Multiple Task Learning for Quantitative Structure Activity Relationship Learning: Use of a Natural Metric

Presented by: Noureddin Sadawi

Department of Computer Science
Brunel University - London

September 11, 2015

Teams

- **University of Manchester**
  - Prof Ross D. King
  - Dr Ivan Olier
- **Brunel University - London**
  - Dr Larisa Soldatova
  - Dr Crina Grosan
  - Dr Noureddin Sadawi
- **University of Dundee**
  - Prof Andrew Hopkins
  - Dr Jeremy Besnard
  - Dr Richard Bickerton
  - Dr Willem van Hoorn

EPSRC
Engineering and Physical Sciences Research Council
EP/K030582/1

Brunel University London

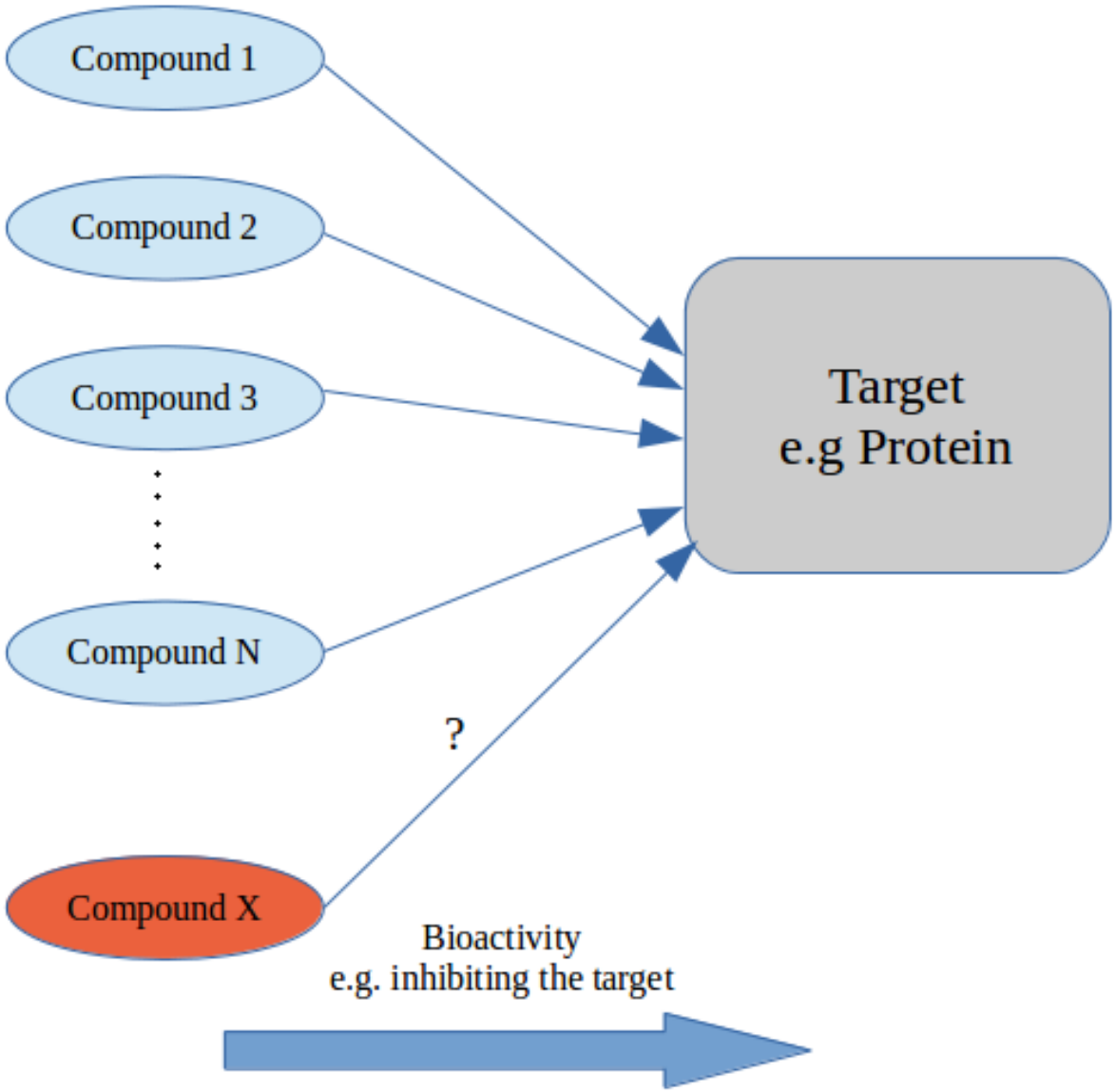
The Physical Problem

- We wish to use small molecules (Drugs) to modulate the biological activity of proteins (Targets), and thereby treat a disease
