.47 | | 0.75 |
| | 256 | 0.68 | 0.92 | 0.93 | 0.48 | | 0.75 |
| | 512 | 0.68 | 0.91 | 0.92 | 0.49 | | 0.75 |
| | 1024 | 0.68 | 0.92 | 0.93 | 0.50 | | 0.76 |
| | 2048 | 0.68 | 0.92 | 0.93 | 0.50 | | 0.76 |
| Morgan2 | 32 | 0.70 | 0.91 | 0.93 | 0.30 | | 0.71 |
| | 64 | 0.68 | 0.91 | 0.94 | 0.46 | | 0.75 |
| | 128 | 0.68 | 0.92 | 0.94 | 0.47 | | 0.75 |
| | 256 | 0.68 | 0.92 | 0.94 | 0.49 | | 0.76 |
| | 512 | 0.68 | 0.92 | 0.94 | 0.50 | | 0.76 |
| | 1024 | 0.68 | 0.92 | 0.94 | 0.51 | | 0.76 |
| | 2048 | 0.68 | 0.92 | 0.94 | 0.52 | | 0.76 |
| Morgan3 | 32 | 0.70 | 0.90 | 0.92 | 0.27 | | 0.70 |
| | 64 | 0.70 | 0.91 | 0.94 | 0.45 | | 0.75 |
| | 128 | 0.70 | 0.91 | 0.94 | 0.48 | | 0.76 |
| | 256 | 0.70 | 0.92 | 0.94 | 0.50 | | 0.76 |
| | 512 | 0.70 | 0.91 | 0.94 | 0.50 | | 0.76 |
| | 1024 | 0.70 | 0.91 | 0.94 | 0.51 | | 0.77 |
| | 2048 | 0.70 | 0.92 | 0.94 | 0.51 | | 0.77 |
| FMorgan1 | 32 | | 0.38 | 0.38 | 0.35 | 0.39 | 0.38 |
| | 64 | | 0.41 | 0.40 | 0.39 | 0.43 | 0.41 |
| | 128 | | 0.44 | 0.43 | 0.41 | 0.46 | 0.44 |
| | 256 | | 0.46 | 0.44 | 0.43 | 0.48 | 0.45 |
| | 512 | | 0.46 | 0.45 | 0.44 | 0.48 | 0.45 |
| | 1024 | | 0.45 | 0.44 | 0.43 | 0.48 | 0.45 |
| | 2048 | | 0.45 | 0.44 | 0.43 | 0.48 | 0.45 |
| FMorgan2 | 32 | | 0.46 | 0.45 | 0.44 | 0.48 | 0.46 |
| | 64 | | 0.45 | 0.43 | 0.45 | 0.47 | 0.45 |
| | 128 | | 0.45 | 0.44 | 0.50 | 0.48 | 0.47 |
| | 256 | | 0.45 | 0.44 | 0.50 | 0.48 | 0.47 |
| | 512 | | 0.45 | 0.43 | 0.51 | 0.47 | 0.47 |
| | 1024 | | 0.45 | 0.43 | 0.51 | 0.47 | 0.46 |
| | 2048 | | 0.45 | 0.43 | 0.51 | 0.47 | 0.46 |
| FMorgan3 | 32 | | 0.46 | 0.45 | 0.44 | 0.48 | 0.46 |
| | 64 | | 0.45 | 0.43 | 0.50 | 0.47 | 0.46 |
| | 128 | | 0.44 | 0.42 | 0.50 | 0.46 | 0.46 |
| | 256 | | 0.44 | 0.42 | 0.50 | 0.46 | 0.45 |
| | 512 | | 0.43 | 0.41 | 0.50 | 0.46 | 0.45 |
| | 1024 | | 0.43 | 0.41 | 0.50 | 0.46 | 0.45 |
| | 2048 | | 0.43 | 0.41 | 0.50 | 0.46 | 0.45 |
| RDK | 32 | 0.27 | 0.25 | 0.26 | -0.02 | | 0.19 |
| | 64 | 0.56 | 0.60 | 0.60 | -0.02 | | 0.44 |
| | 128 | 0.62 | 0.72 | 0.72 | 0.17 | | 0.56 |
| | 256 | 0.69 | 0.90 | 0.89 | 0.42 | | 0.73 |
| | 512 | 0.70 | 0.91 | 0.92 | 0.26 | | 0.70 |
| | 1024 | 0.70 | 0.91 | 0.91 | 0.34 | | 0.71 |
| | 2048 | 0.70 | 0.91 | 0.91 | 0.34 | | 0.71 |
| Tanimoto | | | | | | | |
| Morgan1 | 32 | | 0.92 | 0.91 | 0.84 | 0.93 | 0.90 |
| | 64 | | 0.93 | 0.92 | 0.87 | 0.94 | 0.91 |
| | 128 | | 0.92 | 0.91 | 0.87 | 0.94 | 0.91 |
| | 256 | | 0.93 | 0.91 | 0.88 | 0.94 | 0.92 |
| | 512 | | 0.93 | 0.91 | 0.89 | 0.94 | 0.92 |
| | 1024 | | 0.93 | 0.92 | 0.89 | 0.94 | 0.92 |
| | 2048 | | 0.93 | 0.92 | 0.89 | 0.94 | 0.92 |

Table B.3: Average Cross-Validated Coefficient of Determination of the Tuned SVR Models Built on Molecular Fingerprints and Tanimoto Kernel Descriptors with Bit Lengths from 32 to 2048 (Continued)

| Descriptor | Bit Length | SVR Kernel | | | | | Mean |
|---|---|---|---|---|---|---|---|
| | | Linear | Polynomial | RBF | Sigmoid | Precomputed | |
| Morgan2 | 32 | | 0.93 | 0.92 | 0.84 | 0.93 | 0.91 |
| | 64 | | 0.93 | 0.91 | 0.90 | 0.94 | 0.92 |
| | 128 | | 0.92 | 0.89 | 0.91 | 0.94 | 0.91 |
| | 256 | | 0.91 | 0.89 | 0.91 | 0.94 | 0.92 |
| | 512 | | 0.91 | 0.89 | 0.92 | 0.94 | 0.92 |
| | 1024 | | 0.91 | 0.89 | 0.92 | 0.94 | 0.92 |
| | 2048 | | 0.91 | 0.89 | 0.92 | 0.94 | 0.92 |
| Morgan3 | 32 | | 0.92 | 0.91 | 0.82 | 0.91 | 0.89 |
| | 64 | | 0.92 | 0.91 | 0.90 | 0.94 | 0.92 |
| | 128 | | 0.91 | 0.89 | 0.91 | 0.94 | 0.91 |
| | 256 | | 0.90 | 0.88 | 0.92 | 0.94 | 0.91 |
| | 512 | | 0.90 | 0.87 | 0.92 | 0.94 | 0.91 |
| | 1024 | | 0.90 | 0.86 | 0.92 | 0.94 | 0.91 |
| | 2048 | | 0.90 | 0.86 | 0.93 | 0.94 | 0.90 |
| FMorgan1 | 32 | | 0.38 | 0.38 | 0.35 | 0.39 | 0.38 |
| | 64 | | 0.41 | 0.40 | 0.39 | 0.43 | 0.41 |
| | 128 | | 0.44 | 0.43 | 0.41 | 0.46 | 0.44 |
| | 256 | | 0.46 | 0.44 | 0.43 | 0.48 | 0.45 |
| | 512 | | 0.46 | 0.45 | 0.44 | 0.48 | 0.45 |
| | 1024 | | 0.45 | 0.44 | 0.43 | 0.48 | 0.45 |
| | 2048 | | 0.45 | 0.44 | 0.43 | 0.48 | 0.45 |
| FMorgan2 | 32 | | 0.46 | 0.45 | 0.44 | 0.48 | 0.46 |
| | 64 | | 0.45 | 0.43 | 0.45 | 0.47 | 0.45 |
| | 128 | | 0.45 | 0.44 | 0.50 | 0.48 | 0.47 |
| | 256 | | 0.45 | 0.44 | 0.50 | 0.48 | 0.47 |
| | 512 | | 0.45 | 0.43 | 0.51 | 0.47 | 0.47 |
| | 1024 | | 0.45 | 0.43 | 0.51 | 0.47 | 0.46 |
| | 2048 | | 0.45 | 0.43 | 0.51 | 0.47 | 0.46 |
| FMorgan3 | 32 | | 0.46 | 0.45 | 0.44 | 0.48 | 0.46 |
| | 64 | | 0.45 | 0.43 | 0.50 | 0.47 | 0.46 |
| | 128 | | 0.44 | 0.42 | 0.50 | 0.46 | 0.46 |
| | 256 | | 0.44 | 0.42 | 0.50 | 0.46 | 0.45 |
| | 512 | | 0.43 | 0.41 | 0.50 | 0.46 | 0.45 |
| | 1024 | | 0.43 | 0.41 | 0.50 | 0.46 | 0.45 |
| | 2048 | | 0.43 | 0.41 | 0.50 | 0.46 | 0.45 |
| RDK | 32 | | 0.26 | 0.26 | 0.26 | 0.27 | 0.26 |
| | 64 | | 0.60 | 0.60 | 0.45 | 0.60 | 0.56 |
| | 128 | | 0.71 | 0.69 | 0.57 | 0.71 | 0.67 |
| | 256 | | 0.88 | 0.87 | 0.79 | 0.89 | 0.86 |
| | 512 | | 0.91 | 0.89 | 0.85 | 0.92 | 0.89 |
| | 1024 | | 0.90 | 0.89 | 0.88 | 0.92 | 0.90 |
| | 2048 | | 0.89 | 0.87 | 0.89 | 0.92 | 0.89 |

Table B.4: Average Cross-Validated RMSE of the Tuned SVR Models Built on Molecular Fingerprints and Tanimoto Kernel Descriptors with Bit Lengths from 32 to 2048

| Descriptor | Bit Length | SVR Kernel | | | | | Mean |
|---|---|---|---|---|---|---|---|
| | | Linear | Polynomial | RBF | Sigmoid | Precomputed | |
| Fingerprints | | | | | | | |
| Morgan1 | 32 | 15.2 | 7.8 | 7.4 | 22.2 | | 13.2 |
| | 64 | 15.5 | 8.0 | 7.5 | 19.9 | | 12.7 |
| | 128 | 15.5 | 8.0 | 7.5 | 19.8 | | 12.7 |
| | 256 | 15.5 | 7.9 | 7.4 | 19.6 | | 12.6 |
| | 512 | 15.5 | 8.0 | 7.5 | 19.5 | | 12.6 |
| | 1024 | 15.5 | 7.9 | 7.3 | 19.2 | | 12.5 |
| | 2048 | 15.5 | 7.9 | 7.3 | 19.2 | | 12.5 |
| Morgan2 | 32 | 15.0 | 8.1 | 7.1 | 22.8 | | 13.2 |
| | 64 | 15.5 | 8.0 | 6.7 | 20.1 | | 12.6 |
| | 128 | 15.5 | 7.8 | 6.7 | 19.9 | | 12.5 |
| | 256 | 15.5 | 7.8 | 6.6 | 19.5 | | 12.4 |
| | 512 | 15.5 | 7.8 | 6.6 | 19.2 | | 12.3 |
| | 1024 | 15.5 | 7.8 | 6.6 | 19.0 | | 12.2 |
| | 2048 | 15.5 | 7.8 | 6.6 | 19.0 | | 12.2 |
| Morgan3 | 32 | 15.0 | 8.5 | 7.9 | 23.3 | | 13.7 |
| | 64 | 15.0 | 8.0 | 6.9 | 20.2 | | 12.5 |
| | 128 | 15.0 | 8.0 | 6.8 | 19.8 | | 12.4 |
| | 256 | 15.0 | 7.9 | 6.7 | 19.3 | | 12.2 |
| | 512 | 15.0 | 8.0 | 6.6 | 19.4 | | 12.3 |
| | 1024 | 15.0 | 8.0 | 6.6 | 19.1 | | 12.2 |
| | 2048 | 15.0 | 7.9 | 6.6 | 19.1 | | 12.2 |
| FMorgan1 | 32 | | 21.4 | 21.6 | 21.9 | 21.2 | 21.5 |
| | 64 | | 20.9 | 21.1 | 21.4 | 20.6 | 21.0 |
| | 128 | | 20.4 | 20.5 | 21.0 | 20.0 | 20.5 |
| | 256 | | 20.1 | 20.3 | 20.5 | 19.8 | 20.2 |
| | 512 | | 20.1 | 20.3 | 20.5 | 19.7 | 20.2 |
| | 1024 | | 20.1 | 20.3 | 20.5 | 19.7 | 20.2 |
| | 2048 | | 20.1 | 20.3 | 20.5 | 19.7 | 20.2 |
| FMorgan2 | 32 | | 20.1 | 20.3 | 20.4 | 19.7 | 20.1 |
| | 64 | | 20.3 | 20.5 | 20.3 | 19.8 | 20.2 |
| | 128 | | 20.2 | 20.5 | 19.3 | 19.7 | 19.9 |
| | 256 | | 20.2 | 20.5 | 19.2 | 19.7 | 19.9 |
| | 512 | | 20.2 | 20.5 | 19.2 | 19.8 | 19.9 |
| | 1024 | | 20.3 | 20.6 | 19.1 | 19.8 | 20.0 |
| | 2048 | | 20.3 | 20.6 | 19.1 | 19.8 | 19.9 |
| FMorgan3 | 32 | | 20.0 | 20.2 | 20.3 | 19.7 | 20.1 |
| | 64 | | 20.3 | 20.6 | 19.2 | 19.9 | 20.0 |
| | 128 | | 20.4 | 20.7 | 19.2 | 20.0 | 20.1 |
| | 256 | | 20.5 | 20.8 | 19.2 | 20.0 | 20.1 |
| | 512 | | 20.5 | 20.9 | 19.2 | 20.1 | 20.2 |
| | 1024 | | 20.5 | 20.9 | 19.3 | 20.1 | 20.2 |
| | 2048 | | 20.5 | 20.9 | 19.2 | 20.1 | 20.2 |
| RDK | 32 | 23.3 | 23.6 | 23.5 | 27.6 | | 24.5 |
| | 64 | 18.0 | 17.2 | 17.3 | 27.6 | | 20.0 |
| | 128 | 16.9 | 14.4 | 14.5 | 24.9 | | 17.7 |
| | 256 | 15.1 | 8.8 | 8.9 | 20.8 | | 13.4 |
| | 512 | 15.0 | 8.1 | 7.9 | 23.5 | | 13.6 |
| | 1024 | 15.0 | 8.2 | 8.1 | 22.2 | | 13.4 |
| | 2048 | 15.0 | 8.4 | 8.3 | 22.1 | | 13.5 |
| Tanimoto | | | | | | | |
| Morgan1 | 32 | | 7.8 | 8.3 | 10.9 | 7.3 | 8.6 |
| | 64 | | 7.3 | 7.9 | 9.9 | 6.9 | 8.0 |
| | 128 | | 7.5 | 8.2 | 9.8 | 6.9 | 8.1 |
| | 256 | | 7.3 | 8.0 | 9.4 | 6.8 | 7.9 |
| | 512 | | 7.3 | 8.0 | 9.2 | 6.8 | 7.8 |
| | 1024 | | 7.1 | 7.8 | 9.0 | 6.5 | 7.6 |
| | 2048 | | 7.1 | 7.8 | 9.0 | 6.5 | 7.6 |

Table B.4: Average Cross-Validated RMSE of the Tuned SVR Models Built on Molecular Fingerprints and Tanimoto Kernel Descriptors with Bit Lengths from 32 to 2048 (Continued)

| Descriptor | Bit Length | SVR Kernel | | | | | Mean |
|---|---|---|---|---|---|---|---|
| | | Linear | Polynomial | RBF | Sigmoid | Precomputed | |
| Morgan2 | 32 | | 7.2 | 7.5 | 10.9 | 7.1 | 8.2 |
| | 64 | | 7.4 | 8.2 | 8.8 | 6.7 | 7.8 |
| | 128 | | 7.9 | 8.9 | 8.3 | 6.7 | 7.9 |
| | 256 | | 8.0 | 9.0 | 8.0 | 6.6 | 7.9 |
| | 512 | | 8.0 | 9.0 | 7.9 | 6.6 | 7.9 |
| | 1024 | | 8.0 | 9.1 | 7.8 | 6.6 | 7.9 |
| | 2048 | | 8.0 | 9.2 | 7.8 | 6.6 | 7.9 |
| Morgan3 | 32 | | 8.0 | 8.1 | 11.5 | 8.1 | 8.9 |
| | 64 | | 7.6 | 8.3 | 8.7 | 6.9 | 7.9 |
| | 128 | | 8.2 | 9.2 | 8.0 | 6.9 | 8.1 |
| | 256 | | 8.4 | 9.5 | 7.7 | 6.8 | 8.1 |
| | 512 | | 8.6 | 9.8 | 7.6 | 6.9 | 8.2 |
| | 1024 | | 8.8 | 10.1 | 7.5 | 6.9 | 8.3 |
| | 2048 | | 8.8 | 10.1 | 7.5 | 6.9 | 8.3 |
| FMorgan1 | 32 | | 21.4 | 21.6 | 21.9 | 21.2 | 21.5 |
| | 64 | | 20.9 | 21.1 | 21.4 | 20.6 | 21.0 |
| | 128 | | 20.4 | 20.5 | 21.0 | 20.0 | 20.5 |
| | 256 | | 20.1 | 20.3 | 20.5 | 19.8 | 20.2 |
| | 512 | | 20.1 | 20.3 | 20.5 | 19.7 | 20.2 |
| | 1024 | | 20.1 | 20.3 | 20.5 | 19.7 | 20.2 |
| | 2048 | | 20.1 | 20.3 | 20.5 | 19.7 | 20.2 |
| FMorgan2 | 32 | | 20.1 | 20.3 | 20.4 | 19.7 | 20.1 |
| | 64 | | 20.3 | 20.5 | 20.3 | 19.8 | 20.2 |
| | 128 | | 20.2 | 20.5 | 19.3 | 19.7 | 19.9 |
| | 256 | | 20.2 | 20.5 | 19.2 | 19.7 | 19.9 |
| | 512 | | 20.2 | 20.5 | 19.2 | 19.8 | 19.9 |
| | 1024 | | 20.3 | 20.6 | 19.1 | 19.8 | 20.0 |
| | 2048 | | 20.3 | 20.6 | 19.1 | 19.8 | 19.9 |
| FMorgan3 | 32 | | 20.0 | 20.2 | 20.3 | 19.7 | 20.1 |
| | 64 | | 20.3 | 20.6 | 19.2 | 19.9 | 20.0 |
| | 128 | | 20.4 | 20.7 | 19.2 | 20.0 | 20.1 |
| | 256 | | 20.5 | 20.8 | 19.2 | 20.0 | 20.1 |
| | 512 | | 20.5 | 20.9 | 19.2 | 20.1 | 20.2 |
| | 1024 | | 20.5 | 20.9 | 19.3 | 20.1 | 20.2 |
| | 2048 | | 20.5 | 20.9 | 19.2 | 20.1 | 20.2 |
| RDK | 32 | | 23.4 | 23.4 | 23.5 | 23.3 | 23.4 |
| | 64 | | 17.2 | 17.3 | 20.3 | 17.2 | 18.0 |
| | 128 | | 14.8 | 15.1 | 17.8 | 14.6 | 15.6 |
| | 256 | | 9.5 | 9.8 | 12.4 | 9.1 | 10.2 |
| | 512 | | 8.4 | 8.9 | 10.5 | 7.9 | 8.9 |
| | 1024 | | 8.5 | 9.2 | 9.4 | 7.8 | 8.7 |
| | 2048 | | 8.8 | 9.7 | 9.2 | 7.8 | 8.9 |

Table B.5: Average Cross-Validated Coefficient of Determination of the Tuned Tree-Based Models Built on Molecular Fingerprints and Tanimoto Kernel Descriptors with Bit Lengths from 32 to 2048

| Descriptor | Bit Length | Tree-based Models | | | Mean |
|---|---|---|---|---|---|
| | | Decision Tree | Gradient Boosting | Random Forest | |
| Fingerprints | | | | | |
| Morgan1 | 32 | 0.86 | 0.90 | 0.92 | 0.90 |
| | 64 | 0.87 | 0.91 | 0.93 | 0.90 |
| | 128 | 0.88 | 0.91 | 0.93 | 0.91 |
| | 256 | 0.88 | 0.91 | 0.93 | 0.91 |
| | 512 | 0.88 | 0.91 | 0.93 | 0.91 |
| | 1024 | 0.88 | 0.91 | 0.94 | 0.91 |
| | 2048 | 0.88 | 0.91 | 0.94 | 0.91 |
| Morgan2 | 32 | 0.87 | 0.91 | 0.93 | 0.90 |
| | 64 | 0.86 | 0.91 | 0.93 | 0.90 |
| | 128 | 0.86 | 0.91 | 0.93 | 0.90 |
| | 256 | 0.86 | 0.91 | 0.93 | 0.90 |
| | 512 | 0.88 | 0.91 | 0.93 | 0.91 |
| | 1024 | 0.88 | 0.92 | 0.94 | 0.91 |
| | 2048 | 0.88 | 0.92 | 0.94 | 0.91 |
| Morgan3 | 32 | 0.84 | 0.90 | 0.92 | 0.89 |
| | 64 | 0.89 | 0.91 | 0.93 | 0.91 |
| | 128 | 0.85 | 0.91 | 0.93 | 0.90 |
| | 256 | 0.85 | 0.91 | 0.93 | 0.90 |
| | 512 | 0.87 | 0.91 | 0.93 | 0.90 |
| | 1024 | 0.88 | 0.91 | 0.93 | 0.91 |
| | 2048 | 0.88 | 0.92 | 0.93 | 0.91 |
| FMorgan1 | 32 | 0.14 | 0.31 | 0.23 | 0.23 |
| | 64 | 0.15 | 0.35 | 0.24 | 0.25 |
| | 128 | 0.17 | 0.39 | 0.28 | 0.28 |
| | 256 | 0.18 | 0.39 | 0.28 | 0.28 |
| | 512 | 0.17 | 0.39 | 0.28 | 0.28 |
| | 1024 | 0.18 | 0.39 | 0.29 | 0.29 |
| | 2048 | 0.18 | 0.39 | 0.29 | 0.29 |
| FMorgan2 | 32 | 0.12 | 0.37 | 0.25 | 0.25 |
| | 64 | 0.12 | 0.37 | 0.25 | 0.25 |
| | 128 | 0.12 | 0.37 | 0.26 | 0.25 |
| | 256 | 0.12 | 0.38 | 0.26 | 0.25 |
| | 512 | 0.12 | 0.38 | 0.26 | 0.25 |
| | 1024 | 0.12 | 0.38 | 0.25 | 0.25 |
| | 2048 | 0.12 | 0.38 | 0.25 | 0.25 |
| FMorgan3 | 32 | 0.12 | 0.37 | 0.26 | 0.25 |
| | 64 | 0.11 | 0.36 | 0.25 | 0.24 |
| | 128 | 0.12 | 0.37 | 0.25 | 0.25 |
| | 256 | 0.12 | 0.37 | 0.25 | 0.25 |
| | 512 | 0.11 | 0.38 | 0.25 | 0.25 |
| | 1024 | 0.11 | 0.37 | 0.25 | 0.24 |
| | 2048 | 0.12 | 0.38 | 0.25 | 0.25 |
| RDK | 32 | 0.27 | 0.28 | 0.27 | 0.28 |
| | 64 | 0.58 | 0.62 | 0.59 | 0.60 |
| | 128 | 0.45 | 0.74 | 0.63 | 0.61 |
| | 256 | 0.84 | 0.89 | 0.90 | 0.88 |
| | 512 | 0.86 | 0.90 | 0.93 | 0.90 |
| | 1024 | 0.87 | 0.90 | 0.93 | 0.90 |
| | 2048 | 0.85 | 0.90 | 0.93 | 0.89 |
| Tanimoto | | | | | |
| Morgan1 | 32 | 0.45 | 0.81 | 0.75 | 0.67 |
| | 64 | 0.45 | 0.82 | 0.76 | 0.68 |
| | 128 | 0.31 | 0.81 | 0.70 | 0.61 |
| | 256 | 0.39 | 0.82 | 0.73 | 0.65 |
| | 512 | 0.43 | 0.81 | 0.73 | 0.66 |
| | 1024 | 0.44 | 0.83 | 0.75 | 0.67 |
| | 2048 | 0.43 | 0.83 | 0.76 | 0.67 |

Table B.5: Average Cross-Validated Coefficient of Determination of the Tuned Tree-Based Models Built on Molecular Fingerprints and Tanimoto Kernel Descriptors with Bit Lengths from 32 to 2048 (Continued)

| Descriptor | Bit Length | Tree-based Models | | | Mean |
|---|---|---|---|---|---|
| | | Decision Tree | Gradient Boosting | Random Forest | |
| Morgan2 | 32 | 0.59 | 0.84 | 0.80 | 0.74 |
| | 64 | 0.44 | 0.83 | 0.75 | 0.67 |
| | 128 | 0.41 | 0.81 | 0.72 | 0.64 |
| | 256 | 0.43 | 0.81 | 0.72 | 0.65 |
| | 512 | 0.37 | 0.80 | 0.72 | 0.63 |
| | 1024 | 0.44 | 0.82 | 0.74 | 0.67 |
| | 2048 | 0.46 | 0.82 | 0.75 | 0.68 |
| Morgan3 | 32 | 0.55 | 0.82 | 0.78 | 0.72 |
| | 64 | 0.53 | 0.83 | 0.78 | 0.72 |
| | 128 | 0.48 | 0.83 | 0.76 | 0.69 |
| | 256 | 0.43 | 0.82 | 0.74 | 0.66 |
| | 512 | 0.37 | 0.79 | 0.72 | 0.63 |
| | 1024 | 0.40 | 0.80 | 0.73 | 0.65 |
| | 2048 | 0.41 | 0.80 | 0.72 | 0.64 |
| FMorgan1 | 32 | 0.14 | 0.31 | 0.23 | 0.23 |
| | 64 | 0.15 | 0.35 | 0.24 | 0.25 |
| | 128 | 0.17 | 0.39 | 0.28 | 0.28 |
| | 256 | 0.18 | 0.39 | 0.28 | 0.28 |
| | 512 | 0.17 | 0.39 | 0.28 | 0.28 |
| | 1024 | 0.18 | 0.39 | 0.29 | 0.29 |
| | 2048 | 0.18 | 0.39 | 0.29 | 0.29 |
| FMorgan2 | 32 | 0.12 | 0.37 | 0.25 | 0.25 |
| | 64 | 0.12 | 0.37 | 0.25 | 0.25 |
| | 128 | 0.12 | 0.37 | 0.26 | 0.25 |
| | 256 | 0.12 | 0.38 | 0.26 | 0.25 |
| | 512 | 0.12 | 0.38 | 0.26 | 0.25 |
| | 1024 | 0.12 | 0.38 | 0.25 | 0.25 |
| | 2048 | 0.12 | 0.38 | 0.25 | 0.25 |
| FMorgan3 | 32 | 0.12 | 0.37 | 0.26 | 0.25 |
| | 64 | 0.11 | 0.36 | 0.25 | 0.24 |
| | 128 | 0.12 | 0.37 | 0.25 | 0.25 |
| | 256 | 0.12 | 0.37 | 0.25 | 0.25 |
| | 512 | 0.11 | 0.38 | 0.25 | 0.25 |
| | 1024 | 0.11 | 0.37 | 0.25 | 0.24 |
| | 2048 | 0.12 | 0.38 | 0.25 | 0.25 |
| RDK | 32 | 0.27 | 0.27 | 0.27 | 0.27 |
| | 64 | 0.58 | 0.60 | 0.59 | 0.59 |
| | 128 | 0.39 | 0.66 | 0.59 | 0.55 |
| | 256 | 0.56 | 0.79 | 0.75 | 0.70 |
| | 512 | 0.44 | 0.79 | 0.72 | 0.65 |
| | 1024 | 0.38 | 0.79 | 0.70 | 0.62 |
| | 2048 | 0.37 | 0.77 | 0.70 | 0.61 |

Table B.6: Average Cross-Validated RMSE of the Tuned Tree-Based Models Built on Molecular Fingerprints and Tanimoto Kernel Descriptors with Bit Lengths from 32 to 2048

| Descriptor | Bit Length | Tree-based Models | | | Mean |
|---|---|---|---|---|---|
| | | Decision Tree | Gradient Boosting | Random Forest | |
| Fingerprints | | | | | |
| Morgan1 | 32 | 10.1 | 8.5 | 7.7 | 8.8 |
| | 64 | 9.7 | 8.4 | 7.2 | 8.4 |
| | 128 | 9.6 | 8.3 | 7.1 | 8.3 |
| | 256 | 9.5 | 8.3 | 7.1 | 8.3 |
| | 512 | 9.4 | 8.2 | 7.1 | 8.2 |
| | 1024 | 9.6 | 8.1 | 6.8 | 8.2 |
| | 2048 | 9.5 | 8.1 | 6.8 | 8.1 |
| Morgan2 | 32 | 9.9 | 8.4 | 7.3 | 8.5 |
| | 64 | 10.1 | 8.1 | 7.1 | 8.4 |
| | 128 | 10.2 | 8.0 | 7.2 | 8.5 |
| | 256 | 10.2 | 8.1 | 7.2 | 8.5 |
| | 512 | 9.6 | 8.2 | 7.1 | 8.3 |
| | 1024 | 9.5 | 7.9 | 7.0 | 8.1 |
| | 2048 | 9.4 | 7.8 | 6.9 | 8.0 |
| Morgan3 | 32 | 10.9 | 8.6 | 7.6 | 9.0 |
| | 64 | 9.1 | 8.3 | 7.2 | 8.2 |
| | 128 | 10.4 | 8.1 | 7.5 | 8.7 |
| | 256 | 10.4 | 8.1 | 7.4 | 8.6 |
| | 512 | 9.9 | 8.1 | 7.2 | 8.4 |
| | 1024 | 9.6 | 8.0 | 7.0 | 8.2 |
| | 2048 | 9.4 | 7.9 | 7.0 | 8.1 |
| FMorgan1 | 32 | 25.3 | 22.6 | 23.9 | 23.9 |
| | 64 | 25.2 | 22.0 | 23.7 | 23.6 |
| | 128 | 24.8 | 21.3 | 23.1 | 23.1 |
| | 256 | 24.8 | 21.3 | 23.1 | 23.1 |
| | 512 | 24.8 | 21.3 | 23.1 | 23.1 |
| | 1024 | 24.7 | 21.2 | 23.1 | 23.0 |
| | 2048 | 24.7 | 21.2 | 23.1 | 23.0 |
| FMorgan2 | 32 | 25.6 | 21.7 | 23.5 | 23.6 |
| | 64 | 25.6 | 21.6 | 23.6 | 23.6 |
| | 128 | 25.6 | 21.6 | 23.5 | 23.6 |
| | 256 | 25.6 | 21.5 | 23.5 | 23.5 |
| | 512 | 25.6 | 21.6 | 23.5 | 23.6 |
| | 1024 | 25.5 | 21.5 | 23.6 | 23.5 |
| | 2048 | 25.6 | 21.5 | 23.6 | 23.6 |
| FMorgan3 | 32 | 25.6 | 21.6 | 23.5 | 23.6 |
| | 64 | 25.7 | 21.8 | 23.6 | 23.7 |
| | 128 | 25.6 | 21.6 | 23.6 | 23.6 |
| | 256 | 25.6 | 21.6 | 23.6 | 23.6 |
| | 512 | 25.7 | 21.5 | 23.5 | 23.6 |
| | 1024 | 25.7 | 21.7 | 23.6 | 23.6 |
| | 2048 | 25.5 | 21.6 | 23.6 | 23.6 |
| RDK | 32 | 23.3 | 23.1 | 23.2 | 23.2 |
| | 64 | 17.7 | 16.8 | 17.5 | 17.3 |
| | 128 | 20.2 | 13.8 | 16.7 | 16.9 |
| | 256 | 10.8 | 8.9 | 8.6 | 9.4 |
| | 512 | 10.2 | 8.5 | 7.4 | 8.7 |
| | 1024 | 10.0 | 8.4 | 7.3 | 8.6 |
| | 2048 | 10.4 | 8.4 | 7.4 | 8.7 |
| Tanimoto | | | | | |
| Morgan1 | 32 | 20.2 | 11.8 | 13.6 | 15.2 |
| | 64 | 20.2 | 11.6 | 13.5 | 15.1 |
| | 128 | 22.6 | 12.0 | 14.9 | 16.5 |
| | 256 | 21.3 | 11.6 | 14.1 | 15.7 |
| | 512 | 20.6 | 11.8 | 14.0 | 15.5 |
| | 1024 | 20.3 | 11.3 | 13.5 | 15.0 |
| | 2048 | 20.5 | 11.3 | 13.5 | 15.1 |

Table B.6: Average Cross-Validated RMSE of the Tuned Tree-Based Models Built on Molecular Fingerprints and Tanimoto Kernel Descriptors with Bit Lengths from 32 to 2048 (Continued)

| Descriptor | Bit Length | Tree-based Models | | | Mean |
|---|---|---|---|---|---|
| | | Decision Tree | Gradient Boosting | Random Forest | |
| Morgan2 | 32 | 17.6 | 11.0 | 12.1 | 13.5 |
| | 64 | 20.4 | 11.4 | 13.6 | 15.1 |
| | 128 | 21.0 | 11.9 | 14.5 | 15.8 |
| | 256 | 20.5 | 11.9 | 14.4 | 15.6 |
| | 512 | 21.6 | 12.2 | 14.4 | 16.1 |
| | 1024 | 20.5 | 11.6 | 13.8 | 15.3 |
| | 2048 | 19.9 | 11.5 | 13.7 | 15.0 |
| Morgan3 | 32 | 18.4 | 11.6 | 12.8 | 14.2 |
| | 64 | 18.5 | 11.1 | 12.7 | 14.1 |
| | 128 | 19.7 | 11.4 | 13.4 | 14.8 |
| | 256 | 20.6 | 11.7 | 13.9 | 15.4 |
| | 512 | 21.7 | 12.4 | 14.5 | 16.2 |
| | 1024 | 21.1 | 12.2 | 14.0 | 15.8 |
| | 2048 | 20.9 | 12.3 | 14.4 | 15.9 |
| FMorgan1 | 32 | 25.3 | 22.6 | 23.9 | 23.9 |
| | 64 | 25.2 | 22.0 | 23.7 | 23.6 |
| | 128 | 24.8 | 21.3 | 23.1 | 23.1 |
| | 256 | 24.8 | 21.3 | 23.1 | 23.1 |
| | 512 | 24.8 | 21.3 | 23.1 | 23.1 |
| | 1024 | 24.7 | 21.2 | 23.1 | 23.0 |
| | 2048 | 24.7 | 21.2 | 23.1 | 23.0 |
| FMorgan2 | 32 | 25.6 | 21.7 | 23.5 | 23.6 |
| | 64 | 25.6 | 21.6 | 23.6 | 23.6 |
| | 128 | 25.6 | 21.6 | 23.5 | 23.6 |
| | 256 | 25.6 | 21.5 | 23.5 | 23.5 |
| | 512 | 25.6 | 21.6 | 23.5 | 23.6 |
| | 1024 | 25.5 | 21.5 | 23.6 | 23.5 |
| | 2048 | 25.6 | 21.5 | 23.6 | 23.6 |
| FMorgan3 | 32 | 25.6 | 21.6 | 23.5 | 23.6 |
| | 64 | 25.7 | 21.8 | 23.6 | 23.7 |
| | 128 | 25.6 | 21.6 | 23.6 | 23.6 |
| | 256 | 25.6 | 21.6 | 23.6 | 23.6 |
| | 512 | 25.7 | 21.5 | 23.5 | 23.6 |
| | 1024 | 25.7 | 21.7 | 23.6 | 23.6 |
| | 2048 | 25.5 | 21.6 | 23.6 | 23.6 |
| RDK | 32 | 23.3 | 23.3 | 23.2 | 23.2 |
| | 64 | 17.7 | 17.2 | 17.5 | 17.5 |
| | 128 | 21.2 | 15.9 | 17.5 | 18.2 |
| | 256 | 18.1 | 12.3 | 13.7 | 14.7 |
| | 512 | 20.4 | 12.4 | 14.4 | 15.8 |
| | 1024 | 21.5 | 12.5 | 15.0 | 16.3 |
| | 2048 | 21.6 | 13.1 | 14.8 | 16.5 |

### Descriptors Derived from Molecular Graphs

Table B.7: Average Cross-Validated Performance of the Linear Models Built on the WL Kernel

| WL Depth | Linear Models | | | | | Mean |
|---|---|---|---|---|---|---|
| | Linear Regression | Lasso | Ridge | Elastic Net | Bayesian Ridge | |
| Mean $R^2$ | | | | | | |
| 2 | <-1.00 | 0.85 | 0.86 | 0.63 | 0.86 | <-1.00 |
| 3 | 0.90 | 0.92 | 0.93 | 0.66 | 0.93 | 0.87 |
| 4 | 0.91 | 0.92 | 0.93 | 0.67 | 0.93 | 0.87 |
| 5 | 0.92 | 0.93 | 0.93 | 0.67 | 0.93 | 0.87 |
| 6 | 0.93 | 0.93 | 0.93 | 0.67 | 0.93 | 0.88 |
| 7 | 0.93 | 0.93 | 0.93 | 0.66 | 0.93 | 0.88 |
| 8 | 0.93 | 0.93 | 0.93 | 0.65 | 0.93 | 0.87 |
| 9 | 0.93 | 0.93 | 0.93 | 0.64 | 0.93 | 0.87 |
| 10 | 0.93 | 0.93 | 0.93 | 0.63 | 0.93 | 0.87 |
| Mean RMSE (%) | | | | | | |
| 2 | >100.0 | 10.5 | 10.2 | 16.7 | 10.3 | >100.0 |
| 3 | 8.6 | 7.5 | 7.3 | 16.0 | 7.4 | 9.4 |
| 4 | 8.0 | 7.5 | 7.3 | 15.7 | 7.4 | 9.2 |
| 5 | 7.6 | 7.4 | 7.3 | 15.7 | 7.3 | 9.1 |
| 6 | 7.4 | 7.3 | 7.3 | 15.7 | 7.3 | 9.0 |
| 7 | 7.3 | 7.3 | 7.2 | 15.9 | 7.3 | 9.0 |
| 8 | 7.3 | 7.3 | 7.2 | 16.1 | 7.2 | 9.0 |
| 9 | 7.3 | 7.3 | 7.3 | 16.3 | 7.3 | 9.1 |
| 10 | 7.3 | 7.3 | 7.3 | 16.6 | 7.3 | 9.1 |

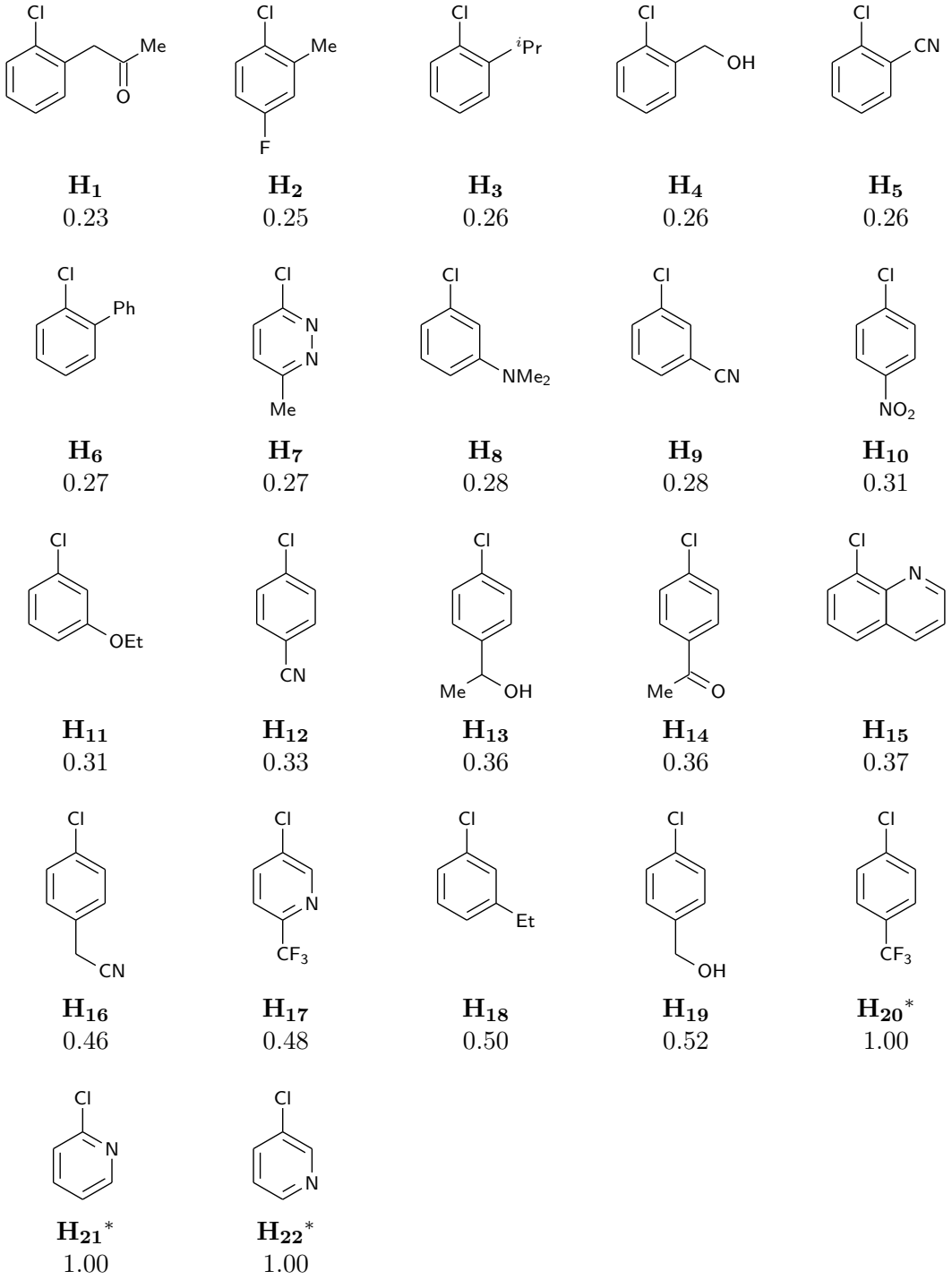
Table B.8: Average Cross-Validated Performance of the SVR Models Built on the WL Kernel

| WL Depth | SVR Kernel | | | | Mean |
|---|---|---|---|---|---|
| | Polynomial | RBF | Sigmoid | Precomputed | |
| Mean $R^2$ | | | | | |
| 2 | 0.92 | 0.93 | 0.60 | 0.85 | 0.83 |
| 3 | 0.93 | 0.92 | 0.68 | 0.93 | 0.86 |
| 4 | 0.93 | 0.92 | 0.72 | 0.92 | 0.87 |
| 5 | 0.93 | 0.92 | 0.83 | 0.92 | 0.90 |
| 6 | 0.93 | 0.92 | 0.85 | 0.93 | 0.91 |
| 7 | 0.92 | 0.92 | 0.87 | 0.93 | 0.91 |
| 8 | 0.92 | 0.91 | 0.88 | 0.93 | 0.91 |
| 9 | 0.92 | 0.91 | 0.89 | 0.93 | 0.91 |
| 10 | 0.92 | 0.91 | 0.89 | 0.93 | 0.91 |
| Mean RMSE (%) | | | | | |
| 2 | 7.5 | 7.4 | 17.2 | 10.5 | 10.7 |
| 3 | 7.4 | 7.5 | 15.5 | 7.4 | 9.5 |
| 4 | 7.5 | 7.6 | 14.3 | 7.5 | 9.2 |
| 5 | 7.4 | 7.7 | 11.2 | 7.5 | 8.5 |
| 6 | 7.4 | 7.7 | 10.4 | 7.4 | 8.2 |
| 7 | 7.5 | 7.8 | 9.9 | 7.3 | 8.1 |
| 8 | 7.6 | 8.0 | 9.4 | 7.3 | 8.1 |
| 9 | 7.7 | 8.2 | 9.1 | 7.3 | 8.1 |
| 10 | 7.8 | 8.4 | 8.9 | 7.3 | 8.1 |

Table B.9: Average Cross-Validated Performance of the Tree-Based Models Built on the WL Kernel

| WL Depth | Tree-Based Models | | | Mean |
|---|---|---|---|---|
| | Decision Tree | Gradient Boosting | Random Forest | |
| Mean $R^2$ | | | | |
| 2 | 0.59 | 0.84 | 0.82 | 0.75 |
| 3 | 0.58 | 0.84 | 0.80 | 0.74 |
| 4 | 0.58 | 0.86 | 0.80 | 0.75 |
| 5 | 0.51 | 0.86 | 0.78 | 0.72 |
| 6 | 0.55 | 0.86 | 0.80 | 0.74 |
| 7 | 0.54 | 0.87 | 0.79 | 0.73 |
| 8 | 0.61 | 0.86 | 0.81 | 0.76 |
| 9 | 0.59 | 0.85 | 0.80 | 0.75 |
| 10 | 0.56 | 0.85 | 0.79 | 0.73 |
| Mean RMSE (%) | | | | |
| 2 | 17.4 | 10.9 | 11.7 | 13.3 |
| 3 | 17.5 | 10.8 | 12.0 | 13.5 |
| 4 | 17.5 | 10.3 | 12.1 | 13.3 |
| 5 | 19.1 | 10.2 | 12.8 | 14.1 |
| 6 | 18.2 | 10.0 | 12.3 | 13.5 |
| 7 | 18.4 | 10.0 | 12.4 | 13.6 |
| 8 | 16.9 | 10.1 | 11.8 | 12.9 |
| 9 | 17.4 | 10.4 | 12.1 | 13.3 |
| 10 | 17.9 | 10.5 | 12.6 | 13.7 |

## B.4 Prospective Buchwald-Hartwig Reactions




**H$_1$** 0.23 | **H$_2$** 0.25 | **H$_3$** 0.26 | **H$_4$** 0.26 | **H$_5$** 0.26

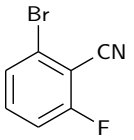
**H$_6$** 0.27 | **H$_7$** 0.27 | **H$_8$** 0.28 | **H$_9$** 0.28 | **H$_{10}$** 0.31

**H$_{11}$** 0.31 | **H$_{12}$** 0.33 | **H$_{13}$** 0.36 | **H$_{14}$** 0.36 | **H$_{15}$** 0.37

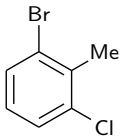
**H$_{16}$** 0.46 | **H$_{17}$** 0.48 | **H$_{18}$** 0.50 | **H$_{19}$** 0.52 | **H$_{20}$*** 1.00

**H$_{21}$*** 1.00 | **H$_{22}$*** 1.00

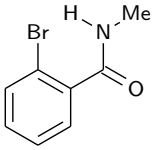
Figure B.5: Aryl chlorides in the prospective Buchwald-Hartwig reactions. *Molecules present in the Buchwald-Hartwig dataset, compiled by Doyle *et al.*[32] H$_n$, key; number, maximum Tanimoto similarity score (with the Morgan2 fingerprint) to the aryl halides in the Buchwald-Hartwig reactions.



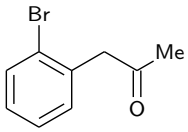

**H23**
0.19



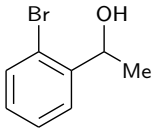

**H24**
0.21



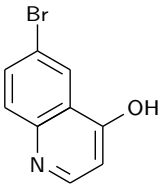

**H25**
0.22



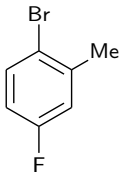

**H26**
0.23



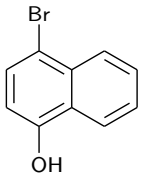

**H27**
0.24



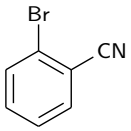

**H28**
0.24



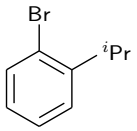

**H29**
0.25



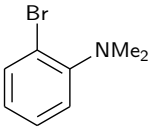

**H30**
0.25



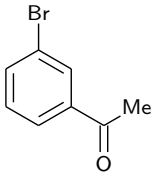

**H31**
0.26



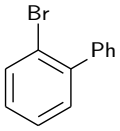

**H32**
0.26



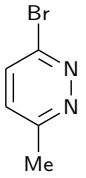

**H33**
0.26



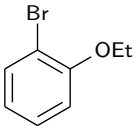

**H34**
0.27



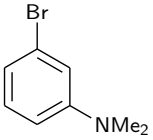

**H35**
0.27



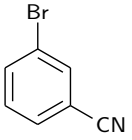

**H36**
0.27



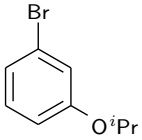

**H37**
0.28



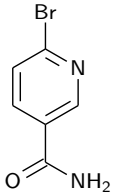

**H38**
0.28



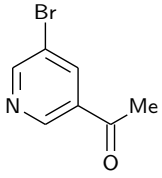

**H39**
0.28



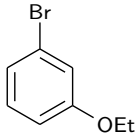

**H40**
0.30



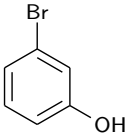

**H41**
0.30



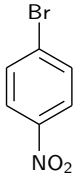

**H42**
0.30



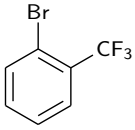

**H43**
0.30



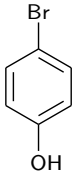

**H44**
0.31



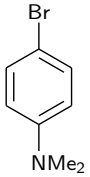

**H45**
0.31



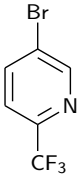

**H46**
0.36



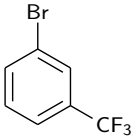

**H47**
0.38



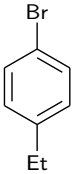

**H48**
0.39




**H49**
0.42

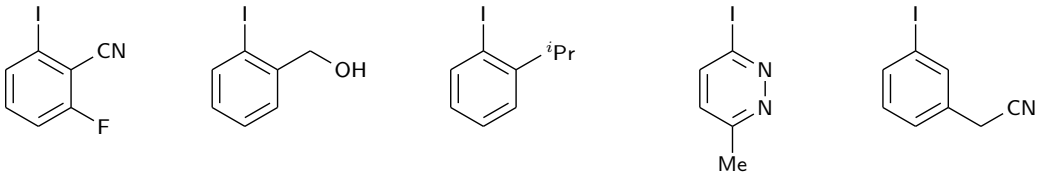



**H50**
0.50

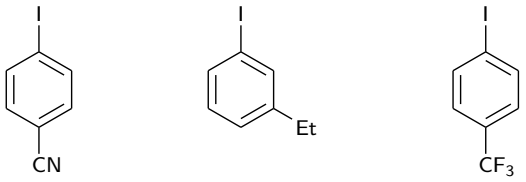



**H51***
1.00

Figure B.6: Aryl bromides in the prospective Buchwald-Hartwig reactions. *Molecules present in the Buchwald-Hartwig dataset, compiled by Doyle *et al.*[32] $H_n$, key; number, maximum Tanimoto similarity score (with the Morgan2 fingerprint) to the aryl halides in the Buchwald-Hartwig reactions.

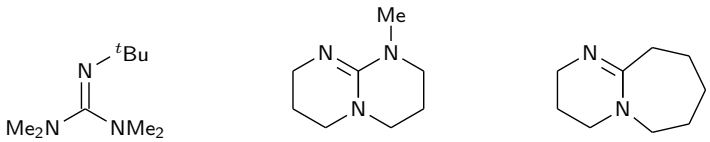



**H52** 0.23 | **H53** 0.26 | **H54** 0.26 | **H55** 0.27 | **H56** 0.29




**H57** 0.33 | **H58** 0.50 | **H59*** 1.00

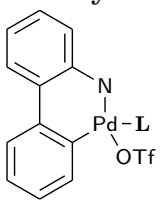
Figure B.7: Aryl iodides in the prospective Buchwald-Hartwig reactions. *Molecules present in the Buchwald-Hartwig dataset, compiled by Doyle *et al.*[32] $H_n$, key; number, maximum Tanimoto similarity score (with the Morgan2 fingerprint) to the aryl halides in the Buchwald-Hartwig reactions.



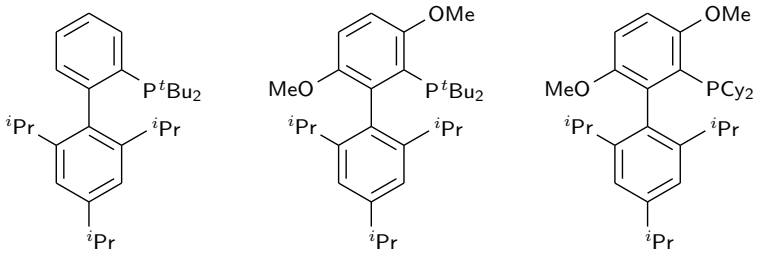

**BTMG*** **MTBD*** **DBU**

Figure B.8: Bases in the prospective Buchwald-Hartwig reactions. *Molecules present in the Buchwald-Hartwig dataset, compiled by Doyel *et al.*[32]

**Pd Catalyst:**

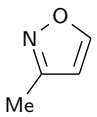



**L:**




**t-BuXPhos*** **t-BuBrettPhos*** **BrettPhos**

Figure B.9: Catalyst ligands in the prospective Buchwald-Hartwig reactions. *Molecules present in the Buchwald-Hartwig dataset, compiled by Doyel *et al.*[32]




**3-methylisoxazole***

Figure B.10: Additive in the prospective Buchwald-Hartwig reactions. *Molecules present in the Buchwald-Hartwig dataset, compiled by Doyel *et al.*[32]

## B.5 Quantum Chemical Descriptors of the Prospective Reactions

Quantum chemical descriptors were calculated for the 49 aryl halides and the base, DBU, present in the prospective reactions. Density functional theory (DFT) geometry optimizations and property calculations were performed using the Spartan '14 software package with a combination of the B3LYP exchange-correlation functional and 6-31G(d) basis set.[236,237] Properties, including electrostatic charges, vibrational frequencies, and $^{13}C/^{1}H$ NMR chemical shifts, were calculated at the optimized geometries. Improved tolerances and thresholds were set throughout, using the keywords: SCFTOLERANCE = VERYHIGH, GRADIENTTOLERANCE = 0.000005, DISTANCETOLERANCE = 0.00002, and BIGGRID. Additional molecular descriptors including the molecular weight, the energy of the HOMO and LUMO orbitals, and the total dipole moment, as well as Quantitative Structure-Activity Relationship (QSAR) descriptors for molecular volume, surface area, ovality, electronegativity, and hardness were also extracted for each species. The NMR and electrostatic descriptors were extracted for the set of six shared atoms previously determined for the aryl halide reagent class in the work by Doyle *et al.*[32] (Chapter 4 Figure 4.1), with C1-C4 and H1-H2 numbered giving the C1 label to the carbon bound to the heaviest halide for which the associated atom pattern exists. In cases where the two sides of the aryl halide are not symmetrically equivalent and two potential C2/C3 and H1/H2 atoms exist, the labels have been assigned to the atoms on the side with the lowest total mass. Vibrational frequencies and intensities were extracted for the three shared vibrations previously determined for the aryl halide reagent class by Doyle *et al.*,[32] and have been identified by visual inspection of the atomic displacements along each normal coordinate. Electrostatic descriptors were extracted for shared nitrogen atom (N1) in the base reagent class (Chapter 4 Figure 4.1).

