Cell-line Specific Drug Synergy Prediction Using Machine Learning Based Model

Nure Tasnina

Ph.D. Student
Computer Science and Applications
Virginia Tech

29 November, 2020

What is Drug Synergy

- **Synergy:** When the combined effect of drugs is greater than the additive effect expected from the individual drugs.
- **Antagonism:** The opposite of synergy i.e. when the combined effect of drugs is less than what would be expected.

What is Synergy Score

Synergy Score indicates the degree of synergy between drugs.

- Loewe Additivity: $E_{Loewe} = \frac{E_{min}+E_{max}[(d_A+d_B)/m]^{\lambda}}{1+[(d_A+d_B)/m]^{\lambda}}$
  - Score $> 0 \Rightarrow$ *Synergistic*
  - Score $< 0 \Rightarrow$ *Antagonistic*

Problem Formulation

Given, experimentally derived synergy score (e.g. Loewe Additivity) between a number of drug pairs in a particular cell line, predict synergy between new drug pairs in that cell line.

$$f(Drug_A, Drug_B, cell\_line\_x) = s$$

Why is it an Important Problem to Solve

- Drug combination has been proven useful in treating many diseases (e.g. cancer) as it has the potential to increase efficacy, decrease detrimental side effects and overcome drug resistance.
- Enormous drug space preclude the possibility of systematic experimental exploration of drug combinations.
- Hence, computational approach becomes a necessity.

Previous Works

Paper	Model	Dataset	Feature	Evaluation Metric	Performance
Preuer et. al.	Fully connected Deep Neural Network	O'Neil	Chemical structure of drugs, gene expression for cell lines	MSE, Pearson's correlation	Better than Gradient Boosting, Random Forests, SVM and Elastic Nets
Jiang et. al.	GCN based encoder and matrix factorization based decoder	O'Neil	Drug target interaction, Protein protein interraction	AUC (after implementing threshold on predicted score) Pearson's correlation	Claimed to perform better than DeepSynergy
Celebi et. al.	Linear regression, Lasso, support vector machine (SVM), random forest, and XGBoost (Best)	AstraZeneca Dream Challenge data	Drug: Chemical structure, target, protein domain, interaction network Cell line: Gene expression, Mutational profile, Copy number, Drug Monotherapy	Weighted average Pearson correlation coefficient (WAPCC)	WAPCC = 0.39 by XGBoost Model
Kim et. al.	Multi-modal, multi-tasking Deep Neural Network	DrugComb	Chemical structure of drugs, SMILES encoding gene expression for cell lines	AUC (after implementing threshold on predicted score)	Claimed to perform better than DeepSynergy

My Approach: Supervised Machine Learning Model

I considered this problem as a regression based supervised machine learning problem.

- **Task:** Predict the synergy score between two drugs in a particular cell line.
- **Evaluation:** Evaluate the model's performance using Root Mean Squared Error(RMSE). 5-fold cross validation setup.
- **Experience:** The model will learn from the labeled feature vector where the feature vector will contain drug specific and/or cell line specific features and the label will be the synergy score.

Data

- **DrugCombDB:** Source of synergy scores of drug pair data
  - number of drug combinations = 448555
  - number of unique drugs = 2887
  - number of human cancer cell lines = 124

Data

- **Therapeutic Target Database(TTD) :** Source of Drug Target data
  - number of drug targets $= 3,419$
- **DrugBank:** Additional source of Drug Target data
  - number of drugs $= 9,591$
  - number of drug-targets $= 4,115$

