RNAlign: Alignment of tumor and cell line transcriptomes using conditional VAEs

Anonymous authors
Paper under double-blind review

Abstract

Preclinical cancer models such as cancer cell lines (CL) are central to cancer research but can poorly represent tumor samples due to fundamental differences like stromal cell contamination or in-vitro adaptation. This hinders the translation of new biomarkers or therapeutics into the clinical setting, leading to false leads, failed clinical trials, and the need for expensive multiomics pipelines to reconcile data sets. In this work, we build on conditional variational auto-encoders (CVAE) to enable the direct comparison or selection of the most representative CL for cancer research. We introduce RNAlign (pronounced *RNA-align*), a CVAE framework with novel regularization techniques, to enable pan-cancer alignment of tumor and CL gene expression profiles. The resulting learned transformation achieves state-of-the-art removal of the most significant differences between the model types, while preserving biologically important subtype information. This framework is extendable to other tumor models such as organoids and can be directly integrated into existing workflows to guide clinical precision medicine.

1 Introduction

Tumor models such as cell lines (CL) play a key role in understanding how tumors develop and respond to various perturbations. The genomic, transcriptomic, and epigenetic features of CLs have been extensively cataloged, establishing them as a platform for systematic discovery and subsequent testing of genetic and chemical vulnerabilities (Ghandi et al., 2019). Promising biomarkers identified through these studies are then advanced to animal or human models. However, few biomarkers successfully make this transition due to the challenges of imperfect translation to clinical settings (Butler, 2008; Seyhan, 2019; Lieu et al., 2013).

Direct comparison of tumors to CLs would enable matching patient profiles with appropriate vulnerabilities for precision medicine (Luebker et al., 2017; Najgebauer et al., 2018). Large-scale efforts such as the Cancer Genome Atlas Program (TCGA) and the Cancer Cell Line Encyclopedia (CCLE) have allowed these comparisons, though many of these efforts have been limited to single cancer types (Barretina et al., 2012; Weinstein et al., 2013; Virtanen et al., 2002; Kao et al., 2009; Marie Vincent & Postovit, 2017). The use of genomic data is often hampered by a lack of matched normal samples, so a framework to perform transcriptomic mapping of tumors to CLs would be beneficial to the use of existing data sets and downstream analyses. Tumor transcriptomics has been successful in distinguishing cancer types, subtype discovery, and identification of potential anti-cancer agents (Sørlie et al., 2001; Wigle et al., 2002; Aran et al., 2015; Yu et al., 2019). However, combining tumor and CL data poses several significant challenges. Firstly, cancer CL libraries represent an incomplete sampling of real-world cancer diversity and may be wrongly annotated (Sharifnia et al., 2017). Furthermore, cell culture conditions and ongoing genomic instability contribute to the discrepancy between CLs and tumors (Aran et al., 2015). Secondly, a major problem faced in tumor transcriptomics is the presence of contaminating stromal and immune cells at variable proportions (Buess et al., 2007). These non-cancer cells do not merely contribute additively to the expression counts – they are shown to participate in cross-talk with cancer cells (Elenbaas & Weinberg, 2001).

The disparity between tumors and CLs generates inefficiencies at each stage of drug development. Biomarkers discovered in models may ignore the variability in tumors and lead to clinical failures. To mitigate such costly failures in the drug research and development (R&D) pipeline, emerging alternatives are being explored, such as the use of advanced models (e.g. organoids) that better mimic

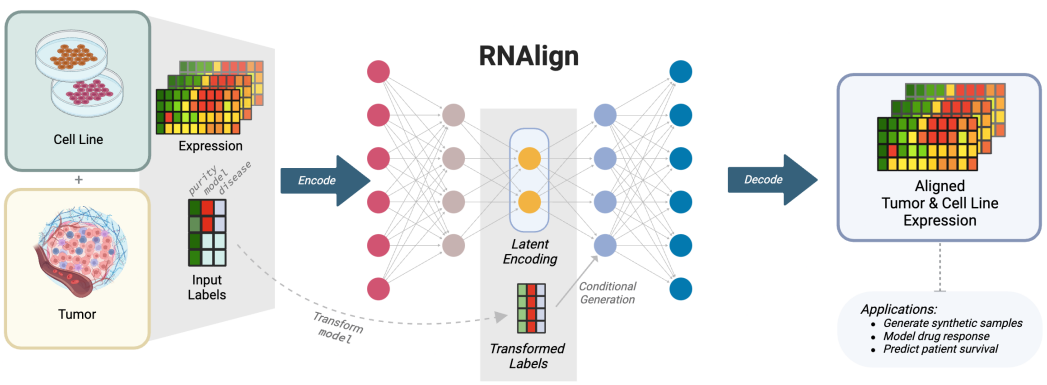

Figure 1: RNAlign, a CVAE (Sohn et al., 2015) with additional regularization, is trained on RNA-seq data from both tumor and CL, along with their associated class labels. Light gray boxes demarcate the concatenated inputs into the encoder and decoder. To align the tumor and CL data, during inference, model and purity labels to the trained decoder are homogenized.

the tumor micro-environment, along with consortium-led initiatives aimed at standardizing data generation processes across models (Koc et al., 2022). However, neither of these approaches leverage existing datasets that have already been extensively characterized, thereby precluding potentially significant clinical findings.

We hypothesize that a conditional variational auto-encoders (CVAE) framework can learn common biological patterns across different model types, and use conditional generation to align tumor and CL data (Sohn et al., 2015). In this work, we introduce RNAlign, a CVAE enhanced with novel regularization techniques for latent space disentanglement and conditional generation. We demonstrate that it outperforms existing methods in aligning a pan-cancer dataset, while preserving biological variability between cancer subtypes.

2 Related Works

Several works have been proposed to address the misalignment between existing tumor and CL data. CELLector features a multi-omics approach to evaluate and guide the selection of the appropriate in-vitro cancer model (Najgebauer et al., 2018). Newman et al. (2015) directly adjust expression values to remove the disparity between models, requiring the expression profiles of the contributing cell types. Zhang et al. (2020a) carry out batch correction using provided class annotations. Aran et al. (2015) address tumor heterogeneity in downstream analyses by using purity as a co-variate during differential expression analysis. Yu et al. (2019) use linear projection to remove stromal contamination and estimate cancer expression. Finally, to enable direct comparisons between tumor and CL data, Celligner is an unsupervised alignment method that performs a two-step statistical transformation to remove systematic differences between CL and tumor profiles (Warren et al., 2020).

3 The RNAlign Framework

3.1 Modeling Gene Expression and Class Labels using Regularized CVAEs

The RNAlign framework uses the probabilistic generative model CVAE to approximate the log-likelihood of the data x which is generated by the distribution $p_\theta(x|z,y)$ conditioned on the latent variable z and fully observed class labels y, using variational inference to maximize the Evidence Lower Bound (ELBO) (Kingma, 2013; Sohn et al., 2015; Esmaeili et al., 2018; Debbagh, 2023). The training loss for the CVAE decomposes into a reconstruction term and a KL divergence term:

$$\mathcal{L}_{\text{CVAE}} = -\mathbb{E}_{q(z|x,y)}[\log p(x|z,y)] + \beta D_{KL}(q(z|x,y)||p(z))$$

The hyper-parameter β on the KL divergence term allows fine-tuning of the trade-off between the two terms (Burgess et al., 2018). Additionally, the generative processes for class labels $p(y)$ (e.g. tumor or CL) and latent variables $p(z)$ are assumed to be independent. This encourages disentangling class information from the latent variable z, enabling conditional generation of new samples. Figure 1 depicts the CVAE comprising of an encoder $q(z|x, y)$ and a decoder $p(x|z, y)$. The encoder takes as input gene expression values x and class values y and outputs a latent variable z. $p(z)$ is parameterized as an isotropic Gaussian distribution with mean μ and standard deviation δ. During training, the latent variable is sampled as $z \sim \mathcal{N}(\mu, \delta)$ and concatenated with y as input to the decoder, to output gene expression values. CVAEs are suitable for the alignment task as they are able disentangle class labels from latent space (Zhang et al., 2020b). This architecture allows one to transform the class of a given sample by explicitly modifying the class label y (e.g. "CL" $\rightarrow$ "tumor" or "tumor" $\rightarrow$ "CL") and concatenating it to z_μ at the decoder.

