Designing Target-Specific Datasets for Regioselectivity Predictions on Complex Substrates

Jules Schleinitz[1], Alba Carretero-Cerdán,[1,2,◊] Anjali Gurajapu,[1,◊] Yonatan Harnik,[3] Gina Lee,[1] Amitesh Pandey,[1] Anat Milo[3,*] and Sarah Reisman[1,*]

[1]*The Warren and Katharine Schlinger Laboratory for Chemistry and Chemical Engineering, Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, United States.*

[2]*Division of Theoretical Chemistry & Biology, CBH School, KTH Royal Institute of Technology, Teknikringen 30, S-10044 Stockholm, Sweden*

[3]*Department of Chemistry, Ben-Gurion University of the Negev, Beer-Sheva 841051, Israel.*

[◊]These authors contributed equally and are listed alphabetically.

*Corresponding authors

ABSTRACT: The development of machine learning models to predict the regioselectivity of C(sp^3)–H functionalization reactions is reported. A dataset for dioxirane oxidations was curated from the literature and used to generate a model to predict the regioselectivity of C–H oxidation. To assess whether smaller, intentionally designed datasets could provide accuracy on complex targets, a series of acquisition functions were developed to select the most informative molecules for the specific target. Active learning-based acquisition functions that leverage predicted reactivity and model uncertainty were found to outperform those based on molecular and site similarity alone. The use of acquisition functions for dataset elaboration significantly reduced the number of datapoints needed to perform accurate prediction, and it was found that smaller, machine-designed datasets can give accurate predictions when larger, randomly selected datasets fail. Finally, the workflow was experimentally validated on five complex substrates and shown to be applicable to predicting the regioselectivity of arene C–H radical borylation. These studies provide a quantitative alternative to the intuitive extrapolation from "model substrates" that is frequently used to estimate reactivity on complex molecules.

Introduction

Data science and machine learning (ML) tools have recently been used to provide quantitative guidance for aspects of synthetic organic chemistry that historically have been largely driven by expert chemical intuition. ML models have been developed to predict reaction conditions,[1,2] reaction yields,[3–7] and reaction selectivity.[8,9] There is great interest in the development of ML models that predict the regioselectivity of direct C–H functionalization reactions, that are controlled by the innate reactivity of the substrate and/or reagent – rather than by a directing group. Predictive models can derisk direct C–H activation in the late-stage of a multistep synthesis campaign, aid synthetic planning,[10] and provide rationale for late-stage diversification efforts. Recent reports have primarily focused on predicting the site of arene C–H functionalization, including arene radical borylation,[11] electrophilic aromatic substitution,[12,13] Ir-catalyzed borylations,[14] and others.[15–17] Despite developments in models for product prediction that can identify the right reaction when provided with reagents and substrates, they remain approximative in challenging regioselectivity situations.[18] More generally, models that predict the regioselectivity of C(sp^3)–H functionalization remain underdeveloped.

The key challenge toward predicting the regioselectivity of innate C(sp^3)–H functionalization in complex molecules is that there are typically multiple sites within the molecule with apparently similar reactivity (Fig. 1a). For certain reactions, expert rules can allow the chemist to intuit the most likely site of reaction. For example, White and coworkers have used an informer library to develop rules that aid prediction of the most likely site of Fe-catalyzed C–H oxidation.[19] Nonetheless, a quantitative model for predicting the site of C(sp^3)–H functionalization could improve deployment of such methods in complex systems that possess multiple similar sites. As an example, Sigman and Davies reported a logistic regression analysis to predict the site of carbene C(sp^3)–H functionalization of silyl ethers based on calculations of ground-state chemical properties of the possible reaction sites.[20]

Along with the opportunities presented by predicting the regioselectivity of innate C(sp^3)–H functionalization, a major obstacle is the development of a dataset to support this task, especially on complex

substrates. In principle, the most straightforward solution to improve model performance would be to increase dataset size and leverage high-throughput experimentation (HTE). HTE has been effectively leveraged to sample *reaction condition* space for a few substrates for which calibration curves can be prepared to enable quantification of the products.[21] However, sampling *substrate* space can be more challenging, often requiring large up front investments of time to generate calibration curves of possible reaction outcomes.[2,22] This challenge is exacerbated when considering innate C(sp^3)–H functionalization since it is not clear *a priori* what the site of reaction will be. Despite progress in automated structure prediction from NMR spectra, the deconvolution of NMR mixtures is currently limited to aromatic compounds.[23] Thus, the purification, characterization, and assignment of the site of C(sp^3)-functionalization on a complex molecule often becomes the rate-limiting step in dataset generation (Fig. 1b).

Additionally, a recent report on C(sp^2)–H borylation shows that, even with an HTE dataset, regioselectivity predictive models may fail when molecules of interest are far from the training set distribution.[11] Despite significant progress in designing substrate scopes to assess the domain of applicability of new methods,[24,25] the extrapolation to complex substrates often remains challenging. The difficulty associated with quantifying the applicability domain and extrapolation capabilities of ML models renders their use on complex targets risky (Fig. 1b).

Here we report a method to efficiently train ML models to predict the regioselectivity of innate C(sp^3)–H functionalization reactions on complex targets. To overcome the experimental constraints that limit dataset size and avoid inaccurate extrapolation from the training set, this approach focuses on target-specific dataset generation. A strategy was developed to select the most informative dataset to accurately predict regioselectivity on the target molecule using acquisition functions (Fig. 1c). Acquisition functions (AFs) are policies on how to choose the next substrate to evaluate. This approach seeks to minimize the distribution shift between training data and the molecule of interest to provide high-performance models in a low-data regime. As a proof-of-concept, we use a small literature dataset to develop a model to predict the regioselectivity of dioxirane C(sp^3)–H oxidation. We then show that the use of AFs enables the design of substrate-specific subsets of the dataset that improve model accuracy on a given substrate compared to models trained on the whole dataset. This task was formulated with the goal of predicting the regioselectivity of C(sp^3)–H oxidation in a complex molecule by performing reactions on commercially available small substrates. We demonstrate that the use of AFs to recommend the substrates improves model performance relative to random selection, and the strategy was experimentally validated on a set of five complex molecules. Finally, we show this general workflow can be applied to a literature dataset for radical arene C–H borylation, demonstrating the generality of the approach. Taken together, the findings illustrate a strategy that both leverages literature data and significantly reduces the number of experiments required to develop high-performing models of regioselectivity.

