

# 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<sub>50</sub> 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<sub>50</sub> (mg/kg) values
- Four types of administration:

Oral

Intravenous

Intraperitoneal

Subcutaneous

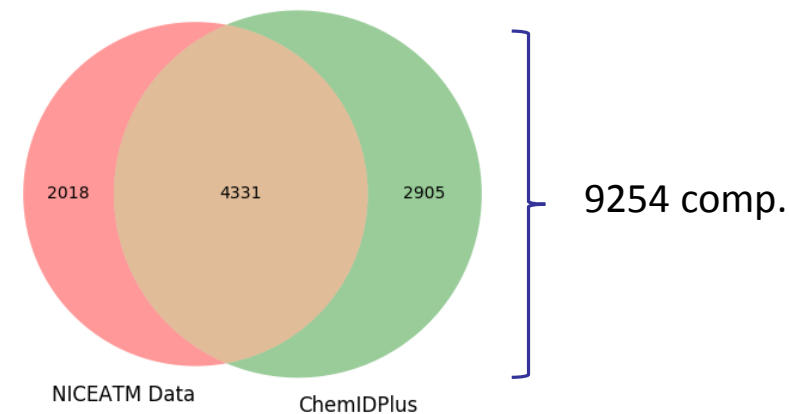
\* <http://chem.sis.nlm.nih.gov/chemidplus/>

# Data set preparation

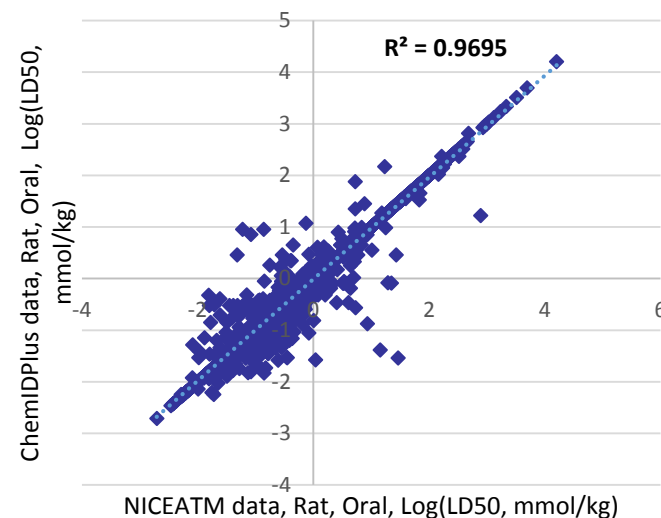
- All structures were standardized/normalized
- All duplicate compounds were eliminated
- Mixtures and salts were removed
- LD<sub>50</sub> values in mg/kg were converted to mmol/kg

