scientific reports

OPEN

Robust evaluation of deep learning-based representation methods for survival and gene essentiality prediction on bulk RNA-seq data

Baptiste Gross[1,2,2]✉, Antonin Dauvin[1,2,2], Vincent Cabeli[1,2,2], Virgilio Kmetzsch[1], Jean El Khoury[1], Gaëtan Dissez[1], Khalil Ouardini[1], Simon Grouard[1], Alec Davi[1], Regis Loeb[1], Christian Esposito[1], Louis Hulot[1], Ridouane Ghermi[1], Michael Blum[1], Yannis Darhi[1], Eric Y. Durand[1,2,2] & Alberto Romagnoni[1,2,2]

Deep learning (DL) has shown potential to provide powerful representations of bulk RNA-seq data in cancer research. However, there is no consensus regarding the impact of design choices of DL approaches on the performance of the learned representation, including the model architecture, the training methodology and the various hyperparameters. To address this problem, we evaluate the performance of various design choices of DL representation learning methods using TCGA and DepMap pan-cancer datasets and assess their predictive power for survival and gene essentiality predictions. We demonstrate that baseline methods achieve comparable or superior performance compared to more complex models on survival predictions tasks. DL representation methods, however, are the most efficient to predict the gene essentiality of cell lines. We show that auto-encoders (AE) are consistently improved by techniques such as masking and multi-head training. Our results suggest that the impact of DL representations and of pretraining are highly task- and architecture-dependent, highlighting the need for adopting rigorous evaluation guidelines. These guidelines for robust evaluation are implemented in a pipeline made available to the research community.

Keywords RNAseq, Representation learning, Deep learning, Survival prediction, Gene essentiality, Benchmarking

Precision medicine and the development of new therapies require accurate disease diagnosis and outcome prediction. The field of omics research has experienced an unprecedented data revolution fueled by high-throughput technologies, enabling the generation of high-dimensional omics data at an exponential pace. This wealth of data provides interesting opportunities to unravel the molecular landscape of diseases, including cancer, and emphasizes the need for robust computational approaches to extract meaningful insights. In particular, RNA sequencing (RNA-seq) is now ubiquitous in molecular biology and oncology[1] and was shown to be the most informative omics modality for predicting phenotypes of interest such as patient survival[2] or gene essentiality in cell lines[3].

In parallel, deep learning-based representation learning approaches have shown remarkable potential in analyzing complex data, ranging from images to texts[4–6]. These methods, powered by artificial neural networks, excel at capturing intricate patterns, detecting subtle relationships, and making accurate predictions. Applying deep representation learning techniques (DRL) to RNA-seq data for cancer research holds the potential to revolutionize our understanding of cancer progression, classification, and treatment response.

Therefore, the integration of deep learning-based approaches within the field of omics research holds immense promise for advancing our understanding of cancer biology[7]. Nonetheless, despite the vast potential of DRL algorithms and demonstrated success in vision and Natural Language Processing (NLP) domains, they still face

[1]Owkin, Inc., New York, NY, USA. [2]These authors contributed equally: Baptiste Gross, Antonin Dauvin, Vincent Cabeli. Eric Y. Durand and Alberto Romagnoni jointly supervised the work. ✉email: baptiste.gross@owkin.com

challenges in surpassing traditional tree-based methods on tabular data[8]. Importantly, their application to omics data remains underexplored when considering gene expression matrices derived from bulk RNA-seq.

Typical tasks associated with omics data like survival or gene essentiality predictions present unique challenges, involving high-dimensional feature spaces with limited sample sizes, a need to account for batch effects and variations in data generation procedures[9,10]. They require various intricate steps like normalization, scaling, dimensionality reduction (DR), and the selection of prediction models and training frameworks. Moreover, the presence of noisy and heterogenous labels, such as survival and censoring information, further complicates the analysis. Generally, these tasks can be sensitive to overfitting and exhibit significant variability across different datasets and tasks.

Reducing the dimensions of genomic data is often a privileged option to help address such difficulties. Non-DL dimensionality reduction methods have proven effective in analyzing omics data for guided feature selection[11], deconvolution of bulk RNA-seq data[12], clustering[13,14] or prediction of clinical endpoints[15]. With the rise of deep learning, several papers have explored new representation learning methods for cancer transcriptome data analysis such as auto-encoder architecture variations for unsupervised tasks such as biomarker identification[16,17], subtyping[18] or as a preliminary step to reduce dimensions before supervised tasks including gene essentiality[3] or drug response predictions[19–22]. In single-cell data, the higher amount of data points allowed the development of more complex methods inspired from self-supervised learning[23,24], graph-based methods[25] and more recently from large language models[26–28] for novel cell type discovery and cell type annotation. Nonetheless, despite numerous publications showcasing the potential of deep representation learning, there is a scarcity of comprehensive benchmarks in the context of cancer research focusing on deep-learning methods for bulk RNA-seq due the multitude of potential tasks. Existing benchmarks showcase limited improvement over linear baselines when using deep learning methods in phenotype prediction or clustering[29,30], but do not cover tasks such as gene essentiality prediction or pan-cancer pretraining approaches. This lack of reliable comparisons seemingly originates from the difficulty of accurately assessing models' performances in a simple cross-validation setting[31,32], which becomes even more complex when adding hyperparameters (HPs) tuning, small dataset sizes, noisy labels and confounding factors, which are ubiquitous in omics for oncology[33].

This impedes the development and evaluation of these methods, limiting their reliability and usage. It has been argued that the rapid growth of deep learning in other applications can be partly attributed to the widespread adoption of clear benchmarks, like ImageNet[34] for visual deep learning or more recently the CASP competition for protein structure prediction[35].

In this paper, we investigated and robustly benchmarked the performance of different prediction models and training approaches on bulk RNA-seq data in the context of cancer research. In particular, we studied the impact of the representation method choice on the performance of 11 cancer-specific survival prediction tasks, a pan-cancer survival prediction task on 33 combined cohorts and a gene essentiality prediction task on hundreds of cell lines. We implemented deep representation models that were used successfully on bulk RNA-seq data such as auto-encoders[16] (AE), pre-trained auto-encoders[3,20] (PreAE), graph neural networks[36] (GNN) and multi-head auto-encoders[37] (MHAE) as well as models often used to analyze single-cell data such as variational auto-encoders[38] (VAE) or masking auto-encoders[39] (MAE) that are also known to perform well on tabular data in other fields[40]. Given the limited size of data points in bulk transcriptomics dataset, we compared these models to AE trained on augmented data with Gaussian Noise (DA-GN) to mimic the classical augmentation process used in computer vision[41]. These DRL models were compared to baseline models: transcripts per millions (TPM) gene counts, referred to as Identity, and Principal Component Analysis (PCA). Our evaluation encompassed preprocessing steps such as normalization, scaling, and feature selection, and provided a fully reproducible framework to evaluate their impact as well as the one of random splits during cross-validation and hyperparameters (HPs) tuning strategy on different prediction tasks.

This study highlights the considerable variability in results due to data splits and the limited impact of the choice of the representation model or training framework on the performances on the considered tasks.

Results

Overview

In this work, we benchmarked the performance of different representations of bulk RNA-Seq data on survival prediction tasks on TCGA dataset and gene essentiality prediction on DepMap data. In particular, we compared the results obtained with standard baseline representations (Identity and PCA) with those obtained with DRL architectures such as AE, VAE, MAE, MHAE, DA-GN, GNN based on prior knowledge and PreAE. An overview of the different tasks, metrics and models considered is shown in Table 1.

The benchmarking pipeline (Supplementary Fig. S1) was based on repeated holdout cross-validation processes and addressing issues such as unfair budget for HP tuning, overfitting or performing hardly generalizable statistical tests on particular data splits (See "Methods" section and Supplementary materials).

