Semi-supervised Omics Factor Analysis (SOFA) disentangles known sources of variation from latent factors in multi-omics data

Tümay Capraz[1,2], Harald Vöhringer[2-5], Wolfgang Huber[2-3]

[1]Collaboration for joint PhD degree between EMBL and Heidelberg University, Faculty of Biosciences, Heidelberg, Germany
[2]Genome Biology Unit, EMBL, Heidelberg, Germany
[3]Molecular Medicine Partnership Unit (MMPU), Heidelberg, Germany
[4]Department of Medicine V, Hematology, Oncology and Rheumatology, University Hospital Heidelberg, Heidelberg, Germany
[5]Department of Hematology and Oncology, University Hospital Düsseldorf, Düsseldorf, Germany

Abstract

Group Factor Analysis is a family of methods for representing patterns of correlation between features in tabular data[1]. Argelaguet et al. identify latent factors within and across modalities[2]. Often, some factors align with known covariates, and currently, such alignment is done post hoc. We present Semi-supervised Omics Factor Analysis (SOFA), a method that incorporates known sources of variation into the model and focuses the latent factor discovery on novel sources of variation. We apply it to a pan-gynecologic multi-omics data set from The Cancer Genome Atlas (TCGA), where we guide the model with cancer type labels and discover an independent factor representing an immune infiltration vs proliferation transition axis. The inferred factor is predictive of treatment outcomes. We further use SOFA to identify microglial subpopulations during adolescence associated with cell migration and inflammatory response in a single-cell multi-omics data set (RNA- and ATAC-seq) from the human cerebral cortex. SOFA simplifies the discovery of novel patterns and structures in multi-omics data.

Correspondence

Tümay Capraz
EMBL Heidelberg
Meyhofer Str. 1
69117 Heidelberg
tuemay.capraz@embl.de

Introduction

Efforts at understanding biological systems benefit from integrating data from multiple measurement technologies at “omics” scale. Multi-omics studies have revolutionized our ability to map different layers of regulation and cellular identity in multicellular tissues[3–6].

In the analysis of multi-modal datasets, factor models are used to capture the relationships between observed features and a smaller set of latent factors, which explain the patterns of correlation across thousands of samples. Approaches applied to multi-omics data include canonical correlation analysis (CCA)[7], group factor analysis (GFA)[1], multi omics factor analysis (MOFA)[2], iCLuster[8], and MuVI[9]. These methods are unsupervised, in the sense that they do not utilize categorical or continuous sample-level label information (hereafter guiding variables) to inform the factorization. While latent factors can be aligned with guiding variables post-hoc, joint modeling of multi-omic and guiding variables would promote both consistency and interpretability. This is because, by assigning specific factors to guiding variables *a priori*, we can capture sources of variation related to the variable of interest, thus freeing the remaining factors to effectively represent yet unexplained variability. Consequently, the analyst can focus on unguided factors, facilitating the discovery of novel drivers of variation.

The joint modeling task has been addressed in the unimodal setting. Principal component regression (PCR) uses PCA factors as covariates in a regression model to predict guiding variables[10]. However, PCR does not guarantee that the most relevant features for the prediction task are captured, since the PCA only considers relations between the predictors, but not the guiding variable. Partial least squares (PLS) simultaneously decomposes a high-dimensional set of predictors and guiding variables, aiming to find factors that capture the maximum covariance between the two[11,12]. SPEAR[13] and DIABLO[14] enable the incorporation of guiding variables of interest in the multimodal setting, but like PLS, both methods regress out guiding variables across all factors instead of allocating specific ones, limiting the ability to study factors that capture unexplained sources of variation.

We present Semi-supervised Omics Factor Analysis (SOFA), a probabilistic factor model, that allows analysts to jointly model multi-modal omics data and sample-level information. SOFA performs a low-rank decomposition of multi-omics data, partitioning the latent variables into guided factors associated with guiding variables and unguided factors that remain free from such associations. The guided factors capture axes of variation related to the guide variable, while the unguided factors are free to explain remaining variance in the data. In this way, a priori known biological and technical drivers of variation can be regressed out from the latent space, simplifying the interpretation of the inferred factors. This enables a more streamlined analysis, where the analyst can focus on the unguided factors to identify novel mechanisms and processes driving the variation in their multi-omics data sets. On the other hand, guided factors identify features that are associated with the guiding variables.

We show that SOFA identifies factors that are predictive of survival in the pan-gynecologic cohort of The Cancer Genome Atlas (TCGA)[15]. Applying SOFA to cancer cell lines of the cancer dependency map[16,17] (DepMap) reveals factors associated to processes influencing

drug response and CRISPR sensitivity. Finally, we highlight how SOFA can be applied on single-cell multi-omics data to disentangle variation in microglial cells[18]. We provide the SOFA method as a Python package along with comprehensive tutorials.

Results

The SOFA Model

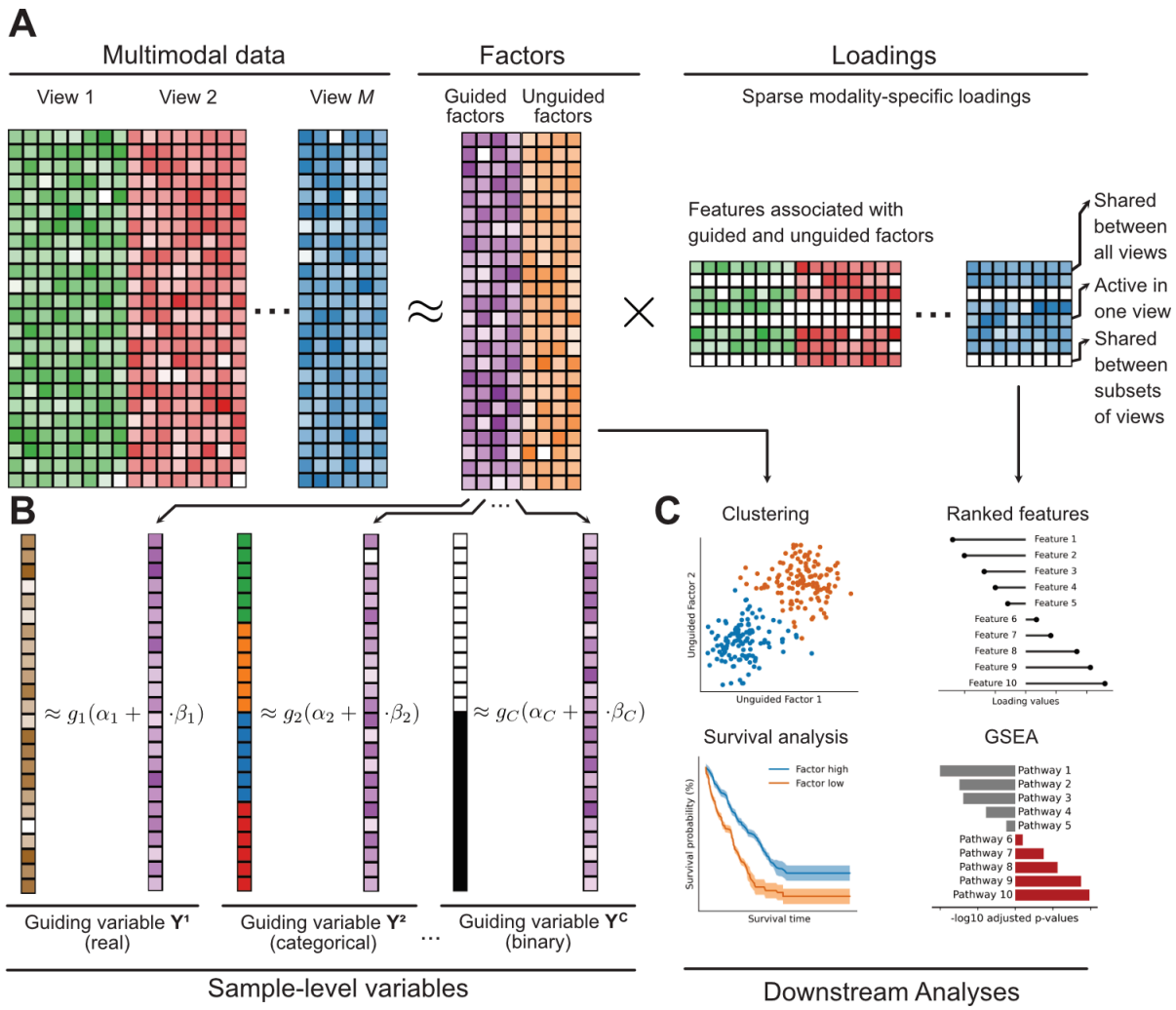

Figure 1: **A**: The SOFA model. Input data modalities X are decomposed into shared latent factors Z and modality specific feature loadings W. A hierarchical horseshoe prior shrinks the loading weights on the view, factor and feature level. This leads to factors shared between all views, factors active in one view and factors shared between subsets of views. **B:** Guided factors are annotated with the guide variables via linear regression. **C:** Unguided factors capture potentially novel sources of variation. We can inspect the loadings to identify which features or gene sets drive the factors and we can inspect the factors and perform analyses on the sample level such as clustering or survival analysis.

We present SOFA, a statistical method that builds on MOFA[2] for the semi-supervised integration of multi-modal omic data with guiding variables. Given a set of real-valued data

matrices containing multi-omic measurements from overlapping samples (also called views), along with sample-level guiding variables that capture additional properties such as batches or mutational profiles, SOFA extracts an interpretable lower-dimensional data representation, consisting of a shared factor matrix and modality-specific loading matrices (**Fig. 1A**). SOFA explicitly assigns a subset of factors to explain both the multi-omics data and the guiding variables (guided factors) (**Fig. 1B**), while preserving another subset of factors exclusively for explaining the multi-omics data (unguided factors). Importantly, this feature allows the analyst to discern variation that is driven by known sources from novel, unexplained sources of variability (**Methods**, **Fig. 1C**).

