Supporting Information

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

Chloe Wilson,*[a] María Calvo,[a] Stamatia Zavitsanou,[a] James Somper,[a] Ewa Wieczorek,[a] Tom Watts,[a] Jason Crain[b] and Fernanda Duarte*[a]

[a]Physical and Theoretical Chemistry Laboratory, University of Oxford, 12 Mansfield Road, Oxford, OX1 3TA, UK, e-mail: fernanda.duartegonzalez@chem.ox.ac.uk, chloe12345wilson@hotmail.com;

[b]IBM Research, The Hartree Centre, STFC Laboratory, Sci-Tech Daresbury, Warrington, WA4 4AD, UK

1 Dataset analysis

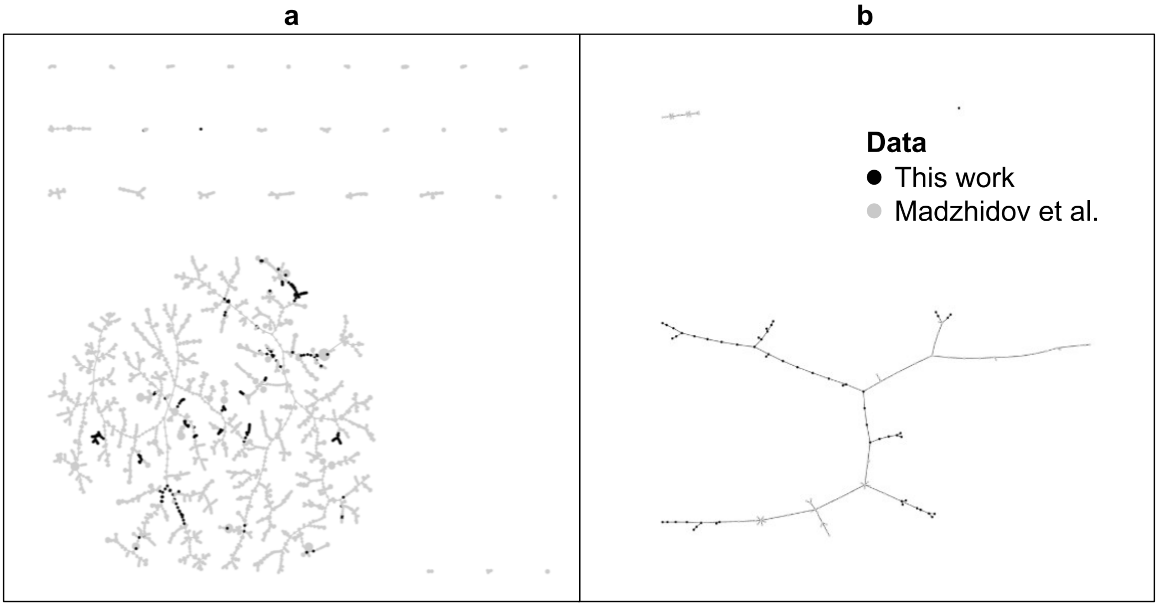


Figure 1: **Tree maps (TMAPs) of the training and test data utilised in this work. a.** The total training set of 4862 S_N2 reactions. 4666 of these were compiled by Madzhidov et al.[1] (shown in grey), and 196 were added in the current work to increase the chemical diversity in the training data (shown in black). **b.** The total test set of 129 S_N2 reactions. 73 of these were compiled by Madzhidov et al.[1] (shown in grey), and 56 were added in the current work to represent a broader area of chemical space (shown in black). Interactive versions of each TMAP are provided on GitHub (see *Data availability* in the main text).

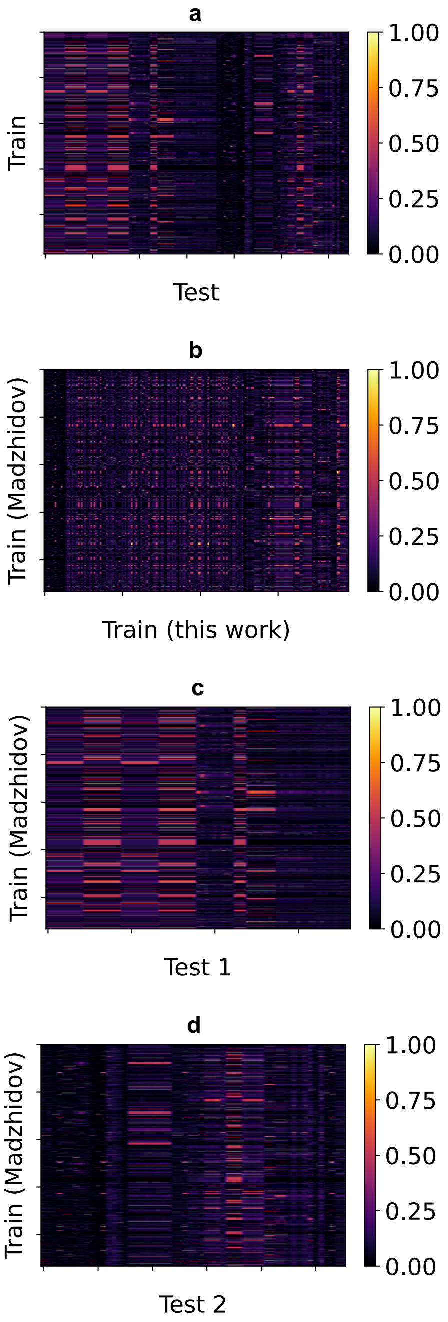


Figure 2: **Tanimoto similarity maps. a.** Tanimoto similarity of the 4862 reactions used to train the RF and BERT models in this work to the total 129 test reactions. **b to d.** Tanimoto similarity of the 4666 training reactions compiled by Ref[1] to: the 196 training reactions manually curated in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

1.1 Relative RMSE decrease from diversifying the training data

The relative decrease in RMSE from diversifying the training data is visualised in Figure 3 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 and -0.1 ± 0.2 $\log k$ for phosphine). 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 an importance in the upper quartile for the test data) for 100% of test reactions in electrolyte.

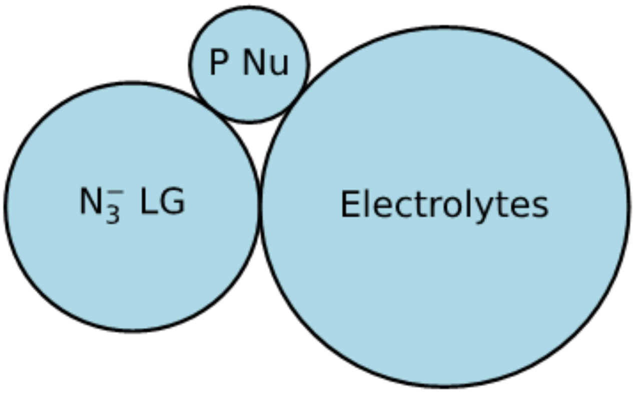


Figure 3: Relative RMSE decrease (represented by the circle radii) from diversifying the training data, evaluated using reactions in Test 2 with electrolyte solvent (12 reactions), azide LGs (N_3^- LG, 5 reactions), and phosphine nucleophiles (P Nu, 4 reactions).

2 Model training

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 4**a**, evaluated using the total test data Test 1 + Test 2), epoch number (BERT only, Figure 4**b**, evaluated using validation data), or hyperparameter optimisation trial (Figure 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,

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

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). Eq. 1 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 BERT, and Grid Search for 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 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 $-\frac{\partial RMSE}{\partial x} < 0.01$ 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.

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 1: Optimised hyperparameter values in the BERT model.

