#ODSC

ODSC West, San Francisco
November 3, 2017

PSA Lab

MICHIGAN STATE UNIVERSITY

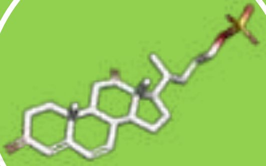
Building Hypothesis-driven Virtual Screening Pipelines for Millions of Molecules

Sebastian Raschka

sebastianraschka.com

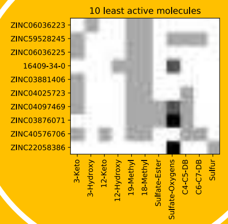
@rasbt

Domain Knowledge



Tools

Statistical Analyses & Machine Learning



By the beginning of the twentieth century, the Great Lakes were the richest freshwater fishery in the world […]
But those good years were soon gone.

Dennis, Jerry. *The Living Great Lakes: Searching for the Heart of the Inland Seas*. Macmillan, 2003.

https://en.wikipedia.org/wiki/Great_Lakes#/media/File:Great_Lakes_from_space_crop_labeled.jpg

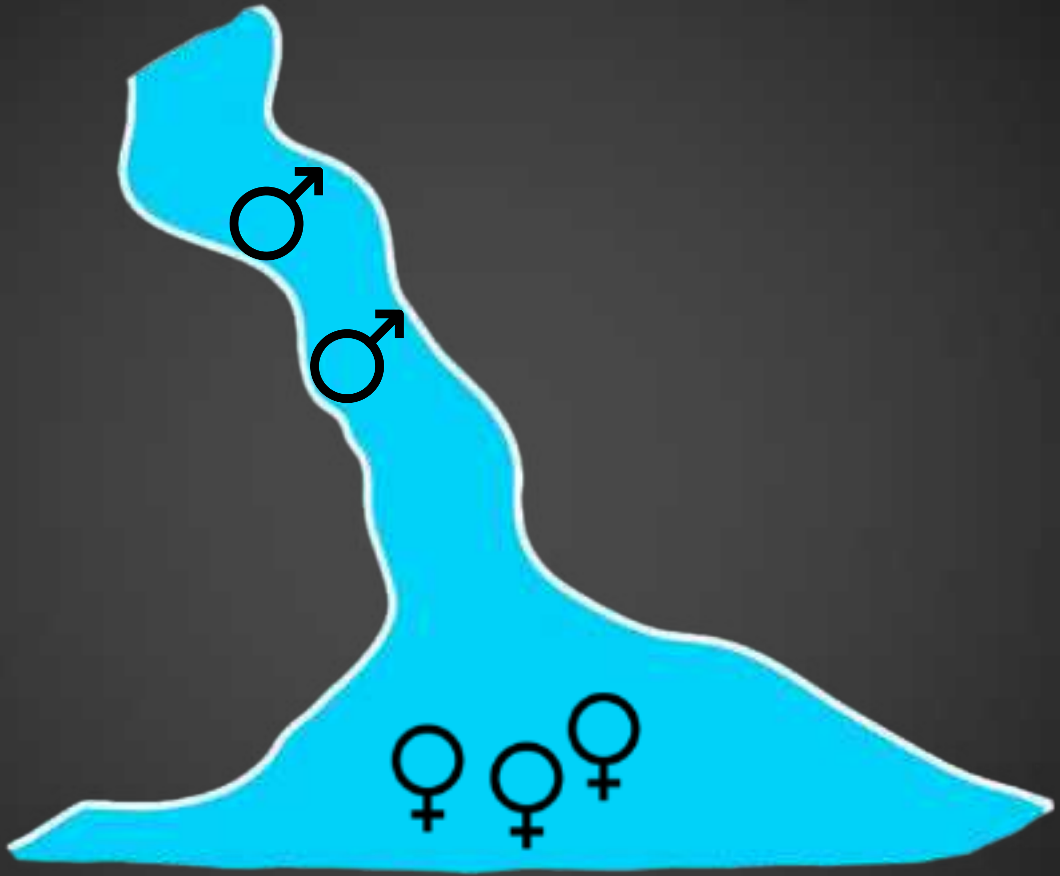
The opening of the Welland Canal […] allowed ships from all over the world to come to the upper lakes […] But nobody could have foreseen that the canal would also allow entry to a most unwelcome visitor, the sea lamprey.

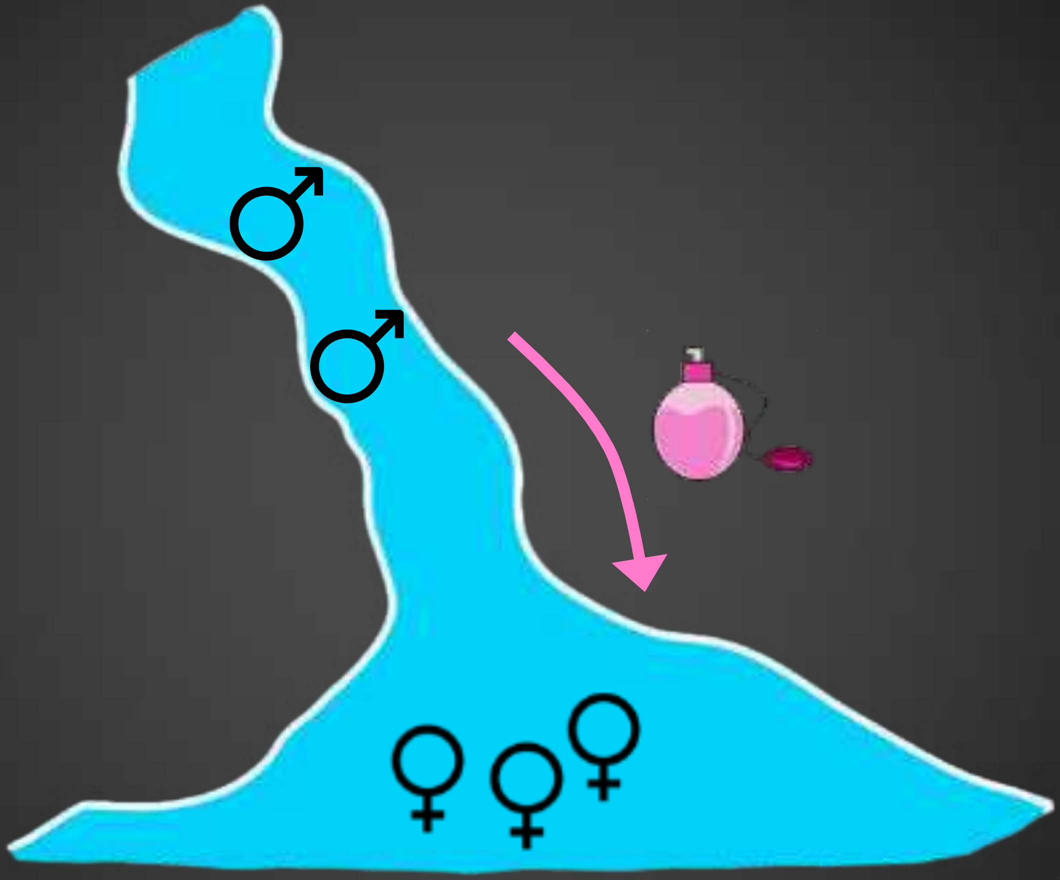
Dennis, Jerry. *The Living Great Lakes: Searching for the Heart of the Inland Seas*. Macmillan, 2003.

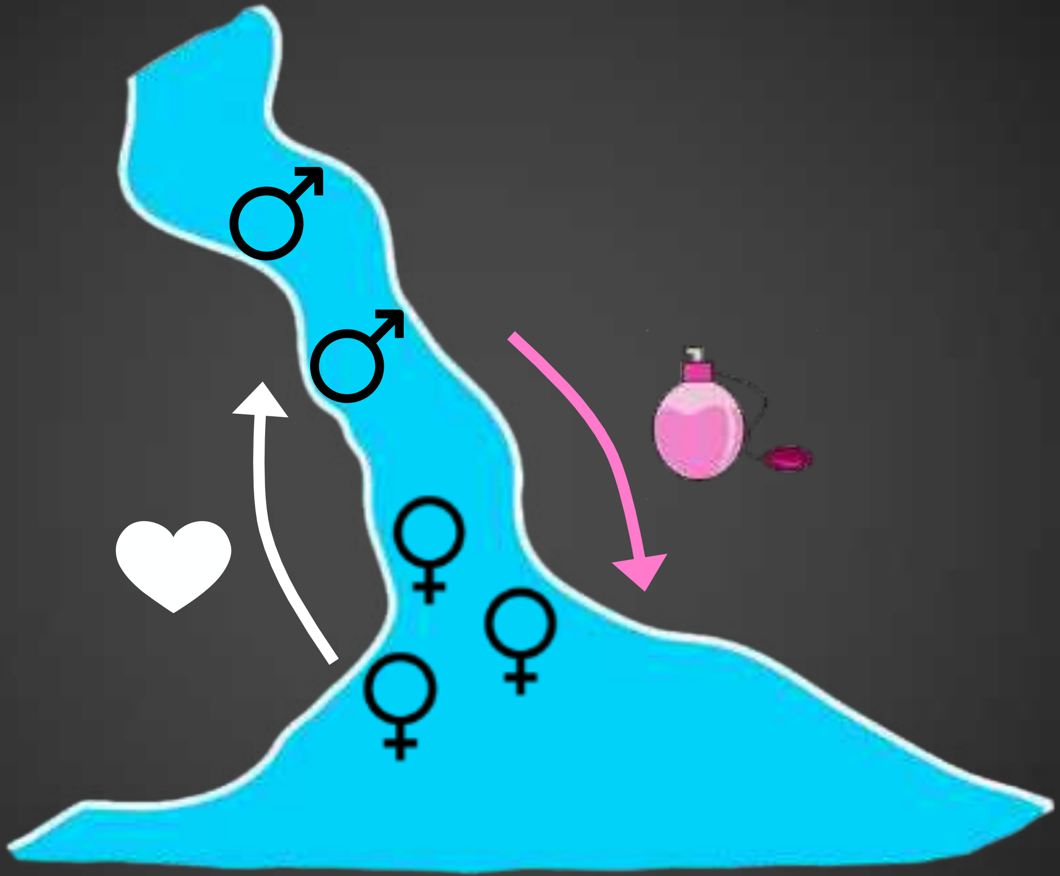
https://en.wikipedia.org/wiki/Welland_Canal#/media/File:Welland_Canal_aerial.png

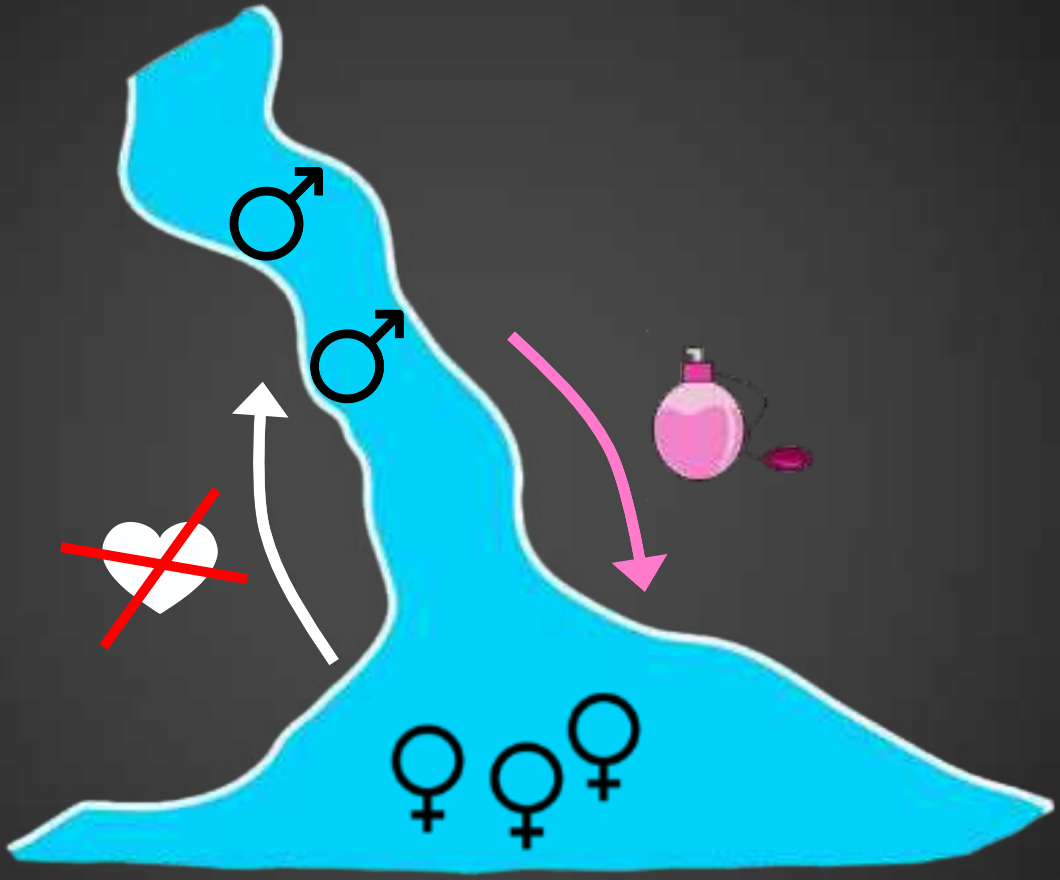

https://en.wikipedia.org/wiki/Sea_lamprey#/media/File:Sea_lamprey_on_brown_trout_flipped.jpg

