LEARNING DRUG PERTURBATIONS VIA CONDITIONAL MAP ESTIMATORS

Benedek Harsanyi, Marianna Rapsomaniki & Jannis Born
AI for Scientific Discovery
IBM Research Europe
8803, Rüschlikon, Zurich, Switzerland
jab@zurich.ibm.com

ABSTRACT

The growing availability of single-cell perturbation data called for novel methods to capture treatment response. While early attempts employed autoencoders, neural optimal transport (OT) emerged as a more principled alternative because it inherently accommodates the challenges of unpaired data induced by cell destruction during data acquisition. However, neural OT relied on casting the problem to convex regression which induced practical challenges during training. The recently introduced Monge Gap overcomes these challenges through a simple and architecturally agnostic regularizer. While successful, this approach lacks an intrinsic mechanism for generating maps *conditional* on covariates present in perturbation response studies (e.g., dosage, time, drug, or cell type). Here, we extend the Monge Gap and propose `CMonge`, an approach that learns Monge maps conditionally on arbitrary context vectors. It is based on a two-step training procedure combining an autoencoder with a Monge map estimator. We show its value for predicting single-cell perturbation responses, conditional to a drug, a drug dosage, or both. We verify that our conditional models achieve comparable results to the condition-specific state-of-the-art and observe that it particularly excels at capturing higher moments of distributions. Importantly, `CMonge` learns from data aggregated across conditions which exploits cross-task benefits and allows to generalize to unseen conditions with promising performance.

1 INTRODUCTION

Understanding how cells change states in response to different stimuli is a long-standing question in biology, with broad implications across a myriad of diseases. Single-cell RNA sequencing (scRNA-seq) coupled to high-throughput screening (Srivatsan et al., 2020) allow to capture the response of heterogeneous cell populations to thousands of drug perturbations at once. As the cost of these experiments remains high, *in silico* modeling of perturbation responses has emerged as an appealing alternative. Early methods to predict perturbations at a single-cell level relied on perturbation autoencoders and latent space arithmetics (Lotfollahi et al., 2019; 2023; Hetzel et al., 2022). One main challenge of perturbation modeling is the fact that the scRNA-seq data acquisition is destructive to the cells, resulting in unpaired measurements of unperturbed and perturbed cell populations which motivates the use of optimal transport (OT) to model cells as probability distributions. The central question of OT is to find a mapping between a source and a target distribution while minimizing the cost of displacement. Cuturi (2013) proposed to solve the entropic regularized version instead of the traditional Monge formulation. By using Sinkhorn's matrix scaling algorithm, it is possible to compute OT maps efficiently between samples, which motivates to use Wasserstein distance inspired losses in a deep learning setting (Genevay et al., 2018). Recently, neural OT emerged, to parameterize the transportation map via a neural network architecture. By leveraging Brenier's theorem, a unique dual potential exists, which has a gradient equal to the transportation map. This potential can be represented as a convex function, which gave rise to neural solvers based on input convex neural networks (ICNNs) (Amos et al., 2017; Makkuva et al., 2020). Building on these results, Bunne et al. (2023) suggested to model the scRNA-seq perturbation prediction task as a neural OT problem leveraging ICNNs. The method was extended to model perturbations conditioned on a context variable (Bunne et al., 2022). However, one drawback of the dual approach is that Brenier's theorem (1987)

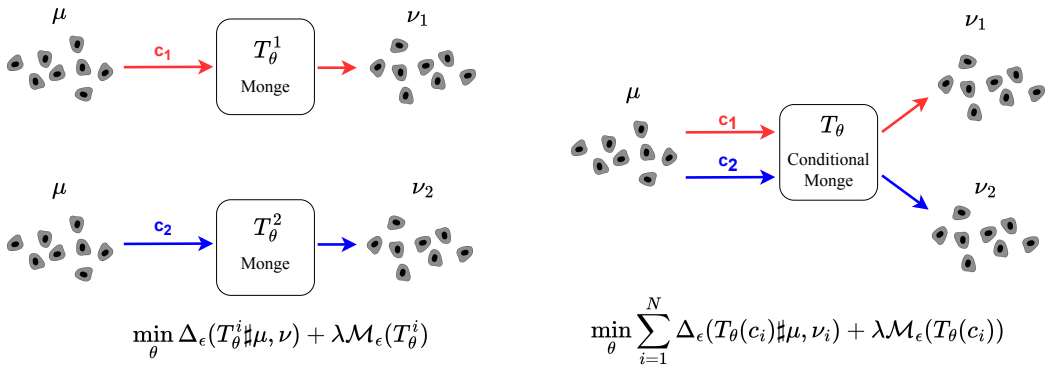

Figure 1: Overview of the CMonge approach. Instead of learning local maps for each perturbation separately (see left), we propose to model perturbation responses via a global estimator that can be conditioned on a potentially unseen context c_i at inference time (see right).

only works when the cost function is the squared Euclidean distance. Moreover, training an ICNN to model the dual potentials poses many architectural and training challenges: a special weight initialization scheme is required to make the initial gradients meaningful, and the non-negativity of the weights exacerbates training. Furthermore, similarly to generative adversarial networks (GANs), the dual training suffers from instability due to its min-max loss function. Recently, flow-based (Tong et al., 2023; Pooladian et al., 2023) or regularization-based approaches (Uscidda & Cuturi, 2023) and even quantum computing techniques (Mariella et al., 2023; Basu et al., 2023) have been proposed to overcome some of these challenges. Among those, the *Monge Gap* (Uscidda & Cuturi, 2023) stands out due to its simplicity: it simply employs a regularizer to estimate OT maps with any ground cost c. The Monge Gap allows to directly parameterize T and optimize the debiased version of the primal objective, the Sinkhorn divergence along with the Monge gap, to ensure c-optimality and fit a mapping between the source and the target distribution. However, the learned maps are local (*i.e.*, unconditional), implying that distinct models are fitted for each perturbation, which has several shortcomings:

(1) Data for each perturbation is needed, so no inference can be made for new perturbations;
(2) The computational cost of training separate models can be significant;
(3) There is no inductive bias to accommodate any covariates present in the data;
(4) Potential cross-task benefits arising from training concurrently on drugs with comparable perturbation effects cannot be exploited.