We enhance class learning by proposing two novel regularization terms which control the relationship between the latent variables z and the class labels y. These regularization terms encourage disentanglement of y from the latent space and improve conditional generation from z (Appendix Table 5). First, to discourage encoding class information in the latent space z, we penalize the distance correlation between the latent variables z and the class variables y (Székely et al., 2007). In contrast to Pearson correlation, distance correlation ($\mathcal{R} \in [0, 1]$) captures non-linear dependencies, and equals zero if and only if the variables are independent. The distance correlation loss between the input class variables and the latent variables is defined as: $\mathcal{L}_{cor} = \mathcal{R}(y, z)$. This encourages the latent variable z to be independent of the class label y.

Furthermore, given the tendency of CVAEs to ignore class labels, we increase sensitivity of the decoder to changes in the class label by imposing a loss based on the L2 norm of the gradient of the ELBO with respect to y, which requires an additional back-propagation step to compute. The loss is defined as: $\mathcal{L}_{grad} = -\|\nabla_y \mathcal{L}_{ELBO}\|_2$. The total loss to be minimized during training is: $\mathcal{L}_{\text{TOTAL}} = \mathcal{L}_{\text{CVAE}} + \lambda_1 \mathcal{L}_{cor} + \lambda_2 \mathcal{L}_{grad}$.

3.2 Experimental settings

We implement the CVAE in Pyro (Bingham et al., 2018) with a symmetric encoder and decoder structure, comprising fully connected layers. ReLU activation and dropout layers are used between each of the hidden layers. The model was trained for 1000 epochs using the Adam optimizer (Kingma, 2014), along with a scheduler that reduces the learning rate by a factor of 10 at epochs 500 and 750. The number of hidden layers, latent dimensionality, dropout percentage, and learning rate are treated as hyper-parameters. An annealing schedule is used during training for the KL divergence loss term β to prevent posterior collapse (Fu et al., 2019). Random hyper-parameter sampling (n=20) was carried out to determine good values for the aforementioned hyper-parameters, including the parameters β and λ in the total loss function (Bergstra & Bengio, 2012). The model with the lowest total loss on the test set was chosen. The search ranges used to sample each hyper-parameter are recorded in Appendix Table 2.

The input to the encoder, x, is normalized log transcripts per million (log TPM) expression values for the 3000 genes with highest median absolute deviation (Appendix A.2). The input is further normalized to have zero mean and unit variance. y comprises one-hot encoded cancer type, sample type (CL vs tumor), and purity values for each sample estimated by PUREE (Revkov et al., 2023). Purity, the proportion of cancer cells in a sample, can be reliably and consistently estimated from gene expression data using computational methods (Aran et al., 2015). Cancer type (e.g. Breast Cancer) and sample type (e.g. CL) labels were used as input class labels for training and were treated as fully observed variables as this information is available for all samples.

4 Results

4.1 RNAlign improves global alignment of pan-cancer CL and tumor data

We trained RNAlign on 12,236 tumor samples from the TCGA, TREEHOUSE and TARGET datasets, along with 1,249 CL samples from CCLE (Goldman et al., 2018). Despite consistent processing of the input data, the marked differences between the raw expression values for tumors

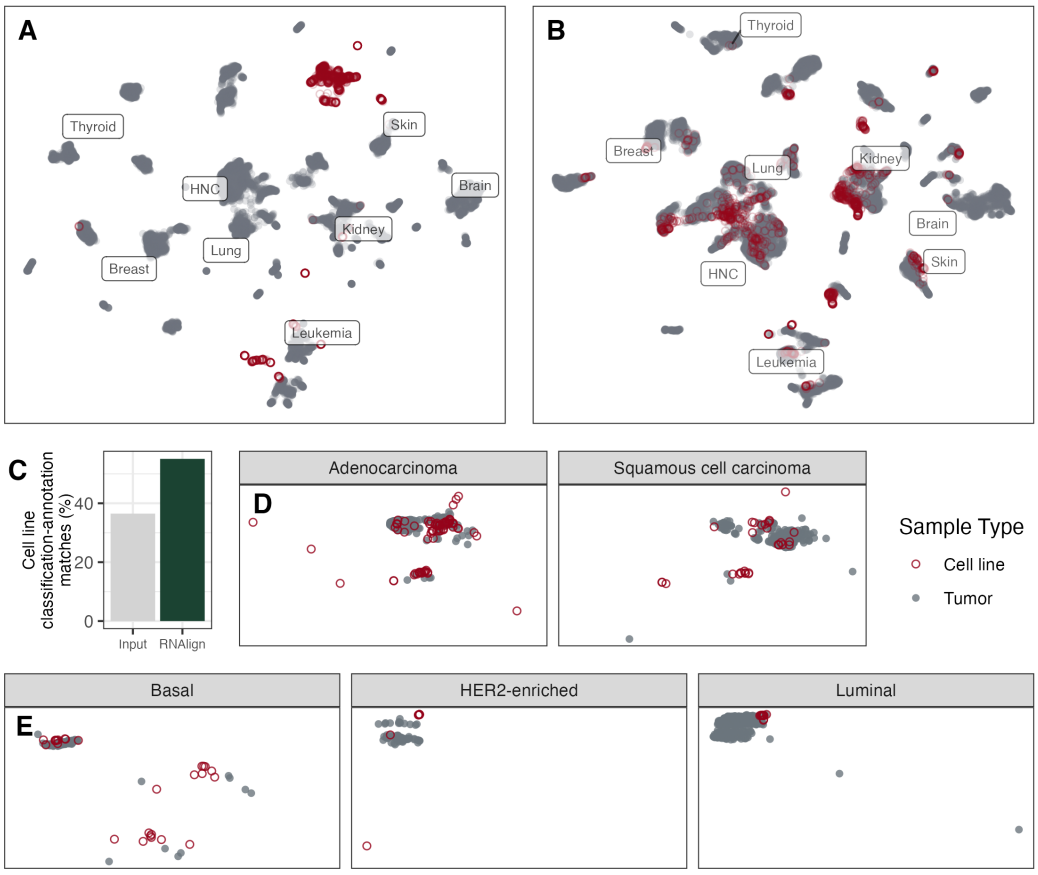

Figure 2: UMAP projections of (a) raw input data and (b) RNAlign-transformed data for 12,236 tumors and 1,249 CLs reflect major differences in input expression values from CLs and tumors, and few CLs cluster with their relevant cancer types. Transformation of the dataset results in clustering of tumors and CLs in their respective cancer types. Out of 26 cancer types, the clusters for the largest 8 are labeled. (c) Aligned data increases the median percentage of CLs matched to their appropriate tumor cancer-type. (d) Adenocarcinoma (n_{tumor}= 516; n_{cl}= 79) and Squamous cell carcinoma (n_{tumor}= 498; n_{cl}= 30) NSCLC samples, after transformation with the same pan-cancer model, show alignment of tumors and CLs to their respective disease types. (e) Transformed BRCA tumors and CLs cluster together in their respective subtypes. Basal (n_{tumor}= 190; n_{cl}= 27), HER2-enriched (n_{tumor}= 81; n_{cl}= 15), and Luminal (n_{tumor}= 770; n_{cl}= 14) subtypes are shown. The NSCLC and BRCA data are subsets of the pan-cancer aligned data.

and CLs of the same cancer types are shown in Figure 2A. Fundamental differences such as the presence of contaminant cells in tumor samples and CL in-vitro adaptations impede direct comparisons. To demonstrate this, we compared the cancer type of each CL sample to the majority cancer type of the 25 nearest tumor samples (Appendix A.10). Only 36.5% of CL samples were predominantly surrounded by tumor samples of the same cancer type (Figure 2C), highlighting inherent disparities between sample types. Per-cancer type percentages are displayed in Appendix Figure 4.

