Title

Application of Machine Learning and Data Science in Synthetic Organic Chemistry

Author

Wang, Jason

Publication Date

2025

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Application of Machine Learning and Data Science in Synthetic Organic Chemistry

A dissertation submitted in partial satisfaction of the requirements

for the degree Doctor of Philosophy in Chemistry

by

Jason Yiheng Wang

2025

ABSTRACT OF THE DISSERTATION

Application of Machine Learning and Data Science in Synthetic Organic Chemistry

by

Jason Yiheng Wang

Doctor of Philosophy in Chemistry

University of California, Los Angeles, 2025

Professor Abigail Gutmann Doyle, Chair

Chapter 1 describes the development of Auto-QChem, an automated, high-throughput and end-to-end density functional theory (DFT) calculation tool that can generate quantum chemical descriptors for organic molecules. We discuss in detail the design and implementation of Auto-QChem, as well as its current functionalities. We also review literature examples in synthetic organic chemistry where Auto-QChem-derived descriptors were applied in machine learning (ML) models to accelerate methodology development.

Chapter 2 describes the design, implementation and application of reinforcement learning bandit optimization models in chemistry reaction optimization, where generally applicable reaction conditions were identified via efficient condition sampling and evaluation of experimental feedback. In addition to performance benchmarking on existing reaction datasets in literature, we also experimentally investigated a palladium-catalyzed imidazole C–H arylation reaction, an

aniline amide coupling reaction and a phenol alkylation reaction. In all three cases, bandit optimization models identified most generally applicable yet not well studied reaction conditions for the respective reaction.

Chapter 3 describes the discovery and characterization of multiple *N*-(hetero)aryl, *N*-benzyl and *N*-alkyl derivatives of the 9-mesityl-3,6-di-*tert*-butyl-10-phenyl acridinium photocatalyst. The catalytic performances of these catalysts as photo-oxidant or photo-reductant (via *in situ* generated acridine radical) were compared in three model reactions. We also identified improved catalytic conditions for a previously reported cyanoarene-catalyzed nucleophilic amination reaction using a synthesized *N*-cycloheptyl acridinium catalyst with up to 98% reaction yield.

The dissertation of Jason Yiheng Wang is approved.

Anastassia N. Alexandrova

Kendall N. Houk

Alexander Michael Spokoyny

Abigail Gutmann Doyle, Committee Chair

University of California, Los Angeles

2025

Table of Contents

List of Figures

List of Tables

Acknowledgements

My journey started at Columbia during my freshman year (2016), when I attended a symposium where faculties introduced their research to undergraduate students. Tom had just moved to Columbia a year ago from Colorado State and started his talk on C–H activation with a structure of methane on the slide: one carbon atom, four hydrogen atoms and four sticks. He said (and I paraphrase): "Wouldn't it be amazing if we can individually functionalize every single one of these C–H bond? We would be able to build every molecule just from methane!" At the time, freshman me thought that was indeed amazing. We have a lot of methane; we can make every single molecule from it and synthetic chemistry is solved! Obviously, that wasn't the case, as Tom proceeded to show his actual slides (taken from students, of course) with the reaction, where you need amine with an electron withdrawing group, exactly one C–H bond (to prevent over-functionalization) exactly four carbons away (1,5–HAT), iridium photocatalyst ($1000 per gram) and many other things. Did I mention you also need a blue LED lamp used in fish tanks? I felt slightly deceived and disappointed; what about the other 25 C–H bonds in that molecule? How are they going to get functionalized?

Nonetheless, I walked up to Tom after the symposium and introduced myself. I told him I am interested in doing research in his group, and he asked me to come to his office at a certain day and time to do a quick interview. I do not remember the details of that conversation, but I must have answered his questions quite well. He led me into the lab afterwards and introduced me to a graduate student in the group, Ben Ravetz, and made me promise that I will work in the group for at least six months. That six months eventually turned into three and a half years, and I became good friends with many people in the Rovis group over that time: Ben (who taught me everything I know in lab, good and bad), Melissa, Erik, Isra, Corey, Nick, Vanessa… It was a great

environment with a group of people that had a lot of fun together (too many after midnight parties in the conference room) but were also extremely hard-working and passionate about chemistry. I really enjoyed my time there, and regret that it was cut short by the COVID-19 pandemic. Tom has been an amazing mentor even after I graduated, and I will always appreciate his support, guidance and the fact that he took a chance on me and gave me my first opportunity.

About when I started working in the Rovis group, I also decided that I was going to double major in computer science, in addition to biochemistry which I had always wanted to. I was genuinely intrigued by all the machine learning talks and was curious to find out what the fuss is all about. My computer science studies were mostly independent from chemistry and what I was working on in the Rovis group (which was a series of cobalt-catalyzed alkyne hydrofunctionalization). In the summer of my sophomore year, Abby came to Columbia to give a talk on the new machine learning paper she had just published, which uses machine learning model trained on chemical descriptors to predict the reactivities of C–N cross-coupling (*Science* **2018**, *360*, 186-190). I was not thinking about graduate school at the time but was still very excited to see that a cutting-edge computer science tool gets its first application in synthetic organic chemistry.

And so it begins, I started brainstorming about research ideas and potential applications of machine learning in chemistry. I had some ideas about training machine learning models to predict the reactivities of photoredox reactions. I decided on a test reaction, which is a red-light up-conversion system with conjugated porphyrins (photosensitizer) and perylenes (annihilator). My proposal includes making and testing a library of different porphyrins with various functional groups and see if I can derive any structural-reactivity relationships with machine learning predictions. But if you know anything about porphyrins, you will know that they are really difficult

to make, especially for an undergraduate student. Every step of the synthesis felt like mining gold, and I gave up after a few months of trying. I did not give up on the general idea of this proposal though and continued to work on a similar idea in graduate school, this time on cyanoarene photocatalysts. But history does repeat itself, and we ran into synthetic challenges again with cyanoarenes… Third time is a charm though, and we did find a class of catalysts that can be made quite easily. If you keep reading, you'll find out what that is in the next 350 pages.

It was mental exercises like this that made me more curious about all the new research on applying machine learning models in organic chemistry research, which eventually led me to Abby's group at Princeton in 2020, and now at UCLA since 2021. Abby has been incredibly supportive of my often-wacky research ideas since the very start and has given me freedom to explore anything that piqued my interest. I have worked on and gained so much knowledge in areas from nickel organometallics to reinforcement learning, which is not possible without her eclectic mix of scientific knowledge, receptiveness to new ideas and encouragement to do impactful research. In addition to being one of the most kind, compassionate and considerate people I know, Abby is also extraordinarily dedicated to her entire research group of students and staffs. I feel very fortunate to have been part of her research group and will greatly miss the intellectual environment where I can freely explore. Lastly, I am also thankful for her move to UCLA. Although I enjoyed my time at both Princeton and UCLA, life in Los Angeles is just a little bit more exciting than that in central New Jersey.

In addition to Abby, I also want to acknowledge Professors Ken Houk, Alex Spokoyny, and Anastassia Alexandrova for serving on my graduate committee and reviewing my thesis, as well as their guidance throughout my graduate school journey.

While I cannot name every single person, I must acknowledge the many people I overlapped with in the Doyle lab, who made graduate school an enjoyable five years of my life. Many people have mentored and inspired me scientifically when I first started graduate school: Sam, Jesus, Andrzej… Special shout-out to Sam, who has always inspired me with his passion and curiosity and has also become a close friend. Speaking of friends, I'm also thankful for the friendships I developed over the years, particularly the three (now) New Yorkers, Stavros, Makeda, (aforementioned) Sam. I will cherish the fond memories we made during our many trips in both Los Angeles and New York city, remember the hours-long conversations at Marea, and be grateful for the fact I can call you a close friend. I also want to acknowledge MJ, who not only worked alongside me (figuratively and literally) and helped me finish a years-long project but also became one of my closest friends and went on many adventures with me, including a thrilling trip to the Hoover dam. Also, the other New Jersey refugees: Will, Stephen, Wendy, Shivaani and Garrison, and will miss the camaraderie of our first year in LA setting up pool parties and an empty lab. I also had the privilege of working with many younger students: Daniel, Flora, Neyci, Braden, Mimi…All of them have gone on to become great scientists, and I can't wait to see the exciting research they will produce in the future. To the rest of the lab, many of whom I can call a good friend: Judah, Erin, Maddy, Winston, Paris, Alex… I will miss all the fun we had in and out of lab and wish you all my best in wherever the future takes you.

Lastly, I want to thank my parents and many other families and friends for their unconditional support over the years, even though most of them have absolutely no idea (nor do they care) what my Ph.D. work is about. Words cannot adequately describe my love and gratitude.

Financial support for my graduate research in the Doyle group has been generously provided by the following agencies and organizations: Bristol Myers Squibb (BMS), the BMS Graduate Fellowship in Synthetic Organic Chemistry, the Princeton Innovation Fund, the Princeton Catalysis Initiative, NIH-NIGMS (R35 GM126986), the Dreyfus Program for Machine Learning in the Chemical Sciences and Engineering, the United States National Science Foundation (NSF) Office of Advanced Cyberinfrastructure (OAC-2118201), the NSF under the CCI Center for Computer Assisted Synthesis (CHE-1925607, CHE-2202693), UCLA Dissertation Year Award and general departmental and university support.

Chapter 1 is a modified version of Żurański, A. M.*; Wang, J. Y.*; Shields, B. J.; Doyle, A. G. Auto-QChem: an Automated Workflow for the Generation and Storage of DFT Calculations for Organic Molecules. *React. Chem. Eng.* **2022**, *7*, 1276−1284. (*: equal contribution) DOI: 10.1039/D2RE00030J. A.M.Z., J.Y.W., B.J.S. are responsible for the software development and manuscript. A.G.D. is the principal investigator.

Chapter 2 is a modified version of 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.; Hao, B.; Valle, D. 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*, 1025–1033. DOI: 10.1038/s41586-024-07021-y. J.Y.W. and A.G.D. designed the overall research project. J.Y.W. designed and implemented optimization models and algorithms with inputs from J.M.S., J.L., J.E.T., B.J.S. and A.G.D.; J.M.S., B.J.S., J.L., J.E.T., J.Y.W. and A.G.D. designed and planned reaction scopes for the C–H arylation reaction, the amide coupling reaction and the phenol alkylation reaction. J.M.S., S.K.K., M.-J.T., D.L.G., M.P., D.N.P., B.H., D.D., S.D., A.F., G.G.Z.,

S.M. and J.P. carried out high-throughput experiments and authentic product synthesis for the three reactions. J.Y.W. wrote the paper with inputs from all authors.

Chapter 3 is a modified version of a manuscript Wang, J. Y.*; Fan, F.*; Ruos, M. E.*; Adao Gomes, L.; Lavin, M.; O'Connor, T. J.; Lopez, S. A.; Doyle, A. G. Diversification of acridinium photocatalysts: property tuning and reactivity in model reactions. *Tetrahedron Lett.* **2025** Submitted. (*: equal contribution) J.Y.W., F.F., M.E.R., L.A.G., M.L., T.J.O.C. are responsible for the experimental and computational studies. S.A.L. and A.G.D. are principal investigators.

Biographical sketch

Education

University of California, Los Angeles

- Ph.D. in Chemistry (anticipated: Winter 2025)
- GPA: 4.0

Columbia University in the City of New York

- B.A. in Biochemistry, B.A. in Computer Science (June 2020)
- GPA: 3.8

Professional and Academic Experiences

Graduate student researcher

- Advisor: Professor Abigail G. Doyle
- July 2021–present
- University of California, Los Angeles

Graduate teaching assistant

- Department of Chemistry, UCLA
  - CHEM 14A General Chemistry I (undergraduate level)
  - CHEM 14D Organic Chemistry II (undergraduate level)
  - CHEM 143A/243A Physical Organic Chemistry I (graduate level)

Graduate student researcher

- Advisor: Professor Abigail G. Doyle
- September 2020 – June 2021
- Princeton University

Undergraduate teaching assistant

- Department of Mathematics, Columbia University
  - MATH UN1101 Calculus I (undergraduate level)
  - MATH UN1102 Calculus II (undergraduate level)
  - MATH UN1201 Calculus III (undergraduate level)

Undergraduate student researcher

- Advisor: Professor Tomislav Rovis
- January 2017 – June 2020
- Columbia University

Selected Awards and Honors

- Undergraduate Research Fellowship, Société de Chimie Industrielle (2018)
- Merck Future Talent Program (Kenilworth, NJ) (2019)
- Dean's List, Columbia College (2016-2020)
- Bristol Myers Squibb Graduate Fellowship in Organic Chemistry (2023-2024)
- Genentech Graduate Student Symposium in Chemical Research Awardee (2024)

- UCLA Dr. Yuh Guo Pan Excellence in Research Award (2024)
- UCLA Dissertation Year Fellowship (2024-2025)

Publications

1. Ravetz, B. D.; **Wang, J. Y.**; Ruhl, K. E.; Rovis, T. Photoinduced Ligand-to-Metal Charge Transfer Enables Photocatalyst-Independent Light-Gated Activation of Co(II). *ACS Catal.* **2019**, *9*, 200−204. DOI: 10.1021/acscatal.8b04326.
2. **Wang, J. Y.***; Żuranński, A. M.*; Shields, B. J.; Doyle, A. G. Auto-QChem: an Automated Workflow for the Generation and Storage of DFT Calculations for Organic Molecules. *React. Chem. Eng.* **2022**, *7*, 1276−1284. (*: equal contribution) DOI: 10.1039/D2RE00030J.
3. Newman-Stonebraker, S. H.; **Wang, J. Y.**; Jeffrey, P. D.; Doyle, A. G. Structure−Reactivity Relationships of Buchwald-Type Phosphines in Nickel-Catalyzed Cross-Couplings. *J. Am. Chem. Soc.* **2022**, *144*, 19635–19648. DOI: 10.1021/jacs.2c09840.
4. **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.; Hao, B.; Valle, D. 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*, 1025–1033. DOI: 10.1038/s41586-024-07021-y.
5. **Wang, J. Y.**; Doyle, A. G. 'Bandit' algorithms help chemists to discover generally applicable conditions for reactions. *Nature* **2024**. DOI: 10.1038/d41586-024-00446-5.
6. Romer, N. P.; Min, D. S.; **Wang, J. Y.**; Walroth, R. C.; Mack, K. A.; Sirois, L. E.; Gosselin, F.; Zell, D.; Doyle, A. G.; Sigman, M. S. Data Science Guided Multi-objective Optimization of a Stereoconvergent Nickel-Catalyzed Reduction of Enol Tosylates to Access Tri-substituted Alkenes. *ACS Catal.* **2024**, *14*, 4699–4708. DOI: 10.1021/acscatal.4c00650.
7. **Wang, J. Y.**; Newman-Stonebraker, S. H. *CSD Communication*, **2024**, CCDC 2339820. DOI: 10.5517/ccdc.csd.cc2jjs2y.

Selected Presentations

- NSF Institute for Data Driven Dynamical Design (ID4), Princeton University (October 2022)
- The UK Catalysis Hub and Center for Rapid Online Analysis of Reactions (ROAR) online seminar (November 2022)
- NSF Center for Computer Assisted Synthesis (C-CAS), semi-annual meetings (2022-2024)
- Caltech, Pasadena, CA (March 2024)
- Genentech Graduate Student Symposium in Chemical Research, South San Francisco, CA (May 2024)
- Bristol Myers Squibb Graduate Student Symposium, Lawrenceville, NJ (May 2024)
- Columbia University, New York, NY (May 2024)
- American Chemical Society Division of Organic Chemistry Graduate Student Symposium, Charlottesville, VA (July 2024)

Chapter 1. Auto-QChem: an automated workflow for the generation and storage of DFT calculations for organic molecules

1.1 Introduction

Data-driven synthetic chemistry has witnessed rapid growth in recent years owing to advances in computing power, software, and algorithms, coupled with an increase in data availability from experiment and computation. The recent developments in machine learning, artificial intelligence and other data-driven approaches in organic chemistry has demonstrated their potentials as complementary and quantitative approaches for reactivity and selectivity predictions,[1,2] synthesis planning[3] and mechanistic studies.[4] Importantly, the application of machine learning models in organic chemistry requires effective representations of chemical structures.[5] Machine learning models trained with chemical descriptors often offer enhanced interpretability compared to molecular fingerprints and various learned representations.[6-10] In particular, features derived from density function theory (DFT) calculations are more closely associated with physical and chemical attributes of molecules, thus enabling improved mechanistic understandings. Therefore, these features serve as good candidates for building statistical and machine learning models. However, DFT calculations often require vast computing resources and proficiency in the operation of various software tools, which presents a significant barrier to experimental chemists. These problems are exacerbated by the number of DFT calculations required to featurize datasets that are sufficient for modern machine learning models. An automatic, high-throughput DFT calculation framework has the potential to accelerate the workflow and facilitate the computation of chemical descriptors by non-experts.

Many tools have been developed to automate high-throughput DFT calculations, such as AFLOW,[11] pymatgen,[12] MAST,[13] Atomate,[14] QMflows,[15] Nexus,[16] and AiiDA[17,18]. However, most of these tools are designed to facilitate material science research and are not well-suited for small organic molecules. Downstream applications in machine learning models also require a framework to extract and store a large amount of information from DFT calculation results. Databases containing DFT-calculated properties of materials and small molecules[19-22] have also been developed, usually with an underlying high-throughput workflow clearly defined. For example, the open-access VERDE materials database[22] provides numerous calculated photophysical properties of π-conjugated organic molecules. Such databases usually provide exceptional data access through APIs and web interfaces but end users often do not have direct access to the calculation pipelines. Beyond functionalities, the simplicity and ease of use for non-experts is also an important consideration. The objectives and limitations of current systems prompted us to implement a framework specifically designed for usage requirements of synthetic organic chemists.

1.2 Results and discussions

1.2.1 Overall design and implementation of Auto-QChem

A successful and robust high-throughput DFT calculation framework requires several key functionalities: (a) the ability to generate input files with user specifications for selected quantum chemistry software, (b) an interface with high performance computing (HPC) clusters for the submission and retrieval of jobs with error correction mechanisms, and (c) an analysis workflow to automatically extract information from calculation results. More specifically, we are interested in an end-to-end framework that can generate DFT-derived features directly from string

representations (such as SMILES[23]) of organic molecules in a high-throughput fashion, as well as provide storage and convenient access to processed data.

With these goals in mind, we developed Auto-QChem, an automated software package that streamlines DFT calculations for organic molecules. Starting from string representations of molecules, Auto-QChem performs initial conformational searches, manages DFT calculations on local HPC cluster, and facilitates cloud data storage and access via a web interface (*vide infra*).

The Auto-QChem framework is written in Python 3;[24] DFT calculations are performed with Gaussian 16;[25] the database is powered by MongoDB;[26] and the database web interface is written in Python Dash web framework.[27] Both the database and the web interface are hosted on a common AWS (Amazon Web Service) cloud server.[28] The code base is publicly hosted on a GitHub repository (https://github.com/doyle-lab-ucla/auto-qchem) together with its functional documentation and a series of user manuals showcasing example usage. The database web interface is publicly available at https://autoqchem.org/. The framework is modularized such that all operations can be performed from a single Jupyter notebook.[29]

1.2.2 Computational workflow

The workflow of Auto-QChem (**Fig. 1**) starts with a set of molecules represented as SMILES strings. Each SMILES string is first converted to a RDKit[30] molecule object. With a user-defined limit on the maximum number of conformers generated, Auto-QChem performs a conformational search for each molecule using one of the following configurable force field methods: (a) a genetic algorithm for stochastic conformer search implemented in OpenBabel,[31] (b) ETKDG distance geometry algorithm[32] implemented in RDKit. In practice, the RDKit implementation is more commonly used and set as the default option due to its robustness and RDKit's overall ease of installation and use compared to OpenBabel. An explicit option to sample

conformers of large molecules was also later added to cope with the instability when searching for molecules with many conformational possibilities using RDKit.

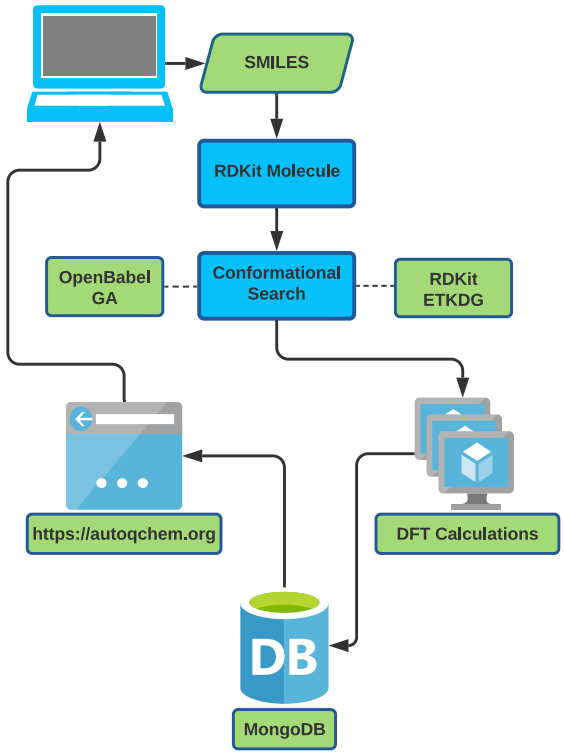


Fig. 1 Computational workflow of Auto-QChem.

By default, the following calculation workflow is applied: (a) geometry optimization; (b) frequency and thermochemical analysis, including vibrational frequency, molecular volume, natural population analysis (NPA) and nuclear magnetic resonance (NMR) calculations; and (c) a time-dependent DFT calculation for vertical excited state transitions. DFT calculation parameters such as functionals, basis sets and solvation models can be specified by the user. For each conformer, an input file with calculation specifications and atomic coordinates is generated and submitted to a Slurm scheduler[33] or SGE scheduler for DFT calculation with Gaussian on a local computer cluster (Slurm scheduler is used by Princeton's cluster, and SGE scheduler is used by UCLA's cluster). If a calculation runs out of time or memory, it can be resubmitted with a higher

time or resource limit using the last geometry checkpoint. Calculations with unspecified errors will be ignored.

Upon successful completion of the DFT calculations, duplicate conformers are removed from the ensemble with a configurable root-mean-square deviation (RMSD) threshold (0.35 Å by default). For each unique conformer, both molecule-wide and atomic numeric descriptors are extracted from Gaussian output files (the exact name and definitions of these descriptors are listed in the Auto-QChem GitHub repository https://github.com/doyle-lab-ucla/auto-qchem). These numeric descriptors and Gaussian output files are then uploaded to the Auto-QChem database.

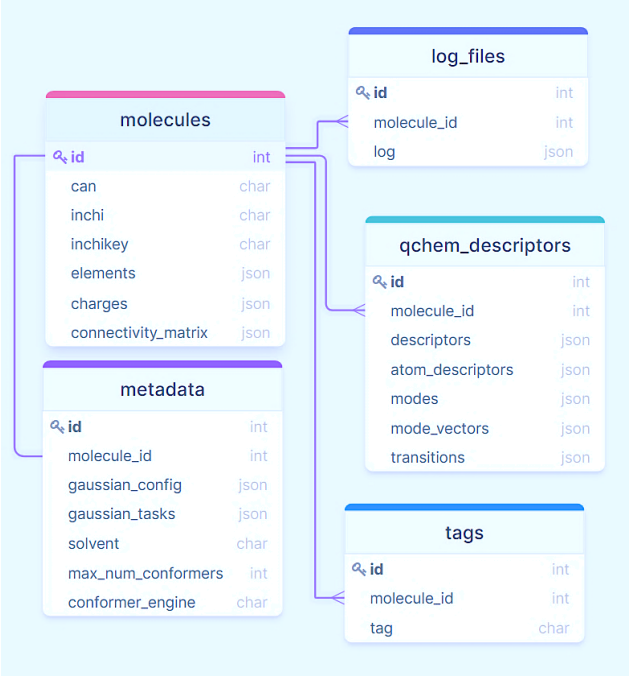


Fig. 2 Collection schema of Auto-QChem database.

1.2.3 Database

Data is organized into 5 collections (tables) to support queries and retrieval of the data (**Fig. 2**):

- **molecules**: master collection that stores information of individual molecules, such as string representations (SMILES, InChI, InChIKey), atomic coordinates, charges, and connectivity matrices.
- **metadata**: one-to-one auxiliary collection that stores the configuration of calculation for each molecule.
- **log_files**: many-to-one collection of raw output files of the calculations (one per conformer).
- **qchem_descriptors**: many-to-one collection of extracted numeric descriptors (one per conformer).
- **tags:** many-to-one collection that stores individual project name tags for easier retrieval and better organization of data.

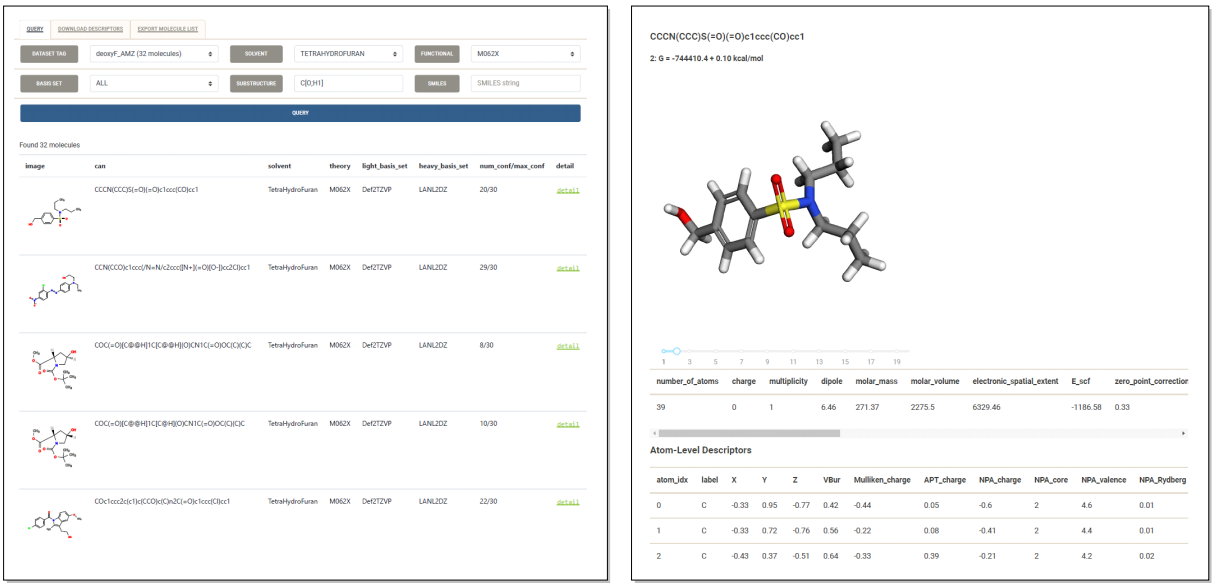


Fig. 3 Query view (left) and the molecule view (right) of the web interface.

Molecules are indexed such that a particular molecule along with its metadata must be unique, thus disallowing repeated calculations of one molecule with the same calculation configurations. However, calculations of the same molecule with a different configuration (e.g.,

different solvents, different basis sets) are allowed. Prior to generation of DFT jobs, Auto-QChem warns users if the requested calculation has already been performed and exists in the database. If a calculation of the same molecule with same computational configuration does exist, Auto-QChem will skip the calculation by default.

1.2.4 Queries and data retrieval

Data can be viewed and retrieved from the web interface hosted at https://autoqchem.org. There are two views available:

- **query view**: a view that allows for web queries of the database and downloads of descriptor sets. The query form contains the following filters: dataset name tags, solvents, functionals, basis sets, SMARTS substructure and SMILES strings.
- **molecule view**: an interactive display of the structures of all calculated conformers for one molecule, as well as tabulated numeric descriptors (an example is shown in **Fig. 3**).

After a successful query, a selection of numeric descriptor sets can be downloaded with the following configurations:

- **global**: molecular descriptors, such as HOMO/LUMO energy, dipole moment and molecular weight.
- **substructure atomic**: atomic descriptors from substructure searches. When a substructure is used for the query, atoms from substructure matches are identified in a consistent order and their atomic descriptors (e.g., NMR shifts, partial charges, buried volume) are extracted.

- **common core atomic**: atomic descriptors for the maximum common substructure within a dataset of molecules. The common core is determined using the FMCS (Find Maximum Common Substructure) algorithm[34] implemented in RDKit.[35]
- **min max atomic**: minimum and maximum values for each atomic descriptor over all atoms.
- **transitions**: top 10 excited state transitions ordered by oscillation strength and associated excited-state properties.

Because multiple conformers exist for the same molecule, for each molecular or atomic properties there will be values corresponding to individual conformers. By default, Boltzmann-weighted average of all conformers is calculated for each numeric descriptor and treated as feature vectors for each molecule. Different weighting options can be specified when exporting descriptors, for example, arithmetic average, the properties from the lowest (or highest) energy conformer only.

1.3 Applications of Auto-QChem

1.3.1 Substrate scope design in Ni/photoredox methodology development

As one example of AutoQChem's use in a synthetic chemistry context,[36] a team from our group, led by Dr. Stavros Kariofillis, developed a Ni/photoredox catalyzed alkylation reaction of aryl halides using acetals as alcohol-derived aliphatic radical sources.[37] To evaluate the generalizability of this methodology, the team set out to design a representative, diverse, and unbiased aryl bromide substrate scope through an unsupervised learning approach with DFT-derived featurization. An initial set of aryl bromides (molecular weight < 400) was generated through a Reaxys® search, which yielded around 290,000 candidates. After applying additional filters, such as commercial availability, spectroscopic data availability and functional group

compatibility, 2683 aryl bromides remain for DFT calculation. Some preliminary studies suggested that common featurization approaches, such as molecular fingerprints and cheminformatics descriptors, are often insufficient to represent electronic and steric features of substrates relevant to reactivity sites, necessitating the use of DFT-derived featurization.

With Auto-QChem, low-energy conformers were generated directly from SMILES strings for all aryl bromides. Gaussian jobs of generated conformers were then submitted to a connected HPC cluster. Successful calculations were logged and uploaded to the Auto-QChem database, along with 168 electronic and steric features (HOMO/LUMO energy, dipole moments, atomic volume, etc.) extracted from Gaussian log files. It is worth noting that, using Auto-QChem, DFT calculations of this size can be completed within a few days with minimal human intervention.

After feature preprocessing,[44] the remaining 95 features were used for hierarchical clustering to generate 15 clusters[43] and the molecules closest to the center of each cluster were chosen as the substrate scope (**Fig. 4b**). The final substrate scope includes a wide array of functional groups (such as esters, nitriles, chlorides), substitution patterns (mono-, di- and tri-substitution) and sterically varied substituents (ortho-, meta- and para-substitution). By comparing substrates from Ni/photoredox literature with the selected substrate scope, the team discovered that most aryl bromide substrates from literature examples are only present in a few clusters, while others (primarily clusters possessing multi-substituted aryl bromides) are significantly unexplored (**Fig. 4a**). This approach allows for study of chemical space coverage in the literature and identification of areas where high versus low yields are generally obtained. Unlike traditional substrate scopes in the literature, where selection usually happens in an arbitrary and subjective fashion, a machine learning-designed substrate scope with quantum mechanically informed descriptors is better suited for evaluating the generality of a reaction without human bias (**Fig. 4c**).

A systematic selection of substrates also enabled the training of regression models without selection bias to formulate predictive generalizations from DFT-derived features. It was discovered that electronegativity of the aryl bromide was highly correlated with yield. Using electronegativity as a predictive feature, a generalized additive model (GAM) was trained and validated with additional substrates. Similar models trained with literature substrates were less accurate and did not generalize well during validation This analysis demonstrated that a systematically designed substrate scope can effectively evaluate the generality of a reaction, as well as reveal reactivity trends for a larger population of substrates.

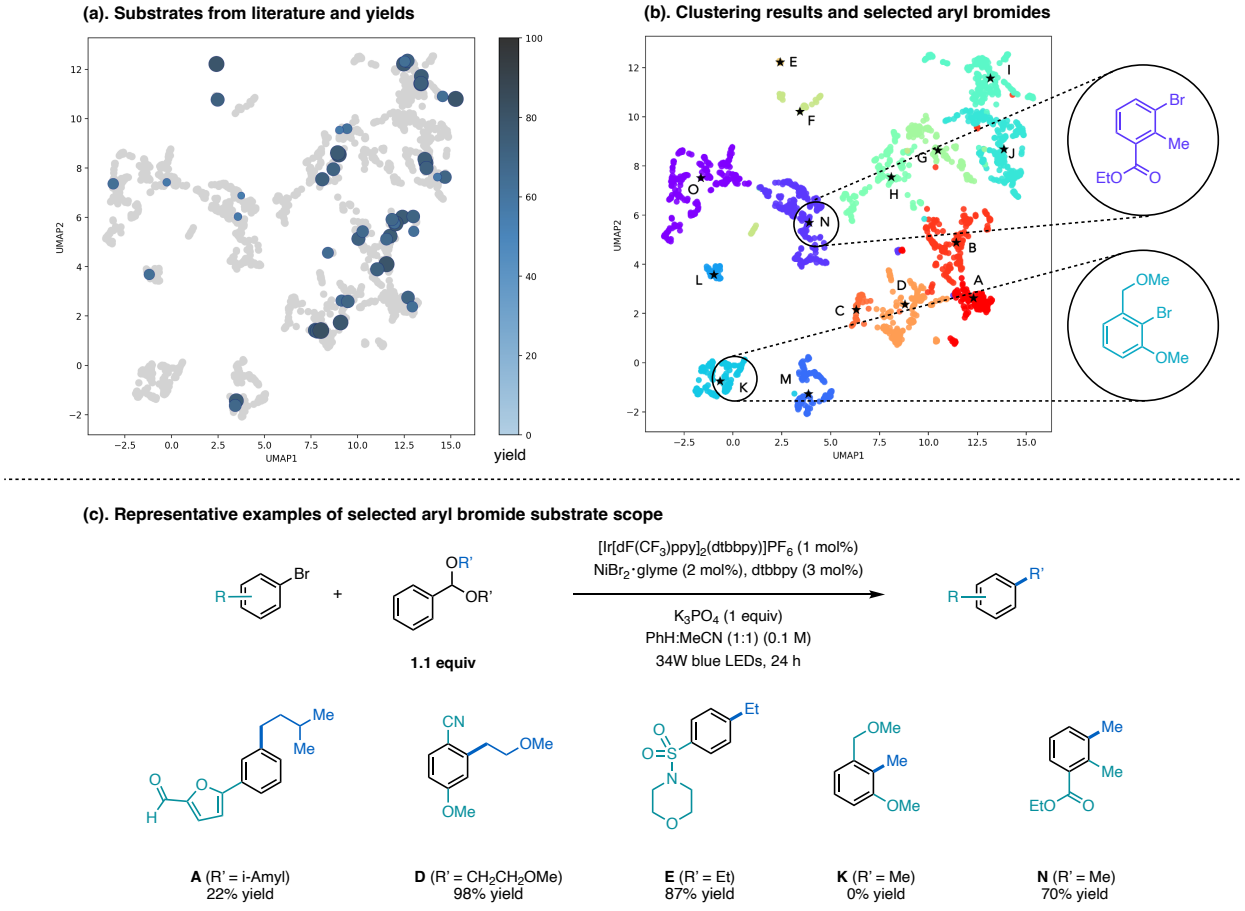


Fig. 4 Substrate scope design in a Ni/photoredox methodology development.

1.3.2 Ligand parametrization and enantioselectivity prediction in nickel catalysis

In another example, a team in our group, led by Dr. Will Lau, developed a Ni/photoredox-catalyzed enantioselective cross-electrophile coupling of aryl iodides and styrene oxides.[38] The optimal ligand, a chiral biimidazoline (BiIm) ligand, was discovered only after extensive screening of common chiral amine bidentate ligands. Bioxazoline (BiOx) ligands previously used in our asymmetric reductive coupling of aziridines[39] resulted in good enantioselectivity but low to moderate yield of the product. To understand the key features of BiIm ligands that affect reactivity and enantioselectivity of this reaction, we sought to use statistical modeling with physical and chemical descriptors from DFT calculations.

A total of 20 BiOx and 9 BiIm ligands was selected to be studied. The team collected enantioselectivity data under standard reaction condition with a model substrate (**Fig. 5a**). Under the hypothesis that ligand environments will likely affect the computed features, DFT calculations were performed for all the ligands under three different environments: free ligand, ligand bound to a tetrahedral nickel difluoride complexes and ligand bound to a square planar nickel oxidative addition complex (**Fig. 5b**). As a potential limitation, Auto-QChem (and most conformer-generating software) cannot reliably generate conformers for transition metal complexes,[45] especially for group 10 metals like nickel. As a result, all the initial conformers for nickel-bound ligand were manually generated and submitted for DFT calculation. Auto-QChem's descriptor extraction module was still used to extract electronic and atomic volume features from output files. Importantly, a multivariate linear regression analysis showed that, although they give a worse fit for the data, features derived from free ligands were sufficient for a descriptive linear regression model. From the regression model, NBO_{C4}, NBO_{N1} and polarizability independently affect $\Delta\Delta G^{\ddagger}$, suggesting that electronic, rather than steric attributes of BiIm ligands govern the enantioselectivity

of this reaction (**Fig. 5c**). This study demonstrated how insights from regression modeling with DFT-derived features can afford a mechanistic probe of complex catalytic reactions.

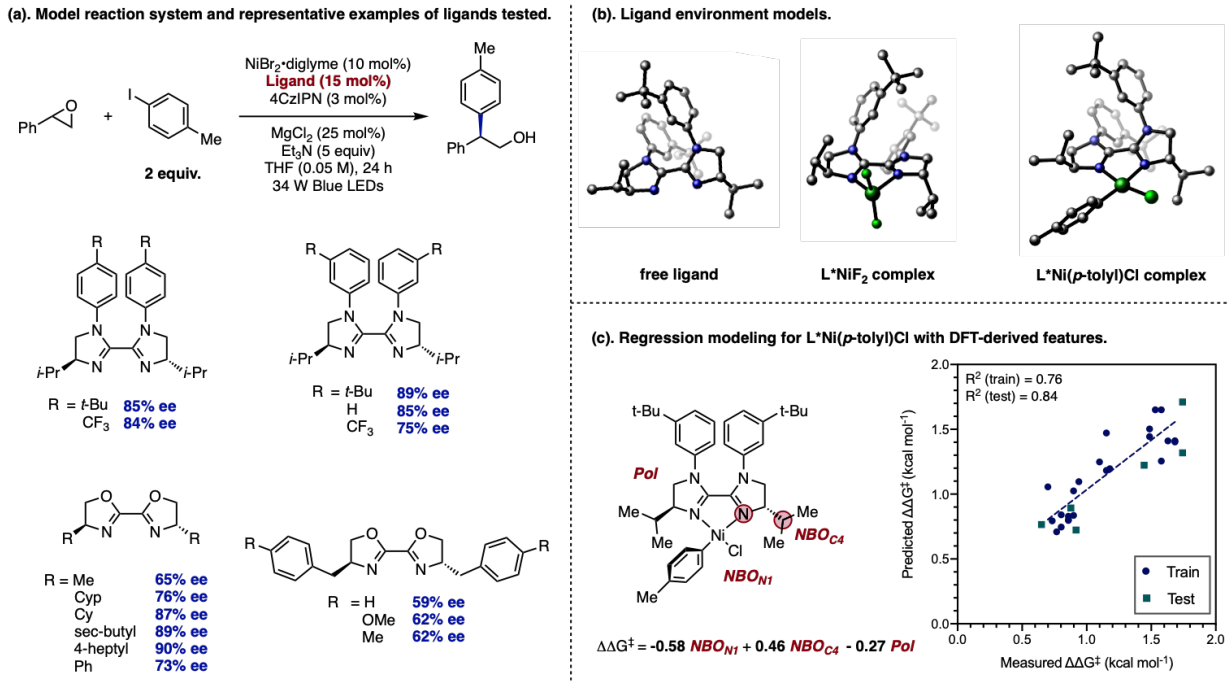


Fig. 5 Ligand parametrization and enantioselectivity prediction in nickel catalysis.

1.3.3 Reaction condition optimization via Bayesian optimization

The optimization of reaction conditions is often tedious and time-consuming in methodology development campaigns. In the pursuit of conditions that provide the highest yield for reactions of interest, chemists often rely on empirical knowledge and qualitative understandings of the current optimization progress to design the next experiment. Typical approaches include the adoption of known conditions from literature, design of experiments (DoE), or more time- and resource-intensive methods such as high-throughput experimentations (HTE) and in-depth mechanistic studies. For individual reaction components, the lack of quantitative assessment of their effects on reaction yield usually requires running many combinations of the conditions, which in turn limits the size of chemical space explored during optimization.

In a recent study by our group and BMS (Bristol Myers Squibb), led by Dr. Benjamin Shields,[40] we demonstrated the application of Bayesian optimization, a sequential design algorithm for global optimization of black-box functions, in efficient reaction condition optimization. A software framework, EDBO (Experimental Design via Bayesian Optimization), was developed, where a Bayesian optimization algorithm was integrated into real-time laboratory experimentations (**Fig. 6**). After a reaction space is defined, initial experiments are selected via clustering or other sampling approaches. Chemists run the suggested reactions in lab, analyze the results when reactions finish and input reaction yield into the system. Bayesian optimization algorithms use new results to update the prior and form a new posterior distribution over the objective function. An acquisition function is constructed with the new posterior to determine new query points (new reactions to run). This optimization loop is repeated until the desired yield or resource limit is reached.

During the development of the Bayesian optimization framework, its performance was evaluated by comparing simulation results to human decision-making benchmarks obtained with large HTE reaction datasets. Bayesian optimization requires each reaction component to be translated into a suitable numeric representation. The effects of different featurizations (DFT-derived features, molecular descriptors such as Mordred,[41] and one-hot encoding) were tested on optimization convergence. DFT calculations for hundreds of molecules contained in these reaction datasets were completed with an early version of Auto-QChem,[42] which greatly simplified the workflow. Compared to other featurizations, DFT features offer more efficient learning curves and consistent performance in terms of worst-case loss. Using DFT-derived features, it was showed that Bayesian optimization outperformed human decision-making baselines established by expert

chemists. The performance benefits obtained with DFT-derived features further validate the necessity of high-throughput DFT featurization frameworks like Auto-QChem.

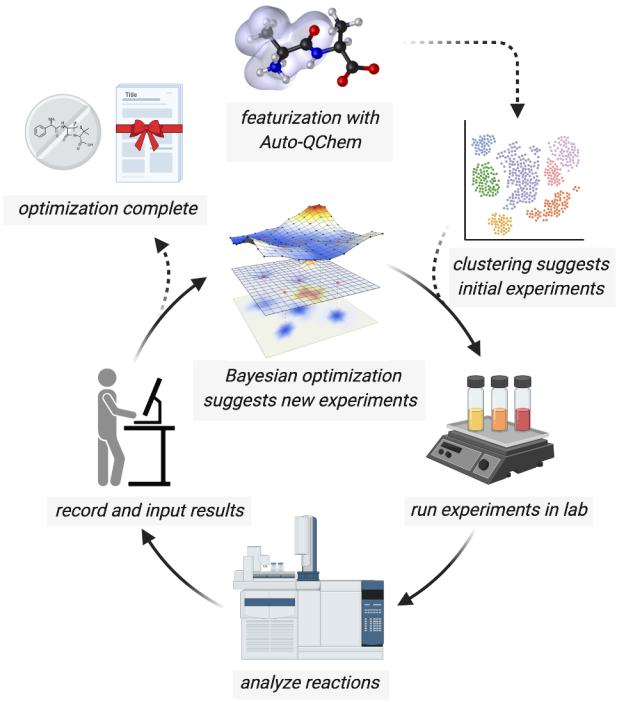


Fig. 6 The optimization workflow of EDBO.

1.3.4 Other applications

In addition to the three applications of Auto-QChem described above, there has been some new developments since the publication (2022) in using Auto-QChem and DFT descriptors to model catalytic reactivities. For example, in a study from our group led by Dr. Wendy Williams on nickel-catalyzed cross-electrophile coupling of aziridines and aryl iodides,[46] Auto-QChem was used to featurize thousands of aryl and heteroaryl iodide substrates. A diverse substrate scope was similarly constructed as discussed in **1.3.1**, and a simple yet accurate linear regression model was fit to model the substrate effect on catalytic reactivities. HOMO (Highest Occupied Molecular Orbital) energy of the aryl iodide and $\%V_{bur}$ (percent buried volume) of the iodine atom were the two features required for the linear regression model, highlighting both the electronic and steric

requirements of this reaction. Another recent publication by our group, led by Dr. Shivaani Gandhi, models catalytic reactivities of a copper-catalyzed Chan-Lam reaction with sulfonamides and aryl boronic acids.[47] Auto-QChem was again applied to featurize sulfonamide substrates and construct a diverse substrate scope. The Chan-Lam reaction was evaluated with multiple substrates and conditions combinations via HTE. DFT features derived with Auto-QChem and labels from the experimental dataset were used to build a neural networks model that can accurately predict the reaction yields of Chan-Lam reaction, allowing the screening of reaction conditions *in silico*.

Beyond works from our group, other research groups have also started to adopt Auto-QChem into reaction modeling. For example, Auto-QChem derived DFT-descriptors were benchmarked against other featurization methods in a nickel-catalyzed cross-coupling reaction dataset,[48] as well as a Suzuki-Miyaura cross-coupling reaction dataset.[49] In some cases, Auto-QChem derived DFT features were observed to enhance the predictive model performance when coupled with other featurization methods.

1.4 Conclusions and outlooks

In conclusion, we developed Auto-QChem, an automated, high-throughput and end-to-end DFT calculation workflow. Designed to facilitate the increasing applications of machine learning models in organic chemistry, Auto-QChem generates DFT-derived molecular and atomic features starting from simple string representations of the molecules. After initial conformational searches, each conformer is submitted to a local computer cluster for DFT calculations with user-specified configurations. Cluster jobs are managed directly through Auto-QChem with error-correcting mechanisms. Successful calculation results and extracted DFT features are then uploaded to a database. A web interface (https://autoqchem.org) is also available for convenient data access. We also present three distinct studies from our group where Auto-QChem was used to featurize a large

set of molecules and greatly simplified the workflow in reaction modeling. Some more recent examples that highlight the application of Auto-QChem after the publication of this work in 2022, both from our group and other research groups, were also discussed.

We would also like to highlight some limitations of Auto-QChem at the present stage and outline some future directions. First, as mentioned in **1.3.2**, Auto-QChem lacks the ability to generate accurate conformers for transition metal complexes and molecules with non-canonical bonds. Such problems are not unique to Auto-QChem as we leverage external programs such as RDKit to handle conformational searches. We are actively seeking improvement and experiment with other conformational search software that can alleviate such problems.

Another important functionality of Auto-QChem is the ability to manage jobs on HPC clusters. Currently, Auto-QChem only supports Slurm scheduler and SGE-type scheduler. More specifically, Auto-QChem supports job syntax for Princeton University's computer cluster and UCLA's Hoffman2 cluster. Due to the lack of access to other computational clusters, and the fact that each cluster usually requires a specific job syntax for allocating resources and running jobs, integration of other cluster job schedulers will require some modifications to existing code. It is possible for experienced users to modify Auto-QChem in the current stage to work with other computational clusters. For example, some researchers from University of Michigan and Caltech have successfully done so.

As a software package, Auto-QChem requires regular maintenance and troubleshooting. All known bugs that have not been fixed are recorded on the GitHub repository as issues waiting to be worked on. The website, database application and AWS server also need regular maintenance to ensure a reliable user experience, especially on scale. We are already observing operational difficulties with increasingly large datasets, which will require application update and server

migration. Efforts have also been made to improve Auto-QChem as a Python package with callable, programmable methods capable of more complicated operations for experienced users.

For the major functionality updates, we will continue to include external packages and automate the calculation of additional electronic and steric features that are not currently supported by Auto-QChem, such as Hirshfeld charges and Sterimol parameters. Significant work might be necessary to retroactively re-calculate these parameters for molecules that are already existing in the database. Barring any quality control issues, we also intend to invite other users to upload data to Auto-QChem. Like previously mentioned, modified versions of Auto-QChem have been hosted in other research institutions, and some external users have already been contributing to the database. With enough data on hand, we would also like to train machine learning models with existing data to predict DFT-level features for similar molecules,[50] which will address the speed bottleneck of DFT calculations in our workflow.

1.5 References

1. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**,1134–1140.
2. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.
3. M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.
4. S. Zhao, T. Gensch, B. Murray, Z. L. Niemeyer, M. S. Sigman and M. R. Biscoe, *Science*, 2018, **362**, 670–674.
5. L. David, A. Thakkar, R. Mercado and O. Engkvist, *J. Cheminform.*, 2020, **12**, 56.
6. S. Jaeger, S. Fulle and S. Turk, *J. Chem. Inf. Model*, 2018, **58**, 27–35.

7. S. Kearnes, K. McCloskey, M. Berndl, V. Pande and P. Riley, *J. Comput.-Aided Mol. Des.*, 2016, **30**, 595–608.
8. W. Coley, R. Barzilay, W. H. Green, T. S. Jaakkola and K. F. Jensen, *J. Chem. Inf. Model*, 2017, **57**, 1757–1772.
9. R. D. Hull, S. B. Singh, R. B. Nachbar, R. P. Sheridan, S. K. Kearsley and E. M. Fluder, *J. Med. Chem.*, 2001, **44**, 1177–1184.
10. M. Olivecrona, T. Blaschke, O. Engkvist and H. Chen, *J. Cheminform.*, 2017, **9**, 48.
11. S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko and D. Morgan, *Comput. Mater. Sci.*, 2012, **58**, 218–226.
12. S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Comput. Mater. Sci.*, 2013, **68**, 314–319.
13. T. Mayeshiba, H. Wu, T. Angsten, A. Kaczmarowski, Z. Song, G. Jenness, W. Xie and D. Morgan, *Comput. Mater. Sci.*, 2017, **126**, 90–102.
14. K. Mathew, J. H. Montoya, A. Faghaninia, S. Dwarakanath, M. Aykol, H. Tang, I. Chu, T. Smidt, B. Bocklund, M. Horton, J. Dagdelen, B. Wood, Z.-K. Liu, J. Neaton, S. P. Ong, K. Persson and A. Jain, *Comput. Mater. Sci.*, 2017, **139**, 140–152.
15. F. Zapata, L. Ridder, J. Hidding, C. R. Jacob, I. Infante and L. Visscher, *J. Chem. Inf. Model*, 2019, **59**, 3191–3197.
16. J. T. Krogel, *Comput. Phys. Commun.*, 2016, **198**, 154–168.
17. S. P. Huber, S. Zoupanos, M. Uhrin, L. Talirz, L. Kahle, R. Häuselmann, D. Gresch, T. Müller, A. V. Yakutovich, C. W. Andersen, F. F. Ramirez, C. S. Adorf, F. Gargiulo, S. Kumbhar, E.

Passaro, C. Johnston, A. Merkys, A. Cepellotti, N. Mounet, N. Marzari, B. Kozinsky and G. Pizzi, *Sci. Data*, 2020, **7**, 300.

18. M. Uhrin, S. P. Huber, J. Yu, N. Marzari and G. Pizzi, *Comput. Mater. Sci.*, 2021, **187**, 110086.
19. S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl and C. Wolverton, *npj. Comput. Mater.*, 2015, **1**, 15010.
20. K. Choudhary, K. F. Garrity, A. C. E. Reid, B. DeCost, A. J. Biacchi, A. R. H. Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, A. Davydov, J. Jiang, R. Pachter, G. Cheon, E. Reed, A. Agrawal, X. Qian, V. Sharma, H. Zhuang, S. V. Kalinin, B. G. Sumpter, G. Pilania, P. Acar, S. Mandal, K. Haule, D. Vanderbilt, K. Rabe and F. Tavazza, *npj. Comput. Mater.*, 2020, **6**, 173.
21. G. A. Smith, D. Altarawy, L. A. Burns, M. Welborn, L. N. Naden, L. Ward, S. Ellis, B. P. Pritchard and T. D. Crawford, *WIREs Comput. Mol. Sci*., 2021, **11**, e1491.
22. B. G. Abreha, S. Agarwal, I. Foster, B. Blaiszik and S. A. Lopez, *J. Phys. Chem. Lett.*, 2019, **10**, 6835–6841.
23. Weininger, *J. Chem. Inf. Model*, 1988, **28**, 31–36.
24. Python Software Foundation, https://www.python.org, (accessed January 2022).
25. Gaussian 16, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E.

Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.

26. MongoDB, https://www.mongodb.com, (accessed January 2022).
27. Dash Python User Guide, https://dash.plotly.com, (accessed January 2022).
28. Amazon Web Services, https://aws.amazon.com, (accessed January 2022).
29. T. Kluyver, in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, ed. F. Loizides and B. Schmidt, IOS Press, Amsterdam, 2016, pp. 87-90.
30. RDKit: Open-source cheminformatics, https://www.rdkit.org/, (accessed January 2022).
31. N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminform.*, 2011, **3**, 33.
32. S. Riniker and G. A. Landrum, *J. Chem. Inf. Model*, 2015, **55**, 2562–2574.
33. Slurm workload manager, https://slurm.schedmd.com, (accessed January 2022).
34. Dalke and J. Hastings, *J. Cheminform.*, 2013, **5**, O6.
35. rdkit.Chem.fmcs.fmcs module, https://www.rdkit.org/docs/source/rdkit.Chem.fmcs.fmcs.html, (accessed January 2022).
36. S. K. Kariofillis, S. Jiang, A. M. Żurański, S. S. Gandhi, J. I. Martinez Alvarado and A. G. Doyle, *J. Am. Chem. Soc.,* 2021*,* DOI: 10.1021/jacs.1c12203.
37. S. K. Kariofillis, B. J. Shields, M. A. Tekle-Smith, M. J. Zacuto and A. G. Doyle, *J. Am. Chem. Soc.*, 2020, **142**, 7683–7689.

38. S. H. Lau, M. A. Borden, T. J. Steiman, L. S. Wang, M. Parasram and A. G. Doyle, *J. Am. Chem. Soc.*, 2021, **143**, 15873–15881.

39. B. P. Woods, M. Orlandi, C.-Y. Huang, M. S. Sigman and A. G. Doyle, *J. Am. Chem. Soc.*, 2017, **139**, 5688–5691.

40. B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, *Nature*, 2021, **590**, 89–96.

41. H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, *J. Cheminform.*, 2018, **10**, 4.

42. Auto-QChem, https://github.com/b-shields/auto-QChem, (accessed January 2022).

43. 15 is the number of clusters at which the maximum and stable Silhouette score was reached.

44. Preprocessing includes scaling, outlier removal, removal of features with low variance and correlation analysis.

45. Software that specifically focuses on conformer generation for transition-metal complexes do exist, such as molSimplify: E. I. Ioannidis, T. Z. H. Gani and H. J. Kulik, *J. Comput. Chem.*, 2016, **37**, 2106–2117.

46. W. L. Williams, N. E. Gutiérrez-Valencia and A. G. Doyle, *J. Am. Chem. Soc.*, 2023, **145**, 24175–24183.

47. S. S. Gandhi, G. Z. Brown, S. Aikonen, J. S. Compton, P. Neves, J. I. M. Alvarado, I. I. Strambeanu, K. A. Leonard and A. G. Doyle, *ACS Catal.*, 2025, 2292–2304.

48. J. Schleinitz, M. Langevin, Y. Smail, B. Wehnert, L. Grimaud and R. Vuilleumier, *J. Am. Chem. Soc.*, 2022, **144**, 14722–14730.

49. P. Raghavan, A. J. Rago, P. Verma, M. M. Hassan, G. M. Goshu, A. W. Dombrowski, A. Pandey, C. W. Coley and Y. Wang, *J. Am. Chem. Soc.*, 2024, **146**, 15070–15084.

50. B. C. Haas, M. A. Hardy, S. S. S. V, K. Adams, C. W. Coley, R. S. Paton and M. S. Sigman, *Digital Discovery*, 2024, **4**, 222–233.

Chapter 2. Identifying general reaction conditions via bandit optimization

2.1 Introduction

Chemists have long sought robust synthetic methods that can be applied to a wide variety of substrates.[1–3] However, methodologies are generally developed and optimized with only one or a few model substrates to circumvent synthetic and analytical constraints. These "optimized" conditions are subsequently applied to a substrate scope, usually with higher yielding substrates preferentially reported. Reaction conditions optimized for a single substrate are not guaranteed to be applicable to other molecules with distinct structural features. Despite the increased efficiency of reaction optimization enabled by automated reaction systems[4–10] and optimization algorithms,[11–20] this phenomenon still significantly hampers the adoption of newly developed methodologies in synthetic chemistry.[21,22] Further optimization for different target substrates is typically required, and pharmaceutically relevant molecules with high structural complexity might not even be compatible with existing conditions at all.[23] Most work to date has focused on retroactively evaluating the general applicability of developed methodologies via substrate scope design. One approach is to cluster commercial substrates into groups with unsupervised learning models, from which a representative substrate scope can be constructed by sampling from all groups.[24] Another approach involves expert-designed scopes intended to test substrate compatibility relevant in pharmaceutical synthesis.[25] Additive screening is also a prevalent strategy to assess condition applicability.[26,27] Identifying incompatible additives with problematic structural features or ones that facilitate the desired transformation can provide insights that can enhance the general applicability of a reaction method.

Nevertheless, *post hoc* analyses of applicability do not change the reaction conditions derived from prior optimization. *De novo* optimization processes that can directly yield generally

applicable conditions are highly sought. Recent advances in asymmetric catalysis have started to address this problem, where chiral ligands/catalysts that enable highly stereoselective transformations for a broad range of substrates are identified through multi-substrate screening combined with mechanistic studies and data science approaches during methodology development.[28–31] However, unlike asymmetric catalysis where catalyst/ligand effects predominantly affect stereoselectivity, optimization of reactivity is a multi-dimensional problem that involves both chemical (e.g., catalysts, bases, solvents) and physical components (e.g., temperature, wavelengths, voltage, time). External factors, such as reaction vessels and set-up procedures, can also have significant effects on reactivity. Despite advancement in high-throughput experimentation (HTE) technologies, exhaustive examination of all aspects of a chemical reaction remains difficult and expensive to carry out. Such a problem is exacerbated by the simultaneous survey of a sizable scope of diverse substrates necessary to correctly identify conditions that are broadly applicable, which can result in lengthy authentic product synthesis campaigns and appreciable analytical challenges. Judicious selection of experiments is therefore imperative to efficiently explore the reaction space during optimization. A notable recent example from Burke, Aspuru-Guzik, and Grzybowski aimed to find more general sets of conditions for a Suzuki-Miyaura cross-coupling reaction with aryl halides and aryl *N*-methyliminodiacetic acid (MIDA) boronates.[32] Bayesian optimization was used to select experiments that were carried out by a robotic system, which greatly alleviates synthetic challenges. After the initial benchmarking and down-selection of reaction conditions prior to optimization, exploration of over 50% of the reaction space identified conditions more general than a previously published standard condition. This important advance notwithstanding, a universal reaction optimization model targeting general

applicability, especially one with an efficient experiment selection strategy that can also be easily incorporated into the workflow of bench chemists, has not yet been realized.

In this study, we show that reinforcement learning (RL) models can effectively guide chemists to the most generally applicable conditions for a given substrate scope without prior experimental data on the reaction system. We designed a discrete optimization framework with experiment selection strategies that target condition generality, as quantified by average reactivity (albeit other distribution metrics can be used). Through performance benchmarking on four existing reaction datasets, we demonstrate that the implemented reinforcement learning model and its underlying algorithms reach high accuracies for identifying optimal general conditions in all cases, while being adaptable, scalable, and data efficient. To further substantiate the optimization framework, we also validated the learning model on three unseen chemical transformations.

2.2 Results and discussions

2.2.1 Model design and development

The multi-armed bandit problem[33] is a RL problem that resembles many characteristics of the generality optimization problem in chemistry. In the classic formulation, a casino player is presented with a series of slot machines, each with a fixed but different reward distribution that is also initially unknown. With a limited budget, the objective of the player is to maximize overall winnings by recognizing and playing the slot machine with better payouts. Reconciling the classic exploration-exploitation tradeoff, the player must efficiently allocate limited resources to balance the exploration of rarely played machines and the exploitation of current best options. In a reaction optimization campaign, chemists often need to choose from many options for reaction conditions to maximize certain objectives with limited initial knowledge of how they will perform on a wide

range of substrates (**Fig. 7a**). Finite experimental resources must be efficiently allocated to each reaction condition in consideration of a similar exploration-exploitation tradeoff: current best conditions derived from empirical knowledge are usually exploited, while new conditions are explored in hopes of discovering novel and more effective methods. The similar characteristics of both problems prompted us to adapt solutions to the multi-armed bandit problem (often called bandit optimization algorithms) for a generality optimization problem in chemistry.

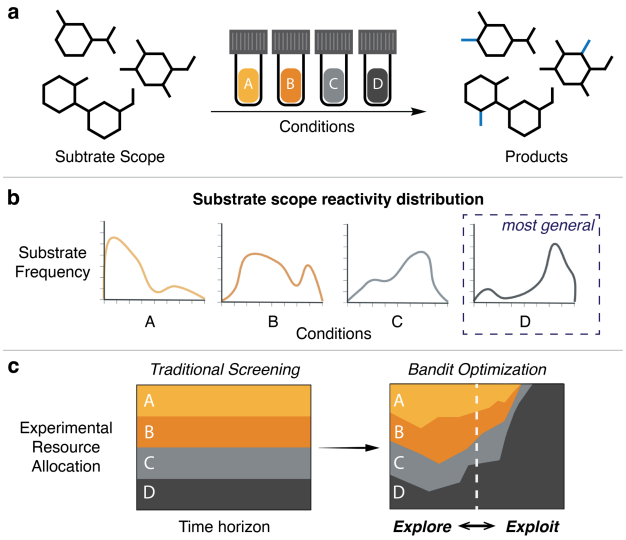


Fig. 7 Optimize for most general condition with bandit optimization.

We designed an optimization framework where reaction conditions are treated as options (arms) to explore, with substrates being the underlying population for each option. Using reaction yield as an example of an optimization objective, the same substrate scope is expected to exhibit different reactivity behaviors under different conditions, resulting in unique reward distributions for each arm (**Fig. 7b**). The treatment of condition variables as discrete arms allows for versatile interpretations of conditions. Unlike design of experiments (DOE) or Bayesian optimization where a high-dimensional search space needs to be defined to cover all combinations of condition components,[34] our approach allows arms to be defined to cover one condition dimension (e.g.,

solvent) or many dimensions (e.g., catalyst/ligand/base/solvent combinations). In other words, it is possible to accommodate different precisions required in a single optimization campaign, ranging from comparing full sets of conditions established in literature to fine-tuning a specific reaction component. This approach bypasses the need to re-define high-dimensional spaces when pivoting objectives during optimization and recycles existing data by representing reaction results covered by one arm as samples from that distribution. Incorporating substrates into a distribution also means no explicit search space needs to be defined, and the algorithm can adjust its estimation of each condition's distribution by continuing to sample that condition. This feature allows for both the elimination of ineffective arms and the expansion of substrate scope on the fly during optimization. The latter is especially important in application, as the generality of a reaction condition is highly dependent on the scope it is applied to.

Leveraging algorithms formulated for multi-armed bandit problems, we implemented the optimization framework in Python specifically aimed at identifying generally applicable reaction conditions. Fundamentally, bandit algorithms balance exploration and exploitation of conditions and efficiently allocate experimental resources to conditions that exhibit higher reactivity (**Fig. 7c**). Our implementation centers around a reaction scope object that can create substrate scopes with possible conditions, interface with bandit algorithms, propose and record experimental results, predict yields for unrun reactions, and recommend general conditions (**Fig. 8**). We implemented bandit algorithms for both binary rewards (e.g., reactivity thresholds) and continuous rewards (e.g., numeric reaction yields). Bandit algorithms that optimize for continuous rewards are not commonly studied compared to those designed for binary rewards, and their behaviors in real-world datasets can often deviate from theoretical performance analyses. To address these limitations, we adapted existing algorithms, such as Thompson sampling[35] and Bayes UCB (Upper

Confidence Bound) algorithms,[36] with gaussian priors to accommodate continuous rewards. Effective algorithm classes were identified through extensive benchmarking with synthetic data, as well as empirical modifications and hyperparameter selections that are beneficial to algorithm performance. Multiple approaches to support batch proposing and updating were also implemented to allow parallel experimentation in practice (see Section 2.4 for details on algorithm development). Unlike optimization frameworks that involve costly fitting of Gaussian processes and neural networks as surrogate models,[37] our framework is also lightweight and computationally efficient, written almost in pure Python with minimal dependencies. This advantage not only enhances software performance in a production environment but also allows us to extensively simulate the learning model with existing datasets to statistically evaluate its effectiveness.

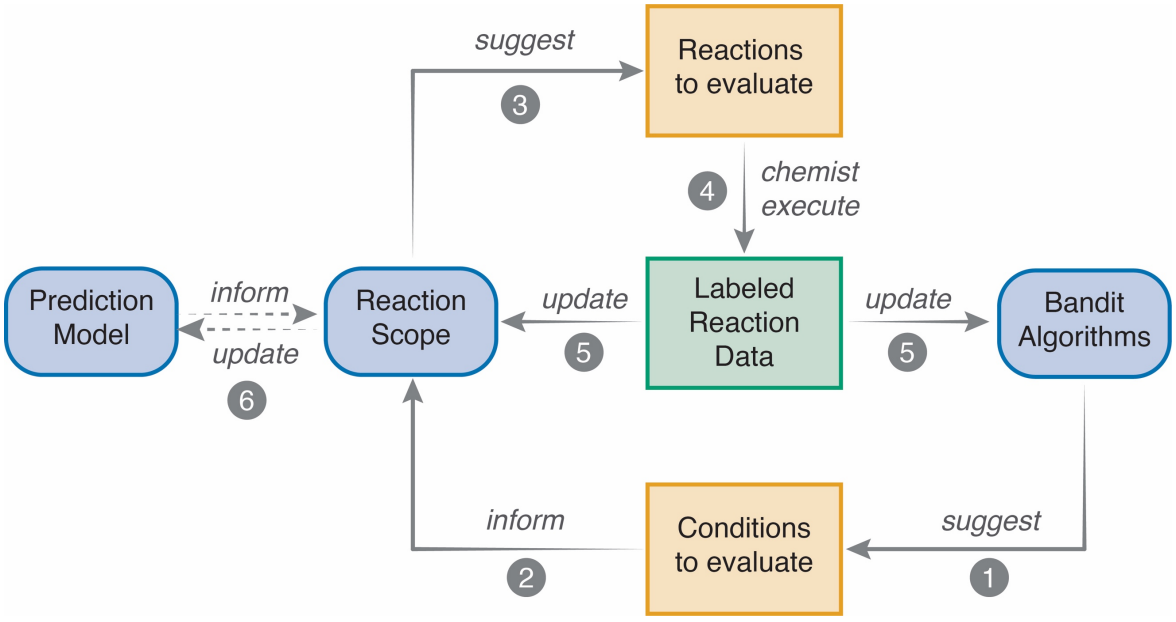


Fig. 8 Model architecture and workflow of bandit algorithms during reaction optimization.

2.2.2 Performance testing with chemistry reaction datasets

We simulated the optimization model on three previously published real-world chemistry reaction datasets consisting of a variety of conditions applied to a broad scope of substrates: a nickel-catalyzed borylation dataset previously investigated by Bristol Myers Squibb (BMS),[38] a deoxyfluorination dataset from the Doyle group,[39] and a Buchwald-Hartwig C–N cross-coupling dataset,[40] all with the aim of finding the most general conditions (**Fig. 9a**). The optimization targets range from ligand (borylation), base–sulfonyl fluoride combination (deoxyfluorination) to full sets of catalytic conditions (C–N cross-coupling). In addition to numeric reaction yields (deoxyfluorination), we used other reactivity metrics, including pass/fail responses (borylation) and normalized UPLC-MS (Ultra-Performance Liquid Chromatography-Mass Spectrometry) ion counts (C–N cross-coupling) to represent scenarios when calibrated reaction yields are not available. For every dataset, the most general conditions are first determined through analyses of reaction yield distributions (**Fig. 9c**; see Section 2.4.8 for detailed yield analyses on all datasets). Optimization runs were then simulated by iteratively allowing suitable algorithms to propose experiments and providing algorithms with actual experimental results. After each round, the learning model updated its beliefs for the reaction scope, and this process was continued until a specified number of experiments was reached. This simulation process was repeated many times (e.g., 500) and the top-*n* accuracy was calculated as the relative frequency of the learning model correctly identifying top-*n* conditions across all simulations.

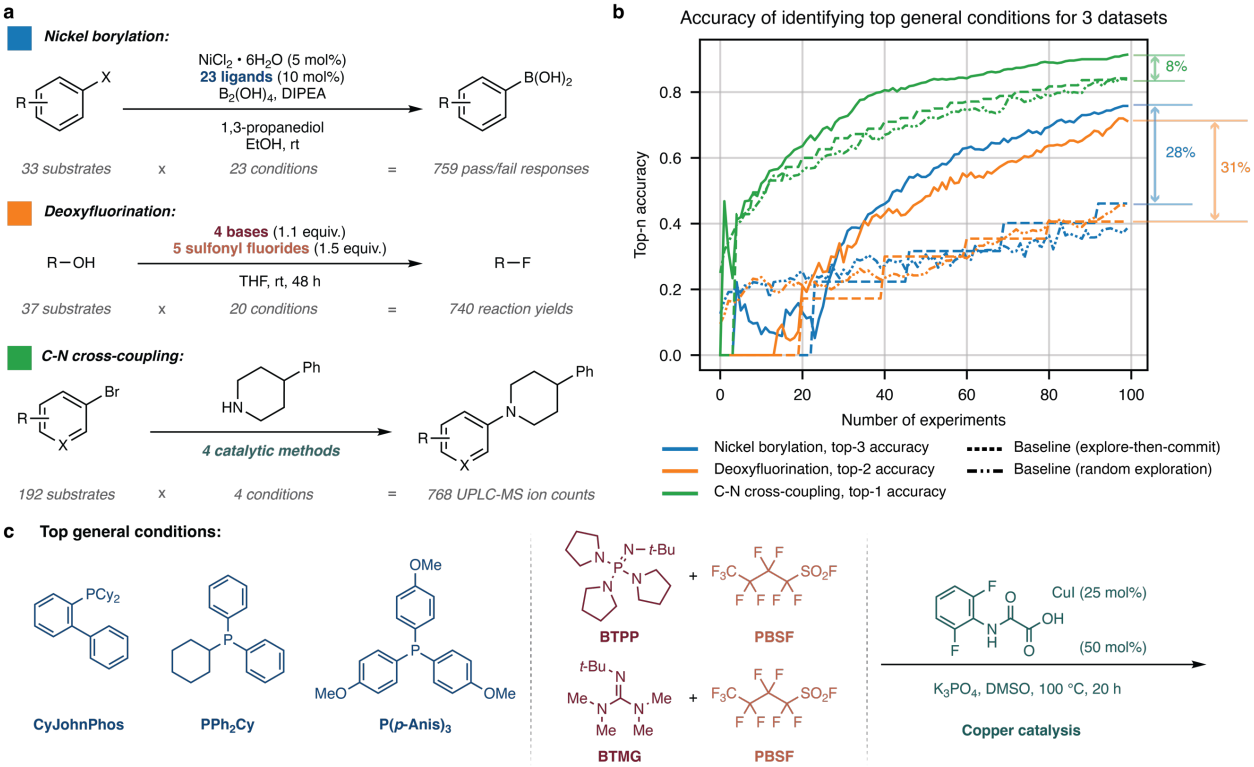


Fig. 9 Testing the bandit optimization framework on three datasets with different objectives and condition complexities.

To confirm that meaningful learning took place with the developed model, we established two baselines for comparison for each dataset. One of them is the pure exploration baseline where conditions are randomly selected for evaluation, the accuracy of which is equivalent to a random guess. The other baseline strategy, explore-then-commit (ETC), tries each condition a fixed number of times during the exploration stage and then continuously exploits the best empirical option. ETC is similar to how chemists traditionally optimize reactions, where screens are conducted for one reaction dimension with other parameters fixed and the best option is then exploited. Compared to the two established baselines, our model achieved good accuracies within 100 experiments for all three datasets, with substantial improvements over pure exploration

baseline (63%–73%) and ETC baseline (14%–32%) (**Fig. 9b**). These results validated that our learning model can be successfully translated to chemistry reaction data and is accurate in finding the most general conditions for various reactions. Different condition precisions and optimization objectives can also be accommodated through the flexible design of the optimization framework.

Compared to model performance, data efficiency is often overlooked during computational model development. RL models can be especially data-hungry and computationally expensive. While access to large amounts of data is possible in certain scenarios, the execution of a reaction has always been the bottleneck in chemistry reaction optimization, especially in a batch experiment setting where experiments are conducted sequentially. Notably, our learning model does not require any pre-training or initialization and is inherently data efficient. To further demonstrate the model's effectiveness at low data availability, we tested a large-scale Buchwald-Hartwig C–N cross-coupling HTE dataset previously published by the Doyle group and Merck.[41] After removing incomplete reaction entries, this dataset contains 300 unique combinations of aryl halides and isoxazole additives, 4 ligands in the form of palladium pre-catalysts and 3 organic bases, totaling 3600 experiments (**Fig. 10a**). MTBD as base, with *t*-BuXPhos (**L2**), *t*-BuBrettPhos (**L3**) and AdBrettPhos (**L4**) as ligands are the top three most general conditions based on average yield across the substrate scope (**Fig. 10b**). Various algorithms were again simulated with random starts, and the average accuracies of identifying the top three conditions were tracked throughout the simulations. Meaningful learning was achieved by most algorithms tested when compared to the plain annealing ε-greedy algorithm. The best-performing Bayes UCB algorithm achieved >90% accuracy on average after exploring only 2% of the scope (72 reactions) and converged to >95% accuracy after 100 reactions (or 2.8% of the scope) (**Fig. 10c**).

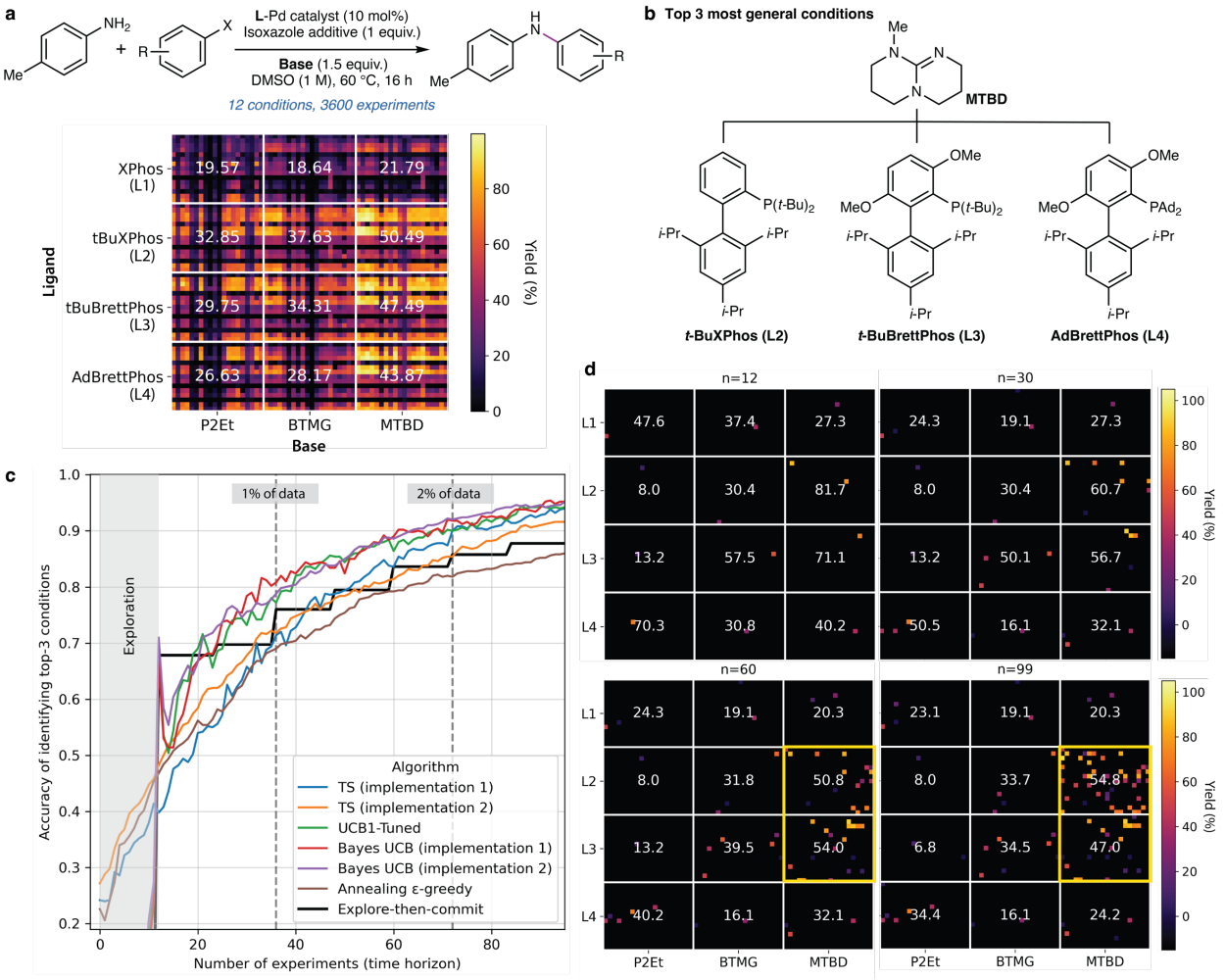


Fig. 10 Testing the bandit algorithms on a previously published C–N cross-coupling reaction dataset.

We also probed algorithm behaviors in detail by visualizing optimization progresses at specific time points during one simulation. Without any deliberate selection of favorable results, we used the first simulation run with Bayes UCB algorithm to visualize the experiments selected at four different time points, as well as the current empirical average yields for each condition combination at each time (**Fig. 10d**). During the exploration stage (up to n=12, n refers to the number of experiments), the algorithm sampled one experiment for each condition combination to

gain a preliminary understanding. The learning model subsequently balanced the exploration of conditions with limited data and the exploitation of conditions with high reactivity. The increased sampling of a particular condition provides a more accurate estimate of its average yield and better informs the algorithm of the potential gain if this condition was to be chosen again. At n=60, **L3**–MTBD (54.0% empirical yield) still had a higher empirical average than **L2**–MTBD (50.8% empirical yield). At n=99, the algorithm has corrected this inaccurate estimation by continuously sampling **L2**–MTBD, which turned out to have a higher empirical average (54.8% empirical yield, vs. 47.0% empirical yield for **L3**–MTBD). **L2**–MTBD (*t*-BuXPhos–MTBD) is therefore correctly identified as the most general condition in this dataset.

2.2.3 Optimization study 1: palladium-catalyzed C–H arylation reaction

Literature datasets that probe the effects of conditions on a scope of substrates often contain only a singular dimension of substrates.[40] However, in many chemical transformations, best exemplified by cross-coupling reactions, two or more substrate components are usually present in the scope. A reaction dataset with many diverse substrates pairings and calibrated reaction yields for all products under the same environment, one that is also sufficiently large for modeling, would be ideal to evaluate the performance of generality optimization algorithms in a regime where multiple substrate dimensions simultaneously interact with conditions. Due to the lack of such datasets in the literature, we decided to collect a palladium-catalyzed imidazole direct C5–H arylation dataset that satisfies these requirements. Compared to cross-coupling methods commonly

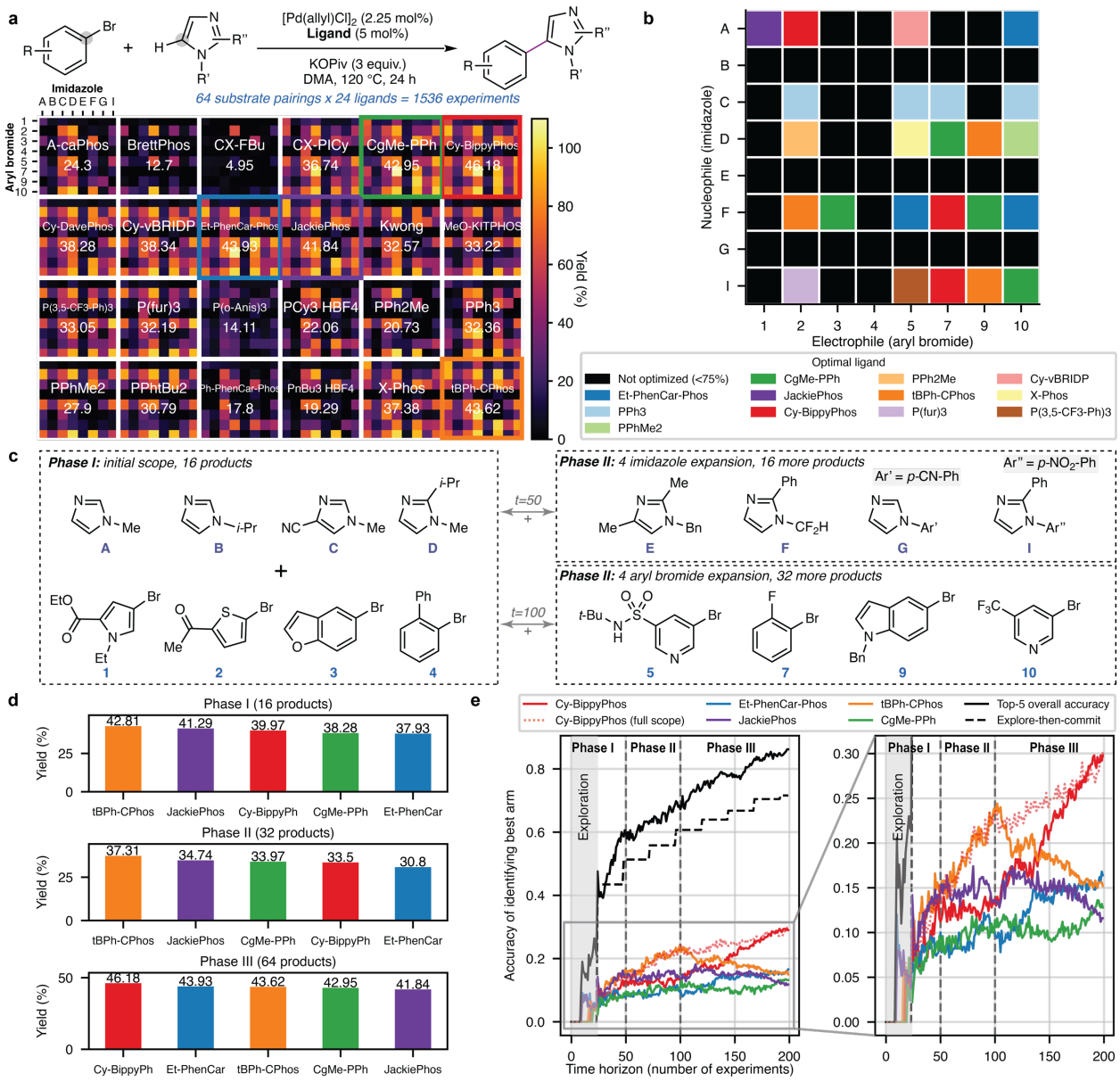


Fig. 11 Optimization studies of a palladium-catalyzed C–H arylation reaction.

studied in the form of large reaction datasets, direct C–H arylation bypasses the need for pre-functionalization or potentially unstable coupling partners. Building upon a C–H arylation dataset investigated in a prior collaboration between the Doyle group and BMS,[15] where conditions were extensively surveyed with a single pair of substrates, we expanded the substrate dimensions of both imidazoles and aryl bromides and specifically studied ligand effects with an expanded ligand scope. Commercial imidazoles and aryl bromides were clustered using *k*-medoids clustering, and

representative molecules from each cluster were selected based on expert knowledge to cover various functionalities and substitution patterns. An extended ligand scope was selected from the BMS monophosphine ligand library, which includes most of the ligands in the previous dataset. A total of 64 unique C5-arylated imidazole products were generated from 8 imidazoles and 8 aryl bromides, each evaluated with 24 ligands yielding 1536 total reactions (**Fig. 11a**).

We first retrospectively analyzed the dataset by mimicking a traditional model substrate approach, where ligands are screened with a model substrate (or product) to identify the highest-performing ligand as optimal. For each of the 64 products in the scope, we filtered out products (40 out of 64) that did not achieve a reaction yield above 75% (these reactions are usually considered as "not optimized" in practice). For the rest of the products, the highest-yielding ligand was selected (**Fig. 11b**). 12 out of 24 ligands in the scope can be considered as "optimal" with different substrate pairings. Most of these ligands, however, are non-optimal when considering all 64 products. The most notable example, PPh_3, is the optimal ligand for imidazole **C** with multiple aryl bromides, but its average yield over all products is only 32.4%, compared to the 46.2% for CyBippyPhos. Moreover, our previous HTE study of C–H arylation,[15] where imidazole **C** and aryl bromide **7** were used as model substrates to evaluate 1984 different reaction conditions including 14 monophosphine ligands, identified CgMe-PPh as the optimal ligand almost exclusively (19 out of top 20 conditions, with the only other ligand being PPh_3). These analyses highlight that a traditional screening approach with model substrates, even after significant exploration of the condition space, does not usually produce a satisfying condition. The derived "optimal" conditions can be biased and misleading, often with poor general applicability. In contrast, simulating our learning model with this dataset showed an 85% top-5 accuracy (**Fig. 11e**, compared to the 21% random exploration baseline), and a >95% top-9 accuracy on average after 200 experiments (see

Section 2.4.8 for detailed simulation studies for this reaction). Non-optimal ligands, such as PPh_3, are almost always excluded from consideration by the model, thus reducing bias when choosing general conditions.

One key advantage of the bandit optimization model is that no search space needs to be explicitly defined. Reactivity responses from various substrates are treated as feedback from the environment that the algorithm is learning from. This means that the substrate scope, as part of a dynamic environment, can arbitrarily change on the fly and the model can learn these changes continuously based on the feedback it receives during optimization. It is common in practice to expand the substrate scope and further evaluate a developed method's utility, which will affect how generally applicable a condition is and possibly affect the optimization model's ability to select such conditions. To test the learning model's performance in this problem setting, we designed a test scenario where both the imidazole and aryl bromide scopes available to the algorithm were restricted at first and expanded on the fly during optimization. Four imidazoles (**A**, **B**, **C**, **D**) and four aryl bromides (**1**, **2**, **3**, **4**) constituted the initial scope, defined as Phase I. After 50 experiments in Phase I, the imidazole scope was expanded to include four additional imidazoles (**E**, **F**, **G**, **I**), creating 16 new potential products in Phase II. After 50 experiments in Phase II, the aryl bromide scope was expanded again to include four additional aryl bromides (**5**, **7**, **9**, **10**), creating 32 new potential products in Phase III (**Fig. 11c**). While Phase I and II experience similar rankings for the top 5 ligands, the relative order changes in Phase III after the addition of four aryl bromides (**Fig. 11d**). During optimization simulations, the individual accuracies over time for each of the top 5 ligands were tracked and compared (**Fig. 11e**). The model correctly identified the initial ligand reactivity rankings in Phase I and II. Crucially, when the reactivity ranking was changed in Phase III, the algorithm did not overcommit and successfully adjusted its belief in

ligand performance by increasingly sampling Cy-BippyPhos (red) and Et-PhenCarPhos (blue), the top two performing ligands. The previous top ligands, tBPh-CPhos (orange) and JackiePhos (purple), were downgraded by the algorithm in Phase III. We also compared the accuracy of Cy-BippyPhos under a substrate expansion regime with the accuracy of Cy-BippyPhos obtained from a separate optimization where the full substrate scope is always available for the algorithm to sample from. Although the initial accuracies understandably differed due to the different reactivity distributions in Phase I and II, the end accuracies at experiment 200 are similar despite the differences in the initial sampling pools. The model is capable of learning a changing substrate scope through continued sampling, while not overcommitting to any prior beliefs. The same level of performance can also be achieved in the same time frame regardless of the substrate scope expansion, further highlighting the developed model's efficiency.

2.2.4 Optimization study 2: amide coupling reaction

Due to the prevalence of amide bond structures in biological systems and pharmaceutical compounds, amide coupling reactions are the most commonly employed reactions in medicinal and process chemistry.[42] Carboxylic acids are often preferred as inexpensive and abundant starting materials. Their chemical stability, while desirable on account of the ease of handling on scale, necessitates activation by coupling reagents, usually through in situ formation of an acid halide or anhydride. Despite the vast number of activators (>200) developed for amide coupling reactions,[43] chemists often resort to a few routine reagents on the basis of their proven reliabilities.[44] However, the efficacy of these coupling reagents when applied to specific target substrates is still difficult to assess a priori, especially for the challenging coupling with weakly nucleophilic anilines. Aniline deactivation from the aromatic system, as well as accompanying steric and electronic demands

from various substituents, complicates the selection of productive coupling reagents. Other aspects of reaction conditions, such as bases and solvents, can also affect reactivity.

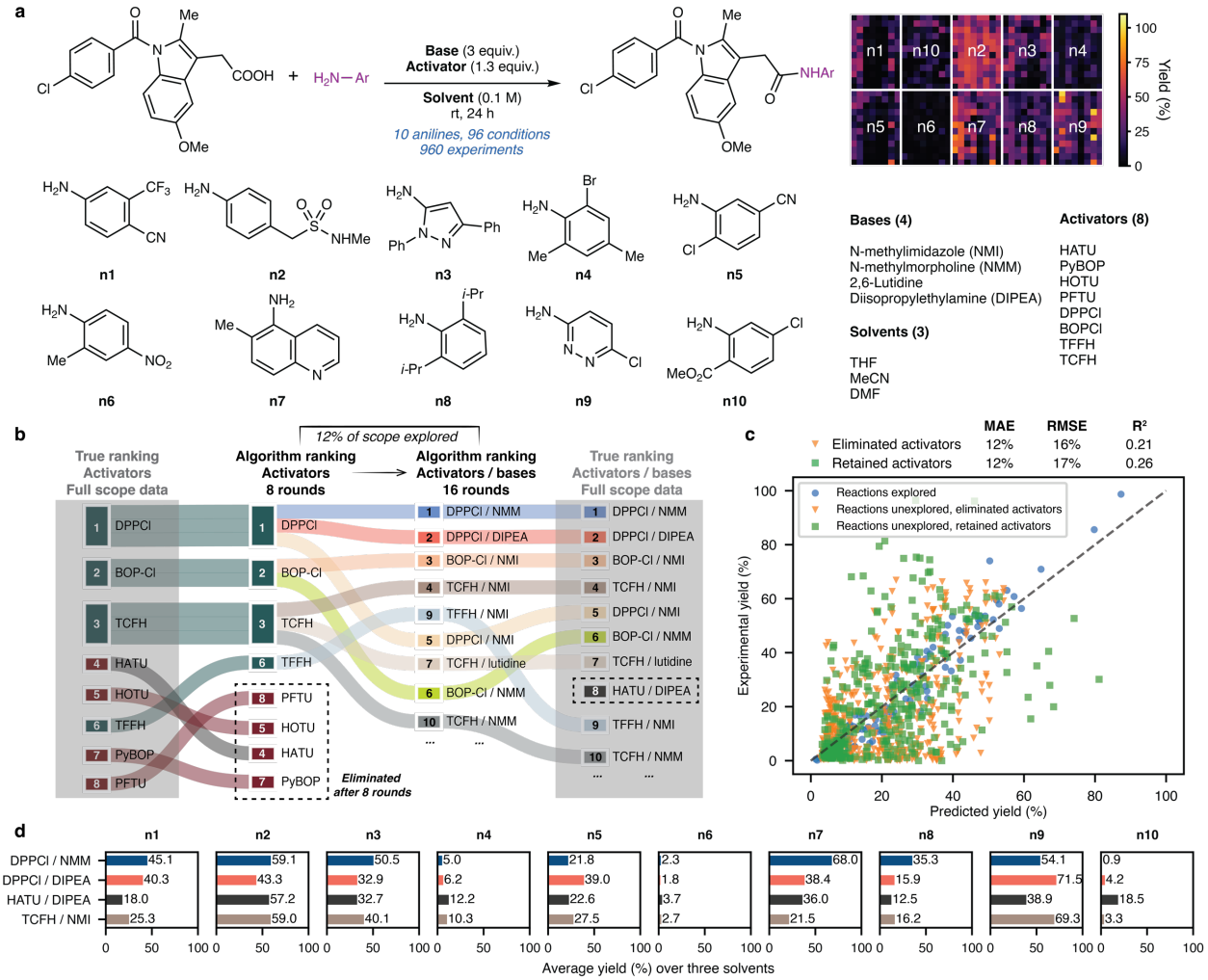


Fig. 12 Optimization studies of an amide coupling reaction with anilines.

Using the late-stage functionalization of indomethacin, a commonly prescribed nonsteroidal anti-inflammatory drug (NSAID), as an example, we sought to demonstrate our model's ability to identify generally applicable amide coupling conditions when faced with a diverse scope of aniline substrates and reaction conditions. Starting from a commercial library of anilines, we generated dense vector embeddings for all molecules using mol2vec[45] and clustered

them into ten groups using *k*-means clustering. One representative aniline was chosen from each cluster to constitute the aniline scope, which encompasses combinations of various heterocycles (quinolines, pyrazoles, pyridazines), electronically deactivating groups (nitriles, nitros, trifluoromethyls), sterically demanding *ortho*-substitutions, and potentially problematic functional groups (aryl chlorides/bromides, sulfonamides, esters). A series of eight amidation reagents, including aminiums, uroniums, (halo)phosphoniums, and phosphinic halides, were investigated as part of the condition scope, as well as four common organic bases and three solvents (**Fig. 12a**).

For the defined reaction scope, we attempted to identify the most general activators and bases for the selected scope and used the solvent dimension as a way of minimizing anomalous experimental observations. We first aimed to filter out less-effective activators by setting the optimization objective to activators alone. Unlike simulation studies where real-time feedback was immediately provided for each proposed experiment, batch experiments are necessary in practice to maximize time efficiency, resulting in a delayed feedback setting. After each proposal of batch experiments, predicted results for these experiments, which came from a separately trained supervised learning model with existing data, were continuously supplied to the bandit algorithm until experimental feedback became available. After 8 rounds of initial experiments (5 experiments per round), activators were ranked by reactivity based on the model's beliefs, and the bottom four activators (PFTU, HOTU, HATU, PyBOP) were eliminated. For the four remaining activators (DPPCl, BOP-Cl, TCFH, TFFH), the optimization objective was modified to activator–base combinations. Relevant data for the four activators retained were recycled and incorporated as knowledge of the new objective by the optimization model. After 16 additional rounds of experiments, all activator–base combinations were again ranked by projected reactivity (top nine

conditions are shown in **Fig. 12b**). Overall, about 12% of the reaction scope were experimentally explored following the model's suggestions.

