**FDMine: a graph mining approach to predict and evaluate food-drug interactions**

Md. Mostafizur Rahman1, Srinivas Mukund Vadrev1, Arturo Magana-Mora2, Jacob Levman1 and Othman Soufan1

*1Department of Computer Science, St. Francis Xavier University, Nova Scotia, Canada*

*2Saudi Aramco, EXPEC Advanced Research Center, Drilling Technology Team, Dhahran, 31311, Saudi Arabia.*

**Additional File 1**

1. **DrugBank Dataset Extraction**

The following data flow diagram (Figure S1) is showing the extraction procedure of the DrugBank dataset. The main DrugBank database is big and contains a lot of information. We used ‘dbparser v1.2.0’ a package of ‘R’ programming language, and parse the whole dataset to get our project related information dbparser1. In this project we have considered only the ‘approved’ drug group and ‘small molecule’ type drugs information. To complete this job, we used ‘dplyr’ a package of ‘R’ programming language. After minimizing the main DrugBank dataset, we had 1,683 unique approved small molecule drug information and 788 unique approved molecule drug who has metabolism related interactions.

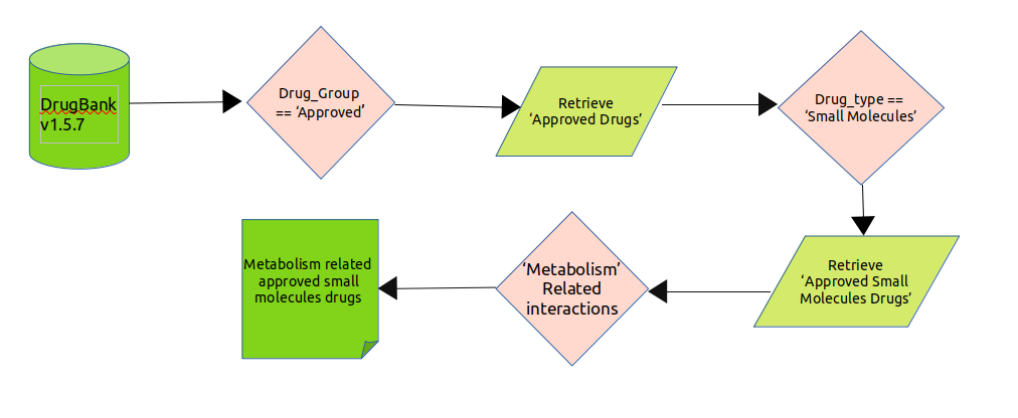


Figure S1: DrugBank Dataset Extraction Procedure

1. **FooDB Dataset Extraction**

For preparing our food dataset, we have downloaded the FooDB Version 1.0 dataset in JSON format foodb. We used ‘R’ programming language to parse the database for data pre-processing. We required "tidyverse" and "dplyr" packages of version 1.3.0 and 1.0.5 respectively for this work. Figure S2 represents the block diagram of exclusion criteria we have used for food data processing.

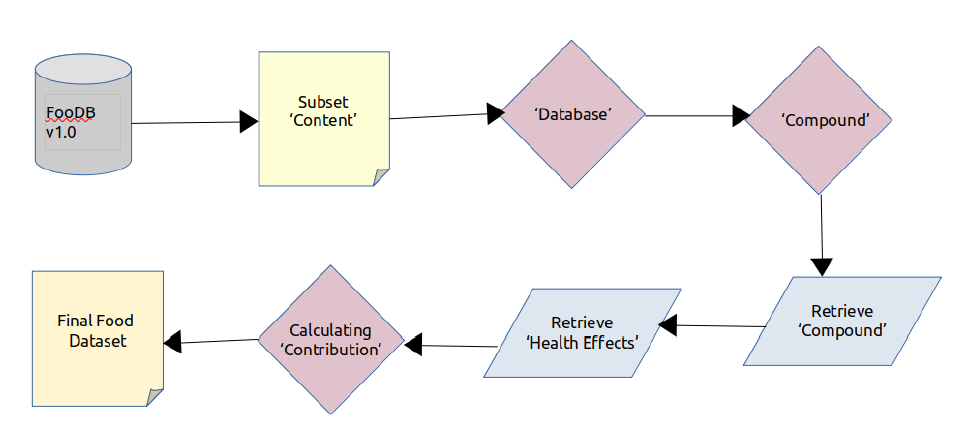


Figure S2: FooDB Dataset Extraction Procedure

1. **Structure Similarity Profile**

A structural similarity profile (SSP) is a feature vector that contains a unique numerical representation after acquiring structural features of individual drugs and food components. The SSP contains pairwise structural similarity scores obtained from the comparison among all the 788 approved small molecule drugs of DrugBank and 568 unique food compounds. Structural similarity between a pair (i.e. drug-drug, food-food, and drug-food) was measured by the Tanimoto coefficient. The Tanimoto coefficient is an appropriate choice for calculating the structure similarity based on the fingerprint. The Tanimoto coefficient is defined as the number of common chemical fingerprints compared to the number of all chemical fingerprints of the two drugs. Chemical fingerprints of each drug were calculated by Morgan/Circular Fingerprints.

We have used a Python package ‘RDKit’ (version 2020.09.2) to calculate the SSP of individual pairs. We followed 4 steps to calculate the SSP:

Step-1: Check whether the given SMILES are correct or not. Because at the time of parsing the main SMILES structure could have been destructed or may contain some noise.

Step-2: Make a list of mols from the SMILES structures.

Step-3: Make a list of fingerprints (fp) from the mols.

Step-4: Compare all fp pairwise without creating any duplicates. Here, duplicate means, we did not check the same drug-drug and food-food fingerprints. Because, the similarity between same drug-drug and food-food will be the highest score (1).

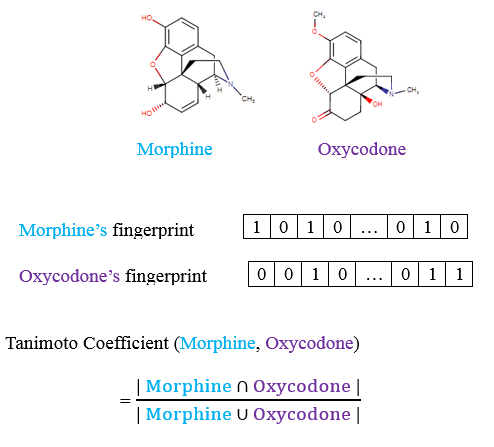


Figure S3: Calculating Structure Similarity Profile Using Tanimoto Coefficient

1. **Food Compound’s Contribution**

We calculated the food compound contribution using the given formula. Food Compound's contribution has a huge impact to predict the food-drug interactions. In the Table S1 this food contains 4 different compounds. All of these compounds does not have the same contribution. So, this is not fair to use the score from SSP directly in the models. So, we use the contribution of the food compound to predict FDIs efficiently.

