

A PROTOTYPES OF COVARIANCE MATRIX AND THEIR PROPERTIES

Here we present derivations for the properties shown in **Table 1**.

Covariance matrix $\Sigma(\mathcal{X}_\phi) = L + \frac{1}{n}J$. Let us begin with the first covariance matrix, $\Sigma(\mathcal{X}_\phi) = L + \frac{1}{n}J$. Its expected smoothness index under the condition where $\mathbb{E}(\mathcal{X}_\phi) = \mathbf{0}$ can be calculated following Eq. (2) in the main text

$$\mathbb{E}(\mathcal{S}(\phi)) = \text{tr}(L\Sigma(\mathcal{X}_\phi)), \quad (15)$$

$$= \text{tr}\left[L\left(L + \frac{1}{n}J\right)\right]. \quad (16)$$

In the situation where \mathcal{G} is a connected graph, we can apply Gutman & Xiao (2004); Chebotarev & Shamis (2006); Van Mieghem et al. (2017)

$$LL^\dagger = L^\dagger L = I - \frac{1}{n}J, \quad (17)$$

where I is the unit matrix, to derive

$$\mathbb{E}(\mathcal{S}(\phi)) = \text{tr}[L^2 + L(I - L^\dagger L)], \quad (18)$$

$$= \text{tr}(L^2 + L - LL^\dagger L). \quad (19)$$

Because the Moore–Penrose pseudoinverse satisfies Barata & Hussein (2012)

$$LL^\dagger L = L, \quad (20)$$

we can eventually reformulate Eq. (19) to obtain

$$\mathbb{E}(\mathcal{S}(\phi)) = \text{tr}(L^2), \quad (21)$$

the exact results shown in **Table 1**. Eq. (21) suggests that the expected smoothness of mapping ϕ is fully determined by the graph topology properties conveyed by graph Laplacian L if $\mathcal{X}_\phi \sim \mathcal{N}(\mathbf{0}, L + \frac{1}{n}J)$.

Given covariance matrix $\Sigma(\mathcal{X}_\phi) = L + \frac{1}{n}J$, variables $X_\phi(i)$ and $X_\phi(j)$ will evolve inversely (i.e., stronger negative covariance) if nodes v_i and v_j are connected by an edge with larger weight. Therefore, the defined Gaussian variable, $\mathcal{X}_\phi \sim \mathcal{N}(\mathbf{0}, L + \frac{1}{n}J)$, is more applicable to node heterogeneity and local structure descriptions, where edge weights measure node difference.

Covariance matrix $\Sigma(\mathcal{X}_\phi) = L^\dagger + \frac{1}{n}J$. Then we turn to the second covariance matrix, $\Sigma(\mathcal{X}_\phi) = L^\dagger + \frac{1}{n}J$. Under the same condition introduced above, we can derive its expected smoothness index shown in **Table 1**

$$\mathbb{E}(\mathcal{S}(\phi)) = \text{tr}(L\Sigma(\mathcal{X}_\phi)), \quad (22)$$

$$= \text{tr}\left[L\left(L^\dagger + \frac{1}{n}J\right)\right], \quad (23)$$

$$= \text{tr}(I) - \frac{1}{n}\text{tr}(J) + \text{tr}[L(I - L^\dagger L)], \quad (24)$$

$$= n - 1, \quad (25)$$

where Eqs. (24-25) are obtained by applying Eq. (17) and Eq. (20) subsequently. Eq. (25) suggests that the expected smoothness of mapping ϕ in a graph characterized by $\mathcal{X}_\phi^\heartsuit \sim \mathcal{N}(\mathbf{0}, L^\dagger + \frac{1}{n}J)$ is independent of graph topology and fully determined by graph size.

To understand what kind of information is captured by $\Sigma(\mathcal{X}_\phi) = L^\dagger + \frac{1}{n}J$, we need to consider the precision matrix Q (the inverse of covariance matrix) of Gaussian variable $\mathcal{X}_\phi^\heartsuit \sim \mathcal{N}(\mathbf{0}, L^\dagger + \frac{1}{n}J)$

$$Q := \Sigma(\mathcal{X}_\phi)^{-1} = L + \frac{1}{n}J. \quad (26)$$

Eq. (26) is derived according to the invertible property of $L + \frac{1}{n}J$ Xiao & Gutman (2003); Chebotarev & Shamis (2006)

$$\left(L + \frac{1}{n}J\right)^{-1} = L^\dagger + \frac{1}{n}J. \quad (27)$$

Given that the partial correlation between $X_\phi(i)$ and $X_\phi(j)$, the actual values of \mathcal{X}_ϕ on nodes v_i and v_j , is determined by matrix Q Rue & Held (2005)

$$\text{corr}(X_\phi(i), X_\phi(j) | \mathcal{X}_\phi \setminus \{X_\phi(i), X_\phi(j)\}) := -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}} \quad (28)$$

$$= -\frac{L_{ij} + \frac{1}{n}}{\sqrt{(L_{ii} + \frac{1}{n})(L_{jj} + \frac{1}{n})}}, \quad (29)$$

variables $X_\phi(i)$ and $X_\phi(j)$ will have a stronger partial correlation if nodes v_i and v_j are connected by an edge with larger weight (i.e., a larger value of $-L_{ij}$). Therefore, Gaussian variable $\mathcal{X}_\phi^\heartsuit \sim \mathcal{N}(\mathbf{0}, L^\dagger + \frac{1}{n}J)$ is more applicable to node homogeneity and global structure descriptions, where edge weights represent node similarity.

B NECESSARY DETAILS OF CLASSIC GRAPH OPTIMAL TRANSPORT

Here we elaborate necessary details of classic graph-signal-based optimal transport approaches (e.g., GOT Petric Maretic et al. (2019); Petric Maretic (2021) and fGOT Maretic et al. (2022)). These approaches represent graphs $\mathcal{G}_a(V_a, E_a)$ and $\mathcal{G}_b(V_b, E_b)$ as variables $\mathcal{X}_\phi^a \sim \mathcal{N}(\mathbf{0}, \Sigma_a)$ and $\mathcal{X}_\phi^b \sim \mathcal{N}(\mathbf{0}, \Sigma_b)$ to formalize the optimal transport problem.

Permutation matrix. Classic approaches consider an optimal transport problem whose solution is constrained as a permutation matrix $M \in \mathbb{R}^{|V_b| \times |V_a|}$. Applying permutation M on graph \mathcal{G}_b , we can make \mathcal{G}_b evolve towards \mathcal{G}_a via

$$M \circ \mathcal{X}_\phi^b \sim \mathcal{N}(\mathbf{0}, M^T \Sigma_b M). \quad (30)$$

The permutation matrix set to which M belongs is defined as

$$\mathcal{M} = \left\{ M \in \{0, 1\}^{|V_a| \times |V_b|} \left| \forall i, \sum_j M_{i,j} = 1, \forall j, \sum_i M_{i,j} = 1 \right. \right\}, |V_b| = |V_a|, \quad (31)$$

$$\mathcal{M} = \left\{ M \in [0, \infty)^{|V_a| \times |V_b|} \left| \forall i, \sum_j M_{i,j} = |V_a|^{-1}, \forall j, \sum_i M_{i,j} = |V_b|^{-1} \right. \right\}, |V_b| \neq |V_a|, \quad (32)$$

where Eq. (31) is the standard definition of permutation matrix set Mena et al. (2018) and Eq. (32) is a generalization proposed in fGOT Maretic et al. (2022).

