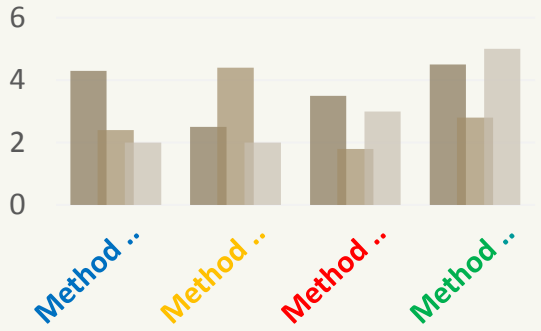
Way2Drug PREDICTIVE SERVICES
Understanding Chemical-Biological Interactions

XXVII Symposium on Bioinformatics and Computer-Aided Drug Discovery

Benchmarks for interpretation of QSAR models

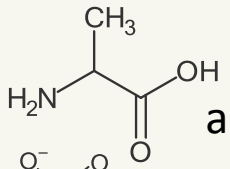


Maria Matveieva, *Pavel Polishchuk*

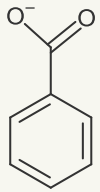
Palacky University in Olomouc, Czech Republic

Interpretation of models (structural)

dataset

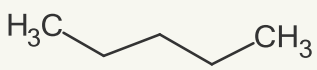


active



inactive

…

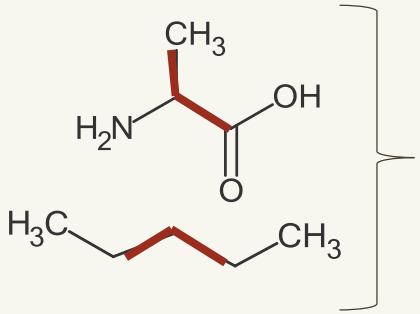


active

Modelling

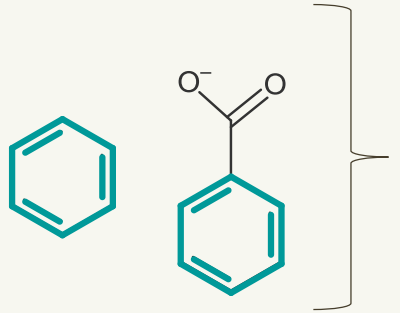
(Calculate descriptors)

PATTERN/REGULARITY:



active

…



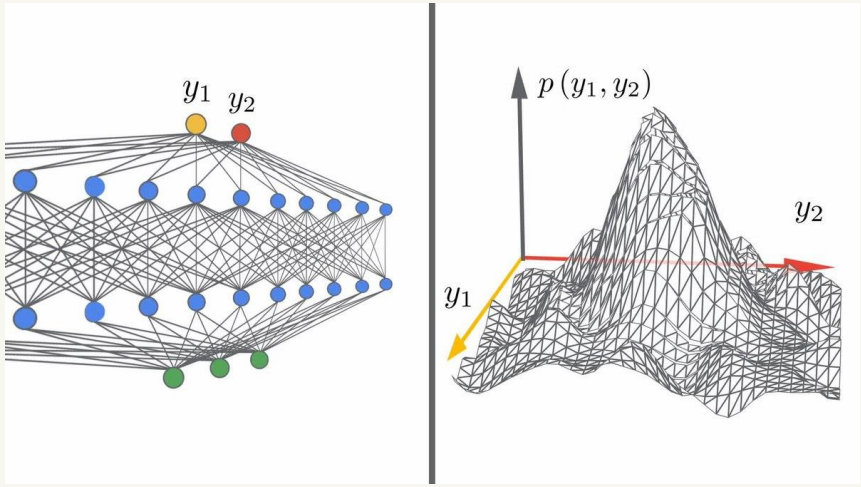
inactive

Why to interpret?

