Quantitative structure–reactivity relationships for synthesis planning: The benzhydrylium case

Maike Vahl,[†] Johannes V. Diedrich,[†,‡] Maike Mücke,[†,‡] and Jonny Proppe[*,†]

[†]*Institute of Physical and Theoretical Chemistry, TU Braunschweig, Braunschweig*

[‡]*Institute of Physical Chemistry, Georg-August University, Göttingen*

E-mail: j.proppe@tu-bs.de

Abstract

Selective and feasible reactions are top targets in synthesis planning, both of which depend on the reactivity of the molecules involved. Mayr's approach to quantifying reactivity has greatly facilitated the planning process, but reactivity parameters for new compounds require time-consuming experiments. In the past decade, data-driven modeling has been gaining momentum in the field as it shows promise in terms of efficient reactivity prediction. However, state-of-the-art models use quantum chemical data as input, which prevents access to real-time planning in organic synthesis. Here, we present a novel data-driven workflow for predicting reactivity parameters of molecules that takes only structural information as input, enabling *de facto* real-time reactivity predictions. We use the well-understood chemical space of benzhydrylium ions as an example to demonstrate the functionality of our approach and the performance of the resulting quantitative structure–reactivity relationships (QSRRs). Our results suggest that it is straightforward to build low-cost QSRRs that are accurate, interpretable, and transferable to yet unexplored systems within a given scope of application.

Introduction

The ability to predict whether two compounds will react and, if so, how fast, is essential for synthesis planning. The estimation of relative rates and hence, selectivity, is equally vital as they determine the likelihood of unwanted side reactions taking place during the synthesis process. Knowledge of these fundamental variables is rooted in the reactivity of the molecules involved. They allow chemists to make informed decisions about which reactions to pursue, thereby saving time, resources, and effort in the laboratory. Traditionally, determining reactivity has relied on experimental trial and error or referencing existing literature and databases, which is time-consuming and limits the scope of exploration.[1,2]

By leveraging advances in hardware, algorithms, and data science, a plethora of new efficient tools for planning organic syntheses has become available in the past two decades.[3–11] These new techniques can rapidly provide valuable insights into the reactivity of molecules, enabling chemists to make informed decisions during routine synthesis design. This approach has the potential to significantly accelerate the discovery and development of novel compounds, drugs, and materials, benefiting a wide range of scientific disciplines.

Our group currently explores the feasibility of real-time prediction of reactivity parameters by quantitative structure–reactivity relationships (QSRRs).[12–14] We aspire to build an interactive platform on which users can query arbitrary organic compounds and receive instant feedback, including site-specific reactivity information and uncertainty estimates[15] to ensure reliability und practical benefit. With the ability to assess reactivity in real-time, chemists can efficiently evaluate a vast number of potential reactions and choose the most promising ones for further investigation.

Here, we present a proof of principle using the chemical space of benzhydrylium ions as an example. The benzhydrylium ion and its derivatives write a success story in terms of quantifying chemical reactivity. Driven by the attempt to systematize the use of carbocations in organic synthesis, Mayr and coworkers studied reactions of olefins with benzhydrylium ions.[16,17] Mayr's team was astonished when they found that the relative reactivity of most

alkenes is independent of the reactivity of the benzhydrylium ion they react with.[18,19] Eventually, Mayr and Patz proposed a simple expression containing only three empirical parameters to compute the rate constant of polar bimolecular reactions in solution,[20]

$$\log k(20\ ^\circ\mathrm{C}) = s_\mathrm{N}(N + E) \tag{1}$$

Here, E, N, and s_N represent electrophilicity, nucleophilicity, and a nucleophile-specific sensitivity parameter, respectively. As they proceeded, Mayr and his team found that the Mayr–Patz equation (1) is also valid for many other classes of nucleophiles and electrophiles. To date, reactivity parameters have been determined for 352 electrophiles (E) and 1264 nucleophiles (N, s_N), which can be accessed via *Mayr's Database of Reactivity Parameters*.[21,22] A brief explanation of how these parameters are determined experimentally[23,24] is given in Boxes 1 and 2 of ref.[14]

Because synthesis and kinetic experiments are time-consuming and resource-intensive, attempts have been made to determine reactivity parameters by thermochemical calculations based on density functional theory (DFT). However, they have not yet prevailed over the experimental approach, also because of accuracy issues. In a recent uncertainty quantification study,[15] we confirmed that the average accuracy of experimental rate constants corresponding to reactions of olefins with benzhydrylium ions is higher — deviation in k below one order of magnitude — than that achievable with standard DFT calculations. Even high-performing functionals result in average barrier height errors of at least 2 kcal mol^{-1},[25] translating to a deviation in k of one to two orders of magnitude at 20 °C assuming validity of the Eyring equation.[26] Ultimately, neither of the two approaches (experiment vs. DFT) is suitable for the efficient prediction of reactivity parameters. This is one reason why data-driven or machine-learning (ML) algorithms have gained much attention in this context as they are capable to yield fast predictions by interpolating between available data.[14]

In supervised ML, relationships between descriptors (input variables) and targets (output

variables) are learned by means of regression (continuous target) or classification (discrete target). Aside from the expensive acquisition of targets (i.e., experimental reactivity parameters), the generation of descriptors can constitute a critical bottleneck of the ML workflow. For instance, previous data-driven studies have mostly relied on quantum molecular properties (QMPs) as descriptors,[27–37] meaning that each prediction is preceded by quantum chemical calculations, which occupy almost 100% of the overall prediction time.

While QMPs are among the most informative descriptors,[33] we target *fast* descriptor generation that avoids quantum chemical calculations as much as possible. For this purpose, we focus on *structural* descriptors in this work. Structural descriptors are direct representations of the connectivity/graph or the three-dimensional structure of a molecule.[38] There are two principal types of structural descriptors: General (application-agnostic) descriptors, which are applicable to a broad range of structure classes, but rather difficult to interpret. On the other hand, application-specific descriptors are rather simple to interpret but not generalizable to cases outside the domain of application. Here, we investigate the merits and drawbacks of both types of descriptors.

In general, structural descriptors are much higher in dimensionality than QMPs. As a rule of thumb, the greater the dimensionality of a descriptor, the more data is needed to uncover the underlying QSRR. However, the number of reactivity parameters in Mayr's database is limited. Therefore, we hypothesize that the structural descriptors examined here are too high-dimensional to be directly linkable with the relatively small number of available reactivity parameters. To meet this challenge, we propose a two-step workflow for building QSRR models from structural descriptors (Fig. 1).

Assume a set of K available reactivity parameters that is too small to build an accurate QSRR model based on a high-dimensional structural descriptor. Further assume that a set of $L \gg K$ reactivity parameters would be necessary to achieve the desired accuracy. Then, if we could identify a less expensive surrogate quantity that correlates well with the reactivity parameter of interest, it would be possible to build such a two-step QSRR model. In this

work, we propose QMPs to serve as surrogate quantities. The use of QMPs may seem like a contradiction to the goal of avoiding them, as stated above. However, given a set of M molecules of interest, only $L \ll M$ of which are equipped with QMPs, we have avoided (multiples of) $M - L$ quantum chemical calculations, leading to substantial computational savings.

Summarizing: In step 1, high-dimensional structural descriptors are linked with a small number of QMPs. The training set size of step 1 is L. In step 2, the same QMPs are linked with the actual reactivity parameters. The training set size of step 2 is K. Both steps are based on multivariate linear regression (MLR)[39] to facilitate interpretation of the results. The group of Sigman has popularized the method for physical organic chemistry[40] and it has already been applied by Orlandi et al.[35] for predicting and understanding Mayr's nucleophilicity parameter N.

In this work, we apply the novel two-step QSRR workflow to a dataset of $M = 3570$ benzhydrylium ions (Fig. 3), for only $K = 27$ of which an electrophilicity parameter E is available. At the same time, these 27 systems cover a wide range of reactivity, $-10.04 < E < 8.02$, spanning almost 20 orders of magnitude. Their electrophilic center, an carbenium ion, can be tuned by distant substituents. As a result, the reactivity of benzhydrylium ions can be dominantly attributed to electronic effects, leading to unambiguous E parameters. These electrophiles are therefore particularly well suited for building quantitative nucleophilicity scales for a variety of organic compounds.[1]

After an overview of the data and methods used in this work, the potential of our two-step workflow in terms of real-time reactivity prediction is evaluated. In particular, the MLR models for both steps of the workflow are analyzed with respect to performance and interpretability.

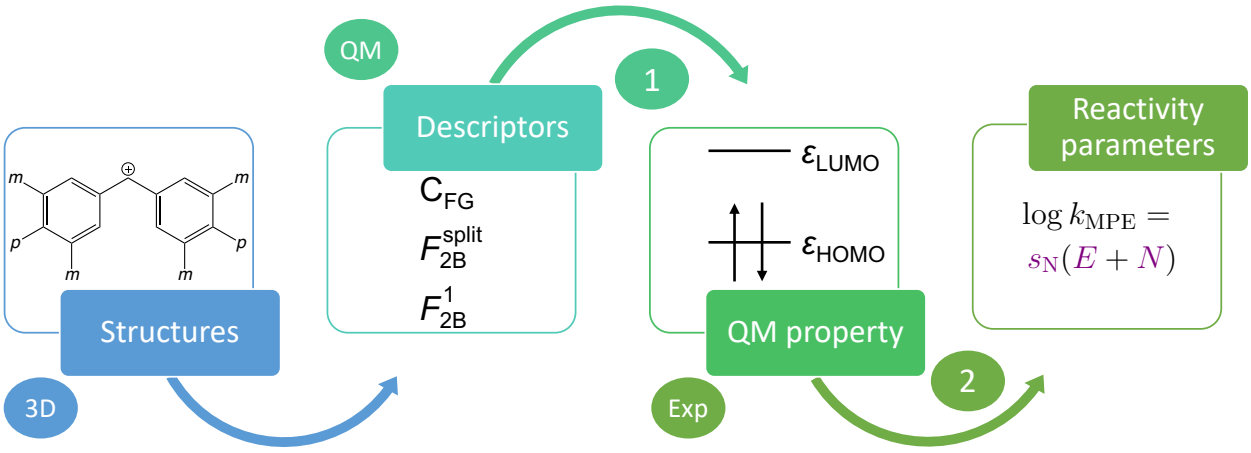

Figure 1: Schematic description of our two-step workflow for predicting reactivity parameters from molecular structures. In step 2 of the workflow, the relationship between a suitable quantum molecular (QM) property and the reactivity parameter of interest is learned. Since QM properties require expensive quantum chemical calculations, we seek to replace their calculation in step 1 of the workflow. For this purpose, the three-dimensional (3D) molecular structures are initially transformed into machine-learnable descriptors, before the relationships between these descriptors and the QM property are learned.

