Tianfan Fu (Georgia Institute of Technology) tfu42@gatech.edu

Machine learning for drug discovery and development

My research interest is **machine learning for drug discovery and development (ML4Drug)**. In general, my research can be grouped into two categories: (1) developing cutting-edge machine learning algorithms for essential drug discovery and development tasks (Section 2); (2) building up foundation for ML4Drug to benefit the whole community (Section 3). The uniqueness of my research can be summarized as follows.

- I develop cutting-edge machine learning algorithms for drug discovery and development — an emerging, life-critical and interdisciplinary field with incredible opportunities for innovation and impact. I have published my works in both top machine learning venues (NeurIPS, ICLR, KDD, AAAI, IJCAI) and domain-specific journals (Cell Patterns, Nature Chemical Biology, Bioinformatics). (Section 2)
- My work has a significant industrial impact. For example, my Hierarchical Interaction Network (HINT) paper [11] for clinical trial outcome prediction and DeepPurpose paper [4] for drug repurposing have been deployed by IQVIA, one of the largest clinical research organizations (CROs) in the world.
- I work closely with domain experts to solve real-world applications. For example, I have collaborated with Prof. Connor W. Coley (a chemist and Assistant Professor at MIT) on five drug discovery papers. I have also collaborated with Dr. Lucas Glass (Vice President at IQVIA) on five drug discovery & development papers.
- I contribute to ML4Drug community by writing a comprehensive textbook [18], curating data hubs [5, 6], generating benchmarks [11, 3, 4, 5], co-founding the AI4Science workshop (at NeurIPS and ICML), and writing AI4Science review paper [19]. (Section 3)

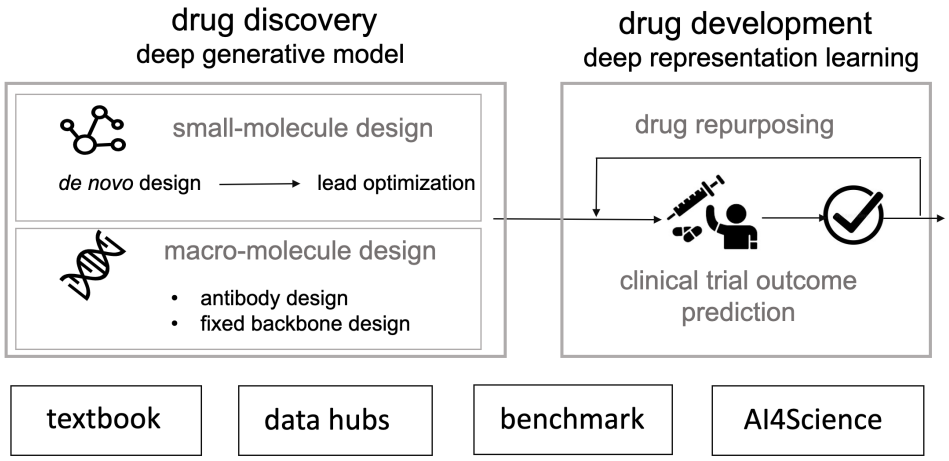

Figure 1: My research focuses on developing cutting-edge ML algorithms for drug discovery and development, including (i) **deep generative models** for both small-molecule design [2, 1, 17, 14, 15] and macro-molecule design [13, 12], (ii) **deep representation learning** for both clinical trial outcome prediction [11] and drug repurposing [16, 4]. Also, I contribute to the research community by writing a textbook [18], curating ML-ready data hubs [5, 6] and benchmarks [11, 3, 4, 5, 6], and building up the AI4Science community [19].

1 Motivation

Novel types of safe and effective drugs are needed to meet the medical demands of billions worldwide and improve human life quality. Bringing a novel drug to the market currently takes 13-15 years and between $2-3 billion, on average. On the other hand, with the huge volumes of open data produced in

biomedical & healthcare research and algorithmic development, machine learning (ML) for drug discovery and development is a quickly emerging field with incredible opportunities for innovation and impact. The global market of AI/ML for drug is estimated to reach $4 billion by 2027 from $0.6 billion in 2022, at a compound annual growth rate of 45.7% during the foreseen period. However, some challenges remain; for example, it is computationally prohibitive to navigate the astronomically large drug space, and it is hard to capture the interaction between multimodal data for clinical trial predictive modeling.

