MOREL: MULTI-OMICS RELATIONAL LEARNING

ICLR 2022

Arman Hasanzadeh Ehsan Hajiramezanali, Nick Duffiel and Xiaoning Qian

Presenter: Qingmei Wang

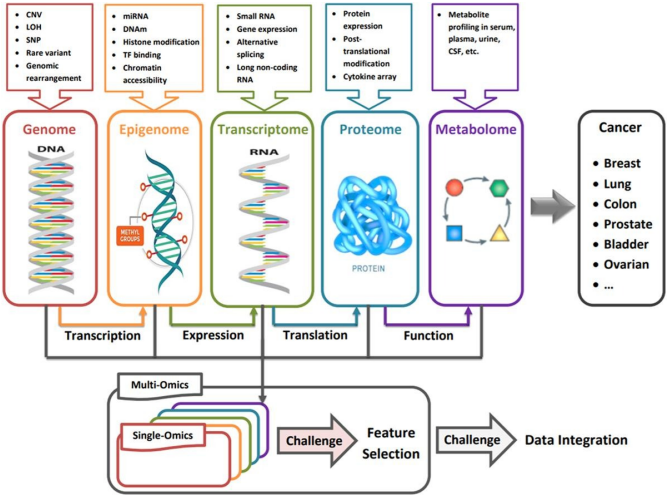
October 19, 2023

Overview

- Preliminaries
- Problem Formulation and Notations
- Method
- Experiments
- Conclusion

What's Multi-omics Data

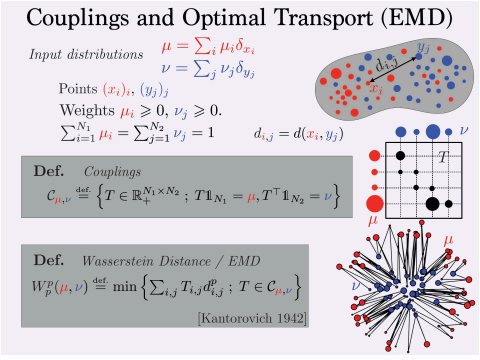
"Multi-omics" refers to multiple "omics" datasets: Genomics, Epigenomics, Transcriptomics, Proteomics, and Metabolomics.

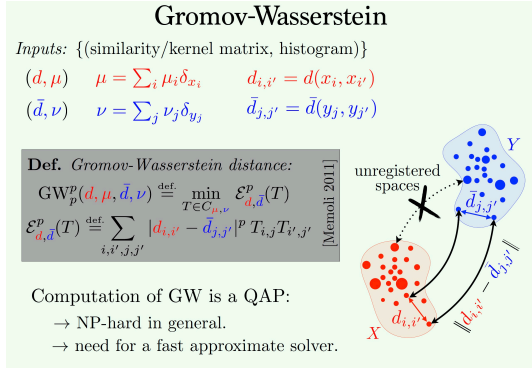


Molecular Relational Learning

- Molecular Interaction Prediction: Crucial for discovering or designing new molecules.
- Drug-Drug Interaction Prediction: Learning whether a combination of two drugs will produce side effects.

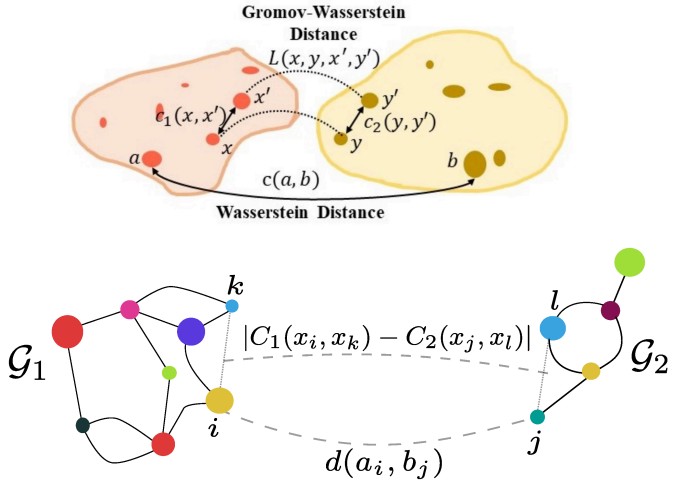
Wasserstein Distance & Gromov-Wasserstein Distance





- WD quantifies the geometric discrepancy between two probability distributions.
- GWD has been proposed as a natural extension of WD when a meaningful transportation cost between the distributions cannot be defined.

