Research Statement

Chengxi Zang
Department of Population Health Sciences
Weill Cornell Medicine, Cornell University
September, 2020

My long-term career goal is to become a global leader in health AI/data science with a vision of empowering healthcare, drug discovery & development by deep analytics and the insights mined from widely existed structured and dynamic data (e.g. molecules, electronic health records, social media, etc.). The need for effective management and mining of structured and dynamic data, and integrating data-driven insights with domain knowledge is critical to achieving my career goal, and my objective is to develop and apply a suite of data mining and machine learning approaches to address critical challenges in information integration, knowledge acquisition and deep analytics for the healthcare practice, drug discovery & development.

I got my Ph.D. degree with honor (top 3) from Computer Science Department, Tsinghua University, in January 2019. During my Ph.D. period, I visited the Center for Complex Network Research (CCNR), Northeastern University, Boston, as a visiting scholar from July 2017 to May 2018. I did my postdoc training at the Weill Cornell Medicine, Cornell University, since February 2019. During my postdoc period, I also got a postdoc fellowship in Boehringer Ingelheim Pharmaceuticals, Inc., since March 2020. My interdisciplinary experience has equipped me with the unique and necessary skills to achieve my career goal. I highlight my achievements below.

- *Publications*: 12 peer-reviewed conference papers (9 as the first or corresponding author), 4 peer-reviewed conference workshop papers (2 as the first author), and 2 peer-reviewed journal articles (1 as the first author), (See my curriculum vitae here http://www.calvinzang.com/file/Chengxi_Zang-CV.pdf).
- *Awards:* 1) One Best paper candidate (ICDM 2018), and 2) One Best paper award (DLGMA 2020) (Acronyms: ICDM stands for IEEE International Conference on Data Mining, DLGMA stands for AAAI 2020 Deep Learning on Graphs: Methodologies and Applications Workshop. ICDM, AAAI are the top conferences in data mining and machine learning. Deep Learning on Graphs workshop is one of the largest workshops held in AAAI and KDD. Peer-reviewed, in computer science, papers in top conferences have higher quality than most journals. Some reasonable public CS conference ranking lists: Ref1, Ref2, Ref3).

1. Research Overview

My research lies in the intersection between data science, machine learning and healthcare. My research theme is to develop *interpretable, robust, fair,* and *intelligent* approaches that can effectively mine insights from *big structured* and *dynamic data* which widely exist in *healthcare systems*, *drug discovery & development, social media,* and then validate and incorporate such insights in *healthcare practice*, *drug discovery & development*, and *social science*. In the following, I will summarize my research from *algorithm*, *application,* and *system* aspects.

Algorithm. I have been working on developing novel computational algorithms for analyzing various kinds of (structured and dynamics) healthcare data, including Molecules, Electronic Health Records (EHRs), Pharmaceutical Research and Development data. 1) One of my major research focus is AI for drug discovery, i.e., designing novel and optimized molecular graphs driven by big chemical data and deep graph generative models. Designing novel and optimized molecular graphs is the major goal of the drug discovery process (including lead discovery and lead optimization steps) which takes up to 5 years and costs hundreds of millions of US dollars. To generate novel and optimized molecular graphs is a very difficult problem, because the drug-like chemical space is too large ($\sim 10^{60}$) to search, to optimize discrete molecular graphs is a hard combinatorial problem, and further to evaluate them are very expensive. My early research on drug discovery is an invertible deep graph generative framework (MoFlow) that achieves many state-of-the-art results in molecular graph generation and optimization. The related research has also resulted in KDD 2020 [1]. 2) My second focus is patient screening driven by real-world clinical data, i.e., to find targeted patient cohorts by real-world EHR or claim data. Big real-world EHRs and claims offer great opportunities to screen patients for clinical trials and to complement the knowledge gained from the whole drug discovery, development, and post-market monitoring process. However, this is a very difficult problem because the patient records are sparse, longitudinal, heterogeneous and noisy, labeling targeted patients by domain experts are

very expensive, and the screening algorithms should be interpretable and have the ability to incorporate rich knowledge from domain experts. My early research on patients screening driven by real-world clinical data is an interpretable and knowledge-rich data-driven algorithm for screening likely Borderline Personality Disorder (BPD) patients from real-world EHR systems for the clinical trial recruitment in Boehringer Ingelheim Pharmaceuticals, Inc.