Table S1: Calculating the contribution of a food compound in a food

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Food\_ID\_Compound\_ID\_Compound\_Name** | **Original Content of a Compound in the food** | **The total original content in the food** | **Contribution of the compound in the food (0 ~ 1)** | **Contribution of the compound in the food (%)** |
| FOOD00005 \_ FDB000633 \_ Kaempferol | 4.10000 | 6.500 | 0.630769230769231 | 63.0769230769231 |
| FOOD00005 \_ FDB002602 \_ Cyanidin | 0.00000 | 6.500 | 0.000 | 0.00 |
| FOOD00005 \_ FDB002798 \_ Apigenin | 0.00000 | 6.500 | 0.000 | 0.00 |
| FOOD00005 \_ FDB011904 \_ Quercetin | 2.40000 | 6.500 | 0.369230769230769 | 36.9230769230769 |

1. **Disjoint and Joint Graph**

When a graph dataset contains multiple subgraph this is known as disjoint graph network. We found our main dataset contains disjoint graph networks. Over this disjoint graph network we applied eight different link prediction algorithms. After that, we created a joint graph network from the same disjoint graph network. To make the joint graph network, we chose any node (randomly) from the subgraphs of the disjoint graph network and add an edge among all subgraphs. Also, we gave a very small edge weight 0.00001 to the new connected edge.

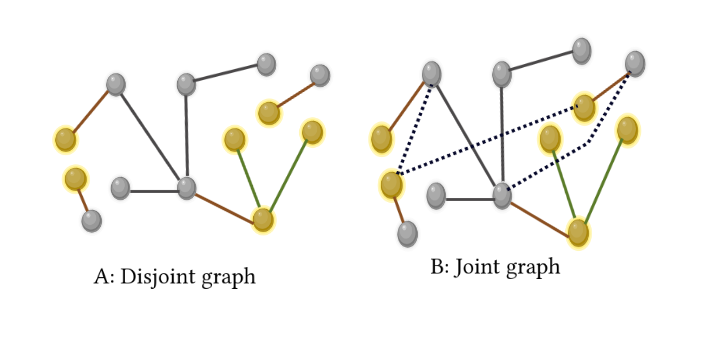


Figure S4: Disjoint and Joint Graph

In the Figure S4 the left network (A: Disjoint graph) contains 3 subgraphs. To make (B: Joint graph) we chose any node from subgraphs and connect them with 0.00001 edge weight. Dashed line indicates the new edge created to convert the disjoint graph into joint graph.

1. **Precision@Top**

We have taken the idea of the precision@k. Instead of taking directly the value of k we chose top 1%, top 2%, and top 5% based on the score given by link prediction algorithm. For example, for the following Figure S5, we have 10 newly links. So, top 1% is the first (Link-1). We are able to match the first link. So, the precision@top-1 is 1. Again, for top-2, we have 2 links (Link-1 and Link-2). Between these 2 links we are able to match 1 and unable to match 1. So, precision2top-2 is 0.5. Same for top-5.

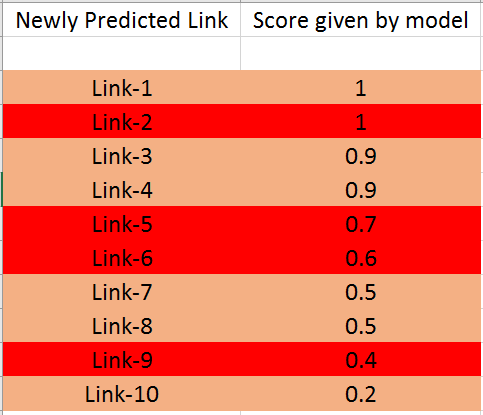


Figure S5: Precision@Top

1. **Comparison of the Precision@top rate over eight link prediction methods and two different graph networks**

