Multitask Deep learning modelling of rodent acute toxicity

04/11/2018

Alexey Zakharov

NCATS

Acute toxicity data set

NICEATM web site:

- 6734 compounds with LD_{50} values measured on rats, oral administration

In addition, we have collected the following data from **ChemIDPlus** web-site*:

- About 50,000 chemical structures with data on acute rodent (mouse and rat) toxicity
- Toxicity endpoints are expressed in LD_{50} (mg/kg) values
- Four types of administration:
  - Oral
  - Intravenous
  - Intraperitoneal
  - Subcutaneous

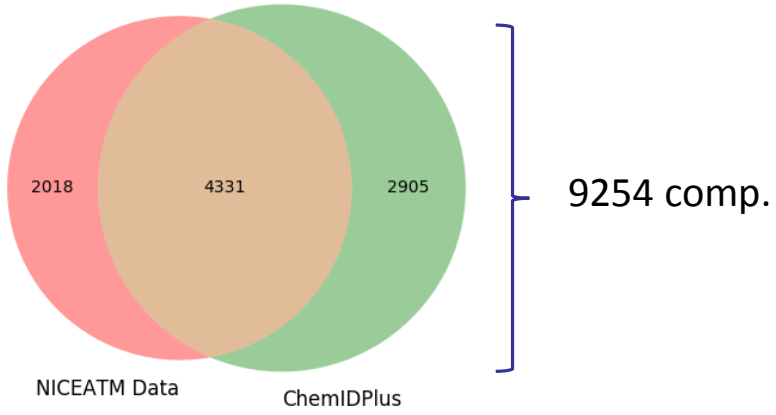
* **http://chem.sis.nlm.nih.gov/chemidplus/**

Data set preparation

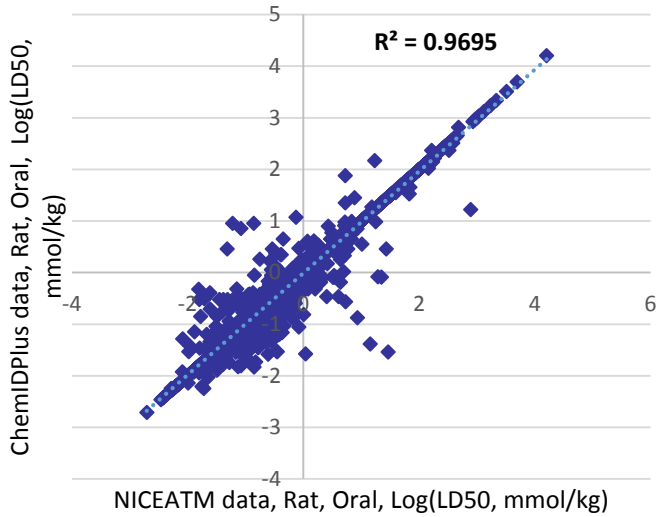
- All structures were standardized/normalized
- All duplicate compounds were eliminated
- Mixtures and salts were removed
- LD_{50} values in mg/kg were converted to mmol/kg

Activity	Number of compounds after curation
Rat Oral LD_{50} (NICEATM)	6,349
Rat Intraperitoneal LD_{50}	4,297
Rat Intravenous LD_{50}	2,360
Rat Oral LD_{50} (ChemIDPlus)	7,241
Rat Subcutaneous LD_{50}	1,456
Mouse Intraperitoneal LD_{50}	29,564
Mouse Intravenous LD_{50}	15,494
Mouse Oral LD_{50}	16,525
Mouse Subcutaneous LD_{50}	5,609

