pKcalculator: A pK\textsubscript{a} predictor for C–H bonds

Rasmus M. Borup, Nicolai Ree and Jan H. Jensen\textsuperscript{*}

Abstract
Determining the pK\textsubscript{a} values of various C–H sites in organic molecules offers valuable insights for synthetic chemists in predicting reaction sites. As molecular complexity increases, this task becomes more challenging. This paper introduces pKcalculator, a quantum chemistry (QM)-based workflow for automatic computations of C–H pK\textsubscript{a} values, which is used to generate a training dataset for a machine learning (ML) model. The QM workflow is benchmarked against 695 experimentally determined C–H pK\textsubscript{a} values in DMSO. The ML model is trained on a diverse dataset of 775 molecules with 3910 C–H sites. Our ML model predicts C–H pK\textsubscript{a} values with a mean absolute error (MAE) and a root mean squared error (RMSE) of 1.24 and 2.15 pK\textsubscript{a} units, respectively. Furthermore, we employ our model on 1043 pK\textsubscript{a}-dependent reactions (aldol, Claisen, and Michael) and successfully indicate the reaction sites with a Matthew’s correlation coefficient (MCC) of 0.82.

Introduction
Over the years, the ability to selectively break a C–H bond to create new connections has attracted increasing interest [1]. While past methods allowed for C–H transformations in simple molecules, recent synthetic protocols [2] enable selective C–H activation and diversification in larger molecules. This has, for example, attracted the pharmaceutical industry to implement such C–H transformations to diversify different types of molecules ranging from small drug-like molecules to intermediates and lead compounds. Especially late-stage functionalization is a promising emerging field that allows chemists to efficiently explore the chemical space in complex molecules by exchanging a C–H bond with different functional groups to modify the biological activity of drugs [2]. However, pinpointing which C–H bond is reacting can be challenging. Grzybowski and co-workers recently addressed this gap by predicting pK\textsubscript{a} values for C–H bonds in dimethyl sulfoxide (DMSO) using a graph convolutional neural network (GCNN) [3]. Using a mix of experimental and computed pK\textsubscript{a} data, they achieved a mean absolute error (MAE) of 2.1 pK\textsubscript{a} units. Lee
and co-workers also addressed this problem by creating a general machine learning (ML) model using either a neural network or XGBoost. They trained on experimental pKₐ values in 39 solvents from the “internet Bond-energy Databank” (iBonD). Thus, they could predict the lowest pKₐ value for a wide range of molecules that contain bonds such as N–H, O–H, C–H, S–H, and P–H. However, they reported a scarcity of nonaqueous pKₐ values and achieved a MAE of 1.5 pKₐ units for the solvent DMSO using XGBoost [4,5]. Unfortunately, neither the Grzybowski group nor the Lee group have made their models generally available to other users.

Inspired by the efforts of the Grzybowski group and the Lee group, we have developed Kalkulator, a quantum chemistry (QM)-based workflow for the automatic computation of C–H pKₐ values in DMSO. The computed C–H pKₐ values are then used to generate training data for an ML model using LightGBM [6]. The QM-based workflow and the ML model are freely available under the MIT license.

Methods

Datasets

We compile a dataset of 732 experimental pKₐ values in DMSO from two different sources, Bordwell [7] and iBonD [4]. The Bordwell dataset contains experimental C–H pKₐ values in DMSO from 419 molecules. For the iBonD database, we select experimental C–H pKₐ values in DMSO for 313 molecules. As the iBonD database only contains an image of each molecule, we employ the “Deep Learning for Chemical Image Recognition” software (DECIMER v. 2.0), developed by Rajan and co-workers [8–10]. While DECIMER converts molecular images into SMILES, manual intervention is required to ensure the SMILES string correctly represents the molecule. Finally, to mirror the dataset by Roszak et al. [3], we also incorporate 43 heterocycles without experimental pKₐ values from Shen et al., leaving us with a dataset of 775 compounds [11]. This dataset will be used to calculate QM pKₐ values using our QM workflow described in the next section.

We also create a dataset from Reaxys that contains 1043 pKₐ-controlled reactions. These reactions include 584 aldol, 408 Claisen, and 51 Michael reactions. This dataset is used as an out-of-sample dataset to see how well our ML model predicts the reaction site. Additionally, we use six pharmaceutical intermediates that undergo selective borylation to compare our QM workflow and ML model with experimentally determined reaction sites.