Sinkhorn operator. Mathematically, the *Sinkhorn operator* functions as an iterative normalization approach of the rows and columns of a matrix Mena et al. (2018). Its definition is given as

$$\zeta^0(M) = \exp(M), \quad (33)$$

$$\zeta^k(M) = \mathcal{T}_c \circ \mathcal{T}_r(\zeta^{k-1}(M)), \forall k \geq 1, \quad (34)$$

$$\zeta(M) = \lim_{k \rightarrow \infty} \zeta^k(M), \quad (35)$$

where $\mathcal{T}_r(X)$ denotes the row normalization

$$\mathcal{T}_r(X) = X \oslash (X \mathbf{1} \mathbf{1}^T), \forall X, \quad (36)$$

and $\mathcal{T}_c(X)$ denotes the column normalization

$$\mathcal{T}_c(X) = X \oslash (\mathbf{1} \mathbf{1}^T X), \forall X. \quad (37)$$

Note that operator \oslash denotes the element-wise division and $\mathbf{1}$ is a vector of ones.

Continuous approximation and implicit entropy regularization via the Sinkhorn operator.

As shown in Mena et al. (2018), the *Sinkhorn operator* can search the solution of a combinatorial assignment problem between an arbitrary permutation matrix $M \in \mathcal{M}$ and the doubly-stochastic matrices in the *Birkhoff polytope* \mathcal{B} Mena et al. (2017). In other words, the *Sinkhorn operator* finds an appropriate continuous approximation of M by elements in \mathcal{B} Mena et al. (2018). More specifically, previous works suggest

$$\varsigma(M/\tau) = \operatorname{argmax}_{B \in \mathcal{B}} [\operatorname{tr}(B^T M) + \tau \mathcal{H}(B)], \quad \forall \tau \in (0, \infty), \quad (38)$$

where $\mathcal{H}(\cdot)$ denotes the entropy Mena et al. (2018). The combinatorial assignment is constrained by an entropy regularization controlled by τ . In the case where τ approaches to zero, one can approximately find an ideal continuous approximation of permutation matrix M Mena et al. (2018)

$$\operatorname{argmax}_{B \in \mathcal{B}} \operatorname{tr}(B^T M) \simeq \lim_{\tau \rightarrow 0^+} \varsigma(M/\tau). \quad (39)$$

Please note that the entropy regularization effects in Eq. (36) are implied by the continuous approximation of permutation matrix M rather than by the optimal transport process itself. The optimization objective defined by Eq. (6) in the main text still belongs to pure optimal transport. To distinguish this type of entropy regularization from the entropy regularization explicitly defined for optimal transport in Eq. (8), we refer to the entropy regularization introduced by continuous approximation via the *Sinkhorn operator* as *implicit entropy regularization* while the entropy regularization explicitly defined as a part of optimal transport objective is referred to as *explicit entropy regularization*.

C ENTROPY-REGULARIZED OPTIMAL TRANSPORT BETWEEN GRAPHS

As we have described in the main text, we begin to formalize our approach by explicitly considering the entropy-regularized optimal transport between variables \mathcal{X}_ϕ^a and \mathcal{X}_ϕ^b Cuturi (2013)

$$\operatorname{EO}_\varepsilon(\mathcal{X}_\phi^a, \mathcal{X}_\phi^b) = \inf_{\gamma} \left(\int_{\Omega \times \Omega} \|x - y\|_2^2 d\gamma(x, y) + \varepsilon \int_{\Omega \times \Omega} \log \left(\frac{d\gamma}{d\rho_a d\rho_b} \right) d\gamma \right), \quad \text{s.t. } \gamma \in \Gamma_{ab} \quad (40)$$

where parameter $\varepsilon \in [0, \infty)$ denotes the entropy regularization magnitude. Similar to GOT Petric Maretic et al. (2019); Petric Maretic (2021) and fGOT Maretic et al. (2022), we also constrain the solution of Eq. (37) as a permutation matrix $M \in \mathcal{M}$ and continuously approximate it via the *Sinkhorn operator*.

The closed-form expression of $\operatorname{EO}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b)$. Based on the Gaussian properties of our proposed graph representation, we can follow the approach introduced in Mallasto et al. (2021) to derive an closed-form expression of $\operatorname{EO}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b)$. For convenience, we first introduce three frequently used matrices

$$K_{ab}^\varepsilon = I + \sqrt{I + \frac{16}{\varepsilon^2} \Sigma_a \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)}, \quad (41)$$

$$K_{aa}^\varepsilon = I + \sqrt{I + \frac{16}{\varepsilon^2} [\Sigma_a]^2}, \quad (42)$$

$$K_{bb}^\varepsilon = I + \sqrt{I + \frac{16}{\varepsilon^2} [\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)]^2}, \quad (43)$$

where notion I denotes the unit matrix. Based on Eqs. (38-40), we can analytically derive

$$\begin{aligned} & \operatorname{EO}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b) \\ &= \operatorname{tr} \left(\Sigma_a + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right) - \frac{\varepsilon}{2} (\operatorname{tr}(K_{ab}^\varepsilon) - \log \det(K_{ab}^\varepsilon) + |V_a| \log 2 - 2|V_a|), \end{aligned} \quad (44)$$

where $\det(\cdot)$ denotes the determinant.

Derivations of the closed-form expression of $\text{EO}_\varepsilon \left(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b \right)$. Below, we sketch the derivation process of Eq. (44). More mathematical proofs can be seen in Mallasto et al. (2021).

FIRST STEP. To offer a clear vision, we begin with analyzing $\text{EO}_\varepsilon \left(\mathcal{X}_\phi^a, \mathcal{X}_\phi^b \right)$ in Eq. (40). As suggested by Borwein et al. (1994); Csiszár (1975); Mallasto et al. (2021), Eq. (40) has a unique minimizer (referred to as the entropic transport plan)

$$\gamma^\varepsilon(x, y) = \alpha^\varepsilon(x) \beta^\varepsilon(y) \exp(-\varepsilon \|x - y\|_2^2) \rho_a(x) \rho_b(y), \quad \forall x, y \in \Omega, \quad (45)$$

where

$$\alpha^\varepsilon(x) \int_\Omega \rho_b(y) \beta^\varepsilon(y) \exp(-\varepsilon \|x - y\|_2^2) dy = 1, \quad \forall x \in \Omega, \quad (46)$$

$$\beta^\varepsilon(y) \int_\Omega \rho_a(x) \alpha^\varepsilon(x) \exp(-\varepsilon \|x - y\|_2^2) dx = 1, \quad \forall y \in \Omega. \quad (47)$$

