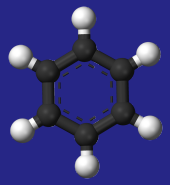
Aromaticity
Mann & Costa
Does it smell?
Why to care?
How to do?
Does it work?
More to do?
A sketch
Take home

Does it smell? Aromaticity Prediction in Molecule Graphs

Martin Mann and Fabrizio Costa

Bioinformatics
University of Freiburg



What the hell is . . .

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

Does it work?

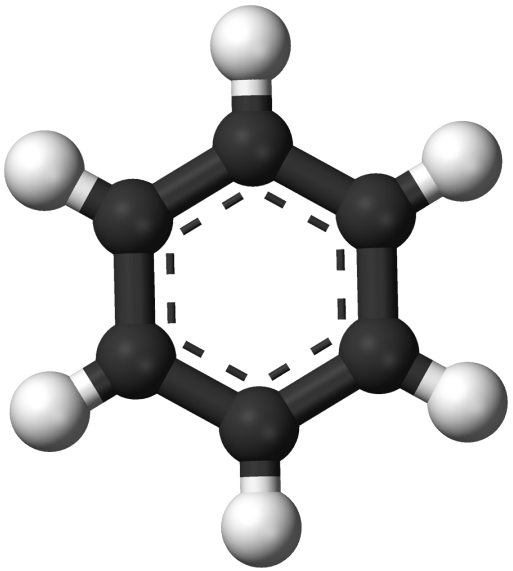
More to do?

A sketch

Take home

. . . Aromaticity ?

Smells good . . .



. . . looks good !

What the hell is Aromaticity?

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

Does it work?

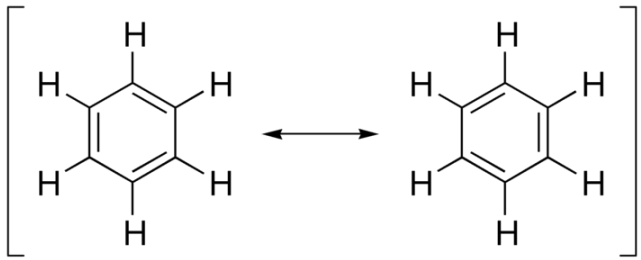
More to do?

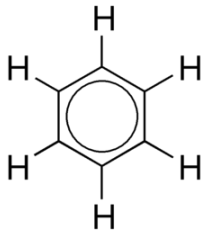
A sketch

Take home

Wikipedia says

- "...a chemical property in which a conjugated ring exhibits a stabilization stronger than expected"
- "...electrons are free to cycle around circular arrangements of atoms which are alternately single- and double-bonded" (e.g. benzene)





What the hell is Aromaticity?

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

Does it work?

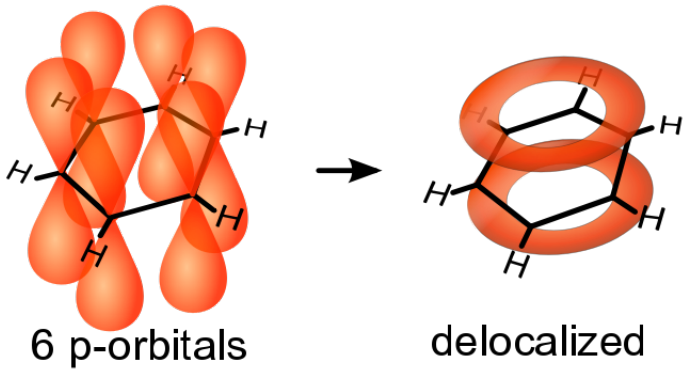
More to do?

A sketch

Take home

Hückel (1931) and von Doering (1951) say

- "...a cyclic ring is aromatic when the number of its π-electrons equals $4n + 2$ where $n \geq 0$"
- only for single ring molecules and $0 \leq n \leq 6$
- rule fails for many molecules



What the hell is Aromaticity?

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

Does it work?

More to do?

A sketch

Take home

Stanger (2009) asks

FEATURE ARTICLE www.rsc.org/chemcomm | ChemComm

What is. . . aromaticity: a critique of the concept of aromaticity—can it really be defined?

