Bigger is not always better: evaluating target-specific dataset design strategies for regioselectivity prediction on complex molecules

Anjali Gurajapu
California Institute of Technology
Pasadena, CA 91125
agurajap@caltech.edu

Jules Schleinitz
California Institute of Technology
Pasadena, CA 91125
jul@caltech.edu

Alba Carretero Cerdán
KTH Royal Institute of Technology
Stockholm, Sweden
albacc@kth.se

Yonatan Harnik
Ben-Gurion University
Be'er Sheva, Israel
harniky@post.bgu.ac.il

Carolyn Ruan
California Institute of Technology
Pasadena, CA 91125
cruan@caltech.edu

Carolyn Ruan
California Institute of Technology
Pasadena, CA 91125
cruan@caltech.edu

Gina Lee
California Institute of Technology
Pasadena, CA 91125
glee3@caltech.edu

Amitesh Pandey
California Institute of Technology
Pasadena, CA 91125
amitesh@caltech.edu

Anat Milo
Ben-Gurion University
Be'er Sheva, Israel
anatmilo@bgu.ac.il

Sarah E. Reisman
California Institute of Technology
Pasadena, CA 91125
reisman@caltech.edu

Abstract

There has been growing interest in using ML models for prediction of reaction yields and selectivity in synthetic chemistry. However, the difficulty and cost of generating experimental data has proved a roadblock in creating practical models for these tasks. For this reason, rational dataset design strategies are emerging in the field, typically limited to clustering approaches to sample the overall chemical space broadly. However, in many real-world contexts like synthetic route planning, the chemist is often narrowly interested in accurate predictions on a specific, known target. As such, we propose a contrasting dataset design strategy that exploits knowledge of the target to create small models focused on local regions of chemical space. We design a series of acquisition functions that consider model uncertainty, several metrics of chemical similarity, and varying degrees of dataset diversity. We find that an active learning strategy that selects training molecules similar to uncertain regions of the target outperforms approaches that consider target similarity alone. Target-focused data sets significantly reduced data requirements; in fact, these smaller datasets could achieve accuracy on targets where larger, diversity-oriented or randomly selected data sets failed. Evaluation

Preprint.

was performed on two literature datasets of C–H functionalization reactions, along with experimental validation on five complex targets. To conclude, we discuss our ongoing work in developing a stopping criterion for the active learning loop to enable a full experimental implementation of this workflow.

1 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. There is great interest in the development of ML models that predict the regioselectivity of direct C–H functionalization reactions, which are controlled by the innate reactivity of the substrate and/or reagent rather than by a directing group. These predictive models can derisk direct C–H activation in the late stage of a multistep synthesis campaign, aid synthetic planning,Guillemard et al. [2021] and provide rationale for late-stage diversification efforts. Notably, for these tasks, there is a proposed, difficult-to-access complex target of interest for the chemist. This target has been designed, but is only worth a lengthy synthetic campaign if the subsequent reaction proceeds as desired. In this common scenario, conducting a few experiments on simple substrates to train an accurate, target-specific model may cost less time and money than synthesizing the complex target only to discover that it does not undergo the desired transformation.

A major obstacle in predicting regioselectivity is the development of a dataset to support the task, especially on complex substrates. Generating experimental data for direct C–H activation involves elucidation of potentially multiple products per substrate, as well as quantification of these products. High-throughput experimentation (HTE)Mennen et al. [2019] is ill-suited to generate data that explores substrate space due to the large upfront time investment required to obtain calibrated yields for many unique reaction products, even in contexts where the expected product is known.Wang et al. [2024], McDonald et al. [2024] Thus, the purification, characterization, and assignment of the site of C-H activation on a complex molecule often becomes the rate-limiting step in dataset generation.

Additionally, a recent report shows that, even with an HTE dataset, regioselectivity prediction models may fail when molecules of interest are far from the training set distribution.Nippa et al. [2023] Despite significant progress in designing substrate scopes to assess the domain of applicability of new methods,Rana et al. [2024], Dreher and Krska [2021] 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. To overcome the experimental constraints that limit dataset size and avoid inaccurate extrapolation from the training set, **we propose a dataset acquisition approach that focuses on target-specific dataset generation, based on both chemical similarity and model uncertainty.**

1.1 Related Works

The notion of assessing the local chemical space where a model can predict accurately has previously appeared in the regioselectivity prediction literature. Guan et al. [2023] develop a prediction tool that triggers a domain-independent, time-intensive quantum mechanical calculation, when the query molecule is outside the domain of a machine learning model (with domain assessed by model confidence). Caldeweyher et al. [2023] develop a model that dynamically mixes a partial least squares (PLS) prediction and a more extrapolative neighboring substituent penalty, de-prioritizing the PLS prediction when the query molecule has low Tanimoto similarity to the training set.

Developing training datasets based on target similarity has appeared in other chemistry contexts, namely for quantum chemistry,Lemm et al. [2023] property prediction, Kim et al. [2024], and gas content evaluation.Yu et al. [2021] The latter two reports cluster the available training data using k-means clustering and develop models specific to each cluster. The former work, given a target query, suggests the nearest N neighbors as training molecules. Since the initial publication of this work, Reid and coworkers have published a radius-based random forest regression algorithm for the synthetic chemistry context that uses a similarity threshold to assess whether a training point should be included when building a model for a given target.Betinol et al. [2025] Our approach is the first that we know of for synthetic chemistry, and moreover introduces an active-learning component to the dataset design, incorporating model uncertainty on the target as a consideration along with target similarity.

1.2 Contributions

We report similarity- and uncertainty-aware dataset design methods to efficiently train ML models to predict the regioselectivity of innate C–H functionalization reactions on complex targets. Our contributions are as follows:

- We design a suite of acquisition functions (AFs) for target-oriented training set selection that consider molecule- and atom-level similarity, model uncertainty, model predictions, and overall training set diversity when selecting the most informative training points.
- We develop models to predict the regioselectivity of dioxirane C(sp^3)–H oxidation, a reaction employed in complex molecule synthesisKanda et al. [2020] that has not previously been modeled.
- We evaluate the acquisition workflow on two literature reaction datasets, finding that the AFs which performed best considered both the model uncertainty on the target as well as atom-level similarity.
- We show that the AFs achieve accuracy at smaller datasets than random selection. On a dataset of 135 reactions, select AFs achieve top-1 accuracy 40-50 datapoints earlier than random selection.
- We demonstrate that small, well-designed datasets can achieve accuracy when large datasets fail. Of the 50 target molecules studied, 12 that could not be predicted accurately using random selection at any dataset size were successfully predicted using a smaller, AF-selected training set.

2 Regioselectivity prediction of C(sp^3)–H Oxidation

2.1 Dataset

As a proof-of-concept, we focused on dioxirane-mediated C–H oxidation reactions, which are controlled by the substrate's innate reactivity. We mined reaction regioselectivity data for dimethyl-dioxirane (DMDO) and trifluoromethyl-dioxirane (TFDO), curating reports providing detailed information about yields and selectivity (complete list in Appendix A.1). After preprocessing the dataset (Appendix A.2), 185 unique reactions remained and were used for further modeling. We noticed that (a) reaction conditions vary little across the dataset (Appendix A.3) and (b) reports showed that TFDO and DMDO maintained the same regioselectivity.Curci et al. [2006] 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.

2.2 Modeling

We framed the regioselectivity modeling task as learning a mapping $f : x - - > y$, where x is a vector representation of a C-H site and y is the experimental selectivity. To build a C-H site representation without extensive computational cost, we leveraged semi-empirical methodsBannwarth et al. [2019] and machine-learned descriptorsS. V. et al. [2023] for the C-H featurization. Specifically, we computed descriptors encoding the steric (Sterimol, buried volume), electronic (C and H charge), and local atomic (hybridization, neighboring atom-types) environments of the C–H bond. In addition, the bond dissociation energy (BDE) was computed. Full details are provided in Appendix A.4. These descriptors were benchmarked against several machine learning models (random forest, K-nearest neighbors, linear regression, support vector regression, Gaussian process regression).

Model performance was evaluated on two tasks: a leave-one-out (LOO) cross-validation and a train-test split on molecule complexity. The latter task was designed to understand how our models performed on the complex targets of interest when trained only 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). In terms of molecular structure, the complex molecule dataset consists of 7 di- and tri-peptides, 3 taxol derivatives, 3 macrocycles, 22 steroids, and 15 miscellaneous compounds (Appendix A.5). To put the modeling results in context, a heuristic 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,

	Random Forest	KNN	Linear Reg	SVR	GPR
BDE	55.46 ± 1.0	55.14	**56.22**	9.19	44.43 ± 0.22
Sterics	**64.76 ± 0.58**	57.84	55.14	57.3	41.24 ± 2.11
Electronics	72.86 ± 0.72	**79.46**	38.38	65.41	60.38 ± 0.8
Local Environment	55.84 ± 0.59	56.76	51.35	47.03	**57.89 ± 0.8**
xTB-Morfeus	**79.08 ± 0.8**	71.89	54.05	58.92	46.16 ± 2.31
Human-selected	**77.95 ± 0.63**	73.51	60.0	64.86	40.0
Model-selected	**<u>79.95 ± 0.89</u>**	73.51	72.97	69.19	43.41 ± 2.82

Table 1: Benchmarking of C-H bond descriptors against model type on LOO cross-validation. **Bold** numbers indicate the best-performing model per descriptor type. **Bold** and <u>underlined</u> indicates the best-performing model overall. "Human-selected" are a set of descriptors chosen using chemical intuition, and "model-selected" are a set of descriptors chosen using RF feature importance.