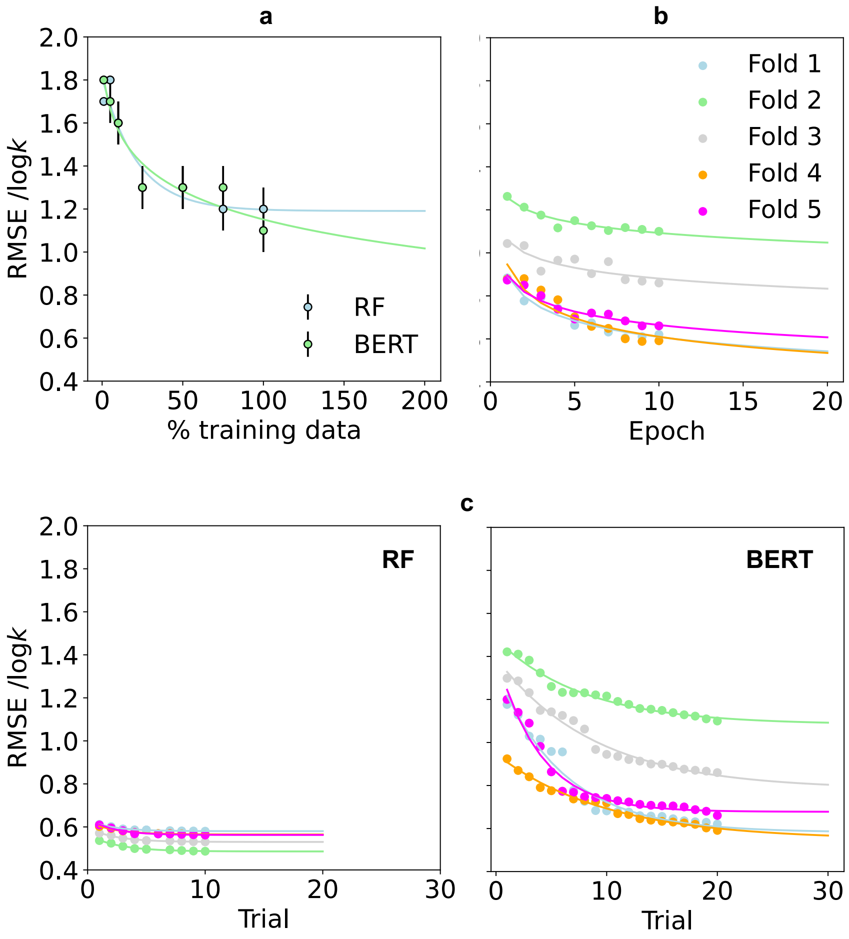
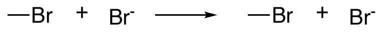
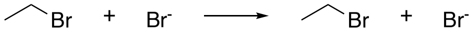
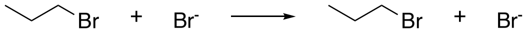
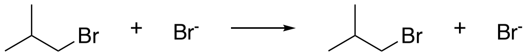
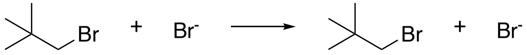
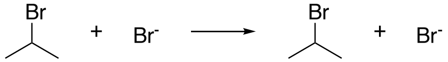
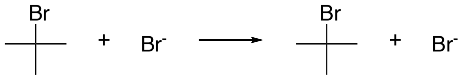
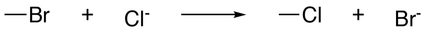
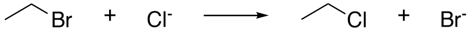
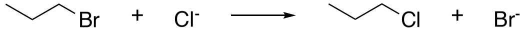
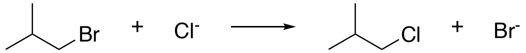
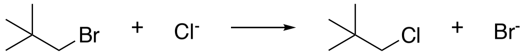
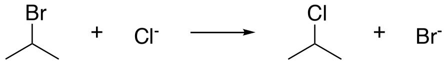
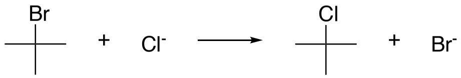
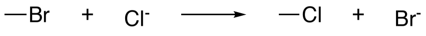
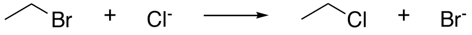
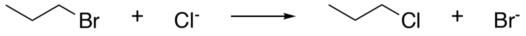


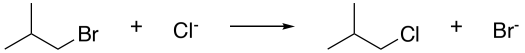
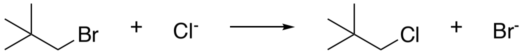
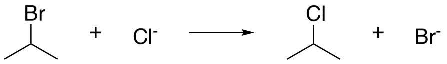
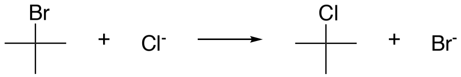
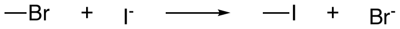
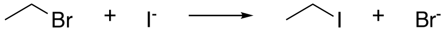
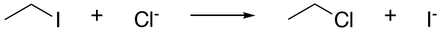
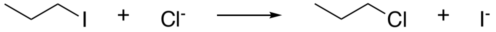
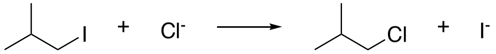
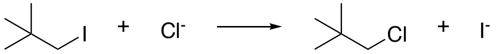
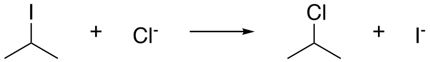
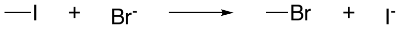
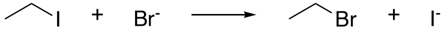
Figure 4: **Learning curves** depicting convergence in the RF and BERT models with respect to: training set size (**a**), epoch number (BERT only, **b**), and hyperparameter optimisation trial (**c**). **a** was evaluated using the total test data Test 1 + Test 2 and the mean RMSE taken over the five CV folds, with the standard error providing an uncertainty estimate. **b** and **c** were evaluated per fold using validation data.

3 DFT

DFT calculations were carried out using the ORCA suite of programs (version 4.2.1),[2] interfaced to autodE (version 1.1)[3] 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 2). Low MW reactions in single-component solvents were chosen owing to their reasonable computational cost and compatibility with the CPCM implicit solvent model.

Table 2: Dataset for DFT calculations.[a]

ID	Reaction	Solvent	Total MW / Da	$\Delta G^{\ddagger}_{\mathbf{exp}}$[b] / kcal mol^{-1}	$\Delta G^{\ddagger}_{\mathbf{DFT}}$ / kcal mol^{-1}	$\log k_{\mathbf{exp}}$	$\log k_{\mathbf{DFT}}$[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.[4] [b]Calculated using the Eyring equation (Equation 1 in the main text).

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 Equation 2 in the main text. 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, and a symmetry correction of $RT\ln(\sigma_r)$, where σ_r is the rotational symmetry number, were added to the calculated free energies. 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 Reactions **18**, **19**, and **22** in Table 2 using Otherm[5] 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.

PBE0 is widely recommended for obtaining accurate structures of organic molecules[6–8] and calculating vibrational frequencies.[9] Therefore, it is an appropriate LOT for the geometry optimisation and frequency calculations of $\text{S}_\text{N}2$ 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.[10] PBE0 has been shown to produce reliable geometries[7] and vibrational frequencies[9] 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,[7,11] was utilised alongside a triple-zeta basis set.

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

4 Accurate and inaccurate predictions

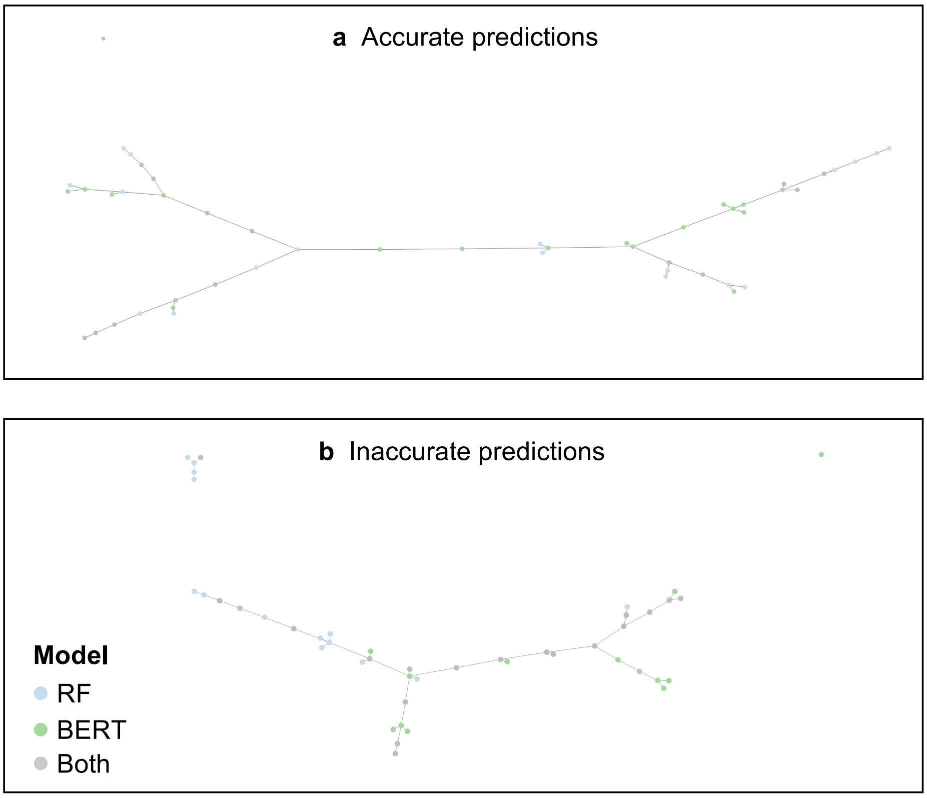
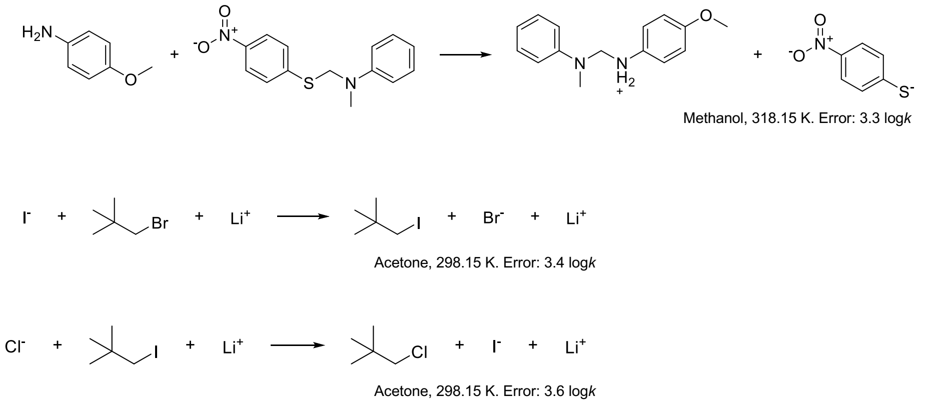


Figure 5: **Accurate (a) and inaccurate (b) predictions** of the RF (shown in blue) and BERT (shown in green) models, visualised as TMAPs. Here, accurate and inaccurate predictions are defined as those with the 25% smallest, and largest, error in $\log k$, respectively. Note that the peripheral data 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* in the main text).

