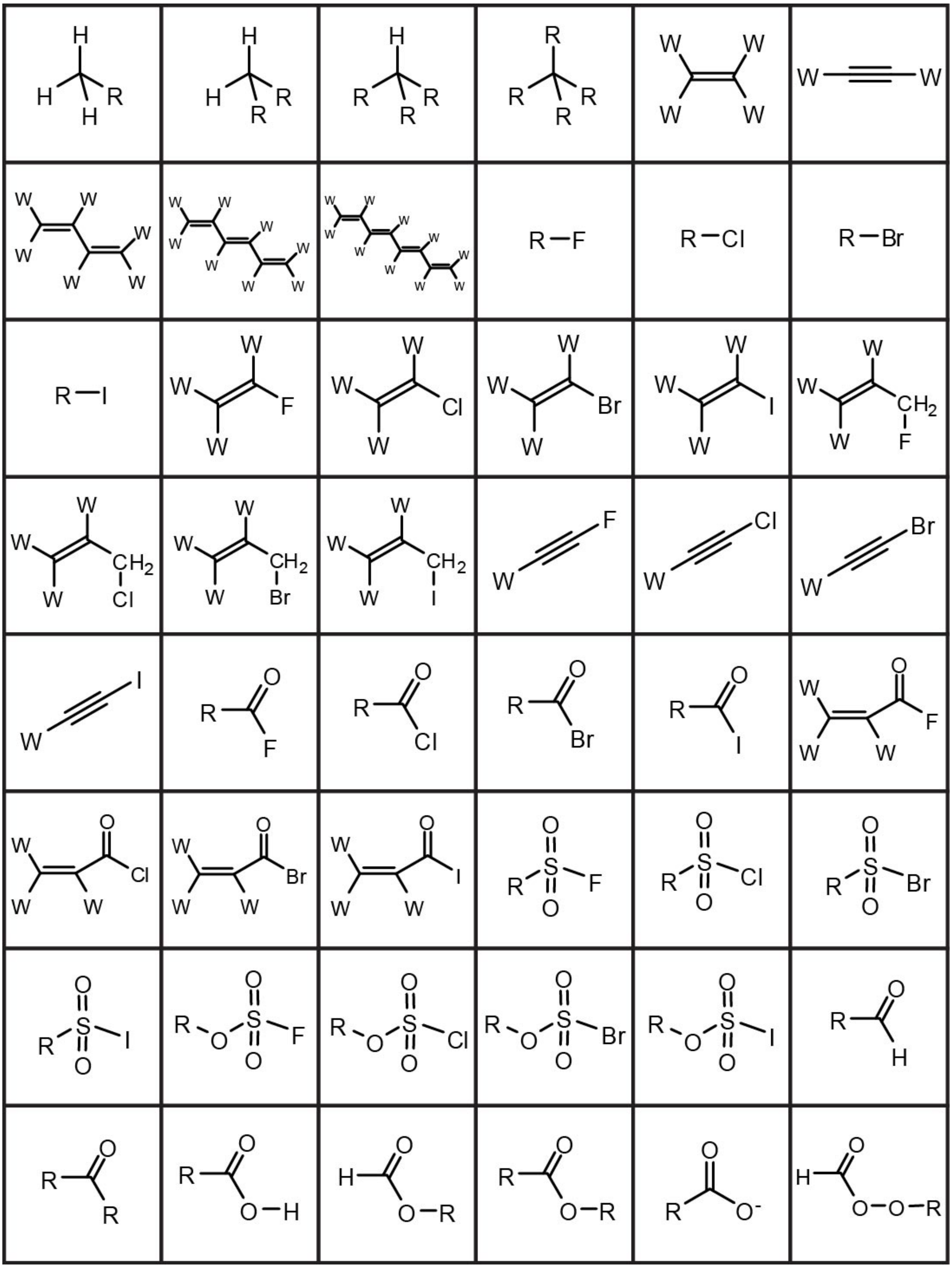
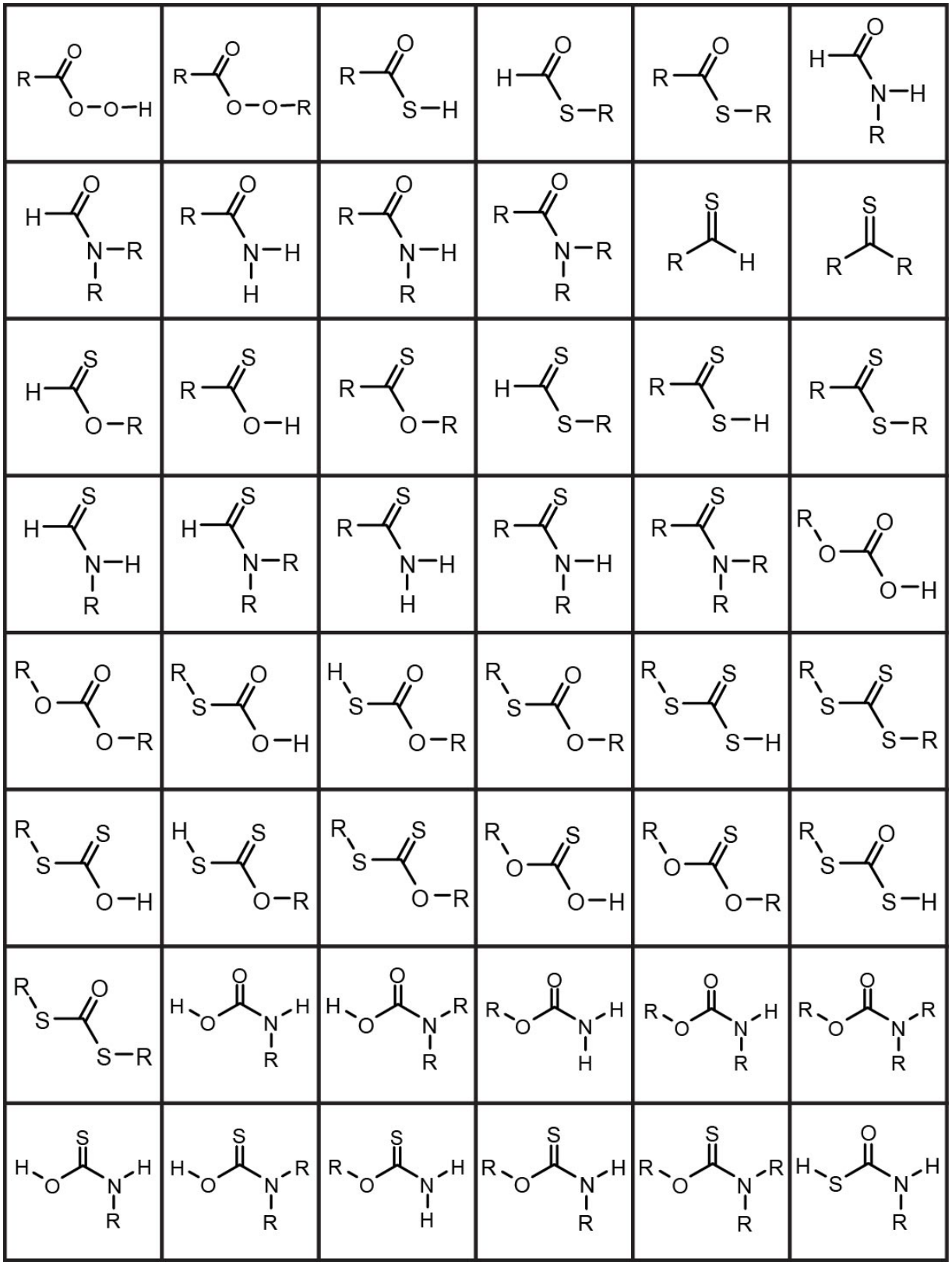
Supplementary Information for Manuscript entitled ***Estimating Chemical Reactivity and Cross-Influence from Collective Chemical Knowledge*** by S. Soh, Y. Wei, B. Kowalczyk, C.M. Gothard, B. Baytekin, N. Gothard and B.A. Grzybowski*

