Supplemental materials for: "A Message Passing framework with Multiple data integration for miRNA-Disease association prediction"

Thi Ngan Dong[1,*], Johanna Schrader[1], Stephanie Mücke[2], and Megha Khosla[3]

[*]dong@l3s.de
[1]L3S Research Center, Leibniz University of Hannover, Hannover, Germany
[2]TRAIN Omics, Translational Alliance in Lower Saxony, Hannover, Germany
[3]Delft University of Technology (TU Delft), Netherlands

ABSTRACT

In this supplementary file, we present the details regarding our compared models in Section 1. Section 2 provides the details corresponding to our data acquisition, pre-processing, benchmarked datasets generation, and negative sampling strategies. Finally, in Section 3, we offer a user guide for the website where we encapsulate all our model-generated predictions for 1,618 miRNAs and 3,679 diseases and some related domain knowledge.

1 The compared models

EPMDA[1] EPMDA first builds a heterogeneous graph in which nodes are miRNAs and diseases, and the edges are built up from miRNA-miRNA, disease-disease Gaussian Interaction Profile Kernel (GIP)[2] similarities and miRNA-disease known associations). They then proposed a graph topology-based edge feature extraction technique operating on the constructed graph. The extracted features will then be fed as input to a Multiple Layer Perceptron (MLP) classifier, which will assign 1 or 0 labels to an input edge. Though EPMDA proposed a graph-based feature extraction approach, their input graph is still constructed from the miRNA and disease GIP similarities. Therefore, EPMDA is still classified as a hand-crafted-based method.

NIMGCN[3] NIMGCN proposes an end-to-end learning framework that operates on a heterogeneous network $\mathscr{G}$, which is built up from miRNA-disease known associations, the pre-calculated miRNA functional similarity (MISIM), and the disease semantic similarities. Two GCNs followed by two non-linear transformation layers are employed to learn the latent representation for miRNAs and diseases separately. Though NIMGCN's contribution lies in the graph-based feature transformation technique, the building blocks for the $\mathscr{G}$ are still pre-calculated similarities or hand-crafted features.

DBMDA[4]**.** DBMDA also employs hand-crafted features. The model can be separated into two modules: the unsupervised feature transformation and the Rotation Forest classifier. The unsupervised feature transformation consists of two auto-encoders whose tasks are to learn the hidden representation for a miRNA-disease pair from the miRNA functional similarities (retrieved from MISIM), disease semantic similarities, and miRNA sequence similarities. The encoded representation will then be fed as input into a Rotation Forest classifier whose job is to predict potential miRNA-disease associations.

NEMII[5] NEMII employs Structural Deep Network Embedding (SDNE) to learn the embeddings for miRNAs and diseases from a bipartite graph built up from known miRNA-disease association information. The learned embedding will then be concatenated with the features extracted from the miRNA family and disease semantic similarity to form the input to a Random Forest classifier. Since NEMII still makes use of disease semantic similarities, it is a hybrid technique.

NEMII does not work effectively for large datasets with many new diseases for two main reasons. On the one hand, the Rotation Forest classifier is a tree-based ensemble model whose complexity is at least $O(mn^2 logn)$[6], where m is the feature size, and n is the number of training samples. In NEMII, feature space grows along with the number of miRNAs and diseases. For a large dataset with 3,679 diseases, NEMII is extremely expensive to run with the input feature size of several thousand. On the other hand, as new diseases appear as unconnected nodes in the bipartite input graph, NEMII cannot learn any meaningful structural embeddings for them. We do not have the results available for NEMII in the inductive test setting for those reasons.

DIMIG 2.0[7] DIMIG 2.0 is a semi-supervised approach that treats miRNA-disease association prediction as a multi-class classification problem where diseases are the labels. They do not use miRNA-disease association during the training process.