Overlap Analysis

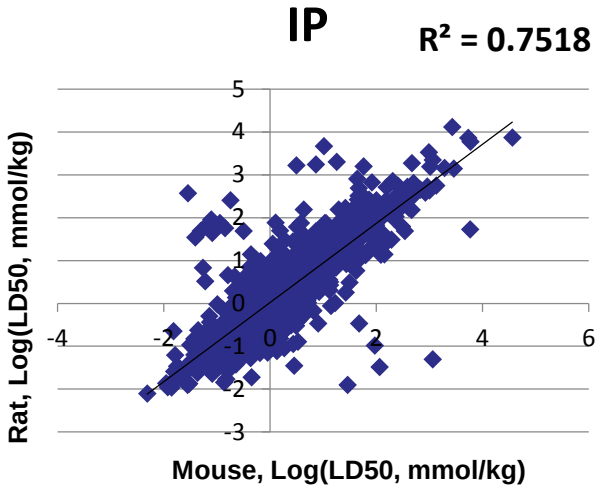


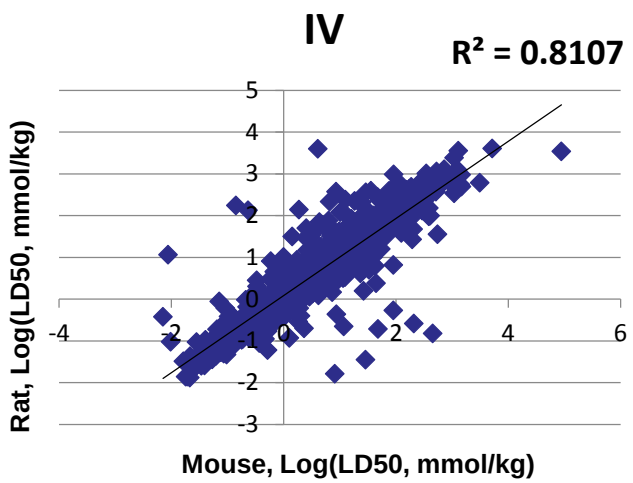
Reproducibility

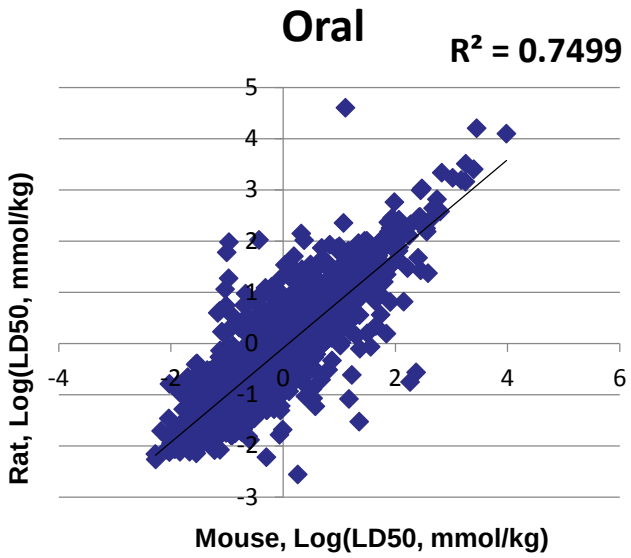


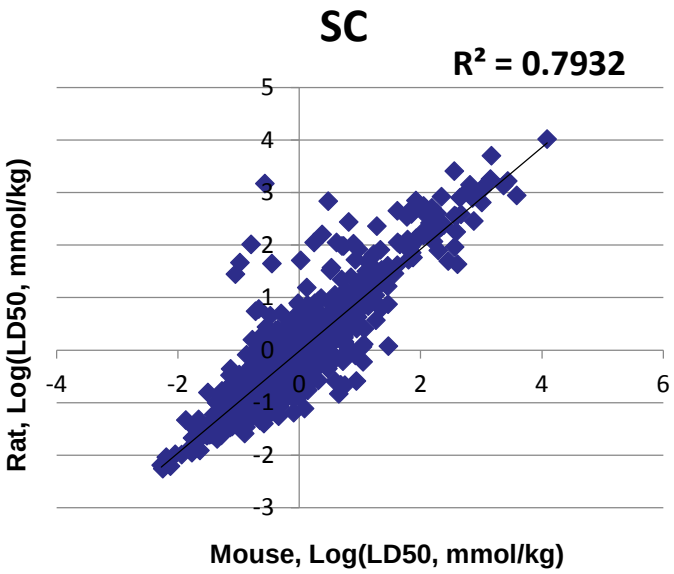
NIH National Center for Advancing Translational Sciences

Interspecies comparisons

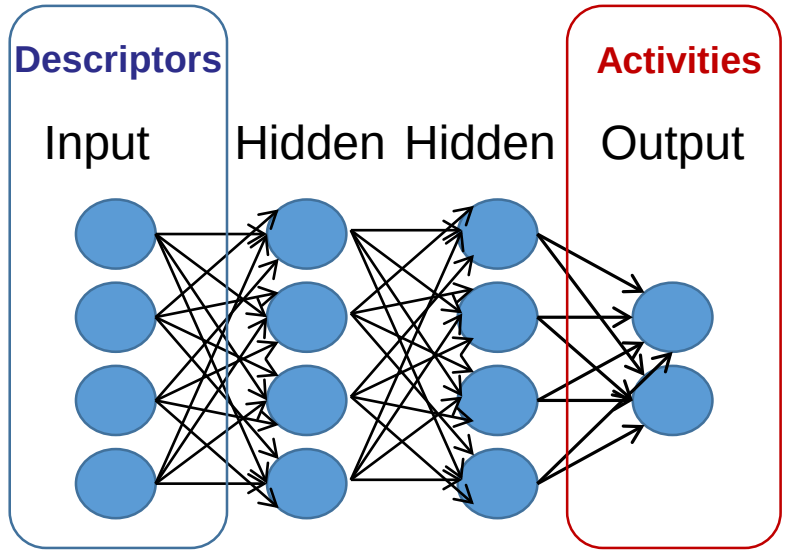








Multitask Deep learning (MDL)