Amnon Stanger*

Received (in Cambridge, UK) 25th September 2008, Accepted 12th January 2009
First published as an Advance Article on the web 9th February 2009
DOI: 10.1039/b816811c

"Therefore, with the current state of knowledge, the answer to the question posed in the title has to be negative." (Stanger,2009)

Why to bother with Aromaticity?

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

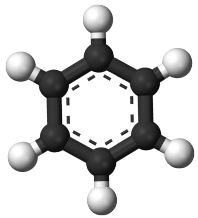
Does it work?

More to do?

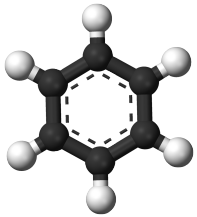
A sketch

Take home

- . . . large effect on the physicochemical properties
  - $\Rightarrow$ reactivity estimation
  - $\Rightarrow$ energy calculation

E( **) = ???**

- . . . hinders canonicalization
  - $\Rightarrow$ database search / fingerprints
  - $\Rightarrow$ compact representation (e.g. SMILES)

 $\Rightarrow$ **c1ccccc1**

Aromaticity Perception

The state-of-the-art = rule-based

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

Does it work?

More to do?

A sketch

Take home

Current tools:

- Daylight
- Marvin
- OpenBabel

DAYLIGHT
Chemical Information Systems, Inc.

ChemAxon

Methods:

- rule-based (e.g. Hückel-rule)
- pattern-based
- explicit exception handling

$\Rightarrow$ **no consistent prediction !**

Knows the vulture . . .

. . . and so it's proven, since **If he knows it smells!**

Aromaticity

Mann & Costa

Does it smell?

Why to care?

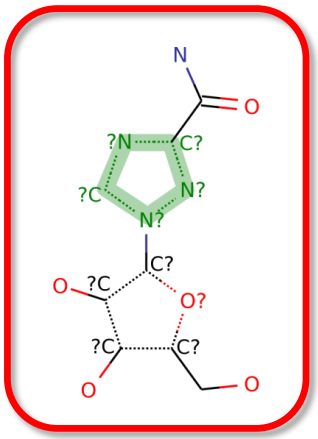
How to do?

Does it work?

More to do?

A sketch

Take home



Yes

No

Other options . . . ???

Aromaticity

Mann & Costa

Does it smell?

Why to care?

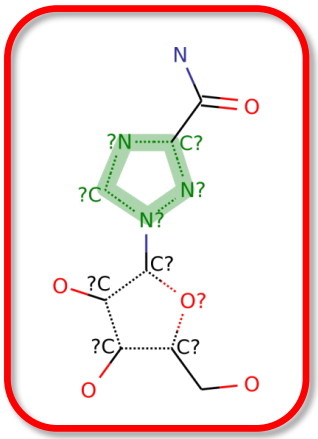
How to do?

Does it work?

More to do?

A sketch

Take home



Yes

No

Aromaticity Perception

NEW idea : data-driven approach !

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

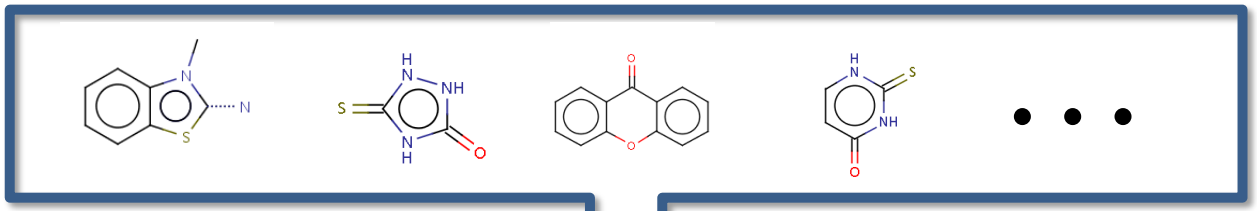
Does it work?

More to do?

