King-Smith Group Working Guide

In this document you will find the useful links, guides, and procedures for the King-Smith Group. As we are an interdisciplinary group, you will be expected to understand synthetic chemistry and machine learning, thus the length of this document. This packet is meant to serve as both a starting up guide and something you can refer back to as you progress in your PhD. As the group grows, this guide will likely change and update. Those changes will be communicated to the group and a live version of this document will be made continually available for reference.

Remember that learning is a continual process – there is no shame in asking for more clarification or feeling confused. We are all here to support you!

Table of Contents

Group Conduct

The King-Smith Group is a welcoming space for all of its members, irrespective of their length of tenure in the group. We all come from different parts of the world, cultures, and backgrounds. While this can occasionally lead to miscommunications, our diversity is something to be celebrated. The King-Smith Group is one that is respectful and professional; no discrimination or harassment will be tolerated. If you feel that someone is displaying bullying behavior towards yourself or another member, and you feel you are unable to resolve it, you must inform me as soon as possible. This statement holds even if that person is not a member of our group.

What can You Expect from Me:

I aim to be present in my office (Rm 241) from 8:30 - 18:00, except on Tuesdays when I typically leave in the late afternoon. You are free to knock on my door. If I am not in the office for whatever reason, please do not hesitate to reach out *via* email (emma.king-smith@ed.ac.uk) or by WhatsApp ([redacted]). Unfortunately, Rm 241 is a communal office, so if you wish to have a private and/or confidential conversation with me, please let me know so that we can book a suitable location.

I will respond to communication within 1 business day or if urgent, within 24 hours. If you haven't heard back from me after that time, please re-send your email or text. Remember that you are not wasting my time – it is my job to support you and your research.

What I Expect from You:

I expect the members of my group to be diligent in their research, problem solving where they can and reaching out for assistance when they need it. Every member and I will have biweekly meetings that will be ~45 min in length, on average. In our biweekly meetings, I do not expect a cohesive narrative, but I do want to see that you are actively engaging with your research and have a plan for the upcoming weeks. I expect you to show me bad results and for the research to be a team effort.

Evening and weekend work is not an expectation, so long as there are no immovable deadlines and you are managing your time well. That being said, research often benefits from a "little extra". For example, checking that your code is still running over the weekend and no problems have arisen or staying with a critical reaction that is taking longer than usual.

Group Meetings:

Group meetings will occur once every 2 weeks and the time is TBD, but likely sometime in the late afternoon. At the moment, I expect them to be 60-90 min in length, with a maximum of two people presenting per session. At the beginning of the meeting, there will be an opportunity to discuss any issues that have arisen over the past week. These may include problems with accessing the high performance computing facility, near misses in the lab, or equipment failures. We should all be on time for group meetings and your presentations should be rehearsed at least once prior to your presentation.

Joint Lawrence-King-Smith Group Meetings:

We have joint group meetings with the Lawrence Group every Thursday from 8:30 - 10:15 in Rm 235 unless otherwise specified. **NOTE: New members will not have out-of-hours access yet so they will need to be let into the building.** These group meetings are a great way to keep on top of the chemistry literature and to practice your fundamental chemistry reasoning. Attendance is mandatory.

Lab Safety:

Like all UK labs, we implement Control of Substances Hazardous to Health (CoSHH) forms prior to any new lab work. All CoSHH forms must be approved by me before you begin your experiment and the form must be displayed on your fumehood. If you are at all ever unsure about a procedure, particularly if it involves pyrophoric agents, **ask someone**. All chemists can tell you about a horrible lab accident that they have personally witnessed – do not become another anecdote! Also remember that increasing the scale of your reaction also comes with potentially unforeseen risks. You must submit another CoSHH form if the scale of your reaction exceeds the previously approved size. Our CoSHH portal can be found here:

https://uoe.sharepoint.com/sites/chemcollab-coshh/SitePages/King-Smith%20Group.aspx?csf=1&web=1&e=Ff2yxc&CID=2abdbdfa-fb0d-48a0-aed8-c7b64668b643

Note that you must be on UoE internet to access sharepoint sites.

Standard working hours for lab work are 8:30 - 17:30. If you have not been approved for lone work or out-of-hours work you must stop all wet lab work at 17:30. Additionally, members who are not PDRAs or PhDs will not be given permission for out-of-hours wet lab work, although you are free to do desk work after 17:30. 5P students will be assessed for out-of-ours lab access on a case-by-case basis.

Sick Leave:

If you are ill, please stay home. We have members of staff who are immunocompromised and will be greatly affected if infected. In the same vein, do not work when you are sick – you will only delay your recovery. Edinburgh does have an extended sick leave policy, which if, after consultation with me, we feel is the best course of action, we will implement.

This goes as well for mental health, please take the time you need to recover and come back to research fresh and motivated.

Annual Leave, Holiday Break, & Lab Cleanup:

Every member of staff is given 32 of paid time off in addition to the holiday break (8 days). Note that these 32 days include bank holidays. Prior to you going on leave, please get my approval for the dates of your absence. I very rarely refuse the time off. The university is closed over Christmas and New Years, so no work can be conducted during that time.

Once a year, we will have a lab cleanup and inventory prior to the holiday lab shut down. I expect all members to contribute to this effort, even if your work is mostly computational. Details will be made known at a date closer to the end of the year. We will also have a festive celebration afterwards, but attendance for this is optional.

General Resources

Internet:

The most reliable internet is *via* eduroam. You can access it by following these steps:

https://information-services.ed.ac.uk/computing/desktop-personal/wifi-networking/configure-device

Edinburgh Shuttle & Lothian Buses:

A convenient way to move from the central campus to Kings Buildings during term time is the Edinburgh Shuttle. It is free to everyone who has a UoE badge and the schedule can be found here:

https://transport.ed.ac.uk/public-transport/shuttle-bus

Alternatively, the #9, #24, and #38 Lothian buses stop close to Kings Buildings. Live information can be found on the Lothian Bus & Tam App:

https://www.lothianbuses.com/app/

Chemistry Stores:

This is the storefront for purchasing general chemicals, consumables, PPE, and stationary. Note that stationery and cleaning items are free of charge.

https://uoe.sharepoint.com/sites/chem-stores

Purchasing:

Items not available through Chemistry Stores can be purchased through People & Money (P&M) under Procurement > Purchase Requisitions.

https://www.ed.ac.uk/staff/services-support/hr-and-finance/people-and-money-system

Anything that cannot be found through the default search bar must go through additional approval.

WARNING: The default codes in P&M purchasing are NOT our codes! To avoid spending on unapproved funds please use the codes provided.

Group Website:

Our website is https://www.kingsmithgroup.com/. If you are comfortable, please send me a headshot and a short ~100-word blurb about yourself. And of course keep in touch after you depart!

Dilbert & PhD Comics:

What's life without some simple pleasures? Great comics to indulge. 'Nuff said.

https://dilbert-viewer.herokuapp.com

https://phdcomics.com/

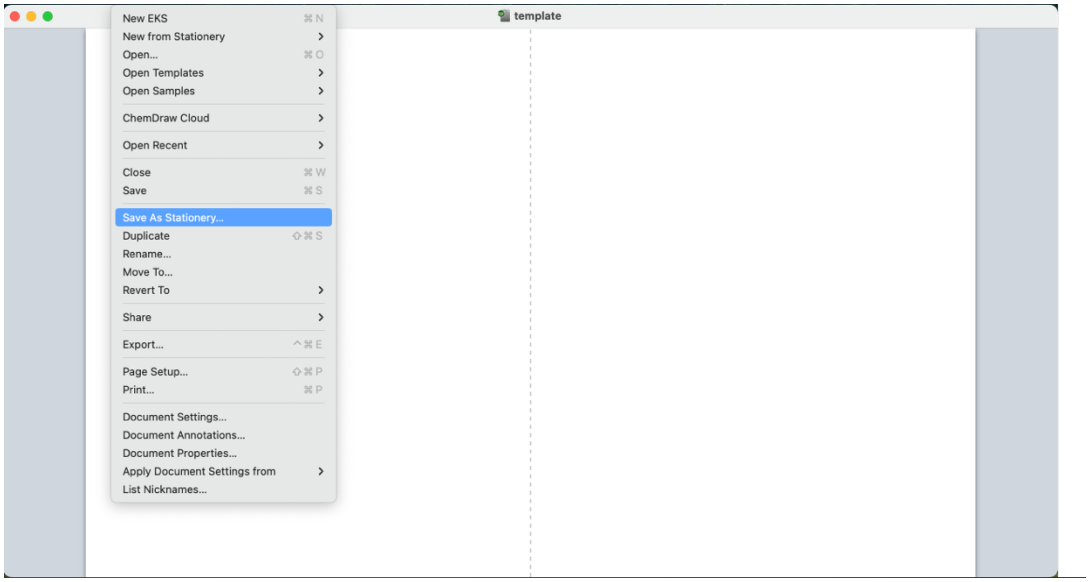
Chemistry Resources

ChemDraw:

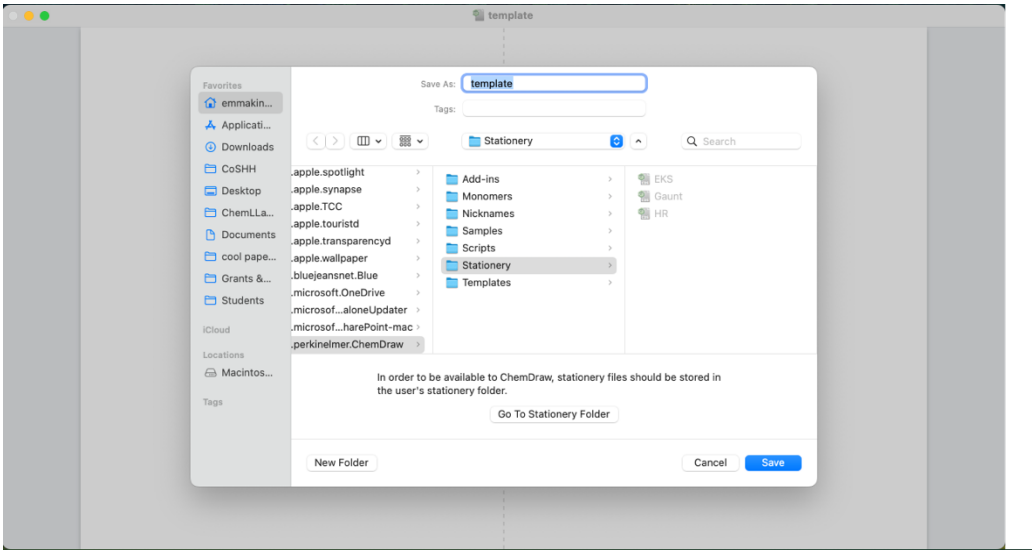
ChemDraw can be downloaded from the University of Edinburgh link:

https://uoe.sharepoint.com/sites/chem-it/SitePages/School-Provided-Software.aspx#chemdraw

The ChemDraw template that I use (and happen to find very aesthetically pleasing) was sent along with your introductory email, but I will make it available on our website as well. To set it as your stationary (e.g. default) open the template in ChemDraw, go to File > Save as Stationary.



You may name it whatever you like. Opening new ChemDraws will default to this style.



MNova:

NMR and Mass Spec analysis tool. This can be downloaded from the University of Edinburgh link:

https://uoe.sharepoint.com/sites/chem-it/SitePages/School-Provided-Software.aspx#mnova

NMR Training:

NMR training must be done prior to running your first experiment. Details can be found here:

http://nmr-server.chem.ed.ac.uk/lowfield/automation-training.html

Master Organic Chemistry:

A website that concisely explains fundamental organic chemistry concepts. For those who are dipping their toes into chemistry for the first time, this is a great first stop.

https://www.masterorganicchemistry.com/

Chemistry By Design:

A great resource to practice your synthesis – 4,000 syntheses and counting!

https://chemistrybydesign.oia.arizona.edu/app/

Fukuyama/Yokoshima Mechanisms:

Difficult to very difficult mechanism practice covering a variety of chemistry. Rest assured they come with answer sheets.

https://www.ps.nagoya-u.ac.jp/lab_pages/natural_products/problem-e.html

Computational Resources

Git & GitHub Access:

We will be using GitHub to store our code which requires git. A guide to git installation and basic git commands can be found here:

https://www.w3schools.com/git/git_workflow.asp?remote=github

Prior to publication of papers, all code should be on a private repository. This means that all members will need an SSH key to access other members' code. All code should be committed at the end of the day to prevent catastrophic losses. Approximately once a week (depending on how frequently coding is being debugged / created) you should push to your remote (GitHub repository).

A free GitHub account can be made by following this guide:

https://docs.github.com/en/get-started/start-your-journey/creating-an-account-on-github

SSH Key access can be created by following these steps:

https://docs.github.com/en/authentication/connecting-to-github-with-ssh/generating-a-new-ssh-key-and-adding-it-to-the-ssh-agent

https://docs.github.com/en/authentication/connecting-to-github-with-ssh/adding-a-new-ssh-key-to-your-github-account

Make sure you remember your password!

EDDIE Access:

EDDIE is Edinburgh's free-to-access computing service for all staff. It is a decent place to work through your code and test small bits of it, but it isn't suitable for the deep learning we will do later on. Documentation for setting up your EDDIE login can be found below and it was also be linked on our group site:

https://uoe.sharepoint.com/sites/DRSFacilitation/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2FDRSFacilitation%2FShared%20Documents%2FGettingStartedOnEddie%2Epdf&parent=%2Fsites%2FDRSFacilitation%2FShared%20Documents&p=true&ga=1

This is an EDDIE wiki that will come in handy later:

https://www.wiki.ed.ac.uk/pages/viewpage.action?spaceKey=ResearchServices&title=Eddie

EIDF GPU Access:

The deep learning will be done through the Edinburgh International Data Facility's (EIDF) GPU service. You will need an SSH key for remote desktop access but you can re-use the key you made for your GitHub account. You will need to create a EPCC SAFE account and request a Machine Account. Instructions can be found here:

https://epcced.github.io/safe-docs/safe-for-users/

The service uses kubernetes to manage the GPUs and computational jobs. They offer a nice guide on getting started with kubernetes:

https://docs.eidf.ac.uk/services/gpuservice/training/L1_getting_started/

Note that kubernetes uses Docker images. Docker can be downloaded by following the instructions here:

https://docs.docker.com/desktop/

Miniconda:

Running code can be a messy business – new packages have instabilities with others, updates to certain functions may break whole systems, versions of Python have major impacts on the types of packages you can install. All of this means we want to run things in virtual environments: clean spaces that we can setup just the way we need and tear it down if we need to. Miniconda is what we will be using and can be downloaded and installed here:

https://www.anaconda.com/docs/getting-started/miniconda/install

Pycharm & Other Interactive Development Environments (IDEs):

IDEs are a great way to get going quickly on the coding portion of your project. They offer nice features like debugging, running your code, and autofilling function names, parentheses, indents, etc. Personally, I use PyCharm, but you may use something else (just know that my assistance will be limited to my knowledge of your IDE).

https://www.jetbrains.com/pycharm/

For any IDE, you will need to specify the python interpreter / virtual environment. I find it's easiest to manage these environments by first creating a conda environment in the command line, then loading that environment into the IDE.

IDEs are, however, optional. You can do the same thing with just a text editor and a Linux terminal.

Juypter notebooks are a quick and easy way to do things, but they will not have the memory needed to run most deep learning models, so I don't advise relying on them.

Machine Learning Resources

Python:

The language we will be using for the majority of our applications is Python. In the beginning, coding is a harsh mistress and will give you many many many unintelligible error messages. This is normal! Error messages can be Googled (look for StackExchange answers) or put into the LLM of your choice (GPT, Claude, DeepSeek, etc.).

For an introduction to coding, the university offers a crash course (Data-Driven Chemistry) which can be accessed here:

https://git.ecdf.ed.ac.uk/data-driven-chemistry/data-driven-chemistry-student/-/blob/master/Session1_introduction/Session_1.1.ipynb?ref_type=heads

GeeksforGeeks is a good resource as well that offers both introductory and step-by-step guides for specific python functions.

https://www.geeksforgeeks.org/python/python-programming-language-tutorial/

Remember that the best way to learn a new language (including a computer language) is exposure! Courses can help improve your fundamentals but ultimately your learning will begin when you start trying to code on your project.

A Math-Free Introduction:

I put together a course for a math-free introduction to machine learning. Admittedly, I am a little biased, but I think it serves as a gentle guide to the fundamental concepts of classic machine learning.

https://github.com/emmaking-smith/ML_for_Chem/tree/master

Machine Learning with Math:

Given how dense math symbolism can be, I find that it's easier to learn the mathematical aspects of machine / deep learning through lectures rather than text.

