Tutorial: Data Mining Methods for Drug Discovery and Development

Cao Xiao
Analytics Center of Excellence, IQVIA
Cambridge, MA, USA

Jimeng Sun
Georgia Institute of Technology
Atlanta, GA, USA

ABSTRACT

In silico modeling of medicine refers to the direct use of computational methods in support of drug discovery and development. Machine learning and data mining methods have become an integral part of in silico modeling and demonstrated promising performance at various phases of the drug discovery and development process. In this tutorial we will introduce data analytic methods in drug discovery and development. For the first half, we will provide an overview about related data and analytic tasks, and then present the enabling data analytic methods for these tasks. For the second half, we will describe concrete applications of each of those tasks. The tutorial will be concluded with open problems and a Q&A session.

CCS CONCEPTS

• **Computing methodologies → Neural networks**; **Learning latent representations**; • **Applied computing → Health informatics**.

KEYWORDS

Data Mining, Drug Discovery, Drug Safety

ACM Reference Format:
Cao Xiao and Jimeng Sun. 2019. Tutorial: Data Mining Methods for Drug Discovery and Development. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'19).* ACM, NY, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332273

1 TARGET AUDIENCE AND PREREQUISITE

The target audiences are graduate students, researchers, scientists and practitioners in both academia and industry who are interested in how data mining and machine learning can be applied in drug discovery and development. The audience should be familiar with basic data mining and machine learning concepts, and ideally some basic experiences in deep learning.

2 TUTORS

Cao Xiao is the Director of Machine Learning at Analytics Center of Excellence of IQVIA. She is leading IQVIA's North America machine learning team to drive next generation healthcare AI. Her team works on various projects on disease modeling and in silico drug modeling (e.g., adverse drug reaction detection, drug repositioning and de novo design). Her research focuses on using machine learning and data mining approaches to solve diverse real world healthcare challenges. Particularly, she is interested in phenotyping on electronic health records, data mining for in-silico drug modeling, biomarker discovery and patient segmentation for neurodegenerative diseases. Her research has been published in leading AI conferences including KDD, NIPS, ICLR, AAAI, IJCAI, SDM, ICDM, WWW and top health informatics journals such as Nature Scientific Reports and JAMIA. Prior to IQVIA, she was a research staff member in the AI for Healthcare team at IBM Research from 2017 to 2019 and served as member of the IBM Global Technology Outlook Committee from 2018 to 2019. She acquired her Ph.D. degree from University of Washington, Seattle in 2016.

Jimeng Sun (corresponding tutor) is an associate professor of College of Computing at Georgia Tech. Prior to Georgia Tech, he was a research staff member at IBM TJ Watson Research Center. His research focuses on data mining for healthcare, especially in developing tensor factorization, deep learning methods, and large-scale predictive modeling systems. He published over 120 papers with h-index 54 and filed over 20 patents. He has received SDM/IBM early career research award 2017, ICDM best research paper award in 2008, SDM best research paper award in 2007, and KDD Dissertation runner-up award in 2008. In 2019, he was recognized as Top 100 AI Leaders in Drug Discovery and Advanced Healthcare. Dr. Sun received B.S. and M.Phil. in Computer Science from Hong Kong University of Science and Technology in 2002 and 2003, M.Sc and PhD in Computer Science from Carnegie Mellon University in 2006 and 2007.

3 TUTORIAL OUTLINE

Motivation: Data mining for drug discovery and development is a fast growing area which combines massive multi-modal biomedical data with data mining/machine learning technologies for identifying new drugs or new indications of existing drugs [4]. Among others, the in silico modeling that builds computational models to understand biological behaviors of drugs, attract the most interests, and achieved promising results in alleviating the issues of cost, length and risks faced with traditional methods for drug discovery and development [4]. As it becomes an emerging area of data mining, we want to provide a comprehensive overview of its latest developments and open challenges to the data mining community. Thus, in this tutorial we will introduce state-of-the-art data mining models for solving drug discovery and development tasks.

The tutorial will be given in the following order:

KDD'19, August 2019, AK

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.1145/3292500.3332273

3.1 Overview (20 minutes)

Using those computational methods to facilitate drug development applications is a fast growing area in research.
The relevant tasks for in silico modeling for drug discovery and development include the following:

(a) **Molecule property prediction** aims at identifying molecules' therapeutic effects and toxicity given the molecule structure data. This is achieved by predicting the bioactivity of molecules such as potency and binding affinity to a target.
(b) **Molecule generation for *de novo* design** (i.e., designing an entirely new molecule from scratch) aims at using generative models to produce chemically correct structures thus for alleviating the experimental discovery effort.
(c) **Drug repositioning** aims at leveraging drug similarity to find novel indications for existing drugs for shortening of drug development cycle.
(d) **Adverse drug reaction (ADR) or drug-drug interaction (DDI) detection** is to leverage large amounts of biomedical data to inform more targeted clinical safety tests, which is critical to the success of drug development.

The available data to support the aforementioned tasks

3.2 Methods (55 minutes)

We will discuss foundational methods for tackling these tasks.

(1) **Generative models.** We will focus on variational autoencoders (VAE) and generative adversarial networks (GAN) that are well suited for *molecule generation for de novo design.*
(2) **Deep representation learning.** We will present major deep neural networks for general *biomedical data encoding*.
(3) **Reinforcement learning.** We will mainly talk about policy gradient (PD) methods that is often used in sequential *molecule generation for de novo design* to incorporates domain-specific rules and improve the effectiveness of generation.
(4) **Graph Embedding** especially graph convolutional networks (GCN) [10] learns improved embeddings of medical concepts related to drugs, such as drugs, proteins.