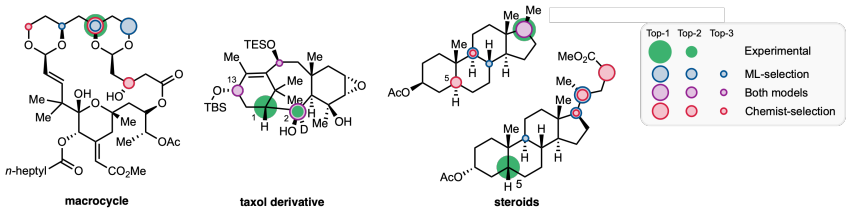


Figure 1: Regioselectivity predictions on selected complex targets of RF trained on small molecules.

and primary (Baseline LOO: 38%, Complex-split: 12%). Top-1 accuracy was used as the evaluation metric.

	Random Forest	KNN	Linear Reg	SVR	GPR
BDE	**22.0**	**22.0**	**22.0**	**22.0**	16.0
Sterics	36.6 ± 2.0	28.0	**40.0**	36.0	1.4 ± 0.9
Electronics	**47.8 ± 1.1**	44.0	8.0	32.0	30.0
Local Environment	24.4 ± 0.8	26.0	20.0	22.0	**28.0**
xTB-Morfeus	**47.6 ± 1.8**	26.0	26.0	28.0	28.0 ± 18.3
Human-selected	**48.6 ± 1.8**	42.0	24.0	24.0	4.0
Model-selected	51.0 ± 2.2	**<u>54.0</u>**	34.0	50.0	25.0 ± 25.0

Table 2: Benchmarking of C-H bond descriptors against model type on simple-complex molecule train-test split. **Bold** numbers indicate the best-performing model per descriptor type. **Bold** and <u>underlined</u> indicates the best-performing model overall. "Human-selected" are a set of descriptors chosen using chemical intuition, and "model-selected" are a set of descriptors chosen using RF feature importance.

Random forest (RF) provided the best balance in performance across both tasks (2, 1). The RF models significantly outperform the rule-based baseline on the LOO task (˜80% top-1 accuracy for the best-performing models versus heuristic baseline 38%) and when predicting on large molecules (˜50% top-1 accuracy versus heuristic baseline 12%).

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%). Error analysis of the different molecules in the complex target dataset reveals the following:

- On the so-called miscellaneous molecules, good performance is achieved (13/15 correct top-1), perhaps due to a large proportion of molecules containing reactive benzylic sites. Good performance is also achieved on the macrocycle class.

- On peptides, the model also performs well, likely due to the small number of more-reactive tertiary C–H sites compared to primary and secondary sites. The main source of error is differentiation between isopropyl groups (4/7 rank the most-reactive site correctly, and the rest predict it as the second-most reactive)
- In the steroid class, 7/11 steroids having a $C_{5\alpha}$–H configuration are predicted correctly, while the reactive site of the 5β-steroids is never ranked higher than top-4. Challenges in distinguishing this reactivity might stem from the featurization failing to adequately capture the stereochemistry on the ring, which has been shown to play a crucial role in determining the selectivity of dioxirane-mediated oxidations.Zou et al. [2013]
- The C1 position in the taxol derivatives was difficult for the model to identify (Fig. 1). 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.Kanda et al. [2020] 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 Kanda et al. [2020] in similar substrates.

Aside from potential weaknesses in the featurization, we hypothesized that poor prediction of some complex targets stems from under-representation of their C–H bonds in the overall training set. Instead of undertaking a prohibitively expensive experimental campaign to balance the C-H site representation in the dataset, we describe the development of an algorithmic approach for selecting the most informative dataset for each individual target.

3 Acquisition Functions for Target-Specific Dataset Design

3.1 Proposed Workflow

The dataset selection algorithm uses an acquisition function to select a tailored dataset for each target one training point at a time. In the proposed workflow, the AF is used to score potential candidate molecules and the best-scored candidate is subjected to the reaction under study. This additional data point is used to refine the predictions further. If needed, the cycle is repeated.

3.2 Design Principles

Acquisition functions to select candidate training molecules were implemented based on five design principles: (1) substructure-level molecular similarity, (2) site-level bond similarity, (3) training set diversity, (4) use of model information, and (5) ease of experimental validation.

Substructure similarity AFs: Substructure similarity was quantified as the number of atoms in the largest shared common substructure of the target molecule and the candidates (maximum common substructure or MCS). AF-2-1 (or AF_{SC}) considers only similarity to target and naively selects the candidate molecule with the largest shared substructure to the target. To incorporate diversity considerations, spectral clustering was performed on substructure-based RDKit fingerprints to divide the molecules into ten clusters. A diversity-only strategy was implemented (AF-3), with no consideration of similarity to target, where the AF would select a molecule randomly from each cluster, alternating between all clusters. Hybrid diversity- and similarity-based AFs were designed (AF-4-1), which select the molecule with the highest substructural similarity score from each cluster, rather than selecting randomly.

Site similarity AFs: C–H site similarity was computed for all target-candidate C–H pairs as the Euclidean distance between their feature vectors. Then, to arrive at a single score per candidate carbon, each candidate site was labeled with its best similarity to a target site. The maximum of these labels was taken as the candidate score (AF-6, or AF_{CH}). To incorporate diversity, k-means clustering was used to divide the C–H sites of the candidate molecule pool into ten clusters. An exclusively diversity-based method to include molecules that represent the full C–H space of the candidate molecules was implemented (AF-8). In the selection process, one molecule from each cluster was sequentially added, prioritizing the most representative molecules in the cluster. The most representative molecule from each cluster was determined to be the molecule possessing the most C–H sites belonging to that cluster. A hybrid diversity- and similarity-based AF was also designed, which aimed to select a pool of molecules that matched all sites of the target (AF-9). In this selection

strategy, each C–H site of the target was analyzed successively and the candidate with the most similar C–H site was selected.

Active learning AFs: Given that we wanted to design the smallest, most informative datasets, we anticipated that including model insights would reduce redundancy in molecule selection. Thus, an additional AF (AF-1 or AF_{AL}) was designed that integrates the predictions of the model and its uncertainty.

The AF score is given as the weighted mean of all target-candidate C–H similarities($\frac{1}{d(j,i)}$, where $d(j,i)$ is the Euclidean distance between the candidate C-H site j and the target C-H site i). An adjustable parameter d_b was added, ensuring the scoring function is always well-defined. The similarity of each C-H site i of the target is weighted by the product of the model uncertainty $\delta(reg,i)$ and the predicted reactivity $r(reg,i)$ on that site. The uncertainty is calculated as the standard deviation of the reactivity predictions of an ensemble of 10 models, and the predicted reactivity is given as the mean.

$$score_{AF-1} = \sum_{i \in target} r(reg,i) * \delta(reg,i) * \left(\frac{1}{n_{cand}} \sum_{j \in candidate} \frac{1}{d(j,i)+d_b} \right),$$

where n_{cand} is the total carbon count of the candidate molecule

Consequently, the selection was biased toward molecules that reduce model uncertainty, while focusing on improving the model's accuracy at the reactive centers of the target. The uncertainty consideration should provide an incentive towards diverse candidate selection, so no additional diversity-based selection was implemented. As a final consideration, with an eye towards experimental validation in the future, incorporating a selection bias towards easy-to-characterize substrates seemed important to avoid extensive, time-consuming experimental analysis. As a simple proxy for ease of characterization, the scores of the active learning acquisition function were divided by the carbon count of the candidate molecules, effectively biasing selection towards candidates that have fewer carbons (AF-10).

$$score_{AF-10} = \sum_{i \in target} r(reg,i) * \delta(reg,i) * \left(\frac{1}{n_{cand}^2} \sum_{j \in candidate} \frac{1}{d(j,i)+d_b} \right)$$

3.3 Experiment Setup

To evaluate the AFs without extensive synthetic burden, the acquisition workflow was simulated on the literature dataset of dioxirane-mediated oxidations. In this simulated workflow, AFs score candidates from the "simple" set of 135 reactions, and the best-scored reaction is added to the training set until the candidate pool is empty. A training trajectory is generated by fitting a random forest model on the training set at each time point. Since the AFs are target-specific, there will be one training trajectory per AF for each of the 50 complex target molecules in the dataset. The training molecule with the largest substructural overlap with the target was used to initiate sampling. The performance of each AF was measured by the number of experiments required to have a consistent top-1 accuracy on the target for at least 10 iterations. Random selection was used as a baseline.

