

Kernel Methods for Predicting Yields of Chemical Reactions

Alexe L. Haywood,^a Joseph Redshaw,^a Magnus W. D.

Hanson-Heine,^a Adam Taylor,^b Alex Brown,^b Andy M. Mason,^b

Thomas Gärtner^c and Jonathan D. Hirst^{a*}

^a *School of Chemistry, University of Nottingham, University Park, Nottingham, NG7 2RD, UK.* ^b *GlaxoSmithKline, Gunnels Wood Rd, Stevenage, SG1 2NY, UK.*

^c *Machine Learning Group, TU Wien Informatics, Vienna, Austria*

Abstract

The use of machine learning methods for the prediction of reaction yield is an emerging area. We demonstrate the applicability of support vector regression (SVR) for predicting reaction yields, using combinatorial data. Molecular descriptors used in regression tasks related to chemical reactivity have often been based on time-consuming, computationally demanding quantum chemical calculations, usually density functional theory. Structure-based descriptors (molecular fingerprints and molecular graphs) are quicker and easier to calculate, and are applicable to any molecule. In this study, SVR models built on structure-based descriptors were compared to models built on quantum chemical descriptors. The models were evaluated along the dimension of each reaction component in a set of Buchwald-Hartwig amination reactions. The structure-based SVR models out-performed the quantum chemical SVR models, along the dimension of each reaction component. The applicability of the models was assessed with respect to similarity to training. Prospective predictions of unseen Buchwald-Hartwig reactions are presented for synthetic assessment, to validate the generalisability of the models, with particular interest along the aryl halide dimension.

1 Introduction

Advances in medicinal chemistry rely on the discovery and synthesis of novel molecules. Time, cost and efficiency pressures in the pharmaceutical industry are key drivers in accelerating drug design and development. 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.¹⁻³ This attention includes the design and optimisation of small molecules. The availability of large reaction datasets and high-performance computing have been key in the development of computer-aided chemistry,⁴ for example in: molecular design,⁵ retrosynthetic planning tools,⁶⁻¹⁰ reaction prediction¹⁰⁻¹² and the optimisation of reaction conditions.¹³⁻¹⁵

Whilst the prediction of biological activities and molecular properties using quantitative structure-activity or structure-property relationship (QSAR/QSPR) models has been well-studied,^{1,16} 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 up opportunities to generate reaction data for use in machine learning.¹⁷⁻¹⁹

Support vector machines (SVM) are a supervised learning technique that use labelled training data to predict the label of unlabelled data.²⁰ It can be applied to classification and regression problems, whereby the label is either a class/category or continuous value, respectively. For non-linear relationships, SVMs use a kernel function to map data from an input space to a high-dimensional feature space, where classification or regression is performed linearly. The kernel function computes the inner product in the feature space directly, without applying the non-linear transformations at a higher computational cost. Different types of kernels have been assessed for both classification and regression problems related to chemo-²¹ and bioinformatics.²²⁻²⁶ Applications of SVMs in chemistry include bioactivity prediction, toxicity-related properties and physicochemical property prediction.^{1,26-29}

A dataset consisting of chemical structures or reactions must be converted to a machine readable format before presented to a machine learning algorithm. Molecular descriptors are based on the structural, physicochemical, electronic, or topological nature of molecules. Quantum chemical descriptors are common for the prediction of chemical reactivity.^{19,30-32} They have also been used to

build kernel-based QSAR and QSPR models, employing the Gaussian radial basis function (RBF) kernel.^{33–35} Site-specific, atomic properties including NMR shifts, vibrational frequencies, vibrational intensities and partial atomic charges have been used, along with global descriptors such as HOMO (Highest Occupied Molecular Orbital) energies, LUMO (Lowest Unoccupied Molecular Orbital) 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.^{31,32} 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 require overlapping, common structural features within the molecules.^{19,30,31} Reaction components that consist of a large variety of molecules with no key shared atoms between them all, require alternative representations such as structure-based descriptors.

A chemical hashed fingerprint defines the two-dimensional topology of a molecule in the form of a vector of binary bits. For example, MACCS Keys³⁶ depict the presence or absence of a set of predefined structural fragments, while other fingerprints consider each atom and its local environment. Morgan circular fingerprints³⁷ encode the neighbourhood within a particular radius of each atom, whereas RDKit fingerprints³⁸ encode topological paths up to a specified path length. Molecular fingerprints are fast and easy to calculate, making them a popular choice for representing molecules. They are established in machine learning for virtual screening³⁹ and have emerged in the prediction of reaction conditions.^{13,14} Sandfort et al.⁴⁰ 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.⁴⁰ Fingerprints have also been utilised in kernel-based QSAR/QSPR relationship models, using the Tanimoto or RBF kernel.^{27–29}

Molecular graphs are another two-dimensional representation that depict the atoms and bonds within molecules as a set of nodes and edges. 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 SVMs.⁴¹ Kriege et al.²³ give a detailed overview of graph kernels and provide guidelines to aid researchers in the

identification of successful kernels for different applications. Molecular graphs have been used in combination with deep learning to generate graph convolutional network models for reaction prediction,¹² retrosynthetic route design⁷ and the prediction of reaction conditions.⁴²

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. Small 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 the chemical intuition and experience of expert synthetic chemists, 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.^{43,44} 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/gas chromatography-mass spectrometry (LCMS/GCMS). There are challenges associated with such high throughput chemistry. These include the handling of very small volumes of liquid, 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.^{43,44} It is also a lower cost alternative for generating reaction data with which to build machine learning models.^{19,32,40,45}

An open-source combinatorial dataset, including reaction yields, was reported by Doyle et al.¹⁹ 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-methylaniline and 15 aryl halides, under varying reaction conditions (Fig. 1). 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 industry.⁴⁶⁻⁴⁸ The aromatic amine products are important building blocks for the synthesis of small drug-like molecules.⁴⁹ However, this key transformation can 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.⁴⁹ Potentially inhibitory isoxazole additives were included in the Buchwald-Hartwig reactions to simulate the effect of drug candidates 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.⁵⁰ This allowed assessment of the sub-structures' effect on reaction performance, without the need to synthesise and isolate isoxazole (or other) containing aryl halides as a prior step to performing the coupling reactions.

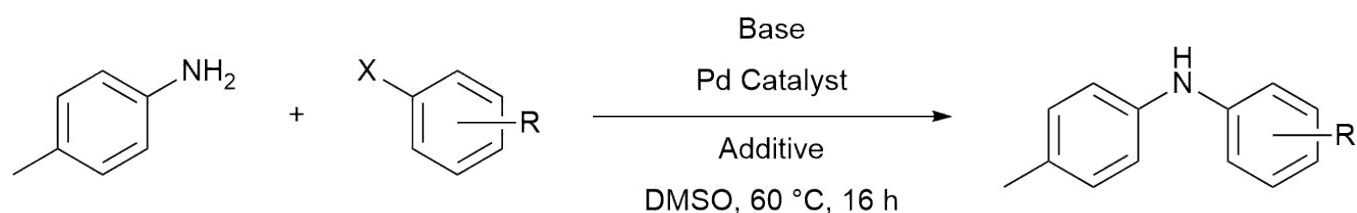


Fig. 1 Buchwald-Hartwig amination reaction.¹⁹

The dataset reported by Doyle et al. was used to build machine learning models to predict reaction yield.¹⁹ The reactions were represented using quantum chemical descriptors for each reaction component (aryl halide, additive, base and catalyst ligand). 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.⁵¹ 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 encodings⁵² can be used as a baseline to validate model performance and reveal potential patterns within the training data that may be fitted by models built on chemically meaningful descriptors. This one-hot encoding of a reaction simply denotes the presence or absence of each molecule in the form of a vector and encapsulates no information beyond this. For a 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 quantum chemical descriptors.⁵²

A more appropriate assessment of model generalisability is to test models with unseen molecules, not present in training, an out-of-sample test.⁵¹ A set of reactions containing specific molecules (one or more reaction components) are withheld from model training and used to assess the predictive ability of the trained model. It is important to ensure models are trained on reactions that cover a broad range of chemical space and observed variables. Doyle et al.¹⁹ designed out-

of-sample test sets by splitting the reactions along the high-throughput plates, where each plate contained a different set of additives. The random forest model built on quantum chemical descriptors was trained using the reactions on plate **1** and **2**, then tested using plate **3**. Chuang and Keiser identified that alternative splits of the plates resulted in a much lower performance, suggesting the random forest model built on quantum chemical descriptors was limited.⁵² Splitting the data along plate lines was also not a reliable way to assess model generalisability as each plate did not cover an even spread of chemical reactivity (Fig. S6a†). Out-of-sample test sets were therefore redesigned using activity ranking, along the additive dimension.⁵³ The mean yield of the reactions containing each additive were ranked from lowest to highest. The highest and lowest yielding additives were included in all training sets. Test sets were constructed from the remaining additives by taking every fourth molecule. This was repeated three more times to create a total of four test sets. Designing test sets using activity ranking ensured the model was trained on a range of reaction yields.⁵³ The quantum chemical random forest model showed good generalisability across the additive dimension, with a mean coefficient of determination (R^2) of 0.69 and root-mean-squared-error (RMSE) of 14.9% in the additive ranked test. Doyle et al. did not perform out-of-sample tests using activity ranking along the aryl halide dimension.

