Gene Regulatory Network (GRN) Inference: Introduction

Sergio Peignier

Section 1

Definitions

Defining the Landscape

- Understand GRN inference $\rightarrow$ Define core concepts:
  - Gene Expression
  - Regulation of Gene Expression
  - Gene Regulatory Networks
- In-depth semantic analysis of recent surveys:
  - Saint-Antoine et al. (2019, 2020)
  - Rybakov et al. (2024)
  - Da et al. (2025)
  - Huynh-Thu et al. (2018)
  - Zhao et al. (2021), Wang et al. (2022)

Gene Expression

Mechanism vs. Measurement

- **Central Dogma of Molecular Biology:** Universal information storage medium flows unidirectionally:

$$\text{DNA} \rightarrow \text{mRNA} \rightarrow \text{Proteins.}$$

- Two distinct epistemological definitions of **Gene Expression**:

1. **Processual Definition:** The dynamic biochemical *mechanism* through which genetic information is read and converted.
2. **Quantitative Definition:** The resulting *output*, measurable abundance or concentration levels of mRNAs and proteins at a given time point.

Regulation of Gene Expression

- **Regulation Condition:** An entity regulates gene expression if it directly **participates** in product generation or alters final product abundance.

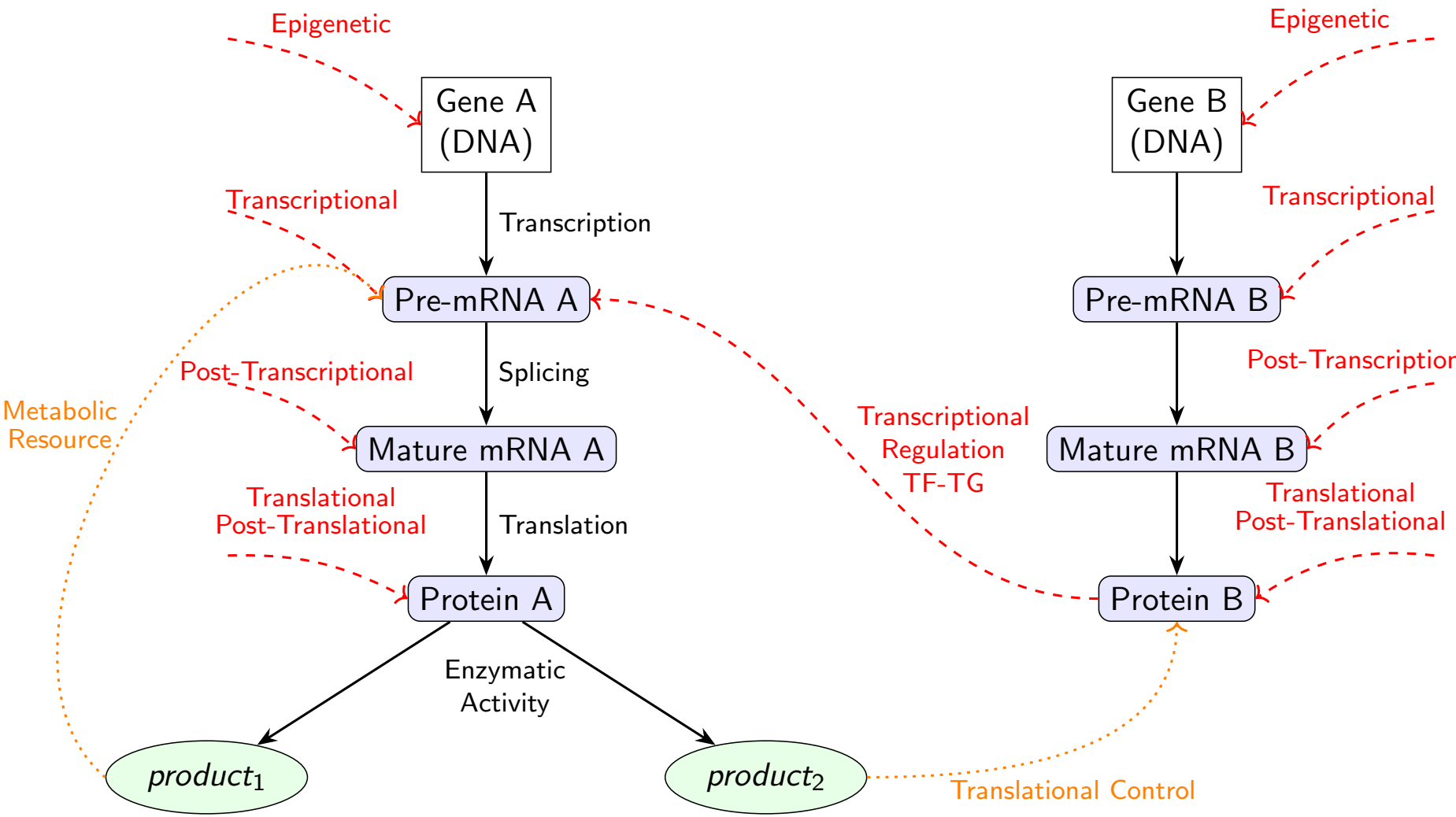
- **Multilayered Nature:** Gene regulation is universally recognized as indirect and multi-scale:

  - **Epigenetic Level:** DNA methylation and histone acetylation altering chromatin architecture.

  - **Transcriptional:** Mainly ensure by Transcription Factors (TFs), specialized proteins binding to regulatory DNA sequences of their Target Genes (TGs) to modulate transcription rates.

  - **Post-Transcriptional/Translational:** Splicing, mRNA stability, translation efficiency, and protein turnover.

