258th ACS National Meeting San Diego Aug 2019

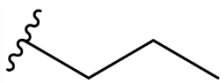
Measuring R group similarity using medicinal chemistry data

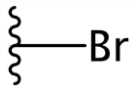
Noel O'Boyle and Roger Sayle

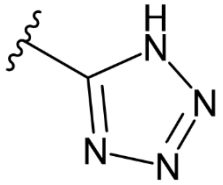
NextMove Software

INTRODUCTION

PREDICTING THE PAST







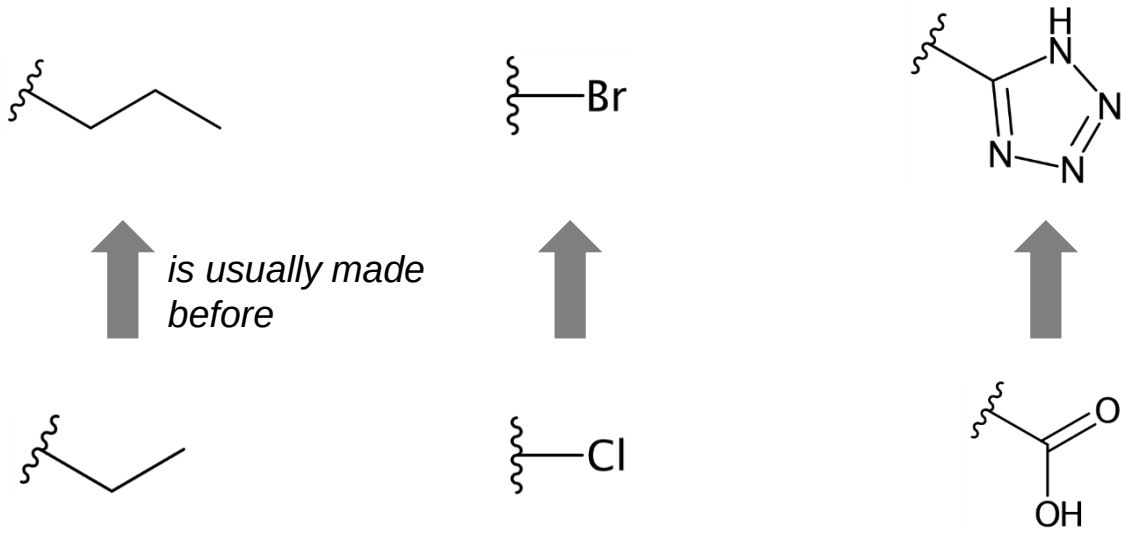
is usually made before

?

?

?

PREDICTING THE PAST



- Given a med chem project containing a molecule with a Br, it is likely that a Cl was tried at that position at an earlier stage
- This relationship between Cl and Br is a measure of similarity
  - A type of similarity relevant to medicinal chemistry projects
  - A type of similarity undetected by molecular fingerprints
- Can we find all pairs of R groups similarly related and exploit these relationships?

TIME SERIES DATA

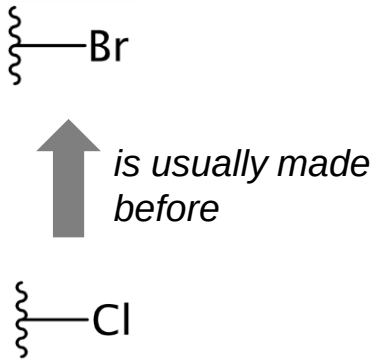
- The ideal dataset is time series data from one or more large pharmaceutical companies over multiple projects
  - Use date of registration to track project progression
  - Caveats: compounds may have been registered in a previous project, parallel syntheses may obscure conceptual order
- For each project, identify all matched pairs
  - For matched pairs of Br/Cl, was the chlorine analog generally made first? (etc…)

TIME SERIES DATA

- The ideal dataset is time series data from one or more large pharmaceutical companies over multiple projects
  - Use date of registration to track project progression
  - Caveats: compounds may have been registered in a previous project, parallel syntheses may obscure conceptual order
- For each project, identify all matched pairs
  - For matched pairs of Br/Cl, was the chlorine analog generally made first? (etc…)
- Unfortunately….
  - We don't have access to pharmaceutical data
  - ChEMBL/patents don't include time series information
    - Though I did try looking at papers published by the same author over time

CO-OCCURRENCE INFORMATION

- But ChEMBL/patents **do** have a large number of independent datasets
  - Each is a subset of data from a medicinal chemistry project
  - Co-occurrence of R group analogs within the same subset is informative
- If Br is generally preceded by Cl, then:
  - Molecules with Br will often co-occur with the Cl analog
  - Analogs with Cl will be more frequent than analogs with Br
    - Since after making/testing Cl, Br may not be made



FINDING CO-OCCURRENCE DATA

CHEMBL25

- Extract molecules associated with (non-duplicate) IC_{50}, EC_{50} or K_i data, where *doc_type* is publication or patent
- Within each document, fragment to find single-cut matched series
- Normalise R groups to the most common form using a tautomer hash
  - (see my other talk*)
- Multiple co-occurrences within the same document only count as a single observation

* CINF 55: ***Making a hash of it: The advantage of selectively leaving out structural information*** (https://nextmovesoftware.com/talks.html)