Methods

Data set

Mayr's database[21,22] comprises electrophilicity parameters for 33 benzhydrylium ions, six of which are annulated and therefore removed for the following analysis. Table 1 and Fig. 2 show the remaining $K = 27$ electrophiles.

For this study, a combinatorial data set of benzhydrylium ion derivatives was generated based on the unsubstituted ion **19**. Its four *meta* (m) and two *para* (p) positions, which are shown in Fig. 3, are suitable substitution sites. By avoiding substitution of the *ortho* positions, the steric situation at the carbenium ion is preserved and, hence, its electrophilicity is predominantly caused by electronic substituent effects. We considered only substituents of benzhydrylium ions available in Mayr's database (see Table 1), 13 in total: –F, –Cl, –Me, –OMe, –OPh, –N(Me)$_2$, –N(Me)(Ph), –N(Ph)$_2$, –*N*-pyrrolidino, –*N*-morpholino, –

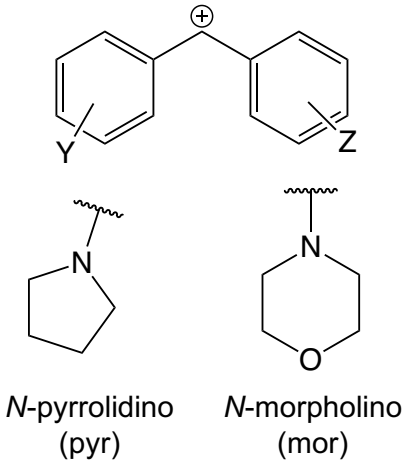

Figure 2: Top: The benzhydrylium scaffold with substituents Y and Z, which are specified in Table 1. Bottom: Lewis structures of the N-pyrrolidino and N-morpholino substituents.

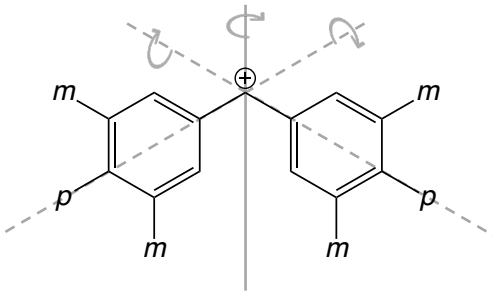

Figure 3: The benzhydrylium scaffold with emphasized *meta* positions m and *para* positions p. Rotation axes leading to (approximately) isoenergetic structures are displayed in gray.

N(Me)(CH_2CF_3), –N(Ph)(CH_2CF_3), and –CF_3. The Lewis structures of the –N-pyrrolidino and –N-morpholino groups are shown in Fig. 2. Only –F and –Cl were selected as possible m-substituents to avoid steric hindrance with the p-substituents, while all of the above-mentioned substituents were selected as possible p-substituents.

Next, all possible substitution combinations of these functional groups were generated, leading to $3^4 \cdot 14^2 = 15876$ structures (counting hydrogen as 3^{rd} m-substituent and 14^{th} p-substituent, respectively). If structures could be converted into each other by the C_2 rotation axis, shown as gray solid line in Fig. 3, only one of them was kept. In addition, tests have shown that assuming a symmetry axis passing through the bond between the carbenium ion and the aromatic rings (dashed gray lines in Fig. 3) is a reasonable approximation (see

Table 1: Benzhydrylium ion derivatives of Mayr's database considered in this work. Substituents Y and Z (cf. Fig. 2) as well as electrophilicity parameters E^1 are listed. See Table S6 for reference electrophile names.

ID	Y	Z	E^1	ID	Y	Z	E^1
1	4-(*N*-pyrrolidino)	Y	−7.69	**15**	4-Me	H	4.43
2	4-N(Me)$_2$	Y	−7.02	**16**	4-F	Y	5.01
3	4-N(Me)(Ph)	Y	−5.89	**17**	4-F	H	5.20
4	4-(*N*-morpholino)	Y	−5.53	**18**	3-F, 4-Me	Y	5.24
5	4-N(Ph)$_2$	Y	−4.72	**19**	H	Y	5.47
6	4-N(Me)(CH$_2$CF$_3$)	Y	−3.85	**20**	4-Cl	Y	5.48
7	4-N(Ph)(CH$_2$CF$_3$)	Y	−3.14	**21**	3-F	H	6.23
8	4-OMe	Y	0.00	**22**	4-(CF$_3$)	H	6.70
9	4-OMe	4-OPh	0.61	**23**	3,5-F$_2$	H	6.74
10	4-OMe	4-Me	1.48	**24**	3-F	Y	6.87
11	4-OMe	H	2.11	**25**	3,5-F$_2$	3-F	7.52
12	4-OPh	4-Me	2.16	**26**	4-(CF$_3$)	Y	7.96
13	4-OPh	H	2.90	**27**	3,5-F$_2$	Y	8.02
14	4-Me	Y	3.63				

SI Section "Examination of rotational symmetry"). The resulting duplicate molecules were removed as well. The final data set therefore consists of $M = 3570$ structures, in which the 27 aforementioned reference structures are present.

Descriptors

For the data set under investigation, two problem-specific descriptors have been developed. All descriptors considered in this work are represented as vectors, the individual elements of which are referred to as *features*. See SI Section "Descriptor properties" for useful requirements for the development/choice of a suitable descriptor.

The counting descriptor $\mathbf{C_{FG}}$ reflects the number of each functional group (FG) at the *meta* positions (group 1) and the *para* positions (group 2), as well as the number of substituent combinations regarding the *meta* positions located at the same ring (group 3) and regarding the *para* positions adjacent to *meta* positions (group 4). The last two groups ensure that the descriptor is a unique description of the substitution pattern.

In Fig. 4, a schematic description of C_{FG} is shown. The hydrogen atom is neglected as

FG. Considering all possible substituents ($m = 2$ and $p = 13$), the descriptor dimension is composed of $m = 2$ features for group 1, $p = 13$ features for group 2, $m \cdot (m + 1)/2 = 3$ features for group 3, and $m \cdot p = 26$ features for group 4. This application-specific descriptor features 44 dimensions. It can only be applied to this specific data set. At the same time, it is an easy-to-interpret descriptor.

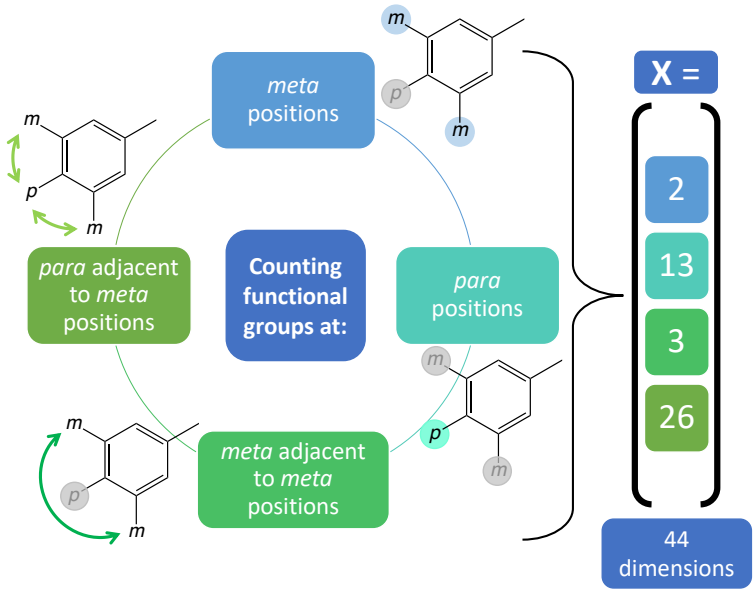

Figure 4: Schematic description of the counting descriptor C_{FG}. The number of occurrences is counted individually for every substituent or combination of substituents according to the categories shown. The right-hand side represents the number of descriptor dimensions (44 in total) occupied by the different categories.

