Предикција штеште интеракције лекова помоћу графовске неуронске мреже

Марко Његомир
Факултет техничких наука
Универзитет у Новом Саду
Нови Сад, Србија
marko.njegomir@uns.ac.rs

***Апстракт*— Коришћење више лекова у терапији може довести до штетних дејстава која су резултат интеракције лекова. Циљ овог рада је да се коришћењем графовске неуронске мреже уграђене у енкодер трансформер архитектуре изврши предикција штетне интеракције лекова. На овај начин се могу одабрати најбољи кандидати за даља клиничка тестирања. Слој графовске неуронске мреже који се користи у овом раду је PNAConv [37] који користи више агрегационих функција са циљем да се боље експлоатише графовска структура молекула лекова. Скуп података који је коришћен је преузет из рада [8] и добијен је филтрирањем TWOSIDES side-effects скупа података који садржи 964 типа ефеката настала штетном интеракцијом лекова. Решење се евалуира коришћењем "AUROC" (Area Under the Receiver Operating Characteristic curve) метрике. Она носи информацију колико модел добро дискриминише између позитивних примера и негативних примера.**

Кључне речи—Графовска неуронска мрежа; трансформер архитектура; полифармација; штетна интеракција лекова;

I. Увод

У савременој медицини, лекови играју важну улогу у лечењу болести. Када је пацијент захваћен комплексном болешћу која утиче на више система у организму или када пати од више различитих болести, неопходно је прописати комбинацију више лекова [14]. Посебни ефекти сваког лека већином су добро испитани уз ригорозне контроле и правила која морају бити испоштована да би се лек пустио у употребу. Међутим, тестирање комбинација лекова представља проблем због неопходности тестирања сваког новог лека са свим постојећим лековима да би се утврдиле могуће штетне интеракције [11]. Ова зависност је асимптотски квадратна, што значи да би број парова лекова које треба тестирати експоненцијално порастио са бројем лекова у употреби. Овакво тестирање при прављењу новог лека је неизводљиво. Додатно, велика количина нових лекова се ставља у промет, док се мања количина лекова повлачи из употребе, што укупно повећава број доступних лекова [11]. Такође, често постоје значајне разлике у лековима који су доступни за лечење болести у различитим државама, што значи да није довољно тестирати само интеракцију лекова који су доступни у једној држави [12].

При изради новог лека, жељено је пронаћи већ постојеће лекове који највероватније имају негативне интеракције у комбинацији са новим леком. На тај начин, у фази клиничких испитивања новог лека, могуће је највеће потенцијалне штетне интеракције тестирају. Додатно, у случају откривања нових штетних интеракција међу лековима, било би корисно пронаћи већ постојеће лекове који би највероватније испољили такве негативне интеракције, и користити их као кандидате за тестирање.

Информације о штетним интеракцијама лекова могу помоћи у смањењу броја људи који су изложени њима. У Сједињеним Америчким Државама, 15% становништва пати од последица штетних интеракција лекова [1]. Ове интеракције могу имати озбиљне последице, као што показују подаци да 2-5% старих људи заврши у болници као резултат штетних интеракција лекова [2][3][4]. Хоспитализације изазване штетним интеракцијама лекова у генералној популацији износе 1% [5], али треба имати на уму да старење популације може довести до раста ових бројева. Годишњи трошак лечења ових штетних интеракција износи 177 милијарди долара [6].

У овом раду је циљ направити предикцију штетних интеракција лекова помоћу модела који би требало да научи да предвиди да ли два улазна лека имају потенцијално штетну интеракцију. Приликом развијања нових лекова, модел би могао да се користи за упаривање новог лека са свим осталим лековима и предвиђање штетних интеракција. Овај поступак би помогао да се изаберу најбољи кандидати за клиничка тестирања штетних интеракција.

У циљу предвиђања штетних интеракција лекова потребно је направити модел који може да распозна парове лекова са и без штетних интеракција. Графовска структура атома молекула користи се за повезивање парова лекова и представља класификациони чвор. Модел користи PNAConv слој унутар енкодера Трансформер архитектуре да би научио да врши предикције о штетним интеракцијама. У циљу бољих предикција, сваки чвор једног лека повезује се са свим чворовима другог лека, што омогућава моделу да узме у обзир информације о свим атомима оба лека. Сама предикција се врши на основу вредности чвора који повезује графове атома два лека, а која се добија након пропуштања података кроз модел.

У овом истраживању коришћен је скуп података који је филтриран из TWOSIDES скупа података [16], а који укључује 964 типа штетних ефеката насталих интеракцијом лекова. Укључено је укупно 4 576 785 парова лекова са штетном интеракцијом, али је касније скуп података допуњен са паровима лекова који немају штетне интеракције, те је коначан скуп података садржао 9 153 570 парова лекова. Како би се евалуирала ефикасност модела, користи се метрика “AUROC” (Area Under the Receiver Operating Characteristic curve), која приказује колико добро модел раздваја позитивне и негативне примере.