3.4 Evaluation of Acquisition Strategies

To evaluate the AFs across the dataset, the difference in performances of the AF relative to random selection was computed over the subset of complex molecules that were predicted accurately by either random selection or the AF considered. The AF_{AL}, AF_{SC}, and AF_{CH} respectively spared 50, 51, and 40 data points on average compared to random selection (Fig 2b.). In the case where the AF did not provide improvement above the random selection, it was typically because random selection afforded an accurate prediction with fewer than 20 data points. In other words, the largest gains using the AF strategy were realized on targets that were most difficult to predict. Moreover, we observed that 27 to 31 targets were predicted accurately using AF-1 and AF-10, whereas random selection only predicts 19 correctly (Fig. 2c). This further suggests that a small but intentionally designed dataset can give better performance than larger ones for this type of task.

This analysis also provided some insight into the necessity of factoring in diversity versus similarity considerations (Fig. 2d) in dataset design. At least on this dataset, the pure diversity-based strategies

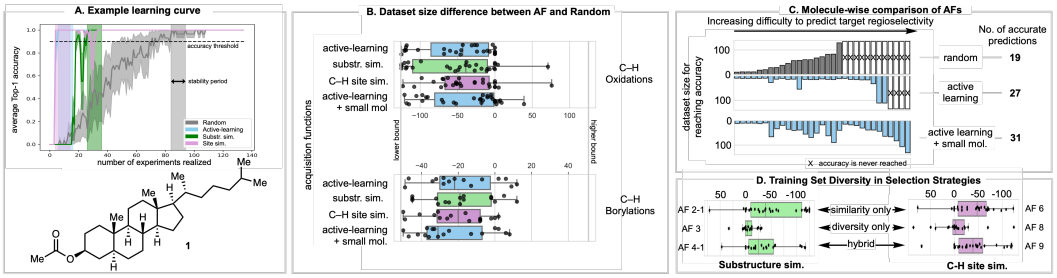


Figure 2: a) Example learning curve on steroid. b) Comparison of AFs with random selection on both literature datasets. c) Molecule-wise comparison of AF_{AL} with random selection. d) Performance of AFs with varying degrees of diversity selection.

(AF-3, AF-8) underperformed compared to their counterparts with similarity considerations. Indeed, their performance is comparable to random selection. This supports the conclusion that in low-data contexts where targets are known, optimizing for training set diversity alone is unwise. The hybrid approaches (AF-4-1 and AF-9) provided comparable, if slightly lower performances than their pure similarity-based counterparts. The literature dataset is a fairly diverse sampling space due to publishing pressures that aim to maximize novelty. It is possible that on more redundant datasets, these hybrid approaches may become more effective, though further testing is needed.

4 Additional Validation

4.1 Experimental Validation on Unseen Complex Targets

Encouraged by the results on molecules mined from literature, we tested whether similar gains could be observed on complex molecules outside of the dataset. Molecules were sourced from the 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.

Target molecules were selected that reflected the synthetic interests of total synthesis chemists and were anticipated to challenge the model to choose between similarly reactive sites. 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. Adamantane **5** was a product of a synthetic methodology project in the lab and requires the model to prioritize between tertiary and benzylic positions.

On 4/5 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 data points. Molecules **2**, **4**, and **5** can achieve stable accuracy within a dozen data points depending on the choice of AF. The improvement is especially pronounced for targets **3** and **4**, where stable accuracy cannot be achieved under random selection in 135 data points. The model struggles with (+)-sclareolide, perhaps weighting tertiary positions over electronic features. Longer range interactions, such as the deactivation of the top-ranked tertiary site by the lactone, seem to not be 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.

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

4.2 Literature Validation on C(sp^2)-H Radical Borylation

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

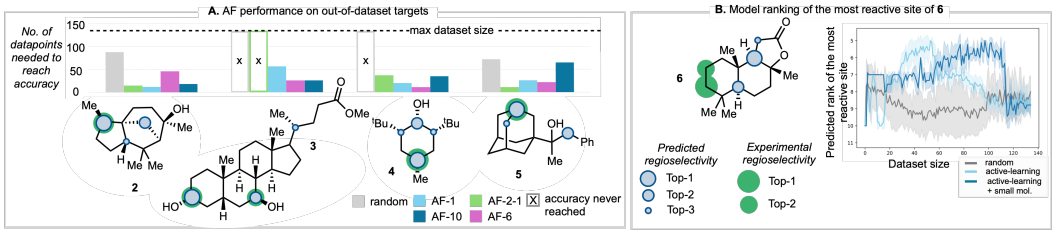


Figure 3: a) Bar plots of the performances of four AFs against random selection on the correctly predicted experimental targets. b) Learning curve depicting how the model's ranking of the most reactive site for (+)-sclareolide evolves with dataset size.

datasetNippa et al. [2023] (82 reactions including 22 large targets), filtered to include only reactions conducted under the same conditions. On this reaction, AF_{AL}, AF_{SC}, and AF_{CH} beat the random baseline, which is consistent with what was observed for C(sp^3)–H oxidation. 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 selection (Fig 2b). Additionally, molecules that could not be predicted accurately using 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%).

5 Studies on Stopping Criteria

A key component for full experimental implementation of the workflow described in 3.1 is assessment of when the acquisition loop can terminate. Our efforts thus far explore model uncertainty on the target, Laws and Schütze [2008] prediction stability, Bloodgood and Vijay-Shanker [2009] and AF scores as properties that can predict whether the model has achieved accuracy. As similarity-based thresholding has performed well in other predictive tasks in synthetic chemistry, we envisioned that the similarity-based AF scores could be an informative metric to track. Uncertainty was measured as the inverse of the fraction of models in the ensemble that agree on the top-ranked carbon, and prediction stability was calculated as the number of models that agree on the top-ranked carbon over a window of n training steps. These properties were paired with a gradient-, value-, or percentage-based threshold at which the acquisition loop would be stopped.

We define two objectives for a good stopping criterion: 1) **Goodness**: accuracy at the stop index, and 2) **Lateness**: area under the curve at the stop index. The ideal stopping criterion will have high goodness and low lateness. As a baseline, we compare to stopping at 10 training data points, which was the most common dataset size at which accuracy was reached across AFs. For reference, the lateness associated with stopping at the first instance of stable accuracy on AF-10 is 0.11.

	Baseline	Prediction Stability	Uncertainty	AF Score
Mean Goodness	0.31	**0.41**	0.12	0.33
Mean Lateness	0.02	0.03	**0.0004**	0.01

Table 3: Benchmarking of stopping criteria on AF-10. **Bold** numbers indicate the best-performing property per objective

From this initial work, we have the following findings:

- We observe modest improvement over the fixed baseline. With the AF score property on AF-10, we see improvement in both objectives simultaneously, but the gains are quite limited.
- A common failure point is that predictions always stabilize and uncertainty always falls, even on molecules that are never predicted accurately. For molecules that are predicted well, they tend to reach that accuracy quite early and therefore there is little differentiation from the baseline.

To build upon this, we are exploring additional properties, e.g., cross-validation accuracy on the training set, and the pairwise distances between carbon reactivity predictions as this distribution should become bimodal as the model differentiates reactive and unreactive sites. Additionally, we have ˜1100 training trajectories of 135 training steps. We are interested in whether a learning task can be framed around predicting whether acquisition should stop or not, given the training trajectory up to that point.

6 Conclusion

A reaction-agnostic acquisition-function based strategy for target-specific dataset design is reported. The approach presented is effective 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 data points. Furthermore, this work demonstrates that AF-designed datasets can provide accuracy on more targets than larger, randomly acquired datasets; an improvement of 24% and 23% is reported for the two datasets respectively. An experimental validation on a set of five complex targets was performed and confirmed the trends observed on the literature data. To conclude, efforts towards developing a stopping criterion for terminating the active learning loop are included.

References

Waldemar Adam, Cong-Gui Zhao, and Kavitha Jakka. *Dioxirane Oxidations of Compounds other than Alkenes*, chapter 1, pages 1–346. John Wiley Sons, Ltd, 2008. ISBN 9780471264187. doi: https://doi.org/10.1002/0471264180.or069.01. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/0471264180.or069.01`.

Gregorio Asensio, Gloria Castellano, Rossella Mello, and M. E. González Núñez. Oxyfunctionalization of aliphatic esters by methyl(trifluoromethyl)dioxirane. *The Journal of Organic Chemistry*, 61(16):5564–5566, 1996. doi: 10.1021/jo9604189. URL `https://doi.org/10.1021/jo9604189`.

Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of Chemical Theory and Computation*, 15 (3):1652–1671, 2019. doi: 10.1021/acs.jctc.8b01176. URL `https://doi.org/10.1021/acs.jctc.8b01176`. PMID: 30741547.

Isaiah O. Betinol, Aleksandra Demchenko, and Jolene P. Reid. Evaluating predictive accuracy in asymmetric catalysis: A machine learning perspective on local reaction space. *ACS Catalysis*, 15(8):6067–6077, 2025. doi: 10.1021/acscatal.5c01051. URL `https://doi.org/10.1021/acscatal.5c01051`.