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

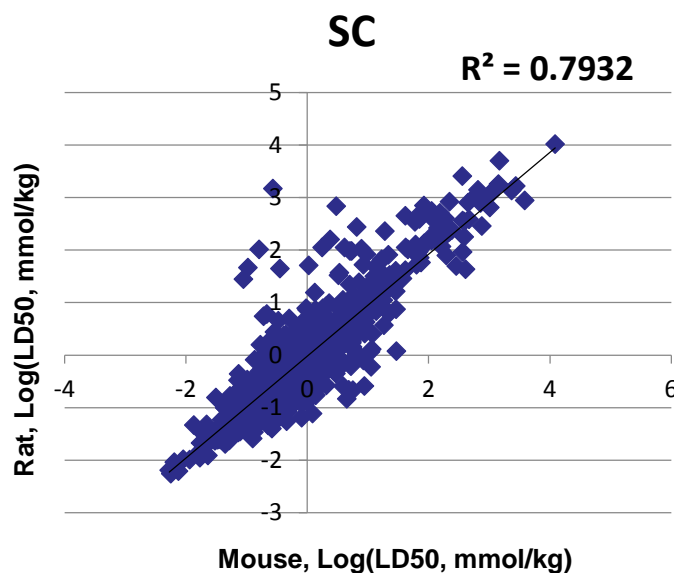
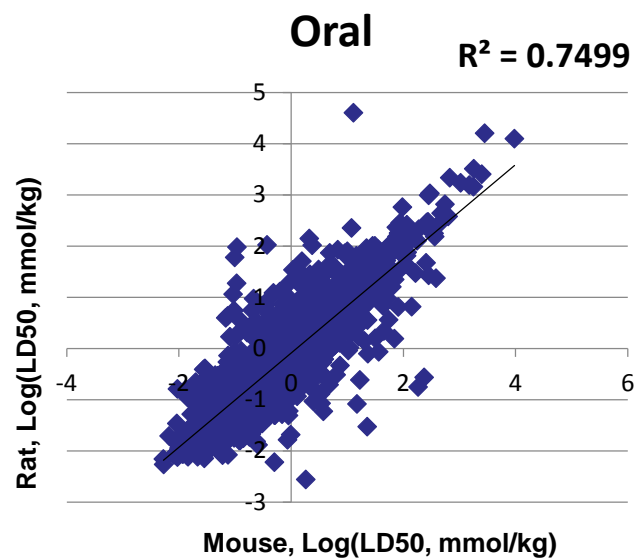
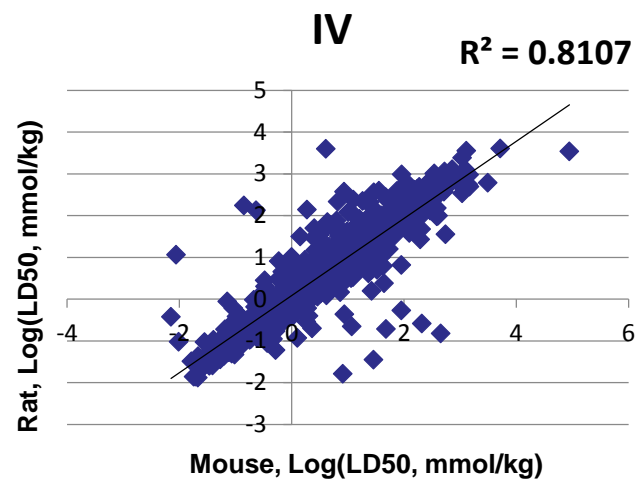
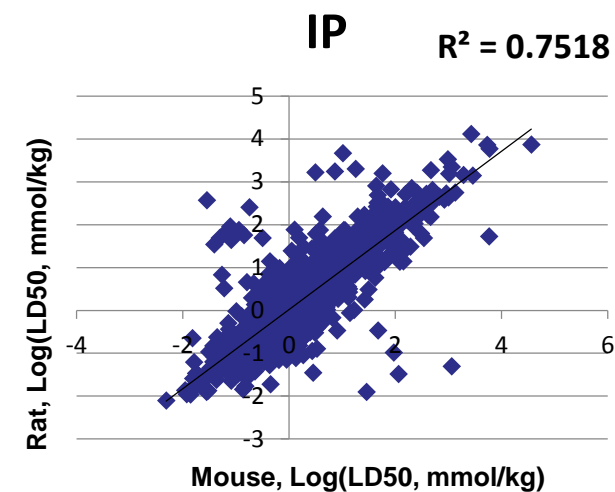
## Overlap Analysis



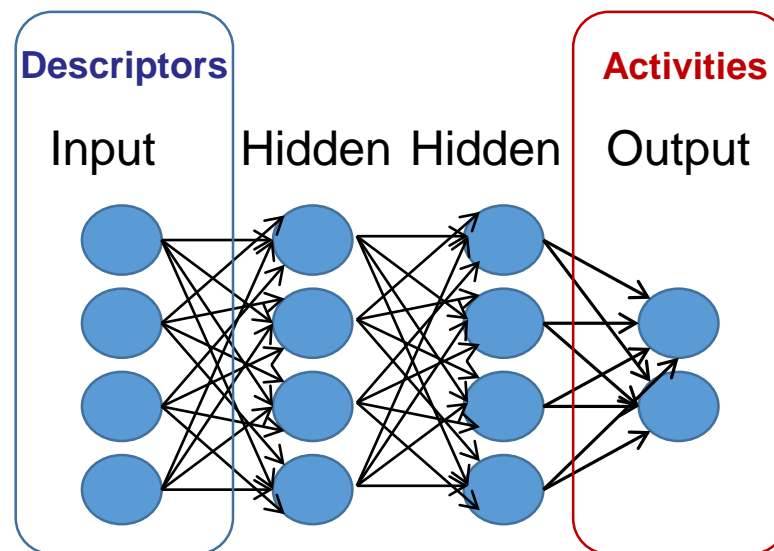
## Reproducibility



# Interspecies comparisons



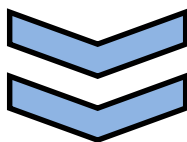
## Multitask Deep learning (MDL)



