Using Deep Learning to Predict Molecule Activity with Its Structure

Is Deep Learning an Evolutionary or Revolutionary Solution?

Junshui Ma

In collaboration with Yuting Xu, Andy Liaw, Robert P. Sheridan and Vladimir Svetnik

Merck & Co. Inc.

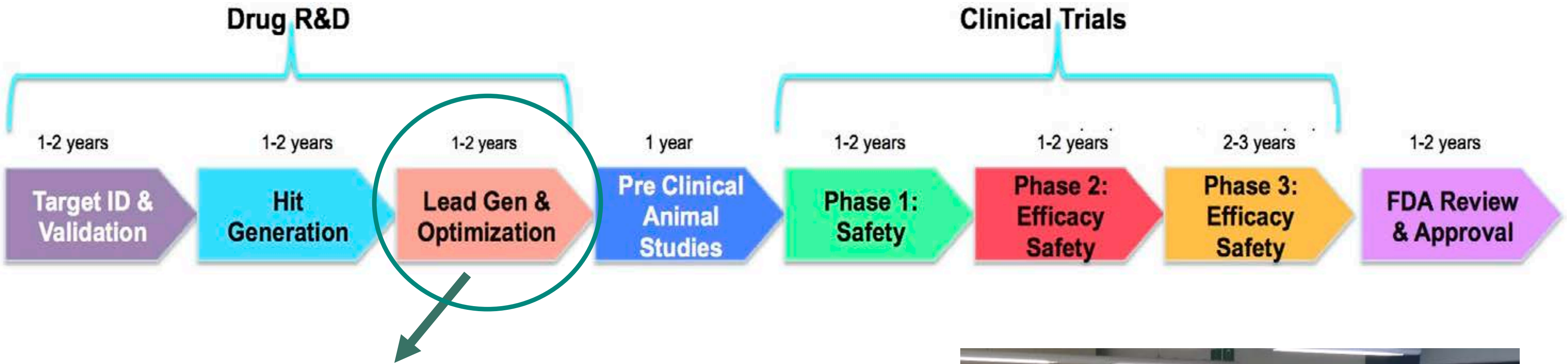
NISS-Merck Meetup, April 25, 2018

Outline

- Background: Drug Development and QSAR**
- Deep Neural Net (DNN) for QSAR:
  - Does “Deep” help?
  - Why Multi-task DNN works?
- Summary and Discussion

** **QSAR**(**Q**uantitative **S**tructure and **A**ctivity **R**elationship): A research area to study the relationship between a molecule’s structure and its chemical and biological activities.

Drug Development (Small Molecules)



Medicinal Chemistry Capability

- Lead molecule identification
- Lead molecule optimization

✓ Target potency

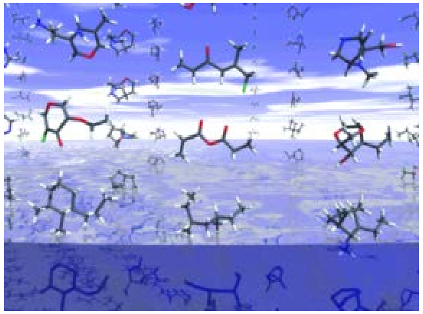
✓ ADME (Absorption, Distribution, Metabolism, Excretion)

✓ Toxicity

✓ …

i.e. Molecules' chemical/biological ***activities***

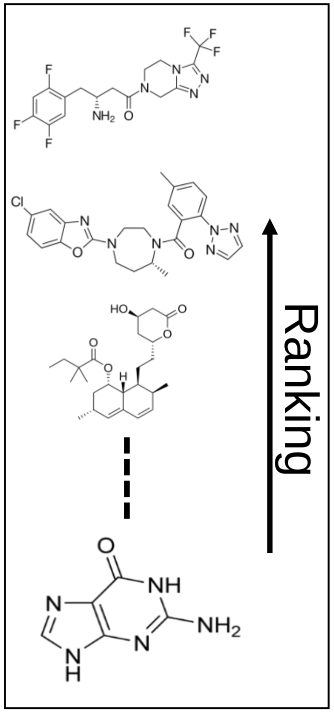
QSAR: Quantitative Structure & Activity Relationship



Molecule Activities by **Lab Experiment**

Computer Predictive QSAR Models

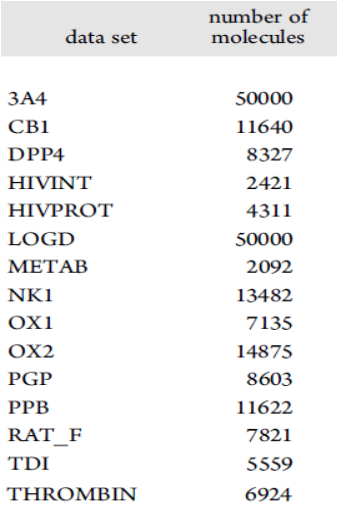
Molecule Activities by **Model Prediction**



Correlation(Lab, Computer) : 0.30 ~ 0.91

Merck QSAR Kaggle Challenge (2012)

15 Diverse Merck QSAR Datasets



data set	number of molecules
3A4	50000
CB1	11640
DPP4	8327
HIVINT	2421
HIVPROT	4311
LOGD	50000
METAB	2092
NK1	13482
OX1	7135
OX2	14875
PGP	8603
PPB	11622
RAT_F	7821
TDI	5559
THROMBIN	6924