Two creators I think do an excellent job of this are StatQuest and 3Blue1Brown.

https://www.youtube.com/watch?v=Gv9_4yMHFhI&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF

https://www.youtube.com/watch?v=aircAruvnKk&list=PLcCe-ymWq77ow42k4-ZrLzlM3F7Ha7smT

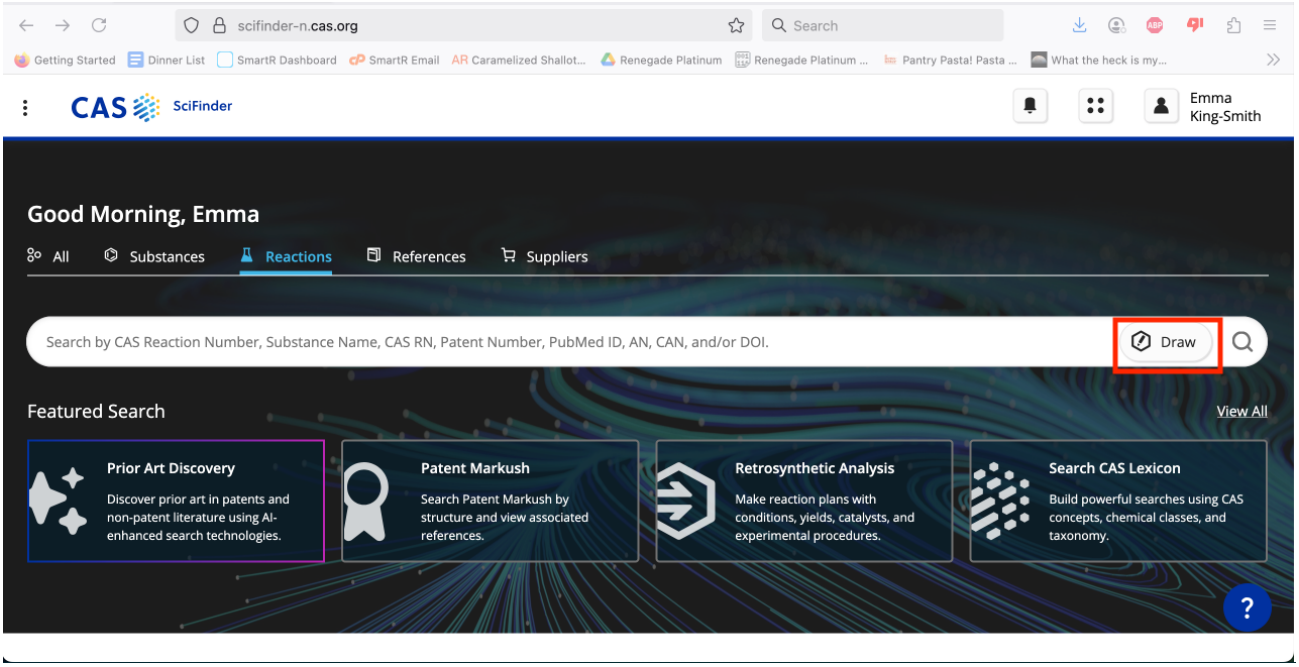
https://www.youtube.com/watch?v=fNk_zzaMoSs

I highly recommend watching through the latter's deep learning and linear algebra playlists (linked above). They are usually 10 - 20 min long and very well put together.

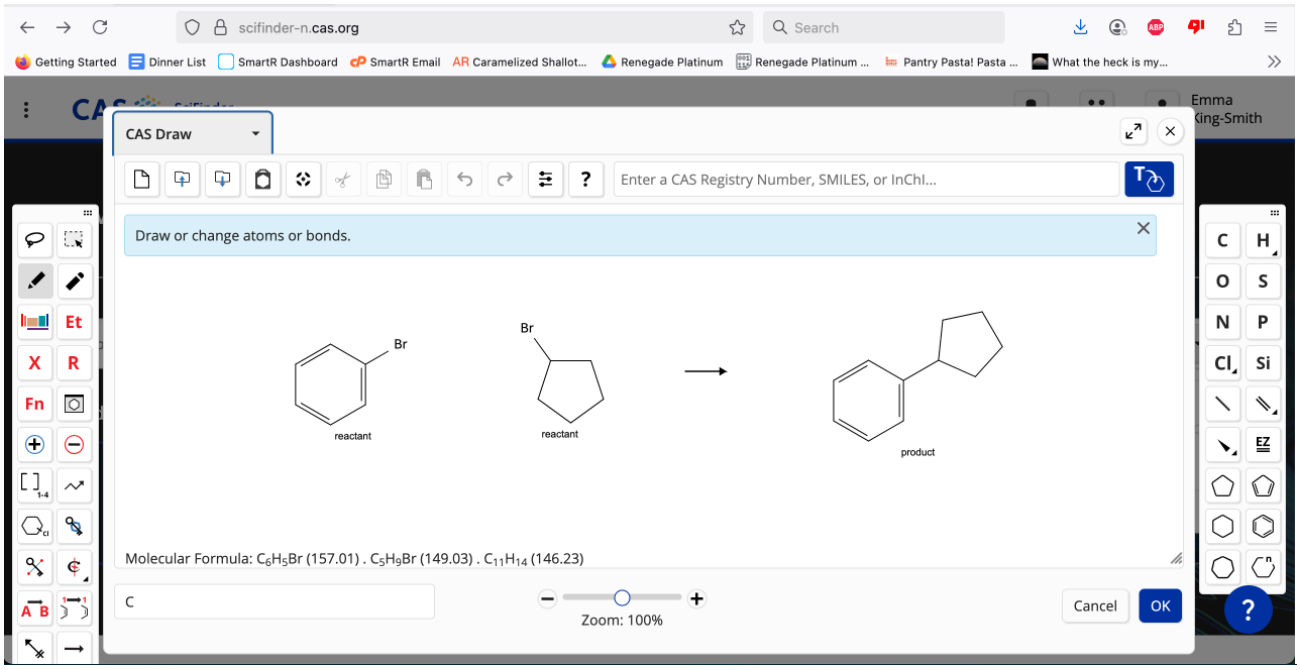
Searching Reaxys & SciFinder

One of the best resources for a chemist is Reaxys and SciFinder. Remember that these are two different databases that are run by two different companies, so they will not have identical information. It's always worthwhile checking both.

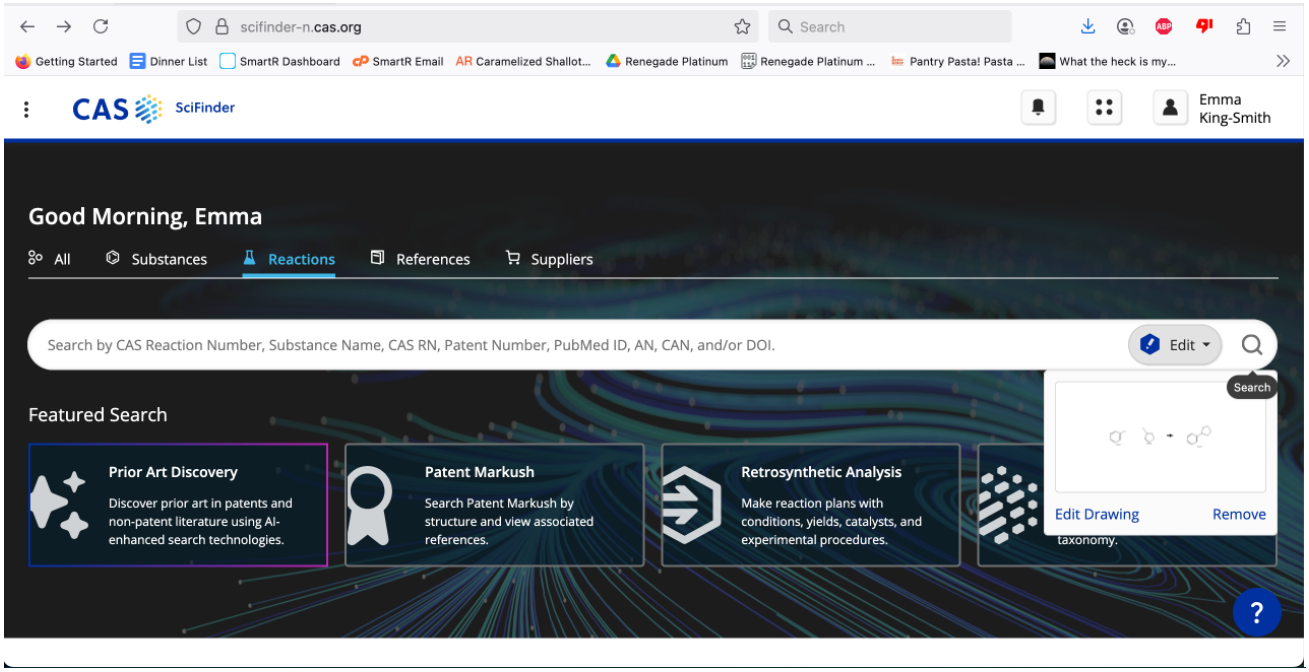
Starting with SciFinder, click the "Draw" button.



