Patterns

CellPress
OPEN ACCESS

People of Data

Integrating supercomputing and artificial intelligence for life science

Jiahua Rao,[1,*] Shuangjia Zheng,[1,2,*] and Yuedong Yang[1,3,*]
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510000, China
[2]Galixir Technologies Ltdr, Beijing 100000, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, Guangzhou 510000, China
*Correspondence: raojh6@mail2.sysu.edu.cn (J.R.), zhengshj9@mail2.sysu.edu.cn (S.Z.), yangyd25@mail.sysu.edu.cn (Y.Y.)
https://doi.org/10.1016/j.patter.2022.100653

Jiahua Rao and Shuangjia Zheng are Ph.D. students in Prof. Yang's lab (Supercomputing And AI for Life science, SAIL Lab) at Sun Yat-sen University. They recently developed an interpretable framework to quantitatively assess the interpretability of Graph Neural Network (GNN) and made comparison with medicinal chemists. Their meaningful benchmarking and rigorous framework would greatly benefit development of new interpretable methods in GNNs.

What would you like to share about your background (personal and/or professional)?

Yuedong Yang: By trained as a computational biologist, I have been working on protein structure prediction and related applications for more than 20 years. In the early years, I employed ab initio simulations through molecular dynamics due to limited availability of experimental data. With the development of high-throughput biotechnology, the amount of biomedical data is exponentially increasing, and I gradually turned to data-driven strategy by firstly statistical approaches and then machine learning. In 2013, I moved from USA to Australia and got a chance to collaborate with colleagues who have survived the last winter of artificial intelligence. Our collaborations led to the successful prediction of protein torsional angles,[1] also one of the earliest works to open up the door for deep learning-based protein structure prediction. In 2017, I joined the school of computer science and engineering in the Sun Yat-sen University, China, where I have built up a multi-disciplinary team to develop supercomputing and AI algorithms for drug design and multi-omics big data analysis. Benefiting from my versatile experience, my group has frequently collaborated with many hospitals, medical schools, pharmaceutical industry, etc. Hence, we are motivated to develop a comprehensive computational platform for biomedical applications in both academics and industry.

Jiahua Rao: I am a Ph.D. student of Sun Yat-sen University, advised by Prof. Yang. Before joining SAIL Lab, I was trained in Computer Science as well as Mathematical Science. I realized my love for research when I joined the SAIL lab and discovered the world of artificial intelligence for science. My research interest is to develop deep learning methods for solving computational problems in the natural sciences, especially in bioinformatics and drug discovery. At the moment, I'm developing interpretable benchmarks and methods for enhancing the interpretability of graph neural networks,[2] which could be a key to accelerating drug discovery. I am also working to integrate prior knowledge data (e.g. knowledge graph, multi-modal data) into deep learning models for next-generation interpretability.

Shuangjia Zheng: I am a final year Ph.D. candidate in computer science and have finished my undergraduate and M.Phil training in pharmaceutical science. Because of the interdisciplinary training, I have always been drawn to research projects in which I can solve fundamental problems in drug discovery while also push the state-of-the-art techniques in machine learning. I also work closely with the pharmaceutical industry and therefore have a sense of the challenges being faced. When I am not working, I'm also a big fan of e-sports and a fingerstyle guitar player.

What motivated you to become a (data) researcher? Is there anyone/anything that helped guide you on your path?

YY: My Ph.D. training involved molecular dynamics simulations of proteins. Although theoretically beautiful, the simulations suffer from relatively poor potential functions, and simulated results usually deviate from experimental observations due to accumulative errors. During my postdoc training, I utilized experimental data to design statistical potential function for protein interaction, which showed a great step to match with the observed. Further applications of machine learning techniques, such as SVM, enabled me to develop bioinformatics methods for truly predicting or guiding biological experiments. With the successes of utilizing deep learning techniques, I am deeply devoted to the data-driven paradigm.

JR: The rise of Artificial intelligence (AI) has inspired widespread applications and innovations in many fields, and AI quickly becomes one of my favorite parts of research. My research interest comes from my mentor and colleagues having trained me during my research career. When I did a research internship in the SAIL lab during my undergraduate, I became completely fascinated by the power of AI to solve computational problems in the natural sciences. I enjoy communicating and sharing my thoughts with my colleagues from different academic backgrounds. My enthusiasm for solving scientific problems with AI models motivates me to become a data researcher.

