TRANSFORMING PREDICTIONS INTO TESTABLE HYPOTHESES: THE CASE OF POLAR ORGANIC REACTIVITY

Jonny Proppe*
Institute of Physical Chemistry
Georg-August University
37077 Göttingen, Germany
*jproppe@uni-goettingen.de

Johannes Kircher
Institute of Physical Chemistry
Georg-August University
37077 Göttingen, Germany

February 24, 2021

ABSTRACT

Herbert Mayr's research on reactivity scales tells a success story of how polar organic synthesis can be rationalized by a simple empirical relationship. In this work, we propose an extension to Mayr's reactivity approach that is rooted in uncertainty quantification (UQ). It transforms the *unique* values of reactivity parameters (s_N, N, E) into value *distributions*. Through uncertainty propagation, these distributions can be exploited to quantify the uncertainty of bimolecular rate constants. Our UQ-based extension serves three purposes. First, predictions of polar organic reactivity can be transformed into testable hypotheses, which increases the overall reliability of the method and guides the exploration of new research directions. Second, it is also possible to quantify the discriminability of two competing reactions, which is particularly important if subtle reactivity differences matter. Third, since rate constant uncertainty can also be quantified for reactions that have yet to be observed, new opportunities arise for benchmarking computational chemistry methods (benchmarking *under uncertainty*). We demonstrate the functionality and performance of the UQ-extended reactivity approach at the example of the 2001/12 reference data set released by Mayr and co-workers [*J. Am. Chem. Soc.* **2001**, *123*, 9500; *J. Am. Chem. Soc.* **2012**, *134*, 13902]. As a by-product of the new approach, we obtain revised reactivity parameters for the electrophiles and the nucleophiles of the reference set.

***K*eywords** Chemical kinetics · Electrophilicity · Nucleophilicity · Reactivity scales · Uncertainty quantification

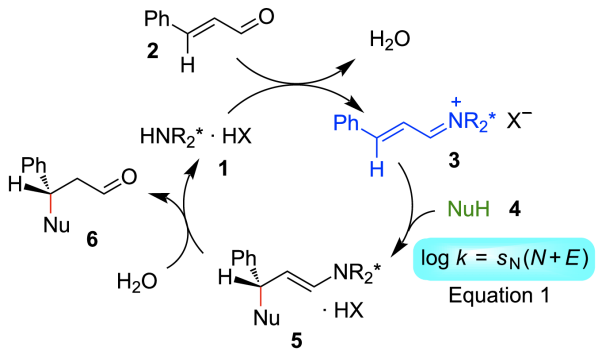
Introduction

Polar organic reactions are ubiquitous in Nature and the chemical industry. Synthesis planning involving reactions of this kind relies on two fundamental questions (among others): whether nucleophilic attack takes place on a relevant time scale, and whether this time scale interferes with that of another reaction in which either the same nucleophile or the same electrophile participates. The answers to both questions revolve around the quantification of reaction rates — absolute ones in the former case, relative ones in the latter case. For instance, in iminium-activated reactions (Scheme 1) [1], it is important that nucleophile **4** is strong enough (in absolute terms) to attack the intermediate iminium ion (**3**) but also weak enough (in relative terms) not to react with the precursor carbonyl compound (**2**).

Herbert Mayr and co-workers provided unambiguous evidence that a simple empirical relationship, known as the Mayr–Patz equation (MPE), addresses scenarios of this kind reliably [2],

$$\log k_\mathrm{exp} \approx \log k_\mathrm{MPE} = s_\mathrm{N}(N + E) \tag{1}$$

We define $\log k \equiv \log_{10} k_2(20\ ^\circ\mathrm{C})$ for the sake of brevity. Here, the decadic logarithm of the bimolecular rate constant measured at 20 °C ($\log k_\mathrm{exp}$) is approximated as the sum of two reactivity parameters (nucleophilicity N and electrophilicity E), multiplied by a nucleophile-specific sensitivity factor (s_N). The MPE allows for semi-quantitative predictions of bimolecular rate constants in a remarkable range of about $-5 < \log k < 8$. The philicities (N and E) of



Scheme 1: Generic mechanistic proposal for the catalytic cycle of an iminium-activated reaction [1]. This figure was published in Tetrahedron, 71, H. Mayr, Reactivity Scales for Quantifying Polar Organic Reactivity: The Benzhydrylium Methodology, 5095–5111, Copyright Elsevier (2015).

the species involved in reactions verifying this relationship cover a range of 30–40 orders of magnitude, which can be considered a unique achievement given that the accuracy of k_{MPE} is within a factor of 10 to 100. On the basis of these results, Mayr formulated an uncertainty principle of organic reactivity: the *accuracy* of k_{MPE} and chemical *diversity* cannot be maximized at the same time. Even though higher accuracy can be reached if one considers a narrower range of chemical species, the small errors in k_{MPE} appear impressive given the diversity of Mayr's reactivity database [3], which currently comprises reactivity parameters for 1227 nucleophiles and 330 electrophiles.

In this work, we introduce uncertainty quantification (UQ) into Mayr's reactivity approach. This combined approach, which we will make openly available [4], enables users to perform *virtual* measurements of $\log k$, which are reported as expectation $\pm$ deviation — just like *physical* measurements. Usually, virtual measurement uncertainty (or prediction uncertainty) is significantly larger than physical measurement uncertainty, which can be attributed to a more comprehensive list of uncertainty components including parameter uncertainty, model discrepancy, and numerical noise [5, 6, 7]. A key feature of our UQ approach is the transformation of *unique* values of reactivity parameters into value *distributions*, which can be translated — via uncertainty propagation — into value distributions of $\log k_{\mathrm{MPE}}$. We argue that quantitative knowledge of uncertainty in k_{MPE} enhances the already powerful reactivity approach by Mayr, for three reasons.

First, virtual measurements of $\log k$ (expectation $\pm$ deviation) represent testable statistical hypotheses. That is, one can quantify an $x\%$ confidence interval of $\log k_{\mathrm{MPE}}$ and count how often $\log k_{\mathrm{exp}}$ is located within that interval (ideally $x\%$). According to the *Guide to the Expression of Uncertainty in Measurement* [8], it is recommended to express uncertainty as a 95% confidence interval. This recommendation is supported by the community [9, 10, 11]. Such reporting standards help to identify shortcomings, thereby increasing the overall reliability of Mayr's reactivity approach and guiding the search for new research directions (e.g., proposing measurements of yet unobserved reactions).

Second, in synthesis planning, where subtle reactivity differences may matter (cf. discussion to Scheme 1), our UQ-based approach can support the decision-making process. The larger the overlap of two $\log k_{\mathrm{MPE}}$ distributions corresponding to competing reactions, the less certain one can discriminate between the two. The overlap itself (a value between zero and one) can be considered a quantitative measure of this kind of discriminability. The more the overlap tends toward zero (one), the more (less) certain it is that one can predict the relative species flux through competing channels.

Third, since rate constant uncertainty can also be quantified for reactions that have yet to be observed, new opportunities arise for benchmarking computational chemistry methods [12, 13]. Even if the experimental benchmark for a reaction of interest is not yet available — which, so far, severely constrained the domain of application for theoreticians — our UQ approach still enables benchmarking, but *under uncertainty*. This way, the diversity of benchmark sets can be increased remarkably, which we anticipate to accelerate method developments in theoretical and computational chemistry.

To explore the potential of UQ for chemical research, we build upon previous work by Proppe and Reiher [13], addressing Mössbauer spectroscopy [7, 14], dispersion corrections to density functional theory [15, 16], and reaction kinetics [17, 18]. This foundation will support our endeavor to pave the way for a novel approach to determining reactivity parameters with steadily increasing accuracy. For demonstration purposes, we selected more than 200 reactions of the two reference data sets published by Mayr and co-workers [19, 20], which cover a wide range of $\log k$ values (-3.6 to $+8.0$).

Methods

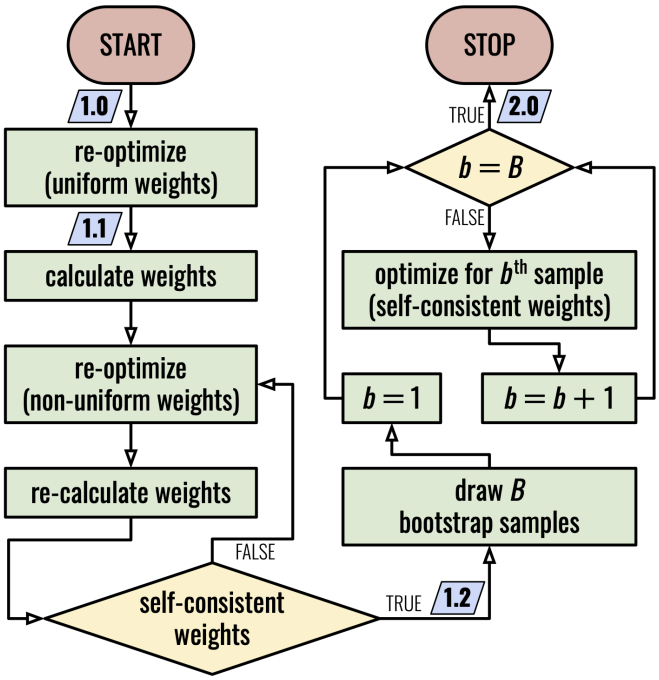
Optimization of Reactivity Parameters

We employed the following objective function for optimizing reactivity parameters,

$$\Delta^2 = \sum_{r=1}^{R} w_r \cdot [\delta_r(\log k)]^2 \tag{2}$$

$$\delta_r(\log k) = \log k_{\text{exp},r} - \log k_{\text{MPE},r} \tag{3}$$

Here, $\delta_r(\log k)$ and w_r are the *residual* and the *weight* of the rth reaction (R reactions in total), respectively. We employed the basin-hopping algorithm by Wales and Doyle [21] as implemented in `SciPy 1.5.0` [22] for minimizing the objective function. We used the default settings of the basin-hopping algorithm except for the argument `niter`, which we set from `niter=100` to `niter=1`. In the original optimization studies [19, 20], a special case of this objective function was employed, where all weights are uniformly distributed, i.e., $w_r = w_{r'}$ for all possible values of r and $r' \neq r$. In this study, non-uniform weights were determined on the basis of discrepancy weighting [7, 11, 23] (see Appendix A). Discrepancy weighting is an iterative procedure that refines weights until they no longer change. The resulting self-consistent weights were then used to determine parameter uncertainty on the basis of Bayesian bootstrapping [24]. This technique simulates drawing new samples from the underlying but unknown population by assuming that the data set at hand itself is the population. Consequently, only available reaction data is used to draw samples, each of which yields slightly different reactivity parameters (see Appendix B for more details on this technique). The full optimization workflow is summarized in Scheme 2.



Scheme 2: Flowchart illustrating our approach to optimizing reactivity parameters. Version labels (in blue rhomboid boxes) represent a hierarchy of distinct parametrizations.

