MULTI-TASK LEARNING OF CONDITIONAL MEAN OPERATORS: APPLICATIONS TO DYNAMICAL SYSTEMS AND UNCERTAINTY QUANTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating conditional statistics and learning representations of a population of conditional distributions are central problems in many data-driven applications, including uncertainty quantification and dynamical systems analysis. Conditional mean operators (CMOs), a class of linear operators between function spaces, resolve these objectives by providing access to a broad class of conditional statistics. However, existing methods typically estimate each CMO independently or constrain it to prespecified function spaces, thereby preventing the exploitation of shared structure across related distributions.
In this work, we posit that related CMOs share finite-dimensional input and output function spaces, and are specialized for each task with a linear operator mapping these spaces. Based on this hypothesis, we introduce MTL-CMO, a multi-task framework that jointly learns shared function spaces and task-specific operators across multiple datasets. We further introduce T-CMO, a transfer learning method that reuses the shared spaces to estimate, in closed form, the operator of a new conditional distribution. We establish statistical guarantees quantifying the benefits of jointly learning the shared function spaces. Our experiments demonstrate that learning shared function spaces improves uncertainty quantification across a broad range of conditional distributions and, when applied to Langevin and plasma dynamics, yields compact representations of complex dynamics that retain physically meaningful information and enable parameter identification.

1 INTRODUCTION

Characterizing the distribution of a target variable Y given a conditioning variable X is key to many statistical learning problem. Beyond solely predicting the conditional expectation, access to the conditional distribution allows the computation of a broad range of conditional statistics, including moments, quantiles, or confidence regions, typically required for uncertainty quantification (Ghanem et al., 2017). Similarly, for stochastic dynamical systems, the conditional distributions of future states given the current state characterizes the dynamic and provides access to physically meaningful quantities like transition probabilities, relaxation times, or invariant measures (Pavliotis, 2014).
The conditional mean operator (CMO) provides a natural operator-theoretic representation of a conditional distribution by mapping an observable f of Y to its conditional expectation:

$$\left[\mathsf{E}_{\mathrm{Y|X}}\, f\right](x) = \mathbb{E}\left[f(Y) \mid X = x\right].$$

By varying f, a single CMO provides access to a broad range of conditional statistics, making it a versatile representation of conditional distributions. For dynamical systems, CMOs are particularly informative: their spectral decompositions reveal coherent dynamical modes associated with characteristic timescales and frequencies. Existing approaches to CMO estimation primarily rely on kernel methods with either fixed kernels (Song et al., 2009; Muandet et al., 2017) or learned feature representations (Shimizu et al., 2024; Kostic et al., 2024). However, accurately estimating these operators generally requires large datasets, making their estimation challenging when only limited observations are available (Hertel et al., 2026).
However, in many applications, data are collected from a population of related conditional distributions arising, for instance, from different physical parameters, experimental conditions, or subjects.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Although the conditional distributions vary across the population, their associated CMOs may share common input and output function spaces, while differing through the linear transformations acting between these spaces. This structure naturally motivates a multi-task approach (Caruana, 1997): the shared function spaces can be learned jointly across tasks, while retaining a task-specific operator for each conditional distribution. Sharing information in this manner can improve the estimation of individual CMOs (Maurer et al., 2016). Moreover, the learned function spaces can be transferred to a new task, for which only the corresponding low-dimensional operator must be estimated, thereby improving data efficiency (Tripuraneni et al., 2020). Nevertheless, existing CMO estimators predominantly treat conditional distributions independently, while existing multi-task approaches to conditional modeling are generally designed for specific applications and do not provide a unified operator-learning framework applicable to both uncertainty quantification and dynamical systems.

Contributions. In this work, **(1)** we introduce MTL-CMO, a multi-task framework that jointly learns a population of CMOs. Each task is represented by a task-specific linear operator acting between shared, learnable input/output function spaces. **(2)** We subsequently introduce T-CMO, a transfer-learning estimator that reuses the learned function spaces to estimate the CMO of a new task in closed-form. **(3)** We establish statistical guarantees quantifying the benefits of jointly learning the shared input/output function spaces for CMO estimation. **(4)** We evaluate MTL-CMO and T-CMO on uncertainty-quantification and dynamical-systems benchmarks in both multi-task and transfer-learning settings. Across these experiments, our methods are competitive or outperform state-of-the-art baselines, including methods specifically designed for the corresponding applications.

Paper organization. Section 2 introduces the necessary background. Section 3 presents the MTL-CMO and T-CMO frameworks, establishes their statistical guarantees, and positions them with respect to related work. Section 4 presents the numerical experiments, before we conclude the paper.

2 Background on CMO: definition, estimation and applications

Conditional Mean Operator (CMO) (Fukumizu et al., 2004). Consider a pair of random variables (X, Y) on the product space $\mathcal{X} \times \mathcal{Y}$ with joint distribution ρ and marginals μ and ν respectively. We assume that ρ is absolutely continuous w.r.t. the product of marginals, i.e. $\rho \ll \mu \otimes \nu$, with density function square integrable, i.e. $\mathrm{d}\,\rho\,/\mathrm{d}(\mu \otimes \nu) \in L^2_{\mu\otimes\nu}(\mathcal{X} \times \mathcal{Y})$. Then, the conditional mean operator $\mathsf{E}_{\mathrm{Y|X}} : \mathrm{L}^2_\nu(\mathcal{Y}) \to \mathrm{L}^2_\mu(\mathcal{X})$ is well defined and verifies for any $f \in \mathrm{L}^2_\nu(\mathcal{Y})$ and $x \in \mathcal{X}$:

$$[\mathsf{E}_{\mathrm{Y|X}}\, f](x) = \mathbb{E}[f(Y) \mid X = x] = \int_{\mathcal{Y}} f(y) p(x, y)\, \nu(dy), \tag{1}$$

where $\mathrm{L}^2_\mu(\mathcal{X})$ and $\mathrm{L}^2_\nu(\mathcal{Y})$ are separable Hilbert space. Since $\mathsf{E}_{\mathrm{Y|X}}$ is an Hilbert-Schmidt operator, and $\| \mathsf{E}_{\mathrm{Y|X}} \|_{\mathrm{op}} = 1$ with $\mathsf{E}_{\mathrm{Y|X}}\, \mathbf{1}_\mathcal{Y} = \mathbf{1}_\mathcal{X}$, it admits a singular value decomposition (SVD) such that:

$$\mathsf{E}_{\mathrm{Y|X}} = \textstyle\sum_{i\geq 0} \sigma_i u_i \otimes v_i, \qquad (\sigma_0, u_0, v_0) = (1, \mathbf{1}_\mathcal{X}, \mathbf{1}_\mathcal{Y}) \tag{2}$$

where $\{u_i\}_{i\geq 0}$ and $\{v_i\}_{i\geq 0}$ are singular functions forming orthonormal basis, $\{\sigma_i\}_{i\geq 0}$ the decreasing sequence of positive singular values, and $\mathbf{1}_\mathcal{X}$ an indicator function. Hence, the conditional mean operator can be approximated arbitrary well, and by the Eckart-Young-Mirsky Theorem the best rank-r approximation is the truncation $\mathsf{E}^{\mathrm{r}}_{\mathrm{Y|X}} = \sum_{i=0}^{r-1} \sigma_i u_i \otimes v_i$ with approximation error $\| \mathsf{E}_{\mathrm{Y|X}} - \mathsf{E}^{\mathrm{r}}_{\mathrm{Y|X}} \|_{\mathrm{op}} = \sigma_r$. It follows that for any $f \in \mathrm{L}^2_\nu(\mathcal{Y})$ and $x \in \mathcal{X}$, the conditional expectation can be approximated with:

$$\big[\mathsf{E}_{\mathrm{Y|X}}\, f\big](x) \approx \Big[\mathsf{E}^{\mathrm{r}}_{\mathrm{Y|X}}\, f\Big](x) = \mathbb{E}_\nu[f(Y)] + \textstyle\sum_{i=1}^{r-1} \sigma_i u_i(x)\, \mathbb{E}_\nu[v_i(Y) f(Y)], \tag{3}$$

Operator estimation with Neural Conditional Probability. The Neural Conditional Probability framework (NCP) learns a finite-rank approximation of a conditional mean operator from paired observations (Kostic et al., 2024). It first introduce the deflated operator $\mathsf{D}_{\mathrm{Y|X}} = \mathsf{E}_{\mathrm{Y|X}} - \mathbf{1}_\mathcal{X} \otimes \mathbf{1}_\mathcal{Y}$ to remove the trivial singular component. For a latent dimension r, NCP parametrizes with neural networks the approximation of $\mathsf{E}_{\mathrm{Y|X}}$ through its truncated SVD

$$\mathsf{E}^\theta_{\mathrm{Y|X}} = \mathbf{1}_\mathcal{X} \otimes \mathbf{1}_\mathcal{Y} + \mathsf{D}^\theta_{\mathrm{Y|X}} = \mathbf{1}_\mathcal{X} \otimes \mathbf{1}_\mathcal{Y} + \textstyle\sum_{i=1}^{r} \sigma^\theta_i u^\theta_i \otimes v^\theta_i, \tag{4}$$

with nontrivial singular values $\sigma^\theta \in [0, 1)^r$, and singular functions $u^\theta(x) \triangleq [u^\theta_1(x), \ldots, u^\theta_r(x)]^\top$ and $v^\theta(y) \triangleq [v^\theta_1(y), \ldots, v^\theta_r(y)]^\top$. Denoting $q_\theta(x, y) = \sum_{i=1}^{r} \sigma^\theta_i u^\theta_i(x) v^\theta_i(y)$, the deflated operator

is estimated by minimizing a regularized problem with a data-fitting term measuring the discrepancy between the estimated and true deflated operators

$$\begin{aligned}\mathcal{L}(\theta) &= \| \mathsf{D}^{\theta}_{\mathrm{Y|X}} - \mathsf{D}_{\mathrm{Y|X}} \|^2_{\mathrm{HS}} - \| \mathsf{D}_{\mathrm{Y|X}} \|^2_{\mathrm{HS}} \\ &= \mathbb{E}_{\mu\otimes\nu}[q^2_\theta(X,Y)] - 2(\mathbb{E}_\rho[q_\theta(X,Y)] - \mathbb{E}_{\mu\otimes\nu}[q_\theta(X,Y)]),\end{aligned} \tag{5}$$

and a regularization encouraging the functions to be centered and orthonormal in their L^2 spaces.

Application: conditional uncertainty quantification. Beyond conditional expectations, the CMO provides access to conditional probabilities. Indeed, for any measurable set $B \subseteq \mathcal{Y}$, the indicator function $\mathbf{1}_B \in \mathrm{L}^2_\nu(\mathcal{Y})$, and the conditional probability can be evaluated with

$$\mathbb{P}(Y \in B \mid X = x) \approx \widehat{\mathbb{P}}(Y \in B \mid X = x) = \left[\widehat{\mathsf{E}^{\theta}_{\mathrm{Y|X}}} \mathbf{1}_B\right](x). \tag{6}$$

Consequently, an estimated CMO can be used to characterize the uncertainty of Y conditionally on $X{=}x$, for instance through a $(1-\alpha)$ conditional confidence region $C_\alpha(x) \subseteq \mathcal{Y}$ satisfying $\left[\mathsf{E}_{\mathrm{Y|X}} \mathbf{1}_{C_\alpha(x)}\right](x) = \mathbb{P}(Y \in C_\alpha(x) \mid X = x) \geq 1-\alpha$. For scalar Y, taking $f_t(y) = \mathbf{1}_{\{y \leq t\}}$ yields $\widehat{F}(t \mid x) = \widehat{\mathbb{P}}(Y \leq t \mid X = x)$ with conditional quantiles verifying $Q_\alpha(x) = \inf\{t \mid \widehat{F}(t \mid x) \geq \alpha\}$.

Application: representation of dynamical systems. Consider a dynamical system that is a time-homogeneous Markov process $(X_t)_{t\geq 0}$ with invariant measure π. For any lag $t > 0$, the transition from X_0 to X_t is characterized by the transition operator $[A_t f](x) = \mathbb{E}[f(X_t) \mid X_0 = x]$ with $f \in L^2_\pi(\mathcal{X})$. Hence, each T_t is precisely the CMO of the pair (X_0, X_t) and describes how observables of the future state depend on the initial state. Under suitable assumptions on the process, the family $\{T_t\}_{t>0}$ forms a semi-group admitting an infinitesimal generator $G = \lim_{t\to 0^+}(A_t f - f)/t$ with a discrete spectrum such that for any $t > 0$

$$[A_t f](x) = [\exp(tG) f](x) = \textstyle\sum_{j\geq 0} e^{t\lambda_j} \langle f, g_j \rangle f_j(x), \tag{7}$$

where λ_j and (f_j, g_j) denote the eigenvalues and corresponding right and left eigenfunctions of G. In particular, the transition operators $\{A_t\}_{t>0}$ share the spectral modes of the infinitesimal generator, while their eigenvalues evolve with the lag as $e^{t\lambda_j}$. Estimating these CMOs from sampled trajectories consequently provides access to the spectral structure of the underlying generator (Koopman, 1931; Singer & Spilerman, 1976; Schmid, 2010). The generator representation is central to data-driven dynamical systems analysis, as its eigenfunctions and eigenvalues encode characteristic spatial structures and temporal scales of the dynamics. Since errors in the estimated transition operators propagate to the generator, accurate CMO estimation is essential for reliably recovering the underlying dynamics. For interested readers, we provide further details on operator-theoretic representation of dynamical systems and their estimation in Appendix A.

3 Multi-task and Transfer learning for CMO

3.1 Multi-task learning of conditional mean operators

MTL strategy & assumptions. We propose a multi-task approach to estimate CMOs from observations of related joint distributions. The approach learns input and output function spaces shared across tasks, while each task retains its own operator representation within those spaces.

Formally, we consider K pairs of random variables $(X^{(k)}, Y^{(k)}) \sim \rho_k$, with marginals μ_k and ν_k. We denote by $\mathsf{E}^{(\mathrm{k})}_{\mathrm{Y|X}} : L^2_{\nu_k}(\mathcal{Y}) \to L^2_{\mu_k}(\mathcal{X})$ the associated conditional mean operator and $\mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}}$ its corresponding deflated operator. We make the following assumptions:

(A1) Hilbert-Schmidt operator. For any $k \in [K]$, $\mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}}$ is a Hilbert-Schmidt operator and its singular spectrum is concentrated on a small number r_k of dominant components, so that it is well approximated by its rank-r_k SVD truncation

$$\mathsf{D}^{(\mathrm{k}),\mathrm{r_k}}_{\mathrm{Y|X}} = \textstyle\sum_{i=1}^{r_k} \sigma_i^{(k)} u_i^{(k)} \otimes v_i^{(k)}. \tag{8}$$

(A2) Shared functional spaces. We assume that the dominant singular functions of the different tasks can be represented within shared finite-dimensional functional spaces. More precisely, there exist two d-dimensional functional spaces, with $d \geq \max_k r_k$, such that for any $k \in [K]$ and $i \leq r_k$

$$u_i^{(k)} \in \mathrm{span}\{\phi_1, \ldots, \phi_d\}, \qquad v_i^{(k)} \in \mathrm{span}\{\psi_1, \ldots, \psi_d\}. \tag{9}$$

Thus, although each task may have a different effective rank and distinct singular functions, its dominant dependence structure is represented within the same pair of functional spaces. This shared representation provides the inductive bias that enables information to be shared across tasks.

Multi-task low-rank operator parametrization. We represent the shared functional spaces with two dictionaries of neural-networks

$$\Phi_\theta(x) \triangleq \{\phi_1^\theta(x), \dots, \phi_d^\theta(x)\} \qquad \Psi_\theta(y) \triangleq \{\psi_1^\theta(y), \dots, \psi_d^\theta(y)\}. \tag{10}$$

Although these dictionaries are shared across tasks, the orthogonality constraints associated with the singular decomposition are task-dependent: for each task k, the left/right singular functions of the delfated operator $\mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}}$ are orthonormal in $\mathrm{L}^2_{\mu_\mathrm{k}}(\mathcal{X})$ and $\mathrm{L}^2_{\nu_\mathrm{k}}(\mathcal{Y})$ respectively. Consequently, a direct parametrization of each operator through its singular value decomposition, as in the NCP framework (see Section 2), would require enforcing different orthogonality constraints for every task, making the optimization cumbersome. To circumvent these task-specific constraints, we parametrize each deflated operator through an unconstrained low-rank factorization:

$$\mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}} = \mathsf{S}^{(k)}_{\Phi_\theta} \mathbf{A}^{(k)} \mathbf{B}^{(k)\top} \mathsf{S}^{(k)*}_{\Psi_\theta} \tag{11}$$

where $(\mathbf{A}^{(k)}, \mathbf{B}^{(k)}) \in (\mathbb{R}^{d \times r_k})^2$. Here, $\mathsf{S}^{(k)}_{\Phi_\theta} : \mathbb{R}^d \to \mathrm{L}^2_{\mu_\mathrm{k}}(\mathcal{X})$ and $\mathsf{S}^{(k)}_{\Psi_\theta} : \mathbb{R}^d \to \mathrm{L}^2_{\nu_\mathrm{k}}(\mathcal{Y})$ denote the synthesis maps associated with the shared neural dictionaries; they map any $\alpha \in \mathbb{R}^d$ to the functions

$$[\mathsf{S}^{(k)}_{\Phi_\theta}\alpha](x) \triangleq \textstyle\sum_{j=1}^d \alpha_j \phi_j^\theta(x), \qquad [\mathsf{S}^{(k)}_{\Psi_\theta}\alpha](y) \triangleq \textstyle\sum_{j=1}^d \alpha_j \psi_j^\theta(y). \tag{12}$$

Importantly, despite the fact that synthesis maps are identical across tasks, their adjoints depend on the task-specific inner products. In particular, the adjoint of $\mathsf{S}^{(k)}_{\Psi_\theta}$ is given by

$$\mathsf{S}^{(k)*}_{\Psi_\theta} f = \left(\langle \psi_j^\theta, f \rangle_{\nu_k}\right)_{j \in [d]}, \qquad f \in \mathrm{L}^2_{\nu_\mathrm{k}}(\mathcal{Y}). \tag{13}$$

Retrieving task-specific SVD from the low-rank parametrization. For each task k, the singular value decomposition of the deflated operator can be recovered *post hoc* by centering and whitening the shared dictionaries with their task-specific Gram matrices that can be estimated from data

$$\mathbf{G}^{(k)}_{\Phi_{\theta,c}} \triangleq \mathbb{E}_{\mu_k}\left[\Phi^{(k)}_{\theta,c}(X)\Phi^{(k)}_{\theta,c}(X)^\top\right], \qquad \mathbf{G}^{(k)}_{\Psi_{\theta,c}} \triangleq \mathbb{E}_{\nu_k}\left[\Psi^{(k)}_{\theta,c}(Y)\Psi^{(k)}_{\theta,c}(Y)^\top\right], \tag{14}$$

with $\Phi^{(k)}_{\theta,c}(x) = \Phi_\theta - \mathbb{E}_{\mu_k}[\Phi_\theta(X)]$ and $\Psi^{(k)}_{\theta,c}(y) = \Psi_\theta - \mathbb{E}_{\nu_k}[\Psi_\theta(Y)]$. Consider the matrix SVD

$$\mathbf{G}^{(k)1/2}_{\Phi_{\theta,c}} \mathbf{A}^{(k)} \mathbf{B}^{(k)\top} \mathbf{G}^{(k)1/2}_{\Psi_{\theta,c}} = \mathbf{U}^{(k)} \operatorname{Diag}(\boldsymbol{\sigma}^{(k)}) \mathbf{V}^{(k)\top}. \tag{15}$$

Denoting $\mathbf{u}_i^{(k)}$ and $\mathbf{v}_i^{(k)}$ the ith columns of $\mathbf{U}^{(k)}$ and $\mathbf{V}^{(k)}$, the singular functions are given by

$$\widetilde{u}_i^{(k)} = \mathsf{S}^{(k)}_{\Phi_{\theta,c}} \mathbf{G}^{(k)-1/2}_{\Phi_{\theta,c}} \mathbf{u}_i^{(k)}, \qquad \widetilde{v}_i^{(k)} = \mathsf{S}^{(k)}_{\Psi_{\theta,c}} \mathbf{G}^{(k)-1/2}_{\Psi_{\theta,c}} \mathbf{v}_i^{(k)}, \tag{16}$$

where $\mathbf{G}^{(k)-1/2}_{\Phi_{\theta,c}}$ and $\mathbf{G}^{(k)-1/2}_{\Psi_{\theta,c}}$ are square root Moore–Penrose inverses. By construction, the functions $\{\widetilde{u}_i^{(k)}\}_{i\in[r_k]}$ and $\{\widetilde{v}_i^{(k)}\}_{i\in[r_k]}$ are orthonormal in their respective space $\mathrm{L}^2_{\mu_\mathrm{k}}(\mathcal{X})$ and $\mathrm{L}^2_{\nu_\mathrm{k}}(\mathcal{Y})$, and therefore yield the task-specific deflated operator with singular decomposition

$$\tilde{\mathsf{D}}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}} = \textstyle\sum_{i=1}^{r_k} \sigma_i^{(k)} \widetilde{u}_i^{(k)} \otimes \widetilde{v}_i^{(k)}. \tag{17}$$

Thus, the proposed factorization avoids imposing orthogonality constraints during training while retaining access to an orthonormal, task-specific spectral representation after optimization.

Multi-task learning protocol (MTL-CMO). At the population level, we jointly estimate the K deflated operators by solving

$$\min_{\theta, \{\mathbf{A}^{(k)}, \mathbf{B}^{(k)}\}_{k=1}^K} \sum_{k=1}^K \underbrace{\| \mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}} - \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} \|^2_{\mathrm{HS}} - \| \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} \|^2_{\mathrm{HS}}}_{\mathcal{L}^{(k)}} + \lambda \underbrace{(\|\mathbf{A}^{(k)}\|^2_{\mathrm{F}} + \|\mathbf{B}^{(k)}\|^2_{\mathrm{F}})}_{\mathcal{R}^{(k)}}. \tag{18}$$

The data-fitting term $\mathcal{L}^{(k)}$ measures the discrepancy between the estimated deflated operator and its population counterpart, following the NCP framework (see Section 2). However, unlike NCP, the

regularization $\mathcal{R}^{(k)}$ does not include orthogonality penalizations, it only imposes a ridge regularization on the matrices $(\mathbf{A}^{(k)}, \mathbf{B}^{(k)})$ to resolve the scaling ambiguity of the factorization.

From an empirical perspective, given a set of observations $\{(x_i^{(k)}, y_i^{(k)})\}_{i\in[n_k]}$, we denote $\Phi_\theta(\mathbf{X}^{(k)}) = \{\Phi_\theta(x_i^{(k)})\}_{i\in[n_k]}$ and $\Psi_\theta(\mathbf{Y}^{(k)}) = \{\Psi_\theta(y_i^{(k)})\}_{i\in[n_k]}$ the feature matrices. Considering, the matrices

$$\mathbf{Q}^{(k)} = \Phi_\theta(\mathbf{X}^{(k)})\mathbf{A}^{(k)}\mathbf{B}^{(k)\top}\Psi_\theta(\mathbf{Y}^{(k)})^\top \quad \text{and} \quad \mathbf{H}^{(k)} = \mathbf{I}_{n_k} - 1/n_k\mathbf{1}\mathbf{1}^\top, \tag{19}$$

the data-fitting loss $\mathcal{L}^{(k)}$ admits an unbiased empirical estimator defined by

$$\widehat{\mathcal{L}^{(k)}} \triangleq \frac{\|\mathbf{Q}^{(k)}\|_{\mathrm{F}}^2 - \|\operatorname{Diag}(\mathbf{Q}^{(k)})\|_2^2}{n_k(n_k-1)} - \frac{2}{n_k-1}\operatorname{Tr}(\mathbf{H}^{(k)}\mathbf{Q}^{(k)}). \tag{20}$$

The multi-task training procedure and the pseudo-code can be found in Appendix B.

Related work. Standard multi-task representation learning methods often share an input feature map across task-specific predictors (Caruana, 1997; Maurer et al., 2016). MTL-CMO extends this principle to operators by learning input and output function spaces, whose corresponding synthesis maps Equation (12) act in spaces defined by each task's marginals. This differs from conditional mean embedding, which typically uses a prescribed reproducing kernel Hilbert space (Song et al., 2009; Muandet et al., 2017), and from NCP, which estimates function spaces but for only a single CMO (Kostic et al., 2024). A detailed related work is provided in Appendix C.

3.2 Transfer learning of conditional mean operators in closed form (T-CMO)

We now suppose that the shared functional space have been learned and are now frozen: $\Phi_{\widehat{\theta}}$ and $\Psi_{\widehat{\theta}}$. We aim to estimate the CMO $\mathsf{E}^{\text{new}}_{\mathrm{Y|X}}$ associated to a new pair of random variable $(X^{\text{new}}, Y^{\text{new}})$. Rather than directly enforcing the low-rank structure, we parametrize the deflated operator with a matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$ such that:

$$\mathsf{D}^{\text{new}}_{\mathrm{Y|X}} = \mathsf{S}^{\text{new}}_{\Phi_{\widehat{\theta},c}}\mathbf{M}\mathsf{S}^{\text{new}*}_{\Psi_{\widehat{\theta},c}}. \tag{21}$$

Denoting $\mathbf{C}^{\text{new}}_{\Phi_{\widehat{\theta},c}\Psi_{\widehat{\theta},c}} = \mathbb{E}_{\rho_{\text{new}}}[\Phi_{\widehat{\theta},c}(X^{\text{new}})\Psi_{\widehat{\theta},c}(Y^{\text{new}})^\top]$ the cross-covariance matrix, the loss

$$\mathcal{L}^{\text{new}} = \operatorname{Tr}\left(\mathbf{G}^{\text{new}}_{\Psi_{\widehat{\theta},c}}\mathbf{M}^\top\mathbf{G}^{\text{new}}_{\Phi_{\widehat{\theta},c}}\mathbf{M}\right) - 2\operatorname{Tr}\left(\mathbf{C}^{\text{new}\top}_{\Phi_{\widehat{\theta},c}\Psi_{\widehat{\theta},c}}\mathbf{M}\right), \tag{22}$$

yields a quadratic problem in $\mathbf{M}$ with closed-form solution

$$\mathbf{M}^* = \mathbf{G}^{\text{new}^{-1}}_{\Phi_{\widehat{\theta},c}}\mathbf{C}^{\text{new}}_{\Phi_{\widehat{\theta},c}\Psi_{\widehat{\theta},c}}\mathbf{G}^{\text{new}^{-1}}_{\Psi_{\widehat{\theta},c}}. \tag{23}$$

In a post-hoc phase, the SVD of the deflated operator can be retrieved following the procedure presented in the previous section and then truncated to retrieve its low-rank modeling.

3.3 Statistical guarantees

We now quantify the benefit of jointly learning the shared spaces through a statistical learning rate. We provide below a simplified result whose complete statement and proof is available in Appendix G, alongside additional results, including a guarantee for transfer to a new task.

Theorem 1 (Informal, see Theorem 2)**.** *Assume K tasks with n observation each, dictionary classes $(\Phi_\Theta, \Psi_\Theta)$ with d bounded neural functions, and uniformly bounded task-specific matrices $\mathbf{A}^{(k)}$ and $\mathbf{B}^{(k)}$. With high probability, the trained model $\widehat{h}=(\widehat{\theta}, \{\widehat{\mathbf{A}}^{(k)}, \widehat{\mathbf{B}}^{(k)}\}_{k\in[K]})$ satisfies*

$$\frac{1}{K}\sum_{k=1}^{K} \| \mathsf{D}^{(\mathrm{k}),\widehat{\theta}}_{\mathrm{Y|X}} - \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} \|^2_{\mathrm{HS}} - \mathcal{A}_K \leq \tilde{\mathcal{O}}\left(\frac{\sqrt{d}\big(\operatorname{Comp}(\Phi_\Theta) + \operatorname{Comp}(\Psi_\Theta)\big)}{\sqrt{nK}} + \frac{d}{\sqrt{n}}\right) \tag{24}$$

where $\mathcal{A}_K$ is the irreducible risk, $\operatorname{Comp}(\cdot)$ is the Gaussian complexity of a dictionary class and $\tilde{\mathcal{O}}$ hides the logarithmic term, and a multiplicative factor depending on the boundedness constants.

