Patterns

CellPress
OPEN ACCESS

People of Data

Toward ordered -omics data science: Researchers on the magic of turning metagenomic chaos into image-like patterns

Wan Xiang Shen[1,2,*] and Yu Zong Chen[1,3,*]
[1]Bioinformatics and Drug Design (BIDD) Group and Center for Computational Science and Engineering, Department of Pharmacy, National University of Singapore, Singapore 117559, Singapore
[2]Department of Chemistry, Faculty of Science, National University of Singapore, Singapore 117543, Singapore
[3]The State Key Laboratory of Chemical Oncogenomics, Key Laboratory of Chemical Biology, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, PR China
*Correspondence: wx.shen@nus.edu.sg (W.X.S.), chenyuzong@sz.tsinghua.edu.cn (Y.Z.C.)
https://doi.org/10.1016/j.patter.2022.100673

Wan Xiang Shen, a postdoctoral researcher at National University of Singapore, and Yu Zong Chen, the PI of the Bioinformatics and Drug Design (BIDD) group, have developed an AI pipeline for enhanced deep learning of metagenomic data. Their *Patterns* paper highlights the advantages of unsupervised data restructuring in microbiome-based disease prediction and biomarker discovery. They talk about their view of data science and the backstory of the article published in *Patterns*.

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

Wan Xiang Shen: I received my BS in life science and my MS in cheminformatics. After graduation, I spent 3 years working as a data scientist and algorithm researcher at Tsingdata D-LAB and Megvii (Face++). During this time, I worked on a range of projects, including recommendation technology, facial detection and recognition, agricultural data science, time series analysis, and bank management data. This experience highlighted the importance of data science in diverse fields and helped me develop my skills as a senior data scientist.

I graduated from the pharmacy PhD program at the National University of Singapore (NUS), where I was part of the Bioinformatics and Drug Design (BIDD) group in Prof. Chen's lab. I have been trained in pharmaceutical science and bioinformatics and have become interested in the intersection of computer science, data science, healthcare, and therapeutics. I am currently a postdoc researcher in the chemistry department at NUS, where I am working on an end-to-end automatic drug design, synthesis, and bio-screening drug discovery pipeline. My research focuses on developing novel methodologies for deep learning of high-dimensional biomedical data and on exploring data-driven approaches to facilitate biomedical research and accelerate the drug discovery process.

Yu Zong Chen: My PhD study was in the field of statistical physics and nonlinear science, and my postdoctoral research was in the field of computational biophysics. Immediately after my postdoctoral career, I entered the fields of bioinformatics, computer-aided drug design, and biomedical AI, in which I've worked until now for 30 years. My current research interests are biomedical AI with particular focus on drug discovery, healthcare, and bio-modeling.

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

WXS: I was drawn to data research because of the opportunity to use programming and data science techniques to solve complex problems in diverse fields. My main motivation is the ability to apply these skills to tackle important scientific and technological challenges and to make meaningful contributions to our understanding of the world around us.

My PhD advisor, Prof. Chen, has been a major influence on my path as a researcher. He encouraged me to challenge myself academically and to explore interesting problems in bioinformatics. Through our discussions and my own research, I became interested in the potential of statistical analysis and machine learning to unlock new insights and advance our knowledge in various fields. This experience helped me develop a passion for data research and motivated me to continue pursuing this path.

YZC: Solving interesting problems always excites me. As an experienced scholar in the field of bioinformatics and biomedical AI, I always try to think about scientific and technological problems in a nonlinear way, and the nonlinear problems have inspired me a lot and guided my path step by step in my academic journey. My rich experiences in interacting with academic and industry experts also enable me to focus on practical problems most relevant to biomedicine and drug discovery.

What is the definition of data science in your opinion? What is a data scientist?

WXS: In my view, data science is a true science, not just a useful tool or an engineering discipline. A data scientist is a scientist who seeks to uncover the patterns, knowledge, or insights hidden within data and to use these discoveries to solve practical problems across a range of fields, such as disease diagnosis in biomedicine and the design of new bioactive molecules in drug discovery. I believe that domain knowledge is essential for a data scientist, as it can help to better interpret the results and uncover the underlying patterns and knowledge in the data. Without this

expertise, a data scientist may be misled by superficial patterns in the data or overlook important insights. I believe that data scientists will play increasingly important roles in solving complex scientific and engineering problems, such as in multidisciplinary drug discovery, where data scientists are collaborating with medicinal chemists and biologists to accelerate the drug discovery process.

