ICGEB-TRAIN
17 May, 2019
Bled

Goldsmiths
UNIVERSITY OF LONDON

META-QSAR AND MULTI-TASK QSAR LEARNING

DR LARISA SOLDATOVA AND THE META-QSAR TEAM

READER IN DATA SCIENCE, DIRECTOR OF MSC DATA SCIENCE ONLINE PROGRAMME

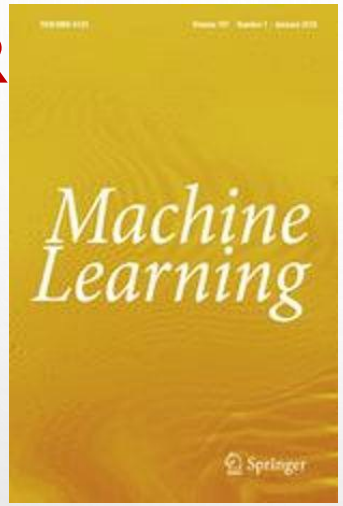
DEPARTMENT OF COMPUTING, GOLDSMITHS, UNIVERSITY OF LONDON

EMAIL: L.SOLDATOVA@GOLD.AC.UK

OUTLINE OF THE TALK

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

PART I: META-QSAR



Ivan Olier, Noureddin Sadawi, G. Richard Bickerton, Joaquin Vanschoren, Crina Grosan, Larisa Soldatova, Ross D. King
Meta-QSAR: a large-scale application of meta-learning to drug design and discovery
2018, *Machine Learning.* 107/1, 285-311.

PART II: MULTI-TASK QSAR LEARNING

New (unpublished)

PART I: META-QSAR

MOTIVATION

DEVELOPING A NEW DRUG IS SLOW AND EXPENSIVE:

- DRUG DEVELOPMENT IS SLOW, GENERALLY TAKING MORE THAN 10 YEARS.
- THE AVERAGE COST TO BRING A NEW DRUG TO MARKET IS ~ 2 BILLION US DOLLARS (DIMASI ET AL. 2015)
- TROPICAL DISEASES SUCH AS MALARIA, SCHISTOSOMIASIS, CHAGAS' DISEASE, ETC., WHICH KILL MILLIONS OF PEOPLE AND INFECT HUNDREDS OF MILLIONS OF OTHERS ARE 'NEGLECTED' (IOSET & CHANG, 2011)
- PRESSURE TO SPEED UP DEVELOPMENT, BOTH TO SAVE LIVES AND REDUCE COSTS.
- A SUCCESSFUL DRUG CAN EARN BILLIONS OF DOLLARS A YEAR, EVEN ONE EXTRA WEEK CAN BE OF GREAT FINANCIAL SIGNIFICANCE

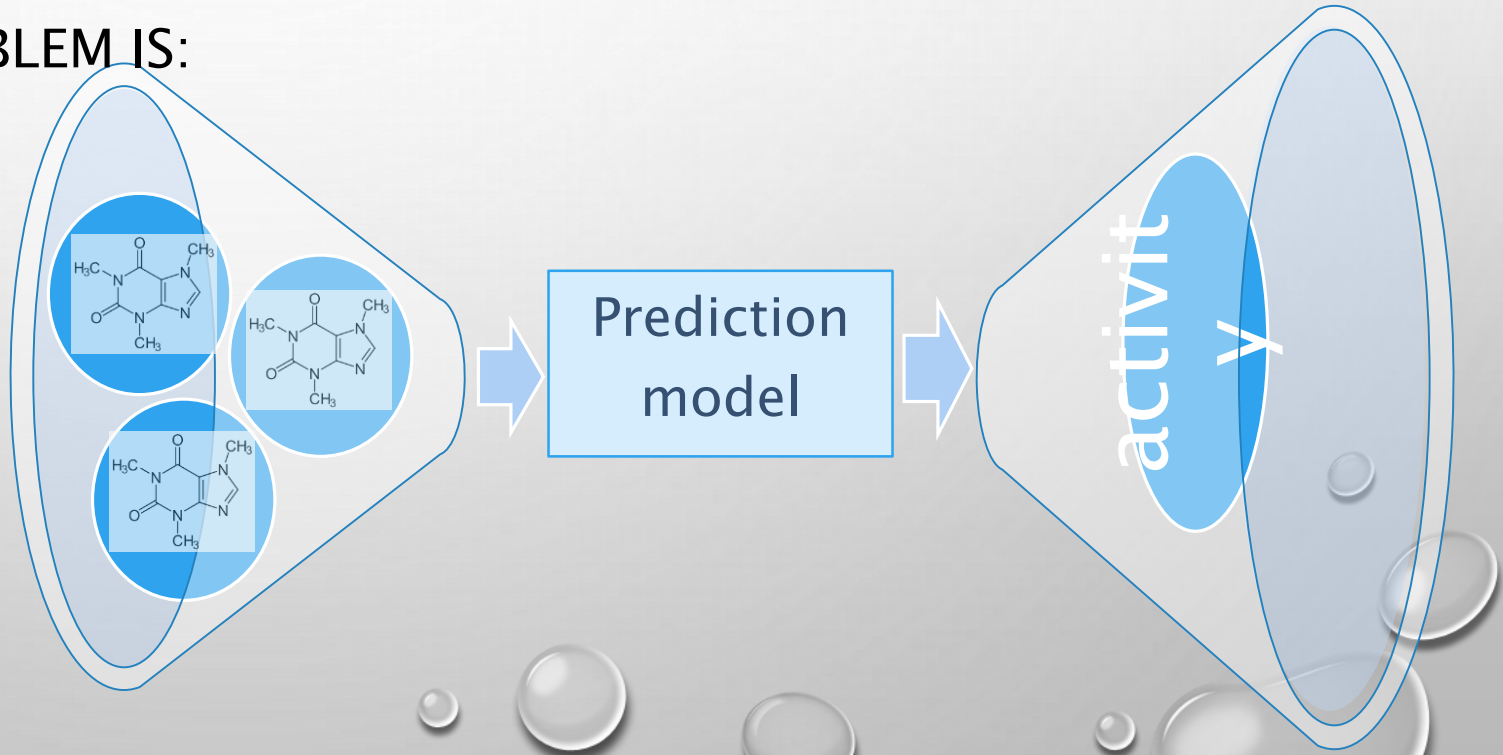
DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2015). The cost of drug development [letter to the editor]. *New England Journal of Medicine*, *372*(20).

Ioset, J. R., & Chang, S. (2011). Drugs for Neglected Diseases initiative model of drug development for neglected diseases: Current status and future challenges. *Future Medicinal Chemistry*

QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP (QSAR)

- QSAR IS A FUNCTION THAT PREDICTS A COMPOUND'S BIOACTIVITY FROM ITS STRUCTURE
- THE STANDARD QSAR LEARNING PROBLEM IS:

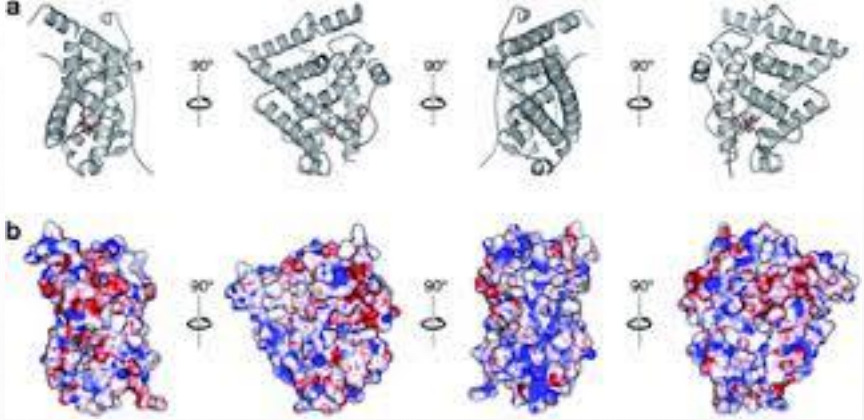
GIVEN A TARGET (USUALLY A PROTEIN) AND A SET OF CHEMICAL COMPOUNDS (SMALL MOLECULES) WITH ASSOCIATED BIOACTIVITIES (E.G. INHIBITING THE TARGET), LEARN A PREDICTIVE MAPPING FROM MOLECULAR REPRESENTATION TO ACTIVITY.



QSAR LEARNING – ‘NO FREE LUNCH’

- ALMOST EVERY FORM OF STATISTICAL AND MACHINE LEARNING METHOD HAS BEEN APPLIED TO LEARNING QSARS
- THERE IS NO AGREED SINGLE BEST WAY OF LEARNING QSARS
- META-LEARNING: WHAT LEARNING IS BETTER IN WHAT SCENARIOS
- MOTIVATION: TO UNDERSTAND THE PERFORMANCE CHARACTERISTICS OF THE MAIN (BASELINE) MACHINE LEARNING METHODS CURRENTLY USED IN QSAR LEARNING.