The interpretation of SOFA's unguided factors is analogous to that of PCA or MOFA, as these factors capture the major sources of variation in the data after accounting for the variability explained by the guided factors, allowing for the stratification of samples into distinct subgroups. SOFA's loading weights link each factor to its most relevant molecular features. Components associated with guided factors reveal features connected to the respective guiding variables, while loading weights associated with unguided factors highlight features linked to novel and unexplained sources of variation. Notably, SOFA employs a hierarchical horseshoe prior[19] on the loading weights, which allows it to distinguish between variation that is shared across modalities and variation that is present in a single modality.

We implemented SOFA as a hierarchical Bayesian factor model, ensuring scalability to large multi omic data sets by employing stochastic variational inference (**Methods**). We tested SOFA on a range of simulations to assess its performance and robustness (**Methods**, **Fig. S1**) and provide SOFA as a Python package via Pypi. In summary, SOFA's factorization framework enables to disentangle known and unknown sources of variation in high-dimensional multi omic datasets.

SOFA identifies cancer type independent immune infiltration vs proliferation axis in the TCGA pan-gynecologic data set

We applied SOFA to a pan-gynecologic cancer multi-omics data set of the cancer genome atlas (TCGA) project, which profiled the transcriptome (mRNA), proteome, methylome, miRNA and mutations of 2599 samples from five different cancers[15]. Additionally, the study includes metadata, the clinical endpoints progression free survival (PFI), and overall survival (OS). We used SOFA to infer 12 latent factors that could be potentially predictive of survival, while guiding a subset of factors with the cancer type labels.

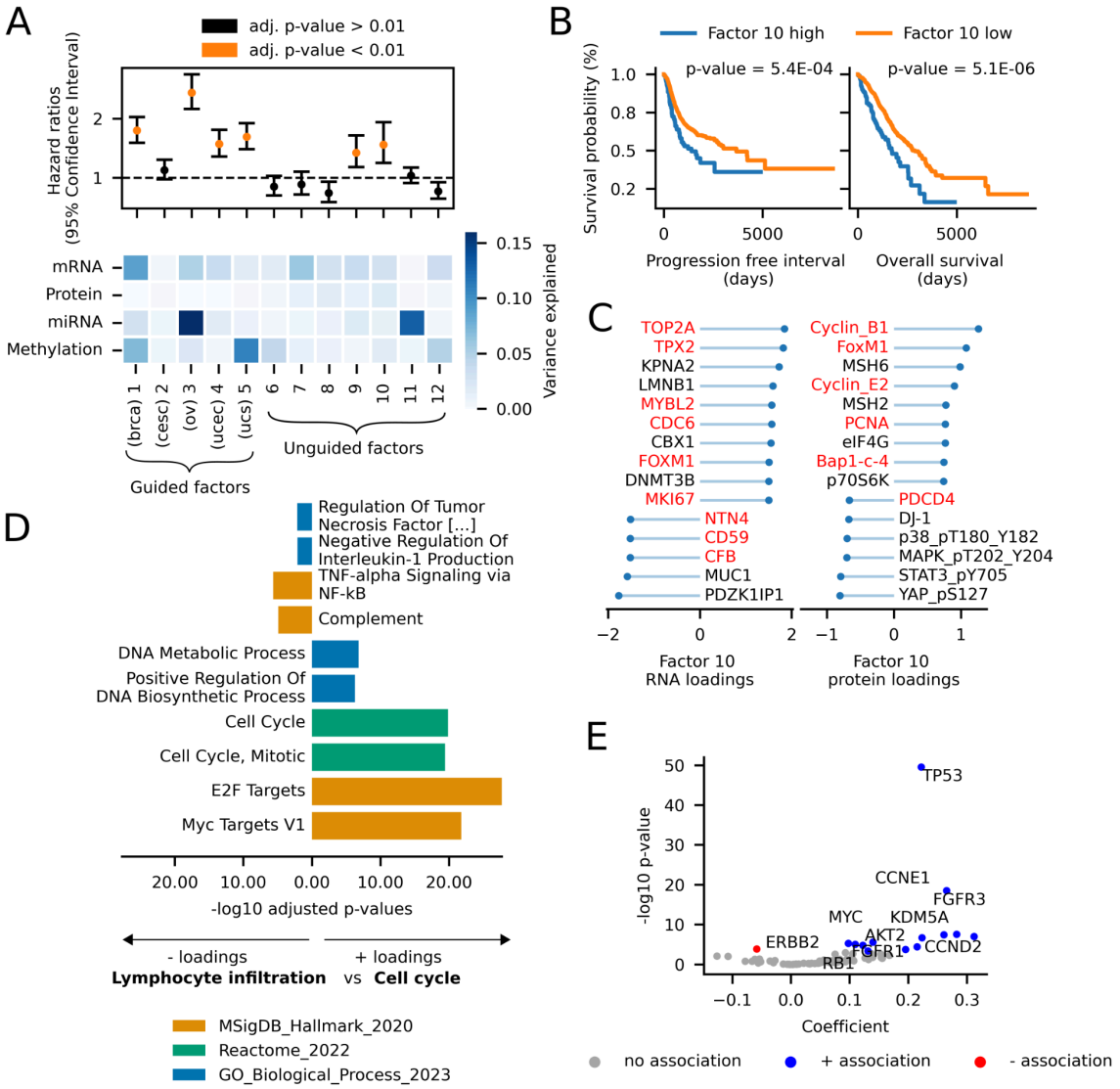

Figure 2: A: Top: Hazard ratios from univariate Cox regression for each factor on overall survival. Hazard ratios > 1 indicate that positive factor values are associated with higher risk, while hazard ratios < 1 indicate that positive factor values are associated with lower risk. Hazard ratios are colored if adjusted p-values are lower than 0.01. P-values were adjusted for multiple testing using the Benjamini-Hochberg (BH) method. Bottom: Variance explained by each factor for each modality. **B**: Kaplan-Meier curves showing overall survival (left) and progression free survival (right) for samples with factor values < 50th and > 95th percentiles of Factor 10. **C**: Features with the highest absolute loading weights for factor 10 in transcriptomics (left) and protein loadings (right). **D**: Gene set overrepresentation analysis for the transcriptomics features with the highest absolute loadings of factor 10. The -log10 adjusted p-values are multiplied by the sign of the loadings to reflect the direction of the association with the factor. **E**: Volcano plot of associations between Factor 10 and mutations, tested using a two sample t-test. BH-adjusted p-values < 0.05 are highlighted in color. Genes colored in blue are significantly associated with higher and genes colored in red with lower factor values. Cancer types shown: brca: breast invasive carcinoma; cesc: Cervical squamous cell carcinoma and endocervical adenocarcinoma; ov: Ovarian serous cystadenocarcinoma; ucec: Uterine Corpus Endometrial Carcinoma; ucs: Uterine Carcinosarcoma

We hypothesized that guiding the SOFA factorization with cancer type information would capture the cancer-specific variation, thus facilitating the discovery of cancer subtype independent drivers of variation. To this end, we performed a SOFA rank 12 decomposition, using five guided factors with binary labels for the cancer type and seven unguided factors (Fig. 2A). To evaluate the importance of each factor, we calculated the fraction of variance explained by each factor across the input modalities (Fig. 2A). Most factors (11/12) explained variance across multiple modalities, indicating that they captured correlated patterns across the different multi-omics measurements. Factor 3 (Ovarian serous cystadenocarcinoma) explained the highest fraction of variance in the miRNA data. miRNA expression was previously reported to play an important role in ovarian cancer and was suggested to serve as a biomarker[20,21]. Notably, the guided factors successfully separated samples based on their corresponding cancer type labels (**Fig. S2F**) and captured biologically relevant loading features. Factor 1 representative for breast cancers showed breast cancer associated proteins AR[22], PR[23] and ER-alpha[24] among the largest proteomics loading weights (**Fig. S2D**). Factor 3 representative for ovarian cancers had *BCAM*[25], *MALAT1*[26] *and WT1*[27] among the highest transcriptomics loadings weights (**Fig. S2E**).

Next, we confirmed that the unguided factors did not capture variance associated with the five cancer types. To this end, we used the *scib* package[28] to perform clustering on the latent factors and calculate the Adjusted Rand Index (ARI) between the cluster labels and cancer type labels for 10 independent initializations. The ARI measures how well the clusterings align with the cancer type labels; a higher ARI corresponds to greater overlap. We compared a guided SOFA model (all 12 factors), only the 7 unguided factors of the guided SOFA model, an unguided (null) SOFA model of rank 12 and an unguided (null) SOFA model of rank 7. Indeed, the 7 unguided factors of the guided model had the lowest ARI, suggesting that the variance associated with the five cancer types was successfully removed by guidance (**Fig. S2C**). This was further corroborated when looking at a t-stochastic neighbor embedding (tSNE)[29] based on the unguided factors (6-12) (**Fig. S2A**).

We then asked whether the SOFA factors could serve as predictors of survival by performing univariate cox regression for each factor on overall survival (OS, **Fig 2A**). This analysis revealed that the cancer-type associated factors were negatively associated with OS, indicating that higher values of these factors corresponded to a lower survival rate. Notably, Factor 1, linked to breast cancer, and Factor 3, associated with ovarian serous cystadenocarcinoma, were significantly linked to poor survival (HR = 1.8, 95% CI: 1.6-2.03 and HR=2.44 95% CI: 2.17-2.75). Conversely, only two out of seven unguided factors showed significant associations with OS, suggesting that most of these factors do not strongly influence survival outcomes. Among the unguided factors predictive of poor survival, we further investigated Factor 10 (HR=1.56 95% CI: 1.25-1.94, **Fig. 2A**). Stratifying samples by Factor 10 values, followed by non-parametric survival curve estimation, revealed substantial differences in median survival times between groups (OS: Median survival time 1701 vs. 2888 days, PFI: 1396 vs 3669 days, **Fig. 2B**).

