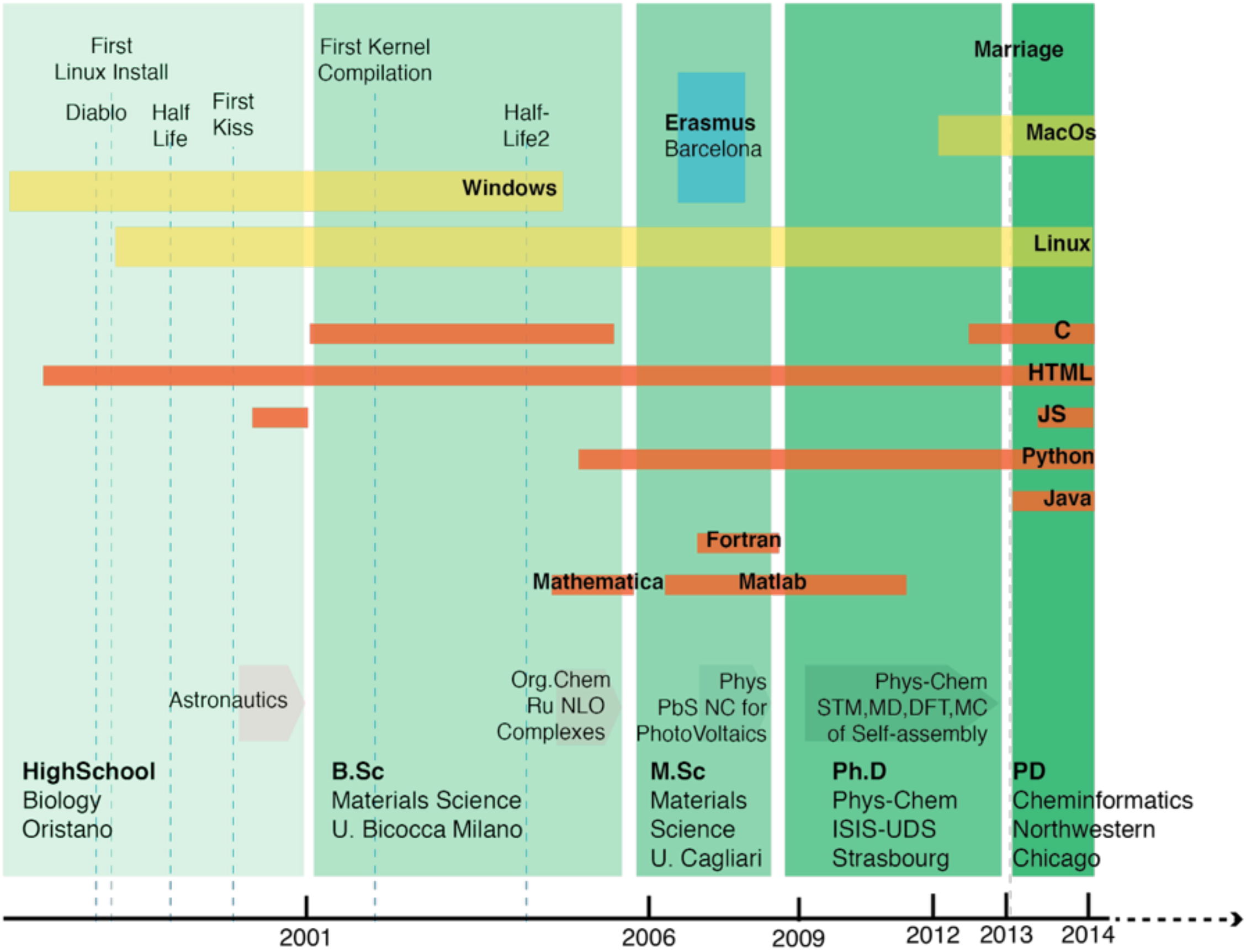


Andrea Cadeddu, Chematica Developer and PostDoc, Grzybowski group
K231 andrea.cadeddu@northwestern.edu

Chemistry is a language

Atoms are letters, molecules are the words, supramolecular entities are the sentences and the chapters.