Fused Gromov-Wasserstein



FGW distance is an optimization problem consisting of a Wasserstein term and a GW term

Problem Formulation and Notation

Aims to infer the interrelations among entities, i.e. features, across all of the views.

- multiple views, $\mathcal{V}$, of data are given
- assume that the structure information is available for some of the views $\mathcal{V}_s \subset \mathcal{V}$
- $\mathcal{V}_u = \mathcal{V} \backslash \mathcal{V}_s$ are unstructured
- $\mathfrak{G}_s = \left\{\mathcal{G}^{(v)}\right\}_{v \in \mathcal{V}_s}$ as the set of graphs for structured views
- $\mathfrak{A}_s = \left\{\mathbf{A}^{(v)}\right\}_{v \in \mathcal{V}_s}$ as adjacency matrices
- $\mathcal{X}_s = \left\{\mathbf{X}^{(v)}\right\}_{v \in \mathcal{V}_s}$ as the set of node attributes for structured views
- $\mathcal{X}_u = \left\{\mathbf{X}^{(v)}\right\}_{v \in \mathcal{V}_u}$ as the set of data for unstructured views
- N_v denotes the number of nodes in structured views and number of features for unstructured views

Problem Formulation and Notation

- MoReL infers the interactions among the nodes in $\mathfrak{G}_s$ and features in $\mathcal{X}_u$
- Inter-relations by a multi-partite graph with $\sum_{v \in \mathcal{V}} N_v$ nodes and a multi-adjacency tensor $\mathcal{A} = \left\{ \mathbf{A}^{(vv')} \right\}_{v,v' \in \mathcal{V}, v \neq v'}$, where $\mathbf{A}^{(vv')}$ is the $N_v \times N_{v'}$ bi-adjacency matrix between views v and v'

MOREL Generative Model

- Define a hierarchical Bayesian model for MoReL with three sets of latent variables:
- $\mathcal{H} = \mathcal{H}_s \cup \mathcal{H}_u = \left\{\mathbf{H}^{(v)}\right\}_{v \in \mathcal{V}_s \cup \mathcal{V}_u}$, which captures the (hidden) structural information
- $\mathcal{A}$, which encodes the interaction among features across views
- $\mathcal{Z} = \mathcal{Z}_s \cup \mathcal{Z}_u = \left\{\mathbf{Z}^{(v)}\right\}_{v \in \mathcal{V}_s \cup \mathcal{V}_u}$, which summarizes the feature/attribute specific information
- The joint probability of observations and latent variables factorizes as follows:

$$p_\theta\left(\mathcal{X}_u, \mathcal{X}_s, \mathfrak{A}_s, \mathcal{H}, \mathcal{A}, \mathcal{Z}\right) =$$
$$p_{\theta_x}\left(\mathcal{X}_u \mid \mathcal{Z}_u\right) p_{\theta_x}\left(\mathcal{X}_s \mid \mathcal{Z}_s\right) p_{\theta_g}\left(\mathfrak{A}_s \mid \mathcal{H}_s\right) p_{\theta_z}(\mathcal{Z} \mid \mathcal{H}, \mathcal{A}) p_{\theta_a}(\mathcal{A} \mid \mathcal{H}) p(\mathcal{H}).$$

FGW Distance.

- Given two structured probability distributions, $\mathbf{\Lambda} \in \mathcal{P}(\mathbb{X})$ and $\mathbf{\Delta} \in \mathcal{P}(\mathbb{Y})$, FGW is defined as follows:

$$
\begin{aligned}
\mathcal{D}_{\mathrm{FGW}}(\mathbf{\Lambda}, \mathbf{\Delta}) &= \alpha\, \mathcal{D}_{\mathrm{W}}(\mathbf{\Lambda}, \mathbf{\Delta}) + \beta\, \mathcal{D}_{\mathrm{GW}}(\mathbf{\Lambda}, \mathbf{\Delta}) \\
&= \alpha \inf_{\pi_w \in \Pi(\mathbb{X} \times \mathbb{Y})} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \pi_w}[c^{(\mathbb{X}\mathbb{Y})}(\mathbf{x}, \mathbf{y})] \\
&\quad + \beta \inf_{\pi_{gw} \in \Pi(\mathbb{X} \times \mathbb{Y})} \mathbb{E}_{(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}') \sim \pi_{gw}}[\,|\!|\, c^{(\mathbb{X})}(\mathbf{x}, \mathbf{x}') - c^{(\mathbb{Y})}(\mathbf{y}, \mathbf{y}')\, |\!|\,],
\end{aligned}
\tag{3}
$$