### Issues Calculating Quantum Chemical Descriptors for the Aryl Iodides

The Spartan files used in the calculations of the quantum chemical descriptors by Doyle *et al.*[32] are given in the GitHub repository: `https://github.com/doylelab/rxnpredict`.[2] According to the supporting information, the Spartan DFT calculations (including NMR calculations) were performed using Spartan '14 V1.1.4. Their program used a series of scripts that submitted the calculations via the command line (B3LYP/6-31G(d)). There is no explicit reference to the use of a pseudopotential when performing the calculations on the aryl iodides.

The 6-31G(d) was used for all molecules except the aryl iodides, which used a mixed basis set: 6-31G(d) and LANL2DZ>kr (Figure B.11). The LANL2DZ>kr pseudopotential and basis set was used for the iodine atom and gave a chemical NMR shift of 424.51ppm when using the 1-ethyl-4-iodobenzene molecule as a test case.

(a) 1-bromo-4-(trifluoromethyl)benzene/M0001/output

```
SPARTAN '14 Quantum Mechanics Driver:  (Win/64b)

Job type: Geometry optimization.
Method: RB3LYP
Basis set: 6-31G(D)
Number of shells: 56
Number of basis functions: 193
Multiplicity: 1
Parallel Job: 4 threads
```

(b) 1-iodo-4-(trifluoromethyl)benzene/M0001/output

```
SPARTAN '14 Quantum Mechanics Driver:  (Win/64b)

Job type: Geometry optimization.
Method: RB3LYP
Basis set: 6-31G* & LANL2DZ>Kr
Number of shells: 53
Number of basis functions: 172
Multiplicity: 1
Parallel Job: 4 threads
```

Figure B.11: The Spartan output file for (a) 1-bromo-4-(trifluoromethyl)benzene and (b) 1-iodo-4-(trifluoromethyl)benzene.[2]

An analogous NMR calculation was performed, using Win/64b Spartan '14 V1.1.8, through the standard graphical user interface. The output geometry of 1-ethyl-iodobenzene was used as the structure and resulted in the following error message, as shown in Figure B.12, "the calculation failed: NMR not allowed for ECP atoms". The same calculation was performed using an up to date developers copy of Q-Chem from September 2020, as Q-Chem constitutes the back-end for most of the Spartan software package. The calculation also failed, with the error "NMR code does not handle pseudopotentials". The same calculation performed for a third time with a much earlier Win/64b Spartan '12 V1.1.0, using the graphical user interface, did run. However, it gave a chemical shift of 588.11ppm for the Iodine atom, which is an error of 163.60 (ca. 28%) compared to the result published by Doyle *et al.*[32]. In this work, aryl iodides were included in the initial model development for consistency with the open dataset, but, due to the ambiguity in the quantum chemical calculations, were not included in the yield predictions of the prospective reactions.

## B.6 Diversity of the Buchwald-Hartwig Dataset

### B.6.1 Chemical Reactivity

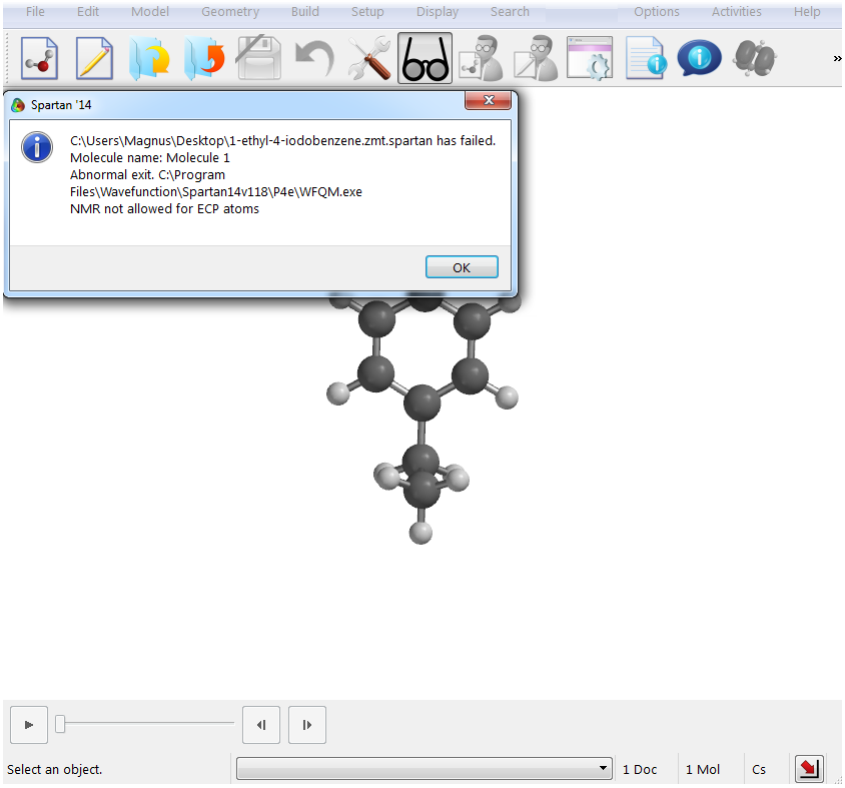



Figure B.12: Error message from running an analogous NMR calculation on the 1-ethyl-iodobenzene output geometry.[2]

(a)

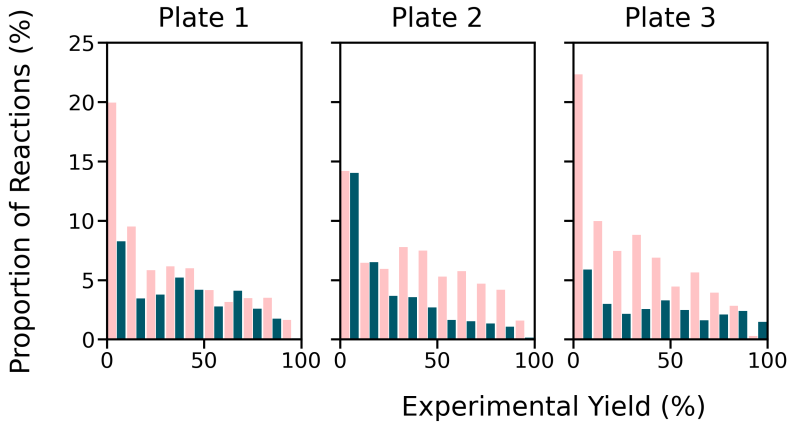
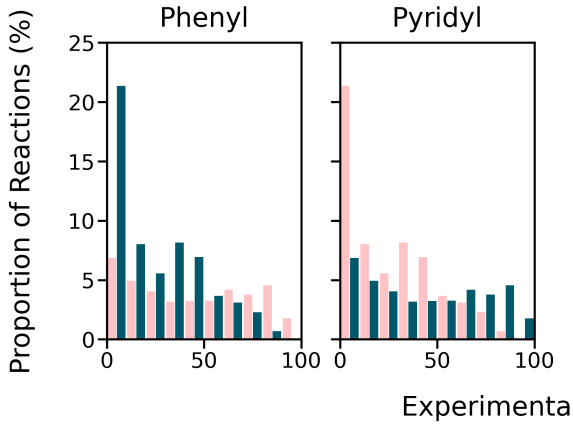
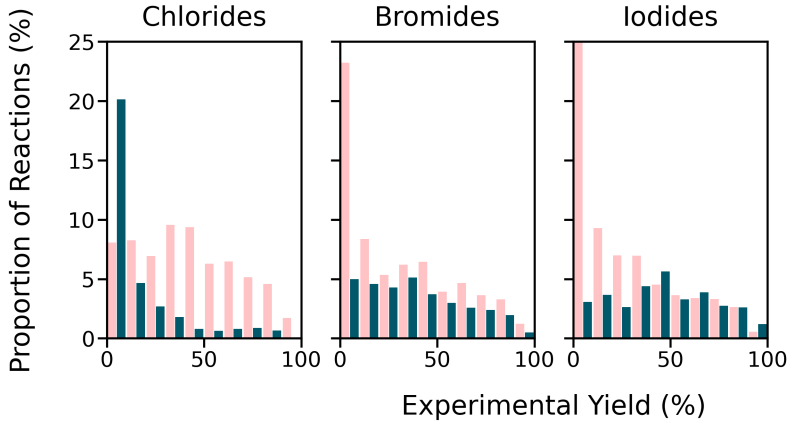


Plate 1
Plate 2
Plate 3
Proportion of Reactions (%)
Experimental Yield (%)
0
5
10
15
20
25
0
50
100

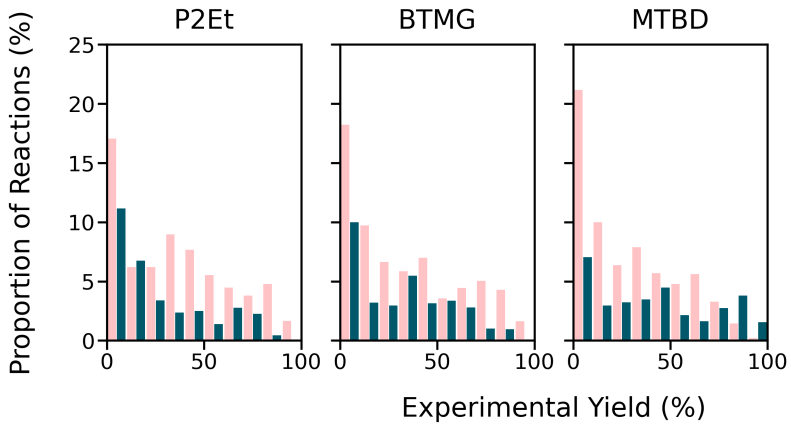
(b)


Phenyl
Pyridyl
Proportion of Reactions (%)
Experimental Yield (%)
0
5
10
15
20
25
0
50
100

(c)


Chlorides
Bromides
Iodides
Proportion of Reactions (%)
Experimental Yield (%)
0
5
10
15
20
25
0
50
100

(d)




(e)

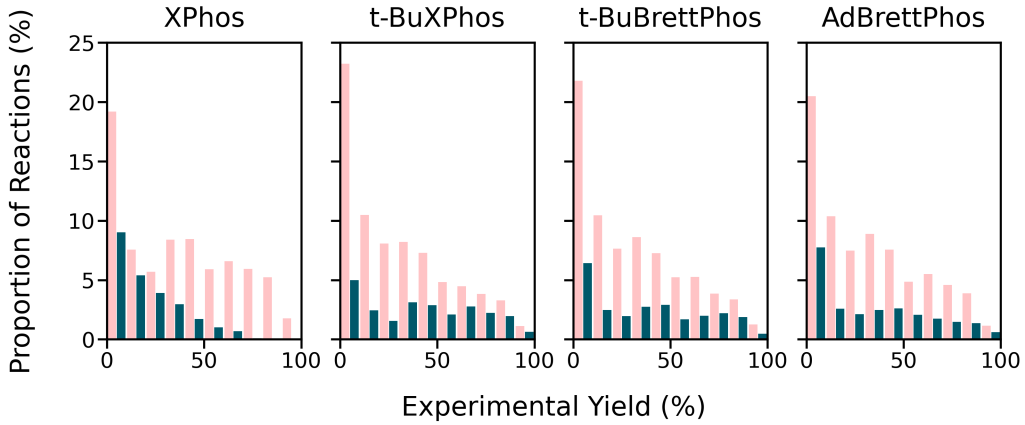



Figure B.13: Distributions of experimental yield of the training data (pale pink bars) and test data (dark blue bars) in the test sets designed without activity ranking. Test Sets were split by (a) high-throughput plates, (b) aryl halide ring type, (c) aryl halide halide type, (d) base and (e) ligand.

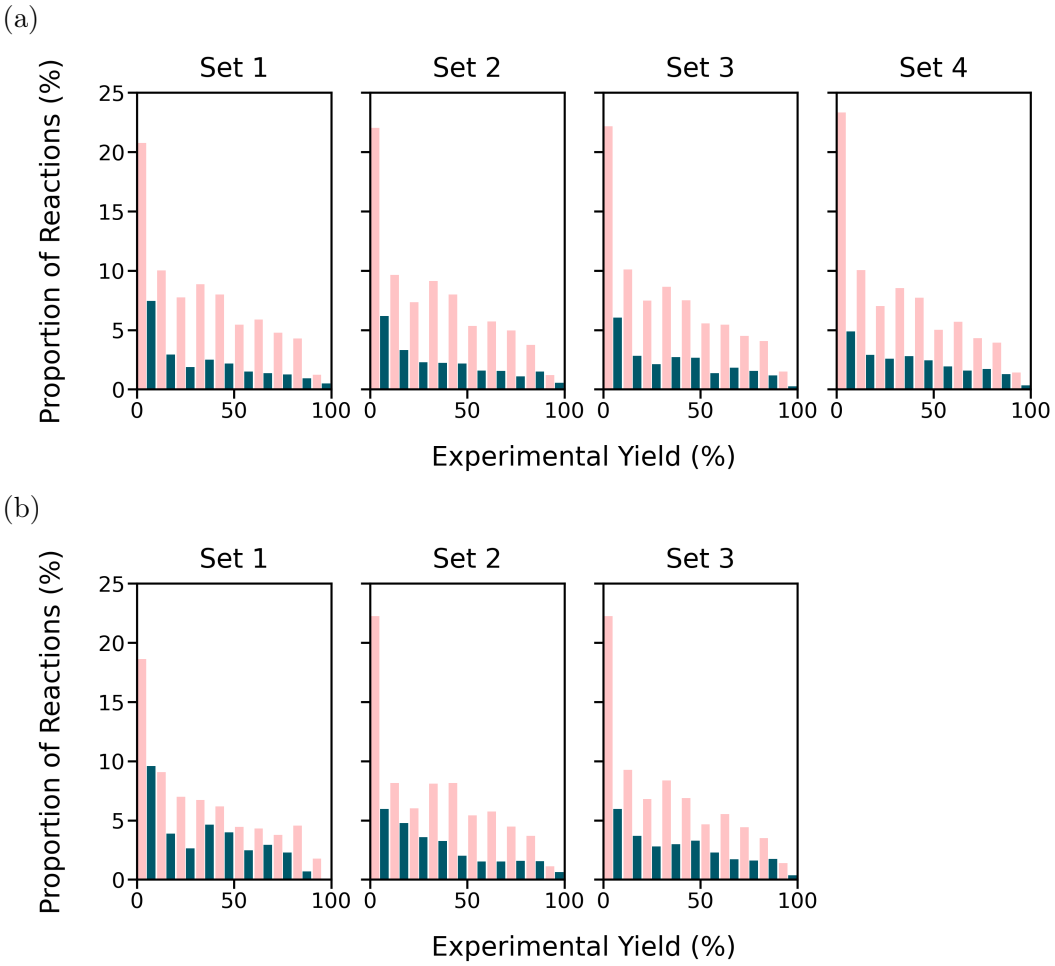



Figure B.14: Distributions of experimental yield of the training data (pale pink bars) and test data (dark blue bars) in the (a) additive and (b) aryl halide, activity ranked tests.

### B.6.2 Domain of Applicability

Table B.10: Maximum Similarity to Training Scores for the Additive and Aryl Halide Ranked Tests

| Key | Name | Maximum Similarity | |
|---|---|---|---|
| | | Raw | Binned |
| Additive | | | |
| 10 | benzo[*c*]isoxazole | 0.31 | 0.30-0.35 |
| 15 | benzo[*d*]isoxazole | 0.31 | 0.30-0.35 |
| 14 | methyl-isoxazole-5-carboxylate | 0.37 | 0.35-0.40 |
| 1 | 4-phenylisoxazole | 0.39 | 0.35-0.40 |
| 5 | 3-methylisoxazole | 0.40 | 0.40-0.45 |
| 3 | 3-phenylisoxazole | 0.40 | 0.40-0.45 |
| 8 | 5-methylisoxazole | 0.41 | 0.40-0.45 |
| 12 | 3,5-dimethylisoxazole | 0.44 | 0.40-0.45 |
| 17 | 3-methyl-5-phenylisoxazole | 0.44 | 0.40-0.45 |
| 20 | 5-methyl-3-(1*H*-pyrrol-1-yl)isoxazole | 0.44 | 0.40-0.45 |
| 18 | *N*,*N*-dibenzylisoxazol-5-amine | 0.49 | 0.45-0.50 |
| 2 | 5-phenylisoxazole | 0.50 | 0.50-0.55 |
| 11 | ethyl-5-methylisoxazole-4-carboxylate | 0.50 | 0.50-0.55 |
| 16 | 5-(2,6-difluorophenyl)isoxazole | 0.50 | 0.50-0.55 |
| 23 | ethyl-3-methoxyisoxazole-5-carboxylate | 0.51 | 0.50-0.55 |
| 4 | ethyl-3-methylisoxazole-5-carboxylate | 0.56 | 0.55-0.60 |
| 6 | ethyl-5-methylisoxazole-3-carboxylate | 0.58 | 0.55-0.60 |
| 9 | ethyl-isoxazole-3-carboxylate | 0.58 | 0.55-0.60 |
| 21 | methyl-5-(furan-2-yl)isoxazole-3-carboxylate | 0.64 | 0.60-0.65 |
| 22 | methyl-5-(thiophen-2-yl)isoxazole-3-carboxylate | 0.64 | 0.60-0.65 |
| Aryl Halide | | | |
| 8 | 1-bromo-4-ethylbenzene | 0.33 | 0.30-0.35 |
| 7 | 1-chloro-4-ethylbenzene | 0.39 | 0.35-0.40 |
| 9 | 1-ethyl-4-iodobenzene | 0.39 | 0.35-0.40 |
| 10 | 2-chloropyridine | 0.48 | 0.45-0.50 |
| 11 | 2-bromopyridine | 0.48 | 0.45-0.50 |
| 15 | 3-iodopyridine | 0.48 | 0.45-0.50 |
| 13 | 3-chloropyridine | 0.55 | 0.50-0.55 |
| 14 | 3-bromopyridine | 0.55 | 0.50-0.55 |
| 1 | 1-chloro-4-(trifluoromethyl)benzene | 0.52 | 0.50-0.55 |
| 2 | 1-bromo-4-(trifluoromethyl)benzene | 0.52 | 0.50-0.55 |
| 3 | 1-iodo-4-(trifluoromethyl)benzene | 0.52 | 0.50-0.55 |
| 6 | 1-iodo-4-methoxybenzene | 0.52 | 0.50-0.55 |
| 5 | 1-bromo-4-methoxybenzene | 0.60 | 0.55-0.60 |

## B.7 Out-of-Sample Tests: Without Activity Ranking

### B.7.1 Additive Test: Plate Split

#### Grid Search Cross-Validated Performance

Table B.11: Grid Search Cross-Validated Performance for the Models in the Additive Test: Plate Split

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **P1** | **P2** | **P3** | Mean | **P1** | **P2** | **P3** | Mean |
| One-hot | Linear Regression | 0.69 | 0.72 | 0.68 | 0.70 | 14.2 | 14.6 | 15.9 | 14.9 |
| | Lasso | 0.69 | 0.73 | 0.69 | 0.70 | 14.2 | 14.5 | 15.7 | 14.8 |
| | Ridge | 0.69 | 0.73 | 0.69 | 0.70 | 14.2 | 14.5 | 15.7 | 14.8 |
| | Bayesian Ridge | 0.69 | 0.73 | 0.69 | 0.70 | 14.2 | 14.5 | 15.7 | 14.8 |
| | SVR - Linear | 0.69 | 0.72 | 0.68 | 0.70 | 14.3 | 14.6 | 15.7 | 14.8 |
| | SVR - Polynomial | 0.89 | 0.90 | 0.91 | 0.90 | 8.3 | 8.7 | 8.4 | 8.5 |
| | SVR - RBF | 0.91 | 0.91 | 0.92 | 0.91 | 7.8 | 8.2 | 8.0 | 8.0 |
| | SVR - Sigmoid | 0.57 | 0.61 | 0.54 | 0.57 | 16.9 | 17.2 | 18.9 | 17.7 |
| | Gradient Boosting | 0.89 | 0.90 | 0.90 | 0.90 | 8.4 | 8.7 | 8.8 | 8.6 |
| | Random Forest | 0.90 | 0.89 | 0.91 | 0.90 | 8.0 | 9.0 | 8.3 | 8.4 |
| Quantum Chemical | Linear Regression | 0.69 | 0.72 | 0.68 | 0.70 | 14.2 | 14.5 | 15.7 | 14.8 |
| | Lasso | 0.69 | 0.72 | 0.68 | 0.70 | 14.3 | 14.6 | 15.7 | 14.9 |
| | Ridge | 0.69 | 0.73 | 0.69 | 0.70 | 14.2 | 14.5 | 15.7 | 14.8 |
| | Bayesian Ridge | 0.69 | 0.72 | 0.68 | 0.70 | 14.3 | 14.6 | 15.7 | 14.9 |
| | SVR - Linear | 0.69 | 0.72 | 0.68 | 0.70 | 14.3 | 14.6 | 15.7 | 14.8 |
| | SVR - Polynomial | 0.90 | 0.91 | 0.89 | 0.90 | 8.1 | 8.2 | 9.1 | 8.5 |
| | SVR - RBF | 0.91 | 0.92 | 0.88 | 0.90 | 7.6 | 8.0 | 9.5 | 8.4 |
| | SVR - Sigmoid | 0.47 | 0.47 | 0.44 | 0.46 | 18.7 | 20.1 | 20.9 | 19.9 |
| | Gradient Boosting | 0.92 | 0.92 | 0.92 | 0.92 | 7.5 | 7.7 | 7.8 | 7.7 |
| | Random Forest | 0.93 | 0.92 | 0.93 | 0.93 | 6.7 | 7.6 | 7.3 | 7.2 |
| Fingerprints: MACCS | Linear Regression | 0.62 | 0.43 | 0.59 | 0.55 | 15.8 | 20.1 | 17.8 | 17.9 |
| | Lasso | 0.66 | 0.67 | 0.62 | 0.65 | 15.0 | 15.9 | 17.2 | 16.0 |
| | Ridge | 0.66 | 0.67 | 0.62 | 0.65 | 15.0 | 15.9 | 17.2 | 16.0 |
| | Bayesian Ridge | 0.66 | 0.67 | 0.62 | 0.65 | 15.0 | 15.9 | 17.2 | 16.0 |
| | SVR - Linear | 0.66 | 0.67 | 0.62 | 0.65 | 15.0 | 15.9 | 17.3 | 16.1 |
| | SVR - Polynomial | 0.88 | 0.87 | 0.86 | 0.87 | 9.0 | 10.1 | 10.5 | 9.9 |
| | SVR - RBF | 0.87 | 0.86 | 0.85 | 0.86 | 9.2 | 10.3 | 10.7 | 10.1 |
| | SVR - Sigmoid | 0.29 | 0.28 | 0.22 | 0.26 | 21.7 | 23.5 | 24.6 | 23.3 |
| | Gradient Boosting | 0.90 | 0.90 | 0.89 | 0.90 | 8.2 | 8.7 | 9.3 | 8.7 |
| | Random Forest | 0.92 | 0.93 | 0.93 | 0.93 | 7.1 | 7.4 | 7.5 | 7.3 |
| Fingerprints: Morgan1 | Linear Regression | 0.67 | 0.68 | 0.50 | 0.62 | 14.8 | 15.6 | 19.0 | 16.5 |
| | Lasso | 0.67 | 0.70 | 0.67 | 0.68 | 14.6 | 15.2 | 16.1 | 15.3 |
| | Ridge | 0.67 | 0.70 | 0.67 | 0.68 | 14.7 | 15.2 | 16.1 | 15.3 |
| | Bayesian Ridge | 0.67 | 0.70 | 0.67 | 0.68 | 14.6 | 15.2 | 16.1 | 15.3 |
| | SVR - Linear | 0.67 | 0.69 | 0.67 | 0.68 | 14.7 | 15.3 | 16.1 | 15.4 |
| | SVR - Polynomial | 0.91 | 0.92 | 0.91 | 0.91 | 7.5 | 8.0 | 8.3 | 7.9 |
| | SVR - RBF | 0.93 | 0.93 | 0.92 | 0.92 | 6.9 | 7.6 | 8.0 | 7.5 |
| | SVR - Sigmoid | 0.47 | 0.48 | 0.41 | 0.45 | 18.8 | 20.0 | 21.4 | 20.1 |
| | Gradient Boosting | 0.91 | 0.92 | 0.91 | 0.91 | 7.6 | 8.0 | 8.2 | 8.0 |
| | Random Forest | 0.93 | 0.93 | 0.93 | 0.93 | 6.8 | 7.5 | 7.1 | 7.1 |
| Fingerprints: RDK | Linear Regression | 0.68 | 0.71 | 0.67 | 0.69 | 14.6 | 15.0 | 15.9 | 15.2 |
| | Lasso | 0.69 | 0.73 | 0.69 | 0.70 | 14.2 | 14.5 | 15.7 | 14.8 |
| | Ridge | 0.69 | 0.73 | 0.69 | 0.70 | 14.2 | 14.5 | 15.7 | 14.8 |
| | Bayesian Ridge | 0.69 | 0.73 | 0.69 | 0.70 | 14.2 | 14.5 | 15.7 | 14.8 |
| | SVR - Linear | 0.69 | 0.72 | 0.68 | 0.70 | 14.3 | 14.6 | 15.7 | 14.8 |
| | SVR - Polynomial | 0.91 | 0.91 | 0.91 | 0.91 | 7.6 | 8.5 | 8.2 | 8.1 |
| | SVR - RBF | 0.92 | 0.91 | 0.92 | 0.91 | 7.3 | 8.5 | 8.1 | 8.0 |

**Table B.11** Grid Search Cross-Validated Performance for the Models in the Additive Test: Plate Split (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **P1** | **P2** | **P3** | Mean | **P1** | **P2** | **P3** | Mean |
| | SVR - Sigmoid | 0.31 | 0.27 | 0.27 | 0.28 | 21.4 | 23.6 | 23.8 | 22.9 |
| | Gradient Boosting | 0.91 | 0.91 | 0.91 | 0.91 | 7.7 | 8.3 | 8.2 | 8.1 |
| | Random Forest | 0.93 | 0.92 | 0.92 | 0.92 | 6.8 | 7.8 | 7.6 | 7.4 |
| Tanimoto: MACCS | Linear Regression | 0.87 | 0.89 | 0.89 | 0.88 | 9.1 | 9.1 | 9.4 | 9.2 |
| | Lasso | 0.91 | 0.92 | 0.91 | 0.91 | 7.7 | 8.0 | 8.5 | 8.1 |
| | Ridge | 0.92 | 0.92 | 0.91 | 0.92 | 7.5 | 7.9 | 8.2 | 7.9 |
| | Bayesian Ridge | 0.91 | 0.90 | 0.89 | 0.90 | 7.8 | 8.6 | 9.2 | 8.5 |
| | SVR - Polynomial | 0.90 | 0.91 | 0.91 | 0.91 | 8.0 | 8.1 | 8.5 | 8.2 |
| | SVR - RBF | 0.89 | 0.91 | 0.90 | 0.90 | 8.4 | 8.5 | 9.0 | 8.6 |
| | SVR - Sigmoid | 0.69 | 0.72 | 0.64 | 0.68 | 14.2 | 14.6 | 16.8 | 15.2 |
| | SVR - Precomputed | 0.91 | 0.91 | 0.89 | 0.90 | 7.8 | 8.4 | 9.1 | 8.4 |
| | Gradient Boosting | 0.79 | 0.80 | 0.74 | 0.78 | 11.9 | 12.3 | 14.2 | 12.8 |
| | Random Forest | 0.69 | 0.76 | 0.64 | 0.70 | 14.3 | 13.5 | 16.8 | 14.9 |
| Tanimoto: Morgan1 | Linear Regression | 0.93 | 0.93 | 0.93 | 0.93 | 6.8 | 7.3 | 7.2 | 7.1 |
| | Lasso | 0.93 | 0.93 | 0.94 | 0.93 | 6.6 | 7.1 | 7.1 | 6.9 |
| | Ridge | 0.93 | 0.94 | 0.94 | 0.94 | 6.6 | 7.0 | 7.0 | 6.9 |
| | Bayesian Ridge | 0.94 | 0.94 | 0.94 | 0.94 | 6.5 | 7.0 | 7.0 | 6.8 |
| | SVR - Polynomial | 0.92 | 0.93 | 0.92 | 0.92 | 7.2 | 7.6 | 7.8 | 7.5 |
| | SVR - RBF | 0.90 | 0.91 | 0.90 | 0.90 | 8.0 | 8.4 | 8.7 | 8.4 |
| | SVR - Sigmoid | 0.89 | 0.89 | 0.88 | 0.88 | 8.5 | 9.3 | 9.8 | 9.2 |
| | SVR - Precomputed | 0.93 | 0.94 | 0.94 | 0.94 | 6.6 | 6.8 | 7.1 | 6.8 |
| | Gradient Boosting | 0.80 | 0.82 | 0.79 | 0.80 | 11.5 | 11.7 | 12.8 | 12.0 |
| | Random Forest | 0.71 | 0.75 | 0.70 | 0.72 | 13.9 | 13.7 | 15.3 | 14.3 |
| Tanimoto: RDK | Linear Regression | 0.91 | 0.90 | 0.91 | 0.90 | 7.9 | 8.9 | 8.3 | 8.4 |
| | Lasso | 0.91 | 0.90 | 0.92 | 0.91 | 7.6 | 8.6 | 7.9 | 8.0 |
| | Ridge | 0.91 | 0.91 | 0.92 | 0.91 | 7.5 | 8.4 | 7.8 | 7.9 |
| | Bayesian Ridge | 0.92 | 0.91 | 0.92 | 0.92 | 7.5 | 8.4 | 7.8 | 7.9 |
| | SVR - Polynomial | 0.90 | 0.89 | 0.90 | 0.90 | 8.1 | 9.1 | 8.7 | 8.6 |
| | SVR - RBF | 0.89 | 0.88 | 0.89 | 0.88 | 8.6 | 9.6 | 9.4 | 9.2 |
| | SVR - Sigmoid | 0.86 | 0.85 | 0.85 | 0.85 | 9.6 | 10.8 | 10.8 | 10.4 |
| | SVR - Precomputed | 0.91 | 0.90 | 0.92 | 0.91 | 7.6 | 8.6 | 8.0 | 8.1 |
| | Gradient Boosting | 0.79 | 0.78 | 0.78 | 0.78 | 11.9 | 12.9 | 13.1 | 12.6 |
| | Random Forest | 0.72 | 0.72 | 0.71 | 0.72 | 13.6 | 14.6 | 14.9 | 14.4 |
| WL | Linear Regression | 0.91 | 0.91 | 0.93 | 0.92 | 7.6 | 8.3 | 7.3 | 7.7 |
| | Lasso | 0.92 | 0.92 | 0.93 | 0.92 | 7.3 | 8.0 | 7.2 | 7.5 |
| | Ridge | 0.92 | 0.92 | 0.94 | 0.93 | 7.2 | 7.8 | 7.1 | 7.4 |
| | Bayesian Ridge | 0.92 | 0.92 | 0.93 | 0.93 | 7.3 | 7.8 | 7.1 | 7.4 |
| | SVR - Polynomial | 0.92 | 0.92 | 0.93 | 0.92 | 7.3 | 8.0 | 7.1 | 7.5 |
| | SVR - RBF | 0.91 | 0.91 | 0.93 | 0.92 | 7.5 | 8.2 | 7.4 | 7.7 |
| | SVR - Sigmoid | 0.84 | 0.82 | 0.83 | 0.83 | 10.3 | 11.7 | 11.6 | 11.2 |
| | SVR - Precomputed | 0.92 | 0.92 | 0.93 | 0.92 | 7.4 | 8.0 | 7.1 | 7.5 |
| | Gradient Boosting | 0.85 | 0.85 | 0.85 | 0.85 | 9.9 | 10.9 | 10.6 | 10.5 |
| | Random Forest | 0.78 | 0.77 | 0.78 | 0.78 | 12.1 | 13.3 | 13.1 | 12.8 |

### Training Set Performance

Table B.12: Training Set Performance for the Models in the Additive Test: Plate Split

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **P1** | **P2** | **P3** | Mean | **P1** | **P2** | **P3** | Mean |
| One-hot | Linear Regression | 0.70 | 0.73 | 0.69 | 0.71 | 14.0 | 14.3 | 15.4 | 14.6 |
| | Lasso | 0.70 | 0.74 | 0.69 | 0.71 | 14.0 | 14.2 | 15.4 | 14.6 |
| | Ridge | 0.70 | 0.74 | 0.69 | 0.71 | 14.0 | 14.2 | 15.4 | 14.6 |
| | Bayesian Ridge | 0.70 | 0.74 | 0.69 | 0.71 | 14.0 | 14.2 | 15.4 | 14.6 |

**Table B.12** Training Set Performance for the Models in the Additive Test: Plate Split (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | P**1** | P**2** | P**3** | Mean | P**1** | P**2** | P**3** | Mean |
| | SVR - Linear | 0.70 | 0.73 | 0.69 | 0.71 | 14.0 | 14.3 | 15.5 | 14.6 |
| | SVR - Polynomial | 0.97 | 0.97 | 0.97 | 0.97 | 4.6 | 4.5 | 4.6 | 4.6 |
| | SVR - RBF | 0.99 | 1.00 | 1.00 | 0.99 | 3.0 | 0.9 | 1.0 | 1.6 |
| | SVR - Sigmoid | 0.58 | 0.63 | 0.56 | 0.59 | 16.7 | 16.8 | 18.5 | 17.3 |
| | Gradient Boosting | 0.95 | 0.96 | 0.96 | 0.96 | 5.6 | 5.6 | 5.6 | 5.6 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.8 | 3.1 | 2.8 | 2.9 |
| Quantum Chemical | Linear Regression | 0.70 | 0.73 | 0.69 | 0.71 | 14.0 | 14.3 | 15.5 | 14.6 |
| | Lasso | 0.70 | 0.73 | 0.69 | 0.71 | 14.1 | 14.3 | 15.5 | 14.6 |
| | Ridge | 0.70 | 0.74 | 0.69 | 0.71 | 14.0 | 14.2 | 15.4 | 14.6 |
| | Bayesian Ridge | 0.70 | 0.73 | 0.69 | 0.71 | 14.1 | 14.4 | 15.5 | 14.7 |
| | SVR - Linear | 0.70 | 0.73 | 0.69 | 0.71 | 14.0 | 14.3 | 15.5 | 14.6 |
| | SVR - Polynomial | 0.96 | 0.97 | 0.96 | 0.96 | 5.1 | 5.1 | 5.7 | 5.3 |
| | SVR - RBF | 0.98 | 0.99 | 0.97 | 0.98 | 3.3 | 3.3 | 4.9 | 3.8 |
| | SVR - Sigmoid | 0.47 | 0.46 | 0.44 | 0.46 | 18.7 | 20.3 | 20.9 | 20.0 |
| | Gradient Boosting | 0.97 | 0.97 | 0.97 | 0.97 | 4.8 | 4.7 | 4.7 | 4.7 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.2 | 2.6 | 2.4 | 2.4 |
| Fingerprints: MACCS | Linear Regression | 0.67 | 0.67 | 0.62 | 0.65 | 14.9 | 15.8 | 17.3 | 16.0 |
| | Lasso | 0.67 | 0.68 | 0.63 | 0.66 | 14.8 | 15.7 | 17.1 | 15.9 |
| | Ridge | 0.67 | 0.68 | 0.63 | 0.66 | 14.8 | 15.7 | 17.1 | 15.9 |
| | Bayesian Ridge | 0.67 | 0.68 | 0.63 | 0.66 | 14.8 | 15.7 | 17.1 | 15.9 |
| | SVR - Linear | 0.67 | 0.68 | 0.63 | 0.66 | 14.9 | 15.7 | 17.1 | 15.9 |
| | SVR - Polynomial | 0.92 | 0.91 | 0.90 | 0.91 | 7.2 | 8.5 | 9.0 | 8.2 |
| | SVR - RBF | 0.92 | 0.90 | 0.89 | 0.91 | 7.2 | 8.6 | 9.1 | 8.3 |
| | SVR - Sigmoid | 0.32 | 0.32 | 0.25 | 0.30 | 21.2 | 22.9 | 24.1 | 22.7 |
| | Gradient Boosting | 0.93 | 0.93 | 0.92 | 0.93 | 6.6 | 7.1 | 7.7 | 7.1 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.3 | 2.5 | 2.4 | 2.4 |
| Fingerprints: Morgan1 | Linear Regression | 0.67 | 0.70 | 0.63 | 0.67 | 14.7 | 15.1 | 17.0 | 15.6 |
| | Lasso | 0.68 | 0.71 | 0.68 | 0.69 | 14.5 | 15.0 | 15.9 | 15.1 |
| | Ridge | 0.68 | 0.71 | 0.68 | 0.69 | 14.5 | 15.0 | 15.9 | 15.1 |
| | Bayesian Ridge | 0.68 | 0.71 | 0.68 | 0.69 | 14.5 | 15.0 | 15.9 | 15.1 |
| | SVR - Linear | 0.68 | 0.70 | 0.67 | 0.69 | 14.5 | 15.1 | 15.9 | 15.2 |
| | SVR - Polynomial | 0.96 | 0.96 | 0.95 | 0.96 | 5.2 | 5.9 | 6.1 | 5.7 |
| | SVR - RBF | 0.98 | 0.97 | 0.97 | 0.97 | 4.0 | 4.9 | 5.2 | 4.7 |
| | SVR - Sigmoid | 0.49 | 0.52 | 0.44 | 0.48 | 18.4 | 19.2 | 20.9 | 19.5 |
| | Gradient Boosting | 0.95 | 0.96 | 0.95 | 0.96 | 5.5 | 5.7 | 6.0 | 5.7 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.3 | 2.5 | 2.4 | 2.4 |
| Fingerprints: RDK | Linear Regression | 0.70 | 0.72 | 0.67 | 0.70 | 14.0 | 14.6 | 15.9 | 14.8 |
| | Lasso | 0.70 | 0.74 | 0.69 | 0.71 | 14.0 | 14.2 | 15.4 | 14.6 |
| | Ridge | 0.70 | 0.74 | 0.69 | 0.71 | 14.0 | 14.2 | 15.4 | 14.6 |
| | Bayesian Ridge | 0.70 | 0.74 | 0.69 | 0.71 | 14.0 | 14.2 | 15.4 | 14.6 |
| | SVR - Linear | 0.70 | 0.73 | 0.69 | 0.71 | 14.0 | 14.3 | 15.5 | 14.6 |
| | SVR - Polynomial | 0.97 | 0.96 | 0.96 | 0.97 | 4.6 | 5.3 | 5.3 | 5.1 |
| | SVR - RBF | 0.98 | 0.97 | 0.97 | 0.97 | 3.8 | 4.9 | 4.9 | 4.5 |
| | SVR - Sigmoid | 0.35 | 0.32 | 0.31 | 0.33 | 20.8 | 22.9 | 23.2 | 22.3 |
| | Gradient Boosting | 0.96 | 0.97 | 0.97 | 0.97 | 4.9 | 4.9 | 5.2 | 5.0 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.3 | 2.6 | 2.5 | 2.4 |
| Tanimoto: MACCS | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 0.99 | 0.99 | 0.99 | 0.99 | 2.3 | 2.3 | 2.6 | 2.4 |
| | Ridge | 0.99 | 0.99 | 1.00 | 0.99 | 2.8 | 3.2 | 1.7 | 2.6 |
| | Bayesian Ridge | 0.98 | 0.97 | 0.97 | 0.97 | 3.9 | 4.8 | 5.2 | 4.6 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 1.3 | 0.9 | 1.0 | 1.1 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.0 | 1.0 |
| | SVR - Sigmoid | 0.73 | 0.75 | 0.67 | 0.72 | 13.4 | 13.8 | 16.1 | 14.4 |
| | SVR - Precomputed | 0.99 | 0.98 | 0.98 | 0.98 | 2.7 | 3.8 | 4.4 | 3.6 |
| | Gradient Boosting | 0.99 | 1.00 | 0.95 | 0.98 | 3.1 | 1.9 | 6.5 | 3.8 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 3.3 | 3.6 | 3.8 | 3.5 |
| Tanimoto: Morgan1 | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 0.9 | 0.9 | 0.9 |
| | Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 0.6 | 0.7 | 0.7 | 0.7 |

**Table B.12** Training Set Performance for the Models in the Additive Test: Plate Split (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | P**1** | P**2** | P**3** | Mean | P**1** | P**2** | P**3** | Mean |
| | Bayesian Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.2 | 0.8 | 1.0 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 1.1 | 1.2 | 1.0 | 1.1 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.0 | 1.0 |
| | SVR - Sigmoid | 0.90 | 0.89 | 0.88 | 0.89 | 8.1 | 9.2 | 9.6 | 9.0 |
| | SVR - Precomputed | 1.00 | 0.99 | 1.00 | 1.00 | 1.8 | 2.2 | 0.9 | 1.6 |
| | Gradient Boosting | 0.99 | 1.00 | 1.00 | 1.00 | 2.4 | 0.5 | 0.6 | 1.2 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 3.4 | 4.1 | 4.4 | 3.9 |
| Tanimoto: RDK | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 1.5 | 1.4 | 1.4 | 1.4 |
| | Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 1.4 | 1.6 | 1.5 | 1.5 |
| | Bayesian Ridge | 0.99 | 0.99 | 1.00 | 0.99 | 2.1 | 2.4 | 1.9 | 2.1 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 0.9 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 1.0 | 1.0 | 1.0 |
| | SVR - Sigmoid | 0.86 | 0.84 | 0.84 | 0.85 | 9.7 | 11.1 | 11.0 | 10.6 |
| | SVR - Precomputed | 1.00 | 1.00 | 1.00 | 1.00 | 1.2 | 1.3 | 1.0 | 1.2 |
| | Gradient Boosting | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 0.7 | 0.6 | 0.8 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 3.2 | 4.1 | 4.1 | 3.8 |
| WL | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 1.7 | 1.6 | 1.5 | 1.6 |
| | Ridge | 0.99 | 0.99 | 1.00 | 1.00 | 1.9 | 2.1 | 1.6 | 1.9 |
| | Bayesian Ridge | 0.98 | 0.98 | 0.99 | 0.99 | 3.3 | 3.4 | 2.0 | 2.9 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 0.9 | 0.9 | 0.9 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 1.0 | 0.9 | 0.9 |
| | SVR - Sigmoid | 0.84 | 0.82 | 0.82 | 0.83 | 10.3 | 11.7 | 11.8 | 11.3 |
| | SVR - Precomputed | 1.00 | 1.00 | 1.00 | 1.00 | 1.5 | 1.7 | 1.1 | 1.5 |
| | Gradient Boosting | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 0.6 | 1.8 | 1.1 |
| | Random Forest | 0.99 | 0.98 | 0.99 | 0.99 | 2.9 | 3.4 | 3.2 | 3.2 |

### Test Set Performance

Table B.13: Test Set Performance for the Models in the Additive Test: Plate Split

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | P**1** | P**2** | P**3** | Mean | P**1** | P**2** | P**3** | Mean |
| One-hot | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | 0.48 | -0.01 | 0.61 | 0.36 | 21.5 | 24.3 | 16.2 | 20.6 |
| | Ridge | 0.51 | 0.02 | 0.60 | 0.38 | 20.8 | 23.9 | 16.3 | 20.3 |
| | Bayesian Ridge | 0.51 | 0.02 | 0.60 | 0.38 | 20.8 | 23.9 | 16.3 | 20.3 |
| | SVR - Linear | 0.51 | -0.01 | 0.60 | 0.37 | 20.8 | 24.2 | 16.3 | 20.5 |
| | SVR - Polynomial | 0.58 | 0.15 | 0.68 | 0.47 | 19.4 | 22.3 | 14.6 | 18.8 |
| | SVR - RBF | 0.65 | -0.13 | 0.64 | 0.39 | 17.7 | 25.6 | 15.4 | 19.6 |
| | SVR - Sigmoid | 0.37 | 0.08 | 0.45 | 0.30 | 23.7 | 23.1 | 19.2 | 22.0 |
| | Gradient Boosting | 0.62 | 0.01 | 0.71 | 0.45 | 18.3 | 24.1 | 14.0 | 18.8 |
| | Random Forest | 0.67 | -0.13 | 0.70 | 0.41 | 17.1 | 25.7 | 14.1 | 18.9 |
| Quantum Chemical | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | 0.46 | 0.14 | -1.73 | -0.38 | 22.0 | 22.3 | 42.7 | 29.0 |
| | Ridge | 0.40 | 0.04 | -1.82 | -0.46 | 23.1 | 23.6 | 43.3 | 30.0 |
| | Bayesian Ridge | 0.47 | 0.02 | -2.06 | -0.52 | 21.7 | 23.9 | 45.2 | 30.2 |
| | SVR - Linear | 0.36 | 0.03 | -1.99 | -0.53 | 23.9 | 23.8 | 44.6 | 30.8 |
| | SVR - Polynomial | 0.49 | -0.24 | -1.52 | -0.42 | 21.3 | 26.9 | 41.0 | 29.7 |
| | SVR - RBF | 0.53 | -0.05 | 0.03 | 0.17 | 20.5 | 24.7 | 25.5 | 23.5 |
| | SVR - Sigmoid | 0.22 | 0.09 | 0.40 | 0.24 | 26.3 | 22.9 | 20.0 | 23.1 |
| | Gradient Boosting | 0.67 | 0.12 | 0.74 | 0.51 | 17.1 | 22.6 | 13.1 | 17.6 |
| | Random Forest | 0.67 | 0.16 | 0.80 | 0.54 | 17.0 | 22.1 | 11.6 | 16.9 |

**Table B.13** Test Set Performance for the Models in the Additive Test: Plate Split (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | P**1** | P**2** | P**3** | Mean | P**1** | P**2** | P**3** | Mean |
| Fingerprints: MACCS | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | 0.23 | 0.02 | 0.41 | 0.22 | 26.1 | 23.8 | 19.8 | 23.2 |
| | Ridge | 0.16 | 0.06 | 0.48 | 0.24 | 27.3 | 23.3 | 18.6 | 23.1 |
| | Bayesian Ridge | 0.21 | 0.06 | 0.49 | 0.25 | 26.6 | 23.4 | 18.4 | 22.8 |
| | SVR - Linear | 0.14 | 0.03 | 0.47 | 0.21 | 27.7 | 23.7 | 18.8 | 23.4 |
| | SVR - Polynomial | 0.35 | 0.03 | 0.40 | 0.26 | 24.0 | 23.7 | 20.0 | 22.6 |
| | SVR - RBF | 0.35 | 0.06 | 0.54 | 0.32 | 24.0 | 23.3 | 17.6 | 21.6 |
| | SVR - Sigmoid | 0.05 | -0.06 | 0.10 | 0.03 | 29.1 | 24.8 | 24.5 | 26.1 |
| | Gradient Boosting | 0.55 | 0.04 | 0.51 | 0.37 | 20.0 | 23.6 | 18.1 | 20.6 |
| | Random Forest | 0.64 | 0.01 | 0.66 | 0.44 | 17.8 | 24.0 | 15.0 | 18.9 |
| Fingerprints: Morgan1 | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | 0.50 | 0.08 | 0.54 | 0.37 | 21.1 | 23.2 | 17.6 | 20.6 |
| | Ridge | 0.52 | 0.05 | 0.47 | 0.35 | 20.7 | 23.4 | 18.7 | 21.0 |
| | Bayesian Ridge | 0.52 | 0.06 | 0.50 | 0.36 | 20.7 | 23.4 | 18.2 | 20.8 |
| | SVR - Linear | 0.51 | 0.03 | 0.55 | 0.36 | 20.9 | 23.8 | 17.3 | 20.6 |
| | SVR - Polynomial | 0.65 | 0.10 | 0.75 | 0.50 | 17.5 | 22.8 | 12.8 | 17.7 |
| | SVR - RBF | 0.65 | 0.11 | 0.77 | 0.51 | 17.6 | 22.8 | 12.3 | 17.6 |
| | SVR - Sigmoid | 0.25 | 0.15 | 0.36 | 0.25 | 25.8 | 22.3 | 20.6 | 22.9 |
| | Gradient Boosting | 0.60 | 0.12 | 0.72 | 0.48 | 18.9 | 22.6 | 13.6 | 18.4 |
| | Random Forest | 0.65 | 0.20 | 0.80 | 0.55 | 17.6 | 21.6 | 11.5 | 16.9 |
| Fingerprints: RDK | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | 0.52 | -0.30 | 0.42 | 0.21 | 20.7 | 27.5 | 19.7 | 22.6 |
| | Ridge | 0.55 | -0.13 | 0.59 | 0.34 | 20.0 | 25.6 | 16.4 | 20.7 |
| | Bayesian Ridge | 0.55 | -0.13 | 0.59 | 0.34 | 20.0 | 25.6 | 16.4 | 20.7 |
| | SVR - Linear | 0.55 | -0.20 | 0.60 | 0.32 | 19.9 | 26.4 | 16.4 | 20.9 |
| | SVR - Polynomial | 0.66 | -0.23 | 0.62 | 0.35 | 17.5 | 26.7 | 15.9 | 20.0 |
| | SVR - RBF | 0.65 | -0.18 | 0.64 | 0.37 | 17.6 | 26.1 | 15.5 | 19.7 |
| | SVR - Sigmoid | 0.09 | -0.05 | 0.10 | 0.05 | 28.4 | 24.7 | 24.4 | 25.8 |
| | Gradient Boosting | 0.66 | -0.38 | 0.42 | 0.24 | 17.4 | 28.3 | 19.6 | 21.8 |
| | Random Forest | 0.71 | -0.35 | 0.52 | 0.29 | 16.2 | 28.0 | 17.8 | 20.7 |
| Tanimoto: MACCS | Linear Regression | 0.39 | -0.21 | 0.55 | 0.24 | 23.3 | 26.5 | 17.4 | 22.4 |
| | Lasso | 0.41 | -0.18 | 0.57 | 0.27 | 23.0 | 26.2 | 16.8 | 22.0 |
| | Ridge | 0.41 | -0.11 | 0.56 | 0.29 | 22.8 | 25.5 | 17.2 | 21.8 |
| | Bayesian Ridge | 0.43 | -0.07 | 0.61 | 0.32 | 22.5 | 25.0 | 16.0 | 21.2 |
| | SVR - Polynomial | 0.46 | -0.09 | 0.60 | 0.33 | 21.9 | 25.2 | 16.2 | 21.1 |
| | SVR - RBF | 0.49 | -0.05 | 0.61 | 0.35 | 21.4 | 24.7 | 16.1 | 20.7 |
| | SVR - Sigmoid | 0.44 | -0.01 | 0.56 | 0.33 | 22.2 | 24.2 | 17.2 | 21.2 |
| | SVR - Precomputed | 0.41 | -0.08 | 0.61 | 0.31 | 23.0 | 25.1 | 16.1 | 21.4 |
| | Gradient Boosting | 0.53 | -0.20 | 0.51 | 0.28 | 20.5 | 26.4 | 18.1 | 21.7 |
| | Random Forest | 0.55 | -0.19 | 0.43 | 0.26 | 20.0 | 26.3 | 19.4 | 21.9 |
| Tanimoto: Morgan1 | Linear Regression | 0.64 | 0.15 | 0.78 | 0.52 | 17.9 | 22.2 | 12.0 | 17.4 |
| | Lasso | 0.64 | 0.15 | 0.79 | 0.52 | 17.9 | 22.2 | 12.0 | 17.4 |
| | Ridge | 0.64 | 0.15 | 0.79 | 0.53 | 17.9 | 22.2 | 12.0 | 17.4 |
| | Bayesian Ridge | 0.64 | 0.15 | 0.79 | 0.53 | 17.9 | 22.2 | 12.0 | 17.3 |
| | SVR - Polynomial | 0.64 | 0.19 | 0.76 | 0.53 | 18.0 | 21.7 | 12.5 | 17.4 |
| | SVR - RBF | 0.63 | 0.20 | 0.75 | 0.52 | 18.3 | 21.6 | 13.0 | 17.6 |
| | SVR - Sigmoid | 0.61 | 0.14 | 0.76 | 0.50 | 18.5 | 22.4 | 12.6 | 17.8 |
| | SVR - Precomputed | 0.64 | 0.15 | 0.79 | 0.52 | 17.8 | 22.3 | 12.0 | 17.4 |
| | Gradient Boosting | 0.69 | 0.04 | 0.72 | 0.48 | 16.5 | 23.6 | 13.8 | 18.0 |
| | Random Forest | 0.64 | -0.05 | 0.70 | 0.43 | 18.0 | 24.7 | 14.1 | 18.9 |
| Tanimoto: RDK | Linear Regression | 0.61 | 0.09 | 0.69 | 0.46 | 18.5 | 23.0 | 14.5 | 18.7 |
| | Lasso | 0.61 | 0.05 | 0.68 | 0.45 | 18.5 | 23.5 | 14.5 | 18.8 |
| | Ridge | 0.62 | -0.02 | 0.69 | 0.43 | 18.5 | 24.4 | 14.4 | 19.1 |
| | Bayesian Ridge | 0.62 | -0.02 | 0.69 | 0.43 | 18.5 | 24.3 | 14.4 | 19.1 |
| | SVR - Polynomial | 0.59 | 0.00 | 0.67 | 0.42 | 19.0 | 24.2 | 14.7 | 19.3 |
| | SVR - RBF | 0.58 | 0.00 | 0.66 | 0.41 | 19.4 | 24.1 | 15.1 | 19.5 |
| | SVR - Sigmoid | 0.60 | -0.05 | 0.64 | 0.40 | 18.7 | 24.7 | 15.5 | 19.6 |
| | SVR - Precomputed | 0.62 | -0.02 | 0.69 | 0.43 | 18.5 | 24.4 | 14.4 | 19.1 |
| | Gradient Boosting | 0.57 | 0.06 | 0.60 | 0.41 | 19.6 | 23.4 | 16.3 | 19.7 |

**Table B.13** Test Set Performance for the Models in the Additive Test: Plate Split (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | P**1** | P**2** | P**3** | Mean | P**1** | P**2** | P**3** | Mean |
| | Random Forest | 0.66 | -0.12 | 0.33 | 0.29 | 17.4 | 25.5 | 21.2 | 21.4 |
| WL | Linear Regression | 0.65 | 0.14 | 0.73 | 0.51 | 17.7 | 22.4 | 13.3 | 17.8 |
| | Lasso | 0.65 | 0.12 | 0.73 | 0.50 | 17.7 | 22.6 | 13.4 | 17.9 |
| | Ridge | 0.64 | 0.13 | 0.73 | 0.50 | 17.8 | 22.5 | 13.5 | 17.9 |
| | Bayesian Ridge | 0.64 | 0.13 | 0.73 | 0.50 | 17.8 | 22.4 | 13.5 | 17.9 |
| | SVR - Polynomial | 0.63 | 0.15 | 0.72 | 0.50 | 18.2 | 22.3 | 13.7 | 18.0 |
| | SVR - RBF | 0.62 | 0.15 | 0.71 | 0.49 | 18.5 | 22.2 | 13.8 | 18.2 |
| | SVR - Sigmoid | 0.61 | 0.14 | 0.70 | 0.48 | 18.6 | 22.3 | 14.2 | 18.4 |
| | SVR - Precomputed | 0.64 | 0.13 | 0.73 | 0.50 | 17.8 | 22.5 | 13.5 | 17.9 |
| | Gradient Boosting | 0.65 | 0.02 | 0.68 | 0.45 | 17.6 | 23.9 | 14.6 | 18.7 |
| | Random Forest | 0.68 | -0.14 | 0.55 | 0.36 | 16.9 | 25.7 | 17.3 | 20.0 |