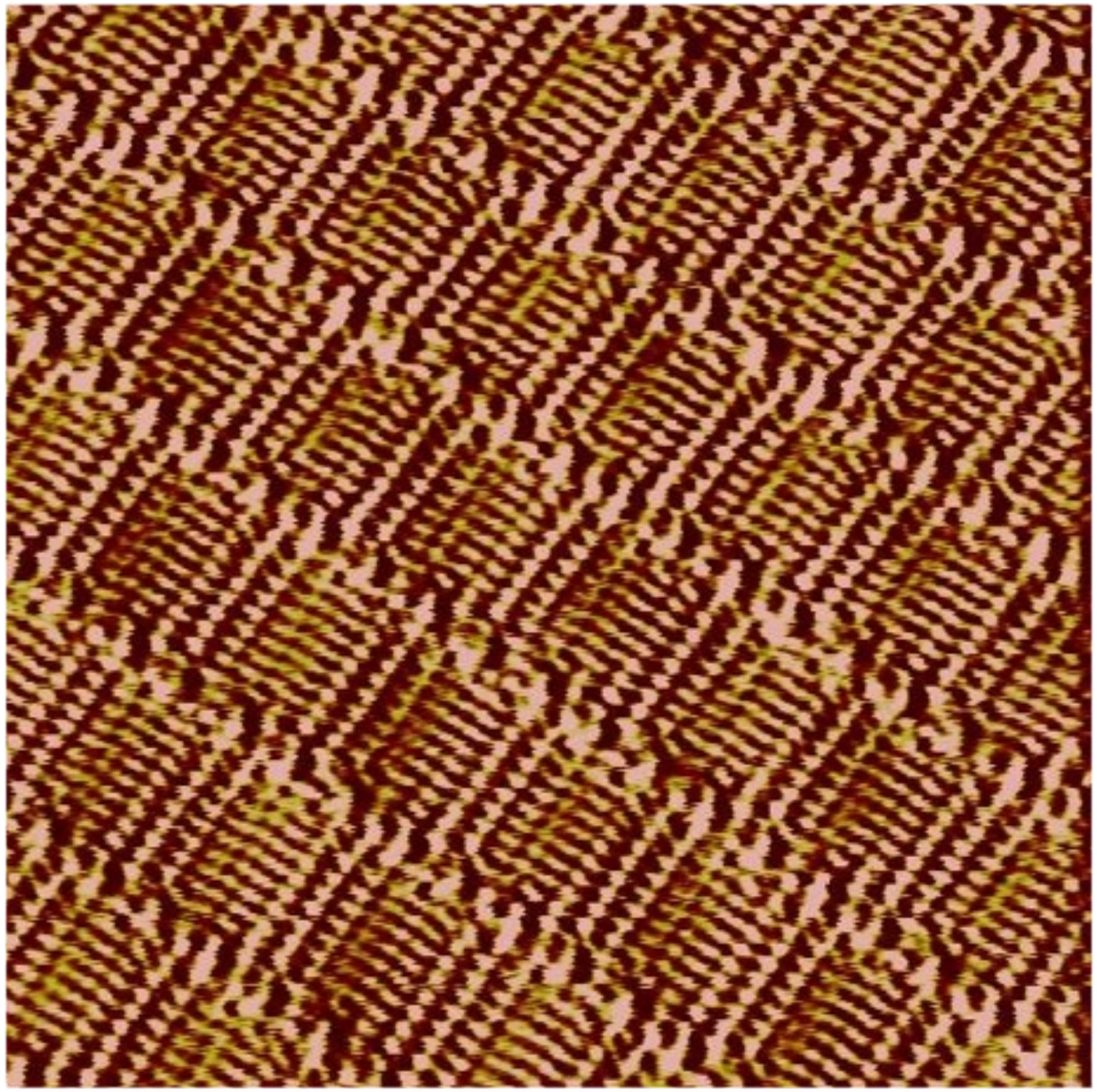
Jean Marie Lehn

1 1.01	6 12.01	7 14.01	8 16.00
H	**C**	**N**	**O**
Hydrogen	Carbon	Nitrogen	Oxygen

Bacon ipsum dolor sit amet pastrami ham hock ground round short loin leberkas tongue. Hamburger shankle strip steak, bacon sirloin shank cow capicola short ribs biltong doner ground round. Ball tip t-bone venison, fatback sirloin landjaeger beef. Pancetta prosciutto meatball, salami kielbasa kevin turducken andouille drumstick cow pork loin shoulder rump tongue tail. Pig andouille jowl, hamburger spare ribs bacon boudin.

Chicken capicola kielbasa jerky tri-tip, short ribs spare ribs beef biltong jowl beef ribs bacon. Frankfurter filet mignon pork loin short ribs, shoulder boudin leberkas tail sirloin. Shankle ham hock spare ribs, tongue kielbasa beef fatback pancetta ball tip chicken. Biltong sirloin chuck chicken ribeye short loin kielbasa tri-tip salami capicola prosciutto ball tip venison ham drumstick. Landjaeger beef boudin bacon rump.

Tenderloin swine pork loin fatback. Beef ribs bresaola fatback chicken kevin prosciutto pancetta rump sausage biltong pork. Pancetta swine venison turducken kielbasa meatloaf landjaeger prosciutto rump frankfurter. Sirloin chuck turducken, jerky brisket kevin salami pork belly rump shankle boudin venison. Beef tenderloin hamburger pork chop short ribs pancetta ground round short loin capicola kielbasa cow meatball. Flank pig shoulder kielbasa beef fatback short ribs bresaola. Spare ribs venison meatloaf, doner meatball kevin hamburger corned beef t-bone strip steak.



Arachidic acid on HOPG ca 20x20nm

What is a word in Chemistry?

Let’s take 10K sentences from Wikipedia

Let’s take 200 among those 10K.

Let’s combine them.

From today's featured article

The **Bohemian Waxwing** is a passerine bird that breeds in the northern forests of Eurasia and North America. It has mainly buff-grey plumage, black face markings and a pointed crest. Its wings are patterned white and bright yellow, and some feather tips have the red waxy appearance that give this species its English name. Its breeding habitat is coniferous forests, usually near water. The pair build a lined cup-shaped nest in a tree or bush for a

The Bohemian Waxwing is a passerine bird that breeds in the northern forests of Eurasia and North America.

Its breeding habitat is coniferous forests, usually near water.

TheBohemianWaxwingisapasserinebirdthatbreedsin-thenorthernforestsofEurasiaandNorthAmerica.