THE RATIONAL FOR META-QSAR LEARNING



META-QSAR LEARNING SHOULD BE SUCCESSFUL BECAUSE ALTHOUGH ALL THE DATASETS HAVE THE SAME OVERALL STRUCTURE, THEY DIFFER IN:

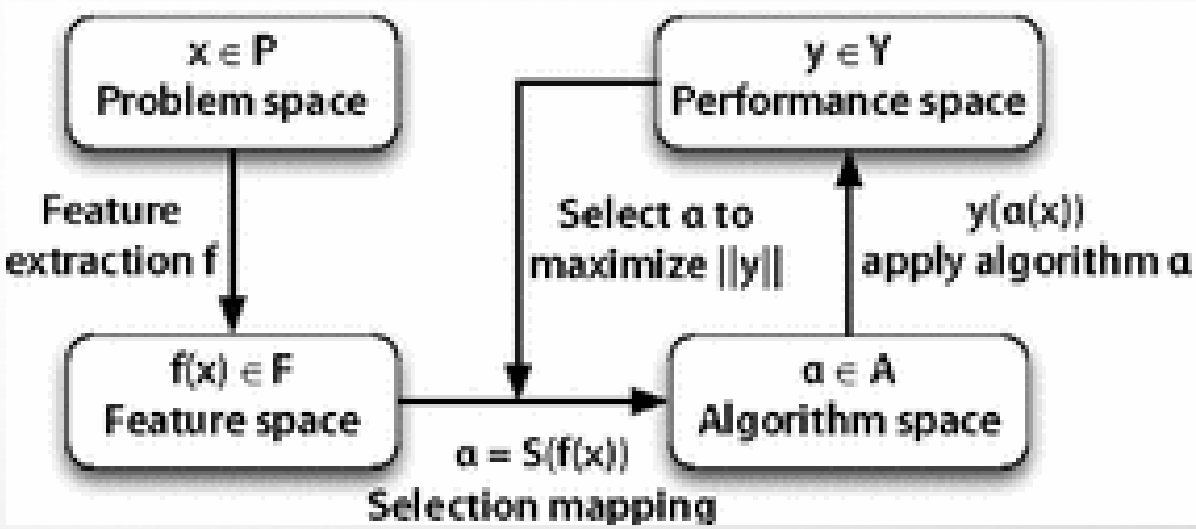
- THE NUMBERS OF DATA POINTS (TESTED CHEMICAL COMPOUNDS),
- IN THE RANGE AND OCCURRENCE OF FEATURES (COMPOUND DESCRIPTORS), AND
- IN THE TYPE OF CHEMICAL/BIOCHEMICAL MECHANISM THAT CAUSES THE BIOACTIVITY.

THESE DIFFERENCES INDICATE THAT DIFFERENT MACHINE LEARNING METHODS ARE TO BE USED FOR DIFFERENT KINDS OF QSAR DATA.

META-LEARNING APPROACH

- **META LEARNING** IS A SUBFIELD OF MACHINE LEARNING WHERE LEARNING ALGORITHMS ARE APPLIED ON METADATA ABOUT MACHINE LEARNING EXPERIMENTS (SCHAUL, 2010).
- A GENERAL FRAMEWORK FOR **ALGORITHM SELECTION**:

1. PROBLEM SPACE *P:* IN OUR CASE THE SPACE OF 8292 QSAR DATASETS
2. FEATURE SPACE *F:* EACH QSAR DATASET IN *P* HAS A SET OF MEASURABLE CHARACTERISTICS (META-FEATURES)
3. ALGORITHM SPACE *A:* BASE-LEVEL LEARNING ALGORITHMS, IN OUR CASE A SET OF 18 REGRESSION ALGORITHMS.
4. PERFORMANCE SPACE *Y:* EMPIRICALLY MEASURED PERFORMANCE, E.G. RMSE OF EACH ALGORITHM *A* ON EACH OF THE QSAR



Rice's framework for algorithm selection.
Rice, J. R. (1976). The algorithm selection problem. *Advances in Computers*, *15*, 65118.

Schaul, Tom; Schmidhuber, Jürgen (2010). "Metalearning". Scholarpedia. 5 (6): 4650

META-LEARNING TASK

THE TASK IS:

- FOR ANY GIVEN QSAR PROBLEM $x \in P$, SELECT THE BEST COMBINATION OF QSAR AND MOLECULAR REPRESENTATION $a \in A$ THAT MAXIMIZES A PREDEFINED PERFORMANCE MEASURE $y \in Y$.

TWO META-LEARNING APPROACHES

- **CLASSIFICATION PROBLEM**:

TO LEARN A MODEL THAT CAPTURES THE RELATIONSHIP BETWEEN THE PROPERTIES OF THE QSAR DATASETS (META-DATA) AND THE PERFORMANCE OF THE REGRESSION ALGORITHMS;

TO PREDICT THE MOST SUITABLE ALGORITHM FOR A NEW DATASET

- **RANKING PROBLEM**: TO FIT A MODEL THAT RANKS THE QSAR COMBINATIONS BY THEIR PREDICTED PERFORMANCES.

BASELINE QSAR LEARNING: ALGORITHMS

- SELECTED 18 REGRESSION ALGORITHMS, INCLUDING LINEAR REGRESSION, SUPPORT VECTOR MACHINES, ARTIFICIAL NEURAL NETWORKS, REGRESSION TREES, AND RANDOM FORESTS.
- EXPERIMENTED WITH BASELINE REGRESSION ALGORITHMS TO INVESTIGATE THEIR EFFECTIVENESS ON QSAR PROBLEMS

BASELINE QSAR LEARNING: DATASETS

THE CHEMBL DATABASE: HTTPS://WWW.EBI.AC.UK/CHEMBL/

- THE DATA: INFORMATION ON THE DRUG TARGETS, THE STRUCTURES OF THE TESTED COMPOUNDS, THE BIOACTIVITIES OF THE COMPOUNDS ON THEIR TARGETS.

THE KEY ADVANTAGES OF USING CHEMBL FOR META-QSAR ARE:

(a) IT COVERS A LARGE SPACE OF TARGETS

(b) THE DIVERSITY OF THE CHEMICAL SPACE

(c) THE HIGH QUALITY OF THE INTERACTION DATA.

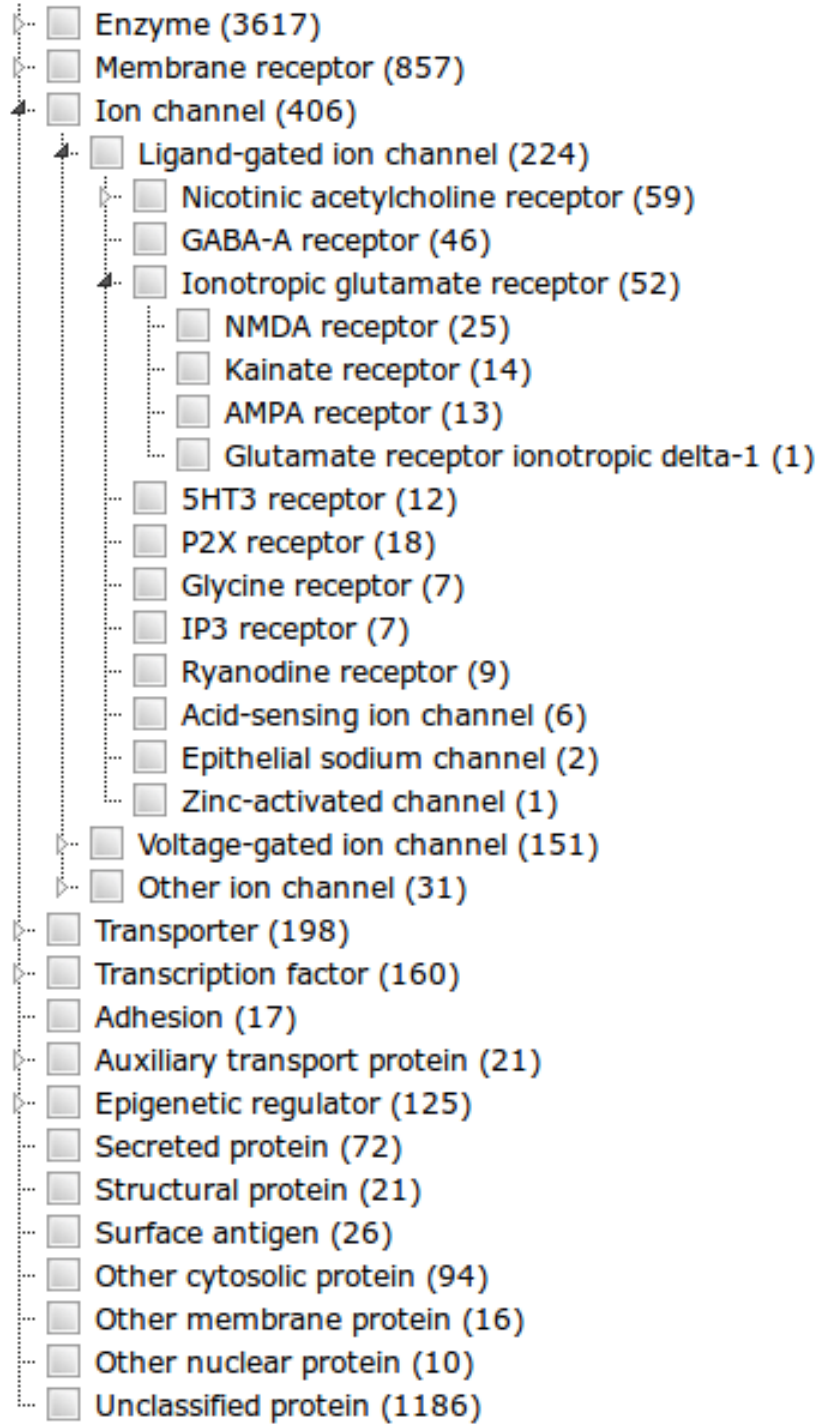
THE MAIN WEAKNESS: FOR SOME TARGETS, INTERACTION DATA ON ONLY A RELATIVELY SMALL NUMBER OF COMPOUNDS

BASELINE QSAR LEARNING: DATASETS

- WE EXTRACTED 2764 TARGETS FROM CHEMBL
- THE NUMBER OF CHEMICAL COMPOUNDS PER TARGET: FROM 10 TO ABOUT 6000
- ASSOCIATED BIOACTIVITIES: IC50, EC50, KI, KD AND THEIR EQUIVALENTS
- BIOACTIVITIES HAVE BEEN NORMALISED BY OUR COLLABORATORS FROM THE UNIVERSITY OF DUNDEE.
- THE SIMPLIFIED MOLECULAR-INPUT LINE-ENTRY SYSTEM (SMILES) REPRESENTATION