- Drugs modulate target activity by specifically binding to the target. Binding to other targets may cause side-effects

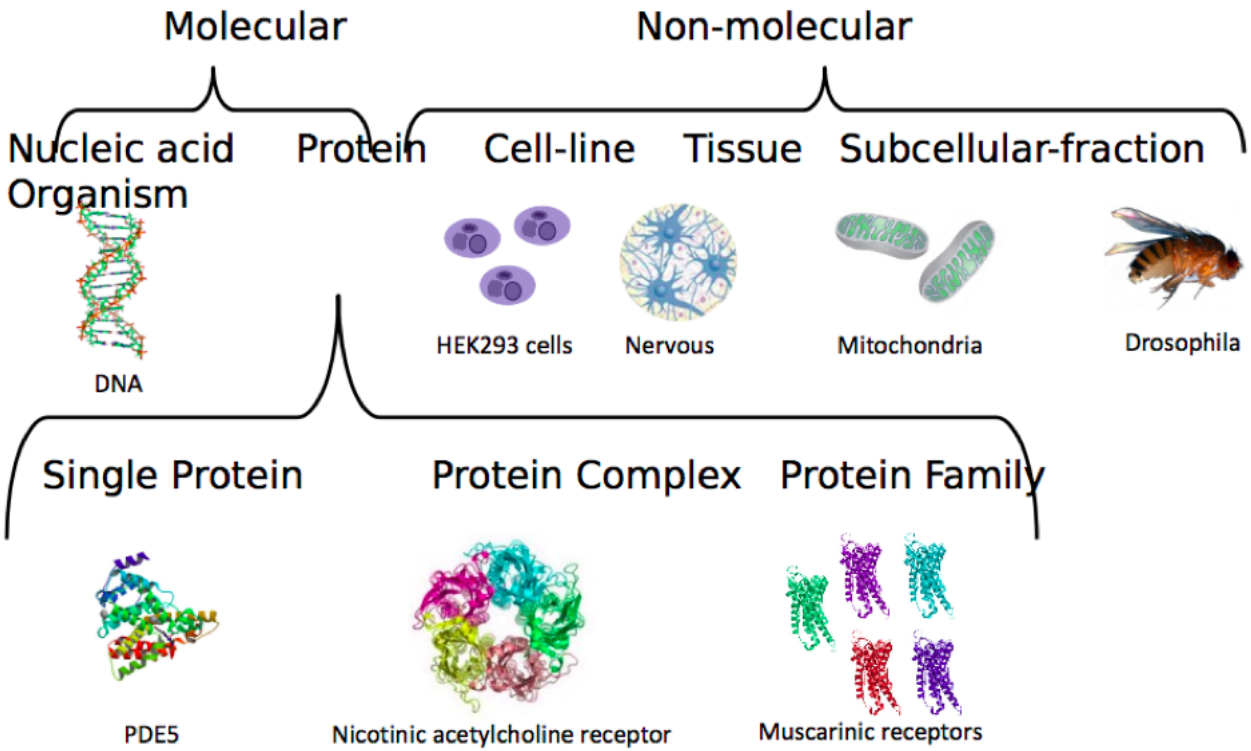
Quantitative Structure-Activity Relationship (QSAR)

- The biological activity of drugs is (largely) dictated by their properties
- Descriptors $\rightarrow$ Mathematical Models $\rightarrow$ Analysis and Prediction of Drug Activity
- Uses a set of molecules whose activity in a particular experiment is known
- Given such set, a QSAR model correlates these activities with properties of molecules in the set (regression)
- Used to guide the synthesis of more potent drugs

Quantitative Structure-Activity Relationship (QSAR)