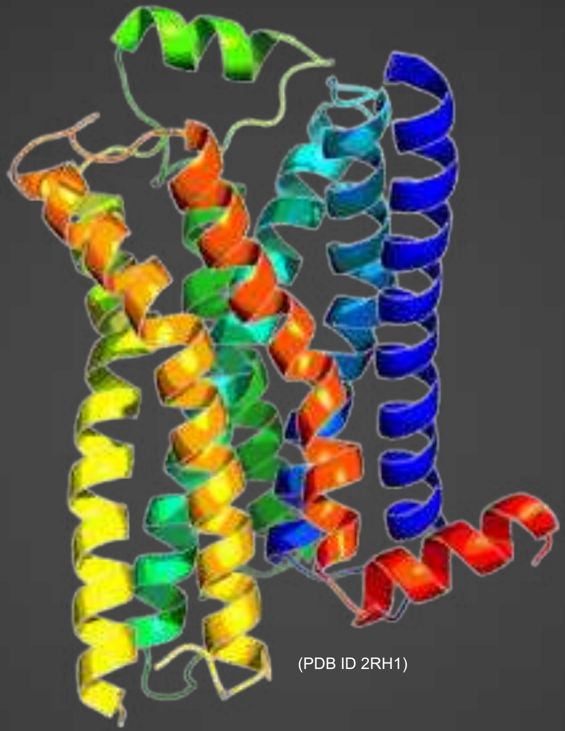
Invasive Species Control





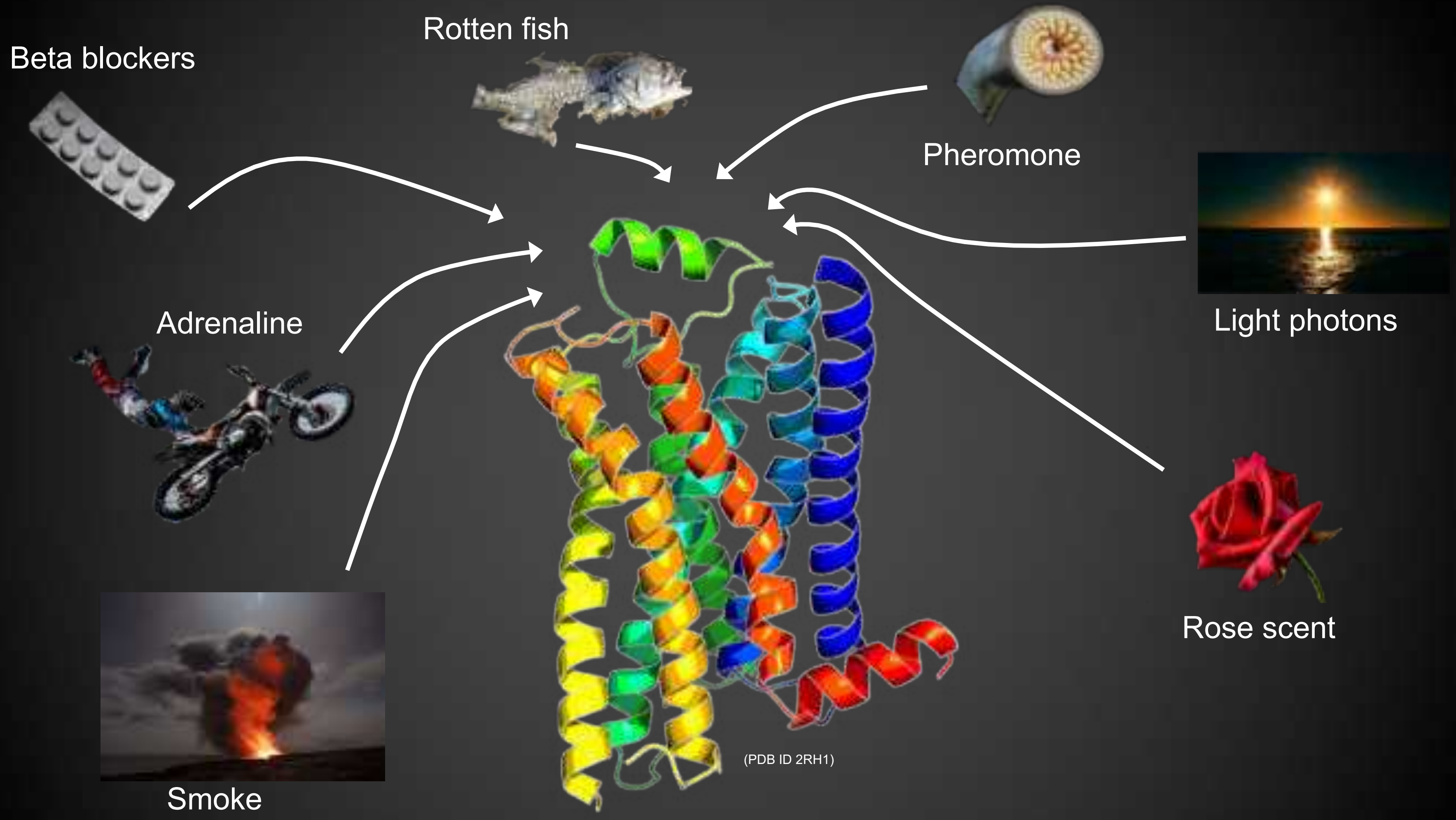






(PDB ID 2RH1)

G protein-coupled receptors: sense diverse chemicals

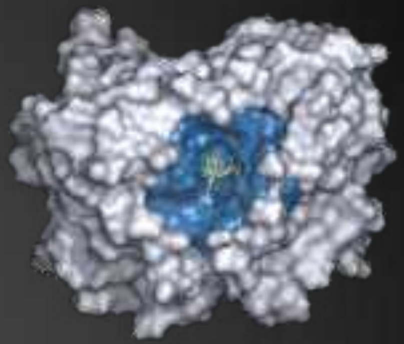

Rotten fish
Beta blockers
Pheromone
Light photons
Adrenaline
Rose scent
Smoke
(PDB ID 2RH1)