Quantification of Uncertainty in log k_{MPE}

We define the *model error* as the root-mean-square error of the residuals,

$$\text{RMSE} \equiv \varepsilon = \sqrt{R^{-1} \sum_{r=1}^{R} [\delta_r(\log k)]^2} = \sqrt{\mu^2 + \sigma^2} \tag{4}$$

It is equivalent to the definition of $\log \sigma$ by Mayr and co-workers (see Footnote 58 [19]). In the case of uniform weights, $w_r = R^{-1}$ for all $r = 1, ..., R$, the squared model error, ε^2, equals Δ^2. The model error combines information on both the *model bias* (μ) and *model dispersion* (σ). The model bias or mean error (ME) represents the centroid of the residuals and is an estimate of the overall systematic error in $\log k_{\text{MPE}}$,

$$\text{ME} \equiv \mu = R^{-1} \sum_{r=1}^{R} \delta_r(\log k) \tag{5}$$

The model dispersion represents the scatter of the residuals and is reflected by the root-mean-square deviation,

$$\text{RMSD} \equiv \sigma = \sqrt{R^{-1} \sum_{r=1}^{R} \left[\delta_r(\log k) - \mu\right]^2} \tag{6}$$

Under the assumption of normally distributed residuals, model dispersion represents the model's contribution to prediction uncertainty, i.e., the uncertainty in $\log k_{\text{MPE}}$. The second contribution to prediction uncertainty is parameter uncertainty, which can be estimated from the ensemble of bootstrap samples generated in the course of our optimization workflow. Since each bootstrap sample (B in total) yields slightly different reactivity parameters, we obtain an empirical distribution for each parameter. Uncertainty propagation is straightforward. For a given reaction, each bootstrap sample yields a slightly different $\log k_{\text{MPE}}$ value, leading again to an empirical distribution. We define the parameter-related uncertainty in $\log k_{\text{MPE}}$ of the rth reaction as

$$\beta_r = \sqrt{B^{-1} \sum_{b=1}^{B} \left[\log k_{\text{MPE},r}^{(b)} - B^{-1} \sum_{b=1}^{B} \log k_{\text{MPE},r}^{(b)}\right]^2} \tag{7}$$

Assuming normally distributed variables and independence of the two uncertainty contributions [25], the prediction uncertainty (95% confidence) corresponding to the rth reaction can be estimated as

$$U_{.95,r} = 1.96 \cdot U_r = 1.96 \cdot \sqrt{\sigma^2 + \beta_r^2} \tag{8}$$

Data Selection

All 304 reactions (in dichloromethane) of the 2001 and 2012 studies were considered [19, 20]. This pool (Charts 1 and 2, Table 1) encompasses 33 benzhydrylium ions (electrophiles) and 45 π-nucleophiles, and covers a wide range of $\log k_{\text{exp}}$ values (-3.6 to $+9.2$). The two *anchor* species are **E15** ($E = 0.00$) and **N7** ($s_{\text{N}} = 1.00$) [20]; their parameters E and s_{N}, respectively, were kept fixed throughout. We excluded those reactions from the optimization procedure (30 in total) for which $\log k_{\text{exp}} > 8$ as the MPE (Eq. 1) loses its validity in that regime (diffusion limit). We also excluded those reactions from the optimization procedure (47 in total) that were not measured according to the standard protocol: measurement at 20 °C plus least-squares fit of absorbance data to a single exponential. While we do not doubt the quality of these 47 data points, we still neglect them in this study as we attempt to remove potential sources of bias to draw conclusions from our UQ analysis that are as unambiguous as possible.

Since we do not know the true values of the experimental rate constants, we rely on an overdetermined system (more equations than unknowns). Therefore, we introduced and applied the *2E3N rule*. First, every non-anchor electrophile (single free parameter, E) needs to participate in at least two observed reaction. Second, every non-anchor nucleophile (two free parameters, s_{N} and N) needs to participate in at least three observed reactions. Third, the two anchor species

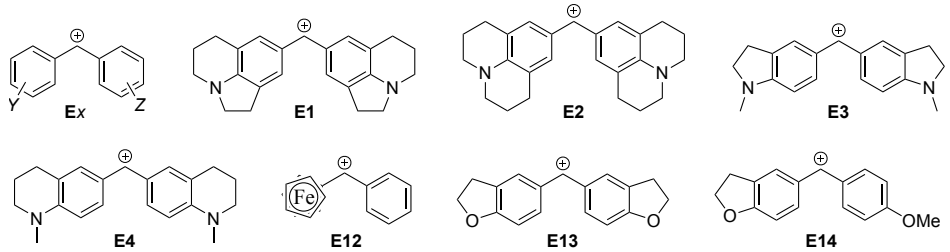


Chart 1: The 2001/12 reference set of electrophiles (benzhydrylium ions). Substituents Y and Z for all electrophiles collectively addressed as **E**x ($x =$ **5–11**, **15–33**) are specified in Table 1. The 2021 reference set comprises the same systems except for species **E33**.

Table 1: Specification of substituents for benzhydrylium ions of the reference set (Chart 1).

	Y	Z		Y	Z
E5	4-(*N*-pyrrolidino)	Y	**E21**	4-Me	H
E6	4-$N(Me)_2$	Y	**E22**	4-F	Y
E7	4-N(Me)(Ph)	Y	**E23**	4-F	H
E8	4-(*N*-morpholino)	Y	**E24**	3-F, 4-Me	Y
E9	4-$N(Ph)_2$	Y	**E25**	H	Y
E10	4-N(Me)(CH_2CF_3)	Y	**E26**	4-Cl	Y
E11	4-N(Ph)(CH_2CF_3)	Y	**E27**	3-F	H
E15	4-MeO	Y	**E28**	4-(CF_3)	H
E16	4-MeO	4-PhO	**E29**	3,5-F_2	H
E17	4-MeO	4-Me	**E30**	3-F	Y
E18	4-MeO	H	**E31**	3,5-F_2	3-F
E19	4-PhO	H	**E32**	4-(CF_3)	Y
E20	4-Me	Y	**E33**	3,5-F_2	Y

(**E15**: no free parameters; **N7**: single free parameter, N) need to participate in at least one (**E15**) or two (**N7**) observed reactions. Additionally, we required a fully connected network of reactions such that each reactivity parameter is a function of the full set of observed reactions. We also relaxed all fixed reactivity parameters of non-anchor species (**N1**–**N3**, **E1**–**E13**, **E16**–**E20** [20]). While relaxation increases the number of species violating the 2E3N rule, reliable UQ requires the elimination of all recognizable sources of systematic errors [8]. As each reactivity parameter is biased to an unknown degree (e.g., due to the finite size of the reference set), that bias would propagate through the network of reactions in the case of parameter fixation. This argument does not apply to the fixed parameters of the anchor species as their values merely scale/shift all other reactivity parameters by a constant value.

Applying parameter relaxation and the exclusion criteria mentioned above ($\log k_{\text{exp}} > 8$, non-standard protocol, violation of 2E3N rule, isolated subnetworks) left us with 212 valid reactions shared among 32 electrophiles and 36 nucleophiles (Fig. 1). This set of reactions represents 102 free reactivity parameters, which were optimized as per Scheme 2. For 30 of the 212 valid reactions, we extracted detailed experimental data from the supplementary material of the 2001/12 studies to quantify measurement uncertainty. For each reaction, there exists a series of observed rate constants, k_{obs} (ordinate), measured with respect to different excess nucleophile concentrations, [**N**] (abscissa). The slope of a linear regression model, $f_k([\mathbf{N}])$, represents the bimolecular rate constant k_2,

$$k_{\text{obs}} \approx f_k([\mathbf{N}]) = k_2[\mathbf{N}] + \text{constant} \tag{9}$$

Here, we applied Bayesian linear regression [26] as implemented in `Scikit-learn 0.23.1` [27] as it additionally yields uncertainty estimates of the regression coefficients. The uncertainty associated with the slope (k_2) represents the experimental standard deviation of the mean [8], which is the accepted definition of measurement uncertainty.

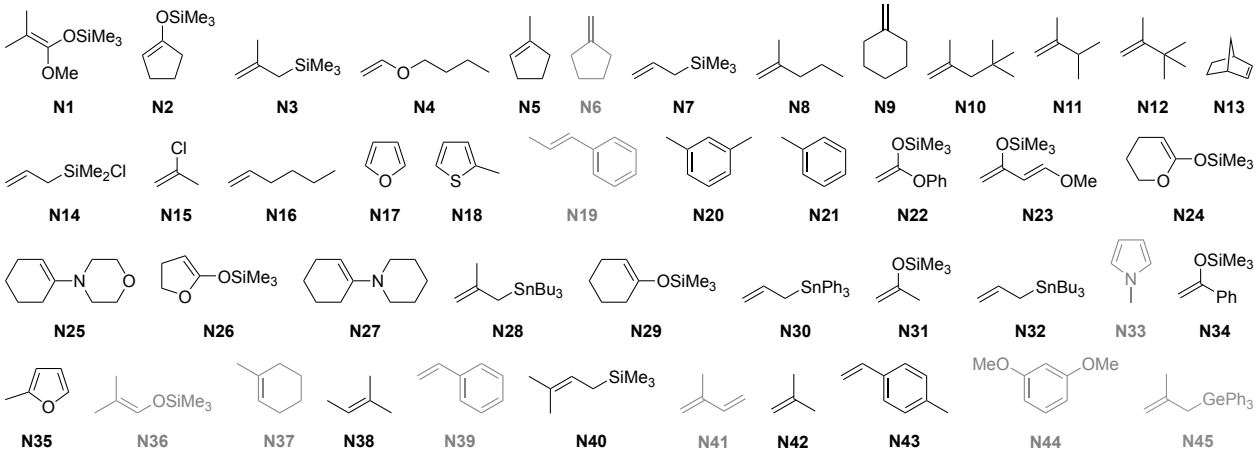


Chart 2: The 2001/12 reference set of nucleophiles (π-systems). The 2021 reference set comprises the same systems except for species **N6**, **N19**, **N33**, **N36**, **N37**, **N39**, **N41**, **N44**, and **N45**.