To understand the biological relevance of Factor 10, we inspected the top loading weights. Negative loadings featured genes like *NTN4*, *CD59* and *CFB*, and proteins such as PDCD4, related to complement system activation, indicating lymphocyte infiltration[30–33](**Fig. 2C**). In contrast, positive loadings included proliferation markers such as *TOP2A* and *TPX2*, as well as Cyclin B1, FoxM1[34–37], suggesting increased proliferation. This was further corroborated

by performing a gene set overrepresentation analysis on the top transcriptomics loadings. Negative loadings were associated with immune system related gene sets, while those with positive loadings were enriched in cell cycle and proliferation related gene sets (**Fig. 2D**).

We tested whether values of Factor 10 were significantly different in mutated vs unmutated samples, using a two-sample t-test. This revealed that samples with mutations in genes related to proliferation such as the tumor suppressor *TP53*[38] and the cell cycle related gene *CCNE1*[39] had significantly higher values of Factor 10. Furthermore, Factor 10 was not exclusively active in one of the cancer types (**Fig. S2B**), indicating that Factor 10 explained variation in the multi-omics data that is independent of the cancer type classification.

Taken together, we found that the guided SOFA factorization effectively delineated cancer type-specific variation from broader biological processes, such cell proliferation or immune cell infiltration, thereby facilitating their identification and subsequent downstream analyses.

SOFA identifies EGFR dependency axis in pan-cancer dependency map, while accounting for known driver mutations and cell growth

Next, we analyzed a pan-cancer cell line data set of the Cancer Dependency Map (DepMap)[17] together with proteomics data from the ProCan-DepMapSanger study[16]. The DepMap project aims to identify cancer vulnerabilities and drug targets across a diverse range of cancer types. The data set includes multi-omics data, encompassing transcriptomics, proteomics, and methylation, as well as drug response profiles for 627 drugs and CRISPR-Cas9 gene essentiality scores for 17485 genes for 949 cancer cell lines across 26 different tissues.

To identify novel biological features significantly associated with CRISPR-Cas9 gene essentiality scores, we performed a rank 20 SOFA decomposition on the DepMap multi-omic data and drug response profiles, while accounting for potential/known drivers of variation such as growth rate, microsatellite instability (MSI) status, *BRAF*, *TP53* and *PIK3CA* mutation and hematopoietic lineage. We chose to exclude the essentiality score data from the SOFA decomposition to be able to independently test factor associations with essentiality scores.

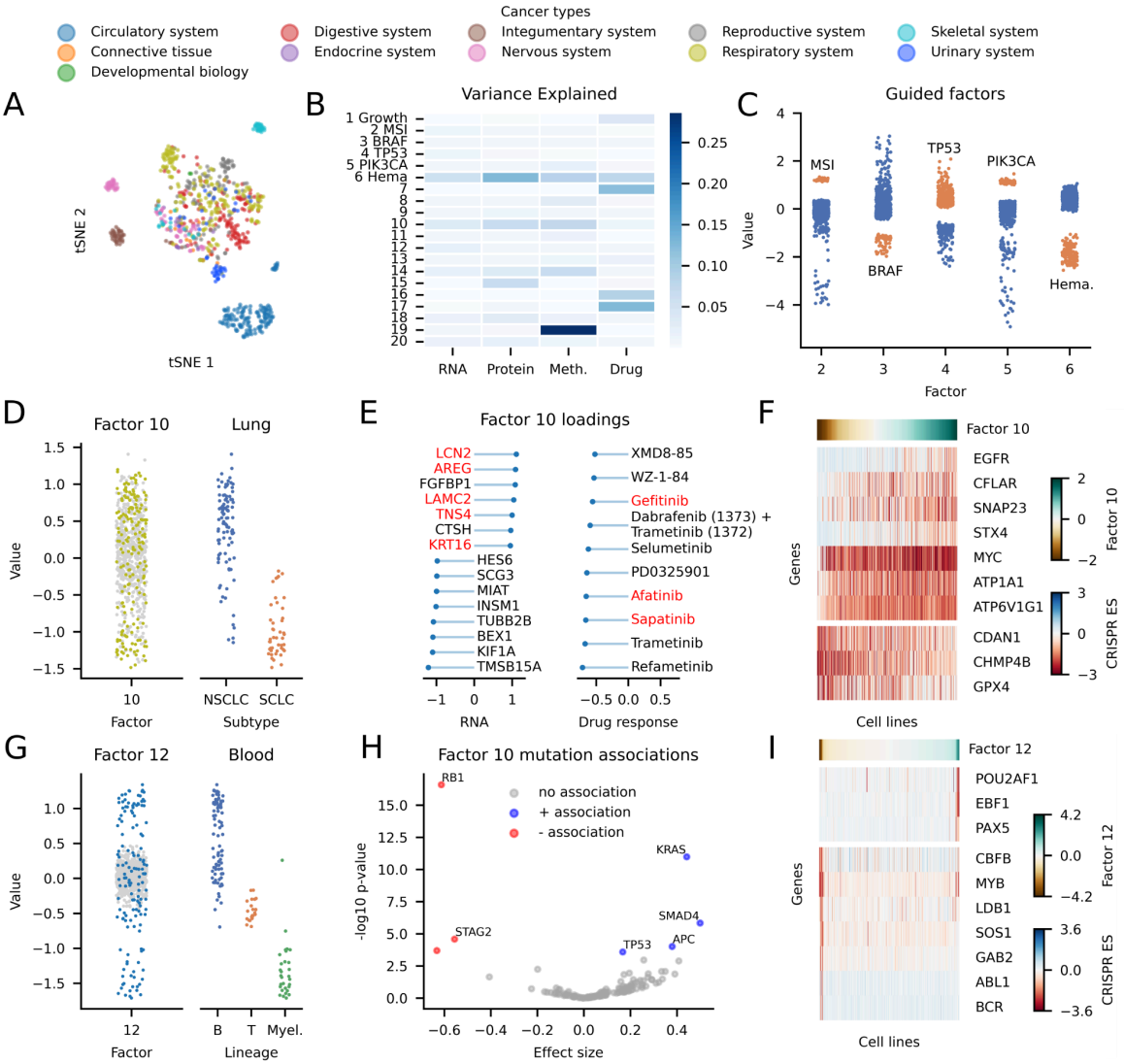

Figure 3: A tSNE of all SOFA factors, colored by cancer types. **B**:Variance explained by each factor for each modality. **C**: Distributions of guided factors colored by their guiding variables. **D**: Distribution of Factor 10 cell lines from the respiratory system are colored (left). Distribution of Factor 10 for cell lines from non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) (right).
E: Highest absolute transcriptomics (left) and drug response (right) loadings of Factor 10. **F**: Heatmap showing CRISPR essentiality scores (ES) of top 10 genes significantly associated with Factor 10, tested using a linear model. The cell lines are ordered based on their values of Factor 10. **G**: Distribution of Factor 12, cell lines from the circulatory system are colored (left). Distribution of Factor 12 for B-cell, T-cell and myeloid cell lines (right). **H**: Volcano plot of associations between Factor 10 and mutations, tested using a two sample t-test. BH-adjusted p-values < 0.05 are highlighted in color. Genes colored in blue are significantly associated with higher and genes colored in red with lower factor values. **I**: Heatmap showing CRISPR essentiality scores (ES) of top 10 genes significantly associated with Factor 12, tested using a linear model. The cell lines are ordered based on their values of Factor 12.

A tSNE of all the inferred factors showed that they captured the differences between cancer types (**Fig. 3A**). Comparing clusterings of the latent factors of the guided model vs a fully unguided model showed similar ARI for the true cancer type labels (**Fig. S3A**). This indicates

that guiding the model does not decrease its ability to capture the molecular differences between cancer types.

We then focused on the guided factors and observed that Factor 6, consistent with previous studies identifying hematopoietic lineage as a major driver of variation in cell lines[16], explained the largest fraction of variance (34%) across all data modalities in this dataset (**Fig. 3B**). Factor 1, guided by growth rate, mostly explained variance in drug response (**Fig. 3B**), which is expected as the natural logarithm of the half maximal inhibitory concentration (lnIC50) drug response is derived from the growth rate of the cell line. **Fig. 3C** shows that the inferred guided factors were predictive of their guiding covariate.

We then subjected the loadings of guided factors to gene set overrepresentation analysis. Expectedly, Factor 3, guided with labels for BRAF mutation, a common mutation in skin cancer[40], was associated with skin cancer related gene sets (**Fig. S3D**). The loadings of Factor 6, guided with hematopoietic lineage, displayed gene sets related to the immune system (**Fig. S3E**).

To test for associations between the guided SOFA factors and essentiality scores, we used factor values as predictors in a linear regression model to explain essentiality scores, and then evaluated the significance of the fitted regression coefficient (**Methods**). Briefly, genes with highly negative essentiality scores are considered essential, as their loss impairs cell viability, while essentiality scores near zero indicate nonessential genes with low impact on viability, and positive scores suggest that loss of the gene may provide survival advantage in certain conditions. Therefore, negative coefficients (effect sizes) indicate that higher factor values correspond to lower essentiality scores, implying a decrease in cell viability. Expectedly, Factor 3 and 4 showed the largest effect sizes for BRAF and TP53, respectively (**Fig. S3F-G**).

We found that among the unguided factors, Factor 10 delineated non-small cell lung cancer (NSCLC) from small cell lung cancer (SCLC) (**Fig. 3D**). Inspecting the loadings of Factor 10, revealed that its highest-ranking gene loadings featured epidermal growth factor receptor (EGFR) regulators such as *LCN2, AREG, LAMC2, TNS4* and *KRT16*[41–45] (**Fig. 3E**). Similarly, the lowest loadings for drug response lnIC50 values indicating sensitivity, were EGFR targeting drugs Gefitinib, Afatinib and Sapatinib (**Fig. 3E**). The EGFR pathway plays an important role in epithelial cancers such as colorectal[46] and lung cancers[47] (in particular in NSCLC)[47] and is associated with tumor progression through prolonged survival, angiogenesis, antiapoptosis, invasion, and metastasis[48]. Indeed, Factor 10 was mostly active in colorectal and lung cancers (**Fig. S3B**). To confirm the association between the EGFR pathway and Factor 10, we tested associations with essentiality scores and found EGFR to be significant (**Fig. 3F**). Additionally, Factor 10 was significantly higher in cell lines with mutations in *TP53, APC, SMAD4* and *KRAS* (**Fig. 3G**). These 4 mutations were previously described to play a role in progression of colorectal cancer[49,50].