Support vector regression (SVR) models have been successful in the prediction of numerical values²⁰ related to QSAR and chemoinformatics.⁵⁴ Although Doyle et al. reported that the random forest method outperformed SVR in an initial in-sample test,^{19,53} we investigate the application of kernel methods further in the prediction of reaction yields, with more rigorous testing. In this study, SVR models are built on quantum chemical descriptors and two types of structure-based descriptors: molecular graphs and molecular fingerprints. The effect of the molecular descriptors on model performance, along the dimension of each reaction component, is investigated. To ensure the reported generalisability of the models is reliable, test sets are designed using activity ranking and the applicability of the models was assessed. A set of prospective reactions are outlined for model validation and predictions of reaction yields are reported prior to experimentation.

2 Methodology

2.1 Dataset

The data used in this study were 4608 single-step reactions reported by Doyle et al.¹⁹ This open access dataset contains the reactants, products, reaction conditions and yields of a single reaction class, the Buchwald-Hartwig amination reaction (Fig. 1). The reactions varied in 23 isoxazole additives, 15 aryl/heteroaryl halides, three bases and four Buchwald ligands (Fig. S1, S2 and S3†). 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;¹⁹ see the ESI† for details. The names of the aryl halide, additive, base and ligand in each reaction were converted to SMILES (Simplified Molecular Input Line Entry Specification) strings.⁵⁵ This was completed using the NCI/CADD Chemical Identifier Resolver API⁵⁶ with the exception of a few unrecognised names, which were drawn and converted to SMILES strings in ChemDraw.

A set of prospective combinatorial reactions was compiled to validate the SVR models. The proposed reactions will be performed experimentally using high-throughput chemistry to identify reaction yields. All possible combinations of 59 aryl halides, three bases, four catalyst ligands and two additives, formed a total of 1416 proposed reactions. Five of the aryl halides are present in the Doyle et al.¹⁹ dataset and will be used as standards. The remaining aryl halides cover ortho, meta and para substituents, with a range of electron withdrawing and electron donating groups (Fig. S4†). The base, DBU, and catalyst ligand, BrettPhos, were selected along with the two higher yielding bases and ligands from the Doyle et al. dataset: MTBD, BTMG, *t*-BuXPhos and *t*-BuBrettPhos (Fig. S3†). The prospective reactions will also be performed without a catalyst to investigate whether the reactions of the *ortho*-substituted halopyridines are proceeding via an alternative reaction pathway. As the aim of these reactions is to assess model generalisability, particularly along the aryl halide dimension, the reactions will be carried out, with and without, a single isoxazole additive: 3-methylisoxazole (Fig. S1†).

2.2 Molecular Descriptors

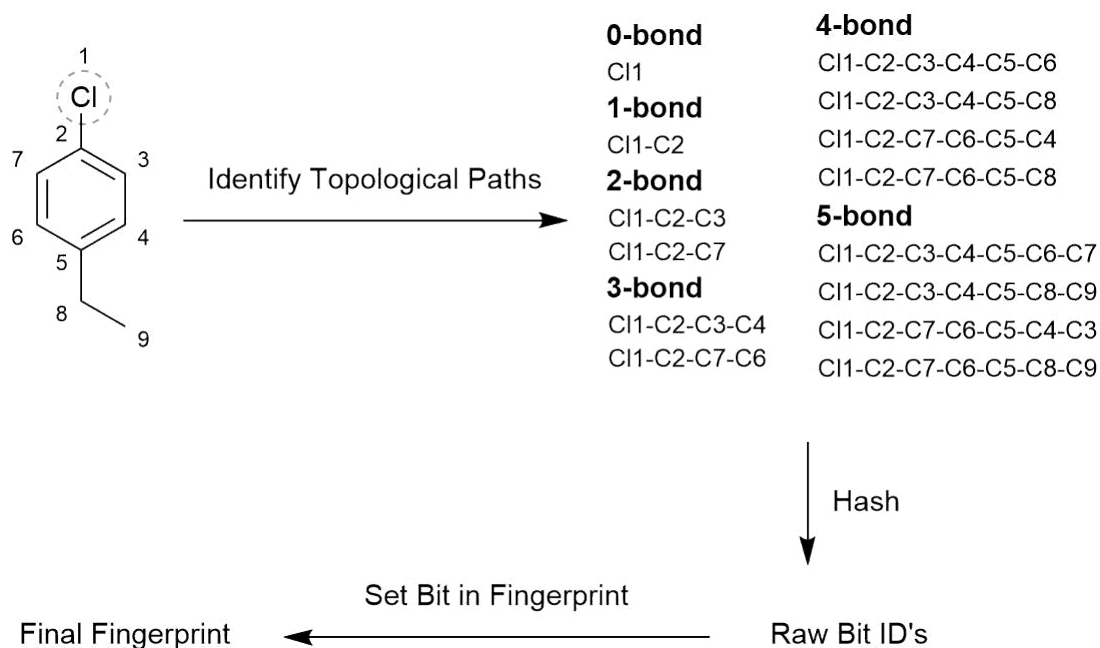
Quantum Chemical Descriptors. A combination of calculated molecular, atomic and vibrational properties formed a set of quantum chemical descriptors for each reaction (Table 1). The quantum chemical descriptors for the Doyle et al. dataset were calculated by Doyle et al.¹⁹ The molecular descriptors included 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. The common molecular vibrational modes across the set of molecules for each reagent class were identified. The vibrational frequencies and infrared transition intensities were calculated for the common modes. The Spartan '14 interface for the Q-Chem quantum chemical software package^{57,58} was used to calculate 120 descriptors per reaction (19, 27, 10 and 64 descriptors for the additive, aryl halide, base and ligand, respectively) using the density functional B3LYP with the 6-31G(d) basis set.^{59,60} The descriptors were standardised by centring the data to have zero mean and scaling to unit variance. The same method was used to calculate the quantum chemical descriptors for the prospective reactions; computational details can be found in the ESI†.

Table 1 Format and Notation of the Quantum Chemical Descriptors, Molecular Fingerprints, Molecular Graphs and One-hot Encodings for a Single Reaction

Reaction Components	Additive	Aryl Halide	Base	Ligand
Quantum Chemical Descriptors	$[D_1^{Ad} \dots D_{19}^{Ad}]$	$[D_1^{AH} \dots D_{27}^{AH}]$	$[D_1^B \dots D_{10}^B]$	$[D_1^L \dots D_{64}^L]$
Fingerprints	$[\dots 0 1 \dots]$	$[\dots 0 1 \dots]$	$[\dots 0 1 \dots]$	$[\dots 0 1 \dots]$
Concatenated Fingerprints	$[\dots 0 1 \dots]$	$[\dots 0 1 \dots]$	$[\dots 0 1 \dots]$	$[\dots 0 1 \dots]$
Graphs	$[G^{Ad}]$	$[G^{AH}]$	$[G^B]$	$[G^L]$
One-hot Encodings	Ad ₁ ⋯ Ad ₂₂	AH ₁ ⋯ AH ₁₅	B ₁ B ₂ B ₃	L ₁ ⋯ L ₄
	$[1 \quad \dots \quad 0]$	$[1 \quad \dots \quad 0]$	$[1 \quad 0 \quad 0]$	$[1 \quad \dots \quad 0]$

Molecular Fingerprints. The topology of molecules can be represented by molecular fingerprints. Three types were implemented using the RDKit package: MACCS Keys,³⁶ RDKit fingerprints³⁸ and Morgan circular fingerprints.³⁷ Fingerprints are hashes (i.e. binary bit vectors) of a specified length, set to 1024-bit for this study (see the ESI† for further discussion), except the MACCS fingerprint which is 167-bit by definition. The bits within a MACCS fingerprint define the

(a)



(b)

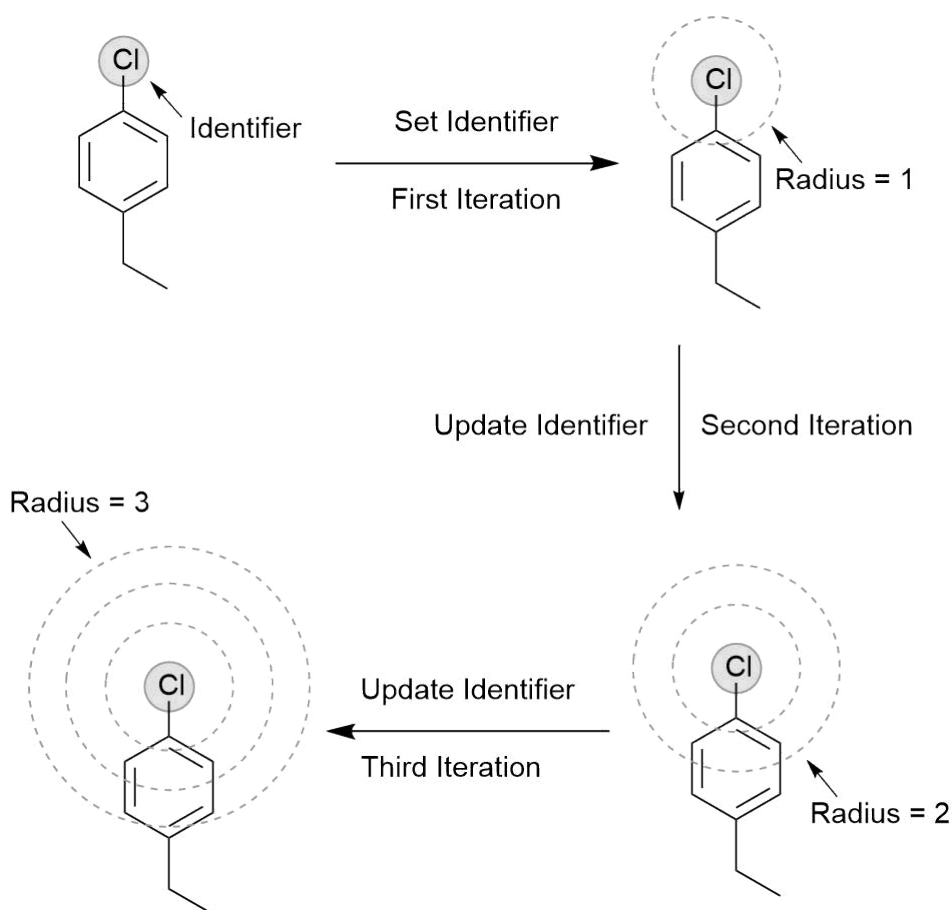


Fig. 2 Schematics of how (a) RDK fingerprints and (b) Morgan fingerprints are calculated, using the chlorine atom in 1-chloro-4-ethylbenzene as an example. These processes would be repeated for each atom in the molecule.

