JASMINE: A powerful representation learning method for enhanced analysis of incomplete multi-omics data

Jenna L. Ballard[1], Zongyu Dai[2], Li Shen[3*], Qi Long[3*],
Alzheimer's Disease Neuroimaging Initiative[†],
Alzheimer's Disease Metabolomics Consortium[††]

[1]Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, 19104, PA, USA.
[2]Graduate Group in Applied Mathematics and Computational Science, University of Pennsylvania, 209 S. 33rd Street, Philadelphia, 19104, PA, USA.
[3*]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Drive, Philadelphia, 19104, PA, USA.

*Corresponding author(s). E-mail(s): li.shen@pennmedicine.upenn.edu; qlong@pennmedicine.upenn.edu;
Contributing authors: jenna.ballard@pennmedicine.upenn.edu; daizy@sas.upenn.edu;

[†]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[††]Data used in preparation of this article were generated by the Alzheimer's Disease Metabolomics Consortium (ADMC). As such, the investigators within the ADMC provided data but did not participate in the analysis or writing of this report. A complete listing of ADMC investigators can be found at: https://sites.duke.edu/adnimetab/team/

Abstract

Integrative analysis of multi-omics data provides a more comprehensive and nuanced view of a subject's biological state. However, high-dimensionality and ubiquitous modality missingness present significant analytical challenges. Existing methods for incomplete multi-omics data are scarce, do not fully leverage both modality-specific and shared information, and produce task-biased representations. We propose JASMINE, a self-supervised representation learning method for incomplete multi-omics data that preserves both modality-specific and joint information and enhances sample similarity structure. JASMINE produces embeddings that achieve superior performance across multiple tasks for two different incomplete multi-omics datasets while requiring only a single round of training per dataset.

Keywords: representation learning, variational autoencoder, multimodal integration, missing data

1 Background

With advances in data collection technologies, storage and computing resources, vast amounts of -omics data, including genetic variation (genomics), gene regulation (epigenomics), gene expression (transcriptomics), protein abundance (proteomics), and metabolite abundance (metabolomics), are becoming available. Each of these modalities sheds light on various biological processes. Multiple modalities are often measured for the same subject, providing more comprehensive information which may aid downstream analyses and improve performance on predictive tasks. One application of interest is to disease pathophysiology, which is complex and multifaceted, resulting from the interactions between processes across many layers, including the molecular, cellular, tissue, organ, and patient levels. Combining the information offered by diverse -omics data modalities may better capture these processes and their interactions which underlie particular phenotypes. This can enhance diagnosis [1], disease subtyping [2], and discovery of disease-relevant biological pathways [3, 4].

Common obstacles to fully leveraging multi-omics data include high dimensionality and incompleteness: some modalities can comprise tens of thousands of features,

and most datasets contain a considerable proportion of samples with missing modalities. The ability to overcome these challenges necessitates strategies to produce clean, compact representations for each sample that are amenable to downstream analyses and prediction tasks.

Many strategies have been devised to integrate multi-omics data for downstream tasks. These can be broadly categorized as early, intermediate, or late integration. Early-integration combines the modalities as a preprocessing step, such as by concatenating the features across data types and using this as input to the model [5–7]. Late integration applies separate models to each data modality and then aggregates the outputs into a final multi-modal prediction [8–10]. These approaches, which handle individual modalities by either simply concatenating their features before inputting them to a model or by applying separate models, may better preserve modality-specific information [5], but they are limited in their ability to fully learn and exploit inter-modality interactions. On the other hand, intermediate integration strategies recognize the modalities as distinct entities at the input level, and then fuse their information while learning inter-modality relationships [5, 11–15].

Another important consideration for multi-omics integration is the information preserved by the representations. Two principles of interest are consensus, which holds that model errors are upper-bounded by the level of disagreement between modalities, and complementarity, which holds that each modality contains critical unique information [5]. Early- and late-integration approaches tend to better preserve modality-specific information, satisfying the complementarity principle. Intermediate integration-based methods can satisfy both complementarity and consensus [16–18] by including modality-specific and shared components, respectively. Some strategies [16, 18] further ensure that the two components contain unique information by imposing orthogonality constraints between them.

Despite the availability of methods for integrating multi-omics data, most leave the issue of missing modalities unaddressed. A common way to handle this is to exclude samples with incomplete data. However, this reduces the sample size which may lead to reduced power in subsequent analysis; and the remaining subset of subjects may not be representative of the original dataset which may lead to biased results. Other techniques impute the missing data as a preprocessing step, but this may bias the imputed values to be similar, which can negatively impact downstream analyses [19]. Some recent multi-modal integration methods handle incomplete data by defining a joint latent space that can be derived from subsets of modalities [17, 20]. Such techniques include product-of-experts (PoE) and mixture-of-experts (MoE), which define the joint latent space as the product [13] and weighted average [21], respectively, of the modality-specific latent distributions. When modalities are missing, they are excluded from the calculation of the joint embedding distribution. Other methods can handle missing data by generating missing modalities from those that are available via cross-encoders [12] or enforcing similarity among the embeddings derived from all modalities [22]. A more recent method, IntegrAO [15], learns omics-specific patient graphs before fusing them to generate low dimensional embeddings in a unified space while using cross-patient similarities to help infer missing information. However, this graph-based method is memory intensive, becoming more prohibitive for both training and inference as sample size increases.

Other limitations of prior methods is that they are often supervised [20, 23–26], resulting in representations that are biased towards a specific task, and requiring re-training for different tasks. The goal of our method is to generate informative, compact representations of multi-omics data that attain superior performance on a variety of downstream tasks. Using a self-supervised approach based on reconstructing the original data from the low-dimensional representations trains the method to preserve the most relevant information in the data while reducing noise. Supervised approaches

may cause the model to ignore important information that is not relevant to the specific prediction task for which they are being trained. By training without a particular prediction task objective, our approach produces task-agnostic representations that excel at multiple supervised tasks while also retaining flexibility for unsupervised tasks such as clustering.

To accomplish these goals, we developed JASMINE (**J**oint **A**nd modality-**S**pecific **M**ultimodal representation learning handling **IN**compl**E**te data), a self-supervised representation learning method that generates compact, task-agnostic embeddings that integrate multi-omics data while handling arbitrarily missing modalities. The main contributions of this work are the following:

1. Modality-specific and joint components preserve both complementary and shared information among the modalities. Orthogonality constraints minimize redundancy among these components.
2. Two types of contrastive learning (CL), at the modality level and sample level, enhance the discriminative structure of the data while reducing irrelevant noise. Modality-level CL enforces consistency among the modalities for a given sample, and sample-level CL utilizes a deep clustering-based approach to ensure that similarity relationships between samples in the original feature space are preserved in the embedding space.
3. Samples with arbitrarily missing modalities are handled using both cross-encoders for modality-specific components and PoE for the joint component of the representation.
4. Self-supervised approach produces one set of representations that can be used for multiple different downstream tasks after a single round of training per dataset.
5. Simulation experiments demonstrated JASMINE's suitability compared to baseline methods for cases when modalities are missing at a high rate, especially as the

number of modalities increases and sample size is limited, and when missingness is unbalanced among ground truth classes.

6. Representations achieved best or close to the best performance on multiple prediction tasks across two different incomplete multi-omics datasets.

2 Results

2.1 JASMINE Framework

We present JASMINE, a variational autoencoder-based framework to learn representations of multi-omics data while handling missing modalities. Training is done in a self-supervised manner, resulting in representations that can be used for a variety of downstream tasks. By incorporating self- and cross-encoders (see Section 5.2), the method preserves modality-specific information while learning pairwise cross-modality relationships. To capture information shared among all modalities, the method additionally learns a joint representation using PoE (see Section 5.2.1), which defines the joint latent space as the product of the modality-specific distributions. The final representation is formed from concatenating the modality-specific and joint representations (see Section 5.2.2). A schematic of JASMINE is presented in Figure 1. Our proposed approach further uses orthogonality constraints to minimize the redundancy of information between the modality-specific and shared components (see Section 5.2.3). It also utilizes modality-level contrastive learning (MCL) and sample-level contrastive learning (SCL) to encourage consistency among embeddings of different modalities from the same sample as well as preserve the similarity structure between samples in the embedding space (see Section 5.2.4). The final objective function comprises loss terms for the encoders, contrastive learning, and orthogonality constraints (see Section 5.2.5). Thus, our approach aims to generate informative representations that preserve both modality-specific and shared information while enhancing the similarity structure among samples. We evaluate the efficacy of the representations generated by our

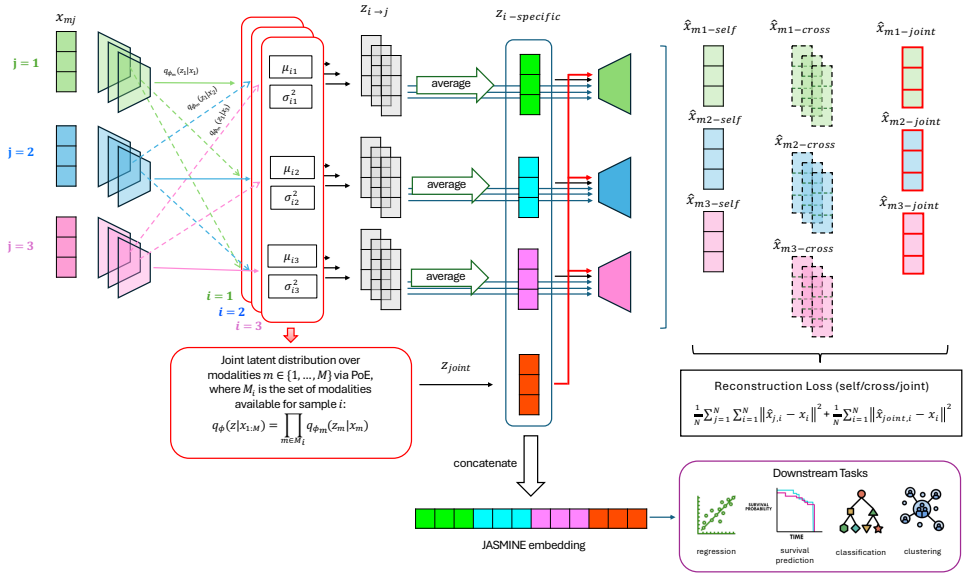


Fig. 1 JASMINE Method Schematic. For data with m modalities, the model consists of m^2 self- and cross- encoders which generate modality-specific representations. To get a single modality-specific representation for a given modality, the representations generated for that modality are averaged. A joint latent space is derived using PoE, from which a joint embedding is generated. Modality-specific and joint representations are concatenated to produce the final JASMINE embedding. These embeddings can be used as input for any desired downstream task (e.g. regression, classification, clustering). JASMINE is learned in a self-supervised fashion based on reconstruction loss. There is one decoder per modality for a total of m decoders, which are used to reconstruct each modality from its corresponding latent representation.

model on classification, clustering and regression tasks in simulated datasets and two different real datasets.

2.2 Experimental Setup

Our experimental pipeline consists of two main steps: (1) generate embeddings using an unsupervised representation learning method and (2) input the embeddings to a prediction model of choice for a given inference problem. In this way, the learned embeddings can be applied to various downstream tasks. More details on the data processing for this pipeline are given in Section 5.6. Here, we apply the embeddings to three types of prediction problems: (1) survival prediction using a Coxnet regression

model; (2) classification using an XGBoost classification model (XGBClassifier); and (3) regression using an XGBoost regression model (XGBRegressor).

We compare JASMINE to multiple baseline embedding frameworks. The baselines include the original data, deep canonical correlation analysis (DCCA) [27], generalized canonical correlation analysis (GCCA) [28], multi-modal variational autoencoder (MVAE) [13], cross-linked unified embedding (CLUE) [12], and IntegrAO [15]. Details about these baseline methods can be found in Section 5.3. A summary of the key features of these methods can be found in Table 1. We evaluate and compare these methods based on various performance metrics. For survival prediction, we use the concordance index for right-censored data based on inverse probability of censoring weights (C-index IPCW) and cumulative dynamic area under the receiver operating characteristic curve (CD AUC). For classification, we use balanced accuracy, area under the receiver operating characteristic curve (AUROC) and average precision score (AP). In the multiclass setting, the macro-average across classes is utilized. For clustering, we assess differential survival via the log-rank test over various numbers of clusters and for five different cancer types. For regression, we used root mean squared error (RMSE), the coefficient of determination (R^2-score), and the Pearson correlation between the predicted and true values (r-pred). More details about these metrics can be found in Section 5.4.

Table 1 Proposed and Baseline Method Summary

Method	# of modalities handled	Handles missing modalities	Linearity	Captures joint modal info	Captures modality-specific info
Original Data	any	no	linear		✓
DCCA	2	no	nonlinear	✓	
GCCA	any	no	linear	✓	
MVAE	any	yes	nonlinear	✓	
CLUE	any	yes	nonlinear		✓
IntegrAO	any	yes	nonlinear	✓	
JASMINE	any	yes	nonlinear	✓	✓

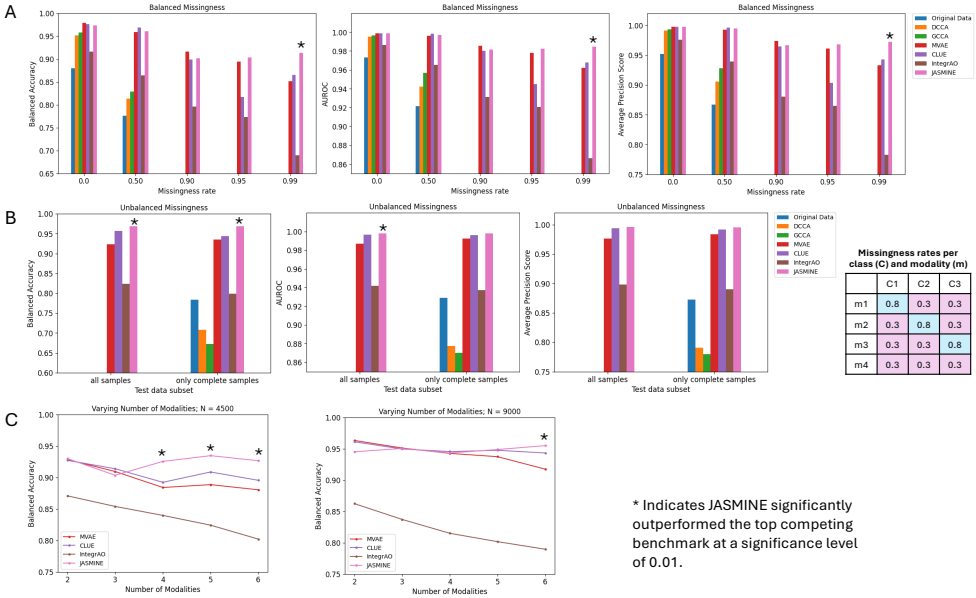
2.3 Evaluation on simulated data

To test the performance of JASMINE under a variety of controlled data settings, we evaluated and compared the performance of JASMINE and the baselines on simulated data. This included varying the rate at which modalities were missing, evaluating the performance when missingness rates depended on the true label, and varying the number of modalities under small and large sample sizes. The synthetic data were simulated to comprise three distinct classes, each with 3,000 samples for a total sample size of 9,000. Each sample had 300 informative features per modality that distinguished the classes as well as an additional 1800 uninformative features per modality which served as random noise. Of the informative features, 30% were simulated to be modality-shared, with a correlation of 0.3 between corresponding features of different modalities; the remaining features had a correlation of 0 between modalities, representing modality-specific information. To simulate missingness, masks were generated under various missingness scenarios to indicate which samples had each modality available. More details on synthetic data generation are available in Section 5.6.1.