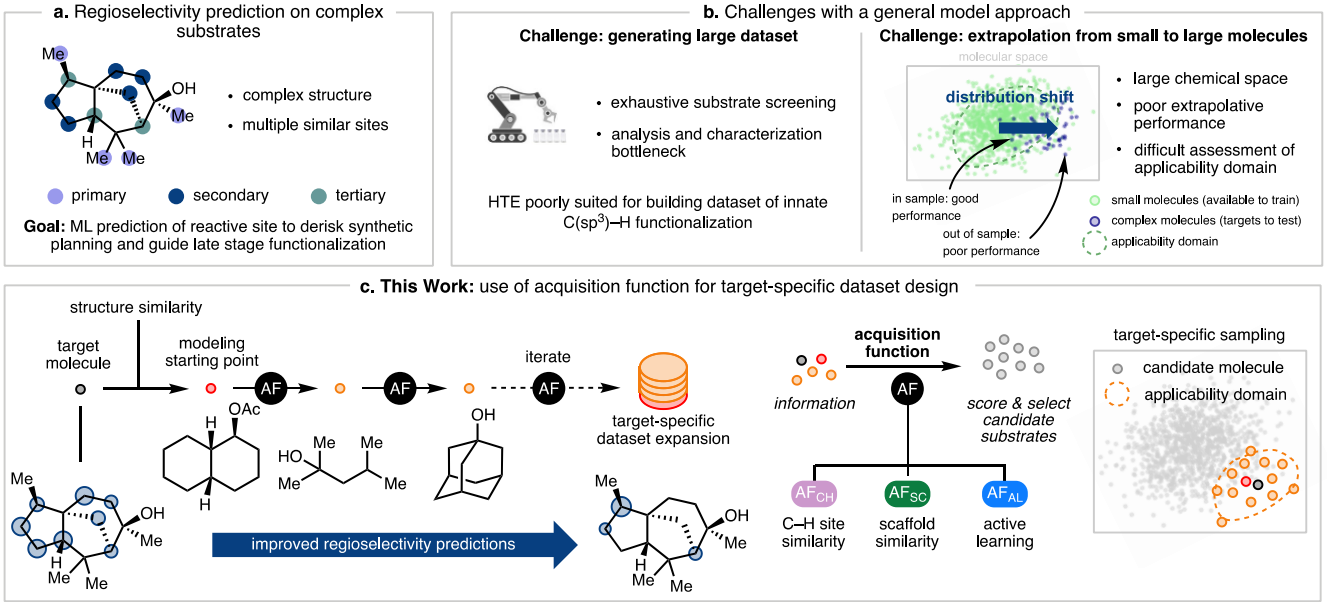

Figure 1. a. Example of late-stage functionalization regioselectivity issues on cedrol. b. Challenges towards building general models. c. This work strategy for complex target regioselectivity prediction.

C(sp^3)–H Oxidation Dataset

As a proof-of-concept, we focused on dioxirane-mediated C–H oxidation reactions (Fig. 2a), which are controlled by the substrate's innate reactivity. We mined reaction regioselectivity data for dimethyl-dioxirane (DMDO) and trifluoromethyl-dioxirane (TFDO),[26–41] curating reports providing detailed information about yields and selectivity. After preprocessing the dataset (details in SI section II.2), 185 unique reactions remained and were used for further modeling. We noticed that (a) reaction conditions vary little across the dataset (details in SI section II.3) and (b) reports showed that TFDO and DMDO maintained the same regioselectivity.[42] Consequently, we decided to leverage data from both dioxirane reagents and rely solely on the description of the C–H bonds, not the reagents, for the design of relevant reaction descriptors.

We began by benchmarking regioselectivity models using different physicochemical descriptors (including steric, electronic and local environment encoding, details in Fig. 2b-c and in SI section III.2) and ML models. Random forest (RF) proved to be best performing model (see SI section IV.3). The performance of regioselectivity models is often evaluated using *n*-fold cross-validation or validation on an expert-designed out-of-sample set. In this context, Nippa *et al.*, elegantly used high-throughput experimentation to validate and enhance a model trained on literature-reported radical borylation reactions.[11] Unsurprisingly, they report that regioselectivity predictions are optimal for substrates similar to those in the training set and a drop in accuracy for more distinct molecules. This common type of distribution shift between the training and test sets is particularly problematic in the context of complex molecule synthesis; structurally complex intermediates proposed for the final steps of a multistep synthesis are inherently "out-of-sample" because they often have never previously been synthesized.[43] The performance of our C–H oxidation model was evaluated by both a leave-one-out (LOO) task and a validation task on complex molecules. The latter was designed specifically to understand how our models performed on the targets of interest while trained on simpler, readily available substrates. The training set contains all molecules with less than 15 carbons (135 molecules) and the test set contains the complex molecules (50 molecules with more than 15 carbons). The complex molecule dataset consists of 7 di- and tri-peptides, 3 taxol derivatives, 3 macrocycles, 22 steroids, and 15 miscellaneous (structures are in SI section II.5).

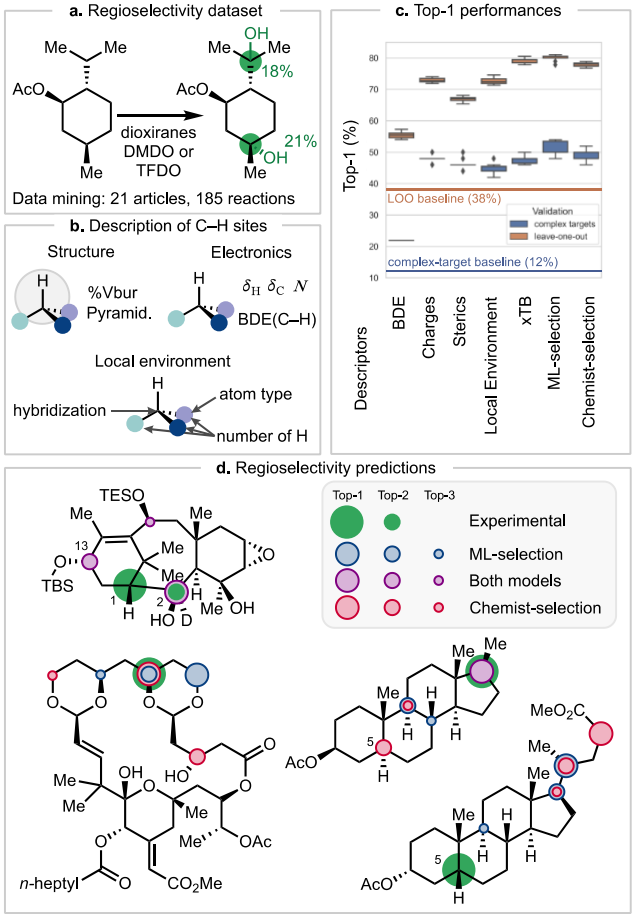

Figure 2. a. Dioxirane C(sp^3)–H oxidation dataset. b. C–H descriptor selection (for further details see SI section III.2). The *Chemist-Selection* has 5 descriptors: BDE, charges of H and C, and %Vbur of H and C of the site. The *ML-selection* has 23 descriptors including charges, BDE, steric, and local environment information. c. Top-1 performances of RF model trained with different sets of descriptors for the C–H bond and validated on leave-one-out and the complex molecules (molecules with more than 15 carbons). d. Regioselectivity predictions of RF trained on the small molecules with the two best sets of features on four selected targets. The experimental observation and the top-1,2,3 of the models are displayed for each target.