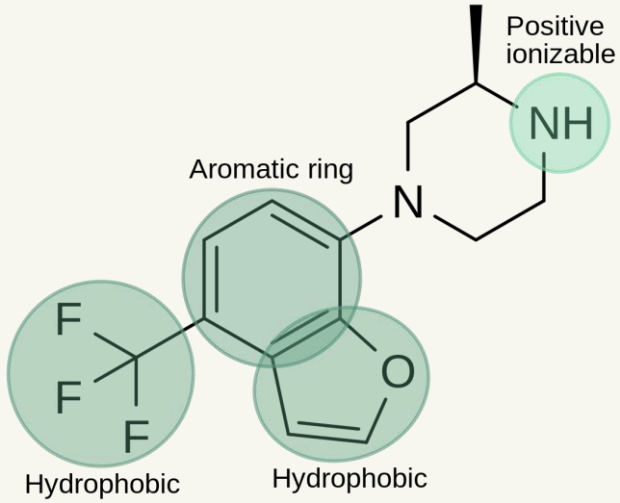
- Knowledge-based validation of the model

“High accuracy = trustworthy?”

“is model right for right reasons?”



- Find *useful features* → structure optimization etc.

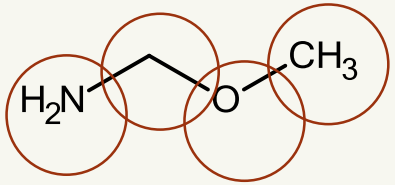


How to interpret?

Zoo of approaches…

Gradient-based: $\frac{\delta\ model}{\delta(x)}$

CAM
Grad-CAM
*Gradient*Input*
…

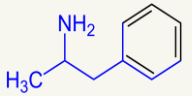


Surrogate methods:
F ≈ ∑ features
LIME (Ribeiro et al. 16)
….

Layer-wise relevance propagation (Bach et al. 15)

By design interpretable methods:
Attention-based neural nets

Subgraph identification (Ying et al. 19)



Perturbation-based:
SPCI (Polishchuk et al. 13)
Similarity maps (Riniker et al. 13)
Feature Importance by permutation (Breiman 01)
…

Integrated Gradient (Sundararajan et al. 17)

Validation: current state

Use “classical“ datasets: solubility, Ames mutagenicity… → Annotated data of different complexity needed

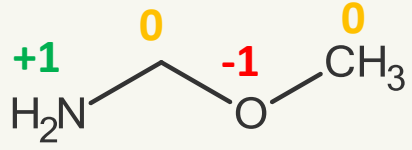
No commonly accepted metrics should reflect methods validity

No systematic comparison ods to trust? date

Which method to choose?
Benchmarking needed!

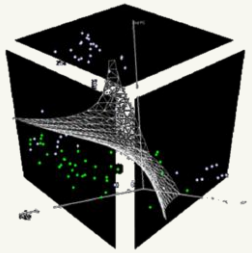
Key idea: synthetic data

Activity defined by labels

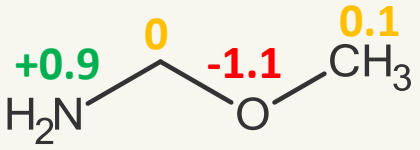


predefined atom labels

Create model



interpret model



calculated contributions

Compare:
1 – 0.9
-1 – (-1.1)
0 – 0.1
…

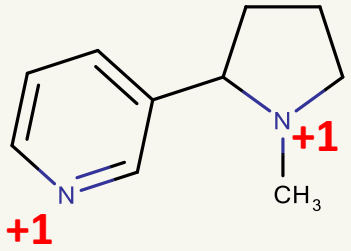
Aims

- Dataset development:
  - simple → complex
  - + metrics development
- Pilot study of applicability of datasets
- Study of different models and descriptors:
  - influence on interpretation quality