II. Рад са подацима

Модул има за циљ претварање података о лековима из TWOSIDES side-effect скупа података у одговарајући облик и учитавање их у граф. Ови подаци се састоје од идентификатора лекова, имена штетних ефеката који настају њиховом интеракцијом и њиховог облика. Како би се ови подаци учитали у граф, потребно је да се идентификатори лекова претворе у информације о структури молекула, што се постиже коришћењем PubChem базе података. Циљ је да се ови подаци претворе у облик који је приходан моделу и да се омогући моделу да ради са њима, а евалуација модела се обавља коришћењем "AUROC" метрике.

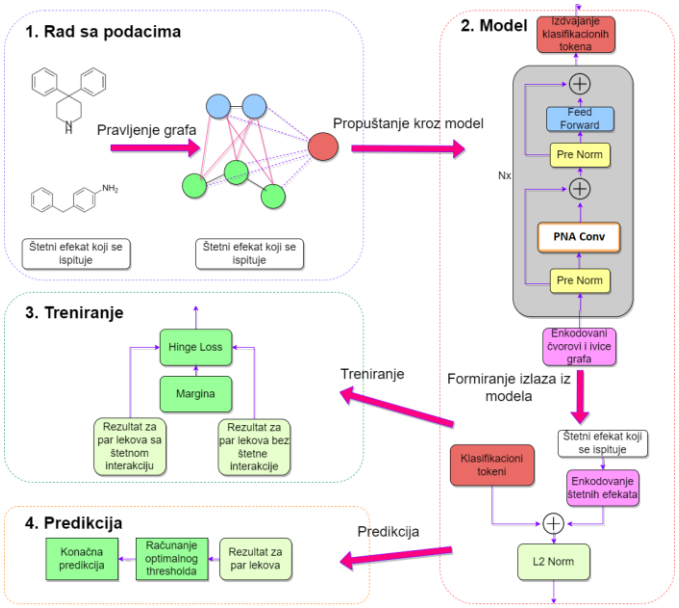
Како би се извршила предвиђања интеракција лекова, потребно је да се подаци о лековима претпроцесирају и конвертују у формат који је погодан за рад са моделом. У том смислу, подаци о лековима извучени су из TWOSIDES скупа података [31] и добијени су у формату идентификатора лекова, штетних ефеката која произлазе из њихове интеракције, типова веза и SMILES стрингова. За претварање ових података у граф чији чворови представљају атоме молекула, а везе између чворова су хемијске везе између атома, коришћен је Data loader [8]. Овај алат обрађује податке и ствара графове атома за сваки појединачан лек из пара лекова. Уз сваки batch, Data loader враћа обележја свих атома у паровима лекова, типове веза, маске за сваки молекул у batch-у, штетни ефекат интеракције лекова који се проверава за сваки пар лекова, као и информације о међусобним везама два молекула.

У овом истраживању, формиран је један граф за сваки пар молекула користећи податке о структури молекула лекова. За сваки граф, листа суседства је формирана како би се добио неусмерен граф молекула лека. Сви атрибути свих атома и ивица парова лекова су садржани у једној матрици атома и једној матрици атрибута ивица. Класификациони чворови су додати за сваки пар молекула и повезани са свим атомима одговарајућег пара молекула. У ову сврху, нови тип веза је уведен у граф да би се разликовале хемијске везе и виртуалне везе у молекулама лекова. Целокупан процес је илустрован на Илустрацији 1.

У експерименту су такође додати и везе између свих атома једног лека са свим атомима другог лека у пару лекова. За ове везе је такође додат нови тип, и због тога модел може да прави разлику између правих веза и виртуелних веза које постоје у графу два лека.

III. Модел

У овом раду се користи графовска трансформер архитектура, која се састоји од повезаног графа атома два лека. Излазни резултат модела, који користи PNAConv слој за механизам пажње, је предикција о потенцијалној штетној интеракцији пара лекова. PyTorch Geometric [20] библиотека се користи за имплементацију слојева граф неуронских мрежа, док се обична PyTorch библиотека [33] користи за остале слојеве. Пре него што се чворови и ивице проследе енкодеру модела, њихова обележја се ембедују. Након тога, ембединг матрице чворова и ивица, као и листа суседства чворова (Илустрација 1), се шаљу енкодеру.



Илустрација 1 Дијаграм тока података са детаљима архитектуре. (1) Рад са подацима укључује прављење графа од атома два лека. (2) Модел се састоји од енкодера трансформера са PNA слојем. (37) У тренирању се користи Hinge loss. (4) Предикција се врши примењивањем оптималног threshold-а за излазе из модела.