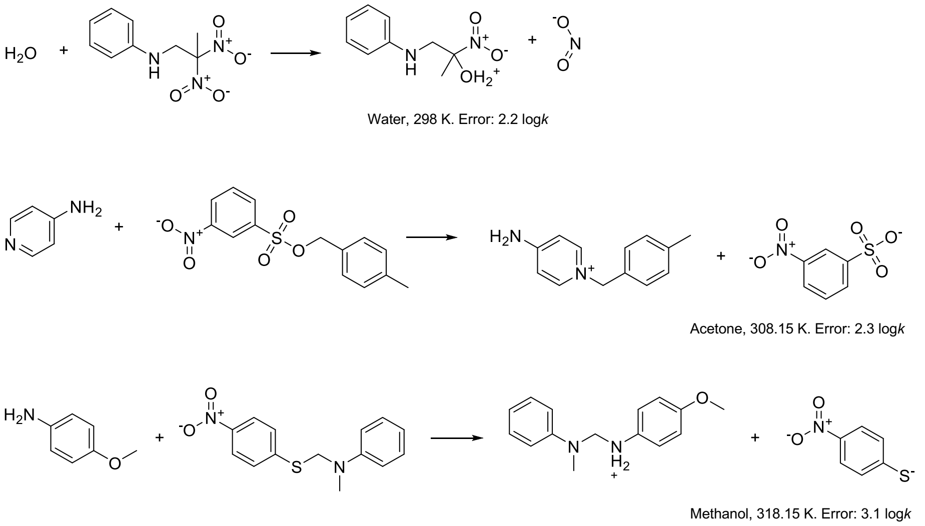


Methanol, 318.15 K. Error: 3.3 log*k*

Acetone, 298.15 K. Error: 3.4 log*k*

Acetone, 298.15 K. Error: 3.6 log*k*

Figure 6: **Error analysis for the RF model.** The three reactions with the largest prediction errors for the RF model.



Water, 298 K. Error: 2.2 log*k*

Acetone, 308.15 K. Error: 2.3 log*k*

Methanol, 318.15 K. Error: 3.1 log*k*

Figure 7: **Error analysis for the BERT model.** The three reactions with the largest prediction errors for the BERT model.

5 Interpretation

5.1 Attention maps and feature importance plots

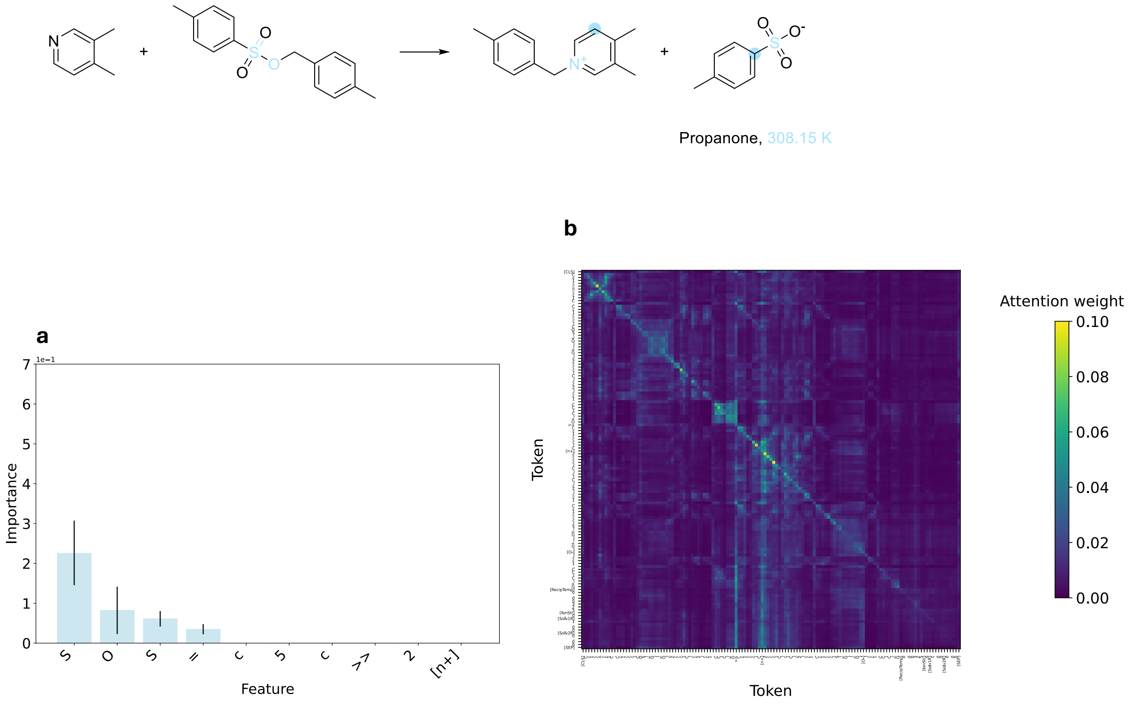


Figure 8: **Feature importance plot (a) and self-attention matrix (b) for one of BERT's accurate predictions (Example 1).** The important features in **a** are highlighted in blue on the reaction scheme. The self-attention matrix in **b** is from the final attention layer, and was averaged over the four attention heads.

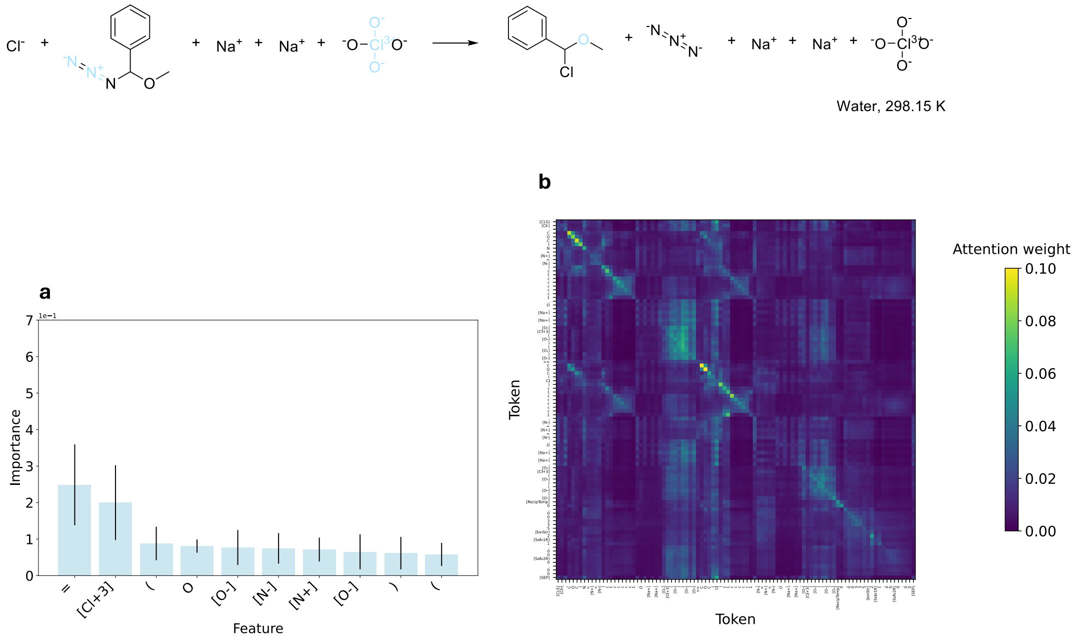


Figure 9: **Feature importance plot (a) and self-attention matrix (b) for one of BERT's accurate predictions (Example 2).** The important features in **a** are highlighted in blue on the reaction scheme. The self-attention matrix in **b** is from the final attention layer, and was averaged over the four attention heads.

