10 Years of Variational Autoencoder: Insights from Cancer Temporal Progression Studies, a Systematic Literature Review

Guillermo Prol-Castelo[1,2], **Davide Cirillo**[1,✉], **and Alfonso Valencia**[1,3]

[1]Barcelona Supercomputing Center (BSC), C/ Jordi Girona 29, 08034, Barcelona, Spain
[2]Department of Medicine and Life Sciences, Universitat Pompeu Fabra, C/Dr Aiguader 88, Barcelona 08003, Spain
[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010, Barcelona, Spain

Deep Learning methods such as Deep Representation Learning (DRL) and, specifically, the Variational Autoencoder (VAE), have been widely used to handle the high dimensionality of available datasets. Hence, these methods have been applied to study cancer through omics data. Cancer is one of the leading causes of death worldwide, and its complex and dynamic nature makes it especially difficult to study. However, the temporal dimension of cancer progression is a target that has not been tackled. In this systematic literature review, we explore the use of DRL, particularly the VAE, in cancer studies with omics data in a temporal context. Our research reveals that the most common uses of the VAE in cancer studies are related to subtyping, diagnosis, and prognosis. Meanwhile, cancer's temporal aspect is often overlooked, namely, because of the lack of longitudinal data. We propose that applying the VAE as a generative model to study cancer in time, for example, focusing on cancer staging, will lead to meaningful advancements in our understanding of cancer.

unaligned data | longitudinal | stages | out-of-sample inference | generative

Correspondence: *davide.cirillo@bsc.es*

1. Introduction

Deep Learning (DL) is capable of capturing complex patterns inherent in health data, including molecular, imaging, and clinical data (1). Given the high-dimensionality of omics data, Deep Representation Learning (DRL) models have emerged as a way to model complex distributions in lower-dimensional representations (2). Moreover, specific DRL methodologies add a generative capability to the model (3). Such is the case of the Variational Autoencoder (VAE),

a DRL method that uses Bayesian principles through variational inference to approximate complex relationships between observations and latent variables (4). The VAE, introduced over ten years ago, consists of an encoder and a decoder; the former infers a low-dimensional representation called the latent space, while the latter learns to reconstruct the latent space into the original data. This architecture makes it particularly useful to both model a real data distribution and generate synthetic data (5).

Given the VAE is able to learn non-linear relationships in heterogeneous, high-dimensional data, its use is appropriate for the study of very complex diseases. Such is the case of cancer, the second most common cause of death worldwide, with an increasing incidence (6, 7). Well-established techniques for data collection, including omics approaches, have produced large amounts of information, also referred to as big data, about cancer (8). High-throughput molecular and imaging omics data are commonly high-dimensional, noisy, sparse, and heterogeneous. These are also highly informative, including data at different levels of granularity: patient- (9) and/or cell-level (10) resolution; information from different biological sources: genes (11), proteins (12), or metabolites (13); and even spatial information (14). Leveraging DRL methods, such as the VAE, and cancer omics, could have a translational impact by discovering clinically relevant biomarkers (15) and patient subgroups (16).

Navigating such a vast amount of information becomes even more challenging due to the fact that cancer is a dynamic process (17), adding a temporal component. However, there are two main problems when considering cancer in a temporal manner. First, cancer progression is not uniform across individuals (18, 19), meaning that even taking a snapshot of the disease at regular intervals may provide varying progression times. Second, the nature of sequencing data involves the destruction of the sample. Both of these reasons imply that any longitudinal datasets capture snapshots of unaligned samples.

Here we present a Systematic Literature Review (SLR) on the use of DRL, paying special attention to the VAE, in cancer studies with omics data in a temporal context. To do so, we consider the following research questions:

1. How has representation learning, particularly the VAE, been used to study cancer with omics data?
2. How do cancer studies take longitudinal data into account?
3. How is cancer's temporal dimension studied with deep learning?
4. Which type of omics are most commonly used in time-related cancer studies?
5. What are the key limitations and challenges in the use of DRL for cancer studies with omics data in a temporal context?

The process of the SLR is summarized in Figure 3 (see Methods for details), and our main insights are summarized in Figure 1. We found 440 papers related to our search terms and reviewed their titles and abstracts. These reveal