SZ: During my undergraduate training, I observed that the drug discovery process was very labor intensive and time consuming. I was wondering how to position these drugs from the infinite chemical/biological space rationally and

intelligently. This sense of curiosity drove me to study data science. My supervisor, Prof. Yang, was very instrumental in helping me be a researcher. He showed me what decent work is and gave me enough freedom to explore the directions I was interested in.

What is the definition of data science in your opinion? What is a data scientist? Do you self-identify as one?

SZ: I think of data science as using domain data to solve scientific problems. A data scientist is someone who has experience in data analysis, processing, and modeling, and can leverage the experience to solve practical problems. I work closely with many talented data scientists, so I don't consider myself a pure data scientist. Data scientists are more like problem solvers, whereas I'm better at asking problems.

JR: There are many different definitions of data science. In my opinion, data science is mining the knowledge and valuable insights from the data and then building solutions to real-world problems. The scientific data is usually of large volume, noisy, and heterogeneous. Therefore, the data analysis could be very difficult and costly, and the knowledge behind data might be hard to uncover. Data scientists are those who have the ability to address these problems from massive amounts of data by developing high-performance scientific methods and algorithms. As for me, I would identify myself as a data analyst and at a point of transition to being a data scientist.

YY: Data science aims to find ways that could be commonly utilized to discover natural rules behind observed data. With the development of deep learning, deep neural network seems to be a solution for big data, particularly with the unified solution for visual and language data. Unfortunately, obtaining "enough" data is usually prevented in natural science due to time-consuming and expensive costs. In contrast, humans usually make correct decisions only from small data. Thus, small or even zero data is an important topic in AI research. Such scenario requires us to fully utilize prior domain knowledge, and sometimes physical rules. Therefore, a true data scientist, in addition to skills in data analysis, also needs to be familiar with domain knowledge for analyzing and solving the problems. From this angle, I can be sorted as a data scientist, although my research involves an increasing weight of theoretical simulations with the help of supercomputers. My dream is to construct a theoretically computational framework for biology, like Newton's laws for physics, and then we might not desire big data any more.

Why did you decide to publish in *Patterns*?

YY: We selected *Patterns* because it could expose my work to data scientists as well as biologists and chemists. Cell Press has gathered top scientists in life science, and I believe our explainable AI study over molecules will attract their attentions.

JR: I like *Patterns* because its goal is to publish groundbreaking research across the full breadth of data science, which means that our publications will be read by a wide audience interested in data science. That's exactly what we've been doing, making our work have more impact on different scientific fields.

SZ: *Patterns* is a decent journal and has a wide audience interested in computational methods, especially in the field of biomedicine.

What barriers have you faced in pursuing data science as a career?

YY: As a data scientist, a big difficulty is to obtain the data. In early years, I attempted to seek collaborations with hospitals for using their biomedical data. No one was interested partly because I was junior. I have to utilize public data and make progress over well-defined benchmark questions. Although I made significant progress in my data mining skills, I had little chance to raise interesting scientific questions. Since my joining the current institute, a comprehensive university affiliated with ten top hospitals, I have quickly built up collaborations with hospitals and related departments. Such collaborations not only provide enriched biomedical data, but also bring domain knowledge to solve the problems and initiate many interesting projects. From my experience, building up effective collaborations is critical for the success of data scientists.

How do you keep up to date with advances in both data science techniques and in your field/domain?

JR: I keep up to date with advances by reading recent papers, attending conferences, and participating in workshops. The discussions with peers from different fields are especially helpful and give me further insights into applying data science to real-world problems.

What is the role of data science in your domain/field? What advancements do you expect in data science in this field over the next 2–3 years?

YY: The high-throughput sequencing techniques have accumulated EB level of data, data analysis becomes indispensable in current biological studies. Though there are many successful implementations of deep learning techniques in biological data,[3] such studies are meeting with bottlenecks due to the multi-modal, high-dimensional, and noisy data from the complicated biological system. This could be conquered from three directions. First, prior knowledge has been widely utilized to interpret data and models, and our group has comprehensively compiled relations between gene, disease, and drug to form a novel knowledge graph, PharmKG. Second, the combination of AI and physical simulations might become more and more important with the increasingly faster supercomputer and better understanding of biological systems. Last but not least, new techniques like automation[4] could be introduced to enable higher throughput of data generation.

What is the fun part of being a data scientist?

JR: The fun part of being a data scientist for me is that I can discover new ideas from large amounts of data to solve those real-world problems. In addition, cross-disciplinary communication is also enjoyable, where we could share different understandings of domain-specific knowledge and data.

What drew you to your current team and topic?