RNAlign transformation enhances global alignment, resulting in more CL samples clustering with tumors of the sample cancer type (Figure 2B). This improvement is reflected in the increase in percentage (55.1%) of CLs that align with their corresponding tumor cluster by cancer type in Figure 2C. This demonstrates that RNAlign enhances alignment between CLs and tumors. By incorporating class labels (e.g. sample purity, model type, disease class) into the CVAE, the latent variable z captures underlying gene expression variations shared across all samples, independent of class labels. To account for fundamental differences between tumors and CLs during inference, each tumor sample is decoded using a modified y variable by setting model = CL and purity = 1.

Method	ΔD	PVCA	ΔkBET
Input	22.56	0.27	0.90
Linear Projection	10.57	0.16	0.91
Celligner	8.59	**0.10**	0.86
RNAlign	**4.75**	0.13	**0.65**

Table 1: RNAlign performs best in metrics that test batch effect removal performance between tumors and CLs. The scores are calculated as the median across cancer types.

4.2 RNAlign transformation preserves biological subtype variability

The presence of disease-specific subtypes is a confounder that complicates pan-cancer alignment. During global alignment, there is no guarantee that subtype-specific variation will be preserved. This variation could instead be erroneously removed, resulting in crude alignment of innately distinct subtypes. However, RNAlign is able to retain subtype variability while aligning sample types. For example, transformation using RNAlign preserves distinct subtype clusters of non-small cell lung cancer (NSCLC) and breast cancer (BRCA) (Figure 2D and 2E, respectively). This highlights the ability of the model to preserve biological variability or local subtype structures. This subtype information is not provided during training, demonstrating that RNAlign is able to model intra-disease variability in an unsupervised manner.

4.3 Benchmarking and ablation analysis show RNAlign's effectiveness in removing sample type disparity

Next, we show that RNAlign outperforms similar methods in pan-cancer alignment of tumor (x_t) and CL (x_{cl}) expression data. To measure the relative performance of each method for a sample type j, we compute three complementary metrics to evaluate batch effect correction (Appendix A.4). First, to measure the compactness and separation of the transformed data per cancer type, we summarize the difference in Euclidean distance between intra-batch pairs (e.g. x_t–x_t, x_{cl}–x_{cl}) and inter-batch pairs (e.g. x_t–x_{cl}). All possible within- and between-sample type pairs are considered for each cancer type. We then compute $\Delta D(x_t, x_{cl}) = D_{intra}(x) - D_{inter}(x_t, x_{cl})$ (see Appendix A.3 for details). We also calculate ΔkBET (Büttner et al., 2019) to measure the local consistency of batch mixing. Higher ΔkBET values indicate small regions where batch effects persist. Lastly, PVCA (Boedigheimer et al., 2008) quantifies the residual variance attributable to batch effects. We calculate the metrics for each cancer type and take the median value. A well-adjusted correction should minimize all three metrics relative to the baseline unadjusted input data. RNAlign tops two out of three metrics (Table 1).

To investigate the effects of the regularization terms or input feature selection on the model performance, we carry out ablation by removing each of the following – $\mathcal{L}_{grad}$, $\mathcal{L}_{cor}$, or purity labels, We then train the model without the regularization or feature in question. Without any of the aforementioned terms, the model exhibits degraded performances of ΔDs of 6.12, 10.70, and 17.34 respectively (Appendix Table 5). RNAlign's performance benefits from both regularization terms; removing either term impairs its ability to remove batch effects. $\mathcal{L}_{grad}$ makes the model more sensitive to changes in the input sample type, while $\mathcal{L}_{cor}$ encourages the model to encode batch-independent biological information into the latent space. Purity is consistently highlighted as a confounding factor in tumor data (Aran et al., 2015) and its inclusion aids the performance of RNAlign.

The nature of CVAEs means that the generation of output data can be conditioned several ways (Sohn et al., 2015). Changing the model and purity labels give the best performance in alignment. We briefly summarize the performance of alternative transformations in Appendix Table 4.

4.4 Failures to align reflect known cancer biology

Established cell lines are known to poorly recapitulate tumor biology due to limited representation of subtypes or transcriptional states; some cancer types may also have highly unique tumor microenvironments in-vio. As such, a biologically relevant alignment should not indiscriminately align cancer types where cell lines are known to be poor proxies of tumors. To evaluate the performance

of RNAlign in this regard, we use the re-annotated cancer type labels for CL as outlined in Section 4.1 to summarize levels of concordance or discordance between aligned CLs and tumors.

High levels of discordance in known tumor biology between CL and tumor samples indeed lead to poor alignment in some cancer types. Though further analyses is needed to assess biological pathways that may contribute to misalignment in these cases, we highlight three cancer types for which the aligned data have low percentages of CL-tumor matching and show that literature review corroborates these results; the poor alignment mirrors known biological discrepancies between CLs and tumors (Appendix Figure 3, Appendix Figure 4).

First, brain CLs (12% of CLs cluster with tumors) exhibit a unique tumor micro-environment absent in CLs, which underlies the weak correlation of brain CLs to tumor samples (Marx, 2024). Numerous brain CLs are also often derived from metastatic or highly aggressive tumors, leading to lineage misrepresentation and poor fidelity to tumor samples (Ghandi et al., 2019). Next, thyroid CLs (8% CL-tumor match) also show poor fidelity to tumor samples. CL mRNA profiles have higher correlation to rare, de-differentiated anaplastic thyroid carcinoma samples, instead of the more common differentiated papillary thyroid carcinoma subtype (Saiselet et al., 2012). Lastly, liver cancer CLs (33% CL-tumor match) group into either hepatocyte-like, which aligns with most HCC tumors, or fibroblast-like clusters that show strong stromal contamination (Fukuyama et al., 2021). Fibroblast-like tumors comprise a heterogeneous population of cancer-associated fibroblasts, (Peng et al., 2022). This split is seen in Figure 3 (liver panel), with fibroblast-like samples clustering in the sparser and more diffuse cluster in the bottom right of the panel. We further summarize the literature for several other cancer types in Appendix A.10. RNAlign does not force arbitrary mixing between model types, and retains biologically salient disparities in these cases; failures of alignment post-transformation are concordant with known literature.

5 Discussion

The disparity between cancer models is a major challenge in translational oncology. RNAlign addresses this issue by globally aligning tumor and CL expression data. We show that the transformed data preserves subtype information without indiscriminately aligning biologically divergent samples. Our novel framework uses regularization terms and informative input features to disentangle known class information from the latent space, producing a more generalized encoding of cancer biology. Inclusion of cancer type labels is key to separating cancer-type-specific information from the latent space, enabling the CVAE to encode broad and consistent patterns of cancer gene expression. Tumor purity, a known confounder, is another important input feature (Aran et al., 2015).

RNAlign enables robust cross-model analyses, allowing preclinical models to be directly used for downstream translational applications. By comparing aligned data, precise CL selection for drug screening is possible, guided by its similarity to patient tumors. This ensures model fidelity, which could improve predictions of clinical responses based on in-vitro data.

A natural extension of RNAlign for translational oncology would be to use the disentangled latent space for survival prediction or drug response tasks. Existing methods use vanilla VAEs to predict patient survival (Apellániz et al., 2024; Rollo et al., 2025). An area for further improvement in RNAlign could be reducing its over-reliance on RNA-seq data and integrating epigenetic, proteomic, or mutational drivers of variation. Similar multi-modal approaches have been demonstrated in ovarian cancer (Hira et al., 2021).

6 Conclusion

We demonstrate that RNAlign, a CVAE framework with novel regularization strategies, disentangles biologically relevant latent features from model-specific variation. RNAlign learns a generalized mapping of cancer biology to successfully harmonize the fundamental differences between tumor and CL transcriptomes, while preserving important subtype-specific variation and known biological incompatibilities. RNAlign enables direct model comparisons to generate robust clinical findings that otherwise require expensive and rigorous R&D pipelines, and its flexibility supports future multi-omics integration and extension to prognostic predictions.

