Modeling causal signal propagation in multi-omic factor space with COSMOS

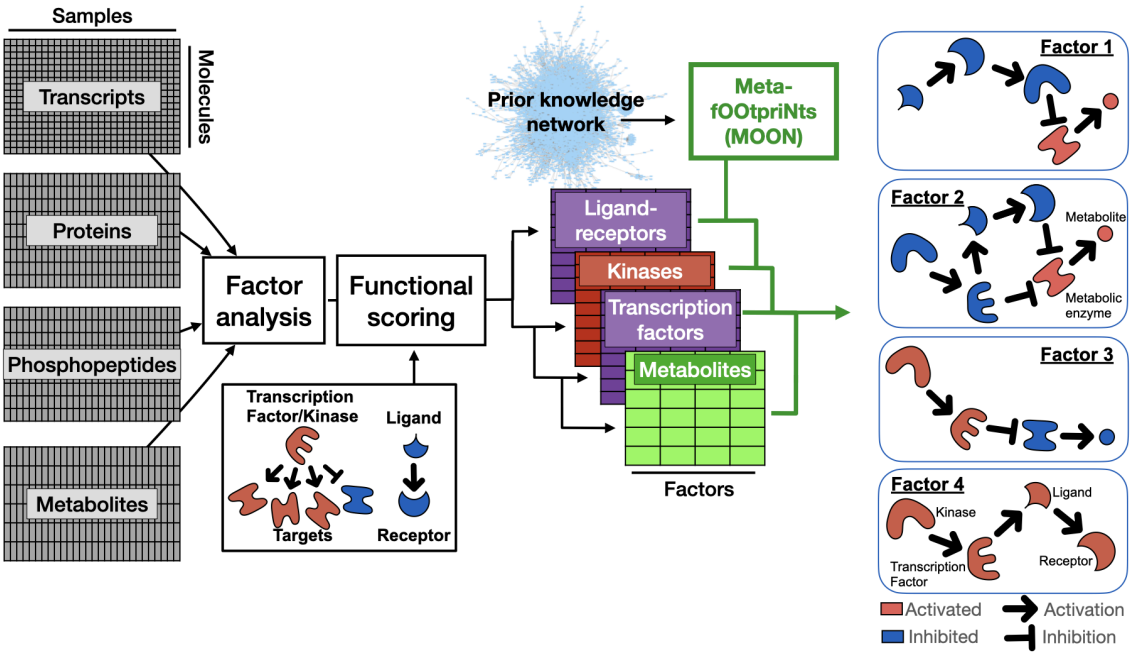
Aurelien Dugourd[1,2], Pascal Lafrenz[3], Diego Mañanes[4], Victor Paton[1], Robin Fallegger[1], Anne-Claire Kroger[1,5], Denes Turei[1], Blerta Shtylla[6], Julio Saez-Rodriguez[1,2,*]

[1] Heidelberg University, Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, Heidelberg, Germany
[2] European Molecular Biology Laboratory, European Bioinformatics Institute(EMBL-EBI), Wellcome Genome Campus, Cambridgeshire, United Kingdom
[3]Center for Molecular Biology Heidelberg (ZMBH), DKFZ-ZMBH Alliance, Heidelberg University, Heidelberg, Germany
[4]Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, Spain
[5]Institut Curie, Inserm U900, Mines ParisTech, PSL Research University, 75248 Paris, France
[6]Pharmacometrics and Systems Pharmacology, Translational Clinical Sciences, Pfizer Research and Development, La Jolla, CA, USA

* Corresponding author: saezlab@ebi.ac.uk

Abstract

Understanding complex diseases requires approaches that jointly analyze omic data across multiple biological layers, including signaling, gene regulation, and metabolism. Existing data-driven multi-omic analysis methods, such as multi-omic factor analysis (MOFA), can identify associations between molecular features and phenotypes, but they are not designed to integrate existing mechanistic molecular knowledge, which can provide further actionable insights. We introduce an approach that connects data-driven analysis of multi-omic data with systematic integration of mechanistic prior knowledge using COSMOS+ (Causal Oriented Search of Multi-Omics Space). We show how factor analysis' output can be used to estimate activities of transcription factors and kinases as well as ligand-receptor interactions, which in turn are integrated with network-level prior-knowledge to generate mechanistic hypotheses about paths connecting deregulated molecular features. Our approach offers an interpretable framework to generate actionable insights from multi-omic data particularly suited for high dimensional datasets such as patient cohorts.



Keywords: factor analysis, Multi-sample analysis, multi-omics datasets, signaling networks, metabolic networks

1. Introduction

Many diseases are the result of complex deregulations of several interconnected biological mechanisms that span over cell-cell communication mediated by ligand and receptors, intracellular signaling, gene regulation, and metabolism. Consequently, measuring multiple types of omics data, such as transcriptomics, proteomics, and metabolomics, is becoming increasingly popular [1–4] to understand such diseases [5,6].

There are many analysis methods that can help researchers extract relevant insights from such complex datasets. Factor analysis is a popular type of dimensionality reduction method for multi-omic data analysis. It provides a data driven estimation of the different axis of variability across samples and modalities [7–12]. Each omic feature is associated with weights related to latent factors, effectively untangling the sources of variability of the dataset. Such an unsupervised approach is particularly powerful when multi-omic datasets are generated from a large number of samples, such as cell line collections and patient cohorts [13–22], where group comparison is not trivial as there can be many different groups in a cohort, and the groups themselves are rarely homogeneous.

While very useful, the resulting multi-omic factors and their associated weights are purely data-driven. They can be further analyzed to provide molecular mechanistic context, and this is typically done by pathway enrichment analysis and/or association with clinical features [23–33]. Besides pathways, Ligand-Receptor (LR) mediated cell-cell communication events can also be explored in the context of such factor weights [34], by scoring the co-association of ligands and their receptors within factors, using a similar statistical framework as pathway analysis [34–36].

However, pathway enrichment and ligand receptor analysis methods are limited in their ability to provide functional and mechanistic insights - they score by analyzing the expression level (or in this case, factor weights) of their genes, which does not always reflect activity (that is, their actual influence on the biological processes their are involved in) [37]. Like the enrichment of expression of pathway components, LR co-expression scoring is only an approximation of cell-cell communication since LR co-expression is not equivalent to an LR biochemical activation; it rather allows us to prioritize which of them is more likely to induce intracellular signaling changes in a given condition [34,38].

Intracellular functional deregulations can be estimated with footprint methods such as Transcription Factors (TF) and kinase activity estimation [39]. They can share the statistical framework of pathway enrichment analysis [40], but focus on the molecules affected by the process of interest e.g. the regulation of

abundance of target transcripts by a transcription factor. Consequently, their output is more biochemically accurate than classic pathway scoring [37]. Therefore, footprint analysis can potentially help us characterize factor analysis results at a biological level as it brings complementary biological insights to pathways and LR analyses.

However, the aforementioned methods only inform about the individual processes associated with latent factors. TFs and LRs can be further connected together using networks derived from existing knowledge about interactions among molecules (prior knowledge networks) [41–48] using methods that rely on various formalisms (such as Integer Linear Programming (ILP) or network diffusion) to find paths in the network that can explain the orchestrated changes in individual processes [48,49]. They are particularly useful to propose mechanistic and testable hypotheses that can link deregulated functional features such as Ligand-Receptor interaction with intracellular signaling, gene regulation, and metabolism. Consequently, they can also theoretically be applied to further characterize biological mechanisms in latent factor space. These network approaches currently suffer from three main limitations. First, their computational complexity can limit their use in contexts where potentially many single samples, contrasts or factors need to be analyzed. Second, the number of mechanistic hypotheses they provide can often be overwhelming, thus making it difficult for researchers to prioritize them for validation and organize them into a coherent and intelligible biological story. Third, these methods inherently rely on the quality of the prior knowledge networks that are used. Errors in the prior knowledge can lead to false molecular interactions being highlighted as relevant to explain a given experimental result.

In this work, we provide an integrated solution to these challenges by developing an approach to bridge multi-omic factors analysis with methods that can put them into the context of biological mechanisms beyond classic pathway enrichment analysis, including footprint analysis and mechanistic network integration. We first demonstrate how factor weights can be used to characterize functional features associated with factors, such as Transcription Factors (TF) activity [50] and LR mediated cell-cell communication events [35]. We also developed a lightweight alternative network scoring procedure for COSMOS [42] called Meta-fOOtprint aNalysis (MOON) that can contextualize prior knowledge networks into mechanistic hypotheses. MOON can generate mechanistic hypothesis spanning over signaling and metabolism, effectively connecting perturbations observed at the level of cells kinase receptors with downstream transcription factor and metabolic de-regulations, as well as interactions between metabolic ligands and chemicals with downstream receptor and signaling cascades. We also assessed the ability of such network scoring procedures to capture relevant regulators of gene expression by applying it to a dataset of transcriptomic changes upon cytokine stimulation [51], and provide

functionalities to facilitate the interpretation of such networks by end users, thereby making it easier to pinpoint errors in prior knowledge. We refer to the new version of COSMOS that incorporates the MOON function and the interface to Factor Analysis as COSMOS+. We used this updated COSMOS+ method to extract coherent mechanistic insight such as a crosstalk between the JAK-STAT pathway, Citrate metabolism and MYC inhibition that was specifically associated with leukemic cell lines in the NCI60 dataset [18]. To make it easier for others to apply the open-source COSMOS+ methods to any multi-omic dataset, we provided a tutorial that cover step by step the analysis presented in this manuscript, as well as cosmosR installation (https://github.com/saezlab/cosmosR)

2. Results

2.1 Presentation of COSMOS+

COSMOS+ bridges mechanistic hypothesis generation with factor analysis. It contextualizes generic resources of prior knowledge spanning protein/protein and protein/metabolite mechanistic interactions with factor weights resulting from the variance decomposition of multi-omic datasets. COSMOS+ is currently tailored to work with multi-omic data sets that contain either only transcriptomics or at least two of the three following types of data: transcriptomics, phospho-proteomics and metabolomics and can be adapted to other scenarios.

Once a multi-omics dataset has been decomposed into a latent factor by a given factor analysis method, such as MOFA [7], each factor, provided it captures variance across multiple omic views, can be further processed through the COSMOS+ method to generate factor-specific mechanistic hypotheses as follows. For each factor, transcription factor activities and ligand-receptor scores are estimated from transcriptomic data, and kinase activities are estimated from phospho-proteomic data. The activities are estimated by modeling each TFs and kinases as a linear regression of their target measurements as function of their mode of regulations (see methods). Potential cell-cell interaction events are scored as a function of the co-regulation of pairs of ligands and receptors, normalized by the background weights of all other genes of the factor considered. To do so, the scores are estimated by modeling each pair of potential ligand and receptor (according to prior knowledge resource of ligand-receptor pairs) as a linear regression of the transcriptomic and/or proteomic measurements as function of their belonging to a given ligand-receptor pair (Figure 1A).

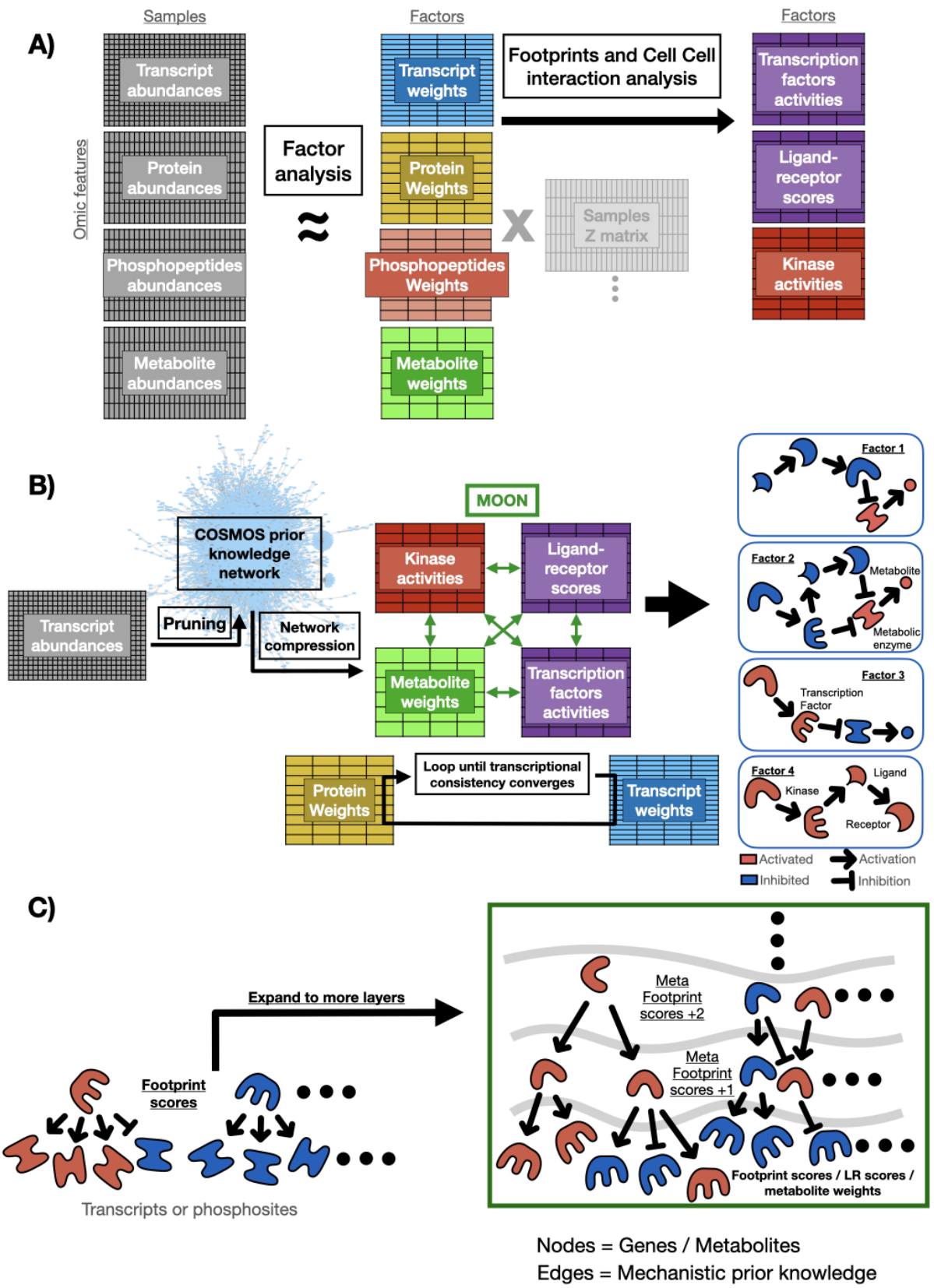


Figure 1: Schematic representation of the COSMOS+ method. *A) footprint and CCC interactions in factor space. Representation of the types of data that are currently primarily handled by COSMOS+, and how they can be processed through factor analysis to generate corresponding footprint (TFs and kinases) and ligand-receptor scores. B) Mechanistic network hypothesis*

generation. Schematic representation of the integration of the feature weights, TF, kinase and LR score with a prior knowledge network. C) Meta-footprint. Schematic representation of the concept of the Meta-fOOtprint aNalysis (MOON), which expands the concept of footprint to upstream layers. Scores are usually estimated using the Univariate Linear Model (ULM) method of decoupleR.

