PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT

DE L'UNIVERSITÉ PSL

Préparée à École Normale Supérieure-PSL

Infer molecular mechanisms at cellular and multicellular scales from single-cell multi-omics data

Soutenue par

Rémi TRIMBOUR

Le 15 septembre 2025

École doctorale nº 515

Complexité du Vivant

Spécialité

Biologie des Systèmes et Bioinformatique

Composition du jury :

Nathalie VIALANEIX Directrice de recherche, INRAE	*Rapporteuse*
Vera PANCALDI Directrice de recherche, CRCT	*Rapporteuse*
Laurence CALZONE Ingénieure de recherche, Institut Curie	*Examinatrice*
Carl HERRMANN Professeur, IPMB Heidelberg	*Examinateur*
Laura CANTINI Chargée de recherche, Institut Pasteur	*Directrice de thèse*
Julio SAEZ-RODRIGUEZ Professeur, Université de Heidelberg	*Directeur de thèse*

Présidente du Jury : Nathalie VIALANEIX

1794 ENS | PSL

RÉSUMÉ

Les organismes multicellulaires reposent sur une coordination précise entre de nombreux types cellulaires. L'identité d'une cellule émerge de l'interaction de plusieurs niveaux de régulation moléculaire. Les progrès récents des technologies de séquençage permettent désormais de profiler plusieurs modalités omiques . Différentes méthodes utiisent ces information pour inférer les interactions moléculaires, mais même les approches intégratives combinant plusieurs modalités ne couvrent pas toutes ces interactions biologiques. Dans cette thèse, je présente deux méthodes basées sur les réseaux multilayer hétérogènes (HMLN), HuMMuS et ReCoN, pour étudier les systèmes cellulaires et multicellulaires. Cette structure conserve l'information spécifique à chaque modalité dans des couches distinctes reliées par des liens inter-couches. Elle peut donc modéliser de nombreuses interactions, englobant régulations intracellulaires et communication entre cellules, avant de les explorer grâce à un algorithme de marche aléatoire. HuMMus et ReCoN peuvent notamment identifier les régulateurs d'un groupe de gènes, et les acteurs qui coordonnent divers types cellulaires lors de réponses complexes.

MOTS CLÉS

Cellules, analyses de données, machine learning, réseaux, génétique, transcriptomique, multi-omique, communication cellulaire.

ABSTRACT

Multicellular organisms require precise coordination across many cell types. A cell's identity arises from the interplay of multiple molecular layers that collectively determine its state and function. Recent advances in single-cell sequencing enable the profiling of these "omics" modalities at the single-cell resolution. Yet transforming these multimodal datasets into mechanistic insight remains a challenge. Even integrative approaches that combine several modalities still ignore the full spectrum of available data. In this thesis, I introduce two flexible frameworks - HuMMuS and ReCoN - to study cellular and multicellular systems, respectively, using heterogeneous multilayer networks (HMLNs). This structure preserves modality-specific information in distinct layers, connected through inter-layer edges. This paradigm enables reconstruction of comprehensive regulatory maps spanning intracellular regulation and cell communication, and exploration via random walk with restarts adapted to different regulatory hypotheses. It can notably identify regulators of specific genes or drivers coordinating different cell types in complex responses.

KEYWORDS

Single-cell, data analysis, networks, multilayers,transcriptomics, multi-omics, gene regulation, cell communication.

Remerciements

Avec une entrée par les études de médecine, ma formation en bioinformatique n'est pas des plus classiques. Merci à toutes les personnes qui m'ont fait confiance malgré ce parcours atypique et m'ont tant appris sur la recherche. Merci à Valérie Lamour et Alain Bessis d'avoir vu ma passion pour la recherche et pour leur implication dans le développement des parcours médecine sciences. Merci à Kirsley pour son rôle dans mon envie d'apprendre la bioinformatique. Merci à Denis, Laura et Julio de m'avoir donné l'opportunité de travailler dans des environnements de recherche particulièrement stimulants. J'aimerais aussi remercier Laurence tout particulièrement, pour son soutien permanent, psychologique et scientifique, dans mes projets et ma thèse. J'ai été particulièrement heureux que tu acceptes de participer à mon comité de suivi, puis à mon jury de thèse et que tu aies pu être présente pour cette conclusion.

Je remercie les membres de mon jury d'avoir accepté d'évaluer mon travail. Encore merci à Laurence, déjà citée plus haut. Merci à Véra pour ses commentaires détaillés et ses questions concernant l'analyse de données génétiques et épigénétiques, auxquelles je réfléchis encore aujourd'hui. Merci à Carl pour ses retours sur les évolutions possibles des réseaux de régulation des gènes, j'espère avoir la chance d'échanger plus à ce sujet dans les prochaines années. Merci à Nathalie pour ces retours et son implication tout au long de cette fin de thèse. Avec du recul, tous les problèmes logistiques que nous avons eus le jour de la soutenance rendent cette journée encore plus mémorable.

À tous mes formidables amis et amies de Paris et de Pasteur, merci pour vos conseils ou simplement votre présence. J'ai hâte d'avoir enfin plus de temps pour prendre soin de vous à mon tour. En particulier mes collègues, Claire, Geert, Jules, Clément, Anna, Daniele, Jérémie et Anthony, vous m'avez beaucoup aidé tout au long de cette thèse. Nos pauses déjeuner/Pedantle resteront un souvenir mémorable, même si vous n'aviez pas vraiment le niveau pour rivaliser - c'est la vérité, c'est écrit par un docteur :).

Thank you to all my wonderful friends from Heidelberg, in particular Francesco, Lorna, Robin, Jovan, Denes, Pau, Sophia, Barbara, Nico, Aurélien, Charlotte, Miguel, Rico, and the other fantastic people I met in Saezlab. You created an amazing working environment at the lab, and I enjoyed every moment we spent together, whether it was over a beer or at a lab meeting.

Merci à ma famille, Papa, Maman et Chloé d'avoir cru en moi pour cette longue aventure. (J'en suis à la moitié, plus que 8 ans d'études !). Vous m'avez donné l'envie d'apprendre, et l'espace pour la développer depuis tout petit.

Merci à Jérémie pour cette amitié constante depuis tant d'années. Nos discussions et tes conseils ont joué un rôle majeur sur ma vie professionnelle et personnelle. Nos soirées jeux étaient aussi la meilleure façon de relâcher la pression de la thèse.

Enfin, merci à Zuzia, pour son immense soutien durant cette fin de thèse. Sans ta direction artistique, tes nombreuses relectures et corrections, cette thèse n'aurait pas été la même. Tu me rappelles chaque jour que cette thèse est un début avant d'être une fin. *Looking forward to seeing what kind of adventure is waiting for us in this New World.* :)

Cette thèse est née du désir de m'investir pleinement dans la recherche médicale, au-delà de son application dans le contexte du soin. J'écris aussi ce court texte pour que sa relecture puisse me rappeler la raison de cet investissement. J'espère pouvoir contribuer à une recherche utile, qui pourra améliorer la santé de tous de manière tangible, sans se faire au détriment du reste du vivant.

Rémi

Glossary

Data and measurement techniques

UMI: Unique Molecular Identifier

scRNA-seq: Single-cell RNA sequencing

scATAC-seq: Single-cell Assay for Transposase-Accessible Chromatin Sequencing

snmC: Single-nucleus methylcytosine sequencing

PCHiC: Promoter-Capture Hi-C

MERFISH: Multiplexed Error-Robust Fluorescence in situ Hybridization

seqFISH: Sequential Fluorescence in situ Hybridization

10x Visium: Slide-based spatial transcriptomics platform

Slide-seq: Bead-based spatial transcriptomics technology

Biological networks

TF: Transcription factor

Peak: Short DNA region defined in scATAC-seq analysis

PPI: Protein-protein interaction

GRN: Gene regulatory network

CCANs: Cis-Coaccessible Networks

CCC: Cell-cell communication

PKN: Prior-knowledge network

RWR: Random walk with restart

SNF: Similarity Network Fusion

MLN: Multilayer network

HMLN: Heterogeneous multilayer network

Data analysis

DE: Differentially expressed (genes)

LSI: Latent Semantic Indexing

kNN: k-nearest neighbors

UMAP: Uniform Manifold Approximation and Projection

t-SNE: t-distributed Stochastic Neighbor Embedding

Biology

HF: Heart failure

NHF: Non-heart failure

ECM: Extracellular matrix

BMMC: Bone-Marrow Mononuclear Cells

PBMC: Peripheral-Blood Mononuclear Cells

Method names

HuMMuS: HeterogeneoUs Multilayer networks for MUlti-omics Single-cell data

ReCoN: Regulatory mechanisms and cell communication network inference

CIRCE: CIs-Regulatory Co-accessiblE network inference

Louvain: Louvain community-detection algorithm

Leiden: Leiden community-detection algorithm

MultiXrank: Random-walk toolkit for universal multilayer networks

SCENIC+: Single-cell regulatory network inference pipeline

Pando: Regulatory network inference method (bulk & single-cell)

CellOracle: GRN inference tool using prior promoter-enhancer knowledge

GENIE3: Gene Network Inference with Ensemble of Trees

Extended summary

Context and contribution of this thesis

The molecular identity of a cell emerges from complex interactions between various molecular regulatory layers. Recent advances in single-cell sequencing technologies now allow measuring these regulatory layers — also called omics — for individual cells. Additionally, the spatial positioning of cells within tissue sections can now be accessed.

Network inference represents a powerful reverse-engineering method to uncover regulatory mechanisms from omics data (e.g., RNA-seq)[1–4]. However, current approaches applied to scRNA-seq data demonstrate limited performance[5,6]. More recently, several methods have been developed to combine different molecular measurements, such as scRNA-seq and scATAC-seq, for inferring interactions between transcription factors (TFs) and genes[7–9]. These methods still face several limitations, notably ignoring other types of single-cell data currently available.

In this thesis, I aimed to leverage complementary information provided by multiple omics to enhance network inference. A significant methodological innovation introduced here is the use of heterogeneous multilayer networks (HMLNs) to integrate multiple omics within a single network, while preserving their independent information.

This thesis revolved around two main objectives:

A - Inference of intracellular mechanisms from multi-omic data.
I proposed a novel method, rigorously evaluated against state-of-the-art methods, and applied to a complex three-modality multi-omic dataset. This method is based on several network layers, each containing interactions specific to one omic type (e.g., genes for scRNA, DNA regions for scATAC). This first step resulted in the development of HuMMuS, now published[10].

B - Inference of multicellular networks.
I subsequently integrated intra- and intercellular interactions via a HMLN, enabling detailed analysis of co-regulation across different cellular environments. I am currently developing ReCoN, a tool for inferring these multicellular molecular mechanisms.

1. Predicting intracellular interactions from multi-omic data

Heterogeneous multilayer networks for multi-omic data integration

A HMLN is a network $M = (V_m, E_m, L),\ m = 1,\ldots, M$ composed of M layers, each containing nodes Vm and intra-layer links $E_m \subseteq V_m \times V_m$. Inter-layer links are defined by L[11]. I selected this structure to integrate molecular regulations within a cell, with each layer representing a biological macromolecule type (e.g., genes, proteins) connected by homogeneous

interactions, and inter-layer links representing interactions between different macromolecular types.

HuMMuS: a tool to reconstruct molecular mechanisms from single-cell multi-omics data

Based on this definition, I developed HuMMuS (HeterogeneoUs Multilayer networks for MUlti-omics Single-cell data), a novel tool to infer regulatory networks from single-cell multi-omic data. This development constitutes the thesis's first achievement, proposing a HMLN-based methodology that integrates additional omics. HuMMuS is open-source and was published in Bioinformatics[10].

As illustrated in Figure 0.1, HuMMuS reconstructs a HMLN comprising three layers: (1) the TF layer containing protein-protein interactions (PPIs) among TFs, (2) the DNA region layer defined by co-accessibility inferred from scATAC-seq data, and (3) the gene layer established from scRNA-seq data.

Exploration of the full network then relies on a random walk with restart (RWR), parameterized according to the desired type of regulation (e.g., TF → target gene, enhancer → target gene).

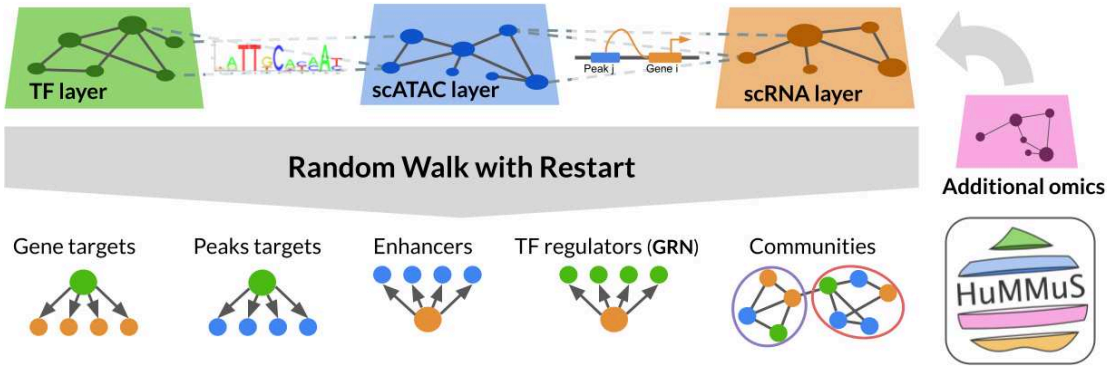


Figure 0.1. Schematic view of HuMMuS workflow. 1) Each layer is coming from a different omics. **2)** Random walks allow to reconstruct different types of outputs.

HuMMuS overcomes the limitations of previous approaches

Thanks to the HMLN structure, HuMMuS presents several advantages over existing methods. It captures both inter-omic (TF-DNA region, DNA region-gene) and intra-omic (TF-TF, DNA region-DNA region, gene-gene) interactions, enabling the identification of additional macromolecule cooperations. For example, TF → target gene interactions can be identified even without knowing all DNA motifs recognized by a TF. Unlike state-of-the-art methods, HuMMuS is also extensible to other omics types, facilitating inference of regulatory mechanisms for additional macromolecule types.

HuMMuS shows superior performance compared to state-of-the-art methods

I compared HuMMuS against several benchmark methods for regulatory network inference, demonstrating superior performance in TF-gene, TF-regulatory region, and enhancer predictions. Following this benchmark, I applied HuMMuS to single-cell murine cortex data (scRNA, scATAC, scmC). This analysis identified relevant TFs regulating distinct neuronal subpopulations.

2. Modeling intercellular interactions from multi-omic data

Multicellular coordination

Numerous pathologies involve multicellular programs, namely coordinated responses between cell types[12–15]. The coordination of these programs relies on molecular entities transmitting necessary information. Recently, computational methods have emerged to study these programs, either within patient cohorts or within tissue samples to understand spatial heterogeneity[15–18]. However, these approaches typically do not explore the underlying molecular mechanisms ensuring coordination.

Our objective here is to identify molecular regulators of these multicellular programs, which could constitute potential targets for precision medicine.

Intercellular communication as a major limitation

Cell-cell communication is currently a central research area in computational biology and a significant challenge for modeling multicellular interactions[19]. Correlations between biological markers can now be quantified across several scales and spatial resolutions (intracellular, juxtacellular, paracrine)[20]. However, these analyses alone cannot formulate mechanistic hypotheses about observed regulations.

ReCoN: extracting molecular mechanisms in multicellular systems

Over recent months, I developed ReCoN (Regulatory mechanism and Cell Communication Network), a tool integrating intracellular mechanisms and intercellular interactions to study multicellular program coordination.

ReCoN models molecular cooperations within and between cells through a HMLN, conferring various advantages. It can predict interactions or regulations absent from classical databases (pathways, receptor targets). Through RWR exploration, ReCoN simultaneously explores molecular regulators (causes) and perturbation consequences.

ReCoN identifies cytokine-induced transcriptomic programs in murine lymph nodes

I evaluated ReCoN's performance for predicting cytokine-induced transcriptomic effects in murine lymph nodes in vivo[21]. Predictions were compared against another method, NicheNet[22], adapted for our case and various partial HMLNs. We quantified the relative contribution of the gene regulatory layer (intracellular cooperations) and the cell communication layer. Both types of information significantly contributed to predicting

cytokine-induced transcriptomic effects, with the complete ReCoN outperforming the adapted NicheNet.

ReCoN predicts transcriptomic changes associated with heart failure from potential key regulators

ReCoN also models complex molecular perturbations. I applied it to study coordinated multicellular responses observed in heart failure, starting from previously identified key regulators[23]. ReCoN again surpassed predictions from the adapted NicheNet.

We observed that cell communication is particularly informative for predicting specific cell type programs, suggesting differential sensitivity to intercellular signaling according to cellular context.

ReCoN identifies the causes and consequences of cardiac fibrosis from involved genes

Cardiac fibrosis is an adaptive phenomenon following trauma such as cardiac arrest or failure, involving fibroblast differentiation and extracellular matrix (ECM) production. ReCoN helped identify intracellular receptors and TFs regulating ECM protein genes within fibroblasts, and intercellular interactions that revealed upstream and downstream cardiac remodeling functions, specific or shared among cell types.

Résumé étendu

Contexte et contribution de cette thèse

L'identité moléculaire d'une cellule résulte des interactions complexes entre différentes couches de régulation moléculaire. Les avancées récentes en technologies de séquençage à cellule unique permettent désormais de mesurer ces couches de régulation — appelées omiques — pour chaque cellule individuellement. Par ailleurs, il est désormais possible d'accéder à la position spatiale des cellules dans des coupes de tissus.

L'inférence de réseaux constitue une méthode de rétro-ingénierie particulièrement efficace pour élucider les mécanismes de régulation à partir de données omiques (par exemple, des données RNA-seq)[1–4]. Toutefois, les approches actuelles appliquées aux données scRNA-seq présentent des performances limitées[5,6]. Récemment, plusieurs méthodes ont été développées pour combiner différentes mesures moléculaires, notamment scRNA-seq et scATAC-seq, afin d'inférer les interactions entre facteurs de transcription (FT) et gènes[7–9]. Ces méthodes présentent encore plusieurs limites, en particulier le fait d'ignorer d'autres types de données single-cell désormais disponibles.

Au cours de cette thèse, j'ai cherché à exploiter l'information complémentaire apportée par différentes omiques afin d'améliorer l'inférence de réseaux. Une innovation méthodologique majeure de cette thèse est l'utilisation de réseaux multi-couches hétérogènes (HMLN) pour intégrer plusieurs omiques dans un seul réseau tout en conservant l'indépendance de leurs informations.

Cette thèse s'est articulée autour de deux objectifs principaux :

A - Inférence des mécanismes intracellulaires à partir de données multi-omiques.
J'ai proposé une méthode novatrice que j'ai rigoureusement évaluée face aux méthodes de référence, et appliquée à un jeu de données multi-omiques complexe à trois modalités. Cette méthode s'appuie sur des couches de réseaux contenant chacune des interactions spécifiques à une omique donnée (gènes pour scRNA, régions d'ADN pour scATAC, etc.). Cette première étape a abouti au développement de HuMMuS, aujourd'hui publié.

B - Inférence de réseaux multicellulaires
J'ai ensuite intégré interactions intra- et intercellulaires par un réseau HMLN, permettant une analyse fine des co-régulations dans différents environnements cellulaires. Je développe actuellement ReCoN, un outil pour inférer ces mécanismes moléculaires multicellulaires.

1. Prédire les interactions intracellulaires à partir de données multi-omiques

Les réseaux multi-couches hétérogènes pour intégrer des données multi-omique

Un HMLN est un réseau $M = (V_m, E_m, L)$, $m = 1,\ldots,M$ composé de M couches, chacune contenant des nœuds V_m et des liens intra-couche $E_m \subseteq V_m \times V_m$. Les liens inter-couches sont définis par L[11]. J'ai choisi cette structure pour intégrer les régulations moléculaires d'une cellule, chaque couche représentant un type de macromolécule biologique (par ex. gènes, protéines) reliée par des interactions homogènes, et les liens inter-couches représentant les interactions entre différents types de macromolécules.

HuMMuS: un outil pour reconstruire les mécanismes moléculaire à partir de données single-cell multi-omiques

A partir de cette définition, j'ai développé HuMMuS (HeterogeneoUs Multilayer networks for MUlti-omics Single-cell data), un nouvel outil permettant d'inférer les régulations à partir de données multi-omiques en cellule unique. Ce développement constitue le premier aboutissement de la thèse, proposant une méthodologie basée sur les HMLN intégrant des omiques additionnelles. HuMMuS est disponible et open-source, et a été publié dans Bioinformatics[10].

Comme illustré en Figure 0.1, HuMMuS reconstruit un HMLN composé de trois couches : (1) la couche des FT, qui contient les interactions protéine-protéine (PPIs) entre FTs, (2) la couche des régions de l'ADN, définit par les co-accessibilités des régions d'ADN inférées à partir de données scATAC-seq, (3) la couche des gènes, établit à partir de données de scRNA-seq.

L'exploration du réseau complet repose ensuite sur une marche aléatoire avec redémarrage (RWR), paramétrée selon le type de régulation recherché (e.g., FT → gène cible, enhancer → gène cible).

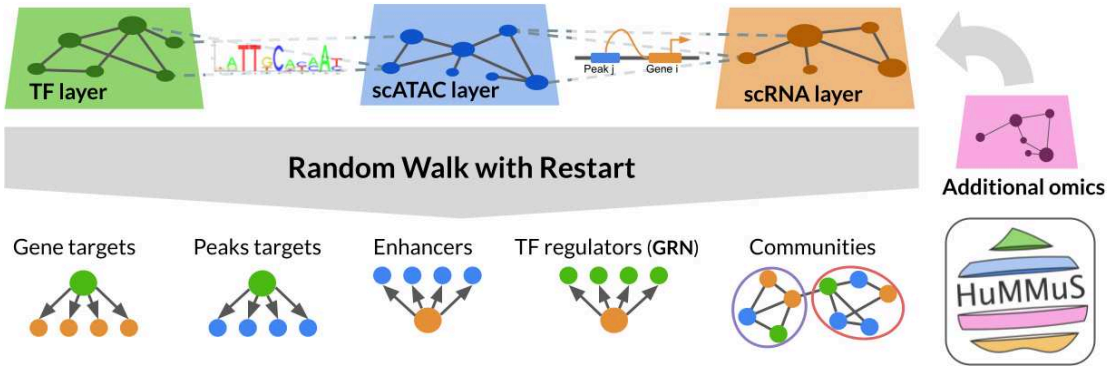


Figure 0.1. Représentation schématique du workflow de HuMMuS 1) Chaque couche est construite à partir d' omiques différentes. **2)** Des marches aléatoires permettent de reconstruire différents résultats en fonction des questions d'intérêt.

HuMMuS dépasse les limites des approches précédentes

Grâce à la structure HMLN, HuMMuS présente plusieurs avantages face aux méthodes déjà existantes. Cette méthode capture à la fois les interactions inter-omiques (FT-région d'ADN,

région d'ADN-gène) et intra-omiques (FT-FT, région d'ADN - région d'ADN, gène-gène). Cela permet de capturer des coopérations supplémentaires entre macromolécules, pour retrouver, par exemple, des interactions TF → gène cible même sans connaître tous les motifs d'ADN reconnu par un FT. Contrairement à l'état de l'art, HuMMuS est aussi extensible à d'autres types d'omiques, permettant donc d'inférer les mécanismes de régulation d'autres types de macromolécules.

HuMMuS affiche de meilleures performances que les méthodes de l'état de l'art

J'ai comparé HuMMuS à plusieurs méthodes de référence pour l'inférence de réseaux de régulation, montrant de meilleures performances pour les prédictions FT-gène, FT-région régulatrice et enhancer. A la suite de ce benchmark, j'ai également appliqué HuMMuS à des données single-cell du cortex murin (scRNA, scATAC, scmC). Cette analyse a permis d'identifier des FT pertinents régulant différentes sous-populations de neurones.

2. Modéliser les interactions intercellulaires à partir de données multi-omiques

Coordination multicellulaire

De nombreuses pathologies impliquent des programmes multicellulaires, c'est-à-dire des réponses coordonnées entre types cellulaires[12–15]. La coordination de ces programmes repose sur des entités moléculaires qui véhiculent l'information nécessaire. Récemment, des méthodes computationnelles ont émergé pour étudier ces programmes, au sein de cohortes de patients ou à l'intérieur de prélèvement tissulaire pour en comprendre l'hétérogénéité spatiale[15–18]. Cependant, ces approches ne s'intéressent généralement pas aux mécanismes moléculaires sous-jacents qui assurent cette coordination.

Notre objectif est ici d'identifier les régulateurs moléculaires de ces programmes multicellulaires, qui pourraient constituer des cibles potentielles pour la médecine de précision.

La communication intercellulaire comme limitation majeure

La communication entre cellules constitue aujourd'hui un axe de recherche central en biologie computationnelle, et un défi majeur pour la modélisation des interactions multicellulaires[19]. Il est désormais possible de quantifier les corrélations entre différents marqueurs biologiques à plusieurs échelles et résolutions spatiales (intracellulaire, juxtacellulaire, paracrine)[20]. Cependant, ces analyses seules ne permettent pas de formuler des hypothèses mécanistiques sur les régulations observées.

ReCoN : extraire les mécanismes moléculaires dans les systèmes multicellulaires

Au cours des derniers mois, j'ai développé ReCoN (Regulatory mechanism and Cell Communication Network), un outil qui intègre à la fois les mécanismes intracellulaires et les interactions intercellulaires, pour étudier la coordination des programmes multicellulaires.

ReCoN modélise les coopérations moléculaires au sein des cellules et entre cellules par un réseau HMLN, ce qui lui confère différents avantages. Cela lui permet de prédire des interactions ou régulations absentes des bases de données classiques (pathways, cibles de récepteurs). Grâce à l'exploration par RWR, ReCoN permet d'explorer à la fois les régulateurs moléculaires (cause) et les conséquences d'une perturbation.

ReCoN identifie les différents programmes transcriptomiques induits par différentes cytokines dans les ganglions lymphatiques murins

J'ai évalué les performances de ReCoN pour retrouver les effets transcriptomiques induits par des cytokines dans des ganglions lymphatiques murins in vivo[21]. Les prédictions ont été comparées à celles obtenues à partir d'une autre méthode, NicheNet[22], adaptée pour notre cas précis, et différents HMLN ne prenant en compte que certaines des informations disponibles. Cela nous a permis de quantifier la contribution relative de la couche de régulations des gènes (coopérations intracellulaires) et de la couche de communication cellulaire. Nous montrons que ces deux types d'information apportent une contribution significative à la prédiction des effets transcriptomiques des cytokines. Alors que les versions ne prenant pas en compte les interactions intercellulaires performant moins bien que la version adaptée de Nichenet, la version complète de ReCoN dépasse ses performances. Nous avons pu évaluer ces performances pour chaque type cellulaire individuellement, et pour l'ensemble du ganglion en même temps.

ReCoN prédit les altérations transcriptomiques associées à l'insuffisance cardiaque à partir de régulateurs clés potentiels

ReCoN est également capable de modéliser l'impact de perturbations moléculaires plus complexes. Je l'ai appliqué à l'étude des réponses multicellulaires coordonnées observées dans l'insuffisance cardiaque, en partant de régulateurs clés identifiés dans[23]. Là encore, ReCoN a surpassé les prédictions issues de la version adaptée de NicheNet.

Nous avons observé que la communication cellulaire est particulièrement informative pour prédire certains programmes spécifiques à certains types cellulaires, ce qui suggère une sensibilité différenciée à la signalisation intercellulaire selon le contexte cellulaire.

ReCoN retrouve des causes et conséquences de la fibrose cardiaque à partir d'une liste de gènes impliquées

La fibrose cardiaque est un phénomène adaptatif du tissu cardiaque à différents traumatismes comme un arrêt ou une insuffisance cardiaque, ou des fibroblastes vont se différencier et produire une importante quantité de matrice extracellulaire (MEC). Elle s'intègre à une reconstruction plus générale du tissu, ou vaisseaux sanguins, muscles et cellules immunitaires s'adaptent également. Nous avons appliqué ReCoN pour comprendre la régulation de la fibrose à deux niveaux. Premièrement, ReCoN a permis de retrouver des récepteurs et FTs qui, à l'intérieur des fibroblastes semblent réguler l'expression de gènes de protéines composant la MEC. En explorant par la suite les

interactions entre cellules, ReCoN a pu identifier d'autres fonctions du remodèlement cardiaque en amont et en aval de la fibrose cardiaque. Ces différentes fonctions peuvent être spécifiques à une type cellulaire (e.g., l'hypoxie dans les cellules endothéliales) ou partagée entre plusieurs (e.g., une transition épithélio-mésenchymateuse partagée entre toutes les cellules).

ReCoN peut donc aussi permettre d'identifier des molécules responsables d'un certain phénotype ou d'une maladie. Ces molécules peuvent être intracellulaires ou dans les cellules environnantes, offrant donc d'autant plus d'interventions potentielles à travers des mécanismes d'action divers.

Short summary

Multicellular organisms require precise coordination across many cell types. A cell's identity arises from the interplay of multiple molecular layers that collectively determine its state and function. Recent advances in single-cell sequencing allow profiling of these "omics" modalities at single-cell resolution. Yet transforming these multimodal datasets into mechanistic insight remains a challenge. Even integrative approaches that combine several modalities still ignore the full spectrum of available data.

In this thesis, I introduce two flexible frameworks to study cellular and multicellular systems using heterogeneous multilayer networks (HMLNs). This structure preserves modality-specific information in distinct layers, connected through inter-layer edges. This paradigm enables reconstruction of comprehensive regulatory maps spanning intracellular regulation and cell communication, and exploration via random-walk algorithms adapted to different regulatory hypotheses. It can notably identify regulators of specific genes or drivers coordinating different cell types in complex responses.

This work unfolds in two main parts:

Intracellular network inference from single-cell multi-omics
I developed *HuMMuS* (HeterogeneoUs Multilayer networks for MUlti-omics Single-cell data), an open-source tool that builds a three-layer HMLN comprising: 1) a transcription factor (TF) layer with known protein–protein interactions between TFs, 2) a chromatin layer—co-accessibility of DNA regions from single-cell ATAC-seq, 3) a gene expression layer—co-regulated genes from scRNA-seq.
Inter-layer edges connect TFs to their binding sites and regulatory regions to target genes. Running a random walk with restart on this network allows HuMMuS to prioritize TF→gene, enhancer→gene, or other regulatory links. Benchmarking against state-of-the-art methods showed that HuMMuS outperforms existing approaches in recovering known TF–gene interactions, TF–regulatory region contacts, and enhancer–target predictions. Applied to a three-modality dataset from the murine cortex (scRNA, scATAC, and single-cell methylation), HuMMuS identified TFs governing specific neuronal subpopulations and revealed novel cooperations undetectable by two-modal methods.

Modeling multicellular coordination via intercellular HMLNs
Many developmental and pathological programs emerge from coordinated actions across cell types, mediated by secreted factors and receptor signaling. To capture these phenomena, I extended HuMMuS's idea to include cell communication, forming a unified network of intra- and intercellular interactions. The resulting tool, *ReCoN* (Regulatory mechanism and Cell Communication Network), integrates cell type–specific networks with ligand–receptor interactions, enabling inference of multicellular regulatory mechanisms.

We applied ReCoN in several biological contexts. In murine lymph nodes, it successfully predicted cytokine-induced transcriptomic programs by jointly leveraging intracellular

regulatory maps and intercellular signaling, outperforming adaptations of other tools. In a heart-failure model, ReCoN identified key regulators driving maladaptive multicellular responses and dissected the contributions of gene regulation versus intercellular signaling. In a study of cardiac fibrosis, ReCoN recovered both intra-fibroblast TFs regulating extracellular-matrix genes and upstream interactions orchestrating fibroblast activation and tissue remodeling.

Together, *HuMMuS* and *ReCoN* propose a new approach to model molecular mechanisms from single-cell multi-omic data. By encoding each modality in an individual layer, this approach preserves distinct molecular interactions while enabling cross-modal discoveries. As single-cell technologies continue to diversify, this framework can readily incorporate proteomics, metabolomics, offering an interesting milestone for future integrative methods.

Court résumé

Les organismes multicellulaires reposent sur une coordination précise entre de nombreux types cellulaires. L'identité d'une cellule émerge de l'interaction de plusieurs niveaux de régulation moléculaire. Les progrès récents des technologies de séquençage permettent désormais de profiler plusieurs modalités « omiques ». Différentes méthodes utiisent ces information pour inférer les interactions moléculaires, mais même les approches intégratives combinant plusieurs modalités ne couvrent pas toutes ces interactions biologiques.

Dans cette thèse, je présente deux méthodes basées sur les réseaux multilayer hétérogènes (HMLN)**,** *HuMMuS* et *ReCoN*, pour étudier les systèmes cellulaires et multicellulaires. Cette structure conserve l'information spécifique à chaque modalité dans des couches distinctes reliées par des liens inter-couches. Elle peut donc modéliser de nombreuses interactions, englobant régulations intracellulaires et communication entre cellules, avant de les explorer grâce à un algorithme de marche aléatoire. HuMMus et ReCoN peuvent notamment identifier les régulateurs d'un groupe de gènes, et les acteurs qui coordonnent divers types cellulaires lors de réponses complexes.

1. Inférence de réseaux intracellulaires à partir de données multi-omiques

J'ai développé *HuMMuS* (HeterogeneoUs Multilayer networks for MUlti-omics Single-cell data), un outil open source qui construit un HMLN à trois couches : 1) la couche des facteurs de transcriptions (FT), qui contient les interactions protéine-protéine (PPIs) entre FTs, 2) la couche des régions de l'ADN, définit par les co-accessibilités des régions d'ADN inférées à partir de données scATAC-seq, 3) la couche des gènes, établit à partir de données de scRNA-seq.

Les arêtes inter-couches relient les TF à leurs sites de liaison et les régions régulatrices à leurs gènes cibles. Une marche aléatoire avec redémarrage (RWR) permet alors à HuMMuS de prioriser les liens TF→gène, enhanceur→gène ou autres interactions. Les benchmarks montrent qu'il surpasse les méthodes de référence pour retrouver les interactions TF–gène, les contacts TF–région régulatrice et les prédictions enhanceur–cible. Sur un jeu de données triplement modal du cortex murin (scRNA, scATAC, méthylation unicellulaire), HuMMuS a révélé les TF contrôlant des sous-populations neuronales spécifiques et mis en évidence des coopérations inédites qu'un modèle bimodal ne détecte pas.

2. Modélisation de la coordination multicellulaire

De nombreux processus développementaux ou pathologiques résultent d'actions concertées entre types cellulaires, médiées par des facteurs sécrétés et le signalement récepteur-ligand. Pour capturer ces phénomènes, j'ai étendu *HuMMuS* afin d'y intégrer la communication cellulaire, formant un large réseau d'interactions intra- et intercellulaires. ReCoN (Regulatory mechanism and Cell Communication Network) permet l'inférence de

mécanismes régulatoires multicellulaires des interactions ligand-récepteurs et des régulations intracellulaires.

ReCoN a été appliqué à divers contextes biologiques. Dans les ganglions lymphatiques murins, il a prédit avec succès les programmes transcriptomiques induits par les cytokines, surpassant d'autres outils. Dans un modèle d'insuffisance cardiaque, *ReCoN* a identifié les régulateurs clés des réponses multicellulaires délétères et disséqué la part relative de la régulation génique et du signalement intercellulaire. Dans une étude sur la fibrose cardiaque, il a mis en évidence à la fois les TF intracellulaires des fibroblastes contrôlant la matrice extracellulaire et les interactions en amont orchestrant l'activation des fibroblastes et le remodelage tissulaire.

HuMMuS et *ReCoN* proposent une nouvelle approche pour modéliser les mécanismes moléculaires à partir de données multi-omiques. Ces méthodologies peuvent facilement être étendues à d'autres omiques et type de macromolécules.

Table of Contents

Chapter 1

Introduction

Cells and organs rely on a complex integration of diverse molecular signals across scales. Recent technological advances now allow us to profile molecular features at the single-cell level, capturing heterogeneity that was previously masked by bulk measurements. However, the countless and essential relationships between these molecular features are not directly accessible; they must be inferred through computational analysis and integration of prior knowledge. The methods to infer these interactions extend progressively, following the improvements in measurement and single-cell technologies. This thesis addresses the methodological and conceptual challenges of inferring molecular mechanisms from single-cell multi-omics data, spanning cellular to multicellular scales.

Content

1.1. Single cell measurements

Single-cell measurements provide an unprecedented view into the molecular heterogeneity of biological systems. Unlike bulk profiling, which averages signals across populations of cells, single-cell techniques allow us to dissect the diversity of cell states, transitions, and regulatory mechanisms within complex tissues. These methods are the foundation for modern computational and systems biology approaches to cellular identity and behavior. In this section, we present an overview of the major single-cell modalities and the associated data analysis techniques.

1.1.1. Single cell sequencing technologies

The exponential interest in single-cell biology arose from the development of high-throughput single-cell sequencing technologies. These platforms measure on a cell-by-cell basis various molecular modalities, such as gene expression, chromatin accessibility, DNA methylation, and protein abundance. The technologies differ in sensitivity, throughput, resolution, and the biological insight they offer. We begin by reviewing these different methods and their specific contributions to our understanding of cellular systems.

1.1.1.1. Single cell transcriptomic

The transcriptome refers to the complete set of RNA transcripts (e.g., messenger RNA, non-coding RNAs) produced in a cell, tissue, or organism at a specific time. It represents all the genes that are actively transcribed into RNA and provides a snapshot of gene expression. The most well-known part of the transcriptome corresponds to the mRNA, as the coding sequences for protein synthesis.

Single-cell RNA sequencing (scRNA-seq) is a technique that analyzes gene expression at the level of individual cells. The process begins by isolating single cells from a tissue sample, often using methods such as droplet encapsulation[24,25] or fluorescence-activated cell sorting (FACS)[26,27]. Each cell's mRNA is then captured, converted into complementary DNA (cDNA), and tagged with unique cell barcodes and unique molecular identifiers (UMIs) to distinguish individual cells and quantify molecular amplification biases[28]. The resulting cDNAs are amplified and prepared for sequencing. High-throughput sequencing platforms read the fragments, producing data that includes gene sequences, cell barcodes, and UMIs. Multiple protocols exist, such as Smart-seq2[29] for full-length transcripts, and 10x Genomics Chromium for high-throughput, 3'-end biased libraries[30]. The sequencing reads are then processed through a pipeline such as CellRanger[24], to identify their mapping gene on the genome and to which cell they belong. The final result is a count matrix m x n, with m is the number of individual mRNA molecules mapped, and n the number of individual cells.

1.1.1.2. Single cell chromatin accessibility

Chromatin accessibility reflects the accessibility of genomic regions to transcription factor binding and other regulatory processes, serving as a proxy for their gene regulatory potential.

Single-cell transposase-accessible chromatin sequencing-seq (scATAC-seq) is a technique that measures chromatin accessibility at the single-cell level. Building upon ATAC-seq technologies[31], it involves isolating individual cells with a similar method as in scRNA-seq[32], applying a transposase enzyme that inserts sequencing adapters into accessible chromatin regions, and then sequencing the extracted fragments. The sequencing reads are aligned to a reference genome, and each read is assigned to its respective cell based on unique cell barcodes. Rather than building the cell-by-region matrix immediately, the aligned reads are aggregated across cells to identify regions with enriched transposase activity and higher number of reads[33,34]. These regions, or peaks, represent candidate regulatory elements important for genes' transcription. Once a set of consensus peaks is established, a cell-by-peak matrix is generated by counting for every cell the number of reads mapping to each region.

1.1.1.3. Single cell methylation

DNA methylation is an epigenetic modification characterized by the addition of a methyl group, typically on the 5′ carbon of cytosine residues and predominantly within CpG dinucleotides. This modification plays a critical role in regulating gene expression, genomic imprinting, and maintaining genome stability.

Single-cell DNA methylation techniques, such as single-cell bisulfite sequencing (scBS-seq)[35] or snmC-seq2[36] enable the profiling of DNA methylation patterns at single-cell resolution. The process begins again with the isolation of individual cells, a technological improvement from bulk methods such as BS-seq. The cells' genomes are then extracted and treated with bisulfite to convert unmethylated cytosines into uracils, while leaving methylated cytosines unchanged. The bisulfite-converted DNA is amplified, incorporating cell-specific barcodes to preserve the cellular origin of each read. Reads are then aggregated across defined genomic regions. Finally, a cell-by-region matrix is constructed, each row representing an individual cell, each column representing a defined genomic region, and matrix entries reflect the methylation level or count of methylated sites.

1.1.2. Multimodal single-cell technologies

Traditional single-cell assays measure one modality (e.g., RNA), which gives only a partial view of the cellular state[37]. Multimodal single-cell technologies have emerged to profile multiple molecular layers from the same cell[38]. These approaches provide a more holistic snapshot of each cell's regulatory state by coupling, for example, transcriptome and surface proteome measurements (as in CITE-seq[39]) or chromatin accessibility with gene expression (as in 10x Multiome or SHARE-seq[40,41]). By capturing different layers together, multimodal assays can reveal how variations in DNA, chromatin and protein levels cooperate in defining cell identity and function. These technologies therefore bridge the

gap between layers of regulation, enabling richer cell atlases than single-modality data alone[38].

1.1.3. Spatial sequencing technologies

Most single-cell RNA-seq requires tissue dissociation, which destroys information about where cells were located in the tissue. Spatial transcriptomics methods retain that spatial context by measuring gene expression in situ on tissue sections[42]. In these approaches, either hybridization-based imaging (e.g., MERFISH[43], seqFISH[44]) or barcoded capture arrays (e.g., 10x Visium[45], Slide-seq[46]) are used to map many transcripts to their original positions. This spatial information is essential to understanding the influence of the environment and neighboring cells on cell behavior; spatial methods therefore allow analysis of expression patterns across tissue architecture[16,20]. For example, spatial profiling has revealed how cells near disease lesions express distinct gene programs, insights that would be lost in bulk or dissociated data. Spatial sequencing complements single-cell measurements by adding cells' positional information, which is essential for studying tissues with complex organization.

1.1.4. Single-cell data analysis

The computational analysis of single-cell data involves several standard stages. As a high-level overview: raw sequencing reads are first processed and quality-checked, producing a cell-by-gene count matrix; then basic analysis (normalization, feature selection, dimension reduction, clustering, and marker-gene identification) is performed; finally, advanced analyses (trajectories, interactions, regulatory networks, etc.) can be done tailored to the question. In practice, most workflows follow a general outline: (1) preprocessing and quality control to filter cells and genes, (2) normalization and correction of technical biases, (3) dimensionality reduction (e.g., PCA, UMAP) to represent cells, (4) cell clustering to discover cell types or states, and (5) differential expression testing to find marker genes between groups. Each step has specialized methods and best practices, as discussed below.

1.1.4.1. Preprocessing

Preprocessing (also called data cleaning) prepares the raw single-cell data for analysis. As described previously, sequencing reads are aligned and counted (e.g., generating unique molecular identifier (UMI) count tables for each gene and cell). Quality control filters out low-quality, artifactual or dead cells. It typically consists in removing cells with very few detected genes or a high fraction of mitochondrial RNA[47–49]. Putative doublets (two cells accidentally captured as one) and ambient RNA contamination are also identified and excluded. After cell filtering, counts are typically normalized to account for differences in sequencing depth or capture efficiency. Common normalization rescales each cell's counts (e.g., by total counts or size factors)[50,51] so that gene expression levels become comparable across cells, even if this step can be deleterious in some contexts[52]. In some pipelines,

additional corrections such as batch effect adjustment or regression of technical covariates are applied to further harmonize the data. The output of preprocessing is a filtered, normalized cell-by-gene matrix ready for downstream analysis.

Data preprocessing is actually essential to correct technical biases, evaluate their quality, and harness their whole potential. Neglected, it can both hide important information and lead to wrong hypothesis. Moreover, and as presented in Chapter 3, different preprocessing might give very different results. Preprocessing strategies must thus be chosen carefully depending on the downstream analysis.

1.1.4.2. Differential analysis

Differential expression analysis identifies genes that vary between groups of cells (for example, between clusters, conditions, or cell types). In single-cell data this typically means statistical testing for "marker" genes that are up- or down-regulated in one cluster relative to others[53,54]. Most single-cell pipelines use nonparametric tests to find genes significantly enriched in one cell population. These tests account for the variability and sparsity in single-cell counts and control for multiple testing. Detecting differentially expressed (DE) genes helps to annotate clusters (by their characteristic markers) and to compare biological conditions (e.g., treated vs. control samples). For example, genes that are consistently higher in one cluster than others can be used as cluster-specific markers or to infer regulatory differences between cell states. It is common to complement gene-by-gene tests with pathway or gene-set enrichment analyses, which interpret the DE genes in terms of biological functions.

1.1.4.3. Cell clustering

Clustering is the process of grouping cells by their expression profiles to reveal putative cell types or subpopulations. After dimensionality reduction (e.g., using principal components), most methods construct a graph of cells based on nearest neighbors and then apply community-detection algorithms (such as Louvain[55] or Leiden clustering[56]) to partition the cells. Each cluster is then assumed to correspond to a distinct cell type or transcriptional state[57]. Clustering is typically unsupervised and scale-free (it can handle thousands of cells), but it often requires choosing parameters (e.g., resolution) to capture the level of granularity needed. Well-separated clusters often match known cell types, whereas more subtle "states" or continuous variation may appear as gradients rather than sharp groups[58]. Once clusters are defined, their identities can be annotated using known marker genes or reference atlases. In practice, clustering combined with visualization (t-SNE[59]/UMAP[60] plots) provides an initial map of the cellular diversity present in the data.

1.1.4.4. Feature relationships

Beyond clustering, single-cell analysis often examines relationships among measured features. A key example is inferring gene regulatory networks or co-expression modules. For instance, methods like GENIE3[61] build "regulons" by identifying sets of genes that share

a transcription factor regulator. Similarly, weighted gene co-expression network analysis (WGCNA)[62,63] and other correlation-based approaches can identify gene or protein[64] modules that vary together across cells. These analyses aim to uncover higher-order structure: how genes co-regulate, which transcription factors drive cell states, or how different measured modalities correlate. Other examples are cell–cell interactions inference (predicting cells synchronizing their behaviors), which also uses relationships among features. In summary, feature-relationship analyses extract cell- and gene-level patterns (networks, pathways) from the single-cell matrix, complementing the cell-level results of clustering and differential testing.

1.1.5. Impact and limits of single-cell data

Single-cell sequencing has revolutionized biology by revealing cellular heterogeneity that bulk measurements cannot grasp. With the ability to profile thousands of genes per cell across thousands to millions of cells, we can now chart the full diversity of cell types and states in a tissue or organism. This unprecedented resolution has led to new discoveries of rare cell populations, developmental trajectories, and context-dependent responses. For example, single-cell studies have identified novel immune cell subsets in cancer, traced cell-lineage relationships in development, and pinpointed cell type-specific drug responses[65–67]. These insights translate into single-cell atlases across organs and species, improved understanding of disease mechanisms, and even new biomarkers for precision medicine[68]. Overall, single-cell data provides unprecedented sensitivity and specificity for uncovering molecular mechanisms in health and disease.

However, single-cell approaches also have important limitations. The data is inherently noisy and sparse. Indeed, many transcripts present in a cell are not detected, leading to dropout (artificial zeros replacing true values) that complicate the analysis[69,70]. Additionally, single-cell assays often capture only certain modalities (e.g., RNA but not protein) and can have a bias against fragile or difficult-to-dissociate cells, limiting completeness. Finally, practical concerns include high experimental cost (per cell and especially per multi-omic or spatial assay), batch effects across experiments, and the need for substantial computing resources to handle large datasets. Analytical challenges such as normalization, integration of batches, and statistical modeling of zero-inflated data remain active research areas[71]. In summary, while single-cell data offers unprecedented insights into cellular and molecular biology, it requires careful interpretation due to technical limitations (dropouts, noise), partial coverage, and computational complexity.

1.2. Molecular and cellular Interactions in biology

Cells orchestrate life's complexity through biological interactions at every scale. Molecules inside a cell constantly bind, signal, and regulate each other, and cells themselves exchange signals with their neighbors. Complex biological functions, such as embryonic development of organisms and tissue homeostasis, emerge from the complex interplay of

these molecular and cellular components[72]. Understanding these interactions – how molecules influence one another through binding or control, and how cells communicate – is essential to explaining how complex functions are built from simpler parts[73,74].

In this section, we discuss the nature of intracellular and intercellular interactions, focusing on regulations and co-operations among molecular components, and their role in coordinating cellular behavior. We will also go through the current data and computational approaches to identify such interactions.

1.2.1. Physical interactions between molecules

One fundamental class of interactions is physical binding between molecules. Proteins often bind to other proteins, forming complexes or triggering each other's activity. Likewise, many proteins bind directly to DNA or RNA. Such physical interactions are the foundations of intracellular biology. They enable enzymes to find their substrates, signaling proteins to relay messages, and genetic regulators to control genes. A classic example is a transcription factor protein binding to a specific DNA sequence to regulate a gene's expression, thereby controlling production of another protein[75]. In general, physical interactions can be transient (e.g., a signaling protein briefly docking to activate an enzyme) or stable (e.g., multiple subunits assembling into a lasting complex). For instance, metabolic enzymes may transiently bind their substrates to catalyze a reaction, whereas the ribosome's many proteins and RNAs assemble into a stable complex to synthesize proteins. In both cases, it is the direct contact between molecular surfaces, through shape and molecular affinities, that allows one molecule to affect another's function.

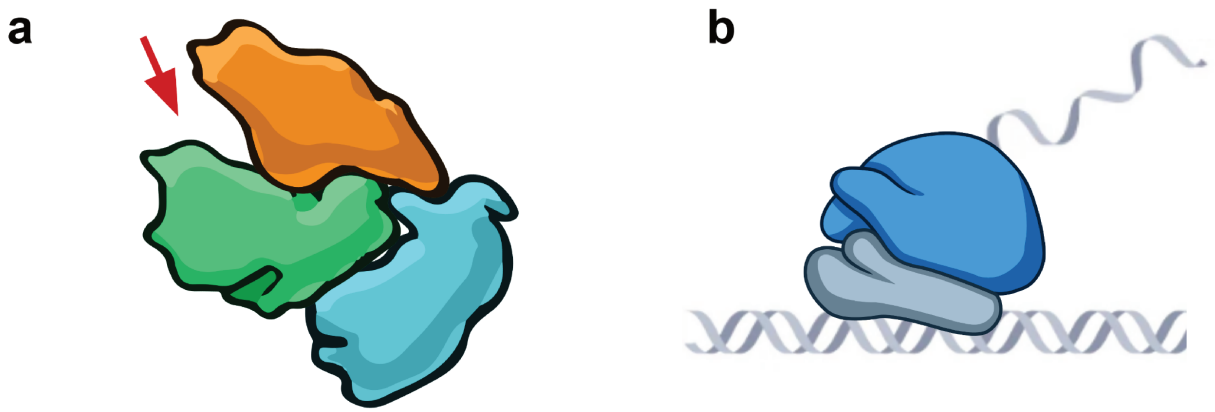


Figure 1.1. Physical interaction with proteins. a) Protein complex forming a "binding pocket", or a specific binding space with a high affinity for specific molecules. The red arrow indicated the binding pocket location. **b)** Interaction between a ribosome – itself a protein complex – and the DNA during the transcription of a gene. *Illustrations of proteins are using Public-Domain entries in the NIH BioArt Source library – bioart.niaid.nih.gov (259, 260, 387, 449).*

Protein–protein interactions (PPIs) underlie most cellular machinery. Many proteins act as molecular "machines" that physically associate in complexes to carry out functions that single subunits could not do alone[76]. For example, the DNA replication apparatus involves

dozens of proteins that form a large complex at the replication fork, each component interacting with others to unwind DNA or synthesize new strands. Similarly, the proteasome is a multi-protein complex whose subunits coordinate to degrade other proteins. Transient PPIs are equally crucial: in cell signaling cascades, an activated kinase enzyme might physically bind a target protein to phosphorylate it (adding a phosphate group) and thereby change the target's activity. This kind of interaction between proteins is the basis for information flow inside cells.