where $\alpha, \beta \in [0, 1]$ are scalar hyper-parameters, $\Pi(\mathbb{X} \times \mathbb{Y})$ is the set of all admissible couplings between $\mathbf{\Lambda}$ and $\mathbf{\Delta}$, and $c^{(\mathbb{X}\mathbb{Y})}$, $c^{(\mathbb{X})}$, and $c^{(\mathbb{Y})}$ are corresponding transportation cost functions. $\mathcal{D}_{\mathrm{FGW}}$ can be further simplified by choosing π_w to be equal to π_{gw} (Chen et al., 2020).

Relational learning via FGW

- FGW distance based decoder for every pair of views, in which each view independently belongs to either structured or unstructured views, i.e. $(v, v') \in \mathcal{V}$
- define the transportation cost functions $c^{(vv')}$ and $c^{(v)}$, and then approximate $\mathcal{D}_{\text{FGW}}$
- define the inter-cost function for the first term of FGW, i.e. $\mathcal{D}_{\text{W}}$, as follows:

$$c^{(vv')}\left(\mathbf{H}_{i,:}^{(v)}, \mathbf{H}_{j,:}^{(v')}\right) = 1 - \sigma\left(\mathbf{H}_{i,:}^{(v)}\left(\mathbf{H}_{j,:}^{(v')}\right)^{T}\right); \quad v, v' \in \mathcal{V}$$

where σ denotes the sigmoid function, and $\mathbf{H}_{i,:}^{(v)}$ represents the structural latent variable of node/feature i in view v

Relational learning via FGW

- For the structured views, define the cost function as a combination of the shortest path distance from the graph and the distance between structural latent variables.Given the normalized shortest path distance matrix between every pair of nodes in the input graph $\mathbf{D}^{(v)}$:

$$c^{(v)}\left(\mathbf{H}_{i,:}^{(v)}, \mathbf{H}_{j,:}^{(v)}\right) = \mathbf{D}^{(v)} \odot \left(1 - \sigma\left(\mathbf{H}_{i,:}^{(v)}\left(\mathbf{H}_{j,:}^{(v)}\right)^{T}\right)\right); \quad \text{for } v \in \mathcal{V}_s$$

  where $\odot$ denotes the Hadamard product.

- For unstructured views, we define the cost function between two features as follows:

$$c^{(v)}\left(\mathbf{H}_{i,:}^{(v)}, \mathbf{H}_{j,:}^{(v)}\right) = 1 - \sigma\left(\mathbf{H}_{i,:}^{(v)}\left(\mathbf{H}_{j,:}^{(v)}\right)^{T}\right); \quad \text{for } v \in \mathcal{V}_u$$

Relational learning via FGW

- Rewrite $\mathcal{D}_{\text{FGW}}$ between two views of data with shared transport matrixas follows:

$$\mathcal{D}_{\text{FGW}}\left(p\left(\mathbf{H}^{(v)}\right), p\left(\mathbf{H}^{(v')}\right)\right)$$
$$\sum_{i=1}^{N_v}\sum_{j=1}^{N_{v'}} \min_{\mathbf{T}_{gw}^{(vv')}\in\Pi} \sum_{\mathbf{H}_{i,:}^{(v)},\mathbf{H}_{j,:}^{(v')},\mathbf{H}_{i,:}^{(v)\prime},\mathbf{H}_{j,:}^{(v')},} \left[\alpha c^{(vv')}\left(\mathbf{H}_{i,:}^{(v)}, \mathbf{H}_{j,:}^{(v')}\right) + \right.$$
$$\left.\beta c^{(v)}\left(\mathbf{H}_{i,:}^{(v)}, \mathbf{H}_{i,:}^{(v)\prime}\right) - c^{(v')}\left(\mathbf{H}_{j,:}^{(v')}, \mathbf{H}_{j,:}^{(v')\prime}\right)\|\right].$$

- To approximate the FGW distance, firstly deploy GW algorithm to obtain $\mathbf{T}_{gw}^{(vv')}$ and $\mathcal{D}_{\text{GW}}$, and then utilize $\mathbf{T}_{gw}^{(vv')}$ along with the defined transportation cost $c^{(vv')}$ to calculate Wasserstein distance term in $\mathcal{D}_{\text{FGW}}$.

