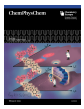
ChemPhysChem

Research Article
doi.org/10.1002/cphc.202200061

Chemistry Europe
European Chemical Societies Publishing

www.chemphyschem.org



Uncertainty Quantification of Reactivity Scales**

Jonny Proppe*[a, b] and Johannes Kircher[a]

According to Mayr, polar organic synthesis can be rationalized by a simple empirical relationship linking bimolecular rate constants to as few as three reactivity parameters. Here, we propose an extension to Mayr's reactivity method that is rooted in uncertainty quantification and transforms the reactivity parameters into probability distributions. Through uncertainty propagation, these distributions can be transformed into uncertainty estimates for bimolecular rate constants. Chemists can exploit these virtual error bars to enhance synthesis planning and to decrease the ambiguity of conclusions drawn from experimental data. We demonstrate the above at the example of the reference data set released by Mayr and co-workers [*J. Am. Chem. Soc.* **2001**, *123*, 9500; *J. Am. Chem. Soc.* **2012**, *134*, 13902]. As by-product of the new approach, we obtain revised reactivity parameters for 36 π-nucleophiles and 32 benzhydrylium ions.

1. 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,[1] it is important that the nucleophile is strong enough (in absolute terms) to attack the intermediate iminium ion but also weak enough (in relative terms) not to react with the precursor carbonyl compound.

Herbert Mayr and co-workers provided unambiguous evidence that a simple empirical relationship, known as the Mayr–Patz equation (MPE),[2] addresses scenarios of this kind reliably,[3]

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

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 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,[4,5] which currently comprises reactivity parameters for 1251 nucleophiles and 345 electrophiles.

In this work, we introduce uncertainty quantification (UQ) into Mayr's reactivity approach. This combined approach, which we made openly available,[6] enables users to perform *virtual* measurements of $\log k$, which are reported as expectation ± 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/inadequacy, and numerical noise.[7–9] A key feature of our UQ approach is the transformation of reactivity parameters into probability distributions, which can be transformed – via uncertainty propagation – into probability 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 ± 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*

[a] Prof. Dr. J. Proppe, J. Kircher
Georg-August University,
Institute of Physical Chemistry
Tammannstrasse 6, 37077 Göttingen (Germany)
[b] Prof. Dr. J. Proppe
Present address:
Technische Universität Braunschweig,
Institute of Physical and Theoretical Chemistry
Gaussstrasse 17, 38106 Braunschweig (Germany)
E-mail: j.proppe@tu-braunschweig.de
Homepage: www.tu-braunschweig.de/en/pci/compmat

[**] *A previous version of this manuscript has been deposited on a preprint server (DOI: 10.26434/chemrxiv-2021-hwh2d-v2)*

Supporting information for this article is available on the WWW under https://doi.org/10.1002/cphc.202200061

Measurement,[10] it is recommended to express uncertainty as a 95% confidence interval. This recommendation is supported by the community.[11–13] 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, our UQ-based approach can support the decision-making process. The larger the overlap of two log k_{MPE} distributions corresponding to competing reactions, the less certain one can discriminate between the two, which makes it more difficult to predict selectivity. The more the overlap tends toward zero (one), the more (less) certain it is that one can predict the relative species flux through competing reaction 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.[14,15] 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 colleagues,[15,16] addressing Mössbauer spectroscopy,[9,17] dispersion corrections to density functional theory,[18,19] reaction kinetics,[20,21] acid-base equilibria,[22] and exchange spin coupling.[23] 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,[24,25] which cover a wide range of log k values (-3.6 to $+8.0$).

1.1. Optimization of Reactivity Parameters

We employed the following objective function for optimizing reactivity parameters,

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

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

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 Doye[26] as implemented in SciPy 1.5.0[27] for minimizing the objective function. It is a global optimization algorithm suited for multivariate non-convex problems. We used the default settings of the basin-hopping algorithm except for the argument `niter`, which we set from `niter=100` to `niter=1` as preliminary tests suggested that a single iteration is sufficient to find optimal reactivity parameters (see Section S1 of the Supporting Information for more details). In the original optimization studies,[24,25] 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 $r' \neq r$. This special case may lead to less-than-optimal results in view of the practitioner's main interest in the reactivity scale approach – the quantitative prediction of (absolute or relative) reaction rates. In 2001, Mayr and co-workers wrote:[24] "Imagine the case that a reaction series, investigated for the elucidation of the reactivity parameters of a structurally unique reagent, matches Eq. 1 only moderately. One would then have to decide whether the benefit of obtaining the new reactivity parameter compensates for the deterioration of the quality of the overall correlation, which is associated with the incorporation of a poorly matching reaction series. An unambiguous decision would often be impossible!" To avoid ambiguity, we argue in favor of a procedure that allows us to include all reaction data but weights them depending on their individual quality. This procedure, which we coin *discrepancy weighting*, assigns an importance (a value between 0 and 1) to each species depending on how well its associated reaction series matches with experimental data (There exist approaches similar to discrepancy weighting, such as the *iteratively reweighted least squares* method[9,13] or the *worst offender* algorithm[28]). These species-specific weights are then combined to yield the reaction-specific weights $\{w_r\}$ of Eq. 2 (see Section 2.4). We point the reader to Appendix A for a full derivation of those weights. Eventually, they can be utilized to determine the uncertainty of reactivity parameters and, as a consequence, of log k_{MPE} on the basis of Bayesian bootstrapping[29] (see Appendix B). The full optimization workflow is summarized in Figure 1.

1.2. Quantification of Uncertainty in log k_{MPE}

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

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

It is equivalent to the definition of log σ by Mayr and co-workers (see Footnote 58 in Ref. [24]). In the case of uniform weights, $w_r = R^{-1}$ for all $r = 1, \ldots, R$, 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_{MPE},

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

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

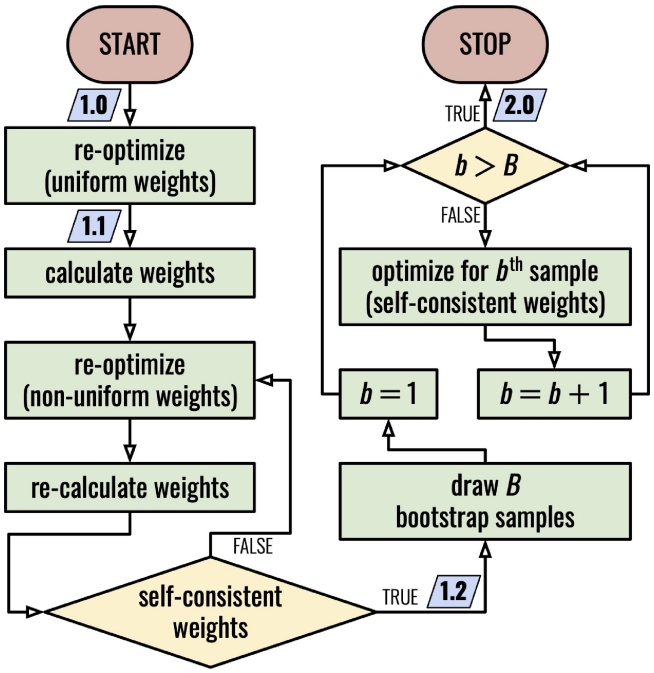


Figure 1. Flowchart illustrating our approach to optimizing reactivity parameters. Version labels (in blue rhomboid boxes) represent a hierarchy of distinct parametrizations.

