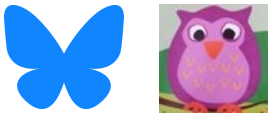
Explainable AI: Understanding, Prediction, and Trust

Jan H. Jensen
Department of Chemistry,
University of Copenhagen



@janhjensen

What's this a picture of?

A. Horse

B. Car

C. Rock

What's this a picture of?

A. Horse

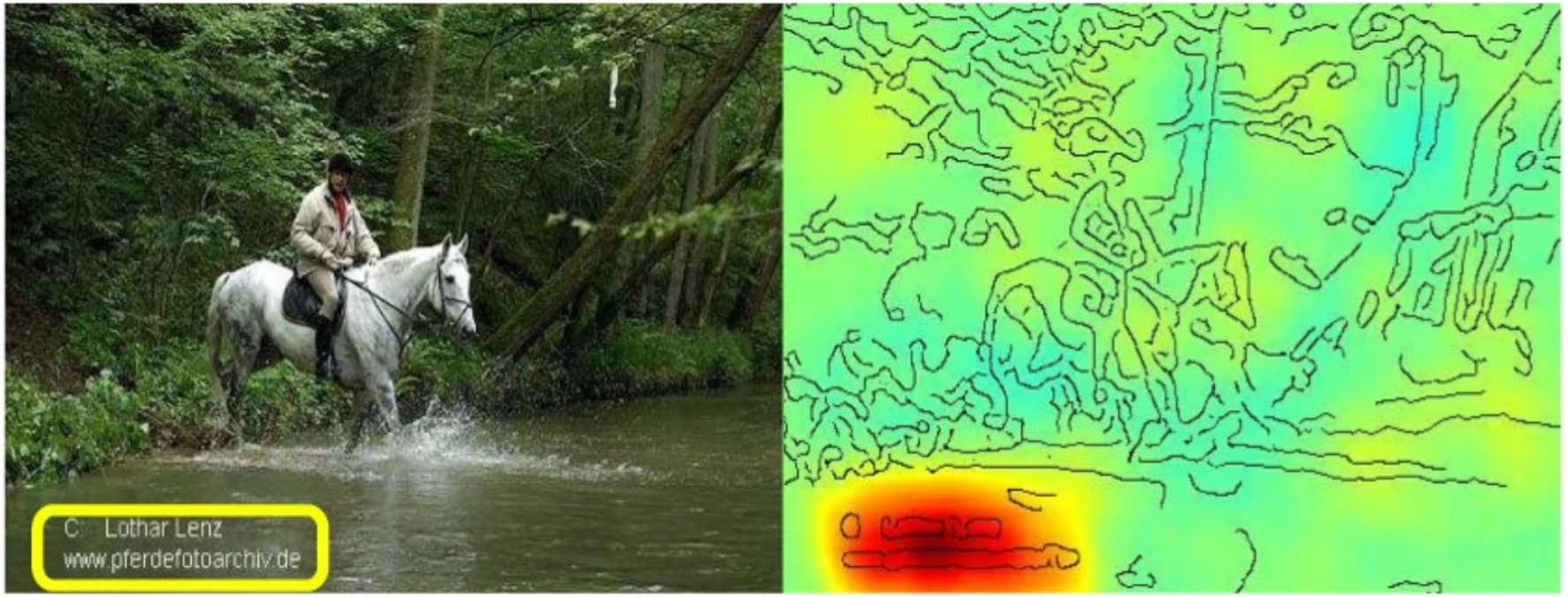
B. Car

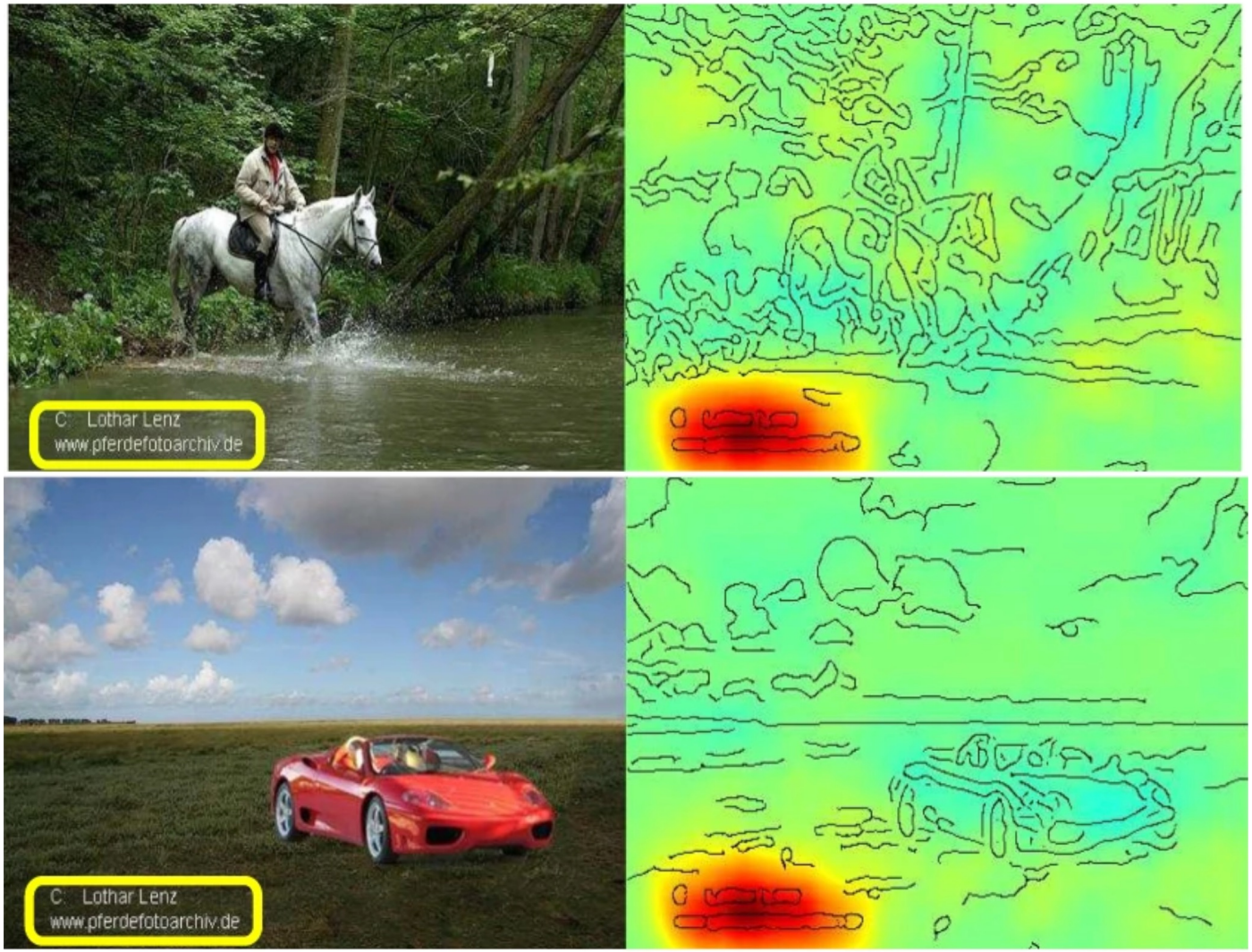
C. Rock

Heatmaps: which pixel contributes most to the decision?