presence or absence of predefined substructures/fragments, called MACCS Keys. A total of 166 public MACCS Keys define the substructures/fragments as SMARTS strings. Note: There are 167 bits due to zero indexing in python, to allow for the original numbering of the MACCS keys (1-166), i.e. every fingerprint begins with a zero.

RDK topological fingerprints and Morgan circular fingerprints are two-dimensional, topological descriptors that define the connectivity of molecules. For the RDK fingerprints, fragments are generated by topological paths starting from each atom up to a predefined path length (number of bonds); the default 7 was used (Fig. 2a). The bond order and neighbour counts of each fragment are used with a hash function to set the bits in the molecular fingerprint. Morgan circular fingerprints, also called Extended Connectivity Fingerprints (ECFPs), initially set an identifier for each atom based on the number of adjacent non-hydrogen atoms, number of bonds to "heavy" atoms, atomic number, atomic mass, number of bonds to hydrogen atoms and whether the atom is in a ring is assigned to each atom. Feature Morgan fingerprints are a variant, also known as the Functional-Class fingerprints (FCFPs), that differ in the assignment of the atom identifier by assigning a code for the atom's role instead, e.g. hydrogen-bond acceptor and donor, aromatic, halogen, basic or acidic. In both types of Morgan fingerprints, each identifier is iteratively updated to include the identifier and bond order of neighbouring atoms (Fig. 2b) up to a specified radius. Each iteration includes a larger circular environment around the atoms. Once the iterations are complete, the identifiers are folded into the length of the bit vector using a hashing function. For the purpose of this study, Morgan and feature Morgan fingerprints with radii of 1, 2 and 3 were investigated.

Graph-based Descriptors. A molecular graph represents the topology of a molecule by a set of nodes corresponding to the atoms, connected by a set of 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.³⁸ This information was parsed to a module within GraKel to generate molecular graph representations.⁶¹

One-hot Encodings. One-hot encodings of chemical reactions are binary vectors that denote the presence or absence of each molecule in the dataset (Table 1). The reactions are represented without using chemically meaningful information and 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.

2.3 Support Vector Regression Models

Machine learning models relating descriptors to reaction yield were developed using the SVR method as implemented in scikit-learn.⁶² In ϵ -SVR, the aim is to find a function that deviates from the observed variables by a maximum of ϵ for each training point. The following values were considered for the ϵ hyperparameter: 1, 5 and 10. To prevent overfitting of the training data, slack variables (ξ) are introduced to allow for errors larger than ϵ . Only the points x_i that fall outside of the ϵ -insensitive tube contribute to the objective function, with their contribution being equal to ξ_i or ξ_i^* .²⁰ Eq. 1 describes the optimisation problem that is solved during the training of the SVR algorithm.

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

The hyperparameter, C , can be tuned to determine the toleration of points outside of ϵ . As C increases, the tolerance increases. The following values were considered for C : 1, 10, 100 and 1000.

SVR uses a kernel function to map input data to a higher-dimensional feature space where regression is performed linearly. The kernel functions (Table 2) explored were: linear, polynomial, sigmoid, Gaussian radial basis function (RBF), Tanimoto and Weisfeiler-Lehman⁶³ (WL). The Sigmoid equation is not a valid kernel but has been successfully applied, see Schölkopf⁶⁴ for further details. The first four kernels were applied to the quantum chemical descriptors, concatenated molecular fingerprints and one-hot encodings using scikit-learn, where the format of the descriptors per reaction is a single vector (Table 1). The equations of these kernels between two reactions (x and y) are shown in Table 2. The molecular fingerprints and molecular graphs are formatted as a single entity per reaction component (Table 1). Tanimoto similarities between fingerprints

were calculated using RDKit. For two molecules in a single reaction component represented by molecular fingerprints (A and B), the Tanimoto similarity^{65,66} is defined as

$$T_{A,B} = \frac{c}{a + b - c}$$

where a and b are the number of bits set in fingerprints A and B , and c is the number of bits set in common in A and B . Although small changes in the structure of small molecules can lead to large changes in the Tanimoto similarity, it is a very well-established measure and thus appropriate for us to consider. Weisfeiler-Lehman subtree (WL) graph kernels⁶³ were calculated using GraKel. Kriege et al.²³ suggest there is little benefit in the combination of the WL graph kernel with non-linear kernels. Therefore the WL graph kernel was used in combination with a linear kernel function and we did not explore beyond this. To calculate the Tanimoto and Weisfeiler-Lehman kernels between two reactions (x and y), the Hadamard product of reaction component kernels was taken (Table 2). This is shown in Eq. 2, where a_i , h_i , b_i and l_i are the additive, aryl halide, base and ligand in reaction i , respectively.

$$k(x,y) = k(a_x, a_y) k(h_x, h_y) k(b_x, b_y) k(l_x, l_y) \quad (2)$$

Table 2 Kernel Equations for Reactions x and y ($\gamma (> 0)$, $c (\geq 0$ for the polynomial kernel) and $d (\geq 0)$ are hyperparameters set to the default in scikit-learn: $\gamma = 1.0/n_{features}$ where $n_{features}$ is the number of features, $c = 1$ and $d = 3$.)

Kernel Name	Equation, $k(x,y)$
Linear	$x^T y$
Polynomial	$(\gamma x^T y + c)^d$
Sigmoid	$\tanh(\gamma x^T y + c)$
RBF	$\exp(-\gamma \ x - y\ ^2)$
Tanimoto	$T_{a_x, a_y} T_{h_x, h_y} T_{b_x, b_y} T_{l_x, l_y}$
WL	$K_{a_x, a_y} K_{h_x, h_y} K_{b_x, b_y} K_{l_x, l_y}$

The descriptors must account for the missing molecules included in the proposed reactions. For the descriptors that represent a reaction in the form of a single vector (i.e. quantum chemical, concatenated molecular fingerprints and one-hot encodings), the bits corresponding to the missing molecules were set to zero. Where descriptors represent a single entity per reaction component, the missing molecules were incorporated in the calculation of the kernel of each reaction component. For example, the graph kernel between two molecules (A and B) represented by molecular graphs

is defined below.

$$K'_{A,B} = \begin{cases} K_{A,B} + 1, & \text{if } A \text{ and } B \text{ are both present} \\ 2, & \text{if } A \text{ and } B \text{ are both missing} \\ 1, & \text{otherwise} \end{cases}$$

If both molecules are present the kernel of the two molecules is the Weisfeiler-Lehman kernel plus one, if both molecules are missing the kernel equals two, otherwise the kernel is equal to one. This method is only applied to the datasets that include missing molecules.

2.4 Model Building and Evaluation

The hyperparameters of the SVR models (ϵ and C) were optimised in scikit-learn by performing an exhaustive grid-search over the specified parameter grid (see section 2.3) on the training set, using five-fold cross-validation. For each train-test split of the data, the training set was 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. This set of hyperparameters was used to build the SVR model on the training set for the particular train-test split.

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. Activity ranking was used to generate the additive and aryl halide test sets, to ensure the models were trained on a range of reaction yields.⁵³ The mean yields of the reactions containing each additive and aryl halide were ranked from lowest to highest. The highest and lowest yielding additives and aryl halides were included in all training sets for the additive and aryl halide tests, respectively. Test sets were constructed of every n^{th} molecule where $n = 4$ for the additives (Table S2) and $n = 3$ for the aryl halides (Table S3). Due to the small number of bases (three) and ligands (four) in the dataset, two leave-one-out test were performed. In the first leave-one-out 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. The performances of the regression models were evaluated by the coefficient of determination (R^2) and RMSE for data

points outside of the training set. The coefficient of determination can be negative if the mean of the data is a better fit to the observed values than the predicted values, i.e. $SS_{res} > SS_{tot}$, see Eq. 3.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{SS_{res}}{SS_{tot}} \quad (3)$$

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 values. The total sum of squares, SS_{tot} is proportional to the variance of the data. All analysis was performed in scikit-learn. SVR models built on one-hot encodings were used as a baseline, for comparison.