Eukaryotic Gene Regulation Complexity



Gene Regulatory Networks: Models vs. Reality

- **Terminological Diversity in Surveys:** Gene Networks, Gene Interaction Networks (GINs), Gene Regulatory Networks (GRNs)...
- **Nodes and Edges Diverge:**
  - **Nodes**: Mostly genes (as objects or random expression variables), but also some surveys include molecular regulators or phenotypes.
  - **Edges**: General mutual interactions or specific TF-TG regulations.
  - **Attributes**: Signed edges (+/-) for activation/inhibition, weighted edges for strength or confidence.
- **The Reification Risk:** Conflating abstract statistical models with physical reality via the "circuit board fallacy" (over-simplifying biology into hardwired microchips).

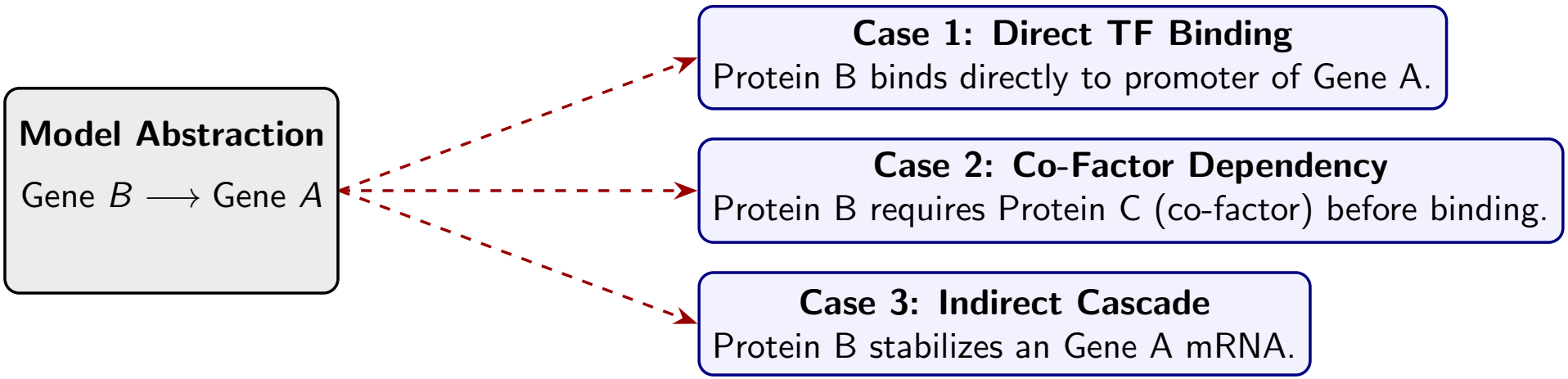
The Epistemology of GRN Inference

Two distinct paradigms shape how inference tasks are approached:

1. **The Realist / Mechanistic Discovery Paradigm:**
   - Aims to retrieve **true** underlying biological regulatory links.
   - Evaluates algorithms against a "true ground truth" network.
2. **The Instrumentalist / Data-Driven Modeling Paradigm:**
   - Views GRNs as **limited** but useful mathematical over-simplifications (Box: *"All models are wrong, but some are useful"*).
   - Focuses on capturing and predicting gene expression patterns rather than reflecting physical ground truth.

GRN Regulation as Causal Claims

- GRN edges implicitly assert underlying causality (Gene $B \rightarrow$ Gene A).
- In reality, interactions involve long, multi-step chains of events.
- Co-factors (e.g., Protein C assisting Protein B) highlight how network topologies depend on human modeling choices.
- A "direct edge" is an epistemological reduction, an abstraction rendering complex biology computationally tractable.



Section 2

Datasets

Gene Expression Matrices: Structure

- Most GRN inference methods rely on **gene expression matrices**.
  - **Columns:** Represent individual genes.
  - **Rows:** Represent experimental samples/conditions.

	TP53	**BRCA1**	**EGFR**	**MYC**
Control 1	5.2	12.1	3.4	8.9
Control 2	4.8	11.5	3.9	9.2
Treatment A	14.3	6.2	10.1	15.6
Treatment B	15.1	5.8	11.0	16.2

- **Gene Products:** Can measure mRNA, proteins, or other products (though *transcriptomic datasets (mRNA)* predominate).

- **Measurement Technologies:**
  - *Microarrays* (developed in the 1990s).
  - *RNA-seq* (leading technology for transcript abundance).

Bulk vs. Single-Cell RNA-seq

Bulk RNA-seq	**scRNA-seq**
Large group/tissue pool.	Cell level.
Dense matrices.	High sparsity and drop-outs.
Coarse-grained description	Fine-grained view.
$\sim 10 \rightarrow 100$ samples	$\sim 10^3 \rightarrow 10^6$ cells
Prone to *curse of dimensionality*	High biological noise

High sparsity in scRNA-seq can hamper inference reliability, sometimes performing comparably to pure random predictors

Overcoming Data Limitations: Preprocessing

To improve GRN inference reliability, **preprocessing** is crucial:

- **Gene Selection:** Given the massive number of genes, restrict analysis to a subset relevant to the biological question:
  - Genes with significantly higher biological variance across samples.
  - Differentially expressed genes.

- **scRNA-seq Corrections:**
  - Normalization.
  - Scale. transformation (log, sqrt, Z-score)
  - Dropout correction.
  - Discretization techniques to minimize noise and sparsity.