Factor 12 picked up the difference between the cell types within the hematopoietic lineage, representing an axis separating B-cells, T-cells and myeloid cells (**Fig. 3H** and **S3C**). The associations of Factor 12 with CRISPR-Cas9 essentiality scores showed that higher factor values (B-cells) were significantly associated with essentiality of B-cell markers *EBF1*, *PAX5*

and *MEF2B*[51,52], while lower factor values (myeloid cells) were significantly associated with essentiality scores of myeloid markers *MYB* and *CBFB*[53,54] (**Fig. 3I**).

Leveraging multi-omics data together with CRISPR-Cas9 essentiality scores, we could show that we were able to identify unguided factors related to EGFR dependency and that explain the difference between hematopoietic cell types. The unguided factors were independent of the drivers of variation we used to guide the model.

SOFA disentangles between and within cell type variation in a single-cell multi-omics data set of the human developing cerebral cortex

We applied SOFA to infer 15 factors from data of a single-cell multi-omics experiment[18], where the authors simultaneously profiled the transcriptome (RNA) and chromatin accessibility (ATAC) of 45549 single cells of the human cerebral cortex at 6 different developmental stages. The authors identified 13 different cell types in the data.

We hypothesized that guiding a subset of SOFA factors with cell type labels could help uncover cell-type independent variation through the unguided factors. To this end, we performed a rank 15 SOFA decomposition on the single-cell multi-omics dataset and guided the first 13 factors with binary cell types labels, while leaving the remaining two factors unguided to capture cell-type independent sources of variation.

Fig. 4A shows the fraction of variance that each factor explained in the transcriptomics and ATAC modalities, with factors guided by microglial and oligodendrocyte labels explaining most variance. The UMAP embedding of all factors in **Fig. 4B** showed that the cells cluster by cell type clearly separating the different cell types.

First, we checked whether the loadings of the guided factors picked up biological signals related to their guiding covariate. Indeed, the guided factors separated cells based on their guiding cell type labels (**Fig. S4A**). The top loadings of the guided factors were cell type specific markers of their respective guiding cell type label. Highest transcriptomics loadings of Factor 1 (astrocytes) included astrocyte specific marker genes such as *SLC1A2,* SLC1A3[55] and *GLUL*[56]. Factor 7 (microglial cells) captured microglia markers *CSF1R*[57], *TYROBP*[58] and *CX3CR1*[59]. Factor 9 (oligodendrocytes) had *MBP*[60], *MOBP*[61] and *PLP1*[62] among the top transcriptomics loadings.(**Fig. S4B-D**).

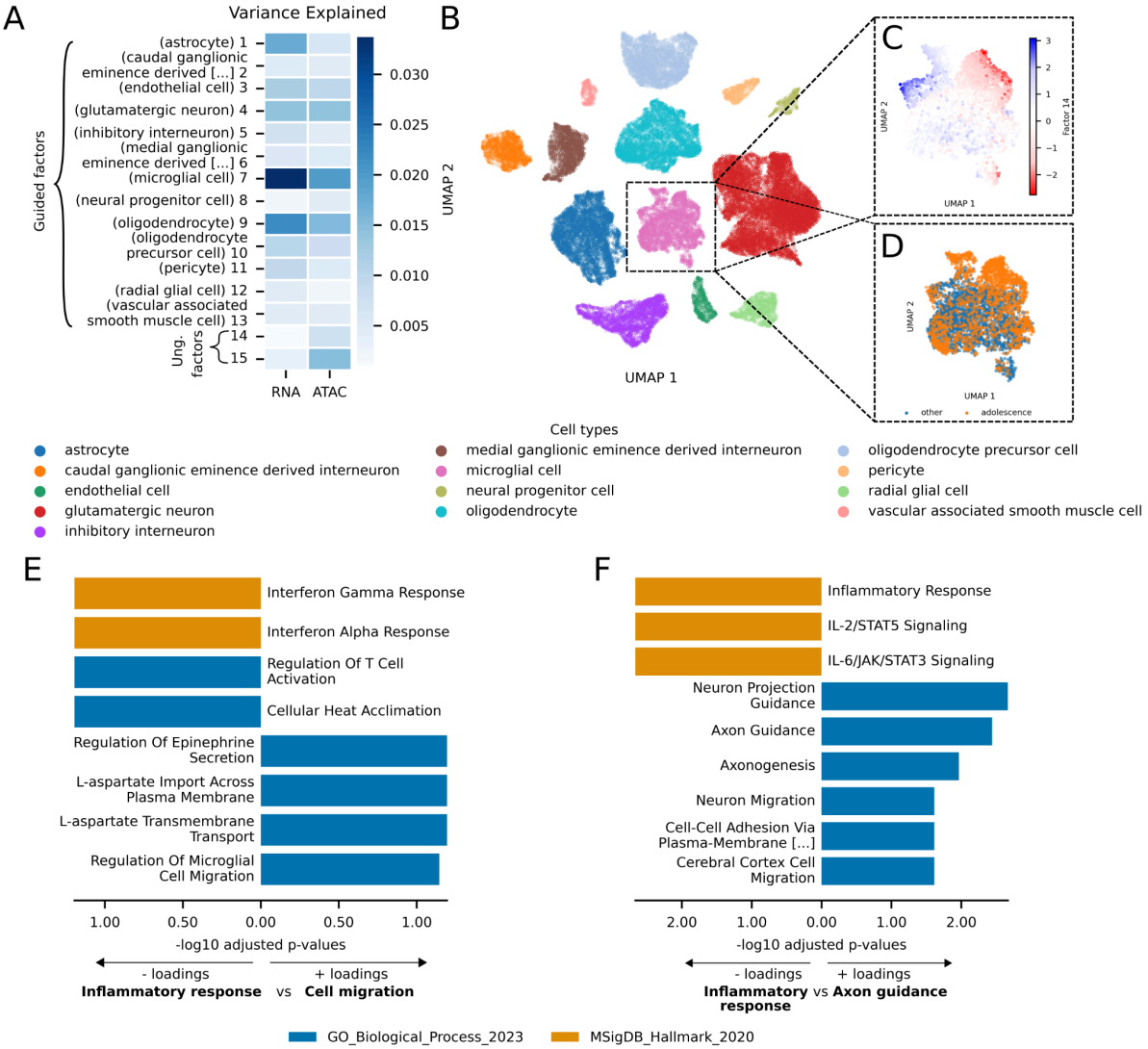

Figure 4: A: Variance explained in RNA and ATAC data modality by each factor. **B**: UMAP based on all factors coloured by cell type. **C**: UMAP of microglial cells coloured by Factor 14. **D**: UMAP of microglial cells colored by adolescence age group. **E**: Gene set overrepresentation analysis for the transcriptomics features with the highest absolute loadings of factor 14. The -log10 adjusted p-values are multiplied by the sign of the loadings to reflect the direction of the association with the factor. **F**: Gene set overrepresentation analysis for the ATAC features with the highest absolute loadings of factor 14. The -log10 adjusted p-values are multiplied by the sign of the loadings to reflect the direction of the association with the factor.

When further inspecting the unguided factors, we found that Factor 14 captured variation within microglial cells and was only active in a subset of microglial cells with relatively high and low factor values (**Fig. 4C**). Gene set overrepresentation analysis on the transcriptomics and ATAC loadings of Factor 14, revealed that the positive loadings are related to microglial migration and transport processes, while the negative loadings play a role in inflammatory response and interleukin signaling (**Fig. 4E** and **4F**). Previous studies using single-cell RNA-seq data from microglia indicated that there is a diverse range of microglial subtypes[63], including populations expressing genes related to inflammatory response[64] and migration[63].

Furthermore, microglia play a critical role in brain development[65], especially during adolescence[66] and Factor 14 was mostly active in cells from the adolescence age group (**Fig. 4C** and **4D**) .

This shows that SOFA is also capable of finding biologically interesting patterns, such as within cell type variation, in single-cell multi-omics data sets.

Discussion

SOFA is a fully interpretable probabilistic framework for the analysis of multi-omics data.
It is based on previous Group Factor Analysis methods such as GFA and MOFA[1,2] and decomposes multi-omics data into shared latent factors that capture shared variance across multiple omics modalities. In contrast to previous latent factor models, it enables the incorporation of a priori known drivers of variation. The drivers of variation can be covariates such as mutations, pathways, cancer or cell types, that most probably do not represent interesting new biological discoveries. While the guided factors explain variation in the molecular omics data that is driven by the guiding variables, the unguided factors can capture the remaining variation, potentially leading to novel biological insights.

Other methods usually involve laborious manual post-hoc annotation of factors or split the data into subsets of cell types, tissues etc., to identify subset specific factors [16,67,68]. SOFA integrates the full data and guiding variables in a single end-to-end analysis, enabling a faster and easier analysis of multi-omics data.

We showed three examples of how SOFA can be used for exploratory data analysis. We identified a biological stratification in the pan-gynecologic cohort of the TCGA[15], that is predictive of survival and independent of cancer type. In a second example, we identified a latent factor related to EGFR cancer dependency enriched in lung and colorectal cancer samples in a pan-cancer cell line analysis of the DepMap[16,17]. Applying SOFA to a single-cell multi-omics data set from cells of the developing human cerebral cortex[18], we show that SOFA can account for inter cell type heterogeneity, while discovering factors explaining intra cell type heterogeneity.

