OXFORD

*Briefings in Bioinformatics*, 2024, **25(3)**, bbae213
https://doi.org/10.1093/bib/bbae213
**Opinion Article**

# Towards multi-omics synthetic data integration

Kumar Selvarajoo [1,2,3,*] and Sebastian Maurer-Stroh [1,2]

[1]Biomolecular Sequence to Function Division, BII, (A*STAR), Singapore, 138671, Republic of Singapore
[2]Synthetic Biology Translational Research Program, Yong Loo Lin School of Medicine, NUS, Singapore, 117456, Republic of Singapore
[3]School of Biological Sciences, Nanyang Technological University (NTU), Singapore 639798, Republic of Singapore
[*]Corresponding author. Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore 138671, Republic of Singapore; Synthetic Biology for Clinical and Technological Innovation (SynCTI), National University of Singapore (NUS), Singapore, 117456, Republic of Singapore; School of Biological Sciences, Nanyang Technological University (NTU), Singapore 639798, Republic of Singapore. E-mail: kumar_selvarajoo@bii.astar.edu.sg

## Abstract

**Across many scientific disciplines, the development of computational models and algorithms for generating artificial or synthetic data is gaining momentum. In biology, there is a great opportunity to explore this further as more and more big data at multi-omics level are generated recently. In this opinion, we discuss the latest trends in biological applications based on process-driven and data-driven aspects. Moving ahead, we believe these methodologies can help shape novel multi-omics-scale cellular inferences.**

***Keywords***: *synthetic data; process-driven; data-driven; machine learning; multi-omics*

Since the turn of the millennium, through multi-omics technological advances, biological and medical fields have been faced with the challenge of dealing with big data generation. Without proper data analytics, these data are enigmatic. Thus, the last two decades of research have been dedicated to understanding variability, noise and bias in omics-wide data to analyze and make sense of the resulting large volume of generally high-quality data. In certain cases, the collection of such data is not simple or straightforward. For example, clinical data contain sensitive patient information, and single-cell omics data can be costly and time consuming to generate. How can we overcome or address such issues?

In the fields of big data and artificial intelligence, synthetic data have been generated and used to investigate human behaviors and pattern recognition [1]. Synthetic data are data that are mainly generated using statistical methodologies or machine learning, i.e. artificially, rather than from actual or experimental events. It is created using algorithms, like data augmentation, and is used for a wide range of applications, including as test data for new products and tools, and for model training and validation without compromising consumer privacy [2]. It has also been shown to be inexpensive, as it reduces the number and time taken for experiments, and combines well with the real data to increase the overall number of observations. Therefore, synthetic data generation has been adopted for almost three decades across a variety of research fields, with more recent applications coming into the clinical and omics fields [3].

Synthetic data generation may be classified into two major categories: process-driven and data-driven methods [4] (Figure 1). In biology, process-driven methods can generate data based on computational or mathematical models of an underlying biochemical process, such as signal transduction or metabolic pathways [5]. Examples include dynamic or kinetic models based on ordinary differential equations, stochastic models based on the Gillespie algorithm or Monte Carlo simulations and agent-based or cellular automata modeling. Here, the models are first developed to explain an observed behavior and then subsequently used to generate simulated or synthetic data using the same model for different conditions or situations. For example, in studying TRAIL signaling for cancer resistance, a dynamic model was used to predict a novel target that significantly enhances cancer cell death [6], which was subsequently tested and validated experimentally in bulk or population cells [7] (Figure 1A). This model was then used to generate 1000 single-cell dynamic data, which are not experimentally plausible due to very small expression values [8]. Thus, the *in silico* or synthetic approach emphasizes the understanding of systems-level effects of interactions between species or agents on the system as a whole, especially in areas where experiments cannot reach.