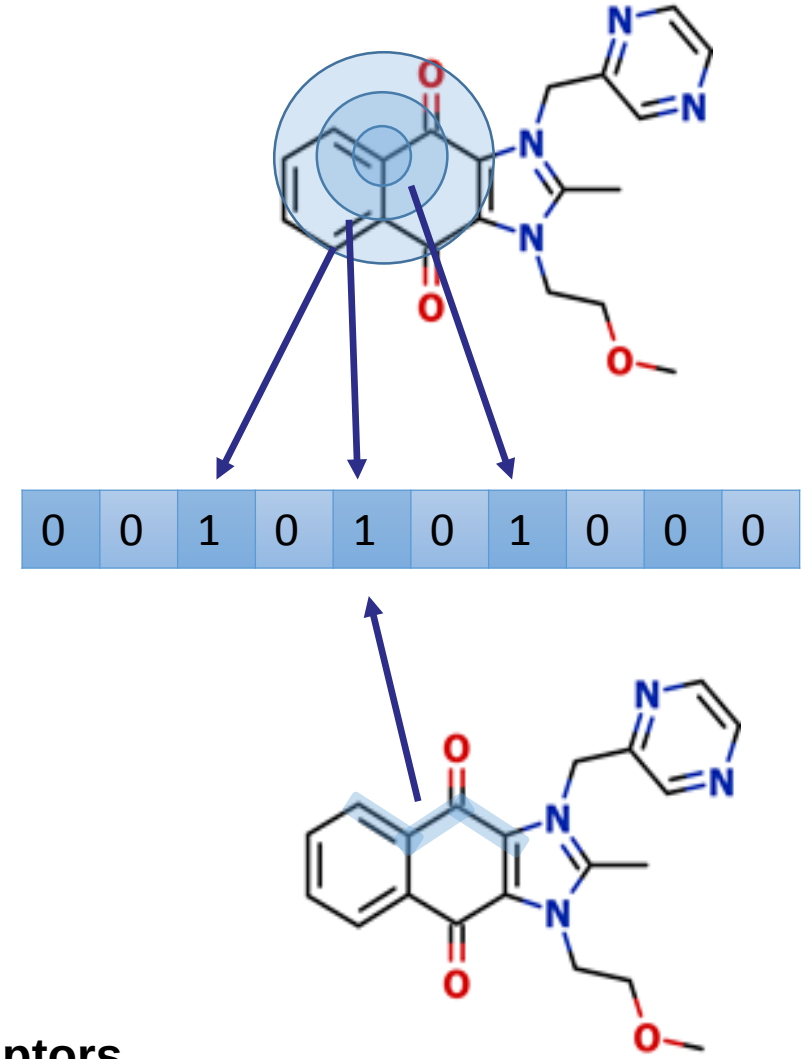


NIH National Center for Advancing Translational Sciences

QSAR Methods

Descriptors:

- RDkit Morgan fingerprints
  Circular fingerprints, 2048 bit, radius 2
- RDkit Avalon fingerprints
  Path based fingerprints, 2048 bit
- RDkit physical-chemical descriptors
  SLogP, topological polar surface area, molecular weight, number of hydrogen bond donors and acceptors

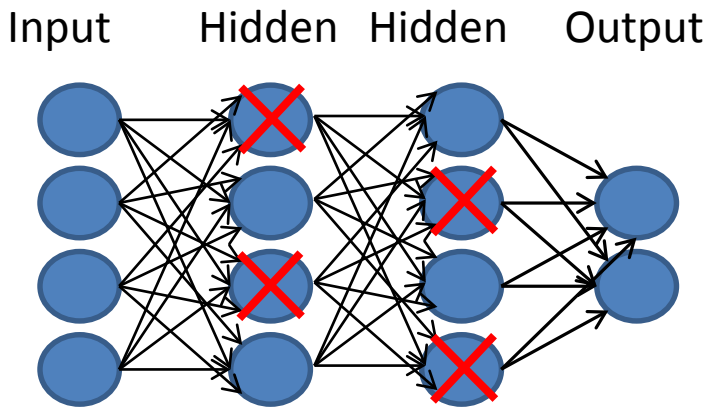


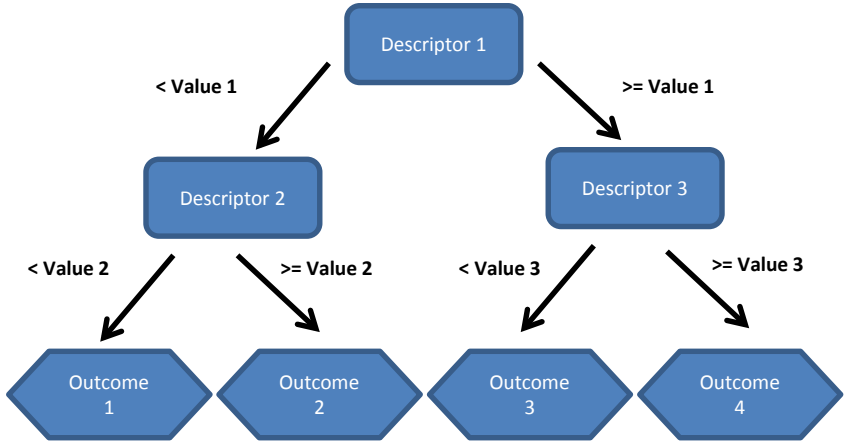
- **Morgan Fingerprint and 5 phys-chem descriptors**
- **Avalon Fingerprint and 5 phys-chem descriptors**