Millions of Molecules

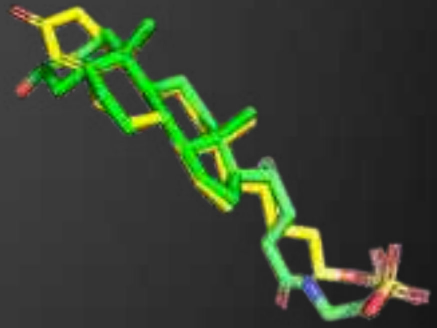
Millions of Molecules

Virtual Screening

Receptor Structure-based

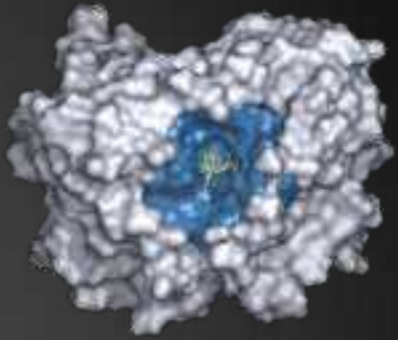


Small Molecule-based

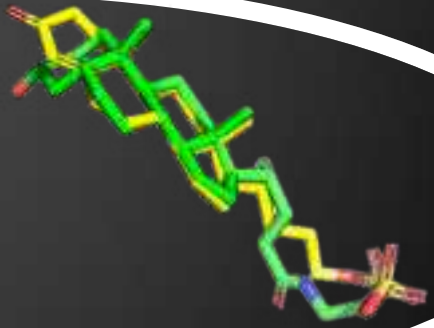


Virtual Screening

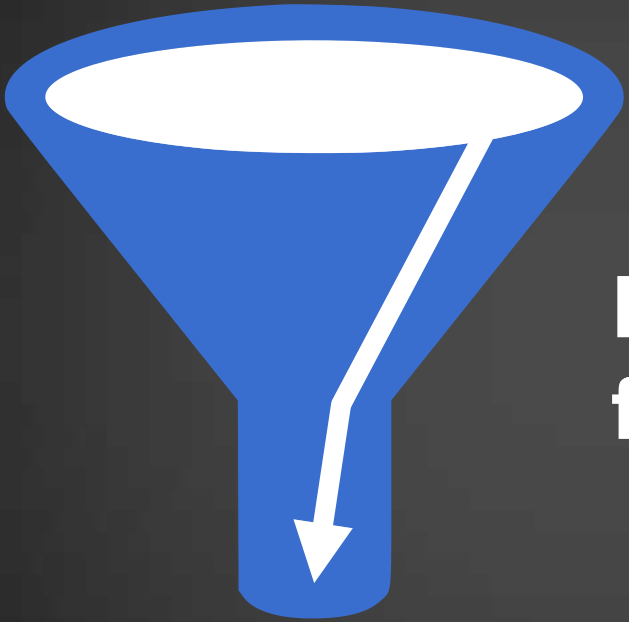
Receptor Structure-based



Small Molecule-based

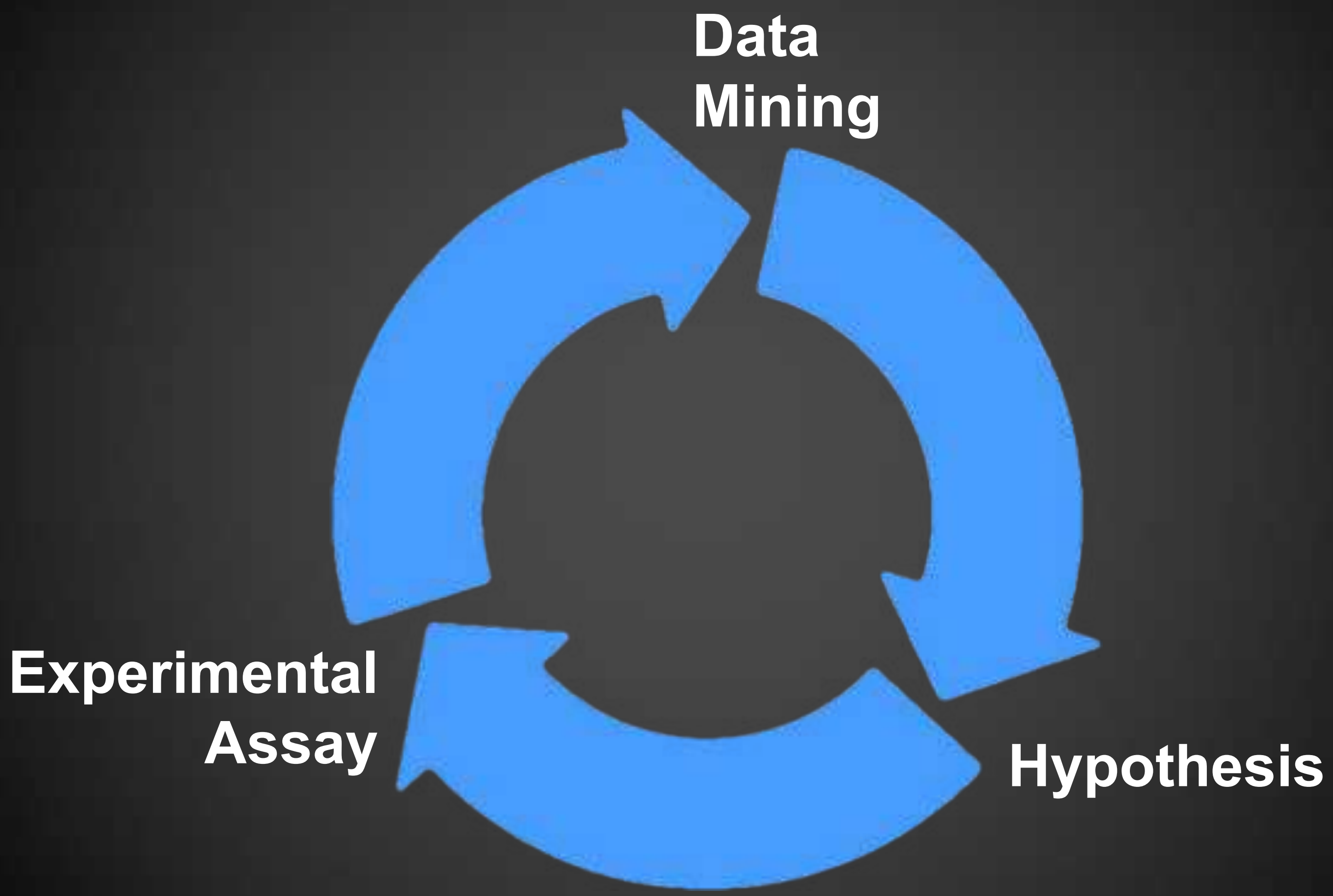


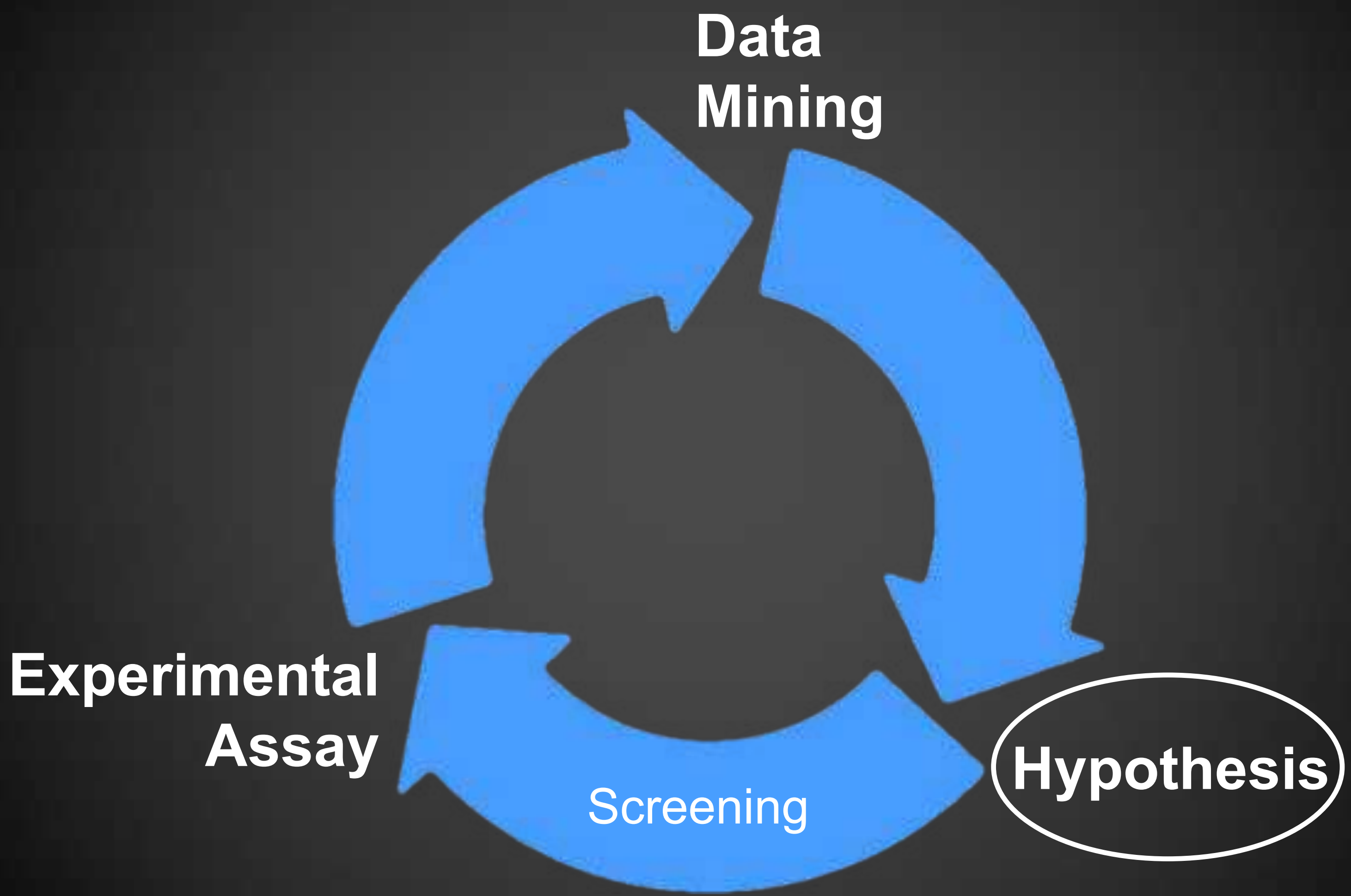
Millions of molecules

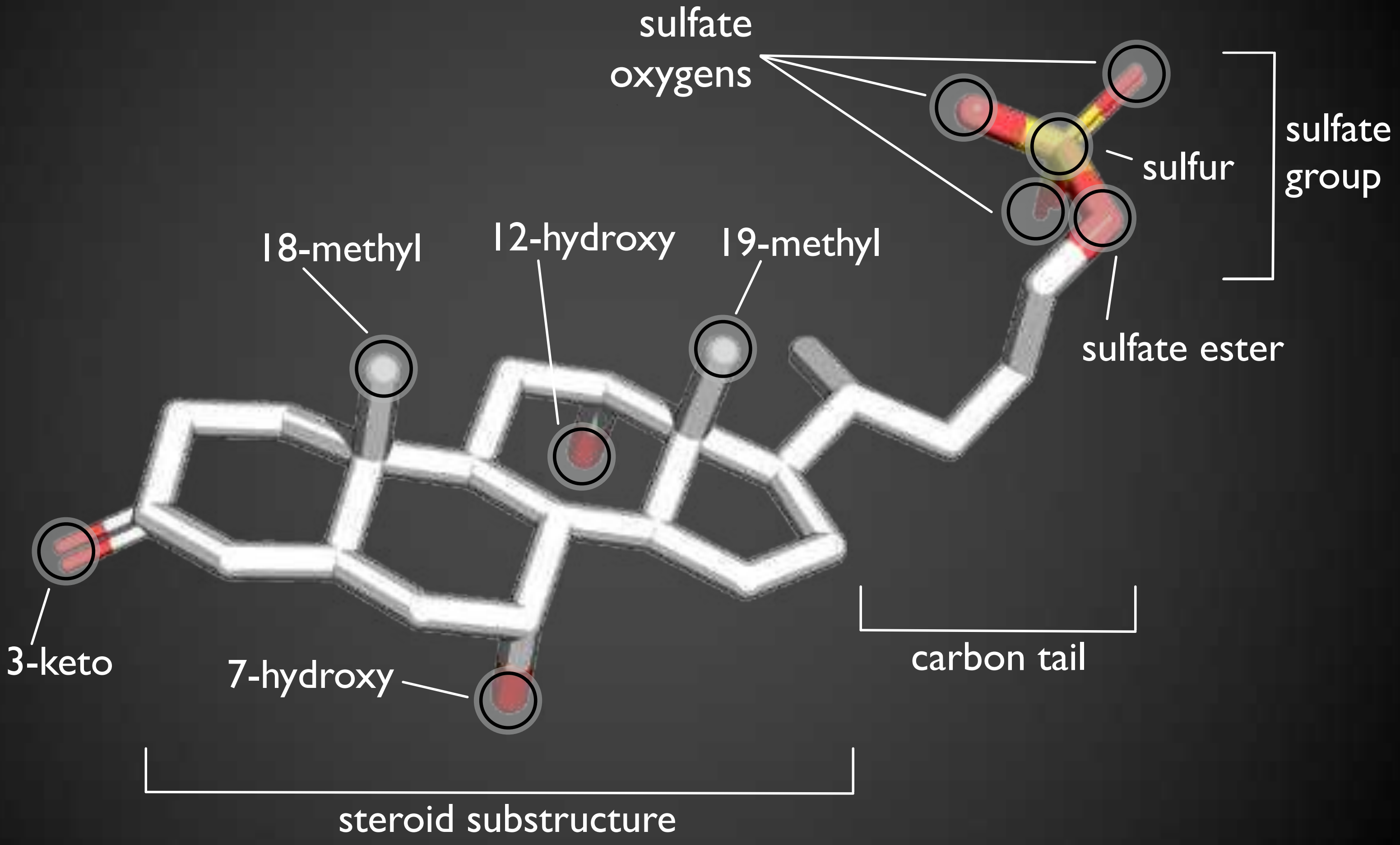


Hundreds of molecules

Hypothesis-based filtering


Data
Mining
Hypothesis
Experimental
Assay


Data
Mining
Experimental
Assay
Screening
Hypothesis

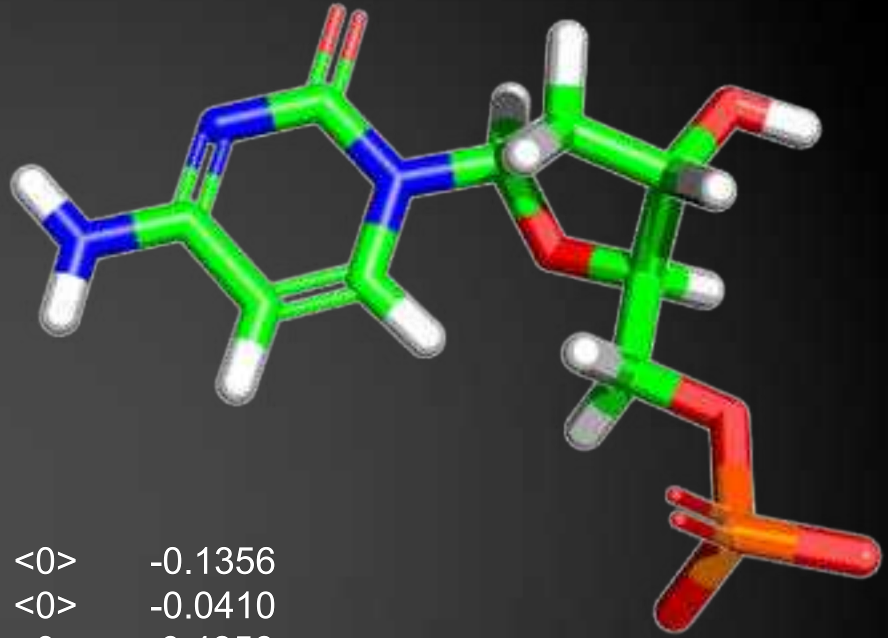

sulfate oxygens
sulfur
sulfate group
sulfate ester
18-methyl
12-hydroxy
19-methyl
3-keto
7-hydroxy
carbon tail
steroid substructure

Tabular Data

BioPandas

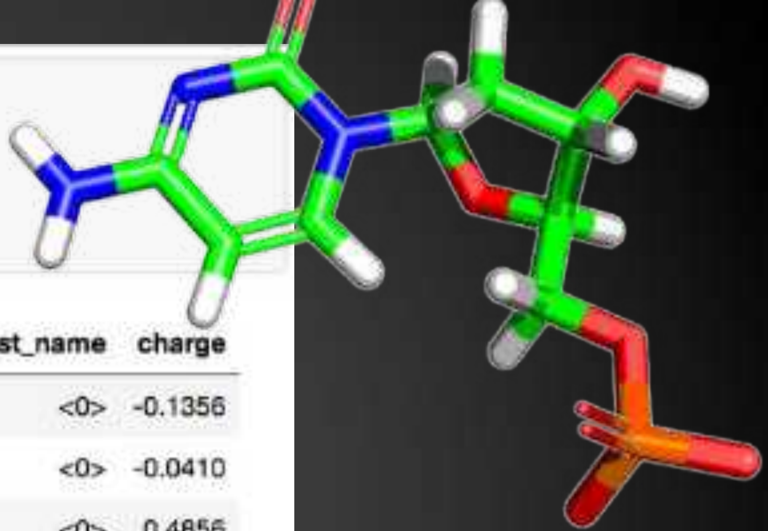
https://rasbt.github.io/biopandas/

Raschka S (2017) BioPandas: Working with molecular structures in pandas DataFrames. J Open Source Softw 2:1–3.



```
@<TRIPOS>MOLECULE
DCM Pose 1
   32    33     0     0     0
SMALL
USER_CHARGES

@<TRIPOS>ATOM
      1 C1          18.8934    5.5819   24.1747 C.2       1 <0>       -0.1356
      2 C2          18.1301    4.7642   24.8969 C.2       1 <0>       -0.0410
      3 C3          18.2645    6.8544   23.7342 C.2       1 <0>        0.4856
      4 C4          16.2520    6.2866   24.7933 C.2       1 <0>        0.8410
      5 C5          15.3820    3.0682   25.1622 C.3       1 <0>        0.0000
…
```

```python
from biopandas.mol2 import PandasMol2

pmol = PandasMol2()
pmol.read_mol2('./molecule.mol2')
pmol.df.head(10)
```



	atom_id	atom_name	x	y	z	atom_type	subst_id	subst_name	charge
0	1	C1	18.8934	5.5819	24.1747	C.2	1	<0>	-0.1356
1	2	C2	18.1301	4.7642	24.8969	C.2	1	<0>	-0.0410
2	3	C3	18.2645	6.8544	23.7342	C.2	1	<0>	0.4856
3	4	C4	16.2520	6.2866	24.7933	C.2	1	<0>	0.8410
4	5	C5	15.3820	3.0682	25.1622	C.3	1	<0>	0.0000
5	6	C6	15.4162	1.8505	26.0566	C.3	1	<0>	0.2800
6	7	C7	16.7283	2.0138	26.8111	C.3	1	<0>	0.2800
7	8	C8	16.0764	4.1199	26.0119	C.3	1	<0>	0.5801
8	9	C9	17.9106	1.3823	26.0876	C.3	1	<0>	0.2800
9	10	N1	17.0289	7.1510	24.0411	N.2	1	<0>	-0.6610

Software

Python explosion blamed on pandas

Data science fad just won't die

By Thomas Claburn in San Francisco 14 Sep 2017 at 20:02 33 SHARE ▼

Not content to bait developers by declaring that Python is the fastest-growing major programming language, coding community site Stack Overflow has revealed the reason for its metastasis.