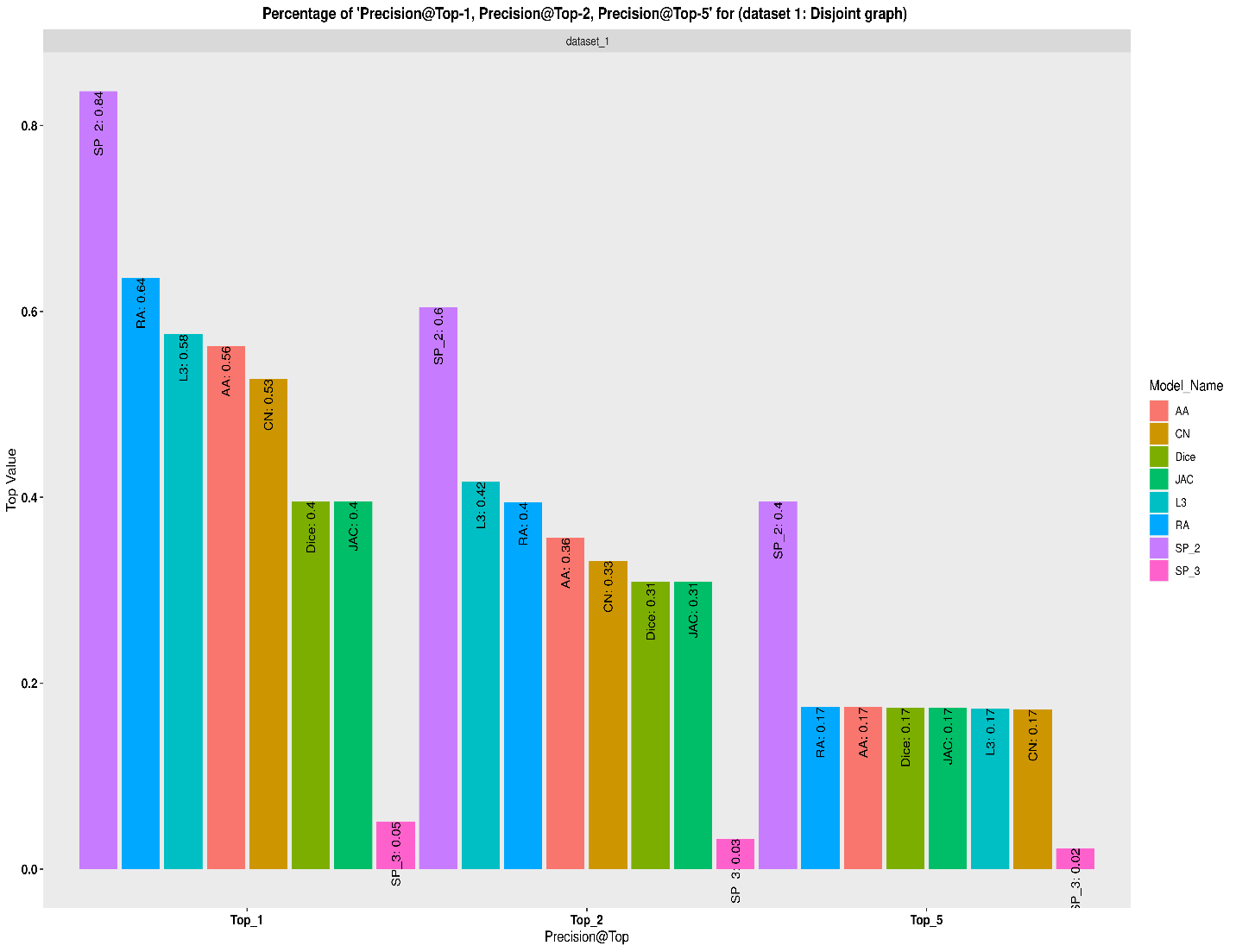


Figure S6: Precision@top comparison of eight different methods over the disjoint graph network

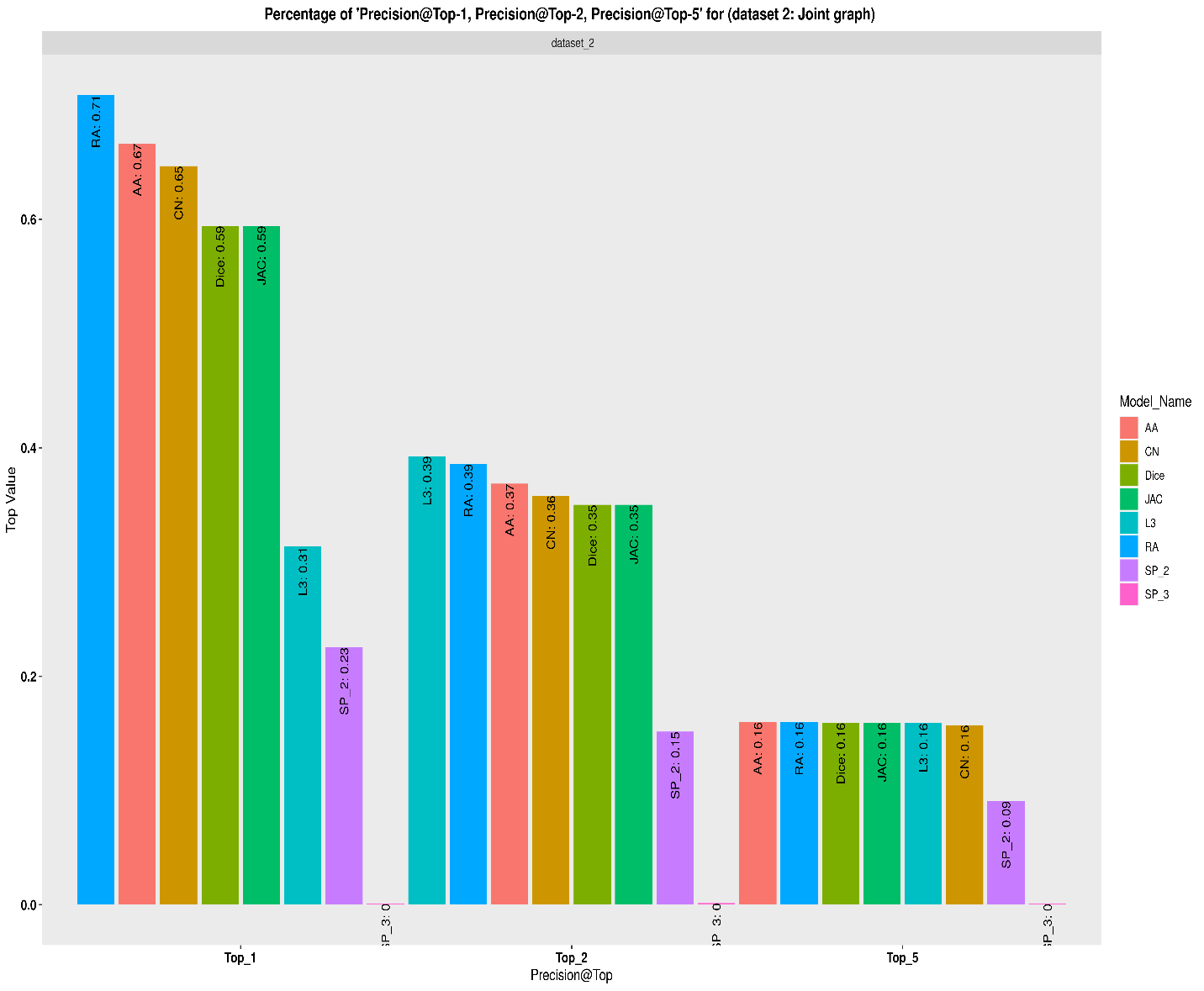


Figure S7: Precision@top comparison of eight different methods over the joint graph network

1. **Area Under the Curve (AUC)**

To evaluate the performance of the link prediction models we used AUROC (Area Under the Receiver Operating Characteristics). This is one of the most important evaluation metrics. To draw AUROC we need true positive rate (TRP) and false positive rate (FPR). TRP and FRP calculated based on the formula given below