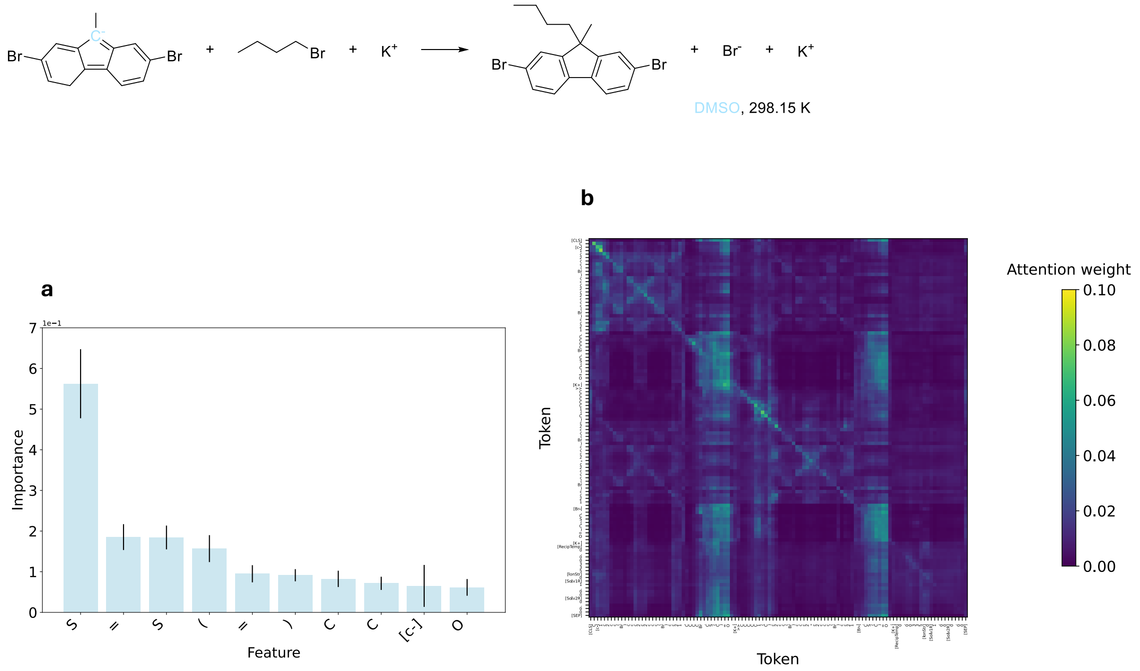


Figure 10: **Feature importance plot (a) and self-attention matrix (b) for one of BERT's inaccurate predictions.** The important features in **a** are highlighted in blue on the reaction scheme. The self-attention matrix in **b** is from the final attention layer, and was averaged over the four attention heads.

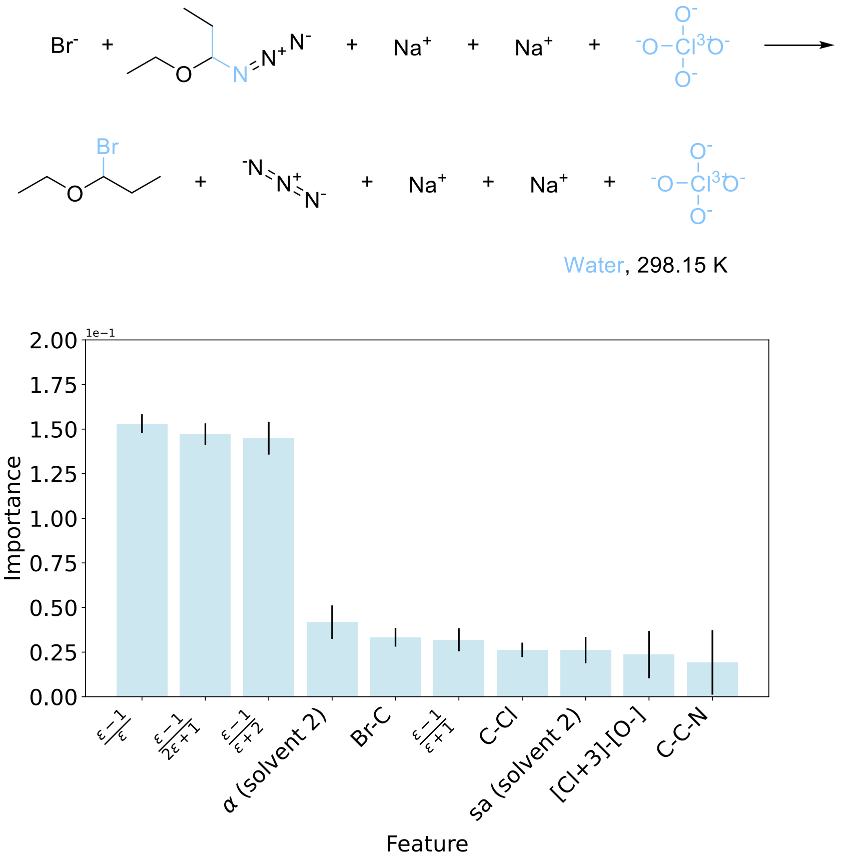


Figure 11: **Feature importance plot for one of RF's accurate predictions (Example 1).** The important features are highlighted in blue on the reaction scheme.

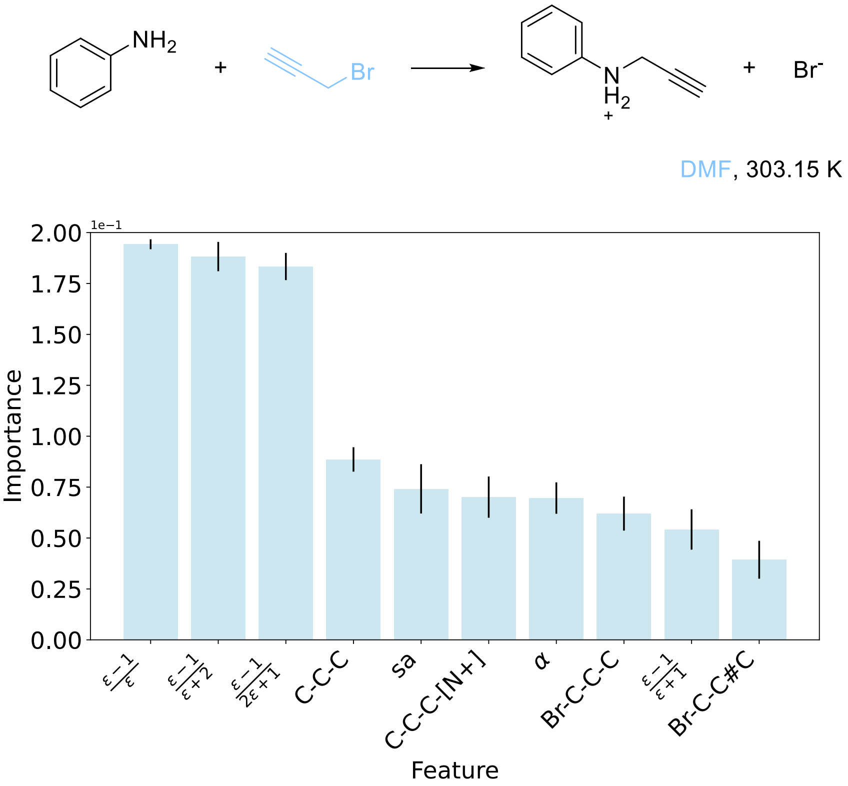


Figure 12: **Feature importance plot for one of RF's accurate predictions (Example 2).** The important features are highlighted in blue on the reaction scheme.

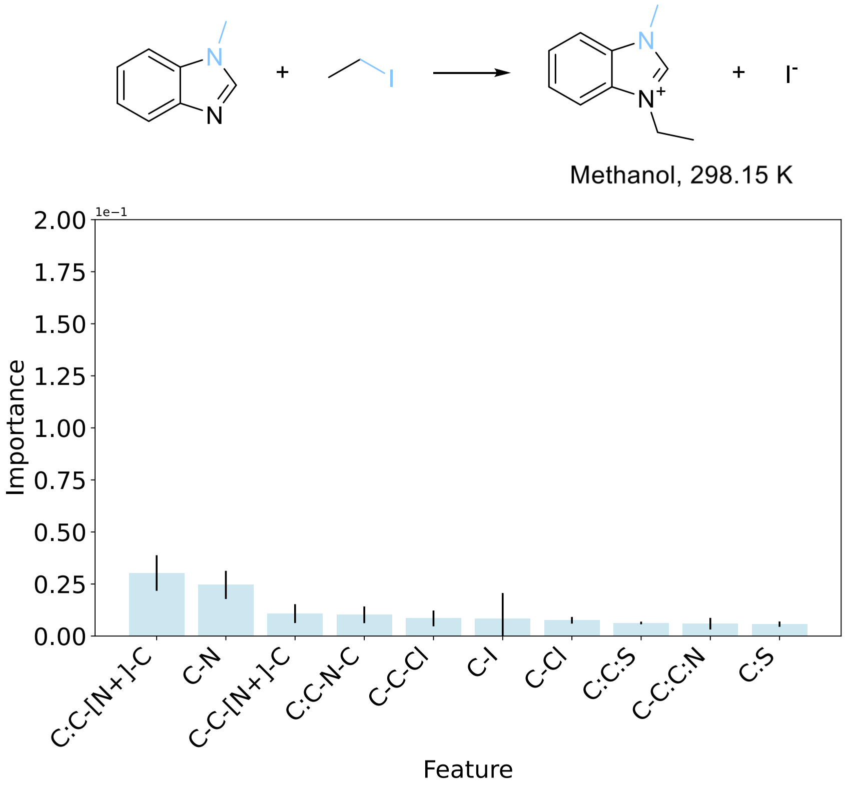


Figure 13: **Feature importance plot for one of RF's inaccurate predictions.** The important features are highlighted in blue on the reaction scheme.