## B.7.2 Aryl Halide Test: Ring Split

### Grid Search Cross-Validated Performance

Table B.14: Grid Search Cross-Validated Performance for the Models in the Aryl Halide Test: Ring Split

| Descriptor | ML Algorithm | $R^2$ | | | RMSE (%) | | |
|---|---|---|---|---|---|---|---|
| | | Phenyl | Pyridyl | Mean | Phenyl | Pyridyl | Mean |
| One-hot | Linear Regression | 0.69 | 0.71 | 0.70 | 16.6 | 12.3 | 14.4 |
| | Lasso | 0.69 | 0.72 | 0.70 | 16.6 | 12.2 | 14.4 |
| | Ridge | 0.69 | 0.72 | 0.70 | 16.6 | 12.2 | 14.4 |
| | Bayesian Ridge | 0.69 | 0.72 | 0.70 | 16.6 | 12.2 | 14.4 |
| | SVR - Linear | 0.68 | 0.72 | 0.70 | 16.6 | 12.2 | 14.4 |
| | SVR - Polynomial | 0.86 | 0.92 | 0.89 | 10.9 | 6.4 | 8.7 |
| | SVR - RBF | 0.88 | 0.93 | 0.91 | 10.3 | 5.9 | 8.1 |
| | SVR - Sigmoid | 0.57 | 0.60 | 0.59 | 19.4 | 14.5 | 16.9 |
| | Gradient Boosting | 0.85 | 0.92 | 0.89 | 11.3 | 6.6 | 8.9 |
| | Random Forest | 0.86 | 0.90 | 0.88 | 11.0 | 7.3 | 9.2 |
| Quantum Chemical | Linear Regression | 0.68 | 0.72 | 0.70 | 16.7 | 12.2 | 14.4 |
| | Lasso | 0.67 | 0.71 | 0.69 | 16.9 | 12.2 | 14.6 |
| | Ridge | 0.68 | 0.72 | 0.70 | 16.7 | 12.2 | 14.4 |
| | Bayesian Ridge | 0.67 | 0.71 | 0.69 | 17.0 | 12.3 | 14.6 |
| | SVR - Linear | 0.68 | 0.72 | 0.70 | 16.7 | 12.2 | 14.5 |
| | SVR - Polynomial | 0.85 | 0.91 | 0.88 | 11.3 | 6.8 | 9.0 |
| | SVR - RBF | 0.88 | 0.91 | 0.89 | 10.3 | 7.0 | 8.6 |
| | SVR - Sigmoid | 0.49 | 0.49 | 0.49 | 21.0 | 16.3 | 18.7 |
| | Gradient Boosting | 0.88 | 0.94 | 0.91 | 10.2 | 5.8 | 8.0 |
| | Random Forest | 0.91 | 0.94 | 0.93 | 8.7 | 5.6 | 7.2 |
| Fingerprints: MACCS | Linear Regression | 0.57 | 0.63 | 0.60 | 19.5 | 14.0 | 16.7 |
| | Lasso | 0.63 | 0.64 | 0.63 | 18.1 | 13.8 | 15.9 |
| | Ridge | 0.63 | 0.64 | 0.63 | 18.1 | 13.8 | 15.9 |
| | Bayesian Ridge | 0.63 | 0.64 | 0.63 | 18.1 | 13.8 | 15.9 |
| | SVR - Linear | 0.62 | 0.64 | 0.63 | 18.2 | 13.8 | 16.0 |
| | SVR - Polynomial | 0.81 | 0.89 | 0.85 | 13.0 | 7.8 | 10.4 |
| | SVR - RBF | 0.79 | 0.88 | 0.84 | 13.4 | 7.9 | 10.6 |
| | SVR - Sigmoid | 0.34 | 0.19 | 0.27 | 24.0 | 20.6 | 22.3 |
| | Gradient Boosting | 0.85 | 0.91 | 0.88 | 11.5 | 6.8 | 9.2 |
| | Random Forest | 0.89 | 0.93 | 0.91 | 9.7 | 5.9 | 7.8 |
| Fingerprints: Morgan1 | Linear Regression | 0.68 | 0.64 | 0.66 | 16.7 | 13.7 | 15.2 |
| | Lasso | 0.68 | 0.67 | 0.68 | 16.6 | 13.2 | 14.9 |
| | Ridge | 0.69 | 0.67 | 0.68 | 16.6 | 13.2 | 14.9 |
| | Bayesian Ridge | 0.68 | 0.67 | 0.68 | 16.6 | 13.2 | 14.9 |
| | SVR - Linear | 0.68 | 0.67 | 0.67 | 16.6 | 13.2 | 14.9 |
| | SVR - Polynomial | 0.87 | 0.92 | 0.90 | 10.7 | 6.3 | 8.5 |
| | SVR - RBF | 0.88 | 0.93 | 0.91 | 10.2 | 5.9 | 8.1 |
| | SVR - Sigmoid | 0.39 | 0.42 | 0.41 | 23.0 | 17.5 | 20.3 |
| | Gradient Boosting | 0.87 | 0.93 | 0.90 | 10.5 | 6.3 | 8.4 |
| | Random Forest | 0.90 | 0.94 | 0.92 | 9.1 | 5.7 | 7.4 |
| Fingerprints: RDK | Linear Regression | 0.65 | 0.70 | 0.68 | 17.5 | 12.5 | 15.0 |
| | Lasso | 0.69 | 0.72 | 0.70 | 16.6 | 12.2 | 14.4 |
| | Ridge | 0.68 | 0.72 | 0.70 | 16.6 | 12.2 | 14.4 |
| | Bayesian Ridge | 0.69 | 0.72 | 0.70 | 16.6 | 12.2 | 14.4 |
| | SVR - Linear | 0.68 | 0.72 | 0.70 | 16.6 | 12.2 | 14.4 |
| | SVR - Polynomial | 0.86 | 0.94 | 0.90 | 11.0 | 5.8 | 8.4 |
| | SVR - RBF | 0.86 | 0.94 | 0.90 | 11.0 | 5.7 | 8.3 |
| | SVR - Sigmoid | 0.23 | 0.29 | 0.26 | 26.0 | 19.3 | 22.7 |
| | Gradient Boosting | 0.86 | 0.93 | 0.90 | 11.0 | 6.0 | 8.5 |
| | Random Forest | 0.90 | 0.93 | 0.92 | 9.4 | 5.8 | 7.6 |
| Tanimoto: MACCS | Linear Regression | 0.84 | 0.90 | 0.87 | 11.9 | 7.1 | 9.5 |
| | Lasso | 0.87 | 0.92 | 0.90 | 10.5 | 6.6 | 8.6 |
| | Ridge | 0.88 | 0.92 | 0.90 | 10.2 | 6.5 | 8.3 |

**Table B.14** Grid Search Cross-Validated Performance for the Models in the Aryl Halide Test: Ring Split (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | RMSE (%) | | |
|---|---|---|---|---|---|---|---|
| | | Phenyl | Pyridyl | Mean | Phenyl | Pyridyl | Mean |
| | Bayesian Ridge | 0.86 | 0.91 | 0.89 | 11.2 | 6.7 | 8.9 |
| | SVR - Polynomial | 0.88 | 0.91 | 0.89 | 10.2 | 7.0 | 8.6 |
| | SVR - RBF | 0.87 | 0.89 | 0.88 | 10.7 | 7.5 | 9.1 |
| | SVR - Sigmoid | 0.57 | 0.66 | 0.61 | 19.5 | 13.3 | 16.4 |
| | SVR - Precomputed | 0.86 | 0.92 | 0.89 | 11.2 | 6.6 | 8.9 |
| | Gradient Boosting | 0.75 | 0.80 | 0.77 | 14.9 | 10.1 | 12.5 |
| | Random Forest | 0.61 | 0.74 | 0.68 | 18.6 | 11.6 | 15.1 |
| Tanimoto: Morgan1 | Linear Regression | 0.91 | 0.93 | 0.92 | 9.0 | 5.9 | 7.4 |
| | Lasso | 0.91 | 0.94 | 0.92 | 8.9 | 5.7 | 7.3 |
| | Ridge | 0.91 | 0.94 | 0.93 | 8.8 | 5.7 | 7.2 |
| | Bayesian Ridge | 0.91 | 0.94 | 0.93 | 8.7 | 5.7 | 7.2 |
| | SVR - Polynomial | 0.90 | 0.92 | 0.91 | 9.3 | 6.5 | 7.9 |
| | SVR - RBF | 0.88 | 0.90 | 0.89 | 10.1 | 7.2 | 8.7 |
| | SVR - Sigmoid | 0.83 | 0.89 | 0.86 | 12.2 | 7.7 | 9.9 |
| | SVR - Precomputed | 0.91 | 0.94 | 0.92 | 8.8 | 5.7 | 7.2 |
| | Gradient Boosting | 0.78 | 0.79 | 0.78 | 14.0 | 10.5 | 12.2 |
| | Random Forest | 0.67 | 0.73 | 0.70 | 17.0 | 11.9 | 14.5 |
| Tanimoto: RDK | Linear Regression | 0.87 | 0.92 | 0.90 | 10.5 | 6.3 | 8.4 |
| | Lasso | 0.88 | 0.93 | 0.91 | 10.1 | 6.0 | 8.1 |
| | Ridge | 0.89 | 0.93 | 0.91 | 9.9 | 5.9 | 7.9 |
| | Bayesian Ridge | 0.89 | 0.93 | 0.91 | 9.9 | 5.9 | 7.9 |
| | SVR - Polynomial | 0.88 | 0.92 | 0.90 | 10.3 | 6.6 | 8.5 |
| | SVR - RBF | 0.86 | 0.90 | 0.88 | 10.9 | 7.2 | 9.0 |
| | SVR - Sigmoid | 0.81 | 0.85 | 0.83 | 13.0 | 8.9 | 11.0 |
| | SVR - Precomputed | 0.88 | 0.93 | 0.91 | 10.1 | 6.0 | 8.1 |
| | Gradient Boosting | 0.77 | 0.82 | 0.79 | 14.3 | 9.8 | 12.1 |
| | Random Forest | 0.63 | 0.75 | 0.69 | 18.0 | 11.5 | 14.8 |
| WL | Linear Regression | 0.88 | 0.93 | 0.91 | 10.3 | 5.9 | 8.1 |
| | Lasso | 0.89 | 0.94 | 0.91 | 9.9 | 5.6 | 7.7 |
| | Ridge | 0.89 | 0.94 | 0.92 | 9.7 | 5.5 | 7.6 |
| | Bayesian Ridge | 0.89 | 0.94 | 0.92 | 9.6 | 5.6 | 7.6 |
| | SVR - Polynomial | 0.89 | 0.94 | 0.91 | 9.7 | 5.8 | 7.7 |
| | SVR - RBF | 0.89 | 0.93 | 0.91 | 9.8 | 6.1 | 8.0 |
| | SVR - Sigmoid | 0.79 | 0.83 | 0.81 | 13.5 | 9.5 | 11.5 |
| | SVR - Precomputed | 0.89 | 0.94 | 0.91 | 9.8 | 5.6 | 7.7 |
| | Gradient Boosting | 0.83 | 0.87 | 0.85 | 12.3 | 8.3 | 10.3 |
| | Random Forest | 0.75 | 0.79 | 0.77 | 14.6 | 10.6 | 12.6 |

### Training Set Performance

Table B.15: Training Set Performance for the Models in the Aryl Halide Test: Ring Split

| Descriptor | ML Algorithm | $R^2$ | | | RMSE (%) | | |
|---|---|---|---|---|---|---|---|
| | | Phenyl | Pyridyl | Mean | Phenyl | Pyridyl | Mean |
| One-hot | Linear Regression | 0.70 | 0.73 | 0.71 | 16.3 | 12.0 | 14.1 |
| | Lasso | 0.70 | 0.73 | 0.71 | 16.3 | 12.0 | 14.1 |
| | Ridge | 0.70 | 0.73 | 0.71 | 16.3 | 12.0 | 14.1 |
| | Bayesian Ridge | 0.70 | 0.73 | 0.71 | 16.3 | 12.0 | 14.1 |
| | SVR - Linear | 0.70 | 0.72 | 0.71 | 16.4 | 12.0 | 14.2 |
| | SVR - Polynomial | 0.97 | 0.98 | 0.97 | 5.4 | 3.0 | 4.2 |
| | SVR - RBF | 0.98 | 1.00 | 0.99 | 4.2 | 0.9 | 2.6 |
| | SVR - Sigmoid | 0.59 | 0.61 | 0.60 | 19.1 | 14.3 | 16.7 |
| | Gradient Boosting | 0.93 | 0.97 | 0.95 | 7.7 | 4.3 | 6.0 |
| | Random Forest | 0.98 | 0.99 | 0.99 | 3.8 | 2.6 | 3.2 |

**Table B.15** Training Set Performance for the Models in the Aryl Halide Test: Ring Split (continued)

| Descriptor | ML Algorithm | $R^2$ | | | RMSE (%) | | |
|---|---|---|---|---|---|---|---|
| | | Phenyl | Pyridyl | Mean | Phenyl | Pyridyl | Mean |
| Quantum Chemical | Linear Regression | 0.70 | 0.73 | 0.71 | 16.4 | 12.0 | 14.2 |
| | Lasso | 0.69 | 0.72 | 0.71 | 16.6 | 12.1 | 14.3 |
| | Ridge | 0.70 | 0.73 | 0.71 | 16.3 | 12.0 | 14.2 |
| | Bayesian Ridge | 0.69 | 0.72 | 0.71 | 16.5 | 12.1 | 14.3 |
| | SVR - Linear | 0.69 | 0.72 | 0.71 | 16.5 | 12.1 | 14.3 |
| | SVR - Polynomial | 0.95 | 0.97 | 0.96 | 6.8 | 4.0 | 5.4 |
| | SVR - RBF | 0.98 | 0.99 | 0.98 | 3.9 | 2.6 | 3.3 |
| | SVR - Sigmoid | 0.51 | 0.49 | 0.50 | 20.8 | 16.4 | 18.6 |
| | Gradient Boosting | 0.97 | 0.98 | 0.97 | 5.4 | 3.4 | 4.4 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 3.0 | 1.9 | 2.4 |
| Fingerprints: MACCS | Linear Regression | 0.53 | 0.64 | 0.59 | 20.3 | 13.7 | 17.0 |
| | Lasso | 0.64 | 0.65 | 0.64 | 17.8 | 13.6 | 15.7 |
| | Ridge | 0.64 | 0.65 | 0.65 | 17.8 | 13.6 | 15.7 |
| | Bayesian Ridge | 0.64 | 0.65 | 0.64 | 17.8 | 13.6 | 15.7 |
| | SVR - Linear | 0.64 | 0.65 | 0.64 | 17.9 | 13.6 | 15.8 |
| | SVR - Polynomial | 0.86 | 0.93 | 0.90 | 11.0 | 6.2 | 8.6 |
| | SVR - RBF | 0.86 | 0.93 | 0.89 | 11.3 | 6.0 | 8.6 |
| | SVR - Sigmoid | 0.24 | 0.23 | 0.24 | 25.9 | 20.1 | 23.0 |
| | Gradient Boosting | 0.92 | 0.95 | 0.93 | 8.4 | 5.4 | 6.9 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 3.2 | 2.0 | 2.6 |
| Fingerprints: Morgan1 | Linear Regression | 0.69 | 0.67 | 0.68 | 16.6 | 13.2 | 14.9 |
| | Lasso | 0.70 | 0.68 | 0.69 | 16.3 | 13.1 | 14.7 |
| | Ridge | 0.70 | 0.68 | 0.69 | 16.3 | 13.1 | 14.7 |
| | Bayesian Ridge | 0.70 | 0.68 | 0.69 | 16.3 | 13.1 | 14.7 |
| | SVR - Linear | 0.70 | 0.67 | 0.69 | 16.4 | 13.1 | 14.7 |
| | SVR - Polynomial | 0.94 | 0.97 | 0.96 | 7.0 | 4.3 | 5.6 |
| | SVR - RBF | 0.96 | 0.98 | 0.97 | 6.3 | 3.3 | 4.8 |
| | SVR - Sigmoid | 0.44 | 0.45 | 0.44 | 22.3 | 17.0 | 19.6 |
| | Gradient Boosting | 0.95 | 0.96 | 0.95 | 7.0 | 4.6 | 5.8 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 3.0 | 1.9 | 2.5 |
| Fingerprints: RDK | Linear Regression | 0.67 | 0.72 | 0.69 | 17.1 | 12.2 | 14.7 |
| | Lasso | 0.70 | 0.73 | 0.71 | 16.3 | 12.0 | 14.1 |
| | Ridge | 0.70 | 0.73 | 0.71 | 16.3 | 12.0 | 14.1 |
| | Bayesian Ridge | 0.70 | 0.73 | 0.71 | 16.3 | 12.0 | 14.1 |
| | SVR - Linear | 0.70 | 0.72 | 0.71 | 16.4 | 12.0 | 14.2 |
| | SVR - Polynomial | 0.95 | 0.98 | 0.96 | 6.8 | 3.3 | 5.1 |
| | SVR - RBF | 0.95 | 0.99 | 0.97 | 6.6 | 2.8 | 4.7 |
| | SVR - Sigmoid | 0.27 | 0.34 | 0.30 | 25.3 | 18.7 | 22.0 |
| | Gradient Boosting | 0.96 | 0.97 | 0.97 | 6.2 | 3.7 | 4.9 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 3.2 | 2.0 | 2.6 |
| Tanimoto: MACCS | Linear Regression | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 |
| | Lasso | 0.99 | 0.99 | 0.99 | 2.6 | 1.9 | 2.3 |
| | Ridge | 0.99 | 0.99 | 0.99 | 2.3 | 2.4 | 2.4 |
| | Bayesian Ridge | 0.95 | 0.98 | 0.97 | 6.3 | 3.1 | 4.7 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.0 | 0.9 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.1 | 1.0 | 1.0 |
| | SVR - Sigmoid | 0.60 | 0.70 | 0.65 | 18.8 | 12.6 | 15.7 |
| | SVR - Precomputed | 0.96 | 0.99 | 0.98 | 5.9 | 2.1 | 4.0 |
| | Gradient Boosting | 1.00 | 1.00 | 1.00 | 1.7 | 1.4 | 1.6 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 4.5 | 2.9 | 3.7 |
| Tanimoto: Morgan1 | Linear Regression | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 0.6 | 0.8 | 0.7 |
| | Ridge | 1.00 | 1.00 | 1.00 | 0.9 | 0.6 | 0.7 |
| | Bayesian Ridge | 1.00 | 1.00 | 1.00 | 1.4 | 0.8 | 1.1 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.4 | 0.9 | 1.2 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.0 | 0.9 | 1.0 |
| | SVR - Sigmoid | 0.86 | 0.86 | 0.86 | 11.2 | 8.6 | 9.9 |
| | SVR - Precomputed | 0.99 | 1.00 | 0.99 | 2.8 | 0.9 | 1.8 |
| | Gradient Boosting | 1.00 | 1.00 | 1.00 | 1.5 | 0.1 | 0.8 |

**Table B.15** Training Set Performance for the Models in the Aryl Halide Test: Ring Split (continued)

| Descriptor | ML Algorithm | $R^2$ | | | RMSE (%) | | |
|---|---|---|---|---|---|---|---|
| | | Phenyl | Pyridyl | Mean | Phenyl | Pyridyl | Mean |
| | Random Forest | 0.97 | 0.98 | 0.97 | 5.0 | 3.4 | 4.2 |
| Tanimoto: RDK | Linear Regression | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 |
| | Lasso | 0.98 | 1.00 | 0.99 | 4.5 | 1.3 | 2.9 |
| | Ridge | 1.00 | 1.00 | 1.00 | 1.9 | 1.3 | 1.6 |
| | Bayesian Ridge | 0.99 | 0.99 | 0.99 | 3.2 | 1.9 | 2.6 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.0 | 0.9 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.0 | 0.9 | 1.0 |
| | SVR - Sigmoid | 0.84 | 0.87 | 0.85 | 12.0 | 8.4 | 10.2 |
| | SVR - Precomputed | 1.00 | 1.00 | 1.00 | 1.4 | 1.0 | 1.2 |
| | Gradient Boosting | 1.00 | 1.00 | 1.00 | 0.7 | 1.5 | 1.1 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 4.5 | 3.1 | 3.8 |
| WL | Linear Regression | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 |
| | Lasso | 0.97 | 1.00 | 0.98 | 4.8 | 1.5 | 3.2 |
| | Ridge | 0.99 | 1.00 | 0.99 | 2.4 | 1.6 | 2.0 |
| | Bayesian Ridge | 0.98 | 0.99 | 0.98 | 4.5 | 2.2 | 3.3 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.0 | 0.9 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.0 | 0.9 | 1.0 |
| | SVR - Sigmoid | 0.79 | 0.83 | 0.81 | 13.6 | 9.5 | 11.5 |
| | SVR - Precomputed | 0.98 | 1.00 | 0.99 | 4.3 | 1.2 | 2.8 |
| | Gradient Boosting | 1.00 | 1.00 | 1.00 | 0.8 | 1.4 | 1.1 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 4.1 | 2.9 | 3.5 |

### Test Set Performance

Table B.16: Test Set Performance for the Models in the Aryl Halide Test: Ring Split

| Descriptor | ML Algorithm | $R^2$ | | | RMSE (%) | | |
|---|---|---|---|---|---|---|---|
| | | Phenyl | Pyridyl | Mean | Phenyl | Pyridyl | Mean |
| One-hot | Linear Regression | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 |
| | Lasso | -0.56 | 0.02 | -0.27 | 28.7 | 29.4 | 29.0 |
| | Ridge | -0.58 | -0.02 | -0.30 | 28.8 | 30.0 | 29.4 |
| | Bayesian Ridge | -0.57 | -0.02 | -0.30 | 28.8 | 30.0 | 29.4 |
| | SVR - Linear | -0.60 | -0.03 | -0.31 | 29.0 | 30.1 | 29.6 |
| | SVR - Polynomial | -0.42 | -0.01 | -0.21 | 27.4 | 29.8 | 28.6 |
| | SVR - RBF | -0.77 | 0.08 | -0.35 | 30.6 | 28.4 | 29.5 |
| | SVR - Sigmoid | -0.64 | -0.13 | -0.39 | 29.4 | 31.5 | 30.5 |
| | Gradient Boosting | -0.98 | 0.07 | -0.45 | 32.3 | 28.7 | 30.5 |
| | Random Forest | <-1.00 | 0.22 | -0.52 | 34.5 | 26.3 | 30.4 |
| Quantum Chemical | Linear Regression | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 |
| | Lasso | <-1.00 | <-1.00 | <-1.00 | >100.0 | 67.4 | >100.0 |
| | Ridge | <-1.00 | 0.15 | <-1.00 | >100.0 | 27.3 | 90.5 |
| | Bayesian Ridge | <-1.00 | 0.12 | <-1.00 | >100.0 | 27.8 | 90.5 |
| | SVR - Linear | <-1.00 | 0.14 | <-1.00 | >100.0 | 27.5 | 95.0 |
| | SVR - Polynomial | <-1.00 | <-1.00 | <-1.00 | >100.0 | 65.6 | >100.0 |
| | SVR - RBF | -0.10 | -1.26 | -0.68 | 24.1 | 44.6 | 34.4 |
| | SVR - Sigmoid | <-1.00 | <-1.00 | <-1.00 | 47.8 | 58.3 | 53.1 |
| | Gradient Boosting | -0.67 | -0.13 | -0.40 | 29.6 | 31.5 | 30.6 |
| | Random Forest | -0.32 | -0.40 | -0.36 | 26.3 | 35.2 | 30.8 |
| Fingerprints: MACCS | Linear Regression | -0.02 | <-1.00 | <-1.00 | 23.2 | >100.0 | >100.0 |
| | Lasso | 0.18 | 0.16 | 0.17 | 20.8 | 27.2 | 24.0 |
| | Ridge | 0.17 | 0.15 | 0.16 | 20.8 | 27.4 | 24.1 |
| | Bayesian Ridge | 0.18 | 0.15 | 0.16 | 20.8 | 27.4 | 24.1 |
| | SVR - Linear | 0.11 | 0.15 | 0.13 | 21.6 | 27.4 | 24.5 |
| | SVR - Polynomial | 0.36 | 0.28 | 0.32 | 18.3 | 25.3 | 21.8 |

**Table B.16** Test Set Performance for the Models in the Aryl Halide Test: Ring Split (continued)

| Descriptor | ML Algorithm | $R^2$ | | | RMSE (%) | | |
|---|---|---|---|---|---|---|---|
| | | Phenyl | Pyridyl | Mean | Phenyl | Pyridyl | Mean |
| | SVR - RBF | 0.46 | 0.21 | 0.34 | 16.8 | 26.3 | 21.6 |
| | SVR - Sigmoid | -0.63 | -0.25 | -0.44 | 29.3 | 33.2 | 31.2 |
| | Gradient Boosting | 0.29 | 0.04 | 0.17 | 19.3 | 29.1 | 24.2 |
| | Random Forest | 0.19 | -0.07 | 0.06 | 20.7 | 30.8 | 25.7 |
| Fingerprints: Morgan1 | Linear Regression | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 |
| | Lasso | -0.98 | -0.02 | -0.50 | 32.3 | 29.9 | 31.1 |
| | Ridge | -0.32 | 0.09 | -0.12 | 26.4 | 28.4 | 27.4 |
| | Bayesian Ridge | -0.32 | 0.09 | -0.11 | 26.3 | 28.4 | 27.3 |
| | SVR - Linear | -0.37 | 0.10 | -0.14 | 26.9 | 28.2 | 27.5 |
| | SVR - Polynomial | -0.25 | 0.15 | -0.05 | 25.6 | 27.4 | 26.5 |
| | SVR - RBF | -0.14 | 0.15 | 0.01 | 24.5 | 27.3 | 25.9 |
| | SVR - Sigmoid | -0.36 | -0.16 | -0.26 | 26.8 | 31.9 | 29.4 |
| | Gradient Boosting | -0.24 | 0.06 | -0.09 | 25.6 | 28.7 | 27.2 |
| | Random Forest | 0.03 | -0.10 | -0.04 | 22.6 | 31.2 | 26.9 |
| Fingerprints: RDK | Linear Regression | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 |
| | Lasso | -0.58 | 0.32 | -0.13 | 28.8 | 24.6 | 26.7 |
| | Ridge | -0.27 | 0.11 | -0.08 | 25.9 | 28.0 | 26.9 |
| | Bayesian Ridge | -0.27 | 0.11 | -0.08 | 25.9 | 28.1 | 27.0 |
| | SVR - Linear | -0.30 | 0.10 | -0.10 | 26.2 | 28.1 | 27.1 |
| | SVR - Polynomial | -0.26 | 0.19 | -0.04 | 25.8 | 26.8 | 26.3 |
| | SVR - RBF | -0.23 | 0.18 | -0.02 | 25.4 | 26.9 | 26.2 |
| | SVR - Sigmoid | -0.39 | -0.32 | -0.35 | 27.0 | 34.1 | 30.6 |
| | Gradient Boosting | -0.73 | 0.36 | -0.18 | 30.1 | 23.7 | 26.9 |
| | Random Forest | -0.35 | 0.39 | 0.02 | 26.6 | 23.1 | 24.9 |
| Tanimoto: MACCS | Linear Regression | -0.19 | 0.17 | -0.01 | 25.0 | 27.1 | 26.1 |
| | Lasso | 0.12 | 0.15 | 0.13 | 21.6 | 27.4 | 24.5 |
| | Ridge | 0.36 | 0.06 | 0.21 | 18.3 | 28.8 | 23.6 |
| | Bayesian Ridge | 0.36 | 0.06 | 0.21 | 18.3 | 28.8 | 23.6 |
| | SVR - Polynomial | 0.22 | -0.02 | 0.10 | 20.2 | 29.9 | 25.1 |
| | SVR - RBF | 0.15 | -0.05 | 0.05 | 21.1 | 30.4 | 25.8 |
| | SVR - Sigmoid | 0.08 | 0.00 | 0.04 | 22.0 | 29.7 | 25.8 |
| | SVR - Precomputed | 0.37 | 0.06 | 0.21 | 18.2 | 28.8 | 23.5 |
| | Gradient Boosting | -0.15 | 0.05 | -0.05 | 24.6 | 29.0 | 26.8 |
| | Random Forest | -0.72 | -0.11 | -0.41 | 30.1 | 31.3 | 30.7 |
| Tanimoto: Morgan1 | Linear Regression | 0.17 | 0.01 | 0.09 | 20.9 | 29.5 | 25.2 |
| | Lasso | 0.11 | 0.01 | 0.06 | 21.6 | 29.6 | 25.6 |
| | Ridge | 0.09 | 0.01 | 0.05 | 21.9 | 29.5 | 25.7 |
| | Bayesian Ridge | 0.09 | 0.01 | 0.05 | 21.9 | 29.5 | 25.7 |
| | SVR - Polynomial | -0.03 | -0.03 | -0.03 | 23.2 | 30.1 | 26.7 |
| | SVR - RBF | -0.07 | -0.05 | -0.06 | 23.7 | 30.4 | 27.1 |
| | SVR - Sigmoid | 0.11 | 0.02 | 0.07 | 21.6 | 29.3 | 25.5 |
| | SVR - Precomputed | 0.08 | 0.01 | 0.05 | 22.0 | 29.5 | 25.7 |
| | Gradient Boosting | <-1.00 | 0.14 | -0.85 | 38.7 | 27.5 | 33.1 |
| | Random Forest | <-1.00 | 0.03 | <-1.00 | 48.1 | 29.3 | 38.7 |
| Tanimoto: RDK | Linear Regression | -0.46 | -0.09 | -0.27 | 27.7 | 31.0 | 29.3 |
| | Lasso | -0.35 | -0.10 | -0.22 | 26.6 | 31.1 | 28.9 |
| | Ridge | -0.32 | -0.09 | -0.21 | 26.4 | 31.0 | 28.7 |
| | Bayesian Ridge | -0.32 | -0.09 | -0.21 | 26.4 | 31.1 | 28.7 |
| | SVR - Polynomial | -0.42 | -0.15 | -0.29 | 27.3 | 31.9 | 29.6 |
| | SVR - RBF | -0.45 | -0.17 | -0.31 | 27.6 | 32.1 | 29.9 |
| | SVR - Sigmoid | -0.23 | -0.06 | -0.15 | 25.5 | 30.6 | 28.0 |
| | SVR - Precomputed | -0.32 | -0.09 | -0.20 | 26.3 | 31.0 | 28.7 |
| | Gradient Boosting | -0.85 | -0.80 | -0.83 | 31.2 | 39.8 | 35.5 |
| | Random Forest | <-1.00 | -0.63 | <-1.00 | 49.7 | 37.9 | 43.8 |
| WL | Linear Regression | -0.20 | 0.15 | -0.02 | 25.1 | 27.4 | 26.2 |
| | Lasso | -0.33 | 0.16 | -0.08 | 26.4 | 27.2 | 26.8 |
| | Ridge | -0.19 | 0.12 | -0.03 | 25.0 | 27.8 | 26.4 |
| | Bayesian Ridge | -0.19 | 0.12 | -0.03 | 25.0 | 27.8 | 26.4 |
| | SVR - Polynomial | -0.14 | 0.06 | -0.04 | 24.5 | 28.8 | 26.6 |

**Table B.16** Test Set Performance for the Models in the Aryl Halide Test: Ring Split (continued)

| Descriptor | ML Algorithm | $R^2$ | | | RMSE (%) | | |
|---|---|---|---|---|---|---|---|
| | | Phenyl | Pyridyl | Mean | Phenyl | Pyridyl | Mean |
| | SVR - RBF | -0.14 | 0.03 | -0.05 | 24.5 | 29.2 | 26.8 |
| | SVR - Sigmoid | -0.26 | 0.14 | -0.06 | 25.8 | 27.5 | 26.6 |
| | SVR - Precomputed | -0.19 | 0.12 | -0.04 | 25.1 | 27.8 | 26.4 |
| | Gradient Boosting | -0.49 | <-1.00 | -0.94 | 28.0 | 45.9 | 36.9 |
| | Random Forest | -0.53 | <-1.00 | <-1.00 | 28.4 | 49.9 | 39.1 |

## B.7.3 Aryl Halide Test: Halide Split

### Grid Search Cross-Validated Performance

Table B.17: Grid Search Cross-Validated Performance for the Models in the Aryl Halide Test: Halide Split

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Aryl Cl | Aryl Br | Aryl I | Mean | Aryl Cl | Aryl Br | Aryl I | Mean |
| One-hot | Linear Regression | 0.68 | 0.71 | 0.67 | 0.69 | 14.4 | 15.1 | 14.8 | 14.8 |
| | Lasso | 0.68 | 0.71 | 0.68 | 0.69 | 14.3 | 14.9 | 14.8 | 14.7 |
| | Ridge | 0.68 | 0.71 | 0.68 | 0.69 | 14.3 | 14.9 | 14.8 | 14.7 |
| | Bayesian Ridge | 0.68 | 0.71 | 0.68 | 0.69 | 14.3 | 14.9 | 14.8 | 14.7 |
| | SVR - Linear | 0.68 | 0.71 | 0.67 | 0.69 | 14.4 | 15.0 | 14.8 | 14.7 |
| | SVR - Polynomial | 0.88 | 0.91 | 0.90 | 0.89 | 8.9 | 8.5 | 8.4 | 8.6 |
| | SVR - RBF | 0.89 | 0.91 | 0.91 | 0.90 | 8.4 | 8.1 | 7.9 | 8.1 |
| | SVR - Sigmoid | 0.58 | 0.60 | 0.55 | 0.58 | 16.5 | 17.7 | 17.5 | 17.2 |
| | Gradient Boosting | 0.87 | 0.90 | 0.90 | 0.89 | 9.1 | 8.7 | 8.4 | 8.7 |
| | Random Forest | 0.87 | 0.90 | 0.88 | 0.88 | 9.1 | 8.7 | 9.1 | 9.0 |
| Quantum Chemical | Linear Regression | 0.68 | 0.71 | 0.68 | 0.69 | 14.4 | 15.0 | 14.8 | 14.7 |
| | Lasso | 0.67 | 0.71 | 0.68 | 0.69 | 14.6 | 15.0 | 14.8 | 14.8 |
| | Ridge | 0.68 | 0.71 | 0.68 | 0.69 | 14.4 | 14.9 | 14.8 | 14.7 |
| | Bayesian Ridge | 0.68 | 0.71 | 0.67 | 0.69 | 14.5 | 15.1 | 14.8 | 14.8 |
| | SVR - Linear | 0.68 | 0.71 | 0.67 | 0.69 | 14.5 | 15.0 | 14.8 | 14.8 |
| | SVR - Polynomial | 0.88 | 0.90 | 0.89 | 0.89 | 8.9 | 9.0 | 8.6 | 8.8 |
| | SVR - RBF | 0.90 | 0.90 | 0.88 | 0.89 | 8.0 | 8.9 | 8.9 | 8.6 |
| | SVR - Sigmoid | 0.50 | 0.56 | 0.34 | 0.47 | 18.0 | 18.6 | 21.1 | 19.2 |
| | Gradient Boosting | 0.89 | 0.92 | 0.91 | 0.91 | 8.4 | 7.8 | 7.7 | 8.0 |
| | Random Forest | 0.92 | 0.93 | 0.92 | 0.92 | 7.4 | 7.5 | 7.3 | 7.4 |
| Fingerprints: MACCS | Linear Regression | 0.61 | 0.64 | 0.59 | 0.61 | 15.9 | 16.7 | 16.7 | 16.4 |
| | Lasso | 0.64 | 0.65 | 0.61 | 0.63 | 15.2 | 16.5 | 16.3 | 16.0 |
| | Ridge | 0.65 | 0.65 | 0.61 | 0.64 | 15.2 | 16.5 | 16.3 | 16.0 |
| | Bayesian Ridge | 0.64 | 0.65 | 0.61 | 0.64 | 15.2 | 16.5 | 16.2 | 16.0 |
| | SVR - Linear | 0.64 | 0.65 | 0.61 | 0.63 | 15.3 | 16.5 | 16.3 | 16.0 |
| | SVR - Polynomial | 0.87 | 0.86 | 0.85 | 0.86 | 9.3 | 10.3 | 10.2 | 9.9 |
| | SVR - RBF | 0.87 | 0.86 | 0.84 | 0.85 | 9.3 | 10.5 | 10.4 | 10.1 |
| | SVR - Sigmoid | 0.33 | 0.26 | 0.21 | 0.27 | 20.9 | 24.0 | 23.1 | 22.7 |
| | Gradient Boosting | 0.88 | 0.89 | 0.88 | 0.88 | 8.9 | 9.4 | 9.1 | 9.1 |
| | Random Forest | 0.92 | 0.91 | 0.90 | 0.91 | 7.3 | 8.4 | 8.2 | 8.0 |
| Fingerprints: Morgan1 | Linear Regression | 0.67 | 0.68 | 0.65 | 0.67 | 14.6 | 15.8 | 15.5 | 15.3 |
| | Lasso | 0.68 | 0.69 | 0.65 | 0.68 | 14.3 | 15.6 | 15.3 | 15.1 |
| | Ridge | 0.68 | 0.69 | 0.65 | 0.68 | 14.3 | 15.6 | 15.3 | 15.1 |
| | Bayesian Ridge | 0.68 | 0.69 | 0.65 | 0.67 | 14.3 | 15.6 | 15.3 | 15.1 |
| | SVR - Linear | 0.68 | 0.68 | 0.65 | 0.67 | 14.4 | 15.6 | 15.3 | 15.1 |
| | SVR - Polynomial | 0.89 | 0.91 | 0.90 | 0.90 | 8.4 | 8.4 | 8.2 | 8.3 |
| | SVR - RBF | 0.91 | 0.92 | 0.91 | 0.91 | 7.8 | 7.8 | 7.9 | 7.8 |
| | SVR - Sigmoid | 0.49 | 0.48 | 0.38 | 0.45 | 18.3 | 20.2 | 20.6 | 19.7 |
| | Gradient Boosting | 0.89 | 0.91 | 0.90 | 0.90 | 8.3 | 8.5 | 8.4 | 8.4 |
| | Random Forest | 0.92 | 0.93 | 0.91 | 0.92 | 7.2 | 7.4 | 7.9 | 7.5 |
| Fingerprints: RDK | Linear Regression | 0.66 | 0.69 | 0.67 | 0.67 | 14.8 | 15.5 | 15.0 | 15.1 |
| | Lasso | 0.68 | 0.71 | 0.68 | 0.69 | 14.3 | 14.9 | 14.8 | 14.7 |
| | Ridge | 0.68 | 0.71 | 0.68 | 0.69 | 14.3 | 14.9 | 14.8 | 14.7 |
| | Bayesian Ridge | 0.68 | 0.71 | 0.68 | 0.69 | 14.3 | 14.9 | 14.8 | 14.7 |
| | SVR - Linear | 0.68 | 0.71 | 0.67 | 0.69 | 14.4 | 15.0 | 14.8 | 14.7 |
| | SVR - Polynomial | 0.88 | 0.91 | 0.90 | 0.90 | 8.7 | 8.2 | 8.1 | 8.4 |
| | SVR - RBF | 0.89 | 0.91 | 0.90 | 0.90 | 8.6 | 8.1 | 8.1 | 8.3 |
| | SVR - Sigmoid | 0.35 | 0.28 | 0.24 | 0.29 | 20.6 | 23.6 | 22.7 | 22.3 |
| | Gradient Boosting | 0.87 | 0.91 | 0.91 | 0.90 | 9.1 | 8.3 | 8.0 | 8.5 |
| | Random Forest | 0.90 | 0.92 | 0.92 | 0.92 | 8.1 | 7.6 | 7.2 | 7.6 |
| Tanimoto: MACCS | Linear Regression | 0.87 | 0.86 | 0.85 | 0.86 | 9.3 | 10.5 | 10.0 | 10.0 |
| | Lasso | 0.88 | 0.89 | 0.88 | 0.88 | 8.7 | 9.4 | 8.8 | 9.0 |
| | Ridge | 0.89 | 0.89 | 0.89 | 0.89 | 8.4 | 9.1 | 8.6 | 8.7 |

**Table B.17** Grid Search Cross-Validated Performance for the Models in the Aryl Halide Test: Halide Split (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Aryl Cl | Aryl Br | Aryl I | Mean | Aryl Cl | Aryl Br | Aryl I | Mean |
| | Bayesian Ridge | 0.88 | 0.88 | 0.88 | 0.88 | 8.7 | 9.6 | 9.0 | 9.1 |
| | SVR - Polynomial | 0.89 | 0.88 | 0.88 | 0.88 | 8.5 | 9.6 | 9.0 | 9.0 |
| | SVR - RBF | 0.87 | 0.87 | 0.87 | 0.87 | 9.2 | 9.9 | 9.4 | 9.5 |
| | SVR - Sigmoid | 0.66 | 0.69 | 0.63 | 0.66 | 14.9 | 15.6 | 15.8 | 15.4 |
| | SVR - Precomputed | 0.88 | 0.89 | 0.88 | 0.88 | 9.0 | 9.4 | 8.9 | 9.1 |
| | Gradient Boosting | 0.75 | 0.75 | 0.72 | 0.74 | 12.8 | 13.9 | 13.8 | 13.5 |
| | Random Forest | 0.65 | 0.71 | 0.67 | 0.68 | 15.1 | 15.0 | 14.9 | 15.0 |
| Tanimoto: Morgan1 | Linear Regression | 0.92 | 0.93 | 0.92 | 0.92 | 7.2 | 7.4 | 7.5 | 7.4 |
| | Lasso | 0.92 | 0.93 | 0.92 | 0.93 | 7.1 | 7.2 | 7.3 | 7.2 |
| | Ridge | 0.92 | 0.93 | 0.92 | 0.93 | 7.1 | 7.2 | 7.3 | 7.2 |
| | Bayesian Ridge | 0.92 | 0.93 | 0.92 | 0.93 | 7.0 | 7.2 | 7.2 | 7.1 |
| | SVR - Polynomial | 0.91 | 0.92 | 0.91 | 0.91 | 7.6 | 7.7 | 7.8 | 7.7 |
| | SVR - RBF | 0.89 | 0.91 | 0.89 | 0.90 | 8.4 | 8.4 | 8.5 | 8.4 |
| | SVR - Sigmoid | 0.87 | 0.88 | 0.86 | 0.87 | 9.0 | 9.6 | 9.7 | 9.4 |
| | SVR - Precomputed | 0.93 | 0.93 | 0.92 | 0.93 | 7.0 | 7.2 | 7.3 | 7.2 |
| | Gradient Boosting | 0.79 | 0.82 | 0.78 | 0.80 | 11.8 | 11.8 | 12.1 | 11.9 |
| | Random Forest | 0.67 | 0.76 | 0.71 | 0.71 | 14.7 | 13.7 | 13.9 | 14.1 |
| Tanimoto: RDK | Linear Regression | 0.87 | 0.92 | 0.90 | 0.90 | 9.1 | 8.0 | 8.1 | 8.4 |
| | Lasso | 0.88 | 0.92 | 0.91 | 0.90 | 8.8 | 7.8 | 7.9 | 8.1 |
| | Ridge | 0.89 | 0.92 | 0.91 | 0.91 | 8.6 | 7.7 | 7.7 | 8.0 |
| | Bayesian Ridge | 0.89 | 0.92 | 0.91 | 0.91 | 8.6 | 7.7 | 7.7 | 8.0 |
| | SVR - Polynomial | 0.87 | 0.91 | 0.90 | 0.89 | 9.3 | 8.2 | 8.2 | 8.6 |
| | SVR - RBF | 0.85 | 0.90 | 0.89 | 0.88 | 9.9 | 8.7 | 8.7 | 9.1 |
| | SVR - Sigmoid | 0.82 | 0.85 | 0.83 | 0.83 | 10.8 | 10.6 | 10.8 | 10.7 |
| | SVR - Precomputed | 0.88 | 0.92 | 0.91 | 0.90 | 8.7 | 7.8 | 7.8 | 8.1 |
| | Gradient Boosting | 0.73 | 0.82 | 0.79 | 0.78 | 13.4 | 11.9 | 12.0 | 12.4 |
| | Random Forest | 0.60 | 0.76 | 0.71 | 0.69 | 16.2 | 13.6 | 14.0 | 14.6 |
| WL | Linear Regression | 0.90 | 0.92 | 0.91 | 0.91 | 8.0 | 7.9 | 7.6 | 7.9 |
| | Lasso | 0.91 | 0.92 | 0.92 | 0.92 | 7.8 | 7.7 | 7.4 | 7.6 |
| | Ridge | 0.91 | 0.93 | 0.92 | 0.92 | 7.6 | 7.5 | 7.3 | 7.5 |
| | Bayesian Ridge | 0.91 | 0.93 | 0.92 | 0.92 | 7.7 | 7.5 | 7.3 | 7.5 |
| | SVR - Polynomial | 0.91 | 0.92 | 0.92 | 0.92 | 7.7 | 7.7 | 7.5 | 7.6 |
| | SVR - RBF | 0.90 | 0.92 | 0.91 | 0.91 | 8.0 | 7.8 | 7.7 | 7.8 |
| | SVR - Sigmoid | 0.81 | 0.83 | 0.81 | 0.82 | 11.0 | 11.4 | 11.3 | 11.2 |
| | SVR - Precomputed | 0.91 | 0.92 | 0.92 | 0.92 | 7.7 | 7.8 | 7.4 | 7.6 |
| | Gradient Boosting | 0.83 | 0.86 | 0.83 | 0.84 | 10.6 | 10.5 | 10.6 | 10.5 |
| | Random Forest | 0.75 | 0.82 | 0.76 | 0.78 | 12.7 | 11.7 | 12.6 | 12.3 |

### Training Set Performance

Table B.18: Training Set Performance for the Models in the Aryl Halide Test: Halide Split

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Aryl Cl | Aryl Br | Aryl I | Mean | Aryl Cl | Aryl Br | Aryl I | Mean |
| One-hot | Linear Regression | 0.69 | 0.72 | 0.69 | 0.70 | 14.1 | 14.7 | 14.6 | 14.5 |
| | Lasso | 0.69 | 0.72 | 0.69 | 0.70 | 14.1 | 14.7 | 14.5 | 14.5 |
| | Ridge | 0.69 | 0.72 | 0.69 | 0.70 | 14.1 | 14.7 | 14.5 | 14.5 |
| | Bayesian Ridge | 0.69 | 0.72 | 0.69 | 0.70 | 14.1 | 14.7 | 14.5 | 14.5 |
| | SVR - Linear | 0.69 | 0.72 | 0.69 | 0.70 | 14.2 | 14.8 | 14.6 | 14.5 |
| | SVR - Polynomial | 0.97 | 0.98 | 0.97 | 0.97 | 4.7 | 4.3 | 4.6 | 4.5 |
| | SVR - RBF | 0.98 | 1.00 | 1.00 | 0.99 | 3.9 | 1.0 | 1.0 | 1.9 |
| | SVR - Sigmoid | 0.58 | 0.61 | 0.56 | 0.58 | 16.5 | 17.4 | 17.3 | 17.1 |
| | Gradient Boosting | 0.95 | 0.96 | 0.95 | 0.95 | 5.7 | 5.5 | 6.0 | 5.7 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 3.1 | 2.9 | 3.1 | 3.0 |

**Table B.18** Training Set Performance for the Models in the Aryl Halide Test: Halide Split (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Aryl Cl | Aryl Br | Aryl I | Mean | Aryl Cl | Aryl Br | Aryl I | Mean |
| Quantum Chemical | Linear Regression | 0.69 | 0.72 | 0.69 | 0.70 | 14.2 | 14.7 | 14.6 | 14.5 |
| | Lasso | 0.68 | 0.72 | 0.68 | 0.69 | 14.4 | 14.8 | 14.6 | 14.6 |
| | Ridge | 0.69 | 0.72 | 0.69 | 0.70 | 14.2 | 14.7 | 14.6 | 14.5 |
| | Bayesian Ridge | 0.69 | 0.72 | 0.68 | 0.70 | 14.3 | 14.9 | 14.6 | 14.6 |
| | SVR - Linear | 0.69 | 0.72 | 0.68 | 0.70 | 14.3 | 14.8 | 14.6 | 14.6 |
| | SVR - Polynomial | 0.95 | 0.96 | 0.95 | 0.95 | 5.6 | 5.4 | 5.9 | 5.6 |
| | SVR - RBF | 0.99 | 0.99 | 0.97 | 0.98 | 3.0 | 2.9 | 4.2 | 3.4 |
| | SVR - Sigmoid | 0.49 | 0.54 | 0.30 | 0.44 | 18.2 | 18.9 | 21.8 | 19.6 |
| | Gradient Boosting | 0.96 | 0.97 | 0.97 | 0.97 | 4.8 | 5.0 | 4.7 | 4.9 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.4 | 2.5 | 2.5 | 2.5 |
| Fingerprints: MACCS | Linear Regression | 0.61 | 0.63 | 0.62 | 0.62 | 16.0 | 17.0 | 16.1 | 16.4 |
| | Lasso | 0.65 | 0.66 | 0.62 | 0.64 | 15.0 | 16.3 | 16.1 | 15.8 |
| | Ridge | 0.65 | 0.66 | 0.62 | 0.64 | 15.0 | 16.3 | 16.1 | 15.8 |
| | Bayesian Ridge | 0.65 | 0.66 | 0.62 | 0.64 | 15.0 | 16.3 | 16.1 | 15.8 |
| | SVR - Linear | 0.65 | 0.66 | 0.61 | 0.64 | 15.1 | 16.3 | 16.2 | 15.9 |
| | SVR - Polynomial | 0.91 | 0.91 | 0.90 | 0.91 | 7.7 | 8.4 | 8.4 | 8.1 |
| | SVR - RBF | 0.91 | 0.91 | 0.90 | 0.90 | 7.5 | 8.5 | 8.4 | 8.2 |
| | SVR - Sigmoid | 0.36 | 0.30 | 0.25 | 0.30 | 20.4 | 23.3 | 22.5 | 22.1 |
| | Gradient Boosting | 0.92 | 0.93 | 0.92 | 0.92 | 7.2 | 7.6 | 7.5 | 7.4 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.4 | 2.7 | 2.8 | 2.6 |
| Fingerprints: Morgan1 | Linear Regression | 0.69 | 0.69 | 0.66 | 0.68 | 14.2 | 15.4 | 15.2 | 15.0 |
| | Lasso | 0.69 | 0.70 | 0.66 | 0.68 | 14.1 | 15.4 | 15.1 | 14.9 |
| | Ridge | 0.69 | 0.70 | 0.66 | 0.68 | 14.1 | 15.4 | 15.1 | 14.9 |
| | Bayesian Ridge | 0.69 | 0.70 | 0.66 | 0.68 | 14.1 | 15.4 | 15.1 | 14.9 |
| | SVR - Linear | 0.69 | 0.69 | 0.66 | 0.68 | 14.2 | 15.4 | 15.2 | 14.9 |
| | SVR - Polynomial | 0.95 | 0.96 | 0.95 | 0.95 | 5.9 | 5.9 | 5.9 | 5.9 |
| | SVR - RBF | 0.96 | 0.97 | 0.96 | 0.97 | 4.8 | 4.9 | 4.9 | 4.9 |
| | SVR - Sigmoid | 0.51 | 0.50 | 0.41 | 0.47 | 17.9 | 19.7 | 20.0 | 19.2 |
| | Gradient Boosting | 0.95 | 0.96 | 0.94 | 0.95 | 5.8 | 5.9 | 6.2 | 6.0 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.4 | 2.4 | 2.7 | 2.5 |
| Fingerprints: RDK | Linear Regression | 0.69 | 0.72 | 0.68 | 0.70 | 14.2 | 14.7 | 14.6 | 14.5 |
| | Lasso | 0.69 | 0.72 | 0.69 | 0.70 | 14.1 | 14.7 | 14.5 | 14.5 |
| | Ridge | 0.69 | 0.72 | 0.69 | 0.70 | 14.1 | 14.7 | 14.5 | 14.5 |
| | Bayesian Ridge | 0.69 | 0.72 | 0.69 | 0.70 | 14.1 | 14.7 | 14.5 | 14.5 |
| | SVR - Linear | 0.69 | 0.72 | 0.69 | 0.70 | 14.2 | 14.8 | 14.6 | 14.5 |
| | SVR - Polynomial | 0.95 | 0.97 | 0.96 | 0.96 | 5.6 | 5.2 | 5.1 | 5.3 |
| | SVR - RBF | 0.96 | 0.97 | 0.97 | 0.97 | 5.1 | 4.7 | 4.5 | 4.8 |
| | SVR - Sigmoid | 0.39 | 0.33 | 0.28 | 0.33 | 20.0 | 22.9 | 22.0 | 21.6 |
| | Gradient Boosting | 0.95 | 0.96 | 0.96 | 0.96 | 5.9 | 5.3 | 5.1 | 5.4 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.6 | 2.5 | 2.5 | 2.5 |
| Tanimoto: MACCS | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 0.99 | 0.99 | 0.99 | 0.99 | 2.4 | 2.6 | 2.5 | 2.5 |
| | Ridge | 0.98 | 0.98 | 0.98 | 0.98 | 3.3 | 3.8 | 3.6 | 3.6 |
| | Bayesian Ridge | 0.96 | 0.96 | 0.96 | 0.96 | 4.8 | 5.9 | 5.2 | 5.3 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 1.0 | 1.0 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.1 | 1.1 | 1.1 |
| | SVR - Sigmoid | 0.68 | 0.72 | 0.66 | 0.69 | 14.3 | 14.9 | 15.1 | 14.8 |
| | SVR - Precomputed | 0.97 | 0.97 | 0.98 | 0.97 | 4.0 | 5.1 | 4.0 | 4.4 |
| | Gradient Boosting | 1.00 | 0.94 | 0.94 | 0.96 | 0.6 | 6.6 | 6.6 | 4.6 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 3.6 | 4.0 | 4.0 | 3.8 |
| Tanimoto: Morgan1 | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 0.9 | 0.9 | 0.9 |
| | Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 0.6 | 0.7 | 0.8 | 0.7 |
| | Bayesian Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 1.1 | 1.4 | 1.1 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 1.1 | 0.9 | 1.2 | 1.1 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 0.9 | 1.0 | 1.0 |
| | SVR - Sigmoid | 0.85 | 0.87 | 0.87 | 0.86 | 9.8 | 10.2 | 9.3 | 9.8 |
| | SVR - Precomputed | 0.99 | 0.99 | 0.99 | 0.99 | 1.9 | 2.1 | 2.4 | 2.1 |
| | Gradient Boosting | 0.99 | 1.00 | 1.00 | 1.00 | 2.2 | 0.6 | 0.6 | 1.1 |