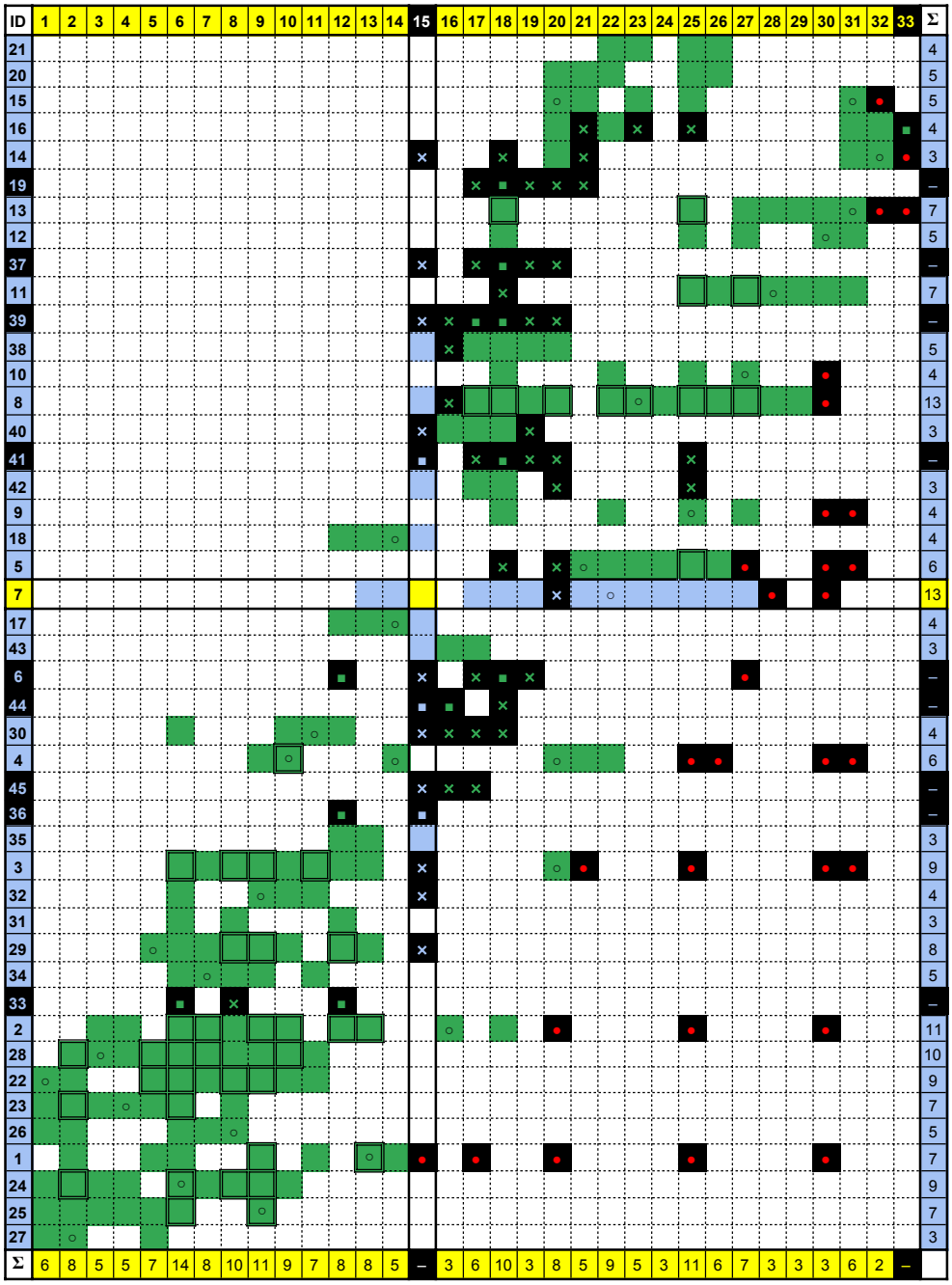


Figure 1: Reaction matrix of the reference set. Identifiers for electrophiles (top row, sorted by E in ascending order from left to right) and nucleophiles (left-most column, sorted by $s_{\mathrm{N}}N$ in ascending order from top to bottom) refer to the nomenclature defined in Charts 1 and 2 as well as Table 1. The bottom row (electrophiles) and the right-most column (nucleophiles) specify how often a species participated in a *valid* reference set reaction. Valid reactions are indicated in the main field (second to second-last row/column) by the background colors yellow (one free parameter), blue (two free parameters), and green (three free parameters). The background color of cells referring to invalid reactions is black. The same color code applies to the top/bottom rows and left-most/right-most columns. White cells of the main field represent unobserved reactions. Filled circles, squares, and crosses represent reactions for which $\log k_{\mathrm{exp}} > 8$, that do not follow the standard protocol, and that correspond to a species violating the 2E3N rule. Unfilled circles represent reactions for which we quantified measurement uncertainty. Cells that exhibit a double-line border were excluded in another series of optimizations for the purpose of hypothesis testing.

Results and Discussion

The structure of this section is reflected by the following roadmap:

I. Reproduction of the 2012 results [20] and quantification of numerical noise.

II. Application of our data selection criteria and re-optimization (uniform weighting), yielding a new set of reactivity parameters referred to as version 1.1.

III. Quantification of measurement uncertainty.

IV. Re-optimization of reactivity parameters (version 1.2) based on non-uniform weights determined through discrepancy weighting.

V. Estimation of empirical parameter distributions via discrepancy-weighted bootstrapping, building the newest set of reactivity parameters (version 2.0).

VI. Quantification and assessment (hypothesis testing) of uncertainty in $\log k_{\mathrm{MPE}}$.

Reproduction of the 2012 Parametrization

I. To validate our optimization procedure, we attempted to reproduce the results of the 2012 parametrization study [20]. The model error ε (Eq. 4) equals 0.13 and is 0.8% smaller than the model error determined in 2012. We find that the absolute difference of 0.17 in the nucleophilicity parameter (N) for **N5** constitutes, by far, the largest deviation. When excluding this nucleophile from the optimization procedure, we still obtain a model error of 0.13, but a decreased model error difference of 0.3% (with the 2021 error being smaller). The largest absolute difference in reactivity parameters that remains equals 0.02. This difference cannot be explained by the truncation of reactivity parameter values (after the second decimal) reported in the original article [20], which we used for this reproduction test. It is possible that the removal of **N5** causes this remaining difference since all reactivity parameters are coupled to each other through the objective function (Eq. 2).

The remaining deviation or a fraction thereof could possibly also be traced back to differences in the optimization algorithms. Mayr and co-workers used proprietary software and, hence, no detailed algorithmic information on the nonlinear optimizer is available. We can, however, estimate the magnitude of numerical noise that emerges from the customized settings of the basin-hopping optimizer. Numerical noise is caused by convergence thresholds, machine precision, etc., and contributes to the model error. For instance, the basin-hopping algorithm requires the specification of a number of iterations (`niter`). Preliminary tests suggested that a single iteration, which we set as default, is sufficient to obtain converged reactivity parameters. To underpin this finding, we re-ran the basin-hopping algorithm with 100 iterations and found a maximum deviation in the reactivity parameters of 5.51×10^{-4}, which is clearly smaller than the remaining maximum deviation of 0.02 with respect to the 2012 parametrization. For three additional arguments of the basin-hopping algorithm (`T`, `stepsize`, and `interval`) we decreased and increased the default values by a factor of 10, respectively. We stopped all runs after a single iteration. None of these runs led to a maximum deviation that was larger than 5.51×10^{-4}. On the basis of these result, we assume that we can safely exclude numerical noise to be the origin of the remaining difference between the 2012 and 2021 parametrizations.

We conclude that we can approximately, but not exactly, reproduce the 2012 results, which we cannot fully resolve. In particular, the disagreement caused by **N5** requires further investigation. Currently, we have no other explanation than a technical problem related to the optimizer employed in the 2012 study, or a typo that was either reported in the 2012 paper or applied in the 2012 optimization procedure.

Revised Reactivity Parameters 1.1: The Effect of Data Selection Criteria

II. We defined several data selection criteria, which led to a decrease of the number of reference electrophiles and reference nucleophiles. To understand the effect of each criterion on the optimization outcome, we applied them sequentially. We did not neglect the 2E3N rule, however, as it was applied implicitly in the 2001/12 studies. Furthermore, we always excluded reactions for which $\log k_{\mathrm{exp}} > 8$ as this constraint was applied consistently in the 2001/12 studies. Full connectivity of the network of reactions was found in all cases.

Without applying any criterion (case 1), we find a deviation (root-mean-square error) between the 2001/12 and 2021 parametrizations of 0.05. This deviation is not zero since the 2001 and 2012 optimizations were conducted successively, i.e., some of the 2001 reactivity parameters were kept fixed during the 2012 optimization. Here, we used the same fixed parameters, but optimized all reactivity parameters (2001 *and* 2012) at the same time. When excluding reactions that were not measured according to the standard protocol (case 2), the root-mean-square error increases to 0.25. We find a similar deviation (0.22) when re-including those reactions, but relaxing all fixed reactivity parameters except for

Table 2: Updated reactivity parameters (2.0) for reference nucleophiles and reference electrophiles. Each value represents the first moment (mean) of the associated empirical parameter distribution obtained through discrepancy-weighted bootstrapping. We also report the original values [19, 20] (1.0), those obtained by relaxing all fixed parameters corresponding to non-anchor species (1.1), and those obtained by relaxation plus discrepancy weighting (1.2). The s_N value (all versions) of the anchor nucleophile **N7** is printed in italics as it was kept fixed during optimization. The anchor electrophile **E15** is not shown as its electrophilicity parameter $E = 0.00$ was kept fixed during optimization. Nucleophiles and electrophiles that have been sorted out according to the criteria outlined in the Methods section (i.e., **N6**, **N19**, **N33**, **N36**, **N37**, **N39**, **N41**, **N44**, **N45**, **E33**) are also not shown. $\mathrm{RMSE}^{(1.0)}$ and $\mathrm{RMSE}^{(2.0)}$ refer to the root-mean-square error with respect to versions 1.0 and 2.0, respectively. The corresponding model errors (Eq. 4) amount to $\varepsilon^{(1.0)} = 0.11$, $\varepsilon^{(1.1)} = 0.09$, $\varepsilon^{(1.2)} = 0.10$, and $\varepsilon^{(2.0)} = 0.10$.