A sketch

Take home

Given: large set of annotated molecules



Predict if a ring is aromatic or not

Data-driven Aromaticity Perception

How to train the SVM?

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

Does it work?

More to do?

A sketch

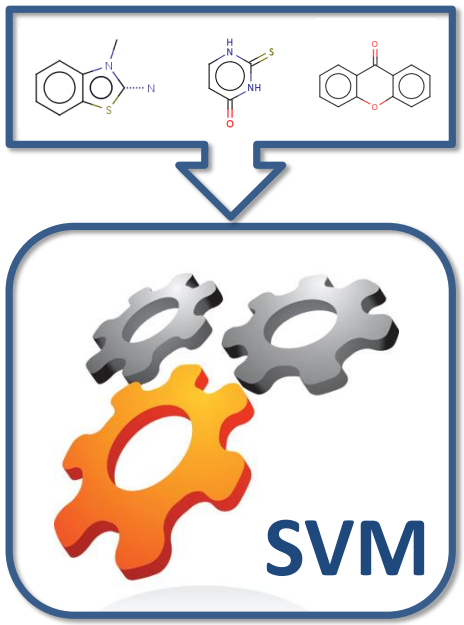
Take home

We need:

1. Molecule representation (graph)
2. Feature description (SVM kernel)

What to encode:

- atoms and bonds
- highlight rings as uncertain
- no hydrogens (not always known)



Data-driven Aromaticity Perception

Molecule graph encoding

Aromaticity

Mann & Costa

Does it smell?

Why to care?

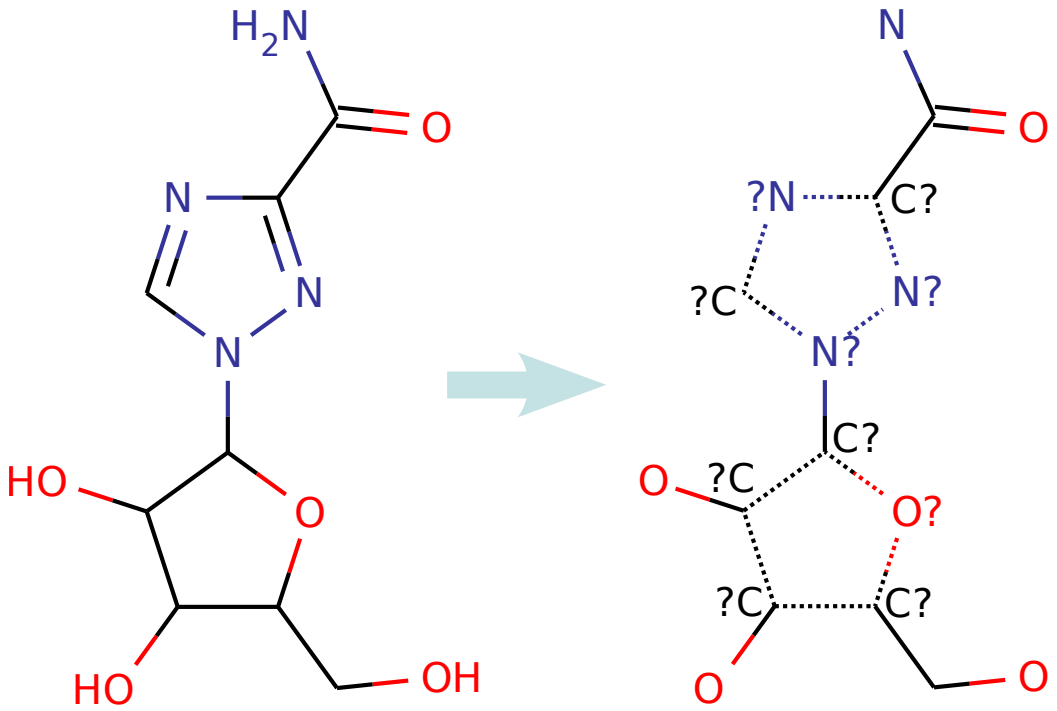
How to do?

Does it work?

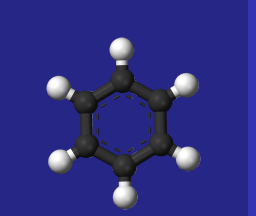
More to do?

