DEU FMD 26(76) (2024) 167-172

Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering
Elektronik/Online ISSN: 2547-958X

RESEARCH ARTICLE / ARAŞTIRMA MAKALESI

Prediction of Associations between Nanoparticle, Drug and Cancer Using Variational Graph Autoencoder

Varyasyonel Çizge Otokodlayıcı Kullanarak Nanoparçacık, İlaç ve Kanser Arasındaki İlişkilerin Tahminlenmesi

Emrah İnan

Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, İzmir Yüksek Teknoloji Enstitüsü, İzmir, TÜRKİYE
*Corresponding Author / Sorumlu Yazar **: emrahinan@iyte.edu.tr

Abstract

Predicting implicit drug-disease associations is critical to the development of new drugs, with the aim of minimizing side effects and development costs. Existing drug-disease prediction methods typically focus on either single or multiple drug-disease networks. Recent advances in nanoparticles particularly in cancer research show improvements in bioavailability and pharmacokinetics by reducing toxic side effects. Thus, the interaction of the nanoparticles with drugs and diseases tends to improve during the development phase. In this study, it presents a variational graph autoencoder model to the cell-specific drug delivery data, including the class interactions between nanoparticle, drug, and cancer types as a knowledge base for targeted drug delivery. The cell-specific drug delivery data is transformed into a bipartite graph where relations only exist between sequences of these class interactions. Experimental results show that the knowledge graph enhanced Variational Graph Autoencoder model with VGAE-ROC-AUC (0.9627) and VGAE-AP (0.9566) scores performs better than the Graph Autoencoder model.

Keywords: *Variational Graph Autoencoder, Nanoparticles, Drug–disease Association*

Öz

Örtük ilaç-hastalık ilişkilerini tahmin etmek, yan etkileri ve geliştirme maliyetlerini en aza indirmek amacıyla yeni ilaçların geliştirilmesi için kritik öneme sahiptir. Var olan ilaç-hastalık tahmin yöntemleri tipik olarak ya tekli ya da çoklu ilaç-hastalık ağlarına odaklanmaktadır. Özellikle kanser araştırmalarında nanoparçacıklardaki son gelişmeler, toksik yan etkileri azaltarak biyoyararlanım ve farmakokinetikte gelişmeler göstermektedir. Bu nedenle, nanopartiküllerin ilaçlar ve hastalıklarla etkileşimi geliştirme aşamasında iyileşme eğilimindedir. Bu çalışmada, hedeflenen ilaç dağıtımı için bir bilgi tabanı olarak nanopartikül, ilaç ve kanser türleri arasındaki sınıf etkileşimlerini içeren hücreye özgü ilaç dağıtım verilerine varyasyonel bir çizge otokodlayıcı modeli sunmaktadır. Hücreye özgü ilaç verme verileri, ilişkilerin yalnızca bu sınıf etkileşimlerinin dizileri arasında var olduğu iki parçalı bir grafiğe dönüştürülür. Deneysel sonuçlar, bilgi çizgesi ile geliştirilmiş Varyasyonel Çizge Otokodlayıcı modelinin VGAE-ROC-AUC (0.9627) ve VGAE-AP (0.9566) skorlarıyla Çizge Otokodlayıcı modelinden daha iyi performans sergilediğini göstermektedir.

Anahtar Kelimeler: *Varyasyonel Çizge Otokodlayıcı, Nanoparçacıklar, İlaç-hastalık İlişkisi*

1. Introduction

Formulation of drugs into nanoparticles gives drugs new functionalities, such as improved bioavailability and pharmacokinetics by reducing toxic side effects, intensively researched in cancer [1]. Various nanocarrier systems are used to focus on the cells and deliver the drug directly to the target without reaching other places, and the enclosed drug quantity can be expanded by guaranteeing drug-carrier attractiveness [2]. For instance, using ionic liquids in the field of nanoparticle synthesis, it has significant antitumor activity against breast cancer cells [3]. To provide more concentrated knowledge for these critical insights, recent advances enable knowledge bases and ontologies to provide highly accessible and easily updated representations of the interoperable data, allowing domain experts to handle relevant records including their relationships [4]. CancerMine is a well-known sample of this kind of knowledge bases, which was extracted from the scientific literature and updated regularly with minimal human effort [5].