To put our results in context, a baseline was designed according to empirical rules-of-thumb on the reactivity of C(sp^3)–H sites which decreases in the following order: benzylic, tertiary, secondary, and primary (see SI section IV.2 for details). Our models significantly outperform this rule-based baseline on the LOO task (~80% top-1 accuracy for the best performing models versus 38% for the baseline detailed in Fig. 2c) and when predicting on large molecules (~50% top-1 accuracy versus 12% top-1 for the baseline). As expected, we observe that predictive performances on the more complex targets are significantly lower than when the models are evaluated as leave-one-out

(performances drop from 80% top-1 to 50% Fig. 2c). Analysis of the different molecules in the complex target dataset reveals that the models achieve good performance on the so-called miscellaneous molecules (13/15 correct top-1) including a large proportion of benzylic positions. Peptides are also well predicted, most likely due to the small number of tertiary C–H sites compared to primary and secondary present in these molecules, though the differentiation between equivalent isopropyl groups is challenging (4/7 correct top-1 and the rest top-2). The models also perform well on the complex macrocycles and most of the steroids. A closer look at the steroids shows that the configuration at C5 plays an important role in the reactivity that the model struggles to grasp (Fig. 2d). Seven out of the 11 steroids having a $C_{5\alpha}$–H configuration are predicted correctly, while the reactive site of the 5β-steroids is never ranked better than top-4. Challenges in distinguishing the reactivity of different ring fusion C–H sites might stem from the model's failure to capture strain release during oxidation, which has been shown to play a crucial role in determining the selectivity of dioxirane-mediated oxidations.[44] We also found that the C1 position in the taxol derivatives was difficult to identify for the model (Fig. 2d). This is likely because our model does not differentiate between hydrogen isotopes. It was shown that the deuteration of the C2 position was crucial to mitigate its oxidation and obtain C1 oxidation as a major product.[45] Even though silyl-ethers and alkenes are both absent from our training set, the reactivity predicted at the C13 position seems reasonable as it has been observed by Baran and co-workers in similar substrates.[45]

We hypothesized that difficulty in predicting the correct site selectivity for some of the complex targets comes from an under-representation of the C–H bonds of the complex molecules in our training set. Whereas the usual solution to this issue would be to augment the size of the dataset, conducting an extensive number of reactions to cover the whole complex molecule space is cost-prohibitive and has little guarantee of being efficient. Thus, the development of an algorithmic approach for selecting the most informative dataset for each individual target is highly desirable.

Acquisition Functions for Target-Specific Dataset Design

Based on these considerations, we set out to increase the accuracy of the predictions of specific targets with a minimal number of experiments. Central to this process is the use of an AF to leverage the model information to select the most insightful examples to generate a tailored dataset for each target. In this workflow, the model-suggested experiments are executed, or selected from a pool of literature data, and each additional datapoint is used to refine the model's predictions. If needed, another round of experiments is suggested, and the cycle can be iterated. Applying this workflow to the C(sp^3)–H oxidation regioselectivity task, AFs are used to score new candidate molecules for their ability to improve the model performance for a specific complex target of interest. Then, the data for the best-scored molecules are added to the training set. Similarly to the full model, the candidate molecules to be added to the dataset contain less than 15 carbons and are commercially available or available in a few steps from commercial materials, whereas the target molecules contain more than 15 carbons.

In a total synthesis context, an expert-designed molecular model substrate is typically chosen by balancing synthetic accessibility and estimated proximity to the target molecule. In a sense, the candidate molecules are analogous to the use of "model substrates" to test the feasibility of a late-stage reaction in complex molecule synthesis. In analogy to expert-designed model substrates, the first AFs were designed using scaffold similarity (AF_{SC} in green, Fig. 3a). The scaffold similarity was quantified as the number of shared atoms in the largest common substructure of the target molecule and the candidates (see SI section V.1 for details). Beyond the ability to select a training set based on trends that may not be intuitive, an added value of this approach is that an AF aggregates information from several "model substrates". This rich information reduces the risk associated with transferring an observation from a single reaction to a more complex target.

Recognizing that the model uses C–H site features as input, AFs were also designed based on site similarity (AF_{CH}, in purple, Fig. 3a). Similarity was computed for all target-candidate C–H pairs as the Euclidean distance between their feature vectors. Then, to aggregate these C–H scores into a molecular score, each candidate site was labeled with its best similarity to a target site, and the maximum of these labels was used as a score for the candidate molecule.

Lastly, we recognized the potential advantage of leveraging the knowledge gained by the model to improve the selection of new training molecules, a process known as active learning. The value of this strategy for chemistry tasks has been demonstrated for modeling computational data[46] and predicting reaction conditions.[1,47] Given that we want to design the smallest, most informative datasets, we anticipated that including model insights would reduce redundancy in molecule selection. Thus, a third AF was designed that integrates the predictions of the model and its uncertainty (evaluated through an ensemble of models). This AF is referred to as the active learning strategy (AF_{AL}, in blue,

Fig. 3a). Because the reactivity and uncertainty were computed per C–H site, this AF is based on site similarity, with scores weighted by the product of predicted reactivity and model uncertainty for each site. Consequently, the selection is biased toward molecules that reduce model uncertainty, while focusing on improving the model's accuracy at the reactive centers of the target. This approach also avoids sampling molecules that would primarily provide information on unreactive sites (a detailed description of the AFs investigated can be found in the sections V.3 of the SI).