A sketch

Take home



- hydrogens are removed
- rings are labeled as uncertain



Data-driven Aromaticity Perception

Molecule graph encoding

Aromaticity

Mann & Costa

Does it smell?

Why to care?

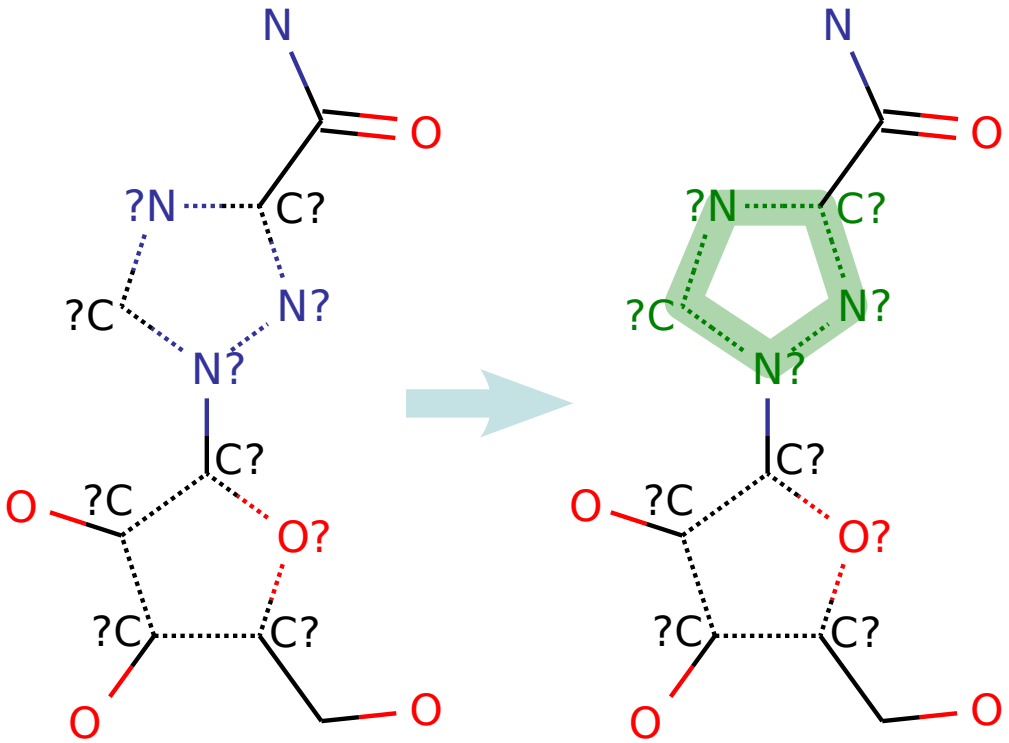
How to do?

Does it work?

More to do?

A sketch

Take home



Ring of interest (train/prediction) is highlighted (extra label)

Data-driven Aromaticity Perception

Feature extraction : NSPDK graph kernel (Costa, 2010)