This result extends the statistical benefit of multi-task representation learning (Maurer et al., 2016) to operator estimation. The key intuition is that all tasks contribute to the learning of the shared function spaces: the cost of learning the neural dictionaries is therefore spread over all nK observations. Once the spaces are learned, each operator still requires its own matrices, giving a task-specific term that depends on $d/\sqrt{n}$ alone.

Table 1: Mean 1-Wasserstein distance to the ground-truth conditional distribution averaged over 100 tasks. Scores in $<\text{mean}> \pm <\text{std}>$ over 10 training seeds. Lower is better. **First** and second.

	Method	CD1	CD2	CD3	CD4
Multi-task	MTL-CMO **(ours)**	**0.088 ± 0.004**	**0.102 ± 0.002**	**0.106 ± 0.003**	0.070 ± 0.001
	Pooled-NCP (Kostic et al., 2024)	0.091 ± 0.002	0.120 ± 0.009	0.139 ± 0.008	0.077 ± 0.008
	MTL-MDN (Bishop, 1994)	0.110 ± 0.001	0.138 ± 0.005	0.216 ± 0.009	0.107 ± 0.005
	DeepJMQR (Rodrigues & Pereira, 2020)	0.091 ± 0.002	0.114 ± 0.005	0.125 ± 0.008	0.066 ± 0.004
Single-task	NCP (Kostic et al., 2024)	0.096 ± 0.001	0.167 ± 0.037	0.185 ± 0.001	0.150 ± 0.002
	CFM (Lipman et al., 2023)	0.112 ± 0.001	0.182 ± 0.021	0.170 ± 0.006	0.150 ± 0.005
	MDN (Bishop, 1994)	0.115 ± 0.002	0.259 ± 0.007	0.265 ± 0.008	0.218 ± 0.003
	NSF (Durkan et al., 2019)	0.109 ± 0.002	0.118 ± 0.013	0.129 ± 0.012	0.075 ± 0.006
	Engression (Shen & Meinshausen, 2025)	0.143 ± 0.001	0.147 ± 0.001	0.129 ± 0.002	0.074 ± 0.001
	Nadaraya–Watson (Nadaraya, 1964)	0.096 ± 0.001	0.152 ± 0.003	0.109 ± 0.002	**0.064 ± 0.001**
	NGBoost (Duan et al., 2020)	0.228 ± 0.003	0.284 ± 0.005	0.278 ± 0.004	0.242 ± 0.006
	FlexCode (Izbicki & Lee, 2017)	0.245 ± 0.001	0.225 ± 0.000	0.321 ± 0.003	0.327 ± 0.003

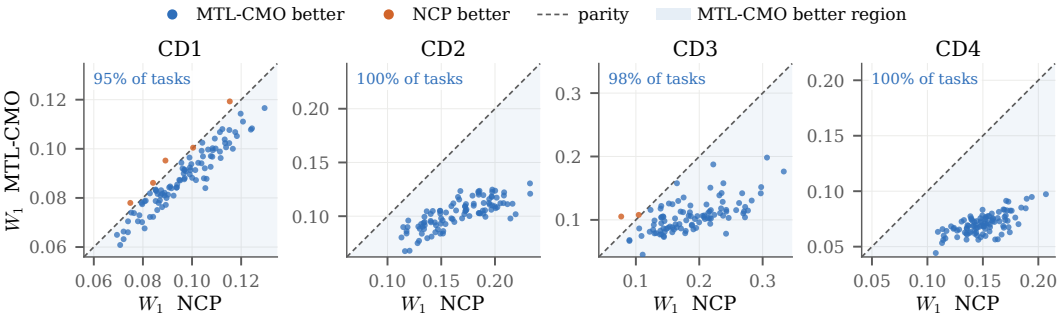

Figure 1: Comparison between MTL-CMO and the single-task NCP for representing CDFs on four conditional distribution families. Performances measured with the 1-Wasserstein to the ground truth.

4 Experimental results

4.1 Uncertainty quantification: Conditional distributions estimation

Experiment objective. We evaluate MTL-CMO, and subsequently T-CMO, performances for uncertainty quantification on population of related synthetic conditional distributions by comparing estimated conditional cumulative distribution functions (CCDFs) to the ground truths.

Datasets. We consider four families of conditional distributions, where the conditioning variable modifies the conditional expectation. Once conditioned, **CD1** and **CD2** are bimodal Gaussian mixtures with task-dependent mixture weights, amplitudes and scales. **CD3** are Student-t distributions with task-dependent location, scale and degrees of freedom. **CD4** are skew-normal distributions with task-dependent location, scale and skewness. While **CD1** has an univariate conditioning others have a multivariate conditioning. Thus, **CD1** to **CD4** progressively introduce complex conditioning, heavy tails, and asymmetric conditional distributions. Families are detailed in Appendix D.1.

Multi-task learning evaluation. We sample $K{=}100$ conditional distributions and generate $n{=}400$ observations for each task and family. We compare MTL-CMO against 3 multi-task and 8 single-task alternatives detailed in Appendix D.5. The experiment is repeated over 10 seeds and Appendix D.2 provides implementation details. Performance is measured with the 1-Wasserstein distance between the estimated and the ground-truth CCDFs averaged over the population and conditioning inputs.

Results are reported in Table 1. MTL-CMO outperforms all competing methods on CD1–CD3 and remains competitive on CD4. Relative to its single-task counterpart, NCP, joint representation learning reduces the Wasserstein error by 8%, 39%, 43%, and 53% on CD1–CD4, respectively. This behavior is further illustrated in Figure 1, which compares the performance of MTL-CMO and to NCP across the four families for different number of training sampels per task. The improve-

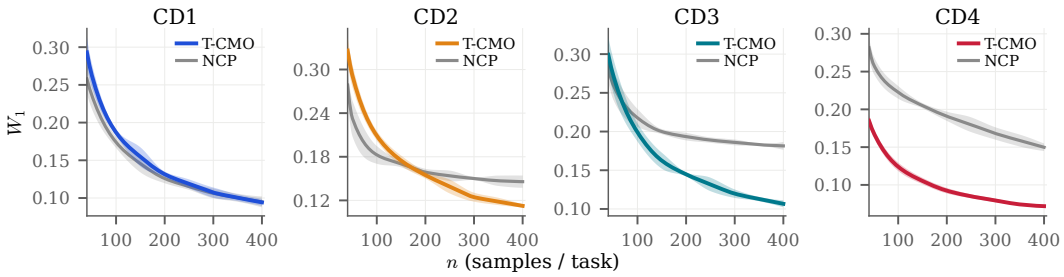

Figure 2: Transfer learning evaluation on unseen conditional distributions. Comparison in 1-Wasserstein between T-CMO and the single-task NCP depending on the number of observations.

Table 2: Reconstruction error between estimated and ground truth operators. Hilbert-Schmidt error in $<\text{mean}> \pm <\text{std}>$ over the 150 unseen systems. Lower is better. **First** and second.

	Trajectory length N				
Method	**5 000**	**10 000**	**20 000**	**40 000**	**80 000**
T-CMO **(ours)**	$\mathbf{0.261 \pm 0.148}$	$\mathbf{0.190 \pm 0.108}$	$\mathbf{0.140 \pm 0.075}$	$\mathbf{0.099 \pm 0.055}$	$\mathbf{0.075 \pm 0.039}$
Single-task NCP (Kostic et al., 2024)	0.297 ± 0.138	0.224 ± 0.098	0.176 ± 0.068	0.138 ± 0.047	0.115 ± 0.031
Pooled NCP (Kostic et al., 2024)	0.306 ± 0.143	0.219 ± 0.102	0.163 ± 0.078	0.118 ± 0.058	$\underline{0.087 \pm 0.041}$
RFF/EDMD (Li et al., 2017)	0.295 ± 0.140	0.233 ± 0.123	0.177 ± 0.069	0.135 ± 0.045	0.109 ± 0.030
RFF/Laplace (Kostic et al., 2025)	0.296 ± 0.141	0.221 ± 0.100	0.172 ± 0.072	0.132 ± 0.051	0.106 ± 0.034
MetaKoopman (Iwata & Kawahara, 2021)	0.288 ± 0.142	0.220 ± 0.100	0.173 ± 0.070	0.136 ± 0.048	0.115 ± 0.034
VAMPNet (Mardt et al., 2018)	$\underline{0.273 \pm 0.145}$	$\underline{0.205 \pm 0.104}$	$\underline{0.154 \pm 0.071}$	$\underline{0.110 \pm 0.051}$	0.088 ± 0.034

ment becomes substantially larger for the more challenging families, which involve multivariate conditioning, heavy tails, or asymmetric distributions. These results highlight the benefit of sharing information across related tasks through shared input/output function spaces, particularly when each task has not enough observations for accurate estimation.

Transfer learning evaluation. We next freeze the input and output function spaces learned during multi-task training. For each family and dataset sizes between 50 and 400, we sample observations from 100 previously unseen conditional distributions. We estimate their conditional CDFs using the closed-form T-CMO estimator and compare it with NCP models trained independently from scratch for each task and sample size.
Results are shown in Figure 2. On the simpler CD1 family, T-CMO and NCP achieve comparable performance. As the conditional families become more complex (CD2–CD4), T-CMO increasingly outperforms NCP as additional target observations become available. The reduced performance of T-CMO in the very low-data regime for some families is consistent with the greater numerical sensitivity of the empirical Gram-matrix inversion when few samples are available. Overall, these results show that function spaces learned from a diverse population of source tasks can be effectively reused to estimate the conditional distributions of previously unseen tasks (Tripuraneni et al., 2020). Implementation details and additional results are provided in Appendix D.4.

4.2 Müller-Brown Langevin dynamics

Müller-Brown dynamical family. The Müller-Brown family is a standard benchmark for slow dynamic of metastable systems and transition-path problems (Müller & Brown, 1979; Zhang et al., 2022). It consists in overdamped Langevin dynamics with two-dimensional energy landscapes comprising three wells separated by energy barriers. Typical trajectories spend long periods within metastable basins and only rarely transition between them. Consequently, the dominant nontrivial eigenfunctions of the Infinitesimal Generator (IG, see Section 2) characterize the principal metastable states, while the corresponding eigenvalues encode the transition timescales. We construct a family of such systems by varying the orientation and depth of one potential well. These variations alter the metastable regions and transition timescales, leading to task-specific spectral decompositions of the corresponding infinitesimal generators.

Experimental protocol. The experiment comprises three stages: (i) learning the shared functional space with MTL-CMO from 750 systems, (ii) finetuning the hyperparameters of the IG estimator with 50 dynamics, and (iii) estimating the IG of 150 unseen dynamics by transferring the function

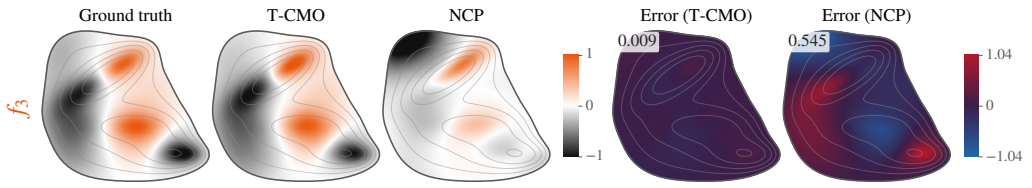

Figure 3: Recovery of 3rd eigenfunction of an unseen Müller-Brown system with T-CMO/Laplace and its single-task counterpart NCP/Laplace. Global error in cosine distance.

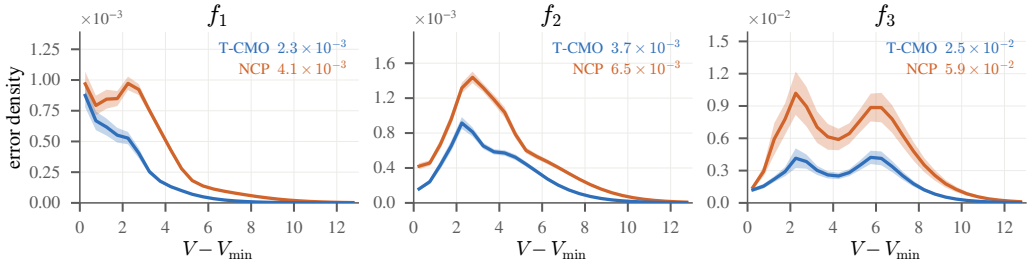

Figure 4: Reconstruction error across eigenfunctions and potential values between T-CMO/Laplace and NCP/Laplace. Error averaged over the 150 unseen systems. Global error in cosine distance.

space with T-CMO. Each system provides a trajectory of 80,000 observations. Transfer performance is evaluated from truncated trajectories from $2k$ to $80k$ observations. Since Langevin dynamics are time-reversible, the IGs are self-adjoint and we consider a single function dictionary, i.e. $\Phi_\theta = \Psi_\theta$. The IGs are estimated with the Laplace estimator (Kostic et al., 2025). Stage (ii) shows that the first three non-trivial components capture 95% of the spectral energy, motivating our focus on rank-3 spectral recovery. For transfer learning, we compare T-CMO/Laplace against: single-task NCP/Laplace, pooled NCP/Laplace, RFF/Laplace, RFF/EDMD, MetaKoopman, and VampNet. Implementation details are provided in Appendix E.1.

Results. The multi-task results reported in Appendix E.2 show that MTL-CMO improves operator reconstruction over its single-task counterpart, NCP. In transfer learning, T-CMO/Laplace achieves the lowest operator reconstruction error (Table 2) and the most accurate recovery of the individual spectral components (Appendix E.3) across all trajectory lengths. More specifically, Figure 3 shows that T-CMO accurately recovers the spatial structure of the third eigenfunction, whereas NCP, exhibits substantially larger localized errors. Figure 4 shows a consistent improvement of T-CMO across all three eigenfunctions and potential values, reducing the overall error by approximately a factor of two. These results demonstrate that transferring the learned function space is particularly beneficial for recovering challenging spectral components of unseen dynamics from limited trajectory data. Additional results can be found in Appendix E.3.

4.3 Turbulent plasma dynamics

Plasma dynamics and reduced-model identification. Turbulent plasmas exhibit strongly nonlinear and multiscale interactions that make high-fidelity simulations computationally demanding. They are therefore often approximated by reduced physical models that retain the dominant mechanisms while enabling faster simulations, which is valuable for applications requiring rapid predictions, such as plasma monitoring and control. Their practical use, however, requires inferring model parameters from observed trajectories, a challenging problem because turbulence and instabilities can obscure their effects on the dynamics (Boeuf & Smolyakov, 2023; Coosemans et al., 2021). This has motivated data-driven operator approaches for extracting physical information from plasma measurements and simulations (Faraji et al., 2024; Taylor et al., 2018).

Experiment objective. We consider Tokam2D (GYSELAX Team, 2026), a plasma simulator based on a reduced two-field model describing the evolution of plasma density and electrostatic potential in a two-dimensional plane. The dynamics are parametrized by g, which controls the inter-

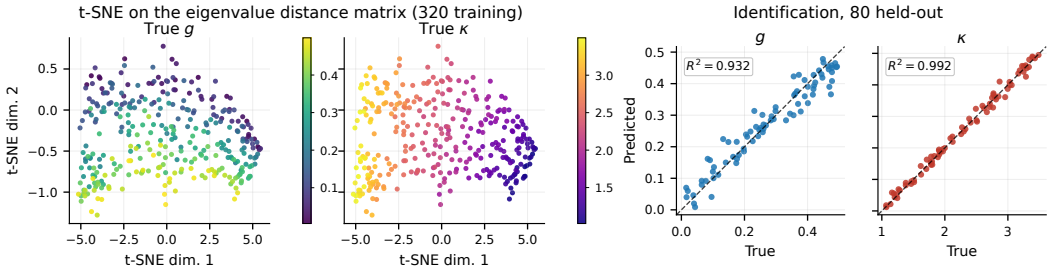

Figure 5: Left: t-SNE on the distance matrix between the 320 training operator spectra colored by g and κ. Right: parameter identification with kernel regression evaluated on the 80 held-out systems.

change coupling between density and vorticity, and κ, which controls the imposed density-gradient drive (Ghendrih et al., 2018; 2022). Varying these parameters produces distinct turbulent regimes. We investigate whether g and κ can be recovered from the spectra of IGs estimated through T-CMO. Appendix F.1 provides a detailed descriptions.

Experimental protocol. We simulate 400 plasma trajectories of 4,093 observations on a 64×64 spatial grid, with (g, κ) sampled uniformly from $[0, 0.5] \times [1, 3.5]$. We train MTL-CMO on 320 trajectories using a convolutional ResNet that produces a dictionary of 128 functions. For IG estimation, we combine T-CMO with the Laplace estimator (Kostic et al., 2025). We regress g and κ from the estimated IG eigenvalues on the training set using kernel ridge regression. On the 80 held-out trajectories of increasing length, we compare our approach against eigenvalues obtained with MIDST (Elul et al., 2024), pooled NCP/Laplace (Kostic et al., 2024), and RFF/Laplace (Brault et al., 2016). Implementation details are provided in Appendix F.2.

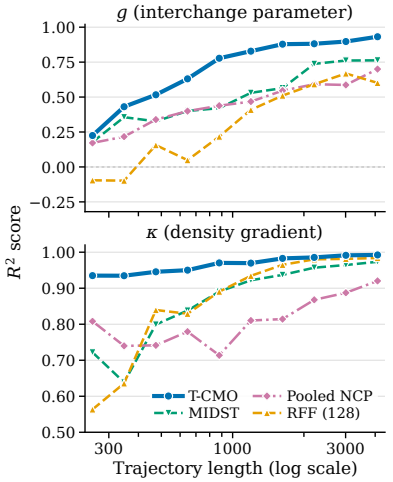

Figure 6: Identification of the parameters g and κ versus trajectory length for T-CMO and competing methods.

Results. Figure 5-left shows a t-SNE embedding of the IG eigenvalue signatures estimated by MTL-CMO on the 320 training systems. Although g and κ are not provided during training, the resulting representations vary smoothly with both parameters. On the 80 held-out systems and using full trajectories, T-CMO/Laplace can be used to predict with R^2=0.932 for g and R^2=0.992 for κ (Figure 5-right). Moreover, Figure 6 shows that it outperforms all baselines across all trajectory lengths with an early near-perfect recovery of κ and rapidly improving the estimation of g compared to other baselines. These results show the benefit of jointly learning and transferring functional spaces for parameter identification in plasma dynamics. In appendices F.3 and F.4 we provide additional results.

5 Conclusion

In this work, we introduced MTL-CMO, a multi-task approach that jointly learns conditional mean operators through shared function spaces and task-specific linear operators. We then proposed T-CMO, which reuses the learned function spaces to estimate the operator of a new distribution in closed form. Our statistical results and experiments on uncertainty quantification and complex dynamical systems demonstrate the benefits of jointly learning the shared function spaces.
In future work we will investigate new strategies to improve the estimation of the task-specific low-rank operators by leveraging the population manifold structure of their matrix representations.

6 CONDITIONS

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing, assist for code implementation and for retrieval and discovery of related works. We have not used generative AI tools for research ideation or execution, draft sections of the paper, generating synthetic datasets and proving mathematical claims. We have reviewed all AI-assisted work. Particularly, we independently reviewed all references suggested by generative AI tools during our related-work search. We also reviewed, tested, and validated all AI-assisted code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The authors have identified no specific ethical concerns associated with this work.

REPRODUCIBILITY STATEMENT

All code required to reproduce our results is included in the supplementary material and will be made publicly available upon acceptance. The proposed methods (MTL-CMO and T-CMO) and their implementation are described in detail in Section 3 and Appendix B. Experimental protocols, implementation details, and hyperparameter settings are reported in Appendix D for the uncertainty quantification experiment, Appendix E for the Langevin experiment, and Appendix F for the turbulent plasma dynamics experiment.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Rie Kubota Ando, Tong Zhang, and Peter Bartlett. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of machine learning research*, 6(11), 2005.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine learning*, 73(3):243–272, 2008.

Christopher M. Bishop. Mixture density networks. Technical Report NCRG/94/004, Aston University, 1994.

Jean-Pierre Boeuf and Andrei Smolyakov. Physics and instabilities of low-temperature $E \times B$ plasmas for spacecraft propulsion and other applications. *Physics of Plasmas*, 30(5):050901, 2023. doi: 10.1063/5.0145536.

Romain Brault, Markus Heinonen, and Florence Buc. Random fourier features for operator-valued kernels. In *Asian Conference on Machine Learning*, pp. 110–125. PMLR, 2016.

Leo Breiman and Jerome H Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598, 1985.

Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

Reinart Coosemans, Wouter Dekeyser, and Martine Baelmans. Bayesian analysis of turbulent transport coefficients in 2D interchange dominated $E \times B$ turbulence involving flow shear. *Journal of Physics: Conference Series*, 1785:012001, 2021. doi: 10.1088/1742-6596/1785/1/012001.

Víctor H. de la Peña and Evarist Giné. *Decoupling: From Dependence to Independence*. Springer, New York, 1999.

Timothée Devergne, Vladimir R. Kostic, Michele Parrinello, and Massimiliano Pontil. From biased to unbiased dynamics: An infinitesimal generator approach. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2404.

Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=pW2Q2xLwIMD.

Tony Duan, Anand Avati, Daisy Yi Ding, Khanh K Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. Ngboost: Natural gradient boosting for probabilistic prediction. In *International Conference on Machine Learning*, pp. 2690–2700. PMLR, 2020.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, volume 32, pp. 7511–7522, 2019.

Yonatan Elul, Eyal Rozenberg, Amit Boyarski, Yael Yaniv, Assaf Schuster, and Alex M Bronstein. Data-driven modeling of interrelated dynamical systems. *Communications Physics*, 7(1):141, 2024.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 109–117, 2004.

Theodoros Evgeniou, Charles A Micchelli, Massimiliano Pontil, and John Shawe-Taylor. Learning multiple tasks with kernel methods. *Journal of machine learning research*, 6(4):615, 2005.

Farbod Faraji, Maryam Reza, Aaron Knoll, and J. Nathan Kutz. Dynamic mode decomposition for data-driven analysis and reduced-order modelling of $E \times B$ plasmas: I. extraction of spatiotemporally coherent patterns. *Journal of Physics D: Applied Physics*, 57:065201, 2024. doi: 10.1088/1361-6463/ad0910.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan): 73–99, 2004.

Thibaut Germain, Rémi Flamary, Vladimir R. Kostic, and Karim Lounici. A spectral-grassmann wasserstein metric for operator representations of dynamical systems. In *International Conference on Learning Representations*, 2026.

Roger Ghanem, David Higdon, and Houman Owhadi (eds.). *Handbook of Uncertainty Quantification*. Springer, 2017. doi: 10.1007/978-3-319-12385-1.

P. Ghendrih, Y. Asahi, E. Caschera, G. Dif-Pradalier, P. Donnel, X. Garbet, C. Gillot, V. Grandgirard, G. Latu, Y. Sarazin, et al. Generation and dynamics of SOL corrugated profiles. *Journal of Physics: Conference Series*, 1125:012011, 2018. doi: 10.1088/1742-6596/1125/1/012011.

Philippe Ghendrih, Guilhem Dif-Pradalier, Olivier Panico, Yanick Sarazin, Hugo Bufferand, Guido Ciraolo, Peter Donnel, Nicolas Fedorczak, Xavier Garbet, Virginie Grandgirard, et al. Role of avalanche transport in competing drift wave and interchange turbulence. *Journal of Physics: Conference Series*, 2397:012018, 2022. doi: 10.1088/1742-6596/2397/1/012018.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 297–299. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/golowich18a.html.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1803–1810, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

GYSELAX Team. Tokam2D: A 2D spectral solver for turbulence schemes, 2026. URL https://github.com/gyselax/tokam2d. Accessed: 2026-05-01.

Minghao Han, Kiwan Wong, Adrian Wing-Keung Law, and Xunyuan Yin. Mako: Meta-adaptive koopman operators for learning-based model predictive control of parametrically uncertain nonlinear systems. *Automatica*, 186:112827, 2026.

Akira Hasegawa and Masahiro Wakatani. Plasma edge turbulence. *Physical Review Letters*, 50(9): 682–686, 1983. doi: 10.1103/PhysRevLett.50.682.

Maximiliano Hertel, Ilja Klebanov, Manuel Schaller, and Karl Worthmann. Verifiable regularity criterion for conditional expectation operators and conditional mean embeddings with applications to nonparametric regression, bayesian inverse problems, and koopman operators. *arXiv preprint arXiv:2608.06155*, 2026.

Tomoharu Iwata and Yoshinobu Kawahara. Meta-learning for koopman spectral analysis with short time-series. *arXiv preprint arXiv:2102.04683*, 2021.

Rafael Izbicki and Ann B Lee. Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 11(2):2800–2831, 2017.

Bernard O Koopman. Hamiltonian systems and transformation in hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

Vladimir Kostic, Pietro Novelli, Andreas Maurer, Carlo Ciliberto, Lorenzo Rosasco, and Massimiliano Pontil. Learning dynamical systems via koopman operator regression in reproducing kernel hilbert spaces. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4017–4031, 2022.

Vladimir Kostic, Karim Lounici, Pietro Novelli, and Massimiliano Pontil. Sharp spectral rates for koopman operator learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 32328–32339, 2023.

Vladimir R. Kostic, Karim Lounici, Grégoire Pacreau, Giacomo Turri, Pietro Novelli, and Massimiliano Pontil. Neural conditional probability for uncertainty quantification. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1950. URL https://papers.nips.cc/paper_files/paper/2024/hash/705b97ecb07ae86524d438abac97a3e2-Abstract-Conference.html.

Vladimir R. Kostic, Karim Lounici, Hélène Halconruy, Timothée Devergne, Pietro Novelli, and Massimiliano Pontil. Laplace transform based low-complexity learning of continuous markov semigroups. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31560–31589. PMLR, 2025.

Guillaume Le Treut, Sarah Ancheta, Greg Huber, Henri Orland, and David Yllanes. Markov-bridge generation of transition paths and its application to cell-fate choice. *Physical Review Research*, 7 (1):013010, 2025. doi: 10.1103/PhysRevResearch.7.013010.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

Qianxiao Li, Felix Dietrich, Erik M Bollt, and Ioannis G Kevrekidis. Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the koopman operator. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(10), 2017.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. VAMPnets for deep learning of molecular kinetics. *Nature Communications*, 9:5, 2018. doi: 10.1038/s41467-017-02388-1.

Andreas Maurer, Massi Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. In *International conference on machine learning*, pp. 343–351. PMLR, 2013.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Mattes Mollenhauer and Péter Koltai. Nonparametric approximation of conditional expectation operators. *arXiv preprint arXiv:2012.12917*, 2020.

Krikamol Muandet, Kenji Fukumizu, Bharath K. Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1–2):1–141, 2017. doi: 10.1561/2200000060.

Klaus Müller and L. D. Brown. Location of saddle points and minimum energy paths by a constrained simplex optimization procedure. *Theoretica chimica acta*, 53(1):75–93, 1979. doi: 10.1007/BF00547608.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*, volume 60 of *Texts in Applied Mathematics*. Springer, 2014. doi: 10.1007/978-1-4939-1323-7.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Filipe Rodrigues and Francisco C Pereira. Beyond expectation: Deep joint mean and quantile regression for spatiotemporal problems. *IEEE transactions on neural networks and learning systems*, 31(12):5377–5389, 2020.

Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656:5–28, 2010.

Xinwei Shen and Nicolai Meinshausen. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3):653–677, 2025.

Eiki Shimizu, Kenji Fukumizu, and Dino Sejdinovic. Neural-kernel conditional mean embeddings. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45040–45059. PMLR, 2024.

Burton Singer and Seymour Spilerman. The representation of social processes by markov models. *American journal of sociology*, 82(1):1–54, 1976.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th annual international conference on machine learning*, pp. 961–968, 2009.

Roy Taylor, J. Nathan Kutz, Kyle D. Morgan, and Brian A. Nelson. Dynamic mode decomposition for plasma diagnostics and validation. *Review of Scientific Instruments*, 89(5):053501, 2018. doi: 10.1063/1.5027419.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International conference on machine learning*, pp. 10434–10443. Pmlr, 2021.

Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26:359–372, 1964.