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

Under the assumption of normally distributed residuals (see Section S3 and Figure S3 of the Supporting Information for validation results), model dispersion represents the model's contribution to prediction uncertainty, i.e., the uncertainty in log k_{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_{MPE} value, leading again to an empirical distribution. We define the parameter-related uncertainty in log k_{MPE} of the rth reaction as

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

Assuming normally distributed variables and independence of the two uncertainty contributions,[30] the prediction uncertainty (95% confidence) corresponding to the rth reaction can be estimated as

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

1.3. Data Selection

All 304 reactions (in dichloromethane) of the 2001 and 2012 studies were considered.[24,25] This pool (Figures 2 and 3, Table 1) encompasses 33 benzhydrylium ions (electrophiles) and 45 π-nucleophiles, and covers a wide range of log k_{exp} values (-3.6 to $+9.2$). The two *anchor* species are **E15** ($E = 0.00$) and **N7** ($s_{\mathrm{N}} = 1.00$);[25] their parameters E and s_{N}, respectively, were kept fixed throughout. We excluded those reactions from the

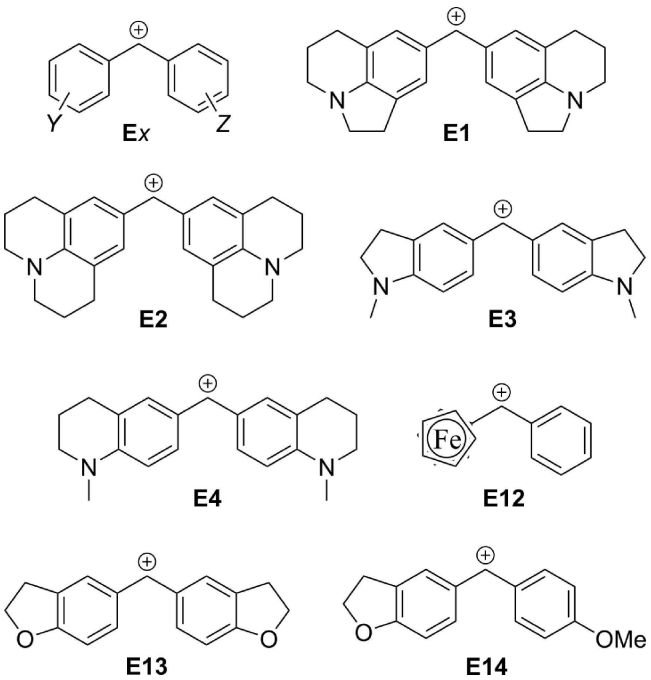


Figure 2. The 2001/12 reference set of electrophiles (benzhydrylium ions). Substituents Y and Z for all electrophiles collectively addressed as **E***x* (x = **5**–**11**, **15**–**33**) are specified in Table 1. The 2022 reference set comprises the same systems except for species **E33**.

Table 1. Specification of substituents for benzhydrylium ions of the reference set (Figure 2).

	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$_2$CF$_3$)	*Y*	**E26**	4-Cl	*Y*
E11	4-N(Ph)(CH$_2$CF$_3$)	*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*

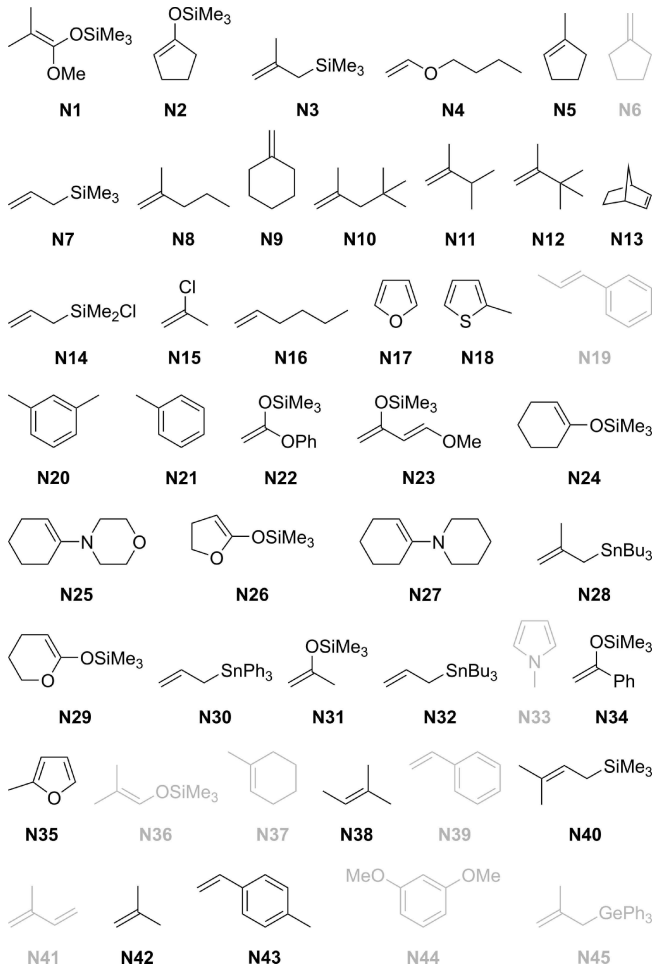


Figure 3. The 2001/12 reference set of nucleophiles (π-systems). The 2022 reference set comprises the same systems except for species **N6**, **N19**, **N33**, **N36**, **N37**, **N39**, **N41**, **N44**, and **N45**.

optimization procedure (30 in total) for which log $k_{exp} > 8$ 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 reactions. 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 (**E15**: no free parameters; **N7**: single free parameter, *N*) need to participate in at least one (**E15**) or two (**N7**) observed reactions. In addition to the 2E3N rule, we required a fully connected network of reactions such that one can traverse from any node (species) to every other node through the edges that represent experimental reaction data (cf. Figure S1). In the 2012 study,[25] the reactivity parameters of several known[24] non-anchor species (**N1**–**N3**, **E1**–**E13**, **E16**–**E20**) were kept fixed. As the 2E3N rule does not apply to non-anchor species with fixed parameters, all systematic errors they embrace will propagate through the reaction network. Reliable UQ, however, requires the elimination of all recognizable sources of systematic error.[10] Therefore, we relaxed all fixed reactivity parameters of non-anchor species, which increased the number of species violating the 2E3N rule.

Applying parameter relaxation and the exclusion criteria mentioned above (log $k_{exp} > 8$, non-standard protocol, violation of 2E3N rule, isolated subnetworks) left us with 212 valid reactions shared among 32 electrophiles and 36 nucleophiles (Figure S1). This set of reactions represents 102 free reactivity parameters, which were optimized as per Figure 1 (Section 2.2). 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 (Section 2.3). For each reaction, there exists a series of observed rate constants, k_{obs} (ordinate), measured with respect to different excess nucleophile concentrations, $[\mathbf{N}]$ (abscissa). The slope of a linear regression model, $f_k([\mathbf{N}])$, represents the bimolecular rate constant k_2,

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

Here, we applied Bayesian linear regression[31] as implemented in Scikit-learn 0.23.1[32] to obtain uncertainty estimates of the regression coefficients (Uncertainty estimates of this kind can also be obtained through ordinary least-squares regression[33]). The uncertainty associated with the slope (k_2) represents the experimental standard deviation of the mean,[10] which is the accepted definition of measurement uncertainty. See Table S2 for experimentally derived values of k_2 and associated uncertainty estimates.