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

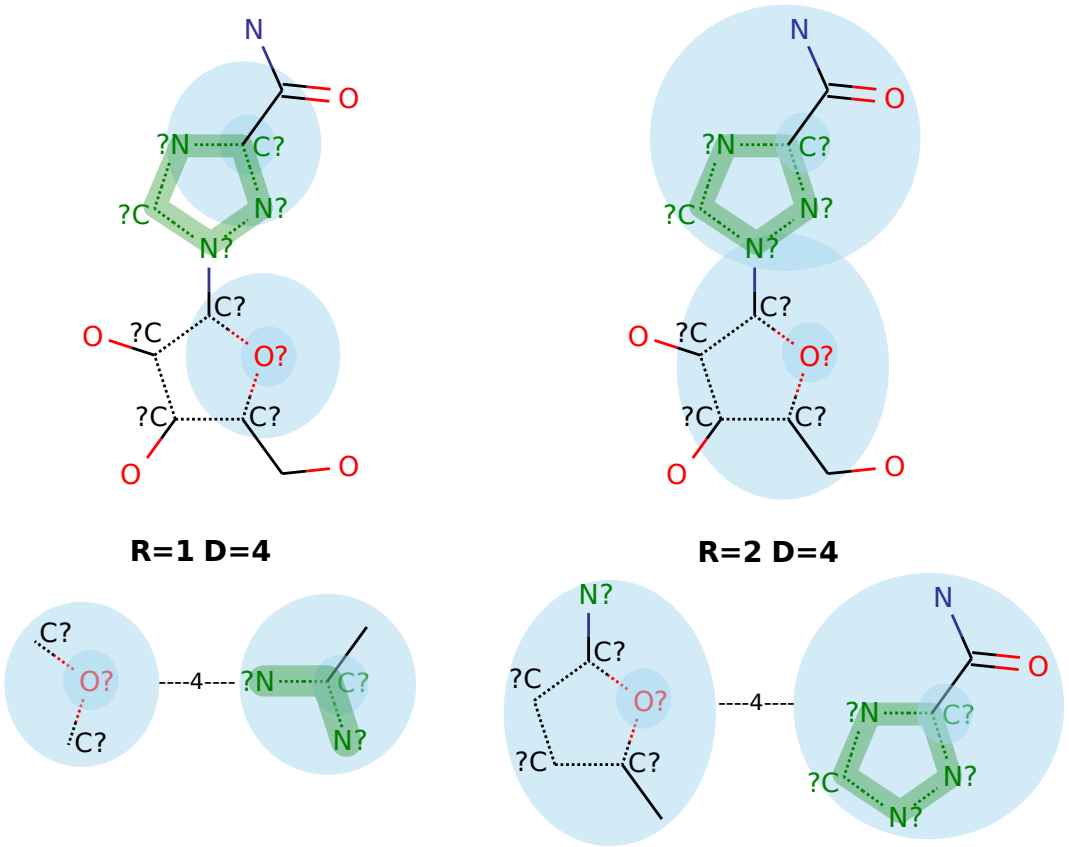
Does it work?

More to do?

A sketch

Take home

Neighborhood **S**ubstructure **P**airwise **D**istance **K**ernel



Each feature gets a hash code $\rightarrow$ bit vector representation

Data-driven Aromaticity Perception

Feature hashing : NSPDK graph kernel (Costa, 2010)

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

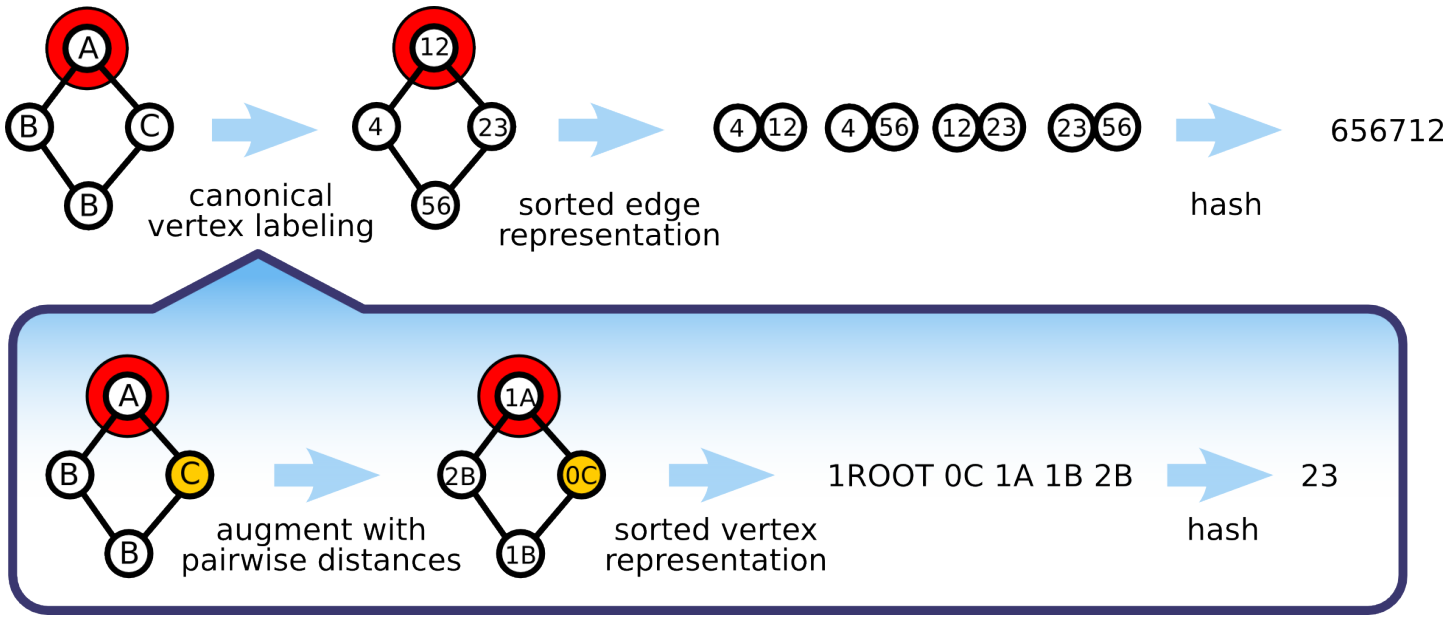
Does it work?