Instead, they use only the known disease-PCG interactions to learn the model parameters. PCG nodes are connected with miRNA nodes in a heterogeneous network based on miRNA-PCG interactions. Learned signals are then propagated through the heterogeneous network to infer miRNAs' labels. DIMIG 2.0 is a feature learning-based approach.

MUCOMID[8] MUCOMID proposes different ways of integrating additional information sources. Similar to MPM, MUCOMID does not rely on secondary or hand-crafted features but employs graph neural networks to learn miRNA, disease, and PCG representations automatically from three information sources: the PCG-PCG interaction, miRNA family, and disease ontology. However, unlike MPM, MUCOMID does not use miRNA-PCG and disease-PCG associations to construct input features. Instead, MUCOMID incorporates such information as additional side tasks to further regularize the model. A dynamic loss balancing technique is employed to train the multitask model in an end-to-end manner.

2 Experimental data and setup

Table 1. The association data statistics where $|n_{md}|$, $|n_m|$, $|n_d|$ refer to the number of associations, miRNAs and diseases respectively.

DATASET	$\|n_{md}\|$	$\|n_m\|$	$\|n_d\|$
HMDD2	4,592	442	309
HMDD3	10,494	742	545
HMDD2 ∪ HMDD3	10,980	742	591
HELD-OUT	4,311	382	226
HELD-OUT2	6,388	697	509
NOVEL-MIRNA	4,734	638	227

2.1 The miRNA-disease association data source

We retrieve the set of miRNA-disease associations from the HMDD v2.0[9] and HMDD v3.0[10] databases. We then perform various pre-processing and filtering steps as described in Section 2.2. In the end, the filtered data for the HMDD v2.0 database (denoted as HMDD2) contains 4,592 known associations between 442 miRNAs and 309 diseases. The filtered data for the HMDD v3.0 database (referred to as HMDD3) includes 10,494 known associations between 742 miRNAs and 545 diseases.

2.2 Data acquisition and pre-processing

As the quantity and quality of the employed data source greatly impact the predictive power of the learned models, apart from the model development, our contribution also lies in the data acquisition and pre-processing. In the following sections, we describe our data acquisition and pre-processing steps.

2.2.1 Disease ID matching

The data deposited in HMDD 2.0 and HMDD 3.0 only provides disease names. Even worse, different names might refer to the same diseases. In addition, to retrieve the disease ontology or disease-PCG associations, we need the diseases' MESH IDs. Therefore, in the first steps of our pre-processing pipeline, we match the HMDD 2.0's and HMDD 3.0's disease names with their corresponding MESH IDs.

In order to do that, we first collect the list of disease IDs, along with their names and synonyms, from the MESH database[11]. We then standardize all disease names and synonyms (remove redundant spaces, quotations, and convert all to lowercase). After that, our disease matcher works as follows: (i) if there is an exact match between the searching disease name and any MESH names/synonyms, then it assigns the corresponding MESH ID to that disease name (ii) otherwise, it outputs a list of names along with their MESH IDs which are the most similar to the searched name and only contain up to several different characters in the character sequence. We later quickly reviewed these lists to increase the data coverage as much as possible.

2.2.2 miRNA name standardization

The HMDD 2.0 and HMDD 3.0 databases store the known associations reported in scientific publications and do not reflect the changes in the miRNA knowledgebase over time. Therefore, the same miRNA might appear with different IDs in the miRNA-disease association databases. To remove unnecessary noise and make the data consistent, we standardize the miRNA IDs according to miRBase[12] - one of the most reliable and popular databases to retrieve miRNAs related information. More specifically, we match multiple miRNAs aliases together and obsoleted IDs to the newly assigned ones according to the data retrieved from miRBase, version 22.1. Table 2 presents the statistics associated with the number of miRNAs and miRNA-disease associations after standardization.

Table 2. Statistics for miRNA-disease data before and after miRNA name standardization. n_m refers to the number of miRNAs while n_{md} denotes the number of miRNA-disease associations.