2. Results and Discussion

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

2.1. Reproduction of the 2012 results.[25]
2.2. Application of the data selection criteria introduced in Section 2.3 and re-optimization (uniform weighting), yielding a new set of reactivity parameters referred to as version 1.1.
2.3. Quantification and assessment of measurement uncertainty.
2.4. Re-optimization of reactivity parameters (version 1.2) based on non-uniform weights determined through discrepancy weighting.
2.5. Estimation of empirical parameter distributions via discrepancy-weighted bootstrapping, building the newest set of reactivity parameters (version 2.0).

2.6. Quantification and assessment of prediction uncertainty in log k_{MPE}.

2.1. Reproduction of the 2012 Parametrization

To validate our optimization procedure, we attempted to reproduce the results of the 2012 parametrization study.[25] Both the model error ε (Eq. 4) and model dispersion σ (Eq. 6) equal 0.13 and are 0.8% smaller than the model error determined in 2012 (see also Table S1 for summary statistics). 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/dispersion of 0.13, but a decreased model error difference of 0.3% (with the 2022 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,[25] 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. In Section S1 of the Supporting Information, we show results that support the hypothesis that numerical errors are not the origin of the remaining difference.

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.

2.2. Revised Reactivity Parameters, Version 1.1: The Effect of Data Cleaning

We defined several data selection criteria (cf. Section 1.3), which led to a decrease of the number of reference electrophiles and reference nucleophiles for the sake of consistency. Furthermore, the previously fixed parameters of some non-anchor species[25] were relaxed. These changes affect the reactivity parameters of the reference species, which play a crucial role as they constitute the basis of determining reactivity parameters for any non-reference (i.e., new) species. Given that many publications refer to the original reference parameters of the combined 2001/12 study, changing them can be considered a critical issue. See Section S2 of the Supporting Information for a detailed analysis on how each criterion affects the optimization outcome.

In Table 2, we report the reactivity parameters of the 2022 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 $s_N N$. 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 2022 reference set decreases by 19% (from 0.11 to 0.09) when employing reactivity parameters of version 1.1 compared to version 1.0.

2.3. Quantification and Assessment of Measurement Uncertainty

Explicit consideration of (physical) measurement uncertainty in optimization procedures is often neglected in scientific studies. However, if its values are widely distributed *and* its magnitude becomes a dominant contribution to the model dispersion σ (Eq. 6), it can significantly alter the optimal values of the parameters under consideration. (Model dispersion and model error are interchangeable terms in this case as the model bias equals zero.) To estimate the importance of explicitly considering measurement uncertainty, we selected 30 of the 212 valid reactions (cf. Figure S1 and Table S2) that represent a diverse set of species and cover a wide range of log k_{exp} values (-2.5 to $+7.8$). We find a positive dependence of the measurement uncertainty, u, on the value of log k_{exp} (Figure 4). Laser flash photolysis experiments,[25] which were carried out to determine k_2 of faster reactions (log $k_{exp} >$ ca. 6), appear to introduce larger measurement uncertainty than conventional and stopped-flow UV/Vis spectrophotometry.[24,25] For the residuals, however, we find no such trend, indicating homogeneous quality of log k_{exp} over the full relevance domain (see also Figure S2).

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

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

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 (1.0),[24,25] 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 Section 1.3 (i.e., **N6**, **N19**, **N33**, **N36**, **N37**, **N39**, **N41**, **N44**, **N45**, **E33**) are also not shown. RMSE$^{(1.0)}$ and RMSE$^{(2.0)}$ refer to the root-mean-square error with respect to versions 1.0 and 2.0, respectively. See Table S1 for the corresponding model errors and related statistics.

	$s_N^{(1.0)}$	$s_N^{(1.1)}$	$s_N^{(1.2)}$	$s_N^{(2.0)}$	$N^{(1.0)}$	$N^{(1.1)}$	$N^{(1.2)}$	$N^{(2.0)}$		$E^{(1.0)}$	$E^{(1.1)}$	$E^{(1.2)}$	$E^{(2.0)}$
N1	0.98	0.84	0.87	0.87	9.00	10.13	9.84	9.84	**E1**	−10.04	−11.23	−10.87	−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	−2.91	−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	−4.24	**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	5.66	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	5.91	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					
RMSE$^{(1.0)}$		0.11	0.10	0.10		0.72	0.54	0.54	RMSE$^{(1.0)}$		0.51	0.38	0.38
RMSE$^{(2.0)}$	0.10	0.07	0.01		0.54	0.22	0.02		RMSE$^{(2.0)}$	0.38	0.15	0.01	

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

We obtain $\bar{u}_{.95} = 2.78 \times 10^{-3}$, which explains only 1.6% of the model dispersion, $\sigma_{.95} = 1.96 \cdot \sigma = 1.70 \times 10^{-1}$. 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 (Figure 4). Assuming a factor of 2.78 for all residuals relative to the associated measurement uncertainties, one would obtain $\bar{u}_{.95}/\sigma_{.95} = (1.96 \cdot 2.78)^{-1} = 18\,\%$. This hypothetically high percentage would correspond to $1.96 \cdot 2.78 = 5.45$ standard deviations of the measurement uncertainty. Given that 1.96 standard deviations already correspond to 95% of the area under a normal distribution, a factor of 5.45 effectively corresponds to 100% of that area. We conclude that we can safely neglect measurement uncertainty in the context of reactivity scales.

2.4. Revised Reactivity Parameters 1.2: The Effect of Discrepancy Weighting

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. 19, see Appendix A) 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 discrepancy weighting. The discrepancy of a model is a measure of its inability to reproduce the reference data within their uncertainty range (here,

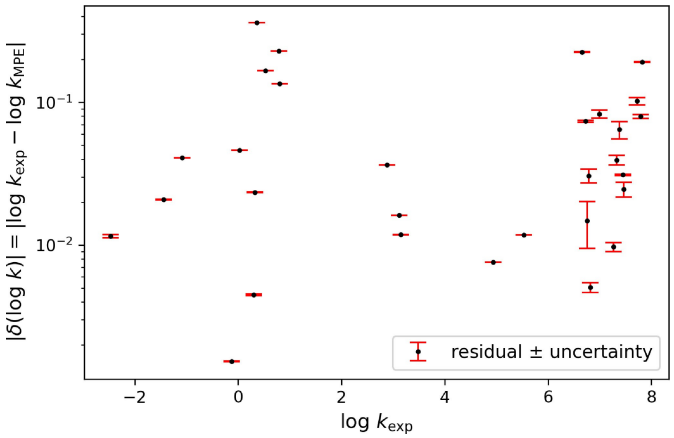


Figure 4. Absolute values of residuals (black dots), $|\delta(\log k)|$, versus log k_{exp} are shown for 30 selected reactions of the 2022 reference set. Version 1.1 parameters were used to calculate log k_{MPE}. Red error bars represent measurement uncertainty (95% confidence), which show a positive trend with respect to log k_{exp} (see also Figure S2 for a scatter plot of the measurement uncertainty versus log k_{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.

