frontiers | Frontiers in Endocrinology

TYPE Editorial
PUBLISHED 05 May 2023
DOI 10.3389/fendo.2023.1180254

OPEN ACCESS

EDITED AND REVIEWED BY
Tom Michoel,
University of Bergen, Norway

*CORRESPONDENCE
Claudio Angione
c.angione@tees.ac.uk

RECEIVED 05 March 2023
ACCEPTED 18 April 2023
PUBLISHED 05 May 2023

CITATION
Angione C, Wang H and Burtt N (2023) Editorial: Artificial intelligence for data discovery and reuse in endocrinology and metabolism.
Front. Endocrinol. 14:1180254.
doi: 10.3389/fendo.2023.1180254

Editorial: Artificial intelligence for data discovery and reuse in endocrinology and metabolism

Claudio Angione[1,2,3]*, Huajin Wang[4] and Noël Burtt[5]

[1]School of Computing, Engineering and Digital Technologies, Teesside University, North Yorkshire, United Kingdom, [2]Centre for Digital Innovation, Teesside University, North Yorkshire, United Kingdom, [3]National Horizons Centre, Teesside University, North Yorkshire, United Kingdom, [4]Open Science & Data Collaborations, Carnegie Mellon University, Pittsburgh, PA, United States, [5]Medical and Population Genetics Program and Metabolism Program, The Broad Institute of MIT and Harvard, Cambridge, MA, United States

KEYWORDS

multi-modal, systems biology, multi-omic integration, mechanism & characterization, machine learning

Editorial on the Research Topic
Artificial intelligence for data discovery and reuse in endocrinology and metabolism

Introduction

As biomedical research has embraced the era of big data, massive amounts of complex multi-omic data are being generated. While there is huge potential in using the rich body of data to make new discoveries, many challenges exist in the dissemination, discovery, and reuse of these data. Artificial Intelligence (AI) and machine learning (ML) technologies have been paramount towards fully extracting value from rich and complex datasets to drive scientific discoveries and clinical decision-making. However, the adoption of AI and ML in endocrinology and metabolic diseases is lagging behind, compared to fields such as cancer genomics (1).

A major part of the challenge comes from the complexity and heterogeneity of data being produced by different omics platforms and research groups. In addition, there is a lack of data standards, data exchange platforms and data processing pipelines that are widely accepted by the community. Due to the complex and multi-faceted mechanisms underlying directly observable phenotypes, identifying multi-omic biomarkers that reflect the interplay between genetic regulation and metabolic response could provide novel insights into cellular functionality. Recent surveys have shown the role that metabolomic profiling plays in increasing the power of clinical variables, but have also highlighted its open challenges (2).

To fully leverage the power of AI to maximize the value of the rich data in endocrinology and metabolism, at least a few key areas need to be addressed. First, aggregated, harmonized, discoverable, and accessible datasets that are suitable for ML and AI applications are in urgent need, especially in the presence of multi-modal data. To this end, developing data-sharing infrastructure, standards, and data curation pipelines is the key to success. Second, it is becoming clear that incorporating mechanistic knowledge into

ML and AI tools will facilitate a biologically-informed interpretation of the predictions. Third, it is crucial to develop easy-to-use tools and visualization methods that can be used by researchers and clinicians not trained as computer scientists to drive scientific discoveries and clinical decisions.

Learning with multiple omic modalities

The articles presented in this Research Topic tackle the issue of learning with multiple omic modalities in a clinical context. Overall, they emphasise that the combination of multiple modalities is more effective than using only one modality in isolation, showing a significant increase in predictive performance. They also address the issue of small sample size, a common drawback of ML studies in omics, where obtaining matched samples across more than one modality remains a challenging task in terms of time and costs involved, with the associated challenges in using deep learning approaches (3, 4).

Feng et al. implement and compare eight ML models for the prediction of lateral lymph node metastasis in patients with papillary thyroid carcinoma, showing that random forest is highly effective and interpretable as a predictive method, but its performance is highly dependent on clinical variables. They also show that combining different modalities (clinical and sonographic in this case) improves the predictive performance.