Michael Bloodgood and K. Vijay-Shanker. A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. In Suzanne Stevenson and Xavier Carreras, editors, *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 39–47, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL `https://aclanthology.org/W09-1107/`.

Paolo Bovicelli, Augusto Gambacorta, Paolo Lupattelli, and Enrico Mincione. A highly regio- and stereoselective c5 oxyfunctionalization of coprostane steroids by dioxiranes: An improved access to progestogen and androgen hormones. *Tetrahedron Letters*, 33(48):7411–7412, 1992a. ISSN 0040-4039. doi: https://doi.org/10.1016/S0040-4039(00)60202-2. URL `https://www.sciencedirect.com/science/article/pii/S0040403900602022`. The International Journal for the Rapid Publication of Preliminary.

Paolo Bovicelli, Paolo Lupattelli, Enrico Mincione, Teresa Prencipe, and Ruggero Curci. 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. *The Journal of Organic Chemistry*, 57(19):5052–5054, 1992b. doi: 10.1021/jo00045a004. URL `https://doi.org/10.1021/jo00045a004`.

Eike Caldeweyher, Masha Elkin, Golsa Gheibi, Magnus Johansson, Christian Sköld, Per-Ola Norrby, and John F. Hartwig. Hybrid machine learning approach to predict the site selectivity of iridium-catalyzed arene borylation. *Journal of the American Chemical Society*, 145(31):17367–17376, 2023. doi: 10.1021/jacs.3c04986. URL https://doi.org/10.1021/jacs.3c04986. PMID: 37523755.

Jack K. Crandall, Ruggero Curci, Lucia D'Accolti, Caterina Fusco, Caterina Fusco, Lucia D'Accolti, and Cosimo Annese. *Methyl(trifluoromethyl)dioxirane*, pages 1–11. John Wiley Sons, Ltd, 2016. ISBN 9780470842898. doi: https://doi.org/10.1002/047084289X.rm267.pub3. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/047084289X.rm267.pub3.

Ruggero Curci, Lucia D'Accolti, and Caterina Fusco. A novel approach to the efficient oxygenation of hydrocarbons under mild conditions. superior oxo transfer selectivity using dioxiranes. *Accounts of Chemical Research*, 39(1):1–9, 2006. doi: 10.1021/ar050163y. URL https://doi.org/10.1021/ar050163y. PMID: 16411734.

Lucia D'Accolti, Cosimo Annese, and Caterina Fusco. Continued progress towards efficient functionalization of natural and non-natural targets under mild conditions: Oxygenation by ch bond activation with dioxirane. *Chemistry – A European Journal*, 25(52):12003–12017, 2019. doi: https://doi.org/10.1002/chem.201901687. URL https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/chem.201901687.

Spencer D. Dreher and Shane W. Krska. Chemistry informer libraries: Conception, early experience, and role in the future of cheminformatics. *Accounts of Chemical Research*, 54(7):1586–1596, 2021. doi: 10.1021/acs.accounts.0c00760. URL https://doi.org/10.1021/acs.accounts.0c00760. PMID: 33723992.

Tarek H. El-Assaad, Jayden Zhu, Anjitha Sebastian, Dominic V. McGrath, Ishita Neogi, and Keshaba N. Parida. Dioxiranes: a half-century journey. *Org. Chem. Front.*, 9:5675–5725, 2022. doi: 10.1039/D2QO01005D. URL http://dx.doi.org/10.1039/D2QO01005D.

Caterina Fusco, Michele Fiorentino, Anna Dinoi, Ruggero Curci, Ralph A. Krause, and Dietmar Kuck. Oxyfunctionalization of non-natural targets by dioxiranes. 2. selective bridgehead dihydroxylation of fenestrindane1. *The Journal of Organic Chemistry*, 61(24):8681–8684, 1996. doi: 10.1021/jo961316l. URL https://doi.org/10.1021/jo961316l.

María E. González-Nuñez, Jorge Royo, Gloria Castellano, Cecilia Andreu, Carmen Boix, Rossella Mello, and Gregorio Asensio. Hyperconjugative control by remote substituents of diastereoselectivity in the oxygenation of hydrocarbons. *Organic Letters*, 2(6):831–834, 2000. doi: 10.1021/ol000017m. URL https://doi.org/10.1021/ol000017m. PMID: 10814435.

Yanfei Guan, Taegyo Lee, Ke Wang, Shu Yu, and J. Christopher McWilliams. Snar regioselectivity predictions: Machine learning triggering dft reaction modeling through statistical threshold. *Journal of Chemical Information and Modeling*, 63(12):3751–3760, 2023. doi: 10.1021/acs.jcim.3c00580. URL https://doi.org/10.1021/acs.jcim.3c00580. PMID: 37272922.

Lucas Guillemard, Nikolaos Kaplaneris, Lutz Ackermann, and Magnus J. Johansson. Late-stage c–h functionalization offers new opportunities in drug discovery. *Nature Reviews Chemistry*, 5(8): 522–545, Jul 2021. doi: 10.1038/s41570-021-00300-6.

Yuzuru Kanda, Hugh Nakamura, Shigenobu Umemiya, Ravi Kumar Puthukanoori, Venkata Ramana Murthy Appala, Gopi Krishna Gaddamanugu, Bheema Rao Paraselli, and Phil S. Baran. Two-phase synthesis of taxol. *Journal of the American Chemical Society*, 142(23):10526–10533, 2020. doi: 10.1021/jacs.0c03592. URL https://doi.org/10.1021/jacs.0c03592. PMID: 32406238.

Jae Young Kim, Salman A. Khan, and Dionisios G. Vlachos. Similarity-based machine learning for small data sets: Predicting biolubricant base oil viscosities. *The Journal of Physical Chemistry B*, 128(48):11963–11970, 2024. doi: 10.1021/acs.jpcb.4c06687. URL https://doi.org/10.1021/acs.jpcb.4c06687. PMID: 39579140.

F. Kovač and A.L. Baumstark. Oxidation of -methylbenzyl alcohols by dimethyldioxirane. *Tetrahedron Letters*, 35(47):8751–8754, 1994. ISSN 0040-4039. doi: https://doi.org/10.1016/S0040-4039(00)78488-7. URL https://www.sciencedirect.com/science/article/pii/S0040403900784887.

Florian Laws and Hinrich Schütze. Stopping criteria for active learning of named entity recognition. In Donia Scott and Hans Uszkoreit, editors, *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 465–472, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL https://aclanthology.org/C08-1059/.

Dominik Lemm, Guido Falk von Rudorff, and O Anatole von Lilienfeld. Improved decision making with similarity based machine learning: Applications in chemistry. *Machine Learning: Science and Technology*, 4(4):045043, Dec 2023. doi: 10.1088/2632-2153/ad0fa3.

Mathieu Lesieur, Claudio Battilocchio, Ricardo Labes, Jérôme Jacq, Christophe Genicot, Steven V. Ley, and Patrick Pasau. Direct oxidation of csp3h bonds using in situ generated trifluoromethylated dioxirane in flow. *Chemistry – A European Journal*, 25(5):1203–1207, 2019. doi: https://doi.org/10.1002/chem.201805657. URL https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/chem.201805657.

Matthew A. McDonald, Brent A. Koscher, Richard B. Canty, and Klavs F. Jensen. Calibration-free reaction yield quantification by hplc with a machine-learning model of extinction coefficients. *Chem. Sci.*, 15:10092–10100, 2024. doi: 10.1039/D4SC01881H. URL http://dx.doi.org/10.1039/D4SC01881H.

Rossella Mello, Michele Fiorentino, Caterina Fusco, and Ruggero Curci. Oxidations by methyl(trifluoromethyl)dioxirane. 2. oxyfunctionalization of saturated hydrocarbons. *Journal of the American Chemical Society*, 111(17):6749–6757, 1989. doi: 10.1021/ja00199a039. URL https://doi.org/10.1021/ja00199a039.

Rossella Mello, Luigi Cassidei, Michele Fiorentino, Caterina Fusco, and Ruggero Curci. Oxidations by methyl(trifluoromethyl)dioxirane. 3. selective polyoxyfunctionalization of adamantane. *Tetrahedron Letters*, 31(21):3067–3070, 1990. ISSN 0040-4039. doi: https://doi.org/10.1016/S0040-4039(00)89027-9. URL https://www.sciencedirect.com/science/article/pii/S0040403900890279.

Steven M. Mennen, Carolina Alhambra, C. Liana Allen, Mario Barberis, Simon Berritt, Thomas A. Brandt, Andrew D. Campbell, Jesús Castañón, Alan H. Cherney, Melodie Christensen, David B. Damon, J. Eugenio de Diego, Susana García-Cerrada, Pablo García-Losada, Rubén Haro, Jacob Janey, David C. Leitch, Ling Li, Fangfang Liu, Paul C. Lobben, David W. C. MacMillan, Javier Magano, Emma McInturff, Sebastien Monfette, Ronald J. Post, Danielle Schultz, Barbara J. Sitter, Jason M. Stevens, Iulia I. Strambeanu, Jack Twilton, Ke Wang, and Matthew A. Zajac. The evolution of high-throughput experimentation in pharmaceutical development and perspectives on the future. *Organic Process Research & Development*, 23(6):1213–1242, 2019. doi: 10.1021/acs.oprd.9b00140. URL https://doi.org/10.1021/acs.oprd.9b00140.