SOFA is a tool for exploratory analysis of multi-omics data sets and can only identify sets of genes or features covarying across multiple modalities and no causal relationships. It is aimed to generate hypotheses that have to be validated using independent test data or in individual experiments. SOFA is a linear method making it fully interpretable, however nonlinear effects within the data can not be captured by SOFA. Furthermore, the current implementation of SOFA is limited to the Gaussian likelihood for the omics views, meaning that e.g. count data from RNA-Seq experiments needs to be appropriately preprocessed and transformed to enable modeling it with a Gaussian likelihood.

Future work could use a combination of linear and non-linear methods (e.g. neural network based) to better capture nonlinear effects that can not be modeled with linear factor analysis alone. We plan to extend SOFA to include the Negative Binomial likelihood to model count data from single-cell transcriptomics experiments and Bernoulli likelihoods to model binary data (e.g. mutation data).

In summary, SOFA is able to discover novel axes of variation while accounting for previously known variation. Its implementation is scalable with runtimes of ~20 minutes on standard GPUs for single cell data sets with ~50k cells, facilitating interactive analyses. It is implemented using the probabilistic programming framework Pyro [69] and is available as a Python package including documentation and analysis examples, making it easy to use.

Methods

The SOFA model

SOFA jointly decomposes a multimodal dataset $\mathbf{X}^m \in \mathbb{R}^{N \times D^m}$ with $m = 1, \ldots, M$ modalities for $i = 1 \ldots N$ samples and $j = 1, \ldots, D^m$ features, and a set of continuous or categorical sample-level covariates (hereafter guiding variables) $\mathbf{Y}^1, \ldots, \mathbf{Y}^C$ where $\mathbf{Y}^c \in \mathbb{R}^N$ or $\mathbf{Y}^c \in \mathbb{N}_0^N$ with $c = 1, \ldots, C$ guiding variables into a lower dimensional shared latent factor matrix and modality specific loading matrices. We model the entries x_{ij}^m of $\mathbf{X}^m$ as normally distributed, given an unobserved mean expression level μ_{ij} and a feature-specific scale parameter σ_j^m, and the entries y_i^c of $\mathbf{Y}^c$ with a mean/location parameter ν_i^c and an optional scale parameter κ^c

$$x_{ij}^m \sim \mathcal{N}(\mu_{ij}^m, \sigma_j^m) \text{ and } y_i^c \sim p^c(\nu_i^c, \kappa^c)$$

where p^c denotes an appropriate likelihood (i.e. Bernoulli, Categorical or Normal) for the c'th guiding variable. In SOFA we express μ_{ij}^m and ν_i^c

$$\mu_{ij}^m = \textstyle\sum_{k=1}^K z_{ik} w_{kj}^m \text{ and } \nu_i^c = g_c(\alpha_c + z_{ic} \cdot \beta_c)$$

where w_{kj}^m represents the entries of the mth loading weight matrix $\mathbf{W}^m \in \mathbb{R}^{K \times D_m}$ consisting of $k = 1, \ldots, K$ components where $C \leq K \leq \min(D_1, \ldots, D_M)$, z_{ik} the elements of a factor weight matrix $\mathbf{Z} \in \mathbb{R}^{N \times K}$ containing the coordinates of each sample with respect to $\mathbf{W}^m$, α_c and β_c an offset and slope parameter and g_c an appropriate inverse link function for the c'th guiding variable.

SOFA is formulated as a Bayesian probabilistic model and uses prior distributions on the model parameters (for a complete derivation of the model see the Supplementary Methods).

Interpretation of the factors

Analogous to the interpretation of factors in PCA, SOFA factors ordinate samples along a zero-centered axis, where samples with opposing signs exhibit contrasting phenotypes along the inferred axis of variation, and the absolute value of the factor indicates the strength of the phenotype. Importantly, SOFA partitions the factors of the low-rank decomposition into

guided and unguided factors: the guided factors are linked to specific guiding variables, while the unguided factors capture global, yet unexplained, sources of variability in the data.

Interpretation of the loading weights

SOFA's loading weights indicate the importance of each feature for its respective factor, thereby enabling the interpretation of SOFA factors. Loading weights close to zero indicate that a feature has little to no importance for the respective factor, while large magnitudes suggest strong relevance. The sign of the loading weight aligns with its corresponding factor, meaning that positive loading weights indicate higher feature levels in samples with positive factor values, and negative loading weights indicate higher feature levels in samples with negative factor values.

Parameter inference

SOFA is formulated as a Bayesian probabilistic model and uses black box variational inference[70] to infer the unobserved parameters. Black box variational inference uses a simpler variational distribution to approximate the intractable posterior. It is a scalable technique that uses stochastic optimization to optimize the evidence lower bound (ELBO) with respect to the variational parameters of the variational distribution. The ELBO is optimized using stochastic variational inference by taking Monte Carlo samples from the variational distribution and computing unbiased but noisy gradients with respect to the variational parameters from the ELBO[70,71]. The ELBO can be used to assess convergence.
For the implementation of SOFA we used the probabilistic programming framework Pyro[69]. If necessary, e.g. for single-cell data, we compute the gradients over subsamples of batches of the observations.

Model selection

The number of factors is an important hyperparameter of SOFA that needs to be determined by running the inference multiple times with a varying number of factors. Factors typically become correlated if the decomposition rank is too high. Additionally, RMSE and variance explained can be used to assess good choices of the number of factors.
Furthermore, guiding variables potentially driving variation in the data, need to be tested in multiple inference runs.

Simulations

To verify that SOFA is able to both infer factors that approximate the multi-omics data and simultaneously guide a subset of the factor, we simulated data from the generative model, inferred a varying number of latent factors and computed the root mean square error between the reconstructed and the simulated data.
Fig. S1A shows the RMSE for the simulated multi-omics data for an unguided and a SOFA model where the first two factors were guided with simulated guide variables. The RMSE of the unguided model was slightly lower than that of the guided model. However, this was only the case for model fits with a lower number of factors than the simulated data. This indicates that if the number of latent factors is at least equal to the number of latent dimensions of the

simulated data, SOFA is able to guide a subset of the factors without losing its ability to fit the multi-omics data.
Furthermore, the RMSE of the guide variables was low for models with the number of factors at least equal to the number of latent dimensions of the data (Fig. S1B). For models with lower number of factors the RMSE had a high variance across model initialization. This indicates that if the dimensionality of the model is too low to approximate the data, the parameter optimization can arrive at local optima that either focus on approximating the multi-omics data or the guide variables.

Data processing

All data processing steps were performed using the *muon* (version 0.1.3)[72] and *anndata* (version 0.8.0)[73] Python packages.

TCGA pan-cancer atlas

All data modalities, survival data and metadata were downloaded using the R package *curatedTCGAdata* (version 1.26.0)[15,74]. All data modalities except the protein array data were log-transformed. All data was centered and scaled to unit variance. Before the analysis, we selected the top 2000 highly variable features of the RNA-seq and methylation data sets.

Pan-cancer DepMap

We analyzed a multi-omic cancer cell line data set with proteomic data from Gonçalves et al.[16] combined with multi-omic and CRISPR essentiality data from Boehm et al.[17] of the same cell lines. Processed and transcript per million (TPM) normalized RNA-seq data, mutation data and integrated CRISPR essentiality scores were obtained from https://cellmodelpassports.sanger.ac.uk/. Processed methylation beta values were obtained from https://www.cancerrxgene.org. Processed and normalized proteomics abundances and drug response lnIC50 data was obtained via figshare from Gonçalves et al.[16]. We log-transformed, centered and scaled the RNA-seq TPM data and the pre-processed methylation beta values and selected the top 2000 highly variable genes prior to the analysis. The proteomics data was imputed with zero, and the top 2000 highly variable proteins were selected. Additionally, we regressed out the sample mean as suggested in Gonçalves et al.[16]. The drug response lnIC50 values were centered and scaled.

Single-cell multi-omics of the human cerebral cortex

The single-cell RNA-seq and ATAC-seq data was obtained from Zhu et al.[18] and downloaded in the h5ad format from https://cellxgene.cziscience.com/. RNA-seq data was normalized to library size and log-transformed. We selected the top 2000 highly variable genes in both modalities. Both datasets were centered and scaled before analysis. Cell type labels were obtained from the h5ad objects.

Downstream analysis

Association testing

Associations between factors and mutations were tested using two sample t-tests. Associations between factors and CRISPR sensitivity scores were tested using a linear model. All p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure[75].

Survival analysis

The survival analysis was performed using the *lifelines* package (version 0.27.7)[76]. Assessment of whether a factor was predictive of overall survival was done by per factor univariate Cox's proportional hazard models. Resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure [75].

Gene set overrepresentation analysis (ORA)

To annotate the factors with biological processes we used the enrichr API [77] provided in the *gseapy* Python package (version 1.0.4) [78]. We used the top 100 features with highest or lowest loadings of a given factor and performed gene set overrepresentation analysis using all genes in the data set as background.

Fraction of explained variance

Assuming the data $\mathbf{X}$ is centered, the fraction of variance that is explained by factor k of modality m was calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \sum_{j=1}^{D^m} (z_{ik} \times w_{k,j}^m)^2}{\sum_{i=1}^{N} \sum_{j=1}^{D^m} x_{ij}^2} \quad .$$

Where the denominator corresponds to the total sum of squares of the centered data $\mathbf{X}$ and the nominator to the residual sum of squares. The fraction can be understood as the variability a given factor k accounts for.

Two dimensional visualizations of the latent space

The t-distributed Stochastic Neighbor Embeddings (tSNE) for 2D visualizations of the latent space were calculated using the *scikit-learn* Python package (version 1.2.1)[79]. Uniform Manifold Approximation and Projection (UMAP)[80] embeddings were computed using the implementation provided in the *scanpy* Python package (version 1.9.3) [81].