Here, we propose `CMonge`, an extension the Monge Gap that learns a global map and can be *conditioned* on different context variables or covariates (cf. Figure 1). Instead of leveraging Brenier's theorem, we contextualize through the primal objective and directly learn the transportation maps between the source and different target measures. Our main contribution is a conditional loss function, which enables to model perturbation responses via global estimators. Our proposed `CMonge` model family is tested on different context variables (drug dosage, drug type and combinations thereof) and different data splits and *in* and *out*-of-distribution settings. We experiment with different encodings including fingerprint-based drug representations and effect-driven drug embeddings.

2 Methods

2.1 Background

The Monge formulation of OT seeks an optimal map $T : \mathbb{R}^d \to \mathbb{R}^d$ between probability measures $(\mu, \nu) \in \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d)$, s.t. T pushes forward μ onto ν, while minimizing a displacement cost:

$$T^* := \arg \inf_{T \sharp \mu = \nu} \int_{\mathbb{R}^d} c(x, T(x)) dx. \tag{1}$$

In practice, when measures are data samples $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{n} \sum_{j=1}^n \delta_{y_j}$, the OT problem is solved through the entropic regularized Kantorovich relaxation, which reads:

$$W_\epsilon(\mu, \nu) := \min_{P \in U_n} \langle P, C \rangle + \epsilon H(P), \tag{2}$$

$$U_n = \{P \in \mathbb{R}_+^{n \times n} : P1_n = \frac{1}{n}1_n, P^T 1_n = \frac{1}{n}1_n\} \tag{3}$$

where $H(P) = -\sum_{i,j} P_{ij} \log(P_{ij})$ is the entropy, with $\epsilon > 0$, P describes the amount of mass flowing between the samples and

$$\langle P, C \rangle = \sum_{i,j} P_{i,j} C_{i,j} \tag{4}$$

is the transportation cost $C_{i,j} = c(x_i, y_j)$. We can construct a differentiable loss function, the Sinkhorn divergence, by debiasing the objective, such that $W_\epsilon(\mu, \mu) = 0$ holds with the modification

$$\Delta_\epsilon(\mu, \nu) = W_\epsilon(\mu, \nu) - \frac{1}{2}(W_\epsilon(\mu, \mu) + W_\epsilon(\nu, \nu)). \tag{5}$$

2.2 Conditional OT Estimators

In a machine learning setting, we learn a parametrized map T_θ, by minimizing the Sinkhorn divergence along with the Monge Gap regularizer (Uscidda & Cuturi, 2023) and $\lambda \geq 0$,

$$\min_\theta \Delta_\epsilon(T_\theta \sharp \mu, \nu) + \lambda \mathcal{M}_\epsilon(T_\theta), \tag{6}$$

where the Monge Gap ensures cost optimality w.r.t. the squared Euclidean distance $c(x, y) = ||x - y||^2$ between the source and transported samples and can be estimated from samples, by

$$\mathcal{M}_\epsilon(T_\theta) = \frac{1}{n}\sum_{i=1}^{n} c(x_i, T_\theta(x_i)) - W_\epsilon(\mu, T_\theta \sharp \mu). \tag{7}$$

However, in a conditional setting, multiple target probability measures are labeled with a context c_i (note the difference from the cost c), such that $(c_i, \mu, \nu_i) \in \mathcal{P}(\mathbb{R}^k \times \mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d)$ for $i \in \{1, \dots K\}$. Instead of learning distinct mappings $T_i \sharp(\mu) \approx \nu_i$ we build a global parametrization $T_\theta : \mathbb{R}^k \times \mathbb{R}^d \to \mathbb{R}^d$ that captures information contextually. Our proposed loss extends Equation 6, by simultaneously optimizing for each of the K conditions

$$\min_\theta \sum_{i=1}^{K} \Delta_\epsilon(T_\theta(c_i) \sharp \mu, \nu_i) + \lambda \mathcal{M}_\epsilon(T_\theta(c_i)). \tag{8}$$

3 Experimental Setup

3.1 Dataset

We evaluate our models on the Sci-Plex3 dataset (Srivatsan et al., 2020) which contains single-cell profiles from three human cancer cell lines (namely A549, K562 , and MCF7) that were exposed to a total of 188 compounds, each one administered at four different doses (10 nM, 100 nM, 1000 nM, and 10000 nM). We test our models on 9 different drugs. Different preprocessing steps (library size normalization, cell and gene filtering, and `log1p` transformation) were inherited from Lotfollahi et al. (2019). The dataset consists of $762,039$ single-cell measurements, out of which $17,565$ belong to the control population, and on average $4,032$ observations to each drug and drug dosage condition. We only consider the 1000 highly-variable genes (HVG), computed on the training set only.

The 4i dataset contains cellular and nuclear measurements of $97,748$ cells ($10,995$ controls) of two lines of melanoma tumors treated with one of 35 cancer therapies each involving $\sim 2,500$ cells. We inherit the preprocessing from Bunne et al. (2022), resulting in 48 features.

3.2 Architecture

Training consists of a two-step procedure. First, we follow Bunne et al. (2023): instead of learning in the gene expression space, we encode the data into a 50-dimensional latent space by training a vanilla autoencoder with an encoder $E_\phi : \mathbb{R}^{1000} \to \mathbb{R}^{50}$ and a decoder $D_\theta : \mathbb{R}^{50} \to \mathbb{R}^{1000}$. Both E_ϕ and D_θ are parameterized by multi-layer perceptrons (MLP). The objective of the network is to minimize the reconstruction loss. Secondly, using the same training set, an OT map is learned between the

encoded unperturbed and perturbed cells by optimizing Equation 8. The map is parameterized by $v_i = T_\varphi(c_i, z_i)$, where $z_i \in \mathbb{R}^{50}$ is the encoded gene expression, and c_i is the context (either a scalar, a sparse or dense vector). Throughout this paper, we will refer to T_φ as the Monge network. Predictions are made in the cell space by shifting with the learned perturbation and decoding the result $D_\theta(E_\phi(x_i) + v_i)$. The weights of the autoencoder are frozen during this phase. We next describe how to encode the contextual information and how to incorporate it into the architecture.