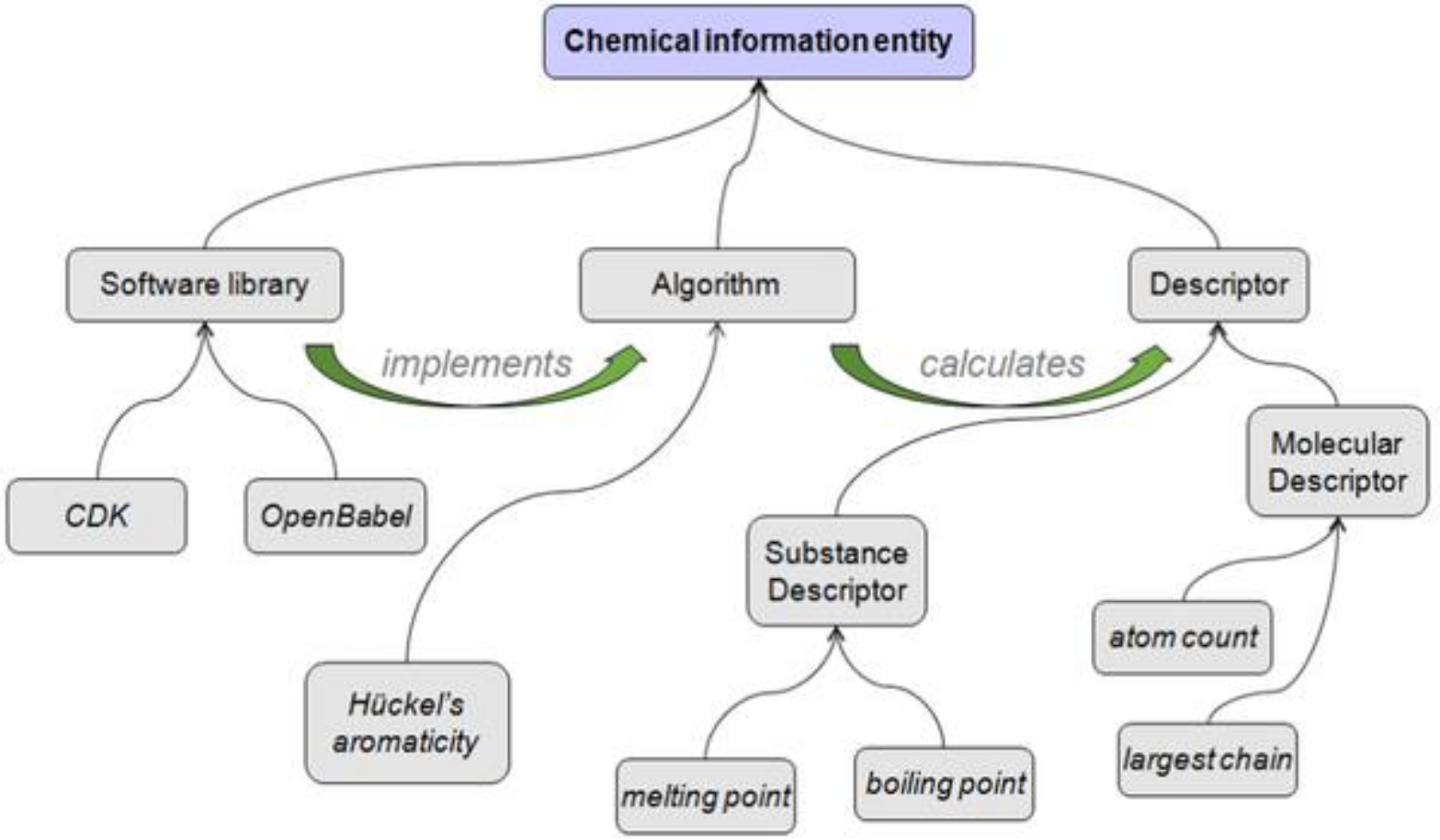
CHEMBL'S CLASSIFICATION OF DRUG TARGETS

- **DRUG TARGET CLASSES**: THE CHEMBL DATABASE CURATORS HAVE CLASSIFIED PROTEIN TARGETS INTO A MANUALLY CURATED FAMILY.
- THE 6-LEVEL HIERARCHY IN CHEMBL20
- FOCUS ON L5
- **DRUG TARGET GROUPINGS**: BASED ON THE PRACTICE THAT INDIVIDUAL PROTEINS CAN BE DESCRIBED BY A RANGE OF DIFFERENT IDENTIFIERS AND TEXTUAL DESCRIPTIONS.
- WE USED 468 DRUG TARGET GROUPS, WITH 2-21 DRUG TARGETS IN A GROUP



MOLECULAR DESCRIPTORS

- THE CHEMINF ONTOLOGY FORMALIZES CHEMINFORMATICS COMPUTATION (HASTINGS ET AL, 2011)
- WE USED **DRAGON** SOFTWARE TO CALCULATE DESCRIPTIONS FROM SMILES WWW.TALETE.MI.IT
- 1447 MOLECULAR DESCRIPTORS
- REDUCED TO 43 BASIC DESCRIPTORS



Hastings et al (2011) The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web. PLOS One

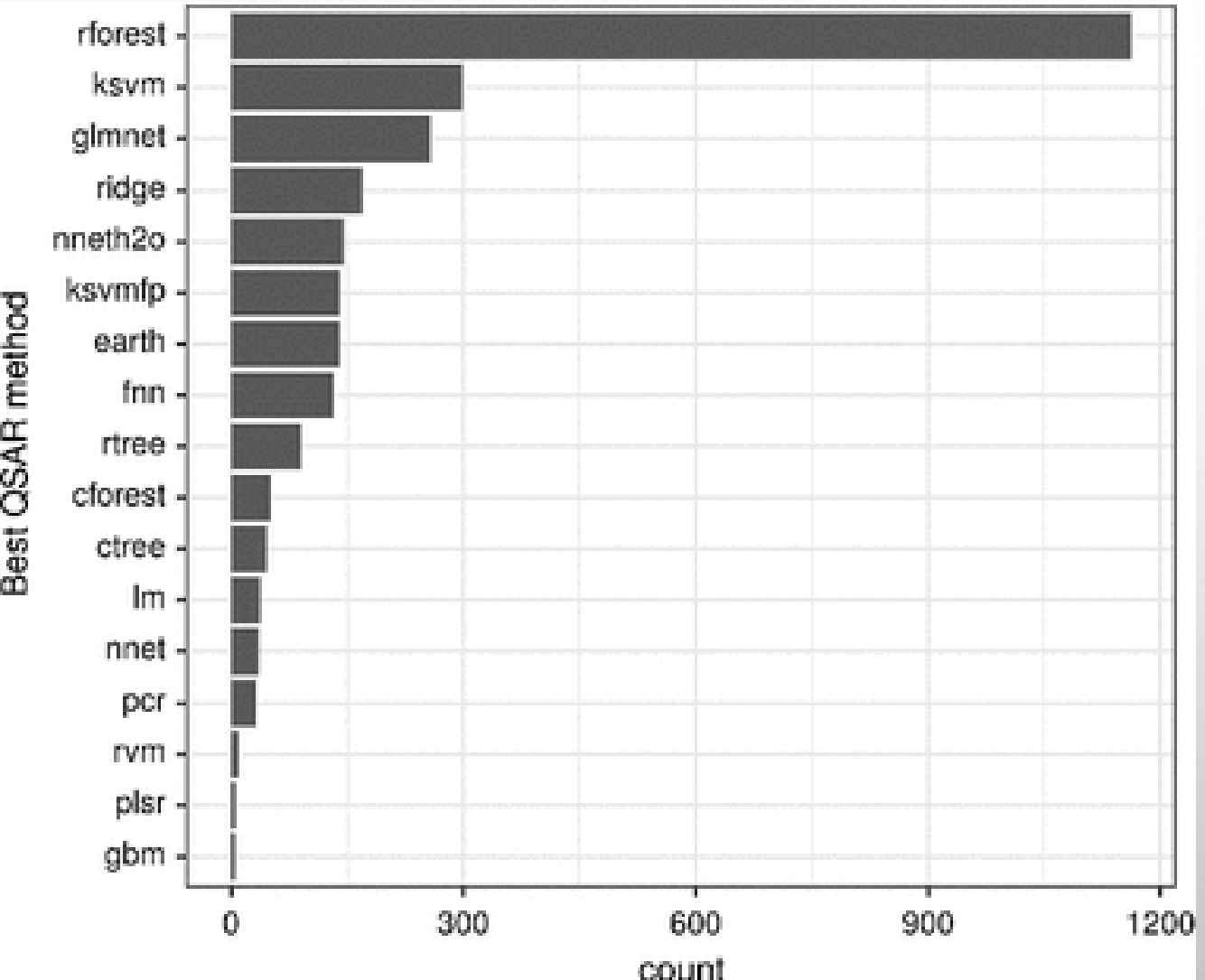
REPRESENTATIONS

WE CONSIDERED THREE REPRESENTATIONS FOR 2764 TARGETS:

- BASIC REPRESENTATION WITH 43 DESCRIPTORS
- FULL REPRESENTATION WITH 1447 MOLECULAR DESCRIPTORS
- FCFP4 FINGERPRINT REPRESENTATION USING THE PIPELINE PILOT SOFTWARE FROM BIOVIA