Baseline methods perform as well as DRL methods on survival prediction tasks

We first considered the survival prediction tasks in TCGA datasets by studying separately the different cancer cohorts. Specifically, we focused on 13 cohorts corresponding to different cancer types with sufficient training size (cf Methods) : Breast Invasive Carcinoma (BRCA), Uterine Corpus Endometrial Carcinoma (UCEC), Kidney Renal Clear Cell Carcinoma (KIRC), Head and Neck Squamous Cell Carcinoma (HNSC), Lung Adenocarcinoma (LUAD), Brain Lower Grade Glioma (LGG), Lung Squamous Cell Carcinoma (LUSC), Skin Cutaneous Melanoma (SKCM), Colon Adenocarcinoma (COAD), Stomach Adenocarcinoma (STAD) and Bladder Urothelial Carcinoma (BLCA). For each of these cohorts, we trained all the representation models considered and used the learned embeddings to predict the overall survival (OS) of each patient thanks to a downstream survival

Tasks	Per-cohort OS	Pan-cancer OS	Gene essentiality	
Description	Considering one tumor type at a time, we aim to predict the overall survival of patients based on the RNA-seq learned representations and accurately rank them by their predicted risk[42]	Same task as Per-cohort OS but pulling all tumors together and ranking patients' predicted risk across all indications[2]	Given cell line RNA-seq learned representations, predict if a given gene of interest is essential to their survivals[3]	
Evaluation metric	C-index	C-index	Overall Spearman Correlation: we evaluate if models can rank all of the cell lines/genes combination correctly	Per-DepOI Spearman Correlation: scores are the average of correlations computed per-DepOI
Models				
Identity	x	x	x	x
PCA	x	x	x	x
AE[16]	x	x	x	x
scVI[38]	x	x	x	x
MHAE[37]	x	x	x	x
MAE[39]	x	x	x	x
DA-GN[41]	x	x	x	x
GNN[36]	x	x	x	x
PreAE[3,20]	x		x	x
PreAE finetuned	x		x	x

Table 1. Overview of the representation models tested over the different tasks included in our benchmark. Models tested for a given task are indicated with an "x". References to existing literature are marked within superscript. Due to the lack of a relevant pretraining dataset for the pancancer task, we did not perform any pretraining experiments on this task.

prediction model, similar for every representation model tested. The performance of the models was evaluated using the concordance index (c-index) that quantifies if a model can output patient risk scores that rank accurately compared to the ground truth (cf Methods). This task is known to be challenging as the survival labels are noisy with a high percentage of censoring and debatable quality as shown in previous work[9]. Additionally model performance varies greatly between cohorts because of difference in datasets quality and low prognosis signal in transcriptomic data compared to other modalities for certain cohorts[2,42]. As shown in Fig. 1, in the indication-specific survival prediction task, we observed that the choice of the representation model had minimal impact on performance, as they consistently achieved similar scores in terms of mean c-index, with less than 10% difference in most cohorts between the best and worst models and close to 15% difference in BRCA or STAD.

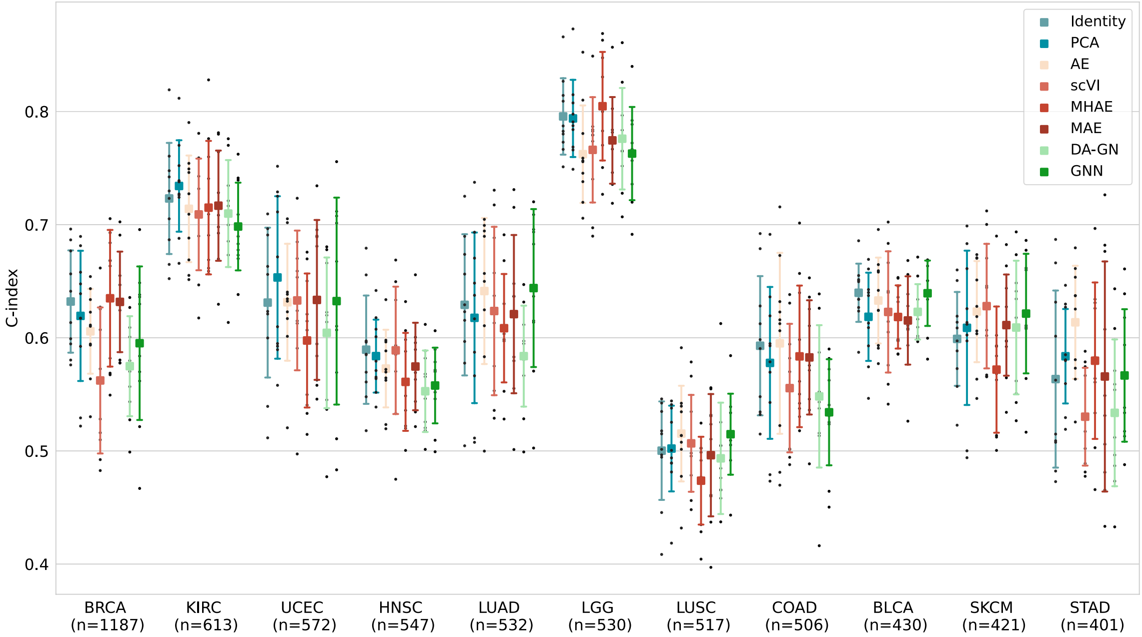

Figure 1. Comparison of performance on per-cohort OS prediction task on different TCGA cohorts for different bulk RNA-seq representation models. Black dots represent c-index results on different test folds. For each model, mean (square) c-index and standard deviation (intervals) are represented. Numbers on x axis labels represent the total number of samples available for this task in each cohort.

Interestingly, Identity and PCA demonstrated excellent performance on most cohorts. Notably, as shown in Fig. 1 and Supplementary Fig. S2, using 75% acceptance criterion on test folds, no model outperformed all the others in at least one cohort. Nonetheless, by using the number of significant pairwise comparisons, Identity was the best representation model choice for BRCA (together with MAE), LUAD, COAD (together with AE) and BLCA while PCA was the best choice for UCEC, KIRK and HNSC. Among the DRL methods, under the same criteria, AE is also the best choice for SKCM and STAD, MHAE for LGG. Notice that these results are not equivalent to standard comparison of mean performance (cfr. Supplementary Table S1), for which for example Identity would only be considered the best model for HNSC and BLCA.

We also considered the OS prediction task in a pan-cancer setup where all cohorts are pulled together, and training and evaluation are done combining the different cancer indications (cf Methods). This task is easier than the previous one: different cancers can easily be classified thanks to their transcriptome, and they have different OS distributions because of their different aggressiveness[9]. For this task, most of our models, including baselines, produced comparable performances when taking into account test set variability. Even though results are not directly comparable because of different evaluation frameworks, State-of-the-art (SOTA) models such as MultiSurv[2] when using solely RNA-seq data exhibits comparable performance, 0.758 (0.735–0.780), to most of our models, including baselines. As shown in Fig. 2, we observed that deep learning methods did not exhibit a clear advantage over all baselines, also in this pan-cancer case. Indeed, PCA, MAE and MH-AE equally emerged as the best-performing methods on pairwise comparison under the 75% criterion acceptance (Supplementary Fig. S3). Once again, mean c-indexes for all representation models differ for no more than 1.5%. Notice that in this case, similar conclusions could have been obtained by comparing mean c-index over test folds (cfr. Supplementary Table S2).

DRL methods showcase small but consistent improvements over baselines for gene essentiality prediction

The third task we considered was the Gene Essentiality prediction task. In the DepMap dataset, CRISPR experiments were conducted on a limited set of genes (n = 1223), disabling them to test if they are essential to the survival or growth of various tumor cell lines. These targeted genes are referred to as "Dependency of Interest" (DepOI) to avoid any confusion with the genes present in the transcriptomic data describing the cell lines. A score called gene essentiality is assigned to every pair of cell lines/DepOIs tested (cf Methods, Dataset). In our task, inspired by the DeepDEP framework, we used the representation models to create embeddings of the cell lines using their transcriptomic data. Downstream prediction models taking as input the combination of a cell line representation and a vector describing the DepOI considered in the experiment were trained to predict the gene essentiality of this combination. For this task, two evaluation metrics were considered: an overall correlation and a per-DepOI correlation in which the models are evaluated specifically for each DepOI and the final score is the average of the 1,223 correlations (cf Methods). As expected, models obtained higher scores on the Overall Spearman correlation than the per-DepOI one (Fig. 3), with differences between the best and the worst models around 0.6% and 8% for the overall and per-DepOI correlations respectively. Indeed, it is a more challenging task to grasp the distinction within a specific gene across various cell lines than it is to do so across all genes and cell lines. To verify that the obtained score ranges were consistent with the original DeepDEP study, we also computed per-DepOI scores using the Pearson correlation. We observed values ranging from 0.268 to 0.294 across the different models, compared to 0.14 for Exp-DeepDEP (RNA-seq model) and 0.17 for DeepDEP (multimodal model), showcasing that all our models were performing well on this task. Interestingly, we see in this task that DRL methods more clearly demonstrated a superior performance compared to baselines

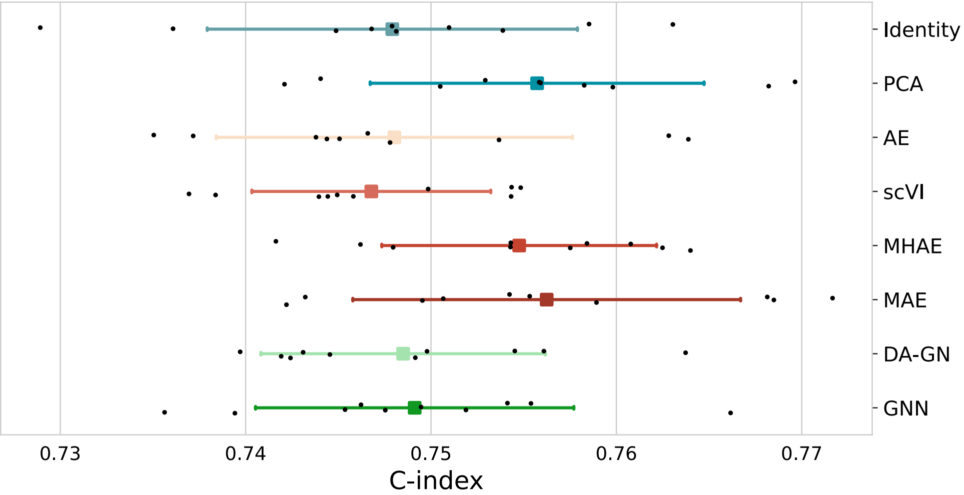

Figure 2. Comparison of performance on pan-cohort OS prediction task on TCGA dataset for different bulk RNA-seq representation models. Black dots represent c-index results on different test folds. For each model, mean (square) c-index and standard deviation (intervals) are represented.