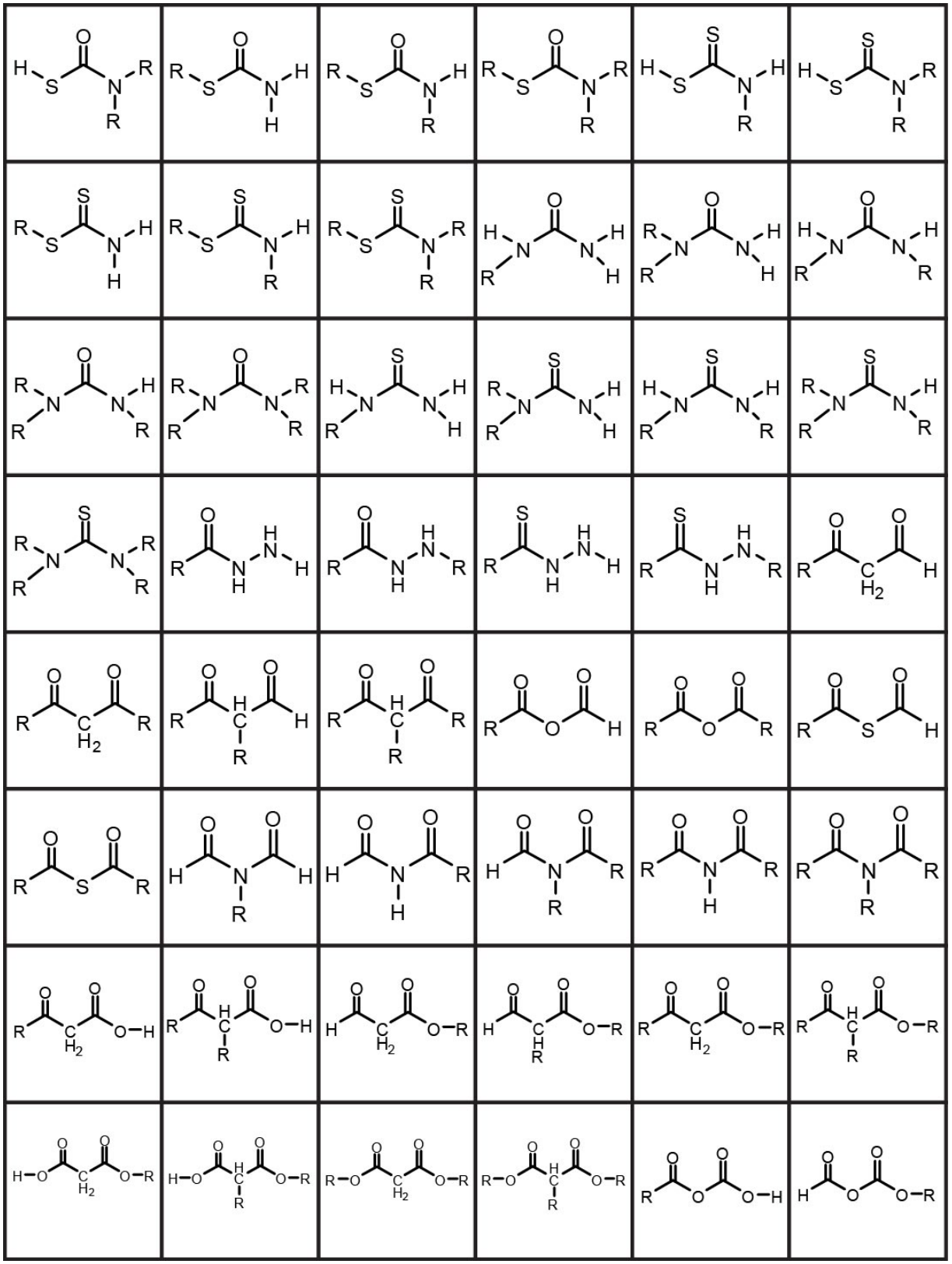
1. Automated identification of functional groups in organic molecules