3.3 Encoding the condition

We test different settings for encoding the conditions. Firstly, we only condition on dosage and attempt to predict perturbation responses per drug. We encode the dosage with the transformation dose $\to \log(\text{dose})$. Secondly, we attempt to train a single model capable of making inferences on any observation, conditioned on both the specific drug and dosage. We consider two approaches for encoding drug information: (i) `CMonge-RDKit`, a fingerprint-based molecular representation, where 194 RDKit features such as atom type, number of bonds, formal charge, atom mass or number of hydrogen atoms, along one-hot-encoded bond features (for full list see Yang et al. (2019)) are extracted from the SMILES representation of the underlying drug, following Lotfollahi et al. (2023), and (ii) `CMonge-MoA`, a data-driven approach, where multidimensional scaling (MDS) embeddings were generated by calculating pairwise Wasserstein distances between individual target populations, following Bunne et al. (2022). This approach ensures that perturbations with similar Modes-of-Action (MoA) effects are accurately represented in close proximity within the embedding. We calculate a 10-dimensional MDS embedding by employing the majorization algorithm `SMACOF` (De Leeuw, 2005) to minimize stress. Following Lotfollahi et al. (2023), we consider an initial drug embedding $h_i \in \mathbb{R}^m$, and the transformed dosage $s_i \in \mathbb{R}$. These vectors are passed to a drug embedder $W_{\text{drug}} : \mathbb{R}^m \to \mathbb{R}^{50}$, and drug-dosage embedder $W_{\text{dose}} : \mathbb{R}^{m+1} \to \mathbb{R}$ networks, which uses the drug and dosage information as well. We get the final embedding by concatenating

$$W_{\text{drug}}(h_i) = z_i^{\text{drug}}, \quad W_{\text{dose}}(h_i, s_i) = z_i^{\text{dose}}, \quad c_i = (z_i^{\text{drug}}, z_i^{\text{dose}}). \tag{9}$$

Our context $c_i \in \mathbb{R}^{50} \times \mathbb{R}$, along with the unperturbed single-cell observation (c_i, z_i) is passed to the Monge network, which is an MLP, with GELU (Hendrycks & Gimpel, 2016) activation functions.

3.4 Evaluation settings

We test our methods in a homogeneous setting, *i.e.*, data among all conditions is aggregated. We use a 80/20 train/test split for each condition. In an OOD setting, we test the model's ability to generalize to unseen dosages. As many genes are left unaffected, instead of evaluating in the 1000-dimensional gene space, we restrict ourselves to only the top 50 marker genes obtained through gene ranking (Wolf et al., 2018). We evaluate with the R^2 metric between the perturbed and predicted feature means and the Wasserstein distance (Equation 2) and report the mean across batches of observations sampled from the test set. The experiments used the same hyperparameters (for more details, see Appendix A.2.1). We train and evaluate models in the following scenarios:

1. `Monge`: As a hypothetical upper bound on performance, we fit separate Monge Gap models, one per drug-dosage pair. These models do not have any context (cf. Uscidda & Cuturi (2023)).
2. `Monge Homo`: To motivate the contextual settings, we fit a homogeneous model for each drug, using data from all four dosages, but discarding the dosage information.
3. `CMonge`: We fit conditional models for each drug with the scalar dose as context. We leave out different dosages during training, thus creating interpolation and extrapolation settings.
4. `CMonge Homo`: We fit conditional models for each drug, where the context is a scalar representing the dose, using a homogeneous split, where each dosage is present in the training set.
5. `CMonge RDKit / MoA`: A single model fitted to all data, combining drug and dosage context. To encode the drug, we compare fingerprints (`RDKit`) to a data-driven approach (`MoA`).

4 Experimental results

In our experiments, we seek to confirm that adding contextual information enhances the performance of inferring single-cell perturbation responses. Among the zoo of perturbation modeling techniques, we first verified our choice of extending the Monge Gap by comparing it against an

ICNN and other methods in an unconditional setting on the 4i data. This experiment revealed superiority of the Monge Gap (cf. Appendix Figure A1 and Table A1). We then trained and evaluated our novel conditional Monge map technique on the SciPlex data, as summarized in Table 1 and Figure 2. We compare the conditional models against unconditional Monge models and the Identity

Table 1: Evaluation of drug effect perturbations, treated with 9 different drugs. Results are compared based on the Correlation Coefficient between the predicted and target feature means (R^2). The average and standard deviation are reported of the 9 experiments per model. Results averaged across nine drugs.

Model	Split	Context		Dosage (nM)			
		Drug	Dose	10	100	1000	10000
Identity				0.747 ± 0.126	0.654 ± 0.260	0.503 ± 0.332	0.227 ± 0.212
Monge	10			0.950 ± 0.019	–	–	–
Monge	100			–	0.934 ± 0.041	–	–
Monge	1000			–	–	**0.960 ± 0.024**	–
Monge	10000			–	–	–	0.977 ± 0.028
Monge	Homo			0.749 ± 0.253	0.767 ± 0.231	0.885 ± 0.098	0.694 ± 0.271
CMonge	10-ood		✓	0.863 ± 0.197	0.944 ± 0.040	0.953 ± 0.039	0.987 ± 0.008
CMonge	100-ood		✓	0.931 ± 0.069	0.931 ± 0.058	0.940 ± 0.052	0.987 ± 0.008
CMonge	1000-ood		✓	**0.960 ± 0.029**	0.958 ± 0.024	0.877 ± 0.118	**0.989 ± 0.007**
CMonge	10000-ood		✓	0.940 ± 0.075	**0.962 ± 0.034**	0.940 ± 0.052	0.526 ± 0.311
CMonge	Homo		✓	0.882 ± 0.185	0.905 ± 0.123	0.904 ± 0.092	0.973 ± 0.026
CM-RDKit	Homo	✓	✓	0.619 ± 0.268	0.690 ± 0.172	0.868 ± 0.056	0.352 ± 0.320
CM-MoA	Homo	✓	✓	0.912 ± 0.039	0.912 ± 0.046	0.902 ± 0.048	0.938 ± 0.079