2 Pushing the Frontier of ML4Drug

I mainly focus on designing cutting-edge ML methods for essential drug discovery and development problems. For drug design, I propose a series of *deep generative models* to design small-molecule and macromolecule drugs (two major drug categories). Also, I design *deep representation learning* methods for predictive drug development problems.

1. Small-molecule drug design. Small-molecule drug, a.k.a. chemical drug, is the widest class of drugs. Small-molecule drug design has two stages: *de novo* design and lead optimization.

• ***De novo* design**. *de novo* drug design aims at producing novel and diverse drug molecules with ideal pharmaceutical properties from scratch. The key challenge is to traverse the discrete chemical space efficiently. Specifically, in order to circumvent the discrete nature of drug molecules and alleviate the brute-force trial-and-error strategy, [2] relaxes the discrete drug molecule into *differentiable scaffolding tree (DST)* to enable the gradient-based numerical optimization to update the differentiable molecule directly, which enable gradient-based optimization on drug molecules. Empirical studies demonstrate the proposed DST method is more sample-efficient and is able to identify desired molecules within thousands of evaluations (oracle calls). The oracle call could be *in vivo* or even *in vitro* experiment and is always costly. This means our method would significantly reduce the cost of drug design. Further, motivated by genetic algorithm's superior but unstable performance (due to random-walk behavior), *Reinforced Genetic Algorithm* [1] was designed to suppress random-walk behavior, which leverages reinforcement learning to prioritize the promising search branches and navigate the discrete space intelligently. The generated molecules can bind tightly to the target proteins that are closely associated with some critical diseases, e.g., the target whose PDB ID is 7l11, which is SARS-COV-2(2019-NCOV) main protease. Also, to quantify uncertainty and explore chemical space thoroughly, Multi-constraint Molecule Sampling (MIMOSA) [15] formulates drug design problem as a sampling problem that samples from the target distribution over the drug space. The desirable drug molecule has larger probability, and then design a Markov Chain Monte Carlo (MCMC) method coupled with pretrained graph neural network to sample from the target distribution. It obtains up to 49% improvement over the strongest baseline.

• **Lead optimization**. Lead optimization aims at enhancing the lead compound (typically the most promising molecule in *de novo* design) via improving its pharmaceutical properties (e.g., diminishing toxicity, and improving absorption) and maintaining its similarity with the lead compound. The key challenge lies in the satisfaction of multiple constraints. To explicitly enhance the similarity constraint, Copy & Refine strategy (CORE) [17] was designed to select existing substructures (substructure is the basic building block) from the input drug molecule with attention mechanism, instead of searching over the entire substructure space. In addition to consistent improvement across multiple tasks, CORE achieves especially outstanding performance in the molecule with rare substructures, with up to 11% gain in success rate. Also, lead optimization requires consistency between the size of the input and output drug molecules. To meet the requirement, Deep Generative Model with Molecule Reward (MOLER) [14] was proposed to impute the constraints into a differentiable loss function into the learning objective. It is a model-agnostic approach that can enhance almost all deep generative models.

2. Macro-molecule drug design. Macro-molecule drugs (a.k.a. biologics) are another mainstream kind of drug. I focus on two core problems: fixed backbone protein design (a.k.a. inverse protein folding) and antibody Complementarity Determining Regions (CDR) loop design. Traditional antibody CDR loop design suffers from poor validity issues of the generated antibody CDR loops. To address this issue, Constrained Energy Model [12] constrains the energy model (one of deep generative models) on the manifold, where all the antibodies are geometrically valid. Thus, it is able to generate 100% valid antibody CDR loops. Existing fixed backbone design methods mostly suffer from the autoregressive generation manner (i.e.,

sequentially generating amino acids). To alleviate the issue, [13] designs a flexible strategy that updates amino acids with adaptive weight, which enables thorough exploration at the uncertain area.