Wu and Zhang apply various bioinformatics methods to identify differentially expressed genes, hub genes and signalling pathways that are potentially important for type 2 diabetes, using data from blood samples of subjects with type 2 diabetes vs healthy controls, downloaded from the GEO database. Further, a pharmacophore target analysis reveals potential drug target genes and pathways for celastrol, a natural phytochemical found to have anti-diabetic properties. The molecular interaction of celastrol and target genes is simulated by AlphaFold2.

Chen et al. study drug metabolism genes differentially expressed in human liver samples with or without NAFLD. Due to the small sample size, they reanalyse two publicly available GEO datasets, as well as previously-collected experimental data from the mouse, to enrich the main dataset and identify nine common differentially expressed genes.

Fu et al. compare five ML algorithms to predict the 52-week blood glucose level in 273 patients with type-2 diabetes, assessing the algorithms in terms of clinical and numerical performance measures. They conclude that XGBoost is the best choice to assist decision-making in the treatment of diabetic patients. They also discuss the challenges introduced by learning with a relatively small sample size.

Taken together, all studies show that focusing on interpreting the predictions generated by ML is a critical topic, especially for clinical applications (5–7). In this context, introducing mechanistic models within ML architectures is likely to represent a step change compared to existing data-driven approaches.

Perspective: mechanism-aware and multi-modal machine learning

Data quality and data scarcity remain major challenges when dealing with the integration of multi-modal data through ML. Documentation, correct labelling and project metadata are as important as the data itself. Furthermore, dealing with missing data is a common issue among different data types or modalities, whether omic, imaging, or clinical data, which reduces the number of useable common samples. The limited number of matched samples in turn fuels the generalizability challenge, since a model with a limited number of samples tends to overfit the data. Transfer learning approaches can mitigate this issue.

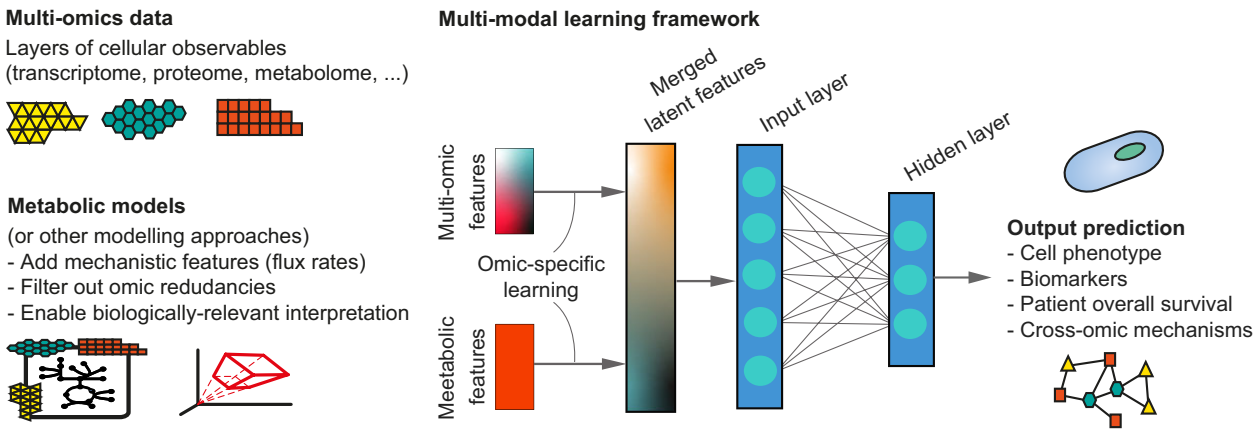

FIGURE 1
Example of a multi-modal learning framework combining omics data with mechanism-drive modelling approaches within a neural network architecture. In an intermediate integration strategy, multi-omics features and mechanistic model-derived features can be used for an independent phase of modality-specific training, before being fused in a single set of latent features. These can then be used within existing ML architectures (e.g. a fully connected neural network) to achieve the final prediction.

It is also important to note that, in complex phenotypes where the interaction between events spanning multiple omic layers is likely to be the main cause of disease progression, traditional multi-omic computational methods based on ML are only able to uncover associations among genes, proteins or other omic components, without offering a mechanistic interpretation. As a result, these methods are not always able to provide the holistic understanding necessary to provide actionable biomarkers.