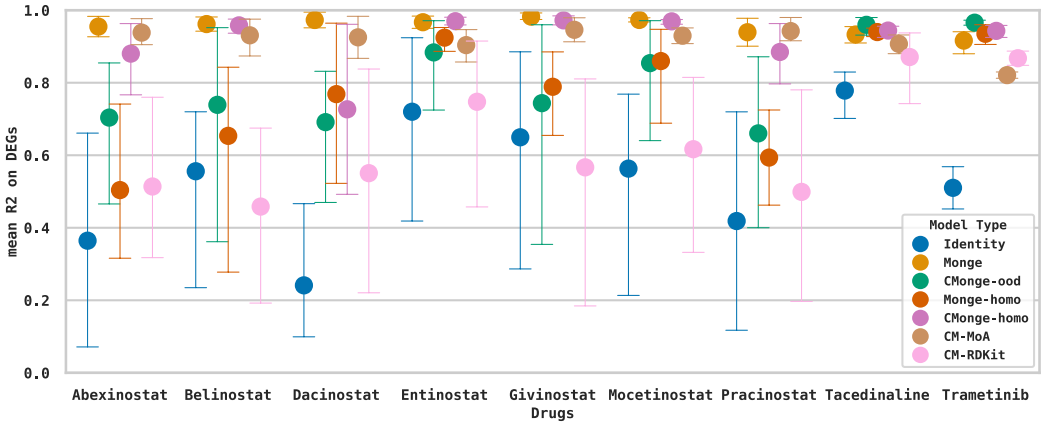

Figure 2: Comparison of the different conditional and unconditional Monge methods, grouped by drug, using the mean R^2 metric. Each point represents the mean performance of the model, out of the four dosages, along with uncertainty around that estimate using error bars.

mapping. In Table 1, the Identity shows that the average drug effect is weak on the DEGs over the lower dosages, while dosage 10000 nM, has significantly stronger effect, making it harder to learn the perturbation responses. The dosage-specific Monge models are highly performant but cannot be applied to other dosages. The homogeneous model (Monge homo) that did not receive any contextual information fails to generalize among different dosages, the model learned the average effect as a response. This motivates us to introduce the Conditional counterpart, CMonge-Homo, which outperforms its homogeneous baseline. Importantly, CMonge-Homo even achieves similar results compared to the upper bound (*i.e.*, the specific Monge models), for example in the 10000nM dosage case. Notably, the inclusion of extra context even enhances performance; the dosage-specific Monge models for 10, 100 and 10000nM are outperformed by CMonge models that were additionally trained on other dosages. This points to strong cross-task learning benefits and suggests

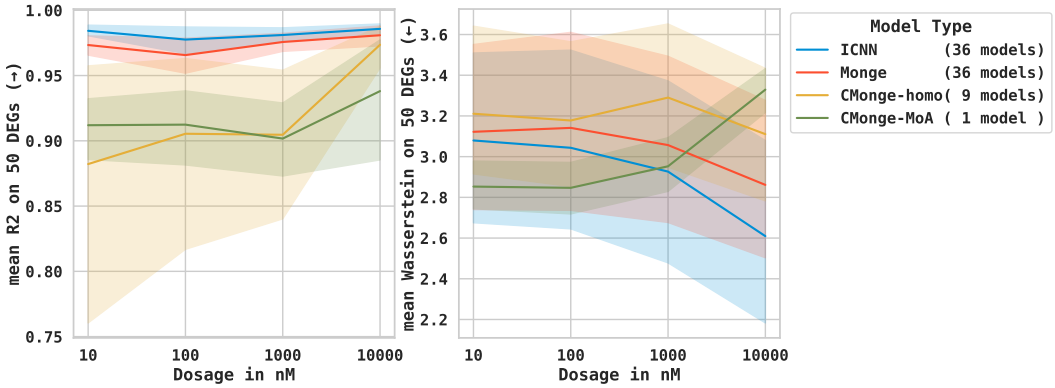

Figure 3: Comparison between the unconditional OT-based models, and the conditional counterparts in the homogeneous setting based on the R2 of feature means (left) and the Wasserstein distance (right).

that conditional Monge models are generally preferable to unconditional models. However, when $10,000$ nM is not included in the training set, the models experience a substantial performance drop, but still outperform the identity prediction. The difficulty of learning to generalize to the highest dosage is, beyond the lower performance of the identity model on this condition, also evidenced by the UMAPs in Figure A4. For all nine drugs, cells from the three cell types form distinct clusters and for a majority of drugs the cells belonging to the $10,000$nM condition sit on the edge of the cell-type-clusters, sometimes even forming their own clusters.

We next sought to investigate the even more challenging setting of learning a global map contextualized on both drug and dosage. Table 1 shows that the 194 dimensional fingerprint seems to introduce too much noise, compared to the 50 dimensional data signal, resulting in poor performance across dosages. On the other hand, the CMonge-MoA model demonstrates strong generalization performance across varying dosages as well as drugs. It outperforms the drug-specific Monge models by a wide margin and is largely better than CMonge-Homo, suggesting cross-task benefits among the nine drugs. In Figure 2 it can be seen that the performance of the fingerprint varies highly among drugs whereas the MoA embeddings consistently yield performance very close to the upper bound (i.e., drug and dosage specific Monge models). However, the MoA, unlike the RDKit features requires availability of a small population of perturbed cells to compute the embedding.

But can we actually replace condition-specific models with a single conditional one? To assess this ambitious question, Figure 3 aggregates performance across drugs and compares the 36 individual ICNN and Monge Gap models to the nine, drug-specific CMonge-Homo models as well as the single CMonge-MoA model. Albeit the unfair comparison of a single model to 36 individual and unconditional models of identical size, the CMonge-MoA achieves a good, yet overall inferior R2 in capturing DEG feature means. It seems that the CMonge models were predominantly driven by the highest dosage (cf. Figure 3) which induced the strongest perturbation effect (cf. Figure A4), but could potentially be mitigated with more careful training. Importantly, however, the CMonge-MoA model captures *better* the higher moments of the distribution than the 36 condition-specific models (as measured by the Wasserstein distance, Figure 3 *right*). This is a critical finding that underlines the advantages of our method and is further supported by the numerical performances (Table A3) and the barplot (cf. Figure A5) of the Wasserstein distance.