Protein–DNA interactions are another vital subset of physical interactions. Transcription factors bind DNA to modulate gene transcription rate, acting as the cornerstone of gene regulation. The classic *lac* repressor in bacteria, discovered by Jacob and Monod[75], binds to the DNA of the *lac* operon to prevent transcription; when lactose is present, the repressor's shape changes so it can no longer bind DNA, allowing the genes to be expressed[75]. Eukaryotes have thousands of transcription regulators, from factors that recognize specific gene enhancers to chromatin proteins that bind and package DNA. These interactions ensure that each gene is expressed only in the right cells at the right time. Physical binding of proteins to DNA can also initiate DNA replication or repair by recruiting necessary enzymes to the correct locations on the genome. To summarize, whenever a cell needs to read, duplicate, or fix its genetic information, it uses proteins that interact directly with the DNA helix or its structuring proteins.

1.2.2. Metabolic and signaling cascades

Beyond direct binding, molecules also interact through biochemical reactions and regulations. In metabolism, for example, one enzyme's product often becomes another enzyme's substrate, forming a chain (or cascade) of transformations. These metabolic interactions link enzymes and metabolites into pathways with specific cellular functions. A simple example is the Krebs cycle, a well established cycle of reactions to produce energy, in the form of ATP, from nutrients[77]. The end product of one reaction step is the starting point for the next one. In a broader sense, the entire cell's metabolism can be described with such interactions, which connect enzymes functionally by the substrates they share. If one enzyme in a pathway is missing or inhibited, its substrate may accumulate and its product will not be produced, affecting all downstream reactions. Enzymatic interactions are not limited to linear chains; they often branch and form feedback loops. A downstream product might circle back to inhibit an earlier enzyme (negative feedback) to maintain balance. In the example of the Krebs cycle, it is helpful to not use all the nutrients available at once. Importantly, molecules do not need to form long-lived complexes, but they encounter each other through diffusion and react dynamically.

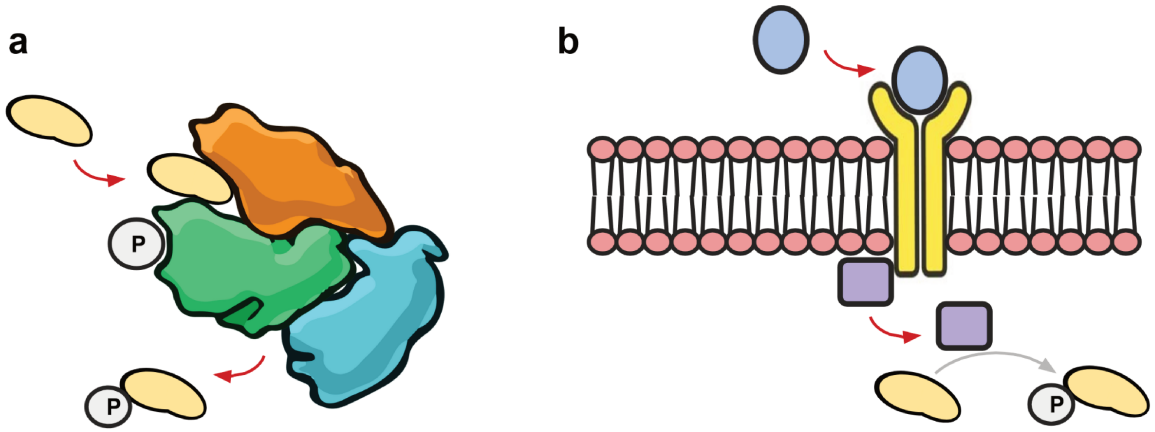


Figure 1.2. Metabolic transformation and signaling mechanisms. a) Schematic representation of a protein phosphorylation (adding a phosphate group on a molecule) through a kinase activity. **b)** Schematic view of a signaling pathway cascade. An external ligand (blue) binds a receptor (yellow), which activates a kinase mediating intracellular effects. *Illustrations of proteins in **a** use Public-Domain entries in the NIH BioArt Source library – bioart.niaid.nih.gov (259, 260, 387).*

These reaction cascades are also essential in integrating external signals. When a signaling molecule (like a hormone or a growth factor) binds to a receptor on the cell membrane, it triggers a chain of enzymatic events inside the cell. The receptor is often an enzyme or a kinase, which upon activation will modify an intracellular protein. That protein might then also act on another, and so on, forming a signal transduction cascade. Each step is an interaction where one molecule's chemical activity (e.g., adding a phosphate group) modifies the next molecule. These cascades can amplify a small external signal into a large response inside the cell because each enzyme can activate many downstream targets. Thus, the signaling mechanisms involve transient modification and binding across time. It overlaps with physical interactions since enzymes often must bind their targets, but emphasizes the dynamical action (catalysis, modification, transformation) more than just binding.

1.2.3. Genetic circuits regulation

Metabolic and signaling cascades often begin and end by the regulation of gene expression, which ultimately lead to the synthesis of the protein involved in these cascades. Another layer of interaction can thus be defined, between genes and their regulators.

In cells, genes do not act nor are transcribed in isolation; one gene's product (usually a protein) often controls the activity of other genes. It is common to simplify these regulations to the impact of proteins, called transcription factors (TFs), on downstream genes. TFs can both positively and negatively regulate gene expression, by enhancing or repressing their transcription respectively. Some of these genes might also encode TF proteins, themselves regulating genes' expression. These regulatory interactions create

circuits governing cell behavior, dynamic and fate by changing genes and protein production.

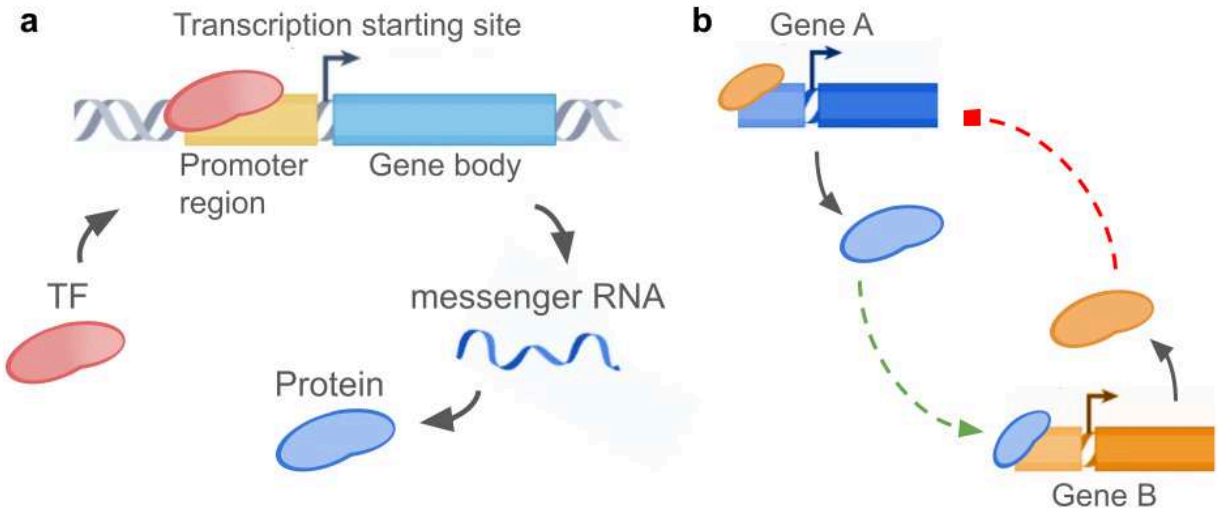


Figure 1.3. Transcription and gene regulatory circuits. a) Schematic representation of the regulation by transcription factors (TF). A TF binds the DNA around a gene promoter or enhancer, triggering the gene transcription by creating space for the RNA polymerase. It produces a messenger RNA, translated later into a protein (e.g., another TF). **b)** Schema of a feedback loop where a gene B represses its own regulator, the gene A. The TF A activates the transcription of the gene B, which produces the TF B and inhibits in return the transcription of the gene A. It can create different dynamics and steady states, depending on the speed of each step.

Cells contain many overlapping gene regulatory circuits, such as feedback loops where a gene product regulates its own transcriptional regulator). Feedback loops can notably produce phenomena like bistability (i.e. a stable “on” or “off” state of a gene, creating a memory of a past signal) or oscillations (i.e. recurring cycles of activity, as seen in circadian rhythm genes)[78]. A variety of other motifs also exist and allow complex logic and inertia in gene expression, such as feed-forward loops, whereas X regulates Y and Z, and Y also regulates Z. Through regulatory interactions, cells can adapt to new information, such as the presence of nutrients or specific conditions of pressure and temperature.

1.2.4. Intercellular interactions and cell–cell communication

So far, we have presented interactions within a single cell. However, in multicellular organisms, cells must also interact with each other to coordinate their activities. No cell lives in complete isolation: neighboring cells exchange chemical signals, adhere to one another, and even directly pass molecular messages. These intercellular interactions are the basis of tissue organization, development, and physiology in multicellular life. In fact, the evolution of multicellularity required the development of complex cell communication mechanisms to ensure that cells could work together for the organism’s benefit[79]. It is possible to differentiate several types of communication, associated with different ranges of actions.

Depending on their roles in the communication, molecules can be named ligands, as the molecule transporting the initial information, or receptors as the first molecule receiving the information in the target cell. Typically, small groups of ligands and receptors have matching shapes and a selective affinity for each other. Both are essential for the communication and their expression presents two complementary ways to modulate cell communication.

1.2.4.1. Paracrine and endocrine communication

One major mode of intercellular interaction is signaling via secreted molecules. Cells can secrete signaling molecules such as hormones, growth factors, or neurotransmitters, which then travel to other cells and bind receptors to influence their behavior[80]. In paracrine signaling, the signals act on nearby cells (e.g., a neuron releasing a neurotransmitter to signal an adjacent neuron). Autocrine communication is defined as a specific subcategory, where the signals are emitted and received by the same cell.

In endocrine signaling, hormones released into the bloodstream can affect distant cells throughout the body (e.g., insulin released by pancreatic β-cells tells muscle cells to uptake glucose[81]). These soluble signals and their receptors form complex interaction networks between cells. A cell often integrates numerous signals, such as inflammatory cytokines and growth-related hormones, which altogether drive its response (grow, move, differentiate, or even die)[82].

1.2.4.2. Juxtacrine and contact-mediated communication

Cells also interact through direct contact. In many tissues, cells are physically joined by adhesion molecules on the cells' surfaces. For example, in epithelial tissues, adjacent cells are glued together by proteins (e.g., cadherins, integrins) that span their membranes and bind one cell to another. This cell adhesion not only provides structural integrity but also carries signals[83,84]. Indeed, some adhesion proteins send signals into the cell when they attach or detach, informing the cell about its environment or neighbors.

Another form of direct interaction is through gap junctions. Gap junctions are physical channels connecting the cytoplasm of adjacent cells, allowing small molecules and ions to pass directly from one cell to another. They enable electrical and metabolic coupling – as in heart muscle, where ionic currents flow through gap junctions to synchronize contractions of cells[85].

1.2.4.3. Specific signaling structures

Finally, some intercellular interactions involve specialized structures or phenomena. Neurons can connect to other neurons via chemical synapses, where the electrical signal of one cell triggers the release of neurotransmitters that quickly bind receptors on the next cell.

In development, certain cells serve as signaling “centers” that release morphogens, in the form of diffusible molecules, creating a concentration gradient through organisms and organs. Surrounding cells detect these gradients and decide fate according to the morphogen level, which leads to spatial patterns following organism's polarity. For instance, the graded Sonic hedgehog signals have an essential role in limb development[86].

In adult tissues, cells constantly signal distress or needs. For example, damaged cells release factors that prompt immune cells to come[87], and oxygen-starved cells induce blood vessel growth[88]. In the immune system, a delicate interplay of cell-cell contacts and signals distinguishes friend from foe to protect the body. Even unicellular organisms exhibit community interactions: bacteria communicate via quorum sensing, secreting small molecules and sensing their concentration to coordinate group behaviors (e.g., biofilm formation[89]) once a threshold population is present.

1.3. Network representations and exploration

Across molecular and cellular scales, nothing in biology works alone. From interactions between smaller components emerge both complex functions and their fine coordination across tissues and organisms. These interactions can be physical bindings, chemical reactions, regulatory influences, or direct cell-cell contacts. Collectively, they form vast interaction networks ensuring the coordination of different processes and underlying both cellular and physiological functions.

In recent decades, biology has embraced a network perspective, recognizing that analyzing interactions between molecular components is essential to understand cells and organisms[72]. This has led to the emergence of new fields such as systems biology and network biology, which treat the cell as an integrated system of interacting components rather than a set of independent molecules. By defining networks of molecular interactions, we can have a better understanding of emergent complex system properties, such as collective behaviors[74] and robustness (i.e. the ability of a system to maintain function despite perturbations).

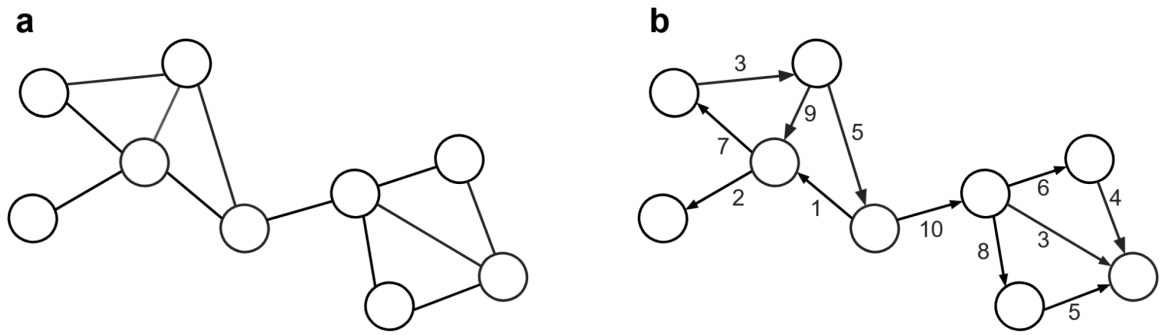


Figure 1.4. Directed and weighted properties of networks. a) Representation of an unweighted and undirected network. **b)** Representation of a directed and weighted network

Networks are mathematical objects to represent interactions between different elements[90]. The networks can be formalised as $G = (V, E)$, with V representing a set of nodes (elements) and $E \subseteq V \times V$ representing a set of edges (interactions) between nodes. These nodes and edges can have additional properties defining different types of networks. Edges can notably be unsigned or signed, depending on whether we want to differentiate positive and negative interactions (e.g., activation and repression). They can also be weighted with a score indicating the strength of the interactions, and directed or undirected to indicate their directionality (e.g., regulation or co-operation).

Additionally, different methods have been developed to analyse large networks, allowing to extract meaningful information such as nodes of interest, hub of highly interacting elements, or general properties and structure of the network. We will next present the main biological networks and networks methods used in this thesis.

1.3.1. Networks in biology

Numerous network types are used to describe biological systems at both molecular and cellular scales. Key examples include 1) protein–protein interaction (PPI) networks, 2) gene regulatory networks (GRNs), 3) signaling networks, and 4) cell communication networks, each capturing a distinct facet of cellular and tissular organization.

1.3.1.1. Protein–protein interaction (PPI) networks

PPI networks contain physical or functional interactions between proteins (e.g., protein complexes, ligand-receptor bindings)[91]. Proteins are then represented as nodes, and their interactions as edges. The edges can be weighted (e.g., confidence) and are usually undirected, since they usually represent co-operations and not regulations between proteins. On a structural note, PPI networks often contain hubs, few central nodes connected to a high number of neighbors.

In the context of this thesis, PPI networks are mostly used to represent complexes in-between TFs and in-between receptors, which can form oligomers to regulate downstream elements.

Until recently, PPI was mostly inferred from direct interaction methods, such as yeast two-hybrid[92] and affinity-purification/mass-spectrometry[93]. Many tools are nowadays developed to computationally infer protein interactions, through prediction of their complex 3D folding. Under the hypothesis that these interactions are largely due to structural properties independent of the cell type, they are stored per species in large databases[64] and reused between different datasets. In parallel, the field of single-cell proteomics data is starting to develop and to gather more interest. This new methodological development promises to uncover in the near future how mutations and diseases can alter PPI networks in individual cells[94].

1.3.1.2. DNA interaction networks

Through DNA folding, portions of the DNA separated by hundreds of thousands of base pairs can actually interact. Several methods propose to reconstruct chromatin interaction networks through chromosome conformation capture methods, such as Hi-C[95] and Micro-C[96]. This concept consists of cross-linking the DNA with formaldehyde, that is, "freezing" the interaction between chromosomal regions that are spatially close enough (~100 nm). Once these interactions are stabilized, DNA can be digested with restriction endonuclease to isolate the complex formed by the two interacting regions, sequence them, and map them to the genome. The older Hi-C version outputs a map of regions of 1-10 million base pairs each, or very large domains. Recent development has allowed the reduction of this granularity to a few nucleosomes for micro-C. In addition, Capture Hi-C has been developed to extract all the regions able to interact with added small probes. It allows the DNA regions to interact with a specific promoter region for example.

Other works adopt a functional look at DNA interactions, predicting cis-regulatory interaction networks. They then often focus on enhancer - promoter interaction, inferring the role of a region in the regulation of a gene promoter transcriptional activity. Some methods measure directly enhancer activity[97,98], while others infer regulations from co-accessibility of the DNA, for example, through scATAC[33,99–101]. Typically, they consider much smaller DNA regions (~100 to 10 thousand base pairs) and consider interactions possible only if their distance on a chromosome is inferior to a threshold limit (~200 thousand to a few million base pairs).

The DNA interaction networks are typically undirected, except when coupled with external information about what regions are regulatory and what regions are regulated. They usually have scores, as the probability or the confidence associated with the interaction.

1.3.1.3. Gene regulatory networks (GRNs)

Gene regulatory networks represent the regulation of gene transcription by TFs. In contrast to PPI networks, GRNs contain direct edges, from TF to genes, that can be signed to discriminate activation and inhibition of transcription. The edges are usually weighted by either the strength of the regulation or its confidence score. While some databases exist to use known TF-gene interaction, the explosion of bulk RNA, and later single-cell RNA sequencing methods, led to the development of a plethora of methods to infer sample and cell type specific GRNs[7–10,61,102–106].

Two paradigms exist to represent cell identity through GRNs. A first group of methods chooses to define a shared GRN encoding all possible interactions between TFs and target genes[7–10]. The difference between cell types and individual cells is then explained in a second step, by the different expression/activity level of the TFs[8,9,103]. A second group defines directly cell type-specific GRNs, which already encodes the activity of all TFs[104–106]. Therefore, it requires to separate from the beginning groups of cells which will be modeled independently.

Regulatory interactions are often mediated by direct physical binding, but they can also be defined functionally (e.g., gene X "interacts" with gene Y if a change in X's activity alters Y's expression, even if indirectly). Through perturbation studies, it is possible to test regulatory influences by perturbing one gene and measuring its effect on other molecules. For example, CRISPR or RNA interference can be used to "knock down" a transcription factor and to identify which genes' expressions vary consequently. Those genes are classified as targets of this transcription factor. Large-scale studies have mapped many interactions, constructing draft gene regulatory maps for organisms like yeast and human cells[107]. These maps show a dense connectivity, whereas each transcription factor can regulate hundreds of genes, and many genes are controlled by multiple regulators, forming a robust web of control.

For genome-wide networks, recent mainstream methods depend on scRNA-seq data. Since single-cell measurements are still extremely noisy despite the recent progress, they filter it with additional information, either measured in the same samples (e.g., scATAC-seq) or from prior knowledge (e.g., physical TF binding around gene TSS with CHIP-seq, TF motifs screening)[108].

1.3.1.4. Signaling networks

Signaling networks in cellular biology are directed graphs representing cascades of molecular reactions[109,110]. Nodes represent signaling molecules (e.g., ions, ligands, receptors, kinases, small GTPases, TFs) and edges symbolise causal biochemical events (e.g., bindings, phosphorylations, post-translational modifications). They are often signed to differentiate activations and inhibitions.

Signaling networks are largely derived from high-confidence pathway databases, compiling manually annotated functional pathways. Other initiatives (e.g., Signor and OmniPath)[111,112] also provide large signaling networks from text mining and literature-based interactions. These interactions can be obtained from different experiments, reflecting those heterogeneous interactions that signaling networks can integrate. It includes notably chemical or genetic perturbation experiments[113], enzymatic modification measurement (e.g., phospho-specific Western blot[114], kinase assays[115]), and physical binding (e.g., co-immunoprecipitation[116], FRET[117]/BRET[118]).

Because of the broad spectrum of interactions that those databases regroup, they are mostly "bulk" network assemblies, not cell type- nor sample-specific.

1.3.1.5. Cell-cell communication networks

Cell-cell communication networks model intercellular signaling as directed graphs, where nodes represent either a cell type or individual cells. The edges can be either a function of all signaling interactions between the involved cells/cell types, or specific to one ligand-receptor couple (e.g., multiplex view, where each pair of cells can be linked in multiple ways)[19].

The current method of studying cell cell communication focuses on known ligand - receptor pairs, filtered and weighted on single-cell RNA-seq[119–122], single-cell proteomic[122,123] or spatial transcriptomic data[19,124,125]. They assume a common backbone network encoding possible intercellular interaction, and personalize it depending on the context (expression and/or location) for individual cells or groups of cells.

These networks do not model only molecular interactions reflecting intrinsic physical properties, but also signal expression and transmission between cells. Those networks can be used to explain tissue and patient differences, which are summarized in the networks themselves. They also propose a different scale to study biological interactions in-tissue and in vivo, less detailed. This led to recent works linking cell communication networks to the previously presented molecular intracellular networks[19,22,126].

1.3.2. Network exploration

Once biological networks are constructed, various analytical strategies can be used to explore network topology and extract meaningful biological insights. Key approaches include topological analysis to find important nodes, community detection to find modules, and network propagation techniques to trace signal flow. We will briefly present some concepts useful to understand this thesis.

1.3.2.1. Topological analysis and centrality measures

Topological analysis involves quantifying the positions or roles of nodes in the network by using centrality measures. Centrality metrics rank nodes by their structural importance and have been widely used to pinpoint key players in biological and sociological processes[127]. For example, the number of connections a node has, also called a node degree, highlights network hubs. In directed networks, a node in-degree and out-degree can be defined as the number of edges reaching or leaving a node, respectively. For a weighted network, it is usually replaced by node strength, which is the sum of the weights of its edges[90]. In PPI networks, hubs tend to be functionally important: it was shown that highly connected proteins are often essential, and removing them can lead to lethal phenotypes[128]. This suggests that the connectivity (degree) is one indicator of a node's criticality in the network. However, a node's degree alone is not always sufficient to capture its biological importance. Therefore, other centrality measures have been developed to characterize different aspects of "importance" in a network.

One such measure is betweenness centrality, which counts how often a node lies on shortest paths between other nodes. Nodes with high betweenness (sometimes called bottlenecks) act as bridges connecting different parts of the network[129]. These bottleneck nodes control the flow of information and often prove to be critical connectors. Notably, betweenness can be an even stronger indicator of essential genes than degree in certain biological networks, particularly directed ones such as GRNs. In other words, a protein that connects multiple pathways or modules may be indispensable despite not having a high degree. Other centrality measures also provide valuable perspectives. For example,

closeness centrality evaluates how close a node is to all others (i.e., measuring how much information transits through this node). In comparison, eigenvector centrality, which is exemplified by the PageRank algorithm, assigns higher scores to nodes connected to other well-connected nodes, capturing an idea of global influence.

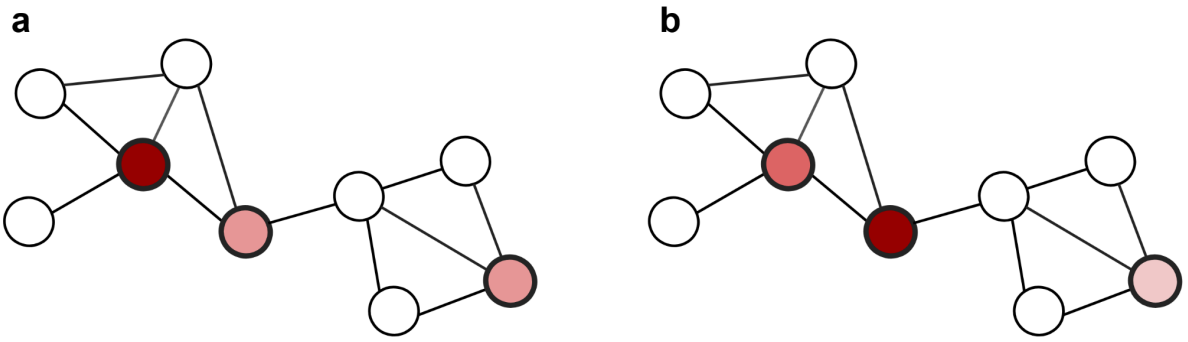


Figure 1.5. Comparison of centrality measurement. a) Node degree illustration highlighting three nodes (heavy border) of a small network. The fill colour indicates the degree and the centrality value ; the darker the tone the higher value. **b)** Node betweenness illustration for the same three nodes.

By using centrality metrics, researchers can identify hub proteins, key connectors, and influential regulators in the network. This topological analysis is particularly useful for finding candidates for experimental validation, such as hub genes in a gene regulatory network or bottleneck proteins that might make good drug targets.

1.3.2.2. Molecular footprints and activities

A powerful, yet simple extension in network biology is the use of molecular footprints and activity as a measure of regulator importance[130,131]. The fundamental premise is that when a regulator (e.g., a TF or a pathway) is active, it induces changes in its downstream targets correlated to the strength of the regulation. These changes can be seen as the “footprint” of a regulator, defined as its different targets and their associated weights.[132,133] The targets are typically the out-degree edges of a regulator in a network, but this definition could be extended to larger ensembles of elements.

From these footprints and measurements of the target elements (e.g., gene expression or phosphorylation changes), each regulator can be associated with a “molecular activity”, predicting its importance for a cell or sample molecular profile[134]. Once calculated, these activities can help identify regulators and pathways driving the cell phenotype, even if those regulators are not themselves differentially expressed.

Unlike purely topological metrics (e.g., centrality or cluster membership) or highly parameterized models, footprint methods yield biologically meaningful summaries such as lists of key regulators (e.g., TFs, kinases, signaling pathways, ligands) that are activated in a given context, derived by leveraging known network connections. These inferred activities directly point to upstream controllers and pathways governing the system’s state, helping

researchers move from a diffused list of omics changes to a clearer picture of "who" is orchestrating those changes. In effect, they bridge between data-driven discovery and mechanistic insight.

1.3.2.3. Community detection and network modules

Many biological networks are organized into modular structures, meaning they consist of groups of nodes densely connected to each other but only sparsely connected to the nodes in other groups[135]. Identifying these groups, also called modules or communities, is essential for understanding the functional organization of a system[136]. In molecular networks such as PPIs, communities often map to known biological functions or complexes. Depending on the size considered for each community, it is possible to identify communities illustrating complexes of several proteins interacting physically with broader biological functions such as cell cycle or apoptosis[137].

To uncover such network modules, researchers use community detection algorithms. Classic methods like the Girvan–Newman algorithm exploit edge betweenness (i.e., removing high betweenness edges that link communities) to iteratively split the network into communities[135]. Other popular approaches (e.g., the Louvain method[55] or hierarchical clustering techniques) optimize a global modularity score that measures the difference in the density of links inside communities versus between them. The result of these analyses is a partitioning of the network into coherent modules, which can then be compared with known biological subsystems. Often, there is a strong agreement: for instance, densely connected clusters in PPI networks have been found to coincide with bona fide protein complexes or signaling[135,137]. Such findings support the idea that modules are the "building blocks" of biological networks and that evolution might favor modular organization for robustness and functional specialization. In network analysis, complex networks can be decomposed into coherent sub-networks and groups of nodes, each potentially corresponding to a specific biological process or complex.

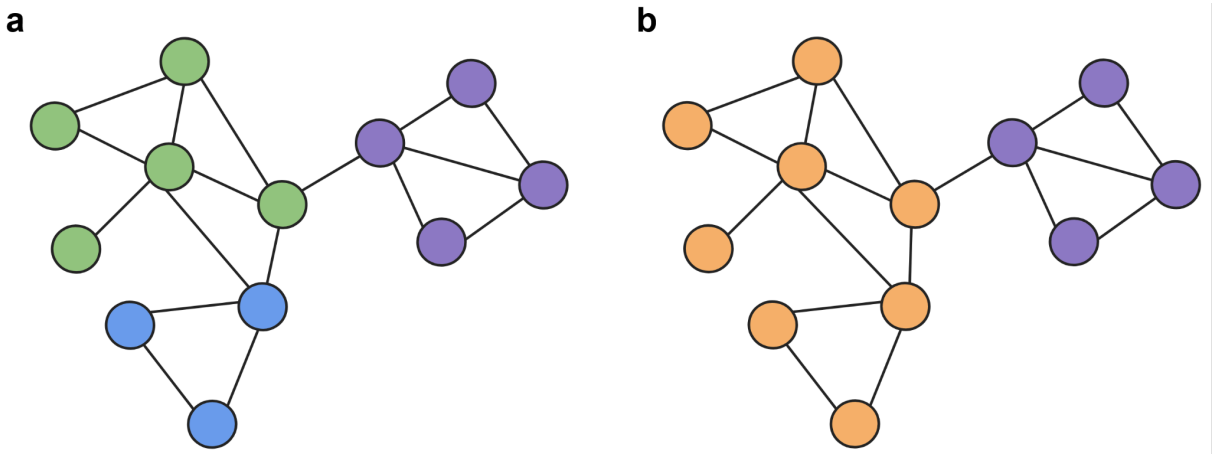


Figure 1.6. Community detection in a network. a - b) Communities detected in the network are indicated with different colors. Depending on the granularity needed, it is possible to identify a different number of communities – three in the panel a) and two in the panel b).

1.3.2.4. Diffusion and network propagation methods

Beyond static topological features, another class of network analysis methods uses diffusion-based propagation to spread information along the connections of the graph. The core idea is to simulate a process where a "signal" starting from a node (or a set of nodes) disperses through the network, similarly to heat diffusion or a random walker exploring a graph[138,139]. This approach transforms a small seed input (e.g., a few genes of interest) into a full, network-wide profile of scores, where each node's score reflects its network proximity or connectivity to the seed nodes. Such network propagation techniques have proved powerful in uncovering relevant nodes that might be missed by local analysis, since they consider the whole network structure. Indeed, numerous algorithms based on random walks, heat diffusion, or even analogies to electrical currents have been successfully applied to identify disease genes, functional modules, and drug targets using networks[139–142].

The main algorithm used nowadays is random walk with restart *(RWR)*. In RWR, a hypothetical walker starts at a seed node and, with each step and according to pre-defined probabilities, it either moves to a random neighbor or returns (restarts) at the seed[138,143–146]. Over many iterations, this process converges to a steady-state distribution of probabilities across the network. Nodes that are "closer" or more strongly connected to the seed accumulate higher probabilities over time, effectively measuring their relative proximity or influence with respect to the seed. Consequently, RWR yields a ranking of nodes based on how easily one can reach them by walking the network starting from the seed. Mathematically, this is equivalent to a diffusion process or a personalized PageRank score centered on the seed nodes[147]. As a result, RWR-based methods significantly outperform simple neighbor-based approaches in identifying genes associated with a disease, when spreading influence from known disease genes through the interactome[145].

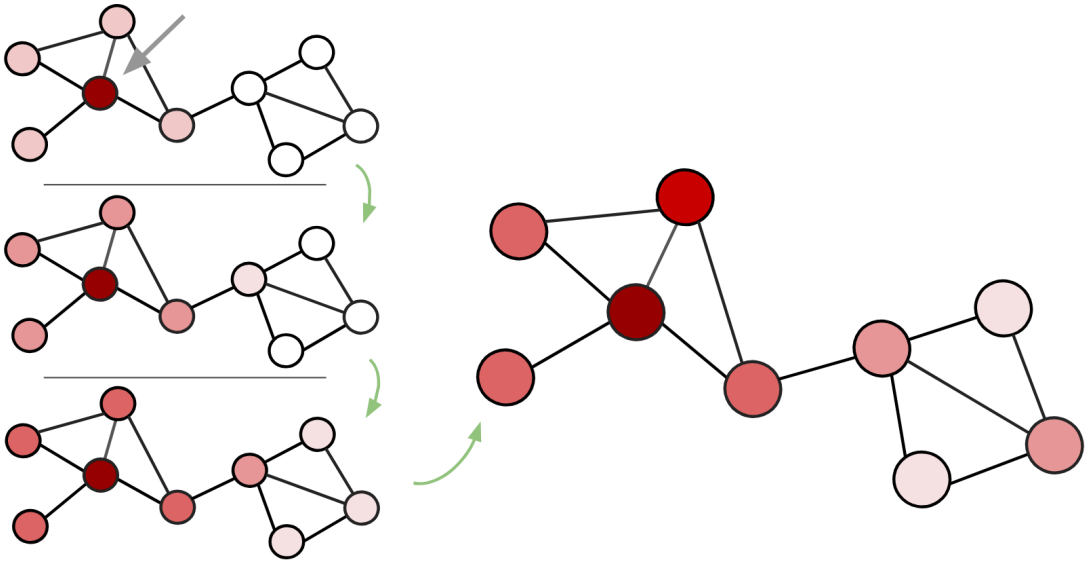


Figure 1.7. Illustration of a diffusion process across a network. Successive steps of the diffusion process are represented, showing the progressive spreading of the information across the network.

The green arrows indicate the order of the steps, and the darkness of the nodes proportional to their scores; the darker red corresponds to the input signal, or seed.

Overall, network propagation integrates information over the network's topology, helping to reduce noise and highlight interconnected nodes. These diffusion algorithms (with different kernels or parameters) are used in network biology to predict new members of pathways, propagate perturbation effects, or find network clusters of high activity. They also trace plausible and interpretable paths of influence in a network, which helps predict how perturbations or signals travel through a biological system, providing a global view of network connectivity beyond immediate neighbors.

1.4. Integrating molecular and cellular networks

Each molecular view and network presented above covers distinct cooperative and regulative aspects of a cell's identity. Their choices depend on the network scale, network type, and input data. Different methods have been developed to regroup some of these views in an unique framework, drawing a more complete cell model. We will here go through different network integration methods and outline their respective benefits.

1.4.1. Network fusion to combine partial molecular views

Complementary networks can exist to describe distinct aspects of complex systems. For example, in biology, it is often considered that the gene expression, protein abundance, or DNA accessibility measurements provide all partial and non-redundant information to describe a cell's fate. Sometimes, the same network can also be inferred from several data types (e.g., extracting PPIs from text mining and from crystallography) that both have partial coverage or important noise. To leverage their complementarity while applying "traditional" network analysis, different methods propose to aggregate these networks into a consensus network. This aggregation is sometimes done intuitively (e.g., keep only the shared interactions between a data-driven network and a literature-based one, or between two patient-specific networks), which can lead to the discarding of relevant information.

Recently, several methods have been proposed to define a consensus network from N independent networks sharing the same set of nodes[148–153]. Among them, similarity network fusion (SNF) has been widely applied in biology. While it was initially thought to identify groups of patients sharing similarities across several networks (e.g., built from different omics measurements), different works have used it to identify consensus feature-networks. It has notably been applied for building consensus co-expression networks[154] and protein-protein similarity networks[155].

While SNF and similar methods are useful to leverage the complementary information of multiple networks, they require them to share a similar set of nodes (e.g., map protein and DNA regions both to gene names). Moreover, it usually assumes that the edge weights are

comparable between networks, which depends on the algorithm chosen to build the independent networks.

1.4.2. The flaws of single layer networks

In social sciences, relationships between individuals can be of different nature and roles (e.g., family, friends, co-workers). Looking at only one of them could not describe the true affinities that organise human societies. Because of the relationships heterogeneity, merging them into a singular network is also challenging (e.g., some can be directed and others indirected, some can evolve through times while others are more stable). To describe those networks, sociologists in the mid-20th century developed the notion of multiplexed interactions[156,157]. Instead of merging all social relationships into one network, each type of relationship is stored in multiple independent networks (or a multiplexed network). Communities can be identified from each of the networks (or layers), and integrated a posteriori through different rules. Notably, this allowed more accurate prediction of the evolution of people's ties over time, taking into account the rule of each type of relationship[157]. The overall ties between people depend on the strength of each type of relationship, but these strengths might not be comparable between different types of interactions. The relative importance of each interaction type might also depend on society's values (e.g., importance of work or nuclear family). Keeping all networks separated addresses both of these limitations.

More recently, multiplex networks[158] (and generalized frameworks such as multilayer network[11]), have been formally defined to describe these ensembles of networks. Nowadays, it is a concept widely used in physics, social sciences and biology.

1.4.3. Multilayer networks

We will go briefly through the mathematical definitions of multilayer networks, which will be used for the works presented in Chapter 2 and Chapter 4.

1.4.3.1. Multilayer networks (MLN)

Multilayer networks (MLN) are structures composed of several "elementary layers" of graphs. They contain a set V of nodes, which can all be represented several times in a set L of layers that will organise the different types of interactions. Each elementary layer $m \in L$ thus contains a subset of node-layer tuples $V_m \subseteq V \times \{m\}$, or node representations. The node representation from each layer can then be connected to any node representation of any layer. We can then list the edge internal to a layer $E_m \subseteq V_m \times V_m$ and the inter-layer edges $E_{(m,n)} \subseteq V_m \times V_n$ for any $(m, n) \in L \times L, m \neq n$. While most of the time, they are treated differently, we can still annotate them altogether as E. We can thus define a multilayer network G such as $G = (V_M, E, V, L)$, with $V_M = \bigcup_{m \in L} (V_m)$ as the union of node-layer representations. Finally, layers can be treated as a group of layers, indicating

shared meaning as in Figure 1.8. (e.g., elementary layers defined on input data category, layers defined on time-points).

In cases where only identical nodes are connected through inter-layer edges, the multilayer network is called a multiplex network.

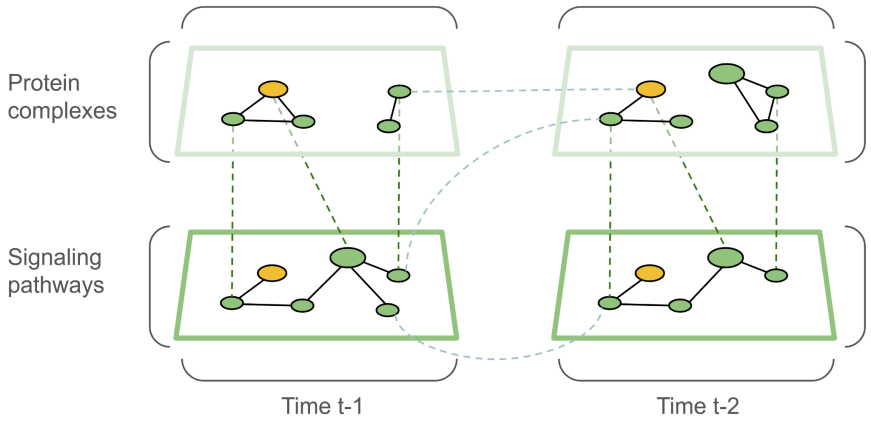


Figure 1.8. Example of a multilayer network. The network has 4 layers (L) which all contain a representation subset (V_m) of the same set of 6 nodes (V). Layers sharing similar meanings are symbolized by brackets (e.g., time snapshot, type of interactions). Continuous lines indicate intra-layer edges and dashed lines indicate inter-layer edges. In yellow, all representations of the same node are indicated, to illustrate its presence in all layers.

1.4.3.2. Heterogeneous multilayer networks (HMLN)

Heterogeneous multilayer networks (HMLN) can be defined as a subset of multilayer networks, where no nodes are shared between groups of layers. However, each group of layers can still be composed of layers sharing the same nodes, depending on the definitions. They are particularly useful in integrating distinct types of nodes with their own interaction mechanisms, such as protein, gene, drugs, diseases, or symptoms in biology. Since the nodes are different, the edges are also heterogeneous and have different meanings (e.g., gene causing a disease, drugs binding a protein). They can additionally follow different attributes : the links from the layer m to the layer n can be directed and weighted, while the links from the layer n to the layer m might be undirected and unweighted.

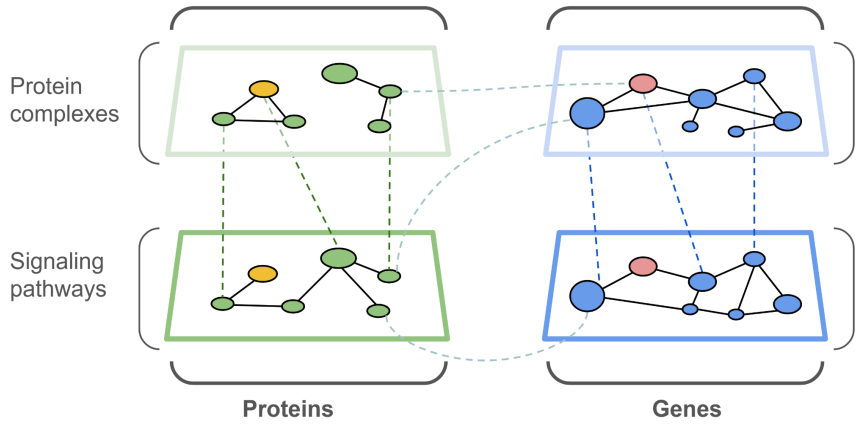


Figure 1.9. Example of a heterogeneous multilayer network. The network has 4 layers (L), divided in two groups which contain non-overlapping sets of nodes – proteins versus genes. Continuous lines indicate intra-layer edges and dashed lines indicate inter-layer edges. All the representations of a protein (in yellow) and of a gene (in red) illustrate their presence in a distinct group of layers.

Several works leverage HMLN to uncover links and communities between various types of biological molecules and categories[10,159–161]. This structure will also be used in the methodologies of both Chapter 2 and Chapter 4.

1.4.3.3. Multilayer networks exploration

MLN and HMLN propose a complex structure containing many nodes and interaction types at once. In contrast to SNF or methods that would simplify the network structure, suppressing intermediary nodes, this method leverages the importance of multiple paths and intermediary nodes, explaining the proximity between nodes.

All classical network metrics, such as node degree, can be adapted to MLN. The simplest way to define it across the whole network consists of layer aggregation. It could be equivalent to summing the degree d_i of each node i representation (i, m), $m \in L$, then subtracting the links in-between node representations to not count them twice (i.e. not count $n \to m$ and $m \to n$) (*see 1.1*).

$$d_i = \sum_{m \in L} d_{(i,m)} - \sum_{m,n \in L,\, n<m} 1_{\{(i,m) \leftrightarrow (i,n)\}} \qquad \textbf{(1.1)}$$

Additionally, the definitions of paths and walks can be adapted from single-layer networks. The main fundamental modification concerns the interpretation of changing or staying inside a layer. Indeed, one could assign a penalty or weight $\gamma_{m,n}$ to any $(m, n) \in L \times L$ transitions between any layers m and n, such as $(m, n) \in L \times L$. In MultiXrank[146], authors defined a random walk with restart algorithm for any type of MLN, and used a transition matrix with jumping probabilities for each pair of layers.

1.4.4. Heterogeneous multilayer networks for biological interactions

Molecular interactions in cells include both unidirected co-operation and directed regulations between different sets of molecules. Single-layer networks offer limited options to integrate all the heterogeneous interactions behind tissue and cellular molecular mechanisms.

1.4.4.1. The importance of complex networks

“All models are wrong but some are useful”. One may still question the interest in complexifying network representation of cellular processes. As presented earlier, life emerges from co-operation between molecules. These molecules can also be involved in several functions themselves (e.g., give an example). In most interventional cellular biology and medicine, the aim is to selectively modify the intensity of one function, conserving the overall stability of the system. For example, it may be desirable to enhance the production of a specific protein (e.g., inflammatory cytokines by immune cells), without altering their metabolism and survival rate. Conserving more interactions and molecular layers allows to picture more accurately the possible adverse consequences on a system of any molecular perturbation.

With the increase of interventional methods and their specificity (e.g., design of protein ligands, inhibiting mi-RNA, CRISPR- KO), new options emerge to development of patient-specific treatment. It is becoming particularly useful to find the best target across omics for each disease treatment development.

1.4.4.2. Overcome sparsity in prior knowledge and data

On HMLN, walkers and distances can be computed using different cross-layer paths. It offers an interesting set-up to overcome the high sparsity in some groups of edges. For example, in the context of inferring TF-TF interactions, evidence about the TF A and TF B dimerization might not be available. However, if both are observed to bind the same DNA regions, this co-binding serves as a strong indicator of potential dimerization. By representing DNA regions as nodes in another layer, it is possible to establish an indirect link through co-bound regions from one TF to the other. Similarly, for drug-protein interactions, using PPIs or protein functional relationships can still identify the impacted protein through alternative paths.

1.4.4.3. Modeling the complexity of gene regulation

Gene regulation and GRN reconstruction are interesting examples of inter-modal molecular interactions integration, for which we developed a HMLN framework in Chapter 2.

Since TF - DNA motif binding information is available, GRNs are often summarized as TF - DNA region - gene triplets. A TF binds a DNA region, which can impact a gene transcription rate through the DNA region - TSS proximity. It allows straightforward modeling where

gene expression depends only on the region accessibility and the TF presence. As a network representation, this is captured by directed edges from TFs to DNA regions, and from DNA regions to genes.

However, many co-operations that are essential in modeling the true process of gene regulation are ignored. Notably, it does not consider **1)** TF -TF interactions, necessary to explain some TFs binding on the DNA, and **2)** the DNA-DNA interactions through DNA folding, explaining the importance of many cis-regulatory DNA regions. Some methods, such as CellOracle, use prior knowledge on promoter-enhancer interactions. But ultimately, it only keeps a set of regulatory regions undifferentiated, which are all directly linked to TFs (i.e. instead of considering triplets – TF-enhancer-promoter). Ultimately, the single-layer network output cannot differentiate these indirect paths from the direct binding on promoters.

In contrast, the HMLN framework avoids suppressing or transforming these edges, providing more complete GRNs.

1.5. Contribution of this thesis

The main contribution of this thesis is the development of HMLN frameworks for predicting molecular mechanisms from single-cell data. These methods integrate different omics data (Chapter 2) and combine both intra- and inter-cellular interactions (Chapter 4) to identify effects and drivers of molecular perturbations across cells and tissues.

Additionally, Chapter 2 introduces a package inferring cis-regulatory DNA region networks, along with some insights about the data preprocessing effect on the predictions.

1.5.1. Main chapters

Model the molecular mechanisms of a cell – HuMMuS

Cellular behaviors emerge from a complex network of molecular interactions. Since the development of single-cell data, numerous methods have been proposed to describe cell complex interplay through gene regulatory networks. However, these methods ignore many types of interactions playing key roles in gene transcription. Chapter 2 proposes a new representation of gene regulatory mechanisms through a heterogeneous multilayer network (HMLN).

Briefly, Chapter 2 integrates the several molecular layers, which contain intra-omics co-operation. Each layer can be inferred from single-cell data (e.g., DNA region layer, gene layer) or prior knowledge (e.g., TF dimerization layer). Layers are connected through inter-omics regulations, such as DNA binding and enhancer-gene effects. Once the HMLN is obtained, it can be explored by random walk with restart to obtain pair-wise scores. These scores can summarize different mechanisms and provide outputs adapted to one's purpose, such as a TF-gene network or DNA region-gene network.

Improve the inference of cis-regulatory DNA interactions – CIRCE

During the development of HuMMuS, different computational challenges emerged due to the size of the single-cell datasets. One of the main limitations is the inference of DNA regions network from single-cell ATAC-seq, which typically contains several million connections. The state-of-the-art method, Cicero, an R package developed in 2018, has been our choice during HuMMuS development. However, it was significantly slowing the network construction while challenging memory usage. As a result, Chapter 3 presents a new implementation of Cicero's algorithm[99] in Python, which optimizes the inference of DNA regions networks.

Chapter 3 introduces both the package CIRCE and a short comparison of different data preprocessing methods. In short, it first updates the strategy of Cicero to compute pseudocells, then infers a network by covariance followed by a graphical lasso correction to sparsify the obtained network. Additionally, the impact of both CIRCE's and Cicero's preprocessing strategies, as well as of data binarization, which is often used in scATAC-seq data preparation, is explored in this Chapter.

Reconstruct molecular programs across cell types – ReCoN

Multicellular behaviors require the coordination of distinct specialized cell types. This coordination involves the exchange of signals in-between cells, also called cell communication. As a consequence, the perturbation of one cell affects neighbouring cells. Reciprocally, the environment and surrounding cells influence a cell's response. Modeling these effects is a crucial step for the understanding of complex diseases with systemic effects. Considering them will help design more specific treatments, either leveraging cell communication to exploit intercellular regulatory mechanisms (e.g., immuno-therapy) or limiting it to reduce adverse effects (e.g., in the context of chemo-therapy). Chapter 4 introduces ReCoN, a method to understand, at the molecular scale, the coordination of cells in response to external and internal perturbations.

Chapter 4 proposes to assemble a HMLN composed of several smaller cell type HMLN. The complete structure allows modeling the effect of different perturbations. It defines direct effect of a perturbation as its effect through direct receptor-binding signal translation, and indirect effect as the effects mediated by surrounding cell types, themselves reacting to the perturbation by emitting secondary messengers. It demonstrates the importance of the indirect effect in predicting the perturbation response of multiple cell types, which is often neglected compared to the direct effect.