After generating representations using each of the embedding methods, an XGBoost Classifier model was trained to predict the class labels. Performance was evaluated via balanced accuracy, AUROC, and AP. These simulation experiments enabled us to determine under which settings JASMINE outperforms the other methods.

2.3.1 JASMINE is particularly suited for handling a high degree of missing modalities and unbalanced missingness patterns

We first tested the effect of modality missingness rate in the four-modality setting by varying the probability that a given sample is missing a particular modality from 0 to 0.99. All samples had equal probability of missing a modality, and missingness was determined independently for each modality. The results are shown in Figure 2A. At high missingness rates of 0.95 and 0.99, JASMINE outperformed all other methods on



	C1	C2	C3
m1	0.8	0.3	0.3
m2	0.3	0.8	0.3
m3	0.3	0.3	0.8
m4	0.3	0.3	0.3

Fig. 2 Simulation experiment results. (A) Balanced missingness setting. Missingness rate was varied from 0 to 0.99 and was the same for all samples and all modalities. For rates 0.90-0.99, methods that do not handle missing data could not be applied due to insufficient samples with all modalities. (B) Unbalanced missingness setting. Missingness rate varied for each modality based on true label. Missingness rates for each modality (denoted m# in the rows) in each class (denoted C# in the columns) are shown in the table on the right. (C) Varying number of modalities. Sparse modalities were added one-by-one, with each additional modality having a missingness rate of 0.95. The values plotted in all graphs represent mean performance over five replicates. Asterisk indicates that JASMINE had significantly better (higher) performance than the second best baseline method at a significance level of 0.01.

all three performance metrics. At lower missingness rates, JASMINE slightly underperformed MVAE and CLUE, but as missingness rate increased to high levels, MVAE and CLUE showed a more considerable decline in performance. JASMINE, MVAE, and CLUE had the top performance in all cases. DCCA and GCCA, which do not handle missing data, outperformed IntegrAO when no samples had missing modalities, but at a missingness rate of 0.5, IntegrAO had better performance, likely due to its ability to handle missingness. Due to their inability to handle incomplete modalities, the original data, DCCA, and GCCA results were not reported for missingness rates above 0.90 due to insufficient complete samples.

We also evaluated performance of these methods when missingness was unbalanced across the classes in the four-modality setting. In this case, the probability of missing a particular modality varied based on a sample's true label. For example, samples in class 1 were more likely to be missing the first modality, samples in class 2 were more likely to be missing the second modality, and samples in class 3 were more likely to be missing the third modality. Here, when a modality was 'more likely' to be missing, its missingness rate was 0.8; otherwise, its missingness rate was 0.3. A table displaying the missingness rates for each modality in each class is shown in Figure 2B. We wanted to determine whether systematic class-based differences in modality missingness patterns would impact the embeddings and their performance on downstream classification. The results shown in Figure 2B demonstrate that JASMINE achieves the best classification performance across all three metrics under these conditions, both when testing on all samples in the test set as well as only the testing samples for which all modalities were available. The fact that JASMINE achieves superior performance when tested only on the complete samples suggests that its top performance is not solely due to learning class-specific modality missingness patterns. After JASMINE, CLUE and MVAE had the best performance, followed by IntegrAO. The original data, DCCA, and GCCA results were only reported on the complete samples due to their inability to handle missing modalities. Of note is that DCCA and GCCA had worse performance than the original data in the unbalanced missingness setting, whereas in the balanced missingness setting and in the absence of missingness in Figure 2A, DCCA and GCCA outperformed the original data.

In addition to varying the modality missingness, we investigated the effect of adding sparse modalities for large and small sample sizes. Here, we started with all samples having the first modality, and then iteratively added additional modalities with missingness rates of 0.95. The results are shown in Figure 2C. At both sample sizes, performance showed a notable decline for MVAE, CLUE, and IntegrAO as

more modalities were included, whereas it remained more stable for JASMINE. At the smaller sample size of 4,500, this difference in performance was more notable, and JASMINE achieved significantly better performance for 4, 5, and 6 modalities. These patterns were also observed at the larger sample size of 9,000, but performance was more similar among MVAE, CLUE and JASMINE. JASMINE still had significantly better performance when there were 6 modalities.

Overall, these results suggest that JASMINE has particularly strong performance relative to the other methods when modalities have a high degree of missingness and when the number of sparse modalities is higher, especially at smaller sample sizes. Additionally, JASMINE was shown to be able to handle unbalanced modality missingness better than the other methods.

2.4 Evaluation of representations learned from TCGA multi-omics data

We applied our methods to mRNA-seq, methylation, miRNA-seq, and functional proteomics reverse phase protein array (RPPA) data from The Cancer Genome Atlas (TCGA) [29]. This dataset comprises 30 cancer types across 10,472 total samples, several of which are missing one or more modalities. Missingness is blockwise, such that either all features or no features are missing for a given modality. JASMINE serves not only as a representation learning technique, but also a means of dimensionality reduction. Therefore, to test our methods' ability to handle a large number of features, we did not perform any prior feature selection or dimensionality reduction. The summary of sample sizes, dimensionality, and labels are available in Table 2A, and more details about TCGA data processing are in Section 5.6.2.

We apply the embeddings generated by our method to two supervised tasks: survival prediction and cancer stage classification. Both are of practical significance: being

Table 2 Data Summary

A. TCGA						
Modality	**mRNA-seq**	**Methylation**	**miRNA-seq**	**RPPA**	**All modalities**	**$\geq$ 1 modality**
# features	20,390	13,130	119	122		
# samples	9,940	9,136	9,797	7,465	6,054	10,472
Observed survival time (days)	**mean $\pm$ sd**	**min**	**Q1**	**med**	**Q3**	**max**
	895.5 ± 979.1	0	294	575	1,155	10,870
# of censored samples			7,267 (69.4%)			
Cancer stage	**I**	**II**	**III**	**IV**	**Total**	
# samples	2,019	2,145	1,648	818	6,630	

B. ADNI						
Modality	**GEX**	**GEN**	**MET**	**PROT**	**All modalities**	**$\geq$ 1 modality**
# features	49,386	935	148	6,496		
# samples	468	1,573	871	735	230	1,665
Diagnosis	**CN**	**MCI**	**AD**	**Total**		
# samples	504	836	325	1,665		
MMSE	**mean $\pm$ sd**	**min**	**Q1**	**med**	**Q3**	**max**
	27.2 ± 2.6	18	26	28	29	30

GEX = gene expression, GEN = genetics, MET = metabolomics, PROT = proteomics
CN = cognitively normal, MCI = mild cognitive impairment, AD = Alzheimer's dementia

able to better predict a cancer patient's survival can help with treatment and palliative care decisions as well as future planning for the patient and their loved ones. Cancer stage classification can help determine the severity of a patient's condition, aiding in prognosis and treatment decision-making. The learned representations can also be used for unsupervised tasks such as clustering to identify cancer subtypes. To assess their discriminative capabilities, we evaluate their ability to form clusters of patients with differential survival.

2.4.1 JASMINE-generated representations achieved superior performance on survival prediction and cancer stage classification in cancer patients

We used the representations learned by our method to train a Coxnet regression model for survival prediction. The results are shown in Table 3A. JASMINE achieved superior c-index IPCW (0.762) and CD AUC (0.798) compared to the baselines, with a 6.6% improvement in c-index and 3% improvement in CD AUC when compared to the concatenated original data features. Notably, the CCA-based methods had the worst performance, followed by IntegrAO and MVAE, and all four were outperformed by the original data. CLUE performed well, indicating the advantage of preserving modality-specific information. JASMINE displayed further performance gains over CLUE, suggesting that preserving joint and specific information while using contrastive learning and orthogonality constraints is beneficial.

We also used the learned representations as input to a XGBoost classifier to predict cancer stage as I, II, III or IV (see Section 5.6.2 for more details). The results are shown in Table 3B. JASMINE again showed superior balanced accuracy (0.497), AUROC (0.743), and average precision score (0.521). These results suggest that the task-agnostic integrated representations learned by our method achieve the goal of obtaining superior performance across diverse tasks. A further advantage of our

method is that it only needs to be trained once on the data and then generates one set of representations that can be utilized for multiple different downstream tasks.

Table 3 Results on TCGA multi-omics data.

A. Survival prediction results		
Method	C-index IPCW	CD AUC
Original Data	0.717 ± 0.028	0.774 ± 0.014
DCCA	0.565 ± 0.139	0.623 ± 0.133
GCCA	0.529 ± 0.131	0.525 ± 0.146
MVAE	0.692 ± 0.013	0.749 ± 0.008
CLUE	0.739 ± 0.019	0.796 ± 0.005
IntegrAO	0.652 ± 0.016	0.673 ± 0.019
JASMINE	$\mathbf{0.762 \pm 0.013}$	$\mathbf{0.798 \pm 0.005}$

B. Cancer stage classification results			
Method	Balanced Acc	AUROC	AP
Original Data	0.415 ± 0.018	0.683 ± 0.011	0.438 ± 0.008
DCCA	0.459 ± 0.013	0.723 ± 0.006	0.487 ± 0.008
GCCA	0.493 ± 0.004	0.733 ± 0.005	0.506 ± 0.009
MVAE	0.469 ± 0.012	0.728 ± 0.006	0.498 ± 0.011
CLUE	0.488 ± 0.011	0.742 ± 0.006	0.518 ± 0.009
IntegrAO	0.444 ± 0.010	0.703 ± 0.008	0.461 ± 0.015
JASMINE	$\mathbf{0.497 \pm 0.007}$	$\mathbf{0.743 \pm 0.004}$	$\mathbf{0.521 \pm 0.008}$

C-index IPCW is the C-index that accounts for high proportion of censoring in the data. CD AUC is the cumulative dynamic AUC for predicting mortality across a range of thresholds. Results are presented as mean ± sd.
AP is average precision, which summarizes a precision-recall curve as the weighted sum of precisions achieved at multiple thresholds. As this is multi-class classification, AUROC and AP are macro-averaged over each class in a one-vs-rest approach. Results are presented as mean ± sd.
Bold font indicates the best performance.

2.4.2 Clustering JASMINE-derived representations suggests biologically meaningful cancer subtypes

In addition to supervised tasks, the representations learned by JASMINE can be used for unsupervised tasks such as clustering. This highlights another advantage to an unsupervised representation learning approach, which does not bias the representations to retain discriminatory information specific to a single task. We investigated

the potential for JASMINE-derived embeddings to stratify patients with a given cancer type into biological meaningful subtypes. We applied K-means clustering to the embeddings for five cancer types: breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), skin cutaneous melanoma (SKCM), and colon adenocarcinoma (COAD) to produce 3-8 clusters. We then compared the p-value of the log-rank test for differential survival among the clusters between JASMINE and the baseline embedding methods. The results are shown in Figure 3. Across cancer types and cluster numbers, clusters derived from JASMINE embeddings more consistently had a high level of survival differentiation compared to the other methods. It often was in the top three methods with highest survival differentiation. This indicates that the JASMINE-derived embeddings preserve information that may be useful for distinguishing cancer subtypes.

2.4.3 The most important embedding feature for cancer survival prediction can be explained by known disease markers

We hypothesize that the JASMINE-derived embeddings attain superior performance by capturing important multi-modal information that effectively distinguishes different subjects. Although the embedding features are not directly interpretable, we performed additional analyses to identify which original features from each modality contributed most to the embedding. JASMINE itself is a self-supervised representation learning method, and thus the embeddings it generates are not specialized for a specific task. Thus, employment of JASMINE for a particular prediction problem utilizes a two-step pipeline: first generating the embeddings and then inputting the representations to a downstream prediction model. To identify the features in each modality driving a particular prediction, we also approach feature interpretation in a two-step manner.

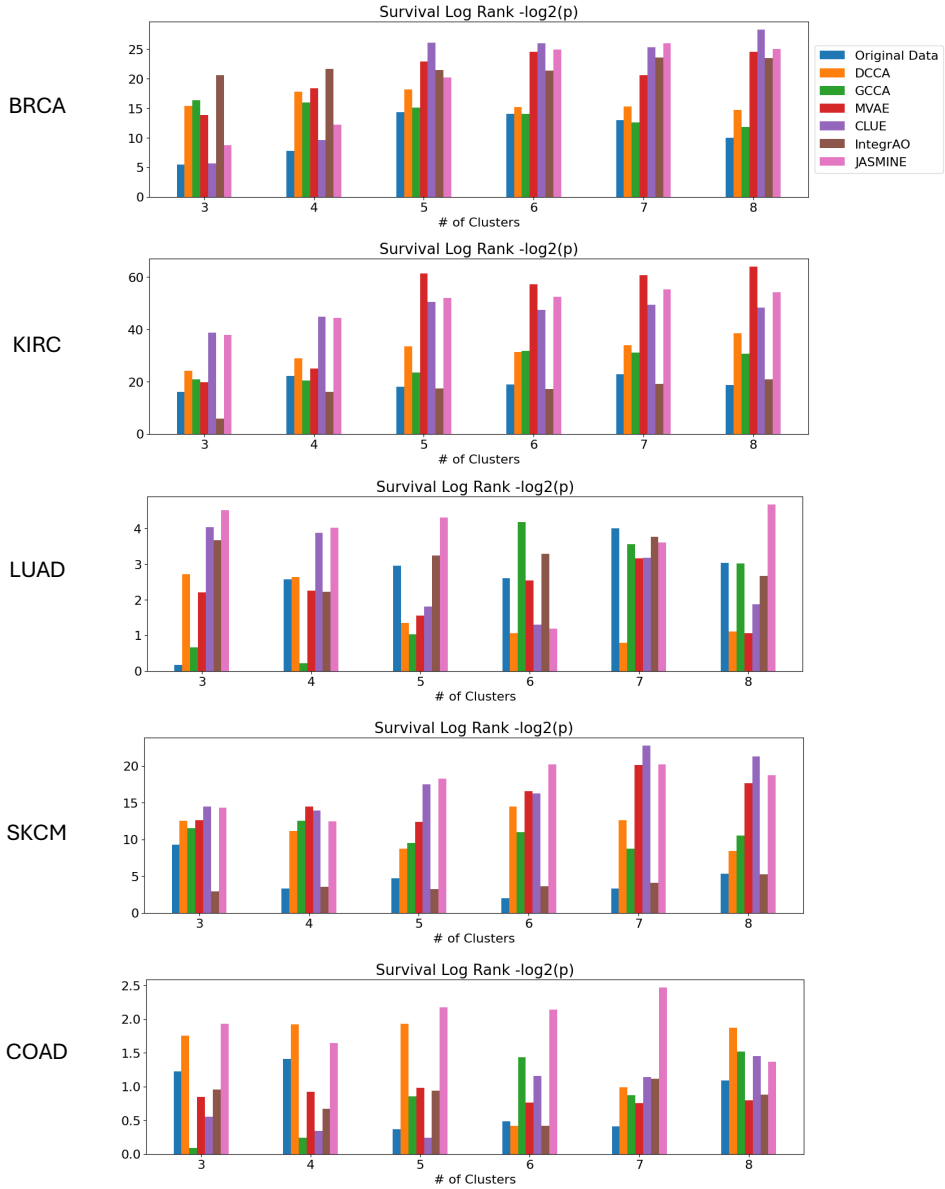


Fig. 3 Clustering Analysis. Log-rank test for differential survival among 3-8 clusters in five cancer types in TCGA: breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), skin cutaneous melanoma (SKCM), and colon adenocarcinoma (COAD). Reported values are mean $-\log_2(p)$ over five replicates, where p is the p-value of the log-rank test.

The first step involves using the downstream prediction model to identify which embedding features are most important for the prediction task. For this experiment, we focused on the cancer survival prediction task. The downstream predictive model for survival prediction was the Coxnet regression model. We selected the embedding feature corresponding to the regression coefficient with the largest absolute value as the most predictive feature. In our experiment, this was the 28th feature of the mRNAseq-specific embedding (mRNAseq28).