The next steps connect the top contributing ligands, receptors, kinases, transcription factors and metabolites of a given factor all together in a consistent network of mechanistic hypotheses using the COSMOS prior knowledge spanning over ligand-receptor interaction, intracellular signaling, transcriptional regulation and metabolic reactions is provided. The prior knowledge network consists of curated molecular interactions coming from the Omnipath database, STITCH database and recon3D human reaction network [52–54], and its building process is now automated and integrated within the OmnipathR package. If transcriptomic data is available, it can be used to perform an initial pruning. This pruning aims at removing mechanisms (e.g. post-translational modification or metabolic reactions) that are mediated by genes that are not expressed in any of the analyzed samples. Then, the meta-network is pruned to only keep mechanisms that are in-between the top contributing ligands, receptors, kinases, transcription factors and metabolites of a given factor within a set number of steps. This step removes elements that have no connecting path to perturbed (approximate non-controllable) or measured nodes (approximate non-observable) [55] . Then, redundant paths in the network (identified from parent nodes that share the exact same set of children nodes) are compressed (Supplementary Figure S1A/B).

Next, the new network meta-footprint method (MOON) is used to score the most consistent mechanisms of the prior knowledge network (PKN) connecting the ligands, receptors, kinases, transcription factors and metabolites (Figure 1B). Briefly, MOON builds upon footprint activity scoring [40] and network diffusion [56,57] and performs an iterative footprint scoring (hence called 'meta-footprint') over a given prior knowledge network, starting from a downstream layer composed of the input nodes (Figure 1C). For example, if we consider a set of scored TFs mapped on a prior knowledge network, we can score nodes directly upstream of those TFs in the network, based on the consistency of the sign of TF activity score they directly regulate. This process can be repeated iteratively over all the nodes of the PKN (provided a node has a path downstream that can reach the input nodes). Thus, to run MOON, the user specifies which set of input will be considered as downstream layer and, optionally, upstream layer. This decision depends on the biological question that is considered. For example, in the case of studying the connections between receptor, kinases and TFs, the downstream layer will be set as the TFs and the upstream layer as receptors (from ligand-receptor pairs) and kinases (estimated from phospho-proteomic data). MOON will iteratively score nodes upstream of the

TFs until it reaches the upstream receptor/kinase layer (if provided, else until it reaches the maximum number of steps allowed by the user), and then compare the MOON score of those receptors/kinases with the upstream input scores (receptors scores and kinase activities). Any receptor/kinase that shows a sign incoherence between its MOON score and the input score/measurement is pruned out along with all incoming and outcoming edges. A transcriptional consistency check can be performed that removes any interaction between TF and inconsistently regulated direct downstream targets (incoherence between sign of the TF activity score and the sign of the downstream measurement/factor weight input). After the interactions are removed, the MOON scoring procedure can be repeated with the new pruned network, until no more interactions are removed. Finally, after scoring the nodes of the network, MOON also specifically prunes out any interaction from the prior knowledge network that would be inconsistent with the MOON node scores (e.g. two nodes with opposite scores connected by a positive interaction). The output of MOON is a multi-omics network representing a set of potential mechanistic interactions consistent with the input data.

2.2 MOON cytokine prediction with the Cytosig dataset

We used MOON to score the nodes of the COSMOS network for each of the 1359 Cytosig [51] ligand expression signatures. Each of the Cytosig ligand expression signatures corresponds to the transcriptional change observed after the application of a given ligand on a cell population. These signatures have been obtained from systematic query and automated processing of several large RNA data repositories. For each of Cytosig's ligand expression signatures, the MOON scores of the corresponding ligand could be estimated in 549 out of 1359 expression signatures across 63 unique ligands (31 ligands were not reachable upstream of transcription factors in the COSMOS prior knowledge within the specified number of steps, therefore their score couldn't be estimated). Out of 63 unique ligands, 47 (75%) of them had a positive MOON score on average across their corresponding signatures (16 of them significantly different from 0, t.test p-value <= 0.05) and the rest a negative score on average (4 of them significantly different from 0, t.test p-value <= 0.05) (Figure 2A). To assess the impact of the time after perturbations as a potential confounding factor, we calculated the correlation between the MOON score of Cytosig ligands and the time delay of RNAseq measurement after perturbation for each signature, as well as between the MOON score and the shortest distance between estimated TF scores and upstream nodes of the network. The correlation between MOON score and time delay was -0.12 (Kendall' Tau coefficient, p-value = 0.002). There was no significant correlation (Kendall's tau = 0.03, p-value = 0.33) between the score of the ligand and the number of steps that separated them from their informative TFs (referred to as "level"). Thus, the time of measurement, but not the steps of the network, seem to influence the MOON score.

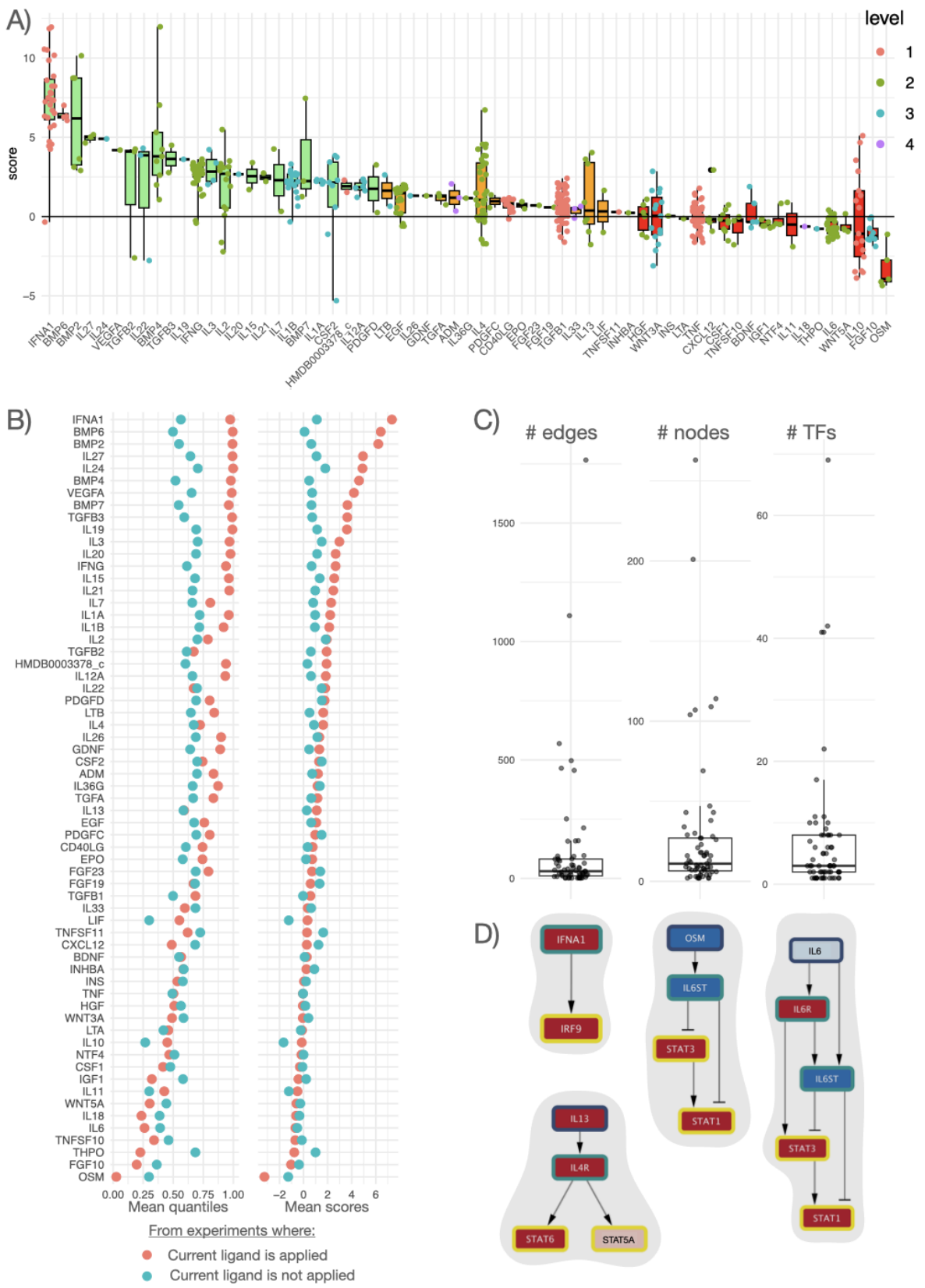


Figure 2: Estimation of the activity of ligands via MOON score on Cytosig data. *A) MOON score of ligands on the corresponding Cytosig experiment. The x-axis represents the ligand, the y-axis represents the MOON score, therefore each dot represents the score of a given ligand MOON score in each experiment where it was applied. When a given ligand is applied in an experiment, its*

corresponding MOON score is expected to be positive. The color of the dots represents the amount of time that passed between the application of a ligand and the measurement of the RNA profile for each Cytosig experiment. The color of the box plot denotes whether the average MOON score for the corresponding ligands is above 1.7 (green), between 1.7 and 0.3 (orange) or below 0.3 (red). B) mean quantiles and scores of ligands in experiments where they were applied compared to experiments where they were not applied. C) Number of edges, number of nodes and number of downstream TF participating in the MOON score estimation of the ligands. Each dot represents a ligand. D) Schematic representation of several ligand scoring networks. Nodes filling in red indicate up-regulation of activity while blue represents down-regulation of activity, in a given experiment. Color of the node border represents the distance from TFs, from yellow (level 0, TF themselves) to dark blue (level 2, 2 steps upstream of the closest TF). Flat arrowheads represent inhibitory interactions, pointy arrowheads represent activatory interactions.

To evaluate the specificity of the ligand scores, we compared their average quantile (within a given contrast) in experiments where cells were treated with them to their average quantile in experiments where cells were treated with other ligands (Figure 2B). For 44 out of 63 unique ligands, average scores were higher in experiments where they were applied compared to experiments where they were not applied, while the 19 others were equal or lower compared to experiments where they were not applied. To facilitate the interpretation of the MOON scores, we extracted a subnetwork for each ligand that contains only the nodes that connect the ligand (upstream regulator) to the nodes of the downstream layers (the most downstream layer being TFs) that were used as targets during each iterative scoring round of MOON. This showed that the ligand scores were estimated from an average of seven TFs (Figure 2C) in the most downstream layer of each respective ligands. The complexity of the MOON scoring network across the ligands varied from very simple networks (2 nodes, 1 edge) to networks with > 1000 of edges in some cases (max = 263 nodes, max = 1766 edges, Figure 2C). Half of the scoring networks of ligands comprised no more than 30 edges, 11 nodes and 3 TFs. As an example, IFNA1 ligand had a very high MOON score on average (7.3). IFNA1 has a single TF, IRF9, as a direct downstream target, (Figure 2D). The MOON scores are defined from their most direct TF targets, to capture first/early effects; other TFs further downstream than IRF9 are not used for the computation of the MOON score. The high MOON scores indicate that treating cell lines with IFNA1 will consistently lead to the activation of IRF9.

OSM ligand seemed to perform very poorly, with an average score of -3.25. OSM's MOON depends on the activity estimation of STAT3 and STAT1 (Figure 2D). We found that there was a critical error in the prior knowledge downstream of OSM, as IL6ST was annotated as an inhibitor of STAT3 and STAT1. IL6ST and OSM are in fact activators of JAK kinases[58], explaining why OSM's score was the opposite of what would be expected. This mistake was corrected in the version of Omnipath of 2024.03.19, and the average score of OSM in this new version is 4.4. This erroneous annotation also explained the seemingly poor

MOON score estimation of the IL6 ligand as well (Figure 2D), which has an average score of 2.0 in the newer version.

An example of a seemingly random performer is IL13 (0.59 average score quantile in experiments where it's applied and 0.59 as well where it's not, Figure 2D), which had relatively heterogeneous MOON scores across experiments, despite having a positive score on average (IL13: mean = 1.06, SD = 2.26). IL13 is estimated mainly from the activity estimations of STAT5A and STAT6 (Figure 2D), and the latter is a canonical outcome of IL13 [59]. While it has a high score in three experiments where it is applied (MOON score > 3), it has a negative score in 4 other experiments, leading to poor average score quantiles. The score did not correlate with either time after collection or cell type, however three out of four of the poorly scored experiments were from the same study (GSE43515). Additional examples with more complex structures are detailed in Supplementary Figure 2, showing cases where larger networks (e.g. TGFB1 and EGFR) may lead to lesser impact of prior knowledge errors and the impact of potentially context specific molecular interactions.

2.3 Footprint analysis of multi-omic factor weights.

To demonstrate its use, we applied COSMOS+ to the NCI60 dataset, consisting of transcriptomics, proteomics and metabolomics across 58 cell lines [18,60]. Preprocessing of the NCI-60 data yielded 6,000 most variable transcripts, 1897 most variable proteins and 139 metabolites measured to be used as inputs for MOFA (see methods, Supplementary Figure S3). While 6,000 transcripts only represent a fraction of all the genes that were measured, they are enough to robustly estimate the activity of >100 transcription factors (Figure 3A). Next, we ran MOFA while progressively constraining the maximum number of factors between 4 and 15, and assessed how robust the results were. We saw that at 9 factors and above, the optimal number of factors that MOFA chose didn't change anymore, consistently settling on 9 factors, regardless of whether the upper factor limit was further increased. (Supplementary Figure S4). The MOFA model with 9 factors could reconstruct different percentages of the data variance for each type of omic data (Figure 3B). 60% of the variance was reconstructed for the RNA data, 19% for the metabolomic data and 23% for the proteomic data. Interestingly, the reconstructed variance is not much higher for the proteomic data compared to the metabolomic data, despite having more than 10 times more quantified features (1897 proteins vs 139 metabolites). This is likely due to the fact that, in general, metabolic abundances are more correlated between each other than protein abundances, due to the entangled metabolic reactions that connect them [61–63], making it easier to capture coordinated patterns despite having less features available to fit the MOFA model (Supplementary Figure S5). Factor 2 and 4 could explain variance in the data across all three omic views.

