SFU SIMON FRASER UNIVERSITY
THINKING OF THE WORLD

Drug-Target Interaction Prediction

Martin Ester

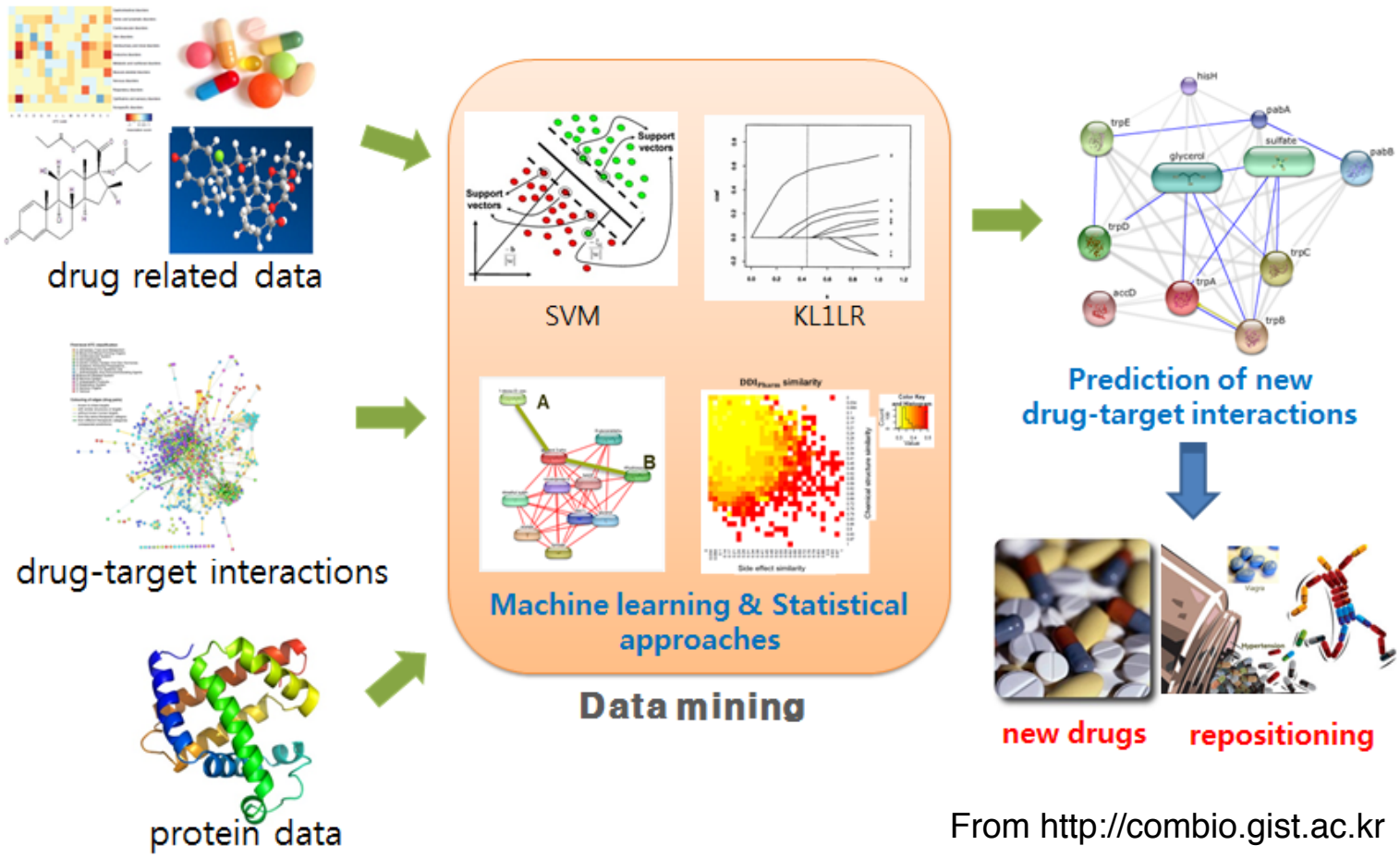
CMPT 884 Spring 2018

- Introduction
- Problem Definition
- Related Work
- SimBoost
  He et al., Journal of Cheminformatics, 2017
- Evaluation
- Future Work

Outline

- Drugs work by interacting with their targets, i.e. disease-causing proteins.
- Finding a compound (candidate drug) that selectively binds to a particular protein is a highly challenging and typically expensive procedure.
- 90% of candidate compounds fail due to cross-reactivity and/or toxicity issues.
- In-silico drug-target interaction prediction can speed up the experimental wet lab work by systematically prioritizing the most potent compounds and help predicting their potential side effects.