Itsbreedinghabitatisconiferousforests,usuallynearwater.

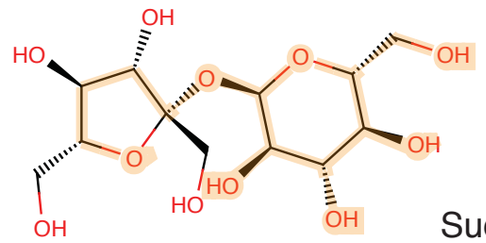
forests

Maximum common substring

Let’s take 10K molecules. (Corpus)

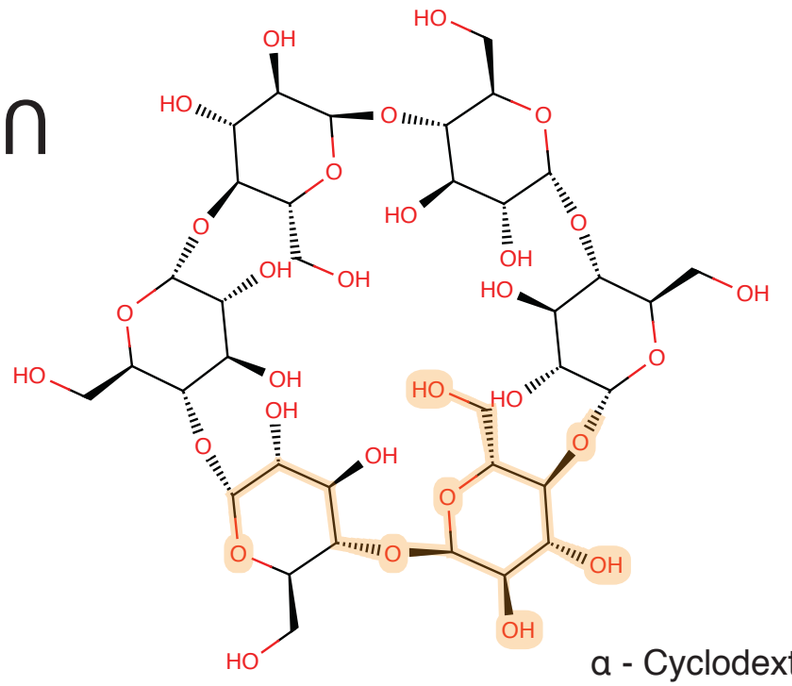
Let’s take 200 among those 10K.

Let’s combine them.