To conclusively evaluate the resulting rankings from our model, we collected experimental results for all remaining reactions not explored during optimization and analyzed true reactivity rankings for activators and activator–base combinations for comparison. The model correctly identified and ranked the top three activators during the activator selection phase. For activator–base combinations, top nine out of ten combinations were identified, with the top four correctly ranked. Interestingly, HATU–DIPEA, one of the most commonly applied amide coupling activator–base combinations,[46] was the only condition not selected in top ten as HATU was eliminated in the initial rounds. Employment of DPPCl (diphenylphosphinic chloride) with NMM or DIPEA yielded the most effective general reactions conditions, ranking number one and two, respectively. Using HATU–DIPEA as a benchmark, the average yields over three solvents (THF, MeCN, and DMF) for DPPCl–NMM and DPPCl–DIPEA for each aniline substrate were also analyzed (**Fig. 12d**). DPPCl–NMM significantly outperformed, or at least matched, HATU–DIPEA for most anilines except **n10**, including highly deactivated anilines (**n1**) and sterically hindered anilines (**n8**). When compared to TCFH–NMI, a reagent combination developed by BMS for challenging amide coupling reaction with non-nucleophilic amines,[47] DPPCl also exhibited superior reactivities for selected anilines (e.g., **n7**). Although not a commonly employed amide coupling reagent, the optimization results suggest that DPPCl can be effective for amide coupling with anilines. These findings can extend to well-established activators not included in the model: for example, in comparison to T3P, a mechanistically similar activator that is much more frequently used in amide coupling, DPPCl can be considered as a promising alternative reagent with exceptional thermal stability[48] and improved atom economy. In fact, effective amide couplings

using DPPCl have been separately investigated by BMS.[49] The desirability of DPPCl-mediated amide coupling in commercial routes has also been demonstrated on multi-kilo scales,[50] further corroborating the optimization model's findings.

Finally, we evaluated the accuracy of the final prediction model from the last round of optimization with measured ground truth data for the full scope. The random forest model was only trained with 12.5% of the data from the reaction scope explored during optimization but exhibits good prediction accuracy for unexplored experiments involving both activators retained and eliminated after initial experimental rounds (12% mean absolute error for both, **Fig. 12c**). The good accuracy of the prediction model under a low-data regime further validates the approach of using a supervised machine learning model to predict experimental results in a delayed-feedback setting during optimization.

2.2.5 Optimization study 3: phenol alkylation reaction

The prevalence of alkyl aryl ethers in natural products and pharmaceuticals has prompted developments in mild and general syntheses of these products. Despite advances in transition-metal catalyzed C–O cross-coupling reactions,[51] traditional approaches, such as Williamson ether synthesis,[52] Mitsunobu etherification[53] and nucleophilic aromatic substitution (SNAr), are still widely used due to their simplicity. However, these reactions usually have limited functional group compatibility. We decided to investigate a base-promoted phenol alkylation reaction with alkyl mesylates, which also suffers from similar substrate applicability issues, with the objective of identifying a more general condition.

Six mesylates and six phenols were selected from commercial databases as substrates with varying structural motifs and complexities. We randomly left out one phenol (**p5**) and one mesylate (**m1**) as external testing substrates and did not include them in the optimization process. As a result,

25 substrate pairings (five phenols × five mesylates) were sampled by the algorithm during optimization, and 11 unseen pairings (those with **p5** and **m1**, including **p5-m1**) were tested after to externally validate the generality of the identified conditions. Six bases (inorganic and organic), two solvents and three temperatures were selected as the condition scope, totaling 36 overall conditions (**Fig. 13a**). Conditions selected by expert medicinal and process chemists at BMS and their corresponding reactivity data were used as a benchmark for the bandit algorithm's decisions and optimization performance.

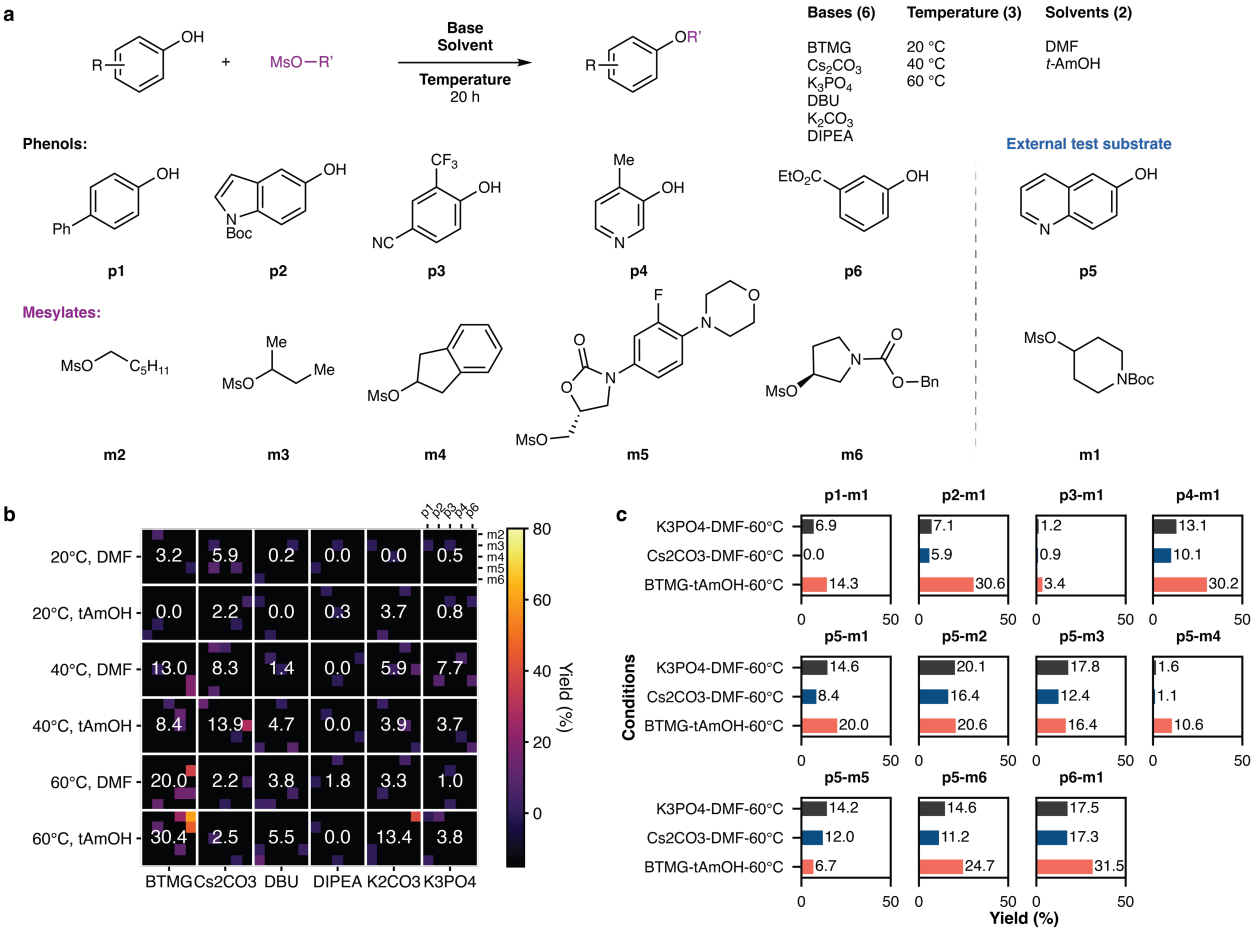


Fig. 13 Optimization studies of a phenol alkylation reaction with mesylates.

Using UCB1-Tuned algorithm, we conducted four rounds of optimization with a total of 90 experiments (36, 18, 18 and 18 for each round, all conducted experiments are included in

Section 2.5.3). The first round of experiments is a uniform exploration of all conditions required by UCB-type algorithms. All conditions were sequentially explored with randomly sampled substrate pairings (21 out of 25 were sampled at this stage). Subsequent rounds of experiments were chosen by the algorithm evaluating different conditions and substrate pairings. After 90 experiments, or 10% of the available reaction scope, the average yields and number of samples for each condition were analyzed (**Fig. 13b**). Significant base (BTMG) and temperature (60 °C) effects on reactivity were observed, with BTMG–*t*-AmOH–60 °C identified as the most generally applicable condition, achieving an average yield of 30.4% over five substrate pairs tested. Two conditions most commonly used and most successful in past HTE datasets at BMS, Cs_2CO_3–DMF–60 °C and K_3PO_4–DMF–60 °C, were selected as benchmark conditions for comparison (see Section 2.5.3 for details on the selection of these conditions). These three conditions were tested on 11 unseen substrate pairings that involve phenol **p5** and mesylate **m1** (**Fig. 13c**). Compared to the benchmark conditions, the algorithmically derived condition, BTMG–*t*-AmOH–60 °C, performed better (or at least comparably) in all except one substrate pairing (**p5**-**m5**). These results showed that bandit algorithms are compatible with continuous parameter optimization and can be used with batch sizes amenable to HTE. Furthermore, validation with unseen substrate pairings showed that the condition identified by the bandit algorithm during optimization is more generally applicable for the reaction scope, even when compared with conditions selected by practicing chemists that performed well in historical datasets.

2.3 Conclusions and outlooks

Our learning model can achieve data-efficient learning at high accuracies and has unique functionalities that we substantiated through the experimental investigations of three chemical transformations. Despite its advances, the optimization framework still has limitations and can be

improved in a few areas. Given the typical experimental budget (100–1,000 experiments) and the efficiency of optimization (2–10% exploration of the scope needed), our approach is not suitable for the evaluation of a scope with thousands of possible conditions. Rather, the condition scope needs to be reduced by expert chemists to selective conditions that show reactivity initially, so that more experimental resources can be spent on sampling substrates. Furthermore, the treatment of reaction conditions as independent arms in a stochastic multi-armed bandit problem setting means that there is no sharing of structural information between arms. Although effective in all our test cases, this approach can be inefficient when more than 100 conditions need to be simultaneously evaluated and significant correlations between conditions are present. Elimination of less effective conditions, as demonstrated in the amide coupling example (optimization study 2), can attenuate this problem. Alternatively, suitable descriptors for conditions could be used to transfer knowledge between similar conditions, but the choice of descriptors is difficult to determine a priori. Finally, although we showed that the learning model can successfully adjust to a changing environment with unseen substrates and correctly identify most general conditions, addition of any new conditions will require additional sampling for the model to have an accurate estimation of their performance. This issue was partially addressed by the inclusion of a real-time supervised learning model, which can be used to extrapolate to unseen conditions and predict their effectiveness, but a more direct approach with knowledge transfer between arms is still desired.

From a theoretical standpoint, there are also other potential directions that can be further explored. First off, a more dynamic problem setting can be explored where a “living model” is always operational and evolving using newly available data on new substrates. One approach for this problem setting is to discount older data points in the reward function, e.g., use a geometric series of a discount factor (older data points will get more discounted), or use a sliding window

and only consider the last *n* data points. The second area to explore is a new sampling strategy for the substrate dimensions. Currently, bandit optimization model will randomly sample a substrate for a selected condition. This is done for several reasons (see section 2.4.7 for details) and have been shown to be more effective than other sampling strategies.[72] However, other substrate sampling strategies can still be explored, e.g., maximize substrate diversity by selecting the substrate that is the most distant (distance metrics based on molecular fingerprints or other representations) from previously selected substrates. Another area to explore is to improve the uncertainty estimation function, which currently relies on the number of substrates sampled with each condition. A recent study showed that uncertainty estimation can be improved with supplemental data from a ML prediction model (which has been implemented and used for a different purpose in this work).[73] In addition to the stochastic bandit setting investigated in this work, other types of bandit problem settings and their solutions can also be applied, such as contextual bandit, combinatorial bandit, and linear bandit.[33] In particular, contextual bandit has been considered, but it is difficult to implement in practice. In other application settings such as website traffic optimization, contextual bandit works by observing a context (such as user interest) and subsequently suggesting options (such as website contents) more suitable for that context. Although this might seem like a suitable approach for generality optimization where the algorithm suggests reaction conditions by observing the substrate context first, chemists usually have the freedom to choose any substrate from the scope during optimization. Therefore, it is more beneficial to strategically sample all the substrates from the start, rendering the contextual bandit approach less relevant.

2.4 Computational section

2.4.1 Bandit optimization algorithms

Most of the algorithms discussed here have been extensively studied in literature. While theoretical studies tend to focus on proving expected regret bound and sample/time complexity of the algorithms, we aim to give a high-level description of these algorithms with minimal usage of mathematical equations and symbols, as well as some empirical observations of the behavior of these algorithms. Not intended as a rigorous technical explanation of the algorithms, these introductions aim to help non-experts (such as those in the field of chemistry) to better understand the logic and underlying principles of these algorithms without the need of analyzing complex mathematical notations.

Bandit optimization algorithms are algorithms designed to solve the multi-armed bandit problem. In a multi-armed bandit problem, the player is faced with multiple arms (or options), each with a different reward distribution that was initially unknown to the player. The player must choose arms strategically to identify the best arm and to maximize cumulative rewards. The general role of a bandit algorithm is to select an action to take next (select an arm to play), based on all past results that have been collected so far. A successful algorithm efficiently exploits known good arms and explores arms with high uncertainty. The information available to an algorithm at each time point ***t*** is all the arms selected at all previous time points, and the rewards returned by playing each of those arms. Different bandit algorithms use this same information in different ways to determine the next action, all with the same objective of identifying the best arm and maximizing overall reward in the long run.

<u>Pure exploration</u>

Pure exploration simply explores all arms randomly throughout the time horizon. At each time point, one arm is randomly selected and played.

Explore-Then-Commit (ETC)

Explore-then-commit is a very similar to traditional A/B testing, where equal amounts of resources are allocated to all options during the initial exploration phase. The best option is selected based on initial data only and exploited throughout the rest of the time horizon. This algorithm is also similar to the traditional reaction optimization approach employed by chemists, where control experiments are performed with one reaction component varying at a time. The "best" reaction parameters (solvent, catalyst, temperature…) are determined from these experiments and exploited for the rest of the optimization campaign.

Formally, ETC is characterized by the number of arms n, and the number of explorations for each arm ***m*** (***m*** is a natural number). Assuming we have a preset value of ***m***, the action at each time point t is chosen as such:

$$\text{Action } A_t = \begin{cases} (t \bmod n) + 1 & \text{if } t \leq mn \\ \arg\max_i \hat{\mu}_i(mn) & t > mn \end{cases}$$

To establish a ETC baseline that can be compared to other algorithms at every time point ***t,*** we simulate our datasets with all possible values of ***m***, limited by the maximum number of experiments allowed ($m \in [1, t_{\max} \mod n]$). Since each round of exploration will take ***n*** experiments (1 experiment for each arm), the accuracy (or other metrics) during a current exploration round is calculated with results from all previous rounds of exploration that are completed, excluding the current, ongoing round. The resulting baseline is a step plot, with accuracy (or other performance metric) being updated every ***n*** experiments. Examples of the

explore-then-commit baseline and more detailed explanations of how the baselines are calculated can be found in Section 2.4.2 for synthetic data, and Section 2.4.8 for reaction data.

ε-greedy

ε-greedy algorithm is an improved version of simple greedy algorithm (exploitation only). ε-greedy incorporates exploration as follows: with a parameter 0<ε<1, at each time point, the algorithm either randomly explores all arms (with probability ε) or exploits the current best arm (with probability 1-ε). In other words:

$$\text{Action } A_t = \begin{cases} \text{a random action} & \text{with probability } \epsilon \\ \arg\max_i \hat{\mu}_i(t) & \text{with probability } 1-\epsilon \end{cases}$$

The obvious limitation of ε-greedy is the necessity of selecting a fixed ε at the beginning of the experiment. If ε is too small, the algorithm does not explore enough at the start and will get tricked by a few positive examples and continuously exploit these sub-optimal options. If ε is too big, the algorithm collects a lot of initial data and figures out the best option quickly but will also continuously explore at later stages of optimization when it is not necessary to do so. Such late-stage exploration wastes resources when the optimal option has been identified.

It is often difficult to know whether a selected parameter will work for the real-world data we have not collected yet. We can only know that a certain ε is better with hindsight knowledge. One solution to this problem is to adaptively adjust ε throughout the time horizon: adopt a big ε at the start to explore all options and gradually decrease ε when exploration is not as necessary in later stages. Such adaptive algorithm is called annealing ε-greedy. The benefit of using annealing ε-greedy is that it eliminates the need to find appropriate ε for each specific use case, while providing somewhat of a performance guarantee.

A common annealing function (also used in our study) is:

$$\epsilon = \frac{1}{\ln(t) + 10^{-7}}$$

Plotting this function reveals that ε decreases over time, and the rate of decrease (first derivative) also decreases, making ε more stable as ***t*** increases (**Fig. 14**). The small number 10^{-7} is used to avoid any division by zero error.

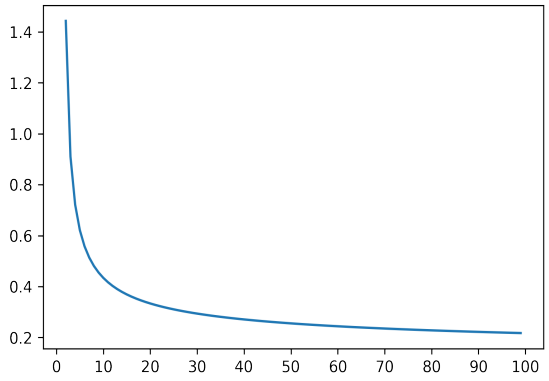


Fig. 14 Annealing function used for annealing ε-greedy algorithm.

Softmax (Boltzmann exploration)

Softmax algorithm assigns each arm a probability that is proportional to the average empirical reward of that arm at each time point.[54] Arms with a higher empirical average reward will have a higher probability to be picked. Specifically, the probability for each arm is modeled using a Boltzmann distribution.

At time point ***t***, probability of selecting arm ***i*** for the next round is updated as follows:

$$p_i(t+1) = \frac{e^{\hat{\mu}_i(t)/\tau}}{\sum_{j=1}^{k} e^{\hat{\mu}_j(t)/\tau}}$$

Typically, a temperature parameter τ is also used to control the randomness of choices. When $\tau \to 0$, algorithm acts as pure greedy where only the arm with the highest empirical average

is picked. When $\tau \to \infty$, the algorithm becomes uniformly random regardless of the current empirical averages. An illustration of the different probability distributions with different τ's from the same empirical average is shown below.

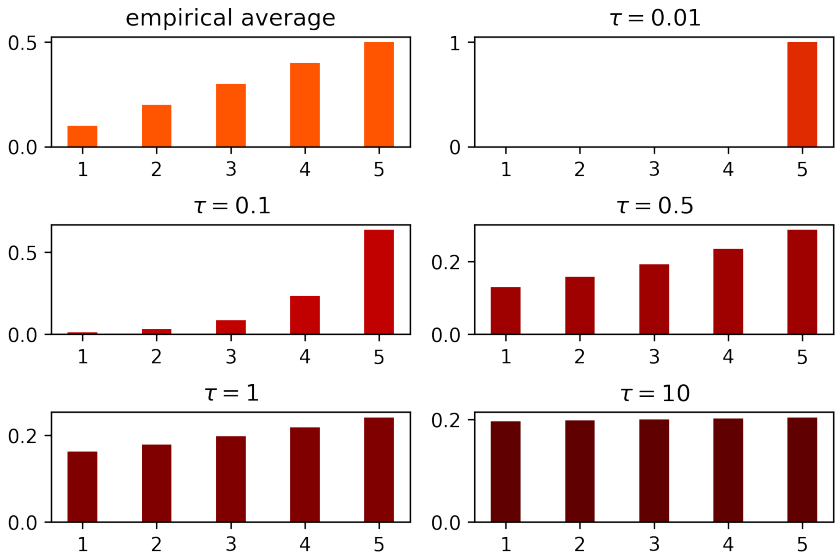


Fig. 15 Probability distribution of five arms with different parameter τ set for softmax.

When selecting τ, it is important to consider the possible range of rewards, as that will affect the scale of appropriate τ's. Similar to annealing ε-greedy, annealing τ is also possible to implement, with a big τ at the beginning to encourage exploration, which slowly decreases to exploit the best options. However, unlike the bounded parameter ε in ε-greedy ($0<\varepsilon<1$), τ does not have an upper bound. This makes finding the appropriate annealing function more difficult. In our testing, finding an appropriate fixed value for τ is usually easier than identifying a suitable annealing function.

Pursuit

Similar to softmax, pursuit algorithm also uses a set of probabilities to guide arm selection.[55] The update rules of probabilities, however, is not directly related to the empirical means. Starting with uniform probabilities, pursuit algorithm re-computes probabilities at each step with a learning rate β (0<β<1). The probability of selecting the current best arm (with the highest empirical average) increases while the probabilities of selecting all other arms decrease.

$$p_i(t+1) = \begin{cases} p_i(t) + \beta(1 - p_i(t)) & \text{if } i = \arg\max_j \hat{\mu}_j(t) \\ p_i(t) + \beta(0 - p_i(t)) & \text{otherwise} \end{cases}$$

Reinforcement Comparison

Reinforcement comparison[54] also uses probabilities to guide arm selection, but in a more complex fashion. First, a set of expected average rewards for each arm ***i*** are updated with empirical average rewards at each time step ***t*** with a learning rate of α (0<α<1).

$$\bar{r}(t+1) = (1 - \alpha)\bar{r}(t) + \alpha r(t)$$

Another set of heuristics called preferences are then updated via comparison between expected and empirical average rewards with a learning rate β (0<β<1).

$$\pi_{j(t)}(t+1) = \pi_{j(t)}(t) + \beta(r(t) - \bar{r}(t))$$

Finally, the set of probabilities used for arm selection is computed with a Boltzmann distribution.

$$p_i(t) = \frac{e^{\pi_i(t)}}{\sum_{j=1}^{k} e^{\pi_j(t)}}$$

Intuitively, more promising arms with empirical average rewards higher than expected average rewards will get a higher preference, resulting in a higher probability of getting selected. Theoretical analysis, to the best of our knowledge, does not exist for this algorithm. In practice, this algorithm can also be difficult to use due to the need of tuning two parameters at the same time. When correctly tuned, however, this algorithm can offer good performance compared to other simpler algorithms (*vide infra*).

Upper Confidence Bound

Most algorithms described above, like ε-greedy and softmax, only keep track of the current rewards for each arm and use that information to determine the next action. As a result, these algorithms can be "gullible": they can easily be fooled with a few unusually good/bad examples for a given arm initially. In other words, only considering empirical means to estimate true means does not account for the uncertainty of such estimation, which can result in less effective optimization. One improvement to address such limitation is to quantify the uncertainty with other information available to the algorithm, for example, the number of samples for each arm. Intuitively, the more times an arm is sampled, we are more confident that the empirical mean for this arm is close to the actual mean, and vice versa.

More specifically, upper confidence bound algorithms uses the strategy of "optimism in the face of uncertainty". Algorithm will be optimistic about any uncertainty present in the estimated mean and regard uncertainty as potential for improvement. As a result, upper confidence

bound algorithms will attempt to select arms with a high combined value of empirical mean and uncertainty. As time goes on, uncertainty for all arms decreases and algorithm can confidently select based on estimated empirical means.

The simplest algorithm in this family is UCB1.[56] At each round, UCB1 updates the upper confidence bound values for all arms. For arm ***j***, the UCB value is a combination of its current empirical mean and the number of samples for this arm compared to the overall number of samples for all arms:

$$\text{UCB1} = \bar{x}_j + \sqrt{\frac{2\log n}{n_j}}$$

After the update for all arms, UCB1 then selects the arm with the highest upper confidence bound, receives the reward for selected arm and updates the UCB values for all arms. Like discussed above, high empirical mean and low sample size can both prompt the algorithm to select a particular arm. Many other variants of upper confidence bound algorithms also exist, each with different confidence interval terms to describe the uncertainty. For example, UCB1-tuned[56] is found to work better in practice with a modified confidence bound:

$$\text{UCB1-tuned} = \bar{x}_j + \sqrt{\frac{\log n}{n_j}\min\{\frac{1}{4}, V_j(n_j)\}}$$

where V is defined as the upper confidence bound for machine *j*'s variances based on current samples:

$$V_j(s) = \frac{1}{s}\sum_{\tau=1}^{s} X_{j,\tau}^2 - \bar{X}_{j,s}^2 + \sqrt{\frac{2\log t}{s}}$$

Other UCB-type algorithms are also implemented in this work, including UCB2,[56] MOSS,[57] UCB-V,[57] DMED.[58] In our testing, we found that UCB1-tuned usually offered the best performance, even when compared to other more advanced algorithms. Therefore, the details of these algorithms are not discussed here and can be found in the respective publications where they were introduced.

Thompson sampling

Thompson sampling is one of the earliest algorithms proposed for bandit problems. The player will maintain a prior distribution for all arms, which gets updated according to Bayes rule with empirical data in each round. The player then samples from the posterior distributions and plays the arm that is optimal based on the sampling results. Exploration of the environment comes from the randomness during the sampling process. At early rounds, with the lack of empirical data, the posterior is not well-concentrated, which results in uniform (more or less) exploration of all arms by the algorithm. With more data collected, each posterior distribution more accurately represents the true distribution for each arm, and the algorithm tends to choose the optimal arms and is less likely to explore (though still possible, since it always samples from the posterior first, rather than simply choosing the posterior with the highest mean). Thompson sampling therefore takes into consideration both the empirical means and the uncertainty in mean estimation with the help of prior distributions.

Operationally, this procedure benefits from an algebraic convenience called conjugate prior. If the posterior distribution and prior distribution belong to the same probability distribution family, the prior is called a conjugate prior for the likelihood function. This gives a closed-form expression for the posterior, which greatly simplifies the update process at each time point. For this reason,

the reward distribution is usually assumed to be a Bernoulli distribution or Normal distribution, where a conjugate prior of Beta distribution or Normal distribution (Normal-Gamma if no fixed variance is assumed) can be used to give a closed form of expression for posterior update.

Assuming Bernoulli reward distributions and using beta distribution as conjugate prior, each arm i is represented by a beta distribution $\text{Beta}(S_i, F_i)$ where S represents the number of successes and F represents the number of failures this arm has seen. For the selected arm, the prior is updated to posterior $\text{Beta}(S_i + \text{reward}, F_i + (1\text{-reward}))$. For the arms not selected, no update happens. When selecting an arm, the algorithm samples from each posterior distributions and chooses the arm based on sample values.

Rewards can also be assumed as a Normal distribution. In this case, the conjugate prior can either be a normal distribution if variance is assumed to be fixed, or a normal-gamma distribution if variance needs to be estimated. A more detailed discussion on Thompson sampling under these situations in later sections.

2.4.2 Bandit algorithms: Monte Carlo simulation testing results with Bernoulli rewards

General remarks on testing frameworks

All implemented algorithms were evaluated with synthetic data first to validate the implementation and identify the optimal parameters and algorithm under different scenarios. The classic testing scenario uses multiple stochastic Bernoulli arms, each with stochastic rewards that follow a Bernoulli distribution with a different probability. In different test scenarios, we adjust the probabilities and the number of arms to determine the appropriate algorithms and their parameters to use under different circumstances.

Because bandit algorithms are active learning algorithms that query for a stochastic reward in real time, each run of the algorithm will give different results. To reduce the effect of randomness

in assessing algorithm performance, algorithms are evaluated with Monte Carlo methods. More specifically, all algorithms are repeatedly run many times and the average metrics across all runs (simulations) for each algorithm are used to establish and compare the performances of different algorithms.

Most of the bandit algorithms studied in literature were developed for arms with Bernoulli rewards (0/1 reward with a probability). Some of the algorithms can be readily applied to continuous rewards such as normally distributed reward, and others need to be modified first. During our testing with synthetic data, we focused on arms with Bernoulli rewards to assess the performance of all algorithms more accurately.

All the algorithms discussed are implemented in Python with a uniform function structure. Simulation testing and analysis functions used to analyze algorithm performance are also provided. These testing and analysis functions are provided as part of the software code. The raw testing logs are also provided, and all the testing results visualized in plots can be reproduced with the raw data.

Performance metrics

Accuracy is defined as the relative frequency (or percentage) of simulations where an optimal arm is played at each time point *t*. An effective algorithm will tend to play the optimal arms more often as time progresses, which increases the accuracy over time. It is worth noting that this definition tends to underestimate the ability of the algorithm to select an optimal arm: at each time point *t*, some instances of the algorithm might be exploring at the time when accuracy is evaluated, which does not affect the identification of optimal arms by the algorithm overall but does lower the accuracy. We modified this definition when we tested these algorithms with chemistry reaction datasets to consider all previous selections by the algorithm (Section 2.4.8).

Average reward is defined as the average reward across all simulations at time point *t*. As more optimal arms are played, the average reward will also increase. Average reward tends to trend similarly with accuracy, and we chose to focus on accuracy in most of the test cases. In practice, the highest achievable average reward will vary in different cases, but the accuracies are always the same scale (0-100%).

Cumulative reward is defined as the average cumulative sum of rewards across all simulations up until time point *t*. We did not implement a related metric, **regret**, which measures the difference between the rewards of action taken and the rewards of optimal action and is often used in theoretical analysis. It is, however, often impossible to calculate regret in a real-world application due to the lack of hindsight knowledge. Due to this limitation, we mainly focus on cumulative reward, which also trends with accuracy in most cases.

Test scenario 1: 5 Bernoulli arms with probabilities [0.1, 0.2, 0.3, 0.4, 0.9]

As discussed above, each Bernoulli arm is assigned a different probability for its Bernoulli reward distribution, with the arm with p=0.9 being the best arm that will produce the highest reward on average. An effective algorithm should find the optimal arm nearly 100% when converged.

For **ε-greedy** algorithms, a small ε (e.g., 0.1) exhibits slow start due to over-exploitation when the best option has not been identified but results in higher accuracy overall. A large ε (e.g., 0.5) is effective at the start, but quickly converges to a lower accuracy due to wasteful exploration when best option has already been identified in later stages. Annealing ε combines the advantages of both: sufficient exploration at the start and exploitation of the best option towards the end.

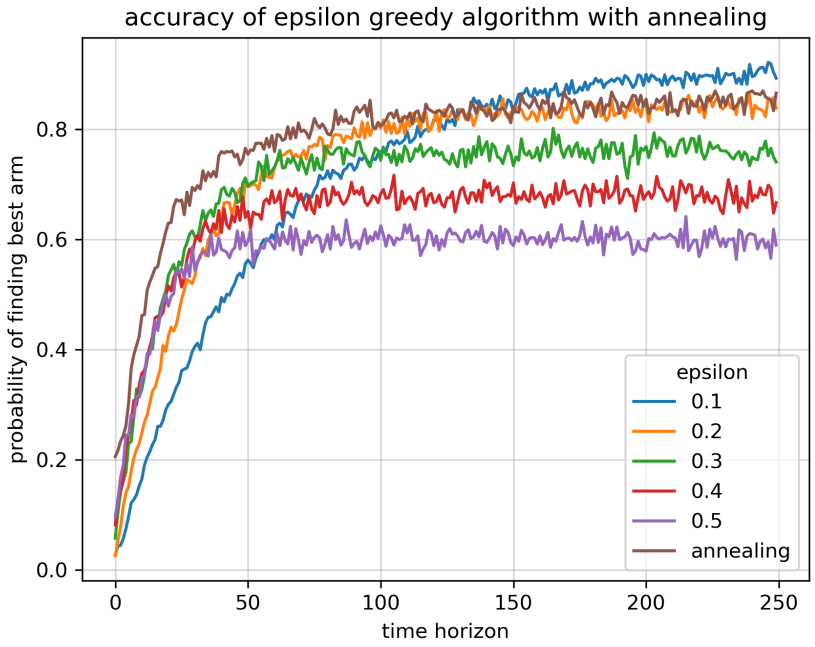


Fig. 16 Accuracy of ε-greedy algorithms with fixed and annealing ε's (test scenario 1).

Softmax algorithm uses scaled empirical averages to make a generally exploitative selection and controls the degree of exploration via a randomness parameter τ. The results for test scenario 1 are shown in **Fig. 17**. At the start, empirical averages are not representative of the true averages and randomness parameters do not matter as much. Therefore, all models perform similarly at the start regardless of τ. Towards the end, models with smaller τ (0.1, 0.2, annealing) perform better. Interestingly, unlike ε-greedy, annealing softmax does not appear to be superior to models with fixed τ's.

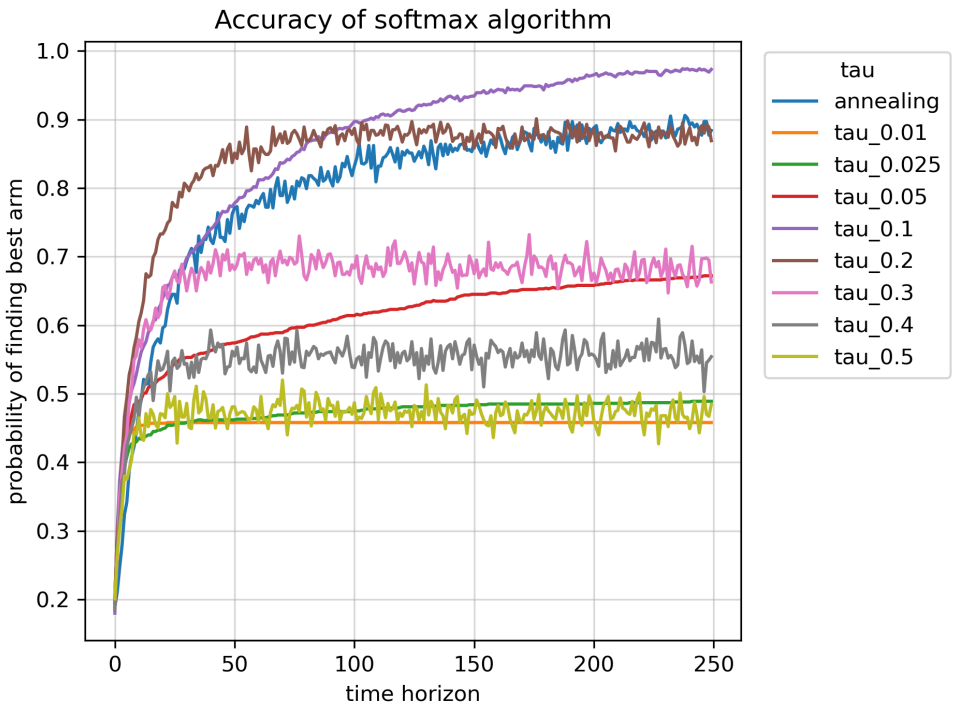


Fig. 17 Accuracy of softmax algorithms with fixed and annealing τ's (test scenario 1).

Pursuit algorithm maintains a probability of selection for each arm and adjusts the probabilities to favor the current optimal arm based on empirical averages. Large learning rates cause the algorithm to converge quickly, often with a subpar accuracy. Smaller learning rates do not converge prematurely but also result in slower learning overall. Only learning rate=0.05 converges to near 100% accuracy within the defined time frame. A 0.025 learning rate also converges to nearly 100% accuracy but suffers from the lack of accuracy initially.

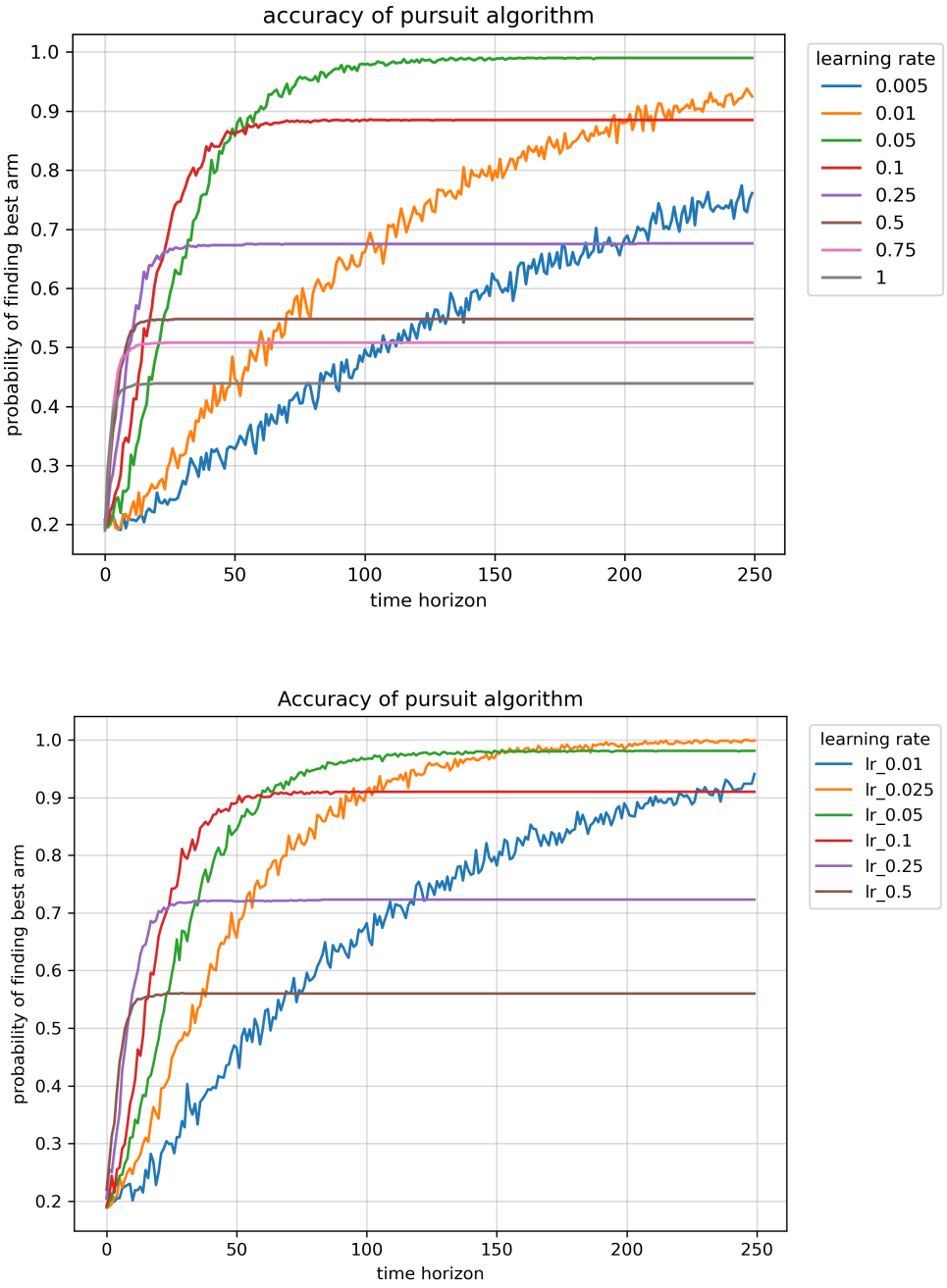


Fig. 18 Accuracy of pursuit algorithms with different learning rates (test scenario 1).

Reinforcement comparison has two parameters, α and β, to tune. Although it's possible to achieve high accuracy, correctly identifying both α and β requires fine tuning and can be very difficult to use in practice with limited prior knowledge. A series of α's and β's are simulated with

test scenario 1, and good results are only obtained after extended tuning. Due to the lack of theoretical studies on this algorithm, it is difficult to rationalize the choice of α and β from the empirical observations.

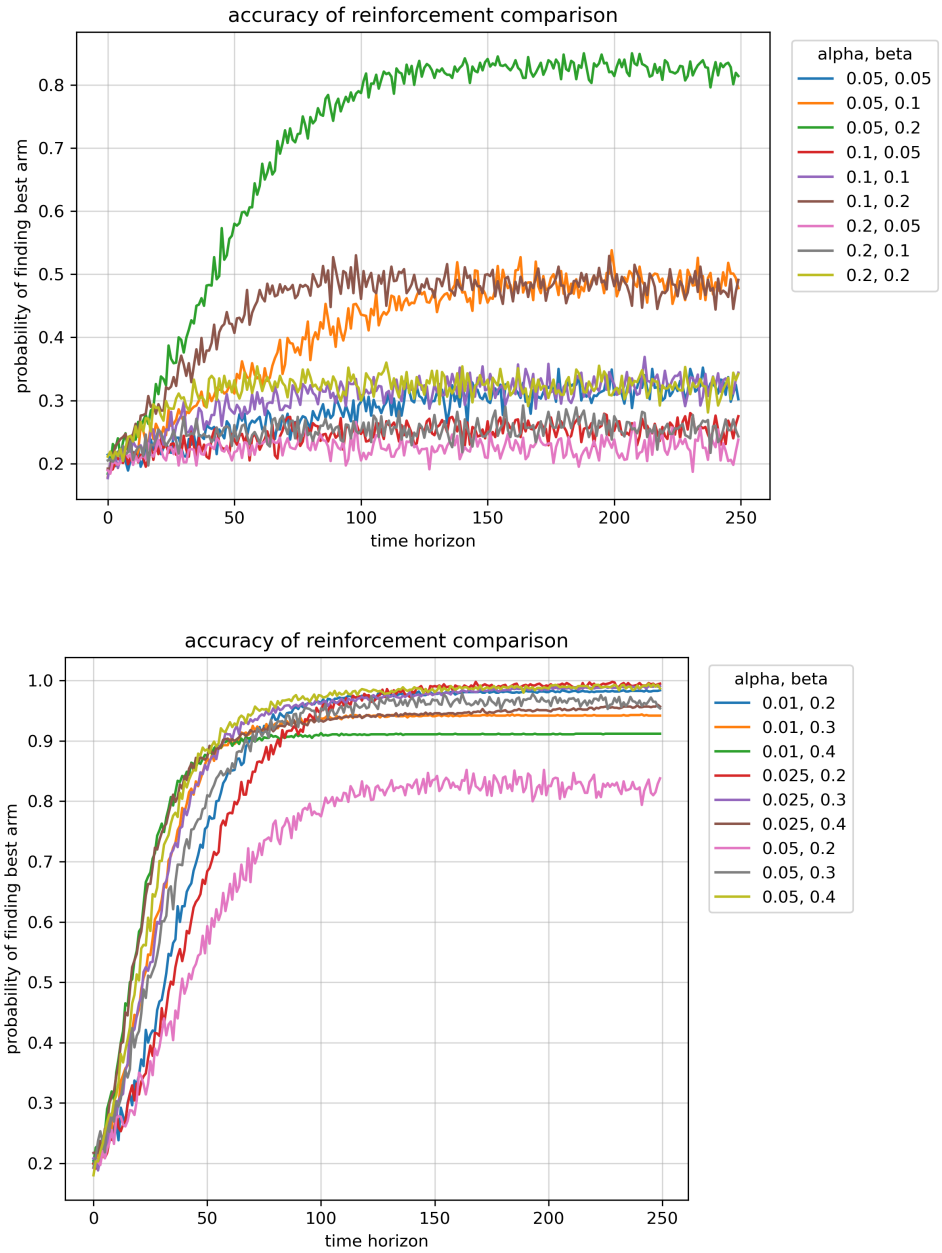


Fig. 19 Accuracy of reinforcement comparison algorithms with different learning rates (test scenario 1).

UCB-type algorithms and **Thompson sampling** do not have any parameters to select, which makes them ideal candidates to use for real-world applications. The only exception is UCB2 algorithms, which has a parameter α to control the confidence interval and the number of repeated samplings for a selected arm. UCB2, by design, will iteratively exploit best option for a period of time and "back off" to explore other options again, even if the optimal arms have been confidently identified. This behavior is not the most ideal for our purposes since we operate in a resource-limited environment and want to minimize unnecessary exploration. Compared to other algorithms, UCB1-tuned and Thompson sampling (beta prior) seems to perform the best (**Fig. 20**).

It is also worth noting that the initial spike that reaches 100% accuracy is caused by the uniform exploration that some of the UCB algorithms require. For these algorithms, every arm is sampled once initially to provide initial data. The implementation of this requirement in our software simply goes down the list and chooses each arm sequentially. At some initial time point *t*, algorithms across all simulations are selecting the same optimal arm, which causes the 100% accuracy artifact.

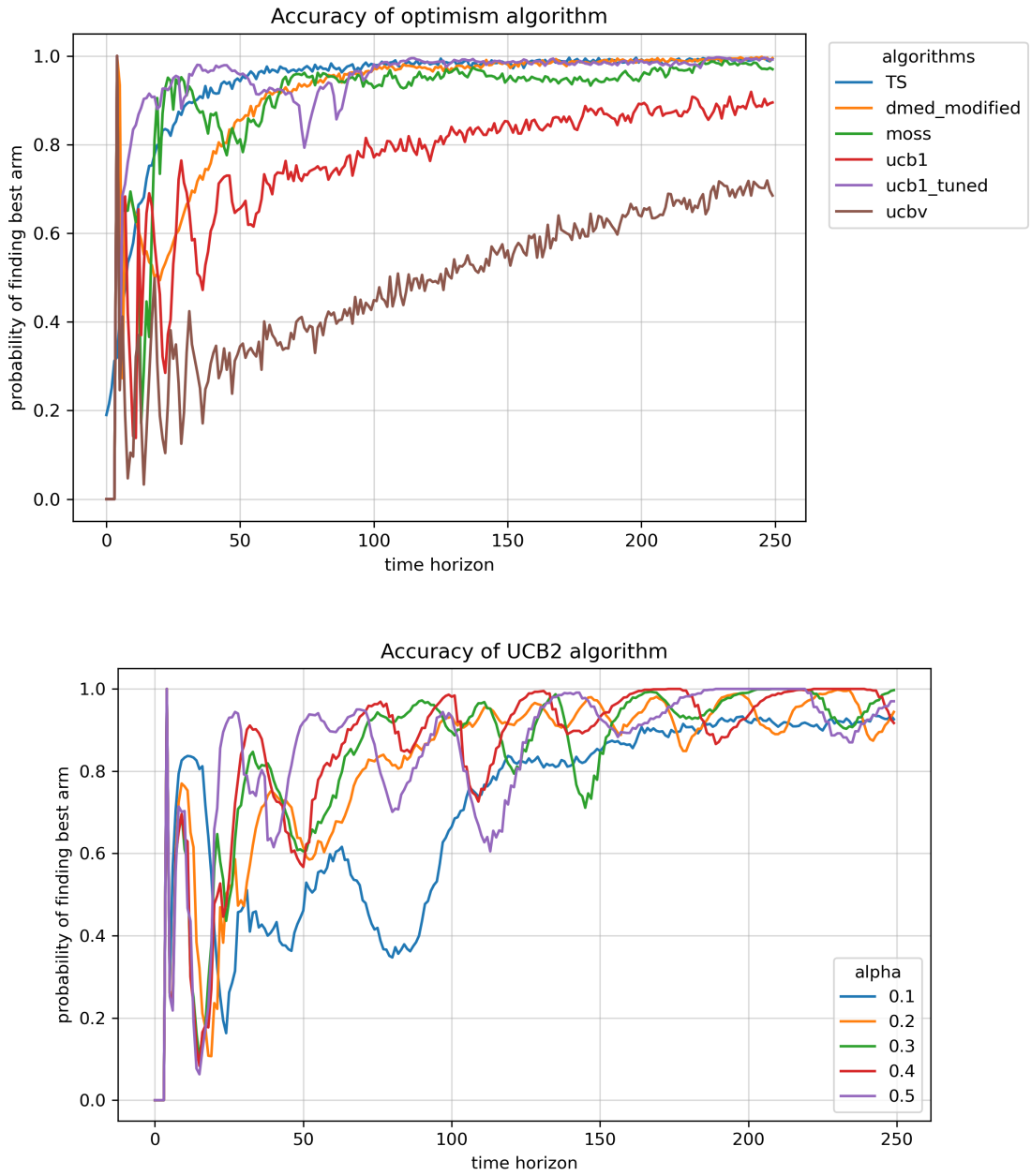


Fig. 20 Accuracy of UCB-type algorithms and Thompson sampling (test scenario 1).

Finally, some of the best-performing algorithms are compared against the explore-then-commit (ETC) algorithm, which we use as a more advanced baseline. Normally, ETC will have a fixed parameter for the number of exploration rounds. To establish a ETC baseline that can be compared to other algorithms, we took a stepwise approach and calculated the ETC accuracy with

the maximum number of exploration rounds possible at each time point. Specifically, for test scenario 1 with 5 arms, ETC baselines are established by progressively exploring 1, 2, 3… times per arm, which equates to 5, 10, 15… samples in total with 5 arms. After each exploration round, the algorithm temporarily commits to the best option based on samples seen so far and re-evaluate after the next round of exploration is complete. After every round of exploration, accuracy is calculated in the same way as the frequency of the true best arm being selected as optimal across all simulations. For example, from t=11 to t=15, the algorithm is committed to the best option determined by 2 rounds of exploration (10 samples), and the accuracy of such option being the true best option is calculated across all simulations. Starting from t=16 to t=20, the algorithm has an updated accuracy with 3 rounds of exploration (15 samples) complete. This process is continued, and the resulting accuracy curve represents the highest ETC accuracy attainable at each time point with the maximum number of exploration rounds.

As shown in the accuracy plot (**Fig. 21**), ETC can be quite effective for simple test cases, with similar levels of performance as UCB1-tuned and Thompson sampling.

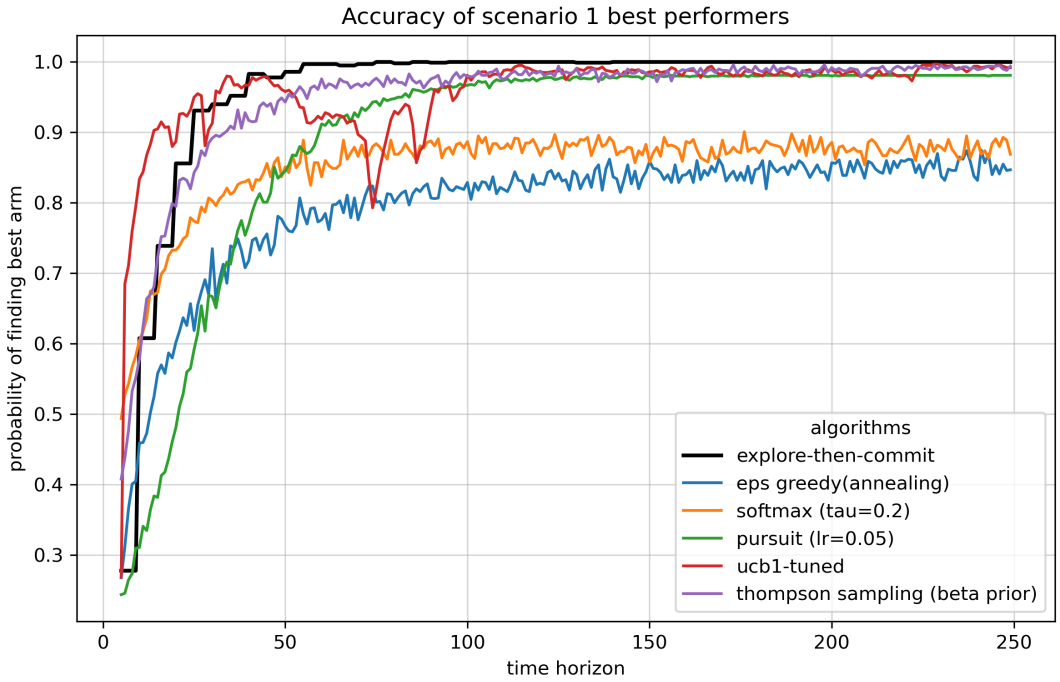


Fig. 21 Accuracy of best-performing algorithms (test scenario 1).

Note: data points from t=0 to t=4 are omitted for clarity, since some of the algorithms require initial explorations and will choose the same arm across all simulations, which will result in a 100% accuracy spike at a random initial time point (t=4 in this case) and might cause confusion. This also applies to all the test scenarios discussed in the following sections.

Test scenario 2: 5 Bernoulli arms with probabilities [0.1, 0.1, 0.1, 0.1, 0.2]

This scenario simulates the situation where there is still a clear best option, but the difference between rewards is minimal. Considering the time horizon specified (250 experiments), an effective algorithm should not converge and should continue to improve. The probability of selecting the best arm at the end of acquisition also indicates how effective an algorithm is.

For **ε-greedy** algorithms, annealing ε-greedy still offers the best performance, despite the overall lower accuracy. Due to the close averages of all arms, none of the algorithms converges before the specified time horizon (t=250).

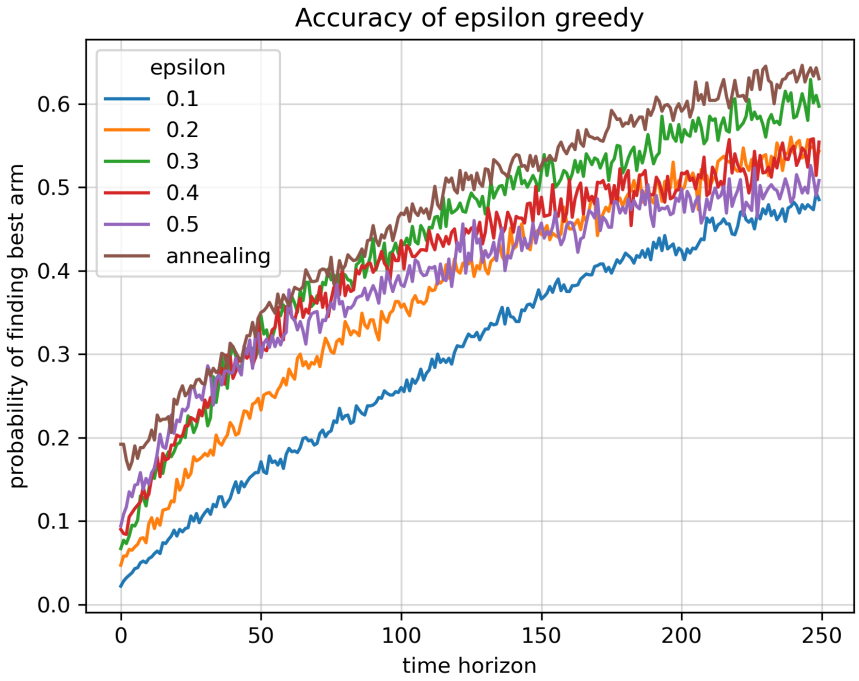


Fig. 22 Accuracy of ε-greedy algorithms with fixed and annealing ε's (test scenario 2).

Softmax algorithms exhibit similar behavior compared to scenario 1, although a smaller τ is required in this case to better differentiate the small differences in averages.

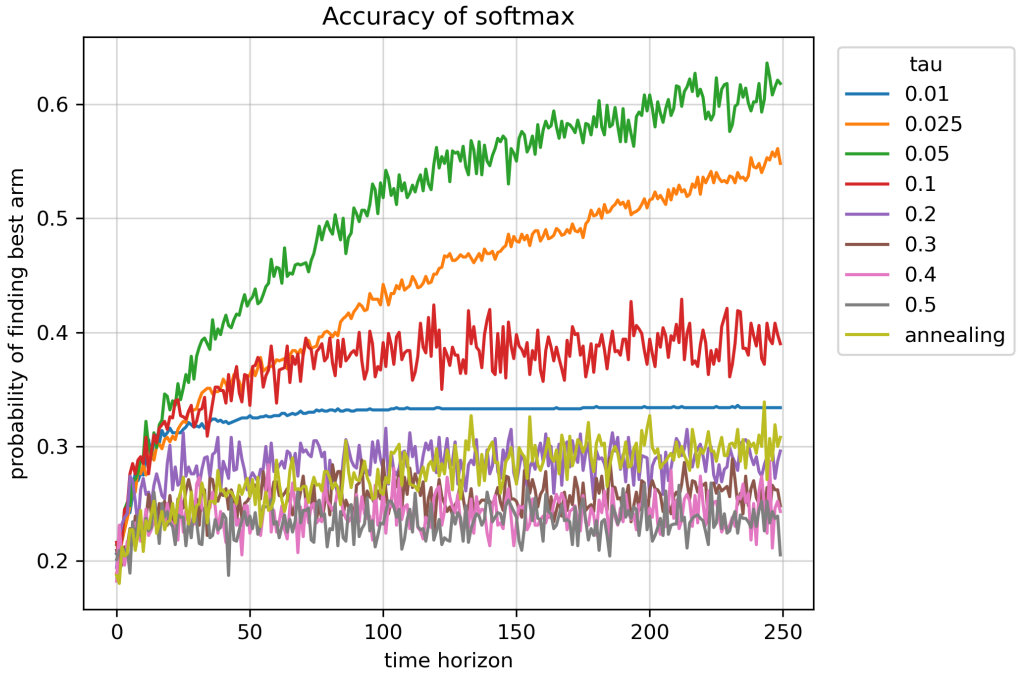


Fig. 23 Accuracy of softmax algorithms with fixed and annealing τ's (test scenario 2).

Pursuit algorithms in this case are also more effective with a very small learning rate. With learning rate bigger than 0.05, algorithms converge prematurely.

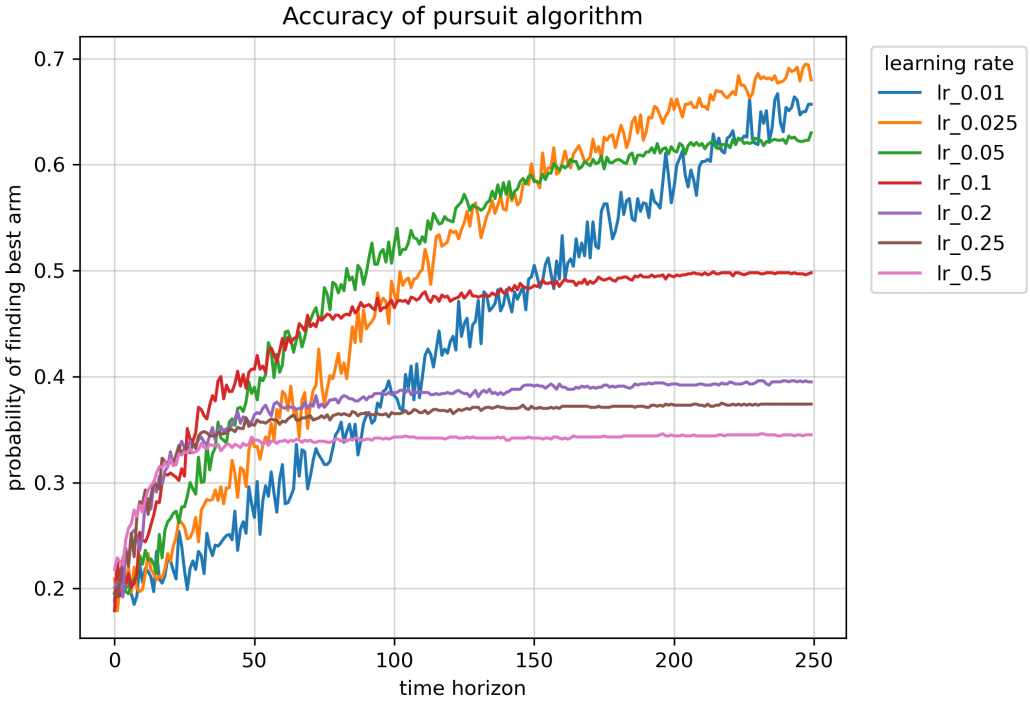


Fig. 24 Accuracy of pursuit algorithms with different learning rates (test scenario 2).

Reinforcement comparison does not offer any discernable trend when it comes to parameter selection. (α=0.01, β=0.4) seems to perform the best. Compared to the optimal parameters identified in test scenario 1, α is much smaller and β is much bigger. Again, it is difficult to identify these parameters a priori, which makes it impractical to use in real time.

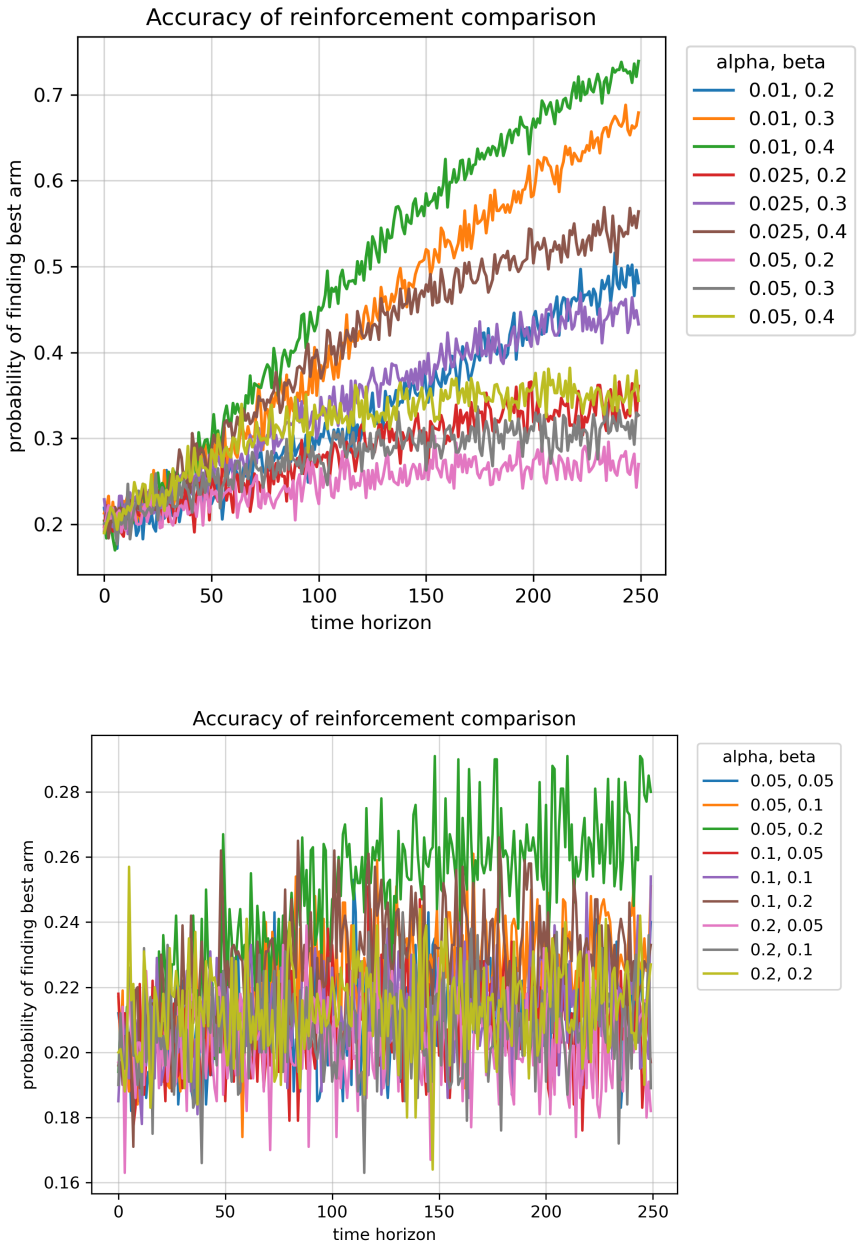


Fig. 25 Accuracy of reinforcement comparison algorithms with different learning rates (test scenario 2).

UCB-type algorithms and **Thompson sampling** offer similar level of performance in this test case, with Thompson sampling and UCB1-tuned being the best performers.

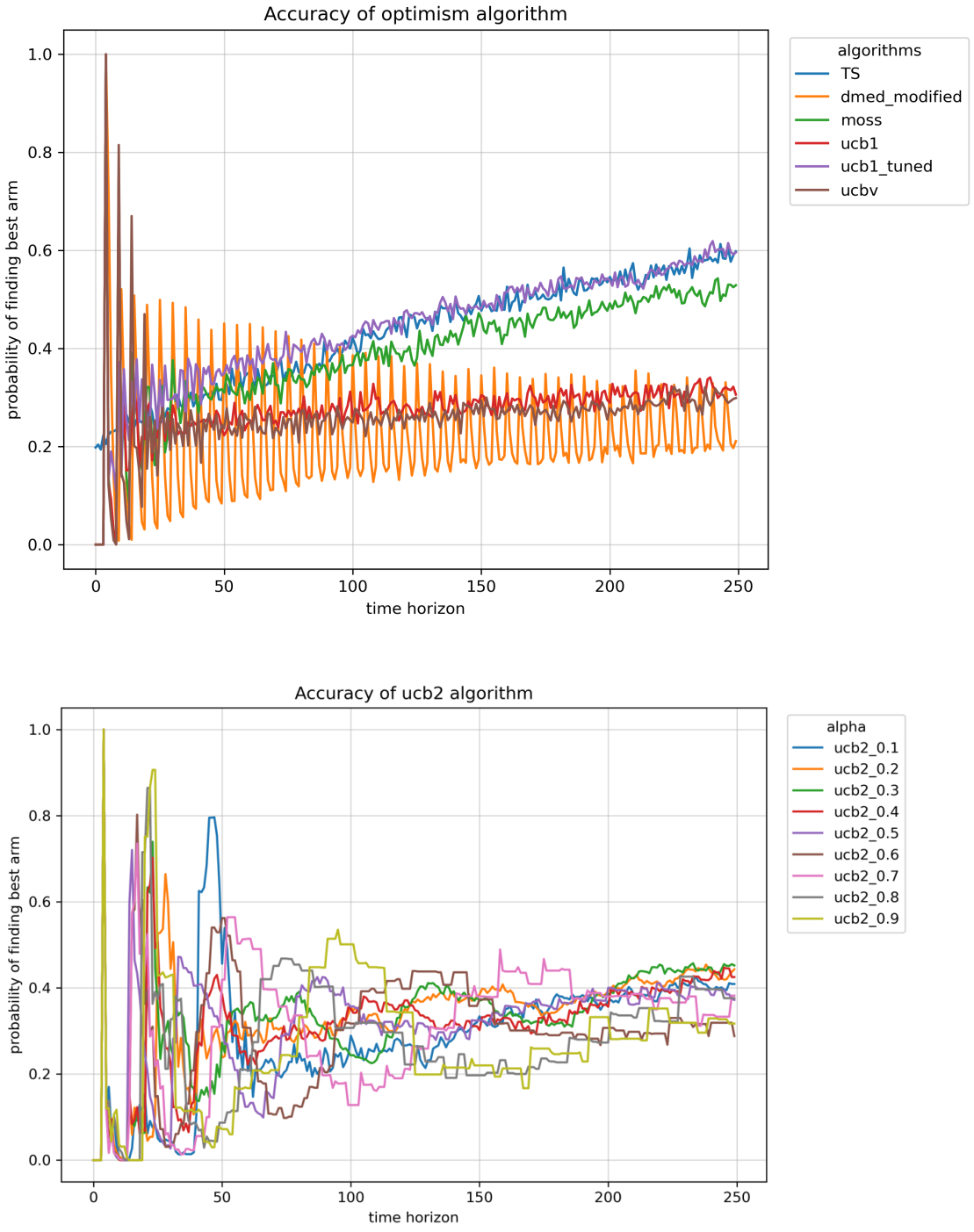


Fig. 26 Accuracy of UCB-type algorithms and Thompson sampling (test scenario 2).

Some of the best performing algorithms in this test scenario were also compared to ETC baseline. As shown by the plot, with a small number of arms (5) and very close arm averages, ETC seems to be the most efficient algorithm. We do note that this limitation is rectified by the introduction of other more advanced algorithms in Section 2.4.3 and 2.4.4.

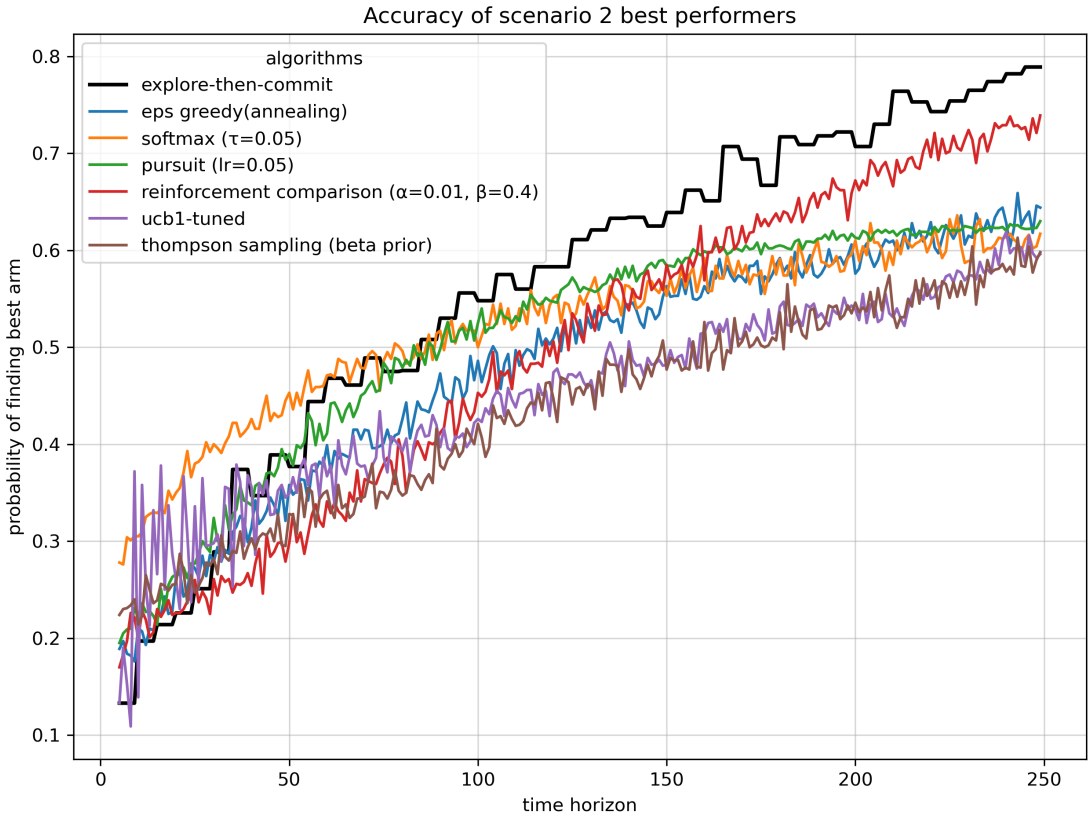


Fig. 27 Accuracy of best-performing algorithms (test scenario 2).

Test scenario 3: 5 Bernoulli arms with probabilities [0.1, 0.25, 0.5, 0.75, 0.9]

This scenario still simulates the situation where a best option is present among five arms, but the average rewards are more evenly distributed between 0 and 1, which makes it slightly more challenging than test scenario 1.

ε-greedy algorithms show similar trends as test scenario 1, with a lower accuracy for all algorithms.

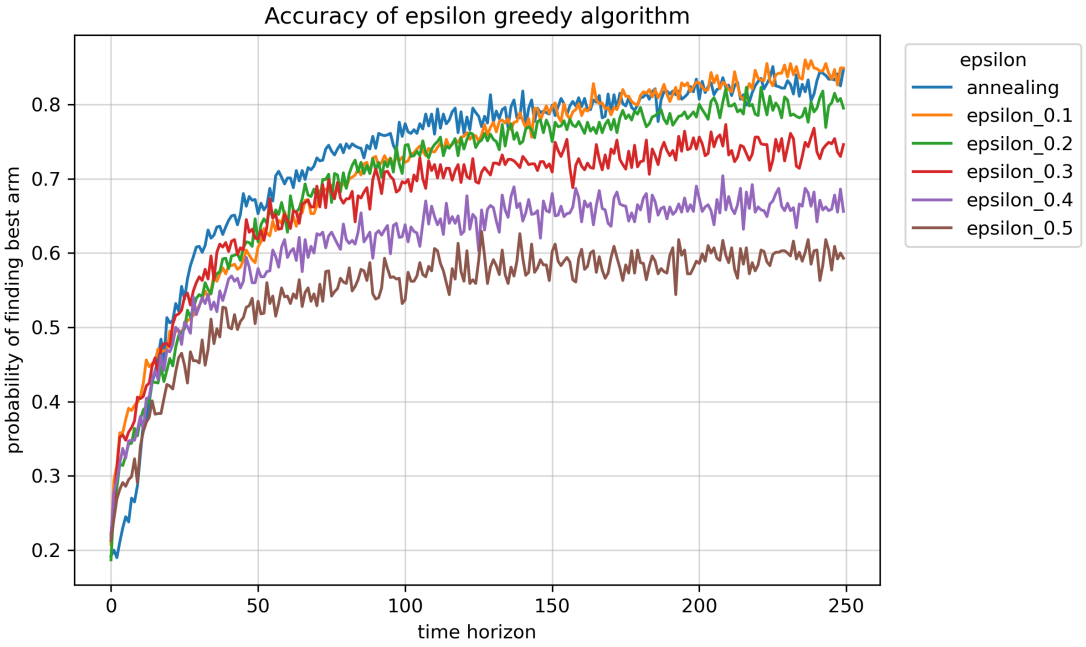


Fig. 28 Accuracy of ε-greedy algorithms with fixed and annealing ε's (test scenario 3).

Softmax algorithm. Annealing softmax in this case actually offers comparable performance to the optimal parameter (τ=0.2), unlike test scenario 1. This shows that annealing softmax can have advantage in a more balanced reward average setting.

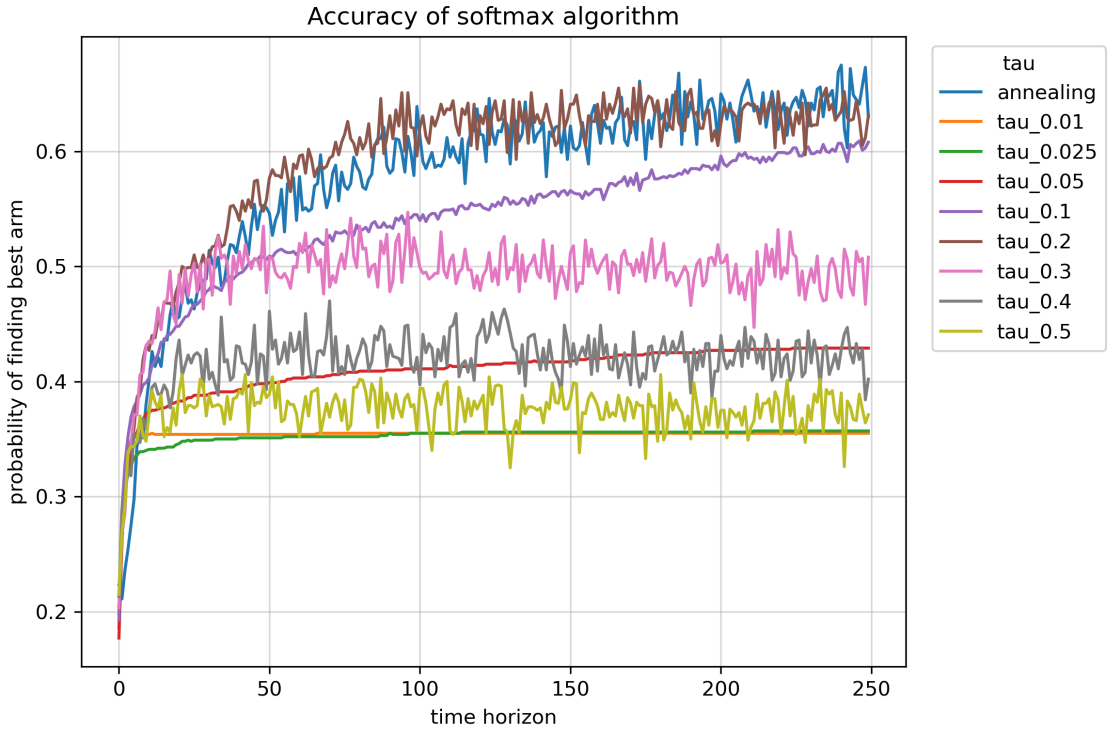


Fig. 29 Accuracy of softmax algorithms with fixed and annealing τ's (test scenario 3).

Pursuit algorithm shows that learning rate has to be small enough (0.025) to converge to a higher accuracy, but big enough (0.05) to achieve good initial accuracy. It will be difficult to properly choose parameter to balance both in practice.

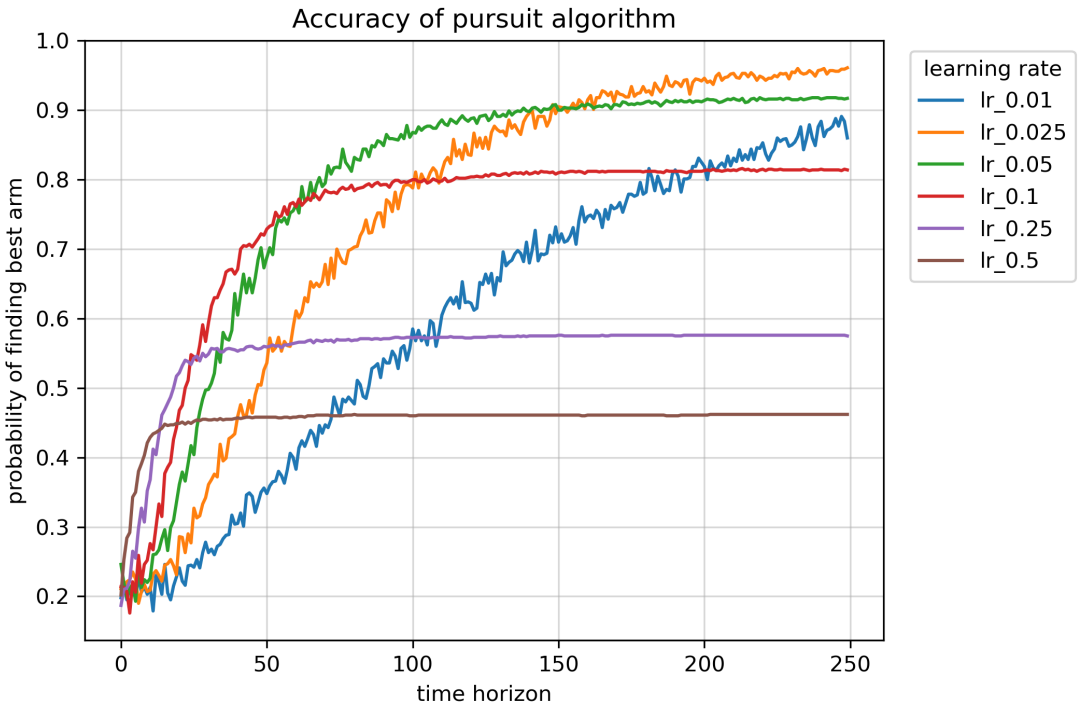


Fig. 30 Accuracy of pursuit algorithms with different learning rates (test scenario 3).

UCB-type algorithms and **Thompson sampling** performed similarly compared to test scenario 1. Thompson sampling and UCB1-tuned appear to be optimal.

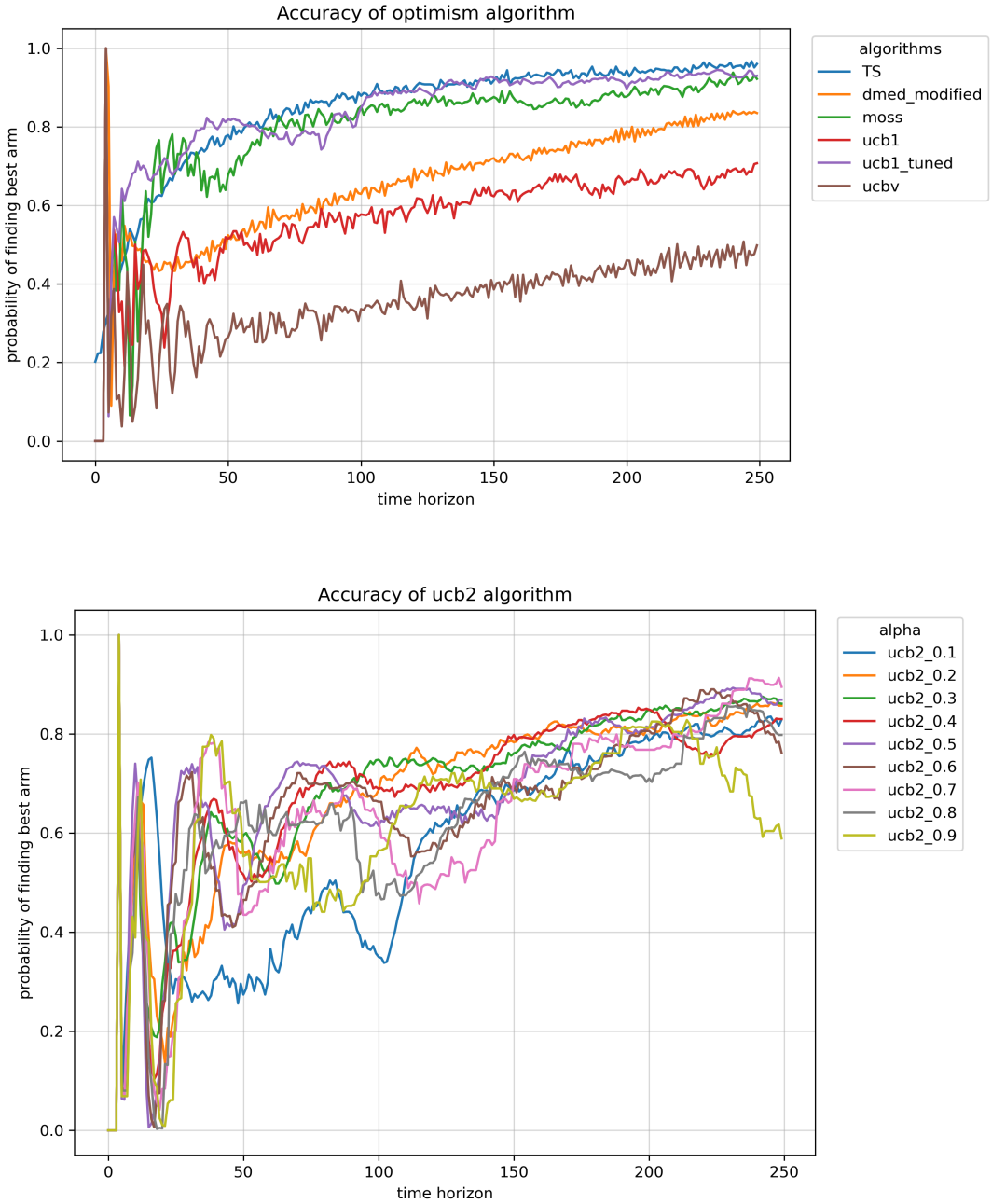


Fig. 31 Accuracy of UCB-type algorithms and Thompson sampling (test scenario 3).

Best performing algorithms in test scenario 3 were again compared with ETC baseline. Compared to test scenario 1, this test scenario is more challenging and top performing algorithms offer advantage over ETC in the initial stages (t<50). However, since there are still only 5 arms in total, ETC quickly catches up and offer comparable performance to the best performing algorithms such as Thompson sampling.

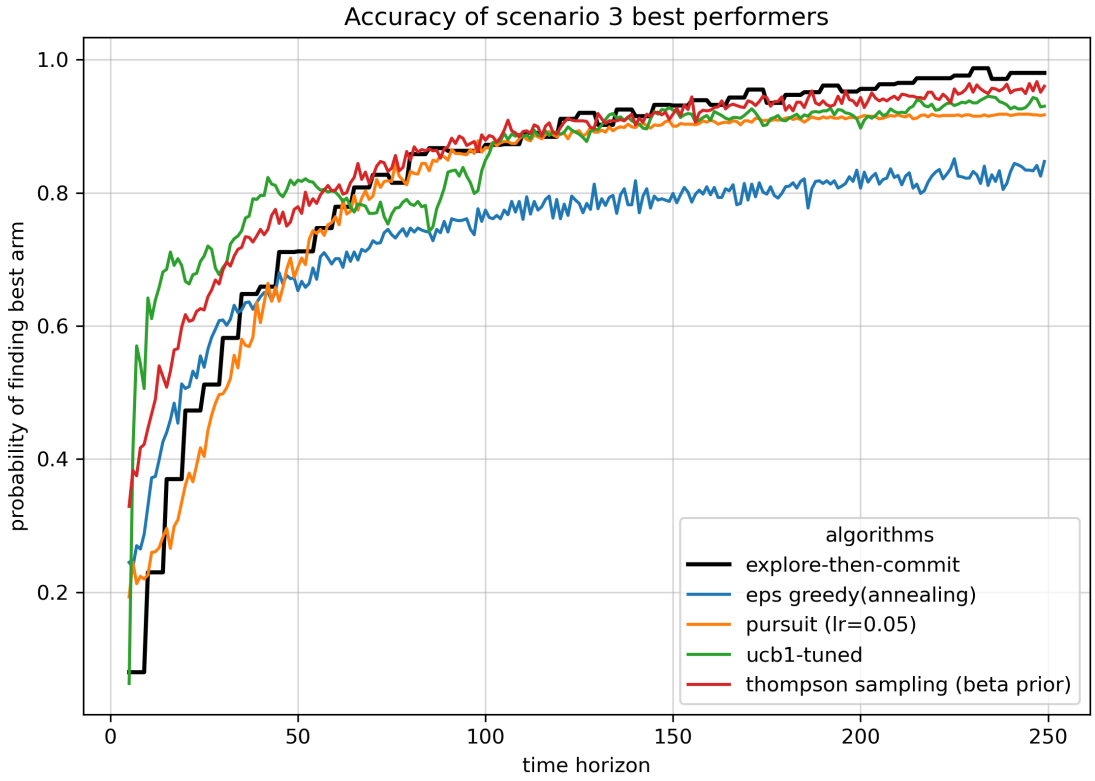


Fig. 32 Accuracy of best-performing algorithms (test scenario 3).

Test scenario 4: 9 Bernoulli arms with probabilities [0.1, 0.2, 0.3, …, 0.8, 0.9]

This scenario simulates the results with evenly distributed average reward similar to test scenario 3, but has more arms compared to scenario 3 to test the effect of increased number of arms. The maximum time horizon is also increased from 250 to 500 to allow algorithms to converge.

Overall, the trends of algorithm performance are very similar to those from test scenario 3, except that the algorithms take longer to converge due to the increased number of arms. ETC algorithms are less efficient in this case due to its uniform exploration of increased number of arms. Top-performing algorithms like Thompson sampling offers definitive advantages.

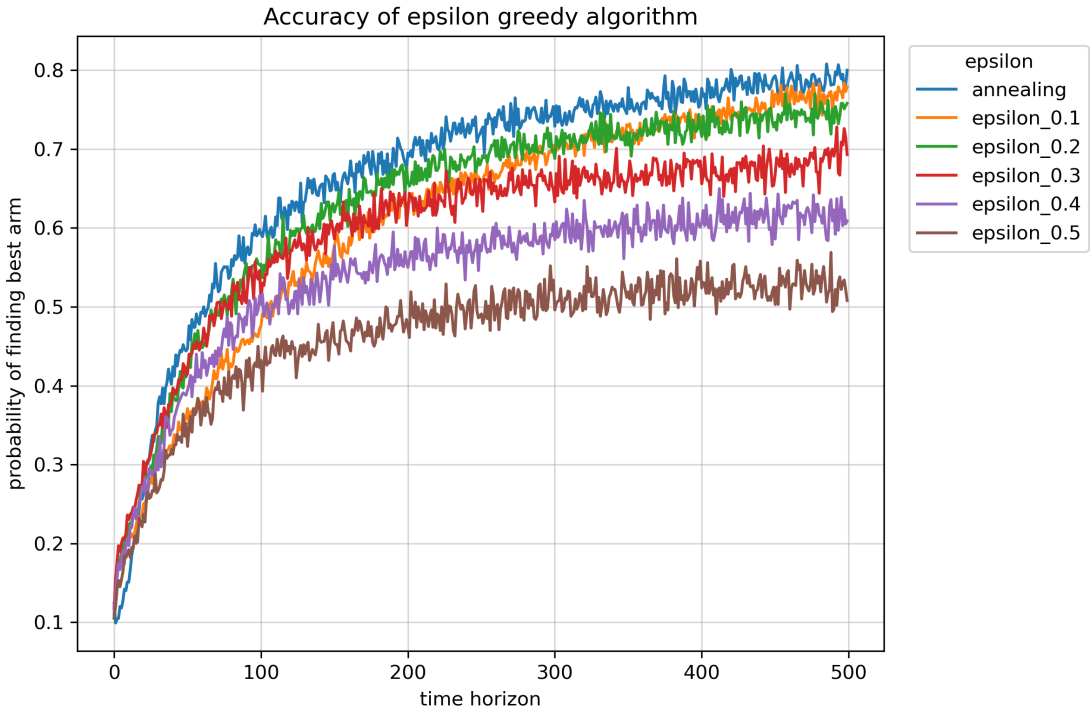


Fig. 33 Accuracy of ε-greedy algorithms with fixed and annealing ε's (test scenario 4).

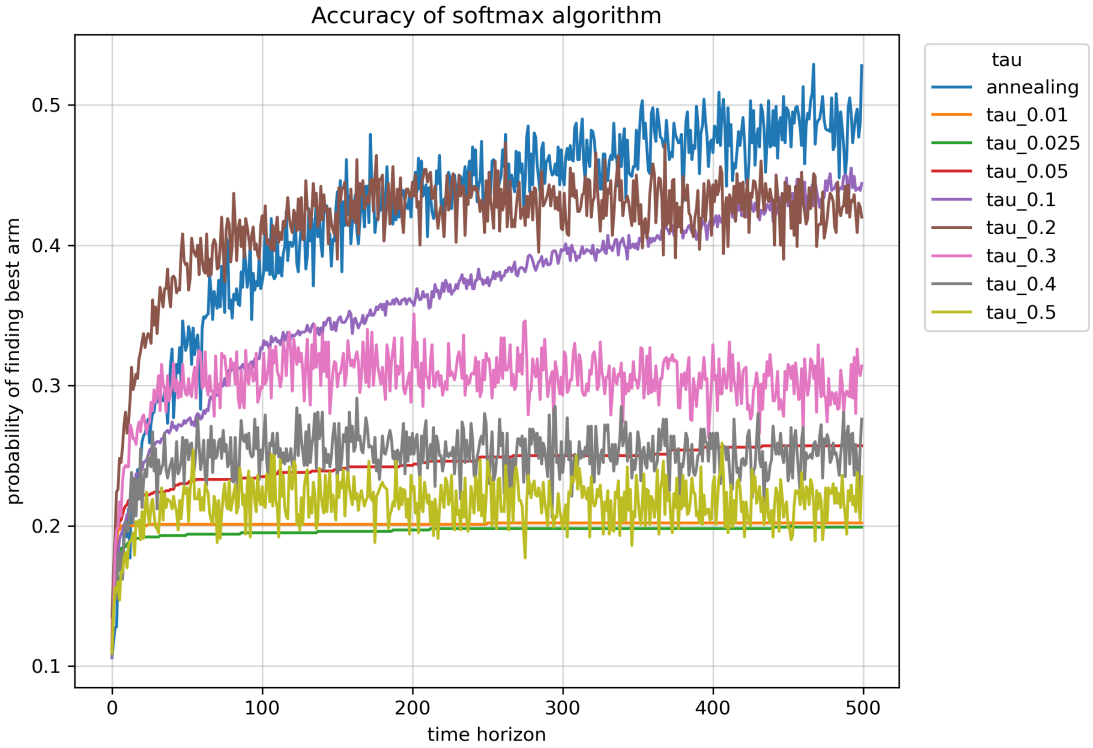


Fig. 34 Accuracy of softmax algorithms with fixed and annealing τ's (test scenario 4).

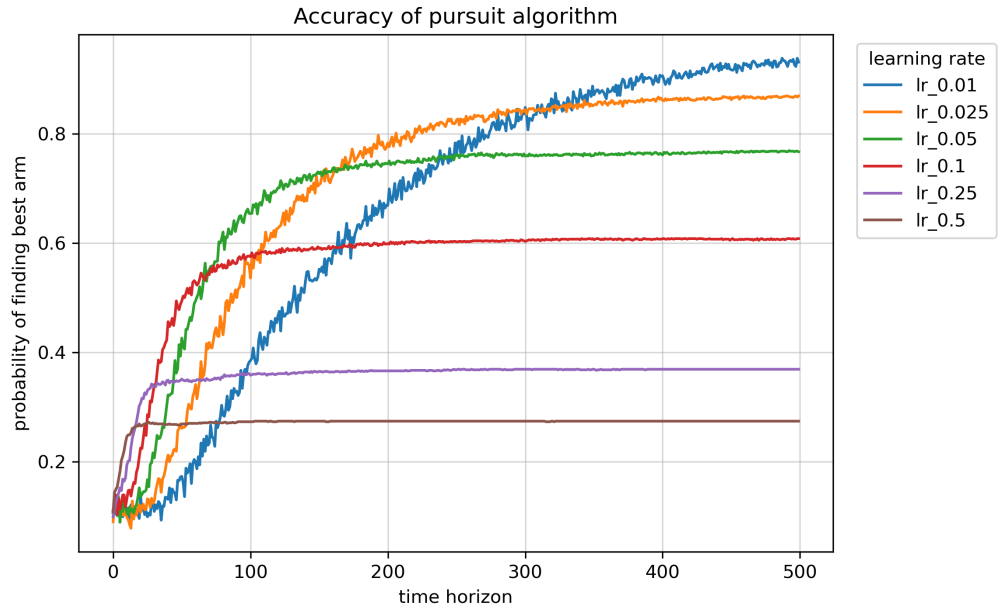


Fig. 35 Accuracy of pursuit algorithms with different learning rates (test scenario 4).

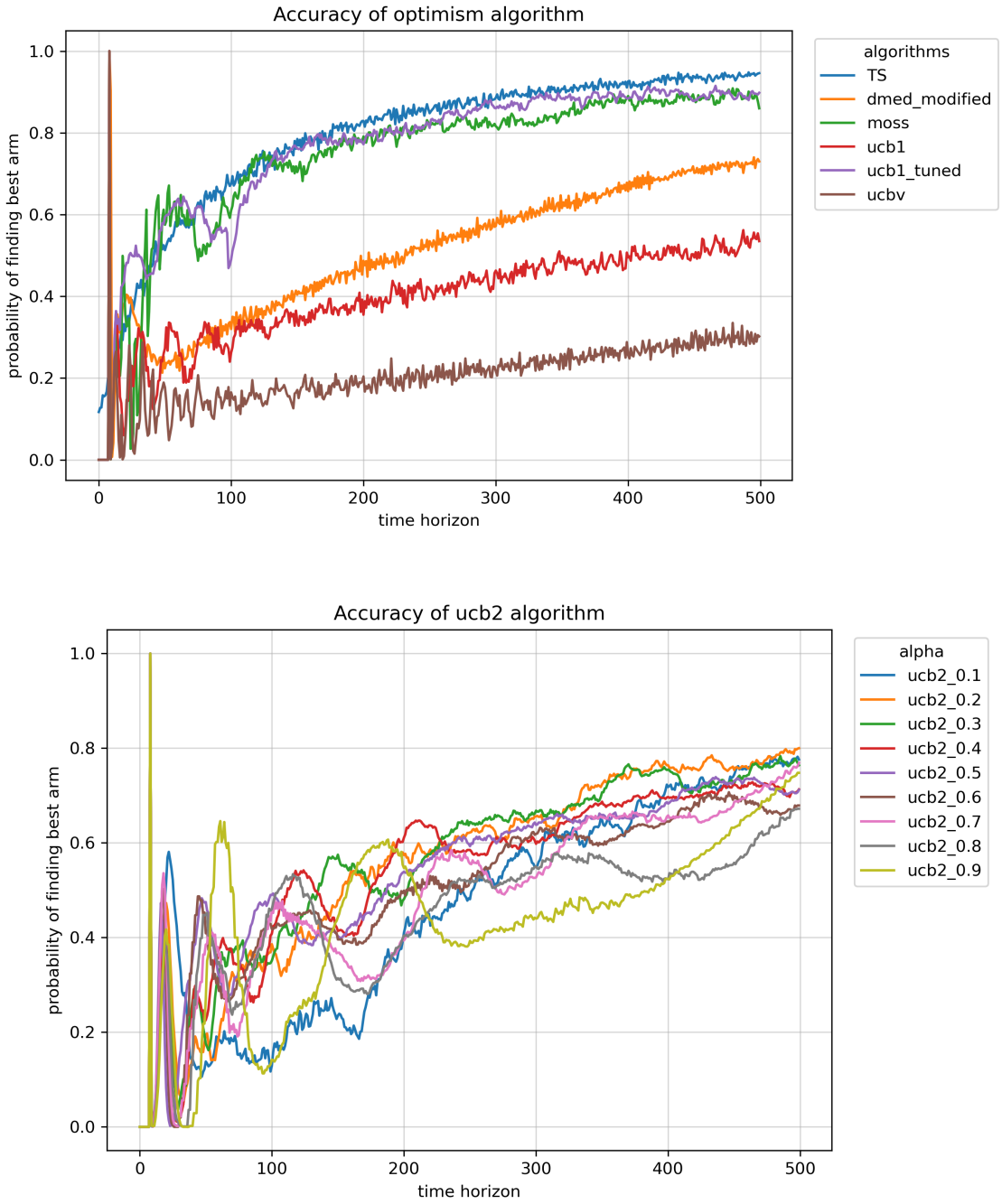


Fig. 36 Accuracy of UCB-type algorithms and Thompson sampling (test scenario 4).