3 Results and Discussion

3.1 Diversity of the Train-Test Splits

The reactions in the Doyle et al.¹⁹ dataset cover a range of yields (Fig. 3), with the majority low yielding (0 to 10%) and few high yielding (90 to 100%). It is important in an assessment of model performance, to split the data into training and test sets that ensure an even spread of chemical reactivity is included in each. Chuang and Keiser have shown that splitting this dataset by high-throughput plate (where all inhibitory additives were present on a single plate) leads to an inaccurate estimation of performance due to an uneven cover of reaction yields (Fig. S6a†).⁵² A similar conclusion is expected when assessing the models along the aryl halide dimension, if split based on halide or ring type (Fig. S6b† and S6c†). Splitting data using activity ranking ensures models are trained and tested on similar distributions of reaction yields (Fig. S7). Details of the activity ranking test sets, including the calculated mean reaction yields are presented in the ESI†.

It is important to assess whether the reactions in the test set are within the domain of applicability. The similarity of test reactions to the training reactions was evaluated using the maximum product of pairwise Tanimoto scores, calculated using the Morgan2 fingerprint, of the reaction components. The maximum similarity to training reactions for the additive and aryl halide ranked tests range from 0.30 to 0.65 and from 0.30 to 0.55, respectively (Fig. 4). The models are expected to predict instances with low maximum similarity scores less accurately than those with high maximum similarity scores.

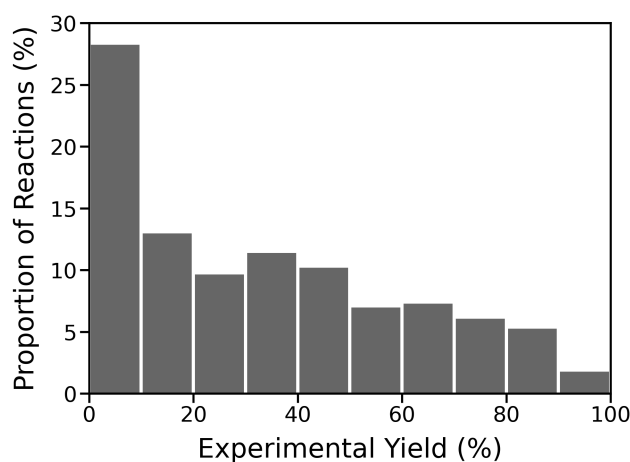


Fig. 3 Distribution of experimental yields, excluding control reactions and reactions containing 5-phenyl-1,2,4-oxadiazole (additive **7**), corresponding to 3955 data points.

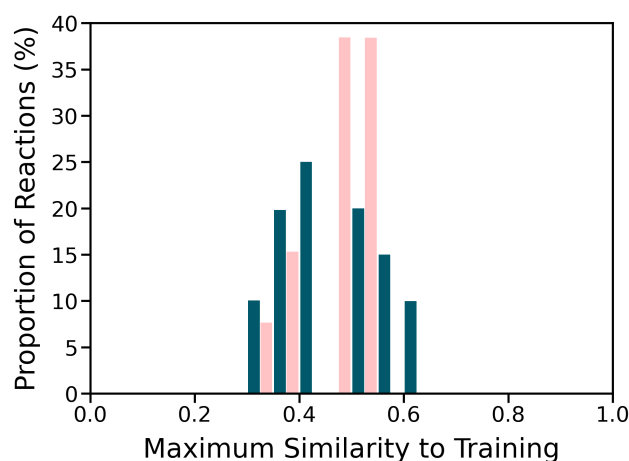


Fig. 4 Distributions of maximum similarity to training reactions for the additive ranked test sets (dark blue bars) and aryl halide ranked test sets (pale pink bars). Maximum similarity to training was calculated using the maximum product of pairwise Tanimoto scores (with the Morgan2 fingerprint) of the reaction components.

3.2 Prediction of Reaction Yield

The performance of the yield prediction models built on quantum chemical descriptors, molecular fingerprints and molecular graphs for the tests without activity ranking and with activity ranking are shown in Table 3 and Table 4, respectively. The performance metrics are reported as the average over the test sets for the specified split of the data. The effect of bit length and type of molecular fingerprint on the performance of the SVR models is presented and discussed in the ESI† (Table S4, Fig. S12, S13, S14, S15). The performance of the SVR 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 The following tests did not take into account the distribution of reaction yields in the training and test sets: additive plate split, aryl halide ring type and halide splits. The lower average performances of these tests (Table 3) in comparison to the activity ranked splits (Table 4) underscore the importance of test set design.^{52,53} 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 (Fig. S6†). The average model performances for the base and ligand leave-one-out tests were modest (Table 3, S11† and S12†). The SVR model built on the topological RDK fingerprint ($R^2 = 0.58$) is the only model to outperform the one-hot encoding model ($R^2 = 0.53$) in the base leave-one-out test. For the ligand leave-one-out test, the quantum chemical model has a poor performance across all ligand test sets (Table S12†). All models have a negative R^2 for the XPhos test set due to the uneven representation of yields in the training and test set (Fig. S5b†), which resulted in overprediction of the reaction yields.

Table 3 Mean Performance Statistics for the Top Reaction Yield Prediction Models Built Using the SVR Algorithm^b

Reaction Component	Split	Descriptor	Kernel	R^2	RMSE (%)
Additive	Plate	Quantum Chemical	Polynomial	0.25 (0.26)	22.6 (1.7)
		Morgan2	Tanimoto	0.54 (0.30)	17.3 (4.6)
		Graphs	WL	0.49 (0.34)	18.2 (4.7)
		One-hot Encodings	Polynomial	0.47 (0.28)	18.8 (3.9)
Aryl Halide	Ring Type	Quantum Chemical	RBF	-0.64 (0.65)	34.0 (12.8)
		MACCS Fingerprint	RBF	0.27 (0.08)	22.6 (5.3)
		Graphs	WL	-0.04 (0.27)	26.4 (1.4)
		One-hot Encodings	Polynomial	-0.21 (0.29)	28.6 (1.7)
	Halide	Quantum Chemical	RBF	-0.71 (0.82)	30.2 (5.1)
		Morgan2 Fingerprint	Tanimoto	-0.22 (0.98)	24.4 (7.9)
		Graphs	WL	-0.40 (1.22)	25.6 (9.8)
		One-hot Encodings	Polynomial	-0.45 (1.10)	26.8 (7.8)
Base	Leave-one-out	Quantum Chemical	RBF	-0.04 (0.08)	27.1 (4.0)
		RDK Fingerprint	RBF	0.58 (0.17)	16.8 (3.5)
		Graphs	WL	0.53 (0.20)	18.0 (5.6)
		One-hot Encodings	RBF	0.53 (0.20)	18.0 (4.8)
Ligand	Leave-one-out	Quantum Chemical	RBF	-0.27 (0.68)	27.2 (2.4)
		Morgan2 Fingerprint	Tanimoto	0.39 (0.68)	16.2 (6.2)
		Graphs	WL	0.38 (0.84)	15.3 (6.2)
		One-hot Encodings	RBF	0.38 (0.70)	16.6 (4.3)

^b R^2 and RMSE statistics are reported in the format "mean (standard deviation)" for the specified test sets.

With Activity Ranking The performance of the yield prediction models built on quantum chemical descriptors, molecular fingerprints and molecular graphs for the additive and aryl halide ranked

tests are shown in Table 4 and Fig. 5. The random forest model built on quantum chemical descriptors from Doyle et al.⁵³ 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.

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 sets a higher baseline for the SVR models (additive split: $R^2 < 0.69$, RMSE $> 15.2\%$; aryl halide split: $R^2 < 0.37$, RMSE $> 20.7\%$) 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 (Fig. S11), base or ligand (Fig. S5) 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 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, that 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 reaction yield, as the linear SVR model predictions show no statistical correlation. Non-linear kernels (polynomial, RBF and sigmoid) were considered, to transform the input data into higher dimensional feature space where regression could be performed linearly. The performance of the quantum chemical, concatenated molecular fingerprints and one-hot encoding SVR models, implementing the polynomial and RBF

kernels, tend to be higher than the linear and sigmoid kernels. The SVR algorithm performs better with structure-based descriptors (molecular fingerprints and molecular graphs) compared to the quantum chemical descriptors. It is encouraging that the Morgan fingerprints capture enough chemical information that they out-perform the quantum chemical descriptors which were adopted by Doyle et al.¹⁹

Table 4 Mean Performance Statistics for the Reaction Yield Prediction Models Built Using the SVR Algorithm and Baseline Random Forest (RF) Models in the Activity Ranked Tests^b