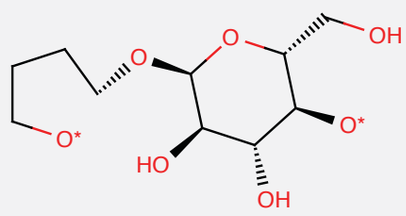
Sucrose

∩



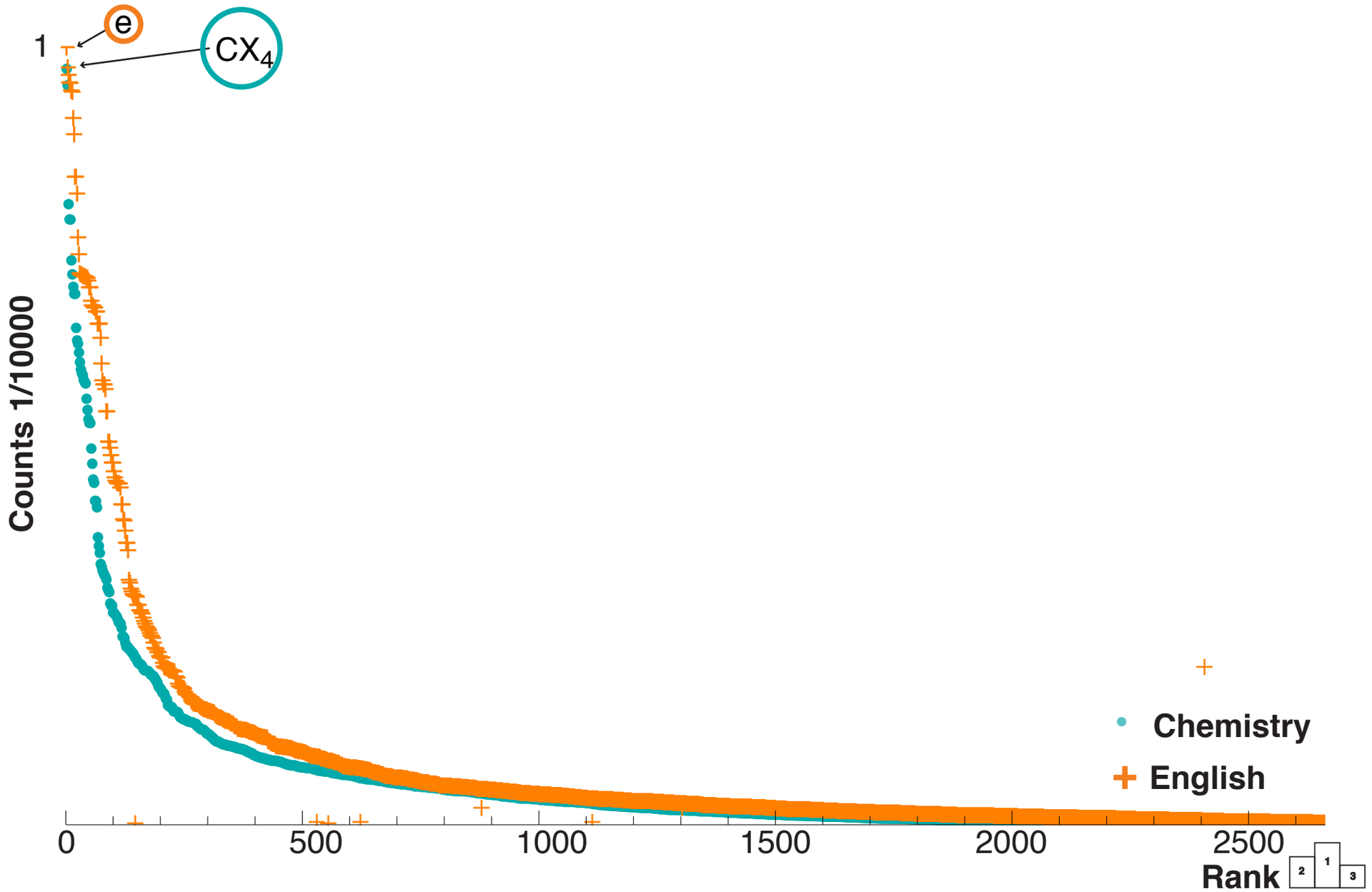
α - Cyclodextrin

=



MCS

Maximum common substructure


e
CX_4
1
Counts 1/10000
0
500
1000
1500
2000
2500
Rank
Chemistry
English

Chemistry behave as a language.

what can we do?

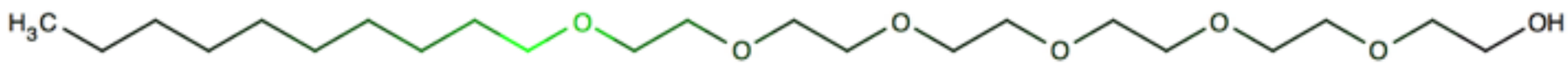
Weight words based on their frequency

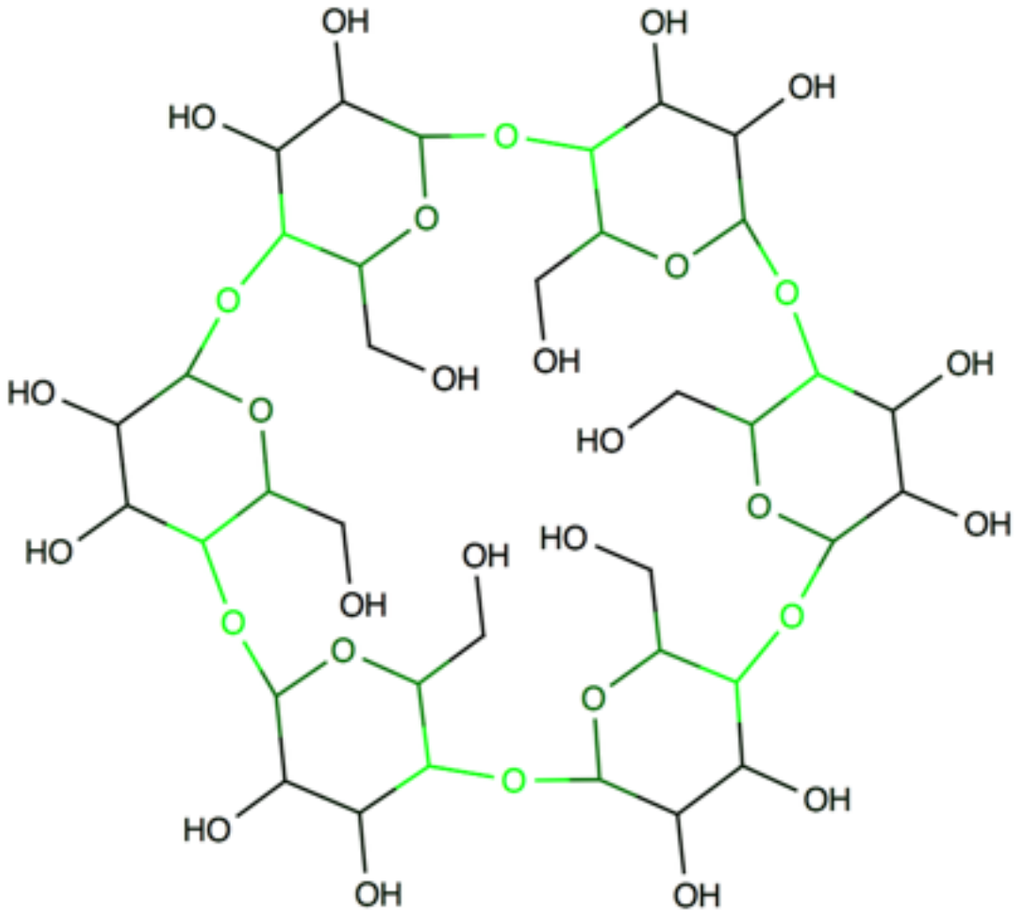
Let’s take a document. (Molecule)

