



Cite this: *Phys. Chem. Chem. Phys.*,  
2022, 24, 26802

# A machine learning approach for predicting the fluorination strength of electrophilic fluorinating reagents†

Vaneet Saini 

The unusual properties of a wide range of organofluorine compounds have provided strong incentives to the scientific community for the development of this field. In parallel to the constantly growing number of organofluorine compounds, an unusually high number of electrophilic N–F fluorinating reagents have emerged as potential fluorinators to achieve fluorine substitution in a simple and efficient manner. Bench stability, crystalline nature and modular synthesis are some of the key characteristics that make them increasingly important in synthetic transformations. In this context, it is important to understand the reactive power of these N–F fluorinating reagents in a quantitative manner. Experimental and DFT investigations to obtain a quantitative understanding of the fluorination power of these reagents are resource intensive, laborious and expensive. Herein, we propose a machine learning approach for predicting the relative power of a wide range of N–F fluorinating reagents by utilizing a simple and fast SMILES-based molecular encoding approach. A neural network algorithm was employed on a novel dataset consisting of four molecular descriptors, two categorical descriptors and 260 data points and was successful in predicting the fluorine plus detachment values for N–F fluorinating reagents belonging to six different categories.

Received 18th July 2022,  
Accepted 17th October 2022

DOI: 10.1039/d2cp03281c

rsc.li/pccp

## Introduction

The importance of fluorine was largely ignored by the pharmaceutical industry until the first half of the 20th century, potentially due its absence in nature's pool, as the majority of drug molecules were either natural product derivatives or inspired by them.<sup>1</sup> Additionally, the general notion that the high oxidation potential of fluorine would make it quite difficult to introduce in organic molecules kept it distanced from drug design and synthesis. However, the advent of 5-fluorouracil and fludrocortisone, which have remarkable biological properties, in the 1950s completely transformed the prevailing perspective regarding the inclusion of fluorine atoms in potential drug candidates.<sup>1</sup> The recent presence of fluorine in over 50% of the blockbuster drugs is suggestive of the growing prominence of organofluorine compounds in the pharmaceutical industry.<sup>2</sup> In fact, it has been recognized that the replacement of hydrogen with fluorine in a potential drug can enhance therapeutic efficiency and impart several beneficial physicochemical characteristics to the molecule by

modulating its  $pK_a$ , lipophilicity, binding affinity, *etc.*<sup>3,4</sup> Consequently, fluorine-containing drugs such as ciprofloxacin, Prozac<sup>®</sup>, Lipitor<sup>®</sup>, *etc.*, have found widespread use in the healthcare sector for the treatment of various ailments. Since the introduction of fluorine imparts several unique properties to the molecule, organofluorine compounds are also prevalent in the agrochemical industry, as evidenced by the fact that around 30% of all agrochemicals are fluorinated compounds.<sup>5</sup>

The widespread importance of fluorine in the pharmaceutical and agrochemical industries has provided a strong incentive to the synthetic community for the development of practical and efficient methods with the aim of deploying fluorine into molecules with greater simplicity. As a consequence, a variety of fluorinating reagents have been devised for the selective installation of fluorine in biologically relevant and material-specific molecules.<sup>6</sup> Amongst the various available reagents, the development of electrophilic fluorinating reagents, especially N–F reagents, has seen tremendous progress, potentially due to the ease with which they can be prepared and handled compared to nucleophilic fluorinating reagents. Additionally, their modular synthesis allows chemists to fine-tune the electronic and steric environment of the reagents to achieve the desired reactivity and selectivity in organic transformations. Consequently, huge numbers of electrophilic fluorinating reagents belonging to different classes

Department of Chemistry & Centre for Advanced Studies in Chemistry, Panjab University, Chandigarh 160014, India. E-mail: vsaini@pu.ac.in

† Electronic supplementary information (ESI) available. See DOI: <https://doi.org/10.1039/d2cp03281c>

have been developed, many of which are now commercially available. These reagents are known to fluorinate a wide range of nucleophiles, for example, activated enols and aromatics, stabilized carbanions, organometallic reagents, *etc.* One of the most popular N-F fluorinating reagents is Selectfluor<sup>TM</sup>, 25 tonnes of which are consumed each year for industrial and laboratory-scale syntheses.<sup>6–9</sup> Other commercially available and widely used N-F fluorinating reagents include *N*-fluoropyridinium salts and *N*-fluorobenzenesulfonimide (NFSI), which were developed by Umemoto<sup>10</sup> and Differding,<sup>11</sup> respectively.

The advent of a wide range of fluorinating reagents necessitates the need to understand the relative fluorination abilities of these reagents, which would not only assist organic chemists in selecting the right reagents for a synthetic transformation, but also promote the design of novel reagents. An initial attempt aimed at quantifying the fluorination ability of N-F fluorinating reagents was made by Gilicinski and co-workers in the year 1992.<sup>12</sup> Their scale was based on the experimental determination of the one-electron reduction potential ( $E_{\text{p}}^{\text{red}}$ ) of ten N-F fluorinating reagents, where molecules with a lower reduction potential have higher fluorination strength. The electrochemical scale was substantiated by correlating the reduction potential with literature yields of the reactions of these reagents with aromatic compounds. Several other groups have employed experimental kinetic approaches to obtain insight into the relative fluorination power of a handful of N-F reagents. As an example, Togni and co-workers ranked seven N-F reagents based on relative kinetic data for the competitive halogenation reaction of a  $\beta$ -keto ester with N-F reagents and *N*-chlorosuccinimide.<sup>13</sup> Similarly, Hodgson and co-workers developed a relative quantitative scale by studying the kinetics of ten electrophilic N-F reagents with a 1,3-dicarbonyl derivative.<sup>14</sup> The major limitation associated with these methods is the use of a limited number of reagents, since empirical determination of the reduction potential or kinetic profile of large number of molecules is a time-consuming and resource-intensive approach. Therefore, several theoretical scales have been proposed that aim to address some of the limitations imposed by experimental procedures. For example, Sudlow and Woolf reported a thermodynamic reactivity scale, and the computationally derived enthalpies were found to agree well with the empirical results.<sup>15</sup> However, this approach was restricted to only 13 closely related reagents, and a less-accurate semi-empirical (AM1) level of theory was used. Inspired by the quantitative scale first proposed by Christe and Dixon<sup>16</sup> for a few  $\text{XF}_n^+$  based fluorinating agents such as  $\text{KrF}^+$ ,  $\text{N}_2\text{F}^+$ ,  $\text{XeF}^+$ , *etc.*, Cheng *et al.* extended this approach to 130 N-F reagents in two of the most common solvents, dichloromethane (DCM) and acetonitrile (MeCN).<sup>6,17,18</sup> The scale is based on computed fluorine plus detachment (FPD) values, which are representative of the heterolytic bond dissociation energies of the N-F bond of the fluorinators. The FPD values were obtained using DFT at the M06-2X/6-311++G(2d,p)//M05-2X/6-31+G(d) level of theory with the inclusion of the SMD solvation model. DFT calculations using advanced methods, such as the one used in this study, can easily become time-limited, as these approaches typically scale

with a computation time complexity of  $O(N^4)$  depending on the molecular size ( $N$ ).<sup>19</sup> Therefore, depending on the system hardware, FPD calculations can take several hours or even days for a medium-sized molecule at this level of theory.