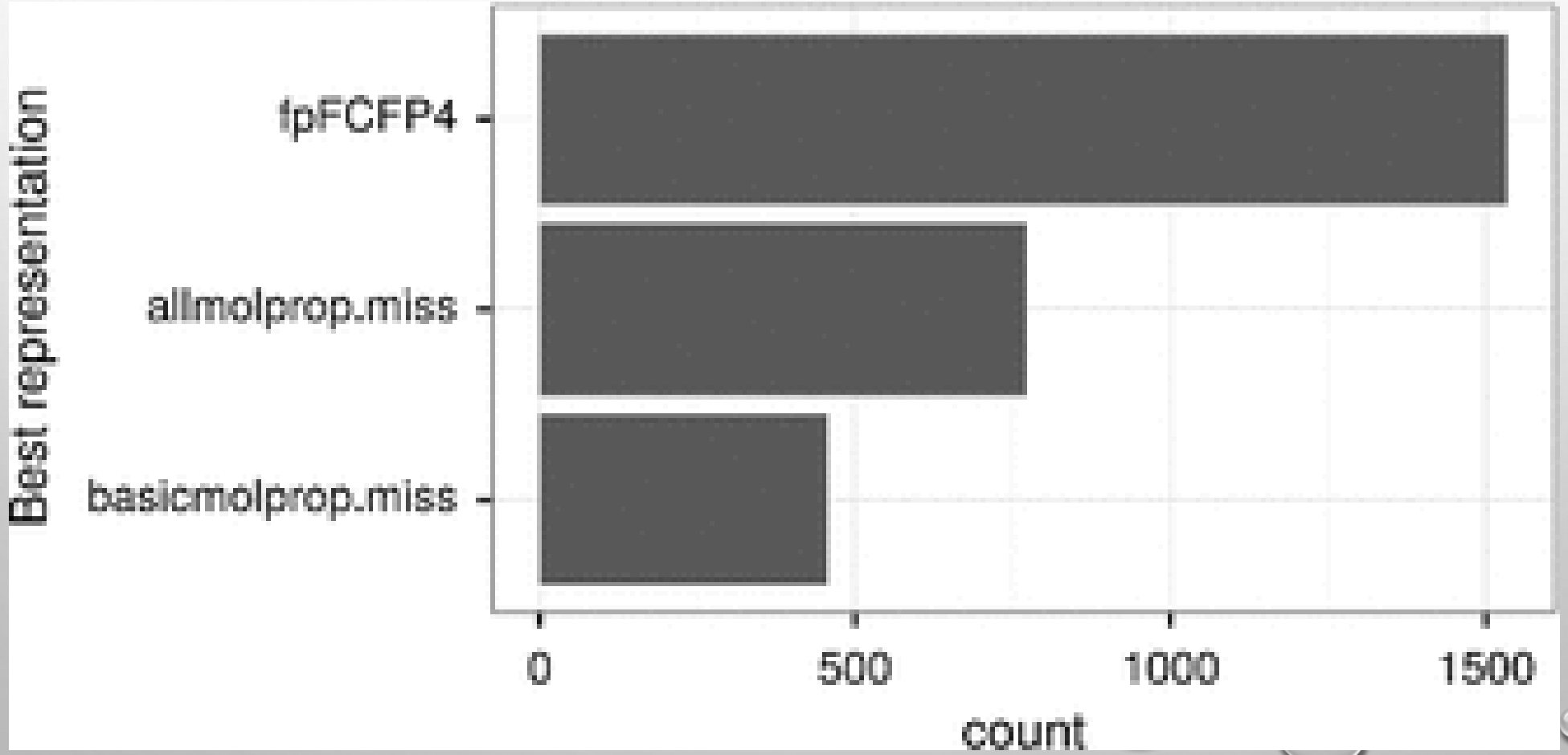
TOTAL **8292 DATASETS**

BASELINE QSAR EXPERIMENTS



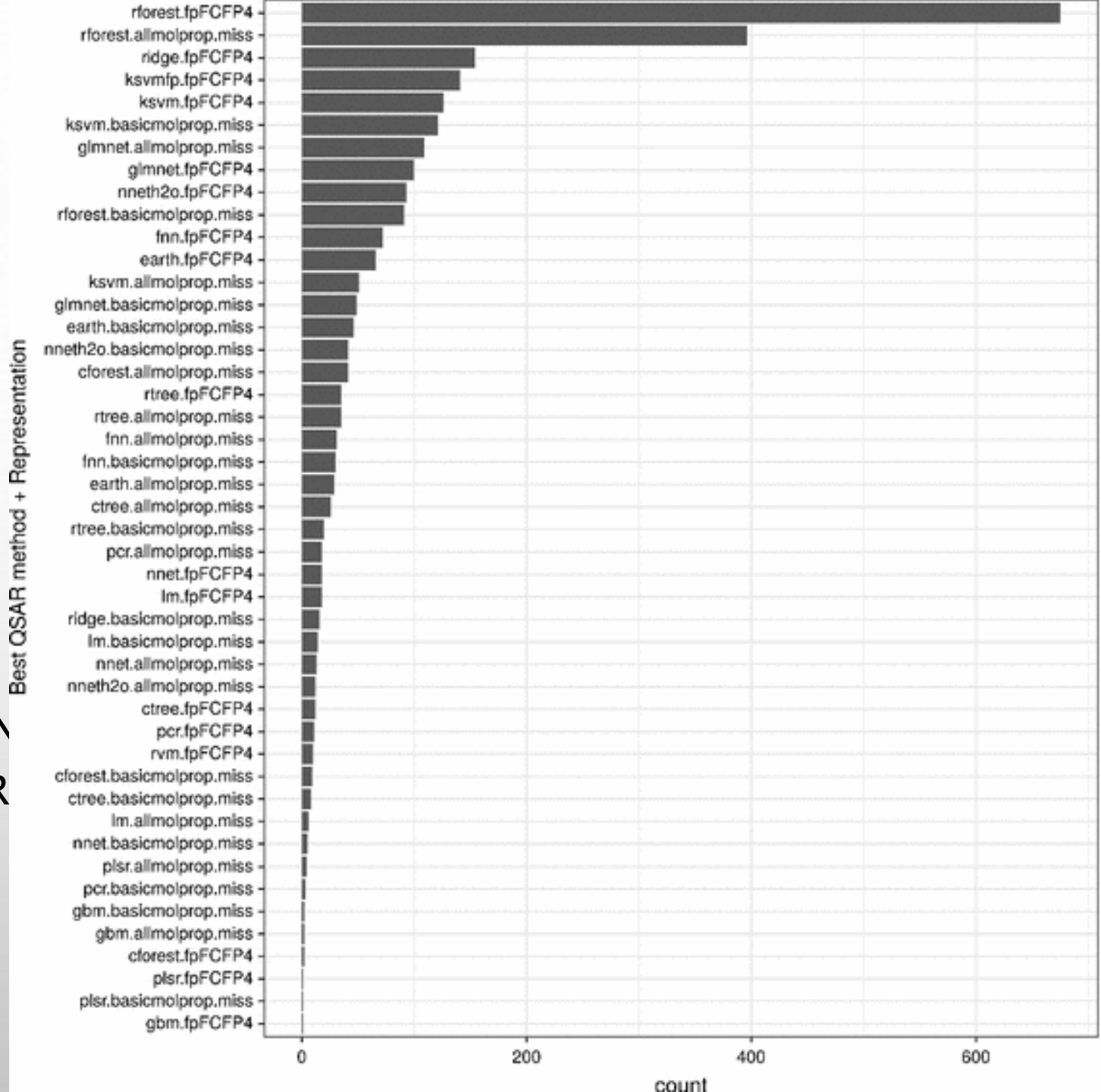
- RANDOM FOREST ('RFOREST') WAS THE BEST PERFORMER IN 1162 TARGETS (OUT OF 2764)
- SVM ('KSVM') – 298 TARGETS
- GLM-NET ('GLMNET') – 258 TARGETS

DATASET REPRESENTATIONS



BASELINE QSAR EXPERIMENTS WITH REPRESENTATIONS

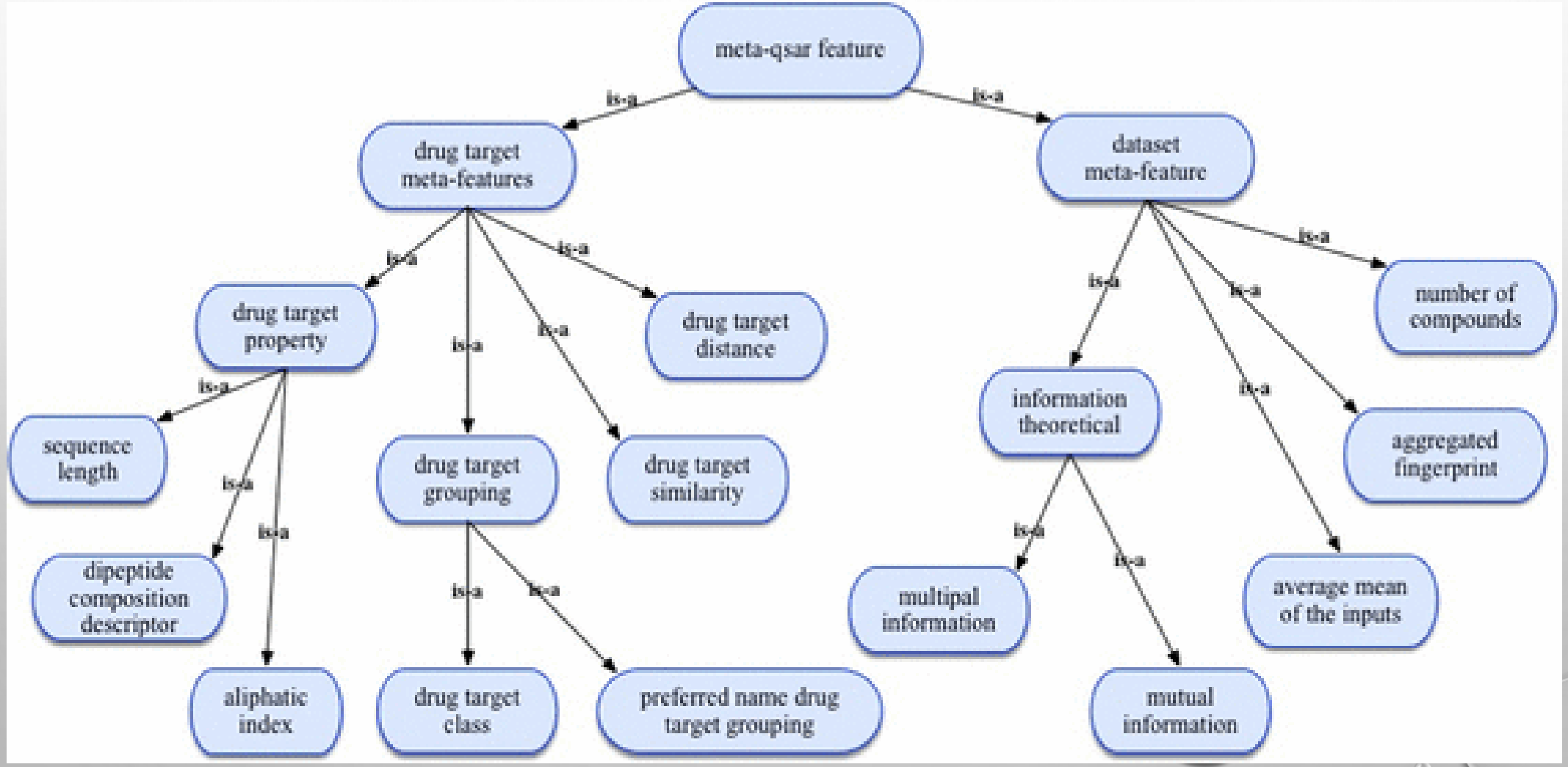
- RANDOM FOREST WITH FCFP4 FINGERPRINTS OR ALL MOLECULAR PROPERTIES WERE THE MOST SUCCESSFUL (675 AND 396 OUT OF 2764 TARGETS, RESPECTIVELY).
- REGRESSION WITH RIDGE PENALISATION AND SVM WITH TANIMOTO KERNEL WERE ALSO SUCCESSFUL WHEN USING THE FCFP4 FINGERPRINT (154 AND 141, RESPECTIVELY).