Meanwhile, we can have the *entropic Kantorovich formulation* of Eq. (40) according to Marino & Gerolin (2020)

$$\begin{aligned} \text{EO}_\varepsilon(\mathcal{X}_\phi^a, \mathcal{X}_\phi^b) = & \sup_{\eta \in L_\varepsilon(\mathcal{X}_\phi^a), \psi \in L_\varepsilon(\mathcal{X}_\phi^b)} \left\{ \int_\Omega \rho_a(x) \eta(x) dx + \int_\Omega \rho_b(y) \psi(y) dy \right. \\ & \left. - \varepsilon \left[\int_{\Omega \times \Omega} \rho_a(x) \rho_b(y) \exp\left(\frac{\eta(x) + \psi(y) - \|x - y\|_2^2}{\varepsilon}\right) dx dy - 1 \right] \right\}, \quad (48) \end{aligned}$$

where $L_\varepsilon(\mathcal{X}_\phi^a)$ and $L_\varepsilon(\mathcal{X}_\phi^b)$ are corresponding classes of Entropy-Kantorovich potentials

$$L_\varepsilon(\mathcal{X}_\phi^a) = \left\{ \eta \mid \eta : \mathbb{R}^{|V_a|} \rightarrow \mathbb{R}, 0 < \int_\Omega \rho_a(x) \exp\left(\frac{\eta(x)}{\varepsilon}\right) dx < \infty \right\}, \quad (49)$$

$$L_\varepsilon(\mathcal{X}_\phi^b) = \left\{ \psi \mid \psi : \mathbb{R}^{|V_b|} \rightarrow \mathbb{R}, 0 < \int_\Omega \rho_b(y) \exp\left(\frac{\psi(y)}{\varepsilon}\right) dy < \infty \right\}. \quad (50)$$

The unique maximizers of Eq. (48), η^ε and ψ^ε , satisfy an analytic relation with the unique minimizer of Eq. (40) Mallasto et al. (2021)

$$\gamma^\varepsilon(x, y) = \exp\left(\frac{\eta^\varepsilon(x) + \psi^\varepsilon(y) - \|x - y\|_2^2}{\varepsilon}\right) \rho_a(x) \rho_b(y), \quad \forall x, y \in \Omega. \quad (51)$$

Such an relation can further relate α^ε and β^ε with η^ε and ψ^ε Mallasto et al. (2021)

$$\eta^\varepsilon(x) = \varepsilon \log \alpha^\varepsilon(x), \quad \forall x \in \Omega, \quad (52)$$

$$\psi^\varepsilon(y) = \varepsilon \log \beta^\varepsilon(y), \quad \forall y \in \Omega. \quad (53)$$

Given these derivations, we can insert Eq. (40) and Eqs. (52-53) into Eq. (48) to derive

$$\begin{aligned} \text{EO}_\varepsilon(\mathcal{X}_\phi^a, \mathcal{X}_\phi^b) = & \varepsilon \left(\int_\Omega \rho_a(x) \eta^\varepsilon(x) dx + \int_\Omega \rho_b(y) \psi^\varepsilon(y) dy \right) \\ & - \varepsilon \left[\int_{\Omega \times \Omega} \rho_a(x) \rho_b(y) \exp\left(\frac{\eta^\varepsilon(x) + \psi^\varepsilon(y) - \|x - y\|_2^2}{\varepsilon}\right) dx dy - 1 \right], \quad (54) \\ = & \int_\Omega \rho_a(x) \log \alpha^\varepsilon(x) dx + \int_\Omega \rho_b(y) \log \beta^\varepsilon(y) dy \\ & - \varepsilon \left[\int_{\Omega \times \Omega} \rho_a(x) \rho_b(y) \alpha^\varepsilon(x) \beta^\varepsilon(y) \exp\left(\frac{-\|x - y\|_2^2}{\varepsilon}\right) dx dy - 1 \right]. \quad (55) \end{aligned}$$

To further simplify Eq. (55), we need to relate α^ε and β^ε with variables $\mathcal{X}_\phi^a \sim \mathcal{N}(\mathbf{0}, \Sigma_a)$ and $\mathcal{X}_\phi^b \sim \mathcal{N}(\mathbf{0}, \Sigma_b)$. Following the idea of Mallasto et al. (2021), we consider to represent α^ε and β^ε as the functions of certain variables

$$\alpha^\varepsilon(x) = \exp(x^T A x + a), \quad \forall x \in \Omega, \quad (56)$$

$$\beta^\varepsilon(y) = \exp(y^T B y + b), \quad \forall y \in \Omega. \quad (57)$$

To analytically derive A , a , B , and b , we first insert Eqs. (52-53) into Eq. (51)

$$\gamma^\varepsilon(x, y) = \alpha^\varepsilon(x) \beta^\varepsilon(y) \exp\left(\frac{-\|x - y\|_2^2}{\varepsilon}\right) \rho_a(x) \rho_b(y), \quad \forall x, y \in \Omega. \quad (58)$$

Because the entropic transport plan needs to have appropriate marginals

$$\rho_a(x) = \int_{\Omega} \gamma^\varepsilon(x, y) dy, \quad \forall x \in \Omega, \quad (59)$$

$$\rho_b(y) = \int_{\Omega} \gamma^\varepsilon(x, y) dx, \quad \forall y \in \Omega, \quad (60)$$

we can have

$$\rho_a(x) = \alpha^\varepsilon(x) \rho_a(x) \int_{\Omega} \beta^\varepsilon(y) \exp\left(\frac{-\|x - y\|_2^2}{\varepsilon}\right) \rho_b(y) dy, \quad \forall x \in \Omega, \quad (61)$$

$$\rho_b(y) = \beta^\varepsilon(y) \rho_b(y) \int_{\Omega} \alpha^\varepsilon(x) \exp\left(\frac{-\|x - y\|_2^2}{\varepsilon}\right) \rho_a(x) dx, \quad \forall y \in \Omega. \quad (62)$$

Eqs. (61-62) directly lead to

$$1 = \exp(x^T A x + a) \int_{\Omega} \exp(y^T B y + b) \exp\left(\frac{-\|x - y\|_2^2}{\varepsilon}\right) \rho_b(y) dy, \quad \forall x \in \Omega, \quad (63)$$

$$1 = \exp(y^T B y + b) \int_{\Omega} \exp(x^T A x + a) \exp\left(\frac{-\|x - y\|_2^2}{\varepsilon}\right) \rho_a(x) dx, \quad \forall y \in \Omega. \quad (64)$$

SECOND STEP. Now, let us relate the above results with $\text{EO}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b)$. We constrain the entropic transport plan γ^ε as a permutation matrix $M \in \mathcal{M}$ and continuously approximate it via the *Sinkhorn operator*. Under this condition, variable \mathcal{X}_ϕ^b evolves towards \mathcal{X}_ϕ^a following

$$\varsigma(M/\tau) \circ \mathcal{X}_\phi^b \sim \mathcal{N}(\mathbf{0}, \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)) \quad (65)$$

in each step. Therefore, we need to update the probability density of \mathcal{X}_ϕ^b following Eq. (65). Based on these settings, we can represent $\text{EO}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b)$ in a similar form of Eq. (55)

$$\begin{aligned} & \text{EO}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b) \\ &= \varepsilon \left(\int_{\Omega} \rho_a(x) \eta^\varepsilon(x) dx + \int_{\Omega} \rho_b^{\varsigma(M/\tau)}(y) \psi^\varepsilon(y) dy \right) \\ & \quad - \varepsilon \left[\int_{\Omega \times \Omega} \rho_a(x) \rho_b^{\varsigma(M/\tau)}(y) \exp\left(\frac{\eta^\varepsilon(x) + \psi^\varepsilon(y) - \|x - y\|_2^2}{\varepsilon}\right) dx dy - 1 \right], \end{aligned} \quad (66)$$

$$\begin{aligned} &= \int_{\Omega} \rho_a(x) \log \alpha^\varepsilon(x) dx + \int_{\Omega} \rho_b^{\varsigma(M/\tau)}(y) \log \beta^\varepsilon(y) dy \\ & \quad - \varepsilon \left[\int_{\Omega \times \Omega} \rho_a(x) \rho_b^{\varsigma(M/\tau)}(y) \alpha^\varepsilon(x) \beta^\varepsilon(y) \exp\left(\frac{-\|x - y\|_2^2}{\varepsilon}\right) dx dy - 1 \right], \end{aligned} \quad (67)$$

$$= \varepsilon \left[\int_{\Omega} \rho_a(x^T A x + a) dx + \int_{\Omega} \rho_b^{\varsigma(M/\tau)}(y^T B y + b) dy \right], \quad (68)$$

where $\rho_b^{\varsigma(M/\tau)}$ denotes the probability density of variable $\varsigma(M/\tau) \circ \mathcal{X}_\phi^b$ in Eq. (65).