Unmasking Clever Hans predictors and assessing what machines really learn

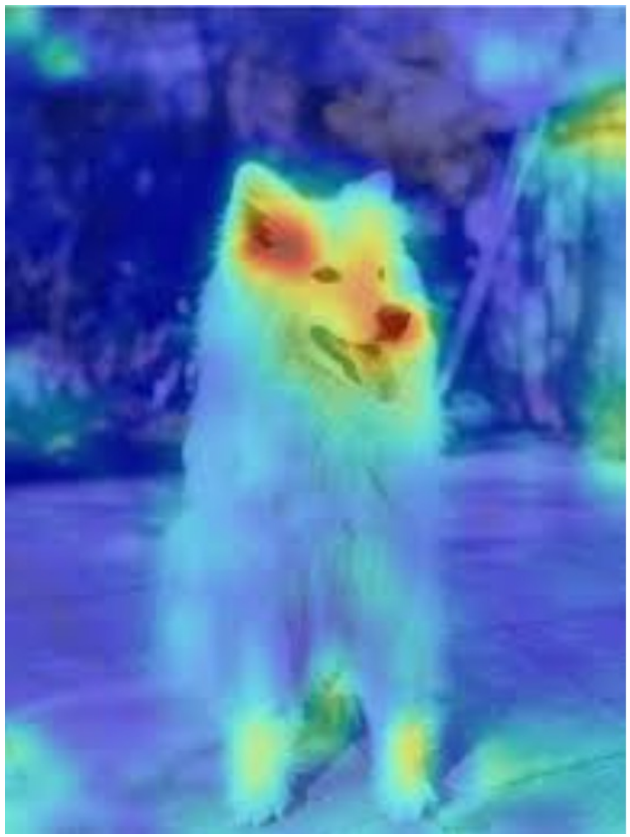
Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek ✉ & Klaus-Robert Müller ✉



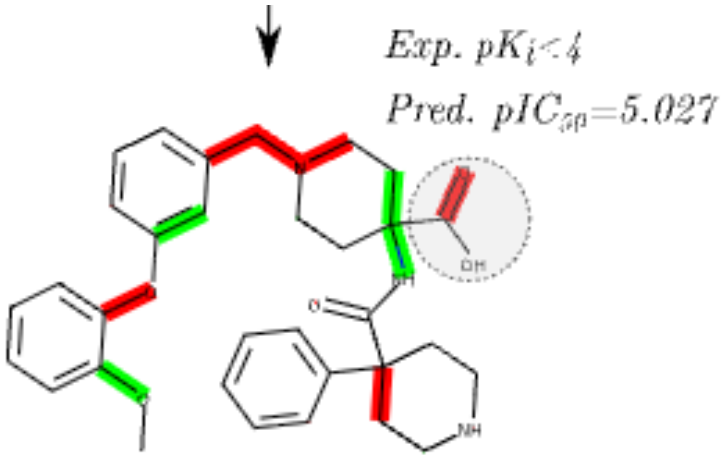

© Lothar Lenz
www.pferdefotoarchiv.de
© Lothar Lenz
www.pferdefotoarchiv.de

Clever Hans

How accurate are molecular heatmaps?



hERG inhibition



Negative contributions

Positive contributions

A ground truth for heatmaps

Prediction of Physicochemical Parameters by Atomic Contributions

Scott A. Wildman and Gordon M. Crippen*

```
from rdkit.Chem.Crippen import MolLogP

print("LogP",MolLogP(sorafenib))
```

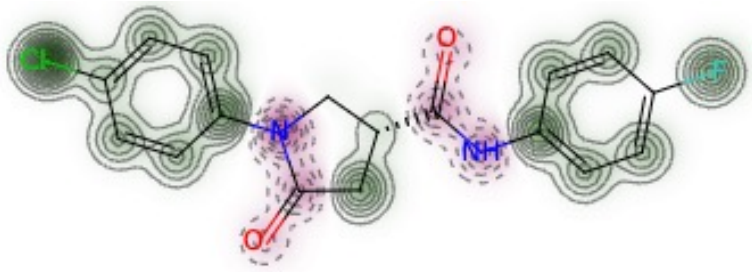
$$\log P = \sum_i n_i a_i$$

Atom types

110 different atom types

n_i = number of atoms of type *i*
a_i = logP contribution of atom type *i*

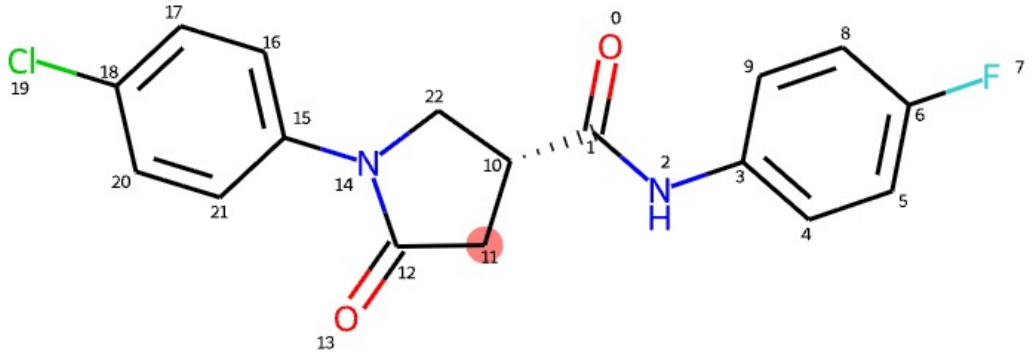
```
#ID SMARTS  "log P" MR  Notes/Questions
C1  [CH4]   0.1441  2.503
C1  [CH3]C  0.1441  2.503
C1  [CH2](C)C   0.1441  2.503
C2  [CH](C)(C)C 0   2.433
C2  [C](C)(C)(C)C   0   2.433
C3  [CH3][N,O,P,S,F,Cl,Br,I]    -0.2035 2.753
C3  [CH2X4]([N,O,P,S,F,Cl,Br,I])[A;!#1] -0.2035 2.753
C4  [CH1X4]([N,O,P,S,F,Cl,Br,I])([A;!#1])[A;!#1]    -0.2051 2.731
C4  [CH0X4]([N,O,P,S,F,Cl,Br,I])([A;!#1])([A;!#1])[A;!#1]   -0.2051 2.731
```



Creating and comparing heatmaps

Step 1: Train Random Forest models with ECFP4 (2048 bit vector) with clogP as ground truth
Step 2: Compute atom contributions for each atom for a particular molecule
Step 3: Compare to clogP atom contributions (normalized dot product and Pearson's *r*)

Training set: 150K
Test set: 5K
Test MSE: 0.58



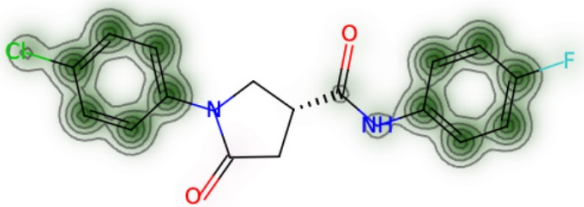
Fragments involving atom 11

Step 2: logP contribution of atom 11 = logP(full FP) – logP(FP with these 10 bits turned off)

