Kinetic predictions for E2 and S_N2 reactions using the BERT architecture: Comparison and interpretation



Chloe A. Wilson

Lady Margaret Hall

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2024

Abstract

The accurate prediction of reaction rates is an integral step in reaction mechanism elucidation and design of synthetic pathways. Traditionally, kinetic parameters have been derived from activation energies obtained from Quantum Mechanics (QM) methods, and more recently, machine learning (ML). Transformer-based models have emerged as a popular ML method for chemical prediction, allowing reactions to be represented as SMILES strings. Despite being state-of-the-art in both reaction classification and yield prediction, Bidirectional Encoder Representations from Transformers (BERT), a type of transformer-based model, have not yet been applied to kinetic prediction. In this Thesis, a BERT model is trained to predict experimental $\log k$ values of E2 and S_N2 reactions, and its predictions are interpreted in the context of known reactivity rules. Initially, BERT is trained to predict $\log k$ of S_N2 reactions and its performance is compared to the state-of-the-art Random Forest (RF) model in terms of accuracy, training time, and recognition of key structural and physical effects (Chapter 3). This model is subsequently fine-tuned to predict $\log k$ of both E2 and S_N2 reactions, and its predictions are interpreted regarding known drivers of mechanistic competition (Chapter 4). The final model exhibits strong performance, achieving near-experimental accuracy and identifying key structural and physical effects. Overall, this Thesis presents a framework for the accurate and interpretable prediction of reaction kinetics, with potential applications in guiding synthetic design.

Declaration

I, Chloe Wilson, declare that the Thesis I am submitting is entirely my own work except where indicated in the text, caption, footnote, or bibliography. This Thesis has not been submitted in whole or in part for any other academic degree or professional qualification.



27th November 2024

Signature *Date*

Acknowledgements

I would like to thank my supervisors, Profs. Fernanda Duarte and Jason Crain, for their guidance and support throughout my DPhil, and for giving me the opportunity to work on this project. The skills I have developed under Fernanda and Jason's mentorship are invaluable, and I will be forever grateful to them for their belief in me. To Fernanda, thank you for welcoming me into your research group and for helping me to grow as a scientist. Fernanda taught me how to shape my ideas and findings into a cohesive project, and consistently encouraged me to refine my work. To Jason, thank you for taking the time to provide valuable technical insights. Jason's role in introducing me to the BERT architecture and IBM Cloud platform have been instrumental in the progression of this project.

I cannot thank my supervisors without also acknowledging my unofficial supervisors: Matina, Ewa, Tom W, and James - past and present members of the Duarte group - without whom, none of this work would have been possible. Matina, for always being just a video call away whenever I needed her. Ewa and Tom, for staying behind after my presentations to offer incredibly detailed (and useful!) advice. And James, for his abundance of scientific knowledge, which he has always kindly shared.

Additional thanks to Matina and Tom W, as well as previous Duarte group members Ali and Tom Y, for their incredible help during my first year of DPhil. They were patient with me while I got to grips with the technical aspects of the project. I am especially grateful for their support during the pandemic, as many of them had not yet met me in person, but still took the time to advise me over video call.

I would also like to thank María Calvo, whom I supervised during her summer project in the Duarte group. The results from the QM calculations performed by María during her summer project are presented in this Thesis (María's contributions are indicated in the relevant sections). Additionally, I would like to thank Matina,

Tom W, Ewa, James, Sara, Wojtek, Bastian, and Fernanda for proofreading and providing feedback on the chapters of this Thesis.

In addition to the Duarte group members mentioned above, I would like to thank Shoubhik and Khan, for their warm friendship. Tomasz, for always checking in. Ally and Hanwen, for our insightful chats over lunch or coffee. Henry, for his unwavering kindness. Bernie, for baking wonderful cakes. Everyone on the badminton chat, for giving me something to look forward to every weekend. And all of the other lovely members (and honorary members) of the Duarte group who made the office a welcoming place to work.

I cannot thank my research group without also acknowledging the wonderful members of Lady Margaret Hall, who provided me with a supportive environment outside of work. I am fortunate to have been part of such a welcoming and inclusive community, and am grateful for the many opportunities to engage with college events.

I also owe a huge amount of gratitude to my partner Joe, who went above and beyond to support me throughout my DPhil. From taking on additional chores at home while I was writing this Thesis, to providing ongoing emotional support. Similarly, I would like to thank Mum, Dad, and my sister Hattie for taking me in during the final weeks of writing this Thesis, and for their help in proofreading the final version. Finally, I would like to thank Nan and Barney for all of their love and support.

Data availability

The data and code for the work entitled "Interpretable predictions of S_N2 kinetics using the BERT and RF architectures: Comparison to known reactivity rules" (Chapter 3) can be accessed via: `https://github.com/C-Wils/InterpretingMLKinetics.git`.

The data and code for the work entitled "Interpretable predictions of E2/S_N2 kinetics using the BERT architecture: Analysing drivers of mechanistic competition" (Chapter 4) can be accessed via: `https://github.com/C-Wils/E2SN2BERT.git`.

Table of Contents

List of Abbreviations xi

Glossary xvi

1 Introduction **1**

1.1 Transition state theory 1

1.1.1 Computational approaches for kinetic prediction 1

1.1.2 Kinetic ML models trained on QM-generated data . . . 2

1.1.3 Kinetic ML models trained on experimental data 17

1.2 BERT 23

1.2.1 BERT in Chemistry 24

1.2.2 BERT in kinetic prediction 25

1.3 Interpretable ML 26

1.3.1 Interpretable ML in Chemistry 27

1.3.2 Interpretable ML in kinetic prediction 30

1.4 E2 and S_N2 reactions 35

1.4.1 S_N2 reactions 35

1.4.2 E2 reactions 38

1.4.3 E2/S_N2 competition 41

1.5 Contributions of this Thesis 43

2 Theory **45**

2.1 QM methods 45

2.1.1 Wave function methods 46

2.1.2 DFT 52

2.1.3 Thermal corrections 53

2.2 ML 58

2.2.1 RF 58

2.2.2 BERT 61
2.3 Interpreting machine-learned predictions 74
2.3.1 Kuz'min prediction contributions 74
2.3.2 IGs 75

3 Interpretable predictions of S_N2 kinetics using the BERT and RF architectures: Comparison to known reactivity rules 77
3.1 Methodology 79
3.1.1 Data 79
3.1.2 ML and QM 83
3.1.3 Calculation times 87
3.1.4 Evaluating model performance 88
3.1.5 Interpretation 91
3.1.6 Dummy model 95
3.2 Results and Discussion 97
3.2.1 RMSE convergence 97
3.2.2 Improving generalisability of the RF 98
3.2.3 Comparison of RF, BERT, and QM 100
3.2.4 Interpreting accurate predictions 103
3.2.5 Interpreting inaccurate predictions 124
3.2.6 Exploring potential limitations 134
3.3 Conclusion 141

4 Interpretable predictions of E2/S_N2 kinetics using the BERT architecture: Analysing drivers of mechanistic competition 143
4.1 Methodology 145
4.1.1 Data 145
4.1.2 ML 147
4.1.3 Interpretation 148
4.1.4 Comparison to Green et al. 150
4.1.5 Literature datasets 152
4.2 Results and Discussion 154
4.2.1 Performance of the E2/S_N2 BERT 154
4.2.2 Interpreting accurate and inaccurate predictions 156
4.2.3 Literature datasets 168
4.3 Conclusion 178

5 Conclusion **180**
5.1 Thesis overview . . . 180
5.2 Implications and future avenues . . . 181
5.2.1 Addressing limitations of the BERT model . . . 182
5.2.2 Increasing generalisability of the BERT model . . . 183
5.2.3 Increasing specificity of the BERT model . . . 184
5.3 Outlook . . . 186

Bibliography **188**

Appendices

A Equations ii
A.1 Softmax . . . ii
A.2 Tanh activation function . . . ii
A.3 Layer normalisation . . . ii

B Derivations iii
B.1 Deriving δ_j for backpropagation . . . iii

C Justifications v
C.1 Justification 1: Evaluating halide nucleophiles . . . v
C.2 Justification 2: Solvent effects in the RF . . . v

D Chapter 1 supplementary material vii
D.1 Train/validation/test splits . . . vii

E Chapter 3 supplementary material ix
E.1 QM methodology . . . ix
E.2 Nucleophilic effects . . . xiv
E.3 Solvent effects . . . xiv
E.4 Product atom mappings . . . xiv
E.4.1 LG effects . . . xiv
E.4.2 Nucleophilic effects . . . xvii
E.4.3 Steric effects . . . xviii
E.4.4 Alyllic effects . . . xix

F Chapter 4 supplementary material xxi
F.1 Training distributions xxi
F.2 Competing reactions in the E2/S_N2 data xxi
F.3 Data imbalance xxi
F.4 Fine-tuning BERT xxiii
F.5 RMSE convergence (Green's data) xxiv
F.6 Interpreting inaccurate predictions xxv
F.6.1 LG effects xxv
F.6.2 Alkyl substituent effects xxvi
F.6.3 Allylic effects xxviii
F.7 Product atom mappings xxix
F.7.1 LG effects xxix
F.7.2 Alkyl substituent effects xxx
F.7.3 Alyllic effects xxxi
F.8 Allylic effects (summing over ring tokens) xxxiii
F.9 Temperature effects xxxiii
F.9.1 S_N2 synthetic data points xxxiii
F.9.2 RF xxxiv
F.10 Literature datasets xxxiv
F.10.1 ΔIG values of reactants xxxiv
F.10.2 $\log k$ distribution for tosylate LGs xxxvii

Appendix bibliography xxxviii

List of Abbreviations

ΔE . . . . . Reaction energy.

$\Delta E^{\ddagger}$ . . . . . Electronic activation barrier.

ΔG . . . . . Reaction free energy.

$\Delta G^{\ddagger}$ . . . . . Gibbs free energy of activation.

ΔH . . . . . Reaction enthalpy.

$\varepsilon_{298.15}$ . . . . . Dielectric constant at 298.15 K.

σ_M . . . . . Standard error of the mean.

AdamW . . . . Adaptive Moment Estimation with decoupled weight decay.

ANN . . . . . Artificial Neural Network.

Ar . . . . . Aryl.

avg. . . . . . Average.

BERT . . . . . Bidirectional Encoder Representations from Transformers.

Bu . . . . . Butyl.

CC . . . . . Coupled Cluster.

CGR . . . . . Condensed Graph of the Reaction.

CPU . . . . . Central processing unit.

CRN . . . . . Chemical Reaction Network.

CV . . . . . Cross-validation.

D . . . . . Hidden dropout probability.

DA . . . . . Diels-Alder.

DFT . . . . . Density Functional Theory.

DMF . . . . . Dimethylformamide

D-MPNN . . . . Directed Message Passing Neural Network.

DMSO . . . . . Dimethylsulphoxide.

dTL . . . . . Diagonal transfer learning.

$E_{\mathbf{pot}}$ . . . . . Potential Energy.

E2 . . . . . Bimolecular elimination.

EI . . . . . Expected Improvement.

ENN . . . . . Equivariant Neural Network.

Et . . . . . Ethyl.

eV . . . . . electron Volts.

EVB . . . . . Empirical Valence Bond.

FMOCs . . . . Frontier molecular orbital coefficients.

FNN . . . . . Feedforward Neural Network.

FVs . . . . . Fisher Vectors.

GBT . . . . . Gradient Boosted Trees.

GELU . . . . . Gaussian Error Linear Unit function.

GGA . . . . . Generalised gradient approximation.

GNN . . . . . Graph Neural Network.

GPR . . . . . Gaussian Process Regression/Gaussian Process Regressor.

GPU . . . . . Graphics processing unit.

HF . . . . . Hartree-Fock.

HOMO . . . . . Highest Occupied Molecular Orbital.

hTL . . . . . Horizontal transfer learning.

hrs . . . . . hours.

IGs . . . . . Integrated Gradients.

IGM . . . . . Ideal Gas Model.

i-**Pr** . . . . . Isopropyl.

IQR . . . . . Interquartile range.

ISIDA . . . . . In SIlico Design and data Analysis.

J . . . . . joule.

k . . . . . Rate constant.

K . . . . . Kelvin.

Kcal . . . . . kilocalorie.

KRR . . . . . Kernel Ridge Regression/Kernel Ridge Regressor.

LDA . . . . . Local density approximation.

LG . . . . . Leaving group.

$\log A$ . . . . . log of the Arrhenius pre-exponential factor.

$\log k$ . . . . . log of the rate constant.

$\log k_{\mathbf{exp}}$ . . . . . Experimental log of the rate constant.

$\log k_{\mathbf{pred}}$ . . . . Predicted log of the rate constant.

LOT . . . . . Level of theory.

LR . . . . . Initial learning rate.

LRP . . . . . Layerwise Relevance Propagation.

LUMO . . . . . Lowest Unoccupied Molecular Orbital.

MA . . . . . Michael Acceptor.

MAE . . . . . Mean Absolute Error.

MC . . . . . Monte Carlo.

MDNN . . . . . Multitask Deep Neural Network.

Me . . . . . Methyl.

mCPBA . . . . *meta*-Chloroperoxybenzoic acid.

ML . . . . . Machine Learning.

MLML . . . . . Masked Language Modelling Loss.

ml-QM-GNN . . Quantum Mechanics augmented Graph Neural Network.

mol . . . . . Mole.

MSE . . . . . Mean Squared Error.

MT . . . . . Molecular Transformer.

MVN . . . . . Multivariate normal.

MW . . . . . Molecular weight.

***n*-Bu** . . . . . Normal butyl.

NLP . . . . . Natural Language Processing.

NN . . . . . Neural Network.

Nu . . . . . Nucleophile.

OHE . . . . . One-hot encodings.

OTs . . . . . Tosylate.

p_P . . . . . Pearson's p value.

p_S . . . . . Spearman's p value.

PCA . . . . . Principal Component Analysis.

PES . . . . . Potential Energy Surface.

Ph . . . . . Phenyl.

Pr . . . . . Propyl.

QM . . . . . Quantum Mechanics.

qNEI . . . . . batched Noisy Expected Improvement.

r_P . . . . . Pearson's r value.

r_S . . . . . Spearman's r value.

RF . . . . . Random Forest.

RMSE . . . . . Root Mean Squared Error.

RNN . . . . . Recurrent Neural Network.

RR . . . . . Ridge Regression/Ridge Regressor.

s . . . . . seconds.

SGD . . . . . Stochastic Gradient Descent.

SHAP . . . . . SHapley Additive exPlanations.

SiRMS . . . . . Simplex Representation of the Molecular Structure.

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

S_N1 . . . . . Unimolecular nucleophilic substitution.

S_N2 . . . . . Bimolecular nucleophilic substitution.

S_NAr . . . . . Aromatic nucleophilic substitution.

SQM/ML . . . . Semiempirical Quantum Mechanics ML.

S_T . . . . . Tanimoto similarity.

STO . . . . . Slater-type orbital.

SVM . . . . . Support Vector Machine.

SVR . . . . . Support Vector Regression/Support Vector Regressor.

TBAF . . . . . Tetra-*n*-butylammonium fluoride.

***t*-Bu** . . . . . Tertiary butyl.

tert . . . . . Tertiary.

TL . . . . . Transfer learning.

TMAP . . . . . Tree map.

TS . . . . . Transition state.

t-SNE . . . . . t-distributed Stochastic Neighbor Embedding.

TST . . . . . Transition state theory.

Glossary

Note that the terms in this glossary are ordered conceptually (rather than alphabetically) to facilitate understanding.

Machine Learning (ML). A field of Artificial Intelligence that involves training a statistical algorithm using observed data to enable it to generalise to new, unseen data.

Features. The set of properties used to describe data for input into an ML model.

Descriptors. The representations used to describe chemical reactions for input into an ML model.

Target value. The property to be predicted by an ML model.

Cross-validation (CV). Evaluation of an ML model using n different train-test splits of the data (where n is any integer). The results are averaged over the n folds, providing a more robust estimate of the model's generalisation to unseen data.

1 | Introduction

1.1 Transition state theory

Reaction rate prediction[1–3] allows chemists to quantitatively elucidate reaction mechanisms, enabling the optimisation of pathways towards desired target compounds.[4–6] According to transition state theory (TST), the experimental rate constant (k) can be deduced from the Gibbs free energy of activation ($\Delta G^{\ddagger}$) using the Eyring equation,[7]

$$k = \frac{k_B T}{h} e^{-\frac{\Delta G^{\ddagger}}{RT}} \tag{1.1}$$

where k_B, h, R and T are the Boltzmann constant, Planck's constant, ideal gas constant and temperature, respectively. TST assumes the formation of an activated complex that exists is in quasi-equilibrium with the reactants. The reaction rate is then defined as the rate of decomposition of this complex to form products. While TST is an approximation, it is widely regarded as a reliable model for estimating rates of elementary reactions, in the absence of high temperatures and quantum tunnelling effects.[8,9]

1.1.1 Computational approaches for kinetic prediction

The $\Delta G^{\ddagger}$ term in Eq. 1.1 can be estimated computationally by employing Quantum Mechanics (QM) methods, most commonly Density Functional Theory (DFT) for its favourable cost-accuracy balance.[10–16] However, DFT approaches often fail to

provide the required chemical accuracy of 1 kcal mol^{-1}, a standard benchmark in computational chemistry due to its alignment with typical experimental errors in thermochemical data and its approximate correspondence to a change in one order of magnitude of k. This failure has been associated with the use of low-level electronic structure methods,[17,18] inaccurate description of entropic contributions,[17–19] and poor description of solvent effects with implicit solvent models.[17,18,20] The scaling of QM methods also renders them impractical for larger systems. On the other hand, reactive Force Fields, such as ReaxFF[21] and the Empirical Valence Bond (EVB)[22] method provide an efficient alternative, however their parametrisation remains time-consuming. Machine Learning (ML) offers a promising alternative to overcome the limitations of the aforementioned methods, either through the use of Machine-learned potentials that model solvents explicitly,[23,24] or ML models that directly predict experimental $\log k$ values[25–35] or QM-generated activation barriers.[34,36–49]

1.1.2 Kinetic ML models trained on QM-generated data

Due to the limited availability of experimental kinetic data, kinetic ML models are often trained using QM-generated activation barriers. Commonly employed QM datasets include the diverse set of gas phase reactions involving C, H, N, and O compiled by Green et al.,[14] and the set of competing E2/S_N2 reactions compiled by von Lilienfeld et al.[13] Also relevant to this discussion are semiempirical Quantum Mechanics ML (SQM/ML) methods, where the SQM-calculated activation barrier is included as feature of the ML model. Each of these contributions are summarised in Figure 1.1**a** and their specifics are detailed in the following sections.

1.1.2.1 Diverse set of gas phase reactions involving C, H, N, and O

Green et al. (2020). One commonly utilised QM dataset is the diverse set of gas phase reactions containing up to 7 heavy atoms involving C, H, N, and O, compiled

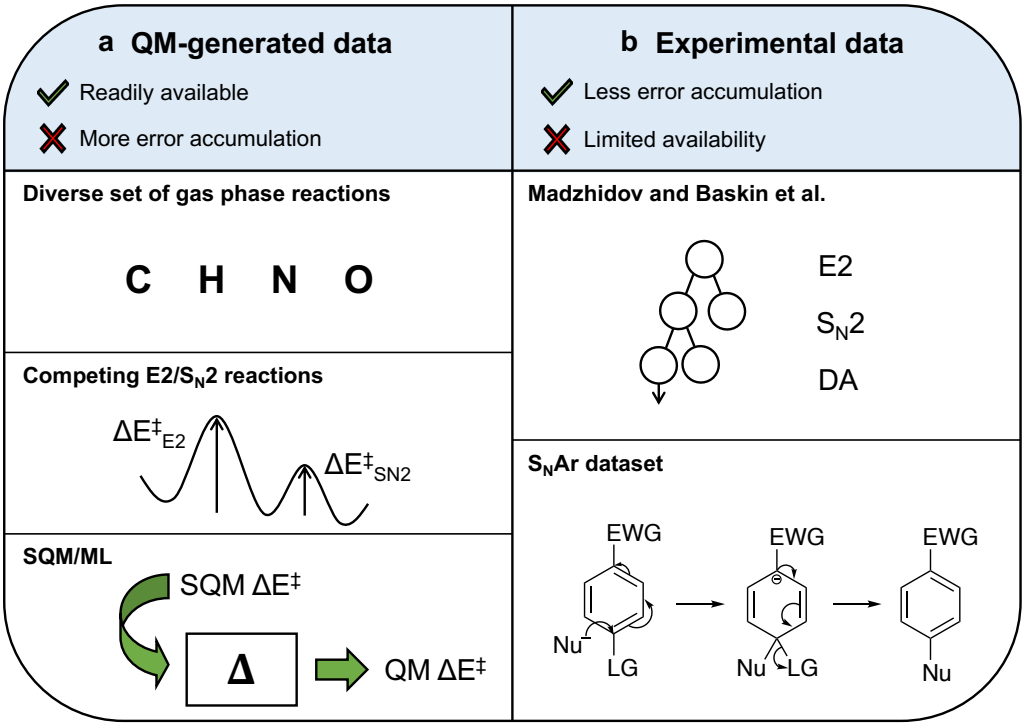


Figure 1.1: Literature overview. a. Due to the limited availability of experimental data, kinetic ML models are often trained using QM-generated activation barriers. Commonly employed QM datasets include the diverse set of gas phase reactions involving C, H, N, and O compiled by Green et al.,[14] and the set of competing E2/S_N2 reactions compiled by von Lilienfeld et al.[13] Recent studies have proposed the use of semiempirical Quantum Mechanics ML (SQM/ML) methods, where the SQM-calculated barrier is included as a model feature. In this approach, the ML is learning the difference (Δ) between the QM-calculated and SQM-calculated values. **b.** To overcome the error accumulation associated with QM-generated data, kinetic ML models can be trained directly on experimental data. Notable advancements in this field include contributions from Madzhidov and Baskin et al.,[25,33] who developed Random Forest (RF) models to predict rates of E2, S_N2, and Diels-Alder (DA) reactions, as well as models trained using the S_NAr dataset compiled by Buttar et al.[32]

by Green et al.[14] This dataset contains 33,000 reactions (16,500 forward and reverse transformations) with electronic activation barriers ($\Delta E^{\ddagger}$) calculated at the B97-D3/def2-mSVP level of theory (LOT), and 24,000 reactions (12,000 forward and reverse transformations) with $\Delta E^{\ddagger}$ calculated at the ωB97X-D3/def2-TZVP LOT. The LOT specifies the QM method used for the calculations, with a higher LOT referring to more accurate, computationally intensive methods, while a lower LOT indicates less accurate, faster approaches. Specifically, B97-D3 and ωB97X-D3 are the DFT functionals employed, and def2-mSVP and def2-TZVP are the basis sets used to describe the electronic wavefunction, with B97-D3/def2-mSVP being a lower LOT than ωB97X-D3/def2-TZVP.

Using their dataset, Green et al. (2020)[36] trained a Directed Message Passing Neural Network (D-MPNN), where reactions are described by molecular graphs, to predict $\Delta E^{\ddagger}$ values at the ωB97X-D3/def2-TZVP LOT. Their approach involved transfer learning (TL), where the model was first pretrained on the larger low LOT dataset and subsequently fine-tuned on the smaller high LOT dataset. This method leverages the pretraining knowledge to enhance the model's learning of the smaller, more accurate dataset. The resulting model achieved a Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), of 3.4 ± 0.3 kcal mol^{-1}, and 1.7 ± 0.1 kcal mol^{-1}, respectively, when evaluated using test data.[a] However, this does not meet the desired chemical accuracy of 1 kcal mol^{-1}. The authors also noted that the true error (with respect to the experimental values) is confounded by the RMSE of the underlying ωB97X-D3/def2-TZVP method, which they calculated to be 2.3 kcal mol^{-1}. Nonetheless, this early work demonstrates the potential of using QM-generated datasets to train kinetic ML models, especially for diverse sets of chemical reactions, where experimental data may be unavailable.

[a] Train/validation/test splits for each dataset are provided in Table D.1 (Appendix D).

Habershon et al. (2022). Green's B97-D3/def2-mSVP dataset was later employed by Habershon et al.[41] in their 2022 study into how errors in ML-predicted $\Delta E^{\ddagger}$ values affect the prediction of macroscopic observables. In their study, Habershon et al. trained an Artificial Neural Network (ANN) using Green's data, utilising molecular fingerprints and QM-calculated reaction energies (ΔE) as descriptors. Their model achieved an RMSE (MAE) of 3.8 (2.8) kcal mol^{-1} based on validation data, showing comparable performance to previously reported models. To assess how these errors impact the prediction of macroscopic observables, the authors performed microkinetics simulations of three Chemical Reaction Networks (CRNs) using their model's predictions, namely, formamide decomposition, aldol reactions, and the decomposition of 3-hydroperoxypropanal. The microkinetics simulations gave mixed results. For the formamide decomposition CRN, the ANN successfully reproduced the same CRN generated using QM-calculated $\Delta E^{\ddagger}$ values. However, there were significant discrepancies for the other two CRNs. For the aldol CRN, the ANN predicted a different ordering of activation energies compared to QM, such that different reactions became more favourable. Moreover, ML-driven simulations of the decomposition of 3-hydroperoxypropanal CRN produced kinetic timescales that were orders of magnitude slower than those from the QM-based simulations. These findings highlight that while ML models can provide reasonably accurate $\Delta E^{\ddagger}$ predictions, errors in these predictions can propagate and significantly impact the accuracy of predicted macroscopic observables.

Green et al. (2022.1). In the same year (2022), Green et al.[34] adapted their 2020 model to describe reactions using a Condensed Graph of the Reaction (CGR). Unlike molecular graphs, which represent single molecules, a CGR combines all species involved in a chemical transformation (reactants and products) into a single graph, providing a more compact representation (see Figure 2.3**a**). This

updated model resulted in an RMSE (MAE) of 6.9 (4.3) ± 0.4 (0.2) kcal mol^{-1}, when evaluated using test data. While this appears less accurate than their 2020 model, the authors attribute the discrepancy to the more rigorous train-test splitting scheme employed in the updated model. This revised approach, which was also employed by Habershon et al., prevents reverse transformations of reactions in the training set from appearing in the test set, providing a more realistic assessment of model performance. Notably, when Green's original model was retrained using this new train-test splitting scheme, it exhibited a larger RMSE (MAE) of 8.0 (5.3) ± 0.4 (0.2) kcal mol^{-1}. These results confirmed that Green's updated model is indeed more accurate than the 2020 version, suggesting CGRs are a more effective descriptor than molecular graphs for the D-MPNN model. To further validate this approach, Green et al. applied their CGR D-MPNN to datasets of competing E2/S_N2 reactions, and S_NAr reactions, with additional details provided later in this chapter.

Green et al. (2022.2). Building on their earlier 2022 improvements, later that year, Green et al.[42] made further refinements to enhance their model's accuracy. They implemented three strategies: 1) fine-tuning the model using higher-accuracy CCSD(T)-F12a/cc-pVDZ-F12 barriers, derived from their original ωB97X-D3/def2-TZVP dataset, 2) incorporating additional descriptors such as ring size and reaction enthalpy (ΔH), and 3) employing a 3D description of reactants and products by means of an equivariant D-MPNN. Fine-tuning the model on the higher-accuracy data lowered the RMSE (MAE) from 6.1 (3.6) kcal mol^{-1} to 5.9 (3.4) kcal mol^{-1}. This evaluation was performed using a single cross-validation (CV) fold, applying the same rigorous train-test splitting criteria as in Green's earlier 2022 model. Further improvements were made by including the ring size and ΔH descriptors, resulting in RMSE (MAE) reductions of 0.2 (0.2) kcal mol^{-1}, and 0.7 (0.2) kcal mol^{-1}, respectively. However, other tested descriptors,

including RDKit molecular feature vectors and QM-generated descriptors, did not enhance the performance of the model.

Interestingly, the 2D CGR D-MPNN outperformed the 3D equivariant D-MPNN, with the latter scoring an RMSE (MAE) of 10.1 (6.2) kcal mol^{-1}. This comparison was based on models trained using the CCSD(T)-F12a/cc-pVDZ-F12 data, without descriptors of ring size or ΔH. The authors noted that the 3D model was trained directly on the higher-LOT data, as using the TL approach employed in the 2D model actually reduced accuracy further. Additionally, the 3D model was trained using only unimolecular reactions ($\approx$ 70% of the dataset) to avoid complications from aligning molecules in 3D space. However, subsequent publications[48] suggest the superior performance of the 2D model is likely due to the inclusion of atom mapping information, which was absent from the 3D model, rather than the difference in dataset size (*vide infra*).

Green's final 2D model, which was fine-tuned using the CCSD(T)-F12a/cc-pVDZ-F12 barriers, and included a description of ring size and ΔH, achieved an RMSE (MAE) of 4.9 (2.9) ± 0.2 (0.2) kcal mol^{-1} when evaluated using test data. This improves upon the earlier 2022 model published by Green et al., demonstrating how the use of higher-LOT training data, combined with a description of ΔH (related to $\Delta E^{\ddagger}$ through the Bell(Brönsted)–Evans–Polanyi principle[50–55]) can lead to more accurate predictions of $\Delta E^{\ddagger}$.

Corminboeuf et al. (2023). In 2023, Corminboeuf et al.[48] sought to overcome the limitations of Green's 3D model by developing an Equivariant Neural Network (ENN) that incorporates atom mapping information into the 3D reaction description. Their model was trained using the 12,000 forward transformations from Green's dataset, with $\Delta E^{\ddagger}$ calculated at the CCSD(T)-F12a/cc-pVDZ-F12 LOT. To evaluate the impact of atom mapping on prediction accuracy, Corminboeuf et al. tested their

model under three different regimes: 1) high quality atom mapping derived from transition state (TS) structures and heuristics, 2) machine-learned atom mappings from RXNMapper,[56] and 3) no atom mapping information. With high quality atom mapping, the model achieved an MAE (RMSE not reported) of 5.0 ± 0.2 kcal mol^{-1} when evaluated using test data, compared to 6.3 ± 0.3 kcal mol^{-1} under the no atom mapping regime. While this improvement is significant, the authors note that in realistic scenarios where the reaction mechanism is not known, atom mapping using heuristics is unfeasible, necessitating the use of the RXNMapper approach. Under this regime, the model's MAE was 6.0 ± 0.3 kcal mol^{-1}, which is equivalent, within error, to the performance without atom mapping. Hence, while atom mapping information can improve the accuracy of $\Delta E^{\ddagger}$ predictions by the 3D ENN, its effectiveness is highly dependent on the quality of the atom mapping.

To further explore the role of atom mapping in predictive accuracy, Corminboeuf et al. tested whether removing atom mapping information from the 2D CGR D-MPNN would similarly reduce its performance. They retrained Green's earlier 2022 model using the same data as for their 3D ENN model, and evaluated its performance across the three atom mapping regimes. Without atom mapping, the 2D CGR D-MPNN exhibited an MAE of 11.9 ± 1.0 kcal mol^{-1}, compared to 5.0 ± 0.5 kcal mol^{-1} with high quality atom mapping, and 6.6 ± 0.3 kcal mol^{-1} with RXNMapper atom mapping. These results confirm that the superior performance of the 2D CGR D-MPNN compared to the 3D equivariant D-MPNN, as observed by Green et al., was likely due to the absence of atom mapping information in the latter.

Building on these findings, Corminboeuf et al. extended their evaluation to two additional datasets. Specifically, they trained their 3D ENN using a dataset of 5269 [3+2] cycloaddition reactions,[16] where the target was $\Delta G^{\ddagger}$ (*vide infra*), and a dateset of 753 enantioselective benzaldehyde propargylation reactions,[4] where

the target was $\Delta E^{\ddagger}$. These additional evaluations revealed that atom mapping had little effect on prediction accuracy for these two datasets, which involve a single reaction mechanism. This suggests that while atom mapping is essential for accurate barrier prediction within diverse sets of reaction classes, it is less important for datasets focused on a single reaction class.

1.1.2.2 Competing E2/S_N2 reactions

von Lilienfeld et al. (2021). Another popular QM dataset is the set of 1286 E2 and 2361 S_N2 gas phase reactions compiled by von Lilienfeld et al.,[13] with $\Delta E^{\ddagger}$ values calculated at the DF-LCCSD/cc-pVTZ//MP2/6-311G(d) LOT. von Lilienfeld et al. (2021)[39] utilised 800 E2 and 1600 S_N2 reactions from this dataset to train two separate kernel ridge regressors (KRRs) to predict $\Delta E^{\ddagger}$ for each reaction mechanism, employing one-hot encodings (OHEs, binary vectors representing features as 1s or 0s based on their presence or absence, respectively) of nucleophiles, leaving groups (LGs), and R groups as descriptors. The resulting models achieved MAEs (RMSEs not reported) of 2.5 kcal mol^{-1} for E2 reactions and 2.2 kcal mol^{-1} for S_N2 reactions, evaluated using test data, displaying comparable performance to the top-performing model trained on Green's data.

Coley et al. (2022). Building on von Lilienfeld's success, in 2022, Coley et al.[40] leveraged their datasets of E2 and S_N2 reactions to train Quantum Mechanics-augmented Graph Neural Networks (ml-QM-GNNs) for $\Delta E^{\ddagger}$ prediction. This architecture predicts QM-generated atom and bond descriptors, such as atomic charges and bond lengths, on-the-fly using a D-MPNN encoder. The QM descriptors are then combined with the structural (graph-based) descriptors as input for $\Delta E^{\ddagger}$ prediction. Coley's ml-QM-GNN models achieved MAEs (RMSEs not reported) of 2.7 ± 0.0 kcal mol^{-1} and 2.8 ± 0.0 kcal mol^{-1} for E2 and S_N2 reactions, respectively, based on validation data. Hence, the OHE KRR models

trained by von Lilienfeld et al. demonstrated marginally better accuracy than the ml-QM-GNNs. To further assess the robustness of the ml-QM-GNN architecture, Coley et al. compared their models' performance to the D-MPNN architecture employed by Green et al. (2020). The D-MPNN displayed similar performance to the ml-QM-GNN, resulting in MAEs of 2.8 ± 0.1 kcal mol^{-1} and 2.9 ± 0.0 kcal mol^{-1} for E2 and S_N2 reactions, respectively. Notably, the D-MPNNs trained on the E2 and S_N2 datasets achieved the same accuracy as the best-performing D-MPNN model trained using Green's data.

Since the ml-QM-GNN involves pooling over reacting atoms, Coley et al. were able to combine the data for E2 and S_N2 reactions to train a single model that predicts $\Delta E^{\ddagger}$ for both mechanisms. This was not possible for the OHE KRR, as the input does not encode reaction pathway information. The E2/S_N2 ml-QM-GNN achieved an RMSE (MAE) of 3.9 (2.9) kcal mol^{-1} when evaluated using test data, matching the performance of the ml-QM-GNNs trained on each individual mechanism. The E2/S_N2 ml-QM-GNN was also shown to have superior generalisability over either of the OHE KRRs, as evidenced by its performance (measured by RMSE) on four test sets containing out-of-sample nucleophiles: 5.4 to 8.7 kcal mol^{-1} for the ml-QM-GNN, compared to 13.3 to 18.6 kcal mol^{-1} for the E2 OHE KRR and 8.6 to 20.2 kcal mol^{-1} for the S_N2 OHE KRR. The greater generalisability of the ml-QM-GNN is to be expected given that the QM-generated descriptors (which are continuous values) allow for extrapolation and interpolation to values not seen during training; something which can not be achieved using OHE (which are binary values).

Green et al. (2022.1). Also in 2022, Green et al.[34] trained their CGR D-MPNN (earlier 2022 version) using the combined dataset of E2 and S_N2 reactions compiled by von Lilienfeld et al. This model resulted in an RMSE (MAE) of

3.6 (2.7) ± 0.1 (0.1) kcal mol^{-1} when evaluated using test data, offering a slightly improved accuracy over Coley's model.

Persson et al. (2024). More recently, in 2024, Persson et al.[49] utilised von Lilienfeld's S_N2 dataset (with barriers recalculated at both the B3LYP/def2-SVP and B3LYP/def2-TZVP LOT) to train an ENN for $\Delta E^{\ddagger}$ prediction. The authors employed frontier molecular orbital coefficients (FMOCs) from the reactant and product complexes as features - alongside atomic numbers and positions, and the reaction free energy (ΔG) - motivated by the chemical interpretability and imposed physical constraints they provide. Persson's model achieved an MAE (RMSE not reported) of 0.021 eV (≈ 0.5 kcal mol^{-1}) for both basis sets, when evaluated using test data. Notably, Persson's model outperformed previous models trained on the same dataset, highlighting the effectiveness of their approach. In addition to $\Delta E^{\ddagger}$, Persson's model also predicts TS FMOCs. Employing the TS HOMO as an example, Persson's model predicted the majority of coefficients within 0.2 of the DFT values, demonstrating its capability in both $\Delta E^{\ddagger}$ and FMOC prediction. The authors also tested their model's applicability to diverse datasets by training it on 14,000 of the reactions compiled by Green et al. (corresponding to those with a single bond forming or breaking), recalculated at the B3LYP/def2-SVP LOT. This model resulted in an MAE of 0.42 eV (≈ 9.7 kcal mol^{-1}), highlighting limitations in the FMOC ENN's applicability to datasets of multiple reaction classes.

1.1.2.3 Other notable QM datasets

Other notable QM-generated datasets include the 1947 iridium-catalysed H_2 activations with $\Delta E^{\ddagger}$ values calculated at the PBE-D3/def2-SVP LOT, compiled by Balcells et al.,[37] and the 5269 [3+2] cycloaddition reactions with $\Delta G^{\ddagger}$ values calculated at the SMD(water) B3LYP-D3(BJ)/def2-TZVP//B3LYP-D3(BJ)/def2-SVP LOT, compiled by Coley et al.[16]

The H_2 activation data was employed by Balcells et al. (2020)[37] to train ANN and Gaussian Process Regressor (GPR) models, using autocorrelation and deltametric functions of atomic properties, and molecular fingerprints (GPR only), as features. The resulting models achieved MAEs of 1.1 kcal mol^{-1} and 0.6 kcal mol^{-1} for the ANN, and GPR, respectively, when evaluated using test data.

Regarding the [3+2] cycloaddition dataset, this was employed by Coley et al.[46] in 2023 to train an ml-QM-GNN model for multitask $\Delta G^{\ddagger}$ and ΔG prediction (using the same featurisation as in their 2022 model), which they used to screen a 10 million reaction search-space for bioorthogonal click potential. The resulting model predicted $\Delta G^{\ddagger}$ with an MAE of 2.4 kcal mol^{-1} when evaluated using test data, and returned > 120,000 candidate bioorthogonal click reactions. Note that this [3+2] cycloaddition dataset was also employed by Corminboeuf et al. to train their 2D CGR D-MPNN and 3D ENN models (*vide supra*).

1.1.2.4 SQM/ML

Grayson et al. (2022). Another significant contribution is the work of Grayson et al.,[43–45] who adopt an SQM/ML approach to predict QM-calculated activation barriers. In this approach, the SQM-calculated barrier is included as a model feature, so that the ML is learning the difference between the QM-calculated and SQM-calculated values (a type of Δ-ML). In 2022, Grayson et al.[43] developed an SQM/ML model to predict the $\Delta G^{\ddagger}$ of C–C bond forming nitro-Michael additions, using a dataset of 1000 reactions with barriers calculated at the IEF-PCM(toluene) ωB97X-D/def2-TZVP LOT. Their model, which utilises a GPR algorithm, incorporates the IEF-PCM(toluene) AM1 $\Delta G^{\ddagger}$ into the feature set, alongside AM1-calculated thermochemical and electronic properties of the Michael acceptor (MA) and TS, such as enthalpy, and HOMO and LUMO energies. This model achieved an MAE (RMSE not reported) of 0.9 ± 0.2 kcal mol^{-1} when

evaluated on an external test set of 35 literature reactions, surpassing the desired chemical accuracy of 1 kcal mol^{-1}.

Grayson et al. (2023.1). In 2023, Grayson et al.[44] expanded their SQM/ML approach by incorporating a TL scheme to predict the $\Delta G^{\ddagger}$ of pericyclic reactions. Employing a Neural Network (NN) architecture and a feature set of SQM-calculated electronic and thermochemical properties (including the AM1 $\Delta G^{\ddagger}$), the authors explored two TL methods: 1) horizontal TL (hTL), which transfers knowledge from one chemical domain to another using the same LOT, and 2) diagonal TL (dTL), which transfers knowledge from one chemical domain to another, going from a lower LOT to a higher LOT. To demonstrate each of these approaches, the authors investigated four regimes: 1) hTL from a dataset of 792 intermolecular DA reactions (dataset A) to a dataset of 273 intramolecular DA reactions[b] (dataset α), 2) hTL from a combined dataset of A + α (dataset B) to a dataset of 103 1,2,4,5-tetrazine DA reactions (dataset β), 3) hTL from dataset B to a sample of 408 reactions from the [3+2] cycloaddition dataset compiled by Coley et al. (dataset γ), and 4) dTL from dataset B,[c] calculated at the ωB97X-D/def2-TZVP LOT, to dataset β, calculated at the higher DSD-PBEP86-D3(BJ)/def2-TZVP LOT. All datasets used in the hTL regimes were (re)calculated at the ωB97X-D/def2-TZVP LOT. Regarding the three hTL regimes, each model achieved an MAE < 1 kcal mol^{-1} when evaluated using test data. Not only is this within chemical accuracy, but it significantly improves on the MAE of the pretrained models (ie., prior to fine-tuning on α, β, and γ), which was 5.1, 1.9, and 8.3 kcal mol^{-1} when evaluated using α, β, and γ test data, respectively. This improvement demonstrates the effectiveness of hTL in enhancing generalisability across different chemical domains. Similarly, dTL from dataset B to dataset β

[b] Training sizes = 174, 66, and 326 for α, β, and γ, respectively.

[c] dTL results are reported for exo reactions while hTL results are reported for endo reactions (both from chemical domain B). The number of exo reactions in domain B is 1109.

reduced the MAE from 10.2 kcal mol^{-1} (for the pretrained model) to 1.2 kcal mol^{-1}, indicating dTL's effectiveness in transferring knowledge across different chemical domains at varying LOTs.

Grayson et al. (2023.2). Also in 2023, Grayson et al.[45] introduced a reformulated approach to barrier prediction, aiming to identify reactions with barriers that fall within the range of a specified target value. Their algorithm trains an (S)QM/ML model on five randomly selected reactions, which is then used to predict barriers for the remainder of the dataset. The reaction whose predicted barrier is closest to the desired target has its real barrier computed using the ground truth LOT; if the predicted barrier is within 1 kcal mol^{-1} of the ground truth, the algorithm is complete. Else, this reaction is added to the training set and the process repeated until a reaction with the desired barrier is found. The authors tested their algorithm using three datasets: 1) a dataset of 1000 aza-Michael additions with $\Delta G^{\ddagger}$ values calculated at the IEF-PCM(water) ωB97X-D/def2-TZVP LOT, 2) Balcells's H_2 activation dataset, and 3) von Lilienfeld's E2 and S_N2 datasets, with $\Delta E^{\ddagger}$ calculated at the MP2/6-311G(d) LOT. In each case, they trained a Ridge Regressor (RR) model using an OHE of interchangeable functional groups, and a lower LOT barrier as features. The lower LOT barrier was calculated using IEF-PCM(water) PM6 for the Michael addition dataset, SVWN/def2-SVP for the H_2 activation dataset, and HF/6-311G(d) for the E2 and S_N2 datasets. For both the Michael addition and H_2 activation datasets, a barrier within 1 kcal mol^{-1} of the target was identified using < 20 training reactions, demonstrating the data efficiency of Grayson's reformulated approach. Notably, the number of training reactions was higher for the E2 and S_N2 datasets (100 for E2 and 48 for S_N2), however, the authors attribute this to the large number of missing barriers from the original data. Importantly, Grayson's reformulated approach

identified E2 and S_N2 barriers within 1 kcal mol^{-1} of the target using significantly fewer training reactions than previous models.

Núñez et al. (2023). Following Grayson et al.'s success with the SQM/ML approach, in 2023, Núñez et al.[47] developed three SQM/ML models for $\Delta E^{\ddagger}$ prediction, utilising the 12,000 (8355 after screening) forward transformations from the ωB97X-D3/def2-TZVP dataset compiled by Green et al. They employed three different architectures, namely, a Multitask Deep Neural Network (MDNN) which also predicts ΔH, Gradient Boosted Trees (GBT), and GPR. The feature set included the PM7 $\Delta E^{\ddagger}$ value, other PM7-calculated descriptors such as the TS imaginary frequency, RDKit cheminformatics descriptors, topological descriptors of the TS, and the number of bonds changed during the reaction. The resulting models achieved MAEs (RMSEs not reported) of 3.7, 3.4, and 3.6 kcal mol^{-1} for the MDNN, GBT, and GPR, respectively, when evaluated using test data. Hence, the predictive performance is comparable between the three architectures. Moreover, each model outperformed previous models trained using Green's ωB97X-D3/def2-TZVP dataset, demonstrating the effectiveness of the SQM/ML approach. It is noted that while the the MDNN predicts both $\Delta E^{\ddagger}$ and ΔH, only $\Delta E^{\ddagger}$ prediction is discussed here, as ΔH is included primarily to enhance the model's accuracy with respect to the $\Delta E^{\ddagger}$ target.

To evaluate the effect of temperature on their models' predictions, Núñez et al. retrained their GBT and GPR models to predict $\Delta G^{\ddagger}$ (calculated from the $\Delta E^{\ddagger}$ values) at 300, 500, and 1000 K. Here, the PM7 $\Delta G^{\ddagger}$ was used as a model feature, instead of $\Delta E^{\ddagger}$. At the two lowest temperatures, the accuracy of $\Delta G^{\ddagger}$ prediction was similar to that for $\Delta E^{\ddagger}$ prediction, with MAEs of 3.7 kcal mol^{-1} (300 K) and 3.9 kcal mol^{-1} (500 K) for GBT, and 4.0 kcal mol^{-1} (300 K) and 4.1 kcal mol^{-1} (500 K) for GPR. However, at 1000 K, these errors increased to 5.4 kcal mol^{-1}

and 5.3 kcal mol^{-1} for the GBT and GPR, respectively. The authors suggest that $\Delta G^{\ddagger}$ prediction at higher temperatures may be improved by employing a multitask ML model to correct the SQM vibrational frequencies.

1.1.2.5 Summary

Analysis of the literature showed how QM-generated datasets have been utilised to train kinetic ML models, particularly for diverse sets of chemical reactions, where experimental data may be unavailable. Most notably, the diverse set of gas phase reactions involving C, H, N, and O, compiled by Green et al. has been employed by a number of groups to train ML models for $\Delta E^{\ddagger}$ prediction. While none of these models have achieved the desired chemical accuracy of 1 kcal mol^{-1}, prediction accuracy has been shown to improve with: inclusion of atom mapping information (Corminboeuf et al., 2023), inclusion of the SQM-calculated $\Delta E^{\ddagger}$ in the feature set (Núñez et al., 2023), and $\Delta E^{\ddagger}$ target values calculated at a higher LOT (Green et al., 2022.2). The latter highlights how uncertainty in the underlying QM method can decrease the performance of the trained ML model. Similarly, it was shown how error accumulation can occur between the underlying QM method and the ML model (Green et al., 2020), and the ML-predicted $\Delta E^{\ddagger}$ values and macroscopic observables (Habershon et al., 2022).

Notably, models trained on the set of competing E2/S_N2 reactions compiled by von Lilienfeld et al. achieved similar prediction accuracies to the best-performing model trained using Green's data (Green et al., 2022.2). Analysis of models trained on von Lilienfeld's data also revealed the importance of including reaction pathway information into the reaction description to predict $\Delta E^{\ddagger}$ across multiple mechanisms (Coley et al., 2022). Additionally, it was shown that continuous-valued descriptors (such as QM-calculated physical properties) result in a more

generalisable ML model than descriptors based on discrete values (such as OHE, Coley et al., 2022).

Overall, the most accurate ML models discussed in this section either adopted an SQM approach (Grayson et al., 2022-2023; Núñez et al., 2023), or utilised FMOCs as features (Persson et al., 2024). Although, the latter was limited by its generalisability to diverse datasets. In particular, the reformulated approach to barrier prediction developed by Grayson et al., which aims to identify reactions with barriers within 1 kcal mol^{-1} of a desired target, seems the most promising in terms of accuracy and data efficiency.

Regarding limitations, ML models trained using QM-generated data are susceptible to error accumulation from the underlying QM method (Green et al., 2020), which can also filter through to the prediction of macroscopic observables (Habershon et al., 2022). Furthermore, the datasets discussed in this section do not contain reactivity information from a variety of solvents and temperatures. To overcome these limitations, kinetic ML models can be trained directly on experimental data, where the target value is $\log k$.

1.1.3 Kinetic ML models trained on experimental data

While QM-generated kinetic data can help address limitations in data availability, the error accumulation from both the ML model and underlying QM method remains a challenge. Furthermore, the QM-generated datasets discussed in Section 1.1.2 do not consider the effect of varying temperature and solvent. To overcome these limitations, kinetic ML models can be trained directly on experimental data, where the target value is $\log k$. Notable advancements in this field include contributions from Madzhidov and Baskin et al.,[25–29,33,35] as well as models trained on the S_NAr dataset compiled by Buttar et al.[32] These contributions are summarised in Figure 1.1**b**, with detailed descriptions provided below.

1.1.3.1 Madzhidov and Baskin et al.

Baskin et al., S_N2 (2011). A significant contribution to this field is the work by Madzhidov and Baskin et al. In 2011, Baskin et al.[25] developed an ANN trained on a dataset of 1924 S_N2 reactions in single solvents, and their experimental $\log k$ values. The reactions were described using structural fragments, and QM-calculated descriptors such as Fukui indices, as well as temperature, and solvent parameters such as the dielectric constant. The resulting model achieved an RMSE (MAE not reported) of 0.6 $\log k$ when evaluated using test data, which is within the desired chemical accuracy of 1 $\log k$ unit. Hence, Baskin's model achieved comparable accuracy to the best-performing S_N2 model trained on QM-generated data (Persson et al., 2024). However, since Baskin's model is trained using experimental data, it is not susceptible to the error accumulation associated with QM-based models.

Madzhidov and Baskin et al., S_N2 (2014). Building on this earlier success, in 2014, Madzhidov and Baskin et al.[26] refined their approach by training a Random Forest (RF) for $\log k$ prediction, utilising a dataset of 1041 S_N2 reactions involving neutral nucleophiles, which were described using In SIlico Design and data Analysis (ISIDA) fragments,[57] in addition to temperature, and solvent parameters. This revised model achieved an RMSE of 0.5 $\log k$ based on validation data, which is comparable to the performance of Baskin's 2011 model. However, the recent model offers the advantage of faster descriptor calculation, as it does not rely on QM-generated descriptors.

Madzhidov et al., S_N2 (2019). Building on their 2014 work, in 2019, Madzhidov et al.[29] trained a Support Vector Regressor (SVR) for $\log k$ prediction utilising a dataset of 4830 S_N2 reactions, including those involving anionic nucleophiles and

solvent mixtures. Consistent with their 2014 work, reactions in Madzhidov's 2019 model were described using ISIDA fragments, alongside reciprocal temperature, and solvent parameters. This newer model resulted in an RMSE of 0.3 log k when tested using validation data, further improving the accuracy of Madzhidov and Baskin's earlier models. That said, when evaluated on an external test set of 105 Menshutkin reactions, the newer model scored an RMSE of 0.8 log k. While this is marginally higher than the model's performance on the validation data, it is still within the desired chemical accuracy of 1 log k unit.

Madzhidov et al., E2 and DA (2015-2017). In addition to the S_N2 models described above, Madzhidov et al. have developed rate prediction models for other single-step organic reactions, including E2[27] and Diels-Alder (DA).[28] In 2015, they trained an SVR for log k prediction using a dataset of 1043 E2 reactions. The reactions were described in the same way as in their S_N2 models, using ISIDA fragments, reciprocal temperature, and solvent parameters. Madzhidov's E2 model resulted in an RMSE of 0.7 log k based on validation data, consistent with the chemical accuracy of their S_N2 models. In 2017, Madzhidov et al. also applied their ISIDA SVR approach to DA rate prediction, employing a dataset of 880 DA reactions and their experimental log k values. The DA model scored an RMSE of 0.8 log k based on validation data, maintaining the chemical accuracy of the E2 and S_N2 models.

Madzhidov et al., S_N2, E2 and DA (2021). In 2021, Madzhidov et al.[33] trained three separate RF models for log k prediction using their S_N2 (2019), E2, and DA datasets. The authors evaluated their models using two distinct CV schemes: 1) solvent-out CV, where reactions in the same solvent are placed in the same CV fold, and 2) transformation-out CV, where reactions of the same transformation are placed in the same CV fold. These schemes tested the models' abilities to

generalise to new solvents, and transformations, respectively, with regular CV serving as a control. For all three mechanisms, prediction accuracy decreased under solvent-out and transformation-out CV compared to regular CV. However, this is to be expected as the models were required to generalise to unseen data in the former two regimes. Notably, the E2 and S_N2 models performed better under the transformation-out scheme than the solvent-out scheme, while the DA model favoured the solvent-out scheme. This discrepancy was attributed to the non-polar DA TS, which is less susceptible to solvent effects than the E2 or S_N2 TS. Importantly, all three models achieved near chemical accuracy (largest RMSE = 1.3 $\log k$) across all CV schemes, suggesting good generalisability to unseen data. Additionally, the S_N2 model was evaluated under the transformation-out scheme using an external test set of 90 Menshutkin reactions, corresponding to transformations not seen during training. The resulting RMSE of 1.0 $\log k$ is comparable to that using validation data (0.8 $\log k$), affirming that transformation-out CV provides a reliable estimate for model performance on unseen data. The transformation-out CV versions of Madzhidov's E2, S_N2, and DA models remain the most recent publicly available implementations.

Madzhidov et al., conjugated models (2023). More recently, in 2023, Madzhidov et al.[35] developed *conjugated* rate prediction models that integrate the Arrhenius equation directly into the ML architecture. Two conjugated models were developed, corresponding to RR and NN architectures. Each model was trained using a multi-task modelling approach to predict $\log k$, $\log A$ (where A is the Arrhenius pre-exponential factor), and $\Delta E^{\ddagger}$, where $\log k$ is calculated from $\log A$ and $\Delta E^{\ddagger}$ by applying the Arrhenius equation. The training dataset consisted of 1849 cycloaddition reactions, described by ISIDA fragments and solvent parameters, with temperature introduced via the Arrhenius equation. Both models accurately predicted $\log k$ and $\Delta E^{\ddagger}$, achieving R^2 values (RMSE

and MAE not reported) of 0.75 ($\log k$) and 0.90 ($\Delta E^{\ddagger}$) for the RR, and 0.71 ($\log k$) and 0.84 ($\Delta E^{\ddagger}$) for the NN, when evaluated using external test data. Meanwhile, $\log A$ was predicted less accurately, resulting in R^2 values of 0.57 and 0.56 for the RR and NN, respectively. These results were consistent with other benchmarked models, including multi-task ($\log k$, $\log A$, *and* $\Delta E^{\ddagger}$) and single-task ($\log k$, $\log A$, *or* $\Delta E^{\ddagger}$) NN and RR models that do not incorporate the Arrhenius equation into their architecture.

1.1.3.2 S_NAr dataset

Buttar et al. (2021). Another significant advancement is the work Buttar et al.,[32] who in 2021 employed a hybrid QM/ML approach to predict $\Delta G^{\ddagger}$ values (calculated from the experimental rate constants) of S_NAr reactions. Similar to the SQM/ML approach adopted by Grayson et al., and Núñez et al., hybrid QM/ML (another type of Δ-ML) predicts the experimentally-derived barrier from the QM-calculated barrier, which is incorporated as a model feature. In the case of Buttar et al., the SMD(solvent) ωB97X-D/6-311+G(d,p) $\Delta G^{\ddagger}$ was added to the feature set, alongside QM-calculated physical properties such as global and local nucleophilicities and electrophilicities. The authors employed GPR as the ML algorithm, which they trained on a dataset of 443 S_NAr reactions. The resulting model achieved an RMSE (MAE) of 1.0 (0.8) kcal mol^{-1} when evaluated using external test data, meeting the desired chemical accuracy of 1 kcal mol^{-1}. Thus, Buttar's model achieved comparable performance to the other models trained on experimental data discussed in this section.

Green et al. (2022.1). Buttar's S_NAr dataset was also employed by Green et al.[34] to train their CGR D-MPNN model (earlier 2022 version) for $\Delta G^{\ddagger}$ prediction. This model achieved an RMSE (MAE) of 1.3 (0.9) ± 0.1 (0.1) kcal mol^{-1} when evaluated using test data, offering comparable accuracy to Buttar's model. Notably,

the CGR D-MPNN trained on Buttar's data significantly outperformed the same model trained using the Green or von Lilienfeld datasets. This may be attributable to the relative simplicity of predicting activation barriers for a single reaction mechanism compared to multiple mechanisms, or the higher quality of Buttar's data compared to the QM-generated datasets, which are subject to uncertainty from the underlying QM method.

1.1.3.3 Summary

Analysis of the literature revealed that kinetic ML models trained using experimental data can achieve chemical accuracy, while also avoiding the error accumulation that can occur in models trained using QM-generated data. In particular, the S_N2, E2 and DA RF models trained by Madzhidov et al. predict $\log k$ within chemical accuracy, while providing a realistic assessment of performance on unseen reactions through the use of transformation-out CV. Another notable example is the S_NAr rate predicting GPR developed by Buttar et al., which achieves chemical accuracy using a hybrid QM/ML approach. However, this method requires calculating DFT $\Delta G^{\ddagger}$ values, making it more computationally expensive than the ISIDA fragment/reaction conditions approach employed by Madzhidov et al. Notably, direct comparison between experimental and QM-generated datasets could be made through the work of Green et al., who trained their CGR D-MPNN model using all three of the Green (QM), von Lilienfeld (QM), and Buttar (experimental) datasets. Tellingly, the CGR D-MPNN performed best when trained using the experimental data. This may have resulted from both Green and von Lilienfeld's datasets containing multiple reaction mechanisms, presenting a more challenging prediction task than the single reaction mechanism in Buttar's dataset. Alternatively, it may reflect the higher quality of Buttar's data compared to the QM-generated datasets, which are subject to uncertainty from the underlying QM method.

Overall, experimental data can be utilised to train ML models for accurate prediction of reaction kinetics. However, models trained using experimental datasets tend to be specific to single reaction mechanisms. To assess the applicability of these models to more diverse datasets, it is important to train ML models using experimental data that encompasses multiple mechanisms.

1.2 BERT

Bidirectional Encoder Representations from Transformers (BERT)[58] is a Natural Language Processing (NLP) model, a type of deep learning model designed to process textual data. BERT relies entirely on a self-attention mechanism, offering a significant advantage over earlier NLP architectures based on Recurrent Neural Networks (RNNs). Unlike RNNs, which process input tokens sequentially, the self-attention mechanism in BERT allow for parallelisation across input tokens, increasing computational efficiency. Furthermore, a BERT model pretrained on a general task can be fine-tuned on several downstream tasks simply by modifying the output layer. The advantage here is that fine-tuning a pretrained model is typically more cost-effective than training from scratch, as the model weights will likely be closer to optimal for the downstream task than randomly initialised weights.

Unlike earlier pretrained transformer architectures, such as the Generative Pretrained Transformer (GPT), which are unidirectional (i.e., tokens can attend only to either preceding or subsequent tokens in the self-attentions layers), BERT is bidirectional, meaning it jointly conditions on both left and right context. This makes BERT suited to a wider range of downstream tasks such as sentence-level (such as natural language inference) and token-level (such as question answering) tasks. Bidirectionality in the BERT model was achieved by pretraining on Masked Language Modelling Loss (MLML), an adaptation of the Cloze task[59] where randomly masked tokens in the input are predicted based on the context.

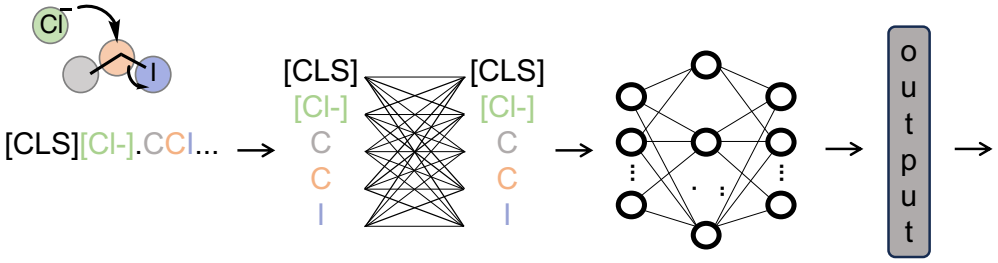


Figure 1.2: BERT overview. The BERT model developed by Laino et al.[60] treats chemistry as a language task, translating SMILES representations of reactions or molecules into target properties using an attention-based neural network (as shown here for a reaction between chloride and ethyl iodide). By modifying the output layer, a BERT model can be fine-tuned for several downstream tasks. Laino et al. successfully applied this approach to both reaction classification,[60] and yield prediction[61] tasks. Here, [CLS] is a special token that is prepended to the SMILES string to learn a global representation of the reaction.