Bailing Zhang. Koopman framework with self-supervised spectral alignment for multi-domain time-series modeling and prediction. *Neurocomputing*, pp. 131109, 2025.

Wei Zhang, Tiejun Li, and Christof Schütte. Solving eigenvalue pdes of metastable diffusion processes using artificial neural networks. *Journal of Computational Physics*, 465:111377, 2022.

Yi Zhang and Jeff Schneider. Learning multiple tasks with a sparse matrix-normal penalty. *Advances in neural information processing systems*, 23, 2010.

A Operator representation of dynamical systems & estimation

A.1 Operator representation of dynamical systems & spectral decomposition

Markov semigroup and transfer operator. Let $(X_t)_{t\geq 0}$ be a time-homogeneous Markov process on a measurable state space $\mathcal{X}$, and invariant measure π. Let $\mathcal{F} \subset \mathcal{L}^2_\pi(\mathcal{X})$ be a space of observables. The evolution of the system is described by the Markov semigroup $\{A_t\}_{t\geq 0}$ acting on $\mathcal{F}$, such as

$$[A_t f](x) = \mathbb{E}[f(X_t) \mid X_0 = x], \quad f \in \mathcal{F}. \tag{25}$$

Infinitesimal generator. The family $\{A_t\}_{t\geq 0}$ is a strongly continuous semigroup satisfying

$$A_0 = \mathrm{Id}, \qquad A_{t+s} = A_t A_s, \quad \forall t, s \geq 0.$$

The infinitesimal generator G associated with $\{A_t\}_{t\geq 0}$ is defined as

$$Gf \triangleq \lim_{t\to 0^+} \frac{A_t f - f}{t}, \tag{26}$$

with domain $\mathcal{D}(G) \subset \mathcal{F}$. The generator is, in general, an unbounded linear operator on $\mathcal{F}$ and characterizes the semigroup through $A_t = \exp(tG)$.

Spectral decomposition via compact resolvent. The spectral analysis of G requires additional structure due to its infinite-dimensional and unbounded nature. Let π be an invariant measure of the process and consider the Hilbert space $\mathcal{L}^2_\pi(\mathcal{X})$.

We assume that G has compact resolvent, i.e., there exists μ_0 in the spectrum of G such that $(\mu_0 \,\mathrm{Id} - G)^{-1}$ is compact on $\mathcal{L}^2_\pi(\mathcal{X})$. Under this assumption, the spectrum of G is discrete and consists of isolated eigenvalues $(\lambda_i)_{i\in\mathbb{N}}$ with finite multiplicity and no accumulation point except possibly at infinity.

In addition, there exists an bi-orthogonal basis $(f_i, g_i)_{i\in\mathbb{N}}$ of $\mathcal{L}^2_\pi(\mathcal{X})$ such that

$$Gf_i = \lambda_i f_i \qquad G^* g_i = \overline{\lambda_i} g_i \qquad \langle f_i, g_j \rangle_{\mathcal{L}^2_\pi(\mathcal{X})} = \delta_{ij} \tag{27}$$

For each isolated eigenvalue λ_i, the associated spectral projector P_i is defined as the projector onto the corresponding eigenspace in $\mathcal{L}^2_\pi(\mathcal{X})$.

In the simple eigenvalue case, we have

$$P_i f = \langle f, g_i \rangle_{\mathcal{L}^2_\pi(\mathcal{X})} f_i. \tag{28}$$

More generally, if λ_i has multiplicity m_i and $(f_{i,k}, g_{i,k})_{k=1}^{m_i}$ is an bi-orthogonal basis of the eigenspace, then

$$P_i f = \sum_{k=1}^{m_i} \langle f, g_{i,k} \rangle_{\mathcal{L}^2_\pi(\mathcal{X})} f_{i,k}. \tag{29}$$

The projectors satisfy

$$P_i P_j = \delta_{ij} P_i, \qquad \sum_i P_i = \mathrm{Id} \tag{30}$$

on the spectral subspace.

With this notation, the generator admits the spectral representation

$$Gf = \sum_i \lambda_i P_i f, \tag{31}$$

for all $f \in \mathcal{D}(G)$ for which the expansion is well-defined, and the semigroup diagonalizes as

$$A_t f = \sum_i e^{\lambda_i t} P_i f. \tag{32}$$

A.2 Estimation through generator resolvent

This section presents the methodology introduced in Kostic et al. (2025).

Learning initial generators via spectral filtering. The first step consists in learning a collection of initial generator atoms directly from trajectory data by estimating spectral components of the infinitesimal generator G. Rather than approximating G through finite differences of the form $(A_{\Delta t} - I)/\Delta t$, which is unstable at small time scales, we rely on spectral filtering based on Toeplitz representations of analytic functions of the generator.

Let $A_{\Delta t} = e^{\Delta t G}$ denote the transfer operator at time step Δt. We consider a Toeplitz symbol

$$T(z) = \sum_{j \in \mathbb{Z}} a_j z^j, \tag{33}$$

and define the associated filtered operator

$$F(G) = T(A_{\Delta t}). \tag{34}$$

Since $A_{j\Delta t} = A_{\Delta t}^j$, the operator $F(G)$ admits the representation

$$F(G) = \sum_{j \in \mathbb{Z}} a_j A_{\Delta t}^j, \tag{35}$$

which can be approximated from data using time-lagged observations. In practice, we use a truncated expansion

$$F_\ell(G) = \sum_{|j| \leq \ell} a_j A_{\Delta t}^j, \tag{36}$$

whose empirical action is represented by a banded Toeplitz matrix acting on the time-ordered data sequence. Thus, analytic functional calculus of the generator reduces to structured linear algebra on trajectory data.

Resolvent and exponential filters. To probe the spectrum of the generator G, we use families of analytic filters.

The key object is the generator resolvent

$$R_\mu = (\mu I - G)^{-1} = \int_0^\infty e^{tG} e^{-\mu t}\, dt, \tag{37}$$

which we approximate by a Toeplitz-weighted sum of transfer operators,

$$R_{\mu,\ell} = \sum_{j=0}^{\ell} a_j A_{\Delta t}^j, \qquad a_j \approx \Delta t\, e^{-\mu j \Delta t}, \tag{38}$$

corresponding to a discretization of the Laplace transform.

Equivalently, using the transfer-operator resolvent, we have

$$(e^\mu I - A_{\Delta t})^{-1} = \sum_{j=0}^{\infty} e^{-(j+1)\mu} A_{\Delta t}^j, \tag{39}$$

which naturally yields Toeplitz coefficients. These filters concentrate spectral information near the shift parameter μ, allowing us to localize different regions of the spectrum.

In addition, exponential and trigonometric filters (e.g. $e^{\Delta t G}$, $\cosh(\Delta t G)$, $\sinh(\Delta t G)$) can be used to emphasize specific spectral structures depending on the nature of the dynamics.

Estimation from data. Algorithm 1 describes the procedure to estimate the infinitesimal generator according to (Kostic et al., 2025).

Algorithm 1 Generator estimation from T-CMO transferred observables

Require: Trajectory $\{x_t^\star\}_{t=0}^{T-1}$, rank-$r$ observable basis $U^\star = \{u_j^\star\}_{j=1}^r$ returned by T-CMO, sampling step Δt, resolvent shift μ, maximum lag ℓ, rank q, regularization γ

1: Evaluate $\mathbf{z}_t^\star \leftarrow \big(u_1^\star(x_t^\star), \dots, u_r^\star(x_t^\star)\big)^\top \in \mathbb{R}^r$
2: Form $\mathbf{Z}^\star \leftarrow (\mathbf{z}_0^\star, \dots, \mathbf{z}_{T-1}^\star)^\top \in \mathbb{R}^{T\times r}$
3: Construct the truncated Laplace weights $a_k \leftarrow \mu\Delta t\, e^{-\mu k\Delta t}, \quad k = 0, \dots, \ell$
4: Form the associated Toeplitz operator $\mathbf{T}_\mu$ and compute $\mathbf{Z}_\mu^\star \leftarrow \mathbf{T}_\mu \mathbf{Z}^\star$
5: Compute $\widehat{\mathbf{C}}^\star \leftarrow \frac{1}{T}\mathbf{Z}^{\star\top}\mathbf{Z}^\star$, $\widehat{\mathbf{H}}_\mu^\star \leftarrow \frac{1}{T}\mathbf{Z}^{\star\top}\mathbf{Z}_\mu^\star$
6: Regularize $\widehat{\mathbf{C}}_\gamma^\star \leftarrow \widehat{\mathbf{C}}^\star + \gamma\mathbf{I}_r$
7: Compute the q dominant generalized eigenvectors of $\widehat{\mathbf{H}}_\mu^\star\widehat{\mathbf{H}}_\mu^{\star\top}\mathbf{v} = \sigma^2\widehat{\mathbf{C}}_\gamma^\star\mathbf{v}$ and collect the normalized directions in $\mathbf{V}_q^\star$
8: Form the reduced resolvent representation $\widehat{\mathbf{R}}_\mu^\star \leftarrow \mathbf{V}_q^{\star\top}\widehat{\mathbf{H}}_\mu^\star\mathbf{V}_q^\star$
9: Compute $\widehat{\mathbf{R}}_\mu^\star\mathbf{r}_i^\star = \nu_i^\star\mathbf{r}_i^\star$
10: Recover $\mathbf{v}_i^\star \leftarrow \mathbf{V}_q^\star\mathbf{r}_i^\star$
11: Recover the generator eigenvalues $\widehat{\lambda}_i^\star \leftarrow \mu\left(1 - \frac{1}{\nu_i^\star}\right)_r$
12: Define the corresponding eigenfunctions $\widehat{f}_i^\star(x) \leftarrow \sum_{j=1}[\mathbf{v}_i^\star]_j\, u_j^\star(x)$
13: **return** $\{\widehat{\lambda}_i^\star\}_{i=1}^q$ and $\{\widehat{f}_i^\star\}_{i=1}^q$

A.3 From T-CMO to estimation of operator representation of dynamical system.

Estimation of a dynamical system's operator representation requires a function space (observable space) that best captures the dynamics as depicted in Section 2 and Appendix A.1. By construction, MTL-CMO, introduced in Section 3, captures function spaces that describe the dominant mode of variation in a population of dynamical systems. These learned function spaces can be adapted via transfer with T-CMO to estimate the operator representation of a new, related dynamical system.

Formally, consider the learned features maps $(\Phi_{\widehat{\theta}}, \Psi_{\widehat{\theta}})$ with MTL-CMO on a population of related dynamical systems. Suppose a sampled trajectory $\{x_i^{\text{new}}\}_{i\in[T]} \subset \mathcal{X}$ from a new dynamical system with invariant measure π_{new}. Then,

1. With T-CMO estimate the new operator $\mathsf{D}_{\mathrm{Y|X}}^{\text{new}}$, see Section 3.2 and Algorithm 3.
2. Derive its singular value decomposition following Section 3.1 (SVD paragraph) and keep the orthonormal function basis $U^{\text{new}} = \{u_i^{\text{new}}\}_{i\in[r_{\text{new}}]}$.
3. Estimate the infinitesimal generator with the function space $\text{span}(U^{\text{new}}) \subset \mathcal{L}^2_{\pi_{\text{new}}}(\mathcal{X})$ with any infinitesimal generator estimator.

Various infinitesimal generator estimators exist, including: (Schmid, 2010; Li et al., 2017; Kostic et al., 2022; 2025). In particular, we describe Kostic et al. (2025) in Appendix A.2.

B Multi-task learning for conditional mean operator: Appendix

We give additional details on the finite-dimensional representation of the task-specific operators, the population objective optimized by MTL-CMO, and the resulting transfer estimator.

Post-hoc Singular Value Decomposition of deflated operators. For each task k, the singular value decomposition of the deflated operator can be recovered *post hoc* by centering and whitening the shared dictionaries. Formally, consider the orthogonal projectors

$$P_{\mu_k} \triangleq I_d - \mathbf{1}_\mathcal{X} \otimes_{\mu_k} \mathbf{1}_\mathcal{X} \qquad P_{\nu_k} \triangleq I_d - \mathbf{1}_\mathcal{Y} \otimes_{\nu_k} \mathbf{1}_\mathcal{Y},$$

where $\mathbf{1}_\mathcal{X}$ and $\mathbf{1}_\mathcal{Y}$ are indicator functions. These projectors center functions, i.e.

$$P_{\mu_k} f = f - \mathbb{E}_{\mu_k}[f(X)] \qquad P_{\nu_k} g = g - \mathbb{E}_{\nu_k}[g(Y)].$$

Assuming an unconstrained operator $\mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta}$, we can define the centered deflated operator:

$$\tilde{\mathsf{D}}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta} \triangleq P_{\mu_k}\, \mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta}\, P_{\mu_k} = \mathsf{S}_{\Phi_{\theta,c}}^{(k)} \mathbf{A}^{(k)} \mathbf{B}^{(k)} \mathsf{S}_{\Psi_{\theta,c}}^{(k)*},$$

with $\Phi_{\theta,c}^{(k)}(x) = \Phi_\theta(x) - \mathbb{E}_{\mu_k}[\Phi_\theta(X)]$ and $\Psi_{\theta,c}^{(k)}(y) = \Psi_\theta(y) - \mathbb{E}_{\nu_k}[\Psi_\theta(Y)]$. In fact, the centered deflated operator is a better estimator of $\mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k})}$. Indeed,

$$\left\| \tilde{\mathsf{D}}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta} - \mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k})} \right\|_{\mathrm{HS}} \leq \left\| \mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta} - \mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k})} \right\|_{\mathrm{HS}},$$

since $P_{\mu_k}\, \mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k})}\, P_{\nu_k} = \mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k})}$ and projectors are contractant maps. Assume the centered Gram-matrices

$$\mathbf{G}_{\Phi_{\theta,c}}^{(k)} \triangleq \mathbb{E}_{\mu_k}\left[\Phi_{\theta,c}^{(k)}(X)\Phi_{\theta,c}^{(k)}(X)^\top\right], \qquad \mathbf{G}_{\Psi_{\theta,c}}^{(k)} \triangleq \mathbb{E}_{\nu_k}\left[\Psi_{\theta,c}^{(k)}(Y)\Psi_{\theta,c}^{(k)}(Y)^\top\right],$$

Consider the matrix SVD

$$\mathbf{G}_{\Phi_{\theta,c}}^{(k)1/2} \mathbf{A}^{(k)} \mathbf{B}^{(k)\top} \mathbf{G}_{\Psi_{\theta,c}}^{(k)1/2} = \mathbf{U}^{(k)}\, \mathrm{Diag}(\boldsymbol{\sigma}^{(k)}) \mathbf{V}^{(k)\top}.$$

Since

$$\mathbf{U}^{(k)\top}\mathbf{U}^{(k)} = \mathbf{I}_{r_k}, \qquad \mathbf{V}^{(k)\top}\mathbf{V}^{(k)} = \mathbf{I}_{r_k},$$

the coefficient matrices

$$\mathbf{C}^{(k)} = \mathbf{G}_{\Phi_\theta}^{(k)-1/2}\mathbf{U}^{(k)}, \qquad \mathbf{D}^{(k)} = \mathbf{G}_{\Psi_\theta}^{(k)-1/2}\mathbf{V}^{(k)}$$

satisfy

$$\mathbf{C}^{(k)\top}\mathbf{G}_{\Phi_\theta}^{(k)}\mathbf{C}^{(k)} = \mathbf{I}_{r_k}, \qquad \mathbf{D}^{(k)\top}\mathbf{G}_{\Psi_\theta}^{(k)}\mathbf{D}^{(k)} = \mathbf{I}_{r_k}.$$

Denoting $\mathbf{u}_i^{(k)}$ and $\mathbf{v}_i^{(k)}$ the ith columns of $\mathbf{U}^{(k)}$ and $\mathbf{V}^{(k)}$, the singular functions are given by

$$\widetilde{u}_i^{(k)} = \mathsf{S}_{\Phi_{\theta,c}}^{(k)} \mathbf{G}_{\Phi_{\theta,c}}^{(k)-1/2} \mathbf{u}_i^{(k)}, \qquad \widetilde{v}_i^{(k)} = \mathsf{S}_{\Psi_{\theta,c}}^{(k)} \mathbf{G}_{\Psi_{\theta,c}}^{(k)-1/2} \mathbf{v}_i^{(k)},$$

where $\mathbf{G}_{\Phi_{\theta,c}}^{(k)-1/2}$ and $\mathbf{G}_{\Psi_{\theta,c}}^{(k)-1/2}$ are square root inverses. By construction, the functions $\{\widetilde{u}_i^{(k)}\}_{i\in[r_k]}$ and $\{\widetilde{v}_i^{(k)}\}_{i\in[r_k]}$ are orthonormal in their respective space $\mathrm{L}^2_{\mu_\mathrm{k}}(\mathcal{X})$ and $\mathrm{L}^2_{\nu_\mathrm{k}}(\mathcal{Y})$, and therefore yield the task-specific singular decomposition

$$\tilde{\mathsf{D}}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta} = \textstyle\sum_{i=1}^{r_k} \sigma_i^{(k)} \widetilde{u}_i^{(k)} \otimes \widetilde{v}_i^{(k)}. \tag{40}$$

Thus, the proposed factorization avoids imposing orthogonality constraints during training while retaining access to an orthonormal, task-specific spectral representation after optimization.

Derivation task-specific theoretical and empirical losses. We start be deriving the loss for each task. Consider the task k, its loss is the discrepancy between the $\mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta}$ and $\mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k})}$:

$$\begin{aligned}
\mathcal{L}^{(k)} &= \| \mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta} - \mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k})} \|_{\mathrm{HS}}^2 - \| \mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k})} \|_{\mathrm{HS}}^2 \\
&= \| \mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta} \|_{\mathrm{HS}}^2 - 2\langle \mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta}, \mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k})} \rangle_{\mathrm{HS}}.
\end{aligned}$$

Let $q_k(x,y) = \sum_{i,j} M_{ij}^{(k)} \phi_i^{(k)}(x)\psi_j^{(k)}(y)$, with $\mathbf{M}^{(k)} = \mathbf{A}^{(k)}\mathbf{B}^{(k)\top}$, be the kernel of $\mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta}$. Since $\mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta} : L^2_{\nu_k} \to L^2_{\mu_k}$ is an integral operator,

$$\begin{aligned}
\| \mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta} \|_{\mathrm{HS}}^2 &= \int q_k(x,y)^2\, d\mu_k(x) d\nu_k(y) \\
&= \mathbb{E}_{\mu_k \otimes \nu_k}[q_k(X,Y)^2] \\
&= \mathrm{Tr}\Big(\mathbf{G}_{\Psi_\theta}^{(k)} \mathbf{M}^{(k)\top} \mathbf{G}_{\Phi_\theta}^{(k)} \mathbf{M}^{(k)}\Big),
\end{aligned}$$

with

$$\mathbf{G}_{\Phi_\theta}^{(k)} \triangleq \mathbb{E}_{\mu_k}\left[\Phi_\theta^{(k)}(X)\Phi_\theta^{(k)}(X)^\top\right] \qquad \mathbf{G}_{\Psi_\theta}^{(k)} \triangleq \mathbb{E}_{\nu_k}\left[\Psi_\theta^{(k)}(Y)\Psi_\theta^{(k)}(Y)^\top\right].$$

Recall that

$$\mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} = \mathsf{E}^{(\mathrm{k})}_{\mathrm{Y|X}} - \mathbf{1}_{\mathcal{X}} \otimes \mathbf{1}_{\mathcal{Y}}, \qquad [\mathsf{E}^{(\mathrm{k})}_{\mathrm{Y|X}} f](x) = \mathbb{E}[f(Y) \mid X = x],$$

so that

$$\langle \mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}}, \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} \rangle_{\mathrm{HS}} = \langle \mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}}, \mathsf{E}^{(\mathrm{k})}_{\mathrm{Y|X}} \rangle_{\mathrm{HS}} - \langle \mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}}, \mathbf{1}_{\mathcal{X}} \otimes \mathbf{1}_{\mathcal{Y}} \rangle_{\mathrm{HS}}.$$

For $u \in L^2_{\mu_k}$ and $v \in L^2_{\nu_k}$,

$$\langle u \otimes v, \mathsf{E}^{(\mathrm{k})}_{\mathrm{Y|X}} \rangle_{\mathrm{HS}} = \langle u, \mathsf{E}^{(\mathrm{k})}_{\mathrm{Y|X}} v \rangle_{\mu_k},$$

and,

$$\begin{aligned}
\langle u, \mathsf{E}^{(\mathrm{k})}_{\mathrm{Y|X}} v \rangle_{\mu_k} &= \int u(x) \mathbb{E}[v(Y) \mid X = x] \, d\mu_k(x) \\
&= \mathbb{E}[u(X) \mathbb{E}[v(Y) \mid X]] \\
&= \mathbb{E}[\mathbb{E}[u(X) v(Y) \mid X]] \\
&= \mathbb{E}_{\rho_k}[u(X) v(Y)] \\
&= \int u(x) v(y) \, d\rho_k(x, y).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\langle \mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}}, \mathsf{E}^{(\mathrm{k})}_{\mathrm{Y|X}} \rangle_{\mathrm{HS}} &= \int q_k(x, y) \, d\rho_k(x, y) \\
&= \mathbb{E}_{\rho_k}\left[q_k(X, Y)\right]
\end{aligned}$$

Similarly,

$$\langle u \otimes v, \mathbf{1}_{\mathcal{X}} \otimes \mathbf{1}_{\mathcal{Y}} \rangle_{\mathrm{HS}} = \langle u, \mathbf{1}_{\mathcal{X}} \rangle_{\mu_k} \langle v, \mathbf{1}_{\mathcal{Y}} \rangle_{\nu_k},$$

which gives

$$\begin{aligned}
\langle \mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}}, \mathbf{1}_{\mathcal{X}} \otimes \mathbf{1}_{\mathcal{Y}} \rangle_{\mathrm{HS}} &= \int q_k(x, y) \, d\mu_k(x) d\nu_k(y) \\
&= \mathbb{E}_{\mu_k \otimes \nu_k}\left[q_k(X, Y)\right]
\end{aligned}$$

Hence

$$\begin{aligned}
\langle \mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}}, \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} \rangle_{\mathrm{HS}} &= \int q_k(x, y) \, d\rho_k(x, y) - \int q_k(x, y) \, d\mu_k(x) d\nu_k(y) \\
&= \mathbb{E}_{\rho_k}\left[q_k(X, Y)\right] - \mathbb{E}_{\mu_k \otimes \nu_k}\left[q_k(X, Y)\right].
\end{aligned}$$

Defining

$$\mathbf{C}^{(k)}_{\Phi_{\theta,c}\Psi_{\theta,c}} = \mathbb{E}_{\rho_k}\left[\Phi^{(k)}_{\theta,c}(X) \Psi^{(k)}_{\theta,c}(Y)^\top\right],$$

we obtain

$$\langle \mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}}, \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} \rangle_{\mathrm{HS}} = \mathrm{Tr}\Big(\mathbf{C}^{(k)\top}_{\Phi_\theta \Psi_\theta} \mathbf{M}^{(k)}\Big).$$

It leads to the theoretical population loss

$$\mathcal{L}^{(k)} = \mathrm{Tr}\Big(\mathbf{G}^{(k)}_{\Psi_\theta} \mathbf{M}^{(k)\top} \mathbf{G}^{(k)}_{\Phi_\theta} \mathbf{M}^{(k)}\Big) - 2\,\mathrm{Tr}\Big(\mathbf{C}^{(k)\top}_{\Phi_\theta \Psi_\theta} \mathbf{M}^{(k)}\Big).$$

From an empirical perspective, given a set of observations $\{(x^{(k)}_i, y^{(k)}_i)\}_{i \in [n_k]}$, we denote $\Phi_\theta(\mathbf{X}^{(k)}) = \{\Phi_\theta(x^{(k)}_i)\}_{i \in [n_k]}$ and $\Psi_\theta(\mathbf{Y}^{(k)}) = \{\Psi_\theta(y^{(k)}_i)\}_{i \in [n_k]}$ the feature matrices. Considering, the matrices

$$\mathbf{Q}^{(k)} = \Phi_\theta(\mathbf{X}^{(k)}) \mathbf{A}^{(k)} \mathbf{B}^{(k)\top} \Psi_\theta(\mathbf{Y}^{(k)})^\top \quad \text{and} \quad \mathbf{H}^{(k)} = \mathbf{I}_{n_k} - 1/n_k \mathbf{1}\mathbf{1}^\top, \tag{41}$$

the data-fitting loss $\mathcal{L}^{(k)}$ admits an unbiased empirical estimator defined by

$$\widehat{\mathcal{L}^{(k)}} \triangleq \frac{\|\mathbf{Q}^{(k)}\|^2_{\mathrm{F}} - \|\,\mathrm{Diag}(\mathbf{Q}^{(k)})\|^2_2}{n_k(n_k - 1)} - \frac{2}{n_k - 1}\,\mathrm{Tr}(\mathbf{H}^{(k)} \mathbf{Q}^{(k)}). \tag{42}$$

Here $(\|\mathbf{Q}^{(k)}\|^2_{\mathrm{F}} - \|\,\mathrm{Diag}(\mathbf{Q}^{(k)})\|^2_2)/(n_k(n_k - 1))$ is the unbiased U-statistic to estimate $\|\,\mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}}\,\|^2_{\mathrm{HS}}$.

Algorithm 2 MTL-CMO learning procedure

Require: Task datasets $\{\mathcal{D}_k\}_{k=1}^K$, system batch size B, within-task sample size L, ranks $\{r_k\}_{k=1}^K$, regularization λ

Initialize shared dictionary parameters θ and $\mathbf{A}^{(k)}, \mathbf{B}^{(k)} \in \mathbb{R}^{d \times r_k}$, $k = 1, \dots, K$
for each optimization step **do**
  Sample $\mathcal{B} \subset \{1, \dots, K\}$ with $|\mathcal{B}| = B$
  for each $k \in \mathcal{B}$ **do**
    Sample L paired observations $\{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^L$ from $\mathcal{D}_k$
    Form $\mathbf{X}^{(k)} \leftarrow (x_i^{(k)})_{i=1}^L$ and $\mathbf{Y}^{(k)} \leftarrow (y_i^{(k)})_{i=1}^L$
    Evaluate $\mathbf{W}^{(k)} \leftarrow \Phi_\theta(\mathbf{X}^{(k)})$ and $\mathbf{Z}^{(k)} \leftarrow \Psi_\theta(\mathbf{Y}^{(k)})$
    Set $\mathbf{H}_L \leftarrow \mathbf{I}_L - \frac{1}{L}\mathbf{1}\mathbf{1}^\top$
    From $\mathbf{Q}^{(k)} \leftarrow \mathbf{W}^{(k)}\mathbf{A}^{(k)}\mathbf{B}^{(k)\top}\mathbf{Z}^{(k)\top}$
    Compute $\widehat{\mathcal{L}}^{(k)} \leftarrow \dfrac{\|\mathbf{Q}^{(k)}\|_{\mathrm{F}}^2 - \|\operatorname{Diag}(\mathbf{Q}^{(k)})\|_2^2}{L(L-1)} - \dfrac{2}{L-1}\operatorname{Tr}(\mathbf{H}_L\mathbf{Q}^{(k)})$
  end for
  Aggregate $\widehat{\mathcal{J}}_{\mathcal{B}} \leftarrow \dfrac{1}{B}\displaystyle\sum_{k\in\mathcal{B}}\left[\widehat{\mathcal{L}}^{(k)} + \lambda\left(\|\mathbf{A}^{(k)}\|_{\mathrm{F}}^2 + \|\mathbf{B}^{(k)}\|_{\mathrm{F}}^2\right)\right]$
  Back-propagate $\widehat{\mathcal{J}}_{\mathcal{B}}$ and update θ and $\{\mathbf{A}^{(k)}, \mathbf{B}^{(k)}\}_{k\in\mathcal{B}}$
end for
return Φ_θ, Ψ_θ and $\{\mathbf{M}^{(k)} = \mathbf{A}^{(k)}\mathbf{B}^{(k)\top}\}_{k=1}^K$

Multi-task optimization. The factorization $\mathbf{M}^{(k)} = \mathbf{A}^{(k)}\mathbf{B}^{(k)\top}$ is scale non-identifiable, since for every $c \neq 0$, $\mathbf{A}^{(k)}\mathbf{B}^{(k)\top} = (c\mathbf{A}^{(k)})(c^{-1}\mathbf{B}^{(k)})^\top$. We therefore regularize both factors and optimize

$$\min_{\theta, \{\mathbf{A}^{(k)}, \mathbf{B}^{(k)}\}_{k=1}^K} \sum_{k=1}^K \left[\widehat{\mathcal{L}^{(k)}} + \lambda\left(\|\mathbf{A}^{(k)}\|_{\mathrm{F}}^2 + \|\mathbf{B}^{(k)}\|_{\mathrm{F}}^2\right)\right].$$

In practice, this objective is optimized stochastically at two levels: each update samples a mini-batch of systems, and then L paired observations within each selected system.