The quantum chemistry-based workflow

Following work by Ree et al. [12–15], we present a fully automated QM-based workflow for computing C–H pKₐ values. A given SMILES string undergoes modifications to produce a list of SMILES for each deprotonated C–H bond. We generate min(1 + 3n_rot, 20) conformers for each SMILES using RDKit (v.2022.09.4) [16,17], where (n_rot) represents the number of rotatable bonds. Each conformer undergoes optimization in dimethyl sulfoxide (DMSO, ε = 47.2) using GFN-FF-xTB [18] and the analytical linearized Poisson–Boltzmann (ALPB) equation [19] as the implicit solvation model. We then remove conformers with relative energies above 3 kcal/mol and select unique conformers by taking the centroids of a Butina clustering using pairwise heavy-atom root mean square deviation (RMSD) with a threshold of 0.5 Å [16,20]. For more information, refer to Supporting Information File 1, section “Selecting unique conformers”.

Subsequently, we re-optimize the remaining conformers in DMSO with GFN2-xTB [21] and the ALPB implicit solvation model to identify the lowest-energy conformer. We then conduct re-optimization in ORCA (v. 5.0.4) [22,23], using the dispersion D4-corrected DFT functional CAM-B3LYP [24,25], the Karlsruhe [26,27] triple-ζ basis set, def2-TZVPD, and the conductor-like polarizable continuum model (CPCM) [28] as the implicit solvation models. CAM-B3LYP is chosen as the optimal functional based on a benchmark study that evaluates the accuracy of different levels of theory, ranging from semiempirical methods (xTB) [21] over composite electronic structure methods (r2SCAN-3c) [29] to DFT methods (CAM-B3LYP) [24,25]. All these methods are evaluated as single-point calculations or optimization and frequency calculations. For comprehensive details, refer to Supporting Information File 1, section “Benchmark study - computational methods”. Hereafter, we check the geometries for imaginary frequencies and use the total thermal energy at 298.15 K. Following the approach of the Grzybowski group [3], we compute the heterolytic dissociation energy through the direct deprotonation reaction, AH_{solv} ⇌ K_{solv}; see Equation 1.

\[
\Delta G^0 = E(A^{-}_{solv}) - E(AH_{solv}).
\]  

For each set of deprotonated C–H sites in a molecule, we determine the minimum heterolytic dissociation energy (\(\Delta G_{min}^{0}\)). Hereafter, we assume a linear relationship between the experimental pKₐ values and \(\Delta G_{min}^{0}\) as this assumption allows us to derive the empirical constants \(a\) and \(b\) and correct any systematic errors; see Equation 2, where \(\Delta G^0\) is replaced by \(\Delta G_{min}^{0}\). After retrieving the empirical constants \(a\) and \(b\), we can determine the QM-computed pKₐ values for all deprotonated C–H sites using Equation 2:

\[
pK_a = a \cdot \Delta G^{0} + b.
\]
Machine learning

The feature descriptor

Recent research shows that the atomic descriptors introduced by Finkelmann et al. [30,31], using charge model 5 (CM5) atomic charges [32], are a great representation of atoms in molecules that can be used in combination with an ML model to predict a variety of properties. These properties encompass the site of metabolism [31,33], the strengths of hydrogen bond donors and acceptors [34-36], and the regioselectivity of electrophilic aromatic substitution reactions [14]. Building on the methodology from Finkelmann et al. [30,31] and Ree et al. [14], we utilize the automated approach to compute CM5 atomic charges from semiempirical tight-binding (GFN1-xTB [37]) calculations. We modify the workflow to enhance the accuracy of the computed CM5 atomic charges. Instead of generating a single random conformer, we produce 20 random conformers from a SMILES string and optimize the structure with molecular mechanics force fields [38] using RDKit [16]. The CM5 atomic charges of the lowest-energy conformer are then used to generate atomic descriptors based on sorting the CM5 charges for a given atom of the input SMILES string. Furthermore, we adjust the shell radius from 5 to 6, improving the performance of the ML model to predict pKa values as detailed in Supporting Information File 1, section “The descriptor”.