**Table B.18** Training Set Performance for the Models in the Aryl Halide Test: Halide Split (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Aryl Cl | Aryl Br | Aryl I | Mean | Aryl Cl | Aryl Br | Aryl I | Mean |
| | Random Forest | 0.98 | 0.98 | 0.97 | 0.98 | 3.9 | 4.2 | 4.2 | 4.1 |
| Tanimoto: RDK | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 1.5 | 1.4 | 1.4 | 1.5 |
| | Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 1.6 | 1.5 | 1.5 | 1.5 |
| | Bayesian Ridge | 0.99 | 0.99 | 0.99 | 0.99 | 2.6 | 2.2 | 2.3 | 2.4 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 0.9 | 1.0 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 0.9 | 0.9 | 1.0 |
| | SVR - Sigmoid | 0.80 | 0.83 | 0.86 | 0.83 | 11.5 | 11.5 | 9.9 | 11.0 |
| | SVR - Precomputed | 1.00 | 1.00 | 1.00 | 1.00 | 1.3 | 1.0 | 1.2 | 1.2 |
| | Gradient Boosting | 0.99 | 1.00 | 0.99 | 0.99 | 2.3 | 0.7 | 2.1 | 1.7 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 3.8 | 3.9 | 4.1 | 3.9 |
| WL | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 1.6 | 1.6 | 1.6 | 1.6 |
| | Ridge | 0.99 | 1.00 | 0.99 | 0.99 | 1.9 | 1.9 | 1.9 | 1.9 |
| | Bayesian Ridge | 0.98 | 0.99 | 0.99 | 0.99 | 3.3 | 3.1 | 3.1 | 3.2 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 0.9 | 0.9 | 0.9 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 0.9 | 0.9 | 0.9 |
| | SVR - Sigmoid | 0.82 | 0.84 | 0.76 | 0.80 | 10.8 | 11.2 | 12.8 | 11.6 |
| | SVR - Precomputed | 1.00 | 1.00 | 1.00 | 1.00 | 1.7 | 1.5 | 1.6 | 1.6 |
| | Gradient Boosting | 1.00 | 0.99 | 1.00 | 1.00 | 0.7 | 2.2 | 0.6 | 1.2 |
| | Random Forest | 0.98 | 0.99 | 0.98 | 0.98 | 3.5 | 3.4 | 3.5 | 3.4 |

### Test Set Performance

Table B.19: Test Set Performance for the Models in the Aryl Halide Test: Halide Split

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Aryl Cl | Aryl Br | Aryl I | Mean | Aryl Cl | Aryl Br | Aryl I | Mean |
| One-hot | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | <-1.00 | 0.43 | -0.13 | -0.56 | 36.5 | 19.1 | 26.9 | 27.5 |
| | Ridge | <-1.00 | 0.36 | -0.13 | -0.52 | 35.4 | 20.1 | 26.9 | 27.5 |
| | Bayesian Ridge | <-1.00 | 0.36 | -0.14 | -0.52 | 35.3 | 20.1 | 27.0 | 27.5 |
| | SVR - Linear | <-1.00 | 0.37 | -0.15 | -0.53 | 35.4 | 20.1 | 27.1 | 27.6 |
| | SVR - Polynomial | <-1.00 | 0.40 | -0.09 | -0.47 | 34.8 | 19.5 | 26.4 | 26.9 |
| | SVR - RBF | <-1.00 | 0.50 | -0.04 | -0.57 | 37.7 | 17.8 | 25.8 | 27.1 |
| | SVR - Sigmoid | <-1.00 | 0.21 | -0.36 | -0.62 | 34.9 | 22.4 | 29.5 | 28.9 |
| | Gradient Boosting | <-1.00 | 0.47 | -0.12 | -0.56 | 36.8 | 18.3 | 26.7 | 27.3 |
| | Random Forest | <-1.00 | 0.59 | 0.40 | -0.72 | 43.1 | 16.1 | 19.6 | 26.3 |
| Quantum Chemical | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | <-1.00 | <-1.00 | <-1.00 | <-1.00 | 30.8 | 74.0 | >100.0 | >100.0 |
| | Ridge | <-1.00 | -0.48 | <-1.00 | <-1.00 | 34.0 | 30.6 | >100.0 | >100.0 |
| | Bayesian Ridge | <-1.00 | -0.50 | <-1.00 | <-1.00 | 33.9 | 30.9 | >100.0 | >100.0 |
| | SVR - Linear | <-1.00 | -0.61 | <-1.00 | <-1.00 | 34.6 | 32.0 | >100.0 | >100.0 |
| | SVR - Polynomial | <-1.00 | -5.42 | <-1.00 | <-1.00 | 35.6 | 63.9 | >100.0 | >100.0 |
| | SVR - RBF | <-1.00 | 0.06 | <-1.00 | -0.95 | 35.1 | 24.5 | 37.2 | 32.3 |
| | SVR - Sigmoid | <-1.00 | -0.34 | <-1.00 | <-1.00 | 33.1 | 29.2 | >100.0 | 86.9 |
| | Gradient Boosting | -0.69 | -0.18 | 0.62 | -0.08 | 27.5 | 27.5 | 15.7 | 23.5 |
| | Random Forest | -0.57 | -0.22 | 0.64 | -0.05 | 26.5 | 27.9 | 15.2 | 23.2 |
| Fingerprints: MACCS | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | <-1.00 | -0.30 | -0.95 | -0.81 | 31.2 | 28.7 | 35.4 | 31.8 |
| | Ridge | <-1.00 | 0.45 | -0.05 | -0.40 | 34.0 | 18.7 | 26.0 | 26.2 |
| | Bayesian Ridge | <-1.00 | 0.45 | -0.05 | -0.39 | 34.0 | 18.7 | 26.0 | 26.2 |
| | SVR - Linear | <-1.00 | 0.44 | -0.12 | -0.44 | 34.4 | 18.9 | 26.8 | 26.7 |
| | SVR - Polynomial | <-1.00 | 0.62 | 0.10 | -0.30 | 34.3 | 15.5 | 24.0 | 24.6 |

**Table B.19** Test Set Performance for the Models in the Aryl Halide Test: Halide Split (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Aryl Cl | Aryl Br | Aryl I | Mean | Aryl Cl | Aryl Br | Aryl I | Mean |
| | SVR - RBF | <-1.00 | 0.63 | 0.09 | -0.29 | 34.0 | 15.4 | 24.2 | 24.5 |
| | SVR - Sigmoid | <-1.00 | -0.02 | -0.73 | -0.68 | 32.1 | 25.5 | 33.3 | 30.3 |
| | Gradient Boosting | <-1.00 | 0.40 | 0.36 | -0.27 | 34.0 | 19.5 | 20.2 | 24.6 |
| | Random Forest | <-1.00 | 0.82 | 0.21 | -0.19 | 34.1 | 10.8 | 22.4 | 22.5 |
| Fingerprints: Morgan1 | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | <-1.00 | 0.38 | 0.42 | -0.42 | 37.0 | 19.9 | 19.2 | 25.3 |
| | Ridge | <-1.00 | 0.18 | 0.30 | -0.37 | 34.1 | 22.8 | 21.1 | 26.0 |
| | Bayesian Ridge | <-1.00 | 0.18 | 0.30 | -0.37 | 34.1 | 22.8 | 21.2 | 26.0 |
| | SVR - Linear | <-1.00 | 0.19 | 0.27 | -0.39 | 34.3 | 22.7 | 21.5 | 26.2 |
| | SVR - Polynomial | <-1.00 | 0.20 | 0.45 | -0.33 | 34.3 | 22.5 | 18.8 | 25.2 |
| | SVR - RBF | <-1.00 | 0.19 | 0.45 | -0.32 | 34.1 | 22.8 | 18.8 | 25.2 |
| | SVR - Sigmoid | <-1.00 | -0.04 | -0.40 | -0.60 | 32.5 | 25.7 | 29.9 | 29.4 |
| | Gradient Boosting | <-1.00 | -0.76 | -0.45 | <-1.00 | 36.2 | 33.5 | 30.4 | 33.4 |
| | Random Forest | <-1.00 | <-1.00 | 0.51 | -0.76 | 34.2 | 37.2 | 17.7 | 29.7 |
| Fingerprints: RDK | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | <-1.00 | 0.38 | 0.43 | -0.45 | 37.7 | 19.8 | 19.0 | 25.5 |
| | Ridge | <-1.00 | 0.48 | 0.15 | -0.38 | 35.2 | 18.1 | 23.3 | 25.5 |
| | Bayesian Ridge | <-1.00 | 0.48 | 0.15 | -0.38 | 35.2 | 18.1 | 23.4 | 25.6 |
| | SVR - Linear | <-1.00 | 0.49 | 0.14 | -0.38 | 35.2 | 18.1 | 23.5 | 25.6 |
| | SVR - Polynomial | <-1.00 | 0.64 | 0.27 | -0.31 | 35.7 | 15.2 | 21.6 | 24.2 |
| | SVR - RBF | <-1.00 | 0.63 | 0.25 | -0.30 | 35.4 | 15.3 | 21.9 | 24.2 |
| | SVR - Sigmoid | <-1.00 | -0.02 | -0.72 | -0.69 | 32.4 | 25.5 | 33.1 | 30.3 |
| | Gradient Boosting | <-1.00 | 0.67 | 0.60 | -0.24 | 36.6 | 14.6 | 16.0 | 22.4 |
| | Random Forest | <-1.00 | 0.75 | 0.56 | -0.18 | 35.7 | 12.5 | 16.8 | 21.7 |
| Tanimoto: MACCS | Linear Regression | <-1.00 | 0.68 | 0.14 | -0.28 | 34.5 | 14.4 | 23.5 | 24.1 |
| | Lasso | <-1.00 | 0.67 | 0.14 | -0.24 | 33.6 | 14.6 | 23.5 | 23.9 |
| | Ridge | <-1.00 | 0.64 | 0.09 | -0.22 | 32.8 | 15.1 | 24.2 | 24.0 |
| | Bayesian Ridge | <-1.00 | 0.63 | 0.08 | -0.23 | 32.7 | 15.4 | 24.2 | 24.1 |
| | SVR - Polynomial | <-1.00 | 0.63 | 0.07 | -0.20 | 32.3 | 15.3 | 24.3 | 23.9 |
| | SVR - RBF | <-1.00 | 0.62 | 0.07 | -0.20 | 32.0 | 15.5 | 24.5 | 24.0 |
| | SVR - Sigmoid | <-1.00 | 0.39 | -0.19 | -0.44 | 33.6 | 19.7 | 27.5 | 27.0 |
| | SVR - Precomputed | <-1.00 | 0.64 | 0.08 | -0.23 | 32.9 | 15.1 | 24.2 | 24.1 |
| | Gradient Boosting | <-1.00 | 0.09 | -0.40 | -0.53 | 32.0 | 24.1 | 29.9 | 28.6 |
| | Random Forest | <-1.00 | -0.66 | -0.85 | -1.07 | 34.7 | 32.5 | 34.4 | 33.9 |
| Tanimoto: Morgan1 | Linear Regression | <-1.00 | 0.21 | -0.07 | -0.39 | 32.1 | 22.4 | 26.1 | 26.9 |
| | Lasso | <-1.00 | 0.21 | 0.21 | -0.32 | 32.5 | 22.4 | 22.5 | 25.8 |
| | Ridge | <-1.00 | 0.19 | 0.21 | -0.29 | 31.9 | 22.7 | 22.5 | 25.7 |
| | Bayesian Ridge | <-1.00 | 0.19 | 0.21 | -0.29 | 31.9 | 22.7 | 22.5 | 25.7 |
| | SVR - Polynomial | <-1.00 | 0.22 | 0.11 | -0.29 | 31.4 | 22.3 | 23.8 | 25.8 |
| | SVR - RBF | <-1.00 | 0.22 | 0.08 | -0.29 | 31.2 | 22.3 | 24.3 | 25.9 |
| | SVR - Sigmoid | <-1.00 | 0.17 | 0.23 | -0.32 | 32.4 | 23.0 | 22.2 | 25.9 |
| | SVR - Precomputed | <-1.00 | 0.20 | 0.20 | -0.29 | 31.9 | 22.6 | 22.6 | 25.7 |
| | Gradient Boosting | <-1.00 | 0.30 | -0.12 | -0.67 | 37.8 | 21.1 | 26.7 | 28.6 |
| | Random Forest | <-1.00 | 0.13 | 0.02 | -0.52 | 34.8 | 23.5 | 25.0 | 27.8 |
| Tanimoto: RDK | Linear Regression | <-1.00 | 0.43 | -0.01 | -0.31 | 32.4 | 19.0 | 25.5 | 25.6 |
| | Lasso | <-1.00 | 0.42 | -0.09 | -0.34 | 32.6 | 19.1 | 26.4 | 26.0 |
| | Ridge | <-1.00 | 0.43 | -0.05 | -0.33 | 32.5 | 19.1 | 25.9 | 25.8 |
| | Bayesian Ridge | <-1.00 | 0.42 | -0.05 | -0.33 | 32.5 | 19.1 | 25.9 | 25.8 |
| | SVR - Polynomial | <-1.00 | 0.36 | -0.11 | -0.36 | 32.3 | 20.1 | 26.7 | 26.4 |
| | SVR - RBF | <-1.00 | 0.33 | -0.14 | -0.38 | 32.3 | 20.6 | 27.0 | 26.6 |
| | SVR - Sigmoid | <-1.00 | 0.41 | -0.03 | -0.33 | 32.6 | 19.4 | 25.7 | 25.9 |
| | SVR - Precomputed | <-1.00 | 0.43 | -0.05 | -0.33 | 32.5 | 19.1 | 25.9 | 25.8 |
| | Gradient Boosting | -0.86 | -0.30 | <-1.00 | -0.93 | 28.8 | 28.8 | 41.0 | 32.9 |
| | Random Forest | <-1.00 | -0.75 | <-1.00 | <-1.00 | 31.2 | 33.3 | 43.4 | 36.0 |
| WL | Linear Regression | <-1.00 | 0.65 | 0.28 | -0.30 | 35.5 | 15.0 | 21.5 | 24.0 |
| | Lasso | <-1.00 | 0.64 | 0.14 | -0.30 | 34.6 | 15.2 | 23.4 | 24.4 |
| | Ridge | <-1.00 | 0.63 | 0.12 | -0.27 | 33.9 | 15.4 | 23.7 | 24.3 |
| | Bayesian Ridge | <-1.00 | 0.63 | 0.12 | -0.27 | 33.9 | 15.4 | 23.8 | 24.4 |
| | SVR - Polynomial | <-1.00 | 0.59 | 0.07 | -0.27 | 33.3 | 16.2 | 24.4 | 24.6 |

**Table B.19** Test Set Performance for the Models in the Aryl Halide Test: Halide Split (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Aryl Cl | Aryl Br | Aryl I | Mean | Aryl Cl | Aryl Br | Aryl I | Mean |
| | SVR - RBF | <-1.00 | 0.56 | 0.04 | -0.28 | 33.1 | 16.7 | 24.7 | 24.8 |
| | SVR - Sigmoid | <-1.00 | 0.59 | 0.08 | -0.32 | 34.4 | 16.1 | 24.3 | 24.9 |
| | SVR - Precomputed | <-1.00 | 0.63 | 0.12 | -0.27 | 33.9 | 15.3 | 23.7 | 24.3 |
| | Gradient Boosting | <-1.00 | 0.53 | 0.11 | -0.15 | 30.6 | 17.2 | 23.8 | 23.9 |
| | Random Forest | <-1.00 | 0.53 | 0.16 | -0.02 | 28.1 | 17.2 | 23.2 | 22.8 |

### B.7.4 Leave-One-Base-Out Test

#### Grid Search Cross-Validated Performance

Table B.20: Grid Search Cross-Validated Performance for the Models in the Leave-One-Base-Out Test

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | BTMG | MTBD | P2Et | Mean | BTMG | MTBD | P2Et | Mean |
| One-hot | Linear Regression | 0.67 | 0.65 | 0.79 | 0.70 | 16.2 | 14.7 | 12.9 | 14.6 |
| | Lasso | 0.68 | 0.66 | 0.79 | 0.71 | 16.0 | 14.5 | 12.8 | 14.4 |
| | Ridge | 0.68 | 0.66 | 0.79 | 0.71 | 16.0 | 14.5 | 12.8 | 14.4 |
| | Bayesian Ridge | 0.68 | 0.66 | 0.79 | 0.71 | 16.0 | 14.5 | 12.8 | 14.4 |
| | SVR - Linear | 0.68 | 0.65 | 0.79 | 0.71 | 16.0 | 14.6 | 12.9 | 14.5 |
| | SVR - Polynomial | 0.89 | 0.86 | 0.94 | 0.89 | 9.6 | 9.3 | 6.8 | 8.6 |
| | SVR - RBF | 0.90 | 0.87 | 0.95 | 0.91 | 9.2 | 8.8 | 6.4 | 8.1 |
| | SVR - Sigmoid | 0.51 | 0.50 | 0.65 | 0.55 | 19.8 | 17.4 | 16.6 | 18.0 |
| | Gradient Boosting | 0.89 | 0.85 | 0.94 | 0.89 | 9.5 | 9.7 | 6.7 | 8.6 |
| | Random Forest | 0.90 | 0.86 | 0.92 | 0.89 | 9.0 | 9.4 | 7.7 | 8.7 |
| Quantum Chemical | Linear Regression | 0.68 | 0.65 | 0.79 | 0.71 | 16.0 | 14.6 | 12.8 | 14.5 |
| | Lasso | 0.68 | 0.65 | 0.78 | 0.70 | 16.1 | 14.7 | 13.1 | 14.6 |
| | Ridge | 0.68 | 0.65 | 0.79 | 0.71 | 16.0 | 14.6 | 12.8 | 14.5 |
| | Bayesian Ridge | 0.68 | 0.65 | 0.78 | 0.70 | 16.2 | 14.7 | 13.0 | 14.6 |
| | SVR - Linear | 0.68 | 0.65 | 0.79 | 0.71 | 16.0 | 14.7 | 12.9 | 14.5 |
| | SVR - Polynomial | 0.87 | 0.84 | 0.93 | 0.88 | 10.1 | 9.9 | 7.2 | 9.1 |
| | SVR - RBF | 0.89 | 0.85 | 0.93 | 0.89 | 9.5 | 9.7 | 7.6 | 8.9 |
| | SVR - Sigmoid | 0.42 | 0.39 | 0.52 | 0.45 | 21.5 | 19.3 | 19.4 | 20.1 |
| | Gradient Boosting | 0.90 | 0.88 | 0.95 | 0.91 | 8.8 | 8.5 | 6.2 | 7.8 |
| | Random Forest | 0.93 | 0.90 | 0.95 | 0.93 | 7.4 | 7.7 | 5.9 | 7.0 |
| Fingerprints: MACCS | Linear Regression | 0.61 | 0.56 | 0.73 | 0.63 | 17.7 | 16.5 | 14.6 | 16.3 |
| | Lasso | 0.62 | 0.58 | 0.73 | 0.64 | 17.6 | 16.1 | 14.5 | 16.0 |
| | Ridge | 0.62 | 0.58 | 0.73 | 0.64 | 17.5 | 16.1 | 14.5 | 16.0 |
| | Bayesian Ridge | 0.62 | 0.58 | 0.73 | 0.64 | 17.6 | 16.1 | 14.5 | 16.0 |
| | SVR - Linear | 0.62 | 0.58 | 0.73 | 0.64 | 17.6 | 16.1 | 14.6 | 16.1 |
| | SVR - Polynomial | 0.85 | 0.82 | 0.91 | 0.86 | 10.9 | 10.4 | 8.3 | 9.9 |
| | SVR - RBF | 0.85 | 0.82 | 0.91 | 0.86 | 11.1 | 10.5 | 8.6 | 10.1 |
| | SVR - Sigmoid | 0.22 | 0.19 | 0.30 | 0.23 | 25.1 | 22.4 | 23.4 | 23.6 |
| | Gradient Boosting | 0.88 | 0.85 | 0.94 | 0.89 | 9.9 | 9.5 | 7.0 | 8.8 |
| | Random Forest | 0.93 | 0.90 | 0.94 | 0.92 | 7.5 | 7.7 | 6.9 | 7.4 |
| Fingerprints: Morgan1 | Linear Regression | 0.65 | 0.63 | 0.76 | 0.68 | 16.8 | 15.1 | 13.7 | 15.2 |
| | Lasso | 0.66 | 0.63 | 0.76 | 0.69 | 16.5 | 15.0 | 13.6 | 15.0 |
| | Ridge | 0.66 | 0.63 | 0.76 | 0.69 | 16.5 | 15.0 | 13.6 | 15.0 |
| | Bayesian Ridge | 0.66 | 0.63 | 0.76 | 0.69 | 16.6 | 15.0 | 13.6 | 15.0 |
| | SVR - Linear | 0.66 | 0.63 | 0.76 | 0.68 | 16.6 | 15.0 | 13.6 | 15.1 |
| | SVR - Polynomial | 0.90 | 0.88 | 0.95 | 0.91 | 8.8 | 8.5 | 6.4 | 7.9 |
| | SVR - RBF | 0.92 | 0.89 | 0.95 | 0.92 | 8.2 | 8.1 | 6.0 | 7.4 |
| | SVR - Sigmoid | 0.40 | 0.38 | 0.52 | 0.43 | 22.1 | 19.6 | 19.4 | 20.4 |
| | Gradient Boosting | 0.90 | 0.88 | 0.95 | 0.91 | 8.9 | 8.7 | 6.4 | 8.0 |
| | Random Forest | 0.94 | 0.91 | 0.95 | 0.93 | 7.0 | 7.5 | 6.3 | 6.9 |
| Fingerprints: RDK | Linear Regression | 0.68 | 0.64 | 0.78 | 0.70 | 16.1 | 14.8 | 13.0 | 14.6 |
| | Lasso | 0.68 | 0.66 | 0.79 | 0.71 | 16.0 | 14.5 | 12.8 | 14.4 |
| | Ridge | 0.68 | 0.66 | 0.79 | 0.71 | 16.0 | 14.5 | 12.8 | 14.4 |
| | Bayesian Ridge | 0.68 | 0.66 | 0.79 | 0.71 | 16.0 | 14.5 | 12.8 | 14.4 |
| | SVR - Linear | 0.68 | 0.65 | 0.79 | 0.71 | 16.0 | 14.6 | 12.9 | 14.5 |
| | SVR - Polynomial | 0.90 | 0.88 | 0.95 | 0.91 | 9.1 | 8.7 | 6.3 | 8.0 |
| | SVR - RBF | 0.90 | 0.88 | 0.95 | 0.91 | 8.9 | 8.6 | 6.2 | 7.9 |
| | SVR - Sigmoid | 0.33 | 0.15 | 0.59 | 0.36 | 23.2 | 22.9 | 17.8 | 21.3 |
| | Gradient Boosting | 0.89 | 0.86 | 0.95 | 0.90 | 9.3 | 9.4 | 6.3 | 8.3 |
| | Random Forest | 0.92 | 0.90 | 0.96 | 0.93 | 7.8 | 7.9 | 5.9 | 7.2 |
| Tanimoto: MACCS | Linear Regression | 0.88 | 0.86 | 0.90 | 0.88 | 10.0 | 9.1 | 8.7 | 9.3 |
| | Lasso | 0.90 | 0.88 | 0.93 | 0.90 | 8.9 | 8.5 | 7.4 | 8.3 |
| | Ridge | 0.90 | 0.89 | 0.94 | 0.91 | 8.7 | 8.2 | 6.9 | 7.9 |

**Table B.20** Grid Search Cross-Validated Performance for the Models in the Leave-One-Base-Out Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | BTMG | MTBD | P2Et | Mean | BTMG | MTBD | P2Et | Mean |
| | Bayesian Ridge | 0.88 | 0.86 | 0.93 | 0.89 | 9.7 | 9.2 | 7.1 | 8.7 |
| | SVR - Polynomial | 0.90 | 0.89 | 0.92 | 0.90 | 9.1 | 8.3 | 7.7 | 8.4 |
| | SVR - RBF | 0.88 | 0.87 | 0.92 | 0.89 | 9.7 | 8.9 | 8.0 | 8.9 |
| | SVR - Sigmoid | 0.66 | 0.57 | 0.74 | 0.66 | 16.6 | 16.2 | 14.1 | 15.6 |
| | SVR - Precomputed | 0.89 | 0.87 | 0.93 | 0.90 | 9.6 | 9.0 | 7.2 | 8.6 |
| | Gradient Boosting | 0.78 | 0.71 | 0.85 | 0.78 | 13.3 | 13.3 | 10.8 | 12.5 |
| | Random Forest | 0.71 | 0.60 | 0.79 | 0.70 | 15.3 | 15.7 | 12.7 | 14.6 |
| Tanimoto: Morgan1 | Linear Regression | 0.93 | 0.92 | 0.94 | 0.93 | 7.3 | 7.1 | 6.6 | 7.0 |
| | Lasso | 0.94 | 0.92 | 0.95 | 0.94 | 7.1 | 7.0 | 6.4 | 6.8 |
| | Ridge | 0.94 | 0.92 | 0.95 | 0.94 | 7.1 | 6.9 | 6.4 | 6.8 |
| | Bayesian Ridge | 0.94 | 0.92 | 0.95 | 0.94 | 7.1 | 6.9 | 6.3 | 6.8 |
| | SVR - Polynomial | 0.93 | 0.91 | 0.94 | 0.92 | 7.7 | 7.6 | 6.8 | 7.4 |
| | SVR - RBF | 0.91 | 0.89 | 0.93 | 0.91 | 8.6 | 8.3 | 7.5 | 8.1 |
| | SVR - Sigmoid | 0.87 | 0.84 | 0.91 | 0.88 | 10.1 | 9.9 | 8.2 | 9.4 |
| | SVR - Precomputed | 0.94 | 0.92 | 0.95 | 0.94 | 7.1 | 7.0 | 6.2 | 6.8 |
| | Gradient Boosting | 0.82 | 0.77 | 0.85 | 0.81 | 12.2 | 11.9 | 10.7 | 11.6 |
| | Random Forest | 0.76 | 0.68 | 0.77 | 0.74 | 13.9 | 13.9 | 13.3 | 13.7 |
| Tanimoto: RDK | Linear Regression | 0.90 | 0.87 | 0.93 | 0.90 | 9.1 | 9.0 | 7.3 | 8.5 |
| | Lasso | 0.90 | 0.88 | 0.94 | 0.91 | 8.8 | 8.7 | 7.0 | 8.2 |
| | Ridge | 0.91 | 0.88 | 0.94 | 0.91 | 8.7 | 8.5 | 6.9 | 8.0 |
| | Bayesian Ridge | 0.91 | 0.88 | 0.94 | 0.91 | 8.6 | 8.4 | 6.9 | 8.0 |
| | SVR - Polynomial | 0.89 | 0.87 | 0.92 | 0.90 | 9.3 | 8.9 | 7.7 | 8.6 |
| | SVR - RBF | 0.88 | 0.86 | 0.91 | 0.88 | 9.8 | 9.3 | 8.4 | 9.2 |
| | SVR - Sigmoid | 0.83 | 0.80 | 0.90 | 0.84 | 11.7 | 11.1 | 9.0 | 10.6 |
| | SVR - Precomputed | 0.90 | 0.88 | 0.94 | 0.91 | 8.8 | 8.7 | 6.9 | 8.1 |
| | Gradient Boosting | 0.77 | 0.75 | 0.82 | 0.78 | 13.5 | 12.3 | 11.8 | 12.5 |
| | Random Forest | 0.71 | 0.69 | 0.75 | 0.71 | 15.4 | 13.9 | 14.0 | 14.4 |
| WL | Linear Regression | 0.91 | 0.89 | 0.95 | 0.91 | 8.5 | 8.2 | 6.5 | 7.7 |
| | Lasso | 0.91 | 0.90 | 0.95 | 0.92 | 8.4 | 7.9 | 6.3 | 7.5 |
| | Ridge | 0.92 | 0.90 | 0.95 | 0.92 | 8.2 | 7.8 | 6.2 | 7.4 |
| | Bayesian Ridge | 0.92 | 0.90 | 0.95 | 0.92 | 8.2 | 7.8 | 6.3 | 7.4 |
| | SVR - Polynomial | 0.91 | 0.90 | 0.95 | 0.92 | 8.3 | 7.9 | 6.5 | 7.6 |
| | SVR - RBF | 0.91 | 0.89 | 0.94 | 0.92 | 8.5 | 8.1 | 6.7 | 7.8 |
| | SVR - Sigmoid | 0.82 | 0.76 | 0.88 | 0.82 | 12.1 | 12.1 | 9.7 | 11.3 |
| | SVR - Precomputed | 0.91 | 0.90 | 0.95 | 0.92 | 8.4 | 8.0 | 6.3 | 7.6 |
| | Gradient Boosting | 0.85 | 0.81 | 0.89 | 0.85 | 11.0 | 10.9 | 9.0 | 10.3 |
| | Random Forest | 0.79 | 0.72 | 0.84 | 0.78 | 13.2 | 13.0 | 11.3 | 12.5 |

### Training Set Performance

Table B.21: Training Set Performance for the Models in the Leave-One-Base-Out Test

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | BTMG | MTBD | P2Et | Mean | BTMG | MTBD | P2Et | Mean |
| One-hot | Linear Regression | 0.69 | 0.67 | 0.80 | 0.72 | 15.7 | 14.3 | 12.6 | 14.2 |
| | Lasso | 0.69 | 0.67 | 0.80 | 0.72 | 15.7 | 14.3 | 12.6 | 14.2 |
| | Ridge | 0.69 | 0.67 | 0.80 | 0.72 | 15.7 | 14.3 | 12.6 | 14.2 |
| | Bayesian Ridge | 0.69 | 0.67 | 0.80 | 0.72 | 15.7 | 14.3 | 12.6 | 14.2 |
| | SVR - Linear | 0.69 | 0.66 | 0.79 | 0.72 | 15.8 | 14.4 | 12.7 | 14.3 |
| | SVR - Polynomial | 0.97 | 0.96 | 0.99 | 0.97 | 4.9 | 4.7 | 3.1 | 4.2 |
| | SVR - RBF | 1.00 | 1.00 | 0.99 | 1.00 | 1.0 | 1.0 | 2.1 | 1.3 |
| | SVR - Sigmoid | 0.52 | 0.52 | 0.65 | 0.56 | 19.7 | 17.2 | 16.4 | 17.8 |
| | Gradient Boosting | 0.95 | 0.94 | 0.98 | 0.95 | 6.6 | 6.1 | 4.2 | 5.6 |
| | Random Forest | 0.99 | 0.98 | 0.99 | 0.99 | 3.2 | 3.2 | 2.7 | 3.0 |

**Table B.21** Training Set Performance for the Models in the Leave-One-Base-Out Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | BTMG | MTBD | P2Et | Mean | BTMG | MTBD | P2Et | Mean |
| Quantum Chemical | Linear Regression | 0.69 | 0.66 | 0.80 | 0.72 | 15.8 | 14.4 | 12.6 | 14.3 |
| | Lasso | 0.69 | 0.66 | 0.79 | 0.71 | 15.9 | 14.5 | 12.9 | 14.5 |
| | Ridge | 0.69 | 0.66 | 0.80 | 0.72 | 15.8 | 14.4 | 12.6 | 14.3 |
| | Bayesian Ridge | 0.69 | 0.66 | 0.79 | 0.71 | 15.9 | 14.5 | 12.8 | 14.4 |
| | SVR - Linear | 0.69 | 0.66 | 0.79 | 0.71 | 15.8 | 14.4 | 12.8 | 14.4 |
| | SVR - Polynomial | 0.95 | 0.94 | 0.98 | 0.96 | 6.1 | 6.0 | 4.4 | 5.5 |
| | SVR - RBF | 0.97 | 0.97 | 0.99 | 0.98 | 5.0 | 4.2 | 3.3 | 4.2 |
| | SVR - Sigmoid | 0.42 | 0.38 | 0.51 | 0.44 | 21.7 | 19.5 | 19.5 | 20.2 |
| | Gradient Boosting | 0.97 | 0.95 | 0.98 | 0.97 | 5.2 | 5.4 | 3.9 | 4.9 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.5 | 2.5 | 2.1 | 2.4 |
| Fingerprints: MACCS | Linear Regression | 0.62 | 0.58 | 0.72 | 0.64 | 17.4 | 16.0 | 14.7 | 16.0 |
| | Lasso | 0.63 | 0.59 | 0.74 | 0.65 | 17.3 | 15.9 | 14.3 | 15.9 |
| | Ridge | 0.63 | 0.59 | 0.74 | 0.65 | 17.3 | 15.9 | 14.3 | 15.9 |
| | Bayesian Ridge | 0.63 | 0.59 | 0.74 | 0.65 | 17.3 | 15.9 | 14.3 | 15.9 |
| | SVR - Linear | 0.63 | 0.59 | 0.74 | 0.65 | 17.4 | 15.9 | 14.4 | 15.9 |
| | SVR - Polynomial | 0.89 | 0.87 | 0.94 | 0.90 | 9.2 | 8.9 | 6.7 | 8.3 |
| | SVR - RBF | 0.89 | 0.87 | 0.94 | 0.90 | 9.3 | 8.9 | 6.8 | 8.4 |
| | SVR - Sigmoid | 0.25 | 0.22 | 0.33 | 0.27 | 24.7 | 21.9 | 22.8 | 23.1 |
| | Gradient Boosting | 0.92 | 0.90 | 0.96 | 0.93 | 8.2 | 7.9 | 5.6 | 7.2 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.6 | 2.5 | 2.3 | 2.5 |
| Fingerprints: Morgan1 | Linear Regression | 0.65 | 0.64 | 0.77 | 0.68 | 16.9 | 14.9 | 13.5 | 15.1 |
| | Lasso | 0.67 | 0.64 | 0.77 | 0.70 | 16.3 | 14.8 | 13.4 | 14.8 |
| | Ridge | 0.67 | 0.64 | 0.77 | 0.70 | 16.3 | 14.8 | 13.4 | 14.8 |
| | Bayesian Ridge | 0.67 | 0.64 | 0.77 | 0.70 | 16.3 | 14.8 | 13.4 | 14.8 |
| | SVR - Linear | 0.67 | 0.64 | 0.77 | 0.69 | 16.4 | 14.8 | 13.4 | 14.9 |
| | SVR - Polynomial | 0.95 | 0.94 | 0.98 | 0.96 | 6.3 | 6.2 | 4.2 | 5.6 |
| | SVR - RBF | 0.97 | 0.96 | 0.98 | 0.97 | 5.2 | 5.3 | 3.4 | 4.6 |
| | SVR - Sigmoid | 0.43 | 0.41 | 0.55 | 0.46 | 21.4 | 19.0 | 18.8 | 19.7 |
| | Gradient Boosting | 0.95 | 0.94 | 0.97 | 0.95 | 6.5 | 6.3 | 4.6 | 5.8 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.4 | 2.5 | 2.2 | 2.4 |
| Fingerprints: RDK | Linear Regression | 0.69 | 0.66 | 0.79 | 0.71 | 15.9 | 14.4 | 12.7 | 14.3 |
| | Lasso | 0.69 | 0.67 | 0.80 | 0.72 | 15.7 | 14.3 | 12.6 | 14.2 |
| | Ridge | 0.69 | 0.67 | 0.80 | 0.72 | 15.7 | 14.3 | 12.6 | 14.2 |
| | Bayesian Ridge | 0.69 | 0.67 | 0.80 | 0.72 | 15.7 | 14.4 | 12.6 | 14.2 |
| | SVR - Linear | 0.69 | 0.66 | 0.79 | 0.72 | 15.8 | 14.4 | 12.7 | 14.3 |
| | SVR - Polynomial | 0.96 | 0.96 | 0.98 | 0.97 | 5.4 | 5.0 | 3.5 | 4.7 |
| | SVR - RBF | 0.97 | 0.97 | 0.99 | 0.97 | 5.2 | 4.4 | 3.1 | 4.3 |
| | SVR - Sigmoid | -0.07 | 0.18 | 0.57 | 0.23 | 29.4 | 22.4 | 18.4 | 23.4 |
| | Gradient Boosting | 0.96 | 0.95 | 0.98 | 0.96 | 6.0 | 5.5 | 4.1 | 5.2 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.6 | 2.6 | 2.0 | 2.4 |
| Tanimoto: MACCS | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 0.99 | 0.99 | 0.99 | 0.99 | 2.6 | 2.6 | 2.5 | 2.6 |
| | Ridge | 1.00 | 1.00 | 0.99 | 0.99 | 1.7 | 1.6 | 3.0 | 2.1 |
| | Bayesian Ridge | 0.96 | 0.95 | 0.98 | 0.96 | 5.7 | 5.3 | 3.9 | 5.0 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.7 | 1.2 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.0 | 1.0 |
| | SVR - Sigmoid | 0.69 | 0.60 | 0.77 | 0.69 | 15.8 | 15.6 | 13.4 | 14.9 |
| | SVR - Precomputed | 0.97 | 0.97 | 0.99 | 0.98 | 5.1 | 4.4 | 3.0 | 4.2 |
| | Gradient Boosting | 0.99 | 0.99 | 1.00 | 0.99 | 2.2 | 2.1 | 1.8 | 2.0 |
| | Random Forest | 0.98 | 0.98 | 0.99 | 0.98 | 3.8 | 3.7 | 3.3 | 3.6 |
| Tanimoto: Morgan1 | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 1.0 | 0.9 | 1.0 |
| | Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 0.8 | 0.8 | 0.7 | 0.7 |
| | Bayesian Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 1.1 | 1.1 | 1.2 | 1.1 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 0.9 | 1.2 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.0 | 1.0 |
| | SVR - Sigmoid | 0.87 | 0.80 | 0.91 | 0.86 | 10.4 | 11.1 | 8.4 | 10.0 |
| | SVR - Precomputed | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 0.9 | 1.9 | 1.3 |
| | Gradient Boosting | 1.00 | 1.00 | 1.00 | 1.00 | 0.8 | 0.6 | 0.5 | 0.6 |

**Table B.21** Training Set Performance for the Models in the Leave-One-Base-Out Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | BTMG | MTBD | P2Et | Mean | BTMG | MTBD | P2Et | Mean |
| | Random Forest | 0.98 | 0.97 | 0.98 | 0.98 | 3.9 | 4.0 | 3.8 | 3.9 |
| Tanimoto: RDK | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 1.5 | 1.4 | 1.5 | 1.5 |
| | Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 1.7 | 1.6 | 1.4 | 1.6 |
| | Bayesian Ridge | 0.99 | 0.99 | 0.99 | 0.99 | 2.5 | 2.4 | 2.1 | 2.3 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.0 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 0.9 | 1.0 | 1.0 |
| | SVR - Sigmoid | 0.83 | 0.83 | 0.86 | 0.84 | 11.9 | 10.3 | 10.4 | 10.8 |
| | SVR - Precomputed | 1.00 | 1.00 | 1.00 | 1.00 | 1.1 | 1.2 | 1.2 | 1.2 |
| | Gradient Boosting | 0.99 | 0.99 | 1.00 | 0.99 | 2.7 | 2.4 | 0.7 | 1.9 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 4.3 | 3.9 | 3.8 | 4.0 |
| WL | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 1.7 | 1.7 | 1.6 | 1.6 |
| | Ridge | 0.99 | 0.99 | 1.00 | 0.99 | 2.1 | 2.0 | 1.7 | 1.9 |
| | Bayesian Ridge | 0.98 | 0.98 | 0.99 | 0.99 | 3.5 | 3.2 | 2.5 | 3.1 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 0.9 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.5 | 1.1 |
| | SVR - Sigmoid | 0.81 | 0.76 | 0.89 | 0.82 | 12.4 | 12.2 | 9.4 | 11.3 |
| | SVR - Precomputed | 1.00 | 1.00 | 1.00 | 1.00 | 1.7 | 1.6 | 1.3 | 1.5 |
| | Gradient Boosting | 1.00 | 1.00 | 1.00 | 1.00 | 0.8 | 0.7 | 1.7 | 1.1 |
| | Random Forest | 0.99 | 0.98 | 0.99 | 0.99 | 3.4 | 3.3 | 3.2 | 3.3 |

### Test Set Performance

Table B.22: Test Set Performance for the Models in the Leave-One-Base-Out Test

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | BTMG | MTBD | P2Et | Mean | BTMG | MTBD | P2Et | Mean |
| One-hot | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | 0.68 | 0.44 | 0.01 | 0.38 | 13.9 | 22.4 | 24.7 | 20.3 |
| | Ridge | 0.69 | 0.50 | 0.22 | 0.47 | 13.7 | 21.2 | 22.0 | 19.0 |
| | Bayesian Ridge | 0.69 | 0.50 | 0.22 | 0.47 | 13.7 | 21.2 | 21.9 | 19.0 |
| | SVR - Linear | 0.69 | 0.50 | 0.20 | 0.46 | 13.8 | 21.1 | 22.2 | 19.0 |
| | SVR - Polynomial | 0.55 | 0.28 | 0.29 | 0.38 | 16.5 | 25.3 | 20.9 | 20.9 |
| | SVR - RBF | 0.76 | 0.57 | 0.26 | 0.53 | 12.1 | 19.7 | 21.3 | 17.7 |
| | SVR - Sigmoid | 0.64 | 0.35 | 0.31 | 0.43 | 14.9 | 24.1 | 20.6 | 19.9 |
| | Gradient Boosting | 0.73 | 0.58 | 0.13 | 0.48 | 12.8 | 19.5 | 23.2 | 18.5 |
| | Random Forest | 0.76 | 0.58 | 0.24 | 0.53 | 12.1 | 19.5 | 21.6 | 17.7 |
| Quantum Chemical | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | 0.62 | -0.59 | 0.31 | 0.11 | 15.2 | 37.8 | 20.7 | 24.6 |
| | Ridge | -0.35 | 0.46 | 0.24 | 0.12 | 28.7 | 21.9 | 21.7 | 24.1 |
| | Bayesian Ridge | -0.37 | 0.66 | 0.24 | 0.18 | 28.9 | 17.4 | 21.6 | 22.6 |
| | SVR - Linear | -0.36 | <-1.00 | 0.20 | <-1.00 | 28.7 | >100.0 | 22.3 | >100.0 |
| | SVR - Polynomial | -2.43 | <-1.00 | -0.52 | <-1.00 | 45.7 | >100.0 | 30.6 | >100.0 |
| | SVR - RBF | 0.15 | -0.49 | -0.55 | -0.30 | 22.7 | 36.5 | 30.9 | 30.1 |
| | SVR - Sigmoid | -0.46 | -0.22 | 0.10 | -0.19 | 29.8 | 33.1 | 23.5 | 28.8 |
| | Gradient Boosting | 0.74 | 0.60 | 0.21 | 0.52 | 12.6 | 19.1 | 22.0 | 17.9 |
| | Random Forest | 0.77 | 0.57 | 0.30 | 0.54 | 11.9 | 19.7 | 20.8 | 17.5 |
| Fingerprints: MACCS | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | 0.47 | 0.50 | -0.06 | 0.30 | 18.0 | 21.2 | 25.5 | 21.6 |
| | Ridge | 0.65 | 0.47 | 0.23 | 0.45 | 14.6 | 21.9 | 21.8 | 19.4 |
| | Bayesian Ridge | 0.65 | 0.47 | 0.23 | 0.45 | 14.6 | 21.9 | 21.7 | 19.4 |
| | SVR - Linear | 0.64 | 0.46 | 0.22 | 0.44 | 14.7 | 22.0 | 22.0 | 19.6 |
| | SVR - Polynomial | 0.71 | 0.62 | 0.39 | 0.57 | 13.3 | 18.4 | 19.4 | 17.0 |

**Table B.22** Test Set Performance for the Models in the Leave-One-Base-Out Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | BTMG | MTBD | P2Et | Mean | BTMG | MTBD | P2Et | Mean |
| | SVR - RBF | 0.74 | 0.44 | 0.44 | 0.54 | 12.6 | 22.4 | 18.5 | 17.8 |
| | SVR - Sigmoid | 0.29 | 0.00 | 0.17 | 0.15 | 20.8 | 29.9 | 22.6 | 24.4 |
| | Gradient Boosting | 0.71 | 0.55 | 0.35 | 0.54 | 13.2 | 20.0 | 20.0 | 17.8 |
| | Random Forest | 0.72 | 0.65 | 0.36 | 0.58 | 13.0 | 17.7 | 19.8 | 16.9 |
| Fingerprints: Morgan1 | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | 0.67 | 0.53 | 0.32 | 0.50 | 14.2 | 20.6 | 20.5 | 18.4 |
| | Ridge | 0.69 | 0.48 | 0.26 | 0.48 | 13.8 | 21.5 | 21.3 | 18.9 |
| | Bayesian Ridge | 0.69 | 0.49 | 0.26 | 0.48 | 13.8 | 21.5 | 21.3 | 18.9 |
| | SVR - Linear | 0.69 | 0.48 | 0.25 | 0.47 | 13.8 | 21.6 | 21.5 | 19.0 |
| | SVR - Polynomial | 0.72 | 0.52 | 0.37 | 0.54 | 13.0 | 20.9 | 19.7 | 17.8 |
| | SVR - RBF | 0.76 | 0.54 | 0.41 | 0.57 | 12.2 | 20.4 | 19.0 | 17.2 |
| | SVR - Sigmoid | 0.52 | 0.19 | 0.30 | 0.34 | 17.2 | 26.9 | 20.7 | 21.6 |
| | Gradient Boosting | 0.69 | 0.58 | 0.34 | 0.54 | 13.6 | 19.3 | 20.2 | 17.7 |
| | Random Forest | 0.71 | 0.54 | 0.34 | 0.53 | 13.2 | 20.3 | 20.2 | 17.9 |
| Fingerprints: RDK | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | 0.52 | 0.55 | 0.01 | 0.36 | 17.0 | 20.1 | 24.6 | 20.6 |
| | Ridge | 0.62 | 0.52 | 0.20 | 0.45 | 15.1 | 20.7 | 22.2 | 19.3 |
| | Bayesian Ridge | 0.62 | 0.52 | 0.20 | 0.45 | 15.1 | 20.7 | 22.2 | 19.3 |
| | SVR - Linear | 0.62 | 0.53 | 0.18 | 0.44 | 15.1 | 20.6 | 22.5 | 19.4 |
| | SVR - Polynomial | 0.65 | 0.68 | 0.19 | 0.51 | 14.5 | 17.0 | 22.3 | 17.9 |
| | SVR - RBF | 0.75 | 0.63 | 0.32 | 0.57 | 12.4 | 18.1 | 20.5 | 17.0 |
| | SVR - Sigmoid | 0.43 | -0.09 | 0.26 | 0.20 | 18.6 | 31.3 | 21.3 | 23.8 |
| | Gradient Boosting | 0.62 | 0.68 | 0.14 | 0.48 | 15.2 | 16.8 | 23.0 | 18.4 |
| | Random Forest | 0.77 | 0.66 | 0.18 | 0.54 | 11.9 | 17.4 | 22.4 | 17.2 |
| Tanimoto: MACCS | Linear Regression | 0.66 | 0.33 | 0.39 | 0.46 | 14.4 | 24.4 | 19.3 | 19.4 |
| | Lasso | 0.66 | 0.34 | 0.40 | 0.47 | 14.3 | 24.4 | 19.3 | 19.3 |
| | Ridge | 0.67 | 0.20 | 0.47 | 0.45 | 14.2 | 26.9 | 18.0 | 19.7 |
| | Bayesian Ridge | 0.67 | 0.20 | 0.47 | 0.45 | 14.1 | 26.8 | 18.0 | 19.6 |
| | SVR - Polynomial | 0.59 | 0.07 | 0.44 | 0.37 | 15.7 | 28.8 | 18.6 | 21.1 |
| | SVR - RBF | 0.55 | 0.04 | 0.41 | 0.33 | 16.5 | 29.3 | 19.1 | 21.6 |
| | SVR - Sigmoid | 0.60 | 0.18 | 0.36 | 0.38 | 15.7 | 27.2 | 19.8 | 20.9 |
| | SVR - Precomputed | 0.68 | 0.19 | 0.47 | 0.45 | 14.0 | 27.0 | 18.0 | 19.7 |
| | Gradient Boosting | 0.58 | 0.28 | 0.42 | 0.43 | 15.9 | 25.4 | 18.9 | 20.1 |
| | Random Forest | 0.39 | 0.30 | 0.16 | 0.28 | 19.2 | 25.1 | 22.8 | 22.4 |
| Tanimoto: Morgan1 | Linear Regression | 0.41 | 0.17 | 0.27 | 0.28 | 19.0 | 27.3 | 21.2 | 22.5 |
| | Lasso | 0.40 | 0.01 | 0.23 | 0.21 | 19.1 | 29.8 | 21.7 | 23.5 |
| | Ridge | 0.41 | -0.04 | 0.28 | 0.22 | 19.0 | 30.5 | 21.0 | 23.5 |
| | Bayesian Ridge | 0.41 | -0.04 | 0.28 | 0.22 | 19.0 | 30.5 | 21.0 | 23.5 |
| | SVR - Polynomial | 0.32 | -0.08 | 0.20 | 0.15 | 20.3 | 31.1 | 22.2 | 24.5 |
| | SVR - RBF | 0.30 | -0.09 | 0.18 | 0.13 | 20.7 | 31.2 | 22.5 | 24.8 |
| | SVR - Sigmoid | 0.49 | 0.01 | 0.33 | 0.28 | 17.7 | 29.9 | 20.3 | 22.6 |
| | SVR - Precomputed | 0.41 | -0.04 | 0.28 | 0.22 | 18.9 | 30.5 | 21.0 | 23.5 |
| | Gradient Boosting | 0.10 | 0.01 | -0.06 | 0.02 | 23.4 | 29.8 | 25.6 | 26.3 |
| | Random Forest | -0.07 | 0.00 | -0.16 | -0.07 | 25.5 | 30.0 | 26.7 | 27.4 |
| Tanimoto: RDK | Linear Regression | 0.63 | 0.44 | 0.17 | 0.41 | 14.9 | 22.5 | 22.6 | 20.0 |
| | Lasso | 0.62 | 0.44 | 0.23 | 0.43 | 15.2 | 22.5 | 21.8 | 19.8 |
| | Ridge | 0.57 | 0.46 | 0.07 | 0.37 | 16.1 | 22.1 | 23.9 | 20.7 |
| | Bayesian Ridge | 0.57 | 0.45 | 0.07 | 0.37 | 16.1 | 22.1 | 23.9 | 20.7 |
| | SVR - Polynomial | 0.47 | 0.36 | 0.01 | 0.28 | 18.0 | 24.0 | 24.7 | 22.2 |
| | SVR - RBF | 0.42 | 0.32 | 0.00 | 0.25 | 18.7 | 24.7 | 24.9 | 22.8 |
| | SVR - Sigmoid | 0.61 | 0.49 | 0.10 | 0.40 | 15.4 | 21.4 | 23.5 | 20.1 |
| | SVR - Precomputed | 0.57 | 0.46 | 0.07 | 0.37 | 16.1 | 22.1 | 24.0 | 20.7 |
| | Gradient Boosting | 0.54 | 0.33 | 0.32 | 0.40 | 16.7 | 24.5 | 20.5 | 20.6 |
| | Random Forest | 0.49 | 0.41 | 0.22 | 0.37 | 17.7 | 23.1 | 21.9 | 20.9 |
| WL | Linear Regression | 0.73 | 0.47 | 0.42 | 0.54 | 12.8 | 21.8 | 18.9 | 17.8 |
| | Lasso | 0.73 | 0.47 | 0.35 | 0.52 | 12.8 | 21.9 | 20.0 | 18.2 |
| | Ridge | 0.76 | 0.40 | 0.40 | 0.52 | 12.2 | 23.3 | 19.2 | 18.2 |
| | Bayesian Ridge | 0.75 | 0.39 | 0.40 | 0.52 | 12.2 | 23.3 | 19.2 | 18.3 |
| | SVR - Polynomial | 0.66 | 0.27 | 0.34 | 0.42 | 14.4 | 25.5 | 20.2 | 20.1 |

**Table B.22** Test Set Performance for the Models in the Leave-One-Base-Out Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | BTMG | MTBD | P2Et | Mean | BTMG | MTBD | P2Et | Mean |
| | SVR - RBF | 0.61 | 0.23 | 0.31 | 0.38 | 15.4 | 26.2 | 20.7 | 20.8 |
| | SVR - Sigmoid | 0.79 | 0.48 | 0.41 | 0.56 | 11.4 | 21.6 | 19.0 | 17.3 |
| | SVR - Precomputed | 0.76 | 0.40 | 0.40 | 0.52 | 12.2 | 23.3 | 19.2 | 18.2 |
| | Gradient Boosting | 0.43 | 0.30 | -0.16 | 0.19 | 18.7 | 25.0 | 26.8 | 23.5 |
| | Random Forest | 0.27 | 0.29 | -0.53 | 0.01 | 21.1 | 25.2 | 30.7 | 25.7 |

## B.7.5 Leave-One-Ligand-Out Test

### Grid Search Cross-Validated Performance

Table B.23: Grid Search Cross-Validated Performance for the Models in the Leave-One-Ligand-Out Test