5 Conclusion

In this paper, we proposed `CMonge`, a novel approach to learn Monge maps conditionally. Instead of contextualizing the Kantorovich dual, we relied on the debiased version of the primal OT loss by extending the Monge Gap. We illustrated our approach on single-cell perturbation response prediction and showcased that our global model can infer cell states across multiple drugs and dosages and even extrapolate to unseen dosages. Our results were especially encouraging, when considering effect-driven embeddings, compared to molecular signatures. Future work could investigate the impact of larger or more expressive architectures than the three-layer MLP utilized in the Monge Gap, e.g., by integrating it with the emerging single-cell foundation models (Cui et al., 2024; Yang et al., 2022).

CODE AVAILABILITY

The source code for reproducing the experiments is available at: https://github.com/AI4SCR/Conditional-Monge.

ACKNOWLEDGEMENTS

We thank Juan Gonzalez-Espitia and Alice Driessen for helpful discussions.

REFERENCES

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pp. 146–155. PMLR, 2017.

Saugata Basu, Jannis Born, Aritra Bose, Sara Capponi, Dimitra Chalkia, Timothy A Chan, Hakan Doga, Mark Goldsmith, Tanvi Gujarati, Aldo Guzman-Saenz, et al. Towards quantum-enabled cell-centric therapeutics. *arXiv preprint arXiv:2307.05734*, 2023.

Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305:805–808, 1987.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional monge maps. *Advances in Neural Information Processing Systems*, 35:6859–6872, 2022.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pp. 1–11, 2024.

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

Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

Jan De Leeuw. Applications of convex analysis to multidimensional scaling. 2005.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Fabian Theis, et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *Advances in Neural Information Processing Systems*, 35:26711–26722, 2022.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, pp. e11517, 2023.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pp. 6672–6681. PMLR, 2020.

Nicola Mariella, Jannis Born, Albert Akhriev, Francesco Tacchino, Christa Zoufal, Eugene Koskin, Ivano Tavernelli, Stefan Woerner, Marianna Rapsomaniki, and Sergiy Zhuk. Quantum theory and application of contextual optimal transport. In *NeurIPS 2023 Workshop Optimal Transport and Machine Learning*, 2023. URL https://openreview.net/forum?id=DW3a9czPGx.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky Chen. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2304.14772*, 2023.

Sanjay R Srivatsan, José L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

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

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

A Appendix

A.1 Benchmarking

We first benchmarked different state of the art methods for unconditional perturbation modeling.

4i dataset We trained each method on each of the 35 therapies with a 80/20 train/validation split. We note that the original scGen (Lotfollahi et al., 2019) relies on a Variational formulation (Kingma & Welling, 2014) formulation. Instead, in our experiments, we follow the setup in Bunne et al. (2022), i.e., we utilized a vanilla AE, and both the encoder and the decoder are parametrized with fully-connected layers. The results in Table A1 and Figure A1 confirm the finding by Uscidda & Cuturi (2023), i.e., the Monge Gap achieves the overall best result with respect to each of the evaluation metrics and also shows lower standard deviation. On Figure A1, the subplot on the Wasserstein distance clearly shows that the Monge model (which directly optimizes this metric) performs consistently, regardless of the perturbation, while the autoencoder is struggling to capture perturbation effects in some cases. On the other hand, the ICNN results are only skewed by one outlier. Remarkably, the optimal transport-based models are outperforming the autoencoder on MMD, R^2, and Drug Signatures, even though they are trained to optimize the primal and dual OT loss.

Table A1: Evaluation of perturbation prediction on the 4i dataset. Average performance is reported over the 35 treatments along with the standard deviation in the 48 dimensional feature space.

Model	Wasserstein	MMD	R^2	Drug Signature	Sinkhorn Div
Monge	$\mathbf{2.345 \pm 0.219}$	$\mathbf{0.009 \pm 0.001}$	$\mathbf{0.901 \pm 0.087}$	$\mathbf{0.333 \pm 0.067}$	$\mathbf{1.812 \pm 0.221}$
AE	2.455 ± 0.396	0.024 ± 0.014	0.804 ± 0.110	0.523 ± 0.150	1.921 ± 0.398
ICNN	2.632 ± 0.512	0.009 ± 0.007	0.878 ± 0.104	0.433 ± 0.406	2.097 ± 0.515
Identity	2.881 ± 0.736	0.033 ± 0.027	0.309 ± 0.209	1.285 ± 0.891	2.347 ± 0.738