The second research area that I've been working on is learning graph (or network) dynamics for analyzing structured and dynamic data gained from complex social systems. 1) My motivating research project is how complex social systems evolve over time at different scales. I studied how WeChat, which is the largest online social network in China, evolves over time at a billion scale. I found that the growth of new users and social links of WeChat social network follows scaling law *over time* [2,3], which contrasts widely accepted word-of-mouth S-shaped growth curve or linear growth curve. I further proposed novel, interpretable and general ordinary differential equations to capture macroscopic growth dynamics of both nodes and links of social networks [2,3,4]. I further propose a stochastic point process to explain the behaviors of adding friends at a microscopic scale [5]. Besides, I studied Tencent Weibo, which is a Twitter-like social media in China, at a million scale, to capture how information spreads on social networks which forms surprisingly complex structural patterns. However, traditional models of information flow cannot generate complex structural patterns in the real world. I proposed a novel data-driven graph-based branching process, which successfully generates real-world complex patterns of information flow [6,7]. One of the related research has also resulted in the *best paper candidate* in ICDM 2018 [8].

2) My WeChat and Weibo projects lead me to a more general problem, namely, learning graph/network dynamics, which widely exists in different complex systems in the real world (e.g. traffic flow in road networks, information flow in social networks, energy flow in biological networks, and bioelectrical flow in brain networks, etc.). However, learning graph dynamics is very difficult because real-world networks are high-dimensional and with complex interactions, dynamics on a network can be continuous-time, nonlinear, and regularly- or irregularly-observed, and structural-dynamic dependencies are difficult to be modeled by simple mechanistic models. My early research on learning graph dynamics is trying to leverage graph (general language for linked systems), ordinary differential equations (ODE, general language for dynamic changes), and deep neural networks (one of the most successful data-driven methods), leading to our graph neural ODE model [8]. Our model is a unified framework to learn continuous-time graph dynamics, structured sequence (regularly observed graph dynamics), and graph semi-supervised learning (a one-snapshot case) [9]. The related research has resulted in KDD 2020 [8] and its workshop version also won the *best paper award* in AAAI 2020 Deep Learning on Graphs: Methodologies and Applications Workshop.

Application.

1) *Borderline Personality Disorder (BoPD).* This disease is characterized by extreme sensitivity to perceived interpersonal slights, an unstable sense of self, intense and volatile emotionality and impulsive behaviors that are often self-destructive, compounding the clinical challenges posed by the severe morbidity, high social costs and substantial prevalence of this disorder in many health-care settings [10]. The motivating problem is to screen likely Borderline Personality Disorder (BoPD) patients from real-world EHRs systems for the clinical trial recruitment of newly developed drug molecules at Boehringer Ingelheim Pharmaceuticals, Inc. I am the leading screening algorithm developer of this project and the developed algorithm is right now being deployed to multiple sites in the US including hospitals, research institutes, and medical schools.

2) *Drug discovery.* AI and big chemical data have the potential to improve drug discovery & development process which are usually lengthy, costly, and prone to failure. I am collaborating with Dr. Fei Wang from Weill Cornell Medicine on deep molecular graph generation and optimization. Early proof-of-concept research [1,11] have shown some promising results.

3) *Social network analysis.* Social networks play a major role in generating and spreading information in modern society and human, group, and social behaviors therein generate big behavior data with rich semantics, which provide great opportunity to investigate critical questions in information flow, fake news, viral memes, epidemics, public health, advertising, recommendations, and so on. I previously worked with the WeChat team funded through Tencent Rhino-Bird Elite Training Program. The related research has resulted in [2,3,4,5,6,7,12,13,14,15].

