Seeking the Optimal Descriptor for S_N2 reactions through Statistical Analysis of Density Functional Theory (DFT) Results

*Lucía Morán-González[a], Maria Besora[b], and Feliu Maseras[a],**

[a] Institute of Chemical Research of Catalonia (ICIQ), The Barcelona Institute of Science and Technology, Avgda. Països Catalans, 16, 43007 Tarragona, Catalonia, Spain

[b] Departament de Química Física i Inorgánica, Universitat Rovira i Virgili, c/Marcel·lí Domingo s/n, 43007 Tarragona, Catalonia, Spain

The bimolecular nucleophilic substitution is one of the fundamental reactions in organic chemistry, yet there is still knowledge to be gained on the role of the nucleophile and the substrate. A statistical treatment of over five hundred DFT-computed barriers for bimolecular nucleophilic substitution at methyl derivatives (S_N2@C) leads to the identification of the numerical descriptors that best represent the entering and leaving ability of 23 different nucleophiles. The treatment is

based on a Singular Value Decomposition (SVD) of a matrix of computed energy barriers. The current work represents the extension to a problem of reactivity of the hidden descriptor methodology that we had previously developed for the thermodynamic problem of bond dissociation energies in transition metal complexes. Analysis of the results shows that a single descriptor is sufficient. Remarkably, this hidden descriptor has different values for entering and leaving abilities and, contrary to expectation, does not correlate especially well with frontier molecular orbitals descriptors. In contrast, it correlates with other thermodynamic and geometric parameters. This statistical procedure can be in principle extended to additional chemical fragments and other reactions.

Introduction

Machine learning,[1,2] big data,[3,4] deep learning[5,6] are becoming household terms in modern chemistry, and for good reason. The combination of data-led approaches and statistics-based methods is being applied with increasing success to a variety of chemical problems such as prediction of homolytic bond dissociation enthalpies (BDE),[7] estimation of barriers for nucleophilic aromatic substitution (S_NAr),[8] and the optimization of selectivity through improved

design of phosphoric acid derivative catalysts for the coupling of imines and thiols.[9] Our group has also proposed a model to estimate the nucleophilicity of the C-H bond of alkanes from the values of a set of six topological descriptors.[10]

A variety of mathematical approaches are being applied for the treatment of data, from linear regression algorithms[11–13] and principal component analysis (PCA)[14] to methods closer to artificial intelligence such as neural networks[15] or random forest treatments.[16,17] Practically all of the treatments rely on the use of descriptors. The basic idea is to discover a mathematical relationship, between the property of interest and the descriptors, and use it to predict the value of the property from those of the descriptors, which are in principle easier to measure. What is a descriptor? A descriptor is a mathematical object that encodes chemical information of a compound. And a wide array of descriptors has been used, from geometrical parameters[18] to electronic features[19]. The underlying assumption is often that if a sufficiently large number of descriptor sets is tested, there will be one, or a combination of them, which gives a good reproduction of a desired property.

We decided a few years ago to attempt a different twist to the problem, and we applied the statistical treatment itself to the identification of the optimal descriptors through the analysis of a relatively large number of computed DFT results. In this way, the hidden descriptor (HD) method

was developed as a tool to find the best possible descriptors for a given problem,[20] as opposed to the standard sets of conventional descriptors that are usually applied. This technique, based on a Singular Value Decomposition (SVD), was applied to the heterolytic bond dissociation energy (BDE) of ligands from transition metal complexes. The strategy resulted in the discovery that the binding properties of a given metal fragment or ligand can be described by a vector of five components. This treatment allowed us a better understanding of the thermodynamics of inorganic compounds in solution.

The same HD methodology is transferred to this work, but now we go one step forward analyzing reactivity through the energy barriers ($\Delta G^{\ddagger}$) of an organic reaction, bimoleclar nucleophilic substitution at sp^3 carbon centers ($S_N2@C$). This is one of the best known elementary reactions in organic chemistry, and its mechanism, summarized in Figure 1, is well understood. A lone pair of the entering group (EG) interacts with the empty σ^* antibonding orbital of the leaving group (LG), which is polarized toward the carbon. This leads to an inversion of configuration at the central carbon. The profile of the reaction may have pre- or post-transition state adducts (Add_1, Add_2) depending on the nature of EG, LG and the solvent, as shown in purple in Figure 1. This mechanistic view has been in part shaped by computational chemistry, as the S_N2 reaction was the topic of early

applications of Hartree-Fock based methods[21,22] and the benchmarking of density functional theory (DFT) application to organic reactivity.[23,24] Specific aspects of the S_N2 reaction have remained the subject of computational studies throughout the years.[25–28] In this context, the contribution from Bickelhaupt and co-workers has led to a better fundamental understanding of the process through the application different techniques, such as the activation strain model, to the analysis of trends involving different nucleophiles, substituents and central atoms.[29]

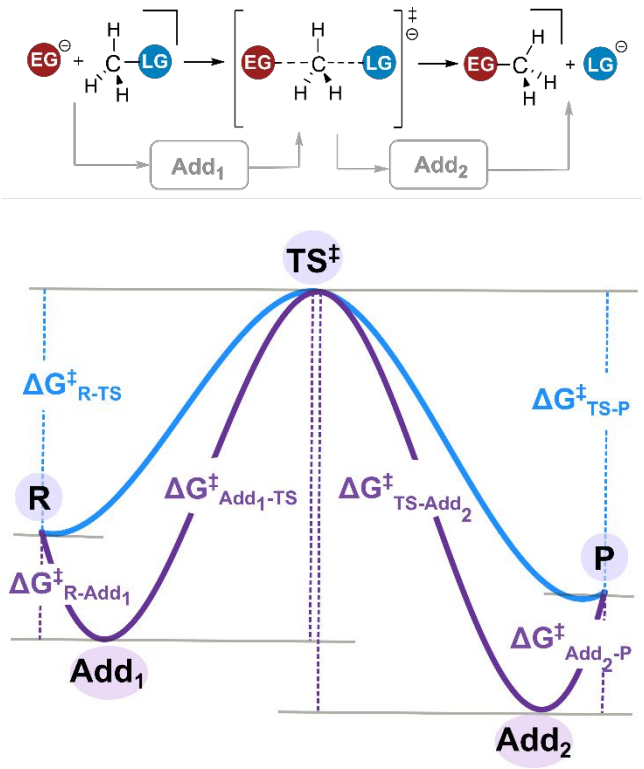

Figure 1. Top: general reaction scheme for S_N2 reactions. EG is the entering group, LG the leaving group, and Add_1 and Add_2 the reactant and product adducts, respectively. Bottom: Free energy

profile of a typical $S_N2@C$ in polar solvent (blue line), and in apolar solvent (purple line), where R are the reactants, TS the transition state, and P the product.

