Article:
Boobier, S, Liu, Y, Sharma, K et al. (4 more authors) (2021) Predicting Solvent-Dependent Nucleophilicity Parameter with a Causal Structure Property Relationship. Journal of Chemical Information and Modeling. acs.jcim.1c00610. ISSN 1549-9596

https://doi.org/10.1021/acs.jcim.1c00610

Predicting Solvent-Dependent Nucleophilicity Parameter with Causal Structure Property Relationship

*Samuel Boobier,[‡] Yufeng Liu,[‡] Krishna Sharma,[‡] David R. J. Hose,[†] A. John Blacker,[‡] Nikil Kapur,[¶] and Bao N. Nguyen[‡]**

[‡]Institute of Process Research & Development, School of Chemistry, University of Leeds, Leeds, LS2 9JT, United Kingdom. Email: b.nguyen@leeds.ac.uk

[¶]School of Mechanical Engineering, University of Leeds, Leeds, LS2 9JT, United Kingdom

[†]Chemical Development, Pharmaceutical Technology and Development, Operations, AstraZeneca, Macclesfield SK10 2NA, UK

ABSTRACT: Solvent dependent reactivity is a key aspect of synthetic science, which controls reaction selectivity. The contemporary focus on new, sustainable solvents highlights a need for reactivity predictions in different solvents. We herein report excellent machine learning prediction of the nucleophilicity parameter *N* in the four most common solvents for nucleophiles in the Mayr's reactivity parameter database (R^2 = 0.93 and 81.6% of predictions within ±2.0 of the experimental values with Extra Trees algorithm). A Causal Structure Property Relationship

(CSPR) approach was utilized, with focus on the physicochemical relationships between the descriptors and the predicted parameters, and on rational improvements of the prediction models. The nucleophiles were represented with a series of electronic and steric descriptors and the solvents were represented with PCA descriptors based on the ACS Solvent Tool. The models indicated that steric factors do not contribute significantly, due to bias in the experimental database. The most important descriptors are solvent-dependent HOMO energy and Hirshfeld charge of the nucleophilic atom. Replacing DFT descriptors with Parameterization Method 6 (PM6) descriptors for the nucleophiles led to an 8.7-fold decrease in computational time, and approximately 10% decrease in the percentage of predictions within ±2.0 and ±1.0 of the experimental values.

1. INTRODUCTION

Predicting reaction selectivity is one of the key cornerstones which underpins synthetic science. The majority of reactions in synthetic processes are under kinetic control, and their selectivity can be predicted through assessment of competitive reaction rates. While rationalizing and predicting stereoselectivity have been consistently demonstrated using molecular modelling of transition states in asymmetric catalysis,[1–3] prediction of chemoselectivity between dissimilar reactions of different kinetic orders remains difficult.[4] This is further hindered by the complexity of reaction conditions and solvent dependence of reaction outcomes.[5–7] On the other hand, selecting the correct solvent and reaction conditions can be a powerful tool in manipulating and controlling reaction selectivity, as recently demonstrated by Vigo and co-workers in a synthesis of Raltegravir (Figure 1).[8]

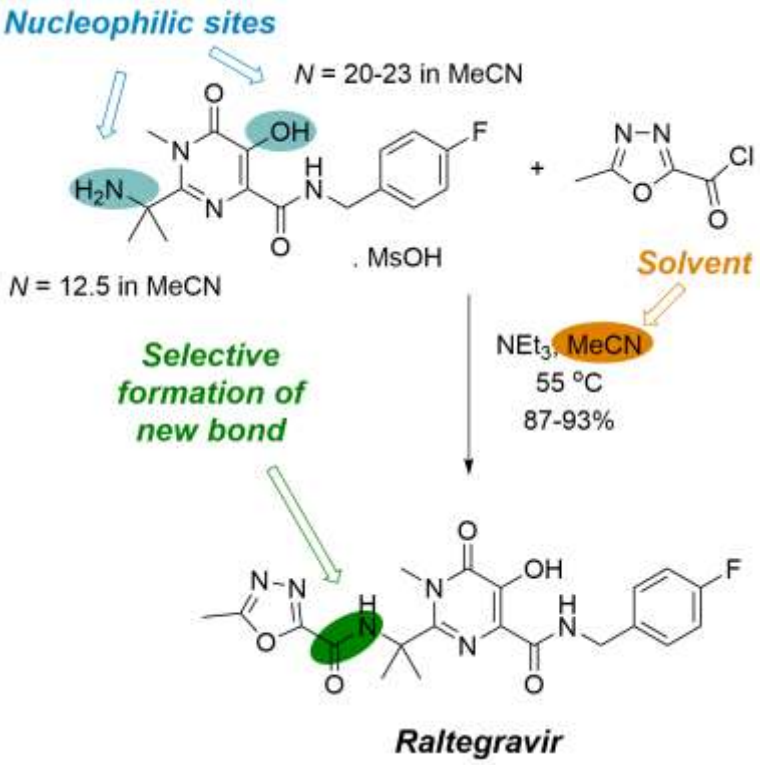

Figure 1. Exploitation of solvent dependent chemoselectivity in the synthesis of Raltegravir.

While significant advances have been made in synthetic route prediction using machine learning and cheminformatics,[9–12] selectivity prediction has not received the same level of attention. Only two recent examples can be found in the literature which report the use of machine learning to predict Mayr's nucleophilicity parameter (*N*) and electrophilicity parameter (*E*).[13,14] These solvent-dependent parameters and s_N (reaction specific constant), when used in Mayr-Patz equation (eq. 1),[15] enable calculation of rate constants between any given pair of nucleophile and electrophile and consequently the selectivity between competing reactions.

$$log_{10}\left(k_{20\,^{o}C}\right) = s_N(N + E) \qquad \text{(eq. 1)}$$

Earlier prediction of Mayr-Patz reactivity parameters were performed using Density Functional Theory (DFT) calculations of the transition states, and were limited to small sub-classes of

nucleophiles.[16,1] Kang and co-workers developed a new model based on neural networks and DFT calculations to successfully predict both nucleophilicity and electrophilicity parameters with reasonable accuracy (R^2 = 0.92, RMSE = 2.28 for *E* and R^2 = 0.81, RMSE = 3.23 for *N*, with no dataset provided for validation and benchmarking).[14] Orlandi et. al. achieved even more impressive accuracy in predicting nucleophilicity parameters (R^2 = 0.94, RMSE = 1.41) employing DFT calculations of molecular properties and a much simpler Multivariate Linear Regression model (MLR).[13] This was accomplished with a small subset of 341 nucleophiles in 7 different solvents from the Mayr reactivity database.[18] However, these data points did not include many classes of nucleophiles, e.g. hydrogen donors, phosphorus nucleophiles, boron-, silicon- and metal-based reagents, and their accompanying complex mechanisms and transition states (Figure 2a).

We report here our development of general prediction models for Mayr's solvent-dependent nucleophilicity parameter which covers all current classes of nucleophiles in the Mayr's reactivity parameter database. We employed our recently formalized Causal Structure Property Relationship (CSPR) approach,[19] which focuses on interpretable machine learning models based on physicochemically guided translation of chemical properties into numerical machine learning inputs. Solvent dependence is implemented through the use of principle components developed for the ACS Solvent Selection Tool by AstraZeneca.[20,21] Thus, a highly accurate prediction model for nucleophilicity parameter *N* in four common solvents was obtained (R^2 = 0.94) for the widest range of nucleophiles in the Mayr's reactivity database. This model was rationally improved through understanding descriptor importance, resulting in 81.6% of the predicted *N* parameters within the experimental errors of the reported values.

2. MATERIALS AND METHODS

2.1. Data curation

Data (1227 nucleophiles in 26 solvents and solvent mixtures) were mined from the online Mayr's reactivity database using Beautiful Soup 4 package in Python 3.[22] Simplified Molecular-Input Line-Entry (SMILES) strings were obtained from Chemical Identifier Resolver (CIRpy).[23] SMILES strings were screened to remove intermediates, salts and complexes. Metal cations, such as lithium or magnesium, were removed. Data points with nucleophilicity parameter measured in dichloromethane (DCM), acetonitrile (MeCN), dimethylsulfoxide (DMSO) and water (the four most common solvents in Mayr's database) were kept. After curation, a dataset of 904 molecules remained with *N* values from -8.80 to 30.82.

2.2. Calculation of descriptors

Solvent Principal Component Analysis (PCA) descriptors were derived from the standard set available from ACS Green Chemistry Institute Solvent Selection Tool.[20,21] The first 5 principal components (***sol_PCA1-5***) were used to represent the solvent.