	$s_\mathrm{N}^{(1.0)}$	$s_\mathrm{N}^{(1.1)}$	$s_\mathrm{N}^{(1.2)}$	$\boldsymbol{s_\mathrm{N}^{(2.0)}}$	$N^{(1.0)}$	$N^{(1.1)}$	$N^{(1.2)}$	$\boldsymbol{N^{(2.0)}}$		$E^{(1.0)}$	$E^{(1.1)}$	$E^{(1.2)}$	$\boldsymbol{E^{(2.0)}}$
N1	0.98	0.84	0.87	0.87	9.00	10.13	9.84	9.84	**E1**	−10.04	−11.23	−10.88	−10.87
N2	0.93	0.84	0.86	0.86	6.57	7.33	7.13	7.12	**E2**	−9.45	−10.59	−10.26	−10.25
N3	0.96	0.86	0.89	0.89	4.41	4.92	4.78	4.77	**E3**	−8.76	−9.78	−9.51	−9.50
N4	0.91	0.86	0.87	0.87	3.76	4.16	4.05	4.05	**E4**	−8.22	−9.18	−8.92	−8.92
N5	1.17	0.98	0.95	0.95	1.18	2.50	2.63	2.64	**E5**	−7.69	−8.60	−8.39	−8.38
N7	*1.00*	*1.00*	*1.00*	*1.00*	1.68	1.78	1.70	1.70	**E6**	−7.02	−7.82	−7.60	−7.60
N8	1.06	1.07	1.05	1.05	0.84	0.87	0.92	0.91	**E7**	−5.89	−6.58	−6.38	−6.38
N9	1.04	1.02	1.01	1.01	1.16	1.40	1.39	1.38	**E8**	−5.53	−6.17	−6.00	−5.99
N10	1.07	1.05	1.04	1.04	0.79	1.02	1.01	1.01	**E9**	−4.72	−5.26	−5.14	−5.13
N11	1.00	0.91	0.98	0.98	0.65	1.47	0.90	0.93	**E10**	−3.85	−4.30	−4.19	−4.18
N12	1.07	1.07	1.09	1.08	0.06	0.20	0.07	0.15	**E11**	−3.14	−3.49	−3.42	−3.41
N13	1.09	1.09	1.10	1.10	−0.25	−0.10	−0.21	−0.21	**E12**	−2.64	−2.97	−2.92	−2.91
N14	1.06	1.25	1.24	1.24	−0.57	−1.31	−1.16	−1.14	**E13**	−1.36	−1.50	−1.37	−1.37
N15	1.97	2.13	2.10	2.08	−3.65	−3.72	−3.72	−3.73	**E14**	−0.81	−0.87	−0.87	−0.87
N16	1.41	1.54	1.51	1.52	−2.77	−2.87	−2.76	−2.77	**E16**	0.61	0.55	0.67	0.68
N17	1.29	1.15	1.18	1.18	1.33	1.48	1.43	1.43	**E17**	1.48	1.41	1.45	1.45
N18	0.99	0.88	0.90	0.92	1.35	1.50	1.47	1.43	**E18**	2.11	1.93	1.98	1.98
N20	2.08	2.28	2.04	2.04	−3.57	−3.66	−3.54	−3.54	**E19**	2.90	2.81	2.81	2.80
N21	1.77	1.88	1.52	1.57	−4.36	−4.36	−4.24	−4.23	**E20**	3.63	3.73	3.59	3.59
N22	0.81	0.72	0.75	0.75	8.23	9.21	8.93	8.92	**E21**	4.43	4.42	4.49	4.50
N23	0.84	0.75	0.78	0.78	8.57	9.58	9.30	9.29	**E22**	5.01	4.96	4.95	4.95
N24	0.86	0.77	0.80	0.80	10.61	11.87	11.47	11.47	**E23**	5.20	5.17	5.28	5.29
N25	0.83	0.74	0.77	0.77	11.40	12.76	12.37	12.35	**E24**	5.24	5.18	5.26	5.25
N26	0.70	0.63	0.65	0.65	12.56	14.07	13.59	13.62	**E25**	5.47	5.40	5.52	5.52
N27	0.81	0.72	0.76	0.76	13.36	14.98	14.42	14.42	**E26**	5.48	5.41	5.47	5.47
N28	0.89	0.79	0.82	0.82	7.48	8.36	8.14	8.14	**E27**	6.23	6.13	6.19	6.19
N29	1.00	0.89	0.91	0.91	5.21	5.82	5.67	5.65	**E28**	6.70	6.61	6.65	6.64
N30	0.90	0.82	0.85	0.85	3.09	3.47	3.40	3.39	**E29**	6.74	6.64	6.69	6.68
N31	0.91	0.82	0.86	0.86	5.41	6.04	5.88	5.87	**E30**	6.87	6.75	6.79	6.78
N32	0.89	0.82	0.85	0.84	5.46	6.07	5.92	5.94	**E31**	7.52	7.31	7.23	7.23
N34	0.96	0.85	0.89	0.89	6.22	6.93	6.73	6.73	**E32**	7.96	7.60	7.52	7.52
N35	1.11	0.99	0.99	1.00	3.61	4.05	3.95	3.92					
N38	1.17	1.11	1.18	1.18	0.65	0.81	0.69	0.69					
N40	1.17	1.38	1.45	1.46	0.90	0.66	0.49	0.49					
N42	0.98	1.09	1.06	1.07	1.11	0.98	1.00	0.98					
N43	1.06	1.11	1.09	1.08	1.70	1.63	1.60	1.63					
$\mathrm{RMSE}^{(1.0)}$		0.11	0.10	0.10		0.72	0.54	0.54	$\mathrm{RMSE}^{(1.0)}$		0.51	0.38	0.38
$\mathrm{RMSE}^{(2.0)}$	0.10	0.07	0.01		0.54	0.22	0.02		$\mathrm{RMSE}^{(2.0)}$	0.38	0.15	0.01	

those of the anchor species (case 3). Consequently, both criteria have a significant effect on the change in reactivity parameters. Applying the criteria of cases 2 and 3 simultaneously (case 4), the root-mean-square error increases to 0.52. This deviation is larger than the sum of deviations caused by these effects individually. Note that for the calculation of the root-mean-square error, we considered only those reactivity parameters of the case-specific reference set (i.e., without removed species).

In Table 2, we report the reactivity parameters of the 2021 reference set, where version 1.0 refers to the original parameters by Mayr and co-workers, and version 1.1 refers to the parameters of case 4. The sensitivity parameter s_N generally decreases, but increases especially for nucleophiles that already exhibited above-average sensitivity values. This behavior is observed, e.g., for **N14**–**N16**, **N20**, and **N21**, the five least reactive nucleophiles of the reference set when sorting by $s_\mathrm{N}N$. The sign of all nucleophilicity parameters N is preserved but their magnitudes significantly increase in almost all cases. This increase is compensated by the increase (decrease) in s_N for nucleophiles with positive (negative) values of N. The large change in the nucleophilicity parameter N of **N5** (causing a change of more than one

order of magnitude in k_2) appears coherent with the findings of the previous subsection. On average, the nucleophilicity parameter N changes by as much as 0.72 units and mostly toward larger values, which is compensated by changes in the electrophilicity parameter E toward consistently smaller values, with an average change of 0.51 units. The model error with respect to the new 2021 reference set decreases by 19% (from 0.11 to 0.09) when employing reactivity parameters of version 1.1 compared to version 1.0.

Model Dispersion vs. Measurement Uncertainty

III. Explicit consideration of (physical) measurement uncertainty is often neglected in optimization studies. However, if its magnitude becomes comparable to the model dispersion σ (Eq. 6), it can significantly alter the optimal values of the parameters under consideration. To estimate the importance of explicitly considering measurement uncertainty, we selected 30 of the 212 valid reactions (cf. Fig. 1) that represent a diverse set of species and cover a wide range of $\log k_{\text{exp}}$ values (-2.5 to $+7.8$). We find a positive dependence of the measurement uncertainty, u, on the value of $\log k_{\text{exp}}$ (Fig. 2). Laser flash photolysis experiments [20], which were carried out to determine k_2 of faster reactions ($\log k_{\text{exp}} >$ ca. 6), appear to introduce larger measurement uncertainties than conventional and stopped-flow UV/Vis spectrophotometry [19, 20]. For the residuals, however, we find no such trend, indicating homogeneous quality of $\log k_{\text{exp}}$ over the full relevance domain.

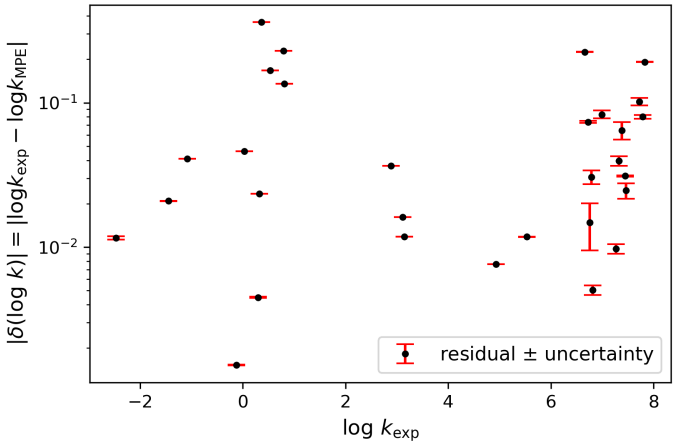


Figure 2: Absolute values of residuals (black dots), $|\delta(\log k)|$, versus $\log k_{\text{exp}}$ are shown for 30 selected reactions of the 2021 reference set. Red error bars represent measurement uncertainty (95% confidence), which show a positive trend with respect to $\log k_{\text{exp}}$. The residuals are consistently larger than their associated 95% confidence intervals (no error bar intercepts with the abscissa), which indicates that measurement uncertainty contributes negligibly to the model error.

We define the average measurement uncertainty (95% confidence) as

$$\bar{u}_{.95} = 1.96 \cdot \bar{u} = 1.96\sqrt{\langle u^2 \rangle} \tag{10}$$

$$\langle u^2 \rangle = \frac{1}{30} \sum_{r=1}^{30} u_r^2 \tag{11}$$

We obtain $\bar{u}_{.95} = 2.78 \times 10^{-3}$, which is significantly smaller than the corresponding model dispersion $\sigma_{.95} = 1.96 \cdot \sigma = 1.70 \times 10^{-1}$. Since uncertainties add up quadratically (assuming they correspond to normally and independently distributed variables), measurement uncertainty provides a negligible contribution of less than 0.2% to the combined uncertainty, $\sqrt{\sigma_{.95}^2 + \bar{u}_{.95}^2}$. A direct comparison of the model residuals (Eq. 3) with individual measurement uncertainties (95% confidence) shows that the former are constantly larger than the latter, from a factor of 2.78 up to several thousands (Fig. 2). Even a factor of 2.78 represents $1.96 \cdot 2.78 = 5.45$ standard deviations compared to 1.96 standard deviations that already correspond to 95% of the area under a normal distribution. We conclude that we can safely neglect measurement uncertainty in the context of reactivity scales.

One should note that distributions of residuals are not necessarily normal/Gaussian. To validate our implicit normality assumption, we compared the empirical distribution of residuals with a Gaussian fit to it (Fig. 3). In relation to

the Gaussian fit, the empirical distribution exhibits more density around its center and toward its tails, and less density in the intermediate regime. Overall, we consider the Gaussian fit a reasonable approximation to the empirical distribution, which is also reflected by the respective 95% confidence intervals: $\sigma_{.95} = 1.70 \times 10^{-1}$ (normal distribution), $Q_{.95} = 1.76 \times 10^{-1}$ (empirical distribution). The latter quantity refers to the distribution of absolute values of the residuals [28, 29].

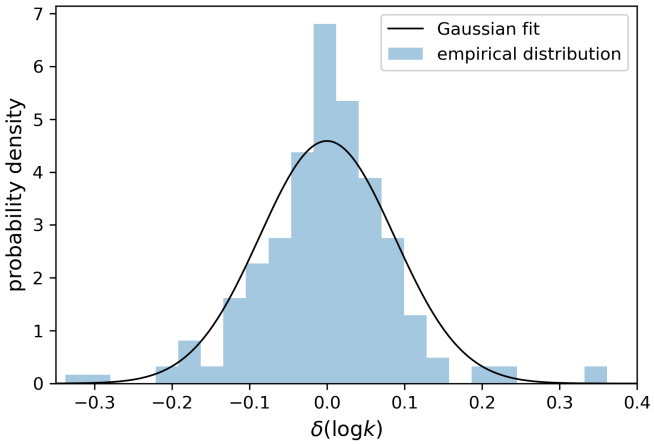


Figure 3: Empirical distribution of residuals (version 1.1), $\delta(\log k)$, versus a Gaussian fit to that distribution.

Revised Reactivity Parameters 1.2: The Effect of Discrepancy Weighting

IV. An insignificant contribution of measurement uncertainty to the model error is not a sufficient condition to neglect non-uniform weights (cf. Eq. 2) in the optimization procedure. Consider the case where the species-specific model dispersion (Eq. 22) of species S is significantly larger than that of the other species. In such a scenario, species S may deteriorate the quality of the overall optimization outcome. One can resolve this situation and process data of potentially heterogeneous quality by applying an iterative re-weighting procedure referred to, by us, as discrepancy weighting. The discrepancy of a model is a measure of its inability to reproduce the reference data within their uncertainty range (here, originating from physical measurements and data post-processing) [7, 11]. The quantification of model discrepancy is an iterative procedure because the weights of the objective function and the species-specific model dispersions are functions of each other. Consequently, the former need to be refined until self-consistency is reached, i.e., until weights and dispersions no longer change. Note that the weights in Eq. 2 refer to reactions and not to species. Hence, for a given reaction, the species-specific model dispersions of the participating nucleophile and electrophile need to be combined to yield a reaction-specific weight. The full procedure is outlined in Appendix A.

