Causal Optimal Transport of Abstractions

Yorgos Felekis   YORGOS.FELEKIS@WARWICK.AC.UK
University of Warwick

Fabio Massimo Zennaro   FABIO.ZENNARO@WARWICK.AC.UK
University of Warwick

Nicola Branchini   N.BRANCHINI@SMS.ED.AC.UK
University of Edinburgh

Theodoros Damoulas   T.DAMOULAS@WARWICK.AC.UK
University of Warwick

Editors: Francesco Locatello and Vanessa Didelez

Abstract

Causal abstraction (CA) theory establishes formal criteria for relating multiple structural causal models (SCMs) at different levels of granularity by defining maps between them. These maps have significant relevance for real-world challenges such as synthesizing causal evidence from multiple experimental environments, learning causally consistent representations at different resolutions, and linking interventions across multiple SCMs. In this work, we propose COTA, the first method to learn abstraction maps from observational and interventional data without assuming complete knowledge of the underlying SCMs. In particular, we introduce a multi-marginal Optimal Transport (OT) formulation that enforces *do-calculus* causal constraints, together with a cost function that relies on interventional information. We extensively evaluate COTA on synthetic and real world problems, and showcase its advantages over non-causal, independent and aggregated OT formulations. Finally, we demonstrate the efficiency of our method as a data augmentation tool by comparing it against the state-of-the-art CA learning framework, which assumes fully specified SCMs, on a real-world downstream task.

Keywords: structural causal models, causal abstractions, causal abstraction learning, causal optimal transport, multi-marginal optimal transport

1. Introduction

Learning relations between models and underlying representations at different levels of granularity is a key challenge across sub-fields of AI as it can enable aggregation of information, transfer learning, emulation via surrogate models, and multi-scale estimation and reasoning e.g. (Weinan, 2011; Somnath et al., 2021; Geiger et al., 2021). Rigorous relationships of abstraction between such models would enable utilising seemingly incompatible data, leading to improved inferences via evidence synthesis and cost savings by minimising the need for extensive data collection.

In causality the notion of abstraction is fundamental for causal representation learning, where causal variables might be abstractions of underlying quantities or when relations are sought between micro- and macro-level models of the same underlying process (Schölkopf et al., 2021; Chalupka et al., 2017). Relations between causal models and estimands across multiple environments have been studied under transportability (Pearl and Bareinboim, 2011) and multi-environment causal analysis (Peters et al., 2016; Yin et al., 2021). A theory of causal abstraction has been formalised (Rubenstein et al., 2017; Beckers and Halpern, 2019; Rischel, 2020; Massidda et al., 2022) through

the definition of a map relating two causal models representing the same system in different levels of detail and a measure of interventional consistency evaluating the discrepancy between the two under interventions (Beckers et al., 2020; Rischel, 2020). This framework has been used in the field of explainability (Geiger et al., 2021), where, given an abstraction, a neural network is trained to behave consistently with an abstracted model. While limited work exists on abstraction learning, (Zennaro et al., 2023) proposed a differentiable programming solution to learn an abstraction between two causal models in the α-abstraction framework of Rischel (2020), but with the strong assumption that the underlying causal models were fully specified. In this work, we lift this restrictive assumption of complete SCM knowledge and study the more realistic setting in which the information available to the modeler is the graph underlying the causal model together with observational and interventional data. We make the following contributions:

- We formalise the problem of learning causal abstractions (CA) from observational and interventional data in the (τ, ω)-framework (Rubenstein et al., 2017).

- We introduce the first method to address this problem without assuming full knowledge of the underlying causal models and showcase its superiority against our multiple developed baselines and prior work (Zennaro et al., 2023) of alternative CA learning frameworks that also assumes fully specified SCMs.

- To do so, we develop a causal Optimal Transport (OT) formulation for abstraction learning, named COTA, where observational and interventional distributions of the base and abstracted models act as marginals in a Kantorovich (1942) joint OT problem with multiple transport plans. Further, we prove the joint convexity of COTA in the induced plans, guaranteeing an optimal solution to the optimization problem.

- We incorporate causal knowledge to the optimisation problem by introducing *do-calculus* constraints and a causally informed cost function. We demonstrate that this enables us to learn better abstraction maps compared to non-causal or independent solutions and also makes COTA a potent data augmentation tool.

2. Background on causality, abstractions, and optimal transport

In this section we introduce basic definitions from the field of causality, causal abstractions, and optimal transport. We use the following standard notation to formalise causal models: boldface capital $\mathbf{X}$ denotes a set of random variables, and capital letter X_i denotes the i-th random variable in $\mathbf{X}$; boldface small $\mathbf{x}$ denotes a set of values realising $\mathbf{X}$; x_i denotes the i-th value in $\mathbf{x}$. We use boldface $\mathbb{P}$ to refer to the underlying probability measures.

2.1. Causality

Definition [SCM (Pearl, 2009)] *A structural causal model $\mathcal{M}$ is a tuple $\langle \mathbf{X}, \mathbf{U}, \mathcal{F}, \mathbb{P}(\mathbf{U}) \rangle$, where $\mathbf{X}$ is a set of N endogenous random variables, each one with domain $\mathrm{dom}[X_i]$, $1 \leq i \leq N$; $\mathbf{U}$ is a set of exogenous random variables each associated with an endogenous variable; $\mathcal{F}$ is a set of structural functions, one for each endogenous variable $X_i \in \mathbf{X}$ defined as $f_i : \mathrm{dom}[\mathrm{PA}(X_i)] \times \mathrm{dom}[U_i] \to \mathrm{dom}[X_i]$ where $\mathrm{PA}(X_i) \subseteq \mathbf{X} \setminus X_i$; $\mathbb{P}(\mathbf{U}) = \prod_{i=1}^{N} \mathbb{P}(U_i)$ is a joint probability distribution over $\mathbf{U}$.*

We make a few standard assumptions on our SCMs. We assume *acyclicity*, implying that the SCM $\mathcal{M}$ entails a directed acyclic graph (DAG) $\mathcal{G}_{\mathcal{M}}$ where nodes correspond to the endogenous variables

$\mathbf{X}$ and edges are defined by the signature of the functions in $\mathcal{F}$ (Peters et al., 2017). [1] We will also assume *faithfulness*, guaranteeing that independencies in the data are captured in the graphical model, and *causal sufficiency*, meaning that there are no unobserved confounders (Spirtes et al., 2000).

Definition *[Interventions (Pearl, 2009)] Given a SCM $\mathcal{M}$, an (exact) intervention $\iota = \mathrm{do}(\mathbf{A} = \mathbf{a})$, where for each endogenous variable $A_i \in \mathbf{A} \subseteq \mathbf{X}$ we have a value $a_i \in \mathbf{a}$ and $a_i \in \mathrm{dom}[A_i]$, is an operator that replaces each function f_i associated with the variable A_i with the constant a_i.*

Graphically, the intervention $\iota = \mathrm{do}(\mathbf{A} = \mathbf{a})$ mutilates the induced graph $\mathcal{G}_{\mathcal{M}}$ by removing the incoming arrows in each node A_i and replacing f_i with the constant a_i. In this way, an intervention defines a new *post-intervention* SCM $\mathcal{M}_\iota$ described by the probability distribution $\mathbb{P}_{\mathcal{M}_\iota}(\mathbf{X})$. Whenever clear from the context, we shorthand $\mathrm{do}(\mathbf{A} = \mathbf{a})$ to $\mathrm{do}(\mathbf{a})$.

Also, sets of interventions are equipped with a natural partially-ordered set (poset) structure[2] with respect to containment: given $\iota = \mathrm{do}(\mathbf{a})$ and $\eta = \mathrm{do}(\mathbf{b})$, $\iota \preceq \eta$ iff $\mathbf{A} \subseteq \mathbf{B}$ and whenever $B_j = A_i$ then $b_j = a_i$ (Rubenstein et al., 2017).

Definition *[Compatibility] Given a set of values $\mathbf{b} \in \mathrm{dom}[\mathbf{B}]$, $\mathbf{B} \subseteq \mathbf{X}$, and an intervention $\iota = \mathrm{do}(\mathbf{a})$ we say that $\mathbf{b}$ and ι are compatible $\mathrm{Cmp}(\mathbf{b}, \iota)$ if $\mathrm{do}(\mathbf{a}) \preceq \mathrm{do}(\mathbf{b})$.*

Thus, a set of values $\mathbf{b}$ such that $\mathrm{Cmp}(\mathbf{b}, \iota)$ is a set of values that agrees with the intervention ι; a set of values $\mathbf{b}$ for which it does not hold $\mathrm{Cmp}(\mathbf{b}, \iota)$, is a setting of $\mathcal{M}$ that is ruled out by ι.

2.2. Causal Abstractions

Causal abstractions formalise relations between low-level (base) and high-level (abstracted) models, enabling causal evidence synthesis and consistent representation learning among them. This allows shifting between varying levels of granularity based on the specific inquiry or available data.

Definition *[τ-ω Exact Transformation (Rubenstein et al., 2017)] Given a base model $\mathcal{M}$ and an abstracted model $\mathcal{M}'$ respectively equipped with posets $\mathcal{I}, \mathcal{I}'$ of interventions, and a surjective and order-preserving map $\omega : \mathcal{I} \to \mathcal{I}'$, a τ-ω transformation is a map $\tau : \mathrm{dom}[\mathbf{X}] \to \mathrm{dom}[\mathbf{X}']$. An exact transformation is a map τ such that*

$$\tau_\# \left(\mathbb{P}_{\mathcal{M}_\iota}(\mathbf{X}) \right) = \mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}(\mathbf{X}'), \ \forall \iota \in \mathcal{I}. \tag{1}$$

For the ω map, *order-preserving* implies that $\iota \preceq \eta \implies \omega(\iota) \preceq \omega(\eta)$ and *surjective* that $\forall \iota' \in \mathcal{I}' \ \exists \ \iota \in \mathcal{I}$ such that $\omega(\iota) = \iota'$. An exact τ-ω transformation is a form of abstraction between probabilistic causal models (Beckers et al., 2020) that ensures commutativity between interventions and transformations: intervening via ι and then abstracting via τ leads to the same result as abstracting first via τ and then intervening via $\omega(\iota)$. Exactness is rare in realistic scenarios due to approximation and uncertainty. Thus, we permit approximate transformations (Beckers et al., 2020; Rischel and Weichwald, 2021) and introduce the concept of average abstraction error.

---

1. Also, assuming the *measurability* of the structural functions in $\mathcal{F}$ we can derive, via a pushforward over the functions in $\mathcal{F}$, the probability distribution $\mathbb{P}_{\mathcal{M}}(\mathbf{X})$ over the endogenous variables.
2. A partially-ordered set (poset) is a pair $(\mathcal{S}, \preceq)$, with a non-empty set $\mathcal{S}$ and reflexive, anti-symmetric, and transitive relation $\preceq$. In a poset, elements are comparable if one precedes the other. Totally-ordered sets are posets, where all their element pairs are comparable.

Definition *[Abstraction error] Let τ be a τ-ω transformation between SCM $\mathcal{M}$ and $\mathcal{M}'$ wrt $\mathcal{I}$ and ω. Given a discrepancy measure $\mathcal{D}$ between distributions, and a distribution q over the intervention set $\mathcal{I}$, we evaluate the approximation introduced by τ as the* abstraction error*:*

$$e(\tau) = \mathbb{E}_{\iota \sim q} \left[\mathcal{D} \left(\tau_\#(\mathbb{P}_{\mathcal{M}_\iota}), \ \mathbb{P}_{\mathcal{M}'_{\omega(\iota)}} \right) \right] \tag{2}$$

We assume a uniform distribution q over $\mathcal{I}$, treating each intervention as equally important. However, a modeller may modify this distribution, assigning varying importance to interventions of particular interest. Fig. 1 (left) shows the commutative diagram induced by such an approximate abstraction.

2.3. Optimal Transport

Optimal Transport theory (Villani et al., 2009) provides a mathematical framework to efficiently redistribute probability mass between distributions by minimising a cost function. Consider two probability measures $\mathbb{P}_{\mathcal{M}}(\mathbf{X}), \mathbb{P}_{\mathcal{M}'}(\mathbf{X}')$ on domains $\mathrm{dom}[\mathbf{X}], \mathrm{dom}[\mathbf{X}']$. When only samples from the measures are available, computational OT resorts to the corresponding empirical measures of them (Peyré et al., 2019), say $\widehat{\mathbb{P}}_{\mathcal{M}}(\mathbf{X}) = \alpha, \widehat{\mathbb{P}}_{\mathcal{M}'}(\mathbf{X}') = \beta$. Thus, we obtain i.i.d. data from the distributions, $\{\mathbf{x}_j\}_{j=1}^N \sim \mathbb{P}_{\mathcal{M}}(\mathbf{X}), \{\mathbf{x}'_i\}_{i=1}^M \sim \mathbb{P}_{\mathcal{M}'}(\mathbf{X}')$ and construct the empirical measures as $\alpha = \sum_{j=1}^N \alpha_j \delta_{\mathbf{x}_j}, \ \beta = \sum_{i=1}^M \beta_i \delta_{\mathbf{x}'_i}$ where δ is the Dirac measure. Note that in principle $\mathrm{dom}[\mathbf{X}], \mathrm{dom}[\mathbf{X}']$ could be either continuous or discrete.

The Monge (1781) formulation of OT then aims to find a map $T : \mathrm{dom}[\mathbf{X}] \to \mathrm{dom}[\mathbf{X}']$ that pushforwards α onto β, the one that minimises:

$$T^\star = \underset{T: \ T_\# \alpha = \beta}{\arg\min} \sum_{j=1}^N c(x_j, T(x_j)) \tag{3}$$

where $c : \mathbf{X} \times \mathbf{X}' \to \mathbb{R}_{\geq 0}$ represents the cost of moving a unit mass from $\mathbf{X}$ to $\mathbf{X}'$. Due to the challenge of unique solution existence of Eq. (3), Kantorovich (1942) introduced a more flexible formulation that seeks to find a coupling matrix, defined as an element of the set of stochastic matrices with given marginals, i.e., $\mathcal{U}(\alpha, \beta) = \{P \in \mathbb{R}_{\geq 0}^{M \times N} : P\mathbb{1}_M = \alpha, P^\top \mathbb{1}_N = \beta\}$. The set $\mathcal{U}(\alpha, \beta)$ is bounded and defined by $M + N$ equality constraints, and therefore is a convex polytope (Peyré et al., 2019). Solving the induced OT problem directly often poses significant computational challenges. To address this issue, *Entropic* OT, an efficient and tractable formulation of OT which incorporates an entropic regularization term, is usually used in practice . This is formalised as:

$$P^\star = \mathrm{OT}_c(\alpha, \beta) = \underset{P \in \mathcal{U}(\alpha, \beta)}{\arg\min} \langle C, P \rangle - \epsilon \mathcal{H}(P) = \underset{P \in \mathcal{U}(\alpha, \beta)}{\arg\min} \sum_{i=1, j=1}^{M,N} C_{i,j} P_{i,j} - \epsilon \mathcal{H}(P) \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product, $C \in \mathbb{R}_{\geq 0}^{M \times N}$ is a cost matrix where each element is constructed with the OT cost function $C_{i,j} = c(\mathbf{x}'_i, \mathbf{x}_j)$ and $\mathcal{H}(P)$ is the discrete entropy of the coupling matrix P with $\epsilon > 0$ a trade-off parameter. The original Kantorovich OT problem is now a special case of Eq. (4) when the entropic regularization parameter ϵ is set to 0. Further details on Optimal Transport, can be found in Appendix H.

3. Abstraction Learning as Multi-marginal Optimal Transport

In this section we formalise the CA learning problem from data as a multi-marginal OT problem, and show how we inject causal information into it. We assume **(a)** access to the causal DAGs of the base
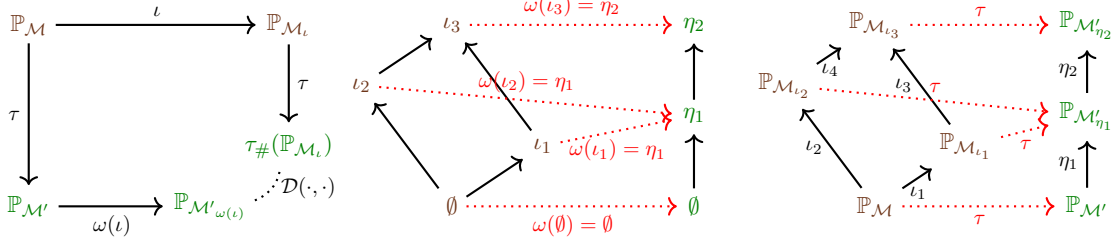
Figure 1: **(Left)** Abstraction commutative diagram. We run two different paths: **(a)** apply ι on the base model $\mathcal{M}$, and then τ; or **(b)** apply τ to get the abstracted model $\mathcal{M}'$, and then $\omega(\iota)$. We compute the distance between $\tau_\#(\mathbb{P}_{\mathcal{M}_\iota})$ and $\mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}$ using a divergence $\mathcal{D}$, indicated by the dotted line. If $\mathcal{D} = 0$, the commutative diagram expresses an exact τ-ω abstraction. **(Middle)** The action of the ω map on a poset of interventions $\mathcal{I} = \{\emptyset, \iota_1, \iota_2, \iota_3, \}$. The map ω is a surjective order-preserving map from $\mathcal{I}$ to $\mathcal{I}'$. **(Right)** The action of τ map on a poset of distributions induced by $\mathcal{I}$. A single map τ pushforwards all the base distributions $\mathbb{P}_{\mathcal{M}_\iota}$ onto abstracted distributions $\mathbb{P}_{\mathcal{M}'_\eta}$.