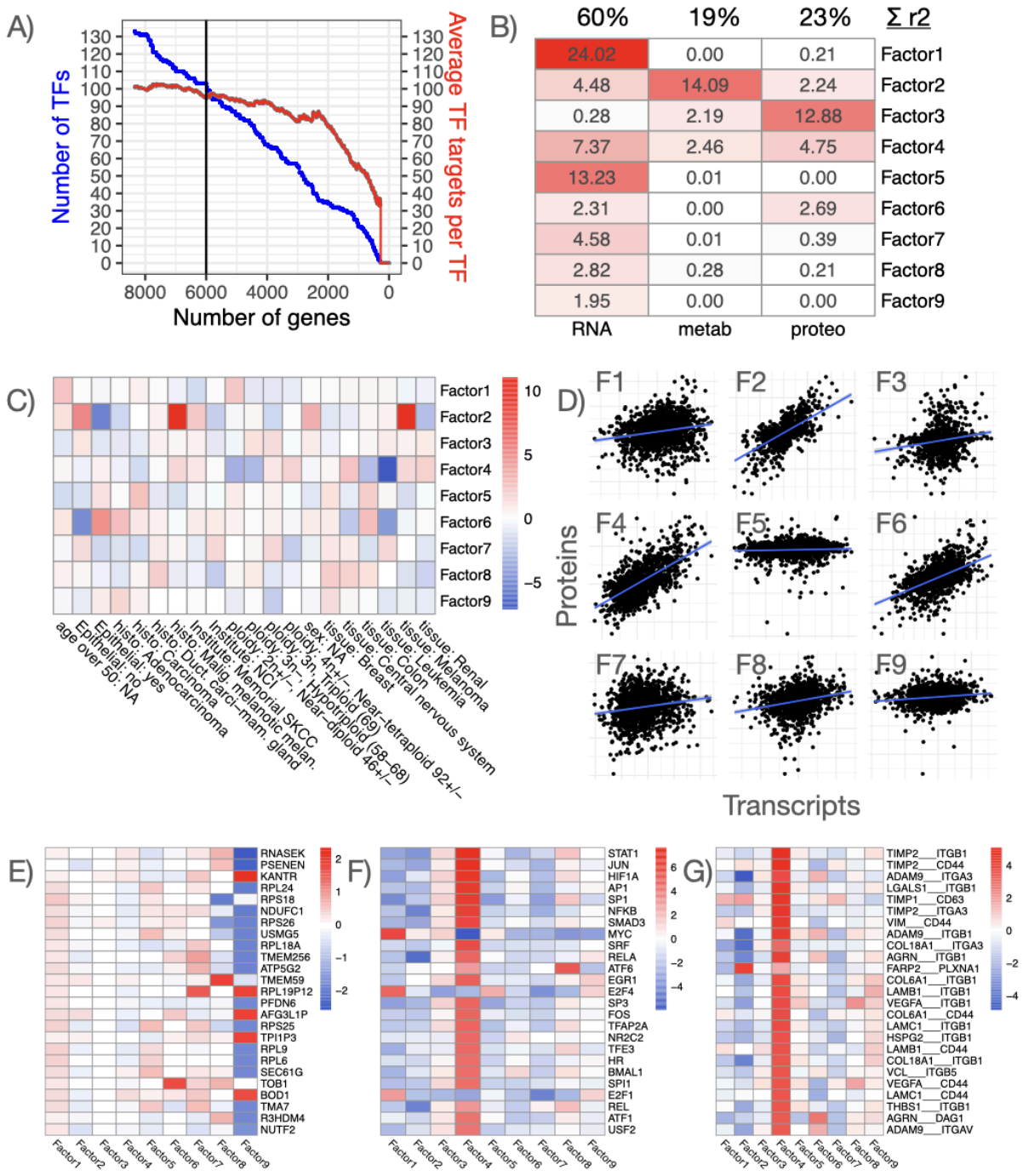


Figure 3: MOFA and footprint analysis results. *A) This plot shows how many TFs have enough measured targets (>10) to estimate their enrichment scores, given how many top variable genes are kept in the analysis. Number of TFs enrichment scores that can be estimated as a function of the number of top variable genes (blue line) and average number of targets per TFs as a function of the number of top variable genes (red line). B) Heatmap of the % of variance explained per omic view and per MOFA factor. C) Enrichment analysis of cell line tissue of origin metadata in each of the MOFA factors. Enrichment scores are computed using the ULM method of decoupleR (t-value of linear regression between the MOFA Z values in the background and each metadata class, such as tissue of origin, age of donor, pathology, etc..). E) Heatmap of the top MOFA RNA features weights across 9 factors. The weights are ordered by their absolute maximum across all factors. F) Heatmap of the top transcription factors associated with MOFA RNA feature weights across 9*

factors. The scores are estimated from RNA weights using the ULM method of decoupleR with the collectrI TF-target regulons and are ordered by their absolute maximum across all factors. G) Heatmap of the top ligand receptors associated with MOFA RNA feature weights across 9 factors. The scores are estimated from RNA weights using the ULM method of decoupleR with LIANA's consensus ligand receptor prior knowledge and are ordered by their absolute maximum across all factors.

We then checked which meta-data features of the NCI60 cell lines were enriched in each factor. MOFA decomposes the data into a Weight matrix and a Z matrix of cell line coordinates in the factor space. The latter can be analyzed to find meta-data features that are associated with the Z matrix coordinates. We used linear regressions (see methods) to model the coordinate of each cell line in the MOFA Z matrix as a function of each binarized clinical feature category (such as Tissue: renal = 0 or 1). Factor 4 was showing a significant negative association with samples of Leukemic origin. We also saw that factor 2 was significantly associated with a Melanoma origin, and negatively associated with an Epithelial origin (Figure 3C).

To get a general sense of the relationship between RNA and protein MOFA weights, we first explored how correlated were the weights of transcripts and of protein abundances (Figure 3D). The correlation between RNA and protein weights was close to 0 for factors 1, 3, 5, 7, 8 and 9. For factor 2, 4 and 6, the correlation was respectively 0.41, 0.42 and 0.29. This showed that the variance reconstructed for factor 2, 4 and 6 across the RNA and proteomic view is, at least partly, capturing the relationship between RNA and the corresponding protein abundance. Of note, the correlation between RNA and protein weights for those factors is of a similar magnitude as usual RNA/protein correlation at the level of bulk data [64,65] (mean correlation for each cell-line across genes between RNA and protein value = 0.41 +- 0.04 s.d.; mean correlation for each gene across cell-lines between RNA and protein value = 0.29 +- 0.24 s.d.) (Supplementary Figure S6).

Then, we set to characterize mechanistically the factors using footprint and LR analysis methods. To do so, we first focused on the RNA weights of each factor (Figure 3E). We saw that factor 9 had much more extreme RNA weight values than other factors. However, this factor only captures a small portion of the variances of the RNA data, and no variance in the other omic views. We then wanted to see if some biological processes could be associated with the RNA weights for each factor. First, we used linear regressions to model RNA weights as a function of their status as targets of each TF of the CollectrI database [50]. This allowed us to obtain scores that represent how consistently high or low are RNA weights of the transcriptional targets of given TFs compared to the background of RNA weights. We can interpret such scores as TFs that are responsible for the transcriptional programs that are captured by given factors.

The TF scores were especially significant for factor 4 (Figure 3F). This indicates that the variance captured by factor 4 is especially consistent with our available prior knowledge about transcriptional regulations. The most expressed TF scores for factor 4 were members of the JAK-STAT, NFKB and hypoxic pathways (STAT1, NFKB, SMAD3, HIF1A). Since factor 4 was negatively associated with cells of Leukemic origins, this may indicate that those pathways are specifically less active in those cells than in cell lines of other origins. Then, we used decoupleR's [40] ULM function again to model RNA weights as a function of their status as members of ligand-receptor (LR) pairs. This allowed us to obtain LR scores that represent how consistently higher or lower are the weights of transcripts coding for LR pairs compared to the weights of other transcripts in each factor. Interestingly, factor 4 again showed the most consistency between RNA weights and prior knowledge about LR interactions (Figure 3G). The most significant LR pairs of factor 4 were involved in processes such as cell adhesion (e.g. CD44, ITGBs, VIM, etc...) (Figure 3C).

2.4 Exploring and interpreting mechanistic multi-omic networks of the Leukemia-Melanoma factor

To complement the TFs and LR interactions associated with factor weights, we used MOON to find network-level hypotheses supported by the RNA, protein and metabolite weights of the MOFA factors. We specifically focused on factor 4, since it captured variance at the level of all three omic views (RNA, proteomic and metabolomic), and was especially consistent with TF-target and LR interaction prior knowledge (Figure 3F,G).
The resulting MOON network (the network after application of a threshold on the moon scores) of factor 4 shows which of the mechanistic interactions available in prior knowledge resources (e.g. the COSMOS PKN) are consistent with the feature weights of factor 4 (Figure 4A). For example, we can see that VIM has a high weight (0.7) in factor 4, while SMAD3, STAT1, STAT3 and JUN have many transcriptional targets that have a high weight in factor 4 (targets mean = 0.13, including VIM). Furthermore, we can also see that MAPK1 (MOON score = 3.2), PRKCA (MOON score = 2.4), ABL1 (MOON score = 2.8) and JAK1 (MOON score = 2.8) have all high MOON scores, which indicates that they are directly upstream of a high number of consistently up-regulated TFs (such as SMAD3, STAT1, STAT3 and JUN).

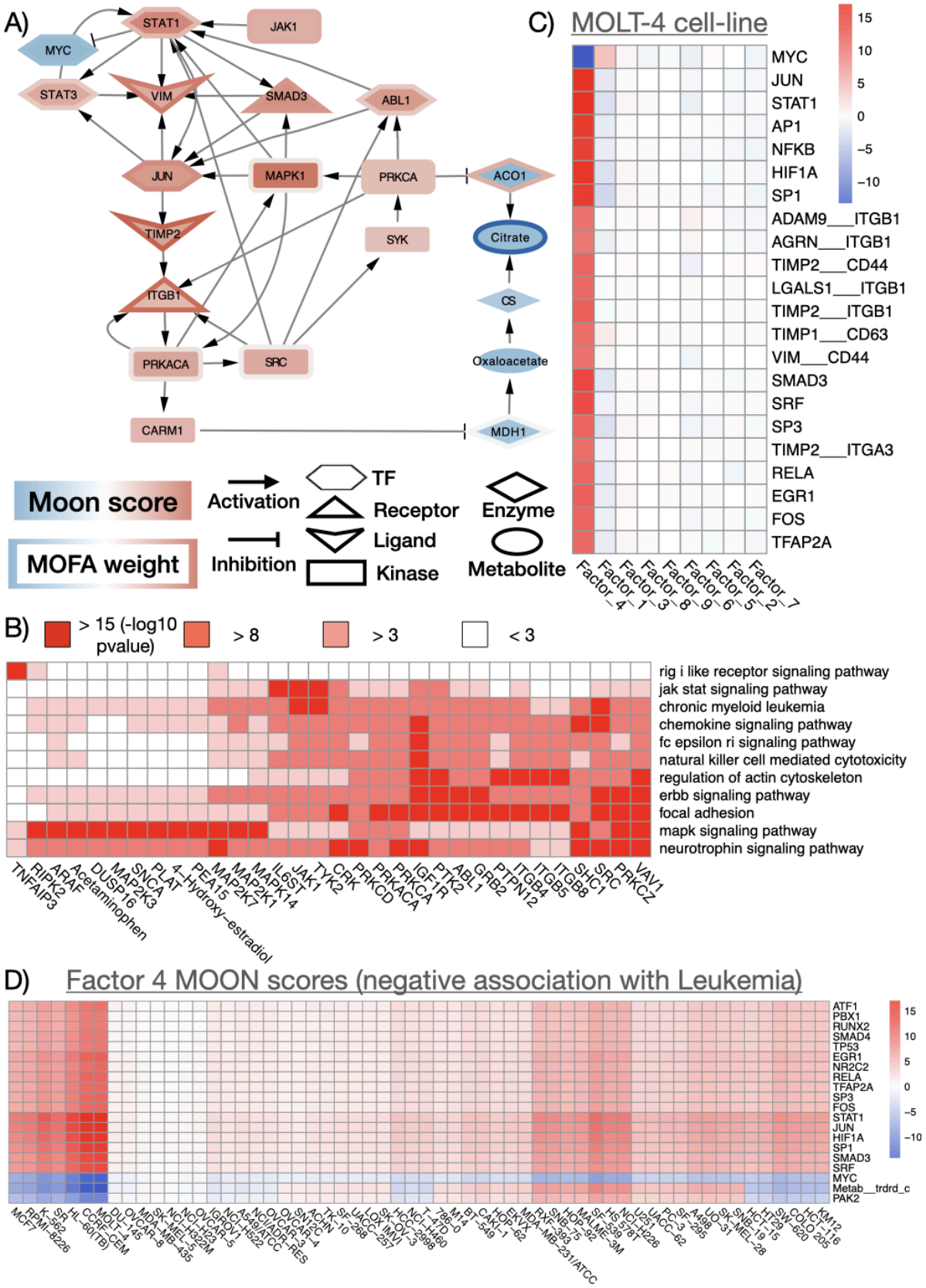


Figure 4: Mechanistic interpretation of a MOFA factor associated with blood cancer cell lines, such as MOLT-4 cell lines. *A) MOON Network connecting the top deregulated TFs and LR interactions of factor 4 based on a signed directed prior knowledge network. B) Heatmap of the top results of the Pathway control analysis. We represent pathways that are significantly*

over-represented downstream of given nodes of the MOON thresholded network C) Heatmap of TF and LR scores estimated for MOLT-4 cell line projected in the factor space. Values represent the t-value of the ULM decoupleR score estimation scaled with the MOFA Z values for MOLT4 across factors. From the scores in factor 4, we see that MOLT-4 cell line seems to have especially high scores for STAT and HIF1A TFs as well as matrix related ligands such as COL18A1 or TIMP1/2. D) Top 20 TF and moon scores estimated from the factor 4 (negatively associated with Leukemia cell lines) RNA reconstructed data across 58 NCI60 cell lines.

Complementarily, their corresponding transcripts also have high weights in factor 4 (MAPK1 = 0.10, PRKCA 0.33, ABL1 = 0.25, JAK1 = 0.43 factor 4 weight). Taken together, factor 4 appears to be characterized by a consistent network of interactions connecting VIM as a downstream transcriptional target of SMAD3, STAT1, STAT3 and JUN, themselves regulated through post-translational regulations by MAPK1, PRKCA, ABL1 and JAK1. Other transcription factors, ligands and receptors also show activities and weights consistent with this signaling and transcriptional regulation, such as MYC, TIMP2 and ITGB1 (Figure 4A). MOON network also points at potential metabolic regulations: ITGB1 and MAK1 can activate PRKACA, which can itself activate CARM1, an inhibitor of the MDH1 metabolic enzyme through methylation (Figure 4A). The inhibition of MDH1 could indirectly explain the downregulation of citrate (citrate = -0.38 factor 4 weight). Finally, PRKCA activation by SYK and SRC downstream of PRKACA can also lead to the inhibition of the citrate conversion activity of the metabolic enzyme ACO1 [66,67] metabolism deregulation.

To find which biological processes are captured in the moon network, we characterized the top scoring nodes of the MOON network (absolute moon scores > 1.5) with the biological pathways that are associated with nodes found closely downstream (for example, within a maximum of 2 steps, see methods: pathway control analysis). Then, we can represent pathways that are significantly over-represented downstream of given top scoring nodes of the MOON thresholded network in a heatmap (Figure 4B). Reassuringly, this analysis found expected control mechanisms, such as JAK1 or IL6ST very significantly controlling the JAK-STAT signaling pathway or ITGB (integrins) family members controlling focal adhesion. It also allows us to propose chemicals that can potentially control signaling pathways such as Acetaminophen or 4-hydroxy-oestradiol controlling the MAPK pathway.

Finally, it is also possible to assess how each individual cell line can be projected in the different factors. This is particularly useful when trying to interpret such factor level mechanistic features at the level of single samples. For example, we can focus on MOLT4, a lymphoblastic cell line, which is the cell line with the most negative coordinate in the factor 4 Z matrix. By multiplying the Z value of MOLT4 with respect to each of the 9 factors with their corresponding weights in the W matrix of MOFA, we obtained a projection of the transcript, protein and metabolite abundance in the factor space (Supplementary Figure S7). Then we

computed TF and lig-receptor scores from the projected transcript abundance and scaled their values with the corresponding factor Z values (Figure 4C). Coherently, we saw that MOLT4 had specifically high scores for the TFs and ligand-receptor interactions that were high in factor 4 (such as STAT1 = 17 scaled ulm t-value or JUN = 16.1 scaled ulm t-value), and vice versa for TFs and ligand-receptor interactions that had low scores in factor 4 (such as MYC = -13.4 scaled ulm t-value). Furthermore, MOLT4 variance does not seem to be captured in any other factors, besides factor 1 showing a milder opposite pattern compared to factor 4 with respect to the RNA weights only (scaled ulm t-values ranging between -2.9 and 4.2 in factor 1 compared to -13.2 to 17 in factor 4). We can also do the same operation but focusing on a given factor rather than a cell line. For example, we reconstructed the RNA data with respect to factor 4 and estimated TF activities, LR scores and subsequently MOON scores connecting the receptors to downstream TFs for each of the 58 cell lines of the NCI60 dataset. This showed that the hypotheses generated for factor 4 were the most relevant for MOLT4 and expectedly other leukemic cell lines such as CCRF-CEM, HL60-(TB), SR, K-562 and RPMI-8226, but also for cell-lines of breast cancer origin such as MCF7 (Figure 4D).