The current contribution intends to complement the previous work on fundamental aspects of the $S_N2@C$ reaction by analyzing trends within a wider range of nucleophiles than what had previously explored. We will accomplish it through the application of a statistical method, the hidden descriptor methodology, to reaction. For that purpose, we computed a set of 529 energy barriers of the $S_N2@C$ reactions in two different solvents: dichloromethane and water. Furthermore, we searched correlations between the derived hidden descriptors and conventional chemical descriptors.

Method for DFT calculations

DFT calculations were performed using the Gaussian 16 program.[30] The geometries were optimized with B3LYP-D3,[31,32] where empirical dispersion was added by means of Grimme dispersion correction.[33] The 6-311+G(d)[34–36] was used for all the elements except for Br and I, where an ECP[37] together with the LANL2DZdp[38] basis set was employed. Solvent effects were introduced using the continuous solvent model SMD[39] for water and dichloromethane. Harmonic

vibrational frequencies were computed on optimized geometries with the default temperature (298.15K) and 1 atm of pressure. All the reported energies correspond to the free energies in solution and in kcal mol^{-1}. A data set collection of computational results is available in the ioChem-BD repository.[40]

Hidden descriptor method

Figure 2 depicts the procedure that we followed as a data workflow. It consists of these steps: (1) design of the data set of reactions to compute, (2) geometry optimization and frequency calculations for every reaction, (3) construction of the energy barrier matrix, (4) execution of a Python code to carry out a SVD, (5) validation of the procedure and selection of the optimal number of descriptors, (6) qualitative analysis of the hidden descriptor results, and (7) correlation between the hidden descriptors and a set of conventional chemical descriptors.

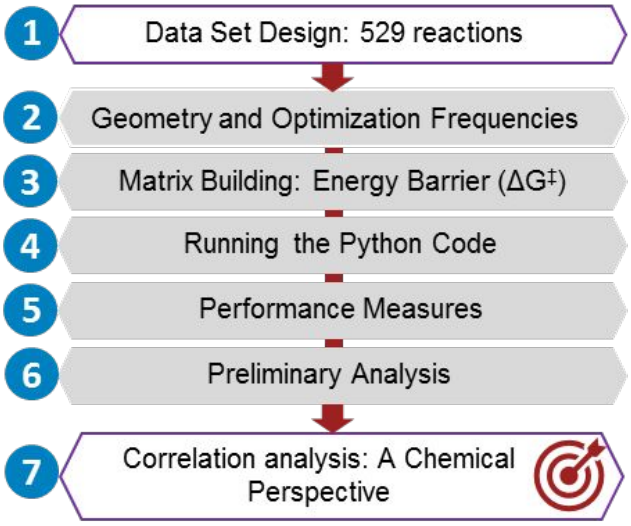

Figure 2. Workflow of the protocol for obtaining hidden descriptors in S_N2@C reaction.

The first step in the treatment is the building of a matrix of free energy barriers. We chose series of twenty three different nucleophiles (listed below) and we computed the barriers for all combinations between them as entering and leaving groups from a methyl center. For each solvent considered, this produced a square matrix with twenty-three rows (one for each entering group) and twenty-three columns (one for each leaving group).

We wrote some in-house Python scripts (supplied in the SI) to perform a SVD[41,42] analysis over the $\mathbf{\Delta G^{\ddagger}}$ matrix following the scheme presented in the Figure 3. The resulting decomposition gave rise to three different matrices related to: i) the entering group (**EG**) (red matrix in Fig. 3), ii) the

weight, defined as the contribution of each hidden descriptor (**W**) (black matrix in Fig. 3), and iii) the leaving group (**LG**^{T}) (blue matrix in Fig. 3). The central matrix (in black) is diagonal.

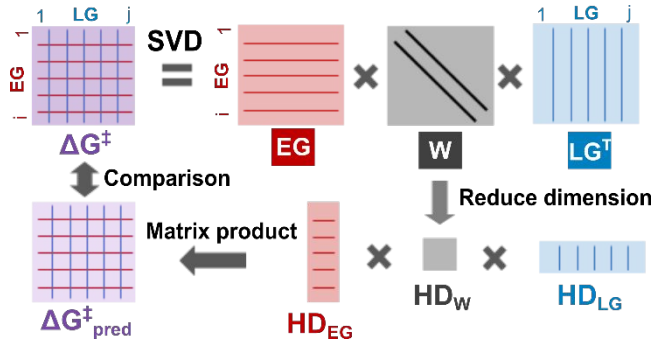

Figure 3. Schematic representation of the hidden descriptor procedure.

In the upper half of Figure 3, the initial Singular Value Decomposition into the product of three matrices is represented. This matrix product is rigorously equal to the initial matrix, albeit more numbers are needed. The interest of the procedure is in the truncation of the matrices (dimensional reduction), as shown in the lower half of Figure 3. The decomposition is hierarchical, the largest weight corresponds to the upper left corner of the black matrix (left part of red matrix, upper part of blue matrix). When we use only one element (one descriptor) from the black matrix, the decomposition warrants that we will obtain the best possible reproduction of the full matrix with a single descriptor. And this descriptor will have associated a single number from the **W** matrix, one column vector from the **EG** matrix and one row vector from the **LG**^{T} matrix. Each of the

elements for these vectors representing a particular nucleophile, either as entering group or as leaving group. The product of the truncated matrices will not reproduce exactly the initial matrix, and the difference between the prediction matrix and the initial one will give an estimation of the validity of the number of necessary descriptors. In our previous study for the case of the BDEs in transition metal complexes we found that five descriptors were necessary for a proper reproduction[20].