1.5.2. Software contributions

The different developments in this thesis led to open-source python and R packages:

- The **HuMMuS package**, developed for Chapter 2, enables to build heterogeneous multilayer networks from single-cell omics data and to predict regulatory scores

through random walk with restart (RWR). The backend code of HuMMuS runs through Python (*hummuspy*) but is called through R. While designed for single-cell RNA-seq and single-cell ATAC-seq data mostly, users can integrate any additional omics of interest.

https://github.com/cantinilab/hummus 

- The **CIRCE package**, developed for Chapter 3, enables the computation of co-accessibility scores between DNA regions from single-cell ATAC data. CIRCE is based on *Cicero*'s algorithm[99], uses the *skggm*[162] implementation of graphical lasso (in C++ and cython) and is implemented for CPU parallelisation.

https://github.com/cantinilab/circe 

- The **ReCoN package**, developed for Chapter 4, enables building heterogeneous multilayer networks for molecular interaction across multiple cell types. Similarly to HuMMuS, it takes single-cell data as input and runs RWR to compute the molecular regulatory scores. When considering external molecules, ReCoN allows to weigh the contribution of direct effect (through direct receptor binding) and indirect effects (through secondary messenger of intermediary cell types).

https://github.com/cantinilab/recon 

1.5.3. List of publications

1.5.3.1. Journal articles

Published

[Chapter 2] Trimbour R., Deutschmann I. M. & Cantini L. 2023. ***Molecular mechanisms reconstruction from single-cell multi-omics data with HuMMuS.*** Bioinformatics, 2024. https://doi.org/10.1093/bioinformatics/btae143

Badia-i-Mompel P., Wessels L., Müller-Dott S., Trimbour R., Ramirez Flores O. R., Argelaguet R. & Saez-Rodriguez J. *2023.* ***Gene Regulatory Network inference in the era of single-cell multiomics.*** Nature Review Genetics **24**, 739–754 (2023). https://doi.org/10.1038/s41576-023-00618-5

Preprints

Badia-i-Mompel P., Casals-Franch R., Wessels L., Müller-Dott S., Trimbour R., Yang Y., Ramirez Flores O. R. & Saez-Rodriguez J. 2024. ***Comparison and evaluation of methods to infer gene regulatory networks from multimodal single-cell data.*** bioRxiv, 2024.12. 20.629764

https://doi.org/10.1101/2024.12.20.629764

Manuscripts in preparation

[Chapter 3] Trimbour R. & Cantini L. ***CIRCE: a fast and scalable python package to predict cis-regulatory DNA interactions from single-cell chromatin accessibility data.*** (expected in 2025)

[Chapter 4] Trimbour R., Ramirez Flores R. O., Saez-Rodriguez J., Cantini L. ***ReCoN reconstructs the molecular mechanisms coordinating multicellular programs.*** (expected in 2025)

1.5.3.2. Presentations

Oral presentations

Scientific Days (*Ecole de l'Inserm - L Bettencourt*) . . . . . Montpellier, December 2024

ECCB (*International Society for Computational Biology*) . . . . . Turku, September 2024

Jounée Boris Ephrussi (*Sorbonne University*) . . . . . Paris, Mai 2024

Computational biology Dept. Day (*Pasteur Institute*) . . . . . Paris, March 2024

Scientific Days (*Ecole de l'Inserm - L Bettencourt*) . . . . . Strasbourg, December 2023

Posters

BC^2 2023 (Swiss Institute of Bioinformatics) . . . . . Bâle, September 2023

Journée Boris Ephrusi (*Sorbonne University*) . . . . . Paris, Mai 2023

Computational biology Dept. Day (*Pasteur Institute*) . . . . . Paris, December 2022

Chapter 2

Intracellular molecular mechanisms reconstruction from single-cell multi-omics data with HuMMuS

Abstract

The molecular identity of a cell results from a complex interplay between heterogeneous molecular layers. Recent advances in single-cell sequencing technologies have opened the possibility to measure such molecular layers of regulation.

Here, we present HuMMuS, a new method for inferring regulatory mechanisms from single-cell multi-omics data. Differently from the state-of-the-art, HuMMuS captures cooperation between biological macromolecules and can easily include additional layers of molecular regulation. We benchmarked HuMMuS with respect to the state-of-the-art on both paired and unpaired multi-omics datasets. Our results proved the improvements provided by HuMMuS in terms of transcription factor (TF) targets, TF binding motifs, and regulatory regions prediction. Finally, once applied to snmC-seq, scATAC-seq, and scRNA-seq data from mouse brain cortex, HuMMuS enabled to accurately cluster scRNA profiles and to identify potential driver TFs.

HuMMuS is available at https://github.com/cantinilab/HuMMuS.

The content of this chapter was published as a journal article.

Rémi Trimbour, Ina Maria Deutschmann, Laura Cantini, *Molecular mechanisms reconstruction from single-cell multi-omics data with HuMMuS*, Bioinformatics, March 2024, https://doi.org/10.1093/bioinformatics/btae143

Contents

2.1. Introduction

Cells within a multicellular organism are remarkably heterogeneous, spanning many different molecular identities[163]. The molecular identity of a cell is the result of a complex interplay among different layers of molecular regulation, all of which can vary because of intrinsic and extrinsic factors. Recent advances in single-cell sequencing technologies have opened the possibility to measure such molecular layers of regulation, a.k.a. omics, at the resolution of the single cell. Examples of omics data currently accessible at single-cell resolution are chromatin accessibility (scATAC), methylation (snmC), expression (scRNA)[164,165]. In addition, sequencing technologies providing the joint profiling of multiple single-cell omics from the same cell have been developed[41,166,167]. Examples of them are 10xGenomics Multiome platform, jointly profiling transcriptome and chromatin accessibility from the same cell, and CITE-seq, simultaneously quantifying cell surface proteins and transcriptome within a single cell[39]. All these data provide the unprecedented opportunity to reveal how different molecular layers interact through complex regulatory mechanisms to define cell identity.

Several methods, co-analysing single-cell omics data to elucidate the regulatory mechanisms that encode cellular identities, have been recently developed[7–9,168–171]. The output of these methods are Gene Regulatory Networks (GRNs), corresponding to graphs linking transcription factors (TFs) with their inferred target genes and/or peaks[5,6,172,173]. The GRNs are obtained by all methods performing TF–peak–gene associations based on binding motif databases [e.g. JASPAR[174]], then filtered through scRNA and scATAC data analysis. All these methods ignore intra-omics cooperation between biological macromolecules, which is crucial in biology. Indeed, TFs can cooperate in the regulation of gene expression by forming dimers and multiple DNA regions can co-regulate the

expression of the same gene. In addition, state-of-the-art methods only consider TF–gene interactions present in binding motif databases and miss all those interactions that are not reported there. Furthermore, all these methods infer GRNs by integrating scRNA and scATAC data, thus ignoring all other complementary layers of molecular regulation (e.g. methylation, proteome). Finally, many methods require either paired data, or perform cell pairing before GRN inference[7,168,169,171]. This is a major limitation, as paired single-cell multi-omics data are still rare and performing cell pairing in dataset profiled from different cells forces a decrease in the size of one of the two datasets thus reducing the richness of its information content.

Here, we introduce HeterogeneoUs Multilayers for MUlti-omics Single-cell data (HuMMuS), a flexible tool based on Heterogeneous Multilayer Networks (HMLNs) to reconstruct regulatory mechanisms from multiple single-cell omics data. HuMMuS considers not only inter-omics interactions (e.g. peak–gene, TF-peak), as done by the state-of-the-art, but also intra-omics ones (e.g. peak–peak, gene–gene, TF–TF) thus allowing to capture cooperation between biological macromolecules. This inclusion of intra-omics interactions allows HuMMuS to explore new TF–gene interactions not present in binding motif databases. In addition, HuMMuS is a flexible framework that can be used both for paired and unpaired single-cell multi-omics data or easily extended to deal with additional omics data, thus not limiting the regulatory mechanisms analysis to only scRNA and scATAC, as it is currently done in the state-of-the-art.

We extensively benchmarked HuMMuS with respect to the state-of-the-art on four independent datasets of scRNA and scATAC. This benchmarking included the prediction of TF targets, TF binding regions, regulatory regions, and the association of its communities with known biological processes. Finally, by applying HuMMuS to unpaired scRNA, scATAC, and snmC data from mouse cortex, we showed that its GRN allows to accurately cluster scRNA profiles and to identify regulators relevant to mouse brain cortex.

HuMMuS is available at https://github.com/cantinilab/HuMMuS as R package, together with a tutorial for its usage.

2.2. Materials and Methods

2.2.1. HuMMuS a new tool for molecular mechanisms reconstruction from single-cell multi-omics data

We developed HeterogeneoUs Multilayers for MUlti-omics Single-cell data (HuMMuS), a new tool for regulatory mechanisms inference from single-cell multi-omics data (Figure 2.1, https://github.com/cantinilab/HuMMuS).

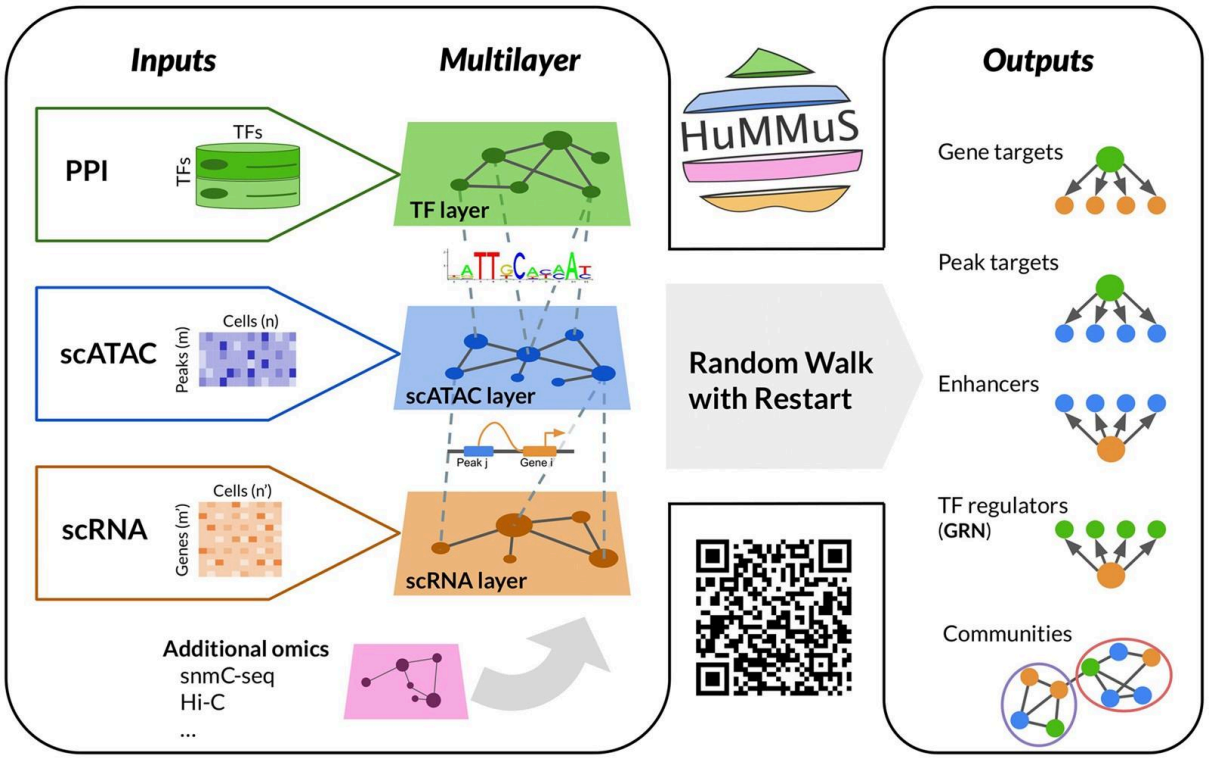


Figure 2.1. Schematic view of HuMMuS workflow.

HuMMuS is based on Heterogeneous Multilayer Networks (HMLNs). A HMLN is a network $M = (V_m, E_m, \mathbf{L})$, $m = 1, ..., M$, composed of M layers, each of them containing different nodes V_m and different intra-layer links $E_m \subseteq V_m \times V_m$. Nodes of different layers are connected by inter-layers links encoded in $\mathbf{L}$[11,146]. As summarized in Figure 2.1, we reconstruct HMLNs composed of three layers: The TF layer, containing unlinked TFs, the scATAC layer containing peak co-accessibility information inferred from scATAC data and the scRNA layer encoding transcriptional regulation inferred from scRNA data. TF interactions were not considered here to compare HuMMuS fairly with respect to the state-of-the-art. An additional version of HuMMuS, called HuMMuS+TF, is also considered in the following to test the effect of TF–TF links on the performances. For all details on the layers' construction see Supplementary Text. Of note, we here focused on this combination of omics data to not advantage HuMMuS by the additional information provided by other single-cell omics data. However, as the HMLN structure is flexible, HuMMuS can easily integrate other single-cell omics data, such as methylation (snmC) or Hi-C data, and additional information on known interactions, such as Protein-Protein interactions in the TF layer to capture TFs cooperativity. Once the HMLN is constructed, HuMMuS uses Random Walks with Restart (RWR)[146] to mine the HMLN and extract different outputs: (i) the prediction of the targets of a TF, based on RWRs starting from each TF in the TF layer and exploring the full network until the scRNA layer; (ii) the prediction of the peaks bound by a given TF, based on RWRs starting from each TF in the TF layer and exploring the scATAC layer; (iii) the prediction of the regulatory regions (proximal and distal enhancers) associated to a given gene, based on RWRs starting in each gene of the scRNA layer and

exploring the scATAC layer; (iv) the reconstruction of Gene Regulatory Networks (GRNs), based on RWRs starting in each gene of the scRNA layer and exploring the full network until the TF layer; (v) the extraction of communities in the GRN, reflecting tightly connected macromolecules in the HMLN frequently involved in the regulation of the same biological process or pathway[72]. Of note, both the prediction of TF targets (output i) and the reconstruction of the GRNs (output iv), in principle lead to a TF–gene network. The choice of reconstructing GRNs by exploring the HMLN from genes to TFs is justified by the need of having a competition among different TFs in the regulation of a gene, as done in most of the GRN inference approaches[5–9,168–170,170,171]. On the contrary, when predicting the targets of a TF, we want to treat each TF independently from the others and make genes compete among themselves.

For this reason, we obtain the output (i) by exploring the HMLN from TFs to genes. See Supplementary Table S1 and Supplementary Figure S1 for a computational comparison between the two approaches and methods for all details on the parameter choice for the RWR and the possible outputs.

Thanks to the use of a HMLN structure, HuMMuS has multiple advantages with respect to the state-of-the-art. First, it captures not only inter-omics interaction (e.g. peak–gene, TF-peak), as done by the state-of-the-art, but also intra-omics ones (e.g. peak–peak, gene–gene, TF–TF). This allows HuMMuS to capture cooperation between biological macromolecules and use it to predict, e.g. TF–gene interactions not present in binding motifs databases. In addition, HuMMuS is a flexible framework, that can be used both for paired and unpaired single-cell multi-omics data or easily extended to deal with additional omics data, thus not limiting the regulatory mechanisms analysis to only scRNA and scATAC, as it is currently done in the state-of-the-art.

In the following we extensively benchmark HuMMuS against SCENIC+, CellOracle and Pando[7–9], being the most famous published works in the field. Interestingly, CellOracle is the only existing method considering some cooperation at the peaks level. In addition, we included GENIE3[61] in the benchmark as a baseline for performances when considering scRNA alone. All the benchmarking is performed on four test cases (see Supplementary Text and Supplementary Table S2): two datasets (called in the following Chen and Liu) of human Embryonic Stem Cells (hESCs), jointly profiled for scRNA and scATAC (i.e. paired data), and two unpaired scRNA and scATAC datasets of mouse Embryonic Stem Cells (mESCs) (called in the following Duren and Semrau). For details on HuMMuS layers structure in these four datasets see Supplementary Table S3. Of note, in Duren and Semrau, being the data unpaired, the scRNA and scATAC information has been profiled from different cells all extracted from mESCs. These last two test cases thus allow to test the impact of cell pairing on the performances of the different methods. The choice of these four test cases is justified by the availability of ChIP-seq and TF perturbation experiments in hESCs and mESCs from McCalla *et al.* (2023)[172]. This additional data, already used in benchmarking works[172], allows indeed to build good ground truths for the different tests presented in the following sections.

2.3. Results

2.3.1. HuMMuS outperforms the state-of-the-art in TF target prediction

We first focused on benchmarking HuMMuS with respect to the state-of-the-art based on the quality of its TF targets predictions. This analysis has been performed on the four test cases presented above, corresponding to scRNA and scATAC profiling of hESCs and mESCs. As ground truth of the TF-targets interactions we used the intersection between ChIP-seq and TF perturbations experiments, as done in McCalla *et al.* (2023)[172]. This choice represents indeed the best estimation of TF targets we can get for real data, as it assures the presence of a binding site for the TF on the promoter of the target gene and, at the same time, a downregulation of the target gene once the TF is knocked down/out.

As described in Figure 2.2a, in each of the four test cases, HuMMuS and the other state-of-art algorithms have been independently applied, a ranking of putative targets for each TF is then identified and compared with the ground truth described above. The ranking of putative gene targets for a TF is obtained for the state-of-the art methods as the list of genes linked to the TF. The genes are ordered according to the weight of their links. For HuMMuS instead, we perform a Random Walk with Restart (RWR) starting from each TF and going across all the HMLN, thus obtaining a ranking of putative target genes based on their closeness to the TF. The overlap for all methods with the ground truth is then analyzed when cutting the ranking at different levels (3, 5, 10, 15, 20, 30, 40, 50, 75, 100).

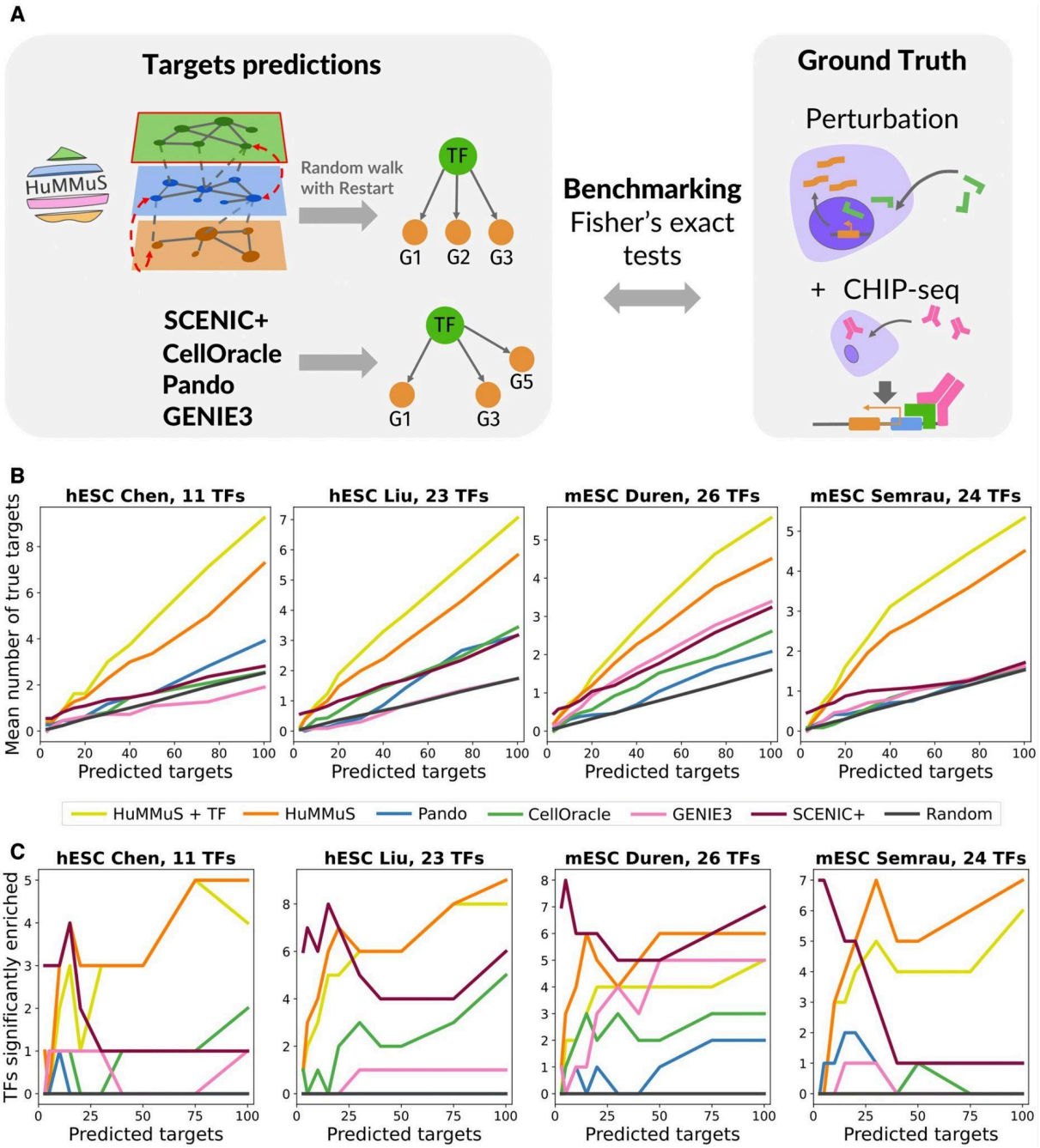


Figure 2.2. Transcription Factor (TF) targets prediction benchmarking. **A)** schematic view of the performed benchmarking. **B)** average number of correctly predicted targets per TF. **C)** number of TFs having a significant amount of correctly predicted targets (Fisher's exact test *P*-value <0.05). In (**B** and **C**) results for different methods are provided: *HuMMuS +TF*, HuMMuS, SCENIC+, Pando, CellOracle, GENIE3, and random.

As shown in Figure 2.2B, HuMMuS outperforms the state-of-the-art in all the four tested datasets at every threshold, except when focusing on the very top of the ranking (3–5 first inferred TF–gene links), where SCENIC+ shows better performances. In addition, the performances of HuMMuS get further improved once including TF–TF interactions in the network (*HuMMuS+TF*). In Semrau the results of state-of-the-art methods are close to random, here represented with a black curve. Of note, even when pairing the cells in the

two unpaired datasets, the performances observed for HuMMuS are not affected (see Supplementary Figure S2). To then test whether the observed performances were driven by a subgroup of TFs or consistent for a high number of them, we computed the number of TFs having a significant number of targets in their top predicted targets (see Supplementary Text for details). As shown in Figure 2.2C, overall, all methods get few TFs with a significant amount of correctly predicted targets. In this test too, HuMMuS gets the best performances in three out of four test cases. Taken together these two results suggest a high potential for HuMMuS in TF targets prediction.

2.3.2. HuMMuS outperforms the state-of-the-art in regulatory region identification

We then benchmarked HuMMuS with respect to the state-of-the-art based on known regulatory regions identification. This benchmark was realized in two steps: first, the ability to predict the peaks bound by a TF is tested; then, the quality of the regulatory regions (proximal and distal enhancers) predicted for each gene is evaluated. As GENIE3 does not provide any information on regulatory regions, it was excluded from this part of the benchmarking.

As shown in Figure 2.3A, to test the quality of the peaks associated with a TF, in HuMMuS we used RWRs from each TF as a proxy of the compatibility between a TF and peaks and filtered the obtained peak ranking at different levels (100%, 80%, 60%, 20%). For SCENIC+, CellOracle and Pando instead, we considered the peaks retained by the model as associated with each TF (see Supplementary Text for details). In CellOracle different peak co-accessibility correlation thresholds have been considered 0.05, 0.2, and 0.8, with the last being the default threshold. We finally compared the predictions obtained by the various methods with the ground-truth composed of ChIP-seq experiments results on the biological system under analysis (mESCs and hESCs) from Hammal *et al.* (2022)[175].

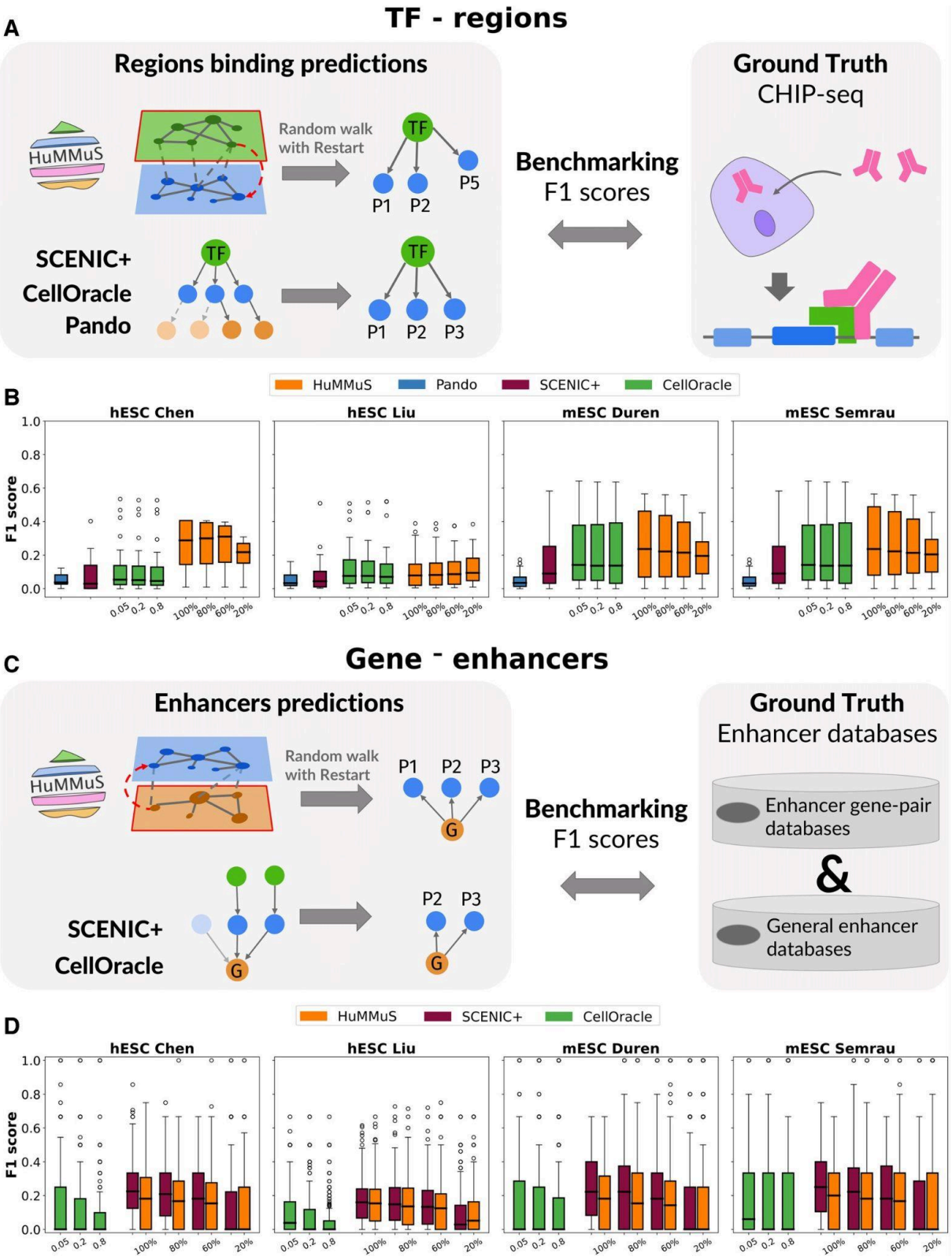


Figure 2.3. Regulatory regions benchmarking. **A)** schematic view of the benchmarking performed for TF-peak associations. **B)** F1 score of the intersection between the ground-truth TF-peak associations and those inferred by Pando, CellOracle, SCENIC+ and HuMMuS; the 100%, 80%, 60%, 20% thresholds of HuMMuS correspond to the number of nodes retained from the predictions. For CellOracle instead, 0.05, 0.2, and 0.8 correspond to the correlation thresholds of the model, with 0.8 being the default one. **C)** schematic view of the benchmarking performed for gene-peak

associations. **D)** F1 score of the intersection between the ground-truth gene-peak associations and those inferred by CellOracle, SCENIC+ and HuMMuS. In (**B**, **D**) results for different methods are provided: HuMMuS, SCENIC+, Pando, CellOracle. The thresholds are the same as those of panel (**B**).

See Supplementary Text for further details on the analysis.

Overall, as shown in Supplementary Figure S3A, HuMMuS identifies more peaks associated with a TF than alternative methods. This result is not surprising as, differently from the state-of-the-art, HuMMuS leverages all the peak layer without constraints neither on genomic windows nor on known TF motifs. This choice of considering TF-peak interactions outside of TF binding motif databases allows to include interactions that are missing in such databases and situations where, due to cooperation between TFs (e.g. condensates), there is a modification in the binding region[176,177]. More interestingly, as shown in Figure 2.3B, once checking the quality of the identified TF-peak associations based on F1 score, HuMMuS outperforms the state-of-the-art in three out of four datasets and it performs comparably to CellOracle in the fourth dataset. Regarding the percentage of true positives (Supplementary Figure S3B), HuMMuS and CellOracle are among the best performing methods in three out of four datasets, once focusing on comparable numbers of tested predictions (1%–10% filtering of HuMMuS). On the other hand, SCENIC+ is among the best performing methods in only one dataset out of four. Overall, these results suggest that considering peak co-accessibility favorably helps reconstruction of TF-peak interactions.

We then focused on the regulatory regions associated with each gene. As shown in Figure 2.3C, in HuMMuS, peaks are ranked based on the RWR starting from the gene. For CellOracle and SCENIC+ instead, the model directly provides a set of peaks associated with a gene. Regarding thresholding, HuMMuS and SCENIC+ were filtered to have a comparable number of predictions (see Supplementary Text), while CellOracle was filtered with different correlation thresholds: 0.05, 0.2, and 0.8, with the last being the default one. The obtained predictions were finally compared with a ground truth composed of gene-regulatory regions associations available from different databases[178–184]. For all details on the analysis, see Supplementary Text. GENIE3 and Pando have been excluded from this analysis as they did not provide an output allowing for this type of evaluation.

As shown in Supplementary Figure S4A, overall HuMMuS gets more enhancers associated with each gene. Again, this result is not surprising given that the intrinsic structure of HuMMuS allows it to predict new peak–gene associations, without genomic window constraints. In addition, as shown in Figure 2.3D HuMMuS and SCENIC+ comparably overperform CellOracle. Same results apply when considering the percentage of true positives (Supplementary Figure S4B). Overall, the obtained results indicate that the enhancers predicted by HuMMuS and SCENIC+ tend to more frequently reflect known ones.

Taken together these two results suggest that HuMMuS can powerfully predict regulatory regions associated with TFs or genes. Also in this case, the results observed for HuMMuS in

the two unpaired data (Duren and Semrau) are not affected by cell pairing (Supplementary Figure S5).

2.3.3. HuMMuS outperforms the state-of-the-art in the biological relevance of its gene communities

We benchmarked HuMMuS with respect to the state-of-the-art based on the biological relevance of their gene communities. Indeed, gene communities in biological graphs have been previously shown to frequently reflect known pathways and biological processes[72,185,186].

As shown in Figure 2.4A, the Louvain algorithm[55] was applied to the HuMMuS GRN and to those of the state-of-the-art and the biological relevance of the obtained communities was evaluated based on the percentage of communities enriched in pathways [KEGG[110,187] and REACTOME[188]] and Gene Ontologies[189,190]. Before running community detection, as most of the GRNs are highly dense (density > 0.8 in half of networks see Supplementary Table S4), a filtering was applied to the links to make all networks equally dense. Regarding the community detection, as the Louvain algorithm depends on the resolution parameter, we here run it with resolution varying in the range 0–2 and choose for each method the resolution giving best performances and a reasonable number of communities (≥10). See Supplementary Text for details on the analysis, Supplementary Table S5 for performances across different resolution values.

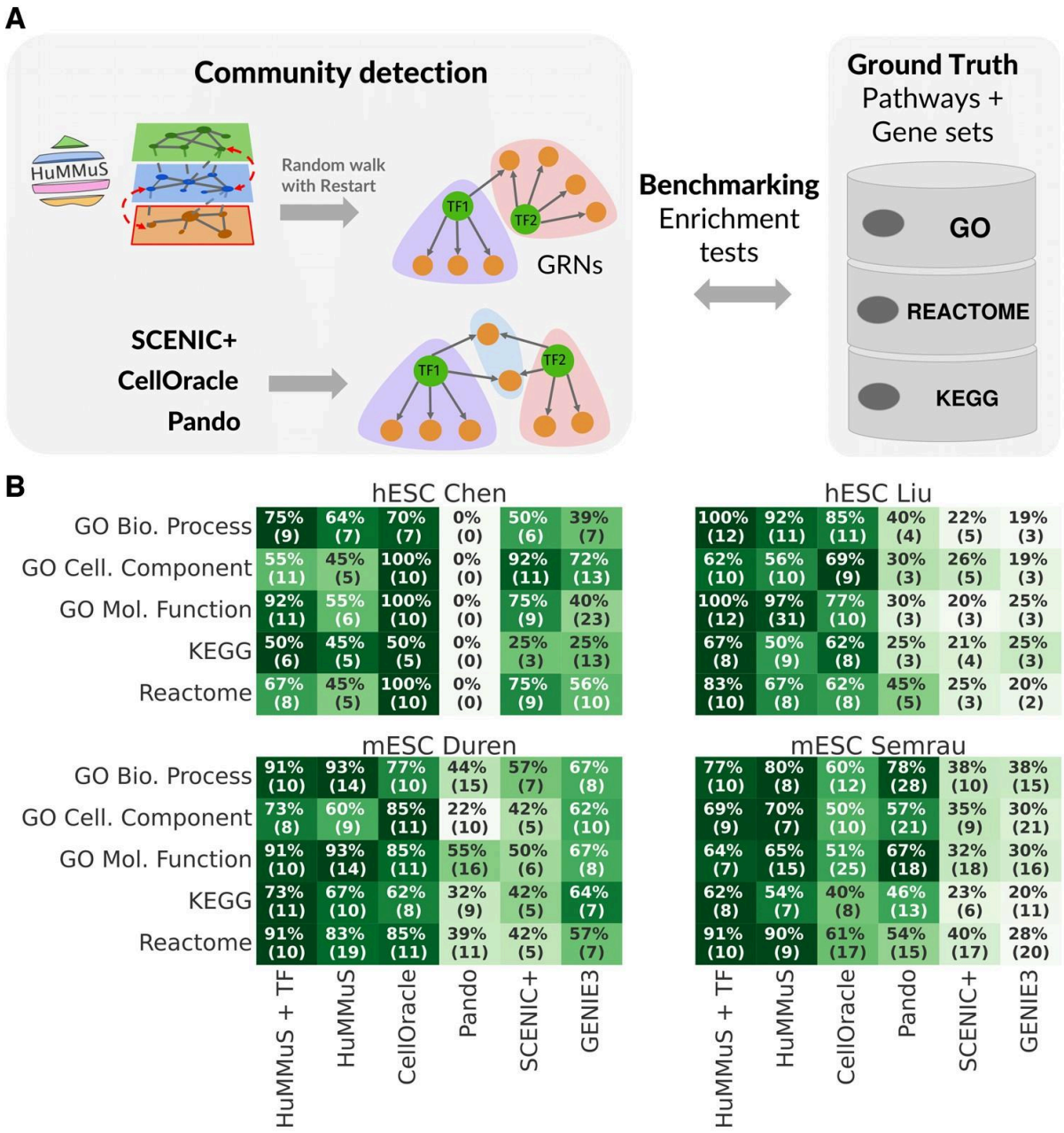


Figure 2.4. Community detection benchmarking. **A)** Schematic view of the benchmarking performed for community detection. **B)** Heatmaps of the percentage of enriched communities in each benchmarked method across the five biological databases. The values reported in the table correspond to the percentage of enriched communities, while those in parentheses are the actual number of enriched communities.

Figure 2.4B shows the results of the comparison. Regarding the number of enriched communities, all methods vary in a range of 5–31 communities, depending on the test case and the database under analysis. Concerning the enrichment in pathways and Gene Ontologies, in three out of four test cases (Liu, Duren, and Semrau), HuMMuS gets the highest percentage of enriched communities in most of the databases. Interestingly, in two out of these three datasets HuMMuS performances get even better once including TF–TF links (see HuMMuS + TF in Figure 2.4B. In the remaining test case (Chen), CellOracle gets

better results. Of note, no evident correlation emerges between the number of identified communities and the performances of the different methods (see Supplementary Table S5).

2.3.4. HuMMuS is robust to unbalanced cell type proportions across omics

Most of the state-of-the-art methods for GRN inference in single-cell multi-omics data require paired data. This requirement is due, on one side, to the use of regression models to infer the interactions, which intrinsically requires paired data, and, on the other side, to the fact that different cell type proportions might impact GRN inference. As HuMMuS is here proposed as a tool that can deal with unpaired data, we evaluated its robustness with respect to unbalanced cell type proportions across omics. For this we employed scRNA[191] and scATAC[192] data profiled from mouse cortical neurons. We only considered three cell populations: MGE, Layer 2/3 and Layer 6; corresponding to a total of 1143 cells. We then tested four scenarios (i) full datasets; (ii) half scRNA cells for Layer 2/3 and everything else unaltered; (iii) half scATAC cells for Layer 6 and everything else unaltered, and (iv) half scRNA cells for Layer 2/3, half scATAC cells for Layer 6 and everything else unaltered. We then used HuMMuS to construct GRNs for all the four scenarios and computed the Spearman correlation between the full dataset (scenario 1) and all others. As shown in Supplementary Figure S6, such correlations resulted to be 0.91–0.95, indicating a robustness of HuMMuS to different cell type propositions across different omics, thus making it particularly suitable for unpaired single-cell data.

Of note, as shown in Supplementary Figure S6, we do not observe the same robustness in the individual layers (Spearman correlations of 0.66–0.68). Thus, further suggesting that the use of RWRs helps to compensate for false and/or missing links in the single layers.

2.3.5. Challenging HuMMuS in mouse cortex profiled for scRNA, scATAC, and snmC

We finally challenged HuMMuS in the reconstruction of molecular mechanisms of the mouse brain cortex. Differently from the state-of-the-art, here for the first time we take into account three single-cell omics data: scRNA[191], scATAC[192], and snmC[193]. The data of size 55 803 cells in scRNA, 2317 cells in scATAC and 3386 cells in snmC are unpaired, obtained by profiling mouse cortical neurons.

Following the HuMMuS pipeline, we reconstructed two HMLNs, one composed of four layers (TF layer, scATAC layer, snmC layer, and scRNA layer; see Figure 2.5A and one composed of three layers (TF layer, scATAC layer, and scRNA layer). The second HMLN is intended to test the added value brought by methylation in the analysis. Then RWRs from the scRNA layer have been used to extract a GRN composed of 637 regulons, each corresponding to a TF and its associated genes ranked by the strength of association[134].

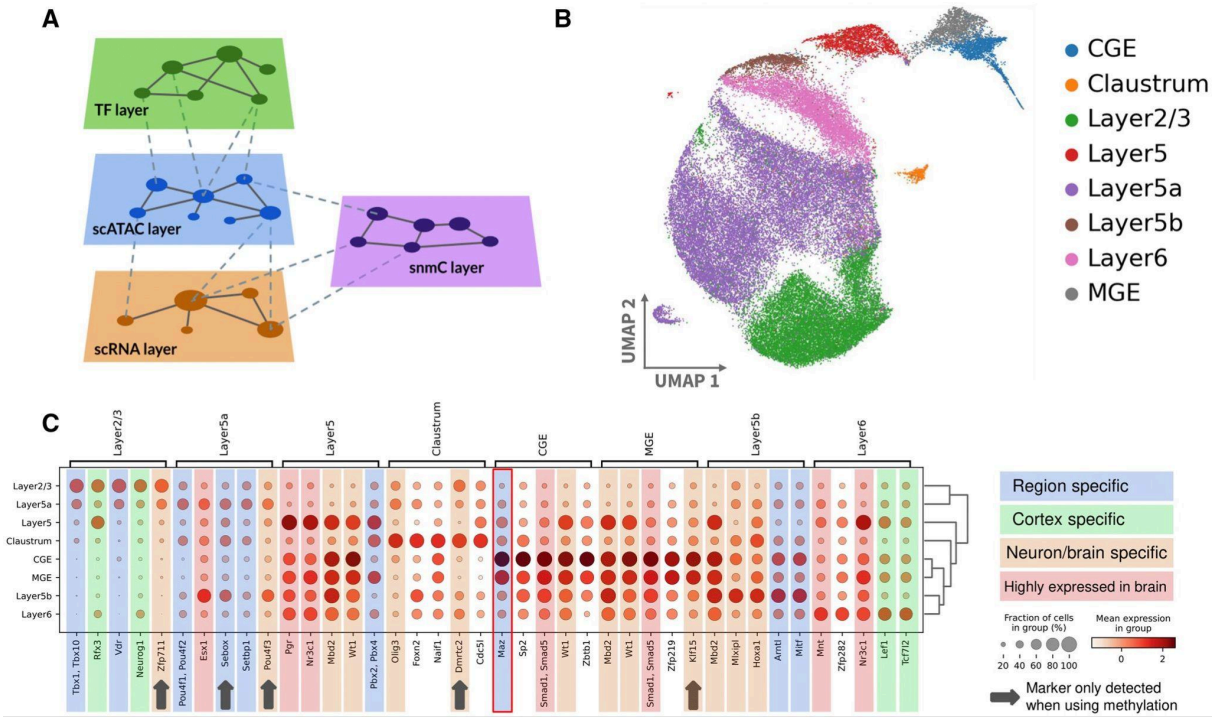


Figure 2.5. Challenging HuMMuS on scRNA, scATAC and snmC from mouse cortex. **A)** HMLN used in HuMMuS to reconstruct regulatory mechanisms from scRNA, scATAC and snmC. **B)** UMAP plot obtained from HuMMuS regulon activity. Cells are colored according to the labels present in their original publication and in previous analyses[191,194]. **C)** Heatmap of activity for the top five TFs per cell population. Colors are used to denote the type of validation available; arrows indicate TFs lost once methylation is excluded from the analysis.

As a first observation, the activity of the obtained regulons, computed according to Badia-i-Mompel *et al.* (2022)[134] and Teschendorff and Wang (2020)[195], is able to correctly cluster the cells according to their area of origin in the mouse cortex (see Figure 2.5B). This suggests that the regulons identified by HuMMuS can nicely recapitulate the known heterogeneity present between the analyzed cells and already reported in Saunders *et al.* (2018)[191] and Cao and Gao (2022)[194]. These conclusions apply with and without the additional methylation layer (see Supplementary Figure S7A).

We then focused on the results obtained with HuMMuS when methylation is included in the multilayer. We then validate in the literature the top five differentially active regulons associated to each cell population (see Figure 2.5C, Supplementary Text for details). Of the obtained 34 regulons, 76% of their TFs have an already reported association with either neurons, cortex, or brain (see Supplementary Table S6). In particular, five of them (Esx1, Pgr, Nr3C1, Smad1/5, Mnt) are reported in the Bgee database[196] as expressed in the brain. Nine of them Zfp711[197], Pou4f3[198], Mbd2[199], Wt1[200], Olig3[201], Dmrtc[202], Mlxipl[203], Hoxa1[204] are documented in publications associating them with either brain or neurons and thirteen of them [Tbx1/Tbx10[205], Rfx3[8,206], Neurog1[207], Vdr[208], Pou4f1/Pou4f2[209], Sebox[210], Setbp1[211], Pbx2/Pbx4[212], Maz[213,214], Arntl[215], Mitf[216], Lef1[217], Tcf7l2[217]] are reported in publications specifically referring to the mouse cortex. Of note, four of these TFs were also already documented to be associated to the specific region of the cortex where HuMMuS found

them to be differentially active. This is the case for Rfx3 and Neurog1, that we find associated with Layer 2/3 and that had been previously associated with this exact brain region[8,206,207,218]. In addition, Lef1 and Tcf17l2 have been documented to be associated with deep layers of the cortex and HuMMuS identifies them in layer 6[217].

Finally, HuMMuS suggests the possible regulatory role of MAZ in CGE-derived cortical inhibitory interneurons. Through bibliographic research MAZ is documented to have a role in neuronal stem cells differentiation and as potential regulator in Purkinje cells, a GABAergic inhibitory neuron population[213,214]. HuMMuS associates it to the Caudal Ganglionic Eminence (CGE) region, producing a high proportion of cortical inhibitory neurons (30%)[219]. In addition, in the top 10% of the 9341 inferred targets of MAZ, we can find Cntnap3, Dlx5, Sp9, Dlx6, Nr2c2ap, Dlx2, Arx, Grik3, all genes documented to be differentially expressed in inhibitory interneurons in The Mouse Organogenesis Cell Atlas (MOCA)[220].

Once methylation is excluded, five TFs are lost: Zfp711 in Layer 2/3, Sebox and Pou4f3 in Layer 5a, Dmrtc2 in Claustrum and Klf15 in MGE. Of note, Zfp711, Pou4f3 and Dmrtc2 had been validated on existing literature to be neuron/brain specific, while Sebox had been validated in the literature to be associated with Layer 5a neurons. The five regulons that are lost once excluding methylation are replaced by the following TFs: Trp63 for Layer 2/3, Myt1l and Olig3 for Layer 5a, Hoxb2 in Claustrum and Plag1 in MGE. Of them, Olig3[201], Plag1[221], and Myt1l[222] have been previously associated with neurons/brain and Hoxb2[223] is a known marker of Claustrum. Altogether these results suggest that methylation has an impact on the selection of the differentially active regulons associated to each cell population. However, whether such an effect is an improvement or not, depends on the cell population under analysis. Indeed, the selection of TFs in Layer 2/3 and Layer 5a improves when methylation is considered, while for Claustrum and MGE the quality of the regulons is higher when methylation is excluded.

2.4. Discussion

Cell identities result from the joint activity of different molecular layers of regulation. These molecular layers can be measured nowadays thanks to single-cell sequencing technologies, such as scRNA, scATAC, and snmC.

Different methods have been recently designed to reconstruct molecular mechanisms from different single-cell omics data. Here we proposed HeterogeneoUs Multilayers for MUlti-omics Single-cell data (HuMMuS), a flexible tool based on Heterogeneous Multilayer Networks (HMLNs) to reconstruct regulatory mechanisms from multiple single-cell omics data. HuMMuS is found to have better performance than the state-of-the-art in the prediction of TF targets, TF binding regions, regulatory regions and in the identification of biologically relevant gene communities. Once applied to the integration of scRNA, scATAC, and snmC data profiled from mouse cortex, HuMMuS identified relevant regulatory mechanisms.

Overall, the main advantages of HuMMuS are the ability to capture intra-omics cooperation between biological macromolecules and its flexibility, allowing to easily integrate additional omics or prior information (e.g. pathway databases) and to work with both paired and unpaired data.

For simplicity, we here only explored inter-layer links based on databases. However, such links could be improved in concrete biological applications considering inter-layer links derived from experimental evidence (e.g. resulting from ChIP-seq experiments instead of generalist motif databases). In addition, further developments of HuMMuS could allow to include additional single-cell data modalities, cell–cell interactions, and interactions from knowledge-based databases (e.g. REACTOME, GO). Finally, we here focused on community detection in GRNs to have a comparable output between HuMMuS and the current state-of-the-art. However, HuMMuS could further include in the future methods for community detection in HMLNs, thus allowing to detect cross-omics communities, providing a better picture of the complex interactions driving some biological processes.

Chapter 3

CIRCE: a scalable python package to predict cis-regulatory DNA interactions from single-cell chromatin accessibility data

Abstract

Single-cell chromatin-accessibility assays now profile hundreds of thousands of cells, challenging existing methods for mapping cis-regulatory interactions.

We present CIRCE, a fast and scalable python package predicting cis-regulatory DNA interactions from single-cell chromatin-accessibility data. CIRCE re-implements the Cicero workflow to analyse single-cell atlases, cutting runtime and memory use by several orders of magnitude. We also provide new options to compute metacells, grouping similar cells to reduce data sparsity.

We benchmarked CIRCE versus Cicero on two datasets, demonstrating the improvement from CIRCE's metacells' strategy. With promoter capture Hi-C data, we also evaluated how DNA interaction predictions are impacted by different preprocessings. The best performance was obtained with the single-cell input data, and we observed a negative impact of Cicero's count normalization. Finally, we demonstrated the scalability of CIRCE by processing a dataset of more than 700000 cells and 1 million DNA regions in less than an hour.

Availability and reproducibility

CIRCE is released as an open-source software under the AGPL-3.0 license. The package source code is available on GitHub at https://github.com/cantinilab/CIRCE, and documented at https://github.com/cantinilab/CIRCE_.

The code to reproduce the presented results is available as a Snakemake pipeline at https://github.com/cantinilab/circe_reproducibility.

The contents of this chapter are under preparation for an article submission.

Contents

3.1. Introduction

Cis-regulatory elements can regulate genes located hundreds of thousands of base pairs away through DNA folding, which brings distal regulatory elements into close spatial arrangement with the gene promoter region to facilitate interactions[95,224]. The regulation of gene expression results from complex and dynamic interaction between all these regulatory regions.

Single-cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) is a powerful technique for studying chromatin accessibility in individual cells. The chromatin accessibility profiles measured allow to understand the role of specific DNA elements, such as enhancers, to define cell state heterogeneity and regulate gene expression[173].

While single-cell ATAC-seq provides insights into chromatin accessibility and potential interactions with transcription factors and the transcriptional machinery, identifying regulatory and interacting regions from chromatin accessibility remains challenging, as it does not inform about DNA conformation.

Several methods took interest in this challenge and propose to infer gene - enhancer connections from single-cell ATAC-seq data alone[33,99,101], single-cell multi-omics data [8,10,225,226], or combining ATAC with other input data types, such as Hi-C data or DNA sequences[227–229]. Among them, Cicero, an algorithm developed for analysing single-cell ATAC-seq data[99], has been widely adopted to uncover cis-regulatory DNA interactions and gene regulatory mechanisms[9,10]. Cicero aims to construct a global map of cis-regulatory interactions considering the distance between regions and the technical effects in measurements.

Cicero was developed when single-cell datasets were relatively small and optimal preprocessing strategies had not been well established. Cicero shows limited resource usage performances when applied to very large high-resolution datasets produced nowadays (e.g. hundred of thousands of cells)[230–232]. Its single-cell preprocessing also relies by default on a clustering of cells from UMAP/t-SNE spaces, which do not accurately preserve distances between observations[233,234]. Furthermore, Cicero is only available as an R package and integrating it in a python workflow, in particular in the broadly used scverse ecosystem, requires hybrid environments and additional effort.

To overcome these limitations, we here introduce CIRCE, a fast and scalable python package to analyse single-cell ATAC datasets and atlases. We update metacells computation strategy and preprocessing following recent literature guidelines. CIRCE re-implements the algorithm proposed in [99] and implemented in the R package Cicero cole-trapnell-lab.github.io/cicero-release. CIRCE allows users to integrate cis-regulatory network computation in scverse workflows[235], and considerably improves computational resources usage for all sizes of datasets. CIRCE runs ~150 faster, allowing CPU parallelisation, and uses significantly less memory.

We also provide some insights on the impact of input data preparation. Aggregating highly similar cells into “metacells” can mitigate the sparsity and sampling noise inherent to single-cell profiles such as scRNA-seq, allowing also to reduce dataset size[236,237]. Because of its extreme sparsity, scATAC is often analysed after metacells computation[99,238]. We compared cis-regulatory interaction predictions from single-cell versus metacell inputs, as well as binarized versus un-binarized counts. Using promoter capture Hi-C data, we observed better predictions from CIRCE metacells than Cicero metacells. We however show that using directly single-cell input data led to even better predictions despite their higher sparsity. While metacells are particularly interesting to condensate information from very large datasets, there could be a trade-off between resource usage and small performance improvement. CIRCE offers both a better performing strategy for metacell computation, and the capacity to use single-cell as input up to atlas-sized dataset by greatly improving the processing speed and memory usage.