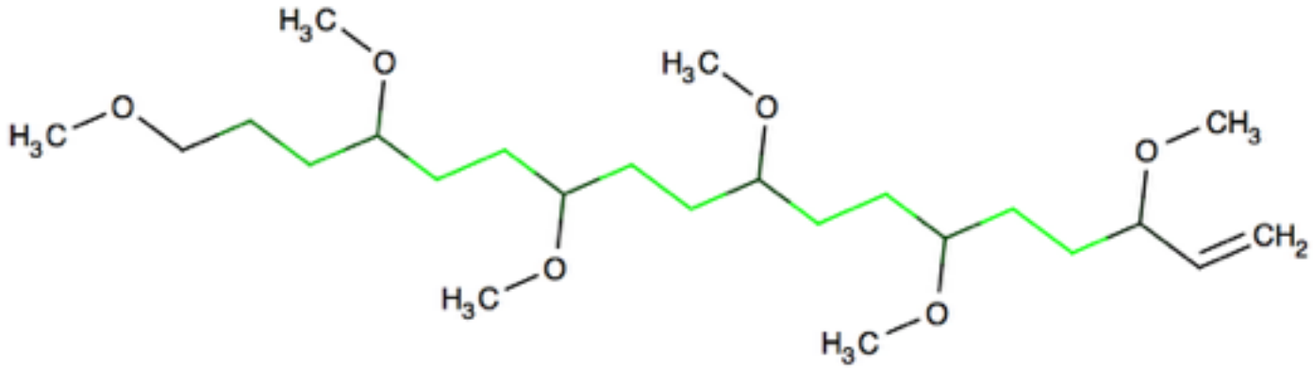
Let’s find of which words is composed.

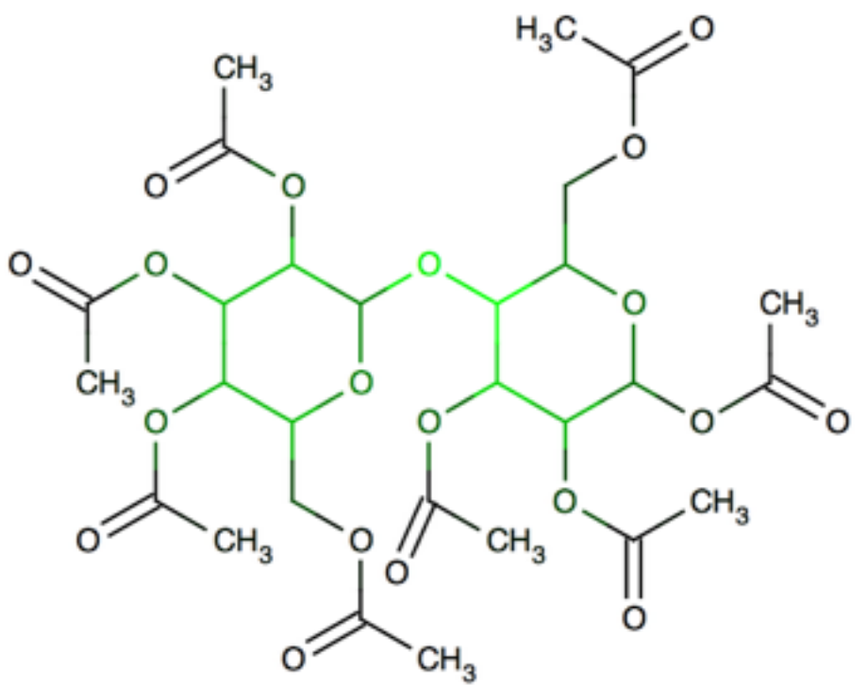
Let’s weight them based on the frequency*.