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ABP | XP | $^t$BBP | $^t$BXP | Mean | ABP | XP | $^t$BBP | $^t$BXP | Mean |
| One-hot | Linear Regression | 0.69 | 0.75 | 0.67 | 0.68 | 0.70 | 14.9 | 14.4 | 15.4 | 15.0 | 14.9 |
| | Lasso | 0.69 | 0.75 | 0.68 | 0.68 | 0.70 | 14.8 | 14.3 | 15.0 | 15.0 | 14.8 |
| | Ridge | 0.69 | 0.75 | 0.68 | 0.68 | 0.70 | 14.8 | 14.3 | 15.0 | 15.0 | 14.8 |
| | Bayesian Ridge | 0.69 | 0.75 | 0.68 | 0.68 | 0.70 | 14.8 | 14.3 | 15.0 | 15.0 | 14.8 |
| | SVR - Linear | 0.69 | 0.74 | 0.68 | 0.68 | 0.70 | 14.9 | 14.4 | 15.1 | 15.0 | 14.9 |
| | SVR - Polynomial | 0.90 | 0.92 | 0.88 | 0.89 | 0.90 | 8.6 | 7.8 | 9.3 | 8.9 | 8.7 |
| | SVR - RBF | 0.91 | 0.93 | 0.89 | 0.90 | 0.90 | 8.2 | 7.5 | 8.9 | 8.5 | 8.3 |
| | SVR - Sigmoid | 0.55 | 0.65 | 0.53 | 0.52 | 0.57 | 18.0 | 16.8 | 18.1 | 18.3 | 17.8 |
| | Gradient Boosting | 0.89 | 0.92 | 0.87 | 0.88 | 0.89 | 8.9 | 8.1 | 9.5 | 9.2 | 8.9 |
| | Random Forest | 0.88 | 0.91 | 0.85 | 0.88 | 0.88 | 9.2 | 8.6 | 10.2 | 9.0 | 9.3 |
| Quantum Chemical | Linear Regression | 0.69 | 0.75 | 0.68 | 0.68 | 0.70 | 14.9 | 14.4 | 15.0 | 15.0 | 14.8 |
| | Lasso | 0.69 | 0.74 | 0.67 | 0.67 | 0.69 | 15.0 | 14.6 | 15.2 | 15.1 | 15.0 |
| | Ridge | 0.69 | 0.75 | 0.68 | 0.68 | 0.70 | 14.9 | 14.3 | 15.0 | 15.0 | 14.8 |
| | Bayesian Ridge | 0.69 | 0.74 | 0.67 | 0.67 | 0.69 | 15.0 | 14.4 | 15.2 | 15.2 | 15.0 |
| | SVR - Linear | 0.69 | 0.74 | 0.68 | 0.67 | 0.70 | 14.9 | 14.4 | 15.1 | 15.1 | 14.9 |
| | SVR - Polynomial | 0.89 | 0.91 | 0.87 | 0.89 | 0.89 | 9.1 | 8.7 | 9.4 | 8.9 | 9.0 |
| | SVR - RBF | 0.88 | 0.89 | 0.88 | 0.89 | 0.89 | 9.2 | 9.2 | 9.3 | 8.6 | 9.1 |
| | SVR - Sigmoid | 0.43 | 0.50 | 0.43 | 0.46 | 0.46 | 20.2 | 20.1 | 20.1 | 19.4 | 20.0 |
| | Gradient Boosting | 0.91 | 0.93 | 0.90 | 0.91 | 0.91 | 8.0 | 7.3 | 8.6 | 8.0 | 8.0 |
| | Random Forest | 0.92 | 0.94 | 0.90 | 0.93 | 0.92 | 7.8 | 6.9 | 8.2 | 7.2 | 7.5 |
| Fingerprints: MACCS | Linear Regression | 0.63 | 0.65 | 0.61 | 0.54 | 0.61 | 16.4 | 16.8 | 16.5 | 17.9 | 16.9 |
| | Lasso | 0.64 | 0.66 | 0.62 | 0.62 | 0.64 | 16.2 | 16.5 | 16.3 | 16.3 | 16.3 |
| | Ridge | 0.64 | 0.66 | 0.62 | 0.62 | 0.64 | 16.2 | 16.5 | 16.3 | 16.3 | 16.3 |
| | Bayesian Ridge | 0.64 | 0.66 | 0.62 | 0.62 | 0.64 | 16.2 | 16.5 | 16.3 | 16.3 | 16.3 |
| | SVR - Linear | 0.63 | 0.66 | 0.62 | 0.62 | 0.63 | 16.2 | 16.6 | 16.4 | 16.3 | 16.4 |
| | SVR - Polynomial | 0.86 | 0.88 | 0.85 | 0.86 | 0.86 | 10.1 | 9.9 | 10.3 | 10.0 | 10.1 |
| | SVR - RBF | 0.85 | 0.87 | 0.84 | 0.85 | 0.86 | 10.2 | 10.2 | 10.5 | 10.2 | 10.3 |
| | SVR - Sigmoid | 0.29 | 0.30 | 0.27 | 0.25 | 0.28 | 22.6 | 23.8 | 22.8 | 23.0 | 23.0 |
| | Gradient Boosting | 0.88 | 0.90 | 0.87 | 0.88 | 0.88 | 9.1 | 8.8 | 9.6 | 9.1 | 9.2 |
| | Random Forest | 0.92 | 0.94 | 0.89 | 0.91 | 0.92 | 7.7 | 7.2 | 8.7 | 7.7 | 7.8 |
| Fingerprints: Morgan1 | Linear Regression | 0.67 | 0.71 | 0.66 | 0.65 | 0.67 | 15.5 | 15.3 | 15.6 | 15.6 | 15.5 |
| | Lasso | 0.67 | 0.72 | 0.66 | 0.66 | 0.68 | 15.3 | 15.1 | 15.5 | 15.4 | 15.3 |
| | Ridge | 0.67 | 0.72 | 0.66 | 0.66 | 0.68 | 15.3 | 15.1 | 15.5 | 15.4 | 15.3 |
| | Bayesian Ridge | 0.67 | 0.72 | 0.66 | 0.66 | 0.68 | 15.3 | 15.1 | 15.5 | 15.5 | 15.3 |
| | SVR - Linear | 0.67 | 0.72 | 0.66 | 0.66 | 0.68 | 15.4 | 15.1 | 15.6 | 15.5 | 15.4 |
| | SVR - Polynomial | 0.91 | 0.92 | 0.90 | 0.90 | 0.91 | 8.2 | 7.9 | 8.6 | 8.3 | 8.2 |
| | SVR - RBF | 0.92 | 0.93 | 0.91 | 0.92 | 0.92 | 7.6 | 7.6 | 8.1 | 7.6 | 7.7 |
| | SVR - Sigmoid | 0.46 | 0.50 | 0.45 | 0.43 | 0.46 | 19.7 | 20.1 | 19.8 | 20.0 | 19.9 |
| | Gradient Boosting | 0.90 | 0.92 | 0.89 | 0.90 | 0.90 | 8.4 | 7.8 | 8.8 | 8.6 | 8.4 |
| | Random Forest | 0.92 | 0.94 | 0.90 | 0.93 | 0.92 | 7.5 | 7.0 | 8.3 | 7.2 | 7.5 |
| Fingerprints: RDK | Linear Regression | 0.68 | 0.74 | 0.68 | 0.67 | 0.69 | 15.1 | 14.4 | 15.1 | 15.2 | 15.0 |
| | Lasso | 0.69 | 0.75 | 0.68 | 0.68 | 0.70 | 14.8 | 14.3 | 15.0 | 15.0 | 14.8 |
| | Ridge | 0.69 | 0.75 | 0.68 | 0.68 | 0.70 | 14.8 | 14.3 | 15.0 | 15.0 | 14.8 |
| | Bayesian Ridge | 0.69 | 0.75 | 0.68 | 0.68 | 0.70 | 14.8 | 14.3 | 15.0 | 15.0 | 14.8 |
| | SVR - Linear | 0.69 | 0.74 | 0.68 | 0.68 | 0.70 | 14.9 | 14.4 | 15.1 | 15.1 | 14.9 |
| | SVR - Polynomial | 0.90 | 0.93 | 0.88 | 0.90 | 0.90 | 8.6 | 7.7 | 9.1 | 8.6 | 8.5 |
| | SVR - RBF | 0.90 | 0.93 | 0.89 | 0.90 | 0.91 | 8.4 | 7.6 | 8.9 | 8.2 | 8.3 |
| | SVR - Sigmoid | 0.40 | 0.48 | 0.34 | 0.37 | 0.40 | 20.7 | 20.5 | 21.5 | 21.0 | 21.0 |
| | Gradient Boosting | 0.89 | 0.92 | 0.88 | 0.89 | 0.90 | 8.7 | 8.0 | 9.3 | 8.8 | 8.7 |
| | Random Forest | 0.91 | 0.93 | 0.90 | 0.92 | 0.92 | 7.8 | 7.3 | 8.5 | 7.5 | 7.8 |
| Tanimoto: MACCS | Linear Regression | 0.86 | 0.90 | 0.84 | 0.88 | 0.87 | 10.1 | 9.1 | 10.6 | 9.2 | 9.7 |
| | Lasso | 0.89 | 0.91 | 0.87 | 0.90 | 0.89 | 9.0 | 8.6 | 9.6 | 8.2 | 8.9 |
| | Ridge | 0.89 | 0.92 | 0.88 | 0.91 | 0.90 | 8.9 | 8.2 | 9.2 | 8.0 | 8.6 |

**Table B.23** Grid Search Cross-Validated Performance for the Models in the Leave-One-Ligand-Out Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ABP | XP | $^t$BBP | $^t$BXP | Mean | ABP | XP | $^t$BBP | $^t$BXP | Mean |
| | Bayesian Ridge | 0.88 | 0.90 | 0.87 | 0.90 | 0.89 | 9.3 | 8.9 | 9.6 | 8.5 | 9.1 |
| | SVR - Polynomial | 0.88 | 0.91 | 0.86 | 0.91 | 0.89 | 9.4 | 8.5 | 10.0 | 8.0 | 9.0 |
| | SVR - RBF | 0.86 | 0.90 | 0.84 | 0.90 | 0.88 | 10.0 | 9.1 | 10.5 | 8.3 | 9.5 |
| | SVR - Sigmoid | 0.68 | 0.67 | 0.66 | 0.67 | 0.67 | 15.3 | 16.3 | 15.6 | 15.3 | 15.6 |
| | SVR - Precomputed | 0.88 | 0.90 | 0.87 | 0.90 | 0.89 | 9.2 | 8.9 | 9.5 | 8.4 | 9.0 |
| | Gradient Boosting | 0.77 | 0.81 | 0.73 | 0.77 | 0.77 | 13.0 | 12.5 | 13.9 | 12.7 | 13.0 |
| | Random Forest | 0.71 | 0.74 | 0.67 | 0.72 | 0.71 | 14.4 | 14.5 | 15.4 | 14.1 | 14.6 |
| Tanimoto: Morgan1 | Linear Regression | 0.93 | 0.94 | 0.91 | 0.93 | 0.93 | 7.2 | 6.9 | 7.9 | 6.8 | 7.2 |
| | Lasso | 0.93 | 0.94 | 0.92 | 0.94 | 0.93 | 7.1 | 6.8 | 7.7 | 6.7 | 7.1 |
| | Ridge | 0.93 | 0.94 | 0.92 | 0.94 | 0.93 | 7.1 | 6.7 | 7.7 | 6.6 | 7.0 |
| | Bayesian Ridge | 0.93 | 0.94 | 0.92 | 0.94 | 0.93 | 7.0 | 6.7 | 7.6 | 6.6 | 7.0 |
| | SVR - Polynomial | 0.92 | 0.93 | 0.90 | 0.92 | 0.92 | 7.6 | 7.6 | 8.4 | 7.4 | 7.7 |
| | SVR - RBF | 0.90 | 0.91 | 0.88 | 0.90 | 0.90 | 8.3 | 8.4 | 9.2 | 8.2 | 8.5 |
| | SVR - Sigmoid | 0.88 | 0.89 | 0.86 | 0.88 | 0.88 | 9.5 | 9.3 | 9.8 | 9.3 | 9.5 |
| | SVR - Precomputed | 0.93 | 0.94 | 0.92 | 0.94 | 0.93 | 7.1 | 6.8 | 7.7 | 6.7 | 7.1 |
| | Gradient Boosting | 0.81 | 0.81 | 0.77 | 0.81 | 0.80 | 11.8 | 12.3 | 12.7 | 11.6 | 12.1 |
| | Random Forest | 0.72 | 0.74 | 0.68 | 0.73 | 0.72 | 14.3 | 14.5 | 15.1 | 13.9 | 14.4 |
| Tanimoto: RDK | Linear Regression | 0.88 | 0.94 | 0.86 | 0.90 | 0.90 | 9.2 | 7.1 | 9.9 | 8.4 | 8.6 |
| | Lasso | 0.89 | 0.94 | 0.87 | 0.91 | 0.90 | 9.0 | 7.0 | 9.6 | 8.1 | 8.4 |
| | Ridge | 0.89 | 0.94 | 0.88 | 0.91 | 0.90 | 8.9 | 6.9 | 9.4 | 8.0 | 8.3 |
| | Bayesian Ridge | 0.89 | 0.94 | 0.88 | 0.91 | 0.91 | 8.8 | 6.9 | 9.3 | 7.9 | 8.2 |
| | SVR - Polynomial | 0.87 | 0.94 | 0.86 | 0.91 | 0.89 | 9.5 | 7.2 | 10.1 | 8.1 | 8.7 |
| | SVR - RBF | 0.86 | 0.93 | 0.84 | 0.90 | 0.88 | 10.2 | 7.5 | 10.7 | 8.3 | 9.2 |
| | SVR - Sigmoid | 0.86 | 0.88 | 0.84 | 0.85 | 0.86 | 10.0 | 9.8 | 10.5 | 10.2 | 10.1 |
| | SVR - Precomputed | 0.89 | 0.94 | 0.88 | 0.91 | 0.90 | 8.9 | 6.9 | 9.3 | 8.0 | 8.3 |
| | Gradient Boosting | 0.74 | 0.87 | 0.72 | 0.83 | 0.79 | 13.6 | 10.2 | 14.2 | 10.9 | 12.2 |
| | Random Forest | 0.64 | 0.83 | 0.63 | 0.76 | 0.71 | 16.0 | 11.7 | 16.3 | 13.0 | 14.3 |
| WL | Linear Regression | 0.92 | 0.94 | 0.90 | 0.91 | 0.92 | 7.8 | 7.0 | 8.4 | 8.0 | 7.8 |
| | Lasso | 0.92 | 0.94 | 0.90 | 0.91 | 0.92 | 7.6 | 6.8 | 8.2 | 7.7 | 7.6 |
| | Ridge | 0.92 | 0.94 | 0.91 | 0.92 | 0.92 | 7.5 | 6.8 | 8.1 | 7.6 | 7.5 |
| | Bayesian Ridge | 0.92 | 0.94 | 0.91 | 0.92 | 0.92 | 7.5 | 6.9 | 8.1 | 7.6 | 7.5 |
| | SVR - Poly | 0.92 | 0.94 | 0.90 | 0.92 | 0.92 | 7.6 | 7.1 | 8.3 | 7.6 | 7.6 |
| | SVR - RBF | 0.92 | 0.93 | 0.90 | 0.91 | 0.91 | 7.8 | 7.4 | 8.6 | 7.9 | 7.9 |
| | SVR - Sigmoid | 0.83 | 0.85 | 0.82 | 0.82 | 0.83 | 11.1 | 11.1 | 11.4 | 11.1 | 11.2 |
| | SVR - Precomputed | 0.92 | 0.94 | 0.90 | 0.91 | 0.92 | 7.6 | 6.9 | 8.3 | 7.8 | 7.6 |
| | Gradient Boosting | 0.86 | 0.88 | 0.83 | 0.85 | 0.85 | 10.1 | 9.9 | 11.0 | 10.3 | 10.3 |
| | Random Forest | 0.80 | 0.84 | 0.74 | 0.78 | 0.79 | 12.1 | 11.6 | 13.6 | 12.4 | 12.4 |

### Training Set Performance

Table B.24: Training Set Performance for the Models in the Leave-One-Ligand-Out Test

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ABP | XP | $^t$BBP | $^t$BXP | Mean | ABP | XP | $^t$BBP | $^t$BXP | Mean |
| One-hot | Linear Regression | 0.71 | 0.76 | 0.69 | 0.69 | 0.71 | 14.6 | 14.1 | 14.8 | 14.8 | 14.6 |
| | Lasso | 0.71 | 0.76 | 0.69 | 0.69 | 0.71 | 14.6 | 14.1 | 14.8 | 14.8 | 14.6 |
| | Ridge | 0.71 | 0.76 | 0.69 | 0.69 | 0.71 | 14.6 | 14.1 | 14.8 | 14.8 | 14.6 |
| | Bayesian Ridge | 0.71 | 0.76 | 0.69 | 0.69 | 0.71 | 14.6 | 14.1 | 14.8 | 14.8 | 14.6 |
| | SVR - Linear | 0.70 | 0.75 | 0.69 | 0.68 | 0.71 | 14.6 | 14.2 | 14.8 | 14.9 | 14.6 |
| | SVR - Polynomial | 0.97 | 0.98 | 0.96 | 0.97 | 0.97 | 4.8 | 3.7 | 5.0 | 4.8 | 4.6 |
| | SVR - RBF | 1.00 | 1.00 | 0.98 | 1.00 | 0.99 | 1.0 | 0.9 | 4.1 | 1.0 | 1.7 |
| | SVR - Sigmoid | 0.55 | 0.67 | 0.53 | 0.53 | 0.57 | 18.1 | 16.5 | 18.2 | 18.2 | 17.7 |
| | Gradient Boosting | 0.96 | 0.96 | 0.94 | 0.94 | 0.95 | 5.7 | 5.5 | 6.6 | 6.4 | 6.1 |

**Table B.24** Training Set Performance for the Models in the Leave-One-Ligand-Out Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ABP | XP | $^t$BBP | $^t$BXP | Mean | ABP | XP | $^t$BBP | $^t$BXP | Mean |
| | Random Forest | 0.99 | 0.99 | 0.98 | 0.99 | 0.99 | 3.2 | 2.9 | 3.5 | 3.0 | 3.2 |
| Quantum Chemical | Linear Regression | 0.70 | 0.75 | 0.69 | 0.69 | 0.71 | 14.7 | 14.2 | 14.8 | 14.8 | 14.6 |
| | Lasso | 0.70 | 0.75 | 0.68 | 0.68 | 0.70 | 14.8 | 14.4 | 15.0 | 15.0 | 14.8 |
| | Ridge | 0.70 | 0.75 | 0.69 | 0.69 | 0.71 | 14.6 | 14.1 | 14.8 | 14.8 | 14.6 |
| | Bayesian Ridge | 0.70 | 0.75 | 0.69 | 0.68 | 0.70 | 14.7 | 14.2 | 14.9 | 15.0 | 14.7 |
| | SVR - Linear | 0.70 | 0.75 | 0.69 | 0.68 | 0.71 | 14.7 | 14.3 | 14.9 | 14.9 | 14.7 |
| | SVR - Polynomial | 0.95 | 0.96 | 0.95 | 0.95 | 0.95 | 5.8 | 5.7 | 6.2 | 5.9 | 5.9 |
| | SVR - RBF | 0.98 | 0.98 | 0.97 | 0.98 | 0.98 | 4.1 | 4.5 | 4.3 | 4.0 | 4.2 |
| | SVR - Sigmoid | 0.42 | 0.50 | 0.41 | 0.46 | 0.45 | 20.4 | 20.1 | 20.4 | 19.5 | 20.1 |
| | Gradient Boosting | 0.97 | 0.97 | 0.96 | 0.96 | 0.96 | 5.0 | 4.7 | 5.4 | 5.2 | 5.1 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 2.6 | 2.3 | 2.7 | 2.4 | 2.5 |
| Fingerprints: MACCS | Linear Regression | 0.63 | 0.64 | 0.63 | 0.62 | 0.63 | 16.4 | 17.2 | 16.2 | 16.2 | 16.5 |
| | Lasso | 0.65 | 0.67 | 0.63 | 0.63 | 0.64 | 16.0 | 16.4 | 16.1 | 16.1 | 16.2 |
| | Ridge | 0.65 | 0.67 | 0.63 | 0.63 | 0.64 | 16.0 | 16.4 | 16.1 | 16.1 | 16.2 |
| | Bayesian Ridge | 0.65 | 0.67 | 0.63 | 0.63 | 0.64 | 16.0 | 16.4 | 16.1 | 16.1 | 16.2 |
| | SVR - Linear | 0.64 | 0.67 | 0.63 | 0.63 | 0.64 | 16.1 | 16.4 | 16.2 | 16.2 | 16.2 |
| | SVR - Polynomial | 0.91 | 0.91 | 0.90 | 0.90 | 0.90 | 8.3 | 8.4 | 8.6 | 8.4 | 8.4 |
| | SVR - RBF | 0.90 | 0.91 | 0.90 | 0.90 | 0.90 | 8.3 | 8.5 | 8.6 | 8.5 | 8.5 |
| | SVR - Sigmoid | 0.33 | 0.34 | 0.30 | 0.27 | 0.31 | 22.0 | 23.2 | 22.3 | 22.6 | 22.5 |
| | Gradient Boosting | 0.92 | 0.93 | 0.91 | 0.92 | 0.92 | 7.6 | 7.4 | 7.9 | 7.7 | 7.7 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 2.5 | 2.4 | 2.8 | 2.6 | 2.6 |
| Fingerprints: Morgan1 | Linear Regression | 0.68 | 0.73 | 0.66 | 0.66 | 0.68 | 15.3 | 15.0 | 15.6 | 15.4 | 15.3 |
| | Lasso | 0.68 | 0.73 | 0.67 | 0.67 | 0.69 | 15.1 | 14.9 | 15.3 | 15.3 | 15.1 |
| | Ridge | 0.68 | 0.73 | 0.67 | 0.67 | 0.69 | 15.1 | 14.9 | 15.3 | 15.3 | 15.1 |
| | Bayesian Ridge | 0.68 | 0.73 | 0.67 | 0.67 | 0.69 | 15.1 | 14.9 | 15.3 | 15.3 | 15.2 |
| | SVR - Linear | 0.68 | 0.73 | 0.67 | 0.66 | 0.68 | 15.2 | 15.0 | 15.4 | 15.3 | 15.2 |
| | SVR - Polynomial | 0.95 | 0.96 | 0.94 | 0.95 | 0.95 | 5.9 | 5.9 | 6.2 | 5.9 | 6.0 |
| | SVR - RBF | 0.97 | 0.97 | 0.96 | 0.97 | 0.97 | 4.8 | 5.1 | 5.3 | 4.9 | 5.0 |
| | SVR - Sigmoid | 0.49 | 0.54 | 0.47 | 0.45 | 0.49 | 19.3 | 19.4 | 19.4 | 19.6 | 19.4 |
| | Gradient Boosting | 0.95 | 0.96 | 0.94 | 0.95 | 0.95 | 6.2 | 6.0 | 6.6 | 6.2 | 6.3 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 2.5 | 2.4 | 2.7 | 2.4 | 2.5 |
| Fingerprints: RDK | Linear Regression | 0.70 | 0.75 | 0.69 | 0.68 | 0.71 | 14.7 | 14.3 | 14.9 | 14.9 | 14.7 |
| | Lasso | 0.71 | 0.76 | 0.69 | 0.69 | 0.71 | 14.6 | 14.1 | 14.8 | 14.8 | 14.6 |
| | Ridge | 0.71 | 0.76 | 0.69 | 0.69 | 0.71 | 14.6 | 14.1 | 14.8 | 14.8 | 14.6 |
| | Bayesian Ridge | 0.71 | 0.76 | 0.69 | 0.69 | 0.71 | 14.6 | 14.1 | 14.8 | 14.8 | 14.6 |
| | SVR - Linear | 0.70 | 0.75 | 0.69 | 0.68 | 0.71 | 14.6 | 14.2 | 14.8 | 14.9 | 14.6 |
| | SVR - Polynomial | 0.96 | 0.98 | 0.96 | 0.96 | 0.97 | 5.0 | 4.2 | 5.4 | 5.2 | 5.0 |
| | SVR - RBF | 0.98 | 0.98 | 0.97 | 0.97 | 0.97 | 4.2 | 3.6 | 4.7 | 4.9 | 4.4 |
| | SVR - Sigmoid | 0.08 | -0.08 | 0.03 | 0.17 | 0.05 | 25.8 | 29.7 | 26.3 | 24.2 | 26.5 |
| | Gradient Boosting | 0.94 | 0.97 | 0.95 | 0.95 | 0.95 | 6.5 | 5.1 | 5.9 | 5.8 | 5.8 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 2.6 | 2.5 | 2.8 | 2.5 | 2.6 |
| Tanimoto: MACCS | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 2.8 | 2.7 | 2.8 | 2.7 | 2.8 |
| | Ridge | 0.98 | 1.00 | 0.98 | 1.00 | 0.99 | 3.7 | 1.6 | 3.8 | 1.6 | 2.7 |
| | Bayesian Ridge | 0.96 | 0.97 | 0.95 | 0.96 | 0.96 | 5.5 | 5.3 | 5.7 | 5.0 | 5.4 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 0.9 | 1.0 | 0.9 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.1 | 1.0 | 1.1 | 1.0 | 1.1 |
| | SVR - Sigmoid | 0.71 | 0.70 | 0.69 | 0.69 | 0.70 | 14.6 | 15.6 | 14.9 | 14.7 | 14.9 |
| | SVR - Precomputed | 0.98 | 0.98 | 0.96 | 0.98 | 0.97 | 4.2 | 4.5 | 5.1 | 4.1 | 4.5 |
| | Gradient Boosting | 0.99 | 0.98 | 0.98 | 0.98 | 0.98 | 3.2 | 4.2 | 3.3 | 3.4 | 3.5 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 3.8 | 3.6 | 3.7 | 3.6 | 3.7 |
| Tanimoto: Morgan1 | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.0 | 0.9 | 1.0 |
| | Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.7 | 0.7 | 0.7 | 0.6 | 0.7 |
| | Bayesian Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.2 | 0.9 | 1.3 | 0.8 | 1.1 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.2 | 1.0 | 0.9 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| | SVR - Sigmoid | 0.87 | 0.90 | 0.87 | 0.86 | 0.87 | 9.7 | 9.2 | 9.6 | 9.9 | 9.6 |

**Table B.24** Training Set Performance for the Models in the Leave-One-Ligand-Out Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ABP | XP | $^t$BBP | $^t$BXP | Mean | ABP | XP | $^t$BBP | $^t$BXP | Mean |
| | SVR - Precomputed | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 0.9 | 0.9 | 2.5 | 0.9 | 1.3 |
| | Gradient Boosting | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 0.7 | 2.2 | 0.7 | 1.0 | 1.2 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 4.0 | 3.9 | 4.1 | 3.8 | 4.0 |
| Tanimoto: RDK | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.3 | 1.4 | 1.4 | 1.5 | 1.4 |
| | Ridge | 1.00 | 1.00 | 0.97 | 1.00 | 0.99 | 1.3 | 1.3 | 4.2 | 1.4 | 2.1 |
| | Bayesian Ridge | 0.99 | 1.00 | 0.99 | 0.99 | 0.99 | 2.8 | 1.9 | 3.3 | 2.3 | 2.6 |
| | SVR - Polynomial | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 0.9 | 1.0 | 0.9 | 1.0 |
| | SVR - Sigmoid | 0.88 | 0.88 | 0.86 | 0.85 | 0.87 | 9.4 | 9.8 | 10.0 | 10.4 | 9.9 |
| | SVR - Precomputed | 1.00 | 1.00 | 0.97 | 1.00 | 0.99 | 1.0 | 1.2 | 4.4 | 1.2 | 1.9 |
| | Gradient Boosting | 0.99 | 1.00 | 0.99 | 1.00 | 0.99 | 2.8 | 0.9 | 2.9 | 1.0 | 1.9 |
| | Random Forest | 0.97 | 0.98 | 0.97 | 0.98 | 0.98 | 4.4 | 3.6 | 4.6 | 3.8 | 4.1 |
| WL | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.8 | 1.7 | 1.6 | 1.6 | 1.7 |
| | Ridge | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 2.0 | 1.8 | 1.9 | 1.7 | 1.8 |
| | Bayesian Ridge | 0.99 | 0.99 | 0.98 | 0.99 | 0.99 | 3.2 | 2.6 | 3.5 | 2.9 | 3.0 |
| | SVR - Poly | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 0.9 | 1.0 | 0.9 | 0.9 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 0.9 | 1.0 | 1.0 | 1.0 |
| | SVR - Sigmoid | 0.83 | 0.85 | 0.79 | 0.79 | 0.82 | 11.0 | 11.0 | 12.1 | 12.1 | 11.6 |
| | SVR - Precomputed | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.6 | 1.6 | 1.5 | 1.3 | 1.5 |
| | Gradient Boosting | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.8 | 0.8 | 0.9 | 1.6 | 1.0 |
| | Random Forest | 0.98 | 0.99 | 0.98 | 0.99 | 0.98 | 3.5 | 3.3 | 3.5 | 3.2 | 3.4 |

### Test Set Performance

The poor performing models that had a mean $R^2$ value less than -1.00 and a mean $RMSE$ greater than 100% were not included in the following table.

Table B.25: Test Set Performance for the Models in the Leave-One-Ligand-Out Test

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ABP | XP | $^t$BBP | $^t$BXP | Mean | ABP | XP | $^t$BBP | $^t$BXP | Mean |
| One-hot | Lasso | 0.68 | <-1.00 | 0.70 | 0.61 | 0.24 | 16.1 | 25.1 | 15.7 | 17.6 | 18.6 |
| | Ridge | 0.70 | <-1.00 | 0.68 | 0.57 | 0.23 | 15.6 | 25.0 | 16.3 | 18.5 | 18.8 |
| | Bayesian Ridge | 0.70 | <-1.00 | 0.68 | 0.57 | 0.23 | 15.6 | 25.0 | 16.3 | 18.5 | 18.8 |
| | SVR - Linear | 0.70 | <-1.00 | 0.68 | 0.56 | 0.21 | 15.5 | 25.5 | 16.4 | 18.7 | 19.0 |
| | SVR - Polynomial | 0.65 | -0.48 | 0.62 | 0.50 | 0.32 | 16.8 | 21.3 | 17.8 | 19.9 | 18.9 |
| | SVR - RBF | 0.87 | <-1.00 | 0.85 | 0.71 | 0.30 | 10.0 | 26.2 | 11.1 | 15.2 | 15.6 |
| | SVR - Sigmoid | 0.60 | -0.78 | 0.53 | 0.40 | 0.19 | 17.9 | 23.4 | 19.7 | 21.9 | 20.7 |
| | Gradient Boosting | 0.83 | <-1.00 | 0.87 | 0.78 | 0.35 | 11.7 | 25.3 | 10.3 | 13.2 | 15.1 |
| | Random Forest | 0.84 | <-1.00 | 0.94 | 0.81 | 0.36 | 11.3 | 25.8 | 7.1 | 12.1 | 14.1 |
| Quantum Chemical | Lasso | 0.53 | -0.83 | 0.23 | <-1.00 | <-1.00 | 19.4 | 23.7 | 25.2 | 67.4 | 33.9 |
| | SVR - RBF | -0.12 | -0.16 | 0.17 | -0.40 | -0.13 | 30.0 | 18.9 | 26.1 | 33.2 | 27.1 |
| | SVR - Sigmoid | <-1.00 | <-1.00 | 0.40 | 0.33 | -0.70 | 51.7 | 26.1 | 22.2 | 23.0 | 30.8 |
| | Gradient Boosting | 0.81 | <-1.00 | 0.88 | 0.65 | 0.28 | 12.5 | 26.0 | 9.9 | 16.7 | 16.3 |
| | Random Forest | 0.90 | <-1.00 | 0.89 | 0.56 | 0.32 | 8.9 | 25.3 | 9.4 | 18.7 | 15.6 |
| Fingerprints: MACCS | Lasso | 0.60 | <-1.00 | 0.64 | 0.61 | 0.13 | 17.9 | 26.8 | 17.1 | 17.5 | 19.8 |
| | Ridge | 0.62 | <-1.00 | 0.64 | 0.40 | 0.11 | 17.5 | 26.2 | 17.1 | 21.7 | 20.6 |
| | Bayesian Ridge | 0.62 | <-1.00 | 0.64 | 0.40 | 0.11 | 17.5 | 26.1 | 17.1 | 21.7 | 20.6 |
| | SVR - Linear | 0.63 | <-1.00 | 0.65 | 0.39 | 0.09 | 17.4 | 26.6 | 17.1 | 22.0 | 20.8 |
| | SVR - Polynomial | 0.84 | <-1.00 | 0.87 | 0.49 | 0.21 | 11.2 | 26.8 | 10.3 | 20.2 | 17.1 |
| | SVR - RBF | 0.84 | <-1.00 | 0.86 | 0.51 | 0.27 | 11.4 | 25.5 | 10.6 | 19.7 | 16.8 |

**Table B.25** Test Set Performance for the Models in the Leave-One-Ligand-Out Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ABP | XP | $^t$BBP | $^t$BXP | Mean | ABP | XP | $^t$BBP | $^t$BXP | Mean |
| | SVR - Sigmoid | 0.26 | <-1.00 | 0.23 | -0.02 | -0.14 | 24.5 | 24.9 | 25.2 | 28.4 | 25.8 |
| | Gradient Boosting | 0.84 | <-1.00 | 0.87 | 0.59 | 0.22 | 11.2 | 27.3 | 10.3 | 18.0 | 16.7 |
| | Random Forest | 0.91 | <-1.00 | 0.91 | 0.82 | 0.33 | 8.5 | 26.7 | 8.6 | 12.0 | 13.9 |
| Fingerprints: Morgan1 | Lasso | 0.67 | <-1.00 | 0.68 | 0.32 | 0.05 | 16.4 | 27.5 | 16.1 | 23.2 | 20.8 |
| | Ridge | 0.58 | <-1.00 | 0.66 | 0.51 | 0.13 | 18.3 | 26.0 | 16.8 | 19.7 | 20.2 |
| | Bayesian Ridge | 0.58 | <-1.00 | 0.66 | 0.51 | 0.14 | 18.3 | 26.0 | 16.8 | 19.8 | 20.2 |
| | SVR - Linear | 0.57 | <-1.00 | 0.65 | 0.49 | 0.10 | 18.6 | 26.6 | 16.9 | 20.1 | 20.6 |
| | SVR - Polynomial | 0.77 | <-1.00 | 0.88 | 0.62 | 0.28 | 13.7 | 25.7 | 9.8 | 17.2 | 16.6 |
| | SVR - RBF | 0.75 | <-1.00 | 0.88 | 0.63 | 0.30 | 14.1 | 25.0 | 9.9 | 17.2 | 16.6 |
| | SVR - Sigmoid | 0.38 | -0.75 | 0.46 | 0.23 | 0.08 | 22.3 | 23.2 | 21.2 | 24.7 | 22.8 |
| | Gradient Boosting | 0.56 | <-1.00 | 0.90 | 0.58 | 0.21 | 18.8 | 25.9 | 9.3 | 18.3 | 18.1 |
| | Random Forest | 0.83 | <-1.00 | 0.91 | 0.64 | 0.30 | 11.8 | 25.8 | 8.5 | 16.9 | 15.8 |
| Fingerprints: RDK | Lasso | 0.48 | -0.99 | 0.71 | 0.35 | 0.14 | 20.5 | 24.7 | 15.5 | 22.7 | 20.9 |
| | Ridge | 0.68 | <-1.00 | 0.72 | 0.34 | 0.17 | 15.9 | 25.1 | 15.1 | 22.9 | 19.8 |
| | Bayesian Ridge | 0.68 | <-1.00 | 0.72 | 0.34 | 0.17 | 16.0 | 25.1 | 15.1 | 22.9 | 19.8 |
| | SVR - Linear | 0.69 | <-1.00 | 0.72 | 0.32 | 0.14 | 15.8 | 25.7 | 15.2 | 23.2 | 20.0 |
| | SVR - Polynomial | 0.86 | <-1.00 | 0.92 | 0.29 | 0.23 | 10.6 | 25.6 | 8.3 | 23.8 | 17.1 |
| | SVR - RBF | 0.87 | <-1.00 | 0.92 | 0.27 | 0.25 | 10.1 | 25.2 | 8.1 | 24.0 | 16.9 |
| | SVR - Sigmoid | 0.10 | <-1.00 | 0.08 | -0.08 | -0.53 | 27.0 | 31.4 | 27.5 | 29.2 | 28.7 |
| | Gradient Boosting | 0.77 | <-1.00 | 0.92 | 0.30 | 0.22 | 13.7 | 25.5 | 8.4 | 23.6 | 17.8 |
| | Random Forest | 0.50 | <-1.00 | 0.93 | 0.04 | 0.11 | 20.1 | 24.9 | 7.5 | 27.6 | 20.0 |
| Tanimoto: MACCS | Linear Regression | 0.91 | -0.79 | 0.94 | 0.56 | 0.41 | 8.3 | 23.4 | 7.1 | 18.6 | 14.3 |
| | Lasso | 0.91 | -0.75 | 0.94 | 0.56 | 0.41 | 8.6 | 23.2 | 7.2 | 18.6 | 14.4 |
| | Ridge | 0.90 | -0.56 | 0.93 | 0.53 | 0.45 | 9.0 | 21.9 | 7.8 | 19.2 | 14.5 |
| | Bayesian Ridge | 0.88 | -0.55 | 0.91 | 0.52 | 0.44 | 9.9 | 21.8 | 8.8 | 19.5 | 15.0 |
| | SVR - Polynomial | 0.90 | -0.39 | 0.93 | 0.48 | 0.48 | 8.8 | 20.7 | 7.5 | 20.2 | 14.3 |
| | SVR - RBF | 0.89 | -0.36 | 0.92 | 0.45 | 0.48 | 9.3 | 20.4 | 8.0 | 21.0 | 14.7 |
| | SVR - Sigmoid | 0.63 | -0.81 | 0.66 | 0.34 | 0.20 | 17.3 | 23.6 | 16.6 | 22.8 | 20.1 |
| | SVR - Precomputed | 0.89 | -0.58 | 0.92 | 0.52 | 0.44 | 9.2 | 22.0 | 8.2 | 19.5 | 14.8 |
| | Gradient Boosting | 0.87 | <-1.00 | 0.89 | 0.38 | 0.26 | 10.4 | 25.3 | 9.6 | 22.1 | 16.9 |
| | Random Forest | 0.84 | <-1.00 | 0.87 | 0.30 | 0.14 | 11.5 | 27.5 | 10.3 | 23.6 | 18.2 |
| Tanimoto: Morgan1 | Linear Regression | 0.73 | -0.71 | 0.93 | 0.62 | 0.39 | 14.7 | 22.9 | 7.6 | 17.4 | 15.7 |
| | Lasso | 0.73 | -0.61 | 0.93 | 0.62 | 0.42 | 14.7 | 22.2 | 7.6 | 17.4 | 15.5 |
| | Ridge | 0.72 | -0.65 | 0.93 | 0.62 | 0.41 | 14.9 | 22.5 | 7.6 | 17.4 | 15.6 |
| | Bayesian Ridge | 0.72 | -0.65 | 0.93 | 0.62 | 0.41 | 14.9 | 22.5 | 7.6 | 17.4 | 15.6 |
| | SVR - Polynomial | 0.67 | -0.53 | 0.90 | 0.59 | 0.41 | 16.3 | 21.7 | 8.9 | 18.0 | 16.2 |
| | SVR - RBF | 0.64 | -0.51 | 0.88 | 0.57 | 0.39 | 17.0 | 21.6 | 10.0 | 18.4 | 16.8 |
| | SVR - Sigmoid | 0.71 | -0.82 | 0.87 | 0.59 | 0.34 | 15.3 | 23.6 | 10.2 | 18.0 | 16.8 |
| | SVR - Precomputed | 0.72 | -0.65 | 0.92 | 0.62 | 0.40 | 14.9 | 22.5 | 7.9 | 17.4 | 15.7 |
| | Gradient Boosting | 0.62 | <-1.00 | 0.85 | 0.54 | 0.22 | 17.5 | 25.5 | 11.1 | 19.0 | 18.3 |
| | Random Forest | 0.58 | -0.98 | 0.85 | 0.38 | 0.21 | 18.3 | 24.7 | 11.0 | 22.2 | 19.0 |
| Tanimoto: RDK | Linear Regression | 0.90 | <-1.00 | 0.94 | 0.26 | 0.27 | 9.1 | 24.9 | 7.3 | 24.2 | 16.4 |
| | Lasso | 0.90 | <-1.00 | 0.94 | 0.26 | 0.27 | 9.1 | 24.8 | 7.1 | 24.2 | 16.3 |
| | Ridge | 0.90 | <-1.00 | 0.94 | 0.26 | 0.27 | 9.1 | 24.8 | 7.1 | 24.3 | 16.3 |
| | Bayesian Ridge | 0.89 | -1.00 | 0.94 | 0.27 | 0.27 | 9.2 | 24.8 | 7.0 | 24.1 | 16.3 |
| | SVR - Polynomial | 0.90 | -0.98 | 0.94 | 0.26 | 0.28 | 9.1 | 24.7 | 7.0 | 24.3 | 16.3 |
| | SVR - RBF | 0.90 | -0.96 | 0.94 | 0.26 | 0.28 | 9.1 | 24.6 | 7.0 | 24.2 | 16.2 |
| | SVR - Sigmoid | 0.81 | <-1.00 | 0.86 | 0.28 | 0.23 | 12.3 | 25.1 | 10.6 | 23.9 | 18.0 |
| | SVR - Precomputed | 0.90 | <-1.00 | 0.93 | 0.25 | 0.27 | 9.1 | 24.9 | 7.5 | 24.4 | 16.5 |
| | Gradient Boosting | 0.76 | -0.73 | 0.89 | 0.27 | 0.30 | 13.9 | 23.0 | 9.3 | 24.0 | 17.5 |
| | Random Forest | 0.56 | -0.95 | 0.87 | 0.10 | 0.15 | 18.8 | 24.5 | 10.4 | 26.7 | 20.1 |
| WL | Linear Regression | 0.78 | -0.94 | 0.91 | 0.76 | 0.38 | 13.3 | 24.4 | 8.7 | 13.9 | 15.1 |
| | Lasso | 0.78 | -0.91 | 0.91 | 0.76 | 0.38 | 13.4 | 24.2 | 8.5 | 13.9 | 15.0 |
| | Ridge | 0.77 | -0.87 | 0.91 | 0.75 | 0.39 | 13.7 | 24.0 | 8.4 | 13.9 | 15.0 |
| | Bayesian Ridge | 0.77 | -0.86 | 0.91 | 0.75 | 0.39 | 13.7 | 23.9 | 8.4 | 13.9 | 15.0 |
| | SVR - Poly | 0.76 | -0.78 | 0.91 | 0.75 | 0.41 | 14.0 | 23.4 | 8.5 | 14.0 | 14.9 |
| | SVR - RBF | 0.75 | -0.73 | 0.91 | 0.75 | 0.42 | 14.2 | 23.1 | 8.5 | 14.0 | 14.9 |
| | SVR - Sigmoid | 0.71 | -0.89 | 0.81 | 0.66 | 0.32 | 15.3 | 24.1 | 12.5 | 16.4 | 17.1 |
| | SVR - Precomputed | 0.77 | -0.88 | 0.91 | 0.75 | 0.39 | 13.6 | 24.0 | 8.6 | 13.9 | 15.0 |

**Table B.25** Test Set Performance for the Models in the Leave-One-Ligand-Out Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ABP | XP | $^t$BBP | $^t$BXP | Mean | ABP | XP | $^t$BBP | $^t$BXP | Mean |
| | Gradient Boosting | 0.61 | -0.67 | 0.82 | 0.78 | 0.38 | 17.8 | 22.6 | 12.1 | 13.3 | 16.5 |
| | Random Forest | 0.28 | -0.79 | 0.90 | 0.83 | 0.30 | 24.1 | 23.5 | 9.0 | 11.7 | 17.0 |

# B.8 Out-of-Sample Tests: With Activity Ranking

## B.8.1 Additive Ranked Test

### Grid Search Cross-Validation Performance

Table B.26: Grid Search Cross-Validated Performance for the Models in the Additive Ranked Test

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | Mean | **1** | **2** | **3** | **4** | Mean |
| One-hot | Linear Regression | 0.69 | 0.72 | 0.69 | 0.69 | 0.70 | 15.0 | 14.4 | 15.4 | 15.2 | 15.0 |
| | Lasso | 0.70 | 0.72 | 0.69 | 0.69 | 0.70 | 15.0 | 14.3 | 15.4 | 15.2 | 15.0 |
| | Ridge | 0.70 | 0.72 | 0.69 | 0.69 | 0.70 | 15.0 | 14.3 | 15.4 | 15.2 | 15.0 |
| | Bayesian Ridge | 0.70 | 0.72 | 0.69 | 0.69 | 0.70 | 15.0 | 14.3 | 15.4 | 15.2 | 15.0 |
| | SVR - Linear | 0.69 | 0.72 | 0.68 | 0.69 | 0.70 | 15.1 | 14.4 | 15.4 | 15.2 | 15.0 |
| | SVR - Poly | 0.90 | 0.90 | 0.89 | 0.90 | 0.90 | 8.6 | 8.6 | 9.2 | 8.7 | 8.8 |
| | SVR - RBF | 0.91 | 0.91 | 0.90 | 0.91 | 0.91 | 8.1 | 7.9 | 8.6 | 8.2 | 8.2 |
| | SVR - Sigmoid | 0.58 | 0.60 | 0.56 | 0.57 | 0.58 | 17.6 | 17.0 | 18.1 | 18.0 | 17.7 |
| | Gradient Boosting | 0.90 | 0.90 | 0.89 | 0.90 | 0.90 | 8.7 | 8.5 | 9.2 | 8.8 | 8.8 |
| | Random Forest | 0.89 | 0.91 | 0.89 | 0.90 | 0.90 | 8.8 | 8.2 | 9.2 | 8.5 | 8.7 |
| Quantum Chemical | Linear Regression | 0.69 | 0.72 | 0.68 | 0.69 | 0.70 | 15.0 | 14.4 | 15.4 | 15.3 | 15.0 |
| | Lasso | 0.69 | 0.72 | 0.68 | 0.69 | 0.70 | 15.1 | 14.4 | 15.4 | 15.3 | 15.0 |
| | Ridge | 0.70 | 0.72 | 0.69 | 0.69 | 0.70 | 15.0 | 14.3 | 15.4 | 15.2 | 15.0 |
| | Bayesian Ridge | 0.69 | 0.71 | 0.68 | 0.69 | 0.69 | 15.1 | 14.4 | 15.4 | 15.3 | 15.1 |
| | SVR - Linear | 0.69 | 0.72 | 0.68 | 0.69 | 0.69 | 15.1 | 14.4 | 15.4 | 15.3 | 15.0 |
| | SVR - Poly | 0.89 | 0.91 | 0.89 | 0.89 | 0.89 | 8.9 | 8.3 | 9.2 | 9.1 | 8.9 |
| | SVR - RBF | 0.90 | 0.92 | 0.89 | 0.90 | 0.90 | 8.4 | 7.7 | 8.9 | 8.8 | 8.5 |
| | SVR - Sigmoid | 0.45 | 0.47 | 0.45 | 0.45 | 0.45 | 20.2 | 19.7 | 20.3 | 20.3 | 20.1 |
| | Gradient Boosting | 0.92 | 0.92 | 0.91 | 0.92 | 0.92 | 7.7 | 7.5 | 8.2 | 7.9 | 7.8 |
| | Random Forest | 0.93 | 0.93 | 0.92 | 0.93 | 0.93 | 7.2 | 7.1 | 7.6 | 7.4 | 7.3 |
| Fingerprints: MACCS | Linear Regression | 0.60 | 0.60 | 0.54 | 0.64 | 0.60 | 17.1 | 16.9 | 18.5 | 16.3 | 17.2 |
| | Lasso | 0.65 | 0.66 | 0.63 | 0.66 | 0.65 | 16.2 | 15.8 | 16.7 | 16.0 | 16.2 |
| | Ridge | 0.65 | 0.66 | 0.63 | 0.66 | 0.65 | 16.2 | 15.8 | 16.7 | 16.0 | 16.2 |
| | Bayesian Ridge | 0.65 | 0.66 | 0.63 | 0.66 | 0.65 | 16.2 | 15.8 | 16.7 | 16.0 | 16.2 |
| | SVR - Linear | 0.64 | 0.66 | 0.63 | 0.65 | 0.65 | 16.2 | 15.8 | 16.7 | 16.1 | 16.2 |
| | SVR - Poly | 0.87 | 0.88 | 0.85 | 0.87 | 0.87 | 9.8 | 9.3 | 10.5 | 9.8 | 9.9 |
| | SVR - RBF | 0.87 | 0.88 | 0.85 | 0.87 | 0.86 | 9.9 | 9.5 | 10.7 | 10.0 | 10.0 |
| | SVR - Sigmoid | 0.31 | 0.28 | 0.27 | 0.26 | 0.28 | 22.6 | 23.0 | 23.5 | 23.6 | 23.2 |
| | Gradient Boosting | 0.89 | 0.90 | 0.88 | 0.90 | 0.89 | 8.9 | 8.5 | 9.6 | 8.8 | 9.0 |
| | Random Forest | 0.93 | 0.93 | 0.92 | 0.93 | 0.93 | 7.1 | 7.1 | 7.7 | 7.4 | 7.3 |
| Fingerprints: Morgan1 | Linear Regression | 0.65 | 0.69 | 0.66 | 0.66 | 0.67 | 16.0 | 14.9 | 16.0 | 15.9 | 15.7 |
| | Lasso | 0.67 | 0.70 | 0.67 | 0.67 | 0.68 | 15.6 | 14.8 | 15.8 | 15.7 | 15.5 |
| | Ridge | 0.67 | 0.70 | 0.67 | 0.67 | 0.68 | 15.6 | 14.8 | 15.8 | 15.7 | 15.5 |
| | Bayesian Ridge | 0.67 | 0.70 | 0.67 | 0.67 | 0.68 | 15.6 | 14.8 | 15.8 | 15.7 | 15.5 |
| | SVR - Linear | 0.67 | 0.70 | 0.66 | 0.67 | 0.67 | 15.6 | 14.9 | 15.9 | 15.7 | 15.5 |
| | SVR - Poly | 0.91 | 0.92 | 0.91 | 0.91 | 0.91 | 8.2 | 7.7 | 8.4 | 8.2 | 8.1 |
| | SVR - RBF | 0.92 | 0.93 | 0.91 | 0.92 | 0.92 | 7.6 | 7.2 | 8.0 | 7.7 | 7.6 |
| | SVR - Sigmoid | 0.49 | 0.48 | 0.45 | 0.45 | 0.46 | 19.5 | 19.6 | 20.3 | 20.4 | 19.9 |
| | Gradient Boosting | 0.91 | 0.92 | 0.90 | 0.91 | 0.91 | 8.2 | 7.8 | 8.6 | 8.3 | 8.2 |
| | Random Forest | 0.93 | 0.93 | 0.92 | 0.93 | 0.93 | 7.1 | 6.9 | 7.5 | 7.4 | 7.2 |
| Fingerprints: RDK | Linear Regression | 0.69 | 0.71 | 0.67 | 0.68 | 0.69 | 15.1 | 14.6 | 15.6 | 15.5 | 15.2 |
| | Lasso | 0.70 | 0.72 | 0.69 | 0.69 | 0.70 | 15.0 | 14.3 | 15.4 | 15.2 | 15.0 |
| | Ridge | 0.70 | 0.72 | 0.69 | 0.69 | 0.70 | 15.0 | 14.3 | 15.4 | 15.2 | 15.0 |
| | Bayesian Ridge | 0.70 | 0.72 | 0.69 | 0.69 | 0.70 | 15.0 | 14.3 | 15.4 | 15.2 | 15.0 |
| | SVR - Linear | 0.69 | 0.72 | 0.68 | 0.69 | 0.70 | 15.1 | 14.4 | 15.4 | 15.2 | 15.0 |
| | SVR - Poly | 0.91 | 0.91 | 0.90 | 0.91 | 0.91 | 8.1 | 8.0 | 8.6 | 8.2 | 8.2 |
| | SVR - RBF | 0.91 | 0.92 | 0.91 | 0.92 | 0.91 | 7.9 | 7.7 | 8.3 | 7.9 | 8.0 |

**Table B.26** Grid Search Cross-Validated Performance for the Models in the Additive Ranked Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | Mean | **1** | **2** | **3** | **4** | Mean |
| | SVR - Sigmoid | 0.23 | 0.36 | 0.45 | 0.33 | 0.34 | 23.8 | 21.5 | 20.4 | 22.4 | 22.0 |
| | Gradient Boosting | 0.91 | 0.91 | 0.90 | 0.91 | 0.91 | 8.3 | 8.0 | 8.9 | 8.3 | 8.4 |
| | Random Forest | 0.93 | 0.93 | 0.92 | 0.93 | 0.93 | 7.4 | 6.9 | 7.8 | 7.5 | 7.4 |
| Tanimoto: MACCS | Linear Regression | 0.90 | 0.87 | 0.87 | 0.89 | 0.88 | 8.4 | 9.6 | 9.8 | 9.1 | 9.2 |
| | Lasso | 0.92 | 0.90 | 0.90 | 0.91 | 0.91 | 7.5 | 8.4 | 8.8 | 8.3 | 8.3 |
| | Ridge | 0.93 | 0.91 | 0.90 | 0.91 | 0.91 | 7.3 | 8.0 | 8.7 | 8.1 | 8.0 |
| | Bayesian Ridge | 0.91 | 0.91 | 0.88 | 0.91 | 0.90 | 8.2 | 8.3 | 9.5 | 8.4 | 8.6 |
| | SVR - Poly | 0.92 | 0.90 | 0.89 | 0.90 | 0.90 | 7.6 | 8.6 | 9.0 | 8.4 | 8.4 |
| | SVR - RBF | 0.91 | 0.89 | 0.88 | 0.90 | 0.89 | 8.0 | 9.0 | 9.5 | 8.8 | 8.8 |
| | SVR - Sigmoid | 0.70 | 0.69 | 0.66 | 0.68 | 0.68 | 14.8 | 15.0 | 15.9 | 15.5 | 15.3 |
| | SVR - Precomputed | 0.92 | 0.90 | 0.89 | 0.91 | 0.90 | 7.9 | 8.4 | 9.2 | 8.3 | 8.5 |
| | Gradient Boosting | 0.78 | 0.79 | 0.76 | 0.77 | 0.78 | 12.8 | 12.3 | 13.3 | 13.1 | 12.9 |
| | Random Forest | 0.71 | 0.68 | 0.69 | 0.72 | 0.70 | 14.7 | 15.2 | 15.2 | 14.4 | 14.9 |
| Tanimoto: Morgan1 | Linear Regression | 0.94 | 0.93 | 0.93 | 0.93 | 0.93 | 6.7 | 7.1 | 7.1 | 7.4 | 7.1 |
| | Lasso | 0.94 | 0.93 | 0.93 | 0.93 | 0.94 | 6.5 | 6.9 | 7.0 | 7.2 | 6.9 |
| | Ridge | 0.94 | 0.93 | 0.93 | 0.93 | 0.94 | 6.5 | 6.9 | 7.0 | 7.1 | 6.9 |
| | Bayesian Ridge | 0.94 | 0.94 | 0.94 | 0.93 | 0.94 | 6.5 | 6.8 | 7.0 | 7.1 | 6.9 |
| | SVR - Poly | 0.93 | 0.92 | 0.92 | 0.92 | 0.92 | 7.3 | 7.5 | 7.8 | 7.8 | 7.6 |
| | SVR - RBF | 0.91 | 0.91 | 0.90 | 0.90 | 0.90 | 8.2 | 8.2 | 8.7 | 8.6 | 8.4 |
| | SVR - Sigmoid | 0.88 | 0.90 | 0.87 | 0.88 | 0.88 | 9.2 | 8.6 | 9.8 | 9.6 | 9.3 |
| | SVR - Precomputed | 0.94 | 0.94 | 0.93 | 0.93 | 0.94 | 6.6 | 6.8 | 7.1 | 7.2 | 6.9 |
| | Gradient Boosting | 0.82 | 0.81 | 0.78 | 0.79 | 0.80 | 11.6 | 11.7 | 12.8 | 12.4 | 12.1 |
| | Random Forest | 0.74 | 0.74 | 0.69 | 0.71 | 0.72 | 13.8 | 13.7 | 15.3 | 14.8 | 14.4 |
| Tanimoto: RDK | Linear Regression | 0.92 | 0.90 | 0.90 | 0.91 | 0.91 | 7.9 | 8.5 | 8.7 | 8.3 | 8.3 |
| | Lasso | 0.92 | 0.91 | 0.91 | 0.91 | 0.91 | 7.7 | 8.1 | 8.4 | 8.0 | 8.0 |
| | Ridge | 0.92 | 0.91 | 0.91 | 0.92 | 0.91 | 7.6 | 8.0 | 8.3 | 7.9 | 8.0 |
| | Bayesian Ridge | 0.92 | 0.91 | 0.91 | 0.92 | 0.92 | 7.6 | 8.0 | 8.2 | 7.9 | 7.9 |
| | SVR - Poly | 0.91 | 0.90 | 0.90 | 0.90 | 0.90 | 8.3 | 8.6 | 8.9 | 8.5 | 8.6 |
| | SVR - RBF | 0.89 | 0.88 | 0.88 | 0.89 | 0.89 | 9.0 | 9.2 | 9.5 | 9.0 | 9.2 |
| | SVR - Sigmoid | 0.85 | 0.86 | 0.85 | 0.85 | 0.85 | 10.6 | 10.0 | 10.7 | 10.5 | 10.4 |
| | SVR - Precomputed | 0.92 | 0.91 | 0.91 | 0.91 | 0.91 | 7.6 | 8.1 | 8.4 | 8.0 | 8.0 |
| | Gradient Boosting | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 12.7 | 12.6 | 12.7 | 12.7 | 12.7 |
| | Random Forest | 0.71 | 0.71 | 0.71 | 0.72 | 0.71 | 14.6 | 14.6 | 14.8 | 14.5 | 14.6 |
| WL | Linear Regression | 0.94 | 0.91 | 0.92 | 0.92 | 0.92 | 6.9 | 7.9 | 8.0 | 7.9 | 7.7 |
| | Lasso | 0.94 | 0.92 | 0.92 | 0.92 | 0.92 | 6.9 | 7.7 | 7.8 | 7.7 | 7.5 |
| | Ridge | 0.94 | 0.92 | 0.92 | 0.92 | 0.93 | 6.8 | 7.5 | 7.7 | 7.5 | 7.4 |
| | Bayesian Ridge | 0.94 | 0.92 | 0.92 | 0.92 | 0.93 | 6.9 | 7.5 | 7.8 | 7.6 | 7.4 |
| | SVR - Poly | 0.94 | 0.92 | 0.92 | 0.92 | 0.92 | 6.9 | 7.6 | 7.9 | 7.7 | 7.5 |
| | SVR - RBF | 0.93 | 0.92 | 0.91 | 0.92 | 0.92 | 7.2 | 7.8 | 8.2 | 7.9 | 7.8 |
| | SVR - Sigmoid | 0.83 | 0.84 | 0.82 | 0.83 | 0.83 | 11.3 | 10.7 | 11.7 | 11.4 | 11.3 |
| | SVR - Precomputed | 0.94 | 0.92 | 0.92 | 0.92 | 0.92 | 6.8 | 7.6 | 7.9 | 7.7 | 7.5 |
| | Gradient Boosting | 0.86 | 0.86 | 0.85 | 0.85 | 0.85 | 10.2 | 10.2 | 10.7 | 10.6 | 10.4 |
| | Random Forest | 0.79 | 0.78 | 0.75 | 0.76 | 0.77 | 12.4 | 12.7 | 13.7 | 13.3 | 13.0 |