META-FEATURES FOR META-QSAR LEARNING

- META-LEARNING ANALYSIS REQUIRES A SET OF META-FEATURES
- WE USED AS META-FEATURES, CHARACTERISTICS OF THE DATASETS CONSIDERED IN THE BASE STUDY AND DRUG TARGET PROPERTIES
- UTILISED A SIMILAR APPROACH EMPLOYED BY CHEMINF ONTOLOGY TO FORMALLY DEFINE META-FEATURES

META-QSAR ONTOLOGY



DATASET META-FEATURES

- MULTIPLE INFORMATION (ALSO CALLED TOTAL CORRELATION) AMONG THE RANDOM VARIABLES IN THE DATASET
- MUTUAL INFORMATION BETWEEN NOMINAL ATTRIBUTES X AND Y. DESCRIBES THE REDUCTION IN UNCERTAINTY OF Y DUE TO THE KNOWLEDGE OF X, AND LEANS ON THE CONDITIONAL ENTROPY $H(Y|X)$
- AVERAGE STANDARD DEVIATION OF THE FEATURES
- SKEWNESS OF THE RESPONSE VARIABLE
- …..

DRUG TARGET META-FEATURES

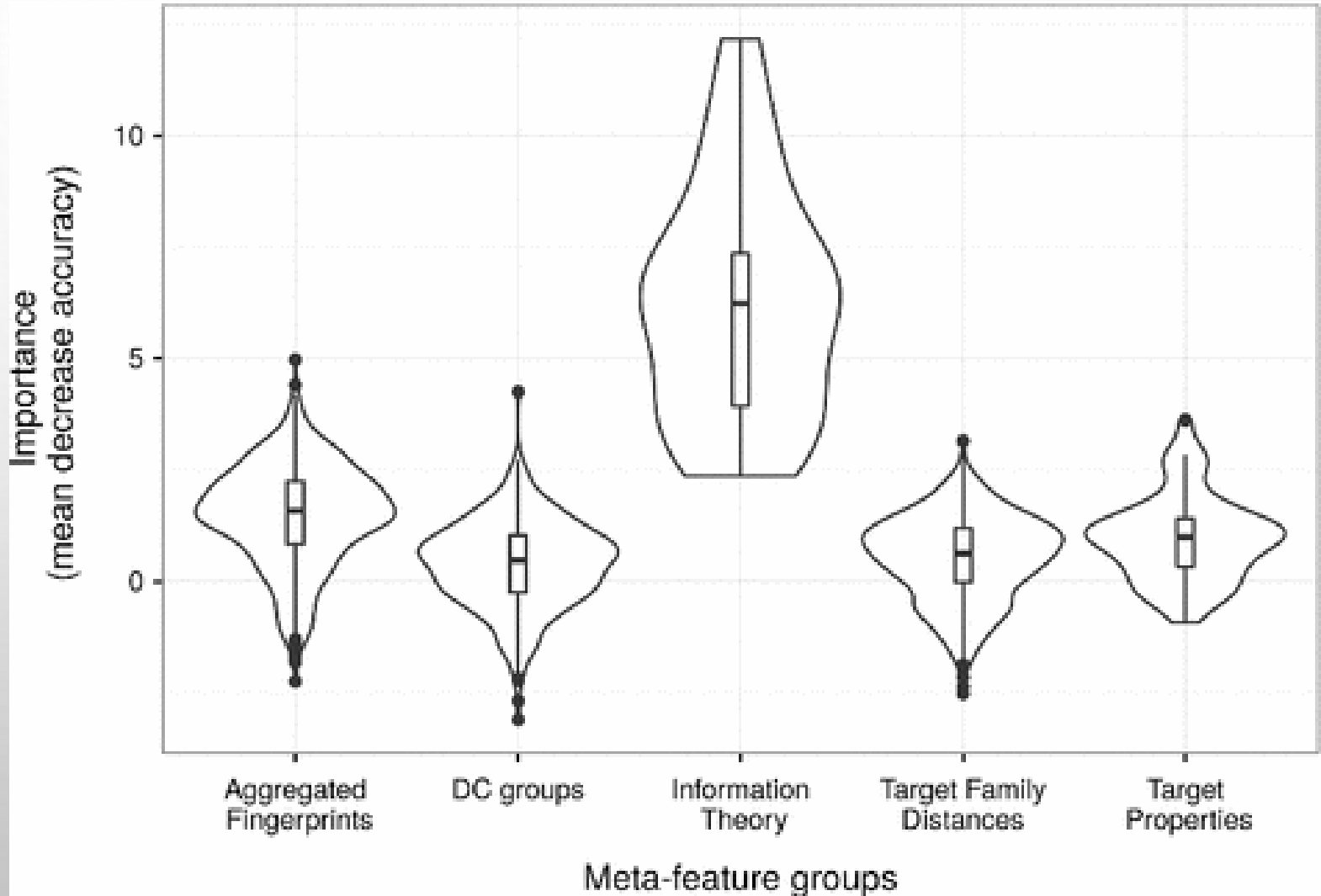
- MOLECULAR WEIGHT – RATIO OF THE MASS OF A MOLECULE TO THE UNIFIED ATOMIC MASS UNIT. SOMETIMES CALLED THE MOLECULAR WEIGHT OR RELATIVE MOLAR MASS
- SEQUENCE LENGTH – THE NUMBER OF AMINO ACIDS IN A PROTEIN SEQUENCE
- HYDROPHOBICITY – THE ASSOCIATION OF NON-POLAR GROUPS OR MOLECULES IN AN AQUEOUS ENVIRONMENT WHICH ARISES FROM THE TENDENCY OF WATER TO EXCLUDE NON-POLAR MOLECULES (NOTE: THERE ARE 38 VARIANTS OF HYDROPHOBICITY)
- THE INSTABILITY INDEX – A PROTEIN WHOSE INSTABILITY INDEX IS SMALLER THAN 40 IS PREDICTED AS STABLE, A VALUE ABOVE 40 PREDICTS THAT THE PROTEIN MAY BE UNSTABLE
- THE ALIPHATIC INDEX – THE RELATIVE VOLUME OCCUPIED BY ALIPHATIC SIDE CHAINS (ALANINE, VALINE, ISOLEUCINE, AND LEUCINE).

DRUG TARGET GROUPINGS

- WE ALSO USED DRUG TARGET GROUPINGS, E.G. 'DRUG TARGET CLASSES', AND 'THE PREFERRED NAME GROUPINGS', AS META-FEATURES.
- WE USED THE 6-LEVEL CHEMBL HIERARCHY TREE TO COMPUTE DISTANCES BETWEEN TARGET FAMILIES AS META-FEATURES FOR THE META-QSAR LEARNING.

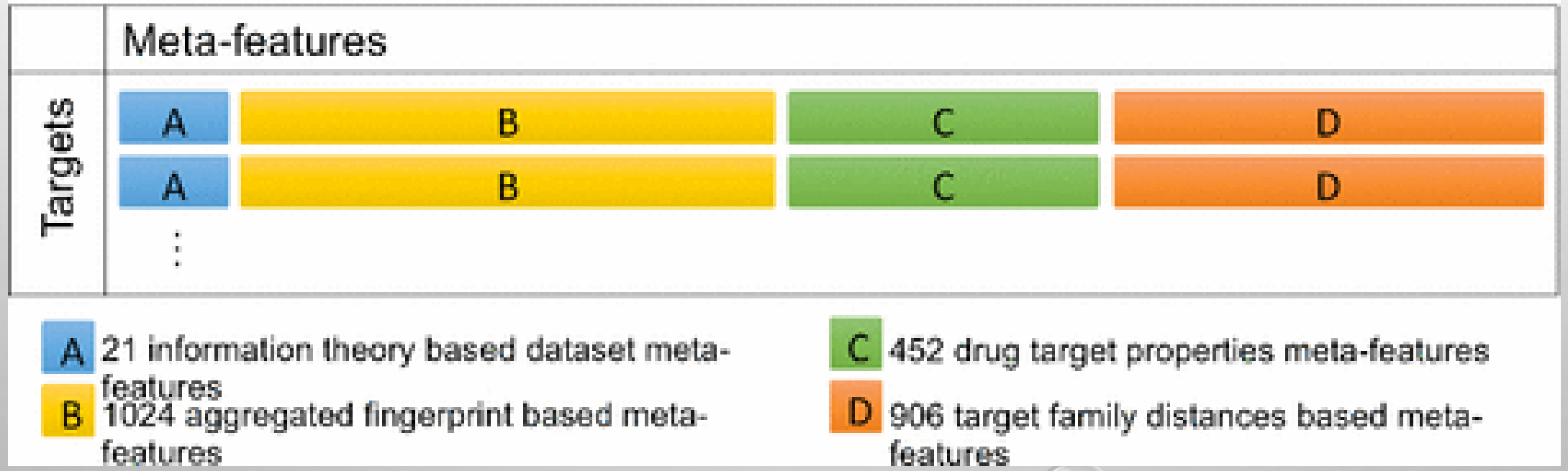
THE IMPORTANCE OF EACH META-FEATURE IN THE CLASSIFICATION TASK

- WE USED THE ALL-CLASSES RANDOM FOREST IMPLEMENTATION TO ESTIMATE THE IMPORTANCE OF EACH META-FEATURE IN THE CLASSIFICATION TASK, AS ESTIMATED USING THE MEAN DECREASE ACCURACY.
- THE INFORMATION THEORY GROUP IS MOST INFLUENTIAL
- ALL GROUPS CONTRIBUTED TO THE TASK