Although the scale based on FPD parameter is a comprehensive scale for determining the relative fluorination strength of a wide range of N-F fluorinating reagents, quantum-chemistry-assisted FPD calculations at an advanced level of theory using proprietary software are time-consuming and expensive. Therefore, there is a need for alternate approaches that would quantitatively predict the fluorination strength of the reagents at a faster pace and lower computational cost, and using open-source software.

Recently, machine learning (ML) methods have emerged as an attractive tool for predicting various chemical and physical properties of organic molecules, such as polarity,<sup>20</sup> solubility,<sup>21,22</sup>  $\text{p}K_{\text{a}}$ ,<sup>23</sup> electrophilicity,<sup>24</sup> nucleophilicity,<sup>25,26</sup> bond dissociation energies,<sup>27,28</sup> *etc.* In fact, the scientific community has started to view them as a potential alternative to expensive quantum chemical methods, as they can provide similar accuracy at a much lower computational cost. These techniques have gained popularity because of their ability to learn complex patterns and predict trends with high accuracy and at a faster pace. For example, a properly trained model can predict the electronic energies of molecules in a fraction of second, which is a significant advancement over traditional methods.<sup>29</sup> Considering the vast amount of data available for the dissociation energies (FPD) of a wide variety of N-F fluorinating reagents, we sought to introduce a ML model to predict FPD values by generating a novel dataset based on SMILES (Simplified Molecular Input Line Entry System) representations of molecules.<sup>30</sup>

## Results and discussion

### Dataset and descriptors

It should be noted that the advantages furnished by these ML approaches comes at the price of big data production, as the architecture requires vast amount of data in the form of descriptors and observations in order to train. In this context, 260 FPD values corresponding to 130 N-F fluorinating reagents in two solvents, namely, dichloromethane (DCM) and acetonitrile (MeCN), were collected from literature sources.<sup>17,18</sup> The N-F fluorinating reagents were categorized into six different categories, namely, *N*-fluorosulfonimides, *N*-fluorosulfonamides, *N*-fluorocarboxamides, *N*-fluoropyridiniums, *N*-fluoroheterocycles, and *N*-fluoroammoniums, as suggested by Cheng and co-workers in their seminal publication (Fig. 1A).<sup>17</sup> The *N*-fluoropyridinium salt category has the greatest number of unique fluorinating reagents with 40 molecules, followed by *N*-fluorosulfonimides and *N*-fluoroammoniums with 34 and 21 molecules, respectively (Fig. 1B). The remaining classes, *i.e.*, *N*-fluorosulfonamides, *N*-fluorocarboxamides, and *N*-fluoroheterocycles, contributed only 20, 10 and 5 molecules, respectively. The FPD values for these fluorinating reagents are typically in the range of 110.9 to 278.4 for MeCN and 112.3 to 290.4 for DCM.

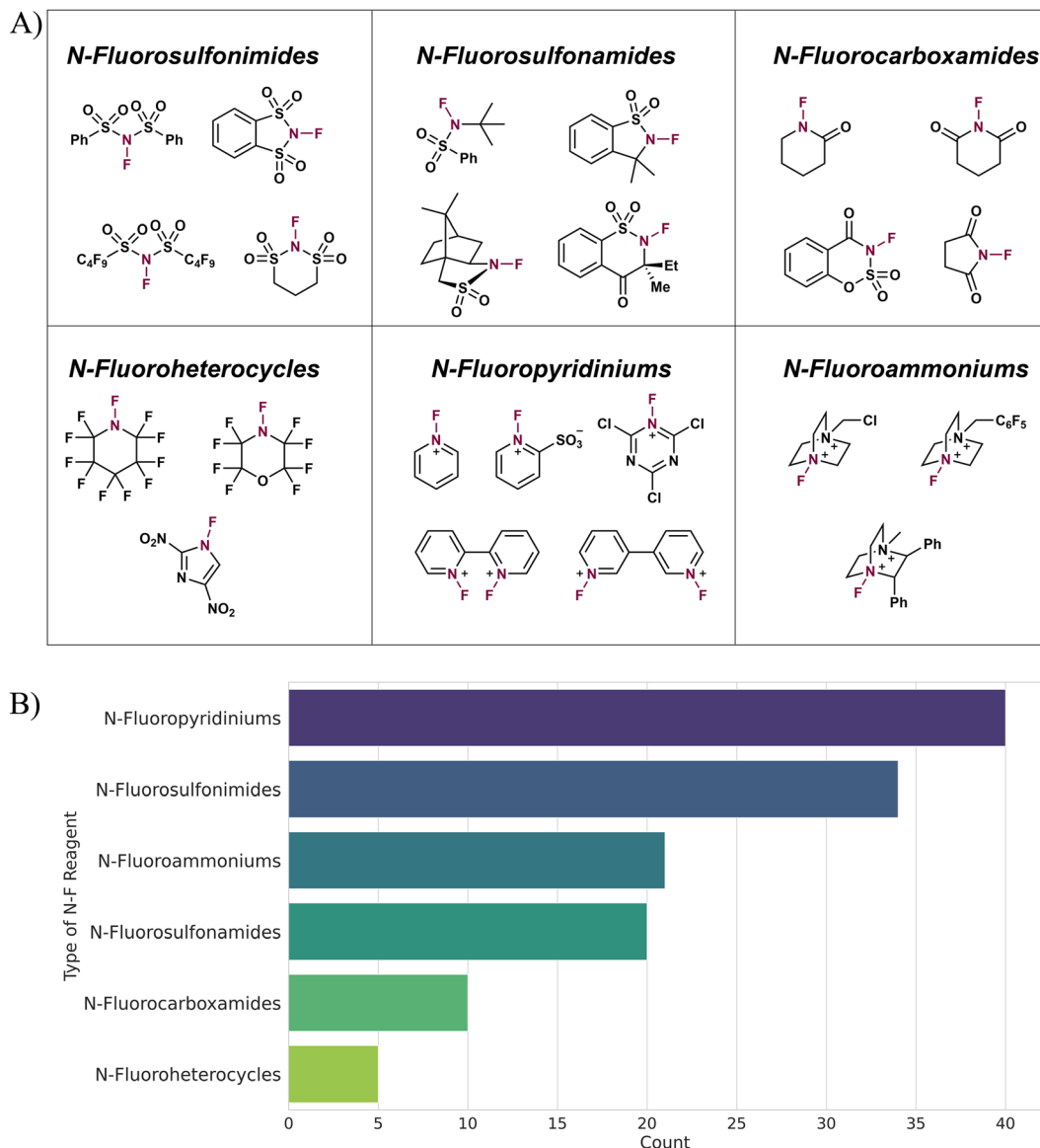


Fig. 1 (A) Representative examples of various N–F fluorinating reagents belonging to 6 categories. (B) Count plot representing the total number of N–F fluorinating reagents belonging to each class.