Database	Before		After	
	n_m	n_{md}	n_m	n_{md}
HMDD 2.0	578	6,401	540	5,909
HMDD 3.0	1,120	15,165	859	12,552

2.3 The transductive testing setup

The transductive testing setup aims at evaluating different models' performances on the set of partially observed miRNAs and diseases. We train each model with the HMDD2 dataset while testing them with the HELD-OUT test set as described below.

Let **M** and **D** denote the set of miRNAs and diseases observed in the HMDD2 dataset, respectively. We construct the HELD-OUT dataset by restricting the set of miRNAs and diseases to **M** and **D** and including only the miRNA-diseases associations, which appear in the HMDD3 dataset but not in the HMDD2 dataset. A mathematical description of HELD-OUT is given below:

$$\text{HELD-OUT} = (\mathbf{M} \times \mathbf{D} \cap \text{HMDD3}) \backslash \text{HMDD2}$$

where $\mathbf{M} \times \mathbf{D}$ denotes the set of all possible pair combinations between miRNAs in **M** and diseases in **D**. Table 1 presents the transductive training and testing data statistics. We generate the negative training and testing samples using the negative sampling strategy given in section 2.5.

2.4 The inductive setting setup

2.4.1 The large independent testing sets

The HELD-OUT2 test set. HELD-OUT2 contains all associations that appear in HMDD3 but not in HMDD2. We devise this dataset to test all models' performance on a large independent test set that contains both new miRNAs and new diseases (with respect to the training data). After preprocessing, HELD-OUT2 contains 6,388 known associations for 697 miRNAs and 509 diseases. Among those, there are 300 new miRNAs and 282 new diseases that do not appear in the training set HMDD2.

The NOVEL-MIRNA test set. The NOVEL-MIRNA test set is a subset of the HELD-OUT2 test set. To construct NOVEL-MIRNA, we remove all associations related to any disease that does not appear in **D**. In the end, NOVEL-MIRNA contains 4,734 known associations for 638 miRNAs and 227 diseases in which there are 256 new miRNAs that do not appear in the training set HMDD2. The data statistics for our large independent test sets are presented in Table 1.

2.4.2 The datasets for new diseases

Table 3. The individual diseases' training and testing dataset statistics where $|E_{train}|$, $|E_{val}|$, $|E_{test}|$ refer to the number of *positive* training, validating and testing samples respectively.

DISEASE	$\lvert E_{train}\rvert$	$\lvert E_{test}\rvert$	DISEASE	$\lvert E_{train}\rvert$	$\lvert E_{test}\rvert$
D001943	10649	331	D005909	10803	177
D015179	10704	276	D001749	10813	167
D013274	10720	260	D012516	10821	159
D008175	10757	223	D010190	10822	158
D011471	10761	219	D006333	10838	142
D002289	10766	214	D002292	10839	141
D010051	10789	191	D003110	10854	126
D008545	10791	189	D015470	10868	112
D005910	10791	189	D002294	10876	104

The inductive setting setup aims at evaluating models' performances on completely new diseases and is described as follows:

- Let H = HMDD2 $\cup$ HMDD3
- We take out the set of diseases $\hat{\mathscr{D}}$ such that each disease $d \in \hat{\mathscr{D}}$ has more than 100 known associations in H. There are 18 such diseases.

- For each disease $d \in \hat{\mathscr{D}}$, a dataset is created as follows: (i) The positive training set includes all known associations in H except those associated with d, (ii) The negative training samples are generated as described in Section 2.5, (iii) We evaluate all models on the ***complete*** testing set where all known associations for d in H form the positive test set and the negative testing samples consist of all possible combinations of d and any miRNA that does not appear in the positive testing set.

Table 3 presents the statistics corresponding to our 18 datasets for new diseases.