Prior Construction

- The paper imposes independent zero-mean unit-variance Gaussian priors on elements of $\mathcal{H}$. The prior for $\mathcal{Z}$ is a multivariate Gaussian distribution whose mean and diagonal covariance matrix are constructed from the inferred multi-partite graph and the structural latent variable $\mathcal{H}$.
- The paper uses two graph neural networks (GNNs)$g_{pz}^{(\mu)}$ and $g_{pz}^{(\sigma)}$ to map $\mathcal{H}$ and $\mathcal{A}$ to the parameters of $p_{\theta_z}(\mathcal{Z})$
- Specifically,

$$p_{\theta_z}(\mathcal{Z} \mid \mathcal{H}, \mathcal{A}) = \prod_{v \in \mathcal{V}_s} \prod_{i=1}^{N_v} p_{\theta_z}\left(\mathbf{Z}_{i,:}^{(v)} \mid \mathcal{H}, \mathcal{A}\right); \quad p_{\theta_z}\left(\mathbf{Z}_{i,:}^{(v)} \mid \mathcal{H}, \mathcal{A}\right) = \mathcal{N}\left(\boldsymbol{\mu}_{pz}^{(v,i)}, \boldsymbol{\sigma}_{pz}^{(v,i)}\right),$$

$$\text{with} \quad \left[\boldsymbol{\mu}_{pz}^{(v,i)}\right]_{v,i} = g_{pz}^{(\mu)}(\mathcal{H}, \mathcal{A}), \quad \left[\boldsymbol{\sigma}_{pz}^{(v,i)}\right]_{v,i} = g_{pz}^{(\sigma)}(\mathcal{H}, \mathcal{A}).$$

Likelihood of observations

Likelihood of observations. To reconstruct the input graphs in the structured views, we assume that the views and edges are conditionally independent. More specifically, we employ an inner-product decoder as follows:

$$p_{\theta_g}(\mathfrak{A}_s \mid \mathcal{H}_s) = \prod_{v \in \mathcal{V}_s} \prod_{i,j=1}^{N_v} p_{\theta_g}\left(\mathbf{A}_{i,j}^{(v)} \mid \mathbf{H}_{i,:}^{(v)}, \mathbf{H}_{j,:}^{(v)}\right);$$

$$p_{\theta_g}\left(\mathbf{A}_{i,j}^{(v)} \mid \mathbf{H}_{i,:}^{(v)}, \mathbf{H}_{j,:}^{(v)}\right) = \mathrm{Ber}\left(\sigma(\mathbf{H}_{i,:}^{(v)} (\mathbf{H}_{j,:}^{(v)})^T)\right).$$

To generate the features in unstructured views and node attributes in structured views, we assume that views are conditionally independent. Hence we can expand the feature reconstruction terms in the the equation (2) as follows:

$$p_{\theta_x}(\mathcal{X}_u \mid \mathcal{Z}_u) = \prod_{v \in \mathcal{V}_u} p_{\theta_x}(\mathbf{X}^{(v)} \mid \mathbf{Z}^{(v)}), \qquad p_{\theta_x}(\mathcal{X}_s \mid \mathcal{Z}_s) = \prod_{v \in \mathcal{V}_s} p_{\theta_x}(\mathbf{X}^{(v)} \mid \mathbf{Z}^{(v)}).$$

We note that p_{θ_x} could also be view specific depending on whether the node attributes/features in a view are discrete or continuous. In our experiments, we have deployed the Gaussian likelihood with the unit variance. The mapping from $\mathcal{Z}$ to the parameters of $p_{\theta_x}(\mathcal{X})$, in our case, the mean of the Gaussian distribution, can be any highly expressive function such as neural networks. We denote these functions by $f_{px}^{(v,s)}$ and $f_{px}^{(v,u)}$.