SOFTWARE AND DATA

We will publish the code to run RNAlign on GitHub.

REFERENCES

Patricia A. Apellániz, Juan Parras, and Santiago Zazo. Leveraging the variational bayes autoencoder for survival analysis. *Scientific Reports*, 14(1), October 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-76047-z. URL http://dx.doi.org/10.1038/s41598-024-76047-z.

Dvir Aran, Marina Sirota, and Atul J. Butte. Systematic pan-cancer analysis of tumour purity. *Nature Communications*, 6(1), December 2015. ISSN 2041-1723. doi: 10.1038/ncomms9971. URL http://dx.doi.org/10.1038/ncomms9971.

Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, Joseph Lehár, Gregory V. Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F. Berger, John E. Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa A. Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H. Engels, Jill Cheng, Guoying K. Yu, Jianjun Yu, Peter Aspesi, Melanie de Silva, Kalpana Jagtap, Michael D. Jones, Li Wang, Charles Hatton, Emanuele Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C. Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P. Mesirov, Stacey B. Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vic E. Myer, Barbara L. Weber, Jeff Porter, Markus Warmuth, Peter Finan, Jennifer L. Harris, Matthew Meyerson, Todd R. Golub, Michael P. Morrissey, William R. Sellers, Robert Schlegel, and Levi A. Garraway. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, March 2012. ISSN 1476-4687. doi: 10.1038/nature11003. URL http://dx.doi.org/10.1038/nature11003.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming, 2018. URL https://arxiv.org/abs/1810.09538.

Michael J Boedigheimer, Russell D Wolfinger, Michael B Bass, Pierre R Bushel, Jeff W Chou, Matthew Cooper, J Christopher Corton, Jennifer Fostel, Susan Hester, Janice S Lee, Fenglong Liu, Jie Liu, Hui-Rong Qian, John Quackenbush, Syril Pettit, and Karol L Thompson. Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. *BMC Genomics*, 9(1), June 2008. ISSN 1471-2164. doi: 10.1186/1471-2164-9-285. URL http://dx.doi.org/10.1186/1471-2164-9-285.

Martin Buess, Dimitry SA Nuyten, Trevor Hastie, Torsten Nielsen, Robert Pesich, and Patrick O Brown. Characterization of heterotypic interaction effects in vitro to deconvolute global gene expression profiles in cancer. *Genome Biology*, 8(9), September 2007. ISSN 1474-760X. doi: 10.1186/gb-2007-8-9-r191. URL http://dx.doi.org/10.1186/gb-2007-8-9-r191.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Declan Butler. Translational research: Crossing the valley of death. *Nature*, 453(7197):840–842, June 2008. ISSN 1476-4687. doi: 10.1038/453840a. URL http://dx.doi.org/10.1038/453840a.

Maren Büttner, Zhichao Miao, F Alexander Wolf, Sarah A Teichmann, and Fabian J Theis. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods*, 16(1):43–49, January 2019.

Mohamed Debbagh. Learning structured output representations from attributes using deep conditional generative models, 2023. URL https://arxiv.org/abs/2305.00980.

Xue-Man Dong, Lin Chen, Yu-Xin Xu, Pu Wu, Tian Xie, and Zhao-Qian Liu. Exploring metabolic reprogramming in esophageal cancer: the role of key enzymes in glucose, amino acid, and nucleotide pathways and targeted therapies. *Cancer Gene Therapy*, January 2025. ISSN 1476-5500. doi: 10.1038/s41417-024-00858-5. URL http://dx.doi.org/10.1038/s41417-024-00858-5.

Brian Elenbaas and Robert A. Weinberg. Heterotypic signaling between epithelial tumor cells and fibroblasts in carcinoma formation. *Experimental Cell Research*, 264(1):169–184, March 2001. ISSN 0014-4827. doi: 10.1006/excr.2000.5133. URL http://dx.doi.org/10.1006/excr.2000.5133.

Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, N. Siddharth, Brooks Paige, Dana H. Brooks, Jennifer Dy, and Jan-Willem van de Meent. Structured disentangled representations, 2018. URL https://arxiv.org/abs/1804.02086.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing, 2019. URL https://arxiv.org/abs/1903.10145.

Keita Fukuyama, Masataka Asagiri, Masahiro Sugimoto, Hiraki Tsushima, Satoru Seo, Kojiro Taura, Shinji Uemoto, and Keiko Iwaisako. Gene expression profiles of liver cancer cell lines reveal two hepatocyte-like and fibroblast-like clusters. *PLOS ONE*, 16(2):e0245939, February 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0245939. URL http://dx.doi.org/10.1371/journal.pone.0245939.

Mahmoud Ghandi, Franklin W. Huang, Judit Jané-Valbuena, Gregory V. Kryukov, Christopher C. Lo, E. Robert McDonald, Jordi Barretina, Ellen T. Gelfand, Craig M. Bielski, Haoxin Li, Kevin Hu, Alexander Y. Andreev-Drakhlin, Jaegil Kim, Julian M. Hess, Brian J. Haas, François Aguet, Barbara A. Weir, Michael V. Rothberg, Brenton R. Paolella, Michael S. Lawrence, Rehan Akbani, Yiling Lu, Hong L. Tiv, Prafulla C. Gokhale, Antoine de Weck, Ali Amin Mansour, Coyin Oh, Juliann Shih, Kevin Hadi, Yanay Rosen, Jonathan Bistline, Kavitha Venkatesan, Anupama Reddy, Dmitriy Sonkin, Manway Liu, Joseph Lehar, Joshua M. Korn, Dale A. Porter, Michael D. Jones, Javad Golji, Giordano Caponigro, Jordan E. Taylor, Caitlin M. Dunning, Amanda L. Creech, Allison C. Warren, James M. McFarland, Mahdi Zamanighomi, Audrey Kauffmann, Nicolas Stransky, Marcin Imielinski, Yosef E. Maruvka, Andrew D. Cherniack, Aviad Tsherniak, Francisca Vazquez, Jacob D. Jaffe, Andrew A. Lane, David M. Weinstock, Cory M. Johannessen, Michael P. Morrissey, Frank Stegmeier, Robert Schlegel, William C. Hahn, Gad Getz, Gordon B. Mills, Jesse S. Boehm, Todd R. Golub, Levi A. Garraway, and William R. Sellers. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569(7757):503–508, May 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1186-3. URL http://dx.doi.org/10.1038/s41586-019-1186-3.

Mary Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Akhil Kamath, Fran McDade, Dave Rogers, Angela N. Brooks, Jingchun Zhu, and David Haussler. The ucsc xena platform for public and private cancer genomics data visualization and interpretation. May 2018. doi: 10.1101/326470. URL http://dx.doi.org/10.1101/326470.

Muta Tah Hira, Muhammed Aminur Razzaque, Claudio Angione, James H. Scrivens, Saladin Sawan, Mosharraf Sarkar, and Muta Tah. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. *Scientific Reports*, 11, 2021. URL https://api.semanticscholar.org/CorpusID:232293178.

Martine J Jager, J Antonio Bermudez Magner, Bruce R Ksander, and Sander R Dubovy. Uveal melanoma cell lines: Where do they come from? (an american ophthalmological society thesis). *Trans. Am. Ophthalmol. Soc.*, 114:T5, August 2016.