Experimental Categorization of Datasets

Datasets are categorized based on their underlying experimental design:

1. **Non-Interventional vs. Interventional (Perturbation):**
   - *Non-interventional:* Purely observational studies.
   - *Interventional:* Direct modification of the biological system (e.g., RNA interference, knock-downs).
2. **One-Time vs. Time Series:**
   - *One-time:* Samples drawn from a single time point.
   - *Time series:* Gene expression measured across sequential time points.

Integrating Complementary Data Sources

Using gene expression matrices alone can be limiting. Modern approaches combine **multiple sources of information**:

- **Complementary Data Types:** ChIP-seq, DNA methylation, TF binding site databases, ATAC-seq, gene ontologies, pathway data, and previously inferred GRNs.
- **Phenotype Datasets (Gene Sets):** Binary matrices mapping genes to phenotypes/pathways.
- **Dual Purpose of External Data:** Incorporate *prior structural knowledge* to constrain model search space.

Modeling Heterogeneous Gene Expression Data

- **The Core Assumption:** Standard inference assumes a dataset can be explained by a *single network structure*.
- **Heterogeneous Scenarios:** When data spans diverse conditions, models should either:
  - Infer multiple distinct GRNs, or
  - Model a network that *rewires across time/conditions*.
- **Model Expressive Power:** Simple linear models fail on complex heterogeneity; advanced models (e.g., Random Forests, Neural Networks) are better suited to capture multi-condition variability.

Section 3

GRN Inference Methods

Overview of Literature Taxonomies

Reviews categorize GRN inference methods with distinct perspectives:

- **Huynh-Thu et al. (2018):** Probabilistic, Dynamical, Multi-network, and Data-driven models.
- **Saint-Antoine et al. (2020) & Rybakov et al. (2024):** Correlation, Mutual Information, Regression, Bayesian, Logical networks, and Differential equations.
- **Zhao et al. (2021):** Model-based, Information theory-based, and Machine learning-based.
- **Da Silva et al. (2025):** Discrete, Continuous, and Mechanistic models.

Most existing taxonomies intertwine **input data properties** with **methodological assumptions**, causing **ambiguities**.

Critique of Existing Categorizations

- **Huynh-Thu:** Grouping by data type (dynamic/multi-network) overlaps with methodological families (probabilistic/data-driven), causing classification conflicts (e.g., Dynamic Bayesian Networks).
- **Saint-Antoine / Rybakov:** A "miscellaneous" category lumps together disparate models like ODEs, Boolean networks, and Gaussian graphical models; ignores input data characteristics.
- **Zhao:** Unclear boundary between "model-based" and "machine learning-based" (both fit models to data); includes correlation under information theory.
- **Da Silva:** Merges most methods into an extremely broad "continuous" category and fails to distinguish mechanistic from continuous ODE models.

Dual-Criterion Categorization

To resolve ambiguities, we propose separating two orthogonal principles:

1. **Input Data:** What kind of data is being processed?
2. **Methodological:** What underlying mathematical model is assumed?

Methodological Characterization Criteria

Five Major **Methodological Families**:

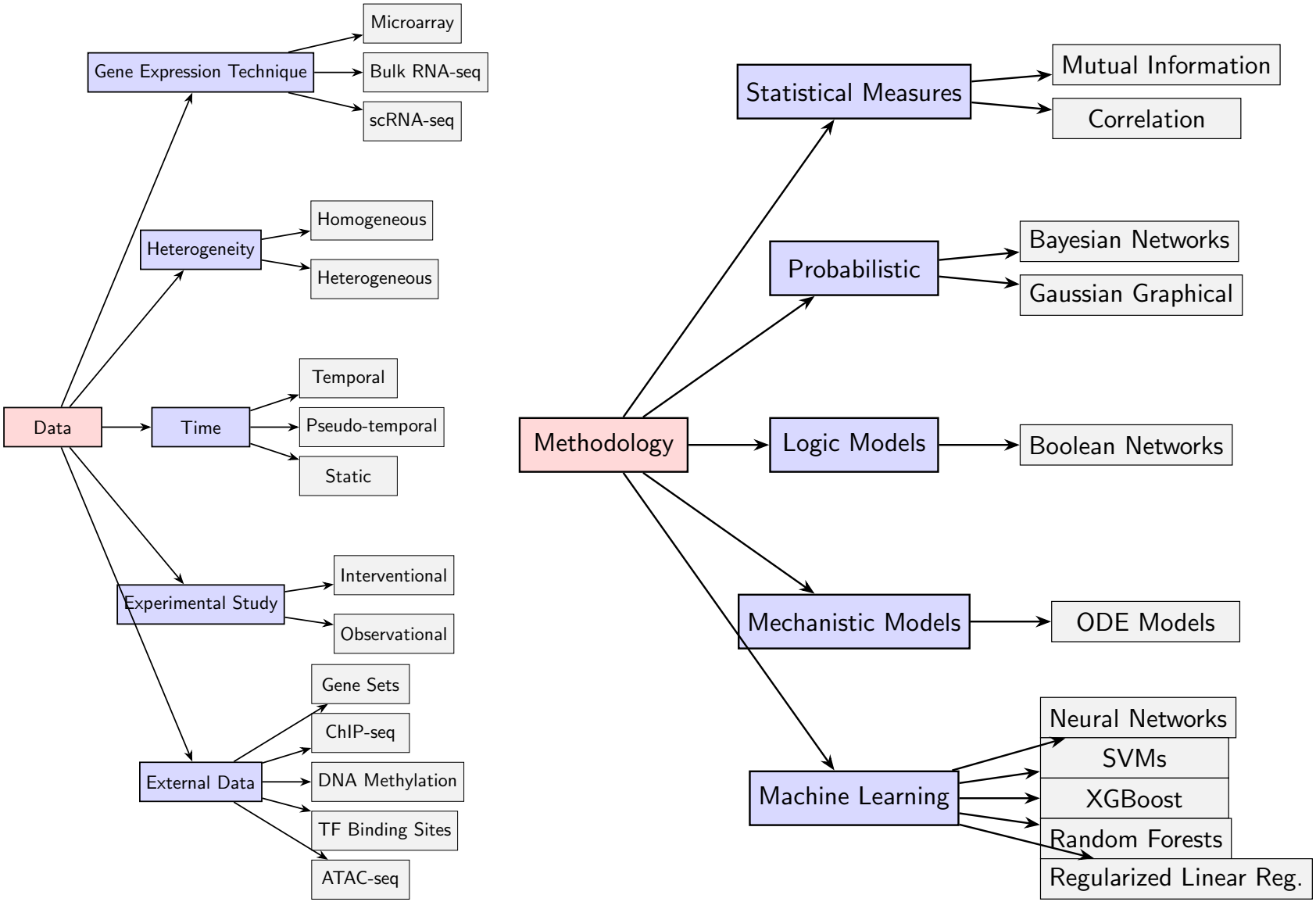
- Statistical Measures (Correlation, Mutual Information)
- Probabilistic Models (Bayesian, Gaussian Graphical)
- Logic Models (Boolean Networks - BNs)
- Mechanistic Models (Ordinary Differential Equations - ODEs)
- Machine Learning Models (Regression, RFs, SVMs, Neural Networks)