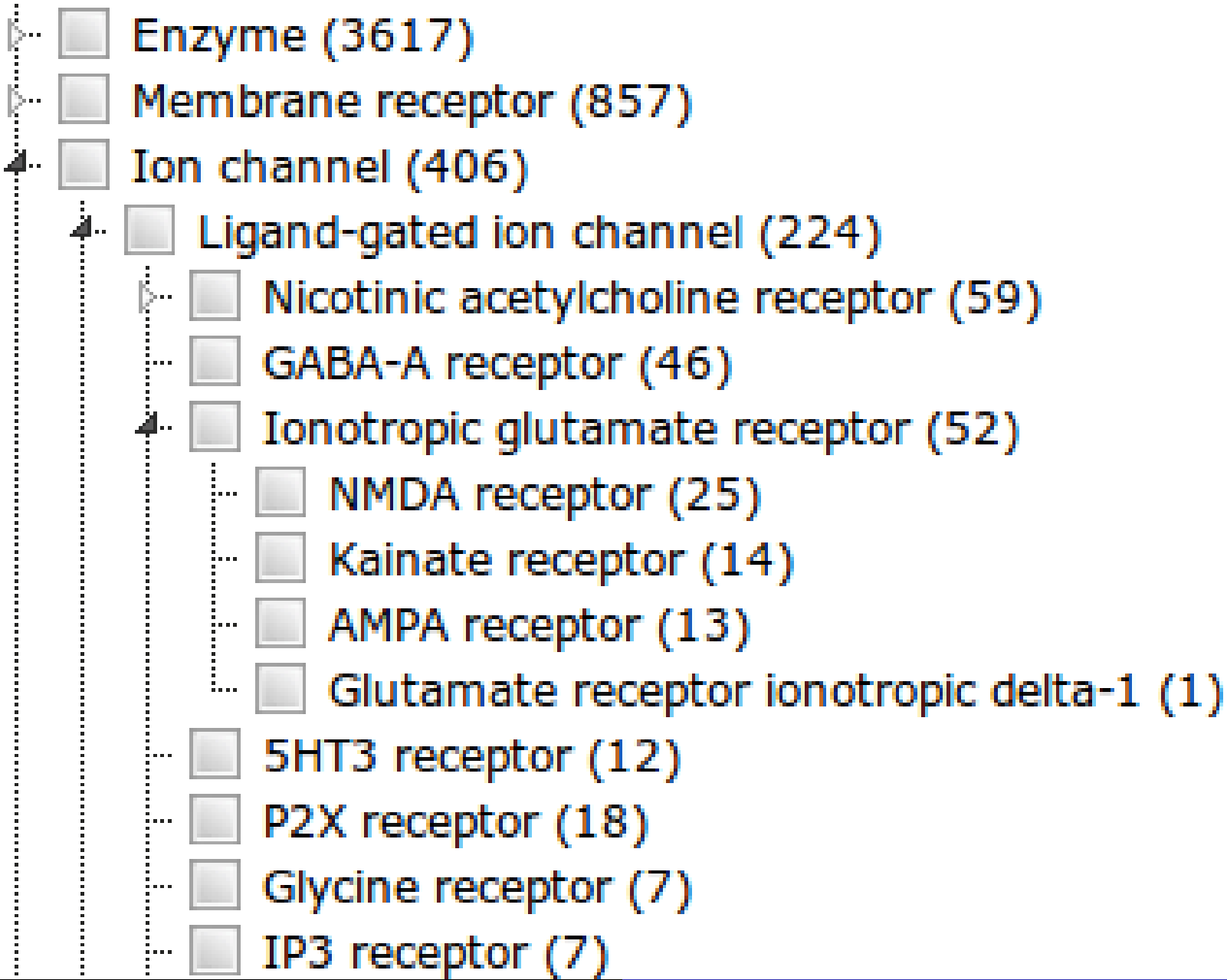
Drug Targets



The ChEMBL Database (v17)

- A freely available and regularly updated resource for drug discovery data (searchable and downloadable)
- Medicinal Chemistry literature is analysed for drug discovery data
- Information on drug targets and the bioactivities of the compounds on those targets
- Currently has information taken from 57,156 publications on: 10,579 targets, 1,411,786 distinct compounds, and 12,843,338 activities
- ChEMBL provides drug target classification/grouping