Method: Feature Extraction

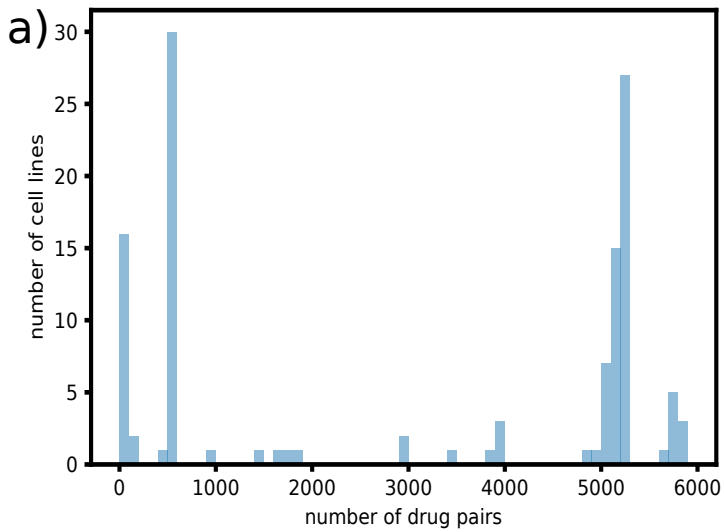
1. **Drug Molecular Features:** Molecular ACCess System (MACCS). MACCS fingerprints contains 166 chemical structures such as the number of oxygens, S-S bonds, ring.
   - Using python package "RDKit" (http://www.rdkit.org) computed MACCSKeys for each drug
2. **Drug Target:** Protein targets of each drug
   - Parsed TTD and DrugBank dataset to map drug names to target proteins.
   - For each drug computed a 0-1 vector of size 3378 (i.e. number of unique targets across TTD and DrugBank) where 1 is for targeted protein, 0 for non targets.

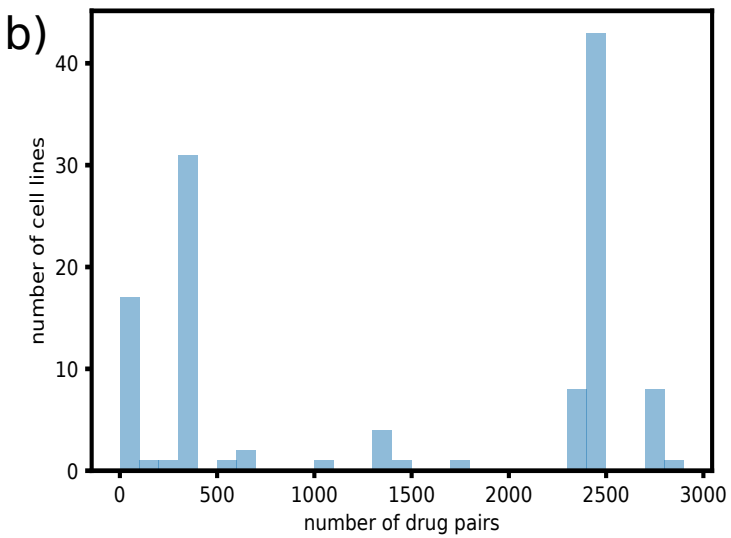
Total number of features for each drug$= 3378 + 166 = 3544$

Method: Combine Features

- To combine the features extracted from different data sources I needed a standard identification system for drugs
- I used PubChemID (https://pubchem.ncbi.nlm.nih.gov/) as the standard ID
- I used python package 'PubChemPy' to retrieve PubChemID for drugs present in three different databases (DrugCombDB, DrugBank and TTD)

Drug Pairs having All Feature Values





- 61 cell lines have scores and feature matrix available for more than 2000 drug combinations
- Number of unique drugs $= 1,956$

Method: Machine Learning Model

Baseline: Support Vector Regression. One SVR model for each cell line.
Reason: Number of features $>>$ Number of Observations/ labeled drug-pairs