The original F_{2B} descriptor was proposed by Pronobis et al.[41] and is a general descriptor including two-body interactions. It is specified for all possible element pairs in the data set including hydrogen atoms. For each unique element combination (x, y), the pairwise sum of inverse internuclear distances, $\{R_{ij}\}$, is calculated without double counting,

$$(F_{2B})_{(x,y)} = \begin{cases} \sum_{ij} \frac{1}{R_{ij}^n} , & x \neq y \\ \sum_{j>i} \frac{1}{R_{ij}^n} , & x = y \end{cases} \tag{2}$$

Therefore, the F_{2B} descriptor takes information on the 3-dimensional molecular structure

into account; as opposed to the C_{FG} descriptor. For more flexibility, the authors introduced different exponents, $n = \{1, ..., 15\}$, resulting in 15 descriptor dimensions per unique element pair. In Fig. 5, a schematic description of the F_{2B} descriptor is shown, which includes the Coulomb-type interactions ($n = 1$) only and is denoted F^1_{2B}. To keep the computational

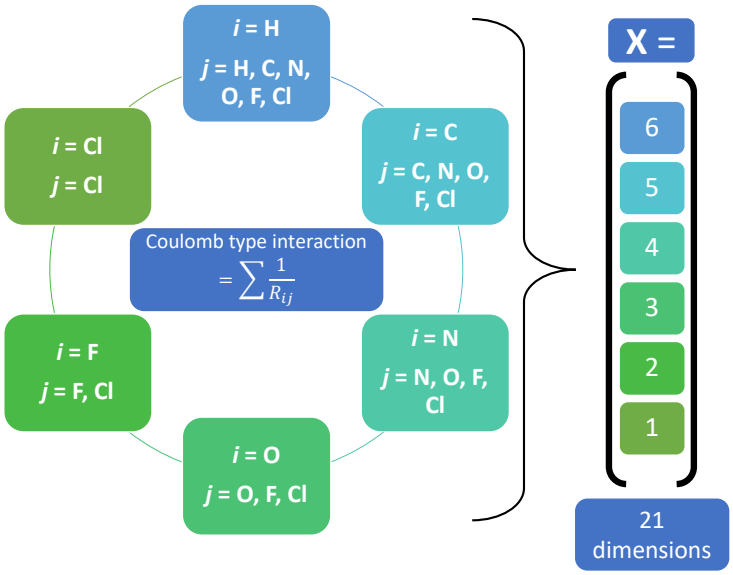

Figure 5: Schematic description of the F^1_{2B} descriptor. The sum of pairwise Coulomb-type interactions is calculated individually for every unique element pair. The right-hand side represents the number of descriptor dimensions (21 in total) occupied by the different categories.

cost of descriptor generation as low as possible, the three-dimensional molecular structures should not originate from expensive quantum chemical structure optimizations. If not otherwise mentioned, the generation of F_{2B}-type descriptors did not include quantum chemical calculations. (The C_{FG} descriptor is independent of the actual three-dimensional structure.) See SI Section "Structure generation" for a detailed description of the automated and quantum-chemistry-free generation of three-dimensional structures.

The F^{split}_{2B} descriptor is an adapted version of the F^1_{2B} descriptor created by us. To include more chemical information, the original F^1_{2B} descriptor is divided into different interaction groups resulting from the benzhydrylium scaffold. For example, carbon atoms

appear in the carbenium ion (C^+) as well as in the phenyl rings (C_{Ph}), and in different *p*-substituents (*p*-C). In F_{2B}^1, the interactions of these carbon atoms with a given second element are summed up into a single feature. By splitting them up in the new descriptor, the interactions are divided among different regions of the molecule, which further helps in the interpretation of results. The interaction groups are shown in Fig. 6. Hydrogen atoms are neglected in all of them. The descriptor dimensions sum up to 44 in total. The dimensionality of this descriptor only coincidentally equals that of the C_{FG} descriptor for the underlying data set. F_{2B}^{split} is an application-specific descriptor. It is easier to interpret than F_{2B}^1 but less universal.

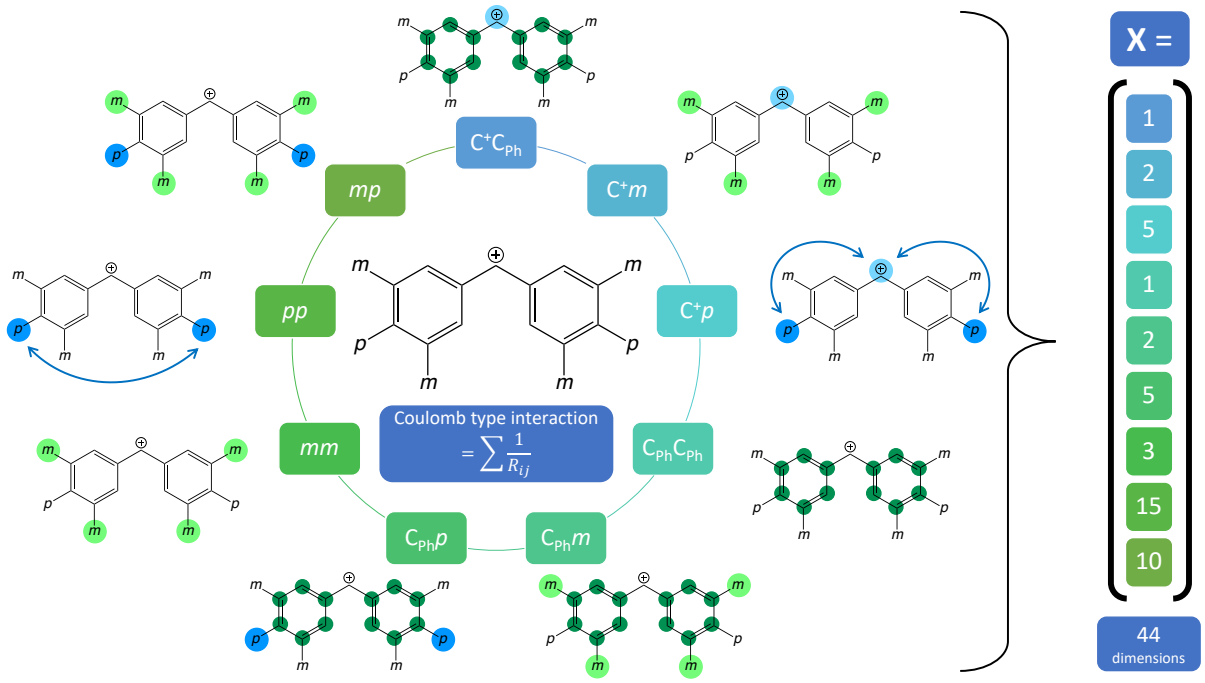

Figure 6: Schematic description of the F_{2B}^{split} descriptor. In each category, the sum of pairwise Coulomb-type interactions is calculated individually for every unique element pair represented by that category. The right-hand side represents the number of descriptor dimensions (44 in total) occupied by the different categories.

Quantum mechanical properties

We selected five QMPs based on conceptual density functional theory[42] as they yielded the most promising results in a data-driven investigation of electrophilicity by Hoffmann et al.[33]

As some of them represent compositions of simpler terms, we partitioned the five QMPs to yield eight QMPs in total, see Table 2. All of them are based on energies of frontier molecular orbitals (FMO), i.e., $\varepsilon_{\mathrm{HOMO}}$ and $\varepsilon_{\mathrm{LUMO}}$, which we obtained from either quantum chemical calculations (Section "Computational methods") or MLR predictions (Section "Results and discussion").

Table 2: List of quantum molecular properties (QMPs) examined in this study, including mathematical definitions. The numbering of the properties corresponds to the ranking determined by Hoffmann et al.[33] Asterisks indicate that the corresponding QMP is important for the definition of other QMPs of this list.

Name		Mathematical definition
Ionisation potential[43]	*	$\mu^{-}_{\mathrm{FMO}} = \varepsilon_{\mathrm{HOMO}}$
Electron affinity[43]	1	$\mu^{+}_{\mathrm{FMO}} = \varepsilon_{\mathrm{LUMO}}$
Global molecular hardness[44]	*	$\eta_{\mathrm{FMO}} = \varepsilon_{\mathrm{LUMO}} - \varepsilon_{\mathrm{HOMO}}$
Electronic chemical potential[45,46]	*	$\mu_{\mathrm{FMO}} = \frac{\mu^{+}_{\mathrm{FMO}}+\mu^{-}_{\mathrm{FMO}}}{2}$
Electrophilicity index[47]	2	$\omega_{\mathrm{FMO}} = \frac{\mu^{2}_{\mathrm{FMO}}}{2\eta_{\mathrm{FMO}}}$
Electroaccepting power[48]	4	$\omega^{+}_{\mathrm{FMO}} = \frac{(3\varepsilon_{\mathrm{LUMO}}+\varepsilon_{\mathrm{HOMO}})^2}{16\mu_{\mathrm{FMO}}}$
Electrodonating power[48]	5	$\omega^{-}_{\mathrm{FMO}} = \frac{(\varepsilon_{\mathrm{LUMO}}+3\varepsilon_{\mathrm{HOMO}})^2}{16\mu_{\mathrm{FMO}}}$
Net electrophilicity[49]	3	$\Delta\omega^{\pm}_{\mathrm{FMO}} = \omega^{+}_{\mathrm{FMO}} + \omega^{-}_{\mathrm{FMO}}$

Metrics

The metrics taken into account in this work are specified with respect to N observations $\{y_i\}$ and corresponding predictions $\{\hat{y}_i\}$. An observation y_i refers to either the electrophilicity parameter E or a QMP of the i^{th} molecule. The mean of $\{y_i\}$ and $\{\hat{y}_i\}$ is denoted $\bar{y}$ and $\bar{\hat{y}}$, respectively. The root-mean-square error (RMSE) is defined as

$$\mathrm{RMSE} = \sqrt{N^{-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \geq 0 \tag{3}$$

The coefficient of determination,[50]

$$R^2 = 1 - \frac{\mathrm{RMSE}^2}{N^{-1}\sum_{i=1}^{N}(y_i - \bar{y})^2} \ , \quad R^2 \in (-\infty, 1] \ , \tag{4}$$

is a strictly monotonically decreasing function of the RMSE. Both RMSE and R^2 are performance metrics. Pearson's correlation coefficient,[51]

$$r = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}} \ , \quad r \in [-1, 1] \ , \tag{5}$$

on the other hand, is a correlation metric. While especially its squared form, $r^2 \in [0, 1]$, is often used as a performance metric, we emphasize that this may be a misconception. Even if two quantities correlate perfectly with each other ($r^2 = 1$), the corresponding R^2 value can be arbitrarily smaller than 1 due to a constant systematic error. Only if the least-squares solution of a linear regression problem is considered, r^2 equals R^2.[50]

Computational methods

The `structure-generator` program was employed for the combinatorial generation of the data set structures in XYZ format with Python (version 3.9.7). After preoptimization with the xTB software (version 6.5.1)[36,52] using GFN2-xTB,[52] CREST (version 2.12)[53,54] was employed to search for the most stable conformer of each molecule with the same settings as before. Full structure optimizations were then carried out with the ORCA program (version 5.0.3)[55,56] using the hybrid meta-GGA exchange–correlation functional TPPSh[57,58] and the D3 dispersion correction with the Becke–Johnson damping function.[59,60] (Note that the optimized structures are not required for the generation of the structural descriptors, see SI Section "Comparison of descriptors: guess structures versus relaxed structures".) The def2-SVP basis set[61] was employed as well as the auxiliary basis set def2/J with the Coulomb integral approximation RIJCOSX.[62] Preliminary tests motivating the choice of functional and basis set are given in SI Section "Development of a quantum chemical protocol".