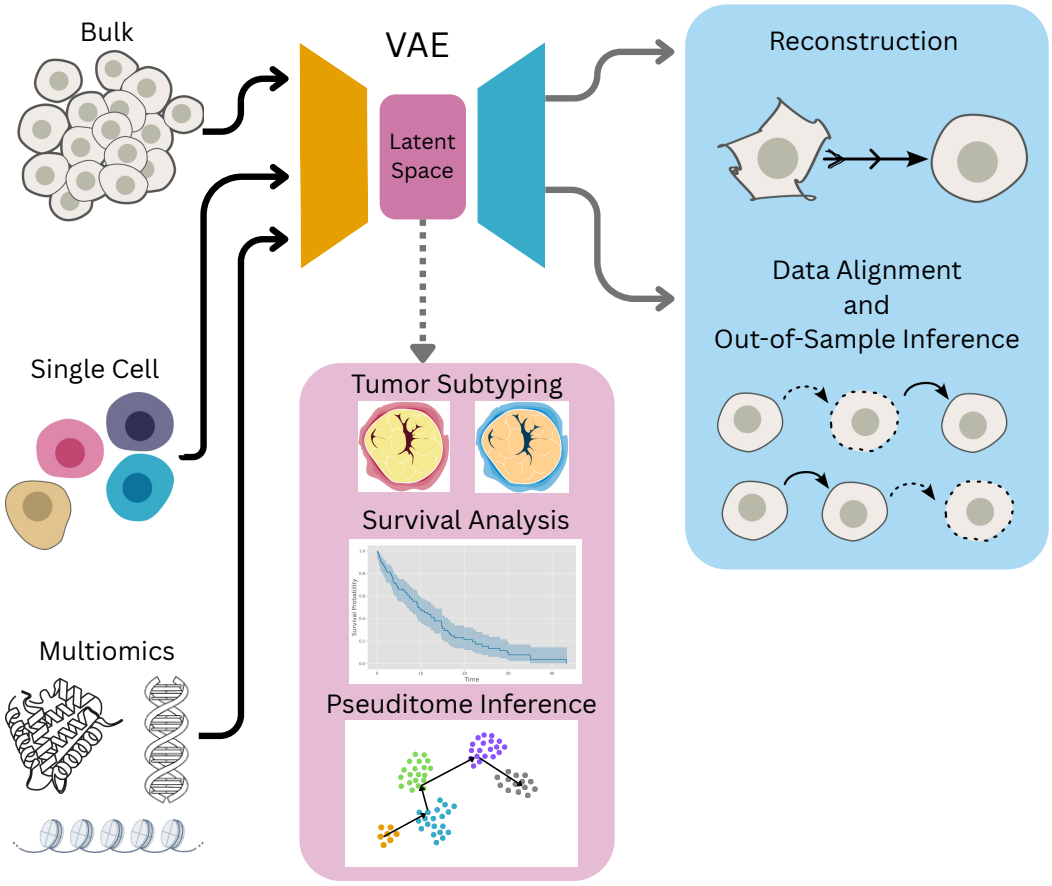

Figure 1. Summary figure of the most common methodologies involving representation learning and cancer studies. On the left side of the figure are shown the most common types of data used to train VAEs particularly—and DRL methods more generally—, in cancer studies. After being passed to the encoder, data is embeded into a lower-dimensional representation, the latent space. The latent space has been commonly used for subtyping, survival anlyses (commonly as part of prognostic analyses), and, when it comes to single-cell data, also for pseudo-time trajectories inference. Possible further applications, shown on the right-hand side of the figure, include leveraging the decoder's generative capabilities to reconstruct noisy data, and align and infer out-of-sample data, which may be used to study cancer progression in time. However, such applications of the decoder remain underexplored in biomedical research.

that the most common use of representation learning and VAEs in cancer studies are related to subtyping, diagnosis, and prognosis. From the abstracts screening, we selected 44 papers for full-text review, of which 21 were included in our final analysis. We learned that single-cell omics, with a recent surge of spatially-resolved data, is the most common type of data used in time-related cancer studies, albeit in pseudotime. There is a lack of longitudinal records in the study of cancer in time, and even the ones available contain unaligned samples. Stages have been considered as a way to provide a time unit for the study of cancer's advancement, but these usually consider different patients. Overall, there is not a clear methodology to align and study cancer through time. We propose the use of generative models, such as the VAE, could provide a way to infer out-of-sample longitudinal data that is aligned on stages.

2. Results

A. VAEs have been used to study cancer diagnosis, prognosis, and subtyping

The most common applications of Deep Representation Learning (DRL) or Deep Learning (DL) in cancer are related to diagnosis, prognosis, or subtyping (Figure 2). Studies of cancer diagnosis aim to distinguish between cancerous and non-cancerous samples (20), or between cancer types (21). Prognosis studies aim to predict the outcome of a disease, which usually involves a survival analysis (22), recurrence prediction (23), or treatment response (24). Meanwhile, subtyping studies aim to identify different subtypes, that is, groups that a certain type of cancer may be separated into (25, 26). All three applications are crucial for developing more personalized treatments that would improve the outcome of the disease.

All of these and other cancer studies rely on the use of high-throughput omics data, which yield information at varying levels of resolution. Bulk data has been assessed through VAEs, providing a way to classify different cancers from TCGA in a reduced dimensional space that is biologically relevant and interpretable (15). Classically, the most accessible data has been bulk data, where the sampled unit is a whole tissue or group of cells. However, the use of single-cell data has been increasing, as it provides more detailed resolution, sampling information from individual cells. For instance, VAEs have been applied to single cell data to detect copy number variations with the objective of probing intratumor heterogeneity (27). Moreover, spatially resolved omics data has recently surged in popularity (28), additionally revealing the spatial distribution of cells in a tissue. Precisely (28) developed an AE-based methodology capable of distinguishing healthy tissue and different breast cancer types. Omics data also vary in the type of information they provide: transcriptomics data reveal gene expression levels, proteomics data reveal protein levels, and metabolomics data reveal metabolite levels. Besides, omics may be studied in combination, known as multiomics data (29, 22). Namely, (16) used an AE to link multi-omic features to the varying survival outcomes in hepatocellular carcinoma. Recent studies have focused on multiomics and single-cell data, with a recent emphasis on spatially resolved data, given they provide very detailed information and, as such, are expected to reveal new insights into cancer biology. (30) used multi-omics, single-cell data to discern rare cell populations in melanoma. The data included in these studies may go beyond molecular data, including imaging data such as radiomics (31) and histological (32) data. Histological images and DRL have been used to cluster the different areas of lung cancer whole-slide images in order to improve diagnosis (33). All in all, it is common to find applications of DRL in general and VAEs in particular to study diagnosis, prognosis, and subtyping of cancers at different levels of resolution, without any particular emphasis on temporal information.

