Chemical Representation Learning for Toxicity Prediction

Greta Markert[1,2], Jannis Born[1,2], Matteo Manica[1], Gisbert Schneider[2], María Rodríguez Martínez[1]

1 – IBM Research Zurich, Switzerland

2 – ETH Zurich, Switzerland

Chemical Representation Learning for Toxicity Prediction

Motivation

Despite seeming promising in pre-clinical studies in animals, more than **30%** of pharmaceuticals **failed in clinical trials** because of their **toxicity in humans**

Kola, Ismail, and John Landis (2004). "Can the pharmaceutical industry reduce attrition rates?." *Nature reviews Drug discovery* 3.8: 711-716.

115 million animals are utilized for clinical experimentation worldwide

Taylor, K., Gordon, N., Langley, G., and Higgins, W. (2008). Estimates for worldwide laboratory animal use in 2005. *Alternatives to Laboratory Animals*, 36(3):327–342.

With the current regulations, **aspirin and paracetamol** would **not** have been **approved**

Hartung, T. (2009). Per aspirin ad astra. . . . *Alternatives to Laboratory Animals*, 37(2 suppl):45–47.

The **intersex rate** of male smallmouth and largemouth bass in the U.S. ranges from **60% to 100%** because of an increase in **estrogenic endocrine disruption**

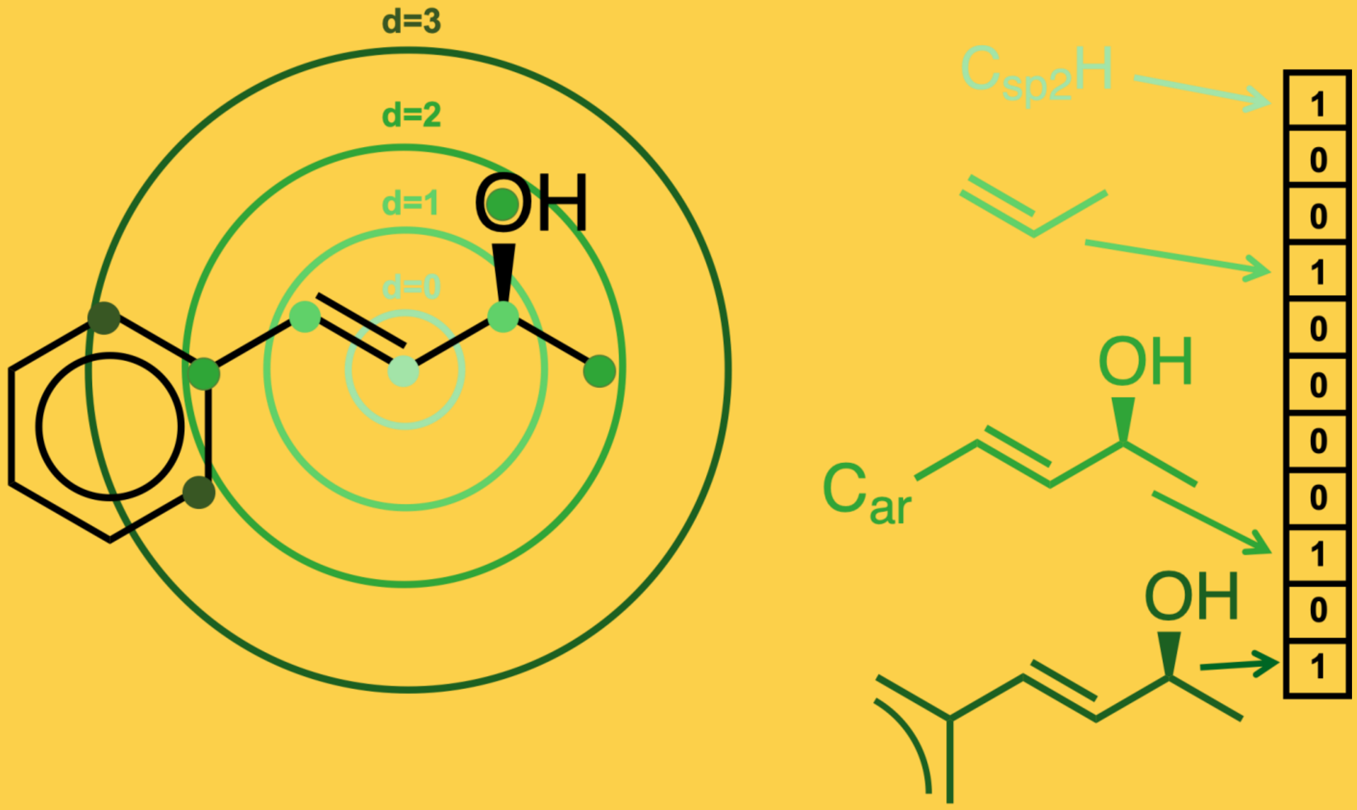
Iwanowicz, L. R et al (2016). Evidence of estrogenic endocrine disruption in smallmouth and largemouth bass inhabiting Northeast US national wildlife refuge waters: A reconnaissance study. *Ecotoxicology and environmental safety*, *124*, 50-59.