More to do?

A sketch

Take home

Each subgraph gets a hash number



Each node (e.g. ⓒ) gets a canonicalization hash number

Data-driven Aromaticity Perception

Training

Aromaticity

Mann & Costa

Does it smell?

Why to care?

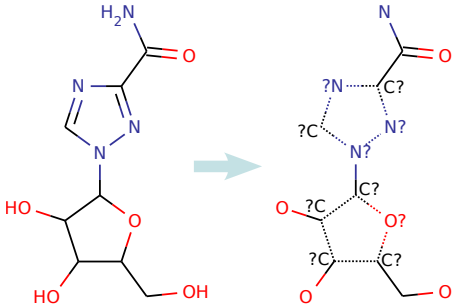
How to do?

Does it work?

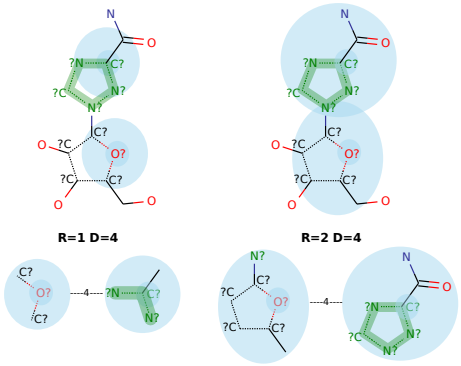
More to do?

A sketch

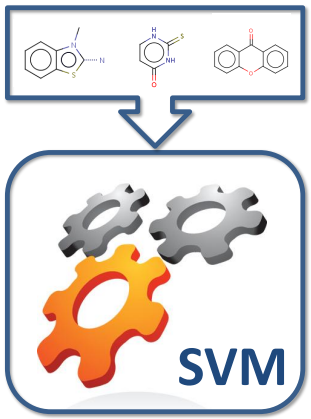
Take home



+



- no hydrogens
- no ring details



- one per ring per molecule (mark)
- NSPDK features
- localized on marked ring

Does it work?
Experimental setup

Aromaticity
Mann & Costa
Does it smell?
Why to care?
How to do?
Does it work?
More to do?
A sketch
Take home

Experimental setup:

1. extract molecules with rings from data base
2. for each molecule:
   - find all rings (Hanser, 1996)
   - for each ring (≤ 15) create features and annotate if aromatic or not

Result: set of "known instances"

Does it work?
Experimental setup

Aromaticity
Mann & Costa
Does it smell?
Why to care?
How to do?
Does it work?
More to do?
A sketch
Take home

Experimental setup:

1. extract molecules with rings from data base
2. for each molecule:
   - find all rings (Hanser, 1996)
   - for each ring (≤ 15) create features and annotate if aromatic or not

   Result: set of "known instances"

3. split data in training und test data
4. evaluate performance via iterations

Does it work?