Despite the crucial benefits of the automated knowledge base construction, it requires time consuming operations during collecting the dataset, model training and tuning, particularly for individual researchers or small labs. To generalise these private efforts, the SPIKE-KBC system proposes an extractive-search based knowledge base curation method without requiring extensive training in text mining and natural language processing [6]. In this case, the SPIKE-KBC system presents the cell specific drug delivery data (CSDD) considering the interactions among nanoparticle, drug and cancer types as a targeted drug delivery knowledge base.

There exist three main categories for predicting drug-disease association methods. The first category leverages disease and drug similarities and their pairwise associations. As an example, to estimate potential drug-disease interactions, Gottlieb et al. [7] introduce a technique concentrating on drug indications by computing drug-drug and disease-disease similarity pairs. Luo et al. [8] propose a random walk-based method leveraging a drug–

DOI:10.21205/deufmd.2024267619

Atıf şekli / How to cite:
Inan, E. (2024). Prediction of Associations between Nanoparticle, Drug and Cancer Using Variational Graph Autoencoder. DEUFMD, 26 (76), 167-172.

Geliş Tarihi / Received: 22.05.2023
Kabul Tarihi / Accepted: 15.08.2023

disease network to predict the association probability. Zhang et al. [9] present a similarity constraint-based matrix factorization approach to estimate the novel effects of drugs.

Another category brings together a wide range of data on drugs and diseases. WGMFDDA [10] performs a graph regularized matrix factorization method to infer potential effects of drugs. LRSSL [11] proposes sparse subspace learning for the prediction of implicit drug indications.

As the last category, recent methods use graph based deep learning models to predict the potential relations drug-disease pairs from the implicit association possibilities. For instance, Zhang et al. [12] propose a bipartite graph neural model and a similarity graph to reach the possible interactions between drugs and diseases. Since these are similarity-based methods to extract new drug-disease relationships, graph neural networks have been widely used. GNDD [13] is a graph neural network-based method for estimating drug-disease pair interactions by covering the complex knowledge between these pairs. Yu et al. [14] present graph neural network method by using only critical drugs and disease properties.

Current drug-disease prediction approaches generally concentrate on single or multiple drug-disease neural models. With recent advances in nanoparticles showing improvements, particularly in cancer research, the interaction of nanoparticles with drugs and diseases has the potential to improve the pipeline of new drugs. In this work, we propose a variational graph autoencoder [15] model to the cell-specific drug delivery data, including the interactions between nanoparticle, drug, and cancer types as a knowledge base for targeted drug delivery.

The rest of the paper is structured as follows. We propose a general perspective of the cell specific drug delivery data used in this study with preprocessing steps, the proposed method, including optimized parameters, and used evaluation criteria as a performance measure in study and results. Section 4 presents the conclusion of the study and denotes the future directions.

2. Materials and Method

This section illustrates the data used in this work and summarizes the methods for predicting CSDD knowledge base.

2.1. CSDD knowledge base

The CSDD knowledge base is based on the SPIKE extractive search engine, which is a sentence-level, contextual and linguistically informed extractive engine [16]. By using this engine, the CSDD knowledge base is extracted by adding layers biomaterial (nanoparticles-incorporated biomaterials [17]), ligand (targeting molecule), target and cancer.

The SPIKE extractive search engine discovers patterns using both dependency graphs and token sequences, as well as Boolean keyword queries. Boolean queries that do not consider the order of keywords or groups of keywords in each text. Sequential queries focus on the distance and order of concepts surrounded by anchor words. Syntactic queries highlight the linguistic representations associated with query words. Considering an example of capturing a sequence-based biomaterial entity in the SPIKE-KBC system [6], it filters "Paragraph:"delivery|targeting|nanomedicine", where "|" symbol refers to alternative keywords. Then a sample query "vehicles such as: *" performs wildcard symbol "*" to match any single word. As an example, the placeholders arg1 and arg2 are used to capture the relationships between biomaterial and drug. To extract the relations, the queries arg1: w={biomaterials2} and arg2: w={FDA_DRUGS_YS1} are defined with the filter drug delivery:("abstract"). For the sample query, 'arg1' and 'arg2' are used to explore if the same sentence contains domain-specific keywords. In this case the number of results is 3431 and there are 1192 relations.

The annotation module comprises these extracted entities and relations and there are 3 annotators to review the extracted output [6]. If there is an approval or rejection of any instance, it ignores additional instances of the same entity-entity pair. Therefore, the final knowledge base instances are not proportional to the exact number of query results.