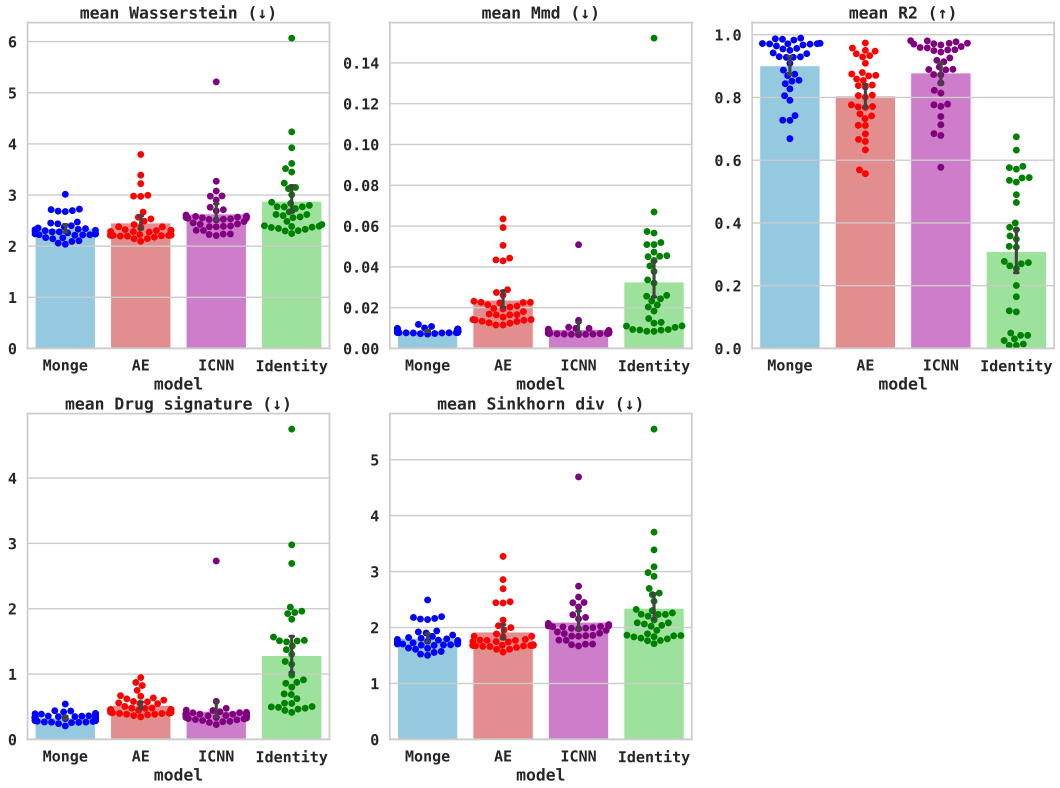

Figure A1: Evaluation of perturbation prediction on the 4i dataset. Each point corresponds to a model trained on one out of the 35 treatments.

SciPlex For each of the nine drug and each of the four dosages we fitted a different model, resulting in 36 models per method. We included the identity mapping as a baseline, which simply predicts the unperturbed cell states. Table A2 and Figure A2 shows the performance of the different methods. Our main evaluation metric is still the correlation coefficient of the HVG feature means (R^2), but we also report the entropy-regularized Wasserstein distance, which is closely related to the objective function of ICNN and Monge models. The ICNN and Monge models used the 50 dimensional latent representation learned by the autoencoder.

All model predictions were decoded and evaluated in the cell space on the 50 HVGs. We observe, that the neural optimal transport based solvers significantly outperforms the autoencoder based approach.

The results of Figure A3 shows that, although the ICNN based solver slightly outperforms the Monge based counterpart, their results are highly correlated, they have almost identical performance with respect to the Wasserstein distance. Combining this with the fact that we are expanding upon the Monge based methodology, we did not include ICNN based benchmarks in Table 1.

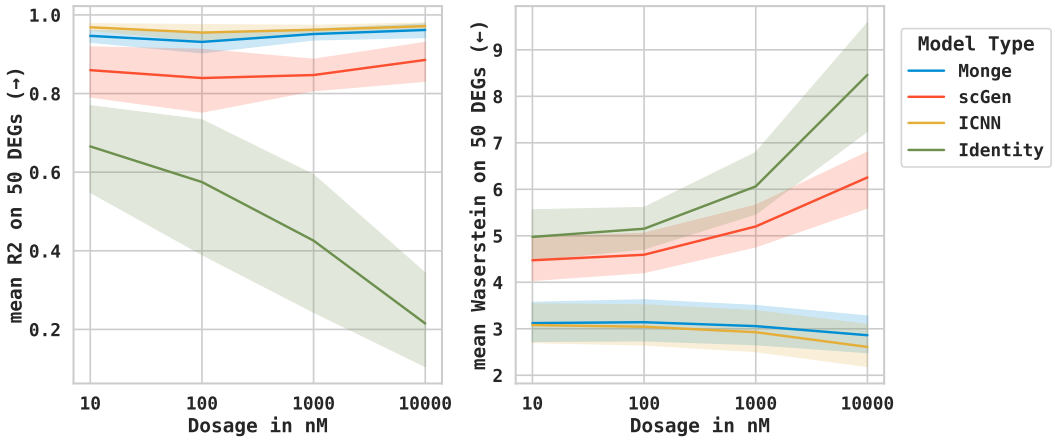

Figure A2: Performance of the benachmarked models for the 4 different dosages, averaged over 9 drugs.

Table A2: Evaluation of drug effect perturbations, treated with 9 different drugs. Results are compared based on the Correlation Coefficient between the predicted and target feature means (R^2). The average and standard deviation are reported of the 9 experiments per model.

Model	Dosage (nM)			
	10	100	1000	10000
Monge	0.947 ± 0.024	0.931 ± 0.043	0.951 ± 0.022	0.962 ± 0.027
scGen	0.859 ± 0.106	0.839 ± 0.134	0.847 ± 0.067	0.886 ± 0.079
ICNN	0.969 ± 0.013	0.955 ± 0.034	0.962 ± 0.017	0.972 ± 0.012
Identity	0.666 ± 0.177	0.575 ± 0.276	0.426 ± 0.288	0.215 ± 0.189

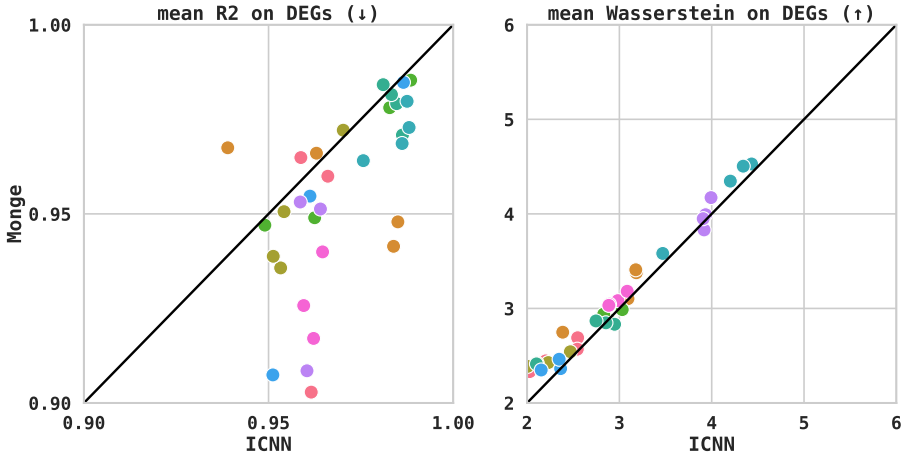

Figure A3: Comparison of neural optimal transport solvers, the scatter plot consists of (x_i, y_i) points, where x_i represents the target metric (R^2 or Wasserstein) obtained by the ICNN solver, and y_i is the performance of the corresponding Monge model on the same drug-dose split. Each drug is denoted with a different color. In case of the R^2 metric, each time an (x_i, y_i) point is under the $x = y$ line, the ICNN outperforms the Monge model.