Generally, N-fluoroammoniums have a wider range of FPD values (110.9–256.2), whereas those of the other fluorinating reagents are narrowly distributed (Fig. S1, ESI†).

Next, the essential step is to encode the molecules in the form of features, for which molecular representation is of utmost importance. The use of atomic coordinates ( $x, y, z$ ) is one method of molecular representation and has been employed for predicting several molecular properties; however, their use is often limited due to the high computational cost.<sup>31,32</sup> Therefore, SMILES-based molecular representation was selected as it is a straightforward way of interpreting a molecule, as atoms, bonds, and branches can be denoted by their symbolic or grammatical specifications.<sup>30</sup> Recently, it has been used as a standard input method for calculating various descriptors in state-of-the-art cheminformatics packages at an encouragingly low computational cost.<sup>33,34</sup> Additionally, it has

been employed in several ML prediction studies, including studies related to *de novo* chemical design,<sup>35</sup> molecular property predictions,<sup>36</sup> screening compounds for drug discovery,<sup>37</sup> *etc.* Thus, in order to use a ML model for predicting the FPDs of a wide range of N–F fluorinating reagents and identify the various features that contribute most to the success of ML algorithm, SMILES-based descriptors were extracted and their quantitative relationship with the FPDs was evaluated. In this work, an open-source Mordred script based on the Python programming language was used for extracting 1613 descriptors.<sup>33</sup> Mordred is integrated with the RDKit platform,<sup>34</sup> another Python package for converting SMILES strings to molecular representations followed by feature extraction to generate various topological, 1D, and 2D descriptors. Therefore, a total of 1613 descriptors were calculated solely from SMILES representations of the 130 unique molecules within a matter of seconds, and were then used

for the predictive modelling study. Both “Type of Reagent” (TOR) and “Solvent” were added to the dataset as categorical descriptors, as the advanced ML algorithm allow us to encode these categorical descriptors into numerical ones *via* a process known as one-hot encoding.

### Feature preprocessing

With the raw dataset constituting 260 FPD values (130 in each solvent), 1613 numerical descriptors calculated from SMILES notations, and 2 categorical descriptors, feature engineering, which aims to decrease the model complexity and increase interpretability without eroding model efficiency and predictive power, was performed. The first step in this direction is eliminating descriptors with any missing values, which reduced the number of features to 1144. The variable reduction process was then followed, which reduces the number of noisy features, increases accuracy and learning efficiency, and simplifies the model. In this method, the correlation-based variable reduction process was employed, which removes redundancy by eliminating highly correlated features. The Pearson correlation coefficient ( $r$ ) was used to identify linearly correlated features, and one of the features with  $r \geq 0.80$  was removed, which left us with a truncated dataset consisting of 378 numerical descriptors. Furthermore, all the features with constant values were removed and the dataset constituting 175 numerical descriptors was subjected to a standardization procedure. Before model evaluation, one hot encoding was performed in order to include TOR and solvent as descriptors, which converted these categorical descriptors into discrete values (0 or 1 depending on the absence or presence of that category).<sup>38</sup>

Since the best algorithm cannot be found analytically, as the accuracy depends on various factors such as the type of problem, number of features and observations in the dataset, quality of the dataset, *etc.*, screening of the most common and popular algorithms must be carried out to find the best one.<sup>39</sup> Therefore, the most common and popular algorithms, each

with a different theoretical base, were evaluated for predicting the FPD values of N-F fluorinating reagents. The coefficient of determination ( $R^2$ ) and root mean square error (rmse) were employed, as they are the most commonly used metrics for ML-based regression studies.<sup>39</sup> The dataset was split, with 85% as the training set and 15% as the test set. For the splitting, two things were considered: (a) the data from different types of reagents should be proportional in both the training and test set (Table S1, ESI†). (b) In order to avoid redundancy, the combination of solvent pair FPD values for a particular molecule were kept in either the training or test set. For example, the FPD values of Selectfluor in both MeCN and DCM were kept in the training set, as otherwise it would have been easier for the model to predict the value in one of the solvents given another value in the training set, leading to a biased model. For model comparison purposes, cross-validation was performed using a five-fold cross validation procedure, which involved splitting the training set into five equally-sized non-overlapping sets.<sup>40</sup> In this approach, one of the sets is held out as a test set, while the remaining sets are used for training purposes. After five iterations, the metrics are computed for each configuration, and the average of these metrics are used for model comparison. In the end, the model trained on the whole training set was evaluated on the external test set (15%) to compare the predictive power of the model. A general workflow of the ML process followed in this study is given in Fig. 2A.

In order to visualize the data constituting 175 numeric descriptors along with the target values, the *t*-distributed stochastic neighbor embedding (*t*-SNE) technique was used which reduces the whole dataset into two dimensions (Fig. 2B).<sup>41</sup> It is an unsupervised ML algorithm that is based on keeping similar points together in a lower-dimensional space. Dimensionality reduction of the whole dataset into two components with respect to the TOR clearly shows clustering of data points. These observations point towards the fact that descriptors belonging to a particular type of fluorinating reagent are inherently analogous.

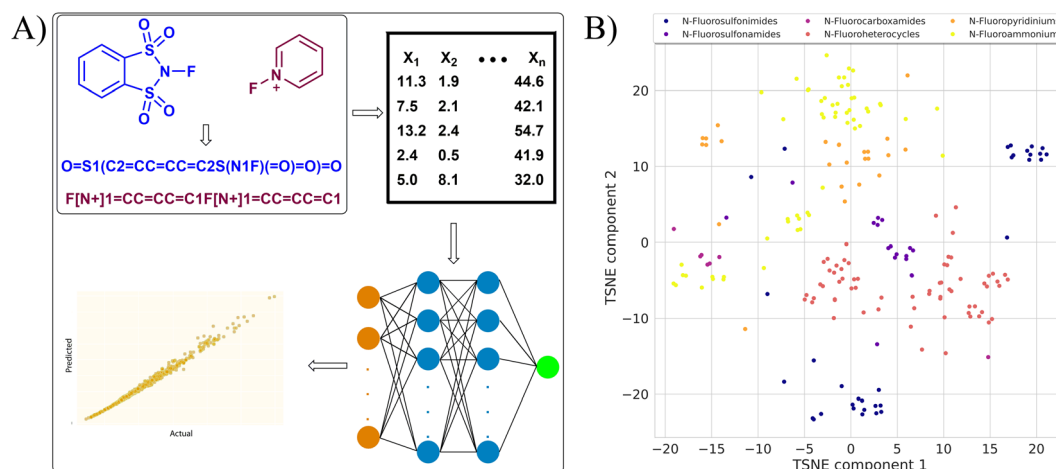


Fig. 2 (A) General workflow of the machine learning approach used in this study. (B) Visualization of the whole dataspace constituting 175 molecular descriptors and FPD values in a two-dimensional space using *t*-SNE.

