Prediction of Boiling Point using Graph Convolutional Neural Networks

By Max A. Maximov

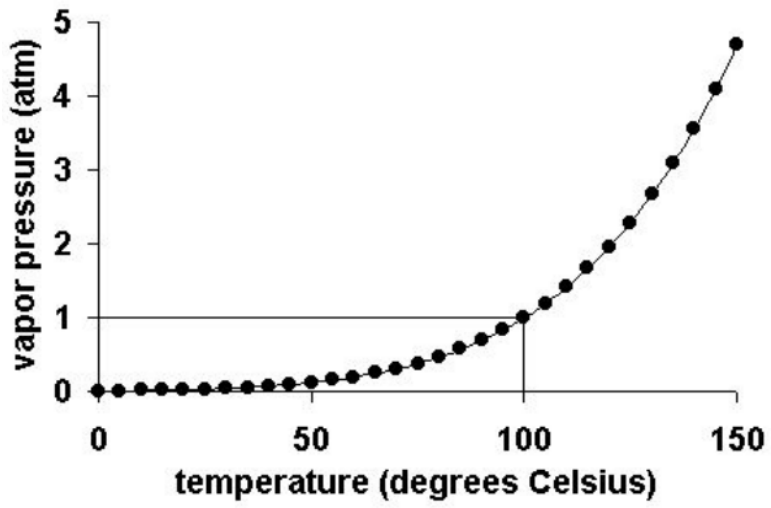
Advisor: Dr. Gennady Gor

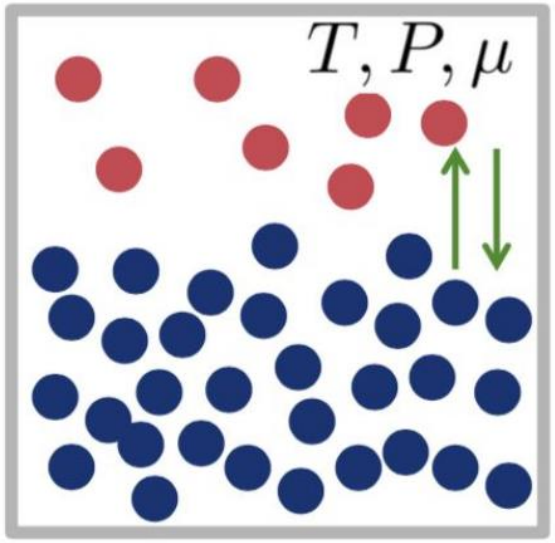
Group meeting presentation

July 2020

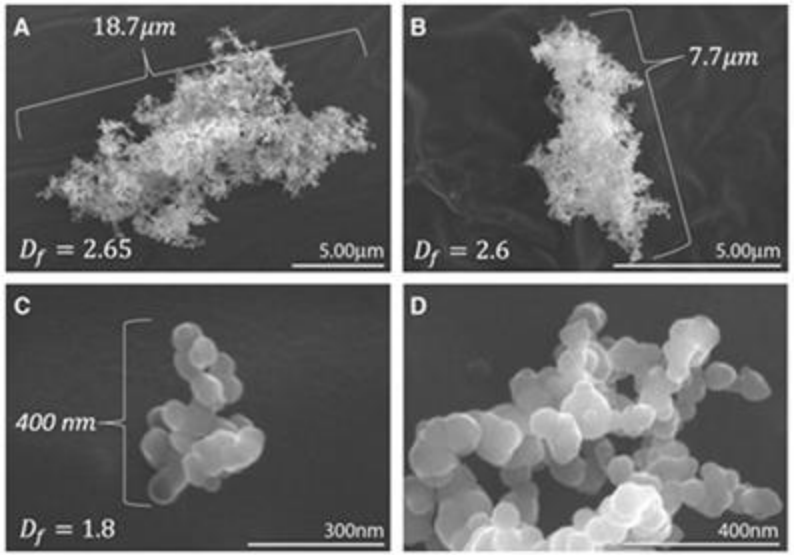
Introduction

- Def. Boiling point is the temperature at which the liquid-vapor or solid-vapor transition occurs
- Depends on
  - Pressure (higher P, higher T)
  - Type of molecule (usually stronger interaction, lower P)





Motivation



- Black carbon from incomplete fuel combustion as environmental pollutant
- Climate forcing affected by chemical interactions
  - Modifications to mixing state and morphology
- **Experiment:** Exposing BC aerosol to supersaturated vapors of different chemicals to understand restructuring of BC aggregates