System. In addition to computational algorithms development and their applications, building software systems implementing these algorithms is a key step to incorporating them into real-world practice and make an impact on everyone's daily life. I have extensive experience in system development as well. Specifically, I am the leading machine learning algorithm developer for screening likely and under-diagnosed Borderline Personality Disorder (BoPD) patients from real-world EHRs systems for the clinical trial recruitment of newly developed drug molecules in Boehringer Ingelheim Pharmaceuticals, Inc. The developed algorithm is right now being deployed to multiple sites in the US including hospitals, research institutes, and medical schools. For deep molecular graph generation and optimization, I led a team with two graduates from Cornell Tech and developed a molecular graph optimization system for drug discovery as illustrated below, and related research resulted in KDD'20 Workshop on Applied Data Science for Healthcare [11]:

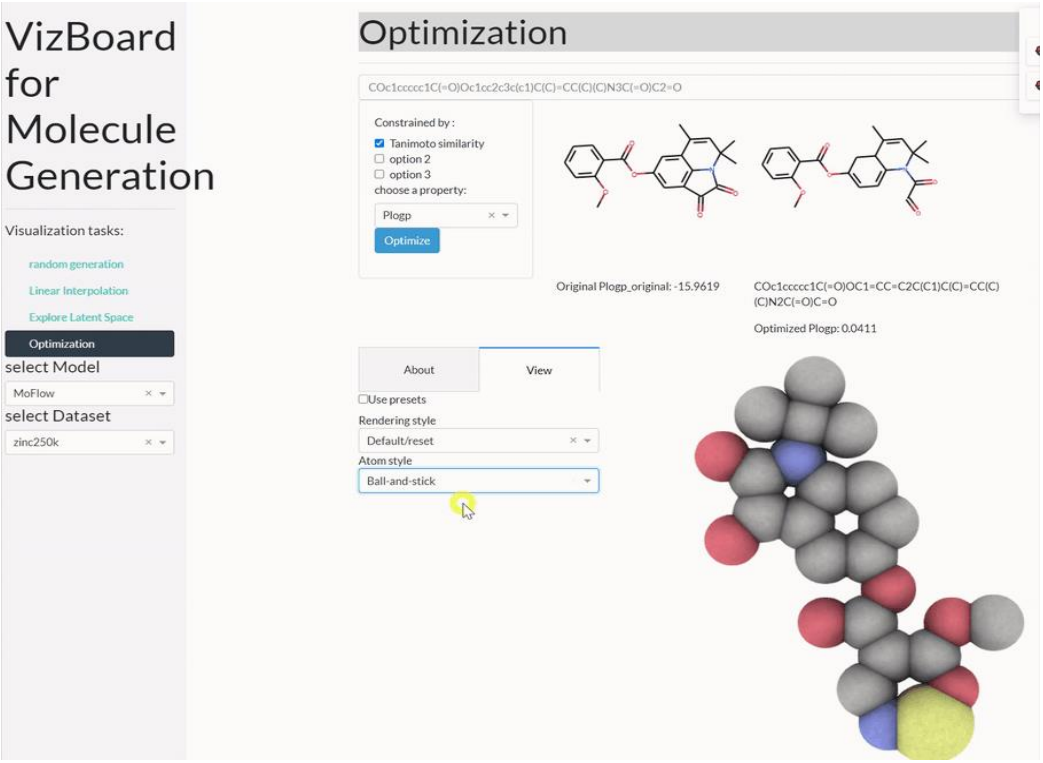

Figure 1. Our developed molecular graph optimization system for drug discovery

2. Vision for Future Research

Similar to how I summarize my existing research, in the following, I will envision my future research in the next 3-5 years on the aspects of the algorithm, application, and system aspects.

