GROUP-10 MUTANT NINJA TURTLES

Multi-Link Prediction analysis on Hetionet's breast cancer data, utilizing Neo4j Knowledge graph capabilities

Blessy Rajan and Shubhangi Kaushik

Full list of author information is available at the end of the article

Abstract

The goal of the project: The goal of this project was to perform multi-link prediction analysis between "Disease"-"Gene", and between "Gene"-"Compound" of breast cancer data from Hetionet using Pykeen. Followed by the biological evaluation of the results obtained utilizing the bioinformatics databases available.
Main results of the project: The predicted links between the "Disease"-"Gene" and "Gene"-"Drug Compound" provide insight into the connection between the disease, gene, and the drug compound not covered in the Hetionet data. Further analysis through bioinformatics databases broadened the perspective into disease treatment.
Personal key learning: Gained knowledge about the Knowledge graphs, and exposure to Neo4J tool. Moreover, the integration of Python with Neo4J enabling the utilization of Python capabilities on top of the querying and visualizations of Neo4J was enriching.
Estimated working hours: 22 hours(5 hours: Executing RotatE, 5 hours: Executing ProjE and QuatE, 3 hours: Executing TransE and ComplEx, 2 hours: Analysis the results, 5 hours: Writing report)
Project Evaluation: 1: (The exposure to Neo4J was really enlightening, as it provided exposure to working with knowledge graphs. Neo4J provided a very informative visual aid to the Predictive analysis results from the Pykeen models.
Challenges Faced: Most of the challenge was faced while making the models run. As the models were running on CPU, the time consumed for running the models even on a very small epoch was a lot.
Word Count: 2344 (including an abstract(234)).

Goal

The goal of this study was to analyze the Multi-link prediction with the help of Pykeen models, between the disease and genes, and drug compound and gene of the breast cancer data from the subset of the Hetionet knowledge graph.
The models from Pykeen chosen for the multi-link prediction were QuatE, ProjE, ComplEx, and TransE.
The first use case for this study was the link prediction between the disease and gene for the breast cancer data in order to find out the genes that are predicted to be upregulated in breast cancer. This prediction builds a relation between the genes and the disease which was not present earlier in the Hetionet data.
The second use case was built on the predicted gene result from the first use case.

The gene which was found to be common in the models was chosen for this use case. In this study, SMC4 was found to be one of the common genes predicted by the Pykeen models. The compounds which downregulate the SMC4 gene were predicted. And further analyses were conducted using bioinformatics databases to solidify the claim.
This study helped to identify the genes that might be upregulated in breast cancer, along with the drug that could help downregulate that particular gene in order to treat breast cancer. Hence, this study could provide insight into how the disease and the drug are connected through genes, which could prove important in drug repurposing.

Data

Hetionet was created using a comprehensive and collaborative methodology to construct an interconnected biomedical network. By gathering data from 29 publicly available resources, a vast repository of data on genes, compounds, diseases, pathways, and other biological elements was compiled. Relationships among these entities were extracted to form an extensive network capturing various biological interactions. The project used an open and collaborative approach, seeking input from domain experts and the scientific community through the Thinklab platform. This transparent and collective effort resulted in Hetionet becoming a valuable resource for biomedical research, drug discovery, and repurposing.
The data used in this study is a subset of the Hetionet dataset, consisting of the data for genes, compounds, and Diseases (Figure 3). The Knowledge Graph helps in visualizing the relation between different nodes, i.e. disease, gene, and compound. The original Hetionet data consists of several nodes, however, the subset used in this study consists of genes, compounds, and diseases. The knowledge graph provides data about whether these elements are up or down-regulating the other element, along with other information like binding, treatment, and many more. The knowledge graph also provides data about how closely are the genes and drug compounds connected to the disease, i.e. whether there is a direct connection or an indirect one.