5.2 Temperature effects

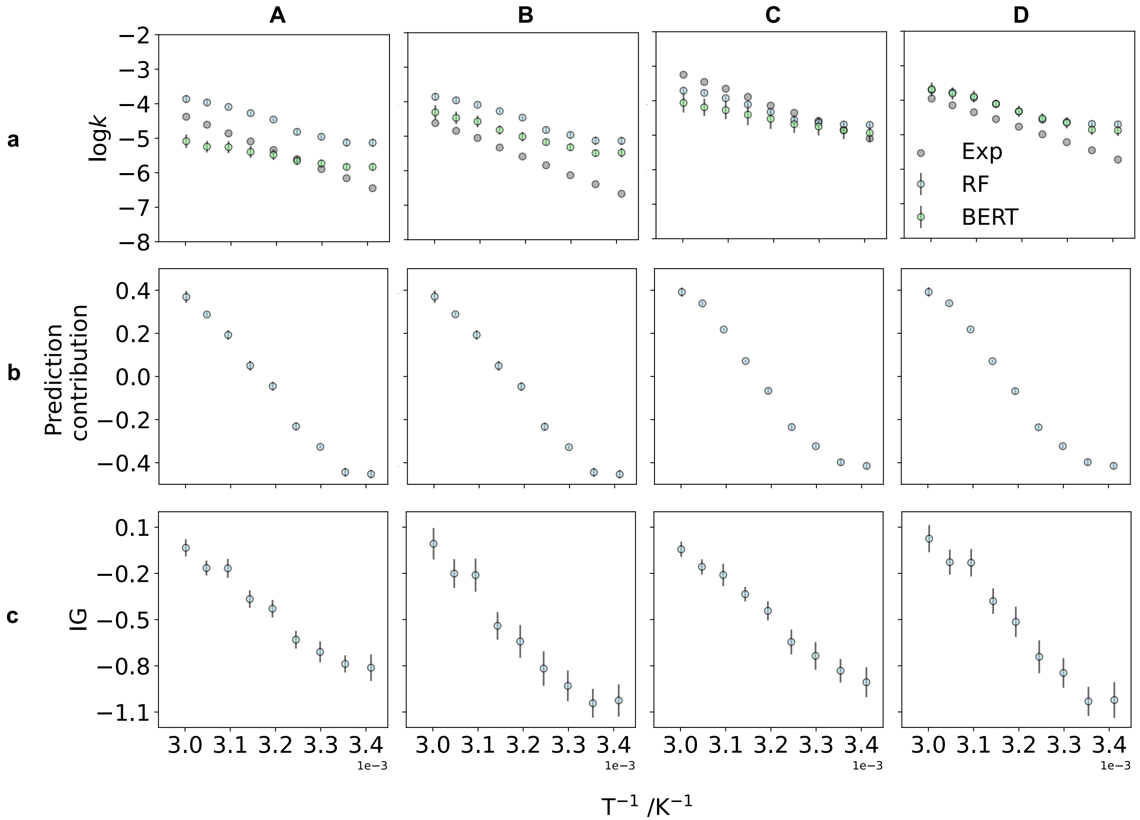


Figure 14: **Analysis of temperature effects. a.** Predicted (Pred) and experimental (Exp) $\log k$ values for each reaction in the test data observed at multiple temperatures. Predicted values are given for both the RF (shown in blue) and BERT (shown in green) models. **b and c.** Feature importances for each reaction in the test data observed at multiple temperatures for the RF (**b**) and BERT (**c**) models. Here, T^{-1} denotes the reciprocal temperature.

5.3 Accurate predictions

5.3.1 Steric effects

In the RF model, C-C and C-C-C fragments decreased $\log k_{\mathrm{pred}}$ in two reactions with unsubstituted centres (shown in Figure 15**a**), corresponding to those where C-C and C-C-C were present in other molecules in the reaction. This was attributed 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 of their product atom mappings) were omitted from the feature set. In this revised model (denoted $\mathrm{RF}_{\mathrm{centre}}$), the C-C and C-C-C features increased $\log k_{\mathrm{pred}}$ in 100% of analysed reactions with unsubstituted centres, including the reactions in Figure 15**a** (Figure 15**b**). 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 $\log k_{\mathrm{pred}}$ in three of RF's accurate predictions where it was absent from the reaction, corresponding to the reaction of 2-amino-1-methylbenzimidazole with methyl iodide in methanol at three temperatures between 298.15 and 308.15 K. Because each of these three examples are of the same transformation, the decrease in $\log k_{\mathrm{pred}}$ with the absence of C-C-C-C was attributed to a spurious correlation learned from similar training scaffolds. Using $\mathrm{RF}_{\mathrm{centre}}$, the C-C-C-C feature either increased $\log k_{\mathrm{pred}}$ or had low impact for 100% of reactions with an unsubstituted centre, including the three reactions of 2-amino-1-methylbenzimidazole with methyl iodide (Figure 15**b**). This affirms that spurious correlations may be prevented by removing noise created by away-from-centre scaffolds, without compromising model accuracy.

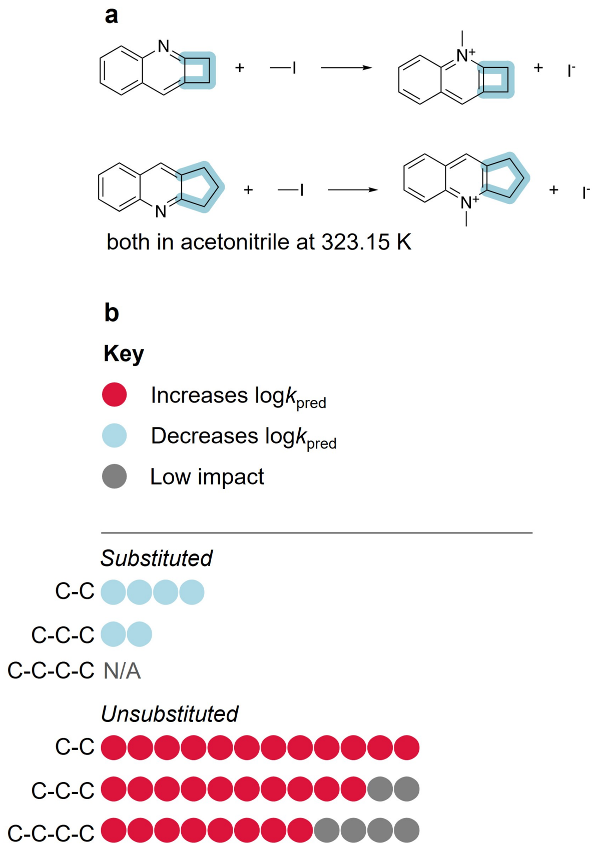


Figure 15: **a.** Reactions where C-C and C-C-C were present in the reaction away from the electrophilic centre (shown in blue). **b.** Signs (positive or negative) of feature importances for each of $\mathrm{RF}_{\mathrm{centre}}$'s accurate predictions where alkyl-substituted and unsubstituted electrophilic centres were high impact. Red and Blue circles denote predictions where the feature importance was positive (increased $\log k_{\mathrm{pred}}$), and negative (decreased $\log k_{\mathrm{pred}}$), 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.

5.3.2 Allylic effects

In the RF model, aromatic groups decreased $\log k_{\mathrm{pred}}$ in two reactions (Shown in Figure 16). This was attributed to their presence in the nucleophile, where they are expected to decrease $\log k$.

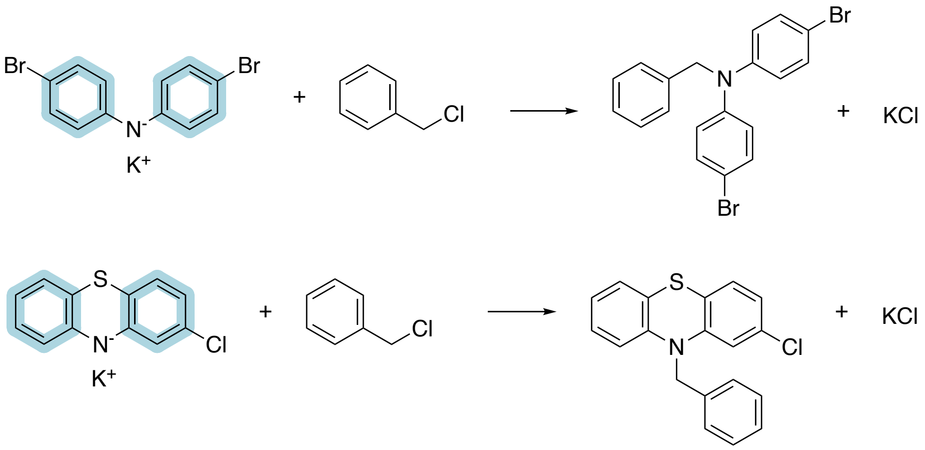


Figure 16: **Reactions where aromatic groups decreased** $\log k_{\mathbf{pred}}$ **in the RF model** attributed to their resonance with the nucleophilic atom (resonating groups shown in blue). Reactions in DMSO at 298.15 K.