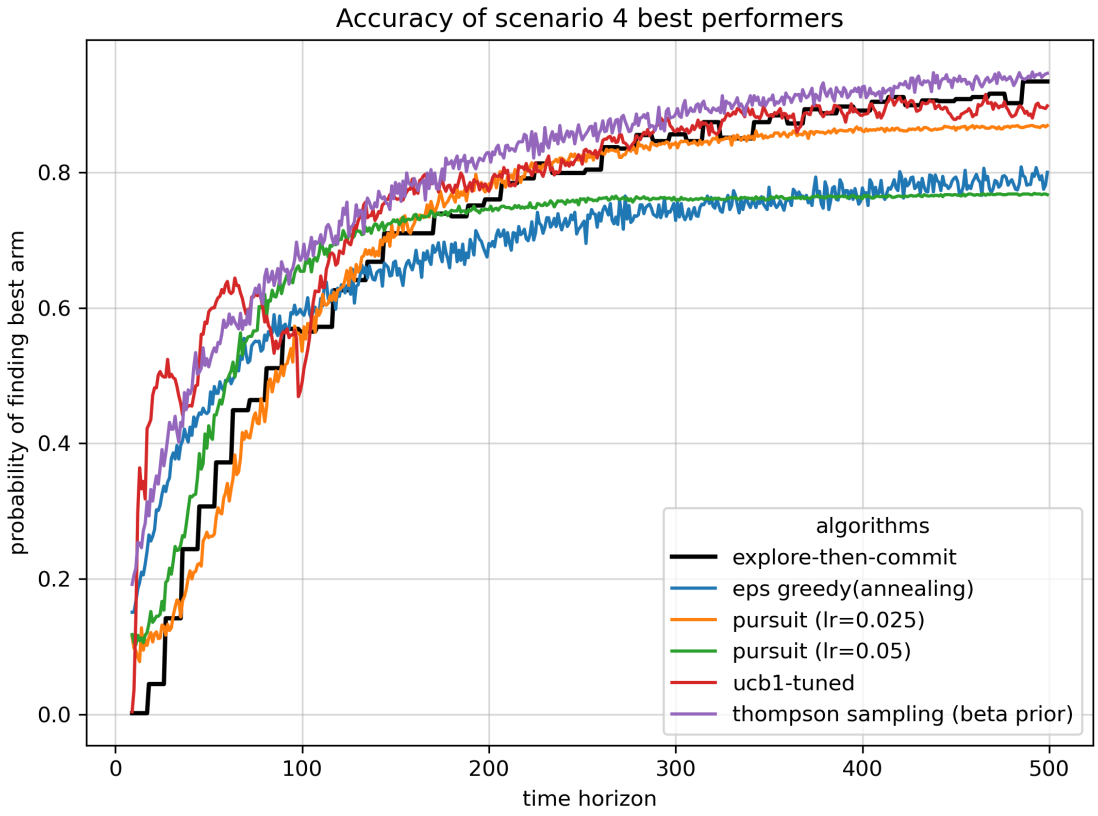


Fig. 37 Accuracy of best-performing algorithms (test scenario 4).

Test scenario 5: 19 Bernoulli arms with probabilities [0.05, 0.1, 0.15, 0.2, …, 0.85, 0.9, 0.95]

This test scenario is similar to test scenario 3 and 4 but further increases the number of arms to 19. The difference in average reward is also smaller (0.05). This test scenario mainly simulates algorithm performance on a bigger scale.

It is worth noting that a lot of the previously effective algorithms are not as effective in this case due to increased number of arms. Thompson sampling is clearly the best algorithm in this case due to its ability to do probability matching and offers significant advantage over ETC baseline in this test scenario.

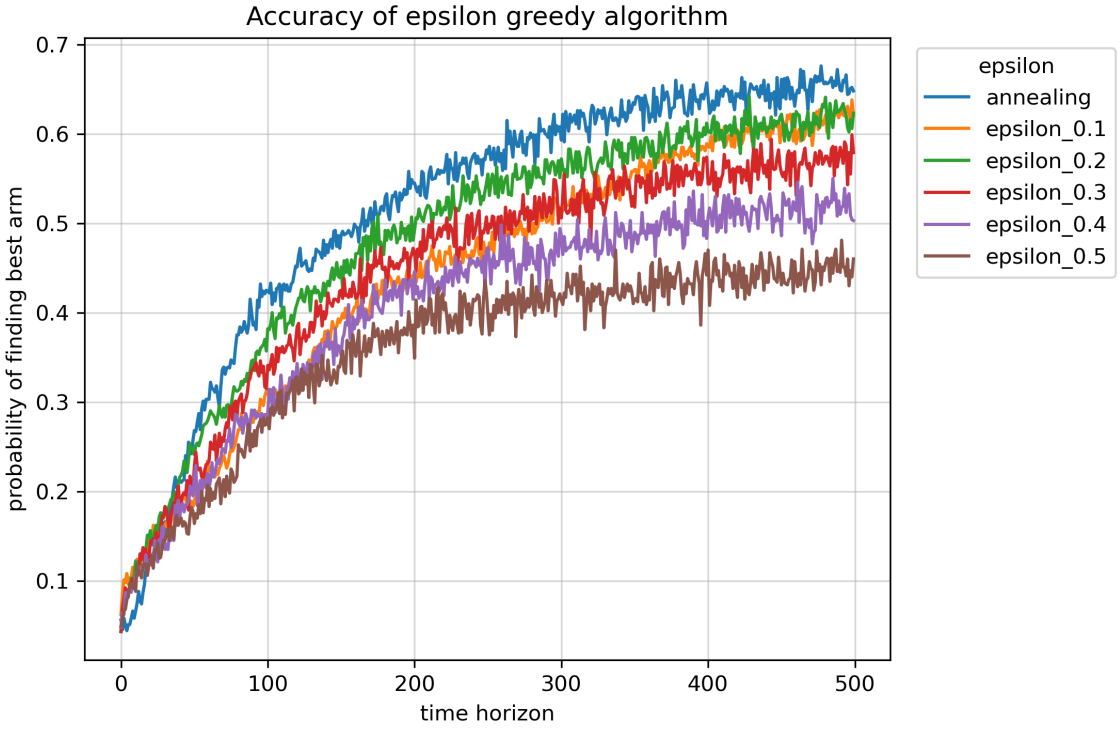


Fig. 38 Accuracy of ε-greedy algorithms with fixed and annealing ε's (test scenario 5).

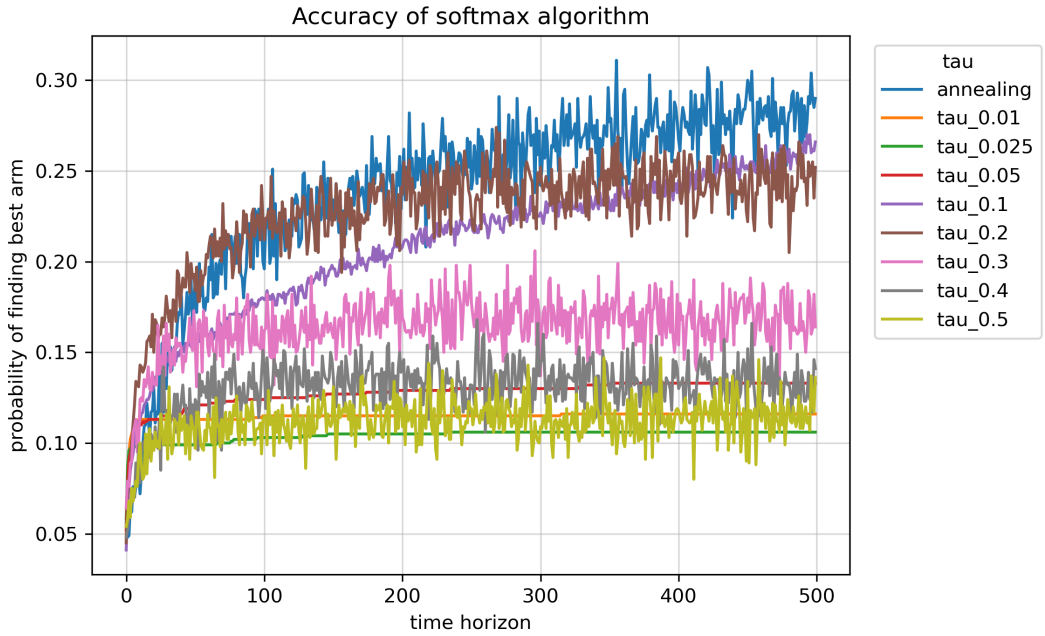


Fig. 39 Accuracy of softmax algorithms with fixed and annealing τ's (test scenario 5).

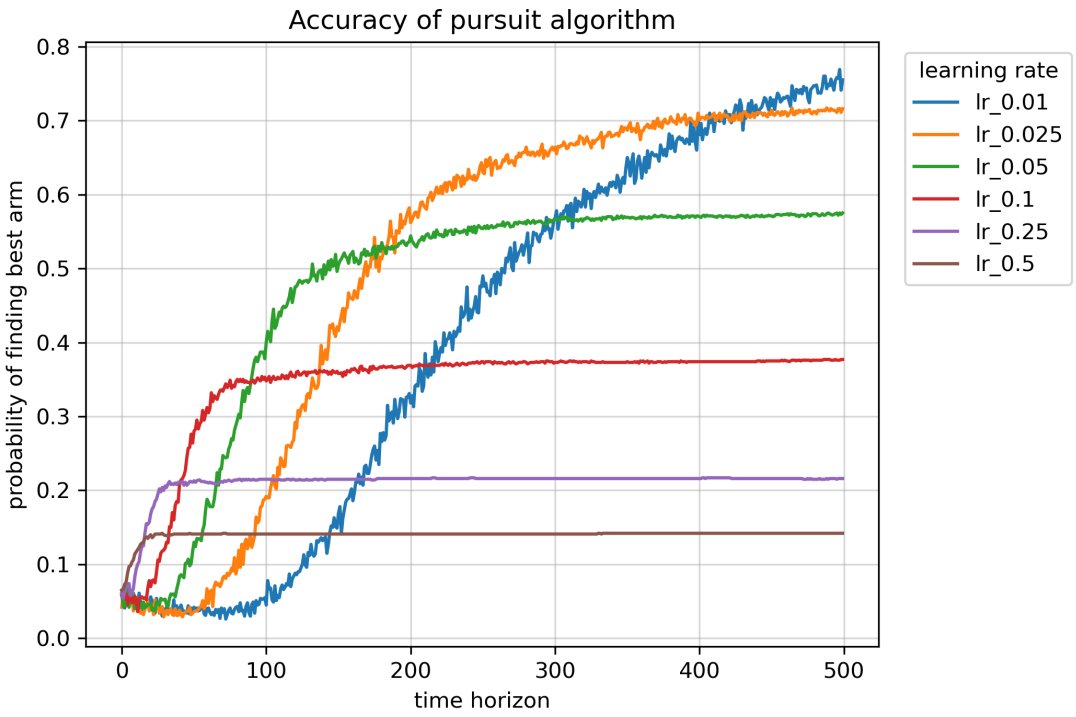


Fig. 40 Accuracy of pursuit algorithms with different learning rates (test scenario 5).

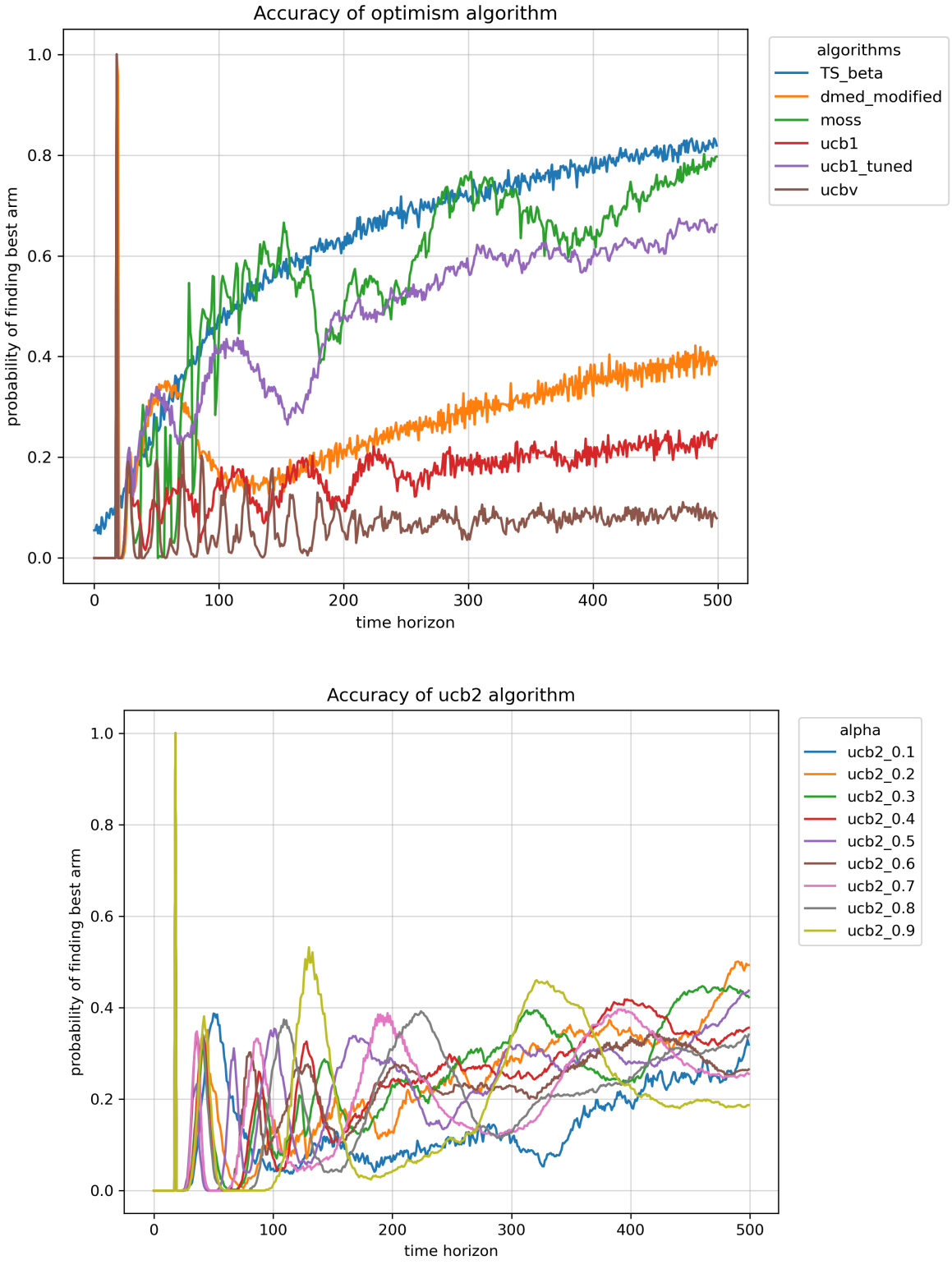


Fig. 41 Accuracy of UCB-type algorithms and Thompson sampling (test scenario 5).

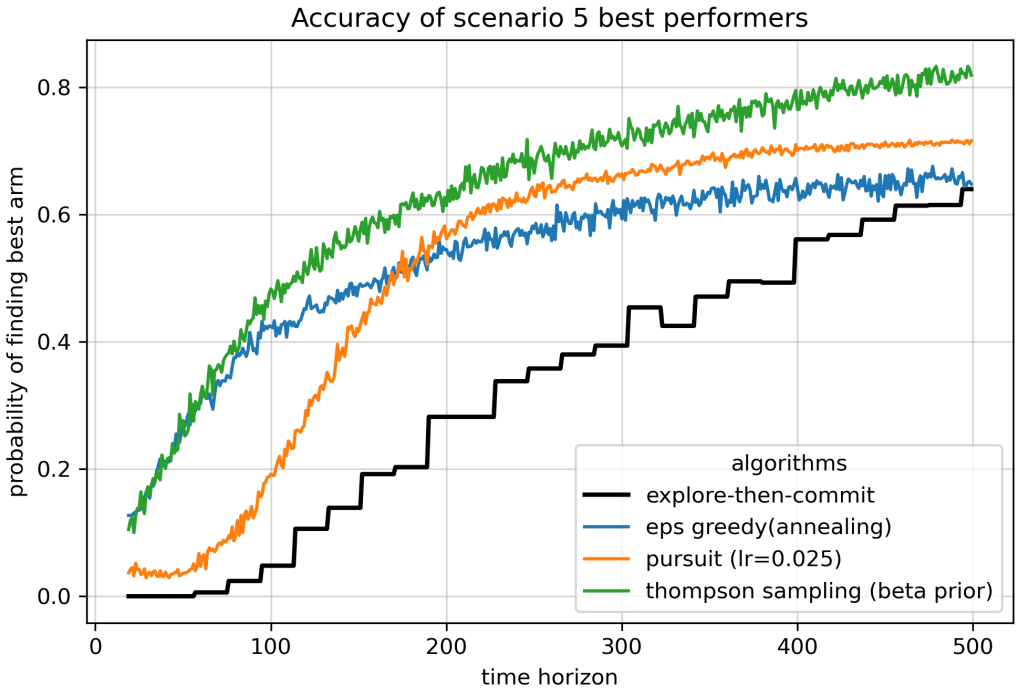


Fig. 42 Accuracy of best-performing algorithms (test scenario 5).

2.4.3 Bandit algorithm modifications: Thompson sampling algorithms with normal priors

For all five test scenarios described above, Thompson sampling seems to be the optimal algorithm. However, in all these test cases Thompson sampling is implemented with a beta conjugate prior since each arm is represented with a Bernoulli distribution. For a real-world chemistry reaction dataset, the reward is a bounded continuous variable. For other algorithms where only empirical means and numbers of arm pulls are considered in updating the algorithm, continuous reward bounded between 0 and 1 (as in the case of reaction yield) can still be used. However, for Thompson sampling with beta prior, the update of beta distribution explicitly uses the binary result (success/fail) and cannot be directly used with reaction yields (or other relevant

metrics). To address this limitation, we implemented Thompson sampling with other conjugate priors under different assumptions about distributions.[59]

Gaussian distribution (unknown mean, known variance)

Assuming the underlying distribution for each arm is a Gaussian distribution with unknown mean but known variance, a gaussian conjugate prior can be used. Assume the variance is $\boldsymbol{\sigma^2}$, the Thompson sampling procedure at each time point $\boldsymbol{t}$, for each arm $\boldsymbol{i}$, will sample $\boldsymbol{\theta_i}$ from:

$$\mathcal{N}(\hat{\mu}_{i,t-1}, \frac{1}{\frac{n_{i,t-1}}{\sigma^2} + \frac{1}{\sigma_0^2}})$$

where σ^2 is the sample variance and σ_0^2 is the variance of the prior.

Assuming $\sigma^2 = \sigma_0^2$, this can be simplified to:

$$\mathcal{N}(\hat{\mu}_{i,t-1}, \frac{\sigma_0^2}{n_{i,t-1} + 1})$$

After sampling, algorithm will then play arm:

$$I_t := \arg\max_i \theta_i$$

The assumption of fixed variance is often difficult to validate in practice, and it is important to note the different effects of over-/under-estimating the variance. We tested this version of Thompson sampling with four test scenarios. In each of the four test scenarios, there are five normally distributed arms with the same fixed means [0.1 ,0.2, 0.3, 0.4, 0.9], but the standard deviations for all arms in these four scenarios are set to 0.1, 0.25, 0.5, 0.75, respectively. For the Thompson sampling algorithm with Gaussian prior, the standard deviation assumption is set to 0.1, 0.25, 0.5, 0.75, and 1 to test the effect of different variance assumption. The accuracy of selecting the correct optimal arm for these three test cases are plotted in **Fig. 43**. Assuming a standard

deviation of 0.25 seems to work better for data with low standard deviations, while assuming a standard deviation of 0.5 offers comparable and even better performance for high standard deviation data settings.

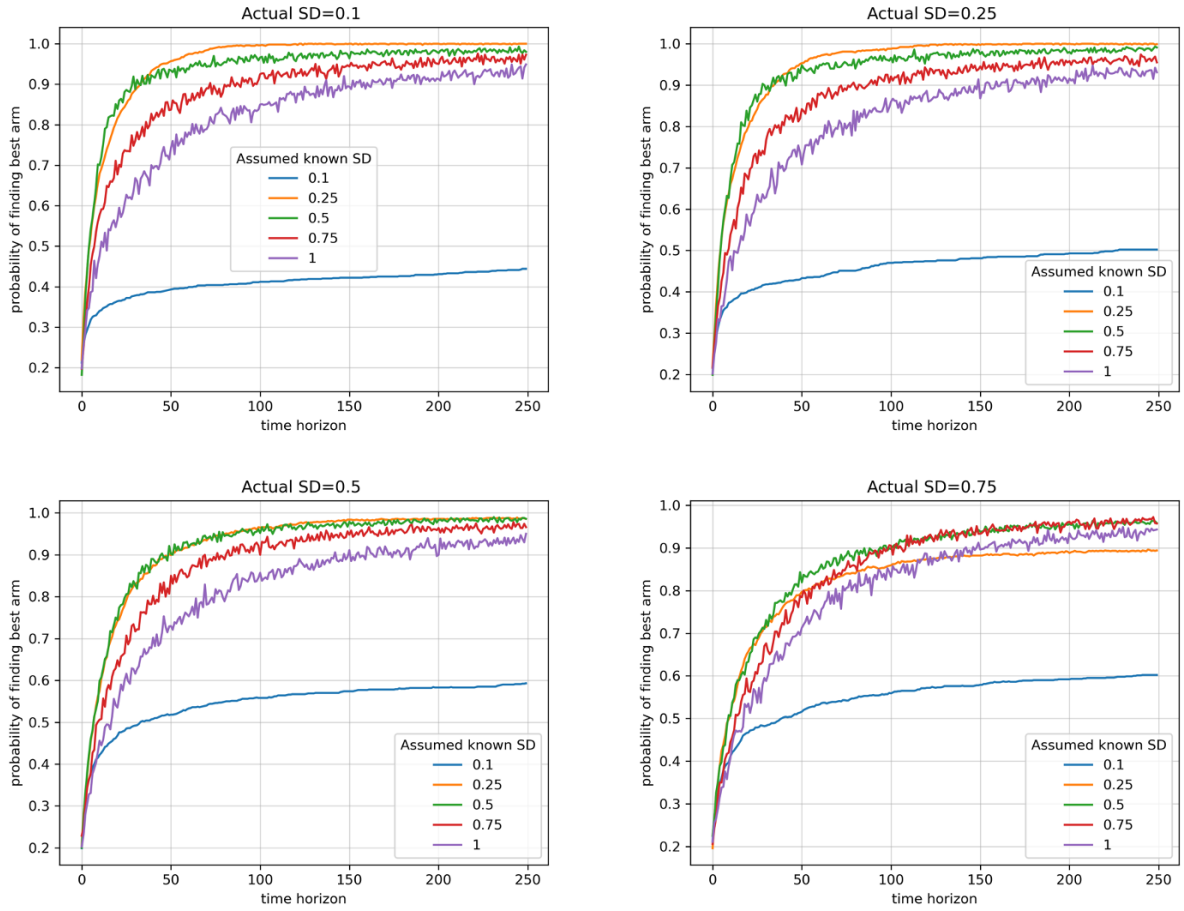


Fig. 43 Accuracy of selecting the best arm with Thompson sampling, assuming different standard deviations, under three different test scenarios each with normally distributed data (same averages, different standard deviation at 0.1, 0.25, 0.5, 0.75).

It has also been suggested that this algorithm can be used not only for Gaussian multi-armed bandit problem, but also general stochastic MAB problems. Problem-specific regret bound has also been proved for general stochastic MAB problems.[60] In our simulations, testing with the

five previously discussed test scenarios in Section 2.4.2, none of the standard deviation settings for gaussian prior performed better than beta prior (**Fig. 44**). However, fortuitously (through a mistake in a previous implementation), we discovered that good results can be obtained by sampling from a slightly different posterior:

$$\mathcal{N}(\hat{\mu}_{i,t-1}, (\frac{1}{n_{i,t-1}+1})^2)$$

This implementation Thompson sampling (referred to as "squared" in the plots) offers similar (or sometimes better) accuracy when used for arms with Bernoulli distribution. It is especially worth noting that we significantly increased accuracy in challenging test scenario 2, where all other algorithms failed to beat ETC baseline. Results of using TS with gaussian prior in the Bernoulli test scenarios are plotted and compared to TS with beta prior in **Fig. 44**.

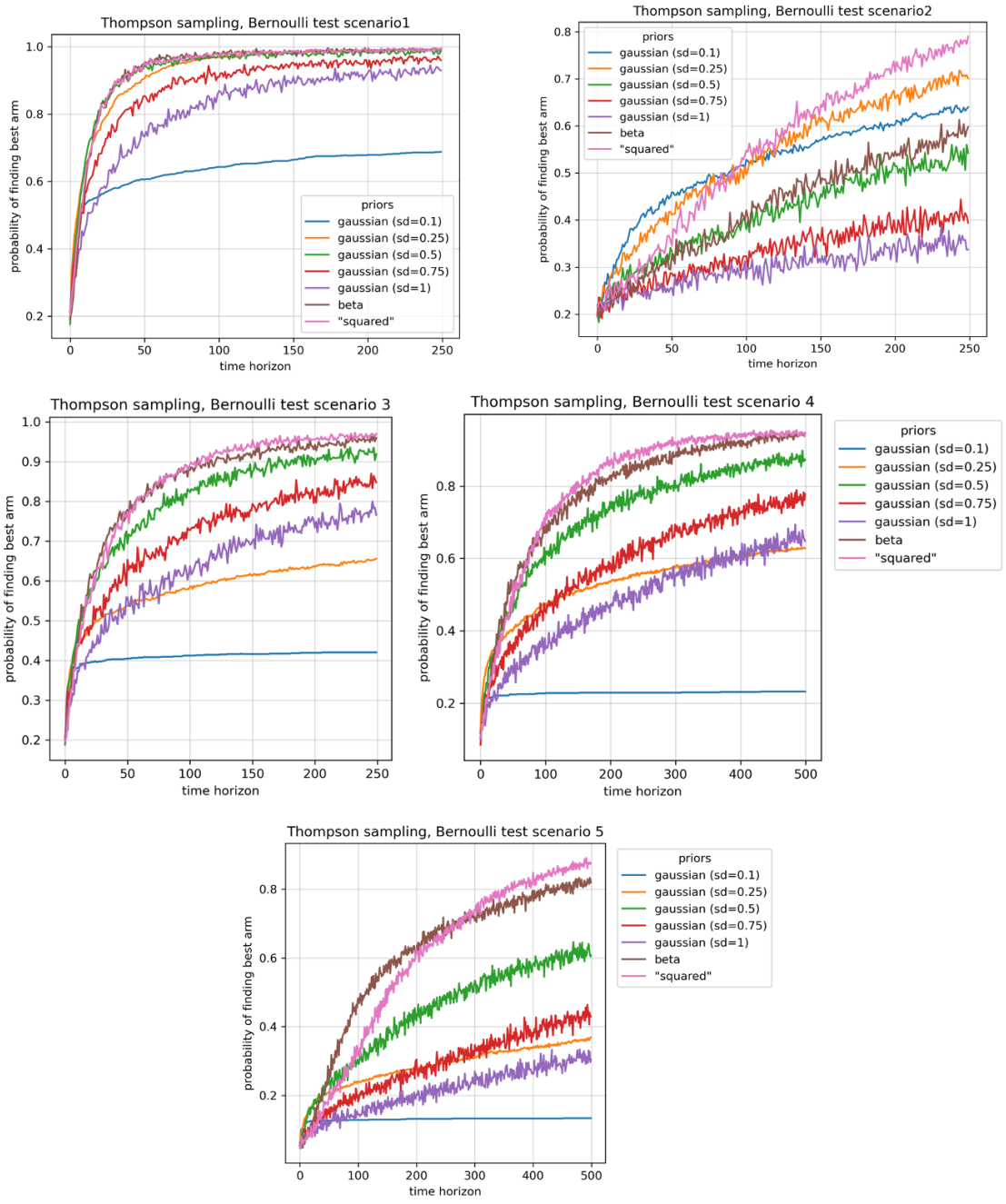


Fig. 44 Comparing performance of Thompson sampling with beta prior and gaussian (normal) prior in five Bernoulli test cases described in Section 2.4.2.

Overall, for arm rewards with a normal distribution, Thompson sampling assuming fixed standard deviation of 0.25 or 0.5 works well, with 0.5 being the most versatile choice. For arm

rewards with a Bernoulli distribution, using a squared variance term for posterior update seems to match the performance (or outperform) Thompson sampling with beta prior.

Gaussian distribution (known mean, unknown variance)

Gaussian conjugate prior for this case also exists, but is less suitable for our problem, as we are primarily interested in estimating the different means for all arms.

Gaussian distribution (unknown mean, unknown variance)

Assuming the underlying distribution for each arm is a Gaussian distribution with unknown mean and unknown variance, a gaussian-gamma conjugate prior can be used (more precisely speaking, this conjugate prior is used when precision is unknown, which is the inverse of variance).[61] The sampling procedure at each time point ***t***, for each arm ***i***, will sample $\boldsymbol{\theta_i}$ from a normal-gamma distribution:

$$\mathcal{N}(\hat{\mu}_{i,t-1}, \frac{1}{\Gamma(\alpha_{i,t-1}, \beta_{i,t-1})})$$

and play arm:

$$I_t := \arg\max_i \theta_i$$

The posterior parameters α and β are updated as follows:

$$\alpha_{i,t} = \alpha_{i,t-1} + \frac{n}{2}$$

$$\beta_{i,t} = \beta_{i,t-1} + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{n\nu}{n+\nu}\frac{(\bar{x} - \mu_0)^2}{2}$$

where n is the number of samples drawn, x_i is each individual sample, μ_0 is empirical mean, and v is the overall number of samples drawn to derive empirical mean. In the case of MAB problems where one arm is pulled once at each time horizon, n=1 is used, the middle sum term equals to zero and $\bar{x}$ is simply the reward.

As for the test case for the implemented algorithms, three test scenarios were used. The means of five arms are fixed at [0.1 ,0.2, 0.3, 0.4, 0.9], but the standard deviations were set to 0.5, 1, and 1.5 respectively. The difference here is that Thompson sampling no longer assumes a fixed variance of 1. The accuracy of selecting the best arm is plotted in **Fig. 45**. The results seem to suggest that the ability to estimate variance does not necessarily translate to higher accuracy.

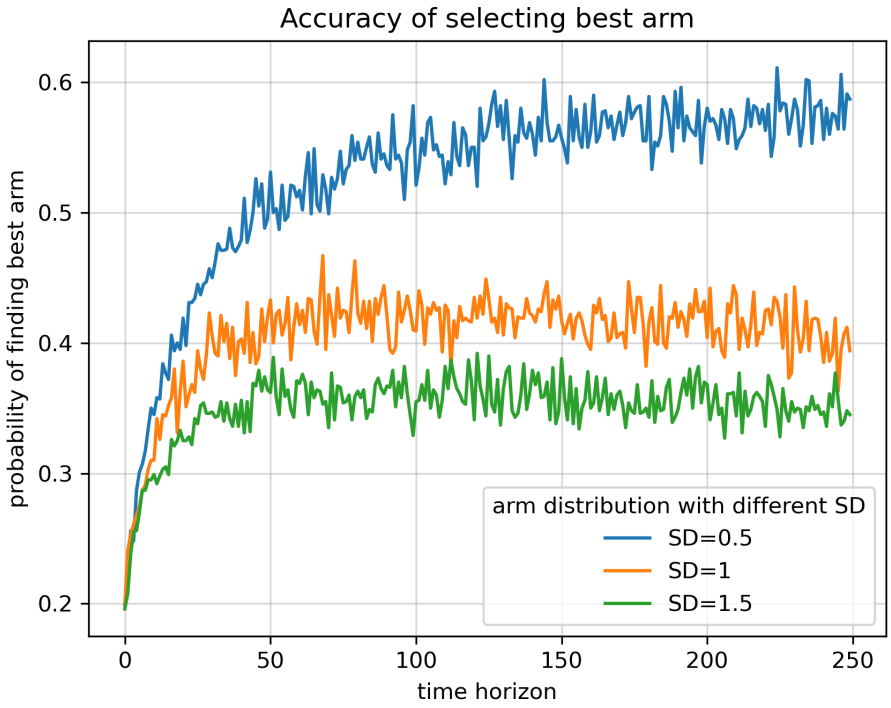


Fig. 45 Accuracy of Thompson sampling (unknown mean, unknown variance) in three test scenarios with normally distributed reward (same mean, different standard deviation)

A closer examination of our implementation and simulation results points the problem to parameter initialization. In this case, α and β for all arms are initialized to 1. This results in a high initial estimate of standard deviation for all arms (~0.7). It is difficult to lower uncertainty for arms that have lower mean quickly, due to the limited number of samples drawn for these arms with low means. At the same time, high uncertainty prompts algorithm to do unnecessary exploration of arms with low means. These factors contributed to the limited accuracy with this algorithm.

Since the yield data we are trying to model falls in the interval [0,1], we used a low standard deviation test scenario to identify suitable initialization for β parameter. With the same five individual mean for five arms as before, [0.1, 0.2, 0.3, 0.4, 0.9], and all arms' standard deviation set to 0.1, we again plot the accuracy of selecting the best arm with different β initialization (**Fig. 46**). Initializing β to 0.1 seems to give the most optimal accuracy.

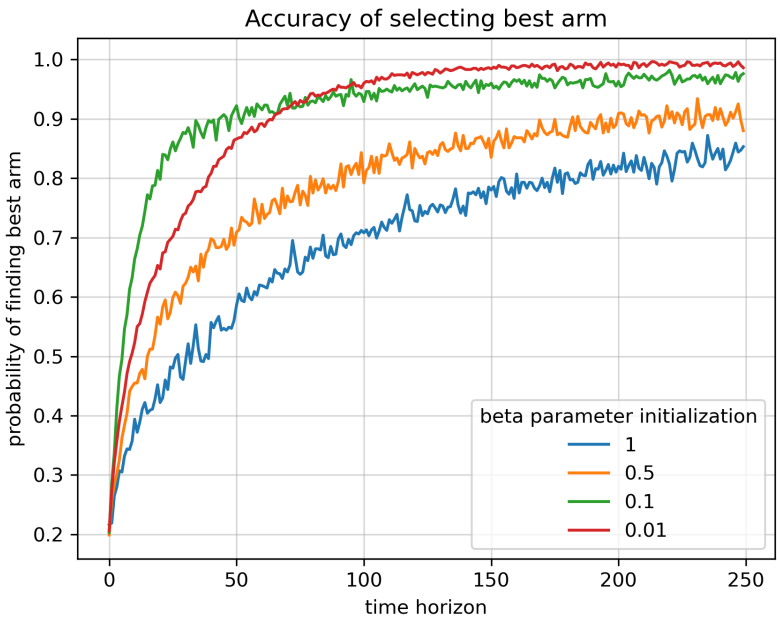


Fig. 46 Tuning β initializations with a low standard deviation (0.1) normal reward test case.

For the same test scenario, if standard deviation for all arms is set to 0.5 (usually above the maximum standard deviation we would observe in a real reaction dataset), initializing β to 0.01 offers a slight advantage compared to 0.1 initialization, but both have similar initial accuracy and converge at about the same time (**Fig. 47**). We opted to use 0.1 as default initial value for β.

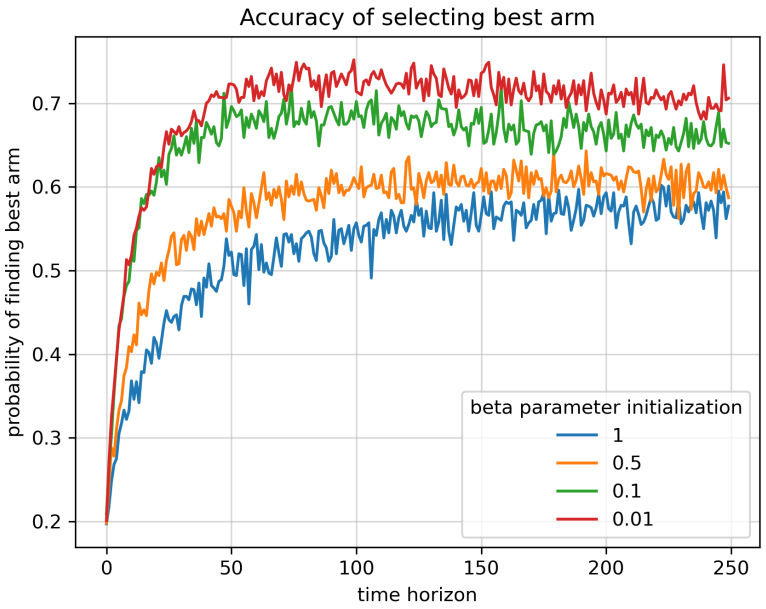


Fig. 47 Tuning β initialization with a high standard deviation (0.5) normal reward test case.

Unlike the previous implementation of Thompson sampling assuming fixed variance, this version of Thompson sampling does not appear to work well for the Bernoulli arm test cases.

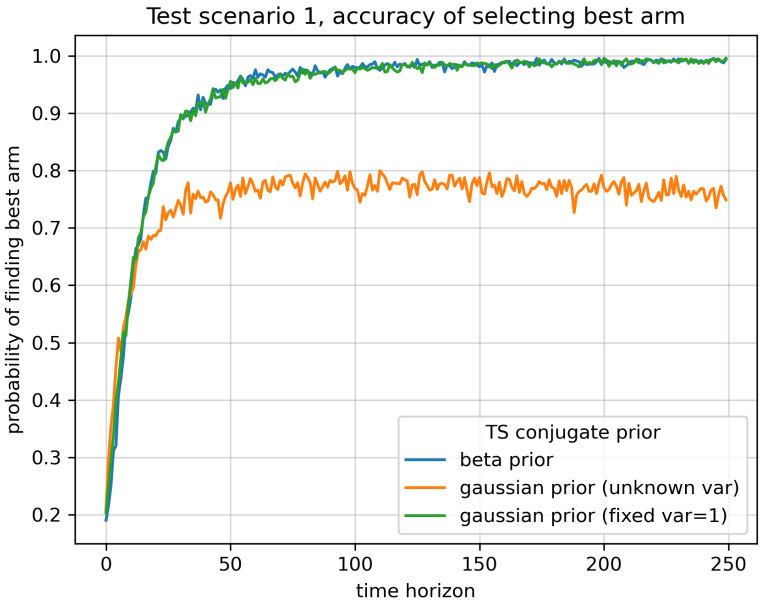


Fig. 48 Comparing three versions of Thompson sampling in Bernoulli test scenario 1

Note: gaussian prior (fixed var=1) uses the "squared" implementation discussed.

Finally, we tested two versions of the Thompson sampling algorithms (assuming fixed variance and not assuming fixed variance) in a test scenario with arms with normally distributed rewards, each with different variances. Five arms are present in this test scenario with [means, standard deviations]: [0.1, 0.2], [0.3, 0.4], [0.5, 0.3], [0.7, 0.1], [0.9, 0.2] (**Fig. 49**).

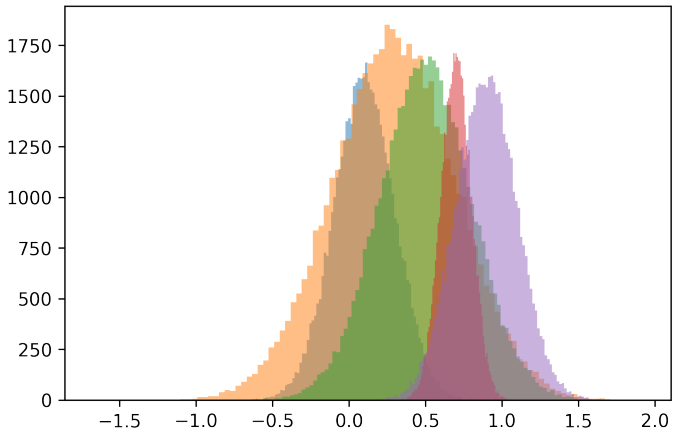


Fig. 49 Visualization of the test case with five arms, each with normally distributed rewards with specified mean and standard deviation.

Simulation results are shown in **Fig. 50**. We tested both unbounded rewards for all arms (left), and bounded [0,1] reward (right) by setting any reward lower than 0 to 0 and setting any reward higher than 1 to 1. We also used ε-greedy algorithm with annealing exploration rate, an effective algorithm demonstrated by Bernoulli arm testing, as a benchmark comparison.

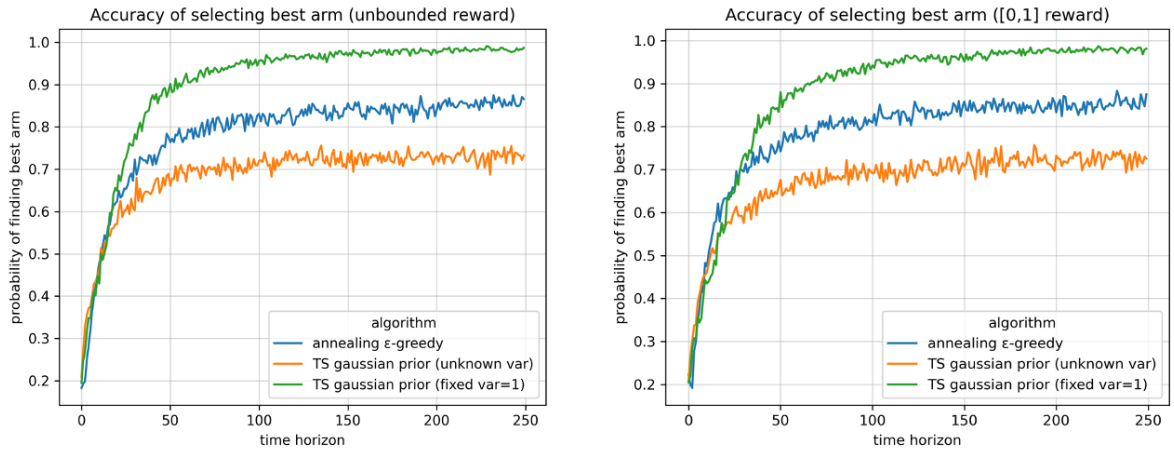


Fig. 50 Comparing the accuracy performance of two versions of Thompson sampling implemented for normal rewards, with ε-greedy as a baseline. Unbounded rewards (left) and [0,1] bounded rewards (right) are both tested.

Thompson sampling with gaussian prior assuming a fixed variance seems to work best for both these situations, outperforming the annealing ε-greedy benchmark. Thompson sampling assuming unknown variance does not seem to offer any advantage, despite its ability to estimate the variance of each arm's distribution. However, plotting the cumulative reward for both situation shows that this algorithm is still capable of achieving the same level cumulative reward performance compared to ε-greedy (**Fig. 51**).

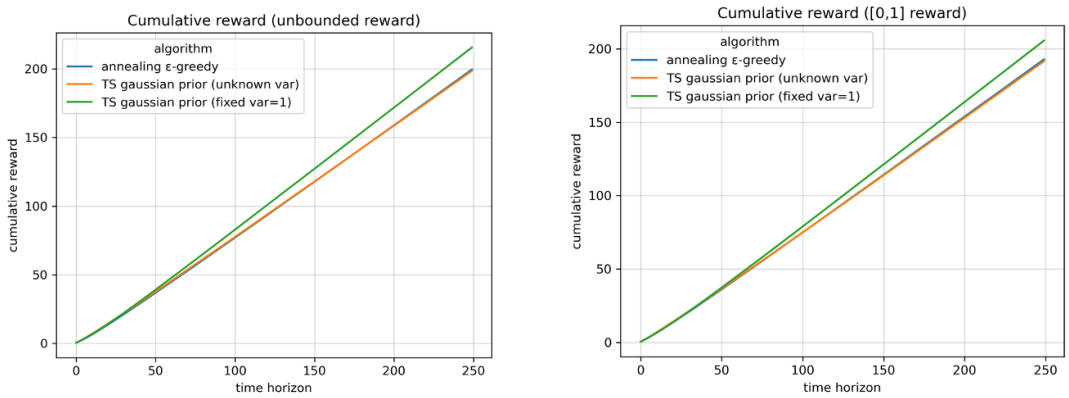


Fig. 51 Comparing the cumulative reward performance of two versions of Thompson sampling implemented for normal rewards, with ε-greedy as a baseline. Unbounded rewards (left) and [0,1] bounded rewards (right) are both tested.

Our testing demonstrates that Thompson sampling with normal conjugate prior assuming a fixed variance is sufficient and often out-performs more complicated normal-gamma conjugate prior which can estimate variance for arms with normal distributions. This algorithm can also be used for arms with Bernoulli distributions and can outperform or offer similar accuracies. However, Thompson sampling with normal-gamma conjugate prior still offers good performance when evaluated with cumulative reward as a metric, so it still warrants consideration when choosing the appropriate algorithm.

Other distributions

We also considered modeling arms as beta distribution or gamma distribution, but the conjugate priors for these distributions can be difficult to derive and compute. These cases are outside the scope of this study and the authors' knowledge in this matter.

2.4.4 Bandit algorithm modifications: Bayesian UCB algorithms with Beta and Normal priors

Not satisfied by the small performance advantages most algorithms have over explore-then-commit baseline, we tried to identify a more effective algorithm. One such algorithm that was later developed after initial synthetic data benchmarking is Bayesian upper confidence bound (UCB) algorithms. Similar to other upper confidence bound algorithms, Bayesian UCB algorithm maintains an upper confidence bound for each arm and selects the arm with the highest UCB value. But unlike other UCB algorithms, Bayesian UCB maintains a prior distribution (which is similar to Thompson sampling) that gets updated after each update. Bayesian UCB uses a fixed quantile function of the posterior distribution as confidence interval and an estimate of uncertainty. We implemented different versions of Bayesian UCB with beta prior and gaussian prior.

Bayesian UCB with standard deviation as confidence bound (beta prior)

The first version of Bayesian UCB simply uses standard deviation as confidence bound, the length of which is controlled by a tunable parameter, *c*. At each update, UCB values are updated as follows:

$$\mathrm{UCB}_j = \bar{x}_j + c\bar{\sigma}_j$$

We first evaluated the different lengths of confidence interval considered in UCB calculations and their effect on optimization. The confidence intervals in Bayesian UCB algorithms are controlled by the number of standard deviations considered. We first tested Bayesian UCB with beta conjugate prior in Bernoulli test scenario 1, 2 and 3 (Section 2.4.2), while considering one, two or three standard deviations as confidence intervals (**Fig. 52**).

For test scenario 1, 1 SD and 2 SD confidence intervals perform very similarly. For test scenario 2 where the means are all very similar, a 1-SD confidence interval seems to be beneficial. For test scenario 3, 2 SD and 3 SD confidence intervals converge to the same accuracy over time, but 3 SD lacks some initial performance. Based on these results, we will use 1 SD confidence interval when all arm means are expected to be very similar, and 2 SD confidence intervals for all other cases.

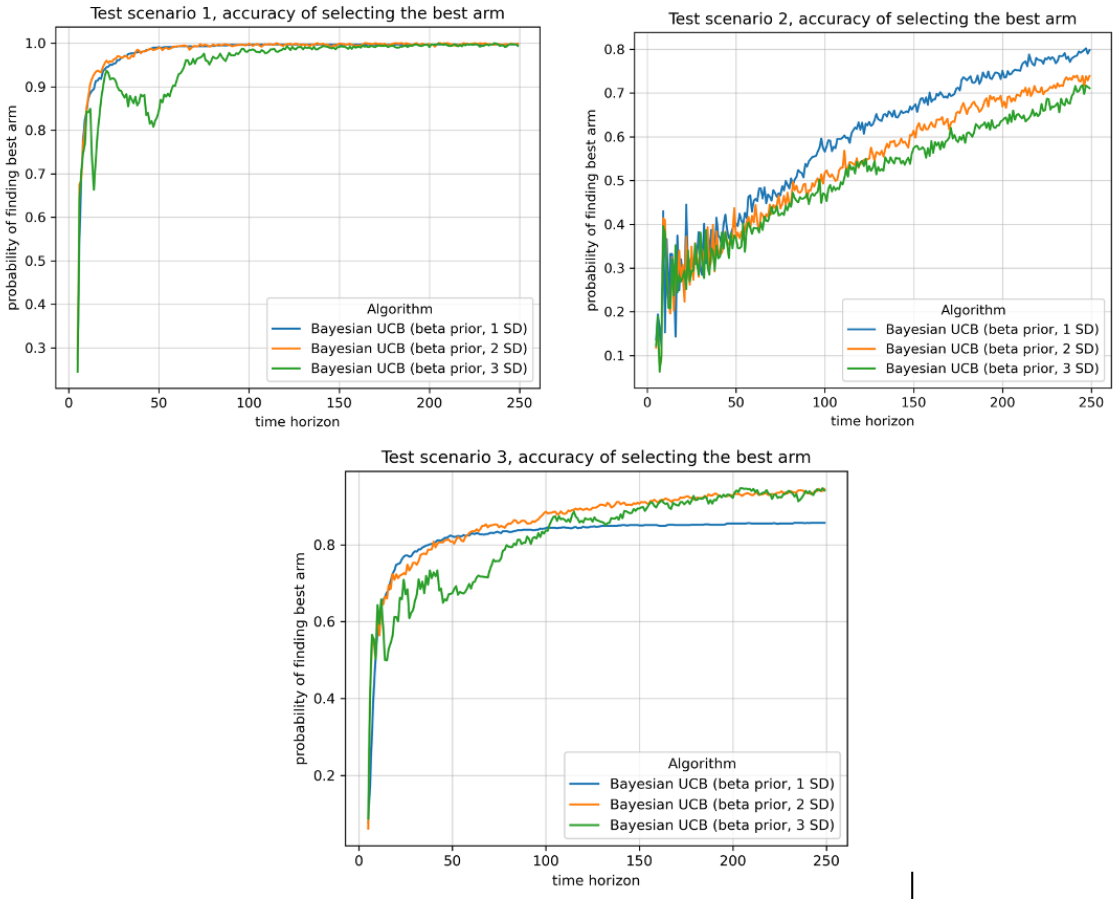


Fig. 52 Bayesian UCB (beta prior) with 1, 2, 3 SDs as confidence interval for Bernoulli test scenarios.

Bayesian UCB with percent point function of posterior distribution (beta prior)

We also implemented the Bayes-UCB algorithm proposed in literature.[62] At each time *t,* the UCB values are updated as follows:

$$q_j(t) = Q(1 - \frac{1}{t(\log n)^c}, \lambda_j^{t-1})$$

where Q(p, λ) is the percent point function (inverse of cumulative distribution function) of the posterior distribution (with distribution parameter λ) for each arm ***j*** at time ***t,*** such that:

$$\Pr(X \leq Q(p, \lambda)) = p$$

In the paper where this algorithm was proposed,[62] the authors also suggest dropping the (log *n*) term (*n* represents the total number of time horizon) and use c=0 in practice. This version uses the same probability "cutoff" for all arms and compares the value at which such probability constraint is satisfied for the posterior distribution for each arm. The arm with the highest such value is chosen as the next arm to play.

This implementation does not have any parameter to tune and is compared to the implementation in the previous section (**Fig. 53**). Other than test scenario 2, using the percent point function defined by each ***t*** does seem to offer slight advantages.

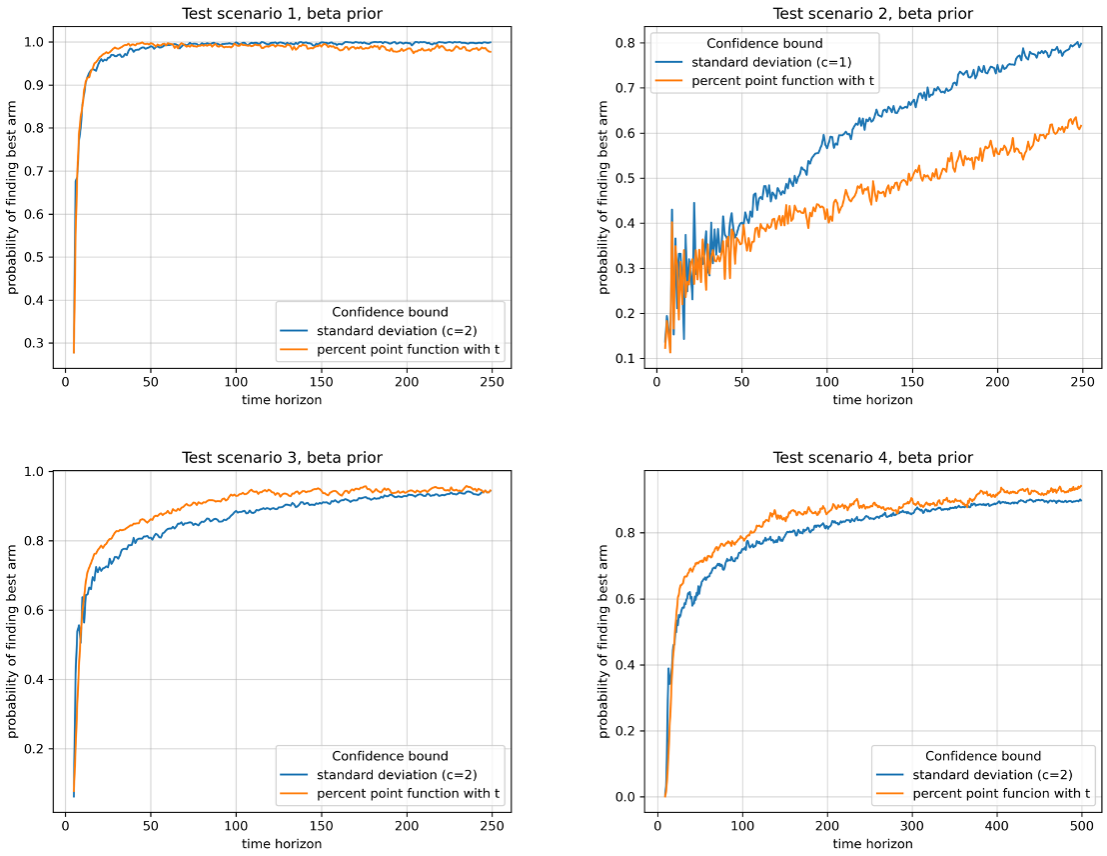


Fig. 53 Comparing two confidence bound implementations with beta prior in test scenario 1-4.

Bayesian UCB (Gaussian priors)

We also implemented Bayesian UCB with gaussian priors based on the implementations with beta priors. For the first approach of using standard deviation as confidence bounds, we only test the case with an assumed variance and set c to 2 (2 standard deviation, a ~95% confidence interval). For the second approach with percent point function, no parameter tuning is needed. We also tested the posterior update with a squared variance term that we used in Section 2.4.3, with c=2 (2 standard deviation confidence interval). All these algorithms are tested with five arms with gaussian reward distributions (means = [0.1, 0.2, 0.3, 0.4, 0.9], all with standard deviation of 0.25 or 0.5), and the results are shown in **Fig. 54**.

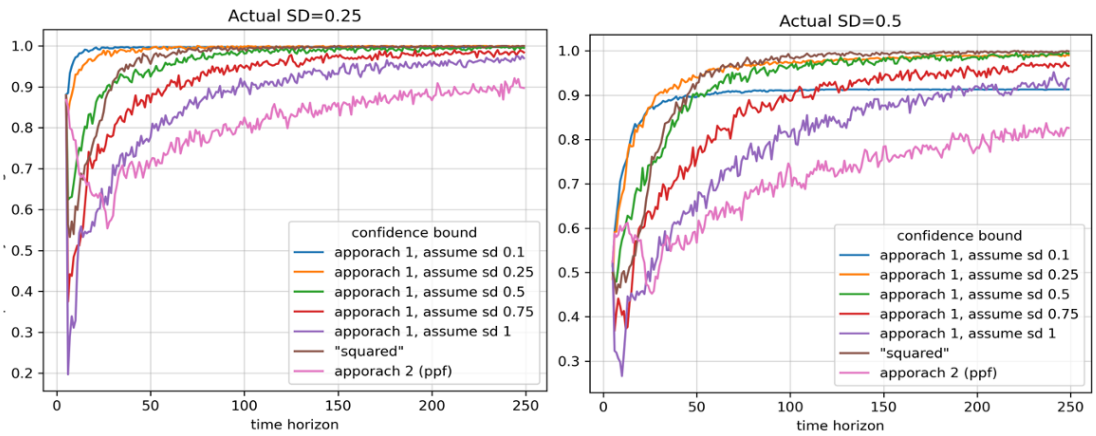


Fig. 54 Testing different approaches of Bayes UCB with Gaussian prior in two Gaussian test scenarios.

Approach 1 which uses 2 standard deviations as confidence bound, with assumed SD of 0.25, seems to work well for both situations. An assumed SD of 0.1 also seems to work for a low standard deviation setting, although 0.25 seems to be the more generally effective choice. The squared variance approach seems to show decent performance and is worth exploring, while approach 2 with percent point function does not seem to work well in this case.

It is also worth noting that Bayes UCB algorithms with Gaussian priors can also be used in Bernoulli test scenarios. Some of the testing results can be found in Section 2.4.5.

2.4.5 Best-performing bandit algorithms in test scenarios with Bernoulli and normal rewards

After modifications made to Thompson sampling as well as Bayes UCB algorithms, we re-tested some of the Bernoulli reward test scenarios as well as two normal reward test scenarios with low and high standard deviations.

Bernoulli test scenario 1

The best performing algorithms are now Bayes UCB algorithms with beta priors, followed by Bayes UCB algorithms with normal priors and Thompson sampling algorithms. Simpler algorithms such as ε-greedy, softmax and pursuit algorithms are not as effective.

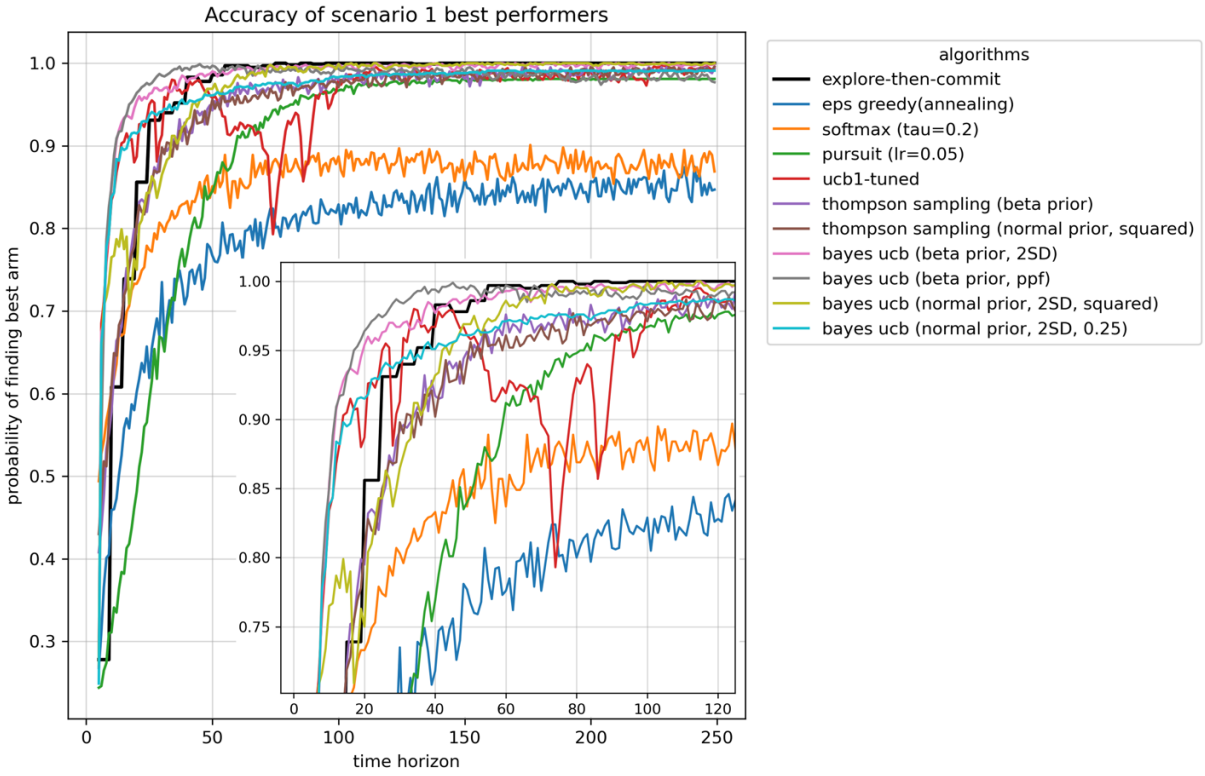


Fig. 55 Bernoulli test scenario 1, updated best performing algorithms.

Bernoulli test scenario 2

Based on results from test scenario 1, we focused mostly on TS and Bayes UCB algorithms for this test scenario due to their effectiveness. Only Bayes UCB with normal priors outperforms the explore-then-commit baseline significantly.

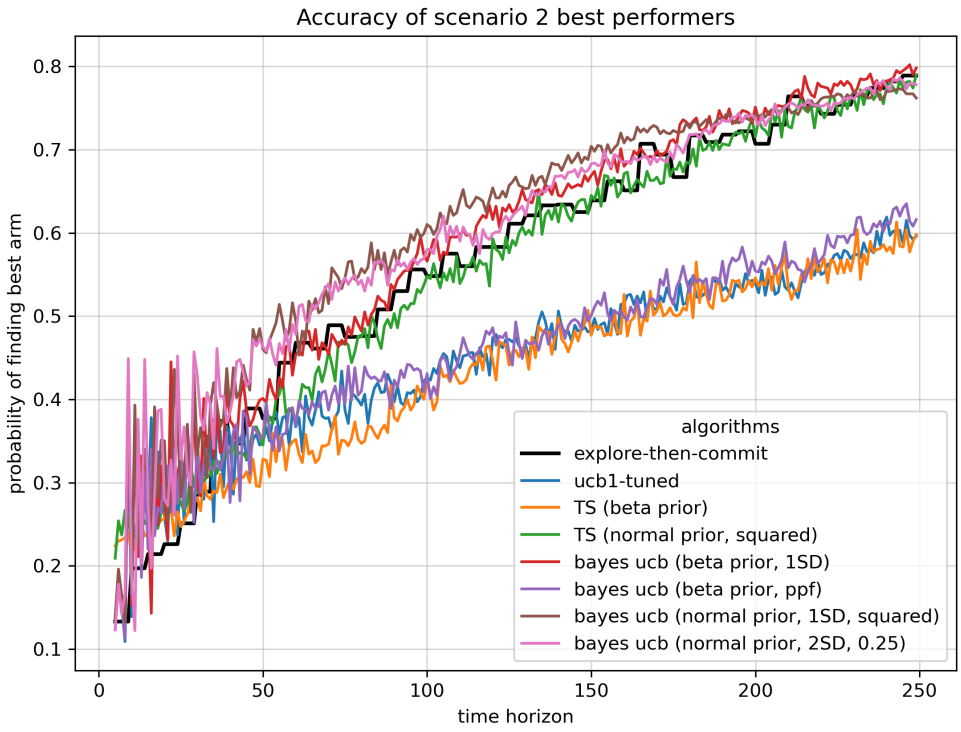


Fig. 56 Bernoulli test scenario 2, updated best performing algorithms.

Bernoulli test scenario 3

For test scenario 3, the performance trends of Bayes UCB and TS algorithms largely follow those in test scenario 1. Bayes UCB (beta prior, ppf) seems to offer the best initial accuracy, while Bayes UCB (normal prior, 2SD, squared) achieves a higher accuracy in later stages. Note: “ppf” refers to the percent point function implementation for Bayes UCB, and “squared” refers to the implementation with a squared variance term.

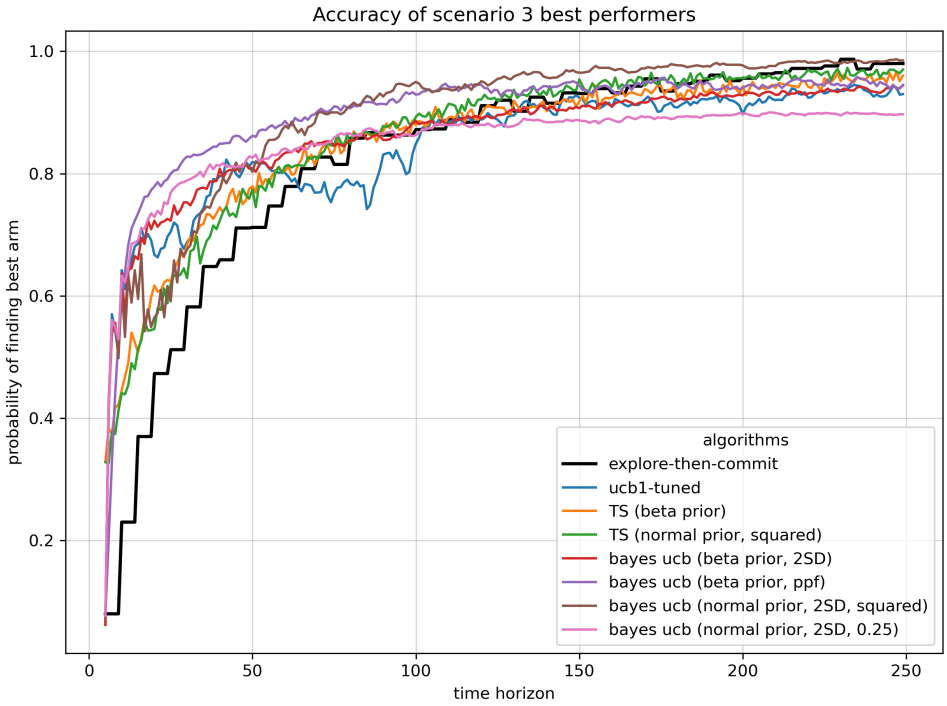


Fig. 57 Bernoulli test scenario 3, updated best performing algorithms.

Normal reward test scenarios

First normal reward test scenarios have five arms, with means [0.1, 0.2, 0.3, 0.4, 0.9] and standard deviation [0.5, 0.5 0.5, 0.5, 0.5]. All possible rewards returned from each arm is also bounded between 0 and 1. For arms with normally distributed reward, TS and Bayes UCB algorithms with normal priors can outperform simple ε-greedy and UCB1-tuned algorithms.

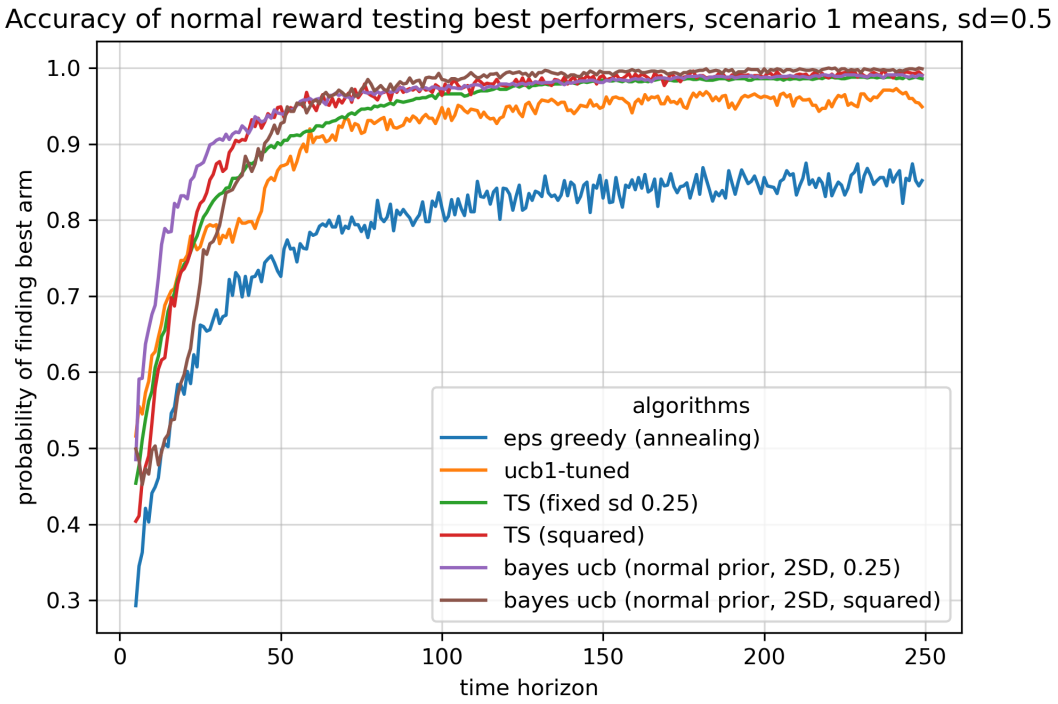


Fig. 58 Best performing algorithms in test scenario with normal rewards, high standard deviation setting (0.5).

The second normal reward test scenarios have five arms, with means [0.1, 0.2, 0.3, 0.4, 0.9] and standard deviation [0.25, 0.25 0.25, 0.25, 0.25]. Similar trends can be observed, with Bayes UCB algorithms being particularly effective.

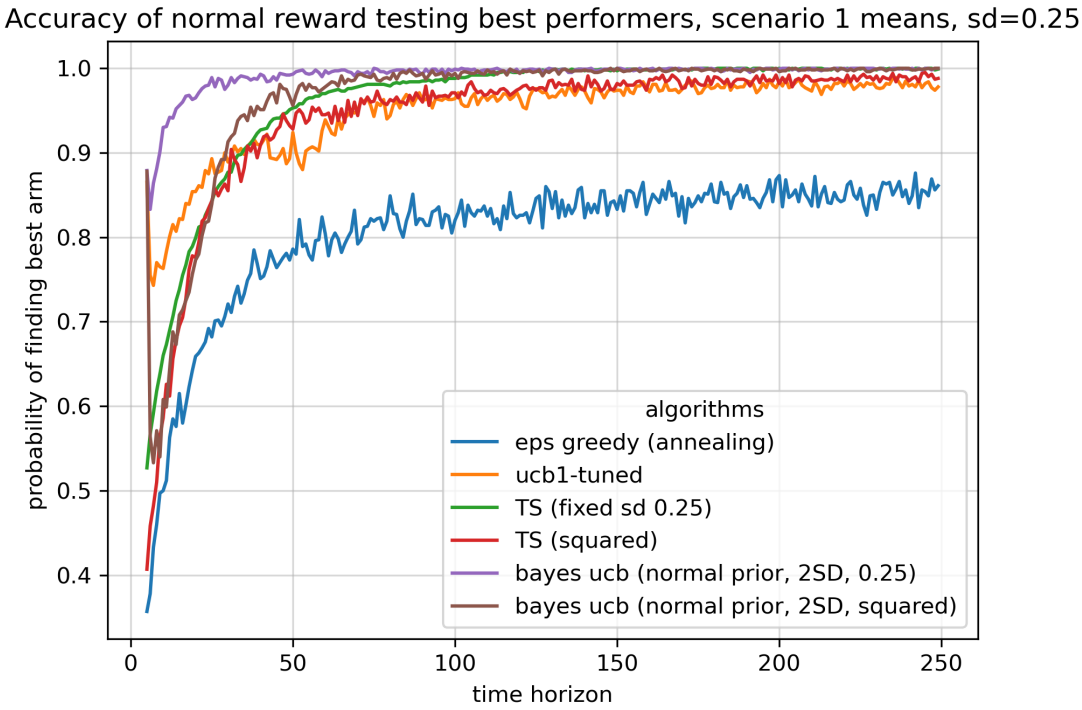


Fig. 59 Best performing algorithms in test scenario with normal rewards, low standard deviation setting (0.25).

2.4.6 Developing learning models and algorithms for batched experiments

Optimization algorithms, including those designed to address multi-armed bandit problems, are mostly sequential optimization algorithms. In other words, at each time point the algorithm outputs one arm to be queried and does not update until the result for that experiment becomes available. For chemistry experiments, which can take hours or days to complete, this creates an obstacle to efficiently conduct optimization in a reasonable, and sometimes constrained, timeframe.

Chemists usually resort to parallel experimentation to improve efficiency, where several experiments exploring different conditions are run at the same time. To allow such parallelization, optimization algorithms need to propose a batch of several experiments and accommodate delayed feedback and only update when all results in batch become available. For the algorithms considered optimal from previous simulation studies with synthetic data, we considered several approaches to accommodate batched experimentation based on the characteristics of the algorithms.

Algorithms with randomness: propose multiple experiments

For algorithms with inherent randomness, it is possible to propose multiple experiments before updating the algorithm. We tested this approach using ε-greedy algorithm with annealing exploration rate, varying the number of experiments proposed in each batch from 1 to 5.

For both scenario 1 and 2, the number of experiments per batch doesn't seem to affect the accuracy or cumulative reward performance (**Fig. 60**, **Fig. 61**).

It's also worth mentioning that, for simulation tests in this and all following sections, the definition of "time horizon" is not actual time, but rather the number of experiments regardless of batch sizes to allow for direct performance comparison. When plotting results, each experiment in one batch is randomly assigned to a time horizon point that covers that batch. For example, for a batch size of 2, batch 1 will cover experiment 1 and 2, and batch 2 will cover experiment 3 and 4… When plotting, experiment 1 can 2 can be assigned to time horizon 1 or 2; experiment 3 and 4 can be assigned to time horizon 3 or 4… Such processing is done because experiments in the same batch are not proposed and updated sequentially.

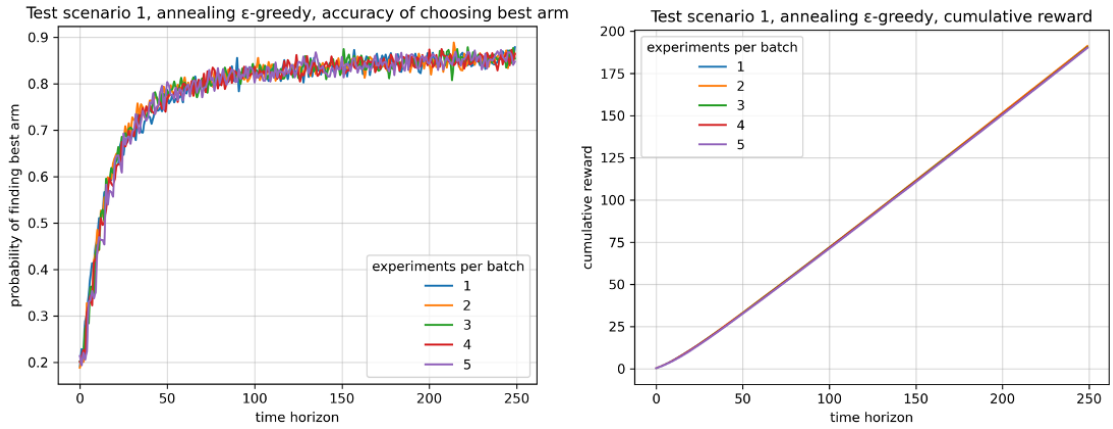


Fig. 60 The effect of batch size on optimization metrics (left: accuracy; right: cumulative reward) using ε-greedy for Bernoulli test scenario 1.

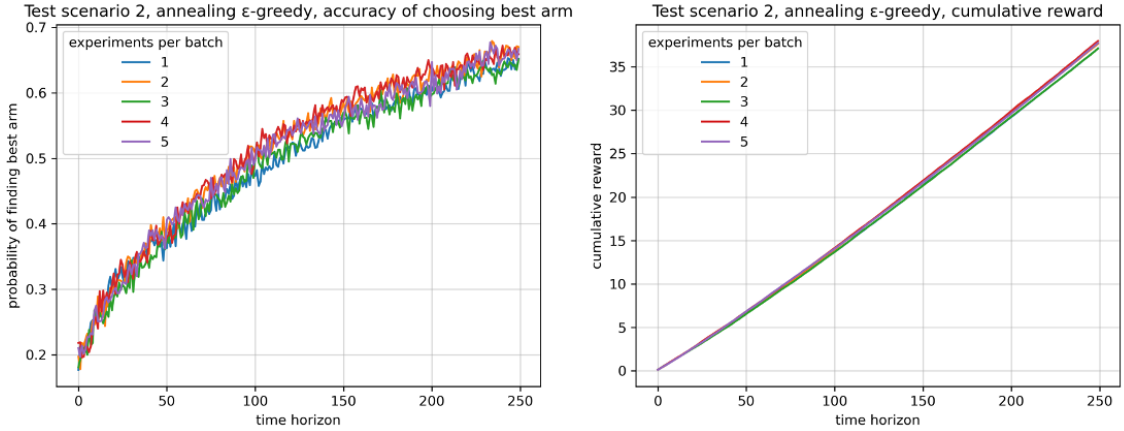


Fig. 61 The effect of batch size on optimization metrics (left: accuracy; right: cumulative reward) using ε-greedy for Bernoulli test scenario 2.

Thompson sampling algorithms: repeatedly sample posterior distribution

Thompson sampling algorithms can also propose multiple experiments in batch by repeatedly sampling from the posterior distribution before updating. The number of experiments does not seem to affect accuracy and cumulative reward, except some minor performance dips, due to the lag in updating the model (**Fig. 62**).

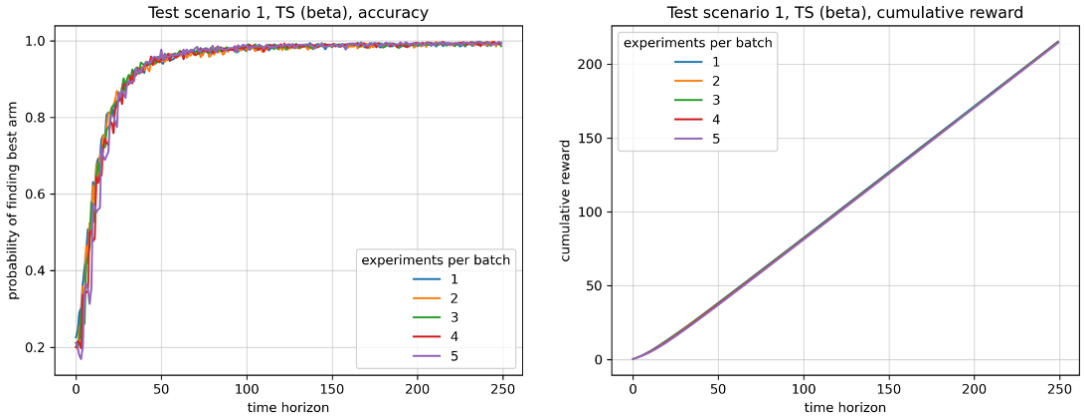


Fig. 62 The effect of batch size on optimization metrics (left: accuracy; right: cumulative reward) using Thompson sampling for test scenario 1.

Other algorithms

Algorithms without randomness or ones that do not sample from posterior distributions are usually deterministic in nature (although not completely deterministic, because each arm will return a stochastic reward, which makes the overall process stochastic). UCB algorithms fall into this category. Because of their deterministic nature, algorithms must select experiments sequentially, which requires some alternative ways of proposing multiple experiments. Different approaches to adapt these algorithms in a batched setting were considered.

- Sample multiple times for each chosen arm

The first approach is to select arms sequentially but run multiple reactions for each selected arm. For example, if optimizing for ligands, at each time, the algorithm will choose one ligand to investigate, and the same ligand can be tested with *n* substrates (*n*=batch size). This approach can be wasteful especially for large batch sizes, since all experiments in the same batch are dedicated to the same selected arm. The potential advantage is that by sampling multiple experiments for

each selected arm, the algorithm will likely converge faster since the mean/variance estimate will be more accurate.

We tested this approach with UCB1-tuned with Bernoulli test scenario 1 and 2. Overall convergence and accuracy do not seem to be affected. In scenario 1, running multiple experiments per selected arm seem to fix the "dip" in accuracy. In scenario 2, more experiments per chosen arm indeed provides a more precise mean estimate, therefore a higher accuracy overtime. It is worth noting that this approach will likely be less effective with a large number of arms (only five in both test cases). It will also be less effective in chemistry reaction optimization with a large batch size, where there are a finite number of substrates to sample from (unlike a Bernoulli distribution, which can be sampled repeatedly).

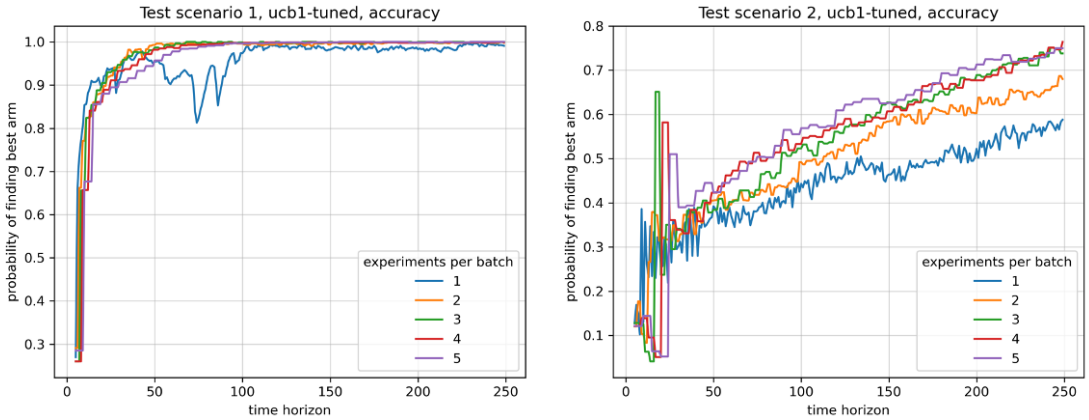


Fig. 63 The effect of batch sizes on optimization metrics using UCB1-tuned for test scenario 1 (left) and 2 (right).

- Initialize multiple algorithms at the same time

The second approach is to initialize n different algorithms at the same time (n = # of experiments per batch). Each algorithm will select one arm and propose one experiment at each round, and all the results are pooled together to update all algorithms, if applicable.

We tested this method with UCB1-tuned, Bayes UCB with beta prior, Bayes UCB with gaussian prior and Thompson sampling with Gaussian prior, again in Bernoulli test scenario 1 and 2. Accuracy is not plotted here since all algorithms select one reaction every round, and it is not reasonable to assign a specific time to a specific selection by one algorithm within each round. But based on cumulative reward, this batched approach achieves similar level of performance as individual algorithms, possibly guaranteeing an average performance level.

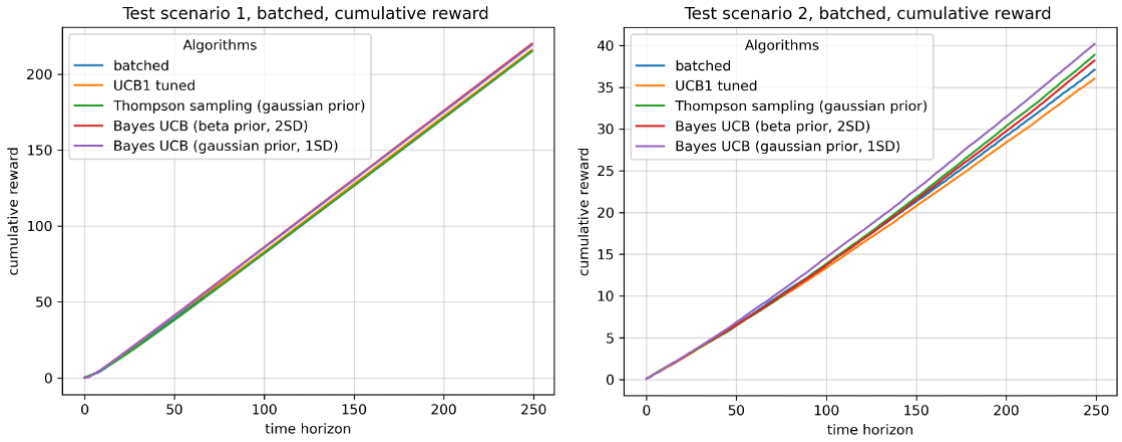


Fig. 64 The effect of batching different algorithms on cumulative reward for Bernoulli test scenario 1 (left) and 2 (right).

- Interpolation with underlying prediction models

The general idea for this approach is similar to kriging and uses an underlying machine learning prediction model that updates after every batch. When algorithm proposes an arm to evaluate and a specific experiment is proposed, instead of waiting for the experiment to be run, algorithm updates itself with the predicted result for the proposed experiment and propose the next arm to evaluate (and the experiment to be run). This process is continued at each round of optimization, until the desired number of experiments has been reached. Then, all proposed

experiments are conducted, and algorithms are updated with the real experimental results, and continue to the next round.

This approach will rely on the prediction model to be effective, especially with low number of training data available. We first tested the feasibility of implementing such prediction model with good prediction accuracy with a small percentage of the scope as training data. We used the deoxyfluorination dataset as a test set, and used random forest model, which has been demonstrated in the original publication[63] to be an effective model, as the prediction model. The random forest model (with default parameters in scikit-learn) was trained with various training set size. A training set size ratio of 10% means that model is trained with 10% of all data in the scope and tested with the remaining 90% of data. Test RMSEs are obtained as the averages of 100 runs with randomly partitioned train/test sets at each ratio. Different featurization methods are also tested and compared: DFT features used in the original publication, one-hot encoding, and a combination of molecular fingerprints (for substrates) and one-hot encoding (for conditions). A linear regression model is also implemented as baseline comparison. As shown in **Fig. 65**, random forest model with fingerprint and one-hot encodings gave similar accuracy performance as the random forest model trained with DFT features, with average RMSE around 20% with only 10% of the data. This is beneficial for our on-the-fly prediction model, as fingerprints are cheap to calculate and time-consuming DFT calculations can be avoided.

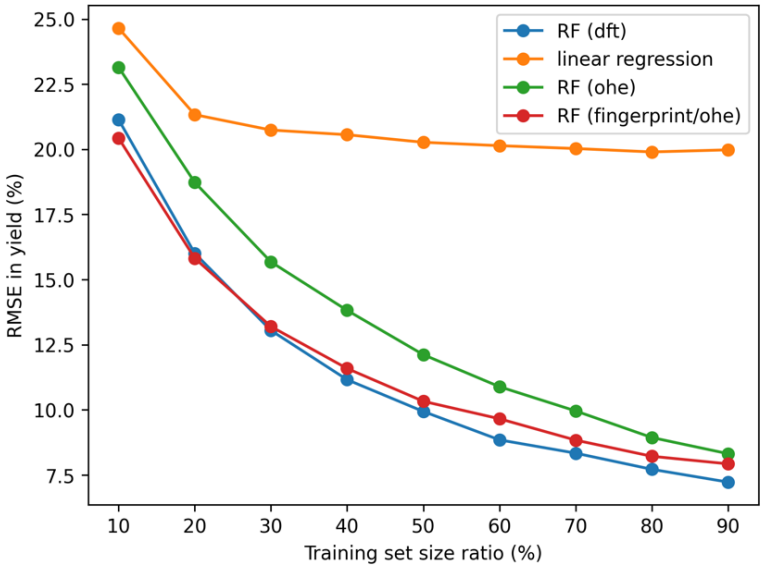


Fig. 65 The effect of training set size on test RMSEs with different models and features. Average of 100 runs on randomly partitioned deoxyfluorination dataset.

After testing prediction model in a low training data setting, and the validation of fingerprints as effective features, we tested the proposed interpolation method during optimization with UCB1-Tuned algorithm on the deoxyfluorination dataset. Substrates are featurized with extended connectivity fingerprints (ECFP) and conditions are featurized with one-hot encoding. After each round, a random forest model is trained with data collected so far and predicts reaction yield for the rest of the scope. Similar to a believer algorithm, this model supplies predicted results to the bandit algorithm, which uses these predicted results to sequentially proposes experiments without actual experimental feedback. We tested the effect of batch sizes on the simulated optimization accuracy. The performances with batch size from 2 to 10 are very similar to the results with batch size of 1 (which is the standard sequential approach, **Fig. 66**, top). Extreme batch sizes (e.g., 50, 100) show lower initial accuracies because a large number of reactions were proposed with limited initial data, but the accuracies catch up if given enough time (**Fig. 66**, bottom). These

results show that experiments can be proposed in batch with ML model interpolation without compromising overall accuracies.

In **Fig. 66**, "time horizon" still represents the number of experiments for all batch sizes to directly compare performance. The exact definition of time horizon, however, is worth discussing in this case. Most bandit algorithms are applied with a batch size of 1, where one arm is selected to query, and immediate feedback is available to the algorithm. In these cases, the definition of time horizon is interchangeable with number of experiments. However, when a batch size of n (n>1) is used, the algorithm does not get any feedback (or experimental result) until all n experiments are proposed and the results of these experiments in the same batch are available. Assuming one batch of reactions take the same amount of time to complete regardless of batch size, the definition of time horizon is better represented with actual time rather than number of experiments. To this end, we plotted the accuracy data in **Fig. 66** with an arbitrary unit of time, assuming one round of experiments take one unit of time (**Fig. 67**). As expected, more experiments per round result in higher accuracies achieved in the same amount of time. However, such effect is less obvious when initial accuracies are low for batch sizes like 50 and 100.

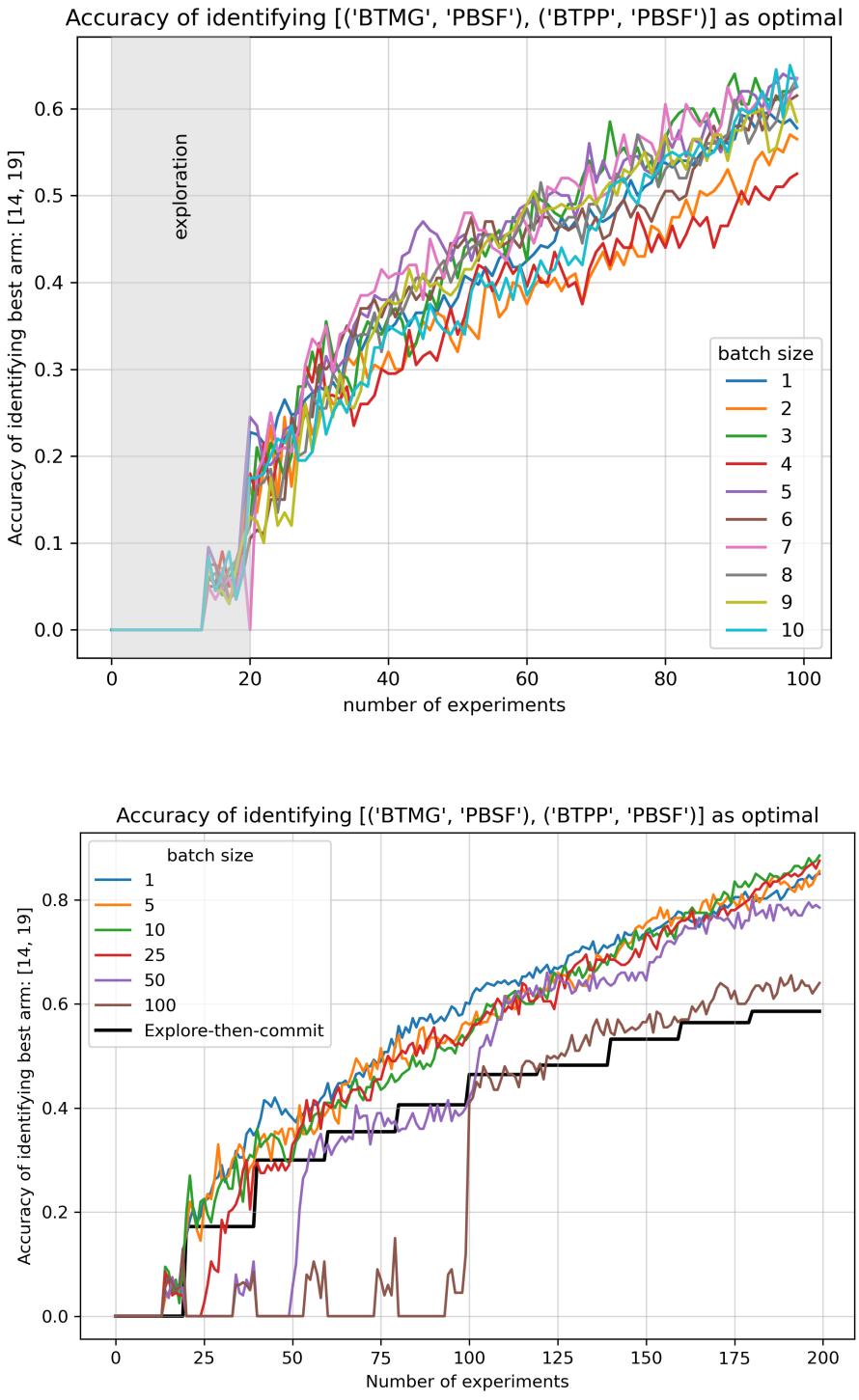
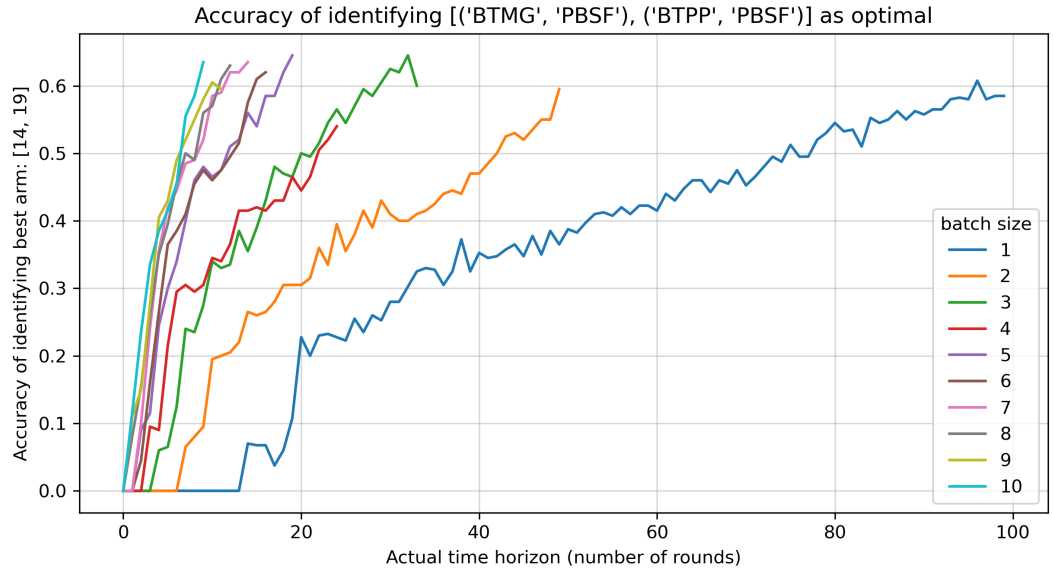


Fig. 66 The effect of batch size on UCB1-Tuned accuracy using a random forest prediction model to interpolate experiment results.