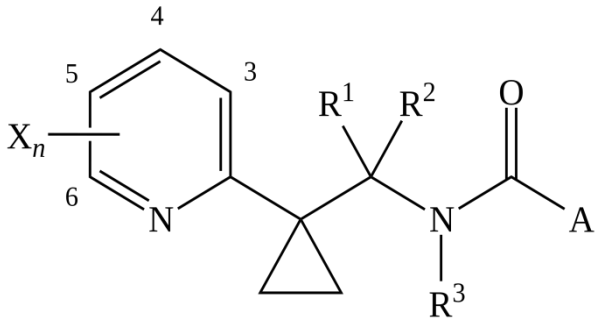
US PATENT APPLICATIONS

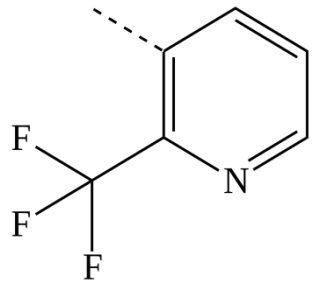
- LeadMine was used to extract 2.9M key molecules from pharmaceutically-related USPTO patent applications
  - Key molecules were those associated with bioactivity or present in R Group tables

TABLE 9

Inhibition of CETP Activity by Examples in Reconstituted Buffer Assay.

Ex. No.	IC_{50} (μM)	Ex. No.	IC_{50} (μM)	Ex. No.	IC_{50} (μM)
249	0.020	419	0.19	425	0.34
244	0.029	230	0.20	514	0.34
634	0.032	248	0.20	237	0.35
221	0.034	266	0.20	399	0.35
229	0.034	378	0.20	645	0.35



No.	X_n	R^1	R^2	R^3	A
1	6-Cl	H	H	H	

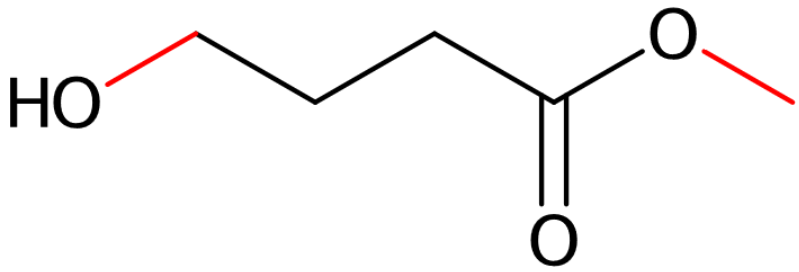
US PATENT APPLICATIONS

- LeadMine was used to extract 2.9M key molecules from pharmaceutically-related USPTO patent applications
  - Key molecules were those associated with bioactivity or present in R Group tables
- Patents were grouped into chemically-related patent families* (CRPFs)
- Molecules within the same patent were fragmented to find matched series
- Multiple co-occurrences within the same CRPF only count as a single observation

* https://nextmovesoftware.com/blog/2017/07/04/chemically-related-patent-families/

FRAGMENTATION SCHEME

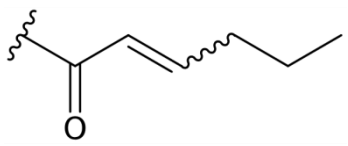
- Intended to generate synthetically-relevant matched series*
- Acyclic single bonds broken:
  - If either end is in a ring OR
  - If the bond is between a non sp^2-hybridised carbon atom and a non-carbon atom
- R group must be ≤ 20 heavy atoms

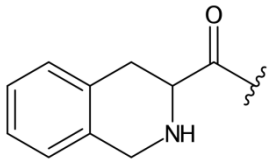


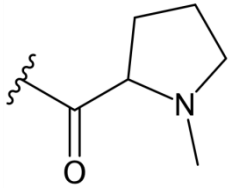
* As used in O'Boyle, Bostrom, Sayle, Gill. ***Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity***. *J. Med. Chem.* **2014**, *57*, 6.