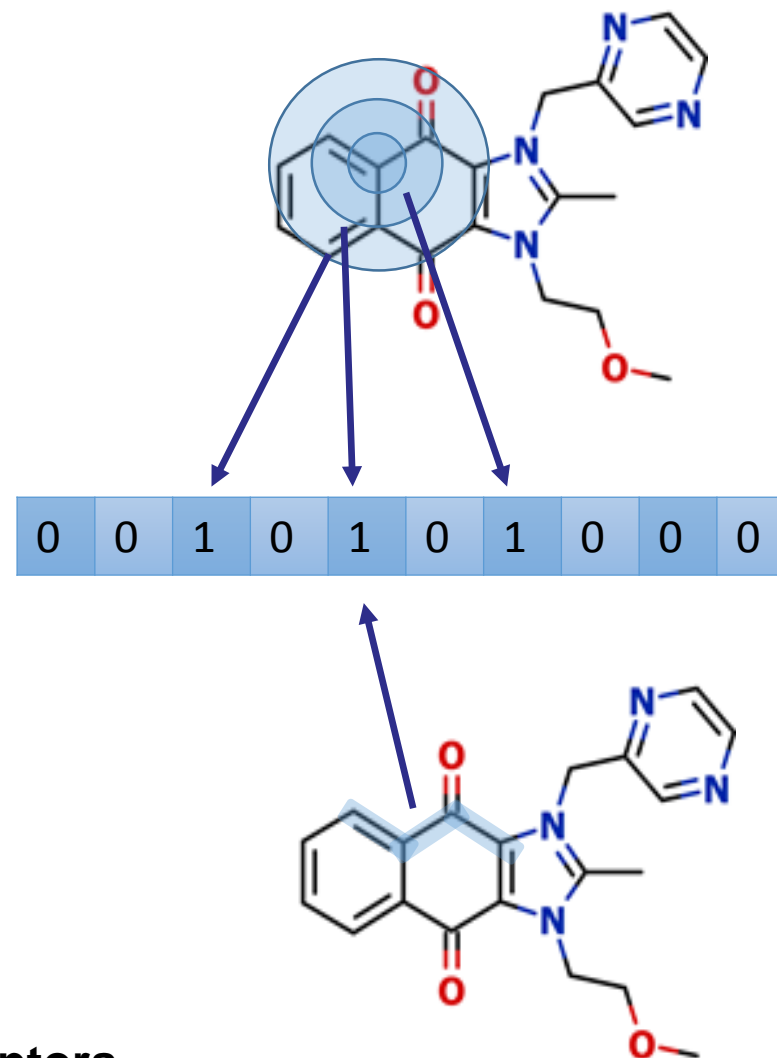
# QSAR Methods

## 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**

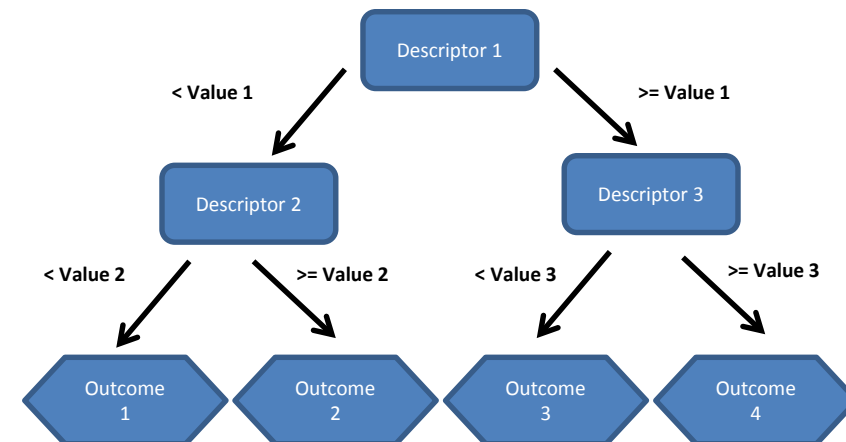
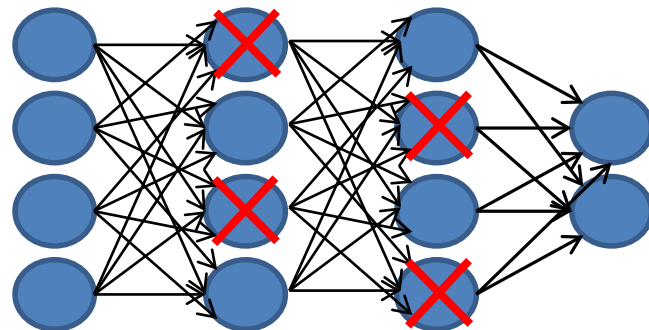