B. Longitudinal data is used to study diseases, but not so much in cancer

The data referenced in the previous section tends to be captured at a single time point. Here we distinguish between primary and metastatic cancer, which show different growth and behavior patterns, and focus on the former. Given

the dynamic nature of primary cancer, longitudinal data is crucial to understand its progression. However, real-patient data is hard to follow up due to issues such as retention difficulties, measurement timing, resource limitations, and ethical considerations. Besides, the process of sequencing tumor samples is also destructive. Both factors render observing the course of cancer in real time not possible, neither at the patient nor the cellular level. Even when the data is collected in a longitudinal manner, cancer progression differs from patient to patient, so the data is unaligned. Possibly for these reasons we have observed the use of longitudinal data in primary cancer studies is scarce.

We found a singular case (34), where extracellular vesicle RNA (evRNA) data from liquid biopsies of colorectal cancer were collected from three cohorts: no recurrence, recurrence, and stable disease. The authors used a DL model to classify temporal subtypes from GSEA pathways. They were able to find patients that changed subtype during follow-up from the liquid biopsy data, ahead of imaging-based diagnosis. This finding confirms an extra layer of complexity in the study of cancer progression: cancer subtypes are time-dependent and may be observed through molecular data.

A study following up on cancer patients is (35), where the authors used electronic health record (EHR) data of patients with pancreatic cancer. They used a model based on gated recurrence units (GRUs) and Transformers to embed event features, encode trajectories in a latent space, and predict a time-dependent risk of occurrence. The study shows how representation learning enables out-of sample predictions when longitudinal, aligned data is available.

The use of longitudinal data and its study with DL is more common in other diseases. During our review, we found examples of longitudinal studies on Cardiovascular Disease (CVD), Alzheimer's Disease (AD), kidney function deterioration, and the effects of radiation. In (36), the authors trained a recurrent neural network on radiomics images of the brain and clinical assessments to learn a representation for binary classification of cognitive decline in AD. (37) used a LSTM-VAE trained on multi-omics (proteomics and metabolomics) to embed time-dependent data obtained during cardiac remodeling. The embeddings were then used to cluster temporal growth patterns, determined to be biologically relevant through KEGG pathway analysis. Time-dependent microarray data in (38) was collected at different time points pre- and post-myocardial infarction to predict the risk of heart failure. They used a ML model to combine interactomic, molecular, and clinical data, used alongside GSEA to discover biomarkers of progress toward myocardial infarction. Another study, (39), assessed progression of kidney function deterioration with histomorphometry data, applying DL for semantic segmentation. The authors used trajectory and pseudo-time unsupervised analysis, adapted from single-cell methodologies, to elucidate patterns in disease progression towards kidney failure. Pseudo-time trajectories recapitulated the patterns of damage observed in histology. The authors of (40) collected longitudinal data of blood samples and radiomics from mice lungs exposed to radiation, using DL for survival analyses from the radiomics data. Altogether, these studies highlight the importance of DL and DRL in the study of disease progression, and the potential of longitudinal data to provide insights into the evolution

of diseases.

The study of cancer evolution with DL and DRL is common in phylogeny, albeit mostly to discern subclones. A 2020 study (41) claimed to be the first to apply DL to study tumor evolution. Specifically, Reinforcement Learning was used to infer the most likely tumor phylogeny, and identify tumor subclones. In (42), the authors used unsupervised learning to learn graph embeddings of relationships between subclones. Hence, phylogeny studies of cancer subtypes offer a way to sort subtypes in terms of evolutionary terms. However, while phylogenetic data and studies reflect the evolutionary history of tumors and the subclone ordering, they lack an explicit temporal resolution the same way that clinically-annotated stages or time-resolved sampling do.

C. Stages as a way to study cancer in a longitudinal manner

Thus far, we have considered actual time as the unit for the study of cancer progression. However, it is rather unfeasible to collect data in large proportions, from different patients, at regular intervals. Besides, tumor evolves differently from patient to patient, so the rate of progression may not be identical for everyone. In this context, stages may be considered as a possible time-unit to study cancer progression. Stages refer to the extent of cancer (43, 44). The most common staging system is the TNM (Tumor, Nodes, and Metastasis), which takes into account the size of the tumor, its spread to nearby lymph nodes, and whether it has metastasized. Thus, there are five stages: stage 0 describes a non-invasive tumor, or carcinoma in situ; stages I through III denote an invasive tumor, the more invasive, the larger the number; while stage IV refers to a metastatic tumor that has spread to distant body parts. Hence, stages provide a standardized and aligned way to monitor the progression of cancer, and may be used as a proxy to study cancer in a longitudinally aligned manner.