3. Discussion

In this study, we introduced COSMOS+, a method to combine factor analysis and Prior Knowledge Networks to extract interpretable mechanistic insights from complex multi-omic experiments. COSMOS+ includes a novel and efficient network scoring procedure meta-footprint (MOON) and provides an interface for factor analysis. It introduces additional updates over the first version COSMOS[42], such as a compression algorithm to remove redundant paths in the network and a method to highlight pathway control mechanisms. We also further curated the metabolic reaction network used to build the prior knowledge network of COSMOS and refactored its building procedure within the OmnipathR package. All those updates facilitate the use of COSMOS+ across a wider range of applications with enhanced interpretation, including the ability to give insights at the level of single samples projected in factor spaces. This allowed us to apply COSMOS+ to a cohort of > 1300 perturbation experiments of the Cytosig dataset [51], and find and fix errors in the prior knowledge that affect the MOON scoring procedure.

We also illustrated how footprints and meta-footprints based on prior knowledge networks can be applied downstream of factor analysis. To do so, we analyzed the transcriptome, proteome and metabolome data of the NCI60 cancer cell lines [18]. We estimated transcription factor activity and ligand-receptor scores from MOFA factor weights and we showed how COSMOS+ points out a potential biological crosstalk between JAK-STAT, ITGB1, PRKCA and citrate metabolism to be a specific regulatory mechanism associated

with leukemia-like cancer cell lines. Finally, we showed how the resulting mechanistic hypothesis network of COSMOS+ is coherent with expected regulatory mechanisms, while hypothesizing novel control mechanisms spanning metabolites and signaling occurring in leukemic cell lines, such as the MAPK pathway being controlled by Acetaminophen or 4-beta-oestradiol. Taken together, these results highlight the synergy between MOFA, footprint and network methods to derive interpretable and actionable biological insights.

Factor-based variance decomposition (i.e. partitioning the variance of a variable (or a set of variables) into components attributable to different sources or factors) has been used for a long time in natural sciences [8], and they are commonly used to analyze omics data [9]. They are particularly suited to handle large cohorts of samples such as cancer patient cohorts [13–17] and cell lines collections [19–22] without explicitly grouping samples. Most of the current efforts to interpret the results of such analyses have focused on characterizing known meta-data variables (e.g. clinical variables) that may correlate with factors, or using pathway ontologies to interpret the feature weights of factors [23–33,68,69]. However, there are many more types of domain knowledge that can potentially be used to interpret feature weights of factors beyond pathway ontologies, such as prior knowledge in the form of footprints and signed-directed networks can help to provide interpretable insights from factor weights. COSMOS+ is agnostic to the variance decomposition method used, as long as they provide a matrix of factor weights.[10], and it can also be applied with other methods instead of MOFA (such as MCFA [10] or MEFISTO [11]).

Footprints and LR analysis can be performed without further downstream network analysis, and various network methods can be used without prior footprint analysis [41,43–48]. Nonetheless, network methods, such as CARNIVAL, that specifically model flows of activity in biological networks particularly benefit from inputs generated from prior footprint analysis [39]. Footprint and LR analysis using factor weights opens the door for factor-specific mechanistic network-level hypotheses, which we demonstrate using the MOON algorithm on the output of a multi-omic factor analysis of the NCI60 dataset. MOON relies on iterative scoring, representing a “greedy search” alternative to global network optimization algorithms based on integer linear programming, such as CARNIVAL[41]. The iterative upstream scoring procedure of MOON has similarities to the scoring approach of causalR [44], the main difference being that MOON scores each node iteratively, by using only direct downstream interactions at each step, while causalR sequentially considers all the nodes that can be reached downstream of a given node (or upstream of a set of deregulated RNA measurements) within increasing maximum number of steps and then aggregates the results obtained with different maximum number of steps. Furthermore, the MOON scoring procedure relies on a linear model while causalR is based on a form of over-representation estimation. Finally, MOON

was developed natively in the context of multi-omic mechanistic hypotheses generation within COSMOS+, which makes it particularly suited to use with the multi-omic factors of MOFA.

We used the Cytosig dataset to evaluate if MOON's scoring procedure yields molecular activity scores (i.e. ligand scores) coherent with ground truth (e.g. The treatment of cells with said ligand). Our results showed that the downstream prior knowledge network of most ligands of the Cytosig dataset were coherent with the measured RNA expression profiles. However, not all ligands were properly scored, and the ground truth was blurred by confounding factors, such as differences in the experimental procedures between RNA profiles (e.g. measurements at different time points after treatment). One of the main benefits of prior knowledge network based methods is that these networks are interpretable, which makes them particularly suited to extract actionable insights from data.[70,71]. Their relationship can be explained with a sentence, for example a SMAD3-VIM interaction can be interpreted as: "SMAD3 regulates the transcription of VIM". Thus, bridging factor-based analysis (e.g. MOFA) and network integration methods (e.g. COSMOS) allows us to put the relationship captured by factor weights in the context of known biological mechanisms such as transcriptional regulations, post-translational modifications (PTM) mediated regulations or metabolite/metabolic enzyme interactions.

Paradoxically, the sheer amount of prior knowledge available about biological molecules can negate the expected interpretability of such an approach[72]. Therefore, there are still significant challenges to extract actionable insights from contextualized prior knowledge networks. To address these challenges, various approaches have been developed such as summarizing metrics (e.g. network flow with node centrality [73]) and matrix representation of graphs [74,75]. In this work, we sought to design an approach that would abstract the network into a matrix format that could also directly inform us about the biological functions regulated by each node of the contextualized network and score the significance of such regulations. We built upon the ideas of network flow[76] and guilt by association[77] while taking the sign of the interactions into account, to intuitively and intelligibly highlight the most significant regulation events captured by a contextualized prior knowledge network.

Besides this, a question that often arises when combining factor analysis and prior knowledge is whether prior knowledge should be used before or after factor analysis. Indeed, there exists other methods that try to directly constrain the factor space with prior knowledge [14,78,79]. Using prior knowledge before performing a factor analysis constrains the considered feature space to the specific subset that is covered by the prior knowledge. For example, we can consider a feature space composed of gene expression measurements and a prior knowledge source being a TF-target network. Then, if TF activity scores

are estimated first using an enrichment analysis, and then factor analysis is performed on the resulting TF scores, the task is effectively to find coordinated axes of variability between TF scores, which themselves are the proxy of the expression of their targets. However, if factor analysis is performed first, and then TF activity scores are estimated on the factor weights, then the task is to find if the resulting factors, that are build using a feature space regardless of whether or not genes are known targets of TFs, are more or less coherent with sets of TF-targets. The latter aims at exploring the variance of the data itself first and then tries to relate it to prior knowledge, while the former tries to find coordinated patterns within an already defined knowledge space. Consequently, when factor analysis is performed first, the latent model can be more affected by the inherent noise of the data, and the latent factors may miss relevant biological signals. On the other hand, performing the factor analysis after constraining the feature space may lead to some association to be missed if they relied on pivotal features that were filtered out. Therefore, the decision about using one approach or the other depends on the expected scope of the latent space that is being constructed. In this study, we specifically aimed to get a comprehensive view of the major source of variations across cell lines with respect to their transcriptomic and metabolomic signatures.

While the approach described here can be useful to help researchers to extract actionable insights from their data, it has several limitations. The first and foremost limitation comes from the use of prior knowledge to support those analyses. Indeed, prior knowledge can never be complete and is rarely error proof [48]. This leads to two types of errors that methods that rely on prior knowledge can make. First, it will potentially miss interactions that underlie the observed data. COSMOS+ is fully constrained within the space of the prior knowledge that is used, and therefore will miss any interaction that isn't covered in the prior knowledge, which is known to be incomplete and biased [48]. The second type of error comes from using interactions that are the product of miss-annotated database entries or flawed experimental sources to interpret the data at hand. This can lead to methods highlighting false interactions as the potential cause of variations observed in the data. Due to the recursive nature of the MOON algorithm, errors in the PKN have a risk to be propagated upward and affect many upstream scores. As we show with the analysis of the Cytosig dataset, this error can be mitigated by cross-checking the source of any critical interactions highlighted by the analysis (consequently, making it easy to perform such cross-checking is crucial) and, if it becomes apparent that the interaction is fictitious, reporting it to the source database, removing it and repeating the analysis without its influence on the pipeline. As the coverage and quality of prior knowledge resources keep improving, COSMOS+ results will become more accurate.

The ability to handle and recover negative feedback loops in prior knowledge networks could be an interesting future direction to explore with such modeling strategies. Indeed, negative feedback loops cannot be recovered by this type of methods unless timepoints are explicitly considered by the network scoring procedure. Recovering a negative feedback loop requires the observation of a given node switching sign between two timepoints. Since COSMOS+ can be applied downstream of factor analysis, it can benefit from the fact that such type of analysis can be applied to time series and spatially resolved data [11]. Beside finding mechanistic networks connecting data that already encode time, the network scoring procedure could be adapted to explicitly take time into account, e.g. using each previous iteration as a starting point for the next one, similar to the PHONEMES-ILP method [49]. Of note, MOON recovers positive feedback loops, because it only prunes out edges from the network when there is an explicit incoherence between the activity sign of multiple interactors.

COSMOS+ could in principle be applied to any type of omic data that can be related to activation or inhibition of node activities in a biological molecular network, such as chromatin accessibility [80], mutational data [81], high throughput protein conformation [82,83], microRNA [84] or other PTMs beside phosphorylation [85,86]. For this, prior knowledge resources that would allow linking such measurements to functional readout are needed (e.g. non-ambiguous miRNA-target interactions, functional annotation of ubiquitination and acetylation, etc…).

To conclude, we present an analytical workflow, implemented in the tool COSMOS+, that combines factor analysis, mechanistic features, and prior knowledge networks to extract meaningful biological insights from multi-omic data. The software and a step-by-step tutorial are available at https://github.com/saezlab/Factor_COSMOS.

4. Methods

4.1 NCI60 Multi-omic data processing

4.1.1 Transcriptomic data

The NCI60 transcriptomic dataset was obtained through the https://discover.nci.nih.gov/cellminer/home.do portal. The data was obtained as log2(FPKM + 1) values. To filter out lowly expressed genes, we set any log2(FPKM+1) value that was lower than 1 (which corresponds to an original FPKM value of 1) as NAs. Then, any gene that had an NA value in more than ⅔ of samples was excluded. Next, genes were ordered by their standard deviation across samples, and the top 6000 genes were kept for further analysis with

MOFA. To decide on the number of transcripts to keep, we tried to reduce it as much as possible while still having enough data to estimate as many transcription factor (TF) activities. We used CollectrI regulons to determine which transcript was a target of a TF, then we checked how many TF would still have at least 10 targets when we progressively reduce the number of transcripts from 8354 to 0 (Figure 1D), from most variable to least variable across cell lines. We saw that at 6000 transcripts, >100 TFs still had at least 10 target transcripts. At 3000 transcripts, that number would go down to > 55 Tfs. Considering this, we settled for 6000 transcripts, so that we would still be able to estimate a good number of TFs for downstream analyses. , at 6000 transcripts, the average number of targets per TF was > 90 (Figure 1D), ensuring relatively robust TF activity estimates.

4.1.2 Proteomic data

The NCI60 proteomic dataset was obtained through the https://discover.nci.nih.gov/cellminer/datasets.do portal. The data was obtained as log10(intensity + 1) from Sequential Window Acquisition of all Theorical MS intensity values (SWATH). Duplicated protein entries were averaged to generate a set of unique proteins with corresponding intensity values. Any log10(intensity+1) value of 0 was converted to NA (which corresponds to an original intensity value of 0). Next, proteins were ordered by their standard deviation across samples, and the top 60% of proteins were kept for further analysis with MOFA.

4.1.3 Metabolomic data

The NCI60 metabolomic dataset was obtained through the https://wiki.nci.nih.gov/display/NCIDTPdata/Molecular+Target+Data portal. The data was obtained as log2(MS intensity + 1). Triplicates were averaged to match the structure of the RNA and proteomic data. Values above 32 were considered outliers and were replaced by NAs. Metabolite identifiers were matched to primary HMDB identifiers manually, one by one.

4.2 MOFA

The different omics were assembled into a single mofa ready input file. For each type of omic data, cross-correlation between features were evaluated using Pearson correlation with pairwise complete observations. The correlation between each transcript and its corresponding protein was also evaluated using Pearson correlation with pairwise complete observations. Mofa was run with the following options: scale_groups = "False", scale_views = "False", likelihoods = c('gaussian','gaussian',,'gaussian'), spikeslab_weights = "True", spikeslab_factors = "False", ard_factors = "True", ard_weights = "True", iter = "1000", convergence_mode = "fast", startELBO = "1", freqELBO = "1", dropR2 = "0.001",

gpu_mode = "True", verbose = "False" and seed = "1". The maximum number of factors was set sequentially between 5 to 15, 20 and 58. The resulting models were compared by correlating their weight matrices (Pearson correlation). Since factors seemed to converge after setting the number of factors to 10 and above, the model with 10 maximum factors parameter was kept for subsequent analysis.
Clinical information about cell lines were discretized when necessary (such as age variable split between > 50 and <= 50 years old). The clinical variables were then converted into a sample-clinical variable edge table. The ULM method of decoupleR was used to model each factor Z values as a function of their belonging or not to each given clinical variable categories. The resulting top enriched clinical variable categories in each factor were thresholded using an absolute t-value of 2 (t-value of ulm's linear coefficient).

4.3 Footprint and ligand-receptor analysis with MOFA weights

Ligand-receptor interactions were obtained from the consensus ressource of the LIANA R package (v 0.1.5), and decomplexified using the decomplexify function. Then, TF target interactions were converted into a ligand-receptor set collection by associating each ligand and receptors to their corresponding ligand-receptor (LR) set. TF-target regulons were obtained using the get_collectri function of the decoupleR R package (v 2.5.2). For each factor, decoupleR's ULM function was used for each TF and LR set to model the RNA MOFA weights as a function of their belonging or not to each TF and LR set, respectively.
For each factor, weights of transcripts and their corresponding protein were correlated using Pearson coefficient.

4.4 COSMOS

The weights of factor 4 were extracted to prepare them as input for the downstream mechanistic hypothesis generation. The prior knowledge network (PKN) was obtained from the COSMOS package (v 1.5.2). to select genes that were consistently regulated at both the level of RNA and protein in each factor, the weight of each gene that had an absolute value of less than 0.2 for their RNA weight and less than 0.05 for their protein weight were set to 0. Metabolite weights that had absolute weight values < 0.2 were also set to 0. The thresholds were estimated from visual inspection of the weight distributions in factor 4.