Introduction



From http://combio.gist.ac.kr

Introduction

Input

- M is a matrix with continuous values where Mi,j represents the binding affinity of drug i and target j.

→ M is very sparse!

- D is a similarity matrix of drugs.
- T is a similarity matrix of targets.

Output

- Prediction of all the non-observed values in M.

Problem Definition

Simulation-based approaches

- Molecular docking of a candidate compound with the protein target is simulated, based on the 3D structure of the target (and the compound).
- Does not work of 3D structure unknown.
- Efficient only when considering a single (or few) targets.

Related Work

Machine learning-based approaches

- Learn from training data.
- Efficient predictions on a larger scale, i.e. for many targets simultaneously.
- (Multitarget) Quantitative Structure–Activity Relationship (QSAR) methods: relate a set of predictor variables, describing the physico-chemical properties of a drug-target pair, to the response variable, representing the existence or the strength of an interaction.

→ feature-based / similarity-based

Related Work

Feature-based ML

- Known drug-target interactions are represented by feature vectors generated by combining chemical descriptors of drugs with descriptors for targets.
- Features:
  sequence descriptors (proteins)
  substructures (compounds)
- Classifiers
  SVM, Naïve Bayes
  DNN

Related Work

Similarity-based ML

- Similarity matrices are computed for both the drug-drug pairs and the target-target pairs.
- Typically, chemical structure fingerprints are used to compute drug similarity, and a protein sequence alignment score is used to compute protein similarity.
- Classifiers