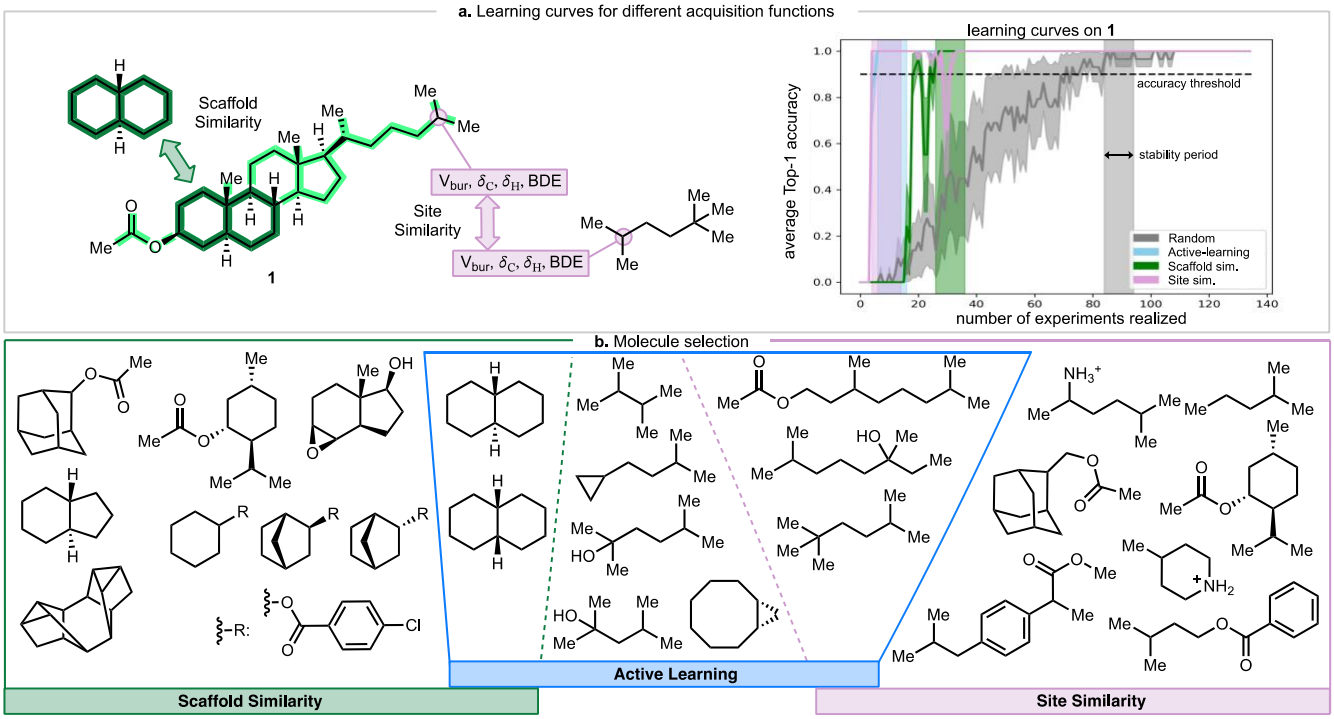

Figure 3. a. Learning curves for 5α-cholestane-3β-ol acetate (**1**) regioselectivity predictions for random selection, AF_{AL}, AF_{SC}, and AF_{CH}. Learning curves are averaged over 30 runs, bold lines represent mean values, and filled areas are two standard deviations. The horizontal dashed line is the accuracy threshold set to determine sampling success. The highlighted vertical colored zone corresponds to the stability zones (colored by AF). b. First 10 molecules selected for **1** by AF_{AL}, AF_{SC}, and AF_{CH}; for the latter we report the result for one run but noticed some variability (19 different molecules were selected for the first 10 over 10 runs).

To evaluate AFs, the sampling library was defined as the 135 molecules that contain less than 15 carbons. To initiate

sampling for each target, the closest molecule with respect to scaffold similarity was selected and each additional training molecule was added one-by-one using an AF. The performance of each AF was measured by the number of experiments required to have a consistent top-1 accuracy during at least 10 iterations (see Fig. 3a).

To better illustrate how the AFs sample the chemical space, the first 10 molecules selected by AF_{AL}, AF_{SC}, and AF_{CH} are shown for the example of cholestanoid **1** (Fig. 3b). Accurate site selectivity prediction is reached in less than 10 data points using the AF_{AL} and AF_{CH}. In contrast, the AF_{SC} and random selection need close to 30 and 80 data points, respectively, to achieve a consistent top-1 accuracy. In the case of the AF_{CH}, for the first 10 selected molecules, all but one contain either a cyclohexane or an isopropyl moiety. Since oxidation occurs at the C25 isopropyl position, the regioselectivity is predicted accurately and requires a very small dataset. In contrast, the AF_{SC} focuses on polycyclic molecules; in this case, only one of the suggested compounds contains an isopropyl group, and the model takes longer to reach the top-1 accuracy. Interestingly, the AF_{AL} strikes a middle ground, selecting compounds that are in the AF_{SC} (e.g. the two decalin compounds) and the AF_{CH} selections. This example showcases how target-specific dataset design can reduce the number of experiments needed to achieve regioselectivity.

To evaluate the generality of the AFs, we averaged the difference in performances of the AF relative to the random selection over the subset of

complex molecules that were predicted accurately by either the random selection or the AF considered (Fig. 4a). The AF_{AL}, AF_{SC}, and AF_{CH} spare 50, 51, and 40 data points, respectively, on average compared to random selection. In the case where the AF does not provide improvement above the random selection, it was typically because random selection afforded an accurate prediction with less than 20 data points (see Fig. 4b, active learning vs random). In other words, the largest gains using the AF strategy were realized on targets that were most difficult to predict. Moreover, we see that 27 to 31 targets are predicted accurately using the different AFs, whereas random selection only predicts 19 correctly (representing up to 24% increase in top-1 accuracy on the complex targets). This further suggests that a small but intentionally designed dataset can give better performances than larger ones for this type of task.[48]

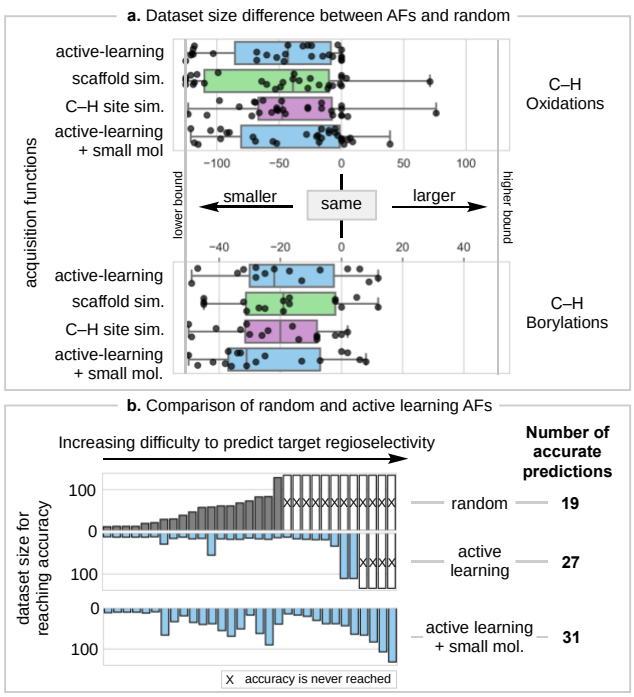

Figure 4. A. Box plots of the performances of four AFs against the random selection. Each point corresponds to a target that was predicted accurately by at least one of the AFs or the random selection. Box plots are displayed for the C–H oxidation and C–H borylation datasets. B. Histograms of the number of datapoints needed per target to reach accuracy using the random, active-learning and active-learning biased toward small-molecule selections on the C–H oxidation dataset. The number of targets accurately predicted for each of these AFs is highlighted, the total number of targets is 50. Exhaustive box plots and histograms for all AFs and both datasets are detailed in SI section V.4.