Descriptor	Kernel	Additive Split		Aryl Halide Split	
		R^2	RMSE (%)	R^2	RMSE (%)
Quantum Chemical	Linear	0.06 (0.41)	26.0 (5.7)	< -1 (>1)	> 100 (> 100)
	Polynomial	0.60 (0.07)	17.3 (1.8)	-0.57 (1.50)	29.9 (13.5)
	RBF	0.53 (0.10)	18.5 (2.3)	0.36 (0.11)	20.8 (2.5)
	Sigmoid	0.35 (0.04)	21.9 (1.2)	0.04 (0.26)	25.2 (2.3)
Morgan1	Linear	0.50 (0.09)	19.3 (2.1)	0.56 (0.05)	17.3 (1.4)
	Polynomial	0.71 (0.11)	14.4 (2.9)	0.68 (0.02)	14.8 (0.9)
	RBF	0.70 (0.12)	14.8 (3.1)	0.70 (0.03)	14.2 (1.1)
	Sigmoid	0.44 (0.05)	20.4 (1.2)	0.40 (0.06)	20.1 (1.2)
	Tanimoto	0.73 (0.12)	13.9 (3.3)	0.65 (0.05)	15.3 (1.0)
Morgan2	Linear	0.57 (0.08)	17.9 (2.0)	0.52 (0.06)	18.0 (1.5)
	Polynomial	0.72 (0.10)	14.4 (2.8)	0.65 (0.05)	15.4 (1.6)
	RBF	0.71 (0.11)	14.4 (3.0)	0.65 (0.06)	15.4 (1.6)
	Sigmoid	0.44 (0.05)	20.4 (1.2)	0.39 (0.03)	20.3 (1.1)
	Tanimoto	0.72 (0.11)	14.4 (2.8)	0.61 (0.03)	16.3 (1.1)
Morgan3	Linear	0.61 (0.09)	17.0 (2.1)	0.52 (0.07)	18.0 (1.8)
	Polynomial	0.73 (0.10)	14.1 (2.8)	0.62 (0.08)	16.0 (2.2)
	RBF	0.73 (0.11)	14.0 (3.0)	0.63 (0.08)	15.9 (2.1)
	Sigmoid	0.43 (0.05)	20.5 (1.1)	0.37 (0.03)	20.7 (1.3)
	Tanimoto	0.70 (0.10)	14.7 (2.5)	0.55 (0.04)	17.4 (1.3)
MACCS	Linear	0.50 (0.11)	22.1 (1.4)	0.52 (0.01)	18.0 (0.7)
	Polynomial	0.71 (0.11)	19.0 (2.1)	0.63 (0.02)	15.8 (0.9)
	RBF	0.70 (0.10)	19.4 (1.9)	0.60 (0.05)	16.5 (1.4)
	Sigmoid	0.44 (0.06)	23.7 (1.1)	0.21 (0.08)	23.0 (0.9)
	Tanimoto	0.56 (0.13)	17.9 (2.7)	0.61 (0.04)	16.3 (1.2)
RDK	Linear	0.55 (0.05)	18.3 (1.1)	0.52 (0.04)	17.9 (1.1)
	Polynomial	0.63 (0.07)	16.6 (1.4)	0.64 (0.09)	15.5 (2.1)
	RBF	0.63 (0.07)	16.6 (1.4)	0.64 (0.09)	15.4 (2.1)
	Sigmoid	0.26 (0.04)	23.4 (0.6)	0.24 (0.07)	22.7 (1.1)
	Tanimoto	0.63 (0.05)	16.6 (1.3)	0.57 (0.13)	17.0 (2.4)
Graphs	WL	0.67 (0.18)	15.3 (4.2)	0.61 (0.08)	16.2 (1.7)
One-hot Encodings	Linear	0.59 (0.05)	17.4 (1.5)	0.31 (0.06)	21.6 (0.5)
	Polynomial	0.68 (0.05)	15.4 (1.5)	0.35 (0.04)	20.9 (0.5)
	RBF	0.69 (0.05)	15.2 (1.6)	0.37 (0.05)	20.7 (0.6)
	Sigmoid	0.49 (0.03)	19.4 (1.0)	0.24 (0.03)	22.6 (0.4)
Quantum Chemical RF	N/A	0.68 (0.11)	15.3 (3.1)	0.21 (0.09)	23.0 (0.6)
One-hot Encodings RF	N/A	0.59 (0.11)	17.4 (2.8)	-0.04 (0.34)	26.2 (3.5)

^b 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 S5†, S6†, S7†, S8†

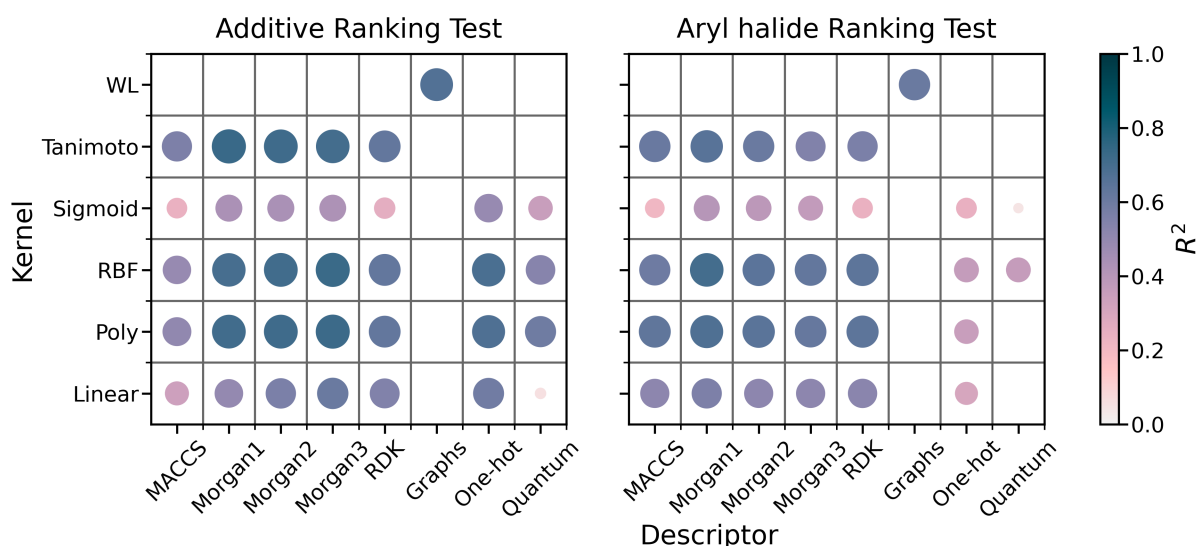


Fig. 5 Predictive accuracy (coefficient of determination) comparison of the SVR models built on quantum chemical descriptors, molecular fingerprints (Morgan1-3, RDKit, MACCS), molecular graphs and one-hot encodings with a range of kernels, in the activity ranked tests. Marker size is proportional to R^2 . Numeric values can be found in Table 4.

The best descriptor-kernel combinations for the additive ranked split were Quantum-Polynomial, Morgan1-Tanimoto, Graphs-WL and One-hot-RBF, with R^2 in the range 0.60 to 0.73 and RMSE 17.3% to 13.9% (Table 4). Each of these best performing models are significantly different to one another, according to the χ^2 test (p -value $< 10^{-5}$ for all combinations, Table S9) under the null hypothesis that the distributions of the residual yield (Fig. S16a) are the same. The best quantum chemical SVR model (Quantum-Polynomial) has a wider distribution of residual yields than the other highest performing SVR models (Fig. S16a), meaning larger associated errors. The random forest algorithm learns more from the quantum chemical descriptors ($R^2 = 0.68$, RMSE = 15.3%) than the SVR algorithm ($R^2 \leq 0.60$, RMSE $\geq 17.3\%$). The performance of the Morgan fingerprints are relatively robust, with only minor variations depending on the choice of kernel and radius. The Morgan1-Tanimoto, Morgan3-Polynomial and Morgan3-RBF SVR models have the highest R^2 of 0.73 and RMSE of 13.9%, 14.1% and 14.0%, respectively. Morgan1-Tanimoto was chosen as the best SVR model built on molecular fingerprints, for further analysis, as it has the lowest RMSE score and required the smallest radius of neighbouring atoms to be encoded in the fingerprints.

Model performances along the aryl halide dimension were lower than along the additive dimension for the baseline and quantum chemical models (Table 4, Fig. 5). Models built on structure-based descriptors had a similar performance to those in the additive ranked test. The best descriptor-kernel combinations for the aryl halide ranked test were Quantum-RBF, Morgan1-RBF,

Graphs-WL and One-hot-RBF. There is a large difference in performance between the structure-based descriptors, with an R^2 of 0.70 and 0.61 for the Morgan1-RBF and Graphs-WL respectively, compared to the Quantum-RBF (0.36) and One-hot-RBF (0.37) models. The similarity in performance between the quantum chemical and one-hot encoding models suggests that the quantum chemical models may only be fitting the intrinsic pattern in the training set and therefore, struggle to extrapolate to unseen aryl halides. In general, there is an even distribution of residual yields around 0% for the best descriptor-kernel combinations (Fig. S16b). However, the models have a tendency 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 (Fig. S16b). This is partially due to the under-representation of higher reaction yields, resulting in poorer model performances (Fig. S17b). This issue is also observed in the additive ranked test (Fig. S17a).