$\mathcal{M}$ and the abstracted $\mathcal{M}'$ models; **(b)** a finite set of interventions $\mathcal{I}$; **(c)** an intervention mapping $\omega : \mathcal{I} \to \mathcal{I}'$; **(d)** samples from the observational and interventional distributions. Thus, each base model intervention $\iota \in \mathcal{I}$ and its image $\omega(\iota) \in \mathcal{I}'$ yield a pair of empirical distributions, denoted as $\pi_\iota = \{(\widehat{\mathbb{P}}_{\mathcal{M}_\iota}(\mathbf{X}), \widehat{\mathbb{P}}_{\mathcal{M}'_{\omega(\iota)}}(\mathbf{X}'))\}$. We define the set of all these pairs as $\Pi_\omega(\mathcal{I})$; see Fig. 1(middle) for the structure of $\Pi_\omega(\mathcal{I})$. Our aim is to learn a single τ-ω transformation from data sampled from the pairs in $\Pi_\omega(\mathcal{I})$. We address this challenge by viewing each pair π_ι as marginals in an Entropic OT problem within the Kantorovich formulation for discrete measures[3]. We compute a plan P^ι for each pair π_ι, thereby leading to a *multi-marginal* optimization problem, made up of $|\Pi_\omega(\mathcal{I})|$ independent OT problems:

$$\boldsymbol{P}^\star = \text{OT}_c(\Pi_\omega(\mathcal{I})) = \underset{\{P^\iota \in \mathcal{U}(\pi_\iota)\}_{\iota \in \mathcal{I}}}{\arg\min} \left\{ \sum_{\iota \in \mathcal{I}} \left\langle C, P^\iota \right\rangle - \epsilon \mathcal{H}(P^\iota) \right\} \tag{5}$$

where $\mathcal{U}(\pi_\iota)$ is the transport polytope of each pair π_ι. As shown in Fig. 1 (right), since we are looking for a single transformation τ, the plans P^ι obtained by solving Eq. (5) have to be aggregated into a single plan $\bar{P}$, from which the map τ can be derived. In our context, we compute the final τ as a stochastic mapping $f_s : \text{dom}[\mathbf{X}] \to \mathcal{A}^{|\text{dom}[\mathbf{X}']|}$, induced from P, where $\mathcal{A}^n = \{\boldsymbol{p} \in \mathbb{R}^n, : p_i \geq 0, \sum_i p_i = 1\}$ the simplex in $\mathbb{R}^n$, to account for uncertainty of the learned abstraction with $n < \infty$ in order to allow the computation of the probability vectors. The stochastic mapping converts the mass allocation, induced by $\bar{P}$, by assigning each base sample to a probability vector, depicting a distribution over the abstracted samples.

Introducing causal knowledge into OT. The optimization problem of Eq. (5) is a collection of independent OT problems. However, we know that the marginals of different plans that correspond to the same model are linked since they are interventional distributions of the same SCM and could be formally related via *do-calculus* operations. Furthermore, standard costs used in OT (e.g. l_1, l_2, l_p) cannot capture a meaningful notion of distance between the domain of the base and the abstracted

---

3. The Kantorovich framework is essential for abstraction problems where marginal distribution dimensions mismatch, necessitating mass splitting between base and abstract points, which renders Monge maps infeasible.

model. For these reasons, we enrich Eq. (5) in two ways: **(a)** by establishing structural causal constraints amongst the different plans able to capture the relation of their marginals and **(b)** by introducing causal knowledge through the definition of a suitable cost function which integrates knowledge from the ω map. We show how such an enrichment transforms the initial problem into a joint causally informed multi-marginal optimization problem.

4. Methodology

In this section we present our methodology: Section 4.1 shows how *do-calculus* (Pearl, 2009) constraints can be incorporated into the optimization problem; Section 4.2 defines a meaningful cost for the `OT` problem; Section 4.3 presents the end-to-end `COTA` algorithm and analyzes its convexity and computational complexity.

4.1. *Do-calculus* constraints for optimal transport of abstractions

As mentioned in Section 3, marginals of different plans can be related via *do-calculus* when they refer to the same `SCM`. We first analyze the intervention set structure to identify comparable interventions.

Definition *[(Maximal) Chain] Given a poset $\mathcal{I}$, a* chain *$\mathcal{I}_q$ is a totally ordered subset of $\mathcal{I}$. Let $\mathcal{C}(\mathcal{I}) = \{\mathcal{I}_1, \ldots, \mathcal{I}_Q\}$ be the set of all chains for $\mathcal{I}$. A chain $\mathcal{I}_q \in \mathcal{C}(\mathcal{I})$ is* maximal *if $\neg\exists$ chain $\mathcal{I}_s \in \mathcal{C}(\mathcal{I})$ such that $\mathcal{I}_q \subset \mathcal{I}_s$.*

Interventions $\iota, \eta \in \mathcal{I}$ are comparable $\iota \preceq_{\mathcal{I}_q} \eta$ if there exists at least one chain $\mathcal{I}_q$ to which they both belong. Notice that comparability in $\mathcal{I}$ extends immediately to $\mathcal{I}'$ because of the order-preservation of ω. Further, we also extend the notion of comparability to transport plans.

Definition *[Comparable plans] Let intervention pairs of distribution π_ι, π_η for $\iota, \eta \in \mathcal{I}$. The induced transport plans P^ι, P^η are* comparable $P^\iota \preceq P^\eta$ iff $\exists\, \mathcal{I}_q \in \mathcal{C}(\mathcal{I})$ such that $\iota \preceq_{\mathcal{I}_q} \eta$.

Marginals of comparable plans can be related via *do-calculus*. Let $\iota = \mathrm{do}(\mathbf{a}), \omega(\iota) = \mathrm{do}(\mathbf{a}')$ and $\eta = \mathrm{do}(\mathbf{b}), \omega(\eta) = \mathrm{do}(\mathbf{b}')$, such that $\iota \preceq_{\mathcal{I}_q} \eta$. Let also the corresponding plans $P^\iota \preceq P^\eta$ be defined over the empirical measures $\widehat{\mathbb{P}}_{\mathcal{M}_\iota}, \widehat{\mathbb{P}}_{\mathcal{M}_{\omega(\iota)}}$ and $\widehat{\mathbb{P}}_{\mathcal{M}_\eta}, \widehat{\mathbb{P}}_{\mathcal{M}'_{\omega(\eta)}}$ respectively. We will show now show how a causal constraint may be derived by equating the relationships given by `OT` and *do-calculus*.

`OT` relationship. The mass conservation constraints $\mathcal{U}(\pi_\iota)$ on P^ι induced by `OT` guarantee that:

$$\overbrace{\widehat{\mathbb{P}}_{\mathcal{M}_\iota}(X_j) = \left(\sum_i P^\iota_{i,j}\right)_j}^{\text{Base}} \;\; \forall j \in \mathrm{dom}[\mathbf{X}] \qquad \overbrace{\widehat{\mathbb{P}}_{\mathcal{M}'_{\omega(\iota)}}(X'_i) = \left(\sum_j P^\iota_{i,j}\right)_i}^{\text{Abstracted}} \;\; \forall i \in \mathrm{dom}[\mathbf{X}'] \quad (6)$$

***do-calculus* relationship.** Causal inference theory (Pearl, 2009, Chapter 3, pp. 73) relates interventional distributions via the *truncated factorization* (or g-formula). Without loss of generality, let π_ι be the pair of observational distributions, where $\iota, \omega(\iota)$ are the null interventions. Then, it holds that:

$$\mathbb{P}_{\mathcal{M}_{\mathrm{do}(\mathbf{b})}}(\mathbf{X}) = \left\{\begin{array}{ll} \frac{\mathbb{P}_{\mathcal{M}}(\mathbf{X})}{\prod_i \mathbb{P}_{\mathcal{M}}(\mathbf{B}_i = \mathbf{b}_i \mid \mathrm{PA}(\mathbf{B}_i))} & \text{if } \mathrm{Cmp}(\mathbf{x}, \mathrm{do}(\mathbf{b})) \\ 0 & \text{otherwise} \end{array}\right\} \text{Base} \quad (7)$$

In our empirical setup, we express Eq. (7) through the minimization of a statistical divergence $d : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}_{\geq 0}$, where D is $|\mathrm{dom}[\mathbf{X}]|$ for the base and $|\mathrm{dom}[\mathbf{X}']|$ for the abstracted model, as follows:

$$d\left(\widehat{\mathbb{P}}_{\mathcal{M}_{\mathrm{do}(\mathbf{b})}}(\mathbf{X}),\ \frac{1}{\prod_i \widehat{\mathbb{P}}_{\mathcal{M}}(\mathbf{B}_i = \mathbf{b}_i \ | \ \mathrm{PA}(\mathbf{B}_i))}\widehat{\mathbb{P}}_{\mathcal{M}}(\mathbf{X})\right) \quad \text{if } \mathrm{Cmp}(\mathbf{x}, \mathrm{do}(\mathbf{b})) \left.\right\} \text{Base} \quad (8)$$

Throughout, we will work with the general class of Bregman divergences (Dhillon and Tropp, 2008). An equivalent relationship to Eq. (8) for the distributions of the abstracted model may be derived.

Integrating *do-calculus* and OT. Finally, in order to express Eq. (8) in terms of the optimization variables P^ι, P^η, we substitute in the mass conservation constraints for both the base and the abstracted models given by Eq. (6):

$$\delta_{\iota,\eta}(P^\iota, P^\eta) := d\left(\left(\sum_i P^\eta_{i,j}\right)_j,\ \frac{1}{(\mathcal{Z}^\eta)_j}\left(\sum_i P^\iota_{i,j}\right)_j\right) \quad \text{if } \mathrm{Cmp}(x_j, \eta). \left.\right\} \text{Base} \qquad (9)$$

$$\delta'_{\iota,\eta}(P^\iota, P^\eta) := d\left(\left(\sum_j P^\eta_{i,j}\right)_i,\ \frac{1}{(\mathcal{Z}^{\omega(\eta)})_i}\left(\sum_j P^\iota_{i,j}\right)_i\right) \quad \text{if } \mathrm{Cmp}(x'_i, \omega(\eta)). \left.\right\} \text{Abstracted}$$

$$(10)$$

where $\mathcal{Z}^\eta, \mathcal{Z}^{\omega(\eta)}$ are the normalizing vectors for the base and the abstracted distributions respectively, induced from Eq. (7); see Appendix A for their derivation in terms of the plans. Instead of computing independently the OT plans as in Eq. (5) we can jointly learn plans that preserve causal relations by incorporating the base and abstracted model distances $\mathcal{D}(P^\iota, P^\eta) = [\delta_{\iota,\eta},\ \delta'_{\iota,\eta}]^\top$ defined over the marginals of two plans.

4.2. A causally informed cost function

The OT cost function captures the transport problem's geometry in order to find the optimal plan. Although in $\mathbb{R}$ costs like l_p can represent the distance between samples, this is not trivial when dealing with samples between causal models. However, the ω map of the τ-ω transformation provides a solution by formally encoding the interventional relationship between samples of two SCMs. In order to compute a distance between samples $\mathbf{x} \in \mathrm{dom}[\mathbf{X}]$ of the base and $\mathbf{x}' \in \mathrm{dom}[\mathbf{X}']$ of the abstracted model, given interventions $\iota = \mathrm{do}(\mathbf{a})$ and $\omega(\iota) = \mathrm{do}(\mathbf{a}')$, we exploit ω to discount the cost of transporting sample $\mathbf{a}$ to $\mathbf{a}'$. We then define $c_\omega : \mathrm{dom}[\mathbf{X}] \times \mathrm{dom}[\mathbf{X}'] \to \mathbb{R}_{\geq 0}$:

$$c_\omega(\mathbf{x}, \mathbf{x}') = |\mathcal{I}| - \sum_{\iota \in \mathcal{I}} \mathbb{1}\left[\mathrm{Cmp}(\mathbf{x}, \iota) \wedge \mathrm{Cmp}(\mathbf{x}', \omega(\iota))\right], \qquad (11)$$

where $\mathbb{1}[a]$ is the indicator function returning one if the condition a is satisfied. The function c_ω discounts the cost of transporting the sample $\mathbf{x}$ to $\mathbf{x}'$ proportionally to the number of pairs $(\iota, \omega(\iota))$ w.r.t. which $\mathbf{x}$ and $\mathbf{x}'$ are compatible. Hence, the larger and more diverse the set of pairs $\Pi_\omega(\mathcal{I})$ is, the more informative the ω-cost will be, thereby enhancing its capacity to convey comprehensive insights into the cost matrix. This sensitivity of c_ω to $\mathcal{I}$ is demonstrated in Section 6. Finally, by construction, C_ω has the advantage of being is invariant to the ordering of the values $\mathbf{x}$ (columns) and $\mathbf{x}'$ (rows). In Appendix C we offer an illustration of a C_ω matrix derived from c_ω and further discuss the construction of ω-costs.

4.3. The causal optimal transport of abstractions (COTA) objective

We now discuss how we can integrate the *do-calculus* constraints discussed in Section 4.1 and the ω-informed cost presented in Section 4.2 in the the OT framework to jointly solve multiple transport problems and learn an abstraction τ. For a given set of pairs $\Pi_\omega(\mathcal{I}_k) = \{\pi_{\iota_1}, \cdots \pi_{\iota_N} \mid \iota_n \in \mathcal{I}_k\}$ where $\mathcal{I}_k$ a maximal chain, we define the objective function of COTA as the following OT problem:

$$\boldsymbol{P}_k^\star = \text{COTA}_c\left(\Pi_\omega\left(\mathcal{I}_k\right)\right) = \underset{\{P^{\iota_n} \in \mathcal{U}(\pi_{\iota_n})\}_{\iota_n \in \mathcal{I}_k}}{\arg\min} \left\{ \kappa \cdot \sum_{\iota_n \in \mathcal{I}_k} \underbrace{\langle C_\omega, P^{\iota_n} \rangle}_{\text{OT}} + \boldsymbol{\lambda}^\top \underbrace{\mathcal{D}(P^{\iota_n}, P^{\iota_{n+1}})}_{\text{do-calculus constraints}} - \mu \cdot \underbrace{\mathcal{H}(P^{\iota_n})}_{\text{entropy}} \right\},$$
(12)

where $\boldsymbol{\lambda} = [\lambda, \lambda']^\top$ and $(\kappa, \boldsymbol{\lambda}, \mu)$ a convex combination, i.e. $\kappa + \lambda + \lambda' + \mu = 1$ for $\kappa, \lambda, \lambda', \mu \geq 0$. Thus, we transformed the initial CA problem of Eq. (5) into a joint multi-marginal OT problem integrated with causal knowledge from different sources. Algorithm 1 presents the end-to end implementation of COTA. The complexity of the algorithm is $\mathcal{O}(N_{\max} + N_{\text{chains}} \cdot (D_{\max}^3 \log D_{\max}))$, where $N_{\max} = \max(N, N')$ with N, N' respectively the number of samples for base and abstracted model, N_{chains} the number of maximal chains, and $D_{\max} = \max(D, D')$ with $D = |\text{dom}[\mathbf{X}]|$, $D' = |\text{dom}[\mathbf{X}']|$. The first term accounts for the complexity of line 3, while the second term accounts for the loop of line 13 and the internal call of the COTA solver. The entropy regularization $\mathcal{H}(\mathcal{P}^{\iota_n})$ allows for a speed up to $\mathcal{O}(D^2)$ with Sinkhorn algorithms (Peyré et al., 2019). In Appendix F we also present Approximate COTA, an approximation that halves the causal constraints $\mathcal{D}(\cdot, \cdot)$ of the optimisation problem of Eq. (12).

Theorem 1 (Convexity of COTA) *The optimization problem given by Eq. (12) is jointly convex in the transport plans $\{P^{\iota_n} \in \mathcal{U}(\pi_{\iota_n})\}_{\iota_n \in \mathcal{I}_k} \forall \mathcal{I}_k \in \mathcal{C}(\mathcal{I})$. When $\mu > 0$, it is strictly convex.*

Proof (sketch) The main assertion that needs to be shown is the joint convexity of the function $\{P^{\iota_n}\}_{\iota_n \in \mathcal{I}_k} \to \kappa \cdot \sum_{\iota_n \in \mathcal{I}_k} \langle C_\omega, P^{\iota_n} \rangle + \boldsymbol{\lambda}^\top \cdot \mathcal{D}(P^{\iota_n}, P^{\iota_{n+1}}) + \mu \cdot \mathcal{H}(P^{\iota_n})$ in all of the plans. [4] To show the main assertion, we use two inequalities following directly from the convexity definition of the single plan function $P^{\iota_n} \to \langle C_\omega, P^{\iota_n} \rangle$ and from the joint convexity of $(P^{\iota_n}, P^{\iota_{n+1}}) \to \mathcal{D}(P^{\iota_n}, P^{\iota_{n+1}})$ in $P^{\iota_n}, P^{\iota_{n+1}}$. Further, note that the entropic regulariser $\mathcal{H}(P^{\iota_n})$ is known to be a strictly convex function of P^{ι_n} (Peyré et al., 2019). Finally, since the function defined by the set of all the plans in $\mathcal{I}_k$ is a summation, combining these two inequalities gives the desired result, with strict convexity holding for $\mu > 0$ due to $\mathcal{H}(P^{\iota_n})$. The full proof is provided in Appendix B. ∎

5. Related work

In this section we briefly review related works of OT within the domain of causality and the multi-marginal techniques akin to our own methodology. There has been increasing interest in applying OT methodology to perform inference in causal models. Regarding treatment effect estimation, Torous et al. (2021) propose estimators of binary treatment effects in the potential outcome framework based on OT to handle high-dimensional covariates; Gunsilius and Xu (2021) address covariate matching in multi-valued treatments via multi-marginal OT. Recently, Tu et al. (2022) propose

---

4. Other conditions, like convex constraints and domain properties, are satisfied by stochastic matrices; see Appendix B.

Algorithm 1 COTA

Require: DAGs for $\mathcal{M}[\mathbf{X}], \mathcal{M}'[\mathbf{X}']$, sets $\mathcal{I}, \mathcal{I}'$ and $\omega : \mathcal{I} \to \mathcal{I}'$, surjective and order-preserving.
Ensure: $\tau : \text{dom}[\mathbf{X}] \to \mathcal{A}^{|\text{dom}[\mathbf{X}']|}$, where $\mathcal{A}^n = \{\boldsymbol{p} \in \mathbb{R}^n : p_i \geq 0, \sum_i p_i = 1\}$.