David F. Nippa, Kenneth Atz, Remo Hohler, Alex T. Müller, Andreas Marx, Christian Bartelmus, Georg Wuitschik, Irene Marzuoli, Vera Jost, Jens Wolfard, and et al. Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning. *Nature Chemistry*, 16(2):239–248, Nov 2023. doi: 10.1038/s41557-023-01360-5.

Takayuki Oritani, Tohru Horiguchi, Masanori Nagura, Qian Cheng, and Toshio Kudo. Chemical oxidation of taxoids with m-cpba and dimethyl dioxirane: Regioselective epoxidation of taxinine j derivatives. *HETEROCYCLES*, 53(12):2629, 2000. doi: 10.3987/com-00-8929.

Debanjan Rana, Philipp M. Pflüger, Niklas P. Hölter, Guangying Tan, and Frank Glorius. Standardizing substrate selection: A strategy toward unbiased evaluation of reaction generality. *ACS Central Science*, 10(4):899–906, 2024. doi: 10.1021/acscentsci.3c01638. URL https://doi.org/10.1021/acscentsci.3c01638.

Shree Sowndarya S. V., Yeonjoon Kim, Seonah Kim, Peter C. St. John, and Robert S. Paton. Expansion of bond dissociation prediction with machine learning to medicinally and environmentally relevant chemical space. *Digital Discovery*, 2:1900–1910, 2023. doi: 10.1039/D3DD00169E. URL http://dx.doi.org/10.1039/D3DD00169E.

Raffaele Saladino, Maurizio Mezzetti, Enrico Mincione, Ines Torrini, Mario Paglialunga Paradisi, and Gaia Mastropietro. A new and efficient synthesis of unnatural amino acids and peptides by selective 3,3-dimethyldioxirane side-chain oxidation. *The Journal of Organic Chemistry*, 64(23):8468–8474, 1999. doi: 10.1021/jo990185w. URL https://doi.org/10.1021/jo990185w.

Gennady V. Shustov and Arvi Rauk. Mechanism of dioxirane oxidation of ch bonds: application to homo- and heterosubstituted alkanes as a model of the oxidation of peptides. *The Journal of Organic Chemistry*, 63(16):5413–5422, 1998. doi: 10.1021/jo9802877. URL `https://doi.org/10.1021/jo9802877`.

Jason Y. Wang, Jason M. Stevens, Stavros K. Kariofillis, Mai-Jan Tom, Dung L. Golden, Jun Li, Jose E. Tabora, Marvin Parasram, Benjamin J. Shields, David N. Primer, and et al. Identifying general reaction conditions by bandit optimization. *Nature*, 626(8001):1025–1033, Feb 2024. doi: 10.1038/s41586-024-07021-y.

Jie Yu, Linqi Zhu, Ruibao Qin, Zhansong Zhang, Li Li, and Tao Huang. Combining k-means clustering and random forest to evaluate the gas content of coalbed bed methane reservoirs. *Geofluids*, 2021(1):9321565, 2021. doi: https://doi.org/10.1155/2021/9321565. URL `https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/9321565`.

Lufeng Zou, Robert S. Paton, Albert Eschenmoser, Timothy R. Newhouse, Phil S. Baran, and K. N. Houk. Enhanced reactivity in dioxirane c–h oxidations via strain release: A computational and experimental study. *The Journal of Organic Chemistry*, 78(8):4037–4048, 2013. doi: 10.1021/jo400350v. URL `https://doi.org/10.1021/jo400350v`. PMID: 23461537.

A Technical Appendices and Supplementary Material

All the code and the data needed to reproduce the results are available on GitHub at the address: (removed to maintain anonymity). A detailed README file is given to facilitate reproduction of the results. Links to the notebooks and python scripts used to reproduce figures and results are provided along with the supporting information.

A.1 List of publications mined for dioxirane regioselectivity reaction data

Data extraction was performed manually from 16 publications as follows Asensio et al. [1996], González-Nuñez et al. [2000], Bovicelli et al. [1992a], Adam et al. [2008], Bovicelli et al. [1992b], Crandall et al. [2016], Fusco et al. [1996], Mello et al. [1989, 1990], D'Accolti et al. [2019], Oritani et al. [2000], Kovač and Baumstark [1994], Lesieur et al. [2019], Saladino et al. [1999], Shustov and Rauk [1998], El-Assaad et al. [2022].

A.2 Dataset preprocessing

The main preprocessing workflow is detailed in the script: preprocess_reactions.py The preprocessing work is as follows. First, data is loaded from the shared Google spreadsheet or any other spreadsheet with the same column names. Second, the following filters are applied:

- Remove the reactions with a "Yes" in the "Discard" column: chemist's decision to remove the reaction (example reasons that caused reactions to be discarded: reactants are not dioxiranes, an additional catalyst is used).
- Drop reactions with nan in the "Reactant_SMILES" column.
- Drop reactions with no selectivity or yield data.
- Canonicalize reactants and products SMILES.
- Concatenate reactant and product to generate rxn_SMILES.
- Map reactant to product atoms with RXN_mapper to generate rxn_mapper_smiles.
- Identify reactive sites.
- Map rxn_mapper_smiles to canonical SMILES.

Third, the results for each pair of reactant-product are combined into a single dictionary with canonical carbon indexes mapped to their corresponding selectivity. We then account for symmetry by looking for equivalent sites and reduce the number of carbons to the unique ones. The number of equivalent carbons present in the molecule then normalizes selectivity. Selectivity is normalized such that the sum over all unique carbons equals 1.

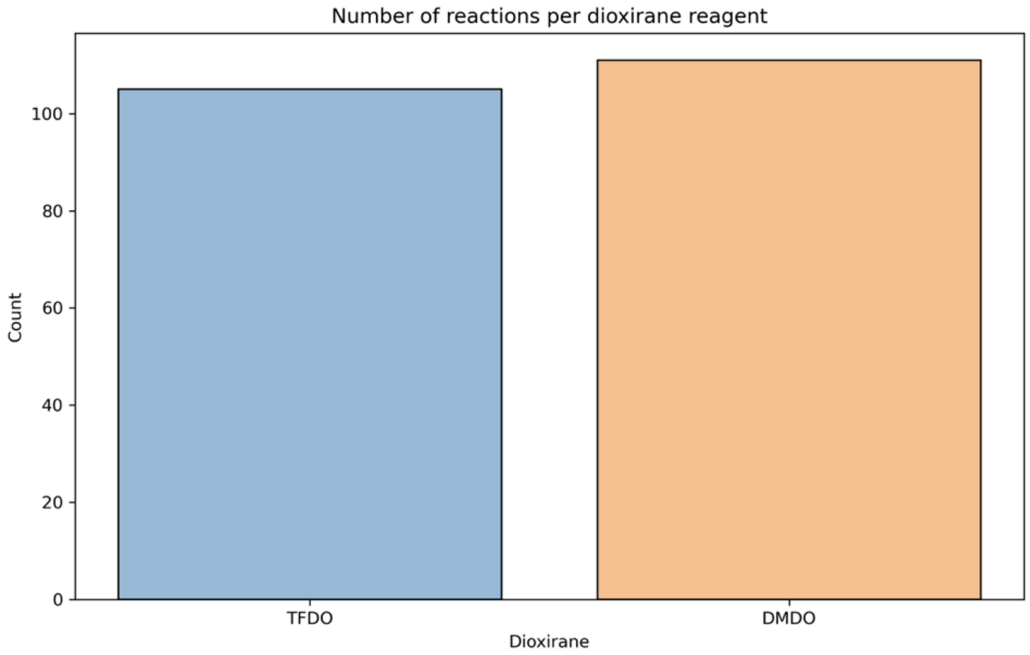


Figure 4: Distribution TFDO/DMDO: 105/111 reactions respectively

Finally, the resulting data is saved in numbered_reaction_1.csv: which has the columns "Reactant_SMILES" giving the canonical SMILES of the substrate and "Selectivity _Reduced" which is a dictionary relating atom_idx in the canonical SMILES to the corresponding selectivity.

Additional filters: Additional filters are described in detail and realized using the script: data/Filter_data.py. The filters implemented are as follows: - All reactions in which the reactant was a mixture of diastereomers that were not specified were filtered out. Racemic mixtures were tolerated. - Amines were used in their protonated form to compute the descriptors because these reactions are usually conducted in the presence of HBF_4 to avoid the reaction of the dioxirane with the amine leading to the formation of the N-oxides instead of C–H activation products. - Directed C–H oxidations were excluded, given the scope of the work is limited to undirected reactions. The data from two articles were excluded because they are examples of intramolecular directed C–H oxidation.