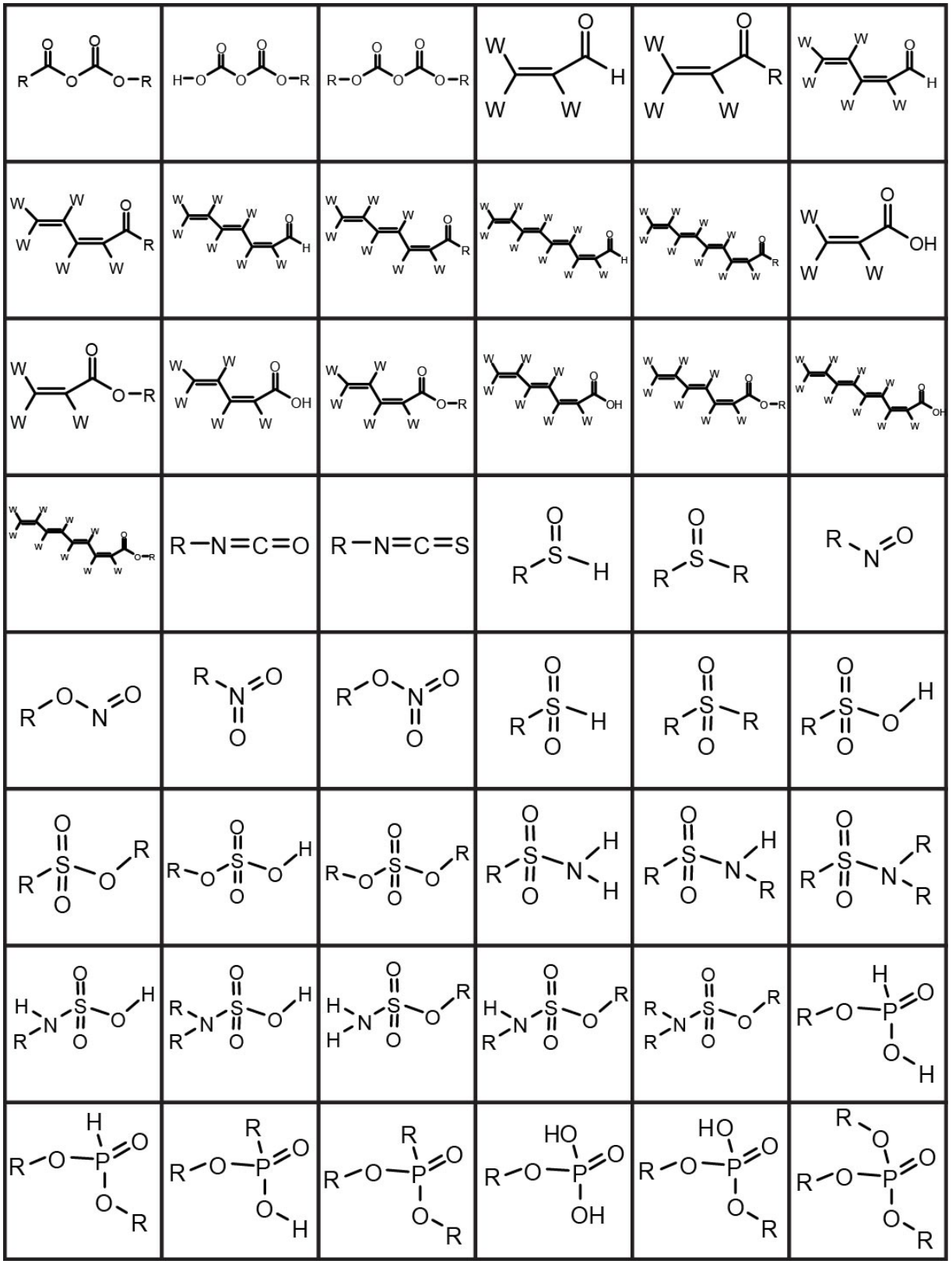
The analysis in this study requires first identifying the various types of functional groups present in organic molecules. Since the numbers of molecules considered are in the millions, this process cannot be performed manually but must, instead, be automated. To this end, we have implemented a computational algorithm (to be published separately), which goes through each molecule atom-by-atom systematically, matching portions of the molecules to specific functional groups listed in Table S1. This table comprises 322 most representative groups found in organic compounds, each with its own distinct reactivity.