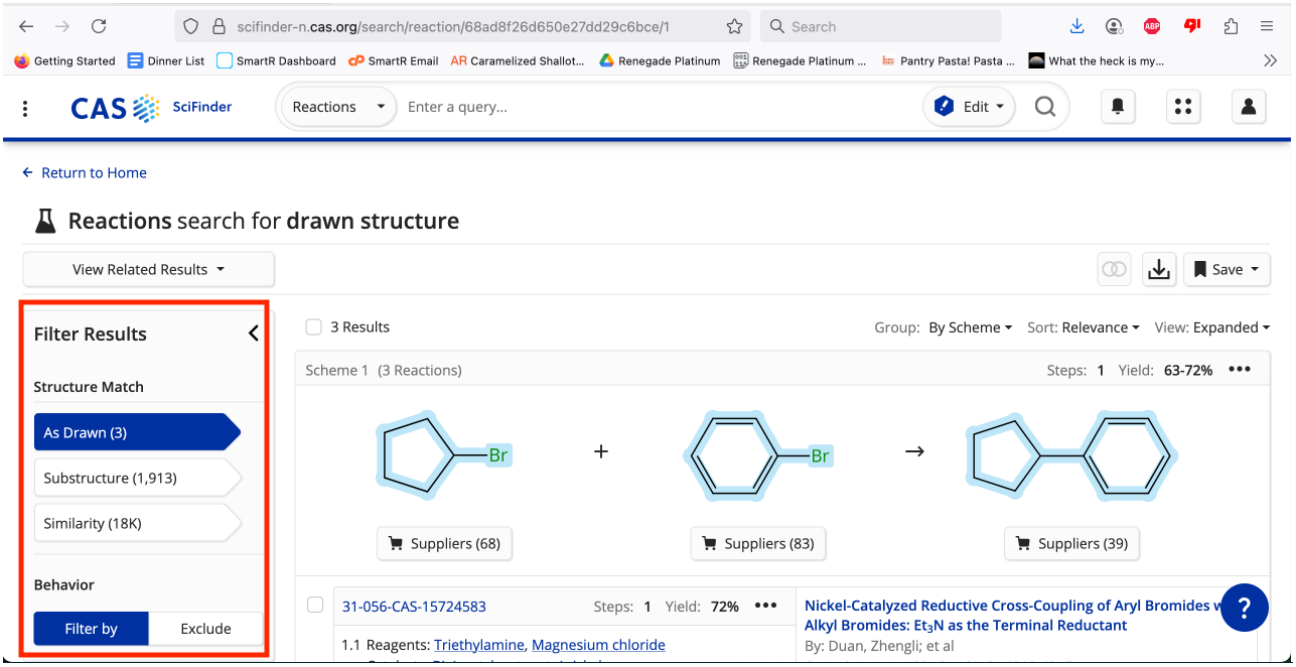
A popup window will appear where you can enter your reaction of interest.



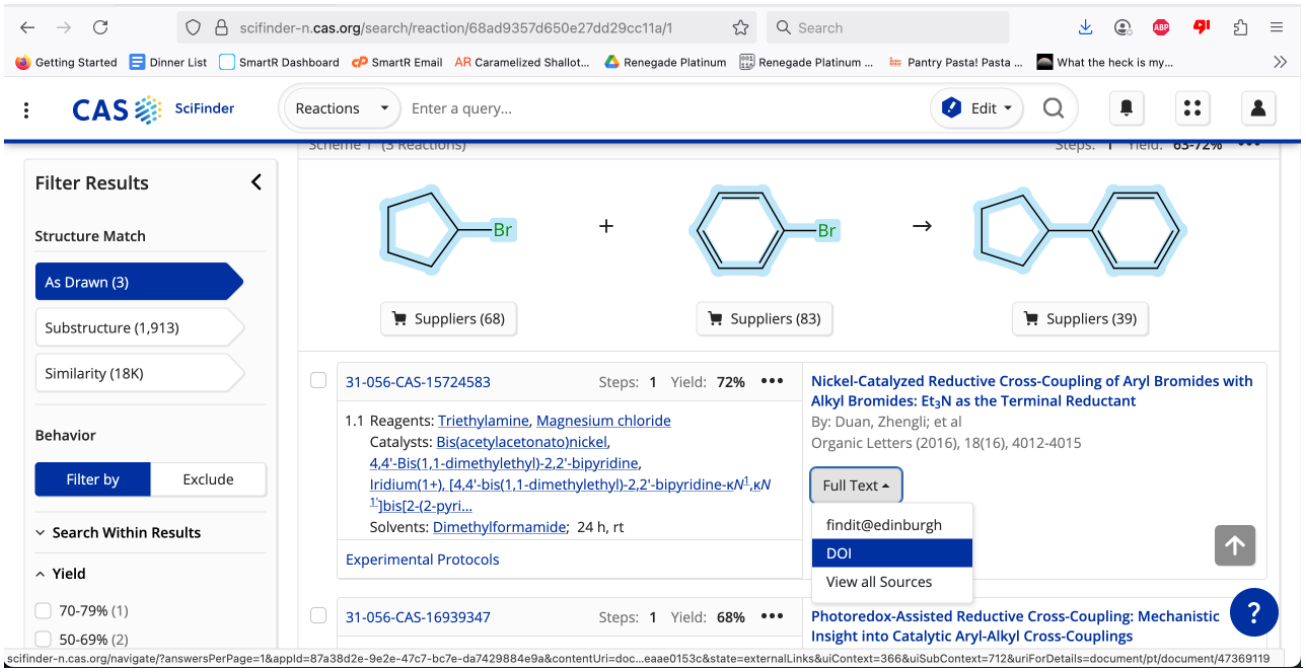
Once it is drawn, click "OK". You will be returned to the home page that now has your reaction of interested underneath the edit button. Press "Search".



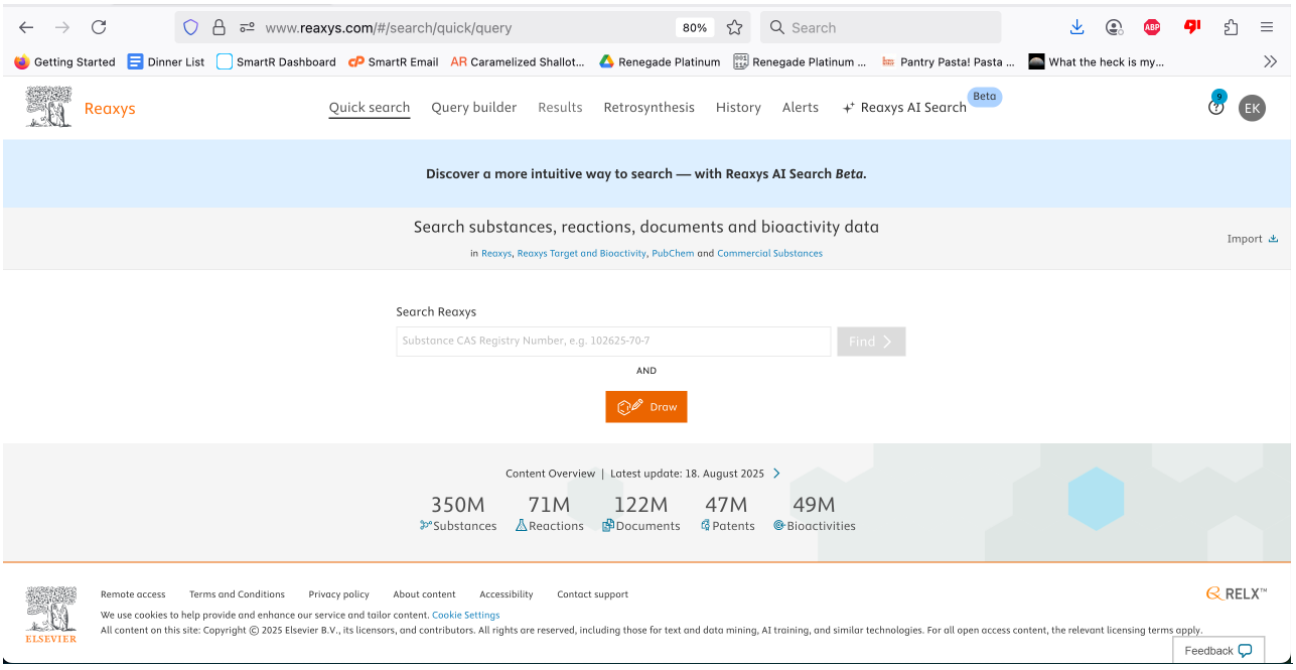
The results of the search will be displayed. You can filter for exact matches, substructure matches, or similar reactions. Scrolling through the filters, you can show results that are from specific journals, use specific reagents, or have specific yields.