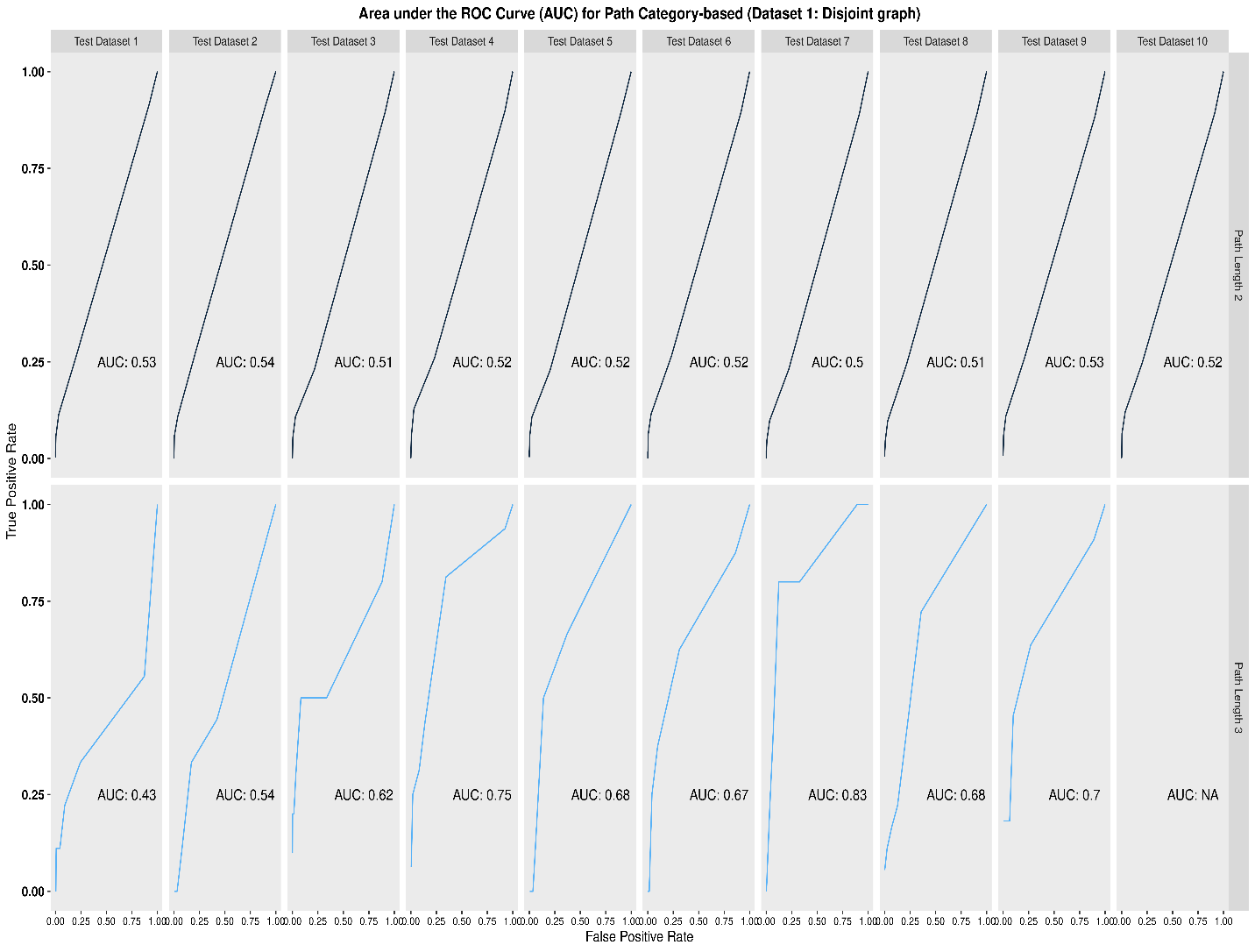


Figure S8: Area Under the Curve (AUC) for path category-based (dataset 1: disjoint graph)

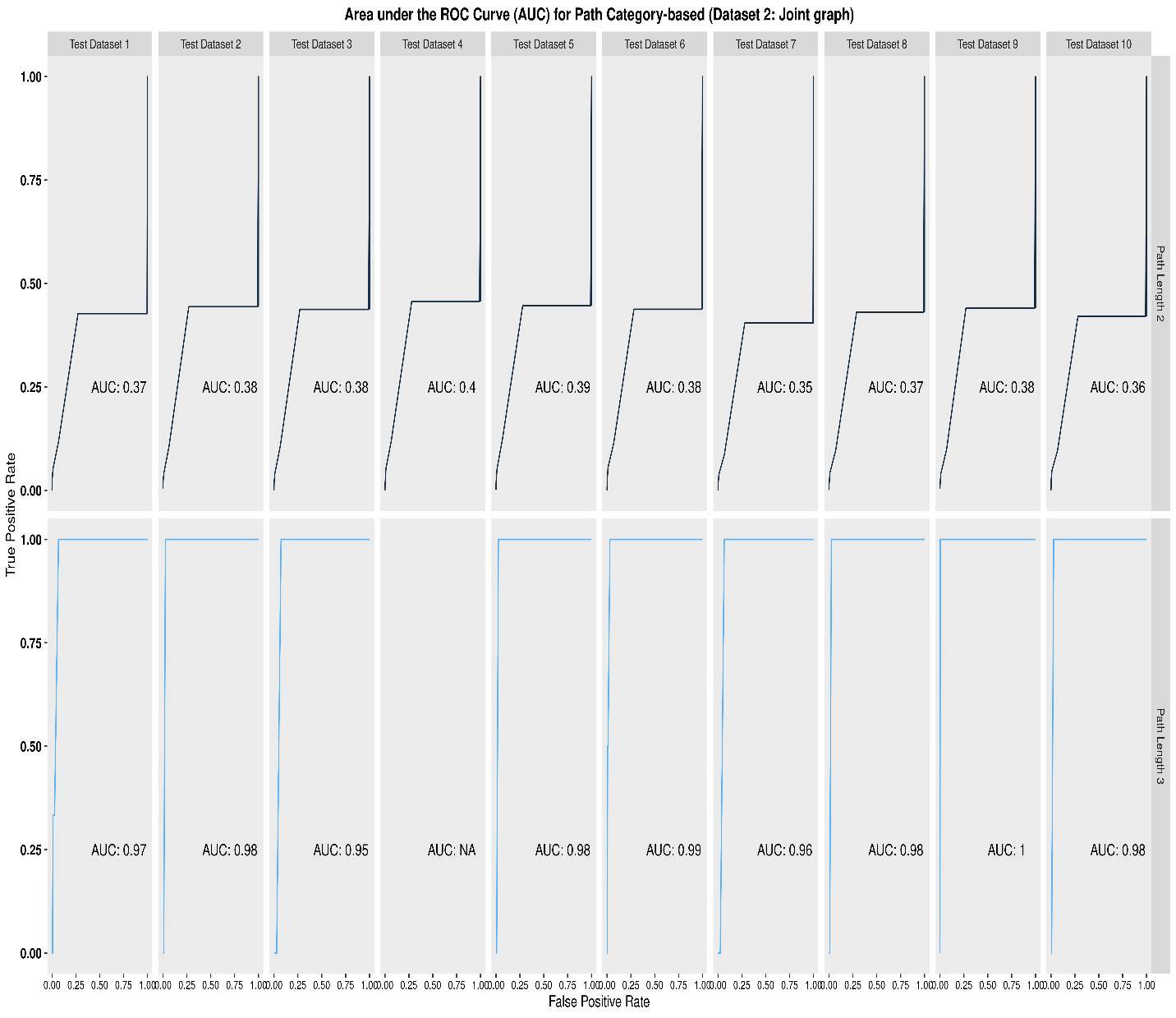


Figure S9: Area Under the Curve (AUC) for path category-based (dataset 2: joint graph)