Highest overlap case: still significant differences

ML

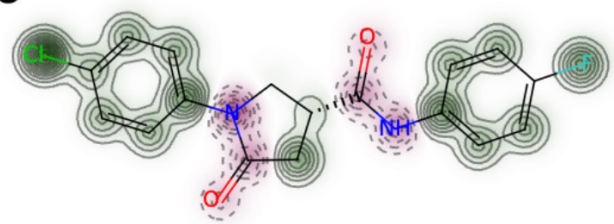
a



error = 0.32
ClogP = 3.47
$\log P_{pred}$ = 3.79±0.64

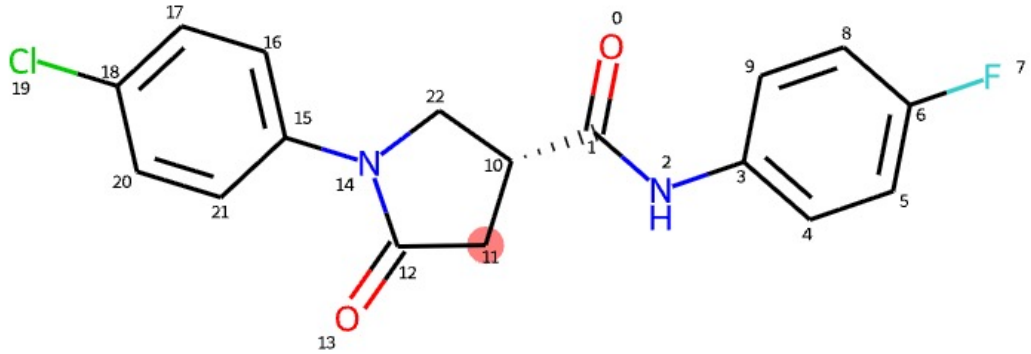
atomic

b



overlap = 0.66
Pearson's r = 0.54

The reason: you are not removing just 1 atom



Fragments involving atom 11

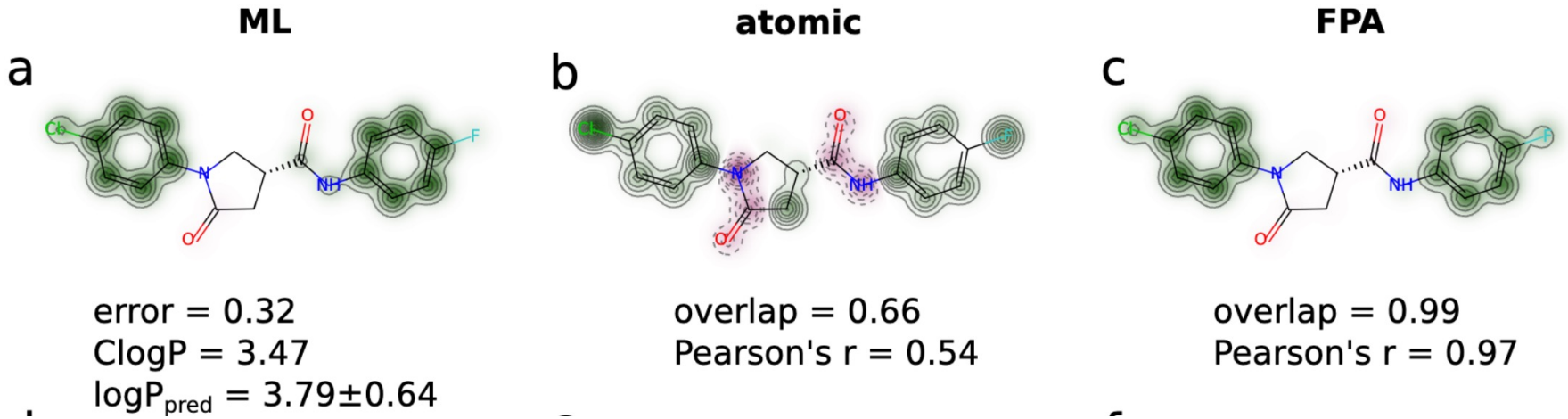
94 231 785 798 926 974 1137 1598 1642 1972

logP contribution of atom 11 = logP(full FP) – logP(FP with these 10 bits turned off)

Finger print adapated ground truth for atom 11 = $\sum \text{clogP}_{\text{fragment}}$

The explanation will depend on the molecular representation

The model “thinks” in terms of fragments instead of atoms



The clogP value is high mainly because of the phenyl rings

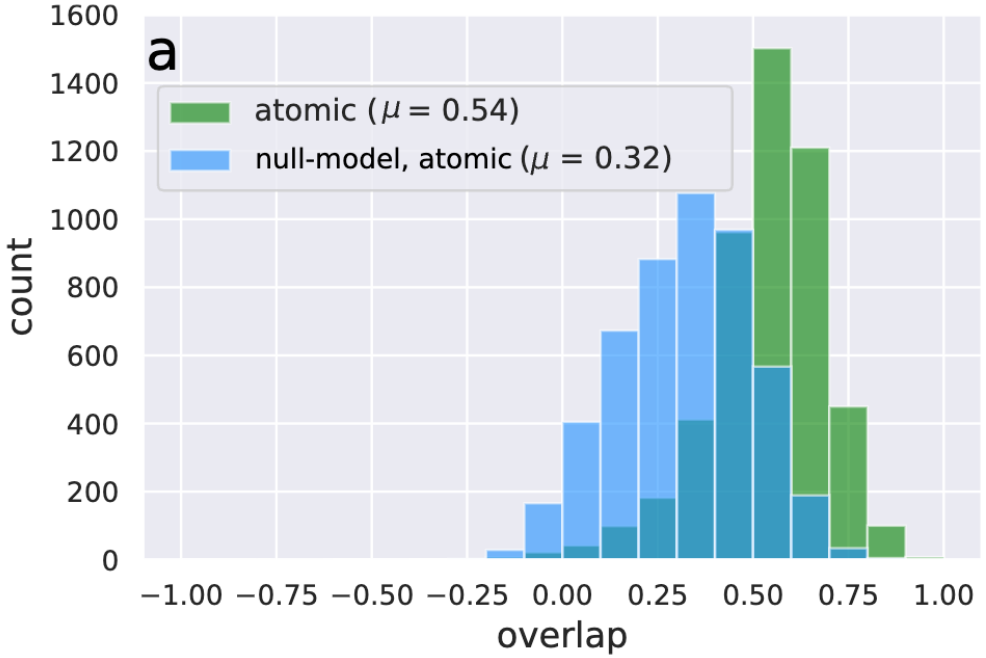
~~Each atom in the central region contributes little to the logP~~

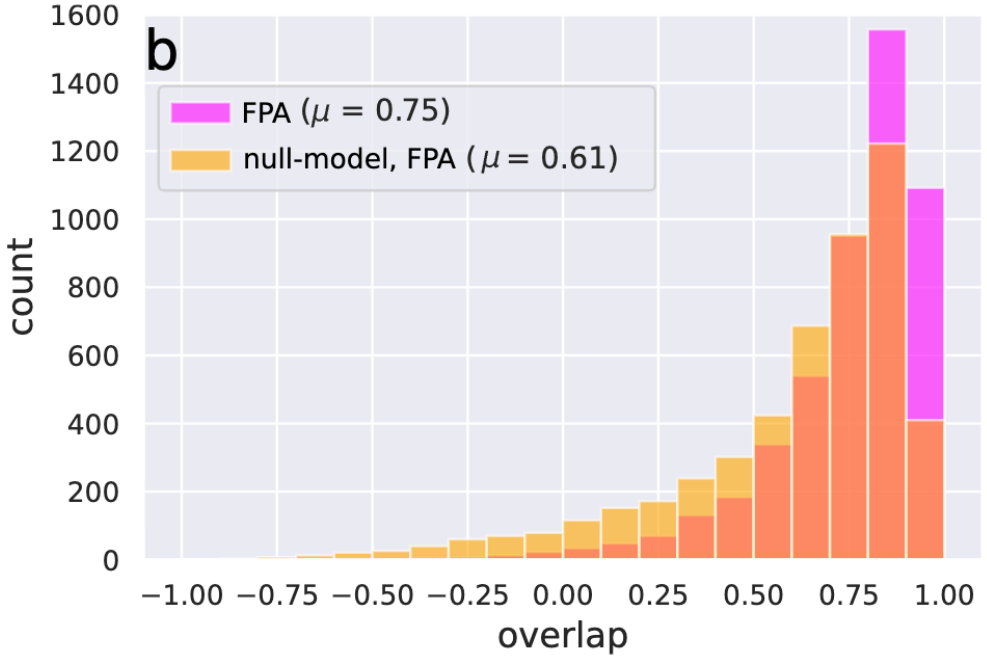
The *net* contribution of the central region to the logP is very small