Nucleophile descriptors were calculated from 3D structures, initially generated in CIRpy,[24] and then optimized in gas and solution phase with Gaussian 09,[25] using M06-2x method and Def2svp basis set with DFT-D3 dispersion correction.[26] Steric descriptors (***N_TCA*** and ***N_BAD***) were calculated by manually placing a Li probe at the approximate angle based on the relevant mechanism and typical Li-X distance from the nucleophilic atom derived from the Cambridge Structure Database (CSD).[27] Fukui descriptors were calculated with Gaussian 09 and multiwfn,[28] using the most nucleophilic atom identified from the Li structures.

2.3 Model building and assessment

Machine learning was performed in Python 3 using the *scikit-learn* package,[29] with the exception of Gaussian Process Regression, which was run using *GPy*.[30] Prior to building models, descriptors were scaled to be between 0 and 1. Models were built with 8 machine learning methods and tested

using a fixed training:test split (90:10) and 10-fold cross validation. For the fixed training:test split, models were built 10 times to get average predictions and metrics (with error bars of 1 standard deviation (SD)). For 10-fold cross validation, this was run 10 times, with average predictions (with error bars of 1 SD). Metrics came from the average of the 10 folds within 10-fold cross validation across a single run. Default parameters were used in most cases, and then subsequently optimized. ANN architecture (*n_estimators* in *scikit-learn*) was explored for the average of 100 runs with a single hidden layer of between 2 and 5000 nodes. SVM parameters were determined via a grid search using 10-fold cross validation of the training set. The number of trees in ExtraTrees (ET) was optimized for the test sets between 1 and 5000. The radial basis function kernel was used for GP and error bars were obtained to 1 SD by obtaining the upper and lower limit which encompassed 68% of the prediction distribution. These parameters are summarized in Table S4 in the Supporting Information.

Table 1. Dataset size, standard deviation (of nucleophilic parameter) and train and test set sizes

Dataset	Dataset Size	SD	Train Set Size	Test Set Size
Full_set	904	7.57	808	96
Solution_set	896	7.52	802	94

3. RESULTS AND DISCUSSION

3.1 Data curation and analysis

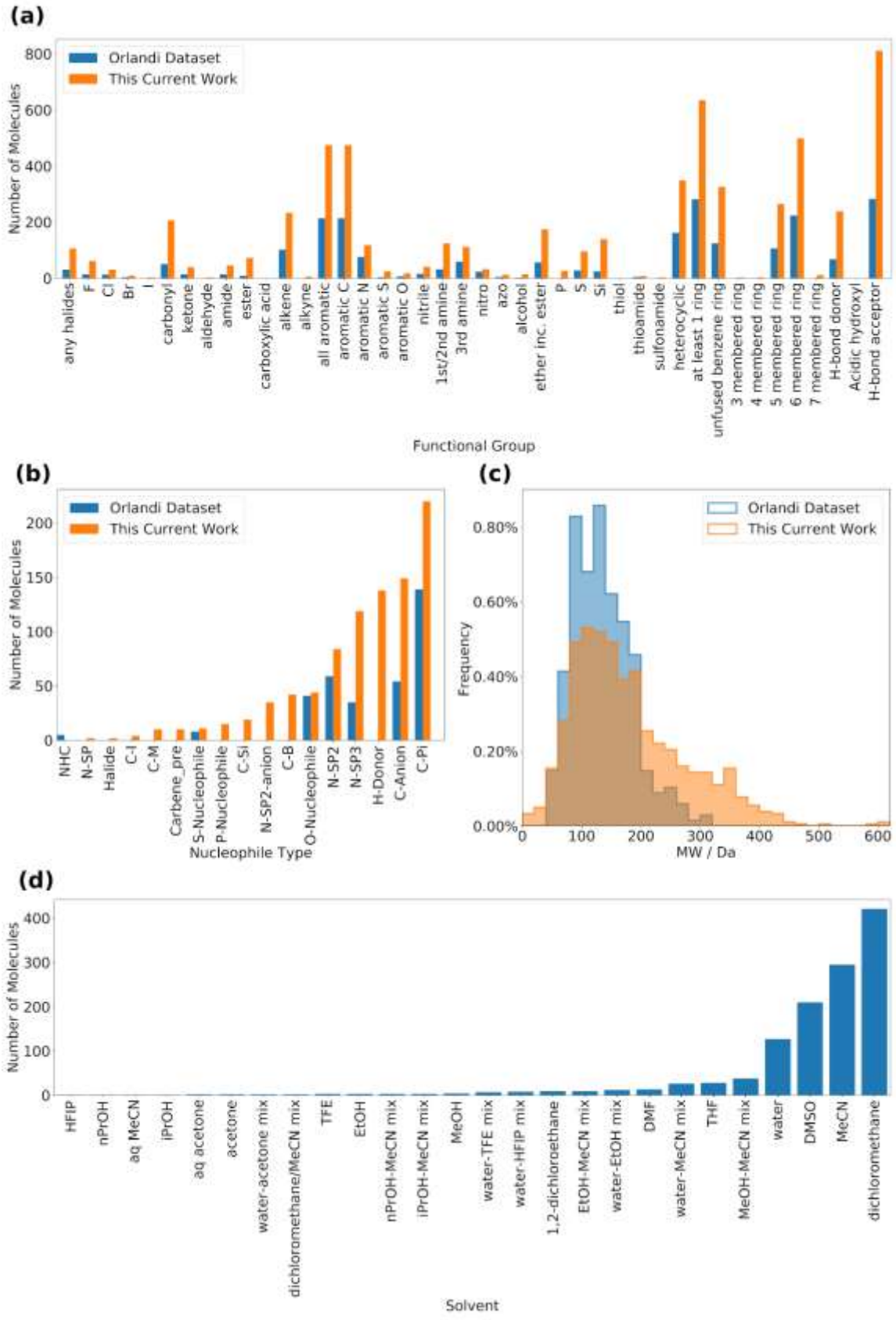
(a)
Orlandi Dataset
This Current Work
Number of Molecules
Functional Group
any halides
F
Cl
Br
I
carbonyl
ketone
aldehyde
amide
ester
carboxylic acid
alkene
alkyne
all aromatic
aromatic C
aromatic N
aromatic S
aromatic O
nitrile
1st/2nd amine
3rd amine
nitro
azo
alcohol
ether inc. ester
P
S
Si
thiol
thioamide
sulfonamide
heterocyclic
at least 1 ring
unfused benzene ring
3 membered ring
4 membered ring
5 membered ring
6 membered ring
7 membered ring
H-bond donor
Acidic hydroxyl
H-bond acceptor
(b)
Orlandi Dataset
This Current Work
Number of Molecules
Nucleophile Type
NHC
N-SP
Halide
C-I
C-M
Carbene_pre
S-Nucleophile
P-Nucleophile
C-Si
N-SP2-anion
C-B
O-Nucleophile
N-SP2
N-SP3
H-Donor
C-Anion
C-Pi
(c)
Orlandi Dataset
This Current Work
Frequency
MW / Da
(d)
Number of Molecules
Solvent
HFIP
nPrOH
aq MeCN
iPrOH
aq acetone
acetone
water-acetone mix
dichloromethane/MeCN mix
TFE
EtOH
nPrOH-MeCN mix
iPrOH-MeCN mix
MeOH
water-TFE mix
water-HFIP mix
1,2-dichloroethane
EtOH-MeCN mix
water-EtOH mix
DMF
water-MeCN mix
THF
MeOH-MeCN mix
water
DMSO
MeCN
dichloromethane

Figure 2. (a) classes of nucleophiles and number of data points in this study and ref. 13; (b) Functional groups distribution in the datasets used in this study and ref. 13; (c) MW distribution in the dataset used by Orlandi and in this study; (d) solvent distribution in the Mayr's database.

The total number of available data points for nucleophilicity in the Mayr's reactivity parameter database is relatively small in machine learning context. Thus, data curation was carried out with the intention of preserving as much usable data as possible. In contrast with the work by Orlandi and co-workers,[13] no class of nucleophile was excluded (Figure 2b and Figure 3). Solvent analysis of the database showed 26 solvents and solvents mixtures used in their measurements. However, most of the data were collected in four solvents: DCM, MeCN, DMSO and water. These make up 86% of the available data. Although 7 solvents were included by Orlandi, the 3 additional solvents added up to 20 data points out of 341 in that study. The low availability of data in other solvents will render validation of predictions in those solvent difficult, particularly with complex non-linear machine learning models. Therefore, only data from these four solvents were adopted in our dataset, which include 904 data points, 808 in the training set and 96 in the test set.

Molecular weight (MW) distribution and functional group analysis of the curated dataset showed that the nucleophiles occupy a reasonably wide chemical space, with many common functional groups present, albeit without highly electrophilic function groups for obvious reasons. Due to the inclusion of third-row element nucleophiles and H-donors, this new dataset has a significant increase in the number of nucleophiles with MW>200 compared to that of Orlandi (Figure 2c). The inclusion of these new classes of nucleophiles means more complex reaction mechanisms will need to be included in the model,[31,32] with appropriate descriptors to describe them.