Bag-Of-Words, *Tf-Idf scores





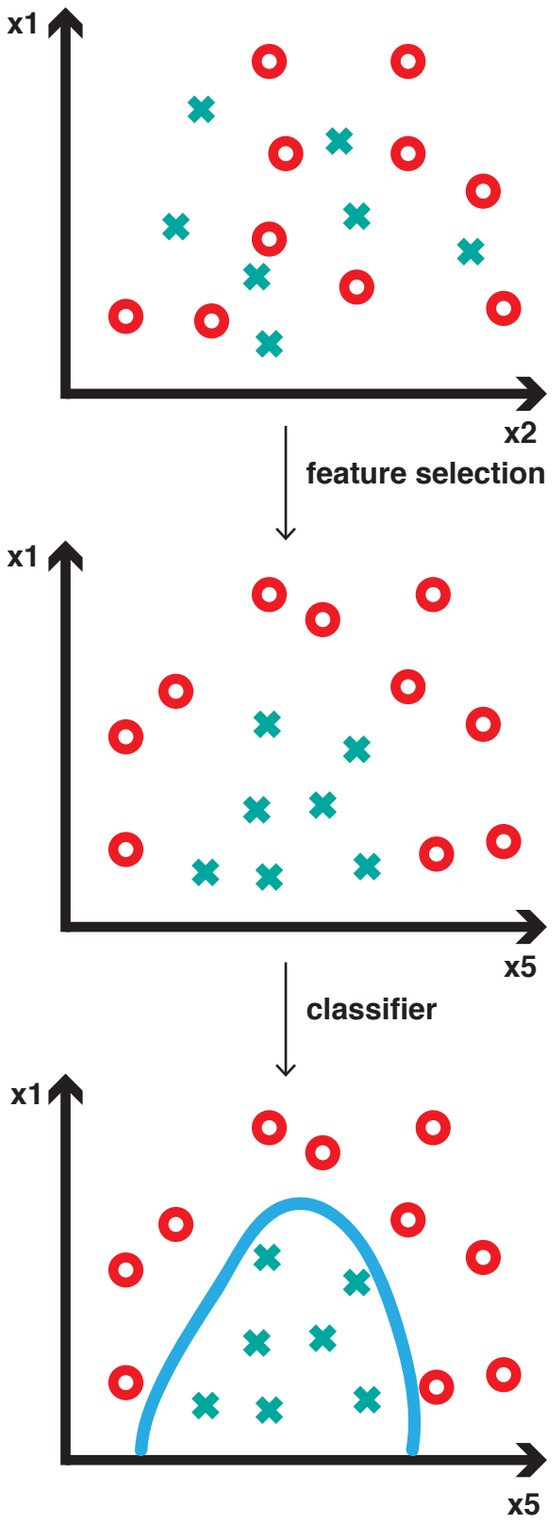




Chemistry behave as a language.

what can we do?