2.5 The negative sampling strategy.

We define the negative pool as the set of all possible combinations of miRNA-disease pairs that do not appear in the set of all known associations. For all training data, we fix the negative:positive ratio to 1:1. For the independent testing sets (HELD-OUT, NOVEL-MIRNA, and HELD-OUT2), we vary the ratio as {1:1, 1:5, 1:10}. For each negative:positive sample rate, we randomly draw 10 subsets from the negative pool and evaluate each method's performance on all those sampled sets to avoid bias and make the comparison as fair as possible. In summary, in the transductive setting, we have 10 train and 10 test sets (corresponding to different negative sample sets). We evaluate each model by training it on all 100 train and test set combinations, each with 2 random model initialization. In total, we report the average results corresponding to 200 experimental runs for the transductive setting.

For the inductive setup, we use the entire set of unknown interactions as negative test samples. We run the model with 10 train sets each time with 2 random initialization of the model. For the inductive setting, we, therefore, report the average results over 20 experimental runs.

2.6 Hyperparameter setup and implementation details

MPM and its variants. We experiment with the number of Message Passing iteration t in [1, 2, 10]. For the feature selection module, we run ReliefF[13] with 20 neighbors and the number of selected features K varying from 50 to 500 with a step size of 50. The results reported in Section 2 in the main paper correspond to $t = 1$ and $K = 100$, which result in the best average AP score among 18 datasets in the inductive test setting. For SDNE, we use the default parameter as suggested by NEMII[5] with the embedding size fixed to 128. The Random Forest classifier is trained with 350 estimators.

Existing benchmarked models. For EPMDA, DBMDA, and NIMGCN, we use the code and setup released in[14]. For NEMII and MUCOMID, we use the same code and setup as published by the authors. For DIMIG 2.0, we follow the same testing strategies employed in[8].

3 An easy-to-use web application

3.1 Biological related features to support biologist justification and verification

As the associated pathway information is more intuitive compared with the list of associated PCGs, we perform pathway and functional enrichment analysis on the list of interacting/associated PCGs for each miRNA/disease and encapsulate the corresponding information into our Windows application. We perform pathway enrichment analysis by using the API provided by Reactome[15] and functional enrichment analysis by using the goscripts package[16]. We retain only pathways and GO terms whose p-values are smaller than 0.05.

3.2 The user guide

We provide an easy-to-use web application (web app) to query MPM's predictions and additional information on the miRNA, disease, and pathway data used in this work: http://software.mpm.leibniz-ai-lab.de/. In the following section, we briefly present a user guide and the functionality of the provided application.

Figure 1 shows the start screen when opening the application tab in the web app and illustrates the main steps to use it. First, the user selects the **i) Main Category** from the tabs at the top of the application, i.e. *miRNA*, *Disease* or *Pathway*, marked with i) in Figure 1. In the next step **ii) Entity Selection**, the user selects a specific entity from that main category by either typing a valid entity name in the search field or by selecting an entity from the drop-down menu. The drop-down menu (which also serves as a search field) is marked with ii) in Figure 1 and opens upon selection. After a specific entity to inspect is selected, the user chooses the **iii) Information Type** they want to display by selecting the corresponding tab, marked with iii) in Figure 1.

Inspecting miRNAs

If the user wants to inspect a specific miRNA, they can choose from *miRNA-Family*, *GO information*, *Pathway Information*, *Disease Associations* (which shows confirmed associations from the data) or *Disease Predictions* (which shows associations that are predicted by MPM) to query the desired information type. The user selects the information type to query by pressing the corresponding tab from the bar below the text field, marked with ii) in Figure 1. *miRNA-Family* will display all miRNAs