Similar to our idea shown in Eqs. (56-64), we can equivalently represent Eqs. (63-64) as

$$1 = \exp(x^T A x + a) \int_{\Omega} \exp(y^T B y + b) \exp\left(\frac{-\|x - y\|_2^2}{\varepsilon}\right) \times \frac{1}{\sqrt{(2\pi)^{|V_a|} \det(\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau))}} \exp\left(-\frac{1}{2} y^T \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) y\right) dy, \forall x \in \Omega, \quad (69)$$

$$1 = \exp(y^T B y + b) \int_{\Omega} \exp(x^T A x + a) \exp\left(\frac{-\|x - y\|_2^2}{\varepsilon}\right) \frac{1}{\sqrt{(2\pi)^{|V_a|} \det(\Sigma_a)}} \times \exp\left(-\frac{1}{2} x^T \Sigma_a x\right) dx, \forall y \in \Omega. \quad (70)$$

Please note that we have a term $|V_a|$ because the matrix size of $\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)$ is same as Σ_a (they all have a size of $|V_a|$). Following the idea of Mallasto et al. (2021), we can have Eqs. (71-72) after some simplification and reorganization of Eqs. (69-70)

$$1 = \frac{\exp(a + b)}{\sqrt{(2\pi)^{|V_a|} \det(\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau))}} \exp\left[x^T \left(A - \frac{1}{\varepsilon} I\right) x\right] \times \int_{\Omega} \exp\left[y^T \left(B - \frac{1}{\varepsilon} I - \frac{1}{2} \left(\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)\right)^{-1}\right) y + \frac{2}{\varepsilon} x^T y\right] dy, \forall x \in \Omega, \quad (71)$$

$$1 = \frac{\exp(a + b)}{\sqrt{(2\pi)^{|V_a|} \det(\Sigma_a)}} \exp\left[y^T \left(B - \frac{1}{\varepsilon} I\right) y\right] \int_{\Omega} \exp\left[x^T \left(A - \frac{1}{\varepsilon} I - \frac{1}{2} [\Sigma_a]^{-1}\right) x + \frac{2}{\varepsilon} y^T x\right] dx, \forall y \in \Omega. \quad (72)$$

Based on the identity that holds for arbitrary C Mallasto et al. (2021)

$$\int_{\Omega} \exp(-x^T C x + b^T x) = \sqrt{\frac{\pi^{|V_a|}}{\det(C)}} \exp\left(\frac{1}{4} b^T C^{-1} b\right), \forall C \in \mathbb{R}^{|V_a| \times |V_a|}, \quad (73)$$

we can transform Eqs. (71-72) into Mallasto et al. (2021)

$$A = \frac{1}{\varepsilon} I + \frac{1}{\varepsilon^2} \left(B - \frac{1}{\varepsilon} I - \frac{1}{2} \left(\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)\right)^{-1}\right)^{-1}, \quad (74)$$

$$B = \frac{1}{\varepsilon} I + \frac{1}{\varepsilon^2} \left(A - \frac{1}{\varepsilon} I - \frac{1}{2} [\Sigma_a]^{-1}\right)^{-1}, \quad (75)$$

$$\exp(a + b) = \sqrt{\det\left(2\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)\right) \det\left(\frac{1}{\varepsilon} I + \frac{1}{2} \left(\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)\right)^{-1} - B\right)}, \quad (76)$$

$$\exp(a + b) = \sqrt{\det(2\Sigma_a) \det\left(\frac{1}{\varepsilon} I + \frac{1}{2} [\Sigma_a]^{-1} - A\right)}. \quad (77)$$

Eqs. (74-75) can be further reformulated as

$$A = \frac{1}{\varepsilon} I + \frac{1}{\varepsilon^2} \left(\frac{1}{\varepsilon} I + \frac{1}{\varepsilon^2} \left(A - \frac{1}{\varepsilon} I - \frac{1}{2} [\Sigma_a]^{-1}\right)^{-1} - \frac{1}{\varepsilon} I - \frac{1}{2} \left(\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)\right)^{-1}\right)^{-1}, \quad (78)$$

$$B = \frac{1}{\varepsilon} I + \frac{1}{\varepsilon^2} \left(\frac{1}{\varepsilon} I + \frac{1}{\varepsilon^2} \left(B - \frac{1}{\varepsilon} I - \frac{1}{2} \left(\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)\right)^{-1}\right)^{-1} - \frac{1}{\varepsilon} I - \frac{1}{2} [\Sigma_a]^{-1}\right)^{-1}. \quad (79)$$

After some reorganization of Eqs. (78-79), we can eventually derive

$$A = \frac{1}{4} \sqrt{[\Sigma_a]^{-1}} \left(I + \frac{\varepsilon}{4} \Sigma_a - \sqrt{I + \frac{16}{\varepsilon^2} \sqrt{\Sigma_a} \varsigma (M/\tau)^T \Sigma_b \varsigma (M/\tau) \sqrt{\Sigma_a}} \right) \sqrt{[\Sigma_a]^{-1}}, \quad (80)$$

$$B = \frac{1}{4} \sqrt{\left(\varsigma (M/\tau)^T \Sigma_b \varsigma (M/\tau) \right)^{-1}} \left(I + \frac{\varepsilon}{4} \varsigma (M/\tau)^T \Sigma_b \varsigma (M/\tau) - \sqrt{I + \frac{16}{\varepsilon^2} \sqrt{\varsigma (M/\tau)^T \Sigma_b \varsigma (M/\tau) \Sigma_a} \sqrt{\varsigma (M/\tau)^T \Sigma_b \varsigma (M/\tau)}} \right) \sqrt{\left(\varsigma (M/\tau)^T \Sigma_b \varsigma (M/\tau) \right)^{-1}}. \quad (81)$$

Meanwhile, we can have

$$\exp(a + b) = \sqrt{\frac{1}{2|V_a|} \det(K_{ab}^\varepsilon)}. \quad (82)$$

In Mallasto et al. (2021), one can see more mathematical proofs of the above derivations.