YZC: Data science complements analytic and experimental science. It is fundamental to scientific research and technology development across many fields. Therefore, it is an interdisciplinary field, involving data processing, data mining, machine learning, computational statistics, and analytics. As an example, building a high-quality bioinformatics database involves multiple tasks of data collection, processing, storage, cross linking, analysis, visualization, and distribution. These tasks require the expertise of a data scientist to collaborate with biological researchers. In my opinion, a qualified data scientist should not only be able to model and analyze data but also have the professional ability to process data to obtain high-quality data.

Why did you decide to publish in *Patterns*?

WXS: I chose to publish in *Patterns* because I believe that the journal aligns well with my interests and research focus in data science. The name of the journal is particularly appealing to me, as it reflects the importance of identifying patterns in data and using these insights to solve complex problems. I was also drawn to the journal's focus on interdisciplinary research and its approach to involving multiple advisors from different fields.

I believe that publishing in *Patterns* will provide a good opportunity for my research to reach a wider audience of researchers and practitioners in the field of data science. The journal's focus on interdisciplinary research and its commitment to involving multiple perspectives in the review process make it an ideal platform for sharing my work and engaging with others who are interested in similar topics. Overall, I believe that publishing in *Patterns* will be a valuable opportunity to share my research and contribute to the field of data science.

YZC: *Patterns* bridges data science solutions to diverse disciplines. We wish to contribute to the development and exploration of this bridge for the advancement of scientific research in biology, medicine, chemistry, and physics and technology development in healthcare, drug discovery, and materials.

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

WXS: As a data/research scientist, I feel it's important to stay up to date, especially in my field where new concepts and methodologies are always emerging. To keep up to date with these advancements, I pay attention to the media, conferences, and latest online papers of journals from time to time. For example, Twitter is a great tool for me to find interesting research and reports; ArXiv is great for me to read the latest research as preprint articles are published in different domains. Also, I often browse GitHub to explore some interesting repositories and discover new research content, then share and discuss the findings with my colleagues.

Which of the current trends in data science seem most interesting to you? In your opinion, what are the most pressing questions for the data science community?

WXS: The representation of complex data is the most interesting problem to me. Data representation is critical for learning and exploring the data and for the development of high-performance models. In the past, the 1D vectors with arbitrary orders are the inputs of conventional machine learning models or fully connected neural networks; however, if we change the order of the input vectors to retrain the model, there is nothing to change with their model performance. This intuitive result tells us that when the data are fed to a model, the order information of the data is ignored. Therefore, the attention mechanisms of a model have emerged in recent years. However, instead of focusing on the context awareness of a model, we try to improve the data representation by exploring the intrinsic feature relationships. Recently, we tried to construct feature relations from the perspective of information theory: using an unsupervised approach to map unordered feature points (data chaos) into a 2D feature map with patterns to enhance the learning efficiency of data.[1,2] The effectiveness of machine learning depends on both the model and the data, and although the model is important, I would love to see a diversity of the data level exploration in the data science community.

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?

WXS: I think data science in bioinformatics and cheminformatics has a very important role. In my research field, from biomedical data ETL (extract, transform, and load) to data visualization, modeling, and analysis, data science is everywhere. For example, in metagenomic studies, data science can help us to do taxonomic profiling based on microbiota-sequencing data. Also, the statistical analysis of metagenomics data can help to identify potential diagnostic biomarkers and find the host-microbial associations. The important microbes identified may be of potential medical or therapeutic value, as a target or a key biomarker for disease diagnosis. In addition, modeling metagenomic data helps us to establish non-invasive diagnostic models in the clinic to detect diseases such as colorectal cancer. In the cheminformatics domain, data science can help us to expand the chemical space of small molecules, to discover bioactive molecules with good potency against specific drug targets. Data science has also been involved in the chemical reaction predictions, to help to establish the digital synthesis pipelines. In the next 2–3 years, I believe there will be a boom in data-driven drug molecule generation in cheminformatics. With the establishment of the AlphaFold Protein Structure Database, structure-based molecular generation will also be further developed.

How did this project you wrote about come to be?