The results from our search highlight a common trend: the difficulty of differentiating cancer stages. A way to take them in consideration, done in (45), is to compare normal tissue against the different stages. Authors used adversarial networks to transform a source sample, normal tissue from TCGA and GTEx, into the feature space of a target sample, the different stages, finding several up- and down-regulated genes across stages. Another way to circumvent this issue is to group stages together (46, 47, 48, 49). In (49), data from TCGA was used to classify either stages I and II or stages II and III across different cancers. The classification was compared when using real data and different DRL techniques, namely PCA, Stacked Denoising AE, and VAE. The authors found that representation learning techniques did not improve classification performances. Authors of (46) grouped stages I and II as an “early” stage, and stages III and IV as a “late” stage from papillary renal cancer data. Stages may be simplified even further to distinguish between normal tissue, non-invasive abnormally-grown tissue, and invasive cancer. (48) used DL To distinguish between normal, neoplasm, and cancer samples from cervical cancer microarray data. Besides molecular data, histological images may also be used to classify stages. As shown by (47), a Convolutional VAE was used to learn embeddings of cell states from chromatin images of breast cancer. The authors considered three stages: non-tumor, Ductal Carcinoma In Situ (DCIS), and Invasive Ductal Carcinoma (IDC). The latent embeddings

were used to infer a pseudo-time ordering of the stages with Partition-based Graph Abstraction (PAGA), a single-cell trajectory inference method, revealing an ordering from non-tumor to tumor progression. Hence, there is a growing interest in the correct characterization of stages, but its hard separability has led to simplifications in the stages considered.

The lack of separability between stages may be due to the usually-unbalanced data available across different stages. In this context, generative modeling may be useful, creating synthetic instances in order to balance out the data. Authors of (50) argued that the use of the Conditional Variational Autoencoder (CVAE) may be used to generate the stages of breast cancer in mRNA and miRNA data from TCGA. They trained the CVAE on the molecular data and used labels of the different stages: I to IV and took solid tissue normal as pseudo-stage 0. The CVAE is then able to generate new, synthetic samples at the trained labels (here, stages). These synthetic stages then showed high separability in a hierarchical clustering analysis. However, the trained CVAE model loss was not reported, and the huge difference between the separability of the real and synthetic data suggests the model may not be properly learning the actual data distribution. When using synthetic data, it is crucial to ensure the model is learning the original data distribution correctly, that the synthetic data is representative of the real data, and to validate the generated data.

D. Single cell omics data is commonly used to infer pseudo-time trajectories of cancer cells

Single-cell data contains more information than bulk data, as it processes individual cells instead of obtaining an average from a tissue sample. In terms of disease progression, single-cell data can more accurately capture the individual differences between cells within the same tissue. Hence, single-cell data analyses open the door to calculating the difference between cells, to then organize them in terms of evolutionary distance. This calculation infers a differentiation trajectory, organizing the cells in a pseudo-time axis, and several methods have been developed in order to do so, such as Monocle (51) and PAGA (52). During our search, we found that single-cell data was more commonly used in terms of trajectory inference than bulk or imaging data. This data was often processed with DRL techniques to generate lower-dimensional embeddings, used instead of the original data to infer pseudo-time trajectories. The VAE was used by (53) to embed data from Glioblastoma and Oligodendroglioma, and these embeddings were used to infer lineage and differentiation trajectories. As such, single-cell data may be seen as a proxy to study the tumor pseudo-time progression.

We also found that most of the results from our search that used single-cell data included Spatial Transcriptomics (ST) data, i.e., how the cells are distributed in the tissue. Besides pseudo-time inference, spatial data also allows to link the different tissue layers. In (54), authors used a Contrast Learning (CL) method to learn latent embeddings of cells and their spots from ST data of the mouse brain. The embeddings were then used for pseudo-time and trajectory inference with PAGA, organizing the different tissue layers. CL was also used by (55) on the mouse visual cortex, combining ST with histological images to learn embeddings of similar spots across tissue regions.

PAGA and Monocle were used for pseudo-time trajectory inference, indicating the progression in the developmental process from the white matter layer towards outer layers. (56) used a VAE to embed a joint representation of spatial proteomics and metabolomics data from B cells in the tonsils, and then used these embeddings for metabolite state clustering. The trajectory inference was able to reconstruct the cell differentiation process of B cells in a spatial domain. (57) used a graph attention autoencoders and multiview collaborative learning to embed ST and histological data from human brain slices. The embeddings were used to identify cancer-related cell states: stemness, migration, and metastasis. (58) combined an autoencoder with a visual transformer to obtain embeddings from spatial Copy Number Variation (CNV) and histology data from six cancer types in order to identify spatial subclones and reveal their evolutionary tree. Altogether, we see the relevance of combining ST data and DRL to relate pseudo-time trajectories with the spatial conformation of healthy and cancerous tissues.

E. Limitations and challenges

From the studies included in our review, we have observed some common trends that limit the study of cancer progression with DRL and DL. First, applications of DL and DRL in cancer studies are mostly centered on subtyping, diagnosis, and prognosis, which hardly take into account the longitudinal aspect of the disease. Studies that do take into account a temporal aspect tend to rely on single-cell data. Hence, the second limitation of the considered studies is that they tend to consider the cell as the fundamental unit of study, so they are not patient-centric. Compared to bulk data, single-cell data adds more per-sample noise, as it includes a level of cellular heterogeneity that is not present in bulk data. Besides, single-cell studies infer a pseudo-time developmental trajectory, calculating the evolutionary distance between cells, but without real temporal data, so the real longitudinal progression of the disease is not considered. Thirdly, stages, which can be seen as a proxy time-unit for cancer's time dimension, are hard to classify. This means it is difficult to extract any meaningful information to infer stage progression.

These limitations imply some challenges in the study of cancer in time. Firstly, the lack of longitudinal data is hence a major challenge. This is due to the need to sample patients at different time points, while also receiving treatment, altering the natural course of the disease. Even when the data is collected in this manner, which is usually more frequent in experimental animal data than human data (59), the degree of cancer advancement differs from patient to patient, meaning the resulting data is unaligned. A second challenge is the lack of a methodology to infer real-time progression. The use of generative models to amplify already-available data may be a way to circumvent this issue. Besides, the use of generative models may also provide a way to infer out-of-sample behavior, effectively adding a temporal dimension to the data. However, this approach leads to a third challenge: the lack of data to validate models. These challenges must be addressed in order to provide a more accurate understanding of cancer progression, and to develop more personalized treatments.