The subsequent descriptor calculations were performed with self-written Python code, which can be accessed through the project-related GitLab repository.[63] After preprocessing the data with Scikit-learn 0.24.2,[64] ordinary least-squares MLR was performed with the same

package.

Results and discussion

The second step (QMP to *E*)

As described in Section "Quantum mechanical properties", eight QMPs were selected, see Table 2. Instead of selecting the single best QMP for our purposes, we propose to use a linear combination of all linearly independent terms contained in the eight pre-selected QMPs (six in total) to build an estimate of the electrophilicity parameter,

$$\begin{aligned} \hat{E} := w_0 + w_1\varepsilon_{\mathrm{L}} + w_2\varepsilon_{\mathrm{H}} + w_3\varepsilon_{\mathrm{L}}^2 + w_4\varepsilon_{\mathrm{H}}^2 \\ + w_5\varepsilon_{\mathrm{L}}\varepsilon_{\mathrm{H}} + w_6(\varepsilon_{\mathrm{L}} - \varepsilon_{\mathrm{H}})^{-1} \approx E \end{aligned} \tag{6}$$

Here, we abbreviated $\varepsilon_{\mathrm{HOMO}}$ and $\varepsilon_{\mathrm{LUMO}}$ as ε_{H} and ε_{L}, respectively. The coefficients w_0 to w_6 were determined by ordinary least-squares MLR and represent the intercept (w_0) and the weights of the linearly independent terms (w_1 to w_6). For training, experimental E parameters of the $K = 27$ reference systems were utilized. In the following, we refer to the optimized model (eq 6) as reference MLR (rMLR) model. Its predicitions $\hat{E}$ approximate the actual electrophilicity parameter E, which is unknown for $M - K = 3543$ of the $M = 3570$ structures considered here. The relative impact of each coefficient $w_{i>0}$ was determined by $|w_i| / \sum_{j>0} |w_j|$ and the results are summarized in Table 3. The coefficients can be directly compared to each other due to standardization of frontier molecular orbital energies ε_{F} (F = L, H),

$$\varepsilon_{\mathrm{F},i} = \frac{e_{\mathrm{F},i} - \mu_{\mathrm{F}}}{\sigma_{\mathrm{F}}} \tag{7}$$

Here, $e_{\mathrm{F},i}$ is the raw frontier molecular orbital energy for the i^{th} molecule obtained from quantum chemical calculations, and μ_{F} and σ_{F} represent mean and standard deviation of raw frontier molecular orbital energies, respectively, for the reference systems. As a consequence,

Table 3: Relative impact, $|w_i|/\sum_j |w_j|$, of the coefficients w_1 to w_6 of the reference MLR (rMLR) model shown in eq (6). The coefficients are dimensionless as we used standardized frontier molecular orbital energies. Pearson's coefficient r refers to the correlation between the respective term and Mayr's E parameter.

Coefficient	Term	Value	Impact	r
w_1	$\varepsilon_{\mathrm{LUMO}}$	-38.1	21.1%	-0.986
w_2	$\varepsilon_{\mathrm{HOMO}}$	$+22.6$	12.6%	-0.968
w_3	$\varepsilon^2_{\mathrm{LUMO}}$	-58.3	32.3%	$+0.980$
w_4	$\varepsilon^2_{\mathrm{HOMO}}$	-6.8	3.8%	$+0.965$
w_5	$\varepsilon_{\mathrm{LUMO}} \cdot \varepsilon_{\mathrm{HOMO}}$	$+54.3$	30.1%	$+0.977$
w_6	$(\varepsilon_{\mathrm{LUMO}} - \varepsilon_{\mathrm{HOMO}})^{-1}$	$+0.2$	0.1%	-0.740

$\varepsilon_{\mathrm{F},i}$ is a dimensionless quantity.

The coefficient of $\varepsilon_{\mathrm{LUMO}}$, w_1, is quite impactful at 21.1% (third highest). The ranking by Hoffmann et al.,[33] shown in Table 2, even suggests $\varepsilon_{\mathrm{LUMO}}$ to be the most impactful among all quantities studied by them (928 in total). The highest relative impact at 32.3% was found for the coefficient of $\varepsilon^2_{\mathrm{LUMO}}$, w_3, and the second highest at 30.1% was found for the product of $\varepsilon_{\mathrm{LUMO}}$ and $\varepsilon_{\mathrm{HOMO}}$, w_5. Both terms were not directly considered in previous regression studies, but are included in the electrophilicity index ω_{FMO} (see Table 2), which has been studied in related contexts[27,29–34] and ranked second by Hoffmann et al. On the other hand, the coefficients associated with the third term of the numerator ($\varepsilon^2_{\mathrm{HOMO}}$) and the denominator ($(\varepsilon_{\mathrm{LUMO}} - \varepsilon_{\mathrm{HOMO}})^{-1}$) of ω_{FMO} are substanstially less impactful, with values of 3.8% (w_4) and 0.1% (w_6), respectively. We draw the conclusion that the highest-impact terms of this analysis play a predominant role in correlating the electrophilicity index ω_{FMO} with the E parameters of benzhydrylium ions.

Fig. 7 shows a plot of the rMLR-predicted $\hat{E}$ parameter versus its experimental analog E for the reference systems. The seven structures with the smallest E values all comprise nitrogen-bonded *para*-substituents. They are associated with a larger deviation of $\hat{E}$ from E than the other structures. We assume that either increased conformational flexibility or size-related increased repulsion with *meta*-substituents (relative to the other functional groups of the data set) is responsible for this trend. Overall, the statistical test set metrics,

$r^2 = R^2 = 0.992$ and RMSE = 0.450, indicate the success of the MLR approach. The

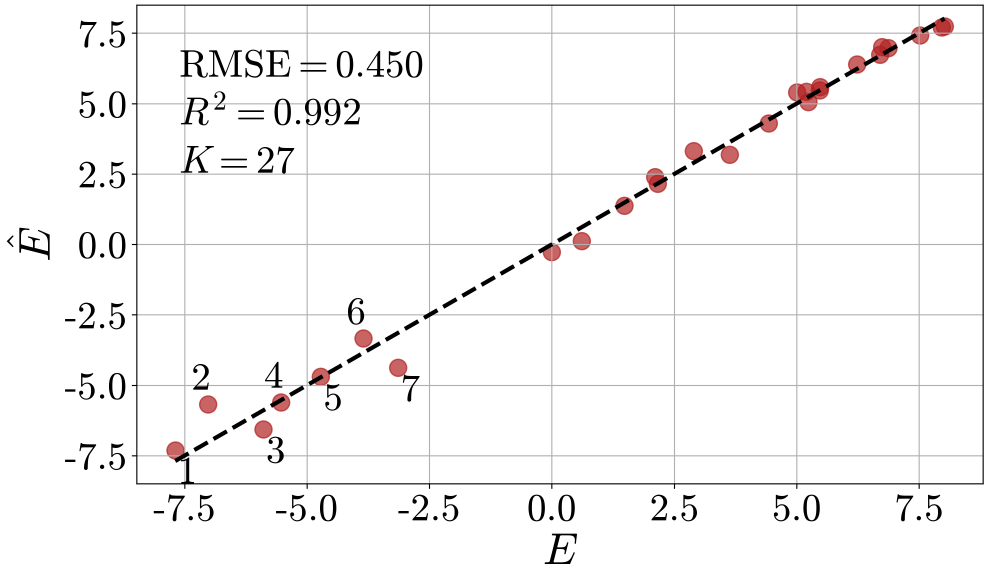

Figure 7: Predicted $\hat{E}$ versus Mayr's E for $K = 27$ reference structures. The coefficients of the rMLR model yielding $\hat{E}$ were optimized with respect to E for the same set of structures. Frontier molecular orbital energies obtained from quantum chemical calculations served as input. See eq (6) for the mathematical definition of the rMLR model.

optimized rMLR model provides a reasonable starting point for the implementation of the overall workflow. Additionally, given the high accuracy of the rMLR model paired with its superior interpretability, we decide against the application of more complex ML models such as neural networks,[65] Gaussian processes,[66] or gradient boosting decision trees.[67] The latter was found to excel other types of ML models in the prediction of E parameters for a range of electrophiles including mostly carbocations and Michael acceptors.[33]

The first step (structure to QMP)

The target of properly connecting the molecular structures with their associated reactivity parameters can be approached via two paths, A and B. In both cases, the goal is to replace the rMLR predictions $\hat{E}$, which require quantum chemical calculations of $\varepsilon_{\mathrm{LUMO}}$ and $\varepsilon_{\mathrm{HOMO}}$, with purely structure-based and hence real-time predictions, $\hat{E}^{\mathrm{A}}$ and $\hat{E}^{\mathrm{B}}$, respectively. We divided the data set into a test set, which includes the 27 reference structures, and a training

set comprising the remaining 3543 structures.

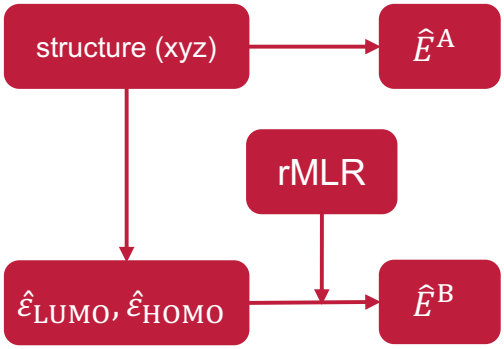

Figure 8: We propose two differents paths to approximate the rMLR model predictions $\hat{E}$, see eq (6), by structural descriptors. In path A, the descriptors are directly mapped onto $\hat{E}$. In path B, the descriptors a mapped onto $\hat{\varepsilon}_{\mathrm{LUMO}}$ and $\hat{\varepsilon}_{\mathrm{HOMO}}$, respectively, before they are plugged into the rMLR model.