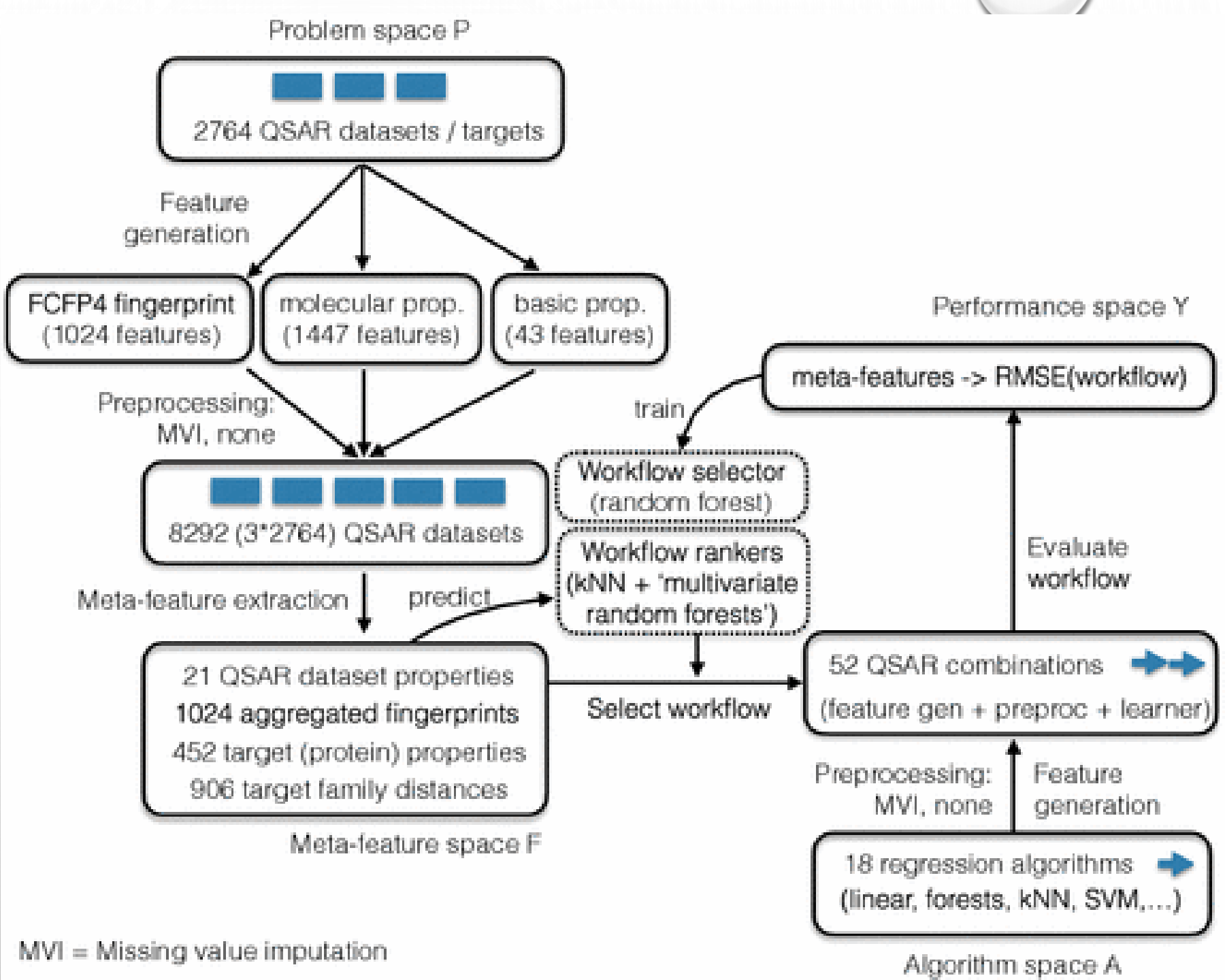
META-QSAR DATASET

- TRAINING META-DATASET
- 2394 META-FEATURES BY 2764 TARGETS



META-LEARNING PIPELINE

THE 52 QSAR COMBINATIONS ARE GENERATED BY COMBINING 3 TYPES OF REPRESENTATION/PREPROCESSING WITH 17 REGRESSION ALGORITHMS, PLUS THE TANIMOTO KSVM WHICH WAS ONLY RUN ON THE FINGERPRINT REPRESENTATION



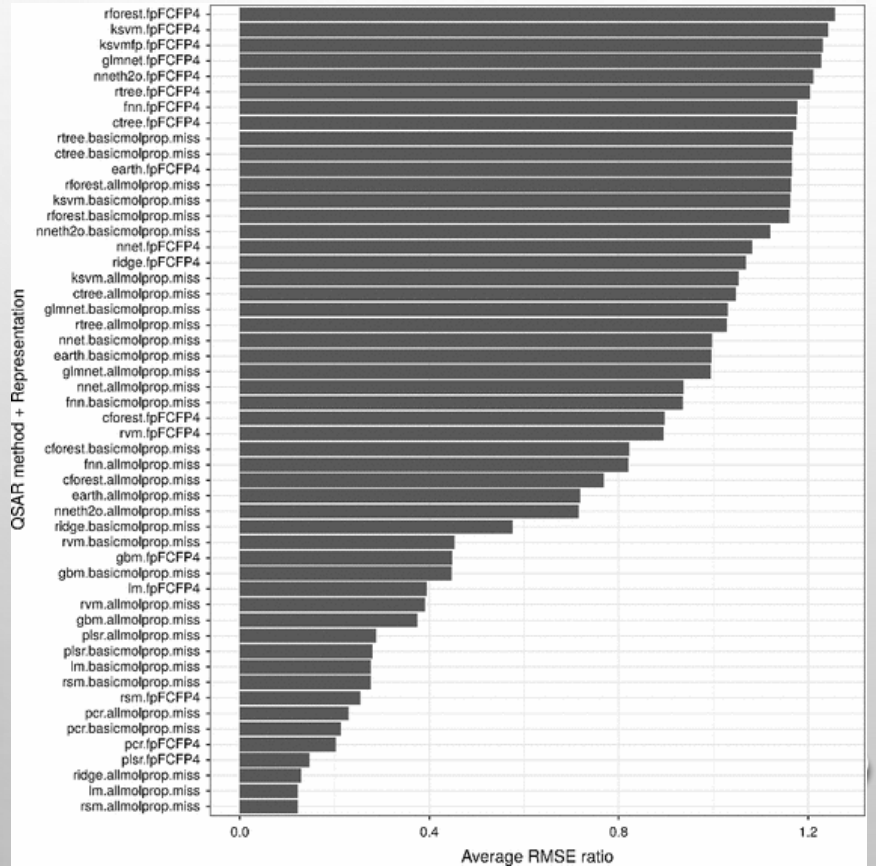
A META-LEARNING CLASSIFICATION AND RANKING

- FOR THE CLASSIFICATION TASK WE USED THE BEST QSAR STRATEGY (COMBINATION OF QSAR METHOD AND DATASET REPRESENTATION) PER TARGET AS THE OUTPUT LABEL
- A META-LEARNING CLASSIFICATION WAS IMPLEMENTED USING A RANDOM FOREST WITH 500 TREES
- FOR THE RANKING TASK, THE QSAR PERFORMANCES (RMSE) WERE USED.
- THE RANKING TASK WAS IMPLEMENTED USING K-NEAREST NEIGHBOUR APPROACH (K-NN) WITH 1, 5, 10, 50, 100, 500, AND ALL NEIGHBOURS; AND A MULTI-TARGET REGRESSION WITH 500 TREES TO PREDICT QSAR PERFORMANCES

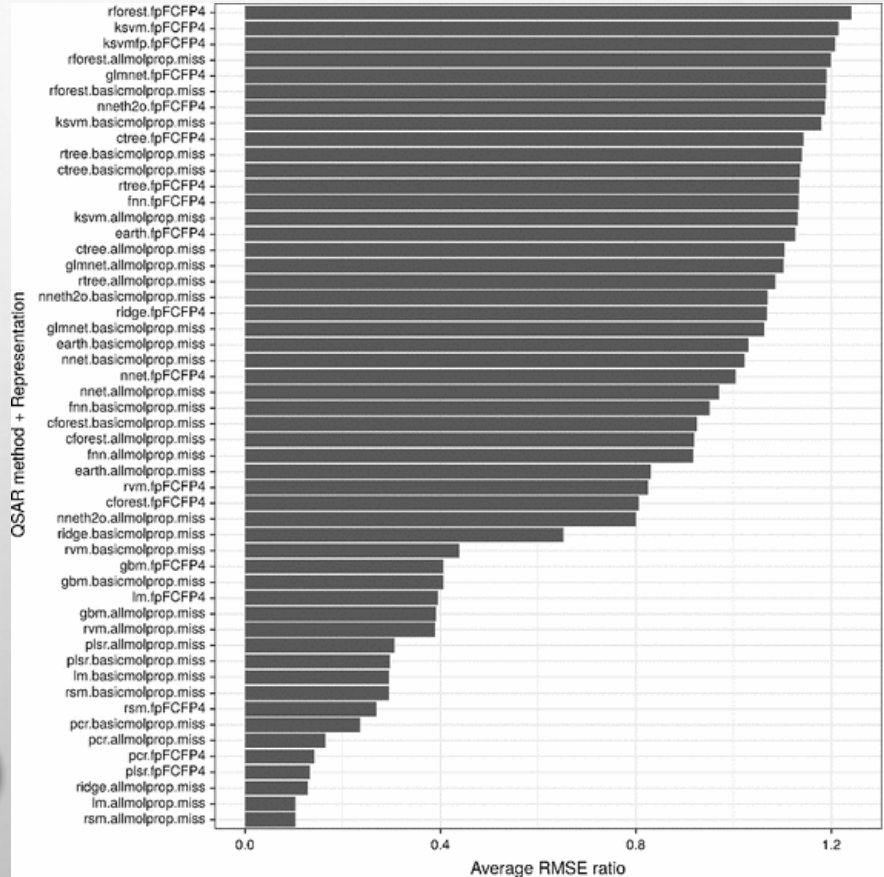
RANKING MODELS

- WE USED THE SPEARMAN'S RANK CORRELATION COEFFICIENT TO COMPARE THE PREDICTED WITH THE ACTUAL RANKINGS
- THE MULTIVARIATE RANDOM FOREST AND 50-NEAREST NEIGHBOURS IMPLEMENTATIONS (MRF AND 50-NN IN THE FIGURE) PREDICTED BETTER RANKINGS, OVERALL.