Results and discussion

Choice of nucleophiles

A fair election of the chemical species is critical for the application of any data-driven approach. Herein, 23 nucleophiles were evaluated, both as entering groups and leaving groups. They were: F^-, Cl^-, Br^-, I^-, HO^-, MeO^-, EtO^-, HOO^-, $HC(=O)OO^-$, TsO^-, TfO^-, HS^-, MeS^-, EtS^-, HSe^-, H_2N^-, $CHOHN^-$, $(CH_3)_3N$, H_2P^-, H_2As^-, $H_5C_6^-$, F_3C^-, NC^-. This set was chosen after inspection of previous studies of this reaction.[43–45] We are aware that the group undergoing the nucleophilic substitution plays a role,[24] however, for the sake of simplicity, we only considered the methyl group as the

centre of the reaction. We combined all the entering and leaving groups resulting in a total of 529 reactions.

Matrix of free energy barriers

We located the minima and transition states at the potential energy surface (PES) of the S_N2@C reactions employing two solvents: water and dichloromethane (DCM), whose dielectric constants (ε) are quite different, 78.35 and 8.93, respectively. The calculations model the solvent using a continuum model; thus, explicit solvent molecules are not considered. We are aware that in water hydrogen bonds between solvent and solute may be important but this is out of the scope of the current study as we aim to study the effect of a generic highly polar solvent.

As mentioned in the Introduction, there are two possible origins for the calculation of free energy barriers. We can measure them from the separate reactants (R) or from the adduct (Add), and this yields two different values: $\Delta G^{\ddagger}_{R\text{-}TS}$ or $\Delta G^{\ddagger}_{Add\text{-}TS}$. The barrier ought to be measured from the lowest energy point, thus only the highest of these two computed energy differences will correspond to the true barrier. This is not known *a priori*, so we carried out a preliminary study with 17 nucleophiles, reported in the Supporting Information. The conclusion was that the energy of the

adduct was above that of the separate reactants in the vast majority of cases, so we decided to focus our discussion on the $\Delta G^{\ddagger}_{R\text{-}TS}$ values.

Figure 4 shows a color-coded representation of the 23x23 matrix with the free energy barriers in water (see Supporting Information for results with $\Delta G^{\ddagger}_{Add\text{-}TS}$). The numeric versions for the matrix in water and in DCM are supplied in the Supporting Information. Entering groups correspond to the rows in Figure 4, and leaving groups to the columns.

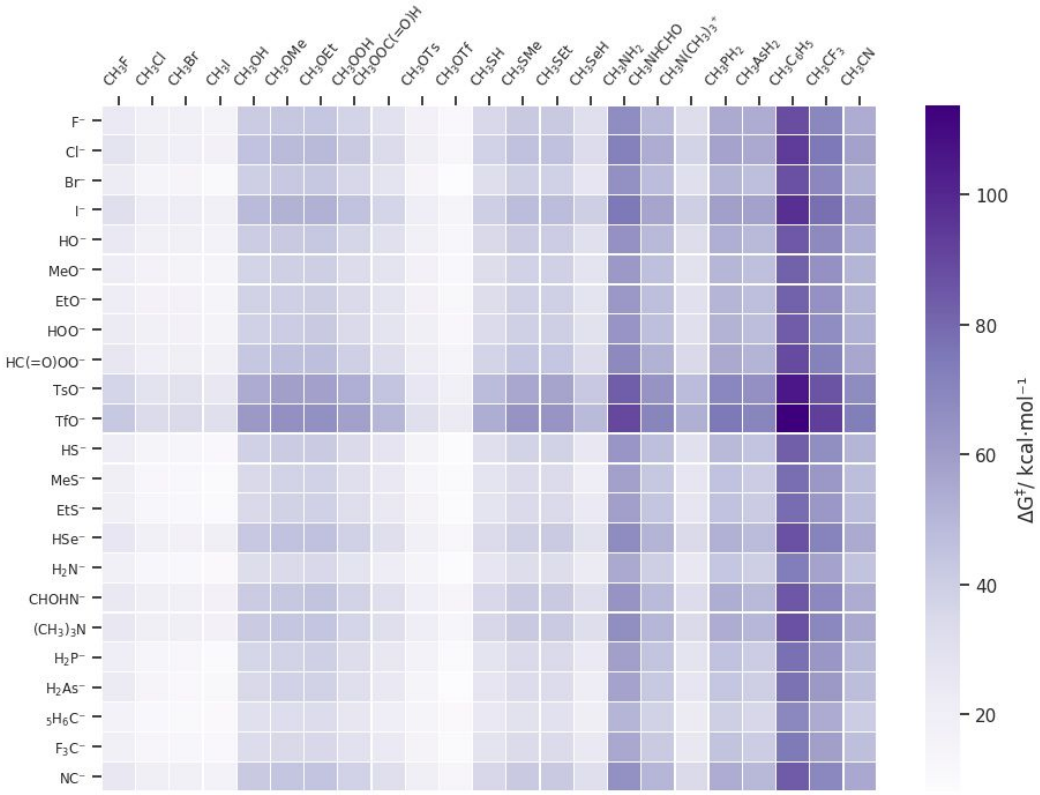

Figure 4. Color-coded version of the matrix of the free energy barriers from the reactants to the transition state ($\Delta G^{\ddagger}_{R\text{-}TS}$).

There are some observations that will be discussed in detail below that are already apparent in Figure 4. The color variation is dominated by columns rather than by rows, indicating a more important role for the leaving group. Some columns are clearly darker than others, confirming for instance that phenyl is a very bad leaving group, while triflate is a very good one. The range of values, indicated in the color code column, extends to barriers higher than 100 kcal/mol, hypothetical reactions that would not take place in experimental conditions. The ability of calculations to quantify these high barriers is an asset for the computational approach in contrast to eventual experimental values. Finally, the main observation is that the existence itself of trends is obvious, which suggests that the problem is well suited for a statistical treatment.