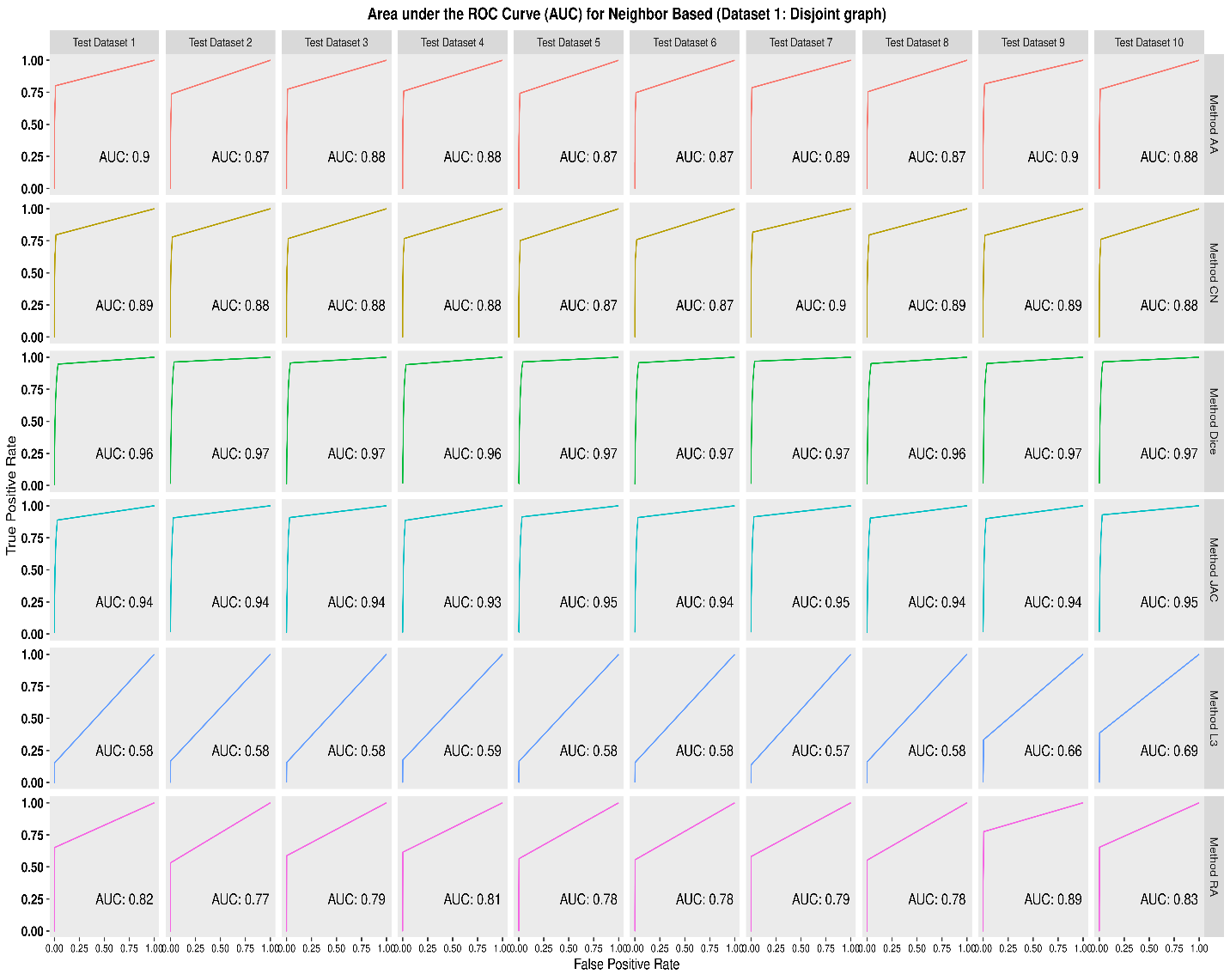


Figure S10: Area Under the Curve (AUC) for neighborhood-based similarity-based (dataset 1: disjoint graph)