https://www.theregister.co.uk/2017/09/14/python_explosion_blamed_on_pandas/

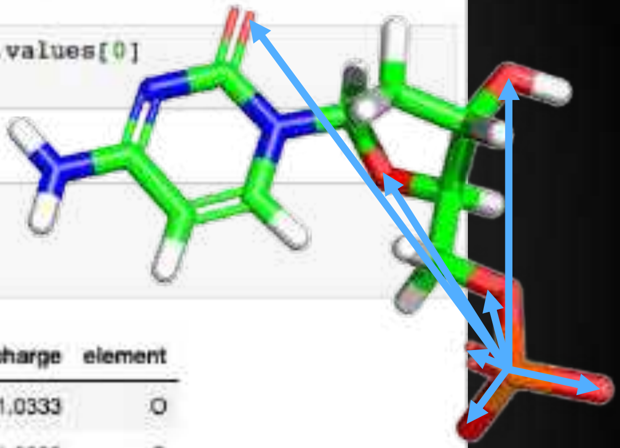
In [2]:
```
pmol.df[pmol.df['atom_type'] == 'P.3']
```

Out[2]:

	atom_id	atom_name	x	y	z	atom_type	subst_id	subst_name	charge
19	20	P1	19.0969	-0.944	25.6653	P.3	1	<0>	1.3712

In [3]:
```
xyz = pmol.df.loc[pmol.df['atom_type'] == 'P.3', ['x', 'y', 'z']].values[0]
xyz
```

Out[3]:
```
array([ 19.0969,  -0.944 ,  25.6653])
```



In [4]:
```
pmol.df['element'] = pmol.df['atom_type'].apply(lambda x: x[0])
df_oxygen = pmol.df[pmol.df['element'] == 'O'].copy()
df_oxygen
```

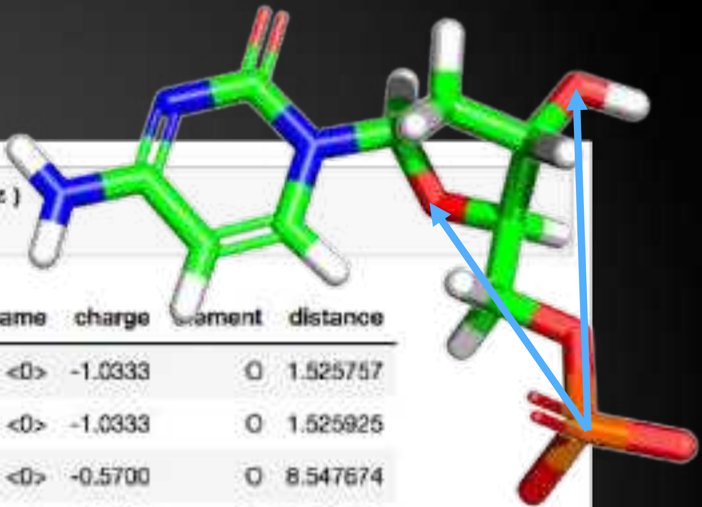
Out[4]:

	atom_id	atom_name	x	y	z	atom_type	subst_id	subst_name	charge	element
12	13	O1	18.7676	-2.3524	26.1510	O.3	1	<0>	-1.0333	O
13	14	O2	20.3972	-0.3812	26.2318	O.3	1	<0>	-1.0333	O
14	15	O3	15.0888	6.5824	25.0727	O.2	1	<0>	-0.5700	O
15	16	O4	18.9314	-0.7527	24.1606	O.2	1	<0>	-1.0333	O
16	17	O5	16.9690	3.4315	26.8994	O.3	1	<0>	-0.5600	O
17	18	O6	14.3223	1.8946	26.9702	O.3	1	<0>	-0.6800	O
18	19	O7	17.9091	-0.0135	26.3390	O.3	1	<0>	-0.5512	O

In [5]:
```python
df_oxygen['distance'] = PandasMol2.distance_df(df_oxygen, xyz)
df_oxygen
```

Out[5]:

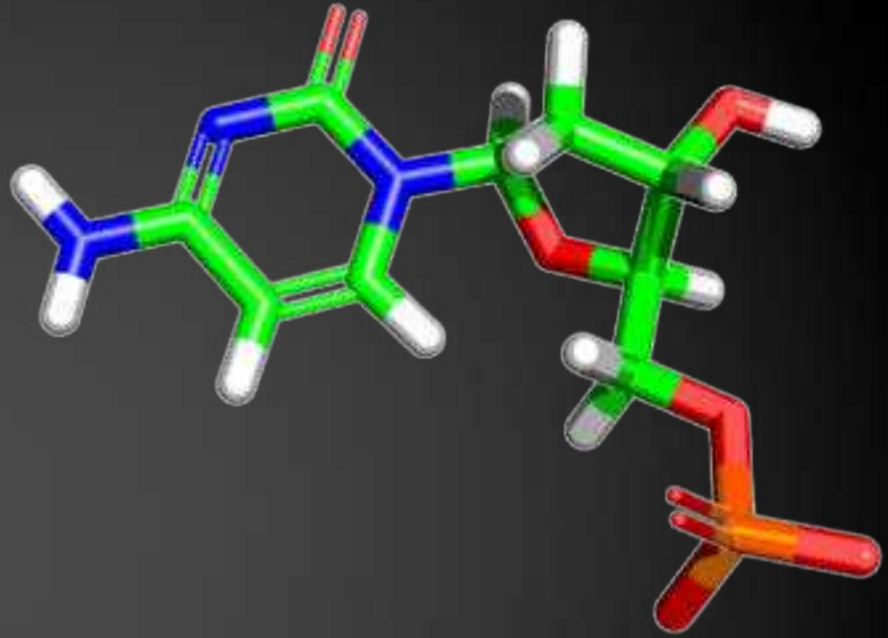
	atom_id	atom_name	x	y	z	atom_type	subst_id	subst_name	charge	element	distance
12	13	O1	18.7676	-2.3524	26.1510	O.3	1	<0>	-1.0333	O	1.525757
13	14	O2	20.3972	-0.3812	26.2318	O.3	1	<0>	-1.0333	O	1.525925
14	15	O3	15.0688	6.5824	25.0727	O.2	1	<0>	-0.5700	O	8.547674
15	16	O4	18.9314	-0.7527	24.1606	O.2	1	<0>	-1.0333	O	1.525814
16	17	O5	16.9690	3.4315	26.8994	O.3	1	<0>	-0.5600	O	5.019558
17	18	O6	14.3223	1.8946	26.9702	O.3	1	<0>	-0.6800	O	5.705893
18	19	O7	17.9091	-0.0135	26.3390	O.3	1	<0>	-0.5612	O	1.652444



In [6]:
```python
df_oxygen[(df_oxygen['distance'] > 3) & (df_oxygen['distance'] < 8)].shape[0]
```

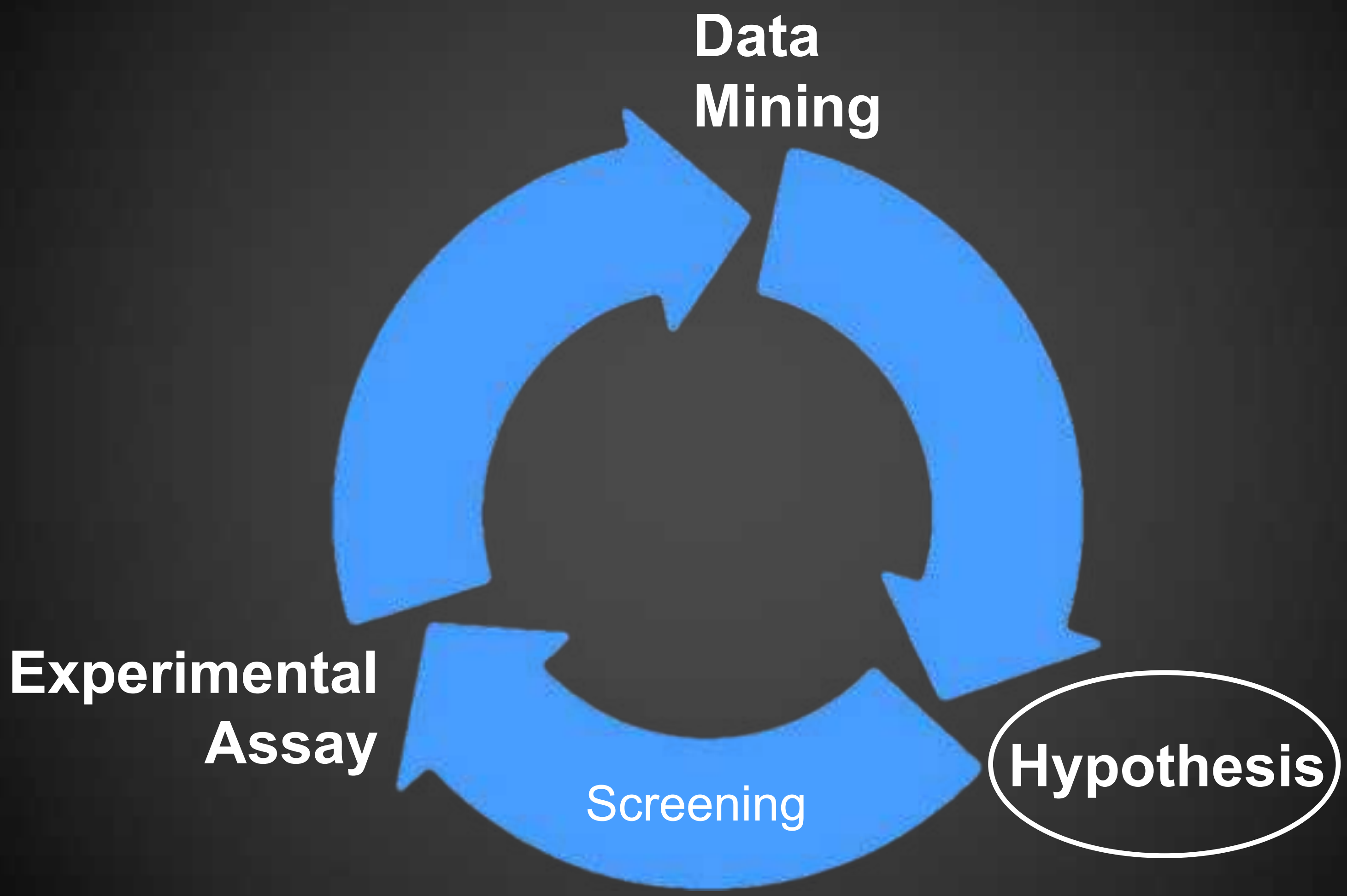
Out[6]: 2

More examples
Multi-mol2 files
Multi-processing
…



BioPandas

http://rasbt.github.io/biopandas/tutorials/Working_with_MOL2_Structures_in_DataFrames