In the second step, we interpret the embedding feature by identifying which original interpretable variables (e.g. genes, miRNAs) from each modality contribute most to those embedding features. We do this in two main steps. First, we cluster the features within each modality into groups of correlated features. Then, for each group, we shuffle the values of all its constituent features and measure the impact on the embedding feature. We identify and select the feature group in each modality that produces the greatest change in the embedding. We expect that these selected interpretable input features are important drivers of the embedding feature of interest and thus can be used to provide a biological explanation of the information it contains. More details on our procedure for identifying the most important input features can be found in Section 5.5.

Applying this approach to the samples from the testing set, we identified top relevant feature clusters for each modality. For mRNAseq, Methylation, and RPPA, the features corresponded to genes, whereas the miRNAseq features corresponded to miRNA molecules. For the top mRNAseq, Methylation, and RPPA gene clusters, we used PANTHER [30] for GO Enrichment Analysis from the Gene Ontology Consortium [31, 32] to determine the top Reactome [33] pathways represented by these genes. For the top miRNA cluster, we used the miRNA Enrichment Analysis and Annotation Tool (miEAA) [34] for over-representation analysis to identify the top

associated diseases from the Human microRNA Disease Database (HMDD) [35]. The top enrichments are shown in Table 4.

Table 4 Top enriched pathways and diseases associated with the most predictive embedding feature for TCGA survival regression.

Modality	Reactome Pathway / HMDD disease	FDR-adjusted p-value
mRNAseq	GABA synthesis	1.06E-02
	↳ GABA synthesis, release, reuptake and degradation	2.42E-10
	↳ Neurotransmitter release cycle	3.16E-15
	↳ Transmission across Chemical Synapses	7.17E-29
Methylation	Negative feedback regulation of MAPK pathway	2.92E-04
	↳ Negative regulation of MAPK pathway	1.25E-03
	Signaling by TCF7L2 mutants	2.36E-02
	↳ Signaling by WNT in cancer	2.34E-03
	↳ Diseases of signal transduction by growth factor receptors and second messengers	1.39E-11
	Signaling by MAP2K mutants	5.84E-03
	↳ Oncogenic MAPK signaling	6.60E-04
miRNAseq	Colorectal Carcinoma	2.28E-04
	Carcinoma, Breast	9.98E-04
	Pancreatic Neoplasms	9.98E-04
	Amyotrophic Lateral Sclerosis	9.98E-04
	Vulvar Squamous Tumor	1.28E-03
	Glioblastoma	1.60E-03
	Carcinoma, Hepatocellular	1.71E-03
	Breast Neoplasms	2.58E-03
RPPA	positive regulation of interleukin-3 production	4.23E-02
	↳ regulation of gene expression	1.61E-05
	↳ regulation of multicellular organismal process	6.30E-07
	↳ positive regulation of gene expression	1.77E-04
	↳ positive reuglation of multicellular organismal process	4.23E-06
	lysyl-tRNA aminoacylation	4.22E-02
	↳ macromolecule metabolic process	8.94E-03
	↳ protein metabolic process	1.03E-02

Reactome [33] pathways are organized hierarchically. Indented entries preceded with arrows indicate sub-pathways.

For mRNAseq, methylation, and RPPA, the top identified gene clusters are enriched in pathways involved in cancer. This is most obvious for the genes identified for the methylation modality, which were enriched for negative feedback regulation of the mitogen-activated protein kinase (MAPK) pathway, signaling by TCF7L2 mutants

including signaling by Wnt in cancer, and signaling by MAP2K mutants including oncogenic MAPK signaling. It is known that the MAPK cascade is involved in human cancer cell survival, malignancy, and resistance to drug therapy [36]. This pathway is commonly deregulated in cancer cells in many different tumor types [37]. The fact that the top genes from the methylation modality were enriched for the oncogenic MAPK signaling pathway clearly suggests an association with MAPK-related cancer mechanisms supported by prior literature [37–40]. Additionally, TCF7L2 is a Wnt transcription factor that is known to suppress cancer cell invasion, and therefore reduced activity of this gene can increase malignancy [41]. It has also been suggested that TCF7L2 mediates resistance to radiotherapy and chemotherapy via its involvement in the Wnt signaling pathway which is known to be involved in cancer [42].

The genes from mRNA that most contributed to the embedding feature were enriched in the GABA synthesis pathway. Synthesis of GABA, a prominent neurotransmitter, has been found to be upregulated in cancer cells, including in non-nervous tissues, and increased GABA levels are associated with poor prognosis in multiple cancer types [43–49]. Prior literature has suggested that this is due to GABA's involvement in a cancer-intrinsic pathway that results in increased tumor cell proliferation and immune cell suppression [43]. Furthermore, GABA receptors can modulate cancer proliferation and migration through known mechanisms [43, 49, 50]. Additionally, deficient GABA transaminase, which is an enzyme that breaks down GABA, is associated with poor prognosis in hepatocellular carcinoma [51, 52].

The genes from RPPA that most contributed to the embedding feature were enriched in pathways involving the positive regulation of interleukin-3 (IL-3) production and lysyl-tRNA aminoacylation. Interleukins and their associated cytokines are important components of signaling pathways for innate and adaptive immune cells and thus play a critical role in cancer development and progression [53]. Additionally, it

has been found that IL-3 acts as a growth factor in many hematologic neoplasms, and its receptor protein CD123 is overexpressed in multiple hematologic cancers [54, 55]. Furthermore, a high expression of CD123 in acute myeloid leukemia is associated with worse prognosis [56]. The lysyl-tRNA aminoacylation pathway was also overrepresented by the top genes from the RPPA modality. Lysyl-tRNA synthetase is a type of aminoacyl-tRNAsynthetase (ARS) which is responsible for aminoacylation, the process by which a specific amino acid is attached to its corresponding tRNA, and plays an essential role in protein synthesis [57]. ARSs maintain cellular and systemic homeostasis, and overexpression or mutations in ARS-encoding genes have been linked to multiple cancer types, including breast, cervical, colorectal, endometrial, head and neck, lung, liver, ovarian, pancreatic, renal, stomach, thyroid, and urothelial cancer, glioma, and melanoma [57–62]. Among the different types of ARSs, lysyl-tRNA synthetase has a relatively high cancer-associated deregulation, can promote metastasis, and is often highly expressed in various cancer cells [63–65]. It has been found to be significantly upregulated in hepatocellular carcinoma [66], breast cancer [67], colorectal cancer [68], and gastric cancer with high expression of KRS in tumor cells associated with poor prognosis [69].

Finally, in the top selected miRNA feature cluster associated with the most predictive embedding feature, many types of cancer, including colorectal carcinoma, breast carcinoma, pancreatic neoplasms, vulvar squamous tumor, glioblastoma, hepatocellular carcinoma, and breast neoplasms, were over-represented. Thus, these results demonstrate that the top genes and miRNAs contributing to the embedding feature most predictive of cancer survival were enriched in pathways related to cancer proliferation, metastasis, and prognosis for multiple types of cancer. This affirms that JASMINE was able to learn an embedding containing information on important cancer biomarkers, aiding in survival prediction.

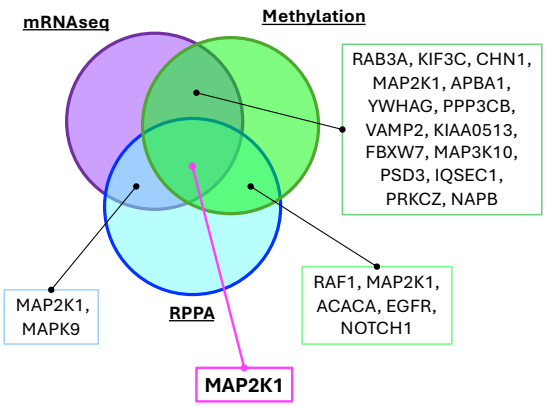
2.4.4 The top embedding feature for cancer survival prediction captures biologically meaningful inter-modality correspondence

In addition to assessing for known predictive biomarkers of cancer survival, we evaluated the inter-modality correspondence of the input variables represented by the top embedding feature for survival prediction (mRNAseq28). To do this, we found the cluster of features within each modality that contributed most to mRNAseq28. We then assessed the overlap in information among the selected clusters across modalities.

For mRNAseq, Methylation, and RPPA, we identified the genes within the top feature clusters that were shared by at least two modalities. For miRNAseq, we identified the target genes for each miRNA molecule in the selected feature cluster that overlapped with at least two of the other modalities. The target genes and their target prediction scores for each miRNA were obtained from miRDB [70, 71]. Target prediction score ranges between 50-100, with higher score indicating more confidence in the prediction. The results are shown in Figure 4.

Among the gene clusters identified for mRNAseq, Methylation, and RPPA, MAP2K1 was found within all three. Two of the miRNAs (hsa-mir-15b and hsa-mir-197) had MAP2K1 as a predicted target gene. In particular, hsa-mir-15b had a high target prediction score of 95. Gain-of-function mutations in MAP2K1 have been observed in multiple cancers including ovarian [72], melanoma [73], colon [73, 74], and lung [75, 76].

Across all miRNA molecules in the cluster, we also see that MAPK9 was frequently a predicted target. MAPK9 is related to MAP2K1 in its involvement in the MAPK signaling pathway, which regulates cellular processes including proliferation, differentiation, transcriptional regulation, and apoptosis [77, 78]. It has been correlated with poor prognosis and tumor progression in various cancers including glioma [79], prostate cancer [80], breast cancer [78], colorectal cancer [81], and renal cell carcinoma



miRNAseq

miRNA ID	Target Genes (target prediction score)
hsa-let-7d	APBA1 (67), MAPK9 (67)
hsa-mir-106b	KIAA0513 (98), PSD3 (55), IQSEC1 (54), MAPK9 (74)
hsa-mir-1287	CHN1 (62), NAPB (58)
hsa-mir-1307	PRKCZ (54)
hsa-mir-130b	KIAA0513 (89), MAPK9 (95)
hsa-mir-15b	**MAP2K1 (95)**, FBXW7 (97), MAPK9 (73), RAF1 (79)
hsa-mir-186	PPP3CB (86), PSD3 (88), NAPB (61), MAPK9 (87)
hsa-mir-191	MAPK9 (56)
hsa-mir-197	KIF3C (83), **MAP2K1 (52)**, FBXW7 (59)
hsa-mir-25	FBXW7 (100), NOTCH1 (80)
hsa-mir-324	KIF3C (74), KIAA0513 (52), PSD3 (71), NAPB (78)
hsa-mir-331	APBA1 (88), RAF1 (50)
hsa-mir-423	IQSEC1 (68), ACACA (67), EGFR (67)
hsa-mir-484	APBA1 (64), KIAA0513 (52)
hsa-mir-505	PPP3CB (66), MAPK9 (51)
hsa-mir-744	PSD3 (68), NAPB (69)

Fig. 4 Inter-modal feature correspondence. On the left, the Venn diagram shows the overlapping genes among the top feature clusters for the mRNAseq, Methylation, and RPPA modalities. The miRNA features are miRNA molecule IDs rather than genes; therefore on the right, the miRNA IDs are listed alongside any of their predicted target genes and corresponding target prediction scores in parentheses for any genes that overlapped with at least two of the other modalities. MAP2K1 is colored in pink and bolded as the gene shared by all modalities. MAPK9 is colored in purple as the most frequently targeted gene across the miRNAs in the top miRNAseq cluster. hsa-mir-15b is bolded as the miRNA molecule that targets both MAP2K1 and MAPK9, with high target scores for both.

[82]. Again, we note that the target prediction score of hsa-mir-15b for MAPK9 was relatively high, at 73. This may suggest an important role particularly for hsa-mir-15b.

The other target genes for hsa-mir-15b that overlapped with at least two of the top mRNAseq, Methylation, and RPPA clusters were FBXW7 (target score = 97) and RAF1 (target score = 79). FBXW7 is a known tumor suppressor gene [83]. It plays a role in the degradation of oncoproteins that are critically involved in oncogenesis, and loss-of-function mutations in this gene are correlated with resistance of tumor cells to treatment and poor prognosis [84–86]. Decreased expression of FBXW7 is observed in multiple cancers, including glioma, lung cancer, liver cancer, urothelial cancer, ovarian cancer, hematopoietic cancers, melanoma, and chronic lymphocytic leukemia [84, 86–90]. RAF1 is a proto-oncogene, and alterations of RAF proteins occurs commonly in cancer [91]. Like MAP2K1 and MAPK9, it is involved in the MAPK signaling pathway, and it has been found to play a role in a variety of solid cancers [92–95].

Thus, the overlap in the most important biological variables across different modalities for the mRNAseq28 embedding feature demonstrated inter-modality correspondence. These consistencies suggest that the cross-encoders that gave rise to the mRNAseq-specific embedding using the information from other modalities were able to capture biologically meaningful relationships among the different modalities. Furthermore, the corresponding variables involved in these relationships were relevant markers for cancer progression and prognosis.

2.5 Evaluation of representations learned from ADNI multi-omics data

To demonstrate the applicability of our method to more than one dataset, we also applied JASMINE to gene expression, genetic, metabolomics, and proteomics data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [96, 97]. There is a total of 1,665 samples, including several that have at least one missing modality. These data are summarized in Table 2B, and more details about ADNI data processing are in Section 5.6.3.

We applied the representations generated by JASMINE to two tasks: diagnosis classification and cognitive score prediction at baseline. Diagnosis classification is an important problem in Alzheimer's disease research and treatment. Alzheimer's disease (AD) is a complex and heterogeneous disease with a wide array of underlying mechanisms and presentations that complicate its characterization and diagnosis. Being able to classify individuals based on their multi-omics data may provide insight into the utility of molecular information for AD detection, guide further research into the molecular signatures underlying AD, and improve diagnostic models. The MMSE (Mini-Mental State Examination) is a commonly used cognitive score which is often used in diagnosing AD. This provides a measure of an individual's cognitive functioning on a more fine-grained scale than a diagnostic label. Predicting MMSE using

multi-omics data may also provide insight into the link between a subject's molecular profile and cognitive functioning, as well as how these relate to AD progression. By evaluating the representations generated by JASMINE on both classification and regression tasks, we also demonstrate a wider range of prediction problems to which our method can be applied.

2.5.1 JASMINE-generated representations achieved superior performance on AD diagnosis classification and cognitive score prediction

We applied JASMINE-derived representations to binary classification between CN and MCI/AD. The results are shown in Table 5A. JASMINE outperformed all baseline methods on all three classification metrics (balanced accuracy = 0.587, AUROC = 0.659, AP = 0.822), again demonstrating the ability of the proposed method to produce representations that have superior performance on a variety of tasks across different datasets. This also suggests that multi-omics data may have some utility in detecting MCI/AD.

We also applied the embeddings to MMSE score regression. The results are presented in Table 5B. Our proposed method had the best RMSE (2.301) and R^2-score (0.010), suggesting more accurate predictions. We note that while DCCA and the original data had the first and third highest r-pred, respectively (DCCA r-pred = 0.198; orig data r-pred = 0.185), the variation in this metric is very high. These methods are also disadvantaged by the fact that DCCA only can take two modalities at a time, and neither can handle missing data. While CLUE has a higher r-pred than the proposed method (CLUE r-pred = 0.192; JASMINE r-pred = 0.180), it has a negative R^2 score (CLUE $R^2 = -0.002$) and worse RMSE (CLUE RMSE = 2.314), indicating that while it can predict relative scores, its predictions are less accurate.

Table 5 Results on ADNI multi-omics data.