ChEMBL's Classification of Drug Target



The Similarity of Drug Targets

- Amino acid sequence of drug targets
- Sequence alignment is used to detect regions of similarity between sequences
- Similar sequences imply that targets are 'homologous' *i.e. evolved from a common ancestor*
- Gives a **metric** of evolutionary similarity/distance that ranges between zero and one, with zero indicating no similarity and one indicating complete similarity

Representing Small Molecules

- A large number of ways to represent molecules have been proposed in chemoinformatics:
  - Bulk properties of the molecules (e.g. LogP - Hydrophobicity, pKa - acid/base)
  - **Fingerprints:** 100s-1000s of boolean attributes that represent the presenece or absence of chemical groups
  - 3-dimensional shapes

The Data we have used

- Each dataset represents a drug target (an organism or species)
- We discarded datasets of size less than 10 so we can perform 10 fold cross-validation
- Attributes are 1024-bit fingerprints

MOL_ID	FP_1	FP_2	...	FP_n	Activity
ID_1	1	0	...	1	6.351
ID_2	0	1	...	0	7.534
...	...	...	...	...	...
ID_22	1	1	...	1	8.001
ID_23	0	1	...	0	6.239

The Idea

The Problem:

- Many datasets are too small (quality of model)
- It is too costly to obtain labeled data

The Proposed Solution:

- Use existing data from related targets where labeled data is aplenty
- One way is to use multiple task learning
- Exploit task relatedness
- Incorporate natural metric

Multiple Task Learning

- Learn tasks jointly instead of separately
- Captures relatedness amongst tasks
- Obtain better models

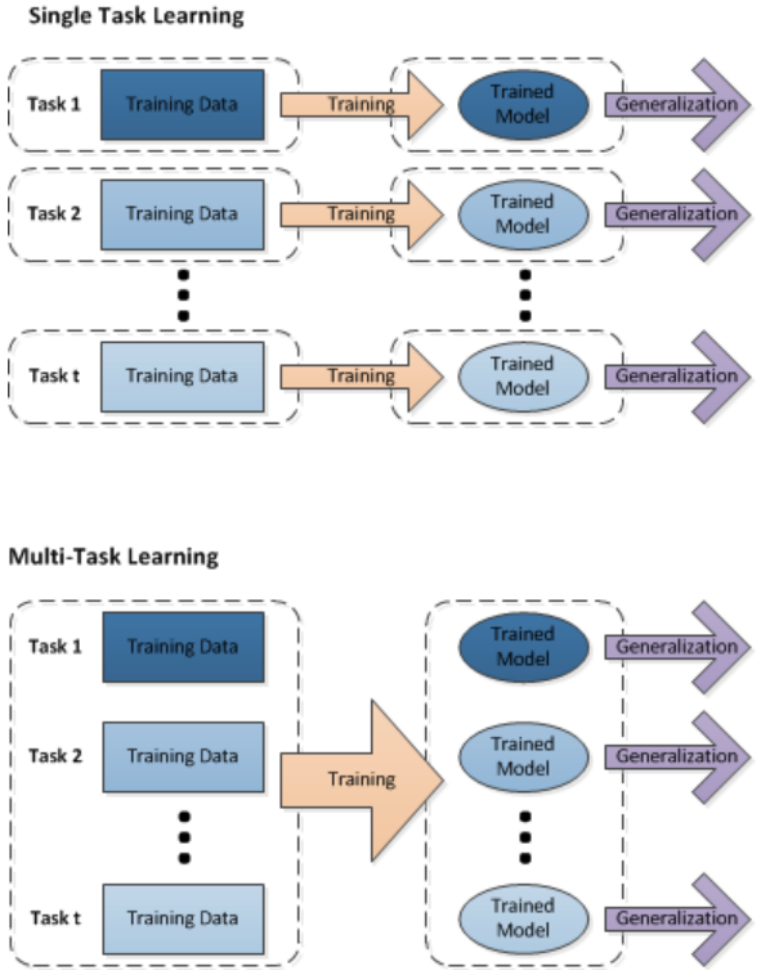


Figure: From SDM 2012 Tutorial by J. Zhou et al

Original Dataset

MOL_ID	FP_1	FP_2	...	FP_n	Activity
ID_1	1	0	...	1	6.351
ID_2	0	1	...	0	7.534
...	...	...	...	...	...
ID_22	1	1	...	1	8.001
ID_23	0	1	...	0	6.239