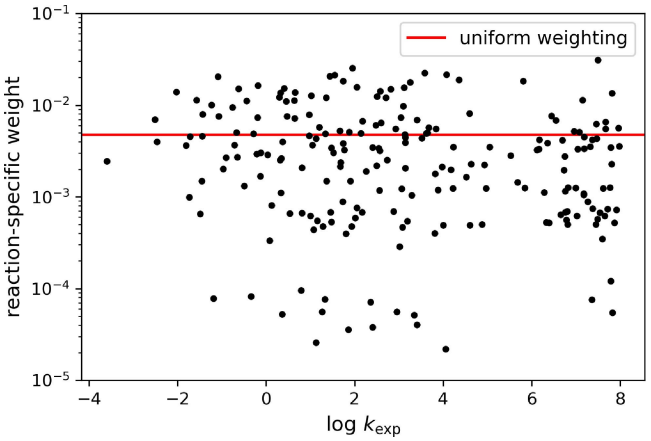


Figure 5. Reaction-specific and non-uniform weights (black dots), $\{w_r\}_{r=1}^{R}$, obtained from discrepancy weighting versus log k_{exp} are shown for all 212 valid reactions of the 2022 reference set. The red baseline represents the case of uniform weighting, i.e., $w_r = R^{-1}$ for all $r = 1, \ldots, R$. The non-uniform weights show no trend with respect to log k_{exp}.

originating from physical measurements and data post-processing).[9,13] 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. See Table S3 for species-specific weights. Reaction-specific weights can be accessed through the project-related GitLab repository.[6]

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. 19) 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 of log k_{exp} versus log k_{MPE} for each species of the reference set. The former condition is also fulfilled as confirmed by $\sigma_S^2/\varepsilon_S^2 \simeq 1$ (cf. Eq. 19) 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 75%. See Table S3 for details. We conclude that discrepancy weighting can be reliably applied in the optimization of reactivity parameters.

Figure 5 shows the weights of all 212 valid reactions as a function of log k_{exp}. They are homogeneously distributed around the red baseline, which represents uniform weights, and show no trend with respect to log k_{exp}. In the previous subsection, we found that measurement uncertainty is overall negligible but increases with log k_{exp}. If measurement uncertainty would, however, contribute significantly to the model error, we would expect a negative trend of the weights with respect to the value of log k. The actual missing trend further supports our conclusion that measurement uncertainty contributes negligibly to the model error.

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.

2.5. Revised Reactivity Parameters 2.0: The Effect of Bootstrapping

Due to the finite size of the reference set, which additionally covers only a fraction of the reaction matrix it spans (cf. Figure S1), 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 referred to as *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 set of 10,000 values per reactivity parameter is referred to as *empirical distribution*.

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 indication of bias that is coherent with the above-mentioned findings. Most of the empirical parameter distributions are symmetric and can be well approximated by a normal distribution, with a tendency of the empirical distribution to be slightly leptokurtic, i.e., it is narrower than the corresponding normal distribution. Parameter uncertainty estimates and histograms of empirical distributions can be accessed through the project-related GitLab repository.[6] A representative example is provided in Figure 6.

Empirical parameter distributions can be exploited in several ways to underpin, improve, and find limitations to the reactivity approach by Mayr. First, the uncertainty in reactivity parameters of non-reference species can be estimated in analogy to Mayr's approach. For a non-reference nucleophile/electrophile, measurements are performed on a series of reactions including reference electrophiles/nucleophiles. Least-squares optimization in accord with Eq. 2 yields the reactivity parameter(s) for the non-reference species. Since 10,000 values are available for s_N, N, and E of the reference species, we can repeat the optimization procedure 10,000 times (which is computationally efficient), resulting in empirical distributions of reactivity parameters also for non-reference species. Second, combining empirical distributions of s_N, N, and E yields an empirical distribution of log k_{MPE} from which its uncertainty can be derived (see Section 2.6 and Figure 6).

We propose a third way to exploiting empirical parameter distributions. A series of theoretical models predicting Mayr-type reactivity parameters were proposed in the past.[34-43] 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.,[42] Table 3 of this work) and electrophilicity E (Hoffmann et al.,[40] 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