A. Diagnosis classification results			
Method	Balanced Acc	AUROC	AP
Original Data	0.490 ± 0.046	0.507 ± 0.045	0.722 ± 0.033
DCCA	0.484 ± 0.024	0.504 ± 0.041	0.706 ± 0.031
GCCA	0.536 ± 0.061	0.577 ± 0.044	0.738 ± 0.039
MVAE	0.566 ± 0.089	0.624 ± 0.085	0.797 ± 0.051
CLUE	0.571 ± 0.028	0.613 ± 0.056	0.765 ± 0.043
IntegrAO	0.498 ± 0.060	0.439 ± 0.074	0.662 ± 0.057
JASMINE	$\mathbf{0.587 \pm 0.050}$	$\mathbf{0.659 \pm 0.044}$	$\mathbf{0.822 \pm 0.028}$

B. MMSE prediction results			
Method	RMSE	R^2-score	r-pred
Original Data	2.302 ± 0.134	0.006 ± 0.118	0.185 ± 0.214
DCCA	2.320 ± 0.109	-0.009 ± 0.041	$\mathbf{0.198 \pm 0.105}$
GCCA	2.428 ± 0.061	-0.103 ± 0.055	-0.035 ± 0.110
MVAE	2.329 ± 0.075	-0.015 ± 0.065	0.143 ± 0.110
CLUE	2.314 ± 0.050	-0.002 ± 0.043	0.192 ± 0.049
IntegrAO	2.421 ± 0.072	-0.096 ± 0.066	-0.021 ± 0.150
JASMINE	$\mathbf{2.301 \pm 0.049}$	$\mathbf{0.010 \pm 0.042}$	0.180 ± 0.070

AP is average precision, which summarizes a precision-recall curve as the weighted sum of precisions achieved at multiple thresholds. Results are presented as mean ± sd.
R^2 score is the coefficient of determination, as computed using the r2_score function in scikit-learn[98]. r-pred is the Pearson correlation coefficient between the predicted and true values. Results are presented as mean ± sd.
Bold font indicates the best performance.

2.6 PoE, MCL, SCL, and orthogonality constraints enhanced performance on multiple tasks.

We performed ablation experiments to evaluate the utility of the main components of JASMINE: PoE for the joint representation (POE), modality-level contrastive learning (MCL), sample-level contrastive learning (SCL), and orthogonality constraints (orthog). We removed one component at a time and evaluated the performance on each of the supervised tasks for both TCGA and ADNI. The results are shown in Table 6.

The full model, consisting of all components, had the best performance across most prediction tasks, including TCGA survival prediction, cancer stage classification, and ADNI diagnosis classification. For survival prediction, although there is a very small increase in CD AUC (+0.63%) when POE is dropped, there is also a large drop in

Table 6 JASMINE ablation experiment results

TCGA Survival Prediction		
	C-index IPCW	CD AUC
Full Model	**0.762 ± 0.013**	0.798 ± 0.005
-POE	0.752 ± 0.016	**0.804 ± 0.009**
-MCL	0.726 ± 0.017	0.789 ± 0.007
-SCL	0.738 ± 0.017	0.791 ± 0.005
-orthog	0.748 ± 0.024	0.798 ± 0.005

TCGA Cancer Stage Classification			
	Balanced Acc	AUROC	AP
Full Model	**0.497 ± 0.007**	0.743 ± 0.004	0.521 ± 0.008
-POE	0.488 ± 0.008	0.745 ± 0.003	0.523 ± 0.005
-MCL	0.477 ± 0.007	0.738 ± 0.003	0.512 ± 0.008
-SCL	0.486 ± 0.007	0.739 ± 0.006	0.516 ± 0.013
-orthog	0.487 ± 0.009	**0.746 ± 0.004**	**0.526 ± 0.007**

ADNI Diagosis Classification			
	Balanced Acc	AUROC	AP
Full Model	**0.587 ± 0.050**	**0.659 ± 0.044**	**0.822 ± 0.028**
-POE	0.566 ± 0.055	0.643 ± 0.091	0.802 ± 0.065
-MCL	0.498 ± 0.028	0.593 ± 0.050	0.790 ± 0.045
-SCL	0.550 ± 0.054	0.611 ± 0.052	0.785 ± 0.032
-orthog	0.518 ± 0.047	0.559 ± 0.035	0.755 ± 0.025

ADNI MMSE Regression			
	RMSE	R^2-score	r-pred
Full Model	2.301 ± 0.049	0.010 ± 0.042	0.180 ± 0.070
-POE	2.391 ± 0.091	−0.070 ± 0.080	0.060 ± 0.135
-MCL	**2.284 ± 0.043**	**0.024 ± 0.037**	**0.225 ± 0.050**
-SCL	2.325 ± 0.094	−0.012 ± 0.080	0.142 ± 0.144
-orthog	2.303 ± 0.079	0.007 ± 0.069	0.198 ± 0.088

Results are presented as mean ± sd. Bold font indicates the best performance.

c-index (-1.3%), suggesting that POE is still important for this task. Removing the MCL component of the model led to the largest drop in performance across all TCGA tasks, suggesting that modality-level contrastive learning was particularly important. Removing SCL similarly led to relatively large drops in performance in most cases. MCL helps to separate distinct samples, while SCL helps to enhance the similarity structure of the original data, allowing for better discrimination between individuals with different survival times and cancer stages. For TCGA cancer stage classification,

dropping orthogonality constraints led to a small improvement in AUROC (+0.40%) and AP (+0.96%), but a much larger drop in balanced accuracy (-2.0%), suggesting that this model component is still beneficial. Using orthogonality constraints to minimize redundancy between the modality-specific and shared components of the embeddings may better preserve important information, aiding in the separation of samples by cancer stage.

On the ADNI dataset, we again see that all components of the model are helpful for achieving higher balanced accuracy, AUROC, and AP for diagnosis classification; removing any of them leads to a considerable drop in performance. For the MMSE prediction task, removing MCL from the model led to the best RMSE (2.284), R^2-score (0.024), and r-pred (0.225), but the full model had second-best RMSE (2.301) and R^2-score (0.024), and removing MCL led to large drops in performance on most other tasks. Overall, the ablation results suggest that all elements of JASMINE are useful across most scenarios, encompassing two different datasets and multiple prediction problems.

2.7 JASMINE achieves an advantageous balance of computational resource consumption and performance

We compared the computational resource usage of JASMINE and the baseline embedding methods. The training times and memory usage of each of these methods are shown in Table 7. All methods except for GCCA utilized A100 GPUs, whereas GCCA was run on a CPU. Furthermore, all but DCCA could be run using 3 or fewer cores. We note that while GCCA has the shortest training time, it still has relatively high peak memory usage and often has worse performance on many tasks. MVAE took many more epochs to converge than CLUE and JASMINE, resulting in a much longer training time. IntegrAO had much higher memory usage due to its reliance on building

patient similarity networks. Overall, CLUE and JASMINE appear to achieve a balance of moderate training time and peak memory usage while attaining high accuracy across multiple tasks.

Table 7 Compute resource comparison

Method	Training time (min)	Peak Memory Usage (GB)	# of cores	GPU/CPU
DCCA	36.12	1.49	23	A100 GPU
GCCA	5.17	3.42	1	CPU
MVAE	320.11	0.65	1	A100 GPU
CLUE	52.12	1.69	3	A100 GPU
IntegrAO	176.30	14.87	2	A100 GPU
JASMINE	58.60	1.72	2	A100 GPU

3 Discussion

We presented JASMINE, a variational autoencoder-based self-supervised representation learning method that generates task-agnostic embeddings from multi-omics data and is capable of handling samples with arbitrarily missing modalities. The representations are learned to incorporate both modality-specific and shared information while minimizing the redundancy between them. Contrastive learning is used to encourage the representations to enhance the similarities and differences between samples in the embedding space. JASMINE thus addresses the limitations of existing multi-omics integration methods by handling data with arbitrarily missing modalities while preserving both modality-specific and shared information and sample similarity structure. Furthermore, the self-supervised nature of JASMINE enables it to learn representations that preserve information in a data-driven, rather than task-biased, manner. Simulation experiments suggest that JASMINE is particularly well-suited for settings where the rate of modality missingness is high and when the missingness patterns differ among ground truth classes. JASMINE also demonstrates stable performance as additional sparsely available modalities are incorporated, whereas other methods show

a notable decline in performance, especially at small sample size. On real data, JASMINE outperforms the baselines across two datasets and multiple prediction tasks, including survival prediction, classification, clustering, and regression. Notably, the three VAE-based methods, MVAE, CLUE, and JASMINE, outperformed the graph-based IntegrAO approach, which tended to surpass the CCA-based methods, on the survival prediction, classification, and regression tasks. Upon further analysis of the JASMINE embeddings of the TCGA data, we found that the embedding feature most important for cancer survival prediction could be explained by known cancer biomarkers that had task-relevant cross-modal correspondence, demonstrating JASMINE's ability to learn and retain biologically meaningful information. JASMINE's superior performance across most supervised tasks in addition to its ability to derive biologically meaningful clusters demonstrates the generalizability of its learned embeddings across a wide set of tasks. The ability to use a single set of embeddings that excel at multiple tasks highlights its efficiency as well as its effectiveness at extracting important information from the data.

The self-supervised nature of JASMINE renders it efficient, as it only needs to be trained once per dataset to produce representations with top performance across a variety of tasks. This mechanism also underlies foundation models, which are trained on vast, often unlabeled datasets and can improve performance on downstream tasks [99–101]. We found that JASMINE-derived embeddings had superior performance on a variety of downstream tasks; given large amounts of data, JASMINE could potentially serve as the backbone architecture for an effective foundation model that can fully leverage the potential of incomplete multi-omics data. Additionally, JASMINE could be fine-tuned using specific prediction objectives to generate task-oriented embeddings with potentially even better performance.

Some limitations of our study include limited sample sizes, particularly in the ADNI dataset. Nonetheless, our results show some promise and are expected to improve

when more samples are available. Another limitation is that JASMINE models the likelihood function for all modalities as multivariate Gaussians, which may be less ideal for certain data types such as allele counts. Expanding JASMINE to enable the decoders to model a broader range of likelihood functions is a potential direction for future work. We also recognize that our use of a manual tuning procedure for the weights of the loss components in the objective function, while designed to efficiently find a parameter configuration with approximately optimal performance, may be less than ideal. Another extension of this work could be to implement a more precise data-driven approach. Additionally, while we performed analyses to help understand the biological underpinnings of the JASMINE embeddings, we acknowledge that the model itself is not inherently interpretable, the explainability of the embeddings remains understudied, and enhancing interpretability is a direction for future research.

JASMINE is also limited to tabular data inputs. In this paper, we focused on applying JASMINE to multi-omics data, which have this format. However, JASMINE could potentially be applied to tabular data from other domains. Furthermore, non-tabular data types, such as images or text, could be vectorized via a preprocessing step and then used as input to our method. In this way, JASMINE could be applied to other types of data from additional disciplines in a relatively straightforward manner. However, utilizing an ad-hoc preprocessing step to vectorize non-tabular modalities may not fully leverage the rich information contained in these data types. Thus, some future directions include extending the model architecture to handle additional modalities such as imaging, spatial transcriptomics, temporal data, and text.

4 Conclusions

We introduced JASMINE, a novel self-supervised representation learning method that produces integrated, low-dimensional embeddings of incomplete multi-omics data that better preserve modality-specific and shared information as well as sample similarity

structure for enhanced performance across a variety of tasks. JASMINE needs to be trained only once per dataset to produce representations with superior performance in multiple scenarios. Upon applying JASMINE to incomplete multi-omics data from two different landmark databases for cancer and Alzheimer's disease, JASMINE produced representations with superior performance across a range of supervised tasks and which clustered into cancer subtypes with better survival differentiation compared to state-of-the-art baseline multi-omics embedding methods. These results exhibit JASMINE's ability to better capture useful information for both clinically relevant prediction tasks and subtype clustering. Its demonstrated information extraction capability also suggests that it could potentially serve as the backbone for an effective foundation model that can harness the full potential of incomplete multi-omics data. This has several implications for the future of biomedical research and clinical care, from enhancing the study of underlying disease mechanisms to improved diagnosis and prognosis.

5 Methods

5.1 Notation and problem

Suppose the input data consists of n samples and m modalities. Then, we can represent the entire input data as a list of the data matrices for each modality: $X = \left[X^{(1)}, X^{(2)}, \ldots, X^{(m)}\right]$. It is possible that some samples are missing a subset of modalities. Thus, the dimensions of each of these modality-specific data matrices can vary. For example, the data matrix for modality j, which is measured for n_j samples and consists of d_j features, has dimension $X^{(j)} \in \mathbb{R}^{n_j \times d_j}$. The goal of our method is to learn integrated representations of the data, preserving both modality-specific and modality-shared information while also allowing for missing modalities. To preserve the most discriminative information in the data while reducing the data dimensionality and avoiding biasing the representations toward a specific downstream task, these

representations are learned in a self-supervised manner. These clean and compact representations can then be used as input for a variety of downstream tasks, including regression and classification.

5.2 Technical details of JASMINE

To accomplish these goals, we propose a variational autoencoder (VAE)-based model to produce integrated embeddings that comprise both modality-specific and modality-shared, or joint, components. The model architecture consists of the following units:

- m self-encoders, $q_{\varphi_j}(z_j|x_j)$, which map a modality j to its own latent space to generate latent embedding z_j.
- $m(m-1)$ cross-encoders, $q_{\varphi_{jk}}(z_k|x_j)$, which map a modality j to the latent space of another modality k to generate latent embedding z_k.
- m decoders, $p_{\theta_j}(x_j|z_j)$ which map a latent embedding for modality j back to its original feature space.

Self-encoders are similar to encoders in a vanilla VAE, enabling the latent representation for a given modality to be derived from the modality itself. However, only having self-encoders is insufficient when there are missing modalities. Therefore, the model also consists of cross-encoders which learn to take one modality j as input and infer the latent space of another modality k. All modality pairings are considered by this model, such that there are $m(m-1)$ cross-encoders. This not only allows the method to generate representations for missing modalities using the modalities that are available, but it also enables the learning of inter-modality relationships.

Additionally, there are decoders for every modality, which takes an embedding from the latent space of a given modality as input, and outputs the parameters of the distribution in the original feature space for that modality. The decoders are used for reconstructing the input during training to ensure that the learned embeddings preserve enough information to be able to recover the original data (measured via

reconstruction loss). For a visual overview of JASMINE, a schematic is given in Figure 1.

The encoders and decoders are feedforward neural networks with a chosen number of layers and hidden dimensions. In our experiments, the encoders had 2 layers, with hidden dimension of 512. The decoder had 1 layer with hidden dimension of 256. The latent representations have dimension 64.

To learn the parameters of these inference networks, we optimize the following objective function, consisting of the evidence lower bound (ELBO) loss, as is traditionally used for training VAEs:

$$\text{ELBO}(x) = \mathbb{E}_{q_\varphi(z|x)} \left[\lambda \log p_\theta(x|z)\right] - \beta \text{KL}\left[q_\varphi(z|x), p(z)\right] \tag{1}$$

where $p(z)$ is the standard Normal prior. Here, we see that the first term seeks to maximize the log likelihood of the data given the latent representation, and is equivalent to minimizing the reconstruction loss between the original data and its reconstruction from its embedding. For the second term, KL is the Kullback-Leibler divergence and is used as a regularization mechanism by encouraging the inferred latent distribution to be close to the prior. λ and β are weights that control the tradeoff between the two terms. For simplicity, JASMINE uses a multivariate Gaussian for the likelihood function in the ELBO for all data modalities.