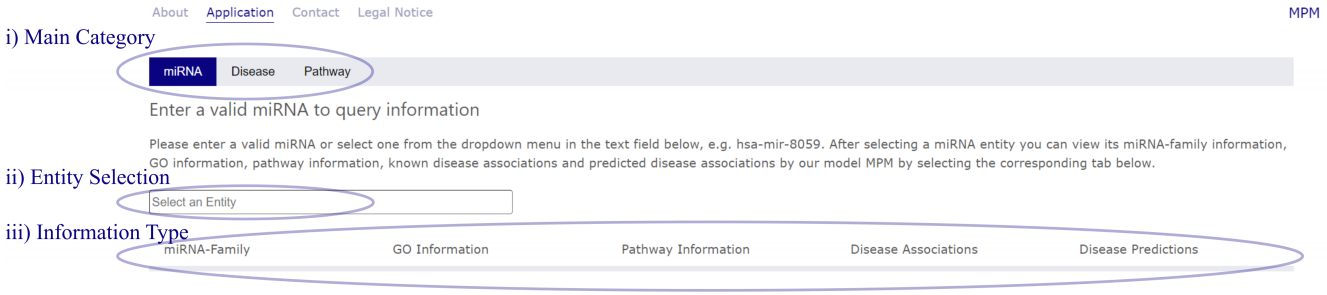

Figure 1. Web app: i) Main Category: First, the user selects whether they want to inspect a miRNA, disease, or pathway by clicking on the corresponding tab at the top of the application. ii) Entity Selection: Then, they choose a specific entity to inspect. Either the user selects one from the drop-down menu or types in the text field using auto-completion. iii) Information Type: Finally, the user selects the information they want to display by selecting the corresponding tab.

that belong to the same family as the selected miRNA entity. *GO information* provides the GO ID with its corresponding name and the belonging p-value for the selected miRNA entity. *Pathway information* will display all pathways and their names that the selected miRNA entity is occurring in. The pathway information additionally provides the corresponding p-value for that miRNA entity in each pathway. The GO and pathway information are sorted ascending by their p-value. The *Disease Associations* option displays all diseases associated with the selected miRNA entity that are known associations from the data. The MeSH ID with the corresponding disease name is provided. Finally, the *Disease Predictions* option provides the predictions made by MPM for the selected miRNA entity. Each predicted associated disease is displayed with its MeSH ID, disease name, and the confidence score of that prediction in descending order. Additionally, the column *Confirmed Association* shows if this specific association was known before, indicated by *yes* and - otherwise.

Inspecting Diseases

A disease can be inspected analogously to a miRNA. Similar to the miRNA category, after selecting a specific disease entity, the disease category as well allows the user to display the *GO Information*, *Pathway Information*, confirmed *miRNA Associations* to the selected disease entity as well as *miRNA Predictions* for the selected disease entity made by MPM. Contrary to the miRNA family information in the miRNA category, in the disease category, the user can display information on the *Disease Ontology*, i.e., the child and parent diseases of the selected disease entity. Example output for predictions on miRNA associations to the disease *Amyloidosis* by MPM is shown in Figure 2. The predicted associated miRNAs are shown in the left column, with their corresponding confidence score in the middle column. The right column indicates whether this association was known from the data before by *yes* or - otherwise.

Inspecting Pathways

When inspecting specific pathways, the user can choose between displaying the most significant miRNAs or diseases corresponding to the selected pathway entity. Figure 3 shows an example query for the pathway *Establishment of Sister Chromatid Cohesion*s most significant *Disease Associations*. The diseases are sorted ascending by their p-value in the right column, with the corresponding disease ID in the left and the disease name in the middle column.

References

1. Dong, Y., Sun, Y., Qin, C. & Zhu, W. Epmda: Edge perturbation based method for mirna-disease association prediction. *IEEE/ACM Transactions on Comput. Biol. Bioinforma.* (2019).
2. van Laarhoven, T., Nabuurs, S. B. & Marchiori, E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* **27**, 3036–3043 (2011).
3. Li, J. *et al.* Neural inductive matrix completion with graph convolutional networks for mirna-disease association prediction. *Bioinformatics* (2020).
4. Zheng, K. *et al.* Dbmda: A unified embedding for sequence-based mirna similarity measure with applications to predict and validate mirna-disease associations. *Mol. Ther. Acids* **19**, 602–611 (2020).
5. Gong, Y., Niu, Y., Zhang, W. & Li, X. A network embedding-based multiple information integration method for the mirna-disease association prediction. *BMC bioinformatics* **20**, 1–13 (2019).