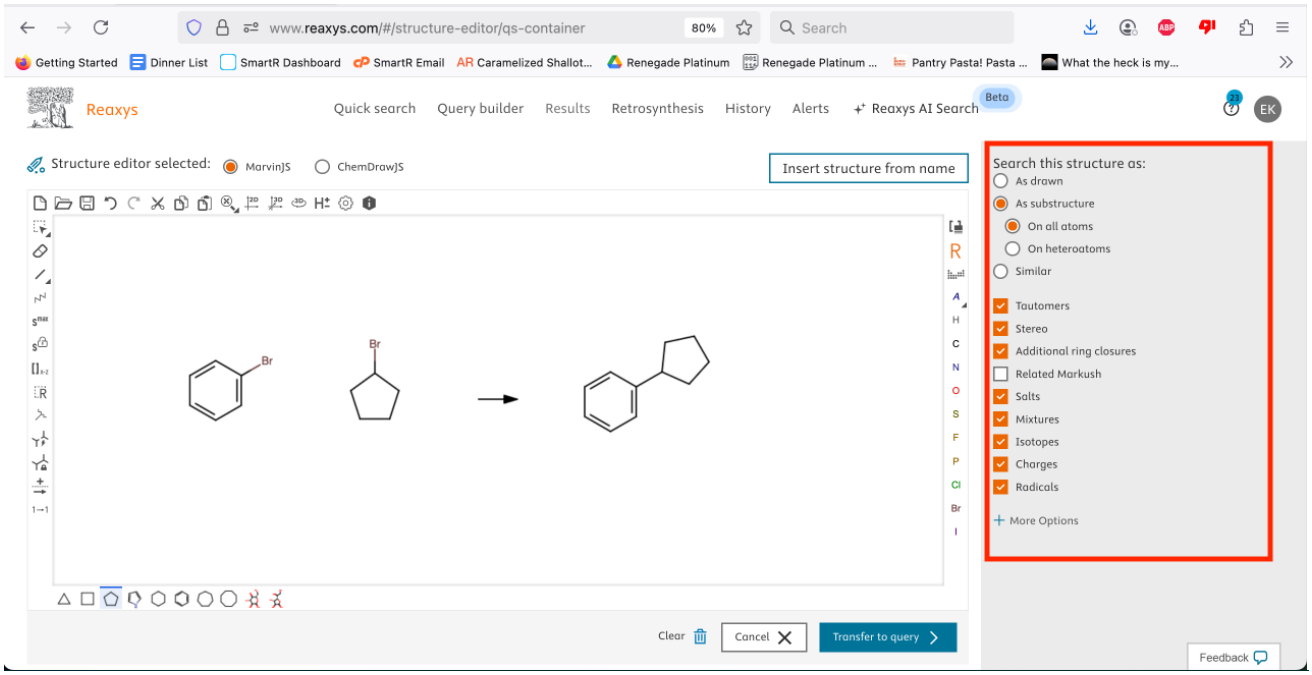
Additionally, you can quickly go to the primary source by clicking "Full Text" and selecting either "findit@edinburgh" or "DOI".



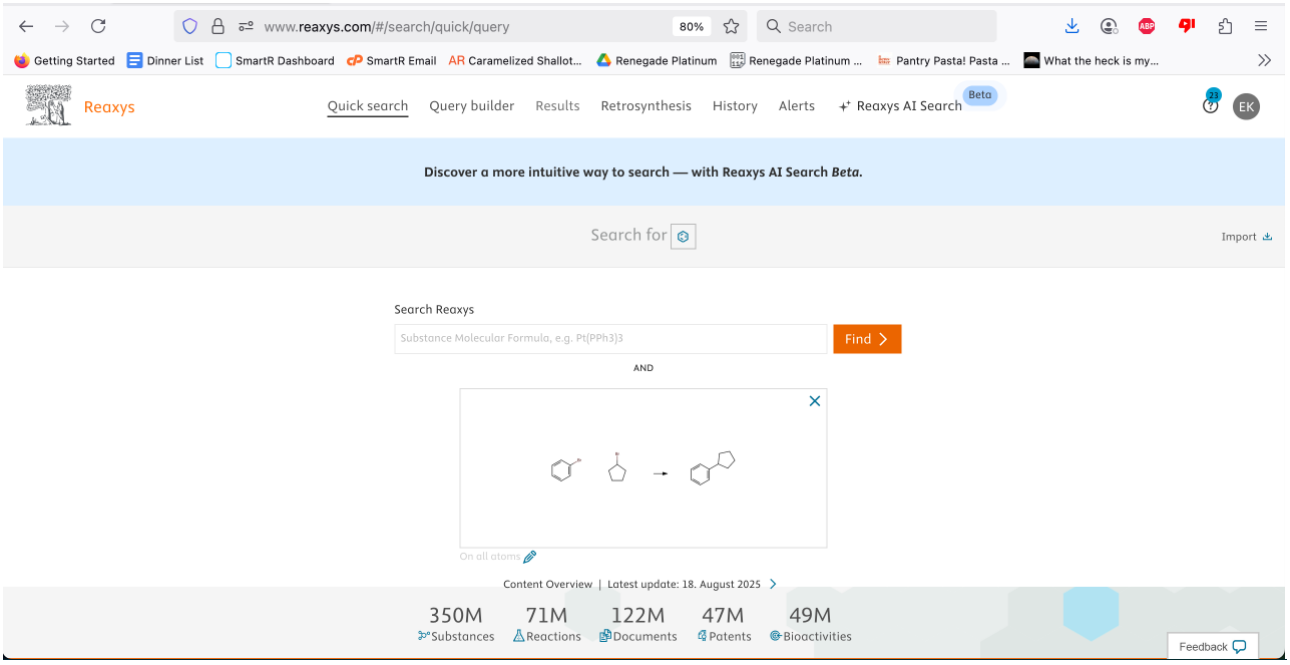
Reaxys employs a very similar search engine. From the homepage, press the "Draw" button.



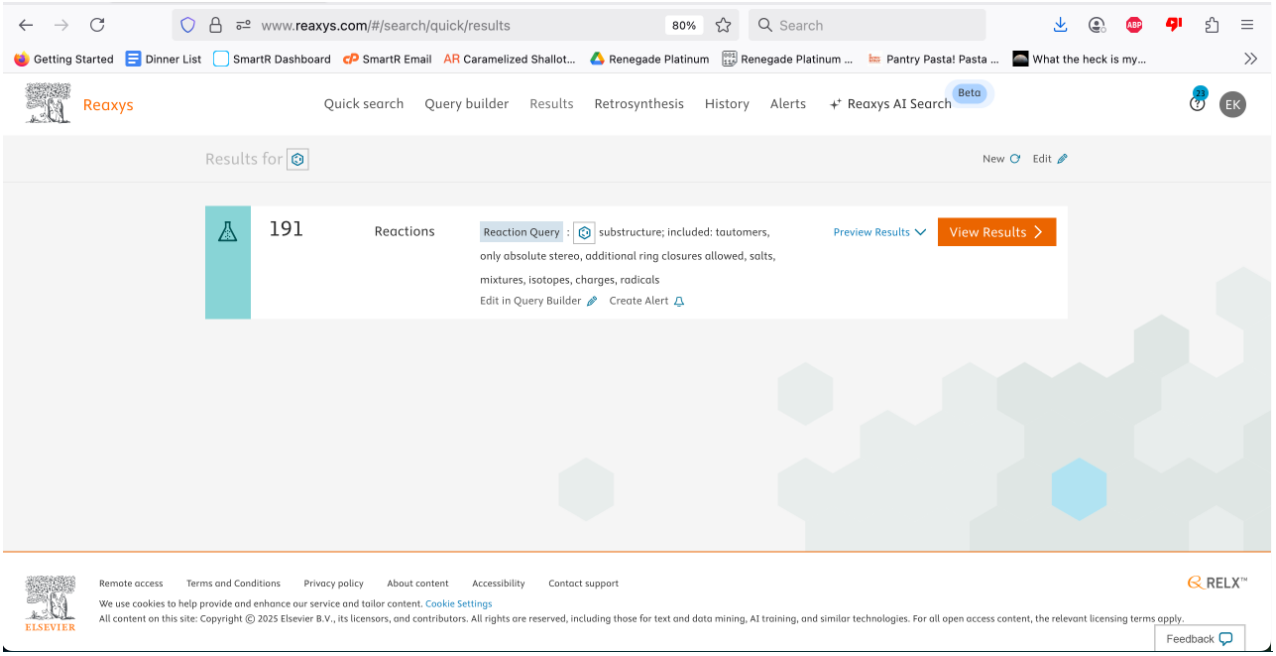
A popup screen will appear where you can draw your reaction of interest. On the right-hand side, you can search for "as drawn" matches (exact matches), "as substructure" matches, or "similar" matches. "As substructure" has the option of matching all atoms or only heteroatoms. You can also select whether or not to include tautomers, stereochemistry, etc. in your matching.