The matrix $\mathbf{Q}^{(k)}$ contains all input–output pairings within task k. Its diagonal entries correspond to matched pairs from ρ_k, whereas its off-diagonal entries cross distinct observations and provide the empirical product-of-marginals term $\mu_k \otimes \nu_k$. For dynamical systems, the diagonal retains the time-lagged transition pairing while the off-diagonal terms break it.

Closed-form transfer. Once the shared dictionaries are fixed, transfer only requires optimizing with respect to $\mathbf{M}$:

$$\mathcal{L}(\mathbf{M}) = \operatorname{Tr}\big(\mathbf{G}_\Psi\mathbf{M}^\top\mathbf{G}_\Phi\mathbf{M}\big) - 2\operatorname{Tr}\big(\mathbf{C}^\top\mathbf{M}\big).$$

Assume that $\mathbf{G}_\Phi \succ 0$ and $\mathbf{G}_\Psi \succ 0$. Since

$$\operatorname{Tr}\big(\mathbf{G}_\Psi\mathbf{M}^\top\mathbf{G}_\Phi\mathbf{M}\big) = \left\|\mathbf{G}_\Phi^{1/2}\mathbf{M}\mathbf{G}_\Psi^{1/2}\right\|_{\mathrm{F}}^2,$$

we have

$$\mathcal{L}(\mathbf{M}) \geq \lambda_{\min}(\mathbf{G}_\Phi)\lambda_{\min}(\mathbf{G}_\Psi)\|\mathbf{M}\|_{\mathrm{F}}^2 - 2\|\mathbf{C}\|_{\mathrm{F}}\|\mathbf{M}\|_{\mathrm{F}}.$$

Thus $\mathcal{L}(\mathbf{M}) \to +\infty$ as $\|\mathbf{M}\|_{\mathrm{F}} \to \infty$. The objective is continuous and coercive, hence admits a global minimizer. Moreover, with $\mathbf{m} = \operatorname{vec}(\mathbf{M})$,

$$\nabla_{\mathbf{m}}^2\mathcal{L} = 2\left(\mathbf{G}_\Psi \otimes \mathbf{G}_\Phi\right) \succ 0,$$

so the minimizer is unique. Finally,

$$\nabla_{\mathbf{M}}\mathcal{L} = 2\mathbf{G}_\Phi\mathbf{M}\mathbf{G}_\Psi - 2\mathbf{C},$$

and the first-order condition gives

$$\mathbf{G}_\Phi\mathbf{M}^\star\mathbf{G}_\Psi = \mathbf{C}, \qquad \boxed{\mathbf{M}^\star = \mathbf{G}_\Phi^{-1}\mathbf{C}\mathbf{G}_\Psi^{-1}.}$$

For a new task, T-CMO therefore estimates only the target-specific operator; the shared dictionaries are not retrained.

Algorithm 3 T-CMO transfer procedure

Require: Target dataset $\mathcal{D}^\star = \{(x_i^\star, y_i^\star)\}_{i=1}^n$, pretrained dictionaries Φ_θ, Ψ_θ, target rank r, Gram regularization ε

Form $\mathbf{X}^\star = (x_i^\star)_{i=1}^n$ and $\mathbf{Y}^\star = (y_i^\star)_{i=1}^n$

Evaluate $\mathbf{W}^\star \leftarrow \Phi_\theta(\mathbf{X}^\star) \in \mathbb{R}^{n\times d}$ and $\mathbf{Z}^\star \leftarrow \Psi_\theta(\mathbf{Y}^\star) \in \mathbb{R}^{n\times d}$

Set $\mathbf{H}_n \leftarrow \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$

Center $\mathbf{W}_c^\star \leftarrow \mathbf{H}_n\mathbf{W}^\star$ and $\mathbf{Z}_c^\star \leftarrow \mathbf{H}_n\mathbf{Z}^\star$

Compute $\widehat{\mathbf{G}}_\Phi^\star \leftarrow \frac{1}{n}\mathbf{W}_c^{\star\top}\mathbf{W}_c^\star$, $\widehat{\mathbf{G}}_\Psi^\star \leftarrow \frac{1}{n}\mathbf{Z}_c^{\star\top}\mathbf{Z}_c^\star$

Compute $\widehat{\mathbf{C}}_{\Phi\Psi}^\star \leftarrow \frac{1}{n}\mathbf{W}_c^{\star\top}\mathbf{Z}_c^\star$

Regularize $\widehat{\mathbf{G}}_{\Phi,\varepsilon}^\star \leftarrow \widehat{\mathbf{G}}_\Phi^\star + \varepsilon\mathbf{I}_d$ and $\widehat{\mathbf{G}}_{\Psi,\varepsilon}^\star \leftarrow \widehat{\mathbf{G}}_\Psi^\star + \varepsilon\mathbf{I}_d$

Recover $\widehat{\mathbf{M}}^\star \leftarrow \widehat{\mathbf{G}}_{\Phi,\varepsilon}^{\star-1}\widehat{\mathbf{C}}_{\Phi\Psi}^\star\widehat{\mathbf{G}}_{\Psi,\varepsilon}^{\star-1}$

Whiten $\widehat{\mathbf{M}}_{\mathbf{W}}^\star \leftarrow \widehat{\mathbf{G}}_{\Phi,\varepsilon}^{\star 1/2}\widehat{\mathbf{M}}^\star\widehat{\mathbf{G}}_{\Psi,\varepsilon}^{\star 1/2}$

Compute the rank-r SVD $\widehat{\mathbf{M}}_{\mathbf{W}}^\star = \mathbf{U}_r^\star \operatorname{Diag}(\boldsymbol{\sigma}^\star)\mathbf{V}_r^{\star\top}$

Recover the singular functions from $\widehat{\mathbf{G}}_{\Phi,\varepsilon}^{\star-1/2}\mathbf{U}_r^\star$ and $\widehat{\mathbf{G}}_{\Psi,\varepsilon}^{\star-1/2}\mathbf{V}_r^\star$

return $\widehat{\mathbf{M}}_{\mathbf{W}}^\star$ and its rank-r singular representation

C Detailed related work

The proposed approach, MTL-CMO, jointly estimates conditional mean operators (CMOs) from data associated with related joint distributions (see Section 3). Because CMOs describe conditional distributions through their action on observables, relevant work spans both general-purpose operator estimation and application-specific methods. We discuss four connections: (i) multi-task representation learning, (ii) single-task CMO estimation, (iii) uncertainty quantification, and (iv) operator-based learning for dynamical systems.

Positioning within multi-task representation learning. MTL-CMO can be considered as a two-sided extension of the one-sided shared representation paradigm of multi-task learning Caruana (1997). A standard formulation of this framework stipulates that each task has a predictor

$$f_k(x) = w_k^\top h_\theta(x),$$

where h_θ is a shared feature map and w_k the task-specific head. The one-sided problem has been studied extensively (Evgeniou et al., 2005; Ando et al., 2005; Argyriou et al., 2008) and typically involves specific regularization (Evgeniou & Pontil, 2004; Zhang & Schneider, 2010; Maurer et al., 2013). Provable statistical guarantees on the benefit of sharing feature maps between tasks have been established (Maurer et al., 2016; Tripuraneni et al., 2021).

MTL-CMO differs in three respects. First, each task-specific object is an operator between function spaces rather than a scalar- or vector-valued predictor. Second, it is two-sided; the method learns shared function spaces on both the input and the output sides of these operators. Third, although the functions themselves are shared across tasks, their associated synthesis maps are defined in the task-specific spaces $\mathrm{L}^2_{\mu_k}(\mathcal{X})$ and $\mathrm{L}^2_{\nu_k}(\mathcal{Y})$, whose inner products depend on the marginals μ_k and ν_k.

Conditional mean operators and their estimation. Conditional mean operators have long been studied; for instance, early work addressed the problem of estimating multiple regression and correlation (Breiman & Friedman, 1985). A prominent body of work, known as conditional mean embedding (CME), focus on the estimation of the operator whose action is restricted to predefined Reproducible Kernel Hilbert Space (RKHS) (Song et al., 2009; Muandet et al., 2017). with classical estimators expressed as regularized kernel ridge regression, whose consistency and statistical rates have been subsequently studied (Grünewälder et al., 2012; Mollenhauer & Koltai, 2020; Hertel et al., 2026). More recently, a hybrid CME estimator combining neural networks to model the output space with a predefined RKHS to model the input space was proposed in (Shimizu et al., 2024). CME approaches have also been extensively used to study dynamical systems, although the choice of function space can affect the recovered spectral information (Kostic et al., 2023). However, CMEs are less well-suited to uncertainty quantification, since evaluating conditional expectations is only possible for observables in the RKHS. To alleviate this issue, Neural Conditional Probability (NCP) instead learns a truncated singular representation of a compact CMO, allowing its action to be eval-

uated on a broader class of observables (Kostic et al., 2024). MTL-CMO builds on this operator perspective while learning functional representations jointly across related tasks.

Multi-task learning for uncertainty quantification. The core idea of multi-task approaches for uncertainty quantification is to jointly learn conditional density functions or conditional cumulative distribution functions. A baseline approach to share information across conditional distributions is to combine a common neural representation with task-specific mixture density networks (Caruana, 1997; Bishop, 1994). Each task then predicts the parameters of its own conditional mixture model. Another approach, DeepJMQR, jointly learns conditional expectations and multiple conditional quantiles to estimate a discrete approximation of one-dimensional conditional transport maps (Rodrigues & Pereira, 2020).
These methods directly parameterize a density or selected statistics. MTL-CMO instead jointly learns operators from which different conditional statistics can be approximated by projecting the corresponding observables, see Equation (3).

Operator-based multi-task learning for dynamical systems. For a stochastic dynamical system, the conditional mean operator that maps an observable of a future state to its conditional expectation given the current state is a Koopman operator at the corresponding time lag. Several multi-task approaches have been proposed to jointly learn dynamics from observed trajectories and by leveraging Koopman operators (Iwata & Kawahara, 2021; Elul et al., 2024; Zhang, 2025; Han et al., 2026). These methods suppose a shared embedding of the trajectories and impose a task-specific linear evolution of the dynamics in the latent space. Some additionally learn a common (Elul et al., 2024; Zhang, 2025) or task-specific (Han et al., 2026) decoder from latent coordinates to the state spaces. In particular, Han et al. (2026) has been proposed for predictive control and therefore includes the control command in the modeling.
Unlike MTL-CMO, these methods are primarily trained to reconstruct or predict trajectories, rather than to minimize a Hilbert-Schmidt discrepancy between operators. Their objectives may therefore favor accurate state predictions without ensuring accurate estimates of the operator's action on other observables. MTL-CMO instead minimizes an empirical objective corresponding to the Hilbert-Schmidt error, which directly estimates the conditional expectation of future states for any observables.

Multi-task conditional mean operator estimation. Standard multi-task representation learning shares an input feature map across task-specific predictors (Caruana, 1997; Maurer et al., 2016). MTL-CMO extends this principle to operators by learning input and output function spaces, whose corresponding synthesis maps Equation (12) act in spaces defined by each task's marginals. This differs from conditional mean embedding, which typically uses a prescribed reproducing kernel Hilbert space (Song et al., 2009; Muandet et al., 2017), and from NCP, which estimates a single CMO (Kostic et al., 2024).

D Uncertainty Quantification Experiments

D.1 Conditional-distribution families

We consider four families of related conditional distributions. For each family, a task k corresponds to a conditional law $P_k(\mathrm{y} \mid \mathbf{x})$, with $\mathbf{X} \sim \mathcal{U}([-2,2]^D)$. Task dependence enters through the parameters of the conditional law.

(CD1) Univariate bimodal Gaussian mixture. For $D = 1$,

$$Y^{(k)} \mid X = x \sim p_k \mathcal{N}\big(a_k \sin x, \sigma_k^2\big) + (1 - p_k)\mathcal{N}\big(-a_k \sin x, \sigma_k^2\big)\,,$$

where

$$p_k \sim \mathcal{U}([0.2, 0.8]), \qquad a_k \sim \mathcal{U}([0.6, 1.0]), \qquad \sigma_k \sim \mathcal{U}([0.5, 0.8]).$$

Equivalently, the mixture component can be represented through a Bernoulli variable $B^{(k)} \sim \mathcal{B}(p_k)$ and $Z^{(k)} = 2B^{(k)} - 1 \in \{-1, 1\}$.

(CD2) Multivariate bimodal Gaussian mixture. For $D = 10$,

$$Y^{(k)} \mid \mathbf{X} = \mathbf{x} \sim p_k \mathcal{N}\big(a_k \sin(\mathbf{w}^\top \mathbf{x}), \sigma_k^2\big) + (1 - p_k)\mathcal{N}\big(-a_k \sin(\mathbf{w}^\top \mathbf{x}), \sigma_k^2\big)\,,$$

with the same distributions for (p_k, a_k, σ_k) as in **CD1**, $\mathbf{w} \in \mathbb{S}^9$, and $\mathbf{w}$ shared across tasks.

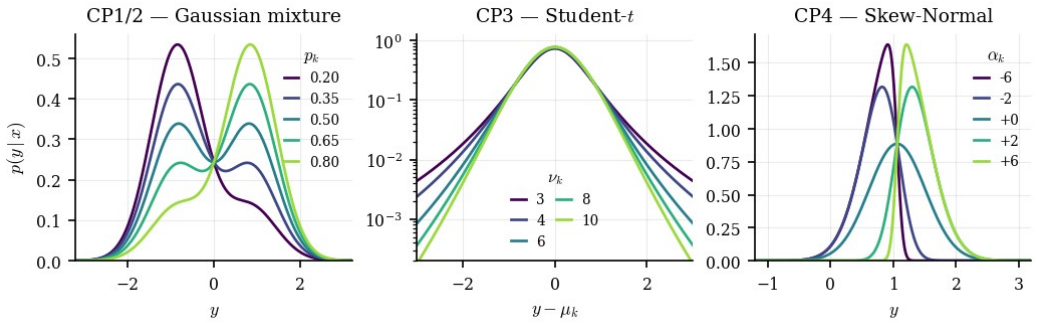

Figure 7: Representative conditional densities from **CD1**, **CD3**, and **CD4** for different conditioning inputs and task parameters.

(CD3) Multivariate Student distribution. For $D = 10$, let

$$\mu_k(\mathbf{x}) = a_k \sin(\mathbf{w}^\top \mathbf{x}) + b_k \cos(\mathbf{w}^\top \mathbf{x}).$$

We define

$$Y^{(k)} \mid \mathbf{X} = \mathbf{x} \sim t_{\nu_k}(\mu_k(\mathbf{x}), \sigma_k)\,,$$

whose conditional density is

$$p_k(y \mid \mathbf{x}) = \frac{\Gamma((\nu_k+1)/2)}{\Gamma(\nu_k/2)\sqrt{\nu_k \pi}\,\sigma_k}\left[1 + \frac{1}{\nu_k}\left(\frac{y-\mu_k(\mathbf{x})}{\sigma_k}\right)^2\right]^{-(\nu_k+1)/2}.$$

The task parameters are

$$a_k \sim \mathcal{N}(1, 0.3^2), \qquad b_k \sim \mathcal{N}(0.3, 0.2^2), \qquad \sigma_k \sim \mathcal{N}(0.5, 0.1^2), \qquad \nu_k \sim \mathcal{U}([3, 10]),$$

The direction $\mathbf{w} \in \mathbb{S}^9$ is shared across tasks.

(CD4) Multivariate skew-normal distribution. We use the same form

$$\mu_k(\mathbf{x}) = a_k \sin(\mathbf{w}^\top \mathbf{x}) + b_k \cos(\mathbf{w}^\top \mathbf{x}),$$

and define

$$Y^{(k)} \mid \mathbf{X} = \mathbf{x} \sim \mathrm{SN}(\mu_k(\mathbf{x}), \sigma_k, \alpha_k)\,.$$

Writing ϕ and Φ for the standard Gaussian density and CDF, respectively, its conditional density is

$$p_k(y \mid \mathbf{x}) = \frac{2}{\sigma_k}\phi\left(\frac{y-\mu_k(\mathbf{x})}{\sigma_k}\right)\Phi\left(\alpha_k \frac{y-\mu_k(\mathbf{x})}{\sigma_k}\right).$$

We sample

$$a_k \sim \mathcal{N}(1, 0.15^2), \qquad b_k \sim \mathcal{N}(0.5, 0.15^2), \qquad \sigma_k \sim \mathcal{N}(0.45, 0.08^2), \qquad \alpha_k \sim \mathcal{N}(1.5, 0.4^2),$$

with σ_k truncated below at 0.1 and $\mathbf{w} \in \mathbb{S}^9$ shared across tasks.

Figure 7 displays representative conditional densities from **CD1**, **CD3**, and **CD4**. We omit **CD2**, which differs from **CD1** only through the higher-dimensional input and the projection $\mathbf{w}^\top \mathbf{x}$.

Experimental protocol. For each family, we generate $K = 100$ tasks with $n = 400$ observations per task. Task parameters are sampled once and kept fixed across repetitions, while the observations are independently resampled. Results are averaged over 10 independent data realizations.

For transfer, we generate 100 additional unseen tasks by independently sampling new task parameters. For each target task, we vary the number of available observations . Transfer results are averaged over 10 independent realizations of the target data.

D.2 MULTI-TASK CMO ARCHITECTURE AND TRAINING

For all four conditional-distribution families, we use the multi-task CMO parametrization introduced . The two shared functional spaces are represented by the neural dictionaries

$$\Phi_\theta(x) = \{\phi_1^\theta(x), \dots, \phi_d^\theta(x)\}, \qquad \Psi_\theta(y) = \{\psi_1^\theta(y), \dots, \psi_d^\theta(y)\}.$$

For each task k, the corresponding deflated conditional operator is parametrized as

$$\mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta} = \mathsf{S}_{\Phi_\theta}^{(k)} \mathbf{A}^{(k)} \mathbf{B}^{(k)\top} \mathsf{S}_{\Psi_\theta}^{(k)*}, \qquad \mathbf{A}^{(k)}, \mathbf{B}^{(k)} \in \mathbb{R}^{d \times r_k}.$$

Hence, $\mathbf{A}^{(k)}\mathbf{B}^{(k)\top}$ defines a rank-r_k factorization of the task-specific operator in the shared functional coordinates. The neural dictionaries are common to all tasks, while the low-rank factors $\mathbf{A}^{(k)}$ and $\mathbf{B}^{(k)}$ capture the task-specific dependence structure.

Neural parametrization. The two dictionaries are parametrized by separate MLPs, both shared across tasks. The input dictionary takes $x \in \mathbb{R}^D$, whereas the output dictionary takes the scalar response $y \in \mathbb{R}$. Each hidden block is composed of a linear layer followed by the activation function and dropout; The dimension d of the learned functional space is given by the width of the last hidden layer. No decoder is used: conditional statistics are obtained directly from the learned CMO. The architecture is adapted to each conditional-distribution family. Table 3 reports the configurations used in the experiments.

Table 3: Neural architecture and operator rank used for the four conditional-distribution families.

	Hidden layers	d	r_k	Activation	Dropout
CD1	3×64	64	20	GELU	0.15
CD2	2×64	64	7	Tanh	0.2
CD3	4×64	64	18	GELU	0.14
CD4	4×64	64	13	Tanh	0.16

Optimization. For each task, the response variable Y is standardized using its empirical mean and standard deviation, while the input variables are left unchanged. The shared dictionaries and the task-specific low-rank factors are optimized jointly with AdamW. We use separate learning rates and weight-decay coefficients for the shared neural parameters and for $\{\mathbf{A}^{(k)}, \mathbf{B}^{(k)}\}_{k=1}^K$.

Table 4: Optimization parameters for the uncertainty-quantification experiments.

	Epochs	Batch size	$\mathrm{lr}_{\mathrm{shared}}$	$\mathrm{lr}_{\mathrm{specific}}$	$\mathrm{wd}_{\mathrm{shared}}$	$\mathrm{wd}_{\mathrm{specific}}$	Grad. clip	Scheduler
CD1	5400	32	5.8×10^{-5}	1.4×10^{-4}	2.2×10^{-3}	2.6×10^{-2}	5.0	None
CD2	4200	64	2.7×10^{-4}	6.4×10^{-5}	3.9×10^{-3}	2.6×10^{-4}	2.0	None
CD3	3600	64	1.2×10^{-4}	1.8×10^{-3}	2.2×10^{-4}	3.4×10^{-6}	1.0	None
CD4	2600	64	3.0×10^{-3}	2.9×10^{-3}	9.1×10^{-6}	4.4×10^{-3}	2.0	Cosine

D.3 CONDITIONAL CDF RECONSTRUCTION

We now detail how a conditional CDF is recovered from the learned CMO. Consider a task k with observations $\mathcal{D}_k = \{(x_j^{(k)}, y_j^{(k)})\}_{j=1}^{n_k}$.

Centering and task-specific SVD. The outputs of the two neural dictionaries are centered independently for each task. We define

$$\overline{\Phi}^{(k)} = \frac{1}{n_k} \sum_{j=1}^{n_k} \Phi_\theta(x_j^{(k)}), \qquad \overline{\Psi}^{(k)} = \frac{1}{n_k} \sum_{j=1}^{n_k} \Psi_\theta(y_j^{(k)}),$$

and

$$\Phi_{\theta,c}^{(k)}(x) = \Phi_\theta(x) - \overline{\Phi}^{(k)}, \qquad \Psi_{\theta,c}^{(k)}(y) = \Psi_\theta(y) - \overline{\Psi}^{(k)}.$$

The corresponding empirical Gram matrices are

$$\widehat{\mathbf{G}}_{\Phi_\theta}^{(k)} = \frac{1}{n_k}\sum_{j=1}^{n_k} \Phi_{\theta,c}^{(k)}(x_j^{(k)})^\top \Phi_{\theta,c}^{(k)}(x_j^{(k)}), \qquad \widehat{\mathbf{G}}_{\Psi_\theta}^{(k)} = \frac{1}{n_k}\sum_{j=1}^{n_k} \Psi_{\theta,c}^{(k)}(y_j^{(k)})^\top \Psi_{\theta,c}^{(k)}(y_j^{(k)}).$$

During multi-task learning, task k is represented in the shared functional coordinates by the low-rank matrix $\mathbf{A}^{(k)}\mathbf{B}^{(k)\top}$. Its orthonormal singular representation is recovered by whitening with the task-specific Gram matrices and computing

$$\widehat{\mathbf{G}}_{\Phi_\theta}^{(k)1/2}\mathbf{A}^{(k)}\mathbf{B}^{(k)\top}\widehat{\mathbf{G}}_{\Psi_\theta}^{(k)1/2} = \mathbf{U}^{(k)}\,\mathrm{Diag}(\boldsymbol{\sigma}^{(k)})\mathbf{V}^{(k)\top}.$$

Denoting by $\mathbf{u}_i^{(k)}$ and $\mathbf{v}_i^{(k)}$ the ith columns of $\mathbf{U}^{(k)}$ and $\mathbf{V}^{(k)}$, the task-specific singular functions are

$$\widetilde{u}_i^{(k)}(x) = \Phi_{\theta,c}^{(k)}(x)\widehat{\mathbf{G}}_{\Phi_\theta}^{(k)-1/2}\mathbf{u}_i^{(k)}, \qquad \widetilde{v}_i^{(k)}(y) = \Psi_{\theta,c}^{(k)}(y)\widehat{\mathbf{G}}_{\Psi_\theta}^{(k)-1/2}\mathbf{v}_i^{(k)}.$$

The resulting rank-r_k deflated operator is

$$\mathsf{D}_{\mathrm{Y|X}}^{(\mathrm{k}),\theta} = \sum_{i=1}^{r_k} \sigma_i^{(k)}\widetilde{u}_i^{(k)} \otimes \widetilde{v}_i^{(k)}.$$

Conditional CDF. For scalar Y, we recover the conditional CDF by taking $f_t(y) = \mathbf{1}_{\{y\le t\}}$, so that

$$F_k(t \mid x) = \mathbb{E}\Big[f_t(Y^{(k)}) \mid X^{(k)} = x\Big].$$

We approximate expectations with respect to the output marginal ν_k using the empirical measure

$$\widehat{\nu}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} \delta_{y_j^{(k)}}.$$

Applying the recovered CMO to f_t gives

$$\widehat{F}_k(t \mid x) = \frac{1}{n_k}\sum_{j=1}^{n_k} \mathbf{1}_{\{y_j^{(k)}\le t\}} + \frac{1}{n_k}\sum_{j:y_j^{(k)}\le t}\sum_{i=1}^{r_k} \sigma_i^{(k)}\widetilde{u}_i^{(k)}(x)\widetilde{v}_i^{(k)}(y_j^{(k)}).$$

The first term is the empirical marginal CDF of $Y^{(k)}$, while the second is the correction induced by the deflated conditional operator and introduces the dependence on x.

For computation, we sort the outputs $y_{(1)}^{(k)} \le \cdots \le y_{(n_k)}^{(k)}$ and define

$$c_j^{(k)}(x) = \sum_{i=1}^{r_k} \sigma_i^{(k)}\widetilde{u}_i^{(k)}(x)\widetilde{v}_i^{(k)}(y_{(j)}^{(k)}).$$

At the mth ordered output,

$$\widehat{F}_k(y_{(m)}^{(k)} \mid x) = \frac{m}{n_k} + \frac{1}{n_k}\sum_{j=1}^{m} c_j^{(k)}(x).$$

The full conditional CDF is therefore obtained by a cumulative sum over the ordered outputs. For an arbitrary threshold t, all observations satisfying $y_j^{(k)} \le t$ are retained.

In practice, small negative empirical masses may arise from numerical estimation. We clip these masses to zero and renormalize them before taking the cumulative sum, ensuring that the resulting estimate is a valid conditional CDF.

Wasserstein evaluation. For each task k, we compare the estimated conditional CDF $\widehat{F}_k(\cdot \mid x)$ with the analytical ground-truth CDF of the corresponding data-generating process using the one-dimensional Wasserstein-1 distance. Both CDFs are evaluated on a uniform grid of 1000 output values covering the relevant support of Y. For each task, the distance is computed at 40 conditioning values uniformly distributed over the input range and averaged over these conditioning points. The resulting scores are then averaged across tasks. All experiments are repeated over 10 independent random seeds.