Machine Learning approaches

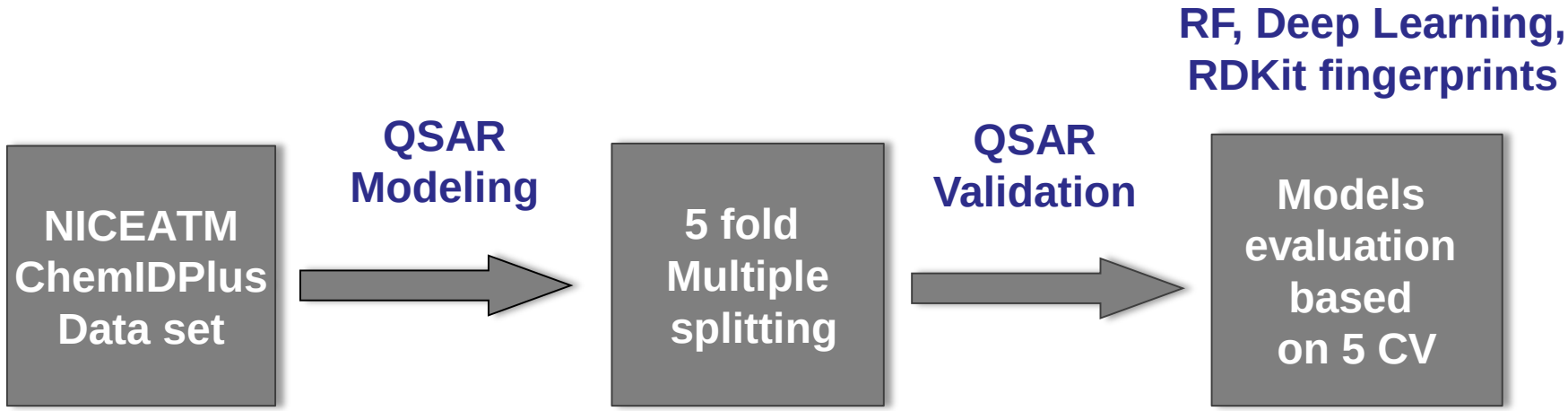
- **Random Forest** as baseline approach: 300 trees, 2048+5 features

- **Multitask Deep Learning**: ReLu, 4 hidden layers, ADAM optimizer, Dropout, Dense layers