Predicted ranking



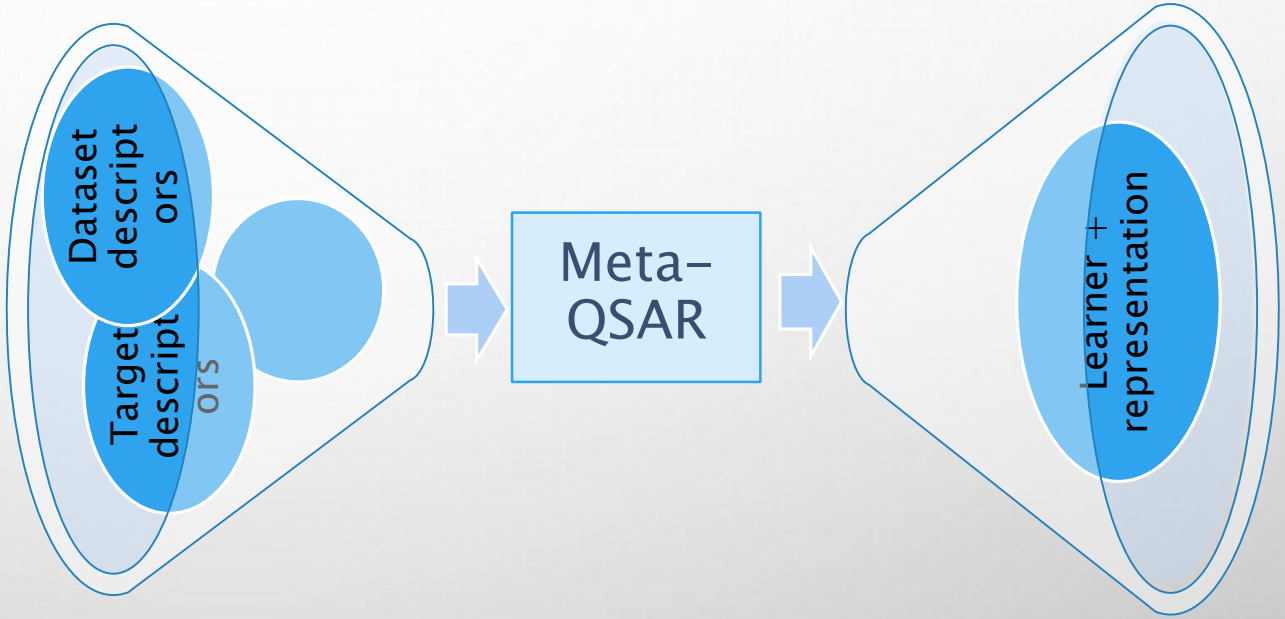
Actual ranking



META-QSAR PERFORMANCE

- PERFORMANCES OF THE BEST SUGGESTED QSAR COMBINATION BY ALL META-QSAR IMPLEMENTATIONS WERE COMPARED WITH AN ASSUMED DEFAULT
- THE DEFAULT (BASELINE) – RANDOM FOREST WITH THE FINGERPRINT MOLECULAR REPRESENTATION (RFOREST.FPFCFP4)
- MOST OF THE META-QSAR IMPLEMENTATIONS IMPROVED OVERALL PERFORMANCE IN COMPARISON WITH THE DEFAULT; THE EXCEPTION OF THE 1-NEAREST NEIGHBOUR

META-QSAR: CONCLUSION



META-LEARNING CAN BE SUCCESSFULLY USED TO SELECT QSAR ALGORITHM/REPRESENTATION THAT PERFORM BETTER THAN THE BEST ALGORITHM/REPRESENTATION (DEFAULT STRATEGY).

PART II: MULTI-TASK QSAR LEARNING

THE PROBLEM

- MANY DATASETS ARE TOO SMALL
- IT IS TOO COSTLY TO OBTAIN LABELED DATA

THE PROPOSED SOLUTION:

- USE EXISTING DATA FROM RELATED TARGETS WHERE LABELED DATA IS APLENTY
- EXPLOIT TASK RELATEDNESS
- INCORPORATE NATURAL METRIC

MULTIPLE TASK LEARNING (MTL)

- Multi-task Learning is different from single task learning in the training (induction) process.
- Inductions of multiple tasks are performed simultaneously to capture intrinsic relatedness.

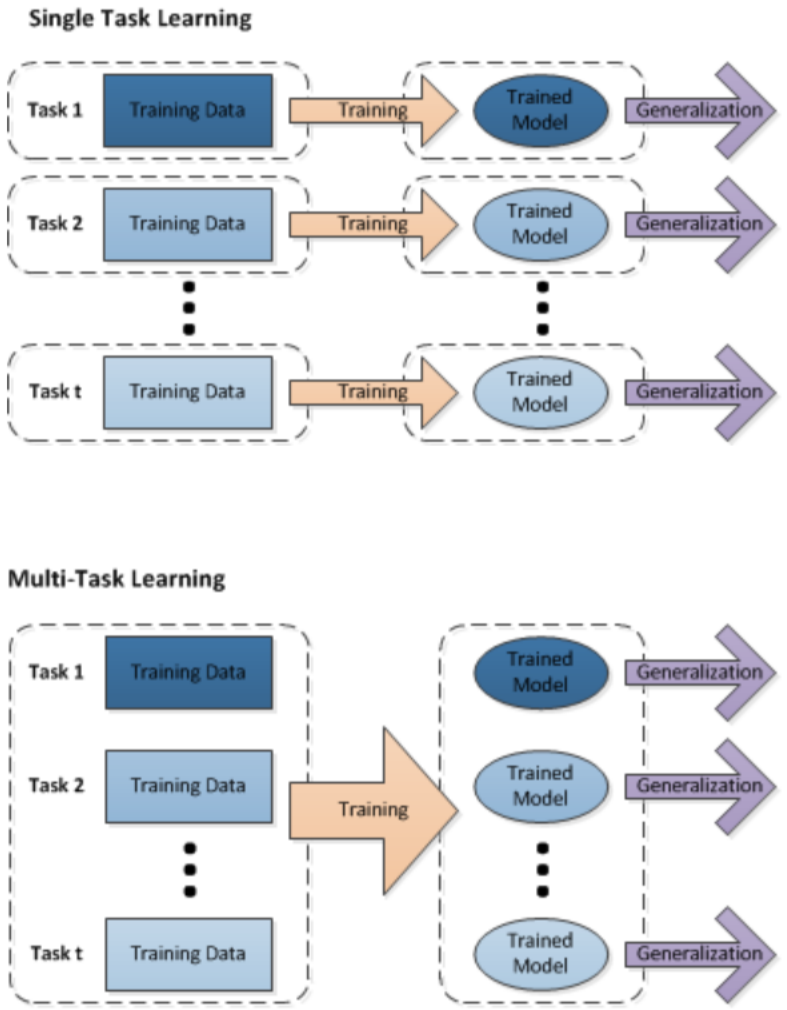


Figure: From SDM 2012 Tutorial by J. Zhou et al

TYPES OF MTL

THERE ARE THREE ASPECTS OF THE TASK RELATEDNESS: FEATURE, PARAMETER, AND INSTANCE, CORRESPONDINGLY – THREE TYPES OF MTL:

1. **FEATURE-BASED MTL** MODELS ASSUME THAT DIFFERENT TASKS SHARE IDENTICAL OR SIMILAR FEATURE REPRESENTATIONS, WHICH CAN BE A SUBSET OR A TRANSFORMATION OF THE ORIGINAL FEATURES.
2. **PARAMETER-BASED MTL** MODELS AIM TO ENCODE THE TASK RELATEDNESS INTO THE LEARNING MODEL VIA THE REGULARIZATION OR PRIOR ON MODEL PARAMETERS.
3. **INSTANCE-BASED MTL** MODELS PROPOSE TO USE DATA INSTANCES FROM ALL THE TASKS TO CONSTRUCT A LEARNER FOR EACH TASK VIA INSTANCE WEIGHTING.

BASELINE: SINGLE TASK LEARNING (STL)

- A SINGLE TASK T_I IS A TASK OF PREDICTING AN ACTIVITY A_I GIVEN A QSAR DATASET OF MOLECULAR STRUCTURES
- DATA: MOLECULAR FINGERPRINTS
- THE FEATURES (1024 BOOLEAN ATTRIBU

MOL_ID	FP_1	...	FP_n	Activity
ID_1	1	...	0	6.45
ID_2	0	...	1	5.98
...	...	...	...	...
ID_111	0	...	1	6.11
ID_112	1	...	1	5.74

STL IMPLEMENTATION

- ALGORITHMS: RANDOM FOREST (100 TREES) ON EACH DATASET
- 10 FOLD CROSS-VALIDATION TO OBTAIN AN ESTIMATE OF THE PERFORMANCE FOR EACH MODEL
- PERFORMANCE METRIC: ROOT MEAN SQUARED ERROR (RMSE)
- SOFTWARE: WEKA 3.7.11 MACHINE LEARNING PACKAGE

FEATURE-BASED MTL (SETTING 1)

- AIM: TO LEARN ALL DRUG TARGETS FOR A PARTICULAR PROTEIN TARGET GROUP (E.G. DHFR) SIMULTANEOUSLY
- CONCATENATE ALL THE DATASETS OF THE SAME GROUP, AND ADD AN EXTRA INDICATOR ATTRIBUTE.