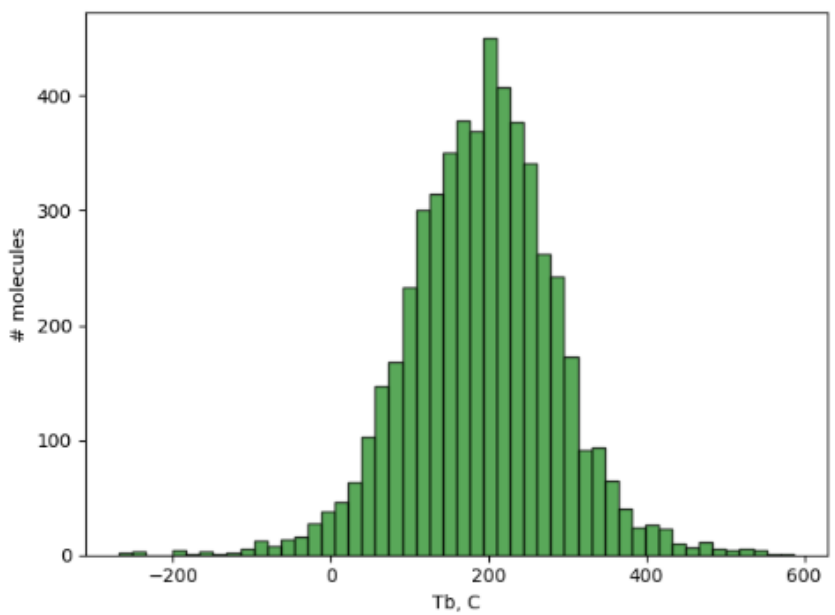
Ways to estimate boing point

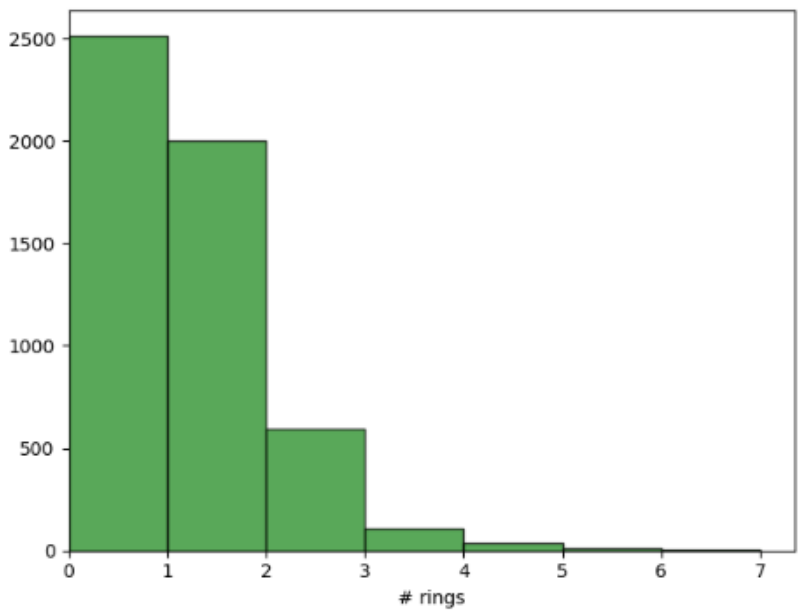
- Experimental data
- Simulation (e.g., HMC-WL)
- Group contribution methods (e.g., UNIFAC)
- Machine Learning

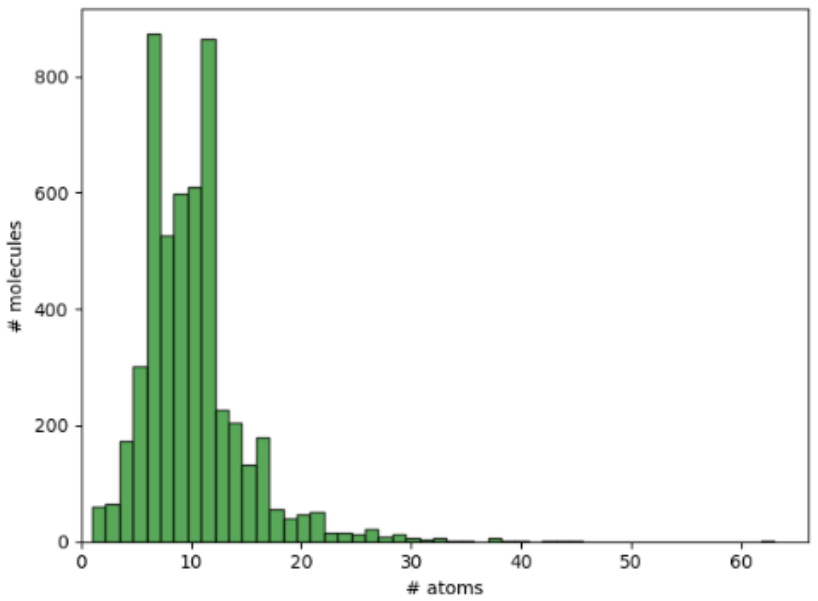
Objective: Can we predict the boiling point from the structure?