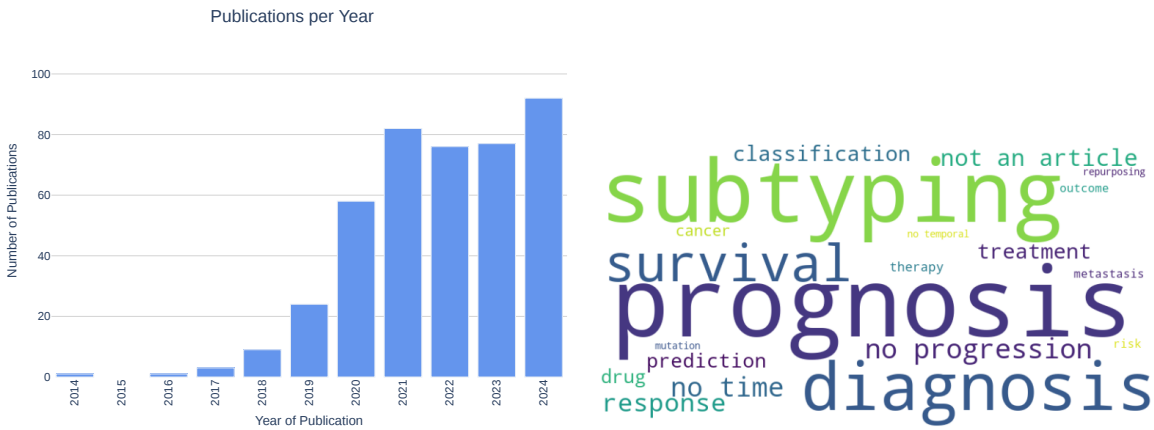

Figure 2. Results from queries. (**A)** Number of publications found, per year. There is a clear increase in relevant publications from 2020. (**B)** Word cloud of exclusion reasons. Most publications use deep and representation learning applied to cancer for subtyping, diagnosis, and studying prognosis and survival.

3. Discussion

DL is a commonly used methodology to learn complex patterns in health data. DRL comprises a specific type of DL methods that have the ability to learn a lower-dimensional representation of complex data distributions. This ability has been combined with generative capabilities in the VAE, a Bayesian-statistics-based representation learning method. Given the complexity of high-throughput data obtained in cancer research, these methods have been applied to study the disease. However, the dynamic nature of cancer implies a temporal component that has not yet been fully explored.

In this study, we conducted a Systematic Literature Review (SLR) on the use of DRL in cancer studies with omics data in a temporal context. To encompass the complexity of the topic, we created a query that was adapted to three different search engines (see Methods). Our search returned 440 papers, of which 21 were relevant to our investigation. We summarized these papers and categorized them into 4 main groups. (1) Discarded papers were analyzed for their exclusion reasons, revealing that the most common applications of DRL and VAEs in cancer are related to subtyping, diagnostics, and prognosis studies. (2) These techniques have been applied in some diseases, taking into consideration a temporal aspect, but their application in longitudinal cancer studies remains limited. (3) Stages may serve as a proxy to align cancer data in a temporal manner. (4) Single-cell omics data is the most frequently used type in cancer studies, with a recent surge in spatially resolved data; however, these studies arrange cells in pseudo-time. Collectively, these findings reveal that the study of cancer progression in time has not yet been adequately explored.

The lack of full matches from our queries may be related to substantial challenges and limitations in the field. Research efforts have predominantly focused on handling the high dimensionality of omics data through embedding inference, which is then utilized for downstream tasks. These tasks rarely address temporal cancer dynamics, except for single-cell studies that employ embeddings to reveal pseudo-time trajectories of cells. However, pseudo-temporal

ordering does not provide further longitudinal details such as clinically diagnosed stages or the elapsed time in the pseudo-temporal ordering. This type of longitudinal information is hard to obtain. Even when available, its acquisition occurs during treatment, altering the natural course of the disease, and involves sample destruction. Consequently, current cancer progression studies often rely on unaligned data.

Recently, there has been an uptick in interest in research of generative models, which can be used to align data, even when considering a temporal component. European projects, such as EVENFLOW, are delving into the application of such methodologies to overcome current limitations in cancer studies. Moreover, synthetic data generation methods based on simulation processes, such as agent-based modeling of cellular dynamics (60, 61), are making strides in integrating physics into generative frameworks to produce more realistic and reliable synthetic data. We believe that these approaches, along with investments in generating appropriate data for validation studies, will pave the way for the eventual development of personalized treatments that consider the individual progression of cancer across patients.

4. Methods

We undertook this systematic literature review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations (62).

A. Search process and selection criteria

We performed a systematic literature review on the Variational Autoencoder applied to omics data and the study of cancer progression through time. Given the complex nature of the topic and the diversity of possible terms related to our research, we considered the following keywords:

- Variational Autoencoder: We considered its abbreviation, VAE, as well. Since the name may not have been explicitly used in the title or abstract, especially during the years after its publication in 2014 (4), we also included the term “representation learning” and “deep learning”.
- omics: We included the main omics data types, such as genomics, transcriptomics, proteomics, and metabolomics, as well as their combination by including multiomics.
- cancer: May also be referred to as tumor.
- time: To include possible references to studies that consider cancer progression through time, such as longitudinal studies, the queries included the keywords progression, evolution, trajectory, time series and longitudinal.