Experimental setup: PubChem polycyclic

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

Does it work?

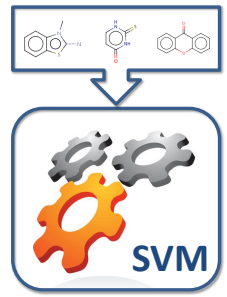
More to do?

A sketch

Take home

Data set:

- NCBI PubChem database
- $\sim 10,000$ *polycyclic* molecules
- ~ 8 rings per molecule



60% training

Training/Test set:

- split training/test $= 60/40\%$
- 10 iterations

40% testing

Does it work?

Experimental setup: PubChem polycyclic

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

Does it work?

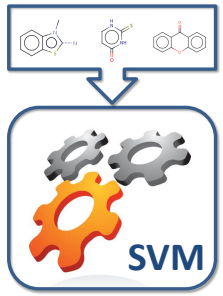
More to do?

A sketch

Take home

Data set:

- NCBI PubChem database
- $\sim 10,000$ *polycyclic* molecules
- ~ 8 rings per molecule



60% training

Training/Test set:

- split training/test $= 60/40\%$
- 10 iterations

40% testing

Result:

- accuracy 97.6% sd(0.2%)
- AUC ROC 0.995 sd(0.001), PR 0.978 sd(0.003)

Does it work?

Experimental setup: PubChem heterocyclic

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

Does it work?

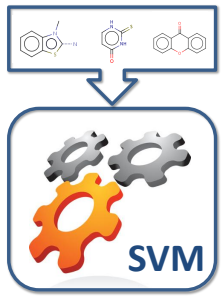
More to do?

A sketch

Take home

Data set:

- NCBI PubChem database
- $\sim 20,000$ *heterocyclic* molecules
- ~ 5 rings per molecule



60% training

Training/Test set:

- split training/test $= 60/40\%$
- 10 iterations

40% testing

Does it work?

Experimental setup: PubChem heterocyclic

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

Does it work?

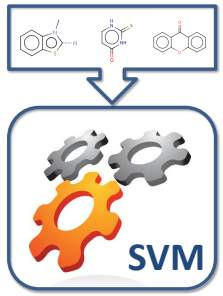
More to do?

A sketch

Take home

Data set:

- NCBI PubChem database
- $\sim 20,000$ *heterocyclic* molecules
- ~ 5 rings per molecule



60% training

Training/Test set:

- split training/test $= 60/40\%$
- 10 iterations

40% testing

Result:

- accuracy 94.5% sd(0.2%)
- AUC ROC 0.985 sd(0.001), PR 0.982 sd(0.001)

Does it work?

Think so . . .

Aromaticity
Mann & Costa
Does it smell?
Why to care?
How to do?
Does it work?
More to do?
A sketch
Take home

High accuracy prediction:

- minimal informed input (no hydrogens, no ring details)
- prediction with high accuracy ($\geq 95\%$)

Does it work?

Think so . . .

Aromaticity
Mann & Costa
Does it smell?
Why to care?
How to do?
Does it work?
More to do?
A sketch
Take home

High accuracy prediction:

- minimal informed input (no hydrogens, no ring details)
- prediction with high accuracy ($\geq 95\%$)

BUT:

- PubChem annotation based on Daylight model
  $\rightarrow$ reproducing OpenEye tool results
- Results for ChEBI database similar (uses Marvin)
- No manually annotated data available . . .
  $\rightarrow$ if available: **Framework is ready!**

Take home messages . . .

Aromaticity
Mann & Costa
Does it smell?
Why to care?
How to do?
Does it work?
More to do?
A sketch
Take home

- Aromaticity is a sloppy defined property
- More continous than YES/NO

Take home messages . . .

Aromaticity

Mann & Costa

Does it smell?

Why to care?

How to do?

Does it work?

More to do?

A sketch

Take home

- Aromaticity is a sloppy defined property
- More continous than YES/NO

- SVM + NSPDK graph kernel enables classification
- Open to regression task
- No good data available (so far)