Experimental validation

Encouraged by the results of this target-specific approach on molecules mined from literature, we sought to test whether we could see similar gains on complex molecules outside of the dataset. To do this, molecules were sourced from our in-house stockroom and an archival library of compounds generated in past projects from the Reisman group. Compounds were then subjected to oxidation by TFDO without significant reaction optimization and resulting isolated yields were used to evaluate prediction accuracy and AF performance.

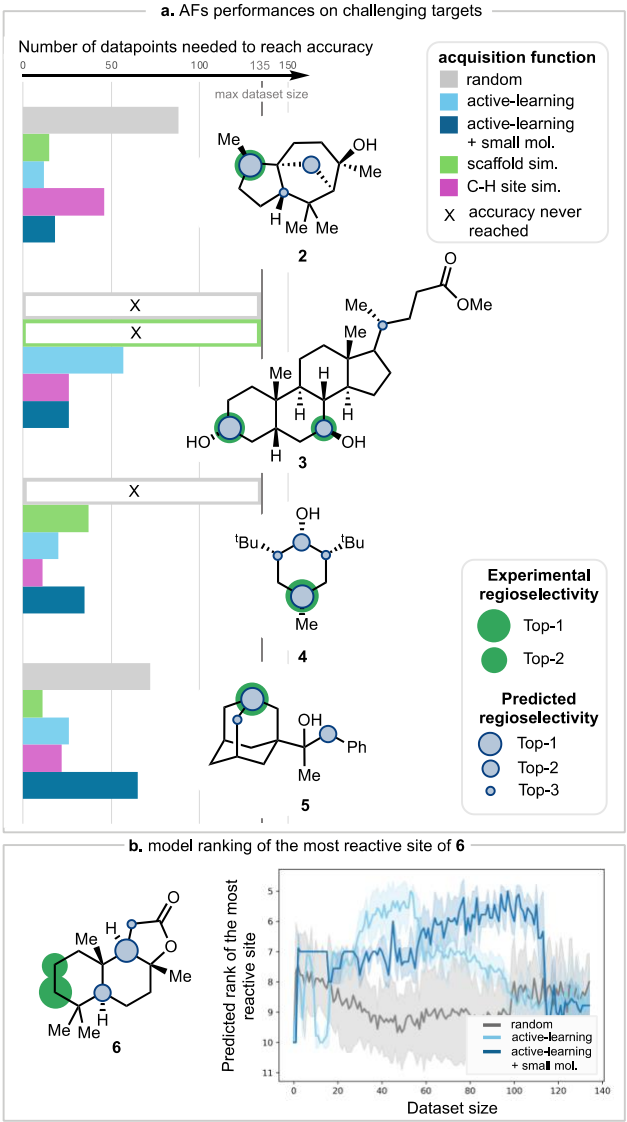

Figure 5. a. Bar plots of the performances of four AFs against random selection on the correctly predicted experimental targets. b. Learning curve depicting how the model's ranking of the most reactive site for (+)-sclareolide evolves with dataset size.

Target molecules were selected that reflected the synthetic interests of the Reisman group and were anticipated to challenge the model to choose between similarly reactive sites (see details SI section VII). These targets included terpenes cedrol (**2**) and (+)-sclareolide (**6**), steroid **3** – which provides interesting competition between carbinol protons – and sterically hindered alcohol **4**, which forces the model to choose between a

tertiary site and a hindered carbinol proton. Although we were interested in validating the model on archival total synthesis substrates, our efforts were restricted to a small set of published molecules without olefins or nitrogen atoms and several candidate compounds gave difficult-to-characterize reaction mixtures (full details in SI section VII). Nonetheless, we were able to characterize the C–H oxidation product of adamantane **5**, a product of a synthetic methods project from 2018,[49] which also required the model to prioritize between tertiary and benzylic positions.

Gratifyingly, we observe that on four of the five targets, the model scores the reactive sites correctly, and the AFs provide stable, accurate predictions at smaller dataset sizes than random selection. For molecules **2**, **3**, **4**, and **5** respectively, the active learning-based AF beats random selection by 76, 78, 115, and 46 datapoints (see Fig. 5a). Molecules **2**, **4**, and **5** can achieve stable accuracy within a dozen datapoints depending on choice of AF. The improvement is especially pronounced for targets **3** and **4**, where stable accuracy cannot be achieved under random selection in 135 datapoints. The model struggles with (+)-sclareolide, perhaps weighting tertiary positions over electronic features (see Fig. 5b). It also appears that longer range interactions, such as the deactivation of the top-ranked tertiary site by the lactone, are not picked up by the selected descriptors. However, even on this difficult-to-predict substrate, the AFs still provide improvement over random – the rank of the most experimentally reactive C–H site is consistently better with the active learning-based AFs than with random selection.

In conclusion, on this validation set, target-specific dataset selection reduces by more than 50% the size of the dataset needed to reach accuracy and increases the accuracy from 2 out of 5 with the random selection to 4 out of 5 using AF_{AL} or AF_{CH}.

External workflow validation

To probe the generality of the target-specific dataset design strategy for regioselectivity predictions, the workflow was repeated on another reaction of interest for late-stage functionalization, the C(sp^2)–H radical borylation. The workflow was applied to a subset of a recently reported borylation reaction dataset[11] (82 reactions including 22 large targets), filtered to include only reactions conducted under the same conditions (detailed in SI section VI). When the learning curves of the models trained on AF-selected molecules were compared to using random selection, significant improvements were observed over random exploration. Interestingly, the AF_{AL}, AF_{SC}, and AF_{CH} were all competitive, which is consistent with what was observed for C(sp^3)–H oxidation (see Fig. 4A). Specifically, in a search space of 60 reactions, AF_{AL}, AF_{SC}, and AF_{CH} spare 18, 16, and 21 data points on average, respectively, compared to random. Additionally, molecules that could not be predicted with top-1 accuracy using the random selection were predicted accurately with the AF strategies (12 were predicted accurately with random selection versus 15, 16, and 16 with AF_{AL}, AF_{SC}, and AF_{CH} respectively – an increase of 14 to 18% in top-1 accuracy).