On the other hand, data-driven methods generate synthetic data that have been trained on actual or observed data. Here, generalized linear regression models and non-linear methods, such as self-organizing maps, random forest and neural networks can be adopted. Notably, a statistical modeling procedure that learns a joint probability distribution is able to generate synthetic data fully with partial real data. More recently, deep learning methods have been used to generate two most popular types of generative AI models today: variational autoencoder (VAE) and generative adversarial network (GAN) models. These newer techniques can improve data utility by feeding models with more and more data. For biological applications, so far, synthetic data has been largely generated for single cell and spatial omics [9, 10]. For example, ACTIVA, an improved VAE model, can generate synthetic transcriptomics data utilizing data augmentation that significantly improves the classification of rare subtypes (Figure 1B, right panels) [11]. Moving forward, these models could include physics

**Kumar Selvarajoo** is a senior principal investigator at BII, heading the Computational Biology & Omics laboratory. He is also an adjunct associate professor at the NUS School of Medicine and NTU School of Biological Sciences.
**Sebastian Maurer-Stroh** is the executive director of BII and an adjunct professor at the department of biological sciences, NUS.
**Received:** April 9, 2024. **Accepted:** April 19, 2024

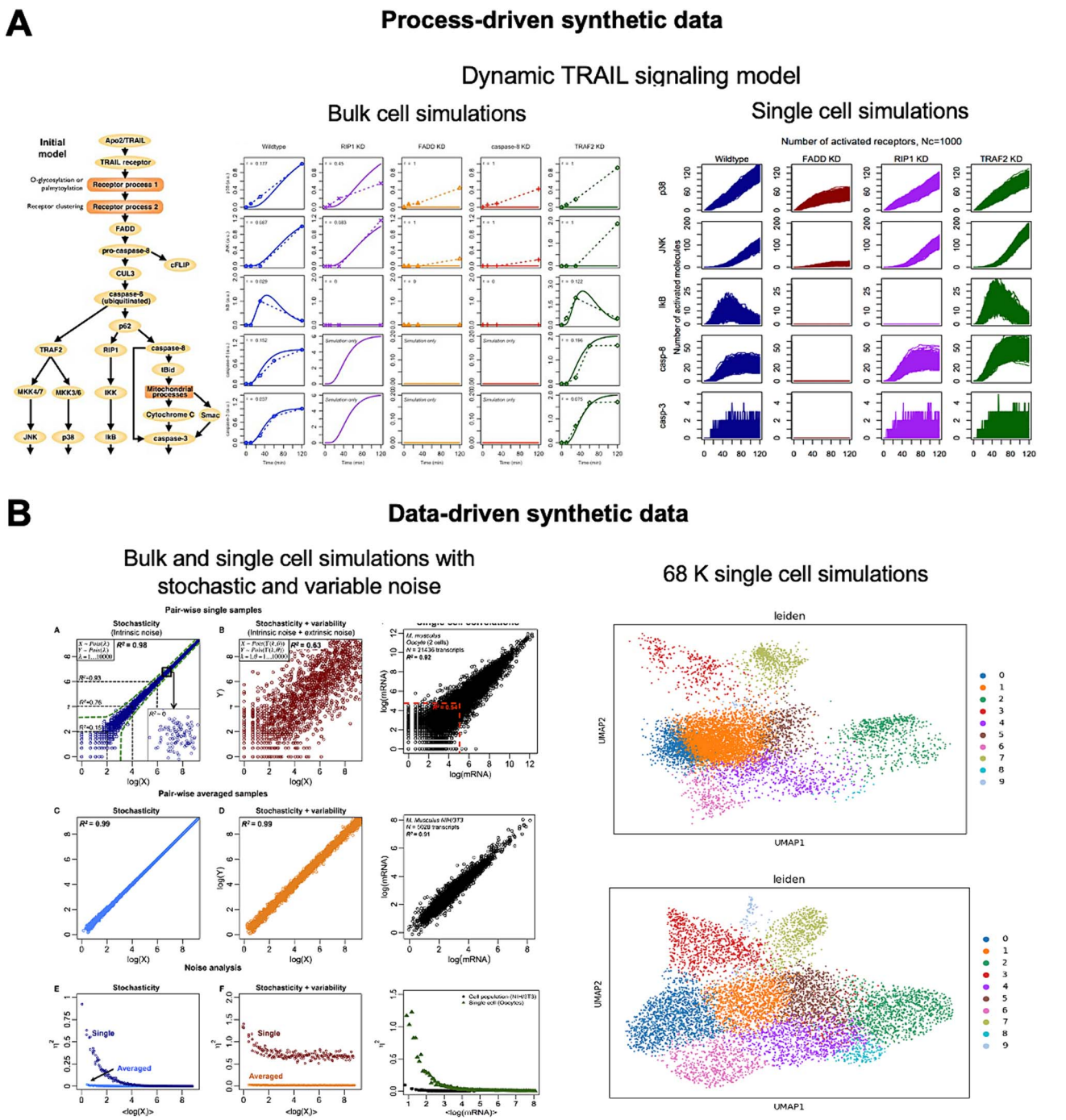


**Figure 1.** Synthetic data generation. (**A**) Process-driven. Left: for bulk cells an initial signaling model is developed using known mechanistic biochemical reactions and the corresponding experimental data for TRAIL signaling [6]. Right: the bulk validated TRAIL model is used to generate 1000 single cells data synthetically [8]. (**B**) Data-driven. Left: statistical models are used to general transcriptome-wide expression data scatter plots for bulk and single cells [10]. Right: a VAE-based model used to generate UMAP plots of 68 K single cells data, top (real) and bottom (synthetic using ACTIVA [11]).

and statistical knowledge, such as using scale-free network, power-law and lognormal distributions observed in biological data, as a means to constraint false prediction and further improve the overall machine learning outcome.

Although useful, as described, it is important to note some of the key limitations of synthetic data generation. First, in real data, we often pick up outliers of interest; synthetic data will not be able to reproduce them easily as they are usually trained to pick up general patterns of the majority data. Second, the quality of synthetic data will highly correlate with the input data, thus, thorough quality checks on the original data are necessary before the machine learning process. Third, user acceptance may be challenging since it is 'learnt' data and not everyone might see or appreciate the benefits. Fourth, in every synthetic data generation, proper skillset, time and effort are key for training and quality check evaluation, which are pinnacle to the overall predictive quality.