MOL_ID	OrganismTID	FP_1	...	FP_n	Activity
ID_1	7	1	...	0	6.45
ID_2	7	0	...	1	5.98
...	...	...	...	...	...
ID_111	10095	0	...	1	6.11
ID_112	10095	1	...	1	5.74

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

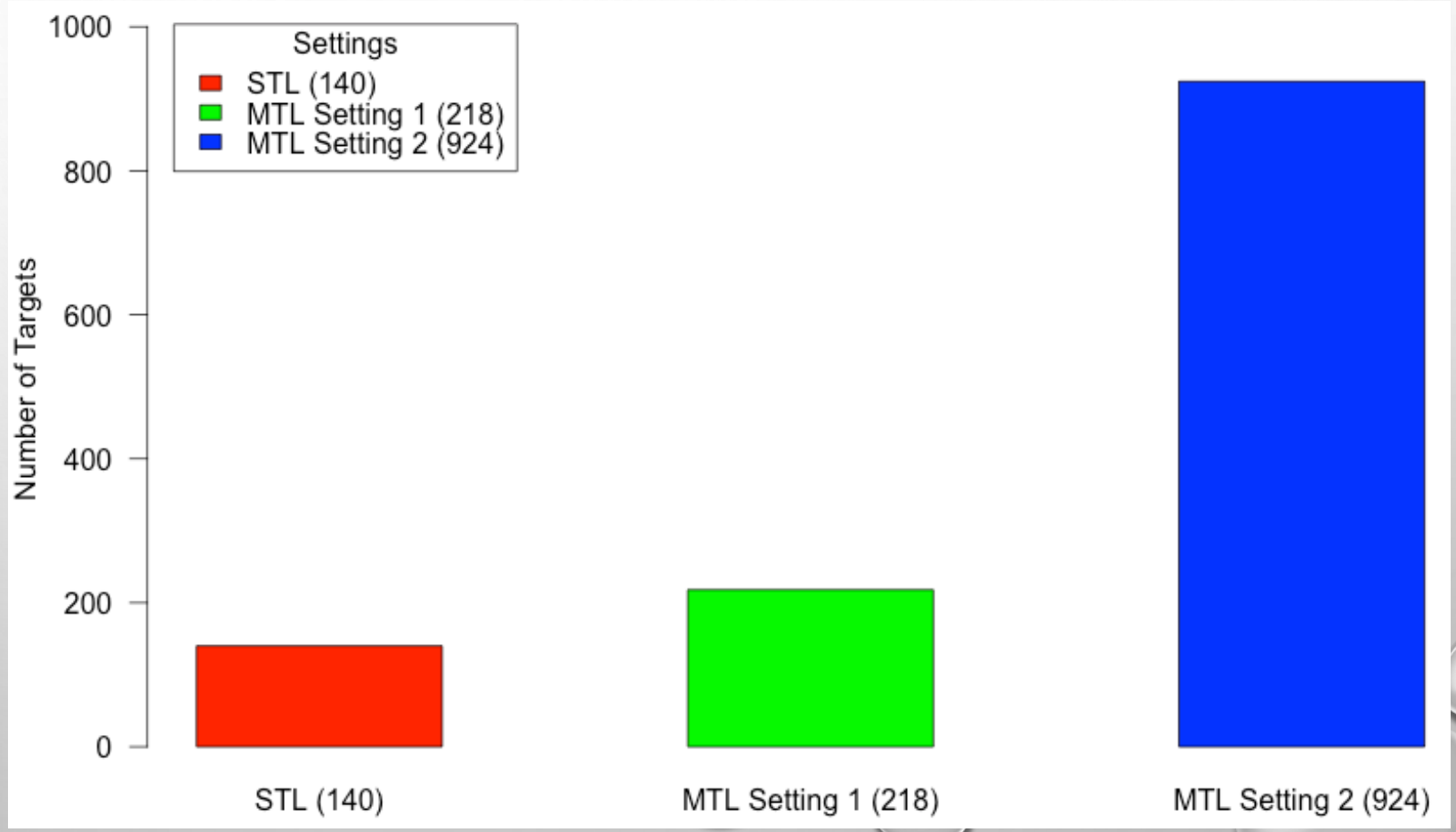
INSTANCE-BASED MTL (SETTING 2)

- CONCATENATE THE *N* DATASETS INTO ONE BIG DATASET
- ADD AN INDICATOR VARIABLE *TID* TO EACH EXAMPLE
- ADD *N* EXTRA VARIABLES TO THE BIG DATASET: *SIMTOTID 1*, *SIMTOTID 2*, ..., *SIMTOTID N*
- VALUES ARE CALCULATED USING SIMILARITIES BETWEEN TARGETS

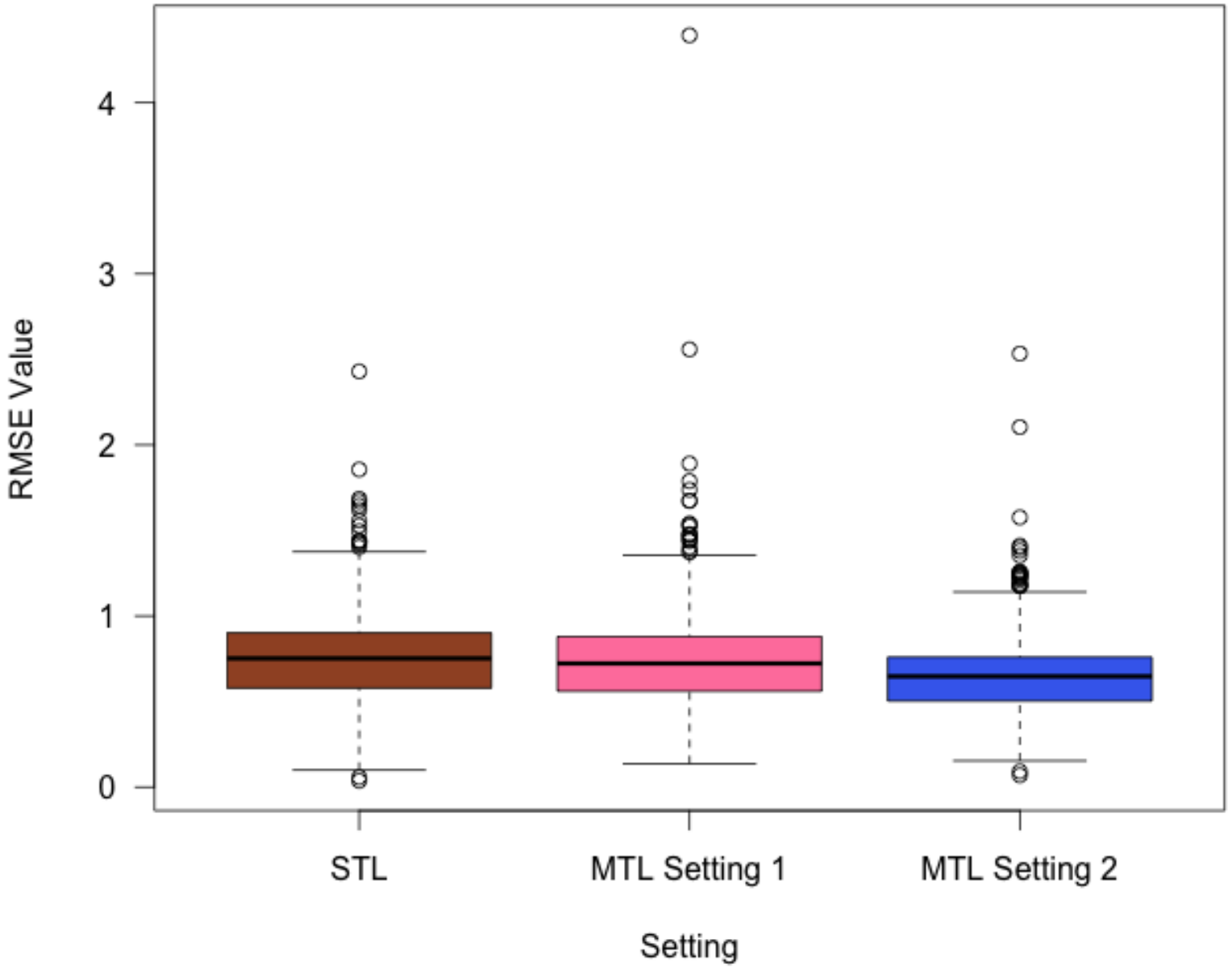
MOL_ID	OrganismTID	SimToOrganism_7	...	FP_1	...	FP_n	Activity
ID_1	7	1	...	1	...	0	6.45
ID_2	7	1	...	0	...	1	5.98
...	...	...	...	...	...	...	...
ID_111	10095	0.3964	...	0	...	1	6.11
ID_112	10095	0.3964	...	1	...	1	5.74

RESULTS FOR L5 TARGET CLASSES

COUNT OF HOW MANY TARGETS EACH ALGORITHMS PERFORMS BETTER THAN THE OTHER TWO ALGORITHMS



BOXPLOT OF RMSE VALUES



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
- BETTER NOT TO STRATIFY BASED ON TARGET ID, USE DISTANCE/SIMILARITY BETWEEN DATASETS

AVAILABILITY

- OPENML: *HTTPS://WWW.OPENML.ORG*

- DATASETS, CODE AND A YOUTUBE VIDEO TUTORIAL
  HTTPS://GITHUB.COM/NSADAWI/MTL-QSAR

ACKNOWLEDGEMENTS

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

Brunel University London

University of Dundee

The University of Manchester

THE META-QSAR TEAM:

IVAN OLIER, NOUREDDIN SADAWI, G. RICHARD BICKERTON, JOAQUIN VANSCHOREN, CRINA GROSAN, LARISA SOLDATOVA, ROSS D. KING

THANK YOU!