Drug design
Preclinical Trials
Clinical Trials
Post-Approval

Chemical Representation Learning for Toxicity Prediction

Chemical Representation

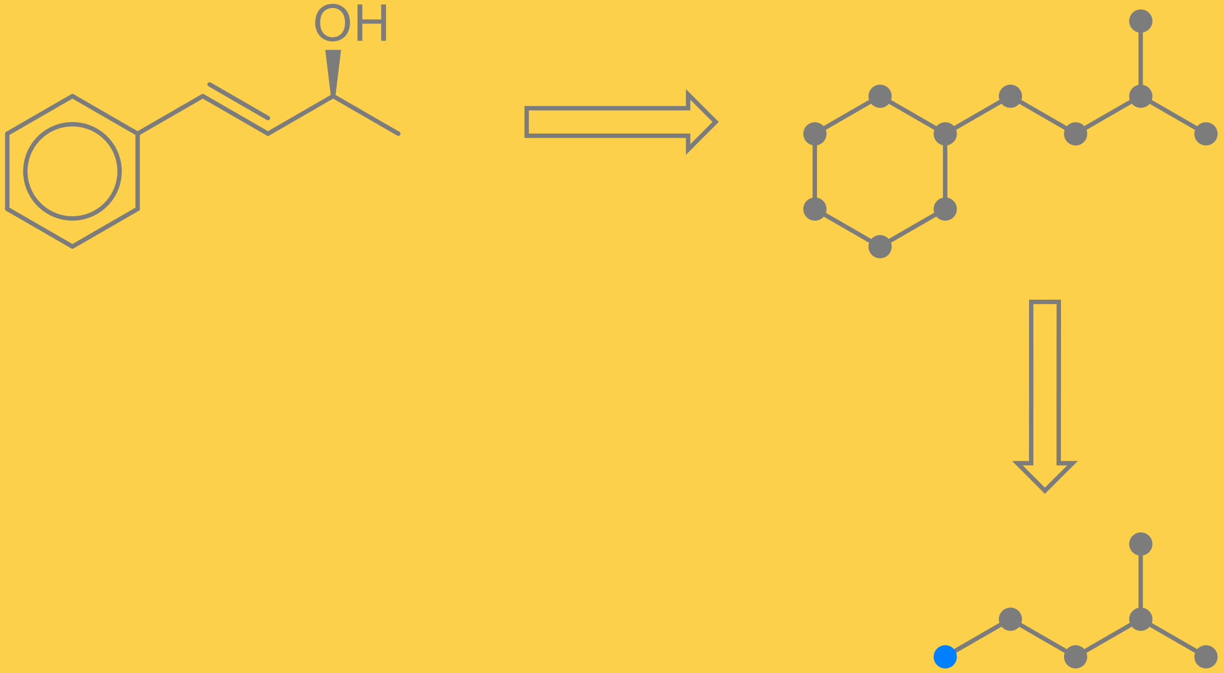
- Fingerprints
- Graphs
- SELFIES
- SMILES