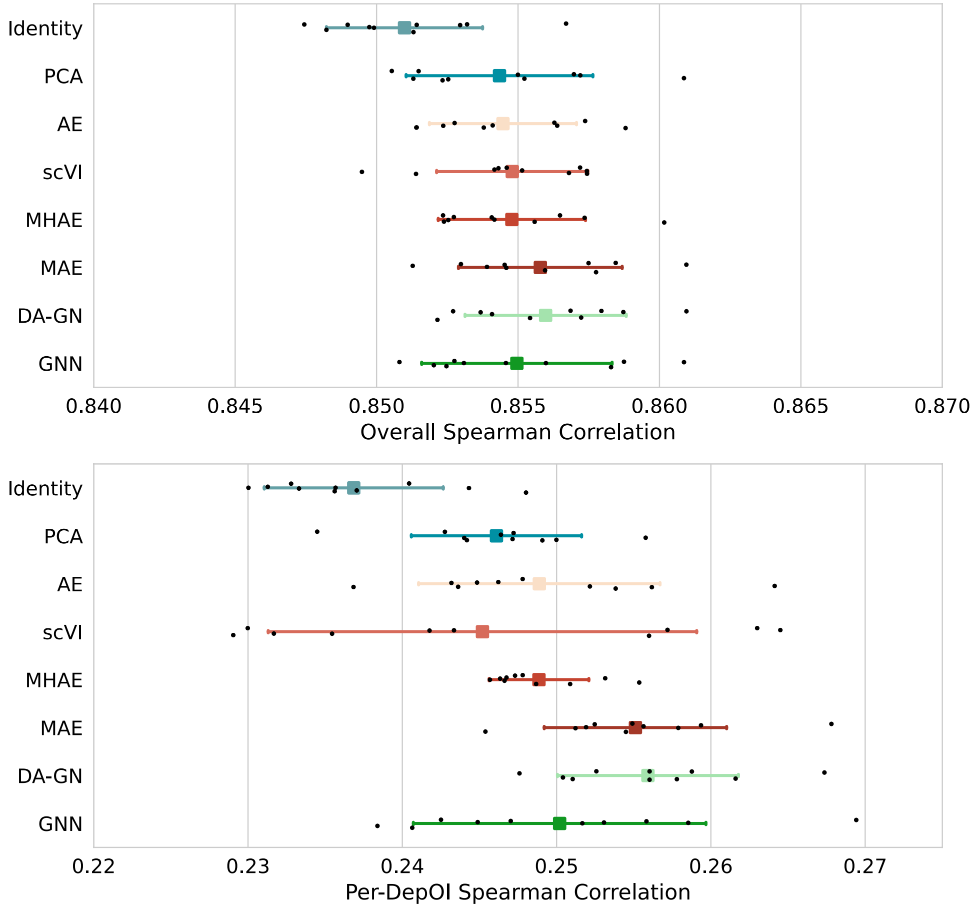

Figure 3. Comparison of performance on gene essentiality prediction task on DepMap dataset for different bulk RNA-seq representation models. (Top panel) Spearman Rank correlation distributions between predicted and observed gene essentialities: black dots represent results on different test folds. For each model, mean (square) Spearman Rank and standard deviation (intervals) are represented. Models are evaluated on their capacity to rank combinations of cell line and DepOI based on their gene essentiality score. (Bottom panel) Same as Top panel, but correlation computed per-DepOI. In this case, models are evaluated on their capacity to rank cell lines based on their gene essentiality score for that particular DepOI. The 1223 correlations are averaged to produce a single score per-DepOI correlation.

on pairwise comparison under the 75% criterion acceptance (Supplementary Fig. S4). In particular, MAE and DA-GN proved to be the most effective approaches for both overall and per-DepOI correlations. Nonetheless, PCA reaches comparable results with best performing DRL methods, with differences of the order of 0.2% for overall correlation and 4% for per-DepOI correlation. Also in this case, mean Spearman correlation over test folds confirm superiority of both MAE and DA-GN over the other methods, and in general of DRL over simple baselines (Supplementary Tables S3, S4).

Top performing models are task-specific and single test set evaluation can lead to false claims of superiority

All in all, we observed that no model clearly outperforms all the others for all tasks and all cohorts. Our study highlighted high variability of performances across different test sets for all tasks and models, showing the importance of adapting a rigorous evaluation framework taking these into account. Notably, we show that performing bootstrapping on a single test set could lead to false claims of superiority for certain models that are not generalizable to other data splits (Fig. 4A).

Nonetheless, as described in the "Methods" section, the acceptance criterion can be used to define another scoring system to evaluate the relative superiority of a model over the others, that by construction is more and more consistent once increasing the number of test folds.

The overall results about relative scoring of the models on the different tasks, using acceptance criterion and pairwise comparison, are shown in Fig. 4B. Under this scoring system, the baseline representation models appear as the best models for survival prediction, Identity for per-cohort task and PCA for pan-cancer (together with MHAE and MAE). On the other hand, all DRL methods can be considered better (or at least equivalent) choices for Gene Essentiality prediction tasks, under both overall and per-DepOI correlation metrics.