**Table 1** Cross-validation metrics for various models screened in this study using 85 : 15 training : test split

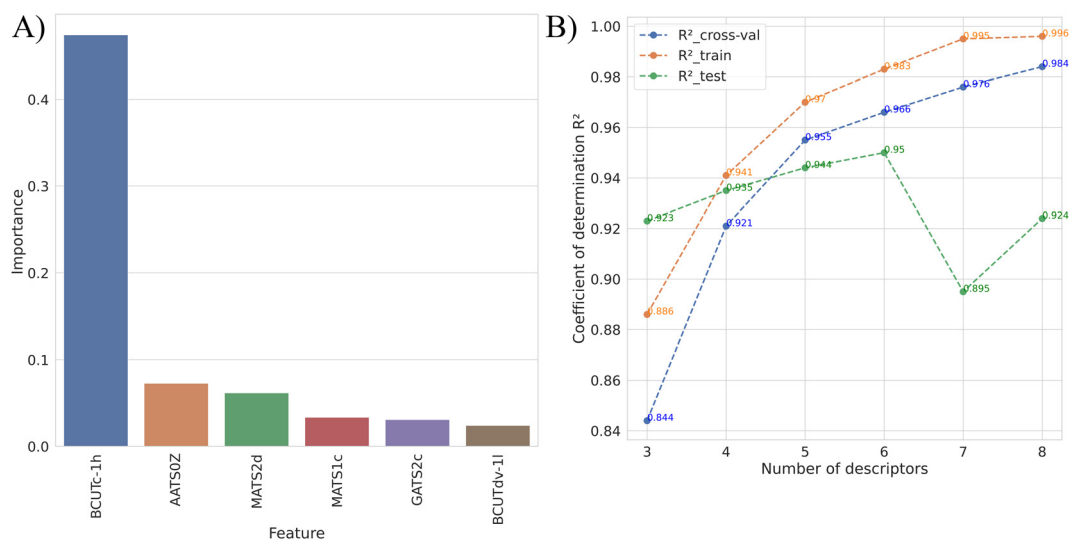
Entry	Model	$R^2$ (cross-validation)	RMSE
1	MLR	-1.382	—
2	PLS	0.719	12.65
3	KNN	0.722	12.47
4	SVM	0.769	11.34
5	AR	0.829	9.99
6	RF	0.932	6.16
7	NN	0.995	1.48

### Model evaluation

The ML algorithms screened in this study were MLR (multiple linear regression), PLS (partial least squares), KNN ( $k$ -nearest neighbor), SVM (support vector machine), AR (AdaBoost regressor), RF (random forest), and NN (neural network).<sup>39</sup> Cross-validation scores are established benchmarks for comparing model accuracy in the initial stages.<sup>40</sup> Therefore, the cross-validation  $R^2$  and rmse scores were compared for model screening. The negative  $R^2$  score observed for the MLR model clearly indicates that the model does not follow the trend of the data (Table 1).<sup>42</sup> PLS is an improved version of MLR and produces better accuracy than MLR, especially when the number of features is comparable to the number of observations. In our case, the PLS model gave significantly improved cross-validation  $R^2$  score of 0.719 over MLR; however, the accuracy was lower than desired. This is an indication that the FPD is not linearly dependent on the descriptors employed in this study and some state-of-the-art ML algorithms must be investigated. The main advantage that ML models have over classical linear models is their ability to handle multicollinearity and a large number of descriptors, which increases model performance at the expense of increasing model complexity. KNN and SVM are well-known cluster-based ML techniques and are appropriate for classification tasks; however, these were investigated in this study to test

their performance in the regression-based problem. These techniques were found to be slightly better than the linear approaches discussed above. In fact, SVM ( $R^2 = 0.769$ ) performed better than KNN ( $R^2 = 0.722$ ) in predicting the FPD values of N-F fluorinating reagents. Ensemble learning algorithms such as AR and RF are important ML algorithms that differ from each other in the manner in which the data is sampled for training purposes. AR and RF clearly gave better performance than the other ML model architectures evaluated in this study.<sup>43</sup> Out of the two tree-based models, RF outperformed AR with a significant increase in the cross-validation score ( $R^2 = 0.829$ , AR;  $R^2 = 0.932$ , RF).

In order to test the performance of deep learning approaches in our study, a NN algorithm was used for predicting the fluorination power of N-F reagents.<sup>44</sup> A multilayer perceptron (MLP)-based NN algorithm with three hidden layers was used.<sup>45</sup> The NN model constitutes an input layer with neurons equivalent to the number of features, followed by hidden layers with 64, 128 and 128 neurons, respectively, and finally an output layer with a single neuron that yields the FPD values in the last step. The hidden layers are equipped with a non-linear activation function, ReLU (rectified linear activation function), which allows the model to learn complex patterns in the dataset, while the output layer is equipped with a linear activation function. The NN algorithm trains the data using a technique known as backpropagation. Initially, random weights are assigned, and input data is propagated to the output layer *via* linear and non-linear combinations of weights. At the end, the loss function is calculated, which is improved over the course of various iterations by adjusting the random weights. This technique is known as backpropagation, and is the essence of NN training. Excitingly, the use of the NN architecture gave high model accuracy as suggested by the  $R^2$  cross-validation score of 0.995 (Table 1).



**Fig. 3** (A) Feature importance representing the contribution of each descriptor to the development of RF model in the decreasing order. (B) Plot describing  $R^2$  scores for the layered approach for the NN model, where the x axis represents the total number of descriptors (3 = 'TOR' + solvent + BCUTc-1h; 4 = 3 + BCUTdv-1l; 5 = 4 + GATS2c; 6 = 5 + MATS1c; 7 = 6 + AATS0Z; 8 = 7 + MATS2d).



Machine learning models have often been criticized for being a black box.<sup>46–48</sup> In fact, in our case, although the state-of-the-art NN model gave significantly enhanced performance compared to the simplistic and interpretable linear models such as MLR and PLS, the use of a large number of features imparts a black-box character to the model, leading to an inherently non-interpretable model. The question arises of whether the model can be simplified without impacting the performance or generalizability. This will lead to an interpretable and explainable model that is more acceptable in the scientific field because of the reluctance to rely on results that are difficult to understand *via* qualitative understanding. Therefore, in order to obtain insight into the features that contribute most to the training of the model, feature importance for the RF model was derived, which provides quantitative information regarding the contribution of each descriptor that is used for learning a specified prediction task. The top six most-contributing descriptors were identified to be BCUTc-1h, AATS0Z, MATS2d, MATS1c, GATS2c and BCUTdv-1l, as can be seen in Fig. 3A.