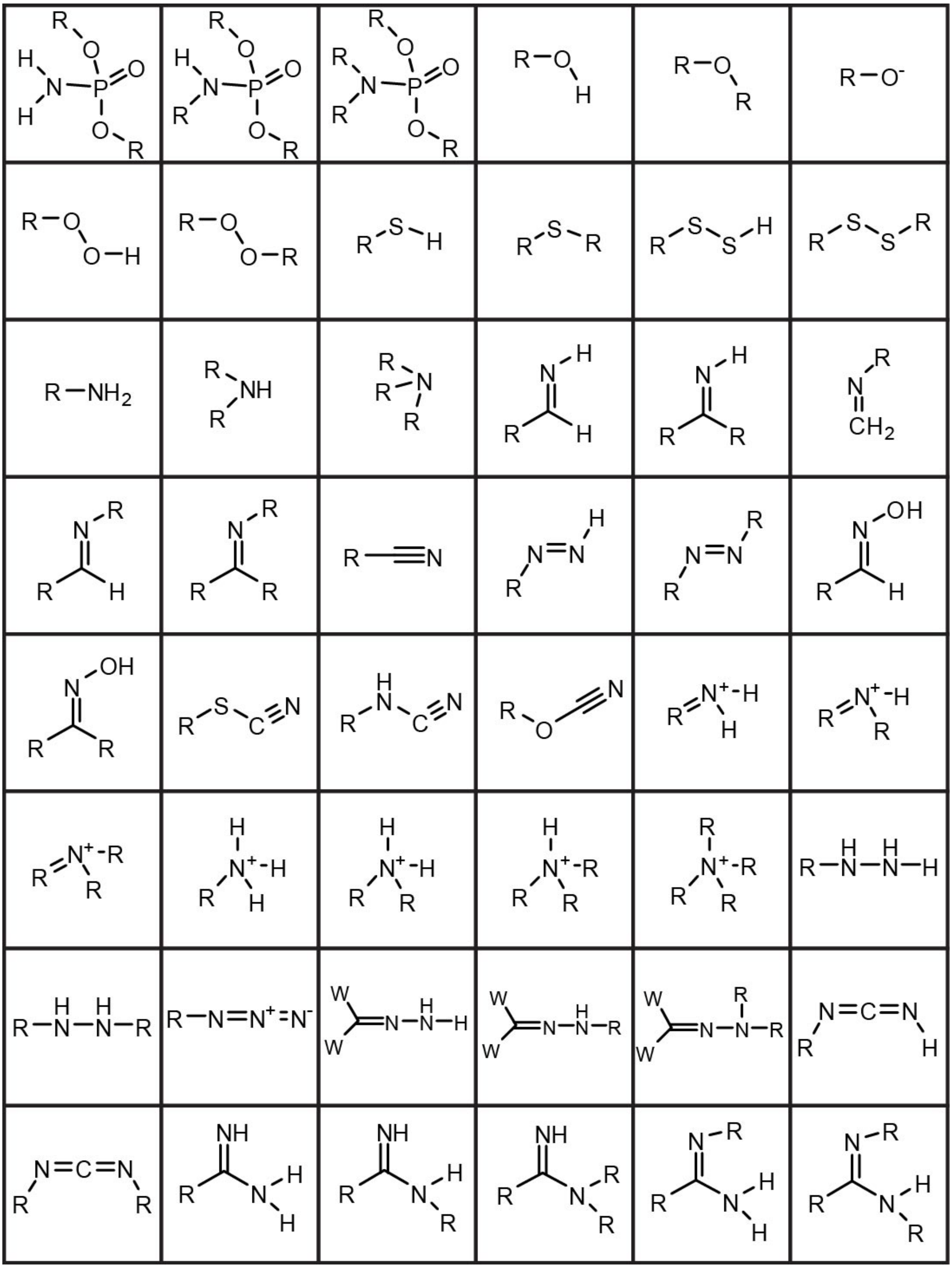
Table S1. List of 322 functional groups



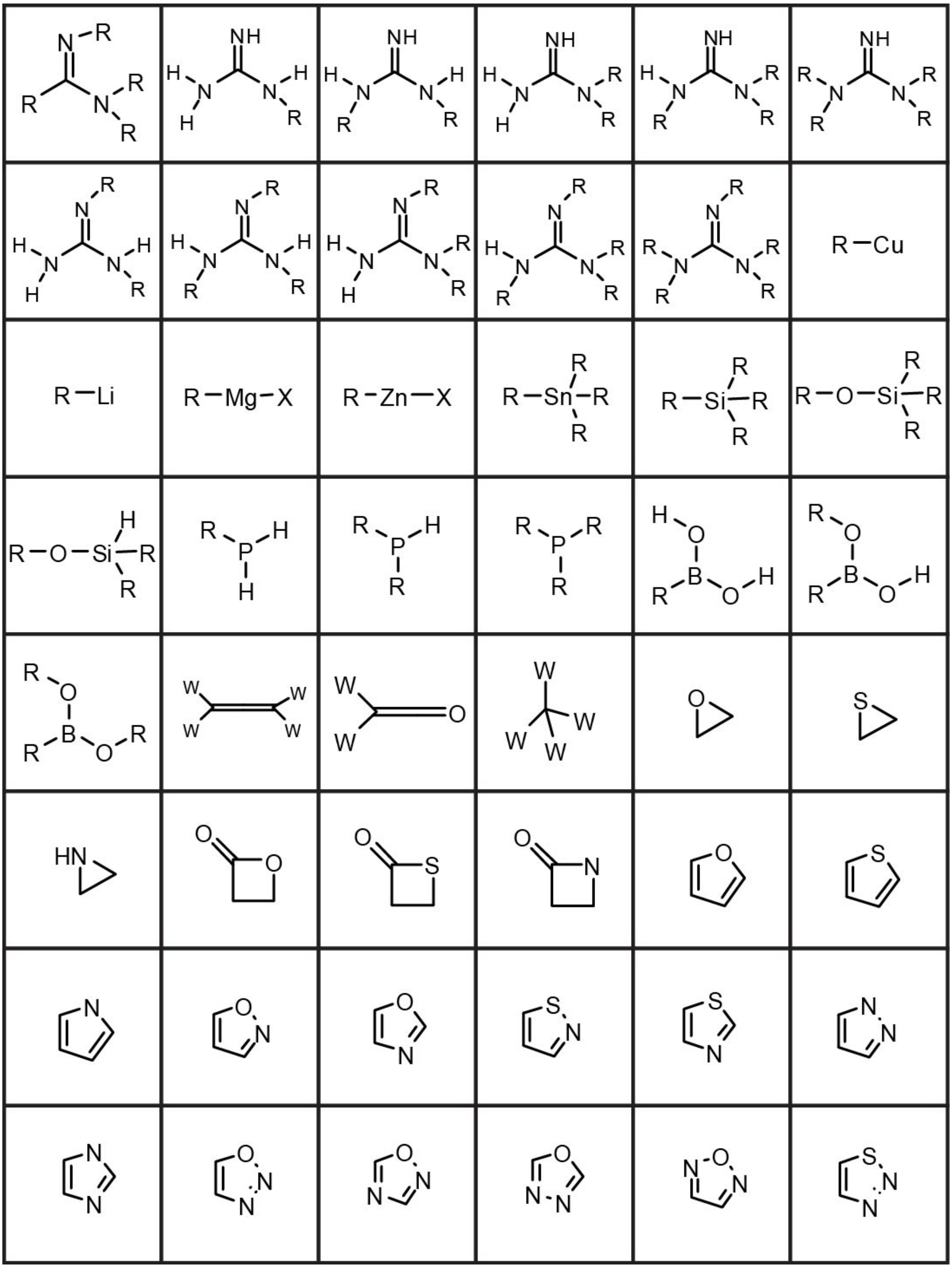


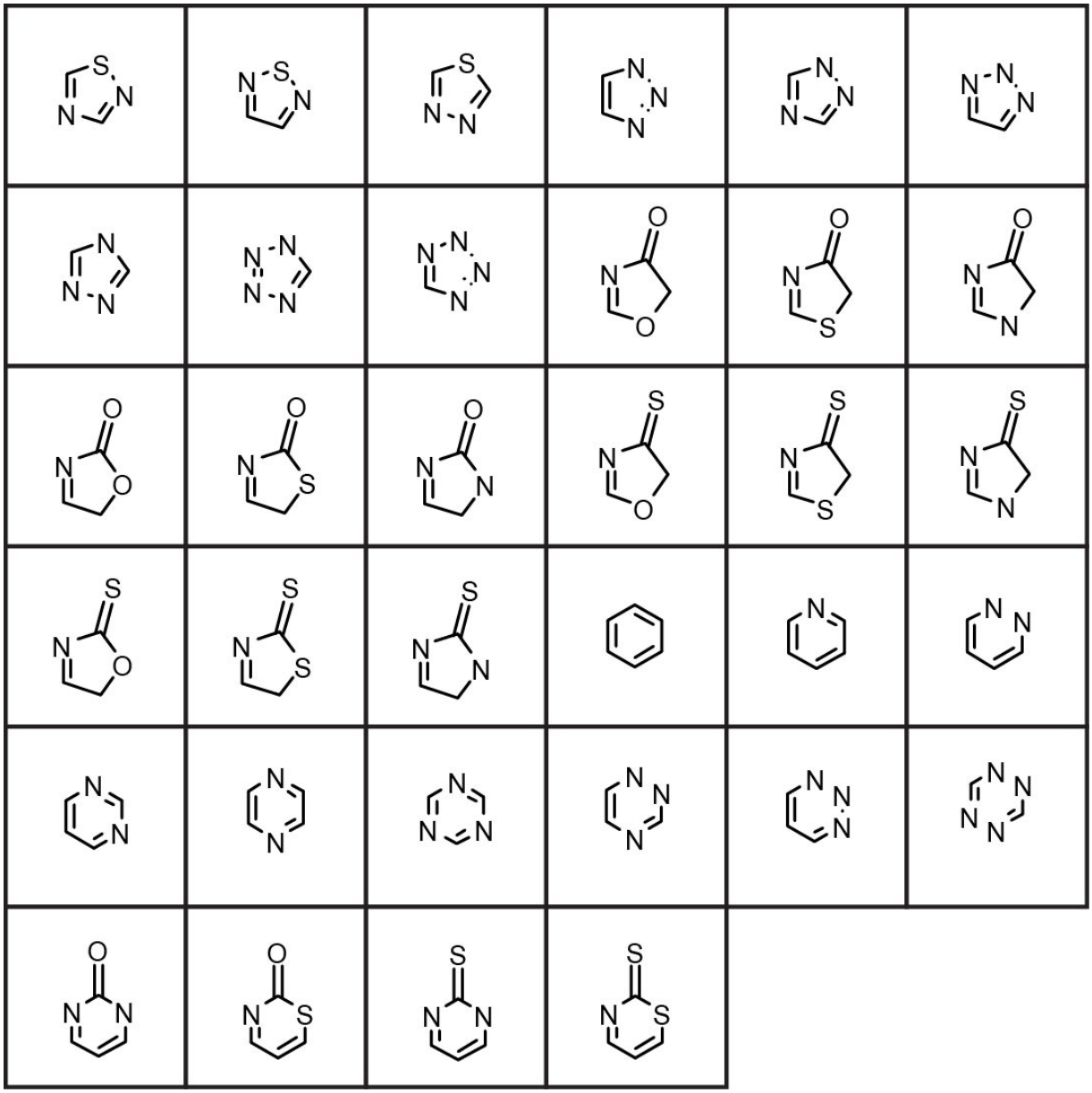






Electronic Supplementary Material (ESI) for Chemical Science
This journal is © The Royal Society of Chemistry 2012





Note 1: "W" represents any atom in the list "H, C, N, O, S, F, Cl, Br, I, P, and B". "R" represents a "C" atom (not "H", for example). "X" represents any halide atom "F, Cl, Br, and I".

Note 2: For functional groups which contain a ring, all the atoms in the ring can be singly bonded to any type of atom outside the ring. These external bonds are not drawn for simplicity.

2. Confidence intervals of reactivity indices

The confidence interval of the reactivity indices indicates the degree of uncertainty associated with the measure and the sample size (i.e., number of reactions on which the index is

based). In order to determine the confidence interval, a search is performed through the database to identify all the molecules containing functional group **A** and the reactions in which these **A**-containing molecules serve as substrates. Within this set of reactions, N_A^{tot}, there can be two outcomes: (1) functional group **A** is reacted or (2) functional group **A** remains unchanged after the reaction. Due