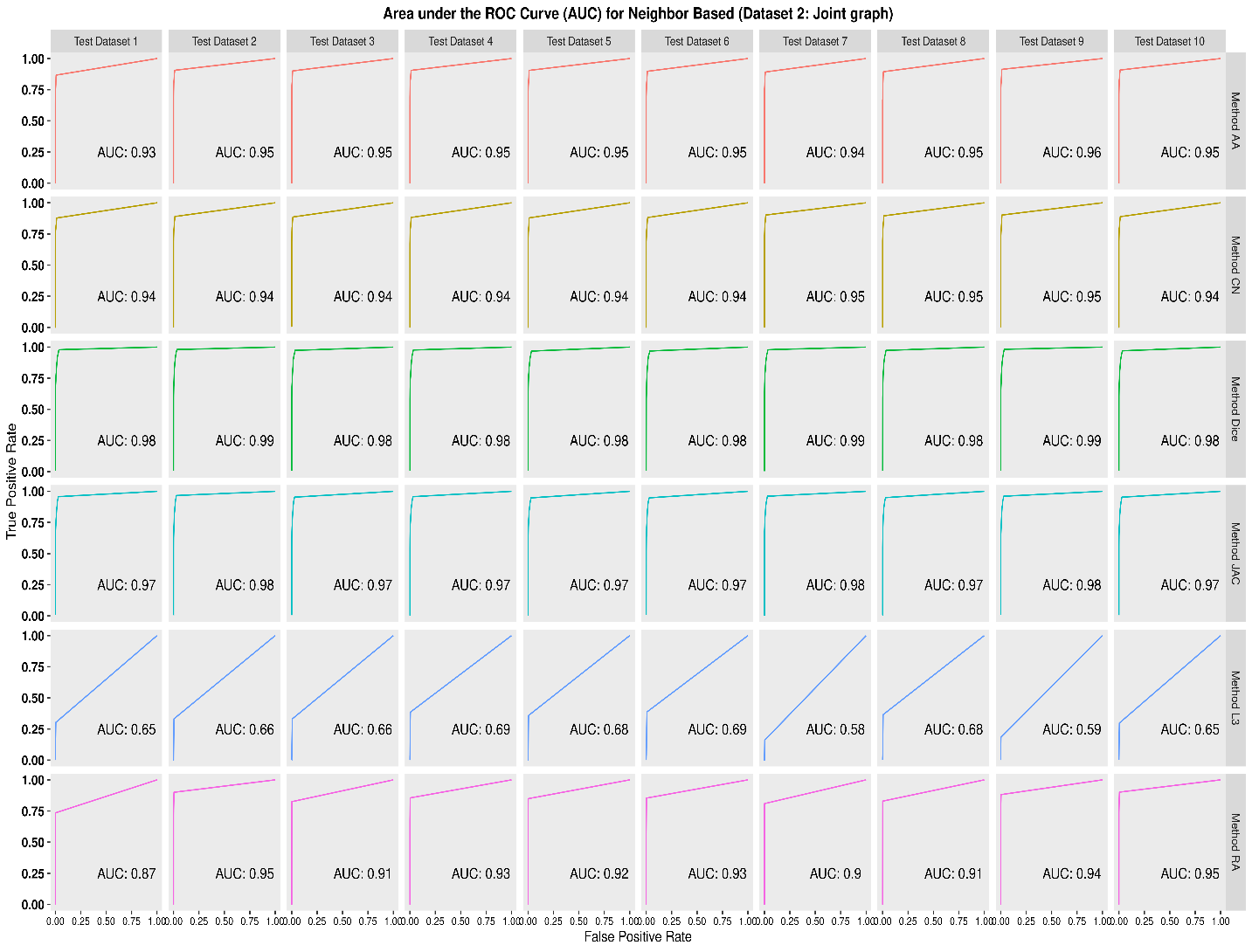


Figure S11: Area Under the Curve (AUC) for neighborhood-based similarity-based (dataset 2: joint graph)

1. **Precision-Recall Curve (PRC)**

A precision-recall curve (also known as a PR Curve) is a graph that shows the precision (y-axis) and recall (x-axis) for various probability thresholds. We calculated precision and recall is as follows

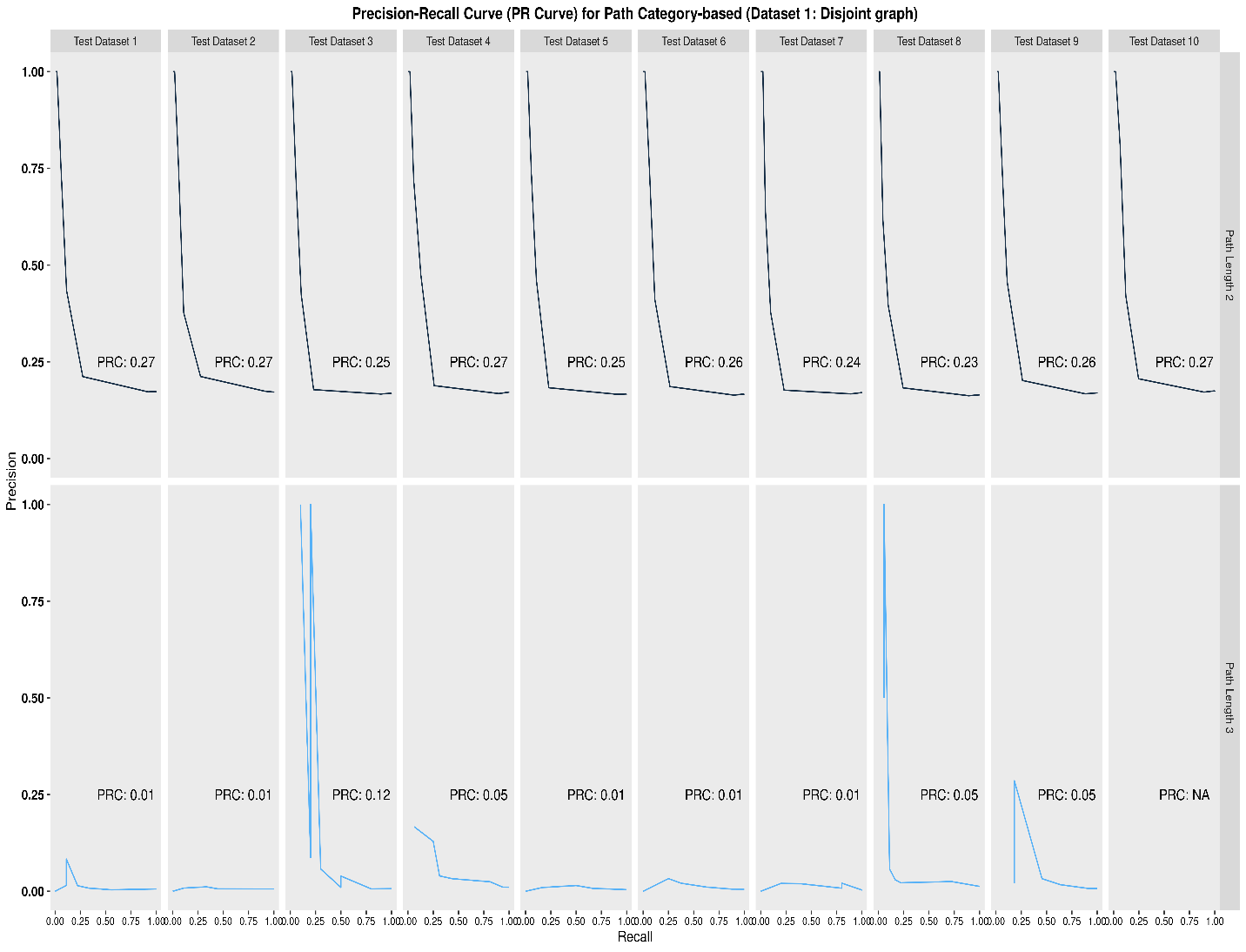


Figure S12: Precision-Recall Curve (PRC) for path category-based (dataset 1: disjoint graph)