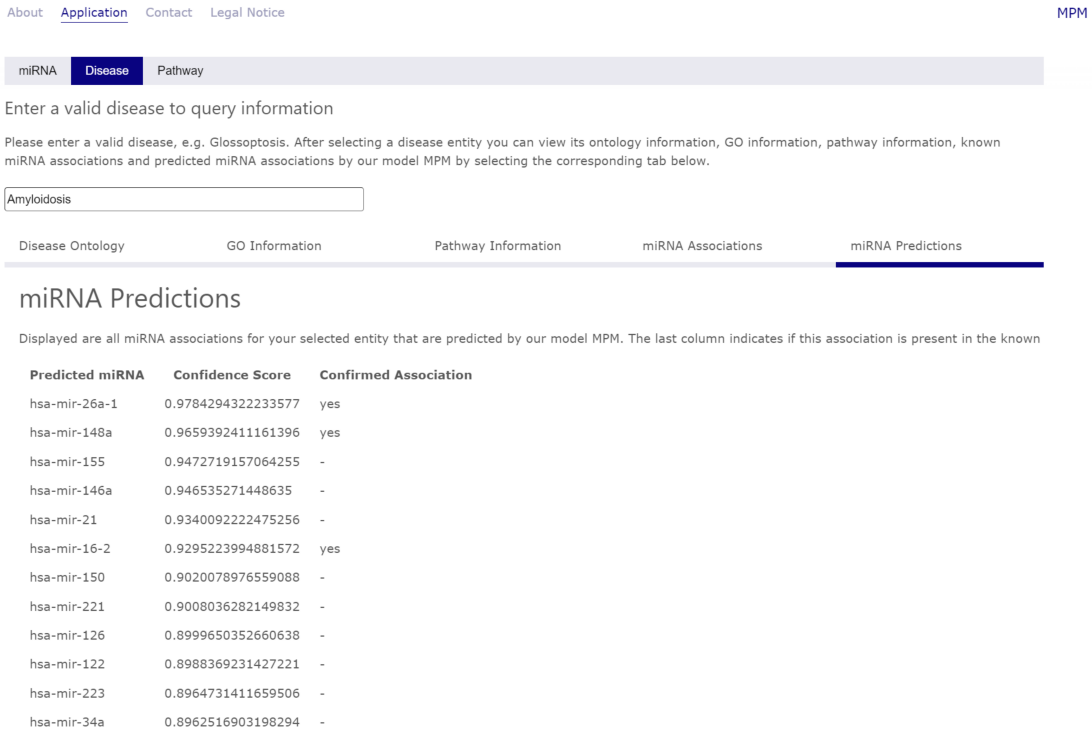

Figure 2. Web app: Display of *predicted miRNAs* for the disease *Amyloidosis*. Shows a list of all *predicted miRNA* associations (left column), with their *confidence score* (middle column) and the indication if this is a *confirmed association* (right column), in which a confirmed association is indicated with *yes* and - otherwise.

6. Bagnall, A. *et al.* Is rotation forest the best classifier for problems with continuous features? *arXiv preprint arXiv:1809.06705* (2018).

7. Pan, X. & Shen, H.-B. Scoring disease-microrna associations by integrating disease hierarchy into graph convolutional networks. *Pattern Recognit.* 107385 (2020).

8. Dong, T. N. & Khosla, M. Mucomid: A multitask convolutional learning framework for mirna-disease association prediction. *arXiv preprint arXiv:2108.04820* (2021).

9. Li, Y. *et al.* Hmdd v2. 0: a database for experimentally supported human microrna and disease associations. *Nucleic acids research* **42**, D1070–D1074 (2014).