Therefore, new hybrid computational methodologies that are both data- and model-driven are needed for novel biomarker discovery, early diagnosis and better prediction of therapeutic targets (8, 9). For instance, multi-modal approaches to integrate multi-omic data with metabolic modelling (Figure 1) have shown promising results with higher accuracy and increased attention for the biological interpretation of ML-derived results (10–12). It seems therefore likely that combining different types of omics data with mechanism-driven models will further improve the ability of ML models to mechanistically characterize a disease.

Another potential direction is the direct incorporation of biological information within the learning process. This could be done by manually changing the structure of the ML architecture, or by adopting a combination of omics depending on the patient's clinical characteristics, e.g. introducing an attention mechanism within the neural network (13). Biomarkers extracted from biologically-informed architectures are likely to have significantly higher potential for survival prognosis and therapeutic role compared to those generated *via* traditional model-agnostic interpretations.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

CA would like to acknowledge The Alan Turing Institute for a Network Development Award, grant TNDC2-100022, and Turing Network Funding, grant D-ELA-013.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Editorial. Why the metabolism field risks missing out on the AI revolution. *Nat Metab* (2019) 1(10):929–30. doi: 10.1038/s42255-019-0133-9

2. Buergel T, Steinfeldt J, Ruyoga G, Pietzner M, Bizzarri D, Vojinovic D, et al. Metabolomic profiles predict individual multidisease outcomes. *Nat Med* (2022) 28 (11):2309–20. doi: 10.1038/s41591-022-01980-3

3. Doan LMT, Angione C, Occhipinti A. Machine learning methods for survival analysis with clinical and transcriptomics data of breast cancer. In: *Computational biology and machine learning for metabolic engineering and synthetic biology*. New York, NY: Springer US (2022). p. 325–93. doi: 10.1007/978-1-0716-2617-7_16

4. Steyaert S, Pizurica M, Nagaraj D, Khandelwal P, Hernandez-Boussard T, Gentles AJ, et al. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nat Mach Intell* (2023) 5:351–62. doi: 10.1038/s42256-023-00633-5

5. Alsinglawi B, Alshari O, Alorjani M, Mubin O, Alnajjar F, Novoa M, et al. An explainable machine learning framework for lung cancer hospital length of stay prediction. *Sci Rep* (2022) 12(1):1–10. doi: 10.1038/s41598-021-04608-7

6. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ digital Med* (2020) 3(1):136. doi: 10.1038/s41746-020-00341-z

7. Zhang Y, Hong D, McClement D, Oladosu O, Pridham G, Slaney G. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *J Neurosci Methods* (2021) 353:109098. doi: 10.1016/j.jneumeth.2021.109098

8. Culley C, Vijayakumar S, Zampieri G, Angione C. A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. *Proc Natl Acad Sci* (2020) 117(31):18869–79. doi: 10.1073/pnas.2002959117

9. Sidak D, Schwarzerová J, Weckwerth W, Waldherr S. Interpretable machine learning methods for predictions in systems biology from omics data. *Front Mol Biosci* (2022) 9. doi: 10.3389/fmolb.2022.926623

10. Magazzù G, Zampieri G, Angione C. Clinical stratification improves the diagnostic accuracy of small omics datasets within machine learning and genome-scale metabolic modelling methods. *Comput Biol Med* (2022) 151:106244. doi: 10.1016/j.compbiomed.2022.106244

11. Lewis JE, Kemp ML. Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance. *Nat Commun* (2021) 12(1):2700. doi: 10.1038/s41467-021-22989-1

12. Chung CH, Chandrasekaran S. A flux-based machine learning model to simulate the impact of pathogen metabolic heterogeneity on drug interactions. *PNAS nexus* (2022) 1(3):pgac132. doi: 10.1093/pnasnexus/pgac132

13. Gong P, Cheng L, Zhang Z, Meng A, Li E, Chen J, et al. Multi-omics integration method based on attention deep learning network for biomedical data classification. *Comput Methods Programs Biomed* (2023) 231:107377. doi: 10.1016/j.cmpb.2023.107377