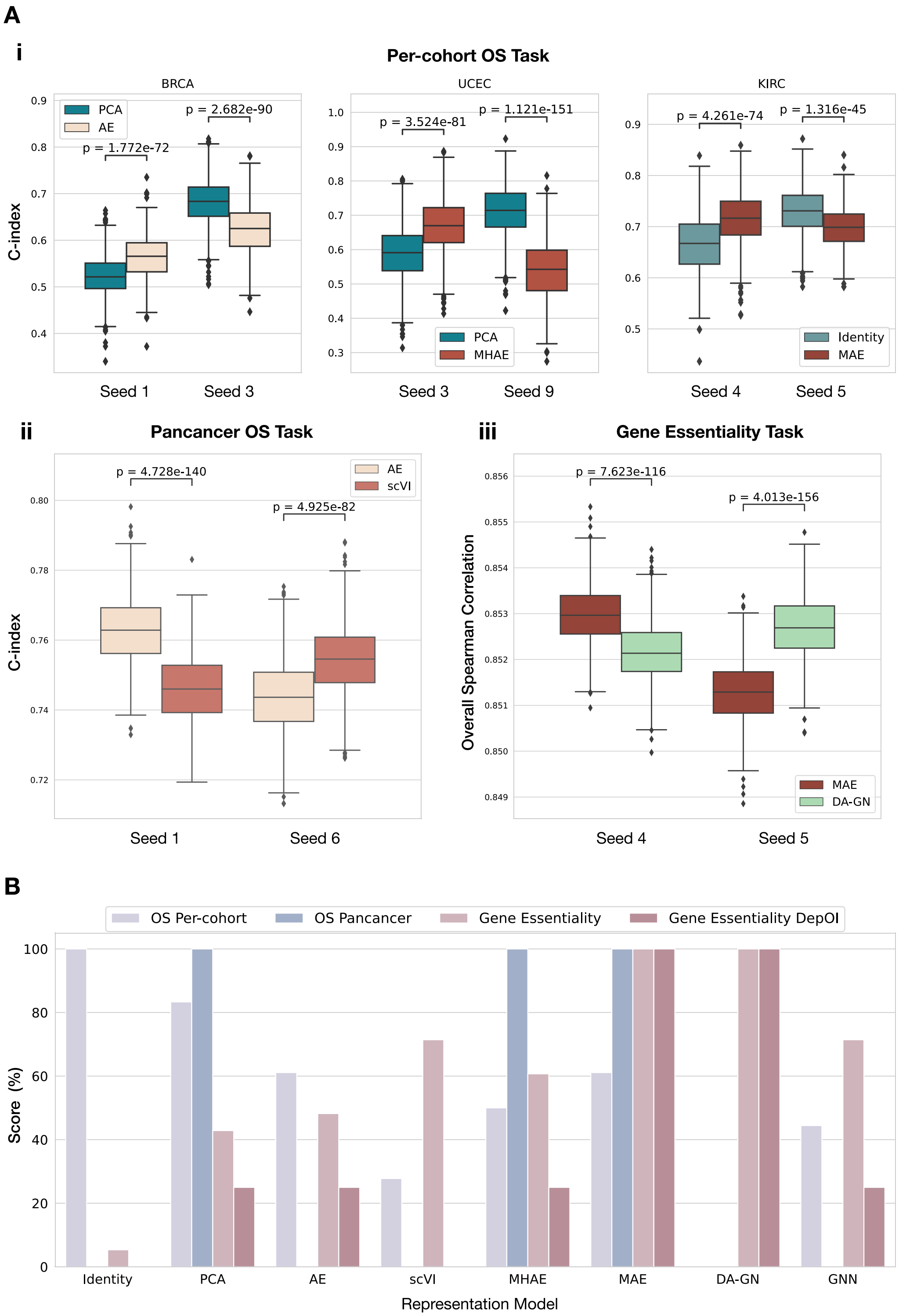

Figure 4. Comparison of model evaluation processes. (**A**) Test metric distributions obtained with bootstrap (n = 1000 per seed) on the test sets generated by different seeds on the downstream tasks. Models were trained following our repeated holdout pipeline, but graphs show results only on arbitrary chosen individual test sets to showcase the limitations of this evaluation process. p-values are computed using Wilcoxon tests. (**i**) Comparison of representation models on 3 example cohorts from the per-cohort OS prediction task. (**ii**) Comparison of AE and scVI on the pan-cancer task. (**iii**) Comparison of MAE and DA-GN on the Gene Essentiality task. (**B**) Overall comparison of the different representation models over the tasks Per-cohort OS Prediction, Pan-cohort OS prediction and gene essentiality (overall metric and per-DepOI). The score for a given model corresponds to the number of winning pairwise comparisons to other models according to our acceptance test criteria. The sum is then normalized by task to obtain the final score and is expressed as a percentage.

Auto-encoders directly trained on the downstream datasets perform as well as pretrained auto-encoders on survival and gene essentiality prediction tasks

As DRL methods are known to perform better on large datasets, we investigated the hypothesis under which they could benefit from pretraining on external datasets, as suggested by the DeepDEP findings showing improved performances of their multimodal models after pretraining auto-encoders on TCGA. For this use case, we focused on the auto-encoder based model on the per-cohort survival prediction and the gene essentiality prediction tasks. For both tasks, the external dataset for pretraining is taken from TCGA, removing the downstream task cohorts for the survival prediction. Specifically for the gene essentiality task, this setting allows us to potentially extend to RNA-seq-only models the claim from DeepDEP that a dataset representing different entities such as patient tumor samples can still be used to train a model to generate consistent embeddings of cell lines expression profiles. Due to the lack of a relevant pretraining dataset for the pancancer task, we did not perform any pretraining experiments on this task.

We first repeated the survival prediction tasks in per-cohort TCGA datasets, by exploring the two different pretraining strategies described in the "Methods" section, on basic auto-encoders. As shown in Fig. 5 and Supplementary Fig. S5, while in terms of mean performances, PreAE finetuned models outperform PreAE on almost all cohorts (with the exception of LUSC and SKCM), with differences of performances between AE and pretrained models ranging from 1 to 11% at most (Supplementary Table S5), the superiority of PreAE finetuned is less clear in terms of 75% acceptance criterion on test folds.

When considering the Gene Essentiality prediction task on the DepMap dataset, the results shown in Fig. 6 and Supplementary Fig. S6 highlight that auto-encoders pre-trained on TCGA samples (details in "Methods" section) struggle to outperform non pre-trained auto-encoders under the acceptance criterion assumed in this paper. This is the case for both pretraining strategies considered in this paper (with or without fine-tuning), and for both overall and per-DepOI correlations. Notice that PreAE (for which only the prediction model on top of the representation obtained on TCGA samples is trained on CCLE, the downstream task dataset) reaches performances which differ from the AE by 0.25% and 5% for the overall and per-DepOI correlation respectively (Supplementary Tables S6, S7). Moreover, retraining on CCLE dataset consistently improves the performance of PreAE for this task. It is to be noted that our setting here is different from DeepDEP as we do not focus on developing an end-to-end model integrating multiple modalities, but we conducted similar experiments on Exp-DeepDEP using the original code baseline (see Supplementary Material for details) and reached similar conclusions (Supplementary Fig. S7 and Supplementary Tables S6, S7).

Discussion

In this paper, we benchmarked different architectures for representation of bulk RNA-Seq data against two classical downstream tasks in cancer research: survival and gene essentiality prediction. Notably, we observed that simple methods for representation like Identity and PCA can achieve comparable or even superior performances with respect to more complex or deep models. In particular, we observed that simple baseline models can be considered as achieving superior performance over other DRL on survival prediction tasks on patient TCGA data, while DRL methods seem to have a slight but consistent advantage over baselines on gene essentiality prediction tasks on cell line DepMap data. This difference might be due to patient data potentially being noisier than cell-line data, introducing more uninformative features to which deep learning architectures are sensible[8]. In the supervised setting and especially for risk models, the signal might be too sparse to be detected and

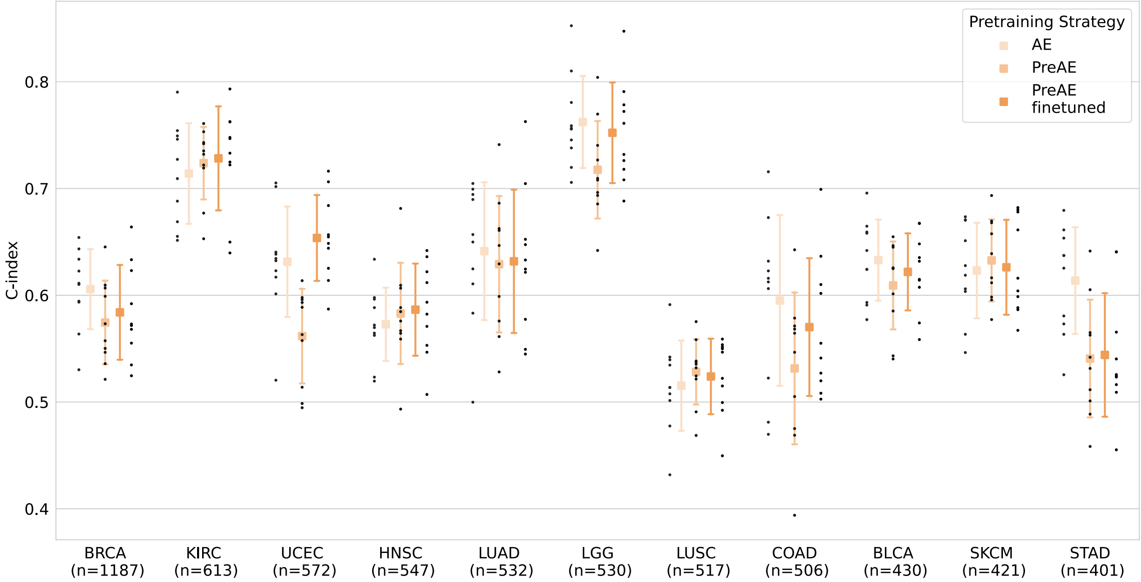

Figure 5. Impact of the pretraining strategy on the performance of auto-encoders on the per-cohort OS prediction task on different TCGA bulk RNA-seq cohorts. Black dots represent c-index results on different test folds. Numbers on x axis labels represent the total number of samples available for this task in each cohort.