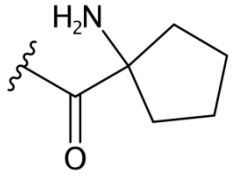
SINGLE-CUT MATCHED SERIES

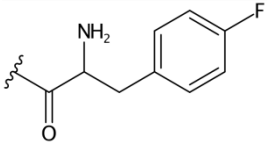
R Groups

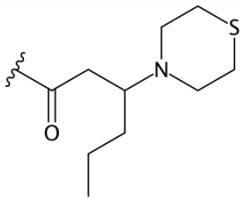


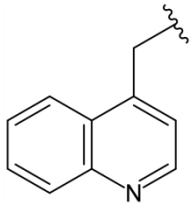


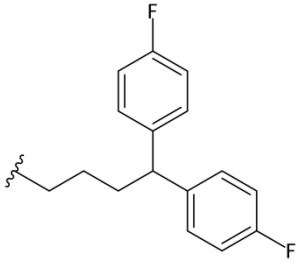


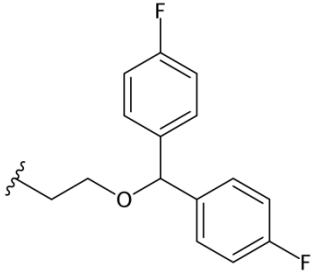


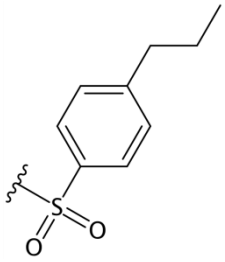




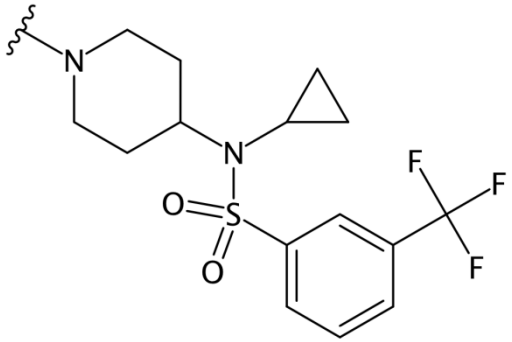






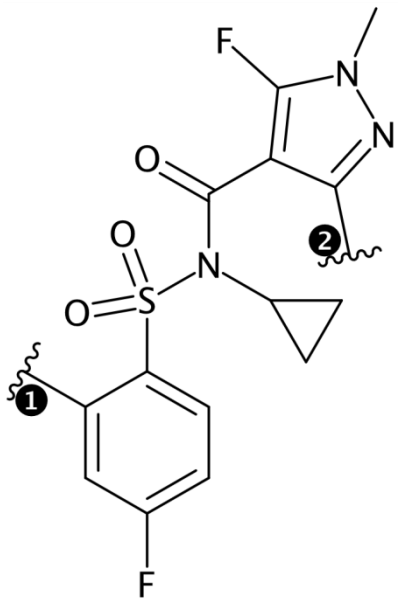


Scaffold



DOUBLE-CUT MATCHED SERIES

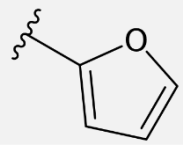
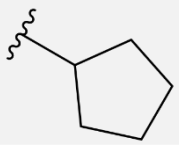
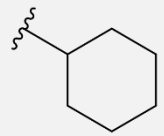
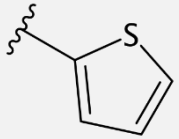
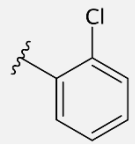
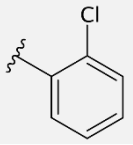
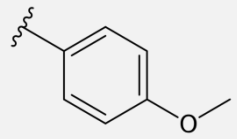
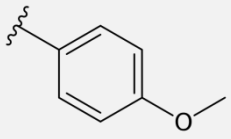
- To maximize recovery of R group co-occurrences, also infer co-occurrences from double-cut data

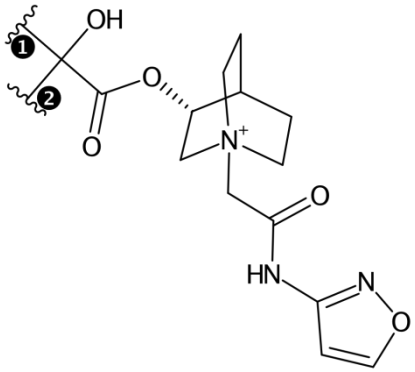


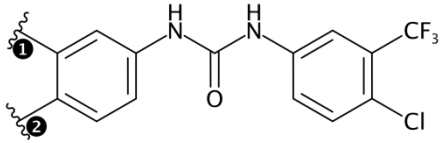
R1	R2
OH	
Br	
OH	
Br	
Cl	
OH	
OH	
Br	

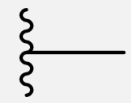
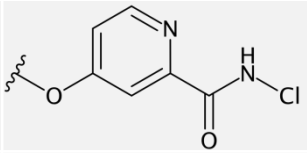
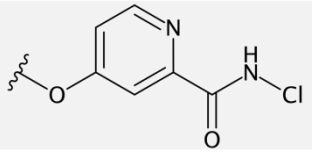
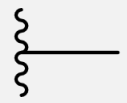
DOUBLE-CUT MATCHED PAIRS

- To maximize recovery of R group co-occurrences, also infer co-occurrences from double-cut data
- Carefully…
  - Where R1 and R2 positions are symmetry-related, discard if R1 ≠ R2
  - If any R1 appears at R2 (or *v.v.*), discard the whole series

R1	R2
	
	
	
	

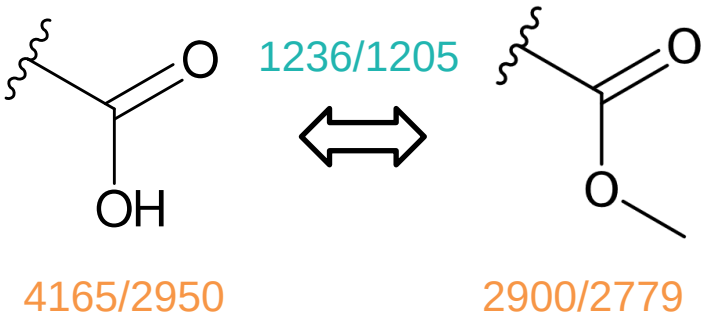




R1	R2
	
	