4.4.1 COSMOS with Meta-fOOtprint aNalysis (MOON)

MOON is a way to score the nodes of a prior knowledge network using successive iterations of the Univariate Linear Model (ULM) method from decoupleR's package over a network, layer by layer, starting from a set of

downstream nodes. At the first iteration of the MOON algorithm, ULM is run using measurements (e.g. metabolomic abundances) or activity scores (e.g. TF activity scores) as input data, and the prior knowledge network as a regulatory network input. Since ULM only considers regulons that have measured downstream targets, only nodes that are directly upstream of the provided measurements of activity scores will be scored at this stage. Thus, the score assigned to nodes at the first iteration represents how their direct targets are significantly up or down-regulated (consistently with the sign of the interactions), compared to the rest of the downstream layer data points. Once those nodes have received a score, they become the input downstream layer for the next scoring iteration, and are removed from the list of possible source nodes of the network to avoid multiple scoring of each node. This process is repeated as long as (1) the iteration number is lower than a user defined limit, (2) there are still source nodes upstream of the last layer that was scored in the network. The output of the MOON function is a dataframe with all the nodes of the input network scored, as well as a corresponding layer index, indicating at which iterations they were scored (that is, how many steps from the original downstream layer input they are).
If an upstream layer input is provided, any node that has a score that is different from the data provided in the upstream layer input is removed from the output dataframe. The algorithm describing the base MOON function is as follow:

<u>Pseudocode for the moon Algorithm</u>

Parameters:

upstream_input: Optional. A named vector representing upstream nodes and their corresponding activity.
downstream_input: A named vector representing downstream nodes and their corresponding activity.
meta_network: A data frame representing a signed, directed network of molecular interactions.
n_layers: An integer representing the number of layers for upstream propagation.
statistic: A string representing the statistical method to use for calculating propagation. Can be "norm_wmean", "wmean", or "ulm".

Algorithm moon(upstream_input, downstream_input, meta_network, n_layers, statistic):

1. Filter `meta_network` to remove rows where the source node is in `downstream_input`. Store this in `filtered_network`.

2. Calculate initial propagation `n_plus_one` based on `downstream_input` and `filtered_network` using the statistical method specified by `statistic` ("norm_wmean", "wmean", or "ulm").

3. Initialize an empty list `res_list` to store the results.
4. Add `n_plus_one` to `res_list`.

5. For `i` from 2 to `n_layers`:
a. Update `filtered_network` to remove rows where the source node is in the last calculated `n_plus_one`.
b. Update `n_plus_one` by running the statistical method specified by `statistic` on the last entry in `res_list` and `filtered_network`.
c. Add the new `n_plus_one` to `res_list`.

6. Combine all results in `res_list` into a single data frame `final_result`.

7. If `downstream_cutoff` is specified, filter `downstream_input` based on this threshold and add it to `final_result`.

8. If `upstream_input` is provided:
a. Filter `final_result` based on the concordance of signs between `upstream_input` and `final_result`.

9. Return `final_result`.

End Algorithm

If RNA data points (here, mofa weights) are provided, a check can be performed to remove any interaction from the network that connects a TF and a downstream gene that has an incoherent expression sign with the TF MOON score. Thus, the moon function and the TF-target coherence check can be run in a loop until the output of the moon doesn't contain any incoherence between TF scores and downstream targets. The algorithm to remove incoherent TF-target interactions is as follow:

<u>Pseudocode for filtering incoherent TF target interactions</u>

Parameters:

- moon_res: The result of the MOON scoring procedure.
- TF_reg_net: The TF target network that was used to compute the TF scores
- meta_network: The complete network data, a data frame representing a signed, directed network of molecular interactions.
-RNA_input: The RNA measurements that were used to compute the TF scores

Algorithm filter_incohrent_TF_target(moon_res, TF_reg_net, meta_network, RNA_input):

1. Initialize `recursive_moon_res` from `moon_res`.
2. Initialize `dorothea_reg` from `TF_reg_net`.

3. Create a data frame `RNA_df from `RNA_input`.

4. Filter `recursive_moon_res` to keep only the rows where the source is in `dorothea_reg$source`. Store the result in `reg_meta`.

5. Merge `reg_meta` with `dorothea_reg` on the "source" column.
- Rename the second column to "TF_score".

6. Merge `reg_meta` with `RNA_df on the "target" and "node" columns.

7. Compute an "incoherent" column in `reg_meta` based on the sign of the product of "TF_score", "RNA_input", and "mor".

8. Create an "edgeID" column in `reg_meta` by concatenating "source" and "target" with an underscore.

9. Extract "edgeID" of incoherent edges from `reg_meta` into `incoherent_edges`.

10. Create an "edgeID" column in `meta_network` by concatenating "source" and "target" with an underscore.

11. Filter out the rows in `meta_network` where "edgeID" is in `incoherent_edges`.

12. Return `meta_network` without the "edgeID" column.

End Algorithm

If there are multiple redundant parallel paths in the network, this can lead to score biases, because two nodes that have the same downstream targets will receive the same score, and thus will count double in the background of the next scoring iteration. To avoid this issue, the PKN can be compressed using a simple procedure that combines any set of nodes that have exactly the same direct downstream direct targets into single virtual nodes that can later be decompressed back into the original nodes using a hash table mapping virtual nodes with the original nodes of the network they are replacing (Supplementary Figure S1B). The algorithm to compress nodes with same children is as follow:

<u>Pseudocode for compressing the network based on nodes sharing the same children</u>

Parameters:

- df: A network dataframe with three columns: source, target and sign
- upstream_input: The upstream inputs for the moon algorithm (upstream nodes)
- downstream_input: The downstream inputs for the moon algorithm (downstream nodes)

Algorithm compress_same_children(df, upstream_input, downstream_input):

1. Extract unique nodes from the source and target columns of `df` into `nodes`.

2. Identify nodes that are parents (i.e., they have children) and store them in `parents`.

3. Create `df_signature` from `df` where the target and sign are concatenated.

4. Generate `node_signatures` for each parent node in `parents` by concatenating "parent_of_" with the sorted list of children and their signs.

5. Identify duplicated node signatures that are not part of `upstream_input` or `downstream_input`. Store them in `dubs`.

6. Extract the parents corresponding to these duplicated signatures into `duplicated_parents`.

7. Update the source column of `df`:
a. Replace any node in the source that has a duplicated signature with the corresponding duplicated parent.

8. Update the target column of `df`:
a. Replace any node in the target that has a duplicated signature with the corresponding duplicated parent.

9. Remove duplicate rows from `df`.

10. Return a list containing:
- "compressed_network": the updated `df`
- "node_signatures": the `node_signatures` list
- "duplicated_signatures": the `duplicated_parents` list

End Algorithm

The algorithm to decompress the network is as follow, were SIF is a Simple Interaction Format network and ATT is a node Attribute data-frame:

Pseudocode for decompressing the solution network

Parameters:

- formatted_res: The result of the MOON scoring procedure.
- meta_network: The complete network data, a data frame representing a signed, directed network of molecular interactions.
- node_signatures: the output of the
- duplicated_parents:

Algorithm decompress_solution_network(formatted_res, meta_network, node_signatures, duplicated_parents):

1. Extract SIF and ATT tables from `formatted_res`.

2. Convert `duplicated_parents` to a data frame `duplicated_parents_df`.

3. Create a data frame `addons` from `node_signatures` excluding those already in `duplicated_parents_df`.

4. Combine `duplicated_parents_df` and `addons` into a single data frame `mapping_table`.

5. Merge `ATT` with `mapping_table` to update the node identifiers in ATT.

6. Use `meta_network` to create an updated SIF table.

7. Filter SIF to keep only the rows where both source and target nodes are present in the updated ATT.

8. Compute the 'Weight' column for the SIF table based on the activities of the source and target nodes.

9. Filter out rows in SIF where 'Weight' is zero.

10. Filter out rows in ATT where 'AvgAct' is zero.

11. Identify 'bad_seeds' - seeds that are not in ATT where 'NodeType' is "P".

12. Remove 'bad_seeds' from both ATT and SIF.

13. Create a new `formatted_res` containing the updated SIF and ATT tables.

14. Return `formatted_res`.

End Algorithm

Finally, to generate a solution network with similar properties as the output network of the classic CARNIVAL pipeline, the node MOON scores can be mapped on the PKN, and each interaction in the network that is not consistent with the sign of the MOON score of it's connecting node is removed from the PKN. It is also possible to define a score threshold to only keep paths connecting upstream and downstream input layer nodes that go through nodes with absolute scores that are strictly higher than the threshold. The algorithm to create a reduced solution network is a follow:

<u>Pseudocode for the reduce solution network function</u>

Parameters:

- moon_res: The result of the MOON scoring procedure.
- meta_network: The complete network data, a data frame representing a signed, directed network of molecular interactions.
- cutoff: an absolute value that will be used as a threshold to prune out any node of the meta_network based on the absolute MOON scores of each corresponding node.
- upstream_input: A boolean indicating whether to keep peers of the upstream node. Defaults to False.
-RNA_input: Optional - The RNA measurements that were used to compute the TF scores
-n_steps: the maximum number of steps allowed between a given pair of upstream and downstream inputs.

Algorithm reduce_solution_network(moon_res, meta_network, cutoff, upstream_input, RNA_input, n_steps):

1. Copy `moon_res` to `recursive_moon_res`.

2. Filter `recursive_moon_res` to include only rows where the absolute score is greater than `cutoff`.

3. Create a vector `consistency_vec` containing the scores from `recursive_moon_res`.

4. Create `res_network` by filtering `meta_network` to only include nodes present in `recursive_moon_res`.

5. Compute a consistency flag for each edge in `res_network` based on `consistency_vec`.

6. Filter `res_network` to include only edges that are consistent.

7. Create a data frame `upstream_input_df` from `upstream_input`.

8. Merge `upstream_input_df` with `recursive_moon_res` and compute a filter-out flag.

9. Extract `upstream_nodes` from `upstream_input_df` that are not filtered out and are present in `res_network`.

10. Reduce `res_network` to only include controllable neighbors within `n_steps` from `upstream_nodes` using COSMOS:::keep_controllable_neighbours().

11. Create SIF from the reduced `res_network`.

12. Create ATT by filtering `recursive_moon_res` to include only nodes present in SIF.

13. If `RNA_input` is not NULL, merge it with ATT.

14. Return a list containing SIF and ATT.

End Algorithm

For the MOON run, the prior-knowledge-network was processed slightly differently. Since the MOON scores are estimated compared to a background distribution, we did not exclude non-regulated TF, genes and metabolites from the network or from the inputs. The PKN was only filtered by removing any gene that wasn't part of the filtered Log2(FPKM+1) dataset, and we kept controllable and observable nodes that were within 6 steps of upstream and downstream inputs in the network. The network was then compressed using the procedure mentioned above.

We ran MOON first between receptors and downstream TF and metabolites, like the first CARNIVAL run. The resulting scores were mapped on the input network and an absolute score threshold of 1.5 (which represents a ULM t-value) was used to generate a reduced solution network comparable to the CARNIVAL solution network.

We then ran MOON a second time, between upstream TF and downstream ligands, using a limit on the number of steps of 1. This essentially allows us to

generate a network that only contains direct interactions between TF and target ligands where the TF activity score is coherent with the sign of the downstream ligand. An absolute score cutoff of 1.5 was also applied to remove non-significant TF-ligand interactions (at least according to this threshold).

The two resulting networks of mechanistic hypotheses resulting from the first and second MOON run were combined by taking the union of their edge and attribute lists into a single network connecting ligand, receptors, TF and metabolites altogether.

4.5 MOON cytokine scoring with cytosig data

4.5.1 Cytosig data collection

The sample-level preprocessed data was downloaded from the FDC platform (https://curate.ccr.cancer.gov/, accessed 22nd April 2022), using the code provided (https://github.com/data2intelligence/FDC_treatment_profile). For each treatment, we calculated a Z-score per gene as follows: the difference of the mean expression between treated and control samples was divided by a smoothed standard deviation of the gene expression of the control samples. The standard deviation was modeled using a locally-weighted regression (LOESS) based on the mean gene expression, using the scikit-misc python package. We kept only perturbations that had at least two matching control samples (as required for computing standard deviations).

4.5.2 Cytosig filtering

We filtered out genes that had no expression measured (NA values) in > 1000 cytosig comparisons. We renamed the cytokines of cytosig to their corresponding gene symbols, or HMDB identifiers for metabolites. Then, we filtered out cytokines that were not present in the COSMOS meta-prior knowledge network.

4.5.3 Cytokine scoring

For each cytosig gene expression signature (composed of cytosig z-scores), we filtered out any unmeasured genes (NA values), then we used decoupleR's ULM method with the Collectri TF-target network to estimate the T-values of linear models for each TFs fitting gene expression z-scores as a function of their mode of regulation (-1, 0 or 1), referred to as the TF score. Then, for each resulting TF score profile, we filtered out specifically the COSMOS prior knowledge network for genes that were not measured in the corresponding gene expression profile. We then filtered out any node of the COSMOS prior knowledge network that couldn't be reached within ten steps upstream of the TFs. Then the COSMOS

prior knowledge network was compressed as explained in (4.4.1). Moon scores were then computed using decoupleR's ULM method as explained in (4.4.1).

4.5.3 MOON scoring networks

To visualize and interpret each MOON score, we implemented a function to recover the sub-network of the prior knowledge network that contains the direct and indirect downstream targets of a given node that were used for its MOON score computation. The algorithm of the function is detailed as follow:

Pseudocode for the Get Moon Scoring Network Function

Parameters:

- upstream_node: The node from which the network analysis starts.
- meta_network: The complete network data, a data frame representing a signed, directed network of molecular interactions.
- moon_scores: A data frame with scores associated with each node in the network, including a 'source' column for node identifiers and a 'level' column for their score levels.
- keep_upstream_node_peers: A boolean indicating whether to keep peers of the upstream node. Defaults to False.

Algorithm get_moon_scoring_network(upstream_node, meta_network, moon_scores, keep_upstream_node_peers = False):

1. Determine the number of steps (`n_steps`) from `upstream_node` based on its moon score level in `moon_scores`.

2. If `keep_upstream_node_peers` is False, update `moon_scores` to exclude peers of the `upstream_node` at the same level.

3. Filter `meta_network` to keep only controllable neighbours of `upstream_node` using a predetermined `n_steps`. Store this in `meta_network_filtered`.

4. Identify downstream inputs by selecting nodes from `moon_scores` that have a level of 0 and are present in the `meta_network_filtered` as targets. Store these nodes in `downstream_inputs`.

5. Further filter `meta_network_filtered` to keep only observable neighbours based on `downstream_inputs` and `n_steps`.

6. Update `moon_scores` to include only nodes present in `meta_network_filtered`, both as sources and targets.

7. If `n_steps` is greater than 1 and `keep_upstream_node_peers` is False, perform a recursive filtering:
a. Initialize `remaining_level` with `n_steps`.
b. While `remaining_level` is greater than or equal to 0:
i. Identify `top_nodes` in `moon_scores` at `remaining_level`.
ii. Identify `child_nodes` in `meta_network_filtered` where the source is in `top_nodes`.
iii. Update `moon_scores` to include nodes in `child_nodes` or not at the immediate lower level.
iv. Update `meta_network_filtered` to include connections from `moon_scores`.
v. Decrement `remaining_level` by 1.

8. Return a list containing two elements: "SIF" with the filtered meta network (`meta_network_filtered`) and "ATT" with the updated moon scores (`moon_scores`).

End Algorithm

4.6 Pathway control analysis

To elucidate the potential pathways modulated by key regulatory nodes in our molecular network, we developed an algorithm termed "Find_Controlled_Pathways." This algorithm commences by identifying a set of "top nodes" from the combined results of the moon propagation algorithm (full_moon_res_combined). These top nodes are distinguished based on an absolute score threshold greater than 1.5 (this value can be set differently at the user's discretion). For each top node, the algorithm employs the COSMOS:::keep_controllable_neighbours function to ascertain downstream nodes within a maximum number interaction steps in the solution network (two steps in this study). If this filtered set of downstream nodes is not empty, an Over-Representation Analysis (ORA) is executed using the piano::runGSAhyper function, utilizing the NABA and KEGG pathway databases in this study. Other pathway set collections can be provided at the user's discretion. The resulting p-values, along with a computed log2 fold ratio, are associated with each node of interest and pathway, and these are stored in a data frame.