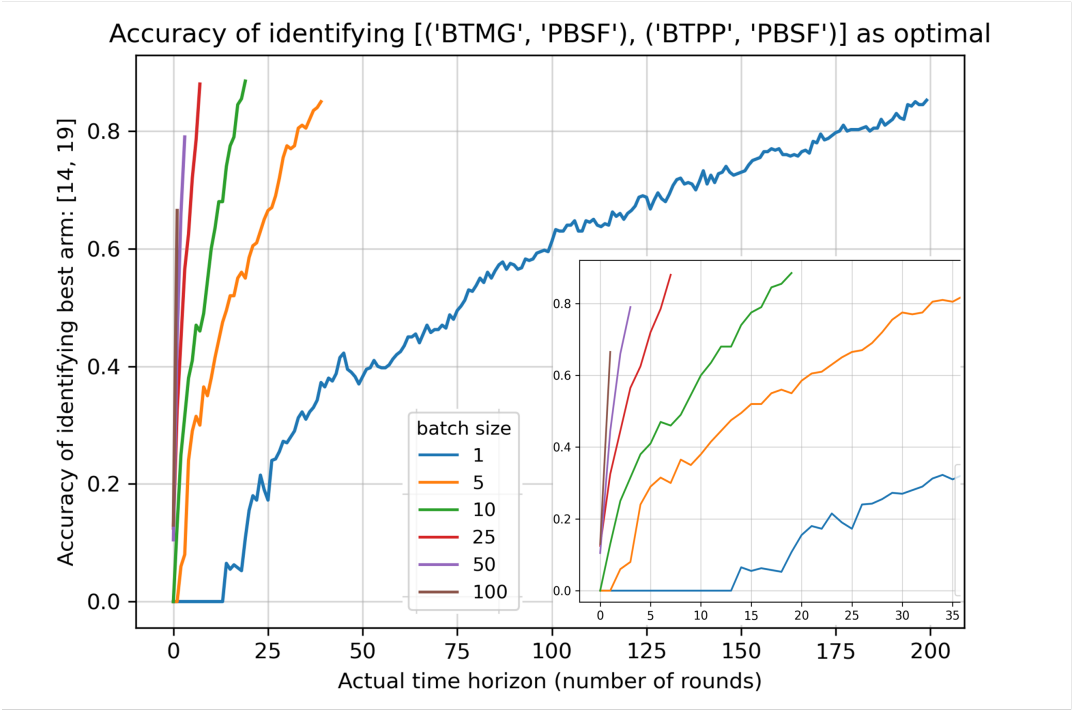


Fig. 67 The effect of batch size on UCB1-tuned accuracy using a random forest prediction model to interpolate experiment results, with the x-axis being actual time, or the number of rounds.

2.4.7 Generality optimization model design for chemistry reaction data

Background

Reaction condition optimization is a closed-loop, interactive problem where the goal-directed learning system's actions uncover characteristics of an uncertain environment and carry consequences that affect its later inputs. Unlike a supervised learning problem where instructive feedback from expert-labeled training data is used to establish models that can identify correct actions in different situations defined by unique features, an optimization problem is a reinforcement learning problem where the model must balance real-time action selection and long-term planning to incrementally learn in uncharted territory. When optimizing an unknown target reaction, chemists take actions in real-time, use evaluative feedback to update beliefs over the initially unknown environment, and continue the process iteratively to reach reactivity or selectivity objectives. Compared to stateless active learning approaches such as Bayesian optimization, one key difference for reinforcement learning models is that the learning agent maintains knowledge of the environment and has an explicit goal directly related to the state of the environment. By design, it can enable efficient optimization by maximizing knowledge learned from limited existing data.

Optimization for generally applicable conditions, where multiple substrates must be simultaneously considered to identify a single set of satisfactory conditions, renders conventional approaches such as Bayesian optimization inefficient. As a stateless algorithm suitable for sub-problems with fixed search spaces, Bayesian optimization sequentially and actively queries select data points to find global optimum of black-box functions, with the assistance of a surrogate model that maps the relationship between inputs and outputs. While effective for single substrate optimization problems, it is not experimentally feasible to individually make separate queries with

selected condition for all substrates in a given scope. Existing approaches with Bayesian optimization[64] have relied on the use of supervised learning models to provide predicted results for such queries. Multi-task Bayesian optimization approach also exist to transfer existing knowledge from one optimization to others.[65] In a general sense, we became interested in using reinforcement learning models to optimize for generally applicable conditions.

Optimization model design with bandit optimization

We envision bandit optimization algorithms as a more suitable approach to identify generally applicable conditions for a reaction scope. Optimization for such conditions aims to find the best condition for a scope of substrate (not just one, as most Bayesian optimization approaches tackle). Fundamentally, our model treats reaction conditions as arms in a multi-armed bandit problem. In a reaction scope with multiple conditions and multiple substrates, each condition will exhibit different reactivities when used on different substrates, resulting in a unique, condition-specific reactivity distribution for each condition. Similar to the classic multi-armed bandit problem and its algorithmic solutions, the player (chemist) will try to select an arm (condition) to evaluate. A reward (reactivity) is returned by sampling the reward distribution (reactivity distribution) of the selected arm (condition). The sampling, in the chemistry reaction case, is done by sampling one reaction with the selected condition that has not been explored yet.

To further demonstrate the reactivity distribution represented by different conditions in a chemistry reaction, we use a deoxyfluorination dataset[63] to visualize such effect. The details of this reaction dataset, as well as other analyses, can be found in Section 2.4.8. This reaction dataset has both sulfonyl fluorides and bases as conditions, and a sizable alcohol substrate scope. Like discussed above, the most obvious and straightforward representation will be that the arms are

represented by sulfonyl fluoride–base conditions, and the distribution for each arm is the different reactivities exhibited by all substrates for each condition. For simplicity, we only visualize three out of all conditions and their reactivity distributions (**Fig. 68**). Each condition has a different yield distribution with regard to the same substrate scope, which is ultimately the distribution that the learning model will sample from after an arm (condition) is selected.

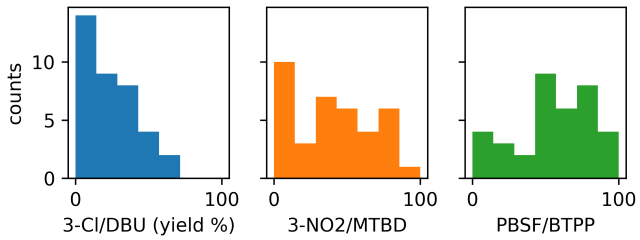


Fig. 68 Reactivity distribution of three base–sulfonyl fluoride conditions in the deoxyfluorination dataset.

The unique aspect of bandit optimization approach, however, is that any reaction component (not necessarily substrates) that is not part of the optimization objective can be easily incorporated into the reward distributions. For the same deoxyfluorination reaction, if we only want to optimize for the most general base, but still maintain the same reaction scope, each base also has their own reactivity distribution with substrates and sulfonyl fluorides included (**Fig. 69**). BTPP has a slightly better distribution than MTBD. Both BTPP and MTBD are significantly better than DBU.

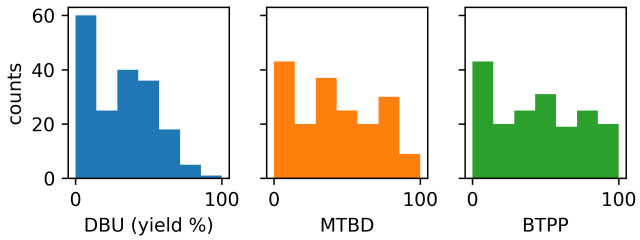


Fig. 69 Reactivity distribution of three bases in the deoxyfluorination dataset.

Similar distributions can be visualized for sulfonyl fluorides (**Fig. 70**). PBSF offers significantly better reactivities than 3-Cl and 3-NO2 phenyl sulfonyl fluoride.

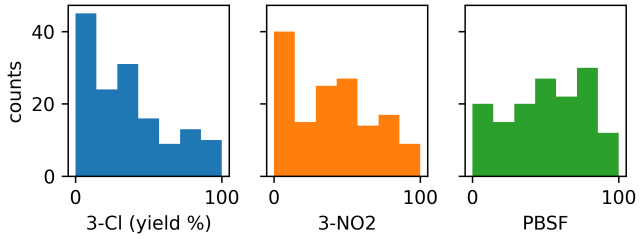


Fig. 70 Reactivity distribution of three sulfonyl fluorides in the deoxyfluorination dataset.

Therefore, for any given reaction scope, each optimization target will have a unique distribution to be sampled from, which includes substrates and may include additional conditions that are not part of the optimization objectives. The goal of the learning model is to efficiently estimate such distributions through sampling, and subsequently suggest best option based on its estimation. As demonstrated, this implementation can enable unique functionalities, such as an evaluation of some of the condition dimensions to eliminate less effective options early on, and then expand to include other condition dimensions to further optimize the reaction (demonstrated in the amide coupling study). Another important functionality is the ability to accommodate

changing substrate scopes. Because substrates are only represented by a distribution, changes in substrates do not directly interfere with the optimizations themselves. Bandit algorithms do not expect a finite search scope anyways, but rather learns from the feedback from the environment (which can be dynamic). Therefore, bandit algorithms can adjust to the changing substrate scope through continued sampling, which was demonstrated by the C–H arylation study.

Experiment selection via sampling from reactivity distributions

One important consideration after a condition is selected by the bandit algorithm is which experiment to run. In its simplest form, this involves choosing a specific substrate to test the condition with. In this study, all substrates selection were done through random sampling of the substrate scope. After a condition is selected, the substrates that have not been explored with that condition is recognized, and one substrate is randomly chosen to run.

This approach is seemingly very simple but can be very effective in practice. Several considerations contributed to the selection of this strategy. First and foremost, the reactions investigated in this study have a sizable but still quite limited substrate scope. The selection of these substrates usually occurs through expert selection to represent diverse structural motifs. With limited number of substrates in the scope with distinct structural features, random sampling is the most efficient way of sampling from the scope. We do acknowledge that in situations such as library synthesis, where a large number of structurally similar molecules exist in the substrate scope, our approach of random sampling can be less efficient. In those situations, an approach where molecules are first grouped by similarity (e.g., through clustering) and sampled with consideration of their cluster labels will be more suitable.

However, it is still worth considering that even structurally similar molecules can exhibit different reactivities under the same conditions. For example, the addition of inconspicuous methyl groups, when *ortho* to the halide for aryl halide substrates, can cause significant steric constraints in cross coupling chemistry and require changes to the conditions to enable effective coupling. Such considerations, though obvious to organic chemists, are difficult to capture through molecular fingerprints or descriptors of substrates that model can understand. These phenomena contribute to the difficulty of optimization especially when trying to find general conditions. In these situations, random sampling might still be the most effective approach, as it does not make any prior assumption of the reactivities.

The nature of the condition space is also worth considering. In our studies, we limited the number of conditions and conditions that might show very similar reactivity trends. If there are many similar conditions and if a particular substrate has been evaluated with one of such conditions, it does not make sense to evaluate the same substrate with very similar conditions. However, this strategy still requires the assumption that similar conditions, when applied to the same substrate, will yield similar results, which may or may not be true in practice. Again, random sampling is probably still the simple and effective approach here.

Within the scope of our study, we tested two other approaches of selecting substrates beyond random sampling with the aid of a supervised learning model. The main idea is that when selecting substrates to run, substrates that might show high reactivities (estimated by a prediction model) is more worthwhile to test than low-yielding substrates, because the former happens less often in discovery (it is still worth pointing out that this assumption might be flawed for reactions that are well established, such as an amide coupling reaction investigated in this study). Therefore, we tested our C–H arylation dataset (details about this dataset can be found in Section 2.4.8) using

the same Bayesian UCB algorithm, but with two other sampling methods after selecting a condition: (1) propose the substrate that gives the highest yield based on model prediction; (2) propose a random substrate out of the top five substrates ranked by predicted yield (to reduce bias). These two approaches are compared with the random sampling approach (**Fig. 71**).

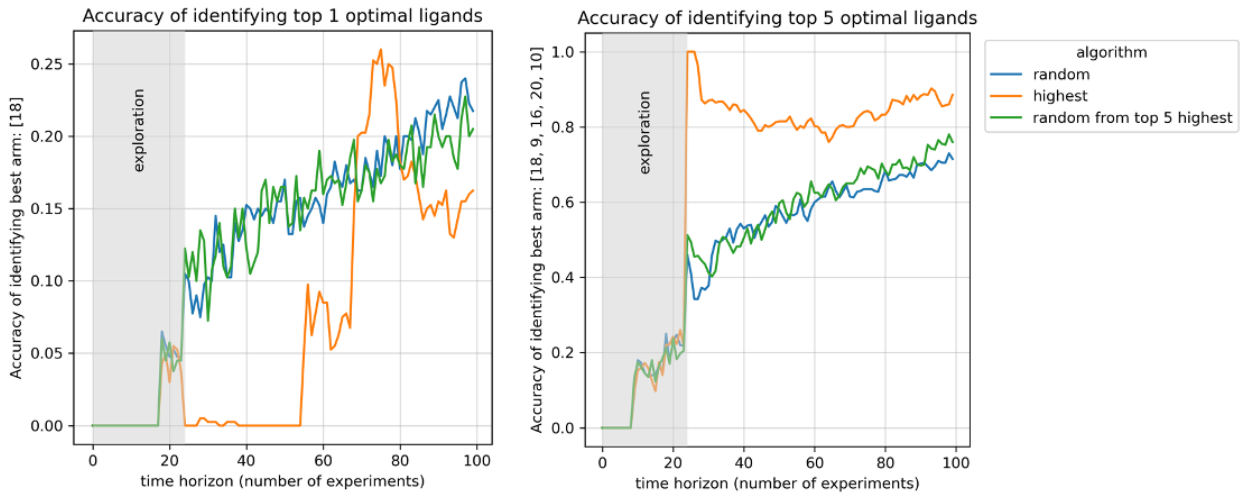


Fig. 71 Comparing substrate sampling methods with C–H arylation data. Top-1 (left) and top-5 (right) accuracy are shown with three methods.

Compared to accuracies from random sampling of substrates, choosing a random substrate out of top five substrates showed similar accuracies, which might be explained by the low accuracy of prediction model. If the prediction model is not accurate (which is likely the case with very low training data initially, where many conditions and substrates have not even been sampled once yet), then the "top five" substrates based on prediction might as well be randomly chosen. Always choosing the substrate that gives the highest predicted yield, however, definitely resulted in bias during optimization. While top-5 accuracy is high with approach (**Fig. 71**, right), the model is not selecting the optimal condition based on the top-1 accuracy (**Fig. 71**, left). It is conceivable that by choosing substrates that gives the highest predicted yield, the model is biasing towards a few

substrates that show higher reactivities and not sampling the rest, which makes it more similar to a less ideal model substrate approach. With these observations, and the fact that both of these approaches require a supervised learning model, we opted to use random sampling of substrates in this study.

Optimization model architecture

The overall architecture of the optimization model is shown in **Fig. 8**. In the simplest scenario where one experiment is evaluated at a time, the bandit algorithm will choose conditions to evaluate first (step 1), which then gets passed on to the reaction scope (which we implemented as a python object that handles many relevant operations, step 2). The reaction scope then suggests a reaction to evaluate with the selected condition and a chosen substrate (step 3). This reaction is executed in lab (step 4), and the labeled reaction data (reaction with yield or other reactivity metric) is used to update both the reaction scope and the bandit algorithm (step 5). In cases where reactions are proposed in batch, a prediction model is trained with all labeled data available in the reaction scope (step 6). In these cases, instead of executing the reaction in lab in step 4, a predicted result from the prediction model is used to "label" the reaction and "update" the reaction scope and bandit algorithm.

After the update, the bandit algorithm suggests the next condition to evaluate (step 1), and this entire process is repeated. It is also worth noting that no stopping criteria are implemented in this case, as we often observe in practice that time or experimental resources usually exhaust first. From empirical observations, exploring around 10% of the reaction scope (e.g., 100 out of 1000 possible experiments) usually gives satisfactory results.

2.4.8 Chemistry reaction dataset: data processing, global analysis, and optimization studies

Overview

Organic chemistry reaction datasets are collected from literature and experiments to test various functionalities of optimization framework. Access to datasets used in this study, test and analysis functions and testing log files are described in this GitHub repository: https://github.com/doyle-lab-ucla/bandit-optimization. The scope of each reaction dataset, as well as the visualization, global analysis and optimization simulation studies for each dataset are presented in the following sections. Overall, seven datasets are included, five of which were previously published and two were experimentally investigated in this study.

It is worth noting that we made a modification to the **accuracy** metric when analyzing chemistry reaction data. Previously, for our simulation studies using arms with Bernoulli and gaussian distribution rewards, accuracy at time point *t* was calculated as the average percentage of times that the optimal arm is selected across all simulations (Section 2.4.2). For simulations with chemistry reaction data, however, we do not only look at which arm is being selected at time point *t*, but rather all previous arms that are selected up until time point *t*. This is due to the limited scope of a chemistry reaction dataset, and the fact that each arm (condition) can only be sampled so many times in a finite scope. The arm sampled the most times up until time point *t* is regarded as the optimal arm chosen by the algorithm, and the percentage of such selection is similarly calculated across all simulations. This modified definition of accuracy matches with what happens in a real optimization campaign, where all historical data are considered when deciding which condition is optimal. All accuracy plots in this section are created with this modified definition.

For the literature datasets we examined, we first determined the most general conditions with all experimental data and tasked our model to "rediscover" these conditions and evaluate the

accuracy. Ideally, for every dataset, there will be one most general condition to be identified by the algorithm based a reactivity metric (such as the number of hits, or average/median yields, as used in this study). However, the top-n conditions often have very close values, and a single best condition cannot be reasonably decided. In these cases, we ranked all conditions from best to worst, and looked for a relatively significant drop in reactivity, and treat all conditions before this drop as optimal conditions. Furthermore, the single best condition based on one metric (e.g., average yield) might also differ from that based on another metric (e.g., number of >80% yield reactions). A top-n accuracy, where n>1, can often accommodate these differences between metrics.

Full descriptions of algorithm abbreviations shown in plots.

Names for algorithms were abbreviated in result plots for clarity. The full descriptions of algorithms used in simulation studies are listed here.

- Algorithms for Bernoulli-type (0/1) reward:
    - **TS (normal prior):** Thompson sampling with normal prior, the "squared" implementation where a standard deviation of 1/(n+1) is used for the normal prior. This is the same algorithm as **TS (squared)** for continuous [0,1] reward.
    - **TS (beta prior):** Thompson sampling with beta prior.
    - **ucb1-tuned:** UCB1-tuned algorithm.
    - **ucb1:** UCB1 algorithm.
    - **Bayes ucb (normal prior):** Bayesian UCB with a gaussian prior with the "squared" implementation in Section 2.4.4, where a standard deviation of 1/(n+1) is used for the normal prior and a 2-standard deviation confidence interval is used for UCB

values. This is the same algorithm as **Bayes ucb (2SD, squared)** for continuous [0,1] reward.

- **Bayes ucb (beta prior, 2SD):** Bayesian UCB with a beta prior that uses mean + 2 standard deviation (2SD confidence interval) as UCB values.
- **Bayes ucb (beta prior, ppf):** Bayesian UCB with a beta prior that uses a percent point function to update UCB values.
- **Annealing ε-greedy:** ε-greedy algorithm with an annealing function used for ε.
- **pure exploration:** exploration, or random selection.
- **explore-then-commit:** explore-then-commit.

- Algorithms for continuous reward from 0 to 1:
  - **TS (squared):** Thompson sampling with normal prior, the "squared" implementation where a standard deviation of 1/(n+1) is used for the normal prior. This is the same algorithm as **TS (normal prior)** for Bernoulli-type reward. This is also the **TS (implementation 1)** in **Fig. 10**.
  - **TS (fixed sd 0.25):** Thompson sampling with normal prior, assuming a known standard deviation of 0.25. This is also the **TS (implementation 2)** in **Fig. 10**.
  - **ucb1-tuned:** UCB1-tuned algorithm.
  - **ucb1:** UCB1 algorithm.
  - **Bayes ucb (2SD, squared):** Bayesian UCB with a gaussian prior with the "squared" implementation in Section 2.4.4, where a standard deviation of 1/(n+1) is used for the normal prior and a 2-standard deviation confidence interval is used for UCB

values. This is the same algorithm as **Bayes ucb (normal prior)** for Bernoulli-type reward. This is also the **Bayes UCB (implementation 1)** in **Fig. 10**.

- **Bayes ucb (2SD, 0.25):** Bayesian UCB with a gaussian prior with assumed standard deviation of 0.25. A 2-standard deviation confidence interval is used for UCB values (Section 2.4.4). This is also the **Bayes UCB (implementation 2)** in **Fig. 10**.
- **ε-greedy:** ε-greedy algorithm with an annealing function used for ε.
- **pure exploration:** exploration, or random selection.
- **explore-then-commit:** explore-then-commit.

Nickel borylation

This dataset is extracted from the publication: “Advancing Base Metal Catalysis through Data Science: Insight and Predictive Models for Ni-Catalyzed Borylation through Supervised Machine Learning.” Stevens, J. M.; Li, J.; Simmons, E. M.; Wisniewski, S. R.; DiSomma, S.; Fraunhoffer, K. J.; Geng, P.; Hao, B.; Jackson, E. W. *Organometallics* **2022**, *41* (14), 1847–1864. [DOI: 10.1021/acs.organomet.2c00089].[66]

The raw data was processed before being used in our optimization simulation studies. Ligand $PnBu_3$•HBF_4 was removed from the scope due to missing yields for some substrates. The electrophile and ligand scope of the dataset is shown in **Fig. 72**.

Reaction yields are visualized with heatmap, with a side-by-side comparison of yields in MeOH and EtOH (**Fig. 73**). Both substrate- and ligand-dependent reactivities can be observed from the heatmap. The differences in yields for EtOH and MeOH were also plotted to compare solvent performance (**Fig. 74**). Notably, for certain substrates (such as **s16**), MeOH offers much higher reactivity compared to EtOH. The exact reason for these reactivity differences is not clear, and likely not due to solubility issues. But overall, MeOH and EtOH exhibit similar reactivity trends across the majority of the substrates.

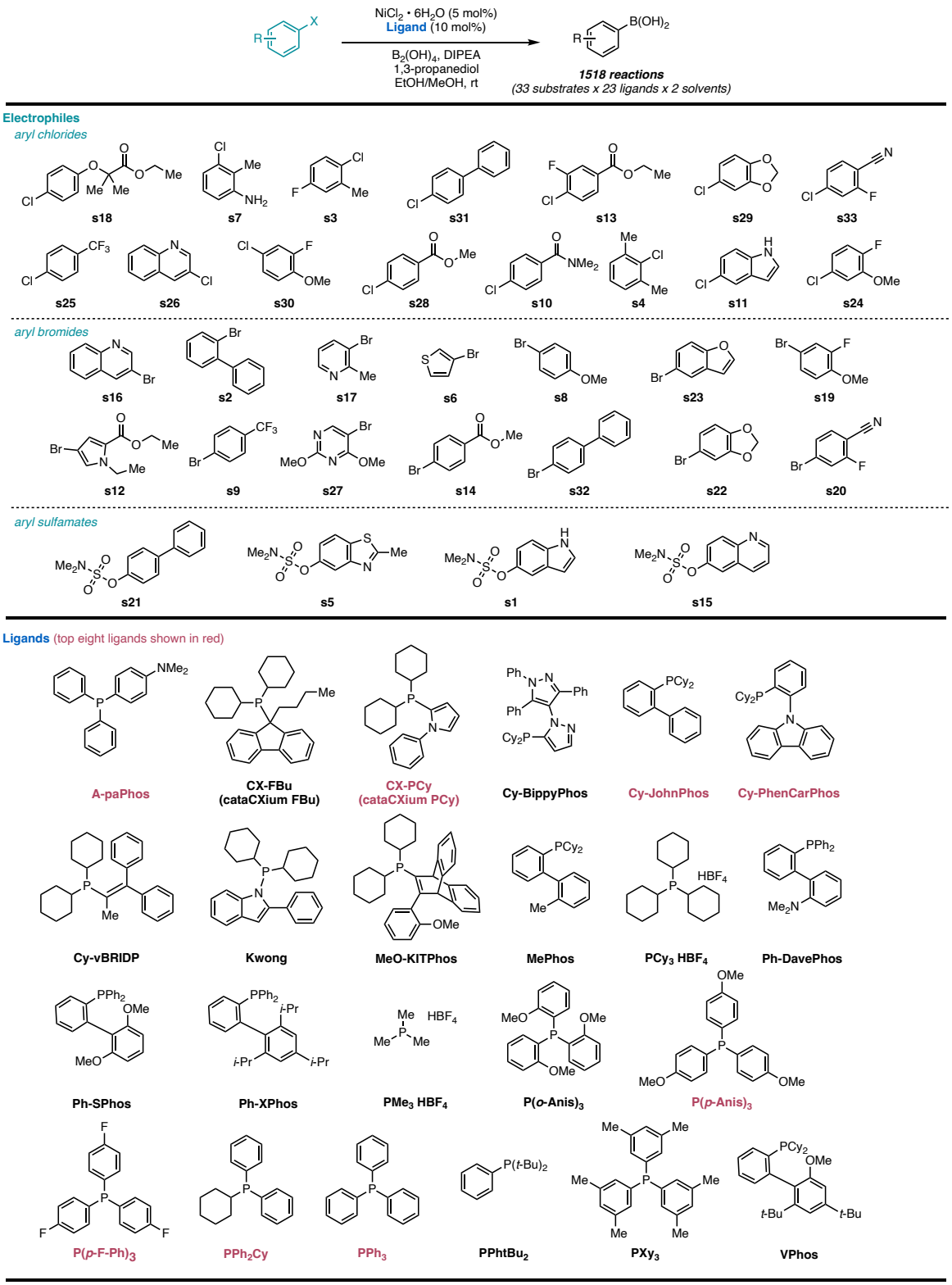


Fig. 72 Nickel-borylation dataset after processing: electrophile scope and ligand scope. Top eight ligands identified are highlighted in red.

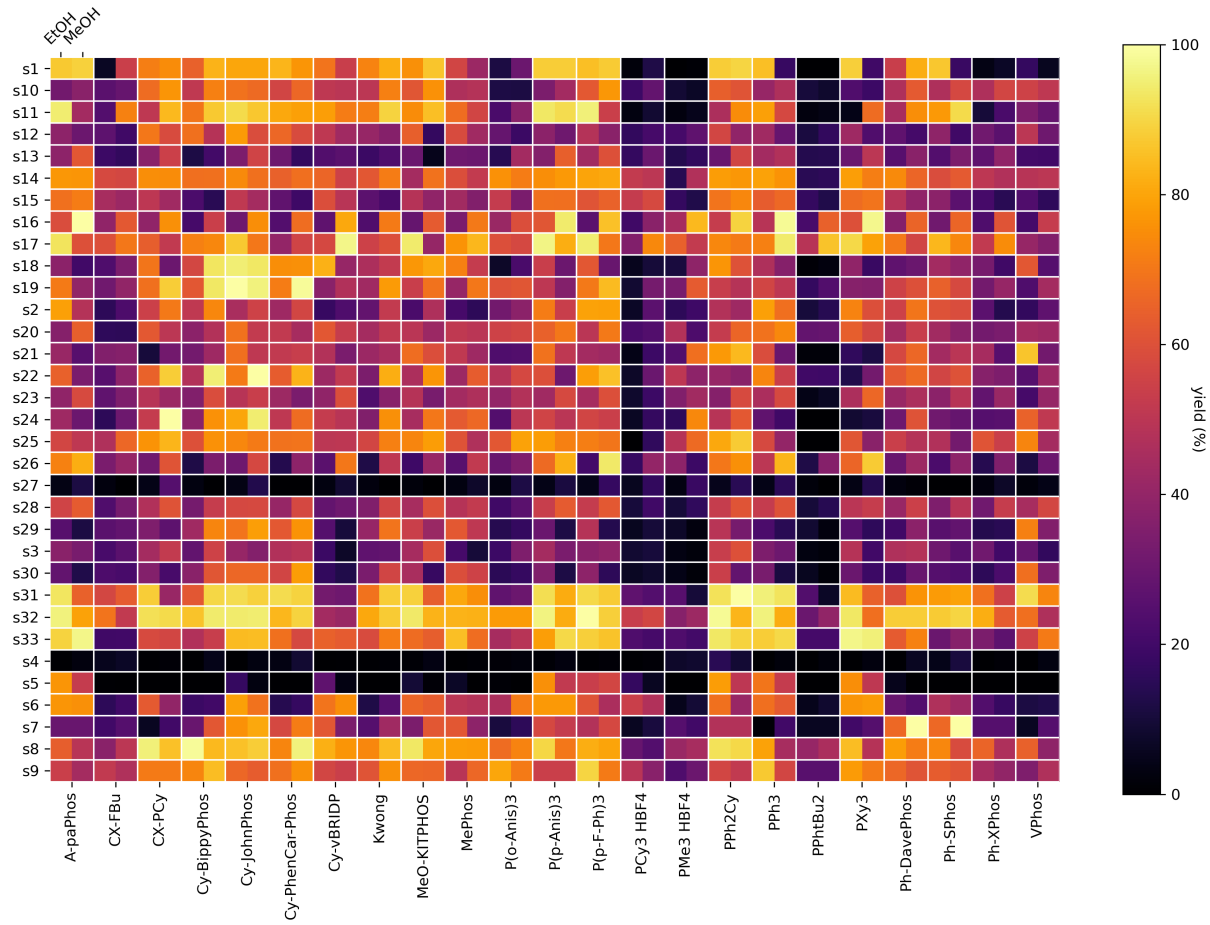


Fig. 73 Visualization of nickel borylation dataset. The y-axis shows electrophiles by their ID, the x-axis shows ligands by names. Reaction yields in both EtOH (left) and MeOH (right) are shown for each electrophile/ligand combination.

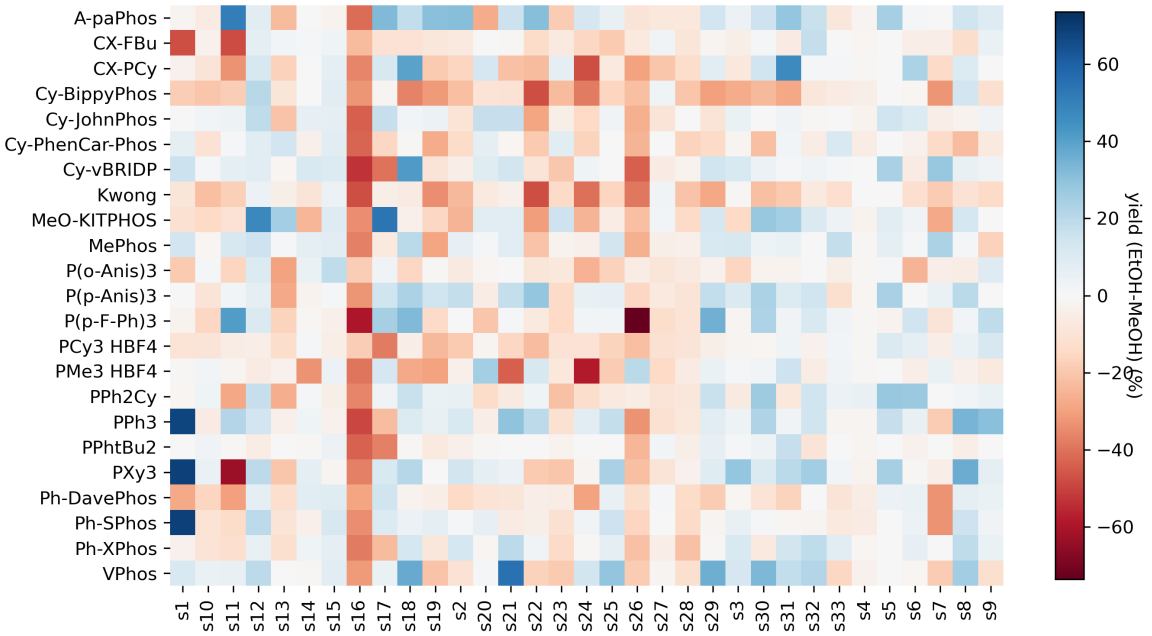


Fig. 74 Visualization of yield difference (EtOH-MeOH) for each electrophile/ligand combination.

For our optimization testing of this dataset with yields as binary rewards, only reactions in EtOH were used, and 50% yield is chosen as a threshold to determine whether a reaction gives satisfying yield or not (hit/no hit). This leaves 759 reactions in total (33 aryl halides, 23 phosphine ligands). The top three and top eight ligands were identified through ranking the number of hits for each ligand across the entire substrate scope (**Fig. 75**, **Fig. 76**). These top ligands align with the top ligands identified in the original publication with standardized Z-score.[66] For the optimization campaign, we only used EtOH data in this hit/no hit (1/0) format with the objective of correctly identifying these top ligands.

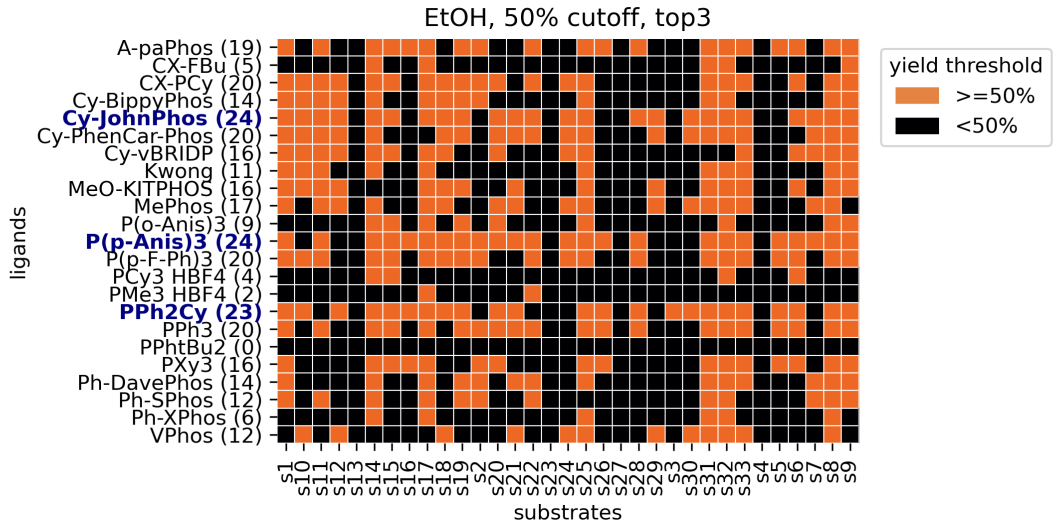


Fig. 75 Top three ligands based on yield threshold (50%) analysis for reactions in EtOH.

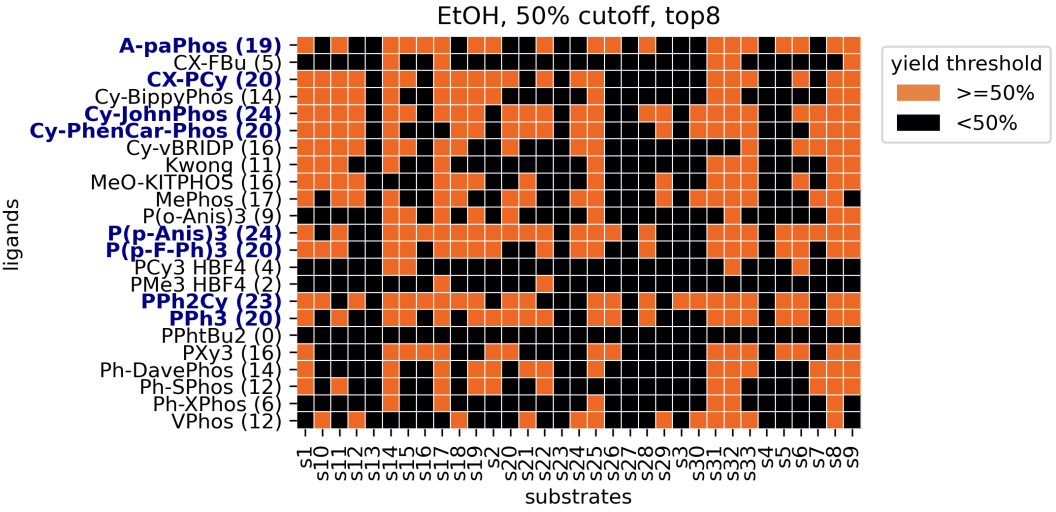


Fig. 76 Top eight ligands based on yield threshold (50%) analysis for reactions in EtOH.

Different algorithms are simulated 500 times with a maximum time horizon of 75 (75 experiments). Accuracy is used as a metric for algorithm performance, where algorithms are tasked to select top-*n* ligands identified through global analysis. Explore-then-commit (ETC) is used as a baseline algorithm. Similar to how the stepwise ETC baseline was calculated for synthetic data (Section 2.4.2), every condition is explored once in each exploration round. At a given time point,

the best condition from all previous, completed exploration rounds is chosen as optimal. This current best condition is temporarily committed until a new round of exploration is completed, which is also when ETC includes the data from the new round and reevaluates the current best condition. The top-n accuracy will get updated after each round, depending on which condition is chosen.

The top-n accuracy is again calculated as the frequency of the identified condition actually being the true optimal condition across all simulations. For ETC, at each time point, the accuracy is the highest ETC accuracy attainable with the maximum number of explorations for each arm. For example, as shown in **Fig. 77**, ETC (black trace) has an accuracy of 0 before t=23, because all 23 ligands are being investigated in the first round of exploration. From t=24 to t=46, the first round of exploration is complete and the accuracy from this round is calculated and shown, while the second round of exploration is underway. In other words, ETC accuracies were updated every 23 experiments (a full round of exploration).

It is also worth noting that, it is actually possible to calculate the accuracy of ETC by enumerating all possible combinations of random samples for each condition, but this calculation becomes exponentially more expensive once the number of samples for each condition reaches 3 and more, which is why we opted for a simulation approach. Because there is no selection of conditions involved, we simulated ETC more extensively (typically 10,000 times) compared to bandit algorithms (typically 500 for reaction dataset) to arrive at a consistent result for the baseline.

The top-3 (**Fig. 77**) and top-8 (**Fig. 78**) accuracies are plotted with various algorithms.

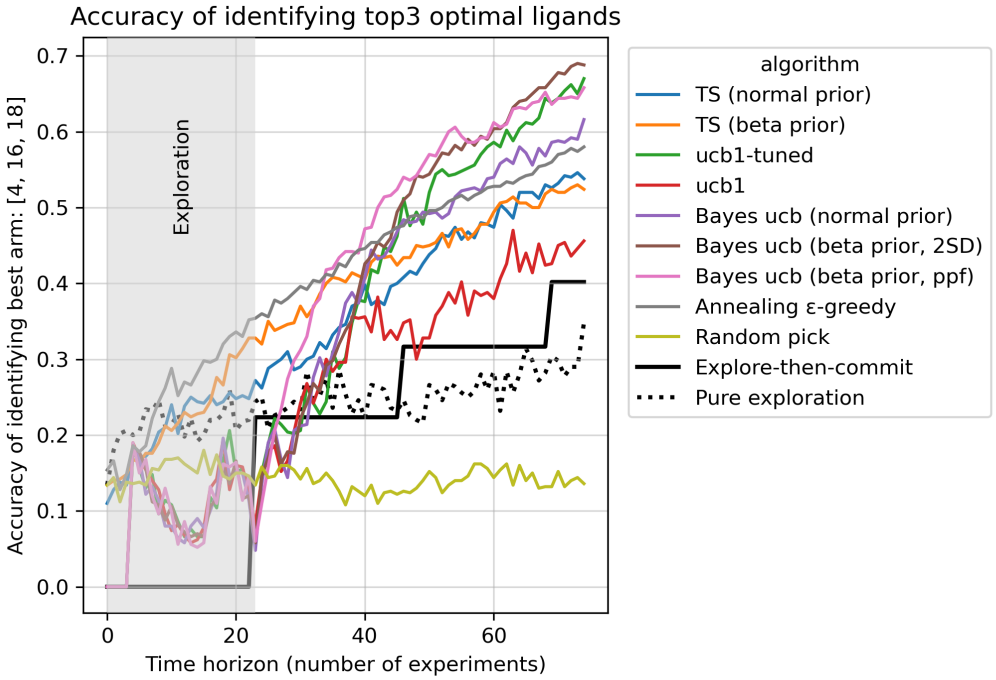


Fig. 77 Top-3 accuracy of identifying optimal ligands in nickel borylation dataset for various algorithms.

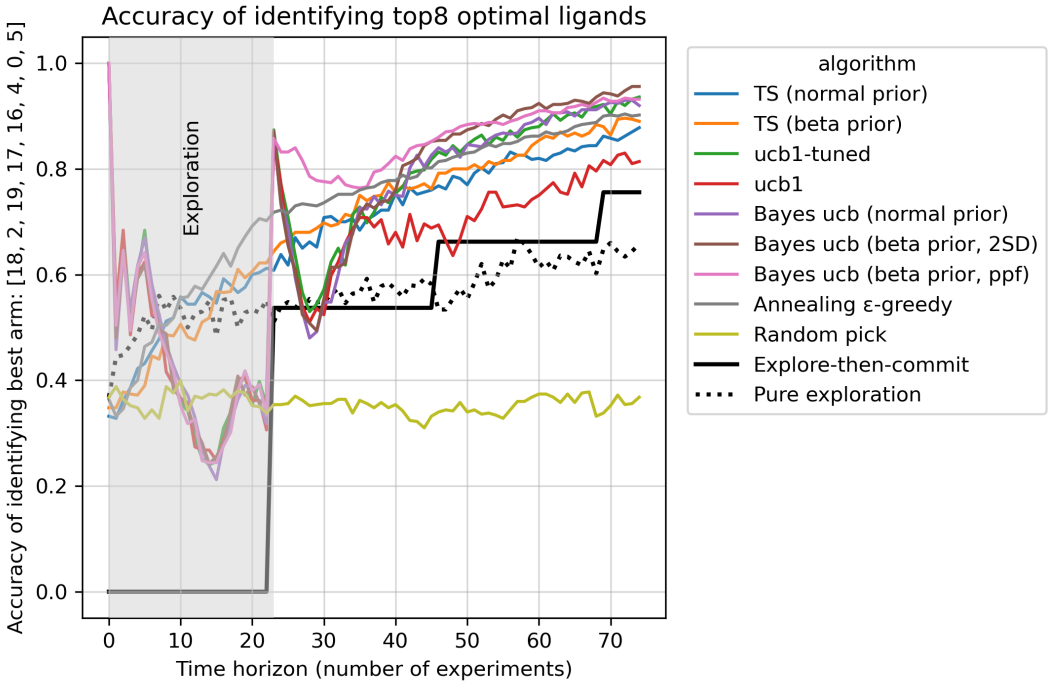


Fig. 78 Top-8 accuracy of identifying optimal ligands in nickel borylation dataset for various algorithms.

Deoxyfluorination

This dataset is extracted from the publication: "Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning." Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. *J. Am. Chem. Soc.* **2018**, *140* (15), 5004–5008. [DOI: 10.1021/jacs.8b01523].[63]

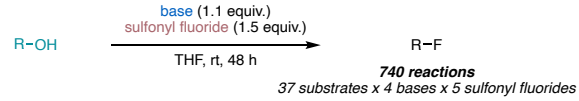
The raw data is used without any additional preprocessing. Some reaction yields reported in the original dataset slightly exceeds 100% due to random analytical errors, all of which we reassigned to 100%. In total, 740 reactions (4 bases, 5 sulfonyl fluorides, and 37 alcohol substrates) are included in the scope (**Fig. 79**). All reaction results (except substrate **S37**) were visualized in a heatmap (**Fig. 80**).

Different metrics were used to evaluate each condition's reaction yields with respect to the entire substrate scope and to determine the most general conditions for the scope. For bases (**Fig. 81**), BTPP and BTMG offer similar performance and outperform MTBD (slightly) and DBU. For sulfonyl fluoride (**Fig. 82**), PBSF significantly outperforms the rest. Base/sulfonyl fluoride combinations are also evaluated with the same set of metrics (**Fig. 83**). Based on these analyses, PBSF/BTPP and PBSF/BTMG are identified as the top-2 most general conditions for this scope.

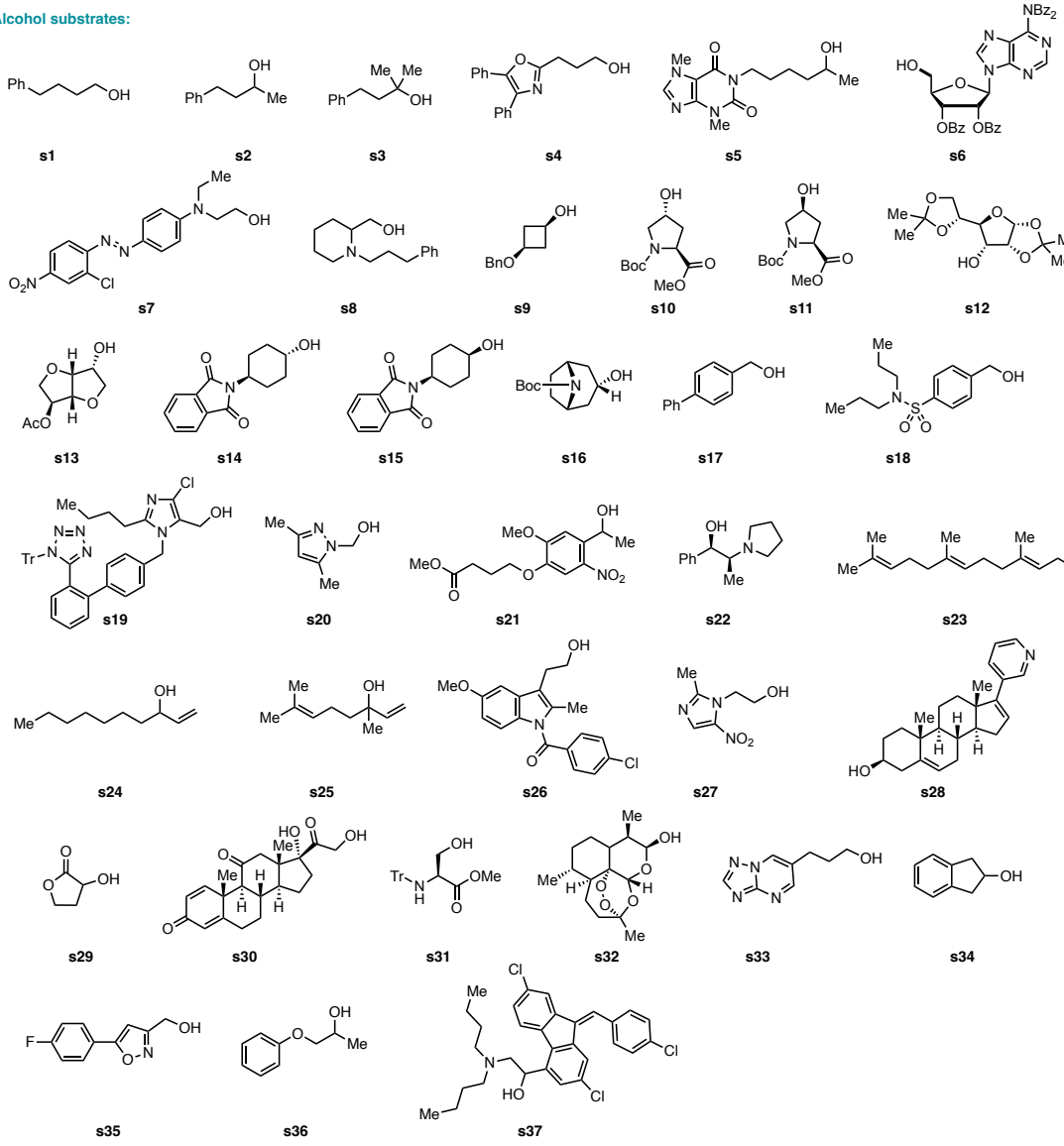
We also visualized optimization results with a traditional model substrate approach. In an optimization campaign for a particular condition component, a model substrate is selected, and reactions are run with different conditions to directly compare their performance with a reactivity threshold (e.g., 75%). For each substrate in our reaction scope, we find the condition that gives the highest yield. If this highest yield is below the yield threshold we set, the optimization is considered as not complete. If this highest yield is above the threshold we set, we choose the condition component that gives this highest yield as optimal. The optimal conditions identified by such model substrate approach are visualized for base (**Fig. 84**), sulfonyl fluoride (**Fig. 85**), and

base/sulfonyl fluoride combination (**Fig. 86**). Around half of the substrate scopes do not give reaction yields higher than the defined threshold of 75%. For the substrates that have conditions with yields above threshold, the "optimal" conditions identified usually cover most of the condition space. In other words, the true optimal condition, or the most general conditions, identified through global analysis for the entire reaction scope are not guaranteed to be found through traditional model substrate approaches.

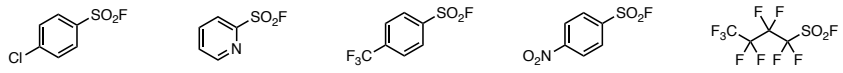
We then tested different algorithms in a continuous [0,1] reward setting (400 simulations, up to 100 total experiments), with the optimization task of finding the optimal base (**Fig. 87**), sulfonyl fluoride (**Fig. 88**) and base/sulfonyl fluoride combinations (**Fig. 89**). Explore-then-commit algorithm was again chosen as the baseline algorithm.



Alcohol substrates:



Sulfonyl fluoride:



Base:

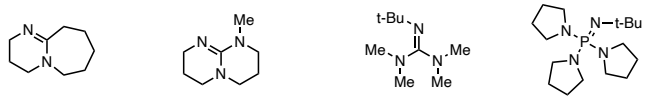


Fig. 79 Deoxyfluorination dataset: alcohol substrates, base and sulfonyl fluoride scope.

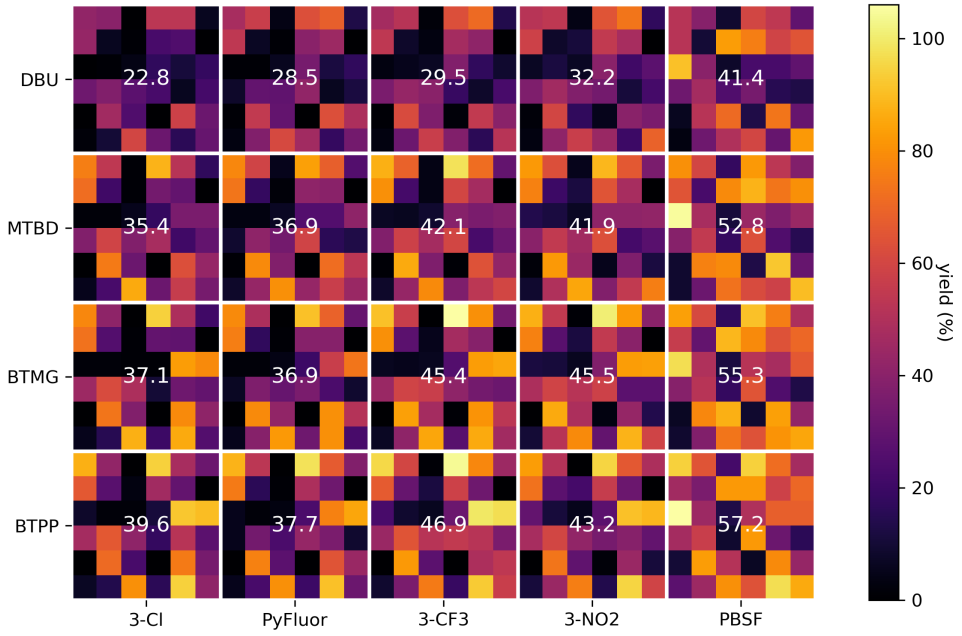


Fig. 80 All results for deoxyfluorination dataset organized by conditions. X-axis represents sulfonyl fluorides, y-axis represents bases. Each colored square represents one substrate, and the same order for substrates are preserved for all condition combination, with S37 omitted for better visualization.

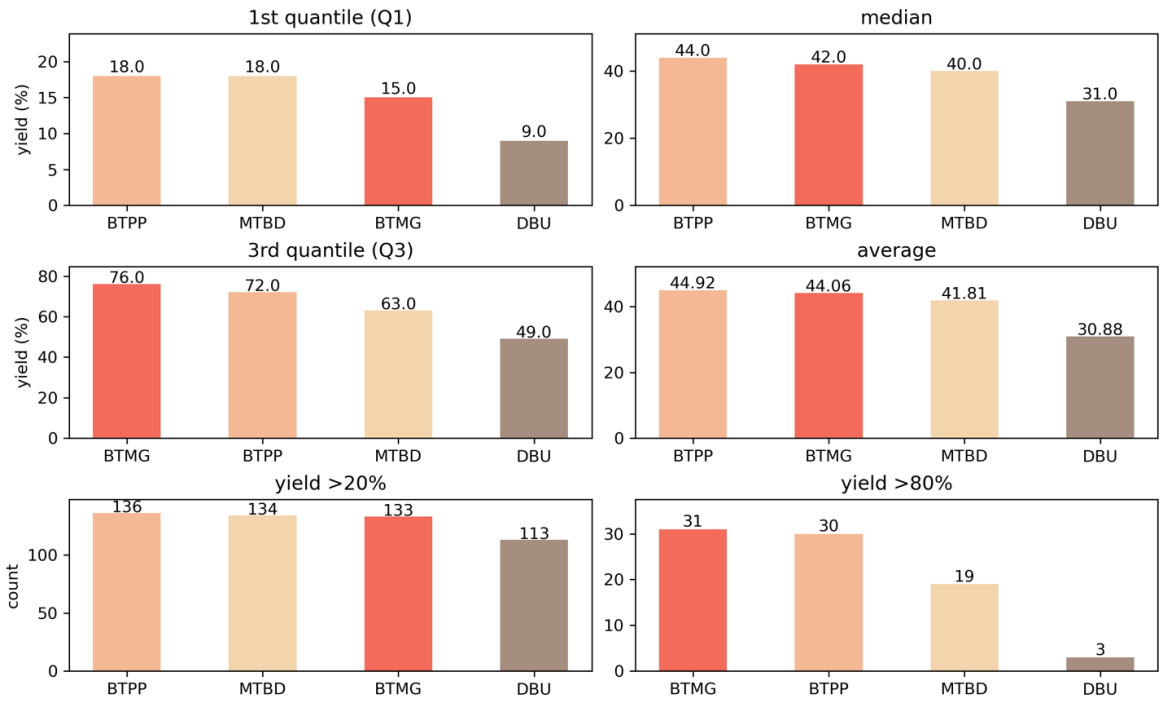


Fig. 81 Different metrics to evaluate base performance in deoxyfluorination dataset (top five for each metric shown).

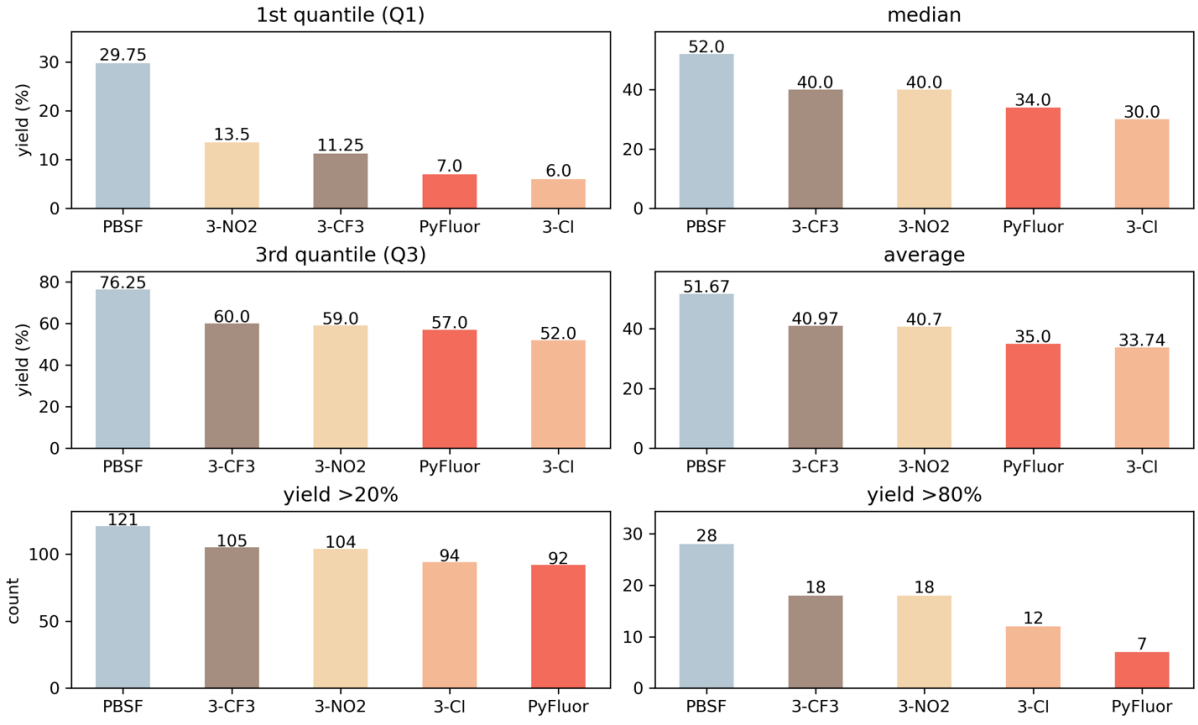


Fig. 82 Different metrics to evaluate sulfonyl fluoride performance in deoxyfluorination dataset (top five for each metric shown).

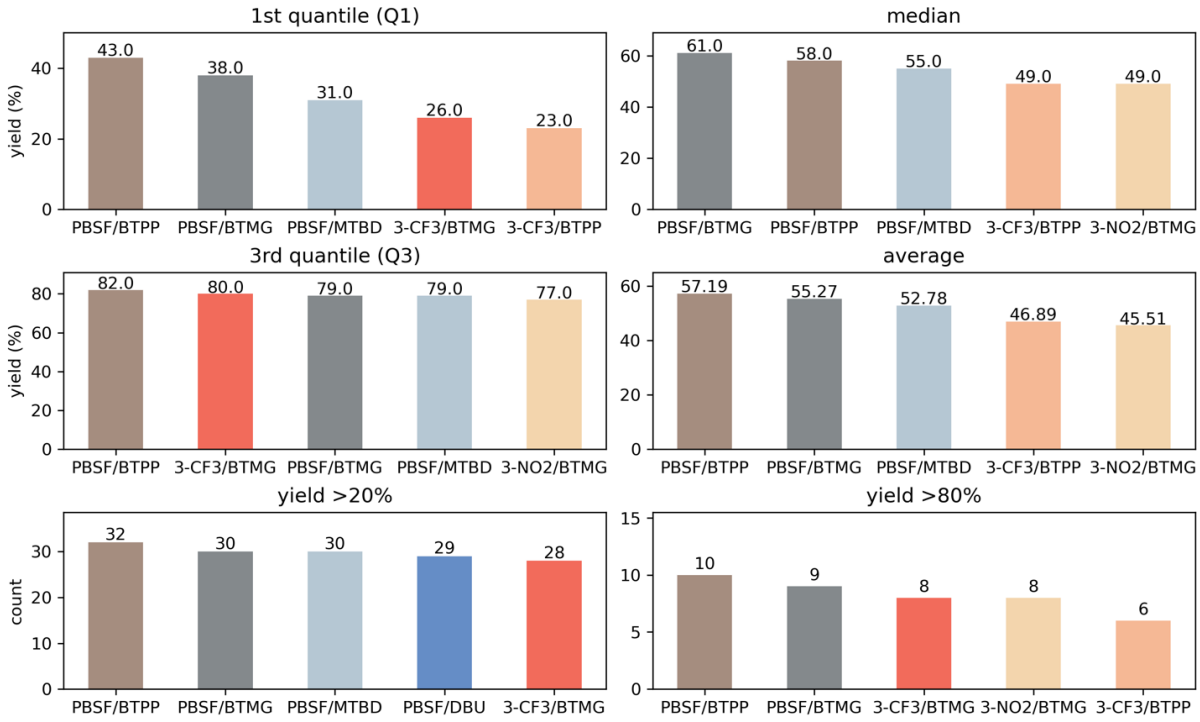


Fig. 83 Different metrics to evaluate base/sulfonyl fluoride combination performance in deoxyfluorination dataset (top five for each metric shown).

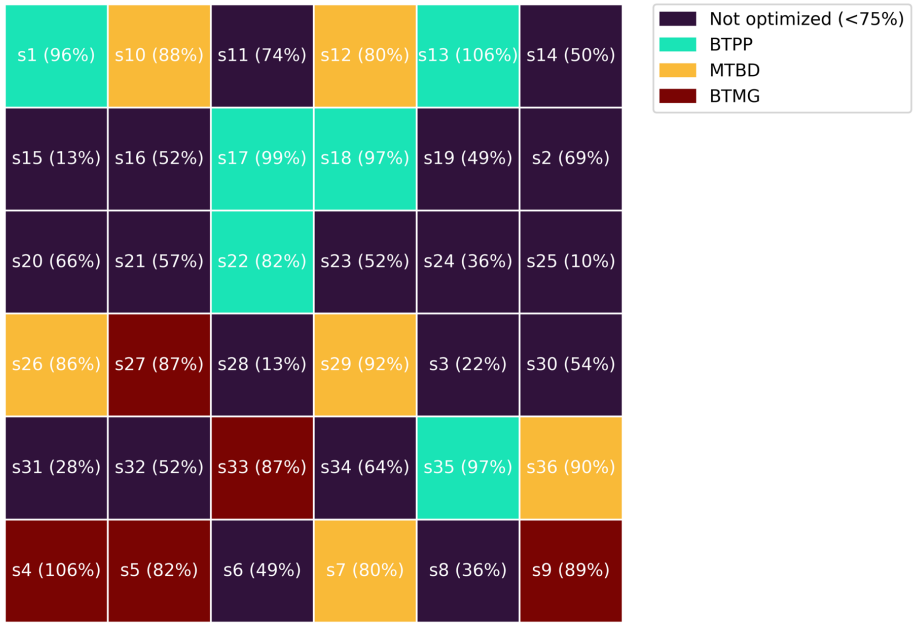


Fig. 84 Optimal base identified through a model substrate approach for deoxyfluorination dataset.

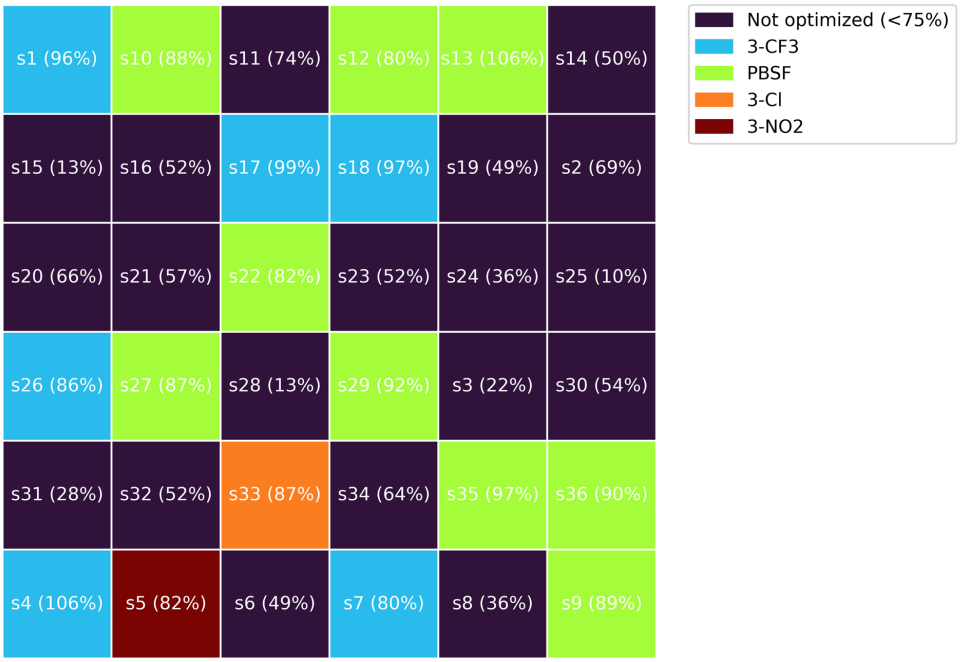


Fig. 85 Optimal sulfonyl fluoride identified through a model substrate approach for deoxyfluorination dataset.

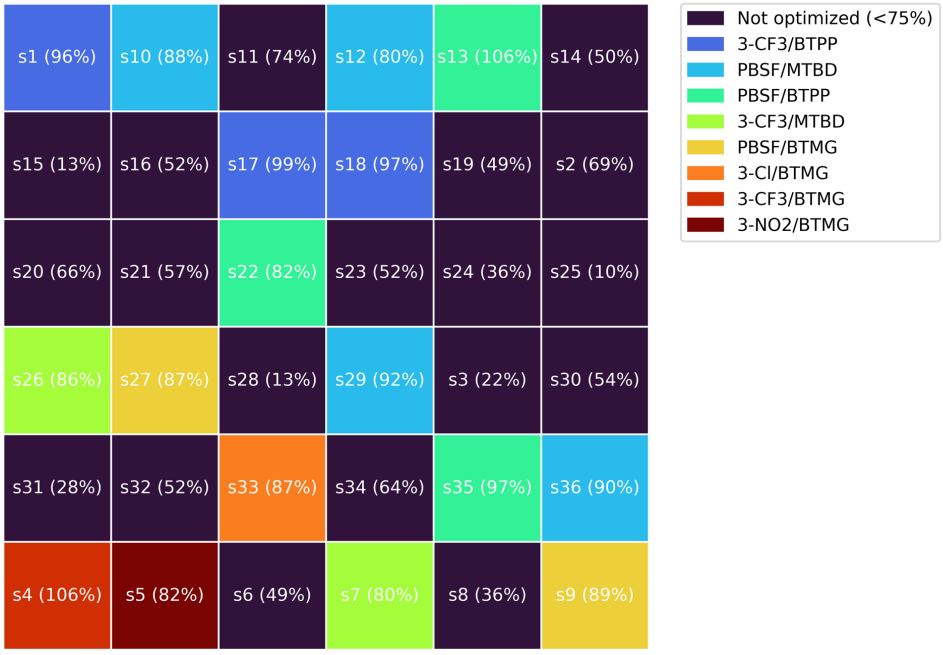


Fig. 86 Optimal base/sulfonyl fluoride combination identified through a model substrate approach for deoxyfluorination dataset.

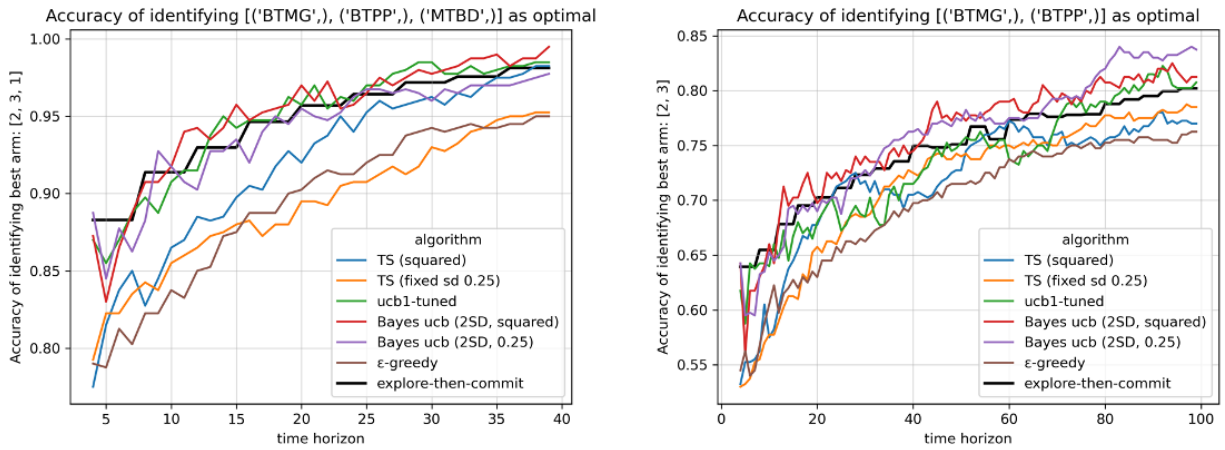


Fig. 87 Accuracy of identifying optimal bases in deoxyfluorination dataset for various algorithms. The plot on the left shows the accuracy of identifying BTMG, BTPP, MTBD in 40 reactions. The plot on the right shows the accuracy of identifying BTMG and BTPP in 100 reactions.

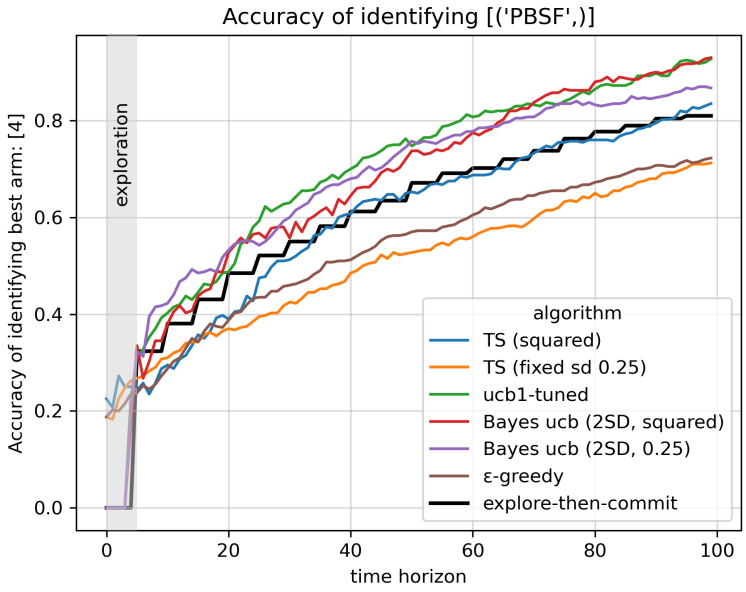


Fig. 88 Accuracy of identifying optimal sulfonyl fluoride (PBSF) in deoxyfluorination dataset for various algorithms.

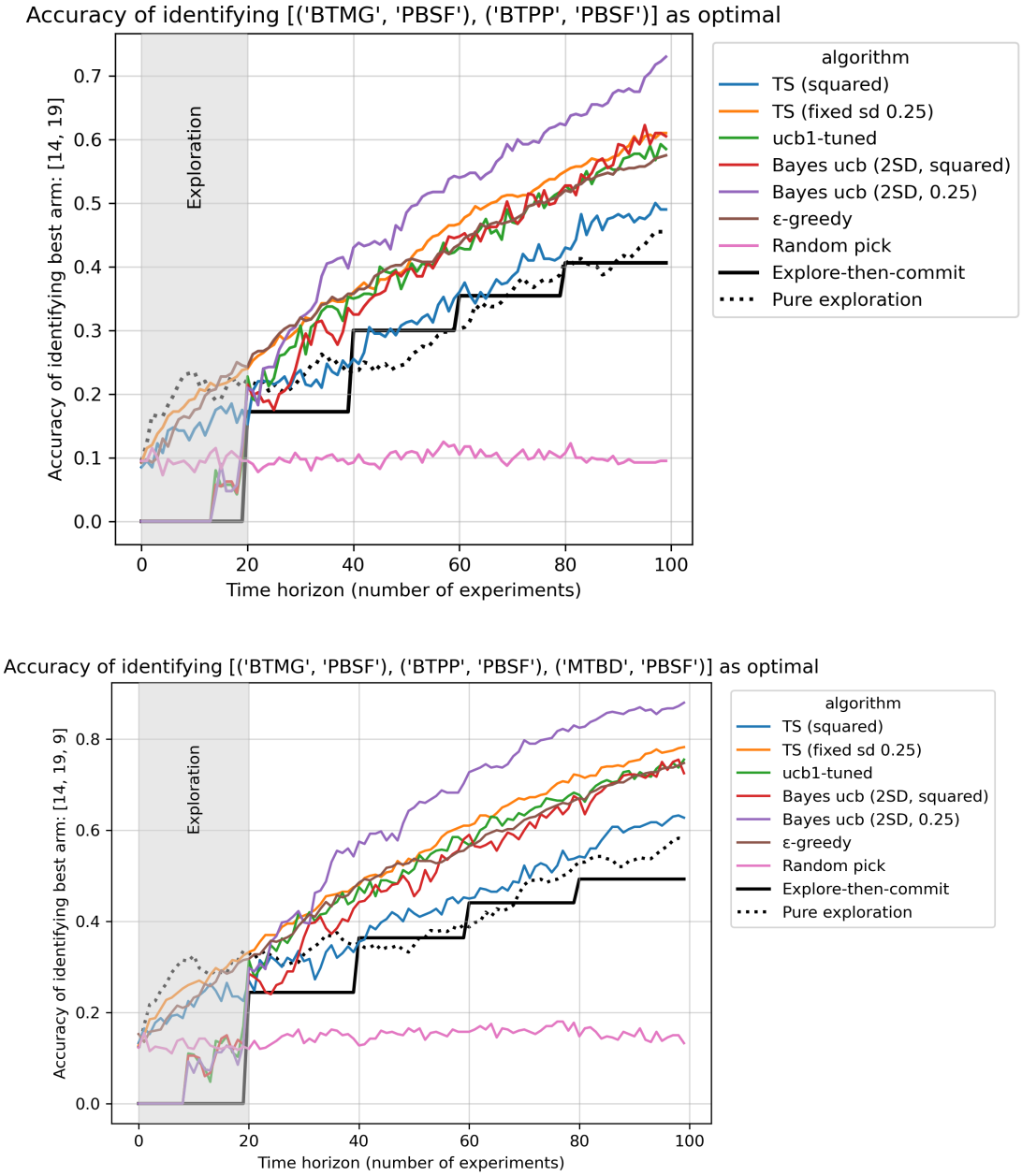


Fig. 89 Top-2 (top) and top-3 (bottom) accuracy of identifying optimal base/sulfonyl fluoride in deoxyfluorination dataset for various algorithms. Top-2: BTMG–PBSF, BTPP–PBSF; top-3: BTMG–PBSF, BTPP–PBSF, MTBD–PBSF.

Buchwald-Hartwig C-N cross-coupling reaction dataset 1 (with yield).

This dataset is extracted from the publication: *"*Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning." Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Science **2018**, 360, 186-190. [DOI: 10.1126/science.aar5169].[67]

The raw data was processed before being used in our optimization simulation studies. Control reactions run without substrates were removed first. Additives that are missing any reaction yields were removed from the scope completely. Finally, substrates and additives are labeled as "s#" and "a#" (#'s are sequentially assigned numbers). This processing leaves 3600 reaction entries overall (15 aryl chloride substrates, 20 additives, 3 based, 4 ligands). Structures for all reaction components are shown in **Fig. 90**. It is worth noting that the isoxazole additives are treated as substrates in this case, because they were used as an alternative way of testing functional group compatibilities if the functional groups present on the isoxazoles were to be present on the actual aryl halide substrates.

All 3600 reactions are visualized with heatmap (**Fig. 91**). The heatmap is divided into sections with each ligand/base combination. For each section, the x axis represents additives and the y axis represents aryl chlorides. Each colored square represents one reaction. A zoomed in visualization for the combination AdBrettPhos–BTMG is shown in **Fig. 92** as an illustration.

A global yield analysis of best-performing conditions (base/ligand) was also conducted (**Fig. 93**). *t*-BuXPhos, *t*-BuBrettPhos, and AdBrettPhos with MTBD as base are the top three most optimal conditions. Various algorithms are simulated (500 simulations, 100 total experiments) and top-1 (**Fig. 94**), top-2 (**Fig. 95**) and top-3 (**Fig. 96**) accuracies are plotted.

In these accuracy plots, a spike in accuracy can often be seen for UCB-type of algorithms at the end of exploration phase. This is due to the fact that UCB-type algorithms require one round

of exploration first (one experiment per arm). Our implementation of this initial exploration behavior is to sample condition arms sequentially based on the order that they are defined (condition 1, condition 2, condition 3…) The top three conditions, MTBD-**L2**, MTBD-**L3**, MTBD-**L4** happen to be the last three conditions defined in this order. Across all simulations, these three conditions are always sampled at the end of the exploration phase, therefore causing a spike in accuracy as they have higher average yields. This spiked accuracy after the exploration round should also be similar to the explore-then-commit accuracy after one round, which is the case as shown in the plots. After the exploration round, the algorithm can freely sample any arm which causes the accuracy to dip. Other algorithms like Thompson sampling or ε-greedy, as shown in the plot, do not produce this artifact as they do not have the exploration requirement.

Note: some results in top-3 accuracy (**Fig. 96**) were presented in **Fig. 10** with the names of the algorithms simplified for clarity. The same algorithm has the same color traces in both figures. For example, TS (implementation 1) in **Fig. 10** corresponds to TS (squared) in **Fig. 96**.

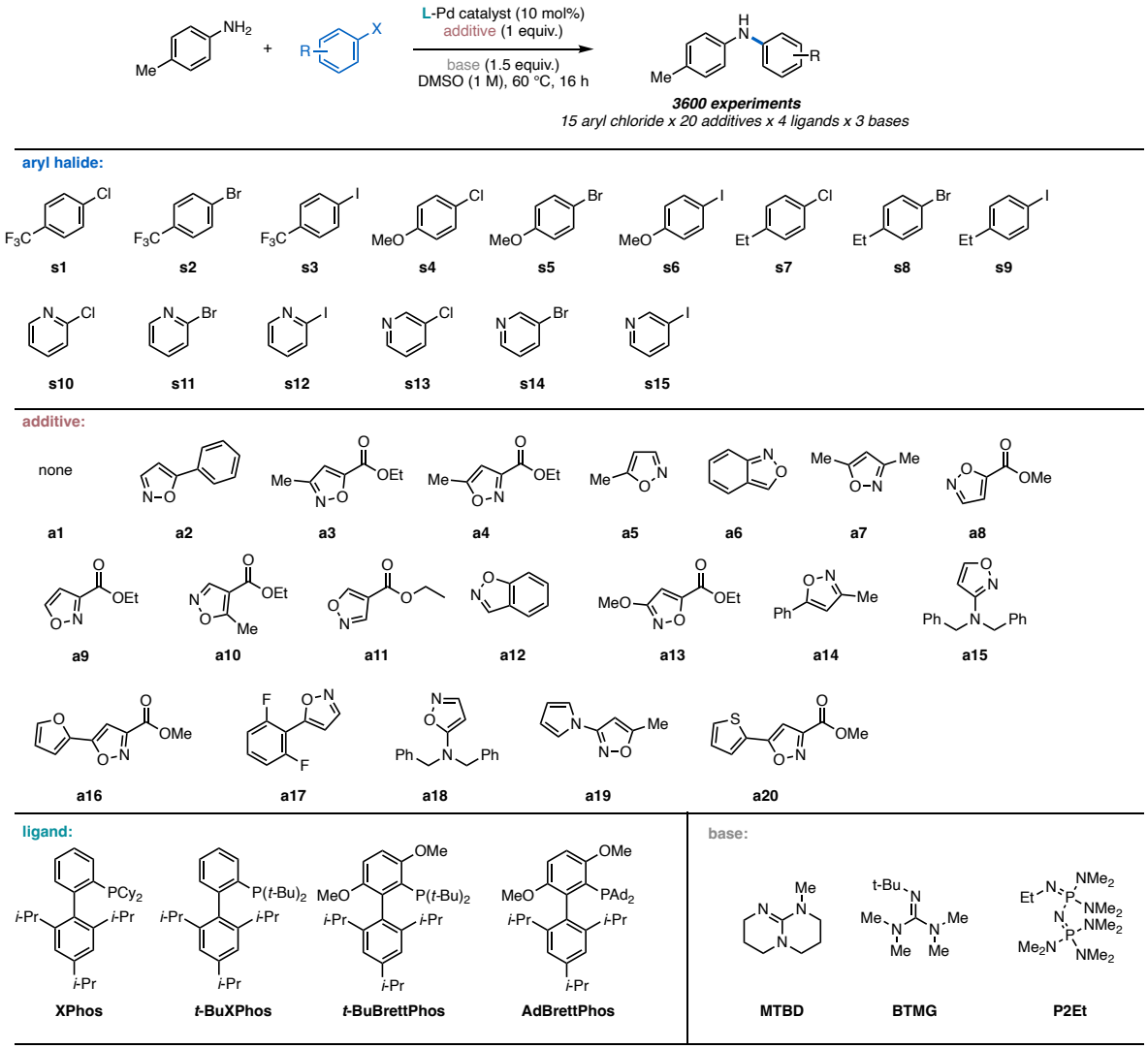


Fig. 90 C-N coupling dataset components: aryl halides, isoxazole additives, ligands and bases.

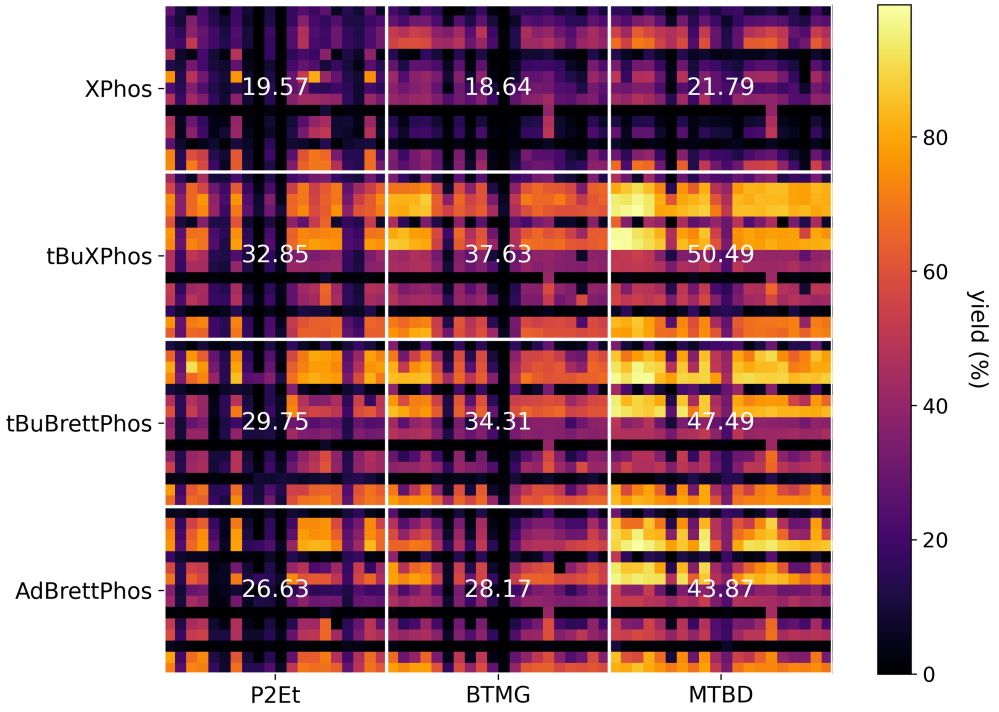


Fig. 91 Heatmap visualization of reaction yields for 3600 Buchwald-Hartwig C-N cross-coupling reactions.

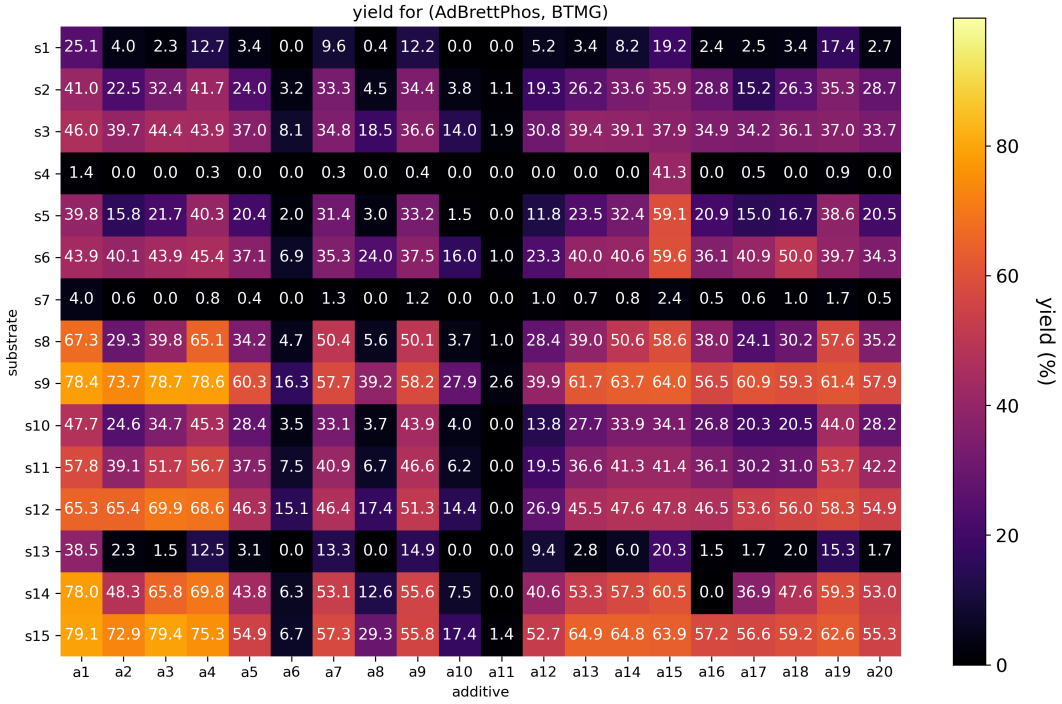


Fig. 92 Heatmap visualization of reaction yields for 300 Buchwald-Hartwig C-N cross-coupling reactions with AdBrettPhos and BTMG.

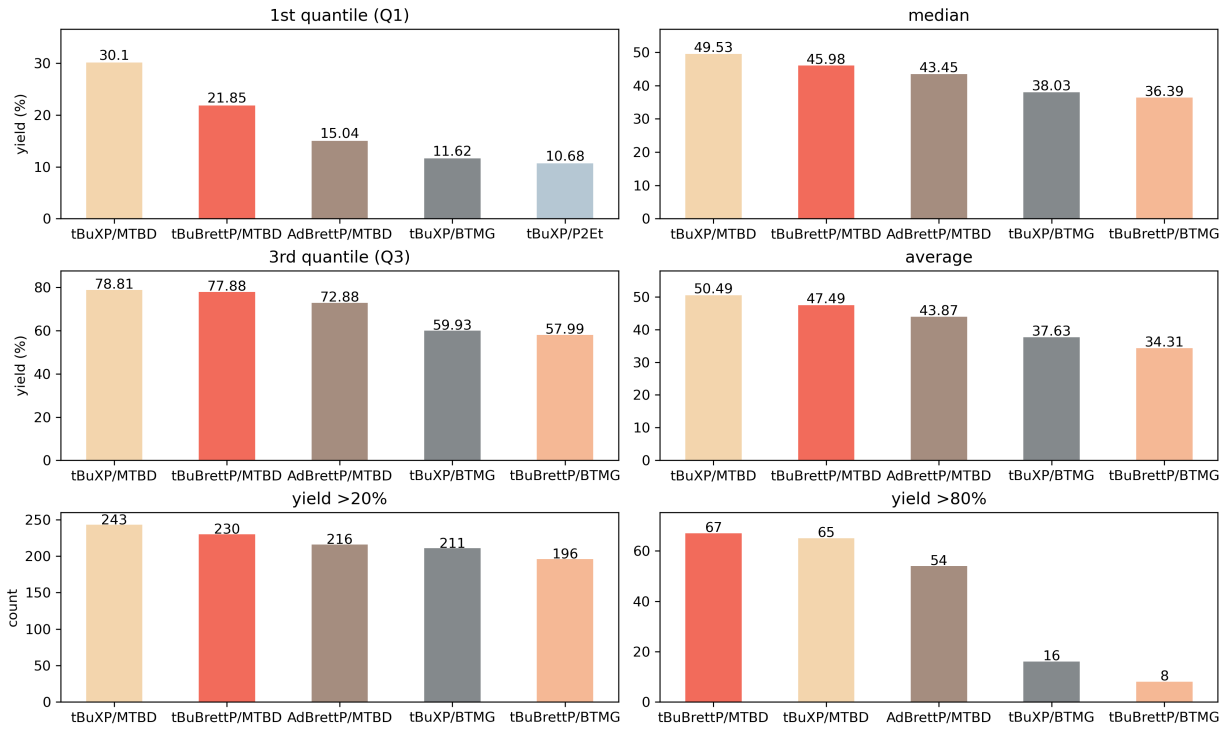


Fig. 93 Different metrics to evaluate ligand/base combination performance in C-N coupling dataset (top five for each metric shown).

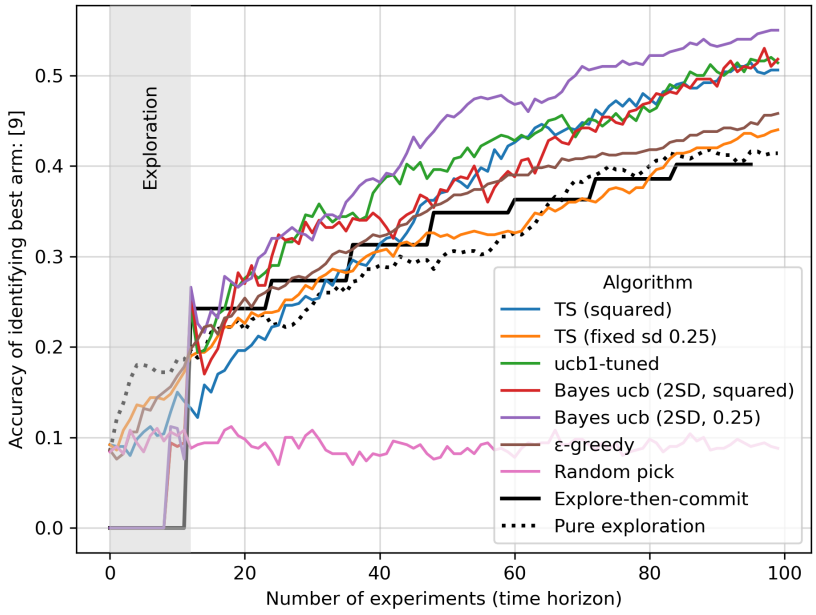


Fig. 94 Top-1 accuracy of identifying optimal base/ligand (MTBD/tBuXPhos) in C-N cross-coupling dataset for various algorithms.

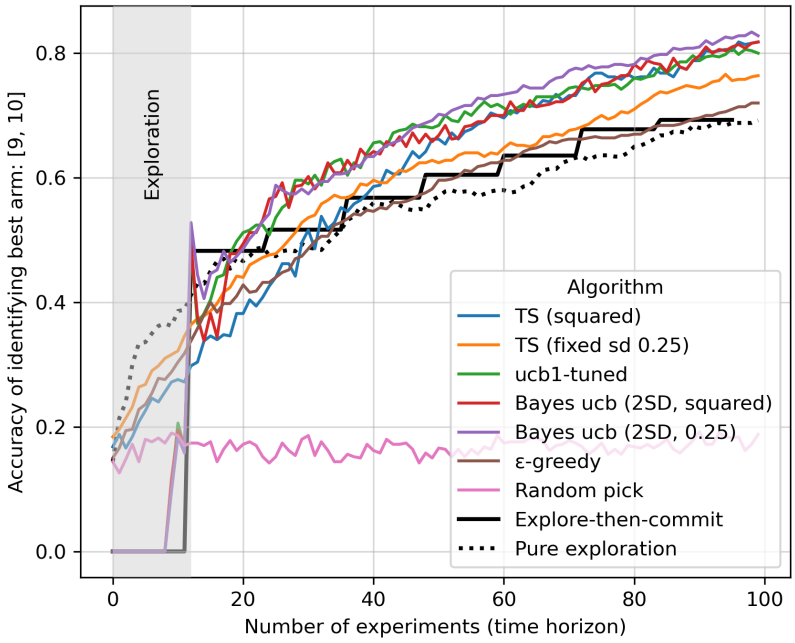


Fig. 95 Top-2 accuracy of identifying optimal base/ligand (MTBD/tBuXPhos, MTBD/tBuBrettPhos) in C-N cross-coupling dataset for various algorithms.

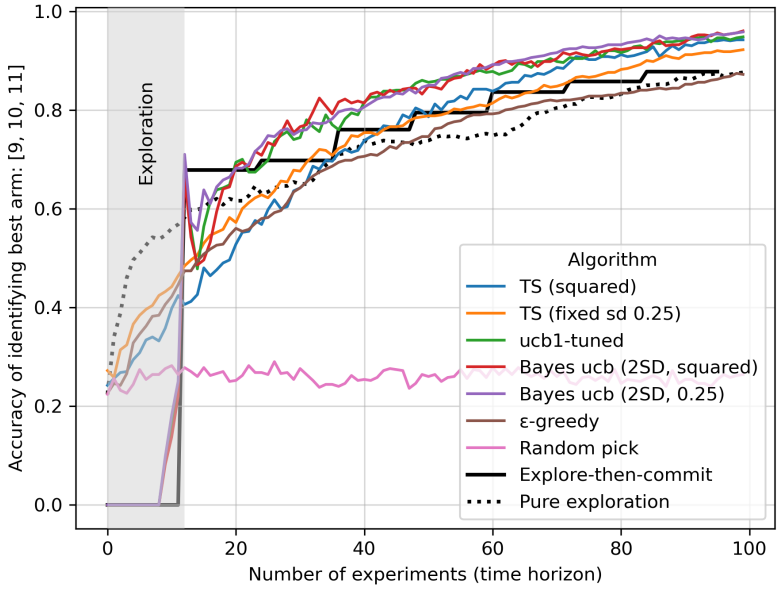


Fig. 96 Top-3 accuracy of identifying optimal base/ligand (MTBD/tBuXPhos, MTBD/tBuBrettPhos, MTBD/AdBrettPhos) in C-N cross-coupling dataset for various algorithms.

Buchwald-Hartwig C-N cross-coupling reaction dataset 2 (no calibrated yield, evaluating four different catalytic methods)

This dataset is extracted from the publication: "Mapping the Dark Space of Chemical Reactions with Extended Nanomole Synthesis and MALDI-TOF MS." Lin, S.; Dikler, S.; Blincoe, W. D.; Ferguson, R. D.; Sheridan, R. P.; Peng, Z.; Conway, D. V.; Zawatzky, K.; Wang, H.; Cernak, T.; Davies, I. W.; DiRocco, D. A.; Sheng, H.; Welch, C. J.; Dreher, S. D. *Science* **2018**, *361* (6402). [DOI: 10.1126/science.aar6236].[70]

For our analysis, we used data from Fig. 3 in the original publication. More specifically, we only used half of the data where the amine partner is fixed, and 192 aryl bromide coupling partners are varied. No calibrated yield was provided for the reactions, but two different analytical metrics, UPLC-MS ion counts and normalized MALDI data, are provided. Since there was poor correlation between the two analytical methods, we only used UPLC-MS ion counts as a readout of reactivity. All ion counts were normalized to [0,1] with the highest ion count in the entire scope being 1. Overall, 768 reactions (4 catalytic conditions, 192 aryl bromide substrates) remain (**Fig. 97**). Because of the large number of substrates used in this reaction, the structures for all aryl bromide coupling partners are not shown here and can be found in the original publication.

The average normalized UPLC-MS ion count responses for each condition is plotted in **Fig. 98**. Surprisingly, copper catalysis conditions perform significantly better than palladium catalysis conditions on average. The two photoredox methods, Ir/Ni and Ru/Ni conditions exhibited almost identical performances, which was expected since these two conditions both rely on the same nickel catalysts for cross coupling and only differ in the photocatalysts used.