Choosing the best Acquisition Function:

To evaluate whether there was a single AF that performed best across all target molecules, the improvement over baseline was compared for the AF_{AL}, AF_{SC}, and AF_{CH} on all large targets in the C–H oxidation (Fig. 4a) and borylation datasets. The performance in terms of dataset size improvement and number of targets accurately predicted were comparable for all three AFs, so there is likely flexibility in which AF will perform best in a given context. However, as illustrated Fig. 3B, both the AF_{SC} and AF_{CH} samplings introduce molecular redundancies. The lack of diversity in the initial suggestions could affect model performances and increase the size of the dataset required to make accurate predictions. We note that using clustering to introduce diversity in the sampling for the AF_{CH} and AF_{SC} did not improve the performances (see SI section V.4). Instead, the AF_{AL}s are intrinsically designed to search for a diversity of compounds in reasonable proximity to the reactive sites. With respect to using an AF_{AL} to sample from a large chemical vendor catalog, we anticipated that biasing sampling towards smaller molecules to limit the difficulties in the characterization of the site of oxidation would be advantageous. This strategy not only led to the selection of smaller molecules but also led to the accurate prediction of more complex molecules 31 against 27 for AF_{AL} on C–H oxidation and 17 against 15 for AF_{AL} on C–H borylation (see AL small mol in Fig. 4a and table S1 and S2 for exhaustive AFs comparison). Therefore, when little to no literature data is available for an initial dataset, we suggest that the AF_{AL} biased toward small molecules is likely to yield the best results for experimental application.

Conclusion

A reaction-agnostic acquisition-function based strategy for target-specific dataset design is reported. The workflow and dataset are publicly available (https://github.com/ReismanLab/regio_dataset_design), enabling further development on the design of descriptors and AFs. The approach presented is efficient in reducing the size of the datasets needed to predict the regioselectivity of complex molecules. Two datasets of reactions: C(sp^3)–H dioxirane oxidation and C(sp^2)–H borylation were used for validation and showed that models trained on datasets designed by the best AFs

needed, respectively, only 30% and 55%, of the data required when trained on randomly selected datapoints. Furthermore, this work demonstrates that acquisition-function designed datasets can provide better accuracy than larger, randomly acquired datasets; an improvement of 24% and 23% is reported for the two datasets respectively.[48] Finally, an experimental validation on a set of five complex targets was performed and confirmed the trends observed on the literature data.

Acknowledgments

A.C.C. acknowledges the Swedish Research Council for a postdoc fellowship (Vetenskapsrådet, VR-2022-06175). A.G. acknowledges support for this work by the National Science Foundation Graduate Research Fellowship under Grant No. (NSF 2139433) and the Hertz Foundation Fellowship. A.M. wishes to thank the Israel Science Foundation for their generous support (grant no. 2252/21) and the Kreitman School of Advanced Graduate Studies for supporting Y.H. with the Chemotech fellowship. G.L. and A.P. acknowledge Caltech for a SURF fellowship. S.E.R. acknowledges the NSF under the CCI Center for Computer-Assisted Synthesis (CHE-2202693) for financial support.

References

(1) Rinehart, N. I.; Saunthwal, R. K.; Wellauer, J.; Zahrt, A. F.; Schlemper, L.; Shved, A. S.; Bigler, R.; Fantasia, S.; Denmark, S. E. A Machine-Learning Tool to Predict Substrate-Adaptive Conditions for Pd-Catalyzed C–N Couplings. *Science* **2023**, *381* (6661), 965–972. https://doi.org/10.1126/science.adg2114.

(2) Wang, J. Y.; Stevens, J. M.; Kariofillis, S. K.; Tom, M.-J.; Golden, D. L.; Li, J.; Tabora, J. E.; Parasram, M.; Shields, B. J.; Primer, D. N.; Hao, B.; Del Valle, D.; DiSomma, S.; Furman, A.; Zipp, G. G.; Melnikov, S.; Paulson, J.; Doyle, A. G. Identifying General Reaction Conditions by Bandit Optimization. *Nature* **2024**, *626* (8001), 1025–1033. https://doi.org/10.1038/s41586-024-07021-y.

(3) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning. *Science* **2018**, *360* (6385), 186–190. https://doi.org/10.1126/science.aar5169.

(4) Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; Zurański, A. M.; Kogej, T.; Norrby, P.-O.; Doyle, A. G.; Chawla, N. V.; Wiest, O. On the Use of Real-World Datasets for Reaction Yield Prediction. *Chem. Sci.* **2023**, *14* (19), 4997–5005. https://doi.org/10.1039/D2SC06041H.

(5) Schleinitz, J.; Langevin, M.; Smail, Y.; Wehnert, B.; Grimaud, L.; Vuilleumier, R. Machine Learning Yield Prediction from NiCOlit, a Small-Size Literature Data Set of Nickel Catalyzed C–O Couplings. *J. Am. Chem. Soc.* **2022**, *144* (32), 14722–14730. https://doi.org/10.1021/jacs.2c05302.

(6) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields Using Deep Learning. *Mach. Learn. Sci. Technol.* **2021**, *2* (1), 015016. https://doi.org/10.1088/2632-2153/abc81d.

(7) Żurański, A. M.; Martinez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **2021**, *54* (8), 1856–1865. https://doi.org/10.1021/acs.accounts.0c00770.

(8) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363* (6424). https://doi.org/10.1126/science.aau5631.

(9) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571* (7765), 343–348. https://doi.org/10.1038/s41586-019-1384-z.

(10) Guillemard, L.; Kaplaneris, N.; Ackermann, L.; Johansson, M. J. Late-Stage C–H Functionalization Offers New Opportunities in Drug Discovery. *Nat. Rev. Chem.* **2021**, *5* (8), 522–545. https://doi.org/10.1038/s41570-021-00300-6.

(11) Nippa, D. F.; Atz, K.; Hohler, R.; Müller, A. T.; Marx, A.; Bartelmus, C.; Wuitschik, G.; Marzuoli, I.; Jost, V.; Wolfard, J.; Binder, M.; Stepan, A. F.; Konrad, D. B.; Grether, U.; Martin, R. E.; Schneider, G. Enabling Late-Stage Drug Diversification by High-Throughput Experimentation with Geometric Deep Learning. *Nat. Chem.* **2024**, *16* (2), 239–248. https://doi.org/10.1038/s41557-023-01360-5.