Considering a randomly selected sample in this knowledge base, the first entity type biomaterial is "alginate" and it connects to the drug "zidovudine", and the reference title for this triple <alginate, CONNECTS_TO, zidovudine> is "Encapsulation of zidovudine in PF-68 coated alginate conjugate nanoparticles for anti-HIV drug delivery [18]". Further, the reference sentence is "In this study, the anti-viral drug <e2>zidovudine</e2> (AZT) was encapsulated inside the amide functionalised <e1>alginate</e1> nanoparticles (AZT-GAAD NPs) using emulsion solvent evaporation method". Thus, it implies different sequences of a biomaterial linked to a specific ligand, and it can deliver a drug to a critical cellular target for a specific cancer type. In this case, the goal of targeted drug delivery in cancer research is to improve the efficacy of anticancer drugs while minimising toxicity. Another direction in this research area is to focus on a nanoparticle that delivers drugs to different types of cancer by binding to a cellular target. As a result of this direction, the CSDD knowledge base extracts all the links between biomaterials, cancers, targets, ligands, and drugs annotated by the four researchers before inserting them into the knowledge base.

Table 1. The statistics of the CSDD knowledge base.

Entity Type	Entity Size
Biomaterial	61
Cancer	53
Cell type	29
Drug	439
Ligand	219
Target	173

To construct the public nanomedicine knowledge base to deliver drugs for specific type of cancers, the CSDD knowledge base compiles entities such as drugs, biomaterials and cell types using SPIKE queries along with knowledge resources such as DrugBank [19] and Human Protein Atlas [20]. It states that it collects 910 drug entities from DrugBank (but there is no further description of the distinctness). They also serve 10 relations between entities including biomaterial-drug, biomaterial-target, and drug-cancer. As denoted in Table 1, this knowledge base contains 61 biomaterials, 53 cancers, 29 cell types, 439 drugs, 219 ligands and 173 targets, linked with 6089 annotated relations. Due to the existence of synonyms and acronyms for many biomaterial and ligand entities, the duration of the knowledge base construction is four to five weeks.

To define relations, they assume that the links between biomaterials, ligands, targets, and drugs might indicate reasonable and critical associations, as biomaterials tend to confer critical intrinsic properties, including protein binding and immune cell evasion. For some cases, explicit relation types exist, e.g., ligand-target or "is_used_to-treat" relation in drug and disease entity pairs. With respect to these implicit and explicit

relation extraction phases, there are critical sequences reflecting entity and relation combinations. In this situation, for instance, the first sequence comprises cell type, ligand, target, cancer, drug, and biomaterial. In the PubMed corpus [21], this sequence reaches the highest number of hits considering the assigned combinations.

2.2. Proposed method

The CSDD knowledge base can be represented as a special case of graphs in the form of sequences. Graph convolutional networks (GCNs) generalize the convolution concept of an image by updating an embedding of a pixel with the aggregated information passed by all other neighbor pixels in non-grid-like structures [22]. In the graph autoencoder (GAE) model, an encoder maps the given graph into a lower dimensional space, and then a decoder rebuilds the given graph from the low-dimensional embedding model. The main objective is then to optimize the model by minimizing the reconstruction loss.

In this study, we employ the Variational GAE (VGAE) [15] as a similar version of the GAE model in which the VGAE leverages a multivariate Gaussian distribution as the output heads of the encoder model, rather than encoding each node as a particular point in the latent embeddings.