Jessica Kao, Keyan Salari, Melanie Bocanegra, Yoon-La Choi, Luc Girard, Jeet Gandhi, Kevin A. Kwei, Tina Hernandez-Boussard, Pei Wang, Adi F. Gazdar, John D. Minna, and Jonathan R. Pollack. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PLoS ONE*, 4(7):e6146, July 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0006146. URL http://dx.doi.org/10.1371/journal.pone.0006146.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Soner Koc, Michael W Lloyd, Jeffrey W Grover, Nan Xiao, Sara Seepo, Sai Lakshmi Subramanian, Manisha Ray, Christian Frech, John DiGiovanna, Phillip Webster, Steven Neuhauser, Anuj Srivastava, Xing Yi Woo, Brian J Sanderson, Brian White, Paul Lott, Lacey E Dobrolecki, Heidi Dowst, Matthew Bailey, Emilio Cortes-Sanchez, Sandra Scherer, Chieh-Hsiang Yang, Maihi Fujita, Zhengtao Chu, Ling Zhao, Andrew Butterfield, Argun Akcakanat, Gao Boning, Kurt Evans, Bingliang Fang, Don Gibbons, Vanessa Jensen, Dara Keener, Michael Kim, Scott Kopetz, Mourad Majidi, David Menter, John Minna, Hyunsil Park, Fei Yang, Brenda Timmons, Jing Wang, Shannon Westin, Timothy Yap, Jianhua Zhang, Ran Zhang, Min Jin Ha, Huiqin Chen, Yuanxin Xi, Luc Girard, Erkan Yucan, Bryce P Kirby, Bingbing Dai, Yi Xu, Alexey Sorokin, Kelly Gale, Jithesh Augustine, Stephen Scott, Ismail Meraz, Dylan Fingerman, Andrew Kossenkov, Qin Liu, Min Xiao, Jayamanna Wickramasinghe, Haiyin Lin, Eric Ramirez-Salazar, Kate Nathanson, Mike Tetzlaff, George Xu, Vashisht G Yennu-Nanda, Rebecca Aft, Jessica Andrews, Alicia Asaro, Song Cao, Feng Chen, Sherri Davies, John DiPersio, Ryan Fields, Steven Foltz, Katherine Fuh, Kian Lim, Jason Held, Jeremy Hoog, Reyka G Jayasinghe, Yize Li, Jinqin Luo, Cynthia Ma, Jay Mashl, Chia-Kuei Mo, Fernanda Rodriguez, Hua Sun, Nadezhda V Terekhanova, Rose Tipton, Brian VanTine, Andrea Wang-Gillam, Mike Wendl, Yige Wu, Matt Wyczalkowski, Lijun Yao, Daniel Cui Zhou, Matthew Ellis, Michael Ittmann, Susan Hilsenbeck, Bert O'Malley, Amanda Kirane, May Cho, David Gandara, Jonathan Reiss, Tiffany Le, Ralph De Vere White, Cliff Tepper, David Cooke, Luis Godoy, Lisa Brown, Marc Dall'Era, Christopher Evans, Rashmi Verma, Sepideh Gholami, David J Segal, John Albeck, Edward Pugh, Susan Stewart, David Rocke, Hongyong Zhang, Nicole Coggins, Ana Estrada, Ted Toal, Alexa Morales, Guadalupe Polanco Echeverry, Sienna Rocha, Ai-Hong Ma, Yvonne A Evrard, Tiffany A Wallace, Jeffrey A Moscow, James H Doroshow, Nicholas Mitsiades, Salma Kaochar, Chong-xian Pan, Moon S Chen, Luis Carvajal-Carmona, Alana L Welm, Bryan E Welm, Michael T Lewis, Ramaswamy Govindan, Li Ding, Shunqiang Li, Meenhard Herlyn, Michael A Davies, Jack Roth, Funda Meric-Bernstam, Peter N Robinson, Carol J Bult, Brandi Davis-Dusenbery, Dennis A Dean, and Jeffrey H Chuang. Pdxnet portal: patient-derived xenograft model, data, workflow and tool discovery. *NAR Cancer*, 4(2), April 2022. ISSN 2632-8674. doi: 10.1093/narcan/zcac014. URL http://dx.doi.org/10.1093/narcan/zcac014.

C. H. Lieu, A.-C. Tan, S. Leong, J. R. Diamond, and S. G. Eckhardt. From bench to bedside: Lessons learned in translating preclinical studies in cancer drug development. *JNCI Journal of the National Cancer Institute*, 105(19):1441–1456, September 2013. ISSN 1460-2105. doi: 10.1093/jnci/djt209. URL http://dx.doi.org/10.1093/jnci/djt209.

Stephen A. Luebker, Weiwei Zhang, and Scott A. Koepsell. Comparing the genomes of cutaneous melanoma tumors to commercially available cell lines. *Oncotarget*, 8(70):114877–114893, December 2017. ISSN 1949-2553. doi: 10.18632/oncotarget.22928. URL http://dx.doi.org/10.18632/oncotarget.22928.

Krista Marie Vincent and Lynne-Marie Postovit. Investigating the utility of human melanoma cell lines as tumour models. *Oncotarget*, 8(6):10498–10509, January 2017. ISSN 1949-2553. doi: 10.18632/oncotarget.14443. URL http://dx.doi.org/10.18632/oncotarget.14443.

Vivien Marx. Closing in on cancer heterogeneity with organoids. *Nature Methods*, 21(4):551–554, March 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02231-8. URL http://dx.doi.org/10.1038/s41592-024-02231-8.

Hanna Najgebauer, Mi Yang, Hayley E. Francies, Clare Pacini, Euan A Stronach, Mathew J. Garnett, J. Saez-Rodriguez, and Francesco Iorio. Cellector: Genomics guided selection of cancer in vitro models. *bioRxiv*, 2018. URL https://api.semanticscholar.org/CorpusID:90842635.

Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453–457, March 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3337. URL http://dx.doi.org/10.1038/nmeth.3337.

Hao Peng, Erwei Zhu, and Yewei Zhang. Advances of cancer-associated fibroblasts in liver cancer. *Biomarker Research*, 10(1), August 2022. ISSN 2050-7771. doi: 10.1186/s40364-022-00406-z. URL http://dx.doi.org/10.1186/s40364-022-00406-z.

Egor Revkov, Tanmay Kulshrestha, Ken Wing-Kin Sung, and Anders Jacobsen Skanderup. Puree: accurate pan-cancer tumor purity estimation from gene expression data. *Communications Biology*, 6(1), April 2023. ISSN 2399-3642. doi: 10.1038/s42003-023-04764-8. URL http://dx.doi.org/10.1038/s42003-023-04764-8.

Cesare Rollo, Corrado Pancotti, Flavio Sartori, Isabella Caranzano, Saverio D'Amico, Luciana Carota, Francesco Casadei, Giovanni Birolo, Luca Lanino, Elisabetta Sauta, Gianluca Asti, Alessandro Buizza, Mattia Delleani, Elena Zazzetti, Marilena Bicchieri, Giulia Maggioni, Pierre Fenaux, Uwe Platzbecker, Maria Diez-Campelo, Torsten Haferlach, Gastone Castellani, Matteo Giovanni Della Porta, Piero Fariselli, and Tiziana Sanavia. Vae-surv: A novel approach for genetic-based clustering and prognosis prediction in myelodysplastic syndromes. *Computer Methods and Programs in Biomedicine*, 261:108605, April 2025. ISSN 0169-2607. doi: 10.1016/j.cmpb.2025.108605. URL http://dx.doi.org/10.1016/j.cmpb.2025.108605.

Manuel Saiselet, Sébastien Floor, Maxime Tarabichi, Geneviève Dom, Aline Hébrant, Wilma C. G. van Staveren, and Carine Maenhaut. Thyroid cancer cell lines: an overview. *Frontiers in Endocrinology*, 3, 2012. ISSN 1664-2392. doi: 10.3389/fendo.2012.00133. URL http://dx.doi.org/10.3389/fendo.2012.00133.

Vishesh Sarin, Katharine Yu, Ian D. Ferguson, Olivia Gugliemini, Matthew A. Nix, Byron Hann, Marina Sirota, and Arun P. Wiita. Evaluating the efficacy of multiple myeloma cell lines as models for patient tumors via transcriptomic correlation analysis. *Leukemia*, 34(10):2754–2765, March 2020. ISSN 1476-5551. doi: 10.1038/s41375-020-0785-1. URL http://dx.doi.org/10.1038/s41375-020-0785-1.