For the weighting procedure to be sound from a statistical perspective, it is important that, after reaching self-consistency, the residuals of species S, $\{\delta_r(\log k)\}_S$, are zero-centered ($\mu_S \simeq 0$, cf. Eq. 22) and randomly distributed, i.e., they show no trend with respect to the absolute value of $\log k$. We find that the latter condition is well met as evidenced by the close-to-one correlation coefficient for all species of the reference set. The former condition is also fulfilled as confirmed by $\sigma_S^2/\varepsilon_S^2 \simeq 1$ (cf. Eq. 22) in most cases, although for a small number of cases, the contribution of σ_S^2 to the overall species error ε_S^2 can be as small as 66%. We conclude that discrepancy weighting can be reliably applied in the optimization of reactivity parameters.

Fig. 4 shows the weights of all 212 valid reactions as a function of $\log k_{\text{exp}}$. They are homogeneously distributed around the red baseline representing uniform weights and show no trend with respect to $\log k_{\text{exp}}$. This finding supports our conclusion that measurement uncertainty contributes negligibly to the model error. Otherwise, we would expect a negative trend of the weights with respect to the value of $\log k$. The revised version 1.2 of reactivity parameters (Table 2) mitigates the upward and downward shifts of N and E to some degree, respectively, but clearly has higher resemblance to version 1.1 than to version 1.0. Consequently, the data selection criteria applied in this study affect the reactivity parameters of the reference set significantly more than discrepancy weighting does.

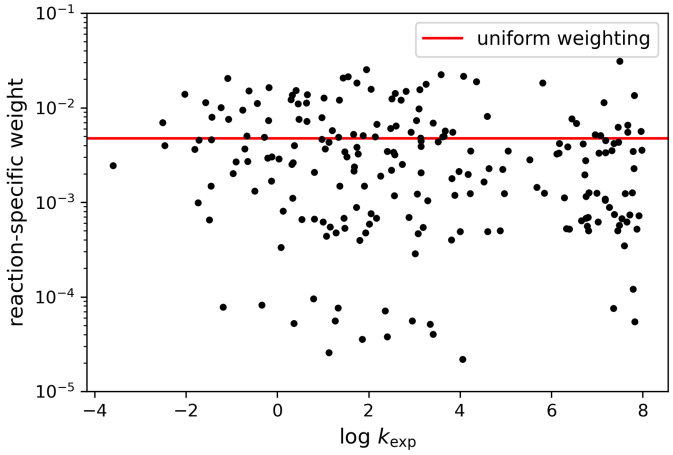


Figure 4: Reaction-specific and non-uniform weights (black dots), $\{w_r\}_{r=1}^{R}$, obtained from discrepancy weighting versus $\log k_{\text{exp}}$ are shown for all 212 valid reactions of the 2021 reference set. The red baseline represents the case of uniform weighting, i.e., $w_r = R^{-1}$ for all $r = 1, ..., R$. The non-uniform weights show no trend with respect to $\log k_{\text{exp}}$.

Revised Reactivity Parameters 2.0: The Effect of Bootstrapping

V. Due to the finite size of the reference set, which additionally covers only a fraction of the reaction matrix it spans (cf. Fig. 1), the optimal values of the reactivity parameters can be expected to carry uncertainty. In order to estimate parameter uncertainty, we applied Bayesian bootstrapping (cf. Appendix B). With this technique, we generated 10,000 synthetic reference sets (bootstrap samples). For each sample, we carried out an individual optimization (using the self-consistent weights of parametrization 1.2), leading to a unique set of optimal reactivity parameters. The sets of 10,000 values per reactivity parameter (102 in total) are referred to as empirical distributions.

For the first time, we can report reactivity parameters that are equipped with quantitative uncertainty measures. We define the first moment (mean) of the empirical parameter distributions as version 2.0 reactivity parameters (Table 2). This most recent parametrization is almost identical to version 1.2, which is indicative of a well-balanced, representative set of reaction data. Uncertainty in s_{N}, N, and E (95% confidence) is located in ranges of 0.02–0.55 (root-mean-square value, RMSV = 0.15), 0.04–1.10 (RMSV = 0.44), and 0.04–0.55 (RMSV = 0.26), respectively. The large uncertainty of 1.10 in the nucleophilicity parameter of **N5** is another indicator of bias (possibly an inconsistency in the documentation of the associated reaction data) that is coherent with the above-mentioned findings.

Empirical parameter distributions can be exploited in several ways to underpin, improve, and find limitations to the reactivity approach by Mayr. First, combining empirical distributions of s_{N}, N, and E yields an empirical distribution of $\log k_{\text{MPE}}$ (see Subsection VI). Second, the uncertainty in reactivity parameters of non-reference species can be estimated in analogy to Mayr's approach. For a non-reference nucleophile, measurements are performed on a series of reactions including reference electrophiles. A calibration model, $f_{\text{E}}(E)$, is constructed which approximates $\log k_{\text{exp}}$ as a function of E,

$$\log k_{\text{exp}} \approx f_{\text{E}}(E) = a + bE \tag{12}$$

Intercept a and slope b are estimated via linear regression, on the basis of which $s_{\text{N}} = b$ and $N = a/b$ can be calculated. Similarly, for a non-reference electrophile, measurements are performed on a series of reactions including reference nucleophiles. A calibration model, $f_{\text{N}}(N|s_{\text{N}})$, is constructed which approximates $\log k_{\text{exp}}/s_{\text{N}}$ as a function of N (parametrized by s_{N}),

$$\frac{\log k_{\text{exp}}}{s_{\text{N}}} \approx f_{\text{N}}(N|s_{\text{N}}) = a + bN \tag{13}$$

Intercept a and slope b are estimated via linear regression, on the basis of which $E = a$ (and $b \simeq 1$) can be calculated. In both cases, 10,000 values are available for (s_{N}, N) and E, respectively, corresponding to reference species. Hence, repeating the calibration 10,000 times (which is computationally efficient) yields empirical distributions of reactivity parameters for non-reference species.

We propose a third and (for now) last way to exploit empirical parameter distributions. A series of theoretical models predicting Mayr-type reactivity parameters were proposed in the past [30, 31, 32, 33, 34, 35, 36, 37, 38]. The predictive power of these models was assessed with respect to some summary statistic (e.g., mean absolute error or root-mean-square error). However, to put the resulting statistics into context, it is necessary to know the uncertainty in the underlying reference values. Ours is the first study providing such uncertainty estimates on a rigorous basis, which allows for assessing previous theoretical work. For instance, regression models were previously employed to predict nucleophilicity N (Orlandi et al. [38], Table 3 of this work) and electrophilicity E (Hoffmann et al. [36], Table 4 of this work) on the basis of quantum-mechanical and empirical descriptors. Version 1.0 reactivity parameters (reference species) and those derived therefrom (non-reference species) served as reference values in both studies. Regarding the reference species, we find that only 21–45% of the predicted reactivity parameters (both N and E) are located inside their 95% confidence intervals, indicating that the theoretical models cannot reproduce the reference values within their uncertainty ranges.

Table 3: Predictions of nucleophilicity, $N^{(\mathrm{O21})}$, by Orlandi et al. [38] for nucleophiles of the reference set. Differences with respect to parametrizations 1.0 (Mayr and co-workers [19, 20]) and 2.0 (this work) are provided, $\Delta N^{(1.0/2.0)} = N^{(1.0/2.0)} - N^{(\mathrm{O21})}$. Parameter uncertainty (95% confidence, assuming normally distributed variables) estimated by us, $\alpha_{.95}(N^{(2.0)})$, is reported. The root-mean-square value (RMSV) as well as the number of differences, $\Delta N^{(1.0/2.0)}$, located inside the 95% confidence interval (NDCI) are provided as summary statistics.

	$N^{(\mathrm{O21})}$	$\Delta N^{(1.0)}$	$\Delta N^{(2.0)}$	$\alpha(N)$		$N^{(\mathrm{O21})}$	$\Delta N^{(1.0)}$	$\Delta N^{(2.0)}$	$\alpha(N)$
N1	10.62	−1.62	−0.78	0.51	**N22**	8.91	−0.68	0.01	0.45
N2	7.93	−1.36	−0.81	0.37	**N25**	11.02	0.38	1.33	0.61
N5	−0.21	1.39	2.85	1.08	**N27**	12.56	0.80	1.86	0.70
N10	0.77	0.02	0.24	0.09	**N34**	5.29	0.93	1.44	0.35
N12	0.02	0.04	0.13	0.52	**N35**	4.37	−0.76	−0.45	0.48
N13	2.58	−2.83	−2.79	0.07	**N38**	2.76	−2.11	−2.07	0.08
N15	−2.09	−1.56	−1.64	0.37	**N42**	1.38	−0.27	−0.40	0.12
N17	2.18	−0.85	−0.75	0.13	**N43**	1.60	0.10	0.03	0.52
N18	1.66	−0.31	−0.23	0.32	RMSV		**1.15**	**1.31**	**0.90**
N20	−3.34	−0.23	−0.20	0.20	NDCI$^{(1.0)}$				**0.32**
N21	−4.07	−0.29	−0.16	0.34	NDCI$^{(2.0)}$				**0.32**

Table 4: Predictions of electrophilicity, $E^{(\mathrm{H20})}$, by Hoffmann et al. [36] for electrophiles of the reference set. Differences with respect to parametrizations 1.0 (Mayr and co-workers [19, 20]) and 2.0 (this work) are provided, $\Delta E^{(1.0/2.0)} = E^{(1.0/2.0)} - E^{(\mathrm{H20})}$. Parameter uncertainty (95% confidence, assuming normally distributed variables) estimated by Hoffmann et al., $\alpha_{.95}(E^{(\mathrm{H20})})$, and by us, $\alpha_{.95}(E^{(2.0)})$, are reported. The root-mean-square value (RMSV) as well as the number of differences, $\Delta E^{(1.0/2.0)}$, located inside the 95% confidence interval (NDCI) are provided as summary statistics.

	$E^{(\mathrm{H20})}$	$\Delta E^{(1.0)}$	$\Delta E^{(2.0)}$	$\alpha_{.95}(E^{(\mathrm{H20})})$	$\alpha_{.95}(E^{(2.0)})$		$E^{(\mathrm{H20})}$	$\Delta E^{(1.0)}$	$\Delta E^{(2.0)}$	$\alpha_{.95}(E^{(\mathrm{H20})})$	$\alpha_{.95}(E^{(2.0)})$
E1	−9.71	−0.33	−1.16	0.24	0.53	**E19**	3.00	−0.10	−0.20	0.33	0.09
E2	−9.78	0.33	−0.47	0.25	0.51	**E20**	3.22	0.41	0.37	0.24	0.14
E3	−8.57	−0.19	−0.93	0.35	0.47	**E21**	4.34	0.09	0.16	0.27	0.14
E4	−8.44	0.22	−0.48	0.24	0.44	**E23**	6.07	−0.87	−0.78	0.20	0.11
E5	−8.04	0.35	−0.34	0.31	0.42	**E24**	5.44	−0.20	−0.19	0.33	0.07
E6	−7.02	0.00	−0.58	0.39	0.39	**E25**	5.15	0.32	0.37	0.20	0.09
E7	−6.16	0.27	−0.22	0.33	0.34	**E26**	5.15	0.33	0.32	0.41	0.08
E8	−6.55	1.02	0.56	0.43	0.32	**E27**	6.10	0.13	0.09	0.24	0.11
E9	−3.80	−0.92	−1.33	0.55	0.29	**E28**	6.70	0.00	−0.06	0.49	0.12
E10	−3.62	−0.23	−0.56	0.43	0.25	**E29**	6.82	−0.08	−0.14	0.35	0.12
E11	−3.06	−0.08	−0.35	0.33	0.23	**E30**	6.69	0.18	0.09	0.25	0.15
E13	−1.17	−0.19	−0.20	0.47	0.12	**E31**	6.73	0.79	0.50	0.29	0.15
E14	−0.70	−0.11	−0.17	0.31	0.16	**E32**	6.53	1.43	0.99	0.35	0.22
E16	0.16	0.45	0.52	0.33	0.07	RMSV		**0.48**	**0.54**	**0.34**	**0.51**
E17	1.25	0.23	0.20	0.35	0.04	NDCI$^{(1.0)}$				**0.31**	**0.45**
E18	2.30	−0.19	−0.32	0.41	0.05	NDCI$^{(2.0)}$				**0.17**	**0.21**