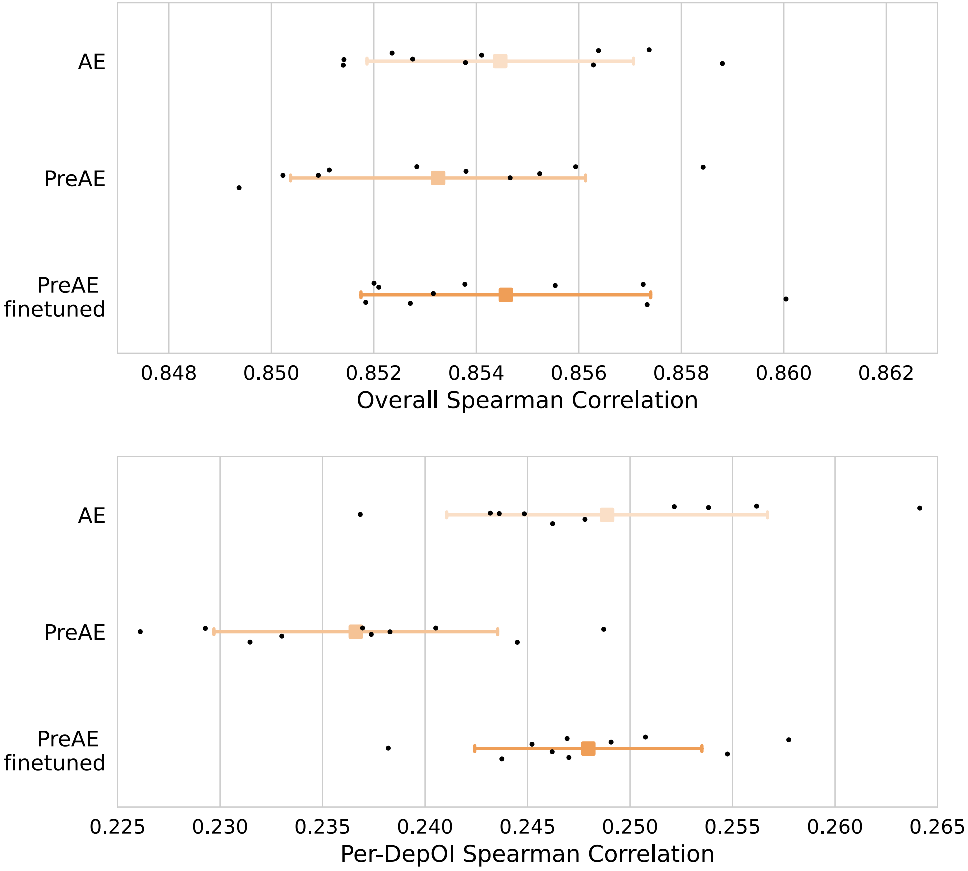

Figure 6. Impact of the pretraining strategy on TCGA samples on the performance of auto-encoders on gene essentiality prediction task on DepMap dataset. (Top panel) Correlation distributions between predicted and observed gene essentialities: black dots represent results on different test folds. (Bottom panel) same as top Left panel, but correlation computed per-DepOI.

amplified by deep learning methods which typically require an order of magnitude more data. Similar behavior was evidenced for example in the multiple myeloma DREAM challenge where a four-parameter model using two clinical features and two genes' expression performed as well as the best model on the validation datasets[43]. Nevertheless, our findings indicated consistent improvements across tasks in auto-encoders when incorporating masking and multi-head techniques, suggesting promising avenues for future research.

Furthermore, we utilized our benchmarking framework to investigate the impact of pretraining on the same models and tasks. We found that pre-trained auto-encoders on patient samples (TCGA) generated representations that achieved comparable gene essentiality performance on cell-lines data to fine-tuned methods, without any specific preprocessing or alignment between TCGA and CCLE datasets. This shows that the representation is able to transfer knowledge and suggests potential for significant computational time savings by pretraining a "feature extractor", a practice common in image processing. However, we observed that pretraining did not lead to significant performance improvements in both overall survival and gene essentiality prediction, even after subsequent fine-tuning.

We also showed that comparing models on a singular test set sampled from the task datasets can lead to misinterpretations even if coupled with statistical significance testing. This highlights the importance of adopting standardized evaluation practices to avoid unwarranted claims of superiority for novel techniques. This seems particularly true when training representation models on patient bulk RNA-seq data and is probably due to the poor ratio samples/features and the intrinsic complexity of the underlying biology. The limited size of the datasets available compared to other fields of machine learning enhance the heterogeneity between the training and the validation splits, explaining partially the high variability of performance across different repetitions for the same model.

In this study, we tackled the issue of reproducibility of results in the field of deep learning on omics data for cancer research. We focused on building a robust, reproducible and fair benchmarking framework to compare different machine learning models built on bulk RNA-seq data with respect to their performance on different downstream tasks. The key elements used in the pipeline we developed (repeated holdout cross-validation and fixed HP tuning budget) are not new per-se[44] but to our knowledge this framework has been used in a limited number of studies[29] evaluating machine learning models on omics data in cancer-specific downstream tasks. Tuning directly on a cross-validation without a held-out test set could lead to inflated performance and justifies the need for a nested cross-validation framework or the proposed repeated holdout test set. We chose the repeated holdout test to be able to provide confidence intervals and to provide a fairer metric (see details in Supplementary Materials). Additionally, we adopted the acceptance criterion from Varoquaux et Colliot[45] to compare model performance, in order to focus on meaningful improvements on finite-sized test sets. This criterion allows us to establish a scoring system for relative model performance, complementing standard mean performance considerations by accounting for variations from non-model factors like random splits during cross-validation and HP tuning strategy. We applied this benchmarking framework to evaluate the performance of different dimensionality reduction methods, including various model architectures and training techniques, on different downstream tasks (survival on patient data and gene essentiality on cell lines) using TCGA and DepMap public pan-cancer datasets.

We focused this work on fair comparisons and ablation studies, ensuring that each model was tested independently without combining different elements such as scaling, normalization, and gene selection. This approach allowed us to identify the strengths and weaknesses of each model more accurately. We believe this framework to be robust and generalizable to other tasks. For instance, while in the original VIME paper[40], it was observed that the addition of mask prediction as a subtask provided improved performances, in our specific use case our ablation study revealed that once applied on bulk RNA-seq data, this addition actually worsened the models' performance and increased training time.

While our study provides valuable insights, we acknowledge some limitations. For instance, we chose not to incorporate multi-omics data in our analysis due to prior research[2,3] indicating marginal gains or lack of statistical significance when integrating such data in related studies. The framework uses a fixed budget HP tuning, which could arguably have been different per model. Larger models and higher HP budgets could potentially yield better performances. The variability of a given model over the different test sets also showcased the sensibility of the models to the datasets considered. Our benchmark is therefore specific to these datasets and their associated tasks, and our results cannot be generalized to other biomedical prediction tasks without additional experiments. Our study also focused on breadth rather than depth, exploring many models without deep parameter studies and tuning. Regarding the comparison with a related method, DeepDEP, we observed that the datasets used in our study were not identical, possibly leading to adaptations in hyperparameters. Nonetheless, we demonstrated that pretraining did not consistently improve performance, supporting our earlier observations. However, while we focused on developing unimodal models using solely RNA-seq data, the gain in performance due to pretraining in DeepDEP is observed on multimodal models. This is potentially hinting that other modalities can more easily benefit from transfer learning and that transcriptomic data might require additional steps such as alignment or batch correction before applying complex DL methods. Lastly, we want to highlight that the highest-performing method on a single metric may not necessarily be the most effective approach in practical applications. Taking into account other relevant considerations, such as model interpretability, can contribute to more informed decisions regarding method selection for specific use cases.

Future development could include extending our tasks by considering out-of-domain evaluation metrics or exploring additional tasks like drug response prediction. The computational scale of the study can be increased, with greater budget for HP tuning and scaling up the number of test sets, and combining models for potentially greater improvements is a natural step. Additionally, different architectures from the regular auto-encoders could potentially leverage the pretraining process more efficiently as well as pretraining on larger bulk RNA-seq datasets[46]. Recent learnings from large pretrained models applied to single-cell data could be applied to our use case, such as different feature preprocessing of RNA-seq[26,27].

In conclusion, we believe our study brings valuable insights to the understanding of deep learning methods for bulk RNA-seq data representations in cancer research. By providing a robust evaluation framework we aim to facilitate fair comparisons of novel approaches against established methods, thus advancing the field towards more reliable and impactful techniques. The limitations identified in this study also offer guidance for future research directions and potential refinements to enhance the application of deep learning in cancer research.

Methods

In this section, we describe all the different elements of the benchmarking pipeline depicted in Supplementary Fig. S1. Additional details can be found in Supplementary Materials. Although our pipeline can be used to evaluate any RNA-seq-based machine learning model, we focused in this study on benchmarking the capacity of various representation models to capture relevant information in bulk RNA-seq data.

Datasets

This study was based on publicly available data from The Cancer Genome Atlas Program (TCGA) and Cancer Cell Line Encyclopedia (CCLE) dataset[47]. Expression data (raw counts, RPKM, TPM) for all TCGA datasets were downloaded from RECOUNT3[48], a publicly available resource concatenating multiple omics datasets aiming at harmonizing preprocessing steps to make comparisons easier between these datasets. For the per-cohort OS prediction task, we focused on the cohorts with less than 90% of patients censored and selected the indications with more than 400 patients with associated RNA-seq samples (BRCA, UCEC, KIRC, HNSC, LUAD, LGG, LUSC, SKCM, COAD, STAD and BLCA). For the pan-cancer task, we selected the 33 cohorts of TCGA, for a total of 10,736 samples. For both OS prediction tasks, we used the OS labels defined by the TCGA Pan-Cancer Clinical Data Resource[9] (TCGA-CDR).

