A Supplementary Results and Comments

A.1 Proof that eqs. (4) and (5) are equivalent when $N_x = N_y = N$

This is the result of a straightforward calculation, exploiting several elementary facts from linear algebra. First, for any matrix $\boldsymbol{A}$ we have that $\|\boldsymbol{A}\|_F^2 = \mathrm{Tr}[\boldsymbol{A}^\top \boldsymbol{A}]$. Second, for any matrices $\{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}\}$ with appropriate dimensions such that the product $\boldsymbol{ABC}$ is defined, we have that $\mathrm{Tr}[\boldsymbol{ABC}] = \mathrm{Tr}[\boldsymbol{CAB}] = \mathrm{Tr}[\boldsymbol{BCA}]$, which is called the *cyclic trace property*. Finally, for any orthogonal matrix $\boldsymbol{Q} \in \mathcal{O}(N)$ we have $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}$.

With these ingredients we can manipulate the squared Procrustes distance as follows:

$$
\begin{aligned}
d_{\mathcal{O}}^2(\boldsymbol{X}, \boldsymbol{Y}) &= \min_{\boldsymbol{Q} \in \mathcal{O}(N)} \|\boldsymbol{X} - \boldsymbol{Y}\boldsymbol{Q}\|_F^2 \\
&= \min_{\boldsymbol{Q} \in \mathcal{O}(N)} \mathrm{Tr}[\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{Q}^\top \boldsymbol{Y}^\top \boldsymbol{Y}\boldsymbol{Q} - 2\boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{Q}] && (\|\boldsymbol{A}\|_F^2 = \mathrm{Tr}[\boldsymbol{A}^\top \boldsymbol{A}]) \\
&= \min_{\boldsymbol{Q} \in \mathcal{O}(N)} \mathrm{Tr}[\boldsymbol{X}^\top \boldsymbol{X}] + \mathrm{Tr}[\boldsymbol{Q}^\top \boldsymbol{Y}^\top \boldsymbol{Y}\boldsymbol{Q}] - 2\,\mathrm{Tr}[\boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{Q}] && (\mathrm{Tr}[\cdot] \text{ is linear}) \\
&= \min_{\boldsymbol{Q} \in \mathcal{O}(N)} \mathrm{Tr}[\boldsymbol{X}^\top \boldsymbol{X}] + \mathrm{Tr}[\boldsymbol{Y}^\top \boldsymbol{Y}\boldsymbol{Q}\boldsymbol{Q}^\top] - 2\,\mathrm{Tr}[\boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{Q}] && (\text{cyclic trace property}) \\
&= \min_{\boldsymbol{Q} \in \mathcal{O}(N)} \mathrm{Tr}[\boldsymbol{X}^\top \boldsymbol{X}] + \mathrm{Tr}[\boldsymbol{Y}^\top \boldsymbol{Y}] - 2\,\mathrm{Tr}[\boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{Q}] && (\text{orthogonality}) \\
&= \mathrm{Tr}[\boldsymbol{X}^\top \boldsymbol{X}] + \mathrm{Tr}[\boldsymbol{Y}^\top \boldsymbol{Y}] - 2 \max_{\boldsymbol{Q} \in \mathcal{O}(N)} \mathrm{Tr}[\boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{Q}] && (\text{first two terms are constant}) \\
&= \mathrm{Tr}[\boldsymbol{X}^\top \boldsymbol{X}] + \mathrm{Tr}[\boldsymbol{Y}^\top \boldsymbol{Y}] - 2\|\boldsymbol{X}^\top \boldsymbol{Y}\|_* && (\text{discussed below})
\end{aligned}
$$

as claimed in the main text. The final step is the comes from the celebrated closed form solution to the orthogonal Procrustes problem, which is comprehensively reviewed by Gower and Dijksterhuis [20]. Briefly, the result can be understood by considering the singular value decomposition $\boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$. Then, due to the cyclic trace property,

$$
\mathrm{Tr}[\boldsymbol{X}^\top \boldsymbol{Y}\boldsymbol{Q}] = \mathrm{Tr}[\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top \boldsymbol{Q}] = \mathrm{Tr}[\boldsymbol{S}\boldsymbol{V}^\top \boldsymbol{Q}\boldsymbol{U}] \tag{16}
$$

This final expression is maximized by setting $\boldsymbol{Q} = \boldsymbol{V}\boldsymbol{U}^\top$, resulting in:

$$
\mathrm{Tr}[\boldsymbol{S}\boldsymbol{V}^\top \boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{U}] = \mathrm{Tr}[\boldsymbol{S}] = \|\boldsymbol{X}^\top \boldsymbol{Y}\|_* \tag{17}
$$

Since the sum of the diagonal elements of $\boldsymbol{S}$ is simply the sum of the singular values of $\boldsymbol{X}^\top \boldsymbol{Y}$ (i.e. equal to the nuclear norm of this matrix).