The New York Times

Scientists See Promise in Deep-Learning Programs

By JOHN MARKOFF

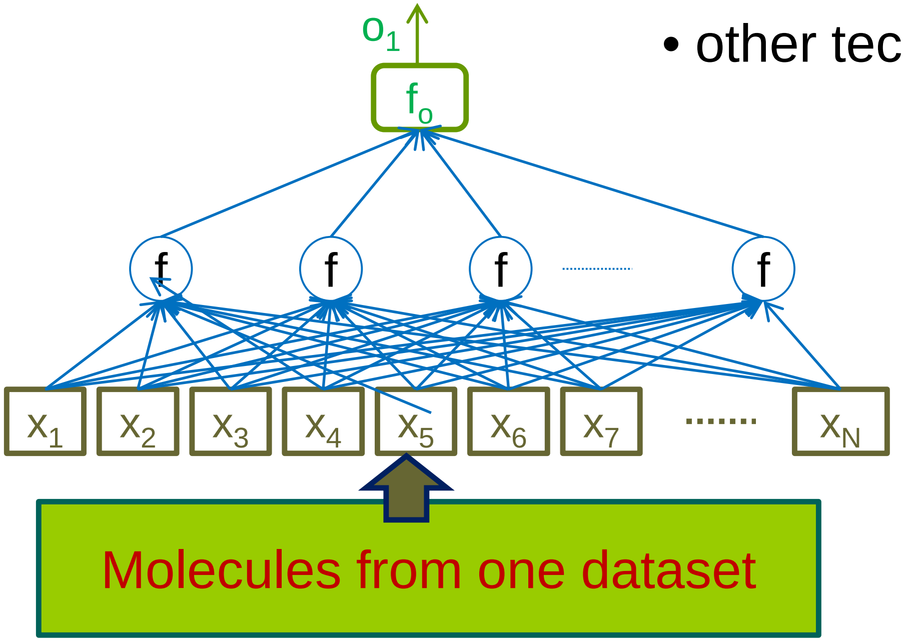
Published: November 23, 2012

Deep Learning (DL) used by the 1st prize winner (*George Dahl, University of Toronto*) beat *Random Forest (RF)*, Merck's internal approach.

Average Correlation *: 0.65 (RF) vs. 0.70 (DL)*

DL is good for QSAR. But is it *revolutionarily* good?

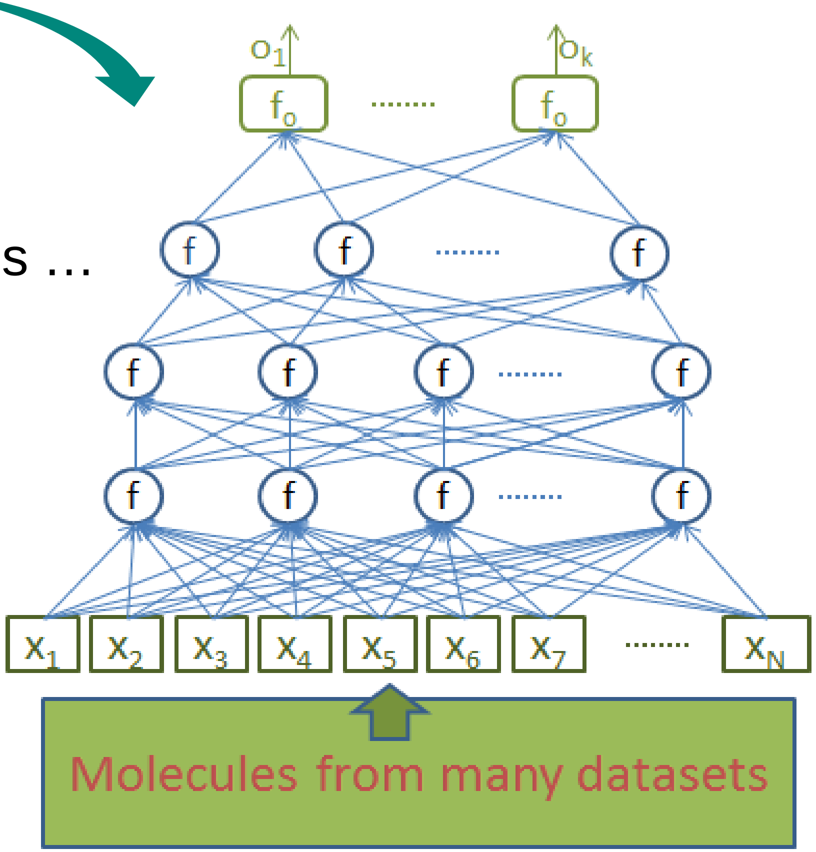
Deep Learning Won Kaggle Competition = Fully-Connected Deep Neural Network (DNN)