Methods

The multi-class link prediction was done using the following models from PyKeen:

ComplEx

ComplEx embedding[9] for link prediction utilizes complex-valued vectors to represent entities and relations in a knowledge base. The relation probabilities between entities are predicted by computing the dot product while keeping anti-symmetry in account. The resulting dot product is a complex number representing the score for a triplet(h, r, t), where h represents head, r represents relationship, and t represents tail entity. The magnitude of the complex number is the predicted probability of the relation between the two entities. The higher the magnitude, the more probable the relation is to exist between the two entities. ComplEx is able to capture and utilize the complex interplay of features, making it a promising approach for diverse applications beyond link prediction.

TransE

TransE[5] employs entity and relationship embeddings in a low-dimensional vector space for prediction. Entities and relationships are represented as vectors, where the relationship vector is derived from the sum of the involved entity vectors. Given a triplet (h, r, t), TransE computes embeddings for each. It then measures the distance between the translated head entity vector and the tail entity vector using an L1 or L2-based scoring function. During training, using stochastic gradient descent, TransE optimizes the embeddings by minimizing the distance for true triplets and maximizing it for false triplets. This approach facilitates effective prediction and knowledge graph completion through learned vector representations.

QuatE

QuatE[11] makes use of quaternions in the analysis, which is a hypercomplex number system that extends the traditional complex number system to four dimensions. In QuatE, predicting the missing entity in a knowledge graph triplet (h, r, t) involves computing a score for each possible tail entity t. The score is obtained by measuring the distance between the quaternion product of the head entity h and the relation matrix r and the quaternion embedding of the tail entity t. The missing entity is predicted as the tail entity with the highest score. The distance metric utilized in QuatE is the L2 distance, considering the real parts of the quaternion embeddings. The framework is trained using a margin-based ranking loss function that aims to maximize the margin between the score of the correct tail entity and the scores of incorrect tail entities.

ProjE

ProjE[8] makes predictions by projecting candidate entities onto a shared embedding space through a combination operator, creating a target vector from input data. Candidate similarity scores are then calculated based on their projection onto the target vector. Training uses a triple set, generating positive and negative candidates through sampling. The model learns to optimize collective scores of candidate entities by adjusting parameters of the combination operator and embedding matrices. For prediction, known entities and relationships are projected onto the shared space, computing similarity scores for candidate entities. The candidate with the highest score is selected as the predicted entity or relationship in the knowledge graph.

Results and Discussion

Initially, the Neo4j graph underwent a transformation to convert to a PyKeen graph. Subsequently, the dataset was partitioned into three subsets: training, testing, and validation, with an allocation ratio of 8:1:1. The knowledge graph embedding model was trained using the methods ComplEx, TransE, QuatE, and projE models.

For use case 1, the head node was considered as the disease "breast cancer", the gene node in the knowledge graph was considered to be the tail node, and the relation to be predicted here was the upregulation of the genes in breast cancer. This prediction was carried out using all the mentioned models and the top 50 candidate genes were selected. The top 50 genes were selected to find the common genes between these

methods.

