Reaction Fingerprints: an early story

Gregory Landrum

NIBR IT

Novartis Institutes for BioMedical Research

RDKit UGM 2013, Hinxton

NOVARTIS

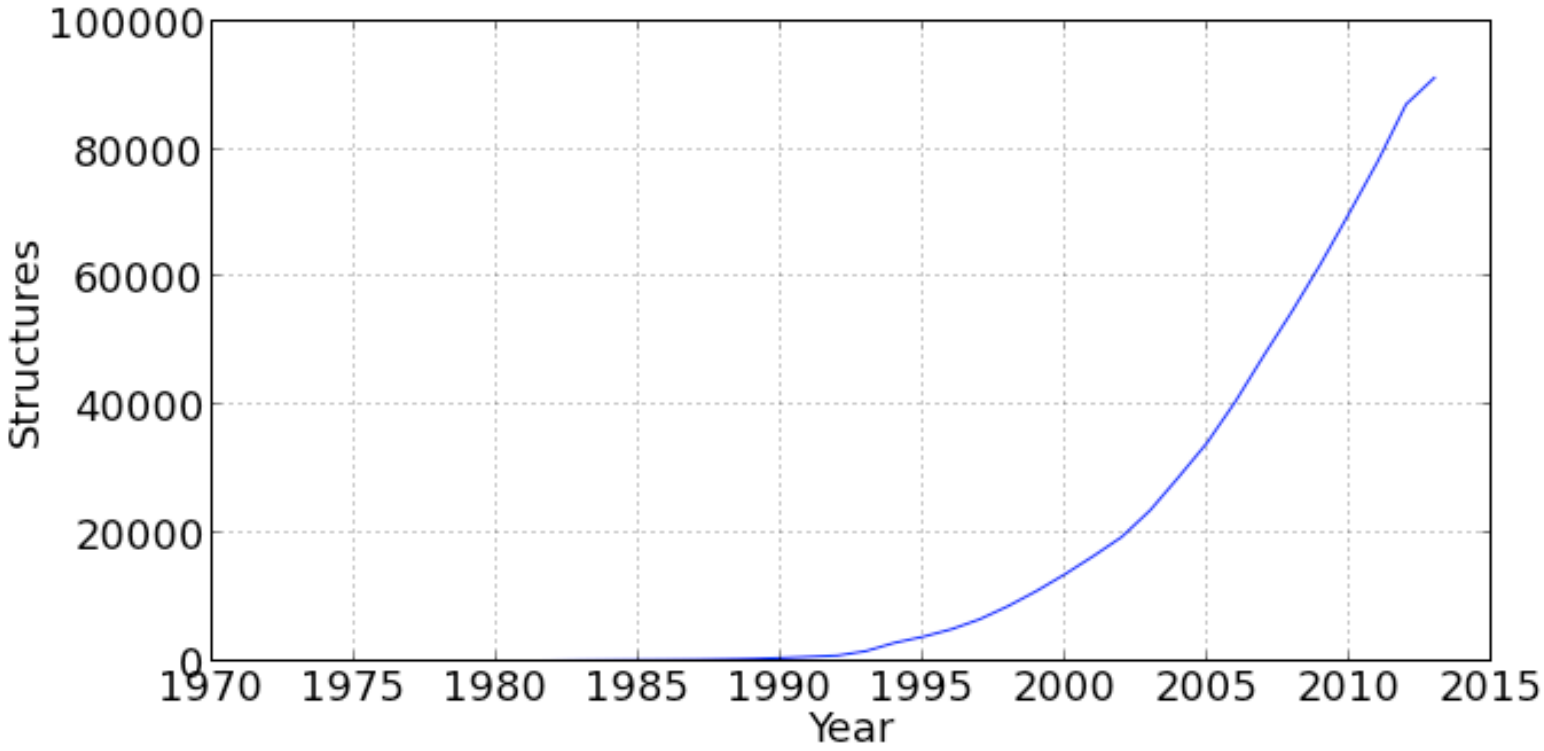
Public data sources in cheminformatics