There is an ELBO term for each of the m^2 encoders, where the first term is calculated using each of the modality-specific decoders to reconstruct the modality for which the latent representation z was generated. The second term is calculated as the KL divergence between each of the m^2 inferences and the prior. Thus, we have

$$\text{ELBO}_{self}(x) = \sum_{j=1}^{m} \mathbb{E}_{q_{\varphi_j}(z_j|x_j)} \left[\lambda \log p_{\theta_j}(x_j|z_j)\right] - \tag{2}$$

$$\beta \mathrm{KL}\left[q_{\varphi_j}(z_j|x_j), p(z_j)\right]$$

$$\mathrm{ELBO}_{cross}(x) = \sum_{j=1}^{m} \sum_{k \neq j} \mathbb{E}_{q_{\varphi_{kj}}(z_{k\to j}|x_k)} \left[\lambda \log p_{\theta_j}(x_j|z_{k\to j})\right] - \beta \mathrm{KL}\left[q_{\varphi_{kj}}(z_{k\to j}|x_k), p(z_{k\to j})\right] \tag{3}$$

Since an embedding space of a given modality can be generated using any of the m modalities via the self- or cross encoders, it is important that these embeddings are consistent regardless of which modality was used to for inference. To do this, we also include a cosine similarity loss term, which ensures that the inferences are aligned in the embedding space. This is given by

$$\mathcal{L}_{cosine} = \sum_{k=1}^{m} \left[1 - \frac{1}{m} \sum_{j=1}^{m} CosSim(z_{jk}, \bar{z}_k)\right] \tag{4}$$

where z_{jk} is the latent embedding of modality k inferred from modality j, $\bar{z}_k$ is the mean of z_{jk} across all $j \in [1, \ldots, m]$ and $CosSim(x, y) = \frac{x \cdot y}{\|x\|\|y\|}$. This aims to maximize the similarity between the embeddings inferred by each of the modalities and their mean.

The embeddings given by the self- and cross-encoders contain modality-specific information. In addition to this, the proposed method identifies information that is consistent among all modalities, which is encoded in a joint embedding and composes the modality-shared component of the final learned representation.

5.2.1 Joint embedding via product-of-experts

Information that is shared among all modalities is also important for defining a given sample. Thus, in addition to modality-specific information obtained via the self- and

cross-encoders, the proposed method learns a joint embedding to capture information that is shared across the modalities.

The strategy used is product-of-experts (PoE)[13], which defines the joint latent distribution to be the product of the modality-specific latent distributions:

$$q(z|x_1, \ldots, x_m) = p(z) \prod_{j=1}^{m} q_{\varphi_j}(z_j|x_j) \tag{5}$$

where $p(z)$ is the standard normal prior and $q(z|x_1, \ldots, x_m)$ is the joint latent space across all modalities $j = 1, \ldots, m$. This strategy handles missing modalities by taking the product over the subset of modalities that are available for a given sample. Let M_i be the set of modalities available for sample i. Let us represent sample i as X_i, where the data of the jth modality for X_i is represented as x_{ij}. Then, the joint latent representation is estimated as

$$q(z|X_i) = p(z) \prod_{j \in M_i} q_{\varphi_j}(z_j|x_{ij}) \tag{6}$$

In order to avoid having to learn all 2^m VAEs to account for all possible combinations of modalities, prior methods have approached this estimation by randomly masking subsets of the modalities in the training data. Since our training data already consists of samples with arbitrarily missing modalities, the model can be trained on these samples to learn to infer the joint space in the presence of incomplete modalities. We define an ELBO loss term for the joint inference as follows:

$$\begin{aligned} \text{ELBO}_{PoE}(x) = \sum_{j=1}^{m} \mathbb{E}_{q(z|X)} \left[\lambda \log p_{\theta_j}(x_j|z)\right] - \\ \beta \text{KL}\left[q(z|X), p(z)\right] \end{aligned} \tag{7}$$

where z is the PoE-derived joint representation, X refers to all modalities available for a given sample (which may only be a subset of all m modalities), $q(z|X)$ is the joint multimodal inference, and $p(z)$ is the standard normal prior for the joint latent space. Note that the first term is equivalent to minimizing the reconstruction loss of each modality j from the joint representation using the corresponding modality-specific decoder, $p_{\theta_j}(x_j|z)$. Since PoE produces a single joint embedding that preserves sufficient information to reconstruct each modality, this training process encourages the joint embedding to contain information that is shared among all modalities.

One of the benefits of the PoE-based formulation of the joint representation space is that the resulting joint distribution can be written in closed-form. Namely, given that we represent each modality-specific embedding space as a Normal(μ_j,Σ_j) distribution, the resulting joint latent space is Normal(μ_{PoE},Σ_{PoE})[102], where:

$$\mu_{PoE} = \left(\sum_{j=1}^{m} \Sigma_j^{-1}\right)^{-1} \left(\sum_{j=1}^{m} \Sigma_j^{-1} \mu_j\right) \tag{8}$$

$$\Sigma_{PoE} = \left(\sum_{j=1}^{m} \Sigma_j^{-1}\right)^{-1} \tag{9}$$

5.2.2 Construction of the final representation

Once the model has obtained the modality-specific representations from the self- and cross-encoders, we obtain final modality-specific representations for each sample by averaging over the representations derived from every available modality, i.e., letting z_{ij} be the final modality-specific representation of modality j for sample i, $z_{i,k\rightarrow j}$ be the representation of modality j derived from modality k, and M_i be the set of modalities available for sample i, we have

$$z_{ij} = \frac{1}{|M_i|} \sum_{k \in M_i} z_{i,k\rightarrow j} \tag{10}$$

This enables the generation of modality-specific embeddings for all modalities for every sample, even when the modality itself is missing.

Let $z_{i,joint}$ be the joint representation for modality i. Then, the final JASMINE representation for sample i, z_i, is the concatenation of each of the modality-specific representations and the joint representation:

$$z_i = concat(z_{i1}, z_{i2}, \ldots, z_{im}, z_{i,joint}) \tag{11}$$

Letting h be the dimension of the latent embeddings, it follows that $z_i \in \mathbb{R}^{h(m+1)}$.

5.2.3 Orthogonality Constraints

Another aim of our representations is to preserve both modality-specific and modality-shared information in the data while minimizing the amount of redundancy between them, in order to produce more information-rich representations. To do this, we also introduce an orthogonality constraint [18] into our model as a loss term. The orthogonality constraint-based loss term is calculated as follows:

$$\mathcal{L}_{orthog} = \sum_{j=1}^{m} avg(\mathbf{A}^{joint} \odot \mathbf{A}_j^{specific}) \tag{12}$$

where $\mathbf{A}$ represent $B \times h$ matrices of the latent representations, B represents the batch size, h represents the dimension of the joint and specific embeddings, and $\odot$ represents element-wise multiplication. The average of this element-wise product provides a measure of the similarity between the joint and modality-specific embeddings, and therefore, minimizing this helps to accomplish our goal of ensuring that the modality-specific embeddings contain different information from the modality-shared embeddings. This helps to maximize the amount of unique information that

is preserved in the final representations, which consist of the concatenation of the modality-specific and joint representations.

5.2.4 Contrastive learning

To better enforce consistency among modalities for a given sample as well as preserve and enhance the structure in the original data for better discriminability in downstream tasks, we introduce two types of contrastive learning (CL), which we refer to as modality-level and sample-level. These losses are applied to the embeddings generated by our model.

Modality-level CL (MCL) promotes similarity between embeddings of modalities from the same sample and difference between embeddings of modalities from different samples to better distinguish different samples while encouraging consensus among the modalities for a given sample. Here, sample affiliation is used to define positive and negative pairs among the modalities: positive pairs consist of embeddings of different modalities from the same sample, and negative pairs are embeddings of modalities from other samples. As there are multiple positive and negative pairs defined for each embedding, we define a multi-way version of contrastive loss based on soft-nearest neighbors loss to calculate MCL. Letting z_{ij} represent the embedding of modality j for the ith sample, and m be the number of modalities, the MCL is calculated as:

$$\mathcal{L}_{MCL} = -\frac{1}{B}\sum_{i=1}^{B}\log\left[\frac{\sum_{k\neq j,k=1:m}\exp\left(-f(z_{ij},z_{ik})/\tau\right)}{\sum_{k\neq j,l=1:B,k=1:m}\exp\left(-f(z_{im},z_{lk})/\tau\right)}\right] \tag{13}$$

where B is the batch size, $f(x,y)$ is the cosine similarity between x and y, and τ is a temperature parameter that controls how close the features are in the representation space. We set $\tau = 0.1$ for all experiments in this study. Note that the numerator consists of the sum of similarity measures over positive pairs, whereas the denominator consists of the sum of similarity measures over all pairs. We seek to maximize the

similarity between positive pairs compared to negative pairs. Thus, minimizing the negative log of this ratio accomplishes this goal.

Sample-level CL (SCL) encourages the embeddings to preserve and enhance the similarity structure between samples, based on the joint component of their embeddings. We do this in an unsupervised manner to avoid biasing the representations toward a specific structure based on a single set of labels. Thus, SCL is based on a deep clustering approach [103, 104] that consists of the following main steps: (1) stochastic data augmentation to transform each input data example randomly resulting in two correlated views of the same example, which we denote as a positive pair, and (2) generation of the joint embedding for the augmented dataset (consisting of original and perturbed data), and (3) computation of the contrastive loss on the augmented data, defining the original data and its perturbed representation as a positive pair. We calculate SCL as:

$$\mathcal{L}_{SCL} = -\frac{1}{B}\sum_{i=1}^{B}\log\left[\frac{\exp(f(z_i, z_i^+)/\tau)}{\sum_{k=1}^{2N}\mathbb{1}_{[k\neq i]}\exp\left(f(z_i, z_k)/\tau\right)}\right] \tag{14}$$

where z_i is the PoE-derived joint component of the embedding, z_i^+ is the augmented version of z_i, and $f(x, y)$ is again defined to be the cosine similarity between x and y. Since the data are tabular, the stochastic data augmentation is performed following SCARF[105]: a fraction of the features are randomly selected to be "corrupted" by replacing them with values sampled with replacement from all possible values for that feature within the batch. The goal of SCL is to learn to keep samples that are similar to one another closer together in the embedding space. By defining the original and perturbed versions of each sample as positive pairs, we can accomplish this goal without predefined cluster labels. One of the hyperparameters associated with this loss term is the proportion of features to randomly perturb in each sample. We chose this value to be 0.25, which worked well in our experiments.

5.2.5 Learning objective

Given the loss terms described previously, the final objective function is

$$\begin{aligned}\mathcal{L}_{JASMINE} = \lambda_{self}\text{ELBO}_{self} + \lambda_{cross}\text{ELBO}_{cross} + \\ \lambda_{cosine}\mathcal{L}_{cosine} + \lambda_{joint}\text{ELBO}_{PoE} + \\ \lambda_{MCL}\mathcal{L}_{MCL} + \lambda_{SCL}\mathcal{L}_{SCL} + \lambda_{orthog}\mathcal{L}_{orthog}\end{aligned} \tag{15}$$

where the λ values are scalar weights on the loss terms. These values were tuned manually on the validation data (see Section 5.2.6 for more details on the tuning procedure).

5.2.6 Parameter tuning procedure

The weights of the loss terms of the objective function were tuned manually. We started with a set of values which put each loss term on a similar scale to enable roughly equal contribution to the overall objective. Then, we explored a range of values above and below each initial parameter value, modifying one variable at a time. Based on the direction of change that led to performance improvements on the validation data, we either explored additional values farther in that direction until performance decreased, or if performance improved before declining for values in a particular direction, we chose the parameter value that produced the optimal validation performance. We then trained the model using all of the optimal parameter values together. After comparing the performance of each of the model configurations, including the models for which one variable was changed at a time as well as the model combining all of the optimal parameter values, we selected the best combination of parameter values to be that with the best performance. We explored 3-5 values for each parameter, resulting in approximately 25-35 configurations per model.

All parameter tuning was done using the portion of the data reserved for training and validation. Each configuration was trained on the training portion of the five train/validation folds and evaluated on the held-out validation fold. The average performance across these train/validation splits was used to determine the best parameter configuration while accounting for sampling noise. After all parameters were tuned, final performance was reported on the reserved test dataset which was not seen by the model during training nor parameter tuning.

5.3 Baseline Methods

We compare our method to multiple baselines, which were used to generate embeddings from multi-modal data in the first part of the pipeline. These embeddings were subsequently input to the prediction models mentioned above. The baselines include the original data, deep canonical correlation analysis (DCCA) [27], generalized canonical correlation analysis (GCCA) [28], multimodal variational autoencoder (MVAE) [13], cross-linked unified embedding (CLUE) [12], and IntegrAO [15].

Original data consists of concatenating the original features across all modalities and treating the resulting vector as the input to the prediction model. The canonical correlation analysis (CCA)-based methods serve as baseline dimensionality reduction methods that produce unified embeddings incorporating information across multiple data sources. Generally, CCA aims to find linear projections of two data sources with maximal correlation, and thus can learn relevant inter-modality relationships. DCCA extends CCA to learn nonlinear relationships but is limited to two modalities at a time. To use DCCA, we learn all pairwise projections and then take the average across all learned projections for each modality before concatenating the averaged projections across modalities. GCCA extends CCA to handle more than two data sources at a time but is still limited to linear projections. For the original data, DCCA, and GCCA, which do not handle missing modalities, we subset the samples to those that have all modalities.

On the other hand, MVAE and CLUE are autoencoder-based methods that can handle missing modalities. MVAE does this using a product-of-experts approach to define a joint embedding. CLUE, on the other hand, does not learn a joint embedding but instead uses self- and cross-encoders to learn modality-specific representations while being able to impute missing modalities. IntegrAO is a recently published graph-based method that constructs omics-specific patient graphs and combines them utilizing graph neural networks to produce unified patient embeddings. It utilizes patient similarity to propagate information to subjects with missing modalities from those that are similar. Since these methods can handle missing modalities, they were trained on all samples. To keep the testing data consistent across all methods for fair comparison, test results for all methods were reported for only the samples with all modalities. However, we note that it is possible for MVAE, CLUE, IntegrAO, and JASMINE to perform inference on samples with missing modalities. A summary of these methods and their features is presented in Table 1.

5.4 Evaluation Metrics

Survival prediction performance metrics. For survival prediction, we used the concordance index for right-censored data based on inverse probability of censoring weights (C-index IPCW) and cumulative dynamic area under the receiver operating characteristic curve (CD AUC). C-index IPCW measures the how well the method predicts the ordering of the survival times while handling a large proportion of censoring (69.4% of the samples in TCGA are censored) by using the inverse probability of censoring weights. It does not depend on the distribution of censoring times in the test data and thus can produce unbiased estimates [98]. CD AUC is analogous to the AUROC in that it converts the inference to a binary classification problem by predicting survival at a given cutoff time and assesses the sensitivity and specificity of the predictions across a range of time thresholds.

Classification performance metrics. For the classification tasks, we evaluated the performance using balanced accuracy, AUROC and average precision score. Balanced accuracy measures the correctness of the predicted classes while weighting each sample by the inverse prevalence of its true class. For AUROC, we used the macro averaged one-versus-rest AUROC across all classes. Average precision score (AP) is analogous to the AUROC in that it summarizes the precision-recall curve by taking the weighted mean of precisions achieved across a range of predicted probability thresholds. Weighting is determined by the increase in recall from the previous threshold [98]. For the multiclass scenario, the macro-average across classes is utilized.

Clustering performance metrics. We applied K-means clustering to the embeddings over a range of cluster numbers. In our experiments, we varied the number of clusters from 3 to 8. In TCGA, we tested for differential survival among clusters of TCGA subjects within five different cancer types. For this, we used the multivariate log-rank test, which tests for differences in survival curves among multiple clusters. We reported $-\log 2(p)$ of this test, where p is the p-value.