### Training Set Performance

Table B.27: Training Set Performance for the Models in the Additive Ranked Test

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | Mean | **1** | **2** | **3** | **4** | Mean |
| One-hot | Linear Regression | 0.70 | 0.73 | 0.69 | 0.70 | 0.71 | 14.9 | 14.1 | 15.2 | 15.0 | 14.8 |
| | Lasso | 0.70 | 0.73 | 0.69 | 0.70 | 0.71 | 14.8 | 14.1 | 15.2 | 15.0 | 14.8 |
| | Ridge | 0.70 | 0.73 | 0.69 | 0.70 | 0.71 | 14.8 | 14.1 | 15.2 | 15.0 | 14.8 |
| | Bayesian Ridge | 0.70 | 0.73 | 0.69 | 0.70 | 0.71 | 14.8 | 14.1 | 15.2 | 15.0 | 14.8 |

**Table B.27** Training Set Performance for the Models in the Additive Ranked Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | Mean | **1** | **2** | **3** | **4** | Mean |
| | SVR - Linear | 0.70 | 0.73 | 0.69 | 0.70 | 0.70 | 14.9 | 14.2 | 15.2 | 15.0 | 14.8 |
| | SVR - Poly | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 4.7 | 4.7 | 4.8 | 4.7 | 4.7 |
| | SVR - RBF | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.0 | 3.3 | 1.0 | 1.0 | 1.5 |
| | SVR - Sigmoid | 0.59 | 0.61 | 0.57 | 0.57 | 0.59 | 17.5 | 16.9 | 17.9 | 17.9 | 17.5 |
| | Gradient Boosting | 0.96 | 0.95 | 0.95 | 0.95 | 0.95 | 5.6 | 6.1 | 6.1 | 5.8 | 5.9 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 3.0 | 2.8 | 3.1 | 2.9 | 2.9 |
| Quantum Chemical | Linear Regression | 0.70 | 0.73 | 0.69 | 0.70 | 0.70 | 14.9 | 14.2 | 15.2 | 15.1 | 14.8 |
| | Lasso | 0.70 | 0.72 | 0.69 | 0.70 | 0.70 | 14.9 | 14.2 | 15.2 | 15.1 | 14.9 |
| | Ridge | 0.70 | 0.73 | 0.69 | 0.70 | 0.71 | 14.8 | 14.1 | 15.2 | 15.0 | 14.8 |
| | Bayesian Ridge | 0.70 | 0.72 | 0.69 | 0.70 | 0.70 | 14.9 | 14.2 | 15.3 | 15.1 | 14.9 |
| | SVR - Linear | 0.70 | 0.73 | 0.69 | 0.70 | 0.70 | 14.9 | 14.2 | 15.2 | 15.0 | 14.8 |
| | SVR - Poly | 0.96 | 0.96 | 0.95 | 0.96 | 0.96 | 5.7 | 5.3 | 5.9 | 5.7 | 5.7 |
| | SVR - RBF | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 3.8 | 3.5 | 4.0 | 3.9 | 3.8 |
| | SVR - Sigmoid | 0.44 | 0.46 | 0.43 | 0.44 | 0.44 | 20.4 | 19.9 | 20.7 | 20.5 | 20.4 |
| | Gradient Boosting | 0.97 | 0.97 | 0.96 | 0.97 | 0.97 | 5.0 | 4.9 | 5.3 | 5.1 | 5.1 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 2.4 | 2.4 | 2.6 | 2.5 | 2.5 |
| Fingerprints: MACCS | Linear Regression | 0.65 | 0.66 | 0.63 | 0.59 | 0.63 | 16.1 | 15.8 | 16.7 | 17.5 | 16.5 |
| | Lasso | 0.65 | 0.67 | 0.64 | 0.66 | 0.66 | 16.0 | 15.6 | 16.5 | 15.9 | 16.0 |
| | Ridge | 0.65 | 0.67 | 0.64 | 0.66 | 0.66 | 16.0 | 15.6 | 16.5 | 15.9 | 16.0 |
| | Bayesian Ridge | 0.65 | 0.67 | 0.64 | 0.66 | 0.66 | 16.0 | 15.6 | 16.5 | 15.9 | 16.0 |
| | SVR - Linear | 0.65 | 0.67 | 0.64 | 0.66 | 0.65 | 16.1 | 15.6 | 16.6 | 15.9 | 16.0 |
| | SVR - Poly | 0.91 | 0.92 | 0.90 | 0.91 | 0.91 | 8.2 | 7.6 | 8.8 | 8.1 | 8.2 |
| | SVR - RBF | 0.91 | 0.92 | 0.90 | 0.91 | 0.91 | 8.2 | 7.7 | 8.9 | 8.2 | 8.2 |
| | SVR - Sigmoid | 0.34 | 0.32 | 0.30 | 0.29 | 0.31 | 22.1 | 22.3 | 22.9 | 23.1 | 22.6 |
| | Gradient Boosting | 0.93 | 0.93 | 0.91 | 0.93 | 0.93 | 7.4 | 6.9 | 8.0 | 7.3 | 7.4 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 2.4 | 2.4 | 2.6 | 2.4 | 2.5 |
| Fingerprints: Morgan1 | Linear Regression | 0.57 | 0.70 | 0.66 | 0.67 | 0.65 | 17.8 | 14.9 | 15.9 | 15.7 | 16.1 |
| | Lasso | 0.68 | 0.71 | 0.67 | 0.68 | 0.68 | 15.4 | 14.7 | 15.7 | 15.5 | 15.3 |
| | Ridge | 0.68 | 0.71 | 0.67 | 0.68 | 0.68 | 15.4 | 14.7 | 15.7 | 15.5 | 15.3 |
| | Bayesian Ridge | 0.68 | 0.71 | 0.67 | 0.68 | 0.68 | 15.4 | 14.7 | 15.7 | 15.5 | 15.3 |
| | SVR - Linear | 0.68 | 0.70 | 0.67 | 0.68 | 0.68 | 15.5 | 14.7 | 15.7 | 15.6 | 15.4 |
| | SVR - Poly | 0.95 | 0.96 | 0.95 | 0.95 | 0.95 | 6.0 | 5.4 | 6.2 | 5.9 | 5.9 |
| | SVR - RBF | 0.97 | 0.97 | 0.96 | 0.97 | 0.97 | 5.0 | 4.4 | 5.2 | 4.8 | 4.8 |
| | SVR - Sigmoid | 0.51 | 0.51 | 0.48 | 0.48 | 0.49 | 19.1 | 19.0 | 19.7 | 19.8 | 19.4 |
| | Gradient Boosting | 0.95 | 0.96 | 0.94 | 0.95 | 0.95 | 6.3 | 5.6 | 6.5 | 6.3 | 6.2 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 2.5 | 2.4 | 2.5 | 2.4 | 2.4 |
| Fingerprints: RDK | Linear Regression | 0.70 | 0.72 | 0.68 | 0.70 | 0.70 | 14.9 | 14.3 | 15.5 | 15.1 | 15.0 |
| | Lasso | 0.70 | 0.73 | 0.69 | 0.70 | 0.71 | 14.8 | 14.1 | 15.2 | 15.0 | 14.8 |
| | Ridge | 0.70 | 0.73 | 0.69 | 0.70 | 0.71 | 14.8 | 14.1 | 15.2 | 15.0 | 14.8 |
| | Bayesian Ridge | 0.70 | 0.73 | 0.69 | 0.70 | 0.71 | 14.8 | 14.1 | 15.2 | 15.0 | 14.8 |
| | SVR - Linear | 0.70 | 0.73 | 0.69 | 0.70 | 0.70 | 14.9 | 14.2 | 15.2 | 15.0 | 14.8 |
| | SVR - Poly | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 4.9 | 4.8 | 5.0 | 4.8 | 4.9 |
| | SVR - RBF | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 4.2 | 3.8 | 4.3 | 4.2 | 4.1 |
| | SVR - Sigmoid | 0.28 | -0.09 | 0.26 | 0.08 | 0.13 | 23.1 | 28.3 | 23.7 | 26.3 | 25.3 |
| | Gradient Boosting | 0.96 | 0.97 | 0.96 | 0.96 | 0.96 | 5.3 | 5.0 | 5.6 | 5.5 | 5.4 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 2.4 | 2.3 | 2.5 | 2.5 | 2.5 |
| Tanimoto: MACCS | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 2.6 | 2.6 | 2.8 | 2.6 | 2.7 |
| | Ridge | 1.00 | 0.99 | 1.00 | 0.99 | 0.99 | 1.4 | 3.2 | 1.7 | 3.2 | 2.4 |
| | Bayesian Ridge | 0.97 | 0.97 | 0.96 | 0.98 | 0.97 | 4.4 | 4.5 | 5.6 | 4.2 | 4.7 |
| | SVR - Poly | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.6 | 1.0 | 0.9 | 1.1 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| | SVR - Sigmoid | 0.73 | 0.72 | 0.69 | 0.71 | 0.72 | 14.1 | 14.2 | 15.2 | 14.8 | 14.6 |
| | SVR - Precomputed | 0.98 | 0.99 | 0.97 | 0.99 | 0.98 | 3.8 | 3.3 | 5.0 | 3.3 | 3.9 |
| | Gradient Boosting | 0.99 | 0.99 | 0.97 | 0.99 | 0.99 | 2.4 | 2.6 | 4.6 | 2.3 | 3.0 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 3.9 | 3.4 | 3.9 | 3.4 | 3.6 |
| Tanimoto: Morgan1 | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.0 | 1.1 | 1.0 |
| | Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.7 | 0.7 | 0.7 | 0.8 | 0.7 |

**Table B.27** Training Set Performance for the Models in the Additive Ranked Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | Mean | **1** | **2** | **3** | **4** | Mean |
| | Bayesian Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 1.2 | 1.0 | 1.2 | 1.1 |
| | SVR - Poly | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 1.2 | 1.0 | 0.9 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| | SVR - Sigmoid | 0.89 | 0.90 | 0.88 | 0.87 | 0.88 | 9.1 | 8.5 | 9.6 | 9.8 | 9.2 |
| | SVR - Precomputed | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 0.9 | 2.0 | 0.9 | 0.9 | 1.2 |
| | Gradient Boosting | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 0.8 | 2.4 | 0.8 | 0.8 | 1.2 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 4.0 | 3.6 | 4.1 | 4.2 | 4.0 |
| Tanimoto: RDK | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.6 | 1.6 | 1.6 | 1.6 | 1.6 |
| | Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| | Bayesian Ridge | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 2.0 | 2.3 | 2.3 | 2.3 | 2.2 |
| | SVR - Poly | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 0.9 | 1.0 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.0 | 0.9 | 1.0 |
| | SVR - Sigmoid | 0.84 | 0.85 | 0.84 | 0.84 | 0.84 | 10.8 | 10.3 | 11.1 | 11.0 | 10.8 |
| | SVR - Precomputed | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.2 | 1.2 | 1.0 | 1.2 | 1.2 |
| | Gradient Boosting | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 0.9 | 0.9 | 1.1 | 0.9 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 4.1 | 3.7 | 4.3 | 3.9 | 4.0 |
| WL | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.7 | 1.8 | 1.9 | 1.9 | 1.8 |
| | Ridge | 1.00 | 0.99 | 0.99 | 0.99 | 1.00 | 1.6 | 1.9 | 2.1 | 2.0 | 1.9 |
| | Bayesian Ridge | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 2.0 | 3.3 | 3.3 | 3.3 | 3.0 |
| | SVR - Poly | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 0.9 | 3.0 | 0.9 | 0.9 | 1.5 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 1.9 | 1.0 | 1.0 | 1.2 |
| | SVR - Sigmoid | 0.83 | 0.84 | 0.82 | 0.83 | 0.83 | 11.2 | 10.9 | 11.7 | 11.4 | 11.3 |
| | SVR - Precomputed | 1.00 | 0.97 | 1.00 | 1.00 | 0.99 | 1.1 | 4.4 | 1.6 | 1.6 | 2.2 |
| | Gradient Boosting | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 2.2 | 0.9 | 0.9 | 0.9 | 1.2 |
| | Random Forest | 0.98 | 0.99 | 0.99 | 0.99 | 0.98 | 3.7 | 3.2 | 3.3 | 3.2 | 3.4 |

### Test Set Performance

The poor performing models that had a mean $R^2$ value less than -1.00 and a mean $RMSE$ greater than 100% were not included in the following table.

Table B.28: Test Set Performance for the Models in the Additive Ranked Test

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | Mean | **1** | **2** | **3** | **4** | Mean |
| One-hot | Lasso | 0.55 | 0.54 | 0.64 | 0.65 | 0.59 | 18.3 | 19.1 | 16.1 | 15.8 | 17.3 |
| | Ridge | 0.56 | 0.54 | 0.64 | 0.63 | 0.59 | 18.1 | 19.1 | 16.2 | 16.2 | 17.4 |
| | Bayesian Ridge | 0.56 | 0.54 | 0.64 | 0.63 | 0.59 | 18.1 | 19.1 | 16.2 | 16.2 | 17.4 |
| | SVR - Linear | 0.56 | 0.54 | 0.64 | 0.63 | 0.59 | 18.0 | 19.1 | 16.1 | 16.2 | 17.4 |
| | SVR - Poly | 0.64 | 0.63 | 0.73 | 0.71 | 0.68 | 16.3 | 17.0 | 14.0 | 14.3 | 15.4 |
| | SVR - RBF | 0.61 | 0.61 | 0.72 | 0.71 | 0.66 | 17.1 | 17.6 | 14.2 | 14.5 | 15.9 |
| | SVR - Sigmoid | 0.47 | 0.46 | 0.52 | 0.50 | 0.49 | 19.8 | 20.6 | 18.5 | 18.8 | 19.4 |
| | Gradient Boosting | 0.64 | 0.61 | 0.74 | 0.75 | 0.68 | 16.4 | 17.5 | 13.7 | 13.4 | 15.3 |
| | Random Forest | 0.49 | 0.49 | 0.65 | 0.71 | 0.59 | 19.5 | 20.0 | 15.7 | 14.4 | 17.4 |
| Quantum Chemical | Lasso | -2.36 | 0.28 | 0.46 | 0.31 | -0.33 | 50.1 | 23.8 | 19.7 | 22.2 | 28.9 |
| | Ridge | -2.71 | -0.14 | 0.47 | -0.16 | -0.63 | 52.6 | 30.0 | 19.6 | 28.7 | 32.7 |
| | Bayesian Ridge | -1.29 | 0.32 | 0.47 | 0.15 | -0.09 | 41.3 | 23.2 | 19.5 | 24.6 | 27.2 |
| | SVR - Linear | -2.35 | -0.16 | 0.45 | -0.16 | -0.56 | 50.0 | 30.3 | 19.8 | 28.8 | 32.2 |
| | SVR - Poly | -0.25 | 0.36 | 0.63 | 0.00 | 0.18 | 30.5 | 22.5 | 16.2 | 26.7 | 24.0 |
| | SVR - RBF | 0.34 | 0.45 | 0.70 | 0.40 | 0.47 | 22.3 | 20.8 | 14.6 | 20.7 | 19.6 |
| | SVR - Sigmoid | 0.37 | 0.33 | 0.38 | 0.39 | 0.37 | 21.8 | 23.0 | 21.1 | 20.9 | 21.7 |
| | Gradient Boosting | 0.64 | 0.54 | 0.71 | 0.77 | 0.67 | 16.4 | 19.2 | 14.4 | 12.7 | 15.7 |

**Table B.28** Test Set Performance for the Models in the Additive Ranked Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | Mean | **1** | **2** | **3** | **4** | Mean |
| | Random Forest | 0.57 | 0.60 | 0.72 | 0.81 | 0.68 | 17.8 | 17.7 | 14.1 | 11.6 | 15.3 |
| Fingerprints: MACCS | Lasso | 0.39 | 0.51 | 0.06 | 0.26 | 0.31 | 21.3 | 19.7 | 26.0 | 23.0 | 22.5 |
| | Ridge | 0.36 | 0.51 | -0.02 | 0.25 | 0.27 | 21.9 | 19.7 | 27.0 | 23.1 | 22.9 |
| | Bayesian Ridge | 0.36 | 0.51 | 0.04 | 0.27 | 0.30 | 21.8 | 19.8 | 26.2 | 22.8 | 22.6 |
| | SVR - Linear | 0.34 | 0.50 | 0.08 | 0.23 | 0.29 | 22.2 | 19.9 | 25.7 | 23.4 | 22.8 |
| | SVR - Poly | 0.24 | 0.67 | 0.52 | 0.44 | 0.47 | 23.8 | 16.0 | 18.6 | 20.0 | 19.6 |
| | SVR - RBF | 0.30 | 0.67 | 0.51 | 0.45 | 0.48 | 22.8 | 16.2 | 18.7 | 19.7 | 19.4 |
| | SVR - Sigmoid | 0.16 | 0.28 | 0.29 | 0.23 | 0.24 | 25.1 | 23.8 | 22.5 | 23.5 | 23.7 |
| | Gradient Boosting | 0.31 | 0.65 | 0.76 | 0.63 | 0.59 | 22.7 | 16.6 | 13.1 | 16.2 | 17.1 |
| | Random Forest | 0.33 | 0.64 | 0.82 | 0.80 | 0.64 | 22.4 | 16.9 | 11.5 | 12.1 | 15.7 |
| Fingerprints: Morgan1 | Lasso | 0.33 | 0.37 | 0.44 | 0.66 | 0.45 | 22.3 | 22.2 | 20.0 | 15.5 | 20.0 |
| | Ridge | 0.35 | 0.49 | 0.53 | 0.64 | 0.50 | 22.1 | 20.0 | 18.4 | 16.0 | 19.1 |
| | Bayesian Ridge | 0.37 | 0.49 | 0.54 | 0.65 | 0.51 | 21.7 | 20.0 | 18.2 | 15.8 | 18.9 |
| | SVR - Linear | 0.33 | 0.48 | 0.54 | 0.64 | 0.50 | 22.3 | 20.3 | 18.2 | 16.0 | 19.2 |
| | SVR - Poly | 0.53 | 0.68 | 0.78 | 0.82 | 0.70 | 18.7 | 16.0 | 12.6 | 11.5 | 14.7 |
| | SVR - RBF | 0.51 | 0.66 | 0.77 | 0.83 | 0.69 | 19.2 | 16.4 | 13.0 | 11.0 | 14.9 |
| | SVR - Sigmoid | 0.35 | 0.43 | 0.48 | 0.45 | 0.43 | 22.1 | 21.3 | 19.3 | 19.8 | 20.6 |
| | Gradient Boosting | 0.47 | 0.48 | 0.76 | 0.78 | 0.62 | 19.9 | 20.3 | 13.2 | 12.5 | 16.4 |
| | Random Forest | 0.52 | 0.49 | 0.84 | 0.86 | 0.68 | 18.9 | 20.1 | 10.7 | 9.9 | 14.9 |
| Fingerprints: RDK | Lasso | 0.39 | 0.54 | 0.45 | 0.53 | 0.48 | 21.4 | 19.1 | 19.8 | 18.3 | 19.7 |
| | Ridge | 0.49 | 0.59 | 0.53 | 0.63 | 0.56 | 19.5 | 18.0 | 18.4 | 16.3 | 18.1 |
| | Bayesian Ridge | 0.49 | 0.59 | 0.53 | 0.63 | 0.56 | 19.5 | 18.0 | 18.4 | 16.3 | 18.0 |
| | SVR - Linear | 0.49 | 0.59 | 0.54 | 0.63 | 0.56 | 19.5 | 18.1 | 18.2 | 16.3 | 18.0 |
| | SVR - Poly | 0.54 | 0.70 | 0.56 | 0.70 | 0.62 | 18.5 | 15.4 | 17.8 | 14.7 | 16.6 |
| | SVR - RBF | 0.56 | 0.68 | 0.57 | 0.70 | 0.63 | 18.0 | 15.8 | 17.5 | 14.6 | 16.5 |
| | SVR - Sigmoid | 0.19 | 0.07 | 0.33 | 0.21 | 0.20 | 24.5 | 27.0 | 21.9 | 23.7 | 24.3 |
| | Gradient Boosting | 0.43 | 0.61 | 0.46 | 0.71 | 0.55 | 20.7 | 17.7 | 19.6 | 14.5 | 18.1 |
| | Random Forest | 0.53 | 0.51 | 0.46 | 0.61 | 0.53 | 18.7 | 19.7 | 19.8 | 16.7 | 18.7 |
| Tanimoto: MACCS | Linear Regression | 0.28 | 0.65 | 0.65 | 0.46 | 0.51 | 23.2 | 16.5 | 15.9 | 19.6 | 18.8 |
| | Lasso | 0.31 | 0.66 | 0.66 | 0.48 | 0.53 | 22.7 | 16.5 | 15.5 | 19.2 | 18.5 |
| | Ridge | 0.30 | 0.66 | 0.66 | 0.50 | 0.53 | 22.8 | 16.5 | 15.6 | 18.9 | 18.5 |
| | Bayesian Ridge | 0.38 | 0.65 | 0.68 | 0.52 | 0.56 | 21.6 | 16.5 | 15.1 | 18.6 | 17.9 |
| | SVR - Poly | 0.33 | 0.66 | 0.71 | 0.53 | 0.56 | 22.4 | 16.4 | 14.3 | 18.3 | 17.8 |
| | SVR - RBF | 0.36 | 0.66 | 0.73 | 0.56 | 0.58 | 21.9 | 16.4 | 13.9 | 17.8 | 17.5 |
| | SVR - Sigmoid | 0.40 | 0.56 | 0.65 | 0.62 | 0.56 | 21.2 | 18.6 | 15.8 | 16.5 | 18.0 |
| | SVR - Precomputed | 0.31 | 0.67 | 0.67 | 0.49 | 0.54 | 22.7 | 16.1 | 15.3 | 19.1 | 18.3 |
| | Gradient Boosting | 0.46 | 0.53 | 0.69 | 0.80 | 0.62 | 20.0 | 19.2 | 14.9 | 12.0 | 16.5 |
| | Random Forest | 0.55 | 0.50 | 0.70 | 0.77 | 0.63 | 18.4 | 19.9 | 14.5 | 12.9 | 16.4 |
| Tanimoto: Morgan1 | Linear Regression | 0.55 | 0.69 | 0.82 | 0.84 | 0.73 | 18.3 | 15.7 | 11.2 | 10.7 | 14.0 |
| | Lasso | 0.57 | 0.69 | 0.83 | 0.84 | 0.73 | 17.9 | 15.6 | 11.1 | 10.7 | 13.8 |
| | Ridge | 0.57 | 0.69 | 0.83 | 0.84 | 0.73 | 17.9 | 15.6 | 11.2 | 10.7 | 13.8 |
| | Bayesian Ridge | 0.57 | 0.69 | 0.83 | 0.84 | 0.73 | 17.9 | 15.6 | 11.1 | 10.7 | 13.8 |
| | SVR - Poly | 0.59 | 0.70 | 0.82 | 0.83 | 0.74 | 17.4 | 15.3 | 11.3 | 11.1 | 13.8 |
| | SVR - RBF | 0.60 | 0.70 | 0.81 | 0.81 | 0.73 | 17.3 | 15.4 | 11.5 | 11.5 | 13.9 |
| | SVR - Sigmoid | 0.54 | 0.66 | 0.80 | 0.82 | 0.70 | 18.5 | 16.5 | 11.9 | 11.4 | 14.6 |
| | SVR - Precomputed | 0.57 | 0.69 | 0.83 | 0.84 | 0.73 | 17.9 | 15.6 | 11.2 | 10.7 | 13.8 |
| | Gradient Boosting | 0.56 | 0.65 | 0.76 | 0.73 | 0.67 | 18.1 | 16.6 | 13.2 | 13.9 | 15.4 |
| | Random Forest | 0.55 | 0.54 | 0.78 | 0.70 | 0.64 | 18.3 | 19.1 | 12.4 | 14.7 | 16.1 |
| Tanimoto: RDK | Linear Regression | 0.58 | 0.65 | 0.62 | 0.70 | 0.64 | 17.6 | 16.6 | 16.4 | 14.7 | 16.3 |
| | Lasso | 0.59 | 0.65 | 0.63 | 0.70 | 0.64 | 17.6 | 16.5 | 16.4 | 14.6 | 16.3 |
| | Ridge | 0.58 | 0.65 | 0.62 | 0.70 | 0.64 | 17.7 | 16.5 | 16.5 | 14.5 | 16.3 |
| | Bayesian Ridge | 0.58 | 0.66 | 0.62 | 0.70 | 0.64 | 17.6 | 16.5 | 16.5 | 14.6 | 16.3 |
| | SVR - Poly | 0.58 | 0.64 | 0.62 | 0.70 | 0.64 | 17.7 | 16.8 | 16.5 | 14.5 | 16.4 |
| | SVR - RBF | 0.58 | 0.63 | 0.62 | 0.70 | 0.63 | 17.7 | 17.0 | 16.5 | 14.7 | 16.5 |
| | SVR - Sigmoid | 0.58 | 0.65 | 0.58 | 0.68 | 0.62 | 17.6 | 16.7 | 17.3 | 15.1 | 16.7 |
| | SVR - Precomputed | 0.58 | 0.65 | 0.62 | 0.70 | 0.64 | 17.7 | 16.5 | 16.5 | 14.5 | 16.3 |
| | Gradient Boosting | 0.56 | 0.54 | 0.59 | 0.59 | 0.57 | 18.1 | 19.0 | 17.1 | 17.2 | 17.8 |
| | Random Forest | 0.57 | 0.54 | 0.57 | 0.53 | 0.55 | 17.8 | 19.1 | 17.5 | 18.2 | 18.2 |

**Table B.28** Test Set Performance for the Models in the Additive Ranked Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | | RMSE (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | Mean | **1** | **2** | **3** | **4** | Mean |
| WL | Linear Regression | 0.42 | 0.68 | 0.78 | 0.81 | 0.67 | 20.9 | 15.9 | 12.7 | 11.5 | 15.2 |
| | Lasso | 0.42 | 0.68 | 0.78 | 0.81 | 0.67 | 20.8 | 15.9 | 12.7 | 11.5 | 15.2 |
| | Ridge | 0.42 | 0.68 | 0.78 | 0.81 | 0.67 | 20.8 | 15.9 | 12.6 | 11.6 | 15.2 |
| | Bayesian Ridge | 0.42 | 0.68 | 0.78 | 0.81 | 0.67 | 20.7 | 15.9 | 12.7 | 11.6 | 15.2 |
| | SVR - Poly | 0.43 | 0.66 | 0.76 | 0.81 | 0.67 | 20.6 | 16.3 | 13.0 | 11.8 | 15.4 |
| | SVR - RBF | 0.44 | 0.66 | 0.75 | 0.80 | 0.66 | 20.5 | 16.5 | 13.3 | 12.0 | 15.6 |
| | SVR - Sigmoid | 0.45 | 0.65 | 0.74 | 0.77 | 0.65 | 20.2 | 16.5 | 13.7 | 12.7 | 15.8 |
| | SVR - Precomputed | 0.42 | 0.67 | 0.78 | 0.81 | 0.67 | 20.8 | 16.0 | 12.7 | 11.5 | 15.3 |
| | Gradient Boosting | 0.50 | 0.66 | 0.76 | 0.74 | 0.67 | 19.3 | 16.3 | 13.0 | 13.5 | 15.5 |
| | Random Forest | 0.45 | 0.58 | 0.75 | 0.76 | 0.64 | 20.2 | 18.3 | 13.4 | 13.0 | 16.2 |

## B.8.2 Aryl Halide Ranked Test

### Grid Search Cross-Validated Performance

Table B.29: Grid Search Cross-Validated Performance for the Models in the Aryl Halide Ranked Test

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | Mean | **1** | **2** | **3** | Mean |
| One-hot | Linear Regression | 0.69 | 0.72 | 0.71 | 0.71 | 15.7 | 14.5 | 14.8 | 15.0 |
| | Lasso | 0.70 | 0.72 | 0.71 | 0.71 | 15.6 | 14.5 | 14.8 | 15.0 |
| | Ridge | 0.70 | 0.72 | 0.71 | 0.71 | 15.6 | 14.5 | 14.8 | 15.0 |
| | Bayesian Ridge | 0.70 | 0.72 | 0.71 | 0.71 | 15.6 | 14.5 | 14.8 | 15.0 |
| | SVR - Linear | 0.69 | 0.72 | 0.71 | 0.71 | 15.7 | 14.6 | 14.9 | 15.0 |
| | SVR - Poly | 0.89 | 0.91 | 0.91 | 0.90 | 9.5 | 8.2 | 8.4 | 8.7 |
| | SVR - RBF | 0.90 | 0.92 | 0.92 | 0.91 | 9.1 | 7.7 | 7.9 | 8.2 |
| | SVR - Sigmoid | 0.58 | 0.61 | 0.60 | 0.60 | 18.4 | 17.1 | 17.5 | 17.6 |
| | Gradient Boosting | 0.88 | 0.91 | 0.91 | 0.90 | 9.6 | 8.2 | 8.4 | 8.7 |
| | Random Forest | 0.89 | 0.91 | 0.92 | 0.90 | 9.5 | 8.1 | 8.0 | 8.5 |
| Quantum Chemical | Linear Regression | 0.69 | 0.72 | 0.71 | 0.71 | 15.7 | 14.5 | 14.9 | 15.0 |
| | Lasso | 0.69 | 0.72 | 0.70 | 0.70 | 15.8 | 14.6 | 14.9 | 15.1 |
| | Ridge | 0.69 | 0.72 | 0.71 | 0.71 | 15.7 | 14.5 | 14.8 | 15.0 |
| | Bayesian Ridge | 0.69 | 0.72 | 0.70 | 0.70 | 15.8 | 14.6 | 15.0 | 15.1 |
| | SVR - Linear | 0.69 | 0.72 | 0.70 | 0.70 | 15.8 | 14.6 | 14.9 | 15.1 |
| | SVR - Poly | 0.89 | 0.90 | 0.89 | 0.89 | 9.5 | 8.6 | 8.9 | 9.0 |
| | SVR - RBF | 0.90 | 0.90 | 0.90 | 0.90 | 9.0 | 8.7 | 8.9 | 8.8 |
| | SVR - Sigmoid | 0.47 | 0.45 | 0.48 | 0.47 | 20.6 | 20.3 | 19.9 | 20.3 |
| | Gradient Boosting | 0.91 | 0.93 | 0.93 | 0.92 | 8.7 | 7.5 | 7.5 | 7.9 |
| | Random Forest | 0.92 | 0.94 | 0.94 | 0.93 | 8.2 | 6.6 | 6.8 | 7.2 |
| Fingerprints: MACCS | Linear Regression | 0.64 | 0.62 | 0.64 | 0.63 | 17.0 | 16.9 | 16.6 | 16.8 |
| | Lasso | 0.65 | 0.65 | 0.66 | 0.66 | 16.7 | 16.1 | 16.1 | 16.3 |
| | Ridge | 0.65 | 0.65 | 0.66 | 0.66 | 16.7 | 16.1 | 16.1 | 16.3 |
| | Bayesian Ridge | 0.65 | 0.65 | 0.66 | 0.66 | 16.7 | 16.1 | 16.1 | 16.3 |
| | SVR - Linear | 0.65 | 0.65 | 0.65 | 0.65 | 16.7 | 16.2 | 16.2 | 16.4 |
| | SVR - Poly | 0.86 | 0.88 | 0.88 | 0.87 | 10.8 | 9.4 | 9.5 | 9.9 |
| | SVR - RBF | 0.85 | 0.88 | 0.88 | 0.87 | 10.9 | 9.6 | 9.6 | 10.0 |
| | SVR - Sigmoid | 0.29 | 0.27 | 0.31 | 0.29 | 23.8 | 23.5 | 22.8 | 23.4 |
| | Gradient Boosting | 0.87 | 0.90 | 0.89 | 0.89 | 10.0 | 8.5 | 9.0 | 9.2 |
| | Random Forest | 0.92 | 0.93 | 0.93 | 0.93 | 7.9 | 7.3 | 7.3 | 7.5 |
| Fingerprints: Morgan1 | Linear Regression | 0.68 | 0.70 | 0.67 | 0.68 | 16.1 | 15.0 | 15.7 | 15.6 |
| | Lasso | 0.69 | 0.71 | 0.68 | 0.69 | 15.8 | 14.8 | 15.5 | 15.4 |
| | Ridge | 0.69 | 0.71 | 0.68 | 0.69 | 15.8 | 14.8 | 15.5 | 15.4 |
| | Bayesian Ridge | 0.69 | 0.71 | 0.68 | 0.69 | 15.8 | 14.8 | 15.5 | 15.4 |
| | SVR - Linear | 0.69 | 0.71 | 0.68 | 0.69 | 15.9 | 14.8 | 15.6 | 15.4 |
| | SVR - Poly | 0.90 | 0.92 | 0.91 | 0.91 | 9.1 | 7.8 | 8.0 | 8.3 |
| | SVR - RBF | 0.91 | 0.93 | 0.92 | 0.92 | 8.5 | 7.3 | 7.5 | 7.8 |
| | SVR - Sigmoid | 0.47 | 0.47 | 0.48 | 0.47 | 20.7 | 20.0 | 19.8 | 20.1 |
| | Gradient Boosting | 0.90 | 0.92 | 0.92 | 0.91 | 9.0 | 7.9 | 7.8 | 8.2 |
| | Random Forest | 0.93 | 0.94 | 0.94 | 0.93 | 7.8 | 7.0 | 6.8 | 7.2 |
| Fingerprints: RDK | Linear Regression | 0.68 | 0.72 | 0.70 | 0.70 | 15.9 | 14.7 | 15.0 | 15.2 |
| | Lasso | 0.70 | 0.72 | 0.71 | 0.71 | 15.6 | 14.5 | 14.8 | 15.0 |
| | Ridge | 0.70 | 0.72 | 0.71 | 0.71 | 15.6 | 14.5 | 14.8 | 15.0 |
| | Bayesian Ridge | 0.70 | 0.72 | 0.71 | 0.71 | 15.6 | 14.5 | 14.8 | 15.0 |
| | SVR - Linear | 0.69 | 0.72 | 0.71 | 0.71 | 15.7 | 14.6 | 14.9 | 15.0 |
| | SVR - Poly | 0.89 | 0.92 | 0.92 | 0.91 | 9.4 | 7.8 | 7.7 | 8.3 |
| | SVR - RBF | 0.89 | 0.92 | 0.93 | 0.91 | 9.3 | 7.5 | 7.5 | 8.1 |
| | SVR - Sigmoid | 0.47 | 0.46 | 0.45 | 0.46 | 20.6 | 20.2 | 20.3 | 20.4 |
| | Gradient Boosting | 0.89 | 0.92 | 0.91 | 0.91 | 9.5 | 7.6 | 8.3 | 8.5 |
| | Random Forest | 0.92 | 0.94 | 0.93 | 0.93 | 8.2 | 6.7 | 7.2 | 7.3 |
| Tanimoto: MACCS | Linear Regression | 0.86 | 0.90 | 0.88 | 0.88 | 10.4 | 8.7 | 9.6 | 9.6 |
| | Lasso | 0.89 | 0.91 | 0.90 | 0.90 | 9.4 | 8.2 | 8.7 | 8.8 |
| | Ridge | 0.89 | 0.91 | 0.90 | 0.90 | 9.2 | 8.0 | 8.4 | 8.6 |

**Table B.29** Grid Search Cross-Validated Performance for the Models in the Aryl Halide Ranked Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | Mean | **1** | **2** | **3** | Mean |
| | Bayesian Ridge | 0.88 | 0.91 | 0.90 | 0.90 | 9.7 | 8.5 | 8.7 | 9.0 |
| | SVR - Poly | 0.89 | 0.91 | 0.89 | 0.90 | 9.3 | 8.3 | 9.0 | 8.9 |
| | SVR - RBF | 0.88 | 0.89 | 0.88 | 0.88 | 9.8 | 8.9 | 9.4 | 9.4 |
| | SVR - Sigmoid | 0.68 | 0.68 | 0.70 | 0.69 | 16.0 | 15.5 | 15.0 | 15.5 |
| | SVR - Precomputed | 0.88 | 0.91 | 0.90 | 0.89 | 9.9 | 8.3 | 8.7 | 9.0 |
| | Gradient Boosting | 0.75 | 0.77 | 0.79 | 0.77 | 14.1 | 13.1 | 12.7 | 13.3 |
| | Random Forest | 0.72 | 0.71 | 0.74 | 0.72 | 15.0 | 14.7 | 14.0 | 14.6 |
| Tanimoto: Morgan1 | Linear Regression | 0.92 | 0.94 | 0.93 | 0.93 | 7.7 | 6.9 | 7.0 | 7.2 |
| | Lasso | 0.93 | 0.94 | 0.94 | 0.93 | 7.6 | 6.8 | 6.9 | 7.1 |
| | Ridge | 0.93 | 0.94 | 0.94 | 0.94 | 7.5 | 6.8 | 6.9 | 7.1 |
| | Bayesian Ridge | 0.93 | 0.94 | 0.94 | 0.94 | 7.5 | 6.8 | 6.9 | 7.0 |
| | SVR - Poly | 0.92 | 0.92 | 0.92 | 0.92 | 8.0 | 7.5 | 7.6 | 7.7 |
| | SVR - RBF | 0.91 | 0.91 | 0.91 | 0.91 | 8.7 | 8.3 | 8.3 | 8.4 |
| | SVR - Sigmoid | 0.87 | 0.89 | 0.89 | 0.88 | 10.1 | 9.1 | 9.2 | 9.4 |
| | SVR - Precomputed | 0.93 | 0.94 | 0.94 | 0.93 | 7.5 | 6.8 | 6.9 | 7.1 |
| | Gradient Boosting | 0.81 | 0.81 | 0.82 | 0.81 | 12.3 | 12.0 | 11.7 | 12.0 |
| | Random Forest | 0.72 | 0.73 | 0.78 | 0.74 | 14.9 | 14.3 | 13.0 | 14.1 |
| Tanimoto: RDK | Linear Regression | 0.89 | 0.92 | 0.91 | 0.91 | 9.4 | 7.9 | 8.0 | 8.5 |
| | Lasso | 0.90 | 0.92 | 0.92 | 0.91 | 9.1 | 7.7 | 7.8 | 8.2 |
| | Ridge | 0.90 | 0.92 | 0.92 | 0.91 | 9.0 | 7.6 | 7.7 | 8.1 |
| | Bayesian Ridge | 0.90 | 0.92 | 0.92 | 0.92 | 8.9 | 7.7 | 7.6 | 8.1 |
| | SVR - Poly | 0.89 | 0.91 | 0.91 | 0.90 | 9.4 | 8.3 | 8.4 | 8.7 |
| | SVR - RBF | 0.88 | 0.89 | 0.89 | 0.89 | 9.9 | 8.9 | 8.9 | 9.2 |
| | SVR - Sigmoid | 0.84 | 0.86 | 0.85 | 0.85 | 11.2 | 10.2 | 10.5 | 10.7 |
| | SVR - Precomputed | 0.90 | 0.92 | 0.92 | 0.91 | 9.1 | 7.7 | 7.8 | 8.2 |
| | Gradient Boosting | 0.80 | 0.80 | 0.81 | 0.80 | 12.8 | 12.1 | 12.0 | 12.3 |
| | Random Forest | 0.71 | 0.73 | 0.75 | 0.73 | 15.2 | 14.3 | 13.8 | 14.4 |
| WL | Linear Regression | 0.91 | 0.93 | 0.93 | 0.92 | 8.6 | 7.4 | 7.4 | 7.8 |
| | Lasso | 0.91 | 0.93 | 0.93 | 0.92 | 8.4 | 7.2 | 7.1 | 7.6 |
| | Ridge | 0.91 | 0.93 | 0.93 | 0.93 | 8.3 | 7.1 | 7.0 | 7.5 |
| | Bayesian Ridge | 0.91 | 0.93 | 0.93 | 0.93 | 8.3 | 7.2 | 7.0 | 7.5 |
| | SVR - Poly | 0.91 | 0.93 | 0.93 | 0.92 | 8.4 | 7.3 | 7.2 | 7.6 |
| | SVR - RBF | 0.91 | 0.92 | 0.93 | 0.92 | 8.6 | 7.5 | 7.4 | 7.8 |
| | SVR - Sigmoid | 0.82 | 0.84 | 0.83 | 0.83 | 11.9 | 10.9 | 11.3 | 11.4 |
| | SVR - Precomputed | 0.91 | 0.93 | 0.93 | 0.92 | 8.5 | 7.2 | 7.2 | 7.6 |
| | Gradient Boosting | 0.85 | 0.86 | 0.86 | 0.86 | 11.1 | 10.1 | 10.4 | 10.5 |
| | Random Forest | 0.79 | 0.79 | 0.80 | 0.80 | 13.0 | 12.4 | 12.1 | 12.5 |

### Training Set Performance

Table B.30: Training Set Performance for the Models in the Aryl Halide Ranked Test

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | Mean | **1** | **2** | **3** | Mean |
| One-hot | Linear Regression | 0.70 | 0.73 | 0.72 | 0.72 | 15.4 | 14.3 | 14.7 | 14.8 |
| | Lasso | 0.70 | 0.73 | 0.72 | 0.72 | 15.4 | 14.3 | 14.7 | 14.8 |
| | Ridge | 0.70 | 0.73 | 0.72 | 0.72 | 15.4 | 14.3 | 14.7 | 14.8 |
| | Bayesian Ridge | 0.70 | 0.73 | 0.72 | 0.72 | 15.4 | 14.3 | 14.7 | 14.8 |
| | SVR - Linear | 0.70 | 0.73 | 0.71 | 0.72 | 15.5 | 14.3 | 14.7 | 14.8 |
| | SVR - Poly | 0.97 | 0.97 | 0.98 | 0.97 | 5.0 | 4.5 | 4.3 | 4.6 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.0 | 1.0 |
| | SVR - Sigmoid | 0.59 | 0.62 | 0.60 | 0.60 | 18.2 | 16.9 | 17.3 | 17.5 |
| | Gradient Boosting | 0.95 | 0.96 | 0.96 | 0.96 | 6.2 | 5.5 | 5.5 | 5.8 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 3.3 | 2.9 | 2.7 | 3.0 |

**Table B.30** Training Set Performance for the Models in the Aryl Halide Ranked Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | Mean | **1** | **2** | **3** | Mean |
| Quantum Chemical | Linear Regression | 0.60 | 0.73 | 0.71 | 0.68 | 17.9 | 14.3 | 14.7 | 15.6 |
| | Lasso | 0.70 | 0.73 | 0.71 | 0.71 | 15.6 | 14.4 | 14.8 | 14.9 |
| | Ridge | 0.70 | 0.73 | 0.72 | 0.72 | 15.5 | 14.3 | 14.7 | 14.8 |
| | Bayesian Ridge | 0.70 | 0.73 | 0.71 | 0.71 | 15.6 | 14.4 | 14.8 | 14.9 |
| | SVR - Linear | 0.70 | 0.73 | 0.71 | 0.71 | 15.5 | 14.3 | 14.7 | 14.9 |
| | SVR - Poly | 0.95 | 0.96 | 0.96 | 0.96 | 6.2 | 5.4 | 5.5 | 5.7 |
| | SVR - RBF | 0.98 | 0.98 | 0.98 | 0.98 | 4.0 | 3.6 | 3.5 | 3.7 |
| | SVR - Sigmoid | 0.44 | 0.43 | 0.46 | 0.44 | 21.2 | 20.9 | 20.2 | 20.7 |
| | Gradient Boosting | 0.97 | 0.97 | 0.97 | 0.97 | 5.2 | 4.9 | 4.9 | 5.0 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.7 | 2.3 | 2.3 | 2.4 |
| Fingerprints: MACCS | Linear Regression | 0.64 | 0.65 | 0.65 | 0.65 | 17.1 | 16.3 | 16.2 | 16.5 |
| | Lasso | 0.66 | 0.67 | 0.66 | 0.66 | 16.5 | 15.9 | 15.9 | 16.1 |
| | Ridge | 0.66 | 0.67 | 0.66 | 0.66 | 16.5 | 15.9 | 15.9 | 16.1 |
| | Bayesian Ridge | 0.66 | 0.67 | 0.66 | 0.66 | 16.5 | 15.9 | 15.9 | 16.1 |
| | SVR - Linear | 0.66 | 0.66 | 0.66 | 0.66 | 16.5 | 15.9 | 16.0 | 16.2 |
| | SVR - Poly | 0.90 | 0.92 | 0.92 | 0.91 | 9.0 | 7.6 | 7.8 | 8.1 |
| | SVR - RBF | 0.90 | 0.92 | 0.92 | 0.91 | 9.1 | 7.6 | 7.7 | 8.2 |
| | SVR - Sigmoid | 0.33 | 0.30 | 0.35 | 0.33 | 23.2 | 23.0 | 22.2 | 22.8 |
| | Gradient Boosting | 0.92 | 0.93 | 0.93 | 0.93 | 8.0 | 7.0 | 7.1 | 7.4 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.7 | 2.5 | 2.4 | 2.6 |
| Fingerprints: Morgan1 | Linear Regression | 0.69 | 0.71 | 0.69 | 0.70 | 15.8 | 14.8 | 15.4 | 15.3 |
| | Lasso | 0.70 | 0.72 | 0.69 | 0.70 | 15.6 | 14.6 | 15.4 | 15.2 |
| | Ridge | 0.70 | 0.72 | 0.69 | 0.70 | 15.6 | 14.6 | 15.4 | 15.2 |
| | Bayesian Ridge | 0.70 | 0.72 | 0.69 | 0.70 | 15.6 | 14.6 | 15.4 | 15.2 |
| | SVR - Linear | 0.70 | 0.72 | 0.69 | 0.70 | 15.7 | 14.6 | 15.4 | 15.2 |
| | SVR - Poly | 0.95 | 0.96 | 0.96 | 0.96 | 6.2 | 5.4 | 5.6 | 5.7 |
| | SVR - RBF | 0.97 | 0.98 | 0.97 | 0.97 | 5.2 | 4.3 | 4.6 | 4.7 |
| | SVR - Sigmoid | 0.50 | 0.51 | 0.51 | 0.51 | 20.0 | 19.4 | 19.3 | 19.6 |
| | Gradient Boosting | 0.95 | 0.96 | 0.96 | 0.95 | 6.5 | 5.8 | 5.8 | 6.0 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.6 | 2.4 | 2.3 | 2.4 |
| Fingerprints: RDK | Linear Regression | 0.70 | 0.73 | 0.71 | 0.71 | 15.5 | 14.3 | 14.8 | 14.9 |
| | Lasso | 0.70 | 0.73 | 0.72 | 0.72 | 15.4 | 14.3 | 14.7 | 14.8 |
| | Ridge | 0.70 | 0.73 | 0.72 | 0.72 | 15.4 | 14.3 | 14.7 | 14.8 |
| | Bayesian Ridge | 0.70 | 0.73 | 0.72 | 0.72 | 15.4 | 14.3 | 14.7 | 14.8 |
| | SVR - Linear | 0.70 | 0.73 | 0.71 | 0.72 | 15.5 | 14.3 | 14.7 | 14.8 |
| | SVR - Poly | 0.96 | 0.97 | 0.97 | 0.97 | 5.5 | 4.8 | 4.6 | 4.9 |
| | SVR - RBF | 0.97 | 0.98 | 0.98 | 0.98 | 4.7 | 3.8 | 3.9 | 4.1 |
| | SVR - Sigmoid | 0.28 | 0.17 | 0.27 | 0.24 | 24.1 | 25.1 | 23.5 | 24.2 |
| | Gradient Boosting | 0.96 | 0.97 | 0.96 | 0.96 | 5.8 | 5.0 | 5.2 | 5.3 |
| | Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 2.7 | 2.3 | 2.4 | 2.5 |
| Tanimoto: MACCS | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 0.99 | 0.99 | 0.99 | 0.99 | 2.8 | 2.4 | 2.4 | 2.5 |
| | Ridge | 1.00 | 0.99 | 0.99 | 0.99 | 1.9 | 3.1 | 3.1 | 2.7 |
| | Bayesian Ridge | 0.96 | 0.97 | 0.97 | 0.97 | 5.8 | 4.4 | 4.6 | 5.0 |
| | SVR - Poly | 1.00 | 1.00 | 0.99 | 1.00 | 1.0 | 1.0 | 2.0 | 1.3 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.2 | 1.1 |
| | SVR - Sigmoid | 0.70 | 0.72 | 0.73 | 0.72 | 15.4 | 14.7 | 14.3 | 14.8 |
| | SVR - Precomputed | 0.97 | 0.99 | 0.98 | 0.98 | 5.2 | 3.2 | 3.7 | 4.0 |
| | Gradient Boosting | 0.94 | 0.97 | 0.99 | 0.97 | 6.9 | 4.4 | 3.1 | 4.8 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 4.1 | 3.7 | 3.7 | 3.8 |
| Tanimoto: Morgan1 | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 0.9 | 1.0 | 0.9 |
| | Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 0.8 | 0.6 | 0.6 | 0.7 |
| | Bayesian Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 1.4 | 0.8 | 0.8 | 1.0 |
| | SVR - Poly | 1.00 | 1.00 | 1.00 | 1.00 | 1.2 | 0.9 | 0.9 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 1.0 | 1.0 |
| | SVR - Sigmoid | 0.87 | 0.90 | 0.89 | 0.89 | 10.1 | 8.7 | 9.2 | 9.4 |
| | SVR - Precomputed | 0.99 | 1.00 | 1.00 | 1.00 | 2.3 | 0.9 | 0.9 | 1.4 |
| | Gradient Boosting | 1.00 | 1.00 | 0.99 | 1.00 | 0.8 | 0.7 | 2.4 | 1.3 |

**Table B.30** Training Set Performance for the Models in the Aryl Halide Ranked Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | Mean | **1** | **2** | **3** | Mean |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 4.2 | 4.0 | 4.1 | 4.1 |
| Tanimoto: RDK | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 1.5 | 1.5 | 1.5 | 1.5 |
| | Ridge | 1.00 | 1.00 | 1.00 | 1.00 | 1.7 | 1.5 | 1.5 | 1.5 |
| | Bayesian Ridge | 0.99 | 0.99 | 0.99 | 0.99 | 2.8 | 2.2 | 2.2 | 2.4 |
| | SVR - Poly | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 0.9 | 1.0 | 1.0 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 | 1.0 | 0.9 | 1.0 |
| | SVR - Sigmoid | 0.84 | 0.85 | 0.81 | 0.83 | 11.5 | 10.8 | 11.9 | 11.4 |
| | SVR - Precomputed | 1.00 | 1.00 | 1.00 | 1.00 | 1.2 | 1.0 | 1.2 | 1.2 |
| | Gradient Boosting | 1.00 | 1.00 | 1.00 | 1.00 | 0.8 | 0.8 | 0.8 | 0.8 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 3.9 | 3.8 | 3.9 | 3.9 |
| WL | Linear Regression | 1.00 | 1.00 | 1.00 | 1.00 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Lasso | 1.00 | 1.00 | 1.00 | 1.00 | 1.6 | 1.7 | 1.8 | 1.7 |
| | Ridge | 0.99 | 1.00 | 1.00 | 1.00 | 2.1 | 1.9 | 1.9 | 1.9 |
| | Bayesian Ridge | 0.98 | 0.99 | 0.99 | 0.99 | 3.7 | 3.1 | 2.7 | 3.2 |
| | SVR - Poly | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 0.9 | 0.9 | 0.9 |
| | SVR - RBF | 1.00 | 1.00 | 1.00 | 1.00 | 0.9 | 0.9 | 0.9 | 0.9 |
| | SVR - Sigmoid | 0.82 | 0.81 | 0.83 | 0.82 | 11.9 | 11.9 | 11.2 | 11.7 |
| | SVR - Precomputed | 1.00 | 1.00 | 1.00 | 1.00 | 1.7 | 1.5 | 1.6 | 1.6 |
| | Gradient Boosting | 1.00 | 1.00 | 1.00 | 1.00 | 0.3 | 0.8 | 0.3 | 0.5 |
| | Random Forest | 0.98 | 0.98 | 0.98 | 0.98 | 3.6 | 3.4 | 3.6 | 3.5 |