Table 5: Transfer to unseen conditional distributions. Wasserstein-1 distance to the ground-truth conditional distribution (mean $\pm$ standard deviation over 10 target-data seeds), averaged over 100 unseen tasks. Lower is better. The last row of each block reports the ratio between single-task NCP and T-CMO.

n_{tr}	50	100	150	200	250	300	350	400
CD1								
NCP	**0.2362 ± 0.0059**	**0.1739 ± 0.0025**	**0.1455 ± 0.0038**	**0.1258 ± 0.0021**	**0.1152 ± 0.0015**	**0.1062 ± 0.0021**	0.1004 ± 0.0018	0.0948 ± 0.0024
T-CMO	0.2629 ± 0.0053	0.1869 ± 0.0015	0.1546 ± 0.0042	0.1315 ± 0.0012	0.1188 ± 0.0020	0.1072 ± 0.0022	0.1004 ± 0.0015	**0.0943 ± 0.0015**
NCP / T-CMO	0.90	0.93	0.94	0.96	0.97	0.99	1.00	1.01
CD2								
NCP	**0.2326 ± 0.0082**	**0.1826 ± 0.0041**	**0.1704 ± 0.0016**	0.1588 ± 0.0015	0.1538 ± 0.0024	0.1500 ± 0.0007	0.1470 ± 0.0024	0.1458 ± 0.0028
T-CMO	0.2913 ± 0.0052	0.2089 ± 0.0022	0.1747 ± 0.0023	**0.1547 ± 0.0016**	**0.1391 ± 0.0032**	**0.1244 ± 0.0021**	**0.1186 ± 0.0017**	**0.1125 ± 0.0007**
NCP / T-CMO	0.80	0.87	0.98	1.03	1.11	1.21	1.24	1.30
CD3								
NCP	**0.2603 ± 0.0062**	0.2176 ± 0.0038	0.1999 ± 0.0011	0.1934 ± 0.0016	0.1889 ± 0.0012	0.1860 ± 0.0012	0.1830 ± 0.0005	0.1815 ± 0.0017
T-CMO	0.2728 ± 0.0074	**0.1990 ± 0.0039**	**0.1624 ± 0.0034**	**0.1447 ± 0.0009**	**0.1317 ± 0.0029**	**0.1198 ± 0.0018**	**0.1128 ± 0.0009**	**0.1067 ± 0.0015**
NCP / T-CMO	0.95	1.09	1.23	1.34	1.43	1.55	1.62	1.70
CD4								
NCP	0.2584 ± 0.0038	0.2225 ± 0.0028	0.2049 ± 0.0012	0.1909 ± 0.0020	0.1797 ± 0.0033	0.1681 ± 0.0027	0.1587 ± 0.0026	0.1498 ± 0.0017
T-CMO	**0.1683 ± 0.0019**	**0.1240 ± 0.0019**	**0.1052 ± 0.0014**	**0.0923 ± 0.0012**	**0.0849 ± 0.0009**	**0.0794 ± 0.0008**	**0.0742 ± 0.0001**	**0.0719 ± 0.0005**
NCP / T-CMO	1.54	1.79	1.95	2.07	2.12	2.12	2.14	2.08

D.4 Transfer to unseen conditional distributions

We next evaluate whether the functional spaces learned jointly across source tasks can be reused to estimate conditional distributions that were not observed during multi-task training. For each conditional-distribution family, we generate a new collection of target tasks independently of the source tasks used to learn the shared dictionaries. The parameters of the target distributions are drawn from the same task-generating distribution as during multi-task training.

Transfer protocol. Given the learned dictionaries $\Phi_{\hat{\theta}}$ and $\Psi_{\hat{\theta}}$, all their parameters are frozen. For a new target task with

$$\mathcal{D}_{\mathrm{tr}} = \{(x_j, y_j)\}_{j=1}^{n_{\mathrm{tr}}},$$

we estimate only the operator associated with this task. The dictionary outputs are centered using the target observations, and we compute the empirical Gram and cross-covariance matrices

$$\widehat{\mathbf{G}}_{\Phi} = \frac{1}{n_{\mathrm{tr}}} \sum_{j=1}^{n_{\mathrm{tr}}} \Phi_c(x_j)^{\top} \Phi_c(x_j), \qquad \widehat{\mathbf{G}}_{\Psi} = \frac{1}{n_{\mathrm{tr}}} \sum_{j=1}^{n_{\mathrm{tr}}} \Psi_c(y_j)^{\top} \Psi_c(y_j),$$

and

$$\widehat{\mathbf{C}} = \frac{1}{n_{\mathrm{tr}}} \sum_{j=1}^{n_{\mathrm{tr}}} \Phi_c(x_j)^{\top} \Psi_c(y_j).$$

The target operator is then recovered in closed form as

$$\mathbf{M}^{\star} = \widehat{\mathbf{G}}_{\Phi}^{-1} \widehat{\mathbf{C}} \widehat{\mathbf{G}}_{\Psi}^{-1}.$$

We subsequently whiten $\mathbf{M}^{\star}$ with the target Gram matrices and compute its truncated SVD, using the same rank as in the corresponding source model. The resulting singular functions are used to reconstruct the target conditional CDF as described in Section D.3. No gradient-based optimization or fine-tuning of the shared dictionaries is performed at transfer time.

Transfer performance. Table 5 reports the Wasserstein-1 error obtained on 100 unseen target tasks as the number of target observations increases. Results are averaged over 10 independent target-data seeds.

The benefit of transferring the shared functional spaces depends strongly on the conditional-distribution family. For CP1, whose conditional structure is comparatively simple, T-CMO remains close to single-task NCP and the two methods reach essentially the same accuracy for large target sample sizes. The advantage of transfer becomes more pronounced for the higher-dimensional and more structured families. On CP2, T-CMO improves over single-task NCP from $n_{\mathrm{tr}} = 200$ onward, while on CP3 the improvement appears already at $n_{\mathrm{tr}} = 100$ and increases steadily with the amount of target data. The strongest effect is observed for CP4, where T-CMO consistently outperforms single-task NCP and achieves approximately a two-fold reduction in Wasserstein error for moderate and large target sample sizes.

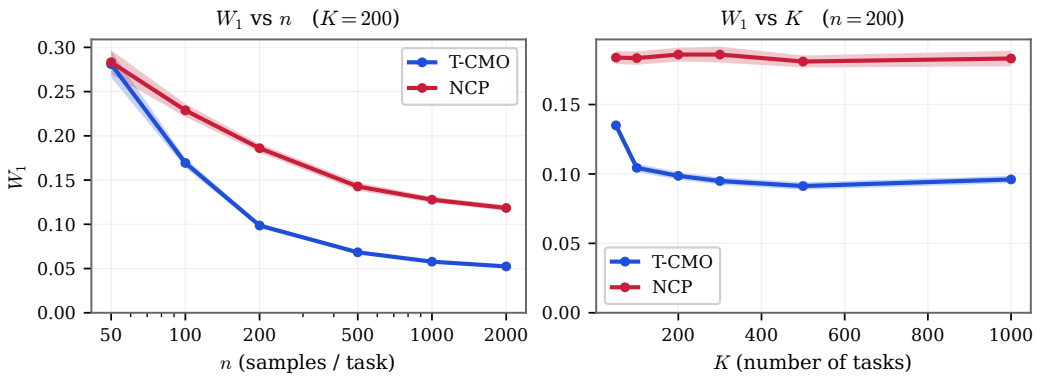

Figure 8: Sensitivity of transfer performance on **CD4**. **Left:** Wasserstein-1 error as a function of the number of observations per source task n, with $K = 200$. **Right:** Wasserstein-1 error as a function of the number of source tasks K, with $n = 200$.

Table 6: Sensitivity of transfer performance on **CD4**. Wasserstein-1 distance to the ground-truth conditional distribution. *Gain* denotes the relative reduction in W_1 of T-CMO with respect to single-task NCP. Lower is better.

	T-CMO **(ours)**	NCP	Gain
Varying n (samples per task), $K = 200$			
$n = 50$	0.281 ± 0.015	0.283 ± 0.013	1%
$n = 100$	0.169 ± 0.005	0.229 ± 0.007	26%
$n = 200$	0.099 ± 0.003	0.186 ± 0.005	47%
$n = 500$	0.068 ± 0.002	0.143 ± 0.005	52%
$n = 1000$	0.058 ± 0.001	0.128 ± 0.003	55%
$n = 2000$	0.052 ± 0.001	0.118 ± 0.003	56%
Varying K (number of tasks), $n = 200$			
$K = 50$	0.135 ± 0.004	0.184 ± 0.005	27%
$K = 100$	0.104 ± 0.003	0.183 ± 0.005	43%
$K = 200$	0.099 ± 0.003	0.186 ± 0.005	47%
$K = 300$	0.095 ± 0.003	0.186 ± 0.006	49%
$K = 500$	0.091 ± 0.003	0.181 ± 0.004	50%
$K = 1000$	0.096 ± 0.003	0.183 ± 0.006	48%

These results suggest that the shared functional spaces are particularly useful when the conditional distribution is difficult to recover from an individual task alone. Multi-task learning provides a representation adapted to the common structure of the task family, so that a new conditional distribution can be identified by estimating only its task-specific operator. For simpler conditional families, where a single-task model can already learn an accurate representation from relatively few observations, the benefit of sharing across tasks is correspondingly smaller.

Sensitivity to the number of tasks and samples. We finally investigate how transfer performance depends on the amount of information available during multi-task learning. We focus on **CD4** and vary either the number of observations per source task n, while fixing $K = 200$, or the number of source tasks K, while fixing $n = 200$. Figure 8 shows the corresponding trends, while Table 6 reports the same results numerically.

Increasing the number of observations per source task substantially improves T-CMO. At $n = 50$, T-CMO and single-task NCP perform almost identically, whereas the relative reduction in Wasserstein error reaches 47% at $n = 200$ and 56% at $n = 2000$. Increasing the number of source tasks has a similar effect: the gain rises from 27% at $K = 50$ to approximately 50% for $K \geq 300$, after which

the performance largely saturates. In contrast, single-task NCP remains essentially insensitive to K, since it does not exploit information from the additional source tasks. Overall, these results show that transfer improves as the shared functional spaces are learned from either more observations per task or a richer collection of related conditional distributions, with diminishing returns once the common representation is sufficiently well identified.

D.5 BASELINES

We compare MTL-CMO against both multi-task and single-task conditional distribution estimators. All methods are trained from the same observations and evaluated using the same conditioning inputs, output grid, and Wasserstein-1 metric described above.

Multi-task baselines. Pooled-NCP is a pooled version of Neural Conditional Probability (NCP) (Kostic et al., 2024). Samples from all source tasks are aggregated and a single NCP model is trained without access to the task identity. This baseline therefore exploits the larger pooled dataset, but does not model task-specific operators.

MTL-MDN combines hard parameter sharing (Caruana, 1997) with Mixture Density Networks (Bishop, 1994). A common neural backbone is shared by all tasks and each task has its own Gaussian-mixture output head. We use two mixture components for CD1 and four components for CD2–CD4.

DeepJMQR follows the joint multi-quantile regression approach of (Rodrigues & Pereira, 2020). We use a shared neural representation together with a task embedding and predict 99 conditional quantiles, trained with the pinball loss. The conditional CDF used for evaluation is recovered from the predicted quantiles.

Single-task baselines. NCP (Kostic et al., 2024) is trained independently on every task and therefore provides the direct task-specific counterpart of our multi-task operator model. Each model only observes samples from its corresponding task.

Conditional Flow Matching (CFM) is the flow-matching objective of (Lipman et al., 2023). A conditional velocity field is trained independently for each task and used to transport samples from the reference distribution to the target conditional distribution. Conditional CDFs are obtained from generated samples.

Mixture Density Network (MDN) (Bishop, 1994) directly parameterizes $p(y \mid x)$ as a Gaussian mixture whose parameters are predicted by a neural network. As for MTL-MDN, we use two Gaussian components for CD1 and four components for CD2–CD4.

Neural Spline Flow (NSF) (Durkan et al., 2019) models the conditional distribution using monotone rational-quadratic spline transformations. We use three spline transforms with eight bins and recover the conditional CDF from samples generated by the fitted flow.

Engression (Shen & Meinshausen, 2025) is a neural distributional regression method trained using the energy score. We use the authors' implementation with input/output standardization and obtain conditional CDFs from samples generated by the fitted model.

NGBoost (Duan et al., 2020) performs probabilistic gradient boosting using the natural gradient. We use a Gaussian predictive distribution, 100 boosting estimators, learning rate 0.1, and depth-2 regression trees as base learners.

FlexCode (Izbicki & Lee, 2017) represents the conditional density in an orthogonal series whose coefficients are estimated through regression. We use a Fourier basis, with Lasso regression for CD1–CD2 and Random Forest regression for CD3–CD4.

Nadaraya–Watson (NW) is a non-parametric kernel baseline based on the classical Nadaraya–Watson estimator (Nadaraya, 1964; Watson, 1964).

Model selection and evaluation. Hyperparameters are selected independently for each conditional-distribution family using held-out validation data; target evaluation tasks are not used for model selection. For neural baselines, architecture, learning rate, regularization, and method-specific parameters are selected using the same validation criterion. For Nadaraya–Watson, only the

input-kernel bandwidth is selected. Generative baselines are evaluated from their generated conditional samples, whereas methods providing an analytical or empirical CDF are evaluated directly. All methods are finally compared on exactly the same conditioning points and output grid.

E LANGEVIN EXPERIMENTS

E.1 TWO-DIMENSIONAL MÜLLER–BROWN FAMILY

Potential family. We consider the parametric Müller–Brown family introduced in the main text, derived from the standard Müller–Brown potential, a widely used benchmark for metastable dynamics and transition-path problems (Devergne et al., 2024; Le Treut et al., 2025).It corresponds to a two-dimensional overdamped Langevin dynamics,

$$d\mathbf{X}_t = -\nabla V_{r,\theta}(\mathbf{X}_t)\, dt + \sqrt{2\beta^{-1}}\, d\mathbf{W}_t, \qquad \beta = 1, \tag{43}$$

where $\mathbf{W}_t$ is a standard two-dimensional Brownian motion. For $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$, the potential is

$$V_{r,\theta}(\mathbf{x}) = \frac{1}{s}\sum_{j=1}^{4} w_j(r) A_j \exp\Big[(\mathbf{x}-\boldsymbol{\mu}_j)^\top \mathbf{H}_j^{(\theta)}(\mathbf{x}-\boldsymbol{\mu}_j)\Big] + c_{\text{wall}}\left[(x_1-x_c)^4 + (x_2-y_c)^4\right]. \tag{44}$$

Here $\mathbf{w}(r) = (1, 1, r, 1)$, $\mathbf{H}_3^{(\theta)} = \mathbf{R}_\theta \mathbf{H}_3 \mathbf{R}_\theta^\top$, and $\mathbf{H}_j^{(\theta)} = \mathbf{H}_j$ for $j \neq 3$, with $\mathbf{H}_j = \begin{pmatrix} a_j & b_j/2 \\ b_j/2 & c_j \end{pmatrix}$. The canonical Müller–Brown coefficients are $\mathbf{A} = (-200, -100, -170, 15)$, $\boldsymbol{\mu}_1 = (1, 0)$, $\boldsymbol{\mu}_2 = (0, 0.5)$, $\boldsymbol{\mu}_3 = (-0.5, 1.5)$, $\boldsymbol{\mu}_4 = (-1, 1)$, $\mathbf{a} = (-1, -1, -6.5, 0.7)$, $\mathbf{b} = (0, 0, 11, 0.6)$, and $\mathbf{c} = (-10, -10, -6.5, 0.7)$.

We use $s = 35$ and sample $r \sim \mathcal{U}([0.75, 1.15])$, $\theta \sim \mathcal{U}([-\pi/4, \pi/12])$. The parameter r modifies the amplitude of the third Müller–Brown component, whereas θ rotates its local quadratic form over a 60° range, while the other three components remain fixed.

A fixed quartic wall is added for numerical confinement, with $c_{\text{wall}} = 0.5$ and $(x_c, y_c) = (-0.25, 0.875)$. Its contribution is negligible in the metastable basins and increases rapidly away from the region of interest, preventing trajectories from exploring numerically irrelevant regions.

Figure 9 illustrates the complementary effects of the two task parameters. Varying r primarily changes the energetic importance of the third basin relative to the other metastable states. In contrast, varying θ rotates its anisotropic quadratic structure, modifying the orientation of the local level sets and of the corresponding drift field. The family therefore preserves the same global three-basin topology while inducing controlled task-dependent changes in the dynamics and associated spectral structure.

Simulation and dataset. For each (r, θ), we simulate the overdamped Langevin dynamics at a sampling rate $\Delta t_{\text{sim}} = 10^{-3}$ and down-sampled the trajectories to the sampling rate $\Delta t_{\text{obs}} = 10^{-2}$. We discard the first $5{,}000$ integration steps and retain $80{,}000$ observations for each system.

We generate 950 systems in total: 750 source tasks are used for multi-task learning, 50 tasks are reserved for model selection, and the remaining 150 systems are kept unseen for transfer evaluation. Performance is evaluated from trajectory prefixes $N \in \{2, 5, 10, 20, 40, 80\} \times 10^3$. All model-selection choices for the competing methods are made on the 50 validation tasks, separately for each trajectory length N.

Reference spectrum. For each task, reference eigenpairs are obtained by finite-difference discretization of the backward Langevin generator on $[-3, 3]^2$ using a 240×240 grid, corresponding to $\Delta x = 0.025$. We retain the first three non-trivial eigenpairs $(\lambda_j, \psi_j)_{j=1}^3$. To avoid numerical instabilities in regions far outside the relevant part of the energy landscape, the potential used to construct the discrete generator is clipped at $V_{\min} + 300$.

The stationary density is $\pi_{r,\theta}(\mathbf{x}) \propto \exp[-\beta V_{r,\theta}(\mathbf{x})]$. Reference and estimated eigenfunctions are normalized in $L^2(\pi_{r,\theta})$, and mode-wise recovery is measured using the sign-invariant cosine error $1 - \cos_\pi(\widehat{\psi}_j, \psi_j)$. We additionally report eigenspace and eigenvalue diagnostics below.

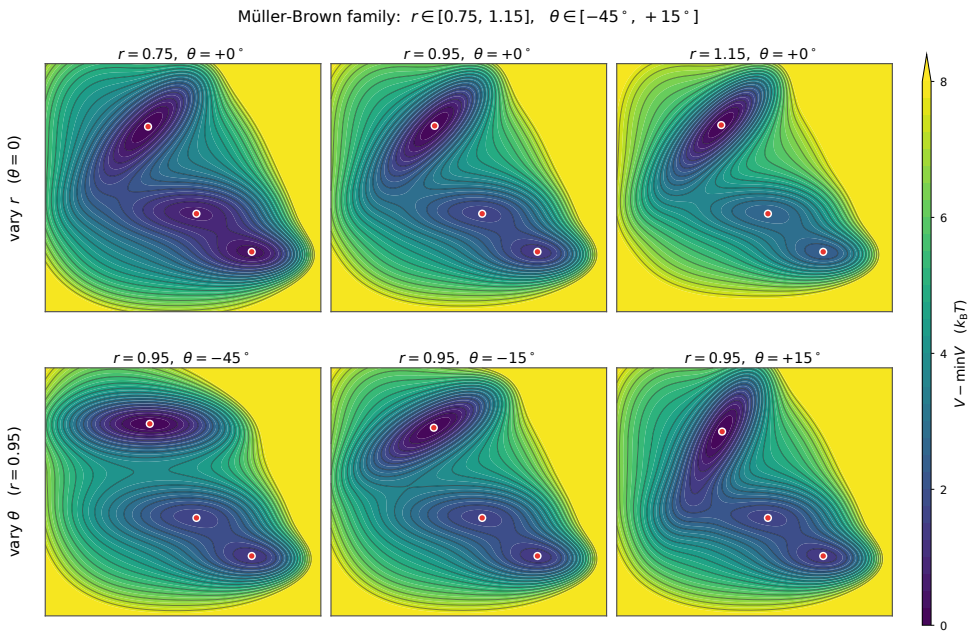

Figure 9: **Parametric Müller–Brown family.** Top: varying r at fixed $\theta = 0^\circ$ modifies the relative depth of the third metastable basin while largely preserving its orientation. Bottom: varying θ at fixed $r = 0.95$ rotates the local geometry of the same basin over a 60° range.

E.2 Shared representation and operator estimation

Multi-task representation learning. We instantiate the multi-task learning procedure of Section 3 with a single shared map $\Phi_\theta = \Psi_\omega = \phi_\theta$, exploiting the reversibility of the Langevin dynamics. The map $\phi_\theta : \mathbb{R}^2 \to \mathbb{R}^{64}$ is parameterized by a three-layer MLP with hidden width 576 and SiLU activations. During the multi-task stage, each system is associated with a rank-20 operator $M^{(k)} = A_k B_k^\top$, $A_k, B_k \in \mathbb{R}^{64\times 20}$. The final shared representation is learned on the 800 training systems. Training uses consecutive states, corresponding to $\tau = 0.01$, with temporal windows of length 2176. The remaining optimization settings are summarized in Table 7.

System-specific operator estimation. For each unseen system, we apply T-CMO using the closed-form transfer procedure of Algorithm 3, with adaptation rank 20 and no gradient updates. The resulting representation is then passed to the resolvent estimator described in Appendix A.2, from which we retain the three slowest non-trivial modes for evaluation.

Spectral-rank diagnostic. We examine the reduced-rank reconstruction error of the whitened lagged operator as a function of the retained rank. As shown in Fig. 10, the error decreases sharply over the first three components and then rapidly saturates. This provides an empirical diagnostic for the low-dimensional spectral structure captured by the first three modes.

Geometry of the learned operators. We next examine whether the task-specific operators learned in the shared functional space retain the structure of the underlying Müller–Brown family. Each recovered rank-3 operator is represented in the common ϕ_θ basis, and pairwise operator distances are computed using SGOT (Germain et al., 2026). We apply t-SNE to the resulting distance matrix and visualize the induced geometry in Figure 11. Importantly, neither r nor θ is provided during learning. Nevertheless, the embedding exhibits smooth gradients with respect to both parameters: operators associated with nearby values of r or θ occupy neighboring regions. This shows that the learned operators retain the structured variation of the underlying dynamical family without supervision from its physical parameters.

Operator estimation accuracy. We next assess whether sharing the functional representation improves operator estimation on the systems used during multi-task learning. Figure 12 compares

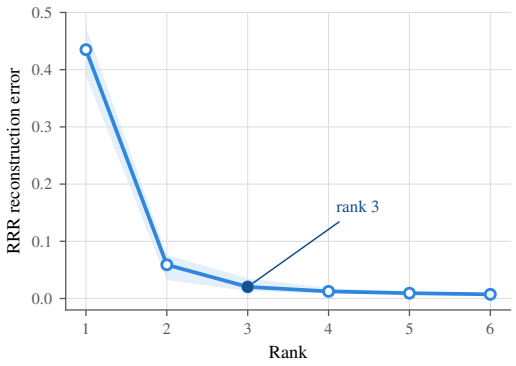

Figure 10: **Spectral-rank diagnostic.** Mean reduced-rank reconstruction error of the whitened lagged operator as a function of the retained rank. Most of the reduction occurs within the first three components, after which the error rapidly saturates.

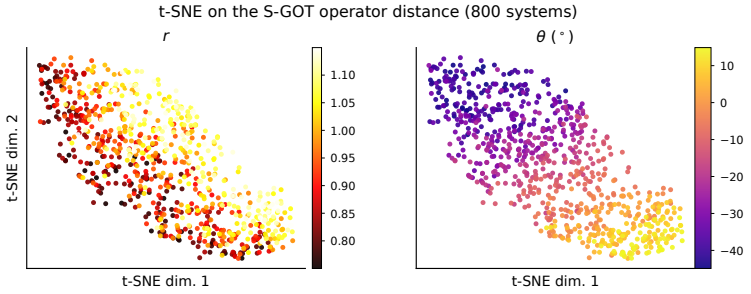

Figure 11: **Geometry of the learned Müller–Brown operators.** t-SNE of pairwise S-GOT distances between operators learned during the multi-task stage, colored by r (left) and θ (right).

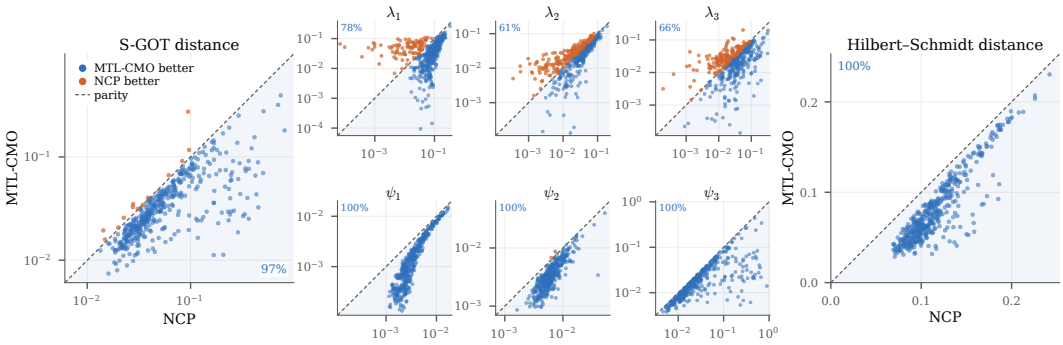

Figure 12: **Multi-task versus single-task CMO learning.** Eigenfunction errors of MTL-CMO versus independently trained single-task CMO models. Points below the diagonal favor MTL-CMO.

MTL-CMO with independently trained single-task CMO models. Most systems lie below the diagonal, showing that the shared representation improves eigenfunction recovery across the task family rather than on only a small subset of systems. The improvement is observed across all three modes and is strongest for the more difficult third mode ψ_3.

Table 7: Architecture, training, and spectral-estimation configuration for the Müller–Brown experiment.

Configuration	Value
Architecture	
Input dimension	2
Shared feature dimension d	64
Hidden layers	3×576
Activation	SiLU
Task-specific rank	20
Multi-task training	
Systems	800
Batch size	40
Window length	2176
Lag	1 step
Epochs	450
Optimizer	AdamW
Network learning rate	3.117×10^{-4}
Operator learning rate	1.780×10^{-4}
Weight decay	4.49×10^{-7}
λ	3.952×10^{-3}
Spectral estimation	
Adaptation rank	20
Spectral rank	3
Resolvent shift μ	4.2
Maximum lag	10,000
Regularization	3.339×10^{-8}

E.3 Transfer learning on unseen systems

Experimental protocol. We evaluate spectral recovery on the 150 unseen Müller–Brown systems. T-CMO adapts the shared representation to each target system using the closed-form procedure of Algorithm 3, with adaptation rank 20 and no gradient updates. Spectral quantities are then extracted with the resolvent-based estimator, and we retain the three slowest non-trivial modes, see Appendix A.3. Performance is evaluated from trajectory prefixes $N \in \{2, 5, 10, 20, 40, 80\} \times 10^3$. All methods receive exactly the same target prefixes; methods requiring task-specific estimation are refitted from the corresponding prefix, whereas transfer and meta-learning methods reuse their source-stage representation according to their respective procedures. Hyperparameters are selected on the 50 validation systems separately for each trajectory length.

Baselines. We compare against six complementary approaches. **Single-task NCP** (Kostic et al., 2024) learns an independent representation from each target trajectory and therefore uses no information from the source systems. **Pooled NCP** (Kostic et al., 2024) trains a single model on the concatenated source trajectories without task identities, testing whether pooling alone provides a transferable representation. **RFF-EDMD** combines fixed random Fourier features (Rahimi & Recht, 2007) with direct EDMD-type operator estimation (Kostic et al., 2022), while **RFF + resolvent** uses the same features with the resolvent-based spectral estimator (Kostic et al., 2025). **MetaKoopman** (Iwata & Kawahara, 2021) explicitly meta-learns Koopman spectral information across related short time series, and **VAMPNet** (Mardt et al., 2018) learns dominant dynamical modes through a neural variational objective.

Quantitative spectral recovery. The mode-wise results show where this improvement arises. For the eigenfunctions (Table 8), T-CMO attains the lowest mean error in 17 of the 18 mode–trajectory combinations. The only exception is ψ_3 at $N = 5{,}000$, where it is second to VAMPNet by 3.1×10^{-3}, well within the reported across-system variability. The first two modes become accurate for several methods as N increases, whereas ψ_3 remains substantially harder and provides the clearest separation between approaches. For the eigenvalues (Table 9), T-CMO achieves the lowest reported

Table 8: Eigenfunction errors for the first three non-trivial modes on the 150 unseen Müller–Brown systems. Values are mean ± standard deviation. Lower is better; **best** and second best per column.