Input Data Characterization Criteria

Five key criteria characterize the **input data used** in GRN inference:

1. **Gene Expression Technology:** Microarray, bulk RNA-seq, or single-cell RNA-seq (scRNA-seq).
2. **Data Heterogeneity:** Homogeneous samples vs. heterogeneous mixtures (e.g., multi-cell-type datasets).
3. **Temporal Dimension:** Static datasets, time series, or pseudo-temporal developmental states.
4. **Experimental Design:** Observational (non-interventional) vs. interventional (perturbations/knock-downs).
5. **External Prior Data:** Integration of ChIP-seq, DNA methylation, ATAC-seq, gene sets, or prior networks.

Methodological Categorization Tree



Methodological Trade-offs

According to comparative literature:

- **Model-based/Mechanistic methods** (ODEs but also BNs):
  - Best suited for small-scale networks.
  - Allow to model and perform *in-silico* experiments.
  - Require careful parameter tuning and robust and rich temporal data.
- **Information theory / Statistical methods**:
  - Ideal for solving steady-state gene expression data efficiently.
- **Machine Learning methods**
  - Excel at scaling up to infer large-scale gene networks by handling complex feature dependencies.

Section 4

GRN Inference Evaluation

Overview

- **The Core Challenge:** Assessing inference quality is notoriously difficult due to a **lack of complete knowledge** regarding true biological regulatory interactions.
- **Two Main Evaluation Paradigms:**
  1. **External Validation:** Compares inferred GRNs against a reference "ground truth" benchmark dataset using binary classification metrics. (Tied to the *Realist* paradigm).
  2. **Internal Evaluation:** Computes metrics from gene expression data alone to check robustness and biological plausibility. (Tied to the *Instrumentalist* paradigm).

External Evaluation: Benchmark Datasets

External validation relies on pre-established reference benchmarks:

- **Curated Datasets:**
  - *STRING:* Functional interactions from experimental results and text-mining.
  - *NonSpecific Networks:* DoRothEA, RegNetwork, and TRRUST (integrating human/mouse TF-TG relationships).
  - *Cell-Type-Specific ChIP-Seq Networks:* ENCODE, ChIP-Atlas, and ESCAPE (binding sites in specific cell types).

- **Synthetic Datasets:**
  - Controlled environments via simulators: GeneNetWeaver (bulk), SERGIO, HARISSA, BoolODE/BEELINE (scRNA-seq).
  - Simulators model stochasticity, transcriptional bursting, technical noise.

External Evaluation and Performance Gaps

- **Binary Classification Framework:**
  - Reference links $\rightarrow$ true positives, missing links $\rightarrow$ false positives.
  - High class imbalance (few positive regulatory links)
  - Preferred metrics:*Precision-Recall curves* and *MCC*
- **The Synthetic-to-Real Performance Gap:**
  - Many algorithms excel on synthetic data but perform close to random predictors on real datasets (e.g., SINCERITIES, SCRIBE, SINGE).
  - Algorithms like *GENIE3, GRNBOOST2, and PIDC* show relatively stable performance on experimental data.
  - General-purpose and scRNA-seq specific methods often struggle to reliably predict network structures (Chen & Mar, 2018; BEELINE evaluations).

Internal Evaluation in GRN Inference

The Challenge

- **Scarcity of Ground Truths:** True networks are typically restricted to model species.
- **The Generalization Gap:** High benchmark performance does *not* guarantee success on novel datasets.

Internal Evaluation assesses model quality using gene expression data alone:

- **Train/Test Splitting (Cross-Validation):** Train on a subset and evaluate generalization on unseen data to retain only robust models.
- **Stability Analysis:** Measure consistency across data subsets; reliable methods tend to uncover stable regulatory links.

Trustability Problem

- **The Trustability Challenge:** Many inferred GRNs are highly sensitive to input datasets, lack biological plausibility, and fail to generalize.
- **Key Issues in Data-Driven Approaches:**
  - Overfitting to specific datasets.
  - Poor generalization to unseen conditions.
  - Violation of known biological constraints and low interpretability.
- **Core Scientific Question:** How can we ensure that inferred GRNs are robust, interpretable, and biologically plausible without relying solely on elusive ground-truth networks?