Regression performance metrics. For regression tasks, we used root mean squared error (RMSE), the coefficient of determination (R^2-score), and the Pearson correlation between the predicted and true values (r-pred). RMSE takes the square root of the mean of squared differences between the true and predicted values across samples. The R^2-score, as computed using the r2_score function in scikit-learn [98], measures the proportion of variation in the data that is explained by the model, i.e., what is left after subtracting the proportion of variation in the data captured by the error in predictions:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{16}$$

where y is the true outcome, $\hat{y}$ is the predicted outcome, and $\bar{y}$ is the expected value of y. This measures how well future samples are likely to be predicted by the model, where 1.0 is the best possible score, and the model that always predicts $\bar{y}$ has an R^2 score of 0 [98].

5.5 Feature Importance Analysis

JASMINE itself is a self-supervised representation learning method, and thus the embeddings it generates are not specialized for a specific task. Furthermore, employment of JASMINE for a particular prediction problem involves a two-step pipeline: first generating the embeddings, and then inputting the representations to a downstream prediction model. To identify the features in each modality driving a particular prediction, we also approach feature interpretation in a two-step manner. This involves first using the downstream prediction model to identify which embedding features are most important for the prediction task. Then, once the top embedding features are identified, we interpret these features by identifying which original interpretable variables (e.g. genes, proteins, miRNAs) from each modality contribute most to those embedding features.

In our experiments, we focused on the cancer survival prediction task. The downstream predictive model for survival prediction was the Coxnet regression model. We first selected the top predictive embedding feature determined by this model by identifying the feature corresponding to the regression coefficient with the largest absolute value.

Each element of the embedding is the result of complex functions of the original inputs and is thus not directly interpretable. To elucidate its potential biological underpinnings, we identified the original interpretable features with the greatest influence on the embedding feature of interest. We determined this by applying a perturbation via shuffling to the original features and measuring the impact on the embedding

feature of interest. This idea is based on permutation feature importance [106, 107], which identifies a feature as important if shuffling its values meaningfully increases prediction error. Since JASMINE is a self-supervised representation learning method, we modified this approach to measure the change in the embedding feature rather than prediction error. This involved two main steps: (1) identifying groups of correlated features within each modality, and for each group, (2) permuting the values of all of its constituent features and measuring the impact on the embedding feature of interest. The features within each modality that produced the greatest change in the embedding feature were interpreted to be the most important contributors to that embedding feature.

Because many of the input features are correlated, we first identified groups of correlated features within each modality before applying the permutation to these feature groups rather than individual features. This served two main purposes: (1) it reduced the computational burden of testing tens of thousands of individual features to testing thousands of feature groups, and (2) it helped avoid the case in which correlated features capturing similar information reduce the impact of perturbing a single important feature, which could result in overlooking truly important features.

To generate groups of correlated features, we computed the correlation matrices between the features in each modality separately. Then, we converted these correlation matrices to distance matrices by subtracting the absolute values of their elements from 1. We then applied agglomerative clustering to the distance matrices to identify clusters of correlated features within each modality [108–110].

For the perturbation step, we applied the shuffling at the cluster-level by permuting the values of all features within a given cluster before evaluating the impact on the embedding feature of interest. This was calculated using the correlation across samples between the value of the generated embedding feature before and after shuffling the input features within the cluster of interest. We expect that if an input feature is

more relevant to the embedding feature, then permuting its values will have a greater impact on the value of the embedding feature, resulting in a lower correlation. We used correlation rather than a measure of magnitude of change to account for differences in scales among the features: a larger scale feature may produce a greater shift in the embedding feature than a smaller scale feature of equal importance.

Overall, this procedure resulted in one group of features for each modality that was designated as most important for a given embedding feature. We then performed pathway and disease enrichment analyses on these features to interpret the underlying biology captured by the embedding. For the top mRNAseq, Methylation, and RPPA gene clusters, we used PANTHER [30] for GO Enrichment Analysis from the Gene Ontology Consortium [31, 32] to determine the top Reactome [33] pathways represented by these genes. The miRNAseq features were miRNA molecule IDs rather than genes; thus, for the top miRNA cluster, we used the miRNA Enrichment Analysis and Annotation Tool (miEAA) [34] for over-representation analysis to identify the top associated diseases from the Human microRNA Disease Database (HMDD) [35].

Additionally, we analyzed overlapping information among the modalities to evaluate whether the embedding captured biologically meaningful inter-modal feature correspondence. For mRNAseq, Methylation, and RPPA, we identified the genes within the top feature clusters that were shared by at least two modalities. For miRNAseq, we identified the target genes for each miRNA molecule in the selected feature cluster that overlapped with at least two of the other modalities. The target genes and their target prediction scores for each miRNA were obtained from miRDB [70, 71]. Target prediction score ranges between 50-100, with higher score indicating more confidence in the prediction. We then assessed the overlapping genes for associations with a prediction task of interest.

5.6 Data processing

The data were split into train, test, and validation sets. We used multiple train/validation splits for tuning hyperparameters and assessing performance variability. For our experiments, there were five train/validation splits, each resulting in one trained model, and a single test dataset. We report the means and standard deviations of the performance metrics on the test data from the five models. For more details on the dataset-specific splits and hyperparameter settings, see Sections 5.6.2 and 5.6.3.

Our experimental pipeline consists of two main steps: (1) generate embeddings using an unsupervised representation learning method and (2) input the embeddings to a prediction model of choice for a given inference problem. To generate embeddings for part (1) of our pipeline, we first trained the models on each training dataset and then used them to generate embeddings for all data samples. For part (2) of the pipeline, we used the embeddings of the training data to train a predictive model on a specific task (e.g. survival prediction, classification, regression). Performance of the predictive model on the embeddings of the validation data were used to tune the hyperparameters of the models. Final results were reported on the embeddings of the test data.

5.6.1 Simulated data

To compare JASMINE and the baseline methods under controlled data scenarios, we applied them to simulated data. These data were simulated as coming from three different classes, each consisting of 3,000 samples for a total sample size of 9,000. For each sample, informative features which distinguished class labels were generated from a multivariate Normal distribution. We simulated six modalities with 300 informative features per modality. Some experiments involved fewer than six modalities; for these, we simply ignored the extra modalities. Our simulated data incorporated both modality-shared and specific informative features. Modality-shared features had

a correlation of 0.3 with the corresponding features from all other modalities, whereas modality-specific features had a correlation of 0.

To generate the data, we first sampled class centroids from a multivariate Normal distribution with zero mean and inter-class variances of 0.1. These centroids represented all informative features from all modalities, for a total of $300 \times 6 = 1,800$ informative features. The correlations among corresponding features of different modalities were set based on whether the features were modality-specific (correlation of 0) or modality-shared (correlation of 0.3). Once the class centroids were generated, we produced individual samples by taking 3,000 random draws from a multivariate Normal distribution with the class centroid as the mean and intra-class variance of 0.3. Again, the correlations among features were set based on whether the features were modality-specific or modality-shared. Once the total feature vectors were generated, we split them into feature sets representing different modalities. For each modality, we generated 1,800 uninformative features from a standard uniform distribution to represent noise in the data. Finally, we applied a nonlinear mapping by squaring the data so that they were not linearly separable. We also generated an additional dataset using the same settings but reducing the number of samples per class to 1,500 for a total sample size of 4,500 for experiments at small sample size.

Our simulation experiments involved altering the modality missingness patterns and the number of modalities. To do this, we generated random masks to indicate which modalities are missing for each sample. Masks were boolean, with True indicating that a modality is available and False indicating that it is missing. We generated masks for three types of experiments: (1) balanced missingness at various rates, (2) unbalanced missingness based on class, and (3) varying the number of modalities.

For balanced missingness at various rates, every sample had the same probability of missing a modality regardless of its true label. For each modality, we generated

a boolean mask with length equal to the number of samples with probability $p \in \{0.0, 0.5, 0.1, 0.95, 0.99\}$ of being False, indicating that the modality is missing.

For unbalanced missingness based on class, we simulated the case when the probability that a given modality is missing is larger for samples within a specific class. In this case, samples in class 1 were missing modality 1 with probability 0.8 but missing the other modalities with probability 0.3. Samples in class 2 were missing modality 2 with probability 0.8 but missing the others with probability 0.3. Samples in class 3 were missing modality 3 with probability 0.8 but missing the remaining modalities with probability 0.3.

Finally, for the experiments varying the number of modalities, all samples were simulated to have modality 1 available. For each following modality, the probability of missing that modality was 0.95. As for the balanced missingness experiment, we generated masks independently for each modality, each with the same missingness probability across all samples.

5.6.2 TCGA

We applied JASMINE and the baseline methods to multi-omics data from The Cancer Genome Atlas (TCGA) [29], consisting of mRNA-seq, methylation, miRNA-seq, and functional proteomics reverse phase protein array (RPPA) modalities.

The mRNA-seq data were obtained using the Affymetrix HG U133, Affymetrix Exon Array and Agilent gene expression platform. Each feature corresponds to the transcript abundance calculated from microarray intensities for a given gene (e.g. CA5BP, ERP27, TP53).

The methylation data represent the normalized methylation β values calculated from methylated and unmethylated probe intensities, quantifying methylation levels at each gene. Thus, each feature corresponds to a gene (e.g. BRCA1, FOXA1, GRP). Genes on chromosomes X or Y, with gene symbol NA, and for which more than 5% of the β values were NA were filtered out.

The miRNAseq data were RPKM log2 transformed gene-level counts calculated from microarray intensities. Each feature represents a gene (e.g. hsa-mir-1247, hsa-mir-216a, hsa-mir-770). All records with NA values were removed.

The RPPA data represent protein expression levels measured via antibody probe and calculated from spot intensities. Each feature corresponds to a gene/protein (e.g. CLDN7|Claudin-7, ERBB3|HER3, CTNNB2|beta-Catenin). Features with NA values for any sample were removed.

All features were standardized to have zero mean and standard deviation of 1 before being input to the models. This ensured that the features were on a similar scale within and across modalities.

This dataset consists of 10,472 total samples, some of which are missing one or more modalities in a blockwise fashion, i.e., all features for a modality are either observed or not observed. We kept the features that were measured for all subjects with the corresponding modality. To test our methods' ability to handle a large number of features, we did not perform any prior feature selection or dimensionality reduction. The summary of sample sizes, dimensionality, and labels are available in Table 2.

For TCGA, we split the data samples into six folds, five of which were used for 5-fold cross-validation. The sixth split was used as a held-out test dataset, on which our final results are reported. For the methods that do not handle missing data (original data, DCCA, GCCA), we subset the data to samples with all modalities available (6,054 samples). For the remaining baseline methods (MVAE, CLUE, IntegrAO) and JASMINE, we use all samples for training.

For cancer stage classification, we used pathologic stage as the label, which is based on clinical staging and histopathologic examination of surgical specimens to determine the extent of cancer spread [111]. There are four main cancer stages, I, II, III and IV, each additionally comprising sub-stages. We removed samples with stages that did not fall into any of these categories due to very little representation of these categories

in the data. We aggregated the sub-stages into the four main stages as the labels for four-class classification.

For training, we used the RMSprop optimizer from PyTorch, learning rate of 0.002 and dropout of 0.2. The model architecture consisted of the following parameter settings: encoder depth of 2, encoder hidden dimension of 512, decoder depth of 1, decoder hidden dimension of 256. The dimension of the learned representations was set to 64 (this setting was used for all baseline methods, except for original data). The best settings for the weights of the loss terms included in our method, which were determined based on performance on the validation data, were found to be $\lambda_{self} = 1.0$, $\lambda_{cross} = 1.0$, $\lambda_{cosine} = 3.0$, $\lambda_{joint} = 1.0$, $\lambda_{MCL} = 1.0$, $\lambda_{SCL} = 0.5$, and $\lambda_{orthog} = 0.5$.

5.6.3 ADNI

We also applied JASMINE to multi-omics data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [96, 97]. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The original goal of ADNI was to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The current goals include validating biomarkers for clinical trials, improving the generalizability of ADNI data by increasing diversity in the participant cohort, and to provide data concerning the diagnosis and progression of Alzheimer's disease to the scientific community. For up-to-date information, see adni.loni.usc.edu.

From ADNI, we utilized gene expression (GEX), genetic (GEN), metabolomics (MET), and proteomics (PROT) modalities. The GEX data were microarray gene expression profile data calculated from microarray intensities for a given transcript.

Expression values were preprocessed via the RMA (Robust Multi-chip Average) normalization method. Each feature corresponds to a transcript, given by the Affymetrix probe ID (e.g. 11715373_a_at, 11723370_s_at, 11729407_a_at). Although the ADNI GEX data are mRNA-seq measurements, they are not equivalent to the mRNA-seq data in TCGA, as these represent probe-level rather than gene-level expression, and they were preprocessed using different normalization methods. The GEN data are allele counts for 935 SNPs derived from key AD GWAS studies [112, 113]. These counts were extracted via PLINK.

The MET data were flow injection analysis (FIA) and ultra performance liquid chromatography (UPLC) data from which represent serum metabolite concentrations. The FIA data were measured for the Glycerophospholipid, Sphingolipid, and Acylcarnitine analyte classes; the UPLC data were measured for the Amino Acid and Biogenic Amine analyte classes. These concentrations were calculated following peak integration and calibration of the mass spectrometry data. Each feature represents a metabolite (e.g. C0, C14.1, C16 for FIA and Ala, Arg, Asn for UPLC).

The PROT data represent median normalized CSF protein levels measured via SomaScan 7K assay and calculated from probe intensities. Each feature represents an analyte ID (e.g. X10589.7, X16916.19, X3336.50), which is convertible to its corresponding gene or protein ID. Feature filtering performed by ADNI included removing aptamers based on thresholding the scale factor, the coefficient of variation, and call rate as well as removing outliers, non-human analytes, and analytes without protein targets. For both the MET and PROT data, features that were missing more than 5% of the samples were removed; for the remaining features, missing values were imputed using the mean of that feature across all samples.

All features were standardized to have zero mean and standard deviation of 1 before being input to the models. This ensured that the features were on a similar scale within and across modalities.

The dataset comprises a total of 1,665 samples, including several that have at least one missing modality. As with the TCGA data, missingness is blockwise such that either all features or no features are missing for a given modality. For gene expression and genetic data, all features were measured for all samples. For metabolomics and proteomics data, we dropped features that were measured for fewer than 95% samples. For the remaining features, we filled missing values with the sample means. For samples with repeated proteomics measurements, we utilized the first measurement available. Baseline diagnoses and cognitive scores were extracted for each patient. The majority of metabolomics and proteomics samples were baseline measurements. Many gene expression measurements were not taken at baseline, but we included these measurements as a proxy for baseline gene expression. The sample sizes, features, modalities, and labels are summarized in Table 2.

For ADNI, due to the smaller sample size, we split the data first into 70% training and 30% testing. Then, within the training set, we generated five random 70%/30% training/validation splits. As with TCGA, for the methods that do not handle missing data (original data, DCCA, GCCA), we subset the data to samples with all modalities available (230 samples). For the remaining baseline methods (MVAE, CLUE, IntegrAO) and our proposed method (JASMINE), we use all samples for training. For fair comparison, all methods are tested only on samples with all modalities.

We applied JASMINE to diagnosis classification and MMSE (Mini-Mental State Examination) prediction. The data contained three main diagnoses: CN (cognitively normal), MCI (mild cognitive impairment), and AD (Alzheimer's dementia). Due to small sample size and class imbalance resulting in relatively few AD samples, we converted the task to a binary classification problem, with CN and MCI/AD as the two classes. MMSE is a cognitive exam consisting of 30 questions that evaluate attention and orientation, memory, registration, recall, calculation, language and ability to draw a complex polygon [114]. Lower scores indicate worse cognitive state, and a score less

than 24 has traditionally been used as a cutoff indicating evidence of dementia. As the score varies from 0 to 30, we posed this as a regression task.