Attila A. Seyhan. Lost in translation: the valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. *Translational Medicine Communications*, 4 (1), November 2019. ISSN 2396-832X. doi: 10.1186/s41231-019-0050-7. URL http://dx.doi.org/10.1186/s41231-019-0050-7.

Tanaz Sharifnia, Andrew L. Hong, Corrie A. Painter, and Jesse S. Boehm. Emerging opportunities for target discovery in rare cancers. *Cell Chemical Biology*, 24(9):1075–1091, September 2017. ISSN 2451-9456. doi: 10.1016/j.chembiol.2017.08.002. URL http://dx.doi.org/10.1016/j.chembiol.2017.08.002.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Eliana B Souto, Joana R Campos, Raquel Da Ana, Carlos Martins-Gomes, Amélia M Silva, Selma B Souto, Massimo Lucarini, Alessandra Durazzo, and Antonello Santini. Ocular cell lines and genotoxicity assessment. *Int. J. Environ. Res. Public Health*, 17(6):2046, March 2020.

Jie Sun, Jie Ding, Han Yue, Binbin Xu, Akrit Sodhi, Kang Xue, Hui Ren, and Jiang Qian. Hypoxia-induced bnip3 facilitates the progression and metastasis of uveal melanoma by driving metabolic reprogramming. *Autophagy*, 21(1):191–209, September 2024. ISSN 1554-8635. doi: 10.1080/15548627.2024.2395142. URL http://dx.doi.org/10.1080/15548627.2024.2395142.

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. 2007.

Therese Sørlie, Charles M. Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B. Eisen, Matt van de Rijn, Stefanie S. Jeffrey, Thor Thorsen, Hanne Quist, John C. Matese, Patrick O. Brown, David Botstein, Per Eystein Lønning, and Anne-Lise Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, September 2001. ISSN 1091-6490. doi: 10.1073/pnas.191367098. URL http://dx.doi.org/10.1073/pnas.191367098.

Mathias Uhlén, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, IngMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-Khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-Henrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, Kalle von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar von Heijne, Jens Nielsen, and Fredrik Pontén. Tissue-based map of the human proteome. *Science*, 347(6220), January 2015. ISSN 1095-9203. doi: 10.1126/science.1260419. URL http://dx.doi.org/10.1126/science.1260419.

Carl Virtanen, Yuichi Ishikawa, Daisuke Honjoh, Mami Kimura, Miyuki Shimane, Tatsu Miyoshi, Hitoshi Nomura, and Michael H. Jones. Integrated classification of lung tumors and cell lines by expression profiling. *Proceedings of the National Academy of Sciences*, 99(19):12357–12362, September 2002. ISSN 1091-6490. doi: 10.1073/pnas.192240599. URL http://dx.doi.org/10.1073/pnas.192240599.

Jinghan Wang, Linfang Li, Keqiang Zhang, Yong Yu, Bin Li, Jiang Li, Zi Yan, Zhenli Hu, Yun Yen, Mengchao Wu, Xiaoqing Jiang, and Qijun Qian. Characterization of two novel cell lines with distinct heterogeneity derived from a single human bile duct carcinoma. *PLoS ONE*, 8(1): e54377, January 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0054377. URL http://dx.doi.org/10.1371/journal.pone.0054377.

Allison C. Warren, Andrew Jones, Tsukasa Shibue, William C. Hahn, Jesse S. Boehm, Francisca Vazquez, Aviad Tsherniak, and James M. McFarland. Global computational alignment of tumor and cell line transcriptional profiles. *Nature Communications*, 12, 2020. URL https://api.semanticscholar.org/CorpusID:214723159.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, September 2013. ISSN 1546-1718. doi: 10.1038/ng.2764. URL http://dx.doi.org/10.1038/ng.2764.

Dennis A Wigle, Igor Jurisica, Niki Radulovich, Melania Pintilie, Janet Rossant, Ni Liu, Chao Lu, James Woodgett, Isolde Seiden, Michael Johnston, et al. Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Research*, 62(11):3005–3008, 2002.

K. Yu, B. Chen, D. Aran, J. Charalel, C. Yau, D. M. Wolf, L. J. van 't Veer, A. J. Butte, T. Goldstein, and M. Sirota. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nature Communications*, 10(1), August 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11415-2. URL http://dx.doi.org/10.1038/s41467-019-11415-2.

Yuqing Zhang, Giovanni Parmigiani, and W Evan Johnson. Combat-seq: batch effect adjustment for rna-seq count data. *NAR Genomics and Bioinformatics*, 2(3), September 2020a. ISSN 2631-9268. doi: 10.1093/nargab/lqaa078. URL http://dx.doi.org/10.1093/nargab/lqaa078.

Ziye Zhang, Li Sun, Zhilin Zheng, and Qingli Li. Disentangling the spatial structure and style in conditional vae. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 1626–1630. IEEE, October 2020b. doi: 10.1109/icip40778.2020.9190908. URL http://dx.doi.org/10.1109/ICIP40778.2020.9190908.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Information Maximizing Variational Autoencoders, May 2018. URL http://arxiv.org/abs/1706.02262. arXiv:1706.02262 [cs].

A Appendix

A.1 Training Hyper-parameters

Table 2 reflects the ranges used for random hyper-parameter searches.

Table 2: Hyperparameters search space for training RNAlign.

Parameter	**Random distribution**
Learning rate	$10^{\text{Uniform}(-5,-2)}$
Number of dense layers to finetune	RandomChoice($[1, 2, 3]$)
Adam Weight decay	$10^{\text{Uniform}(-6,-3)}$
Dropout layer *p*	RandomChoice($[0.2, 0.4, 0.6]$)
Batch size	RandomChoice($[32, 64, 128, 256]$)
Multiplier on purity estimate	Uniform$(0, 100)$
σ prior (reconstruction step)	Uniform$(0.2, 1)$
ϵ prior (reparametrization step)	Uniform$(0.2, 1)$
Purity estimate multiplier	Uniform$(0, 100)$
Beta (KL divergence)	$10^{\text{Uniform}(-1.5,-1)}$
$\lambda_1 \mathcal{L}_{cor}$	$10^{\text{Uniform}(-2,-0.5)}$
$\lambda_2 \mathcal{L}_{grad}$	$10^{\text{Uniform}(-2,-0.5)}$

A.2 Expression data

Expression data for 12,236 tumor samples were downloaded from the UCSC Treehouse Public Data using the Xena browser (`https://xenabrowser.net`), specifically the Tumor Compendium V10 Public PolyA data set (Goldman et al., 2018). Samples are derived from the UCSC Treehouse Childhood Cancer Initiative, the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) program, and The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013). Data for 1,249 CL samples were taken from the DepMap Public 19Q4 file: `CCLE_expression_full.csv`. All expression data were processed using the STAR-RSEM pipeline and are TPM log2-transformed (with a pseudocount of 1 added). We subset expression data to the most variable 3000 genes, using median absolute deviation.

A.3 Calculation of ΔD

The ΔD metric is calculated as follows: For each batch, we compute the pairwise Euclidean distances between all samples of the same type (e.g. x_t–x_t, x_{cl}–x_{cl}). These distances reflect the compactness of the data within each batch. The median of these distances is taken as $D_{intra}(x)$, representing the typical distance between samples of the same type within a batch. We also compute the pairwise Euclidean distances between samples of different types (e.g. x_t–x_{cl}). These distances reflect the separation between batches. The median of these distances is taken as $D_{inter}(x_t, x_{cl})$, representing the typical distance between samples of different types across batches. The use of medians across pairwise Euclidean distances is motivated by the fact that datasets often contain CLs that are mis-annotated, poorly representative of tumors, or exhibit extreme molecular profiles due to long-term culturing artifacts.

For each batch, we calculate the difference between its intra-batch median distance and its inter-batch median distance:

$$\Delta D(x_t, x_{cl}) = D_{intra}(x) - D_{inter}(x_t, x_{cl})$$

To ensure robustness, we compute ΔD for all cancer types in the study and take the median of these values. This aggregation provides a summary measure of batch correction performance across diverse biological contexts, reducing the influence of cancer-specific artifacts.