Pubchem Scrapped Dataset

- Scanned about 1M molecules
- About 5000 molecules had normal boiling point

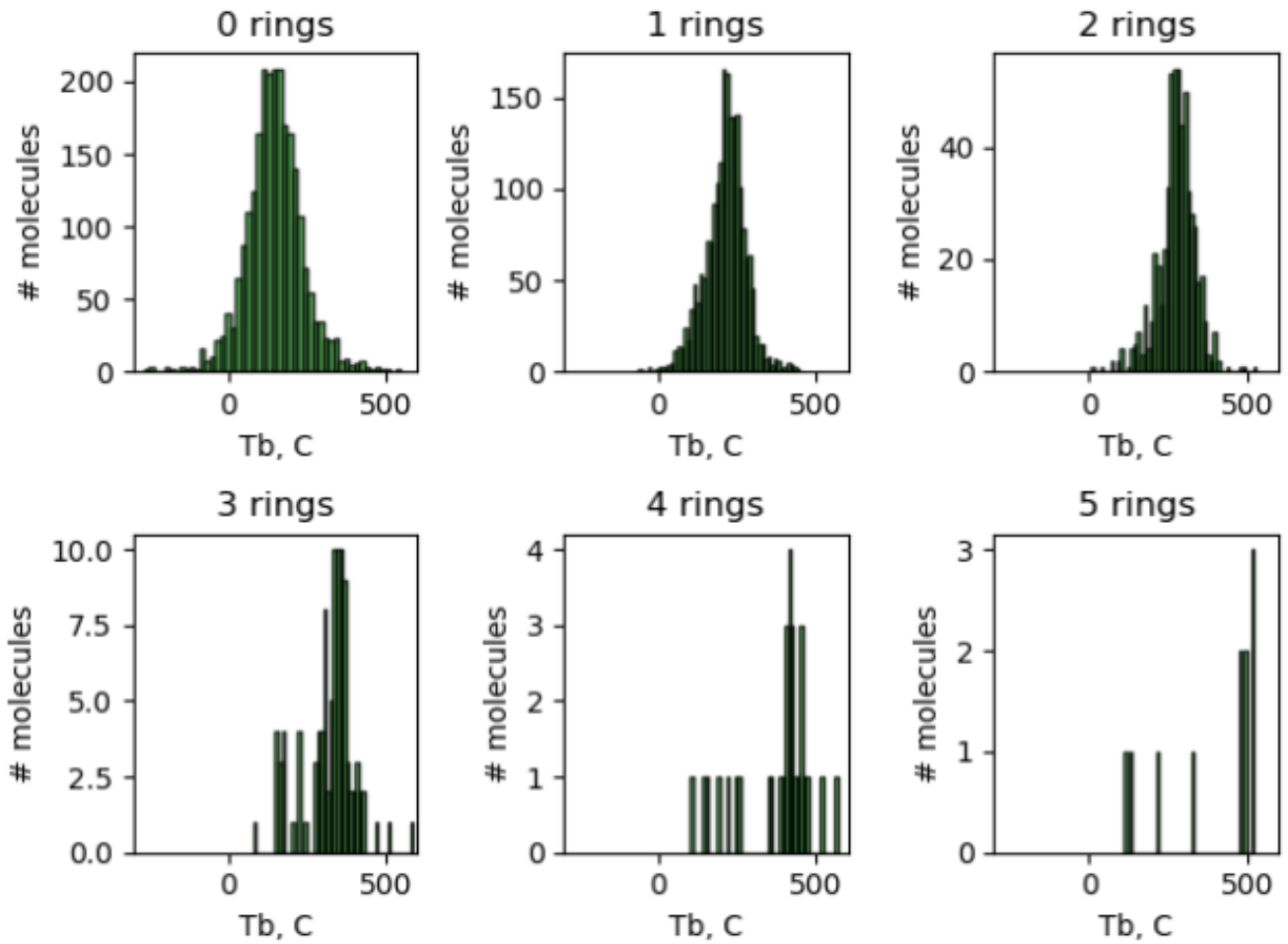






https://pubchem.ncbi.nlm.nih.gov/

Structure vs Normal Boiling Point

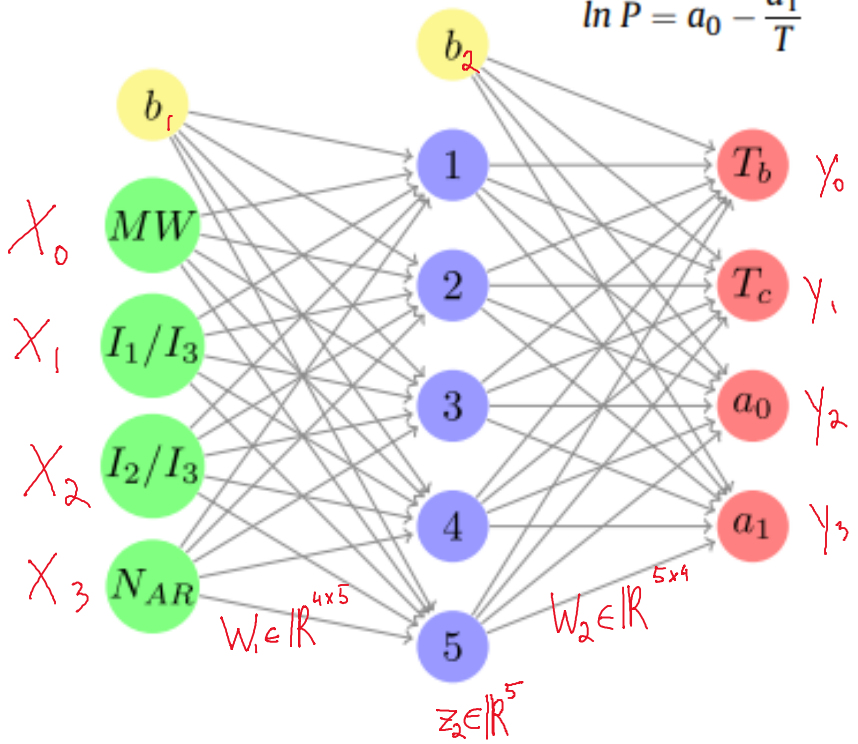


Simple ML model

		MW (g/mol)	I_1/I_3	I_2/I_3	N_{AR}	T_b (K)	T_c(K)	T_b^{exp} (K)
Naphthalene		128	0.27	0.73	2	495	775	473
Anthracene		178	0.16	0.84	3	626	921	614
Phenantrene		178	0.25	0.75	3	630	923	611
Fluoranthene		198	0.31	0.69	3	678	1001	656
Tetracene		228	0.1	0.9	4	742	1041	723
Triphenylene		228	0.5	0.5	4	755	1083	712
Chrysene		228	0.17	0.83	4	747	1054	714
Pyrene		202	0.35	0.65	4	703	1057	666
Benz[a]anthracene		228	0.17	0.83	4	739	1070	708
Benzo[a]pyrene		252	0.22	0.78	5	827	1163	768
Acephenanthrylene		202	0.28	0.72	3	683	995	679
7H-Benzo[c]fluorene		216	0.25	0.75	3	707	1013	671

Groven, Steven D., C. Desgranges, and J. Delhommelle. *Fluid Phase Equilibria* 484 (2019): 225-231.

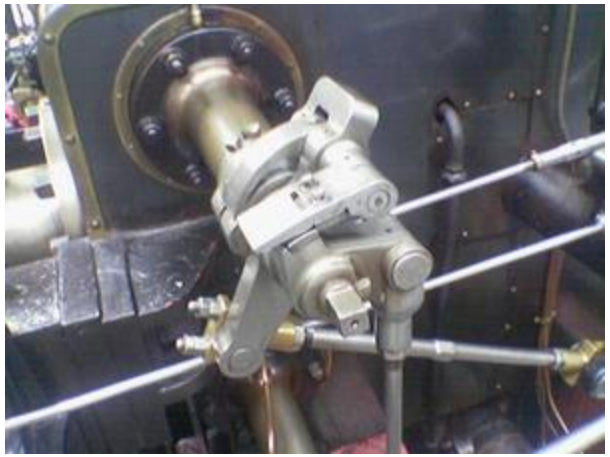
$$\ln P = a_0 - \frac{a_1}{T}$$

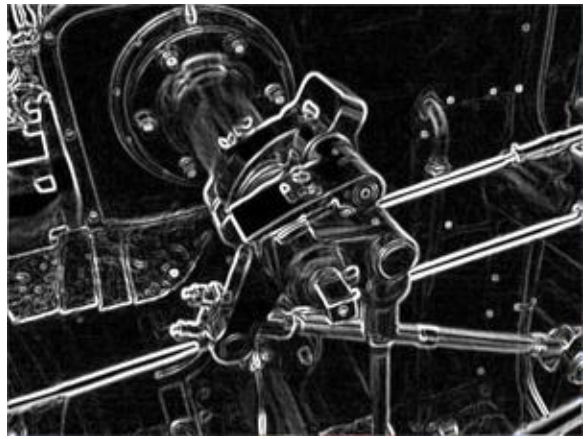


$$\forall i = 0..4 \quad (z_2)_i = \tanh\left((b_1)_i + \sum_{k=0}^{3} (W_1)_{ik} X_k\right)$$