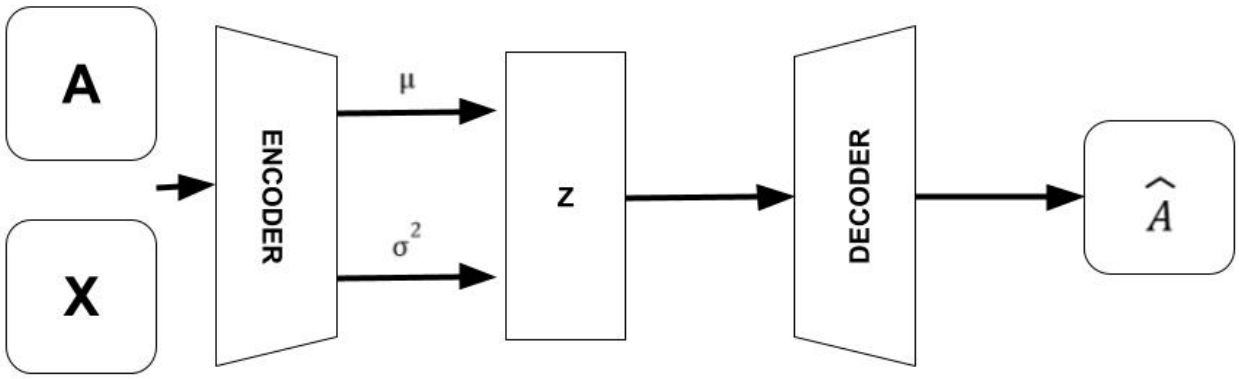

Figure 1. General structure of the VGAE method

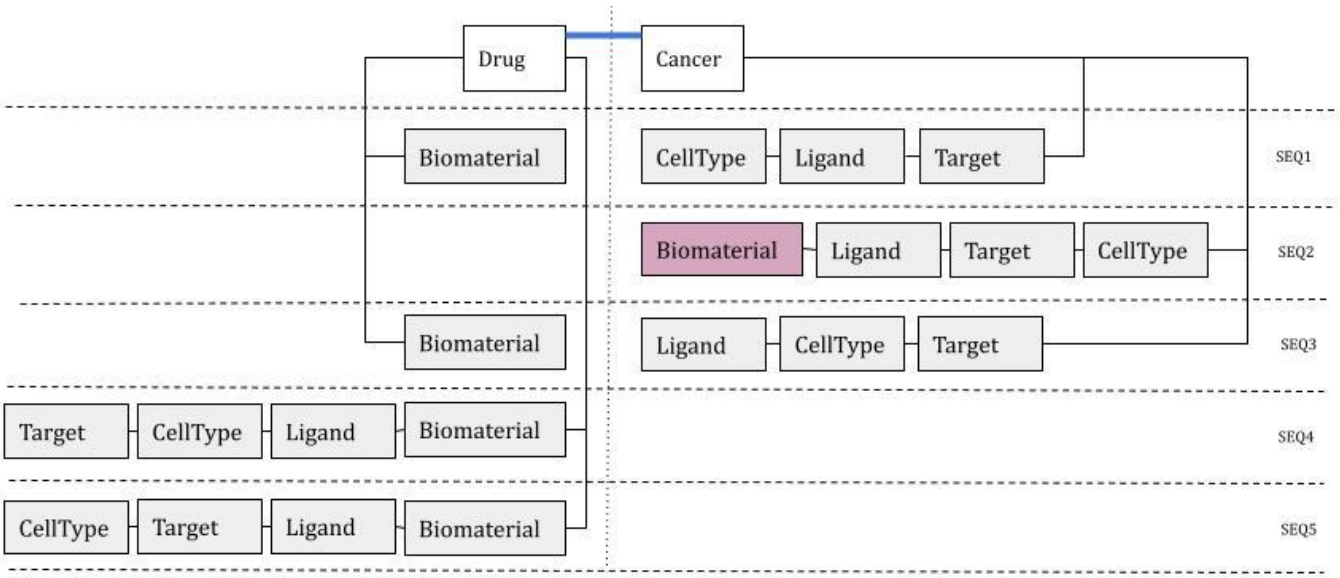

Figure 2. Example sequences of the bipartite graph structure

Traditional approaches solve the link prediction problem by assuming that similar nodes tend to have the probability of edges. These methods usually compute the similarity of nodes through heuristic node similarity scores or a labor-intensive feature extraction task. Rather than employing extensive manual feature extraction methods, the VGAE can learn latent features from the local neighborhood in a new association prediction task leveraging a graph neural network, and subsequently aggregate the pairwise node embeddings to build association representations. As denoted in Figure 2, nodes represent meta-layer classes drugs and cancers combined with 5 different sequences and meta-edges (blue bold line) represent main associations for the link prediction task on the bipartite graph.

We revise the VGAE method of the implementation in the PyTorch Geometric (PyG) library [23]. Particularly, we adapt the encoder model by separating GCN convolutional (Conv) layers to produce the mean and variance distributions. In this case, one of these Conv layers estimates the mean of the distribution, and the other one predicts the standard deviation. There are 3 Conv layers and one dropout in the VGAE class. We employ the default decoder model in the PyG library.