PNAConv слој у енкодеру има способност да искористи више агрегационих функција и да на тај начин боље експлоатише графовску структуру молекула.

Након пропуштања кроз PNAConv слој, подаци стижу до потпуно повезаног слоја, којег чине комбинација линеарних, GELU, и dropout слојева (Илустрација 1).

Следећи слојеви након Layer Normalization слојева у моделу су PNAConv и Fully Connected слојеви. Излаз из ових слојева се спаја са излазом из Layer Normalization слојева и тако се реализују skip конекције (Илустрација 1).

Након проласка кроз енкодер архитектуру, излазни чворови се користе за добијање класификационих токена. За пар лекова, прави се embedding који се додаје на одговарајући embedding класификационог токена. L2

норма се рачуна на добијени резултат и то је коначни излазни резултат модела (Илустрација 1).

IV. Тренирање

Модел је дизајниран тако да извршава предикцију штетних интеракција лекова. За то је неопходно проследити парове лекова моделу. Парови лекова могу бити позитивни (са штетним интеракцијама) или негативни (без штетних интеракција). Модел рачуна резултате за оба типа парова лекова, које затим прослеђује Max Margin (Hinge) функцији са маргином 1. Максимизација маргине се користи због тога што је погодна за contrastive learning приступ који се овде користи. Имплементација Max Margin функције се разликује од изворне формуле, где се max функција замењује са ReLU функцијом. За тренирање модела, користи се ADAM optimizer и backpropagation алгоритам за пропагацију градијената.

Предвиђање резултата модела врши се упоређивањем свих резултата са threshold вредношћу. Резултати који имају вредност већу од threshold-а се класификују као позитивни, док су резултати испод threshold-а негативни (Илустрација 1). Threshold се динамички одређује тако да буде оптималан за AUROC метрику. Након тога, предикције се поређују са тачним лабелама и на основу тог поређења се рачунају метрике за евалуацију.

V. Експеримент

Хардвер на коме су покретани експерименти је Lenovo Legion Y-540 laptop, и има следеће компоненте:

- I7 9750 6-core processor @4.5 GHz
- 16GB ram
- GTX 1660 TI 6GB graphic card
- M2 SSD 512 GB

Хиперпараметри модела:

- Dim 32
- Depth 2
- Learning rate 1e-3
- Epoch 2
- Batch size 100
- Dropout 0.1 за embedding слојеве
- Dropout 0 за слојеве у енкодеру

VI. Евалуација

Евалуација модела се врши користећи десетоструку унакрсну валидацију (стратификовану), где се тест скуп користи за проверу перформанси модела. Међутим, појединачни молекули лекова могу се појавити у оба сета (тренинг и тест), али исти пар лекова није дозвољено да буде у оба сета. Ово је вид трансдуктивног учења, исто као и у раду [8], што омогућава директно поређење резултата.

VII. Резултати

Поређење са другим моделима је урађено на основу AUROC метрике (Табела 1). Резултати модела су преузети из рада [8].

Назив модела	***AUROC***
Drug-fingerprints	0.744
RESCAL	0.693
DEDICOM	0.705
DeepWalk	0.761
Concatenated features	0.793
Експеримент PNA Conv (2 епохе)	0.829
Eksperiment 2 **GatV2 Conv** (**дипломски 30 епоха**)	0.866
Decagon	0.872
MHCADDI	**0.882**

VIII. Дискусија

Резултати после само две епохе су бољи од већине модела, и то указује да би наставак овог експеримента могао имати обећавајуће резултате.

Осим PNAConv слоја, корисно би било испробати и E2(n) еквиваријантни слој. Овај слој би се могао тестирати и на новом скупу података који би садржао и позиционе информације као што је на пример QM9 скуп података.

Референце

[1] Kantor, E.D., Rehm, C.D., Haas, J.S., Chan, A.T. and Giovannucci, E.L., 2015. Trends in prescription drug use among adults in the United States from 1999-2012. Jama, 314(17), pp.1818-1830.

[2] Bénard-Laribière, A., Miremont-Salamé, G., Pérault-Pochat, M.C., Noize, P. and Haramburu, F., 2015. EMIR Study Group on behalf of the French network of pharmacovigilance centres. Incidence of hospital admissions due to adverse drug reactions in France: the EMIR study. Fundam Clin Pharmacol, 29(1), pp.106-111.

[3] Becker, M.L., Kallewaard, M., Caspers, P.W., Visser, L.E., Leufkens, H.G. and Stricker, B.H., 2007. Hospitalisations and emergency department visits due to drug–drug interactions: a literature review. Pharmacoepidemiology and drug safety, 16(6), pp.641-651.

[4] Olivier, P., Bertrand, L., Tubery, M., Lauque, D., Montastruc, J.L. and Lapeyre-Mestre, M., 2009. Hospitalizations because of adverse drug reactions in elderly patients admitted through the emergency department. Drugs & aging, 26(6), pp.475-482.