Chemical Representation Learning for Toxicity Prediction

Chemical Representation

- Fingerprints
- Graphs
- SELFIES
- SMILES

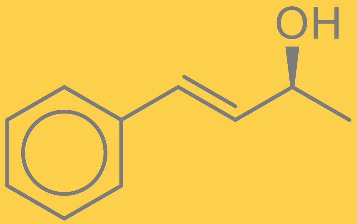


Chemical Representation Learning for Toxicity Prediction

Chemical Representation

- Fingerprints
- Graphs
- SELFIES
- SMILES

SMILES Flavors:



original molecule (“raw”):	c1ccc(/C=C/[C@H](C)O)cc1
canonical (RDKit):	C[C@H](O)/C=C/c1ccccc1
remove stereoinformation:	c1ccc(/C=C/C(C)O)cc1
remove double bond direction:	c1ccc(C=C[C@H](C)O)cc1
kekulization:	C1=CC=C(/C=C/[C@H](C)O)C=C1
explicit bonds:	c1:c:c:c(/C=C/[C@H](-C)-O):c:c1
explicit hydrogens:	[cH]1[cH][cH][c](/[CH]=[CH]/[C@H]([CH3])[OH])[cH][cH]1
augmentation:	C[C@H](O)C=Cc1ccccc1,...
shuffling:	c[C@H]Ccc/C(Cc=Oc)1/)c(,...
SELFIES:	[c][Branch13][Branch21][/C][=C][/C@Hexpl][Branch13][epsilon][C][O][c][c][c][c][c][Ring1][Branch23]