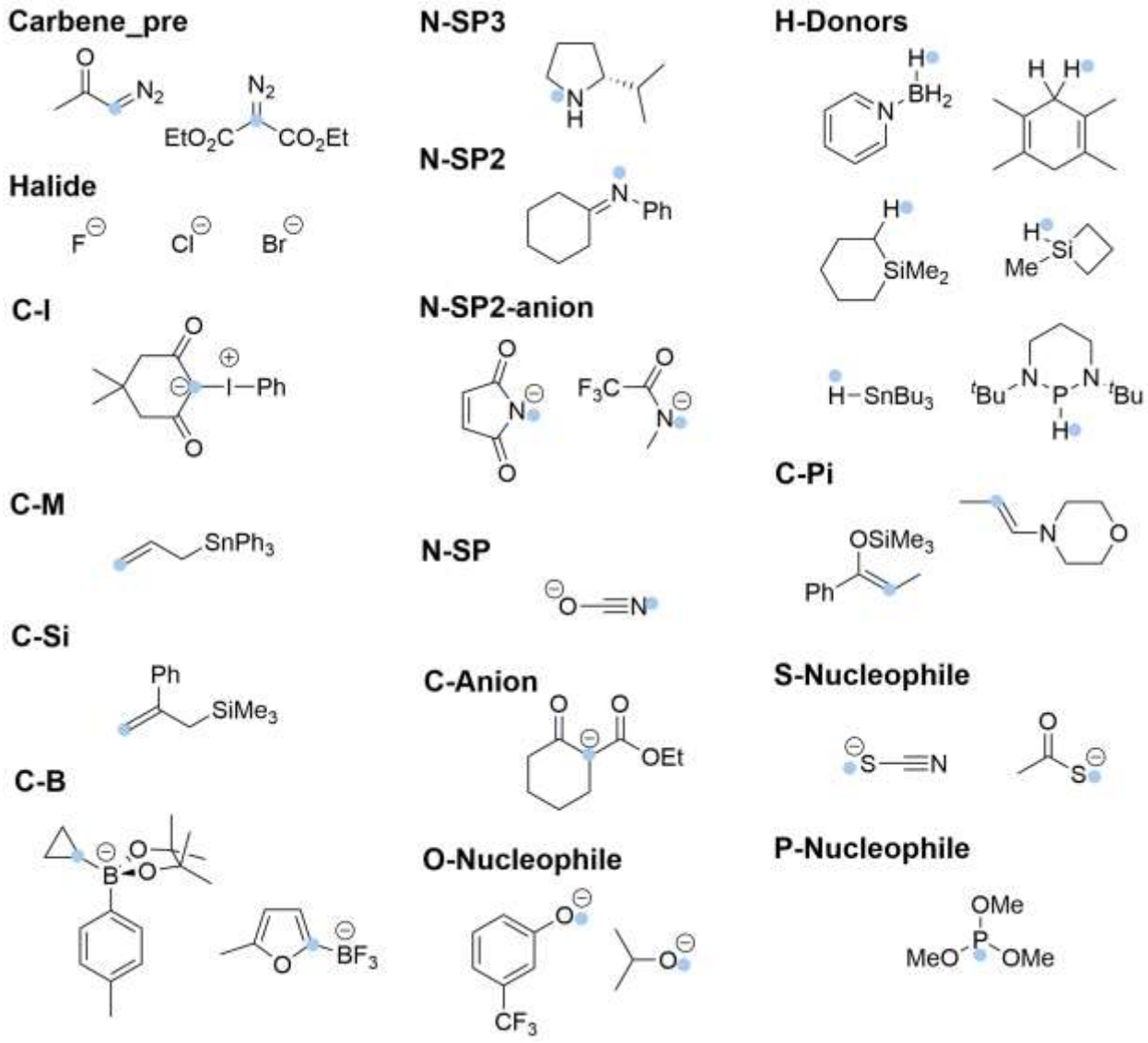

Figure 3. Examples of different classes of nucleophiles in the Mayr's database. The positions which form the new bonds are shown with light blue circles.

3.2 Descriptor development

Previous prediction of nucleophilicity often employed implicit solvation models, as a balance between computational cost and accuracy, in their molecular modeling to represent the effect of solvents.[13,16,17] The impact of these solvation models has been inconclusive. Liu et. al. suggested that solvation does not play an important role in the nucleophilicity parameters of π-nucleophiles.[16]

On the other hand, solvation energy of the nucleophile, protonated-nucleophile and H^+ received a high weighing in the MLR model from Orlandi.[13] Importantly, Buttar and co-workers suggested that the largest error in DFT calculations of activation energy barriers is due to poor performance of solvation models, and compensated that with a combined DFT/machine learning approach.[6]

One of the major developments in reaction solvent selection in recent years has been the development and publication of the ACS Green Chemistry Institute Solvent Selection Tool.[19,20] This model is based on principal component analysis of 70 chemical and physical properties of 272 solvents and has been successfully applied to guide solvent selection in chemical processes. Thus, we decided to build our solvent dependent prediction model based on these PCA descriptors. This has an unique potential advantages: simple expansion of the reactivity prediction to modern green solvents in synthetic processes by changing the solvent descriptors. Specifically, ***sol_PCA1-5*** descriptors were used to represent the properties of the solvent and its influence on nucleophilicity in our models.

For steric hindrance, the Sterimol parameters are the established descriptors in Quantitative Structure-Activity Relationship (QSAR).[33] However, Orlandi and co-workers found that they did not have high weighing in the MLR model.[13] One of the reasons for this could be that the Sterimol parameters were developed for drug design instead of transition state, where interactions over longer distances must be considered. Thus, we decided on using two descriptors based on a cone angle approach at up to 7 Å, similar to that developed for phosphine ligands in organometallic chemistry.[34] To obtain these, a probe Li atom was manually inserted at the relevant angle, based on consensus mechanistic view of these reactions, to the nucleophilic atom. The distance between the nucleophilic atom and Li was standardized as the typical Li-atom bond length, defined as the average bond length found in the CSD (Table S1 in the Supporting Information).[35] ***N_TCA*** is

defined as the sum of the three largest $\widehat{XNuLi}$ angles (where X is another atom within 7 Å distance, approximately twice the typical sum of Van der Waal radii of atoms forming a new bond, and Nu is the nucleophilic atom, Figure 4b). ***N_BAD*** is defined as the single largest maximum $\widehat{XNuLi}$ angle. Atoms are represented as spheres using standard Van der Waal radii and their sizes are included in the calculation of angles.

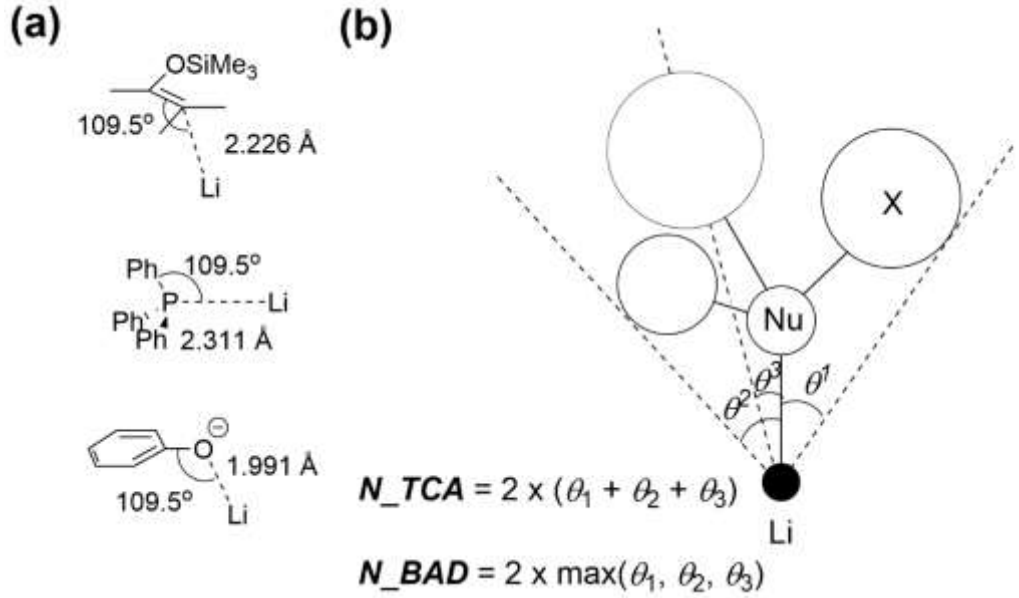

Figure 4. Diagram showing (a) examples of Nu-Li structures used in this study; and (b) the calculation of ***N_TCA*** and ***N_BAD*** descriptors.