5.3.3 Solvent effects

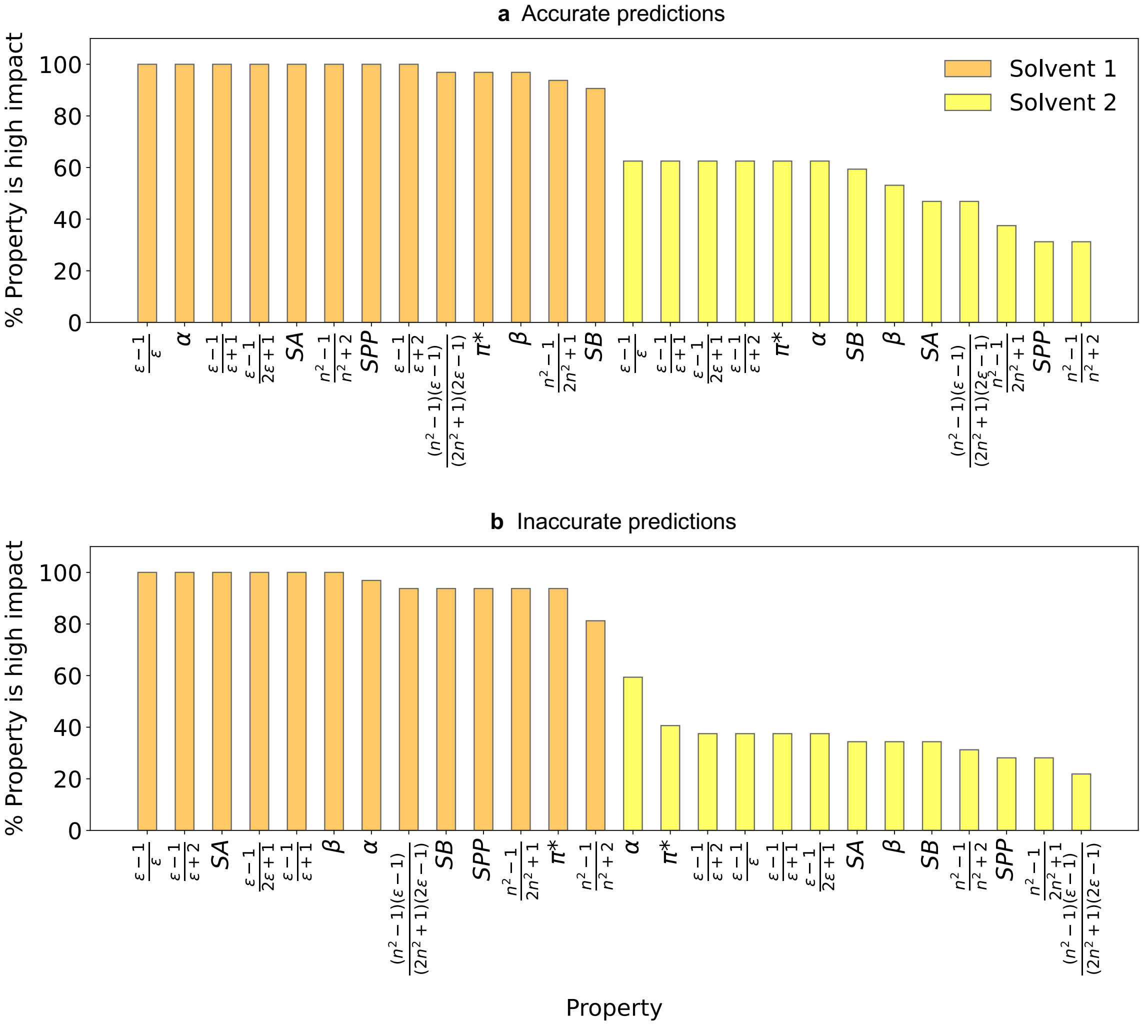


Figure 17: **The percentage of accurate (a) and inaccurate (b) predictions where each solvent property of the RF model was high impact.** Here, ε = dielectric constant, n = refractive index, SA, SB, and SPP = Catalán acidity,[12] basicity[13] and polarity/polarisability[14] constants, respectively, and α, β, and π^* = Kamlet-Taft acidity,[15] basicity,[16] and polarity/polarisability[17] constants, respectively.

Interpreting solvent effects in RF In the RF model, solvent was modelled by 13 properties representing either polarity or proticity. The Spearman's r value (r_S) between each of these properties and $\log k_\mathrm{exp}$ is plotted in Figure 18 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 $\log k_\mathrm{exp}$ varies greatly across the training data, with each solvent property being negatively correlated ($r_\mathrm{S} < 0$) with $\log k_\mathrm{exp}$ in some reactions, and positively correlated ($r_\mathrm{S} > 0$)

with $\log k_{\text{exp}}$ in others. Therefore, there is no consistent relationship between each solvent property and $\log k$ in the experimental data. Accordingly, solvent effects were not analysed for RF.

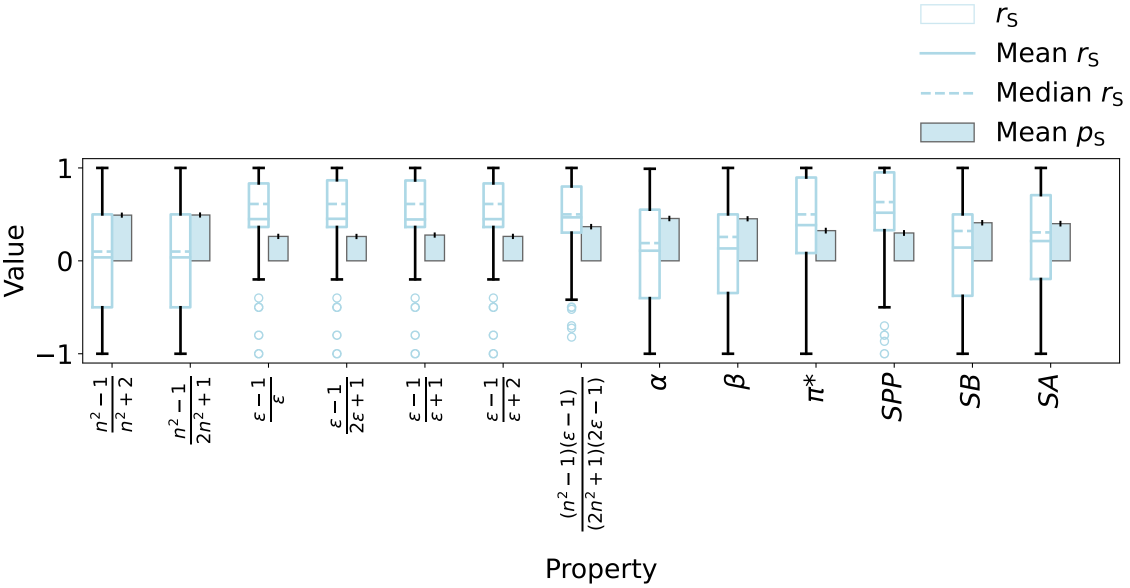


Figure 18: **Spearman's r value (r_S) between each solvent property in RF and $\log k_{\text{exp}}$** for each reaction in the training data that was observed in different solvents with a single solvent component and ionic strength 0. The mean Spearman's p value (mean p_S) is also provided for each solvent property. Here, ε = dielectric constant, n = refractive index, SA, SB, and SPP = Catalán acidity,[12] basicity[13] and polarity/polarisability[14] constants, respectively, and α, β, and π^* = Kamlet-Taft acidity,[15] basicity,[16] and polarity/polarisability[17] constants, respectively.

5.4 Inaccurate predictions

We examined the ability of RF and BERT models to identify key features influencing reactivity. This included evaluating the contributions from each reaction centre: nucleophilic (Nu) atom, leaving group (LG) atom, and electrophilic carbon (C) atoms, as well as temperature (represented as T^{-1}) and solvent polarity and proticity.

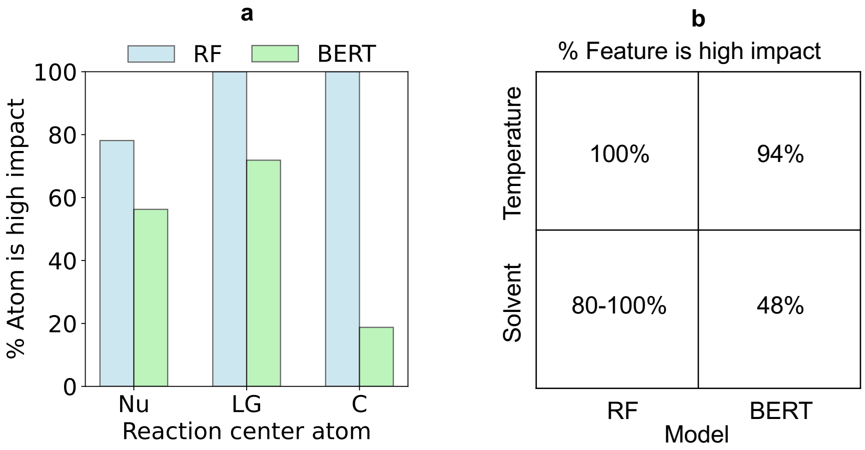


Figure 19: **High impact features of inaccurate predictions. a.** The percentage of inaccurate predictions where the nucleophilic (Nu), leaving group (LG), and electrophilic carbon (C) atoms were high impact features of the RF and BERT models. **b.** The percentage of inaccurate predictions where temperature 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 RF is provided in Figure S17**b**.