The Gene Essentiality prediction experiments were carried out using the CCLE dataset, which contains the expression profiles of hundreds of human cancer cell lines. The dataset was obtained through the DepMap portal, version 22Q4[47], which provided the gene perturbation experiments results used as labels. Expression data (raw and TPM pseudo-counts) and gene signatures files were downloaded directly from the platform and non-transformed TPM data was obtained by applying a reverse transformation of the pseudo-counts file. We used the same selection process as previous studies[3] to define gene dependencies of interest (DepOI) and filtered out cell lines with missing dependencies. This resulted in 1223 DepOIs and 893 cell lines for our experiments, for a total of 1,092,139 combinations with an associated gene essentiality score. All datasets used for this study are available for download in the GitHub repository.

Tasks description

The different representation architectures were evaluated by the performance on the prediction of patients or cell lines labels when using the output of baseline representation models or of DRL ones as input features. These downstream tasks were evaluated on several test sets sampled from the datasets following a repeated holdout cross-validation process (see related section below for more details).

The first task we considered was cancer-specific OS prediction, a common task in cancer biology that can improve the identification of subgroups of patients at risk. In this task, representation and prediction models were trained on single cancer cohorts separately to predict the survival of patients within a particular cancer subtype. Our second task consisted in pan-cancer OS prediction, which is often considered in the literature as a standard for comparing methods and claim state of the art performance of newly developed machine learning models[2]. The models are trained on all the cohorts from TCGA combined in a single dataset, evaluating the capabilities to predict the different survival times of a more heterogeneous population suffering from different indications. Harrell C-index[49] was used as the final metric for both tasks and is referred to as c-index in the paper. The different data splits were stratified to ensure similar censorship levels between train and test sets.

The gene essentiality prediction task described in this paper was inspired by the DeepDEP framework[3], with the modification of predicting gene dependency instead of gene effect as recommended in previous work[3,50] and focusing only on bulk RNA-seq rather than a multi-modal setting. In this context, gene dependency refers to the extent to which a particular gene is necessary for cell survival and growth, which depends on which cell line is considered. Each cell line is represented by its expression data while DepOIs are represented by vectors called fingerprints. Fingerprints are specific encodings that summarize the relevant biological functions of a given gene and represent its involvement in 3115 gene signatures related to chemical and genetic perturbation defined in MSigDBv6.2[51]. We focused only on the cell line representation based on RNA-seq data in our benchmark while the fingerprints were fixed and reduced to 500 dimensions using PCA, an arbitrary choice to balance between computation speed, performance and variance explained (Supplementary Fig. S8). The final dataset for the downstream prediction models is the cartesian product of the cell lines representation dataset and the DepOI fingerprints dataset: for each cell line / gene combination with a gene dependency label, we created the corresponding features by concatenating the cell line representation and the fingerprints representation. To evaluate model performances, we used the Spearman rank correlation. While in the original DeepDEP paper Pearson correlation coefficients were computed, both methods are suited to evaluate this task[52,53]. The Spearman correlation was preferred as it is more robust to outliers on non-normally distributed data[54] and fits better the framework of ranking genes for target selection. The correlation was computed on the whole test set and is referred to as the overall correlation. Following DeepDEP's evaluation, we computed as well a per-DepOI metric: for each gene with available experiments, we calculated the Spearman rank correlation over the different cell lines and averaged the result over all the genes to have one final metric per repetition. We partitioned the data by cell lines, ensuring that a cell line did not overlap between the training and test sets.

Preprocessing datasets and gene selection

The RNA-Seq data was preprocessed following a classical bioinformatics pipeline. While different preprocessing options were tested (data and results not shown) as they can impact downstream analyses[55], we decided to choose a fixed standard choice to keep a reasonable computational budget and compare models all else being equal. For all the experiments without pretraining on external datasets, we selected the 5000 most variable genes on the training sets, applied a logarithmic operation and normalized the data with mean-standard scaling. The exact same transformation was then applied to the test dataset to prevent any potential leak by computing features statistics on the whole dataset. All models used TPM normalized expression data except the variational auto-encoder model model which used raw counts to fit a negative binomial prior distribution (*cf Methods, Representation Models*).

When considering pretraining, we implemented another preprocessing process closer to recent pretraining strategies for foundation models (FM) trained on single-cell expression data (scRNA-seq)[26–28]. While these models can afford little feature selection during pretraining thanks to the vast amount of available data, pretraining on size-limited bulk RNA-seq datasets still requires gene selection to avoid the curse of dimensionality. However, the most variable genes for the pretraining datasets can be different from those of the datasets used to evaluate on the downstream tasks for non-pretrained models. To select genes that were still relevant for the tasks considered in this paper, we built two gene lists based on their variances in the TCGA pan-cancer dataset used for pretraining:

- In the case of the per-cohort OS prediction task, the previous procedure has to be adapted in order to ensure a fair comparison with the non-pre-trained case. We therefore took the union of the top 2,000 most variable genes for each of the 11 cohorts selected in the downstream task to make sure relevant genes per indication were also selected in the final features rather than genes solely linked to cancer type that would be considered when looking at the genes' variances across the whole TCGA dataset. This resulted in a set of 5046 unique gene identifiers.
- For the gene essentiality task, we took the top 5000 most variable genes in TCGA after intersection with the genes present within the CCLE dataset, similarly to DeepDEP's procedure of selecting genes with a standard deviation superior to 1 in TCGA.

The TCGA data used for pretraining was normalized within each fold with mean standard scaling and learned statistics were saved for potential usage on the downstream datasets (pretraining experiments).

Repeated holdout cross-validation framework

In this study, we aim to compare the performance of different representation learning algorithms on downstream survival and gene essentiality prediction tasks using bulk RNA-seq data. Each representation model is trained and used to transform the input expression data before feeding the learned low-dimensional embeddings to a task-specific prediction model, fitted for each representation model tested. To achieve a comprehensive evaluation, we adopt a validation pipeline that focuses on exploring the learning algorithm's variability to diverse hyperparameter settings. Our validation pipeline involves a repeated holdout cross-validation approach[45] in which the dataset is repeatedly split in two to create pairs of training and test sets, comprising 80% and 20% of the original data respectively (Supplementary Fig. S1). For experiments without pretraining, the training sets are used to select jointly the optimal HPs for the representation and prediction models by performing a fivefold cross-validation for a given set of HPs. The HP tuning is performed using a Tree-structured Parzen Estimator (TPE Sampler) implemented by Optuna[56] with a fixed budget of 50 iterations. Then, we select the set of HPs with the best average performance over the validation folds on the downstream tasks to train the representation and prediction models on the whole training set before evaluating it on the test set. This procedure is repeated 10 times to generate a distribution of scores over the different test sets, providing robust performance assessments compared to a single test evaluation.

For pretraining experiments, we performed for each task HP tuning of the pre-trained representation models separately from the prediction models. We used the TPE Sampler with 50 trials to find the architecture choices and regularization factors (Supplementary Table S8) that minimized the reconstruction loss of the auto-encoders on TCGA data. For each trial, a fivefold cross-validation was repeated 5 times to estimate the generalization error by averaging the scores obtained per fold. For the gene essentiality prediction task, all of the 33 cohorts were used during pretraining (for a total of 10,736 samples in the pretraining dataset) while we removed the 11 cohorts of the downstream tasks for the per-cohort OS prediction task (4480 samples in the pretraining dataset).

Acceptance testing and model scoring

Limitations of traditional null-hypothesis significance testing and the cross-validation framework to derive statistical conclusions has been critically examined in previous work[45]. They highlight the unsuitability of standard null-hypothesis significance testing, including *t*-tests, for cross-validation due to the non-independence of runs and the complexity of deriving confidence intervals. Consequently, they advocate for a repeated holdout framework as an alternative to cross-validation.

While it is feasible to derive confidence intervals from the repeated holdout framework, the authors still recommend a different scoring system. They argue that traditional hypothesis testing primarily focuses on the statistical significance of expected improvements in models over an infinite population. This approach, however, is not applicable to studies that concentrate on practical, meaningful improvements on finite-sized test sets. Therefore, following the authors' recommendation, we considered a method superior if it outperforms another 75% of the time.

We proposed a scoring system for all models considered for a given task based on the aforementioned criterion. To evaluate the relative importance of each model, a score was generated by counting the number of significant pairwise comparisons won against other models. When different cohorts were considered for the same task, an average score was obtained from the cohort-specific scores. This value was then normalized by the maximum value obtained by a model on the task, resulting in a score between 0 and 1. This score was then converted into a percentage to rank the different models.

Representation models

We selected state-of-the-art methods from various subfields of DRL: linear models as baselines (Identity, PCA), auto-encoders-like models (AE, VAE), Self-Supervised Learning methods (Masking Auto-Encoders), Semi-Supervised methods (Multi-Head Auto-Encoders), data augmentation techniques coupled with AE, graph-based methods leveraging prior-knowledge (GNN) and training frameworks (pretraining). For each of these categories, one base model was implemented, and we considered the different architecture choices as HPs sweeps in our pipeline as described above. Details about the architectures and HPs ranges (Supplementary Tables S8, S9) are described in Supplementary Materials.