Data preparation and hyperparameter optimization

Building on the procedure outlined by Ree et al. [14], we employ the Optuna framework (v. 3.3.0) [39] to identify optimal hyperparameters for LightGBM regression and classification models [6]. Specifically, the Bayesian optimization technique utilizing the tree-structured Parzen estimator is applied for hyperparameter space exploration. For the regression task, the target value are the QM-computed pKa values. For the binary classification task, which aims to predict the site with the lowest QM-computed pKa value, labels are assigned in the following manner: ‘1’ for the lowest QM-computed pKa value (true site) and ‘0’ for all other QM-computed pKa values. As there is sometimes a slight variation between the pKa value and the other pKa values, we also introduce a tolerance where a pKa value within +1 pKa units or +2 pKa units of the lowest pKa value is accepted as ‘1’ to account for these variations, see Supporting Information File 1, section “Machine learning models” for more information. Further, given the significant imbalance between the two classes (with ‘0’s far outnumbering ‘1’s), the hyperparameter scale_pos_weight is invoked during hyperparameter optimization. Finally, we establish a “null model” for the classification task, wherein all sites are predicted as ‘0’.

The dataset with QM-computed pKa values (775 compounds; 3910 pKa values) is initially split randomly by compound into a training set (80%; 620 compounds; 3121 pKa values) and a held-out test set (20%; 155 compounds; 789 pKa values). For each ML model, we carry out a fivefold randomly shuffled cross-validation. Within each fold, the original training set is further split randomly into a new training set (90% of the original training set) and a validation set (10% of the original training set). This allows us to evaluate different models and estimate their performance. Hereafter, each ML model is trained on our original training set and tested against the held-out test set. Finally, we select the best-performing ML model.

Results and Discussion

Computing pKa values

From section “The quantum chemistry-based workflow” above, we can determine the empirical values a and b in Equation 2. For each set of deprotonated sites in a molecule, we extract the computed ΔG° value and fit it against the experimental pKa values. Hereafter, we convert the computed ΔG° to QM-computed pKa values using Equation 2. We then inspect outliers that exceed an absolute pKa unit difference of 5 pKa units between the experimental pKa value and the QM-computed pKa value. We choose an absolute pKa unit difference of 5 pKa units to ensure that the QM-computed pKa is well above the error that is to be expected on the level of theory we are using (CAM-B3LYP). The observed outliers typically result from one of the following reasons: (i) calculation errors concerning the expected minimum pKa site, (ii) discrepancies between literature structures and database structures, (iii) mislabeled experimental pKa values, or (iv) extrapolated pKa values. Notably, the extrapolated pKa values correspond to compounds beyond the scale measurable in DMSO (pKa > 35) because of the autoprotolysis of DMSO (pKa,DMSO = 35) [40,41]. For more information regarding finding and removing outliers, see Supporting Information File 1, section “Finding outliers”. After multiple iterations, we identified 695 molecules to have reliable experimental pKa values and computed ΔG° values. The values for the computed ΔG° are then fitted against the experimental pKa values, leaving us with empirical constants a and b; see Figure 1. We now use the derived linear regression to convert all computed ΔG° values into QM-computed pKa values for our whole dataset (775 compounds). These values are used as target values for the ML part.

Machine learning models for predicting C–H pKa values

To learn and predict C–H pKa values, we train a LightGBM regression model with our generated dataset containing QM-computed pKa values (775 compounds; 3910 pKa values). Hereafter, we correlate and compare the ML-predicted pKa values and the QM-computed pKa values and achieve a MAE and a RMSE of 1.24 and 2.15 pKa units, respectively, for the held-out test set (155 compounds; 789 pKa values), as illus-
trated in Figure 2. When zooming in on the ML-predicted pK\textsubscript{a} values that are not correlating well with the QM-computed pK\textsubscript{a} values, we find C–H sites that are either bridgeheads or where the negative charge is stabilized by resonance. This may be due to the nature of the chosen descriptor vector based on sorted CM5 atomic charges as it may not take into account, for example, steric strain and charge delocalisation. We discuss this further in Supporting Information File 1, section “Outliers for the test set”.

We then compare our ML model with previously reported ML models for predicting pK\textsubscript{a} values, namely, the GCNN C–H pK\textsubscript{a} predictor by Roszak et al. [3] and the XGBoost pK\textsubscript{a} predictor by Yang et al. [5]. Roszak et al. [3] used a mix of experimental data (414 compounds) [7], manually curated DFT data (212 compounds), and previously reported DFT data (194 C–H sites) [11]; they obtained a MAE of 2.18 pK\textsubscript{a} units for their test set. Yang et al. [5] used filtered entries from the iBonD dataset, comprising 15338 compounds and 19397 pK\textsubscript{a} values across 39 solvents [5]. As they not only predict C–H pK\textsubscript{a} values, we cannot compare our result with their best ML model. However, they also report a holistic six-solvent (HM-6S) XGBoost model in DMSO (9.3% of the data), which most likely contains the majority of C–H pK\textsubscript{a} values. For this XGBoost model, they achieved MAE and RMSE values of 1.53 and 2.35 pK\textsubscript{a} units, respectively. A comparison between our ML model, the GCNN model of Roszak et al., and the model of Yang et al. is shown in Table 1. While a direct comparison with these studies is not feasible because of differing datasets, our model surpasses Roszak et al.’s GCNN model by a MAE of 0.94 pK\textsubscript{a} units and outperforms Yang et al.’s HM-6S model by a MAE of 0.29 pK\textsubscript{a} units.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGBM (this work)</td>
<td>1.24</td>
<td>2.15</td>
</tr>
<tr>
<td>GCNN [3]</td>
<td>2.18</td>
<td>—</td>
</tr>
<tr>
<td>XGBoost HM-6S (DMSO)\textsuperscript{a} [5]</td>
<td>1.53</td>
<td>2.35</td>
</tr>
</tbody>
</table>