SZ: I've been interested in the application of interpretability to pharmaceuticals for a long while. We were one of the first to apply attention mechanisms to solve the black-box problem of QSAR. However, we also

Subset of the Yang Lab (from left to right): Jiahua Rao, Sheng Chen, Shuangjia Zheng, Qianmu Yuan, Lingxue Dai, Haoyang Zhang, Yi Wang, Yuansong Zeng, Yuedong Yang, Zuoyi Wei, Maoling Ding, Zhongyue Zhang

recognized that it is difficult to assess the interpretability quantitatively. Recently, an interesting review[5] highlighted this problem and inspired us to put more efforts into it. We teamed up with an AI drug discovery startup, GALIXIR, to explore this topic deeply. The experienced medicinal chemists in the company gave us a lot of guidance on experimental design and helped us to finish this project.

What's next for the project? What's next for you?

SZ: An upgraded version of this work[2] has now been deployed in the partner company's platform and is playing an essential role in real-world pipelines, particularly in the problems of property prediction and lead optimization. The project is also constantly being updated to introduce better representation models and interpretable modules. I will continue to work on intelligent drug discovery and hope to bring more exciting work to the field.

How did this project you wrote about come to be? Was there a particular result that surprised you, or did you have a eureka moment? How did you react?

JR: The project[2] comes from our general interest in molecular representation learning in graph neural networks, also one of classical cheminformatics tasks. Graph Neural Networks (GNNs) have received increasing attention because of their expressive power on topological data, but they remain criticized for their lack of interpretability. To interpret GNN models, explainable artificial intelligence (XAI) methods have been developed. However, these methods are limited to qualitative analyses without quantitative assessments from real-world datasets, mainly due to a lack of ground truths. Therefore, we established five benchmarks and quantitatively assessed XAI methods on four GNN models, and further made comparisons with medicinal chemists of different experience levels. We found that XAI methods could deliver reliable and informative answers in identifying the key substructures. This finding is exciting, but not the end of the story. We further found that the identified substructures were shown to complement existing classical fingerprints to improve molecular property predictions. We believe that further development of new XAI methods in GNNs would greatly benefit from our meaningful benchmarking and rigorous framework.

REFERENCES

1. Lyons, J., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K., Sattar, A., Zhou, Y., and Yang, Y. (2014). Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. J. Comput. Chem. *35*, 2040–2046. https://doi.org/10.1002/jcc.23718.

2. Rao, J., Zheng, S., Lu, Y., and Yang, Y. (2022). Quantitative Evaluation of Explainable Graph Neural Networks for Molecular Property Prediction. Patterns *3*, 100628.

3. Rao, J., Zhou, X., Lu, Y., Zhao, H., and Yang, Y. (2021). Imputing single-cell RNA-seq data by combining graph convolution and autoencoder neural networks. iScience *24*, 102393. https://doi.org/10.1016/j.isci.2021.102393.

4. Qiu, J., Xie, J., Su, S., Gao, Y., Meng, H., Yang, Y., and Liao, K. (2022). Selective functionalization of hindered meta-C–H bond of o-alkylaryl ketones promoted by automation and deep learning. Chem. https://doi.org/10.1016/j.chempr.2022.08.015.

5. Jiménez-Luna, J., Grisoni, F., and Schneider, G. (2020). Drug discovery with explainable artificial intelligence. Nat. Mach. Intell. *2*, 573–584. https://doi.org/10.1038/s42256-020-00236-4.

About the authors

Yuedong Yang is a Professor in the School of Computer Science and Engineering, Sun Yat-sen University, China. Dr. Yang has published >150 research articles, accompanied with over 50 bioinformatic tools that are widely used by academics and industry. His current research interests focus on developing algorithms for intelligent drug design and multi-scale omics data analysis. In addition, he is responsible for developing one-stop platform for biomedical computations on the "Tianhe-2" supercomputer.

Jiahua Rao is a Ph.D. student in the School of Computer Science and Engineering, Sun Yat-Sen University, advised by Prof. Yuedong Yang. His research interests include deep learning, knowledge graph, multi-omics integration, and drug discovery.

Shuangjia Zheng is a final year Ph.D. candidate in the School of Computer Science and Engineering, Sun Yat-Sen University, supervised by Prof. Yuedong Yang. His research interests lie in deep learning, drug discovery, knowledge graph, and computational biology. He has published >30 publications in high-profile journals and conferences that have been cited over 1,300 times. He also works as the VP of research in GALIXIR, a Chinese AI drug discovery startup.