an aside at the beginning

NOVARTIS

Protein data bank

the exception

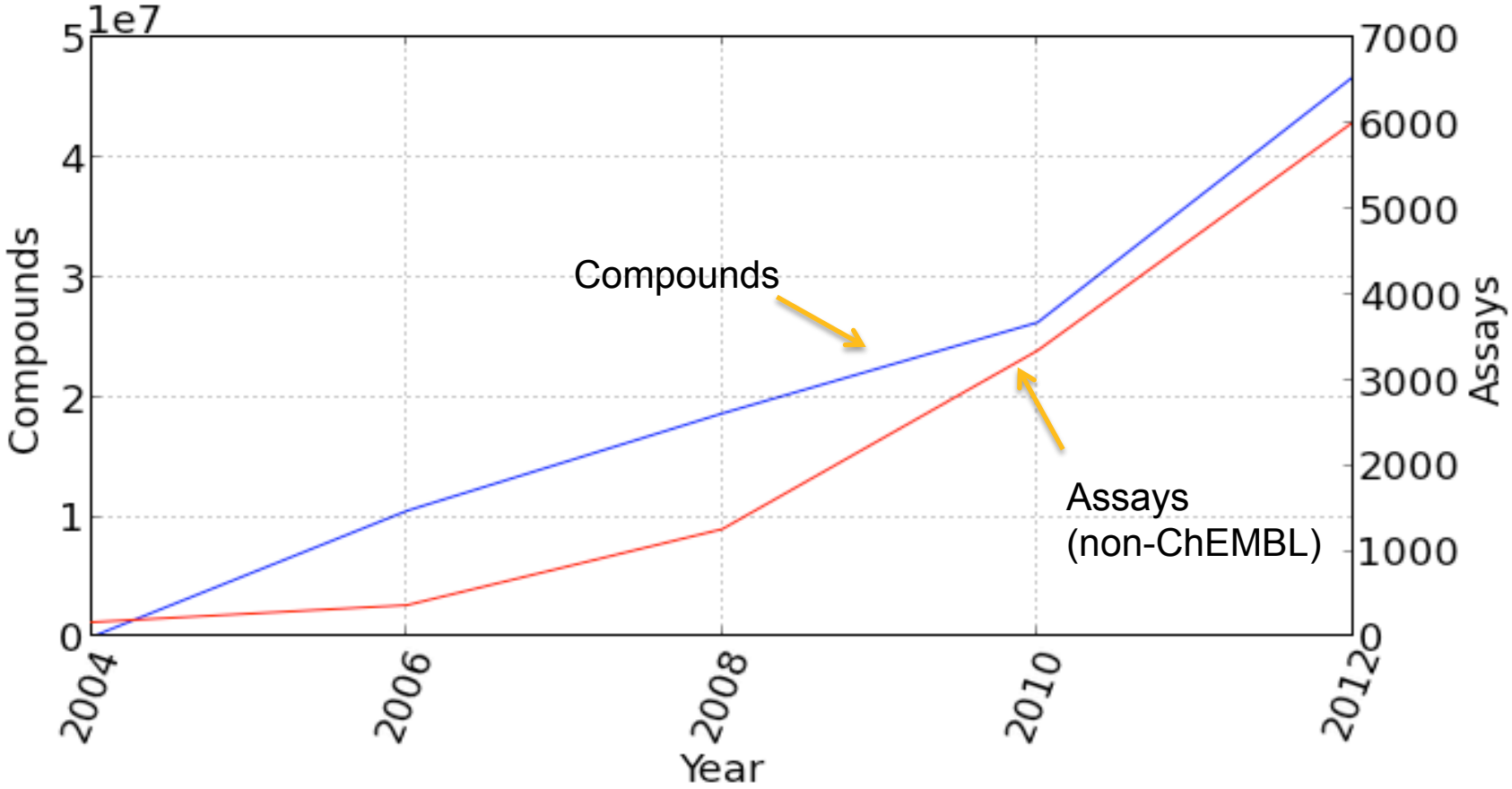


- Crystal structures of proteins
- Deposition is mandatory for publishing protein crystal structures