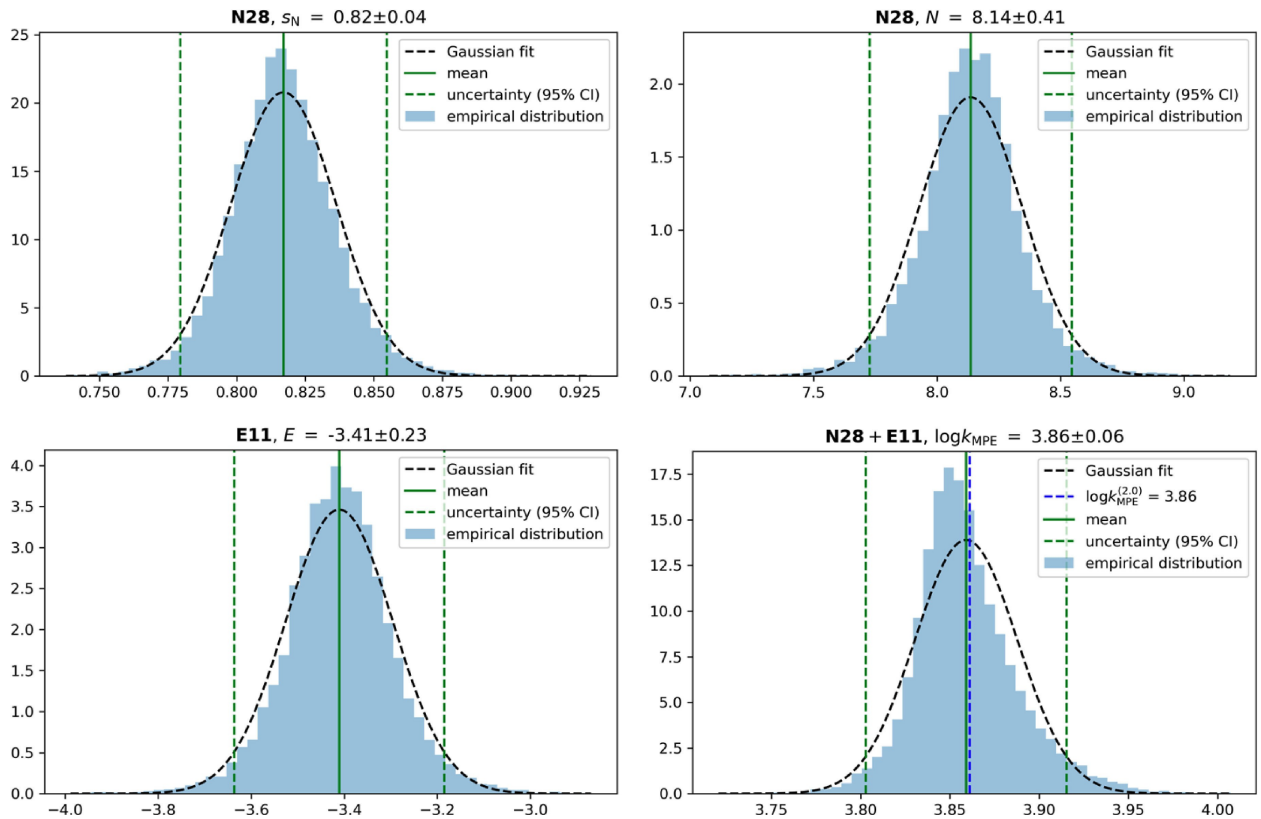


Figure 6. Empirical distributions for nucleophile **N28** (s_N and N), electrophile **E11** (E), and the reaction **N28** + **E11** (log k_{MPE}) obtained from Bayesian bootstrapping. Mean values (green solid line) and symmetric 95% confidence intervals (green dashed line) of the distributions are reported. Corresponding normal distributions are shown as black dashed curves and serve as reference frames. The mean values of the s_N, N, and E distributions (μ_{s_N}, μ_N, and μ_E) correspond to the version 2.0 reactivity parameters reported in Table 2. The blue dashed line in the bottom-right plot represents the value of log k_{MPE} obtained via $\mu_{s_N}(\mu_N + \mu_E)$. It is identical, within two decimals, to the mean value of the asymmetric log k_{MPE} distribution.

Table 3. Predictions of nucleophilicity, $N^{(O21)}$, by Orlandi et al.[42] for nucleophiles of the reference set. Differences with respect to parametrizations 1.0 (Mayr and co-workers[24,25]) and 2.0 (this work) are provided, $\Delta N^{(1.0/2.0)} = N^{(1.0/2.0)} - N^{(O21)}$. 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 percentage of differences, $\Delta N^{(1.0/2.0)}$, located inside the 95% confidence interval (PDCI95) are provided as summary statistics.

	$N^{(O21)}$	$\Delta N^{(1.0)}$	$\Delta N^{(2.0)}$	$\alpha_{.95}(N^{(2.0)})$		$N^{(O21)}$	$\Delta N^{(1.0)}$	$\Delta N^{(2.0)}$	$\alpha_{.95}(N^{(2.0)})$
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**	**0.46**
N20	−3.34	−0.23	−0.20	0.20	$PDCI95^{(1.0)}$				**32%**
N21	−4.07	−0.29	−0.16	0.34	$PDCI95^{(2.0)}$				**32%**

Table 4. Predictions of electrophilicity, $E^{(H20)}$, by Hoffmann et al.[40] for electrophiles of the reference set. Differences with respect to parametrizations 1.0 (Mayr and co-workers[24,25]) and 2.0 (this work) are provided, $\Delta E^{(1.0/2.0)} = E^{(1.0/2.0)} - E^{(H20)}$. Parameter uncertainty (95% confidence, assuming normally distributed variables) estimated by Hoffmann et al., $\alpha_{.95}(E^{(H20)})$, and by us, $\alpha_{.95}(E^{(2.0)})$, are reported. The root-mean-square value (RMSV) as well as the percentage of differences, $\Delta E^{(1.0/2.0)}$, located inside the 95% confidence interval (PDCI95) are provided as summary statistics.

	$E^{(H20)}$	$\Delta E^{(1.0)}$	$\Delta E^{(2.0)}$	$\alpha_{.95}(E^{(H20)})$	$\alpha_{.95}(E^{(2.0)})$		$E^{(H20)}$	$\Delta E^{(1.0)}$	$\Delta E^{(2.0)}$	$\alpha_{.95}(E^{(H20)})$	$\alpha_{.95}(E^{(2.0)})$
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**	**0.35**	**0.26**
E17	1.25	0.23	0.20	0.35	0.04	$PDCI95^{(1.0)}$				**62%**	**45%**
E18	2.30	−0.19	−0.32	0.41	0.05	$PDCI95^{(2.0)}$				**45%**	**21%**

that the theoretical models cannot reproduce the reference values within their uncertainty ranges.

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.[3] 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 by Orlandi et al. and Hoffmann et al. were trained with respect to the currently accepted set of reactivity parameters (version 1.0).[4,5] We would like to raise one issue, though. Hoffmann et al.[40] provided uncertainty estimates for reactivity parameters that are a by-product of their regression framework (Gaussian processes[44]). Only 45–62% of their predictions (with respect to reference species only) fall within their 95% confidence intervals, indicating that their model underestimates parameter uncertainty. We observed this behavior of Gaussian processes in another context[21] and concluded to select kernel functions not only with respect to predictive power; they should also yield statistically reliable results, i.e., about 95% of the predictions should be located inside their 95% confidence intervals.

2.6. Quantification and Assessment of Uncertainty in log k_{MPE}

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_{MPE} values. Histograms and statistics of the corresponding empirical distributions of log k_{MPE} can be accessed through the project-related GitLab repository. All distributions are unimodal, and many of them are clearly asymmetric, see Figure 6 for a representative example. A consequence of this skewness is that log k_{MPE} calculated from version 2.0 reactivity parameters may

not well represent the mean of its empirical distribution, even though we observe such a behavior only for a handful of cases.

From the ensemble of log k_{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_{MPE} (95% confidence) of the full reaction matrix is shown in Figure 7. 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 $-5 < \log k_{MPE} < 8$, 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 $s_N N$). 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_{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