DATA-DRIVEN R GROUP SIMILARITY

FREQUENCY AND CO-OCCURRENCE OF R GROUPS

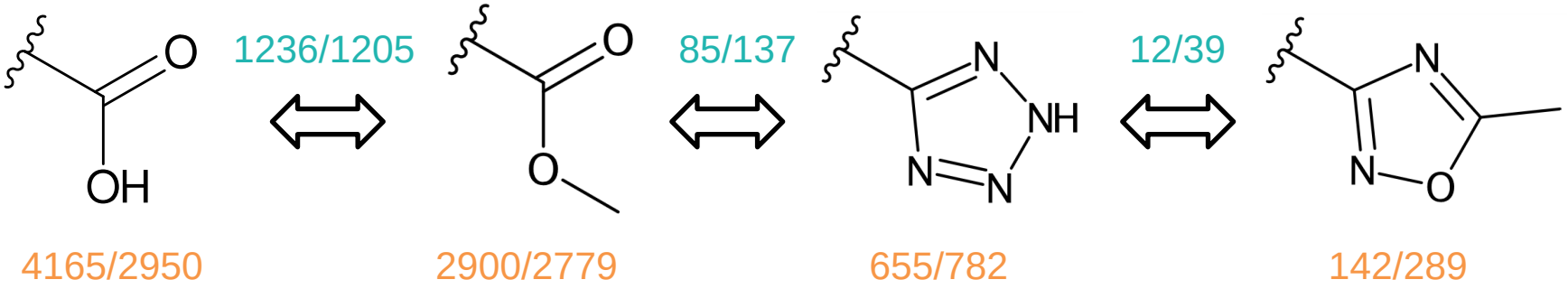


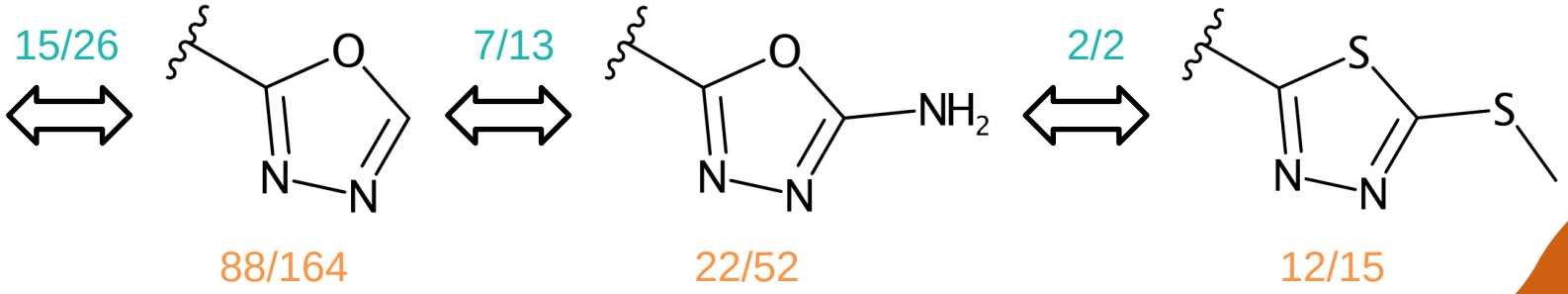
4165/2950 2900/2779

Color key: *frequency*
co-occurrence

Dataset key: ChEMBL/USPTO

FREQUENCY AND CO-OCCURRENCE OF R GROUPS

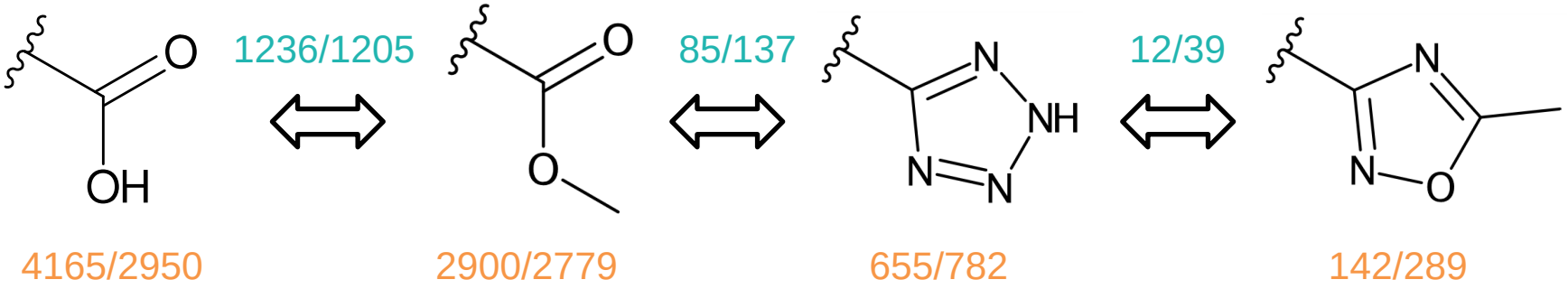


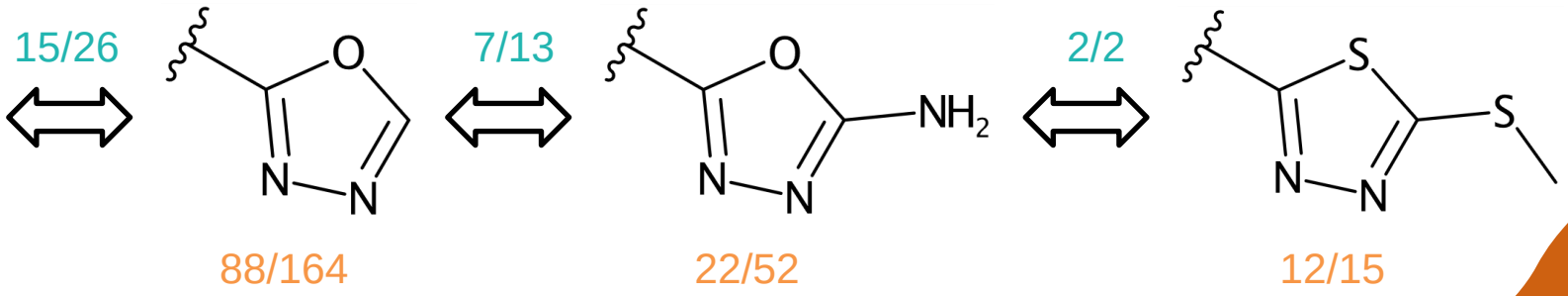


Color key: *frequency*
co-occurrence

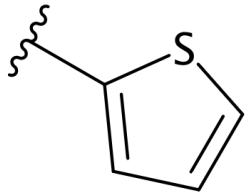
Dataset key: ChEMBL/USPTO

FREQUENCY AND CO-OCCURRENCE OF R GROUPS





Decreasing frequency ≈ likely to be synthesised afterwards (if at all)

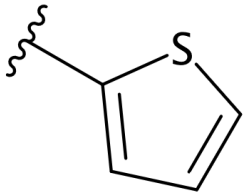
Find the 8 R groups most similar to 

More frequent

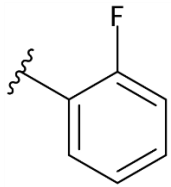
- We need to balance high co-occurrence against high frequency
  - Everything has a high co-occurrence with methyl, for example
- Rank R groups based on co-occurrence divided by frequency

Less frequent

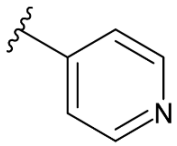
- Rank R groups based on co-occurrence

Find the 8 R groups most similar to 

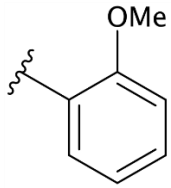
More frequent



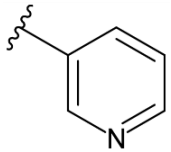
4/6