DATASET DEVELOPMENT

N_count:
all N atoms = +1,
other atoms = 0

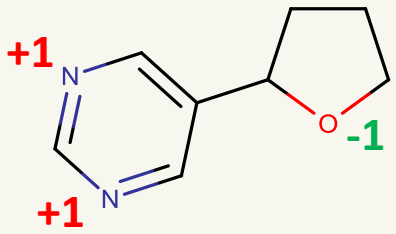
activity=2



Simple additive property

N - O:
all N atoms = +1,
all O atoms = -1,
other atoms = 0

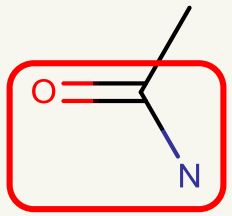
activity=1



Simple additive property
With negative pattern

Amide_count:
all amide groups = +1,
other atoms = 0

activity=1

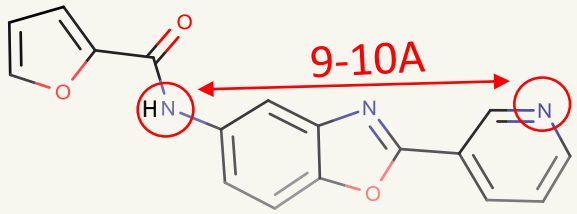


Realistic property, example:
lipophilicity

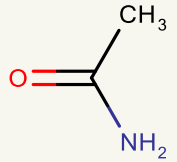
Amide_class:
molecules with
at least 1 amide groups = active

Pharmacophore:
all pharm. centers = +1,
other atoms=0

activity=1



activity=0



Realistic property,
ex.: ligand-receptor interaction

10.000 molecules with different activity
randomly sampled from ChEMBL

Descriptors & models

- Morgan fingerprints (r=2)
- RDKIT fingerprints
- Atom Pairs fingerprints
- Topological torsion fingerprints

×

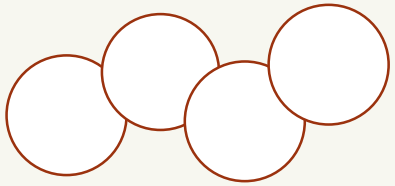
- Random Forest
- Support Vector Machines
- Gradient Boosting
- Partial Least squares

- Graph convolutional NN

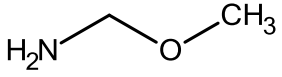
Universal interpretation approach

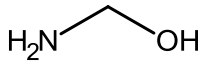
(implemented in SPCI)

Gradient-based:



CAM
Grad-CAM
*Gradient*Input*
...

F(

) - F(

) = Contribution(C)

Surrogate methods:
LIME (Ribeiro et al. 16)
....

Layer-wise relevance propagation (Bach et al. 15)

By design interpretable methods:
Attention-based neural nets

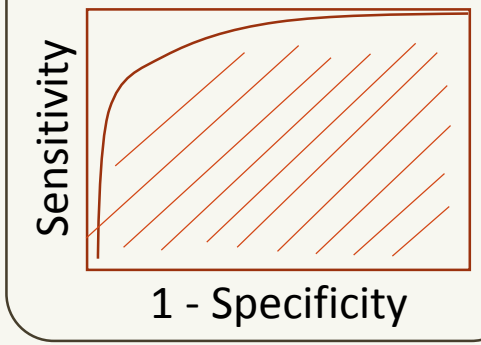
*Subgraph identification (*Ying *et al. 19)*

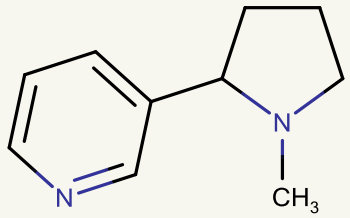
Perturbation-based:
SPCI (Polishchuk et al. 13)
Similarity maps (Riniker et al. 13)
Feature Importance by permutation (Breiman 01)
...