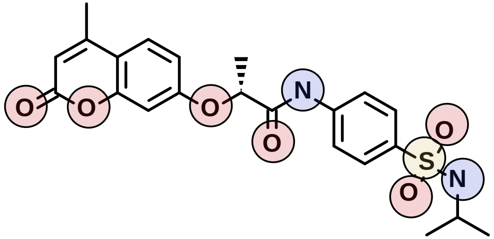

Data
Mining
Experimental
Assay
Screening
Hypothesis

screenlamp

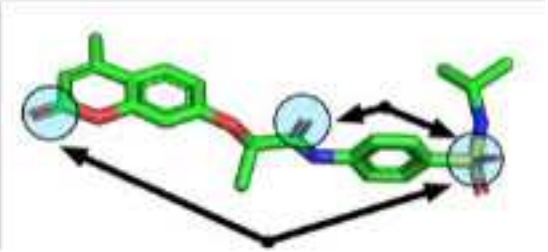
Hypothesis-based Filtering

ID	Weight			Purchasable
12	423 g/mol			✓
363	423 g/mol			✓

General Properties

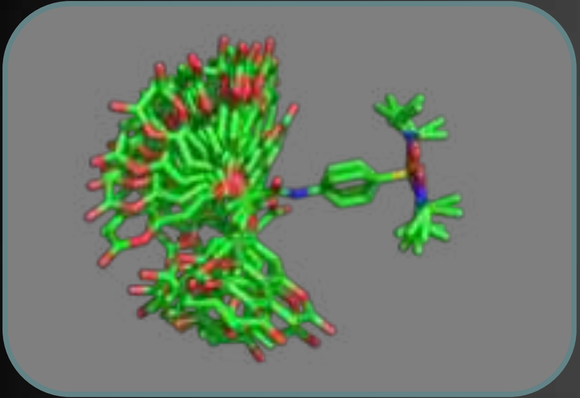


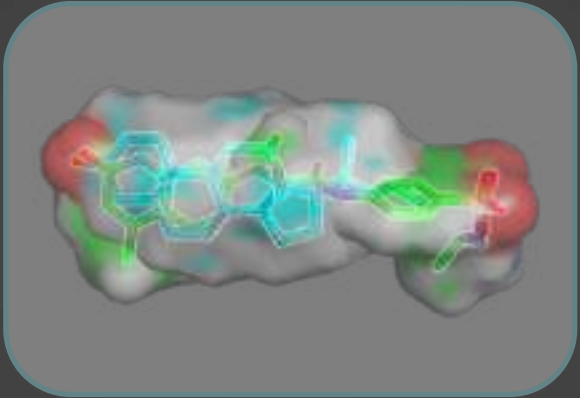
Atom Type Counts

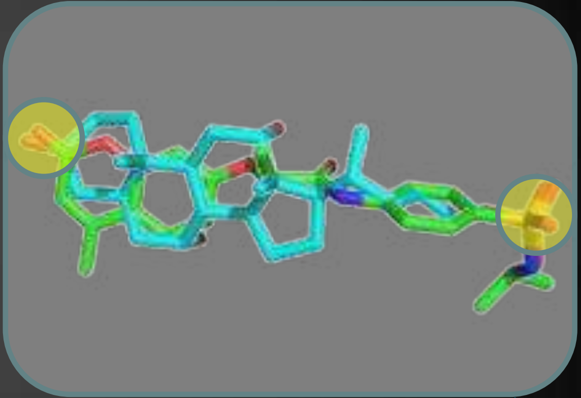


Functional Group Distances

Conformer Overlays and Pharmacophore Matching





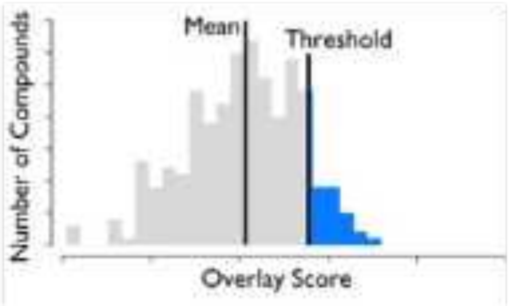


Conformer Sampling

Volumetric & Chemical Overlays

Functional Group Matching

Selection for Experimental Assays



ID	3-keto			12-hydroxy
12	✓			✓
363	✓			✗

Domain Knowledge
Chemical Scaffold
Purity
Price
PAINS
Docking Scores

Overall Similarity Thresholds

Functional Group Matching Patterns

Additional Selection Criteria

pipeline-example-1-config.yaml

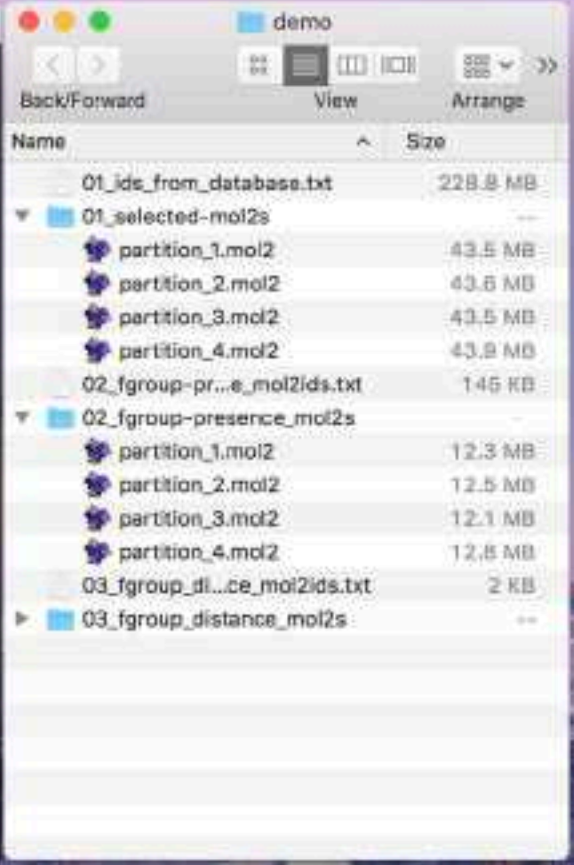
```yaml
general settings:
  screenlamp tools directory: /Users/sebastian/code/screenlamp/tools
  project output directory: /Users/sebastian/code/screenlamp/example-files/example_1/screening-results
  input mol2 directory: /Users/sebastian/code/screenlamp/example-files/example_1/dataset/mol2
  number of cpus: 0 # 0 means all avaible CPUs (recommended)
```

...

```yaml
###############################################################
### Step 03: PREFILTER BY FUNCTIONAL GROUP DISTANCE
###############################################################
functional group distance filter settings:
  # the following selection criteria select all molecules that
  # have an sp2-hybridized sulfur atom (MOL2 atom type S.3 or S.o2)
  # and a keto group (MOL2 atom type O.2), and where the distance between
  # the sulfur and oxygen atoms is between 13 and 20 angstrom
  selection key: ((atom_type == 'S.3') | (atom_type == 'S.o2')) --> (atom_type == 'O.2')
  distance: 13-20
```

...

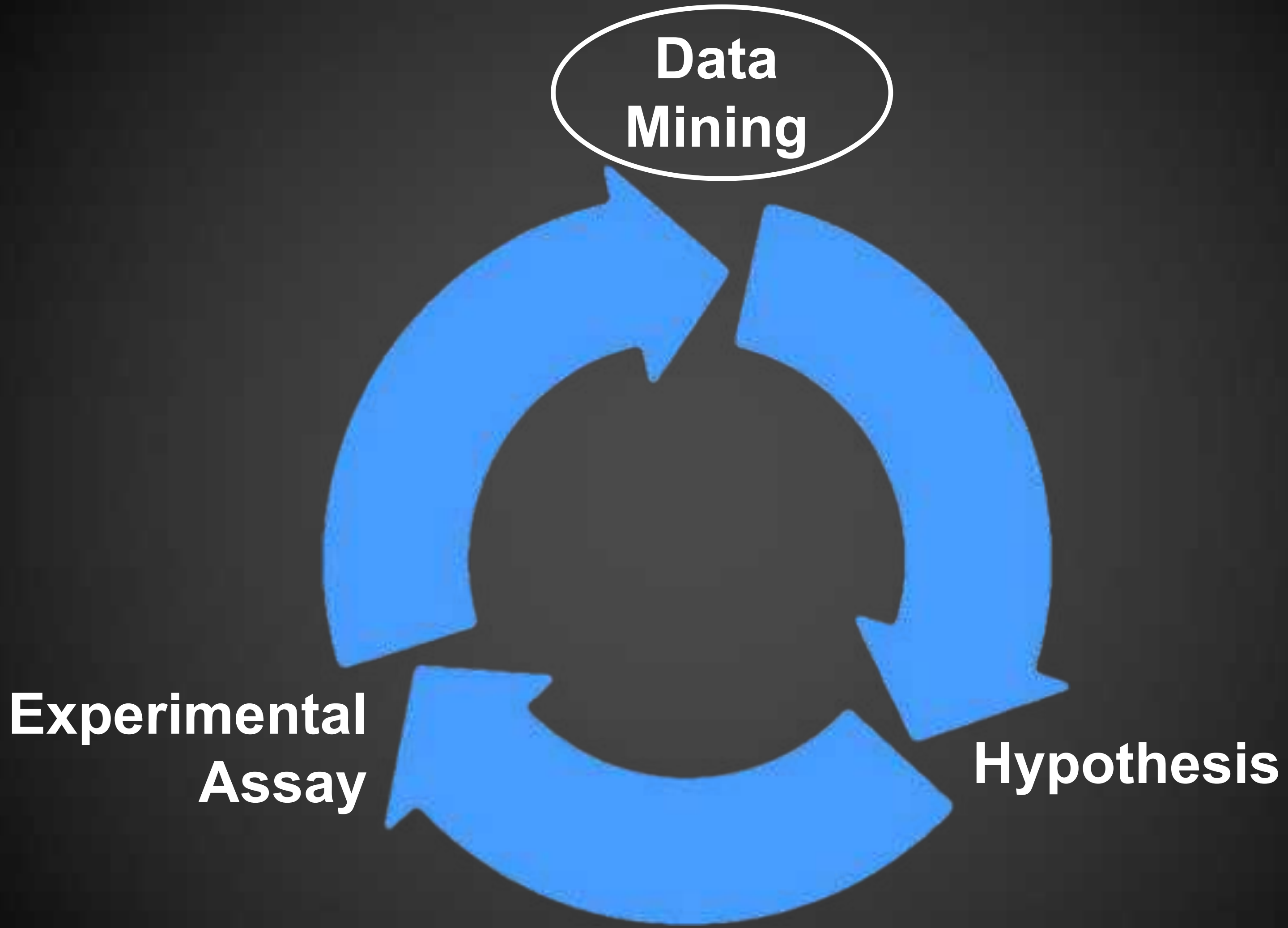
```
Processing partition_3.mol2 | scanned 9848 molecules | 18623 mol/sec
Processing partition_4.mol2 | scanned 9835 molecules | 18812 mol/sec
Finished


SELECTED MOL2s:
Running command:
python /Users/sebastian/code/screenlamp/tools/count_mol2.py --input /Users/sebastian/Desk
top/demo/02_fgroup-presence_mol2s

partition_1.mol2 : 2768
partition_2.mol2 : 2795
partition_3.mol2 : 2746
partition_4.mol2 : 2847
Total : 11156


##############################################
Step 03: PREFILTER BY FUNCTIONAL GROUP DISTANCE
##############################################

Using selection: ["((pdmol.df.atom_type == 'S.3') | (pdmol.df.atom_type == 'S.o2'))", "(p
dmol.df.atom_type == 'O.2')"]
Processing partition_1.mol2 | 230 mol/sec
Processing partition_2.mol2 | 220 mol/sec
Processing partition_3.mol2 | 214 mol/sec
Processing partition_4.mol2
```



screenlamp

Enabling the hypothesis-driven prioritization of ligand candidates in big databases: Screenlamp and its application to GPCR inhibitor discovery for invasive species control (2017).
Raschka S., A. M. Scott, N. Liu. S. Gunturu, M. Huertas, W. Li, and L. A. Kuhn
JCAM (manuscript under revision)

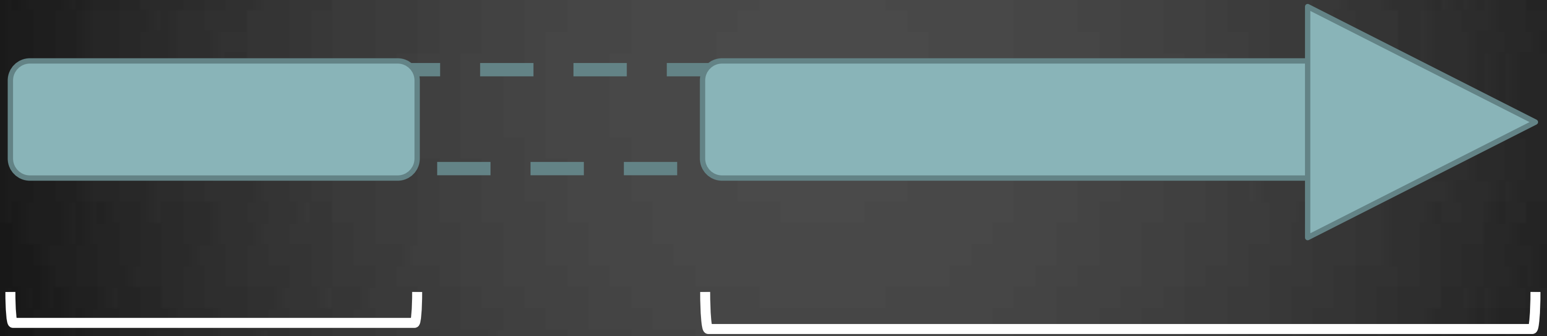
https://psa-lab.github.io/screenlamp


Data
Mining
Experimental
Assay
Hypothesis

```python
df = pd.read_csv('../eog-assay-results.csv')
df.head(10)
```