1.2.1 BERT in Chemistry

BERT models have gained popularity in Chemistry, with applications in molecular discovery,[62,63] reaction classification,[60] and yield prediction.[61] In each case, chemistry is treated as a language task, where SMILES representations of reactions or molecules are translated into target properties via an attention-based neural network (Figure 1.2). One notable example is the work of Laino et al.,[60] who pretrained a BERT model for chemical prediction using MLML on chemical reactions. In this approach, BERT was tasked with predicting SMILES tokens that had been masked from the input sequence. The authors then fine-tuned the pretrained model on both a reaction classification[60] and yield prediction[61] task. Trained on a dataset of $\approx$ 2.6 million reactions from the Pistachio database,[64] Laino's BERT model classified reactions with 98% accuracy when evaluated using test data. Similarly, when trained on a dataset of 3955 Buchwald-Hartwig reactions, Laino's BERT model predicted reaction yields with an R^2 (MAE and RMSE not reported) of 0.95 $\pm$ 0.01 on validation data, outperforming previous models trained on this dataset. This predictive capability was further validated using test data, where the BERT model predicted yields with an R^2 of 0.73. Thus, Laino's BERT

model demonstrated strong performance in both reaction classification and yield prediction, indicating its predictive potential across a variety of chemical tasks.

1.2.2 BERT in kinetic prediction

The hidden representations generated by Laino's pretrained BERT model can be extracted and used as reaction fingerprints. These so-called *rxnfps* have been tested by both Buttar et al.[32] and Green et al.[34] as structural features for input into their respective GPR and CGR D-MPNN models. Buttar et al. found that training their GPR using only rxnfps resulted in an RMSE (MAE) of 1.5 (1.0) ± 0.1 (0.1) kcal mol^{-1} on validation data. This result not only approaches chemical accuracy, but also matches the performance of the GPR trained using DFT-calculated descriptors when the DFT $\Delta G^{\ddagger}$ and other TS features are excluded. Moreover, the GPR trained using rxnfps outperformed, or was comparable to, all other structural descriptors tested, including the ISIDA fragment representations employed by Madzhidov et al. Thus, rxnfps proved to be a robust feature set for encoding reaction information.

Buttar's results were corroborated by Green et al., who compared their CGR D-MPNN trained on Buttar's data to a Feedforward Neural Network (FNN, representative of the FNN component of the D-MPNN) trained using rxnfp representations of the same dataset. Both the CGR D-MPNN and rxnfp FNN achieved similar accuracy, with the rxnfp FFN scoring an RMSE (MAE) of 1.4 (1.0) ± 0.2 (0.1) kcal mol^{-1}, compared to 1.3 (0.9) ± 0.1 (0.1) kcal mol^{-1} for the CGR D-MPNN. Furthermore, consistent with Buttar's results, the FNN trained with rxnfps outperformed, or was on par with, all other fingerprints tested, including ISIDA fragments. These results reinforce Buttar et al.'s assessments that rxnfps are highly effective descriptors for reactivity prediction within single reaction mechanisms. This said, when Green et al. made the same comparisons using their ωB97X-D3/def2-TZVP dataset and von Lilienfeld's E2/S_N2 dataset,

both the CGR D-MPNN, and the FNN model trained with ISIDA fragments, outperformed the FNN trained with rxnfps. This suggests that while rxnfps are suitable descriptors of single reaction mechanisms in reactivity prediction, they are less effective for datasets containing multiple mechanisms. This may stem from the fact that rxnfps are extracted from a pretrained BERT model that has not been fine-tuned for reactivity prediction. During fine-tuning, the representations are updated to minimise the loss with respect to the prediction target, thereby improving their relevance to specific tasks. Consequently, fine-tuning BERT specifically for rate prediction could enhance the suitability of the rxnfp representations for this task across multiple reaction mechanisms. To date, BERT has not been trained specifically for rate prediction.

1.3 Interpretable ML

It has become increasingly desirable for ML models to not only be accurate, but also interpretable. Interpretability can help to identify key features,[25,32,44,45,47] rationalise model limitations,[65] and importantly, ensure predictions are grounded in logical reasoning.[65–67] A classic example of the latter is presented in the work of Müller et al.,[66] who uncovered a spurious correlation between a classifier's ability to recognise horse images, and the presence of a source tag in the corner of those images. The classifier, a Support Vector Machine (SVM) with images represented by Fisher Vectors (FVs),[68,69] achieved state-of-the-art performance, correctly identifying horse images with an accuracy of 80% when evaluated using test data. Utilising Layerwise Relevance Propagation (LRP),[70] which allows pixel contributions towards predictions to be visualised as heatmaps, Müller et al. revealed that the classifier was attending to the source tags, rather than the horses, when making its predictions (Figure 1.3). Consequently, removing the source tag also removed the classifiers ability to recognise horse

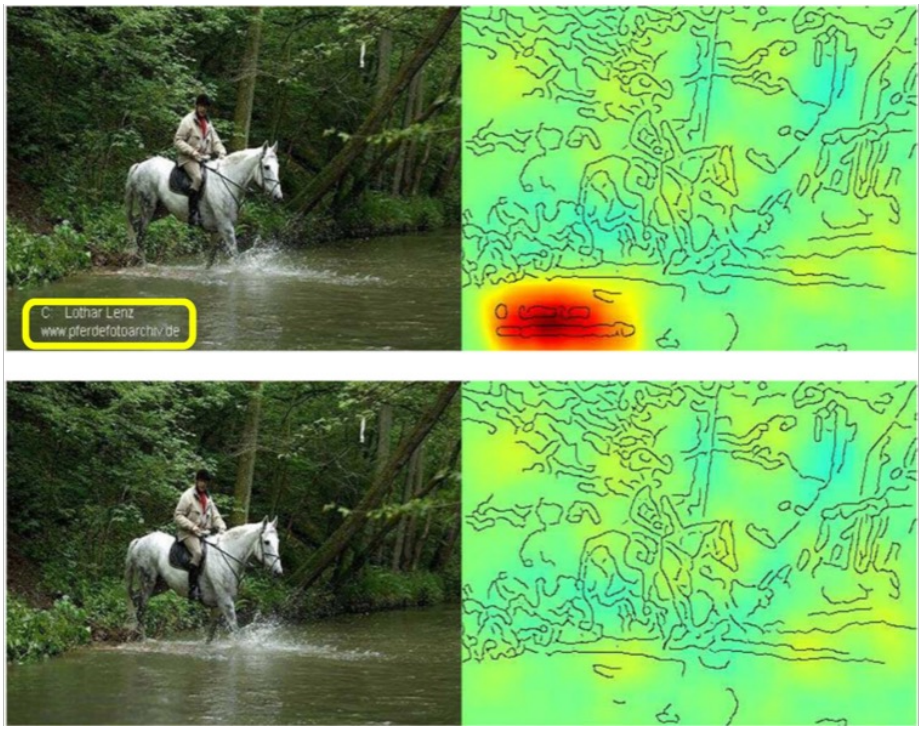


Figure 1.3: The work of Müller et al.[66] Using Layerwise Relevance Propagation,[70] Müller et al.[66] uncovered a spurious correlation between a classifier's ability to recognise horse images, and the presence of a source tag in the corner of those images. The classifier achieved state-of-the-art performance, correctly identifying horse images with an accuracy of 80%. However, Müller et al. revealed that the classifier was attending to the source tags, rather than the horses, when making its predictions. Consequently, removing the source tag also removed the classifiers ability to recognise horse images. Image from Müller et al.[66]

images. These insights provided by Müller et al. underscore the importance of ensuring that predictions are not only accurate but also interpretable. Without interpretability, spurious correlations can lead to overfitted models that are unable to generalise to unseen data.

1.3.1 Interpretable ML in Chemistry

In the context of chemistry, interpretability can help to ensure that machine-learned predictions align with established chemical principles. Two examples relevant to this Thesis are the works of Kuz'min et al.,[67] and Lee et al.[65]

In the work of Kuz'min et al.,[67] the authors developed a new algorithm to interpret predictions made by RF models, using affinity data for 5-HT_{1A} receptors[d] as an example. This algorithm calculates the contribution of each descriptor to the prediction, relative to predicting the mean target value of the training data. The 5-HT_{1A} ligands were described using Simplex Representation of the Molecular Structure (SiRMS),[67,71] which represents molecules as a system of tetratomic fragments. By combining fragment-based descriptors with calculated descriptor contributions, Kuz'min et al. were able to assess the contributions of individual atoms towards 5-HT_{1A} affinity prediction. Using this approach, they found that calculated fragment contributions were consistent with experimental observations. For example, the largest predicted affinity increase among aryl fragments was associated with p-electron donor substituents in the *ortho*-position, consistent with these groups forming hydrogen bonds with 5-HT_{1A} receptors. Conversely, results showed *para*- substituents were predicted to decrease affinity for 5-HT_{1A} receptors, in agreement with them displaying unfavourable steric interactions. Thus, Kuz'min et al. successfully demonstrated how their method for calculating prediction contributions can be used to validate machine-learned predictions against known chemical principles. Kuz'min prediction contributions are compatible with any descriptor set applied to an RF model, and are accessible through the treeinterpreter software.[72]

In contrast, Lee et al.[65] focused on interpreting predictions made by the Molecular Transformer (MT), an NLP model developed in collaboration with Laino et al., and a predecessor to Laino's BERT model. While the BERT model consists of an encoder and output layer, making it applicable to multiple downstream tasks, the MT utilises an encoder-decoder architecture, making it suited to translation tasks, such as predicting products from reactants. In the case of Lee et al., the

[d] A type of serotonin receptor.

MT was tasked with predicting the products of selective chemical reactions, and the model's reasoning for predicting one product over the other was interpreted using two quantification methods. Firstly, the contribution of each SMILES token towards the prediction was calculated using Integrated Gradients (IGs),[73] whose magnitude and sign quantify the token's influence towards a prediction relative to a baseline "." token. Secondly, the most similar training reactions to incorrectly predicted test reactions were identified by calculating the Euclidean distance between their latent space vectors. Applying these methods, the authors presented three examples: 1) reactions that were predicted accurately 'for the right reasons' (epoxidation of alkenes), 2) reactions that were predicted inaccurately (DA reactions), and 3) predictions that were predicted accurately 'for the wrong reasons' (Friedel-Craft's acylation of substituted benzenes).

In the first of these test cases (epoxidation of alkenes), the MT correctly attributed selectivity for the more substituted alkene to the presence of methyl substituents for an example reaction of of 1,2-dimethyl-1,4-cyclohexadiene with mCPBA. This finding was further validated using two literature reactions, including a more challenging example involving a less common oxidising agent (dimethyldioxirane) and competition between two substituted alkenes, one of which was adjacent to an electron-withdrawing group. Overall, these results suggest that the MT's predictions for the epoxidation of alkenes are grounded in sound chemical reasoning. In contrast, the MT was unable to predict the correct product for a DA reaction between acrylonitrile and 2-methoxybuta-1,3-diene, instead predicting an acyclic molecule. To rationalise the reason behind this incorrect prediction, the authors identified the three most similar training reactions. The first of these reactions was erroneous, while the second and third belonged to a different reaction class (C-C bond formation via Grubbs methathesis rather than cycloaddition). This indicates that the MT had not effectively learned the latent representation

of a DA reaction. Further investigation revealed there were only seven [4+2] cycloadditions in the USPTO database[74] on which the MT was trained. This limited training representation likely contributed to the model's inability to correctly predict the DA product.

Regarding the Friedel-Crafts acylations, the MT correctly predicted that the acylation of F-substituted benzene would result in a *para*-product. However, the model incorrectly attributed this to the reagents rather than the F substituent. This was confirmed using adversarial samples, where the *para*-product was still predicted even when F was replaced with a *meta*-directing substituent. Further analysis revealed that the training data contained significantly more examples of *para* reactions than *meta* reactions, suggesting the presence of a scaffold bias in the model. The authors demonstrated the impact of this scaffold bias by applying a Tanimoto similarity (S_T)[75] constraint to the test data. Imposing a similarity constraint ensures the test set is structurally dissimilar from the training set, providing a more realistic assessment of model accuracy on unseen reactions. Using a threshold of $S_T < 0.4$ (where $S_T = 0$ indicates complete dissimilarity and $S_T = 1$ indicates identical structures), the accuracy of the MT dropped from 90% to 75%. A similar decrease in accuracy was observed for the best reported Graph Neural Network (GNN) model, indicating that scaffold bias is not unique to the MT.

Overall, Lee et al. demonstrated the different approaches to interpreting transformer architectures, and how these assessments can be used to validate predictions against known chemical principles while also identifying model limitations.

1.3.2 Interpretable ML in kinetic prediction

Several of the groups introduced in Sections 1.1.2 and 1.1.3 have applied interpretation methods to rationalise and validate their model's predictions.

The simplest of these analyses involve identifying key features that influence predictions. This approach was adopted by Baskin et al.,[25] who highlighted several significant features affecting their NN model's $\log k$ prediction for an example reaction between thiophenylethyl chloride and potassium iodide in acetone. These features included LUMO energy, number of Cl atoms, number of I atoms, temperature, and the solvent dielectric constant. Although the method of calculating feature importance was not specified, this early work demonstrates how even the simplest interpretative techniques can provide insights into the inner-workings of ML models.

Feature importance analyses have also proved popular within the Δ-ML community, to validate the influence of the approximate barriers towards prediction of the more accurate target. This approach was adopted by Buttar et al.,[32] who employed permutation feature importance to identify key features of their hybrid QM/ML model. Permutation feature importance ranks features based on the relative decrease in model accuracy from randomly shuffling their feature values (greatest decrease = most important). Using this method, the authors found that the DFT $\Delta G^{\ddagger}$ feature consistently ranked highest across 10 CV folds, underscoring the significance of the hybrid QM/ML approach. Similarly, groups who employed an SQM/ML approach consistently identified the lower-LOT barrier as the most relevant feature for barrier prediction. This includes Grayson et al.,[43–45] using permutation feature importance, and Núñez et al.,[47] who employed SHapley Additive exPlanations (SHAP) values.[76] The latter measures each feature's contribution to the model output, by applying principles from cooperative game theory.

Identifying key features can also help to validate predictions against known chemical principles. For instance, Coley et al.[40] performed a series of ablation

experiments to assess the relative influence of descriptors of frontier orbital, and electrostatic, interactions towards their ml-QM-GNN's predictions of E2/S_N2 reactivity. Their experiments revealed that descriptors of frontier orbital interactions, such as Fukui function values, were less relevant than descriptors of electrostatic interactions, such as atomic charges and NMR shielding constants. These observations were consistent with a previous qualitative valence bond/DFT analysis which demonstrated that E2/S_N2 competition is primarily driven by the electrostatic interactions in the E2 TS. This hence validated the reliability of Coley's model's predictions.

Similarly, Grayson et al.[43] performed a permutation feature importance analysis to rationalise their 2022 SQM/ML model's learning of C–C bond forming nitro-Michael addition reactivity. Besides the SQM-calculated barrier, which was ranked highest (*vide supra*), other significant features included the Mulliken charge of the carbonyl oxygen, electrostatic descriptors for the R_1–R_4 substituents and core MA atoms, and several steric features including Sterimol parameters.[77,78] Based on these findings, the authors explained how each key feature relates to Michael addition reactivity. For example, the charge of the carbonyl oxygen serves as an indicator of the total electron density in the π-system of each MA; as the oxygen becomes more negative, the interaction between the MA and nucleophile decreases, leading to an increased $\Delta G^{\ddagger}$. Additionally, the identified electrostatic features account for a significant portion of the electrostatic effects in the nitro-Michael addition reaction. Moreover, the relevance of steric features aligns with previous findings that steric effects play a significant role in the rate of Michael addition reactions with glutathione. Thus, feature analysis in Grayson's 2022 model is consistent with known chemical principles. A similar analysis conducted by Grayson et al. for their 2023 model[44] showed that predictions for DA reactions were also consistent with established reactivity concepts. Specifically,

they discovered that both bond forming distances in the TS had a comparable effect on predicted $\Delta G^{\ddagger}$, reflecting the concerted nature of these reactions.

Other groups have employed latent space visualisation to gain insights into their models' chemical reasoning. For instance, Green et al.[36] utilised t-distributed Stochastic Neighbor Embedding (t-SNE) maps[79] to visualise the reaction encodings of their 2020 D-MPNN model in 2D space. Using this method, the authors discovered that their D-MPNN model had learned to cluster unseen reactions by both type and reactivity, thereby illustrating the model's chemical relevance. Similarly, Corminboeuf et al.[48] employed t-SNE maps to visualise the latent space of their 3D ENN and 2D CGR D-MPNN models, which were trained on the diverse set of gas phase reactions compiled by Green et al. Corminboeuf's analysis showed that the quality of the correlation between the latent representations and the target barrier directly corresponded to model performance, validating that their models' predictions were grounded in meaningful chemical insights.

Developing on the insights gained from their t-SNE maps, Green et al.[36] further analysed their 2020 D-MPNN model by investigating how substituting hydrogen atoms with various functional groups affected the model's predictions. Their analysis revealed that the model recognised key chemical principles. For example, substitution of an electronegative group close to the reaction centre had a significant effect on predicted $\Delta E^{\ddagger}$, while any substitution away from the centre had minimal impact. Moreover, the authors supported these findings by examining how the contribution towards predicted $\Delta E^{\ddagger}$ - measured by the norm of the difference fingerprint between the reactant and product encodings - varies with proximity to the reaction centre. This subsequent analysis revealed that contributions towards predicted $\Delta E^{\ddagger}$ diminish as atoms move farther from the centre. Thus, these additional insights reinforce the chemical validity of Green's model's predictions.

Other visualisation techniques for interpreting ML models include Principal Component Analysis (PCA),[80] a dimensionality reduction method that enables high-dimensional data to be visualised in 2D space by representing it through its first two principal components. This method was adopted by von Lilienfeld et al.[39] to visualise the kernel matrix of their E2 and S_N2 barrier predicting OHE KRR models. By examining confidence ellipsoids of the covariance between the first two principal components, they identified intuitive clusters corresponding to LG/nucleophile combinations, thereby demonstrating the chemical relevance of their approach. To analyse their OHE KRRs further, von Lilienfeld et al. plotted the difference between the predicted E2 and S_N2 barriers based on LG, nucleophile and R groups. From this plot, the authors uncovered several key reactivity rules identified by their models. For instance, electron-donating substituents at the reaction centre were predicted to favour the E2 pathway, while disfavouring nucleophilic attack in the S_N2 reaction. These observations align with heuristic reactivity rules, confirming the validity of von Lilienfeld's predictions.

In a novel approach to model interpretation, Persson et al.[49] exploited the multitask prediction functionality of their ENN model to validate its predictions of S_N2 reactivity. Specifically, they compared the predicted TS HOMO coefficients to those calculated using DFT, allowing them to assess the quality of the model's predictions. The authors found that the predicted HOMOs aligned almost perfectly with DFT-calculated values for each of the three example reactions tested. This agreement underscores the chemical validity of the model, providing confidence in its predictions.

In addition to validating predictions, interpretation methods can help to uncover learned mathematical relationships. This was demonstrated by Madzhidov et al.,[35] who explored the temperature dependence in their conjugated and single-task

(non-Arrhenius based) models by observing the change in predicted $\log k$ with varying temperature. To achieve this, the authors compiled a new test set of 61 reactions and 2412 $\log k$ values calculated from the Arrhenius equation at different temperatures. In the conjugated model, these temperatures were incorporated directly into the prediction of $\log k$ via the Arrhenius equation, while in the single-task model, they were included as descriptors. The results indicated that the conjugated model provided more accurate predictions of $\log k$ across the new temperatures compared to the single-task model, with R^2 values of 0.96 and 0.88, respectively. Thus, while both models accounted for temperature effects, the conjugated model did so with greater accuracy.

Interpretability can also help in rationalising model limitations. To understand why their S_N2 RF model (2021) generalised better to new chemical transformations than to new solvents, Madzhidov et al.[33] analysed the features used by the RF to construct its decision trees. They found that, although solvent featured constituted only 5% of the overall feature set, their model often prioritised these features. Therefore, the limited generalisability to new solvents could not be explained by neglecting solvent effects. Instead, the RF primarily used solvent features as branching criteria in the first and second levels of the trees, creating distinct sub-models for each solvent type. Consequently, this segmentation prevented the model from generalising across different solvents.

1.4 E2 and S_N2 reactions

1.4.1 S_N2 reactions

S_N2 reactions are bimolecular processes that involve the backside attack of a nucleophile (Nu) on an electrophilic carbon (C) bonded to a leaving group (LG).

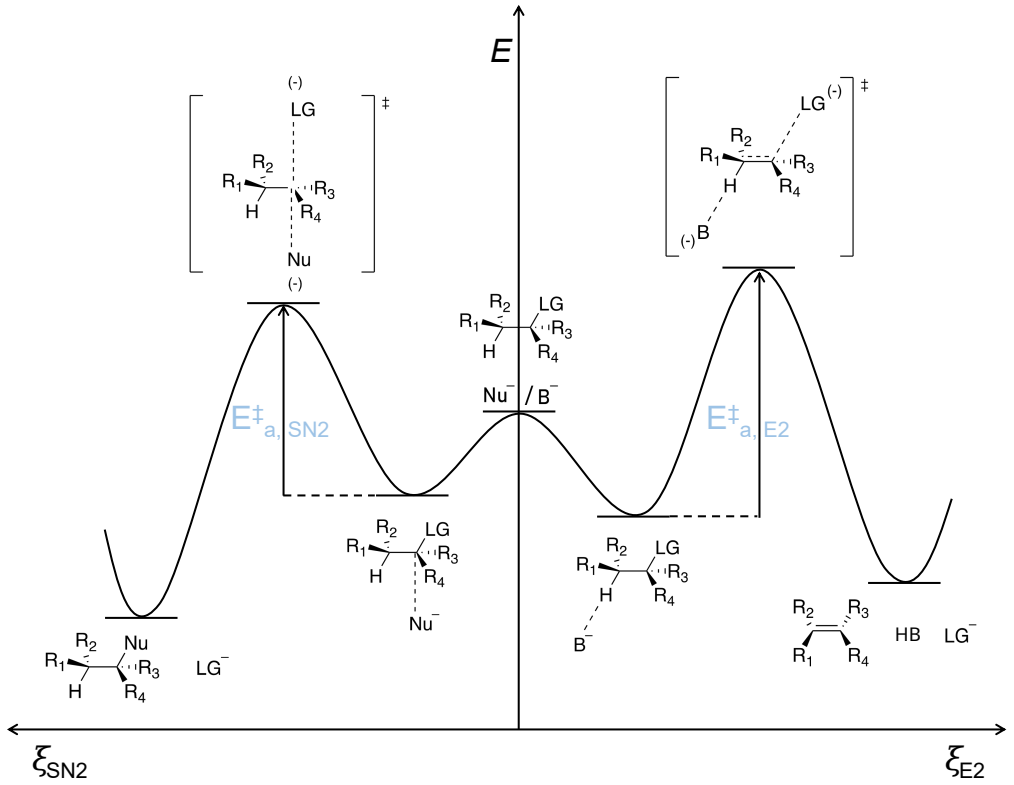


Figure 1.4: Reaction coordinate (ξ) diagram for competing E2/S_N2 reactions. *Left:* S_N2 reactions are bimolecular processes that involve the backside attack of a nucleophile (Nu^-) on an electrophilic carbon bonded to a leaving group (LG). This reaction occurs via a concerted mechanism that proceeds through a trigonal bipyramidal TS, resulting in an inversion of stereochemistry. *Right:* E2 reactions are also bimolecular processes, but instead of performing nucleophilic attack on the electrophilic centre, the nucleophile acts as a base (B^-), abstracting one of the β-hydrogens to form an allylic bond. Like S_N2, the E2 pathway occurs via a concerted mechanism, but with the additional requirement that the LG and β-hydrogen are anti-periplanar. Here, energy is represented by E.

This reaction occurs via a concerted mechanism that proceeds through a trigonal bipyramidal TS (Figure 1.4), resulting in an inversion of stereochemistry.

The S_N2 rate law,

$$S_N2 \text{ rate} = k[\text{Nu}][\text{Substrate}] \tag{1.2}$$

where k is the bimolecular rate constant, is first-order with respect to both the

nucleophile, and substrate, and is second-order overall. The structural and physical factors[81,82] that affect S_N2 rate are described below.

Nucleophilic effects. Eq. 1.2 shows that the S_N2 rate is dependent on the nucleophile. Thus, increasing its nucleophilicity increases the rate of reaction. This may be achieved by using an anionic, rather than a neutral nucleophile, or by decreasing the electronegativity of the nucleophilic atom, for example using a thiol rather than an alcohol. In the latter case, the less electronegative element has a weaker hold of its electrons, making them more available for donation. Regarding size of the nucleophilic atom, this is highly solvent-dependent. In aprotic solvents, smaller atoms, with their higher charge density, are more nucleophilic. However, in protic solvent, hydrogen bonding creates a solvation cage around the nucleophile, making larger atoms (which are less susceptible to these effects) more nucleophilic. Nucleophilicity is also affected by resonant effects, by delocalising electron density away from the nucleophilic centre and decreasing nucleophilicity.

LG effects. S_N2 rate is also enhanced by increasing the stability of the LG. This is achieved by increasing either the size or electronegativity (hence decreasing the basicity) of the LG atom. The premise here is that less basic atoms are less likely to share their electrons, resulting in a weaker C-LG bond. Similarly, resonant effects increase LG stability (e.g., tosylate) by delocalising the electron density across atoms.

Steric effects. Increased steric hinderance at the electrophilic centre results in decreased S_N2 reactivity, by preventing nucleophilic attack. Accordingly, unsubstituted or primary electrophilic centres are preferred over secondary or tertiary centres. Similarly, extensive branching on the nucleophile hinders reaction via the S_N2 pathway.

Allylic effects. S_N2 reactivity can be significantly enhanced by alkene, alkyne and aromatic groups adjacent to the electrophilic centre. This is due to the increased stability of the TS from conjugation between the allylic bond and the electron deficient p orbital at the reaction centre (Figure 1.5**a**). The resulting rate increase is even more prominent for aromatic substituents, which are more effective at π conjugation than isolated allylic bonds.

Carbonyl effects. Among the best groups for promoting S_N2 reactivity are carbonyls adjacent to the electrophilic centre. The π^* orbital from C=O and σ^* from the C-LG combine to form a new, lower energy LUMO, increasing the electrophilicity of both carbons (Figure 1.5**b**).

Solvent effects. Polar, aprotic solvents are favoured for S_N2 reactivity. This solvent type is polar enough to dissolve ionic reagents and stabilise the charged TS, but does not form hydrogen bonds with the nucleophile, allowing it to retain its nucleophilicity.

1.4.2 E2 reactions

E2 reactions are one of several mechanisms that compete with the S_N2 pathway. E2 reactions are also bimolecular processes, but instead of performing nucleophilic attack on the electrophilic centre, the nucleophile acts as a base, abstracting one of the β-hydrogens to form an allylic bond. Much like S_N2, the E2 pathway occurs via a concerted mechanism, however with the additional requirement that the LG and β-hydrogen must be periplanar (Figure 1.4). The anti- rather than the syn- periplanar conformation is favoured, allowing the substituents attached to the C-C bond to be staggered. Accordingly, E2 reactions with one available β-hydrogen are stereospecific, proceeding via the conformation where the LG and β-hydrogen are anti-periplanar. Conversely, E2 reactions where two β-hydrogens

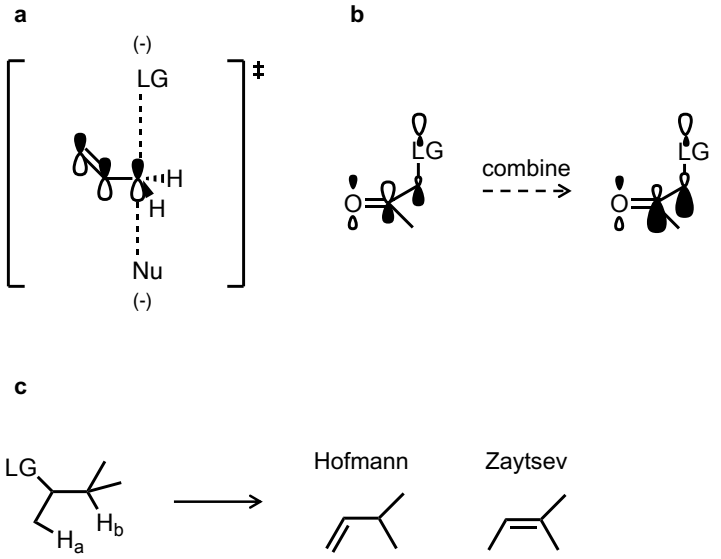


Figure 1.5: Concepts of E2 and S_N2 reactivity. a. Allylic bonds stabilise the S_N2 TS through conjugation with the electron deficient p orbital at the reaction centre. **b.** The π^* orbital from an adjacent carbonyl combines with the σ^* from the C-LG to form a new, lower energy LUMO, increasing the electrophilicity of both carbons. **c.** Hofmann[83,84] and Zaystev's[85,86] rules predict the major product of regioselective E2 reactions. According to Hofmann's rule, the least substituted alkene is the major product (abstraction of H_a). Conversely, Zaystev's rule predicts the most substituted alkene to be the major product (abstraction of H_b).

are available (on the same carbon) are stereoselective, favouring the conformation where an anti-periplanar arrangement of the LG and β-hydrogen coincides with the least unfavourable gauche interactions.

The E2 rate law,

$$\text{E2 rate} = k\,[\text{Base}]\,[\text{Substrate}] \tag{1.3}$$

where k is the bimolecular rate constant, is first-order with respect to both the base, and substrate, and is second-order overall. The structural and physical factors[87,88] that affect E2 rate are described below.

Identity of the base. Eq. 1.3 shows that the E2 rate is dependent on the base. Thus, increasing basicity increases the rate of reaction. This can be achieved by

decreasing the size or electronegativity of the hydrogen acceptor (*vide supra*).

LG effects. Generally, the LG affects E2 reactions in the same way as S_N2 reactions, as it plays a similar role in both mechanisms. An exception to this is cationic LGs, which are discussed in Section 1.4.3.

Steric effects. Unlike S_N2 reactions, E2 reactions are typically promoted by alkyl groups substituting the electrophilic centre. The electron donation provided by these groups help to stabilise the allylic bond forming in the TS. This electron-donating effect often outweighs any steric effects, which are usually negligible for hydrogen abstraction. Exceptions to this are discussed in *Hofmann and Zaytsev rules*, below.

Allylic effects. Similar to S_N2 reactions, E2 reactivity is significantly enhanced by allylic groups adjacent to the reaction centre. In the case of E2, these groups promote reactivity by stabilising the TS through conjugation with the forming allylic bond.

Solvent effects. Like S_N2 reactions, polar, aprotic solvents are favoured for E2 reactivity. These solvents are sufficiently polar to dissolve the anionic reactants while maintaining the basicity of the base.

Hofmann and Zaytsev's rules. E2 reactions with β-hydrogens available on multiple distinct carbons are regioselective. The predominant product of these reactions can be predicted using Hofmann[83,84] and Zaytsev's[85,86] rules (Figure 1.5**c**). Zaytsev's rule predicts that the major product will be the alkene with the most alkyl substituents, as these electron-donating groups enhance the stability of the double bond forming in the TS. Alternatively, Hofmann's rule predicts that the alkene with the least alkyl substituents will predominantly form. This preference is typically observed for reactions with a cationic LG. Cationic LGs

have an electron withdrawing effect on both β-hydrogens, making them more acidic (thus, more readily abstracted by a base). This effect will be dampened by electron-donating substituents, favouring deprotonation at the least substituted carbon. Other factors that favour the formation of the Hofmann product include other electron withdrawing LGs such as F, β-substituents that stabilise negative charge formation on the β-carbon, and extensive branching of the LG, base, or alkyl substituents. The latter destabilises the TS through crowding, which is more prominent in the Zaytsev TS than the Hofmann TS. Additionally, adjacent allylic groups can influence selectivity by promoting the formation of whichever alkene is conjugated with them. Furthermore, additional requirements in cyclic substrates, such as necessity for the trans-diaxial conformation, or prohibiting double bond formation at a bridgehead, can override Hofmann and Zaytsev's rules for determining selectivity.

1.4.3 E2/S_N2 competition

Under kinetic control, the competition between the E2 and S_N2 pathways depends entirely on the relative reaction rates. The structural and physical effects[81,82,87,88] driving this competition are detailed below.

Identity of the nucleophilic base. Intuitively, strong bases that are weak nucleophiles favour the E2 pathway, while strong nucleophiles that are weak bases favour the S_N2 pathway. Additionally, sterically crowded nucleophiles/bases favour the E2 pathway, owing to the reduced impact of steric interference on hydrogen abstraction compared to nucleophilic attack.

LG effects. Cationic LGs favour the E2 pathway by increasing the acidity of the β-hydrogens through electron-withdrawing effects. Additionally, these LGs

often carry a lot of steric bulk (e.g., trimethylamine), disfavouring formation of the S_N2 TS.

Steric effects. Increasing steric hindrance at the electrophilic centre disfavours the S_N2 pathway by preventing nucleophilic attack. Meanwhile, increased substitution by electron-donating alkyl groups favours the E2 pathway by stabilising the double bond forming in the TS. Therefore, unsubstituted and primary substituted centres typically favour S_N2 reactions, while tertiary centres typically favour the E2 reactions (and secondary centres are a mixed case).

Allylic effects. Allylic groups adjacent to the reaction centre promote both E2 and S_N2 reactivity. However, this effect is more prominent in E2 reactions due to conjugation with the forming alkene bond. Hence, the E2 pathway is favoured.

Temperature effects. While higher temperatures increase the rate of both pathways, this effect is more prominent in E2 reactions. This can be understood by studying the thermodynamic relationship between $\Delta G^{\ddagger}$, temperature (T), and the activation entropy ($\Delta S^{\ddagger}$),

$$\Delta G^{\ddagger} = \Delta H^{\ddagger} - T\Delta S^{\ddagger} \tag{1.4}$$

where $\Delta H^{\ddagger}$ is the activation enthalpy. E2 reactions, transitioning from two reactants to three products, have a larger $\Delta S^{\ddagger}$ than S_N2 reactions, where the transition is from two reactants to two products. Therefore, the $-T\Delta S^{\ddagger}$ term in Eq. 1.4 is larger in E2 reactions than in S_N2 reactions. Consequently, higher temperatures having a more significant effect in lowering $\Delta G^{\ddagger}$ for E2 reactions.

1.5 Contributions of this Thesis

This research focuses on the training of accurate and interpretable ML models for kinetic prediction. Specifically, the BERT architecture is employed to predict experimental $\log k$ values for E2 and S_N2 reactions, and its predictions are interpreted in the context of known reactivity rules. This work builds on previous contributions from Madzhidov et al.[33] who predict experimental $\log k$ values for individual reaction mechanisms (including E2 and S_N2) using RF models. The contributions of this Thesis include: 1) comparatively analysing the use of BERT and RF for kinetic prediction, 2) interpreting both RF and BERT models in physical organic chemistry terms, and 3) fine-tuning BERT to predict kinetics for two competing reaction classes.

This research is divided into two sections. In the first section (Chapter 3 of this Thesis), a BERT model is developed to predict experimental $\log k$ values of S_N2 reactions. Its efficacy is compared to the RF model developed by Madzhidov et al.[33] in terms of accuracy, training time, and ability to replicate known reactivity rules. In the second section (Chapter 4), this BERT model is fine-tuned to predict experimental $\log k$ values for both E2 and S_N2 reactions. BERT's learning of the structural and physical factors driving E2/S_N2 competition is then assessed.

Motivated by the success of BERT in reaction classification[60] and yield prediction,[61] this work aims to extend its application to kinetic prediction. The focus is on E2 and S_N2 reactions, owing to the availability of experimental data and well-established reactivity rules that serve as benchmarks for machine-learned predictions. By training BERT on experimental data, this approach avoids the error accumulation commonly associated with QM-generated datasets.

In this work, BERT is fine-tuned to predict the rates of two competing reaction mechanisms, exploring how its accuracy and recognition of electronic and

structural effects evolve when transitioning from single- to multi-class datasets. While previous models trained on QM-generated data have been applied to multiple reaction classes, this work tests the performance of models trained on experimental data in this context. By validating BERT's predictions against known reactivity rules, this work aims to uncover the underlying chemistry behind the model's decisions, thereby increasing chemists' confidence in applying machine-learned kinetics to guide synthetic design.

2 | Theory

2.1 QM methods

As introduced in Section 1.1, the experimental rate constant, k, can be deduced from the Gibbs free energy of activation, $\Delta G^{\ddagger}$, using Eq. 1.1.[7] $\Delta G^{\ddagger}$ can be estimated using Quantum Mechanics (QM) methods by considering it a correction to the electronic activation energy, $\Delta E_{\mathrm{a}}^{\ddagger}$,

$$\Delta G^{\ddagger} = \Delta E_{\mathrm{a}}^{\ddagger} + \text{thermal corrections} \tag{2.1}$$

the nature of the thermal corrections are discussed *vide infra*. $E_{\mathrm{a}}^{\ddagger}$ is calculated as the potential energy difference between a TS and reactant state, respectively defined by a first-order saddle point, and minimum, on the potential energy surface (PES). These points are connected via the minimum energy path. During a geometry optimisation calculation, the TS and reactant state structures are optimised by minimising the gradient of the potential energy (E_{pot}) with respect to the atomic positions, $\boldsymbol{x}$, to satisfy the condition,

$$\frac{\partial E_{\mathrm{pot}}}{\partial \boldsymbol{x}} = 0 \quad . \tag{2.2}$$

The distinction between a TS and minima on the PES lies in the nature of the second partial derivatives of E_{pot} with respect to the atomic positions, contained within the *Hessian* matrix,

$$\boldsymbol{H} = \left[\frac{\partial^2 E_{\text{pot}}}{\partial x_i \partial x_j} \right]_{ij} \tag{2.3}$$

which is positive definite for minima and has exactly one negative eigenvalue for a TS.

To refine E_{pot} at a higher LOT, single point energy calculations are often carried out on the optimised structures. The nature of the QM methods commonly employed to estimate E_{pot} (which are broadly categorised into wave function methods and Density Functional Theory (DFT)) are discussed in the following sections.

2.1.1 Wave function methods

Wave function methods estimate E_{pot} by offering an approximate solution to the time-independent Schrödinger equation,[89–92]

$$\hat{H}\Psi = E\Psi \tag{2.4}$$

where $\hat{H}$ is the Hamiltonian operator and Ψ and E are respectively the wave function and energy describing the system. $\hat{H}$ can be separated into its nuclear and electronic components by applying the Born-Oppenheimer approximation,[93] which allows decoupling of the nuclear and electronic wave functions by treating nuclear motion as stationary relative to electronic motion. The nonrelativistic electronic Hamiltonian is given as,

$$\hat{H}_{\text{elec}} = \hat{T}_{\text{elec}} + \hat{V}_{\text{elec-nuc}} + \hat{V}_{\text{elec-elec}} + \hat{V}_{\text{nuc-nuc}} \tag{2.5}$$

where $\hat{T}$ and $\hat{V}$ denote kinetic and potential energy operators, respectively, and subscripts $\cdot_{\text{elec}}$ and $\cdot_{\text{nuc}}$ denote electronic and nuclear terms. $\hat{T}_{\text{elec}}$ and $\hat{V}_{\text{elec-nuc}}$ both depend on a single set of unfixed coordinates, while $\hat{V}_{\text{elec-elec}}$ depends

on two sets of unfixed coordinates (and $\hat{V}_{\text{nuc-nuc}}$ is constant under the Born-Oppenheimer approximation). Hence, $\hat{T}_{\text{elec}}$ and $\hat{V}_{\text{elec-nuc}}$ are generally solvable, while $\hat{V}_{\text{elec-elec}}$ becomes intractable for systems with more than one electron. Thus, approximations are necessary to simplify its evaluation. One such approximation is the Hartree-Fock self-consistent field.[94,95]

Hartree-Fock self-consistent field. Hartree-Fock (HF) theory[94,95] applies the *mean-field approximation*, allowing $\hat{V}_{\text{elec-elec}}$ to be expressed as,

$$\hat{V}_{\text{elec-elec}} \approx \sum_{i=1}^{N} \nu(X_i) \tag{2.6}$$

where $\nu(X_i)$ is the average repulsive potential felt by an electron with the set of space-spin coordinates X_i, and N is the total number of electrons. The mean-field approximation allows $\hat{H}_{\text{elec}}$ to be rewritten as a sum of one-electron operators known as Fock operators, $\hat{f}(X_i)$,

$$\hat{H}_{\text{elec}} = \sum_{i=1}^{N} \hat{f}(X_i) \quad , \quad \hat{f}(X_i) = -\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M} \frac{Z_A}{d(i,A)} + \nu(X_i) \tag{2.7}$$

where Z_A and M are the nuclear charge of atom A, and the total number of atoms, respectively, $d(i,A)$ is the Euclidean distance between electron i and nucleus A, and ∇^2 denotes the Laplacian operator. Note that the first and second terms of the Fock operator are respectively the one-electron equivalents of the $\hat{T}_{\text{elec}}$ and $\hat{V}_{\text{elec-nuc}}$ terms in Eq. 2.5.

The mean-field approximation considers electrons to be uncorrelated, allowing Ψ to be expressed as a *Hartree product* of one-electron wave functions known as spin-orbitals, $\chi_{a:k}$,

$$\Psi_{\text{HP}}(X_1, X_2, \ldots, X_N) = \chi_a(X_1)\chi_b(X_2)\ldots\chi_k(X_N) \quad . \tag{2.8}$$

The intuition behind the Hartree product arises from the probabilistic relationships $P(X_i) = |\Psi(X_i)|^2$ and $P(B \cap C) = P(B)(C)$ for independent (uncorrelated) events B and C. To satisfy the conditions $\chi_a(X_i)\chi_b(X_j) = \chi_a(X_j)\chi_b(X_i)$ (the indistinguishability of electrons) and $\Psi(X_i, X_j) = -\Psi(X_j, X_i,)$ (the Pauli exclusion principle[96]), the Hartree products are combined antisymmetrically in the form of a *Slater determinant*[97] to provide an ansatz for the wave function,

$$\Psi(X_1, X_2, \ldots, X_N) = N!^{-\frac{1}{2}} \begin{vmatrix} \chi_a(X_1) & \chi_b(X_1) & \ldots & \chi_k(X_1) \\ \chi_a(X_2) & \chi_b(X_2) & \ldots & \chi_k(X_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_a(X_N) & \chi_b(X_N) & \ldots & \chi_k(X_N) \end{vmatrix} \tag{2.9}$$

where $N!^{-\frac{1}{2}}$ is a normalisation constant. Given N electrons and k spin-orbitals, one can construct $\frac{k!}{N!(k-N)!}$ different Slater determinants. However, HF theory only utilises one such determinant.

The spin-orbitals in Eqs. 2.8 and 2.9 are comprised of a radial and spin component. The spin component, which denotes whether the electron is spin-up (α) or spin-down (β), is unaffected by the nonrelativistic Hamiltonian. The radial component is given by the linear combination,

$$\psi(\boldsymbol{r}_i) = \sum_{p=1}^{m} c_{pi}\varphi_p(\boldsymbol{r}_i) \tag{2.10}$$

for an electron with spacial coordinates $\boldsymbol{r}_i$, where c_{pi} denote expansion coefficients and $\varphi_p(\boldsymbol{r}_i)$ denote *basis functions*. The total number of basis functions is given by m. Basis functions are best described by a Slater-type orbital (STO),[98]

$$\varphi(\boldsymbol{r}_i) = \mathcal{N} d(i, A)^{n-1} e^{-\zeta d(i,A)} \tag{2.11}$$

which effectively captures the radial distribution of an electron, i, around a nucleus, A, characterised by a cusp at A and an exponential decay with increasing distance, $d(i, A)$. The rate of exponential decay is controlled by the ζ term, whose magnitude is inversely correlated with diffuseness of the orbital. The n and $\mathcal{N}$ terms in Eq. 2.11 denote the principal quantum number, and a normalisation constant, respectively. Calculating STOs can be computationally demanding, so they are typically approximated using a linear combination of Gaussian orbitals,[99] which can be obtained by squaring the $d(i, A)$ term in the exponent of Eq. 2.11.

The set of employed basis functions forms the *basis set*. Two commonly used basis set families are the Ahlrichs def2 basis sets[100] (denoted def2-xVP, where x = S, TZ, QZ, ..., and P = P, PP) which are recommended for DFT calculations due to their extensive coverage of the periodic table and appropriately designed auxiliary basis sets used in Resolution of Identity approximations (for the efficient evaluation of Coulomb integrals),[101] and correlation-consistent basis sets[102,103] (denoted cc-pVxZ, where x = D, T, Q, ...) which are recommended for high-accuracy wavefunction based calculations due to their systematic convergence of correlation energy which enables convenient complete basis set extrapolation (*vide infra*).[104] For both families of basis sets, the number of basis functions assigned to each electron-occupied space is given by their *zeta* number, e.g., *double*-, *triple*-, and *quadruple*-zeta (DZ, TZ, and QZ) for two, three, and four basis functions, respectively, with accuracy (and computational intensity) increasing with increasing zeta number. The theoretical accuracy limit attainable using an infinite number of basis functions is known as the *complete basis set* limit, and can be approximated by extrapolating the results from calculations performed with increasingly larger finite basis sets. Note that the DZ def2 basis set is *split-valence* (SV), meaning valence electrons are described using two basis functions and core electrons are described using a single basis function. Typically, only valence

electrons are described using basis functions (denoted using 'V' for def2 basis sets >DZ and correlation-consistent basis sets), as these are the electrons that usually participate in bonding. Functions with the next highest orbital angular momentum (known as *polarization* functions) may also be added to the basis set to account for the difference in the space occupied by electrons in molecules compared to in atoms. Polarization functions are denoted using 'P' (for one set of polarization functions) or 'PP' (for two sets of polarization functions) for def2 basis sets, and 'p' for correlation-consistent basis sets. Additionally, *diffuse* functions (typically Gaussian functions with a small ζ) may also be added to the basis set to better describe anions, and long-range interactions such as Van der Waals forces. Diffuse functions are denoted using an appended 'D' for def2 basis sets, and the prefix 'aug-' for correlation-consistent basis sets.

Having established the form of $\hat{H}_{\text{elec}}$ and Ψ, an approximation to the ground state electronic energy can be obtained by applying the *variational principle*,

$$E_{\text{HF}} = \frac{\langle \Psi | \hat{H}_{\text{elec}} | \Psi \rangle}{\langle \Psi | \Psi \rangle} \geq E_0 \tag{2.12}$$

which considers the energy resulting from the HF calculation, E_{HF}, to be an upper bound to the true ground state energy, E_0. Thus, E_{HF} can be optimised with respect to the spin-orbitals to achieve a best estimate of E_0. This is done iteratively until the orbitals are consistent with the field they generate, giving rise to the name *self-consistent field*.

The difference between the HF energy and true ground state energy is known as the *correlation energy*,[105] which arises from the instantaneous interactions between electrons. While the antisymmetric wave function accounts for correlation between electrons of parallel spin (known as exchange correlation), the remaining correlation between the spatial positions of electrons due to their coulomb repulsion

is not accounted for. One approach to introducing this correlation into the wave function, which is utilised in this study, is Coupled Cluster theory.[106,107]

Coupled Cluster theory. Coupled Cluster (CC) theory[106,107] introduces electron correlation into the wave function via the *cluster operator*, $\hat{T}$ (not to be confused with the kinetic energy operator in 2.5), as,

$$\Psi_{\mathrm{CC}} = e^{\hat{T}}\Psi_{\mathrm{HF}} \tag{2.13}$$

where Ψ_{CC} and Ψ_{HF} denote the CC and HF wave functions, respectively. $\hat{T}$ accounts for electron correlation by generating electronic excitations into virtual orbitals, and has the form,

$$\hat{T} = \sum_{m=1}^{M} \hat{T}_m \tag{2.14}$$

where $\hat{T}_m$ is the cluster operator corresponding to the mth level of excitation, and M denotes the maximum excitation level considered. For computational feasibility, Eq. 2.14 is usually truncated at $m = 2$. $\hat{T}_1$ and $\hat{T}_2$ are then given by,

$$\hat{T}_1 = \sum_i \sum_a t_i^a \hat{a}_a^\dagger \hat{a}_i \tag{2.15}$$

$$\hat{T}_2 = \frac{1}{4} \sum_{i,j} \sum_{a,b} t_{ij}^{ab} \hat{a}_a^\dagger \hat{a}_b^\dagger \hat{a}_i \hat{a}_j \tag{2.16}$$

where the t terms are amplitudes, and the $\hat{a}_p^\dagger$ and $\hat{a}_q$ terms are the creation and annihilation operators, respectively, that add an electron to virtual orbital p, and remove an electron from occupied orbital q.

$e^{\hat{T}}$ may be expanded as a Taylor series,

$$e^{\hat{T}} = 1 + \hat{T} + \frac{1}{2!}\hat{T}^2 + \dots \tag{2.17}$$

allowing higher-level excitations to be included implicitly via the higher-order terms. Nonetheless, explicit consideration of triple excitations is typically required to achieve chemical accuracy. Triples are often treated perturbatively, at favourable computational cost to incorporating them directly into the cluster operator (scales as $O(N^7)$ in the number of basis functions rather than $O(N^8)$), giving rise to Coupled Cluster with Singles, Doubles, and perturbative Triples (CCSD(T)),[108] which is utilised in this study.

2.1.2 DFT

Density Functional Theory (DFT) offers an exact solution to the ground state electronic energy by expressing it as a functional of the electron density, ρ. The theoretical foundation for DFT is established within the Hohenberg-Kohn theorems,[109] namely: 1) the external potential in which the electrons move (i.e., the $V_{\text{elec-nu}}$ term in Eq. 2.5) can be uniquely determined from the ground state electron density, and 2) the true ground state energy can be obtained from the energy functional by applying a variational principle, similar to in Eq. 2.12.

The form of the energy functional for practical (Kohn-Sham) DFT[110] is given by,

$$E[\rho(\boldsymbol{r})] = T_{\text{elec}}[\rho(\boldsymbol{r})] + V_{\text{ext}}[\rho(\boldsymbol{r})] + V_{\text{J}}[\rho(\boldsymbol{r}] + V_{\text{xc}}[\rho(\boldsymbol{r})] \tag{2.18}$$

where $T_{\text{elec}}[\rho(\boldsymbol{r})]$ denotes the kinetic energy of a fictitious system of non-interacting electrons with the same density as the interacting system, $V_{\text{ext}}[\rho(\boldsymbol{r})]$ denotes the external potential in which the electrons move, $V_{\text{J}}[\rho(\boldsymbol{r})]$ denotes the mean-field coulomb repulsion experienced by the electrons, and $V_{\text{xc}}[\rho(\boldsymbol{r})]$

denotes the exchange-correlational functional. The exact form of $V_{\mathrm{xc}}[\rho(\boldsymbol{r})]$ is unknown, giving rise to a ladder of approximations of increasing accuracy,[111] ranging from local density approximations (LDAs),[110] which depend entirely on ρ, to double-hybrids,[112,113] which contain a portion of the HF exchange energy and a perturbative correlation treatment (Figure 2.1). Typically, hybrid functionals[114] (rung 4 of the ladder) are employed for their favourable cost-accuracy balance (scales as $O(N^4)$ in the number of basis functions). In this study, the Perdew–Burke–Ernzerhof hybrid functional (PBE0),[115,116]

$$V_{\mathrm{xc}}^{\mathrm{PBE0}} = \frac{1}{4}V_{\mathrm{x}}^{\mathrm{HF}} + \frac{3}{4}V_{\mathrm{x}}^{\mathrm{PBE}} + V_{\mathrm{c}}^{\mathrm{PBE}} \tag{2.19}$$

is employed, which combines HF and PBE[115] (a generalised gradient approximation (GGA), rung 2 of the ladder) exchange energies in a 1:3 ratio, in addition to the PBE correlation energy.

2.1.3 Thermal corrections

G can be estimated from E_{pot} using the ideal gas model (IGM),[121,122] starting from the thermodynamic relationships,

$$G = H - T\Delta S \tag{2.20}$$

$$H = U + RT \tag{2.21}$$

where S, H, and U respectively denote the entropy, enthalpy, and internal energy of the species, and T and R denote the temperature, and ideal gas constant, respectively. By invoking the rigid rotor[123] and Born-Oppenheimer[93] approximations, U and S can be decomposed into their independent contributions,[124]

$$U = E_{\mathrm{ZPE}} + E_{\mathrm{pot}} + E_{\mathrm{trans}} + E_{\mathrm{rot}} + E_{\mathrm{vib}} \tag{2.22}$$

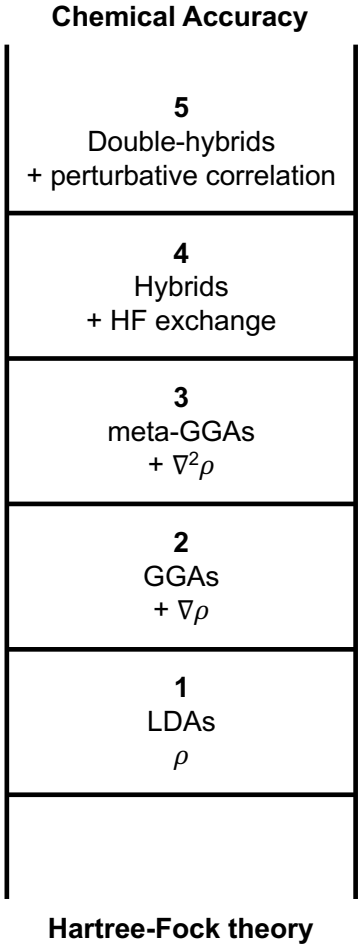


Figure 2.1: Ladder of DFT functionals.[111] Rung 1) Local density approximations (LDAs);[110] entirely dependent on the electron density (ρ). Rung 2) Generalised gradient approximations (GGAs);[117–119] dependent on ρ and its first derivative. Rung 3) meta-GGAs;[120] dependent on ρ and its first and second derivatives. Rung 4) Hybrids;[114] rung 3 + a portion of the HF exchange energy. Rung 5) Double-hybrids;[112,113] rung 4 + a perturbative correlation treatment.

$$S = S_{\text{elec}} + S_{\text{trans}} + S_{\text{rot}} + S_{\text{vib}} \tag{2.23}$$

where E denotes an energy term and subscripts $\cdot_{\text{ZPE}}$, $\cdot_{\text{pot}}$, $\cdot_{\text{elec}}$, $\cdot_{\text{trans}}$, $\cdot_{\text{rot}}$, and $\cdot_{\text{vib}}$ denote zero-point energy, potential energy, electronic, translational, rotational, and vibrational contributions, respectively. Thus, the E_{pot} resulting from the single point energy calculation is incorporated into G through U.

The translational, rotational, and vibrational terms in Eqs. 2.22 and 2.23 can be determined using the statistical thermodynamics equations,[125]

$$E = k_B T^2 \left(\frac{\partial \ln Q}{\partial T} \right)_V \tag{2.24}$$

$$S = k_B \ln Q + k_B T \left(\frac{\partial \ln Q}{\partial T} \right)_V \tag{2.25}$$

where k_B is the Boltzmann constant, and subscript $\cdot_V$ denotes constant volume. Q is the canonical partition function, and is linked to the molecular partition function, q, via,[126]

$$Q = \begin{cases} \frac{q^N}{N!} & \text{trans} \\ \\ q^N & \text{rot,vib} \end{cases} \tag{2.26}$$

where N is the number of particles (= N_A, Avogadro's constant for 1 mole). The molecular partition functions for translational, rotational and vibrational modes are given as,

$$q_{\text{trans}} = \left(\frac{2\pi m k_{\text{B}} T}{h^2} \right)^{\frac{3}{2}} V \tag{2.27}$$

$$q_{\text{rot}} = \frac{\sqrt{\pi}}{\sigma_{\text{r}}} \left(\prod_k \frac{T}{\theta_{\text{rot},k}} \right)^{\frac{1}{2}} \quad , \; \theta_{rot,k} = \frac{h^2}{8\pi^2 k_B I_k} \tag{2.28}$$

$$q_{\text{vib},n} = \frac{1}{1 - e^{-\theta_{\text{vib},n}/T}} \quad , \; \theta_{vib,n} = \frac{h\nu_n}{k_B} \tag{2.29}$$

derived from the particle in a box,[127] rigid rotor,[123,128] and harmonic oscillator[129,130] models, respectively. Here, V is the volume, m and σ_r are respectively the mass, and rotational symmetry number, of the species, and h denotes Planck's constant. $\theta_{rot,k}$ and $\theta_{vib,n}$ define the *characteristic* rotational and vibrational temperatures. The former encompasses the principal moment of inertia I_k, corresponding to rotational axis k, while the latter encompasses the vibrational frequency ν_n, corresponding to normal mode n.

Substituting Eq. 2.27 into Eqs. 2.24 and 2.25, and setting $N = N_A$, the translational terms in Eqs. 2.22 and 2.23 simplify to,[127]

$$E_{\text{trans}} = \frac{3}{2}RT \tag{2.30}$$

$$S_{\text{trans}} = R \ln \left(\frac{q_{\text{trans}}}{N_A} \right) + \frac{5}{2}R \quad . \tag{2.31}$$

Here, the V term encompassed within q_{trans} (see Eq. 2.27) is defined as $\frac{RT}{p^\circ}$ in the gas phase and $\frac{1}{c^\circ}$ in the solvent phase, where p° and c° denote the standard pressure, and concentration, respectively. From observing Eqs. 2.30, 2.31, and 2.27 it is evident that the translational components of G are calculated without relying on QM data.

Substituting Eq. 2.28 into Eqs. 2.24 and 2.25, and setting $N = N_A$, the rotational terms in Eqs. 2.22 and 2.23 simplify to,[128]

$$E_{\text{rot}} = \begin{cases} 0 & \text{if monoatomic} \\ RT & \text{if linear} \\ \frac{3}{2}RT & \text{otherwise} \end{cases} \tag{2.32}$$

$$S_{\text{rot}} = \begin{cases} 0 & \text{if monoatomic} \\ R\ln(q_{\text{rot}}) + R & \text{if linear} \\ R\ln(q_{\text{rot}}) + \frac{3}{2}R & \text{otherwise} \end{cases} . \tag{2.33}$$

Here, the σ_r and I_k terms encompassed within q_{rot} (see Eq. 2.28) are calculated from the optimised geometry of the species.

Substituting Eq. 2.29 into Eqs. 2.24 and 2.25, and setting $N = N_A$, the vibrational terms in Eqs. 2.22 and 2.23 (when summing over normal modes) have the form,[129]

$$E_{\text{vib}} = R \sum_n \frac{\theta_{\text{vib},n}}{e^{\theta_{\text{vib},n}/T} - 1} \tag{2.34}$$

$$S_{\text{vib}} = R \sum_n \frac{\theta_{\text{vib},n}}{T(e^{\theta_{\text{vib},n}/T} - 1)} - \ln(1 - e^{-\theta_{\text{vib},n}/T}) \quad . \tag{2.35}$$

Here, the vibrational frequencies encompassed in $\theta_{\text{vib},n}$ (see Eq. 2.29) are calculated from the optimised geometry of the species, by diagonalising the (mass-weighted) Hessian matrix.[131,132] S_{vib} is typically overestimated for low-frequency modes, so a correction such as Grimme's quasi-RRHO approximation,[133] an interpolation between the rigid rotor and harmonic oscillator, is applied to scale down their contribution. The vibrational frequencies are also utilised in the calculation of E_{ZPE},[134]

$$E_{\text{ZPE}} = \frac{N_A}{2} \sum_n h\nu_n \tag{2.36}$$

as defined by the harmonic oscillator model.[130]

Lastly, the S_{elec} term in Eq. 2.23 is assumed to be zero for ground state molecules and ions.

2.2 ML

While QM methods can provide reasonable estimates of $\Delta G^{\ddagger}$, they are limited in their description of solvent[17,18,20] and entropy.[17–19] Hence, Machine Learning (ML) methods, which enable direct prediction of experimental observables, have emerged as a favourable alternative. The two ML architectures employed in this study, Random Forest (RF) and Bidirectional Encoder Representations from Transformers (BERT), are described in more detail below.

Throughout the following sections, $\log k_{\text{exp}}$ and $\log k_{\text{pred}}$ denote experimental and predicted $\log k$ values, respectively.

2.2.1 RF

2.2.1.1 RF architecture

A Random Forest (RF) is an ensemble of decision trees, where each tree is trained using a random subset of samples or a random subset of the feature space.[135] The RF model[33] employed in this study is composed of 500 decision trees, each trained using a random subset of the feature space. The maximum number of features in each subset is optimised during hyperparameter tuning (*vide infra*).

The decision tree algorithm[136] is depicted in Figure 2.2. Starting at the first node (root), training samples are split into two child nodes based on whether a particular feature meets a specified threshold. The feature and threshold are selected during training as those which give the greatest improvement in Mean Squared Error (MSE),

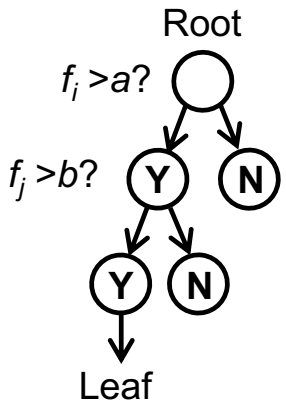


Figure 2.2: Decision tree algorithm.[136] Starting at the root node, training samples are split into child nodes based on whether a particular feature meets a specified threshold. This example shows two data splits: one for feature f_i, with a threshold of a, and one for feature f_j, with a threshold of b. The terminal node of each decision path is referred to as a leaf.

$$\text{MSE} = \frac{1}{N}\sum_{d=1}^{N}(\log k_{\text{exp},d} - \overline{\log k_{\text{exp},c}})^2 \quad . \tag{2.37}$$

Here, N is the total number of training samples at the root node, $\log k_{\text{exp},d}$ is the $\log k_{\text{exp}}$ of data point d, and $\overline{\log k_{\text{exp},c}}$ is the mean $\log k_{\text{exp}}$ of data points at child node c occupied by d. The data continues to be split in this manner until no further MSE improvement can be made, or a user-specified condition has been met (e.g., a maximum number of data splits is reached). No user-specified condition was imposed in this study.