### Test Set Performance

Table B.31: Test Set Performance for the Models in the Aryl Halide Ranked Test

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | Mean | **1** | **2** | **3** | Mean |
| One-hot | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | 0.17 | 0.30 | 0.37 | 0.28 | 22.6 | 22.3 | 21.1 | 22.0 |
| | Ridge | 0.25 | 0.31 | 0.36 | 0.31 | 21.5 | 22.1 | 21.2 | 21.6 |
| | Bayesian Ridge | 0.25 | 0.31 | 0.36 | 0.31 | 21.5 | 22.1 | 21.2 | 21.6 |
| | SVR - Linear | 0.25 | 0.30 | 0.36 | 0.31 | 21.5 | 22.2 | 21.2 | 21.6 |
| | SVR - Poly | 0.32 | 0.35 | 0.39 | 0.35 | 20.6 | 21.5 | 20.6 | 20.9 |
| | SVR - RBF | 0.25 | 0.35 | 0.43 | 0.34 | 21.6 | 21.5 | 20.0 | 21.0 |
| | SVR - Sigmoid | 0.21 | 0.26 | 0.27 | 0.24 | 22.2 | 23.0 | 22.7 | 22.6 |
| | Gradient Boosting | 0.28 | 0.34 | 0.41 | 0.34 | 21.2 | 21.7 | 20.3 | 21.1 |
| | Random Forest | -0.40 | 0.01 | 0.26 | -0.04 | 29.5 | 26.5 | 22.8 | 26.2 |
| Quantum Chemical | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | <-1.00 | -0.60 | -0.97 | <-1.00 | >100.0 | 33.7 | 37.1 | 73.8 |
| | Ridge | <-1.00 | 0.53 | 0.39 | <-1.00 | >100.0 | 18.2 | 20.6 | >100.0 |
| | Bayesian Ridge | <-1.00 | 0.52 | 0.39 | <-1.00 | >100.0 | 18.4 | 20.8 | >100.0 |
| | SVR - Linear | <-1.00 | 0.52 | 0.43 | <-1.00 | >100.0 | 18.5 | 20.0 | >100.0 |
| | SVR - Poly | <-1.00 | 0.36 | 0.03 | <-1.00 | 94.5 | 21.3 | 26.1 | 47.3 |
| | SVR - RBF | 0.48 | 0.50 | 0.25 | 0.41 | 18.0 | 18.8 | 22.9 | 19.9 |
| | SVR - Sigmoid | -0.25 | 0.23 | 0.14 | 0.04 | 27.8 | 23.3 | 24.5 | 25.2 |
| | Gradient Boosting | 0.28 | 0.20 | 0.30 | 0.26 | 21.2 | 23.8 | 22.2 | 22.4 |
| | Random Forest | 0.01 | 0.26 | 0.32 | 0.20 | 24.8 | 22.9 | 21.8 | 23.2 |
| Fingerprints: MACCS | Linear Regression | <-1.00 | 0.49 | 0.49 | <-1.00 | >100.0 | 19.0 | 18.9 | >100.0 |
| | Lasso | 0.50 | 0.54 | 0.50 | 0.51 | 17.5 | 18.2 | 18.8 | 18.2 |
| | Ridge | 0.52 | 0.54 | 0.50 | 0.52 | 17.3 | 18.2 | 18.8 | 18.1 |
| | Bayesian Ridge | 0.52 | 0.54 | 0.50 | 0.52 | 17.3 | 18.1 | 18.8 | 18.1 |
| | SVR - Linear | 0.52 | 0.51 | 0.51 | 0.52 | 17.2 | 18.6 | 18.6 | 18.1 |
| | SVR - Poly | 0.69 | 0.36 | 0.61 | 0.55 | 13.8 | 21.3 | 16.6 | 17.2 |

**Table B.31** Test Set Performance for the Models in the Aryl Halide Ranked Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | Mean | **1** | **2** | **3** | Mean |
| | SVR - RBF | 0.68 | 0.38 | 0.63 | 0.56 | 14.1 | 20.9 | 16.2 | 17.1 |
| | SVR - Sigmoid | 0.16 | 0.31 | 0.18 | 0.21 | 22.9 | 22.2 | 24.0 | 23.0 |
| | Gradient Boosting | 0.64 | 0.67 | 0.64 | 0.65 | 14.9 | 15.4 | 16.0 | 15.4 |
| | Random Forest | 0.55 | 0.69 | 0.01 | 0.42 | 16.6 | 14.8 | 26.3 | 19.3 |
| Fingerprints: Morgan1 | Linear Regression | <-1.00 | 0.48 | <-1.00 | <-1.00 | >100.0 | 19.3 | >100.0 | >100.0 |
| | Lasso | 0.58 | 0.50 | 0.53 | 0.54 | 16.2 | 18.8 | 18.2 | 17.7 |
| | Ridge | 0.60 | 0.50 | 0.54 | 0.55 | 15.8 | 18.8 | 17.9 | 17.5 |
| | Bayesian Ridge | 0.60 | 0.50 | 0.54 | 0.55 | 15.8 | 18.8 | 17.9 | 17.5 |
| | SVR - Linear | 0.59 | 0.50 | 0.55 | 0.55 | 15.9 | 18.8 | 17.7 | 17.5 |
| | SVR - Poly | 0.74 | 0.68 | 0.64 | 0.69 | 12.7 | 15.1 | 15.9 | 14.6 |
| | SVR - RBF | 0.73 | 0.67 | 0.64 | 0.68 | 12.9 | 15.2 | 15.8 | 14.6 |
| | SVR - Sigmoid | 0.39 | 0.45 | 0.31 | 0.39 | 19.4 | 19.7 | 22.0 | 20.4 |
| | Gradient Boosting | 0.65 | 0.64 | 0.66 | 0.65 | 14.7 | 15.9 | 15.4 | 15.4 |
| | Random Forest | 0.61 | 0.71 | -0.02 | 0.43 | 15.6 | 14.4 | 26.7 | 18.9 |
| Fingerprints: RDK | Linear Regression | <-1.00 | <-1.00 | <-1.00 | <-1.00 | >100.0 | >100.0 | >100.0 | >100.0 |
| | Lasso | 0.57 | 0.55 | 0.25 | 0.46 | 16.4 | 18.0 | 22.9 | 19.1 |
| | Ridge | 0.55 | 0.58 | 0.48 | 0.54 | 16.6 | 17.3 | 19.2 | 17.7 |
| | Bayesian Ridge | 0.55 | 0.58 | 0.48 | 0.54 | 16.7 | 17.2 | 19.2 | 17.7 |
| | SVR - Linear | 0.56 | 0.57 | 0.48 | 0.54 | 16.6 | 17.5 | 19.1 | 17.7 |
| | SVR - Poly | 0.68 | 0.73 | 0.53 | 0.64 | 14.2 | 13.9 | 18.3 | 15.4 |
| | SVR - RBF | 0.67 | 0.74 | 0.53 | 0.65 | 14.4 | 13.6 | 18.2 | 15.4 |
| | SVR - Sigmoid | 0.13 | 0.02 | 0.10 | 0.09 | 23.2 | 26.3 | 25.1 | 24.9 |
| | Gradient Boosting | 0.74 | 0.69 | 0.03 | 0.49 | 12.7 | 14.7 | 26.1 | 17.9 |
| | Random Forest | 0.77 | 0.71 | 0.62 | 0.70 | 12.0 | 14.4 | 16.3 | 14.2 |
| Tanimoto: MACCS | Linear Regression | 0.69 | 0.43 | 0.66 | 0.60 | 13.8 | 20.2 | 15.3 | 16.4 |
| | Lasso | 0.69 | 0.46 | 0.67 | 0.61 | 14.0 | 19.7 | 15.1 | 16.3 |
| | Ridge | 0.66 | 0.47 | 0.67 | 0.60 | 14.5 | 19.4 | 15.2 | 16.4 |
| | Bayesian Ridge | 0.65 | 0.49 | 0.67 | 0.60 | 14.8 | 19.0 | 15.3 | 16.3 |
| | SVR - Poly | 0.59 | 0.53 | 0.63 | 0.58 | 16.0 | 18.2 | 16.2 | 16.8 |
| | SVR - RBF | 0.54 | 0.57 | 0.62 | 0.58 | 17.0 | 17.6 | 16.2 | 16.9 |
| | SVR - Sigmoid | 0.49 | 0.65 | 0.46 | 0.54 | 17.8 | 15.7 | 19.4 | 17.6 |
| | SVR - Precomputed | 0.66 | 0.42 | 0.63 | 0.57 | 14.6 | 20.2 | 16.1 | 16.9 |
| | Gradient Boosting | 0.60 | 0.64 | 0.51 | 0.58 | 15.8 | 16.0 | 18.5 | 16.8 |
| | Random Forest | 0.56 | 0.44 | 0.43 | 0.48 | 16.6 | 20.0 | 19.9 | 18.8 |
| Tanimoto: Morgan1 | Linear Regression | 0.66 | 0.73 | 0.63 | 0.68 | 14.5 | 13.7 | 16.1 | 14.8 |
| | Lasso | 0.66 | 0.74 | 0.61 | 0.67 | 14.4 | 13.7 | 16.5 | 14.9 |
| | Ridge | 0.65 | 0.74 | 0.61 | 0.67 | 14.7 | 13.7 | 16.5 | 15.0 |
| | Bayesian Ridge | 0.65 | 0.74 | 0.61 | 0.67 | 14.7 | 13.7 | 16.5 | 15.0 |
| | SVR - Poly | 0.59 | 0.69 | 0.59 | 0.62 | 16.0 | 14.8 | 17.1 | 16.0 |
| | SVR - RBF | 0.55 | 0.67 | 0.56 | 0.59 | 16.7 | 15.4 | 17.5 | 16.5 |
| | SVR - Sigmoid | 0.67 | 0.72 | 0.60 | 0.66 | 14.3 | 14.1 | 16.8 | 15.1 |
| | SVR - Precomputed | 0.65 | 0.73 | 0.61 | 0.67 | 14.7 | 13.7 | 16.5 | 15.0 |
| | Gradient Boosting | 0.55 | 0.38 | 0.40 | 0.44 | 16.7 | 21.0 | 20.5 | 19.4 |
| | Random Forest | 0.49 | 0.29 | -0.10 | 0.23 | 17.8 | 22.5 | 27.8 | 22.7 |
| Tanimoto: RDK | Linear Regression | 0.52 | 0.72 | 0.49 | 0.58 | 17.3 | 14.1 | 18.9 | 16.8 |
| | Lasso | 0.53 | 0.72 | 0.49 | 0.58 | 17.2 | 14.1 | 18.9 | 16.7 |
| | Ridge | 0.51 | 0.72 | 0.49 | 0.57 | 17.4 | 14.1 | 18.9 | 16.8 |
| | Bayesian Ridge | 0.51 | 0.72 | 0.49 | 0.57 | 17.5 | 14.1 | 18.9 | 16.8 |
| | SVR - Poly | 0.42 | 0.66 | 0.43 | 0.50 | 18.9 | 15.6 | 20.0 | 18.2 |
| | SVR - RBF | 0.39 | 0.62 | 0.40 | 0.47 | 19.5 | 16.4 | 20.5 | 18.8 |
| | SVR - Sigmoid | 0.52 | 0.72 | 0.49 | 0.58 | 17.2 | 14.2 | 18.9 | 16.8 |
| | SVR - Precomputed | 0.51 | 0.72 | 0.49 | 0.57 | 17.4 | 14.1 | 18.9 | 16.8 |
| | Gradient Boosting | 0.28 | 0.68 | 0.01 | 0.32 | 21.2 | 15.1 | 26.4 | 20.9 |
| | Random Forest | 0.28 | 0.60 | -0.79 | 0.03 | 21.1 | 16.8 | 35.4 | 24.5 |
| WL | Linear Regression | 0.61 | 0.71 | 0.60 | 0.64 | 15.6 | 14.3 | 16.8 | 15.6 |
| | Lasso | 0.61 | 0.71 | 0.60 | 0.64 | 15.7 | 14.4 | 16.8 | 15.6 |
| | Ridge | 0.60 | 0.70 | 0.60 | 0.63 | 15.7 | 14.5 | 16.8 | 15.7 |
| | Bayesian Ridge | 0.60 | 0.70 | 0.60 | 0.63 | 15.8 | 14.6 | 16.9 | 15.8 |
| | SVR - Poly | 0.57 | 0.66 | 0.57 | 0.60 | 16.3 | 15.5 | 17.4 | 16.4 |

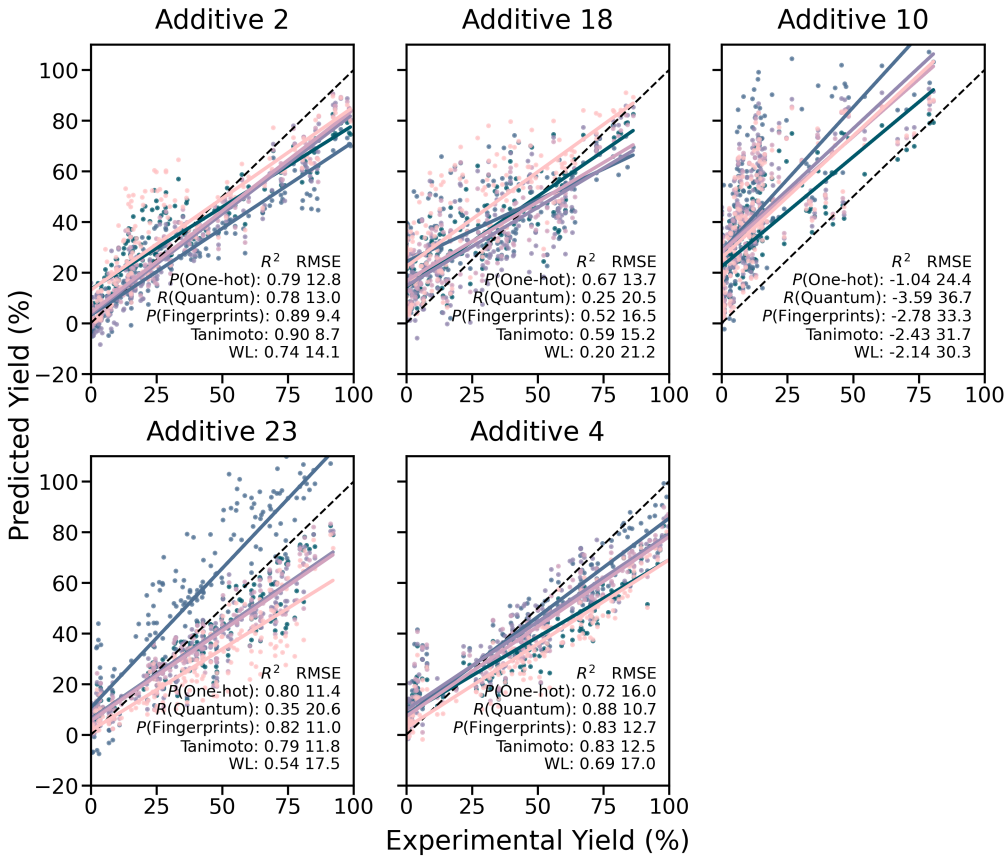
**Table B.31** Test Set Performance for the Models in the Aryl Halide Ranked Test (Continued)

| Descriptor | ML Algorithm | $R^2$ | | | | RMSE (%) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | Mean | **1** | **2** | **3** | Mean |
| | SVR - RBF | 0.55 | 0.64 | 0.55 | 0.58 | 16.6 | 16.1 | 17.7 | 16.8 |
| | SVR - Sigmoid | 0.58 | 0.68 | 0.57 | 0.61 | 16.2 | 15.2 | 17.3 | 16.2 |
| | SVR - Precomputed | 0.60 | 0.70 | 0.60 | 0.63 | 15.7 | 14.5 | 16.8 | 15.7 |
| | Gradient Boosting | 0.21 | 0.29 | 0.53 | 0.34 | 22.1 | 22.4 | 18.2 | 20.9 |
| | Random Forest | -0.12 | 0.09 | -0.01 | -0.01 | 26.4 | 25.4 | 26.7 | 26.1 |

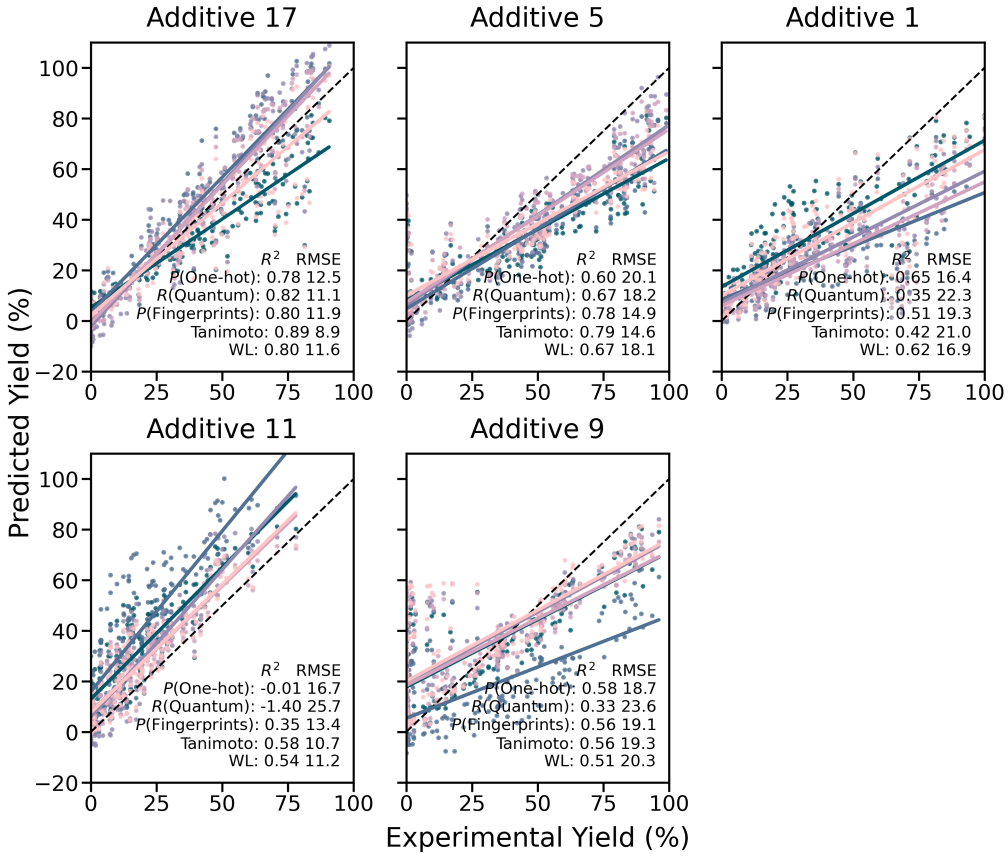
### B.8.3 Domain of Applicability

#### Additive Ranked Tests

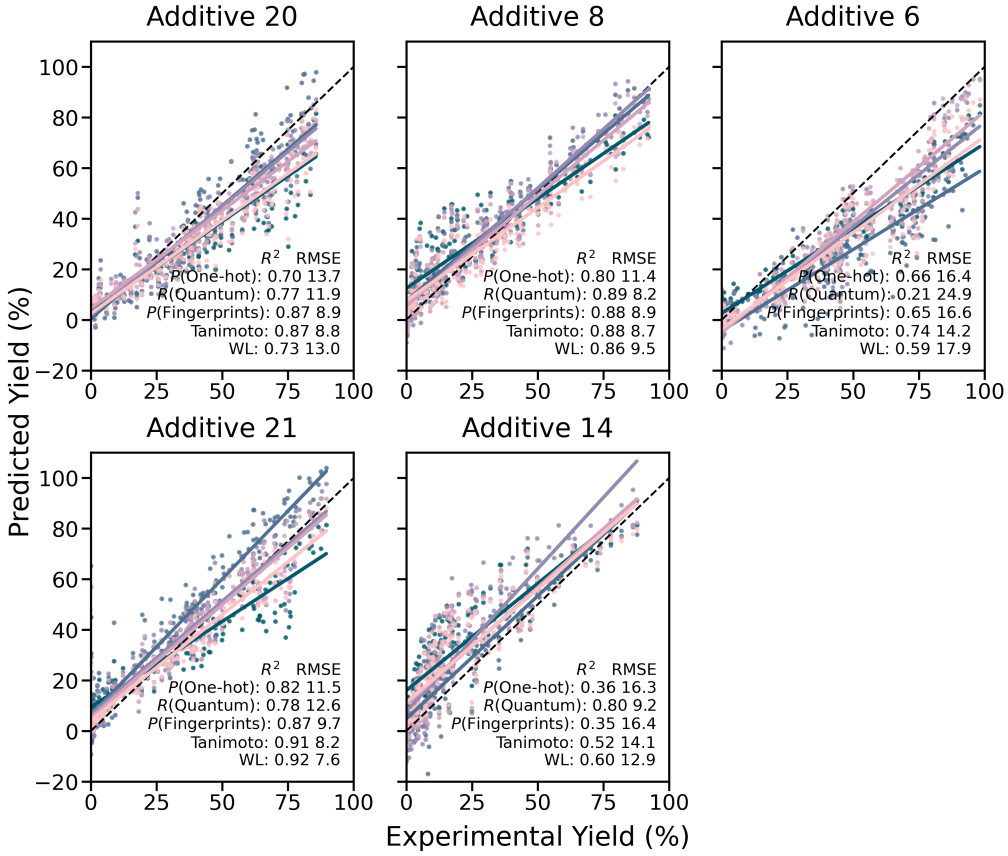
(a)



(b)


Additive 17
$R^2$ RMSE
P(One-hot): 0.78 12.5
R(Quantum): 0.82 11.1
P(Fingerprints): 0.80 11.9
Tanimoto: 0.89 8.9
WL: 0.80 11.6
Additive 5
$R^2$ RMSE
P(One-hot): 0.60 20.1
R(Quantum): 0.67 18.2
P(Fingerprints): 0.78 14.9
Tanimoto: 0.79 14.6
WL: 0.67 18.1
Additive 1
$R^2$ RMSE
P(One-hot): 0.65 16.4
R(Quantum): 0.35 22.3
P(Fingerprints): 0.51 19.3
Tanimoto: 0.42 21.0
WL: 0.62 16.9
Additive 11
$R^2$ RMSE
P(One-hot): -0.01 16.7
R(Quantum): -1.40 25.7
P(Fingerprints): 0.35 13.4
Tanimoto: 0.58 10.7
WL: 0.54 11.2
Additive 9
$R^2$ RMSE
P(One-hot): 0.58 18.7
R(Quantum): 0.33 23.6
P(Fingerprints): 0.56 19.1
Tanimoto: 0.56 19.3
WL: 0.51 20.3
Predicted Yield (%)
Experimental Yield (%)

(c)


Additive 20
$R^2$ RMSE
P(One-hot): 0.70 13.7
R(Quantum): 0.77 11.9
P(Fingerprints): 0.87 8.9
Tanimoto: 0.87 8.8
WL: 0.73 13.0
Additive 8
$R^2$ RMSE
P(One-hot): 0.80 11.4
R(Quantum): 0.89 8.2
P(Fingerprints): 0.88 8.9
Tanimoto: 0.88 8.7
WL: 0.86 9.5
Additive 6
$R^2$ RMSE
P(One-hot): 0.66 16.4
R(Quantum): 0.21 24.9
P(Fingerprints): 0.65 16.6
Tanimoto: 0.74 14.2
WL: 0.59 17.9
Additive 21
$R^2$ RMSE
P(One-hot): 0.82 11.5
R(Quantum): 0.78 12.6
P(Fingerprints): 0.87 9.7
Tanimoto: 0.91 8.2
WL: 0.92 7.6
Additive 14
$R^2$ RMSE
P(One-hot): 0.36 16.3
R(Quantum): 0.80 9.2
P(Fingerprints): 0.35 16.4
Tanimoto: 0.52 14.1
WL: 0.60 12.9
Predicted Yield (%)
Experimental Yield (%)

(d)

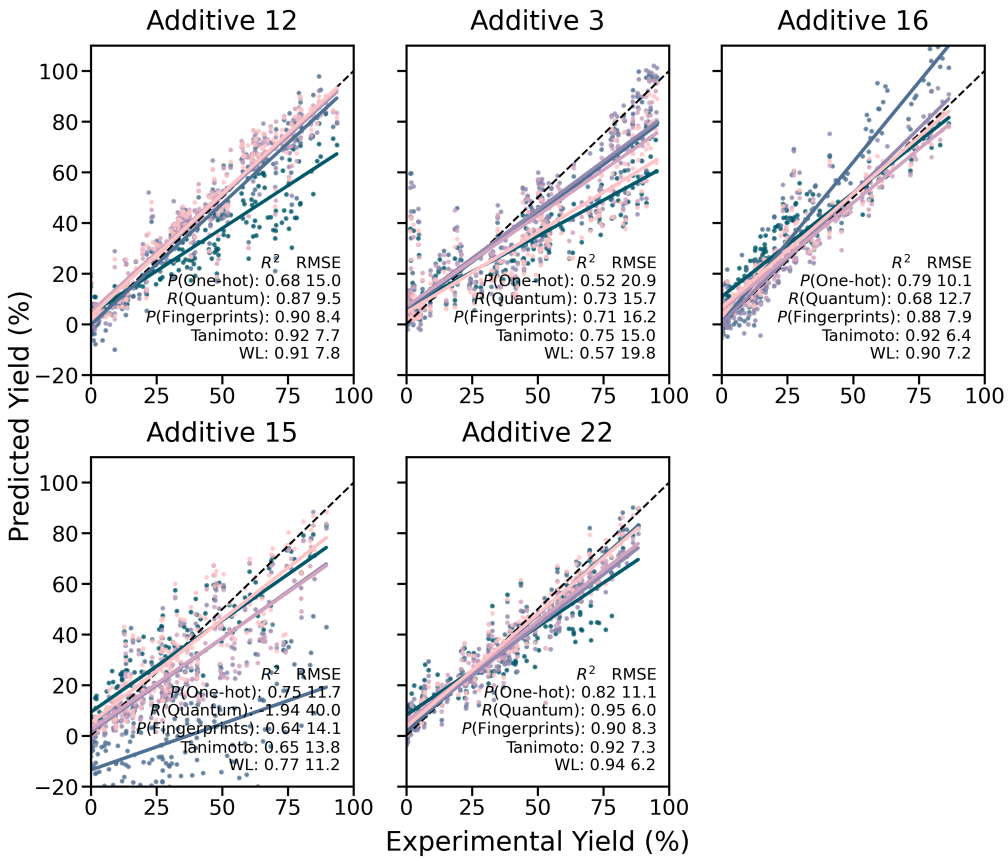
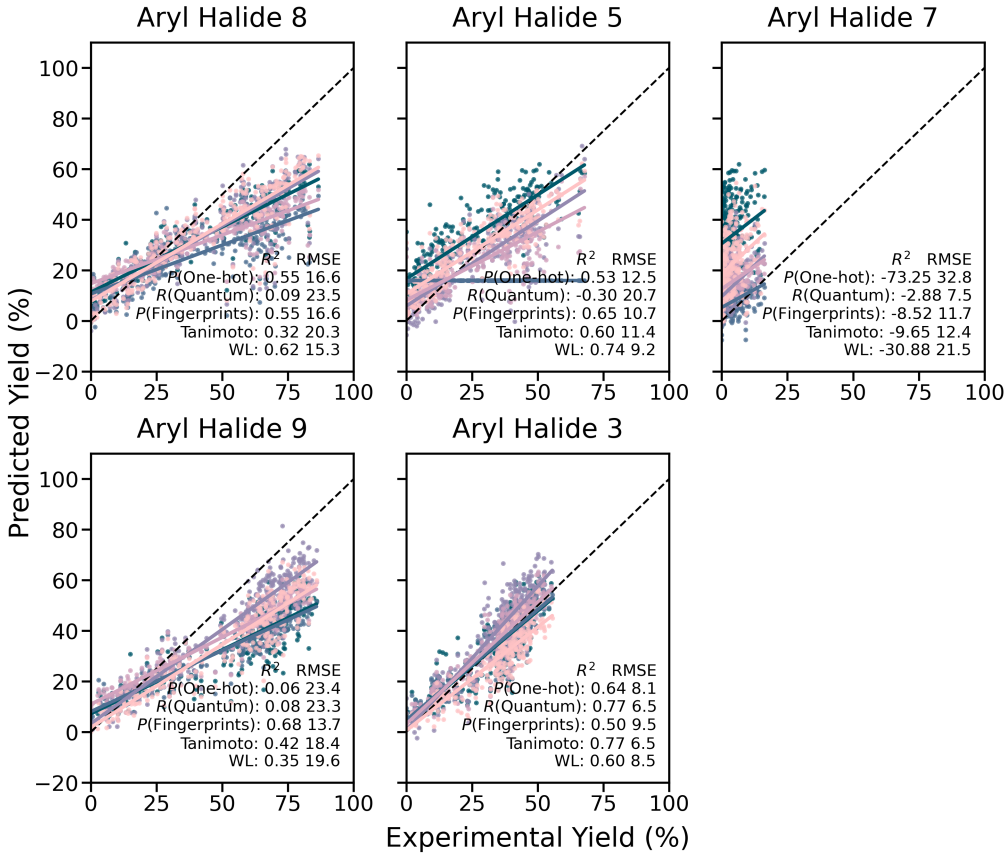
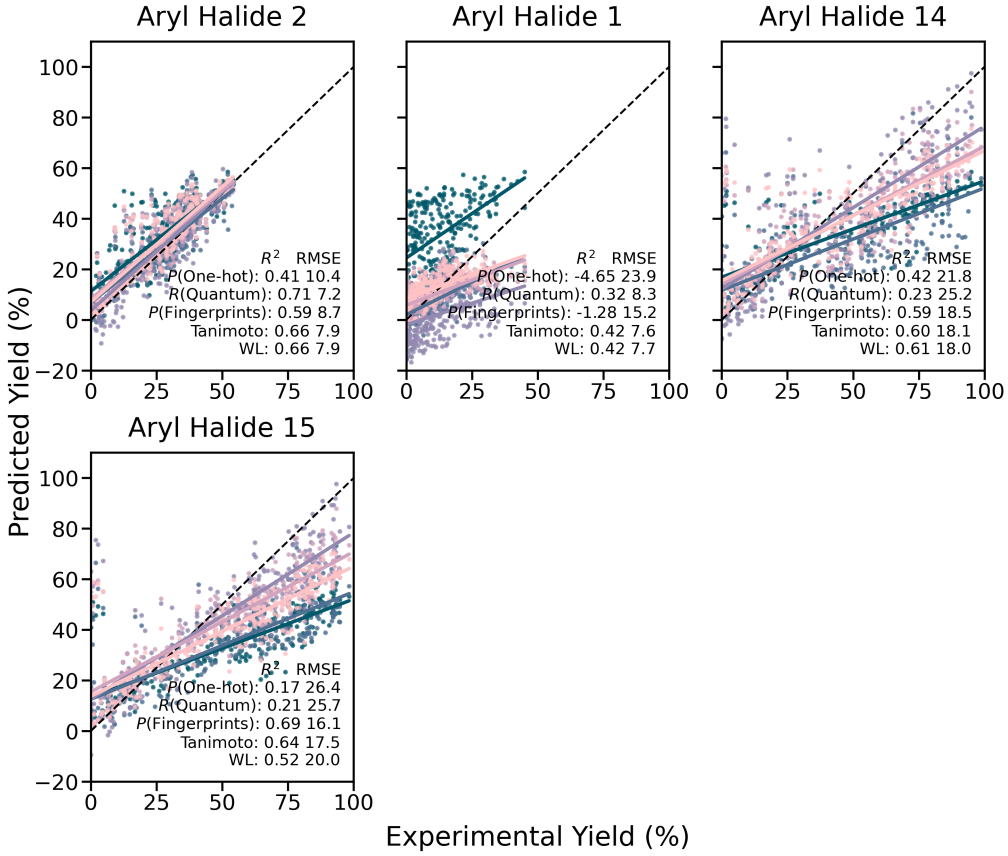



Figure B.15: Predicted yield against observed yield for each additive in the additive ranked test sets one (a) to four (d). $R^2$, coefficient of determination; dashed line, $y = x$; solid line, line of best fit.

(a)


Aryl Halide 8
$R^2$ RMSE
P(One-hot): 0.55 16.6
R(Quantum): 0.09 23.5
P(Fingerprints): 0.55 16.6
Tanimoto: 0.32 20.3
WL: 0.62 15.3
Aryl Halide 5
$R^2$ RMSE
P(One-hot): 0.53 12.5
R(Quantum): -0.30 20.7
P(Fingerprints): 0.65 10.7
Tanimoto: 0.60 11.4
WL: 0.74 9.2
Aryl Halide 7
$R^2$ RMSE
P(One-hot): -73.25 32.8
R(Quantum): -2.88 7.5
P(Fingerprints): -8.52 11.7
Tanimoto: -9.65 12.4
WL: -30.88 21.5
Aryl Halide 9
$R^2$ RMSE
P(One-hot): 0.06 23.4
R(Quantum): 0.08 23.3
P(Fingerprints): 0.68 13.7
Tanimoto: 0.42 18.4
WL: 0.35 19.6
Aryl Halide 3
$R^2$ RMSE
P(One-hot): 0.64 8.1
R(Quantum): 0.77 6.5
P(Fingerprints): 0.50 9.5
Tanimoto: 0.77 6.5
WL: 0.60 8.5
Predicted Yield (%)
Experimental Yield (%)

(b)


Aryl Halide 2
$R^2$ RMSE
P(One-hot): 0.41 10.4
R(Quantum): 0.71 7.2
P(Fingerprints): 0.59 8.7
Tanimoto: 0.66 7.9
WL: 0.66 7.9
Aryl Halide 1
$R^2$ RMSE
P(One-hot): -4.65 23.9
R(Quantum): 0.32 8.3
P(Fingerprints): -1.28 15.2
Tanimoto: 0.42 7.6
WL: 0.42 7.7
Aryl Halide 14
$R^2$ RMSE
P(One-hot): 0.42 21.8
R(Quantum): 0.23 25.2
P(Fingerprints): 0.59 18.5
Tanimoto: 0.60 18.1
WL: 0.61 18.0
Aryl Halide 15
$R^2$ RMSE
P(One-hot): 0.17 26.4
R(Quantum): 0.21 25.7
P(Fingerprints): 0.69 16.1
Tanimoto: 0.64 17.5
WL: 0.52 20.0
Predicted Yield (%)
Experimental Yield (%)

(c)

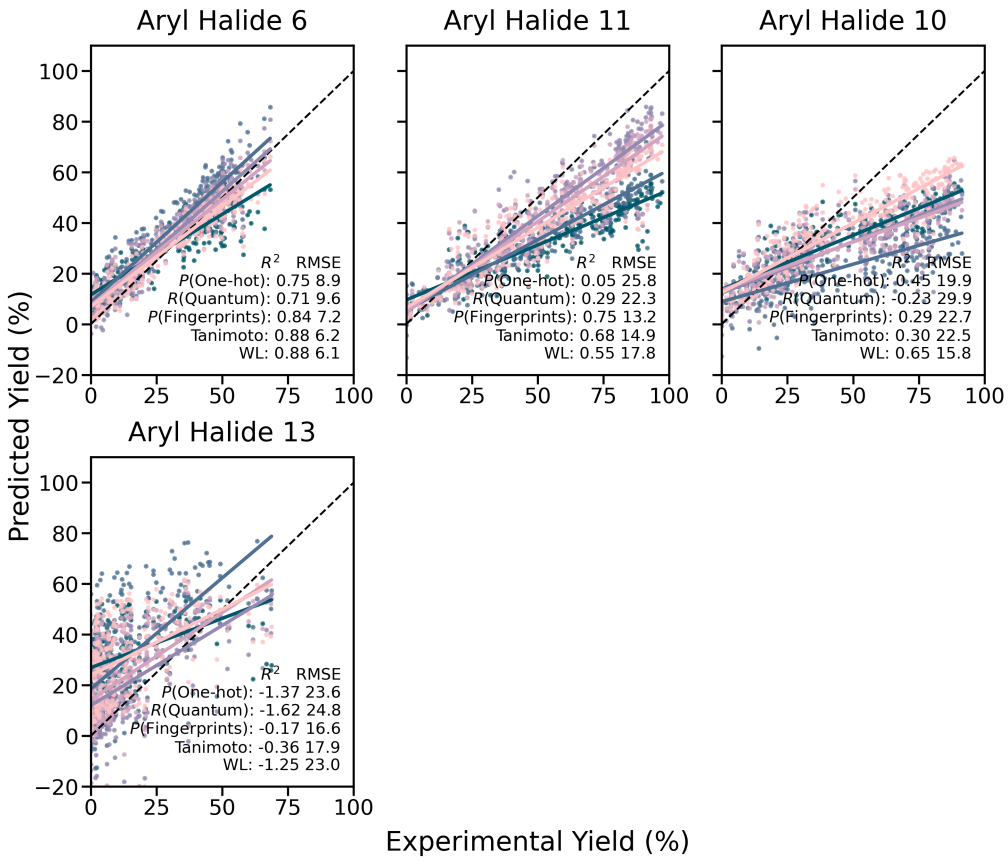



Figure B.16: Predicted yield against observed yield for each aryl halide in the aryl halide ranked test sets one (a) to three (c). $R^2$, coefficient of determination; dashed line, $y = x$; solid line, line of best fit.

## B.9 External Validation

### B.9.1 Training Set Performance

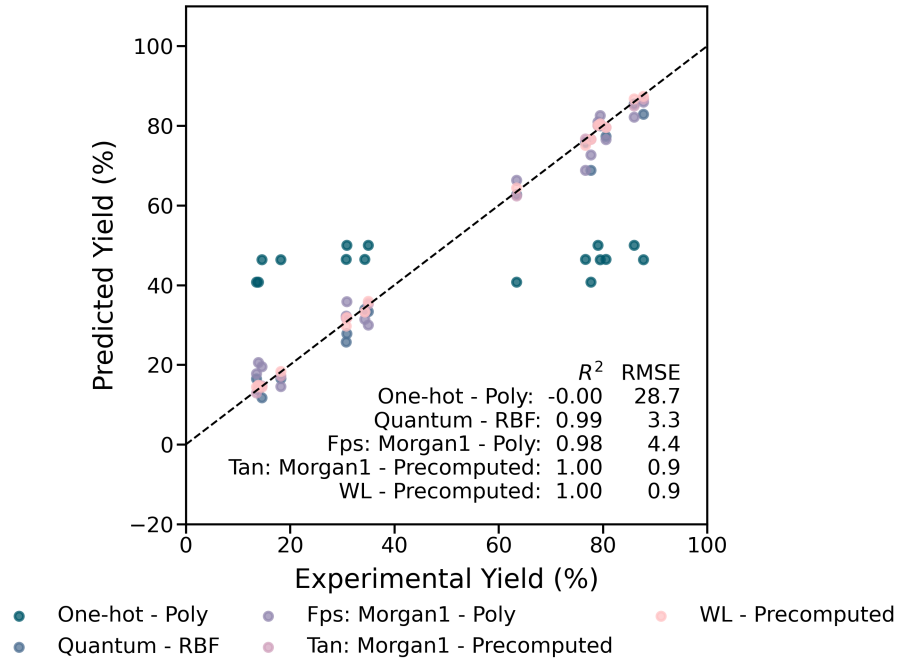



Figure B.17: Predicted yield against observed yield for the 16 reactions present in both the training and test set (subset of the validation reactions). $R^2$, coefficient of determination; RMSE (%), root mean squared error; dashed line, $y = x$.

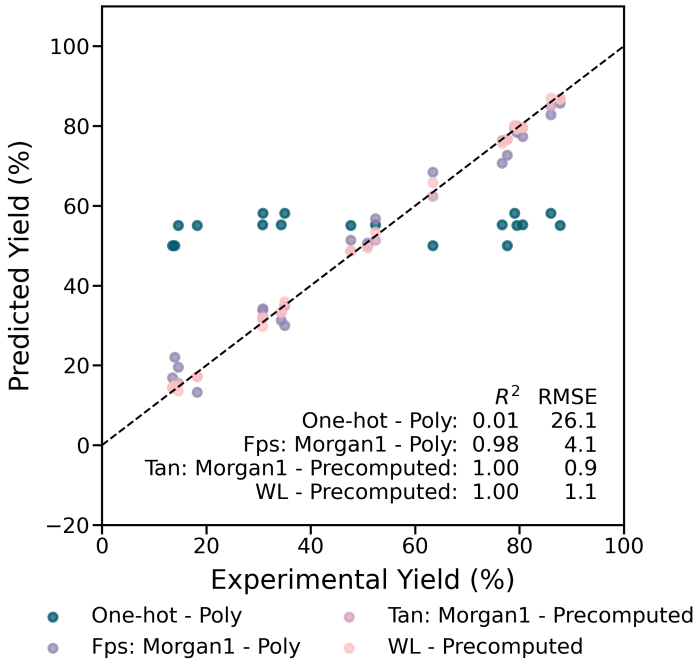



Figure B.18: Predicted yield against observed yield for the 19 reactions present in both the training and test set (all validation reactions). $R^2$, coefficient of determination; RMSE (%), root mean squared error; dashed line, $y = x$.

### B.9.2 Grid Search Cross-Validation

Table B.32: Grid Search Cross-Validated Performance for the SVR Validation Models

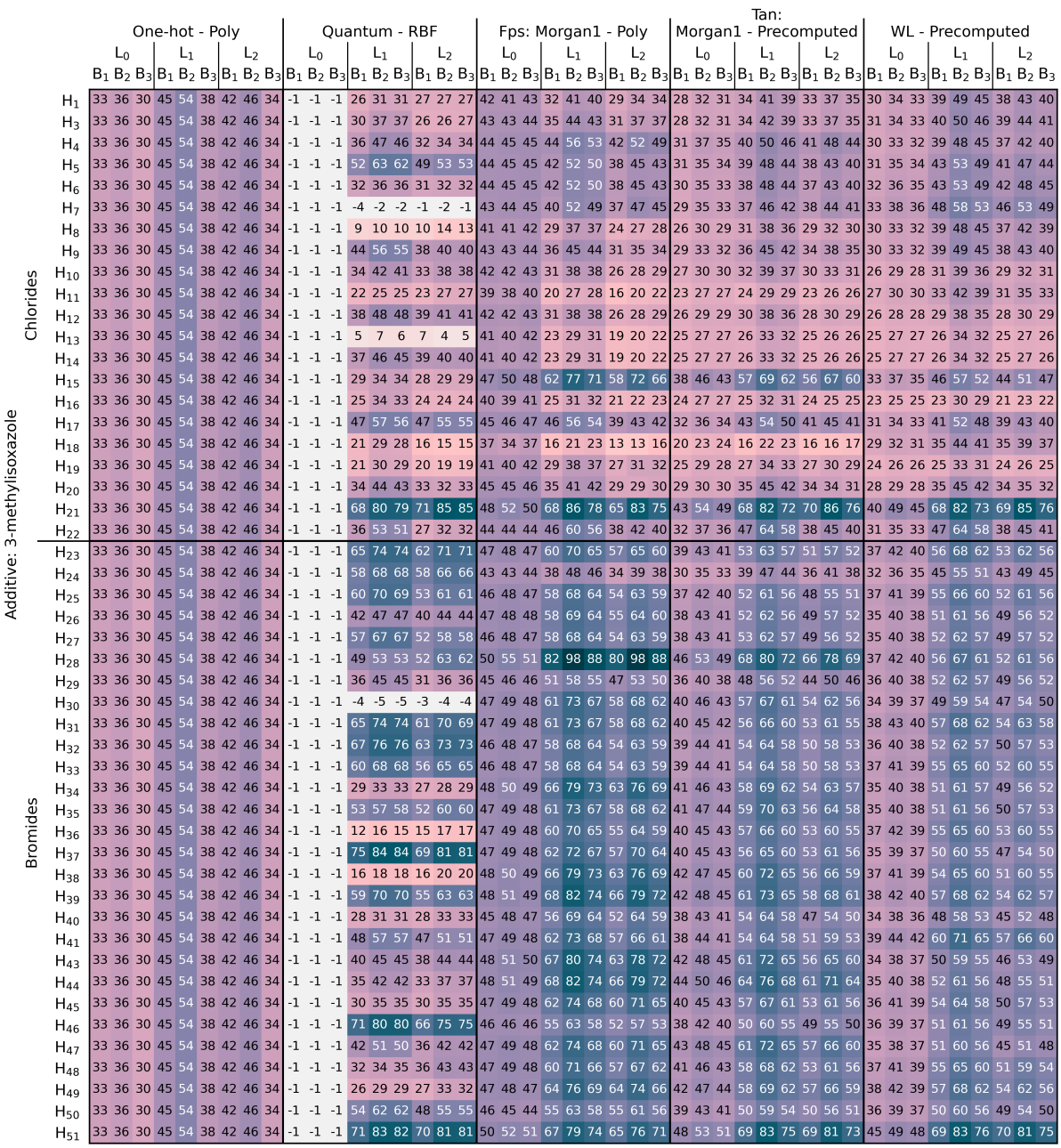
| Validation Test | SVR Model | $R^2$ | RMSE (%) |
|---|---|---|---|
| Subset | One-hot-Poly | 0.90 | 8.4 |
| | Quantum-RBF | 0.86 | 9.9 |
| | Fps: Morgan1-Poly | 0.90 | 8.3 |
| | Tan: Morgan1-Precomputed | 0.92 | 7.2 |
| | WL-Precomputed | 0.91 | 7.8 |
| All | One-hot-Poly | 0.90 | 8.5 |
| | Fps: Morgan1-Poly | 0.91 | 8.2 |
| | Tan: Morgan1-Precomputed | 0.94 | 6.8 |
| | WL-Precomputed | 0.92 | 7.7 |

Table B.33: Best Combination of Hyperparameters for the Prospective SVR Models Identified Using Grid Search Cross-Validation

| Hyperparameter | Possible Values | Validation Test | Model | Value Used |
|---|---|---|---|---|
| C | 1, 10, 100, 1000 | Subset | One-hot-Poly | 100 |
| | | | Quantum-RBF | 1000 |
| | | | Fps: Morgan1-Poly | 1000 |
| | | | Tan: Morgan1-Precomputed | 100 |
| | | | WL-Precomputed | 100 |
| | | All | One-hot-Poly | 100 |
| | | | Fps: Morgan1-Poly | 1000 |
| | | | Tan: Morgan1-Precomputed | 100 |
| | | | WL-Precomputed | 100 |
| epsilon | 1, 5, 10 | Subset | One-hot-Poly | 5 |
| | | | Quantum-RBF | 1 |
| | | | Fps: Morgan1-Poly | 5 |
| | | | Tan: Morgan1-Precomputed | 1 |
| | | | WL-Precomputed | 1 |
| | | All | One-hot-Poly | 5 |
| | | | Fps: Morgan1-Poly | 5 |
| | | | Tan: Morgan1-Precomputed | 1 |
| | | | WL-Precomputed | 1 |

### B.9.3 Prospective Predictions

(a)


Additive: 3-methylisoxazole
Chlorides
Bromides
One-hot - Poly
Quantum - RBF
Fps: Morgan1 - Poly
Tan: Morgan1 - Precomputed
WL - Precomputed
L0 L1 L2
B1 B2 B3
H1 33 36 30 45 54 38 42 46 34 -1 -1 -1 26 31 31 27 27 27 42 41 43 32 41 40 29 34 34 28 32 31 34 41 39 33 37 35 30 34 33 39 49 45 38 43 40
H3 33 36 30 45 54 38 42 46 34 -1 -1 -1 30 37 37 26 26 27 43 43 44 35 44 43 31 37 37 28 32 31 34 42 39 33 37 35 31 34 33 40 50 46 39 44 41
H4 33 36 30 45 54 38 42 46 34 -1 -1 -1 36 47 46 32 34 34 44 45 45 44 56 53 42 52 49 31 37 35 40 50 46 41 48 44 30 33 32 39 48 45 37 42 40
H5 33 36 30 45 54 38 42 46 34 -1 -1 -1 52 63 62 49 53 53 44 45 45 42 52 50 38 45 43 31 35 34 39 48 44 38 43 40 31 35 34 43 53 49 41 47 44
H6 33 36 30 45 54 38 42 46 34 -1 -1 -1 32 36 36 31 32 32 44 45 45 42 52 50 38 45 43 30 35 33 38 48 44 37 43 40 32 36 35 43 53 49 42 48 45
H7 33 36 30 45 54 38 42 46 34 -1 -1 -1 -4 -2 -2 -1 -2 -1 43 44 45 40 52 49 37 47 45 29 35 33 37 46 42 38 44 41 33 38 36 48 58 53 46 53 49
H8 33 36 30 45 54 38 42 46 34 -1 -1 -1 9 10 10 10 14 13 41 41 42 29 37 37 24 27 28 26 30 29 31 38 36 29 32 30 30 33 32 39 48 45 37 42 39
H9 33 36 30 45 54 38 42 46 34 -1 -1 -1 44 56 55 38 40 40 43 43 44 36 45 44 31 35 34 29 33 32 36 45 42 34 38 35 30 33 32 39 49 45 38 43 40
H10 33 36 30 45 54 38 42 46 34 -1 -1 -1 34 42 41 33 38 38 42 42 43 31 38 38 26 28 29 27 30 30 32 39 37 30 33 31 26 29 28 31 39 36 29 32 31
H11 33 36 30 45 54 38 42 46 34 -1 -1 -1 22 25 25 23 27 27 39 38 40 20 27 28 16 20 22 23 27 27 24 29 29 23 26 26 27 30 30 33 42 39 31 35 33
H12 33 36 30 45 54 38 42 46 34 -1 -1 -1 38 48 48 39 41 41 42 42 43 31 38 38 26 28 29 26 29 29 30 38 36 28 30 29 26 28 28 29 38 35 28 30 29
H13 33 36 30 45 54 38 42 46 34 -1 -1 -1 5 7 6 7 4 5 41 40 42 23 29 31 19 20 22 25 27 27 26 33 32 25 26 26 25 27 27 26 34 32 25 27 26
H14 33 36 30 45 54 38 42 46 34 -1 -1 -1 37 46 45 39 40 40 41 40 42 23 29 31 19 20 22 25 27 27 26 33 32 25 26 26 25 27 27 26 34 32 25 27 26
H15 33 36 30 45 54 38 42 46 34 -1 -1 -1 29 34 34 28 29 29 47 50 48 62 77 71 58 72 66 38 46 43 57 69 62 56 67 60 33 37 35 46 57 52 44 51 47
H16 33 36 30 45 54 38 42 46 34 -1 -1 -1 25 34 33 24 24 24 40 39 41 25 31 32 21 22 23 24 27 27 25 32 31 24 25 25 23 25 25 23 30 29 21 23 22
H17 33 36 30 45 54 38 42 46 34 -1 -1 -1 47 57 56 47 55 55 45 46 47 46 56 54 39 43 42 32 36 34 43 54 50 41 45 41 31 34 33 41 52 48 39 43 40
H18 33 36 30 45 54 38 42 46 34 -1 -1 -1 21 29 28 16 15 15 37 34 37 16 21 23 13 13 16 20 23 24 16 22 23 16 16 17 29 32 31 35 44 41 35 39 37
H19 33 36 30 45 54 38 42 46 34 -1 -1 -1 21 30 29 20 19 19 41 40 42 29 38 37 27 31 32 25 29 28 27 34 33 27 30 29 24 26 26 25 33 31 24 26 25
H20 33 36 30 45 54 38 42 46 34 -1 -1 -1 34 44 43 33 32 33 45 45 46 35 41 42 29 29 30 29 30 30 35 45 42 34 34 31 28 29 28 35 45 42 34 35 32
H21 33 36 30 45 54 38 42 46 34 -1 -1 -1 68 80 79 71 85 85 48 52 50 68 86 78 65 83 75 43 54 49 68 82 72 70 86 76 40 49 45 68 82 73 69 85 76
H22 33 36 30 45 54 38 42 46 34 -1 -1 -1 36 53 51 27 32 32 44 44 44 46 60 56 38 42 40 32 37 36 47 64 58 38 45 40 31 35 33 47 64 58 38 45 41
H23 33 36 30 45 54 38 42 46 34 -1 -1 -1 65 74 74 62 71 71 47 48 47 60 70 65 57 65 60 39 43 41 53 63 57 51 57 52 37 42 40 56 68 62 53 62 56
H24 33 36 30 45 54 38 42 46 34 -1 -1 -1 58 68 68 58 66 66 43 43 44 38 48 46 34 39 38 30 35 33 39 47 44 36 41 38 32 36 35 45 55 51 43 49 45
H25 33 36 30 45 54 38 42 46 34 -1 -1 -1 60 70 69 53 61 61 46 48 47 58 68 64 54 63 59 37 42 40 52 61 56 48 55 51 37 41 39 55 66 60 52 61 56
H26 33 36 30 45 54 38 42 46 34 -1 -1 -1 42 47 47 40 44 44 47 48 48 58 69 64 55 64 60 38 43 41 52 62 56 49 57 52 35 40 38 51 61 56 49 56 52
H27 33 36 30 45 54 38 42 46 34 -1 -1 -1 57 67 67 52 58 58 46 48 47 58 68 64 54 63 59 38 43 41 53 62 57 49 56 52 35 40 38 52 62 57 49 57 52
H28 33 36 30 45 54 38 42 46 34 -1 -1 -1 49 53 53 52 63 62 50 55 51 82 98 88 80 98 88 46 53 49 68 80 72 66 78 69 37 42 40 56 67 61 52 61 56
H29 33 36 30 45 54 38 42 46 34 -1 -1 -1 36 45 45 31 36 36 45 46 46 51 58 55 47 53 50 36 40 38 48 56 52 44 50 46 36 40 38 52 62 57 49 56 52
H30 33 36 30 45 54 38 42 46 34 -1 -1 -1 -4 -5 -5 -3 -4 -4 47 49 48 61 73 67 58 68 62 40 46 43 57 67 61 54 62 56 34 39 37 49 59 54 47 54 50
H31 33 36 30 45 54 38 42 46 34 -1 -1 -1 65 74 74 61 70 69 47 49 48 61 73 67 58 68 62 40 45 42 56 66 60 53 61 55 38 43 40 57 68 62 54 63 58
H32 33 36 30 45 54 38 42 46 34 -1 -1 -1 67 76 76 63 73 73 46 48 47 58 68 64 54 63 59 39 44 41 54 64 58 50 58 53 36 40 38 52 62 57 50 57 53
H33 33 36 30 45 54 38 42 46 34 -1 -1 -1 60 68 68 56 65 65 46 48 47 58 68 64 54 63 59 39 44 41 54 64 58 50 58 53 37 41 39 55 65 60 52 60 55
H34 33 36 30 45 54 38 42 46 34 -1 -1 -1 29 33 33 27 28 29 48 50 49 66 79 73 63 76 69 41 46 43 58 69 62 54 63 57 35 40 38 51 61 57 49 56 52
H35 33 36 30 45 54 38 42 46 34 -1 -1 -1 53 57 58 52 60 60 47 49 48 61 73 67 58 68 62 41 47 44 59 70 63 56 64 58 35 40 38 51 61 56 50 57 53
H36 33 36 30 45 54 38 42 46 34 -1 -1 -1 12 16 15 15 17 17 47 49 48 60 70 65 55 64 59 40 45 43 57 66 60 53 60 55 37 42 39 55 65 60 53 60 55
H37 33 36 30 45 54 38 42 46 34 -1 -1 -1 75 84 84 69 81 81 47 49 48 62 72 67 57 70 64 40 45 43 56 65 60 53 61 56 35 39 37 50 60 55 47 54 50
H38 33 36 30 45 54 38 42 46 34 -1 -1 -1 16 18 18 16 20 20 48 50 49 66 79 73 63 76 69 42 47 45 60 72 65 56 66 59 37 41 39 54 65 60 51 60 55
H39 33 36 30 45 54 38 42 46 34 -1 -1 -1 59 70 70 55 63 63 48 51 49 68 82 74 66 79 72 42 48 45 61 73 65 58 68 61 38 42 40 57 68 62 54 62 57
H40 33 36 30 45 54 38 42 46 34 -1 -1 -1 28 31 31 28 33 33 45 48 47 56 69 64 52 64 59 38 43 41 54 64 58 47 54 50 34 38 36 48 58 53 45 52 48
H41 33 36 30 45 54 38 42 46 34 -1 -1 -1 48 57 57 47 51 51 47 49 48 62 73 68 57 66 61 38 44 41 54 64 58 51 59 53 39 44 42 60 71 65 57 66 60
H43 33 36 30 45 54 38 42 46 34 -1 -1 -1 40 45 45 38 44 44 48 51 50 67 80 74 63 78 72 42 48 45 61 72 65 56 65 60 34 38 37 50 59 55 46 53 49
H44 33 36 30 45 54 38 42 46 34 -1 -1 -1 35 42 42 33 37 37 48 51 49 68 82 74 66 79 72 44 50 46 64 76 68 61 71 64 35 40 38 52 61 56 48 55 51
H45 33 36 30 45 54 38 42 46 34 -1 -1 -1 30 35 35 30 35 35 47 49 48 62 74 68 60 71 65 40 45 43 57 67 61 53 61 56 36 41 39 54 64 58 50 57 53
H46 33 36 30 45 54 38 42 46 34 -1 -1 -1 71 80 80 66 75 75 46 46 46 55 63 58 52 57 53 38 42 40 50 60 55 49 55 50 36 39 37 51 61 56 49 55 51
H47 33 36 30 45 54 38 42 46 34 -1 -1 -1 42 51 50 36 42 42 47 49 48 62 74 68 60 71 65 43 48 45 61 72 65 57 66 60 35 38 37 51 60 56 45 51 48
H48 33 36 30 45 54 38 42 46 34 -1 -1 -1 32 34 35 36 43 43 47 49 48 60 71 66 57 67 62 41 46 43 58 68 62 53 61 56 37 41 39 55 65 60 51 59 54
H49 33 36 30 45 54 38 42 46 34 -1 -1 -1 26 29 29 27 33 32 47 48 47 64 76 69 64 74 66 42 47 44 58 69 62 57 66 59 38 42 39 57 68 62 54 62 56
H50 33 36 30 45 54 38 42 46 34 -1 -1 -1 54 62 62 48 55 55 46 45 44 55 63 58 55 61 56 39 43 41 50 59 54 50 56 51 36 39 37 50 60 56 49 54 50
H51 33 36 30 45 54 38 42 46 34 -1 -1 -1 71 83 82 70 81 81 50 52 51 67 79 74 65 76 71 48 53 51 69 83 75 69 81 73 45 49 48 69 83 76 70 81 75