THIRD STEP. Given the expressions of A , B , and $\exp(a + b)$, we are able to analytically derive the closed-form expression of $\text{EO}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b)$. Applying what we have derived above, we can reformulate Eq. (68) as

$$\begin{aligned} & \text{EO}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b) \\ &= \varepsilon \left(\text{tr}(\Sigma_a A) + \text{tr}(\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) B) + a + b \right), \quad (83) \\ &= \frac{\varepsilon}{4} \text{tr} \left[I + \frac{4}{\varepsilon} \Sigma_a - \sqrt{I + \frac{16}{\varepsilon^2} \sqrt{\Sigma_a} \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \sqrt{\Sigma_a}} \right] + \frac{\varepsilon}{4} \text{tr} \left[I + \frac{4}{\varepsilon} \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) - \sqrt{I + \frac{16}{\varepsilon^2} \sqrt{\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \Sigma_a} \sqrt{\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)}} \right] + \frac{\varepsilon}{2} \left(\log \frac{1}{2|V_a|} + \log \det(K_{ab}^\varepsilon) \right), \quad (84) \end{aligned}$$

$$\begin{aligned} &= \frac{\varepsilon}{4} \text{tr} \left[I + \frac{4}{\varepsilon} \Sigma_a - \sqrt{I + \frac{16}{\varepsilon^2} \Sigma_a \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)} \right] + \frac{\varepsilon}{4} \text{tr} \left[I + \frac{4}{\varepsilon} \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) - \sqrt{I + \frac{16}{\varepsilon^2} \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \Sigma_a} \right] + \frac{\varepsilon}{2} \left(\log \frac{1}{2|V_a|} + \log \det(K_{ab}^\varepsilon) \right), \quad (85) \end{aligned}$$

where Eq. (85) is derived from the fact that $\sqrt{C}D\sqrt{C}$ has the same eigenvalues trace with CD for arbitrary square matrices C and D Mallasto et al. (2021). Based on simple reorganization, we can reformulate Eq. (85) as

$$\begin{aligned} & \text{EO}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b) \\ &= \frac{\varepsilon}{4} \text{tr} \left[2I + \frac{4}{\varepsilon} \Sigma_a - K_{ab}^\varepsilon \right] + \frac{\varepsilon}{4} \text{tr} \left[2I + \frac{4}{\varepsilon} \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) - K_{ab}^\varepsilon \right] - \frac{\varepsilon|V_a|}{2} \log 2 \\ & \quad + \frac{\varepsilon}{2} \log \det(K_{ab}^\varepsilon), \quad (86) \end{aligned}$$

$$\begin{aligned} &= \frac{\varepsilon|V_a|}{2} + \text{tr}(\Sigma_a) - \frac{\varepsilon}{4} \text{tr}(K_{ab}^\varepsilon) + \frac{\varepsilon|V_a|}{2} + \text{tr}(\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)) - \frac{\varepsilon}{4} \text{tr}(K_{ab}^\varepsilon) \\ & \quad - \frac{\varepsilon|V_a|}{2} \log 2 + \frac{\varepsilon}{2} \log \det(K_{ab}^\varepsilon), \quad (87) \end{aligned}$$

$$= \text{tr}(\Sigma_a + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)) - \frac{\varepsilon}{2} (\text{tr}(K_{ab}^\varepsilon) - \log \det(K_{ab}^\varepsilon) + |V_a| \log 2 - 2|V_a|), \quad (88)$$

which is same as Eq. (44) and our results in the main text. Thus, we have finished our derivations of the closed-form expression of $\text{EO}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b)$.

D ON THE 2-SINKHORN DIVERGENCE

In this section, we present our derivations of the closed-form expression of the 2-Sinkhorn divergence in our main text (see Eq. (12)).

The closed-form expression of the 2-Sinkhorn divergence. As we have suggested in the main text, the classic entropy-regularized optimal transport problem between graphs defined by $\text{EO}_\varepsilon \left(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b \right)$ inevitably involves a bias Feydy et al. (2019). This bias is created by the non-vanishing auto-correlation terms $\text{EO}_\varepsilon \left(\mathcal{X}_\phi^a, \mathcal{X}_\phi^a \right)$ and $\text{EO}_\varepsilon \left(\varsigma(M/\tau) \circ \mathcal{X}_\phi^b, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b \right)$ when $\varepsilon > 0$ Feydy et al. (2019). Following the idea in previous studies Feydy et al. (2019); Mallasto et al. (2021), we control this bias by replacing $\text{EO}_\varepsilon \left(\mathcal{X}_\phi^a, \mathcal{X}_\phi^b \right)$ with the 2-Sinkhorn divergence

$$\begin{aligned} & \text{SK}_\varepsilon \left(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b \right) \\ &= \text{EO}_\varepsilon \left(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b \right) - \frac{1}{2} \text{EO}_\varepsilon \left(\mathcal{X}_\phi^a, \mathcal{X}_\phi^a \right) - \frac{1}{2} \text{EO}_\varepsilon \left(\varsigma(M/\tau) \circ \mathcal{X}_\phi^b, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b \right). \end{aligned} \quad (89)$$

Based on the closed-form expression of $\text{EO}_\varepsilon \left(\mathcal{X}_\phi^a, \mathcal{X}_\phi^b \right)$ in Eq. (44), we can analytically derive that

$$\text{SK}_\varepsilon \left(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b \right) = \frac{\varepsilon}{4} \left(\text{tr}(K_{aa}^\varepsilon - 2K_{ab}^\varepsilon + K_{bb}^\varepsilon) + \log \left(\frac{\det^2(K_{ab}^\varepsilon)}{\det(K_{aa}^\varepsilon) \det(K_{bb}^\varepsilon)} \right) \right), \quad (90)$$

where K_{aa}^ε , K_{ab}^ε , and K_{bb}^ε are defined in Eqs. (41-43).

The derivations of Eq. (90) is rather simple. Let us formalize terms $\text{EO}_\varepsilon \left(\mathcal{X}_\phi^a, \mathcal{X}_\phi^a \right)$ and $\text{EO}_\varepsilon \left(\varsigma(M/\tau) \circ \mathcal{X}_\phi^b, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b \right)$ in Eq. (89). Similar to Eq. (88), we can have

$$\text{EO}_\varepsilon \left(\mathcal{X}_\phi^a, \mathcal{X}_\phi^a \right) = \text{tr}(\Sigma_a + \Sigma_a) - \frac{\varepsilon}{2} \left(\text{tr}(K_{aa}^\varepsilon) - \log \det(K_{aa}^\varepsilon) + |V_a| \log 2 - 2|V_a| \right), \quad (91)$$

$$\begin{aligned} \text{EO}_\varepsilon \left(\varsigma(M/\tau) \circ \mathcal{X}_\phi^b, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b \right) &= \text{tr} \left(\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right) \\ &- \frac{\varepsilon}{2} \left(\text{tr}(K_{bb}^\varepsilon) - \log \det(K_{bb}^\varepsilon) + |V_a| \log 2 - 2|V_a| \right). \end{aligned} \quad (92)$$

Inserting Eq. (88) and Eqs. (91-92) into Eq. (89), it is trivial to verify the validity of Eq. (90).

Properties of the 2-Sinkhorn divergence. As shown in the main text, the 2-Sinkhorn divergence can take the advantages of both pure optimal transport and *maximum mean discrepancy* Feydy et al. (2019) because it interpolates between them according to entropy regularization magnitude ε Feydy et al. (2019); Mallasto et al. (2021)

$$\text{OT}_2 \left(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b \right) \xrightarrow{\varepsilon \rightarrow 0} \text{SK}_\varepsilon \left(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b \right) \xrightarrow{\varepsilon \rightarrow \infty} \text{MMD}_\varepsilon \left(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b \right), \quad (93)$$

where $\text{MMD}_\varepsilon \left(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b \right) = \|\mathbb{E} \left(\mathcal{X}_\phi^a \right) - \mathbb{E} \left(\varsigma(M/\tau) \circ \mathcal{X}_\phi^b \right)\|_2$ denotes the maximum mean discrepancy Feydy et al. (2019); Mallasto et al. (2021). Below, we present mathematical proofs of this in-between property.