Calculation of hidden descriptors

In the next step, we computed the hidden descriptors following the same procedure outlined in our previous work.[20] Here, the $\Delta G^{\ddagger}_{ij}$ between an entering group *i* and a leaving group *j* will be the element in the ith row and jth column of matrix $\mathbf{\Delta G^{\ddagger}}$. As mentioned above, SVD decomposes a matrix into three different matrices as shown in equation 1, where *n* is the number of studied entering groups (equal to the number of leaving groups), $\mathbf{EG}$, is an $n \times n$ unitary matrix related to

the entering group, $\mathbf{W}$, is a $n \times n$ diagonal matrix for the descriptor weights, and $\mathbf{LG^T}$, is a $n \times n$ matrix for the leaving groups.

$$\mathbf{\Delta G^{\ddagger}} = \mathbf{EG}_{n\times n} \cdot \mathbf{W}_{n\times n} \cdot \mathbf{LG}^{\mathbf{T}}{}_{n\times n} \quad (1)$$

$$\mathbf{\Delta G^{\ddagger}_{pred,}}_{k} = \mathbf{HD}_{k,\mathrm{EG}} \cdot \mathbf{HD}_{k,\mathbf{W}} \cdot \mathbf{HD}_{k,\mathrm{LG}} \quad (2)$$

For dimensionality reduction the key parameters are the number of hidden descriptors, *k,* we want to use. Equation 2 is built by truncation of equation 1, taking the first *k* columns of matrix $\mathbf{EG}$ and the first *k* rows of matrix $\mathbf{LG^T}$. The optimal value for *k* is not defined *a priori*, but has to be adjusted to the problem at hand. The idea is to select the smallest number of descriptors able to reproduce with a sufficient accuracy the values of the matrix. In order to make this choice, we quantify the error associated to each number of descriptors. We can obtain an estimated energy barrier matrix ($\mathbf{\Delta G^{\ddagger}}_{\mathrm{pred},k}$) with the application of equation 2 for a given number of descriptors (*k*). The differences between the exact matrix ($\mathbf{\Delta G^{\ddagger}}$) and the estimated one ($\mathbf{\Delta G^{\ddagger}}_{\mathrm{pred},k}$), will tell which is the error associated to the use of a truncated number of descriptors. In our previous study on the thermodynamics of bond dissociation energies we found out that five descriptors were necessary. For the current case, the number is much smaller.

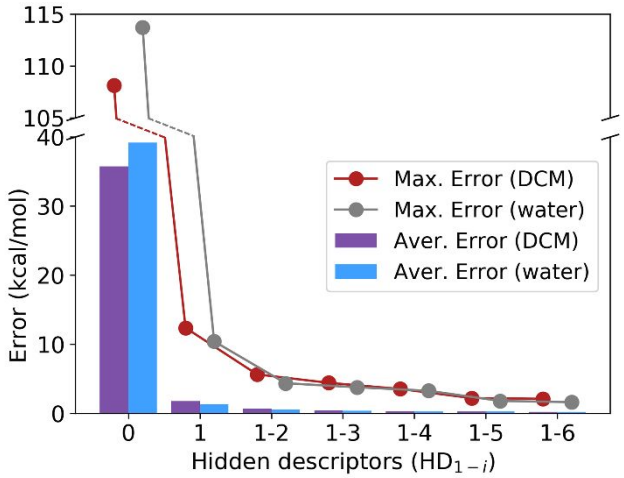

Figure 5. Performance measures for the $\Delta G^{\ddagger}_{pred,k}$ values under DCM (purple bar and red dots) and water (blue bar and grey dots) as a function of the number of hidden descriptors considered ($HD_{1\text{-}i}$). Average errors in bars and maximum error in dots (kcal/mol).

Energy differences (ΔE), root-mean-squared errors (RMSE) and mean absolute errors (MAE) are plotted in the Supporting Information. Figure 5 summarizes the maximum error and average error corresponding to the use of up to six hidden descriptors in both water and DCM. Although solvents are quite different, the performance measurements are similar.

The value for zero hidden descriptors comes from using the average value of the barriers as a guess for each of them. The result for zero descriptors is understandably very bad, with an average error above 30 kcal/mol. But this error is substantially reduced by the introduction of one descriptor. With only one hidden descriptor, HD_1, the $\Delta G^{\ddagger}_{pred,k}$ has an average error lower than 2

kcal/mol for both solvents, which is less than the expected error associated to some DFT functionals.[46,47] The maximum error with the use of single descriptor is 12.4 kcal/mol for DCM and 10.5 kcal/mol for water. Moreover, the improvement associated to the introduction of additional descriptors is quite modest. Therefore, we decide that this reaction is sufficiently described by a single descriptor.

The leading role of the hidden descriptor 1 is confirmed by Table 1, which shows the relative weight of each descriptor in the total barrier. HD_1 is shown to hold roughly the 90% of the total barrier. The other twenty descriptors play a role, but we consider it is so small that it cannot be easily separated from statistical noise.

Table 1. Weight and weight% for water and DCM solvents corresponding to the diagonal values of the W matrices, with respect to the *k* number of the hidden descriptors.

k	Weight wáter	Weight % water	Weight DCM	Weight % DCM
1	1000.92	92.16	934.97	89.28
2	40.18	3.70	53.23	5.08
3	12.12	1.12	14.86	1.42
4	7.63	0.70	9.37	0.89
6	3.15	0.29	3.93	0.38

8	1.99	0.18	3.03	0.29
10	1.74	0.16	2.10	0.20
15	0.81	0.07	1.06	0.10
20	0.56	0.05	0.35	0.03
23	0.04	0.00	0.07	0.01

The need for only one descriptor means that we can assign to each nucleophile two values, one for its entering ability, another one for its leaving ability. We can obtain a good estimate of the energy barrier from $\Delta G^{\ddagger}_{pred} = HD_{1EG} \cdot HD_{1W} \cdot HD_{1LG}$, where HD_{1EG} and HD_{1LG} are the hidden descriptor values that depends on the identity of the entering and leaving groups and HD_{1W} is a constant (HD_{1W}= 1000.92 in water and 934.97 in DCM). We expected bimolecular nucleophilic substitution to be a process chemically simpler than formation of a metal ligand bond, which required five descriptors, but were to a certain extent surprised by this result. The result indicates that the process is chemically simple, but it has some interesting consequences that will be discussed below.

The statistical stability of our matrix of data was confirmed with a series of studies, reported in the SI, where the full set was divided in two subsets, and leave-one-out tests were carried out.

Initial analysis of descriptor HD_1

We report in Table 2 the values of hidden descriptor 1 for each of the nucleophiles. We focus below on the values in water. Values in DCM follow similar trends, and they are reported in the SI. We remark here that the prediction of the barrier in water for a given combination of EG and LG, corresponds to the product of the descriptors for each of the two groups times the weight of this descriptor, which is 1000.92 (Table 1).