3. Drug development (clinical trial). Drug development mainly contains clinical trials that evaluate the safety and effectiveness of the drug on human bodies. I focus on two fundamental tasks: (i) predicting the trial approval rate and (ii) drug repurposing which reuses the approved drugs to cure new disease conditions. Specifically, to obtain an accurate trial approval rate prediction, Hierarchical Interaction Network (HINT) [11] designed a hierarchical interaction network to leverage domain knowledge and capture the interactions between multi-modal trial features (drug molecule, disease codes, and trial protocol). HINT achieves 0.85 F1 scores in phase III trial approval prediction and successfully predicts the failure of some well-known trials. Phase III is the most costly among all the three phases and typically involves 1,000-3,000 volunteers and takes 1-4 years. This high accuracy of HINT model enables clinical scientists to identify the trials that are likely to fail and save lots of time and funding. It has been routinely deployed by IQVIA, one of the largest clinical research organizations (CROs). Further, [16] proposed probabilistic and dynamic drug-disease interaction models to quantify the uncertainty and temporal trends for *drug repurposing*, the repurposed drugs are validated by real-world patient claims records. For example, our model finds the drug Clopidogrel can be reused to treat obstructive hydrocephalus, which is validated by claims data evidence and the literature [10]. Also in drug repurposing, [4] assesses the effect of a series of neural network architectures on drug-target interaction (target is usually associated with certain diseases), which has been routinely used by IQVIA internal teams for drug repurposing.

3 Building up ML4Drug Foundation

Despite rapid growth in recent years, the ML4Drug community still lacks solid foundations, e.g., educational resources, ML-ready datasets, specialized open-source software, etc. To meet the needs of both pharmaceutical and CS/ML researchers, I put lots of effort into building up ML4Drug foundation.

1. Textbook: Machine Learning for Drug Discovery and Development[1]. To fill in the blanks in educational resources, Dr. Cao Xiao, Prof. Jimeng Sun, and I are writing the textbook [18] in this space. The target readers are primarily: (1) graduate (e.g., MS and Ph.D. students) or advanced undergraduate students majoring in computer science, engineering, medicine, chemistry, biology, medical informatics, or biostatistics; (2) data scientists from the pharmaceutical and biotech industries. The book covers both ML and drug discovery & development basics and elaborate ML4Drug tasks with hands-on examples. The book is expected to be published by Springer Press in May 2023.

2. Data hubs: Therapeutics Data Commons (TDC) [5, 6] is a collection of 22 ML-solvable drug discovery tasks, 66 ML-ready datasets with a total of 15M data points. Since its inception in 2021, over 39K scientists worldwide have used TDC. Also, it has 3K active users every month whose backgrounds span disciplines of computer science, chemistry, and biology, indicating that TDC is a broadly interesting initiative.

3. Benchmark. I lead/co-lead the following open-source benchmarks.

• Therapeutics Data Commons (TDC) [5, 6] curated three drug discovery & development benchmarks, including drug design, drug property prediction, and drug synergistic effect prediction.

• DeepPurpose [4] is a deep learning library for drug and target protein modeling, which can predict drug property, protein function, drug-target interaction, protein-protein interaction, etc.

• [11] curates the first public benchmark dataset for general-purpose clinical-trial-outcome predictions, comprising 18K clinical trials, 14K drugs, and 5K diseases. The raw data sources come from drug knowledge bases, disease code databases, historical clinical trial records, and manually curated trial outcome labels.

• Practical Molecular Optimization benchmark [3] thoroughly investigates the performance of 25 drug design algorithms on 23 optimization tasks with a particular focus on sample efficiency.

4. AI for Science (AI4Science) [19]. ML4Drug falls into the broad class of AI4Science due to its inter-

[1] https://ml4drug-book.github.io/

disciplinary nature (chemistry, biology, physics, health). I co-founded the series of AI4Science workshops[2] at leading AI/ML venues (NeurIPS 2021, ICML 2022, and NeurIPS 2022 upcoming) to bring researchers working on AI4Science together and consolidate the fast-growing area into a recognized field. Further, the core organizer team issued an AI4Science review paper [19], which is the first systematic and comprehensive outlook of how AI reshapes scientific discovery, where pharmaceutical science occupies a large proportion.

4 Future Directions

1. Broaden the scope of clinical trial outcome prediction. The current Hierarchical Interaction Network (HINT) model leverages drug molecules (only supporting small-molecule drugs), disease codes, & trial eligibility criteria as the input features and predicts the trial approval rate. I plan to extend the scope of the HINT model from several aspects: (i) supporting the prediction of macro-molecule drugs; (ii) leveraging more trial features, e.g., trial sponsor/location information, trial estimated time, expected number of recruited patients; (iii) predicting more granular trial outcome (e.g., whether the trial would terminate, the failure reason of trial).