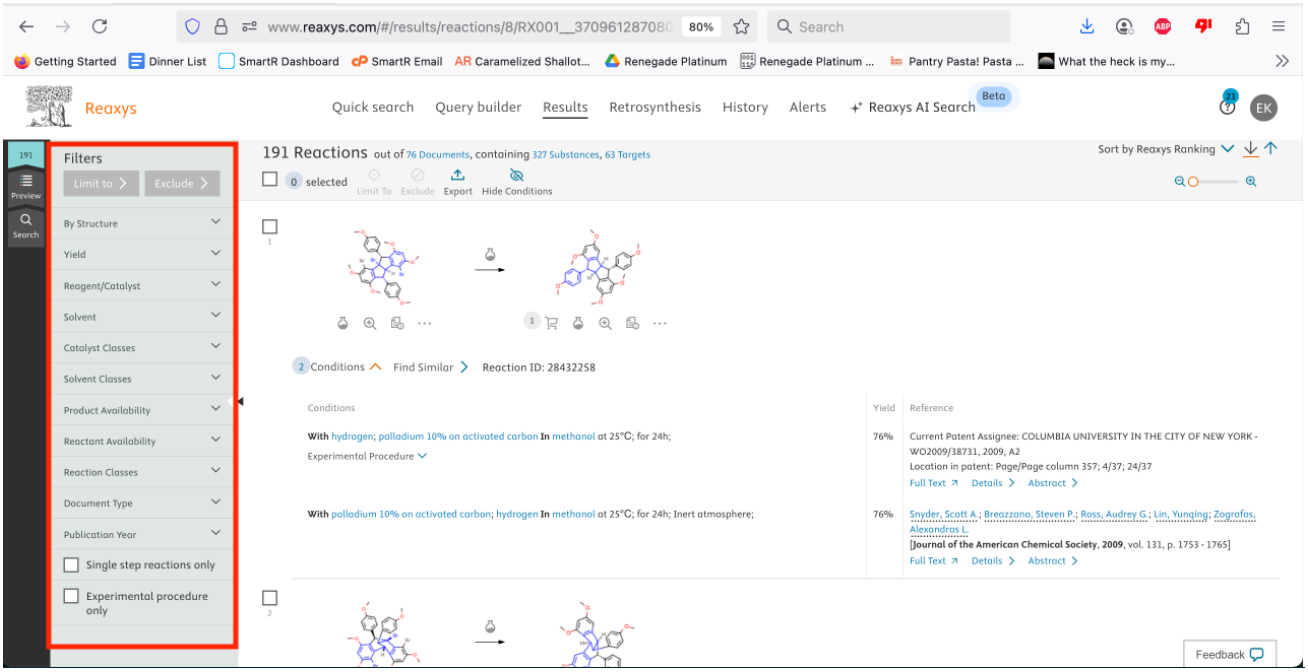
When done, click "Transfer to query" which will bring you back to the homepage.



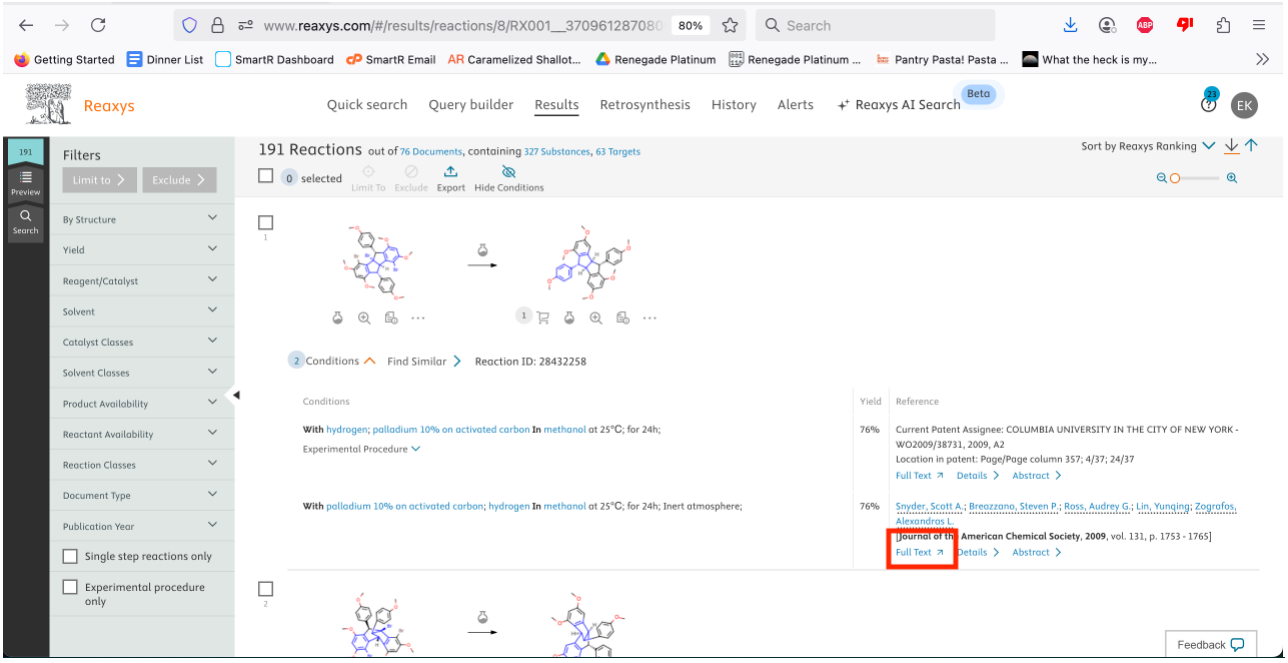
Then click "Find".



We see that 191 reactions match our reaction pattern either exactly or as a substructure. To view results, click "View Results". Then we can perform further filtering by yield, journal, reagent, etc.



You can also easily navigate to the primary source from both. Click on "Full Text" which will redirect you to the article or patent in question.