Molecules from one dataset

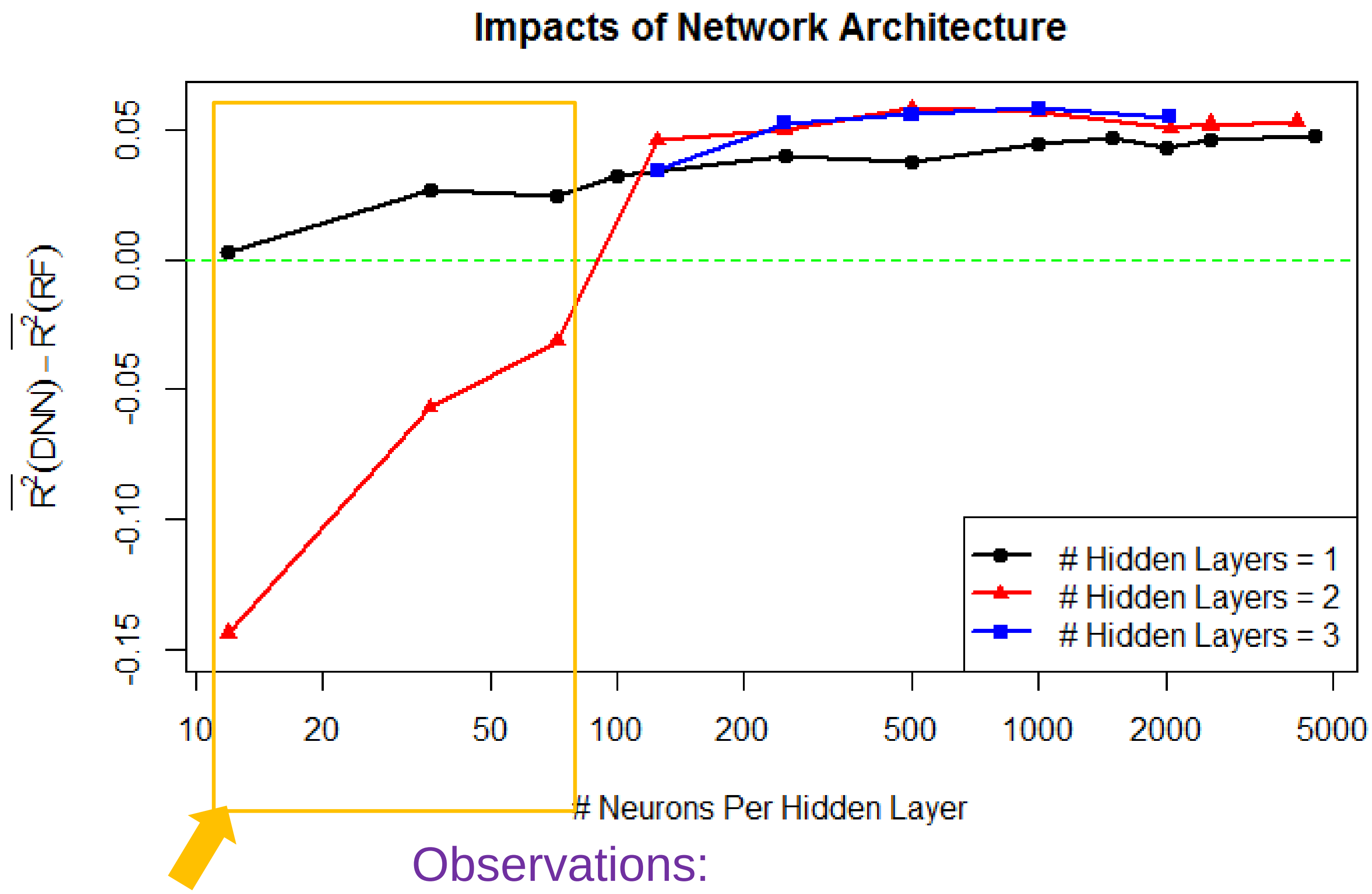
- **Deeper Net**
- **Multi-tasks**
- other techniques …

DNN won Kaggle



Molecules from many datasets

Does “Deep” help?



Neural Net used in the 1980s

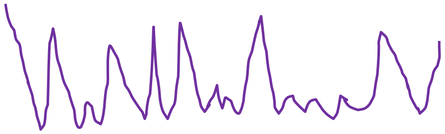
Observations:

1. “Deep” helps, but with a limit, i.e. not > 3-4 layers.
2. “Deeper” requires “wider”

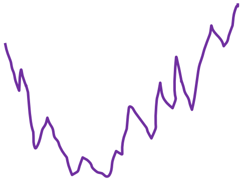
Why "Deep" Helps, But with A Limit?

- Powerful **predictor**
  - ➢ Deep network easily approximates arbitrarily complex prediction functions *
  - ➢ Large and deep network almost guarantees good optimization results **

Large & deep network



vs.