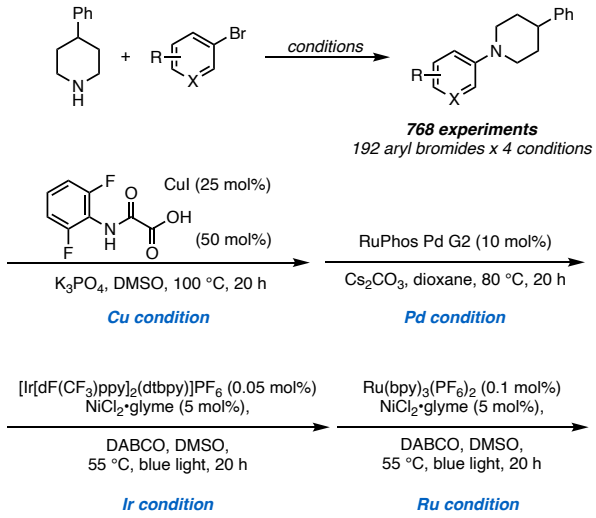


Fig. 97 C-N cross-coupling reaction dataset that evaluates amine scope with four different catalytic conditions.

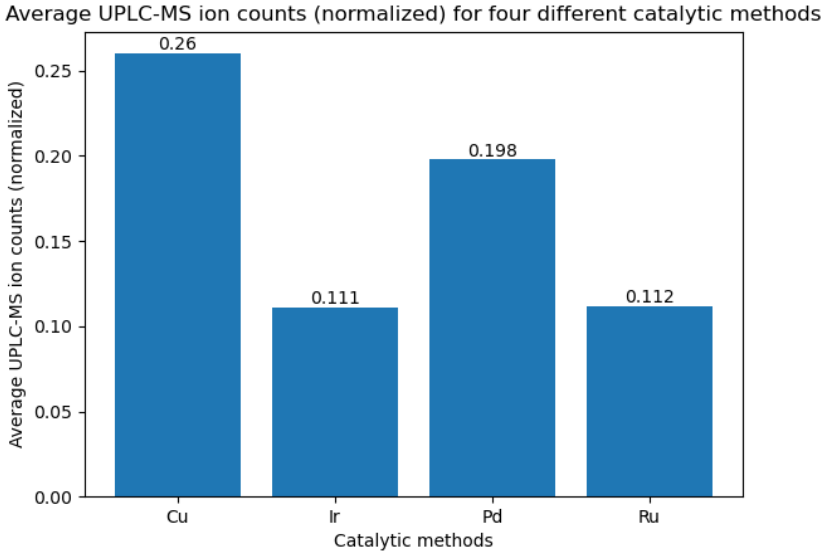


Fig. 98 Average UPLC-MS ion counts (normalized) for four different catalytic methods.

Again, we simulated this dataset with our optimization algorithms, with the objective of identifying copper catalysis condition as the most general condition. Overall, several of the algorithms converged to >95% accuracy within 200 experiments. The top-performing UCB1 and UCB1-tuned algorithms reached 95% accuracy after 100 experiments (**Fig. 99**).

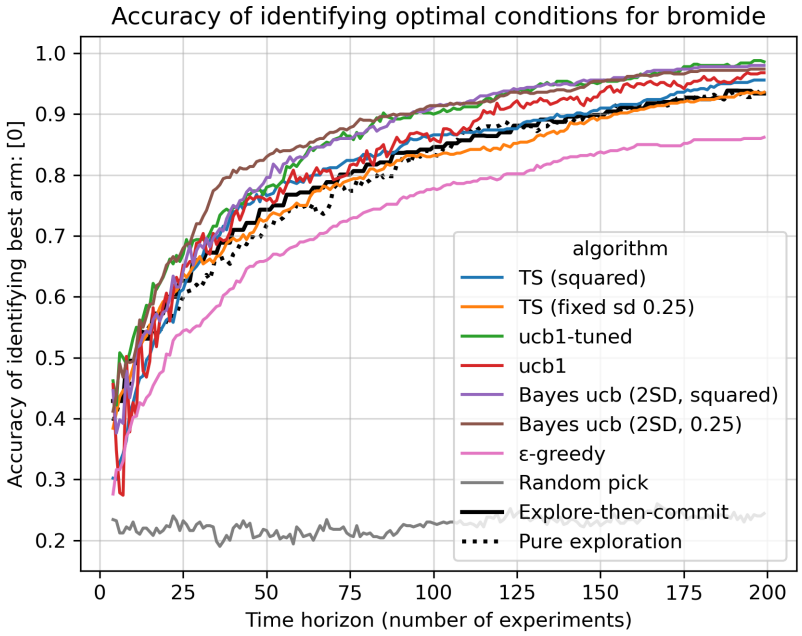


Fig. 99 Top-1 accuracies of identifying optimal condition (copper catalysis conditions) in C-N cross-coupling dataset for various algorithms.

Imidazole C-H direct arylation dataset

Following a previous imidazole C–H direct arylation dataset investigated by the Doyle group and BMS, where 1728 combinations of conditions are evaluated with one pair of coupling partners, we designed a new dataset to focus on substrate effect with both coupling partners (aryl bromides and imidazoles) and their reactivities with different ligands. Evidenced by some of the previous datasets we have simulated in Section 2.4.8, one of the biggest hurdles for a large reaction dataset is to obtain rection yields. Alternative readouts for reactivity, such as additive screening and processed instrument responses are typically used. While simple and quite effective, these readouts still cannot replace actual reaction yields, which allows the direct comparison between reaction conditions and their performances with different substrates. For the C–H arylation dataset,

we decided to explore the substrate dimension extensively with different phosphine ligands and obtain calibrated reaction yields for all reactions.

The first step to establish the dataset is to select the substrate scope (aryl bromides and imidazoles) and ligand scope. We started the selection process with the aryl bromide, imidazole and monophosphine ligand libraries at BMS. By clustering the respective libraries via unsupervised learning approach (e.g., *k*-Medoids clustering with Mordred descriptors, selecting *k* by minimizing silhouette scores), we obtained clusters of structurally similar molecules. For each library, we then manually selected the molecules from each cluster based on their availability and structural features of interest. Overall, 8 aryl bromides and 8 imidazoles are selected as the substrate scope, and 24 monodentate phosphine ligands are selected as the condition scope (**Fig. 100**). All 64 products are synthesized, and all 1536 experiments are carried out and analyzed to obtain calibrated reaction yield. The experimental details of the product synthesis and HTE reactions can be found in 2.5.1.

It is also worth noting that, the original plan for the substrate scope involves 10 aryl bromides and 10 imidazoles. Two of the aryl bromides and two of the imidazoles were later excluded from the substrate scope due to their products' low reactivities and challenging isolations during the authentic product synthesis stage. To ensure consistency in all experimental records, the original labels of substrates were kept for the remaining substrates (hence the non-continuous numeric and alphabetical labels). The substrates excluded are shown in **Fig. 101**.

Results for all experiments are visualized as a heatmap grouped by ligand first, then by substrate pairings (**Fig. 102**). For all 64 substrate pairs, the median and average yields are shown for results with 24 ligands (**Fig. 103**) as a metric of general reactivities. Categorial bar plots for aryl bromides (**Fig. 104**) and imidazoles (**Fig. 105**) with binned yields are also shown. For ligand

performances, a categorical bar plot (**Fig. 106**) and a box plot (**Fig. 107**) are shown, as well as a ranking of top-10 ligand performance based on various metrics (**Fig. 108**). Based on these analyses, Cy-BippyPhos, Et-PhenCarPhos, tBPh-CPhos, CgMe-PPh, and JackiePhos are identified as the top-5 ligands (by average) and used in the optimization studies.

We also visualized optimization results obtained with a model substrate approach, as discussed in the manuscript. The results are shown with a 50% (**Fig. 109**), 75% (**Fig. 110**), and a 90% yield cutoff (**Fig. 111**). Various optimization algorithms were also tested and top-1 (**Fig. 112**), top-5 (**Fig. 113**) and top-9 (**Fig. 114**) accuracies were plotted with ETC baselines (top 1 and top 9 ligands were also determined from average yields in **Fig. 108**).

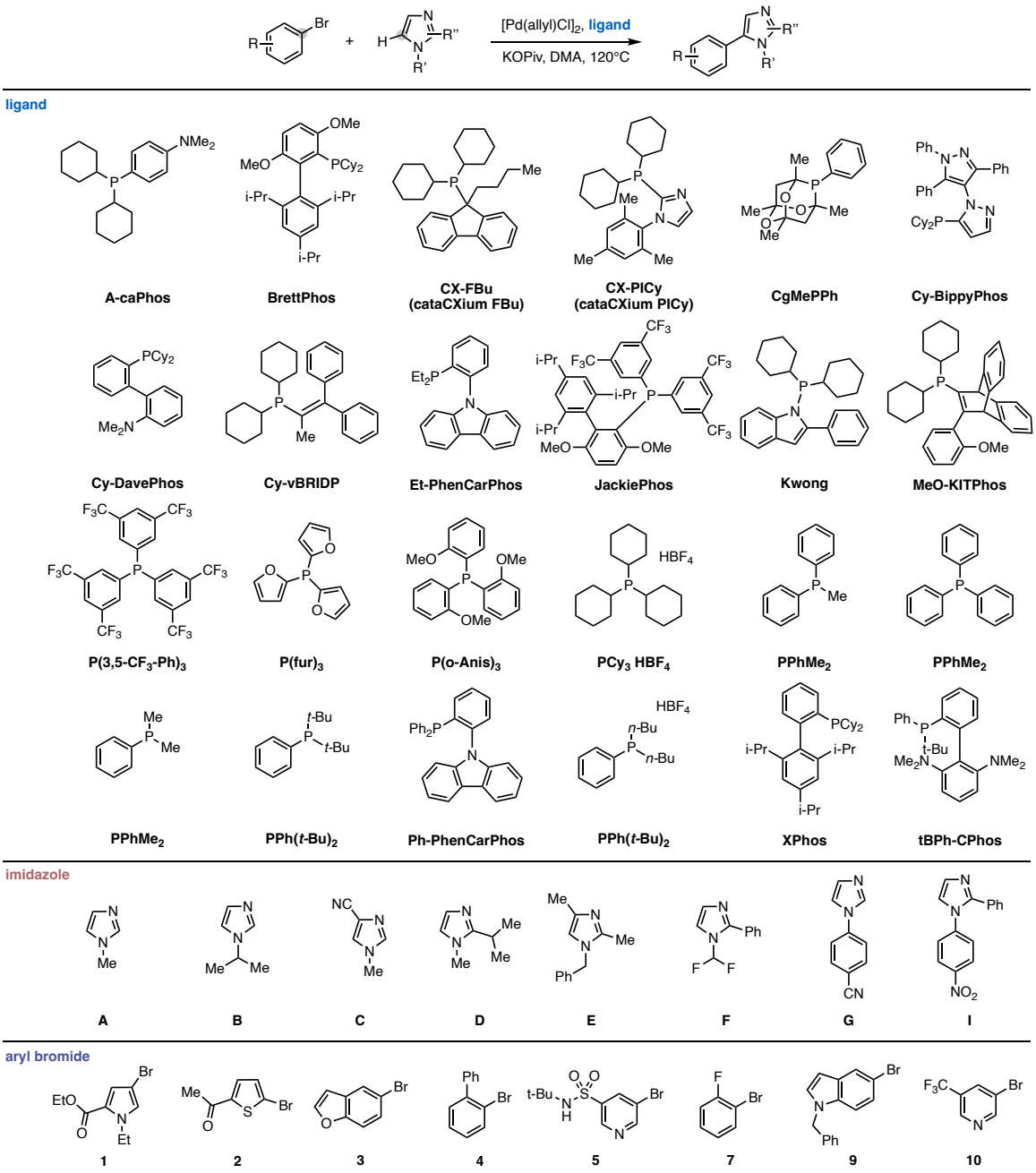


Fig. 100 C–H arylation dataset components: ligands, imidazoles, aryl bromides.

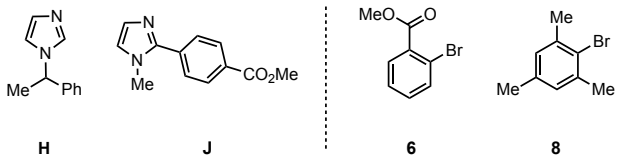


Fig. 101 Two imidazoles and two aryl bromides removed from the planned substrate scope.

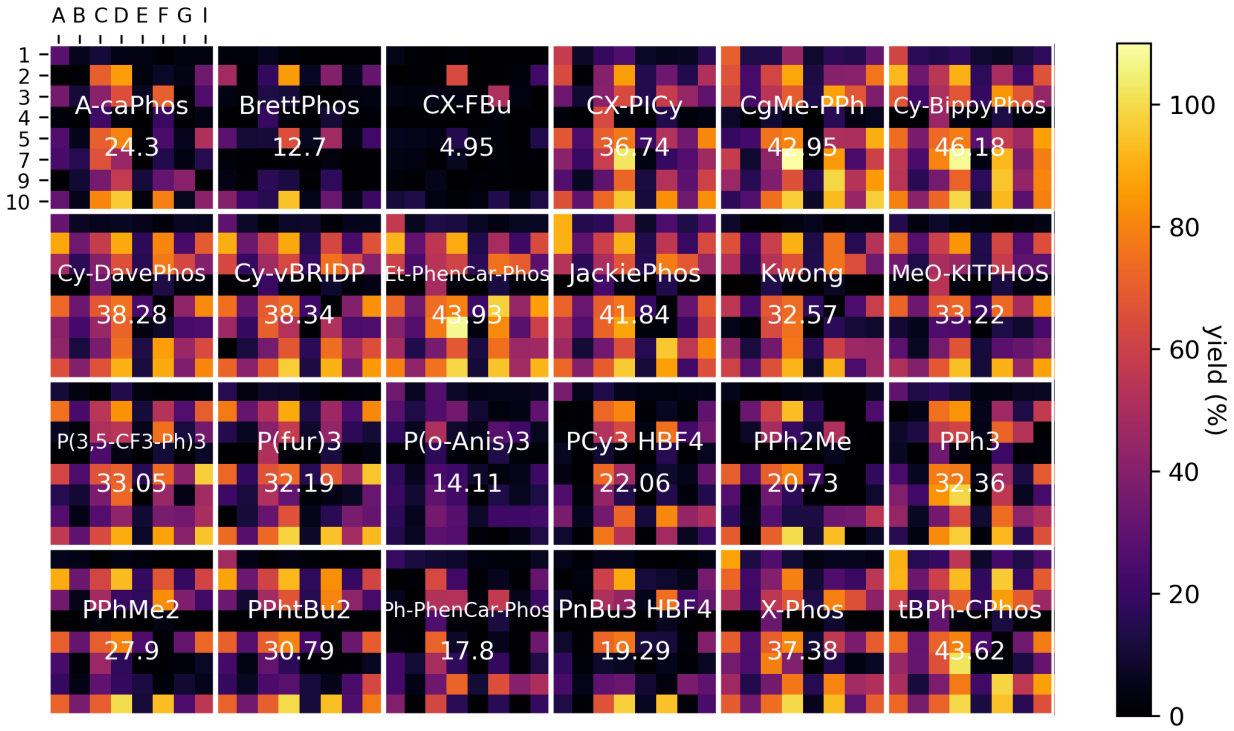


Fig. 102 Heatmap visualization of reaction yields in the imidazole C–H arylation reaction.

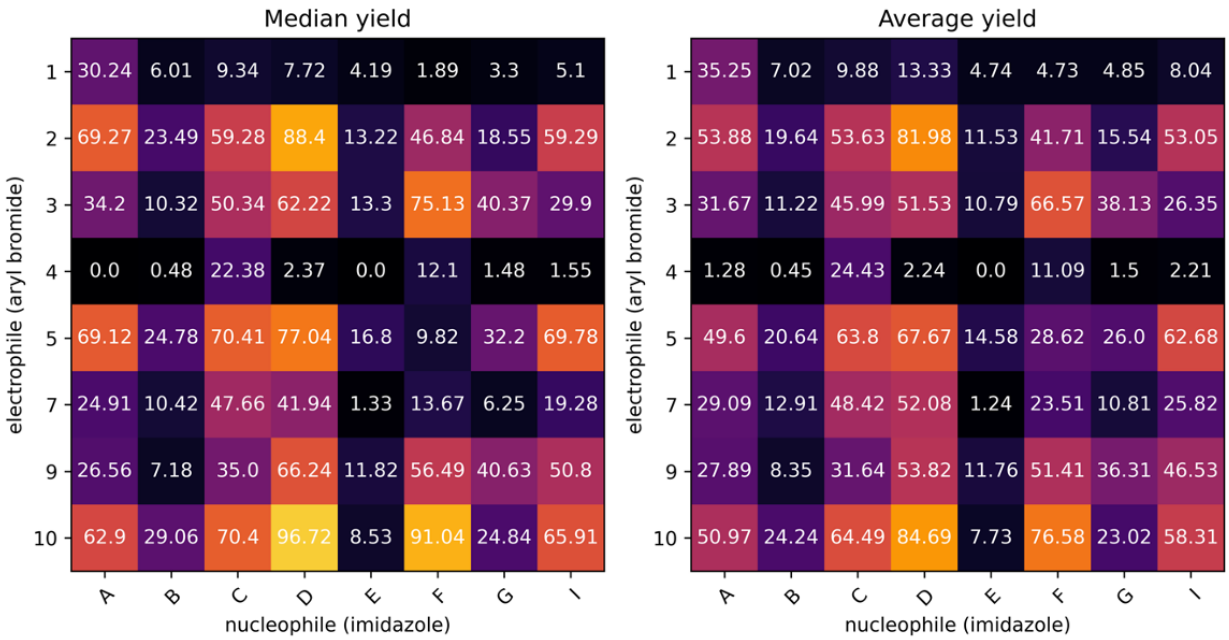


Fig. 103 Median (left) and average (right) reactions yields across 24 ligands for all 64 products in the imidazole C–H arylation reaction.

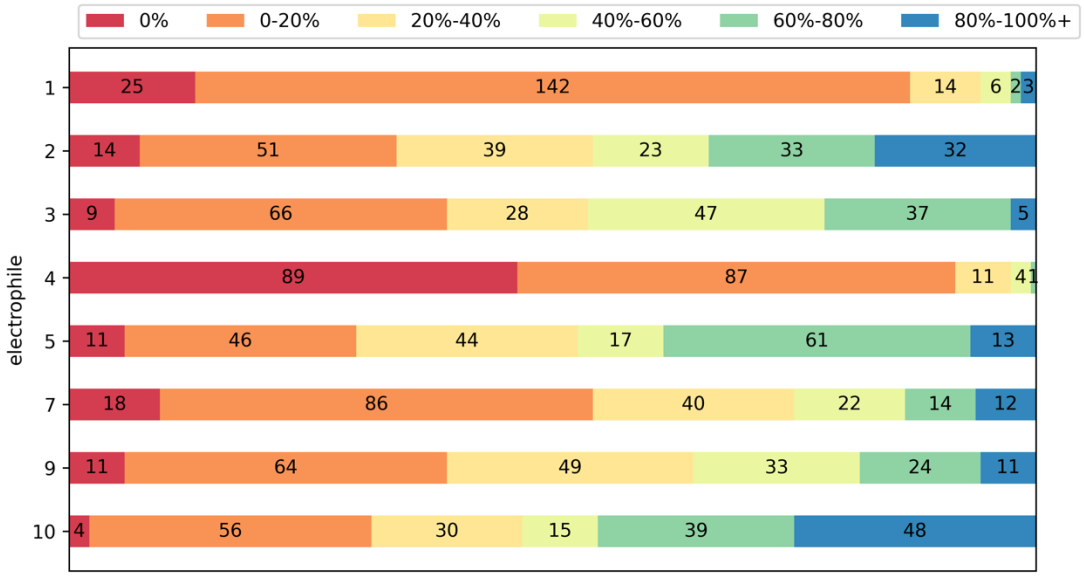


Fig. 104 Categorical bar plot of reaction yields for 8 aryl bromides (electrophiles) in the imidazole C–H arylation reaction.

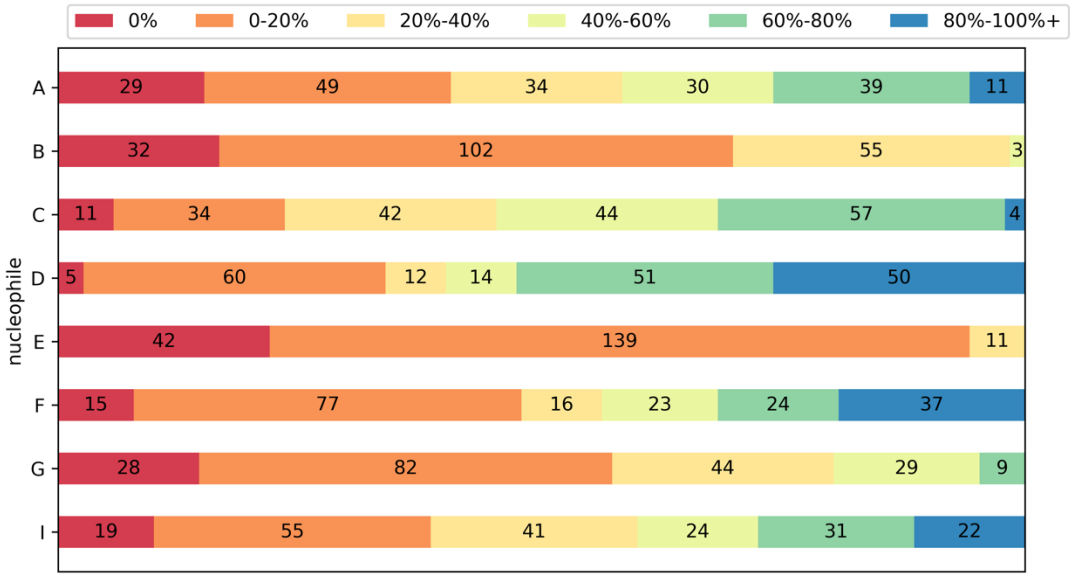


Fig. 105 Categorical bar plot of reaction yields for 8 imidazoles (nucleophiles) in the imidazole C–H arylation reaction.

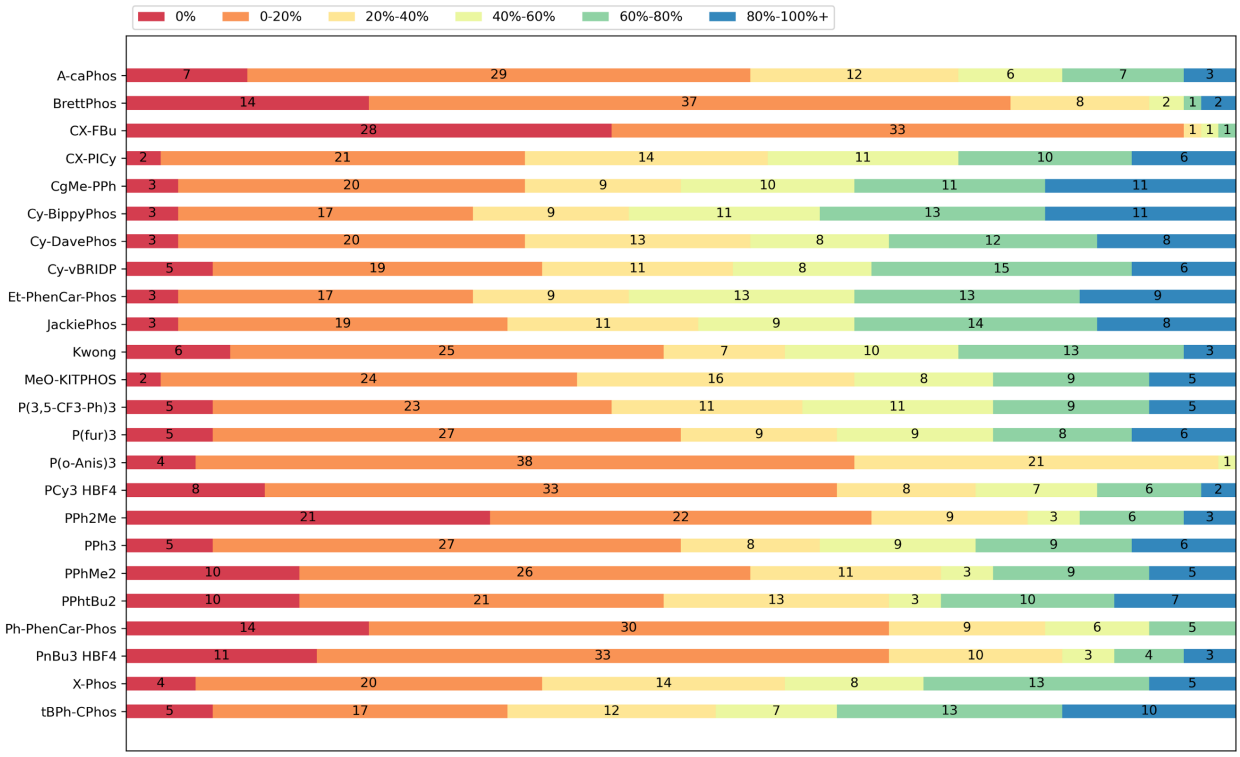


Fig. 106 Categorical bar plot of reaction yields for 24 ligands in the imidazole C–H arylation reaction.

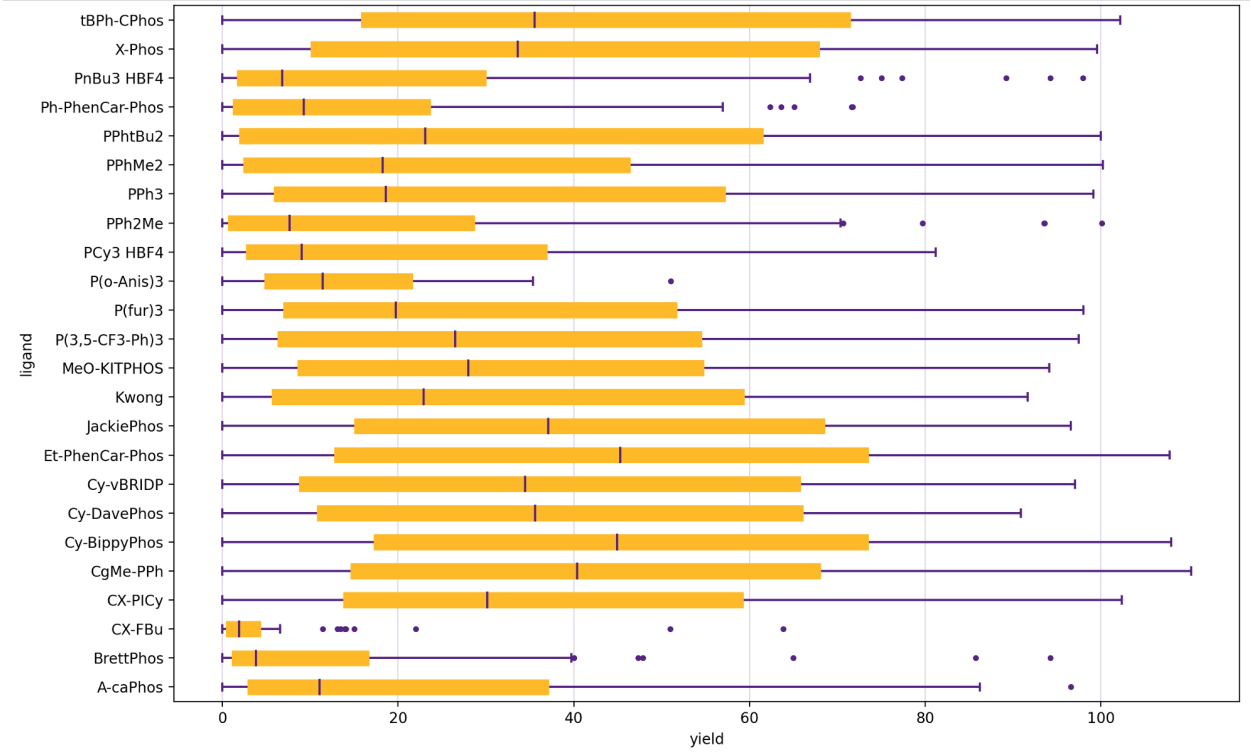


Fig. 107 Box plot of reaction yields for 24 ligands in the imidazole C–H arylation reaction.

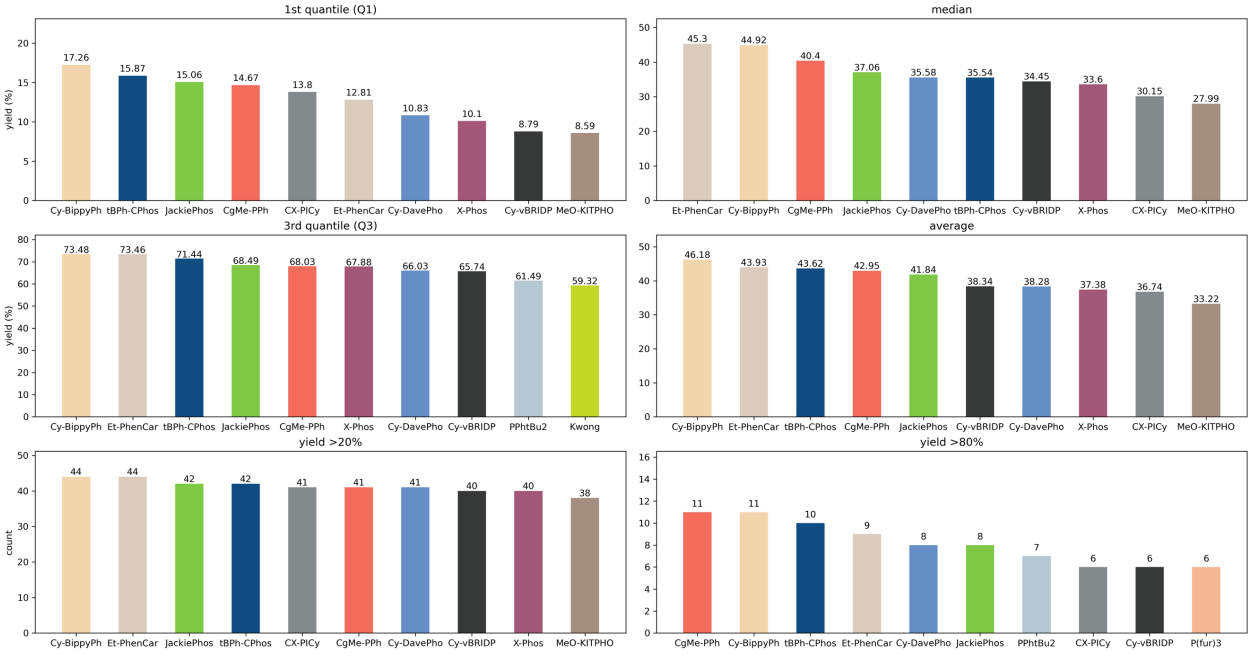


Fig. 108 Different metrics to evaluate ligand performance in the imidazole C–H arylation reaction (top ten for each metric shown).

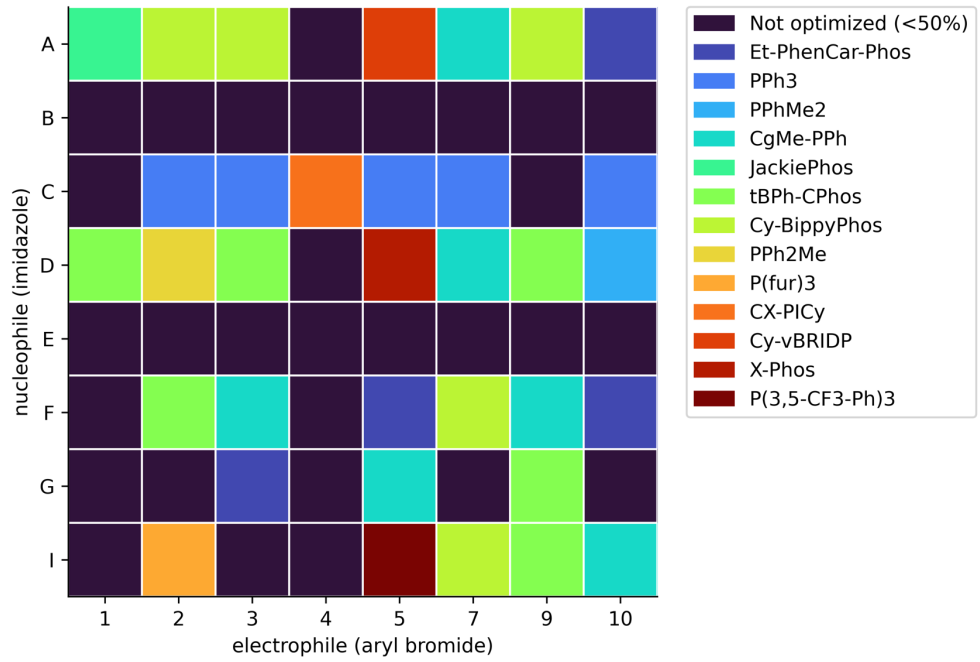


Fig. 109 Model substrate optimization results using a 50% yield cutoff.

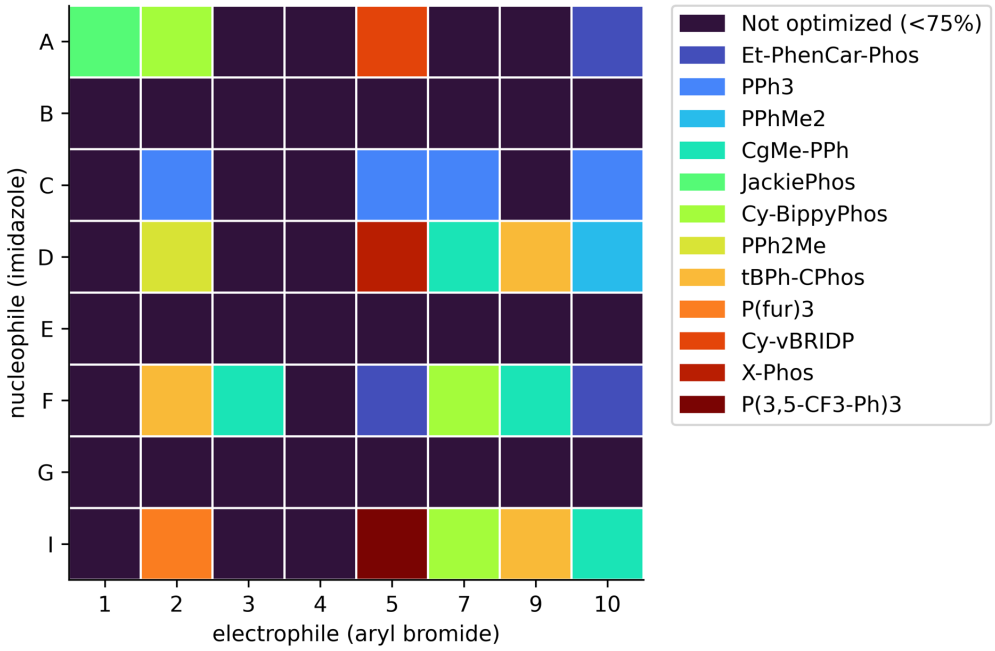


Fig. 110 Model substrate optimization results using a 75% yield cutoff.

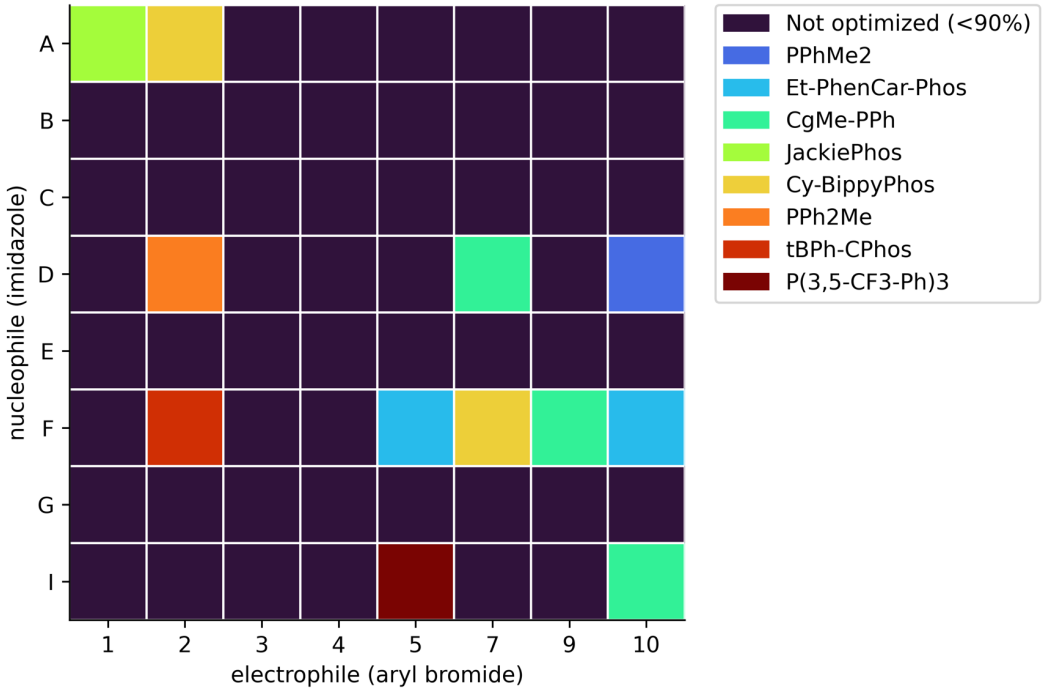


Fig. 111 Model substrate optimization results using a 90% yield cutoff.

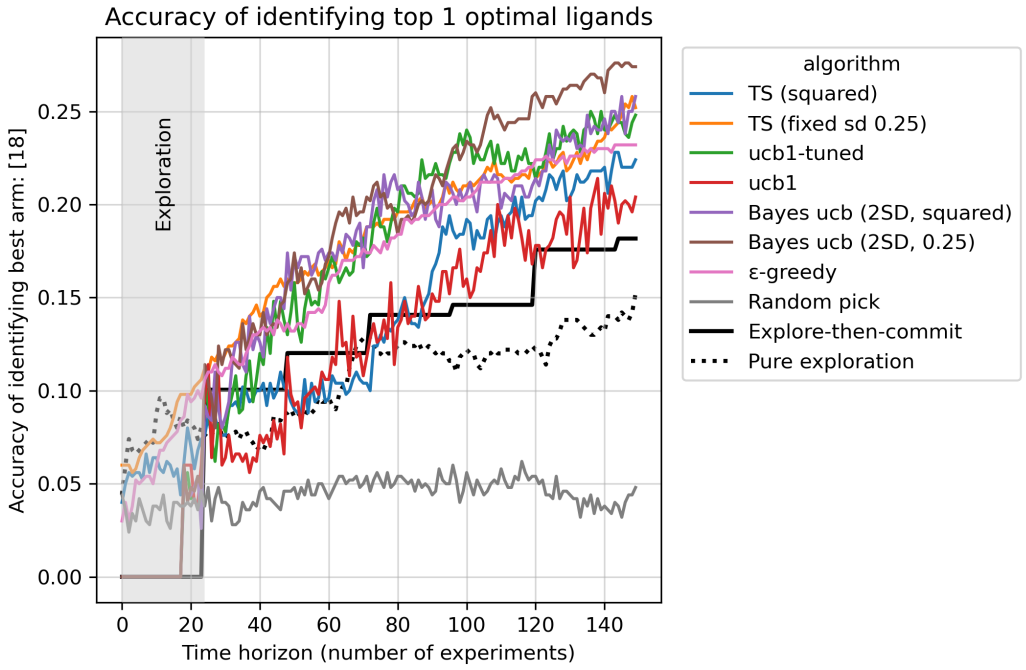


Fig. 112 Model substrate optimization results using a 90% yield cutoff.

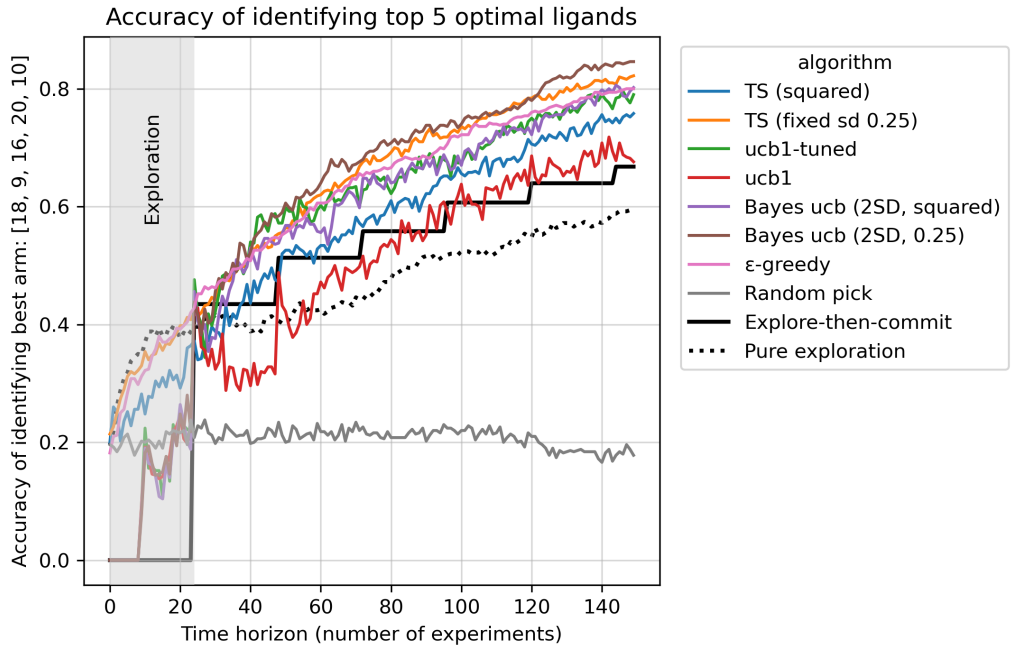


Fig. 113 Top-5 accuracy of identifying optimal ligand in the imidazole C–H arylation reaction for various algorithms.

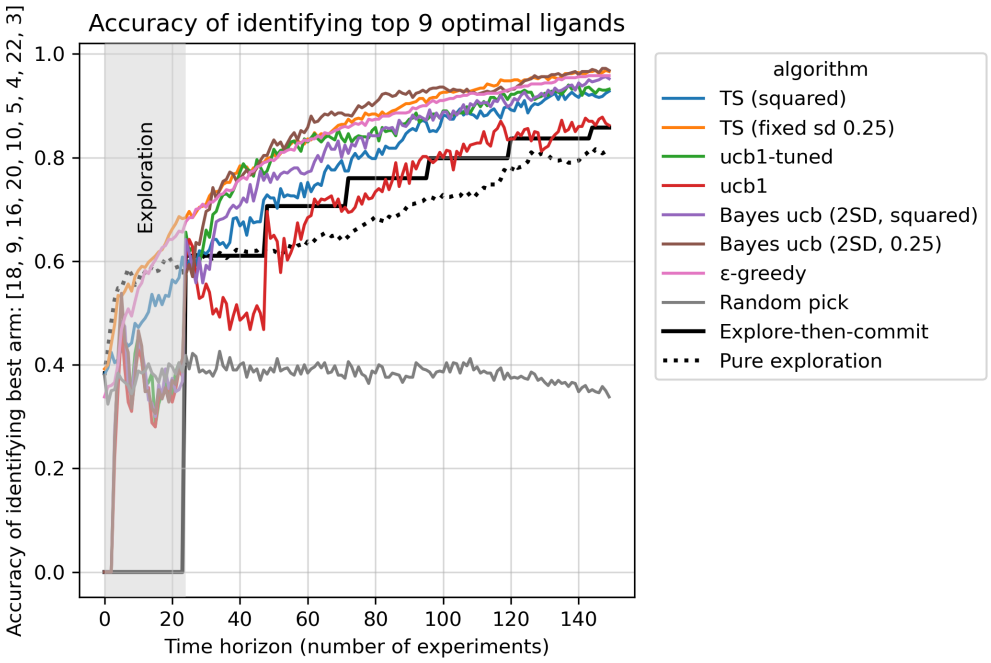


Fig. 114 Top-9 accuracy of identifying optimal ligand in the imidazole C–H arylation reaction for various algorithms.

Amide coupling dataset

For this reaction, we fixed the carboxylic acid core (indomethacin) and varied the aniline coupling partners. Starting from a commercial library of anilines, we generated dense vector embeddings for all molecules using mol2vec[71] and clustered them into ten groups using *k*-means clustering. One representative aniline was chosen from each cluster to constitute the aniline scope, which encompasses combinations of various heterocycles (quinolines, pyrazoles, pyridazines), electronically deactivating groups (nitriles, nitros, trifluoromethyls), sterically demanding *ortho*-substitutions, and potentially problematic functional groups (aryl chlorides/bromides, sulfonamides, esters). A series of eight amidation reagents, including aminiums, uroniums, (halo)phosphoniums, and phosphinic halides, were investigated as part of the condition scope, as well as four common organic bases and three solvents. The selected condition scope was selected manually, mostly based on occurrences in amide coupling literature and their relevance in process and medicinal chemistry. Overall, 10 aniline substrates, 8 activators, 4 bases and 3 solvents were selected as the reaction scope (**Fig. 115**).

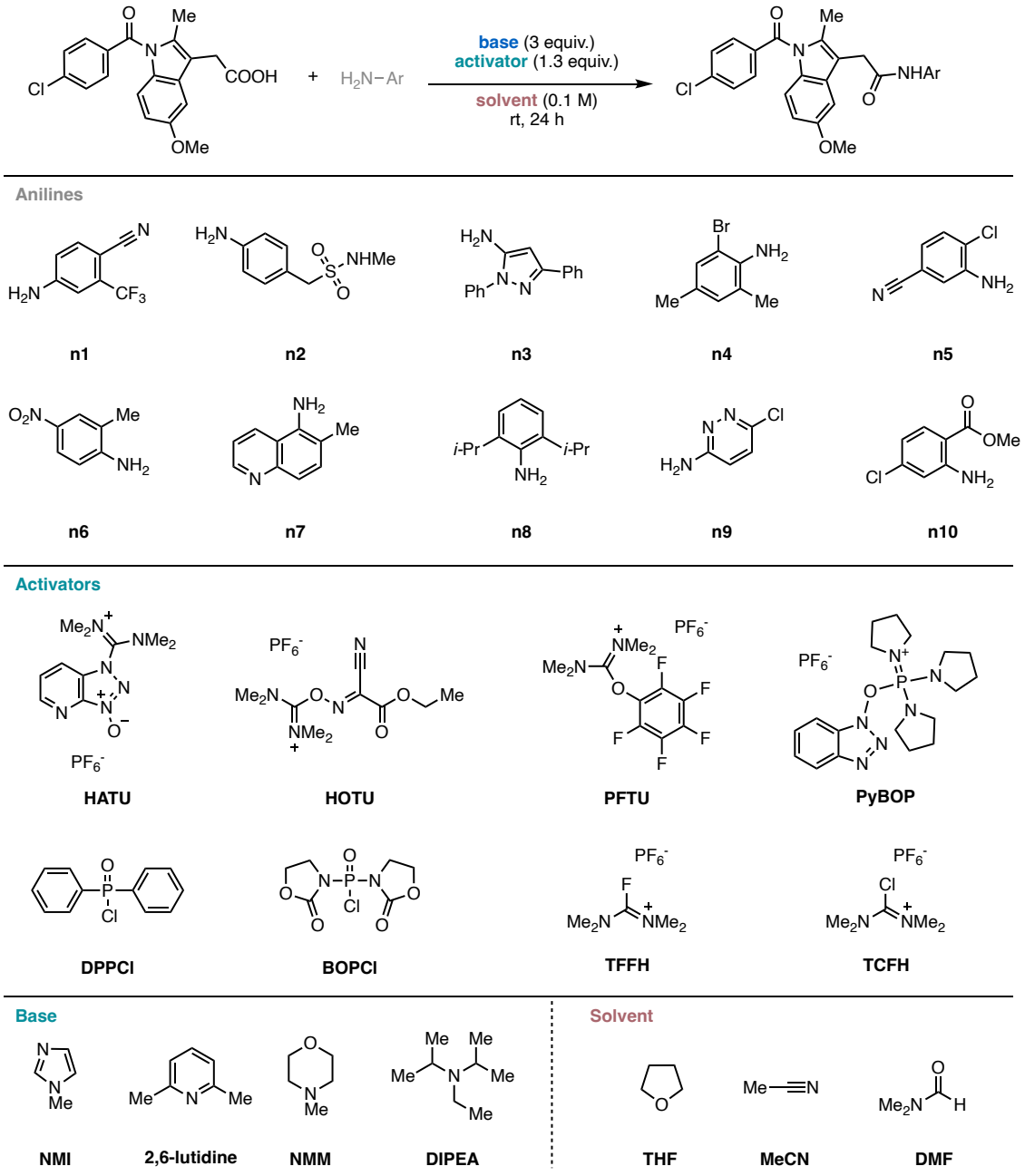


Fig. 115 Amidation dataset components: aniline substrates, activators, bases, solvents.

For the optimization run presented in Section 2.2, the first phase of the optimization was to identify effective activators. With optimization objective set to activators alone, algorithm selects the activator to test, and an experiment using that activator was suggested via random

sampling. For this stage, 8 rounds of experiments (5 experiments per round) were run in batch. The batch proposal was done by a random forest prediction model. For on-the-fly training with prediction models, aniline substrates were encoded with Morgan fingerprints using RDKit's default settings (radius=2, 2048 bits). For bandit algorithms, we used UCB1-Tuned for its generally high performance and lack of parameters that need to be tuned.

The experiment runs are shown in **Table 1**. After 8 rounds of experiments, activators were then ranked based on the number of times they were sampled by the algorithms first, and then further ranked based on their empirical average yields to break ties.

Round	activator	base	substrate	solvent	Exp. yield	Pred. yield
1	PFTU	2,6-Lutidine	n2	THF	0.38	1.0
1	TCFH	1-Methylimidazole	n4	MeCN	0.09	1.0
1	HATU	Diisopropylethylamine	n6	THF	0.04	1.0
1	PyBOP	1-Methylimidazole	n5	MeCN	0.04	1.0
1	BOP-Cl	Diisopropylethylamine	n3	THF	0.11	1.0
2	TFFH	2,6-Lutidine	n4	DMF	0.09	0.0881299999999999
2	DPPCl	2,6-Lutidine	n1	THF	0.15	0.1012220000000000
2	HOTU	Diisopropylethylamine	n6	THF	0.11	0.060454
2	PFTU	2,6-Lutidine	n6	THF	0.01	0.076482
2	BOP-Cl	N-methylmorpholine	n4	DMF	0.17	0.0825559999999999
3	DPPCl	Diisopropylethylamine	n8	MeCN	0.15	0.108348
3	HOTU	2,6-Lutidine	n3	MeCN	0.04	0.106451
3	TCFH	2,6-Lutidine	n9	DMF	0.26	0.115564
3	TFFH	2,6-Lutidine	n3	DMF	0.19	0.107861
3	HATU	2,6-Lutidine	n3	MeCN	0.06	0.089769
4	PyBOP	2,6-Lutidine	n5	MeCN	0.01	0.0517049999999999
4	PFTU	1-Methylimidazole	n3	THF	0.08	0.0862029999999998
4	TCFH	1-Methylimidazole	n9	MeCN	0.52	0.1603820000000000
4	DPPCl	N-methylmorpholine	n3	THF	0.50	0.121883
4	BOP-Cl	2,6-Lutidine	n9	DMF	0.74	0.2069110000000000
5	BOP-Cl	Diisopropylethylamine	n9	THF	0.37	0.5544630000000000
5	BOP-Cl	N-methylmorpholine	n2	MeCN	0.54	0.4037220000000000
5	TCFH	Diisopropylethylamine	n1	MeCN	0.17	0.10918
5	BOP-Cl	1-Methylimidazole	n10	MeCN	0.01	0.0833569999999999
5	DPPCl	N-methylmorpholine	n1	MeCN	0.47	0.337869
6	TFFH	Diisopropylethylamine	n3	DMF	0.47	0.1510380000000000
6	DPPCl	2,6-Lutidine	n3	DMF	0.20	0.17307
6	HOTU	Diisopropylethylamine	n6	DMF	0.05	0.0974269999999998
6	TCFH	1-Methylimidazole	n5	DMF	0.17	0.0714609999999999
6	DPPCl	Diisopropylethylamine	n1	DMF	0.38	0.211555
7	TFFH	N-methylmorpholine	n8	MeCN	0.07	0.3179590000000000
7	TFFH	Diisopropylethylamine	n3	MeCN	0.11	0.2506630000000000
7	HATU	2,6-Lutidine	n4	THF	0.02	0.0829859999999999
7	BOP-Cl	N-methylmorpholine	n3	DMF	0.07	0.3201540000000000
7	PFTU	2,6-Lutidine	n5	THF	0.01	0.0468079999999999
8	DPPCl	N-methylmorpholine	n5	THF	0.18	0.278249
8	PyBOP	2,6-Lutidine	n5	THF	0.00	0.018839
8	TCFH	Diisopropylethylamine	n5	MeCN	0.05	0.087767
8	DPPCl	1-Methylimidazole	n3	MeCN	0.52	0.2706000000000000
8	BOP-Cl	1-Methylimidazole	n7	MeCN	0.50	0.2918940000000000

Table 1 Proposed experiments for activator optimization rounds. Exp. yield: experimental yield. Pre. yield: predicted yield.

After the activator optimization rounds, we selected the top four (out of eight) activators, and optimized activator–base combinations. Optimization was re-initialized with 16 activator–base combinations (4 activators, 4 bases). Relevant existing results from the activator optimization rounds were used as initial data for the new optimization. The bandit algorithm (UCB1-Tuned) selects an activator–base combination to evaluate, and 16 rounds of experiments (5 experiments per round) were run in batch in the same way as the activator optimization round. The experiments run in this phase are shown in **Table 2**. The activator–base combinations were similarly ranked with the number of samples and empirical averages.

Round	activator	base	substrate	solvent	Exp. yield	Pred. yield
9	PFTU	2,6-Lutidine	n2	THF	0.38	1.0
9	TCFH	1-Methylimidazole	n4	MeCN	0.09	1.0
9	HATU	Diisopropylethylamine	n6	THF	0.04	1.0
9	PyBOP	1-Methylimidazole	n5	MeCN	0.04	1.0
9	BOP-Cl	Diisopropylethylamine	n3	THF	0.11	1.0
10	TFFH	2,6-Lutidine	n4	DMF	0.09	0.0881299999999999
10	DPPCl	2,6-Lutidine	n1	THF	0.15	0.1012220000000000
10	HOTU	Diisopropylethylamine	n6	THF	0.11	0.060454
10	PFTU	2,6-Lutidine	n6	THF	0.01	0.076482
10	BOP-Cl	N-methylmorpholine	n4	DMF	0.17	0.0825559999999999
11	DPPCl	Diisopropylethylamine	n8	MeCN	0.15	0.108348
11	HOTU	2,6-Lutidine	n3	MeCN	0.04	0.106451
11	TCFH	2,6-Lutidine	n9	DMF	0.26	0.115564
11	TFFH	2,6-Lutidine	n3	DMF	0.19	0.107861
11	HATU	2,6-Lutidine	n3	MeCN	0.06	0.089769
12	PyBOP	2,6-Lutidine	n5	MeCN	0.01	0.0517049999999999
12	PFTU	1-Methylimidazole	n3	THF	0.08	0.0862029999999998
12	TCFH	1-Methylimidazole	n9	MeCN	0.52	0.1603820000000000
12	DPPCl	N-methylmorpholine	n3	THF	0.50	0.121883
12	BOP-Cl	2,6-Lutidine	n9	DMF	0.74	0.2069110000000000
13	BOP-Cl	Diisopropylethylamine	n9	THF	0.37	0.5544630000000000
13	BOP-Cl	N-methylmorpholine	n2	MeCN	0.54	0.4037220000000000
13	TCFH	Diisopropylethylamine	n1	MeCN	0.17	0.10918
13	BOP-Cl	1-Methylimidazole	n10	MeCN	0.01	0.0833569999999999
13	DPPCl	N-methylmorpholine	n1	MeCN	0.47	0.337869
14	TFFH	Diisopropylethylamine	n3	DMF	0.47	0.1510380000000000
14	DPPCl	2,6-Lutidine	n3	DMF	0.20	0.17307
14	HOTU	Diisopropylethylamine	n6	DMF	0.05	0.0974269999999998
14	TCFH	1-Methylimidazole	n5	DMF	0.17	0.0714609999999999
14	DPPCl	Diisopropylethylamine	n1	DMF	0.38	0.211555
15	TFFH	N-methylmorpholine	n8	MeCN	0.07	0.3179590000000000
15	TFFH	Diisopropylethylamine	n3	MeCN	0.11	0.2506630000000000
15	HATU	2,6-Lutidine	n4	THF	0.02	0.0829859999999999
15	BOP-Cl	N-methylmorpholine	n3	DMF	0.07	0.3201540000000000
15	PFTU	2,6-Lutidine	n5	THF	0.01	0.0468079999999999
16	DPPCl	N-methylmorpholine	n5	THF	0.18	0.278249

16	PyBOP	2,6-Lutidine	n5	THF	0.00	0.018839
16	TCFH	Diisopropylethylamine	n5	MeCN	0.05	0.087767
16	DPPCl	1-Methylimidazole	n3	MeCN	0.52	0.2706000000000000
16	BOP-Cl	1-Methylimidazole	n7	MeCN	0.50	0.2918940000000000
17	TFFH	1-Methylimidazole	n1	THF	0.23	0.197645
17	TCFH	N-methylmorpholine	n9	DMF	0.06	0.3841750000000000
17	BOP-Cl	2,6-Lutidine	n7	MeCN	0.12	0.4059960000000000
17	DPPCl	1-Methylimidazole	n3	THF	0.23	0.4057010000000000
17	TCFH	N-methylmorpholine	n5	MeCN	0.11	0.07122
18	TCFH	2,6-Lutidine	n2	MeCN	0.35	0.3356930000000000
18	TFFH	1-Methylimidazole	n4	DMF	0.05	0.1522020000000000
18	BOP-Cl	2,6-Lutidine	n8	MeCN	0.06	0.178906
18	DPPCl	1-Methylimidazole	n2	MeCN	0.49	0.4196860000000000
18	TFFH	N-methylmorpholine	n5	THF	0.05	0.117228
19	TCFH	2,6-Lutidine	n7	MeCN	0.36	0.2134070000000000
19	TFFH	Diisopropylethylamine	n8	MeCN	0.13	0.1359070000000000
19	DPPCl	1-Methylimidazole	n4	THF	0.07	0.2040260000000000
19	DPPCl	Diisopropylethylamine	n10	MeCN	0.04	0.2281860000000000
19	DPPCl	N-methylmorpholine	n2	MeCN	0.61	0.459644
20	BOP-Cl	1-Methylimidazole	n2	DMF	0.48	0.464601
20	BOP-Cl	Diisopropylethylamine	n6	DMF	0.04	0.0726219999999999
20	DPPCl	N-methylmorpholine	n3	MeCN	0.51	0.4077180000000000
20	BOP-Cl	1-Methylimidazole	n8	DMF	0.30	0.1469760000000000
20	TCFH	2,6-Lutidine	n1	THF	0.36	0.1742770000000000
21	DPPCl	N-methylmorpholine	n9	DMF	0.71	0.3761540000000000
21	BOP-Cl	2,6-Lutidine	n7	THF	0.04	0.2616620000000000
21	DPPCl	2,6-Lutidine	n10	MeCN	0.11	0.0704579999999999
21	DPPCl	N-methylmorpholine	n5	MeCN	0.12	0.1703900000000000
21	TFFH	1-Methylimidazole	n10	DMF	0.02	0.0714489999999999
22	TFFH	2,6-Lutidine	n9	DMF	0.22	0.4203900000000000
22	TCFH	1-Methylimidazole	n7	THF	0.18	0.3021630000000000
22	BOP-Cl	N-methylmorpholine	n6	THF	0.01	0.0960179999999999
22	DPPCl	1-Methylimidazole	n5	DMF	0.12	0.121413
22	TCFH	2,6-Lutidine	n4	THF	0.05	0.104383
23	BOP-Cl	1-Methylimidazole	n9	MeCN	0.45	0.5089850000000000
23	TCFH	Diisopropylethylamine	n2	MeCN	0.18	0.3909040000000000
23	DPPCl	N-methylmorpholine	n10	THF	0.01	0.1178700000000000
23	TFFH	Diisopropylethylamine	n10	THF	0.17	0.0415429999999999
23	BOP-Cl	1-Methylimidazole	n4	DMF	0.06	0.1205600000000000
24	TCFH	N-methylmorpholine	n5	THF	0.27	0.108199
24	TFFH	N-methylmorpholine	n7	DMF	0.10	0.25712
24	DPPCl	Diisopropylethylamine	n3	DMF	0.26	0.3659460000000000
24	BOP-Cl	Diisopropylethylamine	n10	THF	0.01	0.0603559999999999
24	TFFH	2,6-Lutidine	n5	THF	0.01	0.0452339999999999

Table 2 Proposed experiments for activator–base optimization rounds. Exp. yield: experimental yield. Pre. yield: predicted yield.

All reactions in the scope were run with HTE to allow for the direct comparison between optimization ranking and true rankings. Heatmap of results grouped by aniline substrates and conditions are shown in **Fig. 116** and **Fig. 117**. The experimental details of the product synthesis and HTE reactions can be found in Section 2.5.2.

With HTE data for the entire reaction scope, average yields for activators, bases, and solvents (**Fig. 118**) and different metrics were used to analyze activator performance (**Fig. 119**) and activator–base performance (**Fig. 120**, **Fig. 121**). The true rankings for activators and activator–base presented in the manuscript were obtained based on average yields.

We also simulated the full dataset with some optimization algorithms. For activators, top-1 accuracy of identifying DPPCl (**Fig. 122**), top-3 accuracy of identifying DPPCl, BOP-Cl, TCFH (**Fig. 123**) are shown. For activators–bases, top-2 accuracy of identifying DPPCl–NMM, DPPCl–DIPEA is shown (**Fig. 124**).

A more detailed reactivity comparison for DPPCl– NMM and DPPCl–DIPEA, using HATU–DIPEA and TCFH–NMI as baseline, divided by solvents and substrates, is shown in Fig. **Fig. 125**.

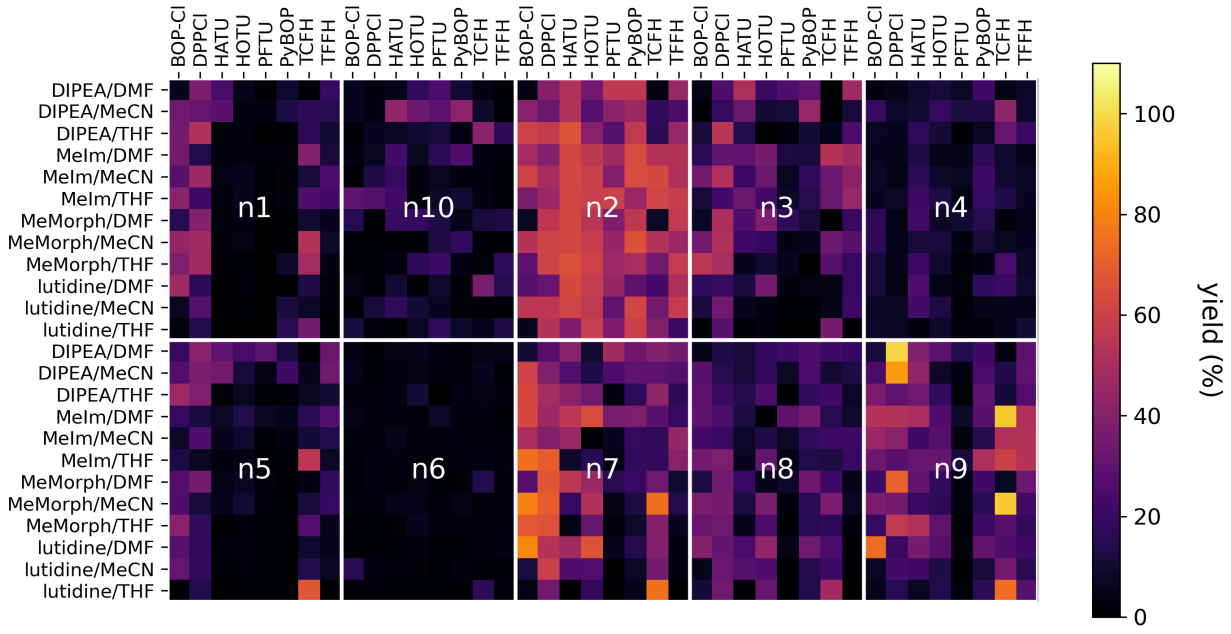


Fig. 116 HTE results for amide coupling reaction (base–solvent, activator).

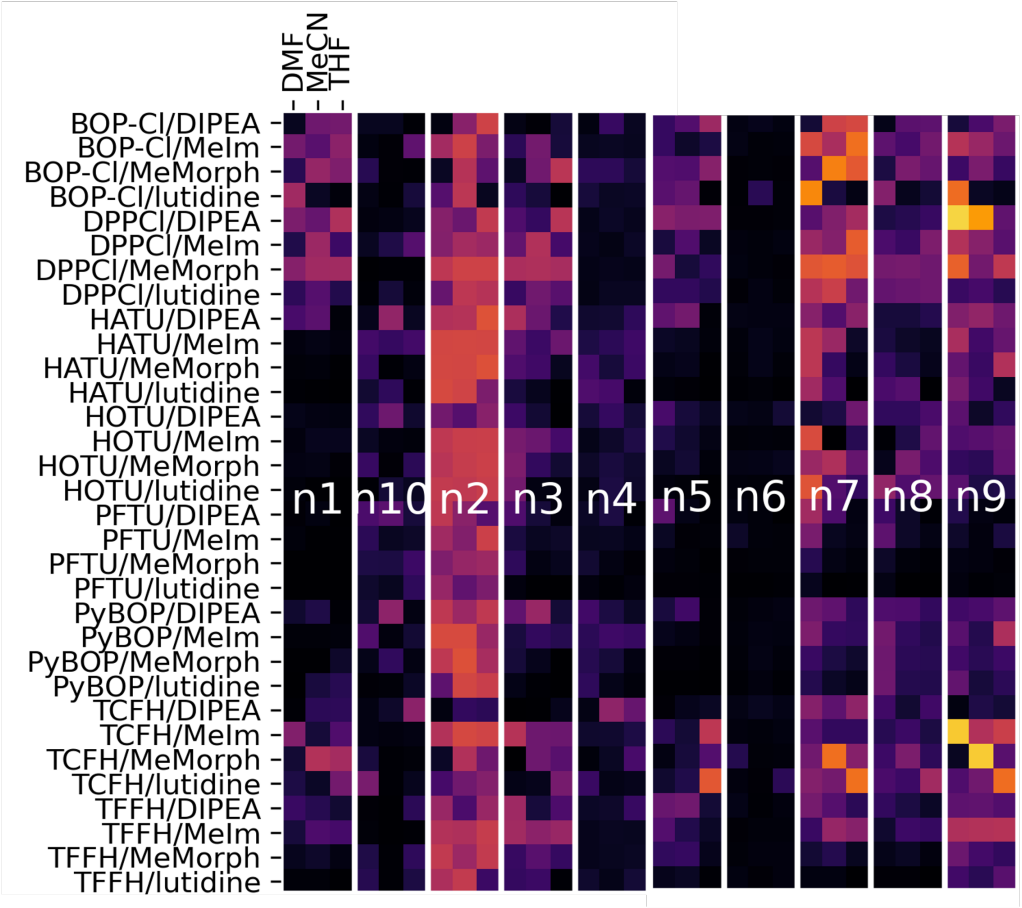


Fig. 117 HTE results for amide coupling reaction (activator–base, solvent).

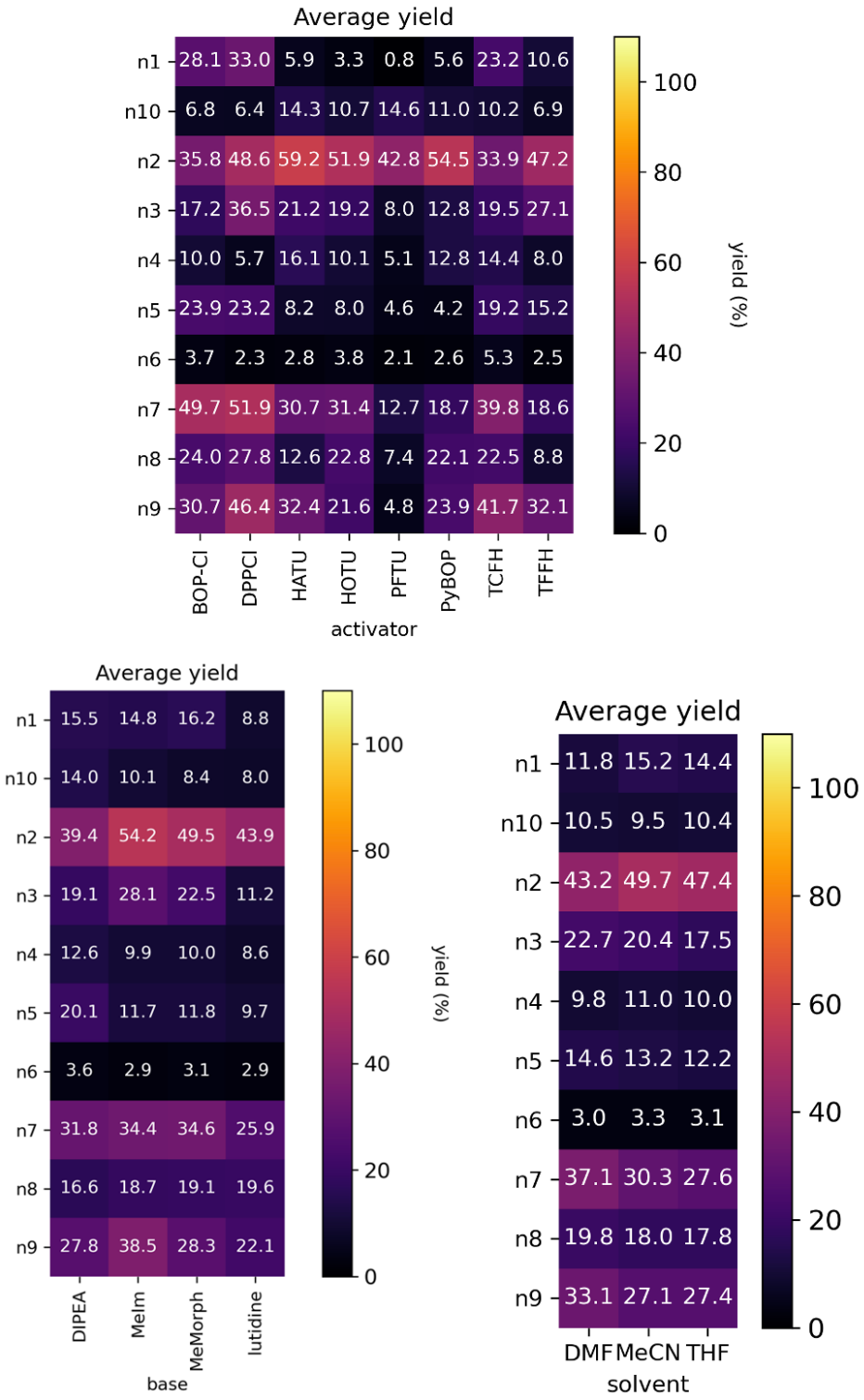


Fig. 118 Average yields of activators, bases, and solvents for each substrate.

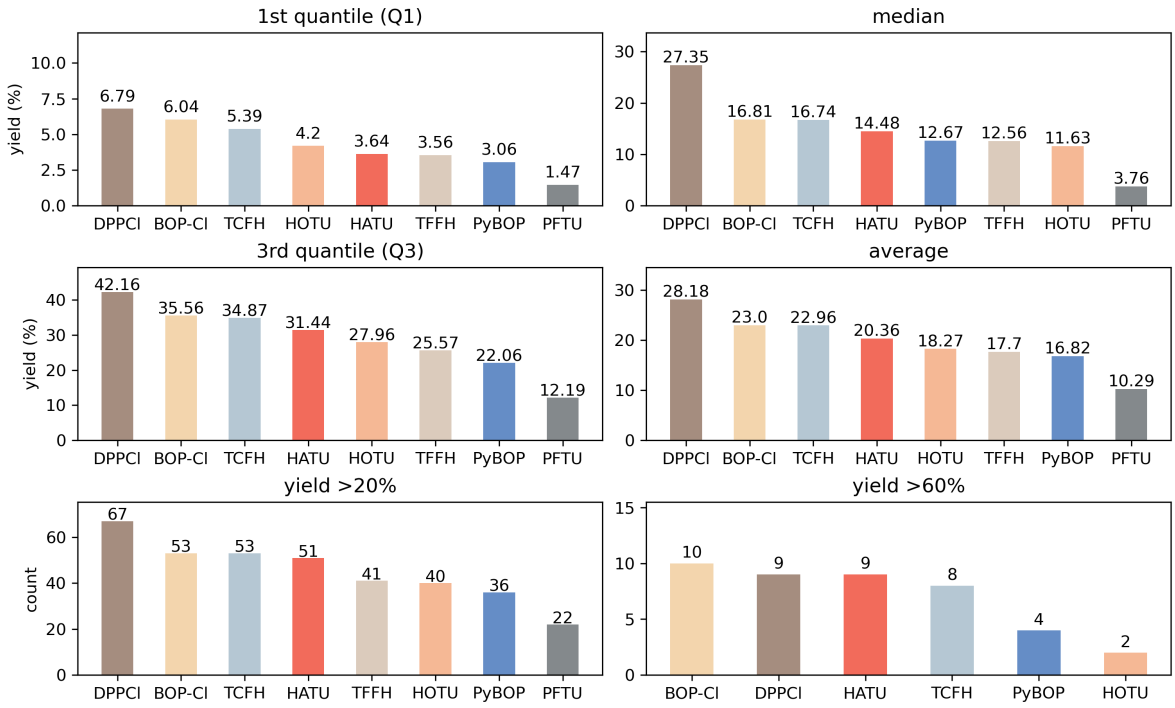


Fig. 119 Different metrics to evaluate activator performance in the amide coupling reaction.

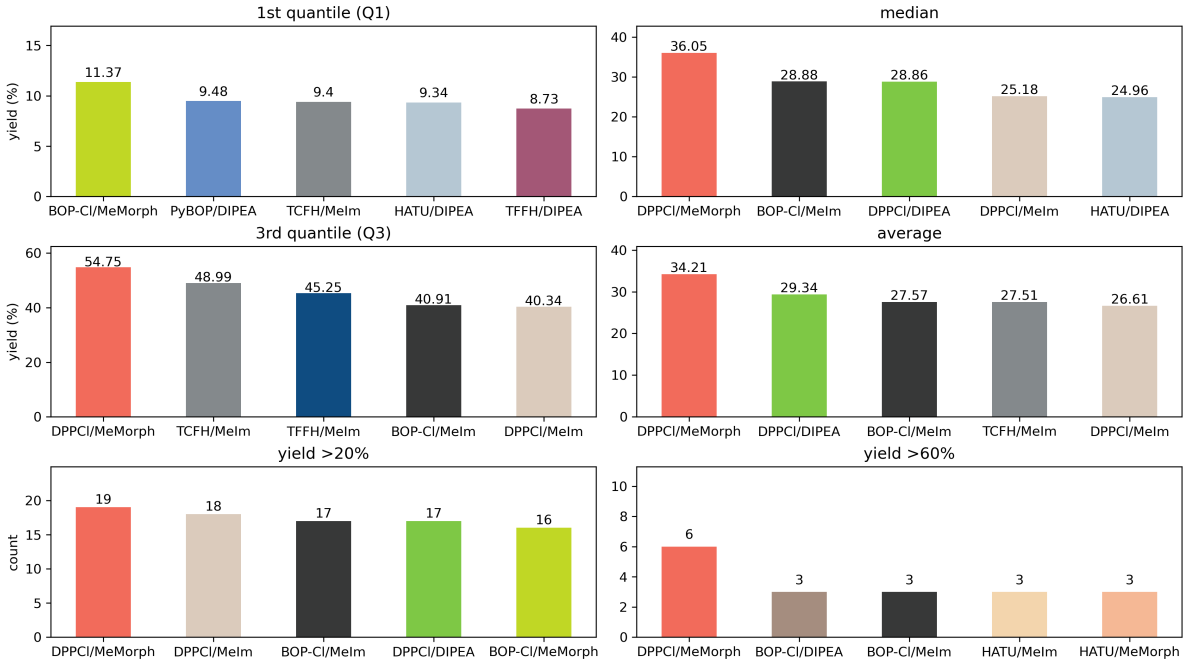


Fig. 120 Different metrics to evaluate activator–base performance in the amide coupling reaction (top 5 plotted).

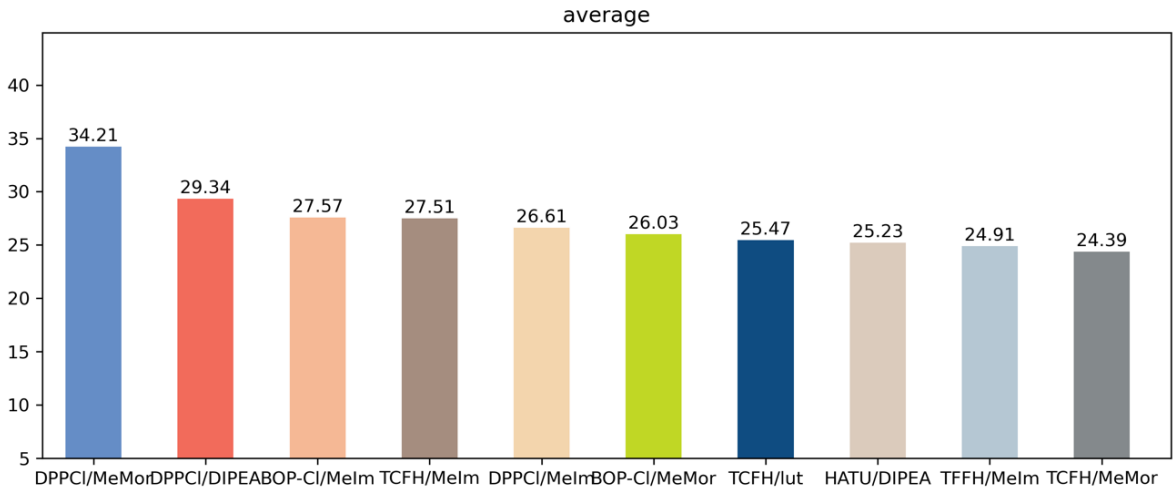


Fig. 121 Top 10 average yields for activator–base in the amide coupling reaction.

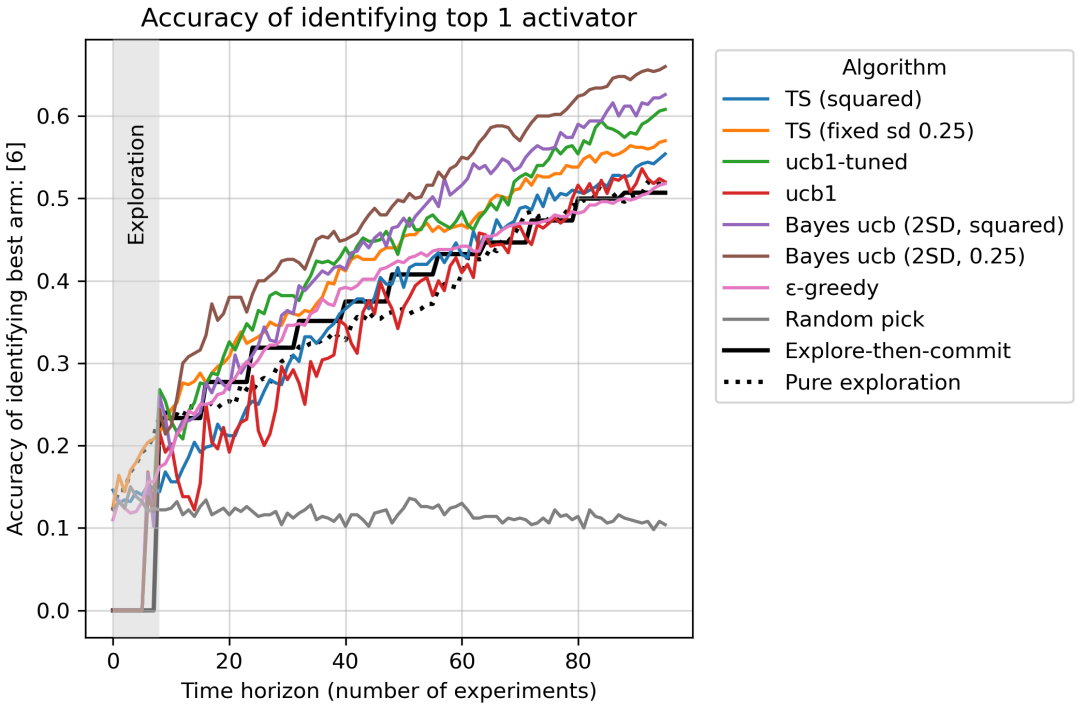


Fig. 122 Top-1 accuracy of identifying optimal activator (DPPCl) in the amide coupling reaction for various algorithms.

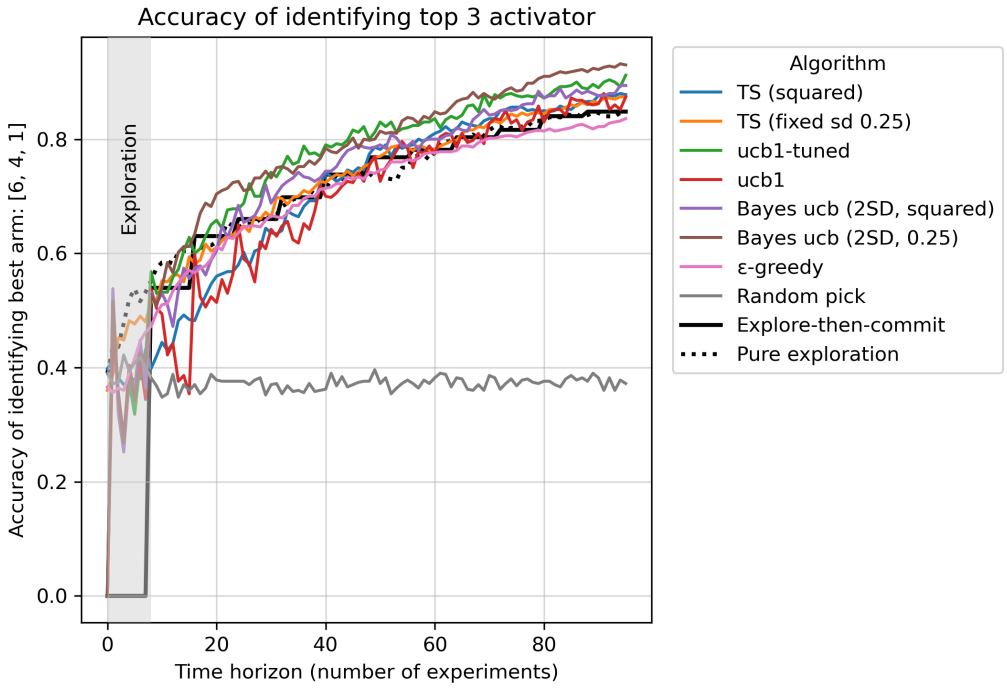


Fig. 123 Top-3 accuracy of identifying optimal activator (DPPCl, BOP-Cl, TCFH) in the amide coupling reaction for various algorithms.

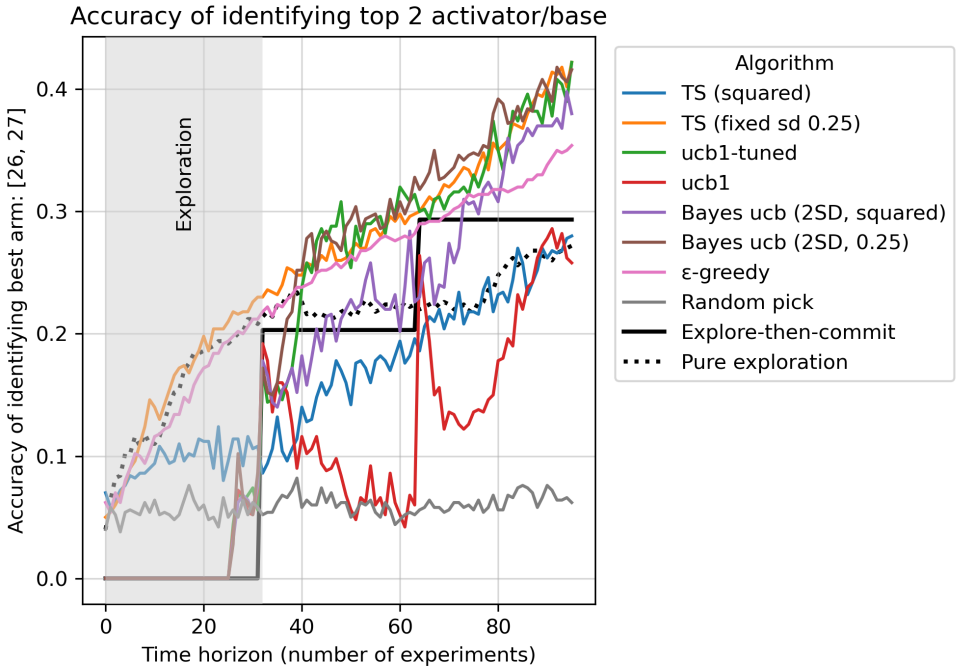


Fig. 124 Top-2 accuracy of identifying optimal activator–base (DPPCl–NMM, DPPCl–DIPEA) in the amide coupling reaction for various algorithms.

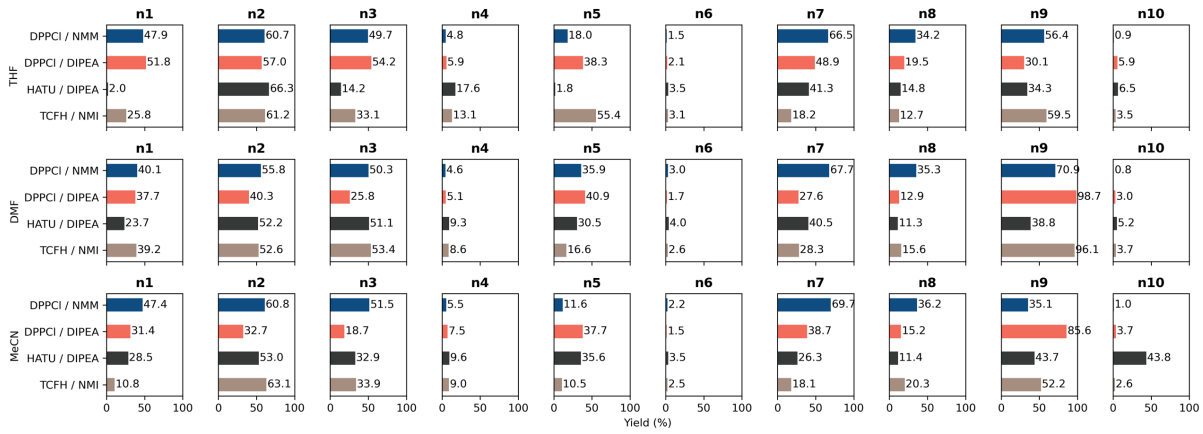


Fig. 125 Yields grouped by solvents for identified conditions of DPPCl–NMM and DPPCl–DIPEA when applied to all ten aniline nucleophiles. HATU–DIPEA and TCFH–NMI were used as baseline comparisons.

2.5 Experimental section

2.5.1 C-H arylation dataset experimentation details

High-Throughput Experimentation

Bulk Ligand Plate Preparation. In a glovebox, 8 mL vials containing 25 μmol ligand were dissolved in 1,2-dichloroethane (2.5 mL) and stirred for 5 min. A 100 μL aliquot of each of the resultant ligand solutions was dispensed to the desired location in the 96 well plate. The solvent was removed *in vacuo* using a Genevac centrifugal evaporator inside the glovebox. Ligand plates were sealed and stored in the glovebox until time of use.

Base Plate Preparation. Potassium Pivalate (6.3 mg, 0.045 mmol)/well was dispensed to each 96 well plate using Unchained Labs Powder Protégé. The vials containing potassium pivalate were stored open in a glovebox for no less than three days prior to use to remove trace amounts of water and then sealed and stored in the glovebox until time of use.

Reaction Execution. In a glovebox, 0.5 M solutions of aryl bromide (0.58 mmol) in *N*,*N*-dimethylacetamide were prepared and dispensed to the ligand vials (40 μL, 0.02 mmol) *via* electronic multi-step pipettor. A solution of $[Pd(allyl)Cl]_2$ (50.0 mg, 0.136 mmol) in *N*,*N*-dimethylacetamide (3.04 mL) was prepared (0.045 M) and dispensed to the vials containing a solution of aryl bromide and ligand (10 μL, 0.45 μmol). The resultant reaction mixtures were sealed and stirred on a shaker block in the glovebox for no less than 30 min. Solutions of imidazole nucleophile (1.15 mmol, 0.27 M) containing 4,4'-di-*tert*-butylbiphenyl (15.3 mg, 0.057 mmol) in *N*,*N*-dimethylacetamide (17.3 mL) were prepared and dispensed to the reaction mixture vials (150 μL, 0.04 mmol). The reaction mixtures stirred for 2 min in the glovebox and 150 μL from each well was transferred to the base plate using a multichannel pipettor. The reactions were then sealed and stirred at the desired temperature 120 °C for 24 h in the glovebox and subsequently cooled to 23 °C. The plate was removed from the glovebox, opened, and diluted to a 900 μL total volume with *N*,*N*-dimethylacetamide. The plate was stirred for 5 min and a 75 μL sample was taken and filtered into an HPLC analysis plate. The filter was rinsed with 400 μL acetonitrile/water (4:1) solution and analyzed by UPLC-MS.

Calibration Curve: A solution of the product marker (0.9 mmol) in 5.40 mL *N*,*N*-dimethylacetamide was prepared. A serial dilution from this solution was performed into vials containing 4,4'-di-*tert*-butylbiphenyl (0.91 mg, 3.42 μmol) to generate solutions that contain 10%, 20%, 40%, 60%, 80%, and 100% of the original product marker *vs* the consistent amount of 4,4'-di-*tert*-butylbiphenyl internal standard. A 75 μL sample of each was taken and filtered into an UPLC-MS analysis plate. The filter was rinsed with 400 μL acetonitrile/water (4:1) solution and analyzed by UPLC-MS.

UPLC-MS Method:

Solvent A: Water with 5% acetonitrile and 0.05% TFA

Solvent B: acetonitrile with 5% water and 0.05% TFA

Gradient: 95% A/B to 5% A/B over 2.5 min, hold 1.0 min at 95% B

Stop Time: 3.5 min

Flow Rate: 0.8 mL/min

Wavelength1: 220 nm

Wavelength2: 254 nm

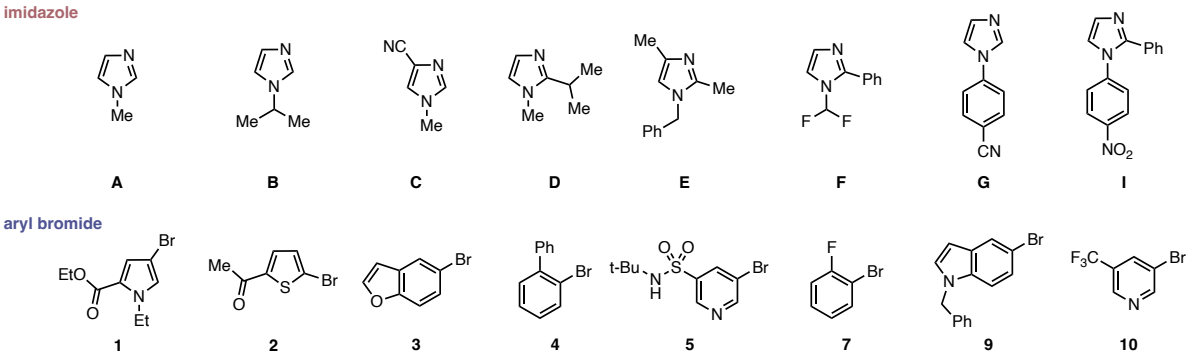
Column: Agilent Poroshell C18 2.7 um 2.1x50mm

Oven Temperature: 40 °C

Dataset Analysis

Substrate scope. 8 imidazoles and 8 aryl bromides were selected as the substrate scope, generating 64 cross-coupled products labeled as <imidazole>-<aryl bromide> (e.g., A3, G5).

Note: The original substrate scope was designed with 10 imidazoles and 10 aryl bromides, but imidazole **H, J** and aryl bromide **6, 8** were later removed from the scope. The original labels from the design were kept to ensure consistency internally.