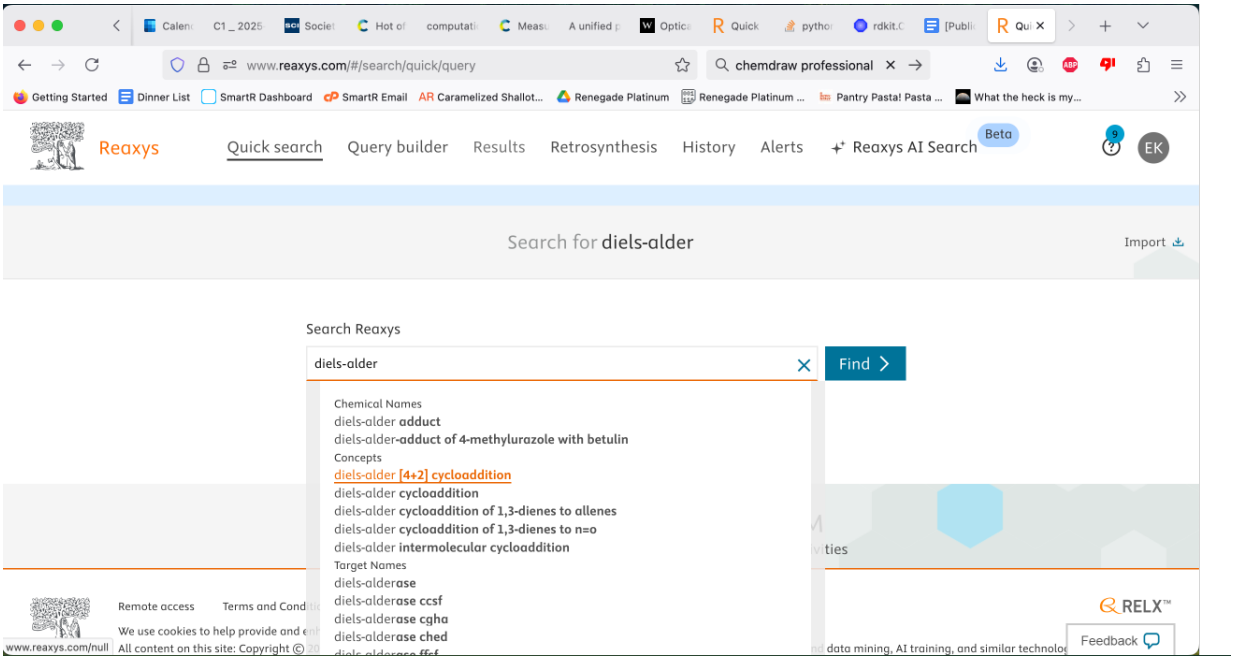


TIP: If you don't get any hits for your reaction in SciFinder or Reaxys, try making it more general: remove rings or functional groups that do not play a part in the reaction. If you get too many irrelevant hits, try adding explicit hydrogens or use the atom mapping functions.

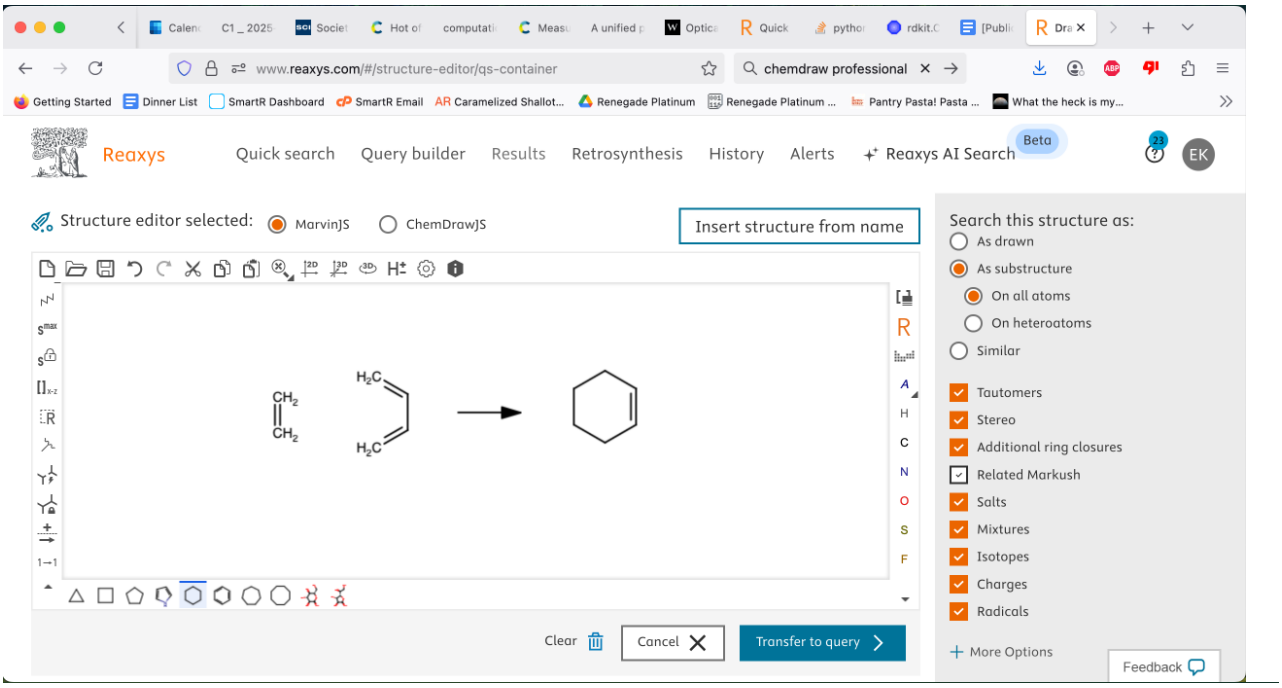
Curating Data from Reaxys

We will often need to augment our own experimental datasets with information from Reaxys. SciFinder is a great resource for synthesis, but their download feature is not as good as Reaxys'. There are a couple of ways to gather information. The first is through the Quick Search feature.

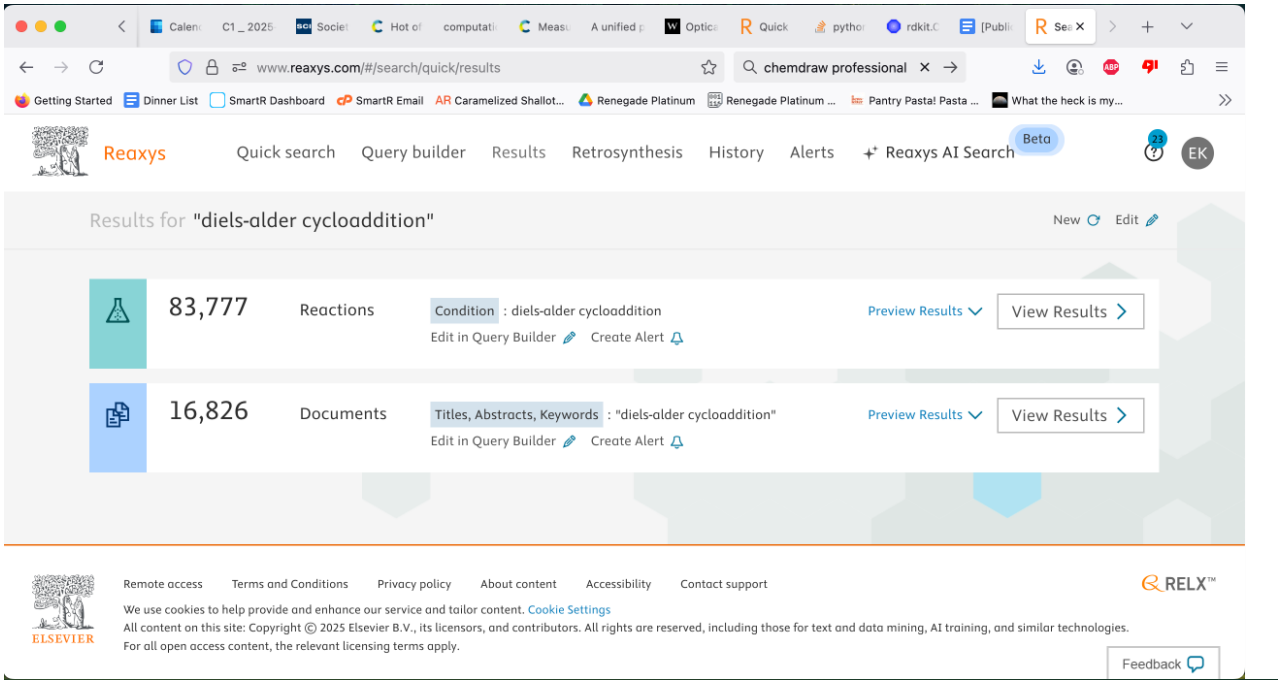
Here we see the potential Quick Search options for "Diels-Alder". When searching for reactions that fit a particular mechanism, you want to select options in "concepts". Click on the concept of interest then click "Find".