NOVARTIS

Pubchem

Evolution

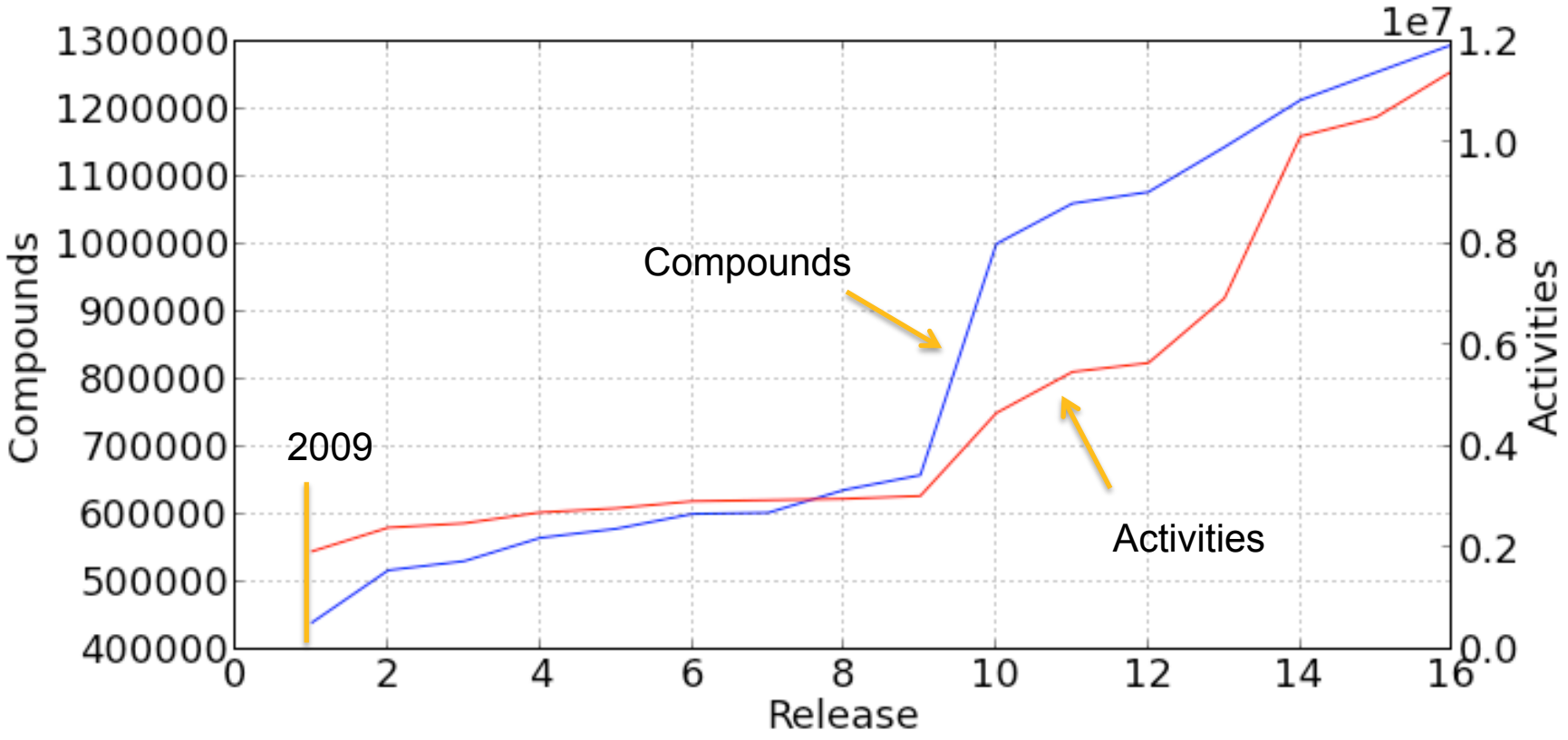


Collection of molecules from vendors and patents together with some assay data, primarily from NIH-funded screening centers.