A.2 Data and experimental details

A.2.1 Hyperparameters of the conditional experiments

In all experiments, we use the AdamW optimizer (Loshchilov & Hutter, 2017), with initial learning rate 10^{-4} and weight decay regularization 10^{-5}. Both the encoder and decoder consist of two hidden layers of 512 dimensions each. The 50-dimensional latent representation is learned through 50 epochs with a batch size of 256. We trained separate autoencoders for the different train-test splits. The Monge network is built out of 4 hidden layers with 64 neurons each. The dose and drug embedders are parameterized with one dense layer. The Euclidean distance is used as displacement cost and the Monge Gap regularizer is set to $\lambda = 10^{-2}$. During the OT training phase, we repeatedly sample a batch 256 observations from the source and target distributions for 1000 iterations for local models (without condition, or conditioned on dosage), while 10000 iterations for the global models (RDKit and MoA). Each batch only contains samples from one context, which is uniformly sampled. All models are implemented using the `OTT-JAX` package (Cuturi et al., 2022).

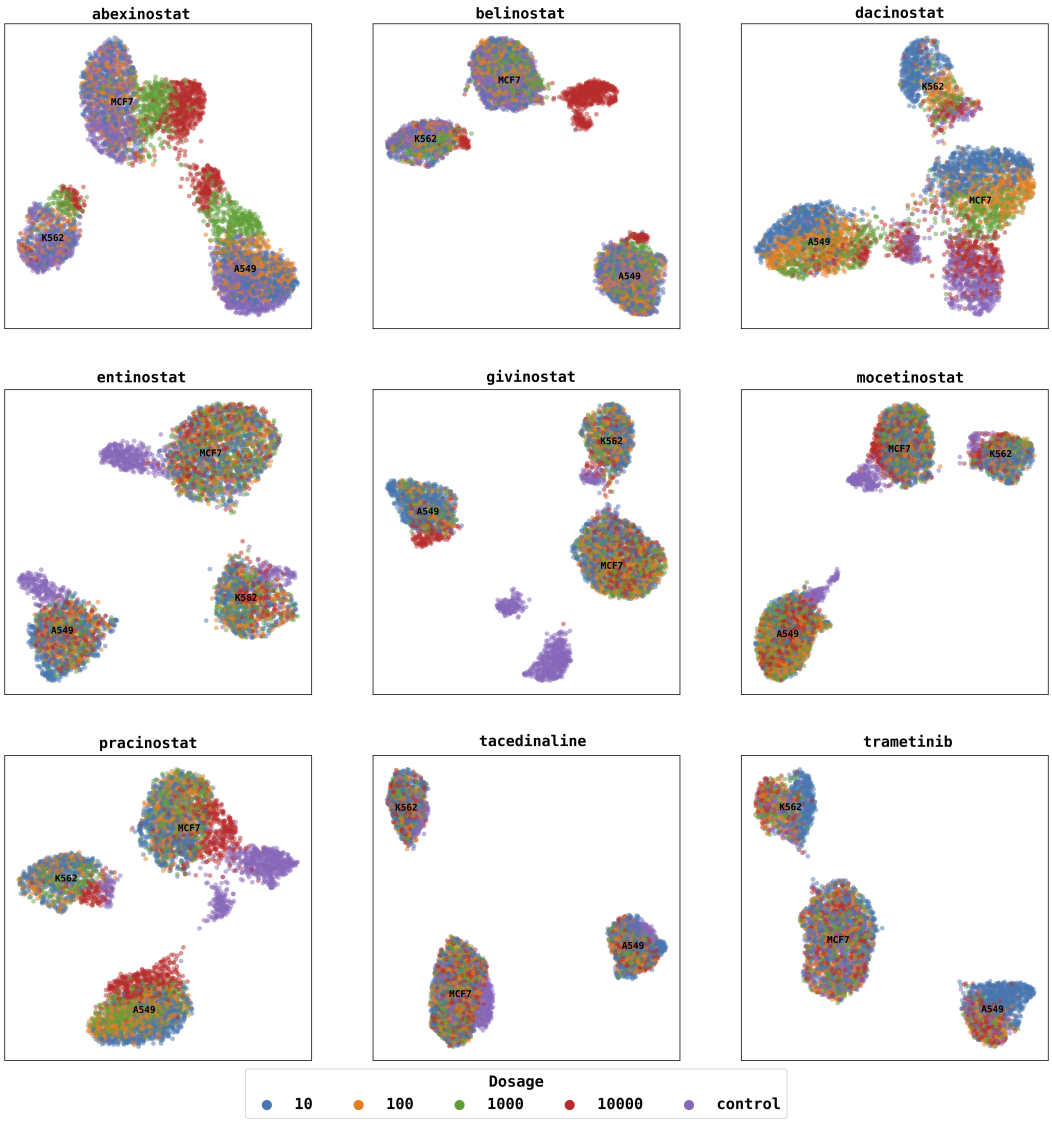

Figure A4: UMAP projection of the 1000-dimensional feature space, filtered on control cells and cells treated with different dosages. We can observe a greater perturbation effect with higher dosage. Moreover, the three clusters are associated with the three different cell types (see black text).

A.3 Additional results on Conditional Monge

Table A3: Evaluation of drug effect perturbations, treated with 9 different drugs. Results are compared based on the **Wasserstein distance** between the predicted and target samples. The average and standard deviation are reported for the 9 experiments per model.