In order to test the impact of reduction of the number of numerical descriptors on the model performance based on the feature importance chart (Fig. 3A), RF and NN models were trained using the most-contributing features along with two categorical descriptors, *i.e.*, TOR and solvent. After model evaluation, the data was trained on the whole dataset, followed by estimation of the predictive power of the model on the external 15% test set. It should be noted that the RF-based feature importance algorithm is based on Gini impurity or variance reduction, which only displays the magnitude of the contribution that each descriptor makes.<sup>43</sup> In this approach, signs are usually ignored, *i.e.*, information regarding the positive or negative contribution of the descriptor cannot be obtained. Therefore, a layered approach was used in which each feature was added sequentially to the dataset and performance metrics were noted for the NN model. As is evident from Fig. 3B, the use of only three descriptors, *i.e.*, BCUTc-1h and two categorical descriptors, gave lower cross-validation and training  $R^2$  scores. The addition of BCUTdv-1l led to a significant increase in the cross-validation and training  $R^2$  scores, and a slight increase in the test  $R^2$  score, indicating the positive influence of the descriptor on the output values. GATS2c and MATS1c are another set of descriptors that contributed positively to the model training, as enhanced cross-validation, training and test  $R^2$  scores were observed. However, the addition of AATS0Z and MATS2d to the dataset negatively impacted the test scores, and the  $R^2$  scores plummeted to 0.909 and 0.927, respectively, for the NN model. A similar trend was observed in the RF model, and complete evaluation metrics are described in Table S2 and Fig. S5 (ESI<sup>†</sup>). Representative metrics for the RF and NN models are shown in Table 2. Clearly, reduction in the number of descriptors from 177 to 8 did not significantly decrease the model accuracy or predictive power for either the RF or NN models. For example, the cross-validation  $R^2$  score of RF for the eight-descriptor-based model is 0.933 with a rmse of 6.27 along with a high training  $R^2$  score of

Table 2 Model evaluation metrics for RF and NN models using different number of descriptors

Model	No. of features	$R^2$ (cross-validation)	RMSE (cross-validation)	$R^2$ (training)	$R^2$ (test)	RMSE (test)
RF	175 + 2	0.932	6.16	0.991	0.796	9.44
	6 + 2	0.933	6.27	0.992	0.787	9.7
	4 + 2	0.916	7.03	0.992	0.915	6.12
NN	175 + 2	0.995	1.48	0.998	0.892	6.87
	6 + 2	0.990	2.28	0.997	0.927	5.70
	4 + 2	0.963	4.04	0.983	0.967	3.84

0.992 and a low test  $R^2$  score of 0.787, which are comparable to those of the 177-descriptor-based model. Similarly, the evaluation of the NN model using eight descriptors gave comparable cross-validation, training and test scores to the 177-descriptor based model. Moreover, fine-tuning of the model with further reduction in the number of descriptors to six led to an increase in the predictive power and generalizability of both the models. Overall, the six-descriptor based NN model gave better predictive power compared to the RF model as suggested by their test  $R^2$  scores of 0.967 (rmse = 3.84) and 0.915 (rmse = 6.12), respectively (Table 2 and Fig. 4).

### Predictions

A comparison of the top five and bottom five predicted test set values with the actual FPD values using the NN model is shown in Fig. 5. The NN model accurately predicts the fluorination strength of N-F fluorinating reagents belonging to the different categories, as shown in Fig. 5 and Table S3 (ESI<sup>†</sup>). As an example, the FPD values of 1-fluoro-4-nitropyridinium and 1-fluoroquinuclidine-1-ium in DCM (entries 1 and 2) can be predicted with an absolute error equal to 0.1. In fact, 28 predictions out of 40 can be predicted with an absolute error less than or equal to 3.0 (Table S3, ESI<sup>†</sup>). Out of the six categories of N-F fluorinating reagents utilized in this study, *N*-fluorosulfonimides and *N*-fluoropyridiniums were the hardest to predict. For example, *p*-bromophenyl- and *p*-(*t*-butylphenyl)-substituted *N*-fluorosulfonamide derivatives gave absolute errors of 4.2 and 4.6, respectively, in DCM solvent (entries 7 and 8). Similarly, *N*-fluoropyridinium derivatives, namely, 3-chloro-1-fluoro-5-(trifluoromethyl)pyridinium-2-sulfonate (entry 9) and 2,6-dicyano-1-fluoropyridinium (entry 10), gave relatively high absolute errors of 6.3 and 12.5, respectively, in DCM solvent. Out of the two solvents used, predicted FPD values gave a slightly lower average absolute error of 2.3 in DCM than in MeCN (absolute error = 2.6). Overall, the NN architecture gave a strong predictive model for predicting the FPD values of a wide range of N-F fluorinating reagents with an average absolute error of 2.6 for all 40 test set values.

### Interpretability

Out of the 175 relevant SMILES-based descriptors, the four molecular descriptors used to furnish a NN model were BCUTc-1h, BCUTdv-1l, GATS2c and MATS1c. BCUTc-1h and BCUTdv-1l are eigenvalue-based descriptors as described by Pearlman *et al.*<sup>49,50</sup>