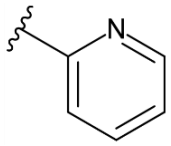
1/2



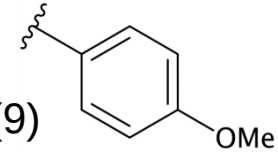
8/7



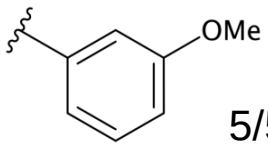
3/1



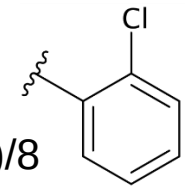
2/3



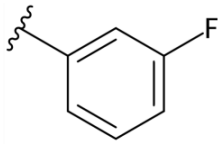
7/(9)



5/5

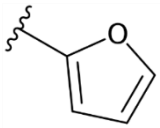


(12)/8

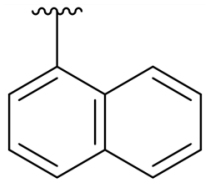


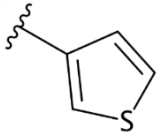
6/4

Less frequent

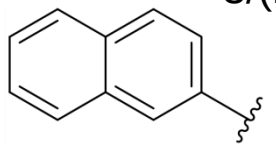


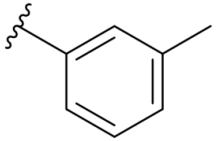
1/1

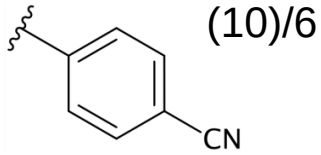

4/7



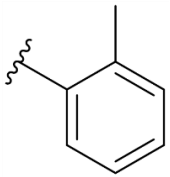
2/2

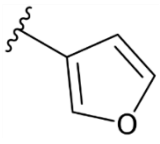

3/(10)


5/4

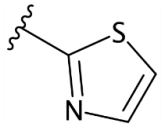


(10)/6


7/3

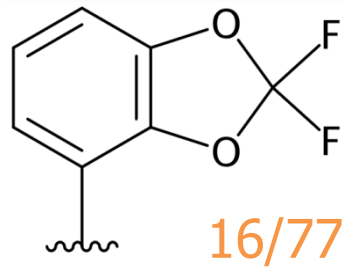


6/5



8/8

Find R groups similar to

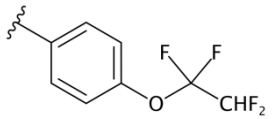


16/77

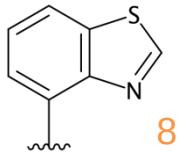
Color key: *frequency* *co-occurrence*

Dataset key: ChEMBL/USPTO

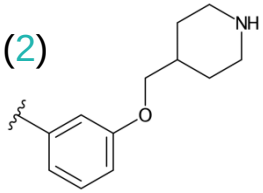
Less frequent (ChEMBL)



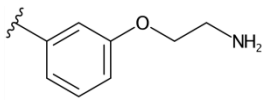
10 (2)



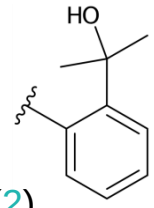
8 (2)



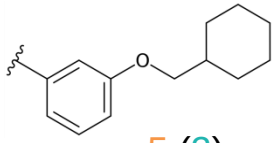
8 (2)



8 (2)