	Trajectory length N					
Method	2 000	5 000	10 000	20 000	40 000	80 000
ψ_1						
T-CMO **(ours)**	$\mathbf{0.0731 \pm 0.0966}$	$\mathbf{0.0294 \pm 0.0414}$	$\mathbf{0.0151 \pm 0.0206}$	$\mathbf{0.0079 \pm 0.0095}$	$\mathbf{0.0040 \pm 0.0053}$	$\mathbf{0.0021 \pm 0.0026}$
RFF-EDMD	0.0982 ± 0.1537	0.0333 ± 0.0427	0.0248 ± 0.0821	0.0098 ± 0.0098	0.0058 ± 0.0054	0.0037 ± 0.0026
MetaKoopman	0.0837 ± 0.1154	0.0327 ± 0.0424	0.0174 ± 0.0209	0.0099 ± 0.0098	0.0058 ± 0.0055	0.0040 ± 0.0029
Single-task NCP	0.0839 ± 0.1071	0.0332 ± 0.0415	0.0177 ± 0.0206	0.0101 ± 0.0095	0.0059 ± 0.0054	0.0041 ± 0.0026
RFF + resolvent	0.0781 ± 0.0896	0.0329 ± 0.0421	0.0170 ± 0.0206	0.0093 ± 0.0095	0.0052 ± 0.0054	0.0033 ± 0.0026
Pooled NCP	0.1013 ± 0.1298	0.0332 ± 0.0425	0.0163 ± 0.0206	0.0083 ± 0.0095	0.0041 ± 0.0054	0.0022 ± 0.0026
VAMPNet	0.0900 ± 0.1310	0.0308 ± 0.0426	0.0164 ± 0.0213	0.0084 ± 0.0098	0.0042 ± 0.0054	0.0023 ± 0.0026
ψ_2						
T-CMO **(ours)**	$\mathbf{0.1174 \pm 0.1652}$	$\mathbf{0.0352 \pm 0.0350}$	$\mathbf{0.0173 \pm 0.0133}$	$\mathbf{0.0089 \pm 0.0063}$	$\mathbf{0.0046 \pm 0.0032}$	$\mathbf{0.0022 \pm 0.0014}$
RFF-EDMD	0.1607 ± 0.1744	0.0710 ± 0.1234	0.0453 ± 0.1162	0.0224 ± 0.0567	0.0085 ± 0.0046	0.0054 ± 0.0021
MetaKoopman	0.1722 ± 0.2028	0.0606 ± 0.0855	0.0361 ± 0.0745	0.0289 ± 0.0972	0.0144 ± 0.0323	0.0079 ± 0.0037
Single-task NCP	0.1890 ± 0.2045	0.0723 ± 0.1072	0.0378 ± 0.0738	0.0235 ± 0.0734	0.0149 ± 0.0656	0.0065 ± 0.0025
RFF + resolvent	0.2013 ± 0.2181	0.0761 ± 0.1102	0.0426 ± 0.0887	0.0295 ± 0.1061	0.0195 ± 0.0932	0.0053 ± 0.0023
Pooled NCP	0.2567 ± 0.2026	0.0890 ± 0.1145	0.0478 ± 0.0982	0.0349 ± 0.1270	0.0191 ± 0.1039	0.0036 ± 0.0023
VAMPNet	0.1559 ± 0.2062	0.0446 ± 0.0417	0.0235 ± 0.0158	0.0144 ± 0.0084	0.0081 ± 0.0044	0.0061 ± 0.0027
ψ_3						
T-CMO **(ours)**	$\mathbf{0.2555 \pm 0.2206}$	0.1378 ± 0.1490	$\mathbf{0.0756 \pm 0.0766}$	$\mathbf{0.0410 \pm 0.0397}$	$\mathbf{0.0263 \pm 0.0283}$	$\mathbf{0.0265 \pm 0.0186}$
RFF-EDMD	0.3527 ± 0.2495	0.2155 ± 0.2180	0.1392 ± 0.1820	0.1216 ± 0.1600	0.0595 ± 0.1098	0.0407 ± 0.0805
MetaKoopman	0.3405 ± 0.2370	0.2073 ± 0.1903	0.1416 ± 0.1534	0.1161 ± 0.1555	0.0805 ± 0.1244	0.0565 ± 0.0950
Single-task NCP	0.3682 ± 0.2335	0.2267 ± 0.2020	0.1593 ± 0.1865	0.1267 ± 0.1765	0.0939 ± 0.1615	0.0593 ± 0.1160
RFF + resolvent	0.3821 ± 0.2407	0.2376 ± 0.2083	0.1775 ± 0.2036	0.1363 ± 0.1851	0.1055 ± 0.1793	0.0678 ± 0.1376
Pooled NCP	0.4675 ± 0.2236	0.2781 ± 0.2165	0.2045 ± 0.2158	0.1502 ± 0.1988	0.1228 ± 0.2024	0.0739 ± 0.1482
VAMPNet	0.3118 ± 0.2496	$\mathbf{0.1347 \pm 0.1388}$	0.0829 ± 0.0813	0.0623 ± 0.0623	0.0410 ± 0.0533	0.0290 ± 0.0321

Table 9: Relative eigenvalue errors for the first three non-trivial modes on the 150 unseen Müller–Brown systems. Values are mean ± standard deviation. Lower is better; **best** and second best per column.

	Trajectory length N					
Method	2 000	5 000	10 000	20 000	40 000	80 000
λ_1						
T-CMO **(ours)**	$\mathbf{0.8288 \pm 2.8035}$	$\mathbf{0.1919 \pm 0.2668}$	$\mathbf{0.1241 \pm 0.1057}$	$\mathbf{0.0916 \pm 0.0824}$	$\mathbf{0.0643 \pm 0.0569}$	$\mathbf{0.0443 \pm 0.0367}$
RFF-EDMD	0.9852 ± 2.8284	0.3160 ± 0.3197	0.2710 ± 0.1682	0.1236 ± 0.1025	0.2239 ± 0.0935	0.1620 ± 0.0706
MetaKoopman	1.0705 ± 3.0177	0.2779 ± 0.4465	0.1842 ± 0.1614	0.1356 ± 0.1260	0.1043 ± 0.0853	0.0870 ± 0.0689
Single-task NCP	1.0360 ± 2.9714	0.2714 ± 0.4275	0.1829 ± 0.1489	0.1346 ± 0.1168	0.1026 ± 0.0793	0.0885 ± 0.0590
RFF + resolvent	1.0123 ± 2.9517	0.2567 ± 0.4141	0.1676 ± 0.1395	0.1202 ± 0.1095	0.0878 ± 0.0731	0.0695 ± 0.0530
Pooled NCP	1.0015 ± 2.9503	0.2490 ± 0.4097	0.1570 ± 0.1313	0.1088 ± 0.1026	0.0766 ± 0.0635	0.0512 ± 0.0458
VAMPNet	0.9758 ± 3.0234	0.2143 ± 0.3208	0.1348 ± 0.1203	0.0951 ± 0.0863	0.0643 ± 0.0570	0.0452 ± 0.0386
λ_2						
T-CMO **(ours)**	$\mathbf{0.2605 \pm 0.4016}$	$\mathbf{0.1140 \pm 0.0923}$	$\mathbf{0.0764 \pm 0.0659}$	$\mathbf{0.0520 \pm 0.0393}$	$\mathbf{0.0354 \pm 0.0304}$	$\mathbf{0.0240 \pm 0.0202}$
RFF-EDMD	0.2830 ± 0.3944	0.1269 ± 0.1085	0.0922 ± 0.0951	0.0564 ± 0.0464	0.0544 ± 0.0382	0.0395 ± 0.0269
MetaKoopman	0.3265 ± 0.5164	0.1515 ± 0.1390	0.1077 ± 0.0933	0.0801 ± 0.0647	0.0649 ± 0.0514	0.0569 ± 0.0388
Single-task NCP	0.3154 ± 0.4623	0.1498 ± 0.1325	0.1035 ± 0.0890	0.0751 ± 0.0617	0.0582 ± 0.0478	0.0482 ± 0.0348
RFF + resolvent	0.3122 ± 0.4538	0.1484 ± 0.1314	0.1021 ± 0.0866	0.0739 ± 0.0619	0.0559 ± 0.0466	0.0440 ± 0.0325
Pooled NCP	0.3111 ± 0.4416	0.1483 ± 0.1302	0.1019 ± 0.0842	0.0714 ± 0.0623	0.0525 ± 0.0454	0.0388 ± 0.0294
VAMPNet	0.3214 ± 0.5216	0.1308 ± 0.1264	0.0891 ± 0.0946	0.0566 ± 0.0492	0.0417 ± 0.0356	0.0310 ± 0.0239
λ_3						
T-CMO **(ours)**	$\mathbf{0.2021 \pm 0.2943}$	$\mathbf{0.0962 \pm 0.0798}$	$\mathbf{0.0646 \pm 0.0545}$	$\mathbf{0.0445 \pm 0.0373}$	$\mathbf{0.0338 \pm 0.0264}$	$\mathbf{0.0249 \pm 0.0215}$
RFF-EDMD	0.2055 ± 0.2940	0.1103 ± 0.1061	0.0837 ± 0.0871	0.0705 ± 0.0791	0.0467 ± 0.0494	0.0313 ± 0.0337
MetaKoopman	0.2753 ± 0.3363	0.1369 ± 0.1217	0.1157 ± 0.0882	0.0913 ± 0.0713	0.0781 ± 0.0621	0.0731 ± 0.0415
Single-task NCP	0.2640 ± 0.3232	0.1353 ± 0.1120	0.1059 ± 0.0865	0.0848 ± 0.0686	0.0681 ± 0.0585	0.0557 ± 0.0428
RFF + resolvent	0.2551 ± 0.3070	0.1360 ± 0.1100	0.1090 ± 0.0903	0.0899 ± 0.0808	0.0708 ± 0.0640	0.0541 ± 0.0477
Pooled NCP	0.2538 ± 0.3021	0.1380 ± 0.1151	0.1133 ± 0.0921	0.0921 ± 0.0858	0.0734 ± 0.0669	0.0539 ± 0.0498
VAMPNet	0.2370 ± 0.4022	0.1096 ± 0.1212	0.0769 ± 0.0622	0.0544 ± 0.0448	0.0446 ± 0.0361	0.0318 ± 0.0270

mean error in every column, up to a tie at displayed precision for λ_1 at $N = 40{,}000$. Thus, transfer improves both the spatial eigenfunctions and the associated relaxation rates.

Transfer versus single-task estimation. To isolate the contribution of the shared representation, we compare T-CMO directly with single-task NCP, which estimates each unseen system independently from the same target trajectory. Across the quantitative tables above, T-CMO has lower mean eigenfunction and eigenvalue errors than the single-task model for every reported mode and trajectory length. The improvement is modest for the easiest mode once trajectories become long, but becomes substantially larger for the harder second and third modes, showing that transfer is most useful when the target spectral structure is difficult to estimate from a single trajectory.

Figure 13 further reveals a clear hierarchy in spectral difficulty. Both methods recover the first two modes relatively accurately, whereas ψ_3 exhibits larger errors and stronger variability across systems. T-CMO reduces the error for all three modes, with the clearest separation on ψ_3. The energy-resolved curves show that the residual error is not spread uniformly over the state space: it

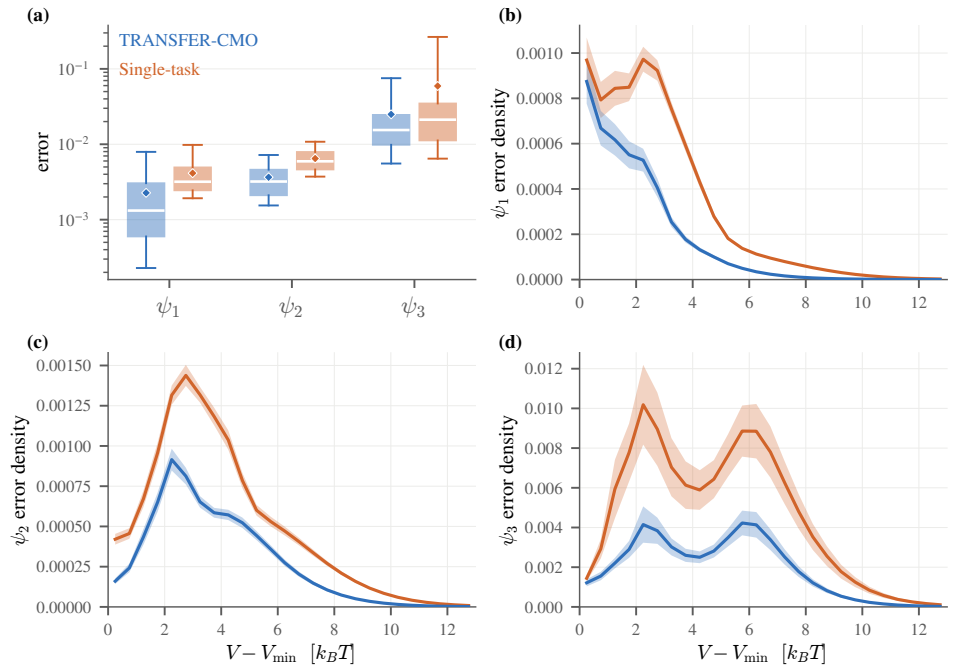

Figure 13: **Transfer versus single-task spectral recovery.** **(a)** Distribution of mode-wise $1 - \cos_\pi(\widehat{\psi}_j, \psi_j)$ errors over unseen systems. **(b–d)** Error distribution across potential-energy levels for ψ_1, ψ_2, and ψ_3.

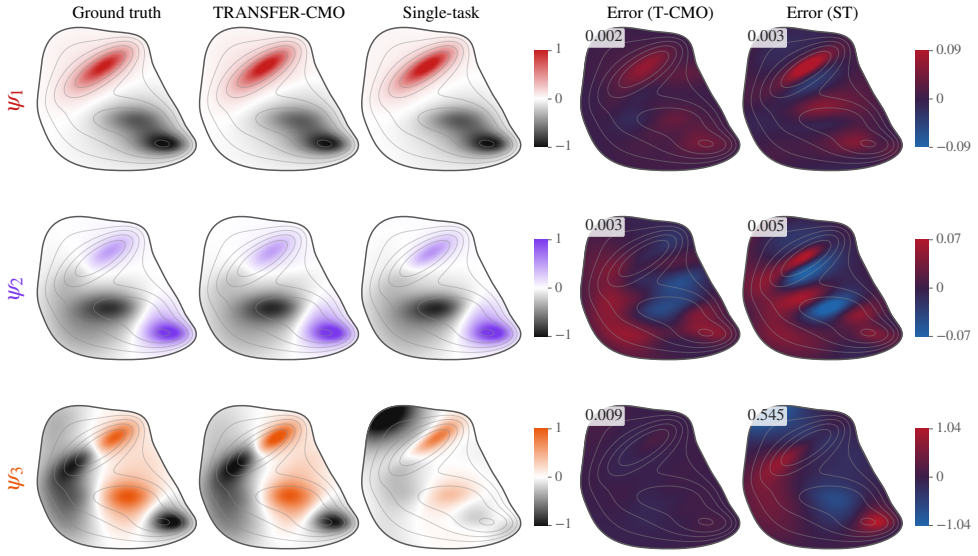

Figure 14: **Spatial recovery of the first three eigenfunctions on a representative unseen system.** From left to right: finite-difference reference, T-CMO, single-task NCP, and the corresponding pointwise errors. Eigenfunctions are sign-aligned and normalized in $L^2(\pi)$.

is concentrated over specific non-minimal energy ranges associated with the more difficult parts of the landscape. T-CMO suppresses these errors across the energy range rather than improving only isolated regions.

Figure 14 provides a spatial view of the same effect. The displayed eigenfunctions are multiplied by $\sqrt{\pi}$, so that $\|f\|^2_{L^2(\pi)} = \int |\sqrt{\pi(\mathbf{x})} f(\mathbf{x})|^2 \, d\mathbf{x}$ and visible spatial discrepancies directly reflect their contribution to the evaluation metric. For ψ_1 and ψ_2, both approaches reproduce the dominant structure, although T-CMO leaves smaller residual errors. The difference is much stronger for ψ_3: T-CMO remains close to the finite-difference reference, whereas the independently learned estimate exhibits substantial spatial distortion. Together with the quantitative results above, this shows that reusing the multi-task representation improves both global operator recovery and the more difficult task-specific spectral components.

F Plasma experiments

This experiment tests whether functional spaces learned from turbulent plasma trajectories, without access to the physical control parameters, recover the physical organization of the Tokam2D family. MTL-CMO first learns the shared representation across source systems; T-CMO then estimates a system-specific operator in closed form, from which we extract a finite-dimensional generator spectrum. The physical parameters (g, κ) are used only afterward as diagnostic variables. The complete experiment is run on a single NVIDIA RTX A6000 GPU.

F.1 Physical setting and plasma data

Physical model. The benchmark is based on a reduced two-field model for edge turbulence, describing the coupled evolution of density fluctuations $n(x, y, t)$ and the electrostatic potential $\phi(x, y, t)$ (Hasegawa & Wakatani, 1983; Ghendrih et al., 2018; 2022). The potential determines the perpendicular $\mathbf{E} \times \mathbf{B}$ flow and the vorticity,

$$\mathbf{V}_E = (-\partial_y \phi, \partial_x \phi), \qquad \Omega = \Delta_\perp \phi, \qquad \Delta_\perp = \partial_{xx} + \partial_{yy}.$$

Since $\mathbf{V}_E$ is divergence free, nonlinear transport can be written as $\mathbf{V}_E \cdot \nabla f = [\phi, f]$, with $[\phi, f] = \partial_x \phi \, \partial_y f - \partial_y \phi \, \partial_x f$.

We consider the gradient-driven configuration, with inverse density-gradient length $\kappa = 1/L_n$. The reduced dynamics are

$$\begin{cases} \partial_t n + \kappa \, \partial_y \phi + [\phi, n] - D_n \Delta_\perp n = -\sigma_n n + \sigma_{n\phi} \phi, \\ \partial_t \Omega + g \, \partial_y n + [\phi, \Omega] - D_\phi \Delta_\perp \Omega = -\sigma_{\phi n} n + \sigma_\phi \phi, \qquad \Omega = \Delta_\perp \phi. \end{cases}$$

The two parameters act through distinct mechanisms. The term $\kappa \, \partial_y \phi$ couples the imposed background density gradient to the radial $\mathbf{E} \times \mathbf{B}$ motion and drives density fluctuations, whereas $g \, \partial_y n$ couples density perturbations back into the vorticity equation and controls the interchange mechanism. Their nonlinear interaction closes the feedback loop $n \to \Omega \to \phi \to \mathbf{V}_E \to n$. Diffusion and linear loss terms are fixed throughout the dataset.

A physically important observable is the radial particle flux $\Gamma_x = n V_{E,x} = -n \, \partial_y \phi$, which measures radial transport induced by the correlation between density fluctuations and the $\mathbf{E} \times \mathbf{B}$ flow.

Simulation family. We generate $M = 400$ simulations using Tokam2D (GYSELAX Team, 2026), with $g \sim \mathcal{U}([0, 0.5])$ and $\kappa \sim \mathcal{U}([1, 3.5])$. All remaining physical and numerical parameters are fixed. Simulations are performed on a 128×128 spatial grid over a square domain of side length 64, with snapshots separated by $\Delta t = 0.05$. Each trajectory contains 4093 stored states. The first 320 systems are used for multi-task learning and the remaining 80 are held out for evaluation.

Figure 15 illustrates the distinct effects of the two physical parameters. Increasing κ strengthens the density-gradient drive and substantially changes the characteristic spatial scale of the potential and vorticity fields, with high-κ regimes exhibiting larger and more coherent structures. The effect of g is qualitatively different: at low κ, increasing g produces a finer and more fragmented vorticity field, consistently with its direct coupling to density perturbations through $g \, \partial_y n$. At larger κ, this effect interacts with the stronger density-gradient forcing. The dependence on (g, κ) is therefore not simply additive; both parameters jointly control the spatial organization of the turbulent flow.

Observed fields and preprocessing. Each state contains three physical channels: density n, vorticity Ω, and radial particle flux Γ_x. The original $3 \times 128 \times 128$ fields are downsampled by a factor of

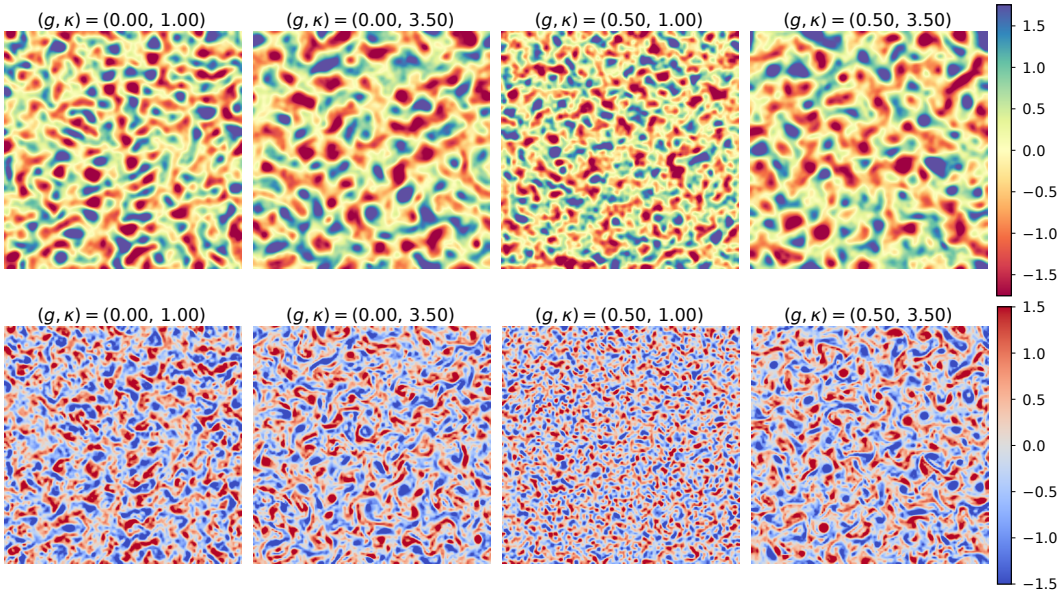

Figure 15: Representative electrostatic potential ϕ (top) and vorticity $\Omega = \Delta_\perp \phi$ (bottom) across the (g, κ) domain. Each field is standardized independently for visualization.

two before being passed to the network, giving $\mathbf{x}_t \in \mathbb{R}^{3\times 64\times 64}$. These observables respectively describe transported density, rotational flow structure, and radial turbulent transport. The electrostatic potential shown in Figure 15 is used only for physical interpretation and is not provided directly to MTL-CMO.

F.2 Shared functional representation

MTL-CMO parameterizes the shared dictionaries Φ_θ and Ψ_θ with a convolutional ResNet. A common encoder maps the three input fields to a 256-dimensional latent representation, followed by two output heads producing $d = 128$ functional features. The encoder contains three residual stages with channel dimensions $(64, 128, 256)$ and $(2, 2, 2)$ residual blocks, using GroupNorm, SiLU activations, and dropout. For each source system, the task-specific operator has rank 80 and is parameterized as $\mathbf{M}^{(k)} = \mathbf{A}^{(k)}\mathbf{B}^{(k)\top}$.

Training follows the MTL-CMO procedure of Algorithm 2 using consecutive states $(x_t^{(k)}, x_{t+1}^{(k)})$. Each optimization step samples 8 systems and 256 transitions per system. Input channels are standardized using statistics computed from the 320 source systems only, while dictionary outputs are centered within each sampled system window before evaluating the multi-task objective.

Table 10: Architecture and optimization configuration for the plasma experiment.

Configuration	Value
Architecture	
Input size	$3 \times 64 \times 64$
Shared feature dimension d	128
Task-specific rank	80
Encoder	Convolutional ResNet
Residual stages	$(64, 128, 256)$
Blocks per stage	$(2, 2, 2)$
Encoder output dimension	256
Output heads	$2 \times (256 \to 128)$
Normalization	GroupNorm (8 groups)
Activation	SiLU
Dropout	0.2
Optimization	
Systems per mini-batch	8
Transitions per system	256
Epochs	120
Optimizer	AdamW
Learning rate	3×10^{-4}
Minimum learning rate	10^{-6}
Weight decay	10^{-4}
Learning-rate schedule	Cosine
Warm-up	10%
Gradient clipping	1

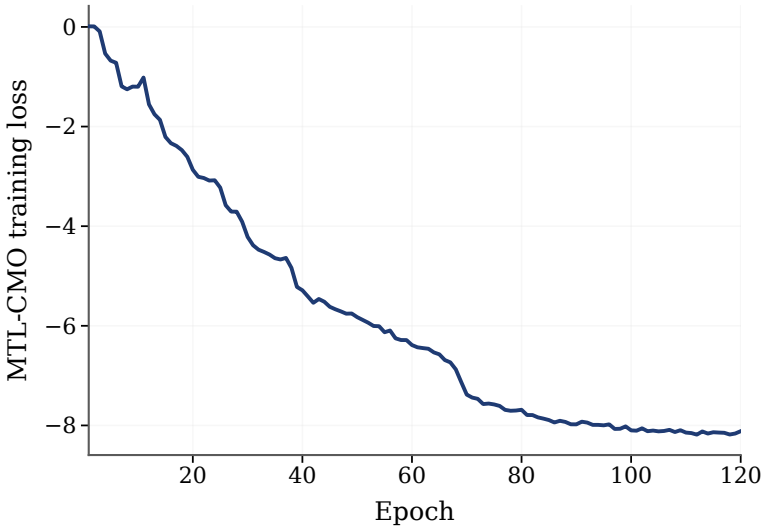

Figure 16: Evolution of the MTL-CMO training objective over 120 epochs.

F.3 System-specific transfer and spectral estimation

Closed-form adaptation. For each plasma system, we apply the T-CMO procedure of Algorithm 3 to the available trajectory. The shared dictionaries are reused without gradient updates, and only the system-specific operator is estimated. The same procedure is applied to source systems when constructing diagnostic representations and to held-out systems for transfer evaluation.

Adaptation rank. We select the transfer rank using a reduced-rank regression diagnostic. For each candidate rank, a one-step predictor is estimated independently on each of the 320 source systems and the reconstruction error is averaged across systems. As shown in Figure 17, the error decreases rapidly and largely saturates around $r = 32$. We therefore use adaptation rank 32 throughout the

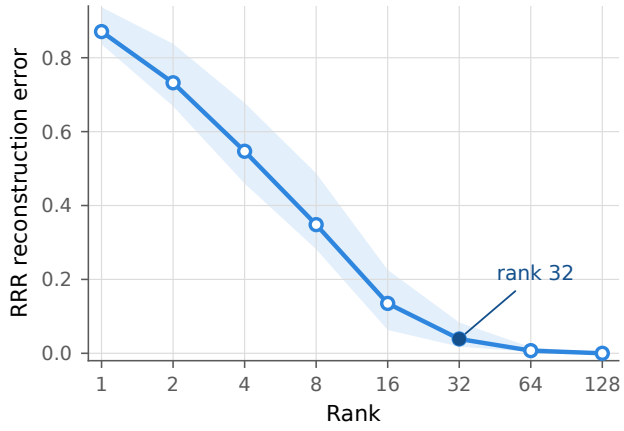

Figure 17: Mean one-step RRR reconstruction error across the 320 source systems versus retained rank. The transfer rank is fixed to 32.

plasma experiments. This is distinct from the rank-80 task-specific operators used during MTL-CMO training.

Generator spectrum. As explained in Appendix A.3, the resulting 32-dimensional adapted trajectories are passed to the generator-resolvent estimator described in Appendix A.2. We retain the 31 slowest recovered generator eigenvalues and represent each system by their real and absolute imaginary parts, yielding a 62-dimensional spectral descriptor. We use resolvent shift $s = 10$ throughout the plasma experiments.

F.4 Physical identification and data efficiency

Compared representations. We compare four dynamical representations. **MTL-CMO/T-CMO** uses the shared ResNet learned across the source family, followed by closed-form T-CMO adaptation and the generator-resolvent pipeline, yielding 62 spectral features. **Pooled NCP** uses the same neural architecture and source data but learns a single pooled representation without task-specific operators; the same adaptation and spectral-estimation pipeline is used afterward. **RFF** replaces the learned functional representation by 128 Gaussian random Fourier features, with bandwidth determined from the source systems, and uses the same downstream spectral estimator. Finally, **MIDST** learns a shared latent dynamical model with a 128-dimensional system-specific representation; for an unseen system its shared parameters are frozen and the system-specific coefficients are adapted by gradient descent. We use lag 10, selected from the source systems.

For physical identification, the spectral coordinates of MTL-CMO/T-CMO, Pooled NCP, and RFF are the real and absolute imaginary parts of the 31 slowest recovered eigenvalues. MIDST uses its native 128-dimensional system-specific representation. Features are standardized using the 320 source systems only. We fit separate RBF kernel-ridge regressors for g and κ, using $\alpha = 10^{-2}$ and $\gamma = 10^{-3}$ for every method, with no method-specific tuning. Performance is measured by R^2 on the 80 held-out systems.