1: **for** $\iota \in \mathcal{I}$ **do:**
2:    $\{\mathbf{x}_j\}_{j=1}^N \sim \mathbb{P}_{\mathcal{M}_\iota}(\mathbf{X}), \{\mathbf{x}'_i\}_{i=1}^M \sim \mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}(\mathbf{X}')$ *# sampling from the true distributions*
3:    $\widehat{\mathbb{P}}_{\mathcal{M}_\iota}(\mathbf{X}) \leftarrow \sum_{j=1}^N \alpha_j \delta_{\mathbf{x}_j}, \widehat{\mathbb{P}}_{\mathcal{M}'_{\omega(\iota)}}(\mathbf{X}') \leftarrow \sum_{i=1}^M \beta_i \delta_{\mathbf{x}'_i}$ *# construct the empirical measures*
4: **for** $j = 0$ **to** N **do:**
5:    **for** $i = 0$ **to** M **do:**
6:       $C_\omega[i,j] \leftarrow |\mathcal{I}| - \sum_{\iota \in \mathcal{I}} \mathbb{1}\left[\text{Cmp}(\mathbf{x_j}, \iota) \wedge \text{Cmp}(\mathbf{x_i}', \omega(\iota))\right]$ *# compute the ω-cost matrix*
7: $\mathcal{C}(\mathcal{I}) \leftarrow$ compute_chains$(\mathcal{I})$ *# compute the set of all maximal chains of $\mathcal{I}$*
8: **for** $\mathcal{I}_k \in \mathcal{C}(\mathcal{I})$ **do:**
9:    $\Pi_\omega(\mathcal{I}_k) \leftarrow \emptyset$
10:    **for** $\iota \in \mathcal{I}_k$ **do:**
11:       $\Pi_\omega(\mathcal{I}_k) \leftarrow \Pi_\omega(\mathcal{I}_k) \cup \{(\widehat{\mathbb{P}}_{\mathcal{M}_\iota}, \widehat{\mathbb{P}}_{\mathcal{M}'_{\omega(\iota)}})\}$ *# compute the set of pairs for every $\mathcal{I}_k$*
12: $\mathcal{P} \leftarrow \emptyset$
13: **for** $\mathcal{I}_k \in \mathcal{C}(\mathcal{I})$ **do:**
14:    $\mathcal{P} \leftarrow \mathcal{P} \cup \{\text{COTA}_c(\Pi_\omega(\mathcal{I}_k)\}$ *# run* COTA *for every $\mathcal{I}_k$ and assemble the set of all plans*
15: **if** COTA$(\widehat{\mathcal{P}})$ **then** $\widehat{\mathcal{P}} \leftarrow \frac{1}{|\mathcal{P}|} \sum_{t=1}^{|\mathcal{P}|} P_t$
16:    **Return:** $\tau \leftarrow f_s(\widehat{\mathcal{P}})$ *# return the final τ as the stochastic mapping $f_s(\cdot)$ of the average plan*
17: **if** COTA$(\widehat{\tau})$ **then** $\widehat{f_s(P_t)} \leftarrow \frac{1}{|\mathcal{P}|} \sum_{t=1}^{|\mathcal{P}|} f_s(P_t)$
18:    **Return:** $\tau \leftarrow \widehat{f_s(P_t)}$ *# return the final τ as the average stochastic mapping $f_s(\cdot)$ of the plans*

---

the use of OT to perform bivariate causal discovery in the context of continuous data and additive noise, with benefits such as avoiding to specify likelihoods and efficient computation due to one-dimensional distributions. Additionally, an extension of OT to multiple marginals is considered in Peyré et al. (2019); Pass (2015); Kostic et al. (2022). In this setting, the problem involves finding couplings between source and target distributions, even in high-dimensional cases. Differently from our formulation, this multi-marginal setup does not consider any relations between the computed transport plans. Finally, in Table 21 of Appendix I we provide a comparative table, showcasing the different proposed approaches in CA learning, including also concurrent research works.

6. Experiments

Throughout the experiments, we investigate the performance of COTA under diverse experimental settings and in different tasks in order to showcase: **(a)** its superiority over non-causal solutions; **(b)** the actual gains of introducing the *do-calculus* constraints into the optimization routine; **(c)** the advantage of the causally informed ω-cost, relative to the size and diversity of the intervention set, compared to standard/non-causal costs; and **(d)** the efficiency of COTA as a data augmentation tool

in a downstream task compared to established state-of-the-art CA learning frameworks. The code and results for all experiments are publicly accessible [5].

COTA. We run COTA considering two Bregman divergences for the distance term $\mathcal{D}$ of Eq. (12), the Frobenius norm (FRO) and the Jensen–Shannon Divergence (JSD). We also run an ablation study in which we replace c_ω in COTA's objective with a conventional cost $c_\mathcal{H}$ based on the Hamming distance and compare the two costs' performance. Finally, regarding the parameter $\boldsymbol{\lambda} = [\lambda, \lambda']^\top$ in Eq. (12), in the main paper we demonstrate the equal weight case where $\lambda = \lambda'$[6] and provide additional results from the more general case of $\lambda \neq \lambda'$ in Appendix E.

Baselines. In addition, we compare our method with three alternative setups which serve as baselines. These configurations consist of non-causal independent solutions of the OT problem, as described in Eq. (5), or barycentric adaptations of the standard OT framework. In particular:

- In Pwise OT we apply OT_c to generate a set of k independent plans $\mathcal{P} = \{P_1, \ldots, P_k\}$ and aggregate them into an average single plan $\widehat{\mathcal{P}}$ and compute $\tau = f_s(\widehat{\mathcal{P}})$. This is equivalent to COTA($\widehat{\mathcal{P}}$) when $\kappa = 1$, $\lambda = 0$ and $\mu = 0$.

- In Bary OT we first compute two barycenters: one of the base model's ($\bar{\boldsymbol{\alpha}}$) and one of the abstracted model's distributions ($\bar{\boldsymbol{\beta}}$) and then solve the standard OT_c (single-pair) problem for $(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}})$, to compute the plan $\bar{P}$, and finally compute $\tau = f_s(\bar{P})$.

- In Map OT we apply OT_c to generate the set of k plans $\mathcal{P} = \{P_1, \ldots, P_k\}$, compute the k independent stochastic maps from them and compute τ as the average of them $\tau = \{\widehat{f_s(P_i)}\}_{i \in [K]}$. This is equivalent to COTA($\widehat{\tau}$) when $\kappa = 1$, $\lambda = 0$ and $\mu = 0$.

Evaluation. Across all scenarios, we assess the learned τ in terms of the abstraction error from Eq. (2) using the JSD ($e_{\text{JSD}}(\tau)$) and the Wasserstein ($e_{\text{WASS}}(\tau)$) distances by employing a leave-one-pair-out procedure to measure the quality and the robustness of the learned abstraction. Specifically, we remove one pair $\pi_i = (\alpha_i, \beta_i)$ from $\Pi_\omega(\mathcal{I})$, learn τ from the remaining pairs, and measure the $e(\tau)$ distance between $\tau_\#(\alpha_i)$ and β_i. All reported measures are presented as the mean and standard deviation over 50 repetitions with a 95% confidence interval. For COTA we select the hyperparameters (κ, λ, μ) via a grid-search of 100 convex combinations.

6.1. Causal Abstraction Learning Simulations

The DAGs alongside their intervention sets for all the following scenarios are presented in Appendix G.

Simple Lung Cancer (STC). This motivating example is made up by a discrete base model with a chain structure (Smoking → Tar → Cancer) and an abstracted that removes the mediator node. We investigate two different scenarios: when the interventions are performed on variables **(a)** without parents (**STC**$_{\text{np}}$) and **(b)** with parents (**STC**$_{\text{p}}$). Table 1 showcases the abstraction error of COTA in **STC**$_{\text{np}}$ and demonstrates its superiority against all the different baselines, both for $e_{\text{WASS}}(\tau)$ and $e_{\text{JSD}}(\tau)$. Also, notice how c_ω returns a lower abstraction error compared to $c_\mathcal{H}$, not only in different settings of COTA, but also with baseline ones. This suggests that indeed the ω-cost can provide more relevant information for learning an abstraction. Further, in Fig. 2 we highlight the impact of the

---

5. github.com/yfelekis/COTA

6. i.e. $\boldsymbol{\lambda}^\top \cdot \mathcal{D}(P^\iota, P^\eta) = [\lambda, \lambda] \cdot \begin{bmatrix} \delta_{\iota,\eta} \\ \delta'_{\iota,\eta} \end{bmatrix} = \lambda \cdot (\delta_{\iota,\eta} + \delta'_{\iota,\eta})$.

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\text{JSD}}(\tau)$	$e_{\text{WASS}}(\tau)$
COTA($\widehat{P}$)	FRO	c_ω	**0.010 ± 0.005**	**0.011 ± 0.003**
		$c_{\mathcal{H}}$	0.087 ± 0.007	0.025 ± 0.001
	JSD	c_ω	0.012 ± 0.006	0.012 ± 0.003
		$c_{\mathcal{H}}$	0.087 ± 0.006	0.025 ± 0.001
COTA($\widehat{\tau}$)	FRO	c_ω	0.013 ± 0.021	0.171 ± 0.001
		$c_{\mathcal{H}}$	0.169 ± 0.005	0.178 ± 0.001
	JSD	c_ω	0.014 ± 0.021	0.171 ± 0.001
		$c_{\mathcal{H}}$	0.169 ± 0.004	0.178 ± 0.001
Pwise OT	–	c_ω	0.013 ± 0.002	0.011 ± 0.002
	–	$c_{\mathcal{H}}$	0.093 ± 0.004	0.039 ± 0.002
Map OT	–	c_ω	0.023 ± 0.022	0.147 ± 0.001
	–	$c_{\mathcal{H}}$	0.169 ± 0.022	0.156 ± 0.001
Bary OT	–	c_ω	0.233 ± 0.142	0.067 ± 0.042
	–	$c_{\mathcal{H}}$	0.323 ± 0.074	0.095 ± 0.039

Table 1: Abstraction error evaluation for the **STC**$_{\text{np}}$ example. The configuration COTA($\widehat{P}$) − FRO − c_ω yields the lowest abstraction error when compared to all other settings and baselines, for both metrics. A "rich" intervention set c_ω effectively captures the correspondence between samples leading to superior performance over $c_{\mathcal{H}}$ in all COTA settings and baselines.

Figure 2: Effect of λ for the **STC**$_{\text{np}}$ example. The ternary plot illustrates a grid-search amongst 1000 convex combinations of (κ, λ, μ) for the COTA($\widehat{P}$) − FRO − c_ω setting. Values of λ close to zero present high abstraction error, demonstrating the benefit of the *do-calculus* constraints in the OT problem. The minimum is reached at (.81, .17, .02) and is denoted with "×".

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\text{JSD}}(\tau)$	$e_{\text{WASS}}(\tau)$
COTA($\widehat{\tau}$)	FRO	c_ω	0.249 ± 0.005	0.135 ± 0.001
		$c_{\mathcal{H}}$	**0.229 ± 0.003**	0.129 ± 0.001

Table 2: Abstraction error for the **STC**$_{\text{p}}$ example with $\mathcal{I} = \{\text{do}(T = 0), \text{do}(T = 1)\}$. Cost $c_{\mathcal{H}}$ outperforms c_ω due to limited intervention set $\mathcal{I}$.
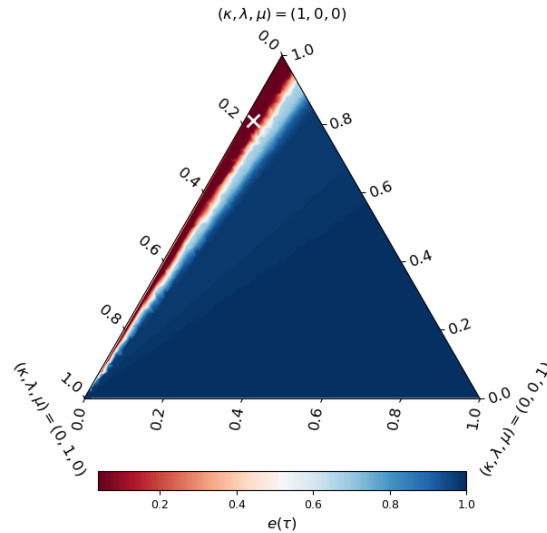
Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\text{JSD}}(\tau)$	$e_{\text{WASS}}(\tau)$
COTA($\widehat{\tau}$)	FRO	c_ω	0.379	0.053
		$c_{\mathcal{H}}$	**0.220**	0.053

Table 3: Abstraction error for the **EBM** example. The conventional cost $c_{\mathcal{H}}$ outperforms c_ω due to a limited intervention set $\mathcal{I}$.

parameter λ, which weighs the *do-calculus* constraints' term in Eq. (12); as the best performing setting of COTA is reached for $\lambda > 0$, this validates the hypothesis that introducing causal constraints helps learning better abstractions. In the **STC**$_{\text{p}}$ scenario, COTA still outperforms the baselines (see Appendix E), although Table 2 shows that $c_{\mathcal{H}}$ returns a lower abstraction error compared to c_ω. We argue that this is likely due to the dependency of ω-cost on the intervention set, which, in this case, comprises only two interventions. As explained in Section 4.2 such a small intervention set leads to an almost-uniform and uninformative ω-cost, and in such case a conventional cost like Hamming might be able to capture certain patterns more efficiently. Visualizations of the induced cost matrices for both functions can be found in Fig. 4 and Fig. 5 in Appendix C.

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\text{JSD}}(\tau)$	$e_{\text{WASS}}(\tau)$
COTA($\widehat{P}$)	FRO	c_ω	0.287 ± 0.014	$\mathbf{0.044 \pm 0.001}$
		$c_\mathcal{H}$	0.287 ± 0.014	0.047 ± 0.001
COTA($\widehat{\tau}$)	FRO	c_ω	$\mathbf{0.263 \pm 0.005}$	0.061 ± 0.001
		$c_\mathcal{H}$	$\mathbf{0.263 \pm 0.006}$	0.061 ± 0.001
Pwise OT	–	c_ω	0.306 ± 0.009	0.045 ± 0.001
	–	$c_\mathcal{H}$	0.387 ± 0.002	0.047 ± 0.001
Map OT	–	c_ω	0.294 ± 0.008	0.054 ± 0.001
	–	$c_\mathcal{H}$	0.350 ± 0.005	0.054 ± 0.001
Bary OT	–	c_ω	0.294 ± 0.047	0.044 ± 0.003
	–	$c_\mathcal{H}$	0.414 ± 0.040	0.046 ± 0.010

Table 4: Abstraction error for **LUCAS**. The configuration COTA($\widehat{\tau}$) − FRO yields the lowest abstraction error for the JSD metric and the COTA($\widehat{P}$) − FRO − c_ω setting for the WASS metric. The ω-cost c_ω outperforms $c_\mathcal{H}$.

Training set	Test set	Zennaro et al. (2023)	COTA
LRCS[$CG \neq k$]	LRCS[$CG = k$]	1.86 ± 1.75	$\mathbf{1.40 \pm 1.39}$
LRCS[$CG \neq k$] +WMG	LRCS[$CG = k$]	0.22 ± 0.26	$\mathbf{0.19 \pm 0.04}$
+WMG[$CG \neq k$]	WMG[$CG = k$]	1.22 ± 0.95	$\mathbf{0.80 \pm 0.55}$

Table 5: MSE of COTA and a SOTA CA framework on a regression task for **EBM**. Augmenting data via the learned abstraction reduces the average error in all different settings compared to the SOTA. We used COTA($\widehat{P}$) − FRO − c_ω with the hyperparameters $(\kappa, \lambda, \mu) = (.2, .5, .3)$ achieving the lowest abstraction error.

LUng CAncer Set (LUCAS)   This model[7] is a large-scale synthetic SCM designed to simulate data related to the study of lung cancer. Table 4 confirms that even on larger and more realistic problems COTA exceeds the baselines in terms of the abstraction error, and c_ω provides results at least as good as $c_\mathcal{H}$. The full table of results can be found in Appendix E.

Electric Battery Manufacturing (EBM)   Finally we compare with the only real-world public dataset that, to the best of our knowledge, has been used for CA learning (Zennaro et al., 2023). This dataset contains data related to electric battery manufacturing collected from two experimental settings. The first setting (WMG) has been modelled through a low-level SCM that captures the effect of a control variable (*comma gap*) on an output (*mass loading*) at multiple spatial locations. The second setting (LRCS) is modelled through a high-level SCM that relates the same control variable to a single output (Cunha et al., 2020). We learn an abstraction map τ using the set of real interventions performed during the collection of the data. Following Zennaro et al. (2023), we use the learned map τ to abstract the WMG data and aggregate them with the LRCS data; then we perform a set of downstream regression tasks. Compared to the SOTA which required full knowledge of the underlying SCM, COTA requires only knowledge of the underlying DAG and provides better results in terms of the Mean Square Error (MSE) in all the proposed setups, as shown in Table 5. In general, COTA is always the top performer compared to the baselines (see Appendix E for the complete table of results), while in Table 3 we highlight a similar behaviour as the one in Table 2 whereby the limited size of the interventional data prevents $c_\mathcal{H}$ to lead to better abstractions compared to c_ω. A complete presentation of the different settings and the data of this case study can be found in Appendix D.

---

7. http://www.causality.inf.ethz.ch/data/LUCAS.html

7. Discussion

In this work, we presented `COTA`, a framework for learning causal abstractions from observational and interventional data through a causally constrained multi-marginal `OT` formulation. Incorporating *do-calculus* constraints and a causally-informed cost c_ω in the optimisation problem led to lower abstraction error compared to non-causal baselines in diverse scenarios. The effectiveness of the c_ω cost was shown to be sensitive to the intervention set; expanding the interventions set improved the performance with the ω-cost compared to conventional non-causal costs like Hamming. Lastly, `COTA` outperform the prior `CA` learning art of Zennaro et al. (2023) when employed as a data augmentation procedure on a real world regression task.

Our work opens up new challenges and directions in both fields of causal abstraction learning and `OT`. First, constrained multi-marginal `OT` settings like `COTA` have been understudied in the literature, and further theoretical work on the guarantees of existence and uniqueness of the estimated maps is needed. Recent methods for joint learning of plans and parameterised maps (Uscidda and Cuturi, 2023; Seguy et al., 2018) to obtain better estimators (Perrot et al., 2016) hold promise in this front. Furthermore, generalising a framework like `COTA` to semi-Markovian `SCMs` presents a significant challenge, because lifting the causal sufficiency assumption may render the estimation of certain causal constraints unidentifiable. Finally, another interesting direction would be to extend `CA` learning frameworks like `COTA` in order to incorporate temporal dependencies and continuous-time models, for example structural dynamical causal models (Bongers et al., 2018).

Acknowledgments

References

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685, 2019.

Sander Beckers, Frederick Eberhardt, and Joseph Y Halpern. Approximate causal abstractions. In *Uncertainty in Artificial Intelligence*, pages 606–615. PMLR, 2020.

Stephan Bongers, Tineke Blom, and Joris M Mooij. Causal modeling of dynamical systems. *arXiv preprint arXiv:1803.08784*, 2018.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991. doi: https://doi.org/10.1002/cpa. 3160440402. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa. 3160440402.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017.