NOVARTIS

ChEMBL

Evolution



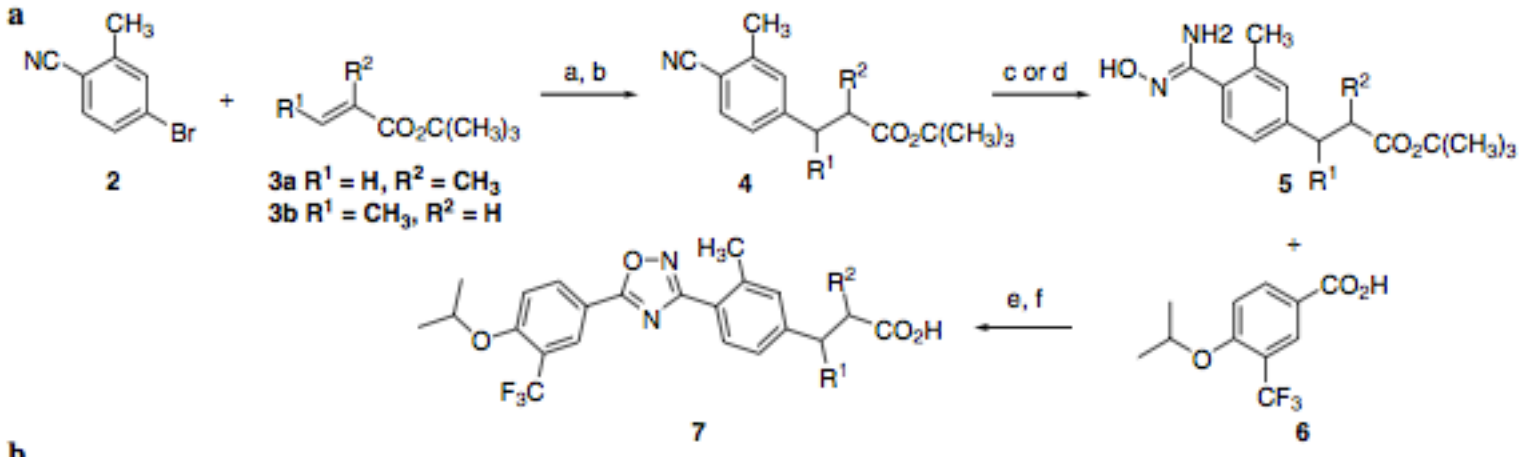
Collection of molecules and assay data curated (primarily) from the literature

NOVARTIS

What about how we made those molecules?

Public reaction data?

- The literature:



- Plenty of data locked up in large commercial databases, very very little in the open

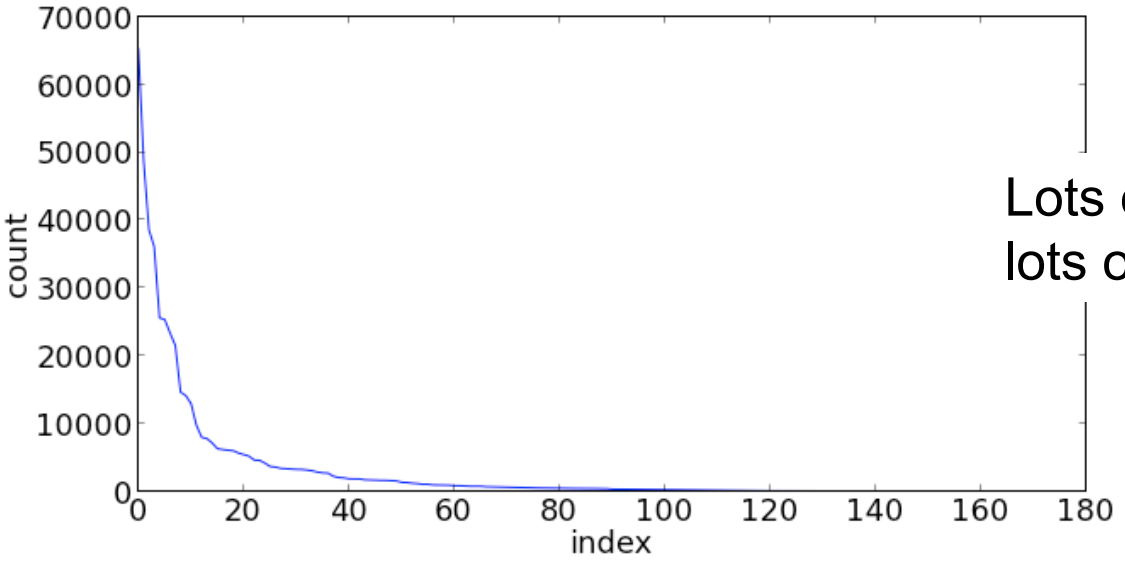
Yan, L. *et al.* SAR studies of 3-arylpropionic acids as potent and selective agonists of sphingosine-1-phosphate receptor-1 (S1P1) with enhanced pharmacokinetic properties. *Bioorganic & Medicinal Chemistry Letters* **17,** 828–831 (2007).