Despite this, synthetic data research has gained interest and momentum recently, and in the near future can be used to generate multi-omics datasets for more integrative biological

applications. Although this may result in more challenges in trusting the entire simulation, machine learning techniques can be further improved to fine-tune the comparison between multi-omics synthetic data and actual experimental data by using sophisticated feature extraction and dimension reduction methods.

**Key Points**

- Computational and machine learning models are playing key roles in biological understanding.
- Synthetic data research is relatively new in biology, and mainly focuses on single omics datasets.
- Further developments for generating multi-omics synthetic datasets for more integrative biological applications are required.

## ACKNOWLEDGEMENTS

The authors thank the Bioinformatics Institute, A*STAR, for funding and support.

## REFERENCES

1. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digit Med* 2023;**6**(1):186.
2. Servia-Rodriguez S, Wang L, Zhao J, *et al.* Privacy-Preserving Personal Model Training. In: *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI), Orlando, FL*, IEEE, NJ, USA, 2018, 153–64.
3. Toussaint PA, Leiser F, Thiebes S, *et al.* Explainable artificial intelligence for omics data: a systematic mapping study. *Brief Bioinform* 2023;**25**(1):bbad453.
4. Goncalves A, Ray P, Soper B, *et al.* Generation and evaluation of synthetic patient data. *BMC Med Res Methodol* 2020;**20**(1):108.
5. Helmy M, Smith D, Selvarajoo K. Systems biology approaches integrated with artificial intelligence for optimized metabolic engineering. *Metab Eng Commun* 2020;**11**:e00149.
6. Piras V, Hayashi K, Tomita M, Selvarajoo K. Enhancing apoptosis in TRAIL-resistant cancer cells using fundamental response rules. *Sci Rep* 2011;**1**(1):144.
7. Hayashi K, Tabata S, Piras V, *et al.* Systems biology strategy reveals PKC$\delta$ is key for sensitizing TRAIL-resistant human Fibrosarcoma. *Front Immunol* 2015;**5**:659.
8. Piras V, Hayashi K, Tomita M, Selvarajoo K. Investigation of stochasticity in TRAIL signaling cancer model. In: *2012 ICME International Conference on Complex Medical Engineering (CME), Kobe, Japan*, IEEE, NJ, USA, 2012, 609–14.
9. Erfanian N, Heydari AA, Feriz AM, *et al.* Deep learning applications in single-cell genomics and transcriptomics data analysis. *Biomed Pharmacother* 2023;**165**:115077.
10. Piras V, Tomita M, Selvarajoo K. Is central dogma a global property of cellular information flow? *Front Physiol* 2012;**3**:439.
11. Heydari AA, Davalos OA, Zhao L, *et al.* ACTIVA: realistic single-cell RNA-seq generation with automatic cell-type identification using introspective variational autoencoders. *Bioinformatics* 2022;**38**(8):2194–201.