binary classification

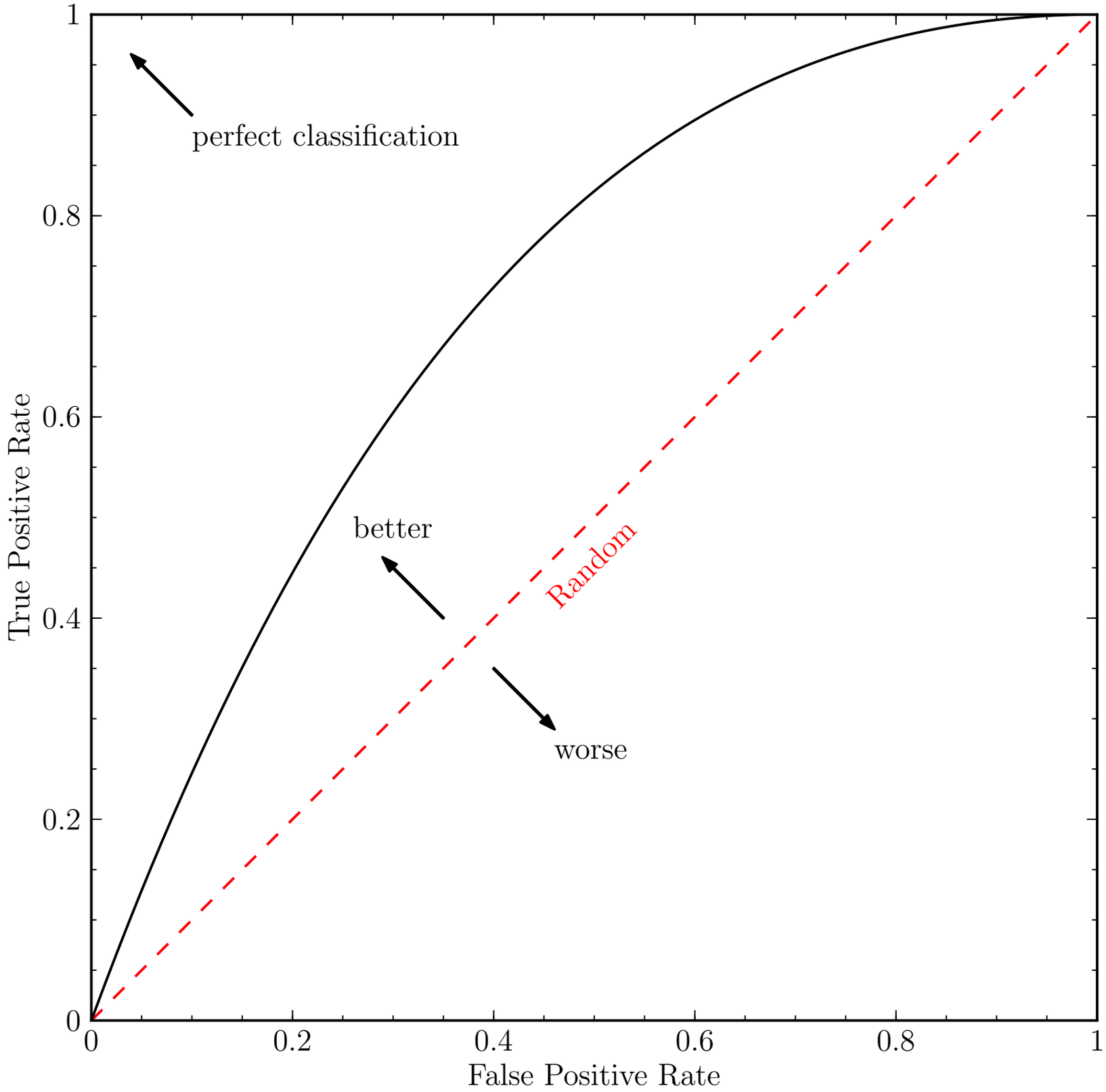


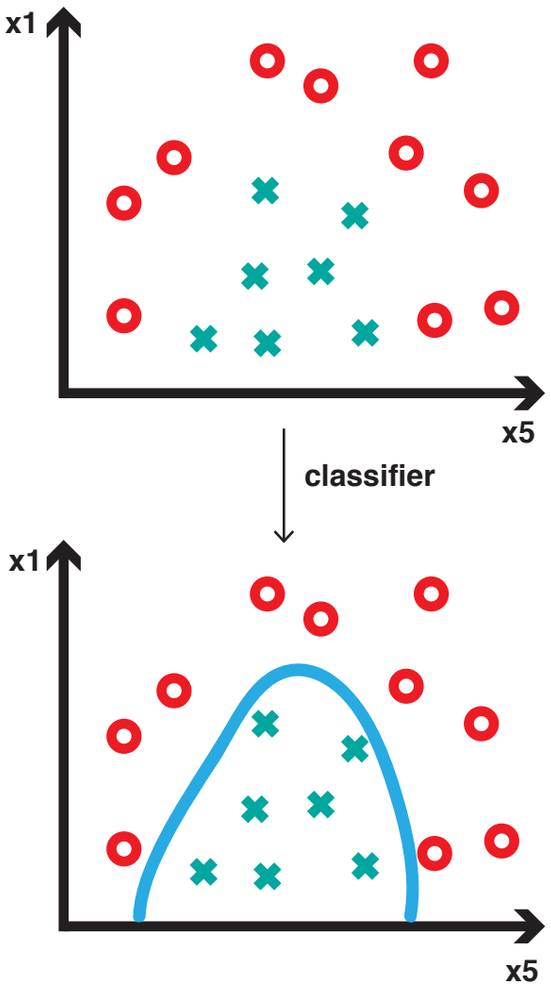
	x_1	x_2	x_3	x_4	x_5	y
a	1	3	3	21	21	0
b	5	8	32	21	53	1
c	8	213	20	55	95	0
d	24	3	59	66	73	1

	x_1	x_2	x_3	x_4	x_5	y
a	1	3	3	21	21	0
b	5	8	32	21	53	1
c	8	213	20	55	95	0
d	24	3	59	66	73	1

$$y = 1 \text{ if } \begin{cases} x_1 < C_5 x_5^2 \\ x_1 > 0 \end{cases}$$

Receiver Operating Characteristic





Let’s take a reaction.

Let’s find which words compose the reactants.

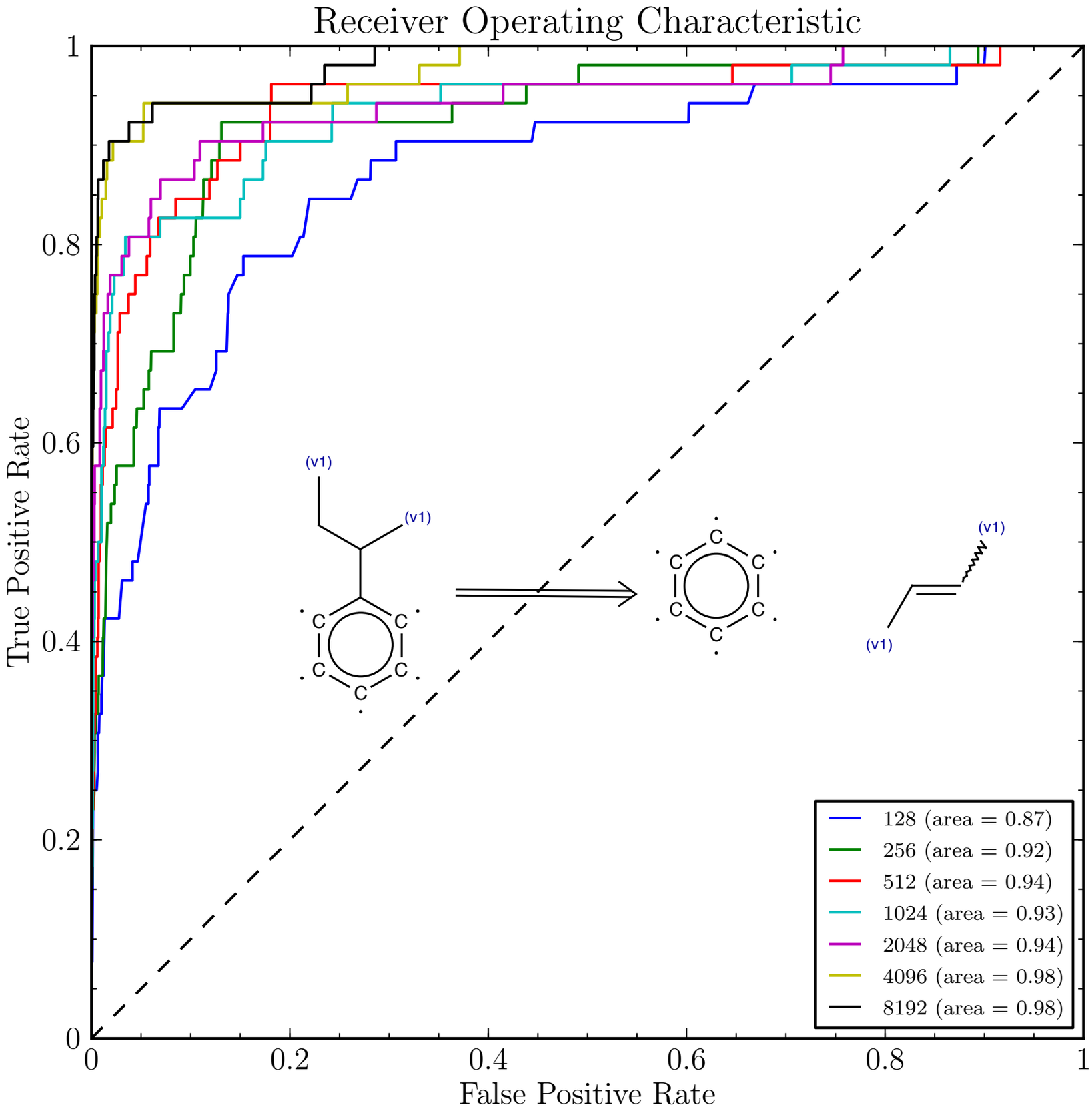
Let’s see which words compose the products.