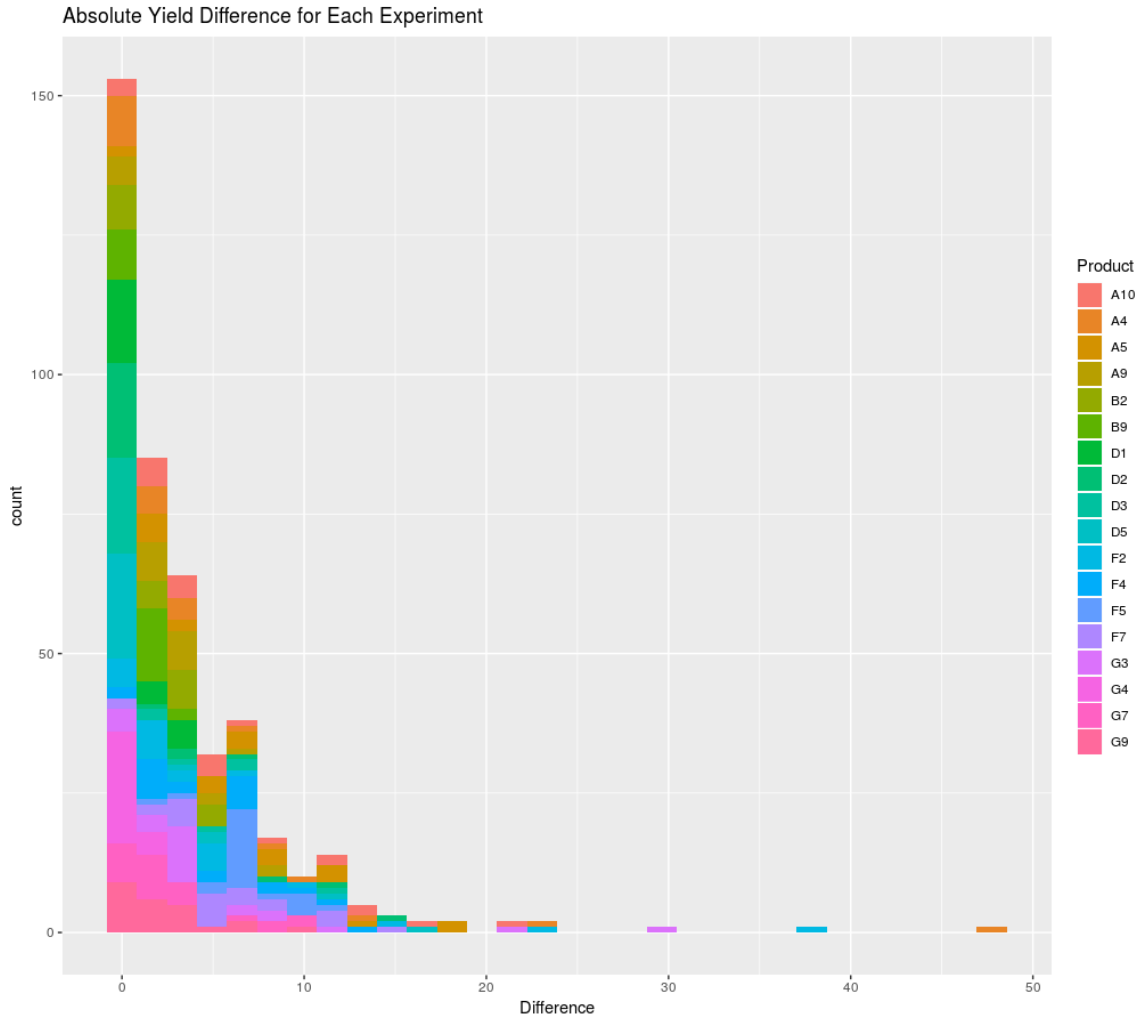


Data Processing. Data processing was completed in R 3.4.4 software installed on Ubuntu 16.04 using the tidyverse 1.2.1 package. The experimental results (Experimental_Data) and

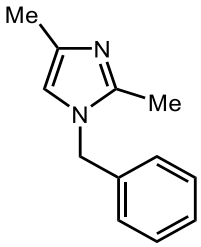
calibration results (Calibration_Data) from the UPLC-MS instrument were imported as .csv files and merged with the experimental design. The relative yields (RelYield_PDT) for the experimental and calibration data were calculated by dividing the area percent of the desired product by the area percent of the internal standard for each entry. The yields were calculated by fitting a linear model on the Calibration_Data and applying to the Experimental_Data using the function below.

```
Model <- function(Calibration_Data, Experimental_Data) {
Model <- lm(Yield ~ RelYield_PDT, Calibration_Data)
Model_coeffs <- coefficients(Model)
Experimental_Data $Yield <- Model_coeffs[1] + Model_coeffs[2]* Experimental_Data $RelYield_PDT
Experimental_Data $Yield <- round(Experimental_Data $Yield, digits = 2)
Experimental_Data
}
```

Duplicated Studies. A total of 18 of 64 possible studies were conducted in duplicate: A4, A5, A9, A10, B2, B9, D1, D2, D3, D5, F2, F4, F5, F7, G3, G4, G7 and G9. Analysis of the data found that the average absolute difference between duplicated runs for the 432-experiment set was 3.6% yield with a standard deviated of 5.0% yield. The yield difference was lower than 10% yield (two standard deviations) for 92% of the experiments.



Preparation of substrates (imidazoles E, F, I)



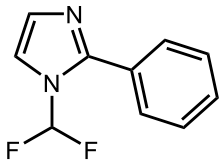
1-Benzyl-2,4-dimethyl-1H-imidazole (**E**).

To a nitrogen-flushed 500 mL round-bottom flask was added 2,4-dimethyl-1H-imidazole (14.5 g, 145 mmol, 1.1 equiv.), potassium carbonate (38.2 g, 276 mmol, 2.1 equiv.), and

acetonitrile (225 mL, 10 mL/g). The mixture was cooled to 0 °C and treated with a solution of benzyl bromide (22.5 g, 132 mmol, 1.0 equiv.) in acetonitrile (22.5 mL, 1 mL/g) over 30 minutes. The reaction mixture was aged at 0 °C for 3 h and warmed to 23 °C and aged for an additional 15 h. The solvent was removed *in vacuo* and was treated with ethyl acetate (225 mL, 10 mL/g). The solution was washed with water (225 mL, 10 mL/g) followed by 10% aqueous sodium chloride solution (225 mL, 10 mL/g). The isolated organic phase was dried over magnesium sulfate, filtered, and concentrated. Purification of the crude by silica gel chromatography (220 g ISCO RediSep-Rf Gold column; 1% to 10% methanol/dichloromethane gradient) afforded the desired product **E** (18.0 g, 97 mmol) in 73% yield as yellow oil.

^{1}H NMR (400 MHz, $CDCl_3$): δ 7.58 - 7.43 (m, 3H), 7.26 (d, *J*=6.8 Hz, 2H), 6.74 (s, 1H), 5.17 (s, 2H), 2.50 (s, 3H), 2.38 (d, *J*=0.9 Hz, 3H)

^{13}C NMR (101 MHz, $CDCl_3$): δ 144.0, 136.6, 136.1, 128.8, 126.5, 125.6, 116.0, 49.4, 13.5, 12.9



1-(Difluoromethyl)-2-phenyl-1H-imidazole (F).

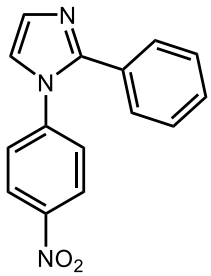
To a 250-mL round bottomed flask was added 2-phenylimidazole (3.1 g, 21 mmol, 1.0 equiv.) and potassium fluoride (2.44 g, 42 mmol, 2.0 equiv.). The flask was transferred to a glovebox under inert atmosphere and treated with acetonitrile (125 mL, 40 mL/g) followed by 1-[[bromo(difluoro)methyl]-ethoxy-phosphoryl]oxyethane (5.6 g, 21 mmol, 1.0 equiv.). The reaction was aged 18 h and the solvent was removed. The crude was redissolved in ethyl acetate

(125 mL, 40 mL/g) and treated with 1 M aqueous hydrochloric acid solution (200 mL, 64 mL/g) and mixed vigorously for five minutes. The organic phase was removed, and the aqueous phase was washed with ethyl acetate (125 mL, 40 mL/g). The isolated aqueous phase was adjusted to pH = 8 with 6 M aqueous potassium hydroxide, treated with ethyl acetate (250 mL, 81 mL/g) and mixed vigorously for five minutes. The isolated organic phase was dried over magnesium sulfate, filtered, and concentrated. Purification by silica gel chromatography (80 g ISCO RediSep-Rf Gold column; 5% to 25% ethyl acetate/heptane gradient) afforded the desired product **F** (2.1 g, 10.0 mmol) in 48% yield as yellow oil.

^{1}H NMR (400 MHz, $CDCl_3$): δ 7.64 - 7.55 (m, 2H), 7.53 - 7.48 (m, 3H), 7.40 (s, 1H), 7.27 - 7.20 (m, 1H), 7.07 (t, *J*=6.0 Hz, 1H)

^{13}C NMR (101 MHz, $CDCl_3$): δ 147.4, 130.3, 130.0, 129.1, 128.9, 115.5, 108.6 (t, *J*=249.8 Hz, 1C)

^{19}F NMR (276 MHz, $CDCl_3$): δ -90.50 (d, *J*=41.4 Hz)



1-(4-Nitrophenyl)-2-phenyl-1H-imidazole (**I**).

To a nitrogen flushed reaction vessel containing potassium carbonate (25.9 g, 187.5 mmol, 2.5 equiv.) and (2-phenylimidazole (10.8 g, 75 mmol, 1.0 equiv.) was added *N,N*-dimethylformamide (108 mL, 10 mL/g) followed by 4-nitrofluorobenzene (11.6 g, 82.5 mmol, 1.1

equiv.). The reaction mixture was heated to 100 °C and aged for 3 h then cooled to 23 °C. Ethyl acetate (216 mL, 20 mL/g) was added and the organic solution was rinsed 3 ×10 wt% aqueous sodium chloride solution (216 mL, 20 mL/g) and the isolated organic phase was concentrated to red oil. Purification by silica gel chromatography (220 g ISCO RediSep-Rf Gold column; 40% to 70% ethyl acetate/heptane gradient) afforded the desired product **I** (14.5 g, 55 mmol) in 73% yield as red solid.

^{1}H NMR (400 MHz, DMSO-d6): δ 8.36 - 8.23 (m, 2H), 7.65 (d, *J*=1.2 Hz, 1H), 7.60 - 7.50 (m, 2H), 7.43 - 7.29 (m, 5H), 7.26 (d, *J*=1.2 Hz, 1H)

^{13}C NMR (101 MHz, DMSO-d6): δ 146.4, 145.9, 143.2, 129.9, 129.4, 128.7, 128.5, 128.5, 126.8, 124.9, 123.4

Authentic product synthesis and characterization

General Procedure A:

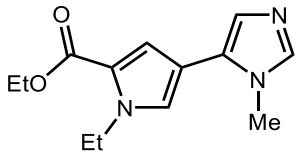
Aryl halide (5 mmol, 1 equiv.) was dispensed to 20 mL vial in a glovebox and treated with a solution of PCy3•HBF4 (92 mg, 0.25 mmol, 5 mol%) in DMA (5 mL). The mixture was stirred for 2 min and treated with a solution of $[(Allyl)PdCl]_2$ (41 mg, 0.11 mmol, 2.25 mol%) in DMA (5 mL) and stirred for no less than 30 min. The reaction mixture was then treated with a solution of the imidazole (10 mmol, 2 equiv.) in DMA (5 mL). The reaction was stirred for 2 minutes and was then poured into a 40 mL vial containing KOPiv (2.15 g, 15 mmol, 3 equiv.). The initial ArBr vial was rinsed with DMA (5 mL) into the reaction vial. The reaction was capped and heated to 140 °C external temperature for 24 h. Upon reaction completion as determined by UPLC-MS, the reaction was cooled to room temperature and filtered through a short plug of celite (this filtration

can be really slow). The filtrate was then concentrated using a Genevac evaporator. The crude material was purified by reversed phase chromatography.

General Procedure B:

Aryl halide (5 mmol, 1 equiv.) was dispensed to 20 mL vial in a glovebox and treated with a solution of PCy3•HBF4 (92 mg, 0.25 mmol, 5 mol%) in DMA (5 mL). The mixture was stirred for 2 min and treated with a solution of $[(Allyl)PdCl]_2$ (41 mg, 0.11 mmol, 2.25 mol%) in DMA (5 mL) and stirred for no less than 30 minutes. The reaction mixture was then treated with a solution of the imidazole (10 mmol, 2 equiv.) in DMA (5 mL). The reaction stirred for 2 minutes and was then poured into a 40 mL vial containing KOPiv (2.15 g, 15 mmol, 3 equiv.). The initial ArBr vial was rinsed with DMA (5 mL) into the reaction vial. The reaction was capped and heated to 140 °C external temperature for 24 h. Upon reaction completion as determined by TLC and/or LC-MS, the crude reaction mixture was cooled to room temperature and subsequently dissolved in 100 mL of H_2O. The solution was then extracted with Et_2O (3 × 50 mL) and concentrated *in vacuo*. The crude residue was then purified by silica gel column chromatography.

A-Series



Ethyl 1-ethyl-4-(1-methyl-1*H*-imidazol-5-yl)-1*H*-pyrrole-2-carboxylate (**A1**).

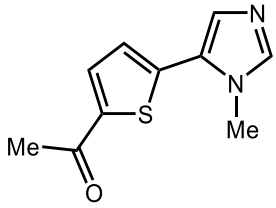
Prepared according to the general procedure B. The title compound was isolated via flash chromatography (100% dichloromethane with 5% triethylamine additive) as an orange oil (88.1 mg, 356 μmol, 7% yield).

^{1}H NMR (500 MHz, $CDCl_3$): δ 7.44 (s, 1H); 7.01 (s, 1H); 7.00 (d, *J* = 2.0 Hz, 1H); 6.94 (d, *J* = 2.1 Hz, 1H); 4.38 (q, *J* = 7.2 Hz, 2H); 4.30 (q, *J* = 7.1 Hz, 2H); 3.67 (s, 3H); 1.43 (t, *J* = 7.0 Hz, 3H); 1.36 (t, *J* = 7.1 Hz, 3H)

^{13}C NMR (126 MHz, $CDCl_3$): δ 160.9; 138.3; 127.6; 127.1; 126.0; 122.6; 116.8; 112.2; 60.2; 44.5; 32.6; 17.1; 14.6

HRMS: (EI+) calculated for $[C_{13}H_{17}N_3O_2+H]^+$ 248.1392, found: 248.1394.

FTIR (ATR, cm^{-1}): 3376.5; 3110.6; 2980.4; 2932.9; 1698.5; 1528.5; 1475.2; 1445.3; 1377.8; 1283.1; 1255.9; 1234.6; 1198.1; 1113.9; 1096.7; 1075.8; 802.7; 652.9



1-(5-(1-Methyl-1H-imidazol-5-yl)thiophen-2-yl)ethan-1-one (A2).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 19 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 2% B, 2-42% B over 23 minutes, then a 0-minute hold at 100% B; Flow Rate: 20 mL/min; Column Temperature: 25 °C.

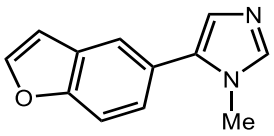
Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation to give the title compound (324 mg, 1.57 mmol, 31% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.94 (d, J = 4.0 Hz, 1H), 7.80 (dd, J = 1.2, 0.6 Hz, 1H), 7.40 (d, J = 4.0 Hz, 1H), 7.35 (d, J = 1.1 Hz, 1H), 3.80 (d, J = 0.5 Hz, 3H), 2.53 (s, 3H).

^{13}C NMR (101 MHz, DMSO-d6): δ 190.5, 142.3, 141.3, 139.0, 134.7, 129.9, 125.8, 125.8, 32.8, 26.3.

HRMS: (EI+) calculated for $[C_{12}H_{10}N_2O+H]^+$ 207.0587, found 207.0603

FTIR (ATR, cm^{-1}): 3086, 1640, 1513, 1490, 1442, 1349, 1274, 1222, 1125, 1069, 1028, 965, 928, 868, 838, 805, 667.



5-(Benzofuran-5-yl)-1-methyl-1H-imidazole (A3).

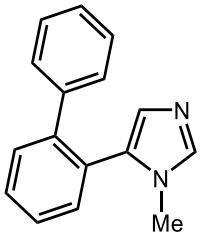
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge Phenyl, 250 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: 22% B over 18 minutes, then isocratic B; Flow Rate: 80 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation to give the title compound (289 mg, 1.46 mmol, 29% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.05 (d, J=2.1 Hz, 1H), 7.75 (d, J=1.6 Hz, 1H), 7.70 (s, 1H), 7.67 (d, J=8.5 Hz, 1H), 7.41 (dd, J=8.5, 1.8 Hz, 1H), 7.02 (s, 1H), 7.00 (s, 1H), 3.67 (s, 3H).

^{13}C NMR (101 MHz, DMSO-d6): δ 153.7, 146.8, 139.3, 132.9, 127.7, 127.2, 124.8, 124.7, 120.8, 111.5, 106.8, 32.2.

HRMS: (EI+) calculated for $[C_{12}H_{10}N_2O+H]^+$ 199.0866, found 199.0883

FTIR (ATR, cm^{-1}): 3093, 3045, 1610, 1576, 1498, 1457, 1253, 1226, 1163, 1110, 1021, 928, 902, 805, 745, 711.



5-([1,1'-Biphenyl]-2-yl)-1-methyl-1H-imidazole (A4).

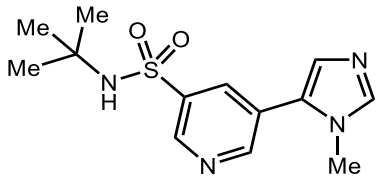
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 17% B, 17-57% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (175 mg, 0.75 mmol, 15% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.59 - 7.35 (m, 5H), 7.34 - 7.21 (m, 3H), 7.16 (d, *J*=7.2 Hz, 2H), 6.76 (s, 1H), 3.00 (s, 3H).

^{13}C NMR (101 MHz, DMSO-d6): δ 141.1, 140.6, 138.2, 131.8, 130.1, 129.1, 128.9, 128.6, 128.3, 128.2, 127.8, 127.6, 127.0, 31.1.

HRMS: (EI+) calculated for $[C_{16}H_{14}N_2+Na]^+$ 257.1049, found 257.1038.

FTIR (ATR, cm^{-1}): 3056, 1490, 1371, 1256, 1058, 1006, 913, 812, 767, 700.



N-(tert-Butyl)-5-(1-methyl-1H-imidazol-5-yl)pyridine-3-sulfonamide (**A5**).

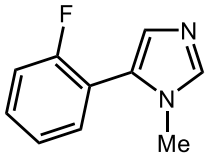
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 17% B, 17-57% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (250 mg, 0.85 mmol, 17% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 9.10 (s, 1H), 9.01 - 8.82 (m, 1H), 8.48 (s, 1H), 7.38 (s, 1H), 7.09 (s, 1H), 3.85 (s, 3H), 1.14 (s, 9H).

^{13}C NMR (101 MHz, DMSO-d6): δ 150.7, 146.0, 142.7, 141.1, 133.0, 128.7, 127.1, 125.1, 53.9, 34.7, 30.2.

HRMS: (EI+) calculated for $[C_{13}H_{18}N_4O_2S+H]^+$ 295.1223, found 295.1247.

FTIR (ATR, cm^{-1}): 3063, 2955, 2922, 2851, 1625, 1543, 1494, 1442, 1394, 1364, 1315, 1230, 1148, 1103, 1051, 1006, 928, 879, 820, 693.



5-(2-fluorophenyl)-1-methyl-1*H*-imidazole (A7).

Prepared according to the general procedure B. The title compound was isolated via flash chromatography (0 to 40% ethyl acetate in hexanes gradient with 5% triethylamine additive) as an orange oil (351 mg, 1.99 mmol, 40% yield).

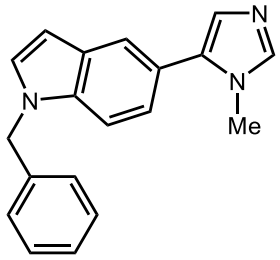
^{1}H NMR (500 MHz, $CDCl_3$): δ 7.56 (s, 1H); 7.30-7.44 (m, 2H); 7.13-7.25 (m, 2H); 7.11 (d, *J* = 1.1 Hz, 1H); 3.59 (d, *J* = 1.4 Hz, 3H).

^{13}C NMR (126 MHz, $CDCl_3$): δ 160.0 (d, *J* = 247.8 Hz); 139.3; 132.1 (d, *J* = 2.9 Hz); 130.5 (d, *J* = 8.2 Hz); 129.7; 127.8; 124.6 (d, *J* = 3.6 Hz); 117.9 (d, *J* = 15.5 Hz); 116.1 (d, *J* = 22.1 Hz); 32.3 (d, *J* = 5.2 Hz).

^{19}F NMR (282 MHz, $CDCl_3$): δ -113.1.

HRMS: (EI+) calculated for $C_{10}H_{10}FN_2$: 177.0823; found: 177.0824.

FTIR (ATR, cm^{-1}): 3278.1, 2930.2, 1640.0, 1556.0, 1488.9, 1477.2, 1447.6, 1419.4, 1227.5, 1207.7, 1110.1, 916.4, 815.3, 758.4, 649.9, 543.0.



1-Benzyl-5-(1-methyl-1H-imidazol-5-yl)-1H-indole (A9).

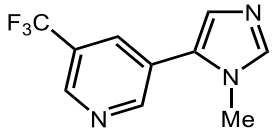
Prepared according to the general procedure A. The title compound was isolated via reversed phase chromatography (373 mg, 1.30 mmol, 26% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.64 (t, J = 1.6 Hz, 2H), 7.57 (d, J = 3.1 Hz, 1H), 7.52 (d, J = 8.5 Hz, 1H), 7.36 – 7.28 (m, 2H), 7.28 – 7.21 (m, 3H), 7.19 (dd, J = 8.5, 1.7 Hz, 1H), 6.94 (d, J = 1.2 Hz, 1H), 6.54 (dd, J = 3.2, 0.8 Hz, 1H), 5.45 (s, 2H), 3.64 (s, 3H).

^{13}C NMR (101 MHz, DMSO-d6): δ 137.2, 136.5, 133.5, 132.3, 128.4, 126.9, 126.9, 125.7, 125.4, 125.0, 120.3, 119.2, 118.5, 108.8, 99.7, 47.6, 30.5.

HRMS: (EI+) calculated for $[C_{19}H_{17}N_3+H]^+$ 288.1495, found 288.1519.

FTIR (ATR, cm^{-1}): 3101, 3026, 1703, 1476, 1446, 1356, 1328, 1267, 1230, 1181, 1110, 1028, 928, 894, 805, 767, 726, 700.



3-(1-Methyl-1H-imidazol-5-yl)-5-(trifluoromethyl)pyridine (A10).

Prepared according to the general procedure A. The crude material was purified via preparative SFC with the following conditions: Column: Chiralpak IA, 250 mm x 30 mm, 5-µm particles; Mobile Phase: 10% Methanol / 90% CO2; Flow Rate: 85 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (689 mg, 3.04 mmol, 61% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 9.04 (br s, 1H), 8.95 (br s, 1H), 8.34 (br s, 1H), 7.83 (s, 1H), 7.34 (s, 1H), 3.76 (s, 3H).

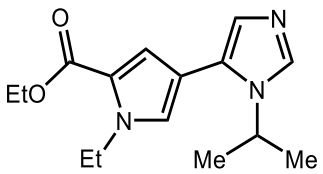
^{13}C NMR (101 MHz, DMSO-d6): δ 151.8, 144.6 (q, *J*=3.7 Hz), 141.2, 131.6 (q, *J*=3.7 Hz), 129.8, 128.1, 126.3, 125.3 (q, *J*=32.3 Hz), 122.2, 32.4.

^{19}F NMR (376 MHz, DMSO-d6): δ -60.9.

HRMS: (EI+) calculated for $[C_{10}H_8F_3N_3+H]^+$ 228.0743, found 228.0754.

FTIR (ATR, cm^{-1}): 2929, 1546, 1494, 1461, 1412, 1345, 1274, 1226, 1170, 1092, 916, 820, 715, 685.

B-Series



Ethyl 1-ethyl-4-(1-isopropyl-1*H*-imidazol-5-yl)-1*H*-pyrrole-2-carboxylate (B1).

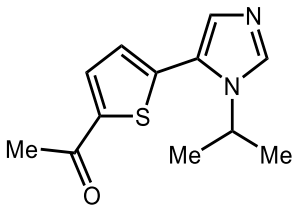
Prepared according to the general procedure B. The title compound was isolated via flash chromatography (5 to 30% acetone in hexanes gradient with 5% triethylamine additive) as a yellow oil (68.8 mg, 250 μmol, 5% yield).

^{1}H NMR (500 MHz, $CDCl_3$): δ 7.62 (d, *J* = 1.1 Hz, 1H); 6.95 (d, *J* = 1.1 Hz, 1H): 6.95 (d, *J* = 2.0 Hz, 1H): 6.90 (d, *J* = 2.0 Hz, 1H); 4.47 – 4.36 (m, 3H); 4.31 (q, *J* = 7.1 Hz, 2H); 1.47 – 1.42 (m, 9H); 1.37 (t, *J* = 7.1 Hz, 3H).

^{13}C NMR (126 MHz, $CDCl_3$): δ 160.9; 134.1; 127.4; 126.8; 126.4; 122.6; 117.6; 112.2; 60.2; 46.9; 44.5; 24.1; 17.1; 14.6.

HRMS: (EI+) calculated for $[C_{15}H_{21}N_3O_2+H]^+$ 276.1706, found 276.1707.

FTIR (ATR, cm^{-1}): 3365; 3110; 2977; 2933; 2854; 1703; 1653; 1477; 1446; 1372; 1288; 1229; 1114; 1096; 1075; 907; 812; 663.



1-(5-(1-Isopropyl-1H-imidazol-5-yl)thiophen-2-yl)ethan-1-one (B2).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 11% B, 11-51% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature:

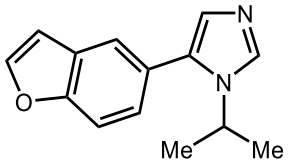
25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation. (406 mg, 1.74 mmol, 35% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.03 (s, 1H), 7.96 (d, *J*=3.9 Hz, 1H), 7.33 (d, *J*=3.9 Hz, 1H), 7.24 (d, *J*=0.9 Hz, 1H), 4.60 (hept, *J*=6.7 Hz, 1H), 2.58 - 2.52 (s, 3H), 1.51 -1.35 (m, 6H).

^{13}C NMR (101 MHz, DMSO-d6): δ 190.6, 143.1, 138.7, 137.3, 134.6, 130.0, 127.1, 124.4, 47.3, 26.3, 23.2.

HRMS: (EI+) calculated for $[C_{12}H_{14}N_2OS+H]^+$ 235.0900, found 235.0915.

FTIR (ATR, cm^{-1}): 3116, 3049, 2978, 2881, 1714, 1640, 1565, 1438, 1356, 1282, 1230, 1039, 965, 928, 808.



5-(Benzofuran-5-yl)-1-isopropyl-1H-imidazole (**B3**).

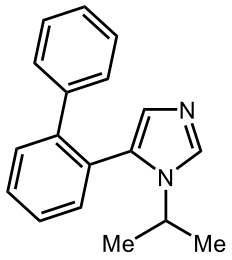
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 19% B, 19-59% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (237 mg, 1.05 mmol, 21% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.07 (d, J=2.2 Hz, 1H), 7.92 (s, 1H), 7.69 (d, J=8.2 Hz, 1H), 7.67 (s, 1H), 7.31 (dd, J=8.5, 1.8 Hz, 1H), 7.02 (dd, J=2.2, 0.9 Hz, 1H), 6.92 (s, 1H), 4.36 (dt, J=13.4, 6.7 Hz, 1H), 1.38 (d, J=6.7 Hz, 6H).

^{13}C NMR (101 MHz, DMSO-d6): δ 153.9, 146.8, 135.0, 132.0, 127.7, 125.6, 125.0, 121.8, 111.5, 106.8, 46.4, 23.5.

HRMS: (EI+) calculated for $[C_{14}H_{14}N_2O+H]^+$ 227.1179, found 227.1201.

FTIR (ATR, cm^{-1}): 3116, 2978, 1703, 1457, 1397, 1371, 1259, 1166, 1110, 1028, 931, 812, 771, 738.



5-([1,1'-Biphenyl]-2-yl)-1-isopropyl-1H-imidazole (**B4**).

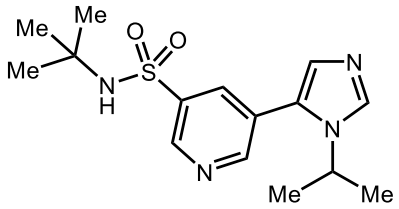
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 19 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 28% B, 28-68% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (124 mg, 0.47 mmol, 9% yield).

^{1}H NMR (500 MHz, DMSO-d6): δ 7.68 (s, 1H), 7.58 - 7.44 (m, 3H), 7.40 (dd, *J*=7.6, 1.0 Hz, 1H), 7.33 - 7.23 (m, 3H), 7.16 (d, *J*=7.0 Hz, 2H), 6.85 (s, 1H), 3.61 (hept, *J*=6.7 Hz, 1H), 0.86 (d, *J*=6.7 Hz, 6H).

^{13}C NMR (126MHz, DMSO-d_6): δ 141.5, 140.8, 134.7, 132.5, 131.2, 130.6, 129.7, 129.2, 128.7, 128.6, 128.2, 128.1, 127.4, 46.8, 23.3.

HRMS: (EI+) calculated for $[C_{18}H_{18}N_2+H]^+$ 263.1543, found 263.1568.

FTIR (ATR, cm^{-1}): 2978, 2929, 1707, 1446, 1394, 1360, 1274, 1218, 1159, 1107, 1006, 887, 834, 767, 741, 697.



N-(tert-Butyl)-5-(1-isopropyl-1H-imidazol-5-yl)pyridine-3-sulfonamide (**B5**).

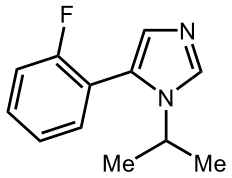
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 14% B, 14-54% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (410 mg, 1.27 mmol, 25% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.97 (d, *J*=2.2 Hz, 1H), 8.85 (d, *J*=2.1 Hz, 1H), 8.18 (t, *J*=2.1 Hz, 1H), 8.06 (s, 1H), 7.17 (d, *J*=1.0 Hz, 1H), 4.34 (hept, *J*=6.7 Hz, 1H), 1.40 (d, *J*=6.7 Hz, 6H), 1.13 (s, 9H).

^{13}C NMR (101 MHz, DMSO-d6): δ 151.8, 145.7, 140.4, 137.0, 133.1, 129.3, 127.2, 126.5, 53.8, 42.1, 29.7, 23.3.

HRMS: (EI+) calculated for $[C_{15}H_{22}N_4O_2S+Na]^+$ 345.1356, found 345.1363.

FTIR (ATR, cm^{-1}): 3063, 2981, 2709, 1546, 1479, 1397, 1319, 1233, 1148, 1103, 1047, 1006, 931, 834, 711, 663.



5-(2-fluorophenyl)-1-isopropyl-1*H*-imidazole (B7).

Prepared according to the general procedure B. The title compound was isolated via flash chromatography (5 to 30% ethyl acetate in hexanes gradient with 5% triethylamine additive) as a red oil (403 mg, 1.98 mmol, 40% yield).

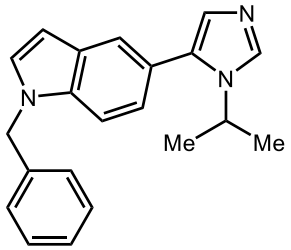
^{1}H NMR (500 MHz, $CDCl_3$): δ 7.71 (s, 1H); 7.46 – 7.36 (m, 1H); 7.32 (td, *J* = 7.5 Hz, 1.9 Hz, 1H); 7.13 – 7.26 (m, 2H); 7.04 (d, *J* = 1.1 Hz, 1H); 4.18 (hept, *J* = 6.7 Hz, 1H); 1.42 (d, *J* = 6.7 Hz, 6H).

^{13}C NMR (126 MHz, $CDCl_3$): δ 160.2 (d, *J* = 247.5 Hz); 135.1; 132.6 (d, *J* = 2.7 Hz); 130.7 (d, *J* = 8.1 Hz); 129.0; 126.4; 124.6 (d, *J* = 3.8 Hz); 118.3 (d, *J* = 15.6 Hz); 116.0 (d, *J* = 21.9 Hz); 47.6 (d, *J* = 2.6 Hz); 24.1

^{19}F NMR (282 MHz, $CDCl_3$): δ -112.8.

HRMS: (EI+) calculated for $[C_{12}H_{13}FN_2+H]^+$ 205.1133; found: 205.1136.

FTIR (ATR, cm^{-1}): 3397; 2980; 2937; 1639.5; 1580; 1556; 1486; 1453; 1372; 1220; 1114; 919; 816; 762; 662.



1-Benzyl-5-(1-isopropyl-1H-imidazol-5-yl)-1H-indole (**B9**).

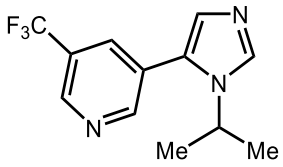
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 32% B, 32-72% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (305 mg, 0.97 mmol, 19% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.87 (s, 1H), 7.58 (d, J=3.2 Hz, 1H), 7.54 (m, 2H), 7.36 - 7.29 (m, 2H), 7.29 - 7.20 (m, 3H), 7.09 (dd, J=8.4, 1.7 Hz, 1H), 6.84 (d, J=0.9 Hz, 1H), 6.54 (d, J=2.8 Hz, 1H), 5.45 (s, 2H), 4.35 (hept, J=6.7 Hz, 1H), 1.36 (d, J=6.7 Hz, 6H).

^{13}C NMR (101 MHz, DMSO-d6): δ 138.1, 135.3, 134.5, 133.1, 130.1, 128.6, 128.5, 127.4, 127.1, 126.4, 122.5, 121.1, 121.0, 110.4, 101.3, 49.2, 46.2, 23.6.

HRMS: (EI+) calculated for $[C_{21}H_{21}N_3+H]^+$ 316.1808, found 316.1831.

FTIR (ATR, cm^{-1}): 3090, 2974, 2929, 1476, 1442, 1390, 1356, 1263, 1181, 1110, 1028, 890, 812, 775, 730, 697.



3-(1-Isopropyl-1H-imidazol-5-yl)-5-(trifluoromethyl)pyridine (B10).

Prepared according to the general procedure A. The title compound was isolated via reversed phase chromatography (106 mg, 0.42 mmol, 8% yield).

^{1}H NMR (500 MHz, DMSO-d6): δ 9.14 - 9.03 (m, 2H), 8.37 (s, 1H), 7.63 (s, 1H), 7.17 (s, 1H), 4.59 (hept, J=6.7 Hz, 1H), 1.47 (d, J=6.7 Hz, 6H).

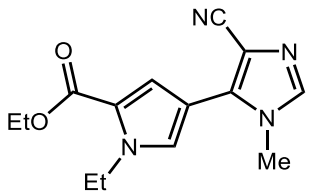
^{13}C NMR (126 MHz, DMSO-d6): δ 153.0, 146.1, 142.2, 133.6, 129.8, 128.0, 125.7 (q, J=32.7 Hz), 123.9 (q, J=1.0 Hz), 119.1, 48.7, 23.8.

^{19}F NMR (470 MHz, DMSO-d6): δ -61.29.

HRMS: (EI+) calculated for $[C_{12}H_{12}F_3N_3+H]^+$ 256.1056, found 256.1078.

FTIR (ATR, cm^{-1}): 3108, 1666, 1427, 1345, 1297, 1177, 1125, 1021, 913, 831, 797, 715.

C-Series



Ethyl 4-(4-cyano-1-methyl-1H-imidazol-5-yl)-1-ethyl-1H-pyrrole-2-carboxylate (C1).

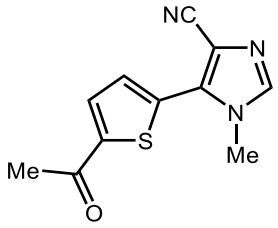
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with 0.1% trifluoroacetic acid; Mobile Phase B: 95:5 acetonitrile: water with 0.1% trifluoroacetic acid; Gradient: a 0-minute hold at 16% B, 16-56% B over 20 minutes, then a 2-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (209 mg, 0.77 mmol, 15% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.85 (s, 1H), 7.69 (s, 1H), 7.21 (s, 1H), 4.39 (q, *J*=7.1 Hz, 2H), 4.27 (q, *J*=7.1 Hz, 2H), 3.76 (s, 3H), 1.38 - 1.26 (m, 6H).

^{13}C NMR (101 MHz, DMSO-d6): δ 159.7, 140.4, 136.2, 128.0, 122.2, 116.6, 116.1, 108.9, 108.2, 59.9, 43.9, 33.1, 16.8, 14.2.

HRMS: (EI+) calculated for $[C_{14}H_{16}N_4O_2+H]^+$ 273.1346, found 273.1358.

FTIR (ATR, cm^{-1}): 3138, 3101, 2978, 2221, 1695, 1599, 1509, 1449, 1379, 1282, 1248, 1207, 1174, 1080, 965, 924, 834, 752.



5-(5-Acetylthiophen-2-yl)-1-methyl-1H-imidazole-4-carbonitrile (C2).

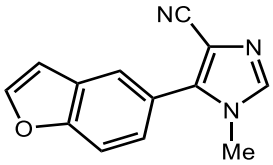
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 1% B, 1-41% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by UV signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (125 mg, 0.54 mmol, 11% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.27 (d, J=3.9 Hz, 1H), 7.23 (s, 1H), 6.87 (d, J=4.0 Hz, 1H), 3.11 (s, 3H), 1.88 (s, 3H).

^{13}C NMR (101 MHz, DMSO-d6): δ 191.7, 146.8, 142.8, 134.8, 134.4, 134.3, 130.9, 115.8, 113.3, 33.9, 26.6.

HRMS: (EI+) calculated for $[C_{11}H_9N_3OS+H]^+$ 232.0539, found 232.0547.

FTIR (ATR, cm^{-1}): 3104, 2228, 1654, 1498, 1442, 1371, 1323, 1274, 1185, 1092, 1043, 965, 935, 849, 820.



5-(Benzofuran-5-yl)-1-methyl-1H-imidazole-4-carbonitrile (C3).

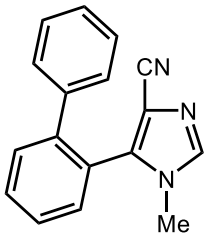
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with 10-mM ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with 10-mM ammonium acetate; Gradient: a 0-minute hold at 11% B, 11-51% B over 20 minutes, then a 2-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (538 mg, 2.41 mmol, 48% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.14 (s, 1H), 7.99 (s, 1H), 7.90 (s, 1H), 7.82 (d, J=8.6 Hz, 1H), 7.51 (d, J=8.6 Hz, 1H), 7.09 (s, 1H), 3.65 (s, 3H).

^{13}C NMR (101 MHz, DMSO-d6): δ 154.7, 147.4, 141.7, 140.8, 127.9, 125.4, 122.5, 120.9, 116.1, 112.1, 110.8, 106.9, 32.8.

HRMS: (EI+) calculated for $[C_{13}H_9N_3O+H]^+$ 224.0818, found 224.0827.

FTIR (ATR, cm^{-1}): 3116, 3041, 2225, 1509, 1457, 1382, 1263, 1203, 1159, 1028, 872, 820, 775, 685.



5-([1,1'-Biphenyl]-2-yl)-1-methyl-1H-imidazole-4-carbonitrile (C4).

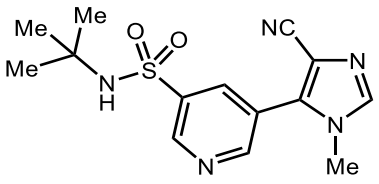
Prepared according to the general procedure A. The title compound was isolated via reversed phase chromatography (174 mg, 0.67 mmol, 13% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.81 (s, 1H), 7.74 - 7.64 (m, 1H), 7.64 - 7.48 (m, 3H), 7.42 - 7.26 (m, 3H), 7.20 - 7.00 (m, 2H), 3.17 (s, 3H).

^{13}C NMR (101 MHz, DMSO-d6): δ 142.5, 141.6, 140.6, 139.9, 132.1, 131.4, 130.9, 129.0, 128.9, 128.5, 128.0, 124.8, 116.0, 112.6, 32.6.

HRMS: (EI+) calculated for $[C_{17}H_{13}N_3+H]^+$ 260.1182, found 260.1188.

FTIR (ATR, cm^{-1}): 3056, 2228, 1505, 1446, 1304, 1241, 1200, 1118, 998, 831, 779, 749, 704, 678.



N-(tert-Butyl)-5-(4-cyano-1-methyl-1H-imidazol-5-yl)pyridine-3-sulfonamide (C5).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B:

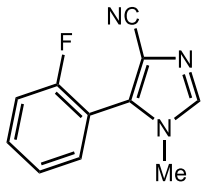
95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 7% B, 7-47% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (960 mg, 3.01 mmol, 60% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 9.12 (s, 1H), 9.03 (s, 1H), 8.42 (s, 1H), 8.10 (s, 1H), 7.97 (s, 1H), 3.73 (s, 3H), 1.15 (s, 9H).

^{13}C NMR (101 MHz, DMSO-d6): δ 151.9, 147.6, 142.0, 140.7, 137.0, 134.3, 123.0, 115.3, 112.3, 54.0, 32.9, 29.7.

HRMS: (EI+) calculated for $[C_{14}H_{17}N_5O_2S+H]^+$ 320.1176, found 320.1186.

FTIR (ATR, cm^{-1}): 3138, 2967, 2862, 2232, 1509, 1468, 1442, 1367, 1323, 1207, 1148, 1103, 995, 853, 808, 752, 700, 663.



5-(2-Fluorophenyl)-1-methyl-1H-imidazole-4-carbonitrile (C7).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 250 mm x 19 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 11.5-minute hold at 20% B,; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals.

Fractions containing the desired product were combined and dried via centrifugal evaporation (271 mg, 1.35 mmol, 27% yield).

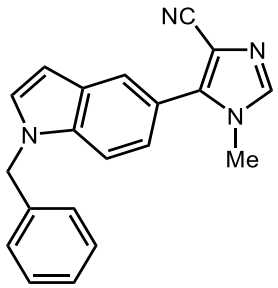
^{1}H NMR (600 MHz, DMSO-d6): δ 8.03 (s, 1H), 7.67 - 7.64 (m, 1H), 7.59 (td, *J* = 7.5, 1.8 Hz, 1H), 7.46 - 7.44 (m, 1H), 7.42 (td, *J* = 7.5, 1.1 Hz, 1H), 3.58 (s, 3H).

^{13}C NMR (101 MHz, DMSO-d6): δ 159.3 (d, *J* = 248.1 Hz), 141.6, 135.6, 133.1 (d, *J* = 8.5 Hz), 132.1 (d, *J* = 1.6 Hz), 125.5 (d, *J* = 3.6 Hz), 116.6 (d, *J* = 21.3 Hz), 115.5, 114.0 (d, *J* = 15.0 Hz), 112.6, 32.8.

^{19}F NMR (376 MHz, $CDCl_3$): δ 113.0.

HRMS: (EI+) calculated for $[C_{11}H_8FN_3+H]^+$ 202.0775, found 202.0781.

FTIR (ATR, cm^{-1}): 31116, 2225, 1502, 1478, 1379, 1304, 1274, 1215, 1189, 1107, 1051, 1002, 946, 872, 820, 764.



5-(1-Benzyl-1H-indol-5-yl)-1-methyl-1H-imidazole-4-carbonitrile (C9).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 27% B, 27-67% B

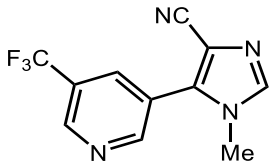
over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation. (665 mg, 2.13 mmol, 43% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.93 (s, 1H), 7.78 (d, *J*=1.3 Hz, 1H), 7.68 - 7.66 (m, 1H), 7.65 (s, 1H), 7.38 - 7.18 (m, 6H), 6.62 (d, *J*=3.2 Hz, 1H), 5.49 (s, 2H), 3.65 (s, 3H).

^{13}C NMR (101 MHz, DMSO-d6): δ 142.9, 140.5, 137.9, 136.0, 130.7, 128.6, 128.4, 127.5, 127.1, 121.9, 121.7, 116.8, 116.5, 110.9, 110.2, 101.7, 49.2, 32.8.

HRMS: (EI+) calculated for $[C_{20}H_{16}N_4+H]^+$ 313.1448, found 313.1454.

FTIR (ATR, cm^{-1}): 3116, 2225, 1505, 1472, 1379, 1334, 1285, 1203, 1174, 1080, 879, 812, 734, 700.



1-Methyl-5-(5-(trifluoromethyl)pyridin-3-yl)-1H-imidazole-4-carbonitrile (C10).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 9% B, 9-49% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (779 mg, 3.09 mmol, 62% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 9.16 (d, *J*=1.2 Hz, 1H), 9.12 (d, *J*=1.8 Hz, 1H), 8.55 (s, 1H), 8.11 (s, 1H), 3.74 (s, 3H).

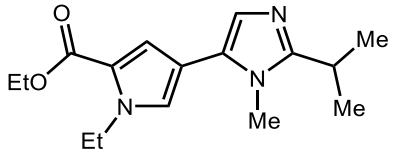
^{13}C NMR (101 MHz, DMSO-d6): δ 153.2, 147.2 (q, *J*=3.7 Hz), 142.0, 136.7, 134.0 (q, *J*=3.7 Hz), 125.4 (q, *J*=32.8 Hz), 124.6, 123.1, 121.9, 115.3, 112.5, 32.9.

^{19}F NMR (376 MHz, DMSO-d6): δ -60.9.

HRMS: (EI+) calculated for $[C_{11}H_7F_3N_4+H]^+$ 253.0696, found 253.0706.

FTIR (ATR, cm^{-1}): 3108, 3056, 2225, 1509, 1461, 1349, 1256, 1148, 1121, 1054, 1010, 916, 857, 827, 749, 711.

D-Series



Ethyl 1-ethyl-4-(2-isopropyl-1-methyl-1H-imidazol-5-yl)-1H-pyrrole-2-carboxylate (D1).

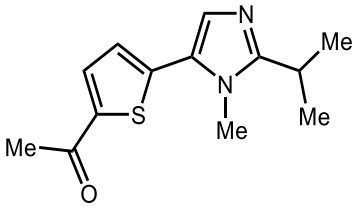
Prepared according to the general procedure A. The title compound was isolated via reversed phase chromatography (152 mg, 0.53 mmol, 11% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.37 (d, J=2.0 Hz, 1H), 6.95 (d, J=2.1 Hz, 1H), 6.78 (s, 1H), 4.33 (q, J=7.1 Hz, 2H), 4.23 (q, J=7.1 Hz, 2H), 3.54 (s, 3H), 3.14 - 2.98 (m, 1H), 1.32 (t, J=7.1 Hz, 3H), 1.28 (t, J=7.1 Hz, 3H), 1.21 (d, J=6.8 Hz, 6H).

^{13}C NMR (101 MHz, DMSO-d6): δ 160.0, 152.4, 126.7, 123.8, 121.3, 116.2, 112.4, 59.6, 48.6, 43.5, 30.5, 25.5, 21.4, 16.9, 14.3.

HRMS: (EI+) calculated for $[C_{16}H_{23}N_3O_2+H]^+$ 290.1863, found 290.1875.

FTIR (ATR, cm^{-1}): 2974, 1699, 1461, 1401, 1282, 1233, 1095, 1017, 801, 760.



1-(5-(2-Isopropyl-1-methyl-1H-imidazol-5-yl)thiophen-2-yl)ethan-1-one (**D2**).

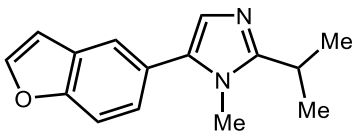
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge Phenyl, 250 mm x 19 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with trifluoroacetic acid; Gradient: a 13-minute hold at 25% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (551 mg, 2.22 mmol, 44% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.94 (d, J=4.0 Hz, 1H), 7.34 (d, J=3.9 Hz, 1H), 7.18 (s, 1H), 3.41 (s, 3H), 3.14 (hept, J=6.8 Hz, 1H), 2.58 - 2.52 (m, 3H), 1.24 (d, J=6.8 Hz, 6H).

^{13}C NMR (101 MHz, DMSO-d6): δ 190.9, 156.0, 142.5, 140.2, 135.2, 128.6, 126.2, 126.1, 31.6, 26.8, 26.0, 21.8.

HRMS: (EI+) calculated for $[C_{13}H_{16}N_2OS+H]^+$ 249.1056, found 249.1062.

FTIR (ATR, cm^{-1}): 2976, 2929, 2870, 1654, 1565, 1513, 1461, 1360, 1315, 1282, 1077, 1036, 991, 939, 827, 797.



5-(Benzofuran-5-yl)-2-isopropyl-1-methyl-1H-imidazole (D3).

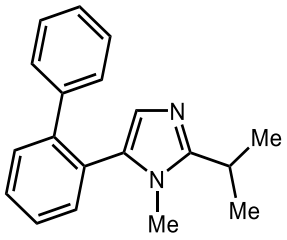
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 15% B, 15-55% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (588 mg, 2.45 mmol, 49% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.04 (d, J=2.1 Hz, 1H), 7.69 (s, 1H), 7.66 (d, J=8.6 Hz, 1H), 7.35 (dd, J=8.5, 1.8 Hz, 1H), 6.99 (dd, J=2.2, 0.9 Hz, 1H), 6.86 (s, 1H), 3.54 (s, 3H), 3.11 (hept, J=6.8 Hz, 1H), 1.26 (d, J=6.8 Hz, 6H).

^{13}C NMR (101 MHz, DMSO-d6): δ 153.4, 152.9, 146.5, 132.5, 127.4, 125.1, 124.9, 120.9, 111.2, 106.6, 30.4, 25.5, 21.2.

HRMS: (EI+) calculated for $[C_{15}H_{16}N_2O+H]^+$ 241.1335, found 241.1345.

FTIR (ATR, cm^{-1}): 3086, 3045, 2974, 1457, 1312, 1129, 1069, 1021, 954, 905, 812, 782, 749.



5-([1,1'-Biphenyl]-2-yl)-2-isopropyl-1-methyl-1H-imidazole (D4).

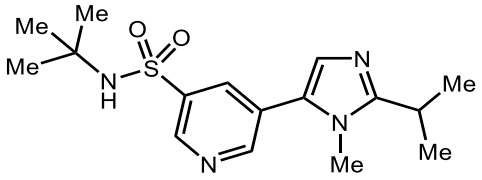
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with 10-mM ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with 10-mM ammonium acetate; Gradient: a 0-minute hold at 26% B, 26-66% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 45 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (940 mg, 3.41 mmol, 68% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.53 (d, J=4.9 Hz, 2H), 7.49 - 7.35 (m, 2H), 7.35 - 7.20 (m, 3H), 7.20 - 7.04 (m, 2H), 6.71 (s, 1H), 2.90 - 2.78 (m, 1H), 2.74 (s, 3H), 1.09 (d, J=6.7 Hz, 6H).

^{13}C NMR (101 MHz, DMSO-d6): δ 152.6, 140.7, 140.5, 131.8, 131.7, 129.9, 129.0, 128.5, 128.5, 128.1, 127.6, 126.9, 125.8, 29.7, 25.4, 21.2.

HRMS: (EI+) calculated for $[C_{19}H_{20}N_2+H]^+$ 277.1699, found 277.1710.

FTIR (ATR, cm^{-1}): 2970, 243, 1707, 1449, 1364, 1285, 1159, 1073, 1006, 954, 883, 834, 745, 704.



N-(tert-Butyl)-5-(2-isopropyl-1-methyl-1H-imidazol-5-yl)pyridine-3-sulfonamide (D5).

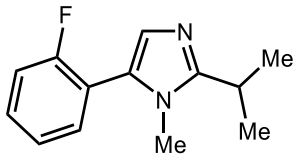
Prepared according to the general procedure A. The crude material was purified via preparative SFC with the following conditions: Column: Chiralpak, 250 mm x 21 mm, 5-µm particles; Mobile Phase: 12% Methanol/ CO2; Flow Rate: 45 mL/min 150 Bar; Column Temperature: 40 C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (970 mg, 2.89 mmol, 58% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.91 (d, J=2.1 Hz, 1H), 8.89 (d, J=2.0 Hz, 1H), 8.22 (t, J=2.1 Hz, 1H), 7.84 (br s, 1H), 7.14 (s, 1H), 3.62 (s, 3H), 3.15 (hept, J=6.8 Hz, 1H), 1.26 (d, J=6.8 Hz, 6H), 1.14 (s, 9H).

^{13}C NMR (101 MHz, DMSO-d6): δ 155.3, 151.0, 144.9, 140.4, 132.1, 128.1, 127.5, 126.7, 53.8, 31.0, 29.8, 25.6, 21.3.

HRMS: (EI+) calculated for $[C_{16}H_{24}N_4O_2S+Na]^+$ 359.1512, found 359.1516.

FTIR (ATR, cm^{-1}): 2967, 2795, 2672, 1543, 1490, 1394, 1323, 1140, 1102, 1006, 834, 704.



5-(2-Fluorophenyl)-2-isopropyl-1-methyl-1H-imidazole (D7).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 16% B, 16-56% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (318 mg, 1.46 mmol, 29% yield).

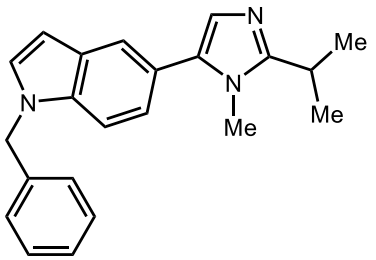
^{1}H NMR (400 MHz, DMSO-d6): δ 7.53 - 7.38 (m, 2H), 7.38 - 7.23 (m, 2H), 6.88 (s, 1H), 3.44 (d, J=1.5 Hz, 3H), 3.12 (hept, J=6.8 Hz, 1H), 1.25 (d, J=6.8 Hz, 6H).

^{13}C NMR (101 MHz, DMSO-d6): δ 161.4 (d, *J*=184.9 Hz), 158.0, 153.7, 131.8 (d, *J*=2.9 Hz), 130.3 (d, *J*=8.8 Hz), 126.8 (d, *J*=1.5 Hz), 124.8 (d, *J*=2.9 Hz), 118.1 (d, *J*=15.4 Hz), 115.9 (d, *J*=22.0 Hz), 30.5, 25.6, 21.4.

^{19}F NMR (376 MHz, DMSO-d6): δ -113.8.

HRMS: (EI+) calculated for $[C_{13}H_{15}FN_2+H]^+$ 219.1292, found 219.1303.

FTIR (ATR, cm^{-1}): 2970, 2929, 1707, 1673, 1464, 1386, 1259, 1159, 1099, 946, 816, 760, 678.



1-Benzyl-5-(2-isopropyl-1-methyl-1H-imidazol-5-yl)-1H-indole (D9).

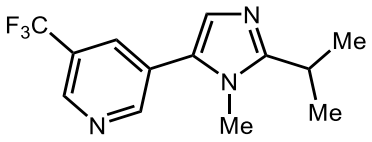
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 31% B, 31-71% B over 15 minutes, then a 0-minute hold at 100% B; Flow Rate: 45 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (938 mg, 2.85 mmol, 57% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.58 (s, 1H), 7.56 (d, J=3.1 Hz, 1H), 7.51 (d, J=8.4 Hz, 1H), 7.35 - 7.28 (m, 2H), 7.28 - 7.19 (m, 3H), 7.14 (dd, J=8.5, 1.7 Hz, 1H), 6.78 (s, 1H), 6.53 (d, J=3.1 Hz, 1H), 5.45 (s, 2H), 3.52 (s, 3H), 3.09 (hept, J=6.8 Hz, 1H), 1.25 (d, J=6.7 Hz, 6H).

^{13}C NMR (101 MHz, DMSO-d6): δ 152.6, 138.2, 135.1, 133.8, 129.9, 128.5, 128.5, 127.4, 127.0, 124.5, 122.2, 121.4, 120.5, 110.3, 101.3, 49.2, 30.6, 25.7, 21.5.

HRMS: (EI+) calculated for $[C_{22}H_{23}N_3+H]^+$ 330.1965, found 330.1973.

FTIR (ATR, cm^{-1}): 2967, 2929, 2870, 1703, 1449, 1390, 1356, 1326, 1259, 1181, 1073, 883, 801, 775, 723, 697.



3-(2-Isopropyl-1-methyl-1H-imidazol-5-yl)-5-(trifluoromethyl)pyridine (D10).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 16% B, 16-56% B over 15 minutes, then a 0-minute hold at 100% B; Flow Rate: 45 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by UV signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (755 mg, 2.8 mmol, 56% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.99 (d, *J*=1.0 Hz, 1H), 8.93 (d, *J*=1.22 Hz, 1H), 8.29 (s, 1H), 7.17 (s, 1H), 3.62 (s, 3H), 3.16 (hept, J=6.8 Hz, 1H), 1.26 (d, J=6.7 Hz, 6H).

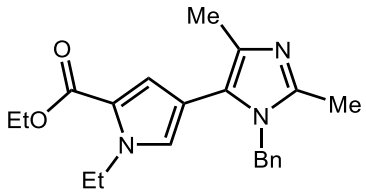
^{13}C NMR (101 MHz, DMSO-d6): δ 155.2, 152.1, 144.3 (q, *J*=4.4 Hz), 131.9 (q, *J*=3.4 Hz), 128.0, 127.8, 126.8, 125.3 (q, *J*=32.3 Hz), 122.2, 30.9, 25.6, 21.3.

^{19}F NMR (376 MHz, DMSO-d6): δ -60.9.

HRMS: (EI+) calculated for $[C_{13}H_{14}F_3N_3+H]^+$ 270.1213, found 270.1225.

FTIR (ATR, cm^{-1}): 3034, 2985, 2940, 1546, 1494, 1326, 1274, 1230, 1177, 1121, 1095, 1047, 950, 820, 711.

E-Series



Ethyl 4-(1-benzyl-2,4-dimethyl-1H-imidazol-5-yl)-1-ethyl-1H-pyrrole-2-carboxylate (E1).

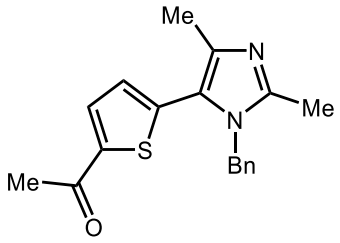
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 250 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 11-minute hold at 41% B; Flow Rate: 80 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by UV signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (192 mg, 0.55 mmol, 11% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.34 - 7.22 (m, 3H), 7.10 (d, *J*=1.7 Hz, 1H), 6.90 (br d, *J*=7.3 Hz, 2H), 6.64 (d, *J*=1.7 Hz, 1H), 5.06 (s, 2H), 4.31 - 4.13 (m, 4H), 2.16 (s, 3H), 2.07 (s, 3H), 1.29 - 1.19 (m, 6H).

^{13}C NMR (101 MHz, DMSO-d6): δ 159.9, 142.9, 137.9, 132.3, 128.7, 128.0, 127.1, 125.7, 121.9, 121.2, 117.3, 111.9, 59.5, 46.4, 43.4, 16.8, 14.2, 13.4, 13.1.

HRMS: (EI+) calculated for $C_{21}H_{25}N_3O_2$+H]$^+$ 352.2020, found 352.2041.

FTIR (ATR, cm^{-1}): 2981, 2933, 1699, 1449, 1416, 1375, 1278, 1237, 1203, 1121, 1017, 857, 797, 760.



1-(5-(1-Benzyl-2,4-dimethyl-1H-imidazol-5-yl)thiophen-2-yl)ethan-1-one (E2).

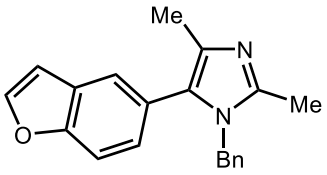
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 250 mm x 30 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 10-minute hold at 35% B; Flow Rate: 80 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by UV signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (553 mg, 1.78 mmol, 36% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.88 (d, *J*=3.9 Hz, 1H), 7.39 - 7.29 (m, 2H), 7.29 - 7.21 (m, 1H), 7.03 (d, *J*=3.9 Hz, 1H), 6.91 (d, *J*=7.5 Hz, 2H), 5.23 (s, 2H), 2.24 (s, 3H), 2.20 (m, 6H).

^{13}C NMR (101 MHz, DMSO-d6): δ 190.5, 145.8, 143.4, 139.4, 137.0, 136.6, 134.3, 128.8, 128.1, 127.3, 125.6, 120.6, 46.8, 26.3, 13.8, 13.1.

HRMS: (EI+) calculated for $[C_{18}H_{18}N_2OS+H]^+$ 311.1213, found 311.1236.

FTIR (ATR, cm^{-1}): 3063, 2914, 1647, 1584, 1498, 1453, 1412, 1356, 1274, 1107, 1073, 1032, 991, 954, 898, 812, 738, 693.



5-(Benzofuran-5-yl)-1-benzyl-2,4-dimethyl-1H-imidazole (E3).

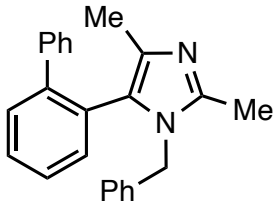
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 22% B, 22-62% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (384 mg, 1.27 mmol, 25% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.03 - 7.93 (m, 1H), 7.65 - 7.54 (m, 1H), 7.54 - 7.46 (m, 1H), 7.35 - 7.22 (m, 2H), 7.22 - 7.16 (m, 1H), 7.16 - 7.08 (m, 1H), 6.97 - 6.88 (m, 1H), 6.88 - 6.78 (m, 2H), 5.10 - 4.93 (m, 2H), , 2.23 - 2.13 (m, 3H), 2.10 - 1.98 (m, 3H).

^{13}C NMR (101 MHz, DMSO-d6): δ 153.8, 146.8, 143.5, 137.6, 132.6, 128.8, 128.3, 127.7, 127.3, 126.4, 125.9, 125.3, 122.9, 111.6, 107.0, 45.3, 13.2, 13.1.

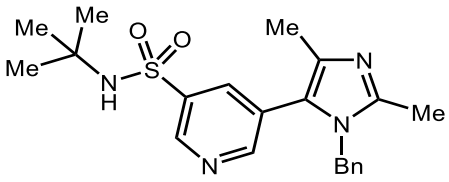
HRMS: (EI+) calculated for $[C_{20}H_{18}N_2O+H]^+$ 303.1492, found 303.1512.

FTIR (ATR, cm^{-1}): 2914, 1707, 1412, 1353, 1263, 1162, 1129, 1021, 875, 812, 767, 738.



5-([1,1'-biphenyl]-2-yl)-1-benzyl-2,4-dimethyl-1*H*-imidazole (E4).

We could not synthesize this product in sufficient amount for characterizations. No reactivity was observed in HTE with any of the 24 ligands.



5-(1-Benzyl-2,4-dimethyl-1H-imidazol-5-yl)-N-(tert-butyl)pyridine-3-sulfonamide (E5).

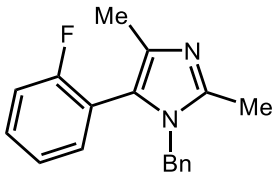
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 19 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 35% B, 35-72% B over 30 minutes, then a 0-minute hold at 100% B; Flow Rate: 20 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (414 mg, 1.04 mmol, 21% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.84 (d, *J*=2.0 Hz, 1H), 8.59 (s, 1H), 7.97 (s, 1H), 7.29 - 7.12 (m, 3H), 6.76 (br d, *J*=7.1 Hz, 2H), 5.11 (s, 2H), 2.27 (s, 3H), 2.06 (s, 3H), 1.00 (s, 9H).

^{13}C NMR (101 MHz, DMSO-d6): δ 152.5, 146.0, 145.9, 141.1, 137.0, 135.6, 134.8, 129.2, 127.9, 127.2, 126.1, 123.8, 54.0, 47.2, 30.2, 13.3, 13.1.

HRMS: (EI+) calculated for $[C_{21}H_{26}N_4O_2S+H]^+$ 399.1849, found 399.1869.

FTIR (ATR, cm^{-1}): 3052, 2967, 2840, 1595, 1550, 1446, 1394, 1326, 1148, 1002, 902, 801, 745, 700.



1-Benzyl-5-(2-fluorophenyl)-2,4-dimethyl-1H-imidazole (E7)

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 250 mm x 30 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 17.5-minute hold at 35% B; Flow Rate: 80 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (103 mg, 0.37 mmol, 7% yield).

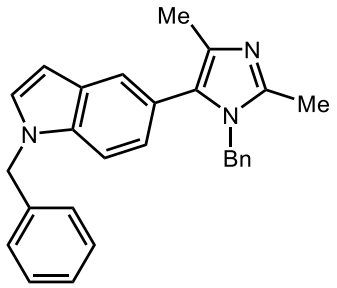
^{1}H NMR (500 MHz, DMSO-d6): δ 7.47 - 7.39 (m, 1H), 7.31 - 7.18 (m, 6H), 6.84 (d, *J*=7.2 Hz, 2H), 5.01 (br s, 2H), 2.24 (s, 3H), 2.01 (s, 3H).

^{13}C NMR (126 MHz, DMSO-d6): δ 161.1, 159.2, 144.8, 137.7, 134.8, 133.0, 131.0 (d, J=8.4 Hz), 129.0, 127.6, 126.3, 125.1 (br d, J=4.2 Hz), 122.0, 118.6, 118.5, 116.4 (d, J=23.0 Hz), 47.2, 13.6, 13.4.

^{19}F NMR (470 MHz, DMSO-d6): δ -113.28.

HRMS: (EI+) calculated for $[C_{18}H_{17}FN_2+H]^+$ 281.1449, found 281.1472.

FTIR (ATR, cm^{-1}): 2922, 1703, 1572, 1490, 1453, 1408, 1356, 1252, 1215, 1013, 879, 816, 760, 693.



1-Benzyl-5-(1-benzyl-2,4-dimethyl-1H-imidazol-5-yl)-1H-indole (E9).

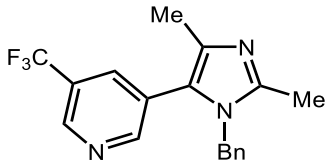
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 19 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 35% B, 35-72% B over 30 minutes, then a 0-minute hold at 100% B; Flow Rate: 20 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (413 mg, 1.06 mmol, 21% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.53 (d, *J*=3.1 Hz, 1H), 7.47 (d, *J*=8.6 Hz, 1H), 7.44 - 7.36 (m, 1H), 7.36 - 7.13 (m, 8H), 6.94 (dd, *J*=8.4, 1.5 Hz, 1H), 6.85 (d, *J*=7.2 Hz, 2H), 6.46 (d, *J*=3.1 Hz, 1H), 5.40 (s, 2H), 5.02 (s, 2H), 2.15 (s, 3H), 2.04 (s, 3H).

^{13}C NMR (101 MHz, DMSO-d6): δ 142.7, 138.1, 137.8, 135.0, 132.0, 129.8, 129.2, 128.6, 128.5, 128.3, 127.4, 127.2, 127.0, 125.7, 123.2, 122.0, 121.3, 110.2, 101.2, 49.2, 46.3, 13.2, 13.2.

HRMS: (EI+) calculated for $[C_{27}H_{25}N_3+H]^+$ 392.2121, found 392.2140.

FTIR (ATR, cm^{-1}): 2933, 1572, 1408, 1356, 1252, 1177, 1077, 1028, 805, 723, 693.



3-(1-Benzyl-2,4-dimethyl-1H-imidazol-5-yl)-5-(trifluoromethyl)pyridine (E10).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 19 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with 0.05% trifluoroacetic acid; Mobile Phase B: 95:5 acetonitrile: water with 0.05% trifluoroacetic acid; Gradient: a 0-minute hold at 13% B, 13-50% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 20 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (171 mg, 0.52 mmol, 10% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.88 (s, 1H), 8.69 (s, 1H), 7.89 (s, 1H), 7.37 - 7.11 (m, 3H), 6.84 (d, *J*=7.2 Hz, 2H), 5.11 (s, 2H), 2.30 (s, 3H), 2.08 (s, 3H).

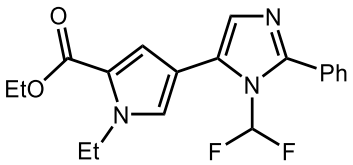
^{13}C NMR (101 MHz, DMSO-d6): δ 153.6, 145.3, 144.7 (q, *J*=4.4 Hz, 1C), 137.1, 135.4, 133.4 (q, *J*=3.7 Hz, 1C), 128.7, 127.3, 127.1, 125.8, 125.0 (q, *J*=32.3 Hz, 1C), 123.1, 122.0, 46.9, 13.1, 13.0.

^{19}F NMR (376 MHz, DMSO-d6): δ -61.1.

HRMS: (EI+) calculated for $[C_{18}H_{16}F_3N_3+H]^+$ 332.1369, found 332.1369.

FTIR (ATR, cm^{-1}): 2926, 1707, 1453, 1408, 1334, 1308, 1252, 1177, 1129, 909, 823, 715.

F-Series



5-(1-(difluoromethyl)-2-phenyl-1H-imidazol-5-yl)-1-ethyl-1H-pyrrol-2-ylpropionate (F1).

Prepared according to the general procedure B. The title compound was isolated via flash chromatography (gradient 0%-20% EtOAc/hexane) as a reddish oil (68.6 mg, 0.19 mmol, 4% yield).

TLC (SiO_2) R_f = 0.17 in 4:1 hexanes/EtOAc.

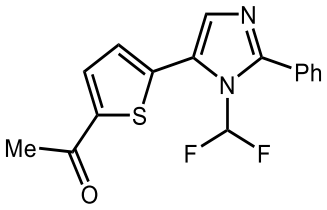
^{1}H NMR (500 MHz, $CDCl_3$): δ 7.77 – 7.63 (m, 2H), 7.53 (dd, J = 4.8, 2.1 Hz, 2H), 7.23 (d, J = 9.5 Hz, 1H), 7.17 – 7.06 (m, 2H), 7.07 – 6.95 (m, 1H), 6.86 (d, J = 2.0 Hz, 1H), 4.42 (q, J = 7.2 Hz, 2H), 4.32 (qd, J = 7.2, 2.5 Hz, 2H), 1.46 (t, J = 7.2 Hz, 3H), 1.38 (td, J = 7.1, 3.3 Hz, 3H).

^{13}C NMR (126 MHz, $CDCl_3$): δ 163.27, 160.67, 130.36 (d, J = 9.6 Hz), 129.23 (d, J = 2.3 Hz), 129.03, 128.50, 127.87 (d, J = 6.5 Hz), 126.86, 122.76, 120.71 (d, J = 6.6 Hz), 117.55, 109.59 (t, J = 251.3 Hz), 95.38, 60.23, 44.55, 16.94, 14.42.

^{19}F NMR (282 MHz, $CDCl_3$): δ -90.59.

HRMS: (EI+) calculated for $[C_{19}H_{19}F_2N_3O_2+H]^+$ 360.1518, found 360.1518.

FTIR (ATR, cm^{-1}): 3119, 2978, 2929, 1699, 1446, 1379, 1323, 1271, 1237, 1066, 924, 823, 764, 697.



1-(5-(1-(difluoromethyl)-2-phenyl-1*H*-imidazol-5-yl)thiophen-2-yl)ethan-1-one (**F2**).

Prepared according to the general procedure B. The title compound was isolated via flash chromatography (gradient 0%-20% EtOAc/hexane) as a light yellow solid (515 mg, 1.62 mmol, 32% yield).

TLC (SiO_2) R_f = 0.10 in 4:1 hexanes/EtOAc.

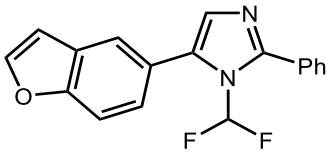
^{1}H NMR (500 MHz, $CDCl_3$): δ 7.67 (dd, J = 5.5, 3.6 Hz, 3H), 7.59 – 7.50 (m, 3H), 7.40 (s, 1H), 7.33 (d, J = 3.9 Hz, 1H), 7.29 – 6.99 (m, 1H), 2.58 (s, 3H).

^{13}C NMR (126 MHz, $CDCl_3$): δ 190.49, 150.27, 144.56, 137.29, 132.76, 131.45, 130.68, 129.23, 129.22, 128.32, 128.21 (t, J = 2.9 Hz), 125.49, 109.70 (t, J = 252.1 Hz), 26.72.

^{19}F NMR (282 MHz, $CDCl_3$): δ -89.76.

HRMS: (EI+) calculated for $[C_{16}H_{12}F_2N_2OS+H]^+$ 319.0711, found 319.0736.

FTIR (ATR, cm^{-1}): 3063, 1651, 1435, 1364, 1278, 1174, 1095, 1043, 972, 939, 812.



5-(Benzofuran-5-yl)-1-(difluoromethyl)-2-phenyl-1H-imidazole (F3).

Prepared according to the general procedure B. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 19 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 28% B, 28-68% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 20 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (456 mg, 1.47 mmol, 29% yield).

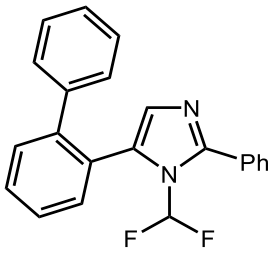
^{1}H NMR (400 MHz, DMSO-d6): δ 8.10 (d, *J* = 2.2 Hz, 1H), 7.83 (d, *J* = 1.7 Hz, 1H), 7.78 – 7.65 (m, 3H), 7.62 – 7.51 (m, 4H), 7.49 (dd, *J* = 8.5, 1.9 Hz, 1H), 7.29 (s, 1H), 7.07 (dd, *J* = 2.2, 1.0 Hz, 1H).

^{13}C NMR (101 MHz, DMSO-d6): δ 154.3, 148.1, 147.1, 133.4, 129.8, 129.7, 129.1, 128.8, 128.8, 127.7, 125.4, 123.5, 121.8, 111.7, 110.0 (t, *J* = 249.8 Hz) 106.9.

^{19}F NMR (376 MHz, DMSO-d6): δ -88.71, -88.86.

HRMS: (EI+) calculated for $[C_{18}H_{12}F_2N_2O+H]^+$ 311.0990, found 311.0999.

FTIR (ATR, cm^{-1}): 3026, 1461, 1367, 1308, 1244, 1185, 1129, 1095, 1043, 946, 890, 816, 771, 734, 700.



5-([1,1'-biphenyl]-2-yl)-1-(difluoromethyl)-2-phenyl-1H-imidazole (F4).

Prepared according to the general procedure B. The title compound was isolated via flash chromatography (gradient 0%-20% EtOAc/hexane) as a light yellow oil (799 mg, 2.31 mmol, 46% yield).

TLC (SiO_2) R_f = 0.25 in 4:1 hexanes/EtOAc

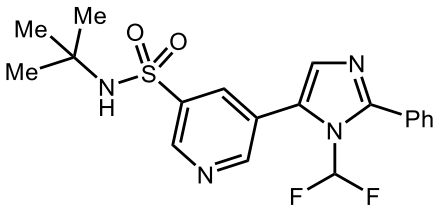
^{1}H NMR (500 MHz, $CDCl_3$): δ 7.66 – 7.56 (m, 3H), 7.56 – 7.49 (m, 3H), 7.49 – 7.39 (m, 3H), 7.35 – 7.28 (m, 3H), 7.25 – 7.21 (m, 2H), 6.93 (s, 1H), 6.73 (t, J = 58.6 Hz, 1H).

^{13}C NMR (126 MHz, $CDCl_3$): δ 147.95, 142.40, 140.18, 132.10, 131.95, 130.54, 130.08, 129.85, 129.83, 129.51, 129.25, 129.03, 128.60, 128.32, 127.51, 127.42, 126.61, 109.10 (t, J = 251.3 Hz).

^{19}F NMR (282 MHz, $CDCl_3$): δ -89.80

HRMS: (EI+) calculated for $[C_{22}H_{16}F_2N_2+H]^+$ 347.1354, found 347.1354.

FTIR (ATR, cm^{-1}): 3049, 1479, 1341, 1252, 1177, 1988, 1054, 943, 834, 738, 697.



N-(tert-butyl)-5-(1-(difluoromethyl)-2-phenyl-1H-imidazol-5-yl)pyridine-3-sulfonamide (F5).

Prepared according to the general procedure B. The title compound was isolated via flash chromatography (gradient 0%-50% EtOAc/hexane) as white solid (491 mg 1.21 mmol, 24% yield).

TLC (SiO_2) R_f = 0.43 in 1:1 hexanes/EtOAc

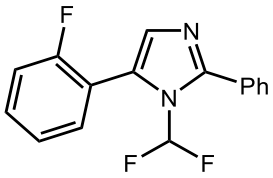
^{1}H NMR (500 MHz, $CDCl_3$): δ 9.13 (d, J = 2.2 Hz, 1H), 8.96 (d, J = 2.1 Hz, 1H), 8.41 (t, J = 2.1 Hz, 1H), 7.69 (dd, J = 6.7, 2.9 Hz, 2H), 7.63 – 7.45 (m, 3H), 7.40 (s, 1H), 7.12 (t, J = 58.5 Hz, 1H), 5.11 (s, 1H), 1.28 (s, 9H).

^{13}C NMR (126 MHz, $CDCl_3$): δ 151.82, 150.66, 147.62, 139.92, 134.64 (t, J = 2.4 Hz), 131.04, 130.98, 129.40, 129.30, 128.02, 127.77, 125.62, 110.04 (t, J = 252.1 Hz), 55.45, 30.16.

^{19}F NMR (282 MHz, $CDCl_3$): δ -88.11.

HRMS: (EI+) calculated for $[C_{19}H_{20}F_2N_4O_2S+H]^+$ 407.1348, found 407.1366.

FTIR (ATR, cm^{-1}): 3071, 2970, 2855, 1550, 1468, 1323, 1252, 1196, 1140, 1099, 1066, 1043, 1002, 831, 771, 704.



1-(difluoromethyl)-5-(2-fluorophenyl)-2-phenyl-1H-imidazole (F7).

Prepared according to the general procedure B (*2.5 mmol scale*). The title compound was isolated via flash chromatography (gradient 0%-20% EtOAc/hexane) as a beige solid (110 mg, 0.38 mmol, 15%).

TLC (SiO_2) R_f = 0.28 in 4:1 hexanes/EtOAc

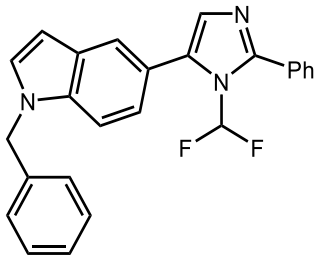
^{1}H NMR (500 MHz, $CDCl_3$): δ 7.75 (dd, J = 6.5, 2.9 Hz, 2H), 7.52 (hept, J = 2.8 Hz, 4H), 7.45 (dtt, J = 11.1, 5.5, 2.7 Hz, 1H), 7.37 – 7.14 (m, 3H), 6.99 (t, J = 58.6 Hz, 1H).

^{13}C NMR (126 MHz, $CDCl_3$): δ 161.04, 159.05, 149.29, 132.07 (d, J = 1.7 Hz), 131.13 (d, J = 8.1 Hz), 130.92, 130.20, 130.10, 129.37, 129.26, 128.87, 124.30 (d, J = 3.8 Hz), 116.02 (d, J = 21.9 Hz), 109.53 (t, J = 251.4 Hz).

^{19}F NMR (282 MHz, $CDCl_3$): δ -90.12 (d, J = 6.2 Hz), -112.70 (t, J = 6.3 Hz).

HRMS: (EI+) calculated for $[C_{16}H_{11}F_3N_2+H]^+$ 289.0947, found 289.0972.

FTIR (ATR, cm^{-1}): 3056, 1681, 1565, 1468, 1356, 1252, 1174, 1058, 1028, 827, 764, 700.



1-benzyl-5-(1-(difluoromethyl)-2-phenyl-1H-imidazol-5-yl)-1H-indole (F9).

Prepared according to the general procedure B (*2.5 mmol scale)*. The title compound was isolated via flash chromatography (gradient 0%-20% EtOAc/hexane) as a red oil (20 mg, 0.05 mmol, 2% yield).

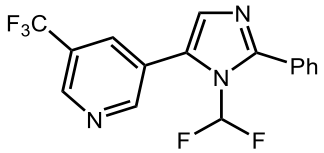
TLC (SiO_2) R_f = 0.20 in 4:1 hexanes/EtOAc

^{1}H NMR (500 MHz, $CDCl_3$): δ 7.89 – 7.73 (m, 3H), 7.66 – 7.57 (m, 1H), 7.59 – 7.46 (m, 3H), 7.42 – 7.36 (m, 1H), 7.37 – 7.27 (m, 3H), 7.25 – 7.19 (m, 2H), 7.18 – 7.12 (m, 2H), 7.00 (dd, J = 59.3, 17.4 Hz, 1H), 6.64 (dd, J = 3.2, 0.8 Hz, 1H), 5.37 (s, 2H).

^{13}C NMR (126 MHz, $CDCl_3$): δ 137.06, 136.43, 135.20, 130.22, 130.15, 129.91, 129.65, 129.16, 128.97, 128.91, 128.73, 127.87, 127.59, 126.83, 123.01, 122.09, 115.58, 110.19, 109.64 – 105.92 (m), 102.28, 50.35.

^{19}F NMR (282 MHz, $CDCl_3$): δ -89.21.

HRMS: (EI+) calculated for $[C_{25}H_{19}F_2N_3+H]^+$ 400.1620, found 400.1617.



3-(1-(difluoromethyl)-2-phenyl-1H-imidazol-5-yl)-5-(trifluoromethyl)pyridine (F10).

Prepared according to the general procedure B (*2.5 mmol scale*). The title compound was isolated via flash chromatography (gradient 0%-33% EtOAc/hexane) as a white solid (359 mg, 1.06 mmol, 42% yield).

TLC (SiO_2) R_f = 0.09 in 4:1 hexanes/EtOAc

^{1}H NMR (500 MHz, $CDCl_3$): δ 9.03 (d, J = 2.1 Hz, 1H), 8.99 – 8.87 (m, 1H), 8.16 (d, J = 2.2 Hz, 1H), 7.85 – 7.64 (m, 2H), 7.66 – 7.47 (m, 3H), 7.37 (s, 1H), 7.12 (t, J = 58.6 Hz, 1H).

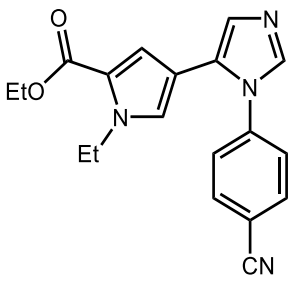
^{13}C NMR (126 MHz, $CDCl_3$): δ 152.33, 150.74, 146.35 (q, J = 4.0 Hz), 133.05 (d, J = 3.5 Hz), 131.43, 130.81, 129.36, 129.23, 128.17, 128.06, 125.70, 124.29, 122.12, 110.01 (t, J = 251.8 Hz).

^{19}F NMR (282 MHz, $CDCl_3$): δ -62.55, -82.22.

HRMS: (EI+) calculated for $[C_{16}H_{10}F_5N_3+H]^+$ 340.0868, found 340.0874.

FTIR (ATR, cm^{-1}): 3034, 1558, 1476, 1371, 1285, 1244, 1196, 1148, 1095, 1054, 834, 775, 700.

G-Series



Ethyl 4-(1-(4-cyanophenyl)-1H-imidazol-5-yl)-1-ethyl-1H-pyrrole-2-carboxylate (G1).

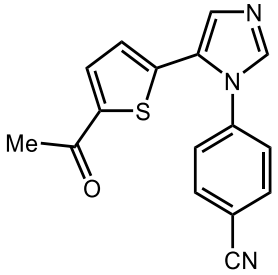
Prepared according to the general procedure A. The title compound was isolated via reversed phase chromatography (23 mg, 0.07 mmol, 1% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ 7.92 – 7.86 (m, 1H), 7.87 – 7.81 (m, 1H), 7.77 (s, 1H), 7.66 – 7.59 (m, 1H), 7.49 (d, *J* = 8.4 Hz, 1H), 7.36 (d, *J* = 1.8 Hz, 1H), 6.75 (d, *J* = 2.0 Hz, 1H), 6.66 (d, *J* = 2.0 Hz, 1H), 4.36 (dq, *J* = 10.4, 7.2 Hz, 4H), 1.42 (dt, *J* = 9.6, 7.2 Hz, 6H).

^{13}C NMR (101 MHz, DMSO-d6): δ 159.8, 140.2, 135.8, 134.2, 133.6, 130.6, 126.5, 120.4, 118.4, 117.7, 116.2, 110.9, 109.0, 59.6, 43.5, 16.9, 14.2.

HRMS: (EI+) calculated for $[C_{19}H_{18}N_4O_2+H]^+$ 335.1503, found 335.1508.

FTIR (ATR, cm^{-1}): 3116, 2989, 2225, 1699, 1606, 1513, 1371, 1300, 1263, 1226, 1185, 1099, 1054, 957, 902, 831, 730.



4-(5-(5-Acetylthiophen-2-yl)-1H-imidazol-1-yl)benzonitrile (G2).

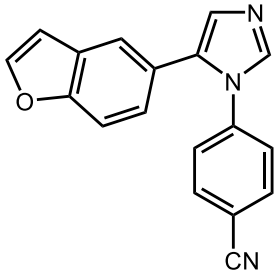
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 12% B, 12-52% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (197 mg, 0.66 mmol, 13% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.11 (s, 1H), 8.08 - 7.98 (m, 2H), 7.80 (d, *J*=4.0 Hz, 1H), 7.69 - 7.58 (m, 2H), 7.54 (s, 1H), 6.93 (d, *J*=3.9 Hz, 1H), 2.47 (s, 3H).

^{13}C NMR (101 MHz, DMSO-d6): δ 190.6, 143.3, 140.8, 139.2, 137.7, 134.3, 133.9, 131.0, 127.5, 127.2, 125.8, 118.0, 111.8, 26.3.

HRMS: (EI+) calculated for $[C_{16}H_{11}N_3OS+H]^+$ 294.0696, found 294.0698.

FTIR (ATR, cm^{-1}): 3101, 3049, 2228, 1651, 1561, 1505, 1464, 1435, 1271, 1218, 1114, 950, 834, 808.



4-(5-(Benzofuran-5-yl)-1H-imidazol-1-yl)benzonitrile (G3).

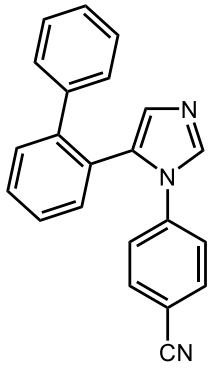
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 21% B, 21-61% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (344 mg, 1.21 mmol, 24% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.10 (s, 1H), 8.02 (d, *J*=2.1 Hz, 1H), 7.97 - 7.87 (m, 2H), 7.56 (d, *J*=8.6 Hz, 1H), 7.51 (s, 1H), 7.48 - 7.40 (m, 2H), 7.28 (s, 1H), 7.04 (dd, *J*=8.6, 1.5 Hz, 1H), 6.95 (s, 1H).

^{13}C NMR (101 MHz, DMSO-d6): δ 153.7, 146.9, 140.0, 139.1, 133.7, 132.3, 129.2, 127.6, 126.1, 124.9, 123.8, 121.2, 118.1, 111.6, 110.4, 106.8.

HRMS: (EI+) calculated for $[C_{18}H_{11}N_3O+H]^+$ 286.0975, found 286.0983.

FTIR (ATR, cm^{-1}): 3090, 1606, 1509, 1457, 1375, 1274, 1207, 1032, 920, 820, 775.



4-(5-([1,1'-Biphenyl]-2-yl)-1H-imidazol-1-yl)benzonitrile (G4).

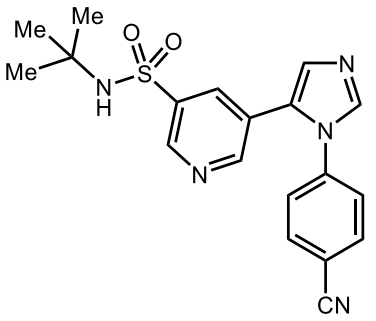
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 30% B, 30-70% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (267 mg, 0.83 mmol, 17% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.57 (s, 1H), 8.33 - 8.22 (m, 3H), 8.22 - 8.12 (m, 2H), 7.99 - 7.90 (m, 1H), 7.89 (s, 1H), 7.87 - 7.79 (m, 1H), 7.74 (t, *J*=7.5 Hz, 2H), 7.34 (d, *J*=8.4 Hz, 2H), 7.27 (d, *J*=7.5 Hz, 2H).

^{13}C NMR (101 MHz, DMSO-d6): δ 140.2, 139.7, 139.2, 137.8, 133.0, 131.4, 131.3, 130.4, 130.1, 129.4, 128.0, 128.0, 127.9, 127.1, 126.6, 123.9, 118.2, 109.2.

HRMS: (EI+) calculated for $[C_{22}H_{15}N_3+H]^+$ 322.1339, found 322.1346.

FTIR (ATR, cm^{-1}): 3104, 3063, 2228, 1602, 1505, 1271, 1107, 1069, 913, 842, 820, 767, 745, 700.



N-(tert-Butyl)-5-(1-(4-cyanophenyl)-1H-imidazol-5-yl)pyridine-3-sulfonamide (G5).

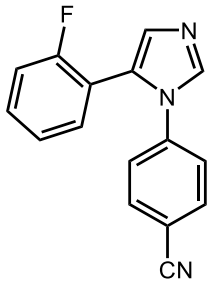
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 14% B, 14-54% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 45 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (251 mg, 0.66 mmol, 13% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.92 (d, *J*=2.0 Hz, 1H), 8.75 (d, *J*=1.8 Hz, 1H), 8.08 - 7.89 (m, 3H), 7.73 (s, 1H), 7.62 (d, *J*=8.4 Hz, 2H), 7.34 (s, 1H), 1.05 (s, 9H).

^{13}C NMR (101 MHz, DMSO-d6): δ 151.3, 146.4, 141.9, 141.0, 140.1, 134.0, 133.1, 130.0, 127.0, 126.3, 124.6, 118.0, 111.4, 53.7, 29.7.

HRMS: (EI+) calculated for $[C_{19}H_{19}N_5O_2S+H]^+$ 382.1332, found 382.1332.

FTIR (ATR, cm^{-1}): 3302, 3145, 3093, 3034, 2974, 2228, 1602, 1561, 1505, 1472, 1427, 1315, 1222, 1140, 1107, 1025, 991, 853, 764, 697, 667.



4-(5-(2-Fluorophenyl)-1H-imidazol-1-yl)benzonitrile (G7).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 19 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with 10-mM ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with 10-mM ammonium acetate; Gradient: a 0-minute hold at 15% B, 15-55% B over 20 minutes, then a 4-minute hold at 100% B; Flow Rate: 20 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (48 mg, 0.18 mmol, 4% yield).

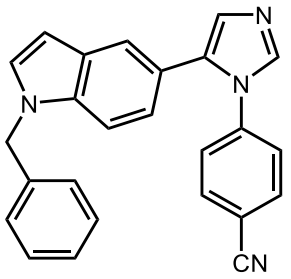
^{1}H NMR (400 MHz, DMSO-d6): δ 8.18 (s, 1H), 7.92 (d, *J*=8.4 Hz, 2H), 7.43 (br d, *J*=8.4 Hz, 3H), 7.38 - 7.32 (m, 1H), 7.30 (s, 1H), 7.28 - 7.21 (m, 1H), 7.21 - 7.13 (m, 1H).

^{13}C NMR (101 MHz, DMSO-d6): δ 159.0 (d, *J*=246.5 Hz), 140.6, 139.8, 134.2, 133.6, 131.9 (d, *J*=2.2 Hz), 131.3 (d, *J*=8.1 Hz), 126.5, 125.4 (br d, *J*=3.7 Hz), 118.5, 117.3 (d, *J*=14.7 Hz), 116.4 (d, *J*=21.3 Hz), 112.1, 110.9.

^{19}F NMR (376 MHz, DMSO-d6): δ -73.8.

HRMS: (EI+) calculated for $[C_{16}H_{10}FN_3+H]^+$ 264.0932, found 264.0945.

FTIR (ATR, cm^{-1}): 3071, 1658, 1602, 1505, 1446, 1274, 1196, 1133, 1107, 1062, 913, 827, 771, 723.



4-(5-(1-Benzyl-1H-indol-5-yl)-1H-imidazol-1-yl)benzonitrile (G9).

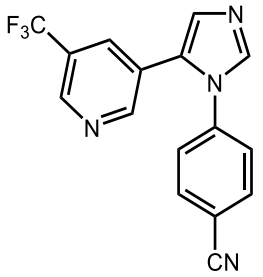
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 33% B, 33-73% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation. (358 mg, 0.96 mmol, 19% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.04 (s, 1H), 7.89 (d, *J*=8.6 Hz, 2H), 7.53 (d, *J*=3.2 Hz, 1H), 7.47 - 7.35 (m, 4H), 7.35 - 7.27 (m, 2H), 7.27 - 7.13 (m, 4H), 6.83 (d, *J*=8.6 Hz, 1H), 6.45 (d, *J*=3.1 Hz, 1H), 5.40 (s, 2H).

^{13}C NMR (101 MHz, DMSO-d6): δ 140.3, 138.6, 138.0, 135.2, 133.6, 133.3, 130.1, 128.6, 128.5, 128.3, 127.4, 127.1, 126.0, 121.9, 120.5, 119.8, 118.2, 110.4, 110.1, 101.4, 49.1.

HRMS: (EI+) calculated for $[C_{25}H_{18}N_4+H]^+$ 375.1604, found 375.1609.

FTIR (ATR, cm^{-1}): 3060, 2228, 1606, 1505, 1461, 1382, 1330, 1267, 1177, 1110, 920, 842, 801, 767, 726, 697.



4-(5-(5-(Trifluoromethyl)pyridin-3-yl)-1H-imidazol-1-yl)benzonitrile (G10).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 19 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with 0.05% trifluoroacetic acid; Mobile Phase B: 95:5 acetonitrile: water with 0.05% trifluoroacetic acid; Gradient: a 0-minute hold at 5% B, 5-45% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 20 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the

desired product were combined and dried via centrifugal evaporation (165 mg, 0.53 mmol, 11% yield).

^{1}H NMR (600 MHz, DMSO-d6): δ 8.98 (br d, 1H, *J* = 1.7 Hz), 8.76 (br d, 1H, *J* = 1.7 Hz), 8.08 (br t, 1H, *J* = 1.7 Hz), 8.02 (br d, 2H, *J* = 8.6 Hz), 7.86 (d, 1H, *J* = 1.4 Hz), 7.65 (br d, 2H, *J* = 8.6 Hz), 7.53 (d, 1H, *J* = 1.4 Hz).

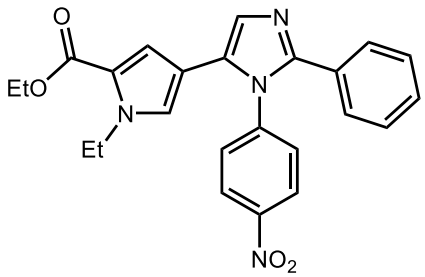
^{13}C NMR (150 MHz, DMSO-d6): δ 158.5 (q, *J* = 35.6 Hz), 152.6, 146.3 (q, *J* = 4.0 Hz), 141.5, 140.3, 134.0, 133.3 (q, *J* = 3.6 Hz), 128.2, 127.2, 124.9, 124.8, 123.2 (q, *J* = 272.7 Hz), 117.9, 111.8.

^{19}F NMR (376 MHz, DMSO) δ -61.37.

HRMS: (EI+) calculated for $[C_{16}H_9F_3N_4+H]^+$ 315.0852, found 315.0862.

FTIR (ATR, cm^{-1}): 3056, 2228, 1606, 1509, 1461, 1405, 1341, 1315, 1207, 1118, 1069, 1021, 909, 849, 760, 708.

I-Series



Ethyl 1-ethyl-4-(1-(4-nitrophenyl)-2-phenyl-1H-imidazol-5-yl)-1H-pyrrole-2-carboxylate (I1).

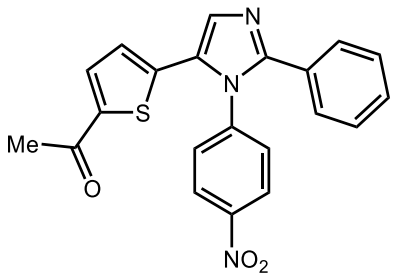
Prepared according to the general procedure A. The title compound was isolated via reversed phase chromatography (203 mg, 0.47 mmol, 9% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.35 - 8.30 (m, *J*=8.8 Hz, 2H), 7.63 - 7.58 (m, *J*=8.8 Hz, 2H), 7.32 - 7.26 (m, 6H), 6.83 (d, *J*=1.8 Hz, 1H), 6.49 (d, *J*=1.8 Hz, 1H), 4.22 - 4.11 (m, 4H), 1.18 (br t, *J*=7.0 Hz, 3H), 1.19 (br t, *J*=7.0 Hz, 3H).

^{13}C NMR (101 MHz, DMSO-d6): δ 159.7, 147.4, 146.4, 142.7, 130.2, 130.1, 129.7, 128.3, 128.2, 126.7, 124.8, 121.2, 116.1, 111.1, 59.6, 43.4, 16.8, 14.1.

HRMS: (EI+) calculated for $[C_{24}H_{22}N_4O_4+H]^+$ 431.1714, found 431.1734.

FTIR (ATR, cm^{-1}): 2987, 2937, 1701, 1597, 1522, 1498, 1481, 1466, 1410, 1379, 1349, 1326, 1312, 1287, 1246, 1216, 1194, 1170, 1123, 1099, 1075, 1011, 963, 929, 903, 855, 825.