Regarding the Table 2, an interesting results is the range of the values, which is significantly narrower for entering ability (0.144 units, from 0.157 to 0.301) than for leaving ability (0.350 units, from 0.062 to 0.412). This means a larger sensibility of the barrier to the nature of the leaving group.

Table 2. Value for HD_1 in water associated to each nucleophile. Separate values are available for entering ability and leaving group.

HD_{1EG} (water)		HD_{1LG} (water)		HD_{1EG} (water)		HD_{1LG} (water)	
TfO^-	0.301	TfO^-	0.062	HOO^-	0.201	HO^-	0.197
TsO^-	0.275	I^-	0.076	EtO^-	0.196	EtS^-	0.198
I^-	0.240	Br^-	0.087	HS^-	0.195	MeO^-	0.209

Cl^-	0.230	Cl^-	0.089	MeO^-	0.195	EtO^-	0.212
$HC(=O)OO^-$	0.219	TsO^-	0.091	H_2P^-	0.183	H_2As^-	0.233
$(CH_3)_3N$	0.213	F^-	0.120	EtS^-	0.181	$CHOHN^-$	0.237
HSe^-	0.212	HSe^-	0.144	MeS^-	0.179	H_2P^-	0.253
F^-	0.211	$HC(=O)OO^-$	0.147	H_2As^-	0.177	NC^-	0.257
NC^-	0.211	$(CH_3)_3N$	0.160	F_3C^-	0.174	H_2N^-	0.315
$CHOHN^-$	0.209	HS^-	0.166	H_2N^-	0.169	F_3C^-	0.330
HO^-	0.207	HOO^-	0.180	$H_5C_6^-$	0.157	$H_5C_6^-$	0.412
Br^-	0.202	MeS^-	0.197				

Moreover, the list of values in Table 2 can be seen as scales of entering ability and leaving ability of the different nucleophiles. A first inspection of these scales shows a reasonable correlation with expectations. The worst leaving group (highest value for the descriptor, 0.412), is phenyl, and the best leaving group (smallest value, 0.062) is triflate. The same groups also repeat in the extreme positions for entering groups, the best entering group is phenyl (0.157) and the worst one is triflate (0.301). It is not surprising to find and inverse correlation between the two rankings: the best EGs ought to be the worst LGs, and *viceversa*. However, on a second inspection we can see that the correlation is far from perfect, as there are many variations in the ordering between the two scales. These differences between the scales for entering and leaving groups prompted us to further examine the correlation between them.

We built vectors from the scales in Table 2, and examined the correlation between them. We summarize the results in Figure 6, where we present the correlation between the different HD_1 fectors. The elements in the diagonal of Figure 6 are obviously 1.00, as they represent the correlation of each vector with itself. The only other value that approaches unity is the correlation between the leaving group ability in water and the leaving group ability in DCM. The facility of a given nucleophile to act as leaving group is thus very similar in both solvents.

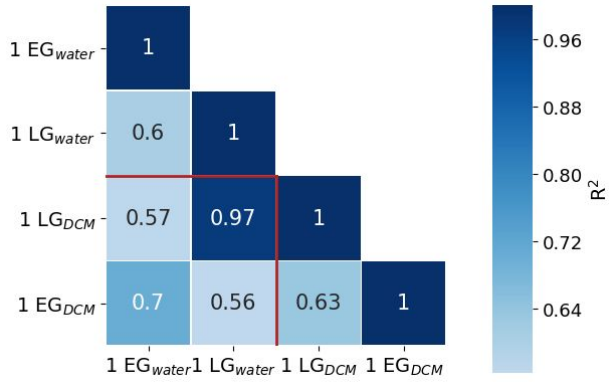

Figure 6. Matrix of the square correlation coefficient (R^2) obtained from the linear regressions between the HD_1 (referred as 1), of EG and LG in water or DCM.

However, all other correlations are significantly smaller, with R^2 coefficients between 0.56 and 0.70. This is remarkable, especially when we deal with the scales for HD_{1EG} and HD_{1LG} in the same solvent. It means that the correlation between the entering and leaving ability of a given

nucleophile is far from perfect. This is not what we expected, and prompted us to further examine the matter in the next section.

Relationship of HD_1 with conventional descriptors

Now we will try to gain chemical insight from the purely statistical result reported above. For this, we will use again statistical tools, as we will analyze the correlation between our values of HD_1 and a variety of conventional descriptors. We will use a large set of 175 conventional descriptors, the full list is supplied in the Supporting Information. Most of them correspond to physical organic chemistry properties which are widely employed in this type of studies.[17,48] Frontier orbital energies, different atomic charges, transition state distances, chemical concepts such as ionization potential, electron affinity, and softness, among others, are all included in the list. This collection of properties enables the description of electronic and steric effects.

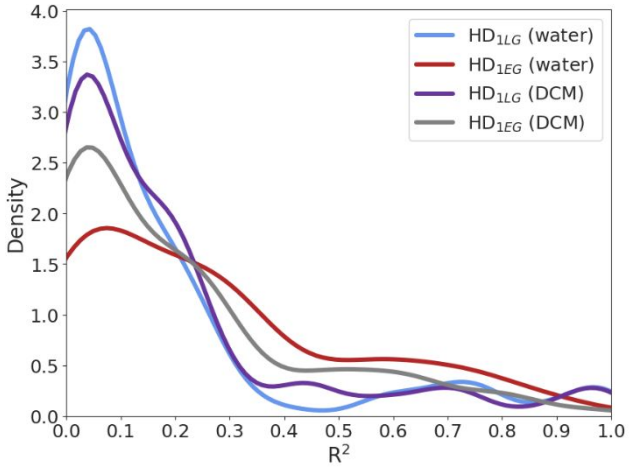

Figure 7. Density plot of the square correlation coefficient (R^2) derived from the linear regression between the conventional descriptors and the HD_{1LG} (water) in blue; HD_{1EG} (water) in red; HD_{1LG} (DCM) in purple; HD_{1EG} (DCM) in grey.

Figure 7 shows the probability density plot of the R^2 derived from the linear regressions between the HD_1 numbers and the conventional descriptors. Overall, most of the models have a R^2 lower than 0.500, and there is a small density of conventional descriptors with a higher R^2 value. Thus, the ones that are highly correlated are well distinguished for the rest.