For this review, we used Google Scholar, Scopus, and Web of Science to search the literature, including results from 2014 through 2024.

Given the keywords, we defined the following queries for each search engine:

- Google Scholar: (“deep learning” OR “representation learning” OR “Variational Autoencoder” OR VAE) AND (omics OR genomics OR transcriptomics OR proteomics OR metabolomics OR “multi-omics” OR multiomics) AND (“cancer” OR “tumor” OR “tumour”) AND (progression OR evolution OR trajectory OR “time series” OR longitudinal) since:2014 before:2025

- Scopus: TS=(“deep learning” OR “representation learning” OR “Variational Autoencoder” OR VAE) AND TS=(omics OR genomics OR transcriptomics OR proteomics OR metabolomics OR “multi-omics” OR multiomics) AND TS=(“cancer” OR “tumor” OR “tumour”) AND TS=(progression OR evolution OR trajectory OR “time series” OR longitudinal) AND PY=(2014-2024)

- Web of Science: TITLE-ABS-KEY((“deep learning” OR “representation learning” OR “Variational Autoencoder” OR VAE) AND (omics OR genomics OR transcriptomics OR proteomics OR metabolomics OR “multi-omics” OR multiomics) AND (“cancer” OR “tumor” OR “tumour”) AND (progression OR evolution OR trajectory OR “time series” OR longitudinal)) AND PUBYEAR > 2013 AND PUBYEAR < 2025

To keep our search within scope and guarantee good quality, we enforced the following inclusion criteria: (1) Peer-review journal articles and conference proceedings; (2) omics data studies; and (3) papers that mathematically study cancer evolution or representational learning to study tumor evolution. For the same reasons, we also apply these exclusion criteria: (1) Opinions, letters to the editor, reviews, systematic reviews, and grey literature; (2) non-English language works.

B. Data extraction

The search on Scopus and Web of Science was performed through the respective websites. Publish or Perish was used to extract the results from Google Scholar. The searches were limited to 200 results for each query. For each query, a table with information about the publication was downloaded from the respective platform. We then identified the duplicated titles and removed them. Next, we read the titles and abstracts, narrowing down our selection. On the remaining entries a full-text screening was performed. The process is summarized in (Figure 3). One reviewer conducted the systematic review process, and a second reviewer assessed the results.

C. Risk of bias

During the search process, we found that most of the journal articles returned by our queries centered on for diagnosis, subtyping, and prognosis of cancer (see Figure 2). This trend is an expected outcome since there exists a great effort toward the early detection of cancers.

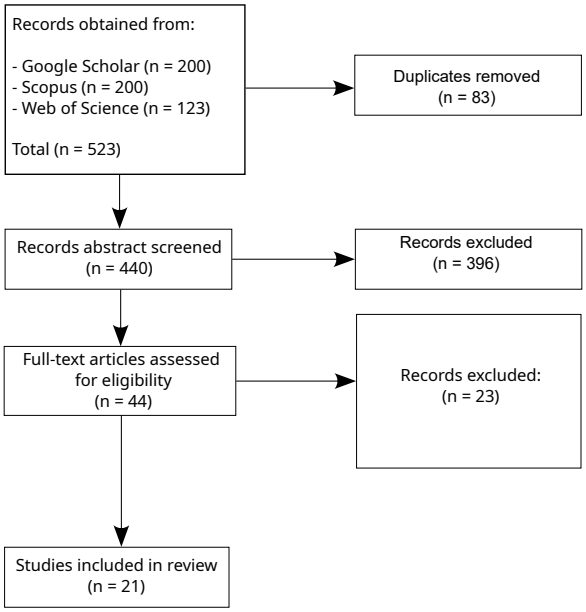

Figure 3. Flow diagram of the systematic literature review process, following PRISMA guidelines.