For the undirected graph G = (V, E) and N represents the size of nodes. As illustrated in Figure 1, A is the adjacency matrix of G, and its D degree matrix, node features as matrix X and Z matrix including stochastic latent variables zi. Finally, a two-layer GCN represents the inference.

$$q(Z|X,A) = \prod_{i=1}^{N} q(z_i|X,A) \quad (1)$$

$$q(z_i|X,A) = \eta(z_i|\mu_i, \mathrm{diag}(\sigma_i^2)) \quad (2)$$

where μ_i represents ith mean vector, diag is the diagonal covariance matrix, and a neural network is to be optimized during the prediction. Overall, the objective becomes the minimization of the form.

$$L = \varepsilon_q[\log p(A|Z) - KL[q(Z|X,A)||p(Z)] \quad (3)$$

We set the hidden size, out channels and epochs to 200, 20 and 30, respectively. The output feature size is 20, the learning rate is set to 0.1, the dropout is 0.5, and the Adam optimizer is used.

2.3. Evaluation criteria

The CSDD knowledge base transformed into a bipartite graph where edges only exist between sequences as denoted in Table 2. Hence, the CSDD knowledge base can be represented as a special case of graphs in the form of sequences. For each sequence, we keep the first class as the initial step in the bipartite graph by using the graph connectivity. Furthermore, the last class represents the target node while predicting links between these sequences. As an example, considering the sequence 1, the representation begins with the class "CellType", followed by "Ligand -> Target -> Cancer -> Drug", and the target class is "Biomaterial" for the overall link prediction task. It follows five different combinations of these representations. As shown in Figure 3, "Drug" and "Cancer" are the meta-layer class representations of the drug-disease association prediction task. For sequence 1, "Biomaterial" links to "Drug" and generates the left side of the bipartite graph as a combined sequence. On the other side, "CellType", "Ligand" and "Target" connect the meta-class "Cancer" and generate another combined sequence. We keep their relations and transform them into such a two-dimensional knowledge base embedding model perspective.

By using the RandomLinkSplit method in the PyG library, we generate train, validation, and test sets. We select 15% and 5% of edges as test and validation edges, respectively. In this situation, it hides some randomly selected edges from the model during the training phase. Since these combinations of sequences are transformed into the link prediction task, the objective is to estimate the output as an actual link or not. Hence, we apply the area under the ROC-AUC curve, and the average precision (AP) as the evaluation metrics for this study.

Table 2. Five different class layering sequences.

Sequence	Representation
SEQ1	CellType, Ligand, Target, Cancer, Drug, Biomaterial
SEQ2	Biomaterial, Ligand, Target, CellType, Cancer, Drug
SEQ3	Ligand, CellType, Target, Cancer, Drug, Biomaterial
SEQ4	Target, CellType, Ligand, Biomaterial, Drug, Cancer
SEQ5	Cancer, CellType, Target, Ligand, Biomaterial, Drug

The ROC-AUC metric represents how well the proposed model predicts whether a positive edge is a positive edge or not. The ROC-AUC metric used by the Sklearn library is closer to 1, which means the model has good positive/negative separability. The AP is the second metric implies the area under the precision-recall curve and summarizes the precision-recall curve as the weighted average of the precisions at each threshold n. The AP measures whether a model can identify all positive edges without misclassifying too many negative edges as positive.

3. Results and Dicussion

To evaluate CSDD knowledge base considering the test set, we generate 5 sequences for classes including cell types, biomaterials, ligands, targets, and cancers. Besides these sequence representations, we keep the entire knowledge base as "ALL" sequence. Since it is a bipartite graph, models learn on their own by leveraging the graph connectivity.

Regarding the ALL sequence, we generate the ROC curve for the GAE model as illustrated in Figure 3. The ROC-AUC curve is an evaluation measure for the classification of the given labels at different threshold values, especially for unbalanced data sets. The ROC is a type of probabilistic curve. On the other hand, the AUC indicates the distinctness degree. Considering the ROC curve, the x-axis and y-axis show the true positive, and the false positive rates, respectively. When the blue line is close to the upper left corner, it indicates that the model is working well.

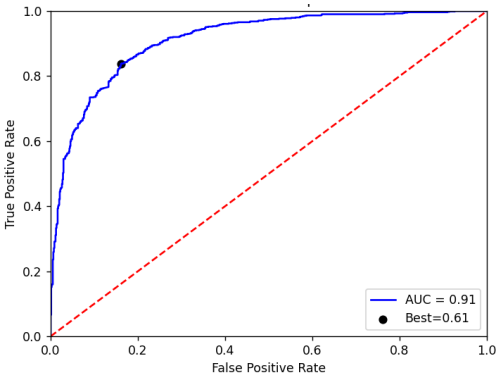

Figure 3. ROC curve for GAE model

As illustrated in Figure 4, the ROC-AUC curve indicates that the AUC score of the VGAE model performs slightly better than the GAE model regarding the sequence ALL representation.