The algorithm aggregates the data across all nodes of interest into a single data frame (pathway_control_df), which is then reshaped to feature each pathway as a row and each node of interest as a column, with the corresponding p-values populating the cells. This comprehensive table serves as a pathway control matrix, enabling the identification of pathways significantly influenced by specific nodes within the regulatory network, and vice-versa. The algorithm for the pathway control analysis is as follow:

<u>Pseudocode to find controlled pathways</u>

Parameters:

- full_moon_res_combined: The result of the moon scoring procedure (or the combined result of multiple mon scoring procedures)
- combined_meta_network_translated: a data frame representing a signed, directed network of molecular interactions with three columns: source, target, sign.
- background_nodes: node IDs that will be used as a background for the over-representation analysis
- pathways: a pathway collection dataframe with two columns: gene and pathway

Algorithm Find_Controlled_Pathways(full_moon_res_combined, combined_meta_network_translated, background_nodes, pathways):

1. Initialize an empty list `pathway_control_set`.
2. Initialize counter `i` to 1.

3. Extract `top_nodes` from `full_moon_res_combined` where the absolute score is greater than 2.

4. For each `node_of_interest` in `top_nodes`:
a. Use `COSMOS:::keep_controllable_neighbours` to find `downstream_nodes` controlled by `node_of_interest` within 2 steps in `combined_meta_network_translated`.
b. If `downstream_nodes` is not empty:
i. Remove `node_of_interest` from `downstream_nodes`.
ii. Filter `downstream_nodes` to only include nodes that are in `background_nodes`.
iii. If the filtered `downstream_nodes` is not empty:
1. Run over-representation analysis (ORA) using `piano::runGSAhyper` on `downstream_nodes` with `background_nodes` as the universe and `pathways` as the gene set collection.
2. Extract results to `res_ORA`.
3. Add a `log2fold_ratio` column to `res_ORA`.
4. Add a `node_of_interest` column to `res_ORA`.
5. Add a `pathway` column to `res_ORA`.
6. Append `res_ORA` to `pathway_control_set`.
7. Increment `i` by 1.

5. Combine all data frames in `pathway_control_set` into a single data frame.

6. Reshape this combined data frame to `pathway_control_df` using `reshape2::dcast`, with pathways as rows and `node_of_interest` as columns, and values being p-values.

7. Set row names of `pathway_control_df` based on the `pathway` column.

8. Remove the first column of `pathway_control_df` containing the pathway names.

9. Return `pathway_control_df`.

End Algorithm

4.7 Cell line projection in factor space

To study how single samples behave in the factor space of MOFA, we reconstruct the RNA data of a given cell line with respect to each factor. To do so, we simply multiply the weights in a given factor with the corresponding value in the MOFA Z matrix for the corresponding cell line and the corresponding factor (Supplementary Figure S5). The algorithm describing this process is as follow:

Pseudocode to Decompose RNA Data for Specific Cell Line

Parameters:

- RNA_weight_cell_line: A data frame containing RNA weights and other information.
- Factor_MOFA: A data frame containing factors from MOFA specific to a cell line.

Algorithm: Decompose RNA Data for Specific Cell Line

1. Merge the RNA_weight_cell_line (the W matrix) and Factor_MOFA (the Z matrix) data frames based on the 'sample' column.
RNA_weight_cell_line <- merge(RNA_weight_cell_line, Factor_MOFA, by.x='sample', by.y='sample')

2. Initialize an empty data frame called cell_line_decomposed_RNA.

3. Loop over the range 1 to number of factors, for each factor:
a. Calculate the product of the RNA weight (column index x+y, x being the factor index and y a number corresponding to potential metadata columns in the dataframe) and the corresponding factor (column index x+y+number of factors) in RNA_weight_cell_line.
b. Store the computed product as a column in cell_line_decomposed_RNA.
cell_line_decomposed_RNA <- as.data.frame(do.call(cbind,lapply(1:9,function(x, RNA_weight_cell_line){
*RNA_weight_cell_line[,x+y] * RNA_weight_cell_line[,x+y+number of factors]*
},RNA_weight_cell_line = RNA_weight_cell_line)))

4. Set the row names of cell_line_decomposed_RNA to the 'feature' column from RNA_weight_cell_line.
row.names(cell_line_decomposed_RNA) <- RNA_weight_cell_line$feature

5. Rename the columns of cell_line_decomposed_RNA by replacing "V" with "Factor_".
names(cell_line_decomposed_RNA) <-
gsub("V","Factor_",names(cell_line_decomposed_RNA))

End

The data frame of decomposed RNA values for the cell line across factors can then be used seamlessly as an input for e.g. footprint analysis. I this case, we estimated TF activities and LR scores using the ULM function of decouleR, as described in Section 4.3.

5. Acknowledgements

We acknowledge funding to J.S.R. by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung BMBF) to support A.D. and D.T. (BMBF, [031L0181B]); by MSCoreSys research initiative research core SMART-CARE (031L0212A) to support A.D.; by Pfizer to support A.D. and R.F.; by HPC/Exascale Centre of Excellence for Personalised Medicine in Europe [PerMedCoE; European Union Horizon 2020 program, grant no. 951773] to support D.T.; D.M. acknowledges funding by the Spanish Government, which supports him through a predoctoral grant (PRE2020-092578 MCIN/AEI/10.13039/501100011033).

Thanks to Attila Gabor for his help maintaining the COSMOS package and setting up the documentation. Thanks to Ricardo Ramirez, Martín Garrido Rodríguez-Córdoba and Pablo Rodríguez Mier for their insightful advice on the manuscript.

6. Code and data availability

CosmosR package and NCI60 mofa-cosmos analysis markdowns can be found at https://saezlab.github.io/cosmosR/, while the source scripts and data can be found at https://github.com/saezlab/Factor_COSMOS. The cytosig benchmark analysis can be found at https://github.com/saezlab/MOON_benchmark. The python implementation of MOON can be found as part of a broader collection of tools at https://github.com/saezlab/networkcommons. The cytosig data was downloaded from the FDC platform (https://curate.ccr.cancer.gov/, accessed 22nd April 2022). The NCI raw data was accessed through NCI60 cellminer: transcriptomics (RNA-seq PMID:31113817), proteomics (SWATH (Mass

spectrometry) PMID:31733513} and metabolomics (LC/MS & GC/MS (Mass spectrometry) DTP NCI60 data).

7. Conflict of interests

JSR reports funding from GSK, Pfizer and Sanofi and fees from Travere Therapeutics, Stadapharm, Owkin and Astex. BS is employed by Pfizer.

8.Authors contributions

AD and JS-R designed the method. AD and PL coded the pipeline and ran the analysis, DM, VP and DT reimplemented parts of the pipeline. RF and ACK helped with the data analysis. BS and JS-R supervised the project. AD wrote the manuscript with help from PL and JS-R.

9. References

1. Chen, C. *et al.* Applications of multi-omics analysis in human diseases. *MedComm* **4**, e315 (2023).
2. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).
3. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform. Biol. Insights* **14**, 1177932219899051 (2020).
4. Conesa, A. & Beck, S. Making multi-omics data accessible to researchers. *Sci. Data* **6**, 251 (2019).
5. Yugi, K., Kubota, H., Hatano, A. & Kuroda, S. Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple "Omic" Layers. *Trends Biotechnol.* **34**, 276–290 (2016).
6. Terakawa, A. *et al.* Trans-omics analysis of insulin action reveals a cell growth subnetwork which co-regulates anabolic processes. *iScience* **25**, 104231 (2022).

7. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
8. Thurstone, L. L. Multiple factor analysis. *Psychol. Rev.* **38**, 406–427 (1931).
9. Velten, B. & Stegle, O. Principles and challenges of modeling temporal and spatial omics data. *Nat. Methods* **20**, 1462–1474 (2023).
10. Brown, B. C. *et al.* Multiset correlation and factor analysis enables exploration of multi-omics data. *Cell Genomics* **3**, 100359 (2023).
11. Velten, B. *et al.* Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nat. Methods* **19**, 179–186 (2022).
12. Frankhouser, D. E. *et al.* State-transition modeling of blood transcriptome predicts disease evolution and treatment response in chronic myeloid leukemia. *Leukemia* **38**, 769–780 (2024).
13. Yang, Z. & Michailidis, G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* **32**, 1–8 (2016).
14. Rau, A. *et al.* Individualized multi-omic pathway deviation scores using multiple factor analysis. *Biostatistics* **23**, 362–379 (2022).
15. Freeman-Cook, K. *et al.* Expanding control of the tumor cell cycle with a CDK2/4/6 inhibitor. *Cancer Cell* **39**, 1404-1421.e11 (2021).
16. Argelaguet, R. *et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
17. Quintero, A. *et al.* ShinyButchR: Interactive NMF-based decomposition workflow of genome-scale datasets. *Biol. Methods Protoc.* **5**, bpaa022 (2020).
18. Su, G., Burant, C. F., Beecher, C. W., Athey, B. D. & Meng, F. Integrated metabolome and transcriptome analysis of the NCI60 dataset. *BMC*

Bioinformatics **12 Suppl 1**, S36 (2011).

19. Wang, H., Wang, X., Xu, L., Cao, H. & Zhang, J. Nonnegative matrix factorization-based bioinformatics analysis reveals that TPX2 and SELENBP1 are two predictors of the inner sub-consensuses of lung adenocarcinoma. *Cancer Med.* **10**, 9058–9077 (2021).
20. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
21. Knowles, D. A., Bouchard, G. & Plevritis, S. Sparse discriminative latent characteristics for predicting cancer drug sensitivity from genomic features. *PLoS Comput. Biol.* **15**, e1006743 (2019).
22. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
23. Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491 (2019).
24. Consiglio, C. R. *et al.* The Immunology of Multisystem Inflammatory Syndrome in Children with COVID-19. *Cell* **183**, 968-981.e7 (2020).
25. Gonçalves, E. *et al.* Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell* **40**, 835-849.e8 (2022).
26. Schlechte, J. *et al.* Dysbiosis of a microbiota-immune metasystem in critical illness is associated with nosocomial infections. *Nat. Med.* **29**, 1017–1027 (2023).
27. Li, L. *et al.* Multi-omics profiling of collagen-induced arthritis mouse model reveals early metabolic dysregulation via SIRT1 axis. *Sci. Rep.* **12**, 11830 (2022).
28. Monaco, G. *et al.* Transcriptome Analysis Identifies Accumulation of Natural Killer Cells with Enhanced Lymphotoxin-β Expression during Glioblastoma

Progression. *Cancers (Basel)* **14**, (2022).

29. Hamsanathan, S. *et al.* Integrated -omics approach reveals persistent DNA damage rewires lipid metabolism and histone hyperacetylation via MYS-1/Tip60. *Sci. Adv.* **8**, eabl6083 (2022).
30. Park, J.-C. *et al.* Multi-Omics-Based Autophagy-Related Untypical Subtypes in Patients with Cerebral Amyloid Pathology. *Adv Sci (Weinh)* **9**, e2201212 (2022).
31. Kwok, A. J. *et al.* Neutrophils and emergency granulopoiesis drive immune suppression and an extreme response endotype during sepsis. *Nat. Immunol.* **24**, 767–779 (2023).
32. Gambacorta, V. *et al.* Integrated multiomic profiling identifies the epigenetic regulator PRC2 as a therapeutic target to counteract leukemia immune escape and relapse. *Cancer Discov.* **12**, 1449–1461 (2022).
33. Mangiante, L. *et al.* Multiomic analysis of malignant pleural mesothelioma identifies molecular axes and specialized tumor profiles driving intertumor heterogeneity. *Nat. Genet.* **55**, 607–618 (2023).
34. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell-cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22**, 71–88 (2021).
35. Dimitrov, D. *et al.* LIANA+: an all-in-one cell-cell communication framework. *BioRxiv* (2023) doi:10.1101/2023.08.19.553863.
36. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
37. Szalai, B. & Saez-Rodriguez, J. Why do pathway methods work better than

they should? *FEBS Lett.* **594**, 4189–4200 (2020).

38. Schäfer, P. S. L., Dimitrov, D., Villablanca, E. J. & Saez-Rodriguez, J. Integrating single-cell multi-omics and prior biological knowledge for a functional characterization of the immune system. *Nat. Immunol.* **25**, 405–417 (2024).
39. Dugourd, A. & Saez-Rodriguez, J. Footprint-based functional analysis of multiomic data. *Current Opinion in Systems Biology* **15**, 82–90 (2019).
40. Badia-I-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances* **2**, vbac016 (2022).
41. Liu, A. *et al.* From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. *NPJ Syst. Biol. Appl.* **5**, 40 (2019).
42. Dugourd, A. *et al.* Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. *Mol. Syst. Biol.* **17**, e9730 (2021).
43. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43–49 (2013).
44. Bradley, G. & Barrett, S. J. CausalR: extracting mechanistic sense from genome scale data. *Bioinformatics* **33**, 3670–3672 (2017).
45. Chowdhury, S. *et al.* DAGBagM: learning directed acyclic graphs of mixed variables with an application to identify protein biomarkers for treatment response in ovarian cancer. *BMC Bioinformatics* **23**, 321 (2022).
46. Massacci, G. *et al.* A key role of the WEE1-CDK1 axis in mediating TKI-therapy resistance in FLT3-ITD positive acute myeloid leukemia patients. *Leukemia* **37**, 288–297 (2023).
47. Rosenberger, G. *et al.* Network-based elucidation of colon cancer drug

resistance by phosphoproteomic time-series analysis. *BioRxiv* (2023) doi:10.1101/2023.02.15.528736.

48. Garrido-Rodriguez, M., Zirngibl, K., Ivanova, O., Lobentanzer, S. & Saez-Rodriguez, J. Integrating knowledge and omics to decipher mechanisms via large-scale models of signaling networks. *Mol. Syst. Biol.* **18**, e11036 (2022).
49. Gjerga, E., Dugourd, A., Tobalina, L., Sousa, A. & Saez-Rodriguez, J. PHONEMeS: Efficient Modeling of Signaling Networks Derived from Large-Scale Mass Spectrometry Data. *J. Proteome Res.* **20**, 2138–2144 (2021).
50. Mueller-Dott, S. *et al.* Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *BioRxiv* (2023) doi:10.1101/2023.03.30.534849.
51. Jiang, P. *et al.* Systematic investigation of cytokine signaling activity at the tissue and single-cell levels. *Nat. Methods* **18**, 1181–1191 (2021).
52. Szklarczyk, D. *et al.* STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* **44**, D380-4 (2016).
53. Türei, D. *et al.* Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol. Syst. Biol.* **17**, (2021).
54. Brunk, E. *et al.* Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat. Biotechnol.* **36**, 272–281 (2018).
55. Saez-Rodriguez, J. *et al.* Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Mol. Syst. Biol.* **5**, 331 (2009).
56. Wu, X., Jiang, R., Zhang, M. Q. & Li, S. Network-based global inference of human disease genes. *Mol. Syst. Biol.* **4**, 189 (2008).
57. Leiserson, M. D. M. *et al.* Pan-cancer network analysis identifies

combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).