NOVARTIS

An emerging area: chemical reactions

Not just what we made, but how we made it

- Text-mining applied to open patent data to extract chemical reactions : 1.12 million reactions[1]
- Reactions classified, when possible, into 156 standard types : >500000 classified reactions[2]



Lots of reactions,
lots of repeats

[1] Lowe DM: "Extraction of chemical structures and reactions from the literature." PhD thesis. University of Cambridge: Cambridge, UK; 2012.
[2] Reaction classification from Roger Sayle and Daniel Lowe (NextMove Software)

NOVARTIS

Got the reactions, what about reaction fingerprints?

Criteria for them to be useful

- Question 1: do they contain bits that are helpful in distinguishing reactions from another?

  Test: can we use them with a machine-learning approach to build a reaction classifier?

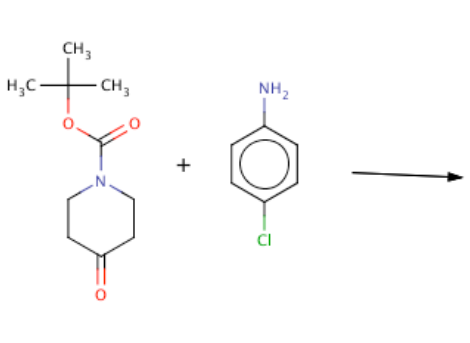
- Question 2: are similar reactions similar with the fingerprints

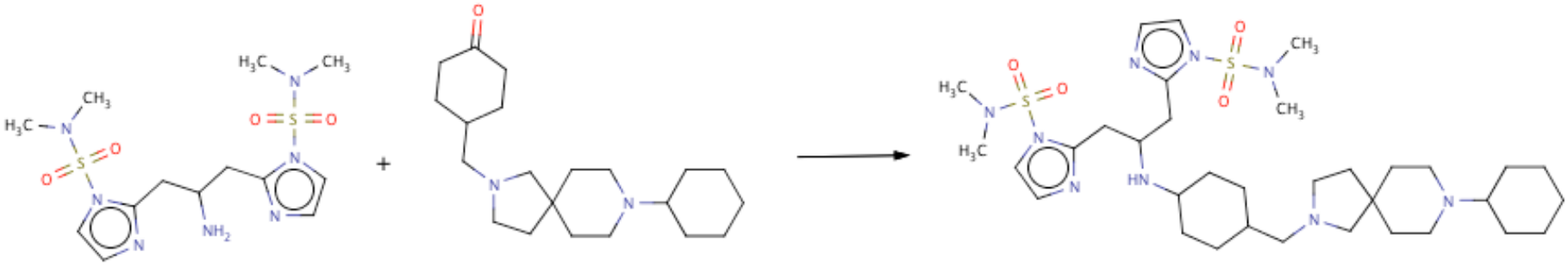
  Test: do related reactions cluster together?

Similarity applied to reactions

What are we talking about?

- These two reactions are both type: “1.2.3 Ketone reductive amination”





It’s obvious that these are the same, right?

Got the reactions, what about reaction fingerprints?

Start simple: use difference fingerprints:

$$FP_{\text{Reacts}} = \sum_{i \in \text{Reactants}} FP_i$$

$$FP_{\text{Products}} = \sum_{i \in \text{Products}} FP_i$$

$$FP_{\text{Rxn}} = FP_{\text{Prods}} - FP_{\text{Reacts}}$$

```
rapfp=None
for ri in range(rxn.GetNumReactantTemplates()):
    m = rxn.GetReactantTemplate(ri)
    fp = AllChem.GetAtomPairFingerprint(m,includeChirality=True)
    if rapfp is None:
        rapfp = fp
    else:
        rapfp += fp

papfp=None
for ri in range(rxn.GetNumProductTemplates()):
    m = rxn.GetProductTemplate(ri)
    fp = AllChem.GetAtomPairFingerprint(m,includeChirality=True)
    if papfp is None:
        papfp = fp
    else:
        papfp += fp

apfp = papfp-rapfp
```

Similar idea here: Ridder, L. & Wagener, M. SyGMa: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites. *ChemMedChem* **3,** 821–832 (2008).