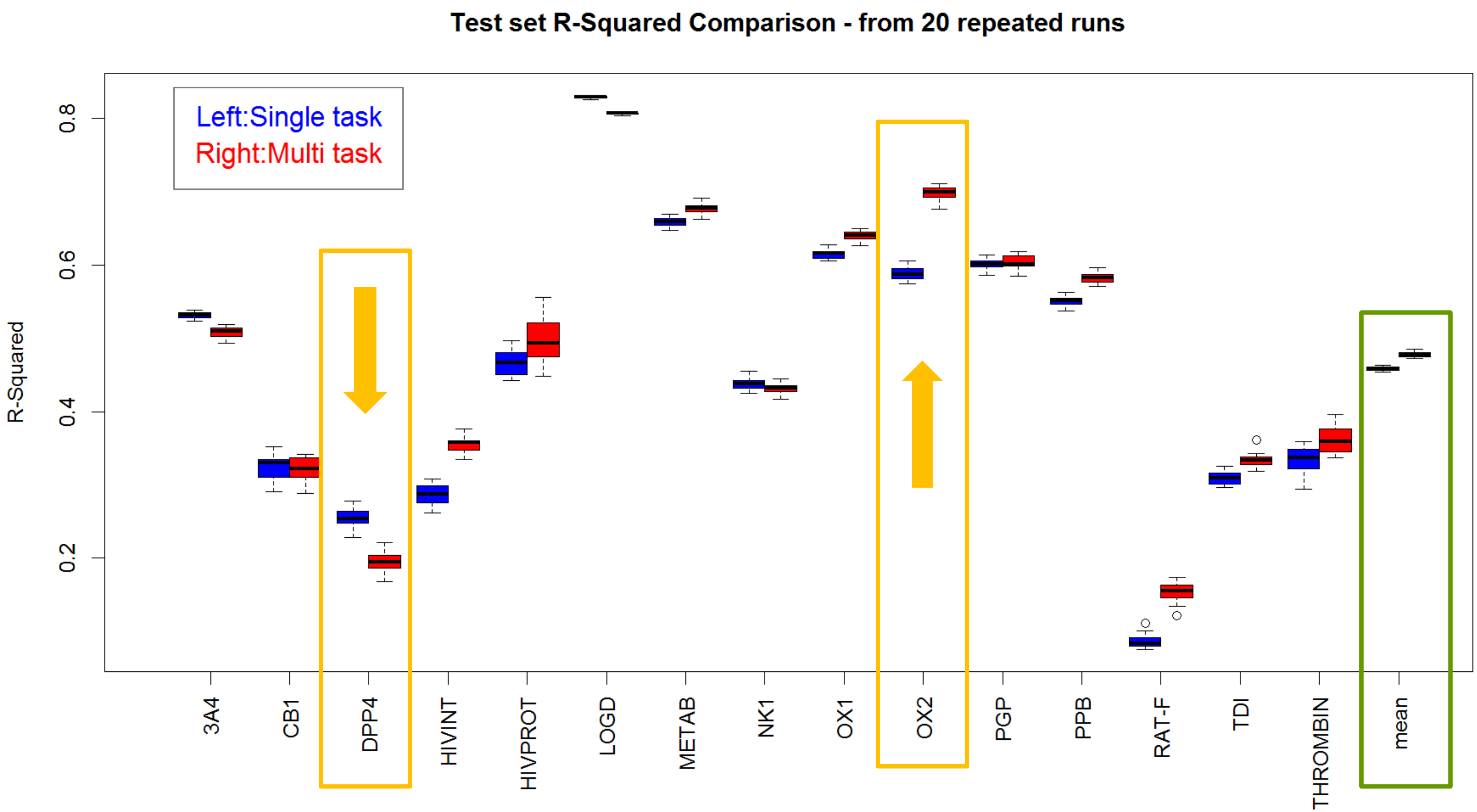


Smaller-scale network

- Ineffective **feature engineering**
  - ➢ QSAR data are molecule descriptors (e.g. AP or DP descriptors, SMILES strings), which are non-redundant, and can defeat DNNs' feature engineering.