# Machine Learning approaches

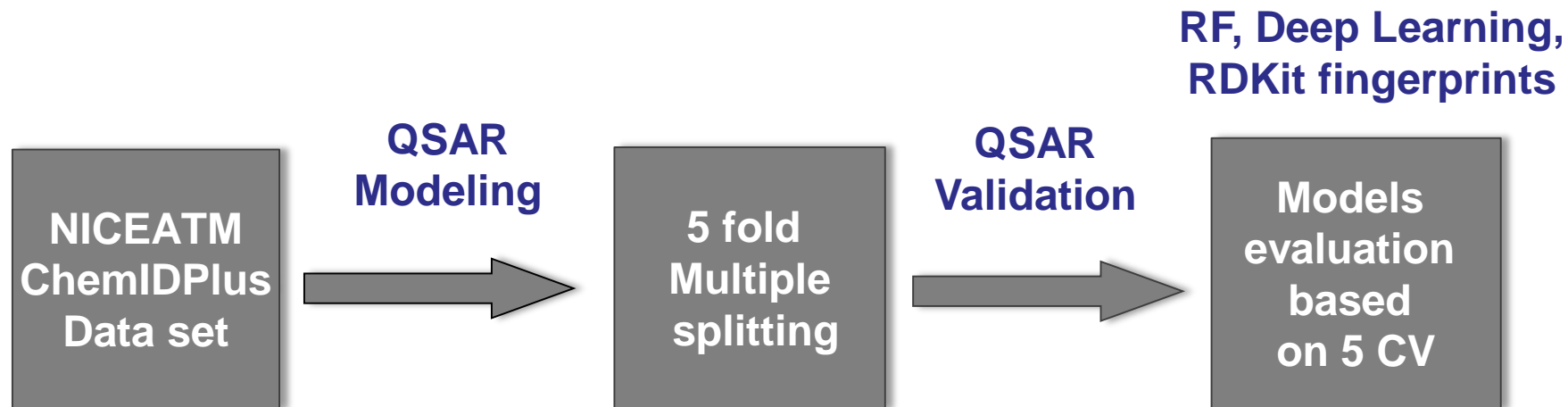


- **Random Forest** as baseline approach: 300 trees, 2048+5 features
- **Multitask Deep Learning**: ReLu, 4 hidden layers, ADAM optimizer, Dropout, Dense layers

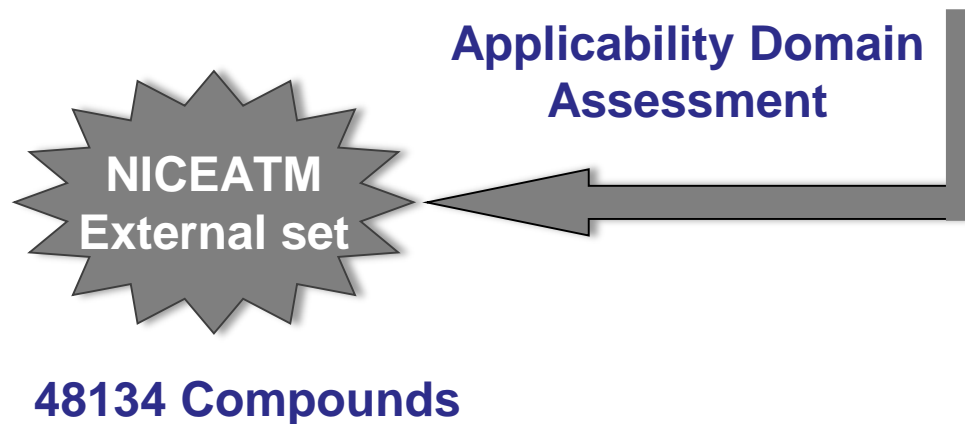
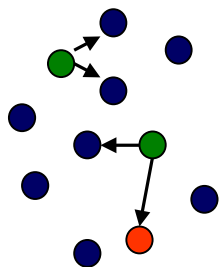
Input Hidden Hidden Output