Model	Split	Context		Dosage (nM)			
		Drug	Dose	10	100	1000	10000
Identity				3.444 ± 0.810	3.511 ± 0.683	4.102 ± 1.000	6.399 ± 1.720
Monge	10			3.120 ± 0.692	–	–	–
Monge	100			–	3.162 ± 0.709	–	–
Monge	1000			–	–	3.163 ± 0.603	–
Monge	10000			–	–	–	3.199 ± 0.465
Monge	homo			3.326 ± 0.531	3.291 ± 0.511	3.175 ± 0.516	4.129 ± 0.981
CMonge	10-ood		✓	3.255 ± 0.592	3.182 ± 0.595	3.125 ± 0.602	2.916 ± 0.545
CMonge	100-ood		✓	3.169 ± 0.622	3.148 ± 0.599	3.108 ± 0.603	$\mathbf{2.906 \pm 0.552}$
CMonge	1000-ood		✓	3.108 ± 0.636	3.101 ± 0.618	3.197 ± 0.543	2.906 ± 0.553
CMonge	10000-ood		✓	3.046 ± 0.611	3.018 ± 0.601	3.059 ± 0.576	4.122 ± 0.745
CMonge	homo		✓	3.210 ± 0.568	3.177 ± 0.585	3.290 ± 0.563	3.110 ± 0.544
CM-RDKit	homo	✓	✓	3.229 ± 0.291	3.143 ± 0.181	3.057 ± 0.238	5.170 ± 1.639
CM-MoA	homo	✓	✓	$\mathbf{2.852 \pm 0.193}$	$\mathbf{2.846 \pm 0.208}$	$\mathbf{2.952 \pm 0.213}$	3.329 ± 0.182

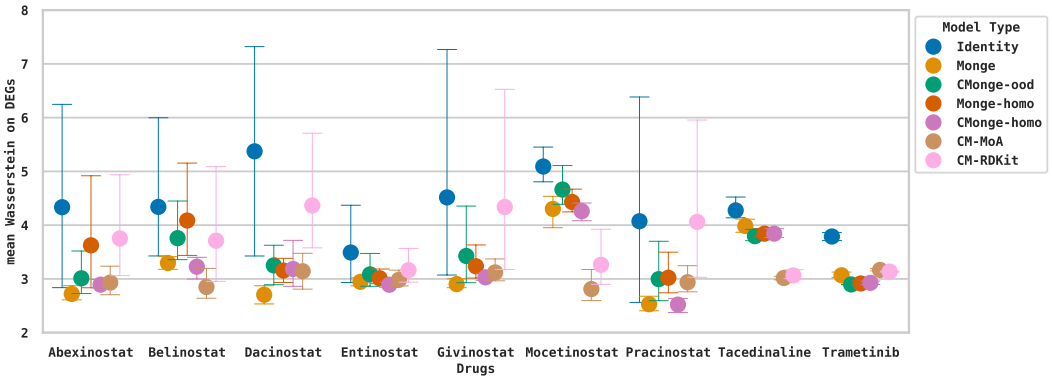

Figure A5: Comparison of the different conditional and unconditional Monge methods, grouped by drug, using the **Wasserstein distance**. Each point represents the mean performance of the model, out of the four dosages, along with uncertainty around that estimate using error bars.

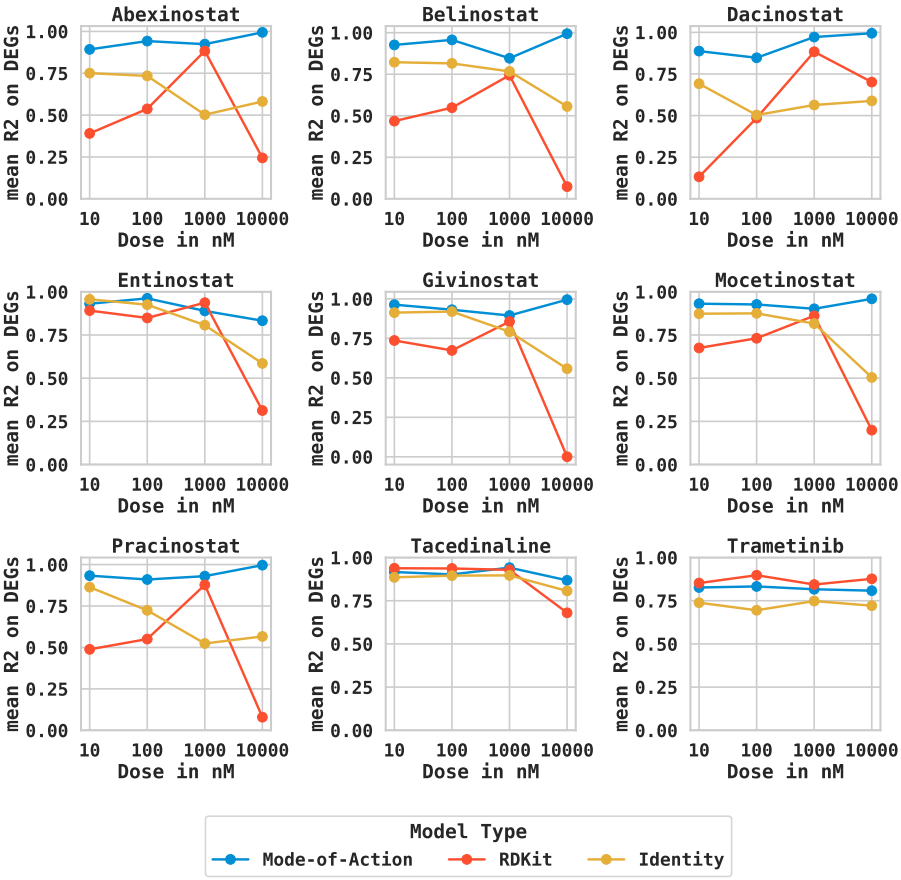

Figure A6: Drug-wise comparison between identity model and the two global conditional models, where we condition based on drug and dosage as well.