A.3 Reaction conditions in the dataset

Using the data mined, we have 216 reactions distributed between TFDO and DMDO reagents as depicted in the figures below. Details on reaction solvent, time, and temperature are provided. Figures can be reproduced using the notebook figures/$04_{r}eaction_{c}onditions.ipynb$.

TFDO reactions are run at lower temperatures on average compared to DMDO, which balances the stronger reactivity of TFDO. DMDO is mostly used in acetone or acetone-containing solvent mixtures, and TFDO is mostly used in trifluoroacetone or trifluoroacetone-containing solvent mixtures. The co-solvent tends to be halogenated solvents, with the exception of acetonitrile (ACN). The reaction times tend to follow similar distributions for both TFDO and DMDO.

A.4 Descriptor Calculation

The featurization is done with the functions written in utils/modelling.py and utils/descriptors.py. It works as follows:

- "prepare_reactivity_mapping" is called on a dataframe containing the list of SMILES whose descriptors will be extracted

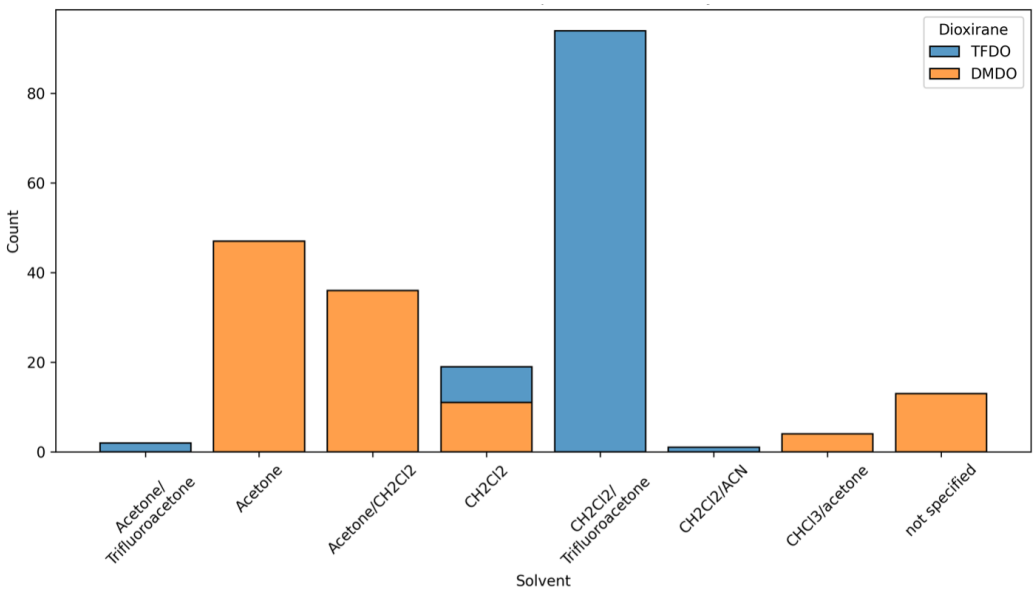


Figure 5: Distribution of reaction solvent

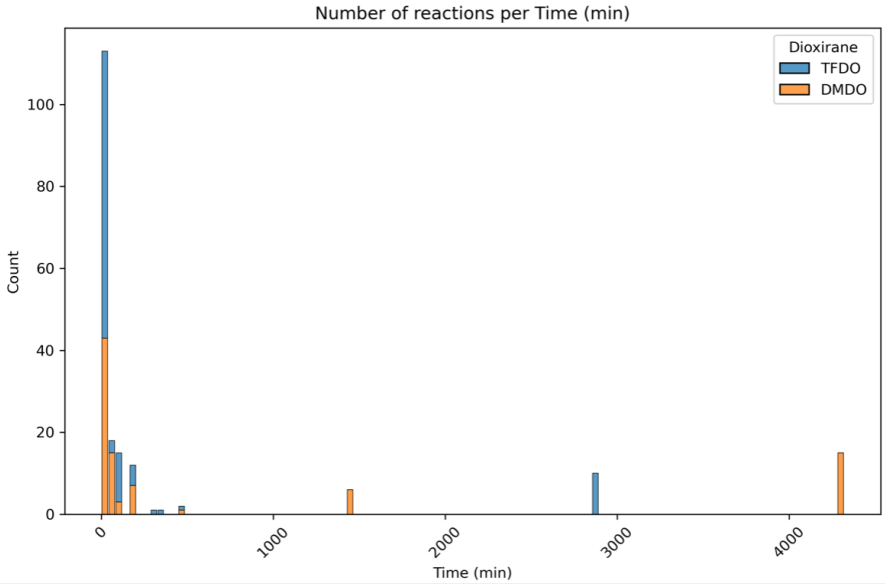


Figure 6: Distribution of reaction time

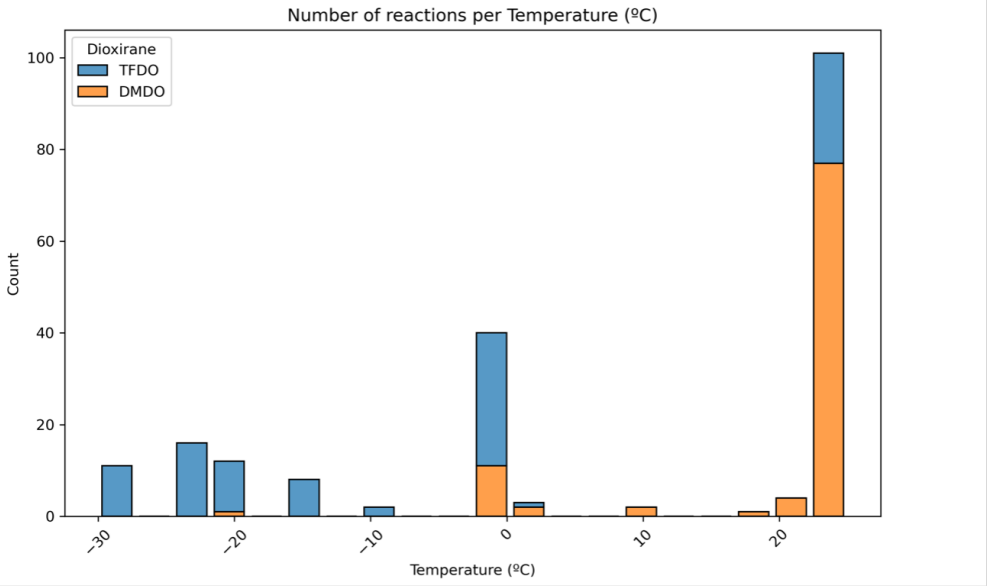


Figure 7: Distribution of reaction temperature

- "prepare_reactivity_mapping" calls “extract_features”, which reads the json files in data/descriptors/smiles_descriptors/ to retrieve the descriptors for each SMILES. If the descriptors are not in the json file, they will be computed on the fly. Note that BDE must be computed separately; a warning is displayed when running the code if BDE has not been computed. Details of BDE calculation is provided in a subsequent section.
- If requested, features are normalized between 0 and 1.
- If threshold is not equal to None, correlated features (with correlation coefficient exceeding the threshold) are dropped
- A data frame is returned for later modeling.

The script data/preprocess_reactions.py is provided to avoid recompiling descriptors when evaluating regioselectivity performances and acquisition functions, by storing the precomputed descriptor dataframes in a subfolder of data/descriptors/. In this script, descriptor types and the input file can be specified through command line arguments.

Descriptors reported: Our modeling is framed such that reactive “carbon” sites are compared and not C–H bonds, because of (1) the difficulty to automatically map C–H bonds from reactants to products, and (2) the fact that some reactions feature a formal CH2 to C=O transformation making it impossible to differentiate the reactivity of the two C–H bonds in the reactant, which is important if these protons are diastereotopic. Therefore, to gather a homogeneous representation for the carbon sites of primary, secondary, and tertiary carbons we decided to report maximum, average, and minimum values of the descriptors that are hydrogen based. The descriptors are computed using the script: utils/descriptors.py.

BDE: Each carbon center is featured with max, min, and average of C–H BDE and Bond Dissociation Free Energies (BDFE) predicted for the carbon center, using the model reported by the Paton group.16 The script we use is provided in data/bdes/compute_bdes.py and requires a python environment that has the TensorFlow package installed.

xTB-Morfeus: The 3D geometries of the molecules are optimized with xTB, then the descriptors are generated using the morfeus-ml python package.

xtb command line options used: xtb temp.xyz –opt extreme –gnf=2 –json

DBSTEP (Sterics): The 3D geometries of the molecules are optimized with xTB, then the descriptors are generated using DBSTEP.

dbstep command: mol = db.dbstep(f"base_cwd/utils/xtb_utils/xtb_f_name/temp.xyz", atom1 = C_idx+1, atom2 = H_idx_dbstep, commandline = True, verbose = False, sterimol = True, volume = True, scan = '2.0:4.5:0.5', measure = 'classic')

Gasteiger (Electronics): Gasteiger charges generated by RDkit.