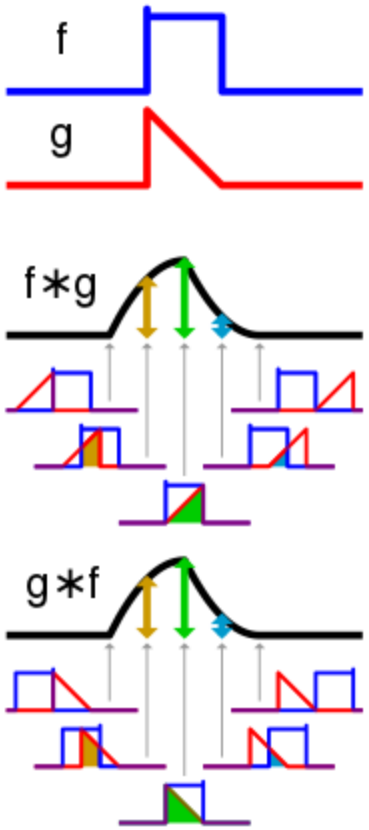
$$\forall j \in \{0,1,2,3\} \quad Y_j = (b_2)_j + \sum_{i=0}^{4} (W_2)_{ji} (z_2)_i$$

Convolutional nets intro: Edge detection





Convolution and Edge filter (Sobel and similar)



$\sum a_{i+k,j+m} w_{ij} =$

$= 10 \cdot 1 + 10 \cdot 1 + 10 \cdot 1 + 0 \cdot 10 + \ldots = 30$

10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0

Vertical

$*$

1	0	-1
1	0	-1
1	0	-1

3×3

$=$

0	30	30	0
0	30	30	0
0	30	30	0
0	30	30	0

Horizontal

10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0
0	0	0	10	10	10
0	0	0	10	10	10
0	0	0	10	10	10

$*$

1	1	1
0	0	0
-1	-1	-1

$=$

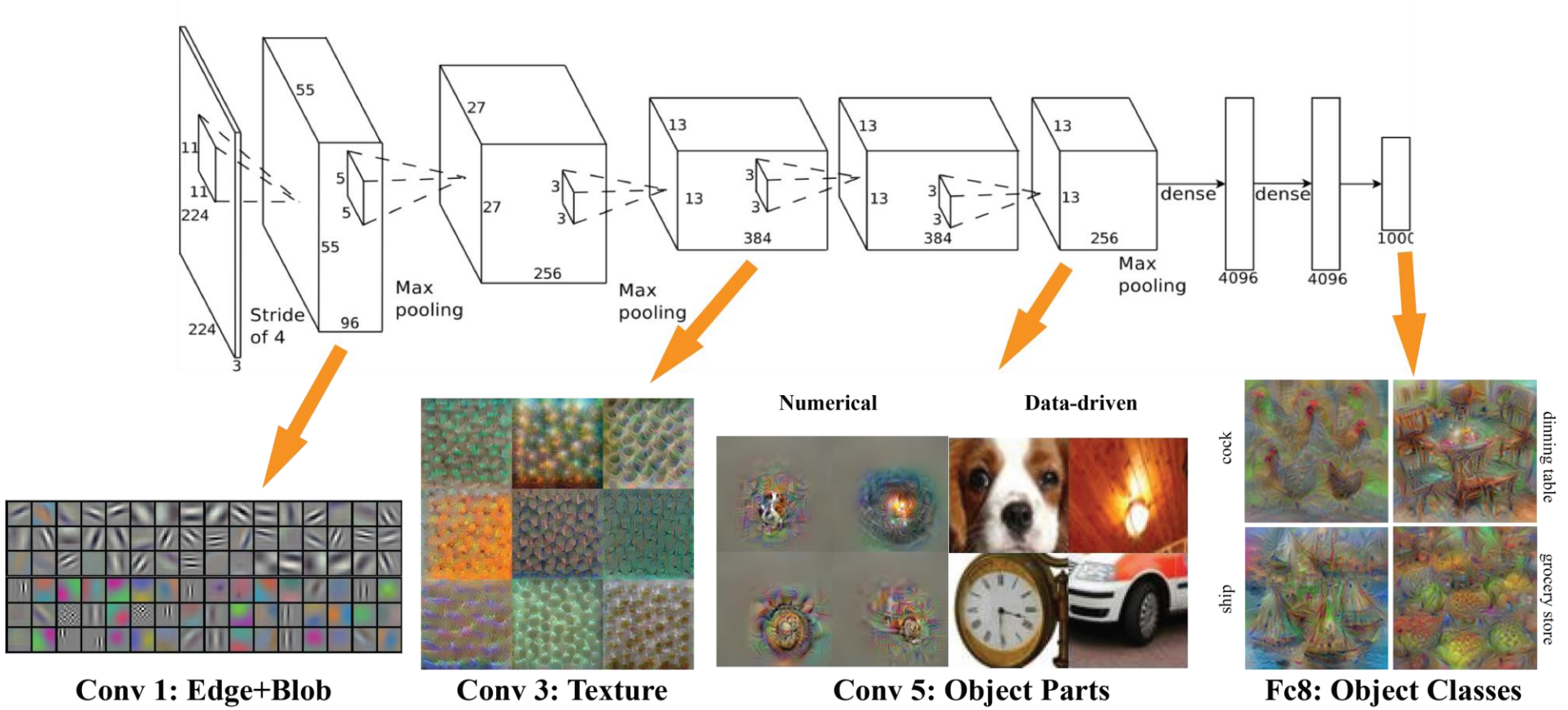
0	0	0	0
30	10	-10	-30
30	10	-10	-30
0	0	0	0

Generic (any slope)

w_1	w_2	w_3
w_4	w_5	w_6
w_7	w_8	w_9

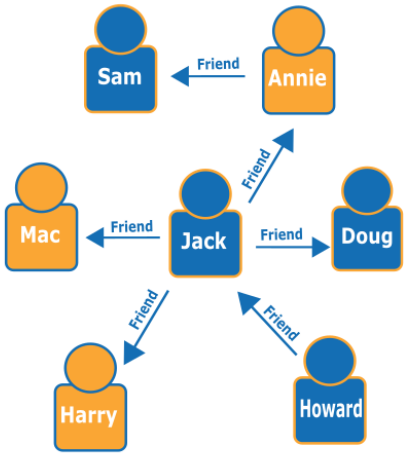
We can learn these parameters from a ref image!