1-(5-(1-(4-Nitrophenyl)-2-phenyl-1H-imidazol-5-yl)thiophen-2-yl)ethan-1-one (I2).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 26% B, 26-66% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature:

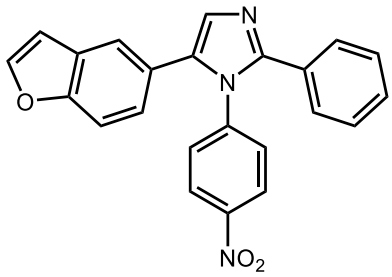
25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (92 mg, 0.24 mmol, 5% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.36 (d, *J*=8.8 Hz, 2H), 7.80 - 7.72 (m, 4H), 7.37 - 7.28 (m, 5H), 6.94 (d, *J*=3.9 Hz, 1H), 2.45 (s, 3H).

^{13}C NMR (101 MHz, DMSO-d6): δ 190.5, 148.9, 147.9, 143.1, 141.6, 137.9, 134.2, 130.7, 130.3, 129.4, 129.0, 128.5, 128.4, 126.9, 125.1, 26.3.

HRMS: (EI+) calculated for $[C_{21}H_{15}N_3O_3S+H]^+$ 390.0907, found 390.0907.

FTIR (ATR, cm^{-1}): 3108, 3067, 1662, 1595, 1565, 1520, 1442, 1349, 1233, 1177, 110, 1073, 965, 857, 801, 771, 689.



5-(Benzofuran-5-yl)-1-(4-nitrophenyl)-2-phenyl-1H-imidazole (I3).

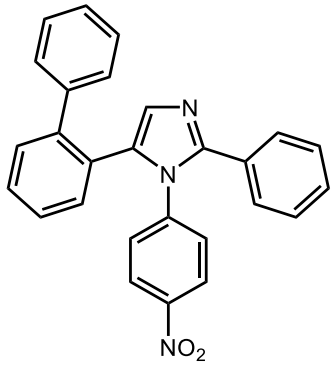
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 33% B, 33-73% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 45 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (380 mg, 1.00 mmol, 20% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.22 (d, *J*=8.8 Hz, 2H), 8.00 (d, *J*=2.1 Hz, 1H), 7.59 - 7.44 (m, 4H), 7.38 (s, 1H), 7.36 - 7.23 (m, 5H), 7.03 (dd, *J*=8.5, 1.4 Hz, 1H), 6.91 (s, 1H).

^{13}C NMR (101 MHz, DMSO-d6): δ 153.7, 147.2, 146.9, 146.9, 142.4, 135.1, 130.2, 130.0, 128.6, 128.5, 128.3, 128.3, 127.4, 125.4, 124.7, 123.9, 121.8, 111.3, 106.8.

HRMS: (EI+) calculated for $[C_{23}H_{15}N_3O_3+H]^+$ 382.1186, found 382.1188.

FTIR (ATR, cm^{-1}): 3090, 3063, 2929, 2855, 1595, 1520, 1464, 1394, 1237, 1189, 1136, 1021, 946, 857, 805, 775, 693.



5-([1,1'-Biphenyl]-2-yl)-1-(4-nitrophenyl)-2-phenyl-1H-imidazole (I4).

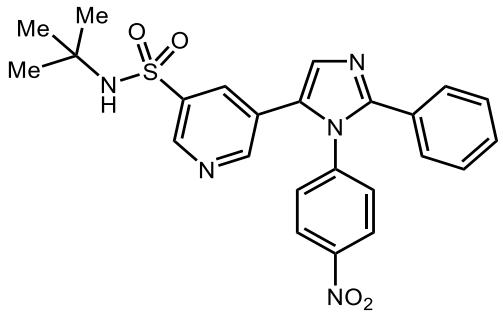
Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 0-minute hold at 41% B, 41-81% B over 20 minutes, then a 0-minute hold at 100% B; Flow Rate: 40 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (793 mg, 1.90 mmol, 38% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 7.86 (d, *J*=8.8 Hz, 2H), 7.62 - 7.52 (m, 1H), 7.52 - 7.40 (m, 2H), 7.37 - 7.21 (m, 5H), 7.21 - 7.01 (m, 5H), 6.76 (br d, *J*=7.3 Hz, 2H), 6.51 (br d, *J*=8.4 Hz, 2H).

^{13}C NMR (101 MHz, DMSO-d6): δ 146.5, 146.0, 141.4, 140.8, 140.0, 134.2, 132.5, 130.1, 130.0, 129.8, 129.8, 128.8, 128.7, 128.5, 128.4, 128.4, 128.0, 127.9, 127.2, 124.0, 45.3.

HRMS: (EI+) calculated for $[C_{27}H_{19}N_3O_2+H]^+$ 418.1550, found 418.1550.

FTIR (ATR, cm^{-1}): 3056, 1595, 1520, 1498, 1446, 1379, 1338, 1174, 110, 1073, 849, 767, 697.



N-(tert-Butyl)-5-(1-(4-nitrophenyl)-2-phenyl-1H-imidazol-5-yl)pyridine-3-sulfonamide (I5).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 200 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 10.5-minute hold at 43% B; Flow Rate: 80 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals.

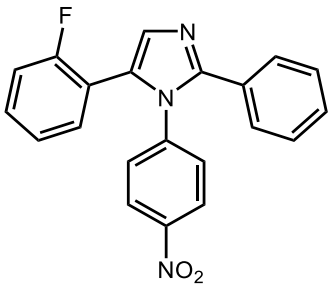
Fractions containing the desired product were combined and dried via centrifugal evaporation (1070 mg, 2.24 mmol, 45% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.86 (d, *J*=2.1 Hz, 1H), 8.78 (d, *J*=2.0 Hz, 1H), 8.28 (d, *J*=8.9 Hz, 2H), 7.76 (br s, 1H), 7.69 - 7.60 (m, 3H), 7.57 (t, *J*=2.0 Hz, 1H), 7.39 - 7.20 (m, 5H), 0.94 (s, 9H).

^{13}C NMR (101 MHz, DMSO-d6): δ 151.5, 148.7, 147.3, 145.9, 141.6, 140.0, 132.5, 130.3, 130.1, 130.0, 129.6, 128.9, 128.7, 128.4, 125.4, 125.1, 53.5, 29.6.

HRMS: (EI+) calculated for $[C_{24}H_{23}N_5O_4S+H]^+$ 478.1544, found 478.1530.

FTIR (ATR, cm^{-1}): 3265, 3067, 2974, 2870, 1595, 1520, 1349, 1308, 1148, 1107, 1025, 857, 767, 689.



5-(2-Fluorophenyl)-1-(4-nitrophenyl)-2-phenyl-1H-imidazole (I7).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge Phenyl, 250 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 12.5-minute hold at 48% B; Flow Rate: 80 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals.

Fractions containing the desired product were combined and dried via centrifugal evaporation (615 mg, 1.71 mmol, 34% yield)

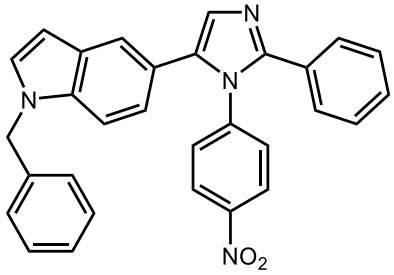
^{1}H NMR (400 MHz, DMSO-d6): δ 8.20 (d, *J*=8.9 Hz, 2H), 7.46 - 7.27 (m, 10H), 7.23 - 7.12 (m, 2H).

^{13}C NMR (101 MHz, DMSO-d6): δ 160.4, 157.9, 147.6, 146.9, 142.1, 132.1, 132.1, 131.0, 131.0, 129.9, 129.2, 128.7, 128.7, 128.6, 128.4, 124.6, 124.6, 116.9, 116.7, 115.8, 115.6. [We could not easily identify doublets from C-F couplings]

^{19}F NMR (376 MHz, DMSO-d6): δ 112.55.

HRMS: (EI+) calculated for $[C_{21}H_{14}FN_3O_2+H]^+$ 360.1143, found 360.1148.

FTIR (ATR, cm^{-1}): 3067, 1595, 1520, 1479, 1390, 1349, 1222, 1174, 1110, 1077, 1028, 857, 760, 704.



1-Benzyl-5-(1-(4-nitrophenyl)-2-phenyl-1H-imidazol-5-yl)-1H-indole (I9).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 250 mm x 30 mm, 5-µm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 11.5-minute hold at 60% B; Flow Rate: 80 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals.

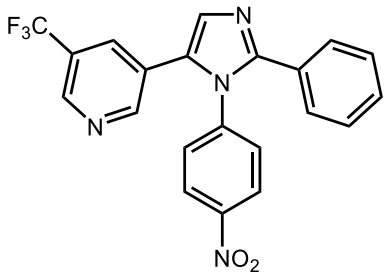
Fractions containing the desired product were combined and dried via centrifugal evaporation (847 mg, 1.80 mmol, 36% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.20 (d, *J*=8.8 Hz, 2H), 7.58 - 7.43 (m, 3H), 7.38 (s, 1H), 7.36 (d, *J*=8.6 Hz, 1H), 7.34 - 7.13 (m, 11H), 6.83 (d, *J*=8.6 Hz, 1H), 6.42 (d, *J*=3.1 Hz, 1H), 5.37 (s, 2H).

^{13}C NMR (101 MHz, DMSO-d6): δ 146.8, 146.7, 142.7, 138.0, 136.2, 135.1, 130.3, 130.0, 130.0, 128.6, 128.5, 128.3, 128.3, 128.2, 127.7, 127.4, 127.1, 124.6, 122.3, 121.1, 119.9, 110.2, 101.4, 49.1.

HRMS: (EI+) calculated for $[C_{30}H_{22}N_4O_2+H]^+$ 471.1816, found 471.1804.

FTIR (ATR, cm^{-1}): 3063, 1595, 1520, 1476, 1345, 1267, 1185, 1095, 946, 857, 801, 775, 700.



3-(1-(4-Nitrophenyl)-2-phenyl-1H-imidazol-5-yl)-5-(trifluoromethyl)pyridine (I10).

Prepared according to the general procedure A. The crude material was purified via preparative LC/MS with the following conditions: Column: XBridge C18, 250 mm x 19 mm, 5-μm particles; Mobile Phase A: 5:95 acetonitrile: water with ammonium acetate; Mobile Phase B: 95:5 acetonitrile: water with ammonium acetate; Gradient: a 11.5-minute hold at 43% B; Flow

Rate: 80 mL/min; Column Temperature: 25 °C. Fraction collection was triggered by MS signals. Fractions containing the desired product were combined and dried via centrifugal evaporation (694 mg, 1.69 mmol, 34% yield).

^{1}H NMR (400 MHz, DMSO-d6): δ 8.86 (s, 1H), 8.63 (s, 1H), 8.38 - 8.23 (m, *J*=8.8 Hz, 2H), 7.85 (s, 1H), 7.72 (s, 1H), 7.67 - 7.55 (m, *J*=8.8 Hz, 2H), 7.42 - 7.20 (m, 5H).

^{13}C NMR (101 MHz, DMSO-d6): δ 152.7, 149.2, 147.8, 145.3 (q, *J*=4.4 Hz), 142.1, 132.9 (q, *J*=3.7 Hz), 130.9, 130.7, 130.6, 130.1, 129.4, 129.2, 128.9, 126.0, 125.5, 125.4, 125.2, 123.7 (q, *J*=272.9 Hz).

^{19}F NMR (376 MHz, DMSO-d6): δ -61.3.

HRMS: (EI+) calculated for $[C_{21}H_{13}F_3N_4O_2+H]^+$ 411.1063, found 411.1062.

FTIR (ATR, cm^{-1}): 3112, 1595, 1524, 1464, 1386, 1345, 1308, 1129, 1092, 909, 857, 775, 697.

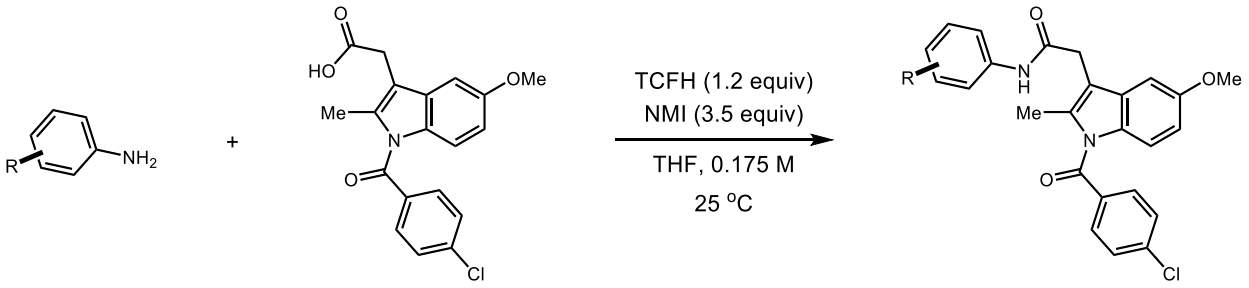
2.5.2 Amidation dataset experimentation details

High-throughput experimentation procedure

To 1 mL vials were added solid amide coupling reagents (26 μmol, 1.3 equiv.) by an automated solid dispensing robot. To the remaining 1 mL vials was added diphenylphosphinic chloride (5 μL, 6.2 mg, 26 μmol, 1.3 equiv.). To each vial was added a 0.2 M stock solution of indomethacin (7.2 mg, 20 μmol, 1.0 equiv.) in the desired reaction solvent (100 μL). The reactions were stirred for 5 minutes then treated with organic bases (60 μmol, 3.0 equiv.) and stirred for 30 minutes. The reaction mixtures were then treated with a 0.3 M stock solution of desired amine (30 μmol, 1.5 equiv.) in the desired reaction solvent (100 μL). The reactions stirred overnight at 25 °C

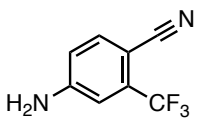
and were diluted with a 0.1 M solution of (4,4')-di-*t*-butylbiphenyl in dimethylformamide (600 µL) and stirred for 5 min. A 10 µL sample of each reaction was diluted into 500 µL 80% acetonitrile/water, filtered, and submitted for UPLCMS analysis.

Authentic product synthesis: procedure and characterization

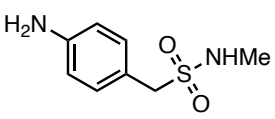


General Procedure: To a 40 mL vial containing indomethacin (1.0 g, 2.8 mmol, 1.0 equiv.) was added TCFH (0.95 g, 3.4 mmol, 1.2 equiv.) followed by THF (16 mL, 0.175 M). The reaction mixture was treated with 1-methylimidazole (0.79 mL, 9.8 mmol, 3.5 equiv.) and stirred for 20 min. To the reaction mixture was added the desired aryl amine (3.6 mmol, 1.3 equiv.). The reaction was stirred for 24 h and concentrated. The resulting crude was crystallized from 3:1 isopropanol/water unless otherwise specified.

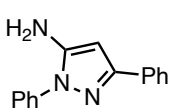
Aniline Substrate scope: Amide product labels (amide #x) corresponds to aniline labels (nx).



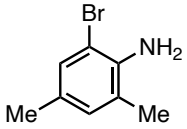
n1



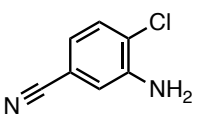
n2



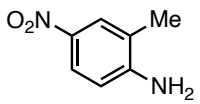
n3



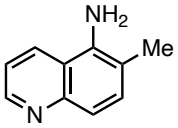
n4



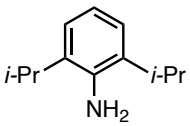
n5



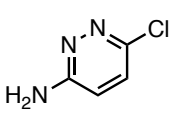
n6



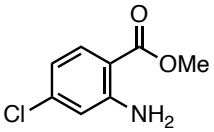
n7



n8

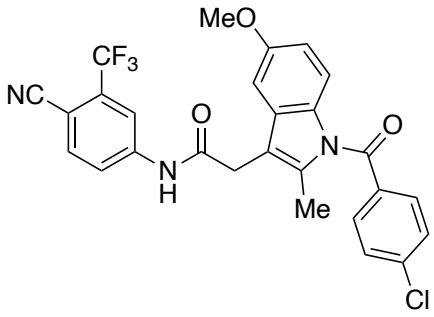


n9



n10

Amide products synthesis and characterization



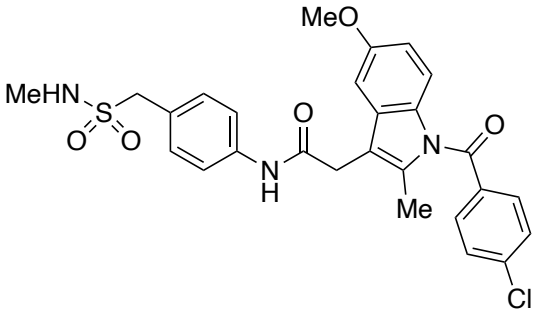
2-(1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1H-indol-3-yl)-N-(4-cyano-3-(trifluoromethyl)phenyl)acetamide (amide #1).

Prepared according to the general procedure. Isolated the product as white solids (641 mg, 1.22 mmol) in 44% yield.

^{1}H NMR (500 MHz, DMSO-d_6): δ 10.98 (s, 1H), 8.30 (d, *J*=2.0 Hz, 1H), 8.09 (d, *J*=8.5 Hz, 1H), 8.00 (dd, *J*=8.5, 2.0 Hz, 1H), 7.70 - 7.62 (m, 4H), 7.14 (d, *J*=2.4 Hz, 1H), 6.93 (d, *J*=8.9 Hz, 1H), 6.72 (dd, *J*=9.0, 2.6 Hz, 1H), 4.34 (br s, 1H), 3.85 (s, 2H), 3.78 (s, 1H), 3.78 - 3.69 (m, 4H), 2.28 (s, 3H), 1.04 (d, J=6.1 Hz, 8H).

^{13}C NMR (126 MHz, DMSO-d_6): δ 169.9, 167.9, 155.6, 143.6, 137.7, 136.5, 135.7, 134.1, 131.6, 131.2, 130.7, 130.2, 129.1, 122.0 (q, *J*=273.8 Hz, 1C), 122.0, 116.4 (q, *J*=4.2 Hz, 1C), 115.7, 114.6, 113.1, 111.2, 101.9, 101.6, 55.4, 32.1, 13.3

HRMS (ESI-TOF): calculated for $C_{27}H_{18}ClF_3N_3O_3$ ([M-H]$^-$): 524.0994, found: 524.0997.



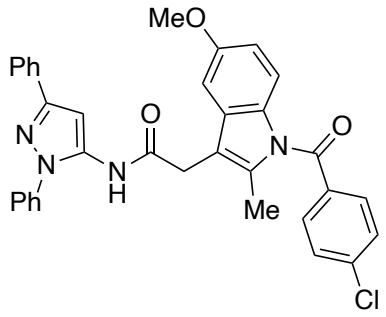
2-(1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1H-indol-3-yl)-N-(4-((N-methylsulfamoyl)methyl)phenyl)acetamide (amide #2).

Prepared according to the general procedure. Isolated the product as yellow solids (469 mg, 0.87 mmol) in 31% yield.

^{1}H NMR (500 MHz, DMSO-d_6): δ 10.30 (s, 1H), 7.72 - 7.67 (m, 2H), 7.67 - 7.63 (m, 2H), 7.63 - 7.59 (m, 2H), 7.38 - 7.25 (m, *J*=8.7Hz, 2H), 7.20 (d, *J*=2.6 Hz, 1H), 6.94 (d, *J*=9.0 Hz, 1H), 6.85 (q, *J*=4.8 Hz, 1H), 6.72 (dd, *J*=9.0, 2.4 Hz, 1H), 4.25 (s, 2H), 3.80 - 3.72 (m, 5H), 2.60 - 2.52 (m, 3H), 2.29 (s, 3H).

^{13}C NMR (126 MHz, DMSO-d_6): δ 168.5, 167.8, 155.5, 138.9, 137.6, 135.4, 134.2, 131.1, 131.1, 130.9, 130.2, 129.0, 125.1, 119.0, 114.5, 114.1, 111.1, 102.0, 55.4, 55.4, 32.0, 28.8, 13.4.

HRMS (ESI-TOF): calculated for $C_{27}H_{25}ClN_3O_5S$ ([M-H]$^-$): 538.1209, found: 538.1210.



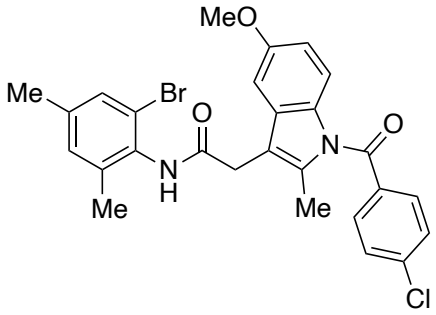
2-(1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1*H*-indol-3-yl)-*N*-(1,3-diphenyl-1*H*-pyrazol-5-yl)acetamide (amide #3).

Prepared according to the general procedure. Isolated the product as white solids (955 mg, 1.66 mmol) in 59% yield.

^{1}H NMR (500 MHz, DMSO-d$_6$): δ 10.10 (s, 1H), 7.86 (d, *J*=7.2 Hz, 2H), 7.72 - 7.56 (m, 4H), 7.51 (d, *J*=6.2 Hz, 2H), 7.43 (t, *J*=7.2 Hz, 2H), 7.39 - 7.29 (m, 4H), 7.12 (d, *J*=2.4 Hz, 1H), 6.97 (d, *J*=9.0 Hz, 1H), 6.91 (s, 1H), 6.75 (dd, *J*=9.0, 2.4 Hz, 1H), 3.83 - 3.70 (m, 5H), 2.21 (s, 3H).

^{13}C NMR (126 MHz, DMSO-d$_6$): δ 169.2, 167.8, 155.6, 150.1, 138.4, 137.6, 137.1, 135.6, 134.1, 132.7, 131.1, 130.7, 130.3, 129.0, 129.0, 128.7, 128.0, 127.4, 125.1, 123.4, 114.5, 113.2, 111.3, 101.9, 100.2, 55.4, 31.0, 13.3.

HRMS (ESI-TOF): calculated for $C_{34}H_{26}ClN_4O_3$ ([M-H]$^-$): 537.1699, found: 537.1705.

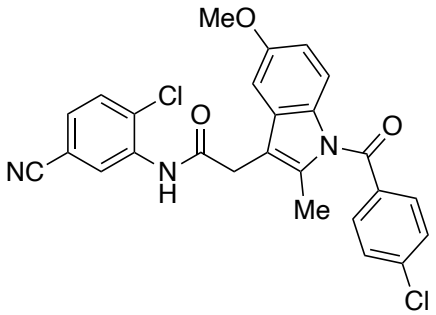


N-(2-bromo-4,6-dimethylphenyl)-2-(1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1H-indol-3-yl)acetamide (amide #4): Prepared according to the general procedure. Isolated the product as off-white solids (1080 mg, 2.00 mmol) in 72% yield.

^{1}H NMR (500 MHz, DMSO-d$_6$): δ 9.53 (s, 1H), 7.74 - 7.67 (m, 2H), 7.67 - 7.61 (m, 2H), 7.32 (s, 1H), 7.22 (d, *J*=2.6 Hz, 1H), 7.05 (s, 1H), 6.99 (d, *J*=9.0 Hz, 1H), 6.73 (dd, *J*=8.9, 2.5 Hz, 1H), 3.77 (s, 3H), 3.74 (s, 2H), 2.29 (s, 3H), 2.25 (s, 3H), 2.08 (s, 3H).

^{13}C NMR (126 MHz, DMSO-d$_6$): δ 168.2, 167.9, 155.5, 138.1, 138.1, 137.9, 137.5, 135.4, 134.3, 132.6, 131.1, 130.9, 130.3, 130.2, 129.0, 122.6, 114.5, 113.9, 111.5, 101.8, 55.4, 30.9, 20.1, 18.4, 13.5.

HRMS (ESI-TOF): calculated for $C_{27}H_{23}BrClN_2O_3$ ($[M+H]^+$): 539.0732, found: 539.0736.



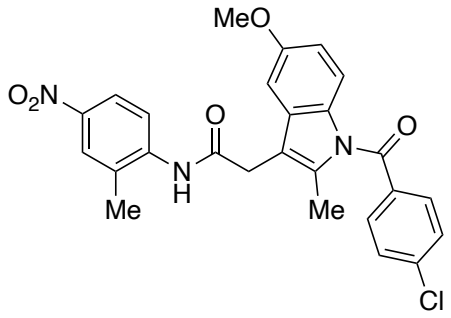
N-(2-chloro-5-cyanophenyl)-2-(1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1H-indol-3-yl)acetamide (amide #5).

Prepared according to the general procedure. Isolated the product as white solids (438 mg, 0.89 mmol) in 32% yield.

^{1}H NMR (500 MHz, DMSO-d$_6$/THF-d$_8$): δ 9.80 (s, 1H), 8.37 (d, *J*=1.8 Hz, 1H), 7.76 - 7.68 (m, 3H), 7.67 - 7.57 (m, 3H), 7.28 (d, *J*=2.4 Hz, 1H), 6.98 (d, *J*=9.0 Hz, 1H), 6.69 (dd, *J*=9.0, 2.4 Hz, 1H), 3.97 (s, 2H), 3.81 (s, 3H), 2.37 (s, 3H).

^{13}C NMR (126 MHz, DMSO-d$_6$/THF-d$_8$): δ 169.5, 168.1, 156.2, 138.2, 136.6, 135.9, 134.7, 131.5, 131.2, 131.1, 130.9, 130.0, 129.2, 129.0, 127.9, 117.9, 114.7, 113.9, 111.5, 110.9, 102.0, 55.5, 31.8, 13.3.

HRMS (ESI-TOF): calculated for $C_{26}H_{18}Cl_2N_3O_3$ ([M-H]$^-$): 490.0731, found: 490.0737.



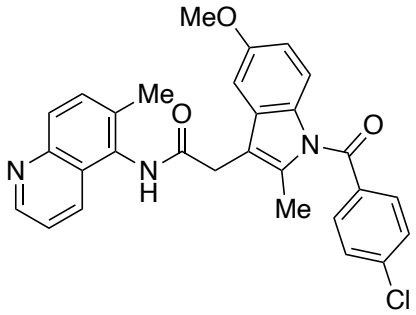
2-(1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1H-indol-3-yl)-N-(2-methyl-4-nitrophenyl)acetamide (amide #6).

Prepared according to the general procedure. Isolated the product as off-white solids (858 mg, 1.74 mmol) in 62% yield.

^{1}H NMR (500 MHz, DMSO-d_6/THF-d_8): δ 9.70 (s, 1H), 8.13 (s, 1H), 8.04 (d, *J*=1.4 Hz, 2H), 7.75 - 7.70 (m, 2H), 7.66 - 7.60 (m, 2H), 7.29 (d, *J*=2.4 Hz, 1H), 6.98 (d, *J*=8.9 Hz, 1H), 6.69 (dd, *J*=9.0, 2.6 Hz, 1H), 3.96 (s, 2H), 3.88 - 3.82 (m, 1H), 3.82 - 3.77 (m, 3H), 2.40 (s, 3H), 2.37 (s, 3H).

^{13}C NMR (126 MHz, DMSO-d_6/THF-d_8): δ 169.0, 167.7, 155.8, 143.3, 143.1, 137.9, 135.5, 134.3, 131.1, 131.0, 130.9, 130.5, 128.9, 125.2, 123.0, 121.4, 114.4, 113.8, 111.1, 101.7, 55.2, 31.7, 17.7, 13.1.

HRMS (ESI-TOF): calculated for $C_{26}H_{21}ClN_3O_5$ ([M-H]$^-$): 490.1175, found: 490.1183.



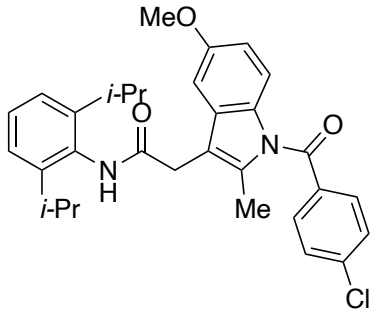
2-(1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1H-indol-3-yl)-N-(6-methylquinolin-5-yl)acetamide (amide #7): Prepared according to the general procedure and isolated by crystallization of the crude from 1:1 heptane/ethyl acetate. Isolated the product as white solids (1180 mg, 2.37 mmol) in 85% yield.

^{1}H NMR (500 MHz, DMSO): δ 9.88 (s, 1H), 8.84 (dd, *J* = 4.1, 1.7 Hz, 1H), 8.23 (ddd, *J* = 8.5, 1.7, 0.9 Hz, 1H), 7.88 (d, *J* = 8.6 Hz, 1H), 7.74 – 7.70 (m, 2H), 7.67 – 7.62 (m, 3H), 7.48

(dd, J = 8.5, 4.2 Hz, 1H), 7.32 (d, J = 2.6 Hz, 1H), 7.02 (d, J = 8.9 Hz, 1H), 6.76 (dd, J = 9.0, 2.6 Hz, 1H), 3.94 (s, 2H), 3.80 (s, 3H), 2.35 (s, 3H), 2.30 (s, 3H).

^{13}C NMR (126 MHz, DMSO): δ 169.2, 167.9, 155.7, 149.6, 146.9, 137.6, 135.6, 134.3, 133.2, 132.0, 131.6, 131.4, 131.2, 130.9, 130.4, 129.0, 127.6, 125.5, 121.3, 114.7, 114.0, 111.5, 101.7, 55.5, 31.1, 18.1, 13.5.

HRMS (ESI-TOF): calculated for $C_{29}H_{23}ClN_3O_3$ ([M-H]$^-$): 496.1433, found: 496.1447.

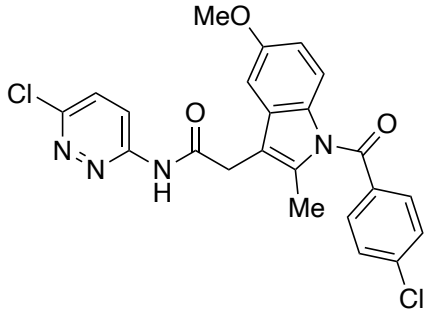


2-(1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1H-indol-3-yl)-N-(2,6-diisopropylphenyl)acetamide (amide #8): Prepared according to the general procedure. Isolated the product as white solids (780 mg, 1.51 mmol) in 54% yield.

^{1}H NMR (500 MHz, DMSO): δ 9.20 (s, 1H), 7.72 – 7.68 (m, 2H), 7.66 – 7.61 (m, 2H), 7.27 (d, J = 2.5 Hz, 1H), 7.22 (dd, J = 8.1, 7.2 Hz, 1H), 7.11 (d, J = 7.7 Hz, 2H), 7.03 (d, J = 9.0 Hz, 1H), 6.75 (dd, J = 9.0, 2.6 Hz, 1H), 3.79 (s, 3H), 3.75 (s, 2H), 3.01 (p, J = 6.9 Hz, 2H), 2.29 (s, 3H), 1.03 (br s, 12H).

^{13}C NMR (126 MHz, DMSO): δ 169.3, 167.9, 155.6, 146.0, 137.5, 135.3, 134.3, 132.6, 131.1, 130.9, 130.4, 129.0, 127.5, 122.7, 114.5, 114.2, 111.4, 101.8, 55.4, 31.2, 27.9, 23.4, 13.4

HRMS (ESI-TOF): calculated for $C_{31}H_{32}ClN_2O_3$ ([M-H]$^-$): 515.2107, found: 515.2100.

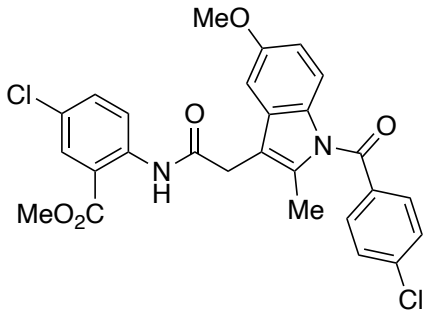


2-(1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1H-indol-3-yl)-N-(6-chloropyridazin-3-yl)acetamide (amide #9): Prepared according to the general procedure and isolated from the crude by crystallization from 1:1 heptane/ethyl acetate. Isolated the product as white solids (900 mg, 1.92 mmol) in 69% yield.

^{1}H NMR (500 MHz, DMSO-d$_6$): δ 11.63 (s, 1H), 8.36 (d, *J*=9.3 Hz, 1H), 7.86 (d, *J*=9.5 Hz, 1H), 7.74 - 7.62 (m, 4H), 7.22 (d, *J*=2.6 Hz, 1H), 6.93 (d, *J*=9.0 Hz, 1H), 6.72 (dd, *J*=9.0, 2.6 Hz, 1H), 3.93 (s, 2H), 3.77 - 3.73 (m, 3H), 2.30 (s, 3H).

^{13}C NMR (126 MHz, DMSO-d$_6$): δ 170.3, 167.8, 155.6, 155.2, 151.1, 137.6, 135.7, 134.1, 131.1, 130.8, 130.2, 130.2, 129.0, 121.4, 114.5, 113.3, 111.2, 101.9, 55.4, 31.6, 13.4.

HRMS (ESI-TOF): calculated for $C_{23}H_{17}Cl_2N_4O_3$ ([M-H]$^-$): 467.0683, found: 467.0701.



Methyl 5-chloro-2-(2-(1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1H-indol-3-yl)acetamido)benzoate (amide #10).

Prepared according to the general procedure. Isolated the product as yellow solids (380 mg, 0.72 mmol) in 26% yield.

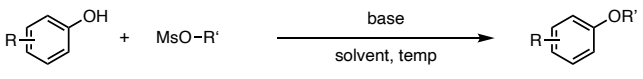
^{1}H NMR (500 MHz, DMSO-d_6/THF-d_8): δ 10.79 (s, 1H), 8.65 (d, *J*=9.2 Hz, 1H), 7.87 (d, *J*=2.6 Hz, 1H), 7.83 (d, *J*=7.6 Hz, 2H), 7.69 - 7.63 (m, 3H), 7.10 (d, *J*=2.6 Hz, 1H), 6.87 (d, *J*=9.0 Hz, 1H), 6.68 (dd, *J*=9.0, 2.6 Hz, 1H), 3.91 (s, 2H), 3.76 (s, 3H), 3.75 (s, 3H), 2.39 (s, 3H).

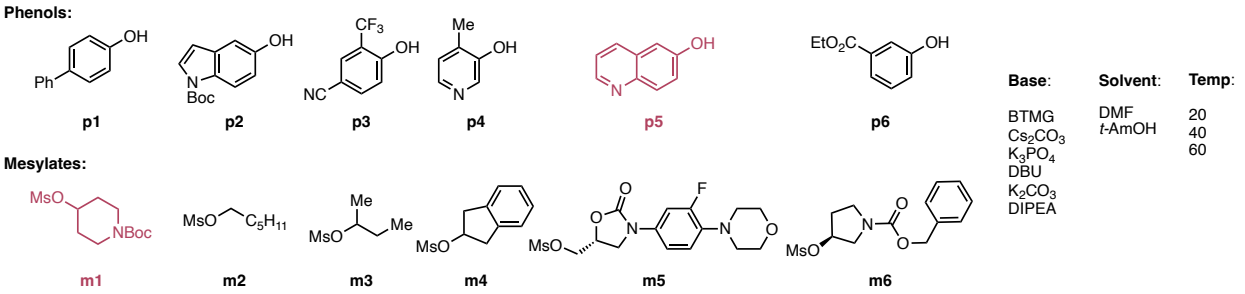
^{13}C NMR (126 MHz, DMSO-d_6/THF-d_8): δ 169.0, 167.8, 166.5, 155.8, 139.2, 138.0, 136.6, 134.1, 133.8, 131.3, 130.6, 130.4, 129.7, 128.9, 126.5, 121.8, 117.4, 114.6, 112.0, 111.4, 101.1, 55.1, 52.4, 33.0, 12.7.

HRMS (ESI-TOF): calculated for $C_{27}H_{21}Cl_2N_2O_5$ ($[M-H]^-$): 523.0833, found: 523.0830.

2.5.3 Phenol alkylation reaction condition optimization and experimentation details

Optimization substrate and condition scope.





We selected 6 phenols and 6 mesylates as the substrate scope. These starting materials are all commercially available and cheap to acquire. The other criteria of selection are structural diversity and synthesizability of the authentic product standards. One phenol (**p5**) and one phenol (**m1**) were randomly left out as external test substrates and are not included in the optimization. Six bases (inorganic and organic), two solvents, and three temperatures commonly investigated in similar reactions are also defined. Overall, 900 experiments (25 substrate pairings, 36 conditions) are available to sample from. The remaining 11 substrate pairings are tested with the algorithm-optimized conditions and benchmark conditions identified from historical dataset at BMS (*vide infra*)

Reaction set up procedure

Stock Solution Preparation. On the benchtop, 4-mL vials containing 400 µmol of mesylate were dissolved in *N*,*N*-dimethylformamide (DMF, 800 µL) or *tert*-amyl alcohol (800 µL) and stirred for 5 min. Similarly, 4-mL vials containing 400 µmol of phenol were dissolved in *N*,*N*-dimethylformamide (DMF, 800 µL) or *tert*-amyl alcohol (800 µL) and stirred for 5 min. The vials were sealed and stored in the freezer until time of use.

Base Plate Preparation. In a glovebox, cesium carbonate (39.1 mg, 0.12 mmol)/well, potassium carbonate (16.6 mg, 0.12 mmol)/well, or potassium phosphate tribasic (25.5. mg, 0.12

mmol)/well was dispensed to each well in a 96 well plate using CHRONECT XPR Solid Dispensing instrument immediately prior to use.

Reaction Execution. In the fume hood, 80 μL of phenol stock solution was dispensed to the appropriate well. Subsequently, liquid bases, including 2-*tert*-butyl-1,1,3,3-tetramethylguadinine (BTMG, 24.2 μL, 0.12 mmol)/well, *N*,*N*-diisopropylethylamine (DIPEA, 20.9 μL, 0.12 mmol)/well, or DBU (17.9 μL, 0.12 mmol)/well was added. The resultant reaction mixtures were sealed and stirred on a shaker block for 5 min. Then, 80 μL of mesylate stock solution was dispensed to the reaction mixture vials. The reactions were then sealed and stirred at the desired temperature (20, 40, or 60 °C) for 20 h in the fume hood. The plate was removed from the shaker, cooled to room temperature, and diluted to an 800 μL total volume with DMF containing 4,4'-di-*tert*-butylbiphenyl (1.07 mg, 4 μmol). The plate was stirred for 5 min and a 20 μL sample was taken and filtered into a UPLC analysis plate. The filter was rinsed with 500 μL acetonitrile/water (4:1) solution and analyzed by UPLC-MS.

Calibration Curve. A solution of 4,4'-di-*tert*-butylbiphenyl (19.18 mg) in 700 μL of DMF was prepared as the internal standard stock solution. 8 μL of the internal standard stock solution was dispensed into five 4-mL vials followed by addition of 3900 μL of DMF. A solution of the product marker (30 μmol) in 300 μL of DMF was prepared. The following volumes of product marker solution (16, 32, 48, 64, and 80 μL) were dispensed into the 4-mL vials to generate solutions that contain 20%, 40%, 60%, 80%, and 100% of the original product marker vs the consistent amount of 4,4'-di-*tert*-butylbiphenyl internal standard. The samples were transferred to UPLC vials for analysis.

UPLC-MS Method.

Solvent A: Water with 5% acetonitrile and 0.05% TFA

Solvent B: Acetonitrile with 5% water and 0.05% TFA

Gradient: 95% A/B to 0% A/B over 1.2 min, hold 0.8 min at 100% B, 0% A/B to 95% A/B over 0.01 min, hold 0.99 min at 95% A/B, 95% A/B to 0% A/B over 0.1 min

Stop Time: 2.0 min

Flow Rate: 0.8 mL/min

Wavelength1: 220 nm

Wavelength2: 254 nm

Column: Agilent Poroshell C18 2.7 µm 2.1x50 mm

Experimental history and data analysis

Four rounds of optimization were conducted, with batch sizes of 36, 18, 18 and 18. UCB1-Tuned was used as the bandit algorithm due to its generally high performance and the lack of tunable parameters. A random forest model is used as the prediction model for batch proposal (*vide supra*). Mesylate and phenol substrates were encoded with Morgan fingerprints using RDKit's default settings (radius=2, 2048 bits). All other parameters except temperature (used directly as normalized values) are one-hot encoded. A Jupyter notebook for all the optimization and analysis for this reaction is included in the GitHub repository.

Since UCB1-Tuned (or any UCB algorithms) requires a uniform exploration of every arm once, the first 36 experiments are therefore proposed sequentially to sample every condition. This is the normal behavior of the algorithm. After the first round, we used a smaller batch size of 18,

which helps the algorithm to converge faster. We also planned to explore 10% of the scope (90 out of 900 reactions) in total, so a batch size of 18 allows us to do exactly three rounds of experiments after the initial round. All 90 experiments, their parameters and the experimental and predicted yields are listed in **Table 3**.

Experiment	Round	Base	Mesylate ID	Phenol ID	Solvent	Temperature	Yield	Predicted Yield
1	1	BTMG	m2	p2	DMF	t20	0.0649	1
2	1	BTMG	m6	p6	DMF	t40	0.1944	1
3	1	BTMG	m5	p6	DMF	t60	0.1435	1
4	1	BTMG	m5	p2	tAmOH	t20	0.0000	1
5	1	BTMG	m2	p3	tAmOH	t40	0.0457	1
6	1	BTMG	m3	p6	tAmOH	t60	0.4584	1
7	1	Cs2CO3	m3	p2	DMF	t20	0.0220	1
8	1	Cs2CO3	m2	p2	DMF	t40	0.1100	1
9	1	Cs2CO3	m6	p3	DMF	t60	0.0050	1
10	1	Cs2CO3	m3	p6	tAmOH	t20	0.0440	1
11	1	Cs2CO3	m5	p4	tAmOH	t40	0.0000	1
12	1	Cs2CO3	m4	p2	tAmOH	t60	0.0467	1
13	1	K3PO4	m3	p1	DMF	t20	0.0104	1
14	1	K3PO4	m3	p4	DMF	t40	0.0561	1
15	1	K3PO4	m3	p3	DMF	t60	0.0000	1
16	1	K3PO4	m3	p3	tAmOH	t20	0.0000	1
17	1	K3PO4	m6	p6	tAmOH	t40	0.0733	1
18	1	K3PO4	m2	p1	tAmOH	t60	0.0000	1
19	1	DBU	m3	p4	DMF	t20	0.0035	1
20	1	DBU	m3	p2	DMF	t40	0.0283	1
21	1	DBU	m6	p1	DMF	t60	0.0352	1
22	1	DBU	m3	p1	tAmOH	t20	0.0000	1
23	1	DBU	m2	p3	tAmOH	t40	0.0061	1
24	1	DBU	m6	p1	tAmOH	t60	0.1243	1
25	1	K2CO3	m3	p1	DMF	t20	0.0000	1
26	1	K2CO3	m4	p3	DMF	t40	0.0305	1
27	1	K2CO3	m2	p4	DMF	t60	0.0452	1
28	1	K2CO3	m2	p4	tAmOH	t20	0.0223	1
29	1	K2CO3	m4	p4	tAmOH	t40	0.0826	1
30	1	K2CO3	m6	p1	tAmOH	t60	0.0000	1
31	1	DIPEA	m5	p1	DMF	t20	0.0000	1
32	1	DIPEA	m5	p3	DMF	t40	0.0000	1
33	1	DIPEA	m4	p1	DMF	t60	0.0000	1
34	1	DIPEA	m2	p2	tAmOH	t20	0.0060	1
35	1	DIPEA	m5	p6	tAmOH	t40	0.0000	1
36	1	DIPEA	m6	p3	tAmOH	t60	0.0000	1
37	2	BTMG	m2	p6	tAmOH	t60	0.6057	0.224301104
38	2	BTMG	m4	p2	DMF	t40	0.0137	0.075508922
39	2	BTMG	m3	p6	DMF	t60	0.3731	0.369240211
40	2	DBU	m5	p6	tAmOH	t60	0.0269	0.083135403
41	2	Cs2CO3	m6	p2	DMF	t40	0.0916	0.053144102
42	2	K2CO3	m6	p6	tAmOH	t40	0.0098	0.081170802
43	2	K3PO4	m5	p4	tAmOH	t40	0.0000	0.019488365
44	2	BTMG	m5	p6	DMF	t20	0.0000	0.136733756
45	2	K3PO4	m5	p6	DMF	t40	0.0851	0.062798968
46	2	BTMG	m2	p4	tAmOH	t60	0.2471	0.084519205
47	2	Cs2CO3	m6	p4	tAmOH	t60	0.0032	0.05532184
48	2	BTMG	m3	p4	tAmOH	t40	0.1264	0.211952683
49	2	K2CO3	m5	p2	DMF	t60	0.0200	0.03243837

50	2	Cs2CO3	m5	p4	tAmOH	t20	0.0000	0.009830195
51	2	DBU	m5	p4	DMF	t60	0.0404	0.020542455
52	2	K2CO3	m4	p6	DMF	t40	0.0955	0.096261114
53	2	DBU	m4	p3	DMF	t40	0.0000	0.030476834
54	2	K2CO3	m6	p2	tAmOH	t20	0.0518	0.030985792
55	3	Cs2CO3	m5	p4	DMF	t20	0.0555	0.007267339
56	3	BTMG	m3	p2	tAmOH	t60	0.1025	0.291007177
57	3	K3PO4	m3	p3	DMF	t20	0.0000	0.00426861
58	3	DBU	m6	p4	tAmOH	t40	0.0970	0.060313346
59	3	DIPEA	m4	p3	tAmOH	t20	0.0000	0.011677747
60	3	Cs2CO3	m4	p6	DMF	t60	0.0383	0.070863802
61	3	DBU	m6	p1	DMF	t20	0.0000	0.047302658
62	3	BTMG	m6	p1	tAmOH	t20	0.0000	0.093006574
63	3	Cs2CO3	m4	p6	tAmOH	t40	0.2851	0.057506405
64	3	K3PO4	m6	p3	DMF	t60	0.0200	0.006561152
65	3	K3PO4	m3	p6	tAmOH	t20	0.0169	0.107335222
66	3	K3PO4	m2	p2	tAmOH	t60	0.0887	0.054641507
67	3	DBU	m6	p2	tAmOH	t20	0.0000	0.061513926
68	3	K2CO3	m4	p3	DMF	t20	0.0000	0.020984125
69	3	K2CO3	m2	p6	tAmOH	t60	0.4033	0.115517229
70	3	DIPEA	m3	p3	DMF	t20	0.0000	0.003372499
71	3	DIPEA	m3	p6	DMF	t40	0.0000	0.085207617
72	3	DIPEA	m2	p3	DMF	t60	0.0361	0.008809454
73	4	DIPEA	m3	p2	tAmOH	t40	0.0000	0.042749176
74	4	DIPEA	m3	p1	tAmOH	t60	0.0000	0.017095519
75	4	BTMG	m6	p2	DMF	t60	0.1301	0.126297472
76	4	K2CO3	m5	p1	tAmOH	t60	0.0000	0.015120335
77	4	Cs2CO3	m2	p1	tAmOH	t40	0.1321	0.041671988
78	4	BTMG	m5	p4	tAmOH	t60	0.1075	0.114965938
79	4	BTMG	m5	p6	DMF	t40	0.1833	0.095352737
80	4	Cs2CO3	m2	p3	DMF	t40	0.0470	0.031703813
81	4	BTMG	m4	p4	tAmOH	t40	0.0792	0.0992472
82	4	BTMG	m5	p4	DMF	t60	0.1523	0.11163743
83	4	DBU	m5	p1	tAmOH	t60	0.0138	0.042679326
84	4	K3PO4	m5	p2	DMF	t40	0.0900	0.040615479
85	4	K2CO3	m2	p4	DMF	t40	0.0502	0.059810674
86	4	DBU	m6	p1	tAmOH	t40	0.0385	0.072366704
87	4	K2CO3	m2	p2	tAmOH	t40	0.0234	0.070843946
88	4	K3PO4	m6	p1	tAmOH	t60	0.0256	0.02905323
89	4	Cs2CO3	m5	p2	DMF	t20	0.0993	0.03385356
90	4	DBU	m6	p2	DMF	t60	0.0392	0.056816471

Table 3 Proposed experiments for phenol alkylation reactions.

Other than the visualization in **Fig. 13**, we also plotted several condition components and their average yields based on the experiments conducted during optimization (**Fig. 126**). A significant base and temperature effect was observed, while the two solvents performed similarly. Overall, BMTG–*t*-AmOH–60 °C achieved the highest average yield.

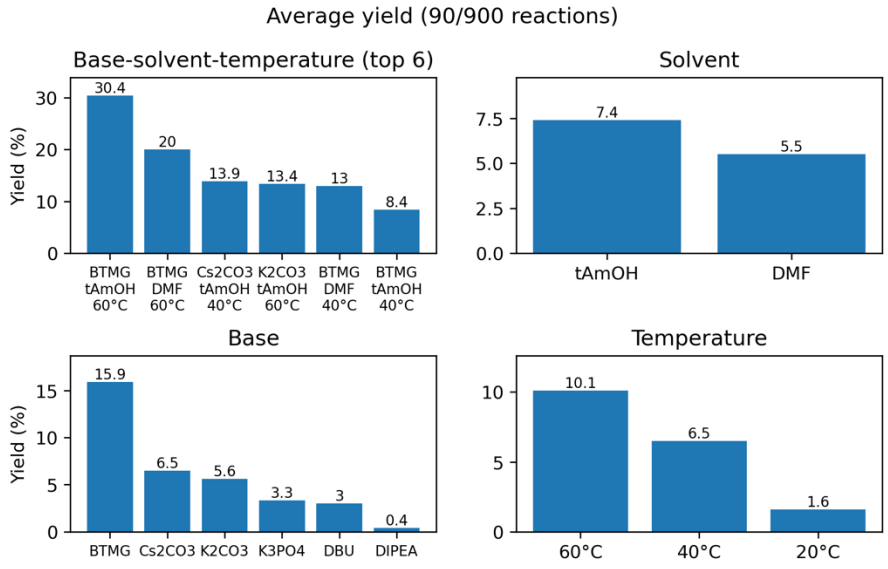


Fig. 126 Average yields of each reaction components from four optimization rounds.

Analysis of historical phenol alkylation data at Bristol Myers Squibb.

A common strategy for the design of HTE studies for the identification of hits for further reaction development and optimization centers around selecting conditions that are prevalent in the chemical literature, offer processing advantages, or present sustainability and cost advantages. Under this scenario, the design of each study is isolated from previous efforts of the same reaction type and historical data is not effectively utilized to inform reagent selection. Thus, ineffective designs propagate over time and the HTE campaigns can become inefficient and resource intensive. The 2016 BMS phenol alkylation screening data demonstrates an example of this scenario. Three separate reactions totaling 288 experiments were investigated during this time period and in 2019 these data were aggregated and statistically evaluated (**Fig. 127**). The central conclusion was that around 60% of the bases selected for these studies showed below average performance across the board and should have been excluded from evaluation in future studies. Furthermore, the high-level analysis showed that the third most effective base, MTBD, had only been utilized in a single

study despite its effectiveness and overall, there was a severe lack of chemical diversity for strong amine bases in the sample set. It became clear through this study that a systematic approach to identifying globally optimal reaction conditions across a chemically diverse pool of reaction conditions could accelerate hit identification through HTE efforts.

For the benchmark conditions, we selected K_3PO_4 and Cs_2CO_3, with DMF as solvent. K_3PO_4–DMF and Cs_2CO_3–DMF achieved the highest Z-scores of product peak area percentage (AP) and have been extensively investigated in these datasets. These two conditions represent conditions chosen with expert knowledge and ones that have demonstrated success in past investigations.

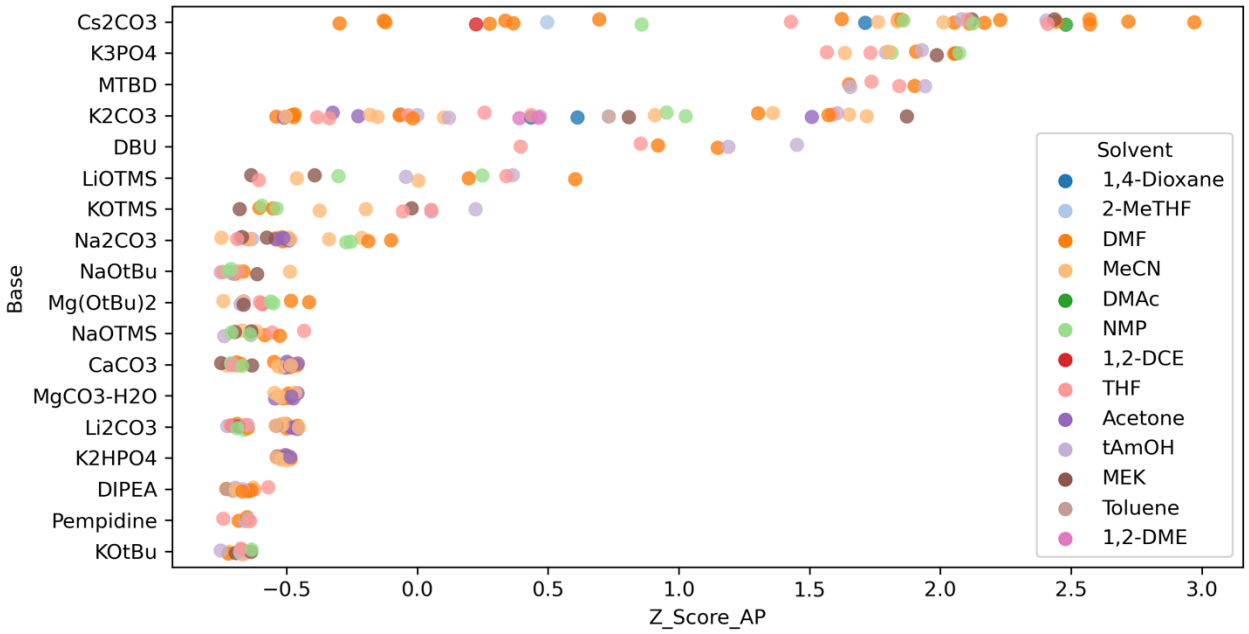
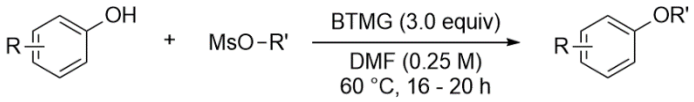


Fig. 127 Bases and solvents investigated in phenol alkylation reactions at BMS. Base is ranked by the highest Z-score of product peak area percentage (AP) achieved.

Authentic product synthesis: procedure and characterization



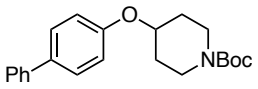
General procedure:

On the benchtop, to a 20 mL vial was added the phenol (1.0 equiv., 2 mmol) followed by 4 mL DMF, and 2-tert-butyl-1,1,3,3-tetramethylguadinine (3.0 equiv., 6 mmol, 98 mass%). The mixture stirred at room temperature for 2 min and was then treated with a solution of mesylate (1.0 equiv., 2 mmol) in DMF (2 mL). The mesylate vial was rinsed with DMF 2 x 1 mL and the contents added to the reaction vial. The reaction was heated to 60 °C and stirred no less than 16 h. UPLC/MS was used to analyze reaction progress. The reaction was split in to a second 20 mL vial (~ 3 mL in each vial) and each vial was treated with 500 μL water and 400 μL glacial acetic acid. Note that the vial will get warm, and some vapor will be generated. The vials were cooled to room temperature and then evaporated on the Biotage V-10 Touch vial evaporator using the very high boil setting. The concentrated mixtures were each dissolved in 20 mL EtOAc and recombined in a separatory funnel, rinsing each vial with 5 mL EtOAc. The organic solution was washed 1 x 1M $KHCO_3$, 2 x 1 M LiCl, 2 x water, 2 x 1M K_2CO_3, and brine (~ 50 mL each). The crude was purified by silica gel chromatography (80 - 220 g ISCO RediSep-RfGold column; 10% to 50% EtOAc/heptane gradient) or preparative SFC.

Product labeling: phenols are labeled from **p1-6**, mesylates are labeled from **m1-6**. The resulting product from the reaction are labeled **m#-p#** accordingly. **m1** and **p5** were selected as out-of-sample test substrates.

Product synthesis and characterization:

m1 series

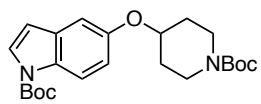


***tert*-butyl 4-([1,1'-biphenyl]-4-yloxy)piperidine-1-carboxylate (m1-p1):** Prepared according to the general procedure. The title compound was isolated via flash column chromatography (silica gel, 40-63 μm, *Silicycle*, 0-40% EtOAc/hexane) as a white solid (160 mg, 453 μmol, 23% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ 7.57 – 7.49 (m, 4H), 7.44 – 7.38 (m, 2H), 7.33 – 7.27 (m, 1H), 7.01 – 6.95 (m, 2H), 4.51 (tt, *J* = 7.2, 3.5 Hz, 1H), 3.72 (ddd, *J* = 13.5, 7.7, 3.8 Hz, 2H), 3.36 (ddd, *J* = 13.5, 7.7, 3.8 Hz, 2H), 1.95 (ddt, *J* = 11.8, 8.0, 3.9 Hz, 2H), 1.79 (dtd, *J* = 13.4, 7.4, 3.8 Hz, 2H), 1.48 (s, 9H).

^{13}C NMR (101 MHz, $CDCl_3$): δ 156.92, 155.01, 140.89, 134.30, 128.87, 128.41, 126.89, 116.52, 79.75, 72.44, 40.81, 30.71, 28.61.

HRMS (ESI-TOF): calculated for $[C_{22}H_{27}NO_3+Na]^+$: 376.1883, Found: 376.1890.

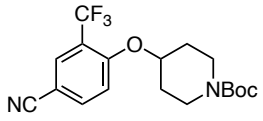


***tert*-butyl 5-((1-(*tert*-butoxycarbonyl)piperidin-4-yl)oxy)-1*H*-indole-1-carboxylate (m1-p2):** Prepared according to the general procedure on 6.00 mmol scale. The title compound was isolated via preparative SFC with the following conditions: Column: Diacel ChiralPak IC, 30 x 250 mm; Temperature: 35 °C; Mobile Phase: 30% EtOH with CO_2; Flow rate: 85 mL/min; Back Pressure: 100 bar; UV Wavelength: 250 nm. The collected fraction was dried in vacuo at ~30°C without any co-solvent (444 mg, 1.07 mmol, 18% yield).

^{1}H NMR (500 MHz, $CDCl_3$): δ 7.95 (br d, J=5.2 Hz, 1H), 7.54 - 7.44 (m, 1H), 6.99 (d, J=2.4 Hz, 1H), 6.86 (dd, J=9.0, 2.4 Hz, 1H), 6.41 (d, J=3.4 Hz, 1H), 4.41 - 4.35 (m, 1H), 3.70 - 3.61 (m, 2H), 3.25 (ddd, J=13.4, 7.9, 3.7 Hz, 2H), 1.90 - 1.81 (m, 2H), 1.74 - 1.61 (m, 2H), 1.59 (s, 9H), 1.40 (s, 9H).

^{13}C NMR (126 MHz, $CDCl_3$): δ 155.0, 153.1, 149.0, 131.5, 130.1, 126.7, 115.9, 115.1, 107.4, 107.0, 79.6, 73.3, 41.0, 30.6, 28.5, 28.2.

HRMS (ESI-TOF): calculated for $[C_{23}H_{32}N_2O_5+Na]^+$: 439.2203, Found: 439.2250.



***tert*-butyl 4-(4-cyano-2-(trifluoromethyl)phenoxy)piperidine-1-carboxylate (m1-p3):** Prepared according to the general procedure on 8.00 mmol scale. The title compound was isolated via preparative SFC with the following conditions: Column: Diacel ChiralPak IC, 30 x 250 mm; Temperature: 35 °C; Mobile Phase: 30% EtOH with CO_2; Flow rate: 85 mL/min; Back Pressure:

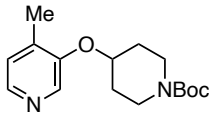
100 bar; UV Wavelength: 250 nm. The collected fraction was dried in vacuo at ~30°C without any co-solvent (1.872 g, 5.05 mmol, 63% yield).

^{1}H NMR (500 MHz, DMSO-d6): δ 8.16 (d, *J* =2.0 Hz, 1H), 8.12 (dd, *J* =8.8, 2.1 Hz, 1H), 7.55 (d, *J* =8.9 Hz, 1H), 4.83 (tt, *J*=8.3, 3.9 Hz, 1H), 3.65 - 3.56 (m, 1H), 3.47 - 3.41 (m, 1H), 3.38 (br s, 1H), 3.17 (br dd, *J* =9.1, 2.2 Hz, 1H), 1.96 - 1.85 (m, 2H), 1.61 (dtd, *J* =13.0, 8.7, 4.0 Hz, 2H), 1.44 - 1.36 (m, 9H).

^{13}C NMR (126 MHz, DMSO-d6): δ 158.7, 154.3, 154.3, 139.1, 132.1 (q, *J*=4.2 Hz), 123.2 (q, *J*=269.6 Hz), 119.2 (q, *J*=37.6 Hz), 116.1, 103.3, 79.4, 73.6, 38.3, 31.7, 28.5.

^{19}F NMR (471 MHz, DMSO-d_6): δ -61.76.

HRMS (ESI-TOF): calculated for $[C_{18}H_{21}F_3N_2O_3+Na]^+$: 393.1396, Found: 393.1444.



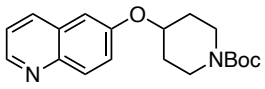
***tert*-butyl 4-((4-methylpyridin-3-yl)oxy)piperidine-1-carboxylate (m1-p4):** Prepared according to the general procedure on a 4.00 mmol scale. The title compound was isolated via silica gel chromatography (silica gel, 40-63 μm, *Silicycle*, 1-10% methanol/dichloromethane) as a yellow oil (433mg, 1.48 mmol, 37% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ 8.15 (s, 1H), 8.08 (d, *J* = 4.7 Hz, 1H), 7.06 (d, *J* = 4.7 Hz, 1H), 4.55 (tt, *J* = 7.0, 3.4 Hz, 1H), 3.64 (ddd, *J* = 13.5, 7.9, 3.8 Hz, 2H), 3.37 (ddd, *J* = 13.5, 7.4,

3.9 Hz, 2H), 2.22 (s, 3H), 1.92 (ddt, *J* = 11.7, 7.6, 3.7 Hz, 2H), 1.78 (dtd, *J* = 13.6, 7.1, 3.7 Hz, 2H), 1.45 (s, 9H).

^{13}C NMR (101 MHz, $CDCl_3$): δ 154.88, 152.32, 142.54, 136.98, 135.51, 125.92, 79.79, 73.01, 40.65, 30.66, 28.53, 15.88.

HRMS (ESI-TOF): calculated for $[C_{16}H_{24}N_2O_3+H]^+$: 293.1860, Found: 293.1862.

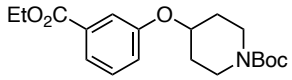


***tert*-butyl 4-(quinolin-6-yloxy)piperidine-1-carboxylate (m1-p5):** Prepared according to the general procedure on 7.00 mmol scale. The title compound was isolated via flash column chromatography (silica gel, 40-63 μm, *Silicycle*, 30-70% EtOAc/hexane) as an orange solid (656 mg, 2.00 mmol, 29% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ 8.78 (dd, *J* = 4.3, 1.6 Hz, 1H), 8.09 – 7.99 (m, 2H), 7.37 (td, *J* = 8.8, 3.5 Hz, 2H), 7.11 (d, *J* = 2.8 Hz, 1H), 4.64 (tt, *J* = 7.2, 3.5 Hz, 1H), 3.74 (ddd, *J* = 12.3, 7.6, 3.8 Hz, 2H), 3.40 (ddd, *J* = 12.7, 7.5, 3.8 Hz, 2H), 2.03-1.96 (m, 2H), 1.89 – 1.77 (m, 2H), 1.48 (s, 9H).

^{13}C NMR (101 MHz, $CDCl_3$): δ 155.38, 154.97, 148.27, 144.49, 134.88, 131.27, 129.40, 123.20, 121.54, 108.14, 79.82, 72.57, 40.61, 30.51, 28.59.

HRMS (ESI-TOF): calculated for $[C_{19}H_{24}N_2O_3+H]^+$: 329.1860, Found: 329.1866.



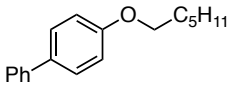
***tert*-butyl 4-(3-(ethoxycarbonyl)phenoxy)piperidine-1-carboxylate (m1-p6):** Prepared according to the general procedure. The title compound was isolated via flash column chromatography (silica gel, 40-63 μm, *Silicycle*, 20-40% EtOAc/hexane) as a colorless oil (162 mg, 464 μmol, 23% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ 7.64 (dt, *J* = 7.7, 1.4 Hz, 1H), 7.57 (dd, *J* = 2.7, 1.5 Hz, 1H), 7.34 (t, *J* = 7.9 Hz, 1H), 7.09 (ddd, *J* = 8.2, 2.6, 1.0 Hz, 1H), 4.54 (tt, *J* = 7.2, 3.5 Hz, 1H), 4.41 – 4.32 (m, 2H), 3.70 (ddd, *J* = 13.5, 7.6, 3.8 Hz, 2H), 3.35 (ddd, *J* = 13.5, 7.7, 3.9 Hz, 2H), 1.93 (ddt, *J* = 11.7, 7.6, 3.8 Hz, 2H), 1.75 (dtd, *J* = 13.6, 7.4, 3.8 Hz, 2H), 1.47 (s, 9H), 1.39 (td, *J* = 7.1, 0.5 Hz, 3H).

^{13}C NMR (101 MHz, $CDCl_3$): δ 166.56, 157.34, 154.98, 132.11, 129.66, 122.40, 121.31, 116.60, 79.77, 72.56, 61.23, 40.73, 30.58, 28.59, 14.47.

HRMS (ESI-TOF): calculated for $[C_{19}H_{27}NO_5+Na]^+$: 372.1781, Found: 372.1796.

m2 series:

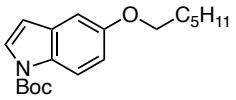


4-(hexyloxy)-1,1'-biphenyl (m2-p1): Prepared according to the general procedure on a 10.00 mmol scale. The title compound was isolated via flash column chromatography (silica gel, 40-63 μm, *Silicycle*, 5-15% EtOAc/hexane) as a white solid (863 mg, 3.93 mmol, 34% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ 7.57 – 7.49 (m, 4H), 7.44 – 7.38 (m, 2H), 7.32 – 7.27 (m, 1H), 7.00 – 6.94 (m, 2H), 4.00 (t, *J* = 6.6 Hz, 2H), 1.81 (dq, *J* = 7.9, 6.6 Hz, 2H), 1.53 – 1.43 (m, 2H), 1.36 (tt, *J* = 7.1, 3.3 Hz, 4H), 0.95 – 0.88 (m, 3H).

^{13}C NMR (101 MHz, $CDCl_3$): δ 158.89, 141.06, 133.69, 128.84, 128.25, 126.86, 126.73, 114.93, 68.26, 31.76, 29.43, 25.90, 22.77, 14.19.

HRMS (ESI-TOF): calculated for $[C_{18}H_{22}O+H]^+$: 255.1743, Found: 255.1735.

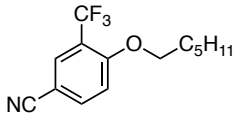


tert-butyl 5-(hexyloxy)-1H-indole-1-carboxylate (m2-p2): Prepared according to the general procedure on a 5.00 mmol scale. The title compound was isolated via silica gel chromatography (silica gel, 40-63 μm, *Silicycle*, 15-50% EtOAc/hexane) as a colorless oil (1.317 g, 4.15 mmol, 83% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ 8.00 (d, *J* = 9.2 Hz, 1H), 7.55 (d, *J* = 3.7 Hz, 1H), 7.02 (d, *J* = 2.4 Hz, 1H), 6.92 (dd, *J* = 9.0, 2.5 Hz, 1H), 6.48 (dd, *J* = 3.7, 0.7 Hz, 1H), 3.99 (t, *J* = 6.6 Hz, 2H), 1.80 (dq, *J* = 8.0, 6.7 Hz, 2H), 1.66 (s, 9H), 1.52 – 1.41 (m, 2H), 1.35 (dp, *J* = 7.0, 3.3 Hz, 4H), 0.95 – 0.86 (m, 3H).

^{13}C NMR (101 MHz, $CDCl_3$): δ 155.49, 149.90, 131.50, 130.01, 126.55, 115.90, 113.69, 107.25, 104.59, 83.55, 68.74, 31.78, 29.51, 28.35, 25.93, 22.77, 14.19.

HRMS (ESI-TOF): calculated for $[C_{19}H_{27}NO_3+H]^+$: 318.2064, Found: 318.2064.



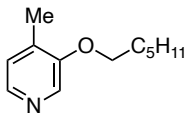
4-(hexyloxy)-3-(trifluoromethyl)benzonitrile (m2-p3): Prepared according to the general procedure on a 5.00 mmol scale. The title compound was isolated via silica gel chromatography (silica gel, 40-63 μm, *Silicycle*, 10-30% EtOAc/hexane) as a pale yellow solid (577 mg, 2.13 mmol, 43% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ 7.85 (d, *J* = 2.1 Hz, 1H), 7.77 (dd, *J* = 8.7, 2.1 Hz, 1H), 7.05 (d, *J* = 8.7 Hz, 1H), 4.11 (t, *J* = 6.3 Hz, 2H), 1.81 (d, *J* = 6.3 Hz, 2H), 1.52 – 1.44 (m, 2H), 1.38 – 1.30 (m, 4H), 0.94 – 0.87 (m, 3H).

^{13}C NMR (101 MHz, $CDCl_3$): δ 160.39, 137.60, 131.50 (q, *J* = 5.4 Hz), 122.57 (q, *J* = 272.9 Hz), 120.28 (q, *J* = 32.2 Hz), 118.11, 113.46, 103.62, 69.61, 31.43, 28.80, 25.47, 22.62, 14.05.

^{19}F NMR (376 MHz, $CDCl_3$): δ -63.35.

HRMS (ESI-TOF): calculated for $[C_{14}H_{16}F_3NO+NH_4]^+$: 289.1522, Found: 289.1522.

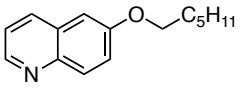


3-(hexyloxy)-4-methylpyridine (m2-p4): Prepared according to the general procedure on a 5.00 mmol scale. The title compound was isolated via silica gel chromatography (silica gel, 40-63 μm, *Silicycle*, 30-70% EtOAc/hexane) as a pale yellow oil (444 mg, 2.30 mmol, 46% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ 8.14 (s, 1H), 8.10 (d, *J* = 4.7 Hz, 1H), 7.05 (dt, *J* = 4.7, 0.7 Hz, 1H), 4.05 (t, *J* = 6.4 Hz, 2H), 2.23 (s, 3H), 1.81 (ddt, *J* = 9.0, 7.8, 6.4 Hz, 2H), 1.54 – 1.41 (m, 2H), 1.35 (tt, *J* = 7.1, 3.2 Hz, 4H), 0.97 – 0.83 (m, 3H).

^{13}C NMR (101 MHz, $CDCl_3$): δ 154.10, 142.29, 135.75, 133.48, 125.49, 68.71, 31.65, 29.42, 25.82, 22.71, 15.74, 14.12.

HRMS (ESI-TOF): calculated for $[C_{12}H_{19}NO–H]^-$: 192.1394, Found: 192.1363.

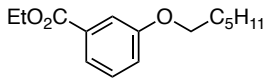


6-(hexyloxy)quinoline (m2-p5): Prepared according to the general procedure on a 5.00 mmol scale. The title compound was isolated via silica gel chromatography (silica gel, 40-63 μm, *Silicycle*, 30-70% EtOAc/hexane) as a red oil (507 mg, 2.21 mmol, 44% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ 8.75 (dd, *J* = 4.3, 1.7 Hz, 1H), 8.05 – 7.94 (m, 2H), 7.35 (ddd, *J* = 14.0, 8.7, 3.5 Hz, 2H), 7.05 (d, *J* = 2.8 Hz, 1H), 4.07 (t, *J* = 6.6 Hz, 2H), 1.90 – 1.79 (m, 2H), 1.56 – 1.45 (m, 2H), 1.37 (dq, *J* = 6.7, 3.5 Hz, 4H), 0.96 – 0.84 (m, 3H).

^{13}C NMR (101 MHz, $CDCl_3$): δ 157.42, 147.93, 144.47, 134.88, 130.90, 129.50, 122.74, 121.42, 105.96, 68.47, 31.73, 29.28, 25.90, 22.75, 14.17.

HRMS (ESI-TOF): calculated for $[C_{15}H_{19}NO+H]^+$: 230.1539, Found: 230.1557.



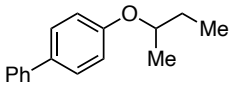
ethyl 3-(hexyloxy)benzoate (m2-p6): Prepared according to the general procedure on a 10.00 mmol scale. The title compound was isolated via silica gel chromatography (silica gel, 40-63 μm, *Silicycle*, 10-25% EtOAc/hexane) as a colorless oil (1.43 g, 5.72 mmol, 57% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ 7.62 (dt, *J* = 7.7, 1.3 Hz, 1H), 7.55 (dd, *J* = 2.7, 1.5 Hz, 1H), 7.32 (t, *J* = 7.9 Hz, 1H), 7.08 (ddd, *J* = 8.2, 2.7, 1.0 Hz, 1H), 4.37 (q, *J* = 7.1 Hz, 2H), 4.00 (t, *J* = 6.6 Hz, 2H), 1.79 (dq, *J* = 7.9, 6.6 Hz, 2H), 1.51 – 1.42 (m, 2H), 1.39 (t, *J* = 7.1 Hz, 3H), 1.37-1.32 (m, 4H), 0.91 (td, *J* = 5.9, 2.6 Hz, 3H).

^{13}C NMR (101 MHz, $CDCl_3$): δ 166.70, 159.25, 131.90, 129.41, 121.86, 119.85, 114.89, 68.36, 61.13, 31.71, 29.31, 25.84, 22.74, 14.47, 14.16.

HRMS (ESI-TOF): calculated for $[C_{15}H_{22}O_3+H]^+$: 251.1642, Found: 251.1640.

m3 series

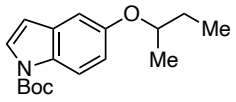


4-(*sec*-butoxy)-1,1'-biphenyl (m3-p1): Prepared according to the general procedure on 2 mmol scale. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 5-50% EtOAc/Heptane) as colorless oil (335 mg, 1480 μmol, 74% yield).

^{1}H NMR (500 MHz, DMSO-d6): δ 7.61 - 7.53 (m, 4H), 7.41 (t, *J*=7.3 Hz, 2H), 7.33 - 7.25 (m, 1H), 6.98 (d, *J*=7.7 Hz, 2H), 4.40 (sxt, *J*=6.0 Hz, 1H),1.71 - 1.54 (m, 2H), 1.23 (d, *J*=6.0 Hz, 3H), 0.92 (t, *J*=7.5 Hz, 3H).

^{13}C NMR (126 MHz, DMSO-d6): δ 157.4, 139.9, 132.2, 128.8, 127.7, 126.6, 126.1, 115.9, 74.1, 28.5, 19.0, 9.5.

HRMS (ESI-TOF): calculated for $[C_{16}H_{18}O+H]^+$: 227.1430, Found: 227.1493.

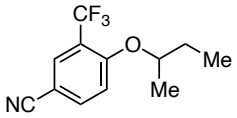


***tert*-butyl 5-(*sec*-butoxy)-1*H*-indole-1-carboxylate (m3-p2):** Prepared according to the general procedure on 2 mmol scale. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 0-100% EtOAc/Heptane) as colorless oil (393 mg, 1360 μmol, 68% yield).

^{1}H NMR (500 MHz, DMSO-d$_6$): δ 7.90 (d, J=9.0 Hz, 1H), 7.61 (d, J=3.7 Hz, 1H), 7.12 (d, J=2.6 Hz, 1H), 6.90 (dd, J=8.9, 2.4 Hz, 1H), 6.60 (dd, J=3.7, 0.6 Hz, 1H), 4.35 (sxt, J=6.0 Hz, 1H), 1.69 - 1.54 (m, 11H), 1.25 - 1.20 (m, 3H), 0.93 (t, J=7.5 Hz, 3H).

^{13}C NMR (126 MHz, DMSO-d$_6$): δ 154.3, 149.5, 131.6, 129.3, 127.0, 115.8, 115.1, 107.8, 106.9, 84.0, 75.3, 29.0, 28.1, 19.6, 10.0.

HRMS (ESI-TOF): calculated for $[C_{17}H_{23}NO_3+H]^+$: 290.1751, Found: 290.1809.



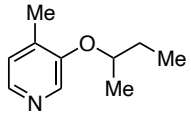
4-(*sec*-butoxy)-3-(trifluoromethyl)benzonitrile (m3-p3): Prepared according to the general procedure on 2 mmol scale. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 5-50% EtOAc/Heptane) as colorless oil (329 mg, 1350 µmol, 68% yield).

^{1}H NMR (500 MHz, DMSO-d$_6$): δ 8.11 (d, *J*=2.0 Hz, 1H), 8.07 (dd, *J*=8.7, 2.1 Hz, 1H), 7.47 (d, *J*=8.9 Hz, 1H), 4.75 (sxt, *J*=6.0 Hz, 1H), 1.72 - 1.57 (m, 2H), 1.26 (d, *J*=6.0 Hz, 3H), 0.91 (t, *J*=7.4 Hz, 3H)

^{13}C NMR (126 MHz, DMSO-d$_6$): δ 159.1, 138.6, 131.5 (q, *J*=4.2 Hz), 122.7 (q, *J*=272.9 Hz), 118.3 (q, *J* = 32.2 Hz), 118.0, 115.4, 102.3, 76.3, 76.2, 28.3, 27.9, 18.4, 8.9.

^{19}F NMR (471 MHz, DMSO-d$_6$): δ -61.84.

HRMS (ESI-TOF): calculated for $[C_{12}H_{12}F_3NO+NH_4]^+$: 261.1209, Found: 261.1270.

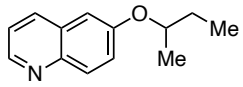


3-(*sec*-butoxy)-4-methylpyridine (m3-p4): Prepared according to the general procedure on 2 mmol scale. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 0-100% EtOAc/Heptane) as colorless oil (103 mg, 620 µmol, 31% yield).

^{1}H NMR (500 MHz, DMSO-d6): δ 8.22 (s, 1H), 8.03 (d, *J*=4.6 Hz, 1H), 7.16 (d, *J*=4.6 Hz, 1H), 4.49 (sxt, *J*=6.0 Hz, 1H), 2.15 (s, 3H), 1.70 - 1.56 (m, 2H), 1.24 (d, *J*=6.0 Hz, 3H), 0.93 (t, *J*=7.5 Hz, 3H)

^{13}C NMR (126 MHz, DMSO-d6): δ 152.6, 141.6, 135.4, 125.5, 125.5, 75.2, 28.7, 19.1, 15.3, 9.4.

HRMS (ESI-TOF): calculated for $[C_{10}H_{15}NO+H]^{+}$: 166.1226, Found: 166.1240.

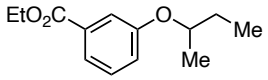


6-(*sec*-butoxy)quinoline (m3-p5): Prepared according to the general procedure on 2 mmol scale. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 0-100% EtOAc/Heptane) as red oil (237 mg, 1178 μmol, 59% yield).

^{1}H NMR (500 MHz, DMSO-d6): δ 8.71 (dd, *J*=4.3, 1.7 Hz, 1H), 8.22 (dt, *J*=7.6, 0.9 Hz, 1H), 7.90 (d, *J*=8.9 Hz, 1H), 7.44 (dd, *J*=8.3, 4.2 Hz, 1H), 7.38 - 7.34 (m, 2H), 4.54 (sxt, *J*=6.0 Hz, 1H), 1.77 - 1.59 (m, 2H), 1.30 (d, *J*=6.1 Hz, 3H), 0.95 (t, *J*=7.4 Hz, 3H).

^{13}C NMR (126 MHz, DMSO-d6): δ 155.5, 147.8, 143.6, 134.6, 130.4, 129.1, 122.8, 121.5, 107.6, 74.4, 28.4, 18.8, 9.5.

HRMS (ESI-TOF): calculated for $[C_{13}H_{15}NO+H]^{+}$: 202.1154, Found: 202.1314



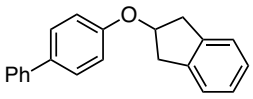
ethyl 3-(*sec*-butoxy)benzoate (m3-p6): Prepared according to the general procedure. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 5-50% EtOAc/Heptane) as colorless oil (283 mg, 1273 µmol, 64% yield).

^{1}H NMR (500 MHz, DMSO-d6): δ 7.52 (dt, *J*=7.8, 1.1 Hz, 1H), 7.47 - 7.37 (m, 2H), 7.21 (ddd, *J*=8.2, 2.7, 1.0 Hz, 1H), 4.51 - 4.39 (m, 1H), 4.31 (q, *J*=7.0 Hz, 2H), 1.73 - 1.52 (m, 2H), 1.32 (t, *J*=7.1 Hz, 3H), 1.24 (d, *J*=6.1 Hz, 3H), 0.93 (t, *J*=7.5 Hz, 3H).

^{13}C NMR (126 MHz, DMSO-d6): δ 165.5, 157.8, 131.3, 129.9, 121.1, 120.5, 115.8, 74.5, 60.7, 28.5, 18.8, 14.1, 9.4.

HRMS (ESI-TOF): calculated for $[C_{13}H_{18}O_3+H]^+$: 223.1329, Found: 223.1355

m4 series

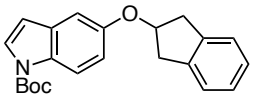


2-([1,1'-biphenyl]-4-yloxy)-2,3-dihydro-1*H*-indene (m4-p1): Prepared according to the general procedure with 2 mmol phenol limiting reagent. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 5-50% EtOAc/Heptane) as off-white solids (92 mg, 321 µmol, 16% yield).

^{1}H NMR (500 MHz, DMSO-d6): δ 7.66 - 7.52 (m, 4H), 7.43 (t, *J*=7.7 Hz, 2H), 7.33 - 7.23 (m, 3H), 7.21 - 7.15 (m, 2H), 7.03 (d, *J*=7.8 Hz, 2H), 5.34 - 5.21 (m, 1H), 3.39 (dd, *J*=16.9, 6.0 Hz, 2H), 3.05 (dd, *J*=16.9, 2.2 Hz, 2H).

^{13}C NMR (126 MHz, DMSO-d6): δ 157.1, 141.1, 140.1, 132.8, 129.2, 128.2, 127.0, 126.8, 126.5, 125.0, 116.1, 77.6, 39.5.

HRMS (ESI-TOF): calculated for $[C_{21}H_{18}O+NH_4]^+$: 304.1696, Found: 304.1756.

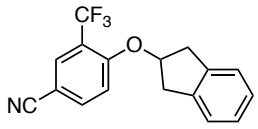


***tert*-butyl 5-((2,3-dihydro-1*H*-inden-2-yl)oxy)-1*H*-indole-1-carboxylate (m4-p2):** Prepared according to the general procedure on 4 mmol scale. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 5-50% EtOAc/Heptane) as colorless oil (200 mg, 780 μmol, 20% yield).

^{1}H NMR (500 MHz, DMSO-d6): δ 7.93 (d, *J*=9.0 Hz, 1H), 7.64 (d, *J*=3.7 Hz, 1H), 7.26 (dd, *J*=5.3, 3.2 Hz, 2H), 7.18 (td, *J*=5.6, 2.8 Hz, 3H), 6.90 (dd, *J*=9.0, 2.4 Hz, 1H), 6.64 (dd, *J*=3.7, 0.6 Hz, 1H), 5.30 - 5.20 (m, 1H), 3.42 - 3.34 (m, 2H), 3.05 (dd, *J*=16.8, 2.4 Hz, 2H), 1.71 - 1.54 (m, 9H).

^{13}C NMR (126 MHz, DMSO-d6): δ 153.6, 149.4, 141.1, 131.5, 129.5, 126.9, 126.8, 124.9, 115.8, 114.4, 107.7, 106.0, 83.9, 77.9, 39.5, 28.0.

HRMS (ESI-TOF): calculated for $[C_{22}H_{23}NO_3+H]^+$: 350.1751, Found: 350.1803.



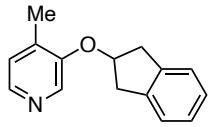
4-((2,3-dihydro-1*H*-inden-2-yl)oxy)-3-(trifluoromethyl)benzonitrile (m4-p3): Prepared according to the general procedure. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 25% EtOAc/Heptane) as orange solid (237 mg, 780 µmol, 39% yield).

^{1}H NMR (500 MHz, DMSO-d$_6$): δ 8.17 - 8.13 (m, 2H), 7.59 (d, J=8.7 Hz, 1H), 7.31 - 7.23 (m, 2H), 7.23 - 7.14 (m, 2H), 5.56 - 5.49 (m, 1H), 3.51 - 3.44 (m, 2H), 3.04 (dd, J=17.1, 2.4 Hz, 2H).

^{13}C NMR (126 MHz, DMSO-d$_6$): δ 159.1, 140.4, 139.1, 132.1 (q, *J* =6.3 Hz), 127.2, 125.0, 123.0 (q, *J* =271.7 Hz), 119.2 (q, *J* =31.4 Hz), 118.4, 116.2, 103.4, 80.1, 39.5.

^{19}F NMR (471 MHz, DMSO-d$_6$): δ -61.78.

HRMS (ESI-TOF): calculated for $[C_{17}H_{12}F_3NO+NH_4]^+$: 321.1209, Found: 321.1212.

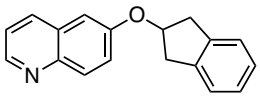


3-((2,3-dihydro-1*H*-inden-2-yl)oxy)-4-methylpyridine (m4-p4): Prepared according to the general procedure on 4 mmol scale. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 0-100% EtOAc/Heptane) as colorless oil (68 mg, 301 µmol, 8% yield).

^{1}H NMR (500 MHz, DMSO-d6): δ 8.34 (s, 1H), 8.09 (d, *J*=4.7 Hz, 1H), 7.27 (dd, *J*=5.4, 3.3 Hz, 2H), 7.22 - 7.12 (m, 3H), 5.42 - 5.28 (m, 1H), 3.42 (dd, *J*=16.9, 6.3 Hz, 2H), 3.05 (dd, *J*=16.9, 2.6 Hz, 2H), 2.05 (s, 3H).

^{13}C NMR (126 MHz, DMSO-d6): δ 152.4, 142.2, 140.7, 135.6, 135.1, 126.7, 125.7, 124.8, 78.5, 39.5, 15.4.

HRMS (ESI-TOF): calculated for $[C_{15}H_{15}NO+H]^+$: 226.1154, Found: 226.1382.

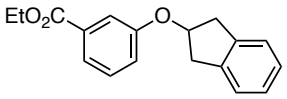


6-((2,3-dihydro-1*H*-inden-2-yl)oxy)quinoline (m4-p5): Prepared according to the general procedure. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 40% EtOAc/Heptane) as yellow oil (410 mg, 1.57 mmol, 78% yield).

^{1}H NMR (500 MHz, DMSO-d6): δ 8.74 (dd, J=4.3, 1.5 Hz, 1H), 8.28 (d, J=8.0 Hz, 1H), 7.91 (d, J=9.2 Hz, 1H), 7.52 - 7.43 (m, 2H), 7.36 - 7.25 (m, 3H), 7.22 - 7.15 (m, 2H), 5.46 - 5.32 (m, 1H), 3.54 - 3.41 (m, 2H), 3.12 (dd, J=16.9, 2.3 Hz, 2H).

^{13}C NMR (126 MHz, DMSO-d6): δ 155.5, 148.4, 144.2, 141.1, 135.3, 131.0, 129.5, 127.0, 125.1, 123.1, 122.1, 108.0, 78.2, 39.6.

HRMS (ESI-TOF): calculated for $[C_{18}H_{15}NO+H]^+$: 262.1226, Found: 262.1272.



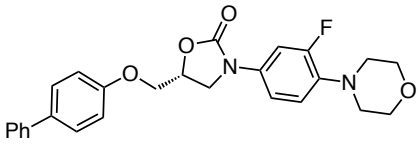
ethyl 3-((2,3-dihydro-1*H*-inden-2-yl)oxy)benzoate (m4-p6): Prepared according to the general procedure. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 25% EtOAc/Heptane) as yellow oil (81.3 mg, 288 μmol, 14% yield).

^{1}H NMR (500 MHz, DMSO-d$_6$): δ 7.55 (dt, J=7.8, 1.1 Hz, 1H), 7.48 - 7.40 (m, 2H), 7.29 - 7.16 (m, 5H), 5.34 - 5.28 (m, 1H), 4.31 (q, J=7.0 Hz, 2H), 3.39 (dd, J=16.9, 6.1 Hz, 2H), 3.03 (dd, J=16.9, 2.2 Hz, 2H), 1.31 (t, J=7.2 Hz, 3H).

^{13}C NMR (126 MHz, DMSO-d$_6$): δ 166.0, 157.7, 141.1, 131.8, 130.5, 127.0, 125.1, 121.8, 120.8, 115.9, 78.1, 61.3, 39.5, 14.6.

HRMS (ESI-TOF): calculated for $[C_{18}H_{18}O_3+H]^+$: 283.1329, Found: 283.1334.

m5 series:



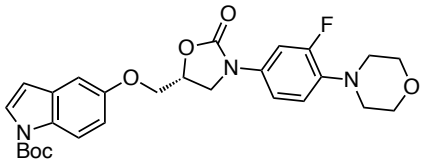
(*R*)-5-(([1,1'-biphenyl]-4-yloxy)methyl)-3-(3-fluoro-4-morpholinophenyl)oxazolidin-2-one (m5-p1): Prepared according to the general procedure on a 10.00 mmol scale. The title compound was isolated via silica gel chromatography (silica gel, 40-63 μm, *Silicycle*, 40-80% EtOAc/hexane) as a white solid (430 mg, 959 μmol, 10% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ 7.57 – 7.47 (m, 5H), 7.45 – 7.38 (m, 2H), 7.34 – 7.29 (m, 1H), 7.17 (ddd, *J* = 8.8, 2.6, 1.2 Hz, 1H), 7.06 – 6.94 (m, 3H), 5.00 (dq, *J* = 10.4, 4.9 Hz, 1H), 4.30 – 4.21 (m, 2H), 4.17 (t, *J* = 8.8 Hz, 1H), 4.05 (dd, *J* = 8.9, 5.9 Hz, 1H), 3.94 – 3.82 (m, 4H), 3.14 – 3.02 (m, 4H).

^{13}C NMR (101 MHz, $CDCl_3$): δ 157.64, 155.70 (d, *J* = 246.5 Hz), 154.37, 140.60, 136.57 (d, *J* = 8.9 Hz), 135.10, 133.38 (d, *J* = 10.4 Hz), 128.91, 128.47, 127.06, 126.91, 119.03 (d, *J* = 4.2 Hz), 115.05, 114.06 (d, *J* = 3.3 Hz), 107.65 (d, *J* = 26.3 Hz), 70.48, 68.15, 67.10, 51.19, 51.16, 47.58.

^{19}F NMR (376 MHz, $CDCl_3$): δ -120.20.

HRMS (ESI-TOF): calculated for $[C_{26}H_{25}FN_2O_4+H]^+$: 449.1871, Found: 449.1877



***tert*-butyl (*R*)-5-((3-(3-fluoro-4-morpholinophenyl)-2-oxooxazolidin-5-yl)methoxy)-1*H*-indole-1-carboxylate (m5-p2):** Prepared according to the general procedure on a 6.00 mmol scale. The title compound was recrystallized with dichloromethane/hexane to provide a white solid (844 mg, 1.65 mmol, 28% yield).

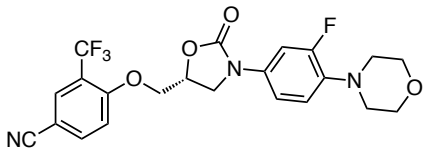
^{1}H NMR (400 MHz, $CDCl_3$): δ 8.03 (d, *J* = 9.0 Hz, 1H), 7.58 (d, *J* = 3.7 Hz, 1H), 7.50 (dd, *J* = 14.4, 2.6 Hz, 1H), 7.17 (ddd, *J* = 8.9, 2.6, 1.2 Hz, 1H), 7.04 (d, *J* = 2.5 Hz, 1H), 6.96 (t, *J*

= 9.1 Hz, 1H), 6.91 (dd, *J* = 9.0, 2.6 Hz, 1H), 6.49 (d, *J* = 3.7 Hz, 1H), 5.03 – 4.96 (m, 1H), 4.30 – 4.22 (m, 2H), 4.16 (t, *J* = 8.9 Hz, 1H), 4.06 (dd, *J* = 8.8, 5.9 Hz, 1H), 3.89 – 3.87 (m, 4H), 3.08 – 3.06 (m, 4H), 1.66 (s, 9H).

^{13}C NMR (101 MHz, $CDCl_3$): δ 155.71 (d, *J* = 246.5 Hz), 154.45, 154.33, 149.75, 136.51 (d, *J* = 8.7 Hz), 133.47 (d, *J* = 10.6 Hz), 131.48, 130.74, 127.02, 119.03 (d, *J* = 4.2 Hz), 116.17, 114.05 (d, *J* = 3.4 Hz), 113.43, 107.65 (d, *J* = 26.3 Hz), 107.12, 105.07, 83.83, 70.62, 68.79, 67.10, 51.21, 51.18, 47.64, 28.33.

^{19}F NMR (376 MHz, $CDCl_3$): δ -120.27.

HRMS (ESI-TOF): calculated for $[C_{27}H_{30}FN_3O_6+H]^+$: 512.2191, Found: 512.2201.



(*R*)-4-((3-(3-fluoro-4-morpholinophenyl)-2-oxooxazolidin-5-yl)methoxy)-3-(trifluoromethyl)benzonitrile (m5-p3): Prepared according to the general procedure on 6 mmol scale. The title compound was isolated via preparative SFC with the following conditions: Column: Diacel ChiralPak IC, 30 x 250 mm; Temperature: 35 °C; Mobile Phase: 30% EtOH with CO_2; Flow rate: 85 mL/min; Back Pressure: 100 bar; UV Wavelength: 250 nm. The collected fraction was dried in vacuo at ~30°C without any co-solvent (672 mg, 1.44 mmol, 24% yield).

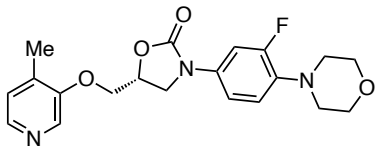
^{1}H NMR (500 MHz, DMSO-d6): δ 8.17 (d, J=8.6 Hz, 1H), 8.15 (s, 1H), 7.55 - 7.46 (m, 2H), 7.19 (dd, J=8.9, 2.0 Hz, 1H), 7.07 (t, J=9.4 Hz, 1H), 5.12 (qd, J=5.8, 3.4 Hz, 1H), 4.59 (dd,

J=11.1, 2.3 Hz, 1H), 4.48 (dd, J=11.1, 3.7 Hz, 1H), 4.22 (t, J=9.3 Hz, 1H), 3.90 (dd, J=9.1, 5.6 Hz, 1H), 3.78 - 3.70 (m, 4H), 3.03 - 2.93 (m, 4H).