Root mean square error for model selection

To assess the quality of the individual model fits we calculated the root mean square error as follows:

$$\mathrm{RMSE} = \frac{\sqrt{\sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{j=1}^{D}{}_m (x_{ij}^m - \hat{x}_{ij}^m)^2}}{\sum_{m=1}^{M} N \times D_m} \quad .$$

Software availability

SOFA is implemented in the Python package *biosofa* and is available on github (https://github.com/tcapraz/SOFA) and PyPI (https://pypi.org/project/biosofa/), along with documentation and example analysis notebooks.

Acknowledgements

We would like to thank Stefan Peidli and Martin Rohbeck for valuable feedback and discussions. We thank EMBL IT services and HPC resources for providing infrastructure to perform computations needed for this work.

Funding

This work was funded by the European Research Council (Synergy Grant DECODE under grant agreement no. 810296).

References

1. Klami, A., Virtanen, S., Leppäaho, E. & Kaski, S. Group factor analysis. *IEEE Trans. Neural Netw. Learn. Syst.* **26**, 2136–2147 (2015).
2. Argelaguet, R. *et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
3. Clark, S. J. *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
4. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
5. Swanson, E. *et al.* Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *Elife* **10**, (2021).
6. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103–1116.e20 (2020).
7. Hotelling, H. RELATIONS BETWEEN TWO SETS OF VARIATES. *Biometrika* **28**, 321–377 (1936).
8. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data

types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–2912 (2009).

9. Qoku, A. & Buettner, F. Encoding Domain Knowledge in Multi-view Latent Variable Models: A Bayesian Approach with Structured Sparsity. in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics* (eds. Ruiz, F., Dy, J. & van de Meent, J.-W.) vol. 206 11545–11562 (PMLR, 25--27 Apr 2023).
10. Jolliffe, I. T. A Note on the Use of Principal Components in Regression. *J. R. Stat. Soc. Ser. C Appl. Stat.* **31**, 300–303 (1982).
11. Abdi, H. Partial Least Squares (PLS) Regression. (2003).
12. Abdi, H. Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdiscip. Rev. Comput. Stat.* **2**, 97–106 (2010).
13. Gygi, J. P. *et al.* A supervised Bayesian factor model for the identification of multi-omics signatures. *Bioinformatics* **40**, (2024).
14. Singh, A. *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055–3062 (2019).
15. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
16. Gonçalves, E. *et al.* Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell* **40**, 835–849.e8 (2022).
17. Boehm, J. S. *et al.* Cancer research needs a better map. *Nature* **589**, 514–516 (2021).
18. Zhu, K. *et al.* Multi-omic profiling of the developing human cerebral cortex at the single-cell level. *Sci Adv* **9**, eadg3754 (2023).
19. Carvalho, C. M., Polson, N. G. & Scott, J. G. Handling Sparsity via the Horseshoe. in *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics* (eds. van Dyk, D. & Welling, M.) vol. 5 73–80 (PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16--18 Apr 2009).
20. Iorio, M. V. *et al.* MicroRNA signatures in human ovarian cancer. *Cancer Res.* **67**, 8699–8707 (2007).

21. Zhao, L., Liang, X., Wang, L. & Zhang, X. The role of miRNA in ovarian cancer: An overview. *Reprod. Sci.* **29**, 2760–2767 (2022).
22. Collins, L. C. *et al.* Androgen receptor expression in breast cancer in relation to molecular phenotype: results from the Nurses' Health Study. *Mod. Pathol.* **24**, 924–931 (2011).
23. Pichon, M. F., Pallud, C., Brunet, M. & Milgrom, E. Relationship of presence of progesterone receptors to prognosis in early breast cancer. *Cancer Res.* **40**, 3357–3360 (1980).
24. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
25. Sivakumar, S. *et al.* Basal cell adhesion molecule promotes metastasis-associated processes in ovarian cancer. *Clin. Transl. Med.* **13**, e1176 (2023).
26. Gordon, M. A., Babbs, B., Cochrane, D. R., Bitler, B. G. & Richer, J. K. The long non-coding RNA MALAT1 promotes ovarian cancer progression by regulating RBFOX2-mediated alternative splicing. *Mol. Carcinog.* **58**, 196–205 (2019).
27. Netinatsunthorn, W., Hanprasertpong, J., Dechsukhum, C., Leetanaporn, R. & Geater, A. WT1 gene expression as a prognostic marker in advanced serous epithelial ovarian carcinoma: an immunohistochemical study. *BMC Cancer* **6**, 90 (2006).
28. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
29. Maaten, L. & Hinton, G. E. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
30. Jiang, Y., Jia, Y. & Zhang, L. Role of programmed cell death 4 in diseases: a double-edged sword. *Cell. Mol. Immunol.* **14**, 884–886 (2017).
31. Yi, L. *et al.* NTN4 as a prognostic marker and a hallmark for immune infiltration in breast cancer. *Sci. Rep.* **12**, 10567 (2022).
32. Sarmoko, Ramadhanti, M. & Zulkepli, N. A. CD59: Biological function and its potential for drug target action. *Gene Reports* **31**, 101772 (2023).

33. Wu, P., Shi, J., Sun, W. & Zhang, H. The prognostic value of plasma complement factor B (CFB) in thyroid carcinoma. *Bioengineered* **12**, 12854–12866 (2021).
34. Earnshaw, W. C., Halligan, B., Cooke, C. A., Heck, M. M. & Liu, L. F. Topoisomerase II is a structural component of mitotic chromosome scaffolds. *J. Cell Biol.* **100**, 1706–1715 (1985).
35. Wittmann, T., Boleti, H., Antony, C., Karsenti, E. & Vernos, I. Localization of the kinesin-like protein Xklp2 to spindle poles requires a leucine zipper, a microtubule-associated protein, and dynein. *J. Cell Biol.* **143**, 673–685 (1998).
36. Yuan, J. *et al.* Cyclin B1 depletion inhibits proliferation and induces apoptosis in human tumor cells. *Oncogene* **23**, 5843–5852 (2004).
37. Fischer, M. & Müller, G. A. Cell cycle transcription control: DREAM/MuvB and RB-E2F complexes. *Crit. Rev. Biochem. Mol. Biol.* **52**, 638–662 (2017).
38. Finlay, C. A., Hinds, P. W. & Levine, A. J. The p53 proto-oncogene can act as a suppressor of transformation. *Cell* **57**, 1083–1093 (1989).
39. Siu, K. T., Rosner, M. R. & Minella, A. C. An integrated view of cyclin E function and regulation. *Cell Cycle* **11**, 57–64 (2012).
40. Maldonado, J. L., Fridlyand, J. & Bastian, B. C. RESPONSE: Re: Determinants of BRAF mutations in primary melanomas. *J. Natl. Cancer Inst.* **97**, 402–402 (2005).
41. Yammine, L., Zablocki, A., Baron, W., Terzi, F. & Gallazzini, M. Lipocalin-2 Regulates Epidermal Growth Factor Receptor Intracellular Trafficking. *Cell Rep.* **29**, 2067–2077.e6 (2019).
42. Stoll, S. W. *et al.* The EGF receptor ligand amphiregulin controls cell division via FoxM1. *Oncogene* **35**, 2075–2086 (2016).
43. Tong, D. *et al.* LAMC2 promotes EGFR cell membrane localization and acts as a novel biomarker for tyrosine kinase inhibitors (TKIs) sensitivity in lung cancer. *Cancer Gene Ther.* **30**, 1498–1512 (2023).
44. Shi, Z.-Z. *et al.* The miR-1224-5p/TNS4/EGFR axis inhibits tumour progression in oesophageal squamous cell carcinoma. *Cell Death Dis.* **11**, 597 (2020).

45. Coulombe, P. A. & Orosco, A. Inhibiting EGFR Signaling Holds Promise for Treating Palmoplantar Keratodermas. *J. Invest. Dermatol.* **143**, 185–188 (2023).
46. Krasinskas, A. M. EGFR Signaling in Colorectal Carcinoma. *Patholog. Res. Int.* **2011**, 932932 (2011).
47. Hsu, P.-C., Jablons, D. M., Yang, C.-T. & You, L. Epidermal Growth Factor Receptor (EGFR) Pathway, Yes-Associated Protein (YAP) and the Regulation of Programmed Death-Ligand 1 (PD-L1) in Non-Small Cell Lung Cancer (NSCLC). *Int. J. Mol. Sci.* **20**, (2019).
48. Spano, J. P. *et al.* Epidermal growth factor receptor signaling in colorectal cancer: preclinical data and therapeutic perspectives. *Ann. Oncol.* **16**, 189–194 (2005).
49. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
50. Malki, A. *et al.* Molecular Mechanisms of Colon Cancer Progression and Metastasis: Recent Insights and Advancements. *Int. J. Mol. Sci.* **22**, (2020).
51. Brescia, P. *et al.* MEF2B Instructs Germinal Center Development and Acts as an Oncogene in B Cell Lymphomagenesis. *Cancer Cell* **34**, 453–465.e9 (2018).
52. Nutt, S. L., Heavey, B., Rolink, A. G. & Busslinger, M. Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. *Nature* **401**, 556–562 (1999).
53. Hoffman, B., Amanullah, A., Shafarenko, M. & Liebermann, D. A. The proto-oncogene c-myc in hematopoietic development and leukemogenesis. *Oncogene* **21**, 3414–3421 (2002).
54. Kundu, M. *et al.* Role of Cbfb in hematopoiesis and perturbations resulting from expression of the leukemogenic fusion gene Cbfb-MYH11. *Blood* **100**, 2449–2456 (2002).
55. DeSilva, T. M., Borenstein, N. S., Volpe, J. J., Kinney, H. C. & Rosenberg, P. A. Expression of EAAT2 in neurons and protoplasmic astrocytes during human cortical development. *J. Comp. Neurol.* **520**, 3912–3932 (2012).
56. Anlauf, E. & Derouiche, A. Glutamine synthetase as an astrocytic marker: its cell type

and vesicle localization. *Front. Endocrinol.* **4**, 144 (2013).