\textsuperscript{a}HM-6S: Table 7 in their paper.

Predicting the lowest C–H pK\textsubscript{a} value

Now that we can fairly accurately predict pK\textsubscript{a} values with our LightGBM regressor, another use case is to be able to identify the C–H site with the lowest pK\textsubscript{a} value to predict the site of reaction. For this purpose, we treat the task as a binary classification and train both a LightGBM classifier and a LightGMB regressor. As described earlier in section “Data preparation and hyperparameter optimization”, the QM-computed pK\textsubscript{a} values

- Predicting the lowest C–H pK\textsubscript{a} value
- We then compare our ML model with previously reported ML models for predicting pK\textsubscript{a} values, namely, the GCNN C–H pK\textsubscript{a} predictor by Roszak et al. [3] and the XGBoost pK\textsubscript{a} predictor by Yang et al. [5]. Roszak et al. [3] used a mix of experimental data (414 compounds) [7], manually curated DFT data (212 compounds), and previously reported DFT data (194 C–H sites) [11]; they obtained a MAE of 2.18 pK\textsubscript{a} units for their test set. Yang et al. [5] used filtered entries from the iBonD dataset, comprising 15338 compounds and 19397 pK\textsubscript{a} values across 39 solvents [5]. As they not only predict C–H pK\textsubscript{a} values, we cannot compare our result with their best ML model. However, they also report a holistic six-solvent (HM-6S) XGBoost model in DMSO (9.3% of the data), which most likely contains the majority of C–H pK\textsubscript{a} values. For this XGBoost model, they achieved MAE and RMSE values of 1.53 and 2.35 pK\textsubscript{a} units, respectively. A comparison between our ML model, the GCNN model of Roszak et al., and the model of Yang et al. is shown in Table 1. While a direct comparison with these studies is not feasible because of differing datasets, our model surpasses Roszak et al.’s GCNN model by a MAE of 0.94 pK\textsubscript{a} units and outperforms Yang et al.’s HM-6S model by a MAE of 0.29 pK\textsubscript{a} units.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGBM (this work)</td>
<td>1.24</td>
<td>2.15</td>
</tr>
<tr>
<td>GCNN [3]</td>
<td>2.18</td>
<td>—</td>
</tr>
<tr>
<td>XGBoost HM-6S (DMSO)\textsuperscript{a} [5]</td>
<td>1.53</td>
<td>2.35</td>
</tr>
</tbody>
</table>

\textsuperscript{a}HM-6S: Table 7 in their paper.

- Predicting the lowest C–H pK\textsubscript{a} value
- Now that we can fairly accurately predict pK\textsubscript{a} values with our LightGBM regressor, another use case is to be able to identify the C–H site with the lowest pK\textsubscript{a} value to predict the site of reaction. For this purpose, we treat the task as a binary classification and train both a LightGBM classifier and a LightGMB regressor. As described earlier in section “Data preparation and hyperparameter optimization”, the QM-computed pK\textsubscript{a} values

- Predicting the lowest C–H pK\textsubscript{a} value
- Now that we can fairly accurately predict pK\textsubscript{a} values with our LightGBM regressor, another use case is to be able to identify the C–H site with the lowest pK\textsubscript{a} value to predict the site of reaction. For this purpose, we treat the task as a binary classification and train both a LightGBM classifier and a LightGMB regressor. As described earlier in section “Data preparation and hyperparameter optimization”, the QM-computed pK\textsubscript{a} values