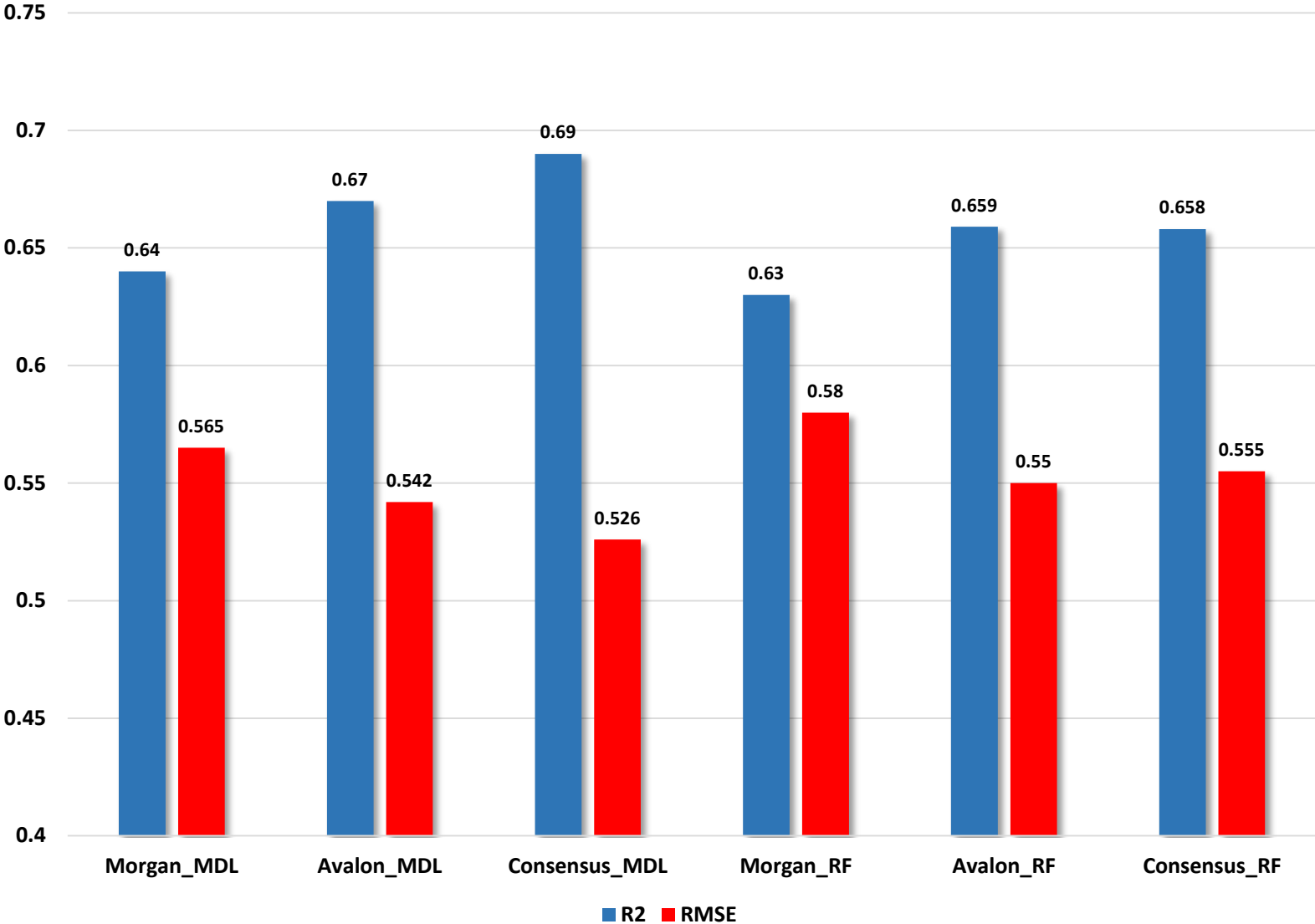
# Modeling workflow



**AD: Tanimoto Similarity**  
calculated between compounds  
using Morgan fingerprints

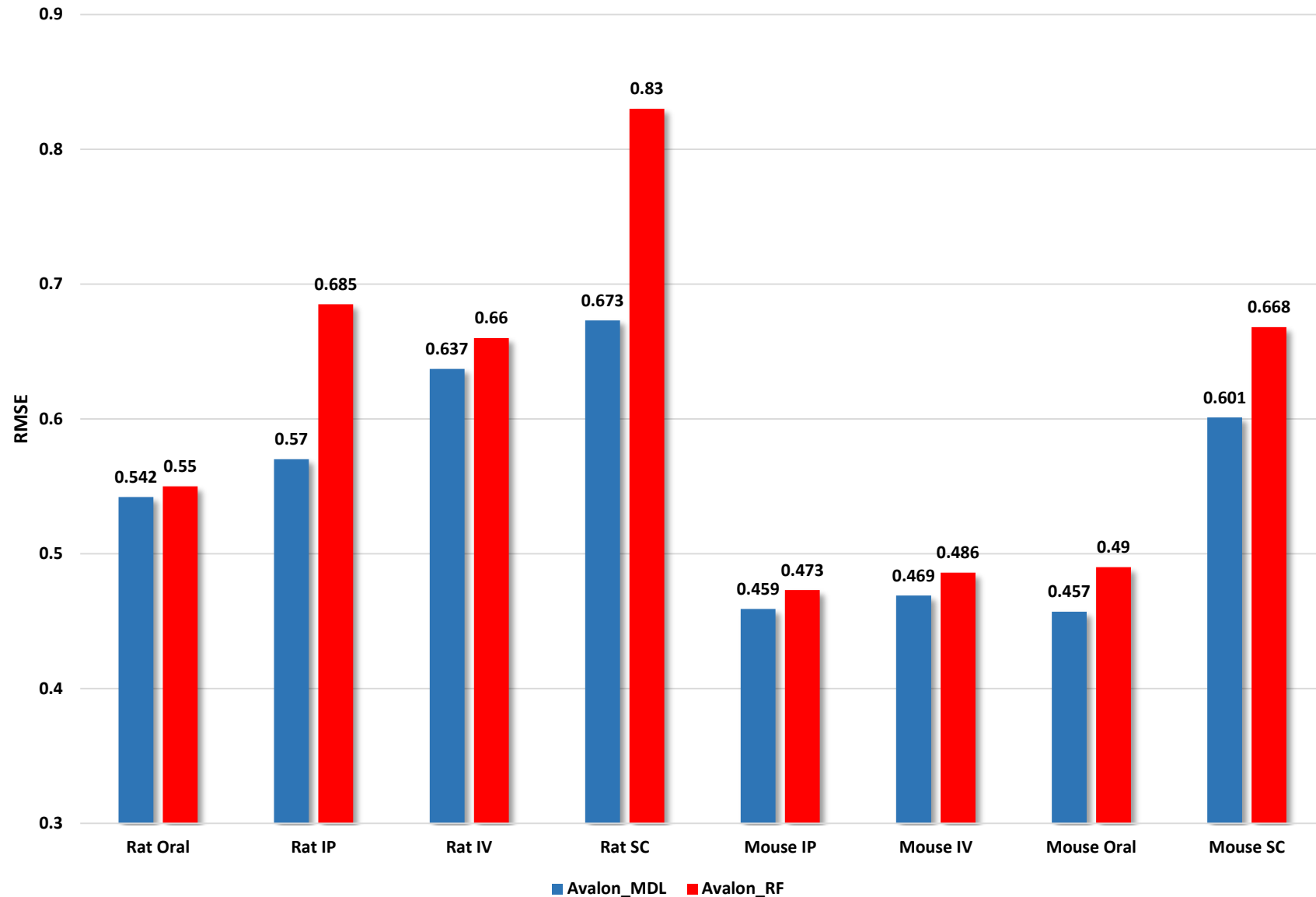


# External 5 fold CV prediction results



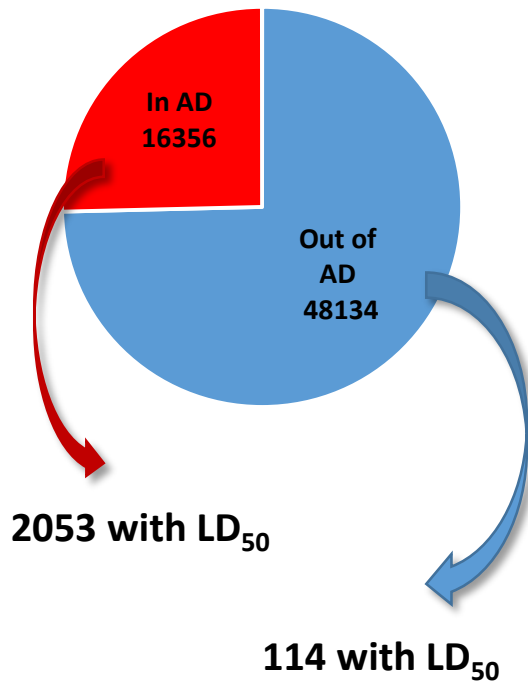


# Can Multitask improve results?

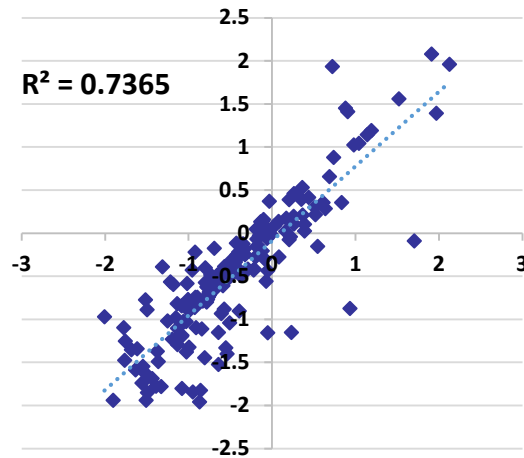