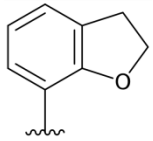
8 (2)

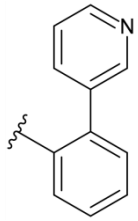


5 (2)

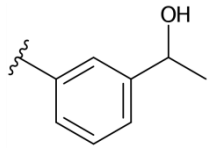
Less frequent (USPTO)

76(7)

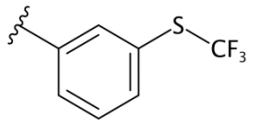




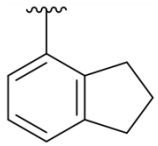
69(5)



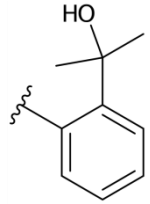
43(6)



65(5)

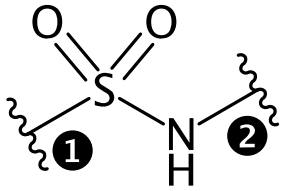


18(6)

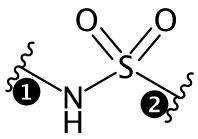


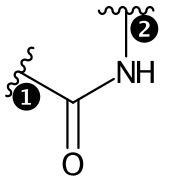
44(5)

plus 200 more, 56 with co-occurrence > 2

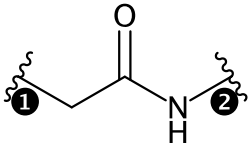
Find the 8 linkers most similar to 

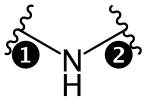
More frequent

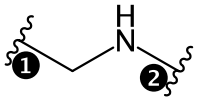
5 

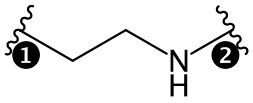
 1

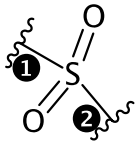
6 

 2

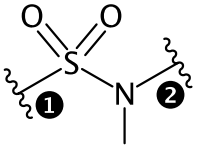
7 

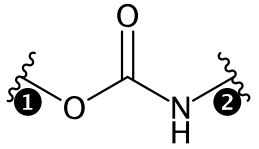
 3

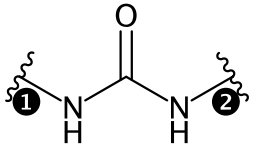
8 

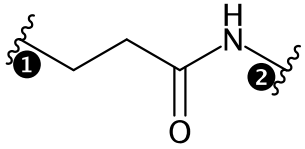
 4

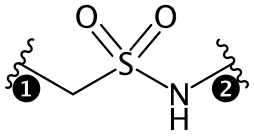
Less frequent

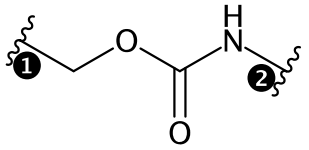
1 

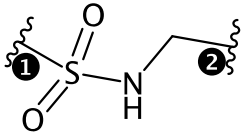
 5

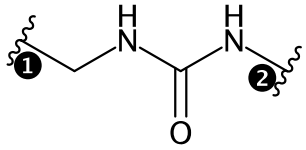
2 

 6

3 

 7

4 

 8

APPLICATIONS

- Ensure that gaps are covered
  - List R groups that should have been tested already
- Idea generation
  - List R groups that are commonly tested afterwards
- Enumerate
  - Generate molecules that are similar in R group space (and implicitly likely to be synthesisable)

- Similarity search…

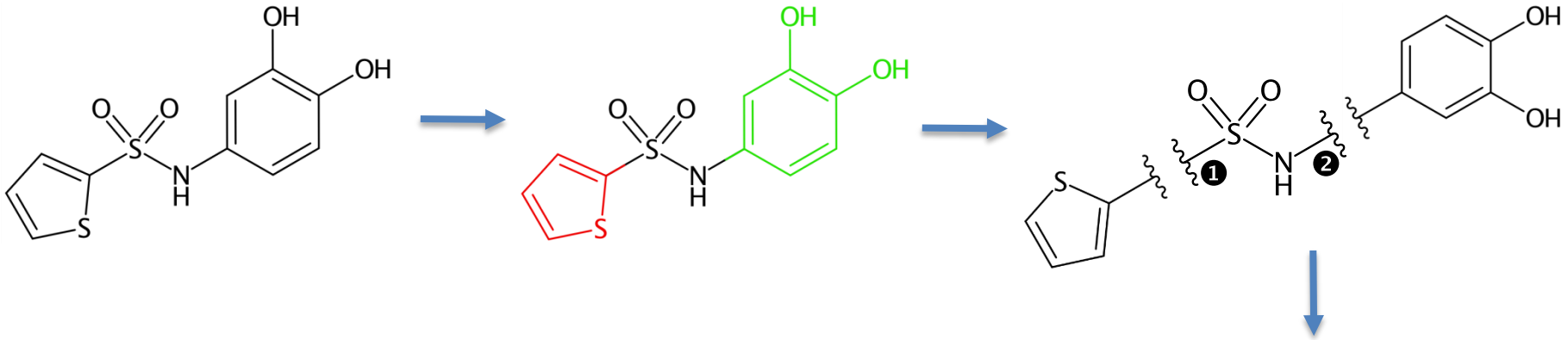
SIMILARITY SEARCH

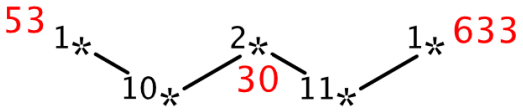
SIMILARITY SEARCH

- Are there molecules available for purchase that are similar in R group/linker space?
  - Worth purchasing and testing to complement in-house efforts
- Does ChEMBL contain bioactivity information for molecules that are similar in R group/linker space?
  - May affect patentability, or provide knowledge of off-target affects
- Are there exemplified compounds in patents* that are similar in R group/linker space?
  - May affect patentability, or provide knowledge of off-target affects