Section 5

Data Driven Inference

Which TFs regulate the TG?

- X_{tf_1}, X_{tf_2} and X_{tf_3}: **Gene expressions** of 3 **TFs**
- X_{tg_0}: **Gene expression** of a **TG**

$$X_{tf_1} = \begin{pmatrix} 1.2 \\ 0.0 \\ 3.3 \\ 0.0 \end{pmatrix} \quad X_{tf_2} = \begin{pmatrix} 1.5 \\ 1.5 \\ 1.5 \\ 1.5 \end{pmatrix} \quad X_{tf_3} = \begin{pmatrix} 0.0 \\ 2.1 \\ 0.3 \\ 4.1 \end{pmatrix} \quad X_{tg_0} = \begin{pmatrix} 0.8 \\ 0.1 \\ 1.4 \\ 0.0 \end{pmatrix}$$

$$tf_1 \xrightarrow{?} tg_0 \qquad tf_2 \xrightarrow{?} tg_0 \qquad tf_3 \xrightarrow{?} tg_0$$

Intuitive answer

- X_{tf_1}, X_{tf_2} and X_{tf_3}: **Gene expressions** of 3 **TFs**
- X_{tg_0}: **Gene expression** of a **TG**

$$X_{tf_1} = \begin{pmatrix} 1.2 \\ 0.0 \\ 3.3 \\ 0.0 \end{pmatrix} \quad X_{tf_2} = \begin{pmatrix} 1.5 \\ 1.5 \\ 1.5 \\ 1.5 \end{pmatrix} \quad X_{tf_3} = \begin{pmatrix} 0.0 \\ 2.1 \\ 0.3 \\ 4.1 \end{pmatrix} \quad X_{tg_0} = \begin{pmatrix} 0.8 \\ 0.1 \\ 1.4 \\ 0.0 \end{pmatrix}$$

$$\boxed{tf_1 \rightarrow tg_0} \qquad tf_2 \not\rightarrow tg_0 \qquad \boxed{tf_3 \rightarrow tg_0}$$

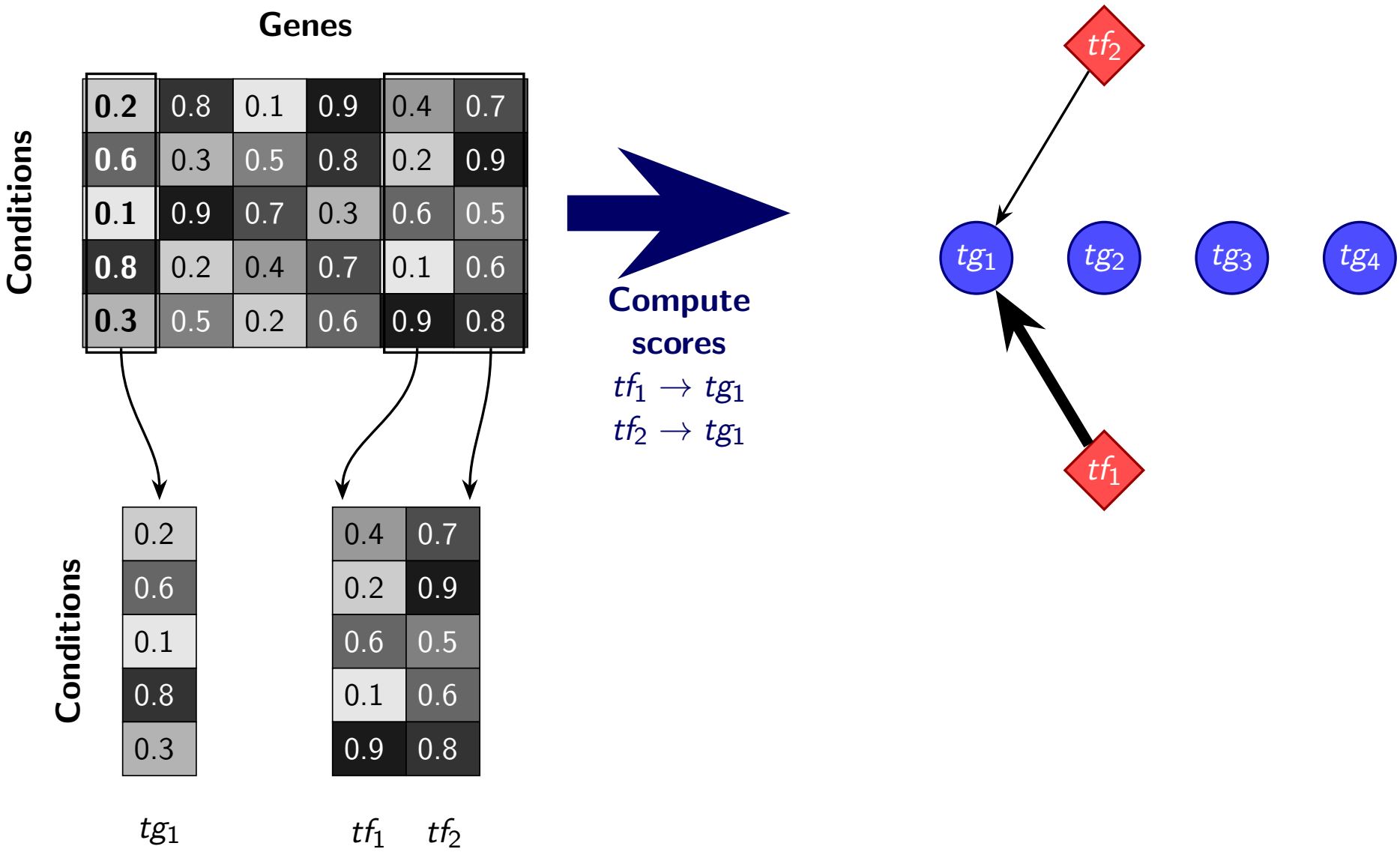
General Data-Driven GRN Inference Methodology

Scope: Applies to both statistical and machine learning approaches.