  Nearest Neighbor

  Kernel SVM

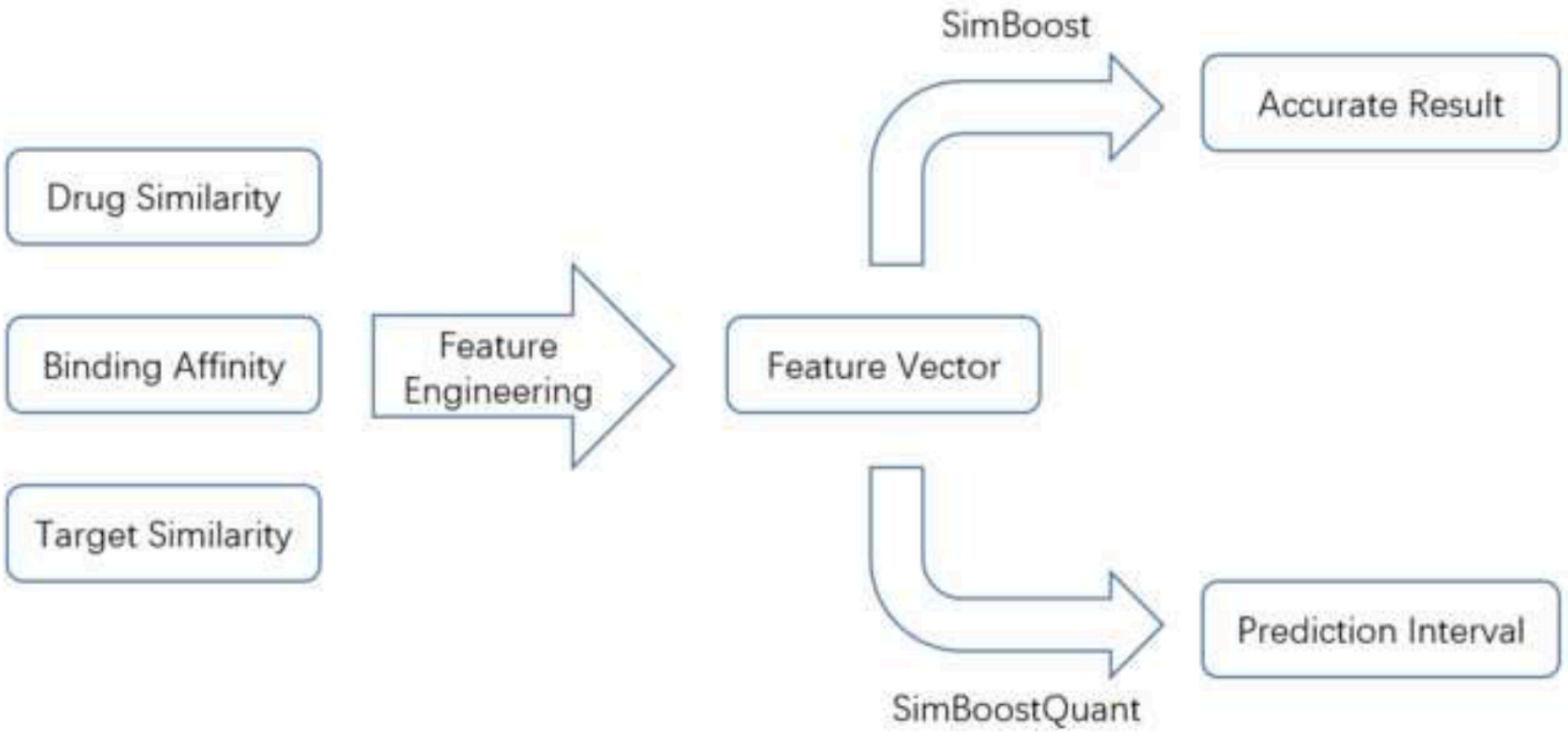
Related Work

Properties

- Uses drug similarities, target similarities, and drug-target interaction data.
- Predicts continuous drug-target binding affinity values.
- Computes a prediction interval in order to assess the confidence of the predicted affinity.

→lacking in state-of-the-art methods

SimBoost


SimBoost
Accurate Result
Drug Similarity
Binding Affinity
Target Similarity
Feature Engineering
Feature Vector
Prediction Interval
SimBoostQuant

SimBoost

Features for single objects (drugs/targets)

- number of observations for the object in *M*
- average binding affinity of object
- average of all similarity scores of the object
- . . .

SimBoost

Features from drug/target network

- We build two networks, one for drugs and one for targets, from D and T, respectively. The nodes are drugs / targets, and an edge between two nodes exists if their similarity is above a user-defined threshold.
- number of neighbours
- average of object-specific features among the k-nearest neighbors of the object
- betweenness, closeness and centrality of the node

SimBoost

Features for drug target pairs

- We build one network for drugs and targets from M. The nodes are drugs and targets, and an edge connects a drug and a target, with the binding affinity as the edge weight.
- latent factors from matrix factorization
- betweenness, closeness and centrality of the node