Integrated Gradient (Sundararajan et al. 17)

Interpretation quality metrics

- **ROC-AUC**
- **[0…1]**





atom	obs	pred
N	+1	+0.9
N	+1	+0.8
C	0	0.1
C	0	0
…	…	…

- **Top-n score: fraction of true atoms in top *n* atoms**
- **[0…1]**

top-n = ∑m / ∑n

Sum over dataset

n – number of true atoms in molecule

m - number of true atoms in top n atoms

- **RMSE**
- **[0…Infinity)**

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i} (y_{i,\mathrm{pred}} - y_{i,\mathrm{obs}})^2}{N}}$$

RMSE = (0.1^2 + 0.2^2 + 0.1^2 + 0^2 + …) / 12

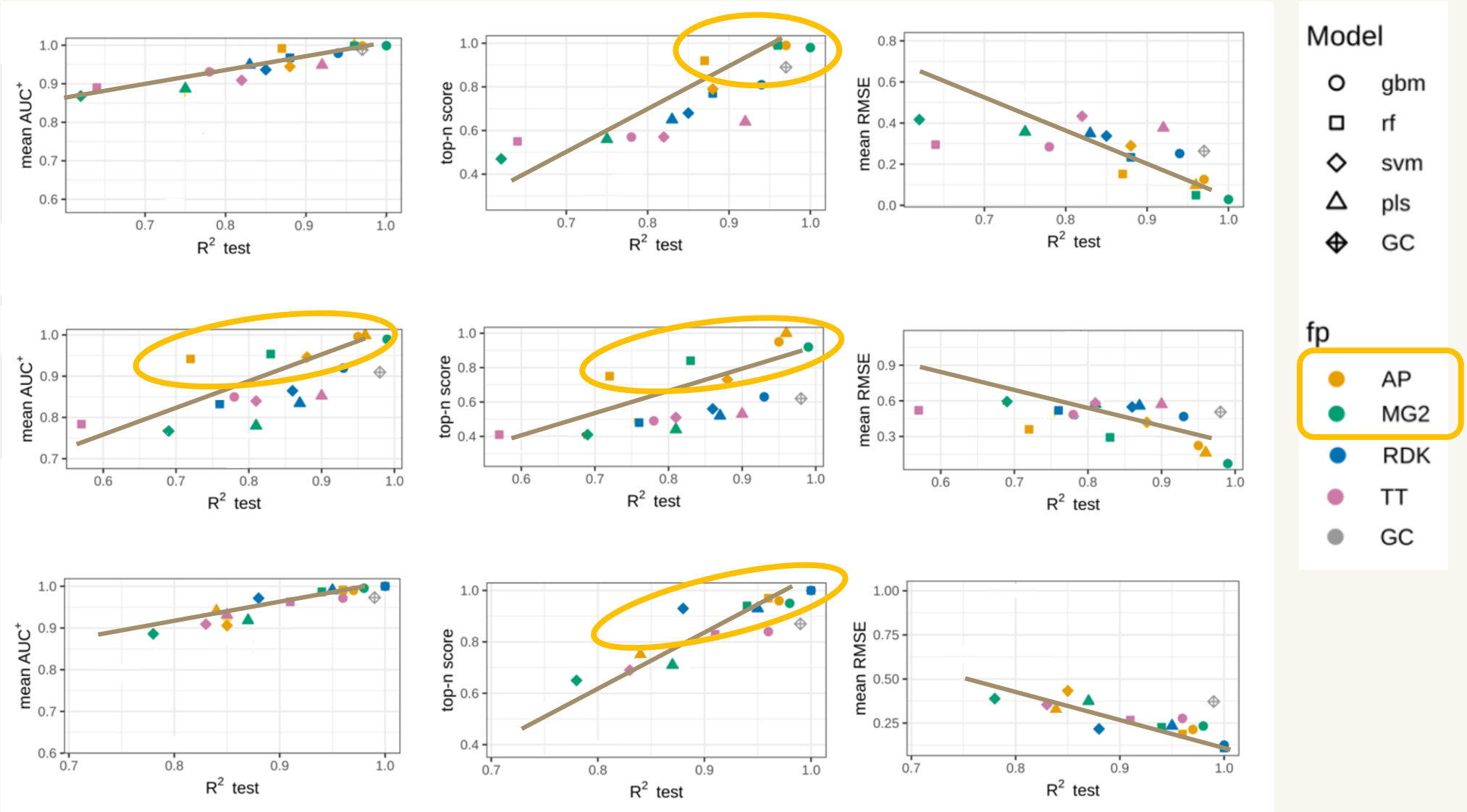
RESULTS OF INTERPRETATION

R^2 vs interpretation performance

N_count

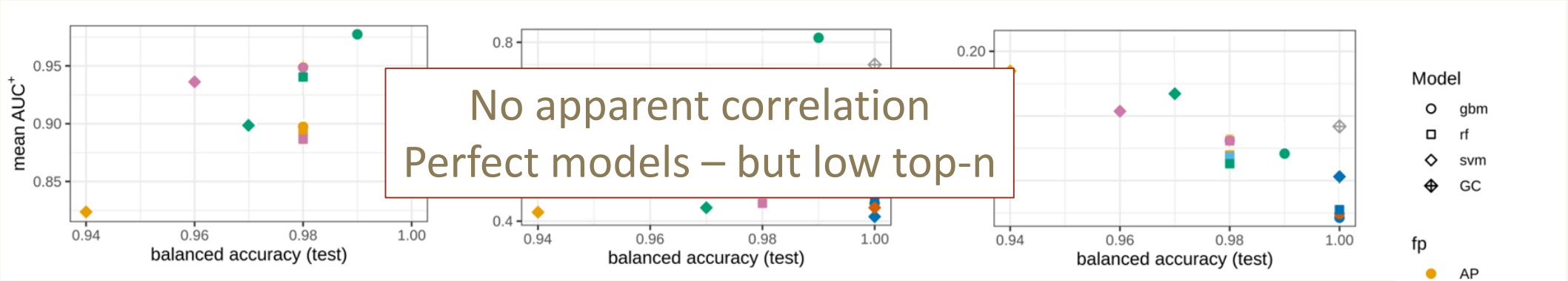
N - O

Amide_count