Are these fingerprints useful?

- **Question 1:** do they contain bits that are helpful in distinguishing reactions from another?

  Test: can we use them with a machine-learning approach to build a reaction classifier?

- Question 2: are similar reactions similar with the fingerprints

  Test: do related reactions cluster together?

Machine learning and chemical reactions

- Validation set:
  - The 40 reaction types with at least 2000 instances from the patent data set
    - 2 separation reaction types removed (11.2 Separation [Resolution] and 11.3 Chiral separation [Resolution])
    - Final: 38 reaction types

- Process:
  - Training set is 200 random instances of each reaction type
  - Test set is 1000 random instances of each reaction type
  - Learning: random forest (scikit-learn)

Learning reaction classes

Results

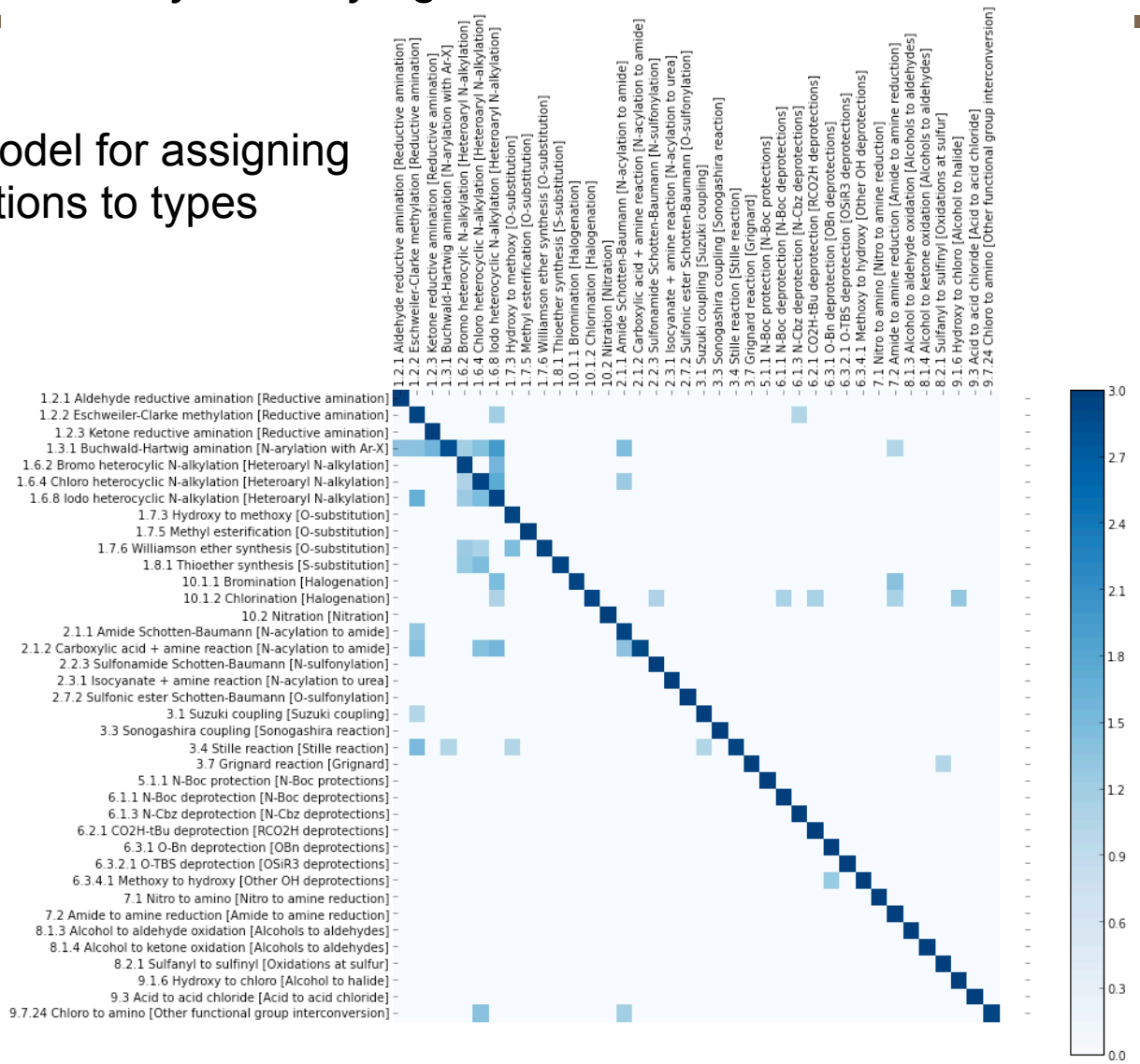
0 0.9890 0.9687 1.2.1 Aldehyde reductive amination [Reductive amination]
1 0.9410 0.8508 1.2.2 Eschweiler-Clarke methylation [Reductive amination]
2 0.9880 0.9518 1.2.3 Ketone reductive amination [Reductive amination]
3 0.6840 0.9434 1.3.1 Buchwald-Hartwig amination [N-arylation with Ar-X]
4 0.9200 0.9037 1.6.2 Bromo heterocylic N-alkylation [Heteroaryl N-alkylation]
5 0.8550 0.8350 1.6.4 Chloro heterocyclic N-alkylation [Heteroaryl N-alkylation]
6 0.8710 0.7400 1.6.8 Iodo heterocyclic N-alkylation [Heteroaryl N-alkylation]
7 0.9440 0.9347 1.7.3 Hydroxy to methoxy [O-substitution]
8 0.9770 0.9879 1.7.5 Methyl esterification [O-substitution]
9 0.9170 0.9797 1.7.6 Williamson ether synthesis [O-substitution]
10 0.9410 0.9731 1.8.1 Thioether synthesis [S-substitution]
11 0.9060 0.9912 10.1.1 Bromination [Halogenation]
12 0.8560 0.9761 10.1.2 Chlorination [Halogenation]
13 0.9710 0.9959 10.2 Nitration [Nitration]