Predictions for queried data points are made by traversing their decision path down each tree and calculating the mean $\log k_{\text{exp}}$ of data points occupying the terminal node (leaf). The predictions output by each tree are then averaged over the entire forest to give the final prediction. This averaging over trees, which are each trained using a random subset of the feature space, creates an ensemble effect designed to reduce overfitting.

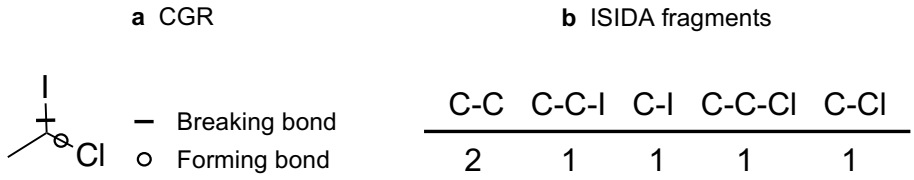


Figure 2.3: Generation of In SIlico Design and data Analysis (ISIDA) fragments[57] for the reaction of chloride with ethyl iodide. **a.** The reactants and products are converted to a Condensed Graph of the Reaction (CGR).[26] Here, dynamic bonds are represented by a circle (forming bond) and dash (breaking bond). **b.** The CGR is converted to ISIDA fragments. Here, all possible sequences of atoms and bonds containing 2-4 atoms have been generated. The counts indicate the number of times each fragment appears in the reaction.

2.2.1.2 ISIDA fragments

The RF model[33] employed in this study utilises In SIlico Design and data Analysis (ISIDA) fragments[57] as structural features of chemical reactions. To generate the ISIDA fragments, reactions are first encoded as a Condensed Graph of the Reaction (CGR), where atoms and bonds are represented by vertices and edges, respectively (Figure 2.3**a**).[26] Unlike molecular graphs, which represent single molecules, a CGR combines all species involved in a chemical transformation (reactants and products) into a single graph, where bonds changed during the transformation are represented by so-called *dynamic bonds*. ISIDA fragments are then computed as counts of molecular fragments in the CGR, such that fragments appearing on both sides of the reaction (i.e., they do not contain a dynamic bond) are counted twice (Figure 2.3**b**).

2.2.1.3 Hyperparameter optimisation with Grid Search

Hyperparameter optimisation of the RF in this study was kept consistent with the methodology used in Ref[33]. Specifically, the max_features hyperparameter, which controls the maximum number of features in each tree's random feature subset, was optimised using the Grid Search method.

Grid Search is an exhaustive search over a predefined set of hyperparameter values to find the combination that achieves the best cross-validated score. In the work of Ref[33] (and thus, in this study), the search space for max_features was defined as [0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 'auto', 'log2', None], and the goal was to minimise the mean MSE over five CV folds. Note that the max_features values correspond to a fraction of the total number of features (n_{feats}), where 'auto' = $\sqrt{n_{\text{feats}}}$, 'log2' = $\log_2 n_{\text{feats}}$, and None = n_{feats}.

2.2.2 BERT

Bidirectional Encoder Representations from Transformers (BERT)[58] is a type of Natural Language Processing (NLP) model, a category of deep learning models that process textual inputs. Prior to being fine-tuned on a specific task, such as rate prediction in this study, BERT models are pretrained on a general task such as Masked Language Modelling Loss (MLML), where the model predicts a masked token from a text sequence. Fine-tuning a pretrained model is typically more cost-effective than training from scratch, as the model weights will likely be closer to optimal for the downstream task than randomly initialised weights. In this study, SMILES strings are used as textual representations of chemical reactions. Here, "»" and "." tokens are used to separate precursors and products, and individual molecules, respectively. The BERT model trained in this study was pretrained on an MLML task using SMILES strings of reactions.[60]

2.2.2.1 BERT architecture

The architecture of the BERT model[58,60,137] utilised in this study is illustrated in Figure 2.4. Firstly, SMILES strings are tokenised (**a**) and each token is assigned a numerical embedding (**b**). The embeddings are then passed to an encoder (**c**), which is comprised of a self-attention mechanism and Feedforward Neural Network (FNN). Lastly, hidden representations output by the encoder

(specifically the hidden representation of a prepended "[CLS]" token, *vide infra*) are fed into a linear prediction layer (**d**). Each component of the BERT model is described in more detail below.

Tokenisation (Figure 2.4**a**)**.** SMILES are input into the BERT model in batches of size d_b (= 16 in this study). Here, they are tokenised and each token assigned an input ID corresponding to its position in the model's vocabulary ($1 \leq \mathrm{ID} \leq N$, where N is the total number of tokens in the vocabulary). IDs corresponding to special "[CLS]" and "[SEP]" tokens are respectively prepended and appended to the sequence. The hidden representation of the former acts as a global representation of the SMILES and is a precursor to the input of the final prediction layer, while the latter denotes the end of the sequence. The ID of special "[PAD]" tokens are appended after the "[SEP]" token to ensure all sequences have the same length of d_t tokens (d_t = 300 in this study).

Embedding (Figure 2.4**b**)**.** The input IDs are fed into an embedding layer, where each ID is assigned a feature embedding of dimension d_f (= 256 in this study). Each token embedding is added to a positional embedding of the same dimension, whose elements correspond to the position of each token in the sequence (from 1 to d_t). Embeddings are learnable parameters of the model.

Encoder (Figure 2.4**c**)**.** The embedding tensor, $\boldsymbol{E} \in \mathbb{R}^{d_\mathrm{b} \times d_\mathrm{t} \times d_\mathrm{f}}$, is passed through three separate linear layers to produce keys ($\boldsymbol{K} \in \mathbb{R}^{d_\mathrm{b} \times d_\mathrm{t} \times d_\mathrm{f}}$), queries ($\boldsymbol{Q} \in \mathbb{R}^{d_\mathrm{b} \times d_\mathrm{t} \times d_\mathrm{f}}$), and values ($\boldsymbol{V} \in \mathbb{R}^{d_\mathrm{b} \times d_\mathrm{t} \times d_\mathrm{f}}$), respectively. The general form of a linear layer in the BERT model is,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^\top + \boldsymbol{b} \tag{2.38}$$

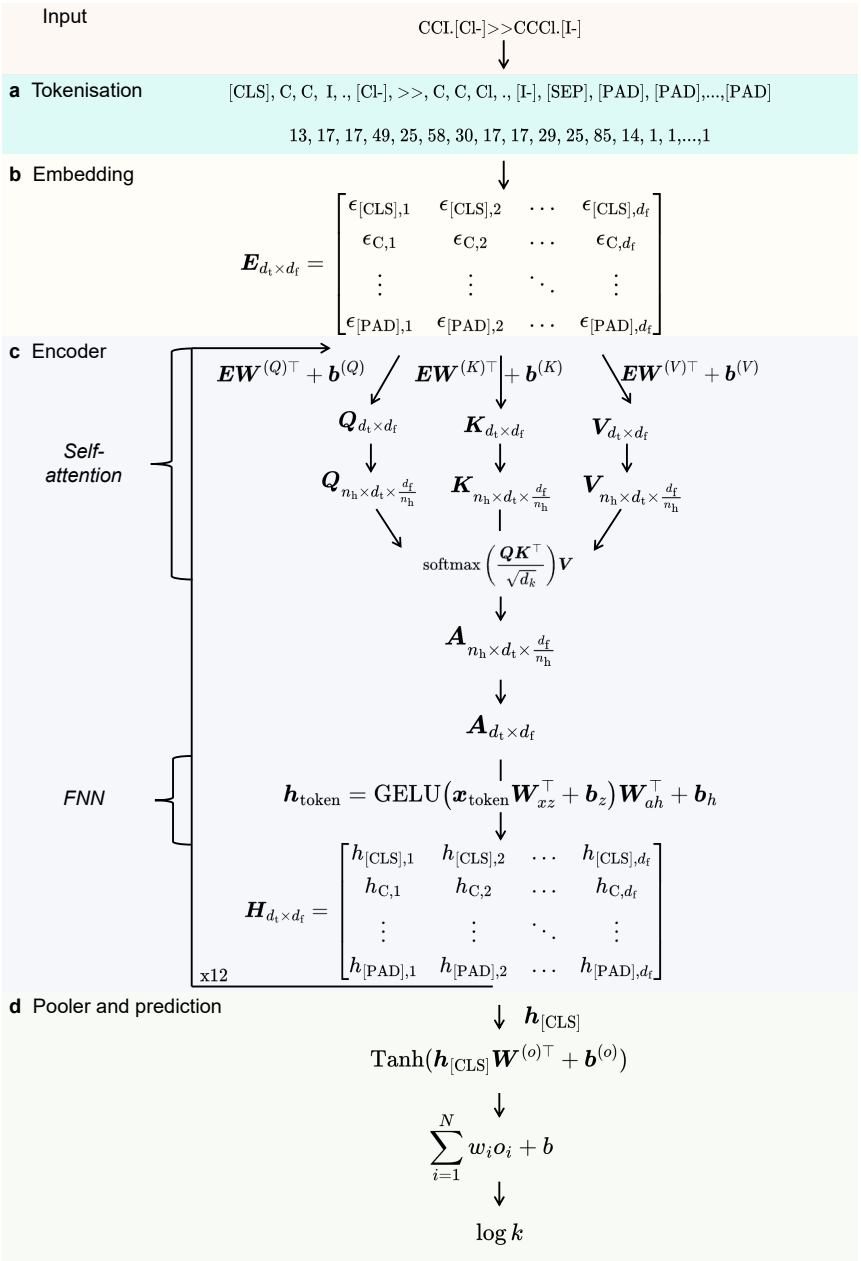


Figure 2.4: BERT architecture[58,60,137] demonstrated using a single input reaction (d_b = 1). **a.** SMILES strings are tokenised and converted to input IDs. **b.** Each ID is assigned a numerical embedding. **c.** The embeddings are passed to an encoder comprised of a *self-attention mechanism* and a *Feedforward Neural Network* (FNN). **d.** The hidden representation output by the encoder corresponding to a prepended "[CLS]" token is passed through a pooler comprised of a linear layer and the Tanh activation function. The outputs of the pooler are then fed through a linear prediction layer.

where $\boldsymbol{Y} \in \mathbb{R}^{*\times d_{\text{out}}}$, $\boldsymbol{X} \in \mathbb{R}^{*\times d_{\text{in}}}$, $\boldsymbol{W} \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$, and $\boldsymbol{b} \in \mathbb{R}^{d_{\text{out}}}$ are outputs, inputs, learnable weights, and learnable biases, respectively. Here, d_{in} and d_{out} respectively denote the number of input and output features, while * indicates any number of dimensions, which are the same for both $\boldsymbol{X}$ and $\boldsymbol{Y}$. $\boldsymbol{K}$, $\boldsymbol{Q}$ and $\boldsymbol{V}$ are subsequently reshaped from $\boldsymbol{K}, \boldsymbol{Q}, \boldsymbol{V} \in \mathbb{R}^{d_{\text{b}}\times d_{\text{t}}\times d_{\text{f}}}$ to $\boldsymbol{K}, \boldsymbol{Q}, \boldsymbol{V} \in \mathbb{R}^{d_{\text{b}}\times n_{\text{h}}\times d_{\text{t}}\times \frac{d_{\text{f}}}{n_{\text{h}}}}$, where n_{h} is the number of attention heads (= 4 in this study). The reshaped tensors are then input into a self-attention layer, which generates combined encodings of each token and their contextual relationships within the sequence. This *multi-headed* self-attention results in a more robust encoding of relationships and nuances within the input sequence compared to performing self-attention on the entire feature space at once.[137]

Self-attention mechanism. Attention outputs, $\boldsymbol{A} \in \mathbb{R}^{d_{\text{b}}\times n_{\text{h}}\times d_{\text{t}}\times \frac{d_{\text{f}}}{n_{\text{h}}}}$, are calculated according to,

$$\boldsymbol{A} = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad . \tag{2.39}$$

Firstly, attention scores are computed as the dot product of $\boldsymbol{Q}$ and $\boldsymbol{K}$. The softmax function (Eq. A.1, Appendix A) is then applied along the key dimension to normalise the attention scores, $\boldsymbol{s} \in \mathbb{R}^{d_{\text{t}}}$, to attention probabilities, $\boldsymbol{p} \in (0, 1)^{d_{\text{t}}}$, where $\sum_{i=1}^{d_{\text{t}}} p_i = 1$. Note that the scaling factor of $\frac{1}{\sqrt{d_{\text{k}}}}$ in Eq. 2.39, where d_{k} is the feature dimension of $\boldsymbol{K}$ (= 64 in this study), is to prevent large attention scores from dominating the softmax output.[137] $\boldsymbol{A}$ is then calculated from the dot product of the attention probabilities and $\boldsymbol{V}$. $\boldsymbol{A}$ is subsequently reshaped from $\boldsymbol{A} \in \mathbb{R}^{d_{\text{b}}\times n_{\text{h}}\times d_{\text{t}}\times \frac{d_{\text{f}}}{n_{\text{h}}}}$ to $\boldsymbol{A} \in \mathbb{R}^{d_{\text{b}}\times d_{\text{t}}\times d_{\text{f}}}$ and passed through a linear layer for input into a positionwise fully connected FNN.

FNN. The equation of the FNN (shown graphically in Figure 2.5) is given by,

$$\boldsymbol{H}_{\text{token}} = \text{GELU}\left(\boldsymbol{X}_{\text{token}}\boldsymbol{W}_{xz}^{\top} + \boldsymbol{b}_z\right)\boldsymbol{W}_{ah}^{\top} + \boldsymbol{b}_h \tag{2.40}$$

where $\boldsymbol{X}_{\text{token}} \in \mathbb{R}^{d_\text{b} \times d_\text{f}}$ and $\boldsymbol{H}_{\text{token}} \in \mathbb{R}^{d_\text{b} \times d_\text{f}}$ are the respective inputs and outputs of the network corresponding to position x of the token dimension. $\boldsymbol{W}_{xz} \in \mathbb{R}^{d_\text{i} \times d_\text{f}}$ and $\boldsymbol{W}_{ah} \in \mathbb{R}^{d_\text{f} \times d_\text{i}}$ are learnable weights and $\boldsymbol{b}_z \in \mathbb{R}^{d_\text{i}}$ and $\boldsymbol{b}_h \in \mathbb{R}^{d_\text{f}}$ are learnable biases. Note that the output of the first linear layer (known as the *intermediate*) has a feature dimension $d_\text{i} > d_\text{f}$ ($d_\text{i} = 512$ in this study), which helps to capture more complex interactions in the data.[137] The GELU function in Eq. 2.40 is the Gaussian Error Linear Unit activation function, which multiplies the weighted output of node i, z_i, by $P(X \leq z_i)$, the standard Gaussian cumulative distribution function evaluated at z_i. In general, activation functions help to capture node interactions and non-linear relationships in the data by multiplying a portion of the previous layer's outputs by 0. GELU does this both deterministically and stochastically, with the latter also helping to reduce network complexity and prevent overfitting.[138] The final network outputs from each sequence position, known as the hidden state, $\boldsymbol{H}$, are input into the next encoder layer and this cycle repeated until a total of n_l encoder layers have been propagated through ($n_\text{l} = 12$ in this study).

Pooler and prediction (Figure 2.4**d**)**.** The final hidden state corresponding to the "[CLS]" token, $\boldsymbol{H}_{[\text{CLS}]}$, is input into a pooler which comprises a linear layer and the Tanh activation function. The latter computes the hyperbolic tangent of each element h_{ij} (Eq. A.2, Appendix A). The outputs of the pooler are input into a linear regressor, which predicts $\log k$ for each data point, d, as,

$$\log k_d = \sum_{i=1}^{d_\text{f}} w_i o_i + b \quad . \tag{2.41}$$

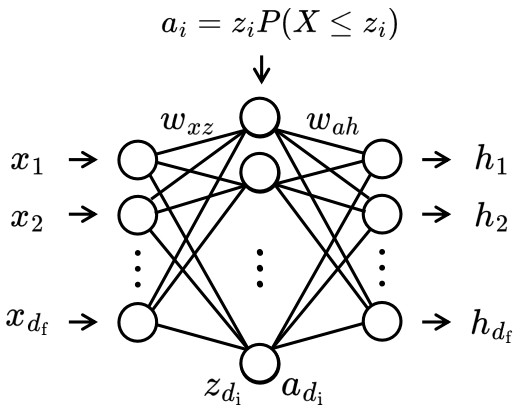


Figure 2.5: The Feedforward Neural Network (FNN) in the BERT model is comprised of two linear layers separated by the GELU activation function. The layers are fully connected and the inputs enter tokenwise. Here, x_i, z_i, a_i, and h_i denote network inputs, weighted inputs, activations, and hidden representations, respectively, w_{ij} denotes the weight connecting nodes i and j, d_f and d_i denote the model feature dimension and intermediate feature dimension, respectively, and the $P(X \leq z_i)$ term in the activation function denotes the standard Gaussian cumulative distribution function evaluated at z_i.

Here, o_i are elements of the pooler output corresponding to data point d, d_f is the feature dimension of the model (= 256 in this study), and w_i and b are learnable weights and a learnable bias, respectively.

Dropout, normalisation, and residual connections. *Dropout.* Dropout adds regularisation to deep learning architectures by stochastically multiplying some layer outputs by zero, which prevents overfitting.[139] In the BERT model, dropout is applied to all fully connected layers, with a probability D, and to the attention probabilities, with a probability D_A. In this study, D was optimised during hyperparameter tuning (*vide infra*), and D_A was set to its default value[60] of 0.1.

Normalisation. BERT normalises layer inputs using layer normalisation (Eq. A.3, Appendix A).[140] Normalising layer inputs reduces training time by preventing internal covariate shift, the change in distribution of network activations as a result of the changing model parameters. Layer normalisation is ideal for small batch sizes, as it normalises across the feature dimension rather than the batch

dimension. In the BERT model, layer normalisation is applied to the embeddings, attention outputs, and outputs of the FNN.

Residual connections. Residual connections sum layer outputs and inputs; this *identity mapping* helps to prevent vanishing or exploding gradients during training.[141] In the BERT model, residual connections are applied to the attention outputs, and outputs of the FNN.

2.2.2.2 Training BERT

BERT is trained over multiple passes of the training data, called epochs. Each epoch contains $\frac{N}{d_\mathrm{b}}$ iterations, where N is the total number of training samples and d_b is the batch size (= 16 in this study). Each iteration is comprised of one forward pass of a data batch through the model and one backwards pass to update the model parameters. During the backwards pass, the loss with respect to the true target values is calculated using the predictions made during the forward pass. The gradient of the loss with respect to the model parameters is then evaluated using backward propagation of errors (backpropagation).[142] Using the computed gradients, the model parameters are updated with an optimisation algorithm such as Adaptive Moment Estimation (AdamW, where W denotes decoupled weight decay, *vide infra*)[143,144] to minimise the loss. The loss function, backpropagation, and AdamW are described in more detail below. Note that in this study, BERT was trained for 10 epochs.

Loss function. The accuracy of BERT's predictions are evaluated using the MSE loss function,

$$\mathrm{MSE} = \frac{1}{d_\mathrm{b}} \sum_{d=1}^{d_\mathrm{b}} (y_d - \hat{y}_d)^2 \tag{2.42}$$

for a batch of d_b data points, where y_d and $\hat{y}_d$ respectively denote $\log k_\mathrm{exp}$ and $\log k_\mathrm{pred}$ of data point d. Note that Eq. 2.42 becomes Eq. 2.37 when substituting $d_\mathrm{b} = N$, $y_d = \log k_\mathrm{exp}$, and $\hat{y}_d = \log k_\mathrm{pred} = \overline{\log k_{\mathrm{exp},c}}$ for the RF model.

Backpropagation. Backpropagation is an efficient application of the chain rule to calculate the gradient of the loss function with respect to the model parameters, $\boldsymbol{\theta}$.[142] As the MSE loss function can be decomposed into a sum over individual error terms (see Eq. 2.42), the gradients can be calculated for each data point and combined at the end by applying the rule $\frac{\partial}{\partial v}\sum f(v) = \sum \frac{\partial}{\partial v} f(v)$. Hence, for an individual data point, d, backpropagation evaluates the gradient of the (squared) error function $E_d = (y_d - \hat{y}_d)^2$ with respect to each model parameter.

To evaluate the gradient of E_d with respect to weight w_{ij} connecting nodes i and j in layers $l-1$ and l of a neural network, one can apply the chain rule,

$$\frac{\partial E_d}{\partial w_{ij}} = \frac{\partial E_d}{\partial a_j}\frac{\partial a_j}{\partial w_{ij}} = \frac{\partial E_d}{\partial a_j}\frac{\partial a_j}{\partial z_j}\frac{\partial z_j}{\partial w_{ij}} \tag{2.43}$$

where a_j and z_j are the activation output, and weighted input, of node j, respectively. By realising that $z_j = \sum_{p=1}^{n} w_{pj} a_p$ (where p is one of the n nodes in layer $l-1$ connected to node j), with only the $w_{ij}a_i$ term depending on w_{ij}, the $\frac{\partial z_j}{\partial w_{ij}}$ term in Eq. 2.43 can be evaluated as a_i, the activation output of node i. Similarly, by noting that $a_j = \mathrm{Act}(z_j)$, where Act is the activation function, the $\frac{\partial a_j}{\partial z_j}$ term in Eq. 2.43 can be evaluated as the first derivative of the activation function, $\mathrm{Act}'(z_i)$. Hence, Eq. 2.43 can be simplified to,

$$\frac{\partial E_d}{\partial w_{ij}} = \delta_j \mathrm{Act}'(z_i) a_i \tag{2.44}$$

$$\delta_j = \begin{cases} \frac{\partial E_d}{\partial \hat{y}}, & \text{if } j = \text{ output node} \\ \sum_{q=1}^{m} \frac{\partial E_d}{\partial a_q}\mathrm{Act}'(z_i) w_{jq} & \text{if } j \text{ in hidden layer} \end{cases} \tag{2.45}$$

where $\delta_j = \frac{\partial E_d}{\partial a_j}$. For j = output node, δ_j can be evaluated as $\frac{\partial E_d}{\partial \hat{y}}$. Else, if j is in a hidden layer, δ_j is evaluated as the total derivative shown in Eq. 2.45, where q is one of the m nodes in layer $l + 1$ receiving output from node j. Here, the chain rule has been applied to $\frac{\partial E_d}{\partial a_j}$, substituting $\frac{\partial z_q}{\partial a_j} = w_{jq}$ and $\frac{\partial a_q}{\partial z_q} = \mathrm{Act}'(z_i)$ (see derivation in Appendix B.1). The w_{jq} and $\mathrm{Act}'(z_i)$ terms in δ_j are known, while $\frac{\partial E_d}{\partial a_q} \equiv \delta_q$ is computed while evaluating the gradients in layer $l + 1$. Hence, backpropagation allows evaluation of the gradients in layer l without recomputing partial derivatives from subsequent layers. Note that the gradient of E_d with respect to biases b_j are calculated in the same way as w_{ij}, by considering b_j as a weight multiplied by an input of 1.

The gradient of the overall MSE with respect to w_{ij} is calculated from the individual gradients $\frac{\partial E_d}{\partial w_{ij}}$ using,

$$\frac{\partial \mathrm{MSE}}{\partial w_{ij}} = \frac{1}{d_\mathrm{b}} \sum_{d=1}^{d_\mathrm{b}} \frac{\partial E_d}{\partial w_{ij}} \tag{2.46}$$

where d_b is the batch size (= 16 in this study).

Adaptive Moment Estimation. During training, the model parameters, $\boldsymbol{\theta}$, are optimised to decrease the MSE using Adaptive Moment Estimation with decoupled weight decay (AdamW),[143,144] a type of Stochastic Gradient Descent (SGD).[145] SGD updates $\boldsymbol{\theta}$ based on a randomly selected batch of data at each iteration, making it more efficient than traditional gradient descent methods that do this using the entire dataset. For each iteration, i, AdamW computes the gradient of the MSE with respect to $\boldsymbol{\theta}$ as,

$$\boldsymbol{g}_i = \nabla \mathrm{MSE}_i(\boldsymbol{\theta}_{i-1}) \tag{2.47}$$

where ∇ denotes the del operator. The exponential moving averages of $\boldsymbol{g}_i$ and $\boldsymbol{g}_i \otimes \boldsymbol{g}_i$ (denoted $\boldsymbol{g}_i^2$), $\boldsymbol{m}_i$ and $\boldsymbol{v}_i$, are then calculated as,

$$\boldsymbol{\mu}_i = \beta\boldsymbol{\mu}_{i-1} + (1-\beta)\boldsymbol{\gamma}_i \ , \ \boldsymbol{\gamma}_i = \begin{cases} \boldsymbol{g}_i & \text{if } \boldsymbol{\mu}_i = \boldsymbol{m}_i \\ \boldsymbol{g}_i^2 & \text{if } \boldsymbol{\mu}_i = \boldsymbol{v}_i \end{cases} . \tag{2.48}$$

Here, β denotes the exponential decay rates used to compute the moving averages, and was set to its default values[146] of 0.9 (when $\boldsymbol{\mu}_i = \boldsymbol{m}_i$) and 0.999 (when $\boldsymbol{\mu}_i = \boldsymbol{v}_i$) in this study. To account for the bias towards zero that can occur from the initialisation of the moving averages, $\boldsymbol{m}_i$ and $\boldsymbol{v}_i$ are corrected by,

$$\hat{\boldsymbol{\mu}}_i = \frac{\boldsymbol{\mu}}{1-\beta^i} \tag{2.49}$$

where β^i denotes β raised to the ith power.

$\boldsymbol{\theta}$ are updated by,

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} - \alpha\left(\frac{\hat{\boldsymbol{m}}_i}{\sqrt{\hat{\boldsymbol{v}}_i} + \varepsilon} + w\boldsymbol{\theta}_i\right) \tag{2.50}$$

where α is the adaptive learning rate, w is the rate of weight decay, and ε, which was set to its default value[146] of 10^{-8} in this study, is for numerical stability. α controls the step size in gradient descent and its initial value was optimised during hyperparameter tuning in this study (*vide infra*). α also decreases with increasing iterations, allowing faster convergence away from the minimum and more precise convergence nearer to the minimum. The $\alpha w\boldsymbol{\theta}_i$ term in Eq. 2.50 denotes the weight decay, which helps prevent overfitting by penalising large weights.[144] The default value[146] of $w = 0$ (no weight decay) was employed in this study.

2.2.2.3 Hyperparameter tuning with Bayesian optimisation

In this study, the initial learning rate (LR) and hidden dropout probability (D) hyperparameters of the BERT model were optimised. This is standard practice, as BERT is typically most sensitive to these two hyperparameters.[61] The hyperparameter set (LR, D) was optimised using Bayesian optimisation,[147,148] an efficient alternative to Grid Search that incorporates knowledge from previous optimisation trials using Bayesian statistics.

Bayes' theorem,[149]

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.51}$$

describes the conditional probability of A given B based on individual probabilities $P(A)$ and $P(B)$ and the conditional probability of B given A, $P(B|A)$. Here, $P(A|B)$, $P(A)$, $P(B)$, and $P(B|A)$ are referred to as the *posterior*, *prior*, *marginal*, and *likelihood*, respectively. Bayesian optimisation applies Bayes' theorem to find the probability distribution over possible *objective functions*, $f(\boldsymbol{x})$, that map hyperparameter sets $x_i = (LR, D)_i$ to the corresponding BERT MSE. The probability distribution is estimated by fitting a *surrogate model* such as a Gaussian Process Regressor (GPR),[150,151] as utilised in this study, to a training set, $T = \{(x_i, y_i)\}_{i=1}^{n}$, consisting of n hyperparameter-MSE pairs from previous trials (where n is the total number of trials thus far). Using the mean objective function from the probability distribution, GPR predicts the BERT MSE for a query set, $Q = \{x_j^*\}_{i=j}^{q}$, consisting of q randomly generated hyperparameter sets, $x_j^* = (LR \in B_{LR}, D \in B_D)_j$, where the default values of $q = 1000$,[152] $B_{LR} = (10^{-6}, 10^{-4})$,[60] and $B_D = (0.05, 0.8)$[60] were applied in this study. The next candidate hyperparameter set, x_{i+1}, is selected from Q by optimising an acquisition function such as Expected Improvement (EI), or an adaptation thereof.[147] BERT

is then trained using x_{i+1} and the resulting $(x_{i+1}, \mathrm{MSE}_{i+1})$ is added to T for the next trial. In total, 20 trials were conducted in this study. A detailed discussion of GPR and the acquisition function are provided below.

GPR. Gaussian Process Regression (GPR)[150,151] models the prior over possible objective functions as a multivariate normal (MVN) distribution,[151]

$$P(\boldsymbol{f}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{\mu}, \boldsymbol{K}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{K}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{\mu})^\top \boldsymbol{K}^{-1}(\boldsymbol{f}-\boldsymbol{\mu})} \quad . \tag{2.52}$$

Here, $\boldsymbol{f} = (f(x_i))_{i=1}^n$ are the objective function values evaluated at $\boldsymbol{x}$, $\boldsymbol{\mu} = (m(x_i))_{i=1}^n$ are the mean objective function values evaluated at $\boldsymbol{x}$, $\boldsymbol{K}$ is the covariance matrix with elements $k_{ij} = k(x_i, x_j)$ where k is the *kernel function*,[153,154] n is the dimension of $\boldsymbol{f}$, and $|\cdot|$ denotes the determinant of $\cdot$. Prior to any observations, $\boldsymbol{\mu}$ defaults to 0, assuming the data is standardised to a zero mean, and $P(\boldsymbol{f}|\boldsymbol{x})$ is defined entirely by $\boldsymbol{K}$. The nature of the kernel function k is described at the end of this section.

The posterior distribution over possible objective functions $\boldsymbol{f}^*$, evaluated at query points $\boldsymbol{x}^*$, is given by the conditional probability,[151]

$$P(\boldsymbol{f}^*|\boldsymbol{f}, \boldsymbol{x}, \boldsymbol{x}^*) = \mathcal{N}(\boldsymbol{K}^{*\top}\boldsymbol{K}^{-1}\boldsymbol{f}, \boldsymbol{K}^{**} - \boldsymbol{K}^{*\top}\boldsymbol{K}^{-1}\boldsymbol{K}^*) \tag{2.53}$$

where elements k_{ij}^* of $\boldsymbol{K}^*$, and k_{ij}^{**} of $\boldsymbol{K}^{**}$, correspond to $k(x_i, x_j^*)$ and $k(x_i^*, x_j^*)$, respectively. $P(\boldsymbol{f}^*|\boldsymbol{f}, \boldsymbol{x}, \boldsymbol{x}^*)$ can be derived from the joint probability $P(\boldsymbol{f}, \boldsymbol{f}^*|\boldsymbol{x}, \boldsymbol{x}^*)$ using the *Marginal and conditional distributions of MVN* theorem,[155] however, the detailed derivation is beyond the scope of this Thesis.

In reality, observations $\boldsymbol{y}$ have added noise compared to the true function values $\boldsymbol{f}$, which is assumed to be normally distributed with a zero mean and variance of

σ_n^2 (= 10^7 in this study). To account for this noise, a factor of $\sigma_n^2 \boldsymbol{I}$ (where $\boldsymbol{I}$ is an identity matrix of the same size as $\boldsymbol{K}$) is added to $\boldsymbol{K}$, giving,[151]

$$P(\boldsymbol{f}^*|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{x}^*) = \mathcal{N}(\boldsymbol{K}^{*\top}[\boldsymbol{K} + \sigma_n^2 \boldsymbol{I}]^{-1}\boldsymbol{y}, \boldsymbol{K}^{**} - \boldsymbol{K}^{*\top}[\boldsymbol{K} + \sigma_n^2 \boldsymbol{I}]^{-1}\boldsymbol{K}^*) \quad . \quad (2.54)$$

The posterior mean and covariance are returned as the GPR prediction and uncertainty, respectively.

The kernel function. The GPR kernel captures data covariance, enabling the generation of smooth prior functions by ensuring that similar inputs yield similar outputs. The similarity between two inputs can be measured by their dot product, and more complex relationships can be captured when computing the dot product in a higher-dimensional feature space. However, this approach can be computationally expensive. For efficiency, the kernel can achieve the same result as computing the dot product in higher-dimensional space without the need for explicit calculation. This is known as the *kernel trick*.[153,154]

The kernel employed in this study,

$$k_{\nu=\frac{5}{2}}(x_i, x_j) = \left(1 + \frac{\sqrt{5}}{l} d(x_i, x_j) + \frac{5}{3l} d(x_i, x_j)^2\right) e^{-\frac{\sqrt{5}}{l} d(x_i, x_j)} \quad (2.55)$$

is a simplified form of the Matérn kernel[156,157] with the ν parameter, whose magnitude is positively correlated with the smoothness of $\boldsymbol{f}$, set to 2.5. Here, $d(x_i, x_j)$ denotes the Euclidean distance between x_i and x_j, and l is the length scale of the kernel, which determines how quickly the correlation between x_i and x_j drops as $d(x_i, x_j)$ decreases. $l > 1$ results in a GPR that underfits to the data and $l << 1$ results in a GPR that overfits.[156]

l is optimised during fitting of the GPR to increase the log marginal likelihood,[158]

$$\log P(\boldsymbol{y}|\boldsymbol{x}, l) = -\frac{1}{2}\boldsymbol{y}^\top [\boldsymbol{K} + \sigma_n^2 \boldsymbol{I}]^{-1}\boldsymbol{y} - \frac{1}{2}\log \left|\boldsymbol{K} + \sigma_n^2 \boldsymbol{I}\right| - \frac{n}{2}\log 2\pi \tag{2.56}$$

where n is the number of observations in $\boldsymbol{y}$, and $|\cdot|$ denotes the determinant of $\cdot$. The marginal likelihood (also referred to as the *evidence*) describes the probability of the model returning $\boldsymbol{y}$, integrated over the function space. The logarithm is taken to simplify its differentiation. The first term in Eq. 2.56 describes the data fit of the model, which increases with decreasing l, the second term is a complexity penalty, which increases with decreasing l, and the final term is a normalisation constant.

Acquisition function. The acquisition function employed in this study is the batched Noisy Expected Improvement (qNEI),[147]

$$\text{qNEI}(Q) = \frac{1}{N}\sum_{m=1}^{N}\left[\max_{j:q} \xi^m(x_j^*) - \max_{i:n} \xi^m(x_i)\right]^+ \tag{2.57}$$

given here for a query set $Q = \{x_j^*\}_{j=1}^q$ and previously evaluated set $T = \{x_i\}_{i=1}^n$. The $\xi^m(x_j^*)$ and $\xi^m(x_i)$ terms represent a Monte Carlo (MC) sample, ξ^m, drawn from the joint posterior over Q and T, evaluated at query point x_j^* and observed point x_i, respectively. N is the total number of MC samples (= 512 in this study), and $[\cdot]^+$ denotes $\max(\cdot, 0)$. Note that the goal of Bayesian optimisation is to minimise the BERT MSE, so here -MSE is maximised.

2.3 Interpreting machine-learned predictions

2.3.1 Kuz'min prediction contributions

Kuz'min prediction contributions,[67] are calculated per data point as,

$$I_i = \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{p\in F}\overline{\log k_{c\in D}^{\text{train}}} - \overline{\log k_{p}^{\text{train}}}\right] \tag{2.58}$$

where I_i is the importance of feature i, D is the set of nodes corresponding to the queried decision path, F is the subset of D corresponding to a decision made on feature i, $\overline{\log k_{p}^{\text{train}}}$ and $\overline{\log k_{c\in D}^{\text{train}}}$ respectively denote the mean $\log k_{\text{exp}}$ of training reactions at parent node p, and at the resulting child node $c \in D$, and t and T respectively denote each tree, and the total number of trees in the RF (T = 500 in this study). The magnitude and sign of the prediction contribution quantify the change in $\log k_{\text{pred}}$ as a result of the feature relative to predicting the mean $\log k$ of the training data ($I > 0$ = increasing $\log k$, $I < 0$ = decreasing $\log k$).

2.3.2 IGs

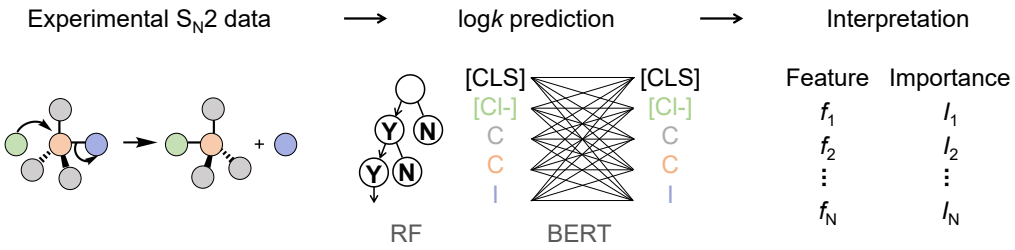
Integrated Gradients (IGs)[65,73] for each token are computed per data point as,

$$\text{IG}_{\text{token}} = \sum_{i=1}^{d_{\text{f}}}(\varepsilon_i - \varepsilon_i')\left[\sum_{k=1}^{m}\frac{\partial f(\boldsymbol{E}' + \frac{k}{m}(\boldsymbol{E} - \boldsymbol{E}'))}{\partial \varepsilon_i}\frac{1}{m}\right] \tag{2.59}$$

where $f(\boldsymbol{E})$ is the model output with respect to the data point embedding, $\boldsymbol{E}'$ denotes a baseline embedding of "." tokens with the same shape as $\boldsymbol{E}$ ($d_{\text{t}} \times d_{\text{f}}$ = 300 × 256 in this study), ε_i and ε_i' respectively correspond to feature i of the token and baseline embeddings, and k corresponds to one step of the m step linear interpolation between $\boldsymbol{E}'$ and $\boldsymbol{E}$ (m = 500 in this study). The gradients quantify the influence a feature has on the model output, while the sum of local gradients over the linear interpolation accounts for features that may have flat gradients in certain regions of space. The $(\varepsilon_i - \varepsilon_i')$ term in Eq. 2.59 is a scaling term to ensure the IGs are on the same scale as $\boldsymbol{E}$. The IGs are summed over features to get a single value per token; the magnitude and the sign of the IG

quantify the influence the token has on prediction relative to the baseline "." token (IG $>$ 0 = increasing $\log k$, IG $<$ 0 = decreasing $\log k$).

3 | Interpretable predictions of S_N2 kinetics using the BERT and RF architectures: Comparison to known reactivity rules



Abstract

The accurate prediction of reaction rates is an integral step in reaction mechanism elucidation and design of synthetic pathways. Traditionally, kinetic parameters have been derived from activation energies obtained from Quantum Mechanics (QM) methods, and more recently, Machine Learning (ML). Transformer-based models have emerged as a popular ML method for chemical prediction, allowing reactions to be represented as SMILES strings. Despite being state-of-the-art in both reaction classification and yield prediction, Bidirectional Encoder Representations from Transformers (BERT), a type of transformer-based model, have not yet been applied to kinetic prediction. In this work, a BERT model is trained to predict experimental log k values of S_N2 reactions, and its performance is compared to the state-of-the-art Random Forest (RF) literature model in terms of accuracy, training time, and ability to replicate known reactivity rules. Both BERT and RF exhibit near-experimental accuracy (RMSE: ≈ 1.1 log k) on similarity-split test data. Additionally, BERT and RF's predictions are interpreted, revealing that both models identify key reaction centre atoms, as well as known electronic, and steric effects. Moreover, each model is found to overcome the other's limitations, such as target value extrapolation and recognition of aromatic effects.

3.1 Methodology

3.1.1 Data

The training and test data utilised in this work can be visualised in Figure 3.1. The training data contains 4862 reactions with 449 unique nucleophiles, 298 unique substrates, 155 unique solvent systems, has a temperature range of 232.15 to 425.15 K, and a log k range of -12.3 to 1.6. The test data contains 129 reactions with 41 unique nucleophiles, 43 unique substrates, 10 unique solvent systems, has a temperature range of 252.15 to 461.00 K, and a log k range of -8.2 to 1.2. All test data has an $S_T < 0.4$ to the training data, so prediction accuracy reflects model performance on novel reactions (Figure 3.2**a**). Here, reaction A is said to have an $S_T > X$ to reaction B if the nucleophile and substrate of A respectively had an $S_T > X$ to the nucleophile and substrate of B, else reaction A is said to have an $S_T < X$ to reaction B. Note that 0.4 is a standard S_T threshold[65] and effectively imposed a similarity constraint without a significant reduction in test set size.

The training data utilised in this work builds upon the S_N2 data compiled by Madzhidov et al.[29] which they used to train their rate prediction RF model.[33] The original dataset consists of 4830 S_N2 reactions and their experimental log k values, which were cleaned in the current work to remove unbalanced reactions (109 reactions), duplicates (46 reactions), and known CV outliers[29] (9 reactions), resulting in 4666 reactions from Ref[33]. The SMILES used to generate the ISIDA fragments were also canonicalised in the current work to improve interpretability (i.e., so each molecular fragment is only represented by one ISIDA fragment). The chemical diversity in the training data was increased by including 196 S_N2 reactions manually curated in the current work.[159–170] Specifically, 83% of the reactions curated in this work have an $S_T < 0.4$ to the reactions from Ref[33], and therefore add structural diversity (Figure 3.2**b**). Furthermore, the reactions curated

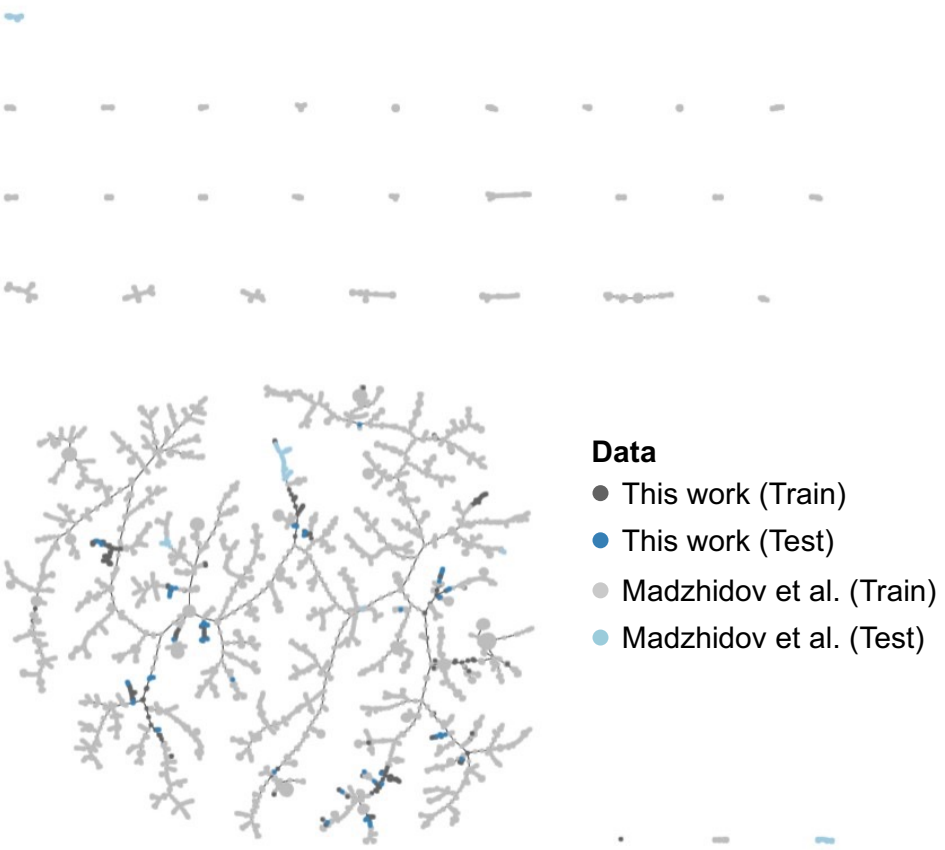


Figure 3.1: Tree map (TMAP) of the training and test data utilised in this work. Shown in grey: The total training set of 4862 S_N2 reactions. 4666 of these were compiled by Madzhidov et al.[29] (light grey), and 196 were added in the current work to increase the chemical diversity in the training data (dark grey). Shown in blue: The total test set of 129 S_N2 reactions. 73 of these were compiled by Madzhidov et al.[29] (light blue), and 56 were added in the current work to represent a broader area of chemical space (dark blue). Interactive versions of each TMAP are provided on GitHub (see *Data availability*).

in this work introduce an additional nucleophile type into the data: phosphines (36 reactions), as well as an additional LG: azide (4 reactions), and an additional solvent type: electrolyte solutions (16 reactions). The $\log k$ range was also increased from -7.7 to 1.6 to -12.3 to 1.6.

The test data utilised in this work is comprised of 73 test reactions compiled by

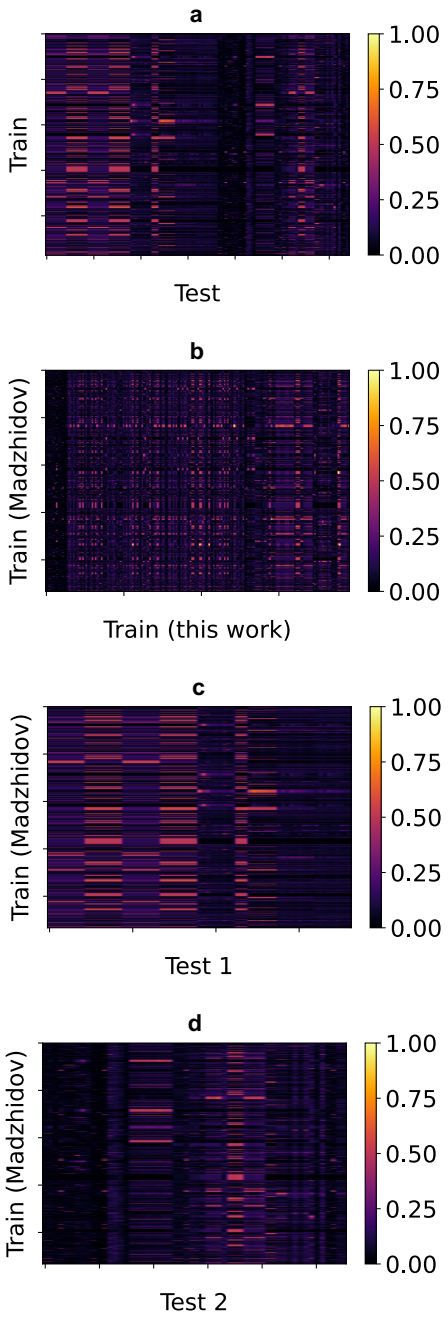


Figure 3.2: Tanimoto similarity maps. a. Tanimoto similarity of the 4862 reactions used to train the RF and BERT models in this work to the total 129 test reactions. **b to d.** Tanimoto similarity of the 4666 training reactions compiled by Ref[29] to: the 196 training reactions manually curated in the current work (**b**), the 73 test reactions compiled by Ref[29] with $S_T < 0.4$ to the training data (Test 1, **c**), and the 56 test reactions manually curated in the current work (Test 2, **d**).

Madzhidov et al.[29] which they used to evaluate their rate prediction RF model[33] (those with $S_T < 0.4$ to the training data, Test 1), and 56 reactions manually curated in the current work (Test 2).[159,161–169] The latter represent an area of chemical space outside the applicability domain of Ref[33]'s training data. Firstly, Test 2 has a low chemical similarity to Ref[33]'s training data, in comparison to Test 1. This was quantified by the percentage of reactions with $S_T < 0.2$ to Ref[33]'s training data: 46% and 12% for Test 2 and Test 1, respectively (Figures 3.2**c** and **d**). A threshold of $S_T < 0.2$ is used here as all test data has an $S_T < 0.4$ to the training data. Secondly, Test 2 contains species not included in Ref[33]'s training data: 12 reactions in electrolyte, 5 reactions with azide LGs, and 4 reactions with phosphine nucleophiles.

The training and test data curated in the current work was compiled using a manual literature search. Data was extracted from 11 papers dating from 1960 to 2006[159–170] which were manually inspected to ensure that the reported rates corresponded to the S_N2 mechanism and excluded reactions outside the intended scope (e.g., micellar or intramolecular reactions). Tabulated data was extracted using Tabula[171] and manually refined to correct artifacts introduced by the extraction process (e.g., omitted decimal points). Textual data (e.g., names of reagents captured outside the table) was manually extracted, while pictorial data (e.g., reactants represented as chemical structures) was replicated using ChemDraw.[172] Reported $\Delta G^{\ddagger}$ values were converted to J mol^{-1} and further converted to k using the Eyring equation. k values reported in 10^x mol^{-1} dm^3 s^{-1} were converted to mol^{-1} dm^3 s^{-1}, temperatures were converted to K, and solvent ratios were converted to mole fractions. Reactions were converted to SMILES by first drawing them in ChemDraw,[172] and then exporting them as SMILES. Training and test sets were created using an initial random 50/50 shuffle

split, followed by removal of test reactions with $S_T > 0.4$ to the training data, and their addition into the training set.

Throughout this chapter, high-dimensional data is visualised using tree maps (TMAPs). TMAPs were generated using the tmap visualisation library (version 1.0.6)[173] and visualised using Faerun (version 0.3.20).[174] Reactions were encoded as MinHash6 fingerprints[175] with the number of permutations set to 512.

3.1.2 ML and QM

3.1.2.1 RF

This work utilises the transformation-out RF model developed by Madzhidov et al.[33] This model is trained using a 5-fold transformation-out CV scheme, where reactions involving the same chemical transformation are placed in different CV folds. Chemical reactions are represented by ISIDA fragments, reciprocal temperature (T^{-1}), and 14 solvent descriptors, including the mole fraction of the primary solvent component, x_1, and 13 properties[a] related to polarity and proticity.

Madzhidov's model was adapted in the current work to explicitly describe the secondary solvent component (using the same properties as for the primary component), as well as solvent ionic strength. This is to account for non-aqueous solvent mixtures and electrolyte solutions in training reactions added in the current work. Additionally, model performance in the current work was calculated by taking the mean RMSE over the five CV folds, allowing the standard error of the mean (σ_M) to be used as an uncertainty estimate. All other features and settings were kept consistent with Ref[33].

[a] Solvent properties in Ref[33]'s RF model: Catalán SA,[176] SB[177] and SPP[178] constants; Kamlet-Taft α,[179] β[180] and π^*[181] constants; 4 functions of dielectric constant (ε): $\frac{(\varepsilon-1)}{\varepsilon}$, $\frac{(\varepsilon-1)}{(2\varepsilon+1)}$, $\frac{(\varepsilon-1)}{(\varepsilon+1)}$, $\frac{(\varepsilon-1)}{(\varepsilon+2)}$; 2 functions of refractive index (n): $\frac{(n^2-1)}{(n^2+2)}$, $\frac{(n^2-1)}{(2n^2+1)}$; and the dielectric constant/refractive index function $\frac{(n^2-1)(\varepsilon-1)}{(2n^2+1)(2\varepsilon-1)}$.

3.1.2.2 BERT

This work also utilises the BERT model developed by Laino et al.,[60] which they pretrained on an MLML task and subsequently fine-tuned for reaction classification. In the current work, Laino's model was further fine-tuned for rate prediction. Five estimators were trained in the current work, corresponding to the five CV folds used in the RF model. For each estimator, the LR and D hyperparameters were optimised between their usual bounds of 10^{-6} and 10^{-4}, and 0.05 and 0.8 (respectively) using Bayesian optimisation implemented in BoTorch (version 0.2.1),[147] interfaced to Ax (version 0.1.9).[182] Only LR and D were optimised during fine-tuning as BERT is typically most sensitive to these two hyperparameters.[61] The number of training epochs was fixed at 10 (consistent with Ref[60]) as this allowed the validation RMSE to converge without leading to overfitting (Section 3.2.1). In contrast to Laino's model, the SMILES input was adapted in this work to include quantitative parameters of the reaction conditions, namely reciprocal temperature, ionic strength, and solvent mole fractions. These parameters were included in particular as the reaction rate depends on them, yet they cannot be implicitly derived from the vanilla reaction SMILES. To incorporate these parameters, their numerical value was appended to the vanilla SMILES using 0-9 and '.' tokens, prefixed by [RecipTemp], [IonStr], [Solv1R], and [Solv2R] tokens, respectively (Figure 3.3). To facilitate the adapted SMILES, the [RecipTemp], [IonStr], [Solv1R], [Solv2R], and 0 tokens were also added to the model tokeniser.

3.1.2.3 QM

QM calculations were carried out with María Calvo, a summer student in the Duarte group (June to September 2021). Calculations using autodE were performed by M. Calvo, while calculations without autodE and frequency shift calculations were performed by myself.

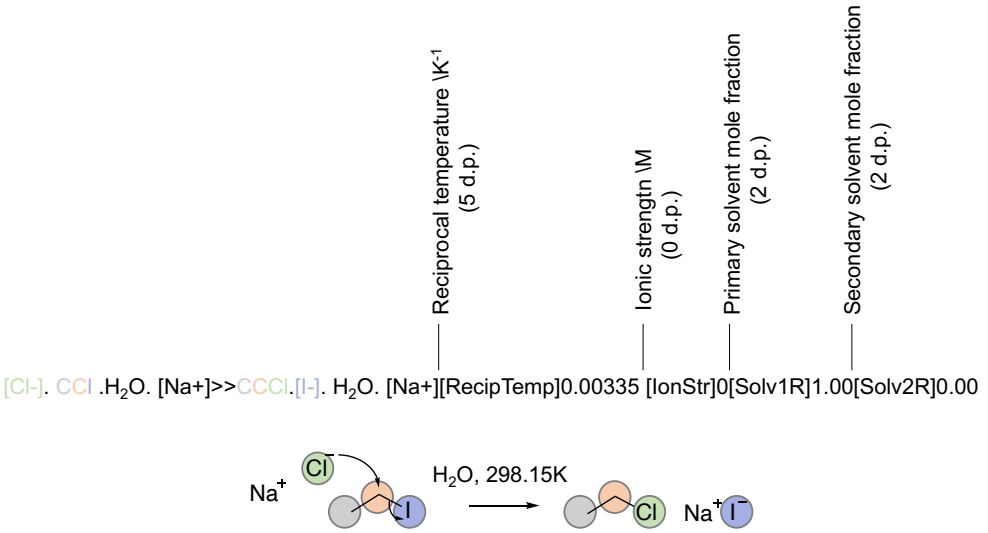


Figure 3.3: Grammar used to construct SMILES in the BERT model demonstrated using the reaction of sodium chloride with ethyl iodide in water at 298.15 K. Molecules follow the order: nucleophile, substrate, solvent, counter-ions on the reactants side, and: major product, bi-product, solvent, counter-ions on the products side. Reciprocal temperature, ionic strength, and solvent mole fractions are appended to the end of the SMILES in the manner depicted.

QM calculations were carried out using the ORCA suite of programs (version 4.2.1),[183] interfaced to autodE (version 1.1)[184] on 30 of the S_N2 reactions manually curated in the current work with a single solvent component and total reactant molecular weight (MW) $\leq$ 235.9 Da (Table E.1, Appendix E.1). Low MW reactions in single-component solvents were chosen owing to their reasonable computational cost and compatibility with the CPCM implicit solvent model.

Geometry optimisations and frequency calculations were carried out at the CPCM(solvent) PBE0-D3BJ/def2-SVP LOT and an approximation to the CPCM(solvent) CCSD(T)/def2-TZVP free energies were made from the PBE0 free energy values using,

$$G_{CCSD(T)} \approx E_{CCSD(T)} + G_{PBE0} - E_{PBE0} \tag{3.1}$$

where E is the single point energy. Default convergence criteria were used corresponding to SCF energy change tolerances of 10^{-8} Eh for geometry optimisations (including TS optimisation + frequency calculations) and 10^{-6} Eh for frequency calculations and single point energy calculations, and an energy change of 5×10^{-6} Eh for the geometry optimisation step. Numerical integration was carried out using a 110-point Lebedev grid, with an integration accuracy of $10^{-4.34}$. Coulomb and HF exchange integrals were evaluated using the RI-J and COSX approximations, respectively, with def2/J being employed as an auxiliary basis set for the former. Free energies were calculated according to Grimme's quasi-RRHO approximation for vibrational entropy. A standard state correction from 1 atm to 1M of $RT \ln(\frac{V_{\text{atm}}}{V_{\text{M}}})$ where V_{atm} and V_{M} are the molar volumes at 1 atm and 1M respectively, was automatically added to the calculated free energies in autodE. A symmetry correction of $RT \ln(\sigma_r)$, where σ_r is the rotational symmetry number, was also added automatically. Where free energies were calculated without using autodE (see Appendix E.1), the standard state and symmetry corrections were added manually. Optimised TS structures contained exactly one imaginary frequency, with magnitudes in the expected range of ≈ 250 cm^{-1} to 450 cm^{-1}. Frequency shift calculations were carried out on the optimised reactant complexes of three reactions (see Appendix E.1) using Otherm[185] to convert their imaginary frequencies to real frequencies. All other optimised reactant complexes had no imaginary frequencies, or imaginary frequencies of magnitude < 30 cm^{-1}, which were disregarded as noise. Further methodological considerations are provided in Appendix E.1.

PBE0 is widely recommended for obtaining accurate structures of organic molecules[186–188] and calculating vibrational frequencies.[189] Therefore, it is an appropriate LOT for the geometry optimisation and frequency calculations of S_N2 reactions in the current work. A D3BJ dispersion correction was added to

account for intermolecular interactions which are likely to form between anionic nucleophiles and polar substrates.[190] PBE0 has been shown to produce reliable geometries[187] and vibrational frequencies[189] when paired with a double-zeta basis set. Therefore, def2-SVP was employed for the calculations using PBE0 in the current work. For higher-level single point energy calculations, CCSD(T), renowned for benchmark-quality electronic structure calculations,[187,191] was utilised alongside a def2-TZVP basis set.

To obtain the QM RMSE in $\log k$, the QM $\Delta G^{\ddagger}$ values (in J mol^{-1}) were converted to $\log k$ using the Eyring equation (Equation 1.1).

3.1.3 Calculation times

BERT and RF training times were defined as the time taken for RMSE convergence over the five CV folds ($t_{5,\text{conv}}$). This was estimated using,

$$t_{5,\text{conv}} \approx 5\eta_{\text{conv}}\theta_{\text{conv}}\frac{t_{\text{fold}}}{\eta_{\text{tot}}\theta_{\text{tot}}} \tag{3.2}$$

where t_{fold} is the time taken for the total number of epochs (η_{tot} = 10 in BERT. This value was set to 1 in the RF) and hyperparameter optimisation trials (θ_{tot} = 20 in BERT and 10 in the RF) in one CV fold (fold 1), and η_{conv} and θ_{conv} are the mean number (over the five CV folds) of epochs (= 9.4 in BERT, Section 3.2.1. This value was set to 1 in the RF) and hyperparameter optimisation trials (= 14.6 in BERT and 2.8 in the RF, Section 3.2.1) taken to achieve convergence. The resulting training times were 256 s for the RF and 190,305 s (52.9 hrs) for BERT on 8 vCPUs of Intel® x86-64 CPU (64 GiB RAM). The BERT training time was 13,311 s (3.7 hrs) on GPU (1 NVIDIA V100 PCIe 16 GB GPU), which is incompatible with the Scikit-learn implementation of RF. Even on GPU, BERT is significantly slower than RF.

BERT and RF prediction times were calculated for five random samples of 30 test reactions (one sample per CV fold), and averaged. QM times refer to the time taken for geometry optimisation and frequency calculations of 30 reactant complexes and their corresponding TSs. The averaging over samples for the ML prediction time is to account for the fact that different samples were used to calculate the ML and QM prediction times, due to some of the lower-MW reactions required for QM calculations failing to meet the $S_T < 0.4$ requirement for the ML test data.

3.1.4 Evaluating model performance

3.1.4.1 Accurate and inaccurate predictions

In this work, *accurate* and *inaccurate* predictions are defined as those with the 25% smallest and 25% largest prediction errors (averaged over the five CV folds), respectively. Using the upper and lower quartiles as thresholds is a standard statistical approach that leverages the data's natural distribution, avoiding arbitrary cut-offs.

3.1.4.2 Uncertainty estimation

Predictions and feature importances in this work are quoted as the mean over the five CV folds, with σ_M providing an uncertainty estimate. σ_M was calculated using,

$$\sigma_M = \frac{\sigma}{\sqrt{n}} \tag{3.3}$$

where σ is the sample standard deviation and n is the sample size (= 5 estimators). In the BERT model, the importances of solvents and substituted electrophilic centres were calculated as the sum of importances over solvent tokens or substituent tokens, respectively. In this case, the associated uncertainties were propagated using,

$$\sigma_{M,\text{prop}} = \sqrt{\sum_{i}^{N} \sigma_{M,i}^{2}} \tag{3.4}$$

where $\sigma_{M,\text{prop}}$ is the propagated uncertainty and N is the total number of uncertainties being propagated over. Eq. 3.4 assumes that the uncertainties are uncorrelated. To verify that the uncertainties being propagated in the current work were uncorrelated, Algorithm 1 was used to calculate the mean Spearman's r value ($\overline{r_S}$) over 50 iterations between randomly sampled pairs of uncertainties for solvent tokens and substituting C tokens. Algorithm 1 returned $\overline{r_S}$ values of 0.24 ± 0.02 and 0.23 ± 0.02 with mean Spearman's p values, $\overline{p_S}$, of 0.65 ± 0.02 and 0.19 ± 0.03 for solvent tokens and substituting C tokens, respectively, verifying that the uncertainties are uncorrelated.