It should be noted that Tables 3 and 4 draw an overly pessimistic picture. On the one hand, both studies included a much larger pool of species than those reported here, comprising several non-reference species. It is well known that

the accuracy of reactivity parameters corresponding to non-reference species is significantly smaller than that observed for reference species [2]. This heterogeneity in accuracy obviously has an effect on theoretical predictions, which we did not take into account in our analysis due to the lack of empirical parameter distributions for non-reference species. On the other hand, our comparison is based on uncertainties corresponding to version 2.0 reactivity parameters, even though the regression models were trained with respect to the currently accepted set of reactivity parameters (version 1.0) [3]. We would like to raise one issue, though. Hoffmann et al. [36] provided uncertainty estimates for reactivity parameters that are a by-product of their regression framework (Gaussian processes [39]). Only 17–31% of their predictions (with respect to reference species only) fall within their 95% confidence intervals, clearly indicating that their model underestimates parameter uncertainty. We observed this behavior of Gaussian processes in another context [18] and concluded to select kernel functions not only with respect to predictive power; they should also yield statistically significant results, i.e., pass a hypothesis test.

Quantification and Assessment of Uncertainty in $\log k_{\text{MPE}}$

VI. Due to the empirical nature of the reactivity parameter distributions, we can propagate uncertainty without assuming some parametrized distribution (e.g., a normal distribution parametrized by mean and variance). That is, for each set of reactivity parameters we obtain one set of $\log k_{\text{MPE}}$ values. From the ensemble of $\log k_{\text{MPE}}$ values for a given reaction we can estimate the contribution of parameter uncertainty to the overall prediction uncertainty (Eq. 8). A heat map comprising uncertainty estimates for $\log k_{\text{MPE}}$ (95% confidence) of the full reaction matrix is shown in Fig. 5. For many of the observed reactions (represented by crosses), the contribution of parameter uncertainty to the overall prediction uncertainty is effectively zero, and model dispersion remains the sole contribution, i.e., $U_{.95} \simeq \sigma_{.95} = 0.21$. For the set of observed reactions, we find prediction uncertainties of 0.21–0.92 (RMSV = 0.25). Taking all combinations of reference nucleophiles and reference electrophiles into account that lie within a range of $-5 < \log k_{\text{MPE}} < 8$, we find a maximum prediction uncertainty of 2.14 (RMSV = 0.50). Consequently, the average accuracy of k_{MPE} that we can expect for any valid combination of reference nucleophile and reference electrophile is within a factor of 10. In most cases, a simple uncertainty pattern can be observed: the larger the distance to an observed reaction in terms of electrophilicity E, the larger the prediction uncertainty (no such trend can be observed with respect to nucleophilicity N or sensitivity-weighted nucleophilicity $s_{\text{N}}N$). This gradual change in uncertainty indicates that information is propagated from observed reactions to similar yet unobserved reactions. We can derive a simple rule for experimental design from this finding: for a given nucleophile, measure $\log k_{\text{exp}}$ for a series of electrophiles that are as equidistant as possible with respect to electrophilicity E.

To assess the quality of our uncertainty estimates, we counted how often the residual of a reaction is located within its 95% confidence interval (hypothesis testing). The result is visualized in Fig. 6A. Only a single residual (or less than 1% of all 212 residuals) is located outside its 95% confidence interval (ideal value: 5%). Hence, our UQ model is rather conservative as it tends to overestimate prediction uncertainty. Overestimation is particularly strong when the contribution of parameter uncertainty to the overall prediction uncertainty tends toward zero. It appears that the model dispersion — a global/constant contribution to the overall prediction uncertainty — is too rough an approximation of the local/reaction-specific model dispersion. Noteworthy, we find a trend between the squared residual, $[\delta(\log k)]^2$, and the squared parameter-related uncertainty in $\log k_{\text{MPE}}$, β^2. We formulated a quadratic regression model, $g(\beta^2)$, to quantify this trend,

$$[\delta(\log k)]^2 \approx g(\beta^2) = a + b_1\beta^2 + b_2\beta^4 \tag{14}$$

Here, a, b_1, and b_2 are the coefficients of the model. We can re-write Eq. 8,

$$U_{.95,r} = 1.96 \cdot \sqrt{(c_r\sigma)^2 + \beta_r^2} \tag{15}$$

where $c_r = 1$ is a uniform weight (not to be confused with the weight w_r of the objective function defined in Eq. 2). The quadratic regression model offers a way to re-define these weights such that $\sum_{r=1}^{R} c_r^2\sigma^2 = \sigma^2 \sum_{r=1}^{R} c_r^2 = \sigma^2 \cdot R$ is a conservation law

$$c_r = \sqrt{R \cdot \frac{g_r(\beta_r^2)}{\sum_{r'=1}^{R} g_{r'}(\beta_{r'}^2)}} \tag{16}$$

Replacing the uniform weights of Eq. 15 with those obtained according to Eq. 16, we obtain an expression of the prediction uncertainty with an effectively local model dispersion (LMD) contribution,

$$U_{.95,r}^{(\text{LMD})} = 1.96 \cdot \sqrt{(c_r\sigma)^2 + \beta_r^2} \tag{17}$$

The resulting hypothesis test is visualized in Fig. 6B. The shape of the updated prediction uncertainty band better reflects the increasing scatter of the residuals (from left to right). Again, a single residual is located outside its 95%

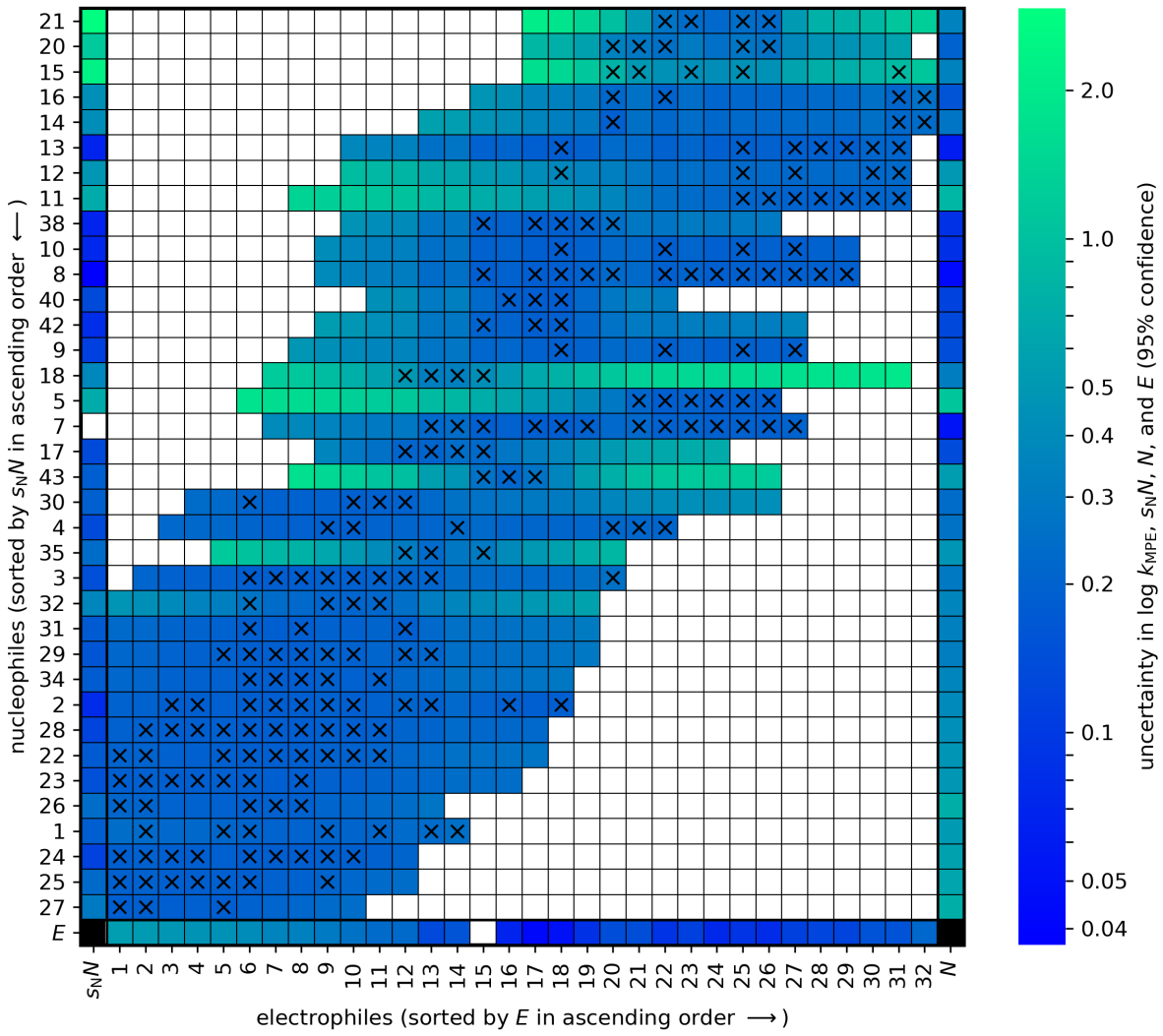


Figure 5: Uncertainty (95% confidence) in $\log k_{\mathrm{MPE}}$, $s_{\mathrm{N}}N$, N, and E. Crosses represent observed reactions. Colored fields without crosses represent unobserved or invalid reactions within a range of $-5 < \log k_{\mathrm{MPE}} < 8$. White fields in the main matrix indicate unobserved or invalid reactions outside that range. White fields outside the main matrix represent anchor species whose reactivity parameters (either s_{N} or E) are fixed.

confidence interval after the update. The UQ model remains conservative, but overall is a much better fit to the actual distribution of residuals.

It should be noted that the hypothesis test is biased somehow and possibly presents an overly optimistic picture as the reactions included in this test were also used to optimize reactivity parameters and quantify prediction uncertainty. As a preliminary test, we split the 212 reference reactions into a training set (166 reactions) and a validation set (46 reactions). We selected the validation reactions (cf. Fig. 1) in a way such that the 2E3N rule was not violated for any species of the 2021 reference set. The training set was subjected to the optimization workflow and subsequent UQ. A hypothesis test (Figs. 6C and 6D) reveals that 9% of the residuals are located outside their 95% confidence intervals. This finding may suggest that our UQ model is too optimistic, but the validation sample size is too small to draw significant conclusions. The decreased training sample size is also problematic: as the number of reactivity parameters remains unchanged, uncertainty estimates are expected be of lower quality than in the previous scenario (Figs. 6A and 6B).