The other DFT descriptors (gas phase) with direct relevance to nucleophilicity were HOMO and LUMO energies (***N_HOMO*** and ***N_LUMO***), dipole moment (***N_DM***), electronic basicity (***N_EB***, the most negative Natural Bond Orbital (NBO) partial charge on non-carbon atom), electronic acidity descriptors (***N_EA*** and ***N_EA_nonH***, the highest positive NBO partial charge on hydrogen atom and non-hydrogen atom).[36,37] For these descriptors, the equivalent values based on solution structure (e.g. ***NS_HOMO***, ***NS_EB***, etc.) in the specified solvents for nucleophilicity measurements were also calculated. Additionally, the solvation energy of the nucleophile was also

calculated with PCM solvation model as ***NS_DeltaG***. In addition, we included condensed Fukui function (f^-) at atom with the most negative Fukui function (***N_fukui***,), condensed Fukui function (f^-) at the nucleophilic atom indicated by the Nu-Li structure (***N_fukui_Li)*** and Hirshfeld charge at the nucleophilic atom indicated by the Nu-Li structure (***N_fukui_charge***).[38] The Fukui descriptors were chosen due to their previous successes in representing local reactivity within organic compounds.[39,40] Due to the wide range of structures of nucleophile in our dataset, both ***N_fukui*** and ***N_fukui_Li*** were initially deemed necessary, even if they are significantly correlated.

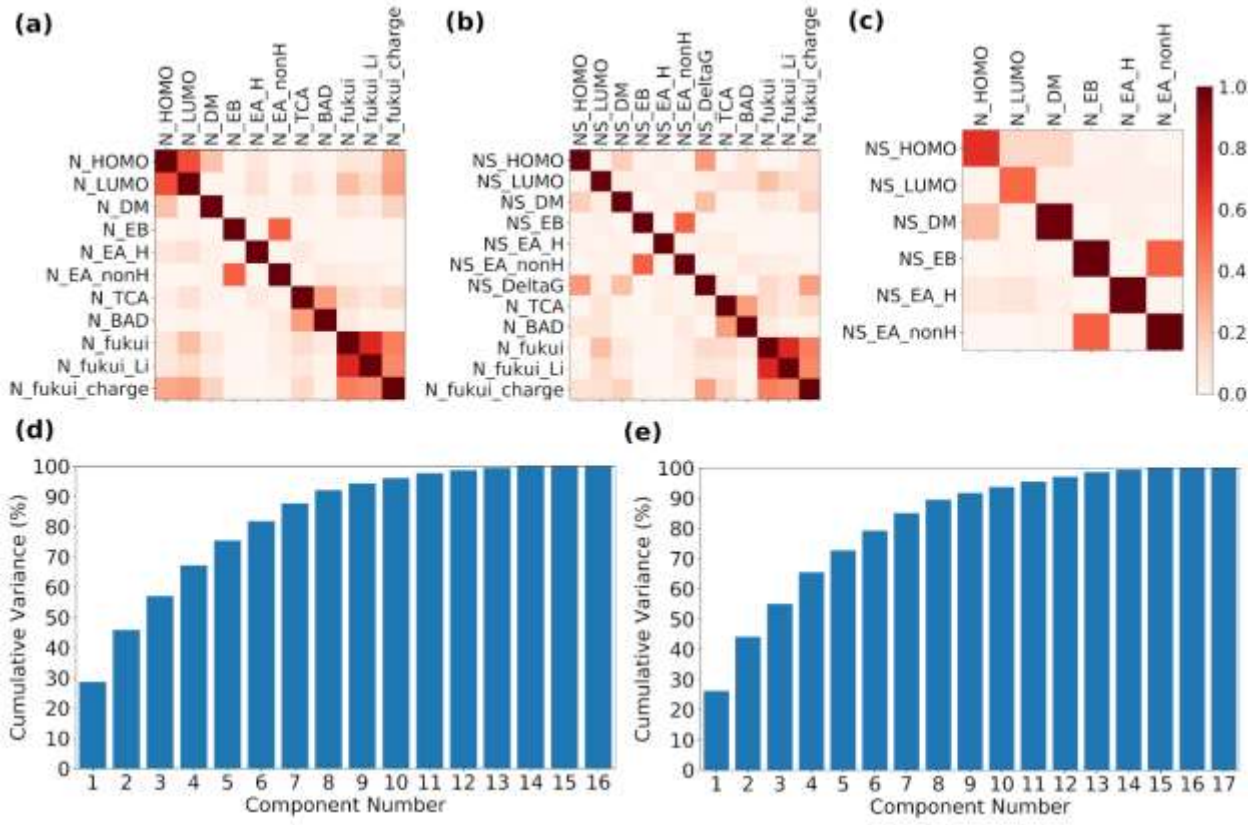

Figure 5. Correlation (R^2) analysis of the nucleophile descriptors in (a) GAS_SET_16; (b) SOL_SET_17; and (c) between GAS_SET_16 and SOL_SET_17; (d) PCA cumulative variance plot for GAS_SET_16; and (e) PCA cumulative variance plot for SOL_SET_17.

Correlation maps between descriptors within GAS_SET_16/***Full_set*** and SOL_SET_17/***Solution_set*** showed that the majority of them are orthogonal, except for the expected correlation between ***N_fukui*** and ***N_fukui_Li*** (Figure 5a-b). The only exception in GAS_SET_16 is moderate correlation between HOMO and LUMO energies, and between ***N_EB*** and ***N_EA_nonH***. Interestingly, there is little correlation between ***NS_HOMO*** and ***NS_LUMO***. This can be explained through examination of ***N_HOMO*** *vs* ***NS_HOMO*** and ***N_LUMO*** *vs* ***NS_LUMO*** relationships (Figure 5c). The R^2 for these are 0.67 and 0.50, respectively, showing the impact of solvent stabilization on the Frontier Molecular Orbital (FMO) energies. The orthogonality between descriptors were confirmed by Principal Component Analysis. The cumulative variance plot (Figure 5d-e) indicated the need for >12 descriptors (including the 5 ***sol_PCA*** descriptors) to capture more than 95% of the variances. This is one of the key aspects of CSPR, which underpins the interpretability of the models.

3.3 Model optimization and interpretation

In order to correctly assess the models, it is important to establish the experimental errors in the measurement of nucleophilic parameters *N*. These were obtained by linear fitting log k_2 *vs* electrophilic parameter *E* from reactions with different electrophiles and determining the intercept with the x-axis. Evaluation of the experimental errors based on the errors in the slope and intercept suggested that the typical experimental error is 1.1 ± 0.9 (Section 8 in the Supporting Information). Thus, in addition to the standard metrics Pearson's R^2 and RMSE, two new metrics were used to assess the models: % of molecules with predicted *N* within ± 1.0 of experimental value (%*N*±1.0) and % of molecules with predicted *N* within ± 2.0 of experimental value (%*N*±2.0). These new metrics reflect the level of noise in the training data, and the practical usability of the predictions

for reaction selectivity prediction, with predicted reaction rate accuracy within $10^{\pm 0.8}$ or $10^{\pm 1.6}$ ($\overline{s_N} \approx 0.8$) times of the observed value.

The size of the Mayr's nucleophilicity parameter dataset excludes effective use of complex deep learning models. Consequently, 8 machine learning algorithms, i.e. MLR, PLS, ANN, SVM, RF, ET, Bag and GP (Table 2, entries 1-8), were applied to a training set of 808 nucleophile/solvent combination and a test set of 96 nucleophile/solvent combinations (***Full_set***, Table 1). Two sets of descriptors (GAS_SET_16 and SOL_SET_17) were used, resulting in 16 initial models. A small number of failed Gaussian optimizations in solution means led to a reduction of training/test set of 802 and 94 entries (***Solution_set***). Model parameters were optimized through iteration and evaluation of model metrics to maximize accuracy while avoiding overfitting (Section 5.2 in the Supporting Information). The results are summarized in Table 2 and 3.