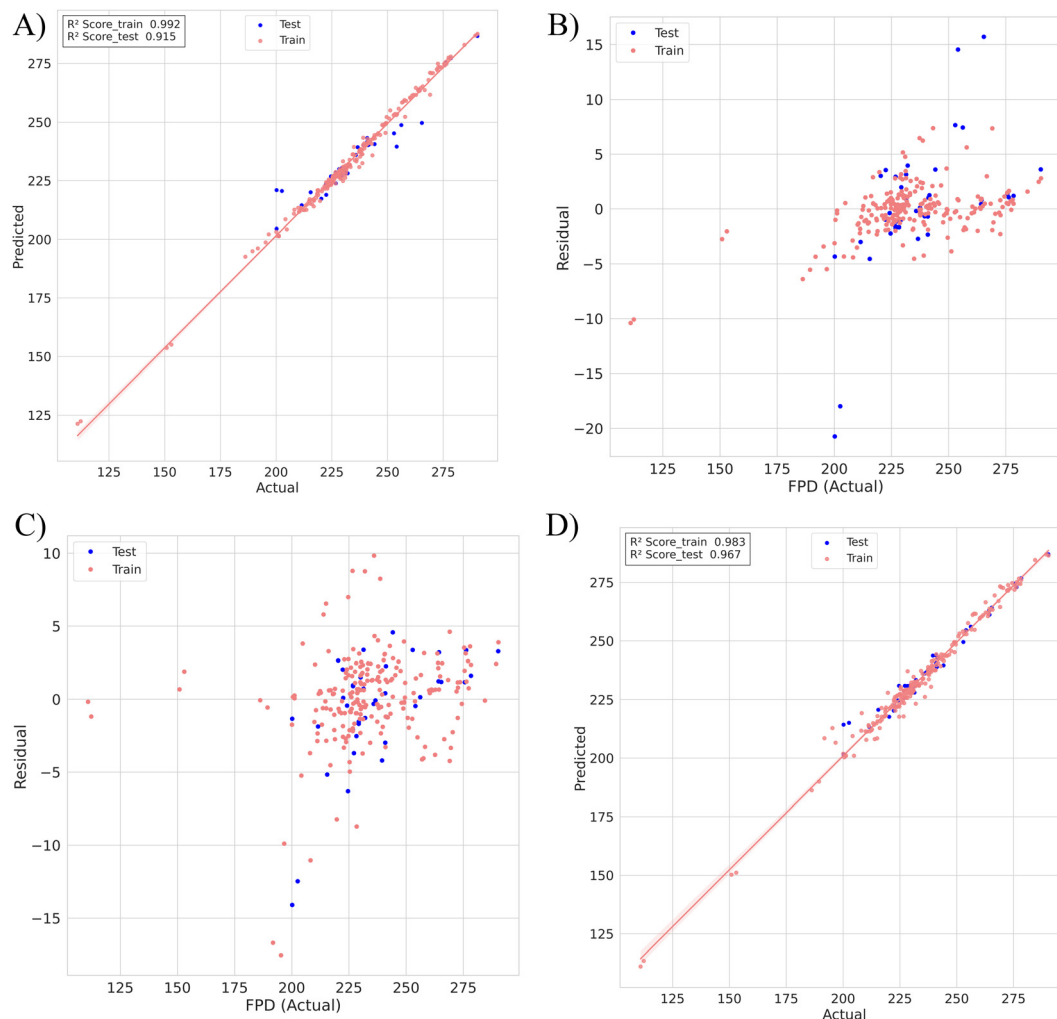


Fig. 4 (A) Regression plot for the RF model. (B) Residual plot showing actual FPD values and residual (actual – predicted) values for the RF model. (C) Regression plot for the NN model. (D) Residual plot showing actual FPD values and residual (actual – predicted) values for the NN model.

These descriptors take into account the connectivity and atomic properties of the molecule, such as atomic weight, partial charges and polarizability. As shown in Fig. 6A, the Pearson correlation matrix of these four descriptors with the FPD values reveals significant correlation of BCUTc-1h descriptor with the FPD values ( $r = -0.74$ ). The negative correlation between the FPD values and BCUTc-1h descriptor is also evident from the scatter plot (Fig. 6B). A similar trend was observed in the feature importance chart, where BCUTc-1h was the most-contributing feature to the RF model development. This is not surprising, as the BCUTc-1h and BCUTdv-1l descriptors take into account polarizability, which is strongly related to the charge distribution of the atoms or molecules. In fact, it is an established fact that the polarizability of molecules directly impacts the bond dissociation energies of the participating atoms.<sup>51</sup> In order to derive chemical intuition from the relevant descriptors used for this approach and to describe how changes in molecular structure impact these descriptor values and hence the reactivity, five electronically varied reagents from the *N*-

fluoropyridinium group were analysed. As shown in Fig. 7, the FPD values decrease moving from electron-releasing to electron-withdrawing substituents on the pyridine ring, leading to an increase in the reactivity. This can be attributed to the decreased electron density on the nitrogen atom of the pyridine ring with the introduction of an electron-withdrawing group such as a trifluoromethyl group, leading to a more polarized N–F bond and hence increased reactivity. Analysis of the BCUTc-1h descriptor, which is the most-contributing feature for our model, reveals a positive correlation with the reactivity, as described before, and can provide first-hand information regarding the reactivity of new molecules. To substantiate this, a hypothetical fluoropyridine molecule containing three trifluoromethyl groups was fed to the model, and FPD prediction was obtained. As is evident from the data in Fig. 7, the BCUTc-1h value is higher because of the presence of highly electron-withdrawing groups, indicating higher reactivity and a lower FPD value (209.6 in MeCN), especially in comparison to the molecule containing only two trifluoromethyl groups.

Entry	N-F Fluorinating Reagent	Actual	Predicted	Absolute Error
1.		222.5 (220.3)	222.4 (217.7)	0.1 (2.6)
2.		256.2 (252.9)	256.1 (249.5)	0.1 (3.4)
3.		235.8 (224.2)	236.1 (224.6)	0.3 (0.4)
4.		237.5 (226.7)	237.9 (228.8)	0.4 (2.1)
5.		240.9 (236.6)	240.5 (236.7)	0.4 (0.1)
6.		276.3 (264.4)	273.0 (261.2)	3.3 (3.2)
7.		239.5 (227.2)	243.7 (230.9)	4.2 (3.7)
8.		244.2 (231.4)	239.6 (228.0)	4.6 (3.4)
9.		224.6 (215.5)	230.9 (220.7)	6.3 (5.2)
10.		202.6 (200.2)	215.1 (214.3)	12.5 (14.1)

Fig. 5 Actual and predicted FPD values along with the absolute error observed for various N-F fluorinating reagents for the NN model. Note: The values are in DCM solvent, and the number in parentheses is the value in MeCN.

The other eigenvalue-based descriptor, namely, BCUTdv-11, does not show any notable correlation with the FPD values, as

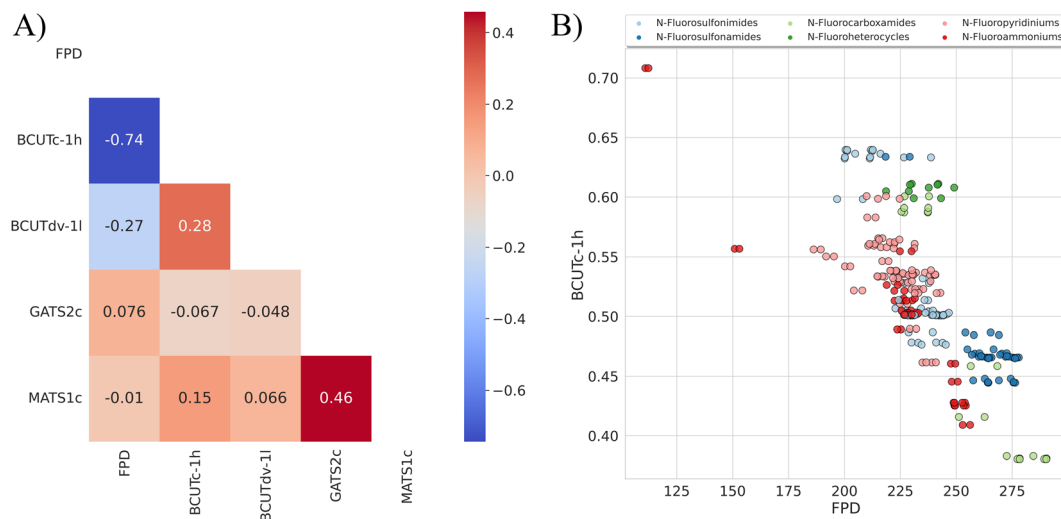
suggested by the scatter plot between them (Fig. S2, ESI†). The values are roughly constant for the reagents in each group of fluorinators. The other two descriptors, namely, GATS2c and MATS1c, which have a minor contribution to the model development and significantly low correlation coefficient with the FPD values, are autocorrelation descriptors.<sup>52</sup> In these descriptors, the atom pair distances are weighed by Gasteiger charges. Since change in the charge distribution with distance impacts the electronegativity of the atoms, their influence on the bond dissociation energies, which are representatives of the FPD values in this study, cannot be neglected and presumably plays a minor role in predicting the FPD values.<sup>53</sup> The scatter plots shown in Fig. S3 and S4 (ESI†) also show random distribution of values, signalling less weightage in modelling the FPD parameter. Their values for the five *N*-fluoropyridinium salts depicted in Fig. 7 also point towards their minor but complex role, as no particular trend is observed with changes in the groups on the molecules.