For training, we used the RMSprop optimizer from PyTorch, learning rate of 0.002 and dropout of 0.2. The model architecture consisted of the following parameter settings: encoder depth of 2, encoder hidden dimension of 512, decoder depth of 1, decoder hidden dimension of 256. The dimension of the learned representations was set to 64 (this setting was used for all baseline methods, except for original data). The best settings for the weights of the loss terms included in our method, which were determined based on performance on the validation data, were found to be $\lambda_{self} = 1.0$, $\lambda_{cross} = 1.0$, $\lambda_{cosine} = 3.0$, $\lambda_{joint} = 1.0$, $\lambda_{MCL} = 2.0$, $\lambda_{SCL} = 0.5$, and $\lambda_{orthog} = 0.5$.

6 Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The TCGA multi-omics data are available in the Broad GDAC Firehose repository, https://gdac.broadinstitute.org/ [115]. Clincal data were files ending in '.clin.merged.picked.txt' from the Clinical_Pick_Tier1. mRNA-seq data were files ending in '.uncv2.mRNAseq_RSEM_all.txt' from the mRNAseq_Preprocess. Methylation data were files ending in '.meth.by_mean.data.txt' from the Methylation_Preprocess. miRNA-seq data were files ending in '.miRseq_RPKM_log2.txt' from miRseq_Preprocess. RPPA data were from files ending with '.rppa.txt' from RPPA_AnnotateWithGene.

The ADNI data are available in the Laboratory of Neuroimaging (LONI) Image & Data Archive (IDA), https://ida.loni.usc.edu/login.jsp [116], but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Baseline diagnoses and cognitive scores were taken from the ADNIMERGE data file for ADNI1/GO/2/3. Gene expression data were taken from the Microarray Gene Expression Profile Data. Genetic data were extracted for 935 SNPs derived from key AD GWAS studies [112, 113]. Metabolomics data were from the ADMC Duke Biocrates p180 Kit Flow injection analysis [ADNIGO,2] and ADMC Duke p180 Ultra Performance Liquid Chromatography [ADNIGO,2]. Proteomics data were from CruchagaLab CSF SOMAscan7k Protein matrix postQC [ADNI1,GO,2].

The code for JASMINE is available under the MIT license on GitHub with the URL https://github.com/jballard28/JASMINE. The software is platform independent and requires Python 3.8 or higher.

Competing interests

The authors declare that they have no competing interests.

Funding

Q.L. discloses support for the research of this work from NIH/NIA [grant numbers R01-AG071174 and RF1-AG063481], and NIH/NCI [grant number U01 CA274576]. L.S. discloses support for the research of this work from NIH/NIA [grant numbers U01-AG068057, U01 AG066833, R01-AG071470 and U19-AG074879].

Authors' contributions

J.L.B., Z.D., and Q.L. conceived the project. J.L.B. and Z.D. developed the methods and implemented the baseline methods. Z.D. conducted initial method implementation and experiments. J.L.B. executed all the analyses, including data preparation and preprocessing, method refinement and implementation, and numerical experiments.

J.L.B., Q.L., and L.S. contributed to the design of the numerical experiments and the interpretation of the results. J.L.B. generated the figures and wrote the manuscript with input from all authors. All authors contributed to editing and revising the manuscript. Q.L. and L.S. jointly supervised the project.

Acknowledgments

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; BristolMyers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

Data collection and sharing for this project was funded by the Alzheimer's Disease Metabolomics Consortium (National Institute on Aging R01AG046171, RF1AG051550 and 3U01AG024904-09S4).

References

[1] Wang, X., Yu, G., Wang, J., Zain, A.M., Guo, W.: Lung cancer subtype diagnosis using weakly-paired multi-omics data. Bioinformatics **38**(22), 5092–5099 (2022)

[2] Mo, Q., Wang, S., Seshan, V.E., Olshen, A.B., Schultz, N., Sander, C., Powers, R.S., Ladanyi, M., Shen, R.: Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences **110**(11), 4245–4250 (2013)

[3] Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P.W., Zhou, X.J.: Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic acids research **40**(19), 9379–9391 (2012)

[4] Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., Stegle, O.: Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. Molecular systems biology **14**(6), 8124 (2018)

[5] Tong, L., Mitchel, J., Chatlin, K., Wang, M.D.: Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. BMC Medical Informatics and Decision Making **20** (2020)

[6] Zhang, X., Xhang, J., Sun, K., Yang, X., Dai, C., Guo, Y.: Integrated multi-omics analysis using variational autoencoders: Application to pan-cancer classification. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2019)

[7] Hira, M.T., Razzaque, M.A., Angione, C., Scrivens, J., Sawan, S., Sarker, M.: Integrated multi-omics analysis of ovarian cancer using variational autoencoders. Scientific Reports **11** (2021)

[8] Sharifi-Noghabi, H., Zolotareva, O., Collins, C.C., Ester, M.: Moli: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics **35** (2019)

[9] Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., Huang, K.: Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nature Communications **12** (2021)

[10] Oh, M., Park, S., Lee, S., Lee, D., Lim, S., Jeong, D., Jo, K., Jung, I., Kim, S.: Drim: A web-based system for investigating drug response at the molecular level by condition-specific multi-omics data integration. Frontiers in Genetics **12** (2020)

[11] Bica, I., Veličković, P., Xiao, H., Liò, P.: Multi-omics data integration using cross-modal neural networks. European Symposium on Artificial Neural Networks 2018 proceedings (2018)

[12] Tu, X., Cao, Z.-J., Xia, C.-R., Mostafavi, S., Gao, G.: Cross-linked unified embedding for cross-modality representation learning. NeurIPS 2022 (2022)

[13] Wu, M., Goodman, N.: Multimodal generative models for scalable weakly-supervised learning. NIPS 2018 **32**, 5580–5590 (2018)

[14] Tong, L., Wu, H., Wang, M.D.: Integrating multi-omics data by learning modality invariant representations for improved prediction of overall survival of cancer. Methods **189** (2021)

[15] Ma, S., Zeng, A.G., Haibe-Kains, B., Goldenberg, A., Dick, J.E., Wang, B.: Moving towards genome-wide data integration for patient stratification with integrate any omics. Nature Machine Intelligence **7**(1), 29–42 (2025)

[16] Chen, Y., Wen, Y., Xie, C., Chen, X., He, S., Bo, X., Zhang, Z.: Mocss: Multi-omics data clustering and cancer subtyping via shared and specific representation learning. iScience **26** (2023)

[17] Lee, C., Schaar, M.: A variational information bottleneck approach to multi-omics data integration. Proceedings of Machine Learning Research **130**, 1513–1521 (2021)

[18] Wang, C., O'Connell, M.J.: Autoencoders with shared and specific embeddings for multi-omics data integration. bioRxiv. Pages: 2024.08.14.607979 Section: New Results. https://doi.org/10.1101/2024.08.14.607979 . https://www.biorxiv.org/content/10.1101/2024.08.14.607979v1 Accessed 2024-11-22

[19] Flores, J.E., Claborne, D.M., Weller, Z.D., Webb-Robertson, B.-J.M., Waters, K.M., Bramer, L.M.: Missing data in multi-omics integration: Recent advances through artificial intelligence. Journal of Agricultural, Biological and Environmental Statistics (2024)

[20] Lotfollahi, M., Litinetskaya, A., Theis, F.J.: Multigrate: single-cell multi-omic data integration. bioRxiv (2022)

[21] Shi, Y., Siddharth, N., Paige, B., Torr, P.H.S.: Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. arXiv. http://arxiv.org/abs/1911.03393 Accessed 2024-03-30

[22] Zuo, C., Dai, H., Chen, L.: Deep cross-omics cycle attention model for joint analysis of single-cell multi-omics data. Bioinformatics **37** (2021)

[23] Moon, S., Lee, H.: Moma: a multi-task attention learning algorithm for multi-omics data interpretation and classification. Bioinformatics **38**(8), 2287–2296 (2022)

[24] Huang, Z., Zhan, X., Xiang, S., Johnson, T.S., Helm, B., Yu, C.Y., Zhang, J., Salama, P., Rizkalla, M., Han, Z., *et al.*: Salmon: survival analysis learning with multi-omics neural networks on breast cancer. Frontiers in genetics **10**, 166 (2019)

[25] Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., Huang, K.: Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nature communications **12**(1), 3445 (2021)

[26] Ahmed, K.T., Sun, J., Cheng, S., Yong, J., Zhang, W.: Multi-omics data integration by generative adversarial network. Bioinformatics **38**(1), 179–186 (2022)

[27] Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. PMLR **27**, 1247–1255 (2013)

[28] Afshin-Pour, B., Hossein-Zadeh, G.-A., Strother, S.C., Soltanian-Zadeh, H.: Enhancing reproducibility of fmri statistical maps using generalized canonical correlation analysis in npairs framework. NeuroImage **40**, 1970–1981 (2012)

[29] Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The cancer genome atlas pan-cancer analysis project **45**(10), 1113–1120 https://doi.org/10.1038/ng.2764

[30] Thomas, P.D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L.-P., Mi, H.: Panther: Making genome-scale phylogenetics accessible to all. Protein Science **31**(1), 8–22 (2022)

[31] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M.,

Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.*: Gene ontology: tool for the unification of biology. Nature genetics **25**(1), 25–29 (2000)

[32] Consortium, T.G.O.: The gene ontology knowledgebase in 2023. Genetics **224**(1), 031 (2023)

[33] Milacic, M., Beavers, D., Conley, P., Gong, C., Gillespie, M., Griss, J., Haw, R., Jassal, B., Matthews, L., May, B., *et al.*: The reactome pathway knowledgebase 2024. Nucleic acids research **52**(D1), 672–678 (2024)

[34] Aparicio-Puerta, E., Hirsch, P., Schmartz, G.P., Kern, F., Fehlmann, T., Keller, A.: mieaa 2023: updates, new functional microrna sets and improved enrichment visualizations. Nucleic acids research **51**(W1), 319–325 (2023)

[35] Cui, C., Zhong, B., Fan, R., Cui, Q.: Hmdd v4. 0: a database for experimentally supported human microrna-disease associations. Nucleic Acids Research **52**(D1), 1327–1332 (2024)

[36] De Luca, A., Maiello, M.R., D'Alessio, A., Pergameno, M., Normanno, N.: The ras/raf/mek/erk and the pi3k/akt signalling pathways: role in cancer pathogenesis and implications for therapeutic approaches. Expert opinion on therapeutic targets **16**(sup2), 17–27 (2012)

[37] Burotto, M., Chiou, V.L., Lee, J.-M., Kohn, E.C.: The mapk pathway across different malignancies: a new perspective. Cancer **120**(22), 3446–3456 (2014)

[38] Dhillon, A.S., Hagan, S., Rath, O., Kolch, W.: Map kinase signalling pathways in cancer. Oncogene **26**(22), 3279–3290 (2007)

[39] Bahar, M.E., Kim, H.J., Kim, D.R.: Targeting the ras/raf/mapk pathway for cancer therapy: from mechanism to clinical studies. Signal transduction and

targeted therapy **8**(1), 455 (2023)

[40] Carrière, A., Cargnello, M., Julien, L.-A., Gao, H., Bonneil, É., Thibault, P., Roux, P.P.: Oncogenic mapk signaling stimulates mtorc1 activity by promoting rsk-mediated raptor phosphorylation. Current biology **18**(17), 1269–1277 (2008)

[41] Wenzel, J., Rose, K., Haghighi, E.B., Lamprecht, C., Rauen, G., Freihen, V., Kesselring, R., Boerries, M., Hecht, A.: Loss of the nuclear wnt pathway effector tcf7l2 promotes migration and invasion of human colorectal cancer cells. Oncogene **39**(19), 3893–3909 (2020)

[42] Zhao, H., Ming, T., Tang, S., Ren, S., Yang, H., Liu, M., Tao, Q., Xu, H.: Wnt signaling in colorectal cancer: pathogenic role and therapeutic target. Molecular cancer **21**(1), 144 (2022)

[43] Huang, D., Wang, Y., Thompson, J.W., Yin, T., Alexander, P.B., Qin, D., Mudgal, P., Wu, H., Liang, Y., Tan, L., *et al.*: Cancer-cell-derived gaba promotes β-catenin-mediated tumour growth and immunosuppression. Nature cell biology **24**(2), 230–241 (2022)

[44] Huang, D., Alexander, P.B., Li, Q.-J., Wang, X.-F.: Gabaergic signaling beyond synapses: an emerging target for cancer therapy. Trends in cell biology **33**(5), 403–412 (2023)

[45] Dong, Y., Wang, G., Nie, D., Xu, Y., Bai, X., Lu, C., Jian, F., Wang, H., Zheng, X.: Tumor-derived gaba promotes lung cancer progression by influencing tams polarization and neovascularization. International Immunopharmacology **126**, 111217 (2024)

[46] Kleinrok, Z., Matuszek, M., Jesipowicz, J., Matuszek, B., Opolski, A., Radzikowski, C.: Gaba content and gad activity in colon tumors taken from

patients with colon cancer or from xenografted human colon cancer cells growing as sc tumors in athymic nu-nu mice. Journal of physiology and pharmacology **49**(2) (1998)

[47] Matuszek, M., Jesipowicz, M., Kleinrok, Z.: Gaba content and gad activity in gastric cancer. Medical science monitor: international medical journal of experimental and clinical research **7**(3), 377–381 (2001)

[48] Mazurkiewicz, M., Opolski, A., Wietrzyk, J., Radzikowski, C., Kleinrok, Z.: Gaba level and gad activity in human and mouse normal and neoplastic mammary gland. Journal of experimental & clinical cancer research: CR **18**(2), 247–253 (1999)

[49] Jayachandran, P., Battaglin, F., Strelez, C., Lenz, A., Algaze, S., Soni, S., Lo, J.H., Yang, Y., Millstein, J., Zhang, W., *et al.*: Breast cancer and neurotransmitters: emerging insights on mechanisms and therapeutic directions. Oncogene **42**(9), 627–637 (2023)

[50] Yan, L., Gong, Y.-Z., Shao, M.-N., Ruan, G.-T., Xie, H.-L., Liao, X.-W., Wang, X.-K., Han, Q.-F., Zhou, X., Zhu, L.-C., *et al.*: Distinct diagnostic and prognostic values of γ-aminobutyric acid type a receptor family genes in patients with colon adenocarcinoma. Oncology Letters **20**(1), 275–291 (2020)

[51] Yang, L., Zhu, J., Wang, L., He, L., Gong, Y., Luo, Q.: A novel risk score model based on gamma-aminobutyric acid signature predicts the survival prognosis of patients with breast cancer. Frontiers in Oncology **13**, 1108823 (2023)

[52] Gao, X., Jia, X., Xu, M., Xiang, J., Lei, J., Li, Y., Lu, Y., Zuo, S.: Regulation of gamma-aminobutyric acid transaminase expression and its clinical significance in hepatocellular carcinoma. Frontiers in Oncology **12**, 879810 (2022)

[53] Briukhovetska, D., Dörr, J., Endres, S., Libby, P., Dinarello, C.A., Kobold, S.: Interleukins in cancer: from biology to therapy. Nature Reviews Cancer **21**(8), 481–499 (2021)

[54] Podolska, M.J., Grützmann, R., Pilarsky, C., Bénard, A.: Il-3: key orchestrator of inflammation. Frontiers in immunology **15**, 1411047 (2024)

[55] Dougan, M., Dranoff, G., Dougan, S.K.: Gm-csf, il-3, and il-5 family of cytokines: regulators of inflammation. Immunity **50**(4), 796–811 (2019)