Geometry of the learned representations. We first examine whether the unsupervised representations organize the source systems according to the underlying physics. For each method, we compute pairwise Euclidean distances between standardized system descriptors and apply classical multidimensional scaling. The values of g and κ are used only afterward to color the embeddings.

The MTL-CMO representation exhibits a continuous organization with respect to both physical parameters. The ordering is particularly pronounced for κ, while g induces a weaker but still structured variation across the embedding. Since neither parameter is observed during representation learning or operator estimation, this organization indicates that the learned spectrum captures physically meaningful variation across turbulent regimes rather than direct supervision from (g, κ).

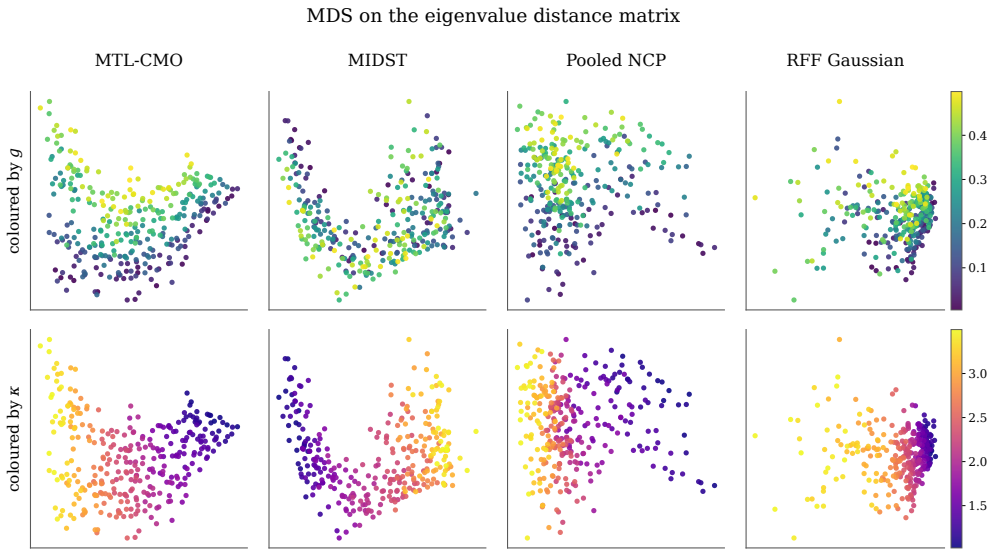

Figure 18: MDS embeddings of the dynamical representations of the 320 source systems, colored by g (top) and κ (bottom).

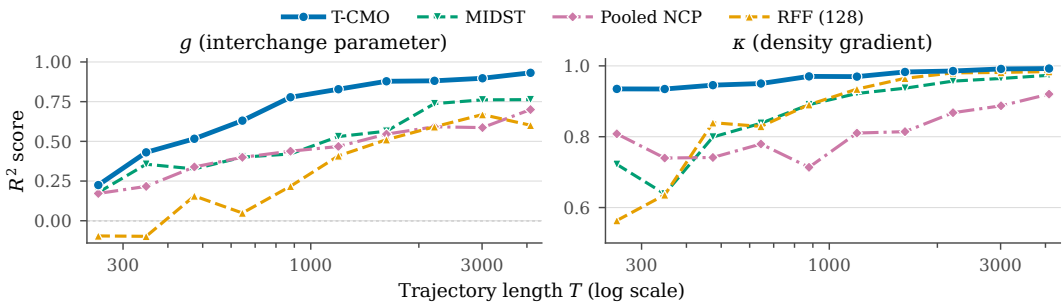

Figure 19: Physical-parameter recovery on held-out systems versus trajectory length for T-CMO and competing representations.

Identification from short trajectories. We next recompute every system-dependent representation from increasingly short trajectory prefixes. The source/held-out split and downstream regression protocol remain unchanged. Figure 19 and show that T-CMO yields the highest R^2 for both parameters at every trajectory length.

For κ, the spectral representation is already highly informative from only 256 frames, with $R^2 = 0.935$, and reaches 0.992 on the complete trajectory. Recovering g is substantially harder: T-CMO increases from $R^2 = 0.225$ at 256 frames to 0.778 at 878 and 0.932 on the full trajectory. This difference is consistent with the spectral geometry above: the density-gradient parameter is strongly expressed in the leading spectral structure, whereas the interchange parameter is encoded through a more distributed signature and benefits more from additional temporal information.

The comparisons separate the contributions of representation learning and multi-task structure. RFF keeps the downstream spectral estimator but removes the learned functional space, while Pooled NCP retains the neural architecture but removes the task-specific multi-task factorization. Their lower performance, especially for g, shows that the gain cannot be explained by the downstream regressor or spectral estimator alone. MIDST benefits from cross-system learning but remains less accurate, particularly for short trajectories and for recovery of the interchange parameter.

Table 11: Held-out R^2 for physical-parameter recovery versus available trajectory length. Lower-dimensional spectral methods use 62 features; MIDST uses its native 128-dimensional representation.

Trajectory length T	256	474	878	1 625	3 008	4 093
Interchange parameter g						
T-CMO **(ours)**	**0.225**	**0.516**	**0.778**	**0.878**	**0.897**	**0.932**
MIDST	0.178	0.326	0.421	0.564	0.761	0.763
Pooled NCP	0.172	0.339	0.438	0.546	0.587	0.700
RFF (128)	−0.096	0.156	0.217	0.511	0.669	0.692
Density-gradient parameter κ						
T-CMO **(ours)**	**0.935**	**0.946**	**0.970**	**0.983**	**0.991**	**0.992**
MIDST	0.722	0.799	0.891	0.937	0.964	0.973
Pooled NCP	0.808	0.741	0.714	0.814	0.887	0.920
RFF (128)	0.563	0.840	0.890	0.965	0.982	0.984

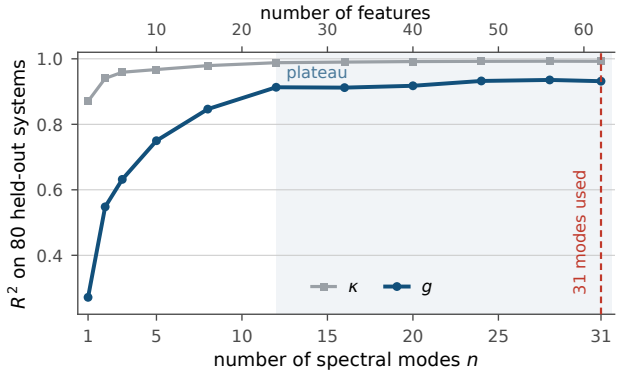

Figure 20: Held-out R^2 versus the number of slow generator modes retained from the T-CMO representation. Each mode contributes two spectral coordinates.

Number of retained spectral modes. We finally vary the number n of slow generator modes supplied to the diagnostic regressor. Each mode contributes two coordinates, $\mathrm{Re}\,\lambda_j$ and $|\,\mathrm{Im}\,\lambda_j|$. Figure 20 shows that κ is largely encoded by the leading modes and saturates rapidly. In contrast, g benefits from a broader part of the slow spectrum and largely plateaus only after roughly a dozen modes. We retain all 31 modes in the main evaluation, yielding 62 spectral features.

RFF dimension sensitivity. The Gaussian RFF baseline exhibits a clear optimum around $D = 128$ on complete trajectories. Increasing the number of random features beyond this value does not improve physical identification.

Table 12: Gaussian RFF sensitivity to the random-feature dimension D on complete trajectories.

Features D	32	64	**128**	256	512	1024
Interchange g	0.388	0.662	**0.714**	0.641	0.639	0.639
Density gradient κ	0.968	0.979	**0.986**	0.985	0.983	0.982

Table 13: Physical-identification evaluation settings.

Parameter	Value
Source systems	320
Held-out systems	80
T-CMO / Pooled NCP spectral features	62
RFF dictionary dimension	128
RFF spectral features	62
MIDST features	128
MIDST lag	10
Regression model	Kernel ridge, RBF
Feature preprocessing	StandardScaler
KRR regularization α	10^{-2}
RBF parameter γ	10^{-3}
Hyperparameter selection	Fixed, no cross-validation
Metric	Held-out R^2

G STATISTICAL GUARANTEES

We study how well MTL-CMO learns the shared representation, and how this depends on the number of tasks. The analysis has two steps. First, we show that the population multi-task loss controls, deterministically, the quality of the learned representation: the error of the best head on each training task, and the error on new tasks that share the same structure (propositions 1 and 2). Second, we bound the population loss of the trained model (section G.1). The cost of learning the shared representation is paid with the total number of samples nK, while each task only pays for its own head.

Setting. For every task $k \in [K]$, we observe n i.i.d. pairs $\{(x_i^{(k)}, y_i^{(k)})\}_{i\in[n]}$ from ρ_k, the K datasets being independent. We assume that ρ_k has a density p_k with respect to $\mu_k \otimes \nu_k$ such that $g_k := p_k - 1 \in L^2(\mu_k \otimes \nu_k)$. Then $\mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}}$ is the integral operator with kernel g_k, i.e. $[\mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} f](x) = \mathbb{E}_{\nu_k}[g_k(x,Y)f(Y)]$, and $\left\|\mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}}\right\|_{\mathrm{HS}} = \|g_k\|_{L^2(\mu_k\otimes\nu_k)}$. We write $h = (\theta, \{\mathbf{A}^{(k)}, \mathbf{B}^{(k)}\}_{k\in[K]})$ for the parameters of the whole model and

$$q_{h_k}(x,y) := \Phi_\theta(x)^\top \mathbf{A}^{(k)} \mathbf{B}^{(k)\top} \Psi_\theta(y)$$

for the kernel learned for task k.

Assumption 1. *There exists $c \geq 1$ such that $\|\Phi_\theta(x)\|_2 \leq c\sqrt{d}$ and $\|\Psi_\theta(y)\|_2 \leq c\sqrt{d}$ for all $\theta \in \Theta$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The heads belong to $\mathcal{W}_\rho := \{(\mathbf{A},\mathbf{B}) : \|\mathbf{A}\|_{\mathrm{F}}^2 + \|\mathbf{B}\|_{\mathrm{F}}^2 \leq 2\rho\}$ for some $\rho > 0$.*

The bound on the features holds, for instance, when the last layer of the networks has a bounded activation. The constraint on the heads is the constrained counterpart of the ridge penalty of section 3. We believe that an analogous guarantee holds for the penalized estimator used in practice, but a complete proof requires significantly more technical work, which we leave for future work. Under assumption 1, $\left\|\mathbf{A}\mathbf{B}^\top\right\|_{\mathrm{F}} \leq \rho$ and $|q_{h_k}| \leq Q_\rho := \rho c^2 d$ by Cauchy-Schwartz.

Estimator and population risk. For task k, the model is the operator $\mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}}$ with kernel q_{h_k}, and the population loss of section 3 is $\mathcal{L}^{(k)}(h) = \left\|\mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}} - \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}}\right\|^2_{\mathrm{HS}} - \left\|\mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}}\right\|^2_{\mathrm{HS}}$. Since the last term

does not depend on h, we measure the performance of the model by

$$\mathcal{E}_k(h) := \left\| \mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}} - \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} \right\|^2_{\mathrm{HS}} = \|q_{h_k} - g_k\|^2_{L^2(\mu_k \otimes \nu_k)},$$
$$\overline{\mathcal{E}}(h) := \frac{1}{K}\sum_{k=1}^{K} \mathcal{E}_k(h), \qquad \mathcal{A}_K := \inf_{h \in \Theta \times \mathcal{W}_\rho^K} \overline{\mathcal{E}}(h), \tag{45}$$

where the second equality holds because $\mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}} - \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}}$ is the integral operator with kernel $q_{h_k} - g_k$. The empirical loss $\widehat{\mathcal{L}}^{(k)}(h)$ of section 3 is an unbiased estimator of $\mathcal{L}^{(k)}(h)$, so that minimizing the multi-task objective amounts to minimizing an empirical version of $\overline{\mathcal{E}}$. We consider any $\widehat{h} = (\widehat{\theta}, \{\widehat{\mathbf{A}}^{(k)}, \widehat{\mathbf{B}}^{(k)}\}_k) \in \Theta \times \mathcal{W}_\rho^K$ such that

$$\frac{1}{K}\sum_{k=1}^{K} \widehat{\mathcal{L}}^{(k)}(\widehat{h}) \le \inf_{h \in \Theta \times \mathcal{W}_\rho^K} \frac{1}{K}\sum_{k=1}^{K} \widehat{\mathcal{L}}^{(k)}(h) + \epsilon_{\mathrm{opt}}. \tag{46}$$

Quality of a representation. For a task k, let $\Phi^{(k)}_{\theta,c} := \Phi_\theta - \mathbb{E}_{\mu_k}[\Phi_\theta(X)]$ and $\Psi^{(k)}_{\theta,c} := \Psi_\theta - \mathbb{E}_{\nu_k}[\Psi_\theta(Y)]$ be the centered dictionaries, and $P^{(k)}_{\Phi_\theta}$, $P^{(k)}_{\Psi_\theta}$ the orthogonal projections in $\mathrm{L}^2_{\mu_\mathrm{k}}(\mathcal{X})$ and $\mathrm{L}^2_{\nu_\mathrm{k}}(\mathcal{Y})$ onto their spans. We measure the quality of the representation θ for task k by the error of the best head on the learned spaces,

$$\eta_k^2(\theta) := \min_{\mathbf{M} \in \mathbb{R}^{d \times d}} \left\| \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} - \mathsf{S}_{\Phi^{(k)}_{\theta,c}} \mathbf{M}\, \mathsf{S}^*_{\Psi^{(k)}_{\theta,c}} \right\|^2_{\mathrm{HS}}. \tag{47}$$

This quantity does not depend on the heads, and it is defined in the same way for a new task, with $\mathsf{D}^{(\mathrm{new})}_{\mathrm{Y|X}}$ in place of $\mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}}$.

Approximation result. Let $\tau_k^2 := \sum_{i > r_k} (\sigma_i^{(k)})^2$ be the tail of the singular values of $\mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}}$. The next proposition shows that the risk (45) controls the quality of the representation on the training tasks.

Proposition 1. *Let $h = (\theta, \{\mathbf{A}^{(k)}, \mathbf{B}^{(k)}\}_k) \in \Theta \times \mathcal{W}_\rho^K$ and $k \in [K]$.*

(i) Best head. $\eta_k^2(\theta) = \left\| \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} - P^{(k)}_{\Phi_\theta}\, \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}}\, P^{(k)}_{\Psi_\theta} \right\|^2_{\mathrm{HS}}$, *and the minimum in* (47) *is attained at the T-CMO solution* $\mathbf{M}_k^* = \mathbf{G}^{(k)+}_{\Phi_\theta} \mathbf{C}^{(k)}_{\Phi_\theta \Psi_\theta} \mathbf{G}^{(k)+}_{\Psi_\theta}$.

(ii) Risk controls the representation. *Let* $T := \mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}}$, *so that* $\mathcal{E}_k(h) = \left\| T - \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} \right\|^2_{\mathrm{HS}}$ *by* (45). *Since T has rank at most r_k,*

$$\tau_k^2 \;\le\; \mathcal{E}_k(h), \qquad \eta_k^2(\theta) \;\le\; \mathcal{E}_k(h).$$

In particular, $\mathcal{A}_K \ge \frac{1}{K}\sum_k \tau_k^2$, *with equality if (A2) is realizable, i.e. if some $h \in \Theta \times \mathcal{W}_\rho^K$ satisfies that q_{h_k} is the kernel of $\mathsf{D}^{(\mathrm{k}),\mathrm{r_k}}_{\mathrm{Y|X}}$ for all k.*

Here $^+$ denotes the Moore–Penrose pseudo-inverse, which coincides with the inverse when the Gram matrices are invertible; it accounts for possibly redundant dictionaries. Item (i) shows that $\eta_k(\theta)$ is the population error of T-CMO with the frozen representation θ, and item (ii) that the risk of any trained model bounds the quality of its representation.

New tasks. A new task shares the representation of the training tasks if its operator factorizes through the same functions. We thus consider (A2) in the following form: there exist $\Phi^\star \colon \mathcal{X} \to \mathbb{R}^{d^\star}$ and $\Psi^\star \colon \mathcal{Y} \to \mathbb{R}^{d^\star}$ such that every task, training or new, satisfies

$$g_k(x,y) = \Phi^\star_{c,k}(x)^\top \mathbf{M}_k\, \Psi^\star_{c,k}(y) \qquad \text{for some } \mathbf{M}_k \in \mathbb{R}^{d^\star \times d^\star}, \tag{48}$$

where $\Phi^\star_{c,k}, \Psi^\star_{c,k}$ are centered with respect to μ_k, ν_k, and $k = \mathrm{new}$ for the new task. Let $\mathbf{G}^{(k)}_{\Phi^\star}$ and $\mathbf{G}^{(k)}_{\Psi^\star}$ be the corresponding covariance matrices. Transferring to any new task of the form (48) requires the training tasks to use all shared directions, which we quantify by

$$\lambda_\Phi := \lambda_{\min}\Big(\frac{1}{K}\sum_{k=1}^{K}\mathbf{M}_k\mathbf{G}^{(k)}_{\Psi^\star}\mathbf{M}_k^\top\Big), \qquad \lambda_\Psi := \lambda_{\min}\Big(\frac{1}{K}\sum_{k=1}^{K}\mathbf{M}_k^\top\mathbf{G}^{(k)}_{\Phi^\star}\mathbf{M}_k\Big).$$

These quantities are the operator counterpart of the task diversity condition used in multi-task representation learning, where the smallest singular value of the matrix collecting the source heads is required to be bounded away from zero (Tripuraneni et al., 2020; Du et al., 2021); here the heads are matrices and each direction is weighted by the covariance of the shared output features. For instance, $\lambda_\Phi, \lambda_\Psi > 0$ as soon as one training task has an invertible head and non-degenerate covariance matrices. For the new task, we let $\sigma_\Phi := \Big\|\mathbf{M}_{\mathrm{new}}\mathbf{G}^{(\mathrm{new})}_{\Psi^\star}\mathbf{M}_{\mathrm{new}}^\top\Big\|_{\mathrm{op}}$ and $\sigma_\Psi := \Big\|\mathbf{M}_{\mathrm{new}}^\top\mathbf{G}^{(\mathrm{new})}_{\Phi^\star}\mathbf{M}_{\mathrm{new}}\Big\|_{\mathrm{op}}$. Finally, the representation is learned where the training inputs lie, which we account for with $\beta_\mu := \max_k \|\mathrm{d}\,\mu_{\mathrm{new}}/\,\mathrm{d}\,\mu_k\|_\infty$ and $\beta_\nu := \max_k \|\mathrm{d}\,\nu_{\mathrm{new}}/\,\mathrm{d}\,\nu_k\|_\infty$; $\beta_\mu = \beta_\nu = 1$ when the marginals coincide.

Proposition 2. *Assume* (48) *with* $\lambda_\Phi, \lambda_\Psi > 0$*. For every* $\theta \in \Theta$*,*

$$\eta^2_{\mathrm{new}}(\theta) \;\le\; C_{\mathrm{new}} \cdot \frac{1}{K}\sum_{k=1}^{K}\eta^2_k(\theta), \qquad C_{\mathrm{new}} := \frac{\beta_\mu\sigma_\Phi}{\lambda_\Phi} + \frac{\beta_\nu\sigma_\Psi}{\lambda_\Psi}. \tag{49}$$

A representation that is accurate on average over the training tasks is thus accurate for every new task sharing it, and since $\eta_{\mathrm{new}}(\theta)$ is the population error of T-CMO (proposition 1(i)), proposition 2 quantifies transfer. The constant C_{new} is small when the training tasks are diverse (large $\lambda_\Phi, \lambda_\Psi$). Both ingredients are necessary: a shared direction that no training task uses cannot be learned, and the learned features are not constrained where no training input lies.

By propositions 1 and 2, all these guarantees follow from a bound on the average risk $\overline{\mathcal{E}}(\widehat{h})$ of the trained model.

G.1 Statistical rate

We measure the complexity of the shared dictionaries on the pooled sample of all tasks. For $m \ge 1$, let $X_l^{(k)} \sim \mu_k$ ($k \in [K]$, $l \in [m]$) be independent and γ_{klj} i.i.d. standard Gaussian variables, and define

$$\mathfrak{G}_{m,K}(\mathcal{F}_\Phi) := \frac{1}{mK}\,\mathbb{E}\sup_{\theta\in\Theta}\sum_{k=1}^{K}\sum_{l=1}^{m}\sum_{j=1}^{d}\gamma_{klj}\,\phi_j^\theta\big(X_l^{(k)}\big), \tag{50}$$

and $\mathfrak{G}_{m,K}(\mathcal{F}_\Psi)$ in the same way with $Y_l^{(k)} \sim \nu_k$. For norm-bounded neural networks, $\mathfrak{G}_{m,K}(\mathcal{F}) \le \mathrm{Comp}(\mathcal{F})/\sqrt{mK}$ with $\mathrm{Comp}(\mathcal{F})$ independent of m and K (Golowich et al., 2018). We use Gaussian rather than Rademacher complexities because the separation between the shared representation and the task-specific heads relies on the chain rule of Maurer et al. (2016), which is based on Gaussian comparison inequalities; the two complexities are equivalent up to a factor $\sqrt{\log(mKd)}$ (Ledoux & Talagrand, 1991, Eq. (4.9)). We set

$$L_\rho := \rho c\sqrt{d}\,(1+Q_\rho), \qquad B_\rho := 4Q_\rho(Q_\rho+3).$$

Theorem 2. *Let assumption 1 hold and* $n \ge 2$ *be even. There exists an absolute constant* $C > 0$ *such that, for any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*,*

$$\overline{\mathcal{E}}(\widehat{h}) \;\le\; \mathcal{A}_K + \epsilon_{\mathrm{opt}} + C\,L_\rho\Big[\underbrace{\mathfrak{G}_{n/2,K}(\mathcal{F}_\Phi) + \mathfrak{G}_{n/2,K}(\mathcal{F}_\Psi)}_{\textit{shared representation}} + \underbrace{c\sqrt{d/n}}_{\textit{heads}}\Big] + 2B_\rho\sqrt{\frac{\log(2\delta^{-1})}{2nK}}. \tag{51}$$

For neural networks, the statistical error is of order

$$L_\rho\,\frac{\mathrm{Comp}(\mathcal{F}_\Phi)+\mathrm{Comp}(\mathcal{F}_\Psi)}{\sqrt{nK}} + L_\rho\,c\sqrt{\frac{d}{n}}.$$

The complexity of the networks, which is typically the dominant term, is paid with the nK samples of all tasks, whereas each head only costs $\sqrt{d/n}$. Learning a single task ($K = 1$) gives $\mathrm{Comp}(\mathcal{F})/\sqrt{n}$: sharing the representation across K tasks divides the dominant term by $\sqrt{K}$. This is the operator counterpart of the benefit of multi-task representation learning of Maurer et al. (2016). The price for sharing is the approximation error $\mathcal{A}_K$, which is small only when the tasks share their dominant singular spaces, i.e. under (A3). As in Kostic et al. (2024), the optimization error is not analyzed.

Combining theorem 2 with propositions 1 and 2 gives the transfer guarantee: with probability at least $1 - \delta$,

$$\eta^2_{\mathrm{new}}(\widehat{\theta}) \le C_{\mathrm{new}} \cdot \mathrm{RHS}(51),$$

where RHS(51) is the right-hand side of (51). The error of T-CMO on a new task thus decreases with the total number of training samples nK, up to the approximation terms.

G.2 Proof of propositions 1 and 2

Lemma 1. *Let* $\mathsf{S}_1 \colon \mathbb{R}^d \to \mathrm{L}^2_{\mu_k}(\mathcal{X})$ *and* $\mathsf{S}_2 \colon \mathbb{R}^d \to \mathrm{L}^2_{\nu_k}(\mathcal{Y})$ *be linear, with* P_1, P_2 *the orthogonal projections onto their ranges. For every Hilbert–Schmidt operator* D, $\min_{\mathbf{M}} \|\mathsf{D} - \mathsf{S}_1 \mathbf{M} \mathsf{S}_2^*\|_{\mathrm{HS}} = \|\mathsf{D} - P_1 \mathsf{D} P_2\|_{\mathrm{HS}}$, *attained at* $\mathbf{M}^* = (\mathsf{S}_1^*\mathsf{S}_1)^+ \mathsf{S}_1^* \mathsf{D} \mathsf{S}_2 (\mathsf{S}_2^*\mathsf{S}_2)^+$. *Moreover,*

$$\|\mathsf{D} - P_1\mathsf{D}P_2\|^2_{\mathrm{HS}} = \|(I-P_1)\mathsf{D}\|^2_{\mathrm{HS}} + \|P_1\mathsf{D}(I-P_2)\|^2_{\mathrm{HS}} = \|\mathsf{D}(I-P_2)\|^2_{\mathrm{HS}} + \|(I-P_1)\mathsf{D}P_2\|^2_{\mathrm{HS}}.$$

Proof. Since $P_i = \mathsf{S}_i(\mathsf{S}_i^*\mathsf{S}_i)^+\mathsf{S}_i^*$, the set $\{\mathsf{S}_1 \mathbf{M} \mathsf{S}_2^*\}$ equals $\{T : T = P_1 T P_2\}$. The map $T \mapsto P_1 T P_2$ is idempotent and self-adjoint for the Hilbert–Schmidt inner product, hence it is the orthogonal projection onto this set, and the closest point to D is $P_1\mathsf{D}P_2$, which corresponds to $\mathbf{M}^*$. The last identities follow from $\mathsf{D} - P_1\mathsf{D}P_2 = (I-P_1)\mathsf{D} + P_1\mathsf{D}(I-P_2) = \mathsf{D}(I-P_2) + (I-P_1)\mathsf{D}P_2$, where in both cases the two terms are orthogonal for the Hilbert–Schmidt inner product. □

Proof of proposition 1. (i) Apply lemma 1 with $\mathsf{S}_1 = \mathsf{S}_{\Phi^{(k)}_{\theta,c}}$ and $\mathsf{S}_2 = \mathsf{S}_{\Psi^{(k)}_{\theta,c}}$, for which $\mathsf{S}_1^*\mathsf{S}_1 = \mathbf{G}^{(k)}_{\Phi_\theta}$, $\mathsf{S}_2^*\mathsf{S}_2 = \mathbf{G}^{(k)}_{\Psi_\theta}$ and $\mathsf{S}_1^* \, \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} \, \mathsf{S}_2 = \mathbf{C}^{(k)}_{\Phi_\theta \Psi_\theta}$.

(ii) Let $T := \mathsf{D}^{(\mathrm{k}),\theta}_{\mathrm{Y|X}}$, so that $\mathcal{E}_k(h) = \left\| T - \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} \right\|^2_{\mathrm{HS}}$ by (45). Since T has rank at most r_k, the Eckart–Young–Mirsky theorem gives $\mathcal{E}_k(h) \ge \tau_k^2$, with equality when q_{h_k} is the kernel of $\mathsf{D}^{(\mathrm{k}),\mathrm{r_k}}_{\mathrm{Y|X}}$. Moreover, let Π denote the projection onto the orthogonal of the constant functions (in $\mathrm{L}^2_{\mu_k}(\mathcal{X})$ or $\mathrm{L}^2_{\nu_k}(\mathcal{Y})$). Since $\mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} = \Pi \, \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} \, \Pi$, we have $\left\| T - \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} \right\|_{\mathrm{HS}} \ge \left\| \Pi T \Pi - \mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} \right\|_{\mathrm{HS}}$, and $\Pi T \Pi = \mathsf{S}_{\Phi^{(k)}_{\theta,c}} \mathbf{A}^{(k)} \mathbf{B}^{(k)\top} \mathsf{S}^*_{\Psi^{(k)}_{\theta,c}}$ is of the form in (47). Hence $\eta^2_k(\theta) \le \mathcal{E}_k(h)$. □

Proof of proposition 2. Fix θ. For a task k (training or new), let P_k and P'_k be the projections of proposition 1 in $\mathrm{L}^2_{\mu_k}(\mathcal{X})$ and $\mathrm{L}^2_{\nu_k}(\mathcal{Y})$, and define the positive semi-definite matrices

$$\mathbf{E}_k := \mathsf{S}^*_{\Phi^\star_{c,k}} (I - P_k) \mathsf{S}_{\Phi^\star_{c,k}}, \qquad \mathbf{F}_k := \mathsf{S}^*_{\Psi^\star_{c,k}} (I - P'_k) \mathsf{S}_{\Psi^\star_{c,k}} \in \mathbb{R}^{d^\star \times d^\star},$$

so that $\mathbf{a}^\top \mathbf{E}_k \mathbf{a} = \left\| (I - P_k) \, \mathbf{a}^\top \Phi^\star_{c,k} \right\|^2_{\mathrm{L}^2_{\mu_k}(\mathcal{X})}$ is the residual of the regression onto the learned dictionary of $\mathbf{a}^\top \Phi^\star_{c,k}$, the centered version under μ_k of the shared function $\mathbf{a}^\top \Phi^\star$, and similarly for $\mathbf{F}_k$. Since the centered features $\Phi^{(k)}_{\theta,c}$ are orthogonal to constants, P_k coincides on centered functions with the projection onto $\mathrm{span}\{\Phi_\theta, 1\}$; as $\mathbf{a}^\top \Phi^\star$ and $\mathbf{a}^\top \Phi^\star_{c,k}$ differ by a constant,

$$\mathbf{a}^\top \mathbf{E}_k \mathbf{a} = \mathrm{dist}_{\mathrm{L}^2_{\mu_k}(\mathcal{X})} \big(\mathbf{a}^\top \Phi^\star, \, \mathrm{span}\{\Phi_\theta, 1\} \big)^2,$$

where neither the function nor the subspace depends on k; only the norm does. The same holds for $\mathbf{F}_k$.