(12) Ree, N.; Göller, A. H.; Jensen, J. H. RegioML: Predicting the Regioselectivity of Electrophilic Aromatic Substitution Reactions Using Machine Learning. *Digit. Discov.* **2022**, *1* (2), 108–114. https://doi.org/10.1039/D1DD00032B.

(13) Tomberg, A.; Johansson, M. J.; Norrby, P.-O. A Predictive Tool for Electrophilic Aromatic Substitutions Using Machine Learning. *J. Org. Chem.* **2019**, *84* (8), 4695–4703. https://doi.org/10.1021/acs.joc.8b02270.

(14) Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Sköld, C.; Norrby, P.-O.; Hartwig, J. F. Hybrid Machine Learning Approach to Predict the Site Selectivity of Iridium-Catalyzed Arene Borylation. *J. Am. Chem. Soc.* **2023**, *145* (31), 17367–17376. https://doi.org/10.1021/jacs.3c04986.

(15) Guan, Y.; Lee, T.; Wang, K.; Yu, S.; McWilliams, J. C. SNAr Regioselectivity Predictions: Machine Learning Triggering DFT Reaction Modeling through Statistical Threshold. *J. Chem. Inf. Model.* **2023**, *63* (12), 3751–3760. https://doi.org/10.1021/acs.jcim.3c00580.

(16) King-Smith, E.; Faber, F. A.; Reilly, U.; Sinitskiy, A. V.; Yang, Q.; Liu, B.; Hyek, D.; Lee, A. A. Predictive Minisci Late Stage Functionalization with Transfer Learning. *Nat. Commun.* **2024**, *15* (1), 426. https://doi.org/10.1038/s41467-023-42145-1.

(17) Li, X.; Zhang, S.-Q.; Xu, L.-C.; Hong, X. Predicting Regioselectivity in Radical C−H Functionalization of Heterocycles through Machine Learning. *Angew. Chem. Int. Ed.* **2020**, *59* (32), 13253–13259. https://doi.org/10.1002/anie.202000959.

(18) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572–1583. https://doi.org/10.1021/acscentsci.9b00576.

(19) Chen, M. S.; White, M. C. A Predictably Selective Aliphatic C–H Oxidation Reaction for Complex Molecule Synthesis. *Science* **2007**, *318* (5851), 783–787. https://doi.org/10.1126/science.1148597.

(20) Boni, Y. T.; Cammarota, R. C.; Liao, K.; Sigman, M. S.; Davies, H. M. L. Leveraging Regio- and Stereoselective C(Sp3)–H Functionalization of Silyl Ethers to Train a Logistic Regression Classification Model for Predicting Site-Selectivity Bias. *J. Am. Chem. Soc.* **2022**, *144* (34), 15549–15561. https://doi.org/10.1021/jacs.2c04383.

(21) Mennen, S. M.; Alhambra, C.; Allen, C. L.; Barberis, M.; Berritt, S.; Brandt, T. A.; Campbell, A. D.; Castañón, J.; Cherney, A. H.; Christensen, M.; Damon, D. B.; Eugenio de Diego, J.; García-Cerrada, S.; García-Losada, P.; Haro, R.; Janey, J.; Leitch, D. C.; Li, L.; Liu, F.; Lobben, P. C.; MacMillan, D. W. C.; Magano, J.; McInturff, E.; Monfette, S.; Post, R. J.; Schultz, D.; Sitter, B. J.; Stevens, J. M.; Strambeanu, I. I.; Twilton, J.; Wang, K.; Zajac, M. A. The Evolution of High-Throughput Experimentation in Pharmaceutical Development and Perspectives on the Future. *Org. Process Res. Dev.* **2019**, *23* (6), 1213–1242. https://doi.org/10.1021/acs.oprd.9b00140.

(22) McDonald, M. A.; Koscher, B. A.; Canty, R. B.; Jensen, K. F. Calibration-Free Reaction Yield Quantification by HPLC with a Machine-Learning Model of Extinction Coefficients. *Chem. Sci.* **2024**, *15* (26), 10092–10100. https://doi.org/10.1039/D4SC01881H.

(23) Venetos, M. C.; Elkin, M.; Delaney, C.; Hartwig, J. F.; Persson, K. A. Deconvolution and Analysis of the 1H NMR Spectra of Crude Reaction Mixtures. *J. Chem. Inf. Model.* **2024**, *64* (8), 3008–3020. https://doi.org/10.1021/acs.jcim.3c01864.

(24) Rana, D.; Pflüger, P. M.; Hölter, N. P.; Tan, G.; Glorius, F. Standardizing Substrate Selection: A Strategy toward Unbiased Evaluation of Reaction Generality. *ACS Cent. Sci.* **2024**, *10* (4), 899–906. https://doi.org/10.1021/acscentsci.3c01638.

(25) Dreher, S. D.; Krska, S. W. Chemistry Informer Libraries: Conception, Early Experience, and Role in the Future of Cheminformatics. *Acc. Chem. Res.* **2021**, *54* (7), 1586–1596. https://doi.org/10.1021/acs.accounts.0c00760.

(26) Asensio, G.; Castellano, G.; Mello, R.; González Núñez, M. E. Oxyfunctionalization of Aliphatic Esters by Methyl(Trifluoromethyl)Dioxirane. *J. Org. Chem.* **1996**, *61* (16), 5564–5566. https://doi.org/10.1021/jo9604189.

(27) González-Nuñez, M. E.; Royo, J.; Castellano, G.; Andreu, C.; Boix, C.; Mello, R.; Asensio, G. Hyperconjugative Control by Remote Substituents of Diastereoselectivity in the Oxygenation of Hydrocarbons. *Org. Lett.* **2000**, *2* (6), 831–834. https://doi.org/10.1021/ol000017m.

(28) Bovicelli, P.; Gambacorta, A.; Lupattelli, P.; Mincione, E. A Highly Regio- and Stereoselective C5 Oxyfunctionalization of Coprostane Steroids by Dioxiranes: An Improved Access to Progestogen and Androgen Hormones. *Tetrahedron Lett.* **1992**, *33* (48), 7411–7412. https://doi.org/10.1016/S0040-4039(00)60202-2.

(29) Adam, W.; Zhao, C.-G.; Jakka, K. Dioxirane Oxidations of Compounds Other than Alkenes. In *Organic Reactions*; John Wiley & Sons, Ltd, 2008; pp 1–346. https://doi.org/10.1002/0471264180.or069.01.

(30) Bovicelli, P.; Lupattelli, P.; Mincione, E.; Prencipe, T.; Curci, R. Oxidation of Natural Targets by Dioxiranes. 2. Direct Hydroxylation at the Side Chain C-25 of Cholestane Derivatives and of Vitamin D3 Windaus-Grundmann Ketone. *J. Org. Chem.* **1992**, *57* (19), 5052–5054. https://doi.org/10.1021/jo00045a004.