2. Designing other types of drugs. In addition to two major categories of drugs (small-molecule and macro-molecule drugs), other types of treatments, e.g., peptide vaccine, messenger RNA (mRNA), gene therapy, have also exhibited their superiority in curing certain diseases, experienced rapid growth recently but are still under-explored by ML. I plan to explore ML's applications in designing novel types of therapies. For example, peptide vaccine design can be viewed as a Covering Integer Programming problem [9], but traditional methods rely heavily on heuristic search and human-curated rule. I plan to leverage reinforcement learning to navigate the discrete space intelligently. Reinforcement learning is able to estimate the expected reward of each search branch and prioritize promising branches. Compared with traditional heuristic search approaches, reinforcement learning methods typically suppress random-walk behavior and exhibit better generalization performance on unseen data.

3. Extending 1D string, 2D graph to 3D geometric structure. Most of the state-of-the-art drug discovery methods learn from 2D molecular graph or 1D protein amino acid sequences. However, both drug and target protein exists in the format of 3D geometric structures. Currently, the algorithmic development, e.g., AlphaFold2 (protein structure prediction) [7] and equivariant neural network [8], enables the modeling of the 3D geometric structure, which can potentially promote a series of small-molecule and macro-molecule modeling tasks, including drug design, drug-target interaction, etc.

References

[1] Fu, Tianfan*, Gao, Wenhao*, Connor W Coley, and Jimeng Sun. Reinforced genetic algorithm for structure-based drug design. In *Neural Information Processing Systems (NeurIPS)*, 2022.

[2] Fu, Tianfan*, Gao, Wenhao*, Cao Xiao, Jacob Yasonik, Connor W Coley, and Jimeng Sun. Differentiable scaffolding tree for molecular optimization. *The International Conference on Learning Representations (ICLR)*, 2022.

[3] Gao, Wenhao*, Fu, Tianfan*, Jimeng Sun, and Connor W Coley. Sample efficiency matters: benchmarking molecular optimization. *Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022.

[4] Kexin Huang, Fu, Tianfan, Lucas Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. Deeppurpose: a deep learning library for drug-target interaction prediction and applications to repurposing and screening. *Bioinformatics*, 2020.

[5] Huang, Kexin*, Fu, Tianfan*, Gao, Wenhao*, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: machine learning datasets and tasks for therapeutics. *NeurIPS Track Datasets and Benchmarks*, 2021.

[6] Huang, Kexin*, Fu, Tianfan*, Gao, Wenhao*, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Artificial intelligence foundation for therapeutic science. *Nature Chemical Biology*, 2022.

[2] https://ai4sciencecommunity.github.io/

[7] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[8] Victor Garcia Satorras, Emiel Hoogeboom, Fabian B Fuchs, Ingmar Posner, and Max Welling. E(n) equivariant normalizing flows. 2021.

[9] Aravind Srinivasan. Improved approximation guarantees for packing and covering integer programs. *SIAM Journal on Computing*, 29(2):648–670, 1999.

[10] Y Tamura et al. Aspirin/clopidogrel. *Reactions*, 1483:18–11, 2014.

[11] Fu, Tianfan, Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. HINT: Hierarchical interaction network for clinical-trial-outcome predictions. *Cell Patterns*, 3(4):100445, 2022.

[12] Fu, Tianfan and Jimeng Sun. Antibody complementarity determining regions (CDRs) design using constrained energy model. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2022.

[13] Fu, Tianfan and Jimeng Sun. SIPF: Sampling method for inverse protein folding. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2022.

[14] Fu, Tianfan, Cao Xiao, Lucas Glass, and Jimeng Sun. MOLER: Molecule-level reward to enhance deep generative model for molecule optimization. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2021.

[15] Fu, Tianfan, Cao Xiao, Xinhao Li, Lucas M Glass, and Jimeng Sun. MIMOSA: Multi-constraint molecule sampling for molecule optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[16] Fu, Tianfan, Cao Xiao, Cheng Qian, Lucas M Glass, and Jimeng Sun. Probabilistic and dynamic molecule-disease interaction modeling for drug discovery. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 404–414, 2021.

[17] Fu, Tianfan, Cao Xiao, and Jimeng Sun. CORE: Automatic molecule optimization using copy and refine strategy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 638–645, 2020.

[18] Fu, Tianfan, Cao Xiao, and Jimeng Sun. Machine learning for drug discovery and development. *Springer (to appear)*, 2023.

[19] Wang, Hanchen*, Fu, Tianfan*, Du, Yuanqi*, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Animashree Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Connor Coley, Yoshua Bengio, and Marinka Zitnik. Enabling scientific discovery with artificial intelligence. In *Submission*, 2022.