### Applicability domain analysis

In order to identify the outliers in the training set and define the reagents residing outside the applicability domain (AD) in the test set, a universal and highly reliable basic theory of standardization was applied.<sup>54</sup> Outliers are generally a small fraction of compounds in the training set whose feature values are drastically different from those of the rest of the data points. If the molecules in the test set are similar to these outliers, the predictions would be below par, as the model is unable to completely capture the features of that small fraction of training set compounds. Therefore, their identification is important, especially in traditional quantitative structure activity relationship studies (QSAR) using multiple linear regression.

From careful analysis of Fig. 1B, one can easily presume that the feature space of *N*-fluoroheterocycles, the class with the smallest number of compounds (four in the training set and one in the test set) would be different from the rest of the data points. To substantiate this point in a statistical manner, the “Applicability Domain (using standardization approach) v1.0” software developed by Roy and co-workers was employed.<sup>54</sup> This tool identifies compounds whose descriptor values lie outside the range of mean  $\pm$  3 standard deviation (SD) and labels them as an outlier (training set) or outside AD (test set). As is evident from Fig. 8, all the 5 *N*-fluoroheterocycles are either above the threshold or at the borderline of the generally accepted range of mean  $\pm$  3SD. For example, compounds **D1a** in the test set and **D1b** in the training set cross the threshold value, and can be considered outside AD and an outlier, respectively. However, it is surprising that despite being outside AD, the NN model efficiently predicts the FPD value of **D1a** with an absolute error of 1.5 in MeCN, which is quite less than the average value of 2.6. The remarkable prediction accuracy for this compound shows the importance of an advanced ML algorithm such as NN, which can easily handle these anomalies in the test set with only few analogous molecules in training. The other three compounds belonging to the same group, **D1c**, **D2** and **D3**, are at the borderline and can be considered distinct





**Fig. 6** (A) Correlation matrix representing the Pearson correlation coefficient for the four descriptors that made it to the final NN model, along with the FPD values. (B) Scatter plot between the FPD values and BCUTc-1h, which is the most-contributing descriptor, colour coded with respect to the type of reagents.

Reagents						
Trend	Increasing Reactivity					Hypothetical Molecule
FPD	236.6	231.7	230.0	223.6	215.1	209.6
BCUTc-1h *10 <sup>3</sup>	461.617	522.975	522.790	536.336	565.602	611.657
BCUTdv-1l *10 <sup>3</sup>	995.030	995.950	2734.427	2812.403	2917.950	2936.052
GATS-2c	1.533	0.978	1.257	0.552	0.520	0.803
MATS-2c	-0.131	0.268	0.036	-0.717	-0.618	-0.495

**Fig. 7** Representative examples of differentially substituted *N*-fluoropyridinium salts along with the FPD values and descriptors used for model building. Note: The BCUTc-1h and BCUTdv-1l descriptor values are multiplied by 10<sup>3</sup> for better readability. FPD values are in MeCN.

from rest of the data points in the training set. Additionally, dinitrogen fluoride **F7** with a unique structure and the lowest FPD value also lies just below the threshold value of 3.0 and can be considered as a borderline outlier. Surprisingly, ammonium fluoride **F8**, which also has a dissimilar molecular structure compared to the rest of the members of the group (*N*-fluoroammoniums) has a SD ( $S_{\text{new}}$ ) value of 2.64, which categorizes it as a normal member of the group.

## Conclusion

In this study, we curated a novel dataset consisting of 260 FPD values of 130 fluorinating reagents in two solvents, namely, DCM and MeCN, and a total of 1613 descriptors generated from SMILES notations of the molecules with the aid of RDKit-based

Mordred script. Out of the various models screened in this study, NN was found to be the best performing model. Tree-based model architecture such as RF also gave high accuracy; however, slight overfitting was observed as suggested by wide difference in the training and test  $R^2$  scores, which were 0.992 and 0.915, respectively. An RF-based feature importance algorithm was used to identify the four molecular descriptors that successfully contributed to the development of an NN-based statistically robust model with high predictive power, as evidenced by the test  $R^2$  score of 0.967 and rmse of 3.84. Overall, the work presented in this study lays an important foundation for the rational design and development of novel N-F reagents and would aid synthetic chemists in their efforts for future evolution of refined, simple and safe synthetic methods for installing fluorine atoms in organic molecules.

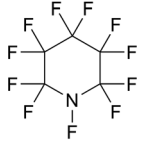
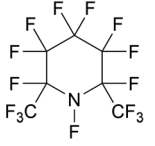
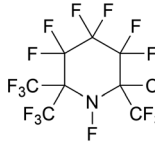
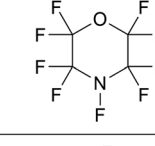
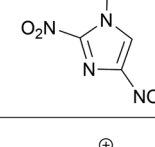
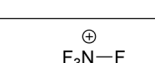
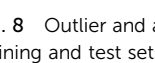
Compound	S <sub>new</sub>	Remark	FPD / Abs. error
 <b>D1a</b>	3.014958	Outside AD	229.3 / 1.5
 <b>D1b</b>	3.00534	Outlier	230.1 / 1.0
 <b>D1c</b>	2.98298	Borderline outlier	237.7 / 4.0
 <b>D2</b>	2.97444	Borderline outlier	231.0 / 3.7
 <b>D3</b>	2.99137	Borderline outlier	218.6 / 1.1
 <b>F7</b>	2.95496	Borderline outlier	110.9 / 0.2
 <b>F8</b>	2.64223	-	150.9 / 0.7

Fig. 8 Outlier and applicability domain analysis of the compounds in the training and test sets. The FPD and absolute error values are in MeCN.

## Methods

The FPD values were obtained from literature sources.<sup>17</sup> All the molecular descriptors were calculated from Mordred script.<sup>33</sup> For machine learning, the Python 3 framework equipped with various libraries, such as Pandas, Numpy, Scikit-learn, Keras, TensorFlow, Matplotlib, and Seaborn, was used. Applicability domain analysis was performed using the “Applicability Domain (using standardization approach) v1.0” software developed by Roy and co-workers.<sup>54</sup>

## Data availability

The datasets and model algorithms can be accessed from this link: <https://github.com/v-saini/Fluorination-Power.git>.

## Conflicts of interest

The authors declare no conflict of interest.

## Acknowledgements

V. S. thanks the Department of Science and Technology, Ministry of Science and Technology, India for a DST-Inspire Faculty grant (DST/INSPIRE/04/2017/002529). V. S. acknowledges the Department of Chemistry, Panjab University, for providing useful resources.