to the binary nature of the outcomes, the number of reactions in which **A** reacts, N_A, follows a binomial distribution. Therefore, the reactivity index, $R_A = N_A / N_A^{tot}$, also follows a binomial distribution with a standard deviation of $\sigma_A = \sqrt{R_A(1-R_A)/N_A^{tot}}$ (see for example[1]).

If N_A^{tot} is in the tens and above, R_A also follows approximately a normal distribution. This approximation (normally a very good one for our study which involves large data sets) facilities the calculation of the 95% confidence interval associated with R_A. If $\hat{R}_A$ is the estimator of the "true" value R_A, the standardized variable of the statistics is $Z = \dfrac{\hat{R}_A - R_A}{\sqrt{R_A(1-R_A)/N_A^{tot}}}$. As such, the 95% probability of finding $\hat{R}_A$ within an interval is $P(-z_{\alpha/2} < Z < z_{\alpha/2}) \approx 0.95$, where $z_{\alpha/2}$ is the 95% limit of the normal distribution. In order to find the confidence limits of R_A, the inequality "<" can be replaced by an equality "=" and the resulting equation can be solved for R_A giving:

$$R_A = \frac{\hat{R}_A + \dfrac{z_{\alpha/2}^2}{2N_A^{tot}} \pm z_{\alpha/2}\sqrt{\dfrac{\hat{R}_A(1-\hat{R}_A)}{N_A^{tot}} + \dfrac{z_{\alpha/2}^2}{4(N_A^{tot})^2}}}{1 + z_{\alpha/2}^2 / N_A^{tot}}$$