A.4 Addressing different aspects of batch effect correction

Our choice of metrics each target a different aspect of batch effect removal:

- ΔD (Euclidean distance difference): Focuses on geometric structure in the data. By comparing intra-batch compactness (e.g. tumor-tumor) to inter-batch separation (e.g. tumor-CL), it directly measures whether corrected data preserves biologically meaningful clusters while minimizing batch-driven distances. This ensures sample-type distinctions (CL vs. tumor) are not over-smoothed.
- $\Delta kBET$ (difference between observed and expected values of kBEt): Evaluates local statistical consistency of batch mixing. It tests whether neighborhoods of cells/samples reflect the expected distribution under ideal correction (e.g. no batch dominance in local regions). This guards against "patchy" overcorrection, where global metrics like ΔD might suggest success, but local biases persist.
- Principal Variance Components Analysis (PVCA): Quantifies the proportion of variance explained by batch after correction. Unlike distance-based metrics, PVCA directly identifies residual technical variability, ensuring batch effects are not just visually reduced but statistically insignificant.

Appendix A.10 discusses the prevalence of cancer types for which CLs which are biologically expected to poorly correlate to tumors and show 0% CL-tumor matches (Appendix Figure 4). For this reason, we take the median of each metric for all cancer types in the study. This way aggregation provides a robust summary measure of batch correction performance across diverse biological contexts, reducing the influence of cancer-specific artifacts.

RNAlign's performance ranking best on ΔD and $\Delta kBET$ suggests it is excellent for mitigating local batch effects and ensuring proper integration at the level of pairwise distances and mixing metrics. However, performing 2nd best in PVCA, suggests that it might be somewhat more aggressive in its correction relative to Celligner, potentially dampening some of the true biological variance. This is a common trade-off in batch correction: achieving strong local integration sometimes risks losing some global biological structure.

A.5 Assessing clustering of CLs and tumors by cancer type

Appendix Figure 3 shows the UMAP representations of individual cancer type annotations in the globally aligned data.

A.6 Disease-specific analysis

To measure the extent of CLs clustering to their appropriate tumors by cancer type, we follow the procedure set out by Warren et al. (2020). Briefly, we re-classify each CL by the most frequently occurring cancer type in its 25 tumor neighbors (defined as those with the highest Pearson correlation).

Appendix Figure 4 highlights the generally poor fidelity of CLs as tumor models; the majority of cancer types has less than 50% of CL cancer type annotations matching those of its neighboring tumor samples.

Better performing cancer types with regards to CL-tumor clustering in the aligned space have more robust cell line representation, and are often derived from samples that have consistent genomic drivers and stable transcriptional states (Ghandi et al., 2019; Warren et al., 2020).

We conducted a literature review to look into cancer types with poor CL matches to tumor samples, collating biological evidence for poor CL representation or availability of tumors.

- Esophageal cancer (0% match in aligned) exhibits significant metabolic reprogramming of glucose, amino acid, and lipid metabolism (e.g. upregulation of HK2, PKM2, and glutaminase), that is lost in CLs during in-vitro adaptation. Tumors retain microenvironment-driven metabolic demands like hypoxia-induced glycolysis, while CLs adopt simplified metabolic states optimized for proliferation (Dong et al., 2025). EC includes squamous cell

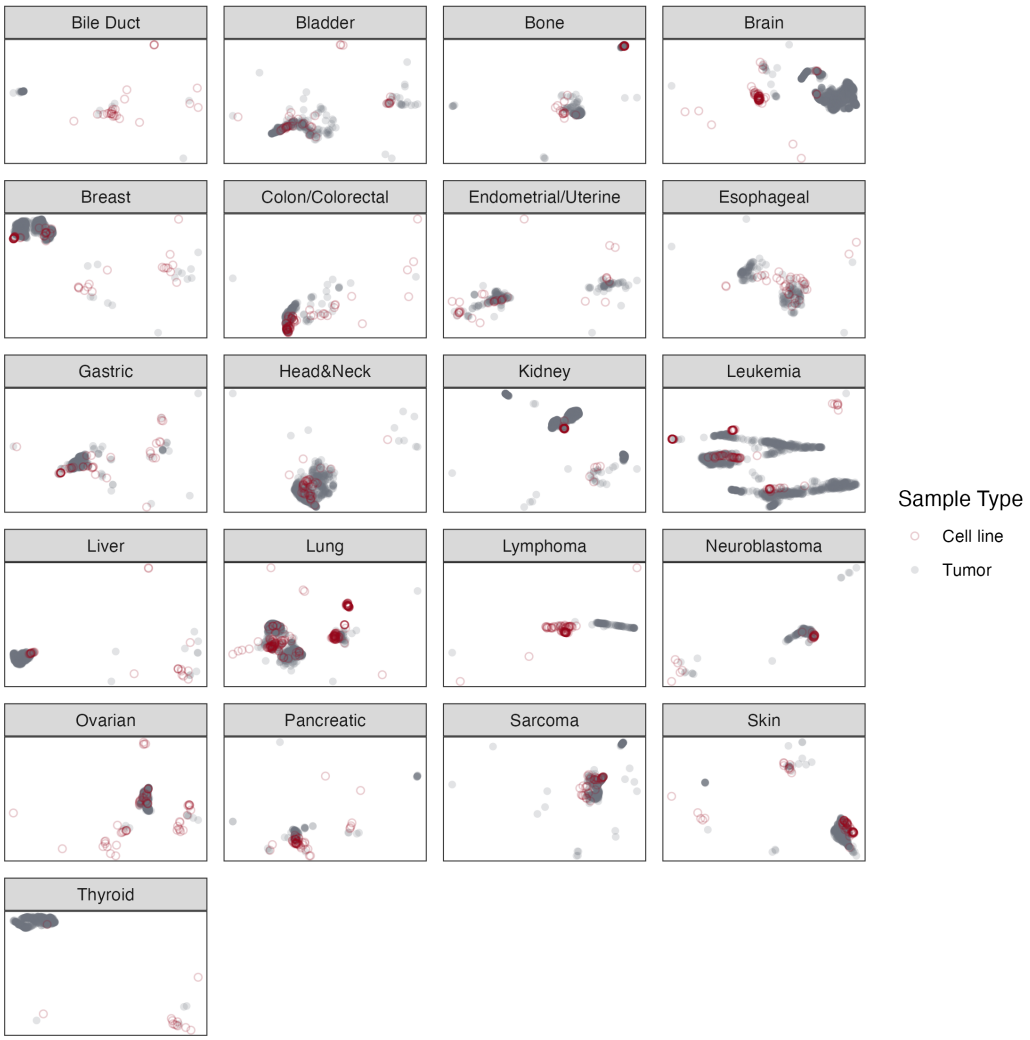

Figure 3: UMAP projections of RNAlign transformed data shows that clustering of CLs and tumors by cancer types improves, but extent of alignment varies across cancer types. Only cancer types with more than 10 samples of each model are displayed.

carcinoma (ESCC) and adenocarcinoma (EAC), which differ in molecular drivers. Some widely-used CLs (e.g. OE33, TE-1) show lower similarity due to subtype misrepresentation (Uhlén et al., 2015).

- Multiple myeloma (0% CL-tumor match) CLs often lack bone marrow stromal interactions like IL-6 signaling, which are critical for tumor survival and gene expression. Additionally, key genomic features are inconsistently represented in CLs, and long-term cultured lines acquire resistance mechanisms (Sarin et al., 2020).
- Subclonal diversity in Cholangiocarcinoma tumors leads to CLs capturing distinct subpopulations For example, EH-CA1a and EH-CA1b are derived from the same tumor but exhibit divergent EMT, MMP, and chemoradiation resistance profiles (Wang et al., 2013). This results in poor fidelity of CLs to in-vivo samples (0% CL-tumor match).
- Eye cancer (0% CL-tumor match) is a rare cancer type with limited availability of well-characterized CLs; the few available are susceptible to genomic drift over long periods of culture (Jager et al., 2016; Souto et al., 2020). Unique micro-environmental factors such as hypoxia, which has been shown to drive metabolic reprogramming in-vivo, are also difficult to recreate in-vitro (Sun et al., 2024).