$$\log P = \sum_{i}^{\text{Atom types}} n_i a_i$$

a_i’s cannot be extracted from this model analysis

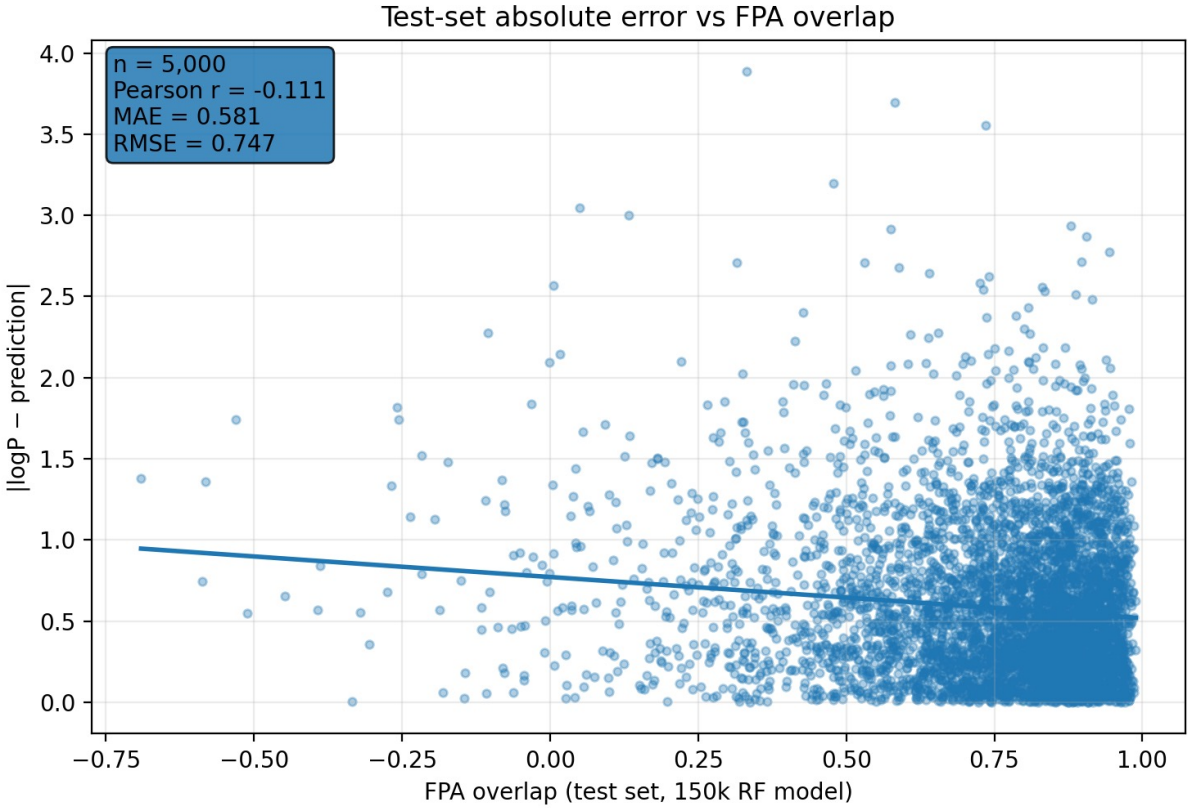
Most heatmaps are accurate (reflect the ground truth)

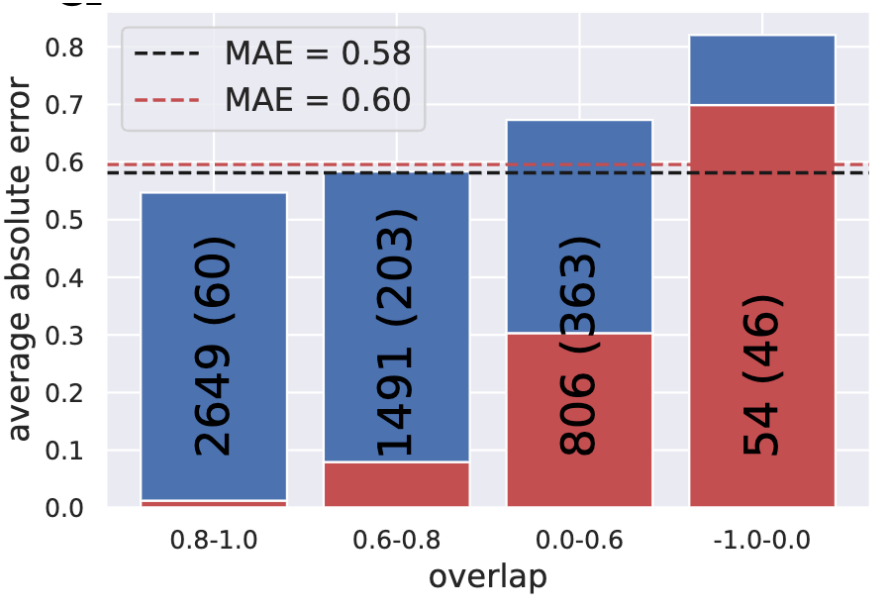




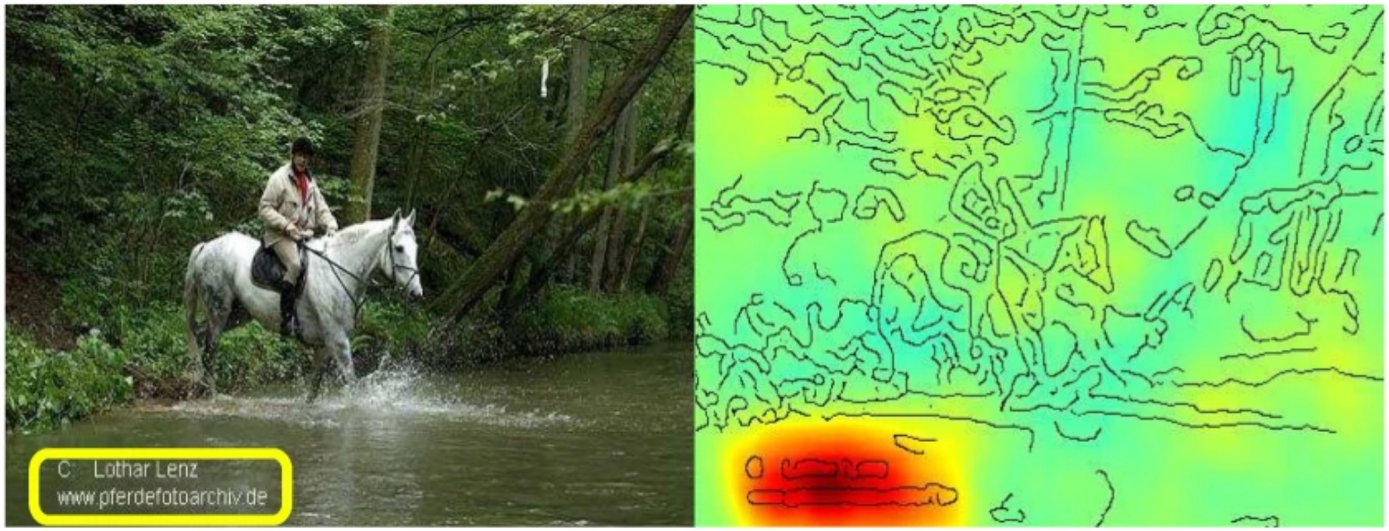
Null model: a_i = clogP/N_{atoms}

No correlation between prediction error and heatmap error





Right prediction, wrong heatmap



A. Horse

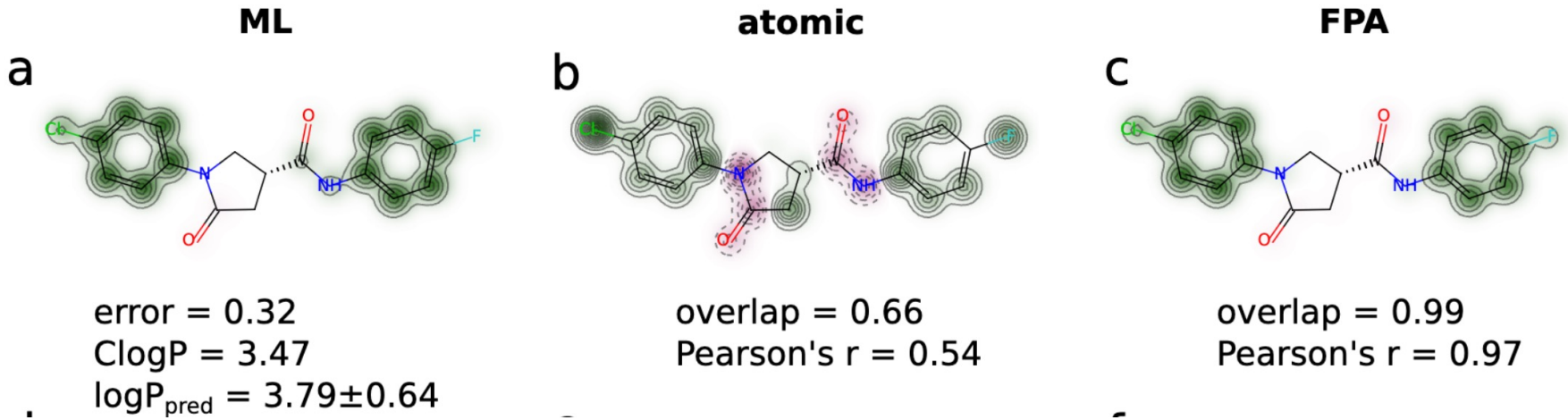
B. Car

C. Rock

Again, this kind of usage assumes that you know what the heatmap is supposed to look like

You can't necessarily learn the chemistry by looking at heatmaps for accurate predictions

Right prediction, right conclusions?