Note that for large N_A^{tot}, $R_A \approx \hat{R}_A \pm z_{\alpha/2}\sqrt{\hat{R}_A(1-\hat{R}_A)/N_A^{tot}}$. Since at 95% level, $z_{\alpha/2}$ = 1.96, the confidence interval is $\mathrm{CI}_A = 1.96\sqrt{\hat{R}_A(1-\hat{R}_A)/N_A^{tot}}$, which depends on both $\hat{R}_A$ and N_A^{tot}.

As an illustration, if $\hat{R}_A$ = 0.90 and the sample size is N_A^{tot} = 100, CI_A = 0.06 or, in other words, $R_A = 0.90 \pm 0.06$. This example shows that the degree of uncertainty at a 95% confidence interval is reasonably small even for a small sample size of a 100.

The same analysis can be applied to R_{AB}. In the case of the cross-influence index $\eta_{AB} = R_{AB} / R_A$, a common engineering approach[2] to assess the propagation of uncertainty in the division is $CI_\eta = \eta_{AB}\left(\dfrac{CI_A}{R_A} + \dfrac{CI_{AB}}{R_{AB}}\right)$. This calculation is applied in Figures 3 and 4 of the main text.

Similarly, for the case where specific reactions are taken into account, the same equation can be used $CI_{\eta}^{S} = \eta_{AB}^{S}\left(\frac{CI_{A}^{S}}{R_{A}^{S}} + \frac{CI_{AB}^{S}}{R_{AB}^{S}}\right)$, where the subscript "*S*" indicates performing the same calculation for the respective parameters with the reduced values of N_{A}^{S} and N_{AB}^{S} obtained from the searches for the specific reactions.