Posterior

Posterior. We model the posterior of the structural latent variables as a Gaussian distribution and infer its parameters independently for each view. More specifically,

$$q_{\phi_h}(\mathcal{H}_u \mid \mathcal{X}_u) = \prod_{v \in \mathcal{V}_u} q_{\phi_h}(\mathbf{H}^{(v)} \mid \mathbf{X}^{(v)}), \qquad q_{\phi_h}(\mathcal{H}_s \mid \mathcal{X}_u, \mathfrak{A}_s) = \prod_{v \in \mathcal{V}_s} q_{\phi_h}(\mathbf{H}^{(v)} \mid \mathbf{X}^{(v)}, \mathbf{A}^{(v)}).$$

We use two GNNs for each structured view, $\{g_{qh}^{(\mu,v)}(\mathbf{X}^{(v)}, \mathbf{A}^{(v)}), g_{qh}^{(\sigma,v)}(\mathbf{X}^{(v)}, \mathbf{A}^{(v)})\}_{v \in \mathcal{V}_s}$, and two fully connected neural networks per unstructured view, $\{f_{qh}^{(\mu,v)}(\mathbf{X}^{(v)}), f_{qh}^{(\sigma,v)}(\mathbf{X}^{(v)})\}_{v \in \mathcal{V}_u}$, to map inputs to the mean and variance of the posteriors. We consider the variational distribution of $\mathcal{Z}$ to be a multivariate Gaussian distribution, and it is factorized as follows:

$$q_{\phi_z}(\mathcal{Z}_u \mid \mathcal{X}_u) = \prod_{v \in \mathcal{V}_u} q_{\phi_z}(\mathbf{Z}^{(v)} \mid \mathbf{X}^{(v)}), \qquad q_{\phi_z}(\mathcal{Z}_s \mid \mathcal{X}_u, \mathfrak{A}_s) = \prod_{v \in \mathcal{V}_s} q_{\phi_z}(\mathbf{Z}^{(v)} \mid \mathbf{X}^{(v)}, \mathbf{A}^{(v)}).$$

We use two GNNs per structured view, $\{g_{qz}^{(\mu,v)}(\mathbf{X}^{(v)}, \mathbf{A}^{(v)}), g_{qz}^{(\sigma,v)}(\mathbf{X}^{(v)}, \mathbf{A}^{(v)})\}_{v \in \mathcal{V}_s}$, and two fully connected neural networks for each unstructured view, $\{f_{qz}^{(\mu,v)}(\mathbf{X}^{(v)}), f_{qz}^{(\sigma,v)}(\mathbf{X}^{(v)})\}_{v \in \mathcal{V}_u}$, in the same fashion as q_{ϕ_h} to infer parameters of q_{ϕ_z}.

Objective Function

$$
\begin{aligned}
\mathcal{L} &= -\,\mathrm{ELBO} + \mathcal{L}_{\mathrm{FGW}} \\
&= \mathbb{E}_{q_{\phi_z}(\mathcal{Z}_u, \mathcal{H}_u \mid \mathcal{X}_u)} \log p_\theta(\mathcal{Z}_u \mid \mathcal{A}, \mathcal{H}) + \mathbb{E}_{q_{\phi_z}(\mathcal{Z}_s, \mathcal{H}_s \mid \mathcal{X}_s, \mathfrak{A}_s)} \log p_\theta(\mathcal{Z}_s \mid \mathcal{A}, \mathcal{H}) \\
&\quad - \mathbb{E}_{q_{\phi_z}(\mathcal{Z}_u \mid \mathcal{X}_u)} \log q_{\phi_z}(\mathcal{Z}_u \mid \mathcal{X}_u) - \mathbb{E}_{q_{\phi_z}(\mathcal{Z}_s \mid \mathcal{X}_s, \mathfrak{A}_s)} \log q_{\phi_z}(\mathcal{Z}_s \mid \mathcal{X}_s, \mathfrak{A}_s) \\
&\quad + \mathbb{E}_{q_{\phi_h}(\mathcal{H}_u \mid \mathcal{X}_u)} \log p(\mathcal{H}_u) + \mathbb{E}_{q_{\phi_h}(\mathcal{H}_s \mid \mathcal{X}_s, \mathfrak{A}_s)} \log p(\mathcal{H}_s) \\
&\quad - \mathbb{E}_{q_{\phi_h}(\mathcal{H}_u \mid \mathcal{X}_u)} \log q_{\phi_h}(\mathcal{H}_u \mid \mathcal{X}_u) - \mathbb{E}_{q_{\phi_h}(\mathcal{H}_s \mid \mathcal{X}_s, \mathfrak{A}_s)} \log q_{\phi_h}(\mathcal{H}_s \mid \mathcal{X}_u, \mathfrak{A}_s)) \\
&\quad + \mathbb{E}_{q_{\phi_z}(\mathcal{Z}_u \mid \mathcal{X}_u)} \log p_{\theta_x}(\mathcal{X}_u \mid \mathcal{Z}_u) + \mathbb{E}_{q_{\phi_z}(\mathcal{Z}_s \mid \mathcal{X}_s, \mathfrak{A}_s)} \log p_{\theta_x}(\mathcal{X}_s \mid \mathcal{Z}_s) \\
&\quad + \mathbb{E}_{q_{\phi_h}(\mathcal{H}_s \mid \mathcal{X}_s, \mathfrak{A}_s)} \log p_{\theta_g}(\mathfrak{A}_s \mid \mathcal{H}_s) + \sum_{v \in \mathcal{V}} \sum_{\substack{v' \in \mathcal{V} \\ v' \neq v}} \mathcal{D}_{\mathrm{FGW}} \left(p(\mathbf{H}^{(v)}),\, p(\mathbf{H}^{(v')}) \right).
\end{aligned}
$$