At first, we prove the left part of Eq. (93) where the 2-Sinkhorn divergence reduces to the 2-Wasserstein distance as $\varepsilon \rightarrow 0$. Inserting Eqs. (41-43) into Eq. (44) and Eqs. (91-92), we can derive

the following limit

$$\begin{aligned} & \lim_{\varepsilon \rightarrow 0} \text{EO}_\varepsilon (\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b) \\ &= \lim_{\varepsilon \rightarrow 0} \left[\text{tr} \left(\Sigma_a + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right) - \frac{\varepsilon}{2} (\text{tr}(K_{ab}^\varepsilon) - \log \det(K_{ab}^\varepsilon) + |V_a| \log 2 - 2|V_a|) \right], \end{aligned} \quad (94)$$

$$\begin{aligned} &= \lim_{\varepsilon \rightarrow 0} \left\{ \text{tr} \left(\Sigma_a + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right) - \frac{\varepsilon}{2} \left[\text{tr} \left(I + \sqrt{I + \frac{16}{\varepsilon^2} \Sigma_a \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)} \right) \right. \right. \\ &\quad \left. \left. - \log \det \left(I + \sqrt{I + \frac{16}{\varepsilon^2} \Sigma_a \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)} \right) + |V_a| \log 2 - 2|V_a| \right] \right\}, \end{aligned} \quad (95)$$

$$\begin{aligned} &= \text{tr} \left(\Sigma_a + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right) - 2 \lim_{\varepsilon \rightarrow 0} \text{tr} \left(\frac{\varepsilon}{4} I + \sqrt{\frac{\varepsilon^2}{16} I + \Sigma_a \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)} \right) \\ &\quad + \lim_{\varepsilon \rightarrow 0} \frac{\varepsilon}{2} \log \det \left(\frac{\varepsilon}{4} I + \sqrt{\frac{\varepsilon^2}{16} I + \Sigma_a \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)} \right) + \lim_{\varepsilon \rightarrow 0} \frac{\varepsilon |V_a|}{2} (\log 2 - \log \varepsilon + 2), \end{aligned} \quad (96)$$

$$= \text{tr} \left(\Sigma_a + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right) - 2 \text{tr} \left(\sqrt{\Sigma_a \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)} \right), \quad (97)$$

$$= \text{OT}_2 (\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b). \quad (98)$$

Similarly, we can further generalize the derivations in Eqs. (94-98) to terms $\text{EO}_\varepsilon (\mathcal{X}_\phi^a, \mathcal{X}_\phi^a)$ and $\text{EO}_\varepsilon (\varsigma(M/\tau) \circ \mathcal{X}_\phi^b, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b)$ in Eq. (89) to obtain

$$\begin{aligned} & \lim_{\varepsilon \rightarrow 0} \text{EO}_\varepsilon (\mathcal{X}_\phi^a, \mathcal{X}_\phi^a) \\ &= \text{tr} (\Sigma_a + \Sigma_a) - 2 \text{tr} \left(\sqrt{\Sigma_a \Sigma_a} \right), \end{aligned} \quad (99)$$

$$= 0, \quad (100)$$

and

$$\begin{aligned} & \lim_{\varepsilon \rightarrow 0} \text{EO}_\varepsilon (\varsigma(M/\tau) \circ \mathcal{X}_\phi^b, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b) \\ &= \text{tr} \left(\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right) \\ &\quad - 2 \text{tr} \left(\sqrt{\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)} \right), \end{aligned} \quad (101)$$

$$= 0. \quad (102)$$

After inserting Eqs. (94-102) into Eq. (89), we can know

$$\begin{aligned} & \lim_{\varepsilon \rightarrow 0} \text{SK}_\varepsilon (\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b) \\ &= \lim_{\varepsilon \rightarrow 0} \text{EO}_\varepsilon (\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b) - \frac{1}{2} \lim_{\varepsilon \rightarrow 0} \text{EO}_\varepsilon (\mathcal{X}_\phi^a, \mathcal{X}_\phi^a) - \frac{1}{2} \lim_{\varepsilon \rightarrow 0} \text{EO}_\varepsilon (\varsigma(M/\tau) \circ \mathcal{X}_\phi^b, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b), \end{aligned} \quad (103)$$

$$= \text{OT}_2 (\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b), \quad (104)$$

which proves the left side of Eq. (93).

Then we turn to the right part of Eq. (93) where the 2-Sinkhorn divergence reduces to the maximum mean discrepancy Feydy et al. (2019); Mallasto et al. (2021) as $\varepsilon \rightarrow \infty$. Following the idea in Mallasto et al. (2021), the proof can be derived in a simple way. One only need to consider the spectral decomposition of a matrix product $\Sigma_a \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)$

$$\Sigma_a \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) = U^T \text{diag} ([\lambda_1, \dots, \lambda_{|V_a|}]) U, \quad (105)$$

where $\text{diag}(\cdot)$ denotes the diagonal. Then, we can have

$$\begin{aligned} & \lim_{\varepsilon \rightarrow \infty} \text{EO}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b) \\ &= \lim_{\varepsilon \rightarrow \infty} \left[\text{tr} \left(\Sigma_a + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right) - \frac{\varepsilon}{2} \left(\text{tr}(K_{ab}^\varepsilon) - \log \det(K_{ab}^\varepsilon) + |V_a| \log 2 - 2|V_a| \right) \right], \end{aligned} \quad (106)$$

$$\begin{aligned} &= \lim_{\varepsilon \rightarrow \infty} \left\{ \text{tr} \left(\Sigma_a + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right) - \frac{\varepsilon}{2} \sum_{i=1}^{|V_a|} \left[-1 + \sqrt{1 + \frac{16}{\varepsilon^2} \lambda_i} \right. \right. \\ &\quad \left. \left. - \log \left(\frac{1}{2} \left(1 + \sqrt{1 + \frac{16}{\varepsilon^2} \lambda_i} \right) \right) \right] \right\}, \end{aligned} \quad (107)$$

$$= \lim_{\varepsilon \rightarrow \infty} \left\{ \text{tr} \left(\Sigma_a + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right) - \frac{\varepsilon}{2} \sum_{i=1}^{|V_a|} \left[-1 + \left(1 + \frac{8}{\varepsilon^2} \right) - \log \left(1 + \frac{4}{\varepsilon^2} \right) \right] \right\}, \quad (108)$$

$$= \lim_{\varepsilon \rightarrow \infty} \left\{ \text{tr} \left(\Sigma_a + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right) - \frac{\varepsilon}{2} \sum_{i=1}^{|V_a|} \left[\frac{8}{\varepsilon^2} - \log \left(1 + \frac{4}{\varepsilon^2} \right) \right] \right\}, \quad (109)$$

$$= \text{tr} \left(\Sigma_a + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right) - \lim_{\varepsilon \rightarrow \infty} \sum_{i=1}^{|V_a|} \left(\frac{4}{\varepsilon} - \frac{\varepsilon}{2} \frac{4}{\varepsilon^2} \right), \quad (110)$$

$$= \text{tr} \left(\Sigma_a + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right) - \lim_{\varepsilon \rightarrow \infty} \sum_{i=1}^{|V_a|} \frac{2}{\varepsilon}, \quad (111)$$

$$= \text{tr} \left(\Sigma_a + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right). \quad (112)$$