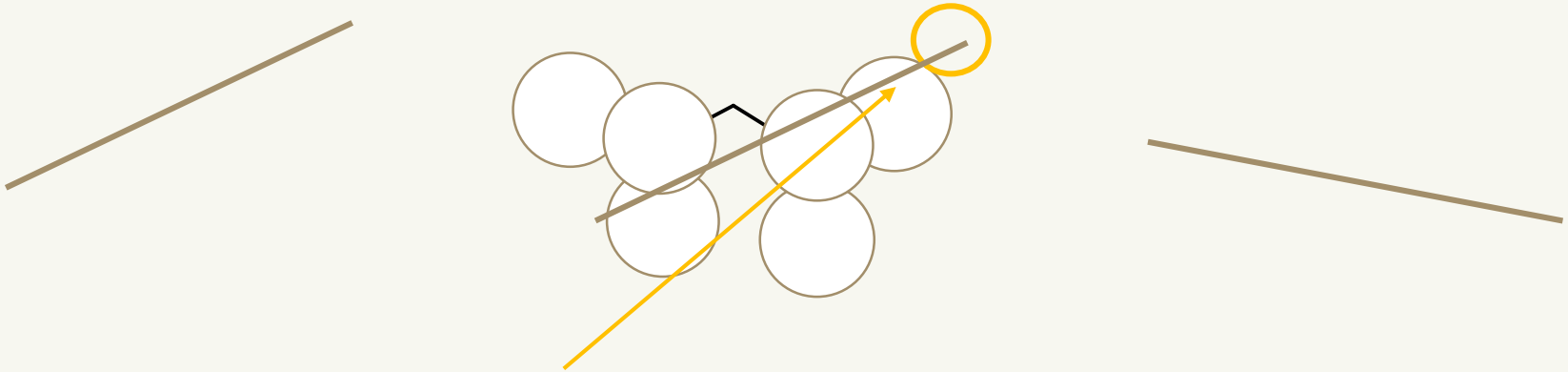


Accuracy vs interpretation performance

Amide_class

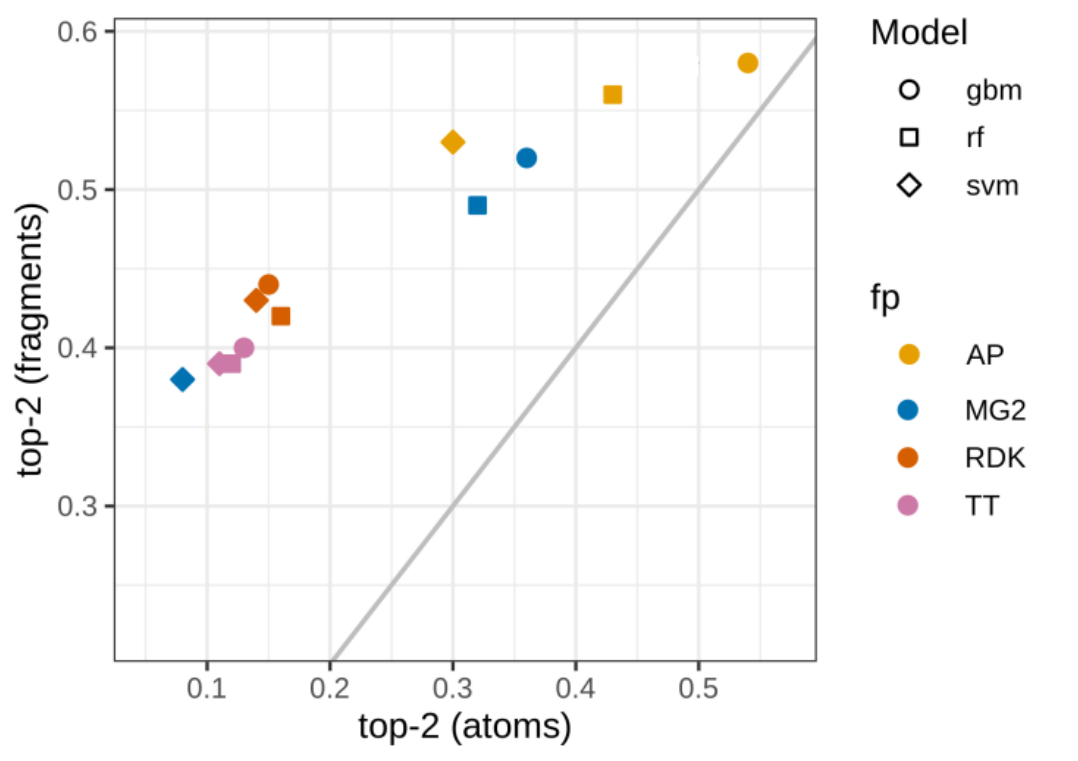


No apparent correlation
Perfect models – but low top-n

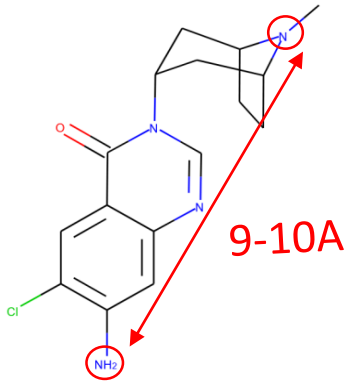


Values are quite low

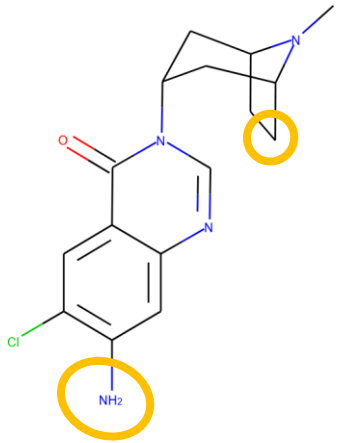
Fragment-based interpretation (pharmacophore)