Path A describes the establishment of a structure–$\hat{E}$ relationship. That is, the output of the rMLR model, $\hat{E}$, is learned directly by another MLR model, which takes structural descriptors instead of QMPs as input. We refer to the predictions of this model as $\hat{E}^{\mathrm{A}}$. The results are shown in Table 4.

Path B includes the training of two separate MLR models; one representing a structure–$\varepsilon_{\mathrm{LUMO}}$ relationship, the other one representing a structure–$\varepsilon_{\mathrm{HOMO}}$ relationship. We refer to the predictions of these models as $\hat{\varepsilon}_{\mathrm{LUMO}}$ and $\hat{\varepsilon}_{\mathrm{HOMO}}$, respectively. Since $\hat{E}$ is a function of only $\varepsilon_{\mathrm{LUMO}}$ and $\varepsilon_{\mathrm{HOMO}}$, substitution of the latter by their MLR-learned analogs ($\hat{\varepsilon}_{\mathrm{LUMO}}$ and $\hat{\varepsilon}_{\mathrm{HOMO}}$) leads to $\hat{E}^{\mathrm{B}}$. Note that the substitution does not alter the optimal coefficients w_0 to w_6 of the rMLR model. The test set performance with respect to $\hat{\varepsilon}_{\mathrm{LUMO}}$, $\hat{\varepsilon}_{\mathrm{HOMO}}$, and $\hat{E}^{\mathrm{B}}$ is reported in Table 4.

Table 4: Test set R^2 values ($K = 27$) for the prediction of $\hat{E}^{\mathrm{A}}$, $\hat{\varepsilon}_{\mathrm{LUMO}}$, $\hat{\varepsilon}_{\mathrm{HOMO}}$, and $\hat{E}^{\mathrm{B}}$ obtained via paths A and B of step 1, respectively, for all three descriptors. The best result is shown in bold for each quantity.

Descriptor	$\hat{E}^{\mathrm{A}}$	$\hat{\varepsilon}_{\mathrm{LUMO}}$	$\hat{\varepsilon}_{\mathrm{HOMO}}$	$\hat{E}^{\mathrm{B}}$
C_{FG}	0.418	**0.930**	**0.613**	0.972
$F_{\mathrm{2B}}^{\mathrm{split}}$	0.693	0.894	0.549	**0.984**
F_{2B}^{1}	**0.941**	0.761	0.449	0.906

Comparing the results of path A and B to each other, the R^2-values of path B are superior to those in path A except for the F_{2B}^{1} case. However, both F_{2B}^{split} and C_{FG} applied in path B excel F_{2B}^{1} in both paths. The highest performance is measured for F_{2B}^{split} in path B. The advantage of F_{2B}^{split} over F_{2B}^{1} not only is its performance but also its interpretability. The same is true for path B over path A as it allows us to understand the E parameter in terms of $\hat{\varepsilon}_{HOMO}$ and $\hat{\varepsilon}_{LUMO}$, respectively.

In Fig. 9, the predicted vs. experimental E parameters are shown for the final workflow settings (path B, F_{2B}^{split}) with respect to the 27 reference systems. A possibility to verify the

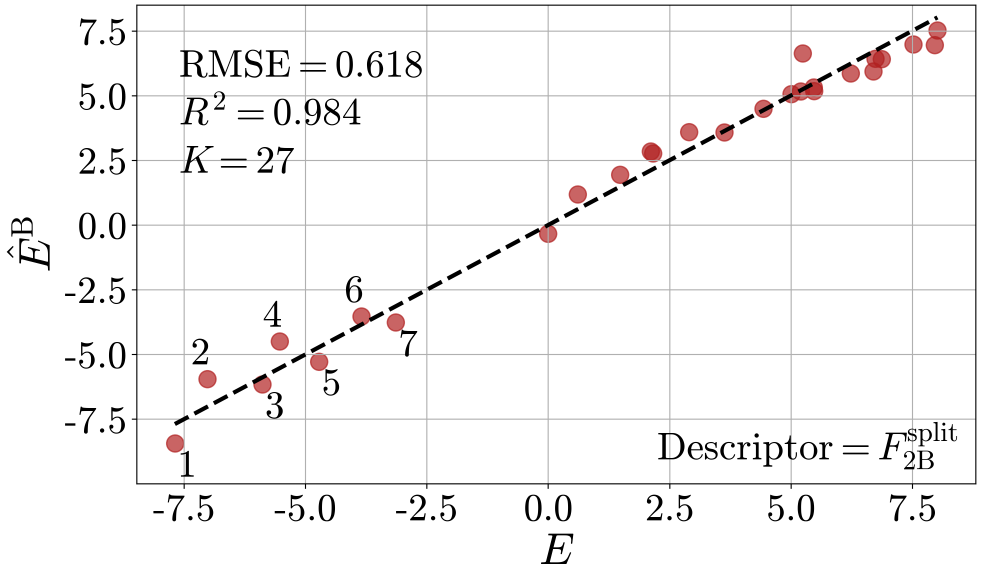

Figure 9: Predicted $\hat{E}^B(F_{2B}^{split})$ versus Mayr's E for $K = 27$ reference structures. The results are based on quantum chemical frontier molecular orbital energies for $M - K = 3543$ benzhydrylium ions.

quality of our approach is to exploit the strict relationship between r^2 and R^2. By coupling different MLR models, the strict relation $r^2 = R^2$ is no longer valid. However, the comparison of both values can be considered as a quality measure: The smaller the deviation between r^2 and R^2, the closer the result is to the least-squares solution. In this case, $r^2 = 0.985$ and $R^2 = 0.984$ are very close, although we use a prediction model which is based on another model, confirming the success of the two-step approach in a least-squares context.

Finally, we would like to know if we really need $M - K = 3543$ systems to obtain a

good prediction of E, or whether a substantially smaller number L is sufficient to identify a QSRR. Learning curves are instructive for this purpose. Due to the lack of reference data, we examine learning curves for $\varepsilon_{\mathrm{HOMO}}$ and $\varepsilon_{\mathrm{LUMO}}$ obtained from quantum chemistry. Since frontier molecular orbital energies have been shown to yield accurate estimates of E (in the form of $\hat{E}$), we consider them adequate surrogate quantities. The results for the $F_{\mathrm{2B}}^{\mathrm{split}}$ are shown in Fig. S8. Significantly steeper learning curves were obtained for the $\mathrm{C_{FG}}$ descriptor, which performed only slightly worse than $F_{\mathrm{2B}}^{\mathrm{split}}$ in the prediction of electrophilicity (path B), see Table 4. The results for $\mathrm{C_{FG}}$ (Fig. 10) suggest that $L \approx 150$ quantum chemical data points are necessary before robust and accurate predictions are obtained. In return, however, $M - L \approx 3420$ or $(M - L)/M \cdot 100\,\% \approx 96\,\%$ quantum-chemistry-free predictions of the electrophilicity parameter E can be made in real-time. This result represents proof of principle that real-time reactivity prediction is possible.

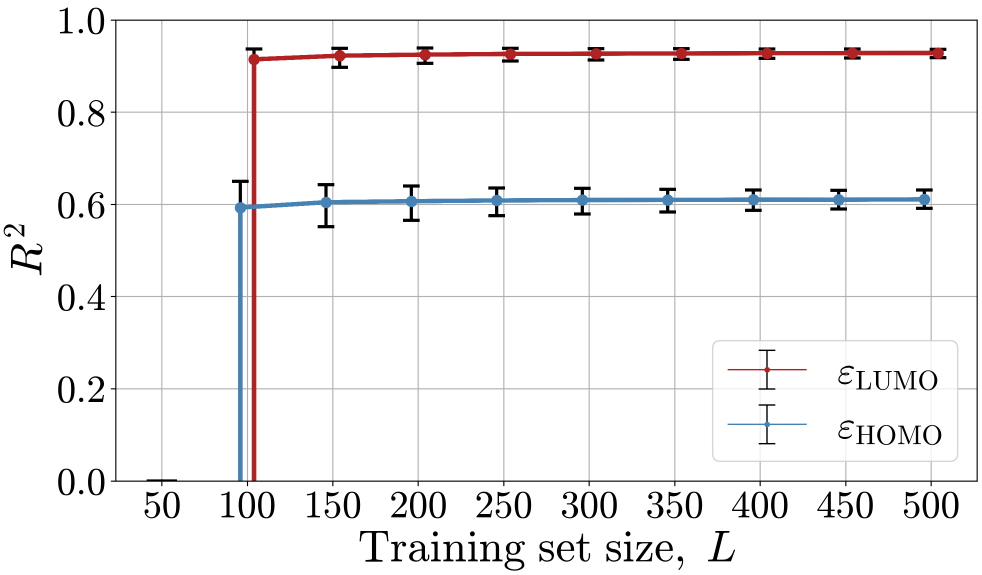

Figure 10: Learning curves for structure-$\varepsilon_{\mathrm{HOMO}}$ and structure–$\varepsilon_{\mathrm{LUMO}}$ relationships based on the $\mathrm{C_{FG}}$ descriptor. The median of the test set R^2 ($K = 27$) is shown (dots) for different training set sizes ($L = 50, 100, ..., 500$). For each size, 1000 MLR models were trained on randomly selected training samples. The error bars represent 95% confidence intervals. The results suggest that real-time reactivity prediction becomes robust and accurate at around $L = 150$.

The direct path (structure to E)

To verify the necessity of our two-step approach, we now examine whether structure–E relationships can be learned directly, i.e., without detouring through quantum chemical calculations. We used the descriptors considered in this work to train MLR models on all available E parameters, which is a small number ($K = 27$).