58. Mosly, D. *et al.* Variation in IL6ST cytokine family function and the potential of IL6 trans-signalling in ERα positive breast cancer cells. *Cell. Signal.* **103**, 110563 (2023).
59. Rolling, C., Treton, D., Pellegrini, S., Galanaud, P. & Richard, Y. IL4 and IL13 receptors share the gamma c chain and activate STAT6, STAT3 and STAT5 proteins in normal human B cells. *FEBS Lett.* **393**, 53–56 (1996).
60. Gholami, A. M. *et al.* Global proteome analysis of the NCI-60 cell line panel. *Cell Rep.* **4**, 609–620 (2013).
61. Camacho, D., de la Fuente, A. & Mendes, P. The origin of correlations in metabolomics data. *Metabolomics* **1**, 53–63 (2005).
62. Saccenti, E., Suarez-Diez, M., Luchinat, C., Santucci, C. & Tenori, L. Probabilistic networks of blood metabolites in healthy subjects as indicators of latent cardiovascular risk. *J. Proteome Res.* **14**, 1101–1111 (2015).
63. Steuer, R. Review: on the analysis and interpretation of correlations in metabolomic data. *Brief. Bioinformatics* **7**, 151–158 (2006).
64. Shankavaram, U. T. *et al.* Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Mol. Cancer Ther.* **6**, 820–832 (2007).
65. Gry, M. *et al.* Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* **10**, 365 (2009).
66. Fillebeen, C., Caltagirone, A., Martelli, A., Moulis, J.-M. & Pantopoulos, K. IRP1 Ser-711 is a phosphorylation site, critical for regulation of RNA-binding and aconitase activities. *Biochem. J.* **388**, 143–150 (2005).
67. Pitula, J. S. *et al.* Selective inhibition of the citrate-to-isocitrate reaction of

cytosolic aconitase by phosphomimetic mutation of serine-711. *Proc Natl Acad Sci USA* **101**, 10907–10912 (2004).

68. Pekayvaz, K. *et al.* Multi-Omic Factor Analysis uncovers immunological signatures with pathophysiologic and clinical implications in coronary syndromes. *medRxiv* (2023) doi:10.1101/2023.05.02.23289392.
69. Meng, C. *et al.* MOGSA: Integrative Single Sample Gene-set Analysis of Multiple Omics Data. *Mol. Cell. Proteomics* **18**, S153–S168 (2019).
70. Lipton, Z. C. The mythos of model interpretability. *Queue* **16**, 31–57 (2018).
71. Lou, Y., Caruana, R. & Gehrke, J. Intelligible models for classification and regression. in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12* 150 (ACM Press, 2012). doi:10.1145/2339530.2339556.
72. Merico, D., Gfeller, D. & Bader, G. D. How to visually interpret biological data using networks. *Nat. Biotechnol.* **27**, 921–924 (2009).
73. Borgatti, S. P. Centrality and network flow. *Soc. Networks* **27**, 55–71 (2005).
74. Henry, N. & Fekete, J.-D. MatrixExplorer: a dual-representation system to explore social networks. *IEEE Trans. Vis. Comput. Graph.* **12**, 677–684 (2006).
75. Bae, J. & Watson, B. Developing and evaluating Quilts for the depiction of large layered graphs. *IEEE Trans. Vis. Comput. Graph.* **17**, 2268–2275 (2011).
76. Ahlswede, R., Cai, N., Li, S. Y. R. & Yeung, R. W. Network information flow. *IEEE Trans Inf Theory* **46**, 1204–1216 (2000).
77. Hou, L., Chen, M., Zhang, C. K., Cho, J. & Zhao, H. Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies. *Hum. Mol. Genet.* **23**, 2780–2790 (2014).
78. Odom, G. J. *et al.* PathwayPCA: an R/Bioconductor Package for Pathway Based Integrative Analysis of Multi-Omics Data. *Proteomics* **20**, e1900409

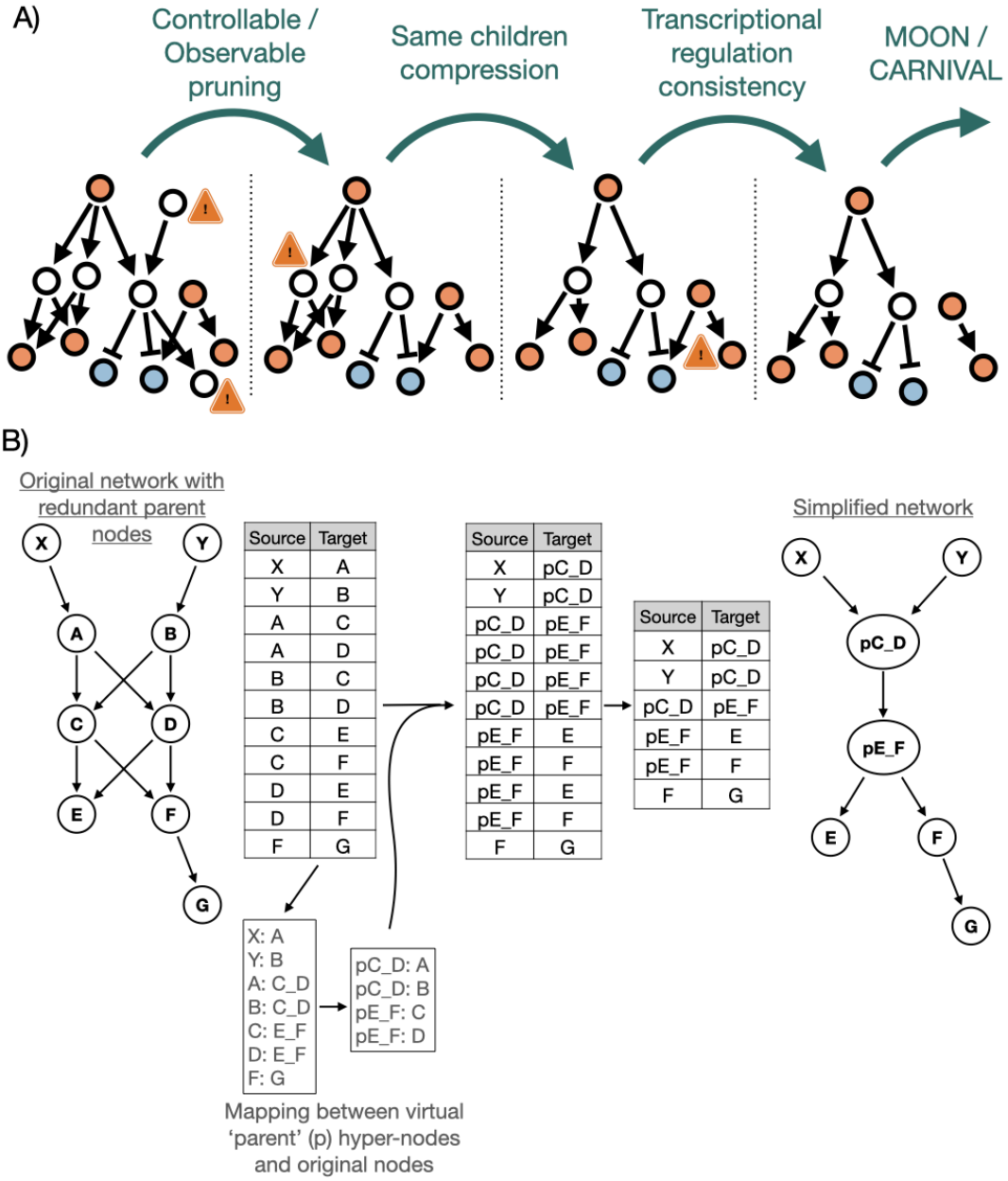
(2020).

79. Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C. & Stegle, O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.* **18**, 212 (2017).
80. Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nat. Protoc.* **17**, 1518–1552 (2022).
81. Wilkerson, M. D. *et al.* Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic Acids Res.* **42**, e107 (2014).
82. Cappelletti, V. *et al.* Dynamic 3D proteomes reveal protein functional alterations at high resolution in situ. *Cell* **184**, 545-559.e22 (2021).
83. Mateus, A. *et al.* Thermal proteome profiling for interrogating protein interactions. *Mol. Syst. Biol.* **16**, e9232 (2020).
84. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
85. Varland, S. *et al.* N-terminal acetylation shields proteins from degradation and promotes age-dependent motility and longevity. *Nat. Commun.* **14**, 6774 (2023).
86. Sun, M. & Zhang, X. Current methodologies in protein ubiquitination characterization: from ubiquitinated protein to ubiquitin chain architecture. *Cell Biosci.* **12**, 126 (2022).
87. Monteleone, G. *et al.* Blocking Smad7 restores TGF-beta1 signaling in chronic inflammatory bowel disease. *J. Clin. Invest.* **108**, 601–609 (2001).
88. Yan, X., Liu, Z. & Chen, Y. Regulation of TGF-beta signaling by Smad7. *Acta Biochim Biophys Sin (Shanghai)* **41**, 263–272 (2009).
89. Gambardella, J. *et al.* Exploiting GRK2 Inhibition as a Therapeutic Option in Experimental Cancer Treatment: Role of p53-Induced Mitochondrial

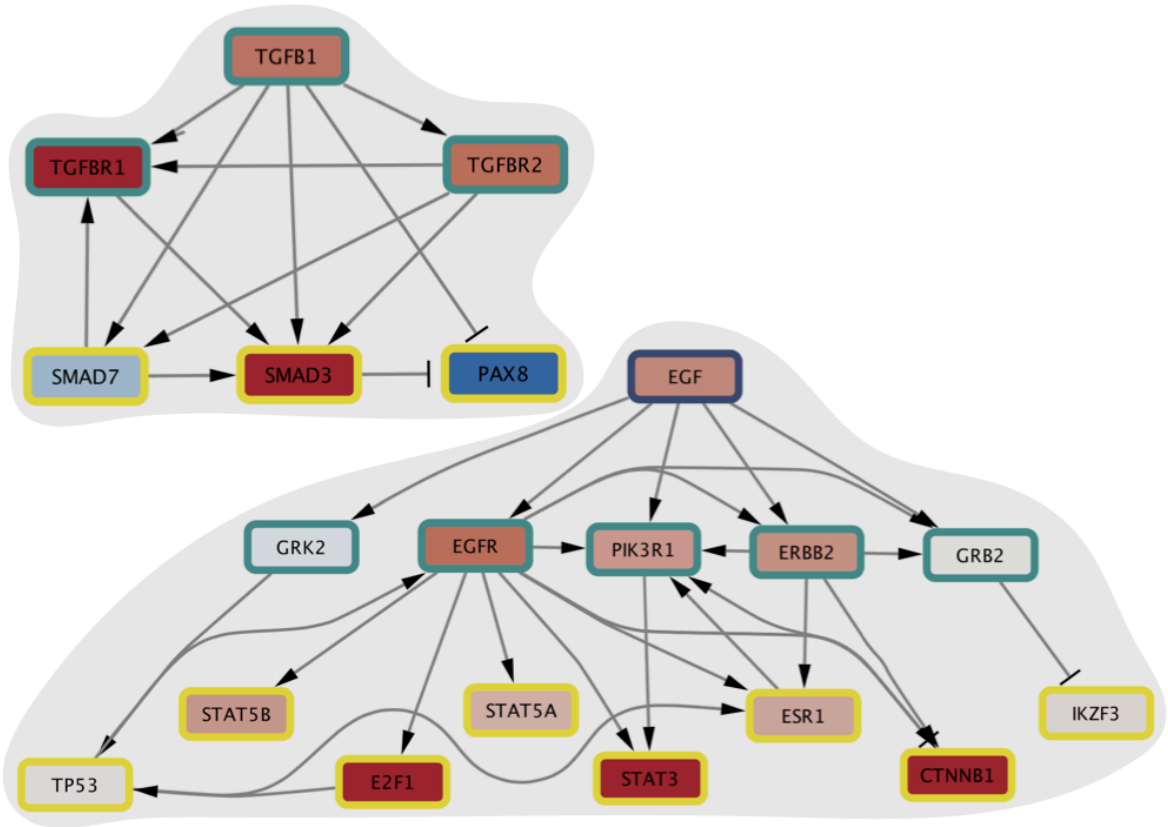
Apoptosis. *Cancers (Basel)* **12**, (2020).

90. Chen, Y., Long, H., Wu, Z., Jiang, X. & Ma, L. EGF transregulates opioid receptors through EGFR-mediated GRK2 phosphorylation and activation. *Mol. Biol. Cell* **19**, 2973–2983 (2008).
91. Yamazaki, T. *et al.* Role of Grb2 in EGF-stimulated EGFR internalization. *J. Cell Sci.* **115**, 1791–1802 (2002).