Another kind of derivation of Eq. (112) can be seen in Mallasto et al. (2021). After repeating the above derivations on terms $\text{EO}_\varepsilon(\mathcal{X}_\phi^a, \mathcal{X}_\phi^a)$ and $\text{EO}_\varepsilon(\varsigma(M/\tau) \circ \mathcal{X}_\phi^b, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b)$ in Eq. (89), we can readily prove

$$\lim_{\varepsilon \rightarrow \infty} \text{EO}_\varepsilon(\mathcal{X}_\phi^a, \mathcal{X}_\phi^a) = \text{tr}(\Sigma_a + \Sigma_a), \quad (113)$$

and

$$\begin{aligned} & \lim_{\varepsilon \rightarrow \infty} \text{EO}_\varepsilon(\varsigma(M/\tau) \circ \mathcal{X}_\phi^b, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b) \\ &= \text{tr} \left(\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) + \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau) \right). \end{aligned} \quad (114)$$

Inserting Eqs. (112-114) into Eq. (89), we can know

$$\lim_{\varepsilon \rightarrow \infty} \text{SK}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b) = 0, \quad (115)$$

which coincides with the maximum mean discrepancy Feydy et al. (2019); Mallasto et al. (2021)

$$\text{MMD}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b) = \|\mathbb{E}(\mathcal{X}_\phi^a) - \mathbb{E}(\varsigma(M/\tau) \circ \mathcal{X}_\phi^b)\|_2 = 0. \quad (116)$$

This result proves the right side of Eq. (93). In fact, what Eq. (116) presents is a special case of the maximum mean discrepancy between Gaussian variables with a same value of the first-order moment Feydy et al. (2019); Mallasto et al. (2021).

E OPTIMIZATION ALGORITHM OF ERGOT PROBLEM WITH THE 2-SINKHORN DIVERGENCE

To improve the comparability of our approach with existing frameworks, we solve the ErGOT problem with the 2-Sinkhorn divergence in Eq. (14) applying the algorithm proposed in GOT Petric Maretic et al. (2019). Such an algorithm approximates the solution of Eq. (14) by the Bayesian exploration and re-parameterization Kingma & Welling (2013); Figurnov et al. (2018). Below, we

Algorithm 1 A stochastic gradient descent algorithm for solving Eq. (123)**Require:** \mathcal{X}_ϕ^a and \mathcal{X}_ϕ^b **Require:** Sampling size $s \in \mathbb{N}^+$, learning rate $\gamma \in (0, \infty)$, parameter $\tau \in (0, \infty)$ for the *Sinkhorn operator*, parameter $\varepsilon \in (0, \infty)$ for entropy regularization, and epoch number $\kappa \in \mathbb{N}^+$ **Require:** Randomly initialized Ω_0 and Σ_0 **while** $t < \kappa$ **do**Randomly sample a set $\{\Xi_t^1, \dots, \Xi_t^s\}$ where each $\Xi_t^i \sim \mathcal{N}(\mathbf{0}, \text{diag}(\mathbf{1}))$

Approximate the gradient of loss function via SGD

$$\begin{aligned} & \nabla \mathbb{E}_{\Xi_t \sim \mathcal{N}(\mathbf{0}, \text{diag}(\mathbf{1}))} (\text{SK}_\varepsilon(\mathcal{X}_\phi^a, (\Omega_t + \Sigma_t \odot \Xi_t) \circ \mathcal{X}_\phi^b)) \\ & \simeq \frac{1}{s} \sum_{i=1}^s \nabla \text{SK}_\varepsilon(\mathcal{X}_\phi^a, (\Omega_t^s + \Sigma_t^s \odot \Xi_t^s) \circ \mathcal{X}_\phi^b), \end{aligned} \quad (117)$$

Define $(\Omega_{t+1}, \Sigma_{t+1})$ by updating (Ω_t, Σ_t) following the approximated gradient in Eq. (117).**end while**

sketch this algorithm with reasonable simplification. More details about this algorithm can be seen in Petric Maretic et al. (2019); Kingma & Welling (2013); Figurnov et al. (2018).

To efficiently solve the optimization problem in our main text

$$\underset{\varsigma(M/\tau) \in \mathbb{R}^{|V_b| \times |V_a|}}{\text{minimize}} \quad \text{SK}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b), \quad \text{s.t. } \varsigma(M/\tau) \text{ is a doubly stochastic matrix}, \quad (118)$$

we can consider an approximation where the algorithm minimizes the expectation of optimization objective with respect to some parameters Θ of a certain probability distribution ρ_Θ

$$\underset{\Theta}{\text{minimize}} \quad \mathbb{E}_{\varsigma(M/\tau) \sim \rho_\Theta} (\text{SK}_\varepsilon(\mathcal{X}_\phi^a, \varsigma(M/\tau) \circ \mathcal{X}_\phi^b)), \quad \text{s.t. } \varsigma(M/\tau) \text{ is a doubly stochastic matrix}. \quad (119)$$

In practice, a common choice of probability distribution ρ_Θ is a multivariate normal distribution Petric Maretic et al. (2019)

$$\rho_\Theta = \prod_{i,j} \mathcal{N}(\omega_{ij}, \sigma_{ij}^2), \quad (120)$$

where $\Theta = (\Omega, \Sigma) \in \mathbb{R}^{|V_a| \times |V_a|} \times \mathbb{R}^{|V_a| \times |V_a|}$. As suggested in Petric Maretic et al. (2019), we can know the following equivalence relation after re-parameterization Kingma & Welling (2013); Figurnov et al. (2018)

$$\varsigma(M/\tau) \sim \rho_\Theta \iff ([\varsigma(M/\tau)]_{i,j} = \omega_{ij} + \sigma_{ij} \xi_{ij}, \quad \xi_{ij} \sim \mathcal{N}(0, 1), \quad \forall (i, j) \in \{1, \dots, |V_a|\}^2). \quad (121)$$

Based on these definitions, we can reformulate Eq. (119) as

$$\underset{\Omega, \Sigma}{\text{minimize}} \quad \mathbb{E}_{\Xi \sim \mathcal{N}(\mathbf{0}, \text{diag}(\mathbf{1}))} (\text{SK}_\varepsilon(\mathcal{X}_\phi^a, (\Omega + \Sigma \odot \Xi) \circ \mathcal{X}_\phi^b)), \quad (122)$$

where notion \odot denotes the Hadamard product and $\mathcal{N}(\mathbf{0}, \text{diag}(\mathbf{1}))$ is the standard multivariate Gaussian distribution. Then, we can approximate the gradient of the above stochastic function by sampling from the standard multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \text{diag}(\mathbf{1}))$

$$\begin{aligned} & \nabla \mathbb{E}_{\Xi \sim \mathcal{N}(\mathbf{0}, \text{diag}(\mathbf{1}))} (\text{SK}_\varepsilon(\mathcal{X}_\phi^a, (\Omega + \Sigma \odot \Xi) \circ \mathcal{X}_\phi^b)) \\ & \simeq \sum_{\Xi \sim \mathcal{N}(\mathbf{0}, \text{diag}(\mathbf{1}))} \nabla \text{SK}_\varepsilon(\mathcal{X}_\phi^a, (\Omega + \Sigma \odot \Xi) \circ \mathcal{X}_\phi^b), \end{aligned} \quad (123)$$

which can be naturally solved by stochastic gradient descent Khan et al. (2017); Petric Maretic et al. (2019). In Algorithm 1, we present necessary details of algorithm designs. One can also see Petric Maretic et al. (2019) for more information.