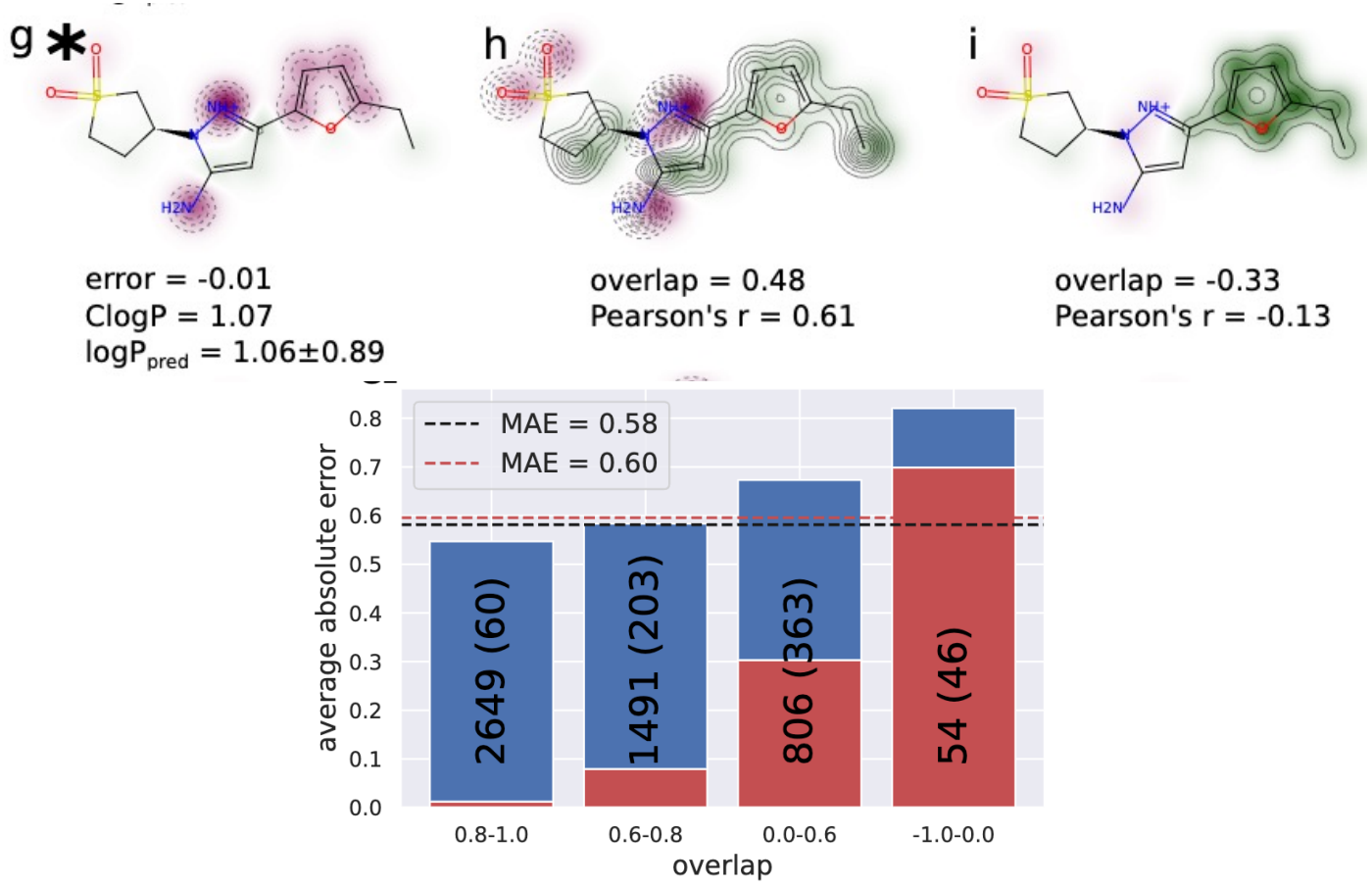


The clogP value is high mainly because of the phenyl rings

The *net* contribution of the central region to the logP is very small

Obviously wrong heatmaps

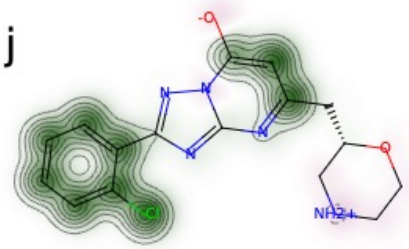
Sign of prediction and sum of “atomic” contributions have opposite signs



Is overlap always a good metric?

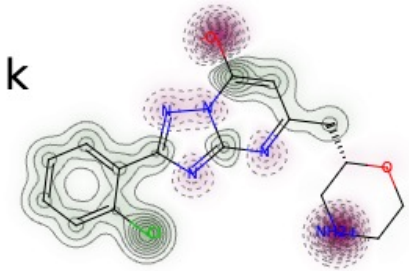
High error, OK overlap

j



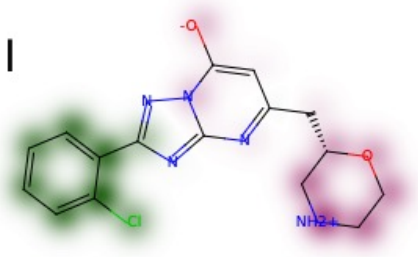
error = 3.56
ClogP = 0.02
$\log P_{pred}$ = 3.58±0.89

k



overlap = 0.49
Pearson's r = 0.63

l



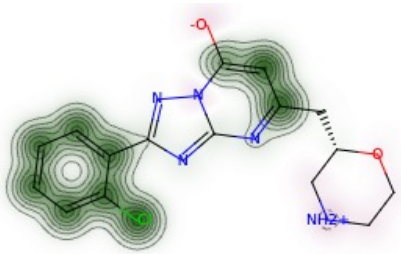
overlap = 0.74
Pearson's r = 0.81

Overlap and *r* emphasises signs, not magnitudes

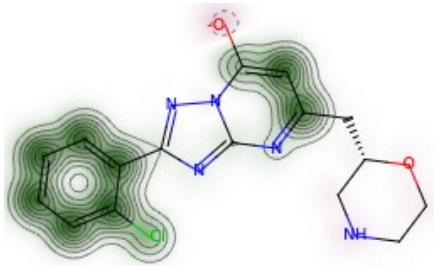
Perhaps RMSE or MAE as an additional metric?