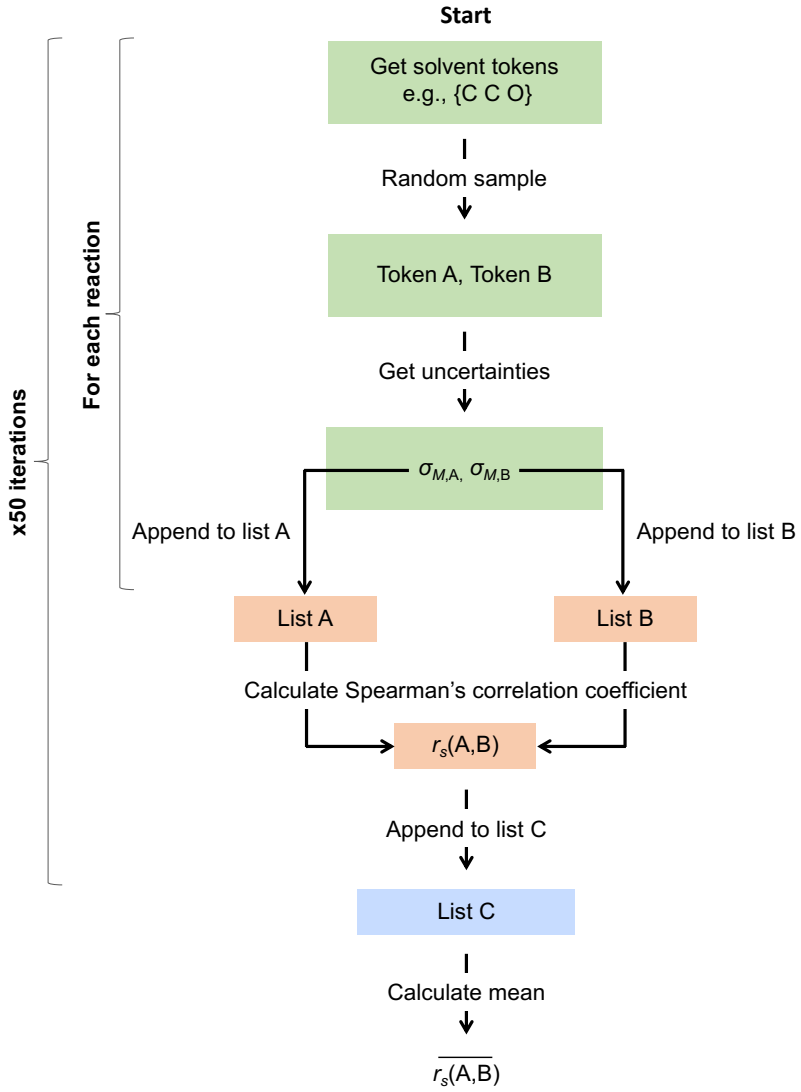
3.1.4.3 Statistical significance

To assess whether the difference in RMSE between the RF and BERT models was statistically significant, a Wilcoxon's matched-pairs signed-rank test[192] was employed. This is the most suitable test for this comparison, where the distributions are numerical, non-parametric, and paired (i.e., the models are trained on the same dataset).[193] The Wilcoxon's p-value was calculated using a permutation test (number of resamples = 9999) to account for the small number of samples (= 5 CV folds) in the distribution. Zeros were handled using the Pratt method,[194] which includes zeros in the ranking procedure but excludes them from the test statistic.

3.1.4.4 RMSE convergence

In this work, the RMSE was said to have converged with respect to a model variable x (where x represents % training data, epoch number, or hyperparameter optimisation trial) when $-\frac{\partial RMSE}{\partial x} < 0.01$.

Algorithm 1 Calculation of the mean Spearman's r value ($\overline{r_S}$) over 50 iterations between randomly sampled pairs of uncertainties associated with the feature importances of SMILES tokens (solvent tokens are used here as an example). Token samples were taken over all accurate and inaccurate predictions where Eq. 3.4 was used to calculate the uncertainty in the feature importance.



Should the RMSE have not converged within the total number of training reactions, epochs, or hyperparameter optimisation trials (denoted x_{tot}), the expected RMSE improvement from training until convergence, ξ, was calculated using,

$$\xi = \frac{\sum_{i=1}^{N} \text{RMSE}_{x_{\text{tot}},i}}{N} - \frac{\sum_{i\in\{c\}} \text{RMSE}_{x_{\text{tot}},i} + \sum_{i\in\{n\}} \text{RMSE}_{\text{conv},i}}{N} \tag{3.5}$$

where $\text{RMSE}_{x_{\text{tot}},i}$ is the RMSE for CV fold i when training with x_{tot}, $\text{RMSE}_{\text{conv},i}$ is the estimated RMSE for fold i when training until convergence, $\{c\}$ is the set of folds where the RMSE had converged within x_{tot}, $\{n\}$ is the set of folds where the RMSE had not converged within x_{tot}, and N is the total number of CV folds (= 5). Here, $\text{RMSE}_{\text{conv},i}$ was estimated by plotting the RMSE against x, extrapolating the graph to $2x_{\text{tot}}$, and identifying the point where $-\frac{\partial RMSE}{\partial x} < 0.01$.

Note that the $-\frac{\partial RMSE}{\partial x} < 0.01$ convergence criterion is not applicable to hyperparameter optimisation using the Grid Search method, employed for the RF model in this work. Unlike Bayesian hyperparameter optimisation (used for the BERT model), where RMSE typically converges with increasing trial number, Grid Search can identify the optimal hyperparameters (i.e., the lowest RMSE) in any single trial. To ensure that the optimal hyperparameters were found among the 10 trials conducted, the trials were ranked from highest RMSE (trial 1) to lowest RMSE (trial 10). If the lowest RMSE appeared in at least three trials, it was assumed to be the global minimum, indicating that the optimal hyperparameters had been identified.

3.1.5 Interpretation

3.1.5.1 Calculating feature importances

Feature importances were calculated using the state-of-the-art method for each model: Kuz'min prediction contributions[67] for the RF and IGs[65,73] for BERT.

Prediction contributions were calculated using treeinterpreter (version 0.2.3),[72] and IGs using LayerIntegratedGradients from Captum (version 0.4.0).[195] The magnitude of a feature's importance corresponds to the feature's influence on $\log k_{\text{pred}}$, while the sign corresponds to whether the feature increased (positive) or decreased (negative) $\log k_{\text{pred}}$.

3.1.5.2 High impact features

In this work, *high impact* features are defined as those with the 25% largest importances (averaged over the 5 CV folds) across the test data, corresponding to a prediction contribution with magnitude ≥ 0.001 in the RF, and IG with magnitude ≥ 0.03 in BERT. Using the upper quartile as a threshold is a standard statistical approach that leverages the data's natural distribution, avoiding arbitrary cut-offs. Features that were not in the upper quartile are referred to as *low impact*. In the RF, reaction centre atoms (Nu, LG, electrophilic C) were considered important if any ISIDA fragment with a substructure match to the atom was high impact, while temperature and solvent properties were said to be important if their numerical value was high impact. Meanwhile in the BERT model, reaction centre atoms were said to be important if the SMILES token representing the atom (or any one of the tokens representing the atom for symmetrical molecules with equivalent Nu, LG, or electrophilic C atoms) was high impact, while temperature was considered important if the [RecipTemp] token or any token corresponding to significant figures of the reciprocal temperature was high impact. Regarding solvent, this was considered a high impact feature of the BERT model if $\geq 50\%$ of solvent tokens on the reactants side were high impact. A threshold of $\geq 50\%$ was employed (as opposed to 100%) to account for the fact that some solvent moieties (such as hydrogen bond donors) are expected to be more critical for reactivity prediction than others. 50% is a standard classification threshold, representing a natural midpoint where majority influence defines a decision boundary.

Reaction centre atoms were identified using autodE (version 1.1).[184] For delocalised nucleophiles with resonant Nu atoms (e.g., nitrite), the Nu atom was assigned to the atom bearing the localised negative charge in the SMILES.

It is noted that each reaction centre atom also has a mapped atom in the products. The importance of these mappings is discussed in Appendix E.4.

3.1.5.3 Analysing structural and physical effects

To assess whether RF and BERT effectively learned key structural and physical effects, this work evaluates whether high impact features align with known reactivity rules including LG, nucleophilic, steric, allylic, temperature, and solvent effects. To analyse structural and physical effects, the feature importances (positive sign = increase $\log k_{\text{pred}}$, or negative sign = decrease $\log k_{\text{pred}}$) of key reaction centre atoms (and bonds) were evaluated. Here, features with an importance of zero within the associated error were categorised as low impact. The importance of each feature is relative to that of a dummy model that predicts the mean $\log k_{\text{exp}}$ of the training data for each test reaction. By definition, all features of the dummy model have an importance of zero, however this is also shown mathematically in Section 3.1.6.

To analyse LG effects, reactions with I, Br, Cl, and F LGs were considered, represented by C-I, C-Br, C-Cl, and C-F fragments in the RF and I, Br, Cl, and F tokens in BERT. Similarly, to analyse nucleophilic effects, reactions with I^-, Br^-, Cl^-, and F^- nucleophiles were considered, represented by C-I, C-Br, C-Cl, and C-F fragments (in the major product) in the RF and [I-], [Br-], [Cl-], and [F-] tokens in BERT. Steric effects were analysed using reactions with alkyl-substituted electrophilic centres, modelled by C-C, C-C-C, and C-C-C-C fragments in the RF, and tokens of electrophilic C and substituting C atoms in BERT. Here, the importance of these features in reactions with unsubstituted centres were used

as a control. Meanwhile, allylic effects were assessed using reactions with alkene, alkyne, and aromatic groups bound to the electrophilic centre, which were represented by C-C=C, C-C≡C, and C-C:C fragments in the RF (where ':' is an aromatic bond), and =, ≡, and c tokens in BERT (where c is an aromatic carbon bonded to the centre). For centres with multiple substituents, the feature importances were summed over the corresponding high impact C, =, ≡, or c tokens in the BERT model (this is not relevant to the RF where features are represented by counts of molecular fragments).

When evaluating temperature effects, feature importances correspond to the importance of the reciprocal temperature feature in the RF, and the sum over importances of SMILES tokens representing temperature (the [RecipTemp] token or any token of the numerical value) in BERT. Regarding solvent effects in the BERT model (solvent effects were not evaluated for the RF, see Justification 2 in Appendix C), the standard threshold of $\varepsilon_{298.15} = 15$ (where $\varepsilon_{298.15}$ is the dielectric constant at 298.15 K) was used to define polar ($\varepsilon_{298.15} > 15$), and non-polar ($\varepsilon_{298.15} < 15$) solvents,[196] while protic and aprotic solvents were defined as those with (protic) or without (aprotic) a proton bonded to a heteroatom. The importance of the solvent was taken as the sum over importances of high impact solvent tokens. Here, the solvent was said to be low impact if none of its tokens were high impact. Note that feature importances were summed over all temperature tokens, but only high impact solvent tokens. This is because solvent effects were analysed by categorising the feature importances into positive (increase $\log k_{\text{pred}}$), negative (decrease $\log k_{\text{pred}}$), or low impact (negligible effect on $\log k_{\text{pred}}$), while temperature effects were evaluated by observing the correlation between feature importance and temperature.

In the RF, fragments containing I, Br, Cl, or F LGs, I^-, Br^-, Cl^-, or F^- nucleophiles, or alkene, alkyne, or aromatic groups that aren't mentioned above were excluded from analysis to avoid confounding effects from other atoms and bonds. Additionally, reactions where C-C, C-C-C, or C-C-C-C fragments contain the product atom mapping of a nucleophilic C^- atom were omitted from the analysis of steric effects in the RF, to avoid confounding nucleophilic effects. For the same reason, reactions where the solvent acted as a nucleophile were omitted from the analysis of solvent effects in the BERT model. Similarly, reactions containing substituting groups other than alkyl were excluded from the analysis of steric effects in both models, as were reactions containing substituting groups other than alkene/alkyne/aromatic and alkyl in the analysis of allylic effects.

It is noted that each atom has a mapped atom in the products. The feature importances of these mappings are discussed in Appendix E.4.

3.1.5.4 Well-learned, poorly-learned, and new features

In this work, *well-learned* and *poorly-learned* features were defined as features that were high impact in accurate predictions but not inaccurate predictions, and features that were high impact in inaccurate predictions but not accurate predictions, respectively. Meanwhile, *new* features were defined as features in the test data that were not in training.

3.1.6 Dummy model

Feature importances calculated in this work are relative to a dummy model which predicts the mean $\log k_{\text{exp}}$ of the training data, $\overline{\log k_{\text{exp}}^{\text{train}}}$ (= -3.4), for each test reaction. By definition, all features of the dummy model have an importance of zero, however this is also shown mathematically for each model below.

RF. Feature importances in the RF were calculated using Eq. 2.58 (repeated here for clarity),

$$I_i = \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{p\in F}\overline{\log k_{c\in D}^{\text{train}}} - \overline{\log k_{p}^{\text{train}}}\right] \tag{2.58}$$

where I_i is the importance of feature i, D is the set of nodes corresponding to the queried decision path, F is the subset of D corresponding to a decision made on feature i, $\overline{\log k_{p}^{\text{train}}}$ and $\overline{\log k_{c\in D}^{\text{train}}}$ respectively denote the mean $\log k_{\text{exp}}$ of training reactions at parent node p, and at the resulting child node $c \in D$, and t and T respectively denote each tree, and the total number of trees in the RF. In the theoretical dummy RF, $\log k_{\text{pred}} = \overline{\log k_{\text{exp}}^{\text{train}}}$, i.e., its value at the root node. Hence, the $\overline{\log k_{c\in D}^{\text{train}}} - \overline{\log k_{p}^{\text{train}}}$ term in Eq. 2.58 equals zero in the dummy RF, resulting in $I_i = 0$ for all features.

BERT. Feature importances in BERT were calculated using Eq. 2.59 (repeated here for clarity),

$$\text{IG}_{\text{token}} = \sum_{i=1}^{d_{\text{f}}}(\varepsilon_i - \varepsilon_i')\left[\sum_{k=1}^{m}\frac{\partial f(\boldsymbol{E}' + \frac{k}{m}(\boldsymbol{E}-\boldsymbol{E}'))}{\partial \varepsilon_i}\frac{1}{m}\right] \tag{2.59}$$

where $f(\boldsymbol{E})$ is the model output with respect to the data point embedding, $\boldsymbol{E}'$ denotes a baseline embedding of "." tokens with the same shape as $\boldsymbol{E}$, ε_i and ε_i' respectively correspond to feature i of the token and baseline embeddings, and k corresponds to one step of the m step linear interpolation between $\boldsymbol{E}'$ and $\boldsymbol{E}$. In the theoretical dummy BERT, $\log k_{\text{pred}} = \overline{\log k_{\text{exp}}^{\text{train}}}$ so does not vary with ε_i. Hence, the $\frac{\partial f(\boldsymbol{E}'+\frac{k}{m}(\boldsymbol{E}-\boldsymbol{E}'))}{\partial \varepsilon_i}$ term in Eq. 2.59 equals zero in the dummy BERT, resulting in IG = 0 for all tokens.

3.2 Results and Discussion

3.2.1 RMSE convergence

To evaluate whether the error had converged in both the BERT and RF models, a convergence criterion of $-\frac{\partial RMSE}{\partial x} < 0.01$ was employed, where x represents some model variable such as % training data (Figure 3.4**a**, evaluated using the total test data Test 1 + Test 2), epoch number (BERT only, Figure 3.4**b**, evaluated using validation data), or hyperparameter optimisation trial (Figure 3.4**c**, evaluated using validation data).

Both RF and BERT achieved convergence within the total number of training reactions (RF: 23% of training data ≡ 1118 reactions, BERT: 18% of training data ≡ 875 reactions), affirming that a sufficient amount of training data was used to train each model. Additionally, BERT achieved convergence within the 10 training epochs for three out of five CV folds (folds 2, 3, and 5), with $-\frac{\partial RMSE}{\partial epoch} = 0.01$ for two folds (folds 1 and 4). To test whether accuracy could be improved by training on folds 1 and 4 until convergence, the expected RMSE improvement, ξ, was calculated using Eq. 3.5. Eq. 3.5 returned an ξ value of 0.0 log k, indicating that accuracy would not be improved by training until convergence. Consequently, the estimators trained for 10 epochs were employed in this work.

Regarding hyperparameter optimisation, this was conducted using Bayesian optimisation for the BERT model, and Grid Search for the RF. For the BERT model, the RMSE converged within the 20 Bayesian optimisation trials conducted for each CV fold, affirming that the optimal hyperparameters (Table 3.1) had been identified. Unlike Bayesian optimisation, where error typically converges with increasing trial number, Grid Search (employed for the RF model) can identify the optimal hyperparameters (i.e., the lowest RMSE) in any single trial. Hence, the

Hyperparameter	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
LR	7×10^{-5}	1×10^{-4}	1×10^{-4}	8×10^{-5}	1×10^{-4}
D	0.05	0.13	0.05	0.05	0.05

Table 3.1: Optimised hyperparameter values in the BERT model.

$-\frac{\partial RMSE}{\partial x} < 0.01$ convergence criterion is not applicable here. Instead, to assess whether the optimal hyperparameters were found among the 10 Grid Search trials conducted, the trials were ranked from highest RMSE (trial 1) to lowest RMSE (trial 10). If the lowest RMSE appeared in at least three trials, it was assumed to be the global minimum, indicating that the optimal hyperparameters had been identified. Indeed, the lowest RMSE appeared in ≥ 7 trials in every CV fold, affirming that the optimal value (max_features = 0.35 for each fold) had been identified.

3.2.2 Improving generalisability of the RF

The dataset utilised in this work builds upon the S_N2 data compiled by Madzhidov et al.[29] which they used to train their rate prediction RF model.[33] Here, the chemical diversity of Madzhidov's data has been augmented by including an additional 196 S_N2 reactions and their experimental $\log k$ values, 83% of which have an $S_T < 0.4$ to the 4666 reactions from Ref[33]. The reactions curated in this work also introduce an additional nucleophile type into the data: phosphines (36 reactions), as well as an additional LG: azide (4 reactions), and an additional solvent type: electrolyte solution (16 reactions).

Notably, diversifying the training data did not impact model performance when evaluated on the test data from Ref[33] (referred to here as Test 1 ≡ 73 reactions), maintaining an RMSE of 1.0 ± 0.0 $\log k$. The most significant improvement of the RF model trained on diversified data was seen on out-of-domain reactions (Test 2 ≡ 56 reactions), where the model's RMSE decreased from 2.0 ± 0.0 $\log k$ (= the RMSE of the dummy model) to 1.4 ± 0.2 $\log k$ (Figure 3.5**a**).

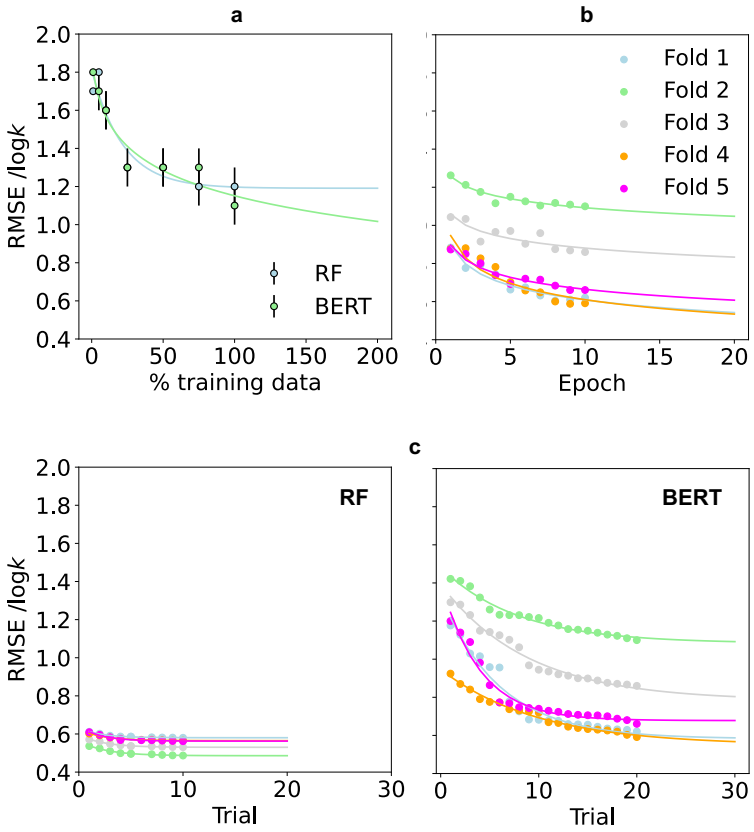


Figure 3.4: Learning curves depicting convergence in the RF and BERT models with respect to: training set size (**a**), epoch number (BERT only, **b**), and hyperparameter optimisation trial (**c**). **a** was evaluated using the total test data Test 1 + Test 2 and the mean RMSE taken over the five CV folds, with the standard error providing an uncertainty estimate. **b** and **c** were evaluated per fold using validation data.

Consequently, augmenting chemical diversity in the training data improves the model generalisability across a broader spectrum of reaction space. In fact, diversifying the training data with only 24% of the reactions curated in this work (with $S_T < 0.4$ to the 4666 training reactions from Ref[33]) was enough for the RMSE to converge to 1.4 $\log k$ when evaluated using Test 2. Accordingly, the revised RF model was employed in this work. Reactions with $S_T > 0.4$ to the diversified data were omitted from all test sets, so the results seen here reflect the model's generalisability to unseen chemistry.

The relative decrease in RMSE from diversifying the training data is visualised in Figure 3.5**b** for the reactions in Test 2 in electrolyte solution (12 reactions), with azide LGs (5 reactions), and with phosphine nucleophiles (4 reactions). The largest decrease was observed for reactions in electrolyte (ΔRMSE = -0.5 ± 0.2 log k, compared to -0.2 ± 0.1 log k for azide LGs and -0.1 ± 0.2 log k for phosphine nucleophiles). This highlights the importance of the ionic strength feature in accurately predicting the reactivity of S_N2 reactions in electrolyte. Indeed, ionic strength proved to be a high impact feature (defined as having a feature importance in the top 25% across the test data) for 100% of test reactions in electrolyte.

3.2.3 Comparison of RF, BERT, and QM

The performance of RF and BERT was assessed based on accuracy and training time. Both models exhibited similar accuracy (RMSE /log k: 1.1 ± 0.1 (BERT) and 1.2 ± 0.1 (RF), Wilcoxon's p = 0.50) on the combined test data of Test 1 and Test 2 (total 129 reactions). However, the time taken to achieve this accuracy was significantly faster for the RF (256 s vs 190,305 s (52.9 hrs), Section 3.1.3).

The absolute prediction errors on the test data are plotted in figure 3.5**d** for the RF and BERT models. The error distributions for both models have a similar MAE (BERT: 0.8 log k, RF: 0.9 log k) and interquartile range (IQR, Both models: 0.4 to 1.1 log k), affirming that they perform comparably.

The prediction accuracy of RF and BERT was also validated against a dummy model that predicts the mean log k_{exp} of the training data for each test point. The dummy RMSE was 2.0 ± 0.0 log k, which is significantly larger than that of the trained BERT and RF. This validates that both models learned about chemical reactivity during the training process.

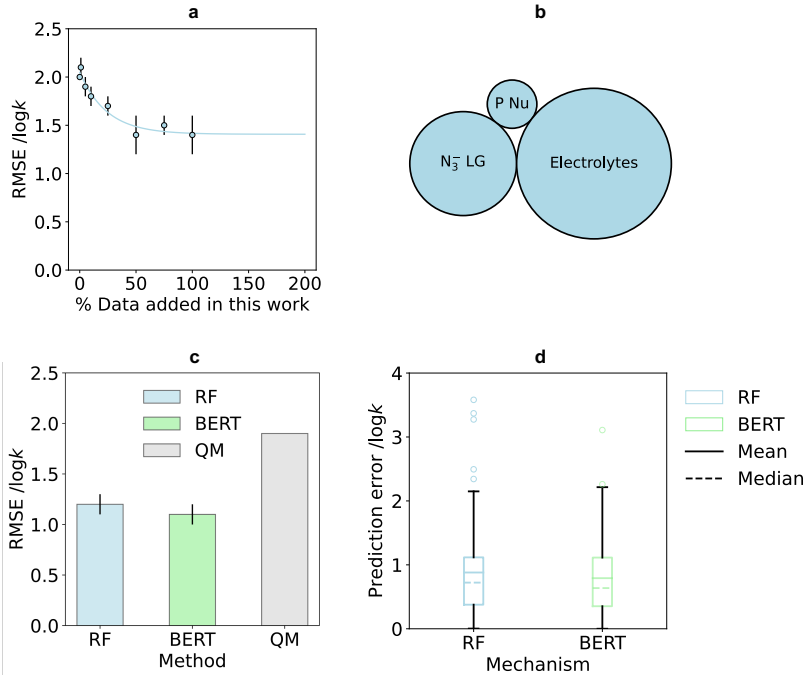


Figure 3.5: Prediction results. a. Learning curve depicting the change in RMSE of Ref[33]'s RF model upon increasing the chemical diversity of the training data in the current work (evaluated using 56 out-of-domain reactions). **b.** Relative RMSE decrease (represented by the circle radii) from diversifying the training data, evaluated using reactions in Test 2 with electrolyte solvent (12 reactions), azide LGs (N_3^- LG, 5 reactions), and phosphine nucleophiles (P Nu, 4 reactions). **c.** RMSE of the RF and BERT models (evaluated using 129 external test reactions) compared to 30 QM calculations carried out at the CPCM(solvent) CCSD(T)/def2-TZVP//PBE0-D3BJ/def2-SVP LOT. To obtain the QM RMSE in $\log k$, the QM $\Delta G^{\ddagger}$ values (in J mol^{-1}) were converted to $\log k$ using the Eyring equation (Equation 1.1). QM calculations carried out with María Calvo. **d.** Distribution of (absolute) prediction errors on the 129 external test reactions for the RF and BERT models.

Both models were also benchmarked against QM-calculated $\Delta G^{\ddagger}$ values at the CPCM(solvent) CCSD(T)/def2-TZVP//PBE0-D3BJ/def2-SVP LOT,[b] with QM exhibiting a higher RMSE of 2.5 kcal mol^{-1} ≡ 1.9 log k (Figure 3.5**c**). The time taken for 30 QM calculations (24,457 s ≡ 6.8 hrs) contrasted starkly with the immediate results (< 1 s) of the same number of predictions by the RF or BERT

[b] QM calculations were carried out with María Calvo, a summer student in the Duarte group (June to September 2021).

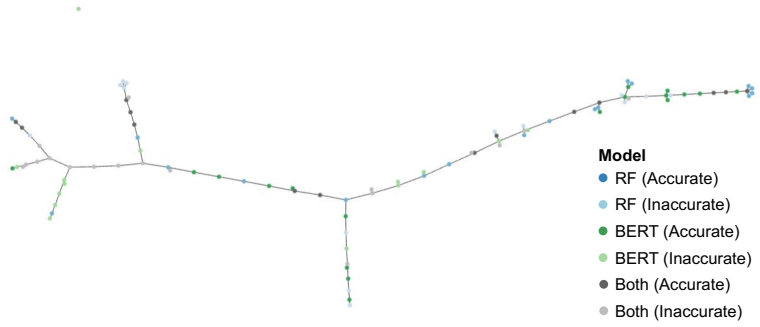


Figure 3.6: Accurate and inaccurate predictions of the RF (shown in blue) and BERT (shown in green) models, visualised as TMAPs. Here, accurate and inaccurate predictions are defined as those with the 25% smallest, and largest, error in log k, respectively. Note that the peripheral data clusters have been artificially repositioned closer to the main branch to fit within the frame. The original interactive versions of each TMAP are provided on GitHub (see *Data availability*).

models. Even when taking into account model training time, the RF training time is significantly less than one QM calculation, while the BERT training time equates to 233 QM calculations when trained using CPU, or 16 QM calculations when trained using GPU (Section 3.1.3). This reinforces the fact that these models are not only more accurate but also faster than traditional electronic structure methods.

Test reactions that were predicted accurately (25% smallest error in $\log k \equiv 32$ reactions), and inaccurately (25% largest error in $\log k \equiv 32$ reactions), by RF and BERT are provided in Figure 3.6. 44% of RF's accurate predictions were also predicted accurately by BERT, while this overlap was 56% for inaccurate predictions. Hence, despite similar predictive accuracy, the models differed in the reactions they accurately or inaccurately predicted, suggesting variations in the underlying relationships learned by each model. Interpreting these underlying relationships is the subject of the remaining text.

3.2.4 Interpreting accurate predictions

To ensure the models' accurate predictions were not spurious, both RF and BERT were examined for their ability to identify key structural and physical features influencing reactivity. This included evaluating the contributions from each reaction centre atom: Nu, LG, and electrophilic C atoms, as well as temperature (represented as T^{-1}) and solvent polarity and proticity. Feature importances calculated in this work are relative to a dummy model that predicts the mean $\log k_{\text{exp}}$ of the training data for each test reaction. By definition, all features of the dummy model have an importance of zero, however this is also shown mathematically in Section 3.1.6.

Both models exhibited consensus in recognising the importance of these reaction centres and conditions. For example, the LG atom emerged as a high impact feature (defined as having a feature importance in the top 25% across the test data) in $>$ 90% of accurate predictions for both RF and BERT, while the Nu atom was found to be a high impact feature in $\geq$ 75% of accurate predictions for both models (Figure 3.7**a**). In contrast, the electrophilic C atom was high impact in 100% of the RF's accurate predictions but only 38% of BERT's accurate predictions. This discrepancy is attributed to the nature of the comparison: in the RF, the electrophilic C is represented by molecular fragments that also include the environment around the electrophilic centre, while in the BERT model it is represented by the electrophilic C token only. Further analysis of learned effects at the electrophilic centre for each model show that RF and BERT both identify reaction centre atoms as key features for predicting reactivity, reinforcing the chemical validity of their accurate predictions (*vide infra*).

Additionally, both models recognised temperature as a high impact feature for 97% of accurate predictions, demonstrating their ability to recognise key physical

features. Regarding solvent, this was described by 13 properties (representing either polarity or proticity) in the RF model, and as a SMILES in BERT. For the RF, all properties pertaining to the primary solvent component were high impact in 90 to 100% of accurate predictions, in agreement with Ref[33]'s assessments that solvent properties are important features of the RF model. Solvent also proved a significant feature of the BERT model, being identified as high impact in 72% of accurate predictions. This reinforces the ability of both models to recognise key physical features on reactivity. BERT and RF's learning of physical features is summarised in Figure 3.7**b**.

To assess whether RF and BERT effectively learned key structural and physical effects, this work evaluates whether high impact features of accurate predictions aligned with known reactivity rules, including LG, nucleophilic, steric, allylic, temperature, and solvent effects. To analyse structural and physical effects, the feature importances of key reaction centre atoms were evaluated. The importance of each feature is relative to that of a dummy model that predicts the mean $\log k_{\text{exp}}$ of the training data for each test reaction. By definition, all features of the dummy model have an importance of zero, however this is also shown mathematically in Section 3.1.6.

3.2.4.1 LG effects

The feature importances of LG atoms contributing to the predicted rate, either increasing (positive sign) or decreasing (negative sign) $\log k_{\text{pred}}$ are presented in Figure 3.8. Here, accurate predictions with halide LGs were evaluated with the distribution of LGs being 16 × I, 6 × Br, 3 × Cl, and 0 × F for RF and 17 × I, 8 × Br, 2 × Cl, and 0 × F for BERT. Note that the total distribution of halide LGs in the test set was 53 × I, 18 × Br, 16 × Cl, and 4 × F. In the RF, the LG

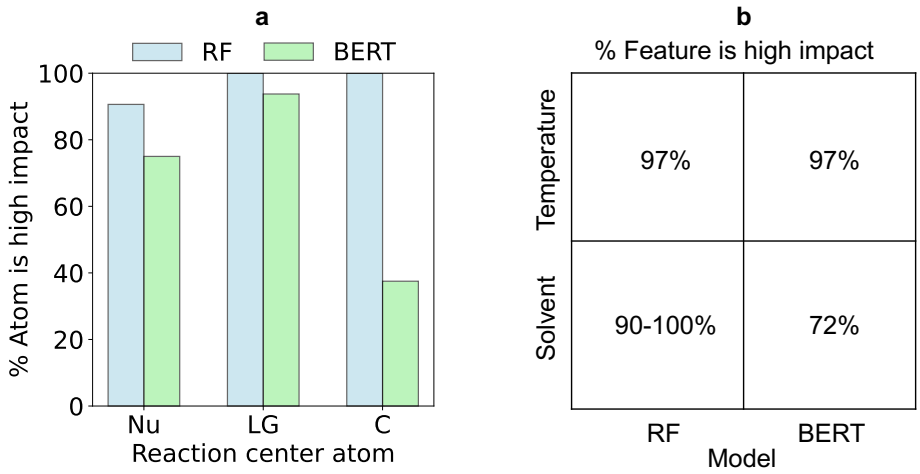


Figure 3.7: High impact features of accurate predictions. a. The percentage of accurate predictions where the nucleophilic (Nu), leaving group (LG), and electrophilic carbon (C) atoms were high impact features of the RF and BERT models. **b.** The percentage of accurate predictions where temperature and solvent were high impact features of the RF and BERT models. A breakdown of the percentage of accurate predictions where each solvent property was high impact in the RF is provided in Appendix E.3.

atoms were represented by C-I, C-Br, C-Cl, and C-F fragments, while in BERT they were represented by I, Br, Cl, and F tokens.

In the BERT model, I and Br respectively increased $\log k_{\text{pred}}$ 100% of the time they were high impact (12 and 5 reactions, respectively), while Cl decreased $\log k_{\text{pred}}$ in both examples with this LG. In the RF model, I increased $\log k_{\text{pred}}$ in 100% of analysed examples, Br had mixed impact: increasing $\log k_{\text{pred}}$ in one reaction, decreasing $\log k_{\text{pred}}$ in two reactions, and having minimal impact in 50% of reactions, and Cl either decreased $\log k_{\text{pred}}$ (1 reaction) or had no significant effect (2 reactions). This suggests that both models identified a positive correlation between halide size and reactivity. Regarding the four test reactions with an F LG, they were neither predicted accurately nor inaccurately by the BERT model. However, they were all predicted inaccurately by the RF. For RF, F decreased $\log k_{\text{pred}}$ in each of these reactions, which is expected given that F is a relatively

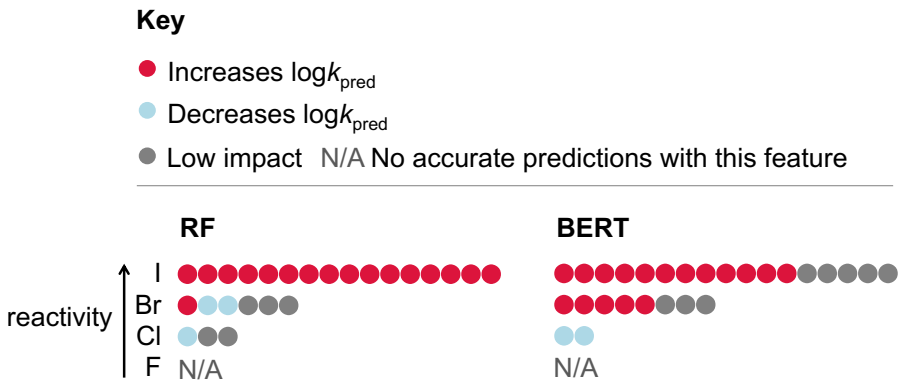


Figure 3.8: Analysis of LG effects. Signs (positive or negative) of feature importances for each of RF and BERT's accurate predictions where the LG was high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature.

poor LG. This suggests that the inaccurate prediction of reactions with an F LG is not attributed to RF's learning of LG effects. A representative example is shown in Figure 3.9 (taken from the test data), which demonstrates that both RF and BERT predicted the correct reactivity trend Cl < Br < I based on mean predicted values (log k_{pred}: -3.2 ± 0.1, -2.7 ± 0.1, -2.6 ± 0.1 for RF and -3.4 ± 0.3, -2.3 ± 0.2, -2.1 ± 0.3 for BERT). The overlap observed for the predicted values for Br and I is attributed to the small difference in experimental values (0.1 and 0.2 log k) compared to the model RMSE of ≈ 1.1 log k. Importantly, both models recognised the trend of increasing reactivity with larger halides. This was also reflected in the feature importances of Cl, Br, and I in this example: -0.22 ± 0.04, 0.02 ± 0.01, 0.24 ± 0.05 for RF and -0.13 ± 0.08, 0.56 ± 0.12, 0.83 ± 0.18 for BERT.

3.2.4.2 Nucleophilic effects

To assess whether RF and BERT effectively learned nucleophilic effects, this work evaluates whether the feature importances of halide nucleophiles in accurate

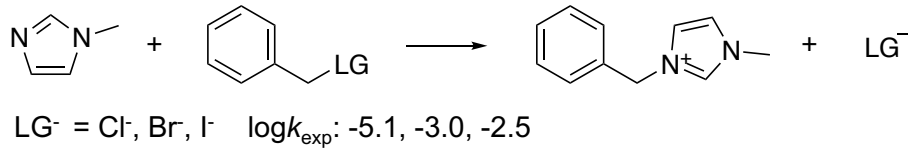


Figure 3.9: Representative example of LG effects from the test data. Reaction in acetonitrile at 298.15 K.

predictions, either increasing $\log k_{\text{pred}}$ (positive sign) or decreasing $\log k_{\text{pred}}$ (negative sign), align with known reactivity trends. This assessment is focused on halide nucleophiles, as they are the only group of nucleophiles in the test data that represents variation in the Nu atom (Justification 1, Appendix C). This evaluation is divided into reactions in protic solvent (reactivity increases with increasing halide size, Figure 3.10), and reactions in aprotic solvent (reactivity decreases with increasing halide size, Figure 3.11). In the BERT model, halide nucleophiles were represented by [I^-], [Br^-], [Cl^-], and [F^-] tokens. Conversely, in the RF model, which has a minimum fragment length of 2 atoms, halide nucleophiles (being monoatomic) were only represented on the products side of the reaction, by C-I, C-Br, C-Cl, and C-F fragments in the major product (see schematic in Figures 3.10 and 3.11). For fair comparison, the feature importances of the I, Br, Cl, and F tokens in the major product are also shown for the BERT model in Figures 3.10 and 3.11.

The distribution of halide nucleophiles (protic solvent) across accurate predictions was $0 \times I^-$, $1 \times Br^-$, $2 \times Cl^-$, and $0 \times F^-$ for both RF and BERT (Figure 3.10, total test distribution: $2 \times I^-$, $2 \times Br^-$, $3 \times Cl^-$, and $0 \times F^-$). While this is too few examples to draw general conclusions from, the predicted increase or decrease in $\log k$ with the nucleophile was chemically reasonable in each of these data points. For instance, Br^- increased $\log k_{\text{pred}}$ in the one RF example and one BERT example, while Cl^- decreased $\log k_{\text{pred}}$ in both RF examples and one of the two BERT examples, suggesting both models recognised the larger halide is more

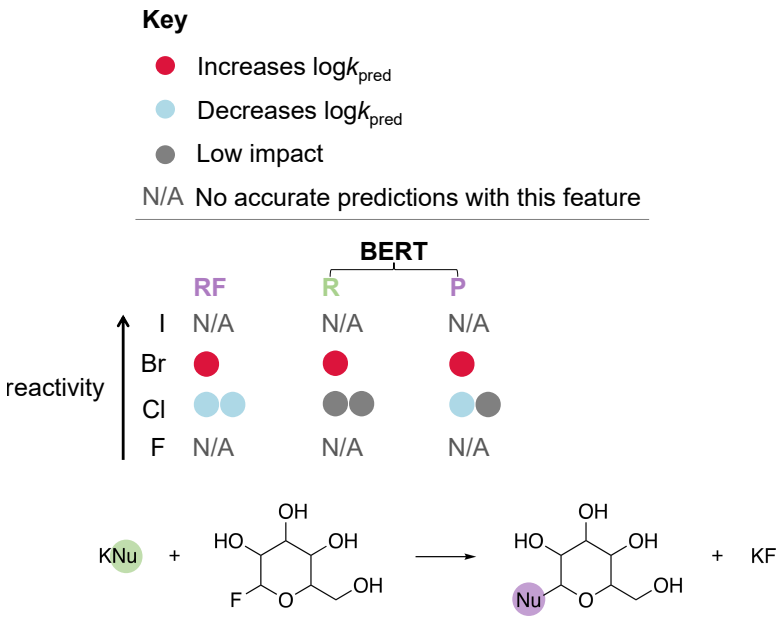


Figure 3.10: Analysis of nucleophilic effects (protic solvent). Signs (positive or negative) of feature importances for each of RF and BERT's accurate predictions where the nucleophile (which was in protic solvent) was high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature. In the BERT model, R and P correspond to reactants and products, respectively, while only the products side is evaluated in the RF.

nucleophilic than the smaller halide in protic solvent. Meanwhile, none of RF's accurate predictions, and only two of BERT's accurate predictions involving halide nucleophiles were in aprotic solvent (Figure 3.11, total test distribution: 2 × I^-, 1 × Br^-, 4 × Cl^-, and 0 × F^-). Hence, nucleophilic effects in aprotic solvent were not analysed for either model.

3.2.4.3 Steric effects

To analyse steric effects, accurate predictions with alkyl-substituted centres and unsubstituted centres were evaluated (Figure 3.12). In the RF, alkyl-substituted centres are modelled by C-C (4 reactions), C-C-C (3 reactions), and C-C-C-C (1 reaction) fragments. The importance of these features in reactions with

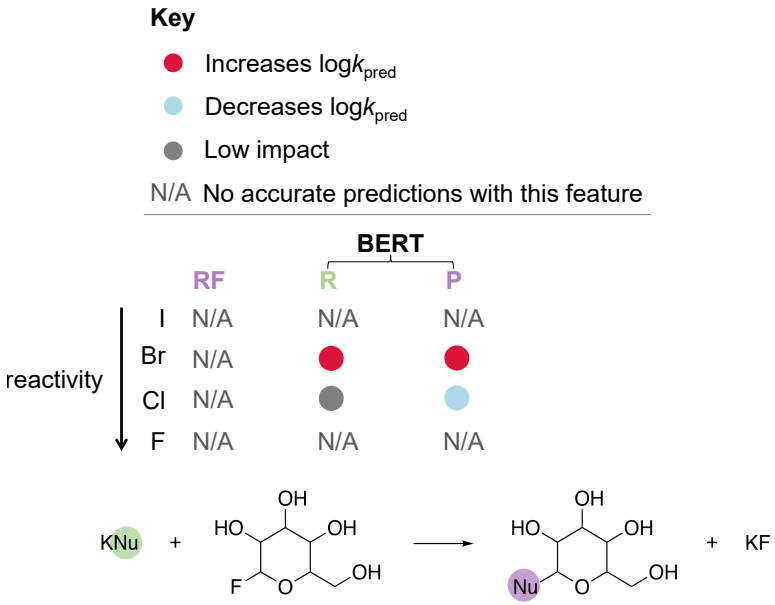


Figure 3.11: Analysis of nucleophilic effects (aprotic solvent). Signs (positive or negative) of feature importances for each of RF and BERT's accurate predictions where the nucleophile (which was in aprotic solvent) was high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature. In the BERT model, R and P correspond to reactants and products, respectively, while only the products side is evaluated in the RF.

unsubstituted centres (where the fragment is either present in the reaction but not at the centre or is not present in the reaction at all) are used as a control, corresponding to 12, 12, and 13 reactions for C-C, C-C-C, and C-C-C-C respectively (note that one reaction with an unsubstituted centre, shown in Figure 3.13**a**, was not used as a control in the analysis of C-C and C-C-C fragments due to these fragments containing the product atom mapping of nucleophilic C^-). Contrarily, alkyl-substituted centres in the BERT model are represented by the electrophilic C and substituting C atoms (13 reactions), while unsubstituted centres are solely represented by the electrophilic C (7 reactions).

Analysis of steric effects shows that in the BERT model, substituted centres

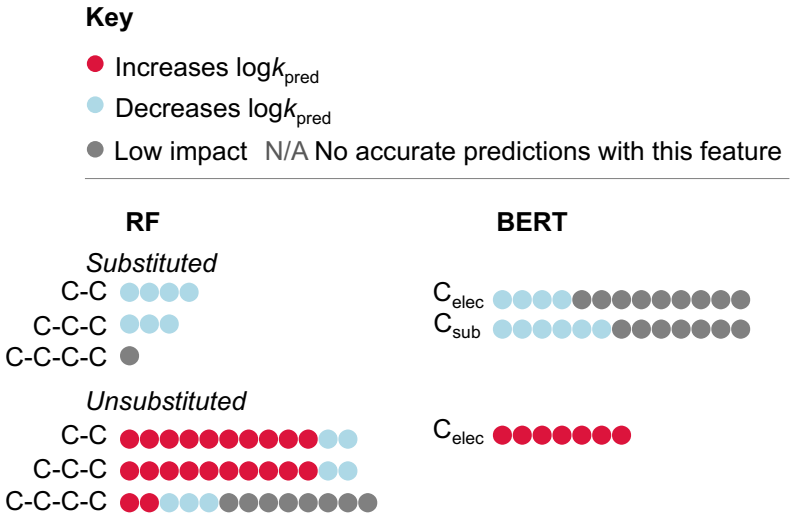


Figure 3.12: Analysis of steric effects. Signs (positive or negative) of feature importances for each of RF and BERT's accurate predictions where alkyl-substituted and unsubstituted electrophilic centres were high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature. C_{elec} and C_{sub} correspond to electrophilic and substituting carbons, respectively.

decreased log k_{pred} in 100% of reactions where they were high impact (6 reactions), while unsubstituted centres increased it. In the RF model, substituted centres modelled by C-C and C-C-C fragments decreased log k_{pred} in 100% of reactions where these fragments were high impact (4 and 3 reactions, respectively), while these features typically increased log k_{pred} for reactions with unsubstituted centres (83% of analysed examples). This is consistent with the knowledge that steric hindrance at the electrophilic centre decreases S_N2 reactivity. This said, C-C and C-C-C also decreased log k_{pred} in two reactions with unsubstituted centres, corresponding to those where C-C and C-C-C were present in other molecules in the reaction (shown in Figure 3.13**b**), suggesting a spurious correlation between the presence/absence of carbon fragments in the reaction and reactivity. This was remediated by removing fragments away from the reaction centre from the feature

set, with no detriment to the overall accuracy of the RF (*vide infra*). Regarding the C-C-C-C fragment, this was found to be low impact in the one example where it was at a substituted centre, and 62% of examples with unsubstituted centres. This is attributed to the limited information about S_N2 reactivity that can be learned from the C-C-C-C fragment, which can be adequately learned from C-C and C-C-C (*vide infra*). Unusually, the C-C-C-C feature also decreased $\log k_{\text{pred}}$ in three examples where it was absent from the reaction, corresponding to the reaction of 2-amino-1-methylbenzimidazole with methyl iodide at three different temperatures (Figure 3.13**c**). Because each of these examples are of the same transformation, the decrease in $\log k_{\text{pred}}$ with the absence of C-C-C-C was attributed to a spurious correlation learned from similar training scaffolds. This was resolved by removing fragments away from the reaction centre from the feature set, without effecting the overall accuracy of the RF (*vide infra*). In conclusion, both models recognised that S_N2 reactivity is decreased by steric hindrance. For the RF, it is recommended that only fragments at the reaction centre are used during training to prevent learning of spurious correlations.

Steric effects are shown for the representative reaction in Figure 3.14**a**. Despite correctly predicting the reactivity order R = *cyclo*-C_6H_{11} < R = (Me, Me) < R = (Pr, H) for this example based on mean predicted values ($\log k_{\text{pred}}$: -0.8 ± 0.5, -0.5 ± 0.6, -0.2 ± 0.7), RF prediction uncertainties are significant (avg. 177.5% of the predicted values), rendering the predictions unreliable. Nevertheless, the feature importances attributed to the R group fragments were increasingly positive from R = *cyclo*-C_6H_{11} to R = (Pr, H) (Figure 3.14**b**), suggesting RF recognised that reactivity increases with decreasing steric hindrance around the electrophilic centre. Conversely, BERT was unable to predict the correct reactivity order, predicting $\log k$ values of -0.4 ± 0.4, -1.1 ± 1.1, and -0.9 ± 0.8 for R = *cyclo*-C_6H_{11}, R = (Me, Me), and R = (Pr, H), respectively, indicating limitations

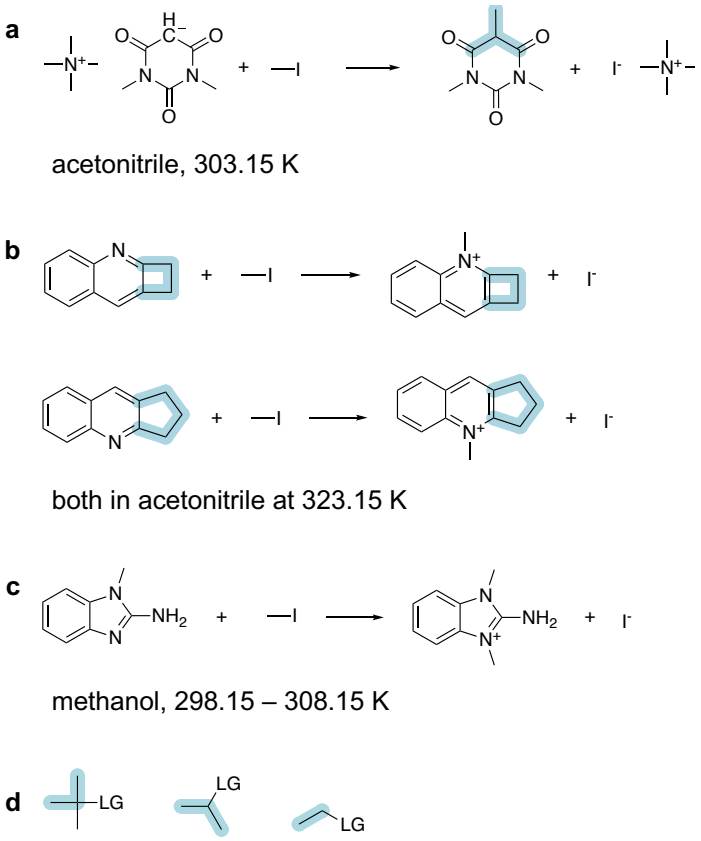


Figure 3.13: Reactions relevant to the discussion of steric effects. a. Reaction omitted from the analysis of C-C and C-C-C fragments due to these fragments containing the product atom mapping of nucleophilic C^- (shown in blue). **b.** Reactions where C-C and C-C-C were present in the reaction away from the electrophilic centre (shown in blue). **c.** Reactions where the C-C-C-C feature decreased $\log k_{\text{pred}}$, despite being absent from the reaction. **d.** The longest carbon fragment at tertiary, secondary and primary electrophilic centres are C-C-C, C-C-C, and C-C, respectively (shown in blue).

in identifying the reactivity differences between substituted centres (Figure 3.14**c**). This was also reflected in the feature importances attributed to the R group tokens, which were all zero within error.

Overcoming spurious correlations in the RF model. In the RF model, C-C and C-C-C fragments decreased $\log k_{\text{pred}}$ in two reactions with unsubstituted centres (shown in Figure 3.13**b**), corresponding to those where C-C and C-C-C were present in other molecules in the reaction (*vide supra*). This was attributed

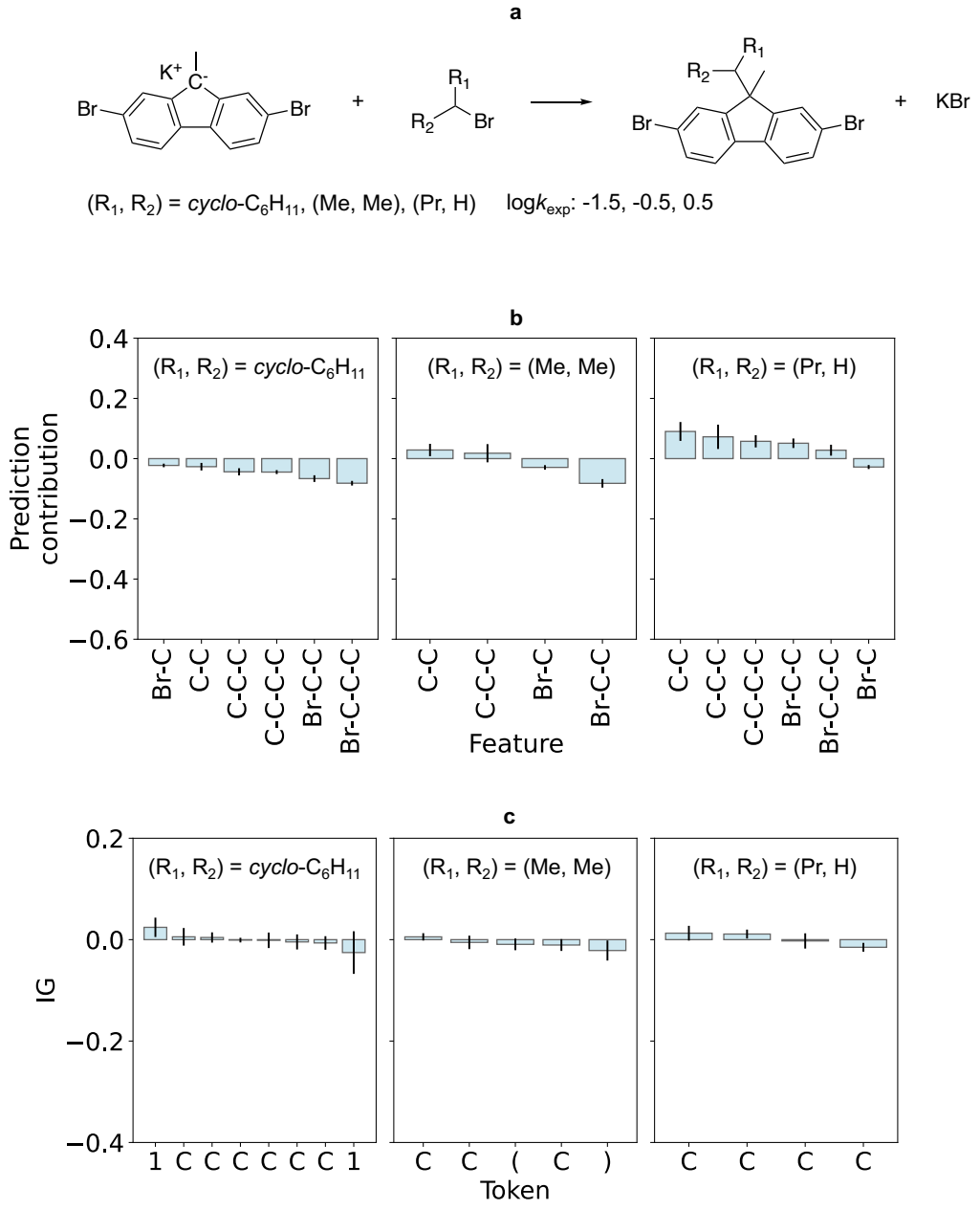


Figure 3.14: a. Representative example of steric effects taken from the test data. Reaction in DMSO at 298.15 K. **b and c.** Feature importances of R groups in the RF (**b**) and BERT (**c**) models for the reaction in **a**.

to a spurious correlation between the presence/absence of carbon fragments in the reaction and reactivity. To test this hypothesis, an identical RF was trained, but ISIDA fragments without a substructure match to one of the reaction centre atoms (Nu, LG, electrophilic C, or one their product atom mappings) were omitted from the feature set. In this revised model (denoted RF_{centre}), the C-C and C-C-C features increased $\log k_{pred}$ in 100% of analysed reactions with unsubstituted centres, including the reactions in Figure 3.13**b** (Figure 3.15). This suggests that removing noise created by away-from-centre scaffolds improves RF's learning of steric effects. Furthermore, this was achieved while maintaining the overall model RMSE of 1.2 ± 0.1 $\log k$.

Also in the RF model, the C-C-C-C feature decreased $\log k_{pred}$ in three of RF's accurate predictions where it was absent from the reaction, corresponding to the transformation in Figure 3.13**c** (*vide supra*). Because each of these three examples are of the same transformation, the decrease in $\log k_{pred}$ with the absence of C-C-C-C was attributed to a spurious correlation learned from similar training scaffolds. Using RF_{centre}, the C-C-C-C feature either increased $\log k_{pred}$ or had low impact for 100% of reactions with an unsubstituted centre, including the three reactions in Figure 3.13**c** (Figure 3.15). This affirms that spurious correlations may be prevented by removing noise created by away-from-centre scaffolds, without compromising model accuracy.

Rationalisation for low impact C-C-C-C fragments. The C-C-C-C fragment was typically a low impact feature in reactions with substituted and unsubstituted centres for the RF. This was attributed to the limited information about S_N2 reactivity that can be learned from the C-C-C-C fragment, which can be adequately learned from C-C and C-C-C. This can be rationalised by considering the longest carbon fragment that defines each of the three degrees of substitution at the electrophilic centre is C-C for primary centres, and C-C-C for secondary and

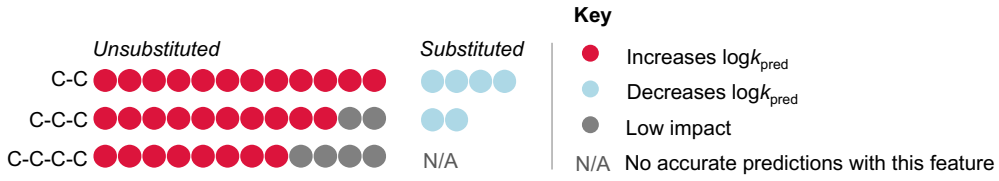


Figure 3.15: Steric effects (RF_{centre}). Signs (positive or negative) of feature importances for each of RF_{centre}'s accurate predictions where alkyl-substituted and unsubstituted electrophilic centres were high impact. Red and Blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature.

tertiary centres (Figure 3.13**d**). Hence, the C-C-C-C fragment only becomes relevant when considering carbons further along the substituting alkyl groups.

3.2.4.4 Allylic effects

It is well known that allylic effects from alkene, alkyne, and aromatic bonds (to the electrophilic centre) increase S_N2 reactivity. Here, accurate predictions with each of these centres were evaluated with the distribution being 0 × alkene, 2 × alkyne, and 6 × aromatic for RF and 2 × alkene, 3 × alkyne, and 4 × aromatic for BERT (Figure 3.16). The total distribution of centres in the test set was 2 × alkene, 4 × alkyne, and 42 × aromatic. Note that carbonyl effects are not discussed here as there was only one example in the test set. Alkene, alkyne, and aromatic bonds at the electrophilic centre were represented by C-C=C, C-C≡C, and C-C:C fragments in the RF (where ':' is an aromatic bond), and =, ≡, and c tokens in BERT (where c is an aromatic carbon bonded to the centre).

In the BERT model, alkyne increased log k_{pred} in 100% of analysed reactions, while alkene groups increased log k_{pred} in one out of two examples, having low impact in the other. Aromatic groups bound to the electrophilic centre were not identified by the BERT model. Conversely, in the RF model, aromatic groups

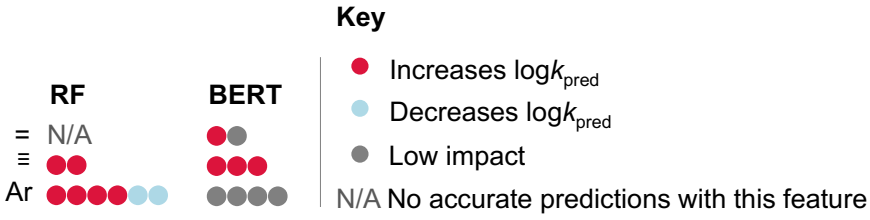


Figure 3.16: Analysis of allylic effects. Signs (positive or negative) of feature importances for each of RF and BERT's accurate predictions where alkene (=), alkyne (≡), and aromatic (Ar) bonds to the electrophilic centre were high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature.

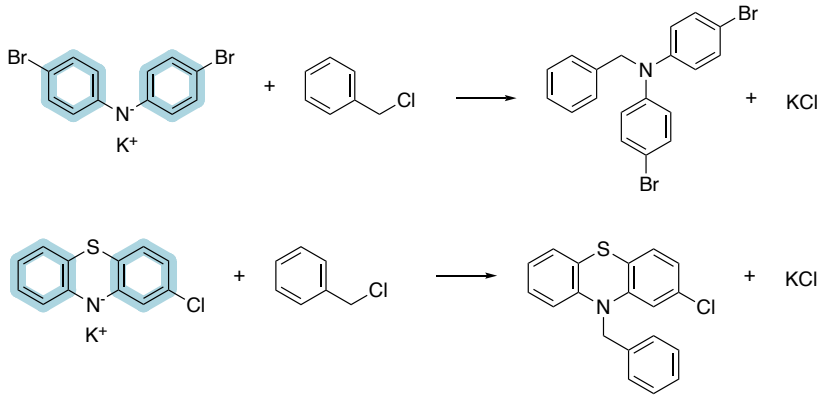


Figure 3.17: Reactions where aromatic groups decreased log k_{pred} in the RF model attributed to their resonance with the nucleophilic atom (shown in blue). Reactions in DMSO at 298.15 K.

exhibited mixed impact, increasing log k_{pred} in four examples and decreasing log k_{pred} in two. However, the latter can be attributed to the fact that aromatic groups in the nucleophile are expected to decrease log k in these two reactions, which are shown in Figure 3.17. Meanwhile, alkyne bonds increased log k_{pred} in both examples where they were bound to the centre in the RF. Regarding the two test reactions with an alkene group, these were neither predicted accurately nor inaccurately by the RF so are not discussed here.