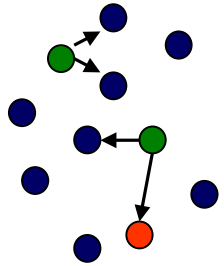


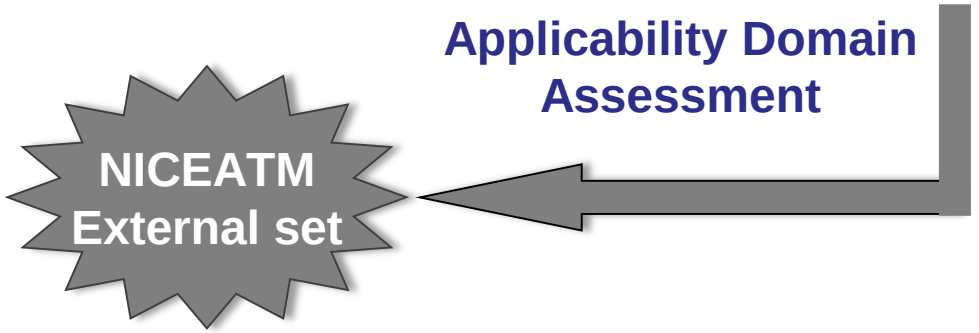


Modeling workflow



AD: Tanimoto Similarity
calculated between compounds using Morgan fingerprints

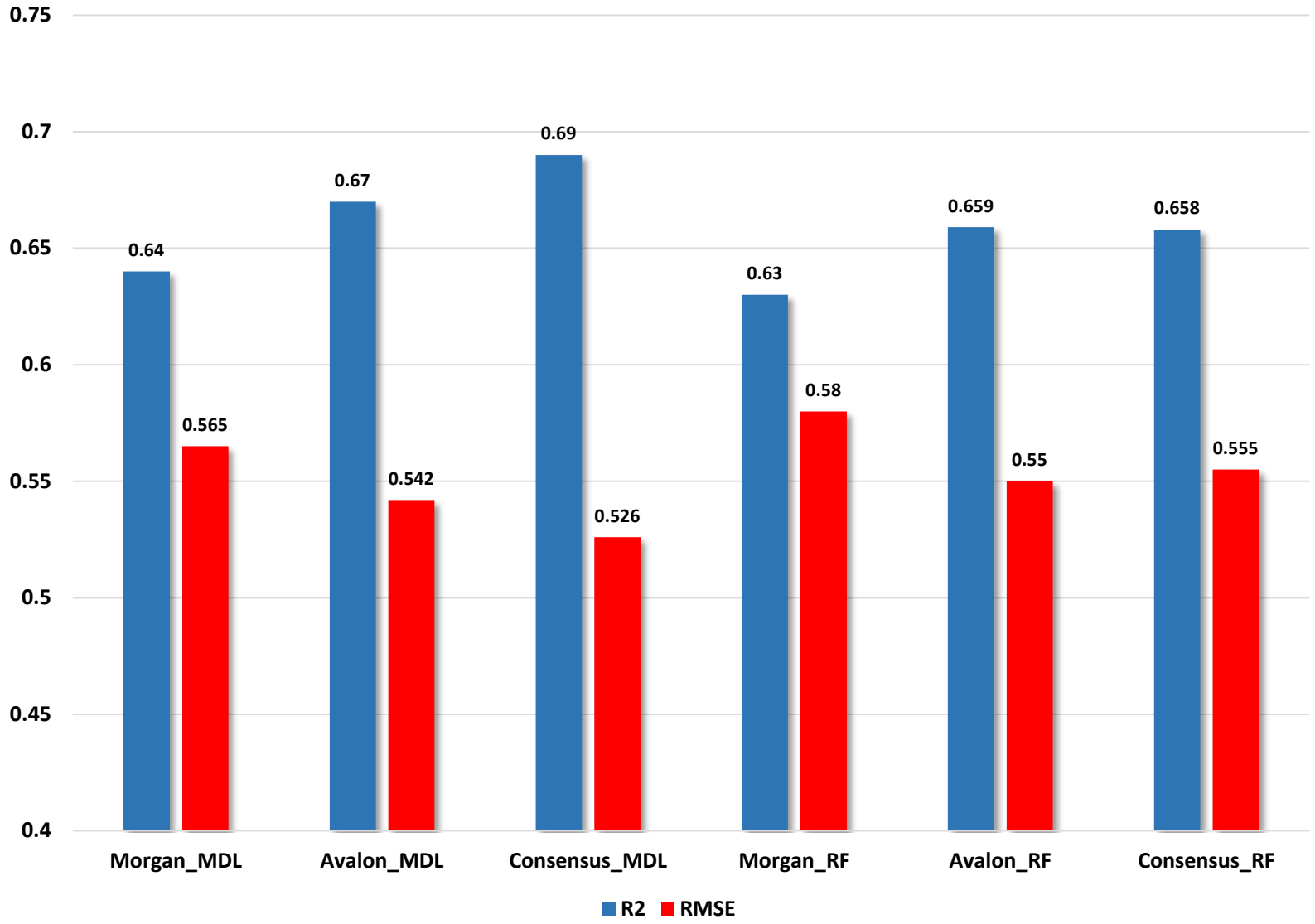