A.2 Proof that eqs. (9) and (10) are equivalent

Recall that we are in the setting where $N_x = N_y = N$. First, for any matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ with columns $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N\}$ we have $\|\boldsymbol{A}\|_F^2 = \sum_{i=1}^N \|\boldsymbol{a}_i\|^2$. Since $\sum_j \boldsymbol{P}_{ij}\boldsymbol{y}_j$ gives column i of the matrix product $\boldsymbol{Y}\boldsymbol{P}$, we have:

$$
\|\boldsymbol{X} - \boldsymbol{Y}\boldsymbol{P}\|_F^2 = \sum_{i=1}^N \|\boldsymbol{x}_i - \sum_{j=1}^N \boldsymbol{P}_{ij}\boldsymbol{y}_j\|^2 \tag{18}
$$

Recall that $\boldsymbol{P}$ is a permutation matrix in the present context. Thus, let $\sigma(i) \in \{1, \ldots, N\}$ denote the index of the unique nonzero element of row i in $\boldsymbol{P}$. Intuitively, $\sigma(i)$ defines the permutation in which column i of $\boldsymbol{X}$ is matched to column $\sigma(i)$ of $\boldsymbol{Y}$. With this notation, we can re-write the expression above:

$$
\sum_{i=1}^N \|\boldsymbol{x}_i - \sum_{j=1}^N \boldsymbol{P}_{ij}\boldsymbol{y}_j\|^2 = \sum_{i=1}^N \|\boldsymbol{x}_i - \boldsymbol{y}_{\sigma(i)}\|^2 \tag{19}
$$

Now define $\delta[i, j]$ as a function that takes in two integers and evaluates to one if $i = j$ and evaluates to zero if $i \neq j$. (This is often called the Kronecker delta function.) We can write:

$$
\sum_{i=1}^N \|\boldsymbol{x}_i - \boldsymbol{y}_{\sigma(i)}\|^2 = \sum_{i=1}^N \sum_{j=1}^N \delta[\sigma(i), j] \cdot \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2 \tag{20}
$$

since the inner sum will evaluate to zero whenever $j \neq \sigma(i)$. Finally, we argue that $\boldsymbol{P}_{ij} = \delta[\sigma(i), j]$. Indeed, $\sum_j \delta[\sigma(i), j]\boldsymbol{y}_j = \boldsymbol{y}_\sigma(i)$ which agrees with $\sum_j \boldsymbol{P}_{ij}\boldsymbol{y}_j = \boldsymbol{y}_\sigma(i)$. Thus, we have shown that:

$$d^2_{\mathcal{P}}(\boldsymbol{X}, \boldsymbol{Y}) = \min_{\boldsymbol{P} \in \mathcal{P}(N)} \|\boldsymbol{X} - \boldsymbol{Y}\boldsymbol{P}\|_F^2 = \min_{\boldsymbol{P} \in \mathcal{P}(N)} \sum_{i,j} \boldsymbol{P}_{ij} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2^2 \tag{21}$$

To arrive at eq. (10), we need to show that we can relax the constraint of the minimization over the permutation group to over the Birkhoff polytope—i.e. to show that:

$$\min_{\boldsymbol{P} \in \mathcal{P}(N)} \sum_{i,j} \boldsymbol{P}_{ij} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2^2 = \min_{\boldsymbol{P} \in \mathcal{B}(N)} \sum_{i,j} \boldsymbol{P}_{ij} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2^2 \tag{22}$$

Here we evoke two well-known results. First, the celebrated Birkhoff–von Neumann theorem states that the vertices of $\mathcal{B}(N)$ are one-to-one matched with the permutation matrices $\mathcal{P}(N)$. Second, the final expression is a linear program since the objective function is linear in $\boldsymbol{P}$ and the constraints are linear (as can be verified from eq. 2). Thus, we evoke a basic fact from the theory of linear programming (see e.g. [5]), which states that, assuming that a finite solution exists, at least one vertex of the feasible set is a solution. Any such vertex is called a "basic feasible solution" and the fact that such solutions exist motivates the well known simplex algorithm for linear programming. Thus, we conclude that relaxing the constraints from $\boldsymbol{P} \in \mathcal{P}(N)$ to $\boldsymbol{P} \in \mathcal{B}(N)$ does not allow us to further minimize the objective, and so eq. (22) is valid. Taking square roots on both sides of eq. (22) proves the result claimed in the main text.