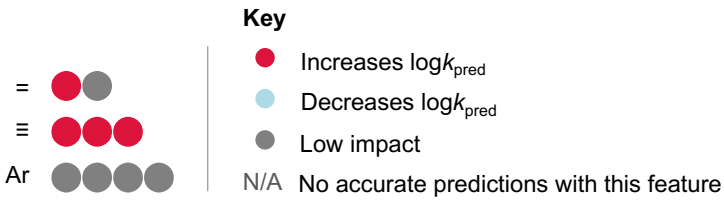


Figure 3.18: Analysis of allylic effects (summing over ring tokens). Signs (positive or negative) of feature importances for each of BERT's accurate predictions where alkene (=), alkyne (≡), and aromatic (Ar) bonds to the electrophilic centre were high impact. In contrast to the results presented in Figure 3.16, here, aromatic effects were evaluated by summing the importances over ring tokens. Red and Blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature.

Importantly, neither model made inaccurate predictions with alkene or alkyne bonds adjacent to the electrophilic centre (Section 3.2.5). This confirms that allylic effects were not a source of error for either model. Aromatic groups increased log k_{pred} in 13 out of RF's 14 inaccurate predictions where they were present, affirming their recognition was not a source of error. However, aromatic groups were low impact in 89% of BERT's inaccurate predictions, suggesting poor learning of aromatic effects as a potential source of prediction error.

To validate that the low importance of aromatic groups in the BERT model wasn't an artifact of only considering the carbon bonded to the electrophilic centre (as opposed to the entire aromatic ring), the importances were also calculated by summing over the ring tokens (Figure 3.18). Aromatic groups were still low impact when summing over the ring tokens, affirming that the results seen here were not biased by the way the importances were calculated.

Allylic effects are shown for one representative example in Figure 3.19**a** (taken from Ref[197]). Both RF and BERT correctly predicted allyl chloride to be more reactive than butyl chloride, with BERT exhibiting larger prediction uncertainty

($\log k_{\text{pred,rel}}$ of allyl chloride/butyl chloride: 1.2 ± 0.3 for RF and 1.1 ± 0.7 for BERT; note predicted values are given as $\log k_{\text{pred,rel}}$ as only experimental k_{rel} values were available for this reaction). Notably, RF attributed a greater $\log k$ increase to the C-C fragment in allyl chloride than to the allylic bond, contrasting with the BERT model, which attributed a greater $\log k$ increase to the allylic bond token in line with chemical principles (Figures 3.19**b** and **c**).

In summary, both models recognise allylic effects, however RF outperforms BERT at recognising aromatic effects.

3.2.4.5 Temperature effects

Reciprocal temperature was introduced as a feature of the BERT model by appending it to the end of the reaction SMILES, and was a continuous-valued feature of the RF. Out of the four reactions in the test data observed at multiple temperatures (Reactions **A** to **D**, Figure 3.20), RF and BERT both correctly predicted the linear decrease in $\log k$ with increasing T^{-1} (negative Pearson's r value ($-r_P$) ≥ 0.97 and Pearson's p value (p_P) = 0.00 for both models). This was seen for both the predicted $\log k$ values and feature importances of T^{-1} (Figure 3.21). Hence, both RF and BERT models successfully captured the mathematical relationship between $\log k$ and temperature. Note that the reactions in Figure 3.20 were taken from the test data and not necessarily predicted accurately or inaccurately.

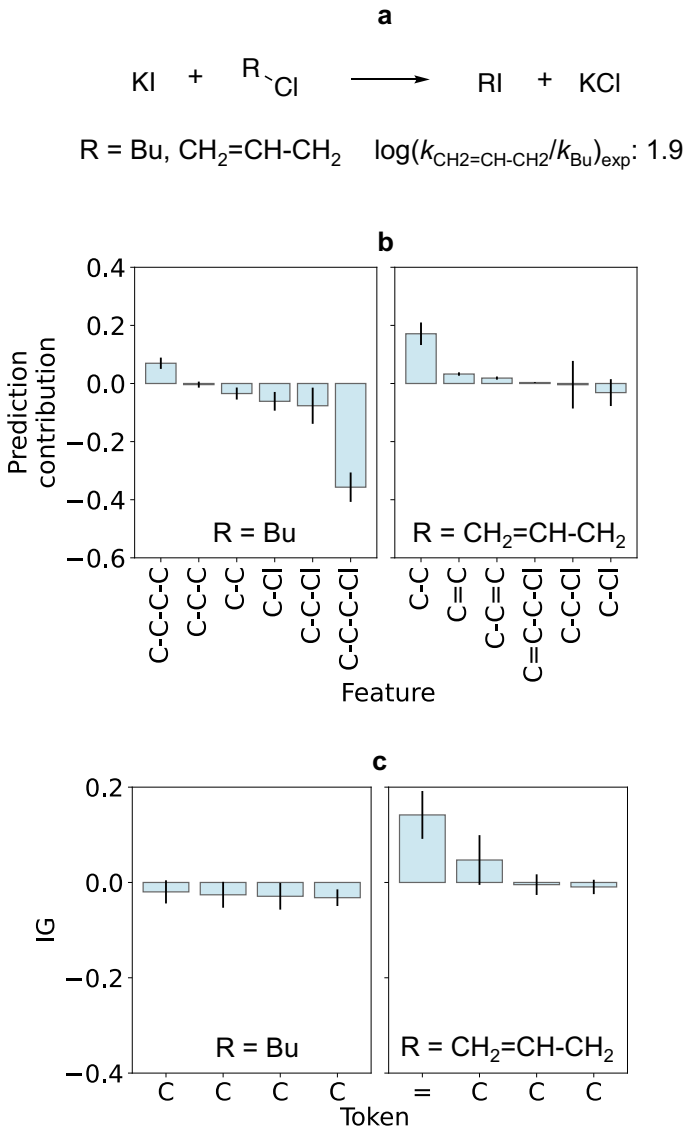


Figure 3.19: a. Representative example of allylic effects from Ref[197]. Reaction in acetone at 323.15 K. **b and c.** Feature importances of R groups in the RF (**b**) and BERT (**c**) models for the reaction depicted in **a**.

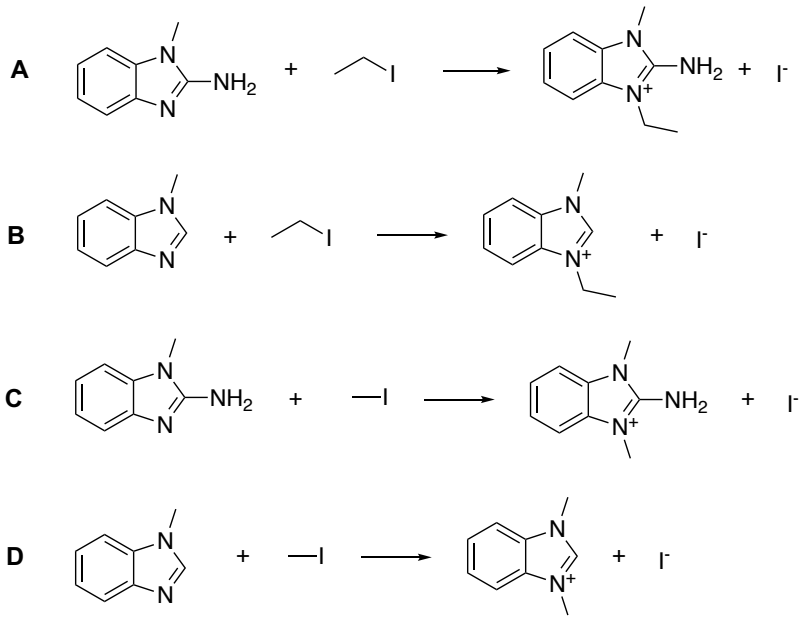


Figure 3.20: Reactions in the test data observed at multiple temperatures. Reactions in methanol at nine temperatures ranging from 293.15 K to 333.15 K.

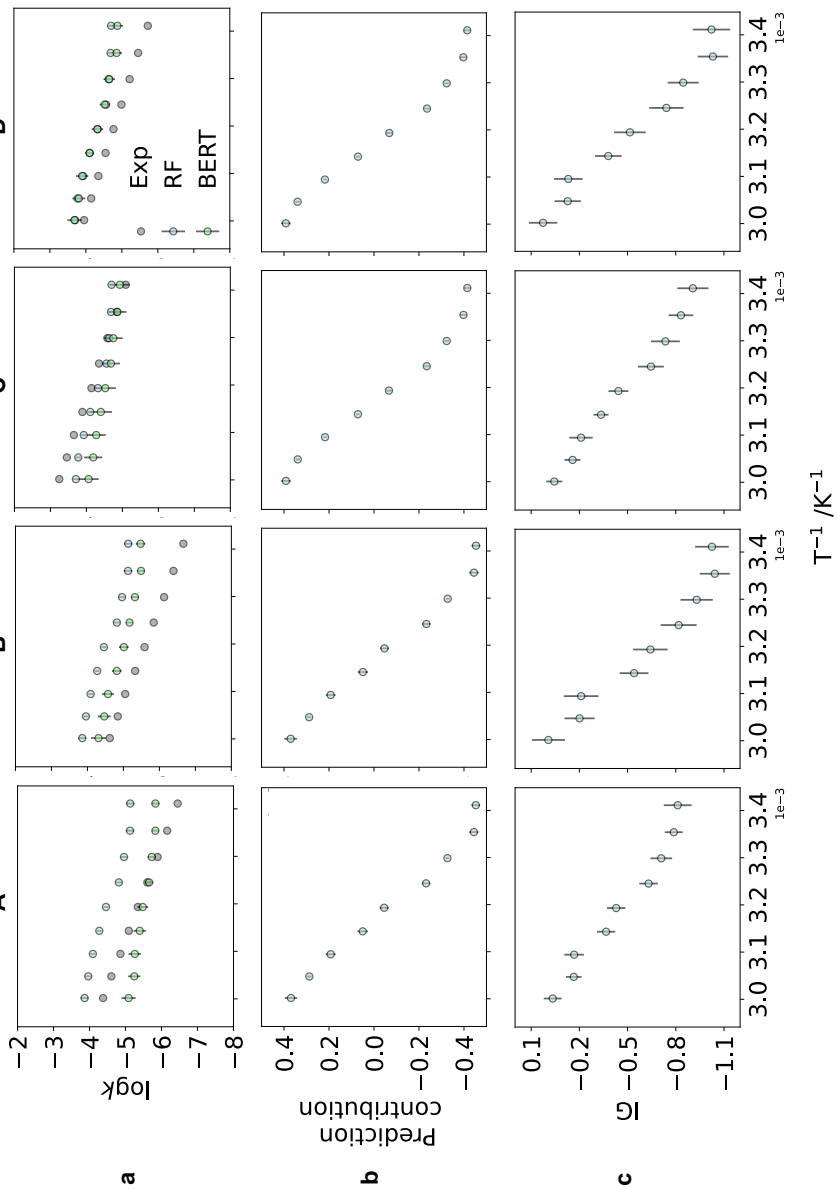


Figure 3.21: Analysis of temperature effects. a. Predicted (Pred) and experimental (Exp) log k values for each of the reactions in Figure 3.20. Predicted values are given for both the RF (shown in blue) and BERT (shown in green) models. **b and c.** Feature importances for each of the reactions in Figure 3.20 for the RF (**b**) and BERT (**c**) models. Here, T^{-1} denotes the reciprocal temperature.

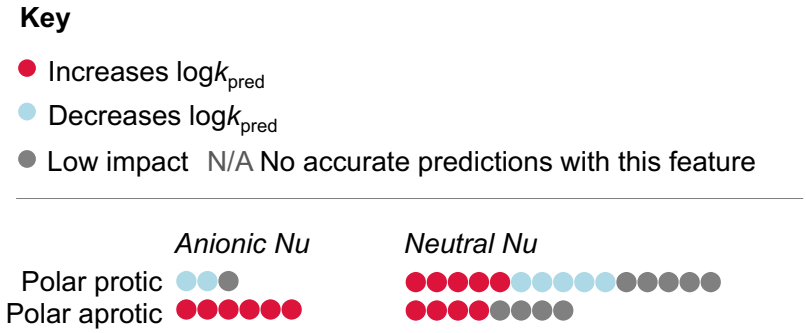


Figure 3.22: Analysis of solvent effects. Signs (positive or negative) of feature importances for each of BERT's accurate predictions where the solvent was high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature.

3.2.4.6 Solvent effects

Solvent was modelled by 13 properties representing polarity and proticity in the RF model, and by SMILES of the solvent molecules in BERT. For the RF, no consistent trend was observed between solvent properties and log k in the experimental data, likely because each solvent differs by multiple properties at once (Justification 2, Appendix C). Hence, solvent effects in the RF have not been evaluated here. As for BERT, solvent effects were evaluated by analysing the total contribution of polar protic ($\varepsilon_{298.15} > 15$ and proton bonded to a heteroatom)[196] and polar aprotic ($\varepsilon_{298.15} > 15$ and no proton bonded to a heteroatom)[196] solvent SMILES in accurate predictions with anionic and neutral nucleophiles, with the distribution of solvents being 3 × polar protic and 6 × polar aprotic for anionic nucleophiles, and 15 × polar protic and 8 × polar aprotic for neutral nucleophiles (Figure 3.22). Note there were no accurate predictions with non-polar solvent ($\varepsilon_{298.15} < 15$),[196] and only one such example in the test set.

BERT predicted polar protic solvents to decrease log k, and polar aprotic solvents to increase log k, in 100% of examples with anionic nucleophiles where the

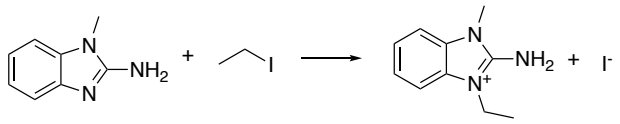


Figure 3.23: Transformation with a neutral nucleophile where polar protic solvent was predicted to decrease log k**.** Reactions in methanol at five temperatures ranging from 298.15 K to 318.15 K.

solvent was high impact (2 reactions for protic, and 6 reactions for aprotic). This may suggest that the BERT model identified the impact of hydrogen bonding on S_N2 reactivity.

Unusually, polar protic solvent was also predicted to decrease log k in 50% of examples with neutral nucleophiles where the solvent was high impact (5 reactions), despite neutral nucleophiles not being as susceptible to hydrogen bonding effects as anionic nucleophiles. However, each of these five reactions was the same type of transformation (reaction of 2-amino-1-methylbenzimidazole with ethyl iodide in methanol, Figure 3.23), suggesting a spurious correlation learned from similar training scaffolds. Importantly, polar protic and polar aprotic solvents were both predicted to increase log k in all other examples with neutral nucleophiles where solvent was high impact (5 reactions for protic, 4 reactions for aprotic), agreeing with the expected behaviour of these systems.

To summarise, BERT demonstrated a general understanding of solvent effects, however this was likely confounded by a spurious correlation for one transformation.

3.2.4.7 Summary

Overall, BERT and RF both recognised LG and temperature effects. Both models recognised steric effects, but it is recommended that only fragments at the reaction centre are used during training of the RF to prevent learning of spurious correlations. Both models also recognised allylic effects, but RF excelled at predicting aromatic effects. BERT recognised solvent effects (albeit confounded

by a spurious correlation for one system), however solvent effects could not be interpreted for the RF. Lastly, although there were too few examples to draw definitive conclusions about each model's learning of nucleophilic effects, both models provided chemically reasonable assessments for the limited examples available. These results suggest that the predictions of both models do align with known reactivity rules. However, caution should be taken when predicting steric effects with the RF, or aromatic or solvent effects with BERT.

3.2.5 Interpreting inaccurate predictions

To ensure that recognition of key structural and physical effects was not a source of error for either model, both RF and BERT's inaccurate predictions were examined. This included evaluating the contributions from each reaction centre atom: Nu, LG, and electrophilic C atoms, as well as temperature (represented as T^{-1}) and solvent polarity and proticity.

Both models identified the importance of the LG atom in reactions they predicted inaccurately, highlighting it as a high impact feature > 70% of the time (Figure 3.24**a**). RF also identified the Nu atom and electrophilic C atom as high impact in > 75% of reactions it predicted inaccurately, suggesting recognition of reaction centre atoms was not a source of error for this model. Conversely, the Nu and electrophilic C atoms were not identified as high impact features so frequently across BERT's inaccurate predictions (Nu: 56% of predictions, C: 19% of predictions). For the electrophilic C, this has been attributed to the nature of the comparison, where the BERT model analyses electrophilic C tokens while RF considers electrophilic centre fragments, which also include the environment around the electrophilic C. Further analysis of learned effects at the electrophilic centre for each model show that RF and BERT both identify reaction centre atoms as key features for predicting reactivity, reinforcing that recognition of the

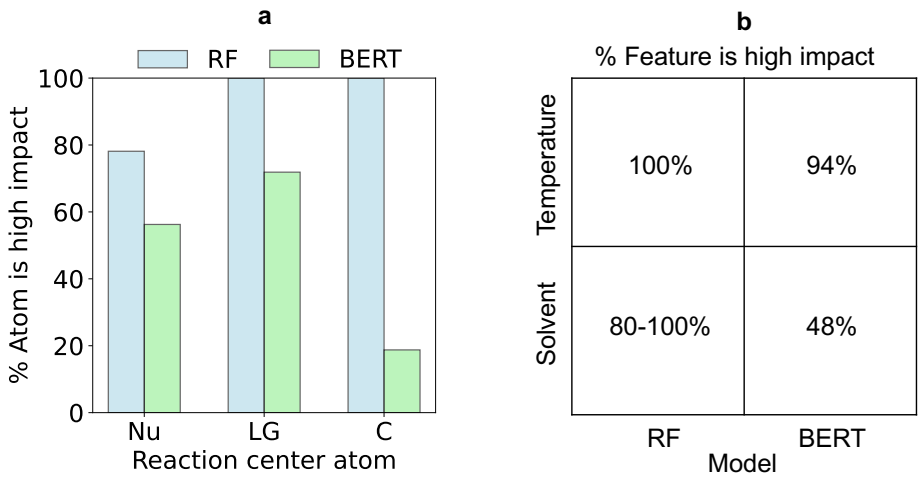


Figure 3.24: High impact features of inaccurate predictions. a. The percentage of inaccurate predictions where the nucleophilic (Nu), leaving group (LG), and electrophilic carbon (C) atoms were high impact features of the RF and BERT models. **b.** The percentage of inaccurate predictions where temperature and solvent were high impact features of the RF and BERT models. A breakdown of the percentage of inaccurate predictions where each solvent property was high impact in the RF is provided in Appendix E.3.

electrophilic centre is not a source of error for either model (*vide infra*). As for the Nu atom, there was no obvious relationship between the type of nucleophile and whether the Nu atom was high impact or low impact (Tables E.2 and E.3, Appendix E.2). Hence, it is possible that underestimating the importance of the nucleophile in these reactions contributed to their inaccurate prediction.

On the other hand, both models recognised temperature as a high impact feature for > 90% of reactions they predicted inaccurately, affirming that identifying this feature was not a source of error in either model (Figure 3.24**b**). Regarding solvent, all properties pertaining to the primary solvent component in the RF model (which represent either polarity of proticity) were high impact in over 80% of reactions that were predicted inaccurately. This reinforces that recognition of key physical features was not a source of error in the RF. Conversely, solvent was a low impact feature in 52% of BERT's inaccurate predictions. Among these,

81% involved a reaction between a neutral nucleophile and sulphonate LG in acetone at 308.15 K, indicating a possible correlation between this nucleophile-LG-solvent-temperature combination and the solvent's low impact. Notably, all training reactions featuring a neutral nucleophile in acetone at 308.15 K involved an iodide LG, rather than a sulphonate LG. This suggests that the inaccurate prediction of similar reactions in the test data may be due to the novelty of the LG, rather than the solvent's significance.

To ensure that learning of key structural and physical effects was not a source of error for either model, this work evaluates whether high impact features of inaccurate predictions aligned with known reactivity rules, including LG, nucleophilic, steric, allylic, and solvent effects.

3.2.5.1 LG effects

The feature importances of LG atoms contributing to the predicted rate, either increasing (positive sign) or reducing (negative sign) $\log k_{\text{pred}}$ are presented in Figure 3.25. Here, inaccurate predictions with halide LGs were evaluated with the distribution of LGs being 7 × I, 4 × Br, 6 × Cl, and 4 × F for RF and 3 × I, 5 × Br, 6 × Cl, and 0 × F for BERT. Note that the total distribution of halide LGs in the test set was 53 × I, 18 × Br, 16 × Cl, and 4 × F. In the RF, the LG atoms were represented by C-I, C-Br, C-Cl, and C-F fragments, while in BERT they were represented by I, Br, Cl, and F tokens.

In the RF model, I increased $\log k_{\text{pred}}$ 100% of the time it was high impact (4 reactions), Br was mostly low impact (75% of the time), Cl decreased $\log k_{\text{pred}}$ 80% of the time it was high impact (4 reactions), and F decreased $\log k_{\text{pred}}$ in 100% of analysed reactions. Similarly, in the BERT model, I increased $\log k_{\text{pred}}$ in both reactions where it was high impact, Br was mostly low impact (60% of the time), and decreased $\log k_{\text{pred}}$ in one reaction and increased $\log k_{\text{pred}}$ in the other, and Cl

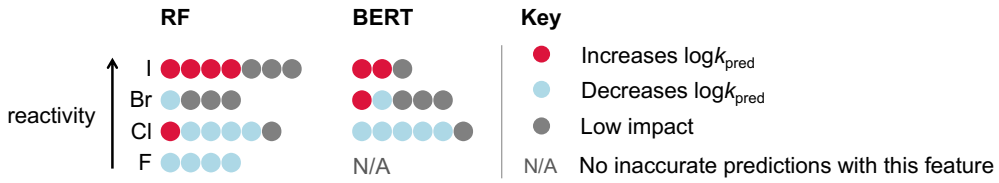


Figure 3.25: Analysis of LG effects (inaccurate predictions). Signs (positive or negative) of feature importances for each of RF and BERT's inaccurate predictions where the LG was high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no inaccurate predictions containing this feature.

decreased log k_{pred} 100% of the time it was high impact (5 reactions). Importantly, both models identified a positive correlation between halide size and S_N2 reactivity, affirming that learning of LG effects was not a source of error for either model.

3.2.5.2 Nucleophilic effects

To confirm that learning of nucleophilic effects was not a source of error for either model, inaccurate predictions with halide nucleophiles were evaluated. This assessment is focused on halide nucleophiles, as they are the only group of nucleophiles in the test data that represents variation in the Nu atom (Justification 1, Appendix C). This evaluation is divided into reactions in protic solvent (reactivity increases with increasing halide size, Figure 3.26), and reactions in aprotic solvent (reactivity decreases with increasing halide size, Figure 3.27). In the BERT model, halide nucleophiles were represented by [I^-], [Br^-], [Cl^-], and [F^-] tokens. Conversely, in the RF model, which has a minimum fragment length of 2 atoms, halide nucleophiles (being monoatomic) were only represented on the products side of the reaction, by C-I, C-Br, C-Cl, and C-F fragments in the major product. For fair comparison, the feature importances of the I, Br, Cl, and F tokens in the major product are also shown for the BERT model in Figures 3.26 and 3.27.

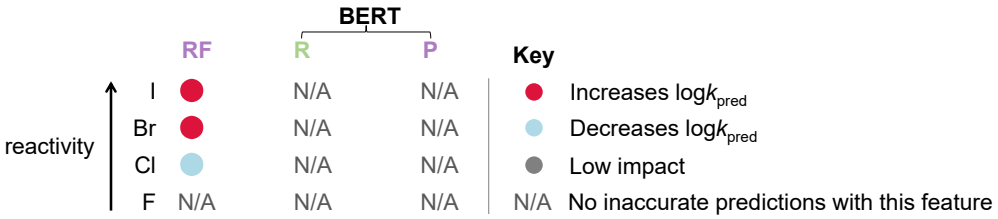


Figure 3.26: Analysis of nucleophilic effects (protic solvent, inaccurate predictions). Signs (positive or negative) of feature importances for each of RF and BERT's inaccurate predictions where the nucleophile (which was in protic solvent) was high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no inaccurate predictions containing this feature. In the BERT model, R and P correspond to reactants and products, respectively, while only the products side is evaluated in the RF.

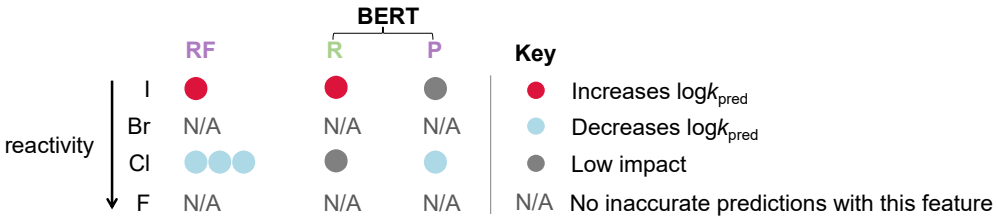


Figure 3.27: Analysis of nucleophilic effects (aprotic solvent, inaccurate predictions). Signs (positive or negative) of feature importances for each of RF and BERT's inaccurate predictions where the nucleophile (which was in aprotic solvent) was high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no inaccurate predictions containing this feature. In the BERT model, R and P correspond to reactants and products, respectively, while only the products side is evaluated in the RF.

The distribution of halide nucleophiles (protic solvent) across inaccurate predictions was 1 × I^-, 1 × Br^-, 1 × Cl^-, and 0 × F^- for RF. Meanwhile, none of BERT's inaccurate predictions involving halide nucleophiles were in protic solvent (Figure 3.26, total test distribution: total test distribution: 2 × I^-, 2 × Br^-, 3 × Cl^-, and 0 × F^-). While there are too few examples to draw general conclusion from, the predicted increase or decrease in log k with the nucleophile

was chemically reasonable in each of RF's inaccurate predictions. For instance, I^- and Br^- increased $\log k_{pred}$ in the one example of each nucleophile, while Cl^- decreased $\log k_{pred}$ in the one example with this nucleophile. This suggests that the RF recognised that reactivity increases with increasing halide size for these examples, which are in protic solvent.

Meanwhile, the distribution of halide nucleophiles (aprotic solvent) across inaccurate predictions was $1 \times I^-$, $0 \times Br^-$, $3 \times Cl^-$, and $0 \times F^-$ for RF and $1 \times I^-$, $0 \times Br^-$, $1 \times Cl^-$, and $0 \times F^-$ for BERT (Figure 3.27, total test distribution: $2 \times I^-$, $1 \times Br^-$, $4 \times Cl^-$, and $0 \times F^-$). While this is too few examples to draw general conclusion from, it is acknowledged that both models predicted Cl^- to decrease $\log k$ (3 examples for RF and one example for BERT), while predicting I^- to increase $\log k$ (one example for each model). Notably, this does not align with the knowledge that reactivity decreases with increasing halide size in aprotic solvents.

3.2.5.3 Steric effects

To confirm that learning of steric effects was not a source of error for either model, inaccurate predictions with alkyl-substituted centres and unsubstituted centres were evaluated (Figure 3.28). Alkyl-substituted centres are modelled by C-C (11 reactions), C-C-C (5 reactions), and C-C-C-C (1 reaction) fragments in the RF, and the electrophilic C and substituting C atoms (9 reactions) in the BERT model. Reactions with unsubstituted centres were not predicted inaccurately by either model.

Analysis of steric effects shows that substituted centres decreased $\log k_{pred}$ 100% of the time they were high impact for both models, which aligns with the understanding that steric hindrance at the electrophilic centre decreases S_N2 reactivity. The fact that no reactions with unsubstituted centres were predicted inaccurately by either model reinforces that learning of steric effects were not a cause of model error.

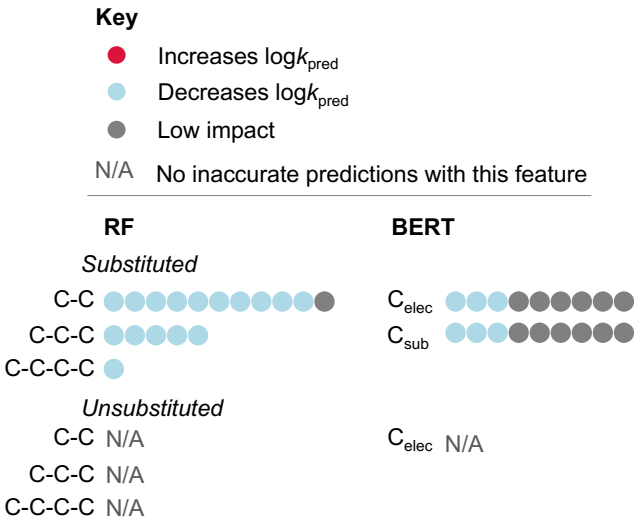


Figure 3.28: Analysis of steric effects (inaccurate predictions). Signs (positive or negative) of feature importances for each of RF and BERT's inaccurate predictions where alkyl-substituted and unsubstituted electrophilic centres were high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no inaccurate predictions containing this feature. C_{elec} and C_{sub} correspond to electrophilic and substituting carbons, respectively.

3.2.5.4 Allylic effects

It is well known that allylic effects from alkene, alkyne, and aromatic bonds (to the electrophilic centre) increase S_N2 reactivity. Here, inaccurate predictions with each of these centres were evaluated to identify whether learning of allylic effects was a source of error for either model (Figure 3.29). Reactions with centres bonded to alkene or alkyne groups were not predicted inaccurately by either model, so the following analysis focuses on centres bonded to aromatic groups, represented by C-C:C fragments in the RF (where ':' is an aromatic bond, 14 reactions), and c tokens in BERT (where c is an aromatic carbon bonded to the centre, 18 reactions). Note the total distribution of centres in the test set was 2 × alkene, 4 × alkyne, and 42 × aromatic.

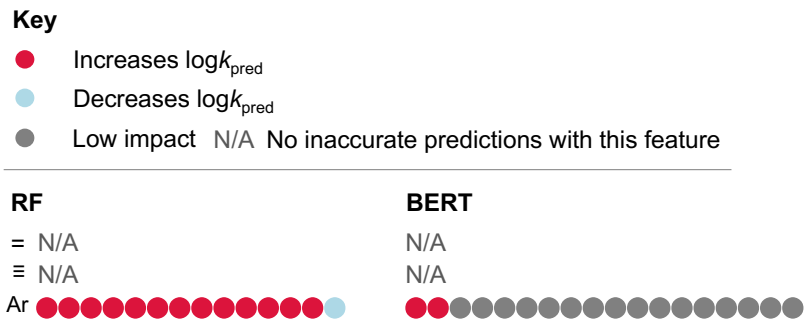


Figure 3.29: Analysis of allylic effects (inaccurate predictions). Signs (positive or negative) of feature importances for each of RF and BERT's inaccurate predictions where alkene (=), alkyne (≡), and aromatic (Ar) bonds to the electrophilic centre were high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no inaccurate predictions containing this feature.

In the RF, aromatic groups increased log k_{pred} in 93% of reactions analysed, affirming the RF had learned this effect. Conversely, aromatic groups were low impact in 89% of BERT's inaccurate predictions, suggesting BERT had not recognised the importance of these groups in enhancing S_N2 reactivity. Importantly, the fact that no reactions with alkene or alkyne groups bound to the centre were predicted inaccurately suggests that learning of allylic effects from these groups did not cause significant error.

3.2.5.5 Solvent effects

Solvent was modelled by 13 properties representing polarity and proticity in the RF model, and by SMILES of the solvent molecules in BERT. For the RF, no clear trend was observed between solvent properties and log k in the experimental data, likely because each solvent differs by multiple properties at once (Justification 2, Appendix C). Hence, solvent effects in the RF have not been evaluated here. As for BERT, solvent effects were evaluated by analysing the total contribution of polar protic ($\varepsilon_{298.15} > 15$ and proton bonded to a heteroatom)[196] and polar

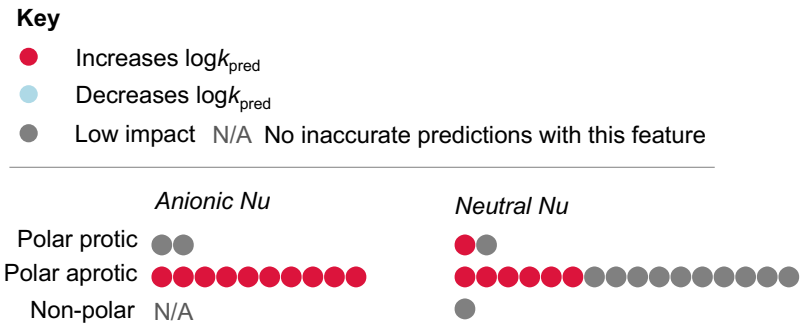


Figure 3.30: Analysis of solvent effects (inaccurate predictions). Signs (positive or negative) of feature importances for each of BERT's inaccurate predictions where the solvent was high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no inaccurate predictions containing this feature.

aprotic ($\varepsilon_{298.15} > 15$ and no proton bonded to a heteroatom)[196] solvent SMILES in inaccurate predictions with anionic and neutral nucleophiles, with the distribution of solvents being 2 × polar protic and 10 × polar aprotic for anionic nucleophiles, and 2 × polar protic and 16 × polar aprotic for neutral nucleophiles (Figure 3.30). Note there was only one test reaction with non-polar solvent ($\varepsilon_{298.15} < 15$),[196] which was predicted inaccurately, and the solvent was low impact.

Polar protic and polar aprotic solvents were both predicted to increase log k in 100% of examples with neutral nucleophiles where solvent was high impact (one of the two reactions with protic solvent and 6 reactions with aprotic solvent), which aligns with the chemical understanding that polar solvent increases S_N2 reactivity in the absence of strong hydrogen bonding to the nucleophile. Polar aprotic solvent was also predicted to increase log k in 100% of examples with anionic nucleophiles, however solvent was a low impact feature for the two reactions with anionic nucleophiles in protic solvent. The latter may have contributed to the erroneous log k prediction for these two reactions (shown in

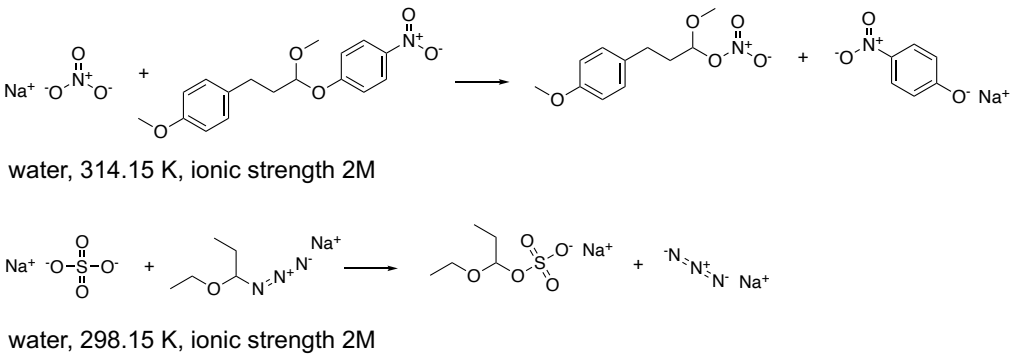


Figure 3.31: Inaccurate predictions with anionic nucleophiles where polar protic solvent was a low impact feature.

Figure 3.31), where solvent hydrogen bonding effects are expected to decrease S_N2 reactivity. Overall, analysis of inaccurate predictions suggests that BERT generally learned solvent effects, however it is possible that hydrogen bonding effects were underestimated in some examples.

3.2.5.6 Summary

Interpreting inaccurate predictions yielded similar results as described for accurate predictions above. Regarding LG effects, both models identified a positive correlation between halide size and S_N2 reactivity across inaccurate predictions, affirming that LG effects were not a source of error for either model. As for steric effects, substituted centres typically decreased $\log k_{\text{pred}}$ in each of RF and BERT's inaccurate predictions, and there were no inaccurate predictions with unsubstituted centres for either model. Hence, recognition of steric effects likely did not contribute to model error. Meanwhile, there were no inaccurate predictions with alkene or alkyne bonds to the electrophilic centre for either model, suggesting recognition of allylic effects from these groups was not a source of error. As for aromatic effects, these were captured by the RF model, but not the BERT model, consistent with the results for accurate predictions. Regarding solvent effects in the BERT model, both polar protic and polar aprotic solvents typically

increased $\log k_{\text{pred}}$ in reactions with neutral nucleophiles, as did polar aprotic solvent in reactions with anionic nucleophiles. However, polar protic solvent was a low impact feature of both inaccurate predictions where it coexisted with an anionic nucleophile. Thus, analysis of inaccurate predictions suggests that BERT generally learned solvent effects, however it is possible that hydrogen bonding effects were overlooked in some examples. Lastly, while there were too few examples to draw definitive conclusions about either model's learning of nucleophilic effects, analysis of available examples suggested that RF identified a positive correlation between halide size and reactivity in protic solvent, aligning with known reactivity rules. Meanwhile, neither model predicted the expected decrease in reactivity with increasing halide size in aprotic solvent.

Overall, analysis of inaccurate predictions indicates that learning of structure-activity relationships was not a significant source of error for either model, with the exception of aromatic effects in the BERT model. Additionally, the agreement between results for accurate and inaccurate predictions affirms the reliability of these evaluations. Two other potential causes of model error, $\log k$ extrapolation and training data representation, are explored in the following section.

3.2.6 Exploring potential limitations

3.2.6.1 log*k* extrapolation

To evaluate how well each model extrapolates to $\log k$ values outside the range of the training data, the relationship between distance in $\log k_{\text{exp}}$ from the training median (training median = $-3.4 \log k$), and prediction error, has been observed for each reaction in the test data (Figure 3.32). Here, the x-axes have been divided into positive distance and negative distance to capture extrapolation to $\log k >$ the training median, and $\log k <$ the training median. For the BERT model, there was little correlation ($r_S = 0.29$, $p_S = 0.00$) between distance from the training

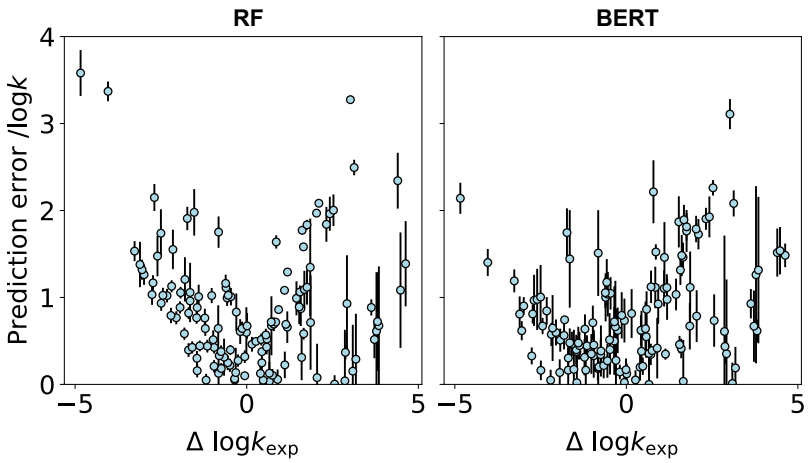


Figure 3.32: Prediction error vs. distance in $\log k_{exp}$ from the training median ($\Delta \log k_{exp}$) for each reaction in the test data, for the RF and BERT models.

median and prediction error, suggesting BERT extrapolates well to $\log k$ values far from the training median. Contrarily, the RF model exhibited a modest correlation ($r_S = 0.40$, $p_S = 0.00$) between absolute distance from the training median and prediction error for $\log k >$ the training median, and a strong positive correlation ($r_S = 0.65$, $p_S = 0.00$) for $\log k <$ the training median. This implies that RF is limited in its ability to extrapolate to $\log k <$ the training median.

In conclusion, both models can extrapolate to $\log k >$ the training median but BERT proved more reliable in extrapolating to $\log k <$ the training median. However, this is to be expected given that BERT has a linear prediction layer, while RF bases its predictions on training averages.

3.2.6.2 Training data representation

One potential limitation of ML models is how well the training data represents the search space. In the current work, two methods are employed to investigate how training data representation affects the accuracy of the RF and BERT models: 1) compare how many training reactions contain well-learned and poorly-learned features (features that were high impact in accurate predictions but not inaccurate

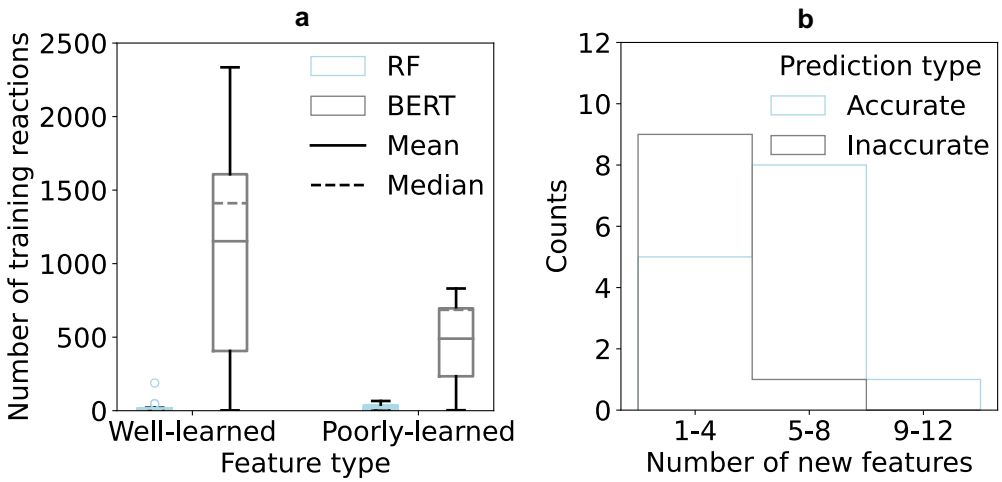


Figure 3.33: Analysis of training data representation. a. The number of training reactions containing each well-learned and poorly-learned feature of the RF and BERT models. **b.** The number of new features (defined as features in the test data that were not in training) in each of RF's accurate and inaccurate predictions.

predictions, and features that were high impact in inaccurate predictions but not accurate predictions, respectively), and 2) compare how many new features (features in the test data that were not in training) were in accurate and inaccurate predictions. The results obtained from these two methods are discussed below.

Well-learned and poorly-learned features. Well-learned and poorly-learned features are provided in Tables 3.2 and 3.3, respectively, with the distribution being 5 × well-learned and 5 × poorly-learned for BERT (5% of model features), and 30 × well-learned and 43 × poorly-learned for the RF (6% and 8% of model features). The number of training reactions containing each of these well-learned and poorly-learned features are presented in Figure 3.33**a**.

In the BERT model, well-learned features were generally better represented in training than poorly-learned features, with well-learned features being in a mean 1152 training reactions (24% of the data), while poorly-learned features were in a mean 490 training reactions (10% of the data). This suggests that better

training representation improves feature learning in the BERT model. This said, BERT can also effectively learn less-represented features, as evidenced by some well-learned features only found in two training reactions. Conversely, in the RF, both well-learned and poorly-learned features were seldom represented during training, with well-learned features being in a mean 16 training reactions (0.3% of the data), and poorly-learned features being in a mean 23 training reactions (0.5% of the data). Hence, there was no correlation between training representation and feature learning in the RF. Overall, these results suggest that RF is more robust towards less-represented features than BERT.

New features. New features are provided in Table 3.4, with the distribution being 37 $\times$ new features in the RF and 1 $\times$ new feature in BERT. The number of new features in accurate and inaccurate predictions is presented in Figure 3.33**b** for the RF model. This comparison couldn't be made for the BERT model because there was only new feature (the [cH-] token) in the test data.

In the RF model, new features were found more frequently in accurate predictions than inaccurate predictions, with 9 accurate predictions containing $\geq$ 5 new features, compared to one inaccurate prediction. This suggests that new features were not a cause of prediction error in the RF model.

The results from analysing new features are unexpected, as these features were frequently found at the reaction centre. Specifically, in all but one of RF's accurate predictions containing new features, $\geq$ 87.5% of the new features were at the centre. Thus, they'd be expected to influence reactivity prediction. However, in each of these examples, other reaction centre features contributed to the prediction (evidenced by 9 – 22% of high impact features being at the centre in each reaction), explaining their accurate prediction.

Overall, the analysis of new features indicates that RF extrapolates effectively to reactions involving features not seen during training. This was attributed to the presence of other reaction centre features that were high impact towards predictions.

Table 3.2: Well-learned features of the RF and BERT models.[a]

RF	BERT
Cl-C:C:N (9)	[NH2+] (4)
Cl-C:C-[N+] (8)	[N+] (3)
Cl-C:C:[N+] (6)	[Na+] (2)
F-C:C-[N+] (5)	9 (2)
F-C:C-N (4)	[PH+] (1)
N:C:C:N (4)	
Br-C-C#C (4)	
[C-]:C:C (3)	
C:[C-]-C (3)	
[C-]-C (3)	
[N+]-C=O (3)	
[N+]:C:C:N (3)	
[As+]-C:C (2)	
C=N (2)	
C-C:C-Cl (2)	
C-[N+]-C=O (2)	
C:C:[C-]-C (2)	
Cl-C:C-[O-] (2)	
C-N=C=O (1)	
C=[N-] (1)	
[N-]=C=O (1)	
Br-C:C:[N+] (1)	
Cl-C:C-O (1)	
C-[As+]-C (1)	
C-C:C-O (1)	
C-C-[N+]-N (1)	
O-C-S (1)	
N=C=O (1)	
N-C-C-O (1)	
Br-C:[N+] (1)	

[a] Numbers in brackets denote the number of accurate predictions each well-learned feature was high impact for.

Table 3.3: Poorly-learned features of the RF and BERT models.[a]

RF	BERT
[O-]-[N+]-[O-] (7)	[c-] (2)
C-O-[N+]-[O-] (7)	[S-] (2)
[N+]-O (7)	8 (1)
O-[N+]=O (7)	6 (1)
I-C-C-N (5)	7 (1)
C-C-P-C (4)	
C-P-C (4)	
Cl-C-C-Cl (4)	
C-P (4)	
Cl-C-C-[N+] (4)	
C-[P+] (4)	
C-C-C#N (3)	
C-C-[P+]-C (3)	
C-C-P (3)	
C-[P+]-C (3)	
N#C-C-N (3)	
N:C:C-N (3)	
[N+]-C-C#N (3)	
[N+]-C-C-N (3)	
C:C-[P+]-C (2)	
C-C-[P+] (2)	
C:C-P-C (2)	
C-C-[I+] (2)	
C:C-P (2)	
N-C-C=O (2)	
C-C-S-C (2)	
C-C-C-[P+] (2)	
C-C-C-P (2)	
C=C-C-[O+] (1)	
I-C-C-[O-] (1)	
C-O-C-[N+] (1)	
Br-C-C-[O-] (1)	
C-N=C (1)	
I-C-C-S (1)	
C:C:C-P (1)	
C=C-Cl (1)	
Br-C-C-I (1)	
C-[I+] (1)	
C:C:C-[S-] (1)	

C-[N+]-N (1)
O-C-C-S (1)
C:C:C-[P+] (1)
C:C-[P+] (1)

[a] Numbers in brackets denote the number of inaccurate predictions each poorly-learned feature was high impact for.

Table 3.4: New features of the RF and BERT models.

RF	BERT
C-N:C-N	[cH-]
C-N:C:N	
N:C-N	
C:N:C-N	
C-[N+]:C-N	
[N+]:C-N	
C-N:C:[N+]	
C:[N+]:C-N	
O-C-O-S	
[N-]-C:C-O	
C-[C-]-C	
C-[C-]-C=O	
C-[C-]-C-O	
[C-]-C=O	
[C-]-C-O	
[C-]-C-O-C	
C-N:C=O	
[C-]:C:N-C	
N:C=O	
[C-]:C:N	
C:[C-]:C:N	
C:[C-]:C=O	
C:N:C=O	
[C-]:C=O	
[C-]:C:N:C	
C-N-C-N	
[N+]-C-[N+]	
[N+]-C-[N+]=O	
[N+]-C-[N+]-[O-]	
N-C-C-[O+]	
[N+]-C-[O+]	

[O+]-C-[N+]=O
[O+]-C-[N+]-[O-]
C-N-C-S
N-C-S
C-S-C-N
C-N-C-[N+]

3.3 Conclusion

In this work, a BERT model was trained to predict experimental log k values of S_N2 reactions and compared its performance to the state-of-the-art RF model from Madzhidov et al.[33] in terms of accuracy, training time, and ability to replicate known reactivity rules. In addition, the dataset of 4830 S_N2 reactions curated by Madzhidov et al.,[29] used to train their RF,[33] was diversified in the current work to expand the model's generalisability to a wider area of chemical space.

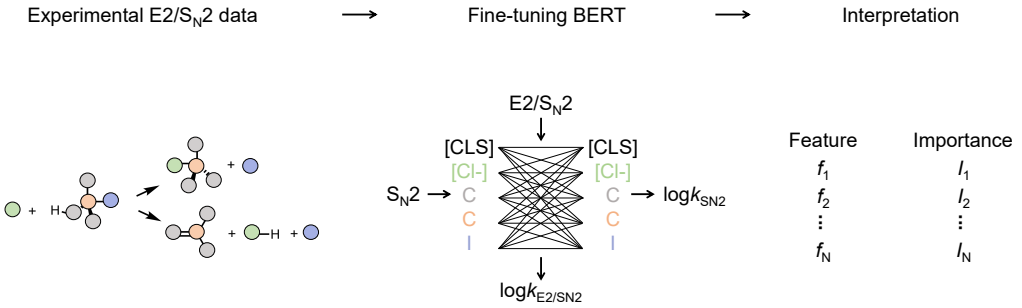
The findings presented here show that increasing the chemical diversity in Madzhidov's training data did indeed broaden the RF's applicability to new areas of chemical space, largely for reactions in electrolyte solution. When compared to a BERT model also trained on the diversified data, both RF and BERT achieved the same prediction accuracy ($\approx$ chemical accuracy), however RF did so significantly faster. Additionally, both models identified key reaction centre atoms and physical features, as well as known LG, steric, allylic, temperature, and solvent effects (BERT only, solvent effects could not be interpreted in the RF), reinforcing the validity of their predictions.

Regarding model strengths and limitations, RF was more robust towards recognising aromatic effects, and underrepresented training features, while BERT was more adept in log k extrapolation. Additionally, it was highlighted that

ISIDA fragment features in the RF should be restricted to the reaction centre to prevent learning of spurious correlations.

The conclusions presented here demonstrate that RF and BERT are both effective models for reactivity prediction, validated by their ability to replicate known chemical principles. Additionally, each model was found to overcome the other's limitations, providing an overall robust predictive framework. Future work should focus on expanding the applicability of these models to a wider range of chemical reactions.

4 | Interpretable predictions of E2/S_N2 kinetics using the BERT architecture: Analysing drivers of mechanistic competition



Abstract

The accurate prediction of reaction rates is an integral step in reaction mechanism elucidation and design of synthetic pathways. Machine Learning (ML) facilitates fast, accurate prediction of experimental rate constants, overcoming limitations in traditional Quantum Mechanics (QM) methods. Transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) are a popular ML method for chemical prediction, allowing reactions to be represented as SMILES strings. In Chapter 3, a BERT model was trained to predict experimental log k values of S_N2 reactions, and its predictions were interpreted in the context of known reactivity rules. In Chapter 4, this model is fine-tuned to predict experimental log k values of both E2 and S_N2 reactions, and its predictions are interpreted regarding known drivers of mechanistic competition. The E2/S_N2 BERT exhibits an RMSE of 1.2 ± 0.1 log k on similarity-split test data, maintaining the performance of BERT models trained on each individual mechanism, and exceeding the accuracy of the best-performing literature model. Additionally, the E2/S_N2 BERT predicts the favoured pathway of competing reactions with 60% accuracy, and achieves 91% accuracy predicting the outcome of adversarial examples. Moreover, the E2/S_N2 BERT successfully identifies key structural and physical effects such as LG, steric, and temperature effects (for temperatures within the training IQR). Model limitations are also addressed, including recognition of aromatic effects, and extrapolation to temperatures outside the training IQR.

4.1 Methodology

Unless specified below, all methodology was consistent with that outlined in 3.1.

4.1.1 Data

The training and test data utilised in this work can be visualised in Figure 4.1. The training data contains 4757 S_N2 reactions with 449 unique nucleophiles, 288 unique substrates, 153 unique solvent systems, a temperature range of 232.15 to 425.15 K, and a log k range of -12.3 to 1.6. The training data also contains 1591 E2 reactions with 50 unique bases, 390 unique substrates, 74 unique solvent systems, a temperature range of 201.15 to 445.95 K, and a log k range of -7.2 to 2.7. The training distribution of temperatures and log k values can be visualised in Figure F.1 (Appendix F.1). The test data contains 127 S_N2 reactions with 41 unique nucleophiles, 42 unique substrates, 10 unique solvent systems, a temperature range of 252.15 to 461.00 K, and a log k range of -8.2 to 1.2. The test data also contains 87 E2 reactions with 32 unique nucleophiles, 49 unique substrates, 11 unique solvent systems, a temperature range of 273.15 to 464.35 K, and a log k range of -6.7 to -0.2. All test data has an $S_T < 0.4$ to the training data, so prediction accuracy reflects model performance on novel reactions.

The training data utilised in this work is comprised of 4757 of the S_N2 training reactions from Chapter 3 (after excluding those with a competing E2 reaction in the dataset), and 1591 E2 reactions compiled by Madzhidov et al.[27] Meanwhile, the test data is comprised of 127 of the S_N2 test reactions from Chapter 3 (after excluding those with a competing E2 reaction in the dataset), and 87 E2 reactions compiled by Madzhidov et al.[27] S_N2 reactions with a competing E2 reaction (in the dataset) were excluded to avoid imbalance between reactions represented by both competing pathways (107 reactions) and reactions represented by a single pathway

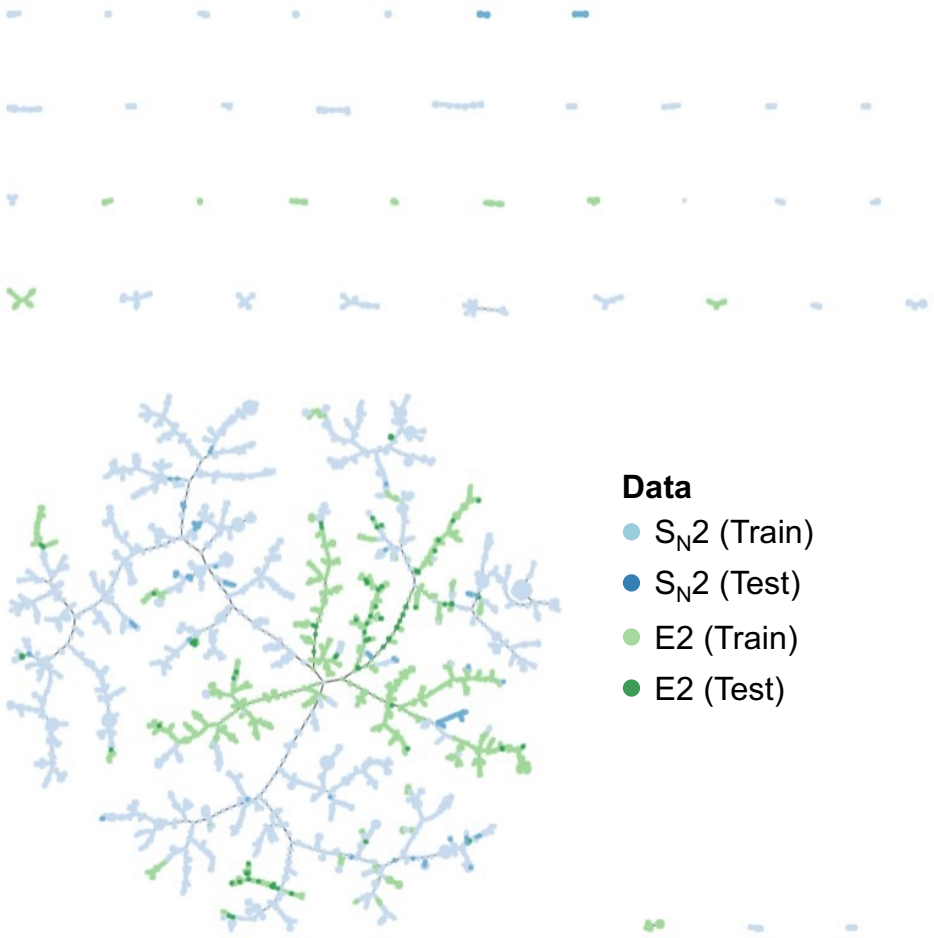


Figure 4.1: Tree maps (TMAPs) of the training and test data utilised in this work. Shown in light colours: The total training set of 4757 S_N2 reactions (light blue) and 1591 E2 reactions (light green). The former correspond to the training data from Chapter 3 (after excluding reactions with a competing E2 reaction in the dataset), and the latter were compiled by Madzhidov et al.[27] Shown in dark colours: The total test set of 127 S_N2 reactions (dark blue) and 87 E2 reactions (dark green). The former correspond to the test data from Chapter 3 (after excluding reactions with a competing E2 reaction in the dataset), and the latter were compiled by Madzhidov et al.[27] Interactive versions of each TMAP are provided on GitHub (see *Data availability*).

(6481 reactions). Competing reactions were omitted from the S_N2 data, rather than the E2 data, due to there being a greater number of S_N2 reactions in the dataset.

The E2 data (train + test) compiled by Madzhidov et al.[27] (originally spanning 1820 reactions and their experimental $\log k$ values) was cleaned in the current work to remove unbalanced reactions (93 reactions), duplicates (7 reactions), and known CV outliers[27] (14 reactions). 26 reactions with a competing S_N2 reaction were also removed from the E2 data for use as an auxiliary test set (Appendix F.2). The remaining 1680 E2 reactions were divided into training and test sets using a random 50/50 shuffle split, followed by removal of test reactions with $S_T > 0.4$ to the training data, and their addition into the training set. A further two test reactions with $S_T > 0.4$ to the new training data (now including the newly added reactions) were removed from the dataset entirely, resulting in 1591 E2 train and 87 E2 test reactions in total.

It is noted that the number of S_N2 reactions in the training data (4757) exceeds the number of E2 reactions (1591), however model accuracy was unaffected by this difference in dataset size (Appendix F.3).

4.1.2 ML

The BERT model trained for S_N2 rate prediction in Chapter 3 ($BERT_{S_N2}$) was fine-tuned in this study to predict experimental $\log k$ values for both E2 and S_N2 reactions, using the dataset of 4757 S_N2 and 1591 E2 reactions outlined in Section 4.1.1. This fine-tuning process was consistent with the methodology used to train $BERT_{S_N2}$, outlined in Section 3.1.2.

The E2/S_N2 BERT is compared to a BERT model trained solely on the dataset of 1591 E2 reactions compiled by Madzhidov et al.[27] ($BERT_{E2}$). Here, the BERT

model developed by Laino et al.[60] (originally trained for reaction classification) was fine-tuned on the E2 data, using the methodology outlined in Section 3.1.2.

The E2/S_N2 BERT is also compared to an RF model trained on the E2/S_N2 data. This RF model was trained following the methodology outlined in Section 3.1.2, but ISIDA fragments without a substructure match to one of the reaction centre atoms (Nu, LG, electrophilic C or one of their product atom mappings for S_N2; and B, LG, electrophilic C, β-C, or one of their product atom mappings for E2) were omitted from the feature set. This approach has been shown to prevent spurious correlations without compromising model accuracy (Section 3.2.4.3).

4.1.3 Interpretation

In addition to the methodology outlined in Section 3.1.5, the following modifications were applied in this study.

4.1.3.1 Analysing structural and physical effects

Identifying reaction centre atoms. The reverse E2 reaction is not supported in autodE (version 1.1),[184] so mappings to reaction centre atoms in the products of E2 reactions were identified using RDKit (version 2021.03.5).[198]

Alkyl substituent effects. In Chapter 3, steric effects in S_N2 reactions were analysed using reactions with alkyl-substituted electrophilic centres. This analysis was extended to E2 reactions in the current work by also considering the electron-donating properties of these substituents. Unlike Chapter 3, which compared substituted and unsubstituted centres, this work analyses primary, secondary, tertiary, and unsubstituted centres. This approach better reflects trends in E2/S_N2 reactivity, where preference for the E2 reaction increases with increasing substitution, while preference for the S_N2 reaction decreases with

increasing substitution. Consistent with Chapter 3, alkyl-substituted centres were modelled by tokens of electrophilic and substituting C atoms. The β-C was not considered a substituting C for the E2 mechanism as it does not contribute to the electron-donating effects being analysed.

Symmetrical substrates. A subset of the examples analysed in this work involve symmetrical substrates with two or more equivalent LG, electrophilic C, or substituting C atoms. In these cases, feature importances were averaged over the equivalent atoms. This was not relevant to Chapter 3, as none of the analysed examples involved symmetrical substrates.