**Kurt Hornik (1991); ** Anna Choromanska, et al. (2014)*

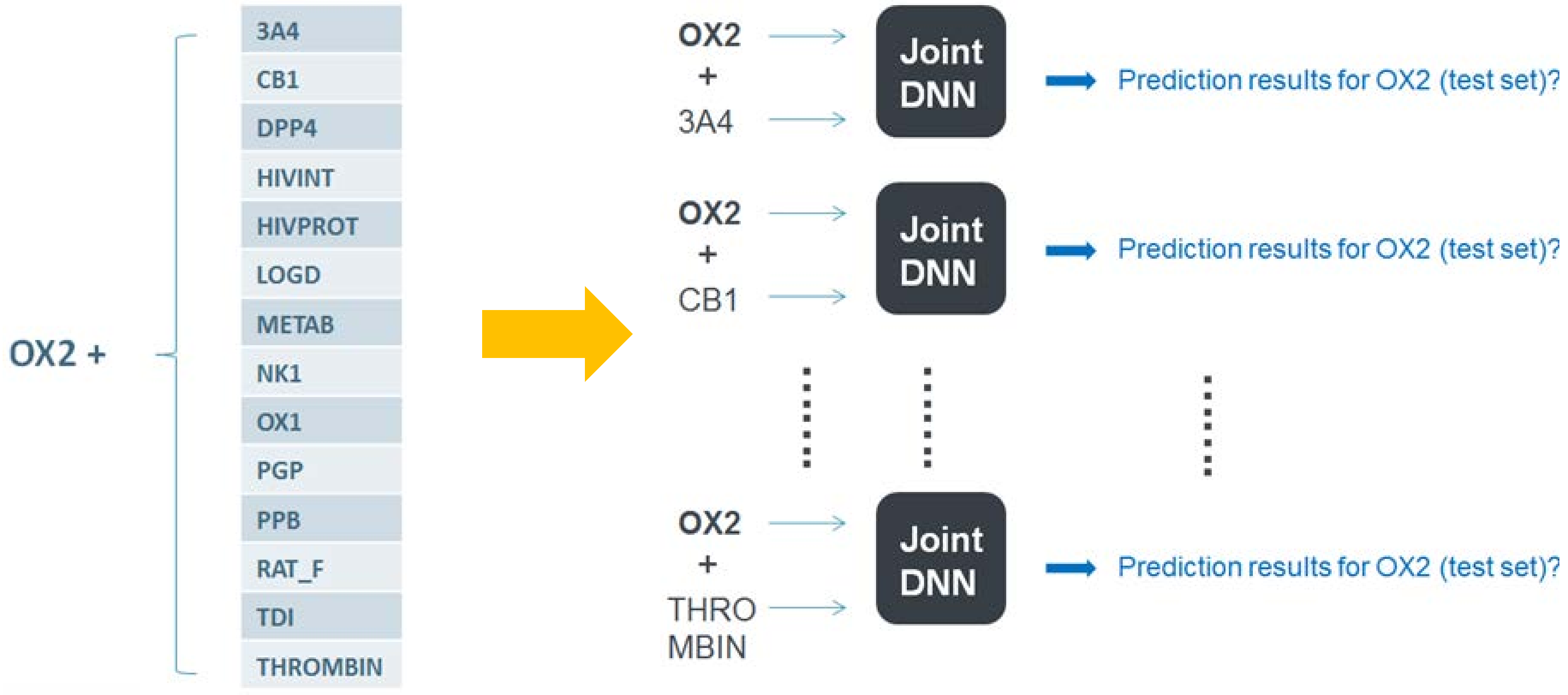
Multi-task DNNs



Box plot reflects the range of a DNN performance due to random initial values.

Public

Pairing OX2 with Each of The Other 14 Datasets



OX2 +

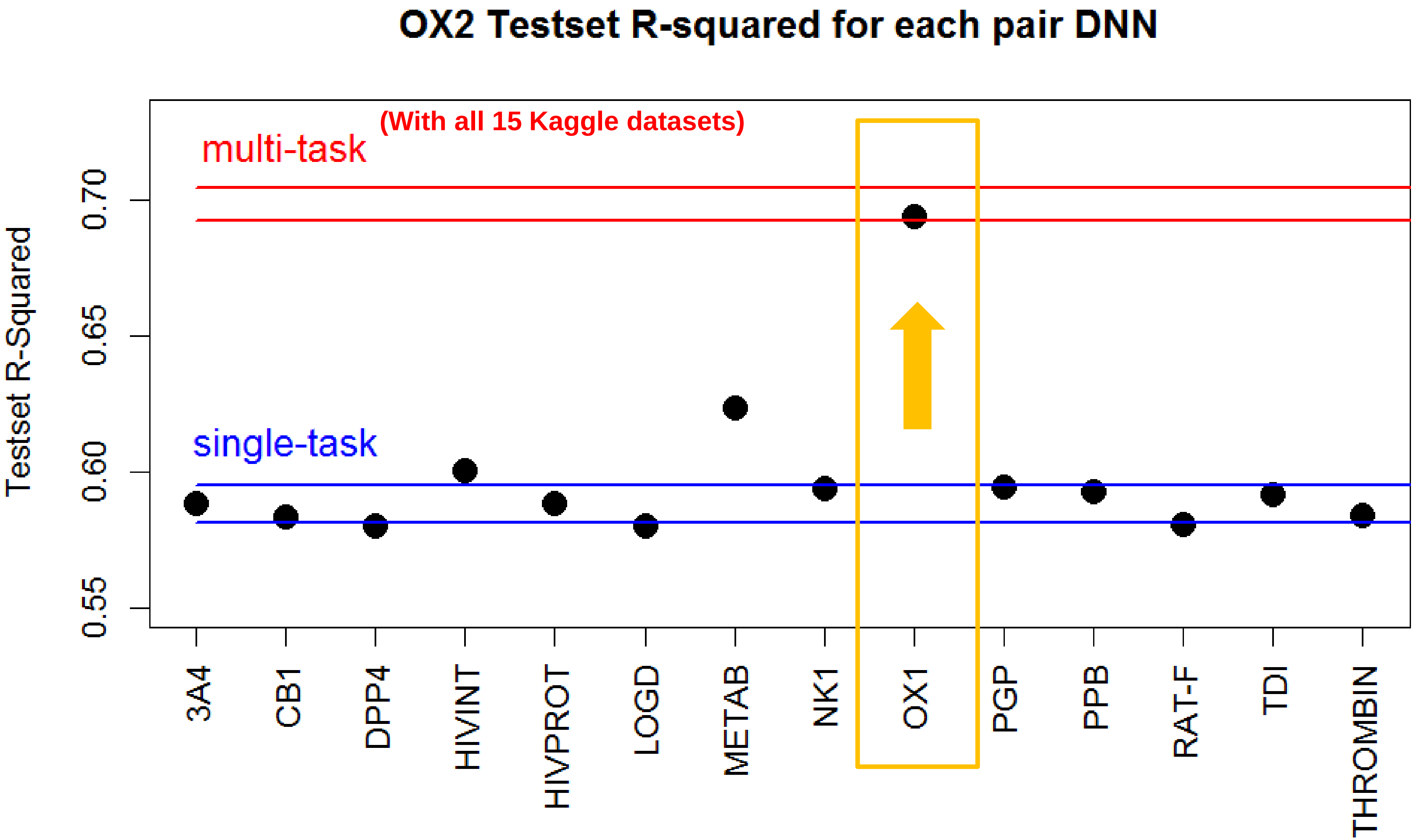
3A4
CB1
DPP4
HIVINT
HIVPROT
LOGD
METAB
NK1
OX1
PGP
PPB
RAT_F
TDI
THROMBIN

OX2 + 3A4 → Joint DNN → Prediction results for OX2 (test set)?