Ricardo Pinto Cunha, Teo Lombardo, Emiliano N Primo, and Alejandro A Franco. Artificial intelligence investigation of nmc cathode manufacturing parameters interdependencies. *Batteries & Supercaps*, 3(1):60–67, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Inderjit S Dhillon and Joel A Tropp. Matrix nearness problems with bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2008.

Joel Dyer, Nicholas Bishop, Yorgos Felekis, Fabio Massimo Zennaro, Anisoara Calinescu, Theodoros Damoulas, and Michael Wooldridge. Interventionally consistent surrogates for agent-based simulators, 2023.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

Florian Gunsilius and Yuliang Xu. Matching for causal effects via multimarginal optimal transport. *arXiv preprint arXiv:2112.04398*, 2021.

L. Kantorovich. On the transfer of masses (in russian). Doklady Akademii Nauk, August 1942.

Armin Kekić, Bernhard Schölkopf, and Michel Besserve. Targeted reduction of causal models, 2023.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017. doi: 10.1109/MSP.2017.2695801.

Vladimir R Kostic, Saverio Salzo, and Massimiliano Pontil. Batch greenkhorn algorithm for entropic-regularized multimarginal optimal transport: Linear rate of convergence and iteration complexity. In *International Conference on Machine Learning*, pages 11529–11558. PMLR, 2022.

Riccardo Massidda, Atticus Geiger, Thomas Icard, and Davide Bacciu. Causal abstraction with soft interventions. *arXiv preprint arXiv:2211.12270*, 2022.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

Mona Faraji Niri, Kailong Liu, Geanina Apachitei, Luis A.A Román-Ramírez, Michael Lain, Dhammika Widanage, and James Marco. Quantifying key factors for optimised manufacturing of li-ion battery anode and cathode via artificial intelligence. *Energy and AI*, 7:100129, jan 2022. doi: 10.1016/j.egyai.2021.100129.

Brendan Pass. Multi-marginal optimal transport: theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 49(6): 1771–1790, 2015.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 247–254, 2011.

Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. *Advances in Neural Information Processing Systems*, 29, 2016.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Eigil F Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. *arXiv preprint arXiv:2103.15758*, 2021.

Eigil Fjeldgren Rischel. The category theory of causal models. Master's thesis, University of Copenhagen, 2020.

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, pages 808–817. Curran Associates, Inc., 2017.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large scale optimal transport and mapping estimation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=B1zlp1bRW.

Vignesh Ram Somnath, Charlotte Bunne, and Andreas Krause. Multi-scale representation learning on proteins. *Advances in Neural Information Processing Systems*, 34:25244–25255, 2021.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Matthew Thorpe. Introduction to optimal transport. 2017. URL https://api.semanticscholar.org/CorpusID:131768046.

William Torous, Florian Gunsilius, and Philippe Rigollet. An optimal transport approach to causal inference. *arXiv preprint arXiv:2108.05858*, 2021.

Ruibo Tu, Kun Zhang, Hedvig Kjellström, and Cheng Zhang. Optimal transport for causal discovery. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=qwBK94cP1y.

Théo Uscidda and Marco Cuturi. The monge gap: A regularizer to learn all transport maps. *arXiv preprint arXiv:2302.04953*, 2023.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

E Weinan. *Principles of multiscale modeling*. Cambridge University Press, 2011.

Kevin Xia and Elias Bareinboim. Neural causal abstractions, 2024.

Mingzhang Yin, Yixin Wang, and David M Blei. Optimization-based causal estimation from heterogenous environments. *arXiv preprint arXiv:2109.11990*, 2021.

Fabio Massimo Zennaro, Máté Drávucz, Geanina Apachitei, W. Dhammika Widanage, and Theodoros Damoulas. Jointly learning consistent causal abstractions over multiple interventional distributions. In *2nd Conference on Causal Learning and Reasoning*, 2023. URL https://openreview.net/forum?id=RNs7aMS6zDq.

Appendix A. Derivation of normalizing vectors

We follow the notation introduced in the main text. For $\iota \preceq \eta$, $\iota = \mathrm{do}(\mathbf{A} = \mathbf{a})$ and $\omega(\iota) = \mathrm{do}(\mathbf{A}' = \mathbf{a}')$ induce a transport plan P^ι, while $\eta = \mathrm{do}(\mathbf{B} = \mathbf{b})$ and $\omega(\eta) = \mathrm{do}(\mathbf{B}' = \mathbf{b}')$ induce a transport plan P^η. Our aim is to express the normalizing vectors $\mathcal{Z}^\eta = \mathbb{P}_\mathcal{M}(\mathbf{B}_i = \mathbf{b}_i \mid \mathrm{PA}(\mathbf{B}_i))$ and $\mathcal{Z}^{\omega(\eta)} = \mathbb{P}_{\mathcal{M}'}(\mathbf{B}'_i = \mathbf{b}'_i \mid \mathrm{PA}(\mathbf{B}'_i))$ in terms of the plan P^ι. These are the conditional probabilities defined in Eq. (7) and can be written as:

$$\mathbb{P}_\mathcal{M}(\mathbf{B}_i = \mathbf{b}_i \mid \mathrm{PA}(\mathbf{B}_i)) = \frac{\mathbb{P}_\mathcal{M}(\mathbf{B}_i = \mathbf{b}_i, \mathrm{PA}(\mathbf{B}_i))}{\mathbb{P}_\mathcal{M}(\mathrm{PA}(\mathbf{B}_i))} \left.\right\} \text{Base} \tag{13}$$

$$\mathbb{P}_{\mathcal{M}'}(\mathbf{B}'_i = \mathbf{b}'_i \mid \mathrm{PA}(\mathbf{B}'_i)) = \frac{\mathbb{P}_{\mathcal{M}'}(\mathbf{B}'_i = \mathbf{b}'_i, \mathrm{PA}(\mathbf{B}'_i))}{\mathbb{P}_{\mathcal{M}'}(\mathrm{PA}(\mathbf{B}'_i))} \left.\right\} \text{Abstracted} \tag{14}$$

Then given $P^\iota \preceq P^\eta$ we express the sub-parts of the equations above in terms of the transportation plan P^ι by defining specific sets of indices on it. Starting from Eq. (13) and the denominator $\mathbb{P}_\mathcal{M}(\mathrm{PA}(\mathbf{B}_i))$, this can be then written as:

$$\mathbb{P}_\mathcal{M}(\mathrm{PA}(\mathbf{B}_i)) = \sum_{[i,j] \in \mathcal{O}_{\iota,\eta,\rho}} P^\iota_{ij}, \tag{15}$$

where $\mathcal{O}_{\iota,\eta,\rho} = \{[i,j] \mid x_j \in \mathrm{dom}[\mathbf{X}] \wedge \mathrm{Cmp}(x_j, \mathrm{do}(\mathrm{PA}_\mathbf{B} = \rho))\}$, for $\rho \in \mathrm{dom}[\mathrm{PA}_\mathbf{B}]$.

Consequently, for the numerator we first define the following set:

$$\mathcal{C}_{\iota,\eta} = \{[i,j] \mid x_j \in \mathrm{dom}[\mathbf{X}] \wedge \mathrm{Cmp}(x_j, \eta)\}$$

and also let the intersection set $\Omega_{\iota,\eta,\rho} = \mathcal{C}_{\iota,\eta} \cap \mathcal{O}_{\iota,\eta,\rho}$ for every $\rho \in \mathrm{dom}[\mathrm{PA}_{\mathbf{B}}]$. Then, we have:

$$\mathbb{P}_{\mathcal{M}}(\mathbf{B}_i = \mathbf{b}_i, \mathrm{PA}(\mathbf{B}_i)) = \sum_{[i,j]\in\Omega_{\iota,\eta,\rho}} P_{i,j}^{\iota} \tag{16}$$

By performing the symmetric operations for the abstracted model in Eq. (14) and get the respective sets $\mathcal{O}'_{\iota,\eta,\rho'}, \mathcal{C}'_{\iota,\eta}$ and $\Omega'_{\iota,\eta,\rho}$ we can then finally define the normalizing vectors in terms of the plan as follows:

$$\mathcal{Z}^{\eta} = \frac{\sum_{[i,j]\in\Omega_{\iota,\eta,\rho}} P_{i,j}^{\iota}}{\sum_{[i,j]\in\mathcal{O}_{\iota,\eta,\rho}} P_{i,j}^{\iota}}, \quad \mathcal{Z}^{\omega(\eta)} = \frac{\sum_{[i,j]\in\Omega'_{\iota,\eta,\rho}} P_{i,j}^{\iota}}{\sum_{[i,j]\in\mathcal{O}'_{\iota,\eta,\rho'}} P_{i,j}^{\iota}} \tag{17}$$

We also present the derivation of the special case in which the intervened variables have *no parents*. Specifically, one can easily see that in this case:

$$\mathcal{O}_{\iota,\eta,\rho} = \{[i,j] \mid x_j \in \mathrm{dom}[\mathbf{X}]\} \quad \text{for } \rho \in \mathrm{dom}[\mathrm{PA}_{\mathbf{B}}]$$

$$\mathcal{O}'_{\iota,\eta,\rho'} = \{[i,j] \mid x'_i \in \mathrm{dom}[\mathbf{X}']\} \quad \text{for } \rho' \in \mathrm{dom}[\mathrm{PA}_{\mathbf{B}'}]$$

This implies that $\Omega_{\iota,\eta,\rho} = \mathcal{C}_{\iota,\eta}$ and similarly $\Omega'_{\iota,\eta,\rho} = \mathcal{C}'_{\iota,\eta}$. Therefore, since $\mathrm{PA}_{\mathbf{B}} = \mathrm{PA}_{\mathbf{B}'} = \varnothing$ then $\mathrm{Cmp}(x_j, \mathrm{do}(\mathrm{PA}_{\mathbf{B}} = \rho)) \; \forall \; x_j \in \mathrm{dom}[\mathbf{X}]$ and $\mathrm{Cmp}(x'_i, \mathrm{do}(\mathrm{PA}_{\mathbf{B}'} = \rho')) \; \forall \; x'_i \in \mathrm{dom}[\mathbf{X}']$, which suggest that $\sum_{[i,j]\in\mathcal{O}_{\iota,\eta,\rho}} P_{i,j}^{\iota} = \sum_{[i,j]\in\mathcal{O}'_{\iota,\eta,\rho'}} P_{i,j}^{\iota} = 1$. Therefore, the normalizing vectors in Eq. (17) now become:

$$\mathcal{Z}^{\eta} = \sum_{[i,j]\in\mathcal{C}_{\iota,\eta}} P_{i,j}^{\iota}, \quad \mathcal{Z}^{\omega(\eta)} = \sum_{[i,j]\in\mathcal{C}'_{\iota,\eta}} P_{i,j}^{\iota} \tag{18}$$

The later relation conveys that in the case of *no parents* for the intervened variable the normalizing vectors become constant vectors for each model and are not different for each of the $x_j \in \mathrm{dom}[\mathbf{X}]$ and $x'_i \in \mathrm{dom}[\mathbf{X}']$ respectively, compared to the general case where we have a unique normalizing constant for each of these entries.

Appendix B. Proof of Theorem 1

Let $D = \mathrm{dom}[\mathbf{X}]$ and $D' = \mathrm{dom}[\mathbf{X}']$. Let $N(k) = N$ (the number of plans depends on the chain k) for simplicity. Hence we have N plans in a particular chain $\mathcal{I}_k$. Also for simplicity of notation, let $P^{\iota n} = P_n, C_\omega = C$. We show the proof for a generic chain, and it holds for any chain in the set of chains. To prove the theorem, we need to show:

1. The set $\underbrace{\mathcal{S}_{D\times D'} \times \ldots \mathcal{S}_{D\times D'}}_{N\text{times}} = \{(P_1, \ldots, P_N) \mid P_n \in \mathcal{S}_{D\times D'}, \text{ for all } k = 1, \ldots, N\}$, where $\mathcal{S}_{D\times D'}$ is the set of stochastic matrices of dimension $D \times D'$, is a convex set.

2. The constraints $\{P_n \in \mathcal{U}(\pi_n), n = 1, \ldots, N\}$ are convex.

3. The function $(P_1, \ldots, P_N) \to \sum_{n=1}^{N} \kappa \cdot \left\langle C, P_n \right\rangle + \boldsymbol{\lambda}^\top \mathcal{D}\left(P_n; P_{n+1}\right) + \mu \cdot \mathcal{H}(P_n)$ is jointly convex in $P_1, \ldots, P_N$. Denote this function by $f^{\text{COTA}}(P_1, \ldots, P_N)$.

Point (1) is a consequence of the fact that each P_n is in the set of stochastic matrices $\mathcal{S}_{D \times D'}$, which is a convex set, and that the Cartesian product $\times$ preserves convexity. Point (2) follows since each $\mathcal{U}(\pi_n)$ is a convex set (Peyré et al., 2019). To prove (3), we are going to use the following facts.

Lemma *The domain of each P_n is a convex set, and the inner product $P_n \to \left\langle C, P_n \right\rangle$ is a convex function in P_n.*

Proof: the domain of P_n is the stochastic matrices as described in point (1) above; the inner product is a convex function in P_n (Boyd and Vandenberghe, 2004) ∎
Further, $\boldsymbol{\lambda}^\top \mathcal{D}(P_n; P_{n+1})$ is jointly convex in the pair (P_n, P_{n+1}), which the next lemma shows.

Lemma *The function $(P_n, P_{n+1}) \to \boldsymbol{\lambda}^\top \mathcal{D}(P_n; P_{n+1})$ is jointly convex in the pair (P_n, P_{n+1})*

$$\mathcal{D}(P_n, P_{n+1}) = [\delta_{n,n+1}, \delta'_{n,n+1}]^\top$$

Given that:

1. P_n, P_{n+1} are stochastic matrices, i.e., they belong to the set $\mathcal{S}_{D \times D'}$.

2. d is a Bregman divergence between probability vectors, which is jointly convex in both its inputs (Dhillon and Tropp, 2008).

3. $\delta(P_n, P_{n+1}) = d(P_n \mathbf{1}, P_{n+1} \mathbf{1})$, $\delta'(P_n, P_{n+1}) = d(P_n^\top \mathbf{1}, P_{n+1}^\top \mathbf{1})$ where $\mathbf{1}$ is a vector of ones of appropriate dimension. Note that $P_n \mathbf{1}$ and $P_n^\top \mathbf{1}$ are the marginals of P_n.

We want to prove that $\boldsymbol{\lambda}^\top \mathcal{D}(P_n, P_{n+1})$ is jointly convex in P_n and P_{n+1}, i.e., for P'_n, P'_{n+1} any stochastic matrices of the corresponding sizes and $\gamma \in [0, 1]$, it holds (due also to linearity of matrix multiplication)

$$\boldsymbol{\lambda}^\top \mathcal{D}(\gamma P_n + (1-\gamma) P'_n, \gamma P_{n+1} + (1-\gamma) P'_{n+1}) \leq \gamma \boldsymbol{\lambda}^\top \mathcal{D}(P_n, P_{n+1}) + (1-\gamma) \boldsymbol{\lambda}^\top \mathcal{D}(P'_n, P'_{n+1}). \quad (19)$$

It is now sufficient to prove that both $\delta(P_n, P_{n+1}), \delta'(P_n, P_{n+1})$ are jointly convex in P_n and P_{n+1}, and convexity of $\boldsymbol{\lambda}^\top \mathcal{D}(\cdot, \cdot)$ follows. We prove the result for δ and an analogous proof holds for δ'.

Expanding:

$$\delta(\gamma P_n + (1-\gamma) P'_n, \gamma P_{n+1} + (1-\gamma) P'_{n+1}) = d((\gamma P_n + (1-\gamma) P'_n)\mathbf{1}, (\gamma P_{n+1} + (1-\gamma) P'_{n+1})\mathbf{1}),$$

given that d is a Bregman divergence and is jointly convex, for any probability vectors $\alpha, \alpha', \beta, \beta'$:

$$d(\gamma \alpha + (1-\gamma)\alpha', \gamma\beta + (1-\gamma)\beta') \leq \gamma d(\alpha, \beta) + (1-\gamma)d(\alpha', \beta')$$

Setting $\alpha = P_n \mathbf{1}$, $\alpha' = P'_n \mathbf{1}$, $\beta = P_{n+1} \mathbf{1}$, and $\beta' = P'_{n+1} \mathbf{1}$, we get (simply by linearity of matrix-vector multiplication):

$$d(\gamma P_n \mathbf{1} + (1-\gamma) P'_n \mathbf{1}, \gamma P_{n+1} \mathbf{1} + (1-\gamma) P'_{n+1} \mathbf{1}) \leq \gamma d(P_n \mathbf{1}, P_{n+1} \mathbf{1}) + (1-\gamma) d(P'_n \mathbf{1}, P'_{n+1} \mathbf{1}).$$

This is equivalent to:

$$\delta(\gamma P_n + (1-\gamma)P'_n, \gamma P_{n+1} + (1-\gamma)P'_{n+1}) \leq \gamma\delta(P_n, P_{n+1}) + (1-\gamma)\delta(P'_n, P'_{n+1}).$$

Thus, if d is a Bregman divergence, making it therefore jointly convex in its two vector inputs, then δ is jointly convex in its two matrix inputs P_n and P_{n+1}. ∎

Now, we resort to the definition of joint convexity of a function $f : \underbrace{\mathcal{S}_{D\times D'} \times \ldots \mathcal{S}_{D\times D'}}_{N\text{times}} \to \mathbb{R}_{\geq 0}$ in the tuple of matrices $(P_1, \ldots, P_N) \in \mathcal{S}_{D\times D'} \times \ldots \mathcal{S}_{D\times D'}$.

Definition 1 A function f is jointly convex in $P_1, \ldots, P_N$ iff, for $\gamma \in [0,1]$ and the domain of $(P_1, \ldots, P_N)$ is a convex set, it holds

$$f(\gamma P_1 + (1-\gamma)P'_1, \ldots, \gamma P_N + (1-\gamma)P'_N) \leq \gamma f(P_1, \ldots, P_N) + (1-\gamma)f(P'_1, \ldots, P'_N). \quad (20)$$

We will show that for $f^{\text{COTA}}(P_1, \ldots, P_N)$, joint convexity follows from convexity of the inner product in P_n and joint convexity in the pairs (P_n, P_{n+1}) of the distance.