For both RF and BERT, results for LG and electrophilic C for inaccurate predictions reflected those for accurate predictions. For example, the LG atom was high impact in $> 70\%$ of inaccurate predictions for both RF and BERT, suggesting both models recognised the importance of this centre (Figure 19**a**). Meanwhile, the electrophilic C atom was identified as high impact in all of RF's inaccurate predictions, but only 19% of BERT's. This discrepancy is attributed to the featurization: in RF, the electrophilic C is represented by molecular fragments that include the environment around the electrophilic centre, while BERT represents it with a single token. Regarding the Nu atom, this was high impact in 78% of RF's inaccurate predictions but only 56% of BERT's. This discrepancy could not be attributed to any additional factor, suggesting BERT may have underestimated the nucleophile's importance in some reactions. On the other hand, temperature was a high impact feature for $> 90\%$ of inaccurate predictions for both models, emphasising their ability to recognise key physical features (Figure 19**b**). In the RF model, where solvent is represented by 13 properties,[1] each of these properties were high impact in $> 80\%$ of inaccurate predictions, in line with ref.[18] In BERT, where solvent is represented by SMILES strings, solvent was high impact in 48% of inaccurate predictions. However, for reactions where solvent had low impact, the inaccurate predictions were more likely due to the novelty of the LG (these reactions have sulphonate LGs, while similar training scaffolds had iodide LGs), rather than the significance of the solvent.

To assess whether RF and BERT effectively learned key structural and physical effects, we evaluate the feature importance of high impact features, including LG, steric, allylic, temperature, and solvent effects, on either increasing (positive sign) or decreasing (negative sign) $\log k_{\text{pred}}$.

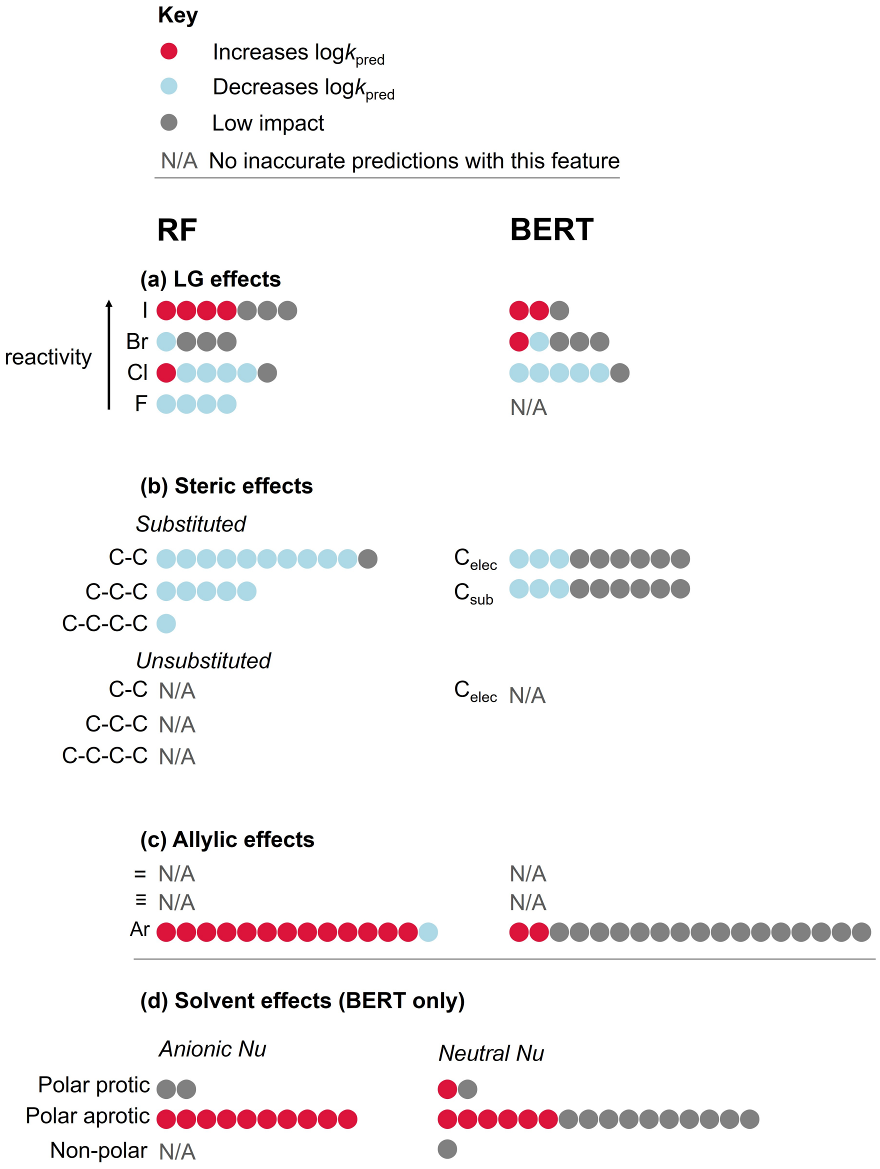


Figure 20: Feature importances for RF and BERT models of inaccurate predictions, defined as the lower quartile of the test data. Red and blue circles denote predictions where the feature increased (positive feature importance) or decreased (negative feature importance) $\log k_{\text{pred}}$. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes that no inaccurate predictions contained this feature. C_{elec} and C_{sub} in **b** correspond to electrophilic and substituting carbons, respectively.

5.4.1 LG effects

The distribution of halide LGs in the test set was 53 × I, 18 × Br, 16 × Cl, and 4 × F; for inaccurate predictions this distribution is: 7 × I, 4 × Br, 6 × Cl, and 4 × F for RF and 3 × I, 5 × Br, 6 × Cl, and 0 × F for BERT (Figure 20**a**).

Overall, both models show a positive correlation between halide size and reactivity (rates: Cl < Br < I), in agreement with the results for accurate predictions. In the RF model, where LG atoms were represented by C-I, C-Br, C-Cl and C-F fragments, I increased $\log k_{\mathrm{pred}}$ in all examples where it was high impact (4 reactions), Br was low impact 75% of the time (3 reactions), Cl decreased $\log k_{\mathrm{pred}}$ 80% of the time it was high impact (4 reactions), and F decreased $\log k_{\mathrm{pred}}$ in all examples. In the BERT model, I increased $\log k_{\mathrm{pred}}$ in both reactions where it was high impact, Br was low impact 60% of the time (3 reactions), and Cl decreased $\log k_{\mathrm{pred}}$ in all cases where it was high impact (5 reactions).

5.4.2 Steric effects

The distribution of alkyl-substituted and unsubstituted centres in the test set was 41 x alkyl-substituted and 24 x unsubstituted; for inaccurate predictions, this distribution is 11 x alkyl-substituted and 0 x unsubstituted for RF and 9 x alkyl-substituted and 0 x unsubstituted for BERT (Figure 20**b**). In RF, alkyl-substituted centres are represented by C-C, C-C-C, and C-C-C-C fragments (where, e.g., C-C-C-C could represent one propyl substitution, or one methyl and one ethyl substitution). Similar fragments located away from electrophilic centres served as a control. In BERT, alkyl-substituted centres are represented by the electrophilic carbon centre and its substituting C atoms.

Our analysis shows that both models recognised that steric hindrance decreases $\mathrm{S_N2}$ reactivity in agreement with the results for accurate predictions. In both models, substituted centres decreased $\log k_{\mathrm{pred}}$ in all examples where they were high impact. Reactions with unsubstituted centres were not predicted inaccurately by either model.

5.4.3 Allylic effects

The distribution of allyl-substituted centres in the test set was 2 × alkene, 4 × alkyne, and 42 × aromatic; for inaccurate predictions this distribution was 0 × alkene, 0 × alkyne, and 14 × aromatic for RF, and 0 × alkene, 0 × alkyne, and 18 × aromatic for BERT (Figure 20**c**). Alkene, alkyne, and aromatic bonds at the electrophilic centre were represented by C-C=C, C-C≡C, and C-C:C fragments in RF (where ':' is an aromatic bond), and =, ≡, and c tokens in BERT (where c is an aromatic carbon bonded to the centre).

Typically, results for inaccurate predictions reflected those for accurate predictions. In the RF model, aromatic groups increased $\log k_{\mathrm{pred}}$ in 93% of examples, suggesting RF recognised that allylic groups increase $\mathrm{S_N2}$ reactivity. Conversely, aromatic groups were low impact in 89% of BERT's inaccurate predictions, suggesting BERT had not recognised their importance for $\mathrm{S_N2}$ reactivity. Importantly, reactions with alkene or alkyne bonds to the electrophilic centre were not predicted inaccurately by either model.