Note that while the β-C was not considered a substituting C in the analysis of alkyl-substituent effects (it does not contribute to the electron-donating effects being analysed), for symmetrical substrates with two or more equivalent β-Cs, the average feature importance of these atoms was included to account for their average electron-donating effect.

Synthetic data points (temperature effects). To analyse temperature effects in E2 reactions, reactivity series from the E2 test data were employed. Each reaction was observed at three temperatures, and 10 synthetic data points were generated to augment each series. To generate the synthetic data, 10 temperatures were randomly selected from a uniform distribution between T_{min} - 25 K and T_{max} + 25 K, where T_{min} and T_{max} are the minimum and maximum observed temperatures in the reactivity series, respectively.

Products. Each atom discussed in the analysis of structural and physical effects also has a mapped atom in the products. The feature importances of these mappings are discussed in Appendix F.7.

4.1.4 Comparison to Green et al.

The E2/S_N2 BERT trained in this work is compared to the best-performing E2/S_N2 rate prediction model in the literature, developed by Green et al.[34] Green's model utilises a CGR D-MPNN architecture and is trained on a dataset of 1264 E2 reactions and 2361 S_N2 reactions (in 5 × 80/10/10 train/validation/test splits), calculated at the DF-LCCSD/cc-pVTZ//MP2/6-311G(d) LOT.

The models were compared using the RMSE reported by Green et al. (in kcal mol^{-1}), which was converted to $\log k$ using,

$$\mathrm{RMSE}_{\mathrm{logk}} = \log e^{\frac{4184\ \mathrm{RMSE}_{\mathrm{kcal}}}{RT}} \tag{4.1}$$

where $\mathrm{RMSE}_{\mathrm{logk}}$ and $\mathrm{RMSE}_{\mathrm{kcal}}$ are the RMSE in $\log k$ and kcal mol^{-1}, respectively, R is the gas constant, and T is the temperature, which was set to 298.15 K. The uncertainty estimate provided by Green et al. (expressed in standard deviations) was converted to σ_M using Equation 3.3 with a population size (n) of 5, corresponding the the number of estimators. To assess whether the difference in RMSE between the two models was statistically significant, a Mann-Whitney U test[199] was employed. This is the most suitable test for this comparison, where the distributions are numerical, non-parametric, and unpaired (i.e., the models are trained on different datasets).[193] The RMSE of each estimator was not reported for Green's model, so this distribution was estimated using the mean value, standard deviation, and population size (= 5). Overall, 100 possible distributions were generated from which the mean Mann-Whitney U p-value was calculated (which itself had a σ_M of 0.00). Here, the p-value was calculated using a permutation test (number of resamples = 9999) to account for the small number of samples (5) in the distribution.

To assess whether the performance gap between the two models resulted from differences in architecture or training data, the BERT model was retrained using Green's dataset. No reactions were removed from Green's data in this work, and CV splits were kept consistent with those used by Green et al. However, stereochemistry, atom map numbers, and explicit hydrogens were removed from the reaction SMILES for consistency with the format used in the current work. The SMILES were also reordered according to the grammar outlined in Figure 3.3. Note that temperature and solvent were not included in Green's dataset, so the SMILES contain reactant and product molecules only. The target value was kept as $\Delta E^{\ddagger}$ for consistency with Green's model, and the resulting RMSE was converted to $\log k$ using Eq. 4.1.

The BERT model developed by Laino et al.[60] (originally trained for reaction classification) was trained directly on Green's data, using the methodology outlined in Section 3.1.2. This approach was taken rather than fine-tuning $\text{BERT}_{\text{S}_\text{N}2}$ due to inconsistencies in reaction representation between Green's dataset and the data used to train $\text{BERT}_{\text{S}_\text{N}2}$, such as differing target values and the inclusion of temperature and solvent in the latter. Further details pertaining to the model's training are provided in Appendix F.5.

To assess whether the difference in RMSE between Green's model, and the BERT model trained on Green's dataset, was statistically significant, a Wilcoxon's matched-pairs signed-rank test[192] was employed following the methodology outlined in Section 3.1.4. This is the most suitable test for this comparison, where the distributions are numerical, non-parametric, and paired (i.e., the models are trained on the same dataset).[193] The RMSE of each estimator was not reported for Green's model, so this distribution was estimated using the mean value, standard deviation, and population size (= 5). Overall, 100 possible distributions

were generated from which the mean Wilcoxon's p-value was calculated (which itself had a σ_M of 0.01).

4.1.5 Literature datasets

The E2/S_N2 BERT was tested on three literature datasets: 1) 10 competing E2/S_N2 reactions from Ref[200] (Figure 4.2**a**), 2) 11 adversarial examples from Ref[201] which proceed via the S_N2 mechanism but not the E2 mechanism (Figure 4.2**b**), and 3) five examples of Hofmann/Zaystev regioselectivity from Refs[202,203] (4 reactions), and the E2 test data (1 reaction) (Figure 4.2**c**). For each reaction, BERT was tasked with predicting the favoured pathway by estimating $\log k$ for both mechanisms. If the favoured pathway had the lower $\log k_{\text{pred}}$, BERT was considered to have predicted it correctly. Furthermore, if the predicted k ratio, $\left(\frac{k_{\text{E2}}}{k_{\text{S}_\text{N}2}}\right)_{\text{pred}}$, was within one order of magnitude of the experimental k ratio, $\left(\frac{k_{\text{E2}}}{k_{\text{S}_\text{N}2}}\right)_{\text{exp}}$, BERT was considered to have predicted this ratio correctly. Given the already limited data sizes, an $S_T < 0.4$ constraint was not applied to the literature datasets. However, particular attention is given to the model's performance on reactions with $S_T < 0.4$ to the training data within each dataset. Note that only reactions with a significant $\log k_{\text{exp}}$ difference ($\Delta \log k_{\text{exp}} \geq 0.1$) were included in each dataset. Specific details pertaining to each dataset are provided below.

The Competing dataset contains 10 E2/S_N2 reactions testing three key factors that drive mechanistic competition: 1) nucleophile/base effects, comparing methoxide to *tert*-butoxide, 2) LG effects, comparing tosylate to Br, and 3) aromatic effects at the β-C. The dataset spans 10 nucleophile/substrate combinations across two different solvents (methanol and *tert*-butanol) and two different temperatures (338.15 and 353.15 K). All nucleophile/substrate combinations were unseen during training, with three having an $S_T < 0.4$ to the training data. To evaluate BERT's learning of the factors driving E2/S_N2 competition, key features were analysed.

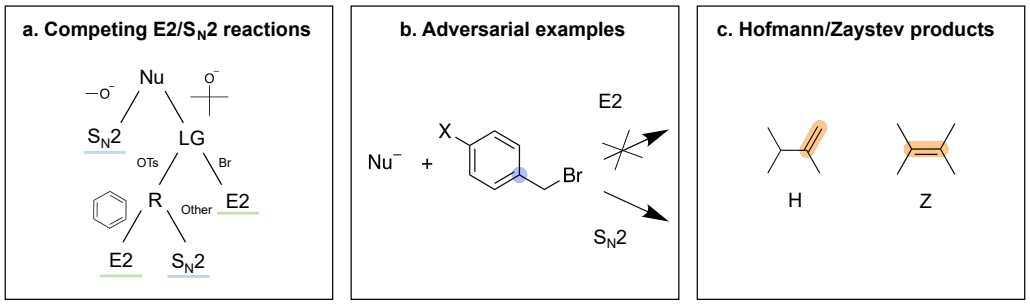


Figure 4.2: Literature datasets used to test the E2/S_N2 BERT model. **a.** 10 competing E2/S_N2 reactions from Ref.[200] These reactions test three effects driving E2/S_N2 competition: 1) nucleophile/base effects , 2) LG effects, and 3) aromatic effects. **b.** 11 S_N2 reactions from Ref.[201] These reactions, which have no β-Hs, cannot proceed through the E2 mechanism, testing BERT's ability to handle adversarial examples. **c.** Five E2 reactions from Refs[202,203] (4 reactions), and the E2 test data (1 reaction). These reactions result in both Hofmann (least substituted) and Zaystev (most substituted) products, testing BERT's recognition of regioselectivity.

Here, the difference in feature importance between the two pathways (ΔIG) was evaluated for relevant functional groups, including methoxide nucleophiles, the *tert*-butyl group of *tert*-butoxide (representing the nucleophile's bulk), tosylate and Br LGs, and the aryl ring of aromatic groups bonded to the β-C. To obtain the importance for each functional group, the IGs of each SMILES token constituting that group were summed.

The Adversarial dataset contains 11 S_N2 reactions of benzyl bromide derivatives, corresponding to four nucleophile/substrate combinations across two different solvents (DMF and ethanol) and a temperature range of 232.15 to 298.15 K. Two of these nucleophile/substrate combinations were present in the training data as S_N2 reactions. However, the adversarial dataset is designed to test BERT's handling of the artificial E2 mechanism, which was not encountered during training.

The Hofmann/Zaystev dataset contains five examples of Hofmann/Zaystev regioselectivity in E2 reactions, corresponding to five nucleophile/substrate combinations across two different solvents (methanol and ethanol). Temperatures were

not provided for four of these examples, so predictions were made at four different temperatures ranging from 273.15 to 348.15 K in these cases. All nucleophile/substrate combinations in the Hofmann/Zaystev dataset were unseen during training, with two having an $S_T < 0.4$ to the training data. To evaluate BERT's learning of the factors driving Hofmann/Zaystev regioselectivity, key features were analysed. Here, the difference in feature importance between the two pathways (ΔIG) was evaluated for relevant functional groups, including substrate Me branches, the *tert*-butyl group of *tert*-butoxide (representing the nucleophile's bulk), and trimethylamine LGs. To obtain the importance for each functional group, the IGs of each SMILES token constituting that group were summed.

4.2 Results and Discussion

4.2.1 Performance of the E2/S_N2 BERT

The BERT model trained for S_N2 rate prediction in Chapter 3 was fine-tuned in this study to predict experimental $\log k$ values for both E2 and S_N2 reactions. This was achieved using a combined dataset of 1591 E2 reactions compiled by Madzhidov et al.[27] and 4757 of the S_N2 training reactions from Chapter 3, after excluding those with a competing E2 reaction in the dataset. The fine-tuned model achieved RMSE convergence in significantly fewer epochs than training directly on the E2/S_N2 data (Appendix F.4), highlighting the effectiveness of this approach.

To evaluate whether prediction accuracy is maintained when combining the E2 and S_N2 data, the performance of the E2/S_N2 BERT was compared to the BERT model from Chapter 3 ($BERT_{S_N2}$), and a BERT model trained solely on the 1591 E2 reactions compiled by Madzhidov et al.[27] ($BERT_{E2}$). Notably, the accuracy of the E2/S_N2 BERT was equivalent to that of the BERT models trained on each individual mechanism, achieving an RMSE of 1.1 ± 0.1 $\log k$ when evaluated

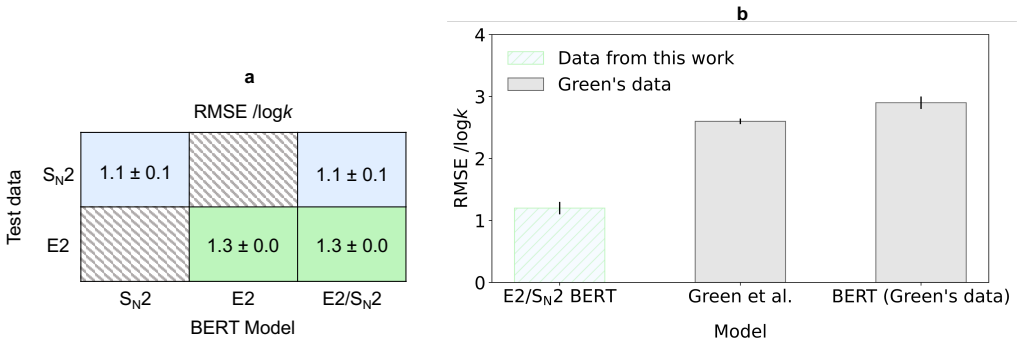


Figure 4.3: Performance of the E2/S_N2 BERT. a. Compared to the BERT model from Chapter 3 ($BERT_{S_N2}$) and a BERT model trained on the 1591 E2 reactions compiled by Madzhidov et al.[27] ($BERT_{E2}$). The former was evaluated using 127 S_N2 test reactions, while the latter was evaluated using 87 E2 test reactions. **b.** Compared to the best-performing E2/S_N2 rate prediction model from the literature developed by Green et al.,[34] and a BERT model trained using Green's dataset. Here, the E2/S_N2 BERT was evaluated using the combined E2 + S_N2 test set (214 reactions), while the RMSE of Green's model was taken from Ref[34] and converted to log k using Eq 4.1. The BERT model trained on Green's data was evaluated using the methodology outlined in Ref.[34]

using S_N2 test data (127 reactions), and 1.3 ± 0.0 log k when evaluated using E2 test data (87 reactions) (Figure 4.3**a**). This suggests that combining the E2 and S_N2 data enhances model generalisability without compromising performance.

When evaluated on the combined E2 + S_N2 test set (214 reactions), the E2/S_N2 BERT achieved an RMSE of 1.2 ± 0.1 log k, outperforming the leading E2/S_N2 rate prediction model in the literature (a CGR D-MPNN developed by Green et al.[34]) by 1.4 ± 0.1 log k (Figure 4.3**b**). This difference was confirmed as statistically significant using a Mann-Whitney U test (p = 0.01). To assess whether the performance gap is attributed to differences in architecture or training data, BERT was retrained using Green's dataset (≡ 1264 E2 reactions and 2361 S_N2 reactions in 5 × 80/10/10 train/validation/test splits). When trained on Green's data, BERT scored an RMSE of 2.9 ± 0.1 log k, significantly larger than when it was trained on the 1591 E2 reactions and 4757 S_N2 reactions from the current work. This suggests that the performance gap between Green's model

and the E2/S_N2 BERT is primarily due to the quality of the training data, with the former trained on QM-generated data, and the latter on experimental data. Notably, both architectures performed comparably when trained on the same dataset ($\Delta_{\text{DMPNN-BERT}}$ RMSE = 0.3 ± 0.1 log k, Wilcoxon's p = 0.10). Note that different statistical tests have been used when comparing models trained on the same dataset (Wilcoxon's signed rank test, paired data), and models trained on different datasets (Mann-Whitney U test, unpaired data).

In the current work, reactions with S_T > 0.4 to the training data were omitted from the test data, so the results seen here reflect the BERT's generalisability to unseen chemistry. This constraint was not applied to Green's data, to maintain consistency with the methodology used in their study.

4.2.2 Interpreting accurate and inaccurate predictions

Test reactions predicted accurately (25% smallest prediction error) and inaccurately (25% largest prediction error) by the E2/S_N2 BERT can be visualised in Figure 4.4**a**. The ratio of E2:S_N2 reactions was 43:57 for both accurate and inaccurate predictions. This is comparable to the ratio of E2:S_N2 reactions in the test data (41:59), suggesting model performance is consistent across both mechanisms. This is also reflected in the error distributions for each mechanism, which have a similar MAE (S_N2: 0.8 log k, E2: 0.9 log k) and IQR (S_N2: 0.3 to 1.1 log k, E2: 0.3 to 1.2 log k) (Figure 4.4**b**).

To assess whether BERT effectively learned the key structural and physical factors driving E2/S_N2 competition, this work evaluates whether high impact features (those with a feature importance in the top 25% across the test data) of accurate and inaccurate predictions aligned with known reactivity rules, including LG, steric, allylic, and temperature effects. To analyse structural and physical effects,

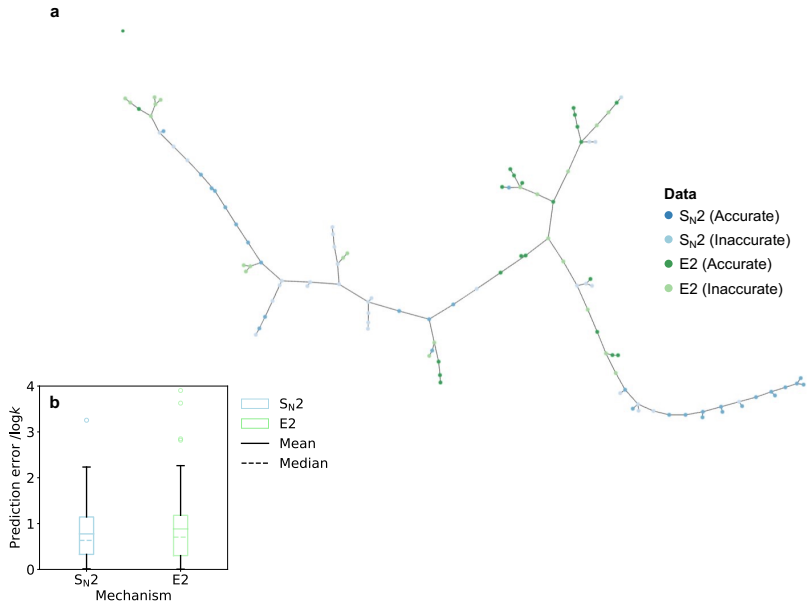


Figure 4.4: Prediction error analysis for the E2/S_N2 BERT. a. Accurate and inaccurate predictions of the BERT model for E2 (shown in green) and S_N2 (shown in blue) reactions, visualised as TMAPs. Here, accurate and inaccurate predictions are defined as those with the 25% smallest, and largest, error in log k, respectively. Note that the peripheral data clusters have been artificially repositioned closer to the main branch to fit within the frame. The original interactive versions of each TMAP are provided on GitHub (see *Data availability*). **b.** Distribution of (absolute) prediction errors on the E2 and S_N2 test data for the BERT model.

the feature importances of key reaction centre atoms were evaluated relative to a dummy model, where feature importance is zero.

4.2.2.1 LG effects

To begin the analysis, halide LGs (represented by I, Br, Cl, and F tokens) were evaluated as a control. While these groups aren't significant drivers of E2/S_N2 competition (reactivity increases with halide size in both mechanisms), they serve as a baseline to test BERT's recognition of reactivity trends when competition isn't a factor. The feature importances of halide LGs contributing to the predicted rate, either increasing (positive sign) or decreasing (negative sign) log k_{pred} are presented in Figure 4.5 for BERT's accurate predictions, with the distribution being

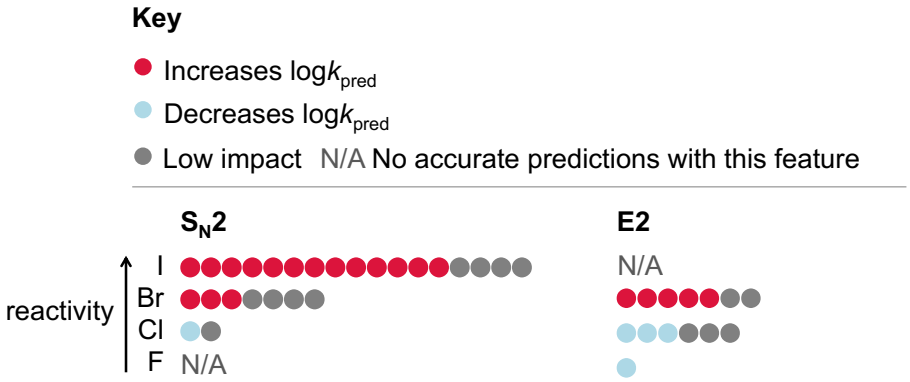


Figure 4.5: Analysis of LG effects. Signs (positive or negative) of feature importances for each of BERT's accurate predictions where the LG was high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature.

17 × I, 7 × Br, 2 × Cl, and 0 × F for S_N2 and 0 × I, 7 × Br, 6 × Cl, and 1 × F for E2. The corresponding analysis for inaccurate predictions is provided in Appendix F.6.1. Note that the total distribution of halide LGs in the test set was 53 × I, 16 × Br, 16 × Cl, and 4 × F for S_N2 and 3 × I, 22 × Br, 22 × Cl, and 2 × F for E2.

Analysis of accurate predictions reveals that I and Br were predicted to increase log k in 100% of S_N2 reactions where they were high impact (13, and 3 reactions, respectively), while Cl decreased log k_{pred} in the one S_N2 example where it was high impact. This suggests that BERT identified a positive correlation between halide size and S_N2 reactivity, consistent with the $BERT_{S_N2}$ model discussed in Chapter 3. Of the four S_N2 test reactions with an F LG, none were predicted accurately by E2/S_N2 BERT. Two of these reactions were predicted inaccurately, but a decrease in log k_{pred} was attributed to the F LG, consistent with this group decreasing S_N2 reactivity (Figure F.6, Appendix F.6.1). Thus, it is unlikely that the inaccurate prediction of these reactions is due to BERT's learning of LG effects.

A positive correlation between halide size and reactivity was also observed in the accurate predictions of E2 reactions. Here, BERT predicted Br to increase $\log k$ in 100% of examples where it was high impact (5 reactions), while Cl and F decreased $\log k_{\text{pred}}$ in 100% of examples where they were high impact (3 reactions, and one reaction, respectively). Of the three E2 test reactions with an I LG, none were predicted accurately by E2/S_N2 BERT. Two of these reactions were predicted inaccurately, but an increase in $\log k_{\text{pred}}$ was attributed to the I LG, consistent with this group increasing E2 reactivity (Figure F.6, Appendix F.6.1). Thus, it is unlikely that the inaccurate prediction of these reactions is due to BERT's learning of LG effects.

Overall, the E2/S_N2 BERT successfully recognised key LG effects across accurate predictions of both E2 and S_N2 reactions. This trend was also observed across BERT's inaccurate predictions (Figure F.6, Appendix F.6.1), reinforcing the reliability of these assessments.

4.2.2.2 Alkyl substituent effects

Alkyl substituents at the electrophilic centre are known drivers of E2/S_N2 competition. Steric crowding from these groups hinders nucleophilic attack, decreasing S_N2 reactivity, while their electron-donating properties stabilise alkene formation in the E2 TS, increasing E2 reactivity.[81,82] To analyse alkyl substituent effects, accurate predictions with primary, secondary, tertiary, and unsubstituted centres were evaluated (Figure 4.6), with the distribution being 8 × unsubstituted, 11 × primary, 1 × secondary, and 0 × tertiary for S_N2 and 2 × primary, 4 × secondary, and 3 × tertiary for E2 (unsubstituted centres, with no β-Hs, cannot undergo E2). The corresponding analysis for inaccurate predictions is provided in Appendix F.6.2. Note that the total distribution of electrophilic centres in the test set was 24 × unsubstituted, 36 × primary, 3 × secondary, and 0 × tertiary for S_N2

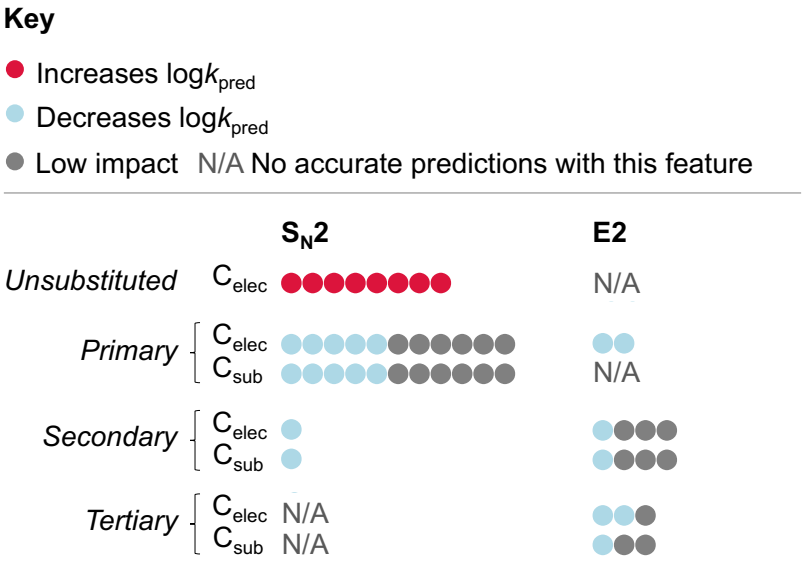


Figure 4.6: Analysis of alkyl substituent effects. Signs (positive or negative) of feature importances for each of BERT's accurate predictions where primary, secondary, tertiary, and unsubstituted electrophilic centres were high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature. C_{elec} and C_{sub} correspond to electrophilic and substituting carbons, respectively.

and 4 × primary, 19 × secondary, and 12 × tertiary for E2. In the BERT model, alkyl-substituted centres are represented by the electrophilic C and substituting C atoms, while unsubstituted centres are solely represented by the electrophilic C. The β-C was not considered a substituting C in the E2 mechanism as it does not contribute to the stabilisation of the forming alkene bond in the TS.

Analysis of accurate predictions reveals that substituted centres decreased log k_{pred} in 100% of S_N2 reactions where they were high impact (5 × primary centres and 1 × secondary centre), while unsubstituted centres increased it. This is consistent with the $BERT_{S_N2}$ model discussed in Chapter 3, indicating that BERT maintained its recognition that steric hinderance decreases S_N2 reactivity, even with the competing E2 mechanism in the training data.

Regarding accurate predictions of E2 reactions, BERT predicted primary, secondary, and tertiary substituted centres to either decrease or have no effect on $\log k$ in 100% of analysed examples. This result is surprising, as E2 reactivity is expected to increase with increasing alkyl substitution. To test whether BERT had overgeneralised its learned relationship between substituted centres and decreased reactivity from S_N2 reactions to E2 reactions, alkyl substituent effects were analysed in $BERT_{E2}$. However, the results were consistent with the E2/S_N2 model, showing either no effect or decreased $\log k_{pred}$ for all substituted centres analysed (Figure 4.7**a**). Hence, the association between alkyl substituents and reduced E2 reactivity cannot be attributed to over-generalisation of S_N2 trends. A similar association[a] was found in an RF model trained on the E2/S_N2 dataset (where substituted centres are represented by C-C, C-C-C, and C-C-C-C fragments at the electrophilic centre, Figure 4.7**b**), indicating a consistent relationship across architectures. To investigate the origin of this relationship, E2 reactivity series in the training data (28 series total) were analysed, revealing a negative correlation ($r_S < 0$) between the substrate van der Waal's volume (a proxy for steric bulk) and $\log k_{exp}$ in 71% of cases. This suggests that BERT's association between alkyl substituents and decreased E2 reactivity reflects a genuine experimental trend linking increased steric crowding to a lower $\log k_{exp}$.

In conclusion, BERT correctly identified that S_N2 reactivity is decreased by steric hindrance at the electrophilic centre. Meanwhile, while BERT's modelling of alkyl substitutents and E2 reactivity did not match the expected promotion of E2 reactivity through alkene bond stabilisation,[81,82] it was consistent with an experimental correlation between increased steric bulk and decreased reactivity in these reactions. The trends discussed here were also observed across BERT's

[a] Note that the three reactions in Figure 4.7**b** where C-C-C-C was attributed to an increase in $\log k_{pred}$ all share the same substrate (*cyclo*-hexylbromide), suggesting this result is substrate-specific rather than indicative of a general trend.

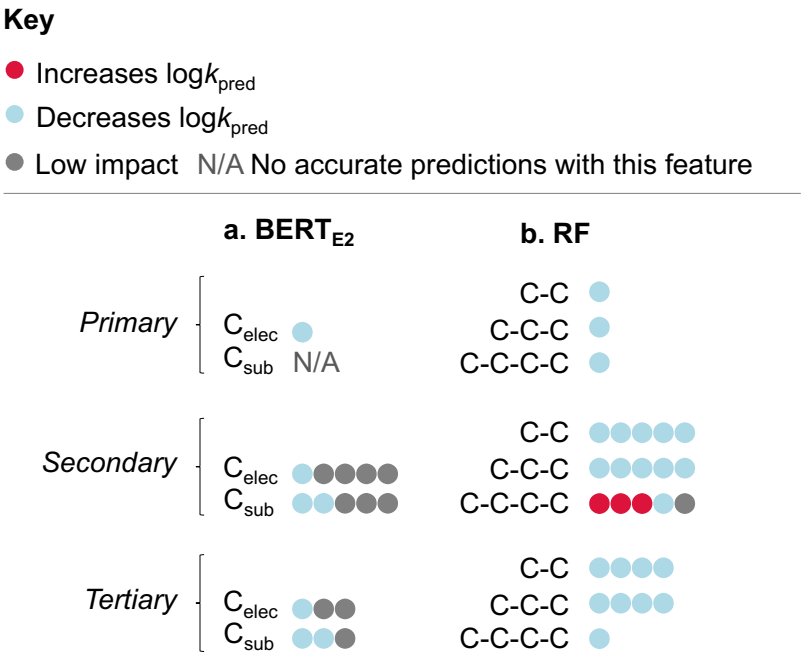


Figure 4.7: Analysis of alkyl substituent effects in E2 reactions (BERT$_{E2}$ and RF). Signs (positive or negative) of feature importances for each of BERT$_{E2}$ and RF's accurate predictions (of E2 reactions) where primary, secondary, and tertiary electrophilic centres were high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature. C_{elec} and C_{sub} correspond to electrophilic and substituting carbons, respectively.

inaccurate predictions (Figure F.7, Appendix F.6.2), reinforcing the reliability of these assessments.

4.2.2.3 Allylic effects

While allylic effects from alkene, alkyne, and aromatic bonds (to the electrophilic centre) increase reactivity of both E2 and S_N2 reactions, they favour the E2 pathway due to conjugation with the forming alkene bond. Here, accurate predictions with each of these allylic centres were evaluated, with the distribution being 1 × alkene, 3 × alkyne, and 5 × aromatic for S_N2 and 0 × alkene, 0 × alkyne, and 3 × aromatic for E2 (Figure 4.8). The corresponding analysis for inaccurate

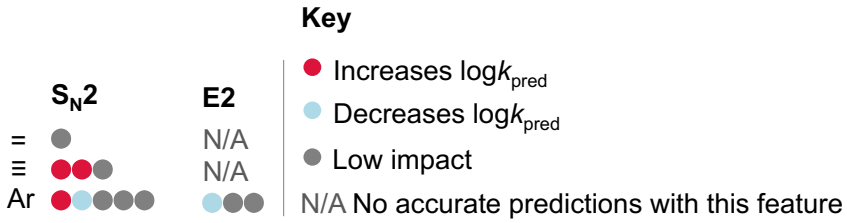


Figure 4.8: Analysis of allylic effects. Signs (positive or negative) of feature importances for each of BERT's accurate predictions where alkene (=), alkyne (≡), and aromatic (Ar) bonds to the electrophilic centre were high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature.

predictions is provided in Appendix F.6.3. The total distribution of centres in the test set was 2 × alkene, 4 × alkyne, and 42 × aromatic for S_N2, and 0 × alkene, 0 × alkyne, and 6 × aromatic for E2. In the BERT model, alkene, alkyne, and aromatic bonds at the electrophilic centre were represented by =, ≡, and c tokens (where c is an aromatic carbon bonded to the centre), respectively.

Analysis of accurate predictions shows that alkyne increased log k_{pred} in 100% of S_N2 reactions where it was high impact (2 reactions), while alkene had low impact in the one S_N2 reaction with this group. While this is too few examples to draw general conclusions from, the association between increased log k_{pred} and alkyne groups at the electrophilic centre may indicate a learned relationship between allylic effects and increased S_N2 reactivity, consistent with the $BERT_{S_N2}$ model discussed in Chapter 3. Importantly, there were no inaccurate S_N2 predictions with alkene or alkyne bonds to the electrophilic centre (Figure F.8, Appendix F.6.3), confirming that allylic effects were not a source of error for this reaction. Conversely, aromatic effects in S_N2 reactions were not identified by the BERT model, with aromatic groups exhibiting mixed impact across accurate predictions (increasing log k_{pred} in one example, decreasing log k_{pred} in one example and

having low impact in three examples), and having low impact across inaccurate predictions (Figure F.8, Appendix F.6.3). This is consistent with the findings for $\text{BERT}_{\text{S}_\text{N}2}$ discussed in Chapter 3.

Regarding E2 reactions, no examples with alkene or alkyne bonds (to the electrophilic centre) were present in the test data. Of the three accurate E2 predictions with an aromatic group at the electrophilic centre, this group decreased $\log k_{\text{pred}}$ in one example and had low impact in the other two. This suggests that BERT was limited in its ability to recognise aromatic effects for E2 reactions, consistent with its performance across $\text{S}_\text{N}2$ reactions. Moreover, aromatic effects were not identified in BERT's inaccurate predictions of E2 reactions, further highlighting its challenge in capturing this effect (Figure F.8, Appendix F.6.3).

In summary, there is evidence to suggest that the BERT model has learned allylic effects from alkyne groups in $\text{S}_\text{N}2$ reactions. However, the limited examples of alkene (in both E2 and $\text{S}_\text{N}2$ reactions) and alkyne (in E2 reactions) groups in the test data meant further analysis was not possible. Additionally, recognition of aromatic effects in either mechanism was identified as a limitation of the BERT model. This is consistent with the findings for $\text{BERT}_{\text{S}_\text{N}2}$ discussed in Chapter 3, further highlighting the model's challenge in capturing this effect. Note this limitation was also observed when considering the entire aromatic ring (as opposed to just the carbon bonded to the electrophilic centre), affirming that the results seen here were not biased by the way the importances were calculated (Figure F.8, Appendix F.6.3).

4.2.2.4 Temperature effects

Reciprocal temperature was introduced as a feature of the BERT model by appending it to the end of the reaction SMILES. Out of the four $\text{S}_\text{N}2$ reactions in the test data observed at multiple temperatures (Reactions **A** to **D**, Figure

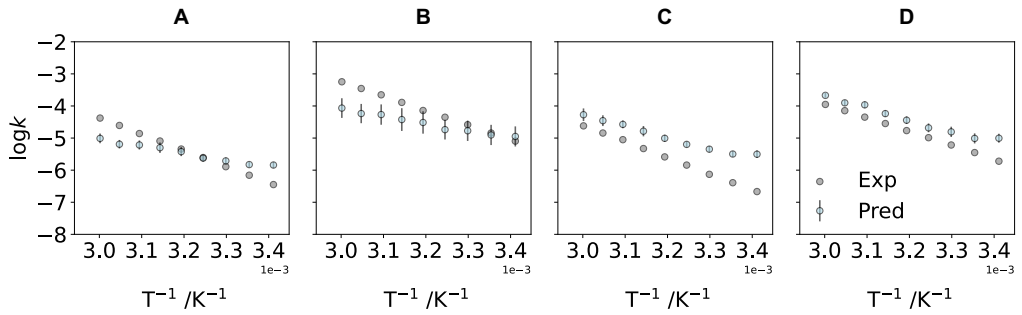
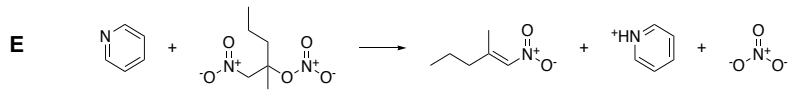


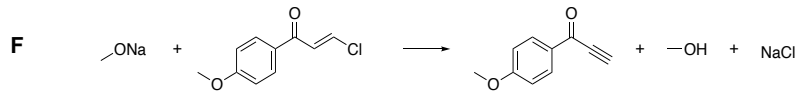
Figure 4.9: Analysis of temperature effects (S_N2). Predicted (Pred) and experimental (Exp) log k values for each of the S_N2 test reactions observed at multiple temperatures (Reactions **A** to **D**, Figure 3.20). Here, T^{-1} denotes the reciprocal temperature.

3.20), BERT correctly predicted the linear decrease in log k with increasing T^{-1} ($-r_P = 0.99$ and $p_P = 0.00$, Figure 4.9). Hence, BERT successfully captured the mathematical relationship between log k and temperature for S_N2 reactions, consistent with the BERT_{S_N2} model discussed in Chapter 3. Note that the reactions in Figure 3.20 were taken from the test data and not necessarily predicted accurately or inaccurately.

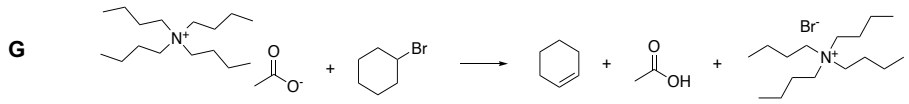
Regarding E2 reactions, there were three examples in the test data observed at multiple temperatures (Reactions **E** to **G**, Figure 4.10; note these reactions were taken from the test data and not necessarily predicted accurately or inaccurately). Each reaction was observed at only three temperatures, so 10 synthetic data points were generated to augment each reactivity series. Out of these three reactions, BERT correctly predicted a linear decrease in log k with increasing T^{-1} ($-r_P \geq 0.93$ and $p_P = 0.00$) within the temperature range 286 to 364 K (3.50×10^{-3} to 2.75×10^{-3} K^{-1}). However, when including temperatures outside of this range, no such correlation was observed. The bounds 286 and 364 K fall $\approx 5\%$ outside the IQR of temperatures in the E2 training data, suggesting limitations in BERT's ability to extrapolate to temperatures outside this range.



1,3-xylene, 370.15 – 393.15 K



methanol, 273.15 – 298.15 K



acetone, 324.15 – 349.55 K

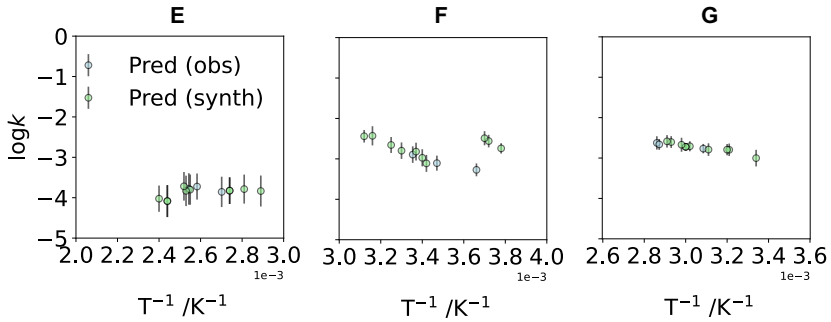


Figure 4.10: Analysis of temperature effects (E2). *Top:* Reactions in the E2 test data observed at multiple temperatures. Reactions are at three different temperatures within the ranges shown. *Bottom*: Predicted (Pred) log k values for Reactions **E** to **G**. Predictions for observed data points (obs) are shown in blue and predictions for synthetic data points (synth) are shown in green. As $\log k_{\text{exp}}$ is not known for the synthetic data points, the ground truth data has been omitted. Note that different x-axis ranges have been used to accommodate different temperature ranges. Here, T^{-1} denotes the reciprocal temperature.

To test whether this phenomenon also occurs for S_N2 reactions, 20 synthetic data points were generated for Reactions **A** to **D**, corresponding to 10 with $T < 286$ K, and 10 with $T > 364$ K (falling outside the IQR of temperatures in the S_N2 training data). Indeed, inclusion of these temperatures reduced the correlation between T^{-1} and $\log k_{\text{pred}}$ significantly ($-r_P \leq 0.69$, Figure F.13 in Appendix F.9.1), suggesting BERT is limited in its ability to extrapolate to temperatures outside the training IQR for both mechanisms.

Notably, this limitation was also observed for an RF model trained on the E2/S_N2 dataset, suggesting a common challenge across both architectures (Figure F.14, Appendix F.9.2). For the RF model, this is likely due to its non-parametric nature where predictions are based on thresholds, making it less effective at handling extreme values. For BERT, the challenge likely arises from how temperature is represented, using discrete tokens rather than continuous values. Nonetheless, both models were able to identify the mathematical relationship between $\log k$ and temperature across the majority temperature range despite this not being explicitly encoded.

Summary. Overall, BERT recognised LG effects in both E2 and S_N2 reactions. BERT identified steric effects in S_N2 reactions, while its modelling of alkyl substitutents effects in E2 reactions were consistent with an experimental correlation between steric bulk and decreased reactivity. There is evidence to suggest that BERT recognised allylic effects from alkyne groups in S_N2 reactions, however, limited examples of alkene and alkyne groups meant further analysis was not possible. Recognition of aromatic effects in both E2 and S_N2 reactions was identified as a limitation. Lastly, BERT identified the linear relationship between logk and reciprocal temperature for temperatures within the training IQR, but was limited in its ability to extrapolate beyond this range. These results suggest that BERT does apply chemical intuition to its modelling of E2/S_N2

reactivity. However, caution should be taken when predicting aromatic effects, or extrapolating to temperatures outside the training IQR. The trends discussed here were also observed across BERT's inaccurate predictions, reinforcing the reliability of these assessments (Appendix F.6).

4.2.3 Literature datasets

The E2/S_N2 BERT was tested on three literature datasets: 1) 10 competing E2/S_N2 reactions from Ref[200], 2) 11 adversarial examples from Ref[201] which proceed via the S_N2 mechanism but not the E2 mechanism, and 3) five examples of Hofmann/Zaystev regioselectivity from Refs[202,203] (4 reactions), and the E2 test data (1 reaction). Here, BERT was tasked with predicting the favoured pathway by estimating $\log k$ for both mechanisms. If the favoured pathway had the lower $\log k_{\text{pred}}$, BERT was considered to have predicted it correctly.

4.2.3.1 Competing reactions

The Competing dataset contains 10 E2/S_N2 reactions testing three key factors that drive mechanistic competition: 1) nucleophile/base effects, comparing methoxide to *tert*-butoxide, 2) LG effects, comparing tosylate to Br, and 3) aromatic effects at the β-C. The dataset spans 10 nucleophile/substrate combinations across two different solvents and two different temperatures. All nucleophile/substrate combinations were unseen during training, with three having an $S_T < 0.4$ to the training data.

Out of the 10 reactions in the Competing dataset, BERT correctly predicted the favoured pathway 60% of the time, including two of the three reactions with $S_T < 0.4$ to the training data. Of the correctly predicted examples, $\Delta \log k_{\text{pred}}$ was significant (≥ 0.1) 66.6% of the time, while $\left(\frac{k_{\text{E2}}}{k_{\text{S}_\text{N}2}}\right)_{\text{pred}}$ was within one order of magnitude of $\left(\frac{k_{\text{E2}}}{k_{\text{S}_\text{N}2}}\right)_{\text{exp}}$ for one such example. Overall, these results are promising,

indicating that BERT can predict the favoured pathway for competing E2/S_N2 reactions. However, there is room for improvement, as BERT still predicts the incorrect pathway 40% of the time. Moreover, while the correct pathway was frequently predicted, $\left(\frac{k_{E2}}{k_{S_N2}}\right)_{pred}$ was accurate only 10% of the time.

To rationalise why BERT predicted the correct pathway for some reactions in the Competing dataset but not others, key features influencing the predicted outcome were analysed for each case. Here, the difference in feature importance between the two pathways (ΔIG) was evaluated for relevant functional groups. A positive $\Delta_{A-B}IG$ indicates the feature was predicted to favour pathway A, while a negative $\Delta_{A-B}IG$ indicates the feature was predicted to favour pathway B. Meanwhile, a ΔIG of zero indicates the feature was insignificant in predicting selectivity. Unless specified, this evaluation was made based on the atom mapping of each functional group in the products, as ΔIG was typically zero in the reactants (Appendix F.10.1). This is intuitive, however, as the products define the mechanistic pathway.

Among the correctly predicted examples were the four reactions in the dataset with a methoxide nucleophile/base, which favour the S_N2 pathway (Figure 4.11). These reactions test BERT's recognition that unhindered nucleophiles (together with a primary electrophilic centre) favour the S_N2 mechanism. For all four of these examples, methoxide was predicted to favour S_N2 ($\Delta_{S_N2\text{-}E2}IG$s: 0.24 ± 0.08, 0.07 ± 0.03, 0.16 ± 0.04, and 0.11 ± 0.05), suggesting BERT correctly associated an unhindered nucleophile with preference for the S_N2 pathway. Other known contributors to S_N2 preference such as the primary electrophilic centre and tosylate LG (present in three examples) were not significant in influencing BERT's predictions ($\Delta IG = 0$ within error). However, the octadecyl group present in the substrate of two reactions (Reactions **H** and **I**) had a notable effect, being predicted to favour S_N2 in both examples ($\Delta_{S_N2\text{-}E2}IG$s: 0.18 ± 0.12 and 0.09 ± 0.06; note this

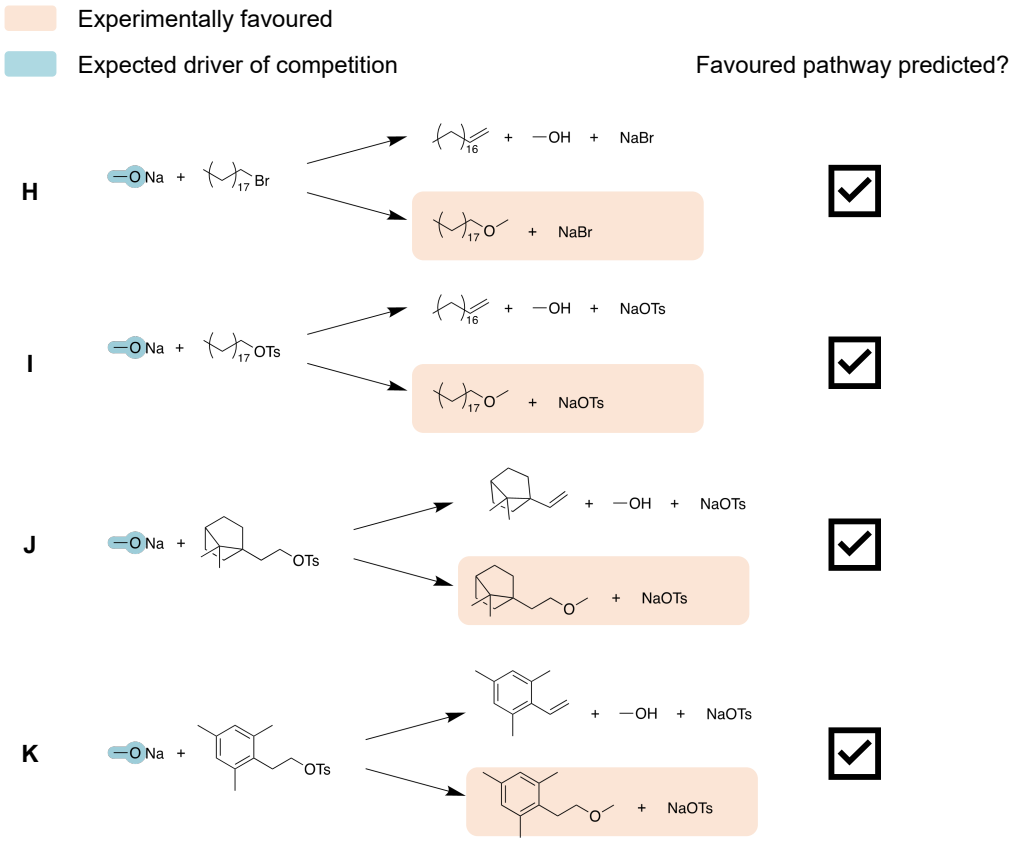


Figure 4.11: Reactions in the Competing dataset with a methoxide nucleophile. These reactions test BERT's recognition that unhindered nucleophiles (together with a primary electrophilic centre) favour the S_N2 mechanism (highlighted in peach). Reactions in methanol at 338.15 K.

pertains to the reactants in one example, and the products in the other). Importantly, BERT identified that an unhindered nucleophile favours the S_N2 mechanism in these four reactions, while also recognising the influence of primary alkyl chains.

Also among the correctly predicted examples were two of the three reactions in the dataset favouring the E2 pathway, which both have a *tert*-butoxide nucleophile/base (Figure 4.12). One of these examples tests BERT's recognition that bulky bases favour the E2 pathway when paired with a halide LG (Reaction **L**), while the other tests BERT's understanding that aromatic groups (bonded to the β-C) promote the E2 pathway in reactions with a bulky base and tosylate LG (which otherwise

favour S_N2) (Reaction **M**). In the aromatic example (Reaction **M**), BERT predicted that the aryl ring favours the E2 mechanism ($\Delta_{\text{E2-S}_N2}IG = 0.11 \pm 0.04$), suggesting it had correctly identified this group's role in promoting the E2 pathway in this reaction. Conversely, BERT had not identified the significance of the *tert*-butoxide base in the halide LG example (Reaction **L**), assigning it an effectively zero ΔIG value. Notably, the most significant contribution in this example came from the 2'-5,5-dimethylbicyclo[2.2.1]heptan-1-yl group ($\Delta_{\text{E2-S}_N2}IG = 0.24 \pm 0.14$), suggesting BERT's prediction in favour of the E2 pathway was largely attributed to steric bulk in the substrate. While bulky substrates are typically known to favour E2 by hindering nucleophilic attack, the analogous reaction with a methoxide nucleophile has been shown to proceed mainly via S_N2 (Reaction **J**, Figure 4.11). Thus, while BERT correctly predicted the favoured pathway for this reaction, its reasoning did not align with known reactivity rules.

A similar link between steric factors in the substrate and predicted E2/S_N2 reactivity was observed in the E2-favouring example that the model predicted incorrectly (Reaction **N**, Figure 4.12). This example, also involving a bulky *tert*-butoxide base and halide LG, was predicted to favour the S_N2 pathway, which BERT attributed to the primary electrophilic centre ($\Delta_{\text{S}_N2\text{-E2}} = 0.09 \pm 0.04$) and octadecyl group in the substrate ($\Delta_{\text{S}_N2\text{-E2}} = 0.19 \pm 0.13$; note this pertains to the reactants). Meanwhile, the *tert*-butoxide base (expected to favour the E2 pathway) was insignificant within error. This suggests that BERT's erroneous prediction in favour of the S_N2 pathway for this reaction stemmed from a chemically intuitive association between primary electrophilic centres and S_N2 reactivity, while ignoring the influence of the bulky base.

Also among the incorrectly predicted competing reactions were the three reactions with a *tert*-butoxide nucleophile/base and tosylate LG that favour the S_N2 pathway

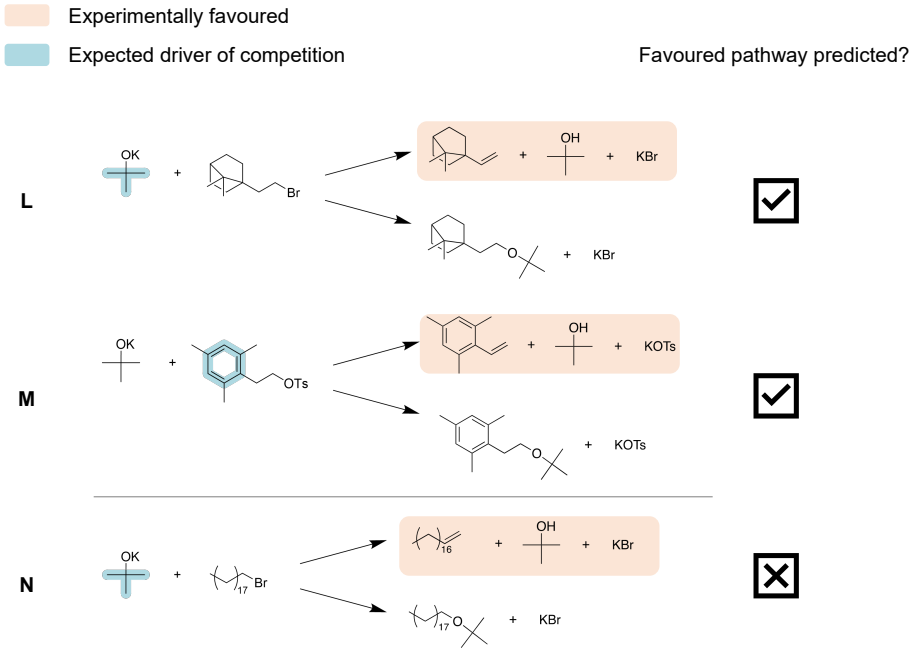


Figure 4.12: Reactions in the Competing dataset that favour the E2 pathway. Reactions **L** and **N** test BERT's recognition that bulky bases favour the E2 pathway (highlighted in peach) when paired with a halide LG, while Reaction **M** tests BERT's understanding that aromatic groups (bonded to the β-C) promote the E2 pathway in reactions with a bulky base and tosylate (OTs) LG (which otherwise favour S_N2). Reactions in *tert*-butanol at 353.15 K.

(Figure 4.13). These reactions test BERT's recognition that the tosylate LG favours the S_N2 mechanism, overriding the influence of the bulky base (which otherwise favours the E2 pathway). For all three of these examples, tosylate had an effectively zero ΔIG value, suggesting BERT had not identified the role of this LG in favouring the S_N2 pathway. However, in two of the three examples (Reactions **P** and **Q**), BERT had correctly identified the bulky base as a significant contributor to the E2 pathway ($\Delta_{\text{E2-S}_N2}IG = 0.10 \pm 0.03$ in both examples). Overall, these results suggest that BERT's erroneous prediction in favour of the E2 pathway for these three reactions stemmed from a chemically intuitive association between bulky bases and E2 reactivity, while ignoring the influence of the tosylate LG. This oversight may be attributed to log k_{exp} values for reactions with tosylate LGs in

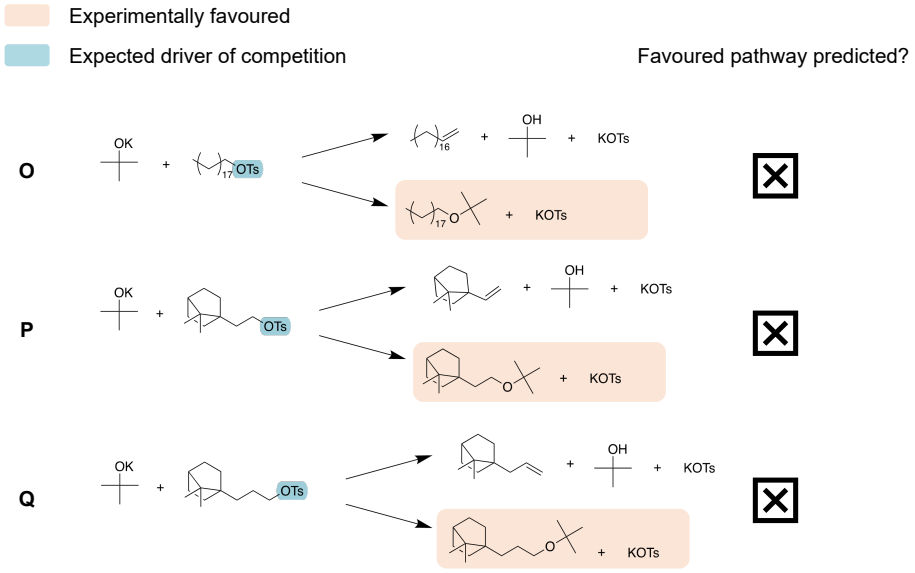


Figure 4.13: Reactions in the Competing dataset with a bulky base and tosylate LG. These reactions test BERT's recognition that the tosylate (OTs) LG favours the S_N2 mechanism (highlighted in peach), overriding the influence of the bulky base (which otherwise favours the E2 pathway). Reactions in *tert*-butanol at 353.15 K.

the training data, which were typically smaller for S_N2 reactions than E2 reactions (IQR: -5.2 to -3.2 for S_N2 and -3.9 to -2.4 for E2, Figure F.15 in Appendix F.10.2). A similar connection was noted in Madzhidov et al.'s early work[29] with the S_N2 dataset (using an SVR algorithm), where $\log k$ was often underpredicted for reactions with substituted tosylate LGs, attributed to the training $\log k$ range.

Overall, BERT recognised that methoxide favours the S_N2 pathway in the four competing examples with this nucleophile. BERT also identified that an aromatic group bonded to the β-C would promote the E2 mechanism in the one Competing example illustrating this effect. Meanwhile, BERT did not recognise that bulky bases favour the E2 pathway in reactions with halide LGs in the two representative competing examples, instead basing its predictions on steric factors in the substrate. Lastly, BERT did not identify the effect of tosylate LGs in promoting the S_N2 mechanism in the three competing examples illustrating this effect. This limitation

was attributed to the unusually small log k values for similar reactions in training. These results suggest that BERT's accurate prediction of competing reactions resulted from its recognition of nucleophilic and aromatic effects (one example), while incorrect predictions resulted from an underestimation of the effects of bulky bases and tosylate leaving groups.

4.2.3.2 Adversarial examples

The Adversarial dataset contains 11 S_N2 reactions of benzyl bromide derivatives, corresponding to four nucleophile/substrate combinations across two different solvents and a temperature range of 232.15 to 298.15 K (Figure 4.14). These reactions, which have no β-Hs, cannot proceed through the E2 mechanism, testing BERT's ability to handle adversarial examples. Note that two nucleophile/substrate combinations in the adversarial dataset were present in the training data as S_N2 reactions. However, this dataset is designed to test BERT's handling of the artificial E2 mechanism, which was not encountered during training.

Out of the 11 reactions in the adversarial dataset, BERT correctly predicted the S_N2 pathway to be favoured 91% of the time, including three of the four reactions where the S_N2 reaction was not in the training data. Moreover, $\Delta \log k_{\text{pred}}$ was significant ($\geq$ 0.1) in every correctly predicted case. These results suggest that BERT effectively handled adversarial examples, correctly identifying that substrates without β-Hs react via the S_N2 pathway. Indeed, the β-C exhibited a positive $\Delta_{S_N2\text{-}E2}IG$ in 70% of correctly predicted examples (and was insignificant in 30%), affirming the predicted selectivity for the S_N2 pathway was attributed to the β-C. Note this evaluation is based on the product mapping of the β-C, as ΔIG was typically zero in the reactants (Appendix F.10.1). This is intuitive, however, as the products define the mechanistic pathway.

Nu	X	Counter-ion	Solvent	Temperature/ K	S_N2 predicted?
			DMF	232.15 – 254.15	☑
			ethanol	273.15, 298.15	☑
			ethanol	298.15	☑
	H		DMF	252.15, 273.15	☑
	H		ethanol	298.15	☑
		Na+	ethanol	273.15	☒

Figure 4.14: Adversarial dataset comprising 11 S_N2 reactions of benzyl bromide derivatives from Ref,[201] spanning four nucleophile/substrate combinations, two different solvents, and a temperature range of 232.15 to 298.15 K. These reactions, which have no β-Hs, cannot proceed through the E2 mechanism, testing BERT's ability to handle adversarial examples. Here, X represents a *para*-substituent on the aryl ring.

It is noted that BERT did not accurately predict $\frac{k_{E2}}{k_{S_N2}}$ for these examples, with the predicted ratio exceeding the experimental ratio (which was set to $\frac{0.1}{99.9}$) by > 1 order of magnitude in all tested cases. However, this is expected, given that no adversarial examples were included in the training set. Incorporating such examples may enhance BERT's ability to accurately quantify $\log k$ in non-reactive scenarios.

4.2.3.3 Hofmann and Zaystev

The Hofmann/Zaystev dataset contains five examples of Hofmann/Zaystev regioselectivity in E2 reactions, corresponding to five nucleophile/substrate combinations across two different solvents (Figure 4.15). Temperatures were not provided for four of these examples, so predictions were made at four different temperatures ranging from 273.15 to 348.15 K in these cases. All nucleophile/substrate combinations in the Hofmann/Zaystev dataset were unseen during training, with two having an $S_T < 0.4$ to the training data.

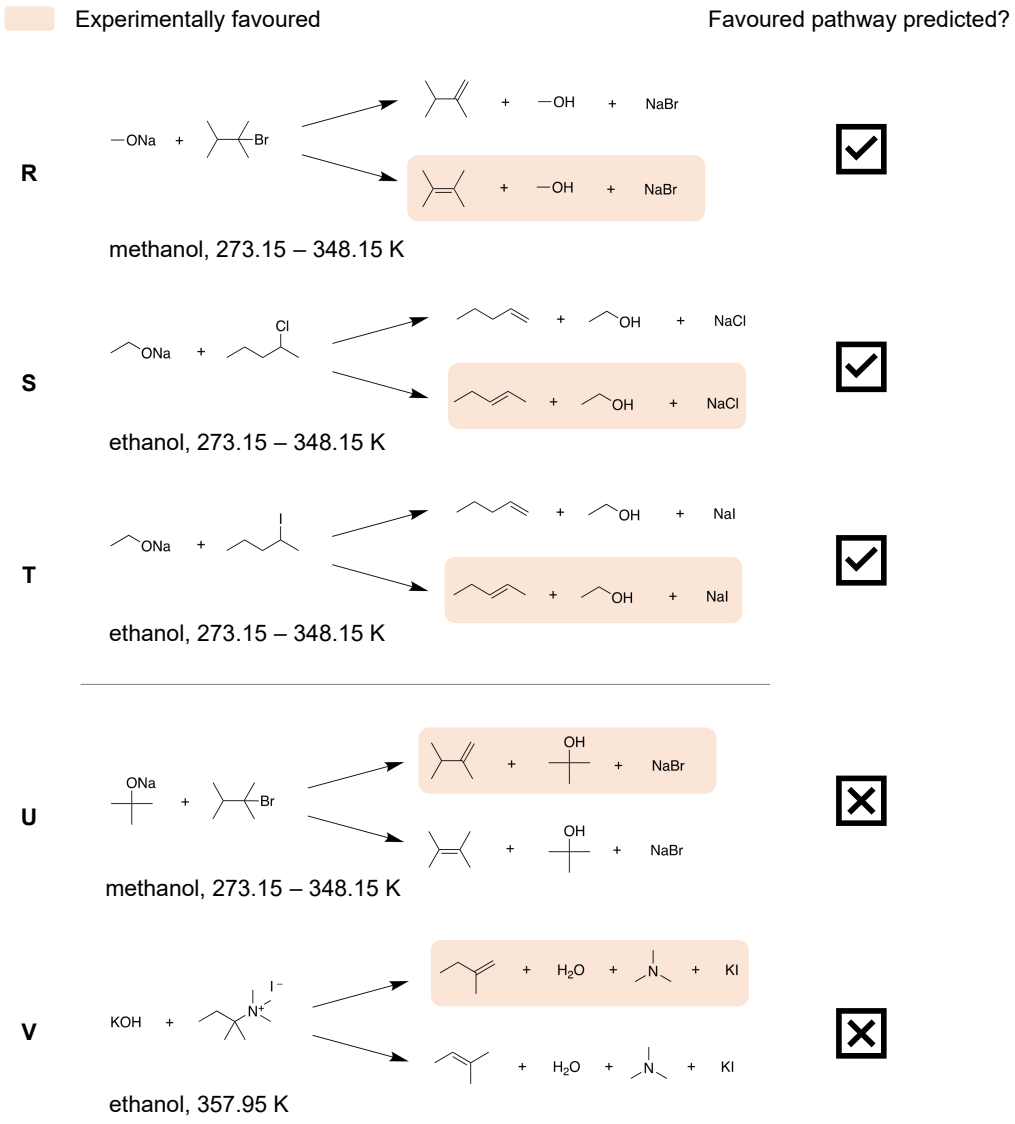


Figure 4.15: Hofmann/Zaystev dataset comprising five E2 reactions from Refs[202,203] (4 reactions), and the E2 test data (1 reaction). The favoured pathway (Hofmann or Zaystev) is highlighted in peach. Where temperatures were not provided, predictions were made at four temperatures ranging from 273.15 to 348.15 K.