* For example, as extracted by LeadMine

FRAGMENTATION TO REDUCED GRAPHS

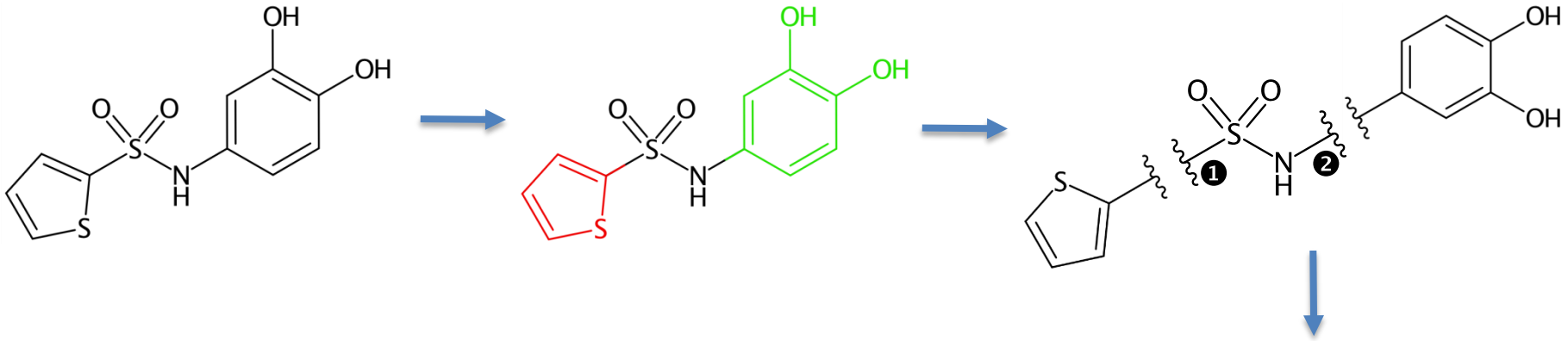


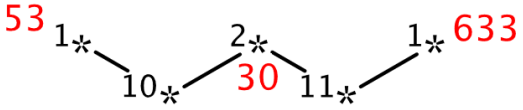


- Note that these reduced graphs:
  - Preserve the substitution position on rings
  - Preserve which end of the linker is attached to which R group

Birchall, K. and Gillet, V.J. (2011) Reduced graphs and their applications in chemoinformatics. In: Bajorath, J., (ed.) Chemoinformatics and Computational Chemical Biology. Methods in Molecular Biology, 672 . Humana Press , 197-212. (http://eprints.whiterose.ac.uk/78616/9/Gillet_reduced-graphs-revised.pdf)

FRAGMENTATION TO REDUCED GRAPHS

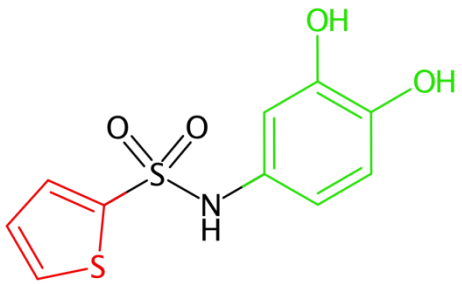




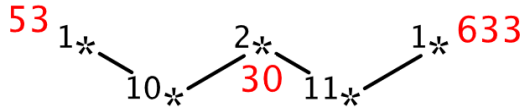
- Note that these reduced graphs:
  - Preserve the substitution position on rings
  - Preserve which end of the linker is attached to which R group
- Reduced graphs (RGs) are stored as SMILES strings
  - Isotopes indicate the node type: 1 for R group, 2 for linker, 10/11 indicate linker locants
  - Map indices indicate the specific identity of the nodes (via a lookup, e.g. R group 53 is 2-thiophenyl)

```
[1*:53][10*][2*:30][11*][1*:633]
```

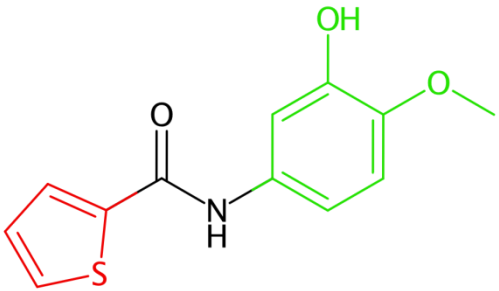
SIMILARITY SEARCH USING RGS

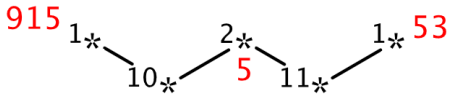


[1*:53][10*][2*:30][11*][1*:633]