Baselines

We considered as baseline representations the Identity and Principal Component Analysis (PCA), widely used for all considered tasks on RNA-seq data. Identity refers to not transforming the input TPM counts and passing the 5000 genes as features to the prediction models directly. The number of components of the PCA was considered a hyperparameter and optimized per fold in our pipeline (Supplementary Table S8).

Auto-encoders

An auto-encoder (AE) is a type of neural network that learns to encode input data into a lower-dimensional representation and then decode it back to the original form, with the goal of reconstructing the input accurately[57]. Given the limited size of our datasets, we focused on small architectures ranging from 0 to 2 hidden layers for the encoder (excluding the representation layer) with 256 to 1024 neurons (more details available in Supplementary Table S8). The decoder part was constructed symmetrically to the encoder. We optimized the AE models using a Mean Squared Error (MSE) Loss and the Adam algorithm[58] and applied rectified linear unit activation

(ReLU) functions between each linear layer. We did the same for other representation models derived from this architecture (MAE, MHAE, DA-GN, PreAE).

Variational auto-encoders: scVI
A variational auto-encoder (VAE) is a type of auto-encoder that incorporates a probabilistic approach, using encoder and decoder networks to generate latent variables and enable sampling from a learned distribution, allowing for generative modeling and capturing underlying data distributions[59].

Specifically, we adapted a popular method to embed scRNA-seq datasets, scVI[38] to bulk RNA-seq data, and assessed the learned representations on our benchmark tasks. scVI is based on a hierarchical Bayesian model where conditional distributions are parametrized by neural networks. As in the VAE, each gene expression profile is encoded through a non-linear transformation into a low dimensional latent vector. This latent representation is then decoded by another non-linear transformation to generate an estimate of the parameters of a Zero-Inflated Negative Binomial. Since bulk RNA-seq data typically does not fit the zero-inflation assumption observed in scRNA-seq, we parameterized scVI's decoder with a Negative Binomial distribution instead.

Masking auto-encoders
Masking in self-supervised learning refers to the process of randomly hiding or removing portions of input data, forcing the model to learn to reconstruct the missing parts, which promotes the discovery of meaningful features and representations. VIME, or Value Imputation and Mask Estimation is a popular masking method for tabular data[40]. In this self-supervised learning framework for tabular data, two pretext tasks are introduced: feature vector estimation and mask vector estimation. The encoder function is trained together with two pretext predictive models to reconstruct a corrupted input sample (feature vector estimation) with MSE loss and estimate the mask vector used to corrupt the sample (mask vector estimation) with a cross-entropy loss. The resulting learned representation captures correlations across different parts of the data, making it useful for downstream tasks. We focused on one re-implementation from the main paper which is similar to the original method where we only included one pretext task: feature vector estimation.

Multi-head auto-encoders
A multi-head auto-encoder (MHAE) includes one or more auxiliary heads to perform supervised tasks during the representation model training. These auxiliary heads are fully-connected neural networks that take the compressed representation as inputs, and predict labels such as overall survival and gene essentiality as outputs. In this study we considered MH auto-encoders trained with an auxiliary head predicting the label of the downstream task, to assess if adding supervision would improve the performance of the auto-encoder.

Graphs neural networks
We used a Graph Neural Network (GNN) architecture to incorporate protein–protein interaction networks (PPIs) as a source of prior knowledge. The STRING PPI database served as the underlying graph on which the bulk RNA-seq data was laid. Each node of the graph represented a gene, with gene expression as a node feature. In this setting, each patient (or sample) was represented as a single graph, and the graph topology over the samples did not vary, only the overlaying signal did. This model is close to the traditional convolutional neural network alternating between convolution and pooling steps but using a graph instead of e.g. pixel coordinates to perform message passing to neighboring features.

Previous studies showcased this kind of model for survival prediction[36,60], and we propose here a modified version intended for representation learning. More details about the implementation are given in the Supplementary Materials.

Data augmentation
Data augmentation is a set of techniques used to reduce overfitting of machine learning models by generating new training data points from the original ones. It is ubiquitous in computer vision, where new images can be obtained by simple transformations such as rotations or cropping[61]. In the case of omics data, there are no obvious equivalents of such transformations. However, certain methods have been successfully applied to single-cell RNA-seq data in the context of contrastive learning by adding noise and simulating dropout events[24]. Specifically, we focused on using data augmentation based on Gaussian Noise (DA GN). The method consists in creating additional samples by copying the data and adding gaussian noise to all the features after they have been standardized with mean and standard normalization. We used a centered normal distribution with standard deviation controlled by a hyperparameter and fixed the number of copies to 4 corrupted samples for one original sample. The target labels for these new data points are simply copied from the original dataset.

Pretraining experiments
The use of bulk RNA-seq data to train representation algorithms is often hindered by the limited number of patients who were screened for a specific condition. To address this issue, pretraining models on larger datasets can help represent the data more accurately. In this study, we tried two different pretraining strategies for both the per-cohort survival prediction and the gene essentiality prediction tasks. In order to better assess the effect of pretraining and decouple it from architecture details, we focused our pretraining experiments on the AE model.

For both downstream tasks, we compared the original AE to two different pretraining strategies, **PreAE** and **PreAE finetuned.** In the case of PreAE, the mean-standard scaling was done using solely the statistics of the TCGA pretraining dataset (as described above, all 33 TCGA cohorts were used for pretraining for gene essentiality prediction task, while we removed the 11 cohorts of the downstream tasks for the per-cohort OS prediction task) and the auto-encoder was used only for inference on the task dataset. For the PreAE finetuned strategy, the scaling was performed using the learned statistics on the training folds of the task dataset while the auto-encoder was also fine-tuned for each fold. These former experiments assess how initializing the weights with pretraining can help on the downstream tasks. Both optimization histories for the pretraining are available in (Supplementary Fig. S9, Supplementary Fig. S10).

Downstream prediction models

For survival prediction tasks, each representation model was combined with a multi-layer perceptron (MLP) optimized with a differentiable Cox loss[62,63] using the Adam algorithm. 20% within each training fold were used for early stopping to prevent overfitting of the MLP. The choice of using MLP rather than Cox linear models was motivated by faster computation times thanks to GPU usage with PyTorch and by the need of a fair comparison with DL models that use nonlinear prediction heads. More details about the hyperparameters of the prediction model are available in the Supplementary Materials (Supplementary Table S9).

For Gene Essentiality a light gradient-boosting machine (LGBM) regressor was trained to predict the essentiality (dependency score) of each gene in a cell line based on the features described in the task section. To train the LGBM regressors effectively, two HPs were considered: the learning rate and the regularization coefficient alpha. The learning rate was set in the range of 0.01 to 0.3 and the regularization coefficient within a range of 0 to 100 for exploring various regularization strengths. Notice that we initially tested both MLP (the choice in DeepDEP paper) and LGBM but ended up keeping the best performing model (Supplementary Table 10).

Data availability

The cancer TCGA data was downloaded from recount3 https://rna.recount.bio/, the associated clinical data from TCGA-CDR hosted on https://gdc.cancer.gov and the cell lines datasets from the DepMap portal https://depmap.org/portal/. The code and the processed data used in our study are available on GitHub: https://github.com/owkin/drl-evaluation