[5] Dechanont, S., Maphanta, S., Butthum, B. and Kongkaew, C., 2014. Hospital admissions/visits associated with drug–drug interactions: a systematic review and meta-analysis. Pharmacoepidemiology and drug safety, 23(5), pp.489-497.

[6] Ernst, F.R. and Grizzle, A.J., 2001. Drug-related morbidity and mortality: updating the cost-of-illness model. Journal of the American Pharmaceutical Association (1996), 41(2), pp.192-199.

[7] Zitnik, M., Agrawal, M. and Leskovec, J., 2018. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(13), pp.i457-i466.

[8] Deac, A., Huang, Y.H., Veličković, P., Liò, P. and Tang, J., 2019. Drug-drug adverse effect prediction with graph co-attention. arXiv preprint arXiv:1905.00534.

[9] Feng, Y.H. and Zhang, S.W., 2022. Prediction of Drug-Drug Interaction Using an Attention-Based Graph Neural Network on Drug Molecular Graphs. Molecules, 27(9), p.3004.

[10] Bai, Y., Gu, K., Sun, Y. and Wang, W., 2020. Bi-level graph neural networks for drug-drug interaction prediction. arXiv preprint arXiv:2006.14002.

[11] Austin, D.H., 2006. Research and development in the pharmaceutical industry. Congress of the United States, Congressional Budget Office.

[12] O'Neill, P. and Sussex, J., 2015. International Comparison of Medicines Usage: Quantitative Analysis from a Swedish Perspective (No. 001611). Office of Health Economics.

[13] Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M. and Leskovec, J., 2020. Open graph benchmark: Datasets for machine learning on graphs. Advances in neural information processing systems, 33, pp.22118-22133.

[14] Bansal, M., Yang, J., Karan, C., Menden, M.P., Costello, J.C., Tang, H., Xiao, G., Li, Y., Allen, J., Zhong, R. and Chen, B., 2014. A community computational challenge to predict the activity of pairs of compounds. Nature biotechnology, 32(12), pp.1213-1222.

[15] Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P. and Bengio, Y., 2017. Graph attention networks. stat, 1050, p.20.

[16] Takeda, T., Hao, M., Cheng, T., Bryant, S.H. and Wang, Y., 2017. Predicting drug–drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge. Journal of cheminformatics, 9(1), pp.1-9.

[17] Rozemberczki, B., Hoyt, C.T., Gogleva, A., Grabowski, P., Karis, K., Lamov, A., Nikolov, A., Nilsson, S., Ughetto, M., Wang, Y. and Derr, T., 2022. ChemicalX: A Deep Learning Library for Drug Pair Scoring. arXiv preprint arXiv:2202.05240.

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

[19] Bronstein, M.M., Bruna, J., Cohen, T. and Veličković, P., 2021. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478.

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

[21] Kipf, T.N. and Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

[22] Brody, S., Alon, U. and Yahav, E., 2021. How attentive are graph attention networks?. arXiv preprint arXiv:2105.14491.

[23] Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R. and Gulcehre, C., 2018. Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261.

[24] Hooker, S., 2021. The hardware lottery. Communications of the ACM, 64(12), pp.58-65.

[25] Hamilton, W.L., 2020. Graph representation learning. Synthesis Lectures on Artifical Intelligence and Machine Learning, 14(3), pp.1-159.

[26] Milgram, S., 1967. The small world problem. Psychology today, 2(1), pp.60-67.

[27] Chen, D., Lin, Y., Li, W., Li, P., Zhou, J. and Sun, X., 2020, April. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 04, pp. 3438-3445).

[28] Orhan, A.E. and Pitkow, X., 2017. Skip connections eliminate singularities. arXiv preprint arXiv:1701.09175.

[29] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[30] Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[31] Tatonetti, N.P., Ye, P.P., Daneshjou, R. and Altman, R.B., 2012. Data-driven prediction of drug effects and interactions. Science translational medicine, 4(125), pp.125ra31-125ra31.

[32] Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A. and Wang, J., 2016. PubChem substance and compound databases. Nucleic acids research, 44(D1), pp.D1202-D1213.

[33] Paszke, A. et al., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035. Available at: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[34] Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[35] Biewald, L., 2020. Experiment tracking with weights and biases. Software available from wandb. com, 2, p.233.

[36] Brandstetter, J., Hesselink, R., van der Pol, E., Bekkers, E. and Welling, M., 2021. Geometric and physical quantities improve e (3) equivariant message passing. arXiv preprint arXiv:2110.02905.

[37] Corso, G., Cavalleri, L., Beaini, D., Liò, P. and Veličković, P., 2020. Principal neighbourhood aggregation for graph nets. Advances in Neural Information Processing Systems, 33, pp.13260-13271.