3.2. Implementation

3.2.1. Workflow and parameters

The methodology of Cicero includes grouping highly similar cells as metacells, before computing the co-accessibility scores[99]. In its last implementation, Cicero proposes by default to compute latent semantic indexing (LSI) reduction from the binarized input data, projecting the cells into few region-topics/dimensions space summarizing the main chromatin patterns. Cicero then computes a second dimensionality reduction on this space (UMAP or t-SNE), from which it calculates the nearest neighbours. It has now been demonstrated that t-SNE or UMAP spaces, especially when keeping very few dimensions, do not conserve distance between observations and thus are not adapted to compute

nearest neighbours[233]. In CIRCE, we propose by default to group the cells directly on a reduction space generated with LSI, while still providing the option to compute an UMAP reduction and re-use their clustering strategy.

The following step of Cicero consists in calculating and correcting covariance between DNA regions into co-accessibility scores. CIRCE reimplements the same algorithm proposed to compute co-accessibility, which uses Graphical Lasso to impose a distance penalization on the covariance of DNA regions (see Figure 3.1.a). To reduce computational complexity, these calculations are performed within a sliding window, where the window size corresponds to the maximum distance considered for cis-interactions. This limit is by default 500kb for human and mouse cis-regulatory interactions in Cicero, a distance also used as a window limit around promoters to find distal enhancers in several other works[179,239]. Distance penalties are computed between every pair of DNA regions, according to the following formula $\rho_{ij} = (1 - d_{ij}^{-s}) \times \alpha$, with d the distance between two regions.

The scaling exponent s of the power-law function estimates the global decay in contact frequency. It has been estimated at 0.75 for the “tension globule” polymer model in human and mouse[240], and a value of 0.85 has been proposed for Drosophila[241]. The parameter α, proportional to the penalty, enforces the sparsity of the final result. It is estimated by selecting random windows and calculating over them the lowest value such that less than 5% of the long range pair of DNA regions in the window (usually >250kb distant DNA regions) and 80% of all pairs are non-zero entries. To simplify the choice of the species-specific parameters in CIRCE (s, window size, distance of long and short range interactions), users can also specify the organism corresponding to their data to select automatically the default literature-based values.

Finally, cis-co-accessibility networks (CCANs) are defined as described in [99], using the Louvain clustering method[55] and the same default parameters.

3.2.2. Implementation

The most popular Graphical Lasso model (the core of CICERO) in python is hosted in the scikit-learn[242] package, but it does not allow pairwise matrix penalty. Consequently, we use the skggm package[243], implementing the QUIC algorithm in a scikit-learn compatible package and offering more freedom on model penalization. CCANs identification is based on the Louvain clustering method implemented in the NetworkX package[244].

CIRCE works directly on AnnData objects[245], making it fully compatible with any scVerse package[235]. All results are stored as a sparse matrix for memory usage optimization, directly in the ‘.obsp’ slot of your input AnnData object, and can easily be extracted as a readable table.

3.3. Performances and comparison with Cicero

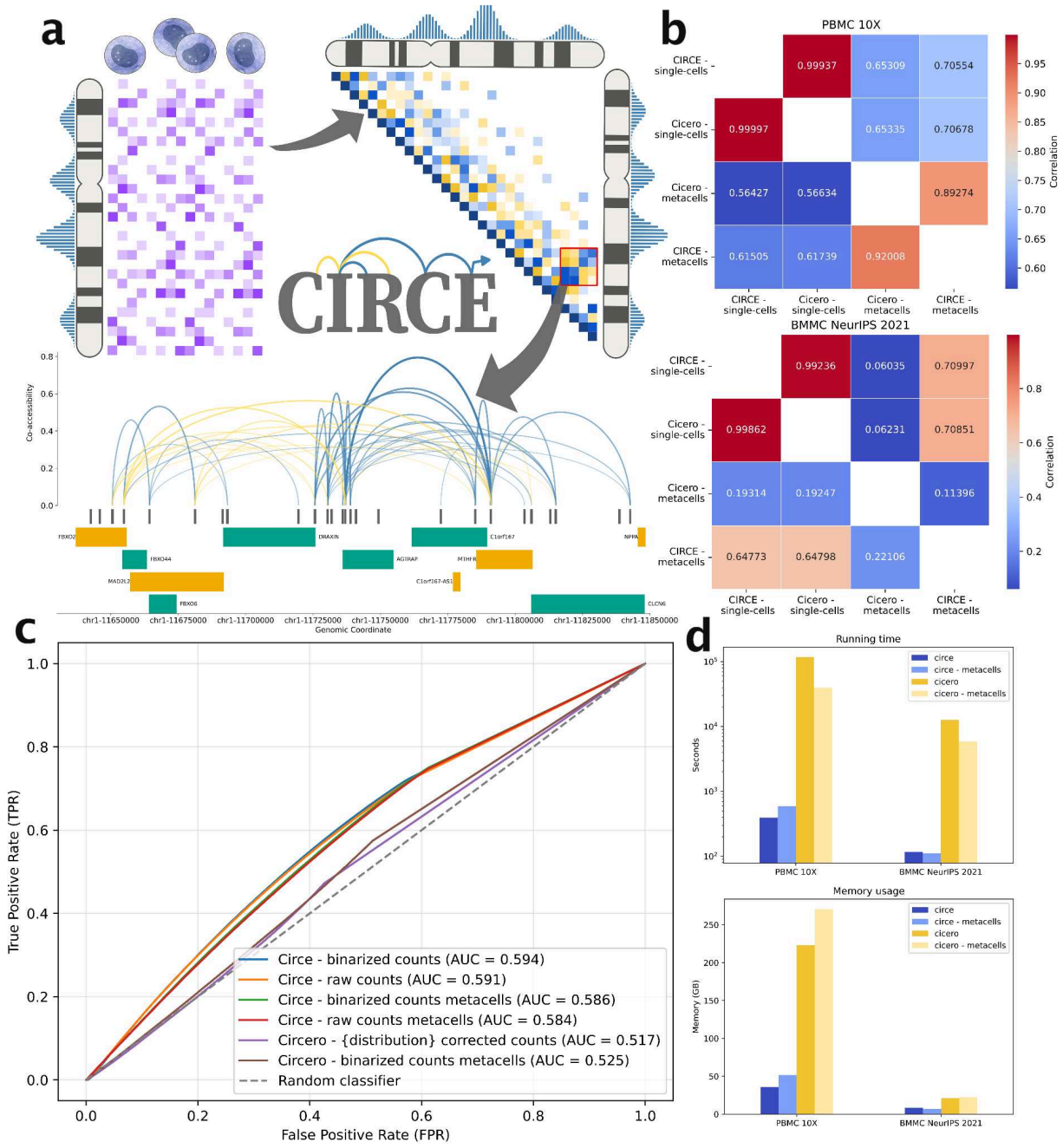


Figure 3.1. CIRCE workflow and performances. a) CIRCE workflow overview. From scATAC-seq data, co-accessibility scores are defined as regularized covariance values between DNA regions, using a graphical lasso model and pairwise distance penalisations. Cis-coaccessible networks (CCANs) can then be extracted with the Louvain community detection method, as modules of DNA regions with high absolute co-accessibility scores. Co-accessibility scores in a CCANs or a DNA window of interest can then be visualised with different graphical options. **b)** Correlation values between the networks obtained from CIRCE, Cicero, and metacells computation on both the BMMC and the PBMC datasets. The upper triangle of the heatmap contains the Spearman correlation, while the lower triangle contains the Pearson correlation. Colour gradient illustrates the correlation values. **c)** ROC curves on the recovering DNA region - promoter interactions obtained from PC-HiC dataset. Different preprocessed inputs are evaluated for each method: CIRCE from raw single-cell counts, binarized single-cell counts, and metacell counts from LSI dimensionality reduction space

on both single-cell count matrices, and Cicero on the corrected count matrices and the metacell count matrices obtained from the binarized corrected counts. **d)** Running-time and memory-usage of CIRCE and Cicero when running on the single-cell or computing metacells.

We compared CIRCE and Cicero with and without generating metacells, on a BMMC dataset (1,771 cells and 110,235 peaks), and a BPMC dataset (9,631 cells and 215,676 peaks). We demonstrated that both implementations were returning very similar results from the same input, then explored the impact of respective preprocessing strategies on inferring enhancer - promoter interactions.

From the same single-cell ATAC-seq input data, we observed almost identical results, with respective Spearman and Pearson correlations of 0.9923 and 0.9986 in the BMMC datasets, and of 0.9993 and 0.9999 on the PBMC dataset (Figure 3.1b). The small difference can be explained from the stochasticity in the algorithms that selects random DNA windows to estimate the parameter α (see Implementation section).

However, we observed substantial differences when using the respective metacell strategies of CIRCE and Cicero. First, CIRCE's metacells were still relatively highly correlated with single-cell inferred networks, with respective Spearman and Pearson correlations of 0.71 and 0.65 on the BMMC dataset, and 0.71 and 0.62 on the PBMC datasets. In contrast, the correlations with Cicero's metacells were lower (Figure 3.1b), with particularly high variability between the datasets. On the BMMC dataset, the correlations with the single-cell data run was especially low, (i.e. Spearman and Pearson correlations respectively 0.06 and 0.19), while the decrease was less marked on the PBMC dataset (i.e. Spearman and Pearson correlations of 0.65 and 0.57).

After observing such differences in the predicted DNA region interactions, we decided to evaluate the networks generated to identify the best strategy. Using a Promoter Capture Hi-C dataset[246], we measured how well each method was recovering enhancer- promoter interactions (Figure 3.1c). Surprisingly, the networks with the highest AUCs were obtained from the binarized single-cell count matrix and the un-binarized count single-cell matrix (respectively 0.594 and 0.591). In comparison, the predictions from the default corrected counts with Cicero's function make_atac_cds had an AUC of 0.517. We tested the default metacell strategy of CIRCE on both raw and binarized counts, both leading to a small decrease of the AUC value, with again the binarized and un-binarized count performing almost identically (0.586 compared to 0.584). The prediction from Cicero's metacells gave an AUC of 0.525, or a small improvement compared to predictions from Cicero's processing single-cell input. Overall, we observed better enhancer - promoter interaction predictions when using the single-cell matrix directly, and without Cicero's count correction. This result suggests that Cicero's count correction might not be the best preprocessing for inferring co-accessibility, while simple binarization does not particularly improve the predictions, as already suggested in [247]. Additionally, while computing metacells can be interesting to reduce the computational time, we observed a small drop in the performances with both tested default strategies.

We also compared running time and memory usage (Figure 3.1d) for CIRCE and Cicero on both datasets, with and without computing metacells. On average, CIRCE ran almost 150 times faster and was using 5.2 times less memory. On the biggest dataset (PBMC 10X - single-cell data), CIRCE ran in 6 min 34 seconds instead of 1 days and 8 hours for Cicero, and used 35.8 Gb of RAM instead of 223.1 Gb.

Finally, to demonstrate CIRCE's scalability and the new possibility it offers, we analysed a human atlas of fetal single-cell chromatin accessibility[232]. CIRCE processed the 720,616 cells and 1,041,455 DNA regions into a co-accessible in 42 min on a HPC. It identified 84,479,373 pairs of co-accessible regions. Since most of the computational time is actually spent in extracting small DNA windows from the initial input (columns of the AnnData input), we recommend to use the csc_matrix format of scipy when analysing huge atlas since it is optimized for column extractions.

3.4. Conclusion

Here we present CIRCE, a python package to predict cis-regulatory DNA interactions from single-cell chromatin accessibility data. CIRCE offers a performant reimplementation of Cicero, adapted to the new challenges of recent single-cell datasets and compatible with the scverse environment. While providing very similar results to Cicero, CIRCE runs much faster and uses less memory, and can highly simplify integration of cis-regulatory DNA interaction networks in python workflows.

We also provide new insights on the preprocessing of single-cell ATAC and the use of metacells to infer co-accessibility with this algorithm. We reveal that binarization is unnecessary and that count normalization can be counter-productive. While different metacells strategies might actually improve the results by reducing noise, we didn't observe such improvement from the tested strategies. They however allow to conserve very good performances with a lower number of cells. CIRCE's optimized implementation will also allow analysing much bigger datasets without having to reduce the number of cells through aggregation.

3.5. Methods

3.5.1. Data preparation

BMMC NeurIPS dataset

The BMMC multiome used here was generated for Open Problems challenge[248] and is accessible under the GEO accession number GSE194122. We extracted the smallest batch of 1,771 cells and kept all the peaks expressed in at least one cell (110,235 peaks).

PBMC dataset

A PBMC multiome dataset was obtained from https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-1-0-0. From the fragment files, we applied the same dataset preparation as in GRETA[249]. The peak calling and merging was done with snapatac2[250]), through the function snap.tl.macs3, snap.tl.merge_peaks and snap.pp.make_preaks_matrix. Only the cells expressing more than 100 genes in the scRNA-seq data were then kept. The final dataset contained 9,631 cells and 215,676 peaks.

PC-HiC dataset

The promoter capture HiC (PCHiC) dataset used to evaluate DNA region interactions prediction in the PBMC dataset is described in [246]. The processed interaction table is available as "PCHiC_peak_matrix_cutoff5.tsv" in the supplemental data S1: https://ars.els-cdn.com/content/image/1-s2.0-S0092867416313228-mmc4.zip.

Fetal human single-cell chromatin accessibility Atlas

The scATAC-seq atlas[251] was downloaded already preprocessed at https://scglue.readthedocs.io/en/latest/data.html under the name Domcke-2020.h5ad.

3.5.2. Reproducibility

The experiments are implemented as a snakemake pipeline containing all the code to reproduce the experiments at https://github.com/cantinilab/circe_reproducibility. All the computations were realised on a HPC equipped Linux Red Hat 8.8 and 2 AMD EPYC 7552 48-Core processors. For the benchmarking, CIRCE and Cicero were executed from their respective singularity containers (available in the reproducibility repository). Resources usage was limited to 20 cores and 430 Gb of RAM through snakemake rule configuration.

Chapter 4

ReCoN reconstructs the molecular mechanisms coordinating multicellular programs

Abstract

In multicellular organisms, cells coordinate to provide a systemic and coherent response to perturbations. However, cells can present very heterogeneous adaptations depending on their precise roles and locations. These complex behaviors emerge from a convoluted interplay between intercellular signals and intracellular regulations. Single-cell technologies opened the possibilities of measuring both coordination and heterogeneity, which led to the inference of gene regulatory networks and cell communication networks. However, most methods focus solely on one of these aspects, only partially recovering in vivo and multicellular behaviors.

We here introduce ReCoN, a framework combining gene regulations and cell communication to provide insights on multicellular coordination. First, ReCoN reconstructs a heterogeneous multilayer network containing both cell type subnetworks and ligand-receptor interactions inferred from single-cell data. Through random walk with restart explorations, ReCoN then infers the response of each cell type to a cell-specific perturbation or an external molecule, such as a gene knock-out or a cytokine respectively. It can also retrieve the molecules driving specific tissue state, such as cell type differentiations or broader systemic disease changes.

ReCoN was evaluated on predicting in vivo response of immune cell types to different cytokines and on recovering cardiac cell type response in heart failure, a condition involving complex multicellular crossplays. Notably, it highlighted the role of indirect effects, where cells emit secondary messengers in response to the initial perturbation to coordinate multicellular transcriptomic responses. ReCoN proposes an actionable modeling of multicellular systems. For example, ReCoN can help design patient-specific molecular treatments by building individual molecular models and by assessing the cellular selectivity of these treatments in vivo.

Keywords: gene regulatory network, cell communication, tissular coordination, molecular analysis, multi-omic, single cell, method, spatial transcriptomics

The contents of this chapter are under preparation as a journal article.

Rémi Trimbour[1], Ricardo O. Ramirez Flores[2,3], Julio Saez-Rodriguez[2,3,#], Laura Cantini[1,#], *ReCoN reconstructs the molecular mechanisms coordinating multicellular programs.*

[1] Institut Pasteur, Université Paris Cité, CNRS UMR 3738, Machine Learning for Integrative Genomics Group, F-75015 Paris, France

[2] 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), Hinxton, Cambridgeshire, U.K.

[#] These two authors jointly supervised this work.

Content

4.1. Introduction and background

The cell's identity depends on internal molecular mechanisms and on the environment[252–254]. In multicellular organisms, the environment notably includes surrounding cells, sending different signals. This organisation alters how cells respond to molecular perturbations compared to isolated settings. The complex coordination that emerges, coupled with intrinsic cell diversity, shapes tissue functions and multicellular programs. Reciprocally, the effect of a perturbation in vivo extends beyond the directly targeted cell, influencing surrounding populations through cascades of molecular interactions. However, conventional perturbation experiments focus on single cell types ignoring the complex intercellular signaling networks governing tissue-level outcomes. Overall, a perturbation influences cells both directly through receptor-mediated pathways and indirectly, via effects on other cells that in turn send secondary signals[255–257]. These indirect routes can significantly shape the final transcriptional response, especially in systemic therapeutic interventions where all cells are simultaneously exposed. Capturing both effects is essential for accurate modeling of biological responses.

Recent computational methods such as MOFAcell[17], DIALOGUE[18], and single-cell interpretable tensor decomposition (scITD)[258] have been developed to characterize how multiple cell types coordinate their gene expression in tissue-level responses. These approaches decompose multi-sample single-cell or spatial transcriptomic data to uncover latent factors. These factors, or multicellular programs, represent coordinated gene expression patterns across distinct cell populations within a tissue, explaining the variability of distinct samples. While these methods excel at describing what tissue-wide gene programs change in concert, they offer limited insight into how such programs are orchestrated at the molecular level. To partially address this mechanistic gap, cell–cell communication inference tools can predict the intercellular signals through static ligand - receptor bindings[120–122]. However, they do not model the upstream and downstream intracellular cascades, nor explain how such signals are integrated and propagated into the complete multicellular transcriptional response. In other words, communication analyses identify who is signaling to whom but not how or what cellular programs are actually coordinat.

Altogether, both sets of methods share a common goal of elucidating multicellular transcriptional coordination, but they diverge in focus: the former delineates broad tissue-wide expression patterns, whereas the latter infers specific intercellular cues underlying those patterns. Moreover, they do not provide a complete picture of the tissue coordination, jointly missing in intracellular regulations. Bridging these complementary perspectives is crucial for fully understanding complex tissue-level responses to perturbations.

We here present ReCoN (Regulatory and Communication Networks), a computational framework designed to model multicellular coordination and responses to perturbation by integrating intracellular mechanisms and cell communication. ReCoN combines tissue- or cell type-specific gene regulatory networks (GRNs) with inferred ligand–receptor communication graphs to predict both direct and propagated transcriptional effects across a tissue. The explicit and detailed network in ReCoN, explored by random walk with restart, allows tracing of how a perturbation spreads through interacting cell types. Unlike prior tools, ReCoN explicitly formulates the direct and the indirect effect, allowing to weigh their contribution.

We demonstrate that ReCoN outperforms existing baselines in both observational and interventional contexts, leveraging the indirect effect described above. We first predicted the cell type responses to different cytokines from the Immune Dictionary, an in vivo perturbation dataset. We also showed that ReCoN can be applied to more complex contexts, retrieving genes relevant for heart failure (HF) from a combination of ligands involved in their coordination. We then explored the regulatory mechanisms behind cardiac fibrosis and HF. We started at the intracellular scale, finding transcription factors and receptors regulating cardiac fibrosis. We then identified biological programs in other cardiac cell types coordinating with fibroblasts upstream and downstream of fibrosis.

4.2. Results

4.2.1. ReCoN: a package to extract molecular mechanisms in multicellular systems

Understanding how cells coordinate distinct molecular programs requires bridging intracellular regulation with intercellular signaling. We present here ReCoN (Regulatory mechanism and Cell Communication Network), a network-based approach to integrate both scales and model multicellular molecular coordination.

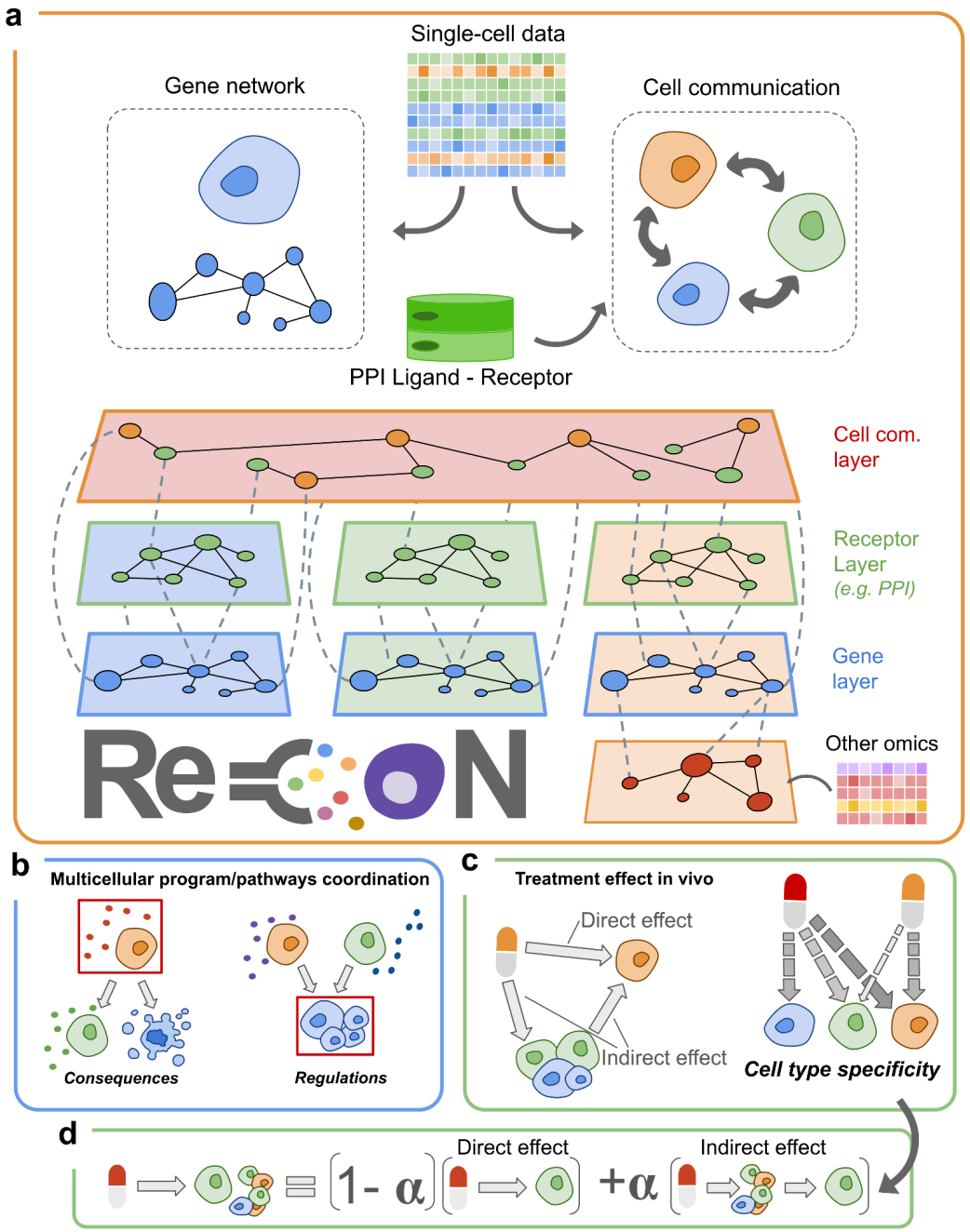


Figure 4.1. An overview of the ReCoN framework and applications. a) General HMLN structure used in ReCoN to connect intracellular network layers and a cell communication layer. Each cell can

contain a different number of layers, illustrated by the asymmetric structure presented. **b)** ReCoN can be used to infer both the upstream driver and downstream target cross-cellular signals of a profile or pathway of interest (grey frame). **c)** ReCoN decomposes the effect of a molecule into a direct and an indirect effect. The direct effect on a cell corresponds to the initial signal transduction, while the indirect one occurs in response to the surrounding cells' signals. **d)** The molecular changes in a cell type after a treatment in vivo result from both the direct and indirect effects. Their contributions are weighted using an α coefficient, and the impact of each cell type i is also modulated by $B_{i \to j}$ coefficients.

ReCoN is a heterogeneous multilayer network framework

ReCoN is a two-step framework: It first builds a multicellular network from single-cell data, then uses a random walk with restart algorithm, MultiXrank[146], to explore how signals propagate within and between cell types.

The multicellular network is based on a heterogeneous multilayer structure (HMLN), which allows the integration of multiple types of biological graphs[10,146,185,259,260] (Figure 4.1a). Formally, this network $N = (V_m, E_m, L)$, $m = 1, ..., M$, is composed of M layers containing different nodes V_m and different intra-layer links E_m. Nodes of different layers are connected by inter-layer links encoded in L[11,261]. Inside this structure, each cell type is represented by a dedicated subnetwork, called a **cell type multilayer**, while cell-cell interactions are contained in a shared **intercellular layer**. Together, they form a unified network of intracellular signaling and cell-cell communication.

The cell type multilayers include two main layers: a gene regulatory layer and a receptor layer. The gene regulatory layer contains transcription factors and their target genes, inferred from scRNA-seq or single-cell multi-omics data. The receptor layer represents membrane proteins and can optionally include receptor–receptor interactions such as dimerization. These two layers are connected by a bipartite graph that links receptors to the genes they influence (see Methods). Each cell type multilayer is constructed independently, making it possible to capture cell type–specific regulatory mechanisms. It is possible to include additional omics layers that would be relevant to a specific context, such as lipidomics or metabolomics networks (Supplementary Notes 1).

The cell type multilayers are connected by a cell communication layer. This layer contains the ligands and receptors of each cell type based on their expression, and models intercellular signaling. Ligands from one cell can interact with receptors in another, forming a tissue-wide signaling network. This network is inferred from scRNA-seq data and integrated into the overall network through two bipartite graphs for each cell type multilayer (see Methods). One connects matching receptors, and the other connecst the ligand to their corresponding genes

Once the network is assembled, ReCoN aims to find the regulators and downstream targets of a molecular signal (Figure 4.1b). ReCoN measures signal propagation through random walk with restart (RWR)[146] (see Methods). This algorithm begins from one or more

molecules of interest, such as a gene, receptor, or pathway, and computes a probability distribution over the entire network. This distribution reflects how strongly each node is connected to the input. The random walk balances steps within and between layers using a transition matrix, exploring the complete multicellular network. General transition matrices define the type of exploration, for discovering regulators or downstream targets (see Methods).

4.2.1.1. ReCoN defines the direct and indirect effects of a perturbation

ReCoN distinguishes between two types of regulatory effects: direct and indirect (see Figure 4.1c). The direct effect captures the influence of the initial signal on a given cell type via its intracellular signal transduction. The indirect effect refers to the influence of intercellular interactions. Surrounding cells are also stimulated by any molecule (e.g., cytokine) and, in turn, release ligands or signals. Such indirect signaling can then affect the target cell through cell–cell communication (CCC).

This distinction is formalized through two types of random walk with restart (RWR) on the heterogeneous multilayer network (see Figure 4.1d). The key difference between them lies in the ability to transition from intracellular layers to the cell communication layer. A parameter γ controls this transition probability, effectively tuning how much influence a cell type has on its surrounding environment.

Formally, the direct effect (M_D) is computed by running a RWR on the full network G, starting from the input molecular profile of interest M_0. This input is a vector encoding the importance of each input node, as the probability of restarting the RWR from them. In this RWR, transitions from intracellular to intercellular layers are blocked ($\gamma = 0$), to capture only the intracellular effects of the input, as shown in (*4.2*):

$$M_D = RWR_{direct}(M_0, \gamma_{direct}, G) \quad \textbf{(4.1)}$$

The result of this step is a vector M_D, representing the distribution of the direct effect across the network. This vector also serves as the input for computing the indirect effects, which model the downstream signaling between cells.

Indirect effects are calculated using additional RWRs that allow transitions between intracellular and cell-communication layers ($\gamma_{indirect} = 0.5$). For each cell type i, ReCoN identifies the genes coding for ligands reached through the direct effect, $mask_i(M_D)$. A RWR is then run from these ligands, capturing the impact of that cell type over the others. Each of these indirect effects is weighted by a cell type-specific influence vector B_i, and the results are aggregated across all emitter cell types.

Finally, the overall system response M_T is obtained by a convex combination of the direct effect and the indirect effects, as shown in (*4.2*):

$$M_T = \alpha \cdot M_D + (1 - \alpha) \cdot \sum B_i \odot RWR_{indirect}(mask_i(M_D), \gamma_{indirect}, G) \quad \textbf{(4.2)}$$

In conclusion, ReCoN aims to bridge gene regulatory networks and cell communication to predict the effect of molecular perturbations. It is, to our knowledge, the first work quantifying and weighing their direct and indirect effects in multicellular systems. ReCoN's HMLN structure leverages both prior knowledge and contextualised networks from single-cell multi-omics data to find new and specific interactions. Through its modular structure, cell types and omics can be easily added to match the hypothesis and environment of interest. Finally, ReCoN exploration can take several molecules as input, modelling their joint effect.

4.2.2. ReCoN predicts cytokine treatment responses in murine lymph nodes

In vivo tissue studies provide a biologically relevant setting to evaluate such models, as they reflect the coordinated interactions among diverse cell types. We therefore evaluated ReCoN performance on the Immune Dictionary dataset, which profiles murine lymph node responses to a panel of cytokine treatments administered in vivo[21]. Transcriptomes from 17 immune cell types were sequenced, offering a rich opportunity to study cell type responses coordination within a complex tissue (see Figure 4.2a).

Since this dataset only contains scRNA-seq data and very few unperturbed cells, we used external data to build the cell type multilayers (see Figure 4.2b). The gene regulatory network was inferred from both a lymph node scRNA-seq dataset and a splenic scATAC-seq dataset from the same mouse inbred strain. The cell communication network was inferred from the Immune Dictionary cells, treated with phosphate-buffered saline (PBS), used as a negative control.

4.2.2.1. Gene regulation and cell communication networks are both informative

To assess the contribution of each context-specific information in ReCoN, we benchmarked four variations of the ReCoN methodology (see Figure 4.2c). These models illustrate the progressive integration of regulatory and communication layers in ReCoN.

We also compared the ligand–gene prior knowledge network (ligand-PKN) from NicheNet[22,262], which has been used to derive the receptor-gene bipartite for ReCoN's models (see Methods). The first ReCoN variation tested, ReCoN-no-context, includes only receptor–gene links derived from the ligand-PKN. It thus ignores GRNs and CCC to isolate the contribution of the context-independent receptor-gene PKN. This minimal configuration also serves to assess how much predictive signal was lost from the ligand-PKN. ReCoN-grn includes a context-specific GRN built from single-cell RNA-seq data from lymph nodes and ATAC-seq data from the spleen. The GRN is shared across all cell types, and no CCC network is included. This model represents the direct effects of cytokine perturbations while adding tissue-specific regulatory context. ReCoN-generic adds a

generic CCC network on top of the GRN, connecting all possible ligand–receptor pairs irrespective of cell type expression. It highlights the contribution of indirect effects arising from cell communication without personalized interaction specificity. Finally, the full ReCoN model incorporates both context-specific GRNs and a personalized CCC network based on cell type-specific ligand and receptor expression. This model captures how different cell types contribute to each other's regulation through expressed ligands and receptors, reflecting multicellular tissue organization.

We first evaluate the models' ability to predict the transcriptomic changes of each cell type individually, through cell type gene rankings (see Supplementary Figure 1). We then verify the models' ability to combine cell type scores across the tissue through a multicellular ranking (see Figure 4.2d). This multicellular evaluation tests whether a model can prioritize strongly perturbed genes from different cell types without bias toward one of them. It thus considers its capacity to combine their different score distributions, even if several cell types present very different response profiles.

The performances of the models are reported as AUROCs. As expected, ReCoN-no-context presented the worst performances in both the individual cell type and multicellular predictions. It underperformed compared to the Nichenet PKN model, with associated Mann-Whitney U test P-values of 3.32e-2 and 6.53e-2, respectively. This illustrates the initial loss of information when inferring the receptor-gene links from the ligand-PKN. The ReCoN-grn outperformed ReCoN-no-context on both metrics (Mann-Whitney U test P-values of 8.13e-4 and 9.32e-3). This significant improvement demonstrates the contribution of the personalized GRN in predicting in-tissue cell type perturbations. It additionally performed slightly better than the ligand-PKN model, thus compensating for the previous information loss. Finally, the ReCoN-generic and ReCoN models had the best performances, significantly better on both metrics than the ReCoN-grn model (ReCoN-generic P-values of 3.88e-6 and 4.61e-3, ReCoN P-values of 5.02e-7 and 3.83e-3). This highlights the necessity of considering the indirect effect through cell communication to accurately predict cell perturbation responses. The full ReCoN, which additionally considers the cell type specificities, showed a modest performance increase on the individual cell type evaluation, and outperformed ReCoN-generic more clearly on the multicellular scores. Although the average AUROC improved from 0.71 to 0.73 with the personalized ReCoN model, the associated P-value of 7.86e-1 indicates that this difference is not statistically significant.

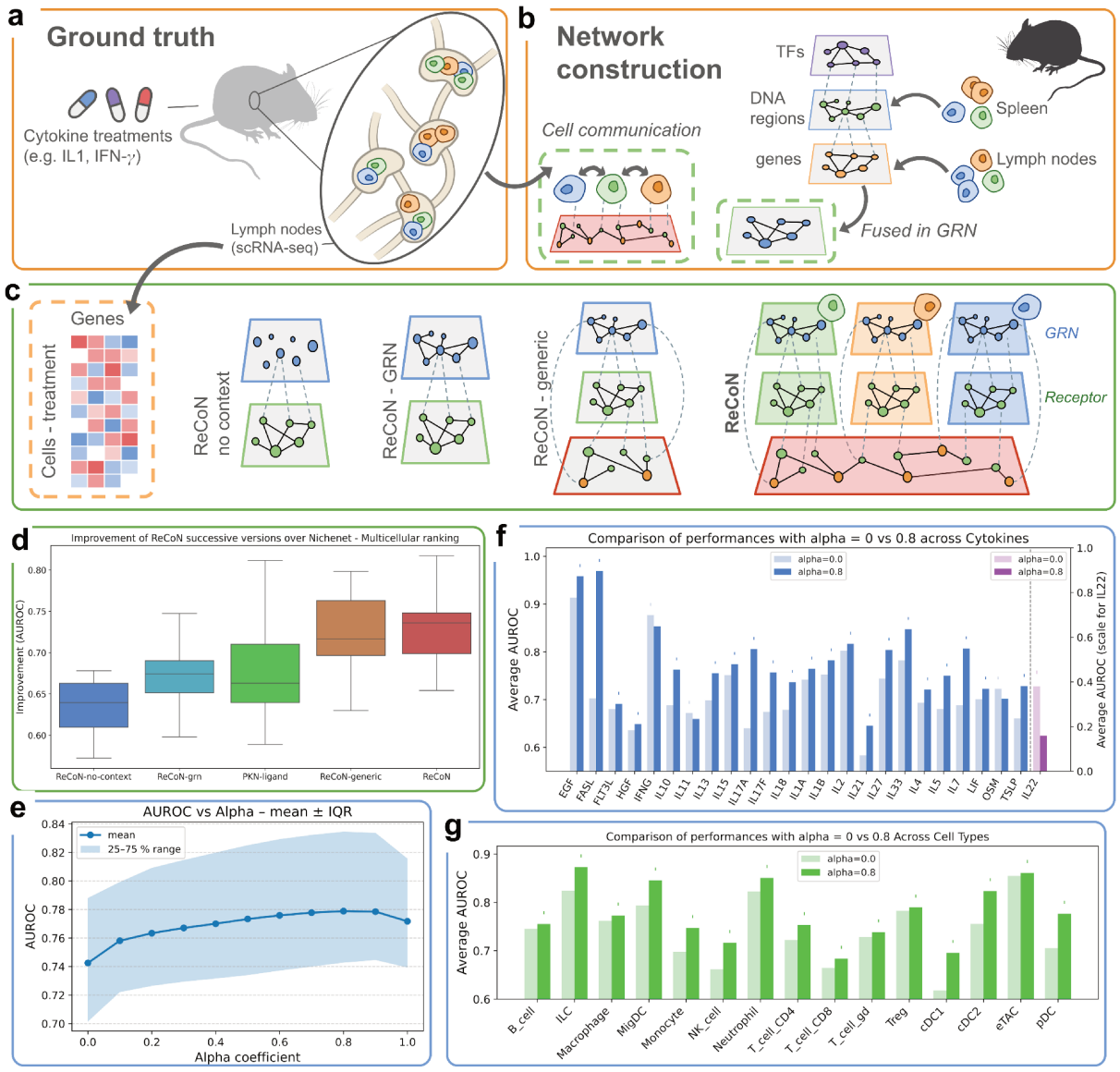


Figure 4.2. ReCoN recovers cell type responses from different cytokine treatments, leveraging cross-cellular effects. a) Schematic representation of the Immune Dictionary dataset used in performance evaluation. **b)** Illustration of the data used to build ReCoN's networks. **c)** Schematic view of the successively enriched networks used in *d*. **d)** Cytokines AUROC distributions for the four successively enriched versions of ReCoN (blue and green) and the PKN model from Nichenet (orange). **e)** Cytokine-cell type pairs AUROC distributions for different α (indicating the contribution of the indirect effect) in ReCoN. **f)** Average AUROCs for each cell type with α = 0.0 (in light green, only direct effect) and α = 0.8 (blue, strong indirect effect); best performance is indicated by a star. **g)** Average AUROCs for each perturbation with α = 0.0 (in light green, only direct effect) and α = 0.8 (blue, strong indirect effect); best performance is indicated by a star. *The mouse and lymph nodes illustrations in **b** come from the NIH BioArt Source library – bioart.niaid.nih.gov/bioart (20, 589).*

Overall, integrating context-specific gene regulatory networks (GRNs) and personalized cell-cell communication (CCC) networks enhances the prediction of cytokine-induced gene expression changes. The inclusion of CCC information, in particular, underscores the significance of indirect effects mediated through intercellular signaling pathways. These

findings highlight the importance of considering both intrinsic regulatory mechanisms and extrinsic communication cues to accurately model multicellular responses.

4.2.2.2. Indirect effects and cell communication explain most of cytokine treatment response

We additionally assessed the relative influence of the cellular environment, and direct signal transduction. We evaluated ReCoN's performance across a range of alpha (α) values, which balance the direct and indirect effects contributions. Overall, the best response predictions averaged per cell type were achieved at α = 0.8, indicating a dominant contribution of indirect signaling via cell–cell communication (see Figure 4.2e). Additionally, the mean AUROC across all cytokine–cell type pairs was 0.76 at α = 0.8, compared to 0.72 at α = 0 (direct effect only), with a Mann–Whitney U test p-value of 1.07×10^{-5}. Consequently, α = 0.8 was selected as the default parameter in subsequent analyses of ReCoN-generic and personalized ReCoN models.

We further investigated the relative impact of indirect effects per cell type. We first measured the optimal α for each cell type. For all, the best performances were associated with an α value above 0.5 (see Supplementary Table 1). We furthermore quantified the predictive power gained by considering cellular coordination (α = 0.8) over the direct effect alone (α = 0) (see Figure 4.2g). Dendritic cells showed the largest gains at α = 0.8, with relative improvements of 12.6%, 10.1%, and 9.0% for pDC, cDC1, and cDC2, respectively. This is consistent with their known roles in orchestrating immune responses and recruiting other cell types[263,264]. Dendritic cells also exhibit high functional heterogeneity and adaptation capacity, notably in response to tumor-specific environments[265]. In contrast, eTACs and Tregs showed the smallest improvements, consistently with their high degree of functional specialization. Both are known for their role in immunosuppression and the inhibition of T cell activation[266,267] Notably, eTAC-mediated $CD4^+$ T cell inactivation has been shown to occur independently of both Tregs and innate inflammatory stimuli[266].

This illustrates the broad and systematic contribution of indirect effects across the studied cell types. The amplitude of this contribution varied between cell types, reflecting differences in their sensitivity to the environment. These variations were consistent with known functional roles, with some cell types showing greater capacity to adapt their responses to contextual signals.

We next examined whether the balance between direct and indirect effects varied between cytokines. Some molecules may exert stronger cell-intrinsic effects, particularly those that drive decisive processes such as apoptosis or proliferation. Among the 22 cytokines tested, 17 (77%) achieved optimal performance with α > 0.6 (see Supplementary Table 2), indicating that indirect effects were generally dominant. IFN-γ and FLT3L showed better predictions at α = 0.2, IL11 at α = 0.2 and OSM and IL22 at α = 0, suggestive of a stronger direct component. IFN-γ performance was an exception, declining markedly at high α values, which could be consistent with its well-documented cell-intrinsic pro-apoptotic activity[268]. FLT3L showed only modest variation, and indirect effects alone still

outperformed the direct effect. FLT3L is primarily involved in immune cell proliferation and migration, amplifying and modulating context-dependent signaling[269]. IL11 and OSM also exhibited minimal change across α values, while IL22 remained poorly predicted regardless of α. Due to very few genes differentially expressed, only cDC1 was conserved and evaluated for IL22, which could also explain the poor performance (see Supplementary Figure 2). When excluding IL22, all cytokines had a relative improvement between -2% and +38%, with an average of 8% using α = 0.8 (see Figure 4.2f).

Overall, cytokine responses reflect a balance between direct and indirect regulation. While indirect signaling predominates across most treatments, the relative contribution of direct effects varies among cytokines. Further investigation of these variations across a broader range of cytokines could enable more precise modeling of the direct–indirect effect balance tailored to each molecule.

4.2.3. ReCoN predicts heart failure transcriptomic changes from potential key regulators.

ReCoN can also model the impact of more complex molecular perturbations. We applied and evaluated ReCoN on retrieving the coordinated multicellular response occurring in heart failure (HF). In ReHeat2, a meta-analysis of human HF transcriptomic data[23], the authors identified widespread shifts in cell type composition and gene expression, including a multicellular program separating HF from non-failing hearts (NFH) (see Figure 4.3a). This program was associated with ligand–cell type interactions, potentially driving disease progression. We focused on these ligand scores to test whether ReCoN could recover their joint regulatory impact. Therefore, we used these ligands as inputs, weighting their influence based on their scores, to model their combined regulatory impact across cell types.

The underlying network was built from a multiome dataset of human left ventricle samples[270]. The gene regulatory layer, shared across cell types, was inferred from the scRNA-seq and scATAC-seq data. The cell–cell communication layer was constructed from the same scRNA-seq data (see Figure 4.3b, Methods). Cell type annotations were adjusted to align with the six major populations described in ReHeat2: cardiomyocytes, fibroblasts, myeloid cells, lymphoid cells, mural cells, and endothelial cells. All ligands from the multicellular factor were used as random walk seeds, with normalized ReHeat2 scores setting their restart probabilities, allowing ReCoN to weight and integrate their contributions.

We compared ReCoN's performance to two reference models derived from NicheNet: the original ligand–target gene network (ligand-PKN) and a receptor-centric version (receptor-PKN), which links receptors to target genes via inferred ligand–receptor interactions. AUROC and AUPR were used to assess model performance in both multicellular and per–cell type gene rankings.

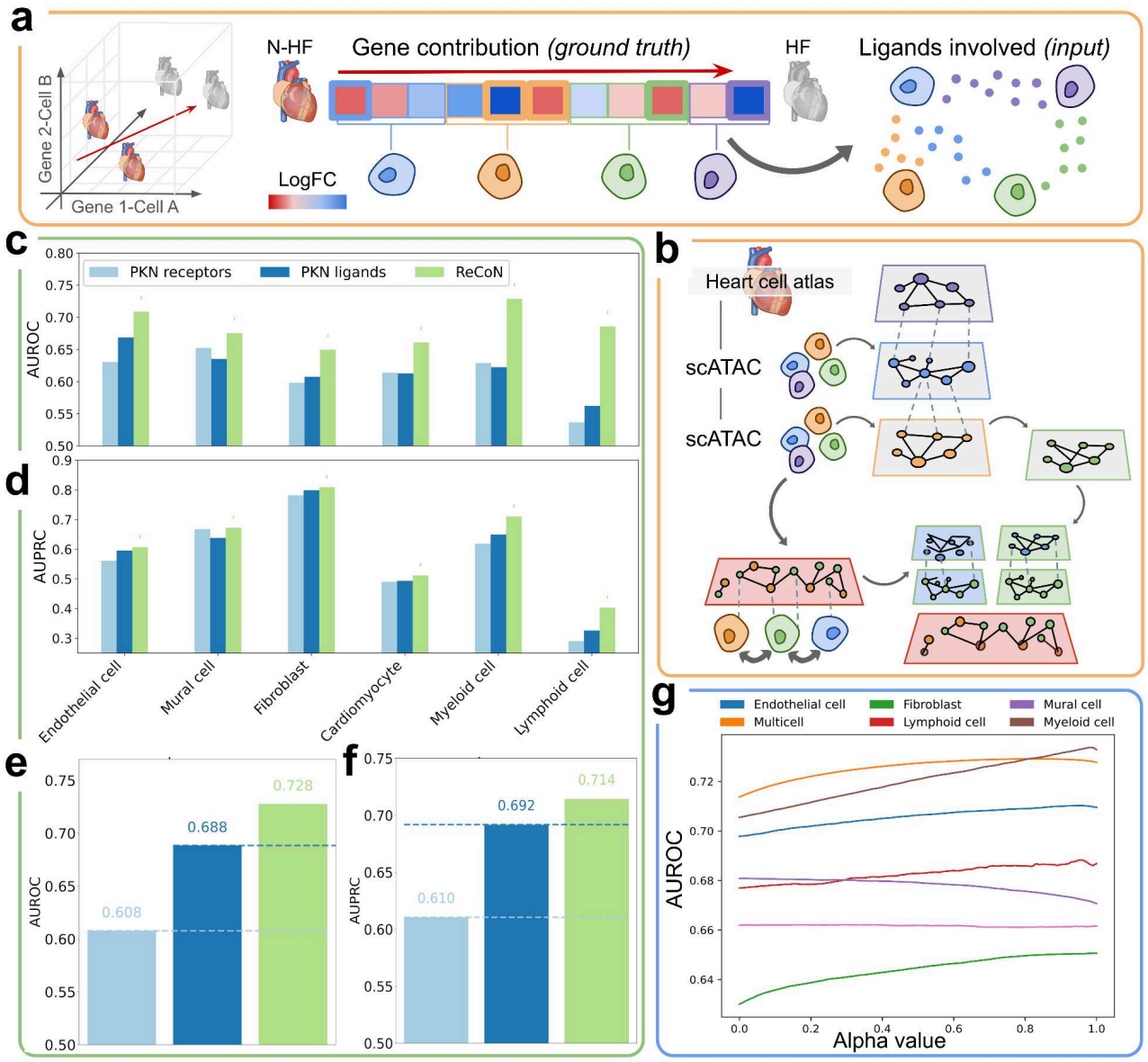


Figure 4.3. ReCoN retrieves the multicellular transcriptomic response to ligands involved in Heart Failure. a) Schematic overview of the heart failure data used as ground truth for model evaluation. **b)** Schematic representation of the reconstructed ventricular heart model reconstructed with ReCoN. **c)** AUROC for predicting differentially expressed genes in individual cell types, comparing ReCoN (green) to two prior knowledge–based models: PKN-ligand (orange) and PKN-receptor (blue). **d)** AUPRC for the same gene prediction task in individual cell types and the same models as in panel c. **e)** AUROC for predicting all differentially expressed genes in the multicellular heart model (i.e., global ranking across all cell types), comparing ReCoN to PKN-based baselines. **f)** AUPRC for the same multicellular prediction task as in panel e. **g)** AUROC as a function of α (the weight of indirect effects) for individual cell type and global multicellular rankings. *The heart illustrations in **a** and **b** come from the NIH BioArt Source library – bioart.niaid.nih.gov/bioart (228).*

In the multicellular ranking, which assesses how well models capture effects across multiple cell types in a comparable manner, ReCoN outperformed both PKN-based approaches in AUROC (see Figure 4.3e) and AUPR (see Figure 4.3f). This ability to jointly evaluate regulatory influence across cell types is essential for identifying selective interventions that affect target populations while sparing others. ReCoN also outperformed the baseline models in the per–cell type rankings, achieving higher AUROCs (see Figure 4.3c) and AUPRCs (see Figure 4.3d) in all six cell types. The AUROC improvement over ligand-PKN ranged from 22% in lymphoid cells to 6.0% in endothelial cells, while the AUPRC improvement ranged from 24% lymphoid again to 1.3% in fibroblasts. Lymphoid cells display especially distinct receptor expression profiles, making them benefit particularly from ReCoN's incorporation of cell-specific signaling context.

We next assessed how ReCoN's performance varied with the relative contribution of the direct and indirect effects. In this setting, the best multicellular and per–cell type rankings were achieved with α values above 0.5, for all cell types except mural cells (see Figure 4.3g). The alpha associated with the best performance on the multicellular ranking was 0.82, close to the 0.8 identified from the Immune Dictionary dataset. It again emphasizes the contribution of indirect signaling and validates this default value by two different systems in mouse and human.

ReCoN consistently outperformed the PKN-based models across both multicellular and per–cell type evaluations. These results reinforce the importance of modeling both direct intracellular regulation and indirect effects mediated by intercellular signaling. By integrating these layers, ReCoN captures the coordinated response to complex perturbations more effectively than static, cell-agnostic frameworks.

Unlike the Immune Dictionary, which captures interventional dynamics with clear temporal resolution, the heart failure analysis is observational and lacks information on signaling chronology. As a result, distinguishing direct from indirect effects is less straightforward: disease-associated ligands may actually initiate cell-intrinsic responses or act downstream to modulate multicellular coordination. Nevertheless, the contribution of indirect effects underscores again the importance of considering cell type interactions to understand their responses even in complex, non-interventional contexts.

4.2.4. ReCoN identifies potential regulators of biological functions in different molecular layers

Heart failure coincides with the activation of diverse biological pathways that disrupt cardiac structure and function. Cardiac fibrosis, a hallmark of this pathological remodeling, is driven primarily by the activation of cardiac fibroblasts and their transition to myofibroblasts, which secrete excessive amounts of extracellular matrix (ECM) proteins. This abnormal ECM accumulation stiffens the myocardium, impairs contractility, and promotes arrhythmias.