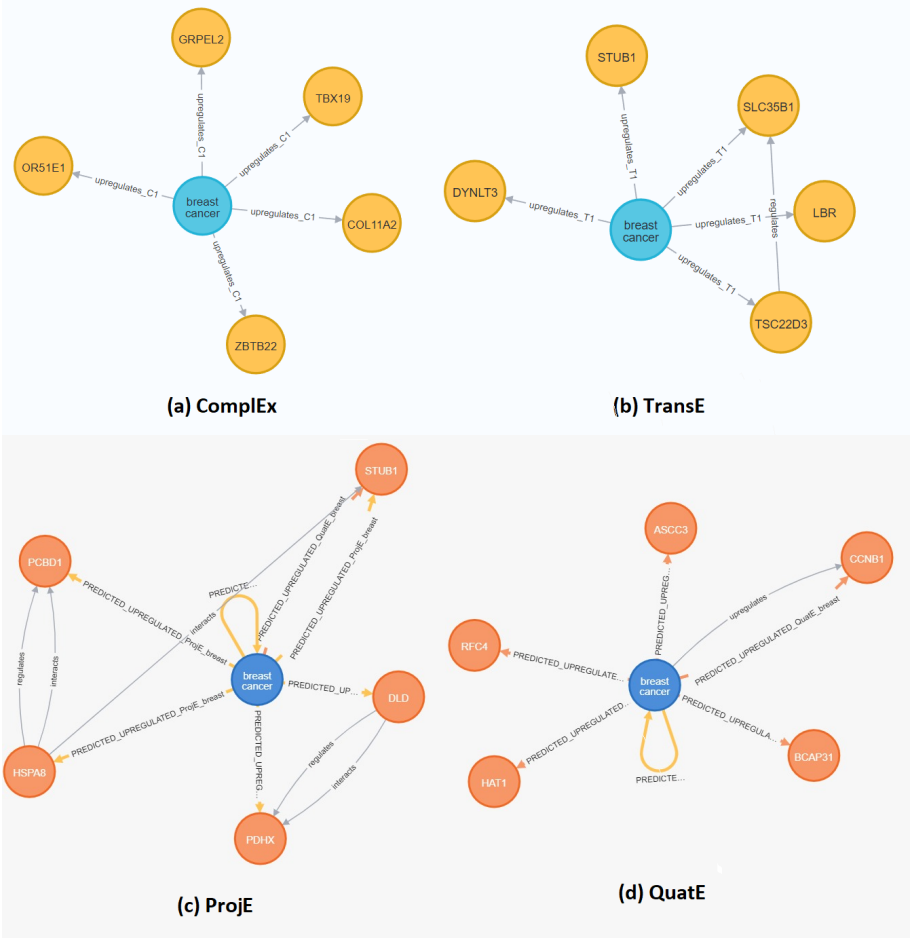

Figure 1: (a) shows the top 5 genes predicted to be upregulated in Breast Cancer by the ComplEx model. The relation between them is denoted by "upregulates_C1", (b) shows the top 5 genes predicted to be upregulated by the TransE model. The relation between them is denoted by "upregulates_T1", (c) shows the top 5 genes predicted to be upregulated by the ProjE model. The relation between them is denoted by "PREDICTED_UPREGULATED_ProjE_breast", (d) showing the top 5 genes predicted to be upregulated by the QuatE model. The relation between them is denoted by "PREDICTED_UPREGULATED_QuatE_breast".

From the top 50 genes predicted to be upregulated in breast cancer by the models, the genes in Table 1 were found to be common.

Gene	ComplEx	TransE	ProjE	QuatE
STUB1	Not Predicted	Predicted	Predicted	Predicted
SMC4	Not Predicted	Predicted(indirect link)	Predicted	Predicted
CCNA2	Not Predicted	Not Predicted	Predicted	Predicted

Table 1: The common genes predicted by the four algorithms. ComplEx algorithm gave very different results in comparison to the rest of the three models, TransE, ProjE and QuatE. STUB1 gene was found to be common in TransE, ProjE and QuatE models. SMC4 gene was CCNA2 was common in ProjE and QuatE models. For TransE, SMC4 was found to be connected to the disease breast cancer via the gene DYNLT3. DYNLT3 was among the top 5 predicted genes by the TransE model. CCNA2 was found to be common in the predictions of ProjE and QuatE models.

As ComplEx did not give any common genes, we decided to investigate the literature regarding the obtained results. Upon examination, it was discovered that none of the genes in the top 5 were associated with Breast Cancer. In contrast, for the other three algorithms, multiple genes from the top 5 results were found to be linked to Breast Cancer based on existing research. For example, DYNLT3[10], which was in the top 5 results for the TransE model, is found to have a role in the proliferation, migration, and invasion of Breast Cancer. From the QuatE model, upregulation of the APOBEC3B gene was found to be connected to the metastases of breast cancer, as compared to the corresponding ER-positive primary tumors[2]. This makes APOBEC3B a promising target for anti-APOBEC3B therapies[2]. From the ProjE model, the upregulation of the SPP1 gene was found to be associated with an increased risk of recurrence in ER+ breast cancer[3].