Heatmaps (and xAI in general) must lead to actionable predictions

Example: error analysis

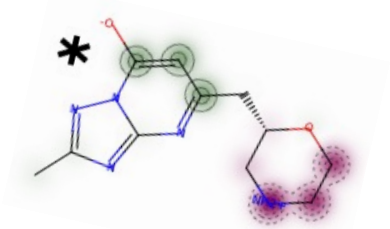


error = 3.56
ClogP = 0.02
$logP_{pred}$ = 3.58±0.89



ΔHeatmap: ~no change
ΔML: +0.1
ΔGround truth: +1.0

✗

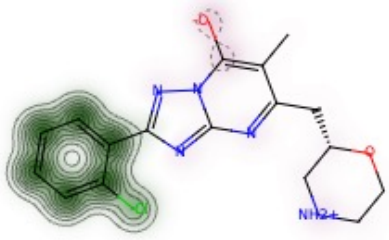


Non-local & non-additive effect

ΔHeatmap: large decrease
ΔML: -3.3
ΔGround truth: -2.0

✓

Overfitting for ionic groups?`



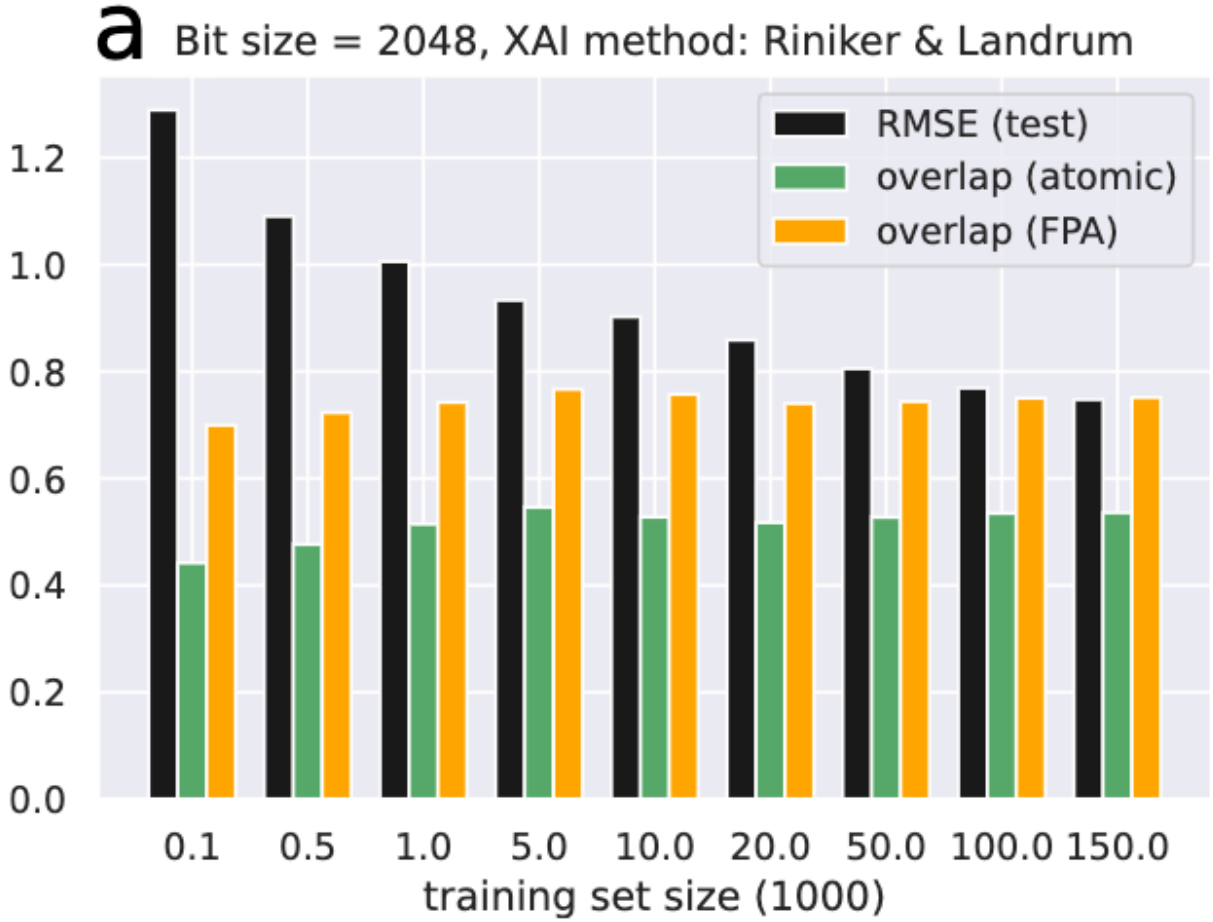
Non-additive effect

ΔHeatmap: increase?
ΔML: -1.2
ΔGround truth: +0.3

✗

Sensitivity to training set size

Converged at ~1000 molecules

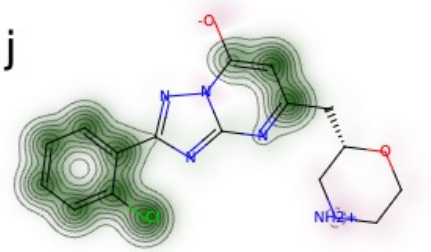


Sensitivity to training set size

heatmaps generally insentive

Training set size

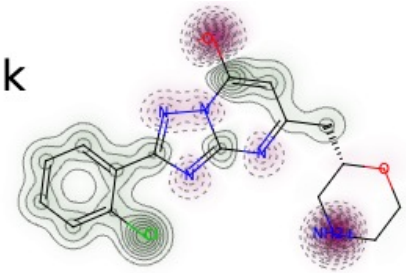
j



150,000

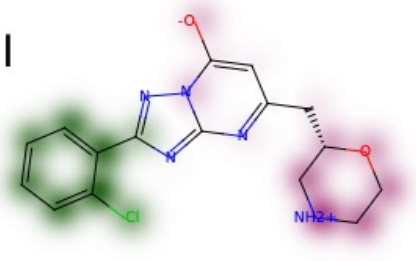
error = 3.56
ClogP = 0.02
$\log P_{pred}$ = 3.58±0.89