The R^2 derived from the relationships between the set of all the descriptors and the HD_1 are gathered in the Supporting Information and the values for some significant ones are collected in Table 3. For the sake of simplicity, we represent X^- as the entering group, Y as the leaving group and CH_3Y as the electrophile.

Table 3. Square correlation coefficient (R^2) between different HD_1 values and selected conventional descriptors. Highest values for each hidden descriptor in bold.

Code	Descriptor	R^2 with HD_{1LG} (water)	R^2 with HD_{1EG} (water)	R^2 with HD_{1LG} (DCM)	R^2 with HD_{1EG} (DCM)
1	HOMO of X^-	0.263	0.630	0.292	0.452
8	LUMO of CH_3Y	0.175	0.306	0.199	0.385

51	ω- (I) of X^-	0.170	0.301	0.059	0.003
60	ω^+ (I) of CH_3Y	0.170	0.698	0.100	0.110
114	ΔV for (I^- + CH3Y → CH_3I + Y^-)	**0.974**	0.732	**0.967**	0.780
130	d(I-C) for TS of (X^- + CH_3I → CH_3X + I^-)	0.701	**0.907**	0.661	**0.839**

We had expected that the descriptors involved in donor-acceptor interactions to play a leading role in S_N2@C reactivity. This is why we introduced in the list the HOMO or LUMO energy values of the chemical species involved, as well as other values related to these frontier molecular orbitals. But to our surprise, they did not show particularly high correlation values in any case.

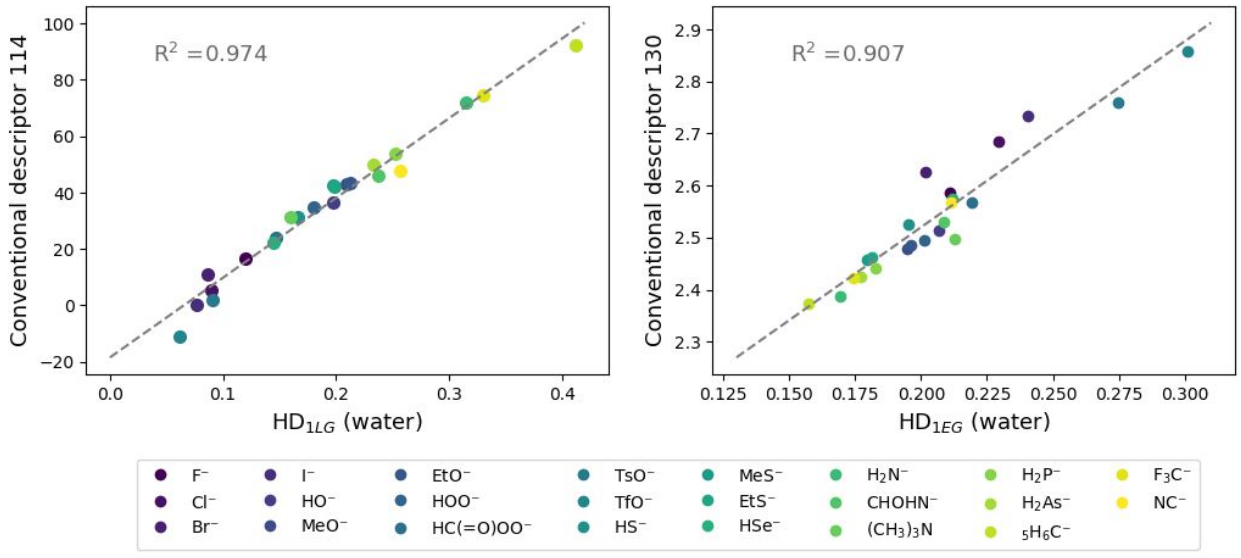

Figure 8. Plot of the values of the HD_{1LG} (left plot), and HD_{1EG} (right plot) in water vs the values of the corresponding conventional descriptor providing the best correlation outcome.

We will start our discussion with HD_{1LG} in water, as it has one of the largest range of descriptor values. The LUMO energy of the CH_3Y molecule, which should be in principle responsible for the reaction presents a very low correlation of 0.175. This may be explained because the LUMO in the reactant will not always correspond to the key orbital in the TS. We also expected to find correlations with an alternative frontier molecular orbital (FMO) feature, as the electrodonating power ω^+, but things were not much better.

Instead, we found that HD_{1LG} in water correlates best with our descriptor number 114, associated to the thermodynamics of the process, defined specifically by the potential energy of the reaction, $I^- + CH_3Y \rightarrow CH_3I + Y^-$. Figure 8 left displays the linear regression between the HD_{1LG} and descriptor number 114. This in fact measures the heterolytic bond dissociation energy of the CH_3-Y bond, as the role of iodine in the equation is identical for all Y groups. We notice that this descriptor concerns potential energy, the corresponding descriptor related to free energy, which was also in the list, has a slightly worse correlation. It is not surprising that stronger bonds are more difficult to cleave, but a correlation of 0.974 is indeed quite high. And it means that this

thermodynamic parameter plays the key role in the barrier for the reaction. In the end, the weakness towards heterolytic cleavage of the C-Y bond is responsible thus for the efficiency of leaving groups of substituents such as triflate, tosylate and the halides. It is worth mentioning that this high correlation between HD_{1LG} and descriptor 114 is also observed for the reaction in DCM. Hence, this dependance on the cleavage energy seems a general feature of the S_N2 process.

We next move our attention to the description of the entering group, represented in the HD_{1EG} scale. We analyze first the results in water. Here, we expected the energy of the HOMO or the electrodonating power ω^- of the entering group, descriptor 1 and 51, respectively, to play a major role, but this is not the case. Within them, only the R^2 of 0.630 for HOMO descriptor is respectable, but is in no way among the highest. The highest value, 0.907, corresponds to conventional descriptor 130, the correlation with HD_{1EG} in water being shown in Figure 8 right. This is a geometrical parameter related to the geometry of a transition state. Specifically, it is the distance between the iodine and the carbon centre in the transition state for the reaction between the nucleophile X^- and CH_3I. This has a more kinetic component, as it is associated to a transition state. The trend is qualitatively logical in the sense that "good" entering groups will have early transition states, with smaller elongation of the C-I distance associated to the leaving group. Still,

it is remarkable that the correlation is with this relatively obscure descriptor and not with more intuitive ones. It is also worth mentioning that the entering group seems to have a smaller effect on the barrier than the leaving group. As mentioned in a previous section, the range of values is narrower, and we see here that the highest correlation values are smaller than for the leaving group. Additional evidence in this direction is provided by the discrepancy between results water and DCM, with an even lower value of 0.839 for R^2 in the latter.

