Chemical Science

ROYAL SOCIETY OF CHEMISTRY

COMMENTARY

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2023, **14**, 10628

All publication charges for this article have been paid for by the Royal Society of Chemistry

DOI: 10.1039/d3sc90185h

rsc.li/chemical-science

A focus on harnessing big data and artificial intelligence: revolutionizing drug discovery from traditional Chinese medicine sources

Mingyu Li[a] and Jian Zhang*[ab]

The advent of big data-driven artificial intelligence (AI) modeling has profoundly impacted the realm of drug discovery. Chen *et al.* (Q. Lv *et al.*, *Chem. Sci.*, 2023, https://doi.org/10.1039/D3SC02139D) have paved a way for modern drug discovery from traditional Chinese medicine (TCM) sources through their efforts over the past decade. They achieved this by creating TCMBank, the most extensive systematic central resource for TCM, which integrates standardized TCM-related big data and streamlines the AI-based drug discovery process.

In the era of big data, data-driven AI modeling has revolutionized drug discovery, transitioning from serendipitous screening to rational design.[1] The utilization of big data and AI in TCM stands as a quintessential example of this transformation, revitalizing the field and continuing to provide a reliable, abundant source for the development of modern pharmaceuticals.[2]

Historically, a significant proportion of medicines were derived from natural herbs. Upon gaining a comprehensive understanding of these herbs' effects, scientists dedicated substantial time and financial resources to high-throughput screening for the active ingredients responsible for their efficacy. Although progress has been slow in recent years, thousands of years of folk practice have explored a vast number of TCM. Over the past several decades, considerable efforts have focused on isolating active ingredients from TCM and investigating their potential targets, culminating in an enormous and intricate repository of TCM-related big data.[3]

The emergence of big data has opened up new possibilities for the modernization of TCM. However, there are several challenges to efficiently leveraging this data, collectively referred to as the "four Vs": velocity, volume, variety and veracity. In terms of velocity, TCM-related information is expanding rapidly, and manual collection considerably lags behind the speed of data generation. Most existing TCM databases suffer from limited data volume, lack of data variety, and slow data velocity, with some not even being updated. This situation necessitates the use of advanced techniques to process data in near-real-time and effectively manage the continuous flow of information. Furthermore, traditional computational modeling methods for drug discovery may not be suitable for handling the vast amount (volume) and diverse types (variety) of data. In particular, when dealing with complex compound herbs and underlying biological mechanisms, the uncertainty (veracity) of the resulting data increases significantly.[4] These challenges call for the development of innovative computational modeling methods to handle and analyze big data.

AI represents a feasible solution to these challenges, primarily due to its robust capacity to automatically capture underlying patterns within existing big data and use the patterns to predict new data.[5] Data-driven modeling is essential for AI performance. This means that the size and quality of the training dataset heavily impact the accuracy of the models, with larger and higher quality datasets typically resulting in more accurate models. Moreover, a single model is more prone to overfitting. It may be too sensitive to specific information in the training set, leading to decreased prediction accuracy and difficulty in generalizing new, unseen data. To combat these issues, ensemble learning (EL) models are developed by combining multiple individual models to achieve better predictive performance and generalization ability.[6]

Therefore, at the core of TCM modernization is the collection of standardized TCM-related big data and the utilization of powerful artificial intelligence techniques that enable innovative modeling tailored to handle heterogeneous big data. Calvin Yu-Chian Chen and co-workers from Sun Yat-sen University have made remarkable strides in this field over the past decade. They have developed TCMBank, the most extensive systematic central resource for TCM (https://doi.org/10.1039/D3SC02139D).[7] This database also

[a]*State Key Laboratory of Medical Genomics, National Research Center for Translational Medicine at Shanghai, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China. E-mail: jian.zhang@sjtu.edu.cn*

[b]*Medicinal Chemistry and Bioinformatics Center, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China*

incorporates an EL-based drug discovery workflow, which assists in identifying potential lead compounds and opportunities for drug repurposing.

Notably, to facilitate efficient big data collection and processing in TCMBank, an AI-based Intelligent Document Identification Module (IDIM) is developed. This module automatically gathers TCM-related information from various sources, including books, articles and TCM-related databases. After manual validation at least twice, a comprehensive TCM network was obtained, comprising 9192 herbs, 61 966 ingredients, 15 179 targets, 32 529 diseases, and their pairwise relationships.