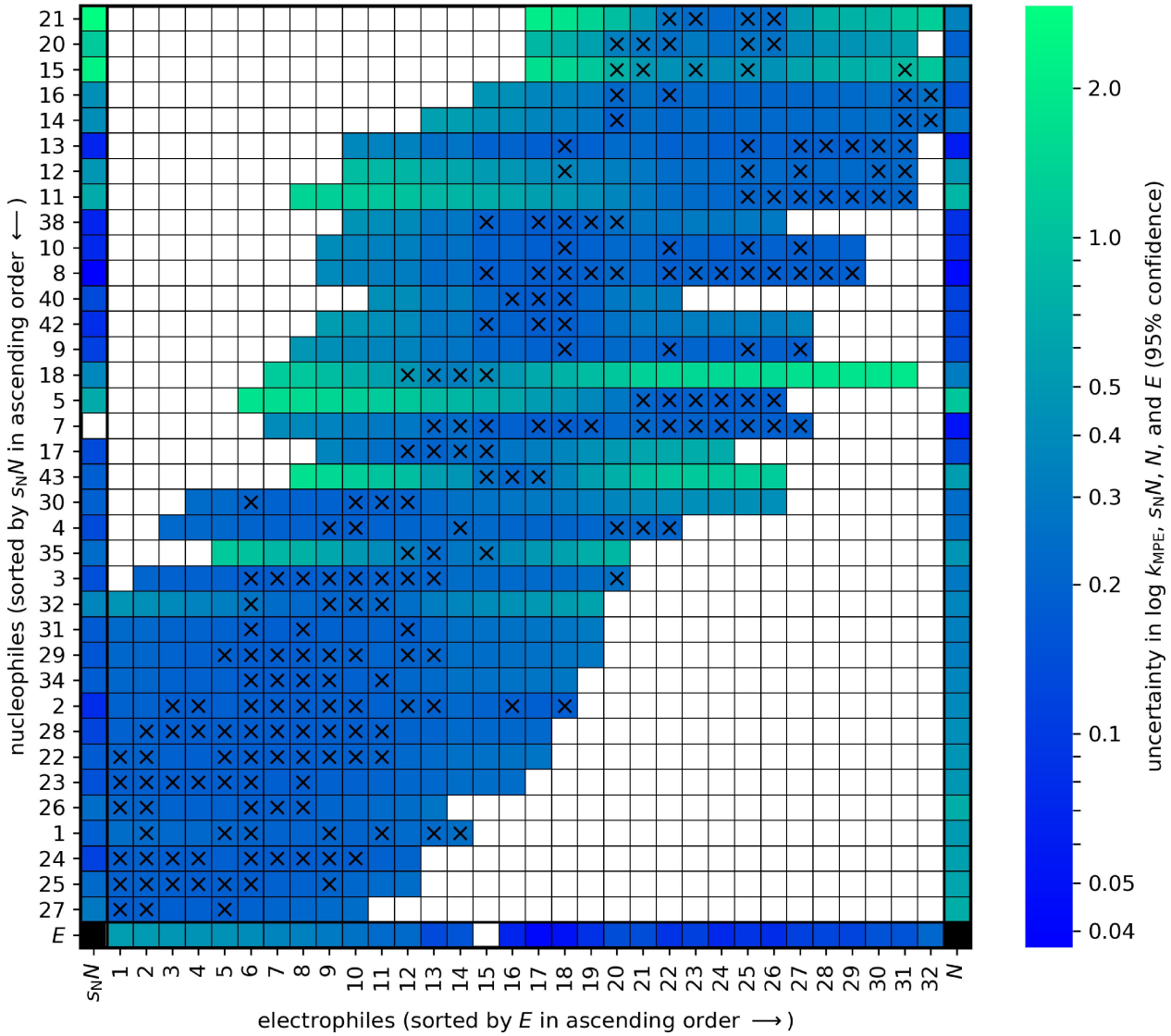


Figure 7. Uncertainty (95% confidence) in log k_{MPE}, $s_N N$, N, and E. Crosses represent observed reactions. Colored fields represent reactions within a range of $-5 < \log k_{MPE} < 8$. White fields in the main matrix indicate reactions outside that range. White fields outside the main matrix represent anchor species whose reactivity parameters (either s_N or E) are fixed.

its 95% confidence interval (hypothesis testing). The result is visualized in Figure 8A. 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 found a trend between the squared residual, $[\delta(\log k)]^2$, and the squared parameter-related uncertainty in log k_{MPE}, β^2 (cf. Eq. 7). This trend is not linear but describes a monotonically increasing function. We found that a quadratic ordinary least-squares regression model, $g(\beta^2)$, appropriately quantifies this trend,

$$[\delta(\log k)]^2 \approx g(\beta^2) = a + b_1\beta^2 + b_2\beta^4 \tag{12}$$

Here, a, b_1, and b_2 are the coefficients of the model. We can generalize Eq. 8 to resolve local model dispersion (LMD),

$$U^{LMD}_{.95,r} = 1.96 \cdot \sqrt{(c_r\sigma)^2 + \beta_r^2} \tag{13}$$

Eqs. 8 and 13 are identical if $c_r = 1$. The weight c_r is 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 the weights of Eq. 13 such that $\sum_{r=1}^{R} c_r^2\sigma^2 = \sigma^2 \cdot \sum_{r=1}^{R} c_r^2 = \sigma^2 \cdot R$ is a conservation law,

$$c_r = \sqrt{R \cdot \frac{g_r(\beta_r^2)}{\sum_{s=1}^{R} g_s(\beta_s^2)}} \tag{14}$$

Replacing the uniform weights with those obtained according to Eq. 14, we obtain an expression of the prediction uncertainty with an effectively local model dispersion. The resulting hypothesis test is visualized in Figure 8B. 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% 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 (R_{train} =166 reactions) and a validation set (R_{val} =46

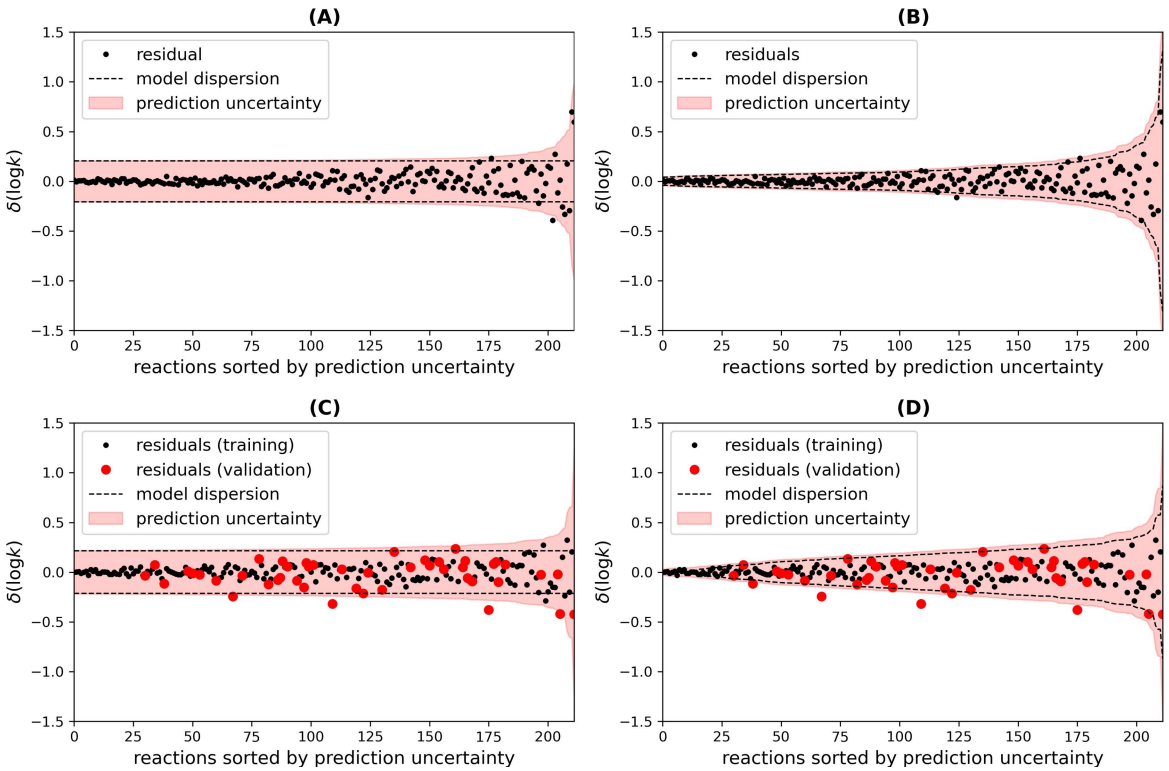


Figure 8. Assessment of uncertainty estimates (95% confidence) for log k_{MPE}. (**A**) Prediction uncertainty is estimated according to Eq. 8. A single residual (or <1% of all residuals) is located outside its 95% confidence interval. (**B**) Prediction uncertainty is estimated according to Eq. 13. 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 (9%) of the 46 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.