We applied the same treatment reported above to higher order descriptors, HD_2, HD_3 and HD_4. The results are detailed in the Supporting Information. None of the relationships established with the conventional descriptors achieved a value of R^2 higher than 0.830. This is further evidence confirming our hypothesis that HD_2, HD_3 and HD_4 are mostly associated with numerical noise, and they do not furnish a significative contribution to the chemical understanding of the reaction.

The main observation resulting from this section is that the descriptor for S_N2@C reactions emerging from our statistical analysis does present a good correlation with some conventional descriptors, but not with the ones we would have expected *a priori*. This raises some questions as to the eventual need to rethink some of our assumptions on this important organic chemistry process.

Conclusion

The statistical analysis of DFT calculations on the transition states of more than 500 bimolecular nucleophilic substitution reactions involving 23 entering groups and 23 leaving groups in two different solvents demonstrates that the free energy barrier can be predicted with an average error below 2 kcal/mol with the use of a single descriptor. Our study leads to the identification of two numerical values for each nucleophile, one for its entering ability and one for its leaving ability, which introduced in a simple arithmetic expression yield a close estimation of the DFT barrier without the need for a QM calculation. In this way, the current work proves the applicability of the hidden descriptor approach, which we had developed for a thermodynamical property in inorganic chemistry, to reactivity in organic chemistry.

The numerical values for the descriptors associated to each nucleophile give moreover insight into fundamental aspects of the S_N2 process. There is a significant asymmetry between the scales for entering and leaving groups, with the identity of the LG having a higher influence on the barrier. More significantly, and contrary to expectation, our computed descriptor shows only a moderate correlation with magnitudes related with frontier molecular orbitals. Instead, it follows

more closely the trends of other thermodynamic and geometrical conventional descriptors. The general validity of these observations will be put to test by further studies on other reactions which are currently under development in our Laboratory.

ASSOCIATED CONTENT

Energy barrier matrices, hidden descriptors matrices, performance measures of the estimation of the energy barriers. List of the conventional descriptors, and the equations and the squared coefficients of the correlations between them and the hidden descriptors.

Funding Sources

We thank for finantial support the CERCA Programme/Generalitat de Catalunya, the MINECO (Grants PGC2018-100780-B-I00 and PID2020-112825RB-100) FEDER funds. L. M.-G. thanks Generalitat de Catalunya for the FI-Agaur predoctoral contract, 2020FI-B 00245.

REFERENCES

(1) Jorner, K.; Tomberg, A.; Bauer, C.; Sköld, C.; Norrby, P. O. Organic Reactivity from Mechanism to Machine Learning. *Nat. Rev. Chem.* **2021**, *5*, 240–255.

(2) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chem. Rev.* **2021**, *121*, 9816–9872.

(3) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 1–7.

(4) Balcells, D.; Skjelstad, B. B. The TmQM Dataset - Quantum Geometries and Properties of 86k Transition Metal Complexes. *J. Chem. Inf. Model* **2020**, *60*, 6135–6146.

(5) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604–610.

(6) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, *577*, 706–710.

(7) St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S. Prediction of Organic Homolytic Bond Dissociation Enthalpies at near Chemical Accuracy with Sub-Second Computational

Cost. *Nat. Commun.* **2020**, *11*, 1–12.

(8) 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*, 1163–1175.

(9) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363*, eaau5631.

(10) Besora, M.; Olmos, A.; Gava, R.; Noverges, B.; Asensio, G.; Caballero, A.; Maseras, F.; Pérez, P. J. A Quantitative Model for Alkane Nucleophilicity Based on C−H Bond Structural/Topological Descriptors. *Angew. Chemie - Int. Ed.* **2020**, *59*, 3112–3116.

(11) Harper, K. C.; Sigman, M. S. Three-Dimensional Correlation of Steric and Electronic Free Energy Relationships Guides Asymmetric Propargylation. *Science* **2011**, *333*, 1875–1878.

(12) Aguado-Ullate, S.; Urbano-Cuadrado, M.; Villalba, I.; Pires, E.; García, J. I.; Bo, C.; Carbó, J. J. Predicting the Enantioselectivity of the Copper-Catalysed Cyclopropanation of Alkenes by Using Quantitative Quadrant-Diagram Representations of the Catalysts. *Chem. - A Eur.*

J. **2012**, *18*, 14026–14036.

(13) Zhao, S.; Gensch, T.; Murray, B.; Niemeyer, Z. L.; Sigman, M. S.; Biscoe, M. R. Enantiodivergent Pd-Catalyzed C–C Bond Formation Enabled through Ligand Parameterization. *Science* **2018**, *362*, 670–674.

(14) Fey, N.; Koumi, A.; Malkov, A. V.; Moseley, J. D.; Nguyen, B. N.; Tyler, S. N. G.; Willans, C. E. Mapping the Properties of Bidentate Ligands with Calculated Descriptors (LKB-Bid). *Dalt. Trans.* **2020**, *49*, 8169–8178.

(15) Friederich, P.; Dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine Learning Dihydrogen Activation in the Chemical Space Surrounding Vaska's Complex. *Chem. Sci.* **2020**, *11*, 4584–4601.

(16) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher., S. D.; Doyle, A. G. Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning. *Science* **2018**, *360*, 186–190.

(17) Li, X.; Zhang, S. Q.; Xu, L. C.; Hong, X. Predicting Regioselectivity in Radical C−H Functionalization of Heterocycles through Machine Learning. *Angew. Chemie* **2020**, *132*,

13355–13361.

(18) Falivene, L.; Cao, Z.; Petta, A.; Serra, L.; Poater, A.; Oliva, R.; Scarano, V.; Cavallo, L. Towards the Online Computer-Aided Design of Catalytic Pockets. *Nat. Chem.* **2019**, *11*, 872–879.

(19) Durand, D. J.; Fey, N. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019**, *119*, 6561–6594.