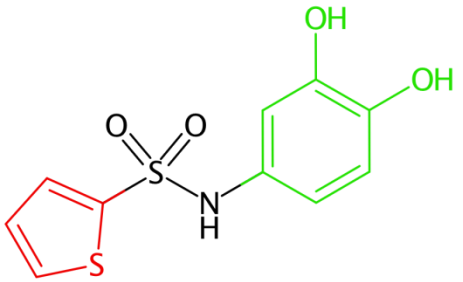
- Convert a database (e.g. ChEMBL) to reduced graphs
- For a given query:
  - Convert to reduced graph
  - Interpret as SMARTS and search the database
  - Score the hits based on distance to the query, a function of whether/where the matching groups appear on the query's list of nearest groups

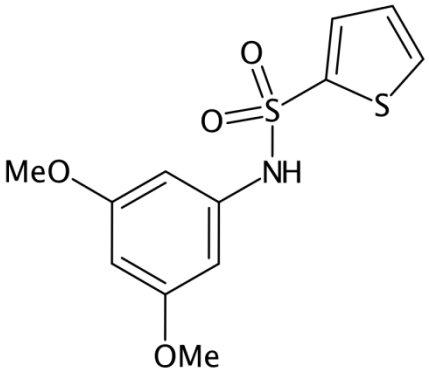


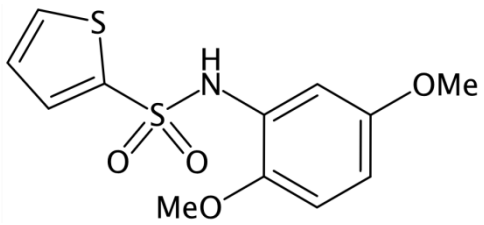


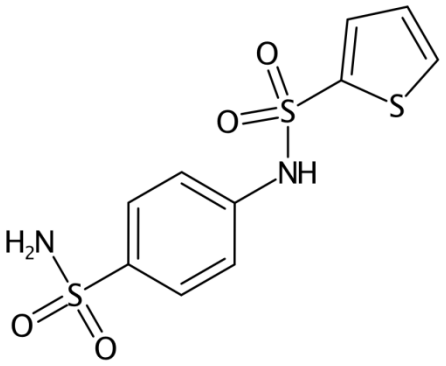
R group 53 is matched (does not contribute to distance)
R group 915 appears at rank 2 in 633's nearest (less frequent) R groups
Linker 5 appears at rank 1 in 30's nearest (more frequent) R groups

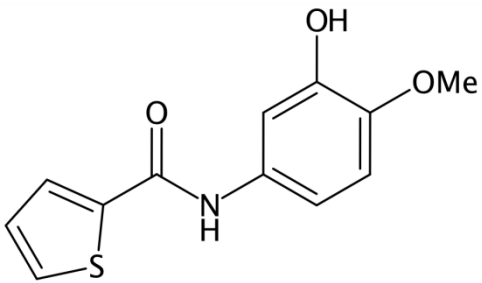
Search ChEMBL for molecules similar to

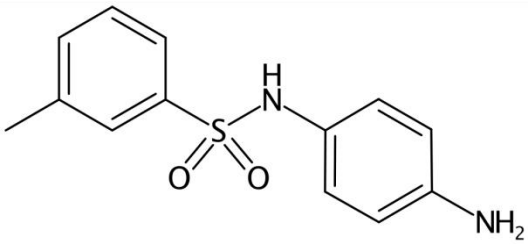


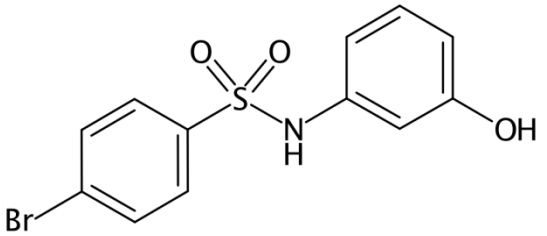


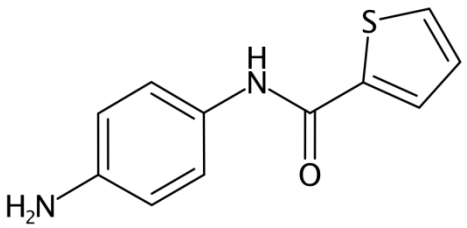


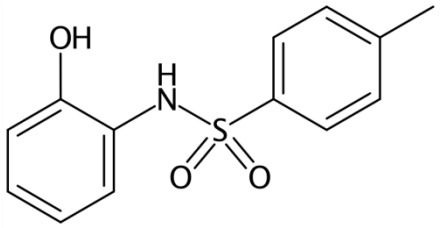


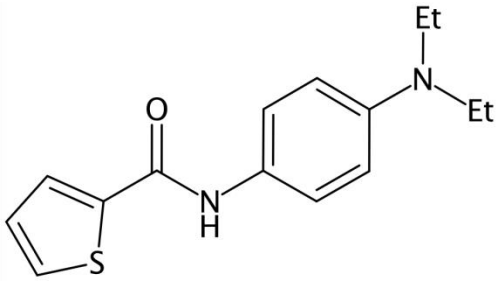












CONCLUSIONS

IN SUMMARY

- Co-occurrence of R groups in matched series from the same paper or patent can be used as an implicit measure of R group similarity
  - For less frequent R groups, patents provide more data
- Relative frequencies of R groups approximate the time dimension
  - Which R groups are typically made *after* versus *before*
- This approach can be used for idea generation, filling gaps, enumeration, and similarity search