The key submodule of IDIM is the biased LexRank module for automatic summarization of crucial sentences and keywords. The term "biased" refers to the incorporation of prior knowledge in the initialization weights to account for the original importance differences of critical sentences or words, instead of initializing all nodes with equal weight within classical LexRank.[8] Prior knowledge for sentence summarization comes from feature fusion, where a multi-layer perceptron classifier is trained to predict the prior probability score of sentences being selected as summaries by using six predefined feature vectors. Subsequently, the entire document is converted into a graph, with nodes assigned the prior scores from the classifier. Edges are connected by the cosine similarity between nodes' feature vectors. Each node's score is updated iteratively, and summaries are generated using high-scoring sentences. Similarly, for keyword extraction, prior knowledge is incorporated with a word graph network derived from public dictionary data. Ultimately, the top k node words with higher node values are chosen as keywords. The feature fusion or prior graph-based biased LexRank has been validated for superior or comparable performance relative to other baselines on popular datasets and practical case studies. Importantly, by combining other AI techniques, such as selenium, pdfplumber, and optical character recognition, to regularly download and parse the latest PDF documents from PubChem, IDIM enables TCMBank to keep pace with the velocity of big data and continuous updates, significantly reducing labor costs.

After constructing TCM-related big data, Chen *et al.* further designed an EL-based drug discovery pipeline by combining molecular docking, EL models, molecular dynamics (MD) simulations, and experimental verification to accelerate drug discovery. The first step is to prepare the target protein sequence and structure, as well as ligand libraries from various sources. Next, Discovery Studio is used to compute and minimize the docking poses of ligands. After that, a ligand-based EL model is used for predicting the negative logarithm of their half-maximal inhibitory concentration (pIC50), which includes feature selection, 12 regression models, and a vote-average strategy. In parallel, a complex-based EL model is developed. This model encodes ligand–target pairs by integrating multiple deep neural networks to obtain embedding features, concatenating them, and finally decoding *via* fully connected layers to output affinity predictions. Resulting candidate ligands are assessed by combining docking scores, pIC50, and affinity predictions. They further utilize MD simulations and cell-based *in vitro* assays to verify the stability and functionality of ligand binding. The reliability of this EL-based drug discovery pipeline has been demonstrated by identification of potential inhibitors for colorectal cancer and Alzheimer's disease.[9,10]

Interestingly, although TCM is a part of natural products (NPs) and shares similar chemical or pharmaceutical properties, it still differs from NPs in certain aspects. Compounds in TCMBank exhibit a statistical trend of chemical properties that have longer tails (*e.g.*, an overdose of rotational bonds). Furthermore, a higher percentage of these compounds exhibit poor absorption, low solubility, and dose-dependent liver injuries, among other concerns. These observations suggest that TCM may not be intuitively friendly to the human body and should be used with caution.

Overall, the TCMBank by Chen *et al.*[7] demonstrates how big data and AI can revolutionize drug discovery from TCM sources. The acquisition of adequate, highly reliable, and issue-specific big data (TCMBank) is a significant factor in the success of AI-assisted drug discovery. AI, in the form of IDIM, enables the constant updating of up-to-date big data, while the EL-based drug discovery workflow holds the potential to significantly enhance efficiency in promoting innovative and rational drug discovery, ultimately generating more high-quality data.

Integrating big data curation and advancements in AI research creates a sustainable paradigm widely applicable in drug discovery. Likewise, our lab has successfully developed the Allosteric database since 2009, using a combination of allosteric big data and AI-driven computational tools to transform the discovery of allosteric modulators from a serendipitous process to a more systematic and rational design.[11,12] This has sparked significant pharmaceutical interest in the field. Moreover, as the large language models (*e.g.*, ChatGPT) continue to grow, incorporating them into the future of data-driven drug discovery holds potential for revolutionizing the field.[13] However, the veracity of available data remains one of the formidable challenges.[1] This is because data is heavily influenced by varying experimental conditions, especially when it comes to drugs operating within complex biological systems. In addition, more efforts are needed to increase AI modeling accuracy and robustness in diverse drug discovery settings. Meanwhile, resource sustainability issues are also becoming a concern in applying AI in drug discovery.[14,15] Therefore, a revolution is needed not only in data utilization but also in methodological design.