The results are shown in Fig. 11. The test set comprises the remaining $M - K = 3543$ structures of the data set. Due to the unavailability of experimental data for these structures, it was assumed that the rMLR model predictions $\hat{E}$ would provide adequate surrogate values for E. The test set metrics, $R^2 = 0.662$ and RMSE = 3.309, confirm that it is not sufficient

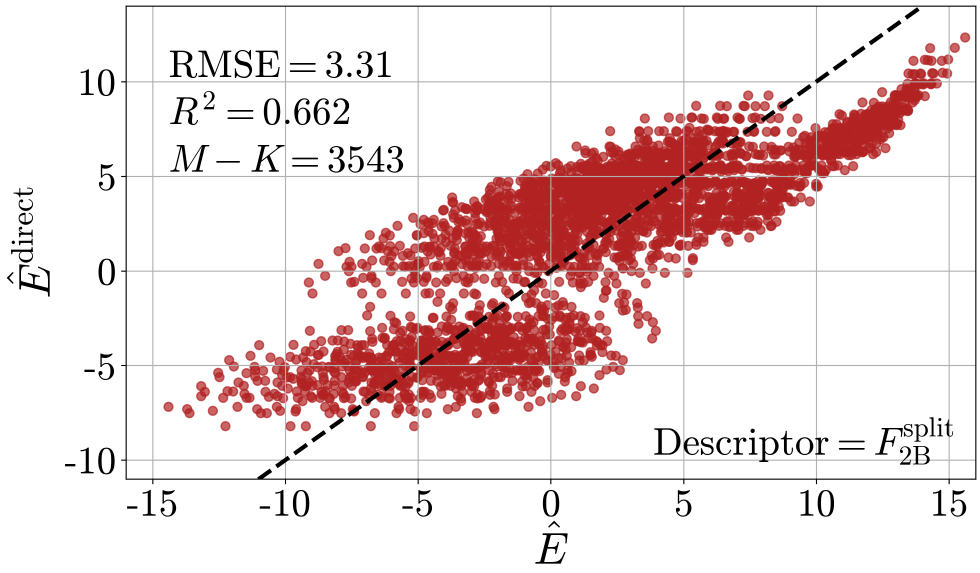

Figure 11: $\hat{E}^{direct}$ versus $\hat{E}$ for $M - K = 3543$ benzhydrylium ions. The results are based on a direct mapping of the F_{2B}^{split} descriptor onto Mayr's E for $K = 27$ reference structures.

to perform reliable predictions based on a training set including only 27 systems.

Chemical insight from linear coefficients

Interpretable models can offer valuable understanding of patterns in data. By grasping which features of a descriptor are important for making predictions, domain experts can gain deeper insights into the problem of interest and potentially make new discoveries.

Here, we are interested in understanding the quantitative and qualitative relationship between molecular structure (in the form of descriptors) and reactivity (E). Recall that we cannot use E directly due to lack of data. We also cannot use $\hat{E}^{\mathrm{A}}$ or $\hat{E}^{\mathrm{B}}$ instead. The correlation between $\hat{E}^{\mathrm{A}}$ and E is poor, and $\hat{E}^{\mathrm{B}}$ is linked with $\hat{\varepsilon}_{\mathrm{LUMO}}$ and $\hat{\varepsilon}_{\mathrm{HOMO}}$, but not with the structural descriptors. However, $\hat{\varepsilon}_{\mathrm{LUMO}}$ and $\hat{\varepsilon}_{\mathrm{HOMO}}$ are in turn linked with them. Taking into account the chemically intuitive correlation between $\varepsilon_{\mathrm{LUMO}}$ and E (see also column r in Table 3), we select $\hat{\varepsilon}_{\mathrm{LUMO}}$ over $\hat{\varepsilon}_{\mathrm{HOMO}}$ for the following analysis.

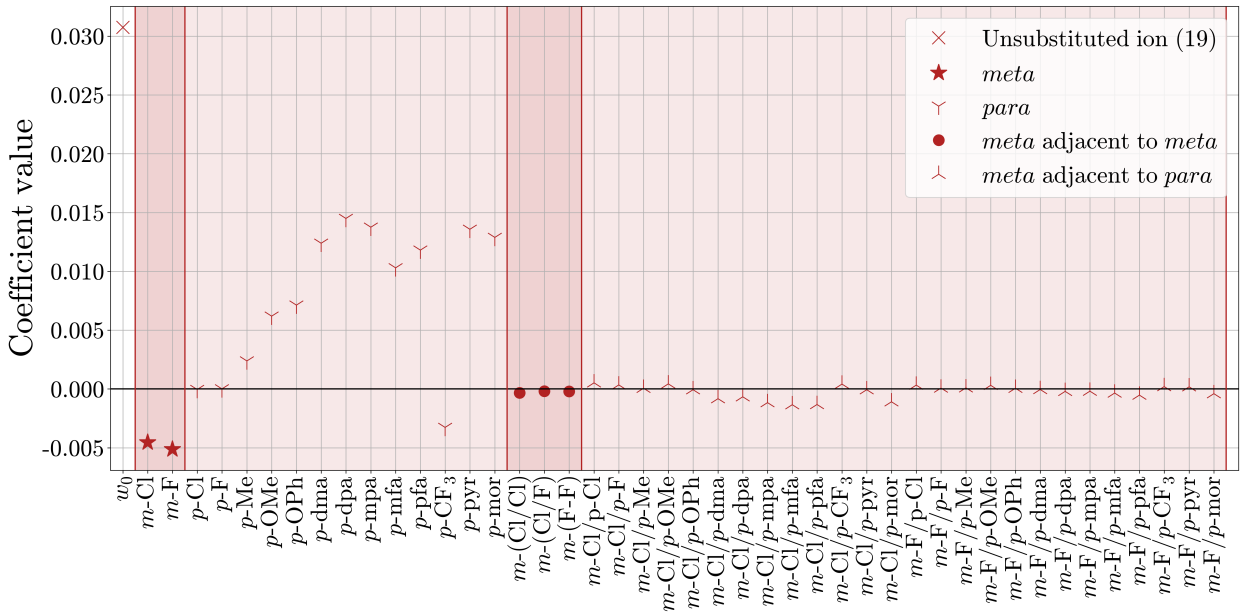

Figure 12: Model coefficients are shown for the regression of $\varepsilon_{\mathrm{LUMO}}$ on $\mathrm{C_{FG}}$. The different reddish sections refer to the categories from Fig. 4. The coefficient names are based on the notation in Mayr's database.[21,22] The following abbreviations are used: *p*-dma, 4-(dimethylamino)phenyl; *p*-dpa, 4-(diphenylamino)phenyl; *p*-mpa, 4-(methylphenylamino)phenyl; *p*-mfa, 4-(methyl(trifluoroethyl)amino)phenyl; *p*-pfa, 4-(phenyl(trifluoroethyl)amino)phenyl.

In Figs. 12 and 13, the regression coefficients are shown for the MLR models linking $\mathrm{C_{FG}}$ and $F_{\mathrm{2B}}^{\mathrm{split}}$ with $\hat{\varepsilon}_{\mathrm{LUMO}}$, respectively. In both cases, the intercept, w_0, is an approximation to $\varepsilon_{\mathrm{LUMO}}$ of the unsubstituted benzhydrylium ion **19** for which all other coefficients are zero.

$\mathbf{C_{FG}}$ (Fig. 12). The different substituents at the *meta* and *para* positions of the benzhydrylium ion have the ability to push/pull electron density in/out of the aromatic rings.

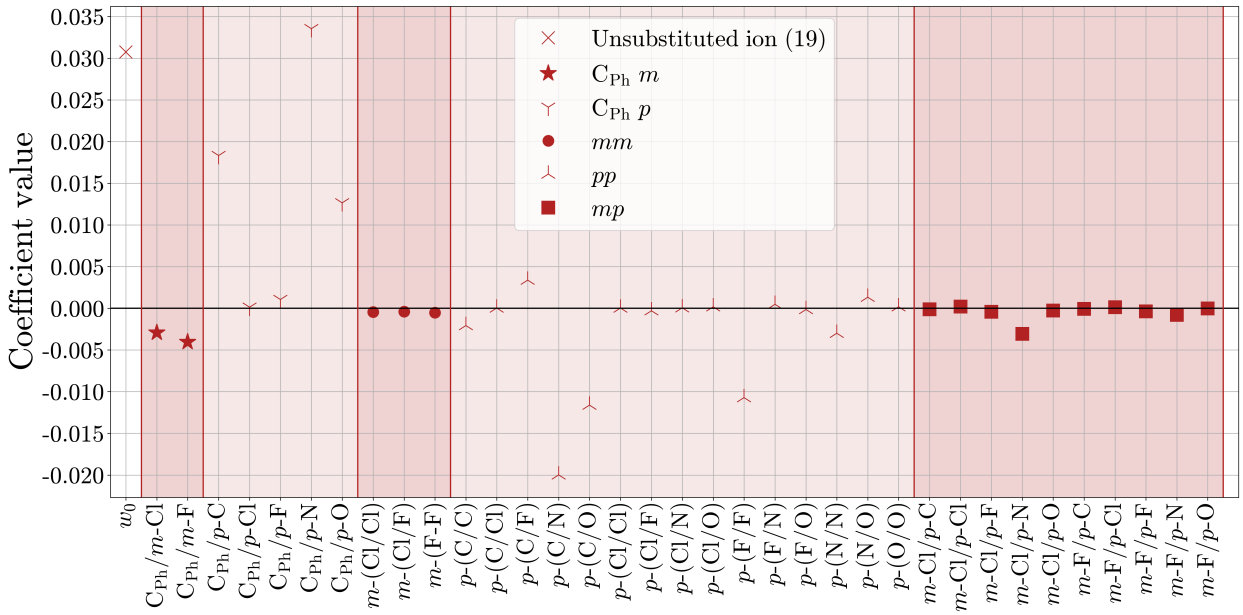

Figure 13: Model coefficients are shown for the regression of $\varepsilon_{\mathrm{LUMO}}$ on $F_{\mathrm{2B}}^{\mathrm{split}}$. The different reddish sections refer to the categories from Fig. 6.