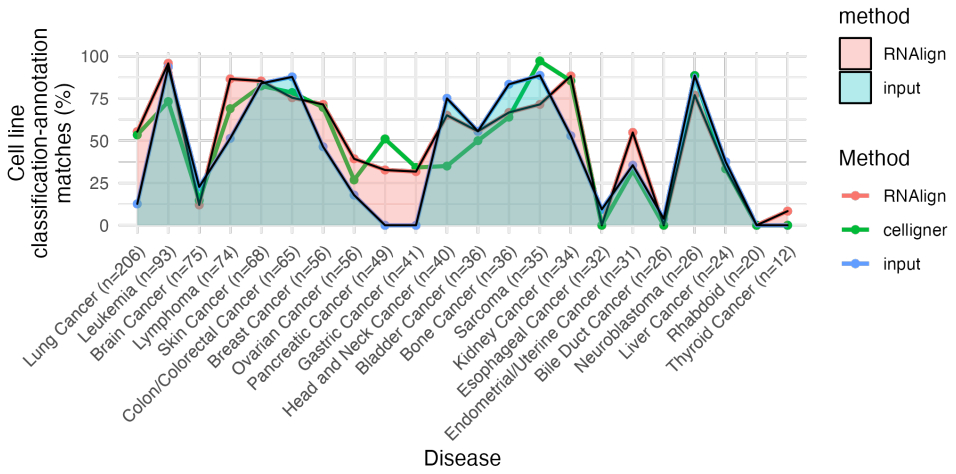

Figure 4: Percentage of cell lines clustering to the equivalent tumors broken down by cancer type. Three datasets are shown, input unaligned data (blue), RNAlign-transformed data (red), and Celligner-transformed (Warren et al., 2020) (green) data. Input- and RNAlign-transformed data are shaded to highlight changes in cancer type-specific classification performance.

The above information highlights that the differences between tumors and CLs is sometimes unique to certain cancer types. Though we show that RNAlign successfully models and removes variation between the model types, it may fail to fully capture disease-specific patterns of variation that can be subtle in pan-cancer analyses. These subtle but crucial signals could be missed due to various reasons, such as insufficient representation across subtypes of the cancer type compared to others, or distinctive characteristics of the tumor micro-environment.

A.7 Unsupervised subtype clustering performance

Table 3 summarizes batch correction scores for the unsupervised/unseen cancer subtypes using different batch correction methods for breast cancer (for annotations basal, HER2-enriched, luminal) and lung cancer (for annotations SCLC, LUAD, LUSC, LCC, Other).

Method	ΔD	PVCA	ΔkBET
	Breast Cancer		
Input	9.87	0.22	0.01
Linear	8.72	0.22	0.02
Celligner	10.67	0.27	0.04
RNAlign	**0.00**	0.03	0.01
	Lung Cancer		
Input	11.68	0.42	0.02
Linear	9.28	0.34	0.03
Celligner	11.49	0.37	0.06
RNAlign	**0.00**	0.14	0.03

Table 3: Comparison of ΔD, PVCA, and ΔkBET scores using different transformation methods for BRCA and NSCLC subtypes.

RNAlign demonstrates superior performance in unsupervised subtype batch correction (Table 3). Its near-zero ΔD values in both breast and lung cancer indicate that the method effectively aligns intra-subtype samples. This is further supported by the low PVCA values, which show a significant reduction in variance attributable to batch effects. Additionally, while RNAlign exhibits excellent local mixing in the breast cancer data as indicated by the low kBET, there is a slight trade-off in local

structure preservation in the lung cancer data. Overall, RNAlign achieves a strong balance between global alignment and variance reduction, with only a minor compromise in local neighborhood mixing in some cases.

A.8 ALTERNATIVE CONDITIONAL GENERATION COMPARISONS

The conditional generation strategy used in this study was to transform the input class labels (model=CL, purity=1), while cancer type labels were kept consistent to the sample annotation. Categorical labels model and cancer type are one-hot encoded. Alternative conditional generation strategies were tested, and the ΔDs of the transformed matrices were measured to assay the additive effect of each label change. Table 4 highlights the conditional generation operations possible through the CVAE framework and their performance using ΔD. All transformation operations result in improved clustering relative to the input data, while both model and purity transformations result in significantly better performance.

Class label transformation	ΔD
Input data	22.56
Model & Tumor (Model = 'CL', Purity = 1)	**4.75**
Purity only (Purity = 1)	9.43
Model only (Model = 'CL')	17.35

Table 4: ΔD for input data, RNAlign transformed data, and similar methods. that test batch effect removal performance between tumors and CLs. The scores are calculated as the median across cancer types.

A.9 ABLATION ANALYSIS

Table 5 summarizes the ΔDs of all methods analyzed in the study, along with equivalent models that each have one aspect of the model ablated – one of L_{grad}, L_{cor} , or purity labels is omitted from each model. All other hyper-parameters are kept the same, and both model & purity labels are transformed for all ablation models. All of the novel features of RNAlign improves alignment of CL and tumor data relative to input data with respect to ΔD. However, any ablation of the model results in significantly worse performance compared to the full model, with the omission of purity labels in the input classes leading to the worst performance.

Method	ΔD
Input data	22.56
RNAlign	**4.75**
RNAlign (*no* L_{grad})	6.12
RNAlign (*no* L_{cor})	10.70
RNAlign (*no purity labels*)	17.34

Table 5: ΔD for input data, RNAlign transformed data, and similar methods. that test batch effect removal performance between tumors and CLs. The scores are calculated as the median across cancer types.

A.10 RNALIGN POTENTIAL DOWNSTREAM APPLICATIONS

A straightforward approach is to use similarity based methods in the transformed space. By identifying the nearest cell line neighbors to a tumor sample, drug response can be inferred based on the behavior of those neighbors with known sensitivity profiles. Another approach could be to build regression or classification models using cell line drug response data on the aligned feature space, which can be applied to patient tumors to estimate treatment efficacy.

Transfer learning or domain adaptation techniques can further refine the predictive performance by adapting insights gleaned from cell line experiments to the nuances of patient data. One transfer

learning strategy would be to pre-train a deep learning model on just the cell line data – using comprehensive drug response labels with the transformed data – to learn robust feature representations. Subsequently, the model could then be fine-tuned using tumor samples, incorporating domain adaptation techniques such as the use of maximum mean discrepancy (Zhao et al., 2018) on the outputs of the latent space to efficiently minimize distribution discrepancies between cell line and tumor features and enhance the model's ability to accurately predict patient drug responses.

A.11 Choice of RNAlign architecture

The use of probabilistic framework over a GANs or GRNs allows uncertainty quantification in the model, which is valuable in biomedical applications where the confidence of the model predictions needs to be taken into account. VAEs also naturally enforce a latent space that is not possible in GANs or GRNs; furthermore, the CVAE architecture has been shown to effectively disentangle factors from this latent space. This regularized latent space then enables identification of shared biological patterns between tumors and CLs, smooth interpolation between latent representations of input samples, and generation of counterfactual or synthetic data based on conditions of interest.

A.12 Future work

Further analyses on RNAlign's performance is required at the local level (i.e intra-cancer type), focusing on two key areas to further validate and refine our transformation method. First, we will evaluate the fidelity of salient biomarker reconstruction in transformed samples to determine whether the relative expression of key biomarkers is preserved across cancer types and subtypes. For example, assessing if the expression patterns of PAM50 gene markers in BRCA samples remain intact after transformation. Secondly, beyond corroboration by literature, a deeper look into the poorly aligned cancer types is also needed, including differential expression and pathway enrichment to identify pathways that contribute to these misalignments.