Table 2. Metrics for models built with GAS_SET_16 descriptors

No.	Method	R^2	RMSE	*%N*±2.0	*%N*±1.0
1	MLR (Multivariate Linear Regression)[a]	0.63	4.74	26.0	12.5
2	PLS (Partial Least Square)[a]	0.63	4.73	26.0	12.5
3	ANN (Artificial Neural Network)[a]	0.86	2.84	59.4	31.8
4	SVM (Support Vector Machine)[a]	0.86	2.82	55.2	34.4
5	RF (Random Forest)[a]	0.89	2.53	71.6	49.2
6	ET (Extra Trees)[a]	0.92	2.18	70.2	49.6
7	Bag (Bagging)[a]	0.89	2.53	72.3	48.1
8	GP (Gaussian Process)[a]	0.82	3.19	59.4	33.3
9	MLR[b]	0.63	4.60	31.4	15.8
10	PLS[b]	0.63	4.60	31.9	15.3
11	ANN[b]	0.84	2.99	58.1	34.5

12	SVM[b]	0.83	3.16	61.2	37.3
13	RF[b]	0.89	2.47	69.8	45.5
14	ET[b]	0.91	2.30	72.4	47.9
15	Bag[b]	0.89	2.47	69.4	45.9
16	GP[b]	0.79	3.41	57.9	33.0

[a]result with training/test split, averaged over 10 runs; [b]results with 10-fold cross validation

Table 3. Metrics for models built with SOL_SET_17 descriptors

No.	Method	R^2	RMSE	*%N*±2.0	*%N*±1.0
1	MLR[a]	0.67	4.24	35.1	13.8
2	PLS[a]	0.67	4.24	35.1	13.8
3	ANN[a]	0.87	2.76	56.9	32.3
4	SVM[a]	0.89	2.33	70.2	44.7
5	RF[a]	0.92	2.05	76.1	49.1
6	ET[a]	0.94	1.84	79.9	53.5
7	Bag[a]	0.92	2.05	76.4	48.7
8	GP[a]	0.87	2.63	68.1	42.6
9	MLR[b]	0.68	4.27	35.8	17.6
10	PLS[b]	0.68	4.27	35.7	17.1
11	ANN[b]	0.87	2.67	64.9	36.2
12	SVM[b]	0.87	2.73	67.9	39.2
13	RF[b]	0.91	2.29	74.7	47.1
14	ET[b]	0.92	2.14	75.9	50.9
15	Bag[b]	0.91	2.29	74.9	46.7
16	GP[b]	0.85	2.96	66.8	40.7

[a]result with training/test split, averaged over 10 runs; [b]results with 10-fold cross validation

MLR and PLS performed poorly with this much larger set of nucleophiles, despite reported success by Orlandi (Figure 2b). The non-linear regression algorithms gave better results. Random-forest-based models outperformed ANN, SVM and GP, with ET giving the best metrics. ET model for GAS_SET_16 gave $R^2 = 0.92$ and RSME = 2.18 (Table 2, entry 6), which is close to the upper limit of the estimated experimental error of 1.1 ± 0.9. Importantly, 70.2% of the predicted *N* values were within ±2 to the experimental values and 49.6% were within ±1.0. Similar metrics were obtained with 10-fold cross validation, albeit with slightly larger standard deviations due to the inclusion of predictions of all data points from 10 separated runs (Table 2, entry 14). The obtained results using SOL_SET_17 descriptors showed significant improvements with all metrics, compared to those obtained with GAS_SET_16 (Table 3). Specifically, the ET model R^2 increased from 0.92 to 0.94 (compared $R^2 = 0.95$ value reported by Orlandi),[13] RMSE decreased from 2.18 to 1.84, %*N*±2.0 and %*N*±1.0 increased from 70.2 and 49.6 to 79.9 and 53.5, respectively (Table 3, entry 6). The dependence on solvent of nucleophilicity parameters was reliably reproduced by the models, with nucleophilicity increasing in the order of DCM < MeCN < water < DMSO (Figure 6c). Few outliers were observed with DCM, DMSO and water. This level of reliability is essential in applying such predictions models to predict solvent-dependent selectivity of synthetic reactions. Similar results were obtained in 10-fold cross validation models, albeit with the expected minor decrease in model metrics due to the more complete test sets.

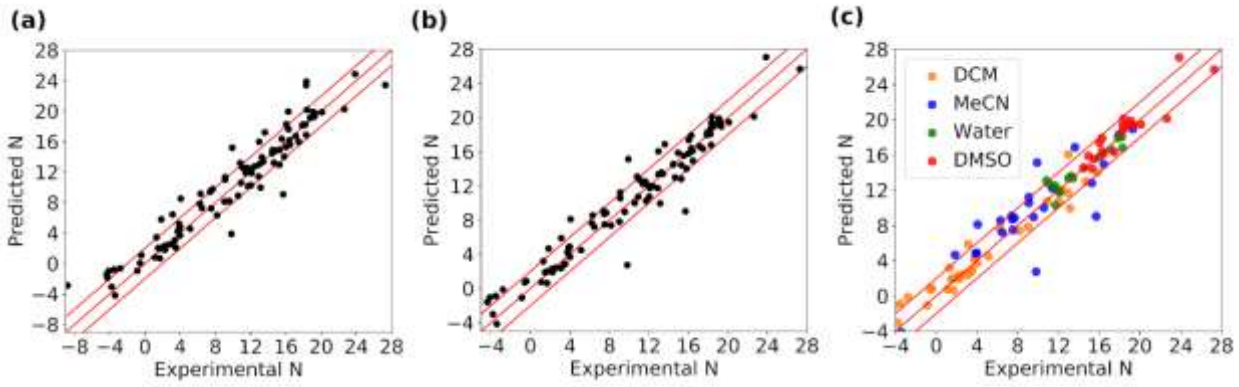

Figure 6. Predicted *vs* experimental nucleophilicity parameters from training/test split models using (a) GAS_SET_16; (b) SOL_SET_17 descriptors; and (c) solvent dependent predictions vs experimental measurements using SOL_SET_17 for nucleophilicity parameters in water (green dots), MeCN (blue dots), dichloromethane (orange dots) and DMSO (red dots).

The dependence of the ET models on the descriptors is summarized in Figure 7a-f. Direct feature importance plot, an option only available with RF based models, showed that the predictions depended on both solvent descriptors and nucleophile descriptors. On the other hand, ***NS_DeltaG***, the additional descriptor in SOL_SET_17, only had a moderate importance (Figure 7b). Other than this, the feature importance plots are strikingly similar for GAS_SET_16 and SOL_SET_17. Amongst the nucleophile descriptors, ***N_HOMO***/***NS_HOMO*** and ***N_fukui_charge*** were the most important. Both are expected from causal relationships between nucleophilicity and molecular and local properties of the nucleophile.[38,41] As feature importance plots are only available to RF based models, the impact of each descriptor is also evaluated by monitoring the model metrics when each descriptor is skipped in turn (Figure 7c-d and section 7.3 of the Supporting Information). The largest negative impacts were observed when ***N_HOMO***/***NS_HOMO*** is skipped. The impact of skipping other descriptors is less conclusive. Consequently, a different test was devised. Every combination of 3 descriptors (560 combinations for GAS_SET_16 and 680 for SOL_SET_17) was

skipped in turn and their impacts on the model metrics (i.e. %N±2.0) were evaluated. This approach would amplify the impact of each missing descriptor. In the ET model built with GAS_SET_16, ***N_HOMO***, ***N_LUMO***, ***N_BAD***, ***N_fukui***, ***N_fukui_Li*** and ***N_fukui_charge*** are the descriptors which are absent in 10 or more of the 50 worst performing models (Figure 7e). In the ET model built with SOL_SET_17, ***NS_HOMO***, ***NS_EB***, ***NS_EA_nonH***, ***N_fukui***, and ***N_fukui_charge*** are the descriptors which are most often absent in the worst performing models. The 50 worst models using SOL_SET_17 are dominated by ***NS_HOMO***, absent in 48 models compared to 11 models for ***N_HOMO***, which renders the importance of other descriptors less reliable (Figure 7f). While both ***N_fukui*** and ***N_fukui_charge*** descriptors are important, ***N_fukui*** is also ambiguously absent in many of the 50 best performing models with both GAS_SET_16 and SOL_SET_17. Finally, the absence of ***N_DM***/***NS_DM*** can be associated with some of the best models, particularly when R^2 is used as the metric (Figure S59 in the Supporting Information).

Interestingly, the identified important descriptors above are all related to the electronic properties of the nucleophiles. The steric descriptors, ***N_TCA*** and ***N_BAD***, did not have large impact on any of the models, in spite of our effort to expand the distance of the steric descriptors. This is similar to the observation by Orlandi and co-workers in their study.[13] A likely explanation is that the experimental data in Mayr's database is self-selecting and biased toward nucleophiles which can react effectively enough for kinetic reaction rate measurements. Thus, the majority of the nucleophiles in the database did not suffer from significant steric hindrance. In order to accurately estimate the influence of steric hindrance on nucleophilicity, additional data on nucleophiles with more significant steric hindrance are needed.