# Analysis of external test set

## Test set coverage

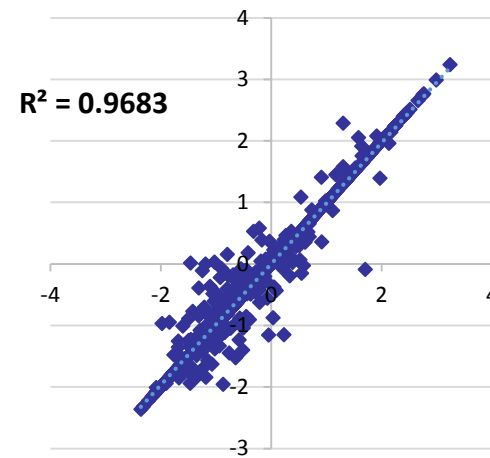


## Data reproducibility, NICEATM training vs test

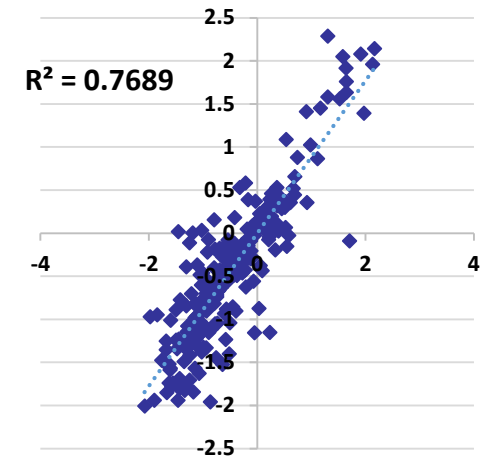


RMSE: 0.417

## Data reproducibility, combined training vs NICEATM test

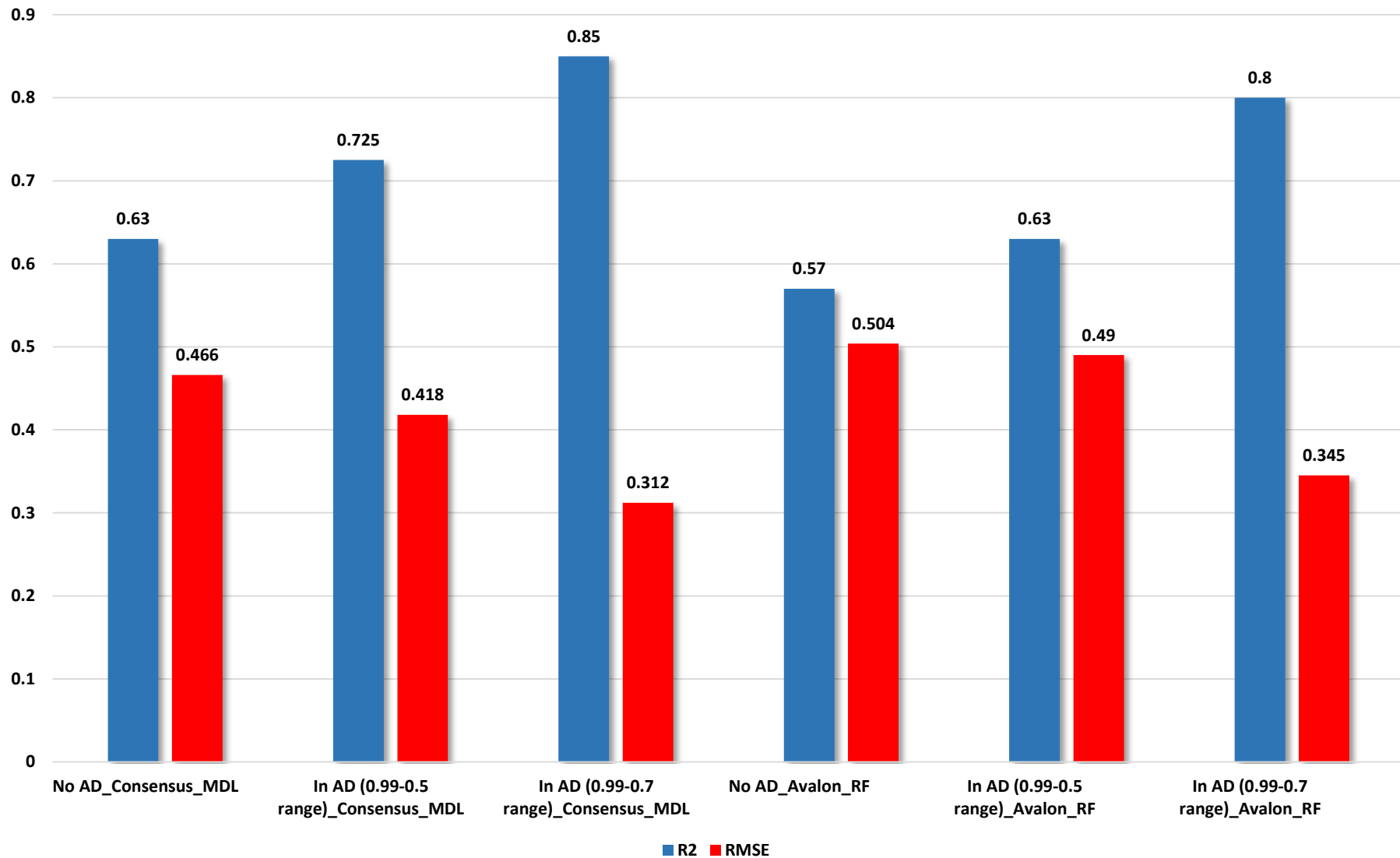


RMSE: 0.161



RMSE: 0.402

# External test set prediction results



# Models Dissemination: NCATS Predictor

U.S. Department of Health & Human Services | National Institutes of Health | National Center for Advancing Translational Sciences

**NIH** National Center for Advancing Translational Sciences

## NCATS Predictor

Introduction | Structure Prediction | Batch Prediction | Models | Resources

Cheminformatics models for acceleration of the translational science and drug discovery projects

NCATS Predictor offers scientific community a virtual screening of drug-like compounds with desirable biological profile and structure optimization of investigated compounds

- 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