Negative regression coefficients correspond to electron-withdrawing groups, reducing the electron density at the carbenium ion and therefore the $\varepsilon_{\mathrm{LUMO}}$, as expected. This results in a larger E parameter (see column r in Table 3). For the underlying data set, large negative coefficients are primarily found for both *meta* substituents, –F and –Cl, and in *para* position for $-CF_3$. The opposite effect is found for positive regression coefficients. They correspond to electron-donating groups increasing the electron density at the carbenium ion, resulting in higher $\varepsilon_{\mathrm{LUMO}}$ values and smaller E parameters. Especially the electron-rich nitrogen- and oxygen-bonded substituents at the *para* positions substantially decrease the E parameter. Compared to the "*meta*" and "*para*" blocks of the C_{FG} descriptor, the coefficients of the "*meta* adjacent to *meta*" and "*meta* adjacent to *para*" blocks are close to zero. They are hence of minor importance for the prediction and interpretation of benzhydrylium reactivity.

$\boldsymbol{F}_{\mathbf{2B}}^{\mathbf{split}}$ (Fig. 13). Contrary to the descriptor composition shown in Fig. 6, some descriptor dimensions were deleted after performing a sensitivity analysis (see SI Section "Sensitivity analysis of the different interaction groups of the $F_{\mathrm{2B}}^{\mathrm{split}}$ descriptor."). All coefficient blocks including the carbenium ion (C^+) were deleted due to strong correlation with those con-

tainting the carbon atoms of the phenyl rings (C_{Ph}). The decision whether to delete the C^+ or C_{Ph} coefficient blocks is arbitrary since both possibilities lead to the same result. Additionally, the coefficient block C_{Ph}/C_{Ph} was deleted, as it is identical for all molecules. The first two coefficient blocks (C_{Ph}/m and C_{Ph}/p) describe the direct interactions of the substituents with the aromatic rings. The closer a substituent's atom (element X) is to the phenyl rings, the greater the effect of the X/C_{Ph} interaction on ε_{LUMO}. Hence, the element that is directly bonded to the phenyl ring is expected to predominantly alter ε_{LUMO}, which is consistent with chemical intuition in many cases. The first coefficient block (C_{Ph}/m) shows the same trend as observed for C_{FG}, with the same explanation. In the second coefficient block (C_{Ph}/p), the interactions to *p*–Cl and *p*–N can be well interpreted since both atoms appear only in one certain position: directly bonded to the aromatic rings. For instance, the strong electron-pushing character of the nitrogen-bonded substituents is reflected by a large positive coefficient value, resulting in a high value of ε_{LUMO} and a low value of E. The C_{Ph}/p interactions to carbon, fluorine, and oxygen, on the other hand, are composed of several possible positions in the molecule. Nevertheless, the general trend in the oxygen interactions can be explained: Oxygen atoms are present in three functional groups (*p*-OMe, *p*-OPh, *p*-mor), all of which are electron donating groups resulting in higher ε_{LUMO} values. In the p/p and m/p coefficient blocks, many different effects overlap, which does not allow for a straightforward interpretation. However, individual insight is still possible. For instance, the negative coefficient of the *p*-(F/F) interaction can be observed only in the electron withdrawing groups containing fluorine (–F, $–CF_3$), because other fluorine atoms (in *p*-mfa and *p*-pfa) are far away and the influence is approximately zero. Regarding the *p*-(N/N) interaction, the possible distance between two nitrogen atoms is always large in this data set, resulting in a descriptor value close to zero. To compare the different regression coefficients, they have to be standardized. However, this may result in artifacts for descriptor dimensions that are close to zero for the entire data set, such as *p*-(N/N).

In summary, the interpretation of C_{FG} is straightforward for each substituent. F_{2B}^{split}, on

the other hand, can reveal details beyond substituent identities. The more universal F_{2B}^{1} descriptor is not nearly as simple to interpret. For instance, no distinction between different fluorine atoms is possible. In general, the more complex the descriptor structure is, i.e., the more different effects overlap in one descriptor dimension, the more difficult the chemical interpretation becomes. This is especially true for the "*meta* adjacent to *para*" block of C_{FG} and the p/p and m/p blocks of F_{2B}^{split}.

The limits of chemical intuition

Finally, we would like to highlight one of the practical benefits of our approach. Fig. 14 shows a fully substituted benzhydrylium ion. It comprises four electron-withdrawing groups (m-Cl) and two electron-donating groups (p-OMe). Does the electrophilicity increase or decrease with respect to the unsubstituted ion **19**? For each individual substituent, the qualitative effect on E can be estimated using chemical intuition, whereas quantitative predictions are already difficult to make at this level. As soon as the effects of several substituents on E overlap, even qualitative estimates are no longer possible. This challenge can be met with the help of a quantitative approach as presented in this study: Without the electron-

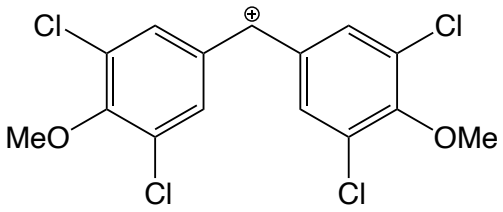

Figure 14: Example of a fully substituted benzhydrylium ion. Does the electrophilicity increase or decrease with respect to the unsubstituted ion **19**?

donating *para* groups, the benzhydrylium ion is more electrophilic ($E = 8.31$) compared to the unsubstituted prototype **19** ($E = 5.32$). Adding p-OMe, its electrophilicity decreases by -4.18 units ($E = 4.13$) and hence even below the prototype's value.

Conclusions and outlook

We have explored the feasibility of real-time data-driven reactivity prediction for routine synthesis planning. In previous data-driven reactivity studies, quantum molecular properties (QMPs) were used to learn quantitative relationships between these properties and Mayr-type reactivity parameters (E, N, s_N). While QMPs are informative quantities, their calculation is computationally intensive, preventing the possibility of a real-time approach.

As an alternative, we have considered here structural descriptors that can be generated in real-time. A combinatorial data set of $M = 3570$ benzhydrylium ions served as domain of application. For only $K = 27$ of these systems, electrophilicity parameters E are available. For each system, three structural descriptors were generated, ranging from application-specific but interpretable ($\mathrm{C_{FG}}$, $F_\mathrm{2B}^\mathrm{split}$) to application-agnostic but less interpretable (F_2B^1). However, a direct mapping of the structural descriptors to E via multivariate linear regression (MLR) is not possible due to lack of data (cf. K). We expect more sophisticated machine-learning models, which are even more data-hungry than linear models, to fail as well.

Instead, we have developed a two-step workflow based on the MLR technique. Step 2 of the workflow resembles previous approaches: a quantitative QMP–E relationship is learned based on K training data points. The QMPs considered here are functions of frontier molecular orbital energies. In step 1 of the workflow, quantitative descriptor–QMP relationships are learned to replace the expensive QMP generation by efficient real-time predictions. We identified $\mathrm{C_{FG}}$ to be the descriptor of choice with respect to the rate of learning. Our analysis suggests that $L \approx 150$ training data points (i.e., quantum chemical calculations) are necessary to make robust and accurate E predictions with a test set R^2-value of approximately 0.98. Hence, we can replace quantum chemical calculations with real-time predictions for almost 96 % of all structures. In summary, the two-step workflow is an effective approach if $K \ll L \ll M$.

A positive side effect of the MLR approach and the use of application-specific descriptors is that they yield interpretable models. Not only could we confirm chemically intuitive trends,

we could also identify quantitative effects of individual substituents and even individual elements on E. Due to the additive nature of the MLR framework, individual functional groups can be simply "clicked" together to predict E parameters for highly substituted benzhydrylium ions, an example where even qualitative estimates based on human expertise are error-prone.

The next challenge on the way to real-time reactivity prediction for arbitrary molecules is to extend our approach to a broader range of structural classes. However, even within the benzhydrylium space, many more substituents are to be explored. We assume that the second step of our approach is — without any further modification — applicable to other functional groups and the resulting substitution patterns. At the same time, we need to overcome the problem of data shortage. With the information provided by data-driven reactivity studies, synthetic chemists can more systematically plan new experiments that, in turn, feed future data-driven campaigns. We invite laboratories around the globe to help us build such experimental–computational feedback loops to accelerate advances in organic synthesis.

Acknowledgement

MV and JP acknowledge funding by Germany's joint federal and state program supporting early-career researchers (WISNA) established by the Federal Ministry of Education and Research (BMBF). The authors thank the research group of Prof. Christoph R. Jacob (TU Braunschweig), especially Dr. Mario Wolter, for computational support and resources. JP thanks Dr. Verena Kraehmer for bringing reactivity scales to his attention.

Supporting Information Available

References

(1) Mayr, H. Reactivity Scales for Quantifying Polar Organic Reactivity: The Benzhydrylium Methodology. *Tetrahedron* **2015**, *71*, 5095–5111.

(2) Mayr, H. Physical Organic Chemistry—Development and Perspectives. *Isr. J. Chem.* **2016**, *56*, 30–37.

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

(4) Struble, T. J. et al. Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis. *J. Med. Chem.* **2020**, *63*, 8667–8682.

(5) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem. Int. Ed.* **2016**, *55*, 5904–5937.

(6) Klucznik, T. et al. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **2018**, *4*, 522–532.

(7) Warr, W. A. A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol. Inf.* **2014**, *33*, 469–476.

(8) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281–1289.

(9) Shen, Y.; Borowski, J. E.; Hardy, M. A.; Sarpong, R.; Doyle, A. G.; Cernak, T. Automation and computer-assisted planning for chemical synthesis. *Nat Rev Methods Primers* **2021**, *1*, 23.

(10) Molga, K.; Szymkuć, S.; Grzybowski, B. A. Chemist Ex Machina: Advanced Synthesis Planning by Computers. *Acc. Chem. Res.* **2021**, *54*, 1094–1106.

(11) Tripp, A.; Maziarz, K.; Lewis, S.; Liu, G.; Segler, M. Re-Evaluating Chemical Synthesis Planning Algorithms. NeurIPS 2022 AI for Science: Progress and Promises. 2022.