SimBoost

Gradient Boosting Regression Trees

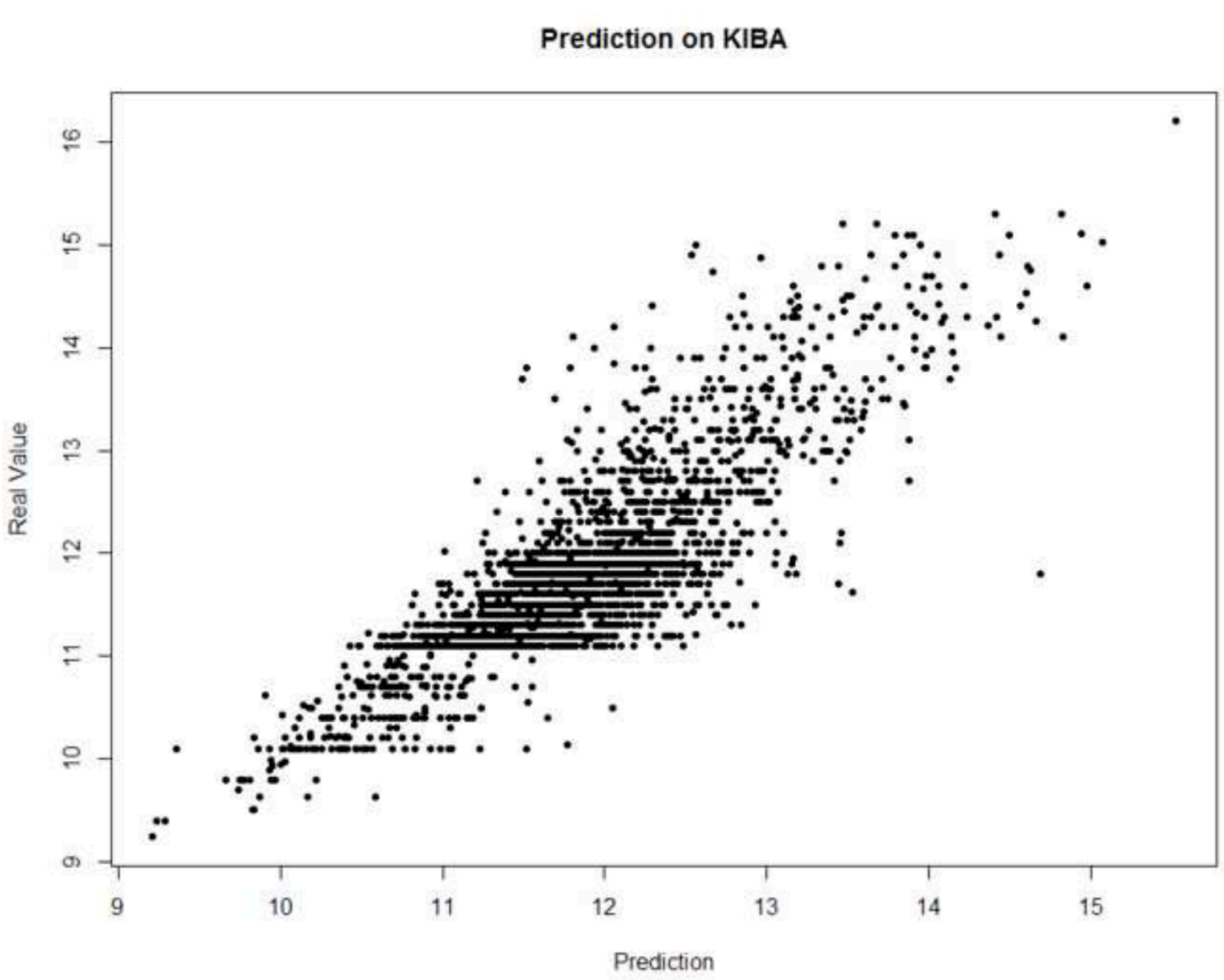
Chosen as regression model because of

- Accuracy: the boosting algorithm is an ensemble model, which trains a sequence of ”weak learners” to gradually achieve a good accuracy.
- Efficiency: the training process can be parallelized, greatly reducing the training time.