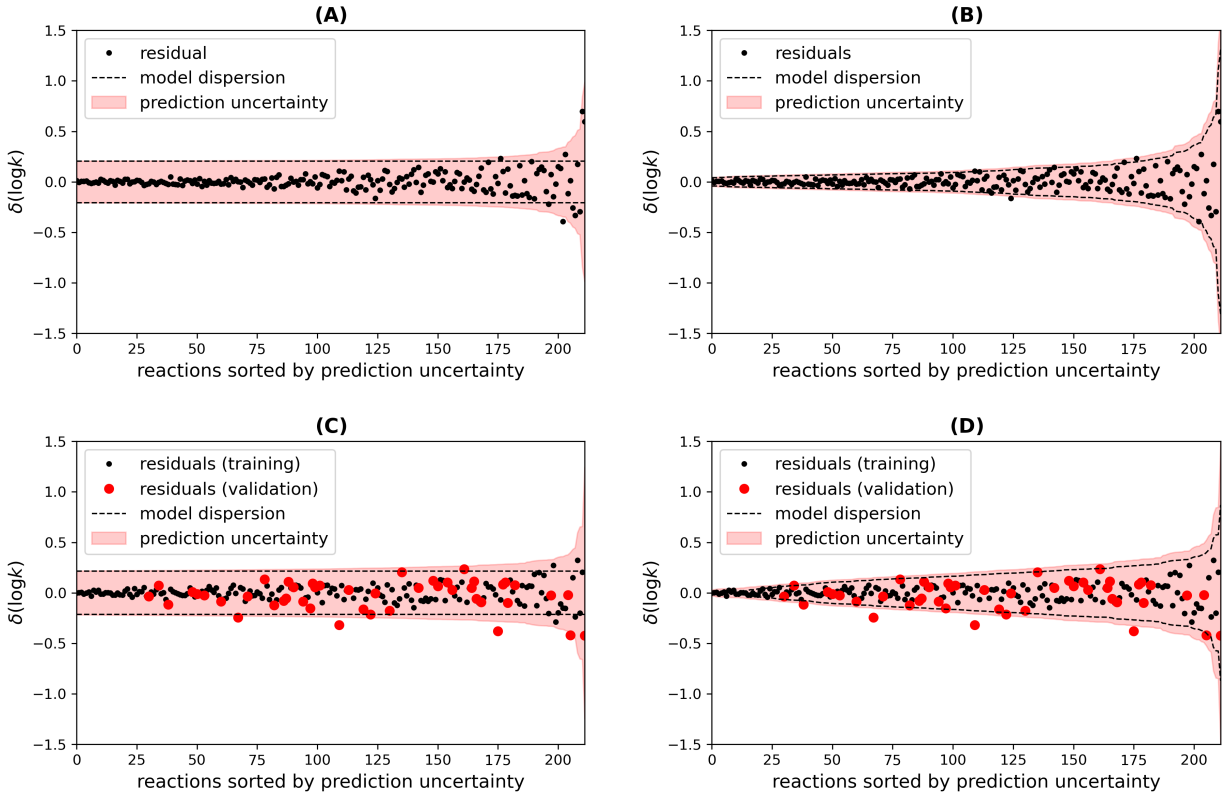


Figure 6: Assessment of uncertainty estimates (95% confidence) for $\log k_{\mathrm{MPE}}$. (**A**) Prediction uncertainty is estimated according to Eq. 8. A single residual (or <1% of all residuals) is located outside its 95% confidence interval. (**B**) Prediction uncertainty is estimated according to Eq. 17. Again, a single residual is located outside its 95% confidence interval. (**C**) Equivalent to (A), but 46 of the 212 reference reactions were excluded from the optimization workflow and subsequent uncertainty quantification. Four of the 46 validation residuals (or 9% of all validation residuals) are located outside their 95% confidence intervals. (**D**) Equivalent to (B), but the same procedure as outlined in (C) was applied. Again, four of the 46 validation residuals are located outside their 95% confidence intervals.

Conclusions and Outlook

We showed that the incorporation of uncertainty quantification (UQ) into the reactivity scale method by Mayr [2] sheds new light on the topic. As a by-product of the UQ-extended reactivity approach, we obtained revised reactivity parameters (102 in total) for 32 electrophiles and 36 nucleophiles of the original reference set [19, 20]. Compared to the original parametrization by Mayr and co-workers, the revised parameters are different by as much as one unit (average changes in s_{N}, N, and E of 0.10, 0.54, and 0.38 units, respectively). It remains to be discussed how these changes could be integrated into Mayr's reactivity database [3]. Since the reactivity parameters of all non-reference species (about 1180 nucleophiles and 300 electrophiles) are derived from the ones of the reference species, our revised set of parameters (version 2.0, cf. Table 2) would have an effect on the entire database.

Most importantly, our UQ-extended reactivity approach allows for *virtual* measurements of $\log k_{\mathrm{MPE}} = s_{\mathrm{N}}(N + E)$, which are reported as expectation $\pm$ deviation — just like *physical* measurements. The random deviation, or prediction uncertainty, associated with $\log k_{\mathrm{MPE}}$ is a reaction-specific quantity. It combines two uncertainty components, one that arises from the mathematical structure of the model (Mayr–Patz equation, MPE) and one that arises from the actual parametrization of that model. The first component, coined model dispersion, equals the standard deviation of the residuals ($\log k_{\mathrm{exp}} - \log k_{\mathrm{MPE}}$) remaining after the optimization of reactivity parameters (s_{N}, N, E), for which we selected the reference set of Mayr's reactivity database [19, 20] . To acknowledge that the quality of $\log k_{\mathrm{MPE}}$ values is not uniformly distributed over the reference species (benzhydrylium ions and π-nucleophiles), we determined weights that reflect this quality and employed them in the optimization procedure.

The second component, coined parameter uncertainty, mainly results from the finite size of the reference set. Two different reference sets can be expected to yield two different sets of optimal (reactivity) parameters. Due to the lack of multiple reference sets, we made use of Bayesian bootstrapping to generate synthetic reference sets from the single reference set at hand. From an ensemble of 10,000 synthetic sets, we obtained empirical distributions for each reactivity parameter, which can be combined to yield uncertainty in $\log k_{\text{MPE}}$. We found prediction uncertainty in $\log k_{\text{MPE}}$ (95% confidence) of 0.21–0.92 for the set of 212 observed reference reactions. For combinations of reference nucleophiles and reference electrophiles that have not yet been observed and lie within a range of $-5 < \log k_{\text{MPE}} < 8$, we found a maximum prediction uncertainty of 2.14. To take into account the possible non-normality of $\log k_{\text{MPE}}$ distributions, we define the following "best practice". For a rough estimation of $\log k$ (which is still expected to be highly accurate in most cases), use the revised reactivity parameters (version 2.0) reported in Table 2. For a critical analysis of $\log k$, we recommend to explicitly calculate the empirical distribution of $\log k_{\text{MPE}}$ [4]. The latter scenario is recommended, e.g., when comparing two competing reactions for which $\log k$ is expected to be similar. We further encourage the community to assess future theoretical predictions of reactivity parameters in the context of parameter uncertainty (as discussed in this study, cf. Tables 3 and 4). Such benchmarks ensure that theoreticians interpret their predictions as critically as possible, but also enable experimentalists to unambiguously evaluate theoretical work.

Uncertainty estimates of $\log k_{\text{MPE}}$ allow us to formulate testable statistical hypothesis, on the basis of which we can assess their quality. We counted how often a residual ($\log k_{\text{exp}} - \log k_{\text{MPE}}$) was located inside the 95% confidence interval of its corresponding $\log k_{\text{MPE}}$ value (the ideal fraction being 95%). We found that >99% of the residuals are located within their 95% confidence intervals when all reactions are used for both parametrization/UQ and hypothesis testing. This finding is indicative of a conservative model, the output of which can be considered an upper bound to the actual prediction uncertainty. When splitting the set of reactions into a training set (166 reactions, subjected to parametrization/UQ) and a validation set (46 reactions, subjected to hypothesis testing), however, we found that 9% of the validation residuals are located outside their 95% confidence intervals. These results appear to be too optimistic but are not yet conclusive due to the small sample size. In future UQ-related work on reactivity scales, the pool of both species and reactions will be increased to draw more robust conclusions. We would also appreciate support by the community in this context. For instance, there are many unobserved combinations ($-5 < \log k_{\text{MPE}} < 8$) present in the reaction matrix of the reference set (Figs. 1 and 5). Measurements of these combinations will further increase the accuracy of reactivity parameters and uncertainty estimates corresponding to reference electrophiles and reference nucleophiles, which are at the heart of Mayr's reactivity scale approach.

In the long run, we aim at deriving reactivity parameters from first-principles calculations, especially for species not yet listed in Mayr's reactivity database. An achievement of this kind would facilitate reactivity predictions to an unprecedented extent due to the resource efficiency and the high automation capacity of computations, thereby reducing experimental expense and accelerating research on polar organic reactivity. Despite their first-principles character, thermochemical calculations are based on approximations that require benchmarking, i.e., an assessment with respect to reference values with well-defined accuracy (here, experimental rate constants, $\log k_{\text{exp}}$). We anticipate that the incorporation of UQ supports our ambition as the reaction matrix spanned by the electrophiles and the nucleophiles of Mayr's reactivity database is rather sparse (cf. Figs. 1 and 5). The vacancies of the reaction matrix (representing unobserved reaction) can be filled by means of our UQ-based approach, allowing for benchmarking *under uncertainty*. The diversity of benchmarkable reactions can be increased remarkably in this way, which increases the significance of conclusions drawn from theoretical studies and is particularly important in the context of data-driven chemical design [40, 41, 42]. Currently, we are benchmarking first-principles models of reactivity against results of this study.

Acknowledgments

J.P. acknowledges funding of this research by the German Research Foundation (DFG) via project 389479699/GRK2455. The authors appreciate generous support by Prof. Dr. Ricardo Mata and thank him, Prof. Dr. Herbert Mayr, Dr. Verena Kraehmer, and Dr. Christopher Stein for insightful discussions and proof-reading of this manuscript.