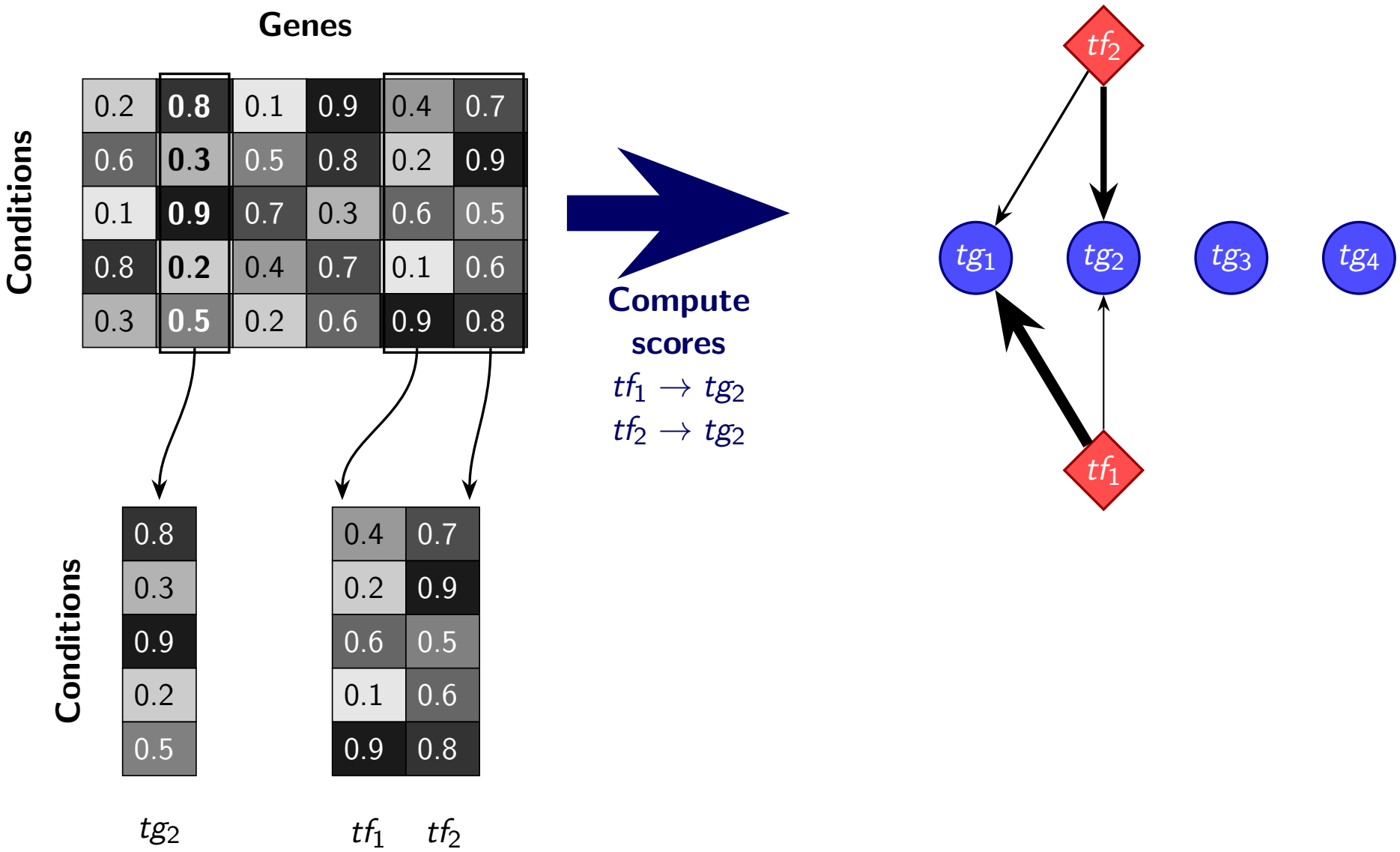
Core Inference Workflow:

1. For each TG, analyze its expression profile against candidate regulators (TFs).
2. **Scoring:** Assign a quantitative score measuring the statistical association between the expression vectors of the TG and its candidate TFs.
3. **Thresholding & Filtering:** Apply a threshold to filter the scored links and construct the final sparse regulatory network.