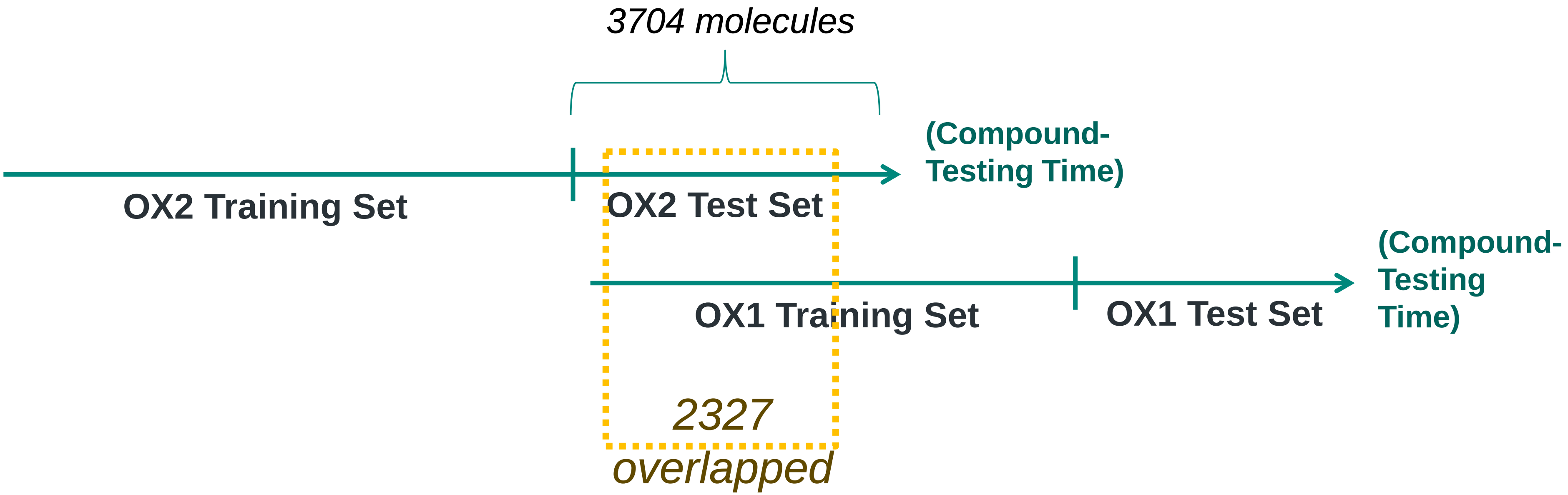
OX2 + CB1 → Joint DNN → Prediction results for OX2 (test set)?

OX2 + THROMBIN → Joint DNN → Prediction results for OX2 (test set)?

OX2 Pairing Results



What happened between OX2 and OX1?



3704 molecules

OX2 Training Set

OX2 Test Set

(Compound-Testing Time)

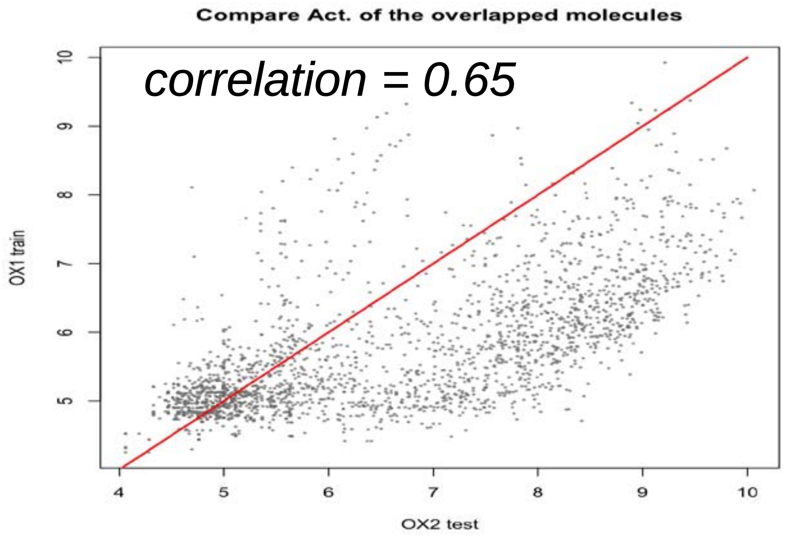
OX1 Training Set

OX1 Test Set

(Compound-Testing Time)

2327 overlapped

The OX1 activities of the 2327 overlapped molecules **positively correlated** with their OX2 activities.