ENV1 (Local Environment): Descriptors for the environment of the carbon in the reactive site are the number of neighbors in the following categories: O, N, H, C, C(sp2), C(sp3), aromatic C.

Any combination of the descriptors above can be made *a posteriori*, and any dataframe with descriptors can be modeled as long as the columns: 'Reactant_SMILES', 'Atom_nº', 'Selectivity', 'Reactive Atom' are present. As such, we also tested the model with the following combinations of descriptors:

Custom (Chemist Selection): We decided to describe the molecules with a simple featurization that makes chemical sense using 5 parameters: %Vbur for C and H, charges for C and H (computed with AIMNET2), and the predicted BDE.

Selected (ML Selection): These descriptors are selected using permutation importance from the RF2 model described in section IV. The permutation importance of each feature in the previously described descriptors is computed and then all the descriptors with an importance of more than 0.1 are concatenated. The selected features are the following:

XTB-Morfeus: 'Buried_Volume_C', 'V_occ_avg', 'Pyramidalization_H_max', 'Buried_Volume_H_max_MFF', 'Pyramidalization_C_MFF', 'Pyramidalization_H_max_MFF', 'Local_Nucleophilicity_H_max', 'dual_H_max'

Local env.: 'num_H', 'num_C', 'n_Csp3'

Electronics: 'gas_charge_H_max', 'gas_charge_C',

Sterics: 'Bmax_2.0_min', 'Bmax_2.0_avg', 'Bmin_3.5_avg', 'L_ch1_avg', 'L_cc1_min', 'Bmin_2.0_avg', 'Bmin_3.0_avg'

BDE: 'bde_avg'

Analysis of the ML selected features revealed that the 23 machine-selected descriptors included the 5 that were chosen *a priori* by experts to build the custom chemist-selected feature set, suggesting that the model was able to extract some relevant reactivity features from the regioselectivity dataset.

A.5 Dioxirane target molecules

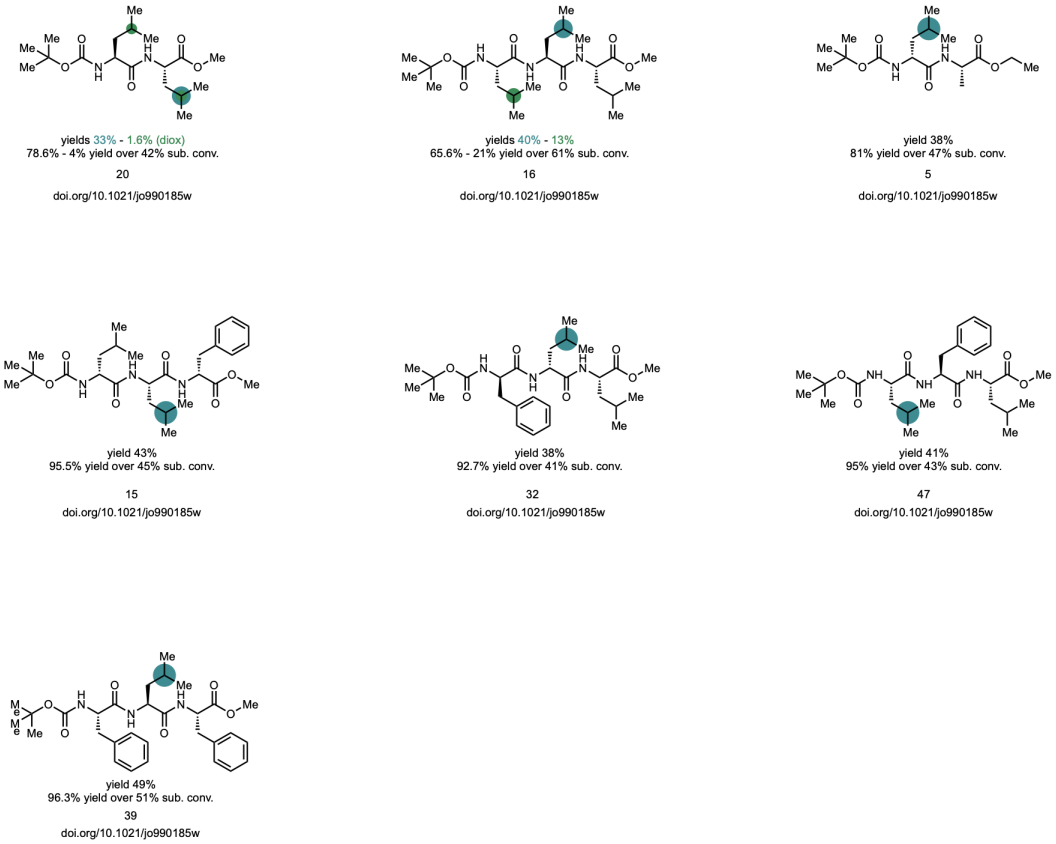


Figure 8: Complex peptide targets

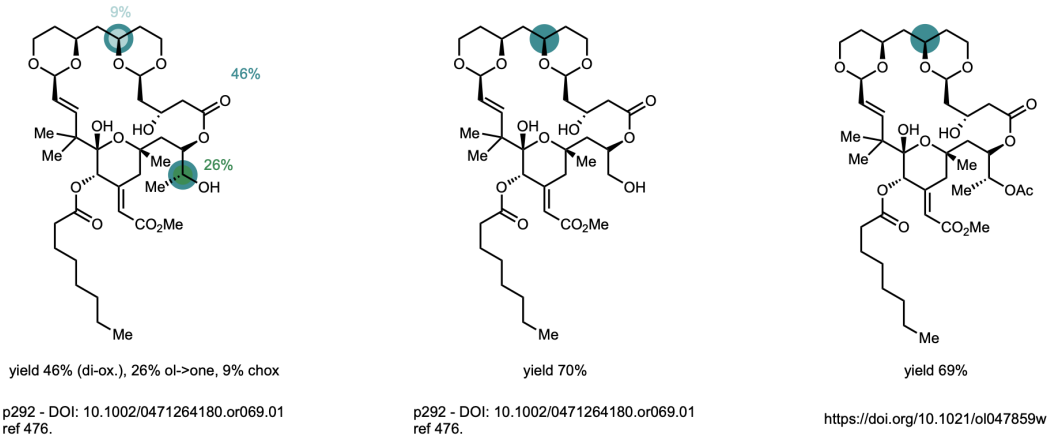
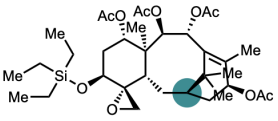
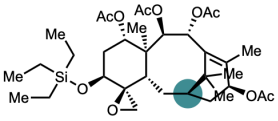


Figure 9: Complex macrocycle targets

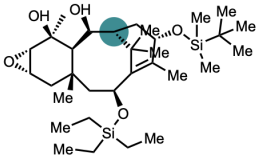




hydroxy-epoxydes were obtained from reacting the corresponding alkene in 51% and 45% yields (96% overall)

Heterocycles 2000, 53, 2629

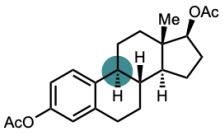
Heterocycles 2000, 53, 2629



49% gram scale

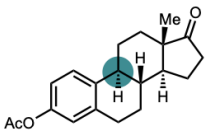
10.1021/jacs.0c03592

Figure 10: Complex taxinine targets



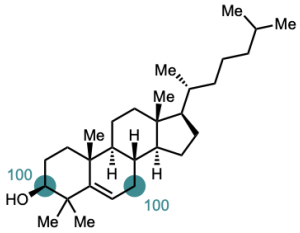
yield 75%

p270 - DOI: 10.1002/0471264180.or069.01
ref 457.



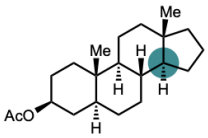
yield 80%

p270 - DOI: 10.1002/0471264180.or069.01
ref 457.



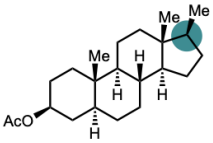
yield 36% - no details about sm

p309 - DOI: 10.1002/0471264180.or069.01
ref 482.



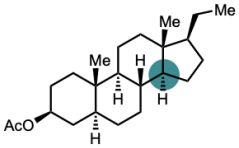
yield 43%
(reported as 50% conv yield 85%)

science/article/pii/S0040403900617392
p271 - DOI: 10.1002/0471264180.or069.01
ref 459.



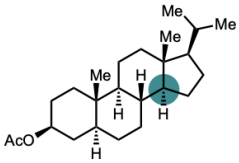
yield 45%
reported as 50% conv and yield 90%

science/article/pii/S0040403900617392
p271 - DOI: 10.1002/0471264180.or069.01
ref 459.