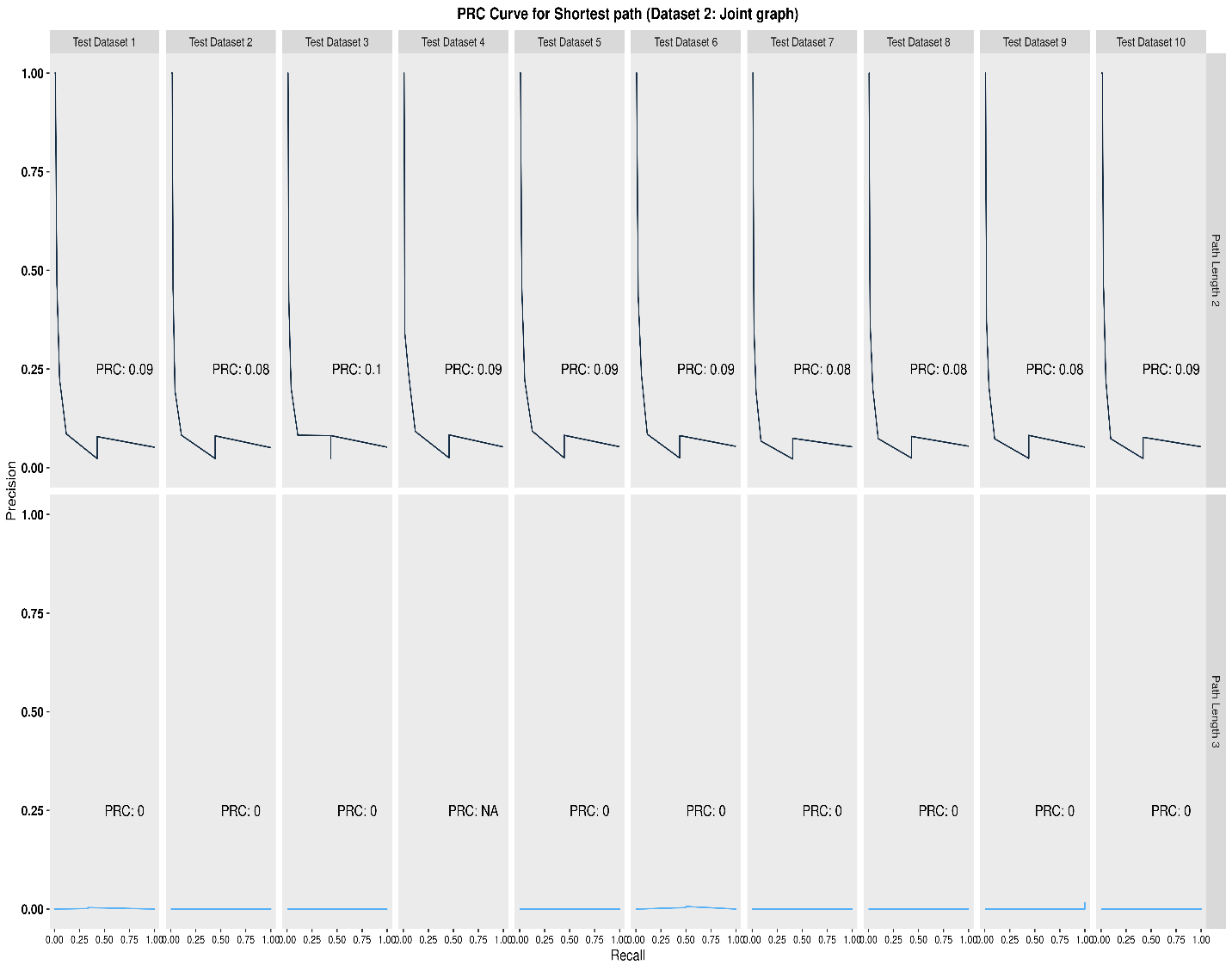


Figure S13: Precision-Recall Curve (PRC) for path category-based (dataset 2: joint graph)