3.3 Prediction Performance by Similarity to Training Data

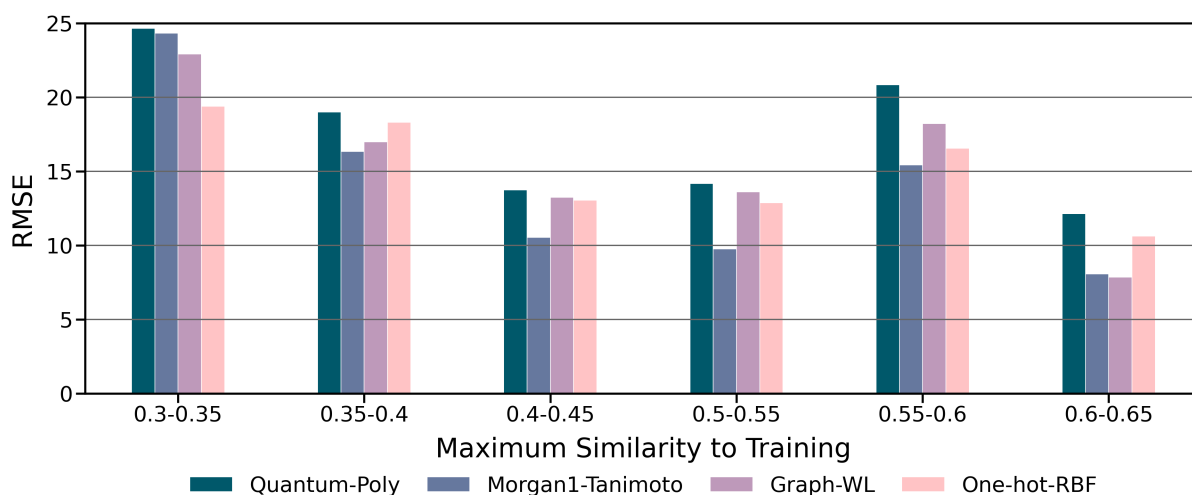
Assessing model performance with respect to maximum similarity to training reactions helps to identify molecules that may be outside the domain of applicability. Maximum similarity to training is defined as the maximum product of pairwise Tanimoto scores (between molecules in the training and test sets) of the reaction components. 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, then the similarity score would only be dependent on the additives in the reactions as shown in Eq. 4. The maximum similarity to training scores of the additives and aryl halides for both activity ranking tests can be found in Table S10 in the ESI†.

$$T_{r_1, r_2} = T_{a_1, a_2} T_{h_1, h_1} T_{b_1, b_1} T_{l_1, l_1} = T_{a_1, a_2} \cdot 1 \cdot 1 \cdot 1 = T_{a_1, a_2} \quad (4)$$

In the additive ranked test, the models performed poorly for reactions in the lowest maximum similarity to training interval, 0.30 to 0.35 (Fig. 6a, S18a†). 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** (Fig. S19d) but

very poor for additive **10** (Fig. S19a). 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 (Fig. 6a, S18a†), as expected. The models have a high RMSE ($\geq 15\%$) for the reactions in the maximum similarity to training intervals 0.35 to 0.40 (additives **1**, **3**, **5**, **14**) and 0.55 to 0.60 (additives **4**, **6**, **9**). This is mainly due to the underprediction of high yielding reactions, which is a result of the under-representation of higher reaction yields (Fig. 3, S17a). The Morgan1-Tanimoto, Graphs-WL and One-hot-RBF SVR models demonstrate good prediction statistics for reactions with a maximum similarity to training greater than 0.35.

(a)



(b)

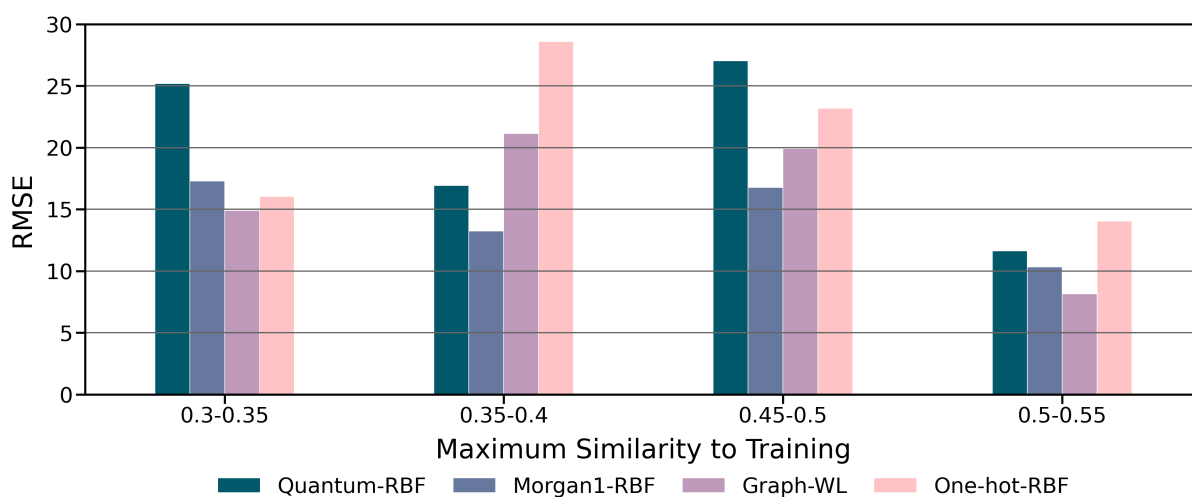


Fig. 6 RMSE against maximum similarity to training for (a) the additive ranked test and (b) the aryl halide ranked test.

For the aryl halide ranked test, there are no obvious trends between maximum similarity to

training and the performance statistics (Fig. 6b, S18b†). The higher yielding reactions containing ethyl substituted aryl halides (0.30 to 0.40) and halopyridines (0.45 to 0.50) are underpredicted by the models (Fig. S20†), due to the under-representation of higher reaction yields (Fig. 3, S17b). Reactions containing the trifluoromethyl and methoxy substituted aryl halides (0.50 to 0.55) are generally predicted well by the models. It is important to consider the coefficient of determination (R^2) and RMSE together, when assessing goodness of fit.⁶⁷ 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 (Fig. S20), the RMSE scores are good ($\leq 15\%$) for at least half of the models.

3.4 Predictions of Prospective Reactions

A set of combinatorial reactions was compiled to validate the generalisability of the SVR models, particularly along the aryl halide dimension. Here, we present predicted yields of these reactions prior to experimentation. The SVR model with the best predictive performance for each descriptor in the aryl halide ranked test was employed: Quantum-RBF, Morgan1-RBF, Graph-WL and the One-hot-RBF baseline. The aryl halides in the prospective reactions cover a range of maximum similarity to training scores between 0.15 to 0.55 (Fig. 7). 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 (Fig. 6b). 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). All models except Morgan1-RBF were identified as the best kernel-descriptor combinations in the base and ligand leave-one-out tests. The Morgan1-RBF model showed comparable correlation to the top SVR model built on molecular fingerprints in both leave-one-out tests (Table S11, S12†). 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.

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.

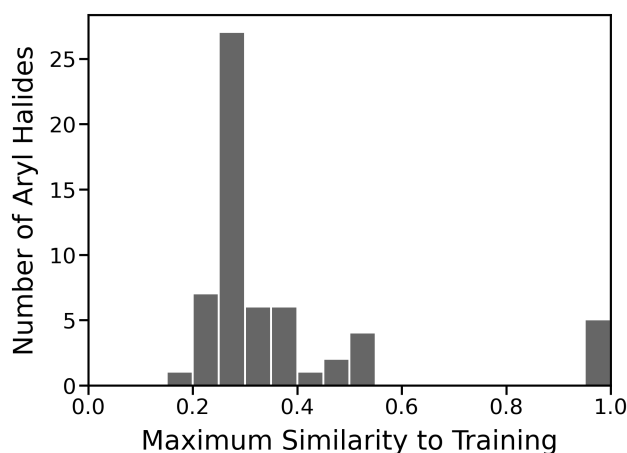


Fig. 7 Distributions of maximum similarity to training for all prospective reactions. Maximum similarity to training was calculated using the maximum pairwise Tanimoto scores (using the Morgan2 fingerprint) of the aryl halides in the training and test set.

These descriptors can be applied to any molecule and are quick and easy to calculate. In this test, the Morgan1-RBF, Graph-WL and One-hot-RBF 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 ESI† for 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 Fig. S21, S22†.

The models built on chemically meaningful descriptors predicted the yield of in-sample reactions, present in both the training and test set, accurately, with $R^2 \geq 0.95$ and $RMSE \leq 6.6\%$ (Fig. S23, S24†). In both tests, the one-hot encodings model showed negligible coefficient of determination ($R^2 \leq 0.02$) between the experimental and predicted yields. An arbitrary number was predicted irrespective of the aryl halide present in the reaction. The predictions were primarily dependent on the type of base in the reaction, as shown by the three clear peaks in the distribution of predicted reaction yield (Fig. 8, S25†). The peaks at approximately 35%, 40% and 50% in the distribution of predicted yield for the subset of proposed reactions (Fig. 8), correspond to DBU,

BTMG and MTBD, respectively (Fig. S26†). The same trend was observed when all prospective reactions were considered (Fig. S27†). There is minimal difference in the distributions of 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 and could not fit any 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.

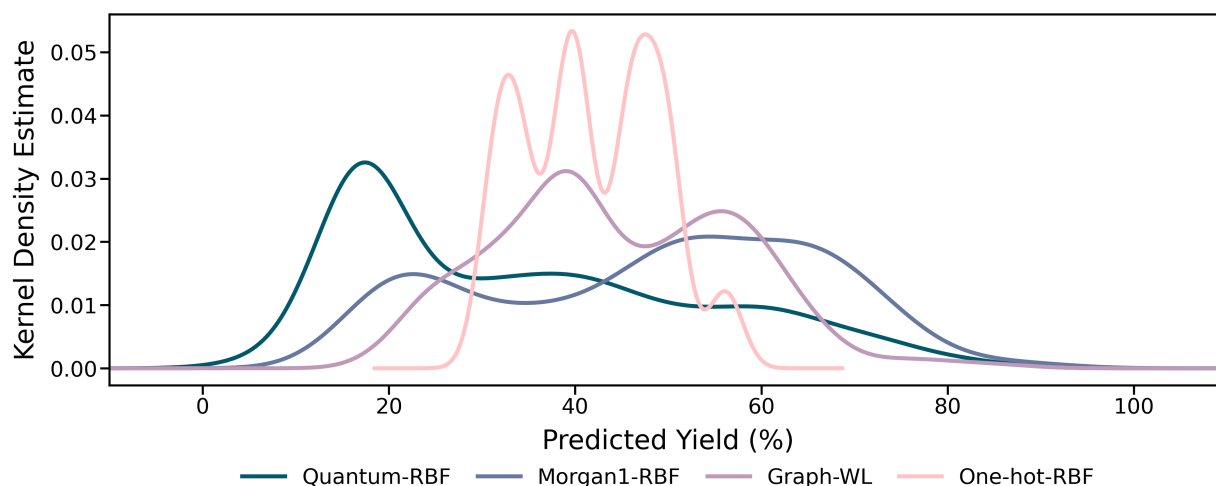


Fig. 8 Distributions of predicted reaction yield for the subset (882) of validation reactions.