k



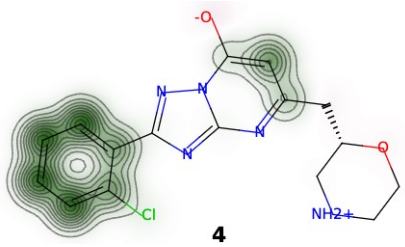
overlap = 0.49
Pearson's r = 0.63

l



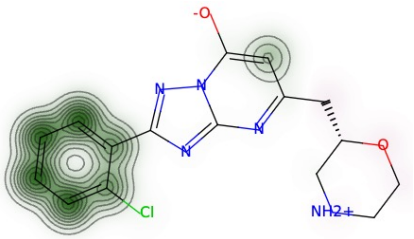
overlap = 0.74
Pearson's r = 0.81

1000



Error = 3.4

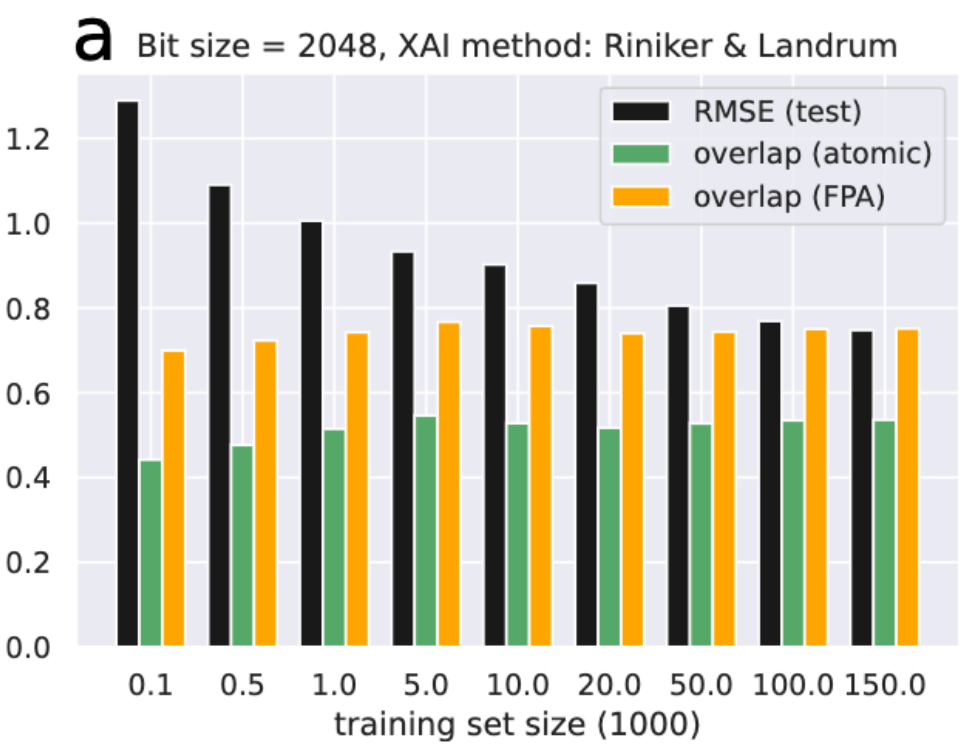
100

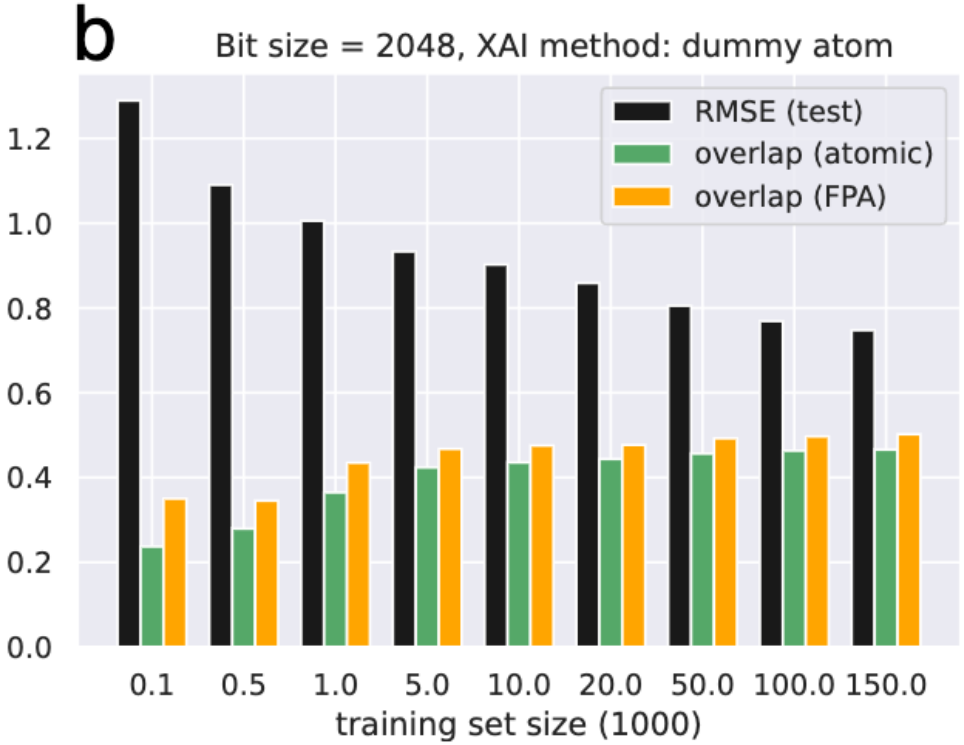


Error = 2.7

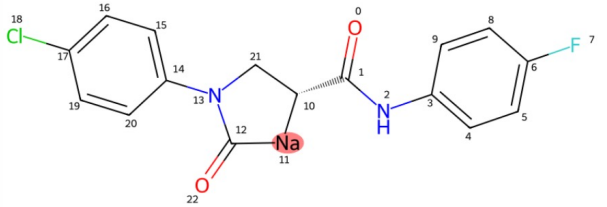
Comparing xAI methods

Dummy atom approach is a bit worse



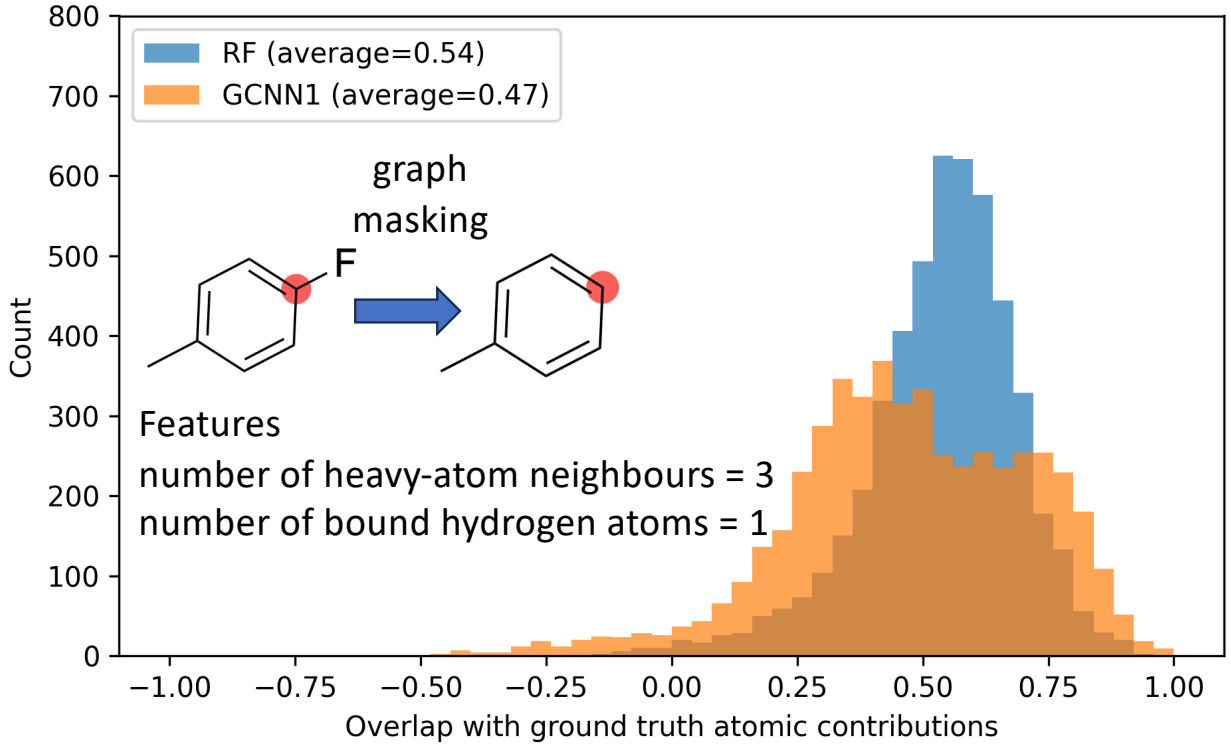


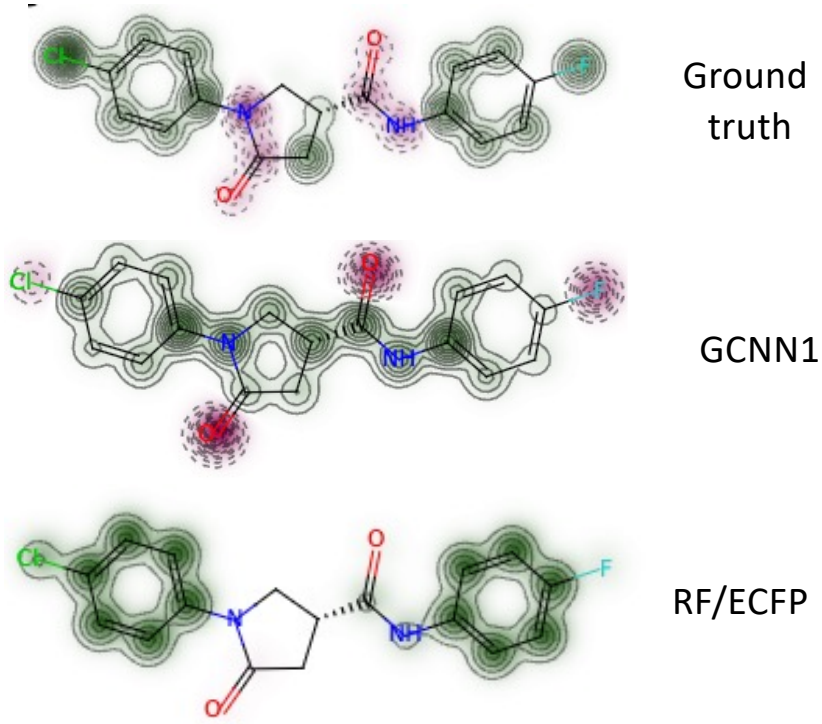
Dummy atoms introduces new FP fragments



Comparing xAI methods

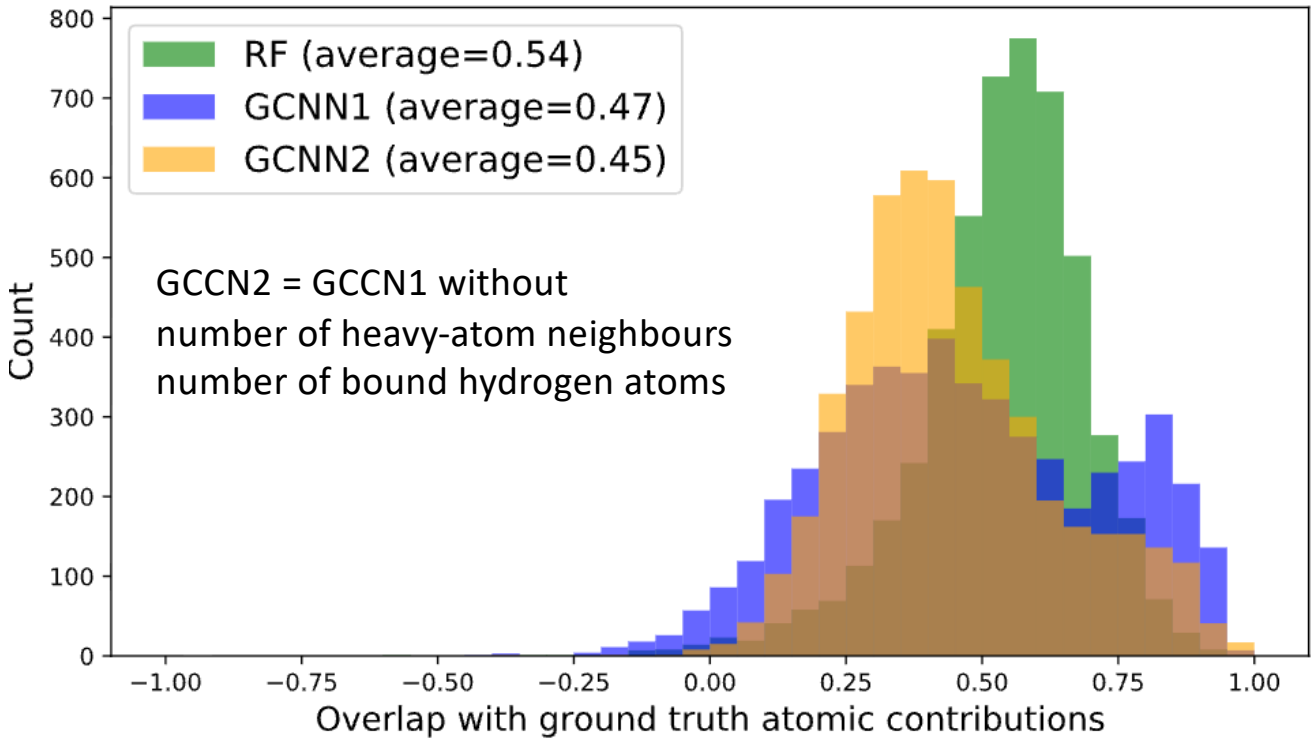
Comparison to graph convolution NNs