(31) Crandall, J. K.; Curci, R.; D'Accolti, L.; Fusco, C.; Fusco, C.; D'Accolti, L.; Annese, C. Methyl(Trifluoromethyl)Dioxirane. In *Encyclopedia of Reagents for Organic Synthesis*; John Wiley & Sons, Ltd, 2016; pp 1–11. https://doi.org/10.1002/047084289X.rm267.pub3.

(32) Fusco, C.; Fiorentino, M.; Dinoi, A.; Curci, R.; Krause, R. A.; Kuck, D. Oxyfunctionalization of Non-Natural Targets by Dioxiranes. 2. Selective Bridgehead Dihydroxylation of Fenestrindane1. *J. Org. Chem.* **1996**, *61* (24), 8681–8684. https://doi.org/10.1021/jo961316l.

(33) Mello, R.; Fiorentino, M.; Fusco, C.; Curci, R. Oxidations by Methyl(Trifluoromethyl)Dioxirane. 2. Oxyfunctionalization of Saturated Hydrocarbons. *J. Am. Chem. Soc.* **1989**, *111* (17), 6749–6757. https://doi.org/10.1021/ja00199a039.

(34) Mello, R.; Cassidei, L.; Fiorentino, M.; Fusco, C.; Curci, R. Oxidations by Methyl(Trifluoromethyl)Dioxirane. 3. Selective Polyoxyfunctionalization of Adamantane. *Tetrahedron Lett.* **1990**, *31* (21), 3067–3070. https://doi.org/10.1016/S0040-4039(00)89027-9.

(35) D'Accolti, L.; Annese, C.; Fusco, C. Continued Progress towards Efficient Functionalization of Natural and Non-Natural Targets under Mild Conditions: Oxygenation by C–H Bond Activation with Dioxirane. *Chem. – Eur. J.* **2019**, *25* (52), 12003–12017. https://doi.org/10.1002/chem.201901687.

(36) Oritani, T.; Horiguchi, T.; Nagura, M.; Cheng, Q.; Kudo, T. Chemical Oxidation of Taxoids with M-CPBA and Dimethyl Dioxirane: Regioselective Epoxidation of Taxinine J Derivatives. *HETEROCYCLES* **2000**, *53* (12), 2629. https://doi.org/10.3987/COM-00-8929.

(37) Kovač, F.; Baumstark, A. L. Oxidation of α-Methylbenzyl Alcohols by Dimethyldioxirane. *Tetrahedron Lett.* **1994**, *35* (47), 8751–8754. https://doi.org/10.1016/S0040-4039(00)78488-7.

(38) Lesieur, M.; Battilocchio, C.; Labes, R.; Jacq, J.; Genicot, C.; Ley, S. V.; Pasau, P. Direct Oxidation of Csp3–H Bonds Using in Situ Generated Trifluoromethylated Dioxirane in Flow. *Chem. – Eur. J.* **2019**, *25* (5), 1203–1207. https://doi.org/10.1002/chem.201805657.

(39) Saladino, R.; Mezzetti, M.; Mincione, E.; Torrini, I.; Paradisi, M. P.; Mastropietro, G. A New and Efficient Synthesis of Unnatural Amino Acids and Peptides by Selective 3,3-Dimethyldioxirane Side-Chain Oxidation. *J. Org. Chem.* **1999**, *64* (23), 8468–8474. https://doi.org/10.1021/jo990185w.

(40) Shustov, G. V.; Rauk, A. Mechanism of Dioxirane Oxidation of CH Bonds: Application to Homo- and Heterosubstituted Alkanes as a Model of the Oxidation of Peptides. *J. Org. Chem.* **1998**, *63* (16), 5413–5422. https://doi.org/10.1021/jo9802877.

(41) *Dioxiranes: a half-century journey - Organic Chemistry Frontiers (RSC Publishing)*. https://pubs.rsc.org/en/content/articlelanding/2022/qo/d2qo01005d (accessed 2024-06-03).

(42) Curci, R.; D'Accolti, L.; Fusco, C. A Novel Approach to the Efficient Oxygenation of Hydrocarbons under Mild Conditions. Superior Oxo Transfer Selectivity Using Dioxiranes. *Acc. Chem. Res.* **2006**, *39* (1), 1–9. https://doi.org/10.1021/ar050163y.

(43) Raghavan, P.; Haas, B. C.; Ruos, M. E.; Schleinitz, J.; Doyle, A. G.; Reisman, S. E.; Sigman, M. S.; Coley, C. W. Dataset Design for Building Models of Chemical Reactivity. *ACS Cent. Sci.* **2023**, *9* (12), 2196–2204. https://doi.org/10.1021/acscentsci.3c01163.

(44) Zou, L.; Paton, R. S.; Eschenmoser, A.; Newhouse, T. R.; Baran, P. S.; Houk, K. N. Enhanced Reactivity in Dioxirane C–H Oxidations via Strain Release: A Computational and Experimental Study. *J. Org. Chem.* **2013**, *78* (8), 4037–4048. https://doi.org/10.1021/jo400350v.

(45) Kanda, Y.; Nakamura, H.; Umemiya, S.; Puthukanoori, R. K.; Murthy Appala, V. R.; Gaddamanugu, G. K.; Paraselli, B. R.; Baran, P. S. Two-Phase Synthesis of Taxol. *J. Am. Chem. Soc.* **2020**, *142* (23), 10526–10533. https://doi.org/10.1021/jacs.0c03592.

(46) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less Is More: Sampling Chemical Space with Active Learning. *J. Chem. Phys.* **2018**, *148* (24), 241733. https://doi.org/10.1063/1.5023802.

(47) Shim, E.; Kammeraad, J. A.; Xu, Z.; Tewari, A.; Cernak, T.; Zimmerman, P. M. Predicting Reaction Conditions from Limited Data through Active Transfer Learning. *Chem. Sci.* **2022**, *13* (22), 6655–6668. https://doi.org/10.1039/D1SC06932B.

(48) Shalit Peleg, H.; Milo, A. Small Data Can Play a Big Role in Chemical Discovery. *Angew. Chem. Int. Ed.* **2023**, *62* (26), e202219070. https://doi.org/10.1002/anie.202219070.

(49) Su, J. Y.; Grünenfelder, D. C.; Takeuchi, K.; Reisman, S. E. Radical Deoxychlorination of Cesium Oxalates for the Synthesis of Alkyl Chlorides. *Org. Lett.* **2018**, *20* (16), 4912–4916. https://doi.org/10.1021/acs.orglett.8b02045.