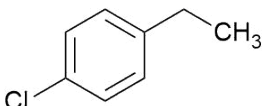
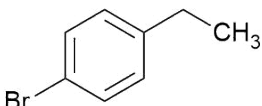
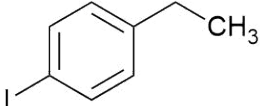
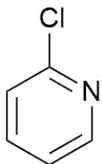
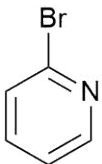
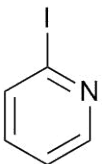
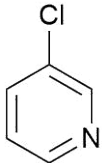
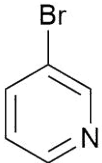
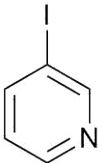
Reactions performed without a catalyst were included in the prospective reactions to evaluate the following synthetic hypothesis; the reactions containing *ortho*-substituted halopyridines are proceeding via an alternative reaction pathway, leading to higher reaction yields. No examples of these reactions were provided in training and therefore, may be beyond the limits of the models. The Quantum-RBF model predicted the yield of reactions without the presence of a catalyst to be 17.1%, irrespective of the other reaction components. Predictions of this arbitrary number suggest these reactions are outside the domain of applicability for the Quantum-RBF model. These predictions largely contributed to the distinct peak in the distribution of predicted yield between 15% and 20% (Fig. 8). The graph-based model predicted a smaller range of yields for reactions performed without a catalyst ($\sim 30\%$) compared to the reactions containing a catalyst ($\gtrsim 50\%$, Fig. S28, S29†). This could indicate a potential limitation in the ability of the Graph-WL model to predict reactions without a catalyst. The chemically meaningful models predicted similar trends in the reactivity of the catalyst ligands (Fig. S28, S29†), 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 (Fig. S30, S31†). 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). This trend is plausible, as it follows the trend in the training reactions (Fig. S6c†). Comparing the mean yield of reactions containing 1-chloro-4-isopropylbenzene (~ 30% to 45%) with a similar alkyl-substituted aryl halide used in training (1-chloro-4-ethylbenzene, ~ 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). 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).

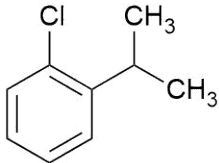
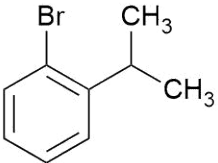
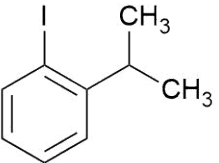
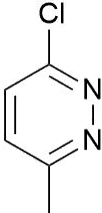
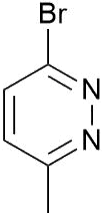
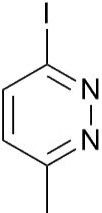
Despite the similar trends between the quantum chemical model and the structure-based models, the predictions are only slightly correlated (Pearson correlation coefficient of 0.53 with Morgan1-RBF and 0.68 with Graph-WL, Fig. S32†). The structure-based models are expected to be more robust than the quantum chemical models for extrapolating to unseen chemical entities. The predictions of the two structure-based models are reasonably correlated and have a Pearson correlation coefficient of 0.90 (Fig. S33†).

Table 5 Mean Experiment Yields of Aryl Halides in the Training Set (Top) and Mean Predicted Yields of Aryl Halides in the Prospective Reactions (Bottom)^c

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

		
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

						
	Subset	All	Subset	All	Subset	All
Quantum-RBF	31.2 (3.6)	N/A	64.6 (6.8)	N/A	N/A	N/A
Morgan1-RBF	39.6 (4.0)	35.3 (9.7)	62.9 (4.3)	56.5 (12.7)	N/A	60.5 (10.5)
Graph-WL	45.9 (3.7)	45.7 (9.0)	57.3 (4.3)	55.2 (10.4)	N/A	56.9 (9.6)
One-hot-RBF	42.5 (7.8)	46.0 (8.8)	42.5 (7.8)	46.0 (8.8)	N/A	46.0 (8.8)
						
	Subset	All	Subset	All	Subset	All
Quantum-RBF	10.8 (5.4)	N/A	25.2 (3.3)	N/A	N/A	N/A
Morgan1-RBF	48.5 (4.9)	42.1 (12.4)	63.5 (4.5)	56.9 (12.3)	N/A	61.7 (9.9)
Graph-WL	51.2 (4.2)	47.0 (10.0)	58.6 (4.3)	56.3 (10.6)	N/A	56.5 (9.9)
One-hot-RBF	42.5 (7.8)	46.0 (8.8)	42.5 (7.8)	46.0 (8.8)	N/A	46.0 (8.8)

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

4 Conclusions

SVR models built on structure-based and quantum chemical descriptors, for the prediction of reaction yield, were compared. The models were applied to a set of Buchwald-Hartwig reactions and the performance was assessed along the dimension of each reaction component. The models built on structure-based descriptors (molecular fingerprints and graphs) demonstrated good prediction statistics and outperformed the quantum chemical SVR models, along the dimension of each reaction component. The models built on molecular fingerprints consistently surpassed the other descriptors in each test, proving fingerprints to be robust descriptors. The moderate performances of the SVR models in the base and ligand leave-one-out tests, suggest they may benefit from including a larger variety of bases and ligands in training. The applicability, ease and quickness of calculating molecular fingerprints makes them particularly attractive (Table 6).

Table 6 Comparison of the Molecular Descriptors used in this Study^d

Descriptor	Representation	Speed	Applicability to molecules	Generalisability	
				Additive	Aryl Halide
Quantum Chemical	Calculated Properties	+	Subset	+	+
Molecular Fingerprints	Structural Topology	+++	All	++++	++++
Molecular Graphs	Structural Topology	+++	All	++	+++
One-hot Encodings	Presence or Absence	++++	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.

Predictions of reaction yield for the proposed reactions, not present in the Doyle et al. dataset, were reported prior to experimentation. Similar trends in the reactivity of the molecules along each reaction component were observed across the chemically meaningful models. The reaction yields predicted by the structure-based models are reasonably correlated. Based on the performance of the models in the preceding tests and the analysis of the predicted yields of the proposed reactions, it is anticipated that the structure-based descriptors will extrapolate better than the quantum chemical model. The reaction yields of the proposed reactions will be attained using high-throughput experimentation, and used to validate and assess the limits of the SVR models.

Overall the results presented show the applicability of the structure-based SVR models to the prediction of reaction yield, across all dimensions of a single reaction class. The machine learning models learnt from a relatively small (a few thousand instances) combinatorial dataset, proving their use in facilitating the optimisation of reaction conditions for the synthesis of new molecules.

In the future, it would be interesting to explore the transferability of the structure-based SVR models to different reaction types or other regression related problems.

Data and Software Availability

The data and code used in this study is available online at https://github.com/alexehaywood/yield_prediction. The prospective reactions and corresponding predictions are included in the supporting information.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/S035990/1] and by the University of Nottingham Green Chemicals Beacon. ALH is supported by EPSRC/NPIF [EP/S515516/1] through a CASE PhD studentship in partnership with GSK. JR is supported by an NPIF PhD studentship through the MRC IMPACT Doctoral Training Programme [MR/S502431/1]. JDH is supported by the Royal Academy of Engineering under the Chairs in Emerging Technologies scheme. We are grateful for access to the University of Nottingham High Performance Computer.

References

- [1] Mitchell, J. B. O. Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 468–481.
- [2] Griffen, E. J.; Dossetter, A. G.; Leach, A. G.; Montague, S. Can we accelerate medicinal chemistry by augmenting the chemist with Big Data and artificial intelligence? *Drug Discov. Today* **2018**, *23*, 1373–1384.
- [3] Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477.

- [4] Haywood, A. L.; Redshaw, J.; Gärtner, T.; Taylor, A.; Mason, A. M.; Hirst, J. D. In *Machine Learning in Chemistry: The Impact Artificial Intelligence*; Cartwright, H., Ed.; The Royal Society of Chemistry, 2020; pp 169–194.
- [5] Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* **2019**, *4*, 828–849.
- [6] Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610.
- [7] Ishida, S.; Terayama, K.; Kojima, R.; Takasu, K.; Okuno, Y. Prediction and Interpretable Visualization of Retrosynthetic Reactions Using Graph Convolutional Networks. *J. Chem. Inf. Model.* **2019**, *59*, 5026–5033.
- [8] Gajewska, E. P.; Szymkuć, S.; Dittwald, P.; Startek, M.; Popik, O.; Mlynarski, J.; Grzybowski, B. A. Algorithmic Discovery of Tactical Combinations for Advanced Organic Syntheses. *Chem* **2020**, *6*, 280–293.
- [9] Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y. Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *J. Chem. Inf. Model.* **2020**, *60*, 47–55.
- [10] Lee, A. A.; Yang, Q.; Sresht, V.; Bolgar, P.; Hou, X.; Klug-McLeod, J. L.; Butler, C. R. Molecular Transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chem. Commun.* **2019**, *55*, 12152–12155.
- [11] Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. “Found in Translation”: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **2018**, *9*, 6091–6098.
- [12] Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **2019**, *10*, 370–377.
- [13] Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **2018**, *4*, 1465–1476.