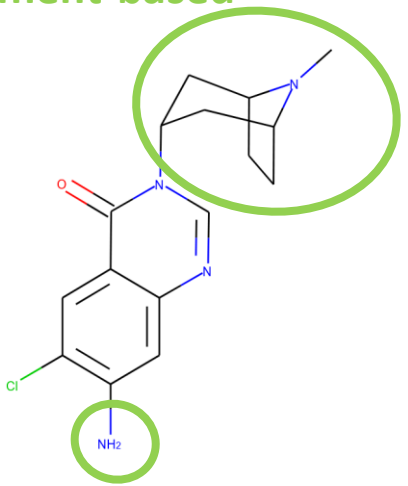
Ground truth



Atom-based

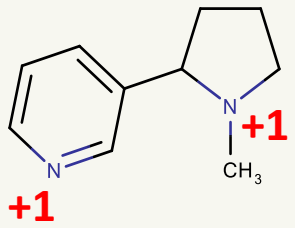


Fragment-based

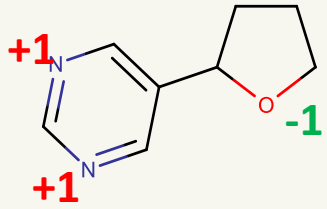


Summary

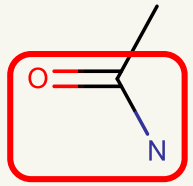
N_count



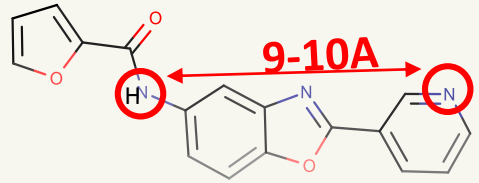
N - O



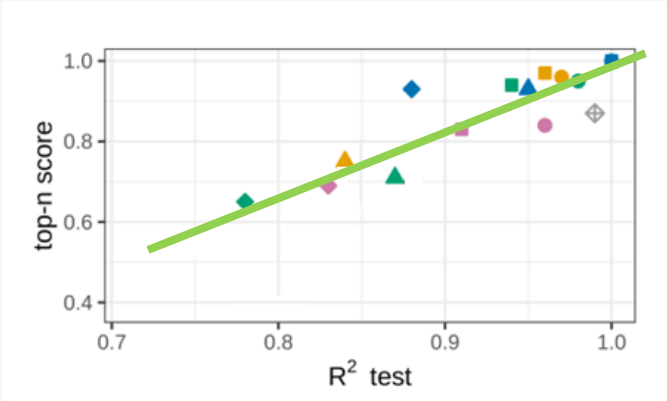
Amide count + Amide classification



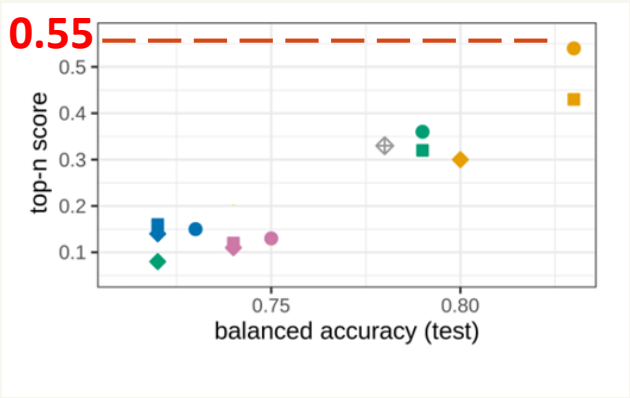
Pharmacophore



Interpretation performance:

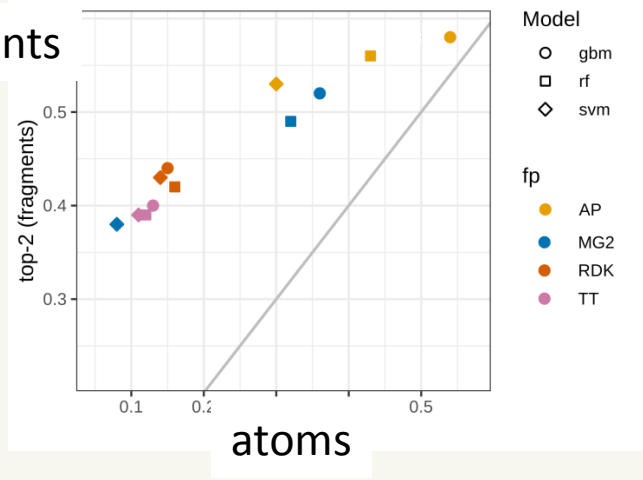


Models performance correlates with interpretation performance



High accuracy models can produce wrong interpretation

fragments



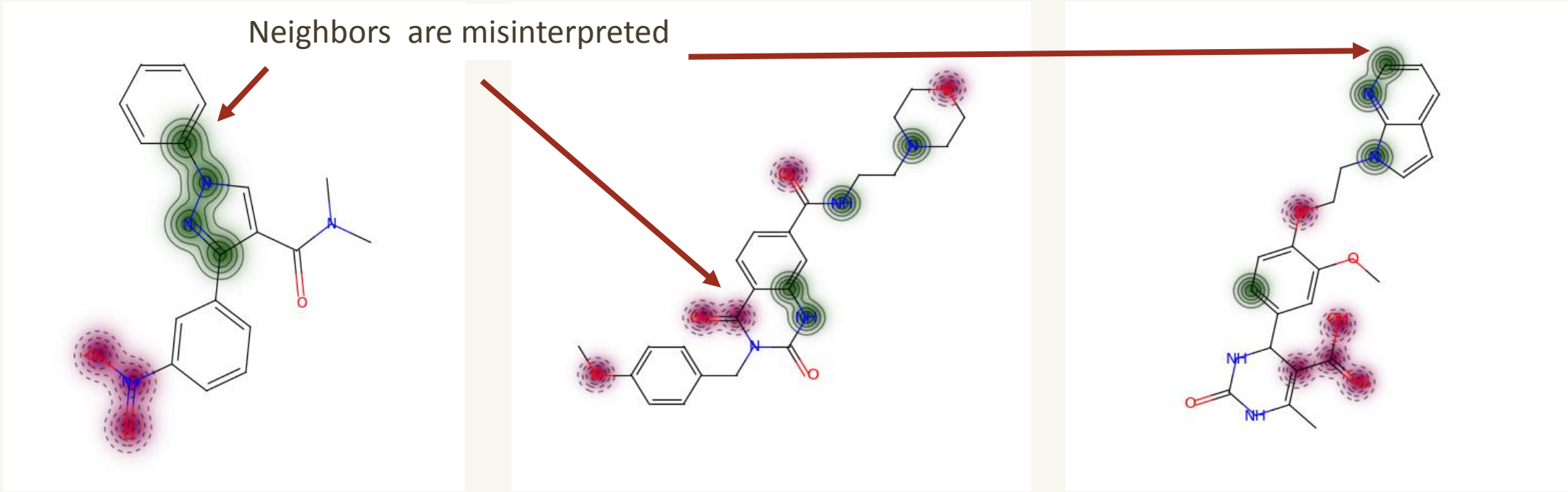
atoms

Thank you for attention!

mariia.matveieva@upol.cz

Examples of misinterpretation: N-O dataset + GC model

R2 model = 0.98 AUC = 0.91 Top-n score = 0.62



100 random molecules were analyzed