Convolutional nets intro: image classification

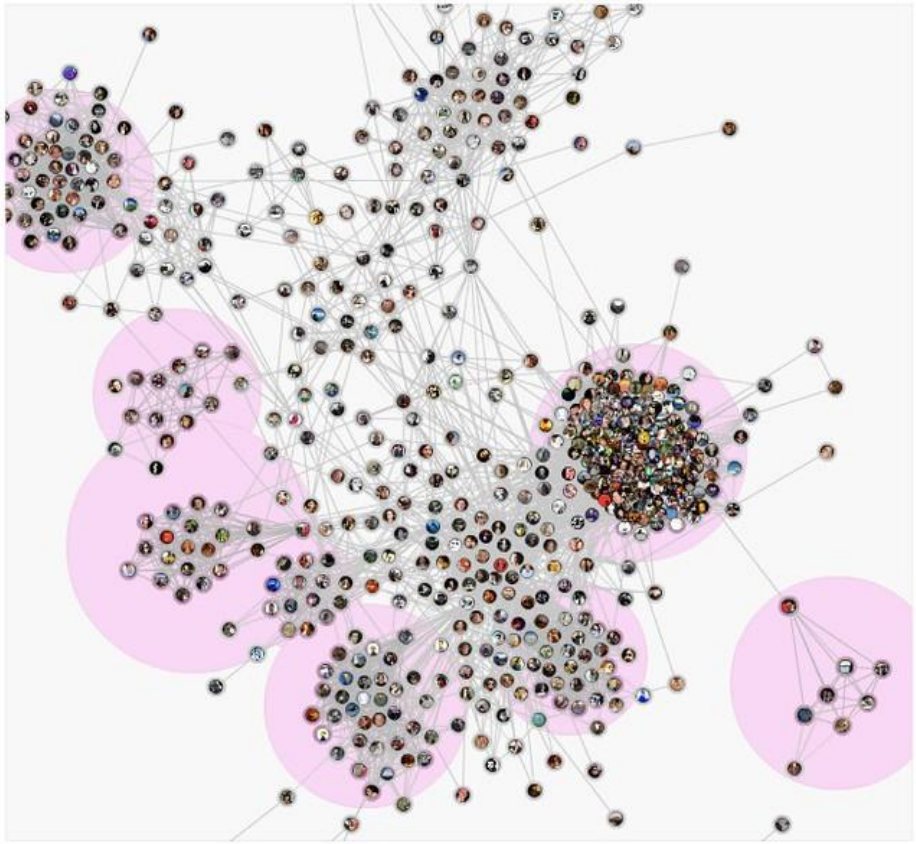


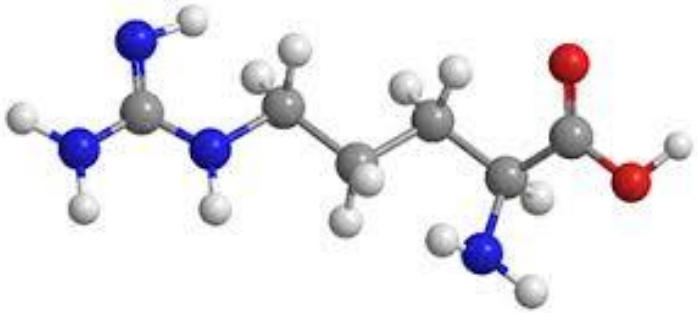
Graph

- G = (V, E) is ordered pair of vertices (V) and edges (E)