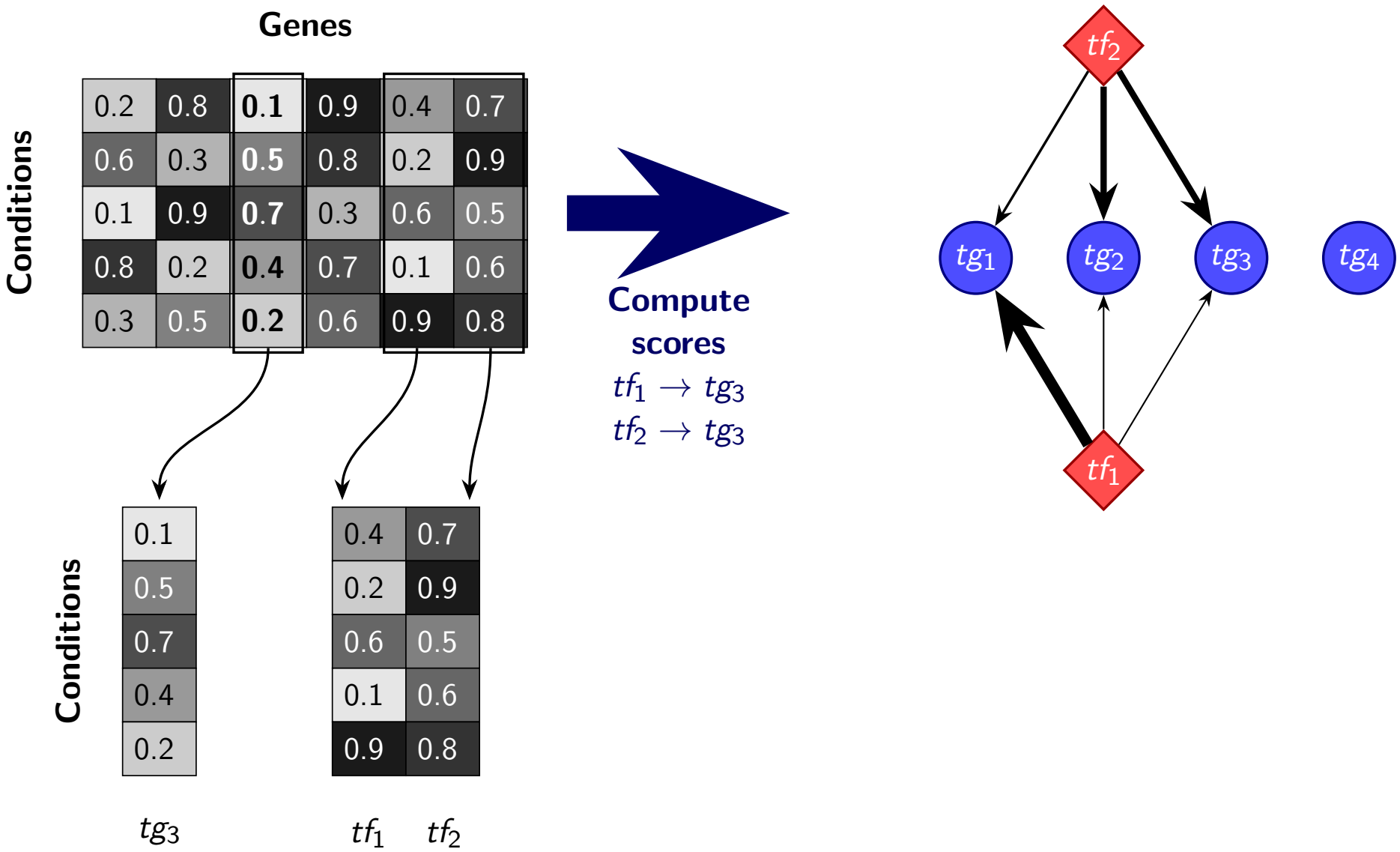
Data Driven Inference: Score Links



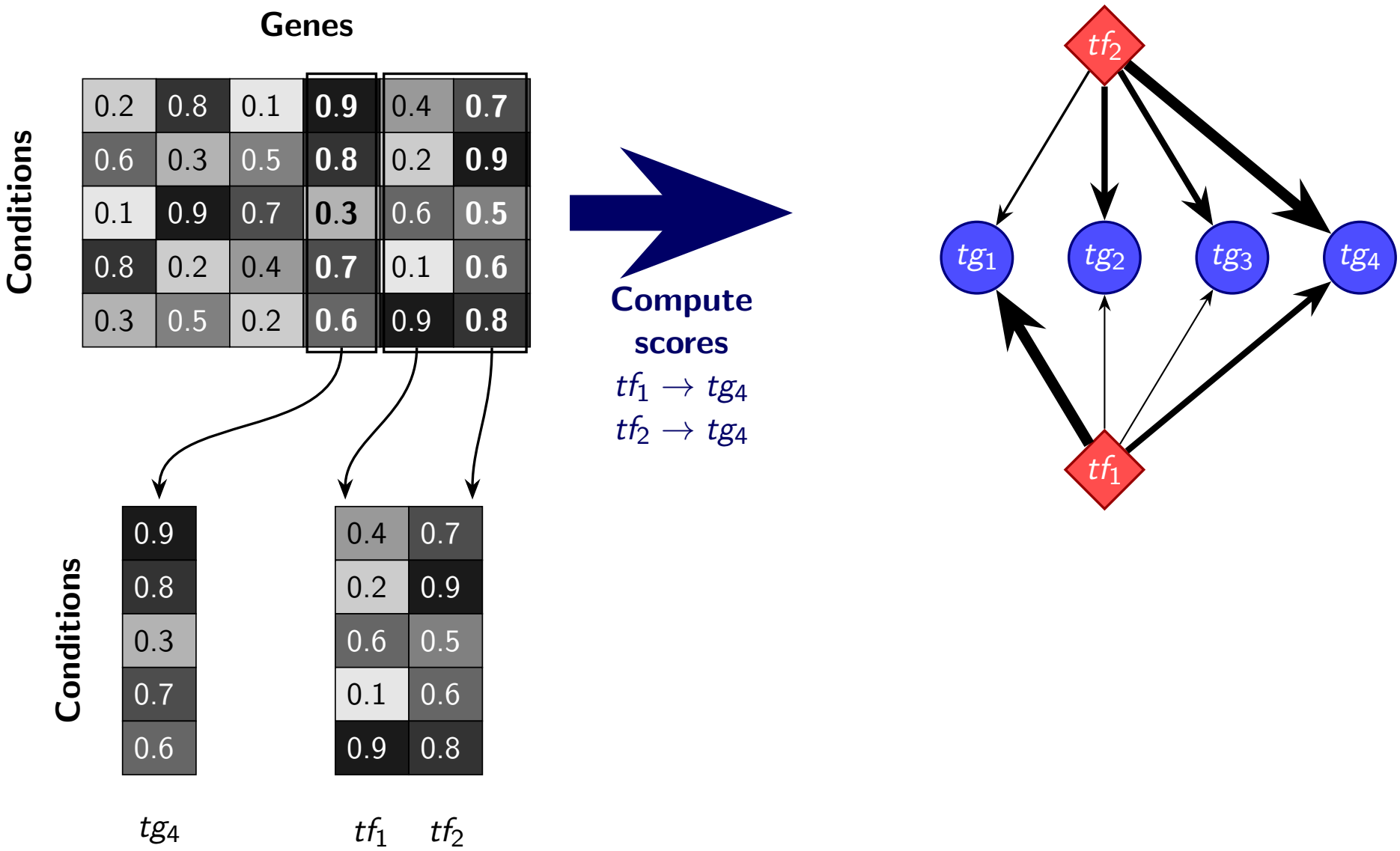
Data Drive Inference: Score Links