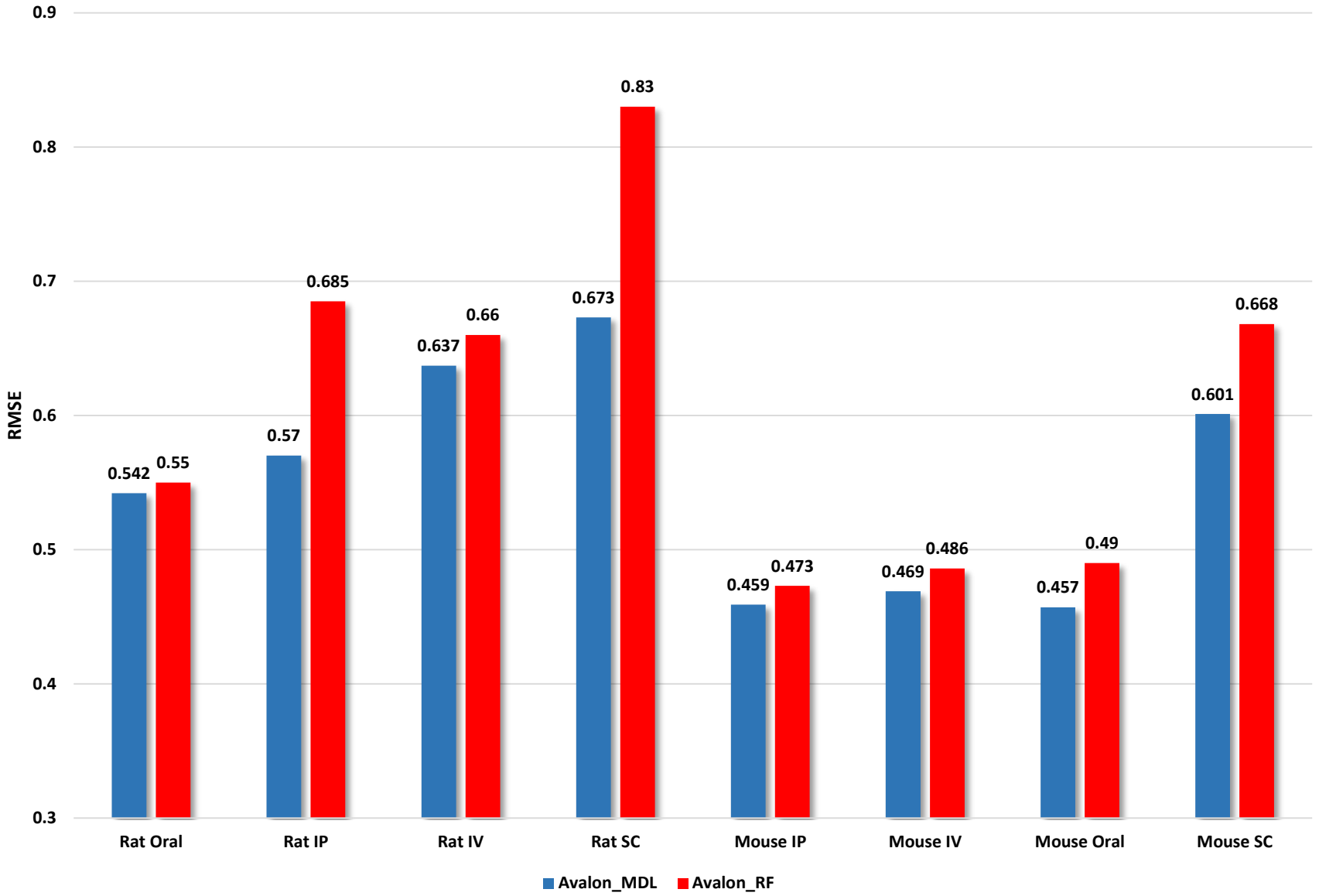
48134 Compounds

NIH National Center for Advancing Translational Sciences

External 5 fold CV prediction results

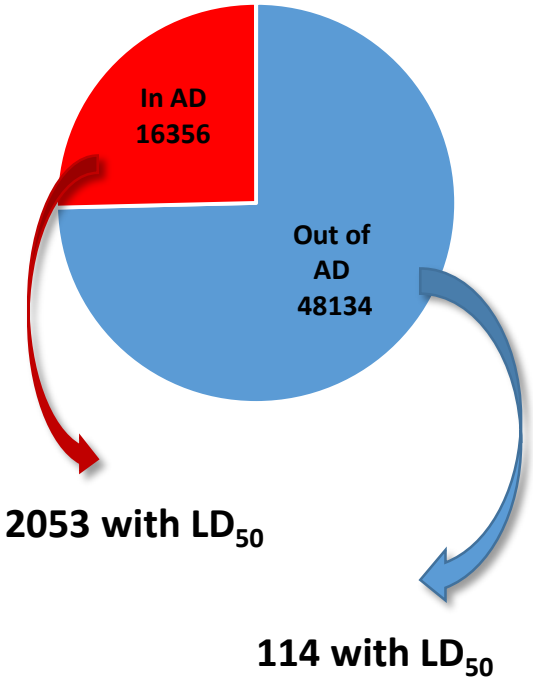


Can Multitask improve results?