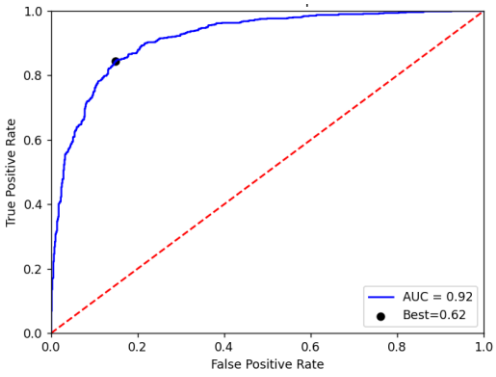

Figure 4. ROC curve for VGAE model

As denoted in Table 3, the VGAE achieves slightly better performance according to the ROC-AUC and AP scores in the sequence ALL representation. Moreover, the SEQ2 starting from the biomaterial as illustrated in Table 2, achieves the highest scores for both models.

Table 3. The experimental results for different sequences.

Sequence	GAE-ROC-AUC	*GAE*-AP	*VGAE*-ROC-AUC	*VGAE*-AP
ALL	0.9129	0.9108	0.92	0.9134
SEQ1	0.8501	0.8614	0.837	0.854
SEQ2	0.9577	0.9506	0.9627	0.9566
SEQ3	0.8018	0.8275	0.811	0.8329
SEQ4	0.7931	0.8251	0.7941	0.8268
SEQ5	0.8413	0.8609	0.837	0.858

the obtained results should be presented, and if necessary, supported with figures, tables, etc. The findings of the study should be compared with relevant literature, and the similarities and differences in the results should be interpreted to highlight the significance of the obtained results.

4. Conclusion

In this study, we propose the effect of different sequences in terms of the GAE and the VGAE models. Experimental setup shows that the VGAE model achieves remarkable results considering the ROC-AUC curve and AP score. Finally, the proposed model reflects the importance of nanoparticles in predicting drug-disease associations, which is one of the essential tasks of drug delivery systems.

Recent advances in nanoparticle research reveal that the interaction of nanoparticles with drugs and diseases has the potential to improve the pipeline of new drugs improvements especially in cancer research. Table 3 shows that the combined sequence "SEQ2" achieves the highest scores, as the only Biomaterial nodes (light magenta) are in the sequence (SEQ2) of the "Cancer" meta-layer class in Figure 2. The rest of the Biomaterial nodes are represented in "Drug" meta-layer class. Hence, it seems to be an indicator of one of the directions to improve the prediction of drug-disease association using nanoparticle interaction data, which is worthy of future investigation.

For the future work, we will enrich the dataset with existing knowledge graphs related to the FDA information as well as other compounds, and we will apply paragraph embeddings [24] and sentence BERT models [25] to employ reference sentences and descriptions as node features.

Ethics committee approval and conflict of interest statement

The authors declare no need for an ethics committee approval and no conflict of interest in this paper.

Acknowledgment

We would like to thank Sumeyra Cigdem Sozer and Cigdem Karakoyun for their detailed feedback and suggestions during the adaptation of the publicly available dataset to the proposed method.

References

[1] Liu, Y., Yang, G., Jin, S., Xu, L., & Zhao, C. X. 2020. Development of high-drug-loading nanoparticles. ChemPlusChem, 85(9), 2143-2157.

[2] Sozer, S. C., Ozmen Egesoy, T., Basol, M., Cakan-Akdogan, G., Akdogan, Y. 2020. A simple desolvation method for production of cationic albumin nanoparticles with improved drug loading and cell uptake. Journal of Drug Delivery Science and Technology. Volume 60, 101931, ISSN 1773-2247. https://doi.org/10.1016/j.jddst.2020.101931.

[3] Akdogan, Y., Sozer, S. C., Akyol, C., Basol, M., Karakoyun, C., Cakan-Akdogan, G. 2022. Synthesis of albumin nanoparticles in a water-miscible ionic liquid system, and their applications for chlorambucil delivery to cancer cells. Journal of Molecular Liquids. Volume 367, Part B, 120575, ISSN0167-7322. https://doi.org/10.1016/j.molliq.2022.120575.