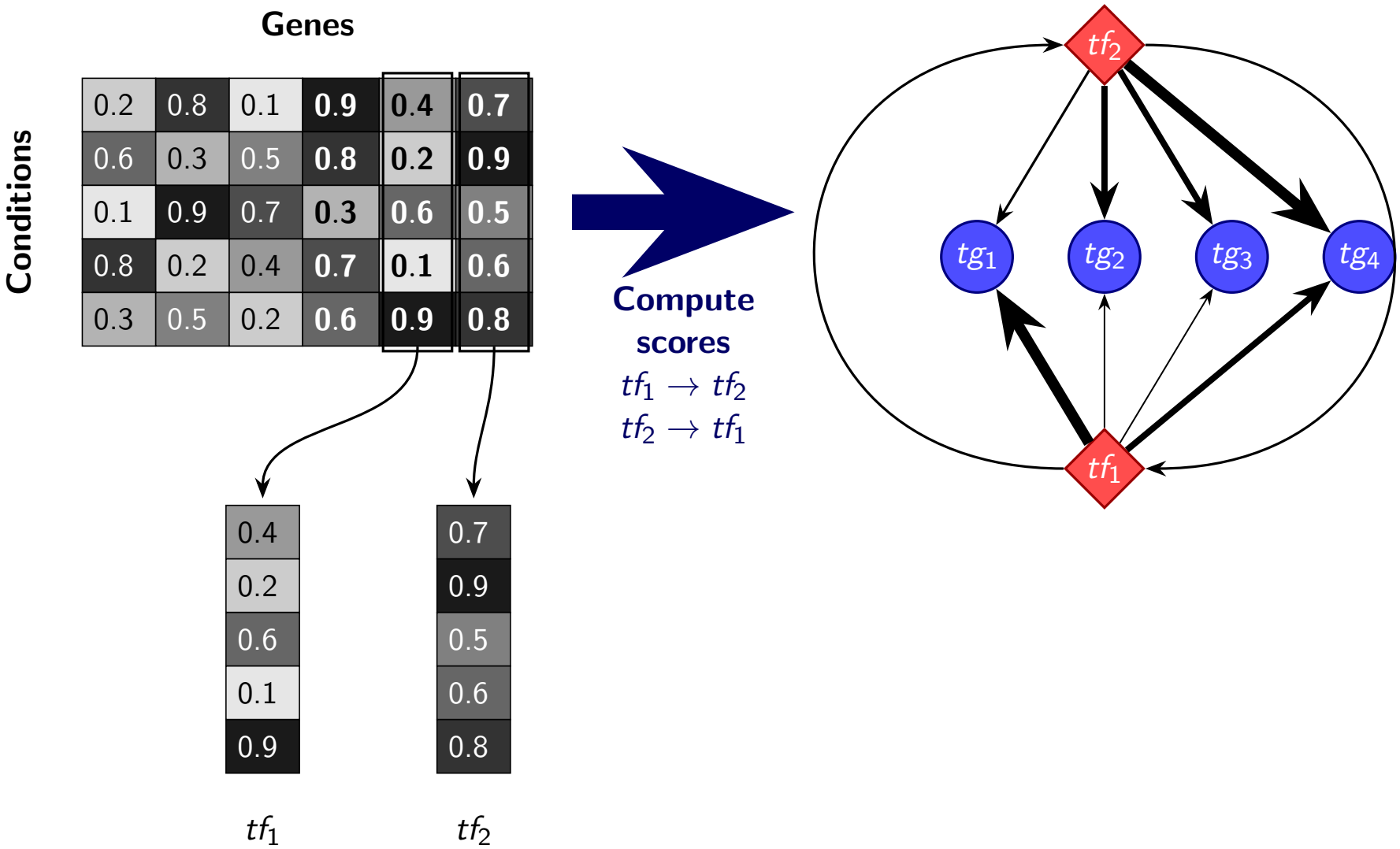
Data Driven Inference: Score links



Data Driven Inference: Score links



Data Driven Inference : Score edges



Data Drive Inference: Filter edges