14 0.9280 0.9143 2.1.1 Amide Schotten-Baumann [N-acylation to amide]
15 0.8160 0.9680 2.1.2 Carboxylic acid + amine reaction [N-acylation to amide]
16 0.9960 0.9774 2.2.3 Sulfonamide Schotten-Baumann [N-sulfonylation]
17 0.9940 0.9871 2.3.1 Isocyanate + amine reaction [N-acylation to urea]
18 0.9930 0.9960 2.7.2 Sulfonic ester Schotten-Baumann [O-sulfonylation]
19 0.9690 0.9808 3.1 Suzuki coupling [Suzuki coupling]
20 0.9960 0.9651 3.3 Sonogashira coupling [Sonogashira reaction]
21 0.9000 0.9688 3.4 Stille reaction [Stille reaction]
22 0.9700 0.9898 3.7 Grignard reaction [Grignard]
23 0.9560 0.9917 5.1.1 N-Boc protection [N-Boc protections]
24 0.9930 0.9315 6.1.1 N-Boc deprotection [N-Boc deprotections]
25 0.9980 0.9122 6.1.3 N-Cbz deprotection [N-Cbz deprotections]
26 0.9920 0.9538 6.2.1 CO2H-tBu deprotection [RCO2H deprotections]
27 0.9890 0.8966 6.3.1 O-Bn deprotection [OBn deprotections]
28 0.9990 0.9852 6.3.2.1 O-TBS deprotection [OSiR3 deprotections]
29 0.9670 0.9719 6.3.4.1 Methoxy to hydroxy [Other OH deprotections]
30 0.9940 0.9660 7.1 Nitro to amino [Nitro to amine reduction]
31 0.9780 0.8956 7.2 Amide to amine reduction [Amide to amine reduction]
32 0.9920 0.9754 8.1.3 Alcohol to aldehyde oxidation [Alcohols to aldehydes]
33 0.9850 0.9676 8.1.4 Alcohol to ketone oxidation [Alcohols to aldehydes]
34 0.9920 0.9773 8.2.1 Sulfanyl to sulfinyl [Oxidations at sulfur]
35 0.9770 0.9635 9.1.6 Hydroxy to chloro [Alcohol to halide]
36 0.9750 0.9750 9.3 Acid to acid chloride [Acid to acid chloride]
37 0.9250 0.9716 9.7.24 Chloro to amino [Other functional group interconversion]

overall >90% accuracy

Machine learning and chemical reactions

Automatically classifying reactions

Build a model for assigning new reactions to types



>90% accuracy

much of the confusion is between related types

NOVARTIS

Are these fingerprints useful?