(12) Muratov, E. N. et al. QSAR without Borders. *Chem. Soc. Rev.* **2020**, *49*, 3525–3564.

(13) 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.

(14) Vahl, M.; Proppe, J. The Computational Road to Reactivity Scales. *Phys. Chem. Chem. Phys.* **2023**, *25*, 2717–2728.

(15) Proppe, J.; Kircher, J. Uncertainty Quantification of Reactivity Scales. *ChemPhysChem* **2022**, *23*, e202200061.

(16) Schneider, R.; Grabis, U.; Mayr, H. Direct Determination of Rate Constants for the Addition of Carbenium Ions to Alkenes. *Angew. Chem. Int. Ed.* **1986**, *25*, 89–90.

(17) Mayr, H.; Schneider, R.; Schade, C.; Bartl, J.; Bederke, R. Addition Reactions of Diarylcarbenium Ions to 2-Methyl-1-Pentene: Kinetic Method and Reaction Mechanism. *J. Am. Chem. Soc.* **1990**, *112*, 4446–4454.

(18) Mayr, H.; Schneider, R.; Grabis, U. Linear Reactivity-Selectivity Correlations in Additions of Diarylcarbenium Ions to Alkenes; a Rebuttal of the Reactivity-Selectivity Principle. *Angew. Chem. Int. Ed.* **1986**, *25*, 1017–1018.

(19) Mayr, H.; Schneider, R.; Grabis, U. Linear Free Energy and Reactivity-Selectivity Relationships in Reactions of Diarylcarbenium Ions with π-Nucleophiles. *J. Am. Chem. Soc.* **1990**, *112*, 4460–4467.

(20) Mayr, H.; Patz, M. Scales of Nucleophilicity and Electrophilicity: A System for Ordering Polar Organic and Organometallic Reactions. *Angew. Chem. Int. Ed.* **1994**, *33*, 938–957.

(21) Mayr, H.; Ofial, A. R. Mayr's Database of Reactivity Parameters. `https://www.cup.lmu.de/oc/mayr/reaktionsdatenbank2/`, last accessed on 26 June 2023.

(22) Mayr, H.; Ofial, A. R. A Quantitative Approach to Polar Organic Reactivity. *SAR QSAR Environ. Res.* **2015**, *26*, 619–646.

(23) Mayr, H.; Bug, T.; Gotta, M. F.; Hering, N.; Irrgang, B.; Janker, B.; Kempf, B.; Loos, R.; Ofial, A. R.; Remennikov, G.; Schimmel, H. Reference Scales for the Characterization of Cationic Electrophiles and Neutral Nucleophiles,. *J. Am. Chem. Soc.* **2001**, *123*, 9500–9512.

(24) Ammer, J.; Nolte, C.; Mayr, H. Free Energy Relationships for Reactions of Substituted Benzhydrylium Ions: From Enthalpy over Entropy to Diffusion Control. *J. Am. Chem. Soc.* **2012**, *134*, 13902–13911.

(25) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.

(26) Eyring, H. The Activated Complex in Chemical Reactions. *J. Chem. Phys.* **1935**, *3*, 107–115.

(27) Pérez, P.; Toro-Labbé, A.; Aizman, A.; Contreras, R. Comparison between Experimental and Theoretical Scales of Electrophilicity in Benzhydryl Cations. *J. Org. Chem.* **2002**, *67*, 4747–4752.

(28) Wang, C.; Fu, Y.; Guo, Q.-X.; Liu, L. First-Principles Prediction of Nucleophilicity

Parameters for π Nucleophiles: Implications for Mechanistic Origin of Mayr's Equation. *Chem. Eur. J.* **2010**, *16*, 2586–2598.

(29) Pereira, F.; Latino, D. A. R. S.; Aires-de-Sousa, J. Estimation of Mayr Electrophilicity with a Quantitative Structure–Property Relationship Approach Using Empirical and DFT Descriptors. *J. Org. Chem.* **2011**, *76*, 9312–9319.

(30) Zhuo, L.-G.; Liao, W.; Yu, Z.-X. A Frontier Molecular Orbital Theory Approach to Understanding the Mayr Equation and to Quantifying Nucleophilicity and Electrophilicity by Using HOMO and LUMO Energies. *Asian J. Org. Chem.* **2012**, *1*, 336–345.

(31) Kiyooka, S.-i.; Kaneno, D.; Fujiyama, R. Intrinsic Reactivity Index as a Single Scale Directed toward Both Electrophilicity and Nucleophilicity Using Frontier Molecular Orbitals. *Tetrahedron* **2013**, *69*, 4247–4258.

(32) Allgäuer, D. S.; Jangra, H.; Asahara, H.; Li, Z.; Chen, Q.; Zipse, H.; Ofial, A. R.; Mayr, H. Quantification and Theoretical Analysis of the Electrophilicities of Michael Acceptors. *J. Am. Chem. Soc.* **2017**, *139*, 13318–13329.

(33) Hoffmann, G.; Balcilar, M.; Tognetti, V.; Héroux, P.; Gaüzère, B.; Adam, S.; Joubert, L. Predicting Experimental Electrophilicities from Quantum and Topological Descriptors: A Machine Learning Approach. *J. Comput. Chem.* **2020**, *41*, 2124–2136.

(34) Lee, B.; Yoo, J.; Kang, K. Predicting the Chemical Reactivity of Organic Materials Using a Machine-Learning Approach. *Chem. Sci.* **2020**, *11*, 7813–7822.

(35) Orlandi, M.; Escudero-Casao, M.; Licini, G. Nucleophilicity Prediction via Multivariate Linear Regression Analysis. *J. Org. Chem.* **2021**, *86*, 3555–3564.

(36) Boobier, S.; Liu, Y.; Sharma, K.; Hose, D. R. J.; Blacker, A. J.; Kapur, N.; Nguyen, B. N. Predicting Solvent-Dependent Nucleophilicity Parameter with a Causal Structure Property Relationship. *J. Chem. Inf. Model.* **2021**, *61*, 4890–4899.

(37) Haas, B. C.; Goetz, A. E.; Bahamonde, A.; McWilliams, J. C.; Sigman, M. S. Predicting Relative Efficiency of Amide Bond Formation Using Multivariate Linear Regression. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119*, e2118451119.

(38) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Physics-Inspired Structural Representations for Molecules and Materials. *Chem. Rev.* **2021**, *121*, 9759–9815.

(39) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and Mechanistic Multivariate Linear Regression Models for Reaction Development. *Chem. Sci.* **2018**, *9*, 2398–2412.

(40) Crawford, J. M.; Kingston, C.; Toste, F. D.; Sigman, M. S. Data Science Meets Physical Organic Chemistry. *Acc. Chem. Res.* **2021**, *54*, 3136–3148.

(41) Pronobis, W.; Tkatchenko, A.; Müller, K.-R. Many-Body Descriptors for Predicting Molecular Properties with Machine Learning: Analysis of Pairwise and Three-Body Interactions in Molecules. *J. Chem. Theory Comput.* **2018**, *14*, 2991–3003.

(42) Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual Density Functional Theory. *Chem. Rev.* **2003**, *103*, 1793–1874.

(43) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L. Density-Functional Theory for Fractional Particle Number: Derivative Discontinuities of the Energy. *Phys. Rev. Lett.* **1982**, *49*, 1691–1694.

(44) Parr, R. G.; Pearson, R. G. Absolute Hardness: Companion Parameter to Absolute Electronegativity. *J. Am. Chem. Soc.* **1983**, *105*, 7512–7516.

(45) Mulliken, R. S. A New Electroaffinity Scale; Together with Data on Valence States and on Valence Ionization Potentials and Electron Affinities. *J. Chem. Phys.* **1934**, *2*, 782–793.

(46) Parr, R. G.; Donnelly, R. A.; Levy, M.; Palke, W. E. Electronegativity: The Density Functional Viewpoint. *J. Chem. Phys.* **1978**, *68*, 3801–3807.

(47) Parr, R. G.; v. Szentpály, L.; Liu, S. Electrophilicity Index. *J. Am. Chem. Soc.* **1999**, *121*, 1922–1924.

(48) Gázquez, J. L.; Cedillo, A.; Vela, A. Electrodonating and Electroaccepting Powers. *J. Phys. Chem. A* **2007**, *111*, 1966–1970.

(49) Chattaraj, P. K.; Chakraborty, A.; Giri, S. Net Electrophilicity. *J. Phys. Chem. A* **2009**, *113*, 10068–10074.

(50) Barrett, G. B. The Coefficient of Determination: Understanding r2 and R2. *Math. Teach.* **2000**, *93*, 230–234.

(51) Rodgers, J. L.; Nicewander, W. A. Thirteen Ways to Look at the Correlation Coefficient. *Am. Stat.* **1988**, *42*, 59–66.

(52) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(53) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2019**, *15*, 2847–2862.

(54) Pracht, P.; Bohle, F.; Grimme, S. Automated Exploration of the Low-Energy Chemical Space with Fast Quantum Chemical Methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(55) Neese, F. The ORCA Program System. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73–78.

(56) Neese, F. Software Update: The ORCA Program System—Version 5.0. *WIREs Comput Mol Sci* **2022**, *12*.

(57) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta–Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(58) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. Comparative Assessment of a New Nonempirical Density Functional: Molecules and Hydrogen-Bonded Complexes. *J. Chem. Phys.* **2003**, *119*, 12129–12137.

(59) 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*, 154104.

(60) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(61) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(62) Weigend, F. Accurate Coulomb-fitting Basis Sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057.

(63) Proppe, J. Quantitative structure–reactivity relationships for synthesis planning: The benzhydrylium case. `https://git.rz.tu-bs.de/proppe-group/qsrr-benzhydrylium`, last accessed on 26 June 2023.

(64) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(65) Aggarwal, C. C. *Neural Networks and Deep Learning: A Textbook*; Springer International Publishing AG: Cham, Switzerland, 2018.

(66) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; The MIT Press: Cambridge (MA), United States, 2006.

(67) Hastie, T.; Tibshirani, R. J.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York (NY), United States, 2009.

TOC Graphic