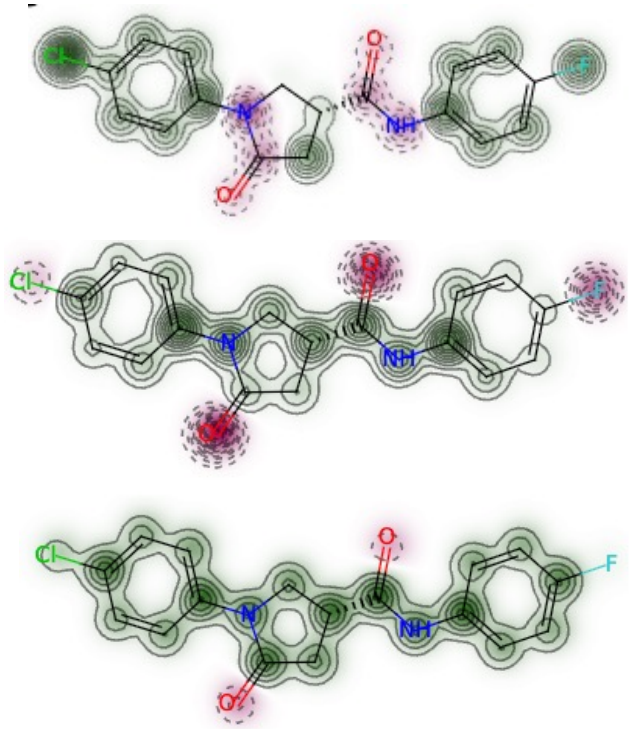




Comparing xAI methods

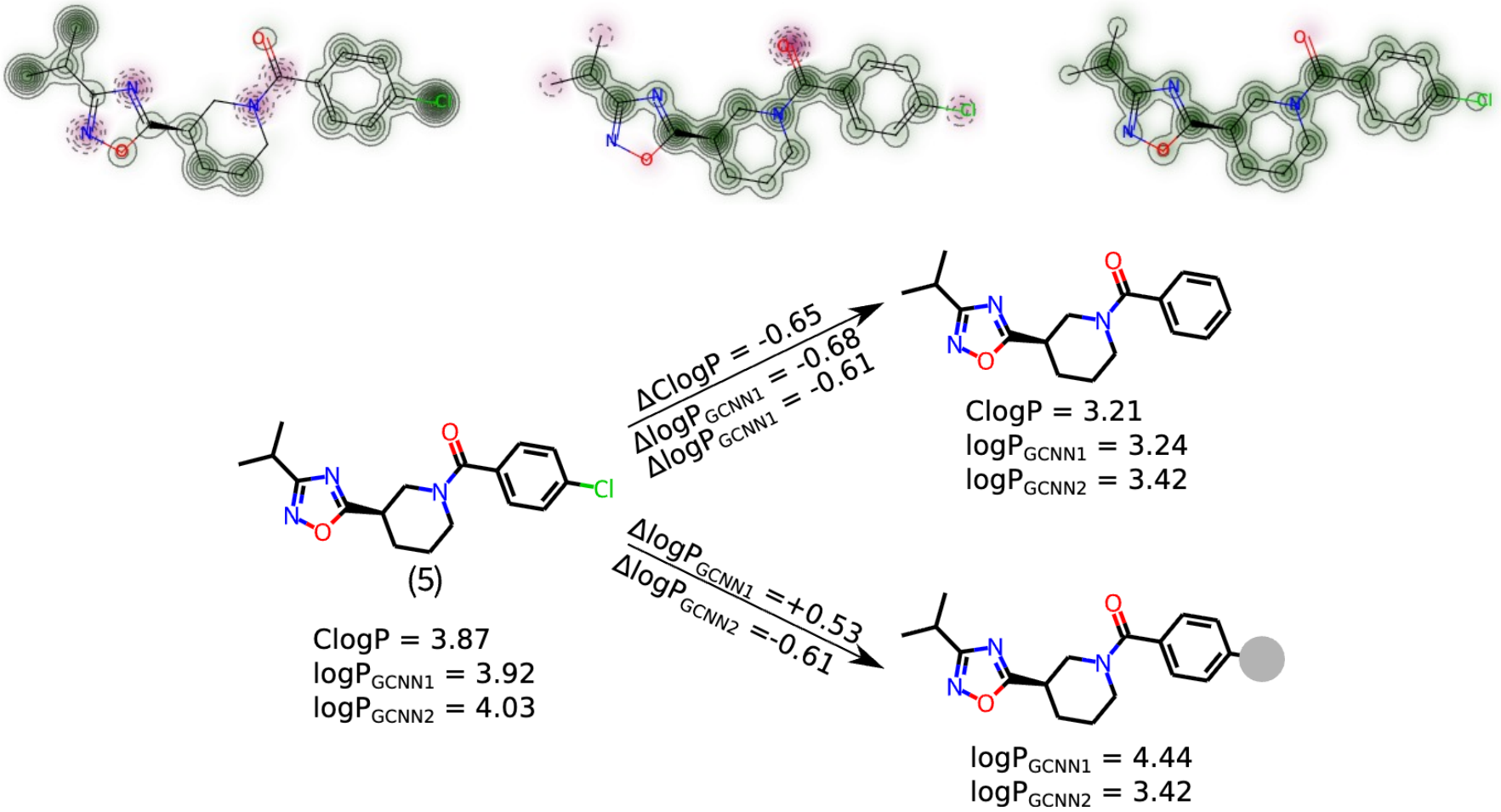
Comparison to graph convolution NNs





Comparing xAI methods

Comparison to graph convolution NNs



Summary

Correct predictions do **not** guarantee correct explanations
(right answer for the wrong reason)

Trust heatmaps only after validation (testing predictions),
not just because the prediction is accurate

The xAI explanation will depend on the molecular representation

A good xAI model should be …
Actionable (predictive)
Understandable (consistent)

Good tool for understanding a model
Questionable tool for understanding chemistry