Among the genes of interest from Table 1, further analysis was decided to be conducted on the "SMC4" gene. The decision to move forward with the "SMC4" gene was based on its biological relevance to breast cancer, obtained from the bioinformatics database like Pubmed. While going through Pubmed, research papers were found providing evidence of upregulation of the "SMC4" gene in breast cancer. This supported the prediction provided by the models used in this study[7]. According to the research paper the upregulation of "SMC4" is related to the increased mortality rate in breast cancer, hence, "SMC4" mRNA level was considered to be a good prognostic biomarker for patients with breast cancer.[7]

In order to have a proper workflow between use case 1 and use case 2, SMC4 gene was considered as the gene, for which the compounds were predicted to downregulate its expression. Here, the gene "SMC4" was considered to be the tail node whereas the compound to be predicted was considered as the head node, the relation was considered to be the downregulation of the gene via the predicted compound. The prediction was carried out using all the mentioned models and the top 50 compounds were selected to find the common genes between the models used.

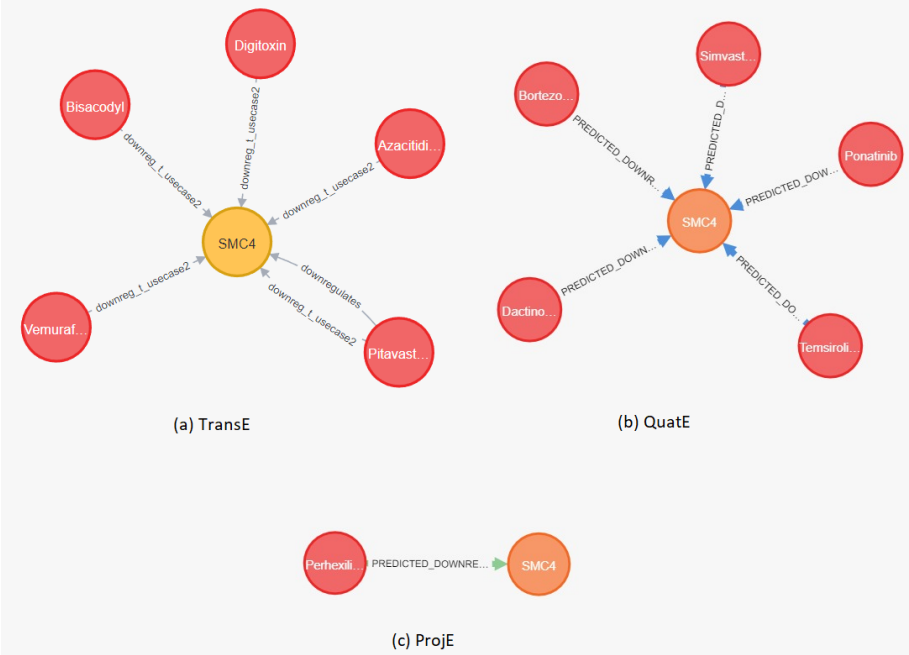

Figure 2: (a) shows the top 5 predicted compounds downregulating SMC4 by the TransE model. The relation between them is denoted by "downred_t_usecase2", (b) shows the top 5 predicted compounds downregulating SMC4 by the QuatE model. The relation between them is denoted by "PREDICTED_DOWNREGULATED_QuatE_breast", (c) shows the only compound downregulating SMC4 by the ProjE model. The relation between them is denoted by "PREDICTED_DOWNREGULATED_ProjE_breast"