2.1 Algorithm

Based on my current research summarized above and my dedication to helping with real-world healthcare, drug discovery & development, I think 1) deep graph generation & optimization models for drug discovery, 2) mining real-world electronic health records (EHR) data, and 3) interpretable and unbiased modelling are the three most important elements to consider next. I will devote my future algorithmic research in these three dimensions as follows:

Deep graph generation and optimization models for drug discovery. Current graph machine learning or graph neural network algorithms majorly focus on inference, embed, encode graphs that take graphs as inputs. However, the reverse procedure, i.e., graph generation & optimization, which are critical techniques for AI-driven drug discovery, is less studied. Graph generation & optimization are very hard tasks that try to generate combinatorial graph structures. What's more, the generated structures should be novel, optimized, and follow some constraints (e.g. valency constraints of valid molecular graphs) in the drug discovery scenario. Specifically, I will consider three aspects of my future research: integrative graph generation, motif-based graph generation, and multi-object graph optimization.

For integrative graph generation, my early work on molecular graph generation [1] has achieved state-of-the-art performance on novelty, validity, uniqueness and drug-likeness optimization. However, some most important properties including binding affinity and selectivity are not considered. We try to integrate gene profiles, protein information, and molecular dynamics into our current molecular graph generative and optimization framework. An integrative graph generation and optimization are critical to the lead discovery and lead optimization.

For motif-based graph generation, we try to generate a molecular graph based on motifs instead of in an atom-wise way. The benefits are motif-based generation have the potential to navigate discrete chemical space further and more efficient, to generate valid molecular graphs, and to inspire multi-object optimization.

For multi-object graph optimization, we try to optimize input graphs into novel and better ones with respect to multiple properties. Such a task is very general in lead optimization.

Mining real-world EHR data. Real-world data (RWD) refers to the patient datasets collected from a wide range of sources except clinical trial cohorts, such as electronic health records (EHR), medical claims, product and disease registries, laboratory test, consumer mobile devices, and social media. RWD EHR have complex and multi-modal structured (e.g. tabular data, ICD codes, medication codes, etc.) or unstructured (clinical notes, images, etc.) data which offer great opportunity to develop cutting-edge machine learning algorithms for real-world healthcare applications (e.g. screening, drug development, risk prognosis & stratification, disease trajectory, subtyping & phenotyping, diagnosis, medical imaging, clinical notes, early warning, real-world evidence, decision support, etc.).

My initial efforts on mining RWD EHR are 1) Screening Borderline Personality Disorder (BoPD) Patients for Clinical Trial Recruitment (1402-0012) in Boehringer Ingelheim *Pharmaceuticals*, Inc. The challenges are BoPD are largely mis- and under-diagnosed, and the supervised labels are very small (hundreds) compared to tens of millions of unlabeled patients. We developed a knowledge-enriched semi-supervised learning framework to learn a screening model with very limited supervision problems; and 2) Predicting Critical Events (ICU transfer, mortality, intubation) in Covid-19 Patients by Contrastive Learning in Mount Sinai *Hospital*. The challenges are positive patients are much less than negative patients. We developed a contrastive learning framework for EHR to deal with imbalanced data problems. Based on our previous efforts, I try to continue developing AI-empowered Predictive Modeling for (interpretable, equal, automatic, privacy-preserved) Prognosis, Screening, Diagnosis, Phenotyping driven by RWD with potential limited supervision, imbalanced labels, and missing data. Our final goals are to integrate our developed predictive models into clinical workflow for clinical decision support.

Interpretable and Unbiased modelling. In healthcare, drug discovery & development problems, model interpretability and model accuracy are equally important. Graph neural networks are the state-of-the-art data-driven methods for structured data, but to interpret them is usually very difficult, and the interpretability can be further worsened by introducing complex graph structures, temporal dimension, and semantics. Specifically, I will consider three aspects in my future research: visual interpretation, mechanistic interpretation, interpretable models.

For visual interpretation, I will leverage prior experience in building effective visualization systems [11] for exploring molecular graph data and deep graph generative models. I will investigate different ways to visualize molecules, e.g., SMILES strings, 2D graphs, and 3D systems. I will also develop interactive ways for users to modify the generated and optimized molecular graphs and get real-time feedback. I will investigate how to interpret deep graph generative models by visualizing the outputs of their hidden layers.