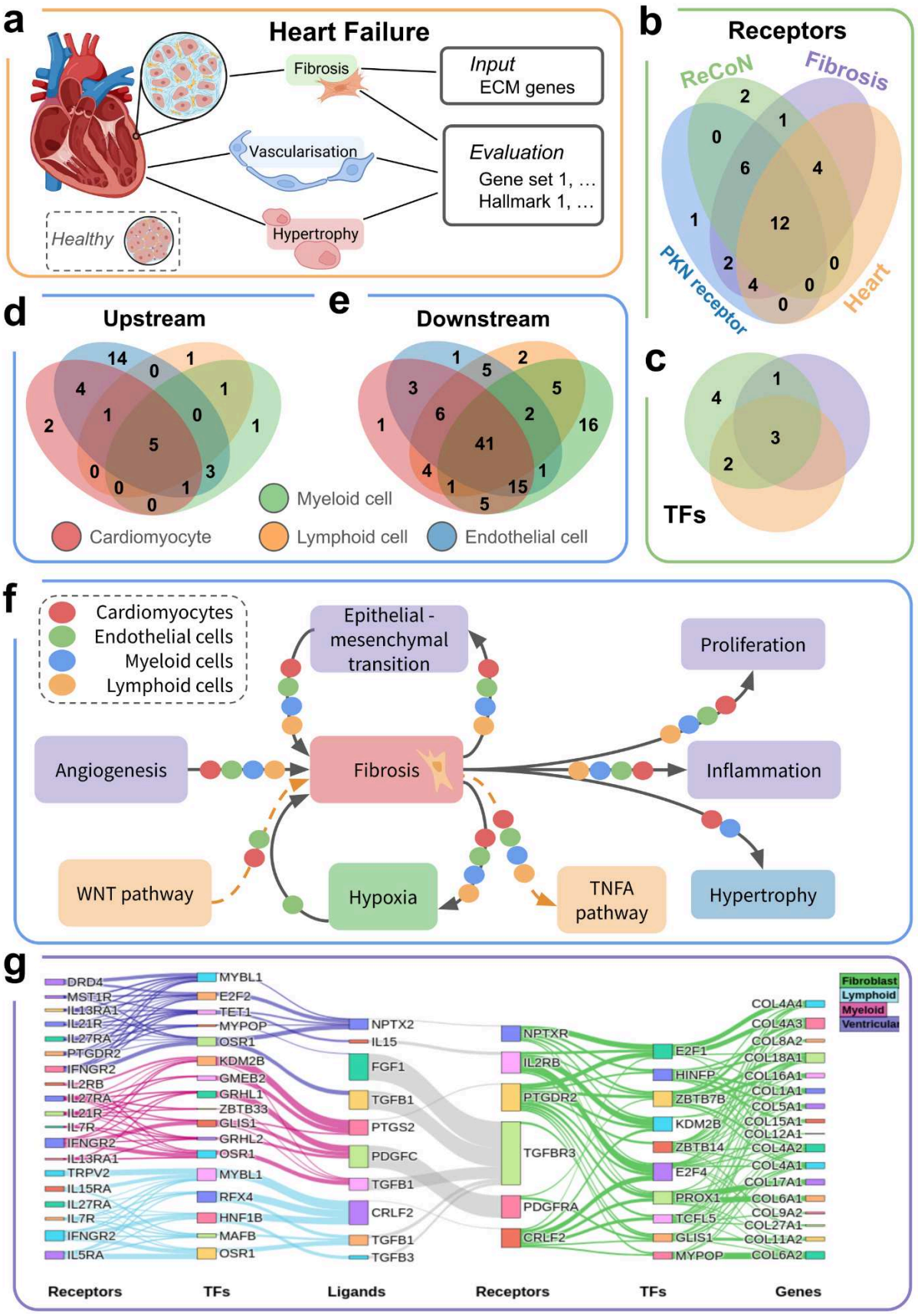


Figure 4.4. Analysis of intracellular regulations and multicellular coordination of cardiac fibrosis. a) Schematic representation of the cardiac fibrosis regulatory exploration with ReCoN. **b)** Venn diagram showing the overlap between receptor sets predicted by ReCoN (green), a PKN

network (blue), and receptors previously implicated in fibrosis (orange) or cardiac diseases (violet). Only the top 25 receptors predicted by ReCoN or PKN have been annotated based on literature; full sizes of fibrosis- and cardiac disease-related receptor sets can therefore not be represented. **c)** Venn diagram showing the overlap between transcription factors (TFs) predicted by ReCoN (green) and those previously implicated in fibrosis (orange) or cardiac diseases (violet). As in panel b, only the top 10 TFs were annotated from literature sources. **d)** Venn diagram showing the intersection of gene sets associated with heart hypertrophy, vascularization, and fibrosis, found to be significantly enriched upstream of cardiac fibrosis in four major cell types: cardiomyocytes (red), myeloid cells (blue), lymphoid cells (orange), and endothelial cells (green). **e)** Corresponding intersections of gene sets significantly enriched downstream of cardiac fibrosis in the same four cell types. **f)** Schematic summary highlighting shared (purple) and cell type-specific (blue, green) biological programs identified as upstream or downstream of cardiac fibrosis. The symbols on the connecting arrows indicate the cell types involved. The most important identified pathways are additionally indicated in orange. **g)** Sankey diagram showing a hierarchical organization of upstream regulators of the "NABA ECM collagens" gene set. Nodes are grouped by molecular type (e.g., transcription factors, receptors, ligands), and links represent the weighted, direct regulatory interactions present in the ReCoN-constructed HMLN. *Illustrations in **a.** were created using BioRender (https://biorender.com).*

In contrast with the previous showcases, we here focused on predicting the regulators instead of the response to a molecular perturbation. We used ReCoN to identify the intracellular molecules regulating the production of ECM proteins in the HF model (see Figure 4.4a). First, we combined 10 gene sets from mSigDB related to ECM production[271], excluding the ones related to cancer propagation. From the listed genes, ReCoN predicted potential upstream TF and receptors from the different cellular layers.

4.2.4.1. ReCoN identified transcription factors related to heart remodelling and fibrosis

Among the top ten TFs predicted by ReCoN, six have been previously linked to fibrosis, and five to heart failure or related cardiac pathologies (see Figure 4.4b, Supplementary Table 3). MYPOP, FBXL10 (also known as KDM2B), and E2F1 were the three highest ranked TFs. All three have been associated with both fibrosis and heart failure, suggesting they could play central roles in regulating cardiac fibroblast activation and ECM protein production. E2F1 has been proposed as a regulator of cardiac fibroblast differentiation[272] and has shown a cardioprotective role in right ventricular failure associated with pulmonary arterial hypertension[273]. FBXL10, an epigenetic regulator responsive to basic fibroblast growth factor[274], has been linked to several cardiac diseases[275], including diabetic cardiomyopathy[276]. This condition is often associated with interstitial fibrosis and can lead to heart failure. It has also been associated with fibroblast metabolic control[277], and appears to be upregulated during the intermediate stages of fibroblast-to-myofibroblast trans-differentiation[278].

MYPOP, also known as *Myb-related transcription factor, Partner of Profilin-1* (PFN-1), interacts with PFN-1 and helps mediate its effects on gene expression[279]. MYPOP has not been directly linked to heart failure. However, PFN-1 has been associated with cardiac

hypertrophy and fibrosis[280], and shown to contribute to fibrosis and cardiac injury in rat models[281]. ReCoN's predictions suggest that MYPOP may be a key mediator of PFN-1's effects in cardiac fibrosis and hypertrophy, aligning with its known molecular interactions.

4.2.4.2. ReCoN leverages prior knowledge to predict receptors regulating ECM gene expression by cardiac fibroblasts.

In addition to TFs, we used ReCoN to identify upstream receptors potentially driving ECM protein expression in cardiac fibroblasts. Predictions were compared to the receptor-PKN baseline, where each receptor was scored by summing its weighted links to target genes (see Figure 4.4c). Among the 25 top-scoring receptors prioritized by ReCoN, 23 were previously implicated in fibrotic processes across various tissues, and 16 could be specifically linked to cardiac fibrosis or related cardiovascular conditions (see Supplementary Table 4). ReCoN's top predictions showed strong agreement with the receptor-PKN, with 18 receptors shared between the two rankings. The prior knowledge model had comparable overlap with the literature, slightly higher with 24 receptors associated with fibrosis and 16 with cardiac-related contexts (see Supplementary Table 5).

We next examined receptors which scores differed the most between ReCoN and the receptor-PKN model, using a metric of "gene movement" across rankings[282]. While the novel receptors prioritized by ReCoN were not more specific to cardiac tissue as we could have expected from the heart-specific GRN in ReCoN's model, they remained enriched for regulators of fibrotic pathways (see Supplementary Table 6, Supplementary Table 7). Eight out of the top ten of these newly highlighted candidates were linked to fibrosis in the literature. This comparison also showed a higher ranking by ReCoN of several receptors previously described as cardiac-specific, refining their annotation to better reflect roles in fibrotic or remodeling processes.

These changes likely stem from ReCoN's integration of context-specific information, enabling it to highlight regulators relevant to the studied condition. However, this specificity may come at the cost of overlooking broadly validated receptors, making it difficult to assess whether the shifts represent true biological refinements or data-driven bias. In our analysis of receptor prioritization, ReCoN highlighted ADGRG1 (also known as GPR56) as a notable candidate, ranking it 12th compared to 50th in the receptor-PKN model. Recent studies have shown that cardiomyocyte-specific deletion of ADGRG1 leads to accelerated cardiac dysfunction, increased inflammation, and higher mortality, underscoring its potential role in heart failure pathogenesis[283]. Conversely, IL1RAP, which ranked higher in the receptor-PKN model, was deprioritized by ReCoN. Blocking IL1RAP has been shown to reduce cardiac inflammation and preserve function in experimental models of myocarditis[284].

Overall, ReCoN effectively recovered many known regulators of fibrosis, showing strong overlap with prior knowledge. At the same time, it reprioritized several receptors based on context-specific features of the data. These changes likely reflect a better adaptation to the studied condition but may also lead to the exclusion of broadly validated targets. These

changes likely come from ReCoN's integration of context-specific information, enabling it to highlight regulators relevant to the studied condition. However, this specificity may come at the cost of overlooking broadly validated receptors, making it difficult to assess whether the shifts represent true biological refinements or data-driven bias.

4.2.5. Cell type specificities and multicellular coordination

Heart failure arises from the co-occurrence of multiple physiological and cellular modifications. We here investigate the specific coordination of major cardiac cell types with cardiac fibroblast and ECM gene activation. Both upstream and downstream of fibrosis, ReCoN predicted genes activated in other cardiac cell types. It could identify genes and pathways activated both differentially and consensually in several major cardiac cell types.

4.2.5.1. Regulatory mechanisms across cell types

We investigated upstream genes connecting other cardiac cell types with fibroblast activation in heart failure. Several known mediators of cardiac fibrosis and remodeling, including TGFB1, NPPB, WNT5B, BMP4, GDF15, and NRG1, were predicted as upstream regulators across multiple cell types. TGFB1, NPPB, and WNT5B were consistently ranked in the top 50 out of 9,971 genes across all interacting cell types, suggesting broad relevance in fibrosis signaling. Other markers, such as BMP4 and GDF15, showed selective prioritization, both ranked out of the top 1500 in lymphoid cells and in the top 20 for all other cell types, suggesting cell type-specific involvement, even among canonical HF drivers. We finally predicted genes specific to a few cell types and often studied in relation to HF. The top-ranked gene in myeloid cells was Oncostatin M (OSM), which was not prioritized in other lineages. OSM is secreted by macrophages and has been implicated in anti-fibrotic regulation, supporting its potential role as a cell type-specific paracrine signal during cardiac remodeling. In contrast, genes such as ADCYAP1 and FGF2 were specifically prioritized in cardiomyocytes and endothelial cells. ADCYAP1 encodes PACAP, a neuropeptide involved in vasodilation with cardioprotective properties in HF[285–287], while FGF2 is a well-known mediator of angiogenesis and cardiac repair[288]. These results emphasize that, while some regulatory signals are broadly shared across cell types, others reflect context-specific interactions between fibroblasts and distinct cardiac populations. They also highlighted some regulatory signals broadly shared across cell types, and identified cell type-selective signals regulating fibroblast activation and cardiac fibrosis.

4.2.5.2. Cellular program coordination

To dissect multicellular coordination of broader functional programs in HF, we performed gene set enrichment on ReCoN's upstream (see Figure 4.4d) and downstream (see Figure 4.4e) predictions in cardiomyocytes, endothelial cells, lymphoid cells, and myeloid cells. Those four lineages were selected because they showed the greatest improvement when including cell-cell communication (see Figure 4.3g).

Shared upstream programs included epithelial–mesenchymal transition (EMT), apical junction remodeling, and angiogenesis across all four cell types (Supplementary Table 8). EMT reflects the acquisition of mesenchymal traits and increased motility that underlie fibroblast activation and matrix deposition. On the other hand, apical junction remodeling indicates changes in cell–cell adhesion and polarity essential for tissue reorganization. Cardiac angiogenesis and fibrosis interact closely to promote cardiac regeneration[289].

Subsequently, the lineage-specific upstream enrichments emerged. Hypoxia-related signaling was strongly enriched only in endothelial cells, consistent with ischemia-driven vascular adaptation in HF (see Figure 4.4f, Supplementary Table 9). Pathway gene sets for androgen response and progesterone inhibition were significant only in endothelial cells, albeit with lower NES, reflecting the specific role of sex-hormone modulation in vascular fibrosis. Cardiomyocytes uniquely enriched mTORC1 signaling, a key regulator of cell growth, protein synthesis, and metabolic adaptation under stress (Source MSigDB). Additionally, both cardiomyocytes and endothelial cells enriched the mitochondrial pathways, reflecting high bioenergetic demand, and WNT signaling, which regulates extracellular matrix gene expression and fibroblast proliferation. All of which indicates a coordinated role in fibrotic remodeling.

Downstream of fibrosis, 41 out of 108 significant gene sets were shared across all lineages, despite showing different enrichment amplitudes. Shared axes corresponded to inflammation, hypertrophy, proliferation, hypoxia, and EMT (see Figure 4.4f, Supplementary Table 10), all important molecular hallmarks of HF. The inflammatory axis included TNFα/NF-κB, TGF-β, IFN-γ/α responses, and generic inflammatory hallmarks, highlighting a global inflammation response and cytokine-driven remodeling. Hypertrophy programs, related to muscle hypertrophy and hypertrophic cardiomyopathy signatures, underline heart enlargement and contractile adaptation. Proliferation terms reflect cell-cycle activation in reparative and immune cells. Hypoxia and EMT remained significant downstream, marking their dual roles as drivers and consequences of fibrosis. Additionally, a gene set of upregulated genes in systolic heart failure was enriched across all lineages, with NES values varying from 1.97 (lymphoid) to 3.04 (cardiomyocytes). This range underlines a shared transcriptional response, while highlighting differences in activation strength among cell types.

We then focused on cardiomyocytes and myeloid cells, which showed the most downstream enrichments. Both were further enriched in physiological and pathological hypertrophy pathways, capturing adaptive versus maladaptive growth responses (Supplementary Table 11). Cardiomyocytes and endothelial cells shared the Cardiac EGF pathway, which promotes fibroblast activation and collagen synthesis in injury contexts. Hedgehog signaling appeared in cardiomyocytes, endothelial, and myeloid cells, reflecting embryonic reactivation and stem-like reprogramming noted in HF, another important hallmark of the condition[23,290]. Additionally, myeloid cells significantly enriched metabolism and protein secretion pathways, which is consistent with macrophage-driven matrix remodeling and cytokine production.

Together, these findings demonstrate that HF is driven by both common multicellular programs, such as inflammation, hypoxia, EMT, and angiogenesis, and lineage-specific mechanisms, including mTORC1 in cardiomyocytes and sex-hormone–linked vascular pathways in endothelial cells. ReCoN successfully captures this nuanced interplay, identifying cell type-specific drivers and consequences of fibrosis. We summarized the specific and generic identified gene programs in Figure 4.4f.

4.2.5.3. Visualisation of multicellular pathways

Analysis with ReCoN relies on the complete multicellular multilayer network exploration. While we allow by default to jump from gene to receptors, it can still be relevant to visualize gene - TF - receptor triplets to extract mechanistic hypotheses underlying ReCoN scores. We provide several functions to visualize the links between downstream genes and their regulators in each level. We thus select the top TFs and receptors regulating a set of intracellular genes, and extract a subnetwork of direct connections between these elements. It is additionally possible to include the impact of other cell types through the top ligands predicted, and their upstream regulators, as presented in Figure 4.4g. In this example, we provide a visualization of the regulations of the gene set “NABA ECM collagens”, which was used earlier as genes involved in fibrosis.

4.3. Discussion

The complexity of tissue biology lies not only in the multitude of cell-specific intrinsic responses but also in the indirect effects that emerge from cell–cell communication. In any multicellular system, a perturbation applied to one cell type can ripple through its neighbors: for example, a cytokine might directly stimulate an immune cell, which then secretes secondary factors that reshape the behavior of surrounding cells. Failing to model these indirect, intercellular effects can lead to an incomplete picture of regulation. Indeed, many diseases and tissue functions arise from coordinated multicellular responses. Recognizing this, we focused on developing a framework to explicitly capture how a perturbation’s influence on one cell is mediated by the responses of others, thereby addressing a fundamental challenge in tissue-level gene regulation.

ReCoN was designed to meet this challenge by modeling gene regulation across multiple cell types simultaneously with a network representation of cell–cell communication from single-cell data. It can thus model indirect regulatory influences in a data-driven way. This approach contrasts with earlier strategies that infer interactions only from ligand–receptor pairs or treat cell types in isolation, sometimes both in parallel[126], which cannot fully unravel the complex, higher-order signaling networks present in real tissues. By jointly analyzing how all cell types respond under a given condition, ReCoN identifies multicellular responses, as coordinated transcriptional changes across different cells, leveraging the concept of indirect effect: for instance, when a given cytokine triggered one leukocyte subset to produce downstream mediators, ReCoN incorporated that indirect influence on other cell types. Isolating direct and indirect effects, ReCoN was used here to provide insights into their respective contribution in predicting in-tissue responses

The advantages of this framework are evident in its strong performance on two distinct biological datasets. In the Immune Dictionary, a recent single-cell atlas mapping how 86 cytokines affect 17 distinct immune cell types in vivo, ReCoN accurately predicted each cell type's gene expression responses to cytokine stimulation, with a significant contribution of the GRN layer. Importantly, it outperformed isolated cell models by capturing indirect effects. Across the board, modeling such cross-talk yielded significantly higher correlation with observed expression profiles, especially for genes regulated via multicellular feedback loops. Similarly, when predicting Heart Failure signatures from the Human Heart Atlas single-cell data, incorporating intercellular interactions led to marked improvements in predictive power. ReCoN better explained gene expression for all studied cell types, correctly capturing many hallmarks of cardiac disease that emerge from intercellular signaling (e.g., BMP4, cardiomyocyte hypertrophy, hypoxia). In both the immune and cardiac contexts, the inclusion of indirect effects proved crucial, underscoring that many key drivers of system-level behavior impact multiple cells through cell-cell communication.

We further demonstrated ReCoN's capabilities in a showcase analysis of multicellular interactions in cardiac fibrosis, a pathological process central to heart failure. The model could infer upstream signals, i.e. identifying which cell types and secreted factors were likely initiating particular gene programs in other cells. It could also predict downstream consequences, i.e. how those intercellular signals altered gene expression in the receiving cells. This comprehensive view allowed us to map a web of coordinated cell–cell responses driving fibrosis. Notably, ReCoN identified distinct but interconnected transcriptional programs activated in cardiomyocytes, endothelial cells, myeloid cells, and lymphoid cells within the fibrotic heart. Cardiomyocytes, for example, reactivated a stress-associated "fetal gene" program, reflecting the well-known hypertrophic and developmental reversion seen in failing hearts. Endothelial cells, by contrast, upregulated an inflammatory and pro-fibrotic program. Each of these cell-specific programs aligns with known roles in heart failure pathology, as prior studies have observed that cardiomyocytes, endothelial cells, myeloid cells, and fibroblasts each adopt unique disease-associated states in failing hearts. This not only recapitulates individual findings of cell-specific changes but integrates them into a coherent multicellular network of cardiac remodeling, yielding testable predictions about which intercellular communications are key causal drivers.

Overall, we also highlight here the value of in vivo perturbations studies, to understand more precisely the indirect effects and their relative amplitudes, which can vary between drugs, environment, and cell types.

4.4. Methods

4.4.1.1. Software and Reproducibility

The notebooks and scripts used to generate the presented results are available at https://github.com/cantinilab/recon_reproducibility, along with the corresponding conda environments and singularity images.

4.4.1.2. Network Construction

ReCoN models context-specific regulation by integrating intracellular gene regulatory interactions with intercellular communication in a heterogeneous multi-layer network. A key requirement is that all edge weights remain strictly positive, ensuring valid probability distributions for diffusion algorithms. Only the nodes that are included in one of the layers are present in the final results, ignoring the ones only present in bipartites.

Gene Regulatory Layer

We inferred a global gene regulatory network using HuMMuS (version 0.1.7), which combines scRNA-seq and scATAC-seq data to predict transcription factor–target gene interactions. However, any GRN inference method based on scRNA-seq alone or both scATAC-seq and scRNA-seq could be used.

With HuMMuS, we first built an HMLN composed of three layers: a TF layer, a scATAC layer containing peak co-accessibility information inferred from scATAC-seq data, and the scRNA layer encoding transcriptional regulation inferred from scRNA-seq data. The co-accessibility network was inferred using CIRCE[291] (version 0.3.4), and we kept all the links with positive scores. The scRNA layer was inferred with GRNBoost2[292] (Python version implemented in Arboreto 0.1.6), and we kept the 50 000 first links. Both layers were then combined with the Python code of HuMMuS, and we explored the HMLN to compute the gene regulatory layer (see Supplementary Notes 2). Finally, only the links with a score above 1.5e-7 were retained in ReCoN's gene regulatory layer.

Receptor–Gene Bipartite Layer

While some took interest in inferring receptor - gene links directly[293], there is very limited direct information compared to the large-scale ligand - target genes publicly available (Cytosig[294], Nichenet[22]). We thus decided to infer receptor - gene links under the assumption that we reconstruct them through a linear matrix equation :

$$R = L \cdot G$$

with ligand–receptor (L) and ligand–gene (G) adjacency matrices from NicheNet v2, accessible at https://zenodo.org/records/7074291. These matrices are accessible for both mouse (lr_network_mouse_21122021.rds, ligand_target_matrix_nsga2r_final_mouse.rds)

and human (lr_network_human_21122021.rds, ligand_target_matrix_nsga2r_final.rds). We retrieved R using non-negative least squares (NNLS) in SciPy[295] (v1.15.2) enforcing to have only positive links as needed in ReCoN. We finally considered all receptor-gene links with a score of 5e-3.

Cell–Cell Communication Layer

The cell-cell communication layer contains nodes defined by a molecule and a cell type that produced it. Intercellular edges were inferred using LIANA+[19] (version 0.1.9) with CellPhoneDB[121] algorithm, without imposing a limit on the initial proportion of a cell type expressing a ligand (parameter expr_prop = 0). This produces a directed, weighted ligand–receptor network with each pair of cell types, using cell type scRNA-seq expression and a prior network of known ligand-receptor bindings. Any method producing non-negative edges could also be used as long as the links contain individual molecules and cell types involved. For the human data (Heart model), we used the "consensus" database of LIANA+ as prior. For the mice data (Lymph node model), we use the ligand-receptor binding database provided by Nichenet for mice. We then retained all interactions strictly positive in all showcases.

4.4.2. Network exploration and signal propagation

4.4.2.1. Random Walk with Restart (RWR)

Random walk with restart (RWR) is a stochastic process consisting of a succession of steps from one node (i.e., the seed) to a neighboring one through the network's edges, with a probability to start again from the seed at each step. RWR can be used to explore HMLNs and to provide a measure of nodes' closeness across the layers, ensuring the existence of a unique stationary distribution[11,296]. RWR strategies have been shown to significantly outperform methods based on local distance measures for the prioritization of gene-disease associations[146,297]. To run the RWR, we here used MultiXrank, a Python package proposing optimized RWR on universal multilayer networks[146].

Different parameters guide the successive steps. First, the restart r determines a trade-off between returning to seed nodes and exploring the network and prevents the random walker from being trapped in dead ends. While the restart probability is typically 0.7 in previous works and HuMMuS[10,146,259,298], the default value in ReCoN is 0.6 to allow a deeper exploration of the network across an important number of layers. If there are multiple seeds, each of them has a relative probability based on its weight to be used as a restart.

We also need to specify the probability of moving from one layer to any layer. It allows you to direct the flow of the exploration and the importance of each layer. By default, in ReCoN, the probability of staying in a layer is 0.5. The probability of jumping to each reachable layer is then 0.5/N, with N the number of reachable layers. Once the layer has been decided, the scores of the reachable nodes in it are normalized such that they can now be used as probabilities to reach them.

4.4.2.2. ReCoN basic explorations

ReCoN has four standard ways of exploring multicellular systems, based on depth and direction. It can first be used to identify upstream and downstream molecules. The 2 exploration directions follow specific flows of information through different transition matrices in-between layers. Downstream explorations allow transition from the cell-cell communication layer to the receptor layer, the receptor layer to the gene regulatory layer, and the gene regulatory layer to the cell-cell communication layer. It represents the expected signal transduction. Upstream explorations allow the exact opposite transitions. Additional layers can be included with their own rules, and could notably allow transition back and forth with other layers (i.e., metabolites to genes and genes to metabolites).

For each exploration direction, it is possible to decide between intercellular exploration (exactly as described above) and intracellular exploration alone. The intracellular exploration excludes cell-cell communication and thus some transitions between layers. In the intracellular downstream exploration, there is no transition from the gene regulatory layer to the CCC layer: genes cannot jump to expressed ligands in the cell communication layer. In the intercellular upstream exploration, transitions from the CCC layer to the gene regulatory layer are not allowed: binding ligands cannot jump to the gene that expresses them. Isolating both exploration depths is useful to understand the specific effect due to direct signal transduction and indirect cell communication.

4.4.2.3. ReCoN predictions

The global effect or context of a perturbation is computed by a combination of its direct and indirect effects. The direct effect corresponds to an intracellular exploration, while the indirect effect is obtained from several intercellular explorations. All three outputs are vectors containing all the nodes of the HMLN and the associated probability of reaching them. These probabilities are interpreted as the weight of the regulations between them and the seeds.

A downside of RWR in a large multilayer network is that nodes of layers that are close to the seeds have mechanically higher weights. Since the indirect effects modelling requires crossing more layers, the direct effect had a much higher contribution. This is also a reason for differentiating direct and indirect effects, allowing for modulating their contribution in a second time. We additionally aim to reduce this effect in the computation of the indirect effect itself. From the direct effect, we first identify downstream genes in the GRN layer that encode secreted ligands. Rather than seeding the indirect RWR on those gene nodes, we assume that they will be translated as protein and seed the walk on the corresponding ligand-produced nodes in the cell–cell communication layer. This places the restart point closer to the receptors and downstream targets in other cells, reducing the score's dilution across the walks. Symmetrically, for the upstream exploration, instead of starting from the binding ligands, we use the gene node linked to them.

4.4.2.4. Cell type proportional contribution

Some cell types have more effect on their surroundings than others. It depends both on their distribution and their ligand expression profiles. By default, each cell type contributes equally to predicting the others. However, it is possible to adjust the Beta coefficient to represent it based on the available information for each dataset. Notably, spatial transcriptomic data could be used to identify the proximity and relative importance of each cell type over the sum of all surrounding cell signals.

4.4.3. Modeling Cytokine Treatments In Vivo (Mouse)

4.4.3.1. Data Sources and Preprocessing

The Immune Dictionary data were downloaded at https://www.immune-dictionary.org as Seurat objects and merged into a MuData object comprising the scRNA-seq profiles of 81 cytokine treatments across 17 lymph node cell types. In these downloadable files, a maximum of 100 cells per cytokine treatment for each cell type were sampled to ensure comparability across cell types for this analysis and 15 cell types were available: B cell, cDC1, cDC2, eTAC, ILC, Macrophage, Migratory DC, Monocyte, Neutrophil, NK cell, pDC, T cell (CD4+), T cell (CD8+), γδ T cell, Treg. All the cells were kept for the subsequent analysis. Only 41 cytokines were present in the prior ligand-receptor database, and 25 had at least one active connection inferred by CellPhoneDB. We use the latter to compare the different models.

4.4.3.2. Identification of Perturbed Genes

For each cytokine–cell type combination, differentially expressed genes were identified with Scanpy's Wilcoxon rank-sum test (FDR P-val < 0.1, $|\log_2 FC| > 1$). In individual cytokine-cell type pairs evaluations, the ones with fewer than two significantly perturbed genes were excluded from downstream ranking analyses. We ended up with 206 cytokine - cell type pair profiles (see Supplementary Figure 2).

4.4.3.3. External Multi-Omic Integration

To refine the gene regulatory layers, we combined scRNA-seq from external murine lymph nodes[299] and scATAC-seq from murine immune cells (CD45+)[300]. The scATAC count matrix is accessible in GEO under accession code GSE242466 as "GSE242466_archr_mouse_immune_cell_atlas.tar.gz". The scRNA-seq data is available at https://cf.10xgenomics.com/samples/cell-exp/7.2.0/4plex_mouse_LymphNode_Spleen_TotalSeqC_multiplex_LymphNode1_BC1_AB1/4plex_mouse_LymphNode_Spleen_TotalSeqC_multiplex_LymphNode1_BC1_AB1_count_sample_filtered_feature_bc_matrix.h5, as part of the following dataset: https://www.10xgenomics.com/datasets/Mixture-of-cells-from-mouse-lymph-nodes-and-spleen-stained-with-totalseqc-mouse-universal-cocktail

Briefly, we filtered out in both matrices the features that were expressed in less than 3 cells, and then limited to only the 16,000 most variable genes. We then filtered out cells with fewer than 3 expressed features. It resulted in the scRNA-seq in 1,789 cells with 13,167 genes, and for the scATAC-seq in 3,759 cells with 254,545 regions.

4.4.3.4. Murine lymph node model reconstruction.

A gene layer was inferred from the external scRNA-seq datasets with GRNBoost2[292] (python implementation in arboreto, v0.1.6), and a DNA region layer was inferred from the scATAC-seq dataset with CIRCE[291] (v0.3.4). Untransformed counts were used as input for both. Both were then combined to infer a GRN with HuMMuS[10] (v0.1.7). The cell communication layer was computed from the Immune Dictionary subset of the cells treated with phosphate-buffered saline solution (PBS).

4.4.3.5. Benchmarking Against Baseline Models

ReCoN performance was compared to NicheNet's ligand-PKN and receptor-PKN models. For each model, AUROC and AUPR were calculated for per–cell type and multicellular rankings. Statistical significance of performance differences was evaluated using two-sided Mann–Whitney U tests (non-Gaussian assumption), corroborated by Wilcoxon signed-rank tests.

4.4.4. Modeling Heart Failure in Humans

4.4.4.1. Multiome Human Heart Cell Atlas Preprocessing

We preprocessed the single-cell multiome data from the human left ventricles as proposed in the GRETA pipeline. The samples were first extracted from the complete dataset. We filtered out the cells expressing fewer than 100 genes, and all genes expressed in fewer than 3 cells. We kept the most variable genes expressed in more than 16,384 cells, and the 65,536 most variable regions. Cell type annotations by different methods are already present in these samples. We kept all the cells whose annotations through unsupervised clustering, followed by marker genes through scANVI, were coherent. The preprocessed dataset contained 25,787 cells, 16,384 genes, and 62,769 peaks.

4.4.4.2. Human Heart model reconstruction.

The gene regulatory layer was then computed similarly to the murine lymph node model, using GRNboost (python v0.1.6) and CIRCE (v0.3.4), and inferring the final GRN combining both layers with the R version of hummus.. The cell types identified from ReHeat2 were used. We thus matched the cell types of the Heart Cell Atlas dataset to the ones of ReHeat2 according to the Supplementary Table 12. The cell communication was then computed between these re-annotated cell types.

4.4.4.3. Heart failure genes and ligands.

To identify the cell type-specific genes associated with HF, we used the MOFAcell scores of the multicellular factor 1 (MCP1) reported in ReHeat2[23]. To have high confidence sets of genes perturbed and unperturbed, we ranked all the scores and considered the top ten percent highest absolute scores as true positives, and the ten percent lowest as true negatives. In parallel, pairs of ligands and receptors with both associated with scores above an absolute gene loading of 0.1 were considered potential driver interactions in HF. The ligands of these pairs were used as seeds for ReCoN and PKN-ligands, while the receptors were used in PKN-receptors.

4.4.4.4. Fibrosis Gene Selection

The genes related to the extracellular matrix were obtained from mSigDb. We merged all the NABA gene sets in human that were not related to cancer : ‘basement_membranes', 'collagens', 'core_matrisome', 'ecm_affiliated', 'ecm_glycoproteins', 'ecm_regulators', 'matrisome', 'secreted_factors’, 'matrisome_associated', 'proteoglycans'. 149 of these genes were present in the gene regulatory layer and used as seeds in the fibrosis analysis showcases.

4.4.4.5. Evaluation of performances in predicting HF differentially expressed genes

Seeds comprised HF-program ligands with normalized scores. RWR influence scores for all genes were computed as described above. We benchmarked against NicheNet’s ligand- and receptor-PKN models using AUROC/AUPR in multicellular and per–cell type contexts.

4.4.4.6. Evaluate receptor ranking shifts between ReCoN and PKN-receptors

We identify the receptor whose activity profile differs most between the ReCoN framework and the PKN-receptor model by treating each receptor’s activity score as a one-dimensional embedding representation, and applying the Gene Mapping Matrix (GMM) approach[282]. Specifically, for each condition (ReCoN or PKN) we form a vector U of receptor scores and compute a normalized distance matrix

$$GMM_{ij} = \frac{|x_i - x_j|}{||U||_2}$$

where $||U||_2$is the Euclidean norm of the score vector. We then quantify each receptor’s “movement” as the Euclidean distance between its corresponding row in the ReCoN GMM and its row in the PKN GMM. The receptor exhibiting the largest movement value is the one whose relative relationships to all other receptors shift most between the two models.

4.4.4.7. Functional Enrichment

We used the gseapy python package to realise the GeneSet Enrichment Analysis (GSEA) of upstream and downstream predictions. We evaluated the enrichment of the MSigDB

Hallmarks (v2025) and the collection of gene sets extracted with the combination of keywords related to hypertrophy, vascularization, and fibrosis (see Supplementary Notes 3). Enrichment p-values were adjusted by Benjamini–Hochberg FDR. Adjusted P-values below 0.1 were considered significant in the subsequent analysis.

4.5. Data accessibility

Processed data and network used to generate the results will soon be available at Zenodo.

Only public data was used to generate the results. The Immune Dictionary data can be downloaded at https://www.immune-dictionary.org/app/home. The murine lymph nodes scRNA-seq data and the murine scATAC-seq data of immune cells used as input for the lymph node model reconstruction can be downloaded on GEO under the accessibility number GSE242466 and on the 10Xgenomics platform https://www.10xgenomics.com/datasets/Mixture-of-cells-from-mouse-lymph-nodes-and-spleen-stained-with-totalseqc-mouse-universal-cocktail, respectively.

The Heart Human Cell Atlas data can be downloaded at https://cellgeni.cog.sanger.ac.uk/heartcellatlas/v2/Global_raw.h5ad (scRNA-seq) and https://cellgeni.cog.sanger.ac.uk/heartcellatlas/v2/Adult_Peaks.h5ad (scATAC-seq). We used the samples with the following IDs: HCAHeartST10773166_HCAHeartST10781063, HCAHeartST10773165_HCAHeartST10781062, HCAHeart9508627_HCAHeart9508819, HCAHeart9508629_HCAHeart9508821, HCAHeart9845431_HCAHeart9917173, HCAHeartST11064575_HCAHeartST11023240, HCAHeartST11064574_HCAHeartST11023239. The genes and ligands weights in the multicellular factor 1 identified in Heart Failure are available at https://doi.org/10.5281/zenodo.15261569. The gene sets were queried and downloaded from https://www.gsea-msigdb.org.

4.6. Code accessibility

ReCoN is publicly available at https://github.com/cantinilab/ReCoN. All the analyses presented in this manuscript were produced with ReCoN version 0.1.0 and will be uploaded to Zenodo. Notebooks and conda environments to reproduce the analyses presented in this manuscript are available at https://github.com/cantinilab/ReCoN_paper.

Preprocessing and gene regulatory layer construction were made through Snakemake[301] pipelines using Singularity[302] containers. Downstream analyses are available as Jupyter Notebooks[303]. All computations were realised on the Pasteur Institute HPC, with 2 AMD EPYC 7552 48-Core Processors, 500 GB of RAM, and Linux Red Hat 8.8.

Chapter 5

Discussion

5.1. Conclusion of the thesis

Heterogeneous multilayer networks offer a new framework for integrating complementary interactions between molecules.

In this thesis, HMLN structures have been leveraged for modelling different scales.

First, it has been applied to inferring molecular mechanisms inside cells. Chapter 2 introduces HuMMuS, a framework and package for modelling the regulation of gene expression by TF and DNA regions. Decomposed into layers that contain specific types of macromolecules (e.g., proteins, DNA, RNA), it considers both intra-omics and inter-omics edges to improve predictions through RWR explorations. While most other state-of-the-art methods based on scRNA-seq and scATAC-seq focus on inter-omics links, usually symbolizing regulation, HuMMuS relies on the assumption that other undirected co-operations are also essential to represent precisely the gene regulatory network. HuMMuS outperformed the most commonly used method in a benchmark considering TF-gene, TF-DNA region, and DNA region-gene links. Additionally, taking into account PPI between TFs improved performances, highlighting the contribution of these links. To further demonstrate the potential of HMLN for future developments, HuMMuS was applied on a three-modality dataset – scRNA-seq, scATAC-seq, and snmC-seq – and efficiently leveraged these third modality. It could be easily extended to upcoming single-cell data and technology developments.

HuMMuS depends on several pieces, which all require important computational resources. Nowadays, scATAC-seq data typically contain hundreds of thousands of individual DNA regions, and inferring their interactions becomes rapidly costly. Initially, HuMMuS used Cicero(source), a state-of-the-art method to infer cis-regulation DNA region interactions. However, this method presents several limitations, both in data pre-processing and computation resources usage. To overcome these challenges, Chapter 3 introduces CIRCE, a new package optimising Cicero's algorithm. This new implementation can compute interactions from large datasets, such as atlases, in less time than most TV show episodes.

Single-cell technologies also allowed to investigate the communication between cells. Building on HuMMuS results, a second HMLN structure was developed to assemble several cell type networks and study their interactions. ReCoN leverages both GRN and cell-communication networks to understand the coordination of cell types in a tissue. ReCoN can predict the impact of an external molecule upon each cell type, considering the signals they will exchange. It can also be applied to more complex setups, such as the

activation of a biological function in one specific cell type, to understand its impact across the tissue. The underlying hypothesis for ReCoN development is that a perturbation doesn't have only a direct effect, but also triggers the production of secondary messengers by each impacted cell, which will also participate in the whole tissue response and coordination. By differentiating such direct and indirect effects, it allows the quantification of their individual contribution in the final predictions. Chapter 4 introduces ReCoN and evaluates the contribution of the cell communication, the gene regulatory network, and the indirect effects to ReCoN's performance. These different elements all showed a significant contribution. The indirect effect also showed a varying contribution across cell types, consistent with the literature and their level of specialization.

Together, these contributions highlight the potential of HMLN to represent biological systems, from detailed modeling of intracellular regulation (HuMMuS) to reconstruction of intercellular programs (ReCoN). The main takeaway I would like to share with whoever would read this thesis is the importance of considering co-operations for biological regulation, and not to consider them only as pair-wise interactions. While it comes with new computational challenges, it offers new opportunities to understand and predict system outcomes. Computational models are still in their emerging area, but they deserve to be heavily invested in to replace animal experimentations, and improve treatment personalisations in the near future.

Open-source packages have been developed for each of the proposed contributions, encouraging their use for understanding various biological conditions and their generalisation to other systems.

5.2. Depth of exploration and restarts as methodological limitations

A major parameter in RWR algorithms is the restart probability, which determines if we continue exploring deeper the network or restart at the seed. This probability is usually a constant, as in the works of this thesis, independent of how many steps have been realised or which steps are possible. In RWR for MLN, a second parameter is the layers transition matrix, which defines the probability of stepping from one layer to another. Similarly, the probabilities of transitioning between layers are constant (i.e. for each node of a layer A, the probability of stepping to B is independent of how many nodes can be reached in this layer).

In the context of HuMMuS, presented in Chapter 1, it means that all TFs will lead to the same number of transcription events, independently of the amplitude of their real effect or their number of targets and associated weights. For example, a DNA region linked to two genes with respective weights of 0.1 and 0.2 will reach them as often as a DNA region linked to them with weights of 10 and 20. A possible solution could be to vary the restart probability from each node, based on the inverse of the sum of the out-degree links, thus penalizing the targets if the sum of their incoming edge weights is low. Recent works took

interest in the subject, proposing notably random walk with extended restart (RWER)[304], that allows choosing or learning (based on a second algorithm) node-specific restart probabilities.

These probabilities could penalize elements less important in the regulatory networks of the cell. It is however possible that elements have different importance on distinct molecular layers. For example, a TF can have a low probability to link directly DNA regions, but a high affinity to form dimers with other TFs. We then would like penalizing the links to target DNA regions without impacting the TF - TF links. A more recent algorithm, CUSTARD[305], has been proposed to penalize links between nodes of different label groups, defining node-restart probabilities while reweighting the network. This method, once adapted to multilayer networks, could be an interesting way to take into account the fine specificities and affinities of each component of our network, or at least their interaction probabilities with each omic layer. It will however first require measurements of these properties, which will probably not be accessible at the cellular scale before important technological development.

5.3. Data quality for network inference and evaluation

5.3.1. Contextualise gene regulatory interactions

Differentiate cell specificities from noise

The performance of current GRN inference methods on scRNA-seq and single-cell multiomics data is modest, with high false positive rates being a common problem. It has been shown that inferred networks often contain many incorrect regulatory links, partly because noise and dropouts mask the true signal[5,249]. The use of prior TF-gene databases to filter these links can help, but these data are often limited to specie-specific information and thus strongly penalize cell type specificities. Moreover, these databases contain biases towards TFs and genes more studied than others in the literature. Indeed, many of these nodes have higher degrees in databases simply because more research and experiments have been carried out to identify their interactions. Current methods are confronted to a trade-off between correcting on generic knowledge – potentially introducing study-biases – and adding data artifacts and noise in their predictions. Overall, it highlights the importance of cell type specific experiments, such as perturbation, to validate network inference methods.

Beyond pair-wise interactions

In contrast, HuMMuS does not filter out information on prior knowledge but proposes to integrate different types of interactions in one large and exhaustive network. It however can't represent more than pair-wise interaction. For example, it cannot represent contextual interactions (i.e. interactions depending on a third molecule's actions in a cell), or TF dimers interaction with a DNA region at once. As an illustration, while if TF A and TF B form a dimer and are linked accordingly in HuMMuS, it is possible to connect any region bound by TF B from TF A too, it does not truly represent their joint action. If we remove the

TF B, the TF A cannot bind these regions anymore, but if we remove TF A, TF B binding will not be affected. This asymmetry denotes the absence of joint action representations. A solution to these problems could be to borrow the conceptual representation of hypergraphs, which introduces edges involving more than two nodes. New tools are developed to explore these generalised graph structures, along with the growing interest in more complex edge definitions[306].

In both cases, more contextual information is necessary on when interactions are observed, in which cell types and cell states.

Validation data of gene regulations in specific conditions

While we jumped technologically from bulk to single-cell data, evaluations of the methods building upon it are still limited by other types of knowledge that are still not cell (not even cell type) specific. This also has repercussions in the evaluation of these methods, since we lack a way to even evaluate cell and cell type-specific network predictions. As used in Chapter 2, we can evaluate GRN on both functional and physical criteria, with perturbation experiments and physical binding observations, respectively. However, the intersection of interactions validated by both metrics is extremely low. It could denote a strong indirect effect in perturbation studies (X perturbs Y, which itself perturbs Z; we infer X perturbs Z), or many physical binding without consequences (X binds the DNA near Y, but it does not affect transcription). In any case, it highlights the incompleteness of our validation tools and the unclear purpose of many gene regulatory inference methods. Are we proposing a fixed network of all the possible interactions, or a snapshot of the current regulatory mechanisms involved in a specific phenotype? Clearer definitions of GRN's purposes and of metrics for their evaluation seem essential in answering which methods could answer each of these very distinct roles[249].

5.3.2. Single-cell and spatial data for cell communication

As stated in the introduction, the distance between two cells is a major constraint in their communication. Spatial data can provide such information and are thus particularly informative to refine cell communication inference. In the current version of ReCoN, the cell communication network is solely based on single-cell data and the average expression of ligand and receptor across cell types. A natural development of ReCoN could be to integrate spatial information to modulate the communication strength between different cell types.

An important concept in spatial data is the concept of niches, or the repetition of cell type arrangement patterns across a tissue. It would be informative to build niche-specific models, differing by the communication strength between cell types and allowing the study of niche-specific behavior. For example, the difference in tissue regrowth after infarction or inhibition of cancer growth probably depends on the tissue environment. ReCoN might predict drivers of these differences in specific cell types, enabling the development of drugs targeting particular tissue regions.

Bibliography

1. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).

2. Silverman, E. K. *et al.* Molecular Networks in Network Medicine: Development and Applications. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **12**, e1489 (2020).

3. Fiscon, G., Conte, F., Farina, L. & Paci, P. Network-Based Approaches to Explore Complex Biological Systems towards Network Medicine. *Genes* **9**, 437 (2018).

4. Ideker, T. & Sharan, R. Protein networks in disease. *Genome Res.* **18**, 644–652 (2008).

5. Kang, Y., Thieffry, D. & Cantini, L. Evaluating the Reproducibility of Single-Cell Gene Regulatory Network Inference Algorithms. *Front. Genet.* **12**, 617282 (2021).

6. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17**, 147–154 (2020).

7. Fleck, J. S. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* 1–8 (2022) doi:10.1038/s41586-022-05279-8.

8. Bravo González-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods* **20**, 1355–1367 (2023).

9. Kamimoto, K. *et al.* Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**, 742–751 (2023).

10. Trimbour, R., Deutschmann, I. M. & Cantini, L. Molecular mechanisms reconstruction from single-cell multi-omics data with HuMMuS. *Bioinformatics* **40**, btae143 (2024).

11. Kivelä, M. *et al.* Multilayer networks. *J. Complex Netw.* **2**, 203–271 (2014).

12. Trigos, A. S., Pearson, R. B., Papenfuss, A. T. & Goode, D. L. How the evolution of multicellularity set the stage for cancer. *Br. J. Cancer* **118**, 145 (2018).

13. Hong, S. & Stevens, B. Microglia: Phagocytosing to Clear, Sculpt, and Eliminate. *Dev. Cell* **38**, 126–128 (2016).

14. Baruch, K. *et al.* PD-1 immune checkpoint blockade reduces pathology and improves memory in mouse models of Alzheimer's disease. *Nat. Med.* **22**, 135–137 (2016).

15. Qi, G., Mi, Y. & Yin, F. Cellular Specificity and Inter-cellular Coordination in the Brain Bioenergetic System: Implications for Aging and Neurodegeneration. *Front. Physiol.* **10**, 1531 (2020).

16. Tanevski, J. *et al.* Learning tissue representation by identification of persistent local patterns in spatial omics data. *Nat. Commun.* **16**, 4071 (2025).

17. Ramirez Flores, R. O., Lanzer, J. D., Dimitrov, D., Velten, B. & Saez-Rodriguez, J. Multicellular factor analysis of single-cell data for a tissue-centric understanding of disease. *eLife* **12**, e93161 (2023).

18. Jerby-Arnon, L. & Regev, A. DIALOGUE maps multicellular programs in tissue from single-cell or spatial transcriptomics data. *Nat. Biotechnol.* **40**, 1467–1477 (2022).

19. Dimitrov, D. *et al.* LIANA+ provides an all-in-one framework for cell–cell communication inference. *Nat. Cell Biol.* **26**, 1613–1622 (2024).

20. Tanevski, J., Flores, R. O. R., Gabor, A., Schapiro, D. & Saez-Rodriguez, J. Explainable multiview framework for dissecting spatial relationships from highly multiplexed data. *Genome Biol.* **23**, 97 (2022).

21. Cui, A. *et al.* Dictionary of immune responses to cytokines at single-cell resolution. *Nature* **625**, 377–384 (2024).

22. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2020).

23. Lanzer, J. D., Flores, R. O. R., Blanco, J. L. & Saez-Rodriguez, J. A cross-study transcriptional patient map of heart failure defines conserved multicellular coordination in cardiac remodeling. 2024.11.04.621815 Preprint at https://doi.org/10.1101/2024.11.04.621815 (2024).

24. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

25. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

26. Bonner, W. A., Hulett, H. R., Sweet, R. G. & Herzenberg, L. A. Fluorescence Activated Cell Sorting. *Rev. Sci. Instrum.* **43**, 404–409 (1972).

27. Herzenberg, L. A. *et al.* The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford. *Clin. Chem.* **48**, 1819–1827 (2002).

28. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).

29. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

30. Orzolek, L. D. Sequencing: 10X Genomics 3' HT Assay for Gene Expression. *Methods Mol. Biol. Clifton NJ* **2822**, 207–226 (2024).

31. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

32. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).

33. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).

34. Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337 (2021).

35. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).

36. Luo, C. *et al.* Robust single-cell DNA methylome profiling with snmC-seq2. *Nat. Commun.* **9**, 3824 (2018).

37. Liu, S. & Trapnell, C. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Research* **5**, F1000 Faculty Rev-182 (2016).

38. Method of the Year 2019: Single-cell multimodal omics. *Nat. Methods* **17**, 1–1 (2020).

39. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).

40. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).

41. Lee, J., Hyeon, D. Y. & Hwang, D. Single-cell multiomics: technologies and data analysis methods. *Exp. Mol. Med.* **52**, 1428–1442 (2020).

42. Williams, C. G., Lee, H. J., Asatsuma, T., Vento-Tormo, R. & Haque, A. An introduction to spatial transcriptomics for biomedical research. *Genome Med.* **14**, 68 (2022).

43. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).

44. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).

45. Maynard, K. R. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).

46. Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).

47. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

48. Su, M. *et al.* Data analysis guidelines for single-cell RNA-seq in biomedical studies and clinical applications. *Mil. Med. Res.* **9**, 68 (2022).

49. Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* 1–23 (2023) doi:10.1038/s41576-023-00586-w.

50. Borella, M., Martello, G., Risso, D. & Romualdi, C. PsiNorm: a scalable normalization for single-cell RNA-seq data. *Bioinformatics* **38**, 164–172 (2021).

51. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).

52. Ahlmann-Eltze, C. & Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nat. Methods* **20**, 665–672 (2023).

53. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

54. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

55. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).

56. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

57. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).

58. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

59. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

60. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).

61. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLOS ONE* **5**, e12776 (2010).

62. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).

63. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

64. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).

65. Jerby-Arnon, L. *et al.* A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. *Cell* **175**, 984-997.e24 (2018).

66. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

67. McFarland, J. M. *et al.* Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat. Commun.* **11**, 4296 (2020).

68. Tabula Sapiens Consortium* *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).

69. Jiang, R., Sun, T., Song, D. & Li, J. J. Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biol.* **23**, 31 (2022).

70. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).

71. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 284 (2018).

72. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).

73. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664 (2002).

74. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).

75. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356 (1961).

76. Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**, 291–294 (1998).

77. Krebs, H. A. The citric acid cycle and the Szent-Györgyi cycle in pigeon breast muscle. *Biochem. J.* **34**, 775–779 (1940).

78. Dourado, H., Kuate, C. A. F., Liebermeister, W. & Lercher, M. J. Growth Mechanics and the emergence of metabolic oscillations in growing cells. 2025.06.24.661369 Preprint at https://doi.org/10.1101/2025.06.24.661369 (2025).

79. Alberts, B. *et al.* Cell Communication. in *Molecular Biology of the Cell. 4th edition* (Garland Science, 2002).

80. Tse, L. H. & Wong, Y. H. GPCRs in Autocrine and Paracrine Regulations. *Front. Endocrinol.* **10**, 428 (2019).

81. Saltiel, A. R. & Kahn, C. R. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* **414**, 799–806 (2001).

82. Bishop, E. L., Gudgeon, N. & Dimeloe, S. Control of T Cell Metabolism by Cytokines and Hormones. *Front. Immunol.* **12**, (2021).

83. Klezovitch, O. & Vasioukhin, V. Cadherin signaling: keeping cells in touch. *F1000Research* **4**, 550 (2015).

84. Hynes, R. O. Integrins: bidirectional, allosteric signaling machines. *Cell* **110**, 673–687 (2002).

85. Guo, Y.-H. & Yang, Y.-Q. Atrial Fibrillation: Focus on Myocardial Connexins and Gap Junctions. *Biology* **11**, 489 (2022).

86. Harfe, B. D. *et al.* Evidence for an Expansion-Based Temporal Shh Gradient in Specifying Vertebrate Digit Identities. *Cell* **118**, 517–528 (2004).

87. Scaffidi, P., Misteli, T. & Bianchi, M. E. Release of chromatin protein HMGB1 by necrotic cells triggers inflammation. *Nature* **418**, 191–195 (2002).

88. Forsythe, J. A. *et al.* Activation of vascular endothelial growth factor gene transcription by hypoxia-inducible factor 1. *Mol. Cell. Biol.* **16**, 4604–4613 (1996).

89. Davies, D. G. *et al.* The involvement of cell-to-cell signals in the development of a bacterial biofilm. *Science* **280**, 295–298 (1998).

90. Barrat, A., Barthélemy, M., Pastor-Satorras, R. & Vespignani, A. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 3747–3752 (2004).

91. Safari-Alighiarloo, N., Taghizadeh, M., Rezaei-Tavirani, M., Goliaei, B. & Peyvandi, A. A. Protein-protein interaction networks (PPI) and complex diseases. *Gastroenterol. Hepatol. Bed Bench* **7**, 17–31 (2014).

92. Young, K. H. Yeast Two-hybrid: So Many Interactions, (in) So Little Time…. *Biol. Reprod.* **58**, 302–311 (1998).

93. Morris, J. H. *et al.* Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions. *Nat. Protoc.* **9**, 2539–2554 (2014).

94. Mansuri, M. S., Williams, K. & Nairn, A. C. Uncovering biology by single-cell proteomics. *Commun. Biol.* **6**, 381 (2023).

95. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

96. Hsieh, T.-H. S. *et al.* Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**, 108–119 (2015).

97. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).

98. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).

99. Pliner, H. A. *et al.* Cicero predicts cis-regulatory DNA interactions from single cell chromatin accessibility data. *Mol. Cell* **71**, 858-871.e8 (2018).

100. Murakami, K., Iida, K. & Okada, M. An Attention-Based Deep Neural Network Model to Detect Cis-Regulatory Elements at the Single-Cell Level From Multi-Omics Data. *Genes Cells* **30**, e70000 (2025).

101. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).

102. Margolin, A. A. *et al.* ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* **7**, S7 (2006).

103. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

104. Zhang, S. *et al.* Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. *Nat. Commun.* **14**, 3064 (2023).

105. Duren, Z., Chen, X., Xin, J., Wang, Y. & Wong, W. H. Time course regulatory analysis based on paired expression and chromatin accessibility data. *Genome Res.* **30**, 622–634 (2020).

106. Wang, L. *et al.* Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. *Nat. Methods* **20**, 1368–1378 (2023).

107. Lee, T. I. *et al.* Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799–804 (2002).

108. Badia-i-Mompel, P. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* 1–16 (2023) doi:10.1038/s41576-023-00618-5.

109. Demir, E. *et al.* The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.* **28**, 935–942 (2010).

110. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).

111. Lo Surdo, P. *et al.* SIGNOR 3.0, the SIGnaling network open resource 3.0: 2022 update. *Nucleic Acids Res.* **51**, D631–D637 (2022).

112. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).

113. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).

114. Mahmood, T. & Yang, P.-C. Western blot: technique, theory, and trouble shooting. *North Am. J. Med. Sci.* **4**, 429–434 (2012).

115. Shults, M. D., Janes, K. A., Lauffenburger, D. A. & Imperiali, B. A multiplexed homogeneous fluorescence-based assay for protein kinase activity in cell lysates. *Nat. Methods* **2**, 277–283 (2005).

116. Harlow, E. & Lane, D. Immunoprecipitation: purifying the immune complexes. *CSH Protoc.* **2006**, pdb.prot4536 (2006).

117. Jares-Erijman, E. A. & Jovin, T. M. FRET imaging. *Nat. Biotechnol.* **21**, 1387–1395 (2003).

118. Angers, S. *et al.* Detection of beta 2-adrenergic receptor dimerization in living cells using bioluminescence resonance energy transfer (BRET). *Proc. Natl. Acad. Sci. U. S. A.* **97**, 3684–3689 (2000).

119. Hou, R., Denisenko, E., Ong, H. T., Ramilowski, J. A. & Forrest, A. R. R. Predicting cell-to-cell communication networks using NATMI. *Nat. Commun.* **11**, 5011 (2020).

120. Raredon, M. S. B. *et al.* Computation and visualization of cell–cell signaling topologies in single-cell systems data using Connectome. *Sci. Rep.* **12**, 4187 (2022).

121. 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).

122. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* **12**, 1088 (2021).

123. Troulé, K. *et al.* CellPhoneDB v5: inferring cell–cell communication from single-cell multiomics data. *Nat. Protoc.* 1–29 (2025) doi:10.1038/s41596-024-01137-1.

124. Zohora, F. T. *et al.* CellNEST reveals cell–cell relay networks using attention mechanisms on spatial transcriptomics. *Nat. Methods* 1–15 (2025) doi:10.1038/s41592-025-02721-3.

125. Dries, R. *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22**, 78 (2021).

126. Bafna, M., Li, H. & Zhang, X. CLARIFY: cell-cell interaction and gene regulatory network refinement from spatially resolved transcriptomics. *Bioinforma. Oxf. Engl.* **39**, i484–i493 (2023).

127. Koschützki, D. & Schreiber, F. Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks. *Gene Regul. Syst. Biol.* **2**, 193–201 (2008).

128. He, X. & Zhang, J. Why Do Hubs Tend to Be Essential in Protein Networks? *PLOS Genet.* **2**, e88 (2006).

129. Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. The Importance of Bottlenecks in Protein Networks: Correlation with Gene Essentiality and Expression Dynamics. *PLOS Comput. Biol.* **3**, e59 (2007).

130. Dugourd, A. & Saez-Rodriguez, J. Footprint-based functional analysis of multiomic data. *Curr. Opin. Syst. Biol.* **15**, 82–90 (2019).

131. Alvarez, M. J. *et al.* Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847 (2016).

132. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.* **29**, 1363–1375 (2019).

133. Schubert, M. *et al.* Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.* **9**, 20 (2018).

134. Badia-i-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinforma. Adv.* **2**, vbac016 (2022).

135. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99**, 7821–7826 (2002).

136. Alcalá-Corona, S. A., Sandoval-Motta, S., Espinal-Enríquez, J. & Hernández-Lemus, E. Modularity in Biological Networks. *Front. Genet.* **12**, 701331 (2021).

137. Lewis, A. C., Jones, N. S., Porter, M. A. & Deane, C. M. The function of communities in protein interaction networks at multiple scales. *BMC Syst. Biol.* **4**, 100 (2010).

138. Page, L., Brin, S., Motwani, R. & Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. http://ilpubs.stanford.edu:8090/422/?utm_campaign=Technical%20SEO%20Weekly&utm_medium=email&utm_source=Revue%20newsletter (1999).

139. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* **18**, 551–562 (2017).

140. Di Nanni, N., Bersanelli, M., Milanesi, L. & Mosca, E. Network Diffusion Promotes the Integrative Analysis of Multiple Omics. *Front. Genet.* **11**, (2020).

141. Sun, L., Yin, Z. & Lu, L. ISLRWR: A network diffusion algorithm for drug–target interactions prediction. *PLOS ONE* **20**, e0302281 (2025).

142. Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for Detecting Significantly Mutated Pathways in Cancer. *J. Comput. Biol.* **18**, 507–522 (2011).

143. Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the Interactome for Prioritization of Candidate Disease Genes. *Am. J. Hum. Genet.* **82**, 949–958 (2008).

144. Tong, H., Faloutsos, C. & Pan, J. Fast Random Walk with Restart and Its Applications. in *Sixth International Conference on Data Mining (ICDM'06)* 613–622 (2006). doi:10.1109/ICDM.2006.70.

145. Navlakha, S. & Kingsford, C. The power of protein interaction networks for associating genes with diseases. *Bioinforma. Oxf. Engl.* **26**, 1057–1063 (2010).

146. Baptista, A., Gonzalez, A. & Baudot, A. Universal multilayer network exploration by random walk with restart. *Commun. Phys.* **5**, 1–9 (2022).

147. Gómez, S. *et al.* Diffusion Dynamics on Multiplex Networks. *Phys. Rev. Lett.* **110**, 028701 (2013).

148. Kumar, A., Rai, P. & Daume, H. Co-regularized Multi-view Spectral Clustering. in *Advances in Neural Information Processing Systems* vol. 24 (Curran Associates, Inc., 2011).

149. Hu, B., Wang, X., Zhou, P. & Du, L. Multi-view Outlier Detection via Graphs Denoising. *Inf. Fusion* **101**, 102012 (2024).

150. Wen, J. *et al.* Highly Confident Local Structure Based Consensus Graph Learning for Incomplete Multi-view Clustering. in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 15712–15721 (2023). doi:10.1109/CVPR52729.2023.01508.

151. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).

152. Xiao, S. *et al.* Graph neural networks for multi-view learning: a taxonomic review. *Artif. Intell. Rev.* **57**, 341 (2024).

153. Rhodes, J. S. & Rustad, A. G. Graph Integration for Diffusion-Based Manifold Alignment. Preprint at https://doi.org/10.48550/arXiv.2410.22978 (2024).

154. Pidò, S., Ceddia, G. & Masseroli, M. Computational analysis of fused co-expression networks for the identification of candidate cancer gene biomarkers. *Npj Syst. Biol. Appl.* **7**, 17 (2021).

155. Zheng, X. *et al.* Fusing multiple protein-protein similarity networks to effectively predict lncRNA-protein interactions. *BMC Bioinformatics* **18**, 420 (2017).

156. Barnes, J. A. Class and Committees in a Norwegian Island Parish. *Hum. Relat.* **7**, 39–58 (1954).

157. White, H. C., Boorman, S. A. & Breiger, R. L. Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions. *Am. J. Sociol.* **81**, 730–780 (1976).

158. Mucha, P. J., Richardson, T., Macon, K., Porter, M. A. & Onnela, J.-P. Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science* **328**, 876–878 (2010).

159. Baptista, A., Gonzalez, A. & Baudot, A. Universal multilayer network exploration by random walk with restart. *Commun. Phys.* **5**, 170 (2022).

160. Lee, B., Zhang, S., Poleksic, A. & Xie, L. Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis. *Front. Genet.* **10**, (2020).

161. Baptista, A., Brière, G. & Baudot, A. Random walk with restart on multilayer networks: from node prioritisation to supervised link prediction and beyond. *BMC Bioinformatics* **25**, 70 (2024).

162. Laska, J. & Narayan, M. skggm 0.2.8: A scikit-learn compatible package for general graphical models. Zenodo https://doi.org/10.5281/zenodo.1413742 (2018).

163. Morris, S. A. The evolving concept of cell identity in the single cell era. *Dev. Camb. Engl.* **146**, dev169748 (2019).

164. Nawy, T. Single-cell sequencing. *Nat. Methods* **11**, 18–18 (2014).

165. Method of the Year 2013. *Nat. Methods* **11**, 1–1 (2014).

166. Mimitou, E. P. *et al.* Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods* **16**, 409–412 (2019).

167. Method of the Year 2019: Single-cell multimodal omics. *Nat. Methods* **17**, 1–1 (2020).

168. Jiang, Y. *et al.* Nonparametric single-cell multiomic characterization of trio relationships between transcription factors, target genes, and cis-regulatory regions. *Cell Syst.* **13**, 737-751.e4 (2022).

169. Kartha, V. K. *et al.* Functional inference of gene regulation using single-cell multi-omics. *Cell Genomics* **2**, 100166 (2022).

170. Skok Gibbs, C. *et al.* High-performance single-cell gene regulatory network inference at scale: the Inferelator 3.0. *Bioinformatics* **38**, 2519–2528 (2022).

171. Ma, A. *et al.* Single-cell biological network inference using a heterogeneous graph transformer. *Nat. Commun.* **14**, 964 (2023).

172. McCalla, S. G. *et al.* Identifying strengths and weaknesses of methods for computational network inference from single-cell RNA-seq data. *G3 Bethesda Md* **13**, jkad004 (2023).

173. Badia-i-Mompel, P. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* (2023) doi:10.1038/s41576-023-00618-5.

174. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2022).

175. Hammal, F., de Langen, P., Bergon, A., Lopez, F. & Ballester, B. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.* **50**, D316–D325 (2022).

176. Guo, Y. & Gifford, D. K. Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. *BMC Genomics* **18**, 45 (2017).

177. Kribelbauer, J. F. *et al.* Context transcription factors establish cooperative environments and mediate enhancer communication. 2023.05.05.539543 Preprint at https://doi.org/10.1101/2023.05.05.539543 (2023).

178. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88-92 (2007).

179. Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).

180. Naville, M. *et al.* Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome. *Nat. Commun.* **6**, 6904 (2015).

181. Bai, X. *et al.* ENdb: a manually curated database of experimentally supported enhancers for human and mouse. *Nucleic Acids Res.* **48**, D51–D57 (2020).

182. Clément, Y., Torbey, P., Gilardi-Hebenstreit, P. & Crollius, H. R. Enhancer–gene maps in the human and zebrafish genomes using evolutionary linkage conservation. *Nucleic Acids Res.* **48**, 2357–2371 (2020).

183. Gao, T. & Qian, J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.* **48**, D58–D64 (2020).

184. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

185. Cantini, L., Medico, E., Fortunato, S. & Caselle, M. Detection of gene communities in multi-networks reveals cancer drivers. *Sci. Rep.* **5**, 17386 (2015).

186. Choobdar, S. *et al.* Assessment of network module identification across complex diseases. *Nat. Methods* **16**, 843–852 (2019).

187. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*

28, 27–30 (2000).

188. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50**, D687–D692 (2022).

189. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

190. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).

191. Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174**, 1015-1030.e16 (2018).

192. atac_v1_adult_brain_fresh_5k -Datasets -Single Cell ATAC -Official 10x Genomics Support. https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_adult_brain_fresh_5k?

193. Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).

194. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022).

195. Teschendorff, A. E. & Wang, N. Improved detection of tumor suppressor events in single-cell RNA-Seq data. *Npj Genomic Med.* **5**, 1–14 (2020).

196. Bastian, F. B. *et al.* The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Res.* **49**, D831–D847 (2021).

197. Kleine-Kohlbrecher, D. *et al.* A functional link between the histone demethylase PHF8 and the transcription factor ZNF711 in X-linked mental retardation. *Mol. Cell* **38**, 165–178 (2010).

198. Zou, M., Li, S., Klein, W. H. & Xiang, M. Brn3a/Pou4f1 regulates dorsal root ganglion sensory neuron specification and axonal projection into the spinal cord. *Dev. Biol.* **364**, 114–127 (2012).

199. Hendrich, B. & Bird, A. Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol. Cell. Biol.* **18**, 6538–6547 (1998).

200. Dame, C. *et al.* Wilms tumor suppressor, Wt1, is a transcriptional activator of the erythropoietin gene. *Blood* **107**, 4282–4290 (2006).

201. Müller, T. *et al.* The bHLH factor Olig3 coordinates the specification of dorsal neurons in the spinal cord. *Genes Dev.* **19**, 733–743 (2005).

202. Casado-Navarro, R. & Serrano-Saiz, E. DMRT Transcription Factors in the Control of Nervous System Sexual Differentiation. *Front. Neuroanat.* **16**, 937596 (2022).

203. Russ, D. E. *et al.* A harmonized atlas of mouse spinal cord cell types and their spatial organization. *Nat. Commun.* **12**, 5722 (2021).

204. Gavalas, A., Davenne, M., Lumsden, A., Chambon, P. & Rijli, F. M. Role of Hoxa-2 in axon pathfinding and rostral hindbrain patterning. *Dev. Camb. Engl.* **124**, 3693–3702 (1997).

205. Flore, G., Cioffi, S., Bilio, M. & Illingworth, E. Cortical Development Requires Mesodermal Expression of Tbx1, a Gene Haploinsufficient in 22q11.2 Deletion Syndrome. *Cereb. Cortex N. Y. N 1991* **27**, 2210–2225 (2017).

206. Callaway, E. M. *et al.* A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86–102 (2021).

207. Dixit, R. *et al.* Neurog1 and Neurog2 control two waves of neuronal differentiation in the piriform cortex. *J. Neurosci. Off. J. Soc. Neurosci.* **34**, 539–553 (2014).

208. Gezen-Ak, D., Dursun, E. & Yilmazer, S. The effects of vitamin D receptor silencing on the expression of LVSCC-A1C and LVSCC-A1D and the release of NGF in cortical neurons. *PloS One* **6**, e17553 (2011).

209. Turner, E. E., Jenne, K. J. & Rosenfeld, M. G. Brn-3.2: a Brn-3-related transcription factor with distinctive central nervous system expression and regulation by retinoic acid. *Neuron* **12**, 205–218 (1994).

210. Cinquanta, M., Rovescalli, A. C., Kozak, C. A. & Nirenberg, M. Mouse Sebox homeobox gene expression in skin, brain, oocytes, and two-cell embryos. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 8904–8909 (2000).

211. Cardo, L. F., de la Fuente, D. C. & Li, M. Impaired neurogenesis and neural progenitor fate choice in a human stem cell model of SETBP1 disorder. *Mol. Autism* **14**, 8 (2023).

212. Golonzhka, O. *et al.* Pbx Regulates Patterning of the Cerebral Cortex in Progenitors and Postmitotic Neurons. *Neuron* **88**, 1192–1207 (2015).

213. Wang, J. *et al.* Regulation of neural stem cell differentiation by transcription factors HNF4-1 and MAZ-1. *Mol. Neurobiol.* **47**, 228–240 (2013).

214. Ning, Z. *et al.* Regulation of SPRY3 by X chromosome and PAR2-linked promoters in an autism susceptibility region. *Hum. Mol. Genet.* **24**, 5126–5141 (2015).

215. Okano, T., Sasaki, M. & Fukada, Y. Cloning of mouse BMAL2 and its daily expression profile in the suprachiasmatic nucleus: a remarkable acceleration of Bmal2 sequence divergence after Bmal gene duplication. *Neurosci. Lett.* **300**, 111–114 (2001).

216. Ohba, K. *et al.* Microphthalmia-associated transcription factor ensures the elongation of axons and dendrites in the mouse frontal cortex. *Genes Cells Devoted Mol. Cell. Mech.* **21**, 1365–1379 (2016).

217. Nagalski, A. *et al.* Postnatal isoform switch and protein localization of LEF1 and TCF7L2 transcription factors in cortical, thalamic, and mesencephalic regions of the adult mouse brain. *Brain Struct. Funct.* **218**, 1531–1549 (2013).

218. Gray, L. T. *et al.* Layer-specific chromatin accessibility landscapes reveal regulatory networks in adult mouse visual cortex. *eLife* **6**, e21883 (2017).

219. Williams, R. H. & Riedemann, T. Development, Diversity, and Death of MGE-Derived Cortical Interneurons. *Int. J. Mol. Sci.* **22**, 9297 (2021).

220. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

221. Alam, S., Zinyk, D., Ma, L. & Schuurmans, C. Members of the Plag gene family are expressed in complementary and overlapping regions in the developing murine nervous system. *Dev. Dyn. Off. Publ. Am. Assoc. Anat.* **234**, 772–782 (2005).

222. Mall, M. *et al.* Myt1l safeguards neuronal identity by actively repressing many non-neuronal fates. *Nature* **544**, 245–249 (2017).

223. Davenne, M. *et al.* Hoxa2 and Hoxb2 Control Dorsoventral Patterns of Neuronal Development in the Rostral Hindbrain. *Neuron* **22**, 677–691 (1999).

224. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).

225. Zhang, L., Zhang, J. & Nie, Q. DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. *Sci. Adv.* **8**, eabl7393 (2022).

226. Sakaue, S. *et al.* Tissue-specific enhancer–gene maps from multimodal single-cell data identify causal disease alleles. *Nat. Genet.* **56**, 615–626 (2024).

227. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).

228. Grover, A. *et al.* UniversalEPI: Harnessing Attention Mechanisms to Decode Chromatin Interactions in Rare and Unexplored Cell Types. 2024.11.22.624813 Preprint at https://doi.org/10.1101/2024.11.22.624813 (2024).

229. Sheth, M. U. *et al.* Mapping enhancer-gene regulatory interactions from single-cell data. 2024.11.23.624931 Preprint at https://doi.org/10.1101/2024.11.23.624931 (2024).

230. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324.e18 (2018).

231. Yao, Z. *et al.* A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature* **598**, 103–110 (2021).

232. Zhang, K. *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985-6001.e19 (2021).

233. Chari, T. & Pachter, L. The specious art of single-cell genomics. *PLOS Comput. Biol.* **19**, e1011288 (2023).

234. Kobak, D. & Linderman, G. C. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat. Biotechnol.* **39**, 156–157 (2021).

235. Virshup, I. *et al.* The scverse project provides a computational ecosystem for single-cell omics

data analysis. *Nat. Biotechnol.* **41**, 604–606 (2023).

236. Baran, Y. *et al.* MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol.* **20**, 206 (2019).

237. Bilous, M., Hérault, L., Gabriel, A. A., Teleman, M. & Gfeller, D. Building and analyzing metacells in single-cell genomics data. *Mol. Syst. Biol.* **20**, 744–766 (2024).

238. Persad, S. *et al.* SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nat. Biotechnol.* **41**, 1746–1757 (2023).

239. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).

240. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E6456-6465 (2015).

241. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).

242. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

243. Laska, J. & Narayan, M. skggm 0.2.7: A scikit-learn compatible package for Gaussian and related Graphical Models. (2017) doi:10.5281/zenodo.830033.

244. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. *Proc. 7th Python Sci. Conf. SciPy 2008* 11–15 (2008) doi:10.25080/TCWV9851.

245. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Wolf, F. A. anndata: Access and store annotated data matrices. *J. Open Source Softw.* **9**, 4371 (2024).

246. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384.e19 (2016).

247. Martens, L. D., Fischer, D. S., Yépez, V. A., Theis, F. J. & Gagneur, J. Modeling fragment counts improves single-cell ATAC-seq analysis. *Nat. Methods* **21**, 28–31 (2024).

248. Luecken, M. *et al.* A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks* **1**, (2021).

249. Badia-i-Mompel, P. *et al.* Comparison and evaluation of methods to infer gene regulatory networks from multimodal single-cell data. 2024.12.20.629764 Preprint at https://doi.org/10.1101/2024.12.20.629764 (2024).

250. Zhang, K., Zemke, N. R., Armand, E. J. & Ren, B. A fast, scalable and versatile tool for analysis of single-cell omics data. *Nat. Methods* **21**, 217–227 (2024).

251. Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).

252. Raghavan, S. *et al.* Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. *Cell* **184**, 6119-6137.e26 (2021).

253. Van Roy, Z. *et al.* Tissue niche influences immune and metabolic profiles to Staphylococcus aureus biofilm infection. *Nat. Commun.* **15**, 8965 (2024).

254. Bain, C. C. & MacDonald, A. S. The impact of the lung environment on macrophage development, activation and function: diversity in the face of adversity. *Mucosal Immunol.* **15**, 223–234 (2022).

255. Fan, Y. *et al.* Ultrafast distant wound response is essential for whole-body regeneration. *Cell* **186**, 3606-3618.e16 (2023).

256. Griffiths, J. I. *et al.* Cellular interactions within the immune microenvironment underpins resistance to cell cycle inhibition in breast cancers. *Nat. Commun.* **16**, 2132 (2025).

257. Salvati, L., Maggi, L., Annunziato, F. & Cosmi, L. Thymic stromal lymphopoietin and alarmins as possible therapeutical targets for asthma. *Curr. Opin. Allergy Clin. Immunol.* **21**, 590 (2021).

258. Mitchel, J. *et al.* Coordinated, multicellular patterns of transcriptional variation that stratify patient cohorts are revealed by tensor decomposition. *Nat. Biotechnol.* 1–10 (2024) doi:10.1038/s41587-024-02411-z.

259. Didier, G., Brun, C. & Baudot, A. Identifying communities from multiplex biological networks. *PeerJ* **3**, e1525 (2015).

260. Bennett, L., Kittas, A., Muirhead, G., Papageorgiou, L. G. & Tsoka, S. Detection of Composite Communities in Multiplex Biological Networks. *Sci. Rep.* **5**, 10345 (2015).

261. Boccaletti, S. *et al.* The structure and dynamics of multilayer networks. *Phys. Rep.* **544**, 1–122 (2014).

262. Sang-aram, C., Browaeys, R., Seurinck, R. & Saeys, Y. Unraveling cell-cell communication with NicheNet by inferring active ligands from transcriptomics data. Preprint at https://doi.org/10.48550/arXiv.2404.16358 (2024).

263. Heras-Murillo, I., Adán-Barrientos, I., Galán, M., Wculek, S. K. & Sancho, D. Dendritic cells as orchestrators of anticancer immunity and immunotherapy. *Nat. Rev. Clin. Oncol.* **21**, 257–277 (2024).

264. Qian, C. & Cao, X. Dendritic cells in the regulation of immunity and inflammation. *Semin. Immunol.* **35**, 3–11 (2018).

265. Michea, P. *et al.* Adjustment of dendritic cells to the breast-cancer microenvironment is subset specific. *Nat. Immunol.* **19**, 885–897 (2018).

266. Gardner, J. M. *et al.* Extrathymic Aire-Expressing Cells are a Distinct Bone Marrow-Derived Population that Induce Functional Inactivation of CD4+ T Cells. *Immunity* **39**, 560–572 (2013).

267. Kondělková, K. *et al.* Regulatory T cells (TREG) and their roles in immune system with respect to immunopathological disorders. *Acta Medica (Hradec Kralove)* **53**, 73–77 (2010).

268. Ossina, N. K. *et al.* Interferon-γ Modulates a p53-independent Apoptotic Pathway and Apoptosis-related Gene Expression*. *J. Biol. Chem.* **272**, 16351–16357 (1997).

269. Momenilandi, M. *et al.* FLT3L governs the development of partially overlapping hematopoietic lineages in humans and mice. *Cell* **187**, 2817-2837.e31 (2024).

270. Kanemaru, K. *et al.* Spatially resolved multiomics of human cardiac niches. *Nature* **619**, 801–810 (2023).

271. Naba, A. *et al.* The matrisome: in silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. *Mol. Cell. Proteomics MCP* **11**, M111.014647 (2012).

272. Liao, R. *et al.* E2F transcription factor 1 (E2F1) promotes the transforming growth factor TGF-β1 induced human cardiac fibroblasts differentiation through promoting the transcription of CCNE2 gene. *Bioengineered* **12**, 6869–6877.

273. Omura, J. *et al.* Identification of Long Noncoding RNA H19 as a New Biomarker and Therapeutic Target in Right Ventricular Failure in Pulmonary Arterial Hypertension. *Circulation* **142**, 1464–1484 (2020).

274. Zacharopoulou, N. *et al.* The epigenetic factor KDM2B regulates cell adhesion, small rho GTPases, actin cytoskeleton and migration in prostate cancer cells. *Biochim. Biophys. Acta Mol. Cell Res.* **1865**, 587–597 (2018).

275. Liu, S. *et al.* Role for the F-box proteins in heart diseases. *Pharmacol. Res.* **210**, 107514 (2024).

276. Yin, L. *et al.* FBXL10 regulates cardiac dysfunction in diabetic cardiomyopathy via the PKC β2 pathway. *J. Cell. Mol. Med.* **23**, 2558–2567 (2019).

277. Janzer, A. *et al.* The H3K4me3 histone demethylase Fbxl10 is a regulator of chemokine expression, cellular morphology, and the metabolome of fibroblasts. *J. Biol. Chem.* **287**, 30984–30992 (2012).

278. Ye, Z., Li, W., Jiang, Z., Wang, E. & Wang, J. An intermediate state in trans-differentiation with proliferation, metabolic, and epigenetic switching. *iScience* **24**, 103057 (2021).

279. Lederer, M., Jockusch, B. M. & Rothkegel, M. Profilin regulates the activity of p42POP, a novel Myb-related transcription factor. *J. Cell Sci.* **118**, 331–341 (2005).

280. Zhao, S. *et al.* Profilin-1 promotes the development of hypertension-induced cardiac hypertrophy. *J. Hypertens.* **31**, 576–586; discussion 586 (2013).

281. Yang, D. *et al.* Profilin-1 contributes to cardiac injury induced by advanced glycation end-products in rats. *Mol. Med. Rep.* **16**, 6634–6641 (2017).

282. Mihajlović, K. Data integration of longitudinal single-cell with multi-omics data to enable precision medicine for complex diseases. *TDX (Tesis Doctorals en Xarxa)* (Universitat Politècnica de Catalunya, 2025). doi:10.5821/dissertation-2117-432840.

283. Einspahr, J. *et al.* Loss of cardiomyocyte-specific adhesion G-protein-coupled receptor G1 (ADGRG1/GPR56) promotes pressure overload-induced heart failure. *Biosci. Rep.* **44**,

BSR20240826 (2024).

284. Lema, D. A. *et al.* IL1RAP Blockade With a Monoclonal Antibody Reduces Cardiac Inflammation and Preserves Heart Function in Viral and Autoimmune Myocarditis. *Circ. Heart Fail.* **17**, e011729 (2024).

285. Toth, D. *et al.* Protective Effects of PACAP in Peripheral Organs. *Front. Endocrinol.* **11**, (2020).

286. Szabó, D. *et al.* PACAP-38 and PAC1 Receptor Alterations in Plasma and Cardiac Tissue Samples of Heart Failure Patients. *Int. J. Mol. Sci.* **23**, 3715 (2022).

287. Racz, B. *et al.* Protective effect of PACAP against doxorubicin-induced cell death in cardiomyocyte culture. *J. Mol. Neurosci. MN* **42**, 419–427 (2010).

288. Virag, J. A. I. *et al.* Fibroblast Growth Factor-2 Regulates Myocardial Infarct Repair. *Am. J. Pathol.* **171**, 1431–1440 (2007).

289. Hara, H., Takeda, N. & Komuro, I. Pathophysiology and therapeutic potential of cardiac fibrosis. *Inflamm. Regen.* **37**, 13 (2017).

290. Rajabi, M., Kassiotis, C., Razeghi, P. & Taegtmeyer, H. Return to the fetal gene program protects the stressed heart: a strong hypothesis. *Heart Fail. Rev.* **12**, 331–343 (2007).

291. cantinilab/Circe. Machine Learning for Integrative Genomics lab (2025).

292. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2019).

293. Barsi, S. *et al.* RIDDEN: Data-driven inference of receptor activity from transcriptomic data. 2024.12.03.626558 Preprint at https://doi.org/10.1101/2024.12.03.626558 (2024).

294. Jiang, P. *et al.* Systematic investigation of cytokine signaling activity at the tissue and single-cell levels. *Nat. Methods* **18**, 1181–1191 (2021).

295. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

296. Brin, S. & Page, L. The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.* **30**, 107–117 (1998).

297. Guala, D. & Sonnhammer, E. L. L. A large-scale benchmark of gene prioritization methods. *Sci. Rep.* **7**, 46598 (2017).

298. Zhao, Z.-Q., Han, G.-S., Yu, Z.-G. & Li, J. Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization. *Comput. Biol. Chem.* **57**, 21–28 (2015).

299. Mixture of Cells from Mouse Lymph Nodes and Spleen Stained with TotalSeq™-C Mouse Universal Cocktail (Next GEM). *10x Genomics* https://www.10xgenomics.com/datasets/Mixture-of-cells-from-mouse-lymph-nodes-and-spleen-stained-with-totalseqc-mouse-universal-cocktail.

300. Simon, M. *et al.* Single-cell chromatin accessibility and transposable element landscapes reveal shared features of tissue-residing immune cells. *Immunity* **57**, 1975-1993.e10 (2024).

301. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

302. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PLOS ONE* **12**, e0177459 (2017).

303. Kluyver, T. *et al.* Jupyter Notebooks – a publishing format for reproducible computational workflows. in *Positioning and Power in Academic Publishing: Players, Agents and Agendas* 87–90 (IOS Press, 2016). doi:10.3233/978-1-61499-649-1-87.

304. Jin, W., Jung, J. & Kang, U. Supervised and extended restart in random walks for ranking and link prediction in networks. *PLOS ONE* **14**, e0213857 (2019).

305. Maxwell, S. & Koyutürk, M. Random walks with variable restarts for negative-example-informed label propagation. *Data Min. Knowl. Discov.* **38**, 4024–4039 (2024).

306. del Genio, C. I. Hypermodularity and community detection in hypergraphs. *Phys. Rev. Res.* **7**, 033045 (2025)

Appendices

Here are the supplementary materials of:

A. “Molecular mechanisms reconstruction from single-cell multi-omics data with HuMMuS” - [*Appendix A*]

B. “ReCoN reconstructs the molecular mechanisms coordinating multicellular programs” - [*Appendix B*]

Appendix A

Supplementary Materials of

Molecular mechanisms reconstruction from single-cell multi-omics data with HuMMuS

*Remi Trimbour[1-2], Ina Maria Deutschmann[2], Laura Cantini[1-2]**

Supplementary Text

HeterogeneoUs Multilayers for MUlti-omics Single-cell data (HuMMuS)

We developed HeterogeneoUs Multilayers for MUlti-omics Single-cell data (HuMMuS), a new tool for regulatory mechanisms inference from single-cell multi-omics data (https://github.com/cantinilab/HuMMuS).

HuMMuS is based on Heterogeneous Multilayer Networks (HMLNs). A HMLN is a network $N = (V_m, E_m, L)$, $m = 1, ..., M$, composed of M layers, each of which contains different nodes V_m and different intra-layer links $E_m \subseteq V_m \times V_m$. Nodes of different layers are connected by inter-layer links encoded in L(Kivelä et al., 2014; Baptista et al., 2022). As summarized in Figure 1, we reconstruct HMLNs composed of three layers: The TF layer, containing unlinked TFs, the scATAC layer containing peak co-accessibility information inferred from scATAC data and the scRNA layer encoding transcriptional regulation inferred from scRNA data. Details on the layers construction are provided below.

Heterogeneous Multilayer Network (HMLN) construction

The standard structure we propose for molecular mechanisms reconstruction with HuMMuS is based on scRNA-seq and scATAC-seq data that do not need to be paired.

TF layer

TFs expressed in the scRNA data and having a known motif according to JASPAR or cisBP databases(Castro-Mondragon et al., 2022; Weirauch et al., 2014) were included in the TF layer. In the presented results, we did not include TF-TF interactions in the TF layer of HuMMuS, to make a fairer comparison with state-of-the-art methods. A second version of HuMMuS, called HuMMuS + TF, is also considered to test the added value brought by TF-TF links. In this case, TFs are linked based on post-translational interactions reported in OmniPath(Türei et al., 2021).

scATAC layer

scATAC data are used in this layer to infer cis-regulatory interactions using Cicero(Pliner et al., 2018). Cicero provides co-accessibility scores between peaks within given windows of the genome. We used 500kb as genomic window size for both human and mouse data, as done in(Kamimoto et al., 2023; Pliner et al., 2018). In addition, Cicero requires defining pseudocells, by averaging groups of N cells. In the following, we used N=50, corresponding to the default Cicero value, with the only exception of the Liu dataset, where too few cells were present, thus requiring N=10. We then filtered the obtained network based on the co-accessibility scores: correlation threshold of zero for all datasets except the last dataset, composed of three omics, where 0.2 is used. The obtained network is undirected and weighted.

scRNA layer

There are many methods to infer gene networks from scRNA data. Though it would be possible to use any network connecting genes without specifically regulatory hypotheses, we here chose to use GENIE3(Huynh-Thu et al., 2010). GENIE3 is indeed one of the most popular methods to infer GRNs from RNA and scRNA data, and it was shown to have better performance than other state-of-the-art tools in(Kang et al., 2021; A et al., 2020). Being the GENIE3 network a complete one, we filtered it to keep only the 10K links with the highest weight. Of note, the network obtained by GENIE3 is here considered as an undirected and weighted network thus allowing a random walk to move from a gene to all other genes co-regulated by a common TF.

TF-peak bipartite

To associate TFs to potential binding regions, we used the function AddMotifs from the Signac package(Stuart et al., 2021) and based on motifmatchr (Schep and University, 2023). This function can, however, be replaced by the users with others, if needed. TF binding motifs were obtained from JASPAR and cisBP databases(Castro-Mondragon et al., 2022; Weirauch et al., 2014). JASPAR motifs were obtained through the JASPAR2020 R package(JASPAR2020). cisBP motifs, already reformatted and deduplicated, were accessed through the chromVARmotifs R package(chromVARmotifs, 2023). To find overlap between TF binding motifs and scATAC-seq peak coordinates, elements were mapped on the genomic sequences from BSgenome.Hsapiens.UCSC.hg38 and BSgenome.Mmusculus.UCSC.mm10 for human and mouse, respectively. The obtained network is unweighted.

Peak-genes bipartite

We finally linked peaks to genes based on the distance of the peak from the transcription starting site (TSS) of the gene. We considered 500 bp before and after the TSS. We chose a small window since we wanted to directly link a gene to only potential promoters and

leave the scATAC layer to give information on more distal regulatory regions, such as enhancers. The obtained network is unweighted.

For the computational time needed to reconstruct the Heterogeneous Multilayer Network (HMLN) with HuMMuS in a dataset of 55K cells scRNA and 9K cells scATAC, see Supp Table 7. After the reconstruction of the HMLN random walk with restart has been used for mining its information.

Random walk with restart (RWR)

Random walk with restart (RWR) is a stochastic process consisting in a succession of steps from one node (i.e. the seed) to a neighboring one through the network's edges, with a probability to start again from the seed at each step. RWR can be used to explore HMLNs and to provide a measure of nodes' closeness across the layers, ensuring the existence of a unique stationary distribution(Kivelä et al., 2014; Brin and Page, 1998). To run the RWR, we here used MultiXrank, a Python package proposing optimized RWR on universal multilayer networks(Baptista et al., 2022).

The RWR of MultiXrank makes at every step three consecutive decisions: (1) it decides whether to restart from the seed or not; (2) it then decides on which layer to go based on different predefined probabilities; (3) it finally decides on which node to move, based on intra-layer links, if we stay in the same layer, and based on inter-links, if we move to another layer. We set the probability to restart from the seed and the probability to jump from one layer to another. The restart probability was set at 0.7 for all the results presented here, being the default value in MultiXrank and also used in other RWR applications(Baptista et al., 2022; Didier et al., 2015; Zhao et al., 2015). Concerning the probability of jumping from one layer to another, we set it to be equiprobable in all layers, including the starting one. This choice is aimed at having each omic contribute equally to the results. Of note, in the HuMMuS package, when possible, we parallelized RWRs to benefit from multi-core usage.

Possible outputs of HuMMuS

The final outputs of HuMMuS are: (i) the prediction of the targets of a Transcription Factor (TF), based on RWRs starting from each TF in the TF layer and exploring the full network until the scRNA layer; (ii) the prediction of the peaks bound by a given TF, based on RWRs starting from each TF in the TF layer and exploring the scATAC layer; (iii) the prediction of the regulatory regions (proximal and distal enhancers) associated to a given gene, based on RWRs starting in each gene of the scRNA layer and exploring the scATAC layer; (iv) the reconstruction of Gene Regulatory Networks (GRNs), based on RWRs starting in each gene of the scRNA layer and exploring the full network until the TF layer; (v) the extraction of communities in the GRN, reflecting tightly connected macromolecules in the HMLN frequently involved in the regulation of the same biological process or pathway(Barabási and Oltvai, 2004).

Benchmarking settings

Datasets and preprocessing

The benchmarking was realized on four datasets: Chen, Liu, Duren, and Semrau (see Supp Table 2). The Chen and Liu datasets consisted of paired single-cell RNA sequencing (scRNA-seq) and single-cell chromatin accessibility profiling (scATAC-seq) data from human embryonic stem cells (hESCs). Duren and Semrau consisted of unpaired scRNA-seq data from mouse embryonic stem cells (mESCs). The Semrau dataset contained only scRNA-seq data; we thus used it together with the Duren's scATAC-seq data. Description of the data and download links can be found in Supp Table 2. Regarding data preprocessing, for both scRNA-seq and scATAC-seq data, we filtered out the features expressed in less than 1% of the cells. Gene counts were then log2-transformed, and peak accessibilities were binarized by replacing the non-null values with 1.

Running the state-of-the-art methods

SCENIC+(Bravo González-Blas *et al.*, 2023)

We first applied cisTopic, initializing a CistopicObject directly from the peak matrix since the fragments files were not available for the four datasets. Topic modelling was realized with the *run_cgs_models* function and all default parameters. To handle data sparsity, accessibility imputation was also done according to SCENIC+ tutorials through the *impute_accessibility* function, with scale_factor = 1e6.

Since the benchmark was realized on single-cell type datasets, we selected important regions for each topic using the Otsu method(Otsu, 1979) and by taking the 3000 top regions per topic. We then used pycisTarget with the precomputed motifs rankings and score per region, and motif annotation databases available at https://resources.aertslab.org/cistarget/databases/mus_musculus/mm10/screen/mc_v10_clust/region_based/ and https://resources.aertslab.org/cistarget/databases/homo_sapiens/hg38/screen/mc_v10_clust/region_based/ for mouse and human datasets, respectively.

The search space around the gene was defined as 150kb upstream/downstream as suggested in the SCENIC+ tutorial, from the gene coordinates of the biomart_host matching each dataset. Finally, the functions *calculate_TFs_to_genes_relationships, calculate_regions_to_genes_relationships*, *build_grn* (min_target_genes=1, rho_threshold=0) *format_egrns* were used successively with all default parameters, except those between parentheses. These two have been lowered to keep more regulations and test different thresholding in downstream evaluations.

Pando(Fleck *et al.*, 2022)

First, unpaired datasets were computationally paired with SCOTv2(Demetci et al., 2022), running *SCOTv2.align* with default parameters (k=50, e=1e-3, balanced=True, rho=5e-2, normalize=True). Following the default Pando pipeline, pseudocells were then aggregated as described in https://github.com/quadbiolab/organoid_regulomes/blob/main/pando/pseudocells.R to reduce data sparsity(Fleck et al., 2022). Motifs were obtained from JASPAR2020 and cisBP, and matched to ATAC peaks with *find_motifs*. The GRN network was finally inferred with *infer_grn* using the parameters suggested in the Pando vignette, plus upstream = 100k, downstream = 100k, and only_tss = TRUE to consider regulatory regions both downstream and upstream of the TSS, as done by the other tools here considered.

CellOracle(Kamimoto *et al.*, 2023)

We applied CellOracle as described in https://github.com/morris-lab/CellOracle. ScATAC-seq datasets were analyzed with Cicero to find co-accessible regions (co-accessibility score > 0.8) in a genomic window of 500kb. Peaks co-accessible with promoters were associated with genes through CellOracle integrate_tss_peak_with_cicero function. Peaks were also scanned with the CellOracle TFinfo function and its default parameters and default motifs to identify TF binding sites, to produce TF-gene edges. Finally, the TF-gene edges were inferred by the get_links function with alpha = 10.

GENIE3(Huynh-Thu *et al.*, 2010)

The R implementation of GENIE3 has been considered here. For both human and mouse datasets, we used the TFs having a known motif in JASPAR2020 or cisBP and expressed in the scRNA-seq data.

TF targets predictions

This first benchmark aims to test the ability of different methods to predict the targets of a Transcription Factor (TF). To do this prediction with HuMMuS, we set the TFs of interest as seeds of the RWR and explored the entire HMLN until the scRNA layer to find their target genes. The probabilities of the RWR have been set as follows: (i) for the default HuMMuS version, from the TF layer, the only option was to move to the scATAC layer (as we have no link in the TF layer). We thus set a probability of 1 in the RWR to move from the TF layer to the scATAC layer. For HuMMuS + TF, we set a probability of 1/2 to stay in the TF layer and 1/2 to move to the scATAC layer; (ii) from the scATAC layer, we could stay on the layer or move either in the TF layer, either in the scRNA layer, we thus set the RWR probability to 1/3 to make all omics have the same relevance; (iii) from the scRNA layer, we could stay on the layer or move up into the scATAC layer we thus set the RWR probability to 1/2 to make all omics have the same relevance. The probability of restart was set to 0.7, the default MultiXrank value. After RWR, we obtained, for each TF, a ranking of putative target genes. The other state-of-the-art methods (CellOracle, GENIE3, Pando) provide a GRN, also corresponding to a list of TF-gene links reflecting a ranking of putative targets per TF. We thus evaluate performances by comparing such rankings with ground-truth TF targets

from(McCalla et al., 2023) that are expressed in the scRNA data. The ground truth in(McCalla et al., 2023) is composed of TF-target gene pairs for both hESCs and mESCs obtained from the intersection of ChIP-seq data and perturbation experiments (impact of TFs KO/KD on gene expression).

For each method (HuMMuS, SCENIC+, CellOracle, GENIE3, Pando) and each TF in the ground-truth, we computed Fisher's exact tests and intersection sizes between the N top target genes and the ground-truth targets, with N varying in (3, 5, 10, 15, 20, 30, 40, 50, 75, 100). For each method, only TFs having at least 100 targets are considered. Finally, intersection performances are averaged across TFs, as TFs can vary from one method to another.

Regulatory regions identification

Predicting the peaks bound by a TF

To predict the peaks bounded by each TF with HuMMuS, we focused on the TF layer and scATAC layer. RWRs were performed from each TF to explore the scATAC layer and find the closest peaks to them according to the RWR. The RWR probabilities were thus set to 1 for going from the TF layer to the scATAC layer (same argument for this as above); ½ to stay in the scATAC layer or move to the scRNA layer, and 1 to go from the scRNA layer to the scATAC layer. The scRNA links are thus not used, and the only scope of the scRNA layer is here to connect peaks associated with the regulation of the same gene. Once obtained a ranking of peaks for each TF, since the output of HuMMuS is a scoring of peaks and not a binary classification, we thresholded the ranking to only keep the top 100%, 80%, 60% or 20% of the ranking as our predictions. We then obtained Pando's TF-peak links from the GRN post regression. TF-peaks links in SCENIC+ were obtained from the pycisTarget predictions. Regarding CellOracle instead, TF-peak links were extracted from the backbone network, since it aggregates the peaks to calculate the TF-gene links. Since the backbone network of CellOracle is weighted according to Cicero, we further considered different Cicero thresholds (0.05, 0.2, 0.8). This list includes the default threshold of 0.8, plus additional lower thresholds since very few connections were kept with the default one. To then evaluate the quality of the obtained predictions, a ground-truth was defined from ReMap2022(Hammal et al., 2022). We thus downloaded the list of the non-redundant peaks bound per TF computed in ReMap2022, using the 37 and 193 experiments available, from hESCs and mESCs respectively. Only ReMap2022 peaks overlapping with the peaks of the scATAC data were considered as part of the ground-truth. Finally, we use F1 scores and proportion of true positives to compare the peaks rankings obtained from the SCENIC+, Pando, CellOracle, and HuMMuS networks and the ground-truth peaks obtained from ReMap2022.

Predicting the regulatory regions (proximal and distal enhancers) associated with a gene

To predict the regulatory regions associated with a gene in HuMMuS, a RWR was computed starting from the gene as a seed. No scRNA link was used, leading to a probability of 1 to go directly to the scATAC layer. Once reaching the scATAC layer, if no restart, the RWR remains in the scATAC layer with probability 1. This solution allows for exploring the peaks associated with a gene based on the scATAC layer and thus potentially regulating the gene. Pando was not considered in this part of the benchmark since it does not infer peak-gene links independently from TF binding. In SCENIC+, peak-gene links were extracted after the regression model. To make the results of SCENIC+ and HuMMuS comparable, we took the same number of predicted enhancers for all the shared genes and filtered these predictions at different percentages (100%, 80%, 60%, 20%). In CellOracle, peak-gene links were extracted from the backbone networks and filtered according to correlation as suggested by the authors. As for TF-regions, we then considered different Cicero thresholds: 0.05, 0.2, and 0.8, with 0.8 being the default value.

The obtained predictions were then compared with a ground-truth based on a combination of six enhancer databases. We first defined a list of potential enhancer-gene interactions from the union of PEGASUS(Clément et al., 2020; Naville et al., 2015), ENdb(Bai et al., 2020) and EnhancerAtlas2.0(Gao and Qian, 2020). We then filtered this list, keeping only the links whose enhancers were present in the union of Fantom5(Forrest et al., 2014), VISTA(Visel et al., 2007), and SCREEN.ENCODE(Moore et al., 2020) databases. Finally, we only kept in the ground-truth enhancers overlapping with the peaks of the scATAC data. The quality of the overlap between predicted regulatory regions and the databases was finally assessed using F1 scores.

Community detection

As community detection methods well-suited for biological HMLN do not exist at the moment, we here compared community detection on the GRN output of HuMMuS vs. the GRNs obtained by the other methods. To obtain a GRN from HuMMuS we run, for each gene, a RWR starting from the gene as seed and arriving up to the TF layer to make TFs compete to regulate it. In the default HuMMuS version, we thus set the probabilities to ½ to stay in the scRNA layer or to jump from it to the scATAC layer, 1/3 to jump to any of the layers from the ATAC one, and a probability of 1 to reach the scATAC layer once reaching the TF layer. In HuMMuS+TF, we used the same RWR probabilities as above, except for the TF layer, where we have a probability of ½ to stay in the layer and ½ to move back to the scATAC layer. Once obtained a GRN also for HuMMuS, we performed community detection on the GRNs of all methods (HuMMuS, SCENIC+, Pando, CellOracle, and GENIE3). Only absolute weights were considered, all networks were filtered to the same density, and community detection was finally realized with the Louvain clustering method(Blondel et al., 2008) from the NetworkX implementation. To find the optimal clustering resolution for each of the methods, we tested 21 values between 0 to 2 with a step size of 0.1 (see Supp Table 5). Only resolutions providing at least 10 communities out of thousands of nodes (see Supp Table 4 for details on the number of nodes per method and dataset) were considered for the following part of the analysis. We considered five different databases to

evaluate the quality of the clustering: GO Cellular Component, GO Biological Process, GO Molecular Function, KEGG 2021 (human) / 2019 (mouse), and Reactome 2016 (Kanehisa and Goto, 2000; Kanehisa et al., 2023; Gillespie et al., 2022; Ashburner et al., 2000; Gene Ontology Consortium, 2021). For each method and resolution, we then used the enrichR package (Kuleshov et al., 2016) to find enriched pathways in each of their communities. We then counted the number and the proportion of communities significantly enriched (p-value < 0.05 in the results presented Figure 2.4) in at least one gene set of the database. For each method, we selected the resolution returning the best performance.