- Predicting the lowest C–H pK\textsubscript{a} value
- Now that we can fairly accurately predict pK\textsubscript{a} values with our LightGBM regressor, another use case is to be able to identify the C–H site with the lowest pK\textsubscript{a} value to predict the site of reaction. For this purpose, we treat the task as a binary classification and train both a LightGBM classifier and a LightGMB regressor. As described earlier in section “Data preparation and hyperparameter optimization”, the QM-computed pK\textsubscript{a} values
n-BuLi is commonly used, which is known to lead to the kinetic control by the type of base used. For the reaction in Scheme 1a, the LE site of reaction occurs at the highlighted circle. Our ML model predicts the pK\textsubscript{a} value here to be 16.4. Again, the ML model correctly predicts the most stable carbanion (lowest pK\textsubscript{a} value), but other factors come into play when synthesizing compounds.

Last, we have an example of the Michael reaction in Scheme 1c. Here, both the experimental site of reaction and the ML-predicted site of reaction match. Our ML model predicts the lowest pK\textsubscript{a} value to be 12.5, whereas the second lowest ML-predicted pK\textsubscript{a} value is 21.9 (the least substituted C–H next to a ketone). For more information, see Supporting Information File 1, section “Outliers for Reaxys”.

When we evaluate our ML models on the whole out-of-sample set, we again find that the regression model (MCC of 0.82) outperforms the classification model (MCC of 0.70) when used as a binary classifier as seen in Table 2. While a direct comparison cannot be made between Roszak et al.’s results [3] and ours, we find our result to outperform theirs with an accuracy of 0.96. In general, it is surprising that the LightGBM regressor outperforms our LightGBM classifier as Ree et al. [14] have shown the opposite to be true for electrophilic aromatic substitutions. However, our regression model serves a dual function, that is, it accurately predicts pK\textsubscript{a} values and identifies the reaction site.

### Prediction of ary1 C–H borylation sites

In the previous section, we showed that our ML model is able to predict the reaction site for pK\textsubscript{a}-dependent reactions. Now, we test the ML model on a more complex reaction type, namely, borylation reactions. Caldeweyer et al. [45] presented a workflow to predict the iridium-catalyzed borylation site of ary1 C–H borylation reactions. For more information, see Supporting Information File 1, section “Outliers for Reaxys”.

#### Table 2: Test set performance metrics: comparison between a LightGBM classifier and a LightGBM regressor for binary classification of the lowest pK\textsubscript{a} site. Reaxys performance metrics: comparison between a LightGBM classifier and a LightGBM regressor for binary classification of the reaction site in Reaxys. The best model is marked in bold.

<table>
<thead>
<tr>
<th>method</th>
<th>ACC</th>
<th>MCC</th>
<th>PPV</th>
<th>TPR</th>
<th>TNR</th>
<th>NPV</th>
<th>ACC</th>
<th>MCC</th>
<th>PPV</th>
<th>TPR</th>
<th>TNR</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>null model</td>
<td>0.80</td>
<td>0.80</td>
<td>1.00</td>
<td>0.80</td>
<td>0.87</td>
<td>1.00</td>
<td>0.80</td>
<td>0.80</td>
<td>0.87</td>
<td>1.00</td>
<td>0.87</td>
<td>1.00</td>
</tr>
<tr>
<td>classifier</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.90</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.92</td>
<td>0.70</td>
<td>0.64</td>
<td>0.85</td>
<td>0.93</td>
</tr>
<tr>
<td>regressor</td>
<td>0.99</td>
<td>0.99</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
<td>1.00</td>
<td>0.96</td>
<td>0.82</td>
<td>0.84</td>
<td>0.84</td>
<td>0.98</td>
<td>0.98</td>
</tr>
</tbody>
</table>

\[\text{ACC: accuracy; MCC: Matthew’s correlation coefficient; PPV: precision/positive predictive value; TPR: recall/true-positive rate; TNR: specificity/true-negative rate; NPV: negative predictive value.} \]

\[\text{All predicted pK_{a} values are “0” to highlight the imbalance of the dataset.} \]
bonds (SoBo) [45] and experimentally validated their approach using six pharmaceutical intermediates from medicinal chemistry programs. In the article, they state that “Iridium catalysts ligated by bipyridine ligands catalyze the borylation of the aryl C–H bonds that are most acidic and least sterically hindered...”[45]. For this reason, we tested both our QM workflow and the ML model to see how well they identify the reaction site when only considering the lowest aromatic C–H pK_a value; see Figure 3. For both methods, we identify the possible site of reaction if the pK_a value is within 1.5 pK_a units of the lowest pK_a value. This is slightly different from our previous approach. However, because of the higher complexity of the reaction and the similarity of aromatic C–H sites, we purposely allow the QM workflow and the ML model to assess more sites as ‘1’ or true site. When the pK_a value is within 1.5 pK_a units, we also ensure that we are within the range of the uncertainty of the QM-computed pK_a values, which have a MAE of 1.48, as discussed in section “Computing pK_a values”.