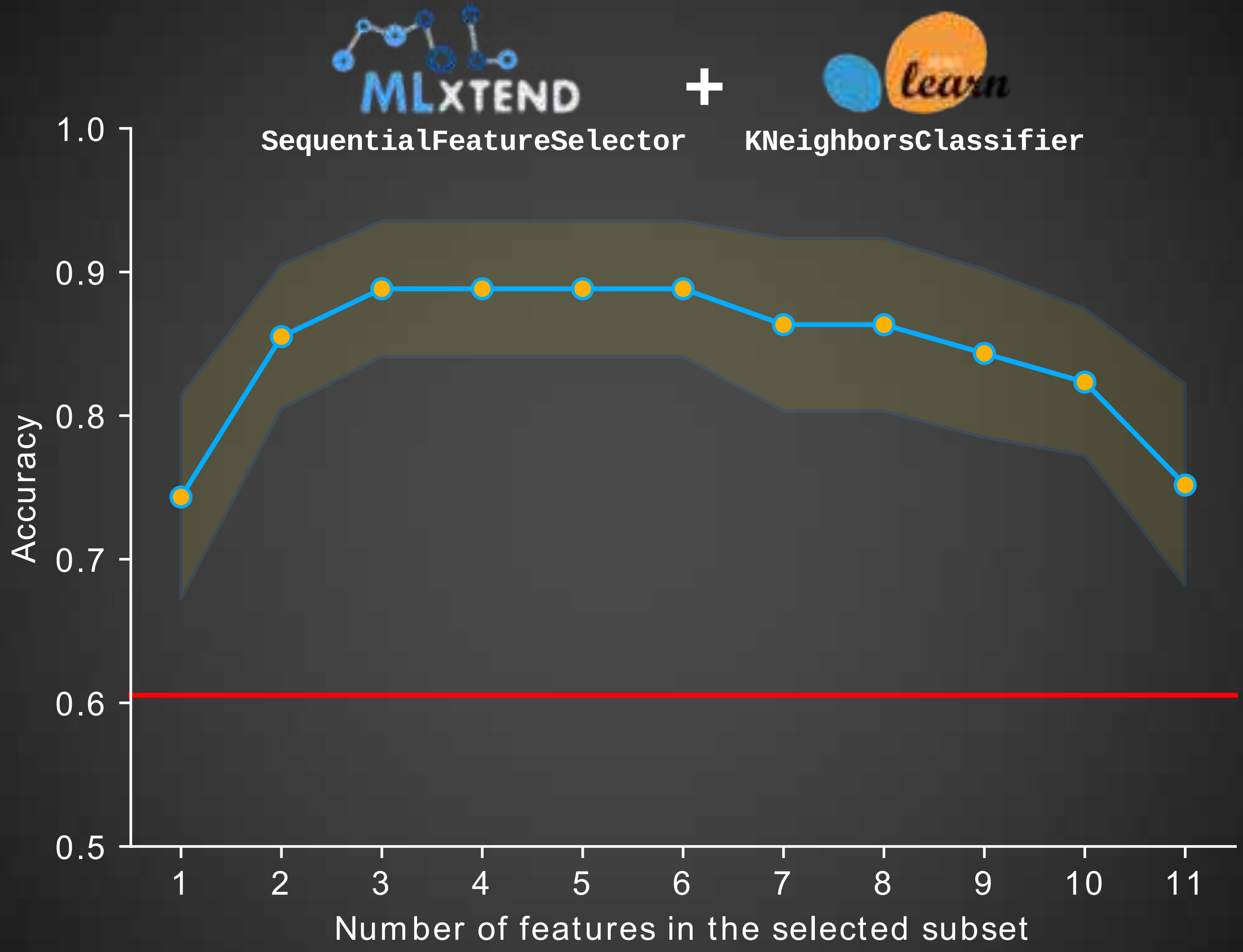
	Molecule ID	Signal-inhibition	3-Keto	3-Hydroxy	12-Keto	12-Hydroxy	19-Methyl	18-Methyl	Sulfate-Ester	Sulfate-Oxygens	...
0	ZINC59528245	0.158	1	0	0	0	1	1	0	1	...
1	ZINC01845398	0.624	0	0	0	0	0	0	0	3	...
2	ZINC01532179	0.686	0	0	0	0	0	0	1	3	...
3	16409-34-0	0.108	0	0	0	1	1	1	0	2	...
4	ENE3	0.897	1	0	0	1	1	1	1	3	...
5	ENE2	0.845	1	0	1	0	1	1	1	3	...
6	ZINC08789094	0.354	1	0	0	0	1	1	0	0	...
7	6785-62-2	0.297	1	0	0	0	0	1	0	2	...
8	ZINC03876071	0.032	0	0	0	0	1	1	1	3	...
9	ZINC71770953	0.483	0	0	0	0	0	0	0	0	...

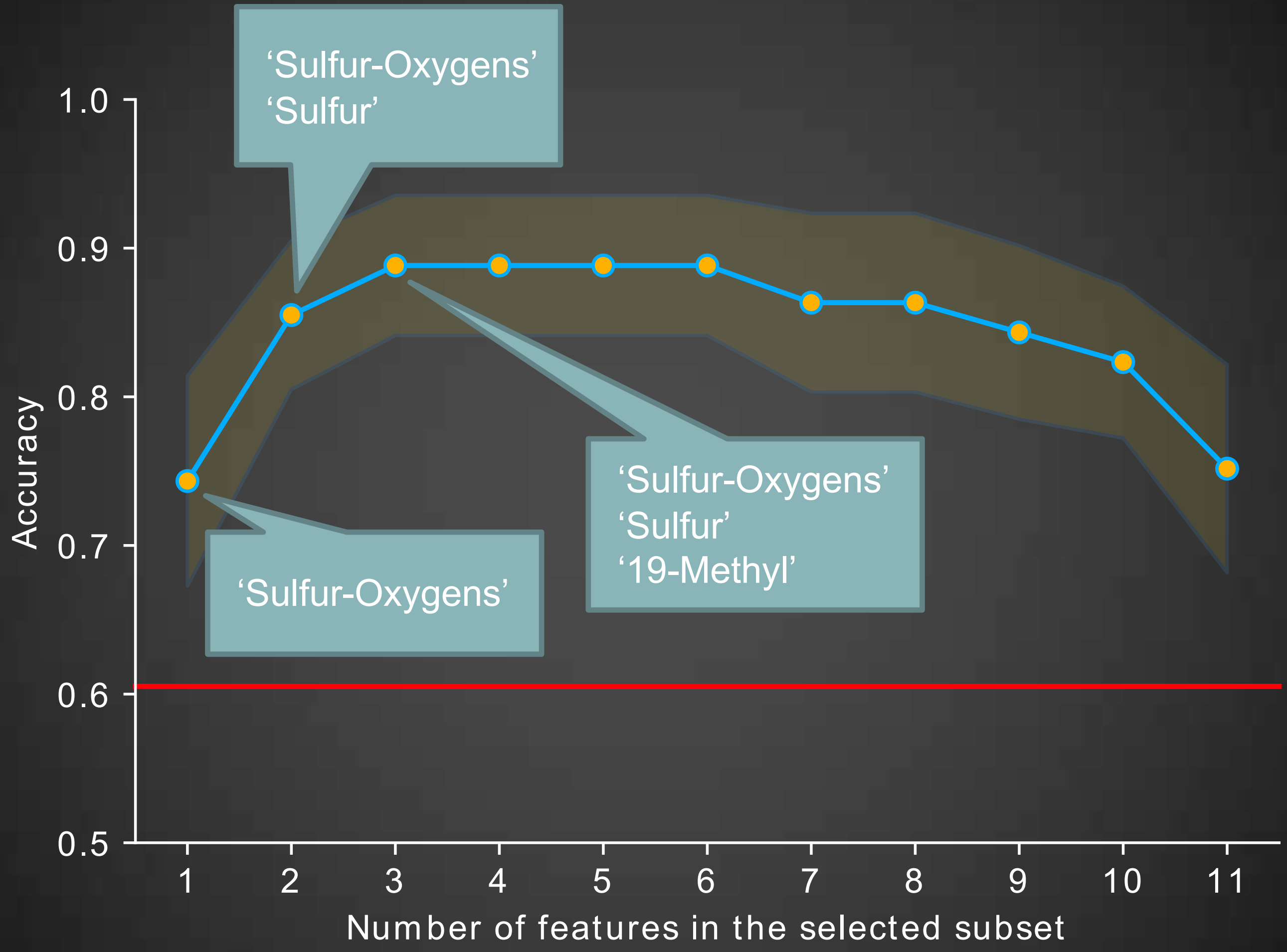
Thresholding Assay Data

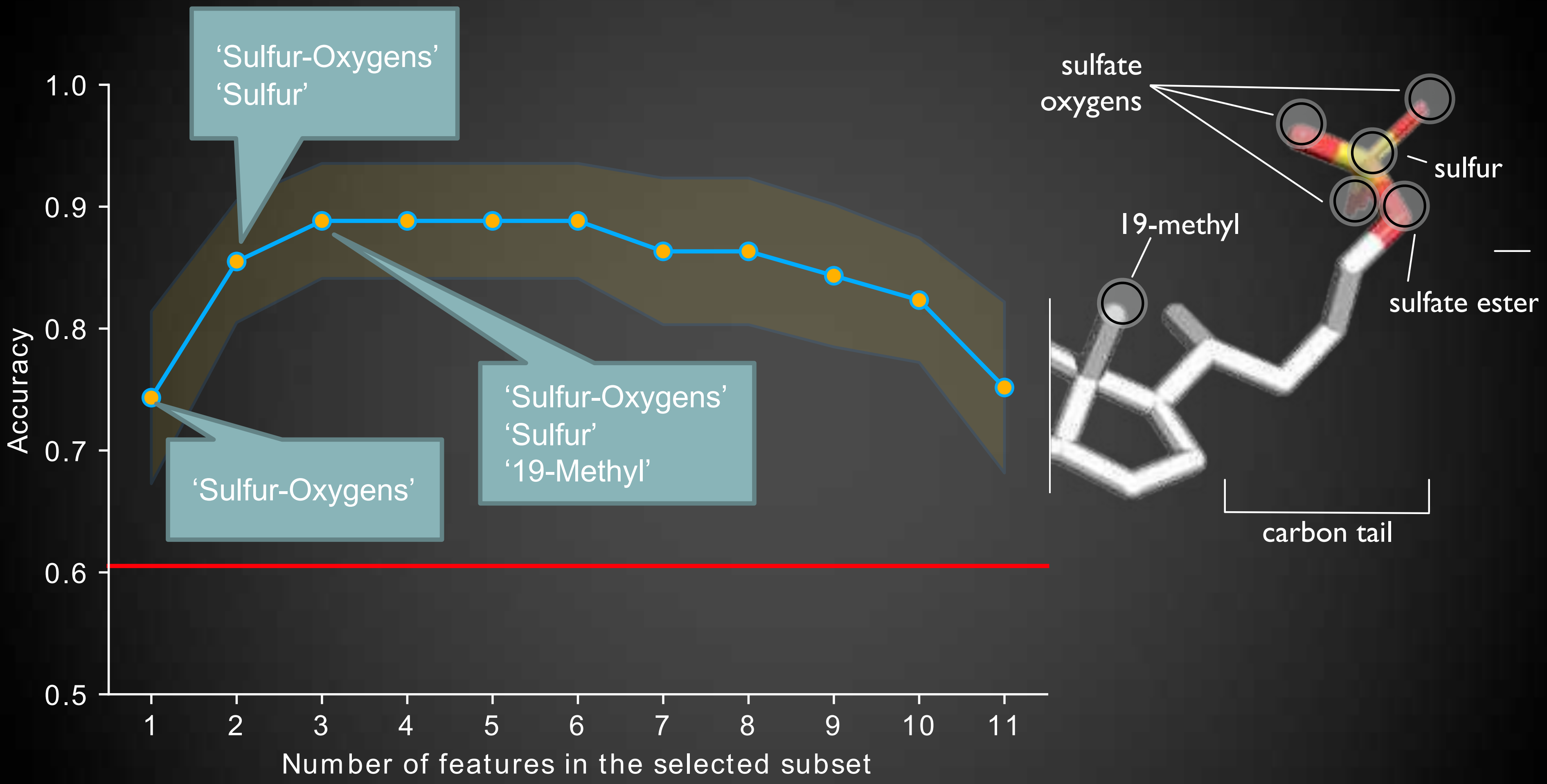
0%  100%

0-25% signal inhibition
= non-active

50-100% signal inhibition
= active

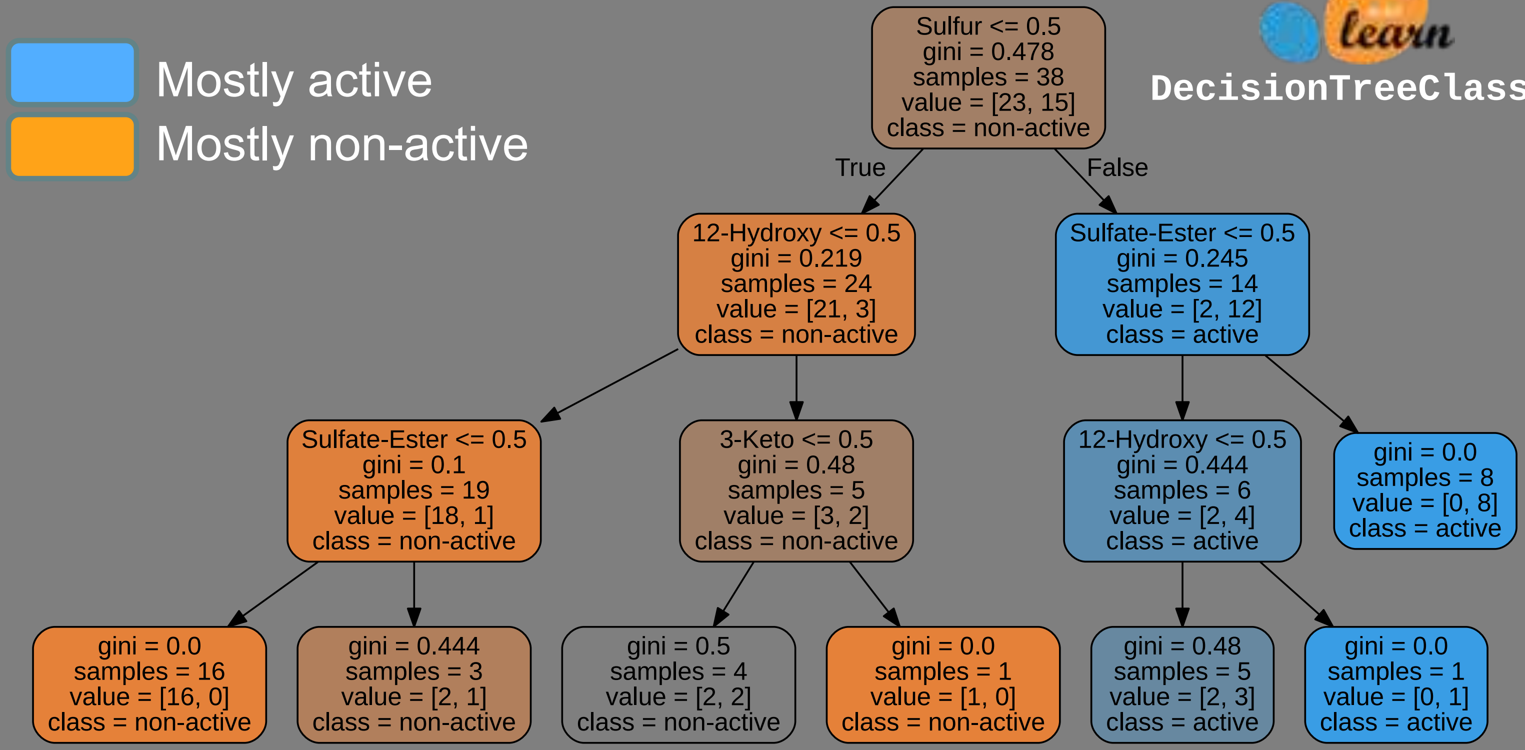

MLXTEND
+
learn
SequentialFeatureSelector
KNeighborsClassifier
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
1
2
3
4
5
6
7
8
9
10
11
Number of features in the selected subset