A Appendix: Discrepancy Weighting

We define the *global* discrepancy d as the square root of the difference between the squared model error (Eq. 4) and the average squared measurement uncertainty (Eq. 11),

$$d = \sqrt{\varepsilon^2 - \langle u^2 \rangle} \tag{18}$$

Note that ε^2 is generally significantly larger than $\langle u^2 \rangle$ and, hence, the square root of a positive number is taken. It is required that the model has been corrected for bias (Eq. 5), such that

$$\varepsilon^2 = \mu^2 + \sigma^2 \approx \sigma^2 \tag{19}$$

In uniformly weighted least-squares optimization (cf. Eq. 2), the model bias is zero by definition. Assuming that measurement uncertainty is also negligible ($\sigma^2 \gg \langle u^2 \rangle$), we can write

$$d^2 = \varepsilon^2 - \langle u^2 \rangle \approx \sigma^2 \tag{20}$$

By design, this approximation also holds true when determining discrepancies for individual species, $S \in \{\mathbf{N1} \cdots \mathbf{N45}, \mathbf{E1} \cdots \mathbf{E33}\}$, i.e.,

$$d_S^2 \approx \sigma_S^2 \tag{21}$$

Since electrophiles and nucleophiles can participate in as few as two or three reactions, we take the *statistical* degrees of freedom of each species explicitly into account,

$$\sigma_S = \sqrt{\nu_S^{-1} \sum_{r \in \mathcal{I}_S} \left[\delta_r(\log k) - \mu_S\right]^2} \approx \sqrt{\nu_S^{-1} \sum_{r \in \mathcal{I}_S} \left[\delta_r(\log k)\right]^2} = \varepsilon_S \tag{22}$$

$$\nu_S = R_S - \gamma_S \tag{23}$$

Here, ν_S constitutes the degrees of freedom of species S, R_S is its number of occurrences, γ_S represents its number of free reactivity parameters, and $\mathcal{I}_S$ is the index set of reactions in which it participates. Hence, the smaller R_S, the larger the effect of γ_S on σ_S will become. The discrepancy $d_{.95}$ of the rth reaction, in which species $S_{\mathrm{N},r}$ and $S_{\mathrm{E},r}$ participate, can then be calculated under the assumption of t-distributed, independent errors,

$$d_{.95,r} = \sqrt{t_{.95,r}^2 \left(\sigma_{S_{\mathrm{N},r}}^2 + \sigma_{S_{\mathrm{E},r}}^2\right)} \tag{24}$$

The reaction-specific t-factor $t_{.95,r}$ corresponds to the folded t-distribution for degrees of freedom ν_r that defines an interval encompassing 95% of the distribution. The degrees of freedom for the rth reaction, ν_r, can be estimated on the basis of the Welch–Satterthwaite equation [43, 44] (in particular, we refer the reader to Eq. 17 of the latter reference),

$$\nu_r = \frac{\left(c_{S_{\mathrm{N},r}} \sigma_{S_{\mathrm{N},r}}^2 + c_{S_{\mathrm{E},r}} \sigma_{S_{\mathrm{E},r}}^2\right)^2}{\frac{c_{S_{\mathrm{N},r}}^2 \sigma_{S_{\mathrm{N},r}}^4}{\nu_{S_{\mathrm{N},r}}} + \frac{c_{S_{\mathrm{E},r}}^2 \sigma_{S_{\mathrm{E},r}}^4}{\nu_{S_{\mathrm{E},r}}}} \tag{25}$$

$$c_S = (\nu_S + 1)^{-1} \tag{26}$$

The reaction-specific degrees of freedom, ν_r are at most as large as the sum of the species-specific degrees of freedom, $\nu_{S_{\mathrm{N},r}}$ and $\nu_{S_{\mathrm{E},r}}$,

$$\nu_r \leq \nu_{S_{\mathrm{N},r}} + \nu_{S_{\mathrm{E},r}} \tag{27}$$

The inverse of the squared discrepancy, $d_{.95,r}^{-2}$, constitutes the weight of the rth reaction. We additionally normalize the weights such that they sum up to one,

$$w_r = \frac{d_{.95,r}^{-2}}{\sum_{r'=1}^{R} d_{.95,r'}^{-2}} \tag{28}$$

In the unweighted case, normalization leads to $w_r = R^{-1}$ for all possible values of r. Normalization does not affect the position of the global minimum of the objective function, but allows for comparability between different sets of weights. It should be noted that discrepancy weighting is an iterative procedure as reaction-specific weights and errors are functions of each other. Hence, we need to update the weights until self-consistency is reached.

B Appendix: Bayesian Bootstrapping

Given R data points, $R-1$ real numbers between zero and one are sampled from a uniform distribution. The numbers 0.0 and 1.0 are added to the tuple of $R-1$ sampled numbers. The tuple is then sorted in ascending order, yielding $q_0 = 0.0 < q_1 < ... < q_{R-1} < q_R = 1.0$. We define $p_r = q_r - q_{r-1}$ as weight of the rth data point (i.e., $p_r = w_r$), which is a number between zero and one. Summing over all weights yields $\sum_{r=1}^{R} p_r = 1$ and, therefore, each weight can be considered the probability of drawing the corresponding data point from the underlying population. Note that if both discrepancy weighting and bootstrapping are applied, the weight of the rth reaction reads

$$w_r = \frac{p_r \cdot d_{.95,r}^{-2}}{\sum_{r'=1}^{R} p_{r'} \cdot d_{.95,r'}^{-2}} \tag{29}$$

We repeat this random procedure B times, representing B bootstrap samples, each characterized by an individual set $\mathcal{P}^{(b)} := \{p_r^{(b)}\}_{r=1}^{R}$. The original sample (the data set at hand) can be characterized by the set $\mathcal{P}^{(0)}$ with a uniform distribution of weights, i.e., $p_r^{(0)} = R^{-1}$ for all possible values of r.

References

[1] Mayr, H.; Lakhdar, S.; Maji, B.; Ofial, A. R. A Quantitative Approach to Nucleophilic Organocatalysis, *Beilstein J. Org. Chem.* **2012,** *8,* 1458–1478.

[2] Mayr, H. Reactivity Scales for Quantifying Polar Organic Reactivity: The Benzhydrylium Methodology, *Tetrahedron* **2015,** *71,* 5095–5111.

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

[4] Proppe, J. "Virtual Measurements of Polar Organic Reactivity", `https://github.com/jproppe/mayr-uq`, 2021; content will be made available as soon as possible.

[5] Kennedy, M. C.; O'Hagan, A. Bayesian Calibration of Computer Models, *J. R. Stat. Soc. B* **2001,** *63,* 425–464.

[6] Pernot, P.; Cailliez, F. A Critical Review of Statistical Calibration/Prediction Models Handling Data Inconsistency and Model Inadequacy, *AIChE J.* **2017,** *63,* 4642–4665.

[7] Proppe, J.; Reiher, M. Reliable Estimation of Prediction Uncertainty for Physicochemical Property Models, *J. Chem. Theory Comput.* **2017,** *13,* 3297–3317.

[8] JCGM, "Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement", 2008.

[9] Taylor, B. N.; Kuyatt, C. E. "NIST Technical Note 1297", 1994.

[10] Ruscic, B. Uncertainty Quantification in Thermochemistry, Benchmarking Electronic Structure Computations, and Active Thermochemical Tables, *Int. J. Quantum Chem.* **2014,** *114,* 1097–1101.

[11] Pernot, P.; Civalleri, B.; Presti, D.; Savin, A. Prediction Uncertainty of Density Functional Approximations for Properties of Crystals with Cubic Symmetry, *J. Phys. Chem. A* **2015,** *119,* 5288–5304.

[12] Mata, R. A.; Suhm, M. A. Benchmarking Quantum Chemical Methods: Are We Heading in the Right Direction?, *Angew. Chem. Int. Ed.* **2017,** *56,* 11011–11018.

[13] Simm, G. N.; Proppe, J.; Reiher, M. Error Assessment of Computational Models in Chemistry, *Chimia* **2017,** *71,* 202–208.

[14] Gallenkamp, C.; Kramm, U. I.; Proppe, J.; Krewald, V. Calibration of Computational Mössbauer Spectroscopy to Unravel Active Sites in FeNC Catalysts for the Oxygen Reduction Reaction, *Int. J. Quantum Chem.* **2021,** *121,* e26394.

[15] Weymuth, T.; Proppe, J.; Reiher, M. Statistical Analysis of Semiclassical Dispersion Corrections, *J. Chem. Theory Comput.* **2018,** *14,* 2480–2494.

[16] Proppe, J.; Gugler, S.; Reiher, M. Gaussian Process-Based Refinement of Dispersion Corrections, *J. Chem. Theory Comput.* **2019,** *15,* 6046–6060.

[17] Proppe, J.; Husch, T.; Simm, G. N.; Reiher, M. Uncertainty Quantification for Quantum Chemical Models of Complex Reaction Networks, *Faraday Discuss.* **2016,** *195,* 497–520.

[18] Proppe, J.; Reiher, M. Mechanism Deduction from Noisy Chemical Reaction Networks, *J. Chem. Theory Comput.* **2019,** *15,* 357–370.

[19] 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.

[20] 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.

[21] Wales, D. J.; Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms, *J. Phys. Chem. A* **1997,** *101,* 5111–5116.

[22] Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, *Nat. Methods* **2020,** *17,* 261–272.

[23] Ruscic, B.; Pinzon, R. E.; Morton, M. L.; von Laszevski, G.; Bittner, S. J.; Nijsure, S. G.; Amin, K. A.; Minkoff, M.; Wagner, A. F. Introduction to Active Thermochemical Tables: Several "Key" Enthalpies of Formation Revisited, *J. Phys. Chem. A* **2004,** *108,* 9979–9997.

[24] Rubin, D. B. The Bayesian Bootstrap, *Ann. Statist.* **1981,** *9,* 130–134.

[25] Hastie, T.; Tibshirani, R.; Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction;* Springer Series in Statistics Springer: New York (NY), United States, 2nd ed ed.; 2009.

[26] Bishop, C. M. *Pattern Recognition and Machine Learning;* Information Science and Statistics Springer: New York (NY), United States, 2006.

[27] Pedregosa, F. *et al.* Scikit-Learn: Machine Learning in Python, *J. Mach. Learn. Res.* **2011,** *12,* 2825–2830.

[28] Pernot, P.; Savin, A. Probabilistic Performance Estimators for Computational Chemistry Methods: The Empirical Cumulative Distribution Function of Absolute Errors, *J. Chem. Phys.* **2018,** *148,* 241707.

[29] Pernot, P.; Savin, A. Probabilistic Performance Estimators for Computational Chemistry Methods: Systematic Improvement Probability and Ranking Probability Matrix. I. Theory, *J. Chem. Phys.* **2020,** *152,* 164108.

[30] 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.

[31] Schindele, C.; Houk, K. N.; Mayr, H. Relationships between Carbocation Stabilities and Electrophilic Reactivity Parameters, E: Quantum Mechanical Studies of Benzhydryl Cation Structures and Stabilities, *J. Am. Chem. Soc.* **2002,** *124,* 11208–11214.

[32] 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.

[33] 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.

[34] 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.

[35] Hoffmann, G.; Tognetti, V.; Joubert, L. On the Influence of Dynamical Effects on Reactivity Descriptors, *Chem. Phys. Lett.* **2019,** *724,* 24–28.

[36] 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.

[37] Mood, A.; Tavakoli, M.; Gutman, E.; Kadish, D.; Baldi, P.; Van Vranken, D. L. Methyl Anion Affinities of the Canonical Organic Functional Groups, *J. Org. Chem.* **2020,** *85,* 4096–4102.

[38] Orlandi, M.; Escudero-Casao, M.; Licini, G. Nucleophilicity Prediction via Multivariate Linear Regression Analysis, *J. Org. Chem.* **2021,** .

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

[40] Aspuru-Guzik, A.; Lindh, R.; Reiher, M. The Matter Simulation (R)Evolution, *ACS Cent. Sci.* **2018,** *4,* 144–152.

[41] dos Passos Gomes, G.; Pollice, R.; Aspuru-Guzik, A. Navigating through the Maze of Homogeneous Catalyst Design with Machine Learning, *Trend Chem.* **2021,** *3,* 96–110.

[42] Pollice, R.; dos Passos Gomes, G.; Aldeghi, M.; Hickman, R. J.; Krenn, M.; Lavigne, C.; Lindner-D'Addario, M.; Nigam, A.; Ser, C. T.; Yao, Z.; Aspuru-Guzik, A. Data-Driven Strategies for Accelerated Materials Design, *Acc. Chem. Res.* **2021,** *54,* 849–860.

[43] Welch, B. L. The Significance of the Difference Between Two Means When the Population Variances Are Unequal, *Biometrika* **1938,** *29,* 350–362.

[44] Satterthwaite, F. E. An Approximate Distribution of Estimates of Variance Components, *Biometrics Bull.* **1946,** *2,* 110–114.