HuMMuS applied to mouse cortex profiled for scRNA, scATAC and snmC

HuMMuS application from HMLN reconstruction to GRN extraction

To illustrate the potential of HuMMuS we used a single-cell dataset of cortical neurons composed of snmC, snATAC-seq, and scRNA-seq. The data were downloaded from (Saunders et al., 2018; atac_v1_adult_brain_fresh_5k -Datasets -Single Cell ATAC -Official 10x Genomics Support; Luo et al., 2017). The snmC dataset was composed of 46,714 genes and 3386 cells; scRNA-seq was composed of 25,299 genes and 55,803 cells, and scATAC-seq was composed of 155,093 peaks and 2317 cells. For scATAC and scRNA, we used preprocessed data in the h5ad files accessible at https://scglue.readthedocs.io/en/latest/data.html under the names Saunders-2018 and 10x-Multiome-Pbmc10k, while for snmC data, we used mCH methylation averaged per gene body (gene_level_mouse.txt) available at https://brainome.ucsd.edu/annoj/brain_single_nuclei/snmcSeq_processed_data.tar.gz and retained only the features expressed in more than 3% of the cells.

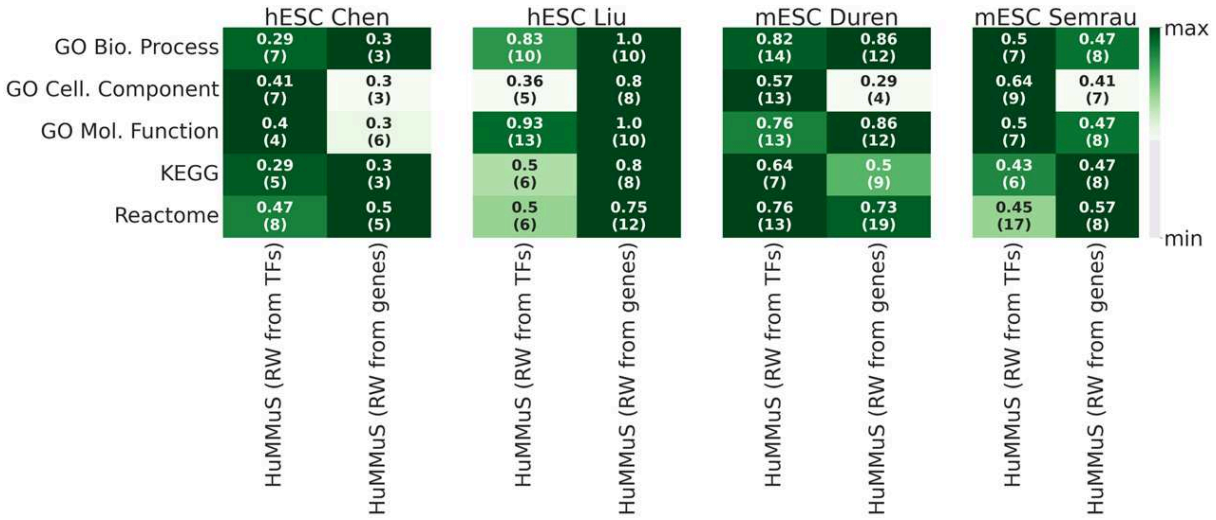
We then used HuMMuS to contract a HMLN consisting of four layers: a TF layer, a snmC layer, a scATAC layer, and a scRNA layer. To follow transcriptional regulation structure, we placed the snmC layer in the middle, connected with the scATAC layer and the scRNA layer. We did not link the snmC layer to the TF layer because TF binding motifs are specific to small regions, making gene bodies too large for precise binding motifs. As in the benchmark, we didn't put links in the TF layer. For the scATAC layer, we used Cicer,o setting a co-accessibility score threshold at 0.2, as almost all correlations were above 0. The scRNA layer was computed with the Python version of GRNBoost2; GENIE3 did not manage to get results on such a big dataset. Then the 50k links with the highest weights were kept. For the snmC layer, since we did not find methods designed to infer networks on methylation data, we used partial correlation from the pingouin0.5.3 python package, accessible at https://github.com/raphaelvallat/pingouin/tree/master. All the links with an absolute corrected correlation above 0.3 were kept. The inter-layer connections not involving the snmC layer were structured as in the benchmark. The connections between the snmC layer and the scATAC layer were set based on the distance of the scATAC peaks from the transcription start site (TSS) of the genes, nodes of the snmC layer (500 bp before and after the TSS). The connections between the snmC layer and the scRNA layer were just based on gene-gene correspondence.

After HMLN construction, using RWR from the gene layer up to the TF layer, we reconstructed a GRN. To give the same importance to each modality, the probability of going to any possible layer was the same. For the scATAC layer, we then have a probability of ¼ to go to each of the other layers or to stay in. For the scRNA layer and the snmC layer, we have a probability of ⅓ to stay in the layer, to move to the scATAC layer, or to move to the other gene-node network. Finally, from the TFs layer, it is only possible to jump to the scATAC layer.

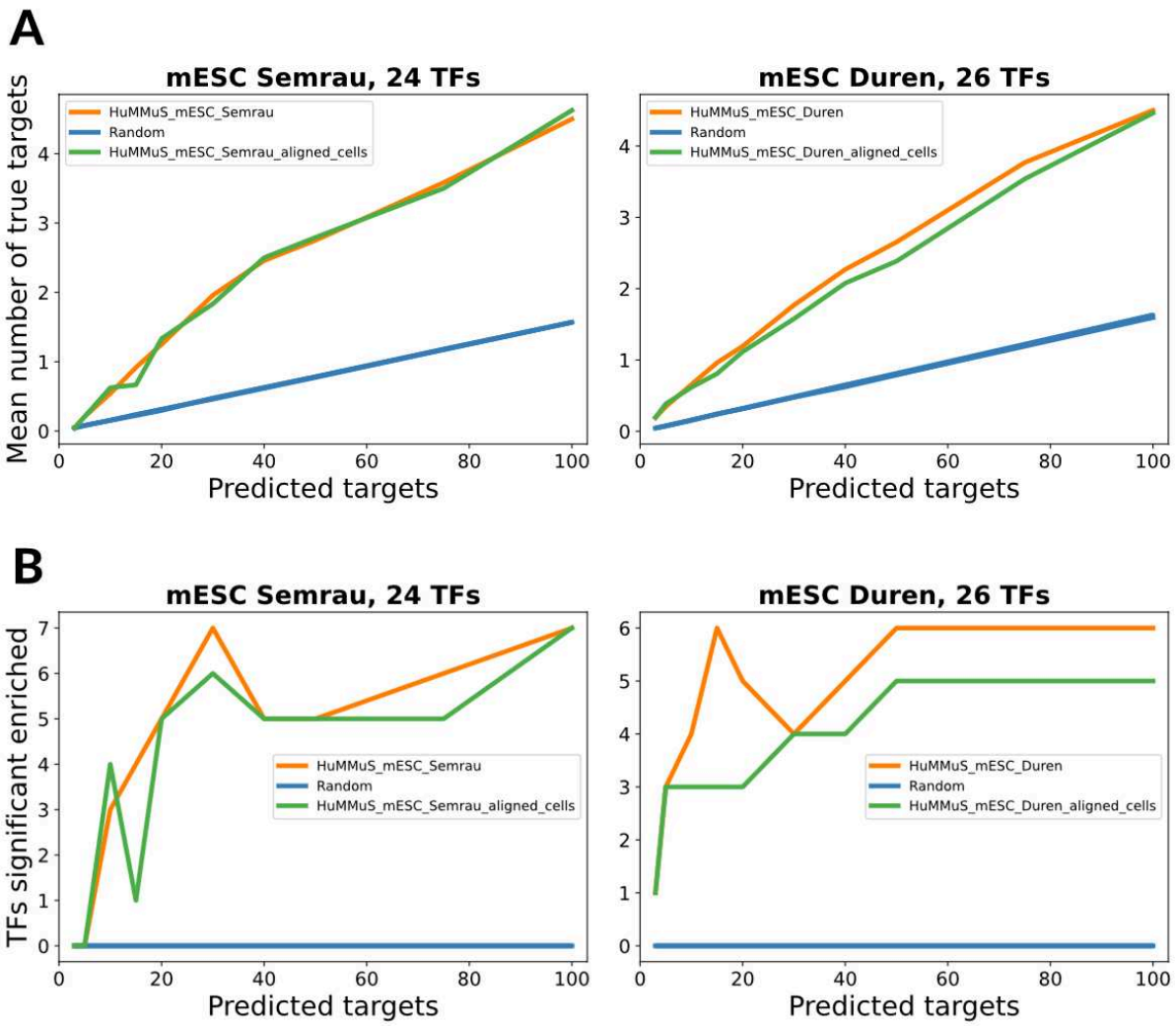
Data analysis with the obtained GRN

Starting from the GRN provided by HuMMuS, we isolated regulons, corresponding to TFs and their linked genes, and evaluated their activity in scRNA data using the unilinear model implemented in Decoupler (Badia-i-Mompel et al., 2022). UMAP was then run on such an activity matrix to test the ability of the obtained regulons to cluster cells according to their cortical neuron sub-population of origin. Finally, TF activities were used to find top marker regulons of each cortical neuron sub-population, focusing on the top 10 regulons per cortical sub-population.

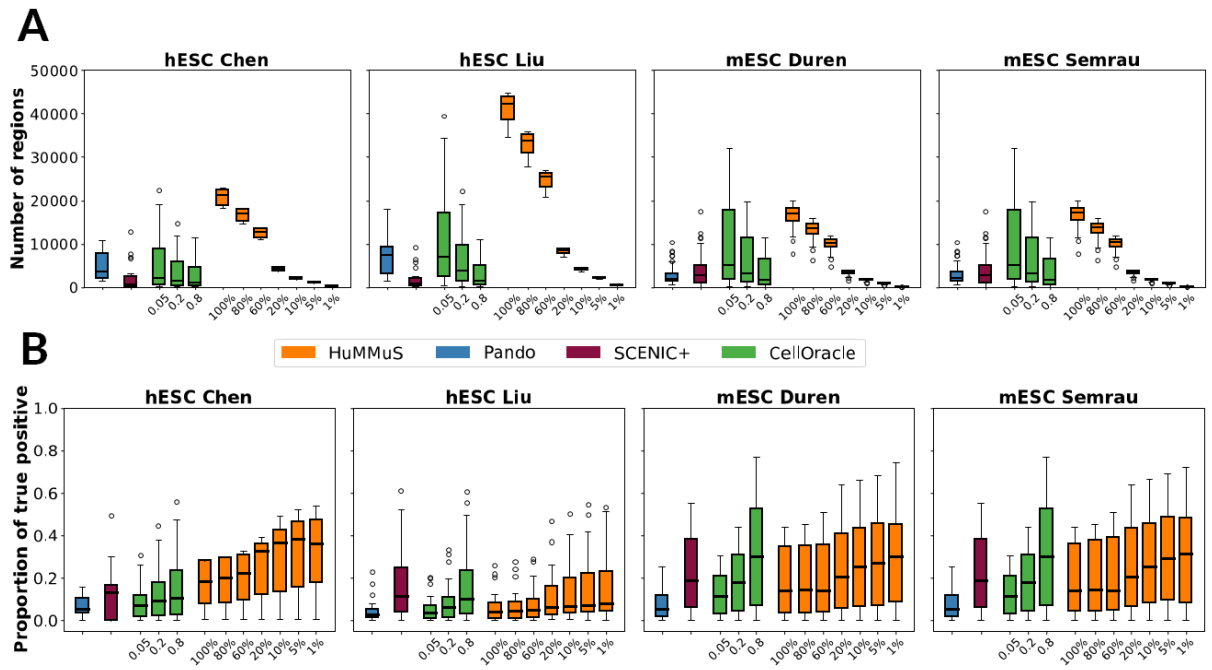
Supplementary Data



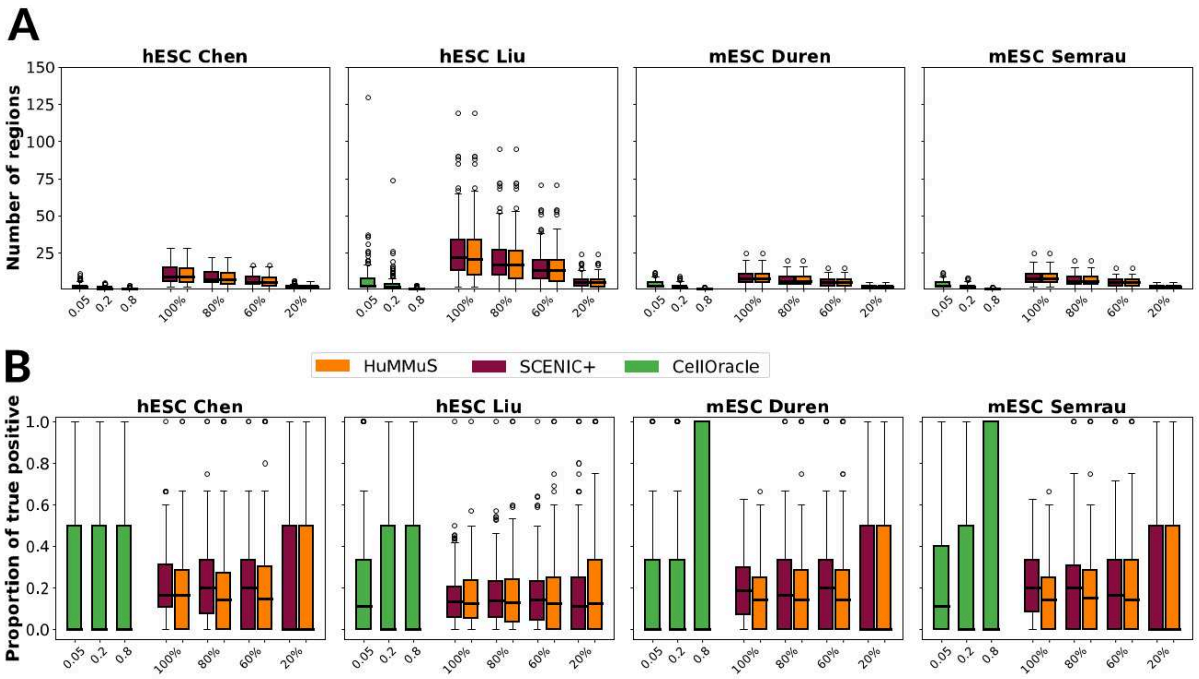
Supplementary Figure 1. Enriched communities from different RWR exploration. (A) Heatmaps of percentage of enriched community when starting random walk with restart from the TFs and from the genes across the five biological databases. The values reported in the table correspond to the percentage of enriched communities, while those in parentheses are the actual number of enriched communities.



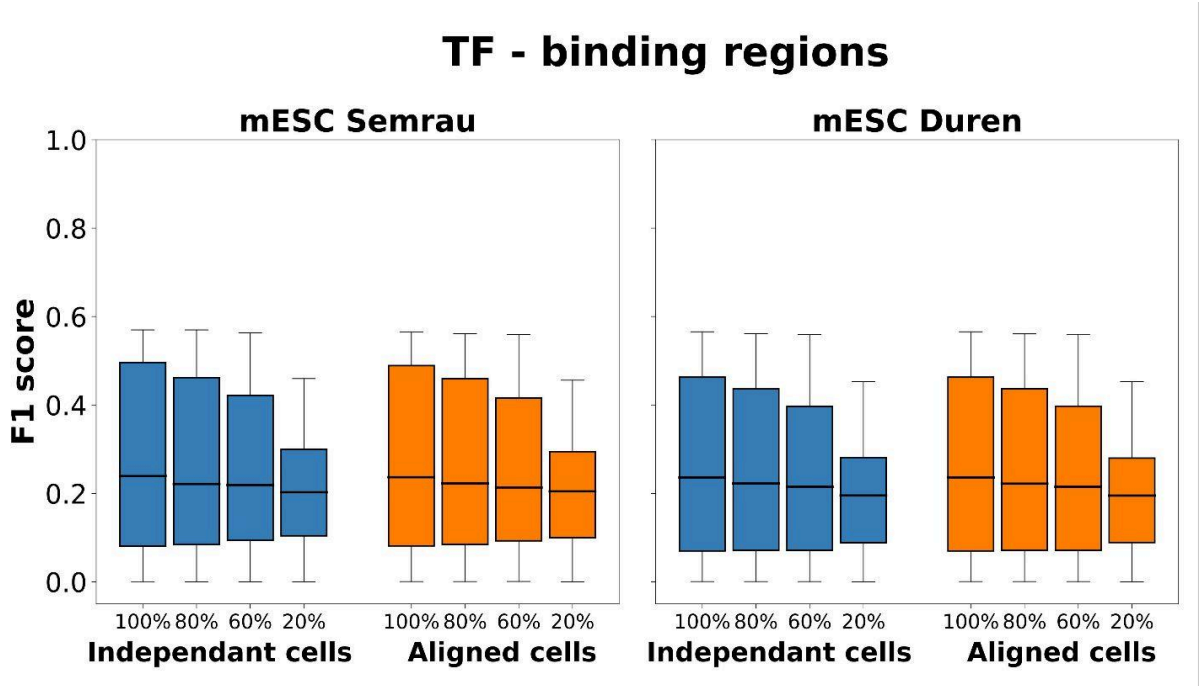
Supplementary Figure 2. Transcription Factor (TF) - target genes prediction with and without cell pairing. (A) Average number of correctly predicted targets per TF. (B) Number of TFs with a significant number of correctly predicted targets (Fisher's exact test p-val<0.05). In (A-B) Colors correspond to different methods: orange (HuMMus on unpaired scATAC+scRNA-seq data), green (HuMMuS on paired scATAC+scRNA-seq data), blue (Random).



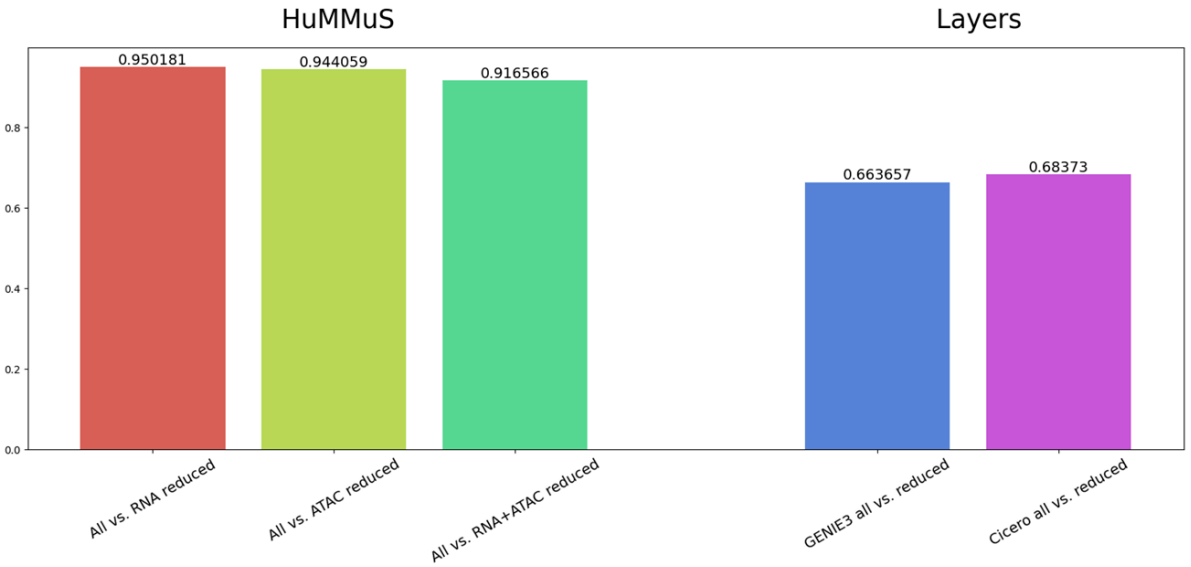
Supplementary Figure 3 - Number of regions and proportion of true positives detected per TF-region predictions' methods. (A) Distribution of the number of binding regions per TF inferred by Pando, CellOracle, SCENIC+ and HuMMuS; (B) Proportion of true positives in the predicted TF - binding regions pairs of the different methods. In (A-B), different colors correspond to different methods: HuMMuS (orange), Pando (blue), CellOracle (green), SCENIC+ (dark red).



Supplementary Figure 4 - Number of regions and proportion of true positives detected per enhancer predictions' methods. (A) Distribution of the number of regulatory regions per gene inferred by CellOracle, SCENIC+ and HuMMuS. (B) Proportion of true positives in the predicted enhancers - gene pairs of the different methods. In (A, B) colors correspond to methods: HuMMuS (orange), CellOracle (green), SCENIC+ (dark red).

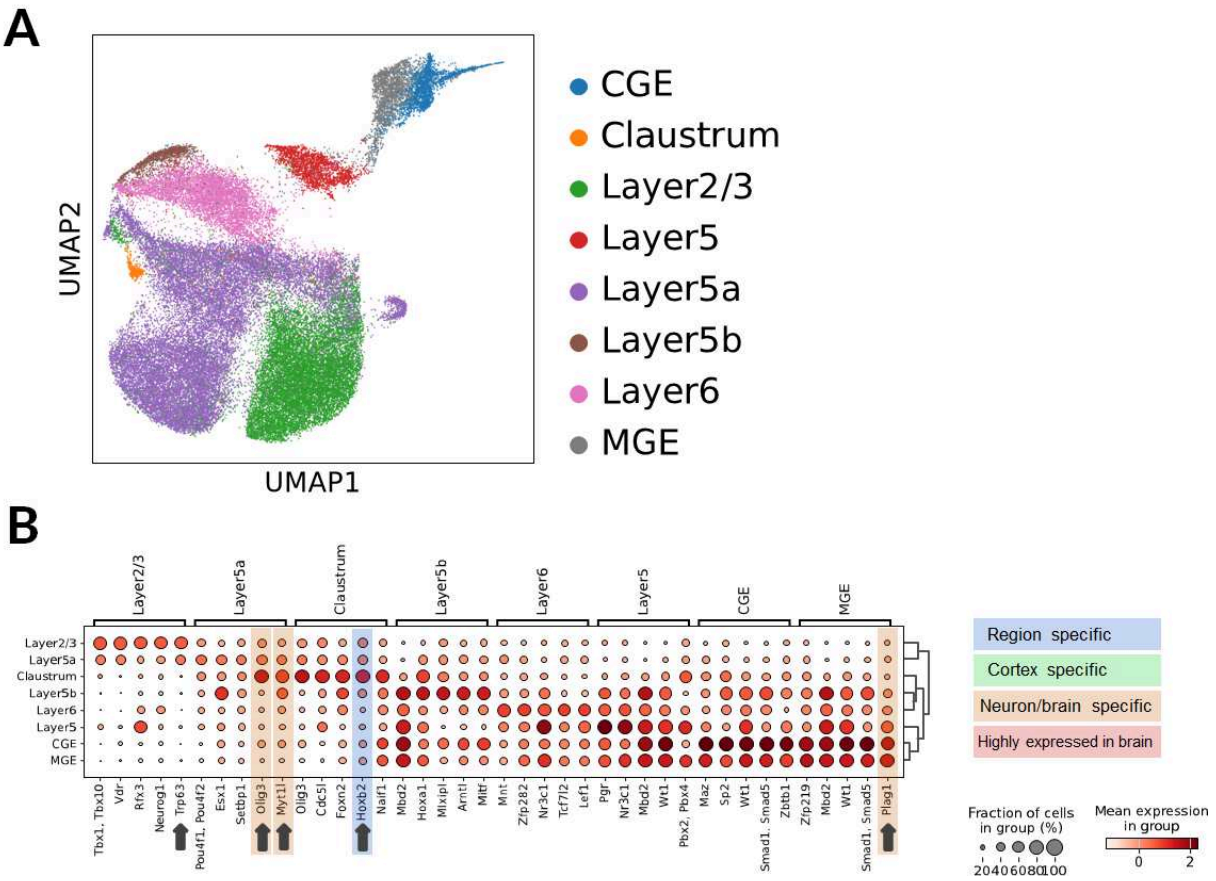


Supplementary Figure 5. Binding regions and regulatory regions prediction with and without cell pairing. F1 score distributions of the intersection between the ground-truth of TF-peak associations and those inferred by HuMMuS. Different colors correspond to different data processing: orange (HuMMuS on unpaired scATAC and scRNA-seq data), blue (HuMMuS on paired scATAC and scRNA-seq data).



Supplementary Figure 6. Spearman correlation between HuMMuS. Spearman correlation coefficient between GRNs inferred by HuMMuS on different subsets of a mouse cortex dataset (scRNA+scATAC-seq unpaired). Correlations have been computed between the complete dataset and 3 subsets of it, removing scRNA-seq cell and/or scATAC-seq measurements (barplots on the left). Correlations between the corresponding scATAC layer and between the corresponding scRNA layer obtained from these subsets have also been computed (barplots on the right). Three cortical neuron locations are present: MGE (184 cells), Layer 2/3 (614 cells), and Layer 6 (345 cells). In the reduced scRNA dataset, half of the Layer 2/3 neurons

have been discarded from the scRNA-seq data. In the reduced scATAC dataset, half of the Layer 6 neurons have been discarded from the scATAC-seq data.



Supplementary Figure 7. HuMMuS results on mouse cortex without methylation. (A) UMAP plot, in regard to Figure 5.B, showing the UMAP of the cells obtained from HuMMuS (without methylation layer) regulon activity. Cells are colored according to the label present in their original publication and in previous analyses. (B) Heatmap of activity for the top five TFs per cell population. Colors are used to denote the type of validation available; arrows indicate TFs lost once methylation is excluded from the analysis.

		TF - target gene (no TF)	TF - target gene (TF)
Chen	GRN (no TF)	**0.6287910691**	
	GRN (TF)		**0.6990991243**
Liu	GRN (no TF)	**0.7700942452**	
	GRN (TF)		**0.6732228101**
Duren	GRN (no TF)	**0.7168694993**	
	GRN (TF)		**0.6675759391**
Semrau	GRN (no TF)	**0.6189075068**	
	GRN (TF)		**0.6724702671**

Supplementary Table 1. Spearman correlation between weighted edges of TF - target gene

networks and GRNs, with/without using TF - TF interactions for each dataset. For each pair of networks, all common non-null edges were used.

A. Benchmark data: Gene expression and chromatin accessibility datasets		
Dataset name	**Associated publication**	**Data accession**
hESC_Chen	(Chen et al., 2019)	NCBI (GSE126074):
scRNA	Obtained by SNARE-seq. cell line mixture SNAREseq cDNA counts and chromatin counts, cell labels for filtering (ftp://ftp.ebi.ac.uk/ pub/ databases/mofa/snare_seq/cell_metadata.txt). Here, we only use H1 cells.	8595 features and 385 cells
scATAC		36954 features and 385 cells
hESC_Liu	(Liu et al., 2019)	https://github.com/hdsu-bioquant/scCAT/blob/master/data/HumanEmbryo/.
scRNA	Obtained by scCAT. Human embryo RNA-seq counts, ATAC-seq counts, and annotation data downloaded from github link	23153 features and 72 cells
scATAC		68952 features and 72 cells
mESC_Duren	(Duren et al., 2018; Zeng et al., 2019)	NCBI (GSE115968): scRNA-seq_RA_D4, and NCBI (GSE115970): scATAC-seq_RA_D4
scRNA	415 scATAC-seq samples generated for the retinoic acid-induced mESC differenation at day 4.	15299 features and 464 cells
scATAC	464 scRNA-seq samples generated for the retinoic acid-induced mESC differenation at day 4.	23176 features and 414 cells
mESC_Semrau	(Semrau et al., 2017)	NCBI (GSM2098553) part of (GSE79578): scrbseq_96h (scRNA-seq), but no scATAC-seq. We use the scATAC-seq from mESC_Duren, NCBI (GSE115970).
scRNA	scRNA seq data generated alone. scATAC-seq data of Semrau et al. has been used as peak layer	10243 features and 384 cells
B. Ground truth table of target genes of transcription factors		
Dataset name	**Data accession link**	**Associated publication**
mESC & hESC TFGT	https://zenodo.org/record/5909090/files/gold_standard_datasets.zip?download=1, Tables used hESC_chipunion_KDUnion_intersect.txt, and mESC_chipunion_KDUnion_intersect.txt, downloaded: 6th April 2022	(Stone et al., 2022)
C. Databases used to construct Ground truth table of enhancers for genes		
Databases name	**Data accession**	**Associated publication**
EnhancerAtlas2.0	http://www.enhanceratlas.org/indexv2.php, Table used (ESC_neuron_EP.txt, ESC_Bruce4_EP.txt, ESC_J1_EP.txt, and ESC_KH2_EP.txt): enhancer-gene interactions - Homo sapiens (hg19: ESC_neuron) and Mus musculus (mm9: ESC_Bruce4, ESC_J1, ESC_KH2), Downloaded: 4th April 2022, Preprocessing: table formatting, convert hg19 to hg38, mm9 to mm10, and ENSEMBL Gene IDs to Gene Symbol	(Gao & Qian, 2020)
ENdb	http://www.licpathway.net/ENdb/, Table used (ENdb_enhancer.txt): All the experimentally confirmed enhancers (hg19 and mm10), Downloaded: 4th April 2022, Preprocessing: convert hg19 to hg38	(Bai et al., 2020)
SCREEN (ENCODE)	http://screen.encodeproject.org, Table used (GRCh38-cCREs.bed.html and mm10-ccREs.bed.html): all human cCREs (hg38) and all mouse cCREs (mm10), Downloaded: 4th April 2022, Preprocessing: none	(The ENCODE Project Consortium et al., 2020)

VISTA	http://enhancer.lbl.gov, Table used (imagedb3.pl.html): all 3281 elements (hg19 and mm9), Downloaded: 4th April 2022, Preprocessing: delete sequences, bring into table form, convert hg19 to hg38, mm9 to mm10	(Visel et al., 2007)
PEGASUS	ftp://ftp.biologie.ens.fr/pub/dyogen/PEGASUS/, Table used (hg19_CNEs_PEGASUS.data.gz): PEGASUS predictions for the human genome (hg19), Downloaded: 4th April 2022, Preprocessing: convert hg19 to hg38, hg19 to mm10, and ENSEMBL Gene IDs to Gene Symbol	(Naville et al., 2015; Clément et al., 2020)
Fantom5	https://slidebase.binf.ku.dk/human_enhancers/presets, Table used (hg19_enhancer_promoter_correlations_distances_cell_type.txt.gz): Enhancer-Promoter Cell Type Associations in 5.Enhancer - FANTOM Robust Promoter associations, Downloaded: 12th May 2022, Preprocessing: convert hg19 to hg38, hg19 to mm10 for enhancer and also promoter regions	(Forrest et al., 2014)

D. Databases for gene enrichment analyses included in enrichR (Chen et al., 2013; Kuleshov et al., 2016; Xie et al., 2021)

Databases name	Data accession	Associated publication
Gene Ontology	GO_Biological_Processes_2021, GO_Cellular_Component_2021, and GO_Molecular_Function_2021	(Ashburner et al., 2000; Gene Ontology Consortium, 2021)
Kyoto Encyclopedia of Genes and Genomes	KEGG_2021_Human and KEGG_2019_Mouse	(Kanehisa & Goto, 2000; Kanehisa, 2019; Kanehisa et al., 2021)
Reactome	Reactome_2016 (citation refers to latest Reactome version not used within enrichR)	(Gillespie et al., 2022)

E. Dataset to test inclusion of 4th layer: Gene expression, chromatin accessibility, and HiC data

Dataset name	Data accession	Associated publication
scRNA mouse cortex	http://download.gao-lab.org/GLUE/dataset/Saunders-2018.h5ad	Saunders et al., 2018
scATAC mouse cortex	http://download.gao-lab.org/GLUE/dataset/10x-ATAC-Brain5k.h5ad	https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_adult_brain_fresh_5k
snmC mouse cortex	http://download.gao-lab.org/GLUE/dataset/Luo-2017.h5ad	Luo et al., 2017

Supplementary Table 2. Collection of publicly available data used in this study. (A) Description of the four datasets to benchmark HuMMuS in respect to state-of-the-art methods. It notably contains accessibility identifier/link and related original publication. **(B)** List of the ground truth used to benchmark target genes of transcription factors for the mESC and hESC datasets described above. **(C)** Databases of enhancers and regulatory regions combined to evaluate methods to retrieve regulatory regions important for each gene. **(D)** List of the databases used to evaluate enrichment of community detected from GRN benchmarking. (E) List and accessibility links to the data used for the 3 omics multilayer reconstruction. We used the table already preprocessed and formatted by Cao et al., 2022.

	Layer / Bipartite	3 omics mouse cortex	hESC Chen	hESC Liu	mESC Duren	mESC Semrau

TF layer (standard HuMMuS)	Number of nodes	717	220	670	607	334
TF layer (from OmniPath)	Number of nodes		432	432	364	364
	Number of edges		1403	1403	1046	1046
scATAC layer	Number of nodes	82108	25102	48207	20779	20779
	Number of edges	302651	96104	439628	72986	72986
scRNA layer	Number of nodes	9349	5095	5712	4520	5695
	Number of edges	50000	10000	10000	10000	10000
snmC layer	Number of nodes	5494				
	Number of edges	558845				
TF --> scATAC bipartite	Number of sources	717	220	670	607	334
	Number of targets	153507	25102	48207	20779	20779
	Number of edges	10685078	619902	3520936	1373435	792771
scATAC--> scRNA bipartite	Number of sources	36494	4989	3624	3036	4772
	Number of targets	16214	3780	2086	2661	4367
	Number of edges	36494	5230	3690	3096	4986
scATAC --> snmC bipartite	Number of sources	35210				
	Number of targets	17284				
	Number of edges	38299				
snmC --> scRNA bipartite	Number of sources	24504				
	Number of targets	24504				
	Number of edges	24504				

Supplementary Table 3. General description of the different multilayer components. This table contains number of nodes and edges in each component of the multilayers analysed in this article

		Number of TFs	Number of genes	Number of edges	Density
hESC_Chen	**CellOracle**	252	5925	369196	0.24731
	HuMMuS	220	5095	1120680	1
	HuMMuS + TF	426	5383	2170332	0.94661
	SCENIC+	536	8595	652237	0.14159
	Pando	220	8152	270817	0.15102
	GENIE3	220	8595	1875712	0.99208
hESC_Liu	**CellOracle**	716	9749	1390371	0.19921
	HuMMuS	670	5733	3801638	0.9899
	HuMMuS + TF	432	5675	2451600	0.98374
	SCENIC+	1387	23153	2884201	0.08977
	Pando	517	15517	232767	0.02902
	GENIE3	670	23153	12424093	0.80094
mESC_Semrau	**CellOracle**	388	7662	706606	0.23772
	HuMMuS	334	5695	1901796	1
	HuMMuS + TF	360	5695	2050200	0.97109

	SCENIC+	738	10243	1157290	0.15311
	Pando	334	9658	695894	0.21575
	GENIE3	334	10243	3300296	0.96477
mESC_Duren	**CellOracle**	664	10473	1337231	0.19231
	HuMMuS	607	4570	2741871	0.98864
	HuMMuS + TF	364	4601	1644271	0.98201
	SCENIC+	1207	15299	2065227	0.11185
	Pando	602	14045	1240683	0.14675
	GENIE3	607	15299	8777184	0.94522

Supplementary Table 4. Density comparison between the GRNs of the different methods. Table containing the number of TFs, genes, edges and the density of the GRNs defined by each method for the benchmark.

			Resolution																					
			0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
hESC_Chen	**CellOracle**	GO_Biological_Process_2021	1	1	1	1	1	1	1	1	2	3	7	10	9	12	18	19	27	33	29	35	38
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	1	2	3	10	17	14	21	16	22	25	24	28	27	36
		GO_Molecular_Function_2021	1	1	1	1	1	1	1	1	2	3	10	16	15	17	17	21	24	33	36	37	36
		KEGG_2021_Human	1	1	1	1	1	1	1	1	2	3	5	4	5	10	8	11	10	11	13	12	10
		Reactome_2016	1	1	1	1	1	1	1	1	2	3	10	14	11	14	17	20	21	23	20	29	32
	Pando	GO_Biological_Process_2021	1	2	2	2	2	2	2	2	2	2	2	3	4	3	4	4	2	2	2	3	3
		GO_Cellular_Component_2021	1	2	2	2	2	2	2	2	2	2	2	2	3	2	3	2	1	1	1	2	2
		GO_Molecular_Function_2021	1	2	2	2	2	2	2	2	2	2	2	3	4	3	4	4	2	2	2	3	3
		KEGG_2021_Human	1	2	2	2	2	2	2	2	2	2	2	1	3	2	3	3	1	1	1	2	2
		Reactome_2016	1	2	2	2	2	2	2	2	2	2	2	2	4	3	4	4	2	2	2	3	3
	GENIE3	GO_Biological_Process_2021	1	1	1	1	1	1	1	1	1	3	7	8	13	15	12	17	17	15	15	20	23
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	1	1	5	13	14	18	18	20	21	20	20	22	20	19
		GO_Molecular_Function_2021	1	1	1	1	1	1	1	1	1	6	6	10	12	18	15	23	21	21	22	25	27
		KEGG_2021_Human	1	1	1	1	1	1	1	1	1	4	4	5	7	6	13	9	10	11	13	10	12
		Reactome_2016	1	1	1	1	1	1	1	1	1	7	10	11	10	10	11	14	17	16	14	20	19
	HuMMuS + TF	GO_Biological_Process_2021	1	1	1	1	1	1	1	1	1	3	4	9	9	11	12	18	22	24	21	23	26
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	1	1	3	5	5	11	8	8	9	11	16	15	14	14
		GO_Molecular_Function_2021	1	1	1	1	1	1	1	1	1	3	7	11	10	13	13	16	16	15	20	15	22
		KEGG_2021_Human	1	1	1	1	1	1	1	1	1	2	5	6	8	7	6	9	10	10	11	8	12
		Reactome_2016	1	1	1	1	1	1	1	1	1	4	5	8	12	15	19	20	22	26	28	21	31
	HuMMuS	GO_Biological_Process_2021	1	1	1	1	1	1	1	1	1	2	4	7	10	7	15	12	18	14	27	24	27
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	1	1	3	4	5	5	7	9	14	13	18	18	20	22
		GO_Molecular_Function_2021	1	1	1	1	1	1	1	1	1	2	5	6	4	9	14	9	12	12	13	20	24
		KEGG_2021_Human	1	1	1	1	1	1	1	1	1	2	5	5	7	8	10	12	8	14	15	15	18
		Reactome_2016	1	1	1	1	1	1	1	1	1	3	5	5	5	8	13	17	12	21	19	21	30
	SCENIC+	GO_Biological_Process_2021	1	1	1	1	1	1	1	1	1	6	7	11	13	9	18	18	16	18	20	24	20
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	1	1	11	12	14	17	18	18	19	22	23	24	23	25
		GO_Molecular_Function_2021	1	1	1	1	1	1	1	1	2	9	12	13	12	12	20	22	21	22	25	22	22
		KEGG_2021_Human	1	1	1	1	1	1	1	1	1	3	5	6	8	7	7	8	6	10	13	9	11
		Reactome_2016	1	1	1	1	1	1	1	1	2	9	12	16	14	19	17	20	19	21	22	23	24
hESC_Liu	**CellOracle**	GO_Biological_Process_2021	1	1	1	1	1	1	2	2	3	5	6	11	15	11	14	22	24	20	27	27	38
		GO_Cellular_Component_2021	1	1	1	1	1	1	2	2	3	5	6	9	13	15	16	17	22	24	21	27	27
		GO_Molecular_Function_2021	1	1	1	1	1	1	2	2	3	5	6	10	14	18	23	20	27	29	34	30	33
		KEGG_2021_Human	1	1	1	1	1	1	2	2	3	4	5	8	10	11	11	11	13	12	15	14	25
		Reactome_2016	1	1	1	1	1	1	2	2	3	4	5	8	8	13	9	13	16	17	20	23	29
	Pando	GO_Biological_Process_2021	1	1	1	1	1	0	1	2	1	2	2	2	2	4	3	3	1	3	3	4	4
		GO_Cellular_Component_2021	1	1	1	1	1	2	1	1	1	1	1	1	3	3	3	3	3	3	3	2	2
		GO_Molecular_Function_2021	1	1	1	1	1	1	2	2	2	2	2	2	2	3	3	2	2	1	2	2	1
		KEGG_2021_Human	1	1	1	1	1	1	0	1	1	2	1	2	2	2	2	3	2	2	2	1	2
		Reactome_2016	1	1	1	2	1	3	1	2	2	3	2	2	3	2	5	4	3	3	4	4	4
	GENIE3	GO_Biological_Process_2021	0	0	0	1	1	1	1	1	1	1	2	2	2	3	2	2	2	3	2	3	2
		GO_Cellular_Component_2021	0	0	0	1	1	1	1	1	1	1	2	2	2	3	2	2	3	2	2	2	2
		GO_Molecular_Function_2021	0	0	0	1	1	1	1	1	1	1	2	3	2	2	1	2	2	3	2	3	3
		KEGG_2021_Human	0	0	0	1	1	1	1	1	1	2	2	3	3	3	3	3	3	3	3	3	3
		Reactome_2016	0	0	0	1	1	1	1	1	2	2	2	2	2	3	2	2	3	2	2	3	3
	HuMMuS + TF	GO_Biological_Process_2021	1	1	1	1	1	1	1	2	3	5	9	12	15	19	21	25	26	34	35	40	43
		GO_Molecular_Function_2021	1	1	1	1	1	1	1	2	3	4	5	7	10	10	11	6	7	10	12	14	20
		KEGG_2021_Human	1	1	1	1	1	1	1	2	3	5	9	12	16	18	22	25	28	35	38	41	42
		Reactome_2016	1	1	1	1	1	1	1	2	3	5	5	8	10	11	8	11	16	14	17	19	18
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	2	3	5	8	10	12	15	15	20	20	25	26	25	29
	HuMMuS	GO_Biological_Process_2021	1	1	1	1	1	1	1	1	2	5	7	11	16	21	22	24	30	41	42	50	47
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	1	2	3	5	6	10	9	9	8	13	10	19	20	19

		GO_Molecular_Function_2021	1	1	1	1	1	1	1	1	2	5	7	11	17	23	24	31	34	42	48	53	54
		KEGG_2021_Human	1	1	1	1	1	1	1	1	2	4	5	5	9	8	8	10	12	19	18	18	24
		Reactome_2016	1	1	1	1	1	1	1	1	2	3	6	8	12	12	15	17	20	24	27	33	33
	SCENIC+	GO_Biological_Process_2021	0	0	0	0	1	2	1	1	3	2	3	5	5	5	5	5	5	4	4	4	4
		GO_Cellular_Component_2021	0	0	0	0	0	2	2	1	1	2	3	4	5	4	4	6	4	6	5	7	5
		GO_Molecular_Function_2021	0	0	0	0	0	1	1	1	1	2	3	3	3	3	3	4	4	5	4	6	4
		KEGG_2021_Human	0	0	0	0	0	1	1	1	1	1	1	4	3	3	3	3	3	3	3	4	3
		Reactome_2016	0	0	0	0	0	2	1	1	2	3	3	4	4	4	4	5	4	6	5	6	6
mESC_Duren	**CellOracle**	GO_Biological_Process_2021	1	1	1	1	1	1	1	1	2	4	4	12	15	24	27	33	37	41	42	46	52
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	1	2	5	4	10	10	17	21	28	26	33	33	34	36
		GO_Molecular_Function_2021	1	1	1	1	1	1	1	1	2	4	4	10	12	25	23	22	31	35	33	40	41
		KEGG_2019_Mouse	1	1	1	1	1	1	1	1	2	5	4	8	9	16	20	17	19	21	23	25	34
		Reactome_2016	1	1	1	1	1	1	1	1	2	5	5	10	17	21	18	27	31	30	35	45	45
	Pando	GO_Biological_Process_2021	1	1	2	2	2	2	2	2	2	2	2	9	7	9	11	18	15	11	21	19	28
		GO_Cellular_Component_2021	1	1	2	2	2	2	2	2	2	2	2	7	4	6	9	12	14	13	17	12	21
		GO_Molecular_Function_2021	1	2	2	2	2	2	2	2	2	2	2	9	8	8	11	16	17	17	18	18	24
		KEGG_2019_Mouse	1	2	2	1	1	2	2	2	2	2	2	6	5	5	6	7	13	12	13	12	14
		Reactome_2016	1	2	2	2	1	2	2	2	2	2	2	9	5	6	11	9	15	13	16	14	19
	GENIE3	GO_Biological_Process_2021	1	1	1	1	1	1	1	1	2	3	2	4	8	9	15	4	13	13	11	16	16
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	1	2	2	2	3	5	10	9	9	12	12	18	21	20
		GO_Molecular_Function_2021	1	1	1	1	1	1	1	1	1	3	3	1	3	8	11	9	16	10	13	18	14
		KEGG_2019_Mouse	1	1	1	1	1	1	1	1	1	3	3	2	3	5	7	5	11	9	9	13	10
		Reactome_2016	1	1	1	1	1	1	1	1	2	3	3	2	6	5	8	10	10	11	15	20	21
	HuMMuS + TF	GO_Biological_Process_2021	1	1	1	1	1	1	1	1	2	4	8	10	12	11	14	19	15	23	23	25	29
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	1	2	4	7	9	7	9	10	15	9	16	15	13	16
		GO_Molecular_Function_2021	1	1	1	1	1	1	1	1	2	4	7	8	8	13	16	14	20	25	20	23	34
		KEGG_2019_Mouse	1	1	1	1	1	1	1	1	2	4	5	8	8	12	11	15	14	15	15	18	21
		Reactome_2016	1	1	1	1	1	1	1	1	2	4	10	10	11	13	17	22	17	20	20	23	26
	HuMMuS	GO_Biological_Process_2021	1	1	1	1	1	1	1	1	2	3	8	9	8	11	16	26	29	30	33	38	51
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	1	2	3	7	7	6	9	10	13	12	18	22	21	25
		GO_Molecular_Function_2021	1	1	1	1	1	1	1	1	2	3	6	6	9	15	12	15	22	28	30	26	34
		KEGG_2021_Human	1	1	1	1	1	1	1	1	2	3	4	7	7	9	13	13	16	22	23	19	19
		Reactome_2016	1	1	1	1	1	1	1	1	2	3	9	8	11	13	17	20	25	28	28	27	33
	SCENIC+	GO_Biological_Process_2021	1	1	1	1	1	1	1	1	2	5	4	5	5	10	7	9	13	13	10	18	19
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	1	2	3	2	4	7	9	7	9	13	12	14	16	14
		GO_Molecular_Function_2021	1	1	1	1	1	1	1	1	2	3	1	3	4	5	4	8	9	11	16	18	17
		KEGG_2021_Human	1	1	1	1	1	1	1	1	2	3	2	2	4	6	2	6	8	7	9	12	14
		Reactome_2016	1	1	1	1	1	1	1	1	2	4	5	5	6	8	7	10	17	12	14	19	21
mESC_Semrau	**CellOracle**	GO_Biological_Process_2021	1	1	1	1	1	1	1	1	2	3	7	10	21	27	36	41	41	48	55	67	73
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	1	2	3	7	11	13	11	22	24	29	33	34	38	44
		GO_Molecular_Function_2021	1	1	1	1	1	1	1	1	2	3	7	11	14	17	23	29	39	42	48	47	61
		KEGG_2019_Mouse	1	1	1	1	1	1	1	1	2	3	7	8	14	15	21	27	25	27	29	29	33
		Reactome_2016	1	1	1	1	1	1	1	1	2	3	6	11	18	20	29	38	33	46	43	45	53
	Pando	GO_Biological_Process_2021	1	1	2	1	2	2	2	2	2	2	2	12	9	15	15	12	13	19	15	11	10
		GO_Cellular_Component_2021	1	2	1	1	1	2	2	2	2	2	2	5	5	9	6	7	8	10	7	7	3
		GO_Molecular_Function_2021	1	2	3	2	2	2	2	2	2	2	2	14	16	14	14	16	15	23	17	17	17
		KEGG_2019_Mouse	1	2	2	1	2	2	2	2	2	2	2	9	8	9	9	8	8	9	7	7	4
		Reactome_2016	1	2	2	1	2	2	2	2	2	2	2	11	9	13	6	11	10	12	9	9	6
	GENIE3	GO_Biological_Process_2021	1	1	1	1	1	2	2	3	3	3	4	5	5	5	5	7	8	8	10	8	14	12
		GO_Cellular_Component_2021	1	1	1	1	1	2	2	3	3	3	4	5	5	5	5	7	10	7	10	12	10
		GO_Molecular_Function_2021	1	1	1	1	1	2	2	3	3	3	4	5	5	6	6	8	7	7	9	13	13
		KEGG_2019_Mouse	1	1	1	1	1	2	2	3	3	3	4	5	5	5	7	7	7	8	9	10	7
		Reactome_2016	1	1	1	1	1	2	2	3	3	3	4	5	5	6	5	7	8	8	8	13	10
	HuMMuS + TF	GO_Biological_Process_2021	1	1	1	1	1	1	1	1	1	3	6	10	11	13	18	23	27	30	37	33	35	40
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	1	3	5	8	9	7	8	12	12	17	18	15	18	20
		GO_Molecular_Function_2021	1	1	1	1	1	1	1	1	3	6	10	12	13	17	23	23	29	35	32	32	36
		KEGG_2019_Mouse	1	1	1	1	1	1	1	1	2	4	8	9	11	13	16	18	21	21	23	25	25
		Reactome_2016	1	1	1	1	1	1	1	1	3	6	10	12	12	15	20	23	24	24	25	27	30
	HuMMuS	GO_Biological_Process_2021	1	1	1	1	1	1	1	1	2	3	7	14	15	19	24	29	29	32	37	41	46
		GO_Cellular_Component_2021	1	1	1	1	1	1	1	1	2	3	5	9	8	11	14	11	14	15	16	17	25
		GO_Molecular_Function_2021	1	1	1	1	1	1	1	1	2	3	8	14	16	17	24	27	32	31	34	40	35
		KEGG_2021_Human	1	1	1	1	1	1	1	1	2	3	5	10	11	14	16	18	22	22	24	27	24
		Reactome_2016	1	1	1	1	1	1	1	1	2	3	8	12	14	19	24	24	26	31	31	34	33
	SCENIC+	GO_Biological_Process_2021	1	1	1	1	1	1	2	2	3	2	5	7	5	7	9	11	13	17	21	24	28
		GO_Cellular_Component_2021	1	1	1	1	1	1	2	2	3	3	5	5	5	5	8	13	15	15	20	22	17
		GO_Molecular_Function_2021	1	1	1	1	1	1	2	2	3	1	5	6	6	5	10	13	19	20	19	22	36
		KEGG_2021_Human	1	1	1	1	1	1	2	2	3	2	5	5	7	9	8	7	10	14	14	20	13
		Reactome_2016	1	1	1	1	1	1	2	2	2	2	5	5	6	5	10	10	10	18	12	19	22

Supplementary Table 5. Number of significantly enriched communities (p value < 0.05) detected for each method, resolution, and database used.