[4] Rubin, D. L., Lewis, S. E., Mungall, C. J., Misra, S., Westerfield, M., Ashburner, M., ... & Musen, M. A. 2006. National center for biomedical ontology: advancing biomedicine through structured organization of scientific knowledge. Omics: a journal of integrative biology, 10(2), 185-198.

[5] Lever, J., Zhao, E. Y., Grewal, J., Jones, M. R., & Jones, S. J. 2019. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nature methods, 16(6), 505-507.

[6] Launer-Wachs, S., Taub-Tabib, H., Goldberg, Y., & Shamay, Y. 2022. Rapid Knowledgebase Construction and Hypotheses Generation Using Extractive Literature Search. bioRxiv, 2022-02.

[7] Gottlieb, A., Stein, G. Y., Ruppin, E., & Sharan, R. 2011. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Molecular systems biology, 7(1), 496.

[8] Luo, H., Li, M., Wang, S., Liu, Q., Li, Y., & Wang, J. 2018. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. Bioinformatics, 34(11), 1904-1912.

[9] Zhang, W., Yue, X., Lin, W., Wu, W., Liu, R., Huang, F., & Liu, F. 2018. Predicting drug-disease associations by using similarity constrained matrix factorization. BMC bioinformatics, 19, 1-12.

[10] Wang, M. N., You, Z. H., Li, L. P., Chen, Z. H., & Xie, X. J. 2020. WGMFDDA: A novel weighted-based graph regularized matrix factorization for predicting drug-disease associations. In Intelligent Computing Methodologies: 16th International Conference, ICIC 2020, Bari, Italy, October 2–5, 2020, Proceedings, Part III 16. Springer International Publishing, 542-551.

[11] Liang, X., Zhang, P., Yan, L., Fu, Y., Peng, F., Qu, L., ... & Chen, Z. 2017. LRSSL: predict and interpret drug–disease associations based on data integration using sparse subspace learning. Bioinformatics, *33*(8), 1187-1196.

[12] Zhang, W., Yue, X., Chen, Y., Lin, W., Li, B., Liu, F., & Li, X. 2017. Predicting drug-disease associations based on the known association bipartite network. In 2017 IEEE international conference on bioinformatics and biomedicine (BIBM) IEEE, 503-509.

[13] Wang, B., Lyu, X., Qu, J., Sun, H., Pan, Z., & Tang, Z. 2019. GNDD: a graph neural network-based method for drug-disease association prediction. In 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) IEEE, 1253-1255.

[14] Yu, Z., Huang, F., Zhao, X., Xiao, W., & Zhang, W. 2021. Predicting drug–disease associations through layer attention graph convolutional network, Briefings in Bioinformatics, 22(4), bbaa243.

[15] Kipf, T. N., & Welling, M. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

[16] Taub-Tabib, H., Shlain, M., Sadde, S., Lahav, D., Eyal, M., Cohen, Y., & Goldberg, Y. 2020. Interactive Extractive Search over Biomedical Corpora. In Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, 28-37.

[17] Fadilah, N. I. M., Isa, I. L. M., Zaman, W. S. W. K., Tabata, Y., & Fauzi, M. B. 2022. The effect of nanoparticle-incorporated natural-based

biomaterials towards cells on activated pathways: a systematic review. Polymers, 14(3), 476.

[18] Joshy, K. S., Susan, M. A., Snigdha, S., Nandakumar, K., Laly, A. P., & Sabu, T. 2018. Encapsulation of zidovudine in PF-68 coated alginate conjugate nanoparticles for anti-HIV drug delivery. International journal of biological macromolecules, 107, 929-937.

[19] Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., ... & Woolsey, J. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic acids research, 34(suppl_1), D668-D672.

[20] Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., ... & Ponten, F. 2017. A pathology atlas of the human cancer transcriptome. Science, 357(6352), eaan2507.

[21] Canese, K., & Weis, S. 2013. PubMed: the bibliographic database. The NCBI handbook, 2(1).

[22] Kipf, T.N., Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. 5th International Conference on Learning Representations, ICLR. Toulon, France, April 24-26, Conference Track Proceedings.

[23] Fey, M., & Lenssen, J. E. 2019. Fast graph representation learning with PyTorch Geometric. arXiv preprint arXiv:1903.02428.

[24] Le, Q., & Mikolov, T. 2014. Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR.

[25] Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982-3992.