In experiment implementation, we also add a numerical acceleration to the algorithm. The acceleration is realized by calculating the 2-Sinkhorn divergence $SK_\varepsilon(\cdot, \cdot)$ using

$$\hat{K}_{ab}^\varepsilon = (K_{ab}^\varepsilon - I)^2 + I = 2I + \frac{16}{\varepsilon^2} \Sigma_a \varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau), \quad (124)$$

$$\hat{K}_{aa}^\varepsilon = (K_{aa}^\varepsilon - I)^2 + I = 2I + \frac{16}{\varepsilon^2} [\Sigma_a]^2, \quad (125)$$

$$\hat{K}_b^\varepsilon = (K_{bb}^\varepsilon - I)^2 + I = 2I + \frac{16}{\varepsilon^2} [\varsigma(M/\tau)^T \Sigma_b \varsigma(M/\tau)]^2, \quad (126)$$

rather than using K_{ab}^ε , K_{aa}^ε , and K_{bb}^ε in Eqs. (41-43). Although we have no strict proof here, the 2-Sinkhorn divergence calculated based on \hat{K}_{ab}^ε , \hat{K}_{aa}^ε , and \hat{K}_{bb}^ε is empirically observed to achieve optimal performance. Because excluding the computation of matrix square root is favorable for algorithm acceleration, we suggest that a strict proof of the validity of Eqs. (124-126) may be a meaningful direction for future exploration.

F EXPERIMENT SETTINGS AND ADDITIONAL RESULTS

Graph alignment. In Fig. 6, we show the experiment results of graph alignment without numerical tricks. All experimental settings are same as Fig. 5 except that we do not add $0.1I$, a scaled unit matrix, to the covariance matrix. The experiment of sample complexity is no longer repeated

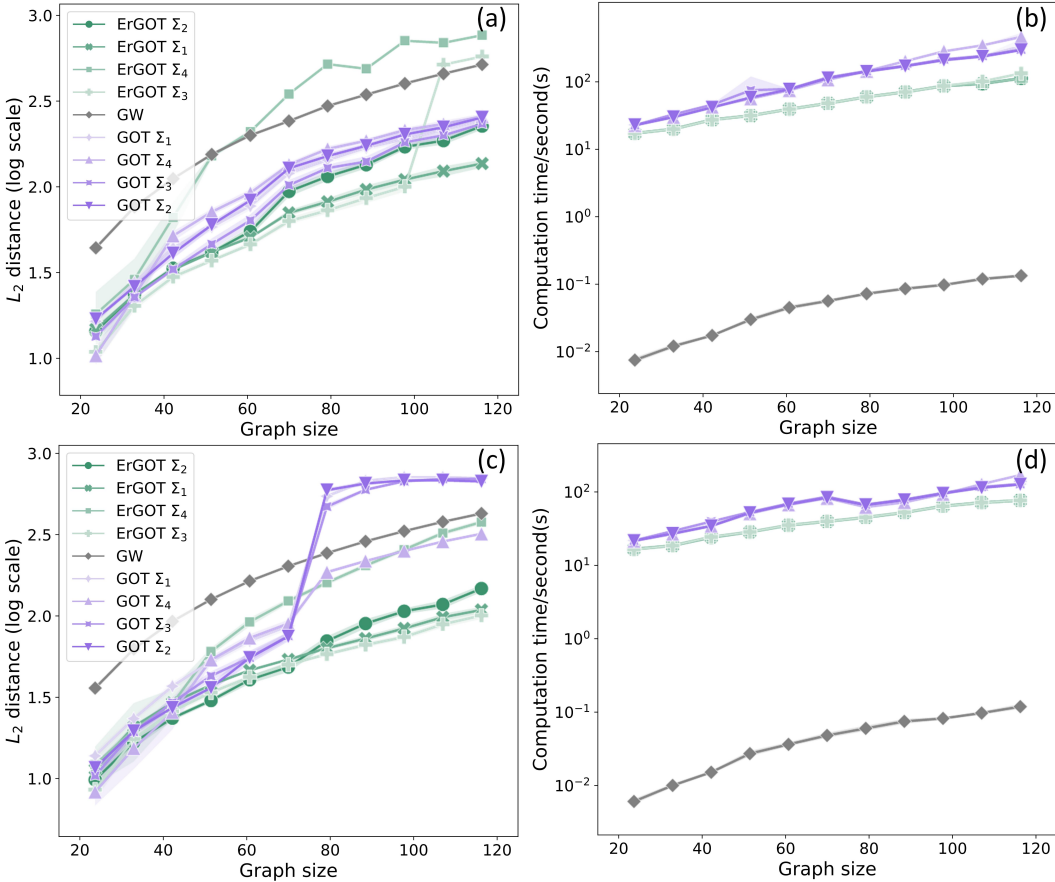


Figure 6: The results of graph alignment (without numerical tricks). (a-b) report the experiment on Erdős-Rényi graphs while (c-d) report the experiment on stochastic block models. (a) and (c) compare ErGOT with GOT and GW on different graph sizes. (b) and (d) measure the time cost of ErGOT (1000 epochs), GOT (1000 epochs), and GW.

because numerical tricks have no influence on it. In **Fig. 6**, we can see the data trends similar to **Fig. 5**. Although the performance of ErGOT and GOT is affected by the absence of numerical tricks, ErGOT still surpasses GOT because ErGOT is more robust in averting numerical problems.

Graph sketching. In **Table 4**, we show the experiment results of graph sketching on three data sets. All experimental settings are same as **Table 2**.

	2X compression			4X compression		
	BZR	MUTAG	Synthie	BZR	MUTAG	Synthie
OTC	79.38 \pm 5.01	81.66 \pm 2.98	50.36 \pm 5.47	-	-	30.29 \pm 4.74
HeavyE	79.23 \pm 6.06	79.7 \pm 6.3	51.60 \pm 10.88	-	-	-
Variation	-	77.27 \pm 3.57	41.93 \pm 3.54	77.02 \pm 2.09	-	-
Algebraic	81.40 \pm 3.31	68.00 \pm 1.85	43.85 \pm 5.64	-	-	-
Affinity	78.65 \pm 4.84	69.45 \pm 3.69	47.84 \pm 2.18	-	-	-
REC	77.08 \pm 3.53	70.00 \pm 5.24	49.32 \pm 7.07	-	-	-
COPT	74.72 \pm 0.69	71.79 \pm 8.10	30.29 \pm .16	78.86 \pm 4.21	84.34 \pm 4.41	35.02 \pm 8.25
ErGOT	78.45 \pm 3.87	73.13 \pm 6.93	42.99 \pm 1.33	77.22 \pm 2.20	73.34 \pm 3.28	46.31 \pm 10.63

Table 4: Graph sketching experiment results on 3 data sets.

We should note that there are multiple missing values in **Table 4** because some algorithms, irrespective of how much effort we devote to fine-tuning them, achieve unreasonably poor performance or meet numerical problems (e.g., return inf or nan during computation) in our experiment. We suggest that these algorithms may be numerically unstable on some data sets. It would be unfair to treat the observed poor performance as their actual capacities. Therefore, we no longer report these results to avoid misleading.

Graph retrieval. In our graph retrieval experiment, graphs in every data set is first compressed to a given graph size. Below, we summarize the settings of graph size:

	MSRC_9	PROTEINS	BZR	MUTAG	MSRC_21C	DHFR	COX2_MD
Compressed size	20	13	7	7	20	30	13

Table 5: Compressed graph size on 7 data sets.