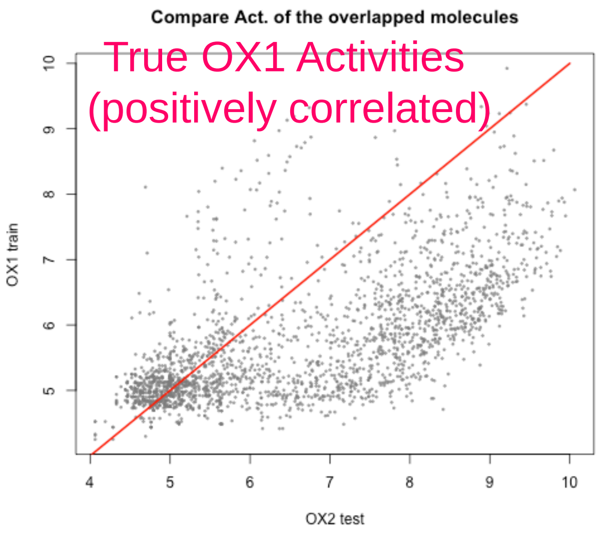
More Questions

- ***Same molecular structure***

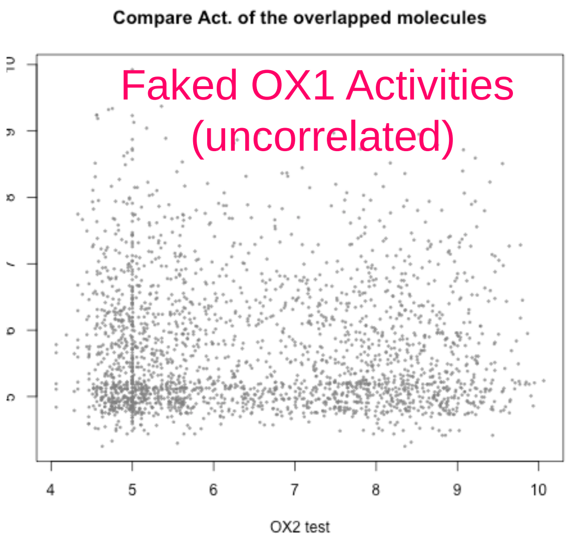
  + *positive- correlated activity* ✓

  + *un- correlated activity* ?

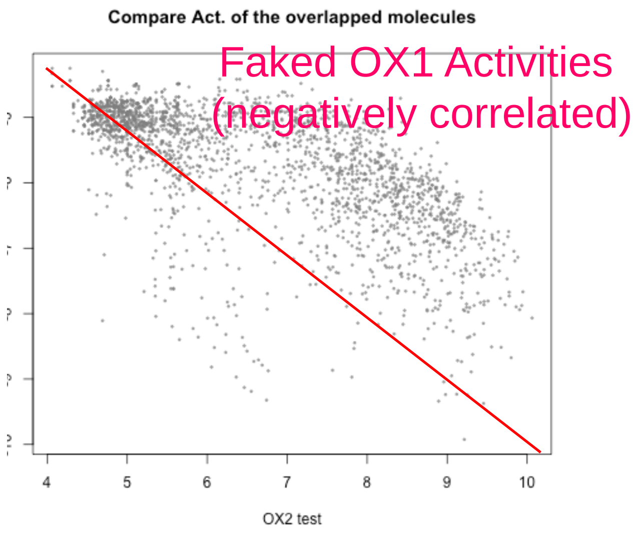
  + *negative- correlated activity* ?


Compare Act. of the overlapped molecules

True OX1 Activities (positively correlated)