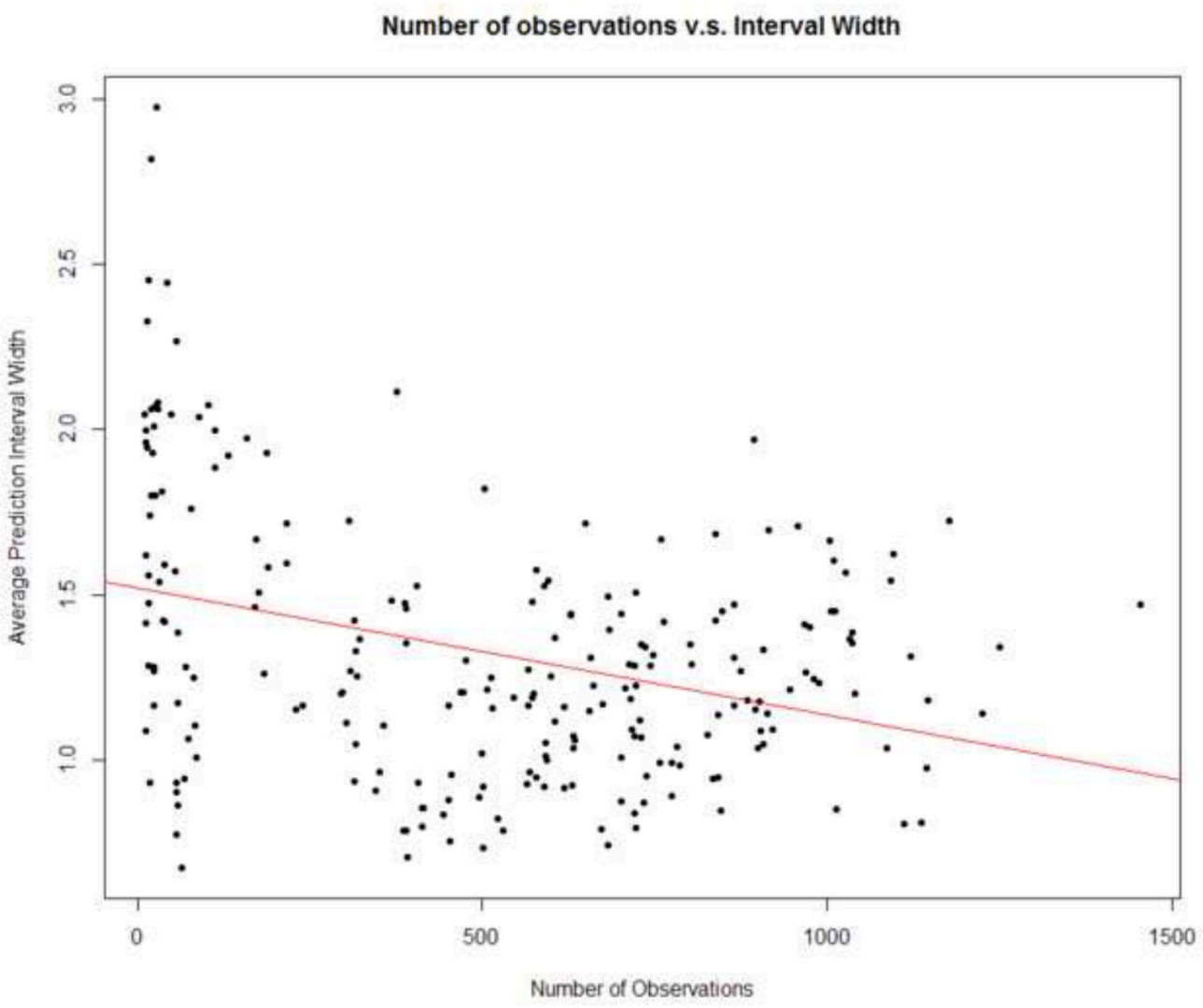
SimBoost

	RMSE	AUC	AUPR	CI
MF	0.382 ± 0.003	0.831 ± 0.002	0.631 ±0.004	0.792 ± 0.001
Continuous *KronRLS*	0.620±0.001	0.884 ± 0.001	0.735 ± 0.001	0.792±0.001
Binary *KronRLS*	-	0.904±0.001	0.7660±0.001	-
SimBoost	**0.204** ± 0.001	**0.907** ± 0.001	**0.782** ± 0.001	**0.847** ± 0.001
SimBoostQuant	0.299± 0.001	0.875 ± 0.001	0.708± 0.002	0.796 ±0.001

Evaluation


Prediction on KIBA
Real Value
Prediction
9
10
11
12
13
14
15
16

Evaluation


Number of observations v.s. Interval Width
Average Prediction Interval Width
3.0
2.5
2.0
1.5
1.0
0
500
1000
1500
Number of Observations

Evaluation

- Prediction for cold-start targets/drugs.
- Combine machine learning approach (first level) with simulation approach (second level).
- Discovery of mechanisms behind drug-target interaction.
- More accurate prediction based on DNN?

Future Work