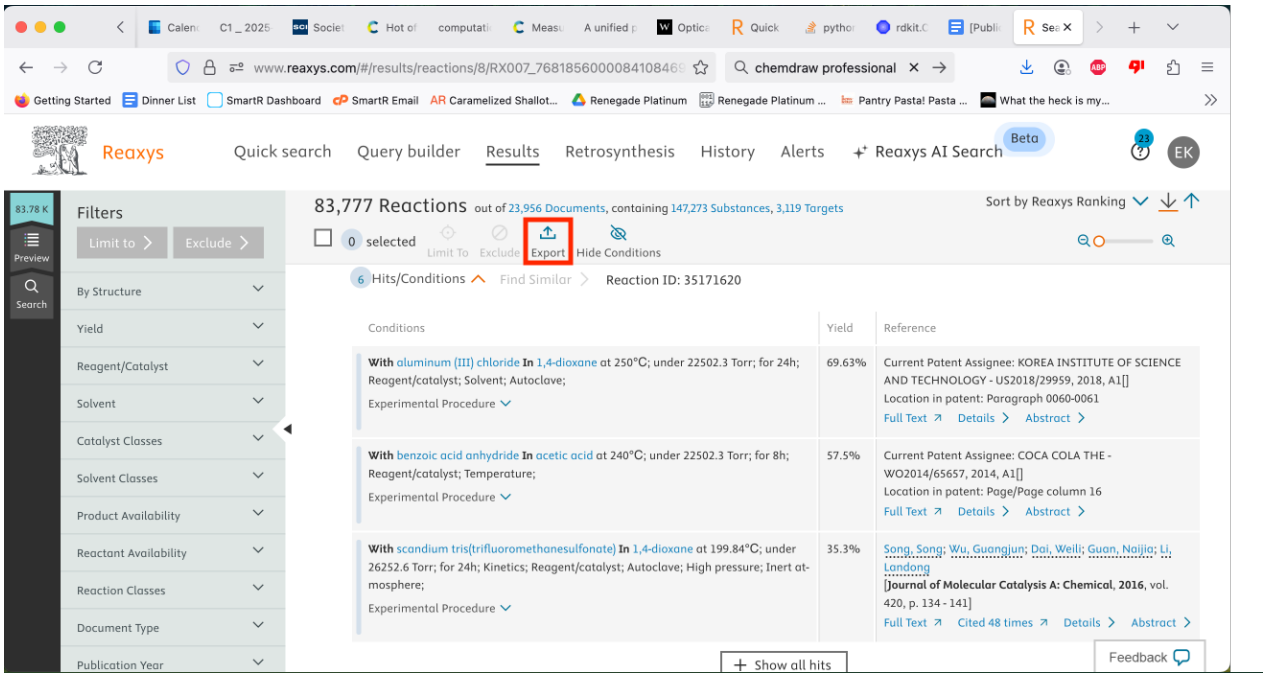
Alternatively, you can draw the reaction of interest and hit "Transfer to query", then "Find".

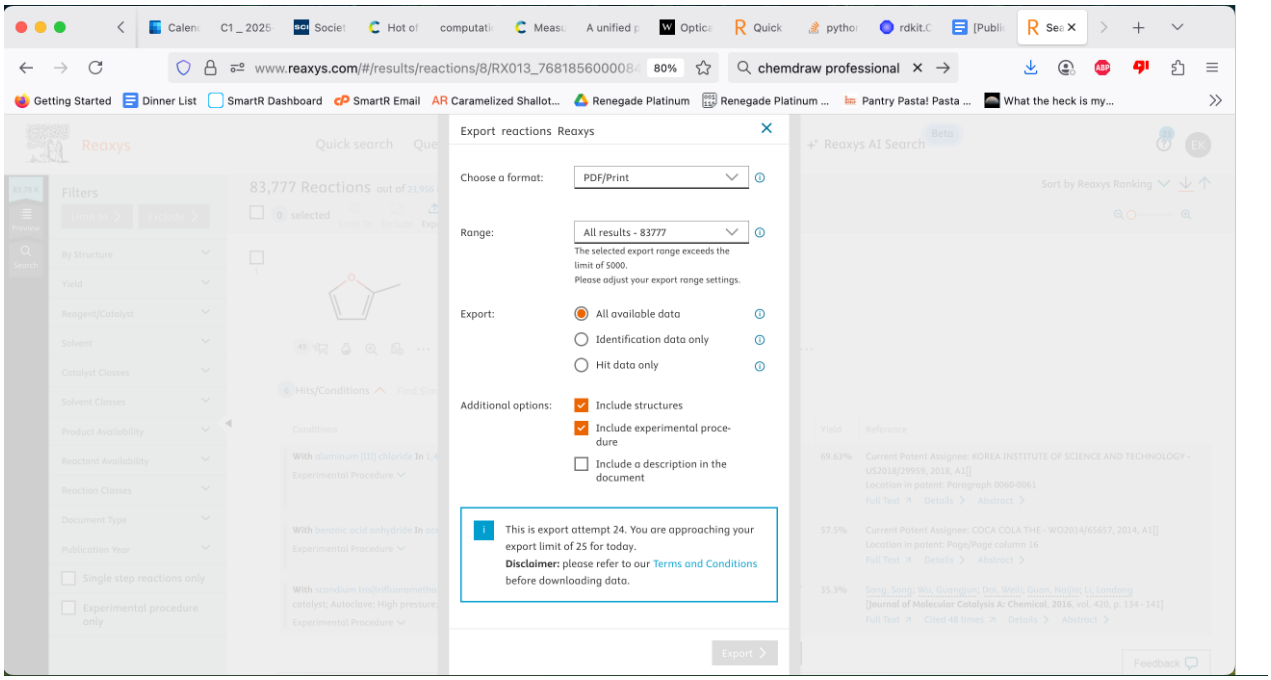


In either case, "Reactions" are what we are interested in. Click "View Results".



All the reactions that Reaxys believes match your query will appear in list form. To export the reactions, click on "Export". A box will pop up.





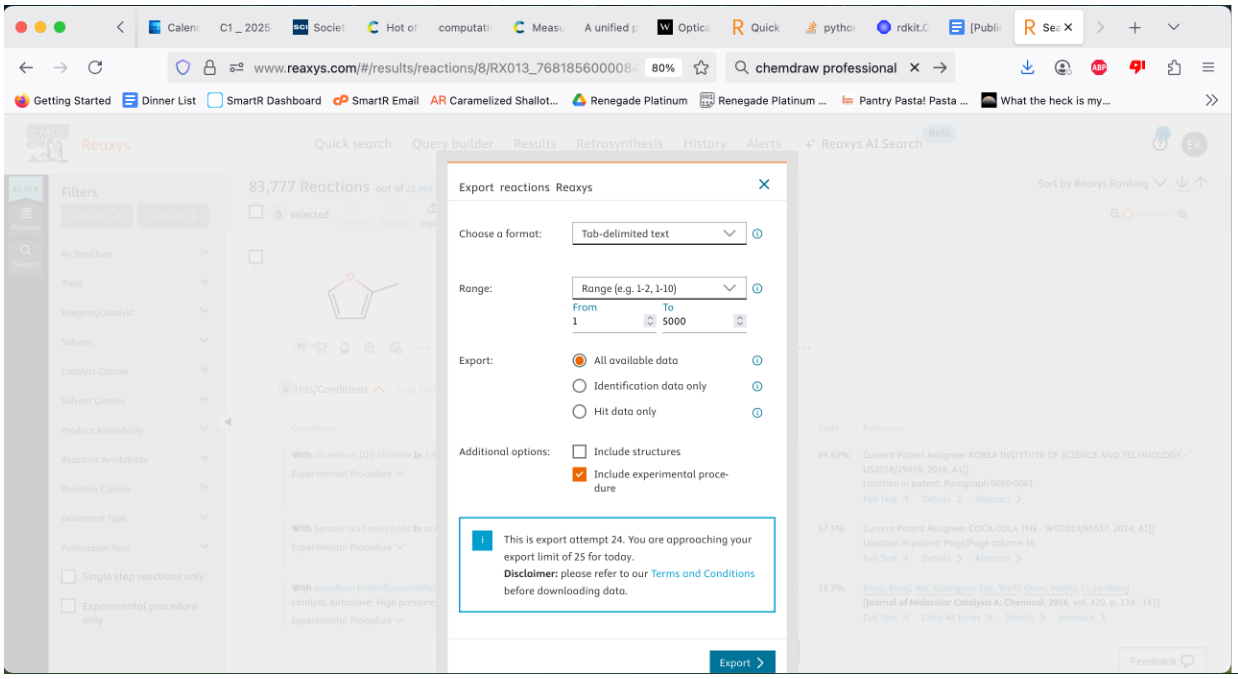
For format, typically we want to export to a tab-delimited text.

For range, we can only export 5,000 entries at a time, so set the range accordingly. It is normal to do 5K entries at a time until you get all of the reported reactions. **NOTE:** You only get 25 downloads a day, so at most you can download 125,000 reactions in a 24-hour period.

For export, we will typically want to get all available data.

We typically include experimental procedures but do not include structures.

Once you have entered all the information, hit "Export".



Your download will start and will let you know when you've finished as a pop up in the bottom right.