## References

- 1 J. Wang, M. Sánchez-Roselló, J. L. Aceña, C. del Pozo, A. E. Sorochinsky, S. Fustero, V. A. Soloshonok and H. Liu, *Chem. Rev.*, 2014, **114**, 2432–2506.
- 2 H. Mei, J. Han, S. Fustero, M. Medio-Simon, D. M. Sedgwick, C. Santi, R. Ruzziconi and V. A. Soloshonok, *Chem. – Eur. J.*, 2019, **25**, 11797–11819.
- 3 S. Purser, P. R. Moore, S. Swallow and V. Gouverneur, *Chem. Soc. Rev.*, 2008, **37**, 320–330.
- 4 E. P. Gillis, K. J. Eastman, M. D. Hill, D. J. Donnelly and N. A. Meanwell, *J. Med. Chem.*, 2015, **58**, 8315–8359.
- 5 T. Furuya, A. S. Kamlet and T. Ritter, *Nature*, 2011, **473**, 470–477.
- 6 N. Rozatian and D. R. W. Hodgson, *Chem. Commun.*, 2021, **57**, 683–712.
- 7 R. E. Banks, S. N. Mohialdin-Khaffaf, G. S. Lal, I. Sharif and R. G. Syvret, *J. Chem. Soc., Chem. Commun.*, 1992, 595–596.
- 8 S. Stavber, M. Zupan, A. J. Poss and G. A. Shia, *Tetrahedron Lett.*, 1995, **36**, 6769–6772.
- 9 P. T. Nyffeler, S. G. Durón, M. D. Burkart, S. P. Vincent and C.-H. Wong, *Angew. Chem., Int. Ed.*, 2005, **44**, 192–212.
- 10 T. Umemoto, K. Kawada and K. Tomita, *Tetrahedron Lett.*, 1986, **27**, 4465–4468.
- 11 E. Differding and H. Ofner, *Synlett*, 1991, 187–189.
- 12 A. G. Gilcinski, G. P. Pez, R. G. Syvret and G. S. Lal, *J. Fluorine Chem.*, 1992, **59**, 157–162.
- 13 P. Y. Toullec, I. Devillers, R. Frantz and A. Togni, *Helv. Chim. Acta*, 2004, **87**, 2706–2711.
- 14 N. Rozatian, I. W. Ashworth, G. Sandford and D. R. W. Hodgson, *Chem. Sci.*, 2018, **9**, 8692–8702.
- 15 K. Sudlow and A. A. Woolf, *J. Fluorine Chem.*, 1994, **66**, 9–11.
- 16 K. O. Christe and D. A. Dixon, *J. Am. Chem. Soc.*, 1992, **114**, 2978–2985.
- 17 X.-S. Xue, Y. Wang, M. Li and J.-P. Cheng, *J. Org. Chem.*, 2016, **81**, 4280–4289.
- 18 M. Li, H. Zheng, X.-S. Xue and J.-P. Cheng, *Tetrahedron Lett.*, 2018, **59**, 1278–1285.
- 19 F. Jensen, *An Introduction to Computational Chemistry*, 1989.
- 20 V. Saini and R. Kumar, *New J. Chem.*, 2022, **46**, 16981–16989.
- 21 D. S. Palmer, N. M. O’Boyle, R. C. Glen and J. B. O. Mitchell, *J. Chem. Inf. Model.*, 2007, **47**, 150–158.
- 22 S. Boobier, D. R. J. Hose, A. J. Blacker and B. N. Nguyen, *Nat. Commun.*, 2020, **11**, 5753.
- 23 R. Roszak, W. Beker, K. Molga and B. A. Grzybowski, *J. Am. Chem. Soc.*, 2019, **141**, 17142–17149.
- 24 G. Hoffmann, M. Balcilar, V. Tognetti, P. Héroux, B. Gaüzère, S. Adam and L. Joubert, *J. Comput. Chem.*, 2020, **41**, 2124–2136.

- 25 V. Saini, A. Sharma and D. Nivatia, *Phys. Chem. Chem. Phys.*, 2022, **24**, 1821–1829.
- 26 S. Boobier, Y. Liu, K. Sharma, D. R. J. Hose, A. J. Blacker, N. Kapur and B. N. Nguyen, *J. Chem. Inf. Model.*, 2021, **61**, 4890–4899.
- 27 P. C. St. John, Y. Guan, Y. Kim, S. Kim and R. S. Paton, *Nat. Commun.*, 2020, **11**, 2328.
- 28 M. Wen, S. M. Blau, E. W. C. Spotte-Smith, S. Dwaraknath and K. A. Persson, *Chem. Sci.*, 2021, **12**, 1858–1868.
- 29 F. A. Faber, A. Lindmaa, O. A. von Lilienfeld and R. Armiento, *Phys. Rev. Lett.*, 2016, **117**, 135502.
- 30 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
- 31 K. T. Schütt, H. E. Saucedo, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, *J. Chem. Phys.*, 2018, **148**, 241722.
- 32 K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller and A. Tkatchenko, *Nat. Commun.*, 2017, **8**, 13890.
- 33 H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, *J. Cheminf.*, 2018, **10**, 4.
- 34 G. Landrum, <http://www.rdkit.org/>, <https://github.com/rdkit/rdkit>, 2016.
- 35 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268–276.
- 36 G. A. Pinheiro, J. Mucelini, M. D. Soares, R. C. Prati, J. L. F. Da Silva and M. G. Quiles, *J. Phys. Chem. A*, 2020, **124**, 9854–9866.
- 37 M. Hirohara, Y. Saito, Y. Koda, K. Sato and Y. Sakakibara, *BMC Bioinf.*, 2018, **19**, 526.
- 38 J. M. Granda, L. Donina, V. Dragone, D.-L. Long and L. Cronin, *Nature*, 2018, **559**, 377–381.
- 39 J. B. O. Mitchell, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2014, **4**, 468–481.
- 40 J. Lever, M. Krzywinski and N. Altman, *Nat. Methods*, 2016, **13**, 703–704.
- 41 L. van der Maaten and G. Hinton, *J. Mach. Learn. Res.*, 2008, **9**, 2579–2605.
- 42 M. Krzywinski and N. Altman, *Nat. Methods*, 2015, **12**, 1103–1104.
- 43 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
- 44 S. Ye, K. Zhong, J. Zhang, W. Hu, J. D. Hirst, G. Zhang, S. Mukamel and J. Jiang, *J. Am. Chem. Soc.*, 2020, **142**, 19071–19077.
- 45 A. A. Kananenka, K. Yao, S. A. Corcelli and J. L. Skinner, *J. Chem. Theory Comput.*, 2019, **15**, 6850–6858.
- 46 R. Rodríguez-Pérez and J. Bajorath, *J. Med. Chem.*, 2021, **64**, 17744–17752.
- 47 R. Dybowski, *New J. Chem.*, 2020, **44**, 20914–20920.
- 48 C. Rudin, *Nat. Mach. Intell.*, 2019, **1**, 206–215.
- 49 R. S. Pearlman and K. M. Smith, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 28–35.
- 50 D. T. Stanton, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 11–20.
- 51 U. Hohm, *J. Chem. Phys.*, 1994, **101**, 6362–6364.
- 52 G. Sliwoski, J. Mendenhall and J. Meiler, *J. Comput.-Aided Mol. Des.*, 2016, **30**, 209–217.
- 53 R. T. Sanderson, *J. Am. Chem. Soc.*, 1975, **97**, 1367–1372.
- 54 K. Roy, S. Kar and P. Ambure, *Chemom. Intell. Lab. Syst.*, 2015, **145**, 22–29.