By taking the convex combinations $\gamma P_1 + (1-\gamma)P'_1, \ldots, \gamma P_N + (1-\gamma)P'_N$ and plugging it in, we get:

$$f^{\text{COTA}}(\gamma P_1 + (1-\gamma)P'_1, \ldots, \gamma P_N + (1-\gamma)P'_N) \quad (21)$$

$$= \sum_{n=1}^{N} \kappa \cdot \left\langle C, \gamma P_n + (1-\gamma)P'_n \right\rangle + \boldsymbol{\lambda}^{\top}\mathcal{D}(\gamma P_n + (1-\gamma)P'_n \,;\, \gamma P_{n+1} + (1-\gamma)P'_{n+1}) \quad (22)$$

$$+ \mu \cdot \mathcal{H}(\gamma P_n + (1-\gamma)P'_n).$$

By combining the fact that the inner product is convex in P_n so $\left\langle C, \gamma P_n + (1-\gamma)P'_n \right\rangle \leq \gamma\left\langle C, P_n \right\rangle + (1-\gamma)\left\langle C, P'_n \right\rangle$, the inequality from Eq. (19), and the (strict) convexity of $\mathcal{H}$, we have

$$f^{\text{COTA}}(\gamma P_1 + (1-\gamma)P'_1, \ldots, \gamma P_N + (1-\gamma)P'_N) \quad (23)$$

$$\leq \sum_{n=1}^{N} \kappa \cdot \left(\gamma\left\langle C, P_n \right\rangle + (1-\gamma)\left\langle C, P'_n \right\rangle\right) + \boldsymbol{\lambda}^{\top}\left(\gamma\mathcal{D}(P_n; P_{n+1}) + (1-\gamma)\mathcal{D}(P'_n; P'_{n+1})\right)$$

$$+ \mu \cdot \left(\gamma\mathcal{H}(P_n) + (1-\gamma)\mathcal{H}(P'_n)\right)$$

$$= \gamma\left(\sum_{n=1}^{N} \kappa\left\langle C, P_n \right\rangle + \lambda\mathcal{D}(P_n; P_{n+1}) + \mu\mathcal{H}(P_n)\right) + (1-\gamma)\sum_{n=1}^{N} \kappa\left\langle C, P'_n \right\rangle + \boldsymbol{\lambda}^{\top}\mathcal{D}(P'_n; P'_{n+1}) + \mu\mathcal{H}(P'_n)$$

$$= \gamma f^{\text{COTA}}(P_1, \ldots, P_N) + (1-\gamma)f^{\text{COTA}}(P'_1, \ldots, P'_N), \text{ which is the definition of convexity in } P_1, \ldots, P_N. \blacksquare$$

Appendix C. `OT` costs

In this section we discuss the mechanics underlying the computations of the OT costs. First, we recall the ω-cost function formula as it is introduced in the Eq. (11). Given two samples $\mathbf{x} \in \text{dom}[\mathbf{X}]$ and $\mathbf{x}' \in \text{dom}[\mathbf{X}']$, and an intervention set $\mathcal{I}$ together with $\omega : \mathcal{I} \to \mathcal{I}'$, we define the ω-cost as the following function:

$$c_\omega(\mathbf{x}, \mathbf{x}') = |\mathcal{I}| - \sum_{\iota\in\mathcal{I}} \mathbb{1}\left[\text{Cmp}(\mathbf{x}, \iota) \wedge \text{Cmp}(\mathbf{x}', \omega(\iota))\right]$$
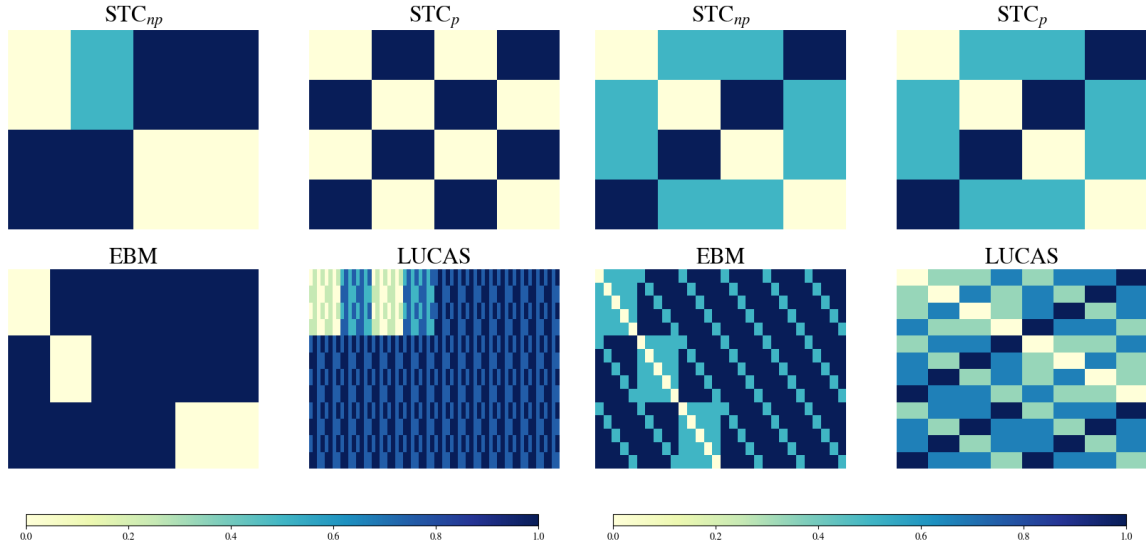
Figure 4: Omega Cost (C_ω)

Figure 5: Hamming cost ($C_\mathcal{H}$)

The function c_ω discounts the cost of transporting samples $\mathbf{x}$ to $\mathbf{x}'$ proportionally to the number of pairs $(\iota, \omega(\iota))$ w.r.t. which $\mathbf{x}$ and $\mathbf{x}'$ are compatible. In Fig. 3 we visualise the construction of the cost matrix C_ω based on the $c_\omega(\cdot, \cdot)$ function of Eq. (11).

As shown in Section 6, for c_ω to be an effective cost function, the intervention set needs to be well-specified and informative regarding the domains of the two models. In such a case, a meaningful cost matrix can be derived and more elements on the induced transport plans will be influenced by causal knowledge. Our experiments confirmed that in scenarios where the intervention set was limited, conventional costs like Hamming $c_\mathcal{H}$ could return lower abstraction errors compared to c_ω. The reason is that, by construction, c_ω will assign maximum values ($|\mathcal{I}|$) for samples for which it has no available information from the ω map. On the other hand, costs like $c_\mathcal{H}$ which spreads its values across the whole domain might be able to capture certain patterns more efficiently. We provide a visualization of the cost matrices for both c_ω and $c_\mathcal{H}$ functions in Fig. 4 and Fig. 5 respectively.

$$
\begin{array}{c}
\begin{array}{cccccc}
\mathbf{x}_1 = 000 & \mathbf{x}_2 = 001 & \mathbf{x}_3 = 010 & \mathbf{x}_4 = 011 & \mathbf{x}_5 = 110 & \mathbf{x}_6 = 110
\end{array}\\
\begin{array}{c}
\mathbf{x}'_1 = 00\\
\mathbf{x}'_2 = 01\\
\mathbf{x}'_3 = 10\\
\mathbf{x}'_4 = 11
\end{array}
\left[
\begin{array}{cccccc}
0 & 0 & 1 & 1 & 2 & 2\\
0 & 0 & 1 & 1 & 2 & 2\\
2 & 2 & 2 & 2 & 2 & 2\\
2 & 2 & 2 & 2 & 2 & 2
\end{array}
\right]
\end{array}
\quad
\begin{array}{l}
\iota_1 : \mathrm{do}(\emptyset) \overset{\omega}{\mapsto} \mathrm{do}(\emptyset)\\[4pt]
\iota_2 : \mathrm{do}(\mathbf{X}[1]=0) \overset{\omega}{\mapsto} \mathrm{do}(\mathbf{X}'[1]=0)\\[4pt]
\iota_3 : \mathrm{do}(\mathbf{X}[1]=0, \mathbf{X}[2]=0) \overset{\omega}{\mapsto}\\
\qquad \mathrm{do}(\mathbf{X}'[1]=0)
\end{array}
$$

Figure 3: ω-informed cost matrix computer with $|\mathcal{I}| = 3$. Notice, how all values $\mathbf{x}$ and $\mathbf{x}'$ are compatible with ι_1 and $\omega(\iota_1)$, while only $\mathbf{x}_1, \mathbf{x}_2$ are compatible with $\mathrm{do}(\mathbf{X}[1]=0, \mathbf{X}[2]=0)$ and $\mathbf{x}'_1$ is compatible with $\mathrm{do}(\mathbf{X}'[1]=0)$

Appendix D. EBM Downstream Task

In our **EBM** scenario we rely on data about battery coating released by the Laboratoire de Réactivité et Chimie des Solides (LRCS) (Cunha et al., 2020) and by the Warwick Manufacturing Group (WMG) (Zennaro et al., 2023). These datasets contain observations about different variables affecting the coating process (*Comma Gap, Mass Loading Position*), as well as observation about a key outcome variable related to the width of the coating (*Mass Loading*). Both groups aim at inferring a machine learning model that would allow them to control the target variable via interventions on the other variables.

Closely following Zennaro et al. (2023), we learn abstractions from the WMG model to the LRCS model in order to merge data collected by the two laboratories and learn a model from the aggregate dataset. We compare our approach COTA with the competing CA learning method proposed in (Zennaro et al., 2023). The latter builds upon the α-abstraction framework established by Rischel (2020), which is briefly summarized in Appendix D.1. To assess the usefulness of abstraction, we solve three extrapolation tasks meant to show how fitting a simple model to the aggregated dataset produced via abstraction guarantees better predictive results. We consider the following three downstream tasks:

- In the first task, we train a regression model on all LRCS samples except for samples belonging to class k. We then test the regression model on LRCS samples belonging to class k. This constitutes the baseline model learned on data from a single laboratory.

- In the second task, we train a regression model on all LRCS samples except for samples belonging to class k and all the abstracted WMG samples. We then test the regression model on LRCS samples belonging to class k. This constitutes a scenario in which we enrich one dataset (LRCS) with data from another laboratory (WMG); moreover, the enriching data provides information on the class k which was not originally observerd in LRCS.

- In the third task, we train a regression model on all LRCS and abstracted WMG samples except for samples belonging to class k. We then test the regression model on LRCS and WMG samples belonging to class k. This constitutes a scenario in which we enrich one dataset (LRCS) with data from another laboratory (WMG); however, the enriching data also lacks observations for the class k.

Our solution outperforms the SOTA in terms of MSE, and confirms that using abstractions to aggregate data may be beneficial for downstream tasks, such as regression tasks. This is particularly true in settings where data are limited because of the cost and complexity of collecting samples, such as in the case of battery manufacturing Niri et al. (2022).

D.1. The α-abstraction framework

This framework draws inspiration from category theory, and assumes two SCMs $\mathcal{M} = \langle \mathbf{X}, \mathbf{U}, \mathcal{F}, \mathbb{P}(\mathbf{U}) \rangle$, $\mathcal{M}' = \langle \mathbf{X}', \mathbf{U}', \mathcal{F}', \mathbb{P}'(\mathbf{U}') \rangle$ with finite sets of endogenous variables, where each variable defined on a finite and discrete domain.

Definition *[Rischel (2020)] Given two SCMs $\mathcal{M}$ and $\mathcal{M}'$, an abstraction α is a tuple $\langle R, a, \alpha \rangle$ where:*

- $R \subseteq \mathbf{X}$ *is a subset of relevant variables in the model $\mathcal{M}$.*

- $a : R \to \mathbf{X}'$ *is a surjective map between variables, from nodes in $\mathcal{M}$ to nodes in $\mathcal{M}'$.*
- α *is a collection of surjective maps* $\alpha_{X'} : \mathrm{dom}[a^{-1}(X')] \to \mathrm{dom}[X']$ *where* $X \subseteq \mathbf{X}$ *and* $X' \subseteq \mathbf{X}'$.

An α-abstraction establishes an asymmetric relation from a base model $\mathcal{M}$ to an abstracted model $\mathcal{M}'$. This definition encodes a mapping on two layers: on a structural or graphical level between the nodes of the DAGs via a, and on a distributional level via the maps $\alpha_{X'}$.

Rischel (2020) introduces a notion of interventional consistency between base and abstracted model, whereby interventional distributions produced in the base and abstracted model are related via the abstraction α; furthermore, a notion of abstraction error, analogous to the one used in this paper, is also proposed. It is then immediate to relate the τ-abstraction and α-abstraction as they both imply comparing distributions generated by base and abstracted model through an abstraction map.

Appendix E. Additional Experiments and Analysis

Here we report the complete experimental results of our simulations. These results corroborate our understanding of the effectiveness of COTA compared to the baseline methods. Regarding the parameter $\boldsymbol{\lambda}^\top$, in Appendix E.1 we demonstrate the complete results from the the equal weight case where $\lambda = \lambda'$ and in Appendix E.2 from the more general case of $\lambda \neq \lambda'$. Notice, that throughout the results, how in $\mathbf{STC_{np}}$ and $\mathbf{LUCAS}$ ω-cost c_ω returns lower abstraction errors compared to Hamming $c_{\mathcal{H}}$, whereas in $\mathbf{STC_p}$ and $\mathbf{EBM}$ the opposite holds true due to the smaller and less diverse intervention sets, as discussed in the main text and in Appendix C.

E.1. Equal weights $(\lambda = \lambda')$.

We present the evaluation results in the case where the parameter $\boldsymbol{\lambda}^\top$ governing the causal constraint term in the COTA optimization objective of Eq. (12) is a constant vector i.e. $\boldsymbol{\lambda}^\top \cdot \mathcal{D}(P^\iota, P^\eta) = [\lambda, \lambda] \cdot \begin{bmatrix} \delta_{\iota,\eta} \\ \delta'_{\iota,\eta} \end{bmatrix} = \lambda \cdot (\delta_{\iota,\eta} + \delta'_{\iota,\eta})$. The complete evaluation for the $\mathbf{STC}$ examples is detailed in Table 6 and Table 8, while results for the $\mathbf{LUCAS}$ and the $\mathbf{EBM}$ examples are presented in Table 10 and Table 12 respectively.

E.2. Different weights $(\lambda \neq \lambda')$.

We present the evaluation results in the case where the parameter $\boldsymbol{\lambda}^\top$ governing the causal constraint term in the COTA optimization objective of Eq. (12) is a non-constant vector i.e. $\boldsymbol{\lambda}^\top \cdot \mathcal{D}(P^\iota, P^\eta) = [\lambda \ \lambda'] \cdot \begin{bmatrix} \delta_{\iota,\eta} \\ \delta'_{\iota,\eta} \end{bmatrix} = \lambda \cdot \delta_{\iota,\eta} + \lambda' \cdot \delta'_{\iota,\eta}$. The complete evaluation for the $\mathbf{STC}$ examples is detailed in Table 7 and Table 9, while results for the $\mathbf{LUCAS}$ and the $\mathbf{EBM}$ examples are presented in Table 11 and Table 13, respectively. Finally, Table 14 demonstrates the MSE of COTA and a SOTA CA framework on a regression task for $\mathbf{EBM}$.

Appendix F. `Approximate COTA`

We can simplify the optimization problem and halve the number of constraints by assuming that the elements of each marginal of the plan inherit the marginal's normalising factor (Occam's razor). This way, we turn $\delta_{\iota,\eta}$ and $\delta'_{\iota,\eta}$ into the element-wise (between the plans) distances

$d\left(\left(\frac{1}{\mathcal{Z}_j^\eta}P_{ij}^\iota\right)_{ij}, \left(P_{ij}^\eta\right)_{ij}\right)$ and $d\left(\left(\frac{1}{\mathcal{Z}_i^{\omega(\eta)}}P_{ij}^\iota\right)_{ij}, \left(P_{ij}^\eta\right)_{ij}\right)$, respectively. Now every pair of elements $P_{ij}^\iota, P_{ij}^\eta$ is related simultaneously through $\mathcal{Z}_j^\eta$ and $\mathcal{Z}_i^{\omega(\eta)}$. Then given $d : \mathbb{R}^{D\times D'} \times \mathbb{R}^{D\times D'} \to \mathbb{R}_{\geq 0}$, we re-express our constraints in a matrix form as:

$$\mathcal{D}(P^\iota, P^\eta) = d\left(\frac{1}{\varphi(\mathcal{Z}_j^\eta, \mathcal{Z}_i^{\omega(\eta)})}P^\iota, P^\eta\right) \quad \text{if } \mathrm{Cmp}(x_j, \eta) \text{ and } \mathrm{Cmp}(x_i', \omega(\eta)) \tag{24}$$

where $\varphi(\cdot, \cdot) : \mathbb{R}^2 \to \mathbb{R}$ is any aggregating function that preserves the correct support of the plans. In our experiments we work with $\varphi(\mathcal{Z}_j^\eta, \mathcal{Z}_i^{\omega(\eta)}) = \min(\mathcal{Z}_j^\eta, \mathcal{Z}_i^{\omega(\eta)})$. We provide the complete abstraction error evaluation results as before alongside the downstream evaluation for the `Approximate COTA` in Appendix F.1.

F.1. `Approximate COTA` Results

We report the complete experimental results for `Approximate COTA`. Once again, we confirm our understanding of the effectiveness of `COTA` compared to the baseline methods. The complete evaluation for the **STC** examples is detailed in Table 15 and Table 16, while results for the **LUCAS** and the **EBM** examples are presented in Table 17 and Table 18, respectively. Furthermore, in $\mathbf{STC}_{\mathrm{np}}$ and **LUCAS** ω-cost c_ω returns lower abstraction errors compared to Hamming $c_\mathcal{H}$, whereas in $\mathbf{STC}_{\mathrm{p}}$ and **EBM** the opposite holds true due to the smaller and less diverse intervention sets, as discussed in the main text and in Appendix C. In addition, in Fig. 6 and Fig. 7 we provide the equivalent simplex plots for $\mathbf{STC}_{\mathrm{np}}$, which illustrate the influence of the parameter λ in the optimization problem. Once again, these plots reinforce the idea that optimal solutions are achieved when λ is greater than 0. Finally, in Table 19, we demonstrate the results regarding the MSE for the downstream task on the **EBM** dataset as before.