Table: Typical QSAR Dataset

Single Task Learning (STL)

- We ran Random Forest (100 trees) on each dataset
- The Features we used are FCFP fingerprints of molecules (1024 Boolean attributes)
- We used 10 fold cross-validation to obtain an estimate of the performance for each model
- We computed Root Mean Squared Error (RMSE) as our performance metric
- We performed all experiments using the WEKA 3.7.11 machine learning package

Multiple Task Learning - Setting 1

1. Let us assume we have a drug target group/class with n datasets (each dataset represents a drug target)
2. Concatenate the n datasets into one big dataset
3. Add an indicator variable *TID* to each example to indicate Target ID
4. Perform stratified 10 fold cross validation using the big dataset
   - Observe: the splits are stratified based on *TID*
   - We used Random Forest with 100 trees
5. Filter predictions using *TID*
6. Compute RMSE

Multiple Task Learning - Setting 1 - Datasets

MOL_ID	TID	FP_1	FP_2	...	FP_n	Activity
ID_1	7	1	0	...	1	6.351
ID_2	7	0	1	...	0	7.534
...	...	...	...	...	...	...
ID_111	95	1	1	...	1	8.001
ID_112	95	0	1	...	0	6.239

Table: Dataset for MTL Setting 1

Multiple Task Learning - Setting 2

1. Concatenate the n datasets into one big dataset
2. Add an indicator variable *TID* to each example to indicate Target ID
3. Add n extra variables to the big dataset:
   SimToTID_1, SimToTID_2, ..., SimToTID_n
4. Fill values of these variables using similarities between targets:
   sim(TID,TID_1), sim(TID,TID_2) ... etc
5. Perform stratified 10 fold cross validation using the big dataset
   - Observe: the splits are stratified based on TID
   - We used Random Forest with 100 trees
6. Filter predictions using TID
7. Compute RMSE

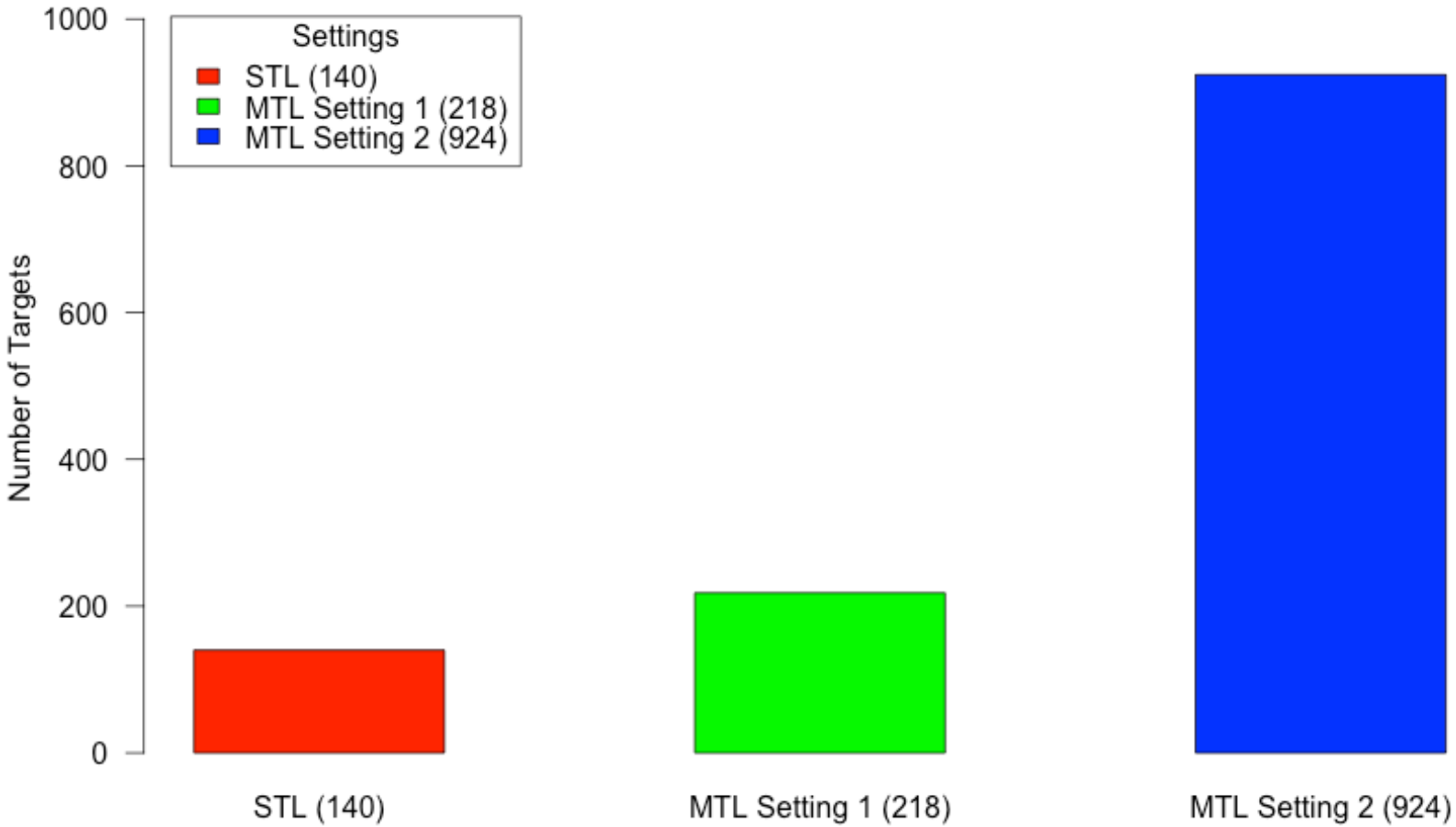
Multiple Task Learning - Setting 2 - Datasets