V – people
E – social connection

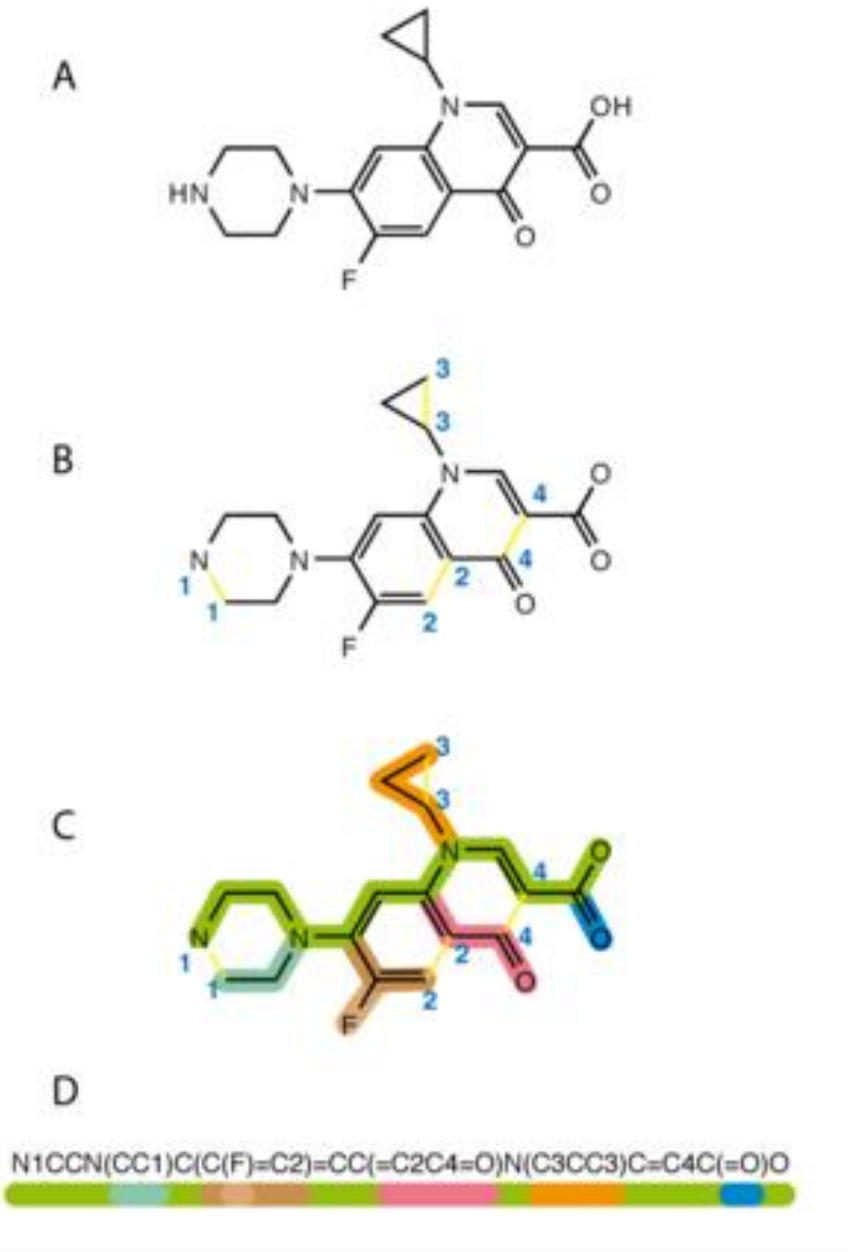




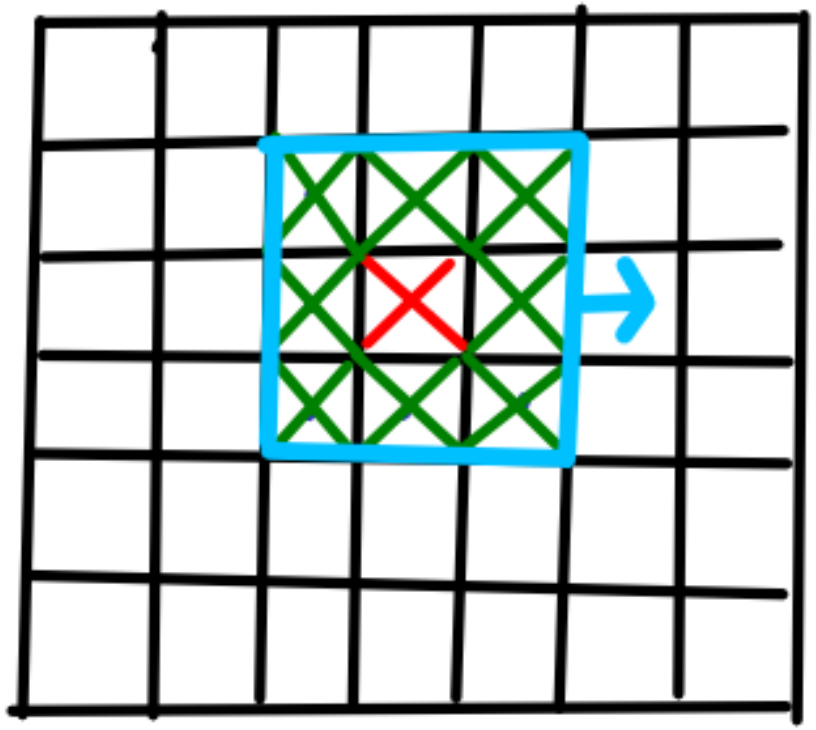
V – atoms
E - bonds

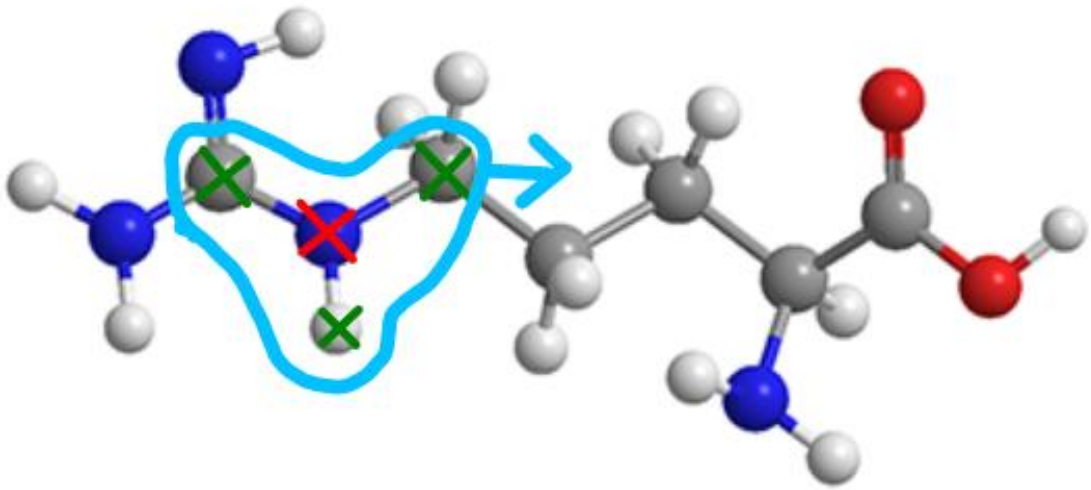
Smiles strings

- String s -> Graph G
- CN=C=O -> CH_3−N=C=O

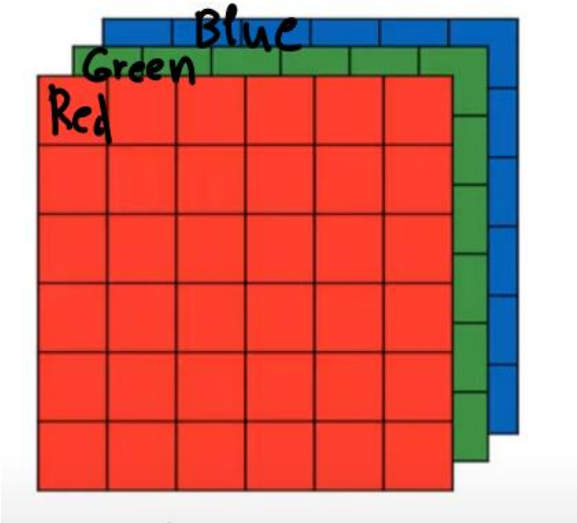


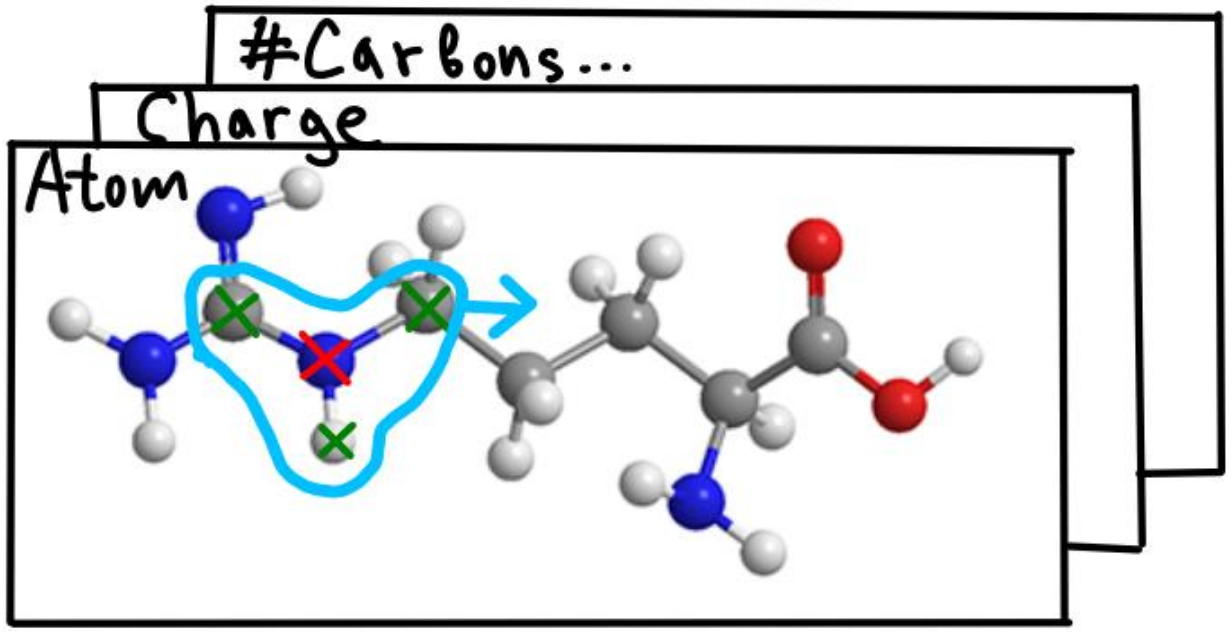
Graph Convolution



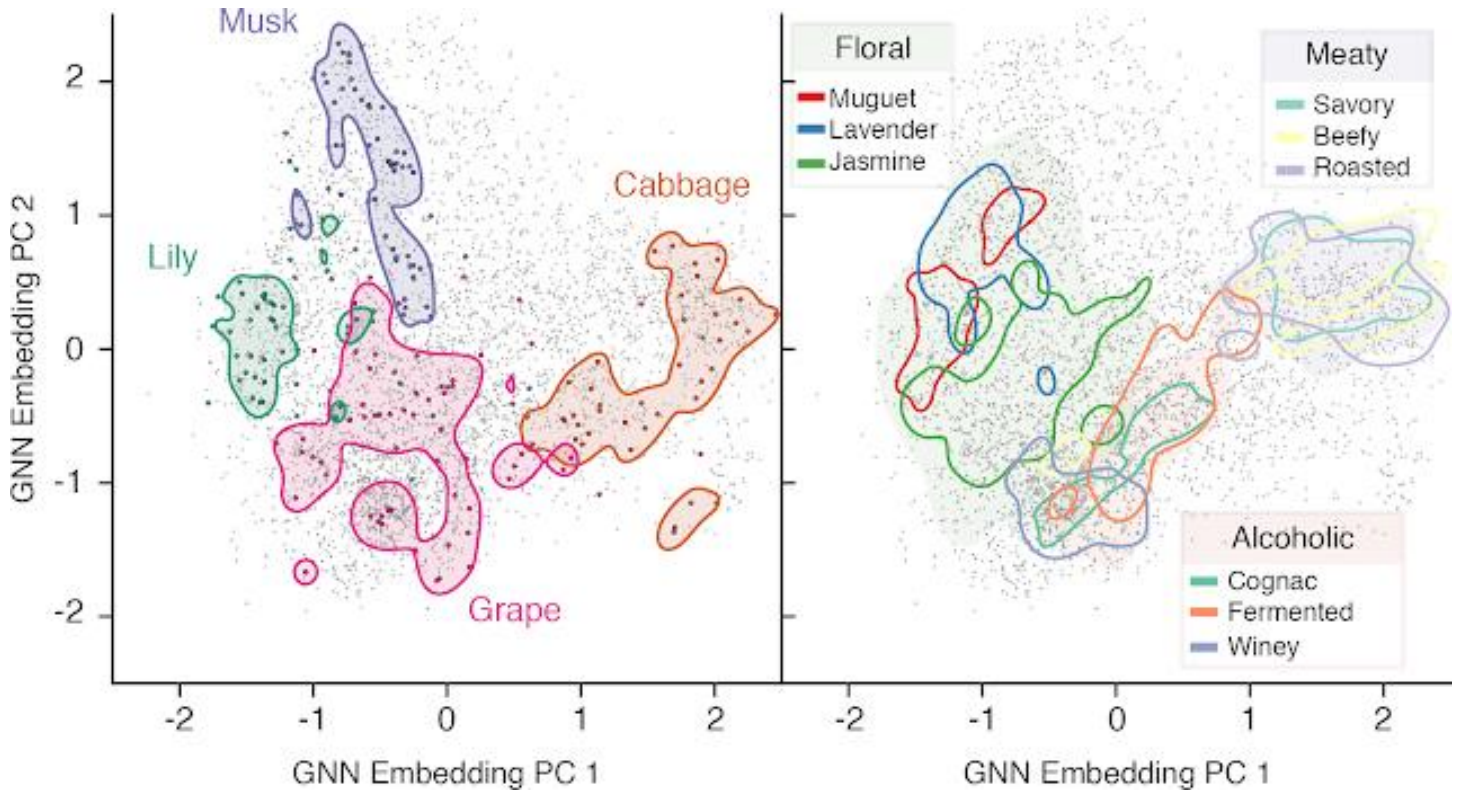


Graph convolution over volume





Learning to smell



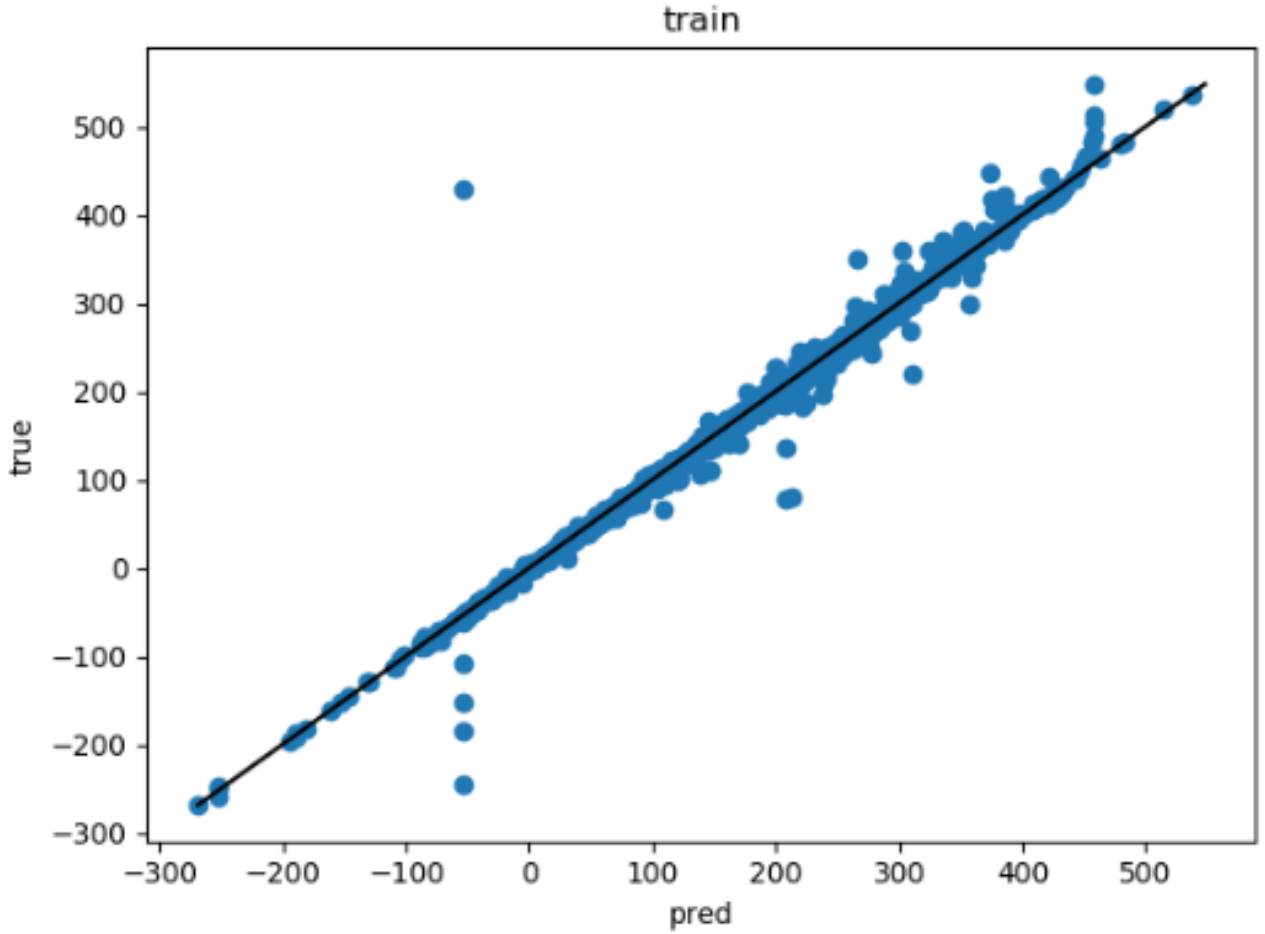
https://ai.googleblog.com/2019/10/learning-to-smell-using-deep-learning.html

Deepchem package

- Based on TensorFlow, rdKit
- Implements Graph convolutions
- Includes 112 featurizers
- Simple: NumSaturatedRings, ExactMolWt
- Extended-Connectivity Fingerprints (Rogers et al, J. Chem. Inf. Model., 2010)

Results