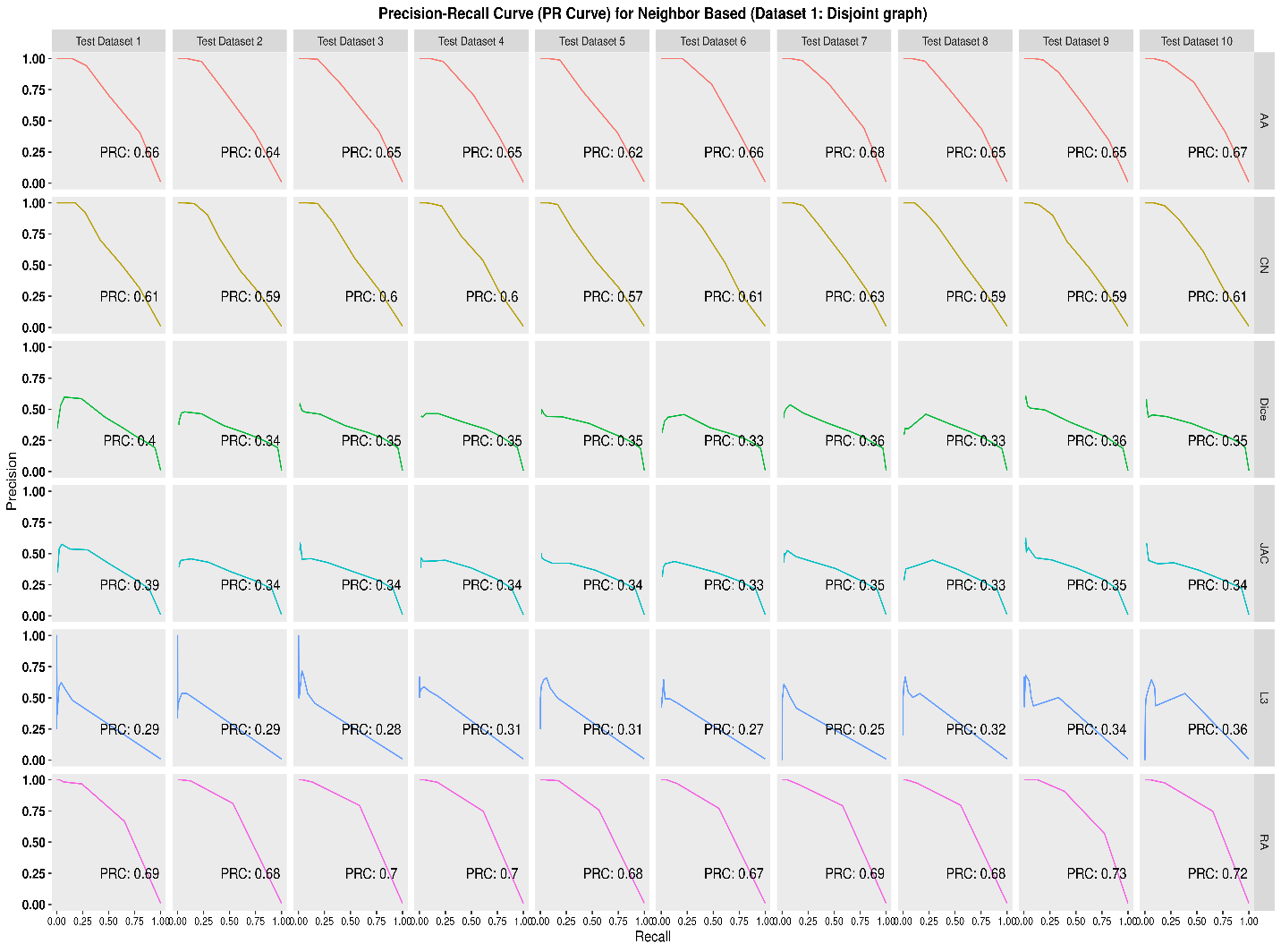


Figure S14: Precision-Recall Curve (PRC) for path neighborhood-based similarity-based (dataset 1: disjoint graph)

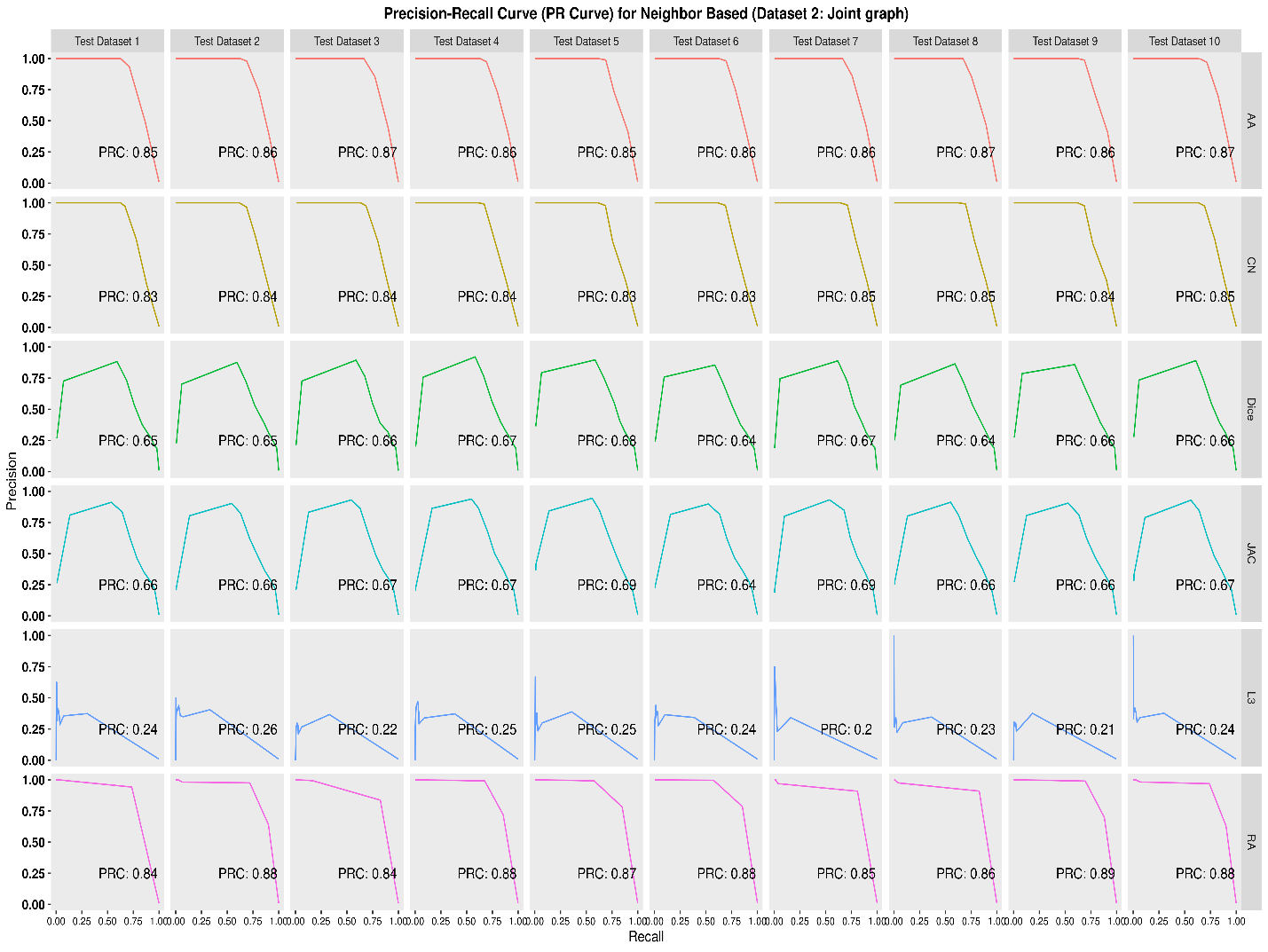


Figure S15: Precision-Recall Curve (PRC) for neighborhood-based similarity-based (dataset 2: joint graph)

1. Table S2 refers the number of links in the graph after applying different food compound contribution score (based on the Tanimoto similarity thresholds 0.6)

**Table S2** Number of links in the graph after applying different food compound contribution score

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Contribution Threshold** | **Total Links** | **DD Links** | **FF Links** | **FD Links** |
| > 0.3 | 155,512 | 2,926 | 152,517 | 69 |
| > 0.4 | 113,348 | 2,926 | 110,400 | 22 |
| > 0.5 | 87,192 | 2,926 | 84,257 | 9 |
| > 0.6 | 67,944 | 2,926 | 65,016 | 2 |

It should be noted that smaller number of FD links does not translate to the possible FDIs that can be examined. We have 87,192 and 92,143 (i.e., for disjoint and joint graphs respectively) possible FDIs when a threshold of > 0.5 is considered.

**Reference**

1. Ali, M. & Ezzat, A.dbparser: ’DrugBank’ Database XML Parser version 1.2.0 from CRAND ainanahan. https://rdrr.io/cran/dbparser/(2020).17