reactions). We selected the validation set reactions (cf. Figure S1) in such a way that the 2E3N rule was not violated for any species of the 2022 reference set. The training set was subjected to the optimization workflow and subsequent UQ. A hypothesis test (Figures 8C and 8D) reveals that 9% of the validation set residuals are located outside their 95% confidence intervals. This finding may suggest that our UQ model is too optimistic. Confidence intervals, however, are sample statistics and as such functions of the underlying data set. To estimate the uncertainty of the 95% confidence interval corresponding to the validation set, we calculated the standard error of a binomial distribution,[31] $p(1-p)/R_{\mathrm{val}} = 4\,\%$ with $p = 9\,\%$. Hence, we cannot reject the compatibility with a 95% coverage, although the validation sample size appears to be too small to draw a robust conclusion. 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 (Figures 8A and 8B). Eventually, we would like to point out that the model dispersion can also be understood as a tunable parameter through which a correct calibration of the prediction uncertainty can be ensured. This approach is known as *parameter uncertainty inflation*.[45]

3. Conclusions

We showed that the incorporation of uncertainty quantification (UQ) into the reactivity scale method by Mayr[3] sheds new light on the topic. As a by-product of the UQ-extended reactivity approach, we obtained revised reactivity parameters for 68 reference species. Compared to the original parametrization by Mayr and co-workers,[24,25] the revised parameters differ by as much as one unit. It remains to be discussed how these changes could be integrated into Mayr's reactivity database.[4,5] Since the reactivity parameters of all non-reference species (about 1200 nucleophiles and 300 electrophiles) are derived from the ones of the reference species, our revised set of parameters would affect the entire database.

Our results suggest that the prediction uncertainty associated with $\log k_{\mathrm{MPE}}$ (95% confidence) amounts to 0.21–0.92 units 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 the relevant range of $-5 < \log k_{\mathrm{MPE}} < 8$, we found a maximum prediction uncertainty of 2.14 units. These numbers reflect the accuracy in k_{MPE} estimated previously.[3] To take into account potential non-normality of the empirical $\log k_{\mathrm{MPE}}$ distributions computed by us, we define the following "best practice". For a rough estimation of $\log k$, which is still expected to be highly accurate in most cases, we recommend to 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 $\log k_{\mathrm{MPE}}$ by an interactive tool that can be accessed through the project-related GitLab repository.[6] 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 for $\log k_{\mathrm{MPE}}$ also allowed us to formulate testable statistical hypotheses, on the basis of which we could assess their quality. The estimates appear to be reliable, but the results 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 needs to 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 ($-5 < \log k_{\mathrm{MPE}} < 8$) present in the reaction matrix of the reference set (Figures 7 and S1). 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, $\log k_{\mathrm{exp}}$). 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. Figures S1 and 7). The vacancies of the reaction matrix (representing unobserved reactions) 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.[46–48] Currently, we are benchmarking first-principles models of reactivity against results of this study.

Appendix A: 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),

$$d = \sqrt{\varepsilon^2 - \langle u^2 \rangle} \tag{15}$$

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

$$\varepsilon^2 = \mu^2 + \sigma^2 \approx \sigma^2 \tag{16}$$

In uniformly weighted least-squares optimization (cf. Eq. 2), the model bias is zero by definition (provided it contains a constant term). Assuming that measurement uncertainty is also negligible ($\sigma^2 \gg \langle u^2 \rangle$), as is shown in Section 2.3, we can write

$$d^2 = \varepsilon^2 - \langle u^2 \rangle \approx \sigma^2 \tag{17}$$

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.,

$$d_S^2 \approx \sigma_S^2 \tag{18}$$

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,

$$\sigma_S = \sqrt{\nu_S^{-1} \sum_{r \in \mathcal{I}_S} [\delta_r(\log\, k) - \mu_S]^2} \approx \sqrt{\nu_S^{-1} \sum_{r \in \mathcal{I}_S} [\delta_r(\log\, k)]^2} = \varepsilon_S \tag{19}$$

$$\nu_S = R_S - \gamma_S \tag{20}$$

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,

$$d_{.95,r} = \sqrt{t_{.95,r}^2 \left(\sigma_{S_{\mathrm{N},r}}^2 + \sigma_{S_{\mathrm{E},r}}^2\right)} \tag{21}$$

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[49,50] (in particular, we refer the reader to Eq. 17 of the latter reference),

$$\nu_r = \frac{\left(c_{S_{\mathrm{N},r}} \sigma_{S_{\mathrm{N},r}}^2 + c_{S_{\mathrm{E},r}} \sigma_{S_{\mathrm{E},r}}^2\right)^2}{\frac{c_{S_{\mathrm{N},r}}^2 \sigma_{S_{\mathrm{N},r}}^4}{\nu_{S_{\mathrm{N},r}}} + \frac{c_{S_{\mathrm{E},r}}^2 \sigma_{S_{\mathrm{E},r}}^4}{\nu_{S_{\mathrm{E},r}}}} \tag{22}$$

$$c_S = (\nu_S + 1)^{-1} \tag{23}$$

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

$$\nu_r \leq \nu_{S_{\mathrm{N},r}} + \nu_{S_{\mathrm{E},r}} \tag{24}$$

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,

$$w_r = \frac{d_{.95,r}^{-2}}{\sum_{s=1}^{R} d_{.95,s}^{-2}} \tag{25}$$

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.

Appendix B: Bayesian Bootstrapping

This technique[29] simulates drawing new samples from an underlying but unknown population by assuming that the data set at hand itself is the population. Consequently, only available data is used to draw samples, each of which yields slightly different parameters.

The following procedure describes sampling from a uniform Dirichlet distribution.[31] 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 $q_0 = 0.0 < q_1 < \ldots < q_{R-1} < q_R = 1.0$. 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

$$w_r = \frac{p_r \cdot d_{.95,r}^{-2}}{\sum_{s=1}^{R} p_s \cdot d_{.95,s}^{-2}} \tag{26}$$

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.

Acknowledgments

J.P. acknowledges funding of this research by the German Research Foundation (DFG) via project 389479699/GRK2455. The authors appreciate advice and artistic input (table-of-contents graphic) by Prof. Ricardo A. Mata and thank him, Prof. Herbert Mayr, Dr. Verena Kraehmer, and Dr. Christopher Stein for insightful discussions and proof-reading of this manuscript. Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in GitLab at https://gitlab.com/jproppe/mayruq, reference number 27888788.