WXS: One of the objectives of my PhD study is to develop new methods to analyze high-dimensional -omics data. But before that, we devised a question about how to learn the unordered data efficiently; for example, we know that convolutional neural networks (ConvNets) are very good at learning structured

Image 1. Professor Chen (left) and Dr. Shen (right) of the bioinformatics and drug design group

image data, but what if the ordered structures of image data are randomized or destroyed? Shuffling the pixels of the images can make the data appear chaotic and remove any existing structure or patterns. This will make it more difficult for the ConvNet model to learn from the data and recognize patterns or features that are relevant to the task. So how to improve the ConvNet model performance even if the data structure or patterns were destroyed? We then tried to introduce a new approach for finding the underlying structure or patterns of the chaotic data that is not relevant to the task; in that case our approach was able to enhance the learning efficiency on the unordered chaotic data. Now back to the high-dimensional biomedical data, we can also regard it as disordered chaotic data. After that, I worked closely with my advisor to design the research and further develop methodology and to analyze the high-dimensional -omics data and finish the project.

Was there a particular result that surprised you, or did you have a eureka moment? How did you react?

WXS: Yes, there was a particular result in the paper that surprised me. It was a finding that the grouping-based multi-channel operation in the metagenomic data restructuring can significantly boost the model performance, and it was completely unexpected. At first, I was skeptical of the result and thought that it might be an error or a statistical fluke. But as I dug deeper into the data and reran the experiments, the finding held up and became more and more clear. I remember feeling a rush of excitement, as in theory, the grouping-based multi-channel operation may make the learning effect of the model worse, because it potentially increases the risk of overfitting. It felt like a true eureka moment, and I couldn't wait to share it with my advisor. I reacted by immediately discussing the finding with my advisor for how to incorporate it into the paper and further explore its implications. Finally, we found that both phenotype-based and genotype-based grouping of microbes in metagenomic data can significantly improve the performance of the disease prediction model. Overall, the surprising results motivated me to continue exploring metagenomic data science and discovering more interesting results.

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

WXS: The next step is to develop the graph representations of the metagenomic data and use a graph neural network (GNN) to learn the representation, because this approach has the potential to overcome the limitations of the current MEGMA 2D representation. By using a non-Euclidean data representation, such as a weighted topological graph, the spatial correlations between feature points can be more accurately preserved. This can provide a more faithful representation of the data and can potentially improve the learning performance of the model. Additionally, using a GNN to learn the representation can provide a more flexible and powerful learning framework, as compared to using a ConvNet-based deep learning model. Overall, developing the graph representations of the metagenomic data and using a GNN to learn the representation is a promising direction for the project and has the potential to improve the performance and capabilities of the model.

As a graduate student amidst the COVID-19 pandemic, I will continue to explore the fields of data science and health and work to advance these fields as a full-time researcher. This will involve seeking out opportunities for postdoctoral fellowships or faculty positions at research institutions, where I can continue to conduct research and advance my career.

REFERENCES

1. Shen, W.X., Liu, Y., Chen, Y., Zeng, X., Tan, Y., Jiang, Y.Y., and Chen, Y. (2022). AggMapNet: enhanced and explainable low-sample omics deep learning with feature-aggregated multi-channel networks. Nucleic Acids Res. *50*, e45. https://doi.org/10.1093/nar/gkac010.
2. Shen, W.X., Liang, S.R., Jiang, Y.Y., and Chen, Y. (2023). Enhanced metagenomic deep learning for disease prediction and consistent signature recognition by restructured microbiome 2D representations. Patterns *4*, 100658. https://doi.org/10.1016/j.patter.2022.100658.

W.X. Shen is currently a postdoctoral researcher in A/Prof. Wujie's Lab in the Department of Chemistry at National University of Singapore (NUS). He received his PhD in bioinformatics from NUS Pharmacy Department in 2022, during which he joined the Bioinformatics and Drug Design (BIDD) group at NUS and focused on the new methods for -omics and pharmaceutical deep learning. Prior to that, he received double BS (2013) degrees in field of life sciences and economics from Northwest University in China, and then received his MS (2016) degree in field of cheminformatics from Tsinghua University, China.

Y.Z. Chen is currently a PI in Shenzhen Bay Lab and Tsinghua Shenzhen International Graduate School, Tsinghua University**.** He received his PhD in physics from University of Manchester, Institute of Science and Technology, UK, in 1989. He also spent 4 years as a postdoc fellow at Biophysics Group, Dept of Phys, Purdue University, Indiana, USA. He then worked as a full-time professor in NUS for over than 20 years. His research in artificial intelligence intersects with topics in herb pair, machine learning, folk medicine, and pattern recognition. His drug discovery research incorporates themes from drug development, target database, druggability, *in silico*, and web browser. His work carried out in the field of bioinformatics brings together such families of science as pharmaceutical sciences, KEGG, peptide binding, epitope, and drug.