Let’s see how this change among many* reactions.

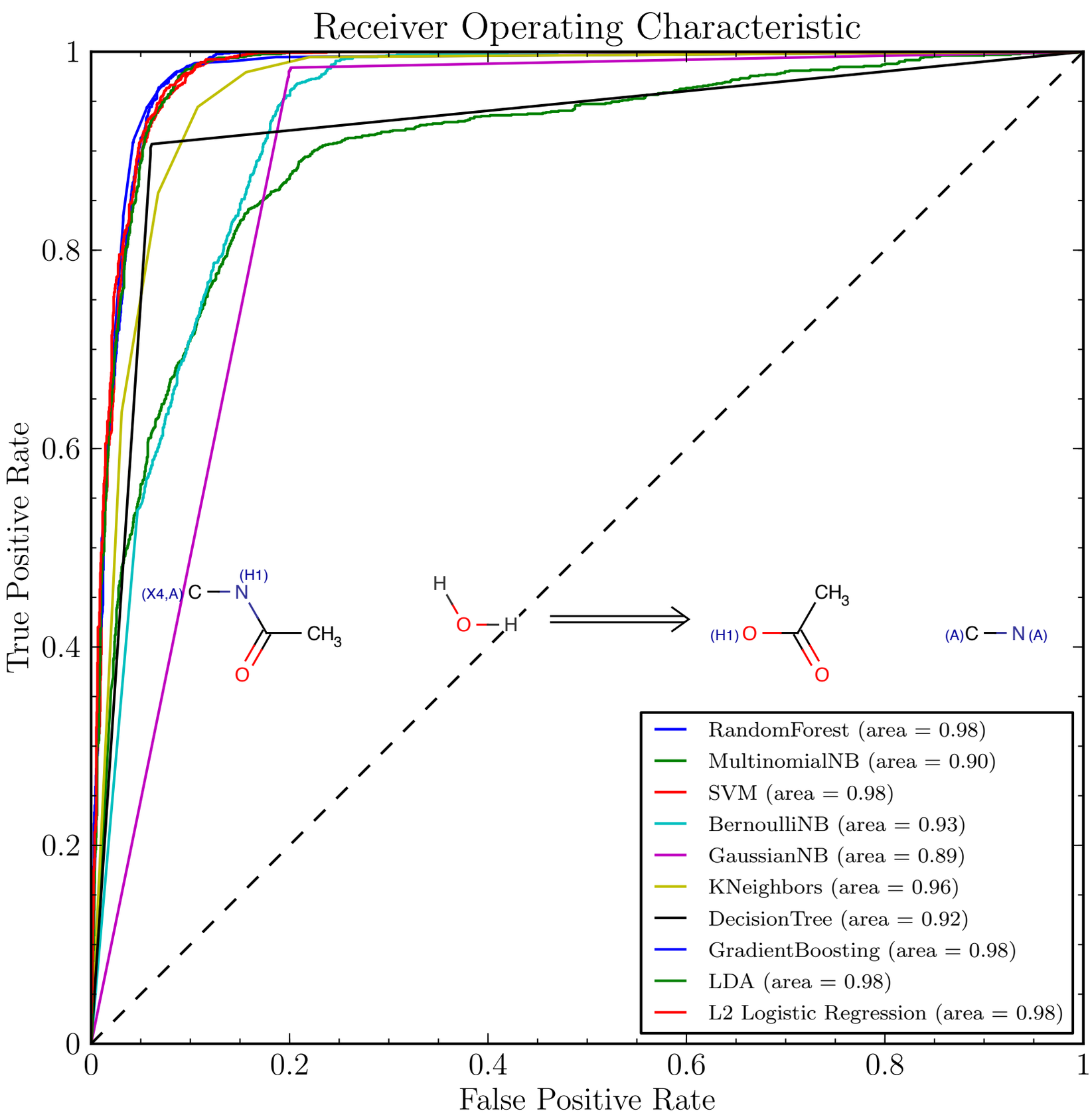
Let’s train the classifier.

*10K

Electrophilic Aromatic Substitution



Amide coupling



Chemistry behave as a language

We can use it to find good disconnections

We can say if a retrosynthesis is possible with Classifiers

Chemistry behave as a language.

So we can use Natural Language Processing.

Acknowledgements

- Bartosz Grzybowski
- Matthew Wampler-Doty
- The Chematica Team

Tf-Idf scores

$$tf(t,d) = 0.5 + \frac{05 \times f(t,d)}{max(f(w,d) : w \in d)}$$

$$idf(t,D) = log \frac{|D|}{|d \in D : t \in d|}$$

Zipf law dependence on number of sentences