Keywords: carbocations · kinetics · nucleophilic addition · reactivity scales · uncertainty quantification

[1] H. Mayr, S. Lakhdar, B. Maji, A. R. Ofial, *Beilstein J. Org. Chem.* **2012**, *8*, 1458–1478.
[2] H. Mayr, M. Patz, *Angew. Chem. Int. Ed.* **1994**, *33*, 938–957; *Angew. Chem.* **1994**, *106*, 990–1010.
[3] H. Mayr, *Tetrahedron* **2015**, *71*, 5095–5111.
[4] H. Mayr, A. R. Ofial, *SAR QSAR Environ. Res.* **2015**, *26*, 619–646.
[5] H. Mayr, A. R. Ofial, *Mayr's Database of Reactivity Parameters*, https://www.cup.lmu.de/oc/mayr/reaktionsdatenbank2/, last accessed on 25 January **2022**.
[6] J. Proppe, *Uncertainty Quantification of Reactivity Scales*, https://gitlab.com/jproppe/mayruq, last accessed on 25 November **2021**.
[7] M. C. Kennedy, A. O'Hagan, *J. R. Stat. Soc. Series B* **2001**, *63*, 425–464.
[8] P. Pernot, F. Cailliez, *AIChE J.* **2017**, *63*, 4642–4665.
[9] J. Proppe, M. Reiher, *J. Chem. Theory Comput.* **2017**, *13*, 3297–3317.
[10] JCGM, *Evaluation of Measurement Data – Guide to the Expression of Uncertainty in Measurement*, **2008**.
[11] B. N. Taylor, C. E. Kuyatt, *NIST Technical Note 1297*, **1994**.
[12] B. Ruscic, *Int. J. Quantum Chem.* **2014**, *114*, 1097–1101.
[13] P. Pernot, B. Civalleri, D. Presti, A. Savin, *J. Phys. Chem. A* **2015**, *119*, 5288–5304.
[14] R. A. Mata, M. A. Suhm, *Angew. Chem. Int. Ed.* **2017**, *56*, 11011–11018; *Angew. Chem.* **2017**, *129*, 11155–11163.
[15] G. N. Simm, J. Proppe, M. Reiher, *Chimia* **2017**, *71*, 202–208.
[16] P. Friederich, F. Häse, J. Proppe, A. Aspuru-Guzik, *Nat. Mater.* **2021**, *20*, 750–761.
[17] C. Gallenkamp, U. I. Kramm, J. Proppe, V. Krewald, *Int. J. Quantum Chem.* **2021**, *121*, e26394.
[18] T. Weymuth, J. Proppe, M. Reiher, *J. Chem. Theory Comput.* **2018**, *14*, 2480–2494.
[19] J. Proppe, S. Gugler, M. Reiher, *J. Chem. Theory Comput.* **2019**, *15*, 6046–6060.
[20] J. Proppe, T. Husch, G. N. Simm, M. Reiher, *Faraday Discuss.* **2016**, *195*, 497–520.
[21] J. Proppe, M. Reiher, *J. Chem. Theory Comput.* **2019**, *15*, 357–370.
[22] J. Uranga, L. Hasecke, J. Proppe, J. Fingerhut, R. A. Mata, *J. Chem. Inf. Model.* **2021**, *61*, 1942–1953.
[23] M. P. Bahlke, N. Mogos, J. Proppe, C. Herrmann, *J. Phys. Chem. A* **2020**, *124*, 8708–8723.
[24] H. Mayr, T. Bug, M. F. Gotta, N. Hering, B. Irrgang, B. Janker, B. Kempf, R. Loos, A. R. Ofial, G. Remennikov, H. Schimmel, *J. Am. Chem. Soc.* **2001**, *123*, 9500–9512.
[25] J. Ammer, C. Nolte, H. Mayr, *J. Am. Chem. Soc.* **2012**, *134*, 13902–13911.
[26] D. J. Wales, J. P. K. Doye, *J. Phys. Chem. A* **1997**, *101*, 5111–5116.
[27] P. Virtanen et al., *Nat. Methods* **2020**, *17*, 261–272.
[28] B. Ruscic, R. E. Pinzon, M. L. Morton, G. von Laszevski, S. J. Bittner, S. G. Nijsure, K. A. Amin, M. Minkoff, A. F. Wagner, *J. Phys. Chem. A* **2004**, *108*, 9979–9997.
[29] D. B. Rubin, *Ann. Statist.* **1981**, *9*, 130–134.
[30] T. Hastie, R. Tibshirani, J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Series in Statistics), Springer: New York (NY), United States, 2nd ed., **2009**.
[31] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics), Springer: New York (NY), United States, **2006**.
[32] F. Pedregosa et al., *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
[33] N. Altman, M. Krzywinski, *Nat. Methods* **2015**, *12*, 999–1000.
[34] P. Pérez, A. Toro-Labbé, A. Aizman, R. Contreras, *J. Org. Chem.* **2002**, *67*, 4747–4752.
[35] C. Schindele, K. N. Houk, H. Mayr, *J. Am. Chem. Soc.* **2002**, *124*, 11208–11214.
[36] C. Wang, Y. Fu, Q. X. Guo, L. Liu, *Chem. Eur. J.* **2010**, *16*, 2586–2598.
[37] F. Pereira, D. A. R. S. Latino, J. Aires-de-Sousa, *J. Org. Chem.* **2011**, *76*, 9312–9319.
[38] L. G. Zhuo, W. Liao, Z. X. Yu, *Asian J. Org. Chem.* **2012**, *1*, 336–345.
[39] G. Hoffmann, V. Tognetti, L. Joubert, *Chem. Phys. Lett.* **2019**, *724*, 24–28.
[40] G. Hoffmann, M. Balcilar, V. Tognetti, P. Héroux, B. Gaüzére, S. Adam, L. Joubert, *J. Comput. Chem.* **2020**, *41*, 2124–2136.
[41] A. Mood, M. Tavakoli, E. Gutman, D. Kadish, P. Baldi, D. L. Van Vranken, *J. Org. Chem.* **2020**, *85*, 4096–4102.
[42] M. Orlandi, M. Escudero-Casao, G. Licini, *J. Org. Chem.* **2021**, *86*, 3555–3564.
[43] D. Kadish, A. D. Mood, M. Tavakoli, E. S. Gutman, P. Baldi, D. L. Van Vranken, *J. Org. Chem.* **2021**, *86*, 3721–3729.
[44] C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press: Cambridge (MA), United States, **2006**.
[45] P. Pernot, *J. Chem. Phys.* **2017**, *147*, 104102.
[46] A. Aspuru-Guzik, R. Lindh, M. Reiher, *ACS Cent. Sci.* **2018**, *4*, 144–152.
[47] G. dos Passos Gomes, R. Pollice, A. Aspuru-Guzik, *Trend Chem.* **2021**, *3*, 96–110.
[48] R. Pollice, G. dos Passos Gomes, M. Aldeghi, R. J. Hickman, M. Krenn, C. Lavigne, M. Lindner-D'Addario, A. Nigam, C. T. Ser, Z. Yao, A. Aspuru-Guzik, *Acc. Chem. Res.* **2021**, *54*, 849–860.
[49] B. L. Welch, *Biometrika* **1938**, *29*, 350–362.
[50] F. E. Satterthwaite, *Biometrics Bull.* **1946**, *2*, 110–114.

Manuscript received: January 25, 2022
Revised manuscript received: February 16, 2022
Accepted manuscript online: February 21, 2022
Version of record online: March 18, 2022