^{13}C NMR (126 MHz, DMSO-d6): δ 159.5, 156.0, 154.3, 154.0, 139.3, 135.9 (d, *J*=8.4 Hz), 133.8 (d, *J*=10.5 Hz), 131.9 (br q, *J* =4.2 Hz), 119.7 (q, *J* =273.8 Hz), 119.6 (br d, *J*=4.2 Hz), 118.7 (q, *J* =31.4 Hz), 115.3, 114.4, 107.0 (d, *J*=26.3 Hz), 104.0, 70.5, 69.9, 66.6, 51.2, 46.5.

^{19}F NMR (471 MHz, DMSO-d_6): δ -61.95, -121.44.

HRMS (ESI-TOF): calculated for $[C_{22}H_{19}F_4N_3O_4+H]^+$: 466.1384, Found: 466.1400



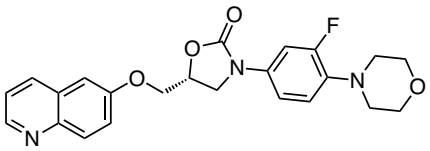
(*R*)-3-(3-fluoro-4-morpholinophenyl)-5-(((4-methylpyridin-3-yl)oxy)methyl)oxazolidin-2-one (m5-p4): Prepared according to the general procedure on a 6.00 mmol scale. The title compound was isolated via silica gel chromatography (silica gel, 40-63 μm, *Silicycle*, 0-10% methanol/dichloromethane) as a pale brown solid (551 mg, 1.42 mmol, 24% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ 8.18 (s, 2H), 7.47 (dd, *J* = 14.4, 2.6 Hz, 1H), 7.15 (ddd, *J* = 8.8, 2.6, 1.2 Hz, 1H), 7.10 (s, 1H), 6.93 (t, *J* = 9.1 Hz, 1H), 5.01 (ddt, *J* = 9.0, 5.3, 3.6 Hz, 1H), 4.37 (dd, *J* = 10.3, 3.6 Hz, 1H), 4.29 (dd, *J* = 10.3, 3.7 Hz, 1H), 4.20 (t, *J* = 8.9 Hz, 1H), 4.03 (dd, *J* = 8.8, 5.3 Hz, 1H), 3.89 – 3.82 (m, 4H), 3.08 – 3.01 (m, 4H), 2.15 (s, 3H).

^{13}C NMR (101 MHz, $CDCl_3$): δ 155.68 (d, *J* = 246.3 Hz), 154.34, 143.63, 136.59 (d, *J* = 8.9 Hz), 136.20, 133.28 (d, *J* = 4.5 Hz), 133.16, 125.87, 119.01 (d, *J* = 4.2 Hz), 113.94 (d, *J* = 3.3 Hz), 107.52 (d, *J* = 26.4 Hz), 70.41, 68.95, 67.08, 51.14, 51.11, 47.20, 15.60.

^{19}F NMR (376 MHz, $CDCl_3$): δ -120.14.

HRMS (ESI-TOF): calculated for $[C_{20}H_{22}FN_3O_4+H]^+$: 388.1667, Found: 388.1670.



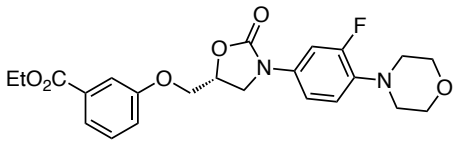
(*R*)-3-(3-fluoro-4-morpholinophenyl)-5-((quinolin-6-yloxy)methyl)oxazolidin-2-one (m5-p5): Prepared according to the general procedure on a 6.00 mmol scale. The title compound was recrystallized with dichloromethane/hexane to provide a pale red solid (411 mg, 970 μmol, 16 % yield).

^{1}H NMR (400 MHz, $CDCl_3$) δ 8.80 (dd, *J* = 4.3, 1.7 Hz, 1H), 8.11 – 8.01 (m, 2H), 7.48 (dd, *J* = 14.3, 2.6 Hz, 1H), 7.43 – 7.33 (m, 2H), 7.17 (ddd, *J* = 8.8, 2.7, 1.1 Hz, 1H), 7.11 (d, *J* = 2.8 Hz, 1H), 6.94 (t, *J* = 9.1 Hz, 1H), 5.05 (ddt, *J* = 8.9, 5.9, 4.6 Hz, 1H), 4.35 (d, *J* = 4.6 Hz, 2H), 4.20 (t, *J* = 8.9 Hz, 1H), 4.06 (dd, *J* = 8.9, 5.9 Hz, 1H), 3.88 – 3.85 (m, 4H), 3.07 – 3.04 (m, 4H).

^{13}C NMR (101 MHz, $CDCl_3$): δ 156.21, 155.67 (d, *J* = 246.6 Hz), 154.28, 148.41, 144.52, 136.66 (d, *J* = 8.9 Hz), 135.32, 133.23 (d, *J* = 10.6 Hz), 131.19, 129.23, 122.22, 121.75, 119.01 (d, *J* = 4.3 Hz), 114.06 (d, *J* = 3.1 Hz), 107.64 (d, *J* = 26.3 Hz), 106.66, 70.33, 68.29, 67.07, 51.15, 51.12, 47.54.

^{19}F NMR (376 MHz, $CDCl_3$): δ -120.12.

HRMS (ESI-TOF): calculated for $[C_{23}H_{22}FN_3O_4+H]^+$: 424.1667, Found: 424.1674.



ethyl (*R*)-3-((3-(3-fluoro-4-morpholinophenyl)-2-oxooxazolidin-5-yl)methoxy)benzoate (m5-p6): Prepared according to the general procedure on a 10.00 mmol scale. The title compound was isolated via silica gel chromatography (silica gel, 40-63 μm, *Silicycle*, 0-10% methanol/dichloromethane) as a white solid (173 mg, 390 μmol, 4% yield).

^{1}H NMR (400 MHz, $CDCl_3$): δ 7.70 (dt, *J* = 7.7, 1.3 Hz, 1H), 7.56 (dd, *J* = 2.7, 1.5 Hz, 1H), 7.49 (dd, *J* = 14.3, 2.6 Hz, 1H), 7.36 (t, *J* = 8.0 Hz, 1H), 7.16 (ddd, *J* = 8.8, 2.6, 1.1 Hz, 1H), 7.10 (ddd, *J* = 8.3, 2.7, 1.0 Hz, 1H), 6.99 (t, *J* = 9.0 Hz, 1H), 5.00 (ddt, *J* = 9.0, 5.9, 4.5 Hz, 1H), 4.38 (q, *J* = 7.1 Hz, 2H), 4.30 – 4.25 (m, 2H), 4.17 (t, *J* = 8.9 Hz, 1H), 4.02 (dd, *J* = 8.9, 5.9 Hz, 1H), 3.91 – 3.85 (m, 4H), 3.10 – 3.05 (m, 4H), 1.39 (t, *J* = 7.1 Hz, 3H).

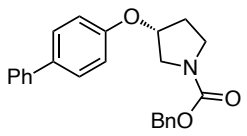
^{13}C NMR (101 MHz, $CDCl_3$): δ 166.29, 158.05, 155.73 (d, *J* = 246.6 Hz), 154.30, 132.22, 129.79, 123.22, 120.00, 119.12 (d, *J* = 3.5 Hz), 114.07 (d, *J* = 3.4 Hz), 107.68 (d, *J* = 26.4 Hz), 70.38, 68.22, 67.08, 61.35, 51.23, 51.20, 47.47, 14.46.

^{19}F NMR (376 MHz, $CDCl_3$): δ -120.14.

HRMS (ESI-TOF): calculated for $[C_{23}H_{25}FN_2O_6+H]^+$: 445.1697, Found: 445.1774.

m6 series:

Note: m6 series compounds exist as rotamers at room temperature. High temperature NMR was used to resolve rotameric peaks.

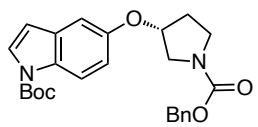


benzyl (*R*)-3-([1,1'-biphenyl]-4-yloxy)pyrrolidine-1-carboxylate (m6-p1): Prepared according to the general procedure. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 25% EtOAc/Heptane) as orange solid (343 mg, 920 μmol, 46% yield).

^{1}H NMR (700 MHz, 100 °C, DMSO-d6): δ 7.59 (dd, *J*=10.6, 8.1 Hz, 4H), 7.43 (t, *J*=7.8 Hz, 2H), 7.38 - 7.35 (m, 4H), 7.34 - 7.27 (m, 2H), 7.03 (d, *J*=8.8 Hz, 2H), 5.12 (s, 2H), 5.06 (dt, *J*=4.4, 2.3 Hz, 1H), 3.69 (dd, *J*=12.1, 4.6 Hz, 1H), 3.59 - 3.49 (m, 3H), 3.00 (s, 1H), 2.21 (dtd, *J*=13.5, 8.9, 4.9 Hz, 1H), 2.15 - 2.08 (m, 1H).

^{13}C NMR (176 MHz, 100 °C, DMSO-d6): δ 156.2, 153.7, 139.4, 136.7, 132.9, 128.2, 127.8, 127.3, 127.2, 126.9, 126.2, 125.7, 115.8, 75.7, 65.5, 51.0, 43.5, 30.1.

HRMS (ESI-TOF): calculated for $[C_{24}H_{23}NO_3+H]^+$: 374.1751, Found: 374.1753.



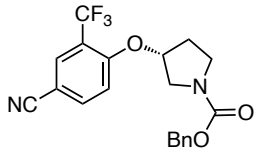
***tert*-butyl (*R*)-5-((1-((benzyloxy)carbonyl)pyrrolidin-3-yl)oxy)-1*H*-indole-1-carboxylate (m6-p2):** Prepared according to the general procedure. The title compound was

isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 25% EtOAc/Heptane) as orange solid (375 mg, 860 µmol, 43% yield).

^{1}H NMR (700 MHz, 100 °C , DMSO-d6): δ 7.97 (d, *J*=9.0 Hz, 1H), 7.61 (d, *J*=3.5 Hz, 1H), 7.38 - 7.33 (m, 4H), 7.33 - 7.26 (m, 1H), 7.16 (d, *J*=2.5 Hz, 1H), 6.94 (dd, *J*=9.0, 2.5 Hz, 1H), 6.60 (d, *J*=3.8 Hz, 1H), 5.14 - 5.09 (m, 2H), 5.01 (dt, *J*=4.6, 2.3 Hz, 1H), 3.66 (dd, *J*=12.0, 4.8 Hz, 2H), 3.59 - 3.50 (m, 2H), 2.21 -2.08 (m, 2H), 1.64 (s, 9H).

^{13}C NMR (176 MHz, 100 °C, DMSO-d6): δ 153.7, 152.6, 148.6, 136.7, 130.8, 129.5, 127.7, 127.1, 126.8, 126.2, 114.9, 113.9, 106.7, 106.4, 83.1, 76.3, 65.6, 51.0, 43.5, 30.1, 27.3.

HRMS (ESI-TOF): calculated for $[C_{25}H_{28}N_2O_5+H]^+$: 437.2071, Found: 437.2058.



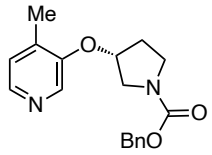
benzyl (*R*)-3-(4-cyano-2-(trifluoromethyl)phenoxy)pyrrolidine-1-carboxylate (m6-p3): Prepared according to the general procedure. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 25% EtOAc/Heptane) as white solid (314 mg, 800 µmol, 40% yield).

^{1}H NMR (700 MHz, 100 °C , DMSO-d6): δ 8.05 - 8.03 (m, 2H), 7.49 (d, *J*=8.5 Hz, 1H), 7.36 - 7.32 (m, 4H), 7.32 - 7.28 (m, 1H), 5.35 (br s, 1H), 5.10 (s, 2H), 3.71 (dd, *J*=12.5, 4.3 Hz, 1H), 3.60 - 3.54 (m, 2H), 3.49 - 3.43 (m, 1H), 2.27 (dtd, *J*=13.9, 9.3, 4.8 Hz, 1H), 2.16 - 2.10 (m, 1H).

^{13}C NMR (176 MHz, 100 °C, DMSO-d6): δ 157.7, 153.7, 137.9, 136.7, 131.0, 127.8, 127.2, 126.8, 122.1 (q, *J*=272.8 Hz), 119.0 (q, *J* =30.9 Hz), 117.1, 115.5, 103.3, 77.7, 65.6, 50.8, 43.4, 39.9, 30.1.

^{19}F NMR (126 MHz, DMSO-d$_6$): δ -61.90.

HRMS (ESI-TOF): calculated for $[C_{20}H_{17}F_3N_2O_3+H]^+$: 391.1264, Found: 391.1266.

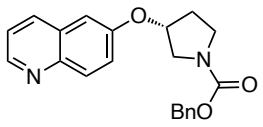


benzyl (*R*)-3-((4-methylpyridin-3-yl)oxy)pyrrolidine-1-carboxylate (m6-p4): Prepared according to the general procedure. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 25% EtOAc/Heptane) as orange solid (193 mg, 620 μmol, 31% yield).

^{1}H NMR (700 MHz, 100 °C, DMSO-d6): δ 8.27 (s, 1H), 8.10 (d, *J*=4.8 Hz, 1H), 7.35 (d, *J*=4.5 Hz, 4H), 7.33 - 7.27 (m, 1H), 7.16 (d, *J*=4.8 Hz, 1H), 5.13 -5.10 (m, 3H), 3.66 (dd, *J*=12.0, 4.5 Hz, 1H), 3.58 - 3.51 (m, 3H), 2.21 (dtd, *J*=13.6, 9.1, 4.8 Hz, 1H), 2.15 - 2.11 (m, 4H).

^{13}C NMR (176 MHz, 100 °C, DMSO-d6): δ 153.7, 151.5, 142.1, 136.7, 134.1, 130.2, 127.8, 127.2, 127.1, 126.8, 125.0, 76.8, 65.5, 51.0, 43.5, 30.3, 14.3, 12.1.

HRMS (ESI-TOF): calculated for $[C_{18}H_{20}N_2O_3+H]^+$: 313.1547, Found: 313.1543.

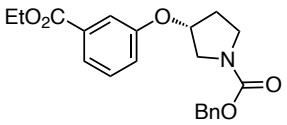


benzyl (*R*)-3-(quinolin-6-yloxy)pyrrolidine-1-carboxylate (m6-p5): Prepared according to the general procedure. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 25% EtOAc/Heptane) as orange solid (395 mg, 1.140 mmol, 57% yield).

^{1}H NMR (700 MHz, 100 °C, DMSO-d6): δ 8.75 (d, *J*=4.0 Hz, 1H), 8.20 (d, *J*=8.0 Hz, 1H), 7.95 (d, *J*=9.0 Hz, 1H), 7.43 (dd, *J*=8.3, 4.3 Hz, 1H), 7.40 (dd, *J*=9.3, 2.8 Hz, 1H), 7.38 - 7.33 (m, 5H), 7.32 - 7.27 (m, 1H), 5.18 (br s, 1H), 5.12 (s, 2H), 3.76 (dd, *J*=12.0, 4.8 Hz, 1H), 3.62 (d, *J*=12.3 Hz, 1H), 3.60 - 3.52 (m, 2H), 2.27 (dtd, *J*=13.6, 9.0, 4.9 Hz, 1H), 2.23 - 2.13 (m, 1H).

^{13}C NMR (176 MHz, 100 °C, DMSO-d6): δ 154.4, 153.7, 147.6, 143.7, 136.7, 134.1, 130.2, 128.5, 127.8, 127.1, 126.9,121.8, 121.0, 108.2, 76.0, 65.5, 51.0, 43.5, 30.1.

HRMS (ESI-TOF): calculated for $[C_{21}H_{20}N_2O_3+H]^+$: 349.1547, Found: 349.1550.



benzyl (*R*)-3-(3-(ethoxycarbonyl)phenoxy)pyrrolidine-1-carboxylate(m6-p6): Prepared according to the general procedure. The title compound was isolated via silica gel column chromatography (80 g ISCO RediSep-RfGold column, 25% EtOAc/Heptane) as clear colorless oil (240 mg, 650 μmol, 33% yield).

^{1}H NMR (700 MHz, 100 °C, DMSO-d6): δ 7.58 (d, J=7.8 Hz, 1H), 7.47 (s, 1H), 7.43 (t, J=8.0 Hz, 1H), 7.35 (d, J=4.3 Hz, 4H), 7.34 - 7.28 (m, 1H), 7.22 (dd, J=8.3, 2.3 Hz, 1H), 5.12 - 5.08 (m, 3H), 4.33 (q, J=7.0 Hz, 2H), 3.69 (dd, J=12.1, 4.6 Hz, 1H), 3.58 - 3.48 (m, 3H), 2.21 (dtd, J=13.6, 9.0, 4.9 Hz, 1H), 2.13 - 2.07 (m, 1H), 1.34 (t, J=7.0 Hz, 3H).

^{13}C NMR (176 MHz, 100 °C, DMSO-d6): δ 165.0, 156.6, 153.7, 136.7, 131.4, 129.4, 127.8, 127.1, 126.9, 121.4, 120.0, 115.8, 76.0, 65.5, 60.2, 50.9, 43.5, 30.0, 13.5.

HRMS (ESI-TOF): calculated for $[C_{21}H_{23}NO_5+H]^+$: 370.1649, Found: 370.1659.

2.6 References

1. Ruiz-Castillo, P. & Buchwald, S. L. Applications of palladium-catalyzed C–N cross-coupling reactions. *Chem. Rev.* **116**, 12564–12649 (2016).
2. Ogba, O. M., Warner, N. C., O'Leary, D. J. & Grubbs, R. H. Recent advances in ruthenium-based olefin metathesis. *Chem. Soc. Rev.* **47**, 4510–4544 (2018).
3. Kolb, H. C., VanNieuwenhze, M. S. & Sharpless, K. B. Catalytic asymmetric dihydroxylation. *Chem. Rev.* **94**, 2483–2547 (1994).
4. Chatterjee, S., Guidi, M., Seeberger, P. H. & Gilmore, K. Automated radial synthesis of organic molecules. *Nature* **579**, 379–384 (2020).
5. Echtermeyer, A., Amar, Y., Zakrzewski, J. & Lapkin, A. Self-optimisation and model-based design of experiments for developing a C–H activation flow process. *Beilstein J. Org. Chem.* **13**, 150–163 (2017).
6. Coley, C. W., Abolhasani, M., Lin, H. & Jensen, K. F. Material-efficient microfluidic platform for exploratory studies of visible-light photoredox catalysis. *Angew. Chem. Int. Ed.* **56**, 9847–9850 (2017).

7. Granda, J. M., Donina, L., Dragone, V., Long, D.-L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377–381 (2018).
8. Hsieh, H.-W., Coley, C. W., Baumgartner, L. M., Jensen, K. F. & Robinson, R. I. Photoredox iridium–nickel dual-catalyzed decarboxylative arylation cross-coupling: from batch to continuous flow via self-optimizing segmented flow reactor. *Org. Process Res. Dev.* **22**, 542–550 (2018).
9. Schweidtmann, A. M. *et al.* Machine learning meets continuous flow chemistry: automated optimization towards the Pareto front of multiple objectives. *Chem. Eng. J.* **352**, 277–282 (2018).
10. Burger, B. *et al.* A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
11. Häse, F., Aldeghi, M., Hickman, R. J., Roch, L. M. & Aspuru-Guzik, A. Gryffin: an algorithm for Bayesian optimization of categorical variables informed by expert knowledge. *Appl. Phys. Rev.* **8**, 031406 (2021).
12. Taylor, C. J. *et al.* Accelerated chemical reaction optimization using multi-task learning. *ACS Cent. Sci.* **9**, 957–968 (2023).
13. Zhou, Z., Li, X. & Zare, R. N. Optimizing chemical reactions with deep reinforcement learning. *ACS Cent. Sci.* **3**, 1337–1344 (2017).
14. Torres, J. A. G. *et al.* A multi-objective active learning platform and web app for reaction optimization. *J. Am. Chem. Soc.* **144**, 19999–20007 (2022).
15. Shields, B. J. *et al.* Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89–96 (2021).

16. Häse, F., Roch, L. M., Kreisbeck, C. & Aspuru-Guzik, A. Phoenics: a Bayesian optimizer for chemistry. *ACS Cent. Sci.* **4**, 1134–1145 (2018).
17. Clayton, A. D. *et al.* Algorithms for the self-optimisation of chemical reactions. *React. Chem. Eng.* **4**, 1545–1554 (2019).
18. Reker, D., Hoyt, E. A., Bernardes, G. J. L. & Rodrigues, T. Adaptive optimization of chemical reactions with minimal experimental information. *Cell Rep. Phys. Sci.* **1**, 100247 (2020).
19. Shim, E. *et al.* Predicting reaction conditions from limited data through active transfer learning. *Chem. Sci.* **13**, 6655–6668 (2022).
20. Gao, H. *et al.* Using machine learning to predict suitable conditions for organic reactions. *ACS Cent. Sci.* **4**, 1465–1476 (2018).
21. Kozlowski, M. C. On the topic of substrate scope. *Org. Lett.* **24**, 7247–7249 (2022).
22. Gensch, T. & Glorius, F. The straight dope on the scope of chemical reactions. *Science* **352**, 294–295 (2016).
23. Dreher, S. D. Catalysis in medicinal chemistry. *React. Chem. Eng.* **4**, 1530–1535 (2019).
24. Kariofillis, S. K. *et al.* Using data science to guide aryl bromide substrate scope analysis in a Ni/photoredox-catalyzed cross-coupling with acetals as alcohol-derived radical sources. *J. Am. Chem. Soc.* **144**, 1045–1055 (2022).
25. Dreher, S. D. & Krska, S. W. Chemistry informer libraries: conception, early experience, and role in the future of cheminformatics. *Acc. Chem. Res.* **54**, 1586–1596 (2021).
26. Collins, K. D. & Glorius, F. A robustness screen for the rapid assessment of chemical reactions. *Nat. Chem.* **5**, 597–601 (2013).
27. Kullmer, C. N. P. *et al.* Accelerating reaction generality and mechanistic insight through additive mapping. *Science* **376**, 532–539 (2022).

28. Wagen, C. C., McMinn, S. E., Kwan, E. E. & Jacobsen, E. N. Screening for generality in asymmetric catalysis. *Nature* **610**, 680–686 (2022).
29. Rein, J. *et al.* Generality-oriented optimization of enantioselective aminoxyl radical catalysis. *Science* **380**, 706–712 (2023).
30. Betinol, I. O., Lai, J., Thakur, S. & Reid, J. P. A data-driven workflow for assigning and predicting generality in asymmetric catalysis. *J. Am. Chem. Soc.* **145**, 12870–12883 (2023).
31. Kim, H. *et al.* A multi-substrate screening approach for the identification of a broadly applicable Diels–Alder catalyst. *Nat. Commun.* **10**, 770 (2019).
32. Angello, N. H. *et al.* Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling. *Science* **378**, 399–405 (2022).
33. Slivkins, A. Introduction to multi-armed bandits. Preprint at https://arxiv.org/abs/1904.07272 (2019).
34. Taylor, C. J. *et al.* A brief introduction to chemical reaction optimization. *Chem. Rev.* **123**, 3089–3126 (2023).
35. Russo, D., Roy, B. V., Kazerouni, A., Osband, I. & Wen, Z. A tutorial on Thompson sampling. Preprint at https://arxiv.org/abs/1707.02038 (2017).
36. Kaufmann, E., Cappe, O. & Garivier, A. On Bayesian upper confidence bounds for bandit problems. *Proceedings of Machine Learning Research* **22**, 592–600 (2012).
37. Snoek, J. *et al.* Scalable Bayesian optimization using deep neural networks. Preprint at https://arxiv.org/abs/1502.05700 (2015).
38. Stevens, J. M. *et al.* Advancing base metal catalysis through data science: insight and predictive models for Ni-catalyzed borylation through supervised machine learning. *Organometallics* **41**, 1847–1864 (2022).

39. Nielsen, M. K., Ahneman, D. T., Riera, O. & Doyle, A. G. Deoxyfluorination with sulfonyl fluorides: navigating reaction space with machine learning. *J. Am. Chem. Soc.* **140**, 5004–5008 (2018).

40. Lin, S. *et al.* Mapping the dark space of chemical reactions with extended nanomole synthesis and MALDI-TOF MS. *Science* **361**, eaar6236 (2018).

41. 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* **360**, 186–190 (2018).

42. Brown, D. G. & Boström, J. Analysis of past and present synthetic methodologies on medicinal chemistry: where have all the new reactions gone? *J. Med. Chem.* **59**, 4443–4458 (2016).

43. El-Faham, A. & Albericio, F. Peptide coupling reagents, more than a letter soup. *Chem. Rev.* **111**, 6557–6602 (2011).

44. Dombrowski, A. W., Aguirre, A. L., Shrestha, A., Sarris, K. A. & Wang, Y. The chosen few: parallel library reaction methodologies for drug discovery. *J. Org. Chem.* **87**, 1880–1897 (2022).

45. Jaeger, S., Fulle, S. & Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model* **58**, 27–35 (2018).

46. Magano, J. Large-scale amidations in process chemistry: practical considerations for reagent selection and reaction execution. *Org. Process Res. Dev.* **26**, 1562–1689 (2022).

47. Beutner, G. L. *et al.* TCFH–NMI: direct access to N-acyl imidazoliums for challenging amide bond formations. *Org. Lett.* **20**, 4218–4222 (2018).

48. Sperry, J. B. *et al.* Thermal stability assessment of peptide coupling reagents commonly used in pharmaceutical manufacturing. *Org. Process. Res. Dev.* **22**, 1262–1275 (2018).

49. Stevens, J. M. *et al.* Leveraging high-throughput experimentation to drive pharmaceutical route invention: a four-step commercial synthesis of branebrutinib (BMS-986195). *Org. Process Res. Dev.* **26**, 1174-1183 (2022).
50. Zheng, B. *et al.* Preparation of the HIV attachment inhibitor BMS-663068. part 6. Friedel–Crafts acylation/hydrolysis and amidation. *Org. Process Res. Dev.* **21**, 1145–1155 (2017).
51. Krishnan, K. K., Ujwaldev, S. M., Sindhu, K. S. & Anilkumar, G. Recent advances in the transition metal catalyzed etherification reactions. *Tetrahedron* 72, 7393–7407 (2016).
52. Fuhrmann, E. & Talbiersky, J. Synthesis of alkyl aryl ethers by catalytic Williamson ether synthesis with weak alkylation agents. *Org. Process Res. Dev.* 9, 206–211 (2005).
53. Swamy, K. C. K., Kumar, N. N. B., Balaraman, E. & Kumar, K. V. P. P. Mitsunobu and related reactions: advances and applications. *Chem. Rev.* 109, 2551–2651 (2009).
54. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: an Introduction*. (MIT Press, 2018).
55. Thathachar, M. A. L. & Sastry, P. S. A new approach to the design of reinforcement schemes for learning automata. *IEEE Trans. Syst., Man, Cybern.* **SMC-15**, 168–175 (1985).
56. Auer, P., Cesa-Bianchi, N. & Fischer, P. Finite-time Analysis of the Multiarmed Bandit Problem. *Mach. Learn.* **47**, 235–256 (2002).
57. Audibert, J.-Y., Munos, R. & Szepesvári, C. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.* **410**, 1876–1902 (2009).
58. Honda, J. & Takemura, A. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Mach. Learn.* **85**, 361–391 (2011).
59. Gelman, A. *et al. Bayesian Data Analysis*. (Chapman and Hall, 2004).
60. Agrawal, S. & Goyal, N. Near-Optimal Regret Bounds for Thompson Sampling. *J. ACM (JACM)* **64**, 30 (2017).

61. DeGroot, M. H. *Optimal Statistical Decisions*. (John Wiley & Sons, 2004).

62. Kaufmann, E., Cappe, O. & Garivier, A. On Bayesian Upper Confidence Bounds for Bandit Problems. in *Proceedings of Machine Learning Research* vol. 22 592–600 (2012).

63. Nielsen, M. K., Ahneman, D. T., Riera, O. & Doyle, A. G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **140**, 5004–5008 (2018).

64. Angello, N. H. *et al.* Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling. *Science* **378**, 399–405 (2022).

65. Taylor, C. J. *et al.* Accelerated Chemical Reaction Optimization Using Multi-Task Learning. *ACS Cent. Sci.* **9**, 957–968 (2023).

66. Stevens, J. M. *et al.* Advancing Base Metal Catalysis through Data Science: Insight and Predictive Models for Ni-Catalyzed Borylation through Supervised Machine Learning. *Organometallics* **41**, 1847–1864 (2022).

67. 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* **360**, 186–190 (2018).

68. Santanilla, A. B. *et al.* Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49–53 (2015).

69. https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.savgol_filter.html.

70. Lin, S. *et al.* Mapping the dark space of chemical reactions with extended nanomole synthesis and MALDI-TOF MS. *Science* **361**, (2018).

71. Jaeger, S., Fulle, S. & Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model* **58**, 27–35 (2018).

72. Schmid, S. P. et al. If optimizing for general parameters in chemistry is useful, why is it hardly done? https://openreview.net/forum?id=ZfL0poiEOe (2024).

73. Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I. & Zrnic, T. Prediction-powered inference. *Science* **382**, 669–674 (2023).

Chapter 3. Diversification of acridinium photocatalysts: property tuning and reactivity in model reactions

3.1 Introduction

In photoredox catalysis, late transition-metal (e.g., iridium, ruthenium) polypyridyl complexes have been commonly employed as photocatalysts, due to their favorable excited redox properties, enhanced photostability and excited state lifetime. However, the high cost and low abundance of these metals have prompted the discovery and application of organic photocatalysts. Organic photocatalysts (OPCs) usually have extended conjugated systems and absorb visible light to reach excited states that can also engage in photoredox catalysis. 9-Mesityl-3,6-di-*tert*-butyl-10-phenylacridinium salt (**1**) and its derivatives have found wide applications in synthetic transformations such as nucleophilic arene and alkene functionalization.[1] A modular synthesis that is amenable to late-stage functionalization[2–8] has enabled access to diverse acridinium derivatives. However, studies on the comparison of their catalytic performances under various reduction and oxidation manifolds have been limited. Using **1** as a template structure, we set out to examine the effects of structural changes on the photophysical properties and various reactivities of acridinium photocatalysts.

3.2 Results and discussions

3.2.1 Underexplored *N*-substitutions for acridinium photocatalysts

Our investigation started by identifying underexplored structural motifs for derivatives of **1**. Prior reports suggest that modifications to the acridinium core do not significantly change the redox potentials of the catalysts (*vide infra*). On the other hand, modifications to the *N*-substituent, while also having little effect on redox potentials, can lead to meaningful changes in excited state

lifetimes. Based on these observations, we synthesized a library of previously unknown catalysts with various *N*-substituents (aryl, heteroaryl, benzyl, alkyl) from xanthylium salts and commercially available amines (**Fig. 128b**).

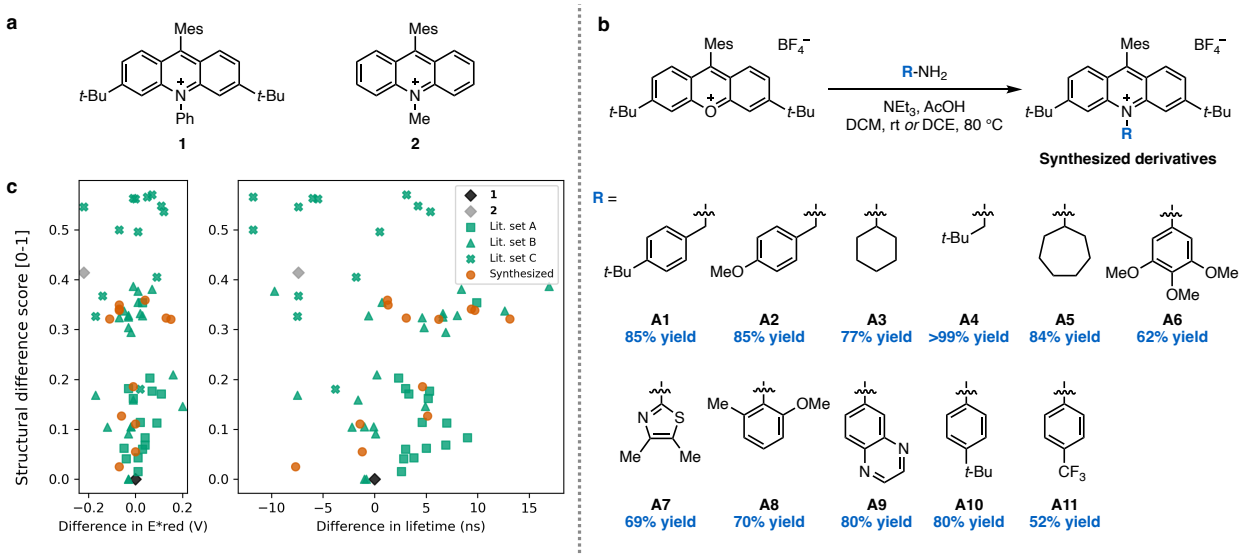


Fig. 128 Syntheses of acridinium photocatalysts and the comparison of photophysical properties.

We characterized the photophysical properties of the synthesized catalysts (**Table 4**), focusing on excited-state reduction potential ($E^*_{red} = E_{0,0} + E_{red}$, vs. SCE) and excited-state lifetime (τ). For comparison, we also compiled properties of acridinium catalysts in three prior reports [2,6,8] that can also be accessed via the same synthetic procedure. To visualize the effect of structural modifications on E^*_{red} and τ, we also defined a structural difference score. This score is calculated as 1 minus the Tanimoto similarity[9] of RDKit fingerprints[10] for a synthesized catalyst and **1**. The higher the structural difference score, the more different a given catalyst structure is compared to **1**. Both synthesized and reported acridinium catalysts are shown as comparison (**Fig. 128c**). Compared to **1** (E^*_{red} = +2.10 V, τ = 13.8 ns[2]), structural modifications have little impact on E^*_{red} (± 0.2 V), while significant changes to excited state lifetimes of over ±10 ns can often be observed.

Compared to derivatives reported in literature, our library features simple changes of *N*-substituents but cover a broad range of property differences.

Catalyst label*	E_{red} (V)	$E_{0,0}$ (eV)	E^*_{red} (V)	τ (ns)
1[a]	-0.56	+2.66	+2.10	13.8
2[b]	-0.49	+2.37	+1.88	6.4
A1	-0.61	+2.64	+2.03	15.09
A2	-0.59	+2.73	+2.14	0.11, 4.41, 15.02
A3	-0.61	+2.64	+2.03	23.52
A4	-0.64	+2.63	+1.99	26.91
A5	-0.63	+2.66	+2.03	23.18
A6	-0.63	+2.67	+2.04	0.15, 4.93, 18.93
A7	-0.45	+2.68	+2.23	0.3, 8.53, 16.85
A8	-0.59	+2.68	+2.09	0.26, 8.85, 18.45
A9	-0.56	+2.66	+2.10	12.38
A10	-0.62	+2.65	+2.03	6.14
A11	-0.56	+2.66	+2.10	12.58

Table 4 Experimentally determined photophysical properties of synthesized acridinium photocatalysts.

*Detailed characterizations and methods of determination can be found in the Supplementary Information. E_{red}: ground state reduction potential, or $E_{1/2}(C/C^-)$, vs. SCE; $E_{0,0}$: Excited-state energy; E^*_{red}: excited-state reduction potential, or $E_{1/2}(C^*/C^-)$, vs. SCE; τ: fluorescence lifetime.
[a] Catalyst properties extracted from prior report.[2]
[b] Catalyst properties extracted from prior reports.[11,12]

To examine the effect of various *N*-substituents on the excited-state nature of acridinium photocatalysts, we performed time-dependent density functional theory (TD-DFT) calculations on optimized S_1 geometry of synthesized acridiniums (see Section 3.4 for more details). These calculations indicate electron transitions from mesityl group or the *N*-substituent to the acridinium core in the excited state. These orbitals have minimal spatial orbital overlap, suggesting that S_1 is an intramolecular charge-transfer (CT) state. We also observed a significant increase in dipole moments from S_0 to S_1 optimized geometries, further supporting the characterization of S_1 as a CT state.

3.2.2 Acridiniums in a S_NAr reaction

Established acridinium photocatalysts such as **1** and **2** are often employed as strong excited-state oxidants. Thus, we decided to test our library of novel acridinium photocatalysts under an oxidative manifold, in the S_NAr reaction of anisoles, as reported by Nicewicz and coworkers.[13] Given that the coupling combination of anisole with 1,3-imidazole was reported to occur in a modest 39% yield with **1**, we were curious to observe how our library of catalysts with modified *N*-substitutions would perform (**Fig. 129a**). Unsurprisingly, catalyst **A10**, being most structurally and electronically like **1**, enabled the reaction to proceed with similar yields. Surprisingly, photocatalysts **A4** and **A5** with alkyl substitution performed the next highest out of our acridinium library, though did not outperform **1**.

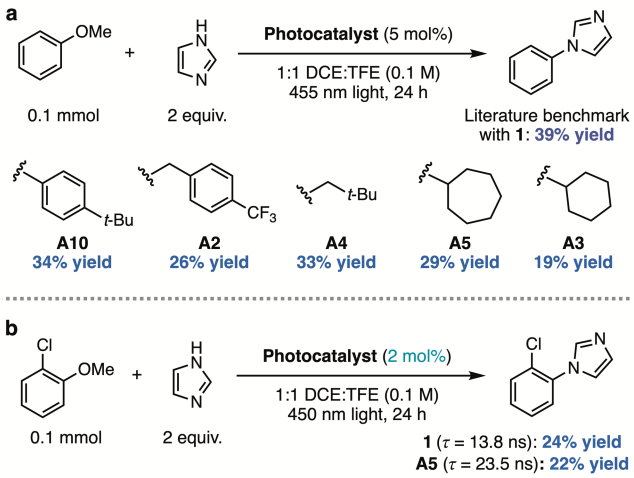


Fig. 129 S_NAr reaction of anisoles with imidazole catalyzed by various acridinium photocatalysts.

Thus, we were interested to observe whether we could exploit the significantly longer lifetime of *N*-alkyl substituted acridiniums by lowering the catalyst loading required. We performed a head-to-head comparison between **A5** and **1** in the reaction of imidazole with *ortho*-chloroanisole, which was reported to proceed with high yield. We found that at a decreased 2 mol% catalyst loading, both acridinium photocatalysts performed comparably to each other (**Fig. 129b**).

Although the overall yield was much lower than with the standard reaction conditions, our results imply that the longer excited state lifetime of *N*-alkyl acridinium photocatalysts could potentially be leveraged in other reactions.

It is also worth noting that all our tested acridinium photocatalysts were competent in this reaction. This finding led us to wonder about the significance of the *N*-phenyl substitution **1**. Therefore, we decided to next screen our library in a different reaction that mechanistically emphasizes its importance (*vide infra*).

3.2.3 Acridiniums in photo-debromination reaction

Nicewicz and coworkers also reported **1** to be a highly competent photocatalyst in a series of dehalogenation reactions, in which the catalyst acts as a super-reductant capable of reducing various aryl halides ($E^*_{ox} = -3.36$ V).[14] Mechanistically, the acridine radical (generated from single electron reduction of the excited state acridinium) can be further irradiated to a twisted intramolecular charge transfer (TICT) state that is a potent reductant. This TICT state was proposed to involve the radical anion being localized on the *N*-phenyl ring, highlighting the importance of this substitution.

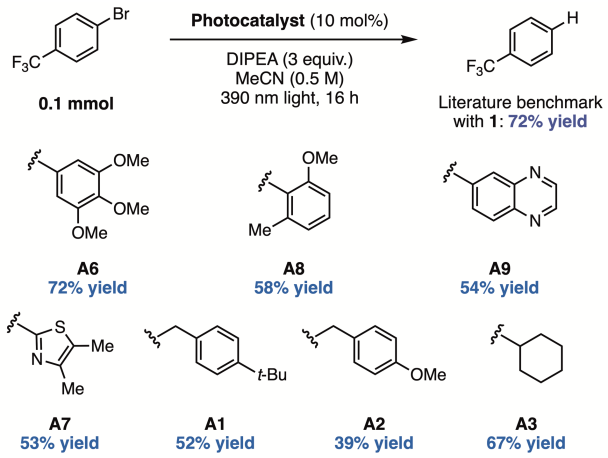


Fig. 130 Reductive debromination reaction catalyzed by various acridinium photocatalysts.

We decided to screen our library of novel acridinium photocatalysts in the dehalogenation of 4-(trifluoromethyl)phenyl bromide (**Fig. 130**). Our initial hypothesis was that *N*-aryl acridiniums would be competent in this reaction, while *N*-alkyl acridiniums, being unable to access the TICT state due to the lack of π-conjugation, would give comparatively lower or no yield. As expected, *N*-aryl substituted acridiniums were competent catalysts in this reaction (**A6**, **A8**, **A9**). However, to our surprise, *N*-heterocyclic (**A7**), *N*-benzyl (**A1**, **A2**), and *N*-alkyl (**A3**) photocatalysts were similarly competent in this reaction, with *N*-cyclohexyl acridinium **A3** giving a yield within error of that reported for **1**. This finding suggests that either the ground state acridine radical of this photocatalyst is itself the reductant, or that its most accessible TICT state would involve the radical anion localized on the mesityl ring of the acridinium core. To test our hypothesis, we performed preliminary TD-DFT calculations on the optimized ground state (D_0) geometry of acridine radical to explore the nature of electronic transitions for the low-lying excited state (D_1). For *N*-alkyl acridiniums (**A3**, **A4**, **A5**), the electronic transition occurs from π_{core} to π^*_{core} with high orbital overlap, indicating that the D_1 is a locally excited state. The character of D_1 in *N*-aryl acridiniums, however, depends on the aryl substituent. Acridiniums with electron-rich aryl substituents (**A1**, **A2**, **A6**, **A8**, **A10**) exhibit electronic transition from π_{core} to an orbital localized

on the core and *N*-substituent, suggesting that D_1 is a mixed local and charge-transfer-state. Conversely, acridiniums with electron-poor aryl substituents (**A7**, **A9**, **A11**), shows electronic transitions from π_{core} to $\pi^*_{N\text{-substituent}}$ and have minimal orbital overlap, indicating that the D_1 is an intramolecular charge-transfer state. Details on the vertical absorption energy and orbitals are available in Section 3.4. Further photophysical studies are still needed to elucidate the nature of the photocatalysts in the photo-reduction reaction.

3.2.4 Acridiniums in C–H amination

The Doyle lab previously reported a novel cyanoarene photocatalyst, CF_3-4-CzIPN, that can engage in oxidative radical-polar crossover (ORPC) to achieve nucleophilic amination of primary and secondary benzylic C(sp^3)–H bonds.[15] Compared to commonly employed 4-CzIPN (E^*_{red} = +1.43 V), cyanoarenes with a high E^*_{red}, represented by CF_3-4-CzIPN (E^*_{red} = +1.91 V), are most beneficial for this reaction by more readily oxidizing the benzylic radical to a carbocation in the radical-polar crossover step. In addition to an extensive screening of cyanoarenes, we have also previously screened acridinium catalyst **2** (E^*_{red} = +1.88 V) with a high throughput photoreactor set up and found that it is effective in catalyzing the amination reaction.[15] Based on these observations, we hypothesized that more optimal acridinium catalysts could be identified for this reaction.

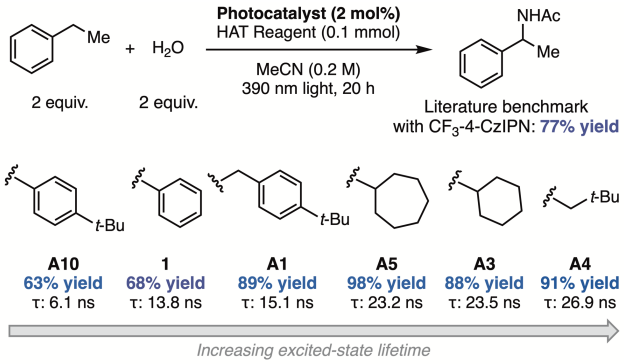


Fig. 131 Nucleophilic C-H amination reaction catalyzed by various acridinium photocatalysts, including two benchmark catalysts CF_3-4-CzIPN and **1**.

On the hypothesis that longer excited state lifetimes will also be beneficial to the reactivity of Ritter amination reaction, we tested *N*-alkyl and *N*-benzyl acridinium catalysts and compared their reactivities with the previously reported best cyanoarene catalyst, CF_3-4-CzIPN, and the commonly used acridinium catalyst **1** (**Fig. 131**). We started first by switching light source from the originally reported 456 nm to a 390 nm Kessil lamp to better match the maximum absorption wavelength of our synthesized acridiniums. Under otherwise identical reaction conditions, we found that *N*-alkyl and *N*-benzyl acridinium catalysts with longer excited-state lifetimes significantly outperformed both **1** (68% yield) and CF_3-4-CzIPN (77% yield). Through this optimization we identified catalyst **A5** as most optimal, which provides the desired product in a near-quantitative 98% yield. As a test of our hypothesis, we also screened *N*-aryl catalyst **A10**, which has a relatively short excited-state lifetime of 6.1 ns and found that the yield decreased significantly.

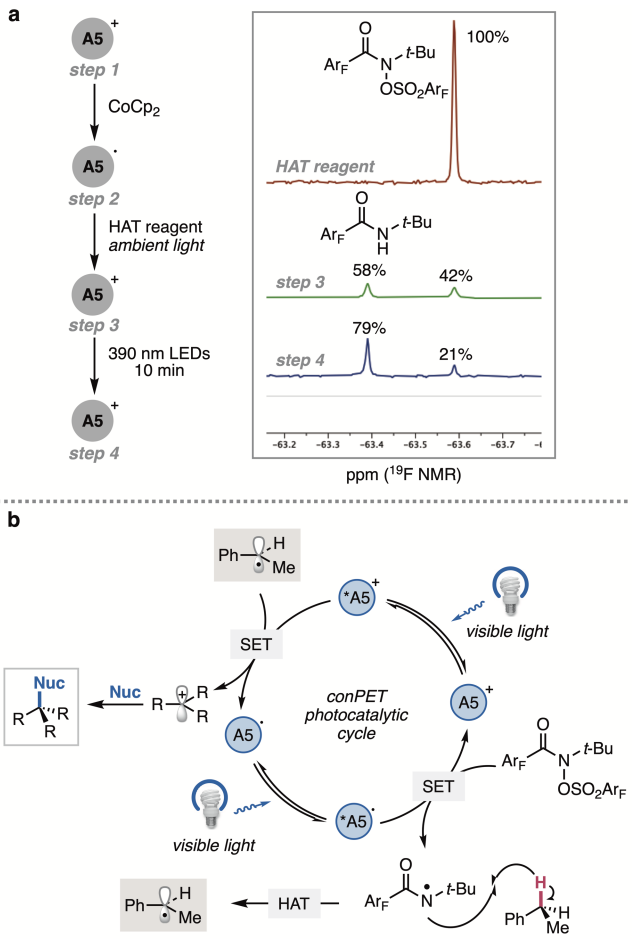


Fig. 132 Mechanistic study on HAT reagent consumption with acridinium photocatalyst **A5** and proposed catalytic cycle.

Following a previously-reported protocol,[12] we subjected a solution of catalyst **A5** to an excess of single-electron reductant cobaltocene to generate its reduced acridine species, **A5**$^{\bullet-}$ (**Fig. 132a**, step 1). Upon addition of this reduced photocatalytic intermediate to a solution of HAT reagent in the absence of light, two signals are observed in the ^{19}F NMR (**Fig. 132a**, step 3). The right peak indicates residual HAT reagent in the reaction mixture. Notably, formation of a new peak (**Fig. 132a**, step 3, left) is observed, consistent with formation and fragmentation of the reduced HAT species, generating a new HAT-derived byproduct. Moreover, when this mixture is then irradiated with a 390 nm light source, we observe further conversion of the HAT reagent to its fragmented byproduct (**Fig. 132a**, step 4). These experiments indicate the photocatalyst is likely

undergoing a reductive quenching cycle to generate the reduced acridine radical species, which is then responsible for single electron transfer (SET) to the HAT reagent, prompting mesolytic fragmentation. Increased conversion of the HAT reagent upon irradiation also suggests that a conPET mechanism may be in effect, wherein the reduced acridine radical undergoes a second photoexcitation event to promote the subsequent SET to the HAT reagent from its photoexcited state (**Fig. 132b**). We hypothesize that the increased excited-state lifetimes of the *N*-alkyl and *N*-benzyl acridiniums increases the kinetic efficiency of the SET steps in the catalytic cycle. Moreover, the high conformational flexibility of catalyst **A5** may help to prevent back-electron transfer from the acridine radical intermediate, a common challenge of many ORPC reactions.[16,17] More in-depth mechanistic and computational studies of these novel photocatalysts are needed to elucidate the exact nature of their increased reactivity.

3.3 Conclusions and outlooks

We have synthesized a library of acridinium photocatalysts featuring underexplored *N*-(hetero)aryl, *N*-benzyl and *N*-alkyl substitutions. We observed a significant effect of *N*-substitutions on the excited-state lifetimes of acridinium photocatalysts. In addition to being competent catalysts in test reactions featuring various oxidative and reductive pathways, the extended excited-state lifetimes have been shown to improve the reactivities of an ORPC reaction, providing new mechanistic insights and future directions in acridinium photocatalyst design.

We have also conducted preliminary studies that might serve as possible directions to explore novel organic photocatalysts further. Previously, we have targeted cyanoarenes as a class of widely used and versatile organic photocatalysts. Cyanoarenes, represented by 4CzIPN, have more balanced excited-state redox properties than acridiniums and can serve as both excited-state oxidant and excited-state reductant. We planned to explore cyanoarenes by the following *in silico*

workflow: 1) Generate virtual libraries of cyanoarene photocatalysts that have not been synthesized by enumerating substituents on the arene core; 2) Calculate photophysical properties (ground state and excited state redox potentials) via TD-DFT; 3) Use calculated photophysical properties to select and synthesize candidate molecules; 4) Test synthesized catalysts in model reactions and determine the correlation between photophysical properties and reactivities; 5) Select additional catalysts from the virtual catalyst library to synthesize and test.

We ran into several challenges with this proposal. First off, TD-DFT cannot reliably and accurately calculate photophysical properties for cyanoarenes. Many calculations time out after days and errors +/- 0.5 V in redox potentials can typically be observed. One potential solution to this problem might be to take a machine learning approach and train a prediction model (on a small dataset of catalysts and their TD-DFT-calculated properties) that can more accurately predict redox potentials faster without relying on TD-DFT. The second issue is that TD-DFT cannot calculate other properties that might also influence reactivities, such as excited-state lifetime. A machine learning approach can also be used in this case to mitigate this issue, although sufficient training data is required and can be difficult to obtain. The last issue is that cyanoarenes can be difficult to synthesize via the traditional S_NAr approach with fluoroarenes and amines. Even simple substituent changes can sometimes significantly alter the reactivity and result in a complex mixture of various products. The proposed approach is not practical without the ability to reliably synthesize selected candidate catalyst molecules, and better synthetic methods to access cyanoarenes (and other organic photocatalysts) are still in need of development.

3.4 Experimental section

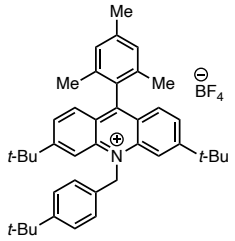
3.4.1 General information

^{1}H nuclear magnetic resonance (NMR) characterization was performed on 400, 500, and 600 MHz spectrometers (101 and 126 MHz for ^{13}C NMR). Chemical shifts for protons are reported in parts per million (ppm) downfield from tetramethylsilane and are referenced to residual protium in the NMR solvent ($CHCl_3$ = 7.26 ppm). Chemical shifts for carbon are reported in parts per million downfield from tetramethylsilane and are referenced to the carbon resonance of the solvent peak ($CDCl_3$ = 77.16 ppm). NMR data are represented as follows: chemical shift (δ ppm), multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, p = pentet, hept = heptet, m = multiplet, br = broad), coupling constant (*J*) in Hertz (Hz). All NMR spectra were taken at 25 °C. High resolution mass spectra were obtained using a Thermo Scientific Thermo Exactive Plus MSD (DART-MS) equipped with an ID-CUBE ion source and a Vapur Interface (Ion Sense Inc.) (atmospheric-pressure chemical ionization, APCI). Ultraviolet-Visible spectroscopy (UV-Vis) was performed with a Shimadzu UV-3101PC spectrophotometer at a sample concentration of 100 μM (MeCN). Fluorescence emission spectra were obtained with a Photon Technologies International QuantaMaster Spectrofluorimeter with 420 nm excitation lights at a sample concentration of 10 μM (MeCN). Time-correlated Single Photon Counting was done via a Horiba FluoroMax Plus Spectrofluorometer at a sample concentration of 5 μM (MeCN). All spectrophotometric samples were prepard in a N_2 glovebox with degassed solvent into a FireflySci Type 41 UV quartz macro cuvette with screw cap (lightpath=10 mm). Cyclic voltammetry (CV) experiments were obtained with a Gamry Interface 1010 Potentiostat/Galvanostat/ZRA instrument and processed using Gamry Echem FrameworkTM and AnalystTM software (working electrode: glassy carbon; reference electrode: Ag/Ag^+; counter electrode: Pt wire; scan rate: 0.1 V/s; sample concentration:

1 mM). All measurements were taken in degassed MeCN with NBu_4PF_6 (0.1 M) as electrolyte at 298 K. Ground state reduction potentials (E_{red}) were identified as half of the absolute maximum current value during the reduction event. Ferrocene (Fc) was used as an internal standard or an external standard. When Fc was used as an internal standard, ferrocene was added into the sample solution and one CV was performed. When Fc was used as an external standard, cyclic voltammetry (CV) was performed with ferrocene only under the same experimental conditions, and its $E_{1/2}$ (vs. Ag/AgCl) was recorded. $E_{1/2}$ = 0.4 V (vs. SCE) for Fc/Fc^+ is used for conversion in this paper.

3.4.2 Syntheses and characterization of acridinium photocatalysts

All acridinium photocatalysts were prepared from xanthylium and amine according to literature precedent.[18] Photocatalysts **A7**, **A8**, **A9**, and **A11** were synthesized using 1,2-dichloroethane as the solvent, at a temperature of 80 °C. All acridinium photocatalysts were bright yellow solids.



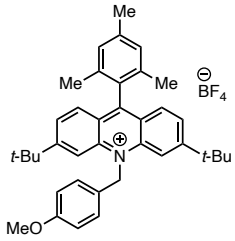
3,6-di-*tert*-butyl-10-(4-(*tert*-butyl)benzyl)-9-mesitylacridin-10-ium tetrafluoroborate (A1).

^{1}H NMR (400 MHz, CD_3CN): δ 8.28 (d, *J* = 1.6 Hz, 2H), 7.93 (dd, *J* = 9.0, 1.6 Hz, 2H), 7.76 (d, *J* = 9.1 Hz, 2H), 7.53 – 7.40 (m, 2H), 7.32 (d, *J* = 8.6 Hz, 2H), 7.24 (s, 2H), 6.55 (s, 2H), 2.48 (s, 3H), 1.75 (s, 6H), 1.40 (s, 18H), 1.29 (s, 9H).

^{13}C NMR (151 MHz, CD_3CN): δ 165.0, 162.4, 152.8, 142.9, 141.1, 136.9, 132.4, 130.6, 129.7, 129.7, 128.4, 127.3, 127.2, 125.4, 115.0, 54.9, 37.6, 35.2, 31.4, 30.6, 21.3, 20.0.

^{19}F NMR (376 MHz, CD_3CN): δ -151.77, -151.82.

HRMS (APCI): calculated for $C_{41}H_{50}N$ ($[M]^+$): 556.3938, found 556.4026.



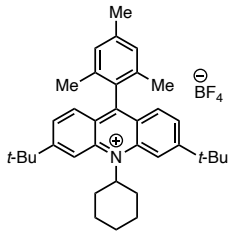
3,6-di-*tert*-butyl-9-mesityl-10-(4-methoxybenzyl)acridin-10-ium tetrafluoroborate (A2).

^{1}H NMR (400 MHz, CD_3CN): δ 8.27 (d, *J* = 1.6 Hz, 2H), 7.93 (dd, *J* = 9.1, 1.6 Hz, 2H), 7.75 (d, *J* = 9.0 Hz, 2H), 7.35 – 7.28 (m, 2H), 7.24 (s, 2H), 7.01 – 6.93 (m, 2H), 6.51 (s, 2H), 3.78 (s, 3H), 2.47 (s, 3H), 1.75 (s, 6H), 1.41 (s, 18H).

^{13}C NMR (151 MHz, CD_3CN) δ 165.0, 162.4, 160.8, 142.9, 141.1, 136.9, 130.6, 129.7, 129.7, 128.7, 128.4, 127.1, 125.4, 115.6, 114.9, 56.0, 54.7, 37.6, 30.6, 21.3, 20.0.

^{19}F NMR (376 MHz, CD_3CN): δ -151.81, -151.87.

HRMS (APCI): calculated for $C_{38}H_{44}NO$ ($[M]^+$): 530.3417, found 530.3502.



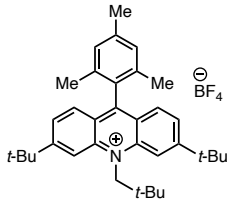
3,6-di-*tert*-butyl-10-cyclohexyl-9-mesitylacridin-10-ium tetrafluoroborate (A3).

^{1}H NMR (500 MHz, $CDCl_3$): δ 8.46 (s, 2H), 7.79 – 7.70 (m, 4H), 7.12 (s, 2H), 5.75 (tt, *J* = 12.7, 3.8 Hz, 1H), 2.84 (qd, *J* = 12.6, 3.8 Hz, 2H), 2.46 (s, 3H), 2.25 (d, *J* = 13.4 Hz, 2H), 2.05 (d, *J* = 13.8 Hz, 1H), 1.85 (qt, *J* = 13.2, 3.6 Hz, 2H), 1.74 (s, 6H), 1.54 (s, 18H), 1.52 – 1.46 (m, 1H).

^{13}C NMR (126 MHz, $CDCl_3$) δ 162.66, 161.57, 142.05, 140.26, 136.06, 129.51, 129.06, 127.14, 125.00, 114.76, 67.07, 37.11, 31.72, 30.62, 26.70, 25.72, 21.46, 20.24.

^{19}F NMR (376 MHz, CD_3CN): δ -151.84, -151.90.

HRMS (APCI): calculated for $C_{36}H_{46}N$ ($[M]^+$): 492.3625, found 492.3699.



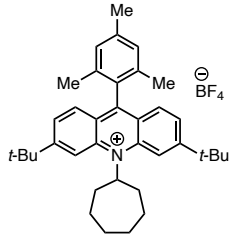
3,6-di-*tert*-butyl-9-mesityl-10-neopentylacridin-10-ium tetrafluoroborate (A4).

^{1}H NMR (500 MHz, $CDCl_3$): δ 8.58 (d, *J* = 1.5 Hz, 2H), 7.75 (dd, *J* = 9.0, 1.4 Hz, 2H), 7.71 (d, *J* = 9.0 Hz, 2H), 7.18 – 7.11 (m, 2H), 5.72 (s, 2H), 2.48 (s, 3H), 1.74 (d, *J* = 36.9 Hz, 6H), 1.53 (s, 18H), 1.09 (s, 9H).

^{13}C NMR (126 MHz, $CDCl_3$) δ 162.76, 160.33, 142.68, 140.19, 136.48, 135.38, 129.86, 129.21, 128.95, 128.48, 127.16, 124.34, 116.35, 77.16, 57.36, 37.22, 36.08, 30.73, 30.04, 21.43, 20.25, 20.04. Peaks split due to the presence of N–C rotamers, compound was unstable to high temperature NMR.

^{19}F NMR (282 MHz, CD_3CN): δ -151.82, -151.87.

HRMS (APCI): calculated for $C_{35}H_{46}N$ ($[M]^+$): 480.3625, found 480.3697.



3,6-di-*tert*-butyl-10-cycloheptyl-9-mesitylacridin-10-ium tetrafluoroborate (A5).

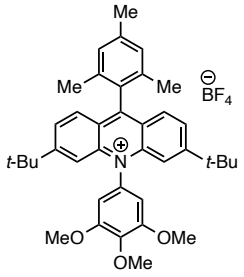
^{1}H NMR (500 MHz, $CDCl_3$): δ 8.38 – 8.32 (m, 2H), 7.80 – 7.71 (m, 4H), 7.12 (s, 2H), 6.02 (tt, *J* = 10.3, 4.7 Hz, 1H), 2.85 (m, 2H), 2.50 (m, 2H), 2.46 (s, 3H), 2.13 (m, 2H), 2.02 – 1.80

(m, 6H), 1.54 (s, 18H). Peaks broad and split due to the presence of N–C rotamers, compound was unstable to high temperature NMR.

^{13}C NMR (MHz, $CDCl_3$) δ 164.23, 161.56, 161.37, 142.37, 140.26, 136.10, 129.51, 129.36, 129.15, 129.06 (corresponds to C–H mesityl), 127.39, 126.95, 124.97, 124.96, 115.88, 113.08, 66.90, 37.14, 33.91, 30.75, 30.60, 28.49, 27.68, 21.40, 20.25. Peaks broad and split due to the presence of N–C rotamers, compound was unstable to high temperature NMR.

^{19}F NMR (282 MHz, CD_3CN): δ -151.83, -151.89.

HRMS (APCI): calculated for $C_{37}H_{48}N$ ($[M]^+$): 506.3781, found 506.3865.



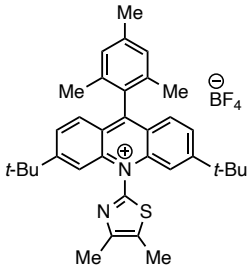
3,6-di-*tert*-butyl-9-mesityl-10-(3,4,5-trimethoxyphenyl)acridin-10-ium tetrafluoroborate (A6).

^{1}H NMR (600 MHz, CD_3CN): δ 7.93 (dd, J = 9.1, 1.7 Hz, 2H), 7.76 (d, J = 9.1 Hz, 2H), 7.56 (d, J = 1.7 Hz, 2H), 7.27 (s, 2H), 7.04 (s, 2H), 3.96 (s, 3H), 3.85 (s, 6H), 2.49 (s, 3H), 1.81 (s, 6H), 1.33 (s, 18H).

^{13}C NMR (151 MHz, CD_3CN): δ 164.1, 162.6, 156.1, 143.4, 141.2, 141.0, 137.0, 133.3, 130.5, 129.7, 128.9, 128.6, 125.0, 116.1, 106.6, 61.4, 57.3, 37.3, 30.3, 21.3, 20.1.

^{19}F NMR (282 MHz, CD_3CN): δ -151.85, -151.90.

HRMS (APCI): calculated for $C_{39}H_{46}NO_3$ ($[M]^+$): 576.3472, found 576.3562.



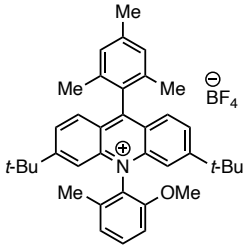
3,6-di-*tert*-butyl-10-(4,5-dimethylthiazol-2-yl)-9-mesitylacridin-10-ium tetrafluoroborate (A7).

^{1}H NMR (600 MHz, CD_3CN): δ 7.96 (dd, *J* = 9.0, 1.7 Hz, 2H), 7.79 (d, *J* = 9.1 Hz, 2H), 7.51 (d, *J* = 1.6 Hz, 2H), 7.25 (s, 2H), 2.65 (s, 3H), 2.55 (s, 3H), 2.48 (s, 3H), 1.79 (s, 6H), 1.36 (s, 18H).

^{13}C NMR (151 MHz, CD_3CN): δ 165.8, 150.7, 149.5, 143.7, 141.4, 137.0, 135.1, 130.2, 129.7, 129.6, 128.9, 125.1, 114.8, 37.5, 30.4, 21.3, 20.1, 14.9, 12.1.

^{19}F NMR (282 MHz, CD_3CN): δ -151.86, -151.90.

HRMS (APCI): calculated for $C_{35}H_{41}N_2S$ ($[M]^+$): 521.2985, found 521.3065.



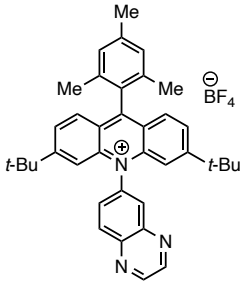
3,6-di-*tert*-butyl-9-mesityl-10-(2-methoxy-6-methylphenyl)acridin-10-ium tetrafluoroborate (A8).

^{1}H NMR (500 MHz, CD_3CN): δ 7.96 (dd, *J* = 9.1, 1.7 Hz, 2H), 7.80 (d, *J* = 8.8 Hz, 3H), 7.42 (d, *J* = 1.7 Hz, 2H), 7.38 – 7.33 (m, 2H), 7.27 (s, 2H), 3.63 (s, 3H), 2.49 (s, 3H), 1.82 (s, 3H), 1.79 (s, 6H), 1.29 (s, 18H).

^{13}C NMR (126 MHz, CD_3CN): δ 165.49, 163.08, 155.23, 142.38, 141.26, 137.61, 136.77, 136.68, 133.90, 130.32, 129.74, 129.70, 129.43, 128.98, 125.16, 124.84, 124.71, 114.28, 112.21, 57.15, 37.24, 30.27, 21.30, 20.00, 16.81. Peaks split due to the presence of N–C rotamers, compound was unstable to high temperature NMR.

^{19}F NMR (282 MHz, CD_3CN): δ -151.84, -151.89.

HRMS (APCI): calculated for $C_{38}H_{44}NO$ ($[M]^+$): 530.3417, found 530.3498.



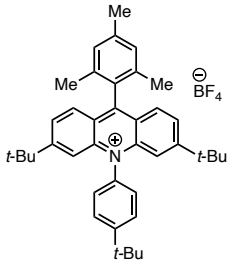
3,6-di-*tert*-butyl-9-mesityl-10-(quinoxalin-6-yl)acridin-10-ium tetrafluoroborate (A9).

^{1}H NMR (600 MHz, CD_3CN): δ 9.18 (d, *J* = 1.8 Hz, 1H), 9.11 (d, *J* = 1.8 Hz, 1H), 8.64 (d, *J* = 8.7 Hz, 1H), 8.53 (d, *J* = 2.3 Hz, 1H), 8.09 (dd, *J* = 8.7, 2.3 Hz, 1H), 7.95 (dd, *J* = 9.1, 1.6 Hz, 2H), 7.83 (d, *J* = 9.1 Hz, 2H), 7.40 (d, *J* = 1.6 Hz, 2H), 7.29 (s, 2H), 2.50 (s, 3H), 1.86 (s, 6H), 1.22 (s, 18H).

^{13}C NMR (126 MHz, CD_3CN): δ 164.7, 163.5, 149.2, 148.6, 144.6, 144.2, 143.4, 141.3, 138.5, 137.1, 137.0, 134.4, 131.2, 131.2, 130.8, 130.4, 129.7, 129.3, 128.7, 125.2, 115.8, 37.3, 30.3, 21.3, 20.1.

^{19}F NMR (282 MHz, CD_3CN): δ -151.86, -151.91.

HRMS (APCI): calculated for $C_{38}H_{40}N_3$ ([M]$^+$): 538.3217, found 538.3304.



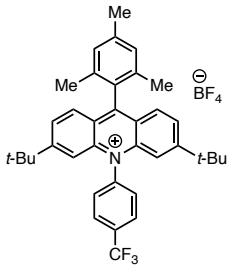
3,6-di-*tert*-butyl-10-(4-(*tert*-butyl)phenyl)-9-mesitylacridin-10-ium tetrafluoroborate (A10).

^{1}H NMR (500 MHz, CD_3CN): δ 7.98 – 7.88 (m, 4H), 7.76 (d, *J* = 8.8 Hz, 2H), 7.60 (d, 2H), 7.41 (s, 2H), 7.25 (s, 2H), 2.49 (s, 3H), 1.80 (s, 6H), 1.50 (s, 9H), 1.34 – 1.23 (m, 18H).

^{13}C NMR (126 MHz, CD_3CN) δ 163.90, 162.44, 156.29, 143.50, 141.13, 137.02, 135.51, 130.44, 129.69, 129.04, 128.97, 128.49, 128.38, 125.11, 115.90, 37.17, 35.91, 31.41, 30.25, 21.28, 20.08.

^{19}F NMR (282 MHz, CD_3CN): δ -151.81, -151.87.

HRMS (APCI): calculated for $C_{40}H_{48}N$ ([M]$^+$): 542.3781, found 542.3867.



3,6-di-*tert*-butyl-9-mesityl-10-(4-(trifluoromethyl)phenyl)acridin-10-ium tetrafluoroborate (A11).

^{1}H NMR (500 MHz, CD_3CN): δ 8.27 (d, *J* = 8.1 Hz, 2H), 7.98 – 7.91 (m, 4H), 7.80 (d, *J* = 9.1 Hz, 2H), 7.31 (d, *J* = 1.6 Hz, 2H), 7.29 – 7.25 (m, 2H), 2.49 (s, 3H), 1.82 (s, 6H), 1.28 (s, 18H).

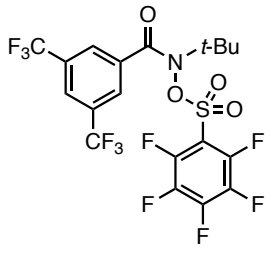
^{13}C NMR (126 MHz, CD_3CN) δ 164.6, 163.5, 143.1, 141.3 (q, *J* = 1.3 Hz), 141.2, 137.0, 133.8 (q, *J* = 32.9 Hz), 130.3, 130.3, 129.7, 129.6 (q, *J* = 3.9 Hz), 129.3, 128.6, 125.1, 124.7 (q, J = 271.7 Hz), 115.5, 37.3, 30.3, 21.3, 20.1.

^{19}F NMR (282 MHz, CD_3CN): δ -63.34, -151.82, -151.87.

HRMS (APCI): calculated for $C_{37}H_{39}NF_3$ ($[M]^+$): 554.3029, found 554.3113.

3.4.3 Representative procedure for C–H Ritter amidation with MeCN/H_2O

Prepared according to a modified literature procedure.[19] In a nitrogen filled glovebox, a 0.5 dram vial was filled with 0.002 mmol photocatalyst, 0.10 mmol HAT reagent (structure shown below), flea-sized stir bar, and MeCN (0.5 mL, 0.2 M). 0.2 mmol of ethylbenzene and H_2O (3.6 μL, 0.20 mmol) were added, and the vial was capped with a screw-cap septa and sealed with parafilm. The vials were brought out of the glovebox and placed on stir plate equipped with a 390nm Kessil lamp (set to 100% intensity) approximately 2cm to the side of the vial. After stirring for 20 hours, a solution of 1,3,5-trimethoxy benzene in $CDCl_3$ (8.4 mg, 0.05 mmol in 1 mL of $CDCl_3$) was added and the yield was determined by ^{1}H NMR relative to the 1,3,5-trimethoxybenzene standard. For 2 vials run on a single stir plate, 1 Kessil lamp was used. For 4 vials run on a single plate, 2 Kessil lamps were used on either side of the stir plate.



HAT reagent

3.4.4 Mechanistic experiments to generate $PC^{\bullet-}$ in a photocatalytic quenching cycle

Reduced photocatalysts were prepared *in situ* using a modified literature procedure.[6]

NMR characterization of $PC^{\bullet-}$ for Mes-Acr-Ph-BF_4 (catalyst **1**)

1. In a nitrogen-filled glovebox, a 10 mM stock solution of catalyst **1** (Mes-Acr-Ph-BF_4) was prepared in CD_3CN.
2. Separately, a 10 mM stock solution of $CoCp_2$ was prepared in CD_3CN.
3. In a 1-dram vial, 300 µL of the photocatalyst solution was added.
4. To this solution was slowly added 600 µL of the $CoCp_2$ stock solution while stirring gently. Upon addition, the solution turned from pale yellow to dark pink/red immediately.
5. The vial was capped and stirred for 10 min. to ensure complete mixing.
6. Following this, the total reaction volume (~900 µL) was passed through a small pad of alumina and added to a J. Young NMR tube. The alumina plug was washed with 1 mL CD_3CN to ensure complete transfer.
7. Separately, a 20 mM stock solution of HAT reagent was prepared in CD_3CN.
8. 150 µL of the HAT reagent stock solution was added to the J. Young NMR tube and the tube was sealed and inverted 3x to ensure complete mixing.
9. The J. Young tube was then brought out of the glovebox and subjected to NMR analysis (step 3, **Fig. 133**).
10. The J. Young tube was then irradiated with a 390nm Kessil lamp for 10 min and then subjected again to NMR analysis (step 4, **Fig. 133**). % conversion of the HAT reagent was quantified using ^{19}F NMR analysis.

11. Separately, 500 μL of the HAT reagent solution was also dispensed into a regular NMR tube and subjected to NMR analysis as a reference (step 1, **Fig. 133**).

NMR characterization of $PC^{\bullet-}$ for Mes-Acr-cycloheptyl-BF_4 (catalyst **A5**)

1. In a nitrogen-filled glovebox, a 10 mM stock solution of catalyst **A5** (Mes-Acr-cycloheptyl-BF_4) was prepared in CD_3CN.
2. Separately, a 10 mM stock solution of $CoCp_2$ was prepared in CD_3CN.
3. In a 1-dram vial, 300 μL of the photocatalyst solution was added.
4. To this solution was slowly added 600 μL of the $CoCp_2$ stock solution while stirring gently. Upon addition, the solution turned from pale yellow to dark pink/red immediately.
5. The vial was capped and stirred for 10 min. to ensure complete mixing.
6. Following this, the total reaction volume (~900 μL) was passed through a small pad of alumina and added to a J. Young NMR tube. The alumina plug was washed with 1 mL CD_3CN to ensure complete transfer.
7. Separately, a 20 mM stock solution of HAT reagent was prepared in CD_3CN.
8. 150 μL of the HAT reagent stock solution was added to the J.Young NMR tube and the tube was sealed and inverted 3x to ensure complete mixing.
9. The J.Young tube was then brought out of the glovebox and subjected to NMR analysis (step 3, **Fig. 133**).
10. The J. Young tube was then irradiated with a 390nm Kessil lamp for 10 min and then subjected again to NMR analysis (step 4, **Fig. 133**). % conversion of the HAT reagent was quantified using ^{19}F NMR analysis.

11. Separately, 500 µL of the HAT reagent solution was also dispensed into a regular NMR tube and subjected to NMR analysis as a reference (step 1, **Fig. 133**).

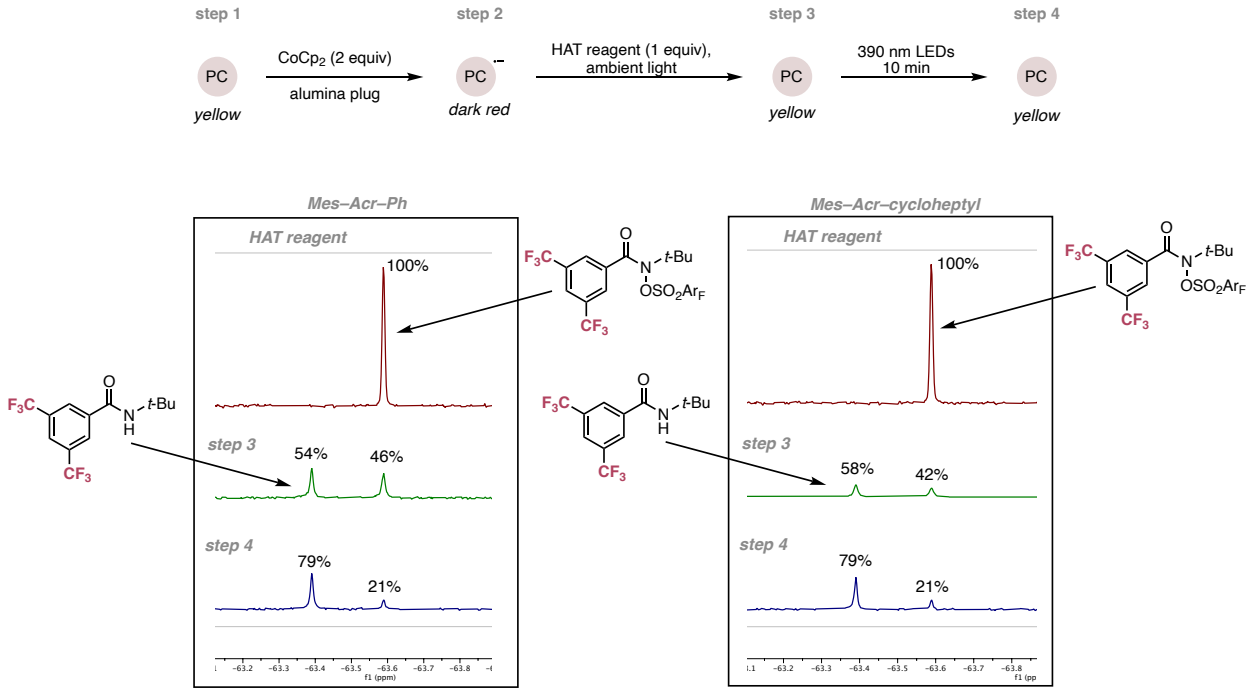


Fig. 133 Mechanistic experiments to generate and investigate the role of acridine radical.

^{19}F NMR spectra of HAT reagent before addition of PC$^{\bullet-}$ (step 1) and after addition of PC$^{\bullet-}$ (step 3). After addition of PC$^{\bullet-}$ to the HAT reagent (step 3), two signals are observed in the ^{19}F NMR. The right peak indicates residual HAT reagent in the reaction mixture. Notably, formation of a new peak (left) is observed, consistent with formation and fragmentation of the HAT$^{\bullet-}$ species, generating a new HAT-derived byproduct and supporting the mechanistic hypothesis that PC$^{\bullet-}$ performs a single-electron reduction on the HAT reagent, prompting mesolytic fragmentation. Moreover, the amount of HAT reagent remaining further decreases after irradiation of the mixture (step 4), demonstrating that the reduced photocatalyst can undergo a second photoexcitation event which increases conversion of the HAT reagent to the fragmentation byproduct. Taken together,

these experiments suggest that for these acridinium photocatalysts a consecutive photoinduced electron transfer (conPET) mechanism may be operative to generate product.

3.4.5 Quantum mechanical calculations

Theoretical calculations with density functional theory (DFT) and time-dependent DFT (TD-DFT) were carried out on the Gaussian 16, Revision B.01 package. Conformers were obtained by performing a conformational search using the Conformer-Rotamer Ensemble Sampling Tool (CREST) at the tight-binding level with GFN-xTB, incorporating explicit solvation and an implicit solvent model (GBSA with acetonitrile).[21] This was followed by single-point energy calculations using N12-SX/6-311+G(d,p),[22,23] re-ranking and full optimization of the 10 lowest-energy conformers with N12-SX/6-311+G(d,p) and implicit solvation with the polarizable continuum model (PCM) using the integral equation formalism variant IEFPCM^{MeCN}.[24] Single point energy calculation was performed to determine the lowest energy conformer. Then optimization calculations with the lowest energy conformer were performed for S_1 and D_0 with IEFPCM^{MeCN}. Frequency calculations were performed to verify the absence of an imaginary frequency for the optimized geometry and confirm that they are a minimum on their potential energy surfaces.

<u>Full Gaussian16 reference:</u>

Gaussian 16, Revision A.03, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R.

Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.

XYZ coordinates of optimized structures and output files

All optimized structures and output files are available at https://doi.org/10.6084/m9.figshare.28346972.v1.

S_1 Vertical emission energy, nature of transition, S_1 character, orbitals and S_0 and S_1 Dipole

Molecule	S_1 Emission Energy (eV)	Nature of Electronic transition	S_1 Character
A1	2.23	$\pi_{Mes} \rightarrow \pi_{core}$	CT*
A2	1.87	$\pi_{R} \rightarrow \pi_{core}$	CT
A3	2.30	$\pi_{Mes} \rightarrow \pi_{core}$	CT
A4	2.34	$\pi_{Mes} \rightarrow \pi_{core}$	CT
A5	2.33	$\pi_{Mes} \rightarrow \pi_{core}$	CT
A6	1.54	$\pi_{R} \rightarrow \pi_{core}$	CT
A7	2.21	$\pi_{Mes} \rightarrow \pi_{core}$	CT
A8	2.27	$\pi_{Mes} \rightarrow \pi_{core}$	CT
A9	2.29	$\pi_{Mes} \rightarrow \pi_{core}$	CT
A10	2.27	$\pi_{Mes} \rightarrow \pi_{core}$	CT
A11	2.22	$\pi_{Mes} \rightarrow \pi_{core}$	CT

Table 5 Vertical emission energies, nature of electronic transition with largest coefficient, and character at N12-SX/6-311+G(d,p) for the optimized geometry of S1 of synthesized acridiniums.

*CT: charge transfer.

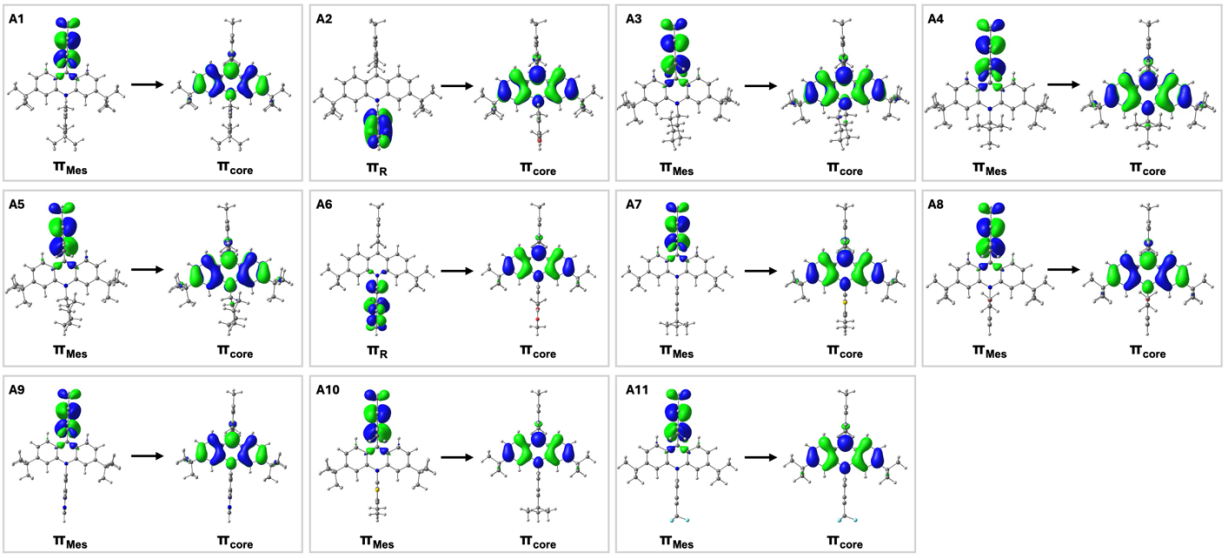


Fig. 134 Orbitals involved emission from S1 → S0 for the S1 optimized geometry with the largest coefficients in the CI expansion. Orbitals are shown with an isosurface value of 0.03 and with mesityl group on the top.

Molecule	S_0 Dipole (D)	S_1 Dipole (D)
A1	3.37	19.5
A2	1.55	22.9
A3	2.66	15.1
A4	3.16	14.4
A5	1.98	15.7
A6	4.15	20.7
A7	2.92	15.8
A8	2.69	15.2
A9	0.67	18.4
A10	0.08	18.0
A11	5.02	22.9

Table 6 Dipole of optimized S0 and S1 geometry at N12-SX/6-311+G(d,p) for the synthesized acridiniums.

TD-DFT calculations were performed on the optimized S_1 geometry of the synthesized acridinium to investigate the nature of electronic transition for $S_1 \rightarrow S_0$ (**Table 5**). In this transition, an electron transition from an orbital predominantly localized in the mesityl groups of most molecules. Exceptions are noted in **A2** and **A6**, where the orbital is localized in the *N*-substituent. The electron transition to an orbital localized in the acridinium core, which exhibits minimal spatial

overlap with the initial orbital. This indicates that the S_1 excited state is an intramolecular charge-transfer (CT) state (**Fig. 134**). Dipole moments were computed for both the S_0 and S_1 optimized geometries. For S_0, the dipole moment ranged from 0.08 to 5.02 D for the, while for the S_1 state, it ranged significantly higher, from 14.4 to 22.9 D (**Table 6**). This substantial increase in charge separation in the S_1 state compared to S_0 further supports the characterization of S_1 as a CT state.

Acridine vertical absorption energy, nature of transition, D_1 character and orbitals

Molecule	$D_0 \rightarrow D_1$ Absorption Energy (eV)	Oscillator Strength	Nature of Electronic transition	D_1 Character
A1	2.34	0.0145	$\pi_{core} \rightarrow \pi^*_{core} / \pi^*_{N\text{-substituent}}$	HLCT*
A2	2.34	0.0147	$\pi_{core} \rightarrow \pi^*_{core} / \pi^*_{N\text{-substituent}}$	HLCT
A3	2.23	0.0059	$\pi_{core} \rightarrow \pi^*_{core}$	LE**
A4	2.14	0.0069	$\pi_{core} \rightarrow \pi^*_{core}$	LE
A5	2.19	0.0057	$\pi_{core} \rightarrow \pi^*_{core}$	LE
A6	2.30	0.0163	$\pi_{core} \rightarrow \pi^*_{core} / \pi^*_{N\text{-substituent}}$	HLCT
A7	2.18	0.0048	$\pi_{core} \rightarrow \pi^*_{N\text{-substituent}}$	CT***
A8	2.32	0.0139	$\pi_{core} \rightarrow \pi^*_{core} / \pi^*_{N\text{-substituent}}$	HLCT
A9	0.93	0.0004	$\pi_{core} \rightarrow \pi^*_{N\text{-substituent}}$	CT
A10	2.26	0.0220	$\pi_{core} \rightarrow \pi^*_{core} / \pi^*_{N\text{-substituent}}$	HLCT
A11	1.98	0.0000	$\pi_{core} \rightarrow \pi^*_{N\text{-substituent}}$	CT

Table 7 Vertical absorption energies, nature of electronic transition with largest coefficient, and character at N12-SX/6-311+G(d,p) for the optimized geometry of acridine on ground state (D0) of synthesized acridiniums.

*HLCT: hybridized local and charge-transfer.
**LE: local excited
***CT: charge transfer

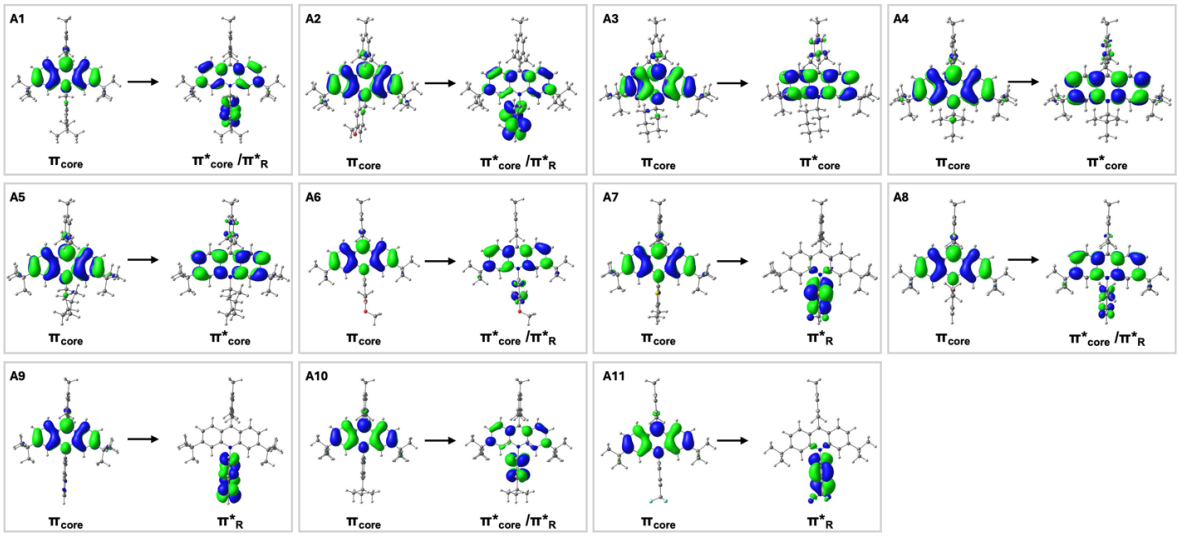


Fig. 135 Orbitals involved absorption from D0 → D1 for the D0 acridine optimized geometry with the largest coefficients in the CI expansion. Orbitals are shown with an isosurface value of 0.03 and with Mesityl group on the top.

3.5 References

1. Singh, P. P.; Singh, J.; Srivastava, V. Visible-Light Acridinium-Based Organophotoredox Catalysis in Late-Stage Synthetic Applications. *RSC Adv.* **2023**, *13* (16), 10958–10986. https://doi.org/10.1039/d3ra01364b.
2. White, A.; Wang, L.; Nicewicz, D. Synthesis and Characterization of Acridinium Dyes for Photoredox Catalysis. *Synlett* **2019**, *30* (07), 827–832. https://doi.org/10.1055/s-0037-1611744.
3. Fischer, C.; Sparr, C. Direct Transformation of Esters into Heterocyclic Fluorophores. *Angew. Chem. Int. Ed.* **2018**, *57* (9), 2436–2440. https://doi.org/10.1002/anie.201711296.
4. Fischer, C.; Sparr, C. Synthesis of 1,5-Bifunctional Organolithium Reagents by a Double Directed Ortho-Metalation: Direct Transformation of Esters into 1,8-Dimethoxy-Acridinium Salts. *Tetrahedron* **2018**, *74* (38), 5486–5493. https://doi.org/10.1016/j.tet.2018.04.060.

5. Hutskalova, V.; Sparr, C. Ad Hoc Adjustment of Photoredox Properties by the Late-Stage Diversification of Acridinium Photocatalysts. *Org. Lett.* **2021**, *23* (13), 5143–5147. https://doi.org/10.1021/acs.orglett.1c01673.
6. Yan, H.; Song, J.; Zhu, S.; Xu, H.-C. Synthesis of Acridinium Photocatalysts via Site-Selective C–H Alkylation. *CCS Chem.* **2021**, *3* (12), 317–325. https://doi.org/10.31635/ccschem.021.202000743.
7. Gini, A.; Uygur, M.; Rigotti, T.; Alemán, J.; Mancheño, O. G. Novel Oxidative Ugi Reaction for the Synthesis of Highly Active, Visible-Light, Imide-Acridinium Organophotocatalysts. *Chem. A Eur. J.* **2018**, *24* (48), 12509–12514. https://doi.org/10.1002/chem.201802830.
8. Cao, Y.-X.; Zhu, G.; Li, Y.; Breton, N. L.; Gourlaouen, C.; Choua, S.; Boixel, J.; Rouville, H.-P. J. de; Soulé, J.-F. Photoinduced Arylation of Acridinium Salts: Tunable Photoredox Catalysts for C–O Bond Cleavage. *J. Am. Chem. Soc.* **2022**, *144* (13), 5902–5909. https://doi.org/10.1021/jacs.1c12961.
9. Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminformatics* **2015**, *7* (1), 20. https://doi.org/10.1186/s13321-015-0069-3.
10. Landrum, G. *Additional Information About the Fingerprints*. https://www.rdkit.org/docs/RDKit_Book.html#additional-information-about-the-fingerprints.
11. Fukuzumi, S.; Kotani, H.; Ohkubo, K.; Ogo, S.; Tkachenko, N. V.; Lemmetyinen, H. Electron-Transfer State of 9-Mesityl-10-Methylacridinium Ion with a Much Longer Lifetime and Higher Energy Than That of the Natural Photosynthetic Reaction Center. *J. Am. Chem. Soc.* **2004**, *126* (6), 1600–1601. https://doi.org/10.1021/ja038656q.

12. Romero, N. A.; Nicewicz, D. A. Mechanistic Insight into the Photoredox Catalysis of Anti-Markovnikov Alkene Hydrofunctionalization Reactions. *J. Am. Chem. Soc.* **2014**, *136* (49), 17024–17035. https://doi.org/10.1021/ja506228u.

13. Tay, N. E. S.; Nicewicz, D. A. Cation Radical Accelerated Nucleophilic Aromatic Substitution via Organic Photoredox Catalysis. *J. Am. Chem. Soc.* **2017**, *139* (45), 16100–16104. https://doi.org/10.1021/jacs.7b10076.

14. MacKenzie, I. A.; Wang, L.; Onuska, N. P. R.; Williams, O. F.; Begam, K.; Moran, A. M.; Dunietz, B. D.; Nicewicz, D. A. Discovery and Characterization of Acridine Radical Photoreductants. *Nature* **2020**, *580* (7801), 76–80. https://doi.org/10.1038/s41586-020-2131-1.

15. Ruos, M. E.; Kinney, R. G.; Ring, O. T.; Doyle, A. G. A General Photocatalytic Strategy for Nucleophilic Amination of Primary and Secondary Benzylic C–H Bonds. *J. Am. Chem. Soc.* **2023**, *145* (33), 18487–18496. https://doi.org/10.1021/jacs.3c04912.

16. Fukuzumi, S.; Ohkubo, K.; Suenobu, T. Long-Lived Charge Separation and Applications in Artificial Photosynthesis. *Acc. Chem. Res.* **2014**, *47* (5), 1455–1464. https://doi.org/10.1021/ar400200u.

17. Fukuzumi, S.; Lee, Y.; Nam, W. Photoredox Catalysis of Acridinium and Quinolinium Ion Derivatives. *Bull. Korean Chem. Soc.* **2025**, *46* (1), 4–23. https://doi.org/10.1002/bkcs.12922.

18. White, A.; Wang, L.; Nicewicz, D. Synthesis and Characterization of Acridinium Dyes for Photoredox Catalysis. *Synlett* **2019**, *30* (07), 827–832. https://doi.org/10.1055/s-0037-1611744.

19. Ruos, M. E.; Kinney, R. G.; Ring, O. T.; Doyle, A. G. A General Photocatalytic Strategy for Nucleophilic Amination of Primary and Secondary Benzylic C–H Bonds. *J. Am. Chem. Soc.* **2023**, *145* (33), 18487–18496. https://doi.org/10.1021/jacs.3c04912.

20. Romero, N. A.; Nicewicz, D. A. Mechanistic Insight into the Photoredox Catalysis of Anti-Markovnikov Alkene Hydrofunctionalization Reactions. *J. Am. Chem. Soc.* **2014**, *136* (49), 17024–17035. https://doi.org/10.1021/ja506228u.

21. Pracht, P.; Bohle, F.; Grimme, S. Automated Exploration of the Low-Energy Chemical Space with Fast Quantum Chemical Methods. *Phys. Chem. Chem. Phys.* **2020**, *22* (14), 7169–7192. https://doi.org/10.1039/c9cp06869d.

22. Frisch, M. J.; Pople, J. A.; Binkley, J. S. Self-Consistent Molecular Orbital Methods 25. Supplementary Functions for Gaussian Basis Sets. *J. Chem. Phys.* **1984**, *80* (7), 3265–3269. https://doi.org/10.1063/1.447079.

23. Peverati, R.; Truhlar, D. G. Screened-Exchange Density Functionals with Broad Accuracy for Chemistry and Solid-State Physics. *Phys. Chem. Chem. Phys.* **2012**, *14* (47), 16187–16191. https://doi.org/10.1039/c2cp42576a.

24. Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105* (8), 2999–3094. https://doi.org/10.1021/cr9904009.