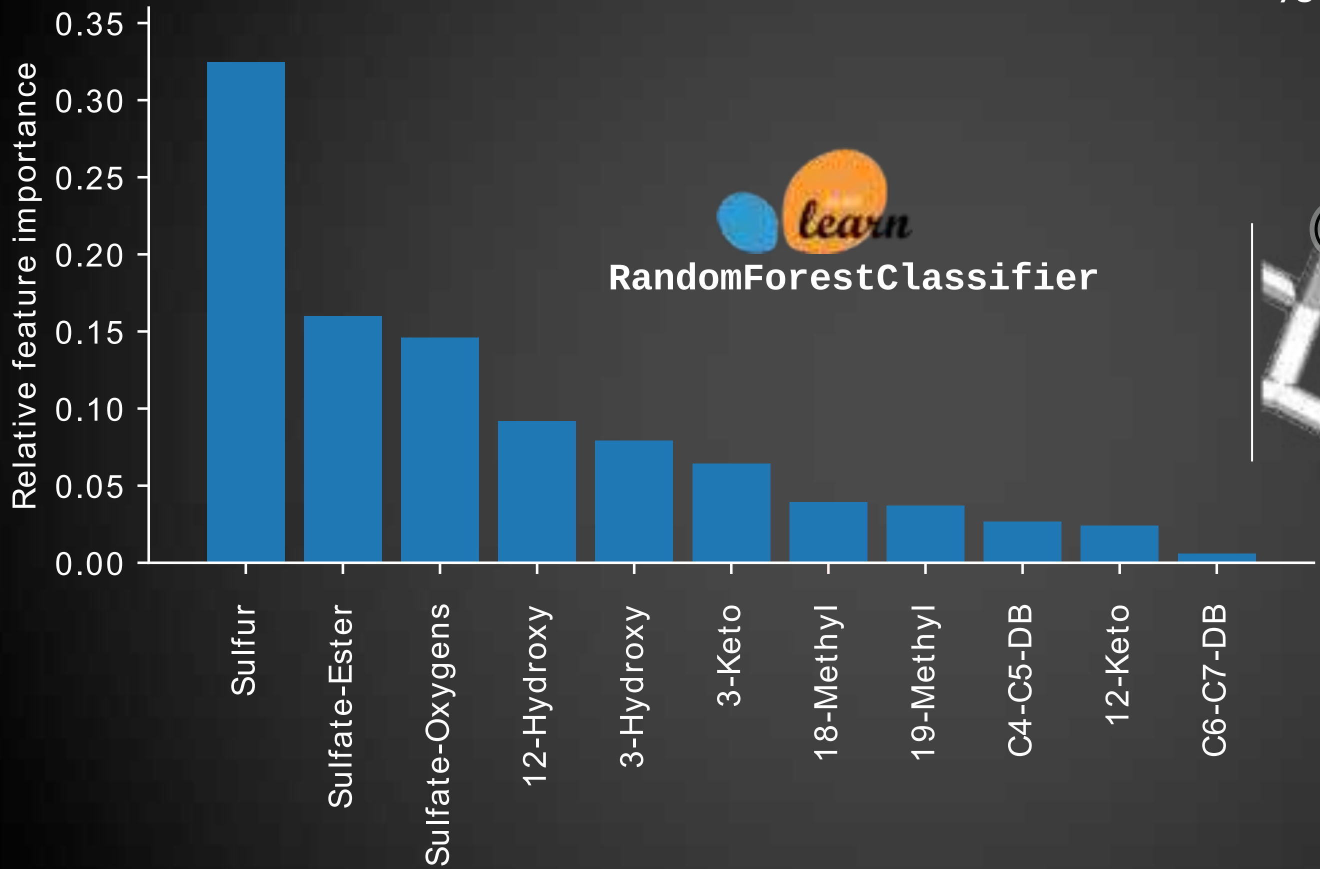

‘Sulfur-Oxygens’
‘Sulfur’
‘Sulfur-Oxygens’
‘Sulfur’
‘19-Methyl’
‘Sulfur-Oxygens’
1.0
0.9
0.8
0.7
0.6
0.5
Accuracy
1
2
3
4
5
6
7
8
9
10
11
Number of features in the selected subset

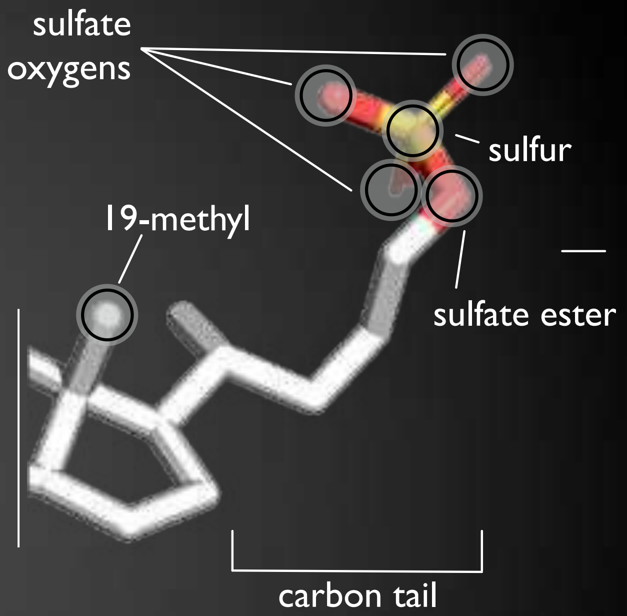


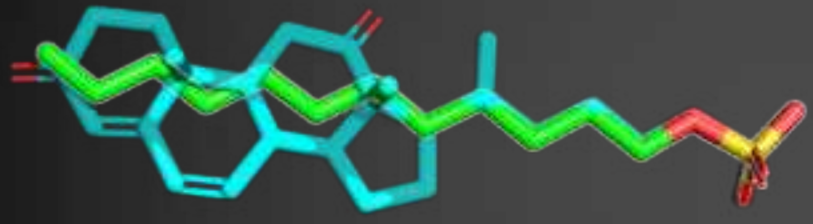
'Sulfur-Oxygens'
'Sulfur'
'Sulfur-Oxygens'
'Sulfur'
'19-Methyl'
'Sulfur-Oxygens'
1.0
0.9
0.8
0.7
0.6
0.5
Accuracy
1 2 3 4 5 6 7 8 9 10 11
Number of features in the selected subset
sulfate oxygens
sulfur
19-methyl
sulfate ester
carbon tail

DecisionTreeClassifier

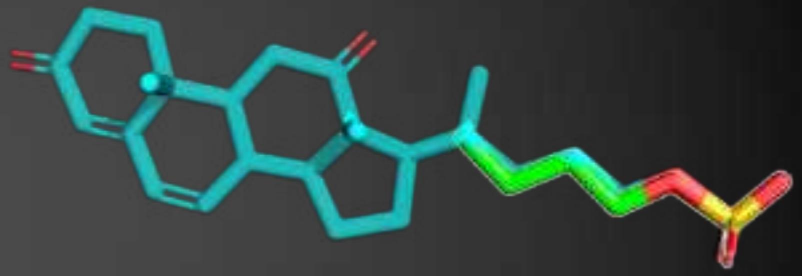



RandomForestClassifier
learn
Relative feature importance
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Sulfur
Sulfate-Ester
Sulfate-Oxygens
12-Hydroxy
3-Hydroxy
3-Keto
18-Methyl
19-Methyl
C4-C5-DB
12-Keto
C6-C7-DB