Chemical Representation Learning for Toxicity Prediction

Chemical Representation

- **Fingerprints**
- **Graphs**
- **SELFIES**
- **SMILES**

Toxicity Prediction

- **Tox21**: environmental toxicity

  → 12’707 compounds, 12 tasks

  → ROC-AUC: 0.877

- **SIDER**: side effects

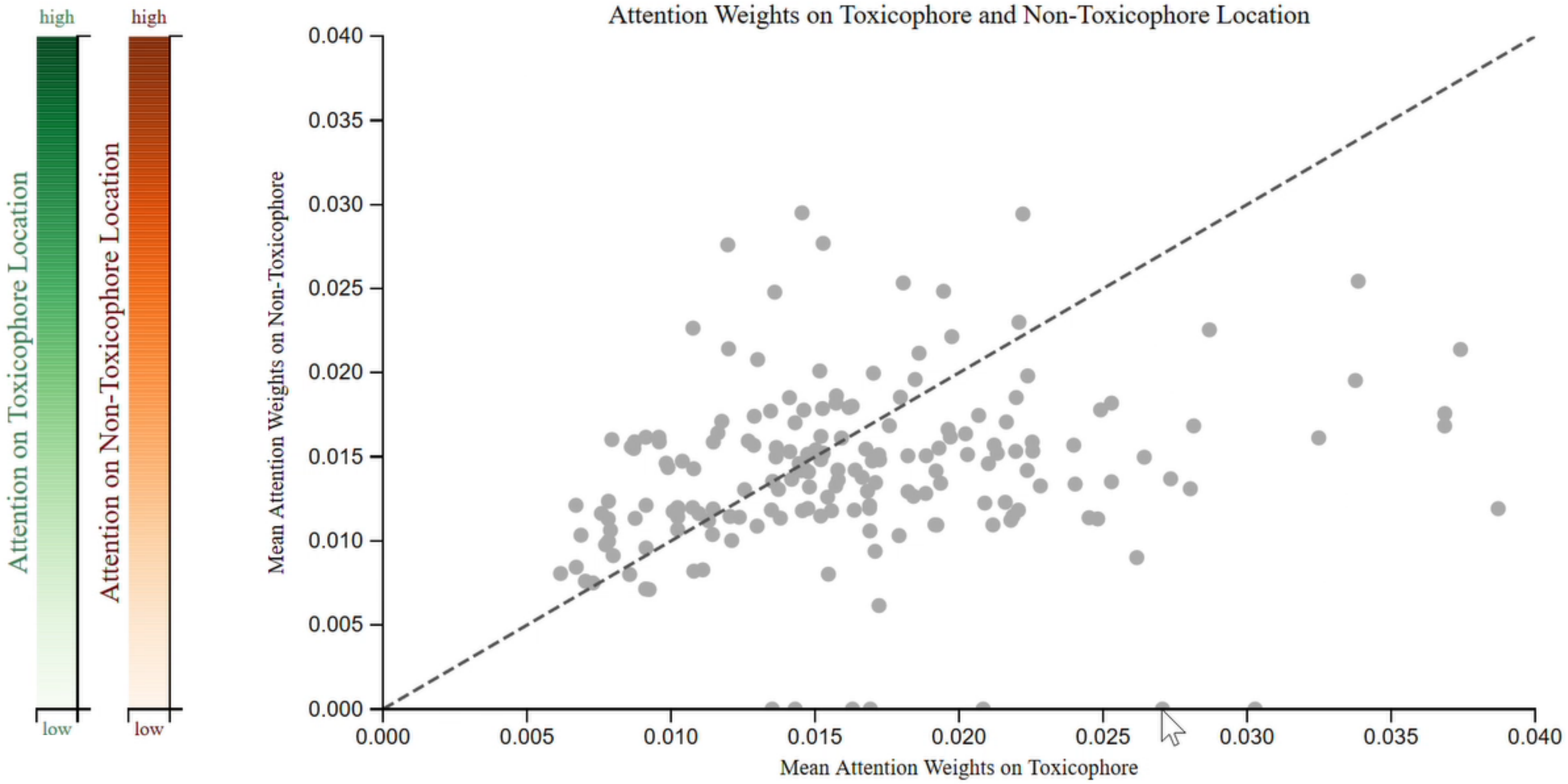
  → 1’430 compounds, 27 tasks

  → ROC-AUC: 0.835

- **ClinTox**: toxicity during clinical trials

  →1’491 compounds, 2 tasks

  →ROC-AUC: 0.983



→ **Mean attention weights from model on toxicophores are significantly higher than on non-toxicophores**

→**Purely data-driven approach**

→**Validation of prediction model**

→**Generation of new toxicophore hypotheses**

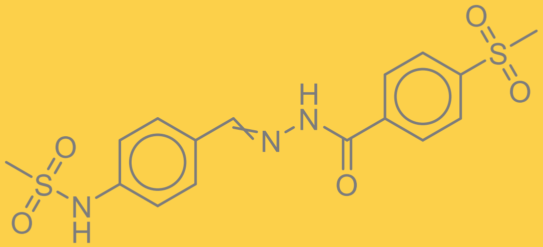
Chemical Representation Learning for Toxicity Prediction

Application

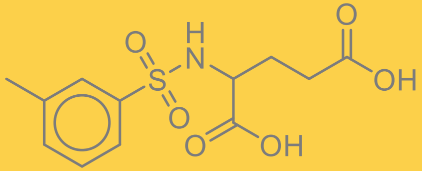
- **Implementation into PaccMann***

  →Generation of efficacious, transcriptomics-specific cancer drugs

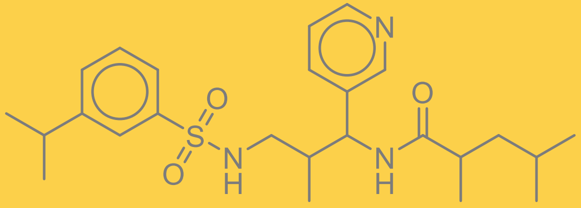
  →Environmental toxicity, side effects and toxicity in clinical trials as critics in generative model



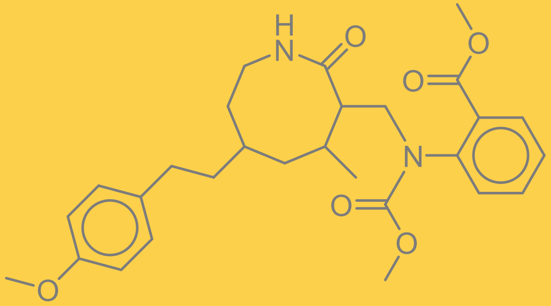
Predicted IC_{50}: 242 nM



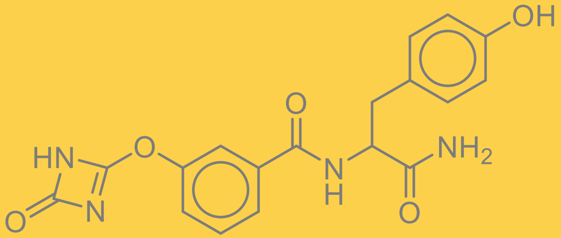
Predicted IC_{50}: 113 nM



Predicted IC_{50}: 241 nM



Predicted IC_{50}: 58 nM



Predicted IC_{50}: 30 nM

→ All these compounds are predicted to be non-toxic for each Tox21 task

*Born, Jannis, et al. "Paccmann rl: Designing anticancer drugs from transcriptomic data via reinforcement learning." *International Conference on Research in Computational Molecular Biology*. Springer, Cham, 2020.