Gene name	Area	Wilcoxon test score / p-val adj	Associated publication / source

Tbx1 Tbx10	Layer 2/3	104.65 / 0	loss of Tbx1 disrupts corticogenesis in mice by promoting premature neuronal differentiation https://pubmed.ncbi.nlm.nih.gov/27005988/
Rfx3	Layer 2/3	102.37 / 0	*We [...] identified [...] TFs with more restricted patterns in specific subclasses, such as Rfx3 [...] (in L2/3 IT)* https://pubmed.ncbi.nlm.nih.gov/34616075/
Vdr	Layer 2/3	102.2 / 0	Expressed in cortical neurons and involved in neurodegeneration, https://pubmed.ncbi.nlm.nih.gov/21408608/
Neurog1	Layer 2/3	98.26 / 0	layer II/III neurons of the piriform cortex. https://pubmed.ncbi.nlm.nih.gov/24403153/
Zfp711	Layer 2/3	96.2 / 0	not studied a lot, involved in brain development https://pubmed.ncbi.nlm.nih.gov/20346720/
Pou4f1 Pou4f2	Layer5a	81.53 / 0	Expressed in [...] the dorsal column of the mesencephalic and pontine central gray, and the lateral interpeduncular nucleus of the brain https://pubmed.ncbi.nlm.nih.gov/7904822/
Esx1	Layer5a	76.13 / 0	expressed highly in midbrain mantle layer (FDR: 4E-4) https://bgee.org/gene/ENSMUSG00000023443?expression=&data_type=IN_SITU
Sebox	Layer5a	71.97 / 0	expressed in cerebral cortex https://www.ncbi.nlm.nih.gov/pmc/articles/PMC16794/
Setbp1	Layer5a	70.6 / 0	expressed in ventricular zone https://molecularautism.biomedcentral.com/articles/10.1186/s13229-023-00540-x
Pou4f3	Layer5a	70.2 / 0	Expressed in brain and DRG https://pubmed.ncbi.nlm.nih.gov/22326227/
Pgr	Layer5	79.42 / 0	expressed in substantia niagra https://bgee.org/gene/ENSMUSG00000031870
Nr3C1	Layer5, Layer6	76.48, 75.96 / 0, 0	expressed in median eminence of neurohypophysis https://bgee.org/gene/ENSMUSG00000024431
Mbd2	Layer5, MGE, Layer5b	67.94, 60.58, 54.90 / 0, 0, 0	highly expressed in brain https://pubmed.ncbi.nlm.nih.gov/9774669/
Wt1	Layer5, CGE, MGE	63.45, 71.84, 60.58 / 0, 0, 0	neurons of DRG and sertolis cells https://pubmed.ncbi.nlm.nih.gov/16467207/
Pbx2 /Pbx4	Layer5	54.88 / 0	regulates patterning of the cerebral cortex in progenitors and post mitotic neurons https://pubmed.ncbi.nlm.nih.gov/26671461/
Olig3	Claustrum	29.18 / 3.54E187	Olig3 coordinates the specification of dorsal neurons in the spinal cord https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1065726/
Foxn2	Claustrum	27.27 / 9.80E-164	
Naif1	Claustrum	27.09 / 1.23E161	
Dmrtc2	Claustrum	26.7 / 4.58E157	*DMRT2, DMRTA1/DMRT4, DMRT3 and DMRTA2/DMRT5 are expressed mainly in cortical regions* https://www.frontiersin.org/articles/10.3389/fnana.2022.937596/full
Cdc5l	Claustrum	26.61 / 5.31E156	
Maz	**CGE**	**73.36 / 0**	**(1) Expressed in Purkinje cells in the brain (at protein level). / (2) driving neurogenesis** (1)https://pubmed.ncbi.nlm.nih.gov/26089202/(2)https://pubmed.ncbi.nlm.nih.gov/22944911
Sp2	CGE	72.6 / 0	
Smad1/Smad5	CGE, MGE	71.92, 60.57 / 0, 0	ventricular zone, FDR: 10E-10 https://www.uniprot.org/uniprotkb/P97454/entry#expression
Zbtb1	CGE	70.19 / 0	
Zfp219	MGE	60.28 / 0	
Klf15	MGE	60.05 / 0	
Mlxipl	Layer5b	48.23 / 0	Expressed in the ventricular and intermediate zones of the developing spinal cord of 12.5 dpc embryos. In later embryos expressed in a variety of tissues. https://www.uniprot.org/uniprotkb/Q99MZ3/entry#expression
Hoxa1	Layer5b	48.19 / 0	Motor neuron axon guidance in development https://pubmed.ncbi.nlm.nih.gov/9367425/
Arntl	Layer5b	46.21 / 0	constitutively expressed in hypothalamus nucleus suprachiasmatic https://pubmed.ncbi.nlm.nih.gov/11207387/
Mitf	Layer5b	46 / 0	Microphthalmia-associated transcription factor ensures the elongation of axons and dendrites in the mouse frontal cortex. https://pubmed.ncbi.nlm.nih.gov/27859996/
Mnt	Layer6	80.83 / 0	Motor neuron expression https://bgee.org/gene/ENSMUSG00000000282
Zfp282	Layer6	78.68 / 0	
Lef1	Layer6	75.96 / 0	deep layers of the cortex, important for normal development https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3825142/
Tcf7l2	Layer6	74.89 / 0	deep layers of the cortex, important for normal development https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3825142/

Supplementary Table 6. Marker TFs/regulons identified by HuMMuS on the cortical mouse

3-omics dataset. This table contains the regions and the Wilcoxon test's statistics associated to the identified marker TFs, and the literature supporting their levels of evidence. Different level of literature-based evidence are indicated by colors : green (marker of the specific subpopulation supported by literature), blue (marker of the cortex or similar area supported by literature), yellow (other brain regions / neuron markers supported by literature), orange (expressed in the brain/cortex according to gene expression databases), gray (no evidence)

		Mouse Cortex Dataset	Chen Dataset
	Size scRNA seq data	25299 genes; 55803 cells	8595 genes; 385 cells
	Size scATAC seq data	155093 peaks; 2317 cells	36954 peaks; 385 cells
Multilayer Creation	Real time	2 h 30 min	41 min 9 sec
	Memory used	50 Gb	26 Gb
	Number of CPUs	90	90
	CPU time	23 h 5 min	3 h 40 min
TF - target genes/regions	Real time per TF (RWR process)	1 min 35 sec	7.51 sec
	Real time all TFs	28 min 25 sec	3 min 58 sec
	Memory used	270 Gb	32 Gb
	Number of CPUs	70	70
	CPU time	24 h 40 min	1 h 47 min
Gene regulatory network	Real time per TF (RWR process)	1 min 40 sec	8.27 sec
	Real time all TFs	9 h 20 min	10 min 02 sec
	Memory used	270 Gb	32 Gb
	Number of CPUs	70	70
	CPU time	548 h 10 min	10 h 44 min

Supplementary Table 7. Resources used to run HuMMuS workflow on different datasets.

References

A,P. et al. (2020) Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nature methods, 17.

Ashburner,M. et al. (2000) Gene Ontology: tool for the unification of biology. Nat Genet, 25, 25–29.

atac_v1_adult_brain_fresh_5k -Datasets -Single Cell ATAC -Official 10x Genomics Support.

Badia-i-Mompel,P. et al. (2022) decoupleR: ensemble of computational methods to infer biological activities from omics data. Bioinformatics Advances, 2, vbac016.

Bai,X. et al. (2020) ENdb: a manually curated database of experimentally supported enhancers for human and mouse. Nucleic Acids Res, 48, D51–D57.

Baptista,A. et al. (2022) Universal multilayer network exploration by random walk with restart. Commun Phys, 5, 1–9.

Barabási,A.-L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet, 5, 101–113.

Blondel,V.D. et al. (2008) Fast unfolding of communities in large networks. J. Stat. Mech., 2008, P10008.

Bravo González-Blas,C. et al. (2023) SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. Nat Methods, 20, 1355–1367.

Brin,S. and Page,L. (1998) The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30, 107–117.

Cao,Z.-J. and Gao,G. (2022) Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nat Biotechnol, 40, 1458–1466.

Castro-Mondragon,J.A. et al. (2022) JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Research, 50, D165–D173.

chromVARmotifs (2023).

Clément,Y. et al. (2020) Enhancer–gene maps in the human and zebrafish genomes using evolutionary linkage conservation. Nucleic Acids Res, 48, 2357–2371.

Demetci,P. et al. (2022) SCOTv2: Single-Cell Multiomic Alignment with Disproportionate Cell-Type Representation. Journal of Computational Biology, 29, 1213–1228.

Didier,G. et al. (2015) Identifying communities from multiplex biological networks. PeerJ, 3, e1525.

Fleck,J.S. et al. (2022) Inferring and perturbing cell fate regulomes in human brain organoids. Nature, 1–8.

Forrest,A.R.R. et al. (2014) A promoter-level mammalian expression atlas. Nature, 507, 462–470.

Gao,T. and Qian,J. (2020) EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Research, 48, D58–D64.

Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res, 49, D325–D334.

Gillespie,M. et al. (2022) The reactome pathway knowledgebase 2022. Nucleic Acids Res, 50, D687–D692.

Hammal,F. et al. (2022) ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. Nucleic Acids Res, 50, D316–D325.

Huynh-Thu,V.A. et al. (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLOS ONE, 5, e12776.

JASPAR2020 Bioconductor.

Kamimoto,K. et al. (2023) Dissecting cell identity via network inference and in silico gene perturbation. Nature, 614, 742–751.

Kanehisa,M. et al. (2023) KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res, 51, D587–D592.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 28, 27–30.

Kang,Y. et al. (2021) Evaluating the Reproducibility of Single-Cell Gene Regulatory Network Inference Algorithms. Front Genet, 12, 617282.

Kivelä,M. et al. (2014) Multilayer networks. Journal of Complex Networks, 2, 203–271.

Kuleshov,M.V. et al. (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res, 44, W90–W97.

Luo,C. et al. (2017) Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science, 357, 600–604.

McCalla,S.G. et al. (2023) Identifying strengths and weaknesses of methods for computational network inference from single-cell RNA-seq data. G3 (Bethesda), 13, jkad004.

Moore,J.E. et al. (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature, 583, 699–710.

Naville,M. et al. (2015) Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome. Nat Commun, 6, 6904.

Otsu,N. (1979) A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics, 9, 62–66.

Pliner,H.A. et al. (2018) Cicero predicts cis-regulatory DNA interactions from single cell chromatin accessibility data. Mol Cell, 71, 858-871.e8.

Saunders,A. et al. (2018) Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. Cell, 174, 1015-1030.e16.

Schep,A. and University,S. (2023) motifmatchr: Fast Motif Matching in R.

Stuart,T. et al. (2021) Single-cell chromatin state analysis with Signac. Nat Methods, 18, 1333–1341.

Türei,D. et al. (2021) Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Molecular Systems Biology, 17, e9923.

Visel,A. et al. (2007) VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res, 35, D88-92.

Weirauch,M.T. et al. (2014) Determination and inference of eukaryotic transcription factor sequence specificity. Cell, 158, 1431–1443.

Zhao,Z.-Q. et al. (2015) Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization. Computational Biology and Chemistry, 57, 21–28.

Appendix B

Supplementary Materials of

ReCoN reconstructs the molecular mechanisms coordinating multicellular programs.

Rémi Trimbour[1], Ricardo O. Ramirez Flores[2,3], Julio Saez-Rodriguez[2,3,#], Laura Cantini[1,#]

[1] Institut Pasteur, Université Paris Cité, CNRS UMR 3738, Machine Learning for Integrative Genomics Group, F-75015 Paris, France

2 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), Hinxton, Cambridgeshire, U.K.

# These two authors jointly supervised this work

Supplementary Text

Supplementary Notes 1. Cell type multilayer reconstruction.

Cell type multilayers are individual heterogeneous multilayers, typically each layer representing different types of macromolecules. In the presented results, cell type multilayer was built from a GRN layer, which contains genes and TFs, and a receptor layer. However, this framework is easily extendable. First, it is possible to add layers for new molecules. The only requirement is to provide at least one bipartite connecting the nodes of this new layer to the other layers. In the case of molecules with roles limited to intracellular mechanisms, the layer would be cell-specific and only connected to the layer of the cell type they belong to. In the case where the layer provides intercellular bridges, it could also connect layers of different cell types. Second, it is possible to integrate layers between the macromolecules already represented in the default structure. For example, receptors could be linked both by ontology and by PPI relationships. In this case, the two layers could be considered as a multiplex. The RWR process can move freely and at no cost between the representations of the same node at each step, as defined by transition probabilities specific to the multilayer. In other words, when exploring this layer, a fourth decision is added to the RWR process, where which layer (i.e. type of interactions) that will be used for the next step is decided.

Supplementary Notes 2 - GRN reconstruction with HuMMuS and hummuspy.

HuMMuS is initially an R package, which relies on Python code for the RWR process, through the MultixRank package(Baptista *et al.*, 2022). However, the computation of the individual layers can be computationally extensive in the default setting, since it relies on GENIE3 and Cicero. To work on bigger datasets, these layers can be computed externally (e.g., with Python packages), before computing the bipartites with the R version of HuMMuS. Once all the layers and bipartites are defined, a GRN can be obtained with either the R package or the Python code directly – *hummuspy*. The two versions provide identical results since they run the exact same Python code in the background. The R version was used for the Heart model, and the Python version was used for the Immune Dictionary. Since we started the development of ReCoN on the Immune Dictionary dataset, the Python version offered us easier integration with downstream analysis to define the default parameters of ReCoN.

Supplementary Notes 3. Combination of keywords used to extract gene sets related to heart failure in MSigDB.

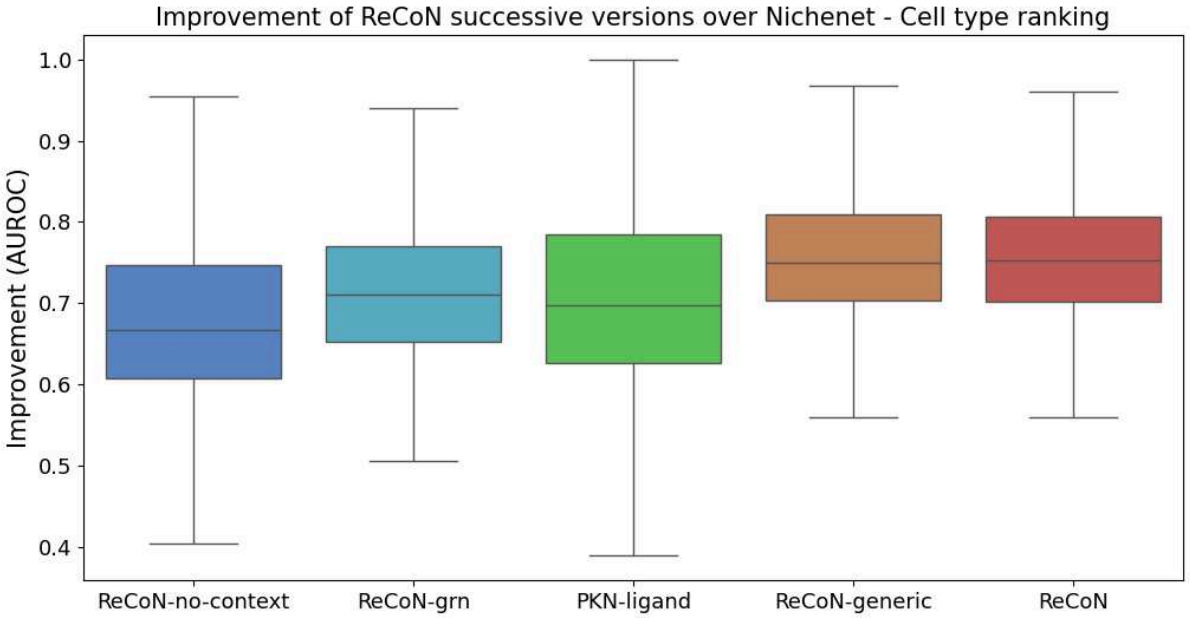
The predictions of multicellular co-operation in heart failure and cardiac fibrosis were evaluated with gene set enrichments. We chose three categories of interest and extracted the related gene sets with the same keyword as in ReHeat2(Lanzer *et al.*, 2024), which provide insights into multicellular co-operation in heart failure.

References

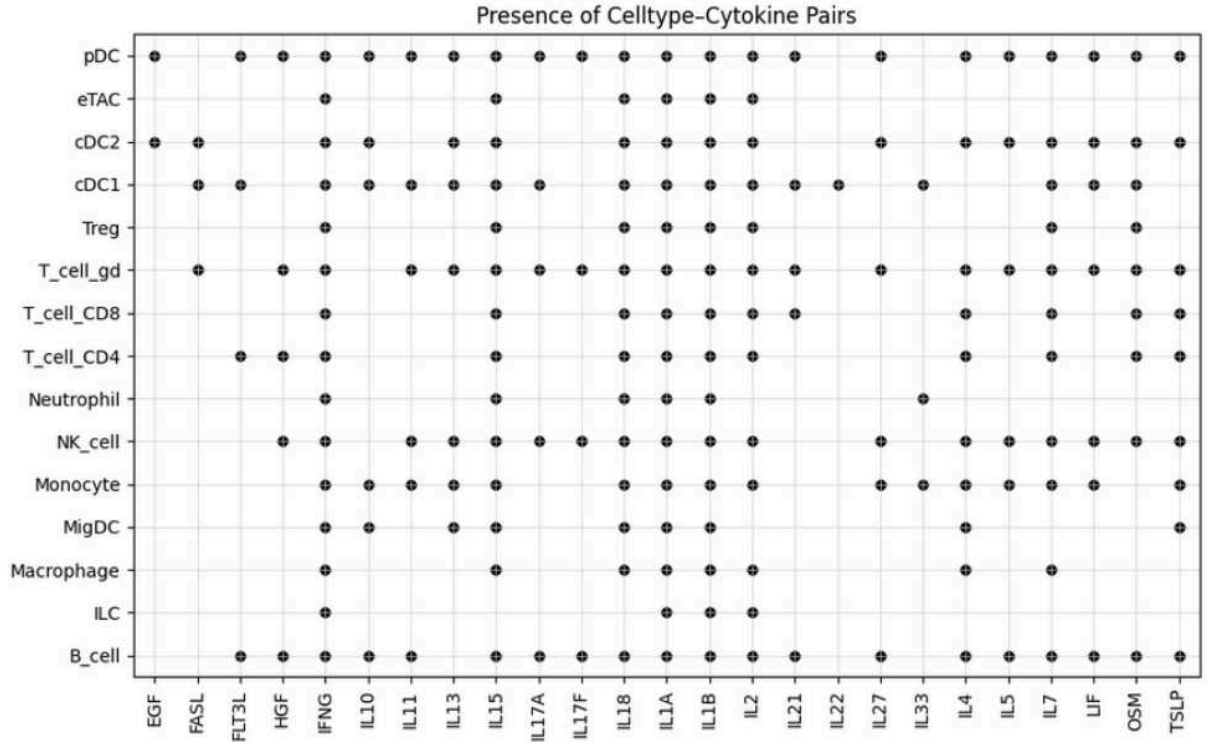
Baptista,A. et al. (2022) Universal multilayer network exploration by random walk with restart. Commun. Phys., 5, 1–9.

Lanzer,J.D. et al. (2024) A cross-study transcriptional patient map of heart failure defines conserved multicellular coordination in cardiac remodeling. 2024.11.04.621815.

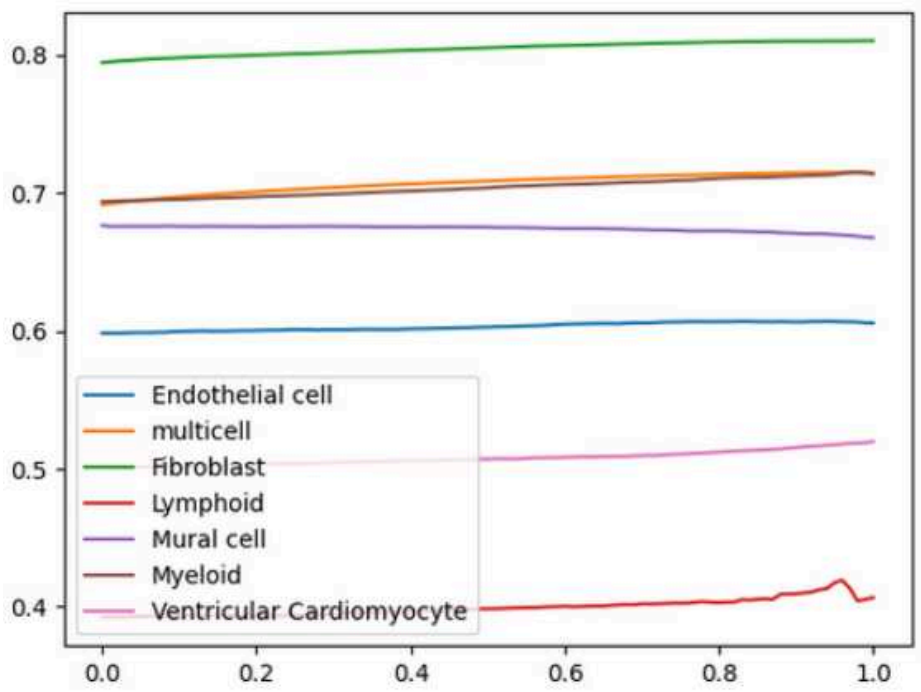
Supplementary Data



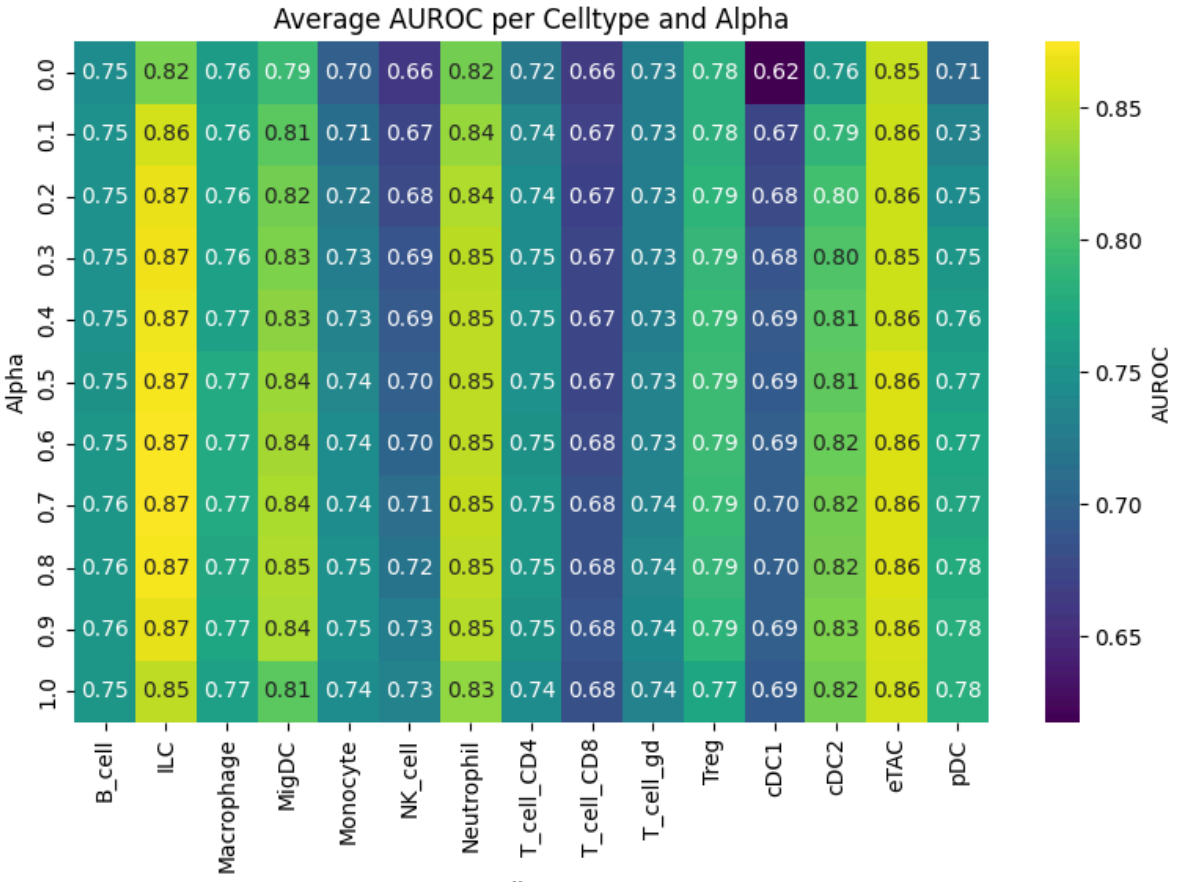
Supplementary Figure 1. ReCoN predictions of transcriptomic responses across each cytokine and cell type pair – cell type ranking. Four successively enriched versions of ReCoN (blue and green) and the PKN model from Nichenet are plotted. Boxplots represent the AUROCS of each model, where each cell type-cytokine pair considered is a value in the boxplot. Outliers are not plotted.



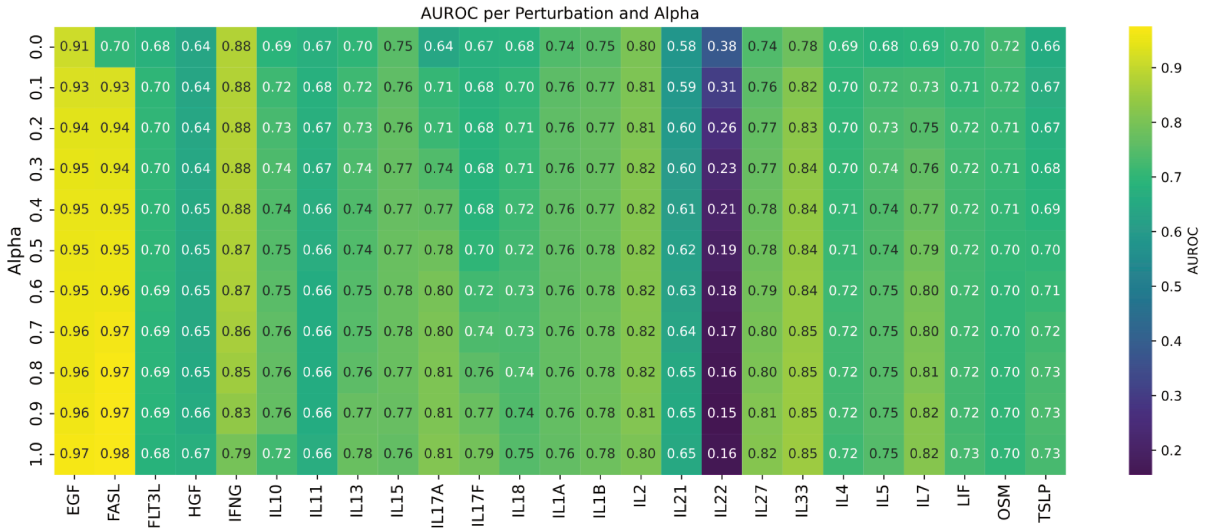
Supplementary Figure 2. Pair of cytokines and cell types used in the study. Only pairs with at least two significantly perturbed genes (see Methods) were considered here and for downstream analysis.



Supplementary Figure 3. AUPRC as a function of α (the weight of indirect effects) for individual cell types and global multicellular rankings in the HF showcase.



Supplementary Table 1. AUROC across every cytokine and different α values – [immune dictionary showcase].



Supplementary Table 2. AUROC across every cytokine and different α values – [immune dictionary showcase].

gene	score	fibrosis	heart	doi
MYPOP	0.001065	Related	Related	https://doi.org/10.1242/jcs.01618, https://doi.org/10.3892/mmr.2017.7446
KDM2B	0.001049	YES	YES	https://doi.org/10.1111/jcmm.14146
E2F1	0.000942	YES	YES	https://doi.org/10.1161/circulationaha.120.047626, https://doi.org/10.1080/21655979.2021.1972194
PROX1	0.000921	NO	YES	https://doi.org/10.1038/s41586-020-2998-x
ZBTB7B	0.000915	YES	NO	https://doi.org/10.1016/s0945-053x(01)00167-6
ZBTB14	0.000904	YES	NO	https://doi.org/10.31083/j.fbl2809205
HINFP	0.000895	NO	NO	—
GLIS1	0.000834	YES	YES	https://doi.org/10.1038/s41421-022-00490-3, https://doi.org/10.1016/j.freeradbiomed.2023.09.037
E2F4	0.000829	YES	NO	https://doi.org/10.1096/fj.201903021rr
TCFL5	0.000819	NO	NO	—

Supplementary Table 3. Annotation of the 10 top TFs predicted by ReCoN – [heart failure showcase]. Receptors are classified as related to fibrosis and to the heart if a publication can justify this link. In violet are the receptors classified as related to both fibrosis and heart condition, in orange are the receptors related to one of the categories, and in red are the receptors related to neither of the categories.

gene	score	fibrosis	heart	doi
DRD4	0.005327	YES	NO	https://doi.org/10.1111/acer.12047
IL27RA	0.004325	YES	YES	https://doi.org/10.1016/j.heliyon.2023.e17099
IFNGR2	0.003781	YES	YES	https://doi.org/10.1007/s10741-013-9393-8
PTGDR2	0.00351	YES	YES	https://doi.org/10.1371/journal.ppat.1011812, https://doi.org/10.1016/j.yjmcc.2022.03.011

IL21R	0.00344	YES	YES	https://doi.org/10.1016/j.ejphar.2022.175482
IL13RA1	0.003424	YES	YES	https://doi.org/10.1161/JAHA.116.005108
GPR182	0.003365	YES	YES	https://doi.org/10.1007/s10456-025-09977-5, https://doi.org/10.1161/jaha.117.007253
TRPV2	0.00332	YES	YES	https://doi.org/10.1038/s41374-019-0349-z
IL7R	0.003092	YES	NO	https://doi.org/10.1172/JCI14685, https://doi.org/10.1186/s12931-022-02077-8
IL17RE	0.003079	YES	Related	https://doi.org/10.3389/fcvm.2024.1470362
IL18RAP	0.002499	YES	NO	https://doi.org/10.1002/hep.32776
ADGRG1	0.002461	YES	YES	https://doi.org/10.1042/bsr20240826, https://doi.org/10.1177/1535370214529395
IL5RA	0.002353	YES	Related	https://doi.org/10.1111/jcmm.18493
MST1R	0.002332	YES	NO	https://doi.org/10.1111/liv.14892
IL15RA	0.002313	YES	NO	https://doi.org/10.3389/fimmu.2024.1404891
CRLF2	0.00224	YES	YES	https://doi.org/10.1096/fj.202302000rr
PDGFRA	0.00221	YES	Related	https://doi.org/10.1016/j.yexcr.2016.10.022
CD180	0.002201	YES	YES	https://doi.org/10.1007/s00441-021-03488-7, https://doi.org/10.3892/mmr.2020.11242
NPTXR	0.002105	NO	NO	–
TGFBR3	0.002091	YES	YES	https://doi.org/10.1111/bph.13166
IL11RA	0.001949	YES	YES	https://doi.org/10.1093/cvr/cvae224
MPL	0.001947	NO	NO	–
MUC5AC	0.001924	YES	NO	https://doi.org/10.1371/journal.pone.0058658, https://doi.org/10.1038/mi.2012.114
IL12RB1	0.001906	YES	YES	https://doi.org/10.3389/fphar.2020.00129
GPR25	0.001896	YES	NO	https://doi.org/10.1111/febs.70117, https://doi.org/10.1016/j.jdermsci.2020.09.010

Supplementary Table 4. Annotation of the 25 top receptors predicted by ReCoN – [heart failure showcase]. Receptors are classified as related to fibrosis and to the heart if a publication can justify this link. In violet are the receptors classified as related to both fibrosis and heart condition, in orange are the receptors related to one of the categories, and in red are the receptors related to neither of the categories.

gene	score	fibrosis	heart	doi
IFNGR2	1.849102	YES	YES	https://doi.org/10.1007/s10741-013-9393-8
DRD4	1.749613	YES	NO	https://doi.org/10.1111/acer.12047
IL13RA1	1.568187	YES	YES	https://doi.org/10.1161/JAHA.116.005108
IL27RA	1.530149	YES	YES	https://doi.org/10.1016/j.heliyon.2023.e17099
IL21R	1.240536	YES	YES	https://doi.org/10.1016/j.ejphar.2022.175482
IL15RA	1.070769	YES	NO	https://doi.org/10.3389/fimmu.2024.1404891
IL5RA	1.057473	YES	Related	https://doi.org/10.1111/jcmm.18493
APCDD1	1.053718	NO	NO	–
TRPV2	1.031249	YES	YES	https://doi.org/10.1038/s41374-019-0349-z
PTGDR2	0.991871	YES	YES	https://doi.org/10.1016/j.yjmcc.2022.03.011

IL18RAP	0.988052	YES	NO	https://doi.org/10.1002/hep.32776
IL7R	0.982419	YES	NO	https://doi.org/10.1172/JCI14685, https://doi.org/10.1186/s12931-022-02077-8
CD40LG	0.893819	YES	YES	https://doi.org/10.1016/j.ijcard.2018.12.076
PDGFRA	0.852159	YES	Related	https://doi.org/10.1016/j.yexcr.2016.10.022
MST1R	0.832997	YES	NO	https://doi.org/10.1111/liv.14892
CRLF2	0.809771	YES	YES	https://doi.org/10.1096/fj.202302000rr
TGFBR3	0.767271	YES	YES	https://doi.org/10.1111/bph.13166
IL1RAP	0.755057	YES	YES	https://doi.org/10.1161/circheartfailure.124.011729
IL17RB	0.744823	YES	NO	https://doi.org/10.1172/JCI14685, https://doi.org/10.1186/s12931-022-02077-8
IL10RA	0.675233	YES	YES	https://doi.org/10.1161/CIRCULATIONAHA.117.027889
IL2RB	0.66985	YES	YES	https://doi.org/10.1161/hypertensionaha.116.07084
TNFRSF13C	0.649738	YES	NO	https://doi.org/10.1126/sciadv.aas9944
GPR25	0.647856	YES	NO	https://doi.org/10.1111/febs.70117, https://doi.org/10.1016/j.jdermsci.2020.09.010
CD180	0.624585	Related	YES	https://doi.org/10.1007/s00441-021-03488-7
IL12RB1	0.591876	YES	YES	https://doi.org/10.3389/fphar.2020.00129
MUC5AC	0.587984	Related	NO	https://doi.org/10.1371/journal.pone.0058658

Supplementary Table 5. Annotation of the 25 top receptors predicted by the PKN-receptor model – [heart failure showcase]. Receptors are classified as related to fibrosis and to the heart if a publication can justify this link. In violet are the receptors classified as related to both fibrosis and heart condition, in orange are the receptors related to one of the categories, and in red are the receptors related to neither of the categories.

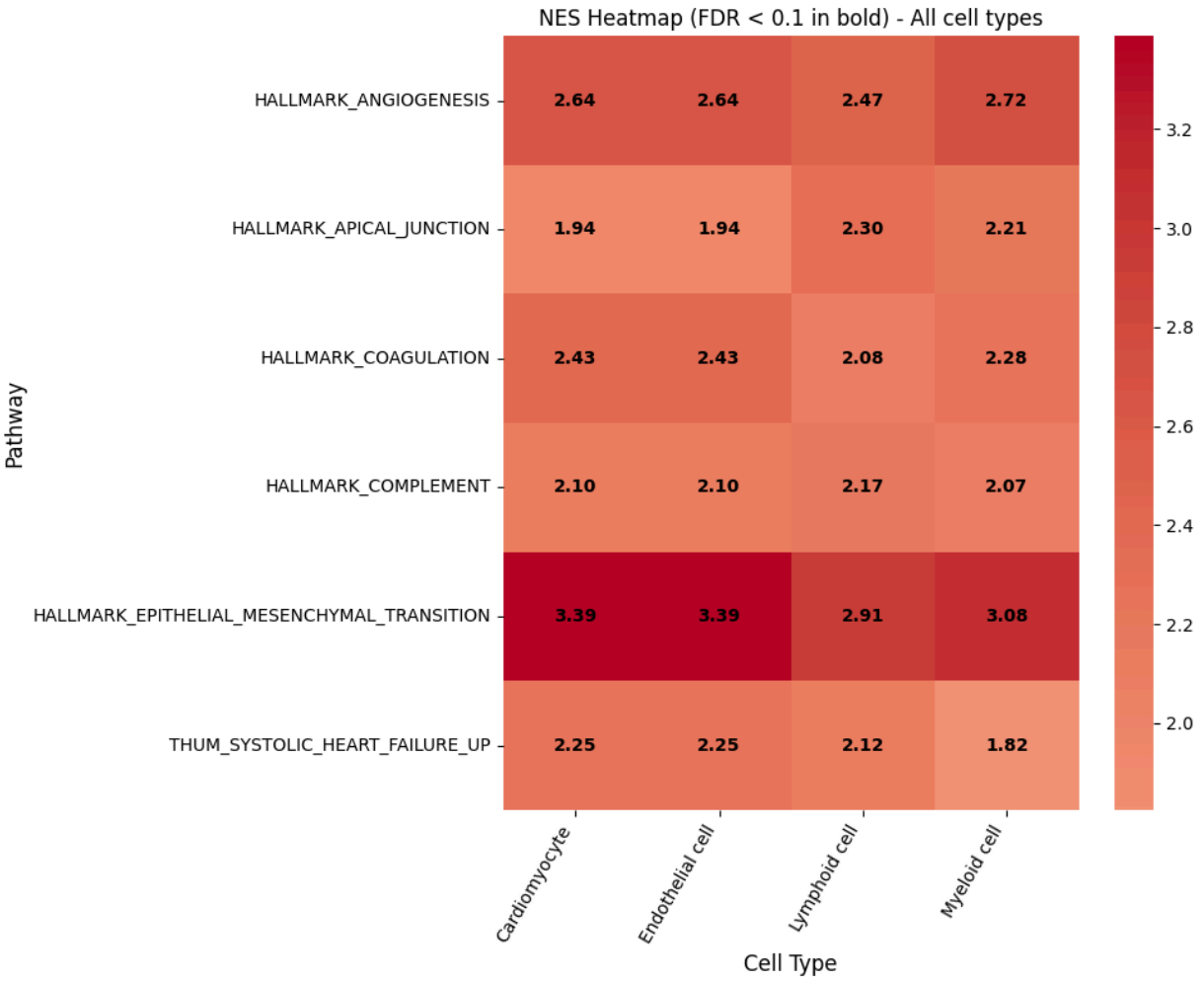
gene	score	ReCoN rank	PKN rank	fibrosis	heart	doi
GPR182	37.167563	7	34	YES	YES	https://doi.org/10.1016/j.jlr.2024.100679
IL17RE	31.16377	10	31	YES	Related	https://doi.org/10.3389/fcvm.2024.1470362
ADGRG1	24.971621	12	50	YES	YES	https://doi.org/10.1042/bsr20240826, https://doi.org/10.1177/1535370214529395
PTGDR2	17.204726	4	10	YES	YES	https://doi.org/10.1371/journal.ppat.1011812, https://doi.org/10.1016/j.yjmcc.2022.03.011
NPTXR	12.26263	19	36	NO	NO	–
DRD4	12.252395	1	2	YES	NO	https://doi.org/10.1111/acer.12047
CDH11	11.894496	50	105	YES	YES	https://doi.org/10.3390/ijms24076549
LRP11	11.498974	48	94	NO	NO	–
TRPV2	11.183415	8	9	YES	YES	https://doi.org/10.1038/s41374-019-0349-z
CD180	10.82372	18	24	YES	YES	https://doi.org/10.1038/cddis.2016.140, https://doi.org/10.1172/jci.insight.160684

Supplementary Table 6. Annotation of the 10 top receptors with the highest gene movement ranked higher in ReCoN than in the PKN-receptor model – [heart failure showcase]. Receptors are classified as related to fibrosis and to the heart if a publication can justify this link. In violet are

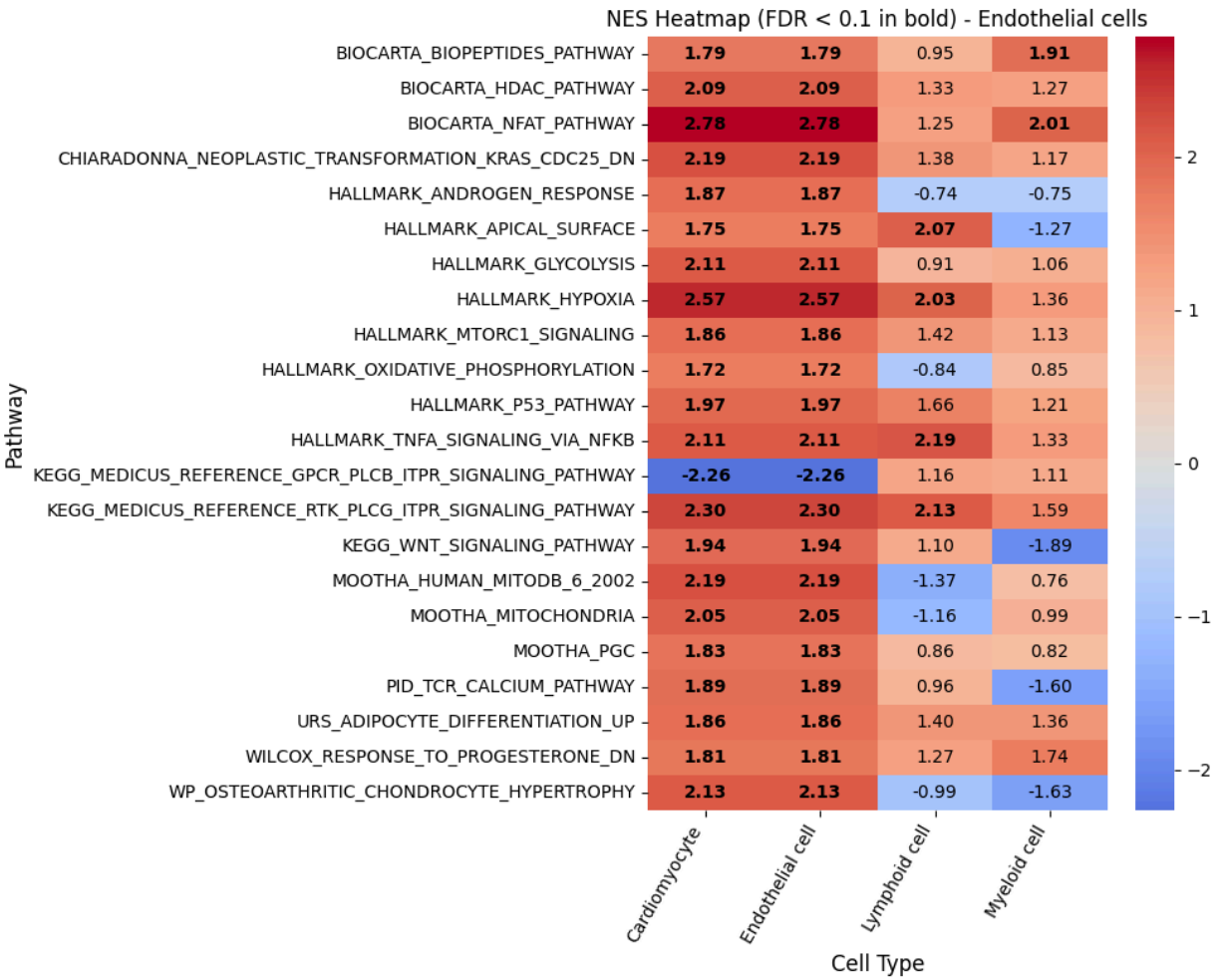
the receptors classified as related to both fibrosis and heart condition, in orange are the receptors related to one of the categories, and in red are the receptors related to neither of the categories.

gene	score	ReCoN rank	PKN rank	fibrosis	heart	doi
IFNGR2	25.24087	3	1	YES	YES	https://doi.org/10.1007/s10741-013-9393-8
APCDD1	22.129364	28	8	NO	NO	–
IL13RA1	16.93084	6	3	YES	YES	https://doi.org/10.1161/JAHA.116.005108
IL1RAP	16.331981	52	18	YES	YES	https://doi.org/10.1161/circheartfailure.124.011729
CD40LG	13.440412	29	13	YES	YES	https://doi.org/10.1016/j.ijcard.2018.12.076
IL10RA	12.16169	54	20	YES	YES	https://doi.org/10.1161/CIRCULATIONAHA.117.027889
OSMR	11.983381	68	35	YES	YES	https://doi.org/10.1186/s12967-023-04163-x
IL15RA	11.862556	15	6	YES	NO	https://doi.org/10.3389/fimmu.2024.1404891
IL5RA	10.333617	13	7	YES	Related	https://doi.org/10.1111/jcmm.18493
CAMK2A	7.037267	128	88	YES	YES	https://doi.org/10.1016/j.jbc.2021.100893

Supplementary Table 7. Annotation of the 10 top receptors with the highest gene movement ranked lower in ReCoN than in the PKN-receptor model – [heart failure showcase]. Receptors are classified as related to fibrosis and to the heart if a publication can justify this link. In violet are the receptors classified as related to both fibrosis and heart condition, in orange are the receptors related to one of the categories, and in red are the receptors related to neither of the categories.

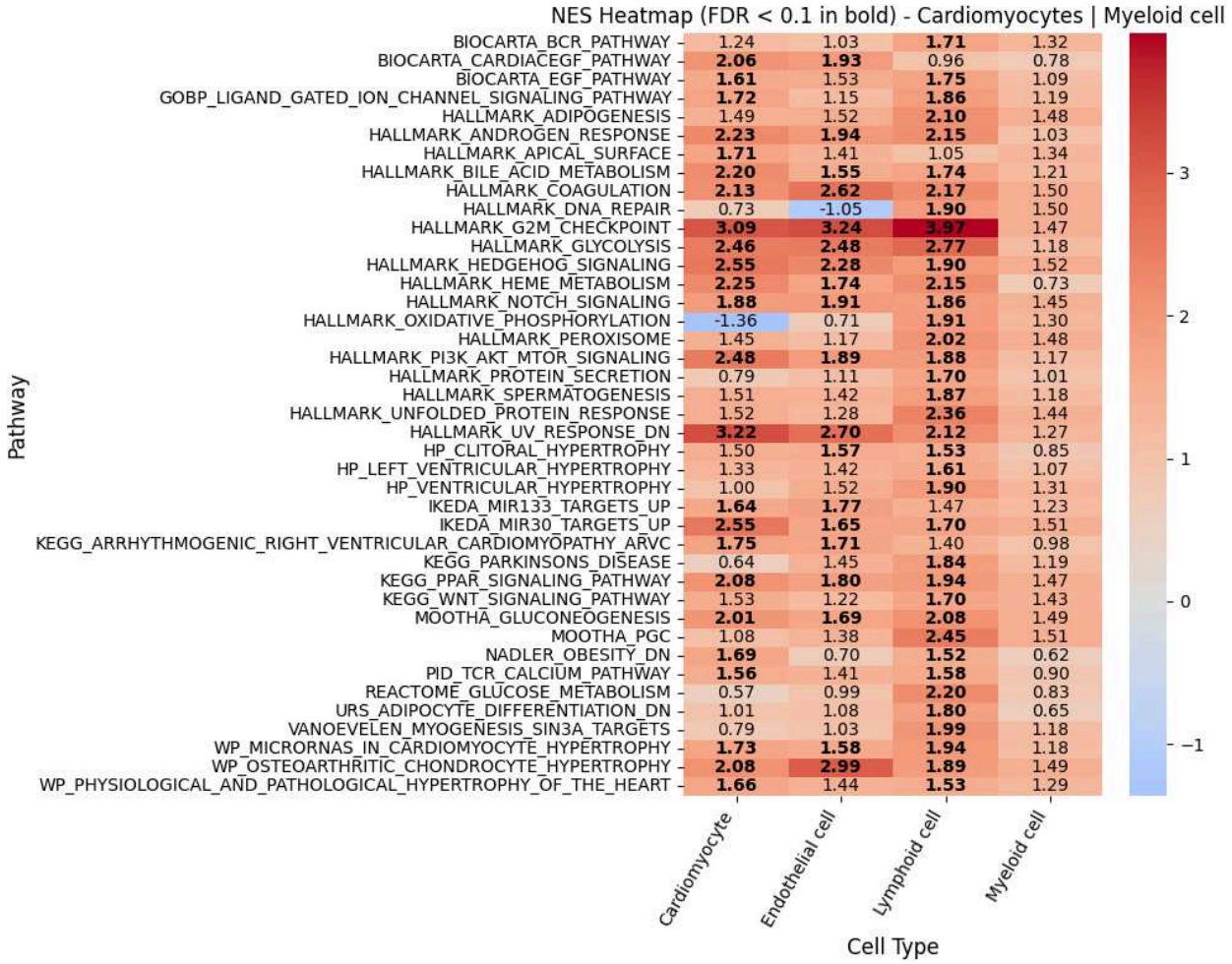


Supplementary Table 8. Enriched pathways in all cardiac cell types upstream of fibrosis genes – [cardiac fibrosis showcase]. Here are all the gene sets enriched in all four cell types in the upstream exploration of section 4.4.4.



Supplementary Table 9. Enriched pathways in endothelial cells upstream of fibrosis genes – [cardiac fibrosis showcase]. Here are all the gene sets enriched in Endothelial cells, in the upstream exploration of section 4.4.4.

Supplementary Table 10. Enriched pathways in all cardiac cell types downstream of fibrosis genes – [cardiac fibrosis showcase]. Here are all the gene sets enriched all four cell types considered in the downstream exploration of section 4.4.4.



Supplementary Table 11. Enriched pathways in cardiomyocytes and myeloid cells downstream of fibrosis genes – [cardiac fibrosis showcase]. Here are all the gene sets enriched in either cardiomyocytes or myeloid cells, and not in lymphoid cells, in the downstream exploration of section 4.4.4.

Heart Cell Atlas	ReHeat2 - HF
CM	Ventricular Cardiomyocyte
Endo	Endothelial cell
Fib	Fibroblast
Lymphoid	Lymphoid
Myeloid	Myeloid
PC	Mural cell
vSMCs	Mural cell

Supplementary Table 12. Cell type matching between ReHeat2 and Heart Cell Atlas datasets – [cardiac fibrosis & heart failure showcase]. The left column corresponds to cell type annotations in the Heart Cell Atlas samples, the right column corresponds to the matching used in our model and in downstream analysis.