The compounds which were found to be in the top 5 of all the models used in this study showed good relation to the treatment of breast cancer. Table 2 denotes the common compound found in TransE and QuatE, although There were no common compounds found between ProjE and the other two models, moreover, the number of predicted compounds using ProjE was only one. Nonetheless, all the drugs found

at the top had relevant data in Pubmed, related to the treatment of breast cancer. For the TransE model, the Vemurafenib compound was found to have a downregulation effect on the genes related to breast cancer. Based on the research paper found in Pubmed[6] Vemurafenib downregulates the expression of RIPK4 accompanied by the 1.9 fold increase of CDK14, and CDK14 downregulates the tumor growth.[6]. For the QuatE model, Sorafenib was found to be one of the top downregulating compounds. It has a promising effect on the downregulation of the secretome genes involved in invasive and angiogenic activity in tumor cells. Although as a single agent, it doesn't show promising effects in breast cancer treatment, however along with other classical chemotherapeutics have shown to be promising in breast cancer treatment.[1]. Sorafenib combined with cyclophosphamide led to smaller tumors compared to either drug alone. Tumors from combined treatment were noted to be well-encapsulated, fibrotic, with a pushing border phenotype and decreased angiogenesis, therefore suggesting a "healed" phenotype[1]. For the ProjE model, Perhexiline led to tumor growth suppression in obesity-associated breast cancer.[4] Although the study states that the perhexiline may act through multiple targets to mediate its anti-tumor effects, nonetheless has the potential to modulate tumor-infiltrating immune cells, which may further enhance its anti-tumor efficacy in vivo.[4] Along with all these analyses a complete graph was built as mentioned in figure 4 to showcase the full network of diseases to compound, with genes playing the role of intermediator nodes. Based on this analysis CCNB2 was found to play a role in the network. This gene was also found to be the common gene predicted in ProjE, QuatE and TransE.

Although the results obtained in this study by utilizing QuatE, TransE, and ProjE models from Pykeen were satisfactory, nonetheless these results could be improved by using more number of epochs while running the models. As the models were running on the CPU, hence the time taken to run each epoch was a lot. Therefore only 3 epochs were used to run each model. In the beginning 20 epochs were used to test the running of the models, however, the system froze and the results were not obtained, eventually it was decided to run the models on the number of epochs that would be able to provide results without disrupting the working of the system.

Compound	TransE	QuatE
Dactinomycin	Predicted	Predicted
Bortezomib	Predicted	Predicted
Simvastatin	Predicted	Predicted

Table 2: The common compound found to be predicted in TransE and QuatE model

Discussion

3 out of the 4 models were able to predict the genes and later drugs of biological significance with the disease Breast Cancer. Interestingly, the model which was expected to perform the best based on its previous performance with other studies[5], ComplEx, performed the worst in our analysis. This could be due to us running the models on 3 epochs. The analysis resulted in valuable outcomes from both TransE and QuatE, successfully predicting genes that are upregulated in breast cancer, as well as identifying compounds that could potentially downregulate the genes

associated to breast cancer. Similarly, ProdE was effective in predicting several significant genes during the analysis. Although ProdE could only predict one significant compound, it still played a crucial role in the context of breast cancer research.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Blessy Rajan:- Did the analysis for the sample model RotatE and the analysis for our project using models QuatE and projE. In the report, abstract, goal and results.
Shubhangi Kaushik:- Did the analysis for the sample model RotatE and the analysis for our project using models TransE and ComplEx. In the report, data, methods and results.
. . .