- Question 1: do they contain bits that are helpful in distinguishing reactions from another?

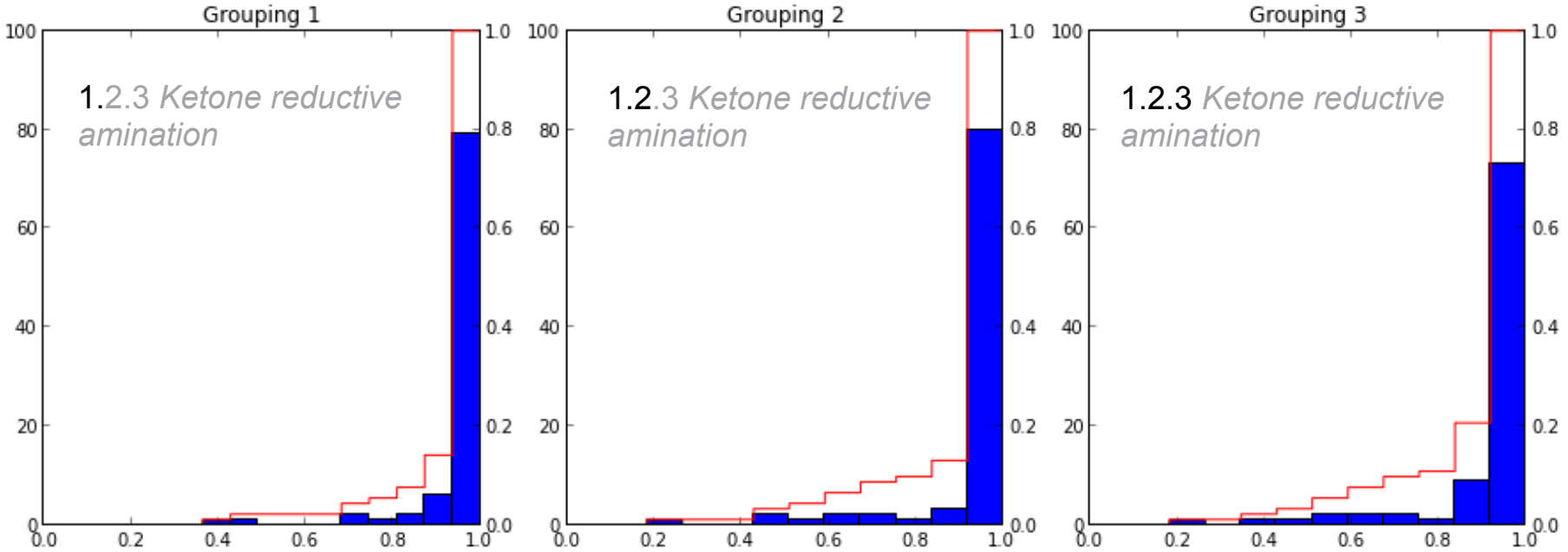
  Test: can we use them with a machine-learning approach to build a reaction classifier?

- **Question 2**: are similar reactions similar with the fingerprints

  Test: do related reactions cluster together?

Clustering reactions

- Reaction similarity validation set:
  - The 54 most common reaction types from the patent data set
  - Look at the homogeneity of clusters



More effort needed, but this is a pretty strong start.

NOVARTIS

Similarity applied to reactions

Can we help classify the remaining 600K reactions?

- Starting point: we have a similarity measure that clusters related reactions together
- We can apply the machine-learning model to the unclassified reactions and see if the original assignment missed any instances
- We can then look for big clusters of unclassified molecules and (manually) assign classes to them.

Acknowledgements

- NIBR:
  - Anna Pelliccioli
  - Sereina Riniker
- NextMove Software:
  - Roger Sayle
  - Daniel Lowe