So far, even overfitted train loss is not that good



Conclusion

- Collected a large dataset
- Need to work on outliers
- It is worth trying other featurizers, many fingerprints are developed
- Reported errors in NIST database

Backup slides

Cool packages

- rdkit
- deepchem

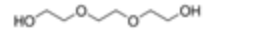
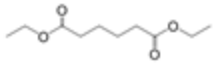
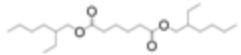
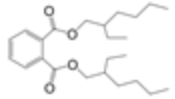
Dataset cleanup

- Removed outliers with Tb > 600 C
- Removed species which decompose at certain temp

ECFP fingerprints

- https://docs.chemaxon.com/display/docs/Extended+Connectivity+Fingerprint+ECFP
- https://pubs.acs.org/doi/pdf/10.1021/ci100050t

Sample experimental data

Coating material	Molecular structure
none	
	Intermediate volatility
TEG	
TEGMBE	
DA	
TDA	
	Low volatility liqu
DOS	
OA	
BEHA	
BEHP	
SA	H_2SO_4

coating material	acronym	molar mass, g mol^{-1}	density, g cm^{-3}	surface tension, mN m^{-1}	boiling point, °C	vapor pressure, Pa[b,c]
triethylene glycol	TEG	150.17	1.12	46.5	287	0.18
triethylene glycol monobutyl ether	TEGMBE	206.28	0.99	31.4	278	0.33
diethyl adipate	DA	202.25	1.01	32.7	245	7.73
tetradecane	TDA	198.38	0.76	26.7	251	2.00
dioctyl sebacate	DOS	426.67	0.91	31.1	436	2.96×10^{-6}
oleic acid	OA	282.46	0.90	32.8	360	7.28×10^{-5}
bis(2-ethylhexyl) adipate	BEHA	370.57	0.92	30.0	417	1.13×10^{-4}
bis(2-ethylhexyl) phthalate	BEHP	390.56	0.99	31.2	384	1.89×10^{-5}
sulfuric acid (H_2SO_4, 69 wt %)	SA	98.08	1.60	73	165[a]	3.87×10^{-7}

Chen, C., Enekwizu, O.Y., Fan, X., Dobrzanski, C.D., Ivanova, E.V., Ma, Y., Gor, G.Y. and Khalizov, A.F. Environ Sci Technol 52.24 (2018): 14169-14179.