57. Hagan, N. *et al.* CSF1R signaling is a regulator of pathogenesis in progressive MS. *Cell Death Dis.* **11**, 904 (2020).
58. Zhang, B. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707–720 (2013).
59. Wolf, Y., Yona, S., Kim, K.-W. & Jung, S. Microglia, seen from the CX3CR1 angle. *Front. Cell. Neurosci.* **7**, 26 (2013).
60. Ainger, K. *et al.* Transport and localization of exogenous myelin basic protein mRNA microinjected into oligodendrocytes. *J. Cell Biol.* **123**, 431–441 (1993).
61. *Oligodendrocytic Basic Protein: Identification of an Encephalitogenic Epitope and Association with Multiple Sclerosis1*.
62. Kim, D., An, H., Fan, C. & Park, Y. Identifying oligodendrocyte enhancers governing Plp1 expression. *Hum. Mol. Genet.* **30**, 2225–2239 (2021).
63. Masuda, T., Sankowski, R., Staszewski, O. & Prinz, M. Microglia Heterogeneity in the Single-Cell Era. *Cell Rep.* **30**, 1271–1281 (2020).
64. Hammond, T. R. *et al.* Single-Cell RNA Sequencing of Microglia throughout the Mouse Lifespan and in the Injured Brain Reveals Complex Cell-State Changes. *Immunity* **50**, 253–271.e6 (2019).
65. Thion, M. S., Ginhoux, F. & Garel, S. Microglia and early brain development: An intimate journey. *Science* **362**, 185–189 (2018).
66. Schalbetter, S. M. *et al.* Adolescence is a sensitive period for prefrontal microglia to act on cognitive development. *Sci Adv* **8**, eabi6672 (2022).
67. Barkley, D. *et al.* Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat. Genet.* **54**, 1192–1201 (2022).
68. Ramirez Flores, R. O., Lanzer, J. D., Dimitrov, D., Velten, B. & Saez-Rodriguez, J. Multicellular factor analysis of single-cell data for a tissue-centric understanding of disease. *Elife* **12**, (2023).
69. Bingham, E. *et al.* Pyro: Deep universal probabilistic programming.

https://www.jmlr.org/papers/volume20/18-403/18-403.pdf.

70. Ranganath, R., Gerrish, S. & Blei, D. Black Box Variational Inference. in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* (eds. Kaski, S. & Corander, J.) vol. 33 814–822 (PMLR, Reykjavik, Iceland, 22--25 Apr 2014).
71. Hoffman, M., Blei, D. M., Wang, C. & Paisley, J. Stochastic Variational Inference. *arXiv [stat.ML]* 1303–1347 (2012).
72. Bredikhin, D., Kats, I. & Stegle, O. MUON: multimodal omics analysis framework. *Genome Biol.* **23**, 42 (2022).
73. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Alexander Wolf, F. anndata: Annotated data. *bioRxiv* 2021.12.16.473007 (2021) doi:10.1101/2021.12.16.473007.
74. Ramos, M. *et al.* Multiomic Integration of Public Oncology Databases in Bioconductor. *JCO Clin Cancer Inform* **4**, 958–971 (2020).
75. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
76. Davidson-Pilon, C. lifelines: survival analysis in Python. *J. Open Source Softw.* **4**, 1317 (2019).
77. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
78. Fang, Z., Liu, X. & GSEApy, G. P. A comprehensive package for performing gene set enrichment analysis in Python., 2023, 39, btac757. *DOI: https://doi.org/10.1093/bioinformatics/btac757*.
79. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **abs/1201.0490**, (2011).
80. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
81. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

Supplementary Figures

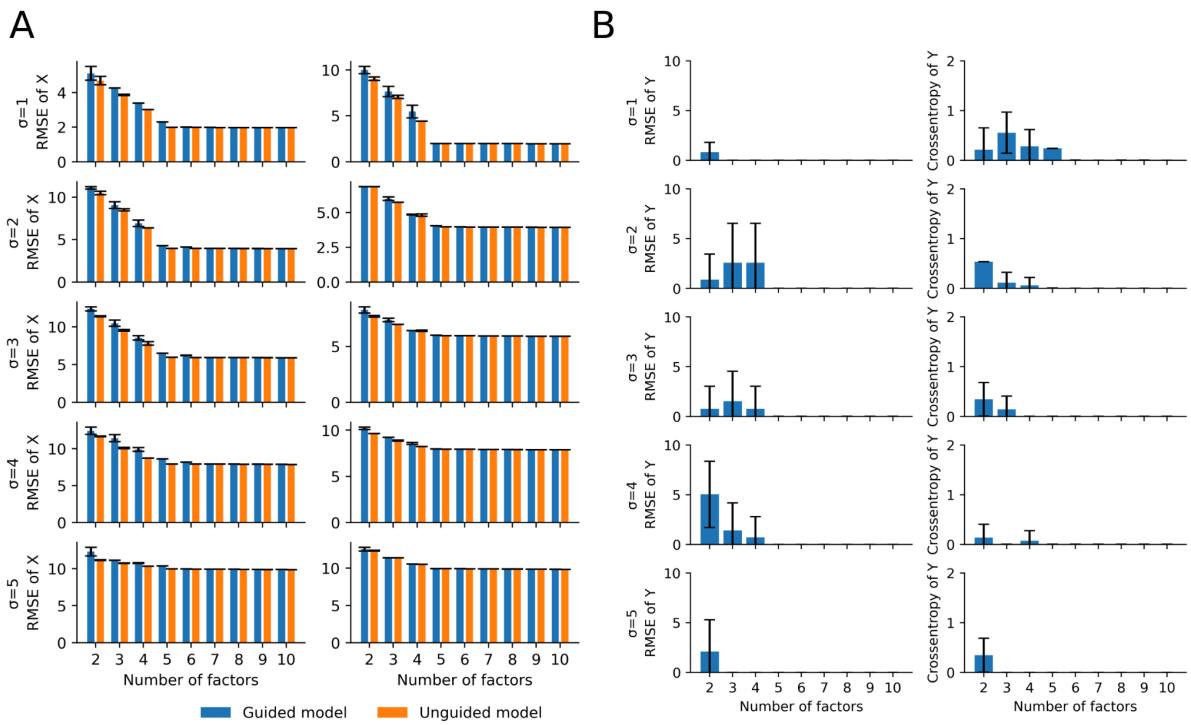

Figure S1: **Simulation Experiments.** In-silico data were simulated according to the generative model of SOFA using a true rank of 5 and two guiding variables for the first (continuous) and second factor (binary). The simulated multi omic data comprised 2 views with each 300 observations and 2000 features. To assess the model fit and selection, we varied the noise variance σ (rows) and performed the SOFA factorization over a range of decomposition ranks with and without the use of guiding variables. **A**: Average root mean square error (RMSE) of true and predicted data averaged over 10 model initializations for the first (left) and second view (right). **B**: Average RMSE (left) and cross entropy (right) of the true and predicted guiding variable averaged over 10 initializations.