MOL_ID	TID	SimToTID_7	...	SimToTID_95	FP_1	...	FP_n	Activity
ID_1	7	1		0.584	1	...	1	6.351
ID_2	7	1		0.584	0	...	0	7.534
...	...	...	...	...	...	...	...	...
ID_111	95	0.584	...	1	1	...	1	8.001
ID_112	95	0.584	...	1	1	...	0	6.239

Table: Dataset for MTL Setting 2

Results for L5 Target Classes

Here we count how many targets each algorithms performs better than the other two algorithms



Sign Test for Results for L5 Target Classes

Table: Pair-wise Sign Test for Results for L5 Target Classes

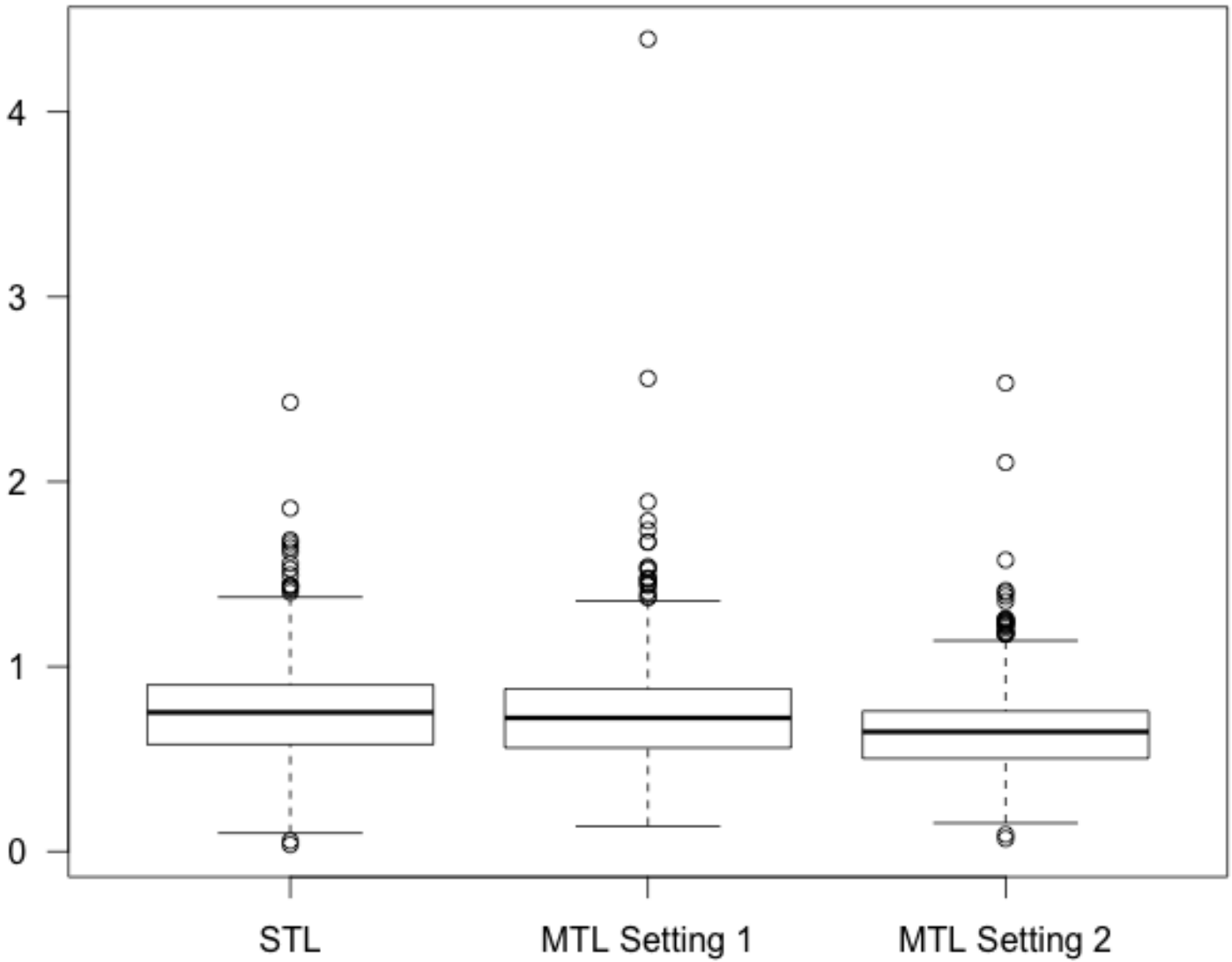
Settings	# +ve	# -ve	# ties
MTL Setting 1 vs STL	782	500	0
MTL Setting 2 vs STL	1081	201	0
MTL Setting 2 vs MTL Setting 1	1043	239	0

A Simple Rank Test for Results for L5 Target Classes

TID	RMSE_STL	RMSE_MTL_1	RMSE_MTL_2
10997	0.933 (3)	0.687 (1)	0.697 (2)
101199	0.997 (3)	0.975 (2)	0.841 (1)
101191	0.805 (3)	0.605 (2)	0.556 (1)
10991	0.936 (3)	0.933 (2)	0.855 (1)
10992	0.680 (1)	0.788 (3)	0.709 (2)
101598	0.622 (3)	0.582 (2)	0.556 (1)
12857	0.711 (1)	1.035 (3)	0.847 (2)
101397	0.267 (3)	0.249 (2)	0.234 (1)
...	...	...	...
AVG RANK	**2.453**	**2.203**	**1.343**

Table: A Simple Rank Test for Results for L5 Target Classes

Boxplot of RMSE Values



Wilcoxon Signed-ranks Test for Results for L5 Target Classes

Table: Pair-wise Wilcoxon Signed-ranks Test for Results for L5 Target Classes

Setting	V	p-value
STL vs MTL Setting 1 medians: 0.752 & 0.722	486824	1.2e-08
STL vs MTL Setting 2 medians: 0.752 & 0.647	743878	2.2e-16
MTL Setting 1 vs MTL Setting 2 medians: 0.722 & 0.647	739764	2.2e-16

Conclusions/Discussion

Conclusions:

- MTL can improve on standard QSAR learning through use of related targets
- MTL QSAR can be improved by incorporating the evolutionary distance of targets

Discussion:

- Do not stratify based on Target ID
- Use distance between targets instead of similarity (distance = 1 - similarity)
- Use distance/similarity between datasets instead of targets