For mechanistic interpretation, my early research shows that by leveraging deep graph neural networks (GNNs) and ordinary differential equations (ODEs), we can model and generate structured and dynamic data derived from healthcare, drug discovery, or social networks and thus improve the model accuracy. I will further explore physical understandings introduced by the ODEs (e.g. diffusion on graphs [8], Hamiltonian system and normalizing flow models [1], dynamical systems [4], etc.). Such mechanistic understandings can help us design novel GNN models with better performance and better interpretability for structured and dynamic data.

For interpretable models, instead of interpreting complex deep models in a post-hoc way, we try to use interpretable models with incorporated domain knowledge to do sensitive clinical tasks (e.g. patient screen, diagnosis, etc.) as we mentioned above. I try to make intrinsically interpretable models (e.g. linear models) more powerful in big-data scenarios and easily incorporated with domain knowledge (e.g. doctors' opinions).

Fairness. Decisions in healthcare are very sensitive to features like gender, age, and ethics, etc. For example, the patient screen algorithm driven by real-world EHR data can be biased with respect to these sensitive features. One example is that there are more young female Borderline Personality Disorder (BPD) patients in our EHR system. I will investigate the algorithm bias of our current screen algorithm and its influence on the clinical trial recruitment. I will also propose updated screen algorithm with fairness constraints.

The research outcomes of the research on above aspects will include publications on top data science and machine learning research venues, such as KDD, ICML, AAAI, IJCAI, etc; as well as funding from NSF (III Core programs), NIH, ONR and industries. I also plan to publish research findings on top clinical journals and top interdisciplinary journals.

2.2 Applications

Despite the algorithm research, having the real-world application context that those algorithms can be developed is more important. Based on my existing experience, I plan to continue focusing my research on drug discovery, Borderline Personality Disorder (BPD) disease, and social media. With the developed methodologies, I am specifically interested in the following problems.

Drug discovery. The main goal of this part of research is to design novel and optimized molecules for pre-clinical and clinical trials. In order to realize this line of research, we need to leverage the large scale chemical, protein data. With our integrative graph generative and optimization models, we try to design novel and optimized molecular graphs in a data-driven way. I will closely collaborate with medicinal chemists to conduct proof-of-concept experimental studies.

Patients screen. The main goal of this part of research is to find likely, say BPD patients from real-world EHR system for future clinical trial recruitments. We try to develop interpretable, fair and domain-knowledge-rich algorithms for this task. We also try to investigate other mental diseases.

Social network analysis. The goal of this part of research is to find dynamical laws of the spreading of information flow (viral new, fake news, memes, etc.) and how to predict and control them in modern social media platforms including WeChat and TikTok. We also try to investigate the fairness issues of the acquisition of information flow.

2.3 Systems

With all the developed algorithms and related contexts, I hope to make them generate impact in routine clinical or drug discovery research and practice. I plan to develop two software systems, one dedicated to drug discovery, the other dedicated to patient screen from real-world EHR system for trial recruitment.

Specifically, as shown in Figure 1, we try to make our current system for drug generation and optimization more user-friendly. Users can type in their seed molecule SIMLEs strings, choose targeted properties and optimization algorithms, and then play with the generated novel and optimized molecular graphs. We try to make interpretability functionalities into our system.

For the patient screen system, I plan to build a BPD patient screen plug that can be integrated into real EHR systems (such as Cerner). The clinicians can select multiple patients and our plug shows if the selected patients are likely BPD patients. The plug can show the distribution and statistics of sensitive demographic features (e.g. age, gender, ethics, etc.), and the clinicians can choose which features to be controlled.

3. Education and Training

I am willing to involve student researchers and postdocs in my research. I am willing to teach both under-graduate-level and graduate-level course with the topic of graph machine learning, or graph machine learning with applications in healthcare. In addition, I will actively create research projects and engage the MS and Ph.D. students in my future research. I will also involve students from other programs (e.g., tri-institute Ph.D. program and MD

programs), other campuses (e.g. I am currently mentoring two graduates from Cornell Tech in my drug discovery project), and other universities. I will also work with other faculty and colleagues to secure NIH and PCORI training grants.