Analysis of the outliers (defined as those with predicted N values outside the experimental N ± 2.0 range) showed that they are evenly distributed across the range of N values (Figure S45 and

S51 in the Supporting Information). The ET model using GAS_SET_16 has outliers in all four solvents, with a slightly higher ratio of outlier nucleophiles in MeCN and a lower ratio in DCM. However, the ET model using SOL_SET_17 had a significant reduction of number of outliers in all solvents, but with much fewer outliers in DMSO, likely due to the high solvation energy of the nucleophiles and transition states in this solvent (Figure 7h and 7j). The distribution of outliers by classes of nucleophile is also revealing (Figure 7g and 7i). A disproportionate number of outliers are in the classes of C-B, H-donor and C-Anion. Two of these are also classes which were absent (C-B and H-donor) in the previous MLR model.[13] These are nucleophiles which either reacts *via* more complex mechanisms or can be significantly influenced by solvation. In fact, the ET model built with SOL_SET_17 resulted in a significant decrease in the number of outliers in C-Anion classes.

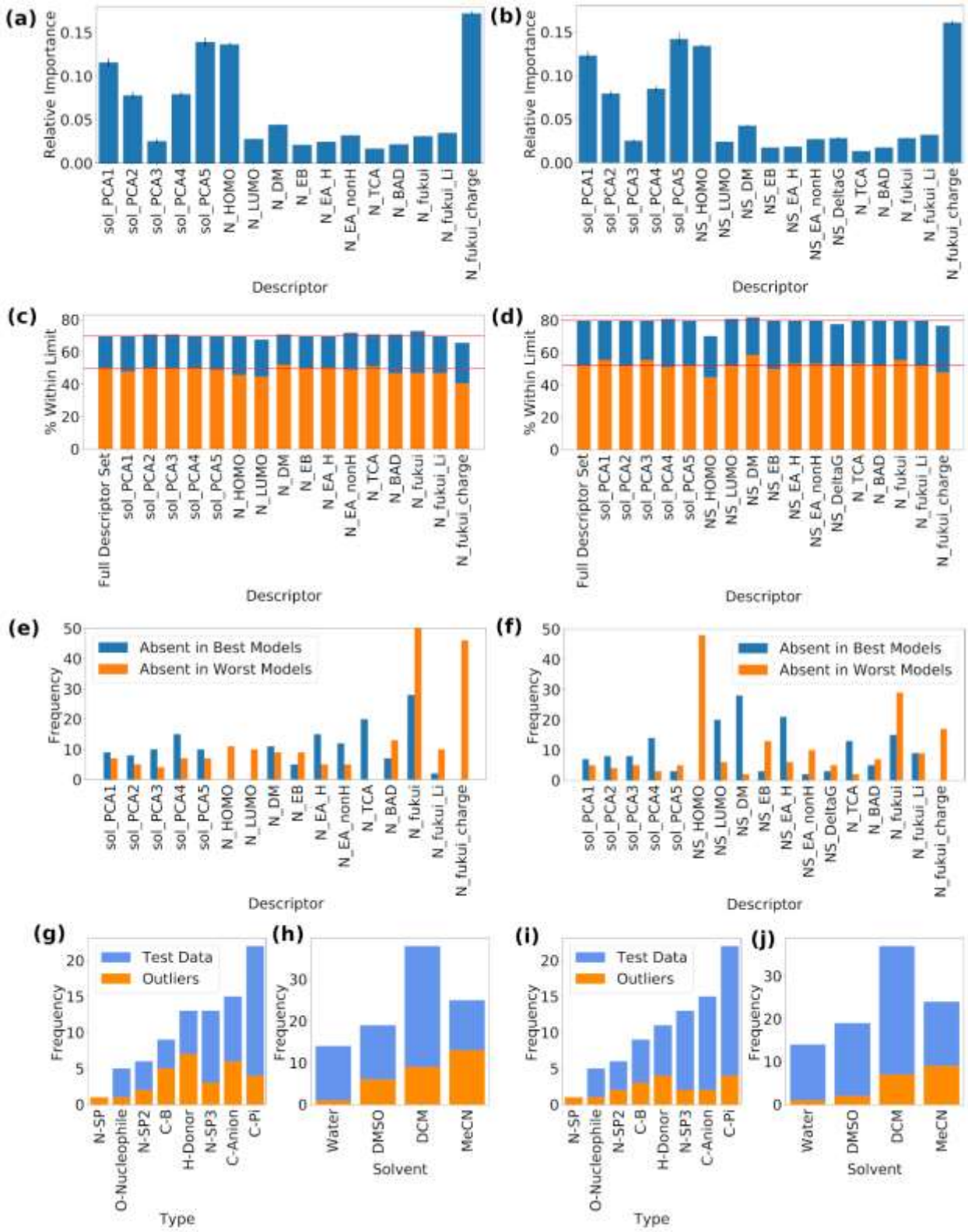
(a)
Relative Importance
0.00
0.05
0.10
0.15
sol_PCA1
sol_PCA2
sol_PCA3
sol_PCA4
sol_PCA5
N_HOMO
N_LUMO
N_DM
N_EB
N_EA_H
N_EA_nonH
N_TCA
N_BAD
N_fukui
N_fukui_Li
N_fukui_charge
Descriptor
(b)
Relative Importance
0.00
0.05
0.10
0.15
sol_PCA1
sol_PCA2
sol_PCA3
sol_PCA4
sol_PCA5
NS_HOMO
NS_LUMO
NS_DM
NS_EB
NS_EA_H
NS_EA_nonH
NS_DeltaG
N_TCA
N_BAD
N_fukui
N_fukui_Li
N_fukui_charge
Descriptor
(c)
% Within Limit
0
20
40
60
80
Full Descriptor Set
sol_PCA1
sol_PCA2
sol_PCA3
sol_PCA4
sol_PCA5
N_HOMO
N_LUMO
N_DM
N_EB
N_EA_H
N_EA_nonH
N_TCA
N_BAD
N_fukui
N_fukui_Li
N_fukui_charge
Descriptor
(d)
% Within Limit
0
20
40
60
80
Full Descriptor Set
sol_PCA1
sol_PCA2
sol_PCA3
sol_PCA4
sol_PCA5
NS_HOMO
NS_LUMO
NS_DM
NS_EB
NS_EA_H
NS_EA_nonH
NS_DeltaG
N_TCA
N_BAD
N_fukui
N_fukui_Li
N_fukui_charge
Descriptor
(e)
Absent in Best Models
Absent in Worst Models
Frequency
0
10
20
30
40
50
sol_PCA1
sol_PCA2
sol_PCA3
sol_PCA4
sol_PCA5
N_HOMO
N_LUMO
N_DM
N_EB
N_EA_H
N_EA_nonH
N_TCA
N_BAD
N_fukui
N_fukui_Li
N_fukui_charge
Descriptor
(f)
Absent in Best Models
Absent in Worst Models
Frequency
0
10
20
30
40
50
sol_PCA1
sol_PCA2
sol_PCA3
sol_PCA4
sol_PCA5
NS_HOMO
NS_LUMO
NS_DM
NS_EB
NS_EA_H
NS_EA_nonH
NS_DeltaG
N_TCA
N_BAD
N_fukui
N_fukui_Li
N_fukui_charge
Descriptor
(g)
Test Data
Outliers
Frequency
0
5
10
15
20
N-SP
O-Nucleophile
N-SP2
C-B
H-Donor
N-SP3
C-Anion
C-Pi
Type
(h)
Frequency
0
10
20
30
Water
DMSO
DCM
MeCN
Solvent
(i)
Test Data
Outliers
Frequency
0
5
10
15
20
N-SP
O-Nucleophile
N-SP2
C-B
H-Donor
N-SP3
C-Anion
C-Pi
Type
(j)
Frequency
0
10
20
30
Water
DMSO
DCM
MeCN
Solvent

Figure 7. Model analysis for ET models: (a) feature importance for model using GAS_SET_16, average of 10 runs; (b) feature importance for model using SOL_SET_17, average of 10 runs; (c) impact of taking out one descriptor on the model *via* %*N*±1.0 (orange) and %*N*±2.0 (blue) metrics, using GAS_SET_16; (d) impact of taking out one descriptor on the model *via* %*N*±1.0 (orange) and %*N*±2.0 (blue), using SOL_SET_17; (e) frequency of descriptors absence in best and worst 50 models, with 3 descriptors removed, based on %*N*±2.0 metrics, using GAS_SET_16; (f) frequency of descriptors absence in 50 best and worst models, with 3 descriptors removed, based on %*N*±2.0 metrics, using SOL_SET_17; (g) distribution of outliers based on classes of nucleophile, using GAS_SET_16; (h) distribution of outliers based on solvent, using GAS_SET_16; (i) distribution of outliers based on classes of nucleophile, using SOL_SET_16; and (j) distribution of outliers based on solvent, using SOL_SET_17.