sulfate oxygens
sulfur
19-methyl
sulfate ester
carbon tail

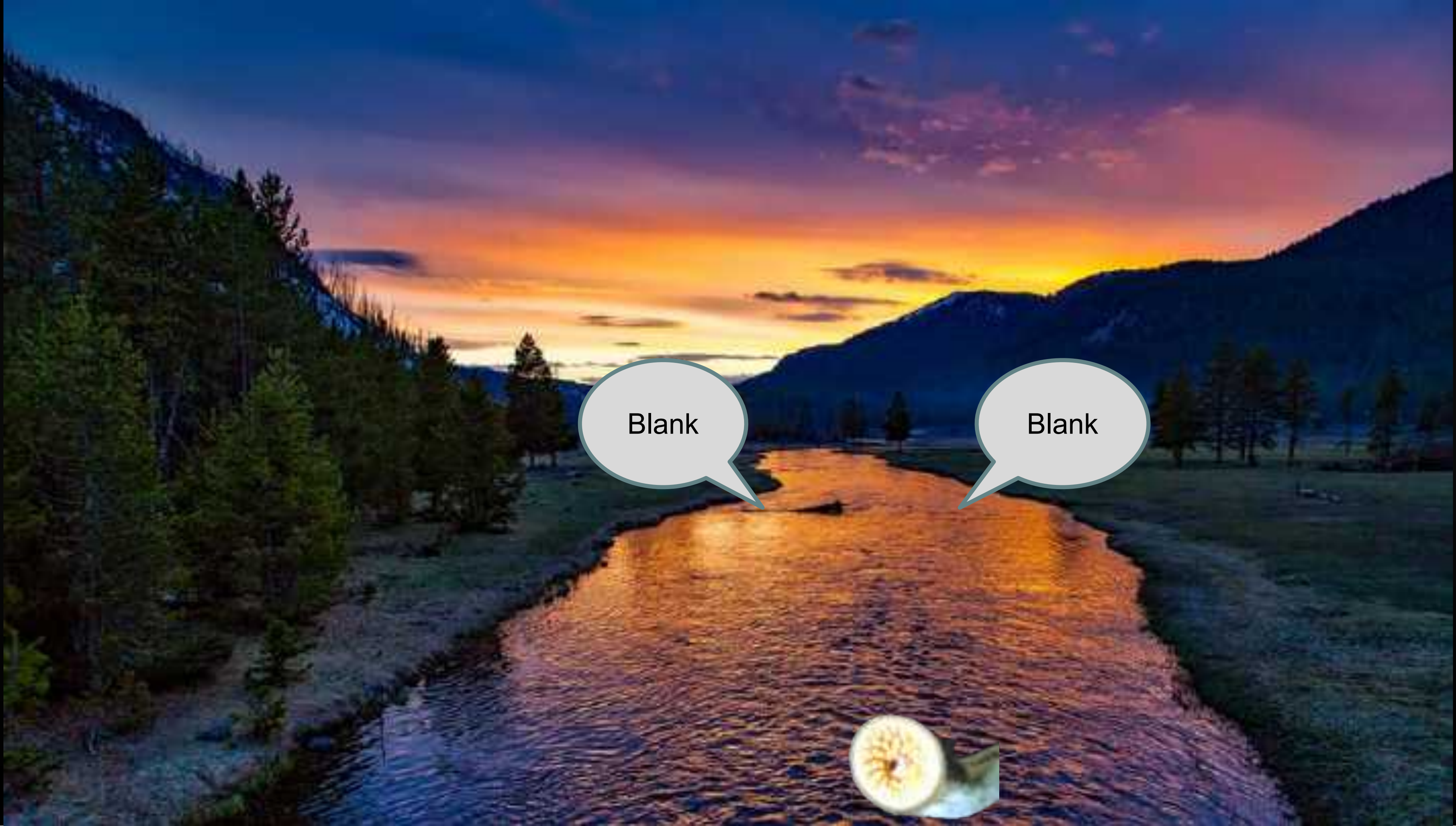


69% signal inhibition



62% signal inhibition

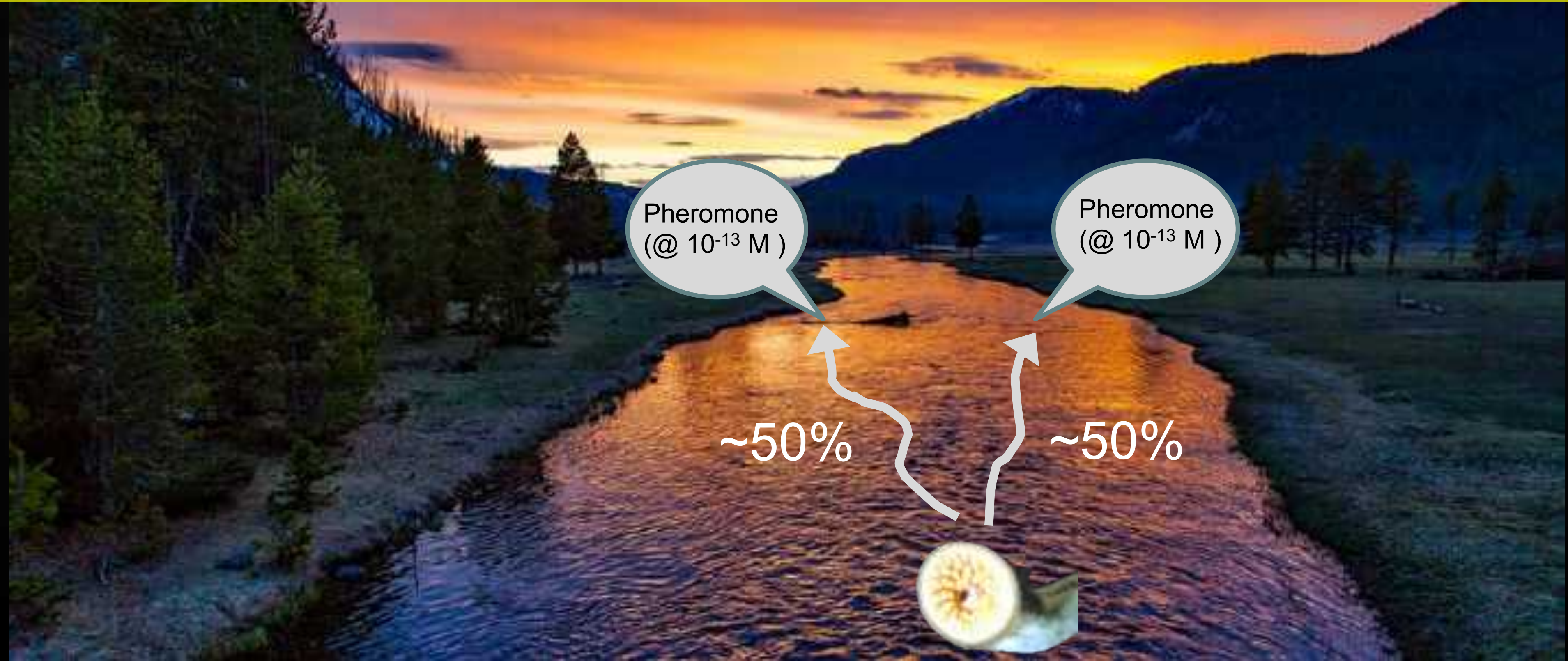
Outcome

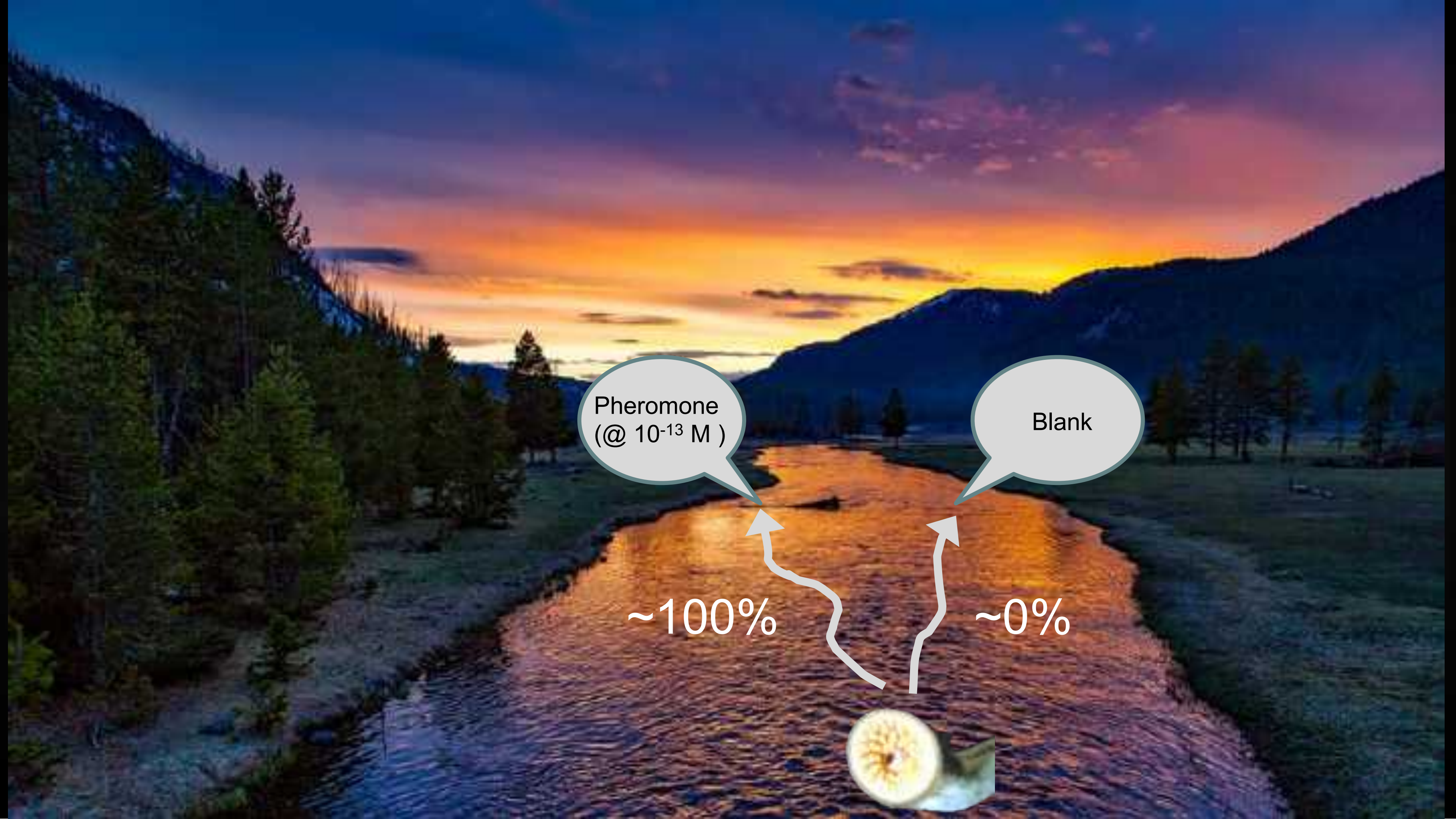

Blank
Blank


Pheromone
(@ 10^{-13} M)
Pheromone
(@ 10^{-13} M)
~50%
~50%

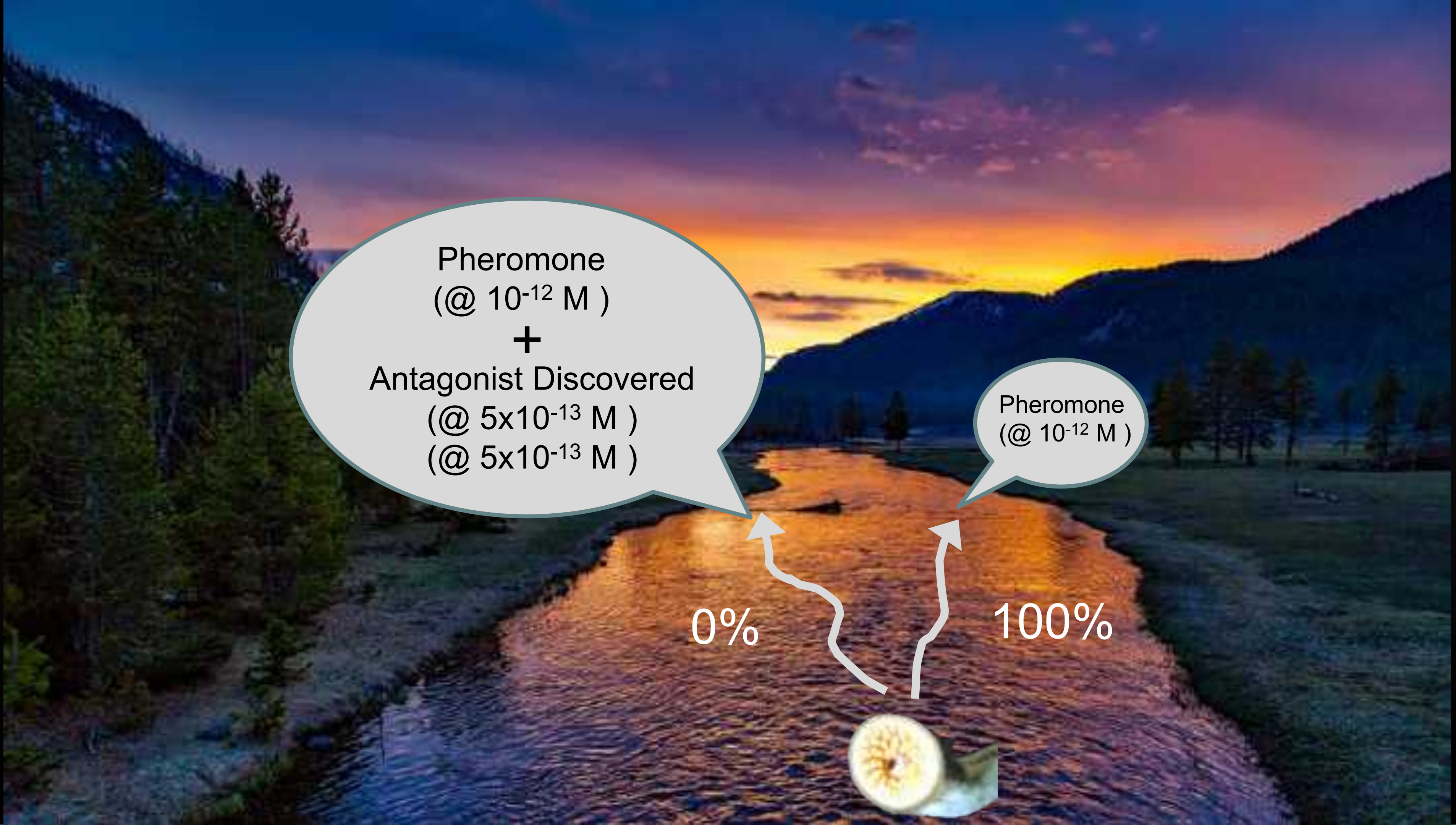
Concentration of gold in the ocean: 4 x 10^{-11} M

(https://web.stanford.edu/group/Urchin/mineral.html)





Pheromone
(@ 10-13 M)
Blank
~100%
~0%



Pheromone
(@ 10^{-12} M)
+
Antagonist Discovered
(@ $5x10^{-13}$ M)
(@ $5x10^{-13}$ M)
Pheromone
(@ 10^{-12} M)
0%
100%

Acknowledgements

Kuhn Lab

Leslie A. Kuhn
Nan Liu
Santosh Gunturu
Jiaxing Chen

Weiming Li Lab

Weiming Li
Anne M. Scott
Mar Huertas

Great Lakes Fishery Commission

Software & Developers

Python (https://www.python.org)

Matplotlib (https://matplotlib.org)

Scikit-learn (http://scikit-learn.org)

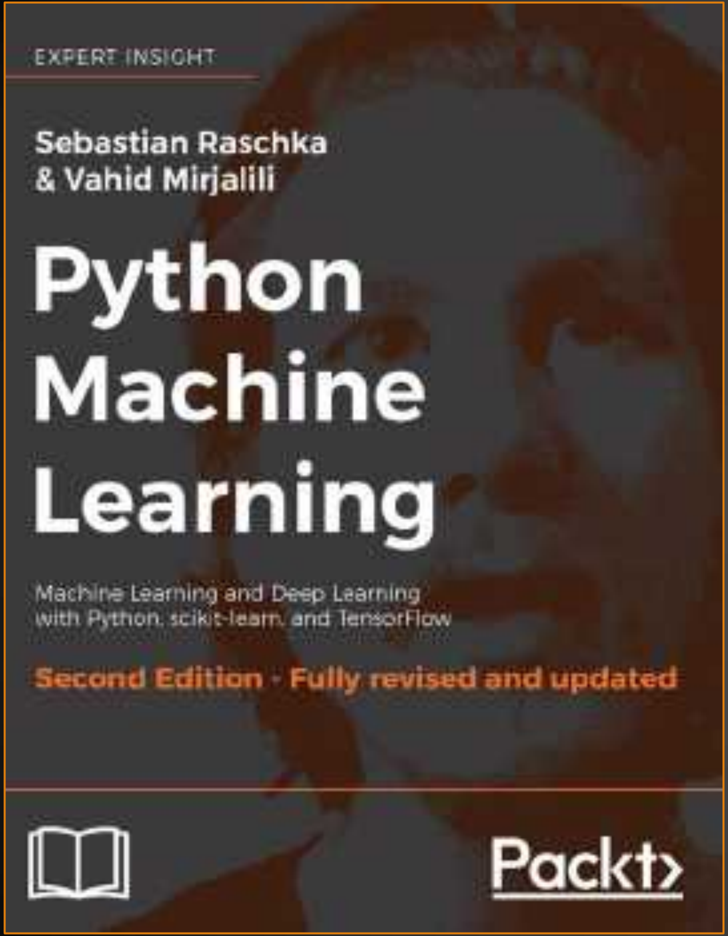
IPython (https://ipython.org)

Jupyter Notebook (http://jupyter.org)

Pandas (https://pandas.pydata.org)

OpenEye (https://www.eyesopen.com)

OpenBabel (http://openbabel.org)



Book signing

Nov 04 (Fri)
1:00 PM
Regency B

Thanks!

Questions?

BioPandas (https://rasbt.github.io/biopandas/)

Screenlamp (https://psa-lab.github.io/screenlamp)