Received: 15 April 2024; Accepted: 8 July 2024
Published online: 24 July 2024

References

1. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: The teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).
2. Vale-Silva, L. A. & Rohr, K. Long-term cancer survival prediction using multimodal deep learning. *Sci. Rep.* **11**, 13505 (2021).
3. Chiu, Y.-C. *et al.* Predicting and characterizing a cancer dependency map of tumors with deep learning. *Sci. Adv.* **7**, eabh1275 (2021).
4. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
5. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (Association for Computational Linguistics, 2019). https://doi.org/10.18653/v1/N19-1423.
6. Misra, I. & Van Der Maaten, L. Self-Supervised Learning of Pretext-Invariant Representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6706–6716 (IEEE, 2020). https://doi.org/10.1109/CVPR42600.2020.00674.
7. Chaudhary, K., Poirion, O. B., Lu, L. & Garmire, L. X. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* **24**, 1248–1259 (2018).
8. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on tabular data?. *Mach. Learn.* https://doi.org/10.48550/ARXIV.2207.08815 (2022).
9. Liu, J. *et al.* An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400-416.e11 (2018).
10. Gönen, M. *et al.* A community challenge for inferring genetic predictors of gene essentialities through analysis of a functional screen of cancer cell lines. *Cell Syst.* **5**, 485-497.e3 (2017).
11. Zhakparov, D. et al. Assessing different feature selection methods applied to a bulk RNA sequencing dataset with regard to biomedical relevance, https://doi.org/10.3929/ETHZ-B-000565782 (2023).
12. Liu, Y. *et al.* Post-modified non-negative matrix factorization for deconvoluting the gene expression profiles of specific cell types from heterogeneous clinical samples based on RNA-sequencing data. *J. Chemom.* **32**, e2929 (2018).
13. Chen, R. *et al.* Large-scale bulk RNA-seq analysis defines immune evasion mechanism related to mast cell in gliomas. *Front. Immunol.* **13**, 914001 (2022).
14. Wei, Q. *et al.* Molecular subtypes of lung adenocarcinoma patients for prognosis and therapeutic response prediction with machine learning on 13 programmed cell death patterns. *J. Cancer Res. Clin. Oncol.* **149**, 11351–11368 (2023).
15. Sauta, E. *et al.* Combining gene mutation with transcriptomic data improves outcome prediction in myelodysplastic syndromes. *Blood* **142**, 1863–1863 (2023).
16. Li, Q. *et al.* XA4C: eXplainable representation learning via autoencoders revealing critical genes. *PLoS Comput. Biol.* **19**, e1011476 (2023).
17. De Weerd, H. A. *et al.* Representational learning from healthy multi-tissue human RNA-Seq data such that latent space arithmetics extracts disease modules. *bioRxiv* https://doi.org/10.1101/2023.10.03.560661 (2023).
18. Withnell, E., Zhang, X., Sun, K. & Guo, Y. XOmiVAE: An interpretable deep learning model for cancer classification using high-dimensional omics data. *Brief. Bioinform.* **22**, bbab315 (2021).
19. He, D., Liu, Q., Wu, Y. & Xie, L. A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening. *Nat. Mach. Intell.* **4**, 879–892 (2022).
20. Chen, J. *et al.* Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. *Nat. Commun.* **13**, 6494 (2022).
21. Dincer, A. B., Celik, S., Hiranuma, N. & Lee, S.-I. DeepProfile: Deep learning of cancer molecular profiles for precision medicine. *bioRxiv* https://doi.org/10.1101/278739 (2018).
22. Rampášek, L., Hidru, D., Smirnov, P., Haibe-Kains, B. & Goldenberg, A. Dr.VAE: Improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics* **35**, 3743–3751 (2019).

23. Shen, H. *et al.* Miscell: An efficient self-supervised learning approach for dissecting single-cell transcriptome. *iScience* **24**, 103200 (2021).
24. Han, W. *et al.* Self-supervised contrastive learning for integrative single cell RNA-Seq data analysis. *bioRxiv* https://doi.org/10.1101/2021.07.26.453730v1 (2021).
25. Li, X. *et al.* Network embedding-based representation learning for single cell RNA-seq data. *Nucleic Acids Res.* **45**, e166 (2017).
26. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
27. Cui, H. *et al.* scGPT: Towards building a foundation model for single-cell multi-omics using generative AI. *bioRxiv* https://doi.org/10.1101/2023.04.30.538439 (2023).
28. Shen, H. *et al.* Generative pretraining from large-scale transcriptomes for single-cell deciphering. *iScience* **26**, 106536 (2023).
29. Smith, A. M. *et al.* Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinform.* **21**, 119 (2020).
30. Cantini, L. *et al.* Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.* **12**, 124 (2021).
31. Bengio, Y. & Grandvalet, Y. No unbiased estimator of the variance of K-fold cross-validation. In *Advances in Neural Information Processing Systems* Vol. 16 (eds Thrun, S. *et al.*) (MIT Press, 2003).
32. Nadeau, C. & Bengio, Y. Inference for the generalization error. *Mach. Learn.* **52**, 239–281 (2003).
33. Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-021-00434-9 (2021).
34. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255, https://doi.org/10.1109/CVPR.2009.5206848 (2009).
35. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins Struct. Funct. Bioinform.* **89**, 1607–1617 (2021).
36. Althubaiti, S. *et al.* DeepMOCCA: A pan-cancer prognostic model identifies personalized prognostic markers through graph attention and multi-omics data integration. *bioRxiv* https://doi.org/10.1101/2021.03.02.433454 (2021).
37. Zhang, X., Xing, Y., Sun, K. & Guo, Y. OmiEmbed: A unified multi-task deep learning framework for multi-omics data. *Cancers* **13**, 3047 (2021).
38. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
39. Fang, Z., Zheng, R. & Li, M. scMAE: A masked autoencoder for single-cell RNA-seq clustering. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btae020 (2024).
40. Yoon, J., Zhang, Y., Jordon, J. & van der Schaar, M. VIME: Extending the success of self- and semi-supervised learning to tabular domain. In *Proc. of the 34th International Conference on Neural Information Processing Systems* (Curran Associates Inc., 2020).
41. Arslan, M., Guzel, M., Demirci, M. & Ozdemir, S. SMOTE and Gaussian noise based sensor data augmentation. In *2019 4th International Conference on Computer Science and Engineering (UBMK)*, 1–5 (IEEE, 2019). https://doi.org/10.1109/UBMK.2019.8907003.
42. Huang, Z. *et al.* Deep learning-based cancer survival prognosis from RNA-seq data: Approaches and evaluations. *BMC Med. Genom.* **13**, 41 (2020).
43. Multiple Myeloma DREAM Consortium *et al.* Multiple myeloma DREAM challenge reveals epigenetic regulator PHF19 as marker of aggressive disease. *Leukemia* **34**, 1866–1874 (2020).
44. Filiot, A. *et al.* Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv* https://doi.org/10.1101/2023.07.21.23292757 (2023).
45. Varoquaux, G. & Colliot, O. Evaluating machine learning models and their diagnostic value. In *Machine Learning for Brain Disorders* Vol. 197 (ed. Colliot, O.) 601–630 (Springer US, 2023).
46. Lachmann, A. *et al.* Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9**, 1366 (2018).
47. Barretina, J. *et al.* The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
48. Wilks, C. *et al.* recount3: Summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol.* **22**, 323 (2021).
49. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
50. Dempster, J. M. *et al.* Extracting biological insights from the project Achilles genome-scale CRISPR screens in cancer cell lines. *bioRxiv* https://doi.org/10.1101/720243 (2019).
51. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
52. Rosenski, J., Shifman, S. & Kaplan, T. Predicting gene knockout effects from expression data. *BMC Med. Genom.* **16**, 26 (2023).
53. Ma, J. *et al.* Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat. Cancer* **2**, 233–244 (2021).
54. Hou, J. *et al.* Distance correlation application to gene co-expression network analysis. *BMC Bioinform.* **23**, 81 (2022).
55. Paton, V. *et al.* Assessing the impact of transcriptomics data analysis pipelines on downstream functional enrichment results. *bioRxiv* https://doi.org/10.1101/2023.09.13.557538 (2023).
56. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. *Mach. Learn.* https://doi.org/10.48550/ARXIV.1907.10902 (2019).
57. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
58. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Preprint at http://arxiv.org/abs/1412.6980 (2017).
59. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. *Mach. Learn.* https://doi.org/10.48550/ARXIV.1312.6114 (2013).
60. Ramirez, R. *et al.* Prediction and interpretation of cancer survival using graph convolution neural networks. *Methods* **192**, 120–130 (2021).
61. Perez, L. & Wang, J. The effectiveness of data augmentation in image classification using deep learning. *Comput. Vis. Pattern Recognit.* https://doi.org/10.48550/ARXIV.1712.04621 (2017).
62. Faraggi, D. & Simon, R. A neural network model for survival data. *Stat. Med.* **14**, 73–82 (1995).
63. Katzman, J. *et al.* DeepSurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 24 (2018).

Acknowledgements

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. We thank Oussama Tchita, Omar Darwiche Domingues and Thomas Chaigneau for their valuable contributions and comments to strengthen our pipeline and coding best practices. We thank Gilles Wainrib for initial ideas and discussions, Nicolas Loiseau for his advice and statistical expertise, Floriane Montanari, Benoît Schmauch, Gilles Wainrib and Jean-Philippe Vert for their detailed proofreading and insightful comments.

Author contributions

B.G., An.D., V.C, V.K., M.B., Y.D and A.R designed the evaluation framework and the different tasks; An.D, B.G, V.K, G.D, V.C, J.E.K, K.O, S.G., Al.D, L.H., R.G., R.L and C.E wrote the code to develop the evaluation pipeline; B.G, An.D, V.C, V.K, J.E.K, G.D, S.G, K.O., Al.D, R.L, C.E, Y.D, A.R analyzed the results; Y.D., A.R. and E.D. coordinated and supervised the work; B.G., An.D., V.C, Y.D, E.D. and A.R wrote the paper with the assistance and feedback of all the other co-authors. All authors reviewed and approved the final manuscript.

Funding

This study has been funded by Owkin, Inc., New York, NY, USA.

Competing interests

All authors are employees of Owkin, Inc., New York, NY, USA.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-67023-8.

Correspondence and requests for materials should be addressed to B.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.