10. Huang, Z. *et al.* Hmdd v3. 0: a database for experimentally supported human microrna–disease associations. *Nucleic acids research* **47**, D1013–D1017 (2019).

11. Bhattacharya, S., Ha-Thuc, V. & Srinivasan, P. Mesh: a window into full text for document summarization. *Bioinformatics* **27**, i120–i128 (2011).

12. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. mirbase: from microrna sequences to function. *Nucleic acids research* **47**, D155–D162 (2019).

13. Kononenko, I. Estimating attributes: analysis and extensions of relief. In *European conference on machine learning*, 171–182 (1994).

14. Dong, T. N. & Khosla, M. Towards a consistent evaluation of mirna-disease association prediction models. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1835–1842 (IEEE, 2020).

15. Fabregat, A. *et al.* The reactome pathway knowledgebase. *Nucleic acids research* **46**, D649–D655 (2018).

16. Moris, P. Python script and package for gene ontology enrichment analysis. https://pmoris.github.io/goscripts/_build/html/source/README.html.

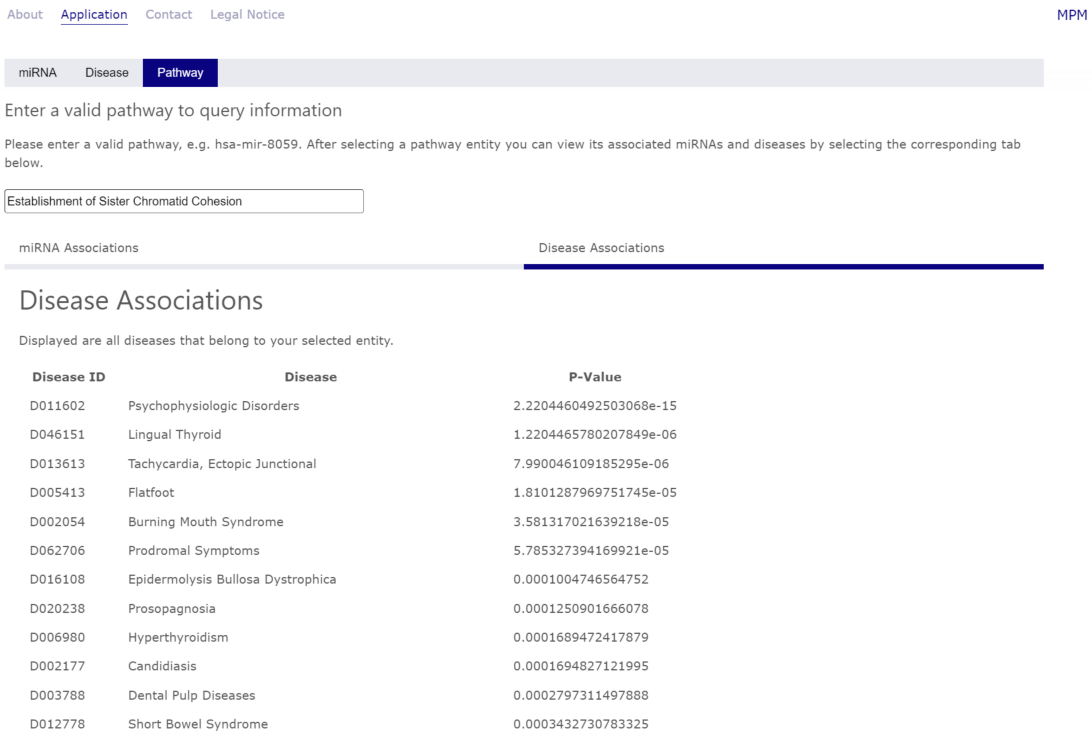

Figure 3. Web app: Display of *most significant diseases* for the pathway *Establishment of Sister Chromatid Cohesion*. The left column shows the diseases mesh ID with the corresponding disease name in the middle column. The p-value for each listed disease is displayed in the right column. The entries are ordered in ascending order by their p-value.