Out of the five reactions in the Hofmann/Zaystev dataset, BERT correctly predicted the favoured pathway 60% of the time, including one of the two reactions with $S_T < 0.4$ to the training data. Of the correctly predicted examples, $\Delta \log k_{\text{pred}}$ was significant (≥ 0.1) at all testing temperatures in one example, and at $\geq 50\%$ of testing temperatures in the other examples. Furthermore, $\left(\frac{k_{\text{E2}}}{k_{\text{S}_\text{N}2}}\right)_{\text{pred}}$ was within one order of magnitude of $\left(\frac{k_{\text{E2}}}{k_{\text{S}_\text{N}2}}\right)_{\text{exp}}$ for all tested cases. Overall, these results are promising, suggesting that BERT can predict Hofmann/Zaystev regioselectivity for E2 reactions. This is further supported by accurate predictions of $\frac{k_{\text{E2}}}{k_{\text{S}_\text{N}2}}$ for these examples. However, there is room for improvement, as BERT still predicts the incorrect product 40% of the time.

Notably, all correctly predicted examples correspond to those where the Zaystev product is favoured, raising the possibility of a training bias towards this pathway. Nonetheless, analysis of training data revealed that 35% (23 reactions) of Hofmann/Zaystev training examples favoured the Hofmann product. While this is lower than the proportion favouring the Zaystev product, the imbalance is not large enough to indicate a bias.

To rationalise BERT's preference for the Zaystev product, key features driving regioselectivity were analysed. Here, ΔIG of relevant functional groups were evaluated for the two Hofmann-favouring examples in the dataset. A positive $\Delta_{\text{Hof-Zay}}IG$ indicates the feature was predicted to favour the Hofmann product, while a negative $\Delta_{\text{Hof-Zay}}IG$ indicates the feature was predicted to favour the Zaystev product. Meanwhile, a ΔIG of zero indicates the feature was insignificant in predicting regioselectivity. Unless specified, this evaluation was made based on the atom mapping of each group in the products, as ΔIG was typically zero in the reactants (Appendix F.10.1). This is intuitive, however, as the products define the mechanistic pathway.

Of the two incorrectly predicted Hofmann-favouring reactions, one tests BERT's recognition that steric crowding from a bulky *tert*-butoxide base and branched 2,3-dimethylbutyl substrate favours the Hofmann product (Reaction **U**), while the other tests BERT's recognition that a cationic trimethylamine LG favours this product (Reaction **V**). In the steric example (evaluated at four temperatures), BERT correctly identified that methyl branches on the substrate would favour the Hofmann product ($\Delta_{\text{Hof-Zay}}IG = 0.16 \pm 0.07$, averaged over the four temperatures), but did not recognise the role of the bulky base ($\Delta IG = 0$ within error). Similarly, BERT did not capture the significance of the cationic LG group in the example illustrating this effect. Overall, these results suggest that BERT's preference towards the Zaystev product in these two examples resulted from its underestimation of the effects driving Hofmann selectivity in these reactions.

4.3 Conclusion

The BERT model trained for S_N2 rate prediction in Chapter 3 (BERT_{S_N2}) was fine-tuned in this study to predict experimental $\log k$ values for both E2 and S_N2 reactions. The E2/S_N2 BERT exhibited comparable accuracy to BERT models trained on each individual mechanism, offering a more generalisable model without compromising performance. Additionally, the E2/S_N2 BERT exceeded the accuracy of the best-performing E2/S_N2 rate prediction model in the literature, which was attributed to the quality of the training data. Furthermore, E2/S_N2 BERT was shown to predict the favoured pathway of competing reactions with 60% accuracy, and achieved 91% accuracy predicting the outcome of adversarial examples.

Regarding model strengths and limitations, the E2/S_N2 BERT identified known LG, steric and temperature effects (for temperatures within the training IQR).

However, recognition of aromatic effects, and extrapolation to temperatures outside the training IQR were identified as limitations.

Overall, BERT successfully predicted rates for both E2 and S_N2 reactions, maintaining performance on each individual mechanism, and improving on literature accuracy. Moreover, BERT correctly identified key structural and physical effects associated with E2/S_N2 reactivity. Future work should aim to overcome the limitations highlighted in this Chapter by enhancing BERT's description of targeted effects. For example, inclusion of the substrate LUMO energy may improve learning of aromatic effects, while a continuous description of temperature may improve extrapolation outside the training range.

5 | Conclusion

5.1 Thesis overview

The accurate prediction of reaction rates is an integral step in reaction mechanism elucidation and design of synthetic pathways. Traditionally, kinetic parameters have been derived from activation energies obtained from Quantum Mechanics (QM) methods, and more recently, Machine Learning (ML). Transformer-based models have emerged as a popular ML method for chemical prediction, allowing reactions to be represented as SMILES strings. Despite being state-of-the-art in both reaction classification and yield prediction, Bidirectional Encoder Representations from Transformers (BERT), a type of transformer-based model, have previously not been applied to kinetic prediction. In this Thesis, a BERT model was trained to predict experimental $\log k$ values of E2 and S_N2 reactions, and its predictions were interpreted in the context of known reactivity rules.

Initially, BERT was trained to predict experimental $\log k$ values of S_N2 reactions and its performance was compared to the state-of-the-art Random Forest (RF) model in terms of accuracy, training time, and ability to replicate known reactivity rules (Chapter 3). Both RF and BERT achieved near-experimental accuracy (RMSE: $\approx 1.1 \log k$) on similarity-split test data, however RF did so significantly faster. Additionally, both models identified key reaction centre atoms and physical features, as well as known LG, steric, allylic, temperature, and solvent effects (BERT only, solvent effects could not be interpreted in the RF), reinforcing the

validity of their predictions. Moreover, each model was found to overcome the other's limitations, with RF excelling at recognising aromatic effects, while BERT was more proficient in $\log k$ extrapolation.

BERT was subsequently fine-tuned to predict experimental $\log k$ values of both E2 and S_N2 reactions, and its predictions were interpreted regarding known drivers of mechanistic competition (Chapter 4). The E2/S_N2 BERT achieved an RMSE of 1.2 ± 0.1 $\log k$, maintaining the performance of BERT models trained on each individual mechanism, and exceeding the accuracy of the best-performing literature model (Green et al.[34]). Furthermore, the E2/S_N2 BERT was shown to predict the favoured pathway of competing reactions with 60% accuracy, and achieved 91% accuracy predicting the outcome of adversarial examples. Regarding model strengths and limitations, the E2/S_N2 BERT identified known LG, steric and temperature effects (for temperatures within the training IQR). However, recognition of aromatic effects, and extrapolation to temperatures outside the training IQR were identified as limitations.

Overall, BERT successfully predicted rates for both E2 and S_N2 reactions, maintaining its performance on each individual mechanism, and improving on literature accuracy. Moreover, BERT correctly identified key structural and physical effects associated with E2/S_N2 reactivity. Future work should address the limitations highlighted in Chapters 3 and 4 by enhancing BERT's description of targeted effects.

5.2 Implications and future avenues

In this Thesis, a BERT model was trained to predict experimental $\log k$ values of E2 and S_N2 reactions, and its predictions were interpreted in the context of known reactivity rules. This work established a framework for the accurate and interpretable prediction of reaction kinetics, with potential applications in guiding

synthetic design. To enhance BERT's efficacy for synthesis planning, future work aims to address the limitations highlighted in this Thesis. Additionally, BERT's application to more complex chemistries is explored, aiming to increase both generalisability (encompassing a wider range of reaction classes) and specificity (providing a more detailed description of individual systems).

5.2.1 Addressing limitations of the BERT model

While BERT successfully identified key LG, steric, and temperature effects (for temperatures within the training IQR), recognition of aromatic effects and extrapolation to temperatures outside the training IQR were identified as limitations. Future work should aim to address these limitations by enhancing BERT's description of each effect. For example, inclusion of the substrate LUMO energy may improve learning of aromatic effects, while a continuous-valued description of temperature may improve extrapolation outside the training range.

Analysis of competing reactions also identified bulky bases and tosylate LGs as potential error sources in the BERT model. For tosylate LGs, this was attributed to the unusually small $\log k$ values for S_N2 reactions with this LG in training, suggesting that including examples with a more representative $\log k$ range may help to address this limitation. This was not the case for bulky bases, which were solely represented by E2 reactions during training. However, including a steric descriptor (such as Sterimol parameters,[77,78] oriented around the axis of the nucleophilic atom) may enhance BERT's learning of this effect.

It is noted that the RF model excelled at recognising aromatic effects, overcoming this limitation of the BERT model. However, BERT was more proficient in $\log k$ extrapolation, with both models limited by temperature extrapolation. Thus, improving BERT to incorporate the strengths of both models (while addressing common challenges) provides a more comprehensive approach.

Similarly, a model trained solely on physical descriptors (e.g., LUMO energies, Sterimol parameters) may seem an attractive option. However, calculating these descriptors involves additional methodological considerations, such as the choice of theoretical method and which atoms to represent in the description. Moreover, this approach requires assumptions about which effects to model *a priori*. By leveraging BERT, the entire system is efficiently described using SMILES, enabling the identification of targeted effects that are not captured by language alone. This results in a more informed, results-driven approach that greatly reduces the complexity of the problem.

5.2.2 Increasing generalisability of the BERT model

To maximise BERT's utility, future work aims to expand the model to a wider range of reaction classes. This *many-class* rate prediction has previously been achieved using the diverse set of gas phase reactions involving C, H, N, and O compiled by Green et al.[14] Despite spanning an extensive 16,500 transformations, this dataset's reliance on QM-generated $\Delta E^{\ddagger}$ values (prone to error accumulation), and its neglect of solvent effects, underscore the need for a diverse dataset based on experimental data. Indeed, the best-performing model trained on Green's data (Green et al., 2022.2[42]) scored an RMSE of 3.6 ± 0.1 $\log k$ (significantly higher than chemical accuracy), affirming higher-quality data is sought. Currently, experimental kinetic data is limited to a few reaction mechanisms (E2,[27] S_N2,[29] cycloadditions,[30,31,204] S_NAr[32]). However, by combining data mining techniques with large reaction databases (such as Pistachio[64,a]), experimental kinetics can be extracted across a wider range of reaction classes.

Recently, the Duarte group gained access to IBM Deep Search,[205] a data-mining tool that utilises NLP to search a vast corpus (~ millions) of preloaded documents

[a] Accessible to the Duarte group.

covering a variety of data sources including patents and open-access publications. By employing Deep Search to mine kinetic data for the 4 million unique reactions in the Pistachio database (spanning 12 reaction superclasses, including heterocycle formations and functional group interconversions[60]), future work aims to curate a diverse dataset of organic reactions and their experimental log k values. This dataset (combined with existing experimental kinetic data[27,29–32,204]) will be used to train BERT to predict rates across a variety of reaction classes, expanding its generalisability beyond the E2/S_N2 reactions explored in this work.

In the absence of sufficient experimental data, BERT may be trained using a multi-fidelity approach.[206,207] In this approach, QM-generated data (low fidelity) is leveraged alongside the experimental data (high fidelity) to increase the size of the training set. Predictions are made for both fidelities so the model can predict experimental log k values without being confounded by the QM error.

5.2.3 Increasing specificity of the BERT model

Whereas generalisability involves expanding BERT to additional reaction classes, specificity involves fine-tuning BERT on a narrowed reaction subset. This enables BERT's application to more complex chemistries requiring a detailed description beyond what the general model provides. Relevant examples include glycosylation and C-F bond formation.

Glycosylation reactions are nucleophilic substitutions at the anomeric centre of carbohydrates. Glycosylation is key to accessing glycans, which are integral to modelling biological processes[208–210] and in the preparation of synthetic vaccines.[211–213] One of the greatest challenges in glycosyl chemistry is directing glycosylation towards the S_N2 limit of the S_N1/S_N2 spectrum, in order to gain stereochemical control.[214,215] Hence, fine-tuning BERT to predict S_N1/S_N2 selectivity in these reactions would be instrumental in their experimental design.

To model stereochemistry in glycosylation reactions, the atom parity notation of SMILES can be leveraged.[216] Here, @@ and @ tokens are respectively used to represent clockwise and anticlockwise arrangement around a stereocentre, with substituent priority going from the left-most bond (highest priority) to the right-most bond (lowest priority) in the SMILES string (Figure 5.1**a**). This approach has been successfully applied by Laino et al.[217] to forward synthesis prediction of carbohydrate reactions in the Molecular Transformer (a predecessor to the BERT model). By combining Laino's dataset of 20,000 carbohydrate reactions (including glycosylation sequences) with a data mining tool such as IBM Deep Search[205] (*vide supra*), future work aims to curate a dataset of glycosylation reactions and their experimental log k values. This dataset will be used to fine-tune BERT to predict rates of glycosylation pathways, allowing S_N1/S_N2 selectivity to be determined. Note, stereochemistry was not modelled in the E2/S_N2 BERT due to the absence of stereochemical information in the data.

Another notable application of S_N2 reactivity is the formation of C-F bonds via S_N2 fluorination. Fluorine substitution is ubiquitous in drug discovery due to its role in modulating lipophilicity, metabolic stability, and bioavailability of drug molecules.[218–220] Naturally, C-F bond formation relies on minimising the E2 pathway to maximise yield of the S_N2 product. Hence, fine-tuning BERT to predict S_N2/E2 selectivity in these reactions would be instrumental in their experimental design. Key advancements in C-F bond formation have exploited hydrogen bond phase transfer catalysis (HB-PTC) to bring solid-phase fluoride salts into solution.[221,222] Here, hydrogen bond donors (e.g., alcohols) coordinate to the fluoride salt with varying stoichiometry. These complexes can be conveniently modelled in BERT using the SMILES sequence "MF.(HBD)X", where MF is the fluoride salt, HBD is the hydrogen bond donor, and X is the coordination stoichiometry (Figure 5.1**b**). By leveraging available experimental data,[221,223]

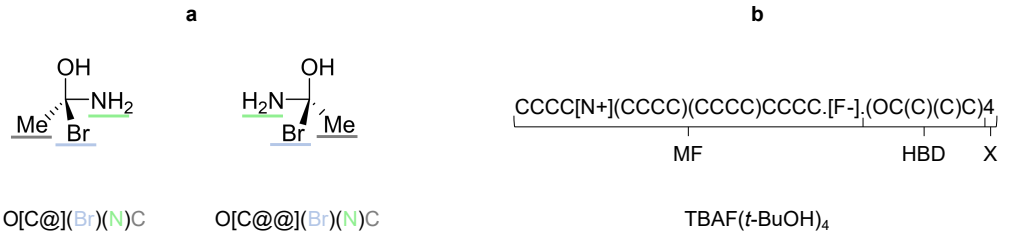


Figure 5.1: SMILES representations of chiral centres (a) and fluoride complexes (b). **a.** In SMILES, stereocentres are defined using @@ and @ tokens to represent clockwise and anticlockwise arrangement of substituents, with substituent priority going from the left-most bond (highest priority, OH in this example) to the right-most bond (lowest priority, Me in this example) in the SMILES string. The highest priority substituent is placed in front of the stereocentre, and clockwise or anticlockwise arrangement is defined going from second-highest priority to lowest priority. **b.** Fluoride complexes can be conveniently modelled in BERT using the SMILES sequence "MF.(HBD)X", where MF is the fluoride salt (tetra-*n*-butylammonium fluoride (TBAF) in this example), HBD is the hydrogen bond donor (*t*-BuOH in this example), and X is the coordination stoichiometry (4 in this example).

future work aims to fine-tune BERT to predict the kinetics of these systems, allowing S_N2/E2 selectivity to be determined.

Overall, this section highlights two promising applications of the BERT model to synthetic design, underscoring the potential for future experimental collaborations.

5.3 Outlook

In this Thesis, a BERT model was trained to predict experimental $\log k$ values for E2 and S_N2 reactions, and its predictions were interpreted in the context of known reactivity rules. The E2/S_N2 BERT exhibited strong performance, achieving near-experimental accuracy and identifying key structural and physical effects. Furthermore, this model demonstrated potential in predicting E2/S_N2 selectivity, achieving 60% accuracy in this task. To address model limitations, inclusion of physical descriptors (e.g., LUMO energies) was suggested to enhance BERT's recognition of targeted effects (e.g., aromatic effects). Additionally, BERT's application to synthetic design was explored, with future efforts aiming

to enhance both generalisability (encompassing a wider range of reaction classes), and specificity (providing a more detailed description of individual systems). To expand BERT to a wider range of reaction classes, more high-quality kinetic data was sought. To address this, both data mining and multi-fidelity methods were proposed as potential solutions. Meanwhile, BERT's applicability to glycosylation and C-F bond formation was discussed, underscoring the potential for future experimental collaborations. Overall, this Thesis presented a framework for the accurate and interpretable prediction of reaction kinetics, offering potential applications in guiding synthetic design.

Bibliography

(1) Jorner, K.; Tomberg, A.; Bauer, C.; Sköld, C.; Norrby, P.-O. *Nat. Rev. Chem.* **2021**, *5*, 240–255.

(2) Komp, E.; Janulaitis, N.; Valleau, S. *Phys. Chem. Chem. Phys.* **2021**, *24*, 2692–2705.

(3) Lewis-Atwell, T.; Townsend, P.; Grayson, M. *WIREs: Comput. Mol. Sci.* **2021**, *12(4)*, e1593.

(4) Doney, A.; Rooks, B.; Lu, T.; Wheeler, S. *ACS Catal.* **2016**, *6*, 7948–7955.

(5) Pereira, O.; Ruth, M.; Gerbig, D.; Wende, R.; Schreiner, P. *J. Am. Chem. Soc.* **2024**, *146*, 14576–14586.

(6) Okada, H.; Maeda, S. *ACS Omega* **2024**, *9*, 7123–7131.

(7) Eyring, H. *Chem. Rev.* **1935**, *17*, 65–77.

(8) Atkns, P.; de Paula, J., *Physical chemistry*, 10th ed.; Oxford University Press: 2014; Chapter 21C.

(9) Ptáček, P.; Šoukal, F.; Opravil, T., *Introduction to the transition state theory*; Oxford University Press: 2014; Chapter 21C.

(10) Ramadhar, T.; Batey, R. *Comput. Theor. Chem.* **2011**, *976*, 167–182.

(11) Li, Y.; Fang, D. *Phys. Chem. Chem. Phys.* **2014**, *16*, 15224–15230.

(12) Li, S.; Fang, D. *Phys. Chem. Chem. Phys.* **2016**, *18*, 30815–30823.

(13) von Rudorff, G.; Heinen, S.; Bragato, M.; von Lilienfeld, A. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045026.

(14) Grambow, C.; Pattanaik, L.; Green, W. *Sci. Data.* **2020**, *7*, 137–144.

(15) Spiekermann, K.; Pattanaik, L.; Green, W. *Sci. Data.* **2022**, *9*, 417–428.

(16) Stuyver, T.; Jorner, K.; Coley, C. *Sci. Data.* **2023**, *10*, 66–79.

(17) Plata, R.; Singleton, D. *J. Am. Chem. Soc.* **2015**, *137*, 3811–3826.

(18) Liu, Z.; Patel, C.; Harvey, J.; Sunoj, R. *Phys. Chem. Chem. Phys.* **2017**, *19*, 30647–30657.

(19) Chan, L.; Morris, G.; Hutchison, G. *J. Chem. Theory Comput.* **2021**, *17*, 2099–2106.

(20) Dehabadi, M.; Saidi, H.; Zafari, F.; Irani, M. *Research Square preprint https://doi.org/10.21203/rs.3.rs-4341960/v1* **2024**.

(21) Van Duin, A.; Dasgupta, S.; Lorant, F.; Goddard, W. *J. Phys. Chem. A* **2001**, *105*, 9396–9409.

(22) Warshel, A.; Weiss, R. *J. Am. Chem. Soc.* **1980**, *102*, 6218–6226.

(23) Young, T.; Johnston-Wood, T.; Deringer, V.; Duarte, F. *Chem. Sci.* **2021**, *12*, 10944–10955.

(24) Zhang, H.; Juraskova, V.; Duarte, F. *Nat. Commun.* **2024**, *15*, 6114–6124.

(25) Kravtsov, A.; Karpov, P.; Baskin, I.; Palyulin, V.; Zefirov, N. *Dokl. Chem.* **2011**, *440*, 299–301.

(26) Madzhidov, T.; Polishchuk, P.; Nugmanov, R.; Bodrov, A.; Lin, A.; Baskin, I.; Varnek, A.; Antipin, I. *Russ. J. Org. Chem.* **2014**, *50*, 459–463.

(27) Madzhidov, T.; Bodrov, A.; Gimadiev, T.; Nugmanov, R.; Antipin, I.; Varnek, A. *J. Struct. Chem.* **2016**, *56*, 1227–1234.

(28) Madzhidov, T.; Gimadiev, T.; Malakhova, D.; Nugmanov, R.; Baskin, I.; Antipin, I.; Varnek, A. *J. Struct. Chem.* **2017**, *58*, 650–656.

(29) Gimadiev, T.; Madzhidov, T.; Tetko, I.; Nugmanov, R.; Casciuc, I.; Klimchuk, O.; Bodrov, A.; Polishchuk, P.; Antipin, I.; Varnek, A. *Mol. Inform.* **2019**, *38*, e1800104.

(30) Glavatskikh, M.; Madzhidov, T.; Horvath, D.; Nugmanov, R.; Gimadiev, T.; Malakhova, D.; Marcou, G.; Varnek, A. *Mol. Inform.* **2019**, *38*, e1800077.

(31) Ravasco, J.; Coelho, J. *J. Am. Chem. Soc.* **2020**, *142*, 4235–4241.

(32) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. *Chem. Sci.* **2021**, *12*, 1163–1175.

(33) Rakhimbekova, A.; Akhmetshin, T.; Minibaeva, G.; Nugmanov, R.; Gimadiev, T.; Madzhidov, T.; Baskin, I.; Varnek, A. *SAR QSAR Environ. Res.* **2021**, *32*, 207–219.

(34) Heid, E.; Green, W. *J. Chem. Inf. Model.* **2022**, *62*, 2101–2110.

(35) Zankov, D.; Madzhidov, T.; Baskin, I.; Varnek, A. *Mol. Inform.* **2023**, *42*, e2200275.

(36) Grambow, C.; Pattanaik, L.; Green, W. *J. Phys. Chem. Lett.* **2020**, *11*, 2992–2997.

(37) Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. *Chem. Sci.* **2020**, *11*, 4584–4601.

(38) Vargas, S.; Hennefarth, M.; Liu, Z.; Alexandrova, A. *J. Chem. Theory Comput.* **2021**, *17(10)*, 6203–6213.

(39) Heinen, S.; von Rudorff, G.; von Lilienfeld, O. *J. Chem. Phys.* **2021**, *155*, 064105.

(40) Stuyver, T.; Coley, C. *J. Chem. Phys.* **2022**, *156*, 084104.

(41) Ismail, I.; Robertson, C.; Habershon, S. *J. Chem. Phys.* **2022**, *157*, 014109.

(42) Spiekermann, K.; Pattanaik, L.; Green, W. *J. Phys. Chem. A* **2022**, *126*, 3976–3986.

(43) Farrar, E.; Grayson, M. *Chem. Sci.* **2022**, *13*, 7594–7603.

(44) Espley, S.; Farrar, E.; Buttar, D.; Tomasi, S.; Grayson, M. *Digit. Discov.* **2023**, *2*, 941–951.

(45) Lewis-Atwell, T.; Beechey, D.; Şimşek, Ö.; Grayson, M. *ACS Catal.* **2023**, *13*, 13506–13515.

(46) Stuyver, T.; Coley, C. *Chemistry* **2023**, *29*, e202300387.

(47) Garcia-Andrade, X.; Garcia Tahoces, P.; Perez-Rios, J.; Martinez Nunez, E. *J. Phys. Chem. A* **2023**, *127*, 2274–2283.

(48) van Gerwen, P.; Briling, K.; Bunne, C.; Somnath, V.; Laplaza, R.; Krause, A.; Corminboeuf, C. *arXiv preprint arXiv:2312.08307* **2023**.

(49) Vijay, S.; Venetos, M.; Spotte-Smith, E.; Kaplan, A.; Wen, M.; Persson, K. *Chem. Sci.* **2024**, *15*, 2923–2936.

(50) Evans, M.; Polanyi, M. *Trans. Faraday Soc.* **1935**, *31*, 875–894.

(51) Evans, M.; Polanyi, M. *Trans. Faraday Soc.* **1936**, *32*, 1333–1360.

(52) Evans, M.; Polanyi, M. *Trans. Faraday Soc.* **1938**, *34*, 11–24.

(53) Bell, R. *Proc. R. Soc. Lond. A.: Math. Phys. Sci.* **1936**, *154*, 414–429.

(54) Brönsted, J.; Pedersen, K. *Z. Phys. Chem.* **1924**, *108*, 185–235.

(55) Brönsted, J. *Chem. Rev.* **1928**, *5*, 231–338.

(56) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. *Sci. Adv.* **2021**, *7*, eabe4166.

(57) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. *Mol. Inform.* **2010**, *29*, 855–868.

(58) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. *arXiv preprint arXiv:1810.04805v2* **2019**.

(59) Taylor, W. *J.Q.* **1953**, *30*, 415–433.

(60) Schwaller, P.; Probst, D.; Vaucher, A.; Nair, V.; Kreutter, D.; Laino, T.; Reymond, J.-L. *Nat. Mach. Intell.* **2021**, *3*, 144–152.

(61) Schwaller, P.; Vaucher, A.; Laino, T.; Reymond, J.-L. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 015016.

(62) Bran, A.; Schwaller, P. *arXiv preprint arXiv:2310.06083v1* **2023**.

(63) Janakarajan, N.; Erdmann, T.; Swaminathan, S.; Laino, T.; Born, J. *arXiv preprint arXiv:2309.16235v1* **2023**.

(64) Pistachio (Nextmove Software). URL: `http://www.nextmovesoftware.com/pistachio.html` (last accessed 19/11/2024).

(65) Kovacs, D.; McCorkindale, W.; Lee, A. *Nat. Commun.* **2021**, *12*, 1695–1703.

(66) Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; Müller, K.-R. *Nat. Commun.* **2019**, *10*, 1096.

(67) Kuz'min, V.; Polishchuk, P.; Artemenko, A.; Andronati, S. *Mol. Inform.* **2011**, *30*, 593–603.

(68) Perronnin, F.; Sánchez, J.; Mensink, T. In *ECCV Conference Proceedings*, 2010, pp 143–156.

(69) Sánchez, J.; Perronnin, F.; Mensink, T.; Verbeek, J. *Int. J. Comput. Vis.* **2013**, *105*, 222–245.

(70) Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; Samek, W. *PloS one* **2015**, *10*, e0130140.

(71) Kuz'min, V.; Artemenko, A.; Polischuk, P.; Muratov, E.; Hromov, A.; Liahovskiy, A.; Andronati, A.; Makan, S. *J. Mol. Model.* **2005**, *11*, 457–467.

(72) Saabas, A. treeinterpreter - python package for interpreting scikit-learn's decision tree and random forest predictions. URL: `https://pypi.org/project/treeinterpreter/` (last accessed 11/6/2024).

(73) Sundararajan, M.; Taly, A.; Yan, Q. In *ICML Conference Proceedings*, 2017; Vol. 70, pp 3319–3328.

(74) Lowe, D. Extraction of chemical structures and reactions from the literature, Ph.D. Thesis, The University of Cambridge, 2012.

(75) Tanimoto, T. *Rapport technique, IBM Internal* **1957**.

(76) Lundberg, S. *arXiv preprint arXiv:1705.07874* **2017**.

(77) Verloop, A.; Hoogenstraaten, W.; Tipker, J. *Drug Des.* **1976**, *7*, 165–207.

(78) Harper, K.; Bess, E.; Sigman, M. *Nat. Chem.* **2012**, *4*, 366–374.

(79) van der Maaten, L.; Hinton, G. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

(80) Pearson, K. *London Edinburgh Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572.

(81) Sykes, P., *A guidebook to mechanism in organic chemistry*, 6th ed.; Pearson Education India: 1986; Chapter 4.

(82) Jones, R., *Physical and mechanistic organic chemistry*, 2nd ed.; Cambridge University Press: 1979; Chapter 7.

(83) Hofmann, A. *Philos. Trans. R. Soc. Lond.* **1851**, *141*, 357–398.

(84) Hofmann, A. *Justus Liebig's Ann. Chem.* **1851**, *78*, 253–286.

(85) Saytzeff, A. *Justus Liebig's Ann. Chem.* **1875**, *179*, 296–301.

(86) Lewis, D. *Bull. Hist. Chem* **1995**, *17*, 21–30.

(87) Sykes, P., *A guidebook to mechanism in organic chemistry*, 6th ed.; Pearson Education India: 1986; Chapter 9.

(88) Jones, R., *Physical and mechanistic organic chemistry*, 2nd ed.; Cambridge University Press: 1979; Chapter 8.

(89) Schrödinger, E. *Ann. phys.* **1926**, *384*, 273–376.

(90) Schrödinger, E. *Ann. Phys.* **1926**, *384*, 489–584.

(91) Schrödinger, E. *Ann. Phys.* **1926**, *385*, 437–490.

(92) Schrödinger, E. *Ann. Phys.* **1926**, *386*, 109–204.

(93) Born, M.; Oppenheimer, R. In *Quantum Chemistry: Classic Scientific Papers*; World Scientific: 2000, pp 1–24.

(94) Hartree, D. *Math. Proc. Camb.* **1928**, *24*, 89–110.

(95) Fock, V. *Z. Phys.* **1930**, *61*, 126–148.

(96) Pauli, W. *Z. Phys.* **1925**, *31*, 765–783.

(97) Slater, J. *Phys. Rev.* **1929**, *34*, 1293–1322.

(98) Slater, J. *Phys. Rev.* **1930**, *36*, 57–64.

(99) Boys, S. *Proc. R. Soc. A: Math. Phys. Eng. Sci.* **1950**, *200*, 542–554.

(100) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(101) Orca Input Library - Basis Sets. URL: `https://sites.google.com/site/orcainputlibrary/basis-sets` (last accessed 12/7/2025).

(102) Dunning Jr, T. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(103) Kendall, R.; Dunning Jr, T.; Harrison, R. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

(104) ccRepo: A correlation consistent basis sets repository. URL: `http://www.grant-hill.group.shef.ac.uk/ccrepo/index.html` (last accessed 12/7/2025).

(105) Löwdin, P.-O. *Phys. Rev.* **1955**, *97*, 1509–1520.

(106) Coester, F.; Kümmel, H. *Nuc. Phys.* **1960**, *17*, 477–485.

(107) Čížek, J. *J. Chem. Phys.* **1966**, *45*, 4256–4266.

(108) Raghavachari, K.; Trucks, G.; Pople, J.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479–483.

(109) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, 864–871.

(110) Kohn, W.; Sham, L. *Phys. Rev.* **1965**, *140*, 1133–1138.

(111) Perdew, J.; Schmidt, K. In *AIP Conference Proceedings*, 2001; Vol. 577, pp 1–20.

(112) Zhao, Y.; Lynch, B.; Truhlar, D. *J. Phys. Chem. A* **2004**, *108*, 4786–4791.

(113) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.

(114) Becke, A. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(115) Perdew, J.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(116) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(117) Langreth, D.; Mehl, M. *Phys. Rev. B* **1983**, *28*, 1809–1834.

(118) Becke, A. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(119) Perdew, J.; Chevary, J.; Vosko, S.; Jackson, K.; Pederson, M.; Singh, D.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*, 6671–6687.

(120) Perdew, J.; Kurth, S.; Zupan, A.; Blaha, P. *Phys. Rev. Lett.* **1999**, *82*, 2544–2547.

(121) McQuarrie, D., *Statistical mechanics*; University Science: 2000.

(122) Maczek, A.; Meijer, A., *Statistical thermodynamics*; Oxford University Press: 2017.

(123) Herzberg, G., *Molecular spectra & molecular structure*; van Nostrand: 1950; Vol. 1: Spectra of diatomic molecules, pp 66–73.

(124) Maczek, A.; Meijer, A., *Statistical thermodynamics*; Oxford University Press: 2017; Chapter 9.

(125) Maczek, A.; Meijer, A., *Statistical thermodynamics*; Oxford University Press: 2017; Chapter 7.

(126) Maczek, A.; Meijer, A., *Statistical thermodynamics*; Oxford University Press: 2017, p 53.

(127) Maczek, A.; Meijer, A., *Statistical thermodynamics*; Oxford University Press: 2017; Chapter 8.

(128) Maczek, A.; Meijer, A., *Statistical thermodynamics*; Oxford University Press: 2017; Chapter 10.

(129) Maczek, A.; Meijer, A., *Statistical thermodynamics*; Oxford University Press: 2017; Chapter 12.

(130) Herzberg, G., *Molecular spectra & molecular structure*; van Nostrand: 1950; Vol. 1: Spectra of diatomic molecules, pp 73–81.

(131) Barone, V. *J. Chem. Phys.* **2005**, *122*, 14108–14108.

(132) Barone, V.; Biczysko, M.; Bloino, J.; Borkowska-Panek, M.; Carnimeo, I.; Panek, P. *Int. J. Quantum Chem.* **2012**, *112*, 2185–2200.

(133) Grimme, S. *Chem. Eur. J.* **2012**, *18*, 9955–9964.

(134) Maczek, A.; Meijer, A., *Statistical thermodynamics*; Oxford University Press: 2017, p 70.

(135) Breiman, L. *Mach. Learn.* **2001**, *45*, 5–32.

(136) Quinlan, J. *Mach. Learn.* **1986**, *1*, 81–106.

(137) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, Ł.; Polosukhin, I. In *NeurIPS Conference Proceedings*, 2017; Vol. 30.

(138) Hendrycks, D.; Gimpel, K. *arXiv preprint arXiv:1606.08415* **2020**.

(139) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.

(140) Ba, J.; Kiros, J.; Hinton, G. *arXiv preprint arXiv:1607.06450v1* **2016**.

(141) He, K.; Zhang, X.; Ren, S.; Sun, J. In *CVPR Conference Proceedings*, 2016, pp 770–778.

(142) Rumelhart, D.; Hintont, G.; Williams, R. *Nature* **1986**, *323*, 533–536.

(143) Kingma, D.; Ba, J. *arXiv preprint arXiv:1412.6980* **2014**.

(144) Loshchilov, I.; Hutter, F. *arXiv preprint arXiv:1711.05101v3* **2019**.

(145) Robbins, H.; Monro, S. *Ann. Math. Stat.* **1951**, 400–407.

(146) Trainer - transformers 4.10.1 documentation (Huggingface.co). URL: `https://huggingface.co/transformers/v4.10.1/main_classes/trainer.html#transformers.TrainingArguments` (last accessed 11/6/2024).

(147) Balandat, M.; Karrer, B.; Jiang, D.; Daulton, S.; Letham, B.; Wilson, A.; Bakshy, E. In *NeurIPS Conference Proceedings*, 2020; Vol. 33, pp 21524–21538.

(148) Frazier, P. *arXiv preprint arXiv:1807.02811v1* **2018**.

(149) Bayes, T. *Biometrika* **1958**, *45*, 296–315.

(150) Rasmussen, C.; Williams, C., *Gaussian processes for machine learning*; 3; MIT press Cambridge, MA: 2006; Vol. 2.

(151) Wang, J. *Comput. Sci. Eng.* **2023**, *25*, 4–11.

(152) Models - Ax platform 0.1.9 documentation. URL: `https://ax.dev/versions/0.1.9/api/models.html#torch-models` (last accessed 11/6/2024).

(153) Mercer, J. *Trans. Roy. Soc. London* **1909**, 415–446.

(154) Boser, B.; Guyon, I.; Vapnik, V. In *COLT Conference Proceedings*, 1992, pp 144–152.

(155) Bishop, C.; Nasrabadi, N., *Pattern recognition and machine learning*; 4; Springer: 2006; Vol. 4; Chapter 2.3.1.

(156) Rasmussen, C.; Williams, C., *Gaussian processes for machine learning*; 3; MIT press Cambridge, MA: 2006; Vol. 2, pp 84–85.

(157) Matérn, B., *Spatial variation*; Springer Science & Business Media: 2013; Vol. 36.

(158) Rasmussen, C.; Williams, C., *Gaussian processes for machine learning*; 3; MIT press Cambridge, MA: 2006; Vol. 2, p 113.

(159) Henderson, W.; Buckler, S. *J. Am. Chem. Soc.* **1960**, *82*, 5794–5800.

(160) Okamoto, K.; Matsuda, H.; Kawaski, H.; Shingu, H. *Bull. Chem. Soc. Jap.* **1967**, *40*, 1917–1920.

(161) DeTar, D.; McMullen, D.; Luthra, N. *J Am Chem Soc* **1978**, *100:8*, 2484–2493.

(162) Bordwell, F.; Mrozack, S. *J. Org. Chem.* **1982**, *47*, 3802–3803.

(163) Bordwell, F.; Hughes, D. *J. Am. Chem. Soc.* **1984**, *106*, 3234–3240.

(164) Rao, T.; Punnaiah, G.; Sundaram, E. *Proc. Indian Acad. Sci.* **1986**, *97*, 55–61.

(165) Amyes, T.; Jencks, W. *J. Am. Chem. Soc.* **1989**, *111*, 7900–7909.

(166) Banait, N.; Jencks, W. *J. Am. Chem. Soc.* **1991**, *113*, 7951–7958.

(167) Kondo, Y.; Urade, M.; Yamanishi, Y.; Chen, X. *J. Chem. Soc., Perkin Trans.2* **2002**, 1449–1454.

(168) Ruff, F. *Internet Electron. J. Mol. Des.* **2004**, *3*, 474–498.

(169) Vlasov, V. *Russ. Chem. Rev.* **2006**, *75*, 765–796.

(170) Ruff, F.; Farkas, O. *J. Org. Chem.* **2006**, *71*, 3409–3416.

(171) Tabula: Extract Tables from PDFs. URL: `https://tabula.technology/` (last accessed 12/7/2025).

(172) ChemDraw: Chemical Drawing Software. URL: `https://revvitysignals.com/products/research/chemdraw` (last accessed 12/7/2025).

(173) Probst, D.; Reymond, J.-L. *J. Cheminf.* **2020**, *12*, 12–24.

(174) Probst, D.; Reymond, J.-L. *Bioinformatics* **2018**, *34*, 1433–1435.

(175) Probst, D.; Reymond, J.-L. *J. Cheminf.* **2018**, *10*, 1–12.

(176) Catalán, J.; Díaz, C. *Liebigs Ann.* **2006**, *1997*, 1941–1949.

(177) Catalán, J.; Díaz, C.; López, V.; Pérez, P.; De Paz, J.-L.; Rodríguez, J.-G. *Liebigs Ann.* **2006**, *1996*, 1785–1794.

(178) Catalán, J.; López, V.; Pérez, P.; Martin-Villamil, R.; Rodríguez, J.-G. *Liebigs Ann.* **2006**, *1995*, 241–252.

(179) Taft, R.; Kamlet, M. *J. Am. Chem. Soc.* **1976**, *98:10*, 2886–2894.

(180) Kamlet, M.; Taft, R. *J. Am. Chem. Soc.* **1976**, *98:2*, 377–383.

(181) Kamlet, M.; Abboud, J.; Taft, R. *J. Am. Chem. Soc.* **1977**, *19:18*, 6027–6038.

(182) Bakshy, E.; Dworkin, L.; Karrer, B.; Kashin, K.; Letham, B.; Murthy, A.; Singh, S. In *NeurIPS Conference Proceedings*, 2018, pp 1–8.

(183) Neese, F. *WIREs: Comp. Mol. Sci.* **2012**, *2*, 73–78.

(184) Young, T.; Silcock, J.; Sterling, A.; Duarte, F. *Angew. Chem.* **2021**, *133*, 4312–4320.

(185) Young, T. duartegroup/otherm: Major Symmetry Improvements, 2020.

(186) Brémond, É.; Savarese, M.; Su, N.; Pérez-Jiménez, Á.; Xu, X.; Sancho-García, J.; Adamo, C. *J. Chem. Theory Comput.* **2016**, *12*, 459–465.

(187) Bursch, M.; Mewes, J.-M.; Hansen, A.; Grimme, S. *Angew. Chem. Int. Ed.* **2022**, *61*, e202205735.

(188) Orca Input Library - Geometry Optimizations. URL: `https://sites.google.com/site/orcainputlibrary/geometry-optimizations` (last accessed 11/6/2024).

(189) Tantirungrotechai, Y.; Phanasant, K.; Roddecha, S.; Surawatanawong, P.; Sutthikhum, V.; Limtrakul, J. *J. Mol. Struct.: THEOCHEM* **2006**, *760*, 189–192.

(190) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comp. Chem.* **2011**, *32*, 1456–1465.

(191) Orca Input Library - Coupled Cluster. URL: `https://sites.google.com/site/orcainputlibrary/coupled-cluster` (last accessed 11/6/2024).

(192) Wilcoxon, F. In *Breakthroughs in statistics: Methodology and distribution*; Springer: 1992, pp 196–202.

(193) Nayak, B.; Hazra, A. *Indian J. Ophthalmol.* **2011**, *59*, 85–86.

(194) Pratt, J. *J. Am. Stat. Assoc.* **1959**, *54*, 655–667.

(195) Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S., et al. *arXiv preprint arXiv:2009.07896* **2020**.

(196) Lowry, T.; Richardson, K., *Mechanism and theory in organic chemistry*; Harper & Row: 1976, p 85.

(197) Clayden, J.; Greeves, N.; Warren, S., *Organic chemistry*, 2nd ed.; Oxford University Press: 2012, p 342.

(198) RDKit: Open-source cheminformatics. URL: `https://www.rdkit.org` (last accessed 11/6/2024).

(199) Mann, H.; Whitney, D. *Ann. Math. Stat.* **1947**, 50–60.

(200) Veeravagu, P.; Arnold, R.; Eigenmann, E. *J. Am. Chem. Soc.* **1964**, *86*, 3072–3075.

(201) Ko, E.; Parker, A. *J. Am. Chem. Soc.* **1968**, *90*, 6447–6453.

(202) Liang-Lin, J.; Locascio, T.; Witherspoon, A.; Gao, S.; Laster, J.; Tripp, M.; Shorb, J.; Cox Jr, C. *J. Chem. Educ.* **2024**, *101*, 2308–2320.

(203) Bunnett, J. *Angew. Chem. Int. Ed. Engl.* **1962**, *1*, 225–235.

(204) Madzhidov, T.; Gimadiev, T.; Malakhova, D.; Nugmanov, R.; Baskin, I.; Antipin, I.; Varnek, A. *J. Struct. Chem.* **2017**, *58*, 650–656.

(205) IBM Deep Search (IBM Research). URL: `https://ds4sd.github.io/` (last accessed 19/11/2024).

(206) Fernández-Godino, M. *arXiv preprint arXiv:1609.07196* **2016**.

(207) Fare, C.; Fenner, P.; Benatan, M.; Varsi, A.; Pyzer-Knapp, E. *NPJ Comput. Mater.* **2022**, *8*, 257.

(208) Ernst, B.; Hart, G.; Sinaÿ, P., *Carbohydrates in chemistry and biology*; Wiley Blackwell: 2008.

(209) Stallforth, P.; Lepenies, B.; Adibekian, A.; Seeberger, P. *J. Med. Chem.* **2009**, *52*, 5561–5577.

(210) Boilevin, J.; Reymond, J.-L. *Synth.* **2018**, *50*, 2631–2654.

(211) Broecker, F.; Seeberger, P. In *Methods in Enzymology*; Elsevier: 2017; Vol. 597, pp 311–334.

(212) Barel, L.-A.; Mulard, L. *Hum. Vaccines Immunother.* **2019**, *15*, 1338–1356.

(213) Mettu, R.; Chen, C.-Y.; Wu, C.-Y. *J. Biomed. Sci.* **2020**, *27*, 1–22.

(214) Adero, P.; Amarasekara, H.; Wen, P.; Bohé, L.; Crich, D. *Chem. Rev.* **2018**, *118*, 8242–8284.

(215) Fu, Y.; Bernasconi, L.; Liu, P. *J. Am. Chem. Soc.* **2021**, *143*, 1577–1589.

(216) Daylight Chemical Information Systems Inc. SMILES - A Simplified Chemical Language. URL: `https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html` (last accessed 19/11/2024).

(217) Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J.-L. *Nat. Commun.* **2020**, *11*, 4874.

(218) Purser, S.; Moore, P.; Swallow, S.; Gouverneur, V. *Chem. Soc. Rev.* **2008**, *37*, 320–330.

(219) Gillis, E.; Eastman, K.; Hill, M.; Donnelly, D.; Meanwell, N. *J. Med. Chem.* **2015**, *58*, 8315–8359.

(220) Meanwell, N. *J. Med. Chem.* **2018**, *61*, 5822–5880.

(221) Engle, K.; Pfeifer, L.; Pidgeon, G.; Giuffredi, G.; Thompson, A.; Paton, R.; Brown, J.; Gouverneur, V. *Chem. Sci.* **2015**, *6*, 5293–5302.

(222) Pupo, G.; Gouverneur, V. *J. Am. Chem. Soc.* **2022**, *144*, 5200–5213.

(223) Yonezawa, T.; Sakamoto, Y.; Nogawa, K. *Jpn. Kokai Tokkyo Koho* **1994**, JP–06316551.

Appendices

A | Equations

A.1 Softmax

The softmax function normalises a vector of n real numbers, $\boldsymbol{x} \in \mathbb{R}^n$, into a probability distribution of n possible outcomes, $\boldsymbol{x} \in (0, 1)^n$ where $\sum_{i=1}^{n} x_i = 1$, by applying the function,

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{A.1}$$

to each element.

A.2 Tanh activation function

The Tanh activation function calculates the hyperbolic tangent of an input x using,

$$\text{Tanh}(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \quad . \tag{A.2}$$

A.3 Layer normalisation

Layer normalisation[1] is applied per data point to a feature x_i using,

$$x_{i,\text{norm}} = \left(\frac{x_i - \bar{\boldsymbol{x}}}{\sigma_x - \varepsilon}\right) w + b \quad . \tag{A.3}$$

Here, $\bar{\boldsymbol{x}}$ and σ_x respectively denote the mean and standard deviation of $\boldsymbol{x}$, w and b are a learnable weight and bias, and ε, which was set to its default value[2] of 10^{-12} in this study, is included for numerical stability.

B | Derivations

B.1 Deriving δ_j for backpropagation

The δ_j term in Eqs. 2.44 and 2.45, representing the partial derivative of error function E_d with respect to the activation output from node j in hidden layer l of a neural network, a_j, is derived in the following manner:[3]

1. Consider E_d to be a function of $\boldsymbol{z} = (z_q)_{q=1}^m$, the weighted inputs into the m nodes in layer $l+1$ receiving input from node j.

2. Express $\frac{\partial E_d(\boldsymbol{z})}{\partial a_j}$ as a total derivative,

$$\frac{\partial E_d(\boldsymbol{z})}{\partial a_j} = \sum_{q=1}^{m} \frac{\partial E_d}{\partial z_q} \frac{\partial z_q}{\partial a_j} \quad . \tag{B.1}$$

3. Apply the chain rule to Eq. B.1 to give,

$$\frac{\partial E_d(\boldsymbol{z})}{\partial a_j} = \sum_{q=1}^{m} \frac{\partial E_d}{\partial a_q} \frac{\partial a_q}{\partial z_q} \frac{\partial z_q}{\partial a_j} \tag{B.2}$$

   where a_q and z_q are the activation output and weighted input of node q, respectively.

4. Realise $z_q = \sum_{r=1}^{k} w_{rq} a_r$ (where r is one of the k nodes in layer l connected to node q) with only the $w_{jq} a_j$ term depending on a_j, so that the $\frac{\partial z_q}{\partial a_j}$ term in Eq. B.2 can be evaluated as w_{jq}.

5. Substitute $\frac{\partial z_q}{\partial a_j} = w_{jq}$ and $\frac{\partial a_q}{\partial z_q} = \text{Act}'(z_i)$ (as derived for $\frac{\partial a_j}{\partial z_j}$ in Eq. 2.44) into Eq. B.2 to yield Eq. 2.45 (repeated here for clarity),

$$\delta_j = \sum_{q=1}^{m} \frac{\partial E_d}{\partial a_q} \text{Act}'(z_i) w_{jq}, \qquad \text{if } j \text{ in hidden layer.} \tag{2.45}$$

C | Justifications

C.1 Justification 1: Evaluating halide nucleophiles

Figure C.1 shows the groups of nucleophiles that were prevalent (i.e., more than one example of each nucleophile) in the test data. Group **i** (halides) represents Nu atom variation, while the other groups represent inductive (groups **ii** and **v**) or resonant (group **iii**) effects. While each of these effects contribute to S_N2 reactivity, it is important to establish that RF and BERT have learned the fundamental effects from varying the Nu atom before evaluating more complex effects such as induction or resonance at the nucleophilic centre.

C.2 Justification 2: Solvent effects in the RF

In the RF, solvent was modelled by 13 properties representing either polarity or proticity. The Spearman's r value (r_S) between each of these properties and $\log k_{exp}$ is plotted in Figure C.2 for each reaction in the training data that was observed in different solvents with a single solvent component and ionic strength 0. The r_S value between each solvent property and $\log k_{exp}$ varies greatly across the training data, with each solvent property being negatively correlated ($r_S < 0$) with $\log k_{exp}$ in some reactions, and positively correlated ($r_S > 0$) with $\log k_{exp}$ in others. Therefore, there is no consistent relationship between each solvent property and $\log k_{exp}$ in the experimental data to compare learned solvent effects in the RF to. Accordingly, solvent effects in the RF were not evaluated in the current work.

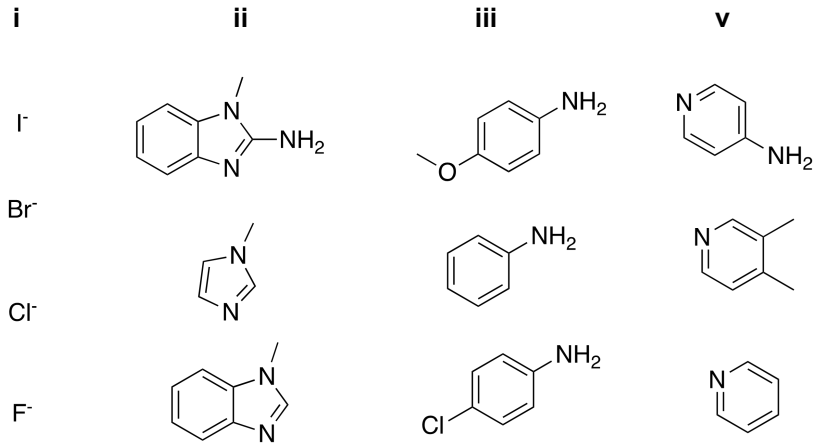


Figure C.1: Groups of nucleophiles that were prevalent in the test data. Here, a group is said to be prevalent if there is more than one example of each nucleophile in the test data.

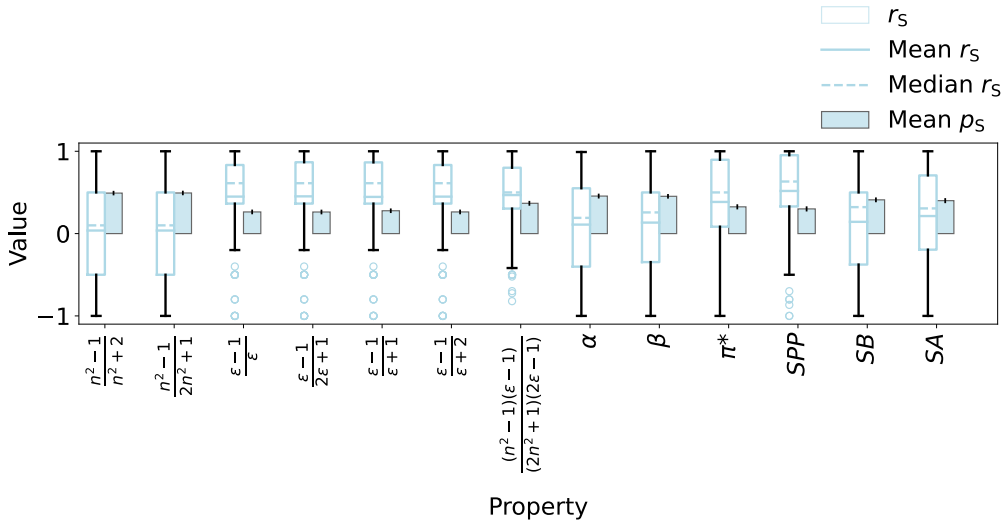


Figure C.2: Spearman's r value (r_S) between each solvent property in the RF and $\log k_{\mathbf{exp}}$ for each reaction in the training data that was observed in different solvents with a single solvent component and ionic strength 0. The mean Spearman's p value (mean p_S) is also provided for each solvent property. Here, ε = dielectric constant, n = refractive index, SA, SB, and SPP = Catalán acidity,[4] basicity[5] and polarity/polarisability[6] constants, respectively, and α, β, and π^* = Kamlet-Taft acidity,[7] basicity,[8] and polarity/polarisability[9] constants, respectively.

D | Chapter 1 supplementary material

D.1 Train/validation/test splits

The train/validation/test splits used in each of the works described in Sections 1.1.2 and 1.1.3 are provided in Table D.1.

Ref.	Train/validation/test
[10]Green et al. (2020)	85/5/10
[11]Habershon et al. (2022)	80/20
[12]Green et al. (2022.1)	80/10/10
[13]Green et al. (2022.2)	85/5/10
[14]Corminboeuf et al. (2023)	90/5/5
[15]von Lilienfeld et al. (2021) E2	800/x
[15]von Lilienfeld et al. (2021) S_N2	1600/x
[16]Coley et al. (2022) E2	800/200
[16]Coley et al. (2022) S_N2	1440/360
[16]Coley et al. (2022) E2 & S_N2	60/20/20
[17]Persson et al. (2024)	80/10/10
[18]Balcells et al. (2020)	80/10/10
[19]Coley et al. (2023)	N/A
[20]Grayson et al. (2022)	64/16/20
[21]Grayson et al. (2023.1) α	174/44/55
[21]Grayson et al. (2023.1) β	66/x/x
[21]Grayson et al. (2023.1) γ	326/x/x
[22]Grayson et al. (2023.2)	N/A
[23]Núñez et al. (2023)	85/5/10
[24]Baskin et al. (2011)	80/10/10
[25]Madzhidov and Baskin et al. (2014)	Not reported
[26]Madzhidov et al. (2016)	80/20
[27]Madzhidov et al. (2017)	80/20
[28]Madzhidov et al. (2019)	80/20
[29]Madzhidov et al. (2021)	80/20
[30]Madzhidov et al. (2023)	72/18/10
[31]Buttar et al. (2021)	72/8/20

Table D.1: Train/validation/test splits used in each of the works described in Sections 1.1.2 and 1.1.3. x denotes this data size was not reported.

E | Chapter 3 supplementary material

E.1 QM methodology

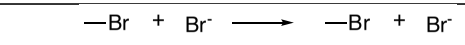
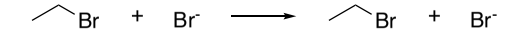
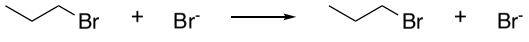
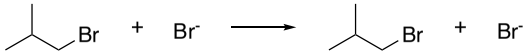
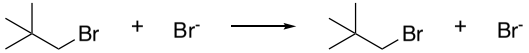
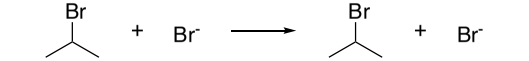
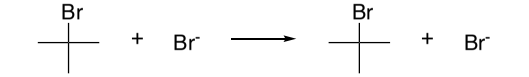
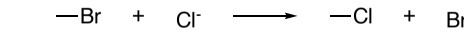
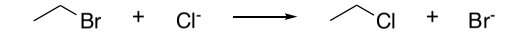
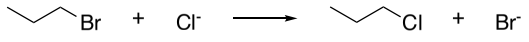
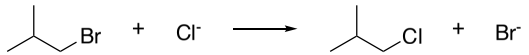
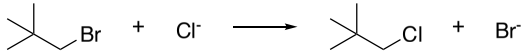
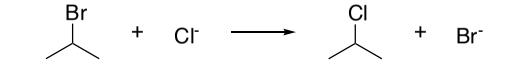
QM calculations were carried out with María Calvo, a summer student in the Duarte group (June to September 2021). Calculations using autodE were performed by M. Calvo, while calculations without autodE and frequency shift calculations were performed by myself.

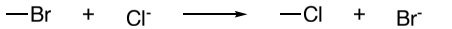
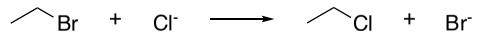
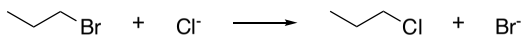
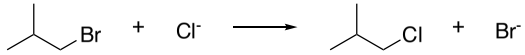
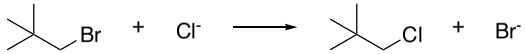
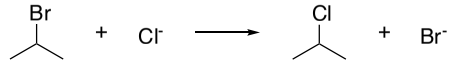
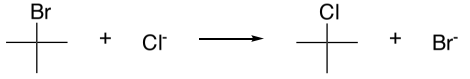
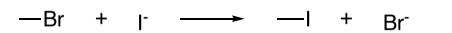
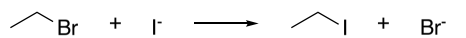
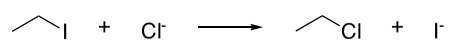
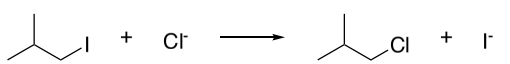
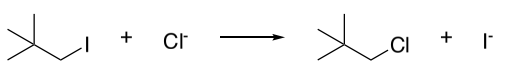
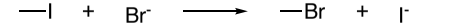
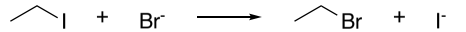
QM calculations were carried out on 30 of the S_N2 reactions manually curated in the current work with a single solvent component and total reactant molecular weight (MW) $\leq$ 235.9 Da, shown in Table E.1.

Methodological considerations. Identity Reactions **1**-**7** were not compatible with autodE, which uses the difference between reactant and product graphs to identify breaking and forming bonds. Hence, ORCA inputs were generated manually for these reactions using optimised xyz coordinates of structural analogues **8** to **14** as initial geometries. TS geometry optimisations for Reactions **18** and **26** did not converge when using initial TS guesses from autodE, each resulting in a structure with two imaginary modes consistent with an elimination pathway (visualised in Chemcraft[32]). The S_N2 TS of each of these reactions was found using optimised xyz coordinates of structural analogue **11** as an initial TS guess and optimising accordingly. Cartesian constraints of the Nu and LG atoms were applied to TS optimisation + frequency calculations of Reactions **14**, **21**, and **24** to prevent 180° angle errors. For these reactions, the optimised xyz coordinates of the path point before the error occurred were used as an initial TS guess, and C-Nu and C-LG bonds were added as internal coordinates. Optimised TS structures contained exactly one imaginary frequency, with magnitudes in the expected range of $\approx$ 250cm^{-1} to 450cm^{-1}. Frequency shift calculations were carried out on the optimised reactant complexes of Reactions **18**, **19**, and

22 using Otherm[33] to convert imaginary frequencies to real frequencies. All other optimised reactant complexes had no imaginary frequencies, or imaginary frequencies of magnitude $< 30\text{cm}^{-1}$, which were disregarded as noise.