(20) Lakuntza, O.; Besora, M.; Maseras, F. Searching for Hidden Descriptors in the Metal-Ligand Bond through Statistical Analysis of Density Functional Theory (DFT) Results. *Inorg. Chem.* **2018**, *57*, 14660–14670.

(21) Keil, F.; Reinhart, A. Theoretical Study of SN2 Reactions . Ab Initio Computations on HF and CI Level. *J. Am. Chem. Soc.* **1976**, *98*, 4787–4793.

(22) Alemán, C.; Maseras, F.; Lledós, A.; Duran, M.; Bertrán, J. Analysis of Solvent Effect on SN2 Reactions by Different Theoretical Models. *J. Phys. Org. Chem.* **1989**, *2*, 611–622.

(23) Deng, L.; Ziegler, T.; Branchadell, V. Potential Energy Surfaces of the Gas-Phase SN2 Reactions X- + CH3X = XCH3 + X- (X = F, Cl, Br, I): A Comparative Study by Density

Functional Theory and Ab Initio Methods. *J. Am. Chem. Soc.* **1994**, *116*, 10645–10656.

(24) Bickelhaupt, F. M.; Swart, M.; Sola, M. Energy Landscapes of Nucleophilic Substitution Reactions%: A Comparison of Density Functional Theory and Coupled Cluster Methods. *J. Comput. Chem* **2007**, *28*, 1551–1560.

(25) Wu, C. H.; Galabov, B.; Wu, J. I. C.; Ilieva, S.; Schleyer, P. V. R.; Allen, W. D. Do π-Conjugative Effects Facilitate Sn2 Reactions? *J. Am. Chem. Soc.* **2014**, *136*, 3118–3126.

(26) Wu, Z.; Rong, C.; Lu, T.; Ayers, P. W.; Liu, S. Density Functional Reactivity Theory Study of SN2 Reactions from the Information-Theoretic Perspective. *Phys. Chem. Chem. Phys.* **2015**, *17*, 27052–27061.

(27) Hamlin, T. A.; Beek, B. Van; Wolters, L. P.; Bickelhaupt, F. M. Nucleophilic Substitution in Solution%: Activation Strain Analysis of Weak and Strong Solvent Effects. *Chem. Eur. J.* **2018**, *24*, 5927–5938.

(28) Gimadiev, T.; Madzhidov, T.; Tetko, I.; Nugmanov, R.; Casciuc, I.; Klimchuk, O.; Bodrov, A.; Polishchuk, P.; Antipin, I.; Varnek, A. Bimolecular Nucleophilic Substitution Reactions: Predictive Models for Rate Constants and Molecular Reaction Pairs Analysis.

Mol. Inf. **2019**, *38*, 1800104.

(29) Hamlin, T. A.; Swart, M.; Bickelhaupt, F. M. Nucleophilic Substitution (SN2): Dependence on Nucleophile, Leaving Group, Central Atom, Substituents, and Solvent. *ChemPhysChem* **2018**, *19*, 1315–1330.

(30) 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. V; 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 Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; 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 16 Revision C.01. Gaussian

Inc. Wallingford CT, 2016.

(31) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(32) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37*, 785–789.

(33) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 1–19.

(34) McLean, A. D.; Chandler, G. S. Contracted Gaussian Basis Sets for Molecular Calculations. I. Second Row Atoms, Z=11-18. *J. Chem. Phys.* **1980**, *72*, 5639–5648.

(35) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XX. A Basis Set for Correlated Wave Functions. *J. Chem. Phys.* **1980**, *72*, 650–654.

(36) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. R. Efficient Diffuse Function-augmented Basis Sets for Anion Calculations. III. The 3-21+G Basis Set for

First-row Elements, Li–F. *J. Comput. Chem.* **1983**, *4*, 294–301.

(37) Wadt, W. R.; Hay, P. J. Ab Initio Effective Core Potentials for Molecular Calculations. Potentials for Main Group Elements Na to Bi. *J. Chem. Phys.* **1985**, *82*, 284–298.

(38) Check, C. E.; Faust, T. O.; Bailey, J. M.; Wright, B. J.; Gilbert, T. M.; Sunderlin, L. S. Addition of Polarization and Diffuse Functions to the LANL2DZ Basis Set for P-Block Elements. *J. Phys. Chem. A* **2001**, *105*, 8111–8116.

(39) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(40) Álvarez-Moreno, M.; De Graaf, C.; López, N.; Maseras, F.; Poblet, J. M.; Bo, C. Managing the Computational Chemistry Big Data Problem: The IoChem-BD Platform. *J. Chem. Inf. Model.* **2015**, *55*, 95–103.

(41) Deprettere, F. SVD and Signal Processing: Algorithms, Analysis and Applications; Elsevier Science Publishers: Amsterdam, 1988.

(42) Stone, J. V. Independent Component Analysis: A Tutorial Introduction; MIT Press:

Cambridge, MA, 2004.

(43) Uggerud, E. Nucleophilicity - Periodic Trends and Connection to Basicity. *Chem. - A Eur. J.* **2006**, *12*, 1127–1136.

(44) Gonzales, J. M.; Cox, R. S.; Brown, S. T.; Allen, W. D.; Schaefer, H. F. Assessment of Density Functional Theory for Model S N 2 Reactions%: CH3X + F- (X = F , Cl, CN, OH, SH, NH2, PH2). *J. Phys. Chem. A* **2001**, *105*, 11327–11346.

(45) Alkorta, I.; Thacker, J. C. R.; Popelier, P. L. A. An Interacting Quantum Atom Study of Model SN2 Reactions (X-···CH3X, X = F, Cl, Br and I). *J. Comput. Chem* **2018**, *39*, 546–556.

(46) Medvedev, M. G.; Ivan S., B.; Jianwei, S.; John P., P.; Konstantin A., L. Density Functional Theory Is Straying from the Path toward the Exact Functional. *Science* **2017**, *355*, 49–52.

(47) Sim, E.; Song, S.; Burke, K. Quantifying Density Errors in DFT. *J. Phys. Chem. Lett.* **2018**, *9*, 6385–6392.

(48) Werth, J.; Sigman, M. S. Linear Regression Model Development for Analysis of Asymmetric Copper-Bisoxazoline Catalysis. *ACS Catal.* **2021**, *11*, 3916–3922.

TOC graphic