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\mathrm{JSD}}(\tau)$	$e_{\mathrm{WASS}}(\tau)$
COTA($\widehat{P}$)	FRO	c_ω	**0.010 ± 0.005**	**0.011 ± 0.003**
		$c_\mathcal{H}$	0.087 ± 0.007	0.025 ± 0.001
	JSD	c_ω	0.012 ± 0.006	0.012 ± 0.003
		$c_\mathcal{H}$	0.087 ± 0.006	0.025 ± 0.001
COTA($\widehat{\tau}$)	FRO	c_ω	0.013 ± 0.021	0.171 ± 0.001
		$c_\mathcal{H}$	0.169 ± 0.005	0.178 ± 0.001
	JSD	c_ω	0.014 ± 0.021	0.171 ± 0.001
		$c_\mathcal{H}$	0.169 ± 0.004	0.178 ± 0.001
Pwise OT	–	c_ω	0.013 ± 0.002	0.011 ± 0.002
	–	$c_\mathcal{H}$	0.093 ± 0.004	0.039 ± 0.002
Map OT	–	c_ω	0.023 ± 0.022	0.147 ± 0.001
	–	$c_\mathcal{H}$	0.169 ± 0.022	0.156 ± 0.001
Bary OT	–	c_ω	0.233 ± 0.142	0.067 ± 0.042
	–	$c_\mathcal{H}$	0.323 ± 0.074	0.095 ± 0.039

Table 6: Abstraction error evaluation for $\mathbf{STC}_{\mathrm{np}}$ of the $\lambda = \lambda'$ case. The COTA($\widehat{P}$) $-$ FRO $- c_\omega$ returns the lower abstraction error for both metrics. The ω-cost c_ω outperforms $c_\mathcal{H}$.

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\mathrm{JSD}}(\tau)$	$e_{\mathrm{WASS}}(\tau)$
COTA($\widehat{P}$)	FRO	c_ω	**0.010 ± 0.006**	**0.010 ± 0.003**
		$c_\mathcal{H}$	0.087 ± 0.006	0.025 ± 0.001
	JSD	c_ω	0.011 ± 0.006	0.012 ± 0.003
		$c_\mathcal{H}$	0.087 ± 0.006	0.025 ± 0.001
COTA($\widehat{\tau}$)	FRO	c_ω	0.012 ± 0.020	0.171 ± 0.001
		$c_\mathcal{H}$	0.169 ± 0.004	0.178 ± 0.001
	JSD	c_ω	0.013 ± 0.019	0.171 ± 0.001
		$c_\mathcal{H}$	0.169 ± 0.004	0.178 ± 0.001
Pwise OT	–	c_ω	0.013 ± 0.002	0.011 ± 0.002
	–	$c_\mathcal{H}$	0.093 ± 0.004	0.039 ± 0.002
Map OT	–	c_ω	0.023 ± 0.022	0.147 ± 0.001
	–	$c_\mathcal{H}$	0.169 ± 0.022	0.156 ± 0.001
Bary OT	–	c_ω	0.233 ± 0.142	0.067 ± 0.042
	–	$c_\mathcal{H}$	0.323 ± 0.074	0.095 ± 0.039

Table 7: Abstraction error evaluation for $\mathbf{STC}_{\mathrm{np}}$ of the $\lambda \neq \lambda'$ case. The COTA($\widehat{P}$) $-$ FRO $- c_\omega$ returns the lower abstraction error for both metrics. The ω-cost c_ω outperforms $c_\mathcal{H}$.

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\mathrm{JSD}}(\tau)$	$e_{\mathrm{WASS}}(\tau)$
COTA($\widehat{P}$)	FRO	c_ω	0.278 ± 0.015	**0.048 ± 0.007**
		$c_\mathcal{H}$	0.241 ± 0.003	**0.048 ± 0.007**
	JSD	c_ω	0.258 ± 0.027	0.054 ± 0.003
		$c_\mathcal{H}$	0.242 ± 0.001	0.054 ± 0.003
COTA($\widehat{\tau}$)	FRO	c_ω	0.249 ± 0.005	0.135 ± 0.001
		$c_\mathcal{H}$	**0.229 ± 0.003**	0.129 ± 0.001
	JSD	c_ω	0.241 ± 0.008	0.135 ± 0.001
		$c_\mathcal{H}$	**0.229 ± 0.004**	0.129 ± 0.001
Pwise OT	–	ω	0.279 ± 0.014	0.091 ± 0.005
	–	$\mathcal{H}$	0.242 ± 0.002	0.067 ± 0.001
Map OT	–	ω	0.250 ± 0.005	0.140 ± 0.001
	–	$\mathcal{H}$	0.229 ± 0.004	0.129 ± 0.001
Bary OT	–	ω	0.318 ± 0.205	0.104 ± 0.061
	–	$\mathcal{H}$	0.272 ± 0.212	0.075 ± 0.058

Table 8: Abstraction error evaluation for $\mathbf{STC}_{\mathrm{p}}$ of the $\lambda = \lambda'$ case. The $c_\mathcal{H}$ settings of COTA($\widehat{\tau}$) return the lower abstraction error for the JSD metric and COTA($\widehat{P}$) $-$ FRO for the WASS metric. $c_\mathcal{H}$ outperforms c_ω due to under-specification of the intervention set.

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\mathrm{JSD}}(\tau)$	$e_{\mathrm{WASS}}(\tau)$
COTA($\widehat{P}$)	FRO	c_ω	0.277 ± 0.018	**0.046 ± 0.007**
		$c_\mathcal{H}$	0.241 ± 0.002	**0.046 ± 0.007**
	JSD	c_ω	0.273 ± 0.009	0.047 ± 0.006
		$c_\mathcal{H}$	0.241 ± 0.003	0.047 ± 0.006
COTA($\widehat{\tau}$)	FRO	c_ω	0.249 ± 0.003	0.135 ± 0.004
		$c_\mathcal{H}$	**0.228 ± 0.004**	0.129 ± 0.001
	JSD	c_ω	0.247 ± 0.004	0.135 ± 0.004
		$c_\mathcal{H}$	**0.228 ± 0.003**	0.129 ± 0.001
Pwise OT	–	ω	0.279 ± 0.014	0.091 ± 0.005
	–	$\mathcal{H}$	0.242 ± 0.002	0.067 ± 0.001
Map OT	–	ω	0.250 ± 0.005	0.140 ± 0.001
	–	$\mathcal{H}$	0.229 ± 0.004	0.129 ± 0.001
Bary OT	–	ω	0.318 ± 0.205	0.104 ± 0.061
	–	$\mathcal{H}$	0.272 ± 0.212	0.075 ± 0.058

Table 9: Abstraction error evaluation for $\mathbf{STC}_{\mathrm{p}}$ of the $\lambda \neq \lambda'$ case. The $c_\mathcal{H}$ settings of COTA($\widehat{\tau}$) return the lower abstraction error for the JSD metric and COTA($\widehat{P}$) $-$ FRO for the WASS metric. $c_\mathcal{H}$ outperforms c_ω due to under-specification of the intervention set.

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\text{JSD}}(\tau)$	$e_{\text{WASS}}(\tau)$
COTA($\widehat{P}$)	FRO	c_ω	0.287 ± 0.014	$\mathbf{0.044 \pm 0.001}$
		$c_\mathcal{H}$	0.287 ± 0.014	0.047 ± 0.001
	JSD	c_ω	0.286 ± 0.014	0.048 ± 0.001
		$c_\mathcal{H}$	0.287 ± 0.014	0.048 ± 0.001
COTA($\widehat{\tau}$)	FRO	c_ω	$\mathbf{0.263 \pm 0.005}$	0.061 ± 0.001
		$c_\mathcal{H}$	$\mathbf{0.263 \pm 0.006}$	0.061 ± 0.001
	JSD	c_ω	$\mathbf{0.263 \pm 0.005}$	0.062 ± 0.001
		$c_\mathcal{H}$	$\mathbf{0.263 \pm 0.006}$	0.062 ± 0.001
Pwise OT	$-$	c_ω	0.306 ± 0.009	0.045 ± 0.001
	$-$	$c_\mathcal{H}$	0.387 ± 0.002	0.047 ± 0.001
Map OT	$-$	c_ω	0.294 ± 0.008	0.054 ± 0.001
	$-$	$c_\mathcal{H}$	0.350 ± 0.005	0.054 ± 0.001
Bary OT	$-$	c_ω	0.294 ± 0.047	0.044 ± 0.003
	$-$	$c_\mathcal{H}$	0.414 ± 0.040	0.046 ± 0.010

Table 10: Abstraction error evaluation for **LU-CAS** of the $\lambda = \lambda'$ case. All settings of COTA($\widehat{\tau}$) return the lowest abstraction error for the JSD metric and the COTA($\widehat{P}$) $-$ FRO $- c_\omega$ setting for the WASS metric. ω-cost c_ω outperforms $c_\mathcal{H}$.

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\text{JSD}}(\tau)$	$e_{\text{WASS}}(\tau)$
COTA($\widehat{P}$)	FRO	c_ω	0.287 ± 0.014	$\mathbf{0.044 \pm 0.001}$
		$c_\mathcal{H}$	0.287 ± 0.014	0.047 ± 0.001
	JSD	c_ω	0.287 ± 0.014	0.045 ± 0.001
		$c_\mathcal{H}$	0.287 ± 0.014	0.047 ± 0.001
COTA($\widehat{\tau}$)	FRO	c_ω	$\mathbf{0.263 \pm 0.005}$	0.060 ± 0.001
		$c_\mathcal{H}$	$\mathbf{0.263 \pm 0.006}$	0.060 ± 0.002
	JSD	c_ω	$\mathbf{0.263 \pm 0.005}$	0.061 ± 0.001
		$c_\mathcal{H}$	$\mathbf{0.263 \pm 0.006}$	0.061 ± 0.001
Pwise OT	$-$	c_ω	0.306 ± 0.009	0.045 ± 0.001
	$-$	$c_\mathcal{H}$	0.387 ± 0.002	0.047 ± 0.001
Map OT	$-$	c_ω	0.294 ± 0.008	0.054 ± 0.001
	$-$	$c_\mathcal{H}$	0.350 ± 0.005	0.054 ± 0.001
Bary OT	$-$	c_ω	0.294 ± 0.047	0.044 ± 0.003
	$-$	$c_\mathcal{H}$	0.414 ± 0.040	0.046 ± 0.010

Table 11: Abstraction error evaluation for **LU-CAS** of the $\lambda \neq \lambda'$ case. All settings of COTA($\widehat{\tau}$) return the lowest abstraction error for the JSD metric and the COTA($\widehat{P}$) $-$ FRO $- c_\omega$ setting for the WASS metric. ω-cost c_ω outperforms $c_\mathcal{H}$.

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\texttt{JSD}}(\tau)$	$e_{\texttt{WASS}}(\tau)$
$\texttt{COTA}(\widehat{P})$	FRO	c_ω	0.379	**0.006**
		$c_{\mathcal{H}}$	0.263	**0.006**
	JSD	c_ω	0.379	**0.006**
		$c_{\mathcal{H}}$	0.263	**0.006**
$\texttt{COTA}(\widehat{\tau})$	FRO	c_ω	0.379	0.053
		$c_{\mathcal{H}}$	**0.220**	0.053
	JSD	c_ω	0.399	0.053
		$c_{\mathcal{H}}$	0.263	0.053
Pwise OT	–	c_ω	0.430	0.027
	–	$c_{\mathcal{H}}$	0.263	0.027
Map OT	–	c_ω	0.408	0.060
	–	$c_{\mathcal{H}}$	0.228	0.053
Bary OT	–	c_ω	0.530	0.019
	–	$c_{\mathcal{H}}$	0.335	0.023

Table 12: Abstraction error evaluation for **EBM** of the $\lambda = \lambda'$ case. The $\texttt{COTA}(\widehat{\tau}) - \texttt{FRO} - c_{\mathcal{H}}$ formulation returns the lower abstraction error for the JSD metric and all the settings of $\texttt{COTA}(\widehat{P})$ for the WASS metric. $c_{\mathcal{H}}$ outperforms c_ω due to under-specification of the intervention set.

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\texttt{JSD}}(\tau)$	$e_{\texttt{WASS}}(\tau)$
$\texttt{COTA}(\widehat{P})$	FRO	c_ω	0.311	**0.006**
		$c_{\mathcal{H}}$	0.263	**0.006**
	JSD	c_ω	0.311	**0.006**
		$c_{\mathcal{H}}$	0.263	**0.006**
$\texttt{COTA}(\widehat{\tau})$	FRO	c_ω	0.369	0.053
		$c_{\mathcal{H}}$	**0.219**	0.053
	JSD	c_ω	0.389	0.053
		$c_{\mathcal{H}}$	0.220	0.053
Pwise OT	–	c_ω	0.430	0.027
	–	$c_{\mathcal{H}}$	0.263	0.027
Map OT	–	c_ω	0.408	0.060
	–	$c_{\mathcal{H}}$	0.228	0.053
Bary OT	–	c_ω	0.530	0.019
	–	$c_{\mathcal{H}}$	0.335	0.023

Table 13: Aabstraction error evaluation for **EBM** of the $\lambda \neq \lambda'$ case. The $\texttt{COTA}(\widehat{\tau}) - \texttt{FRO} - c_{\mathcal{H}}$ formulation returns the lower abstraction error for the JSD metric and all the settings of $\texttt{COTA}(\widehat{P})$ for the WASS metric. $c_{\mathcal{H}}$ outperforms c_ω due to under-specification of the intervention set.

Training set	Test set	Zennaro et al. (2023)	COTA
LRCS$[CG \neq k]$	LRCS$[CG = k]$	1.86 ± 1.75	**1.40 ± 1.39**
LRCS$[CG \neq k]$ +WMG	LRCS$[CG = k]$	0.22 ± 0.26	**0.20 ± 0.02**
LRCS$[CG \neq k]$ +WMG$[CG \neq k]$	LRCS$[CG = k]$ WMG$[CG = k]$	1.22 ± 0.95	**0.48 ± 0.23**

Table 14: MSE of COTA with $\lambda \neq \lambda'$ and a SOTA CA framework on a regression task for **EBM**. Augmenting data via the learned abstraction reduces the average error in all different settings compared to the SOTA. We used $\texttt{COTA}(\widehat{P}) - \texttt{FRO} - c_\omega$ with the hyperparameters $(\kappa, \lambda, \mu) = (.2, .4, .3, .1)$ achieving the lowest abstraction error.

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\text{JSD}}(\tau)$	$e_{\text{WASS}}(\tau)$
COTA($\widehat{P}$)	FRO	c_ω	0.010 ± 0.005	0.010 ± 0.002
		$c_\mathcal{H}$	0.125 ± 0.003	0.011 ± 0.003
	JSD	c_ω	$\mathbf{0.008 \pm 0.001}$	$\mathbf{0.009 \pm 0.001}$
		$c_\mathcal{H}$	0.036 ± 0.011	0.016 ± 0.002
COTA($\widehat{\tau}$)	FRO	c_ω	0.013 ± 0.008	0.171 ± 0.001
		$c_\mathcal{H}$	0.096 ± 0.010	0.175 ± 0.001
	JSD	c_ω	0.013 ± 0.007	0.171 ± 0.001
		$c_\mathcal{H}$	0.145 ± 0.008	0.175 ± 0.001
Pwise OT	–	c_ω	0.013 ± 0.002	0.011 ± 0.002
	–	$c_\mathcal{H}$	0.093 ± 0.004	0.039 ± 0.002
Map OT	–	c_ω	0.023 ± 0.022	0.147 ± 0.001
	–	$c_\mathcal{H}$	0.169 ± 0.022	0.156 ± 0.001
Bary OT	–	c_ω	0.233 ± 0.142	0.067 ± 0.042
	–	$c_\mathcal{H}$	0.323 ± 0.074	0.095 ± 0.039

Table 15: Approximate COTA complete abstraction error evaluation for the $\textbf{STC}_{\text{np}}$ example. The COTA($\widehat{P}$) $-$ JSD $- c_\omega$ setting returns the lower abstraction error for both metrics. The ω-cost c_ω outperforms $c_\mathcal{H}$.

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\text{JSD}}(\tau)$	$e_{\text{WASS}}(\tau)$
COTA($\widehat{P}$)	FRO	ω	0.264 ± 0.001	0.064 ± 0.001
		$\mathcal{H}$	0.241 ± 0.003	0.060 ± 0.004
	JSD	ω	0.259 ± 0.006	$\mathbf{0.051 \pm 0.002}$
		$\mathcal{H}$	0.242 ± 0.001	$\mathbf{0.051 \pm 0.001}$
COTA($\widehat{\tau}$)	FRO	ω	0.248 ± 0.006	0.135 ± 0.001
		$\mathcal{H}$	$\mathbf{0.227 \pm 0.005}$	0.129 ± 0.001
	JSD	ω	0.236 ± 0.002	0.130 ± 0.005
		$\mathcal{H}$	$\mathbf{0.229 \pm 0.006}$	0.129 ± 0.001
Pwise OT	–	ω	0.279 ± 0.014	0.091 ± 0.005
	–	$\mathcal{H}$	0.242 ± 0.002	0.067 ± 0.001
Map OT	–	ω	0.250 ± 0.005	0.140 ± 0.001
	–	$\mathcal{H}$	0.229 ± 0.004	0.129 ± 0.001
Bary OT	–	ω	0.318 ± 0.205	0.104 ± 0.061
	–	$\mathcal{H}$	0.272 ± 0.212	0.075 ± 0.058

Table 16: Approximate COTA complete abstraction error evaluation for the $\textbf{STC}_{\text{p}}$ example. The COTA($\widehat{\tau}$) $-$ FRO $- c_\mathcal{H}$ and COTA($\widehat{\tau}$) $-$ JSD $- c_\mathcal{H}$ settings returns the lower abstraction errors for $e_{\text{JSD}}(\tau)$ and COTA($\widehat{P}$) $-$ FRO $- c_\omega$ and COTA($\widehat{P}$) $-$ JSD $- c_\mathcal{H}$ settings returns the lower abstraction errors for $e_{\text{WASS}}(\tau)$. Table illustrates that $c_\mathcal{H}$ outperforms c_ω due to under-specification of the intervention set.