3. Further examples where functional group **B** "deactivates" **A**

In Figure 3a, b, examples of functional groups **B** activating **A** ($\eta_{AB} > 1$) have been discussed. When **B** deactivates **A** ($\eta_{AB} < 1$), the chemical reasoning is varied. The first example in Figure S2 shows that **B** = – NO_2 group deactivates **A** = benzene ring. This is in good agreement with common chemical knowledge where strongly electron withdrawing nitro group deactivates aromatic systems by decreasing their electron density. The **A** = bromide, **B** = benzene example illustrates the conjugation of electrons from bromide into the benzene ring, resulting in a less reactive bromide group (as compared to alkyl bromide) toward nucleophilic attack. The other two examples demonstrate the high reactivity of the neighboring group **B** surpassing the reactivity of **A**, and so, effectively, deactivating group **A**: In these examples, nucleophilic attack on neighboring acyl chloride and aldehyde groups is more likely than at group **A** (ketone and ester). When **A** and **B** are separated by one carbon atom, deactivation of **A** by electron withdrawing/ donating effects on **B** is less pronounced, since they are spatially further apart. Instead, it can be argued that in these examples, the high reactivity of groups **B** (alkyne, acyl chloride, bromide) toward nucleophilic attack reduces the effective reactivity of **A**. It should be noted here that we discuss mainly the reactivity towards nucleophilic attacks for the previous examples since the most encountered examples of the reactions involving these groups (substitutions of halides, additions to double, triple bonds and carbonyls, as well as reduction reactions) can be viewed, in general, as nucleophilic attacks.

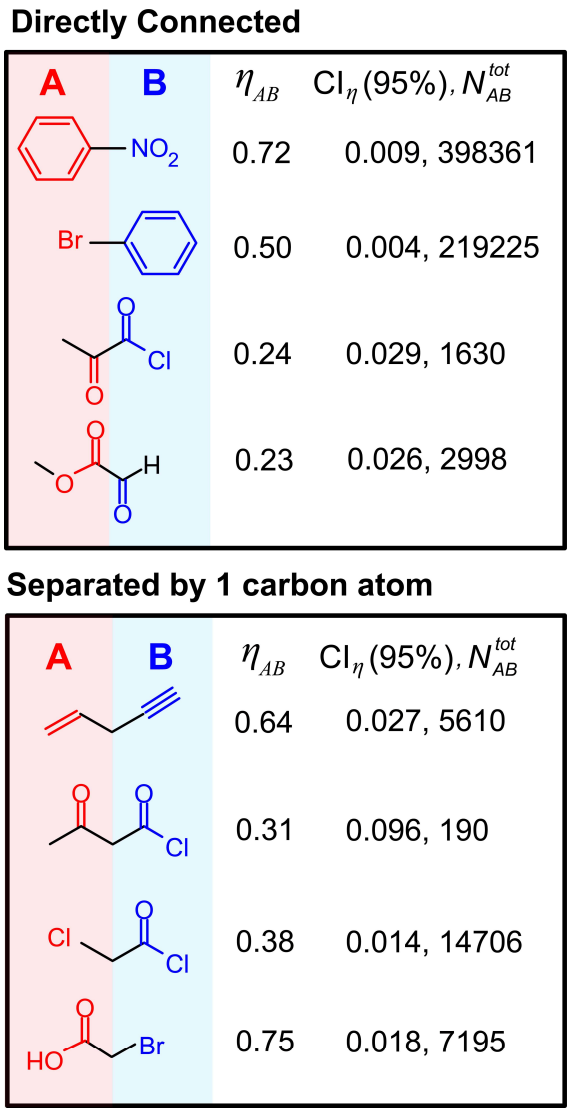


Fig. S1 Examples of the cross-influence index, η_{AB} for cases when functional group **B** deactivates **A** ($\eta_{AB} < 1$).

4. Further examples of effects of reactivity on η_{AB}

As discussed in the main text, when **A** is a highly reactive group itself (e.g., acyl chloride in Fig. 3c), less reactive groups **B** have a marginal effect on the already-high reactivity of **A**. Consequently, the η_{AB} indices are close to unity. This analysis was performed for functional groups **A** and **B** which are directly connected to each other. Figure S3 shows qualitatively similar effects for the case when the functional groups are separated by one carbon atom. As expected, the more reactive group (here, aldehyde), has η_{AB} indices closer to unity than the less reactive ester.

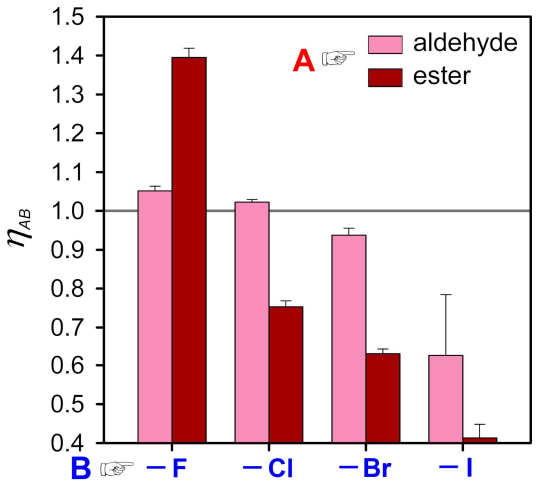


Fig. S2 Effects of functional group's reactivity on the cross-influence index, η_{AB} for the case when groups **A** and **B** are separated by one carbon atom (as opposed to Fig. 3c in the main text, where **A** and **B** are directly connected). The η_{AB} indices for the more reactive group **A** (here, aldehyde), are consistently closer to unity than those for the less reactive ester.