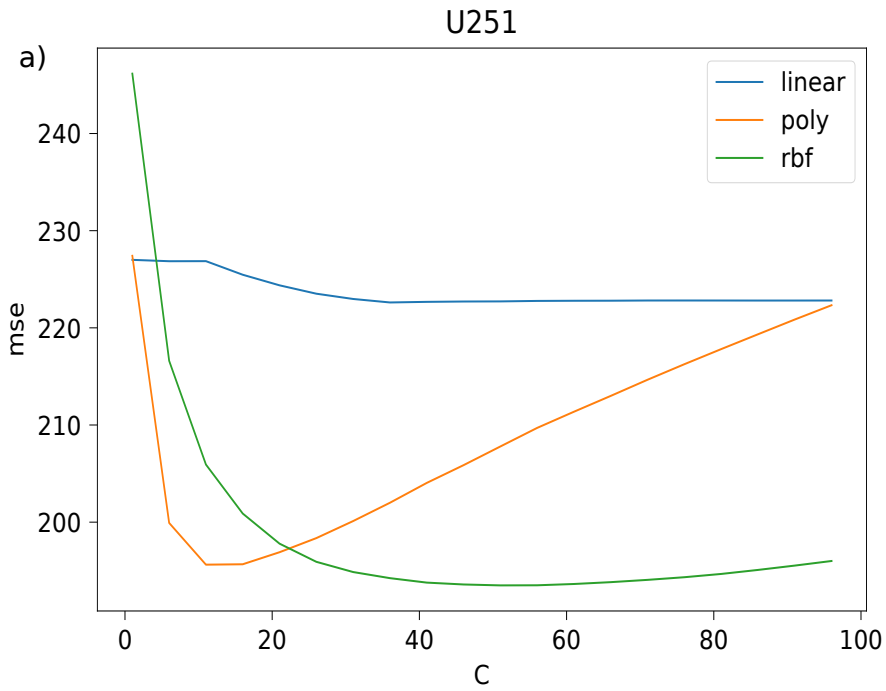
Method: Implementation of SVR

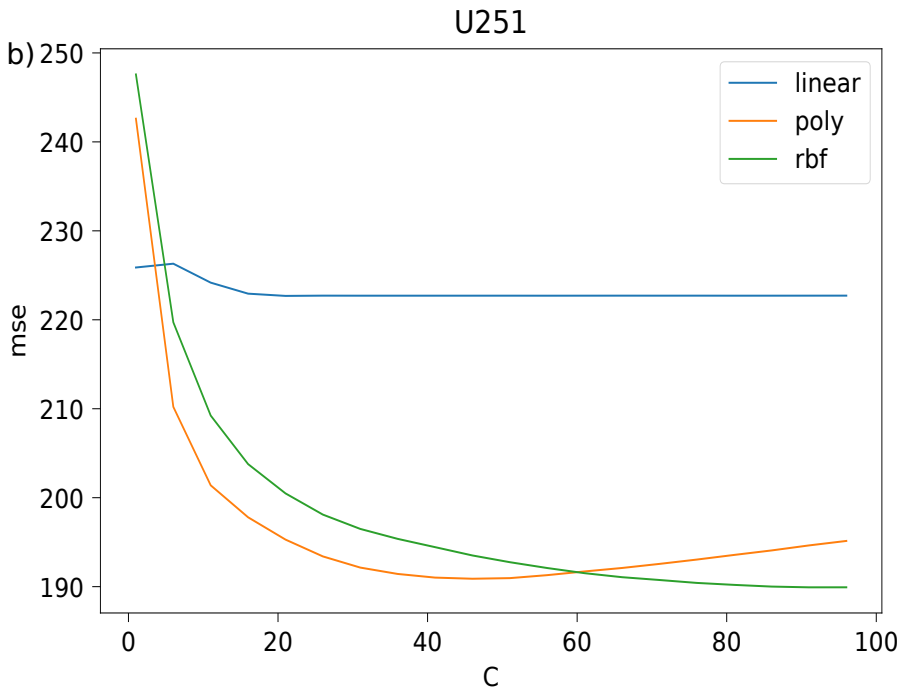
- **Used Library**: sklearn.svm.SVR
- **Train-Test split:** Split dataset into train and test data set with **9:1** ratio
- **Cross Validation:** Doing 5-fold cross validation on training data to tune hyper-parameters: Kernel and Regularization Parameter
  - **Regularization Parameter, C** I am doing GridSearch between values $0 - 50$ to find the optimal value for the regularization parameter (C) of L2 penalty.
  - **Kernel** I am also choosing the best performing kernel among "Linear", "poly", "rbf"

Result: Cross Validation Result

Cell Line	SVR-MACCS		SVR-MACCS-Target		GBR-MACCS	GBR-MACCS-Target
	C	Kernel	C	Kernel	Max Depth	Max Depth
NCI-H226	66	rbf	96	rbf	5	5
HCC-2998	41	rbf	66	rbf	4	5
MDA-MB-435	96	rbf	96	rbf	5	5
DIPG25	11	rbf	16	rbf	3	4
COLO 205	56	rbf	36	poly	5	5
HCT-15	81	rbf	96	rbf	5	5
HOP-92	56	rbf	81	rbf	4	5
TK-10	51	rbf	86	rbf	4	3
U251	51	rbf	91	rbf	5	5
HL-60(TB)	76	rbf	96	rbf	5	5