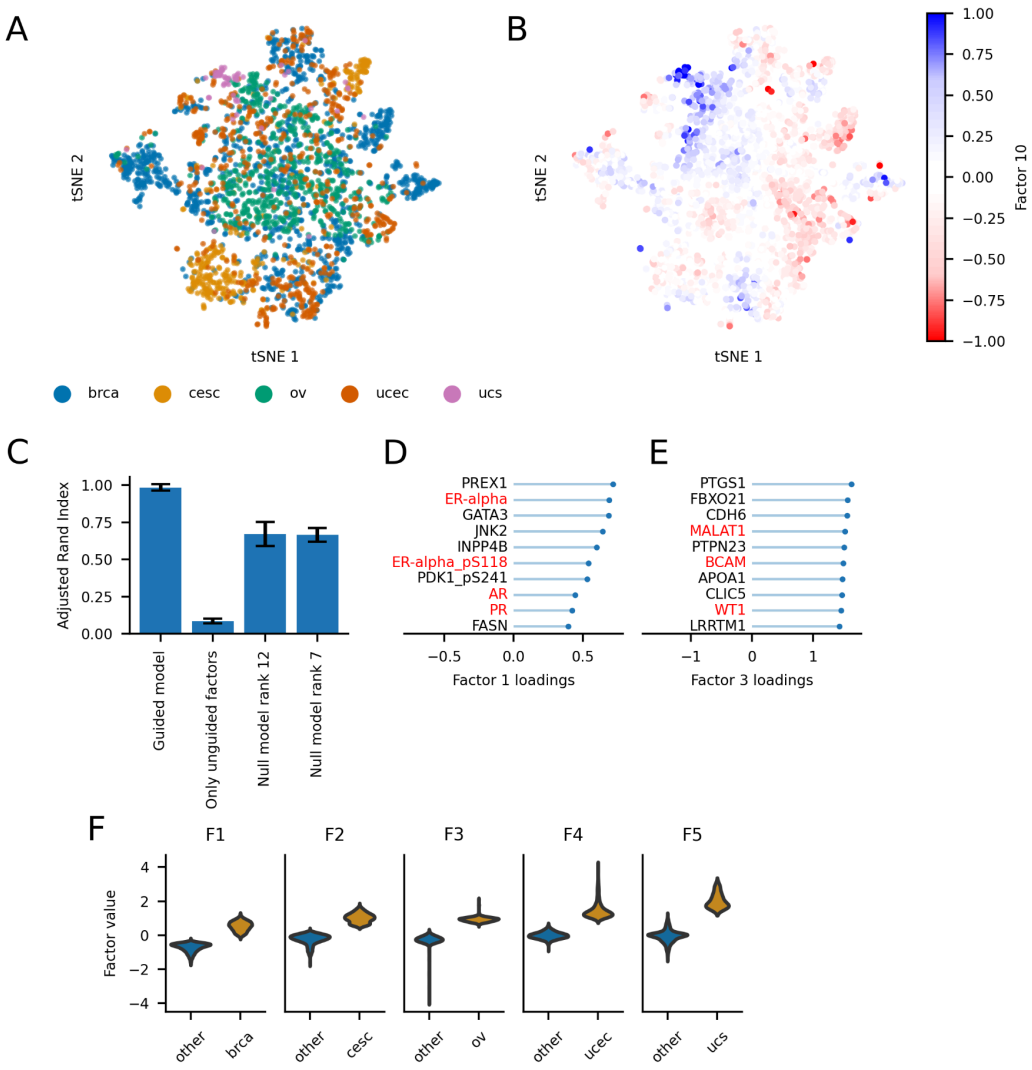

Figure S2: **A**: tSNE of unguided Factors 7-12 colored by cancer type. **B**: tSNE of unguided Factors 7-12 colored by Factor 10. **C**: Adjusted Rand Index (ARI) for clusterings of the guided model, unguided Factors 7-12, unguided null model of rank 12 and unguided null model of rank 7. The clusterings were compared to the true cancer type labels. A higher ARI suggests better overlap of the clustering with the cancer type labels. The ARI was computed for 10 independent model initializations. **D**: Highest protein loadings of Factor 1. **E**: Highest RNA loadings of Factor 3. **F**: **V**iolin plots of each guided factor vs their guiding cancer type label.
brca: breast invasive carcinoma; cesc: Cervical squamous cell carcinoma and endocervical adenocarcinoma; ov: Ovarian serous cystadenocarcinoma; ucec: Uterine Corpus Endometrial Carcinoma; ucs: Uterine Carcinosarcoma

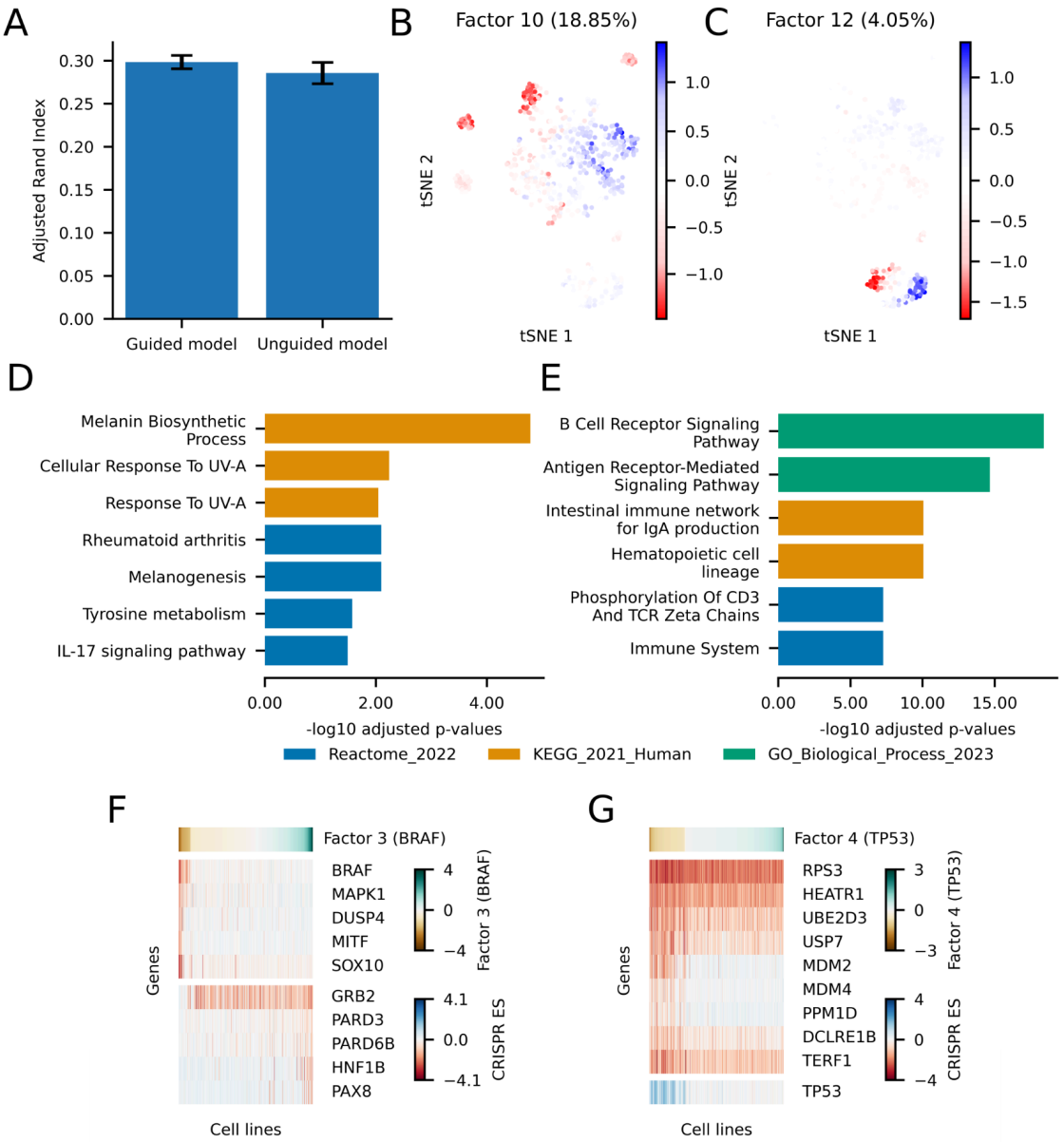

Figure S3: **A**: Adjusted Rand Index (ARI) for clusterings of the factors of a guided model and an unguided null model. The clusterings were compared to the true cancer type labels. A higher ARI suggests higher overlap of the clustering with the cancer type labels. The ARI was computed for 10 independent model initializations. **B-C**: tSNE of all factors coloured by the values of factors 10 and 12. **D**: Gene set overrepresentation analysis of highest RNA loadings of factor 3 (BRAF mutation). **E**: Gene set overrepresentation analysis of highest RNA loadings of factor 6 (hematopoietic lineage). **F**: Heatmap showing CRISPR essentiality scores (ES) of top 10 genes significantly associated with Factor 3 (BRAF mutation), tested using a linear model. The cell lines are ordered based on their values of Factor 3. **G**: Heatmap showing CRISPR essentiality scores (ES) of top 10 genes significantly associated with Factor 4 (TP53 mutation), tested using a linear model. The cell lines are ordered based on their values of Factor 4.

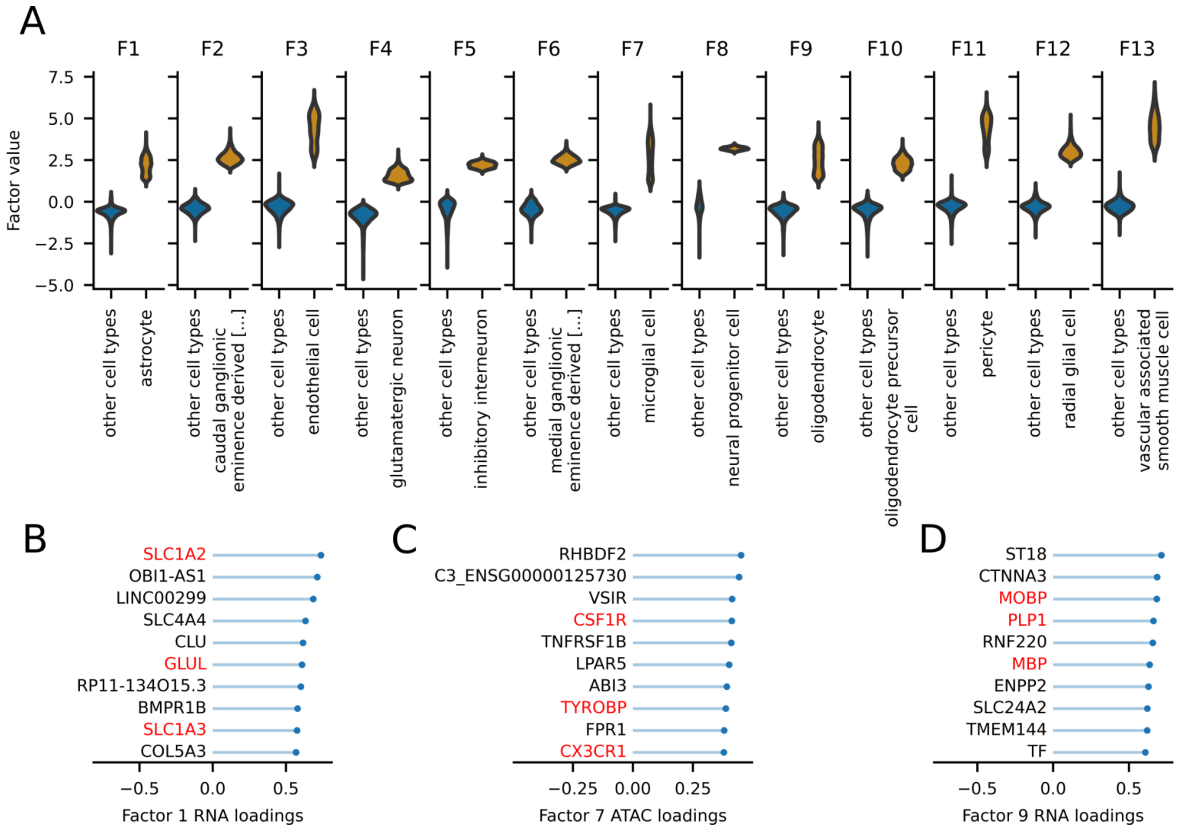

Figure S4: **A**: Violin plots of each guided factor vs their guiding cell type label. **B**: Highest transcriptomics (RNA) loadings of factor 1 (astrocytes). **C**: Highest ATAC loadings of factor 7 (microglial cells). **D**: Highest transcriptomics (RNA) loadings of factor 9 (oligodendrocytes).