Author contributions

J. Z. conceived the project. M. L. and J. Z. wrote the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Generous financial support by grants from the National Natural Science

Foundation of China (No. 81925034) is gratefully acknowledged.

References

1 X. Yang, Y. F. Wang, R. Byrne, G. Schneider and S. Y. Yang, Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery, *Chem. Rev.*, 2019, **119**, 10520–10594.

2 F. I. Saldivar-Gonzalez, V. D. Aldas-Bulos, J. L. Medina-Franco and F. Plisson, Natural product drug discovery in the artificial intelligence era, *Chem. Sci.*, 2022, **13**, 1526–1546.

3 L. Zhang, J. K. Song, L. L. Kong, T. Y. Yuan, W. Li, W. Zhang, B. Y. Hou, Y. Lu and G. H. Du, The strategies and techniques of drug discovery from natural products, *Pharmacol. Ther.*, 2020, **216**, 107686.

4 S. Shilo, H. Rossman and E. Segal, Axes of a revolution: challenges and promises of big data in healthcare, *Nat. Med.*, 2020, **26**, 29–38.

5 H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, A. Anandkumar, K. Bergen, C. P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T.-Y. Liu, A. Manrai, D. Marks, B. Ramsundar, L. Song, J. Sun, J. Tang, P. Veličković, M. Welling, L. Zhang, C. W. Coley, Y. Bengio and M. Zitnik, Scientific discovery in the age of artificial intelligence, *Nature*, 2023, **620**, 47–60.

6 J. Q. Chen, H. Y. Chen, W. J. Dai, Q. J. Lv and C. Y. C. Chen, Artificial Intelligence Approach to Find Lead Compounds for Treating Tumors, *J. Phys. Chem. Lett.*, 2019, **10**, 4382–4400.

7 Q. J. Lv, G. X. Chen, H. H. He, Z. D. Yang, L. Zhao, H. Y. Chen and C. Y. C. Chen, TCMBank: bridges between the largest herbal medicines, chemical ingredients, target proteins, and associated diseases with intelligence text mining, *Chem. Sci.*, 2023, DOI: **10.1039/d3sc02139d**.

8 G. Erkan and D. R. Radev, LexRank: Graph-based lexical centrality as salience in text summarization, *J. Artif. Intell. Res.*, 2004, **22**, 457–479.

9 H. Y. Chen, J. Q. Chen, J. Y. Li, H. J. Huang, X. Chen, H. Y. Zhang and C. Y. C. Chen, Deep Learning and Random Forest Approach for Finding the Optimal Traditional Chinese Medicine Formula for Treatment of Alzheimer's Disease, *J. Chem. Inf. Model.*, 2019, **59**, 1605–1623.

10 G. X. Chen, X. F. Jiang, Q. J. Lv, X. J. Tan, Z. H. Yang and C. Y. C. Chen, VAERHNN: Voting-averaged ensemble regression and hybrid neural network to investigate potent leads against colorectal cancer, *Knowl.-Based Syst.*, 2022, **257**, 109925.

11 X. Liu, S. Lu, K. Song, Q. Shen, D. Ni, Q. Li, X. He, H. Zhang, Q. Wang, Y. Chen, X. Li, J. Wu, C. Sheng, G. Chen, Y. Liu, X. Lu and J. Zhang, Unraveling allosteric landscapes of allosterome with ASD, *Nucleic Acids Res.*, 2020, **48**, D394–D401.

12 D. Ni, Z. T. Chai, Y. Wang, M. Y. Li, Z. T. Yu, Y. Q. Liu, S. Y. Lu and J. Zhang, Along the allostery stream: recent advances in computational methods for allosteric drug discovery, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2022, **12**, e1585.

13 A. M. Bran, S. Cox, A. D. White and P. Schwaller, ChemCrow: augmenting large-language models with chemistry tools, *arXiv*, 2023, preprint, arXiv:2304.05376, DOI: **10.48550/arXiv.2304.05376**.

14 Z. Jia, J. Chen, X. Xu, J. Kheir, J. Hu, H. Xiao, S. Peng, X. S. Hu, D. Chen and Y. Shi, The importance of resource awareness in artificial intelligence for healthcare, *Nat. Mach. Intell.*, 2023, **5**, 687–698.

15 D. Probst, Aiming beyond slight increases in accuracy, *Nat. Rev. Chem.*, 2023, **7**, 227–228.