5.4.4 Solvent effects

Solvent effects were modelled using 13 properties representing polarity and proticity in RF and solvent SMILES in BERT. No correlation between the 13 solvent properties and $\log k$ was observed in the experimental data, so solvent effects are not analysed for RF (Figure S18). In BERT, solvent effects were evaluated by analysing the contribution of polar ($\varepsilon > 15$) protic and aprotic 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 20**d**). Only one inaccurate prediction with non-polar solvent ($\varepsilon < 15$) was obtained, and the solvent was low impact.

BERT consistently predicted that polar solvents increase $\log k$ with neutral nucleophiles (one of two reactions for protic, and 6 reactions for aprotic). For anionic nucleophiles, polar aprotic solvents increased $\log k_{\text{pred}}$ in all examples, however polar protic solvents had low impact.

5.5 Product 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 21 for the RF and BERT models.

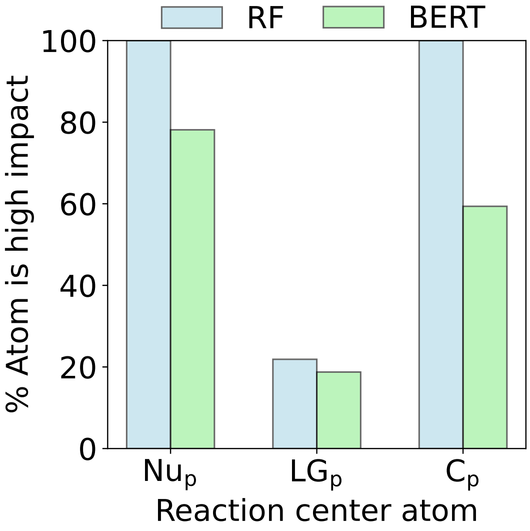


Figure 21: **High impact features of accurate predictions (product mappings).** The percentage of accurate predictions where product mappings to the nucleophilic (Nu_p), leaving group (LG_p), and electrophilic carbon (C_p) atoms were high impact features of the RF and BERT models.

For both RF and BERT, the results for Nu_p and C_p reflected those for Nu and C. For example, Nu_p was high impact in $\geq 75\%$ of accurate predictions for both models. Meanwhile C_p was high impact in all of RF's accurate predictions and $< 60\%$ of BERT's. This discrepancy is attributed to the featurization: in RF, the electrophilic C is represented by molecular fragments that include the environment around the electrophilic centre, while BERT represents it with a single token. Regarding LG_p, this was less important for reactivity prediction than LG, being high impact in $< 25\%$ of accurate predictions for both models. For RF, this was attributed to LG_p typically being monoatomic and therefore not a feature of RF (which has a minimum fragment length of 2 atoms). However, this does not explain LG_p's low importance in the BERT model. Importantly, LG (reactants) was high impact in the BERT model (see *LG effects* in the main text), indicating that BERT effectively recognised the importance of this centre for predicting S_N2 reactivity.

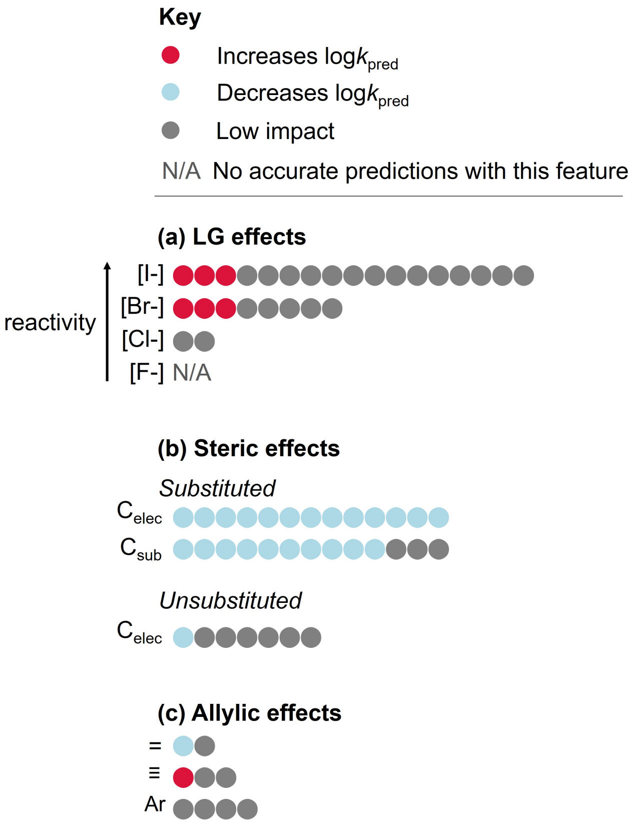


Figure 22: Feature importances (products) for RF and BERT models of accurate predictions, defined as the upper quartile of the test data. Red and blue circles denote predictions where the feature increased (positive feature importance) or decreased (negative feature importance) $\log k_{\mathrm{pred}}$. Grey circles denote predictions where the feature was low impact, or the feature importance was zero within error. N/A denotes that no accurate predictions contained this feature. C_{elec} and C_{sub} in **b** correspond to electrophilic and substituting carbons, respectively.

5.6 LG effects

In BERT, product mappings to halide LGs were represented by [I^-], [Br^-], [Cl^-], and [F^-] tokens. The importances of these mappings are shown in Figure 22**a** for BERT's accurate predictions. Product mappings to halide LGs (which are monoatomic) were not a feature of RF (which has a minimum fragment length of 2 atoms).

In BERT, product mappings to halide LGs were less important for reactivity prediction than in the reactants, with [I^-], [Br^-], and [Cl^-] being low impact 82%, 63%, and 100% of the time.

5.7 Steric effects

In BERT, alkyl-substituted centres are represented by the electrophilic carbon centre and its substituting C atoms. The importance of these mappings is shown in Figure 22**b** for BERT's accurate predictions. In RF, alkyl-substituted centres are represented by C-C, C-C-C, and C-C-C-C fragments with no assignment to reactants or products, so they are not analysed here.

In BERT, product mappings to substituted centres decreased $\log k_{\text{pred}}$ in all examples, in agreement with the results for reactants. Product mappings to unsubstituted centres were less important for reactivity prediction than in the reactants, being low impact 86% of the time.

5.8 Alyllic effects

In BERT, product mappings to 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). The importance of these mappings is shown in Figure 22**c** for BERT's accurate predictions. In RF, product mappings to alkene, alkyne, and aromatic bonds at the electrophilic centre were represented by C-C=C, C-C≡C, and C-C:C fragments in RF (where ':' is an aromatic bond) with no assignment to reactants or products, so are not analysed here.

In BERT, product mappings to alkene, alkyne, and aromatic bonds were less important for reactivity prediction than in the reactants, with =, ≡, and c being low impact in 50%, 67%, and 100% of accurate predictions, respectively.

Bibliography

(1) 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.

(2) Neese, F. *WIREs: Comp. Mol. Sci.* **2012**, *2*, 73–78.

(3) Young, T.; Silcock, J.; Sterling, A.; Duarte, F. *Angew. Chem.* **2021**, *133*, 4312–4320.

(4) DeTar, D.; McMullen, D.; Luthra, N. *J Am Chem Soc* **1978**, *100:8*, 2484–2493.

(5) Young, T. duartegroup/otherm: Major Symmetry Improvements, 2020.

(6) Brémond, É.; Savarese, M.; Su, N.; Pérez-Jiménez, Á.; Xu, X.; Sancho-García, J.; Adamo, C. *J. Chem. Theory Comput.* **2016**, *12*, 459–465.

(7) Bursch, M.; Mewes, J.-M.; Hansen, A.; Grimme, S. *Angew. Chem. Int. Ed.* **2022**, *61*, e202205735.

(8) Orca Input Library - Geometry Optimizations. URL: `https://sites.google.com/site/orcainputlibrary/geometry-optimizations` (last accessed 11/6/2024).

(9) Tantirungrotechai, Y.; Phanasant, K.; Roddecha, S.; Surawatanawong, P.; Sutthikhum, V.; Limtrakul, J. *J. Mol. Struct.: THEOCHEM* **2006**, *760*, 189–192.

(10) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comp. Chem.* **2011**, *32*, 1456–1465.

(11) Orca Input Library - Coupled Cluster. URL: `https://sites.google.com/site/orcainputlibrary/coupled-cluster` (last accessed 11/6/2024).

(12) Catalán, J.; Díaz, C. *Liebigs Ann.* **2006**, *1997*, 1941–1949.

(13) Catalán, J.; Díaz, C.; López, V.; Pérez, P.; De Paz, J.-L.; Rodríguez, J.-G. *Liebigs Ann.* **2006**, *1996*, 1785–1794.

(14) Catalán, J.; López, V.; Pérez, P.; Martin-Villamil, R.; Rodríguez, J.-G. *Liebigs Ann.* **2006**, *1995*, 241–252.

(15) Taft, R.; Kamlet, M. *J. Am. Chem. Soc.* **1976**, *98:10*, 2886–2894.

(16) Kamlet, M.; Taft, R. *J. Am. Chem. Soc.* **1976**, *98:2*, 377–383.

(17) Kamlet, M.; Abboud, J.; Taft, R. *J. Am. Chem. Soc.* **1977**, *19:18*, 6027–6038.

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