Table E.1: Dataset for QM calculations.[a]

ID	Reaction	Solvent	Total MW / Da	$\Delta G^{\ddagger}_{\mathrm{exp}}$ [b] / kcal mol^{-1}	$\Delta G^{\ddagger}_{\mathrm{QM}}$ / kcal mol^{-1}	$\log k_{\mathrm{exp}}$	$\log k_{\mathrm{QM}}$ [b]
1		acetone	174.84	18.6	21.8	-0.9	-3.2
2		acetone	188.87	21.2	22.5	-2.8	-3.7
3		acetone	202.90	21.5	22.8	-3.0	-3.9
4		acetone	216.92	23.2	23.5	-4.2	-4.4
5		acetone	230.95	27.7	29.5	-7.5	-8.8
6		acetone	202.90	23.9	24.6	-4.8	-5.2
7		acetone	216.92	24.7	25.1	-5.3	-5.6
8		acetone	130.39	20.5	18.6	-2.2	-0.9
9		acetone	144.42	22.9	20.9	-4.0	-2.5
10		acetone	158.45	23.2	19.6	-4.2	-1.6
11		acetone	172.47	24.0	22.8	-4.8	-3.9
12		acetone	186.50	29.1	28.7	-8.5	-8.2
13		acetone	158.45	25.2	20.7	-5.7	-2.4

14		acetone	172.47	26.3	20.2	-6.5	-2.0
15		DMF	130.39	17.9	18.8	-0.4	-1.0
16		DMF	144.42	20.1	20.7	-2.0	-2.4
17		DMF	158.45	20.4	20.9	-2.2	-2.5
18		DMF	172.47	22.1	21.4	-3.4	-2.9
19		DMF	186.50	27.2	27.6	-7.1	-7.5
20		DMF	158.45	22.5	23.4	-3.7	-4.4
21		DMF	172.47	24.2	20.6	-4.9	-2.3
22		acetone	221.84	18.2	20.6	-0.6	-2.3
23		acetone	235.87	21.2	24.1	-2.8	-4.9
24		acetone	191.42	22.1	25.2	-3.4	-5.7
25		acetone	205.44	22.4	19.3	-3.6	-1.4
26		acetone	219.47	24.0	20.5	-4.8	-2.2
27		acetone	233.50	28.7	25.8	-8.2	-6.1
28		acetone	205.44	24.1	19.2	-4.9	-1.3
29		acetone	221.84	17.9	18.9	-0.3	-1.1
30		acetone	235.87	19.9	21.7	-1.8	-3.1

MW = Molecular weight, exp = experimental. [a]Data from Ref.[34] [b]Calculated using the Eyring equation.[35]

E.2 Nucleophilic effects

High impact and low impact nucleophiles in BERT's inaccurate predictions are provided in Tables E.2 and E.3 respectively.

E.3 Solvent effects

The percentage of accurate and inaccurate predictions where each solvent property of the RF was high impact are provided in Figure E.1.

E.4 Product atom mappings

Each reaction centre atom (Nu, LG, electrophilic C) has a mapped atom in the products (Nu_p, LG_p, C_p). The percentage of accurate predictions where Nu_p, LG_p, and C_p were high impact features is shown in Figure E.2 for the RF and BERT models.

In both models, Nu_p and C_p proved as important as Nu and C for predicting reactivity, with Nu_p and Nu being high impact in $\geq$ 75% of accurate predictions for both models and C_p and C being high impact in 100% of accurate predictions for the RF, and < 60% of accurate predictions in BERT. The latter was attributed to the nature of the comparison, where the BERT model analyses electrophilic C tokens while RF considers electrophilic centre fragments, which also include the environment around the electrophilic C. On the contrary, LG_p proved less important for reactivity prediction than LG in both models (high impact in < 25% of accurate predictions, compared to > 90% for LG). For the RF, this can be attributed to the fact that LG_p was monoatomic in 78% of accurate predictions, and therefore was not a feature of the RF (for which the minimum fragment length was 2 atoms). However, this does not explain the low importance of LG_p in the BERT model. Importantly, the LG atom on the reactant side was high impact in > 90% of BERT's accurate predictions (Section 3.2.4), emphasising BERT recognises the LG as a key feature in predicting reactivity.

E.4.1 LG effects

In the BERT model, product mappings to halide LGs were represented by [I^-], [Br^-], [Cl^-], and [F^-] tokens, with the distribution of LGs being 17 × I, 8 × Br, 2 × Cl, and 0 × F. Conversely, product mappings to halide LGs (which are

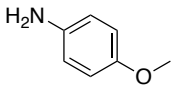
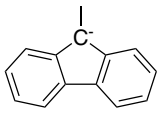
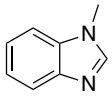
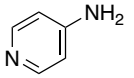
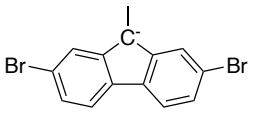
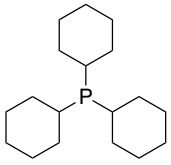
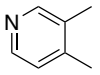
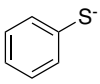
Nucleophile	Times high impact
	1
	1
	1
	1
	1
	2
	6
	1
	1
	2
	1

Table E.2: High impact nucleophiles in BERT's inaccurate predictions.

monoatomic), were not a feature of the RF, for which the minimum fragment length was 2 atoms. Accordingly, the importance of [I^-], [Br^-], [Cl^-], and [F^-] tokens in the BERT model, either increasing $\log k_{\text{pred}}$ (positive sign) or decreasing $\log k_{\text{pred}}$ (negative sign) is presented in Figure E.3 for accurate predictions.

Product mappings to halide LGs in the BERT model were less important for predicting reactivity than their reactant-side counterparts, with I, Br, and Cl being low impact 82%, 63%, and 100% of the time in the products compared to 29%, 38%, and 0% in the reactants. However, this is to be expected as product

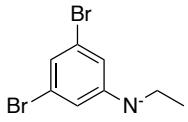
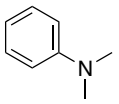
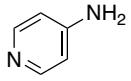
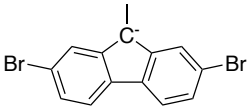
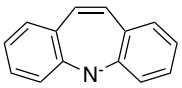
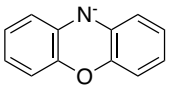
Nucleophile	Times low impact
	2
	1
	1
	1
	1
	1
	1
	2
	1
	1
	1
	1

Table E.3: Low impact nucleophiles in BERT's inaccurate predictions.

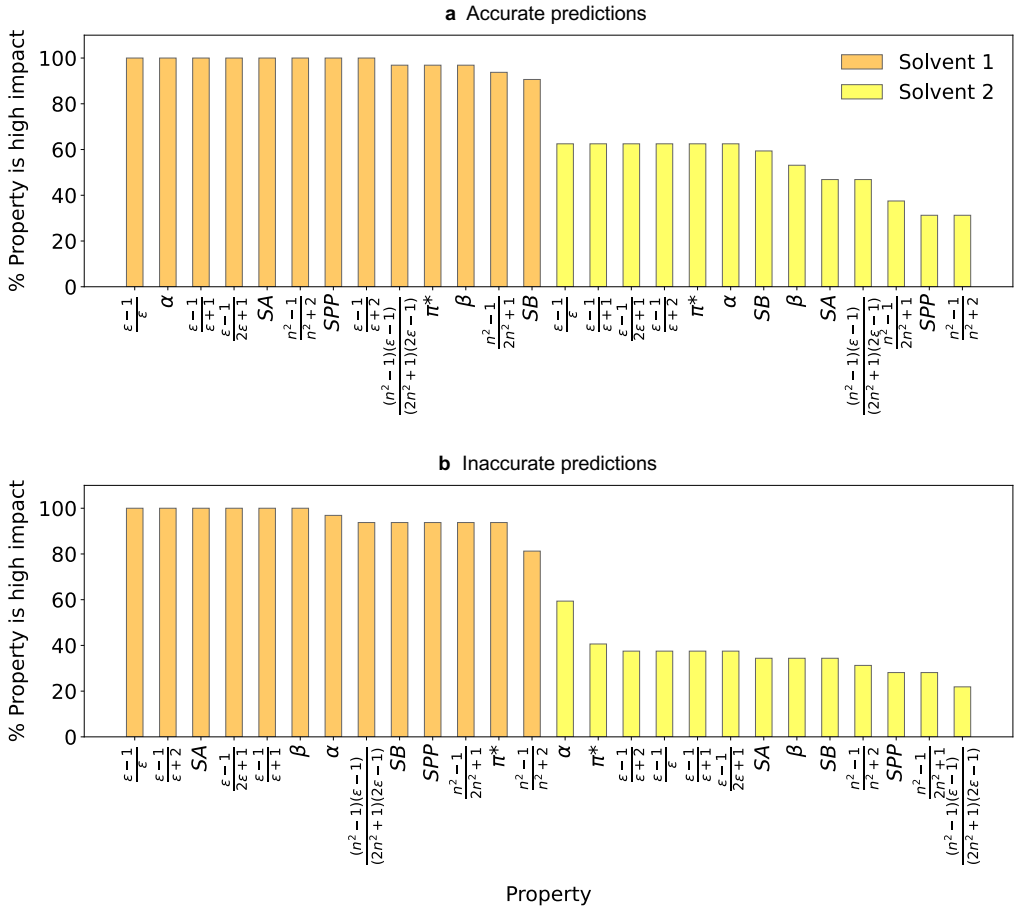


Figure E.1: The percentage of accurate (a) and inaccurate (b) predictions where each solvent property of the RF model was high impact. Here, ε = dielectric constant, n = refractive index, SA, SB, and SPP = Catalán acidity,[4] basicity[5] and polarity/polarisability[6] constants, respectively, and α, β, and π^* = Kamlet-Taft acidity,[7] basicity,[8] and polarity/polarisability[9] constants, respectively.

information is not required to predict reactivity for a single reaction pathway. This said, I and Br (products) increased $\log k_{\text{pred}}$ 100% of the time they were high impact, aligning with the understanding that larger LG atoms increase S_N2 reactivity. This reinforces the conclusion from analysis of reactants that BERT recognises the trend of increasing reactivity with larger halides.

E.4.2 Nucleophilic effects

See Section 3.2.4.2.

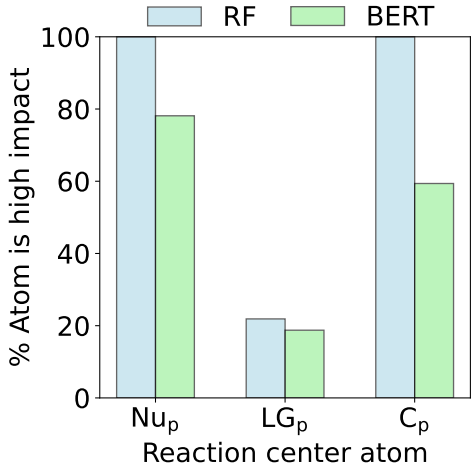


Figure E.2: High impact features of accurate predictions (product mappings). The percentage of accurate predictions where product mappings to the nucleophilic (Nu_p), leaving group (LG_p), and electrophilic carbon (C_p) atoms were high impact features of the RF and BERT models.

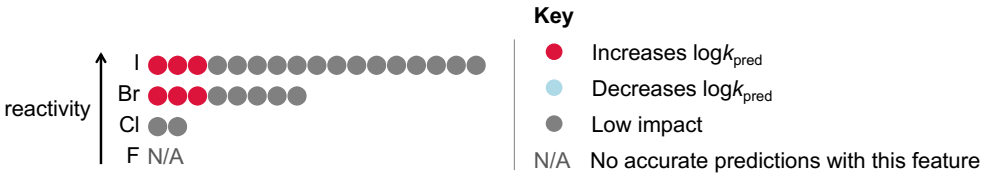


Figure E.3: LG effects (product mappings). Signs (positive or negative) of feature importances for each of BERT's accurate predictions where the product mapping to the LG was high impact. Red and Blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature.

E.4.3 Steric effects

In the BERT model, alkyl-substituted centres were represented on the products side by atom mappings to the electrophilic C and substituting C atoms in the major product (13 reactions), while unsubstituted centres were solely represented by the atom mapping to the electrophilic C (7 reactions). Conversely, alkyl-substituted centres in the RF were represented by C-C, C-C-C, and C-C-C-C fragments (and the importance of these features in reactions with unsubstituted centres used as a control), with no assignment to either reactants or products. Accordingly, the importance of product mappings to the electrophilic C and substituting C atoms in

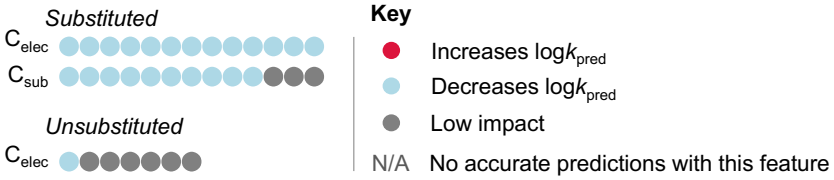


Figure E.4: Steric effects (product mappings). Signs (positive or negative) of feature importances for each of BERT's accurate predictions where product mappings to alkyl-substituted and unsubstituted electrophilic centres were high impact. Red and Blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature. C_{elec} and C_{sub} correspond to electrophilic and substituting carbons, respectively.

the BERT model, either increasing log k_{pred} (positive sign) or decreasing log k_{pred} (negative sign) is presented in Figure E.4 for accurate predictions.

Analysis of steric effects shows that in the BERT model, product mappings to substituted centres decreased log k_{pred} in 100% of examples, reinforcing the conclusion from analysis of reactants that BERT recognises steric hindrance at the electrophilic centre decreases S_N2 reactivity. Conversely, product mappings to unsubstituted centres were less important for predicting reactivity than the reactant-side equivalents (86% low impact compared to 0% for reactants). However, this is logical, as product information is not required to predict reactivity for a single reaction pathway.

E.4.4 Alyllic effects

In the BERT model, product mappings to alkene, alkyne, and aromatic bonds (at the electrophilic centre) were represented by =, ≡, and c tokens in the major product (where the latter is an aromatic carbon bonded to the centre), with the distribution being 2 × alkene, 3 × alkyne, and 4 × aromatic. Meanwhile, alkene, alkyne, and aromatic bonds (at the electrophilic centre) were represented by C-C=C, C-C≡C, and C-C:C fragments in the RF (where ':' is an aromatic bond), with no assignment to either reactants or products. Accordingly, the importance of product mappings to alkene, alkyne, and aromatic bonds (at the electrophilic centre) in the BERT model, either increasing log k_{pred} (positive sign) or decreasing log k_{pred} (negative sign) is presented in Figure E.5 for accurate predictions.

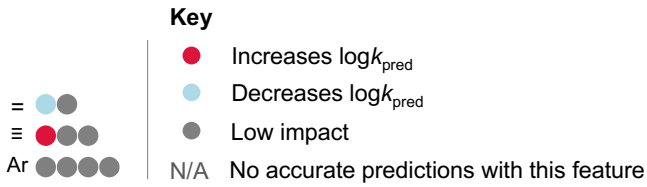


Figure E.5: Allylic effects (product mappings). Signs (positive or negative) of feature importances for each of BERT's accurate predictions where product mappings to alkene (=), alkyne (≡), and aromatic (Ar) bonds to the electrophilic centre were high impact. Red and Blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature.

Product mappings to alkene, alkyne, and aromatic bonds were typically low impact in the BERT model (one of the two reactions with alkene, 67% of reactions with alkyne, and 100% of reactions with aromatic). However, this can be expected as product information is not required to predict reactivity for a single reaction pathway.

F | Chapter 4 supplementary material

F.1 Training distributions

The training distribution of temperatures and $\log k$ values in the E2/S_N2 data can be visualised in Figure F.1.

F.2 Competing reactions in the E2/S_N2 data

In this work, S_N2 reactions with a competing E2 reaction (present in the dataset) were excluded from the E2/S_N2 data to avoid data imbalance (Section 4.1.1). The E2 pathway of 26 of these competing reactions was also removed from the dataset (corresponding to those with $S_T < 0.4$ to the training data) for use as an auxiliary test set. Of these 26 reactions, only four had a significant $\Delta \log k$ (≥ 0.1), shown in Figure F.2. BERT correctly predicted the favoured pathway in two of these examples, corresponding to the reaction of a primary substrate via the S_N2 mechanism (illustrating steric effects, Reaction **W**), and the reaction of a cyclohexyl substrate via the E2 mechanism (illustrating strain effects, Reaction **X**). Notably, the two incorrectly predicted examples were less intuitive, comprising a primary substrate reacting via the E2 mechanism (Reaction **Y**), and a dichloroethane substrate and phenoxide nucleophile, also reacting via the E2 mechanism (Reaction **Z**). To better test BERT's recognition of E2/S_N2 reactivity, three additional datasets were compiled from the literature, with results discussed in Section 4.2.3.

F.3 Data imbalance

It is noted that the number of S_N2 reactions (4757) in the E2/S_N2 training data exceeds the number of E2 reactions (1591). To assess whether this data imbalance impacted model performance, the E2/S_N2 BERT was retrained with the number of S_N2 reactions reduced to 1591 (selected based on lowest latent space similarity

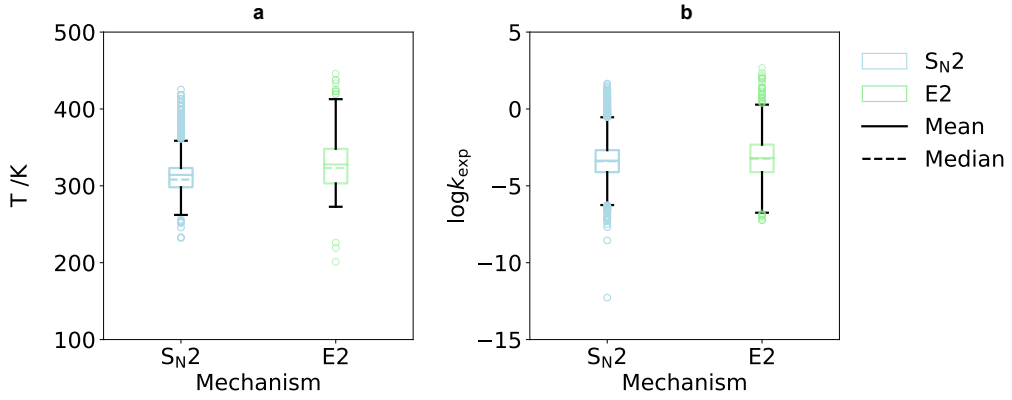


Figure F.1: Training distributions of Temperatures (T, **a**) and log k_{exp} values (**b**) for S_N2 (blue) and E2 (green) reactions in the E2/S_N2 data.

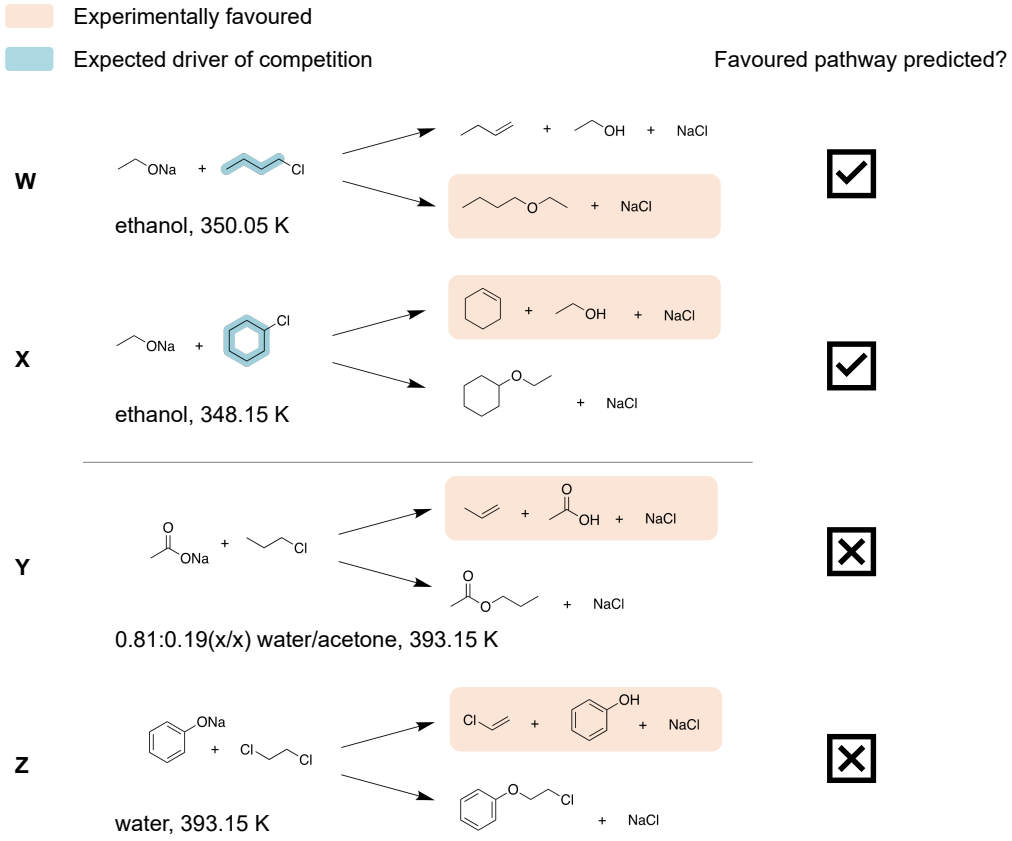


Figure F.2: Auxiliary test set of competing reactions from the E2/S_N2 data with $\Delta \log k \geq 0.1$ and $S_T < 0.4$ to the training data. Reactions **W** and **X** respectively test BERT's recognition of steric, and strain effects, while Reactions **Y** and **Z** have less intuitive outcomes.

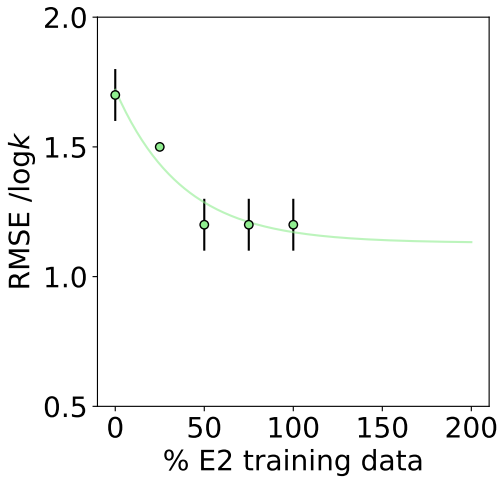


Figure F.3: Learning curve depicting RMSE convergence in the E2/S_N2 BERT with respect to the percentage of E2 data added to the training set (evaluated using the E2 test data). Note that the data point at 0% corresponds to the evaluation of $BERT_{S_N2}$ on the E2 test data.

to the rest of the S_N2 training data). When trained on this balanced dataset, BERT performed comparably to when it was trained on the full, imbalanced dataset (RMSE: 1.2 ± 0.1 log k), suggesting that accuracy was unaffected by the difference in dataset size. Additionally, when evaluated on the E2 test data, the E2/S_N2 BERT (trained on the full, imbalanced dataset) achieved convergence ($-\frac{\partial RMSE}{\partial \%E2} < 0.01$) within the total number of E2 training reactions (Figure F.3), confirming a sufficient number of E2 reactions were used to train the model.

F.4 Fine-tuning BERT

The BERT model trained for S_N2 rate prediction in Chapter 3 ($BERT_{S_N2}$) was fine-tuned in this study to predict experimental log k values for both E2 and S_N2 reactions. This was achieved using a combined dataset of 1591 E2 reactions compiled by Madzhidov et al.[26] and 4757 of the S_N2 training reactions from Chapter 3, after excluding those with a competing E2 reaction present in the dataset. The fine-tuned model achieved RMSE convergence ($-\frac{\partial RMSE}{\partial epoch} < 0.1$) in significantly fewer epochs than training directly on the E2/S_N2 data (avg. 3 epochs compared to 10, Figure F.4), highlighting the effectiveness of this approach.

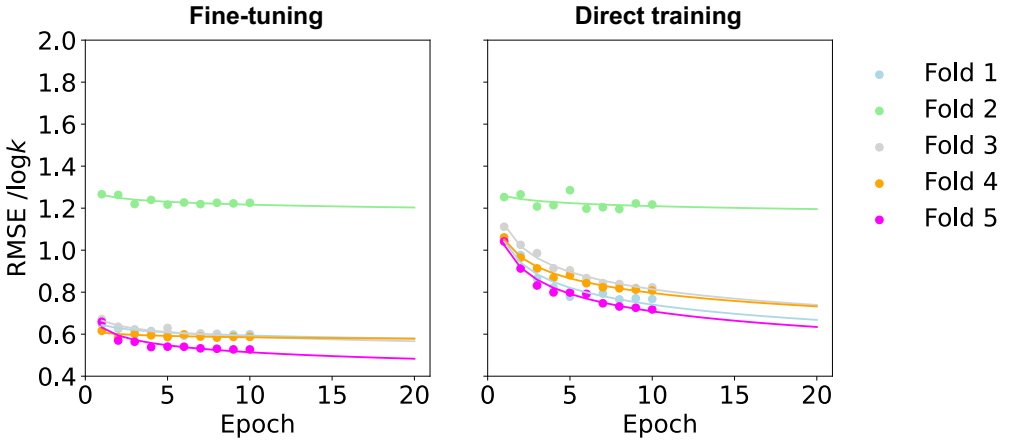


Figure F.4: RMSE convergence with respect to epoch number when fine-tuning the $BERT_{S_N2}$ model, and training directly on the E2/S_N2 data. Evaluated per fold using validation data.

F.5 RMSE convergence (Green's data)

When BERT was trained using Green's data, RMSE convergence (defined as $-\frac{\partial RMSE}{\partial x} < 0.1$, where x is some model variable such as epoch number or hyperparameter optimisation trial) was achieved within 20 hyperparameter optimisation trials for four out of five CV folds (folds 2 to 5), with $-\frac{\partial RMSE}{\partial trial} = 0.01$ for one fold (fold 1, Figure F.5**a**). To test whether accuracy could be improved by optimising fold 1's hyperparameters until convergence, the expected RMSE improvement, ξ, was calculated using Eq. 3.5. Eq. 3.5 returned an ξ value of 0.0 $\log k$, indicating that accuracy would not be improved by optimising until convergence. Consequently, the hyperparameters tuned over 20 trials were employed.

Regarding epoch number, convergence was achieved within 10 epochs for two out of five CV folds (folds 3 and 4), with $-\frac{\partial RMSE}{\partial epoch} = 0.01$ for two folds (folds 2 and 5), and $-\frac{\partial RMSE}{\partial epoch} = 0.04$ for one fold (fold 1, Figure F.5**b**). To test whether accuracy could be improved by training on folds 1, 2, and 5 until convergence, the expected RMSE improvement, ξ, was calculated using Eq. 3.5. Eq. 3.5 returned an ξ value of 0.0 $\log k$, indicating that accuracy would not be improved by training until convergence. Consequently, the estimators trained for 10 epochs were employed.

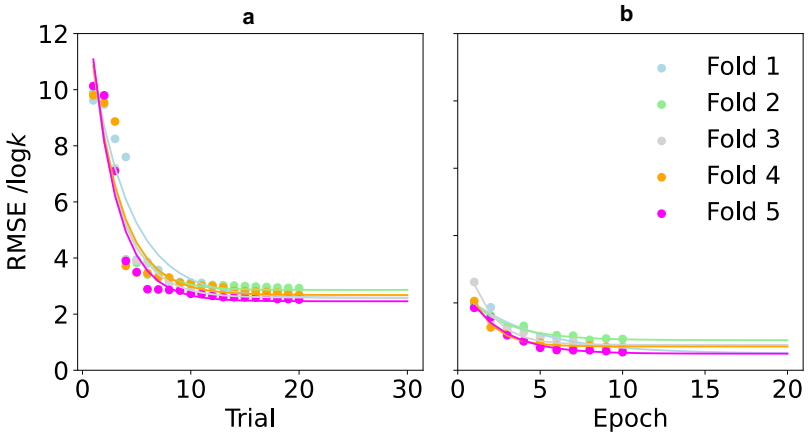


Figure F.5: Learning curves depicting RMSE convergence with respect to hyperparameter optimisation trial (**a**) and epoch number (**b**) for the BERT model trained using Green's dataset. Evaluated per fold using validation data.

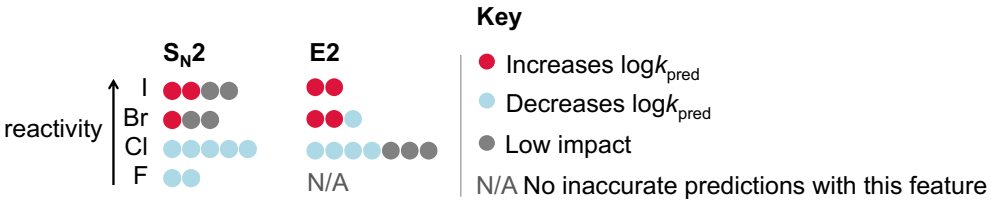


Figure F.6: Analysis of LG effects (inaccurate predictions). Signs (positive or negative) of feature importances for each of BERT's inaccurate predictions where the LG was high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no inaccurate predictions containing this feature.

F.6 Interpreting inaccurate predictions

F.6.1 LG effects

The feature importances of halide LGs contributing to the predicted rate, either increasing (positive sign) or decreasing (negative sign) log k_{pred} are presented in Figure F.6 for BERT's inaccurate predictions, with the distribution being 4 × I, 3 × Br, 5 × Cl, and 2 × F for S_N2 and 2 × I, 3 × Br, 7 × Cl, and 0 × F for E2. Note that the total distribution of halide LGs in the test set was 53 × I, 16 × Br, 16 × Cl, and 4 × F for S_N2 and 3 × I, 22 × Br, 22 × Cl, and 2 × F for E2.

Analysis of inaccurate predictions reveals that I and Br were predicted to increase $\log k$ in 100% of S_N2 reactions where they were high impact (2 reactions, and one reaction, respectively), while Cl and F decreased $\log k_{pred}$ in 100% of analysed S_N2 examples. This suggests that BERT identified a positive correlation between halide size and S_N2 reactivity, consistent with known LG effects.

A positive correlation between halide size and reactivity was also observed in the inaccurate predictions of E2 reactions. Here, BERT predicted I to increase $\log k$ in 100% of analysed examples, while Cl decreased $\log k_{pred}$ in 100% of examples where it was high impact (4 reactions). Meanwhile, Br exhibited mixed impact, increasing $\log k_{pred}$ in two reactions and decreasing $\log k_{pred}$ in one reaction. Thus, the E2/S_N2 BERT successfully recognised key LG effects across inaccurate predictions of both E2 and S_N2 reactions. This is consistent with the conclusions for accurate predictions (Section 4.2.2.1), reinforcing the reliability of these assessments.

F.6.2 Alkyl substituent effects

To analyse alkyl substituent effects, inaccurate predictions with primary, secondary, tertiary, and unsubstituted centres were evaluated (Figure F.7), with the distribution being 0 × unsubstituted, 7 × primary, 1 × secondary, and 0 × tertiary for S_N2 and 1 × primary, 5 × secondary, and 4 × tertiary for E2 (unsubstituted centres, with no β-Hs, cannot undergo E2). Note that the total distribution of electrophilic centres in the test set was 24 × unsubstituted, 36 × primary, 3 × secondary, and 0 × tertiary for S_N2 and 4 × primary, 19 × secondary, and 12 × tertiary for E2. In the BERT model, alkyl-substituted centres are represented by the electrophilic C and substituting C atoms, while unsubstituted centres are solely represented by the electrophilic C. The β-C was not considered a substituting C in the E2 mechanism as it does not contribute to the stabilisation of the forming alkene bond in the TS.

Analysis of inaccurate predictions reveals that substituted centres decreased $\log k_{pred}$ in 100% of S_N2 reactions where they were high impact (3 × primary centres), which aligns with the understanding that S_N2 reactivity is decreased by steric hindrance at the electrophilic centre. Moreover, reactions with unsubstituted centres were not predicted inaccurately by the BERT model, confirming that learning of steric effects was not a cause of model error.

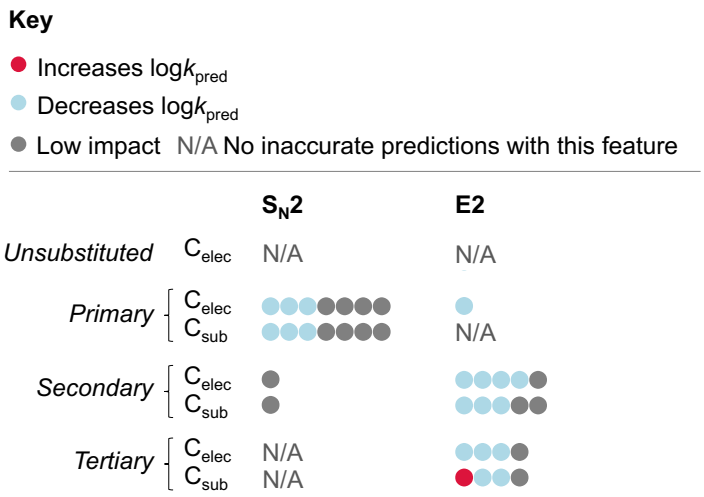


Figure F.7: Analysis of alkyl substituent effects (inaccurate predictions). Signs (positive or negative) of feature importances for each of BERT's inaccurate predictions where primary, secondary, tertiary, and unsubstituted electrophilic centres were high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no inaccurate predictions containing this feature. C_{elec} and C_{sub} correspond to electrophilic and substituting carbons, respectively.

Regarding inaccurate predictions of E2 reactions, BERT predicted primary, secondary, and tertiary substituted centres to decrease log k in 70% of analysed examples. This is consistent with the results for accurate predictions, which was attributed to an experimental correlation between increased steric bulk and decreased reactivity in these reactions (Section 4.2.2.2). Notably, substituting alkyl groups increased log k_{pred} in one reaction with a tertiary centre (corresponding to the reaction of sodium thiophenolate with *tert*-butyl bromide in formamide at 298.15 K), however this is too few examples to indicate a general trend.

In conclusion, BERT correctly identified that S_N2 reactivity is decreased by steric hindrance at the electrophilic centre. Meanwhile, BERT's modelling of alkyl substitutents and E2 reactivity was consistent with an experimental correlation between increased steric bulk and decreased reactivity in these reactions. The conclusions presented here are consistent with those for accurate predictions (Section 4.2.2.2), reinforcing the reliability of these assessments.

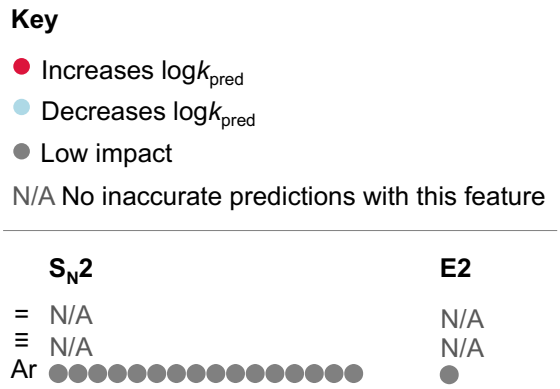


Figure F.8: Analysis of allylic effects (inaccurate predictions). Signs (positive or negative) of feature importances for each of BERT's inaccurate predictions where alkene (=), alkyne (≡), and aromatic (Ar) bonds to the electrophilic centre were high impact. Red and blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no inaccurate predictions containing this feature.

F.6.3 Allylic effects

Inaccurate predictions with alkene, alkyne, and aromatic bonds (to the electrophilic centre) were evaluated, with the distribution being 0 × alkene, 0 × alkyne, and 16 × aromatic for S_N2 and 0 × alkene, 0 × alkyne, and 1 × aromatic for E2 (Figure F.8). The total distribution of centres in the test set was 2 × alkene, 4 × alkyne, and 42 × aromatic for S_N2, and 0 × alkene, 0 × alkyne, and 6 × aromatic for E2. In the BERT model, alkene, alkyne, and aromatic bonds at the electrophilic centre were represented by and =, ≡, and c tokens (where c is an aromatic carbon bonded to the centre), respectively.

Notably, S_N2 reactions with alkene and alkyne bonds (to the electrophilic centre) were not predicted inaccurately by the BERT model. This suggests that recognition of allylic effects from these groups in S_N2 reactions was not a cause of model error. Meanwhile, aromatic groups were low impact in 100% of inaccurately predicted S_N2 reactions (where they were at the electrophilic centre), suggesting BERT had not recognised the significance of these groups in enhancing S_N2 reactivity.

Regarding E2 reactions, no examples with alkene or alkyne bonds (to the electrophilic centre) were present in the test data and only one example with

an aromatic group (at the electrophilic centre) was predicted inaccurately. Accordingly, allylic effects were not analysed for inaccurate predictions of E2 reactions.

In summary, allylic effects (from alkene and alkyne bonds) in S_N2 reactions were not a source or error for the BERT model. However, recognition of aromatic effects in these reactions was identified as a model limitation. This is consistent with the conclusions for accurate predictions (Section 4.2.2.3), reinforcing the reliability of these assessments. Meanwhile, due to limited examples of E2 reactions with alkene, alkyne, and aromatic groups (at the electrophilic centre), allylic effects were not analysed for inaccurate predictions of E2 reactions.

F.7 Product atom mappings

F.7.1 LG effects

In the BERT model, product mappings to halide LGs were represented by [I^-], [Br^-], [Cl^-], and [F^-] tokens, with the distribution of LGs being 17 × I, 7 × Br, 2 × Cl, and 0 × F for S_N2 and 0 × I, 7 × Br, 6 × Cl, and 1 × F for E2. The importance of [I^-], [Br^-], [Cl^-], and [F^-] tokens in the BERT model, either increasing $\log k_{\text{pred}}$ (positive sign) or decreasing $\log k_{\text{pred}}$ (negative sign) is presented in Figure F.9 for accurate predictions.

Product mappings to halide LGs were less important for predicting S_N2 reactivity than their reactant-side counterparts, with I, Br, and Cl being low impact 82%, 100%, and 50% of the time in the products, compared to 24%, 57% and 50% in the reactants. Notably, I increased $\log k_{\text{pred}}$ in the three examples where it was high impact, aligning with the understanding that larger LG atoms increase S_N2 reactivity. Unusually, Cl was predicted to increase $\log k_{\text{pred}}$ in the one S_N2 example where it was high impact (corresponding to the reaction of potassium bis(4-bromophenyl)amide with benzyl chloride in DMSO at 298.15 K), which contradicts the understanding that smaller LG atoms decrease S_N2 reactivity. Importantly, analysis of reactants showed Cl to decrease $\log k_{\text{pred}}$ in this reaction (Section 4.2.2.1), aligning with known LG effects.

Product mappings to halide LGs were similarly insignificant for predicting E2 reactivity, having low impact in 100% of analysed examples. Nonetheless, BERT successfully recognised the influence of these LGs in the reactants (Section 4.2.2.1).

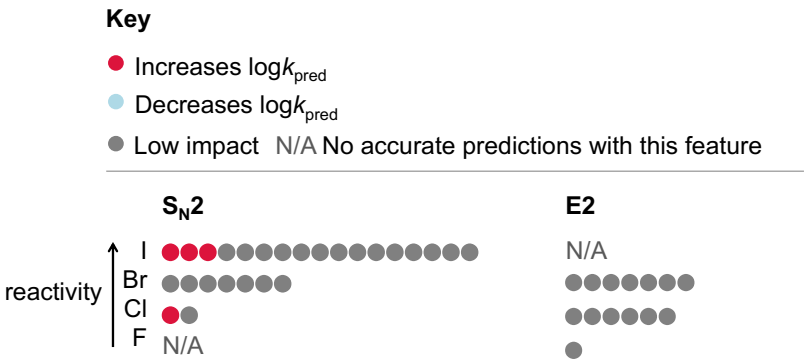


Figure F.9: LG effects (product mappings). Signs (positive or negative) of feature importances for each of BERT's accurate predictions where the product mapping to the LG was high impact. Red and Blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature.

F.7.2 Alkyl substituent effects

In the BERT model, alkyl-substituted centres were represented on the products side by atom mappings to the electrophilic C and substituting C atoms in the major product, while unsubstituted centres were solely represented by the atom mapping to the electrophilic C. The importance of product mappings to the electrophilic C and substituting C atoms in the BERT model, either increasing log k_{pred} (positive sign) or decreasing log k_{pred} (negative sign) is presented in Figure F.10 for accurate predictions, with the distribution being 8 × unsubstituted, 11 × primary, 1 × secondary, and 0 × tertiary for S_N2 and 2 × primary, 4 × secondary, and 3 × tertiary for E2 (unsubstituted centres, with no β-Hs, cannot undergo E2).

Analysis of alkyl substituent effects in S_N2 reactions shows that product mappings to substituted centres decreased log k_{pred} 92% of the time, reinforcing the conclusion from the analysis of reactants that BERT recognises that S_N2 reactivity is decreased by steric hindrance at the electrophilic centre. Unusually, product mappings to unsubstituted centres were also predicted to decrease S_N2 reactivity, contradicting the understanding that these centres typically enhance reactivity in these reactions. Notably, this behaviour was not observed in the $BERT_{S_N2}$ model, with product mappings to unsubstituted centres having low impact in 86% of analysed reactions. It is unexpected that introduction of E2 reactions into the training data led to these centres decreasing log k_{pred}, given that unsubstituted centres do not react via the

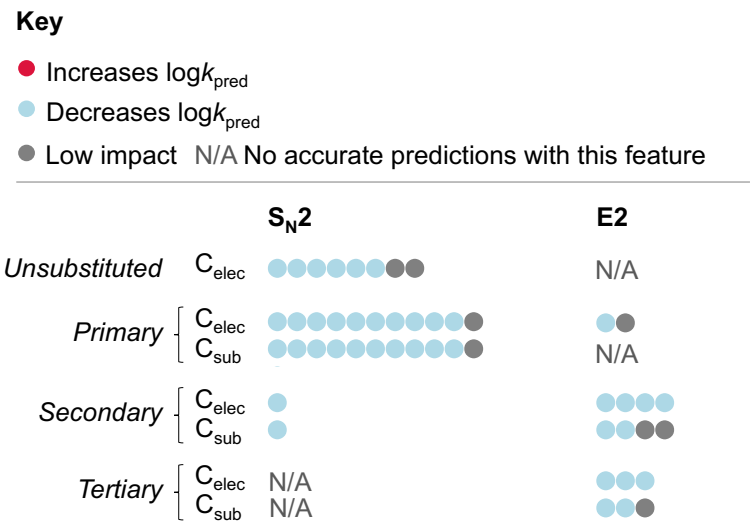


Figure F.10: Alkyl substituent effects (product mappings). Signs (positive or negative) of feature importances for each of BERT's accurate predictions where product mappings to primary, secondary, tertiary, and unsubstituted electrophilic centres were high impact. Red and Blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature. C_{elec} and C_{sub} correspond to electrophilic and substituting carbons, respectively.

E2 mechanism. Importantly, analysis of reactants showed that the E2/S_N2 BERT predicted unsubstituted centres to increase log k_{pred} in 100% of analysed examples, indicating the model had indeed learned this effect (Sections 4.2.2.2 and F.6.2).

Regarding E2 reactions, product mappings to primary, secondary, and tertiary electrophilic centres decreased log k_{pred} in 100% of examples where they were high impact. This is consistent with the results for reactants, which was attributed to an experimental correlation between increased steric bulk and decreased reactivity in these reactions (Section 4.2.2.2).

F.7.3 Alyllic effects

In the BERT model, product mappings to alkene, alkyne, and aromatic bonds (at the electrophilic centre) were represented by =, ≡, and c tokens in the major product (where the latter is an aromatic carbon bonded to the centre), with the distribution being 1 × alkene, 3 × alkyne, and 5 × aromatic for S_N2 and 0 × alkene, 0 × alkyne, and 3 × aromatic for E2. The importance of product mappings

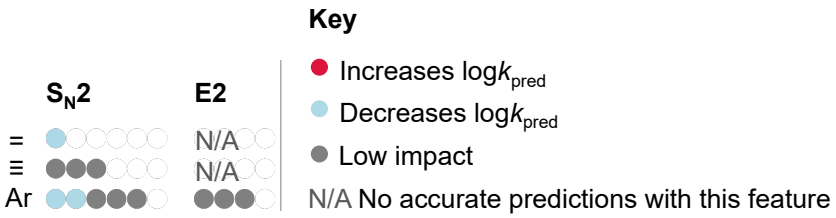


Figure F.11: Allylic effects (product mappings). Signs (positive or negative) of feature importances for each of BERT's accurate predictions where product mappings to alkene (=), alkyne (≡), and aromatic (Ar) bonds to the electrophilic centre were high impact. Red and Blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature.

to alkene, alkyne, and aromatic bonds (at the electrophilic centre) in the BERT model, either increasing log k_{pred} (positive sign) or decreasing log k_{pred} (negative sign) is presented in F.11 for accurate predictions.

Product mappings to alkene, alkyne, and aromatic bonds (to the electrophilic centre) were either low impact or decreased log k_{pred} in 100% of S_N2 reactions analysed. For alkyne bonds, which were predicted to increase log k_{pred} in the analysis of reactants, the product mappings were all low impact. This does not contradict the conclusions from the reactants, but indicates that the products were insignificant in modelling this effect. Conversely, for the one S_N2 example with an alkene bond, this group was low impact in the reactants, but decreased log k_{pred} in the products. While this contradicts the understanding that allylic groups at the electrophilic centre increase S_N2 reactivity, general conclusions cannot be made based on this single example. Meanwhile, product mappings to aromatic groups at the electrophilic centre were either low impact or decreased log k_{pred} in 100% of S_N2 reactions analysed. This reinforces the conclusions from the analysis of reactants that BERT is limited in recognising this effect.

Regarding E2 reactions, no examples with alkene or alkyne bonds (to the electrophilic centre) were present in the test data. Of the three accurate E2 predictions with an aromatic group at the electrophilic centre, the product mapping was low impact 100% of the time, consistent with this group's low significance in the reactants.

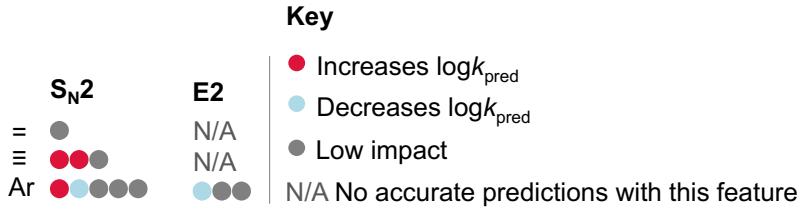


Figure F.12: Analysis of allylic effects (summing over ring tokens). Signs (positive or negative) of feature importances for each of BERT's accurate predictions where alkene (=), alkyne (≡), and aromatic (Ar) bonds to the electrophilic centre were high impact. In contrast to the results presented in Figure 4.8, here, aromatic effects were evaluated by summing the importances over ring tokens. Red and Blue circles denote predictions where the feature importance was positive (increased log k_{pred}), and negative (decreased log k_{pred}), respectively. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes there were no accurate predictions containing this feature.

F.8 Allylic effects (summing over ring tokens)

To validate that the results for aromatic groups were not an artifact of only considering the carbon bonded to the electrophilic centre (as opposed to the entire aromatic ring), the importances were also calculated by summing over the ring tokens (Figure F.12). The results were consistent across both approaches, confirming they were not biased by the way the importances were calculated.

F.9 Temperature effects

F.9.1 S_N2 synthetic data points

While BERT successfully predicted the linear decrease in log k with increasing T^{-1} for E2 reactions within the training IQR, it was limited in its ability to extrapolate beyond this range (Section 4.2.2.4). To test whether this phenomenon also occurs for S_N2 reactions, 20 synthetic data points were generated for the four S_N2 test reactions observed at multiple temperatures (Reactions **A** to **D**, Figure 3.20), corresponding to 10 data points with $T < 286$, and 10 data points with $T > 364$ (falling outside the IQR of temperatures in the S_N2 training data). Indeed, inclusion of these temperatures reduced the correlation between T^{-1} and log k_{pred} significantly ($-r_P \leq 0.69$ compared to $-r_P = 0.99$ without these

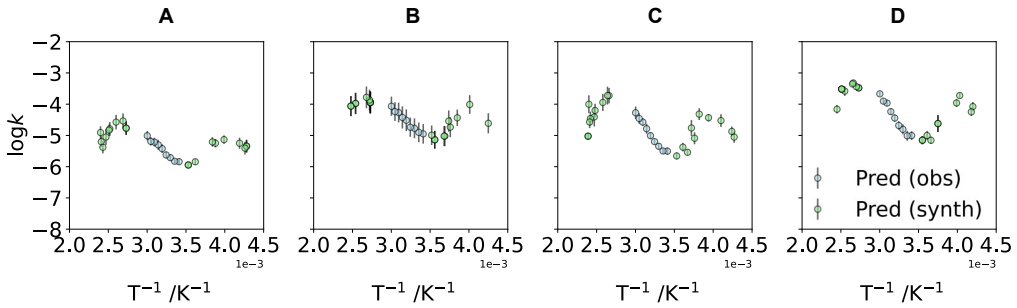


Figure F.13: Analysis of temperature effects (S_N2 synthetic temperatures). Predicted (Pred) $\log k$ values for the four S_N2 test reactions observed at multiple temperatures (Reactions **A** to **D**, Figure 3.20). Predictions for observed data points (obs) are shown in blue and predictions for synthetic data points (synth) are shown in green. As $\log k_{exp}$ is not known for the synthetic data points, the ground truth data has been omitted. Here, T^{-1} denotes the reciprocal temperature.

temperatures, Figure F.13), suggesting BERT is limited in its ability to extrapolate to temperatures outside the training IQR for both mechanisms.

F.9.2 RF

To test whether limitations in extrapolating to temperatures outside the training IQR is unique to the BERT architecture, $\log k$ predictions for each of the S_N2 (Reactions **A** to **D**, Figure 3.20) and E2 (Reactions **E** to **G**, Figure 4.10) test reactions observed at multiple temperatures were also made using an RF model trained on the E2/S_N2 dataset (Figure F.14). Indeed, including the data points with T < 286 and T > 364 (falling outside the training IQR of both mechanisms) reduced the negative correlation between T^{-1} and $\log k$ for both mechnanisms, compared to when these data points were excluded (r_P = -0.85 to -0.87 compared to r_P -0.97 for S_N2, and r_P > 0 compared to r_P = -0.66 to -0.91 for E2). This suggests that extrapolating to temperatures outside the training IQR is a common challenge across both the BERT and RF models.

F.10 Literature datasets

F.10.1 ΔIG values of reactants

Other than the exceptions outlined in table F.1, the ΔIG values of all reactant tokens in the literature datasets were zero within error.

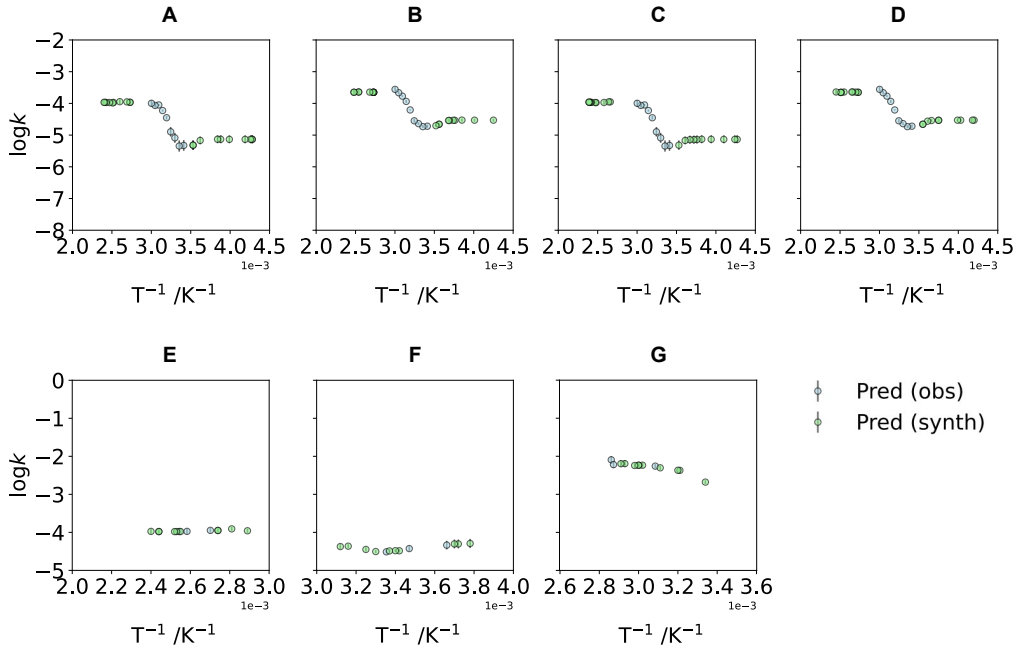


Figure F.14: Analysis of temperature effects (RF). Predicted (Pred) log k values for S_N2 (**A** to **D**) and E2 (**E** to **G**) test reactions observed at multiple temperatures. Predictions for observed data points (obs) are shown in blue and predictions for synthetic data points (synth) are shown in green. As log k_{exp} is not known for the synthetic data points, the ground truth data has been omitted. Note that different x-axis ranges have been used for the E2 reactions to accommodate different temperature ranges. Similarly, different y-axis ranges have been used for E2 and S_N2 reactions to accommodate different log k ranges. Here, T^{-1} denotes the reciprocal temperature.

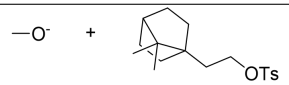
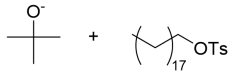
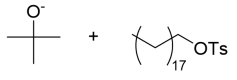
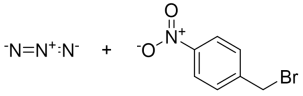
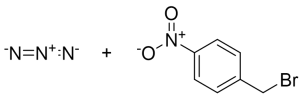
Dataset	Reaction	Token	Token description	$\Delta_{S_N2\text{-}E2}IG$
Competing		2	Ring closure in 2'-5,5-dimethylbicyclo[2.2.1]heptan-1-yl	-0.02 ± 0.01
Competing		C	Carbon in octadecyl	0.01 ± 0.00
Competing		C	Carbon in octadecyl	0.02 ± 0.01
Competing		)	Bracket in 2'-5,5-dimethylbicyclo[2.2.1]heptan-1-yl	-0.01 ± 0.00
Adversarial		O	Oxygen in NO_2	0.03 ± 0.02
Adversarial		=	Double bond in N_3^-	0.04 ± 0.01

Table F.1: Non-zero ΔIG values of reactant tokens in the literature datasets.

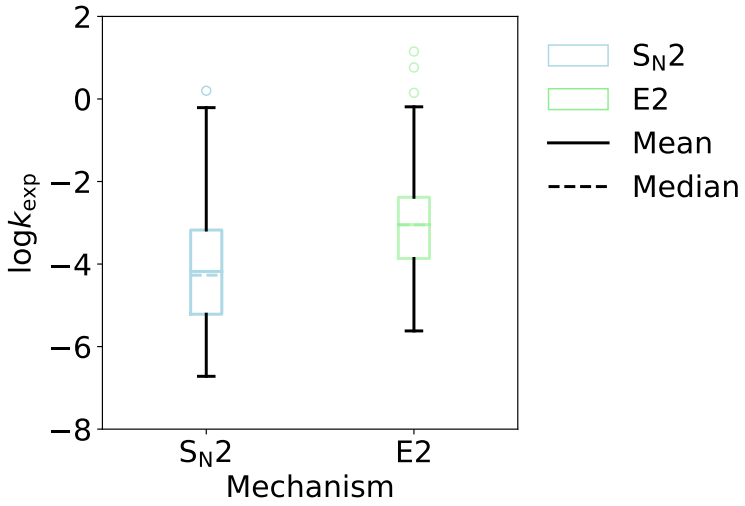


Figure F.15: log k **distribution for reactions with tosylate LGs** in the E2 (green) and S_N2 (blue) training data.

F.10.2 log k distribution for tosylate LGs

Of the three reactions in the Competing dataset with a bulky base and tosylate LG (which favour the S_N2 pathway), BERT had not identified the LG's significance in predicting the reaction outcome (Section 4.2.3.1). This oversight was attributed to log k_{exp} values of training reactions with tosylate LGs, which were typically smaller for S_N2 reactions than E2 reactions (IQR: -5.2 to -3.2 for S_N2 and -3.9 to -2.4 for E2, Figure F.15). A similar connection was noted in Madzhidov et al.'s early work[28] with the S_N2 dataset (using an SVR algorithm), where log k was often underpredicted for reactions with substituted tosylate LGs, attributed to the training data's log k range.

Appendix bibliography

(1) Ba, J.; Kiros, J.; Hinton, G. *arXiv preprint arXiv:1607.06450v1* **2016**.

(2) Schwaller, P.; Probst, D.; Vaucher, A.; Nair, V.; Kreutter, D.; Laino, T.; Reymond, J.-L. *Nat. Mach. Intell.* **2021**, *3*, 144–152.

(3) Rumelhart, D.; Hintont, G.; Williams, R. *Nature* **1986**, *323*, 533–536.

(4) Catalán, J.; Díaz, C. *Liebigs Ann.* **2006**, *1997*, 1941–1949.

(5) Catalán, J.; Díaz, C.; López, V.; Pérez, P.; De Paz, J.-L.; Rodríguez, J.-G. *Liebigs Ann.* **2006**, *1996*, 1785–1794.

(6) Catalán, J.; López, V.; Pérez, P.; Martin-Villamil, R.; Rodríguez, J.-G. *Liebigs Ann.* **2006**, *1995*, 241–252.

(7) Taft, R.; Kamlet, M. *J. Am. Chem. Soc.* **1976**, *98:10*, 2886–2894.

(8) Kamlet, M.; Taft, R. *J. Am. Chem. Soc.* **1976**, *98:2*, 377–383.

(9) Kamlet, M.; Abboud, J.; Taft, R. *J. Am. Chem. Soc.* **1977**, *19:18*, 6027–6038.

(10) Grambow, C.; Pattanaik, L.; Green, W. *J. Phys. Chem. Lett.* **2020**, *11*, 2992–2997.

(11) Ismail, I.; Robertson, C.; Habershon, S. *J. Chem. Phys.* **2022**, *157*, 014109.

(12) Heid, E.; Green, W. *J. Chem. Inf. Model.* **2022**, *62*, 2101–2110.

(13) Spiekermann, K.; Pattanaik, L.; Green, W. *J. Phys. Chem. A* **2022**, *126*, 3976–3986.

(14) van Gerwen, P.; Briling, K.; Bunne, C.; Somnath, V.; Laplaza, R.; Krause, A.; Corminboeuf, C. *arXiv preprint arXiv:2312.08307* **2023**.

(15) Heinen, S.; von Rudorff, G.; von Lilienfeld, O. *J. Chem. Phys.* **2021**, *155*, 064105.

(16) Stuyver, T.; Coley, C. *J. Chem. Phys.* **2022**, *156*, 084104.

(17) Vijay, S.; Venetos, M.; Spotte-Smith, E.; Kaplan, A.; Wen, M.; Persson, K. *Chem. Sci.* **2024**, *15*, 2923–2936.

(18) Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. *Chem. Sci.* **2020**, *11*, 4584–4601.

(19) Stuyver, T.; Coley, C. *Chemistry* **2023**, *29*, e202300387.

(20) Farrar, E.; Grayson, M. *Chem. Sci.* **2022**, *13*, 7594–7603.

(21) Espley, S.; Farrar, E.; Buttar, D.; Tomasi, S.; Grayson, M. *Digit. Discov.* **2023**, *2*, 941–951.

(22) Lewis-Atwell, T.; Beechey, D.; Şimşek, Ö.; Grayson, M. *ACS Catal.* **2023**, *13*, 13506–13515.

(23) Garcia-Andrade, X.; Garcia Tahoces, P.; Perez-Rios, J.; Martinez Nunez, E. *J. Phys. Chem. A* **2023**, *127*, 2274–2283.

(24) Kravtsov, A.; Karpov, P.; Baskin, I.; Palyulin, V.; Zefirov, N. *Dokl. Chem.* **2011**, *440*, 299–301.

(25) Madzhidov, T.; Polishchuk, P.; Nugmanov, R.; Bodrov, A.; Lin, A.; Baskin, I.; Varnek, A.; Antipin, I. *Russ. J. Org. Chem.* **2014**, *50*, 459–463.

(26) Madzhidov, T.; Bodrov, A.; Gimadiev, T.; Nugmanov, R.; Antipin, I.; Varnek, A. *J. Struct. Chem.* **2016**, *56*, 1227–1234.

(27) Madzhidov, T.; Gimadiev, T.; Malakhova, D.; Nugmanov, R.; Baskin, I.; Antipin, I.; Varnek, A. *J. Struct. Chem.* **2017**, *58*, 650–656.

(28) Gimadiev, T.; Madzhidov, T.; Tetko, I.; Nugmanov, R.; Casciuc, I.; Klimchuk, O.; Bodrov, A.; Polishchuk, P.; Antipin, I.; Varnek, A. *Mol. Inform.* **2019**, *38*, e1800104.

(29) Rakhimbekova, A.; Akhmetshin, T.; Minibaeva, G.; Nugmanov, R.; Gimadiev, T.; Madzhidov, T.; Baskin, I.; Varnek, A. *SAR QSAR Environ. Res.* **2021**, *32*, 207–219.

(30) Zankov, D.; Madzhidov, T.; Baskin, I.; Varnek, A. *Mol. Inform.* **2023**, *42*, e2200275.

(31) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. *Chem. Sci.* **2021**, *12*, 1163–1175.

(32) Chemcraft - graphical software for visualization of quantum chemistry computations, Version 1.8, build 682. URL: `https://www.chemcraftprog.com` (last accessed 11/6/2024).

(33) Young, T. duartegroup/otherm: Major Symmetry Improvements, 2020.

(34) DeTar, D.; McMullen, D.; Luthra, N. *J Am Chem Soc* **1978**, *100:8*, 2484–2493.

(35) Eyring, H. *Chem. Rev.* **1935**, *17*, 65–77.