For compound 1, the ML model predicts two low-pK_a sites, indicated by filled circles, of which none corresponds to the experimentally observed site of borylation, indicated by the arrow. However, the QM workflow predicts the correct site as the black ring indicates. Overall, the QM workflow accurately predicts four of the six borylation sites, although, in the case of compounds 2 and 6, there are additional sites with nearly identical pK_a values. In the case of compound 3, most chemists would expect the pK_a of pyrazole C–H sites to be considerably lower than those on the benzene ring, suggesting that factors other than pK_a determine the site of borylation for this compound. In the case of compound 5, the most likely explanation is that the site with the lowest QM-computed pK_a value is sterically hindered compared to the experimentally observed site of borylation. The ML model predicts three borylation sites correctly, but, in the case of compound 5, there are two additional sites with low pK_a values. One failure is for compound 3, where the QM workflow also fails; however, for compounds 1 and 4, the ML model fails, while the QM workflow accurately predicts the site of borylation. This indicates that these compounds are not well represented in the training set.

**Conclusion**

We introduce pKalculator, an automated QM-based workflow that computes C–H pK_a values with a MAE of 1.48 and a
Figure 3: Predicting the site of borylation for a set of six experimentally reported borylation reactions [45]. Arrow: major experimental site/prediction by SoBo; black ring: QM-computed lowest $\text{pK}_a + 1.5$; teal filled circle: ML-predicted lowest $\text{pK}_a + 1.5$.

RMSE of 1.81 when correlating with experimental $\text{pK}_a$ values. We use this method to generate training data for an atom-based regression model that delivers fast and relatively precise predictions with MAE and RMSE values of 1.24 and 2.15, respectively, when correlating with QM-computed $\text{pK}_a$ values. Both methods are freely available under the MIT license. Our workflow can function as a filtering tool for computer-aided synthesis planning for the synthesis of various $\text{pK}_a$-dependent reactions (aldol, Michael, and Claisen), evidenced by its accurate predictions of reaction sites for 1043 reactions (MCC of 0.82). Looking ahead, we aim to explore more reactions that depend on C–H $\text{pK}_a$ values, further enhancing the utility of pKalculator for synthetic chemists. Future iterations will consider factors such as a more extensive and diverse training set, as well as steric hindrance and base reactivity, ensuring even more precise predictions for reaction sites.

Supporting Information
Supporting Information File 1
Additional methods data.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-144-S1.pdf]

Funding
This work was funded by the Independent Research Foundation Denmark (DFF; grant number 1032-00129B).

Conflict of Interest
The authors declare that there are no competing interests.

Author Contributions
Rasmus M. Borup: data curation; formal analysis; investigation; methodology; software; visualization; writing – original draft; writing – review & editing. Nicolai Ree: software; supervision; validation; writing – review & editing. Jan H. Jensen: conceptualization; funding acquisition; project administration; supervision; writing – review & editing.

ORCID iDs
Rasmus M. Borup - https://orcid.org/0000-0002-0878-1345
Nicolai Ree - https://orcid.org/0000-0001-9900-5730
Jan H. Jensen - https://orcid.org/0000-0002-1465-1010

Data Availability Statement
All data that supports the findings of this study is available in the published article and/or the supporting information to this article. The code for the automated workflow and results of the analyzed data are available at https://github.com/jensengroup/pKalculator. Additional data is available at https://sid.erda.dk/sharelink/EyuyjllJdp. The internet Bond-energy Database (iBonD) is accessible for non-profit academic use. Due to licensing restrictions for Reaxys, the Reaxys data cannot be shared. We have provided a list of reaction IDs together with our predictions.

Preprint
A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.26434/chemxrv-2024-56h5h
License and Terms

This is an open access article licensed under the terms of the Beilstein-Institut Open Access License Agreement (https://www.beilstein-journals.org/bjoc/terms), which is identical to the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0). The reuse of material under this license requires that the author(s), source and license are credited. Third-party material in this article could be subject to other licenses (typically indicated in the credit line), and in this case, users are required to obtain permission from the license holder to reuse the material.

The definitive version of this article is the electronic one which can be found at:
https://doi.org/10.3762/bjoc.20.144