3.3 Applications (90 minutes)

During the second half of the tutorial, we will show case recent works for each of these tasks.

(a) Molecule property prediction for drug discovery including AtomNet [16], Coley et al. [5], and Gomes et al. [7].
(b) Molecule generation for *de novo* design refers to generate structurally correct new molecules. Variety of deep generative models are proposed for this task, including geneated based on line representation of drugs [6, 9, 11], and graph representation [2, 12], or using RL [18]
(c) Drug repositioning refers to drugs developed for a particular indication can be efficacious in other indications that share the same pathophysiology [1, 3, 14].
(d) Adverse drug reaction (ADR) or drug-drug interaction (DDI) prediction including [8, 13, 15, 17]

3.4 Future directions and Q&A (15 minutes)

REFERENCES

[1] Alexander Aliper, Sergey Plis, Artem Artemov, Alvaro Ulloa, Polina Mamoshina, and Alex Zhavoronkov. 2016. Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Molecular Pharmaceutics* 13, 7 (2016), 2524–2530. https://doi.org/10.1021/acs.molpharmaceut.6b00248 arXiv:https://doi.org/10.1021/acs.molpharmaceut.6b00248 PMID: 27200455.

[2] Han Altae-Tran, Bharath Ramsundar, Aneesh S. Pappu, and Vijay Pande. 2017. Low Data Drug Discovery with One-Shot Learning. *ACS Central Science* 3, 4 (2017), 283–293. https://doi.org/10.1021/acscentsci.6b00367 arXiv:https://doi.org/10.1021/acscentsci.6b00367 PMID: 28470045.

[3] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. 2008. Drug Target Identification Using Side-Effect Similarity. *Science* 321, 5886 (2008), 263–266. https://doi.org/10.1126/science.1158140

[4] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. 2018. The rise of deep learning in drug discovery. *Drug Discovery Today* 23, 6 (2018), 1241 – 1250. https://doi.org/10.1016/j.drudis.2018.01.039

[5] Connor W. Coley, Regina Barzilay, William H. Green, Tommi S. Jaakkola, and Klavs F. Jensen. 2017. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *Journal of Chemical Information and Modeling* 57, 8 (2017), 1757–1772. https://doi.org/10.1021/acs.jcim.6b00601 arXiv:https://doi.org/10.1021/acs.jcim.6b00601 PMID: 28696688.

[6] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 2224–2232. http://papers.nips.cc/paper/5954-convolutional-networks-on-graphs-for-learning-molecular-fingerprints.pdf

[7] Joseph Gomes, Bharath Ramsundar, Evan N. Feinberg, and Vijay S. Pande. 2017. Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity. *CoRR* abs/1703.10603 (2017). arXiv:1703.10603 http://arxiv.org/abs/1703.10603

[8] Bo Jin, Haoyu Yang, Cao Xiao, Ping Zhang, Xiaopeng Wei, and Fei Wang. 2017. Multitask Dyadic Prediction and Its Application in Prediction of Adverse Drug-Drug Interaction. (2017).

[9] Artur Kadurin, Sergey Nikolenko, Kuzma Khrabrov, Alex Aliper, and Alex Zhavoronkov. 2017. druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Molecular Pharmaceutics* 14, 9 (2017), 3098–3104. https://doi.org/10.1021/acs.molpharmaceut.7b00346 arXiv:https://doi.org/10.1021/acs.molpharmaceut.7b00346 PMID: 28703000.

[10] Thomas N. Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *CoRR* abs/1609.02907 (2016). arXiv:1609.02907 http://arxiv.org/abs/1609.02907

[11] Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. 2017. Grammar Variational Autoencoder. In *ICML*.

[12] Tengfei Ma, Jie Chen, and Cao Xiao. 2018. Constrained Generation of Semantically Valid Graphs via Regularizing Variational Autoencoders. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). 7113–7124.

[13] Tengfei Ma, Cao Xiao, Jiayu Zhou, and Fei Wang. 2018. Drug Similarity Integration Through Attentive Multi-view Graph Auto-encoders. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18).* AAAI Press, 3477–3483. http://dl.acm.org/citation.cfm?id=3304222.3304251

[14] Ioakeim Perros, Fei Wang, Ping Zhang, Peter Walker, Richard Vuduc, Jyotishman Pathak, and Jimeng Sun. [n. d.]. *Polyadic Regression and its Application to Chemogenomics.* 72–80. https://doi.org/10.1137/1.9781611974973.9 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611974973.9

[15] Aleksandar Poleksic and Lei Xie. 2018. Predicting serious rare adverse reactions of novel chemicals. *Bioinformatics* 34, 16 (03 2018), 2835–2842. https://doi.org/10.1093/bioinformatics/bty193

[16] Izhar Wallach, Michael Dzamba, and Abraham Heifets. 2015. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. *CoRR* abs/1510.02855 (2015). arXiv:1510.02855 http://arxiv.org/abs/1510.02855

[17] Cao Xiao, Ping Zhang, W. Chaovalitwongse, Jianying Hu, and Fei Wang. 2017. Adverse Drug Reaction Prediction with Symbolic Latent Dirichlet Allocation. (2017).

[18] Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. 2018. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 6410–6421.