A.3 Relation between soft matching distance and correlation score

Here we show that the optimal soft permutation matrix $\boldsymbol{P} \in \mathcal{T}(N_x, N_y)$ that minimizes the expression in eq. 11 equals the one which maximizes the expression in eq. 12. First, beginning with the minimization problem in eq. 11, we can break the expression into three terms:

$$\begin{aligned}
&\operatorname*{argmin}_{\boldsymbol{P} \in \mathcal{T}(N_x, N_y)} \sum_{ij} \boldsymbol{P}_{ij} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2 && (23)\\
&= \operatorname*{argmin}_{\boldsymbol{P} \in \mathcal{T}(N_x, N_y)} \sum_{ij} \boldsymbol{P}_{ij} (\boldsymbol{x}_i^\top \boldsymbol{x}_i + \boldsymbol{y}_j^\top \boldsymbol{y}_j - 2\boldsymbol{x}_i^\top \boldsymbol{y}_j) && (24)\\
&= \operatorname*{argmin}_{\boldsymbol{P} \in \mathcal{T}(N_x, N_y)} \underbrace{\sum_{ij} \boldsymbol{P}_{ij} \boldsymbol{x}_i^\top \boldsymbol{x}_i}_{\text{(A)}} + \underbrace{\sum_{ij} \boldsymbol{P}_{ij} \boldsymbol{y}_j^\top \boldsymbol{y}_j}_{\text{(B)}} - 2 \underbrace{\sum_{ij} \boldsymbol{P}_{ij} \boldsymbol{x}_i^\top \boldsymbol{y}_j}_{\text{(C)}} && (25)
\end{aligned}$$

Considering term (A) first, we argue that this term is constant with respect to any feasible $\boldsymbol{P}$ since:

$$\sum_{ij} \boldsymbol{P}_{ij} \boldsymbol{x}_i^\top \boldsymbol{x}_i = \sum_{i=1}^{N_x} \left(\sum_{j=1}^{N_y} \boldsymbol{P}_{ij} \right) \boldsymbol{x}_i^\top \boldsymbol{x}_i = \sum_{i=1}^{N_x} \left(\frac{1}{N_x} \right) \boldsymbol{x}_i^\top \boldsymbol{x}_i \tag{26}$$

where the final equality follows from definition of the transportation polytope in eq. (3)—namely, the rows of $\boldsymbol{P}$ each sum to $1/N_x$. We then can make a similar argument for term (B). In particular, since the columns of $\boldsymbol{P}$ each sum to $1/N_y$, we have:

$$\sum_{ij} \boldsymbol{P}_{ij} \boldsymbol{y}_j^\top \boldsymbol{y}_j = \sum_{j=1}^{N_y} \left(\sum_{i=1}^{N_x} \boldsymbol{P}_{ij} \right) \boldsymbol{y}_j^\top \boldsymbol{y}_j = \sum_{j=1}^{N_y} \left(\frac{1}{N_y} \right) \boldsymbol{y}_j^\top \boldsymbol{y}_j \tag{27}$$

In summary, we see that only term (C) is non-constant. That is, we have

$$\begin{aligned}
\operatorname*{argmin}_{\boldsymbol{P} \in \mathcal{T}(N_x, N_y)} \sum_{ij} \boldsymbol{P}_{ij} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2 &= \operatorname*{argmin}_{\boldsymbol{P} \in \mathcal{T}(N_x, N_y)} -2 \sum_{ij} \boldsymbol{P}_{ij} \boldsymbol{x}_i^\top \boldsymbol{y}_j + const. && (28)\\
&= \operatorname*{argmax}_{\boldsymbol{P} \in \mathcal{T}(N_x, N_y)} \sum_{ij} \boldsymbol{P}_{ij} \boldsymbol{x}_i^\top \boldsymbol{y}_j && (29)
\end{aligned}$$

as we have claimed.

A.4 Computational complexity

Computing the soft-matching distance requires solving an optimal transport problem in the discrete setting. The solution to the transportation problem, which is a linear program, can be derived using the network simplex algorithm. With efficient implementations of the simplex algorithm as in the Python Optimal Transport Library, the complexity of solving the linear program is $\mathrm{O}(n^3 \log n)$, assuming that the two representations being compared have n units. Such efficient implementations enable broad application of optimal transport-based solutions in real-world settings.

A.5 Relevance of the Soft-Matching Metric to Disentangled Representation Learning Metrics

It is worth noting that the proposed soft-matching metric can also serve as a valuable tool in the field of disentangled representation learning (DRL) [4] due to its sensitivity to the representational basis. In DRL, the objective is to learn a model that effectively disentangles and makes the underlying generative factors of the data explicit in representational form (i.e. aligned with representational units). Within the DRL literature, various measures have been developed to quantify the alignment between learned representations and ground truth generative factors. Typically, the desiderata involve a combination of different criteria, encompassing the similarity in information content (explicitness) and the degree of one-to-one correspondence between the representational units and generative factors (modularity and compactness) [15]. The soft-matching distance metric offers a unique advantage by simultaneously capturing sensitivity to both of these critical properties.