4. Community Engagement

I have been actively engaged in activities of both data mining and machine learning communities. I have successfully finished two tutorials on the top venues including KDD2020, AAAI2020. I served as program members of AAAI, ICDM, ICKG, CIKM, IJCAI, etc., and reviewers for many top journals including TKDE, TKDE, KAIS, TBD, etc. In the future, I will continue these activities to try to onboard program members of more top venues. I also plan to organize workshops and tutorials on health data mining, graph machine learning for healthcare, AI-driven drug discovery, etc.

Reference

[1] Zang, Chengxi, and Fei Wang. "MoFlow: an invertible flow model for generating molecular graphs." In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 617-626. 2020.

[2] Zang, Chengxi, Peng Cui, and Christos Faloutsos. "Beyond sigmoids: The nettide model for social network growth, and its applications." In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 2015-2024. 2016.

[3] Zang, Chengxi, Peng Cui, Christos Faloutsos, and Wenwu Zhu. "On power law growth of social networks." IEEE Transactions on Knowledge and Data Engineering 30, no. 9 (2018): 1727-1740.

[4] Zang, Chengxi, Peng Cui, Wenwu Zhu, and Fei Wang. "Dynamical Origins of Distribution Functions." In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 469-478. 2019.

[5] Zang, Chengxi, Peng Cui, Christos Faloutsos, and Wenwu Zhu. "Long short memory process: Modeling growth dynamics of microscopic social connectivity." In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 565-574. 2017.

[6] Zang, Chengxi, Peng Cui, Chaoming Song, Wenwu Zhu, and Fei Wang. "Uncovering Pattern Formation of Information Flow." In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1691-1699. 2019.

[7] Lu, Yunfei, Linyun Yu, Tianyang Zhang, Chengxi Zang, Peng Cui, Chaoming Song, and Wenwu Zhu. "Collective Human Behavior in Cascading System: Discovery, Modeling and Applications." In 2018 IEEE International Conference on Data Mining (ICDM), pp. 297-306. IEEE, 2018.

[8] Zang, Chengxi, and Fei Wang. "Neural dynamics on complex networks." In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 892-902. 2020.

[9] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." arXiv preprint arXiv:1609.02907 (2016).

[10] Gunderson, John G., Sabine C. Herpertz, Andrew E. Skodol, Svenn Torgersen, and Mary C. Zanarini. "Borderline personality disorder." Nature Reviews Disease Primers 4, no. 1 (2018): 1-20.

[11] Yang, Karan, Chengxi Zang, and Fei Wang. "Visualizing Deep Graph Generative Models for Drug Discovery." arXiv preprint arXiv:2007.10333 (2020).

[12] Lu, Yunfei, Linyun Yu, Tianyang Zhang, Chengxi Zang, Peng Cui, Chaoming Song, and Wenwu Zhu. "Exploring the collective human behavior in cascading systems: a comprehensive framework." Knowledge and Information Systems (2020): 1-25.

[13] Li, Haoyang, Peng Cui, Chengxi Zang, Tianyang Zhang, Wenwu Zhu, and Yishi Lin. "Fates of Microscopic Social Ecosystems: Keep Alive or Dead?." In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 668-676. 2019.

[14] Zang, Chengxi, Peng Cui, Chaoming Song, Christos Faloutsos, and Wenwu Zhu. "Quantifying structural patterns of information cascades." In Proceedings of the 26th International Conference on World Wide Web Companion, pp. 867-868. 2017.

[15] Lu, Yunfei, Linyun Yu, Peng Cui, Chengxi Zang, Renzhe Xu, Yihao Liu, Lei Li, and Wenwu Zhu. "Uncovering the Co-driven Mechanism of Social and Content Links in User Churn Phenomena." In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3093-3101. 2019.