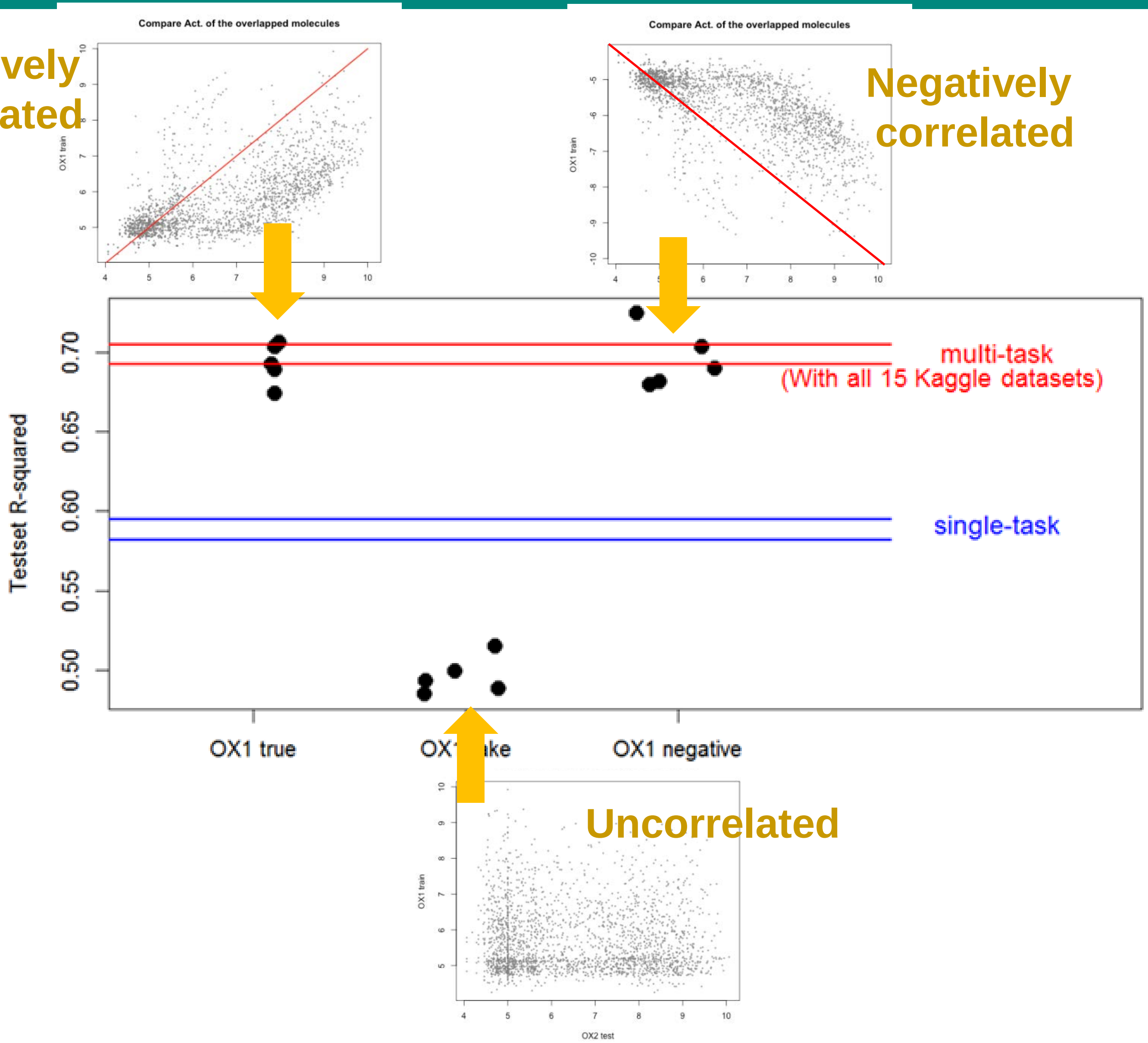

Compare Act. of the overlapped molecules

Faked OX1 Activities (uncorrelated)


Compare Act. of the overlapped molecules

Faked OX1 Activities (negatively correlated)

More Question Answered

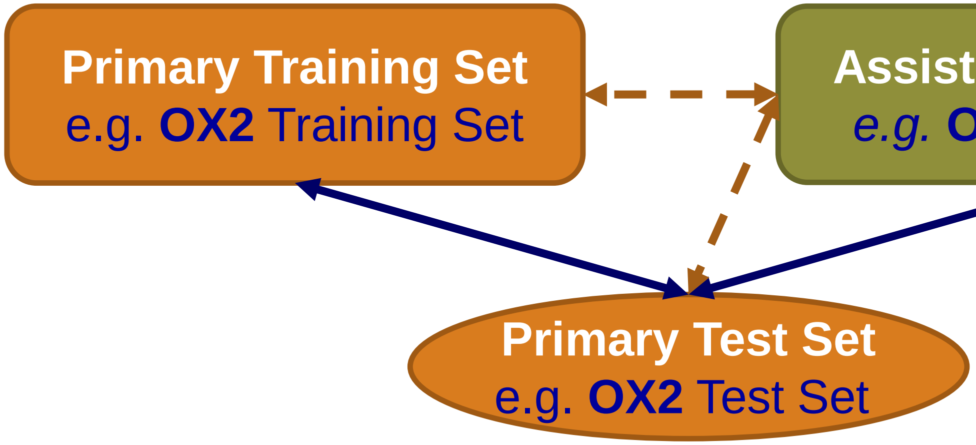


Positively correlated

Negatively correlated

Uncorrelated

Findings regarding Multi-task DNNs



Primary Training Set
e.g. **OX2** Training Set

Assistant Training set
e.g. ***OX1*** *Training Set*

Primary Test Set
e.g. **OX2** Test Set

	Molecular structure ⟷	Molecule Activity ⇠ ⇢	Results
Finding 1	Primary test set molecules are more similar to assistant training set molecules	Primary dataset and assistant dataset have correlated activities (positive or negative)	Improved prediction R^2 for primary test set ⬆
		Uncorrelated biological activities	Decrease prediction R^2 for primary test set ⬇
Finding 2	Primary test set molecules are very different from assistant training set molecules	Correlated or not	No significant change of prediction for primary test set ⬌

Assistant Training Set = Domain Knowledge

- Multi-task DNNs allow us to learn from both the *primary* and an *Assistant Training Set* to boost prediction of the *primary task*, if the *Assistant Training Set is set as:*
  1) *Structure*: identical or very similar to those in the test set of the primary task;
  2) *Activity*: available for experiments related to the primary task.

- Domain knowledge is needed for constructing assistant training sets.

- Multi-task DNNs provide a unique approach for DNNs to incorporate domain expert knowledge.

Summary and Discussion

- Evolutionary vs. Revolutionary: lab-quality reproducibility?

- DNN in its current form is still an evolutionary solution for QSAR.

- Evolutionary ⇒ Revolutionary:
  - Incorporating domain knowledge : Multi-task DNNs can help.
  - Crafting more effective QSAR features: ??