Figure 1: Overall loss function as the sum of the negative variational ELBO and FGW regularization terms

Datasets and Evaluation Metrics

Datasets:

- microbiomemetabolite interactions in cystic fibrosis (CF)
  - Positive accuracy refers to the accuracy of identifying validated interactions with *P.* aeruginosa.
  - Negative accuracy exploits the fact that there should not be any common metabolite targets between known anaerobic microbes (Veillonella, Fusobacterium, Prevotella, and Streptococcus) and notable pathogen P. aeruginosa.
  - Having both higher positive and negative accuracy is desired
- gene-drug interactions in precision medicine
  - prediction sensitivity of identifying known interactions in the test sets
  - the average density of the overall constructed graphs

Experiments

Table 1: Comparison of positive accuracy (in %) on CF dataset at negative accuracy of $> 97\%$.

	SRCA	BCCA	MoReL $_{uu}$	MoReL $_{us}$
Positive accuracy	26.41	28.30 ± 3.21	56.16 ± 1.85	63.77 ± 1.11

Table 2: Comparison of prediction sensitivity (in %) in the precision medicine experiment.

Avg. degree	0.10	0.15	0.20	0.25	0.30	0.40	0.50
SRCA	8.03	12.00	17.15	20.70	26.85	34.93	45.79
BCCA	9.65 ± 0.75	14.34 ± 0.06	18.96 ± 0.42	23.29 ± 0.52	28.22 ± 0.66	38.02 ± 2.15	46.88 ± 1.88
MoReL $_{uu}$	11.29 ± 0.16	15.74 ± 0.62	21.21 ± 0.81	26.20 ± 1.10	30.47 ± 1.07	39.05 ± 0.75	50.19 ± 0.19
MoReL $_{us}$	12.79 ± 0.39	17.51 ± 2.21	22.82 ± 1.01	29.58 ± 1.08	35.05 ± 1.27	45.74 ± 1.75	53.16 ± 0.96

Comparison with Bayrel

Table 3: Positive accuracy (%) on CF dataset.

	BayReL	MoReL $_{ss}$
Positive Acc.	82.70 ± 4.70	89.50 ± 3.29

Table 4: Prediction sensitivity (%) in the precision medicine experiment.

Avg. degree	BayReL	MoReL $_{ss}$
0.4	47.90 ± 0.43	49.24 ± 1.64
0.5	56.76 ± 0.50	58.92 ± 0.40

Table 5: Positive accuracy (%) on CF dataset with unpaired samples.

	BayReL	MoReL $_{us}$
Positive Acc.	31.56	63.24 ± 2.13
Negative Acc.	72	97

Conclusion

- A novel Bayesian deep generative model that efficiently infers hidden molecular relations across heterogeneous views of data.
- To handle unpaired samples across the views of data;
- To combine multiple views from different data sources with any number of missing samples.