5. Two additional examples of diene activation in Diels-Alder reactions relevant to contemporary total synthesis.

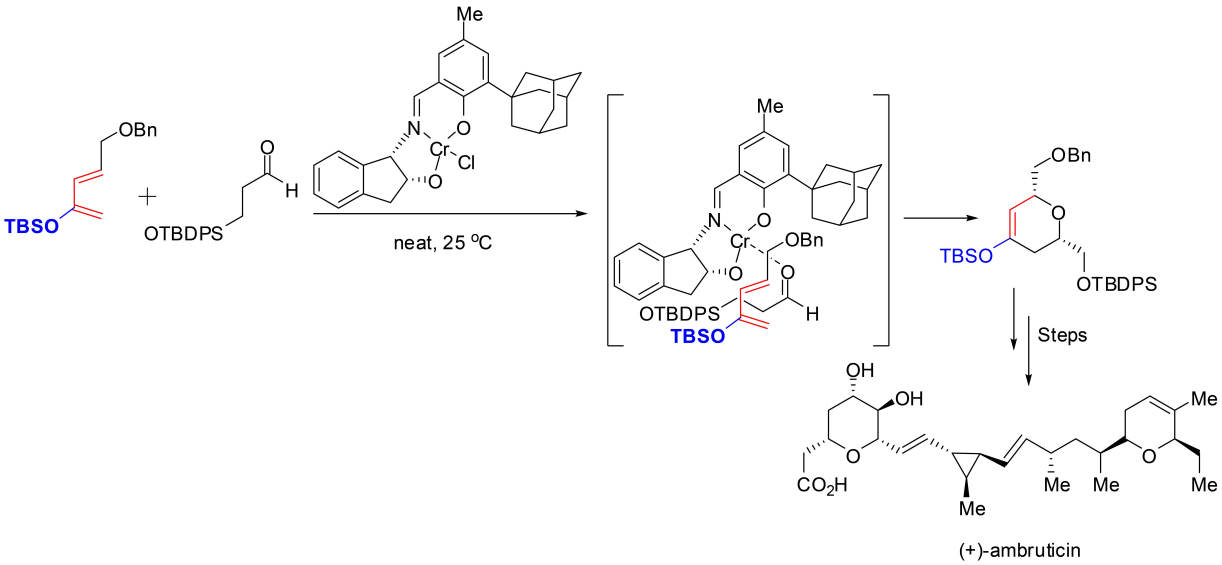


Fig. S3 This example shows the use of bulkier *t*-butyl silyloxy group (TBS) to favor s-cis conformation of diene to form (+)-ambruticin.[3] Alkyl silyloxyl groups are predicted as an activating group (η_{AB} = 1.54) by our ChemGPS.

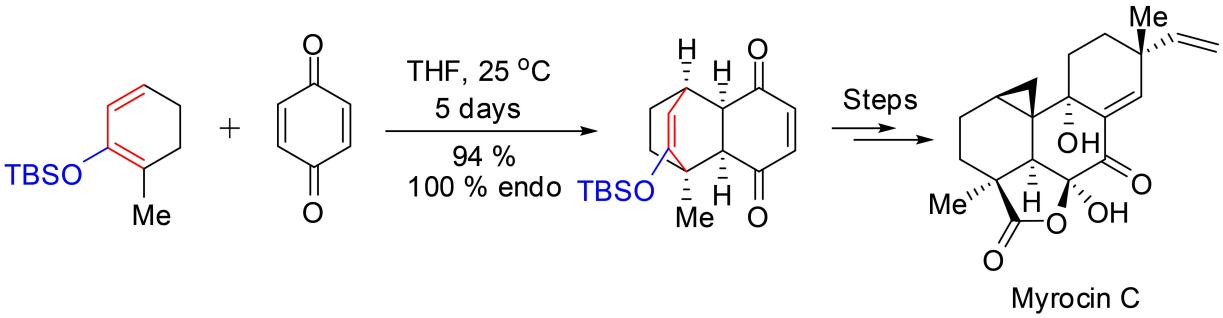


Fig S4 This example shows the use of two electron donating groups methyl and silyloxy to form Myrocin C (an antitumor pentacyclic diterpene)[4,5]

References for Supplementary Information:

1. J. L. Devore, *Probability and statistics: for engineering and the sciences*, Thomson Learning, London, 2004.
2. D. M. Himmelblau and J. B. Riggs, *Basic principles and calculations in chemical engineering*, Prentice Hall, New Jersey, 1989.
3. D. A. Evans, E. J. Olhava, J. S. Johnson and J. M. Janey, *Angew. Chem.-Int. Edit.*, 1998, **37**, 3372-3375.
4. K. C. Nicolaou, S. A. Snyder, T. Montagnon and G. Vassilikogiannakis, *Angew. Chem.-Int. Edit.*, 2002, **41**, 1668-1698.
5. D. Craig, J. J. Shipman and R. B. Fowler, *J. Am. Chem. Soc.*, 1961, **83**, 2885-2891.