List of Abbreviations

AD Alzheimer's Disease. 6

AE Autoencoder. 5, 7

CNV Copy Number Variation. 9

CVAE Conditional Variational Autoencoder. 8

CVD Cardiovascular Disease. 6

DL Deep Learning. 2, 5–7, 9, 11

DRL Deep Representation Learning. 2–9, 11

GSEA Gene Set Enrichment Analysis. 6

KEGG Kyoto Encyclopedia of Genes and Genomes. 6

LSTM Long Short-Term Memory. 6

ML Machine Learning. 6

PAGA Partition-based Graph Abstraction. 8, 9

SLR Systematic Literature Review. 2, 3, 11

ST Spatial Transcriptomics. 8, 9

TCGA The Cancer Genome Atlas. 5, 7, 8

VAE Variational Autoencoder. 2–9, 11

5. Dictionary of Terms

Term	Definition
Autoencoder (AE)	DRL, unsupervised technique that encodes unlabeled data into a lower-dimensional representation, the latent space. It consists of an encoder, which embeds the input data, and a decoder, recreating the input data from the learned embeddings.
Conditional Variational Autoencoder (CVAE)	A VAE used for semi- or fully-supervised learning. In this case, the encoder and decoder are passed the labels of the input data and learn to encode and decode the input data depending on, i.e., conditioned to, the input data.
Convolutional Variational Autoencoder	A VAE that entails the use of convolutional layers in the encoder and decoder, making it suitable to learn latent-spece embeddings of images.
Contrast Learning	A supervised DRL method that learns embeddings of data by comparing, or contrasting, between similar and dissimilar instances within the input data.
Deep Learning (DL)	Discipline within Machine Learning that uses artificial, multilayer neural-networks to learn patterns present in a given input dataset.
Deep Representation Learning (DRL)	Also known as Deep Feature Learning, consists of techniques to automatically estimate a representation of the features included on a given dataset. We use the adjective Deep to remark the fact that we consider Deep-Learnig-based techniques.
Long Short-Term Memory (LSTM)	Specialized DL model designed to learn when to retain or forget relevant information present in sequential data.
Reinforcement Learning (RL)	Area of Machine Learning where an automated agent learns a process of decision-making while interacting with an environment and receiving feedback (a reward or a penalty) from an external interpreter.
Variational Autoencoder (VAE)	DRL technique that follows the AE ability to learn embeddings from unlabled data, albeit considering a variational Bayesian inference to learn a probabilistic latent space.

6. Competing Interest

None declared.

7. Authors Contribution

G.P.C. developed the systematic literature review, and wrote and revised earlier and consolidated versions of the manuscript. A.V. and D.C. conceived the study and oversaw its execution. All authors reviewed and contributed to the final manuscript.

8. Acknowledgements

This work has been supported by the EU project EVENFLOW under Horizon Europe agreement No. 101070430.

9. Bibliography

1. Riccardo Miotto et al. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246, November 2018.
2. Y. Bengio et al. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, August 2013.
3. Cosmin I. Bercea et al. Evaluating normative representation learning in generative AI for robust anomaly detection in brain imaging. *Nature Communications*, 16(1):1624, February 2025.
4. Diederik P Kingma et al. Auto-Encoding Variational Bayes. 2013. Publisher: arXiv Version Number: 11.
5. Yixin Wang et al. Comment: Variational Autoencoders as Empirical Bayes. *Statistical Science*, 34(2), May 2019.
6. Gregory A Roth et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, 392(10159):1736–1788, November 2018.
7. Melissa C. Hulvat. Cancer Incidence and Trends. *Surgical Clinics of North America*, 100(3):469–481, June 2020.
8. Davide Cirillo et al. Big data analytics for personalized medicine. *Current Opinion in Biotechnology*, 58:161–167, August 2019.
9. International Human Genome Sequencing Consortium et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.
10. Fuchou Tang et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, May 2009.
11. Zhong Wang et al. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, January 2009.
12. Scott D. Patterson et al. Proteomics: the first decade and beyond. *Nature Genetics*, 33(S3):311–323, March 2003.
13. Jeffrey R. Idle et al. Metabolomics. *Cell Metabolism*, 6(5):348–351, November 2007.
14. Patrik L. Ståhl et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, July 2016.
15. Eloise Withnell et al. XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data. *Briefings in Bioinformatics*, 22(6):bbab315, November 2021.
16. Kumardeep Chaudhary et al. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clinical Cancer Research*, 24(6):1248–1259, March 2018.

17. Ibiayi Dagogo-Jack et al. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15(2):81–94, February 2018.
18. Kefan Cao et al. Computationally reconstructing the evolution of cancer progression risk, December 2024.
19. Steven A. Frank. Somatic evolutionary genomics: Mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. *Proceedings of the National Academy of Sciences*, 107(suppl_1):1725–1730, January 2010.
20. Innocent Tatchum Sado et al. Early multi-cancer detection through deep learning: An anomaly detection approach using Variational Autoencoder. *Journal of Biomedical Informatics*, 160:104751, December 2024.
21. Tarek Khorshed et al. Deep Learning for Multi-Tissue Cancer Classification of Gene Expressions (GeneXNet). *IEEE Access*, 8:90615–90629, 2020.
22. Kaiwen Tan et al. A multi-omics supervised autoencoder for pan-cancer clinical outcome endpoints prediction. *BMC Medical Informatics and Decision Making*, 20(S3):129, July 2020.
23. Javier Fernandez-Mateos et al. Tumor evolution metrics predict recurrence beyond 10 years in locally advanced prostate cancer. *Nature Cancer*, 5(9):1334–1351, July 2024.
24. Lianmin Zhang et al. Leveraging mitochondrial-programmed cell death dynamics to enhance prognostic accuracy and immunotherapy efficacy in lung adenocarcinoma. *Journal for Immunotherapy of Cancer*, 12(10):e010008, October 2024.
25. Junyu Li et al. Towards artificial intelligence to multi-omics characterization of tumor heterogeneity in esophageal cancer. *Seminars in Cancer Biology*, 91:35–49, June 2023.
26. Asmaa M. Hassan et al. Multi-omics-based Machine Learning for the Subtype Classification of Breast Cancer. *Arabian Journal for Science and Engineering*, 50(2):1339–1352, January 2025.
27. Semih Kurt et al. CopyVAE: a variational autoencoder-based approach for copy number variation inference using single-cell transcriptomics. *Bioinformatics*, page btae284, April 2024.
28. Chang Xu et al. DeepST: identifying spatial domains in spatial transcriptomics by deep learning. *Nucleic Acids Research*, 50(22):e131–e131, December 2022.
29. Bo Yang et al. Deep Subspace Mutual Learning for cancer subtypes prediction. *Bioinformatics*, 37(21):3715–3722, November 2021.
30. Xiaoying Wang et al. MarsGT: Multi-omics analysis for rare population inference using single-cell graph transformer. *Nature Communications*, 15(1):338, January 2024.
31. Shihui Wang et al. A multi-model based on radiogenomics and deep learning techniques associated with histological grade and survival in clear cell renal cell carcinoma. *Insights into Imaging*, 14(1):207, November 2023.
32. Charlie Saillard et al. Pacpaint: a histology-based deep learning model uncovers the extensive intratumor molecular heterogeneity of pancreatic adenocarcinoma. *Nature Communications*, 14(1):3459, June 2023.
33. Adalberto Claudio Quiros et al. Mapping the landscape of histomorphological cancer phenotypes using self-supervised learning on unannotated pathology slides. *Nature Communications*, 15(1):4596, June 2024.
34. Vahid Bahrambeigi et al. Transcriptomic Profiling of Plasma Extracellular Vesicles Enables Reliable Annotation of the Cancer-Specific Transcriptome and Molecular Subtype. *Cancer Research*, 84(10):1719–1732, May 2024.
35. Davide Placido et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature Medicine*, 29(5):1113–1122, May 2023.
36. Garam Lee et al. MildInt: Deep Learning-Based Multimodal Longitudinal Data Integration Framework. *Frontiers in Genetics*, 10:617, June 2019.
37. Neo Christopher Chung et al. Unsupervised classification of multi-omics data during cardiac remodeling using deep learning. *Methods*, 166:66–73, August 2019.