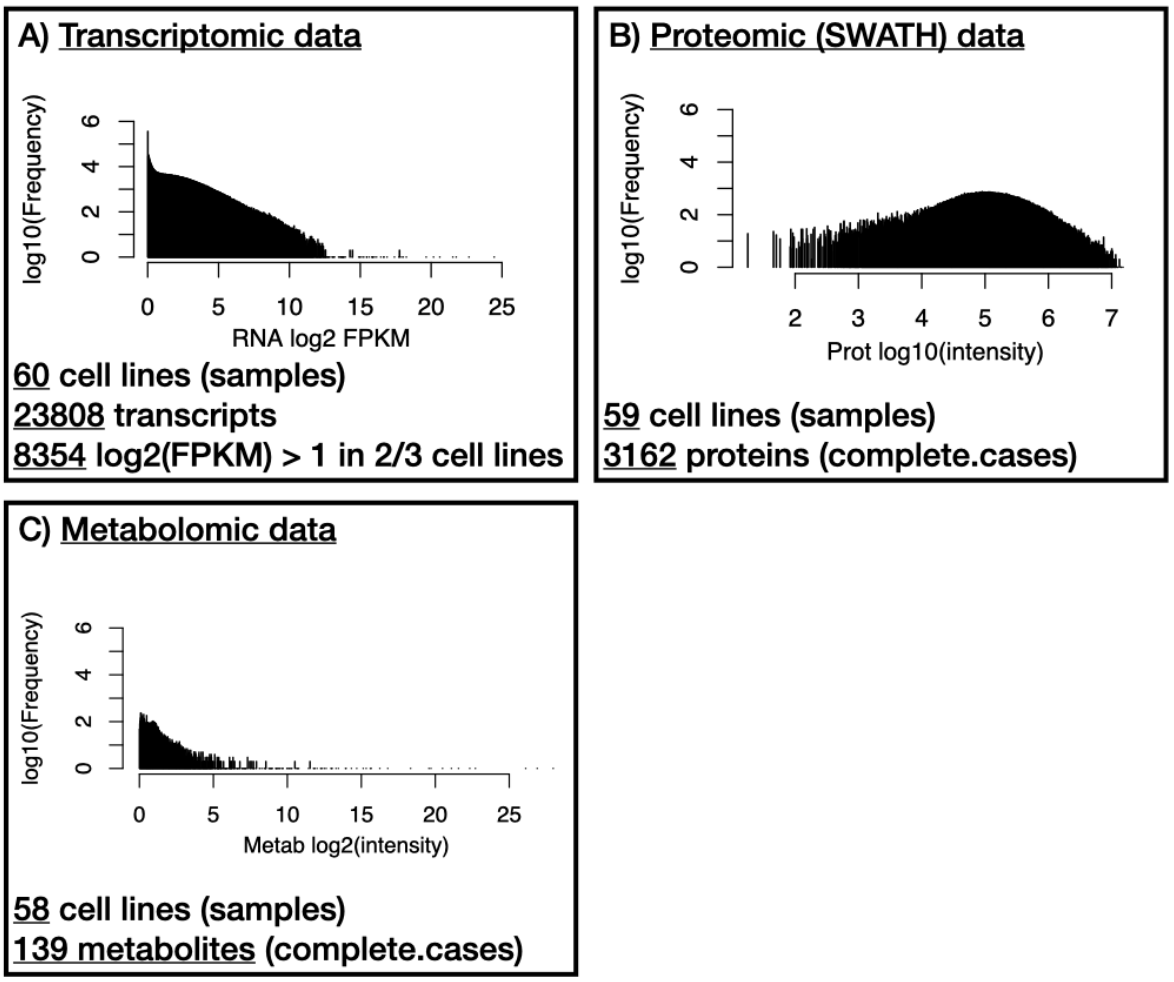
Supplementary Materials



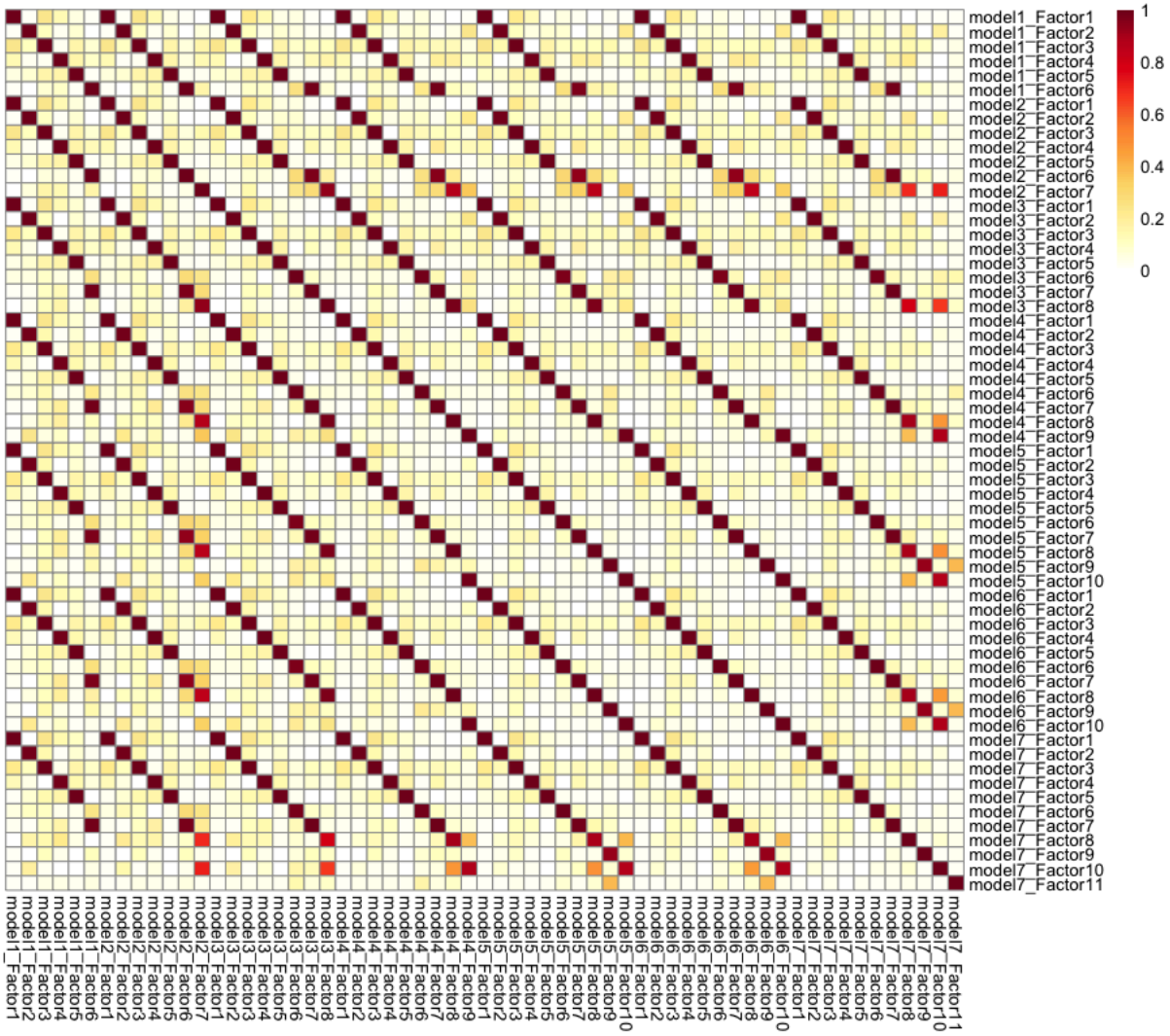
Supplementary Figure S1 *- A) Representation of the prior knowledge pre-processing steps. First, nodes that cannot be reached downstream of the upstream input (controllable) and nodes that cannot be reached upstream of downstream inputs (observables) are pruned out of the network. Then, any nodes that have strictly the same set of direct children nodes are compressed into virtual nodes, preventing redundant paths in the network. Next, a transcriptional consistency check is performed, where any node that is regulated directly by a transcription factor is pruned out if its transcriptional regulation direction doesn't match the activity of said transcription factor. If proteomic data is available, both the transcriptional and proteomic regulation direction have to be consistent. The transcriptional consistency check is also performed after the network scoring, and the network scoring procedure is repeated until the transcriptional consistency check doesn't find any inconsistency anymore. B) Schematic representation of the network compression algorithm.*



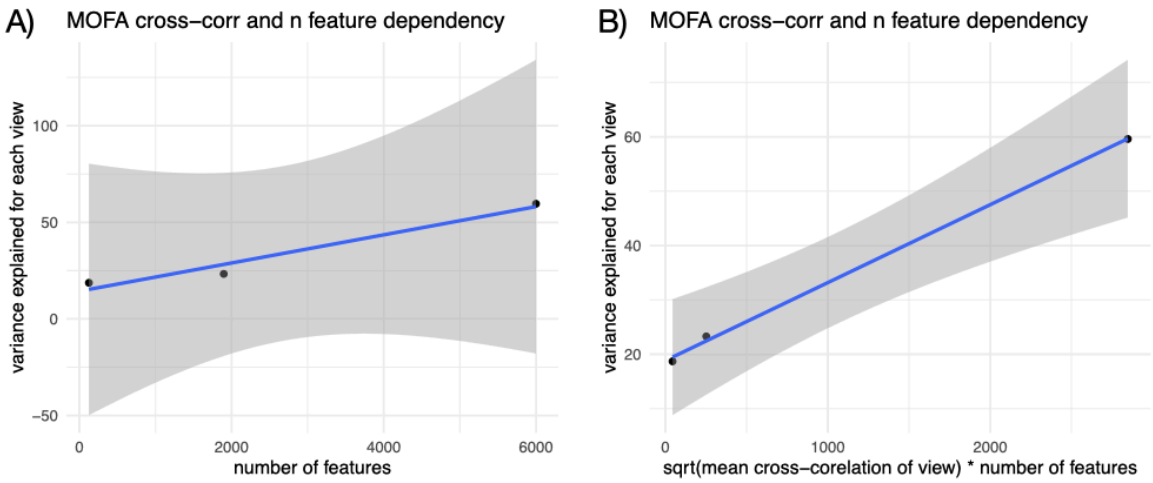
Supplementary Figure S2 - TGFB1 also had relatively heterogeneous MOON scores across experiments, despite having a positive score on average (TGFB1: mean = 0.57, SD = 0.8).TGFB1 and EGF are examples of ligands with more complex MOON score estimation. Indeed, with respect to MOON heuristic, TGFB1 score depends on the activity estimation of SMAD3, SMAD7 and PAX8. While TGFB1 is relatively well scored (69th average quantile in experiments where it's applied and 0.5 in experiments where it's not), SMAD7 activity is systematically inconsistent with respect to the activity of SMAD3 and PAX8. Indeed, SMAD3 is expected to be activated by TGFB1 while PAX8 is expected to be inhibited, and their activity estimation is consistent with this in the example experiment of Figure 2B. However, while SMAD7 should seemingly be activated by TGFB1 according to the prior knowledge network of Omnipath, its activity isn't increased when TGFB1 is applied. According to scientific literature [87,88], SMAD7 is in fact known to be an inhibitor of TGFB1 signaling via negative feedback mechanisms. EGF MOON score, on the other hand, appears to be depending on the activity of STAT5A, STAT5B, E2F1, STAT3, ESR1, CTNNB1 as well as TP53 and IKZF3 (Figure 2B). While STAT5A, STAT5B, E2F1, STAT3, ESR1, CTNNB1 all appear to be coherently deregulated, TP53 and IKZF3 do not seem to follow that trend. According to the prior knowledge network, TP53 and IKZF3 are regulated through GRK2 and GRB2 respectively. The EGF-GRK2-TP53 is documented in the scientific literature [89,90], albeit suggesting that GRK2-TP53 interaction should be an inhibition rather than an activation. EGF-GRB2 is also supported by scientific literature [91], but the GRB2-IKZF3 does not appear to have direct and targeted documented evidence in the scientific literature.



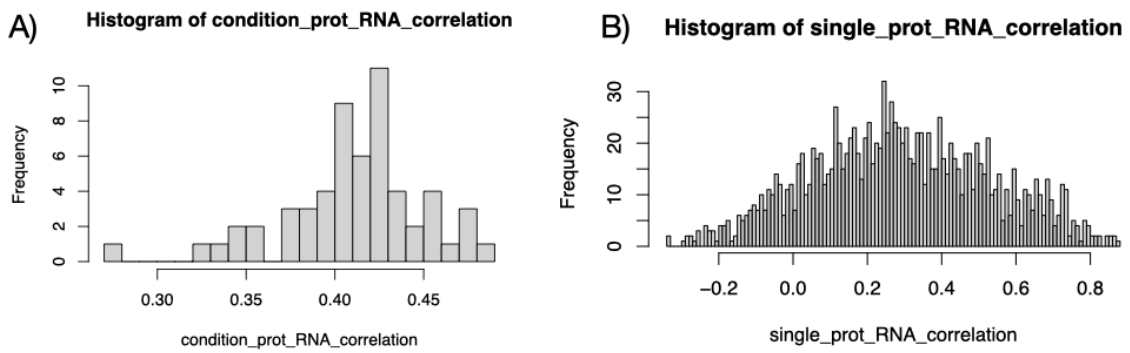
Supplementary Figure S3 *- NCI60 multi-omic data and MOFA. A) Histogram of all the log2 transformed FPKM+1 values of the NCI60 transcriptomic datasets. 60 cell lines have transcriptomic data in single replicates. 23808 transcript have annotated HUGO gene symbols, of which 8354 have log2(FPKM+1) values > 1 in at least ⅔ of the samples. log2(FPKM+1) values < 1 were excluded from the rest of the analysis (see "Prepare_RNA.R"). B) Histogram of all the log10 transformed intensities of the NCI60 proteomic dataset. 59 cell lines have proteomic data in single replicates. 3162 proteins are consistently detected in every cell line (complete cases). Data was obtained directly from cellminer and no further preprocessing was considered necessary (see "Prepare_proteomic.R")). C) Histogram of all the log2 transformed intensity values of the NCI60 metabolomic datasets. 58 cell lines have metabolomic data in triplicates. To remain consistent with the other omic data, triplicates were simply averaged into single cell-line samples. A few extreme outliers (log2(intensities) > 32) were removed. 125 metabolites are consistently detected in every cell line (complete cases) (see "Prepare_metabolomic.R").*



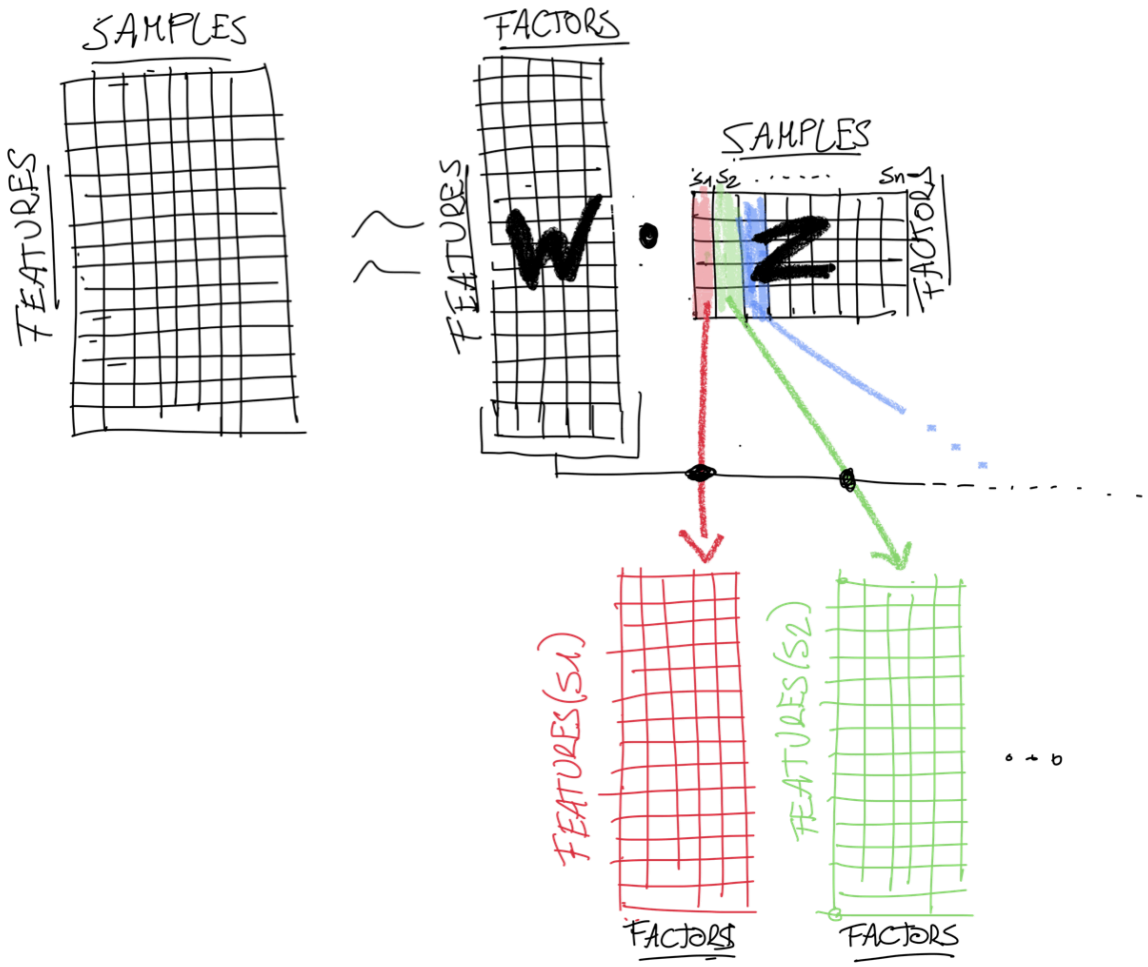
Supplementary Figure S4 *- Correlation heatmap between the factor weights of mofa models with different maximum number of factors (7 to 13 maximum factors).*



Supplementary Figure S5 *- Linear regression between the variance explained by the 9 MOFa factor for each omic view and A) the number of features of each view and B) the number of features of each view multiplied by the square root of the average cross correlation values of each view.*



Supplementary Figure S6 *- A) histogram of correlation coefficients for each cell-line across genes between RNA and protein value. B) histogram of correlation coefficients for each gene across cell-lines between RNA and protein value*



Supplementary Figure S7 *- Schematic representation of the sample matrix reconstruction in factor space.*