- [14] Walker, E.; Kammeraad, J.; Goetz, J.; Robo, M. T.; Tewari, A.; Zimmerman, P. M. Learning To Predict Reaction Conditions: Relationships between Solvent, Molecular Structure, and Catalyst. *J. Chem. Inf. Model.* **2019**, *59*, 3645–3654.
- [15] Maser, M. R.; Cui, A. Y.; Ryou, S.; DeLano, T. J.; Yue, Y.; Reisman, S. E. Multilabel Classification Models for the Prediction of Cross-Coupling Reaction Conditions. *J. Chem. Inf. Model.* **2021**, acs.jcim.0c01234.
- [16] Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **2018**, *23*, 1538–1546.
- [17] Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* **2011**, *10*, 188–195.
- [18] Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **2018**, *559*, 377–381.
- [19] Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **2018**, *360*, 186–190.
- [20] Smola, A. J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput.* **2004**, *14*, 199–222.
- [21] Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR Classifiers Compared. *J. Chem. Inf. Model.* **2007**, *47*, 219–227.
- [22] Kountouris, P.; Hirst, J. D. Predicting β -turns and their types using predicted backbone dihedral angles and secondary structures. *BMC Bioinformatics* **2010**, *11*, 407.
- [23] Kriege, N. M.; Johansson, F. D.; Morris, C. A survey on graph kernels. *Appl. Netw. Sci.* **2020**, *5*, 6.
- [24] Swamidass, S. J.; Chen, J.; Bruand, J.; Phung, P.; Ralaivola, L.; Baldi, P. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics* **2005**, *21*, i359–i368.

- [25] Fröhlich, H.; Wegner, J. K.; Sieker, F.; Zell, A. Kernel Functions for Attributed Molecular Graphs – A New Similarity-Based Approach to ADME Prediction in Classification and Regression. *QSAR Comb. Sci.* **2006**, *25*, 317–326.
- [26] Ivanciuc, O. In *Rev. Comput. Chem.*, volume 23 ed.; Lipkowitz, K. B., Cundari, T. R., Eds.; Wiley-VCH, 2007; Chapter 6, pp 291–400.
- [27] Miyao, T.; Funatsu, K.; Bajorath, J. Exploring Alternative Strategies for the Identification of Potent Compounds Using Support Vector Machine and Regression Modeling. *J. Chem. Inf. Model.* **2019**, *59*, 983–992.
- [28] Miyao, T.; Funatsu, K. Iterative Screening Methods for Identification of Chemical Compounds with Specific Values of Various Properties. *J. Chem. Inf. Model.* **2019**, *59*, 2626–2641.
- [29] Lu, Y.; Anand, S.; Shirley, W.; Gedeck, P.; Kelley, B. P.; Skolnik, S.; Rodde, S.; Nguyen, M.; Lindvall, M.; Jia, W. Prediction of p K a Using Machine Learning Methods with Rooted Topological Torsion Fingerprints: Application to Aliphatic Amines. *J. Chem. Inf. Model.* **2019**, *59*, 4706–4719.
- [30] Singh, S.; Pareek, M.; Changotra, A.; Banerjee, S.; Bhaskararao, B.; Balamurugan, P.; Sunoj, R. B. A unified machine-learning protocol for asymmetric catalysis as a proof of concept demonstration using asymmetric hydrogenation. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 1339–1345.
- [31] Reid, J. P.; Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **2019**, *571*, 343–348.
- [32] Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **2019**, *363*, eaau5631.
- [33] Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. Why Are Some Properties More Difficult To Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P. *J. Chem. Inf. Model.* **2008**, *48*, 220–232.
- [34] Harding, A. P.; Wedge, D. C.; Popelier, P. L. A. p K a Prediction from “Quantum Chemical Topology” Descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 1914–1924.

- [35] Dong, N.; Lu, W.-c.; Chrn, N.-y.; Zhu, Y.-c.; Chen, K.-x. Using support vector classification for SAR of fentanyl derivatives¹. *Acta Pharmacol. Sin.* **2005**, *26*, 107–112.
- [36] Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
- [37] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- [38] Landrum, G. A. RDKit: Open-Source Cheminformatics. <http://www.rdkit.org>.
- [39] Melville, J. L.; Burke, E. K.; Hirst, J. D. Machine Learning in Virtual Screening. *Comb. Chem. High Throughput Screen.* **2009**, *12*, 332–343.
- [40] Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **2020**, *6*, 1379–1390.
- [41] Gärtner, T.; Flach, P.; Wrobel, S. *Learn. Theory Kernel Mach.*; Springer Verlag, 2003; Vol. 277; pp 129–143.
- [42] Marcou, G.; Aires de Sousa, J.; Latino, D. A. R. S.; de Luca, A.; Horvath, D.; Rietsch, V.; Varnek, A. Expert System for Predicting Reaction Conditions: The Michael Reaction Case. *J. Chem. Inf. Model.* **2015**, *55*, 239–250.
- [43] Collins, K. D.; Gensch, T.; Glorius, F. Contemporary screening approaches to reaction discovery and development. *Nat. Chem.* **2014**, *6*, 859–871.
- [44] Buitrago Santanilla, A. et al. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **2015**, *347*, 49–53.
- [45] Henle, J. J.; Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E. Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis. *J. Am. Chem. Soc.* **2020**, *142*, 11578–11592.
- [46] Torborg, C.; Beller, M. Recent Applications of Palladium-Catalyzed Coupling Reactions in the Pharmaceutical, Agrochemical, and Fine Chemical Industries. *Adv. Synth. Catal.* **2009**, *351*, 3027–3043.

- [47] Magano, J.; Dunetz, J. R. Large-Scale Applications of Transition Metal-Catalyzed Couplings for the Synthesis of Pharmaceuticals. *Chem. Rev.* **2011**, *111*, 2177–2250.
- [48] Ruiz-Castillo, P.; Buchwald, S. L. Applications of Palladium-Catalyzed C-N Cross-Coupling Reactions. *Chem. Rev.* **2016**, *116*, 12564–12649.
- [49] Vitaku, E.; Smith, D. T.; Njardarson, J. T. Analysis of the Structural Diversity, Substitution Patterns, and Frequency of Nitrogen Heterocycles among U.S. FDA Approved Pharmaceuticals. *J. Med. Chem.* **2014**, *57*, 10257–10274.
- [50] Collins, K. D.; Glorius, F. Intermolecular Reaction Screening as a Tool for Reaction Evaluation. *Acc. Chem. Res.* **2015**, *48*, 619–627.
- [51] Zahrt, A. F.; Henle, J. J.; Denmark, S. E. Cautionary Guidelines for Machine Learning Studies with Combinatorial Datasets. *ACS Comb. Sci.* **2020**, *22*, 586–591.
- [52] Chuang, K. V.; Keiser, M. J. Comment on “Predicting reaction performance in C–N cross-coupling using machine learning”. *Science* **2018**, *362*, eaat8603.
- [53] Estrada, J. G.; Ahneman, D. T.; Sheridan, R. P.; Dreher, S. D.; Doyle, A. G. Response to Comment on “Predicting reaction performance in C–N cross-coupling using machine learning”. *Science* **2018**, *362*, eaat8763.
- [54] Hasegawa, K.; Funatsu, K. Non-Linear Modeling and Chemical Interpretation with Aid of Support Vector Machine and Regression. *Curr. Comput. Aided-Drug Des.* **2010**, *6*, 24–36.
- [55] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.
- [56] National Cancer Institute Computer-Aided Drug Design (NCI/CADD) group (2009–2020), Chemical Identifier Resolver. 2020; <https://cactus.nci.nih.gov/chemical/structure>.
- [57] Shao, Y. et al. Advances in methods and algorithms in a modern quantum chemistry program package. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172–3191.
- [58] Shao, Y. et al. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* **2015**, *113*, 184–215.

- [59] Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- [60] Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.
- [61] Siglidis, G.; Nikolentzos, G.; Limnios, S.; Giatsidis, C.; Skianis, K.; Vazirgianis, M. GraKeL: A Graph Kernel Library in Python. *J. Mach. Learn. Res.* **2020**, *21*, 1–5.
- [62] Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- [63] Shervashidze, N.; Schweitzer, P.; Jan van Leeuwen, E.; Mehlhorn, K.; Borgwardt, K. M. Weisfeiler-Lehman Graph Kernels. *J. Mach. Learn. Res.* **2011**, *12*, 2539–2561.
- [64] Schölkopf, B. Support Vector Learning. Ph.D. thesis, TU Berlin, 1997.
- [65] Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **2015**, *7*, 20.
- [66] Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science* **1960**, *132*, 1115–1118.
- [67] Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R^2 : Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model.* **2015**, *55*, 1316–1322.