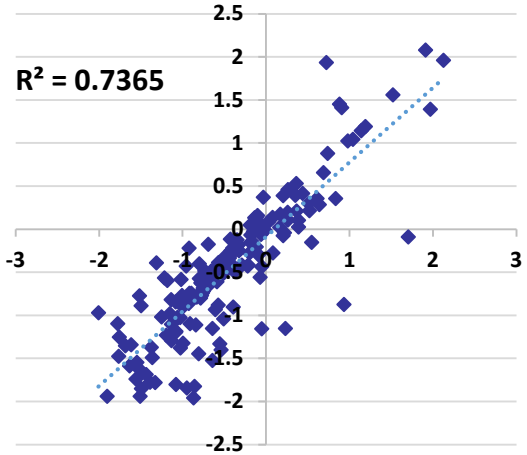


Analysis of external test set

Test set coverage

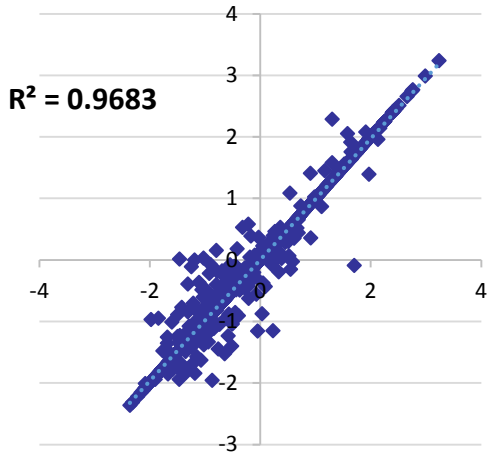


Data reproducibility, NICEATM training vs test

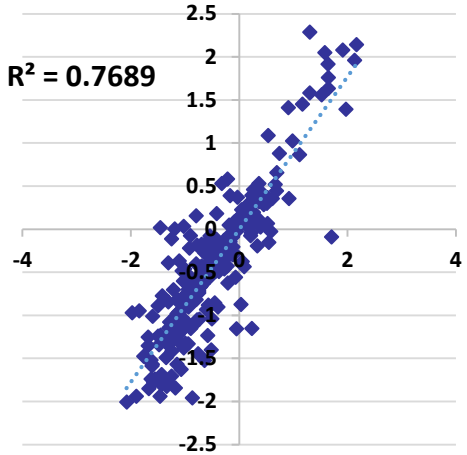


RMSE: 0.417

Data reproducibility, combined training vs NICEATM test



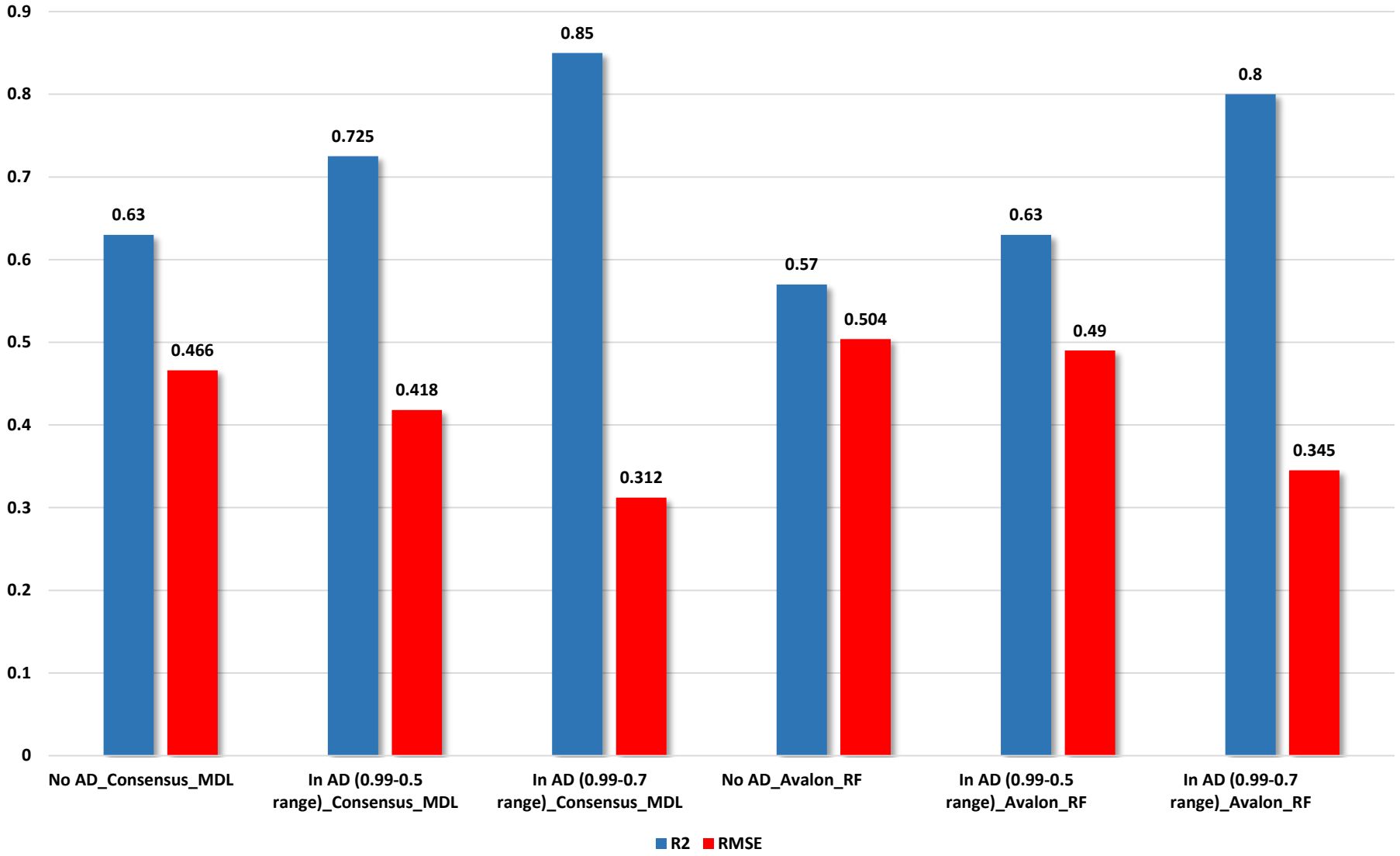
RMSE: 0.161



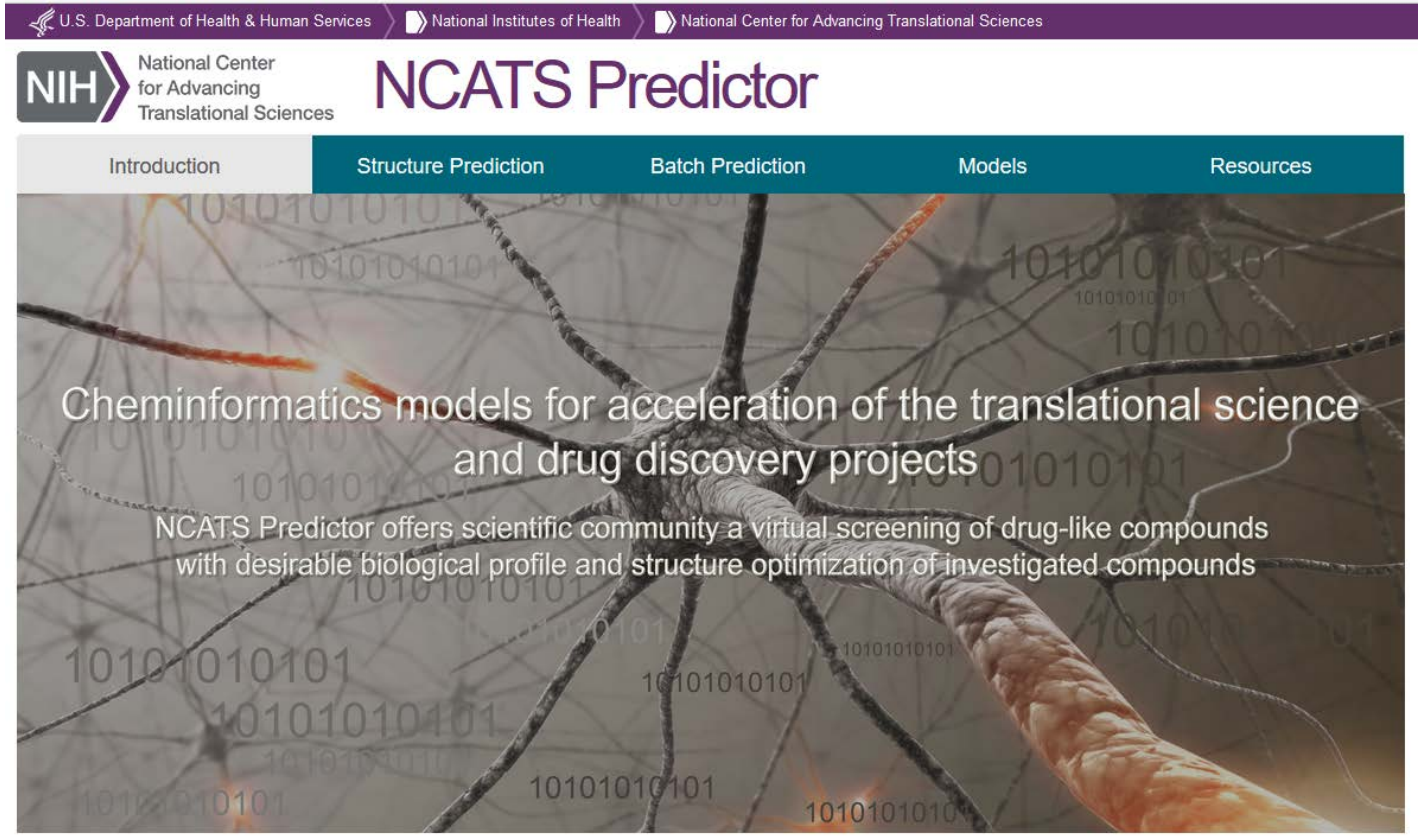
RMSE: 0.402

NIH National Center for Advancing Translational Sciences

External test set prediction results



Models Dissemination: NCATS Predictor



- Predict 1121 biological activities
- Supports SMILES, drug name, images
- Allows to send the batch of compounds
- Show up neighbor activity and structure

https://predictor.ncats.io/

Acknowledgments

National Center for Advancing Translational Sciences (NCATS)

- Tongan Zhao
- Timothy Sheils
- Gergely Zahoranszky-Kohalmi
- Tyler Peryea
- Dac-Trung Nguyen
- Noel Southall
- Anton Simeonov

NIH National Center for Advancing Translational Sciences