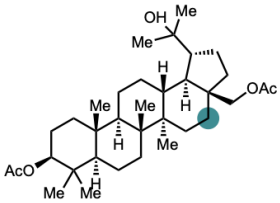
yield 36%
reported as 40% conv and yield 90%

science/article/pii/S0040403900617392
p271 - DOI: 10.1002/0471264180.or069.01
ref 459.



yield 36%
reported as 40% conv and yield 90%

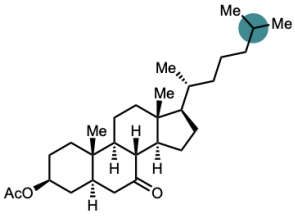
science/article/pii/S0040403900617392
p271 - DOI: 10.1002/0471264180.or069.01
ref 459.



yield 30%, identified as major product

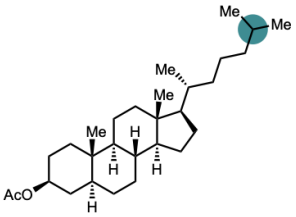
DOI: 10.1002/047084289X.rm267.pub3
(fig 36)

Figure 11: Complex C5-alpha steroid targets



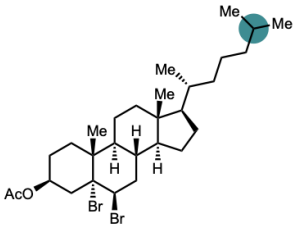
70% based on rsm, 52% yield overall

10.1002/047084289X.rm267.pub3
eq. 34



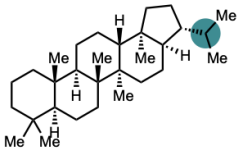
yield 81-85%

J. Org.Chem. 1992, 57, 5052



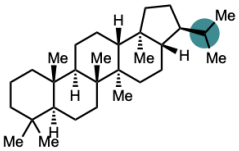
yield 80-88%

J. Org.Chem. 1992, 57, 5052



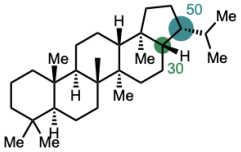
yield (10% recov sm ~70%)

p287 - DOI: 10.1002/0471264180.or069.01
ref 472



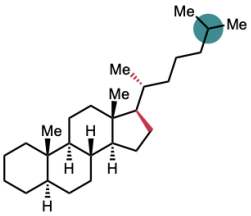
yield (15% recov sm ~70%)

p287 - DOI: 10.1002/0471264180.or069.01
ref 472



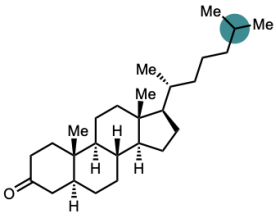
yield (50% - 30%, 15% unidentified polar) 21h
or 12h: (30% - 20% and 30% recov. sm)

p287 - DOI: 10.1002/0471264180.or069.01
ref 472



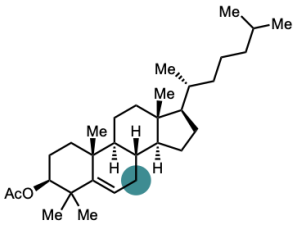
yield 18%, rsm: 79%

10.1002/chem.201805657



yield 85%

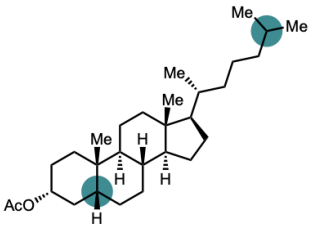
10.1002/047084289X.rm267.pub3
eq. 11



yield 44% - no details about sm

p309 - DOI: 10.1002/0471264180.or069.01
ref 482
Tetrahedron Lett. 1991, 32, 533

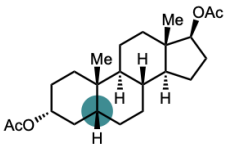
Figure 12: Complex C5-alpha steroid targets (cont.)



no yield reported
(double oxidized product)

https://www.sciencedirect.com/science/article/pii/S0040403900602022
reactant **11**, product = **12**

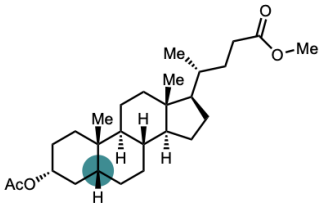
TetrahedronLett. 1992, 33, 7411.



50% conv. 90% isolated yield

https://www.sciencedirect.com/science/article/pii/S0040403900602022
reactant **7**, product = **8**

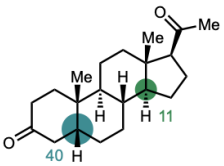
TetrahedronLett. 1992, 33, 7411.



30% conv. 85% isolated yield (DMDO)
70% conv. 85% isolated yield (TFDO)

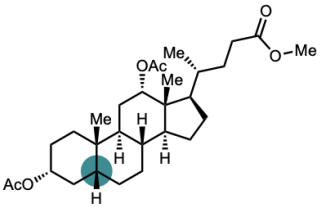
https://www.sciencedirect.com/science/article/pii/S0040403900602022
product = **3**

TetrahedronLett. 1992, 33, 7411.



yield 40%, 11% and 33% sm

p282 - DOI: 10.1002/0471264180.or069.01
ref 458, Synth. Commun. 1993, 23, 135



yield 35% of a “single product”
53% of unreacted sm

p282 - DOI: 10.1002/0471264180.or069.01
ref 458

Figure 13: Complex C5-beta steroid targets

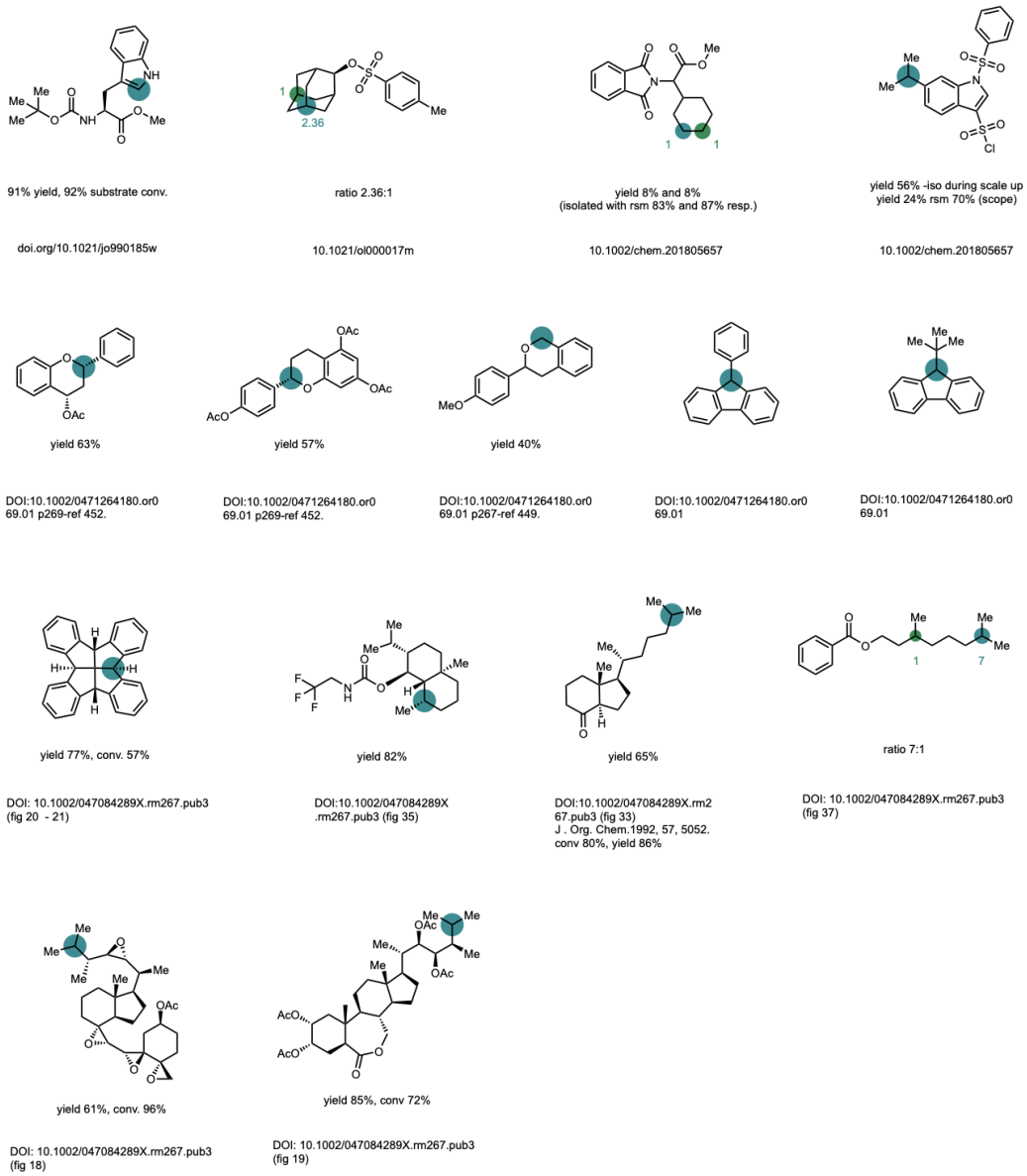


Figure 14: Complex uncategorized targets