[56] Testa, U., Riccioni, R., Militi, S., Coccia, E., Stellacci, E., Samoggia, P., Latagliata, R., Mariani, G., Rossini, A., Battistini, A., *et al.*: Elevated expression of il-3rα in acute myelogenous leukemia is associated with enhanced blast proliferation, increased cellularity, and poor prognosis. Blood, The Journal of the American Society of Hematology **100**(8), 2980–2988 (2002)

[57] Zhou, Z., Sun, B., Huang, S., Yu, D., Zhang, X.: Roles of aminoacyl-trna synthetase-interacting multi-functional proteins in physiology and cancer. Cell Death & Disease **11**(7), 579 (2020)

[58] Yoon, I., Kim, U., Choi, J., Kim, S.: Disease association and therapeutic routes of aminoacyl-trna synthetases. Trends in Molecular Medicine **30**(1), 89–105 (2024)

[59] Park, S.G., Schimmel, P., Kim, S.: Aminoacyl trna synthetases and their connections to disease. Proceedings of the National Academy of Sciences **105**(32), 11043–11049 (2008)

[60] Jin, Y., Huang, X., Wang, Z., Xu, H., Wang, Q., Wang, L.: Aminoacyl-trna synthetase-based prognosis model and exploration of potential mechanisms in hepatocellular carcinoma. Journal of clinical and translational hepatology **11**(4), 877 (2023)

[61] Wang, J., Yang, X.-L.: Novel functions of cytoplasmic aminoacyl-trna synthetases shaping the hallmarks of cancer. The Enzymes **48**, 397–423 (2020)

[62] Wang, J., Vallee, I., Dutta, A., Wang, Y., Mo, Z., Liu, Z., Cui, H., Su, A.I., Yang, X.-L.: Multi-omics database analysis of aminoacyl-trna synthetases in cancer. Genes **11**(11), 1384 (2020)

[63] Kim, S., You, S., Hwang, D.: Aminoacyl-trna synthetases and tumorigenesis: more than housekeeping. Nature Reviews Cancer **11**(10), 708–718 (2011)

[64] Young, H.J., Lee, J.W., Kim, S.: Function of membranous lysyl-trna synthetase and its implication for tumorigenesis. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics **1864**(12), 1707–1713 (2016)

[65] Kim, D.G., Choi, J.W., Lee, J.Y., Kim, H., Oh, Y.S., Lee, J.W., Tak, Y.K., Song, J.M., Razin, E., Yun, S.-H., *et al.*: Interaction of two translational components, lysyl-trna synthetase and p40/37lrp, in plasma membrane promotes laminin-dependent cell migration. The FASEB Journal **26**(10), 4142–4159 (2012)

[66] Zhang, R., Noordam, L., Ou, X., Ma, B., Li, Y., Das, P., Shi, S., Liu, J., Wang, L., Li, P., *et al.*: The biological process of lysine-trna charging is therapeutically targetable in liver cancer. Liver International **41**(1), 206–219 (2021)

[67] Park, S.G., Kim, H.J., Min, Y.H., Choi, E.-C., Shin, Y.K., Park, B.-J., Lee, S.W., Kim, S.: Human lysyl-trna synthetase is secreted to trigger proinflammatory response. Proceedings of the National Academy of Sciences **102**(18), 6356–6361 (2005)

[68] Suh, J.H., Park, M.C., Goughnour, P.C., Min, B.S., Kim, S.B., Lee, W.Y., Cho, Y.B., Cheon, J.H., Lee, K.Y., Nam, D.-H., *et al.*: Plasma lysyl-trna synthetase 1 (kars1) as a novel diagnostic and monitoring biomarker for colorectal cancer.

Journal of clinical medicine **9**(2), 533 (2020)

[69] Kim, B.-h., Jung, W.Y., Lee, H., Kang, Y., Jang, Y.-J., Hong, S.W., Jeong, H.-j., Yoon, S.O.: Lysyl-trna synthetase (krs) expression in gastric carcinoma and tumor-associated inflammation. Annals of surgical oncology **21**, 2020–2027 (2014)

[70] Chen, Y., Wang, X.: mirdb: an online database for prediction of functional microrna targets. Nucleic acids research **48**(D1), 127–131 (2020)

[71] Liu, W., Wang, X.: Prediction of functional microrna targets by integrative modeling of microrna binding and target expression data. Genome biology **20**, 1–10 (2019)

[72] Estep, A.L., Palmer, C., McCormick, F., Rauen, K.A.: Mutation analysis of braf, mek1 and mek2 in 15 ovarian cancer cell lines: implications for therapy. PloS one **2**(12), 1279 (2007)

[73] Murugan, A.K., Dong, J., Xie, J., Xing, M.: Mek1 mutations, but not erk2 mutations, occur in melanomas and colon carcinomas, but none in thyroid carcinomas. Cell cycle **8**(13), 2122–2124 (2009)

[74] Chuang, J., Wang, C., Guo, Y., Valenzuela, V., Wu, J., Fakih, M.: Map2k1 mutations in advanced colorectal cancer predict poor response to anti-egfr therapy and to vertical targeting of mapk pathway. Clinical Colorectal Cancer **20**(1), 72–78 (2021)

[75] Sasaki, H., Hikosaka, Y., Kawano, O., Moriyama, S., Yano, M., Fujii, Y.: Mek1 and akt2 mutations in japanese lung cancer. Journal of Thoracic Oncology **5**(5), 597–600 (2010)

[76] Marks, J.L., Gong, Y., Chitale, D., Golas, B., McLellan, M.D., Kasai, Y., Ding, L., Mardis, E.R., Wilson, R.K., Solit, D., *et al.*: Novel mek1 mutation identified by mutational analysis of epidermal growth factor receptor signaling pathway genes in lung adenocarcinoma. Cancer research **68**(14), 5524–5528 (2008)

[77] Mizuno, S., Ikegami, M., Koyama, T., Sunami, K., Ogata, D., Kage, H., Yanagaki, M., Ikeuchi, H., Ueno, T., Tanikawa, M., *et al.*: High-throughput functional evaluation of map2k1 variants in cancer. Molecular Cancer Therapeutics **22**(2), 227–239 (2023)

[78] Feng, C., Li, P., Liu, P., Wang, B., Li, J., Liu, P.: Lman2 promotes breast cancer tumorigenesis and drug resistance by interacting with mapk9 via activation of the mapk pathway. Cancer Medicine **13**(23), 70448 (2024)

[79] Yang, X., Jia, Q., Liu, X., Wu, W., Han, Y., Zou, Z., Li, M., Fan, D., Song, J., Chen, L.: Mapk9 is correlated with a poor prognosis and tumor progression in glioma. Frontiers in Bioscience-Landmark **28**(3), 63 (2023)

[80] Luo, C.-G., Zhang, J., An, Y.-Z., Liu, X., Li, S.-J., Zhang, W., Li, K., Zhao, X., Yuan, D.-B., AN, L.-Y., et al.: Mapk9 as a therapeutic target: unveiling ferroptosis in localized prostate cancer progression. Biocell **48**(5) (2024)

[81] Mo, S., Pan, B., Qin, L., Wang, S., Wang, Y., Xu, M., Xia, F., Gu, W.: Neurodegeneration promotes tumorigenesis in colorectal cancer: Insights from single-cell and spatial multiomics. JCO Precision Oncology (9), 2400744 (2025) https://doi.org/10.1200/PO-24-00744 https://ascopubs.org/doi/pdf/10.1200/PO-24-00744. PMID:

[82] Zheng, X., Wang, Y., Qiu, X.: Comprehensive analysis of mapk genes in the

prognosis, immune characteristics, and drug treatment of renal clear cell carcinoma using bioinformatic analysis and mendelian randomization. European Journal of Pharmacology **980**, 176840 (2024)

[83] Yeh, C.-H., Bellon, M., Nicot, C.: Fbxw7: a critical tumor suppressor of human cancers. Molecular cancer **17**, 1–19 (2018)

[84] Fan, J., Bellon, M., Ju, M., Zhao, L., Wei, M., Fu, L., Nicot, C.: Clinical significance of fbxw7 loss of function in human cancers. Molecular cancer **21**(1), 87 (2022)

[85] Chen, S., Lin, J., Zhao, J., Lin, Q., Liu, J., Wang, Q., Mui, R., Ma, L.: Fbxw7 attenuates tumor drug resistance and enhances the efficacy of immunotherapy. Frontiers in Oncology **13**, 1147239 (2023)

[86] Qi, Y., Rezaeian, A.-H., Wang, J., Huang, D., Chen, H., Inuzuka, H., Wei, W.: Molecular insights and clinical implications for the tumor suppressor role of scffbxw7 e3 ubiquitin ligase. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 189140 (2024)

[87] Shen, W., Zhou, Q., Peng, C., Li, J., Yuan, Q., Zhu, H., Zhao, M., Jiang, X., Liu, W., Ren, C.: Fbxw7 and the hallmarks of cancer: underlying mechanisms and prospective strategies. Frontiers in Oncology **12**, 880077 (2022)

[88] Zhong, Z., Virshup, D.M.: Recurrent mutations in tumor suppressor fbxw7 bypass wnt/β-catenin addiction in cancer. Science Advances **10**(14), 1031 (2024)

[89] Jardim, D.L., Wheler, J.J., Hess, K., Tsimberidou, A.M., Zinner, R., Janku, F., Subbiah, V., Naing, A., Piha-Paul, S.A., Westin, S.N., *et al.*: Fbxw7 mutations in patients with advanced cancers: clinical and molecular characteristics and outcomes with mtor inhibitors. PloS one **9**(2), 89388 (2014)

[90] Rossi, D.: Fbxw7 is a biologically validated cancer driver gene for cll. Blood, The Journal of the American Society for Hematology **133**(8), 774–776 (2019)

[91] Matallanas, D., Birtwistle, M., Romano, D., Zebisch, A., Rauch, J., Kriegsheim, A., Kolch, W.: Raf family kinases: old dogs have learned new tricks. Genes & cancer **2**(3), 232–260 (2011)

[92] Lim, S.H., Jung, J., Hong, J.Y., Kim, S.T., Park, S.H., Park, J.O., Kim, K.-M., Lee, J.: prevalence of raf1 aberrations in metastatic cancer patients: Real-world data. Biomedicines **11**(12), 3264 (2023)

[93] Bekele, R.T., Samant, A.S., Nassar, A.H., So, J., Garcia, E.P., Curran, C.R., Hwang, J.H., Mayhew, D.L., Nag, A., Thorner, A.R., et al.: Raf1 amplification drives a subset of bladder tumors and confers sensitivity to mapk-directed therapeutics. The Journal of Clinical Investigation **131**(22) (2021)

[94] Dorard, C., Madry, C., Buhard, O., Toifl, S., Didusch, S., Ratovomanana, T., Letourneur, Q., Dolznig, H., Garnett, M.J., Duval, A., *et al.*: Raf1 contributes to cell proliferation and stat3 activation in colorectal cancer independently of microsatellite and kras status. Oncogene **42**(20), 1649–1660 (2023)

[95] Zhang, L., Zhang, S., Cao, X., Shi, J., Zhao, S., Tian, J., Xiao, K., Wang, M., Liu, J., Wang, C., *et al.*: Raf1 facilitates kit signaling and serves as a potential treatment target for gastrointestinal stromal tumor. Oncogene **43**(27), 2078–2091 (2024)

[96] Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., Morris, J.C., Petersen, R.C., Saykin, A.J., Schmidt, M.E., Shaw, L., Shen, L., Siuciak, J.A., Soares, H., Toga, A.W., Trojanowski, J.Q., Alzheimer, s.D.N.I.: The alzheimer's disease

neuroimaging initiative: A review of papers published since its inception **9**(5), 111–194 https://doi.org/10.1016/j.jalz.2013.05.1769 . Publisher: John Wiley & Sons, Ltd. Accessed 2022-09-02

[97] Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Morris, J.C., Petersen, R.C., Saykin, A.J., Shaw, L.M., Toga, A.W., Trojanowski, J.Q., Alzheimer's Disease Neuroimaging Initiative: Recent publications from the alzheimer's disease neuroimaging initiative: Reviewing progress toward improved AD clinical trials **13**(4), 1–85 https://doi.org/10.1016/j.jalz.2016.11.007

[98] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[99] Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., Wang, B.: scGPT: toward building a foundation model for single-cell multi-omics using generative AI, 1–11 https://doi.org/10.1038/s41592-024-02201-0 . Publisher: Nature Publishing Group. Accessed 2024-04-27

[100] Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter **29**(9), 2307–2316 https://doi.org/10.1038/s41591-023-02504-3 . Publisher: Nature Publishing Group. Accessed 2024-12-09

[101] Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence **616**(7956), 259–265 https://doi.org/10.1038/s41586-023-05881-4 . Publisher: Nature Publishing Group. Accessed 2024-12-09

[102] Choudhary, N., Rao, N., Katariya, S., Subbian, K., Reddy, C.: Probabilistic entity representation model for reasoning over knowledge graphs. In: Advances in Neural Information Processing Systems, vol. 34, pp. 23440–23451. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2021/hash/c4d2ce3f3ebb5393a77c33c0cd95dc93-Abstract.html Accessed 2024-11-22

[103] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. arXiv. https://doi.org/10.48550/arXiv.2002.05709 . http://arxiv.org/abs/2002.05709 Accessed 2024-11-22

[104] Wang, X., Jing, L., Liu, H., Yu, J.: Structure-driven representation learning for deep clustering **18**(1), 31–13125 https://doi.org/10.1145/3623400 . Accessed 2024-11-22

[105] Bahri, D., Jiang, H., Tay, Y., Metzler, D.: SCARF: Self-Supervised Contrastive Learning using Random Feature Corruption. arXiv. https://doi.org/10.48550/arXiv.2106.15147 . http://arxiv.org/abs/2106.15147 Accessed 2024-11-22

[106] Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)

[107] Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research **20**(177), 1–81 (2019)

[108] Bansal, N., Blum, A., Chawla, S.: Correlation clustering. Machine learning **56**, 89–113 (2004)

[109] Hsu, H.-H., Hsieh, C.-W., *et al.*: Feature selection via correlation coefficient clustering. J. Softw. **5**(12), 1371–1377 (2010)

[110] Haq, A.U., Zhang, D., Peng, H., Rahman, S.U.: Combining multiple feature-ranking techniques and clustering of variables for feature selection. Ieee Access **7**, 151482–151492 (2019)

[111] Sioutos, N., Coronado, S., Haber, M.W., Hartel, F.W., Shaiu, W.-L., Wright, L.W.: Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information. Journal of biomedical informatics **40**(1), 30–43 (2007)

[112] Jansen, I.E., Savage, J.E., Watanabe, K., Bryois, J., Williams, D.M., Steinberg, S., Sealock, J., Karlsson, I.K., Hägg, S., Athanasiu, L., *et al.*: Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk. Nature genetics **51**(3), 404–413 (2019)

[113] Kunkle, B.W., Grenier-Boley, B., Sims, R., Bis, J.C., Damotte, V., Naj, A.C., Boland, A., Vronskaya, M., Van Der Lee, S.J., Amlie-Wolf, A., *et al.*: Genetic meta-analysis of diagnosed alzheimer's disease identifies new risk loci and implicates aβ, tau, immunity and lipid processing. Nature genetics **51**(3), 414–430 (2019)

[114] Folstein, M.F., Folstein, S.E., McHugh, P.R.: "mini-mental state": A practical method for grading the cognitive state of patients for the clinician **12**(3), 189–198 https://doi.org/10.1016/0022-3956(75)90026-6 . Accessed 2024-12-03

[115] Broad GDAC Firehose. https://gdac.broadinstitute.org/ Accessed 2024-03-28

[116] Crawford: IDA - Image and Data Archive. https://ida.loni.usc.edu/login.jsp Accessed 2024-03-28