38. Oriol Iborra-Egea et al. Deep Learning Analyses to Delineate the Molecular Remodeling Process after Myocardial Infarction. *Cells*, 10(12):3268, November 2021.
39. David L. Hölscher et al. Next-Generation Morphometry for pathomics-data mining in histopathology. *Nature Communications*, 14(1):470, January 2023.
40. Kathryn H. Brown et al. Characterisation of quantitative imaging biomarkers for inflammatory and fibrotic radiation-induced lung injuries using preclinical radiomics. *Radiotherapy and Oncology*, 192:110106, March 2024.
41. Erfan Sadeqi Azer et al. Tumor Phylogeny Topology Inference via Deep Learning. *iScience*, 23(11):101655, November 2020.
42. Monica-Andreea Baciu-Drăgan et al. Oncotree2vec — a method for embedding and clustering of tumor mutation trees. *Bioinformatics*, 40(Supplement_1):i180–i188, June 2024.
43. James R. Egner. AJCC Cancer Staging Manual. *JAMA*, 304(15):1726, October 2010.
44. Yixiao Feng et al. Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes & Diseases*, 5(2):77–106, June 2018.
45. Nicole M. Nelligan et al. Simulating the restoration of normal gene expression from different thyroid cancer stages using deep learning. *BMC Cancer*, 22(1):612, December 2022.
46. Sugi Lee et al. A deep learning and similarity-based hierarchical clustering approach for pathological stage prediction of papillary renal cell carcinoma. *Computational and Structural Biotechnology Journal*, 18:2639–2646, 2020.
47. Xinyi Zhang et al. Unsupervised representation learning of chromatin images identifies changes in cell state and tissue organization in DCIS. *Nature Communications*, 15(1):6112, July 2024.
48. Nguyen Phuoc Long et al. Systematic assessment of cervical cancer initiation and progression uncovers genetic panels for deep learning-based early diagnosis and proposes novel diagnostic and prognostic biomarkers. *Oncotarget*, 8(65):109436–109456, December 2017.
49. Aaron M. Smith et al. Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinformatics*, 21(1):119, December 2020.
50. Haluk Dogan et al. Elucidation of dynamic microRNA regulations in cancer progression using integrative machine learning. *Briefings in Bioinformatics*, 22(6):bbab270, November 2021.
51. Cole Trapnell et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, April 2014.
52. F. Alexander Wolf et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):59, December 2019.
53. Sabrina Rashid et al. Dhaka: variational autoencoder for unmasking tumor heterogeneity from single cell genomic data. *Bioinformatics*, 37(11):1535–1543, July 2021.
54. Nan Chen et al. A signal-diffusion-based unsupervised contrastive representation learning for spatial transcriptomics analysis. *Bioinformatics*, 40(11):btae663, November 2024.
55. Yongshuo Zong et al. conST: an interpretable multi-modal contrastive learning framework for spatial transcriptomics, January 2022.
56. Thomas Hu et al. Single-cell spatial metabolomics with cell-type specific protein profiling for tissue systems biology. *Nature Communications*, 14(1):8260, December 2023.
57. Chunman Zuo et al. Elucidating tumor heterogeneity from spatially resolved transcriptomics data by multi-view graph collaborative learning. *Nature Communications*, 13(1):5962, October 2022.
58. Chia-Kuei Mo et al. Tumour evolution and microenvironment interactions in 2D and 3D space. *Nature*, 634(8036):1178–1186, October 2024.

59. Xiaowei Song et al. Strengths and challenges of longitudinal non-human primate neuroimaging. *NeuroImage*, 236:118009, August 2021.
60. Ahmadreza Ghaffarizadeh et al. PhysiCell: An open source physics-based cell simulator for 3-D multicellular systems. *PLOS Computational Biology*, 14(2):e1005991, February 2018.
61. Miguel Ponce-de Leon et al. PhysiBoSS 2.0: a sustainable integration of stochastic Boolean and agent-based modelling frameworks. *npj Systems Biology and Applications*, 9(1):54, October 2023.
62. Matthew J Page et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, page n71, March 2021.