Author details

References

1. 3 Subapriya Rajamanickam 1 Eva Loranc 1 Pragathi Masamsetti 1 Aparna Gorthi 1 2 July Carolina Romero 1 2 Sonal Tonapi 1 2 Rosangela Mayer Gonçalves 3 Robert L. Reddick 4 Raymond Benavides-5 John Kuhn 4 5 Yidong Chen 6 Alfeu Zanotto-Filho, 1 and 2 Alexander J. R. Bishop1. Sorafenib improves alkylating therapy by blocking induced inflammation, invasion and angiogenesis in breast cancer cells. *Pubmed*, 15, 2018.
2. 2 * Willemijne A. M. E. Schrijver 3 Simone U. Dalm 4 Vanja de Weerd 1 Cathy B. Moelans 3 Natalie ter Hoeve 3 Paul J. van Diest 3 John W. M. Martens-1 2 Anieta M. Sieuwerts, 1 and Carolien H. M. van Deurzen5. Progressive apobec3b mrna expression in distant breast cancer metastases. *Pubmed*, 2017.
3. Lana Othman Scott Montgomery Göran Andersson Elisabet Tina Anna Göthlin Eremo, Kajsa Lagergren. Evaluation of spp1/osteopontin expression as predictor of recurrence in tamoxifen treated breast cancer. *Nature*, 2020.
4. 2 Yoko Tomita 2 3 4 Paul Drew 1 2 Timothy Price 2 3 4 Guy Maddern 1 2 Eric Smith 1 2 3 * Bimala Dhakal, 1 and * Kevin Fenix1, 2. Perhexiline: Old drug, new tricks? a summary of its anti-cancer effects. *Pubmed*, 2023.
5. Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.
6. Methodology Validation Formal analysis Investigation Data curation Writing – original draft Visualization Project administration Funding acquisition 1 Anna A. Brożyna 2 Agnieszka Adamczyk Formal analysis Resources 3 Norbert Wronski Formal analysis 1 Agnieszka Harazin-Lechowska Investigation 3 Anna Muzyk Formal analysis 1 Krzysztof Makuch Formal analysis 4 Michal Markiewicz Formal analysis 4 Janusz Rys Investigation Resources 3 Ewelina Madej, Conceptualization and Methodology Formal analysis Investigation Resources Writing – review editing Supervision Project administration Funding acquisition1 * Agnieszka Wolnicka-Glubisz, Conceptualization. *Vemurafenib and Dabrafenib Downregulates RIPK4 Level*. Pubmed, 2023.
7. Du-ping Huang Min Zheng Rui-min Ma, Fan Yang and Yi luan Wang. The prognostic value of the expression of smc4 mrna in breast cancer. *Pubmed*, 2019.
8. Baoxu Shi and Tim Weninger. Proje: Embedding projection for knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
9. Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR, 2016.
10. Han Wang, Xin Chen, Yanshan Jin, Tingxian Liu, Yizuo Song, Xuejie Zhu, and Xueqiong Zhu. The role of dynlt3 in breast cancer proliferation, migration, and invasion via epithelial-to-mesenchymal transition. *Cancer Medicine*, 2023.
11. Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. Quaternion knowledge graph embeddings. *Advances in neural information processing systems*, 32, 2019.