(b)

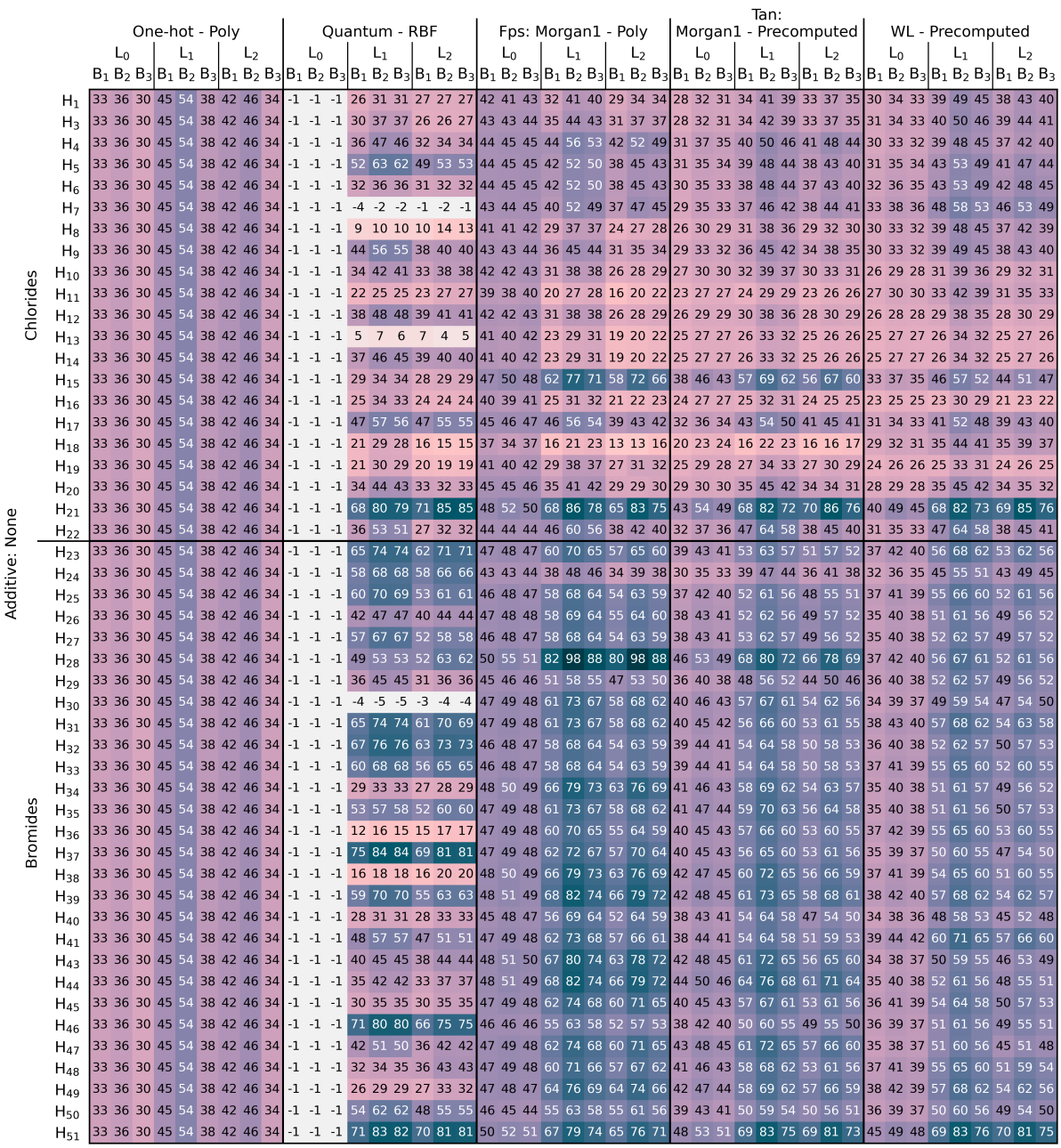
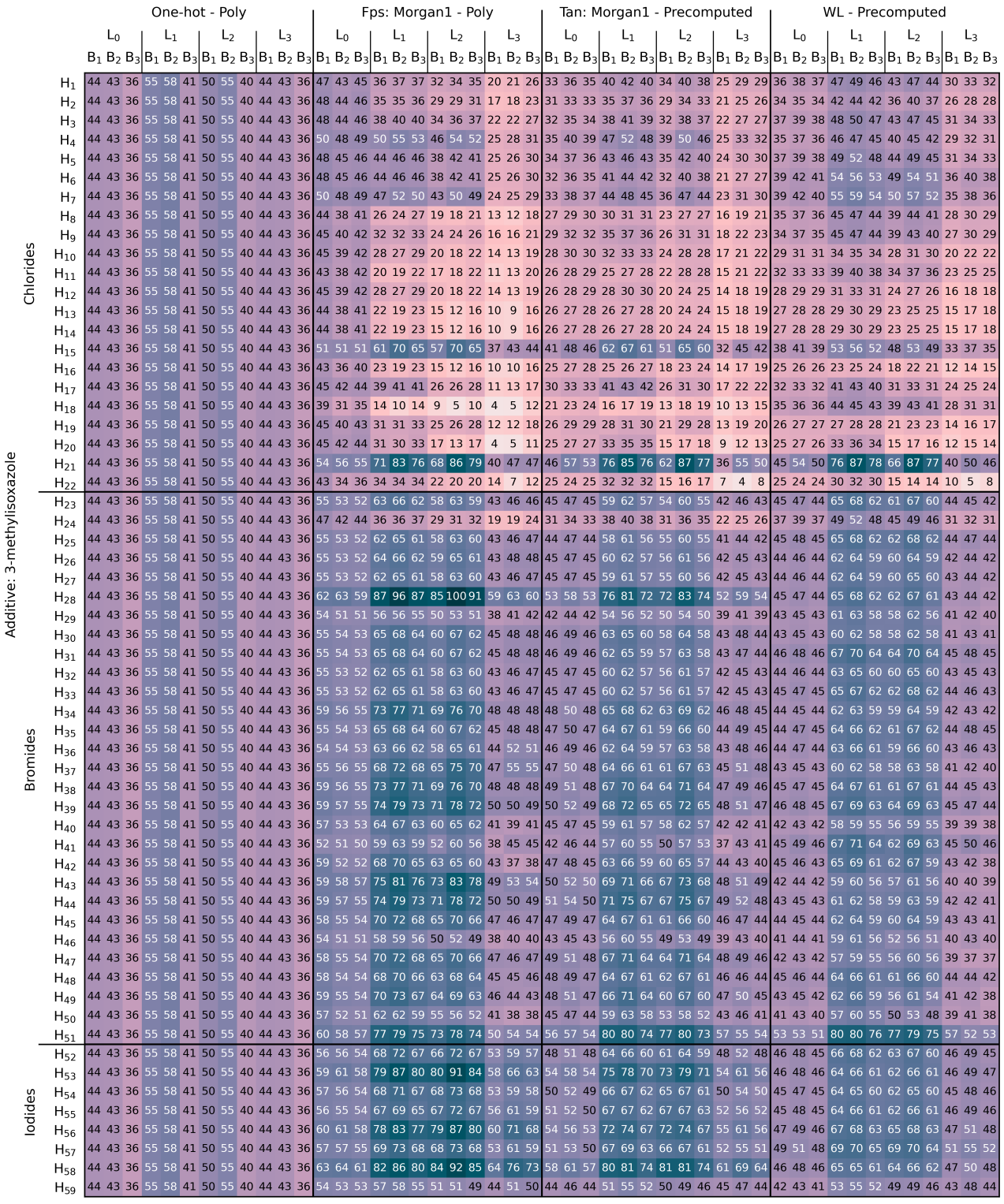



Figure B.19: Predicted reaction yields of the subset of validation reactions (a) with the additive 3-methylisoxazole and (b) without an additive present. $B_1$, BTMG; $B_2$, MTBD; $B_3$, DBU; $L_0$, no ligand; $L_1$, *t*-BuBrettPhos; $L_2$, *t*-BuXPhos; $L_3$, BrettPhos. The keys corresponding to the aryl halides ($H_1$ to $H_{59}$) are shown in Figures B.5, B.6, and B.7.

(a)


One-hot - Poly
Fps: Morgan1 - Poly
Tan: Morgan1 - Precomputed
WL - Precomputed
L0 L1 L2 L3
B1 B2 B3
Additive: 3-methylisoxazole
Chlorides
Bromides
Iodides
H1 44 43 36 55 58 41 50 55 40 44 43 36 47 43 45 36 37 37 32 34 35 20 21 26 33 36 35 40 42 40 34 40 38 25 29 29 36 38 37 47 49 46 43 47 44 30 33 32
H2 44 43 36 55 58 41 50 55 40 44 43 36 48 44 46 35 35 36 29 29 31 17 18 23 31 33 33 35 37 36 29 34 33 21 25 26 34 35 34 42 44 42 36 40 37 26 28 28
H3 44 43 36 55 58 41 50 55 40 44 43 36 48 44 46 38 40 40 34 36 37 22 22 27 32 35 34 38 41 39 32 38 37 22 27 27 37 39 38 48 50 47 43 47 45 31 34 33
H4 44 43 36 55 58 41 50 55 40 44 43 36 50 48 49 50 55 53 46 54 52 25 28 31 35 40 39 47 52 48 39 50 46 25 33 32 35 37 36 46 47 45 40 45 42 29 32 31
H5 44 43 36 55 58 41 50 55 40 44 43 36 48 45 46 44 46 46 38 42 41 25 26 30 34 37 36 43 46 43 35 42 40 24 30 30 37 39 38 49 52 48 44 49 45 31 34 33
H6 44 43 36 55 58 41 50 55 40 44 43 36 48 45 46 44 46 46 38 42 41 25 26 30 32 36 35 41 44 42 32 40 38 21 27 27 39 42 41 54 56 53 49 54 51 36 40 38
H7 44 43 36 55 58 41 50 55 40 44 43 36 50 48 49 47 52 50 43 50 49 24 25 29 33 38 37 44 48 45 36 47 44 23 31 30 39 42 40 55 59 54 50 57 52 35 38 36
H8 44 43 36 55 58 41 50 55 40 44 43 36 44 38 41 26 24 27 19 18 21 13 12 18 27 29 30 30 31 31 23 27 27 16 19 21 35 37 36 45 47 44 39 44 41 28 30 29
H9 44 43 36 55 58 41 50 55 40 44 43 36 45 40 42 32 32 33 24 24 26 16 16 21 29 32 32 35 37 36 26 31 31 18 22 23 34 37 35 45 47 44 39 43 40 27 30 29
H10 44 43 36 55 58 41 50 55 40 44 43 36 45 39 42 28 27 29 20 18 22 14 13 19 28 30 30 32 33 33 24 28 28 17 21 22 29 31 31 34 35 34 28 31 30 20 22 22
H11 44 43 36 55 58 41 50 55 40 44 43 36 43 38 42 20 19 22 17 18 22 11 13 20 26 28 29 25 27 28 22 28 28 15 21 22 32 33 33 39 40 38 34 37 36 23 25 25
H12 44 43 36 55 58 41 50 55 40 44 43 36 45 39 42 28 27 29 20 18 22 14 13 19 26 28 29 28 30 30 20 24 25 14 18 19 28 29 29 31 33 31 24 27 26 16 18 18
H13 44 43 36 55 58 41 50 55 40 44 43 36 44 38 41 22 19 23 15 12 16 10 9 16 26 27 28 26 27 28 20 24 24 15 18 19 27 28 28 29 30 29 23 25 25 15 17 18
H14 44 43 36 55 58 41 50 55 40 44 43 36 44 38 41 22 19 23 15 12 16 10 9 16 26 27 28 26 27 28 20 24 24 15 18 19 27 28 28 29 30 29 23 25 25 15 17 18
H15 44 43 36 55 58 41 50 55 40 44 43 36 51 51 51 61 70 65 57 70 65 37 43 44 41 48 46 62 67 61 51 65 60 32 45 42 38 41 39 53 56 52 48 53 49 33 37 35
H16 44 43 36 55 58 41 50 55 40 44 43 36 43 36 40 23 19 23 15 12 16 10 10 16 25 27 28 25 26 27 18 23 24 14 17 19 25 26 26 23 25 24 18 22 21 12 14 15
H17 44 43 36 55 58 41 50 55 40 44 43 36 45 42 44 39 41 41 26 26 28 11 13 17 30 33 33 41 43 42 26 31 30 17 22 22 32 33 32 41 43 40 31 33 31 24 25 24
H18 44 43 36 55 58 41 50 55 40 44 43 36 39 31 35 14 10 14 9 5 10 4 5 12 21 23 24 16 17 19 13 18 19 10 13 15 35 36 36 44 45 43 39 43 41 28 31 31
H19 44 43 36 55 58 41 50 55 40 44 43 36 45 40 43 31 31 33 25 26 28 12 12 18 26 29 29 28 31 30 21 29 28 13 19 20 26 27 27 27 28 28 21 23 23 14 16 17
H20 44 43 36 55 58 41 50 55 40 44 43 36 45 42 44 31 30 33 17 13 17 4 5 11 25 27 27 33 35 35 15 17 18 9 12 13 25 27 26 33 36 34 15 17 16 12 15 14
H21 44 43 36 55 58 41 50 55 40 44 43 36 54 56 55 71 83 76 68 86 79 40 47 47 46 57 53 76 85 76 62 87 77 36 55 50 45 54 50 76 87 78 66 87 77 40 50 46
H22 44 43 36 55 58 41 50 55 40 44 43 36 43 34 36 34 34 34 22 20 20 14 7 12 25 24 25 32 32 32 15 16 17 7 4 8 25 24 24 30 32 30 15 14 14 10 5 8
H23 44 43 36 55 58 41 50 55 40 44 43 36 55 53 52 63 66 62 58 63 59 43 46 46 45 47 45 59 62 57 54 60 55 42 46 43 45 47 44 65 68 62 61 67 60 44 45 42
H24 44 43 36 55 58 41 50 55 40 44 43 36 47 42 44 36 36 37 29 31 32 19 19 24 31 34 33 38 40 38 31 36 35 22 25 26 37 39 37 49 52 48 45 49 46 31 32 31
H25 44 43 36 55 58 41 50 55 40 44 43 36 55 53 52 62 65 61 58 63 60 43 46 47 44 47 44 58 61 56 55 60 55 41 44 42 45 48 45 65 68 62 62 68 62 44 47 44
H26 44 43 36 55 58 41 50 55 40 44 43 36 55 53 52 64 66 62 59 65 61 43 48 48 45 47 45 60 62 57 56 61 56 42 45 43 44 46 44 62 64 59 60 64 59 42 44 42
H27 44 43 36 55 58 41 50 55 40 44 43 36 55 53 52 62 65 61 58 63 60 43 46 47 45 47 45 59 61 57 55 60 56 42 45 43 44 46 44 62 64 59 60 65 60 43 44 42
H28 44 43 36 55 58 41 50 55 40 44 43 36 62 63 59 87 96 87 85 100 91 59 63 60 53 58 53 76 81 72 72 83 74 52 59 54 45 47 44 65 68 62 62 67 61 43 44 42
H29 44 43 36 55 58 41 50 55 40 44 43 36 54 51 51 56 56 55 50 53 51 38 41 42 42 44 42 54 56 52 50 54 50 39 41 39 43 45 43 61 63 58 58 62 56 41 42 40
H30 44 43 36 55 58 41 50 55 40 44 43 36 55 54 53 65 68 64 60 67 62 45 48 48 46 49 46 63 65 60 58 64 58 43 48 44 43 45 43 60 62 58 58 62 58 41 43 41
H31 44 43 36 55 58 41 50 55 40 44 43 36 55 54 53 65 68 64 60 67 62 45 48 48 46 49 46 62 65 59 57 63 58 43 47 44 46 48 46 67 70 64 64 70 64 45 48 45
H32 44 43 36 55 58 41 50 55 40 44 43 36 55 53 52 62 65 61 58 63 60 43 46 47 45 47 45 60 62 57 56 61 57 42 45 43 44 46 44 63 65 60 60 65 60 43 45 43
H33 44 43 36 55 58 41 50 55 40 44 43 36 55 53 52 62 65 61 58 63 60 43 46 47 45 47 45 60 62 57 56 61 57 42 45 43 45 47 45 65 67 62 62 68 62 44 46 43
H34 44 43 36 55 58 41 50 55 40 44 43 36 59 56 55 73 77 71 69 76 70 48 48 48 48 50 48 65 68 62 63 69 62 46 48 45 44 45 44 62 63 59 59 64 59 42 43 42
H35 44 43 36 55 58 41 50 55 40 44 43 36 55 54 53 65 68 64 60 67 62 45 48 48 47 50 47 64 67 61 59 66 60 44 49 45 44 47 45 64 66 62 61 67 62 44 48 45
H36 44 43 36 55 58 41 50 55 40 44 43 36 54 54 53 63 66 62 58 65 61 44 52 51 46 49 46 62 64 59 57 63 58 43 48 46 44 47 44 63 66 61 59 66 60 43 46 43
H37 44 43 36 55 58 41 50 55 40 44 43 36 55 56 55 68 72 68 65 75 70 47 55 55 47 50 48 64 66 61 61 67 63 45 51 48 43 45 43 60 62 58 58 63 58 41 42 40
H38 44 43 36 55 58 41 50 55 40 44 43 36 59 56 55 73 77 71 69 76 70 48 48 48 49 51 48 67 70 64 64 71 64 47 49 46 45 47 45 64 67 61 61 67 61 44 45 43
H39 44 43 36 55 58 41 50 55 40 44 43 36 59 57 55 74 79 73 71 78 72 50 50 49 50 52 49 68 72 65 65 72 65 48 51 47 46 48 45 67 69 63 64 69 63 45 47 44
H40 44 43 36 55 58 41 50 55 40 44 43 36 57 53 53 64 67 63 60 65 62 41 39 41 45 47 45 59 61 57 58 62 57 42 42 41 42 43 42 58 59 55 56 59 55 39 39 38
H41 44 43 36 55 58 41 50 55 40 44 43 36 52 51 50 59 63 59 52 60 56 38 45 45 42 46 44 57 60 55 50 57 53 37 43 41 45 49 46 67 71 64 62 69 63 45 50 46
H42 44 43 36 55 58 41 50 55 40 44 43 36 59 52 52 68 70 65 63 65 60 43 37 38 47 48 45 63 66 59 60 65 57 44 43 40 45 46 43 65 69 61 62 67 59 43 42 38
H43 44 43 36 55 58 41 50 55 40 44 43 36 59 58 57 75 81 76 73 83 78 49 53 54 50 52 50 69 71 66 67 73 68 48 51 49 42 44 42 59 60 56 57 61 56 40 40 39
H44 44 43 36 55 58 41 50 55 40 44 43 36 59 57 55 74 79 73 71 78 72 50 50 49 51 54 50 71 75 67 67 75 67 49 52 48 43 45 43 61 62 58 59 63 59 42 42 41
H45 44 43 36 55 58 41 50 55 40 44 43 36 58 55 54 70 72 68 65 70 66 47 46 47 47 49 47 64 67 61 61 66 60 46 47 44 44 45 44 62 64 59 60 64 59 43 43 41
H46 44 43 36 55 58 41 50 55 40 44 43 36 54 51 51 58 59 56 50 52 49 38 40 40 43 45 43 56 60 55 49 53 49 39 43 40 41 44 41 59 61 56 52 56 51 40 43 40
H47 44 43 36 55 58 41 50 55 40 44 43 36 58 55 54 70 72 68 65 70 66 47 46 47 49 51 48 67 71 64 64 71 64 48 49 46 42 43 42 57 59 55 56 60 56 39 37 37
H48 44 43 36 55 58 41 50 55 40 44 43 36 58 54 54 68 70 66 63 68 64 45 45 46 48 49 47 64 67 61 62 67 61 46 46 44 45 46 44 64 66 61 61 66 60 44 44 42
H49 44 43 36 55 58 41 50 55 40 44 43 36 59 55 54 70 73 67 64 69 63 46 44 43 48 51 47 66 71 64 60 67 60 47 50 45 43 45 42 62 66 59 56 61 54 41 42 38
H50 44 43 36 55 58 41 50 55 40 44 43 36 57 52 51 62 62 59 55 56 52 41 38 38 45 47 44 59 63 58 53 58 52 43 46 41 41 43 40 57 60 55 50 53 48 39 41 38
H51 44 43 36 55 58 41 50 55 40 44 43 36 60 58 57 77 79 75 73 78 74 50 54 54 56 57 54 80 80 74 77 80 73 57 55 54 53 53 51 80 80 76 77 79 75 57 52 53
H52 44 43 36 55 58 41 50 55 40 44 43 36 56 56 54 68 72 67 66 72 67 53 59 57 48 51 48 64 66 60 61 64 59 48 52 48 46 48 45 66 68 62 63 67 60 46 49 45
H53 44 43 36 55 58 41 50 55 40 44 43 36 59 61 58 79 87 80 80 91 84 58 66 63 54 58 54 75 78 70 73 79 71 54 61 56 46 48 46 64 66 61 62 66 61 46 49 47
H54 44 43 36 55 58 41 50 55 40 44 43 36 57 56 54 68 71 67 68 73 68 53 59 58 50 52 49 66 67 62 65 67 61 50 54 50 45 47 45 64 65 60 62 65 60 45 48 46
H55 44 43 36 55 58 41 50 55 40 44 43 36 56 55 54 67 69 65 67 72 67 56 61 59 51 52 50 67 67 62 67 67 63 52 56 52 45 48 45 64 66 61 62 66 61 46 49 46
H56 44 43 36 55 58 41 50 55 40 44 43 36 60 61 58 78 83 77 79 87 80 60 71 68 54 56 53 72 74 67 72 74 67 55 61 56 47 49 46 67 68 63 65 68 63 47 51 48
H57 44 43 36 55 58 41 50 55 40 44 43 36 57 57 55 69 73 68 68 73 68 53 61 59 51 53 50 67 69 63 66 67 61 52 55 51 49 51 48 69 70 65 69 70 64 51 55 52
H58 44 43 36 55 58 41 50 55 40 44 43 36 63 64 61 82 86 80 84 92 85 64 76 73 58 61 57 80 81 74 81 81 74 61 69 64 46 48 46 65 65 61 64 66 62 47 50 48
H59 44 43 36 55 58 41 50 55 40 44 43 36 54 53 53 57 58 55 51 51 49 44 51 50 44 46 44 51 55 52 50 49 46 45 47 44 42 43 41 53 55 52 49 49 46 43 48 44

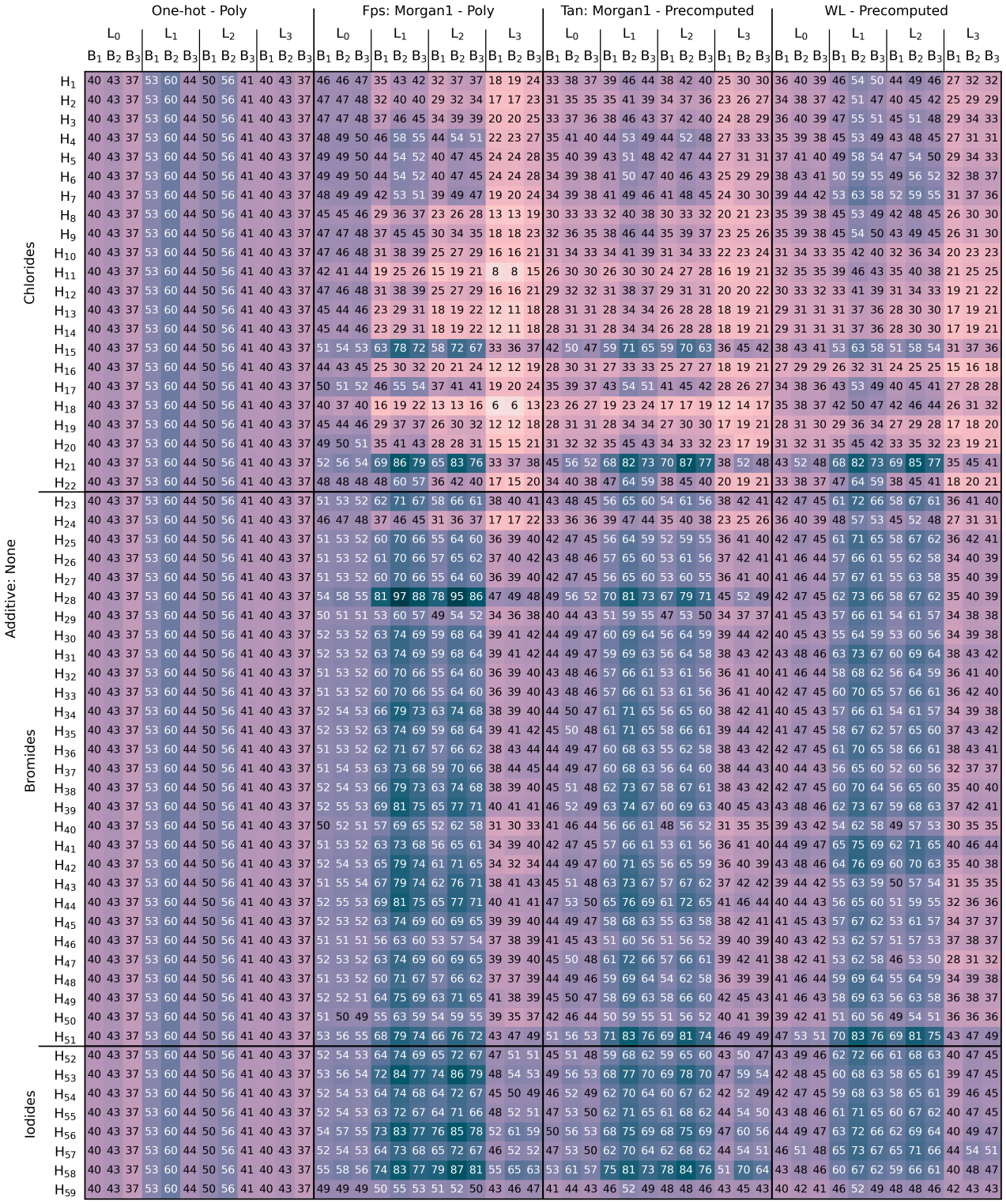



Figure B.20: Predicted reaction yields of the validation reactions (a) with the additive 3-methylisoxazole and (b) without an additive present. $B_1$, BTMG; $B_2$, MTBD; $B_3$, DBU; $L_0$, no ligand; $L_1$, *t*-BuBrettPhos; $L_2$, *t*-BuXPhos; $L_3$, BrettPhos. The keys corresponding to the aryl halides ($H_1$ to $H_{59}$) are shown in Figures B.5, B.6, and B.7.

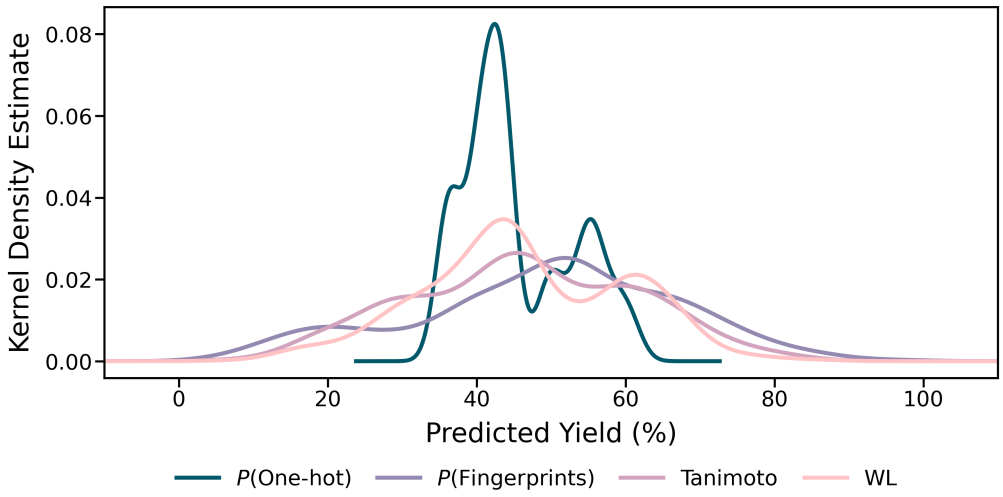



Figure B.21: Distributions of predicted reaction yield for all validation reactions.

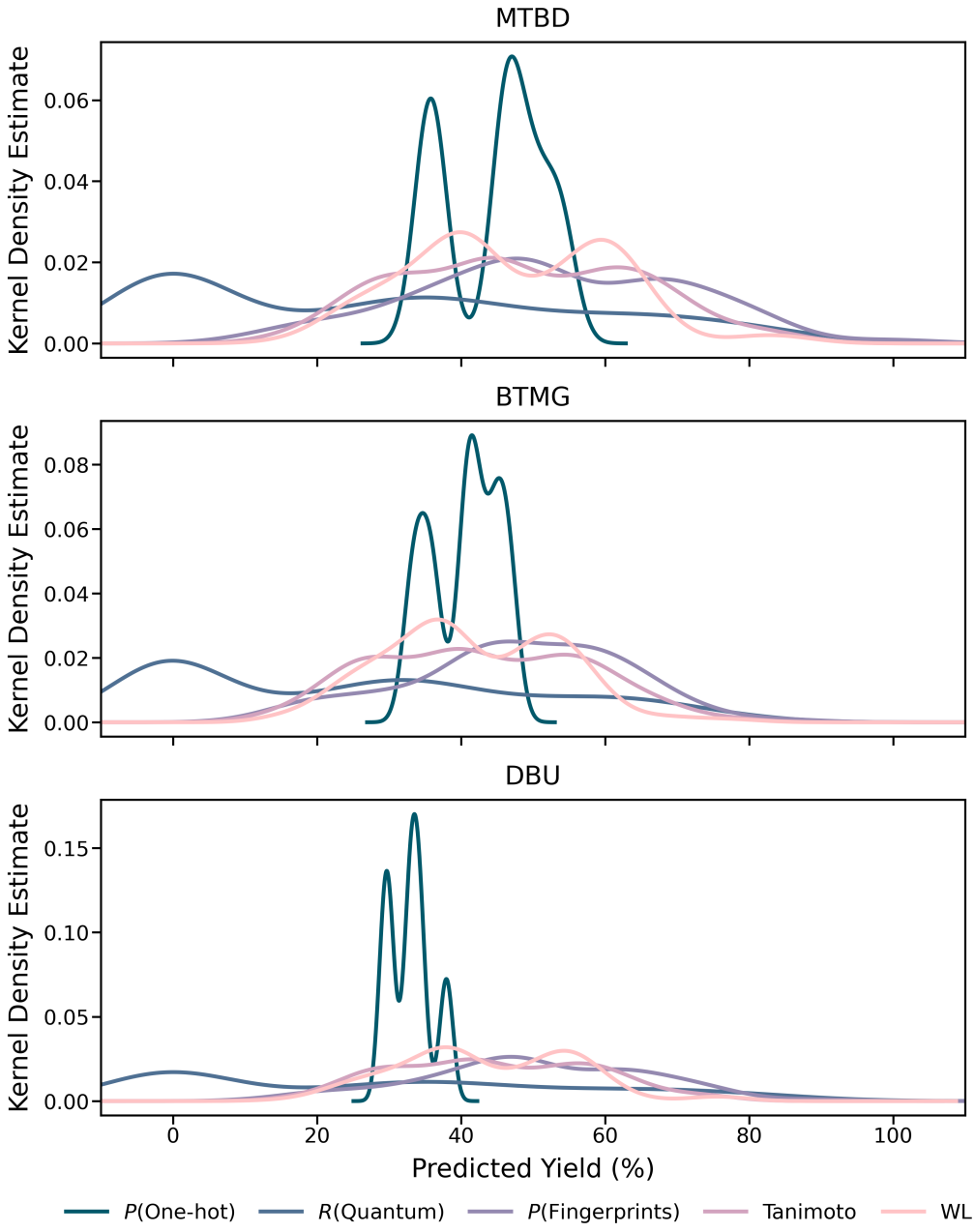



Figure B.22: Distributions of predicted reaction yield for the subset of validation reactions split by base (MTBD, BTMG, DBU).

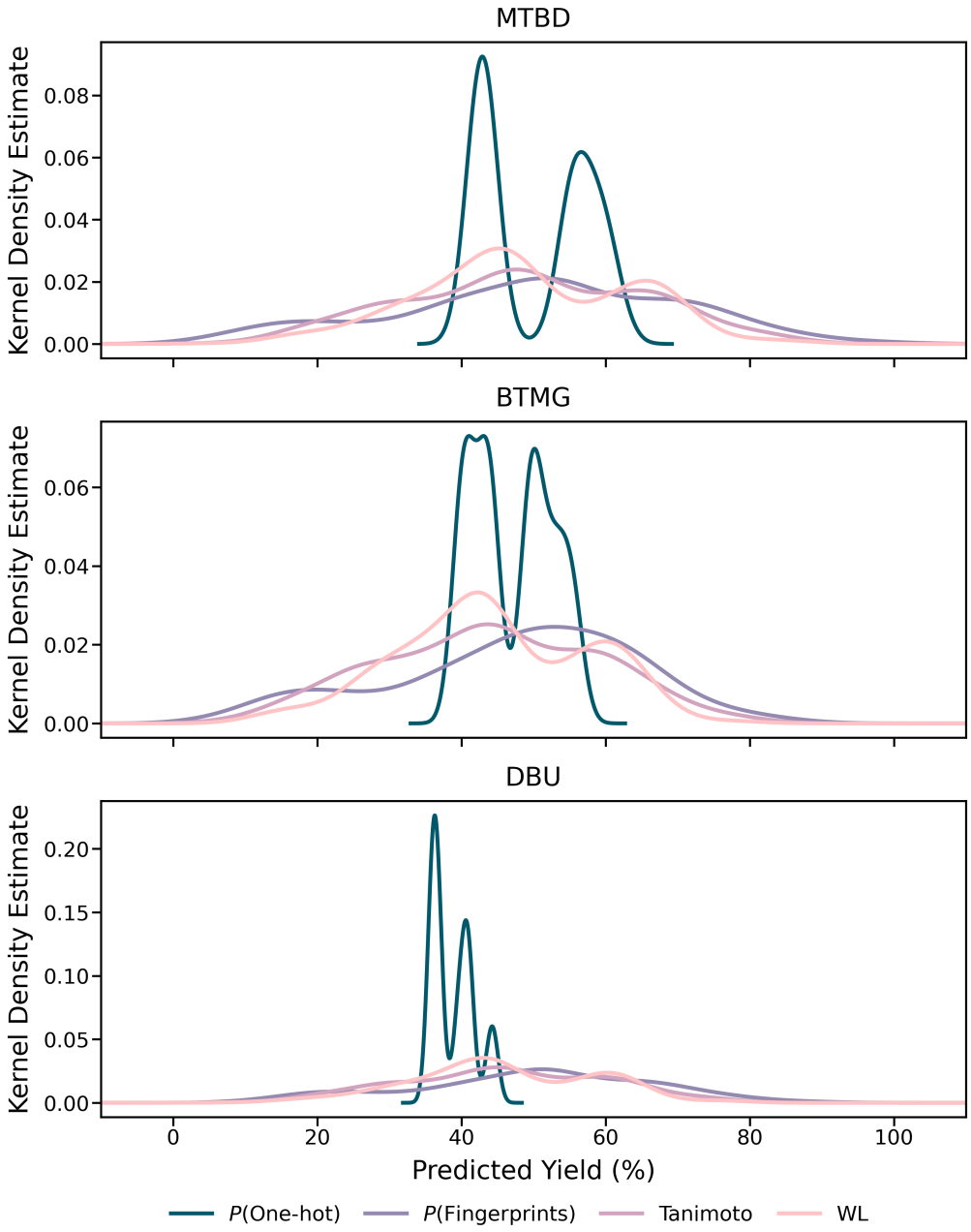



Figure B.23: Distributions of predicted reaction yield for all validation reactions split by base (MTBD, BTMG, DBU).

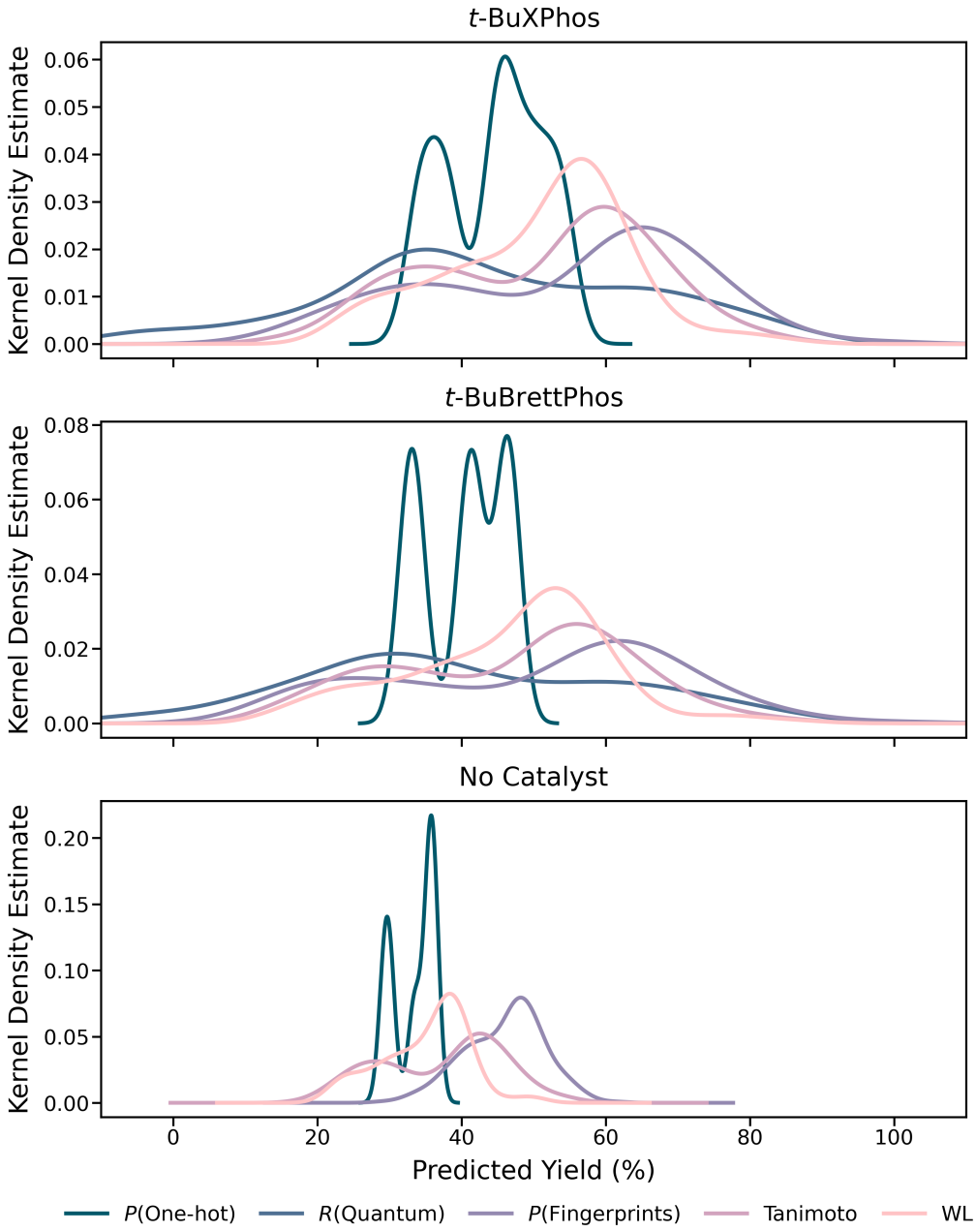



Figure B.24: Distributions of predicted reaction yield for the subset of validation reactions split by catalyst ligand (no catalyst, *t*-BuXPhos, *t*-BuBrettPhos). A distribution for the Quantum-RBF predictions of reactions containing no ligand is not provided as a constant value (-0.64%) was predicted.

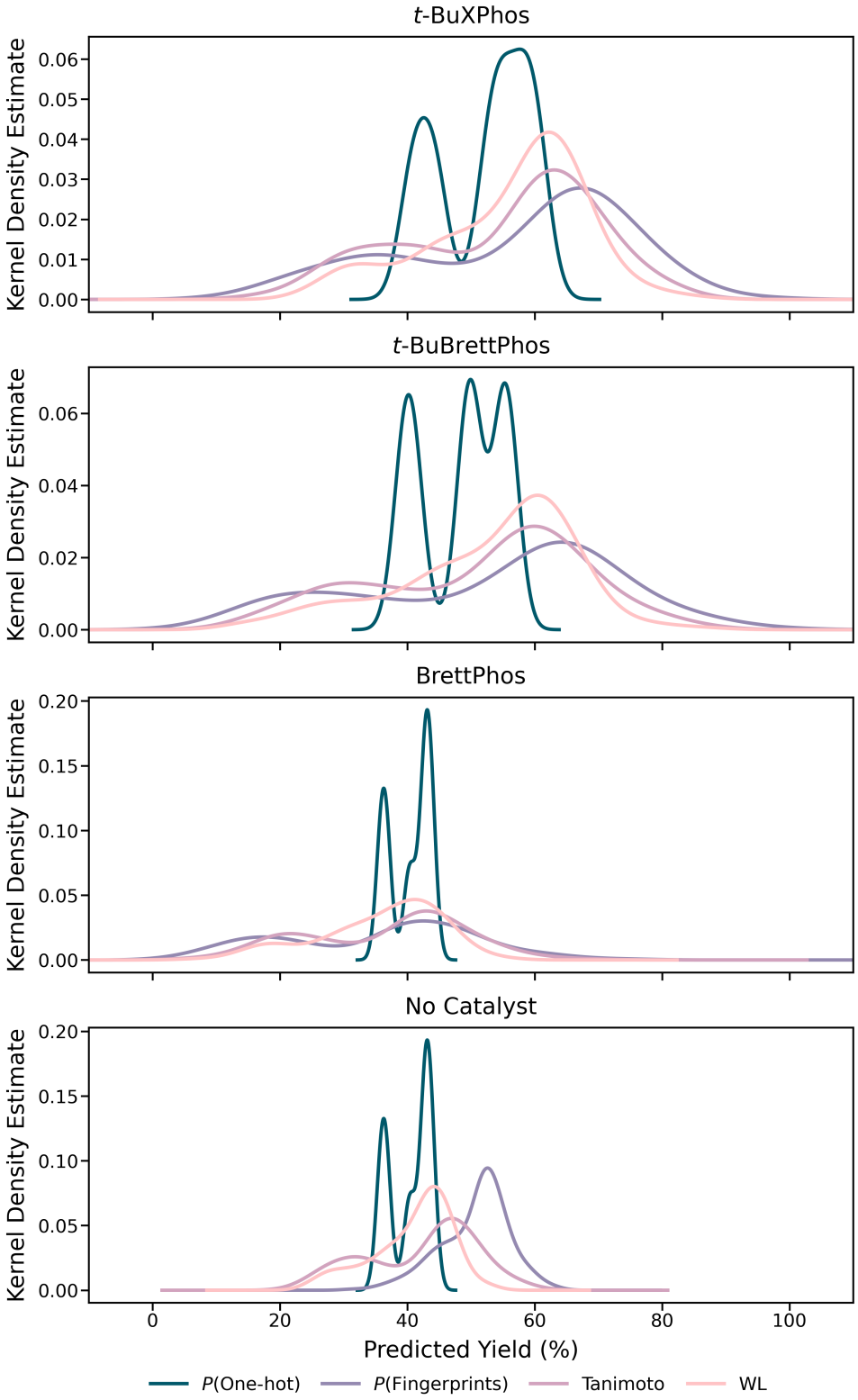



Figure B.25: Distributions of predicted reaction yield for all validation reactions split by catalyst ligand (no catalyst, *t*-BuXPhos, *t*-BuBrettPhos, BrettPhos).

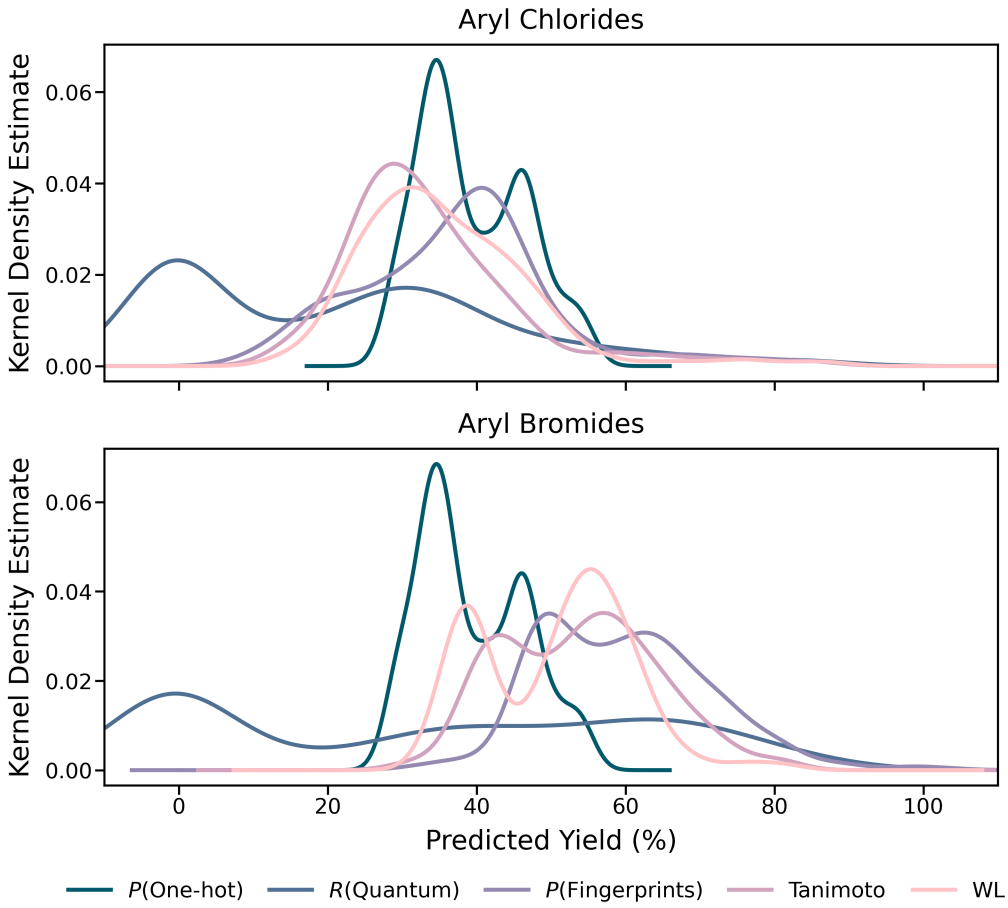



Figure B.26: Distributions of predicted reaction yield for the subset of validation reactions split by halide type (Cl, Br). Reactions performed without a ligand were excluded.

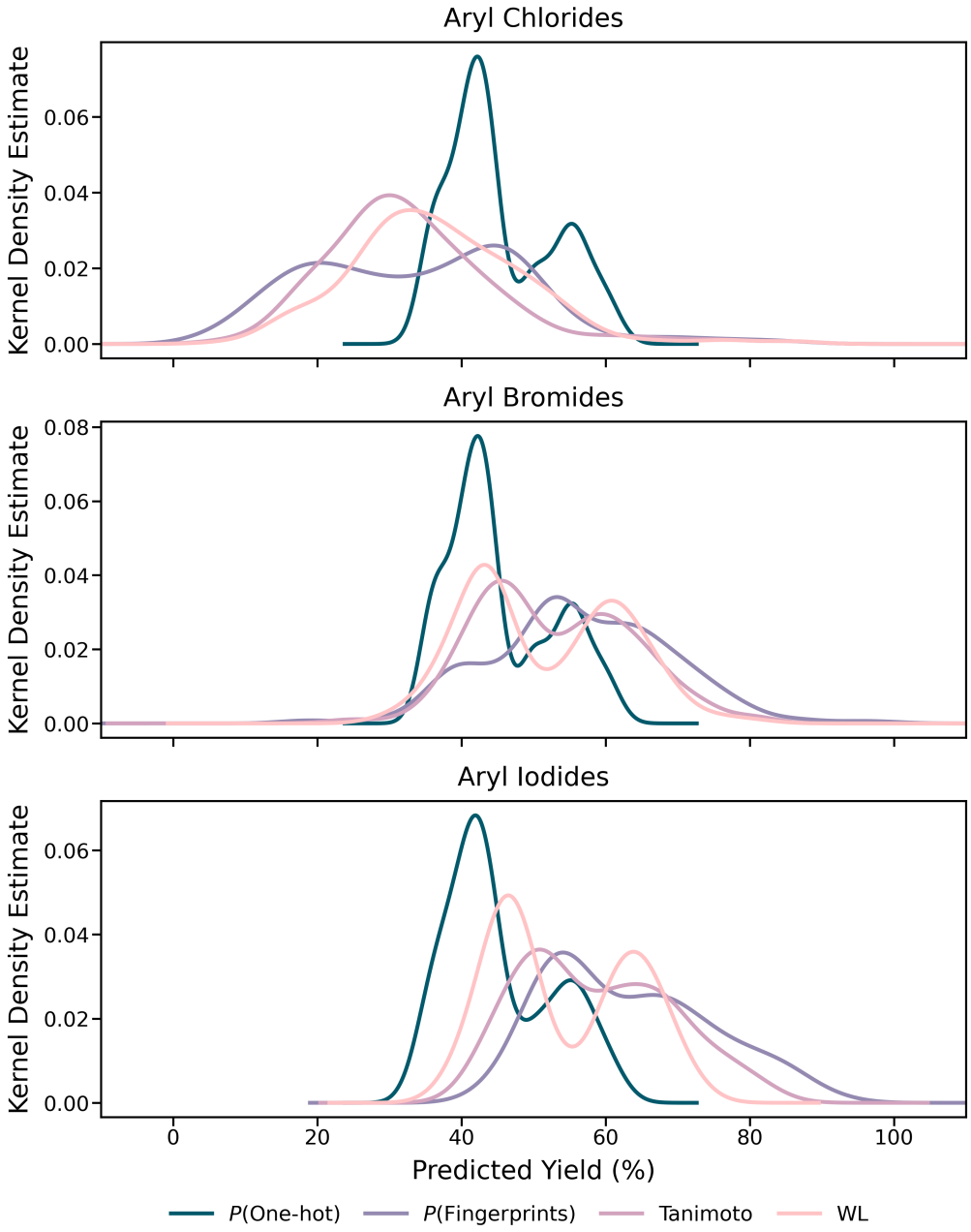



Figure B.27: Distributions of predicted reaction yield for all validation reactions split by halide type (Cl, Br, I).

### B.9.4 Model Comparison

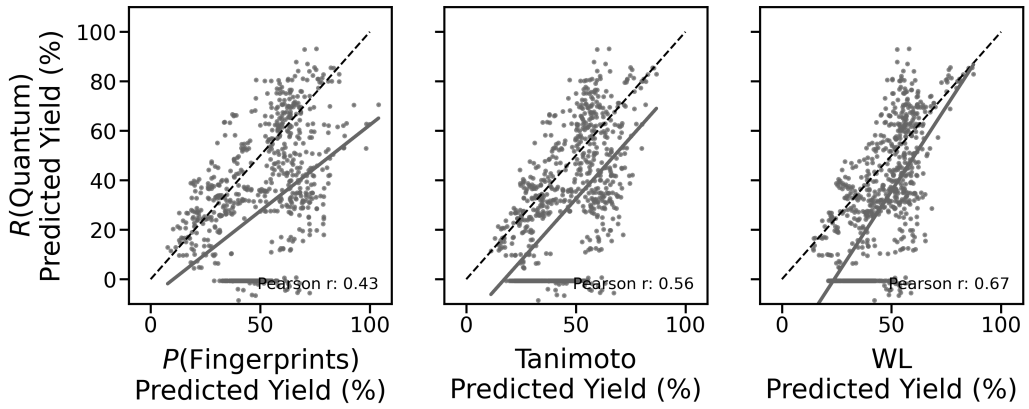



Figure B.28: Comparison between the predicted yield of the quantum chemical models with the structure-based models, and between the structure-based models.

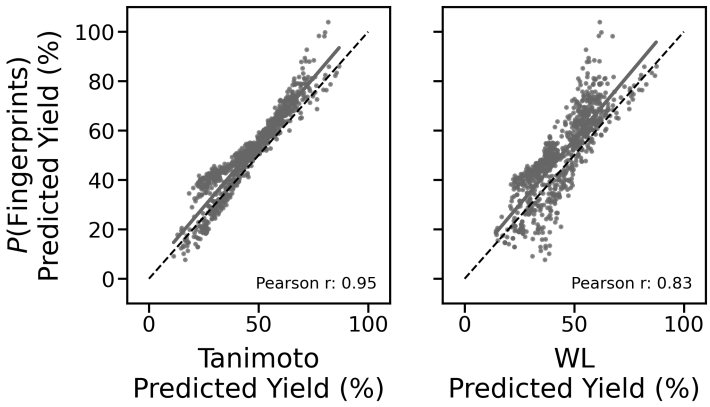



Figure B.29: Comparison between the predicted yield of the structure-based models. Solid line, line of best fit.

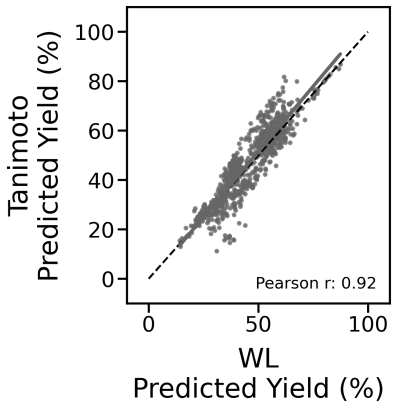



Figure B.30: Comparison between the predicted yield of the structure-based models. Solid line, line of best fit.