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\text{JSD}}(\tau)$	$e_{\text{WASS}}(\tau)$
COTA($\widehat{P}$)	FRO	c_ω	$\mathbf{0.258 \pm 0.003}$	$\mathbf{0.044 \pm 0.001}$
		$c_\mathcal{H}$	0.260 ± 0.004	0.047 ± 0.001
	JSD	c_ω	0.285 ± 0.014	0.045 ± 0.001
		$c_\mathcal{H}$	0.285 ± 0.014	0.046 ± 0.001
COTA($\widehat{\tau}$)	FRO	c_ω	0.259 ± 0.006	0.060 ± 0.003
		$c_\mathcal{H}$	0.259 ± 0.006	0.060 ± 0.003
	JSD	c_ω	0.260 ± 0.007	0.061 ± 0.001
		$c_\mathcal{H}$	0.263 ± 0.004	0.061 ± 0.001
Pwise OT	–	c_ω	0.306 ± 0.009	0.045 ± 0.001
	–	$c_\mathcal{H}$	0.387 ± 0.002	0.047 ± 0.001
Map OT	–	c_ω	0.294 ± 0.008	0.054 ± 0.001
	–	$c_\mathcal{H}$	0.350 ± 0.005	0.054 ± 0.001
Bary OT	–	c_ω	0.294 ± 0.047	0.044 ± 0.003
	–	$c_\mathcal{H}$	0.414 ± 0.040	0.046 ± 0.010

Table 17: Approximate COTA complete abstraction error evaluation for the **LUCAS** example. The COTA($\widehat{P}$) $-$ FRO $- c_\omega$ setting returns the lower abstraction error for both metrics. The ω-cost c_ω outperforms $c_\mathcal{H}$.

Method	$\mathcal{D}$	$\mathcal{C}$	$e_{\text{JSD}}(\tau)$	$e_{\text{WASS}}(\tau)$
COTA($\widehat{P}$)	FRO	c_ω	0.378	$\mathbf{0.006}$
		$c_\mathcal{H}$	0.263	$\mathbf{0.006}$
	JSD	c_ω	0.378	$\mathbf{0.006}$
		$c_\mathcal{H}$	0.263	$\mathbf{0.006}$
COTA($\widehat{\tau}$)	FRO	c_ω	0.389	0.053
		$c_\mathcal{H}$	$\mathbf{0.226}$	0.053
	JSD	c_ω	0.389	0.053
		$c_\mathcal{H}$	$\mathbf{0.226}$	0.053
Pwise OT	–	c_ω	0.430	0.027
	–	$c_\mathcal{H}$	0.263	0.027
Map OT	–	c_ω	0.408	0.060
	–	$c_\mathcal{H}$	0.228	0.053
Bary OT	–	c_ω	0.530	0.019
	–	$c_\mathcal{H}$	0.335	0.023

Table 18: Approximate COTA complete abstraction error evaluation for the **EBM** example. The COTA($\widehat{\tau}$)$-$FRO$-c_\mathcal{H}$ and COTA($\widehat{\tau}$)$-$JSD$-c_\mathcal{H}$ settings returns the lower abstraction errors for $e_{\text{JSD}}(\tau)$ and COTA($\widehat{P}$) for all settings returns the lower abstraction errors for $e_{\text{WASS}}(\tau)$. Table illustrates that $c_\mathcal{H}$ outperforms c_ω due to under-specification of the intervention set collected from the labs.
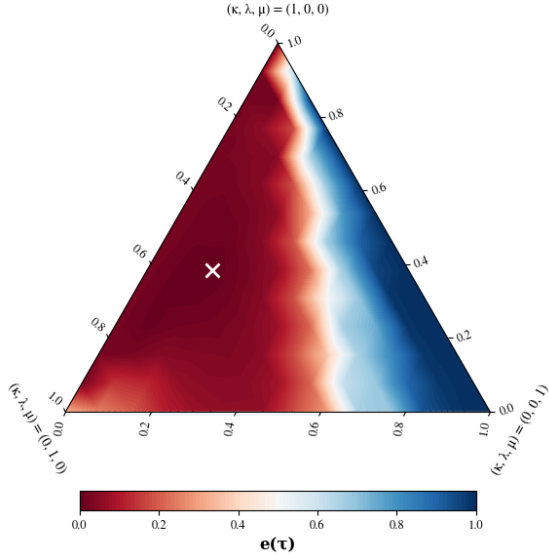
Figure 6: Effect of λ for the **STC**$_{np}$ example in the `Approximate COTA` formulation. The ternary plot illustrates a grid-search amongst 100 convex combinations of (κ, λ, μ) for the `COTA`$(\widehat{P}) - $`FRO`$ - c_\omega$ setting. Values of λ close to zero present high abstraction error, demonstrating the benefit of the *do-calculus* constraints in the `OT` problem. The minimum is reached at $(.38, .46, .16)$ and is denoted with "$\times$"
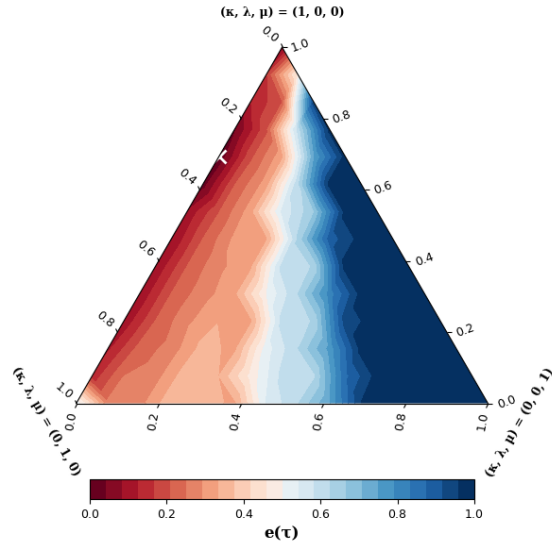
Figure 7: Effect of λ for the **STC**$_{np}$ example in the `Approximate COTA` formulation. The ternary plot illustrates a grid-search amongst 100 convex combinations of (κ, λ, μ) for the `COTA`$(\widehat{P}) - $`JSD`$ - c_\omega$ setting. Values of λ close to zero present high abstraction error, demonstrating the benefit of the *do-calculus* constraints in the `OT` problem. The minimum is reached at $(.61, .38, .01)$ and is denoted with "$\times$"

Training set	Test set	Zennaro et al. (2023)	COTA
LRCS[$CG \neq k$]	LRCS[$CG = k$]	1.86 ± 1.75	$\mathbf{1.40 \pm 1.39}$
LRCS[$CG \neq k$] +WMG	LRCS[$CG = k$]	0.22 ± 0.26	$\mathbf{0.13 \pm 0.07}$
LRCS[$CG \neq k$] +WMG[$CG \neq k$]	LRCS[$CG = k$] WMG[$CG = k$]	1.22 ± 0.95	$\mathbf{0.85 \pm 0.81}$

Table 19: MSE of `Approximate COTA` and a SOTA `CA` framework on a regression task for **EBM**. Augmenting data via the learned abstraction reduces the average error in all different settings compared to the SOTA. We used `COTA`$(\widehat{P}) - $`FRO`$ - c_\omega$ with the hyperparameters $(\kappa, \lambda, \mu) = (.2, .5, .3)$ achieving the lowest abstraction error.

Appendix G. `DAGs` and chains

In this section we provide the complete `DAGs` alongside their corresponding intervention posets and induced chains for all the scenarios that we considered and also visualise the operation of the ω map. For enhanced clarity and ease of navigation between different settings, we have included a concise table presented in Table 20. To provide further insight, the table also offers an analytical description of the interpretation of endogenous variables within each example for both the base and abstracted models.
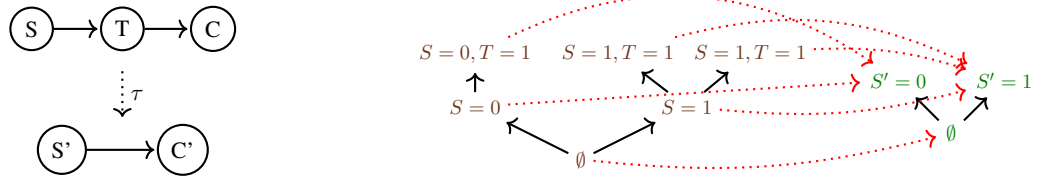
Figure 8: *Simple Lung Cancer* (**STC**) base (top) and abstracted (bottom) `DAGs` alongside their equivalent posets $\mathcal{I}$ and $\mathcal{I}'$ structure for the **STC$_{\text{np}}$** variation. The red arrows represent the $\omega : \mathcal{I} \to \mathcal{I}'$ map.
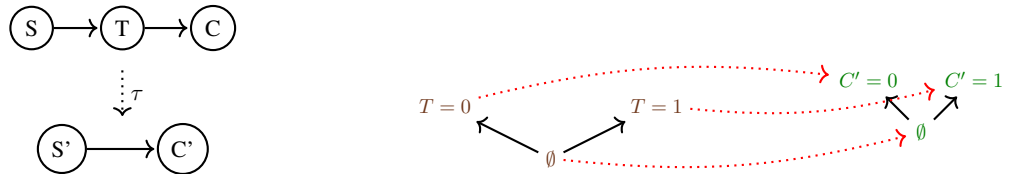
Figure 9: *Simple Lung Cancer* (**STC**) base (top) and abstracted (bottom) `DAGs` alongside their equivalent posets $\mathcal{I}$ and $\mathcal{I}'$ structure in the **STC$_{\text{p}}$** variation. The red arrows represent the $\omega : \mathcal{I} \to \mathcal{I}'$ map.
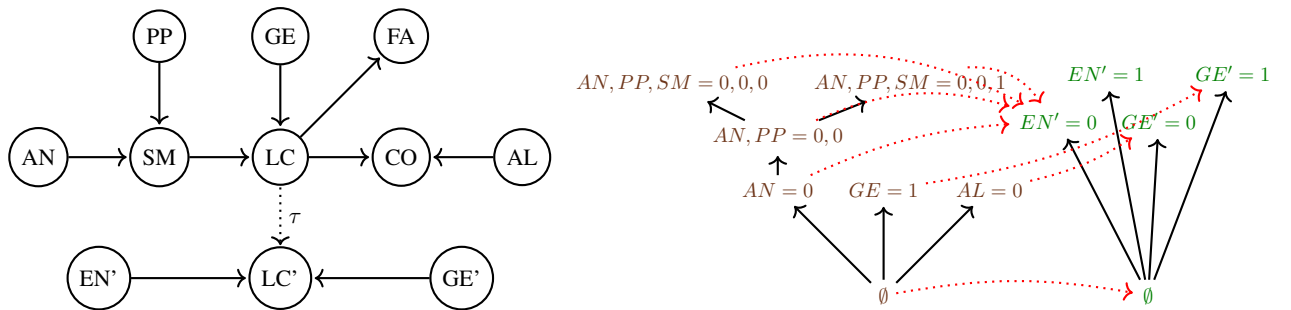

Figure 10: (**LUCAS**) base (top) and abstracted (bottom) `DAGs` alongside their equivalent posets $\mathcal{I}$ and $\mathcal{I}'$ structure. The red arrows represent the $\omega : \mathcal{I} \to \mathcal{I}'$ map.
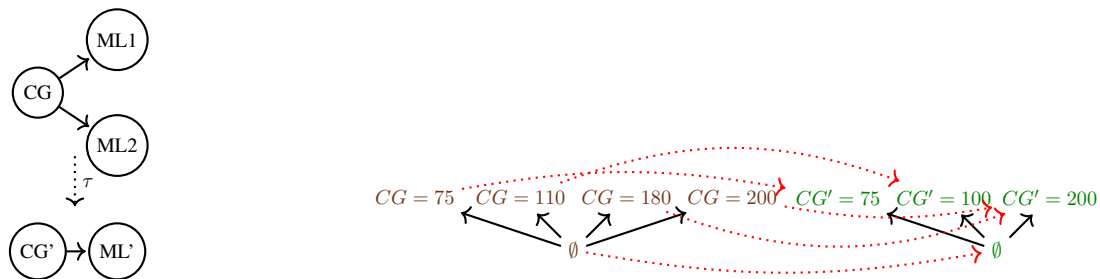
Figure 11: *Electric Battery Manufacturing* (**EBM**) base (top) and abstracted (bottom) DAGs alongside their equivalent posets $\mathcal{I}$ and $\mathcal{I}'$ structure. The red arrows represent the $\omega : \mathcal{I} \to \mathcal{I}'$ map.

Example	**Figure**	**X**	**X′**
STC	Figure 8	**S**: Smoking,	**S′**: Smoking,
	Figure 9	**T**: Tar,	**C′**: Cancer
		C:Cancer	
LUCAS	Figure 10	**AN**: Anxiety, **SM**: Smoking,	**EN′**: Environment,
		GE: Genetics, **PP**: Peer Pressure,	**LC′**: Lung Cancer,
		LC: Lung Cancer, **FA**: Fatigue,	**GE′**: Genetics
		CO: Coughing, **AL**: Allergy	
EBM	Figure 11	**CG**: Comma Gap (lab 1),	**CG′**: Comma Gap (lab 2),
		ML1: Mass loading position 1 (lab 1),	**ML′**: Mass loading (lab 2)
		ML1: Mass loading position 2 (lab 1)	

Table 20: Analytical interpretation of the base $\mathcal{M}$ and abstracted $\mathcal{M}'$ model's endogenous variables.

Appendix H. Optimal Transport

Optimal Transport (OT) theory as surveyed in Villani et al. (2009); Santambrogio (2015) provides a mathematical framework for systematically mapping one probability measure μ to another ν by looking amongst the set of all possible ways to transport the mass from one *source* distribution to a *target* one and selecting the one which minimizes a cost function. The seminal work of Peyré et al. (2019) surveys computational algorithms to solve OT problems in practice. Overall, OT offers a versatile and powerful tool for tackling complex and diverse problems across various fields, from image processing to economics.

H.1. General Measures

Monge formulation   The initial problem formulation was given by Gaspard Monge (1781) and states the following: For two arbitrary measures μ, ν on the Radon spaces[8] $\mathcal{X}$ and $\mathcal{Y}$ respectively, the *Monge problem* seeks to find a *map* $T^* : \mathcal{X} \mapsto \mathcal{Y}$ such that:

$$T^* = \inf_{T} \left\{ \int_{\mathcal{X}} c(x, T(x)) d\mu(x) : T_{\#}\mu = \nu \right\} \tag{25}$$

where $T_{\#}\mu$ is the pushforward function[9] and $c : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ is the cost function representing the cost of moving a unit mass from a location x to a location y. If such a T^* exists and attains the infimum then this is called the *optimal transport map*.
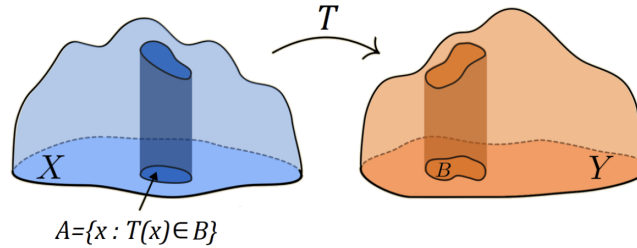

Figure 12: Monge transport map (source: (Kolouri et al., 2017) Figure 1)

Kantorovic formulation   In the initial problem formulation by Monge the map T^* may not always exist, for example, when μ is a Dirac measure but ν is not there is no map to attain the infimum of the optimization problem above.

For this reason, Kantorovich (1942) formulation of the problem relaxes the deterministic approach of Monge and introduces the probabilistic transport idea, which allows the execution of mass splitting from a source toward several targets. Specifically, the Kantorovic problem seeks to find a joint probability measure over the space $\mathcal{X} \times \mathcal{Y}$ or *probabilistic coupling*, which solves the following optimization problem:

$$P^{\star} = \inf_{P} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) dP(x, y) | P \in \mathcal{U}(\mu, \nu) \right\} \tag{26}$$

---

8. Radon space is a separable metric space such that any probability measure on it is a Radon measure

9. Let (X_1, Σ_1, μ) be a measure space, (X_2, Σ_2) a measurable space, and $f : X_1 \to X_2$ a measurable map. Then the following function ν on S_2 is the *pushforward* measure: $\nu(B) = \mu(f^{-1}(B))$ for $B \in \Sigma_2$. We write $f_{\#}\mu = \nu$.

where $\mathcal{U}(\mu, \nu)$ is the collection of all probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ on $\mathcal{X}$ and ν on $\mathcal{Y}$. Namely,

$$\mathcal{U}(\mu, \nu) = \left\{ P \in \mathcal{M}^1_+(\mathcal{X} \times \mathcal{Y}) : \rho_{\mathcal{X}\#}P = \mu, \ \rho_{\mathcal{Y}\#}P = \nu \right\} \tag{27}$$

where, $\rho_{\mathcal{X}\#}$ and $\rho_{\mathcal{Y}\#}$ are the pushforwards of the projections $\rho_{\mathcal{X}}(x, y) = x$ and $\rho_{\mathcal{Y}}(x, y) = y$.

This is an infinite-dimensional linear program over a space of measures. It is clear that this relaxed version of the problem is easier to work with since instead of looking for a map which that associates to each point $x_i \in \mathcal{X}$ a **single** point $y_i = T(x_i) \in \mathcal{Y}$, we are looking for a probability measure with the only constraint to preserve the marginals.

The Kantorovic formulation optimal transport computes the *p-Wasserstein* distance metric $Wp(\cdot, \cdot)$ between two probability distributions μ and ν:

$$W_p^p(\mu, \nu) = \inf_{P \in \mathcal{U}(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} ||x - y||^p dP(x, y), \ p \geq 1 \tag{28}$$

Remarks

- Let T be a transport map between μ and ν, and define $P_T = (id, T)_{\#}\mu$. Then, $P_T \in \mathcal{U}(\mu, \nu)$ is a transport plan between μ and ν.

- Let $\mathcal{X}, \mathcal{Y}$ be two compact spaces, and $c : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous cost function, which is bounded from below. Then Kantorovich's problem admits a minimizer (Thorpe, 2017).

Monge–Kantorovich equivalence for general measures The following theorem ensures that under some relatively simple conditions the Monge problem is feasible, meaning that the infimum of Eq. (25) can be attained and thus, the Kantorovich and Monge formulations are equivalent.