1 Appendix

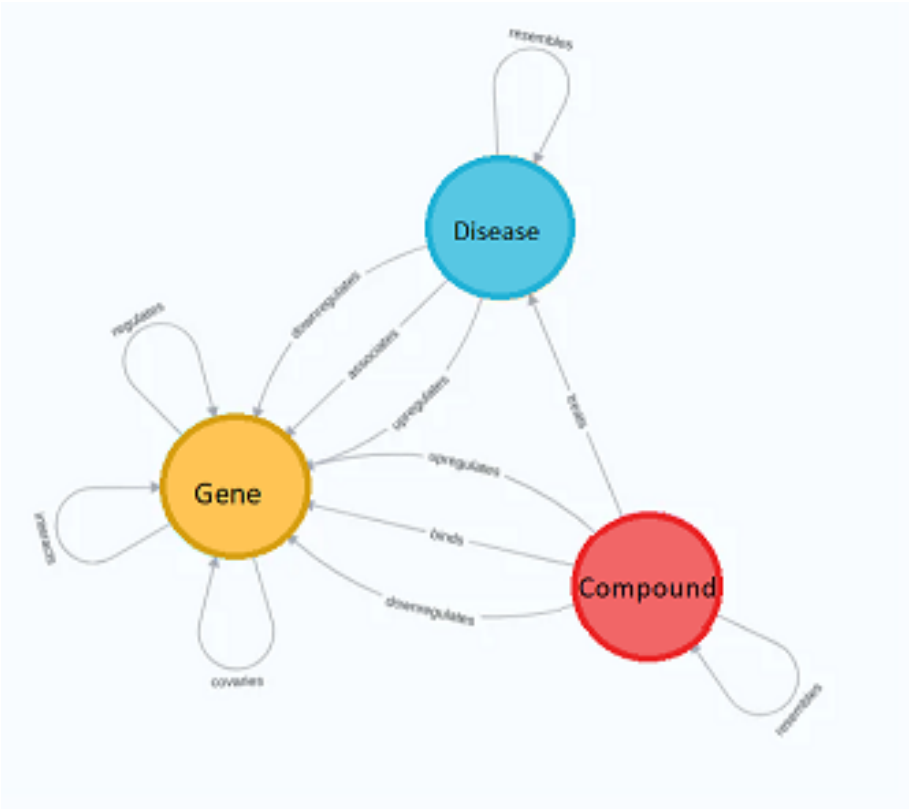

Figure 3: Graph model for the data being used in the analysis. The dataset contains information about Genes, Disease and Compounds and the relationships between them.

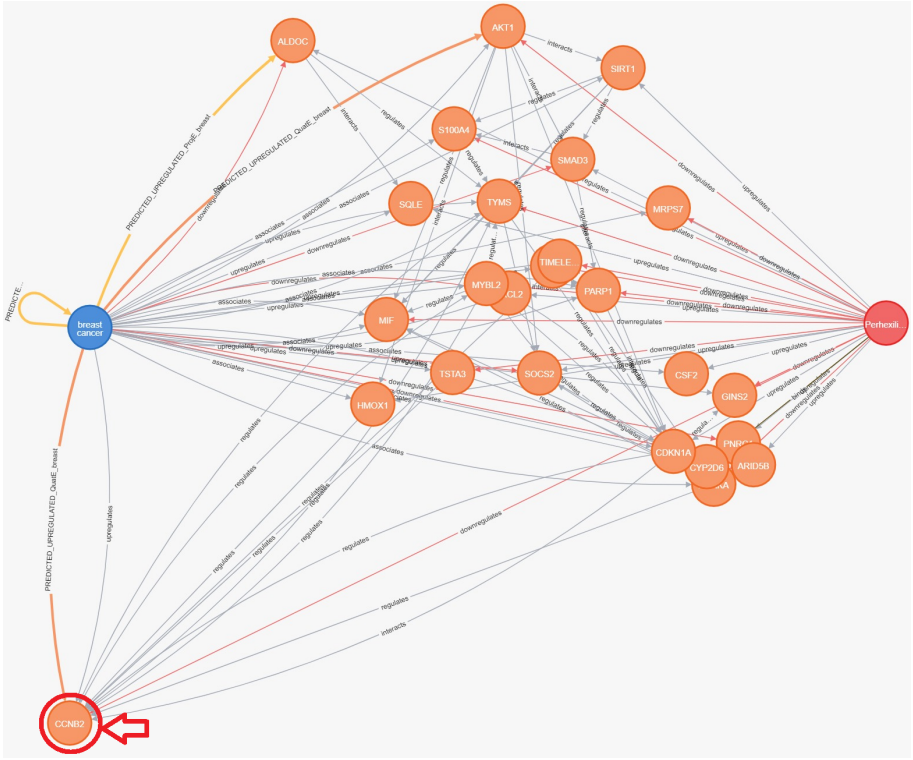

Figure 4: The graph shows the predicted compound Perhixilin is known to dowregulate the gene CCNB2. The same gene was predicted to be upregulating by the disease "breast cancer" by QuatE model.