3.4 Model improvement

The difference between ***N_fukui*** and ***N_fukui_Li*** is the local atoms for which the condensed Fukui functions are calculated. For ***N_fukui***, this is the atom with the most negative Fukui function. For ***N_fukui_Li***, this is the atom which forms the new bond with the electrophile, based on the nucleophile-Li structures employed for the calculation of steric descriptors. Thus, while there is some overlap between these two descriptors, ***N_fukui_Li*** is more directly relevant to the transition state of the reaction between the nucleophile and any given electrophile and ***N_fukui*** can some case give the wrong information. This was reflected in the descriptor dependence analysis from the ET model, wherein ***N_fukui*** has a negative impact on the model (Figure 7e and 7f). In addition, exclusion of ***N_DM***/***NS_DM*** descriptor resulted in a small increase of %*N*±2.0 metrics and a significant increase in %*N*±1.0 metrics (Figure 7e and 7d). Dipole moment may influence solvation energy and nucleophilicity. However, Figure 7c-7d showed that ***NS_DeltaG*** is

not important in the ET model and the local descriptors, e.g. Fukui functions and charges, have much more influences over the nucleophilicity parameter.

Based on these results, the descriptors ***N_DM*** and ***N_fukui*** were removed from GAS_SET_16 and ***NS_DM*** and ***N_fukui*** was removed from SOL_SET_17, giving GAS_SET_14 and SOL_SET_15, respectively. The new ET models using these descriptor sets showed a very minor deterioration in the traditional metrics R^2 and RMSE (Table 4). However, the metrics %*N*±2.0 and %*N*±1.0, which account for the experimental errors in the data, showed significant improvement in the accuracy of the models. Most impressively, the ET model with SOL_SET_15 achieved 81.6% of predictions of the test set within ±2.0 and 58.6% of prediction of the test set within ±2.0 of the experimental *N* values. On the other hand, the overall impact of leaving out ***N_DM/NS_DM*** and ***N_fukui*** on the 10-fold validation models (Table 4, entries 2, 4, 6 and 8) is negligible with a slight increase in RMSE.

Table 4. Metrics for rationally improved ET models

No.	Descriptor set	Modification	R^2	RMSE	%*N*±2.0	%*N*±1.0
1[a]	GAS_SET_16	None	0.92	2.18	70.2	49.6
2[b]	GAS_SET_16	None	0.91	2.30	72.4	47.9
3[a]	GAS_SET_14	Removing ***N_DM*** and ***N_fukui***	0.92	2.21	71.3	51.1
4[b]	GAS_SET_14	Removing ***N_DM*** and ***N_fukui***	0.91	2.30	72.6	48.0
5[a]	SOL_SET_17	None	0.94	1.84	79.9	53.5
6[b]	SOL_SET_17	None	0.92	2.14	75.9	50.9
7[a]	SOL_SET_15	Removing ***NS_DM*** and ***N_fukui***	0.93	1.90	81.6	58.6
8[b]	SOL_SET_15	Removing ***NS_DM*** and ***N_fukui***	0.91	2.17	75.8	51.3

[a]result with training/test split averaged over 10 runs; [b]results with 10-fold cross validation

Finally, the nucleophilic descriptors, except ***N_fukui***, ***N_fukui_Li*** and ***N_fukui_charge***, were generated using PM6 instead of DFT in Gaussian 09. This led to a decrease of CPU time from 3076.0 hours for DFT to 352.0 for PM6 for 896 nucleophiles. Charge descriptors ***N_Ea***/***NS_Ea***, ***N_Eb***/***NS_Eb***, ***N_Ea_NonH***/***NS_Ea_NonH*** were derived from (NBO) population analysis from DFT calculation, which is not possible with PM6. Thus, Mulliken partial charges were used instead.

While little decrease in ET model metrics was observed between GAS_SET_14 and GAS_SET_14_PM6 (Table 4, entry 3 and Table 5, entry 2), a significant deterioration in model accuracy was observed between the ET models using SOL_SET_15 and SOL_SET_15_PM6 (Table 4, entry 7 and Table 5, entry 4). Switching from DFT descriptors to PM6 descriptors led to a decrease of 11.7 in *%N*±2.0 and 7.0 in *%N*±1.0. The much larger decrease in model accuracy observed with SOL_SET_15_PM6 compared to that of GAS_SET_14_PM6 can be attributed to the documented poor accuracy of PM6/PCM solvation model compared to more computationally expensive models, such as HF/SMD.[42,43]

Table 5. Metrics for ET models with semi-empirical PM6 descriptors[a]

No.	Descriptor set	Modification	R^2	RMSE	*%N*±2.0	*%N*±1.0
1	GAS_SET_16_PM6	None	0.92	2.14	68.6	48.1
2	GAS_SET_14_PM6	Removing ***N_DM_PM6*** and ***N_fukui***	0.92	2.13	71.5	49.8
3	SOL_SET_17_PM6	None	0.93	1.98	69.6	49.0
4	SOL_SET_15_PM6	Removing ***NS_DM_PM6*** and ***N_fukui***	0.92	2.02	69.9	51.6

[a]result with training/test split averaged over 10 runs; [b]results with 10-fold cross validation

4. CONCLUSIONS

Highly accurate prediction of nucleophilicity parameter *N* in four most common solvents has been achieved with Extra Trees algorithm across the whole range of nucleophiles in the Mayr's reactivity parameter database (R^2 = 0.93, RMSE = 1.90, %*N*±2.0 = 81.6 and %*N*±1.0 = 58.6, compared to an estimated experimental error of 1.1 ± 0.9).[18] This was achieved using our CSPR approach,[19] which focuses on the causal physicochemical relationships between the orthogonal descriptors and the predicted parameters, and rational improvements of the prediction models. The use of PCA descriptors for the solvents, based on the ACS Solvent Tool,[20,21] which opens up the possibility of rapid expansion to prediction of nucleophilicity in modern green solvents with little direct literature reactivity data. These will underpin prediction of reaction selectivity in different solvents. Analysis of the models showed that steric factors are still under-represented, due to bias in the experimental database. The most important descriptors are solvent-dependent HOMO energy and Hirshfeld charge of the nucleophilic atom. Replacing DFT descriptors with PM6 descriptors for the nucleophiles led to an 8.7-fold decrease in computational time, and approximately 10% decrease in the percentage of predictions within ±2.0 and ±1.0 of the experimental values.

ASSOCIATED CONTENT

Supporting Information: Details and protocols of data curation, calculation and analysis of descriptors, machine learning models optimization and analysis, and examples of Python code.

DATA AND SOFTWARE AVAILABILITY

The dataset from open literature, including calculated descriptors, and the code in this manuscript can be accessed from this link: https://zenodo.org/badge/latestdoi/359866732 and this link:

https://github.com/BNNLab/Nucleophilicity/tree/main/Codes. Citations should refer directly to this manuscript

AUTHOR INFORMATION

Corresponding Author

Dr Bao N. Nguyen Institute of Process Research & Development, School of Chemistry, University of Leeds, Woodhouse Lane, LS2 9JT, United Kingdom.

Present Addresses

†If an author's address is different than the one given in the affiliation line, this information may be included here.

Author Contributions

S.B and B.N.N. co-wrote the manuscript. S.B., Y.L. and K.S. carried out the experiments and machine learning work. B.N., A.J.B, D.H. and N.K. supervised the project. All authors have reviewed and edited the manuscript and contributed to useful discussions.

Funding Sources

S.B. thanks AstraZeneca and the EPSRC for his iCASE studentship. K.S., A.J.B., N.K. and B.N.N. thank the EPSRC for their support through grant number EP/S013768/1.

Notes

Any additional relevant notes should be placed here.

ACKNOWLEDGMENT

This work was undertaken on ARC2, ARC3 and ARC4, part of the High Performance Computing facilities at the University of Leeds, UK.

ABBREVIATIONS

CSPR: Causal Structure Property Relationship

REFERENCES

(1) Peng, Q.; Duarte, F.; Paton, R. S. Computing Organic Stereoselectivity – from Concepts to Quantitative Calculations and Predictions. *Chem. Soc. Rev.* **2016**, *45* (22), 6093–6107. https://doi.org/10.1039/C6CS00573J.

(2) Wheeler, S. E.; Seguin, T. J.; Guan, Y.; Doney, A. C. Noncovalent Interactions in Organocatalysis and the Prospect of Computational Catalyst Design. *Acc. Chem. Res.* **2016**, *49* (5), 1061–1069. https://doi.org/10.1021/acs.accounts.6b00096.

(3) Bahmanyar, S.; Houk, K. N. Origins of Opposite Absolute Stereoselectivities in Proline-Catalyzed Direct Mannich and Aldol Reactions. *Org. Lett.* **2003**, *5* (8), 1249–1251. https://doi.org/10.1021/ol034198e.

(4) Llabrés, S.; Vicente-García, E.; Preciado, S.; Guiu, C.; Pouplana, R.; Lavilla, R.; Luque, F. J. Evolution of a Multicomponent System: Computational and Mechanistic Studies on the Chemo- and Stereoselectivity of a Divergent Process. *Chem. – A Eur. J.* **2013**, *19* (40), 13355–13361. https://doi.org/https://doi.org/10.1002/chem.201302072.