Theorem 2 ((Brenier, 1991)) *For Radon spaces $\mathcal{X}$ and $\mathcal{Y}$ with arbitrary measures μ, ν respectively, if at least one of the two input measures, say μ has a density ρ_μ with respect to the Lebesgue measure, then there exists a unique (up to an additive constant) convex function $\phi : \mathbb{R}^d \mapsto \mathbb{R}$ such that $\nabla \phi$ pushes forward μ onto ν. In other words, there exists a deterministic coupling $P^\star$ as follows:*

$$dP^\star(x, y) = d\mu(x)\delta_{\nabla\phi(x)}(y) \tag{29}$$

Furthermore, if $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(x, y) = ||x - y||^2$ then the optimal $P^\star$ in the Kantorovic formulation is unique and is supported on the graph $(x, T(x))$ of a Monge map $T : \mathbb{R}^d \to \mathbb{R}^d$. More formally,

$$P = (id, T)_{\#}\mu \iff \forall h \in \mathcal{C}(\mathcal{X} \times \mathcal{Y}), \ \int_{\mathcal{X} \times \mathcal{Y}} h(x, y)dP(x, y) = \int_{\mathcal{X}} h(x, T(x))d\mu(x) \tag{30}$$

This means that the map T is uniquely defined as the gradient of a unique convex function ϕ such that $T(x) = \nabla\phi(x)$, where $(\nabla\phi)_{\#}\mu = \nu$.

The two main conclusions from Brenier's theorem are the following:

- In the setting of $\mathcal{W}^2$ with no-singular densities, the Monge problem Eq. (25) and its Kantorovich relaxation Eq. (26) are equivalent (the relaxation is tight).

- An optimal transport map (Monge map) must be the gradient of a convex function. Namely, if $\phi : \mathbb{R}^d \to \mathbb{R}$ convex and $(\nabla\phi)_{\#}\mu = \nu$, then $T(x) = \nabla\phi(x)$ and

$$T^{\star} = \int_{\mathcal{X}} ||x - \nabla\phi(x)||^2 d\mu(x) \tag{31}$$

Various works extended the existence and uniqueness of Monge maps including strictly convex and super-linear costs.

Probabilistic interpretation Both Monge and Kantorovich formulations can be reinterpreted through the prism of random variables (Seguy et al., 2018), Peyré et al. (2019). Consider two complete metric spaces $\mathcal{X}$ and $\mathcal{Y}$ and random variables X and Y. We denote $X \sim \mu$ to say that X is distributed according to the probability measure μ. We can now restate both formulations of the optimal transport problem. Specifically, consider a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ and two random variables $X \sim \mu$ and $Y \sim \nu$ taking values in $\mathcal{X}$ and $\mathcal{Y}$ respectively.

Monge formulation: Find a map $T : \mathcal{X} \to \mathcal{Y}$ which transports the mass from μ to ν while minimizing the transportation cost,

$$\inf_{T} \mathbb{E}_{X \sim \mu} \left[c(X, T_{\#\mu}(X)) \right] \ \text{ s.t } \ T_{\#\mu}(X) \sim Y \tag{32}$$

Kantorovic formulation: Find a coupling $(X, Y) \sim P$ which minimizes the transportation cost and asserts that P has marginals equals to μ and ν,

$$\inf_{P} \mathbb{E}_{(X,Y) \sim P} \left[c(X, Y) \right] \ \text{ s.t } \ X \sim \mu, \ Y \sim \nu \tag{33}$$

H.2. Discrete Measures

In this section, we introduce the notations and the formulation of $\mathtt{OT}$ between discrete distributions. A probability vector is any element $\boldsymbol{\alpha}$ that belongs to the probability simplex Σ_k:

$$\Sigma_k := \left\{ \boldsymbol{\alpha} \in \mathbb{R}^k_{\geq 0} : \sum_{i=1}^{k} \alpha_i = 1 \right\} \tag{34}$$

A *discrete measure* μ with weights $\boldsymbol{\alpha}$ and points $x_1, ..., x_k \in \mathcal{X} \subset \mathbb{R}^d$ is defined as:

$$\mu = \sum_{i=1}^{k} \alpha_i \delta_{x_i} \tag{35}$$

where δ_x is the delta Dirac at position x. This measure is a probability measure if $\mu \in \Sigma_k$.

We are now going to restate the Monge-Kantorovic formulations in the cases of discrete measures. Consider $\mathcal{X} = \{x_i\}_{i=1}^{M} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{y_j\}_{j=1}^{N} \subset \mathbb{R}^d$ with respective (probability) weights $\boldsymbol{\alpha} \in \Sigma_M, \boldsymbol{\beta} \in \Sigma_N$. Thus, we have the discrete probability measures:

$$\mu = \sum_{i=1}^{M} \alpha_i \delta_{x_i} \ \text{ and } \ \nu = \sum_{j=1}^{N} \beta_j \delta_{y_j} \tag{36}$$

Finally, assuming that the cost of transporting a unit of mass from x_i to y_j is $c(x_i, y_j)$ where $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is the *cost function*, this induces a *cost matrix* $C_{ij} = c(x_i, y_j)$.

Monge formulation for discrete measures The Monge formulation of $\mathtt{OT}$ then aims to find a map $T^\star : \mathcal{X} \to \mathcal{Y}$ that push-forwards μ onto ν, by assigning to each x_i a single point y_j. Formally,

$$T^\star = \mathtt{OT}_c^M(\mu, \nu) = \underset{T:\, T_\#\mu=\nu}{\arg\min} \sum_{i=1}^{M} c(x_i, T(x_i)) \tag{37}$$

Kantorovic formulation for discrete measures Following the probabilistic transport approach of Eq. (26) for the general measures, the Kantorovich problem for discrete measures solves the following optimization problem in the form of a convex linear program:

$$P^\star = \mathtt{OT}_c^K(\mu, \nu) = \underset{P \in \mathcal{U}(\mu,\nu)}{\arg\min} \langle C, P \rangle = \underset{P \in \mathcal{U}(\mu,\nu)}{\arg\min} \sum_{i=1,j=1}^{M,N} C_{ij} P_{ij} \tag{38}$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product, expressing the total transportation cost, $C \in \mathbb{R}_{\geq 0}^{M \times N}$ is the cost matrix and $\mathcal{U}(\mu, \nu)$ is the set of joint probability measures with marginals μ and ν and called the *transport polytope* or *coupling set*. In particular, the transport polytope is a convex polytope defined as follows:

$$\mathcal{U}(\mu, \nu) = \left\{ P \in \mathbb{R}_{\geq 0}^{M \times N} : P \mathbb{1}_N = \mu, P^\top \mathbb{1}_M = \nu \right\} = \left\{ P \in \mathbb{R}^{M \times N} : \sum_{j=1}^{N} P_{ij} = \mu, \sum_{i=1}^{M} P_{ij} = \nu \right\} \tag{39}$$

In Fig. 13 we provide a schematic viewed of input measures (μ, ν) and a coupling $\mathcal{U}(\mu, \nu)$ encountered in the case of discrete measures for the Kantorovich $\mathtt{OT}$ formulation for the square euclidean cost $c(x, y) = ||x - y||^2$.

Entropic Optimal Transport Traditional $\mathtt{OT}$, while being a powerful tool, often encounters computational and statistical challenges in high-dimensional spaces. The introduction of entropy into this framework (Peyré et al., 2019) offers an organic solution that facilitates scalability and computational tractability through specific algorithms like Sinkhorn (Cuturi, 2013). Overall, Entropic $\mathtt{OT}$ leverages the principles of information theory, allowing for a more flexible and robust approach to the transportation problem between distributions. Specifically, for discrete measures, Entropic $\mathtt{OT}$ solves the following optimisation problem:

$$P^\star = \mathtt{OT}_c^K(\mu, \nu)_\epsilon = \underset{P \in \mathcal{U}(\mu,\nu)}{\arg\min} \langle C, P \rangle - \epsilon \mathcal{H}(P) = \underset{P \in \mathcal{U}(\mu,\nu)}{\arg\min} \sum_{i=1,j=1}^{M,N} C_{i,j} P_{i,j} - \epsilon \mathcal{H}(P) \tag{40}$$

where $\epsilon > 0$ a trade-off parameter and $\mathcal{H}(P)$ is the discrete entropy of a coupling matrix P is defined as:

$$\mathcal{H}(P) := -\sum_{ij} P_{ij}(\log(P_{ij} - 1)) \tag{41}$$

The idea behind Entropic regularization in optimal transport involves employing a regularization function to derive approximate solutions to the original transport problem of Eq. (38).

Remarks (Peyré et al., 2019)

- $\mathrm{OT}_c^K(\mu, \nu)_\epsilon \xrightarrow{\epsilon \to 0} \mathrm{OT}_c^K(\mu, \nu)$

- $\mathrm{OT}_c^K(\mu, \nu)_\epsilon \xrightarrow{\epsilon \to +\infty} \mu \otimes \nu = \mu\nu^\top$

Finally, it is worth mentioning that, given the strong concavity of $\mathcal{H}$, the objective in Eq. (40) becomes an ϵ-strongly convex function, ensuring the optimization problem $\mathrm{OT}_c^K(\mu, \nu)_\epsilon$ has a unique solution. In Fig. 14 we provide again the optimal coupling $\mathcal{U}(\mu, \nu)$ encountered in the case of discrete measures for the Entropic OT formulation for the square euclidean cost $c(x, y) = ||x - y||^2$.
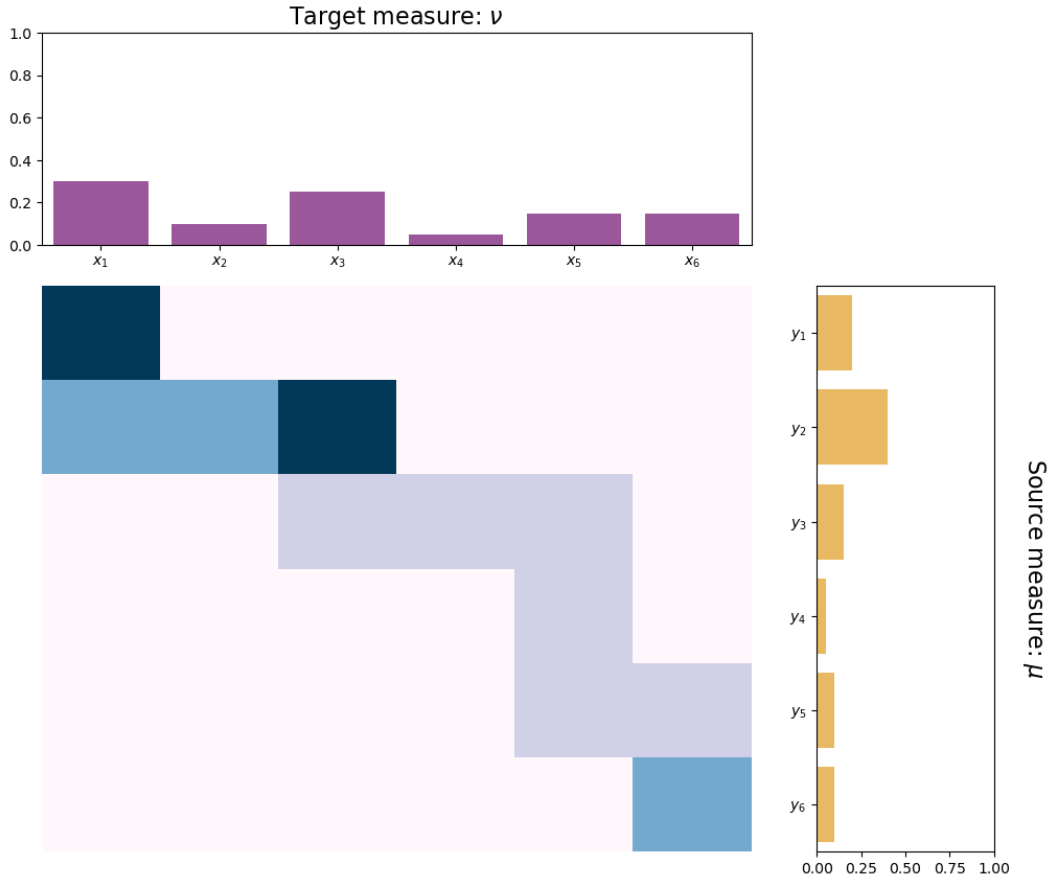

Figure 13: Kantorovic $\left(\mathrm{OT}_c^K(\mu, \nu)\right)$ optimal coupling for two input measures μ, ν. The optimal coupling $P^\star$ belongs to the transport polytope $\mathcal{U}(\mu, \nu)$ and thus preserves the marginals and the total mass, i.e. $\sum_j P_{ij}^\star = \mu$, $\sum_i P_{ij}^\star = \nu$ and $\sum_{i,j} P_{ij}^\star = 1$.
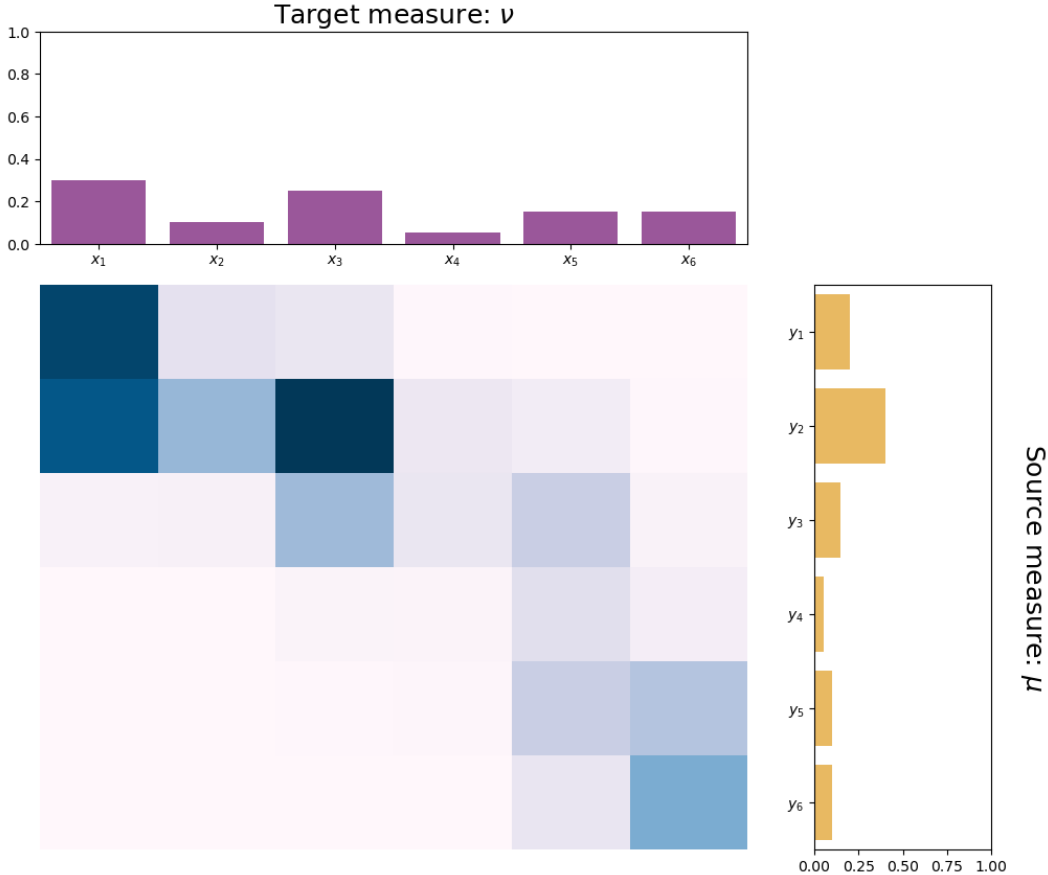
Figure 14: Kantorovic Entropic $\mathtt{OT}$ $\left(\mathtt{OT}_c^K(\mu, \nu)_\epsilon\right)$ optimal coupling for two input measures μ, ν and $\epsilon > 0$. The optimal coupling $P^\star$ belongs to the transport polytope $\mathcal{U}(\mu, \nu)$ and thus preserves the marginals and the total mass, i.e. $\sum_j P_{ij}^\star = \mu$, $\sum_i P_{ij}^\star = \nu$ and $\sum_{i,j} P_{ij}^\star = 1$.

Appendix I. Causal Abstraction Learning: Current State-Of-The-Art

In this section, we present a comparison of various recent methods for Causal Abstraction learning proposed to date, including also concurrent works. These methods have been developed within different CA frameworks. Specifically, COTA alongside Dyer et al. (2023), leverages the exact transformations framework proposed by Rubenstein et al. (2017). Meanwhile, the works by Xia and Bareinboim (2024) and Kekić et al. (2023) draw inspiration from the more constrained framework of constructive τ abstractions introduced by Beckers and Halpern (2019). Additionally, as referenced in the main text, Zennaro et al. (2023) builds upon the α-abstraction framework of Rischel (2020). In Table 21 we aim to align and compare these approaches. A notable observation from the table is the notational equivalence between the α-abstraction (Appendix D.1) and the constructive τ abstraction frameworks. For instance, the intervariable clustering/partitioning introduced by Xia and Bareinboim (2024) closely resembles the a map of the variables, while the intravariable clustering of domains corresponds to the collection α of maps between the realisations of the two models $\mathcal{M}$ and $\mathcal{M}'$.

Framework	$\mathcal{G}_\mathcal{M}$	$\mathcal{F}_\mathcal{M}$	$\mathcal{D}_\mathcal{M}$	Generic τ	Constructive τ		ω	$\mathcal{G}_{\mathcal{M}'}$	$\mathcal{F}_{\mathcal{M}'}$	$\mathcal{D}_{\mathcal{M}'}$
					Variable map $(\mathbb{C}, a)$	Value map $(\mathbb{D}, \alpha_i)$				
COTA	✓	×	✓	†	†		✓	✓	×	✓
Dyer et al. (2023)	×	×	✓	✓	✓		†	✓	†	✓
Zennaro et al. (2023)	✓	✓	✓	−	✓	†	✓	✓	✓	✓
Xia and Bareinboim (2024)	×	×	✓	−	✓	†	✓	†	†	×
Kekić et al. (2023)	✓	×	✓	−	✓	†	✓	✓	†	✓

Table 21: Comparison of different CA learning frameworks. For a given SCM, $\mathcal{F}$ is the set of structural functions/mechanisms, $\mathcal{G}$ is the underlying DAG and $\mathcal{D}$ indicates the data of the SCM. The checkmark symbol $(\checkmark)$ denotes the presence of the provided information, whereas the $(\times)$ symbol the absence of the respective information. The dagger symbol $(\dagger)$ indicates the output/objectives of each method, while the dash symbol $(-)$ denotes the method's incapacity to learn the associated information from the provided assumptions. Finally, a gray symbol $(\checkmark, \dagger)$ featuring a faded area suggests that the information is implied by a more general or stronger assumption.