Result: Hyper Parameter Tuning of SVR on Cell Line U251

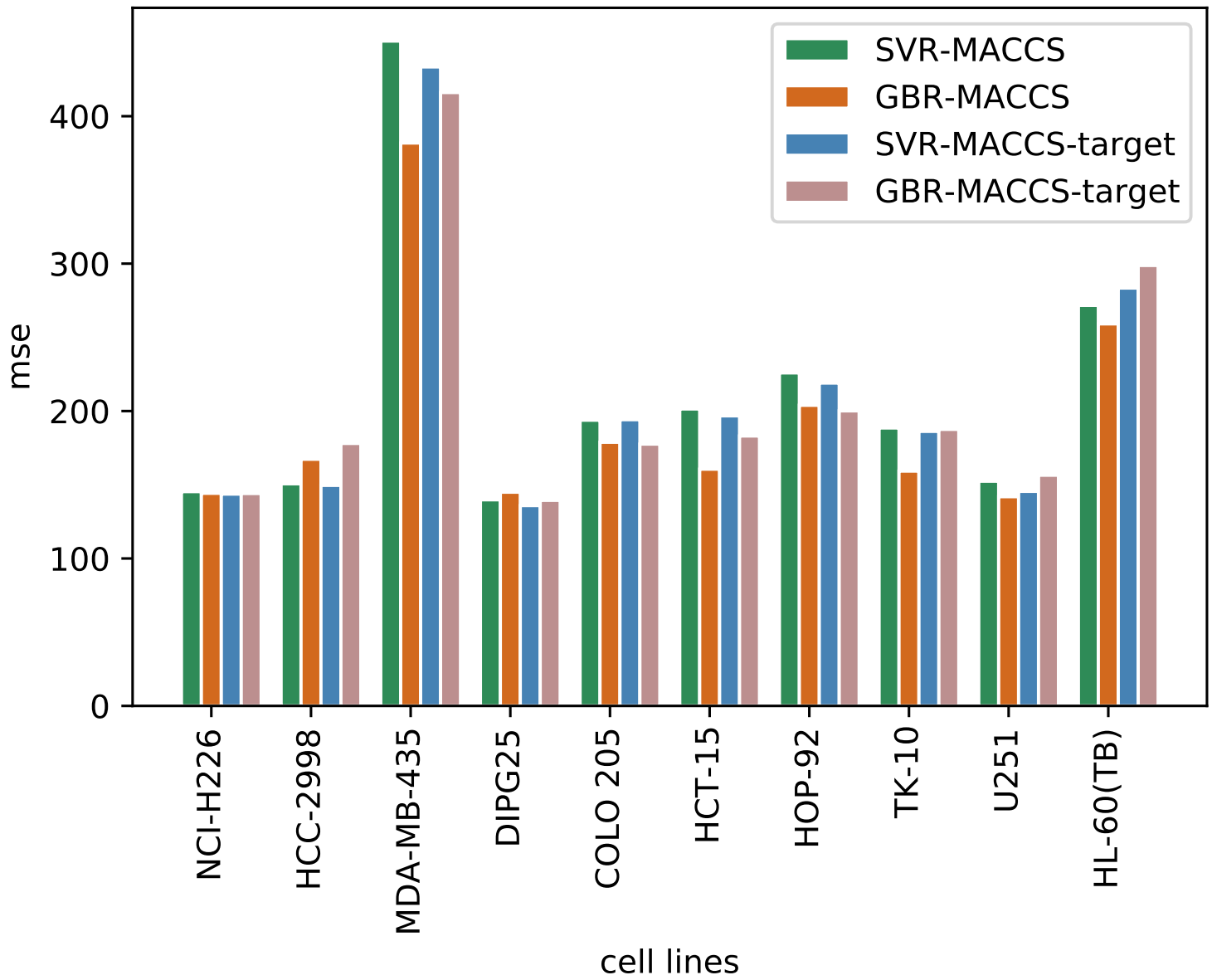




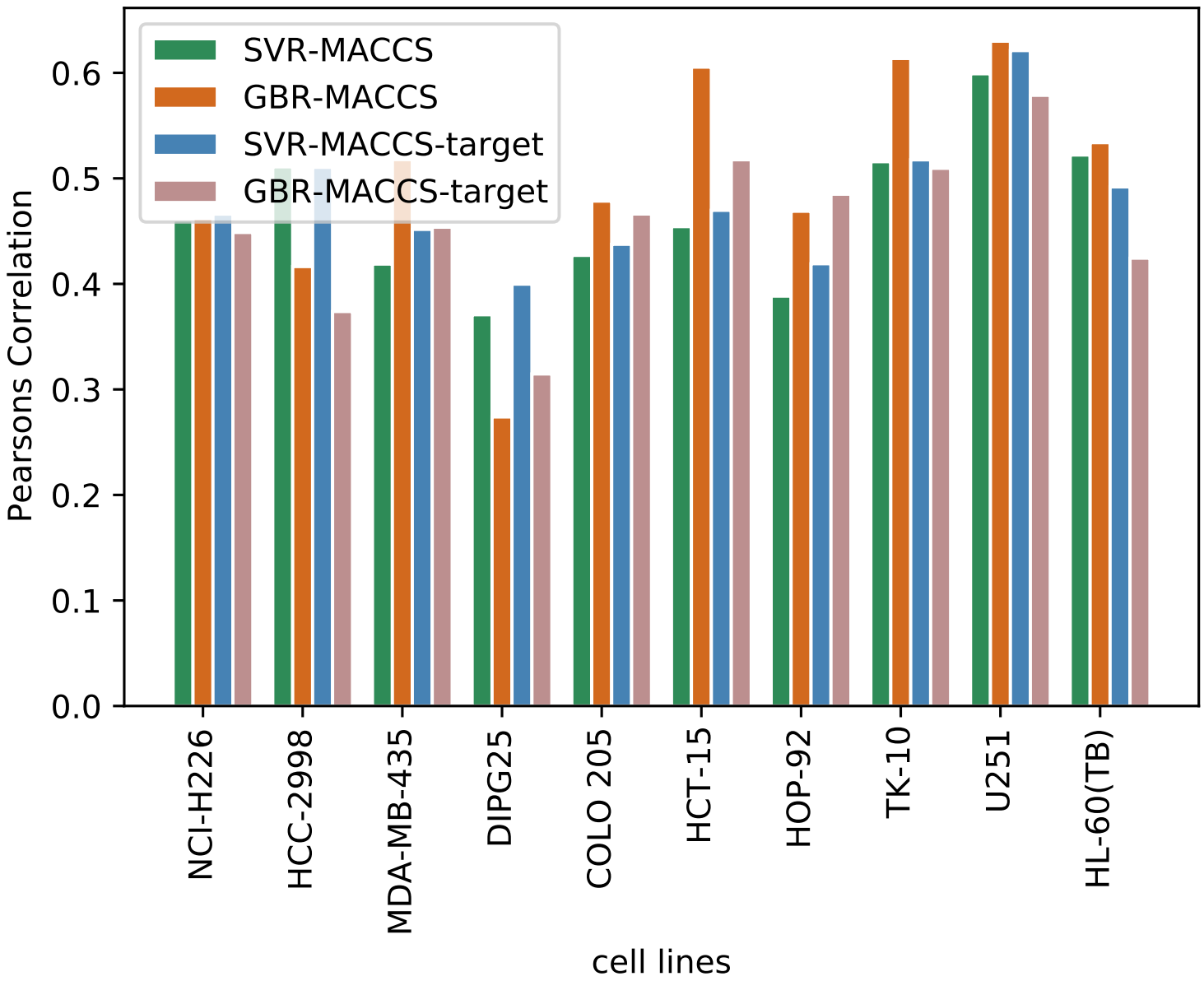
- Trained with MACCS fingerprints

- Trained with MACCS fingerprints and drug-target

Result on Test Dataset: Comparison



Result on Test Dataset: Comparison



Future Work

- I plan to implement Deep Neural Network model to predict synergy score on my extracted features.
- I also plan to incorporate other potential dataset e.g. protein-protein interaction network, pathway information, cell line specific gene expression, and drug-drug interaction network to explore the feature space more.


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
BIYAN
SHUKRIA
TINGKI
YAQHANYELAY
TASHAKKUR ATU
SUKSAMA
EKHMET
MEHRBANI
GRAZIE
MAAKE
PALDIES
KOMAPSUMNIDA
GOZAIMASHITA
EFCHARISTO
JUSPAXAR
BOLZİN
MERCI