(5) Struebing, H.; Ganase, Z.; Karamertzanis, P. G.; Siougkrou, E.; Haycock, P.; Piccione, P. M.; Armstrong, A.; Galindo, A.; Adjiman, C. S. Computer-Aided Molecular Design of Solvents for Accelerated Reaction Kinetics. *Nat. Chem.* **2013**, *5* (11), 952–957.

https://doi.org/10.1038/nchem.1755.

(6) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine Learning Meets Mechanistic Modelling for Accurate Prediction of Experimental Activation Energies. *Chem. Sci.* **2021**, *12* (3), 1163–1175. https://doi.org/10.1039/D0SC04896H.

(7) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **2018**, *4* (11), 1465–1476. https://doi.org/10.1021/acscentsci.8b00357.

(8) Caputo, F.; Corbetta, S.; Piccolo, O.; Vigo, D. Seeking for Selectivity and Efficiency: New Approaches in the Synthesis of Raltegravir. *Org. Process Res. Dev.* **2020**, *24* (6), 1149–1156. https://doi.org/10.1021/acs.oprd.0c00155.

(9) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. – A Eur. J.* **2017**, *23* (25), 5966–5971. https://doi.org/https://doi.org/10.1002/chem.201605499.

(10) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610. https://doi.org/10.1038/nature25978.

(11) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3* (10), 1103–1113. https://doi.org/10.1021/acscentsci.7b00303.

(12) Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y. Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *J. Chem. Inf. Model.* **2020**, *60* (1), 47–55. https://doi.org/10.1021/acs.jcim.9b00949.

(13) Orlandi, M.; Escudero-Casao, M.; Licini, G. Nucleophilicity Prediction via Multivariate Linear Regression Analysis. *J. Org. Chem.* **2021**, *86* (4), 3555–3564. https://doi.org/10.1021/acs.joc.0c02952.

(14) Lee, B.; Yoo, J.; Kang, K. Predicting the Chemical Reactivity of Organic Materials Using a Machine-Learning Approach. *Chem. Sci.* **2020**, *11* (30), 7813–7822. https://doi.org/10.1039/D0SC01328E.

(15) Mayr, H.; Patz, M. Scales of Nucleophilicity and Electrophilicity: A System for Ordering Polar Organic and Organometallic Reactions. *Angew. Chemie Int. Ed. English* **1994**, *33* (9), 938–957. https://doi.org/https://doi.org/10.1002/anie.199409381.

(16) Wang, C.; Fu, Y.; Guo, Q.-X.; Liu, L. First-Principles Prediction of Nucleophilicity Parameters for π Nucleophiles: Implications for Mechanistic Origin of Mayr's Equation. *Chem. – A Eur. J.* **2010**, *16* (8), 2586–2598. https://doi.org/https://doi.org/10.1002/chem.200902484.

(17) DEKA, K.; PHUKAN, P. DFT Analysis of the Nucleophilicity of Substituted Pyridines and Prediction of New Molecules Having Nucleophilic Character Stronger than 4-Pyrrolidino Pyridine. *J. Chem. Sci.* **2016**, *128* (4), 633–647. https://doi.org/10.1007/s12039-016-1057-5.

(18) Mayr's Database of Reactivity Parameters

https://www.cup.lmu.de/oc/mayr/reaktionsdatenbank2/.

(19) Boobier, S.; Hose, D. R. J.; Blacker, A. J.; Nguyen, B. N. Machine Learning with Physicochemical Relationships: Solubility Prediction in Organic Solvents and Water. *Nat. Commun.* **2020**, *11* (1), 5753. https://doi.org/10.1038/s41467-020-19594-z.

(20) Diorazio, L. J.; Hose, D. R. J.; Adlington, N. K. Toward a More Holistic Framework for Solvent Selection. *Org. Process Res. Dev.* **2016**, *20* (4), 760–773. https://doi.org/10.1021/acs.oprd.6b00015.

(21) Roundtable, A. G. P. Solvent Selection Tool https://www.acs.org/content/acs/en/greenchemistry/research-innovation/tools-for-green-chemistry/solvent-tool.html.

(22) Richardson, L. Beautiful soup documentation https://beautiful-soup-4.readthedocs.io/en/latest/.

(23) Swain Matt. Python interface for the Chemical Identifier Resolver https://cirpy.readthedocs.io/en/latest/.

(24) Swain Matt. Python interface for the Chemical Identifier Resolver.

(25) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski,

V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A.; Jr., J. E. P.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 09. Gaussian, Inc.: Wallingford CT 2016.

(26) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1456–1465. https://doi.org/https://doi.org/10.1002/jcc.21759.

(27) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. Sect. B* **2016**, *72* (2), 171–179.

(28) Lu, T.; Chen, F. Multiwfn: A Multifunctional Wavefunction Analyzer. *J. Comput. Chem.* **2012**, *33* (5), 580–592. https://doi.org/10.1002/jcc.22885.

(29) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. %J J. of machine learning research. Scikit-Learn: Machine Learning in Python. **2011**, *12* (Oct), 2825–2830.

(30) GPy: A Gaussian process framework in python http://github.com/SheffieldML/GPy.

(31) Austin, J. D.; Eaborn, C. Organosilicon Compounds. Part XXIX. The Stereochemical Course of the Reaction of a Silicon Hydride with Chlorotriphenylmethane. *J. Chem. Soc.* **1964**, No. 0, 2279–2280. https://doi.org/10.1039/JR9640002262.

(32) Zheng, C.; You, S.-L. Transfer Hydrogenation with Hantzsch Esters and Related Organic Hydride Donors. *Chem. Soc. Rev.* **2012**, *41* (6), 2498–2518. https://doi.org/10.1039/C1CS15268H.

(33) Verloop, A.; Ariens, E. J. *Drug Design*; 1976; Vol. III.

(34) Tolman, C. A. Steric Effects of Phosphorus Ligands in Organometallic Chemistry and Homogeneous Catalysis. *Chem. Rev.* **1977**, *77*, 313.

(35) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. Sect. B* **2016**, *72* (2), 171–179.

(36) Borhani, T. N.; García-Muñoz, S.; Vanesa Luciani, C.; Galindo, A.; Adjiman, C. S. Hybrid QSPR Models for the Prediction of the Free Energy of Solvation of Organic Solute/Solvent Pairs. *Phys. Chem. Chem. Phys.* **2019**, *21* (25), 13706–13720. https://doi.org/10.1039/C8CP07562J.

(37) Famini, G. R.; Wilson, L. Y. Using Theoretical Descriptors in Linear Free Energy Relationships: Characterizing Several Polarity, Acid and Basicity Scales. *J. Phys. Org. Chem.* **1999**, *12* (8), 645–653. https://doi.org/https://doi.org/10.1002/(SICI)1099-1395(199908)12:8<645::AID-POC165>3.0.CO;2-S.

(38) Oláh, J.; Van Alsenoy, C.; Sannigrahi, A. B. Condensed Fukui Functions Derived from Stockholder Charges: Assessment of Their Performance as Local Reactivity Descriptors. *J. Phys. Chem. A* **2002**, *106* (15), 3885–3890. https://doi.org/10.1021/jp014039h.

(39) Gilardoni, F.; Weber, J.; Chermette, H.; Ward, T. R. Reactivity Indices in Density

Functional Theory: A New Evaluation of the Condensed Fukui Function by Numerical Integration. *J. Phys. Chem. A* **1998**, *102* (20), 3607–3613. https://doi.org/10.1021/jp980521x.

(40) Thanikaivelan, P.; Padmanabhan, J.; Subramanian, V.; Ramasami, T. Chemical Reactivity and Selectivity Using Fukui Functions: Basis Set and Population Scheme Dependence in the Framework of B3LYP Theory. *Theor. Chem. Acc.* **2002**, *107* (6), 326–335. https://doi.org/10.1007/s00214-002-0352-z.

(41) Hirshfeld, F. L. Bonded-Atom Fragments for Describing Molecular Charge Densities. *Theor. Chim. Acta* **1977**, *44* (2), 129–138. https://doi.org/10.1007/BF00549096.

(42) Kříž, K.; Řezáč, J. Reparametrization of the COSMO Solvent Model for Semiempirical Methods PM6 and PM7. *J. Chem. Inf. Model.* **2019**, *59* (1), 229–235. https://doi.org/10.1021/acs.jcim.8b00681.

(43) Kromann, J. C.; Steinmann, C.; Jensen, J. H. Improving Solvation Energy Predictions Using the SMD Solvation Method and Semiempirical Electronic Structure Methods. *J. Chem. Phys.* **2018**, *149* (10), 104102. https://doi.org/10.1063/1.5047273.