Step 1 (training tasks). By (48), $\mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}} = \mathsf{S}_{\Phi^\star_{c,k}} \mathbf{M}_k \mathsf{S}^*_{\Psi^\star_{c,k}}$, so that

$$\left\|(I-P_k)\,\mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}}\right\|^2_{\mathrm{HS}} = \mathrm{Tr}\left(\mathbf{E}_k\,\mathbf{M}_k \mathbf{G}^{(k)}_{\Psi^\star}\mathbf{M}_k^\top\right), \qquad \left\|\mathsf{D}^{(\mathrm{k})}_{\mathrm{Y|X}}(I-P'_k)\right\|^2_{\mathrm{HS}} = \mathrm{Tr}\left(\mathbf{F}_k\,\mathbf{M}_k^\top \mathbf{G}^{(k)}_{\Phi^\star}\mathbf{M}_k\right),$$

and both are at most $\eta_k^2(\theta)$ by lemma 1 and proposition 1(i).

Step 2 (change of task). For every f, the projection of f in $\mathrm{L}^2_{\mu_{\mathrm{k}}}(\mathcal{X})$ belongs to $\mathrm{span}\{\Phi_\theta, 1\}$, so its distance to this space in $L^2_{\mu_{\mathrm{new}}}$ is at most $\sqrt{\beta_\mu}$ times its distance in $\mathrm{L}^2_{\mu_{\mathrm{k}}}(\mathcal{X})$. Hence $\mathbf{E}_{\mathrm{new}} \preceq \beta_\mu \mathbf{E}_k$ for every k, and similarly $\mathbf{F}_{\mathrm{new}} \preceq \beta_\nu \mathbf{F}_k$. Using the definition of λ_Φ and Step 1,

$$\lambda_\Phi\,\mathrm{Tr}(\mathbf{E}_{\mathrm{new}}) \le \frac{1}{K}\sum_k \mathrm{Tr}\left(\mathbf{E}_{\mathrm{new}}\mathbf{M}_k\mathbf{G}^{(k)}_{\Psi^\star}\mathbf{M}_k^\top\right) \le \frac{\beta_\mu}{K}\sum_k \mathrm{Tr}\left(\mathbf{E}_k\mathbf{M}_k\mathbf{G}^{(k)}_{\Psi^\star}\mathbf{M}_k^\top\right) \le \frac{\beta_\mu}{K}\sum_k \eta_k^2(\theta),$$

and similarly $\lambda_\Psi\,\mathrm{Tr}(\mathbf{F}_{\mathrm{new}}) \le \frac{\beta_\nu}{K}\sum_k \eta_k^2(\theta)$.

Step 3 (new task). By proposition 1(i) and lemma 1, $\eta^2_{\mathrm{new}}(\theta) \le \left\|(I-P_{\mathrm{new}})\,\mathsf{D}^{\mathrm{new}}_{\mathrm{Y|X}}\right\|^2_{\mathrm{HS}} + \left\|\mathsf{D}^{\mathrm{new}}_{\mathrm{Y|X}}(I-P'_{\mathrm{new}})\right\|^2_{\mathrm{HS}}$, and, as in Step 1,

$$\left\|(I-P_{\mathrm{new}})\,\mathsf{D}^{\mathrm{new}}_{\mathrm{Y|X}}\right\|^2_{\mathrm{HS}} = \mathrm{Tr}\left(\mathbf{E}_{\mathrm{new}}\mathbf{M}_{\mathrm{new}}\mathbf{G}^{(\mathrm{new})}_{\Psi^\star}\mathbf{M}_{\mathrm{new}}^\top\right) \le \sigma_\Phi\,\mathrm{Tr}(\mathbf{E}_{\mathrm{new}}),$$

and $\left\|\mathsf{D}^{\mathrm{new}}_{\mathrm{Y|X}}(I-P'_{\mathrm{new}})\right\|^2_{\mathrm{HS}} \le \sigma_\Psi\,\mathrm{Tr}(\mathbf{F}_{\mathrm{new}})$. We use Step 2 to conclude.

□

G.3 Proof of theorem 2

We use two tools. The first is the chain rule of Maurer et al. (2016), which separates the complexity of a shared representation from the one of the task-specific functions.

Lemma 2 (Maurer et al., 2016, Theorem 13)**.** *Let $\mathcal{H}$ be a class of maps $\mathcal{Z} \to \mathbb{R}^D$ with $0 \in \mathcal{H}$ and with values in the ball of radius R, and let $\mathcal{G}$ be a class of functions $\mathbb{R}^D \to \mathbb{R}$ with Lipschitz constant at most L on that ball and such that $g(0) = 0$ for all $g \in \mathcal{G}$. For $k \in [K]$, let $\zeta_{k1}, \dots, \zeta_{km}$ be i.i.d. from P_k, independent across k, and write $\bar\zeta := (\zeta_{kl})_{k,l}$,*

$$G\big(\mathcal{H}(\bar\zeta)\big) := \mathbb{E}_\gamma \sup_{h\in\mathcal{H}} \sum_{k,l} \langle \gamma_{kl}, h(\zeta_{kl})\rangle, \qquad Q(\mathcal{G}) := \sup_{\mathbf{y}\neq\mathbf{y}'\in(\mathbb{R}^D)^m} \frac{\mathbb{E}\sup_{g\in\mathcal{G}}\sum_{l=1}^m \gamma_l\big(g(y_l)-g(y'_l)\big)}{\|\mathbf{y}-\mathbf{y}'\|_2},$$

where the γ_{kl} are i.i.d. standard Gaussian vectors of $\mathbb{R}^D$ and the γ_l are i.i.d. standard Gaussian variables. Then there exist absolute constants $c_1, c_2 > 0$ such that

$$\mathbb{E}\sup_{h\in\mathcal{H},\, g_1,\dots,g_K\in\mathcal{G}} \frac{1}{K}\sum_{k=1}^K \left(\mathbb{E}_{\zeta\sim P_k}\left[g_k(h(\zeta))\right] - \frac{1}{m}\sum_{l=1}^m g_k\big(h(\zeta_{kl})\big)\right) \le \frac{c_1 L\,\mathbb{E}\,G\big(\mathcal{H}(\bar\zeta)\big)}{mK} + \frac{c_2\,Q(\mathcal{G})\,\sup_{h\in\mathcal{H}}\big\|h(\bar\zeta)\big\|_2}{m\sqrt{K}},$$

where $\big\|h(\bar\zeta)\big\|_2$ denotes the Euclidean norm of the vector $\big(h(\zeta_{kl})\big)_{k,l}$.

Remark 1. *The assumptions $0 \in \mathcal{H}$ and $g(0) = 0$ make the term $G(\mathcal{G}(y_0))$ of the chain rule of Maurer et al. (2016, Theorem 12) vanish at $y_0 = 0$, which is how lemma 2 is stated. They are not essential (see Remark 1 of Maurer et al., 2016): for general classes one simply adds $\inf_{y_0\in\mathcal{H}(\bar\zeta)} \mathbb{E}\,G(\mathcal{G}(y_0))/(mK)$ to the bound. In the proof of theorem 2 below, $g(0) = 0$ holds for every head, and $0 \in \mathcal{H}$ amounts to assuming that $\Phi_\theta \equiv 0$ and $\Psi_\theta \equiv 0$ for some $\theta \in \Theta$.*

The following elementary lemma collects the two quantities that lemma 2 requires for the bilinear heads of MTL-CMO.

Lemma 3. *Let $r, \rho > 0$ and $B_r := \{\mathbf{a} \in \mathbb{R}^d : \|\mathbf{a}\|_2 \le r\}$. For $\|\mathbf{W}\|_{\mathrm{F}} \le \rho$ set $p_{\mathbf{W}}(\mathbf{a}, \mathbf{b}) := \mathbf{a}^\top \mathbf{W} \mathbf{b}$, so that $|p_{\mathbf{W}}| \le \rho r^2$ on $B_r \times B_r$. On $B_r \times B_r$, the classes*

$$\mathcal{G}_{\mathrm{lin}} := \big\{p_{\mathbf{W}} : \|\mathbf{W}\|_{\mathrm{F}} \le \rho\big\}, \qquad \mathcal{G}_{\mathrm{sq}} := \big\{p_{\mathbf{W}}^2 : \|\mathbf{W}\|_{\mathrm{F}} \le \rho\big\}$$

have Lipschitz constant and Q (as defined in lemma 2) at most $\sqrt{2}\rho r$ and $2\sqrt{2}\rho^2 r^3$, respectively. Moreover, if a class $\mathcal{G}$ on B_r^4 is obtained by averaging two members of such a class over disjoint pairs of coordinates, i.e. $g(\mathbf{a}, \mathbf{b}, \mathbf{a}', \mathbf{b}') = \frac{1}{2}\big(\phi(\mathbf{a}, \mathbf{b}) + \phi(\mathbf{a}', \mathbf{b}')\big)$ with ϕ ranging over the class, then the same bounds hold for $\mathcal{G}$.

Proof. Lipschitz constants. For $\|\mathbf{a}\|_2, \|\mathbf{b}\|_2, \|\tilde{\mathbf{a}}\|_2, \left\|\tilde{\mathbf{b}}\right\|_2 \le r$,

$$|p_{\mathbf{W}}(\mathbf{a}, \mathbf{b}) - p_{\mathbf{W}}(\tilde{\mathbf{a}}, \tilde{\mathbf{b}})| \le |(\mathbf{a}-\tilde{\mathbf{a}})^\top \mathbf{W} \mathbf{b}| + |\tilde{\mathbf{a}}^\top \mathbf{W} (\mathbf{b}-\tilde{\mathbf{b}})| \le \rho r\big(\|\mathbf{a} - \tilde{\mathbf{a}}\|_2 + \left\|\mathbf{b} - \tilde{\mathbf{b}}\right\|_2\big) \le \sqrt{2}\rho r \left\|(\mathbf{a}, \mathbf{b}) - (\tilde{\mathbf{a}}, \tilde{\mathbf{b}})\right\|_2,$$

using $\|\mathbf{W}\|_{\mathrm{op}} \le \|\mathbf{W}\|_{\mathrm{F}}$ and Cauchy–Schwarz. Since $|p_{\mathbf{W}}| \le \rho r^2$, we get $|p_{\mathbf{W}}^2 - p_{\mathbf{W}}^2(\dot{})| \le 2\rho r^2 |p_{\mathbf{W}} - p_{\mathbf{W}}(\dot{})| \le 2\sqrt{2}\rho^2 r^3 \left\|(\mathbf{a}, \mathbf{b}) - (\tilde{\mathbf{a}}, \tilde{\mathbf{b}})\right\|_2$.

The quantities Q. Both classes are linear in a matrix parameter:

$$p_{\mathbf{W}}(\mathbf{a}, \mathbf{b}) = \langle \mathbf{W}, \mathbf{a}\mathbf{b}^\top \rangle, \qquad p_{\mathbf{W}}(\mathbf{a}, \mathbf{b})^2 = \langle \mathbf{W} \otimes \mathbf{W}, (\mathbf{a}\mathbf{b}^\top) \otimes (\mathbf{a}\mathbf{b}^\top) \rangle,$$

with $\|\mathbf{W} \otimes \mathbf{W}\|_{\mathrm{F}} = \|\mathbf{W}\|_{\mathrm{F}}^2 \le \rho^2$. For any matrices $\mathbf{M}_1, \dots, \mathbf{M}_m$ and any $v > 0$, Jensen's inequality gives

$$\mathbb{E} \sup_{\|\mathbf{V}\|_{\mathrm{F}} \le v} \sum_{l=1}^m \gamma_l \langle \mathbf{V}, \mathbf{M}_l \rangle \le v\, \mathbb{E} \|\textstyle\sum_l \gamma_l \mathbf{M}_l\|_{\mathrm{F}} \le v \Big(\sum_{l=1}^m \|\mathbf{M}_l\|_{\mathrm{F}}^2 \Big)^{1/2}. \tag{52}$$

Applying (52) with $\mathbf{M}_l = \mathbf{a}_l \mathbf{b}_l^\top - \tilde{\mathbf{a}}_l \tilde{\mathbf{b}}_l^\top$ and $v = \rho$, and using $\left\|\mathbf{a}\mathbf{b}^\top - \tilde{\mathbf{a}}\tilde{\mathbf{b}}^\top\right\|_{\mathrm{F}} \le r\big(\|\mathbf{a} - \tilde{\mathbf{a}}\|_2 + \left\|\mathbf{b} - \tilde{\mathbf{b}}\right\|_2\big) \le \sqrt{2} r \left\|(\mathbf{a}, \mathbf{b}) - (\tilde{\mathbf{a}}, \tilde{\mathbf{b}})\right\|_2$, gives $Q(\mathcal{G}_{\mathrm{lin}}) \le \sqrt{2}\rho r$. For $\mathcal{G}_{\mathrm{sq}}$, write $\mathbf{u} := \mathbf{a} \otimes \mathbf{b}$, so that $\|\mathbf{u}\|_2 \le r^2$ and $\|\mathbf{u} - \tilde{\mathbf{u}}\|_2 \le \sqrt{2} r \left\|(\mathbf{a}, \mathbf{b}) - (\tilde{\mathbf{a}}, \tilde{\mathbf{b}})\right\|_2$; then

$$\left\|\mathbf{u}\mathbf{u}^\top - \tilde{\mathbf{u}}\tilde{\mathbf{u}}^\top\right\|_{\mathrm{F}} \le \big(\|\mathbf{u}\|_2 + \|\tilde{\mathbf{u}}\|_2\big)\|\mathbf{u} - \tilde{\mathbf{u}}\|_2 \le 2\sqrt{2} r^3 \left\|(\mathbf{a}, \mathbf{b}) - (\tilde{\mathbf{a}}, \tilde{\mathbf{b}})\right\|_2,$$

and (52) with $v = \rho^2$ gives $Q(\mathcal{G}_{\mathrm{sq}}) \le 2\sqrt{2}\rho^2 r^3$.

Averaging two blocks. If ϕ has Lipschitz constant L_0 and the class of such ϕ has $Q \le Q_0$, then, writing $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2)$ for the two blocks, $|g(\mathbf{u}) - g(\tilde{\mathbf{u}})| \le \frac{L_0}{2}(\|\mathbf{u}_1 - \tilde{\mathbf{u}}_1\|_2 + \|\mathbf{u}_2 - \tilde{\mathbf{u}}_2\|_2) \le \frac{L_0}{\sqrt{2}}\|\mathbf{u} - \tilde{\mathbf{u}}\|_2$, and likewise $\mathbb{E}\sup_g \sum_l \gamma_l \big(g(\mathbf{u}_l) - g(\tilde{\mathbf{u}}_l)\big) \le \frac{Q_0}{2}\big(\left\|\mathbf{u}^{(1)} - \tilde{\mathbf{u}}^{(1)}\right\|_2 + \left\|\mathbf{u}^{(2)} - \tilde{\mathbf{u}}^{(2)}\right\|_2\big) \le \frac{Q_0}{\sqrt{2}}\|\mathbf{u} - \tilde{\mathbf{u}}\|_2$, where $\mathbf{u}^{(j)}$ collects the j-th blocks. Both constants therefore only decrease. □

The last tool is Hoeffding's representation of a U-statistic as an average of averages of i.i.d. terms (see, e.g., de la Peña & Giné, 1999, Chapter 4). Let $Z_1, \dots, Z_n$ be i.i.d., let s be a symmetric kernel, and let $\mathbb{S}_n$ be the set of permutations of $[n]$. Then

$$\frac{1}{n(n-1)} \sum_{i \ne j} s\big(Z_i, Z_j\big) = \frac{1}{|\mathbb{S}_n|} \sum_{\pi \in \mathbb{S}_n} \frac{2}{n} \sum_{l=1}^{n/2} s\big(Z_{\pi(2l-1)}, Z_{\pi(2l)}\big), \tag{53}$$

where, for each fixed π, the $n/2$ terms of the inner average are i.i.d.

Proof of theorem 2. Step 1 (reduction to a uniform deviation). Let $\bar{h}$ be a fixed element of $\Theta \times \mathcal{W}_\rho^K$, and set

$$\widehat{\overline{\mathcal{L}}} := \frac{1}{K} \sum_{k=1}^K \widehat{\mathcal{L}}^{(k)}, \qquad \overline{\mathcal{L}} := \frac{1}{K} \sum_{k=1}^K \mathcal{L}^{(k)}, \qquad Z := \sup_{h \in \Theta \times \mathcal{W}_\rho^K} \big(\overline{\mathcal{L}}(h) - \widehat{\overline{\mathcal{L}}}(h)\big).$$

Since $\mathcal{E}_k$ and $\mathcal{L}^{(k)}$ differ by a quantity that does not depend on h, (46) gives

$$\overline{\mathcal{E}}(\widehat{h}) - \overline{\mathcal{E}}(\bar{h}) = \overline{\mathcal{L}}(\widehat{h}) - \overline{\mathcal{L}}(\bar{h}) \le Z + \epsilon_{\mathrm{opt}} + \big(\widehat{\overline{\mathcal{L}}}(\bar{h}) - \overline{\mathcal{L}}(\bar{h})\big).$$

It remains to bound Z and the deviation at the fixed point $\bar{h}$.

Step 2 (Concentration argument). With $q = q_{h_k}$ and $q_{ij} = q\big(x_i^{(k)}, y_j^{(k)}\big)$, the empirical loss of section 3 reads

$$\widehat{\mathcal{L}}^{(k)}(h) = \frac{1}{n(n-1)} \sum_{i \neq j} \big(q_{ij}^2 + 2q_{ij}\big) - \frac{2}{n} \sum_{i=1}^{n} q_{ii}, \tag{54}$$

whose expectation is

$$\mathbb{E}_{\mu_k \otimes \nu_k}[q^2] + 2\mathbb{E}_{\mu_k \otimes \nu_k}[q] - 2\mathbb{E}_{\rho_k}[q] = \|q - g_k\|^2_{L^2(\mu_k \otimes \nu_k)} - \|g_k\|^2_{L^2(\mu_k \otimes \nu_k)} = \mathcal{L}^{(k)}(h),$$

where we used $\mathbb{E}_{\rho_k}[q] = \mathbb{E}_{\mu_k \otimes \nu_k}[p_k q]$. By assumption 1, $|q_{h_k}| \le Q_\rho$ for every h and every (x, y), so replacing one sample of task k modifies at most $2(n-1)$ terms of the U-statistic in (54), each by at most $2(Q_\rho^2 + 2Q_\rho)$, and one diagonal term, by at most $2Q_\rho$; hence it modifies $\widehat{\overline{\mathcal{L}}}(h)$ by at most $B_\rho/(nK)$, uniformly in h. Both Z and $\widehat{\overline{\mathcal{L}}}(\bar{h})$ are therefore functions of the nK independent samples with bounded differences $B_\rho/(nK)$, and McDiarmid's inequality together with a union bound gives that, with probability at least $1 - \delta$,

$$Z \le \mathbb{E}Z + B_\rho \sqrt{\frac{\log(2\delta^{-1})}{2nK}} \qquad \text{and} \qquad \widehat{\overline{\mathcal{L}}}(\bar{h}) - \overline{\mathcal{L}}(\bar{h}) \le B_\rho \sqrt{\frac{\log(2\delta^{-1})}{2nK}}.$$

Step 3 (splitting the empirical process). By (54), $\overline{\mathcal{L}}(h) - \widehat{\overline{\mathcal{L}}}(h) = \Delta_U(h) + 2\Delta_V(h)$, where

$$\Delta_U(h) := \frac{1}{K} \sum_{k=1}^{K} \big(\mathbb{E}U_h^{(k)} - U_h^{(k)}\big), \qquad \Delta_V(h) := \frac{1}{K} \sum_{k=1}^{K} \big(V_h^{(k)} - \mathbb{E}V_h^{(k)}\big),$$

with $V_h^{(k)} := \frac{1}{n} \sum_i q\big(x_i^{(k)}, y_i^{(k)}\big)$ and $U_h^{(k)}$ the U-statistic of task k with symmetric kernel

$$s_h\big((x, y), (x', y')\big) := \tfrac{1}{2}\Big(q(x, y')^2 + 2q(x, y') + q(x', y)^2 + 2q(x', y)\Big).$$

Hence $\mathbb{E}Z \le \mathbb{E}\sup_h \Delta_U(h) + 2\,\mathbb{E}\sup_h \Delta_V(h)$, and we bound the two suprema separately.

Step 4 (the U-part). By (53) applied to each task, $\Delta_U(h)$ is, for every h, an average over the K-tuples of permutations $(\pi_1, \dots, \pi_K)$ of

$$\frac{2}{nK} \sum_{k=1}^{K} \sum_{l=1}^{n/2} \Big(\mathbb{E}\big[s_h(\zeta_{kl}^{\pi_k})\big] - s_h(\zeta_{kl}^{\pi_k})\Big), \qquad \zeta_{kl}^{\pi} := \big(Z_{\pi(2l-1)}^{(k)}, Z_{\pi(2l)}^{(k)}\big).$$

The supremum over h of an average is at most the average of the suprema, and each term of the latter has, by exchangeability within tasks and independence across tasks, the same distribution as the term obtained with the identity permutations. Hence $\mathbb{E}\sup_h \Delta_U(h)$ is at most the expected supremum of the same process over the $nK/2$ independent blocks $\zeta_{kl} := \big(Z_{2l-1}^{(k)}, Z_{2l}^{(k)}\big)$, $l \in [n/2]$, to which we apply lemma 2 with $m = n/2$, the shared representation

$$h_\theta(\zeta) := \big(\Phi_\theta(x), \Psi_\theta(y'), \Phi_\theta(x'), \Psi_\theta(y)\big) \in \mathbb{R}^{4d}, \qquad \zeta = \big((x, y), (x', y')\big),$$

and the task-specific functions

$$g_{\mathbf{W}}(\mathbf{a}, \mathbf{b}, \mathbf{a}', \mathbf{b}') := \tfrac{1}{2}\Big(\varphi(\mathbf{a}^\top \mathbf{W} \mathbf{b}) + \varphi(\mathbf{a}'^\top \mathbf{W} \mathbf{b}')\Big), \qquad \varphi(t) := t^2 + 2t, \qquad \|\mathbf{W}\|_{\mathrm{F}} \le \rho,$$

so that $s_h = g_{\mathbf{W}_k} \circ h_\theta$ with $\mathbf{W}_k = \mathbf{A}^{(k)} \mathbf{B}^{(k)\top}$ and $g_{\mathbf{W}}(0) = 0$. It remains to bound the three quantities entering lemma 2.

(a) Range. By assumption 1, h_θ takes values in the ball of radius $R = 2c\sqrt{d}$, and each of the four blocks of $h_\theta(\zeta)$ lies in the ball of radius $r := c\sqrt{d}$ of $\mathbb{R}^d$. Consequently $\sup_h \big\|h(\bar{\zeta})\big\|_2 \le R\sqrt{nK/2}$.

(b) Lipschitz constant and Q. Since $\varphi(t) = t^2 + 2t$, lemma 3 with $r = c\sqrt{d}$ gives, for the Lipschitz constant of $g_{\mathbf{W}}$ and for $Q(\mathcal{G})$, the common bound

$$2\sqrt{2}\rho^2 c^3 d^{3/2} + 2 \cdot \sqrt{2}\rho c\sqrt{d} = 2\sqrt{2}\,\rho c\sqrt{d}\,(1 + Q_\rho) = 2\sqrt{2}\,L_\rho.$$

(c) Gaussian average. Splitting the Gaussian vector $\gamma_{kl} \in \mathbb{R}^{4d}$ into its four blocks, $G(\mathcal{H}(\bar{\zeta}))$ is at most the sum of four Gaussian averages, two of them over the $nK/2$ inputs x drawn from the μ_k and two over the $nK/2$ outputs y drawn from the ν_k, $n/2$ per task. By the definition (50),

$$\mathbb{E}\, G\big(\mathcal{H}(\bar{\zeta})\big) \le 2 \cdot \frac{nK}{2}\Big(\mathfrak{G}_{n/2,K}(\mathcal{F}_\Phi) + \mathfrak{G}_{n/2,K}(\mathcal{F}_\Psi)\Big).$$

Substituting (a), (b) and (c) into lemma 2 with $m = n/2$ yields

$$\mathbb{E}\sup_h \Delta_U(h) \le 4\sqrt{2}\,c_1 L_\rho\Big(\mathfrak{G}_{n/2,K}(\mathcal{F}_\Phi) + \mathfrak{G}_{n/2,K}(\mathcal{F}_\Psi)\Big) + 8c_2 L_\rho\, c\sqrt{\frac{d}{n}}.$$

Step 5 (the V-part). Here the summands are already i.i.d. within each task, so we apply lemma 2 directly with $m = n$, the representation $h_\theta(x, y) = \big(\Phi_\theta(x), \Psi_\theta(y)\big) \in \mathbb{R}^{2d}$, which by assumption 1 takes values in the ball of radius $R = \sqrt{2}c\sqrt{d}$, and the task functions $g_{\mathbf{W}}(\mathbf{a}, \mathbf{b}) = -\mathbf{a}^\top \mathbf{W}\mathbf{b}$, for which $g_{\mathbf{W}}(0) = 0$ and, by lemma 3 with $r = c\sqrt{d}$, both the Lipschitz constant and Q are at most $\sqrt{2}\rho c\sqrt{d} \le \sqrt{2}L_\rho$. As in Step 4(c), $\mathbb{E}\, G(\mathcal{H}(\bar{\zeta})) \le nK\big(\mathfrak{G}_{n,K}(\mathcal{F}_\Phi) + \mathfrak{G}_{n,K}(\mathcal{F}_\Psi)\big)$ and $\sup_h \big\|h(\bar{\zeta})\big\|_2 \le R\sqrt{nK}$. Since a Gaussian average over a sample of size n per task is at most the one over either half of it, $\mathfrak{G}_{n,K} \le \mathfrak{G}_{n/2,K}$, and we obtain

$$\mathbb{E}\sup_h \Delta_V(h) \le \sqrt{2}\,c_1 L_\rho\Big(\mathfrak{G}_{n/2,K}(\mathcal{F}_\Phi) + \mathfrak{G}_{n/2,K}(\mathcal{F}_\Psi)\Big) + 2c_2 L_\rho\, c\sqrt{\frac{d}{n}}.$$

Step 6 (conclusion). Steps 3–5 give $\mathbb{E}Z \le CL_\rho\big[\mathfrak{G}_{n/2,K}(\mathcal{F}_\Phi) + \mathfrak{G}_{n/2,K}(\mathcal{F}_\Psi) + c\sqrt{d/n}\big]$ for an absolute constant C. Combining with Steps 1 and 2, and taking $\bar{h}$ such that $\overline{\mathcal{E}}(\bar{h})$ is arbitrarily close to $\mathcal{A}_K$ (the remaining arbitrarily small term being absorbed in ϵ_{opt}), we obtain (51).

□