The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Semi-Supervised Biomedical Translation with Cycle Wasserstein Regression GANs

Matthew B. A. McDermott
MIT, Cambridge, MA
mmd@mit.edu

Tom Yan
MIT, Cambridge, MA
tyyan@mit.edu

Tristan Naumann
MIT, Cambridge, MA
tjn@mit.edu

Nathan Hunt
MIT, Cambridge, MA
nhunt@mit.edu

Harini Suresh
MIT, Cambridge, MA
hsuresh@mit.edu

Peter Szolovits
MIT, Cambridge, MA
psz@mit.edu

Marzyeh Ghassemi
MIT, Verily,
Cambridge, MA
mghassem@mit.edu

Abstract

The biomedical field offers many learning tasks that share unique challenges: large amounts of unpaired data, and a high cost to generate labels. In this work, we develop a method to address these issues with semi-supervised learning in regression tasks (e.g., translation from source to target). Our model uses adversarial signals to learn from unpaired datapoints, and imposes a cycle-loss reconstruction error penalty to regularize mappings in either direction against one another. We first evaluate our method on synthetic experiments, demonstrating two primary advantages of the system: 1) distribution matching via the adversarial loss and 2) regularization towards invertible mappings via the cycle loss. We then show a regularization effect and improved performance when paired data is supplemented by additional unpaired data on two real biomedical regression tasks: estimating the physiological effect of medical treatments, and extrapolating gene expression (transcriptomics) signals. Our proposed technique is a promising initial step towards more robust use of adversarial signals in semi-supervised regression, and could be useful for other tasks (e.g., causal inference or modality translation) in the biomedical field.

Introduction

Motivation Relative to other fields which have seen recent interest in multi-modal translation (e.g., images to text, or audio to video), the biomedical field lacks large datasets that are "paired" — where two sets of data from the same subject are available (e.g., at different times or in different modalities). Additionally, many biomedical translation tasks involve regressions between source and target domains that are either 1) both representations of some shared, underlying state (e.g. modality translation), or 2) driven by real, bio-physical mechanisms. In either case, we expect both directions of translation to have meaningful, approximately invertible solutions. This makes "cycle" consistency — mapping a particular source example to the target domain and back — desirable. It also means we can leverage the inferred cycle map to reframe the previously independent regression problems as a joint, multi-task learning problem.

Challenge Learning from "extra" unpaired data is valuable in settings where acquiring a large amount of paired data is not feasible. In many clinical settings, paired dataset collection (e.g., patient data pre- and post-treatment) is impossible, as doctors have an ethical imperative to intervene on patients at times inconvenient for dataset curation. In gene expression tasks, obtaining expression levels for transcripts corresponding to the entire genome is expensive, but there are large corpora of smaller snippets, independently measured for the purposes of individual studies.

Goal Our goal is a mechanism of semi-supervised learning for regression problems that 1) leverages large amounts of unpaired data to improve performance on tasks with a scarcity of paired data, and 2) provides an approximately invertible solution from source to target domains. For example, in estimating medical treatment effect, a patient may have data from only a pre- or post- treatment rather than both. In gene expression tasks, the inherent intracorrelations within gene expression profiles implies that we should be able to translate between different subsets of the transcriptome — the set of all total gene transcription products in a cell — in an invertible, non-lossy manner (Eisen et al. 1998).

Solution We design a novel joint regression-adversarial model (CWR-GAN) that uses cycle-consistent generative adversarial networks (GANs) for translation tasks. We demonstrate our method on synthetic datasets to illustrate its key points and analyze its effects on two real-world biomedical datasets: individual treatment effect (ITE) regression based on electronic health record (EHR) data, and transcriptomics (gene expression) extrapolation.

Contributions We develop an end-to-end differentiable architecture that uses adversarial signals for semi-supervised bi-directional translation in the biomedical field. In doing so, we make the following specific contributions:

- We design a cycle-consistent regression adversarial network for semi-supervised regression learning.
- We demonstrate the regularization effect and discriminative performance boost of our method on synthetic data in a semi-supervised setting.
- We evaluate our approach in two diverse real-world biomedical datasets: 1) forecasting the individual treat-

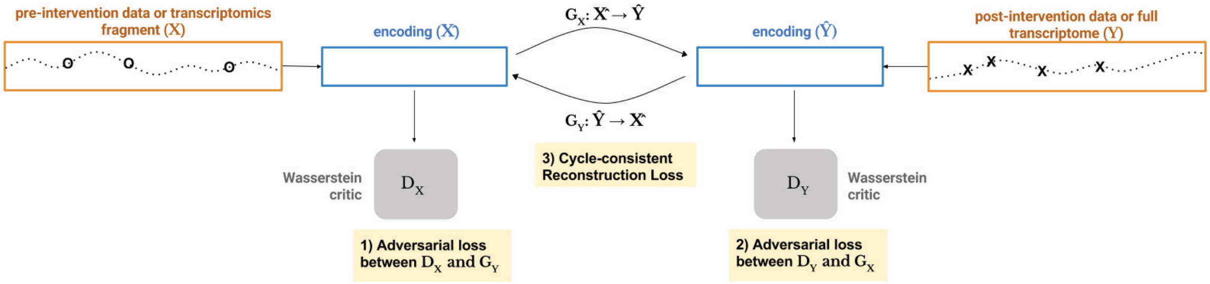
pre-intervention data or transcriptomics fragment (X)
encoding ($\hat{X}$)
G_X: $\hat{X} \rightarrow \hat{Y}$
encoding ($\hat{Y}$)
post-intervention data or full transcriptome (Y)
G_Y: $\hat{Y} \rightarrow \hat{X}$
Wasserstein critic
D_X
3) Cycle-consistent Reconstruction Loss
D_Y
Wasserstein critic
1) Adversarial loss between D_X and G_Y
2) Adversarial loss between D_Y and G_X

Figure 1: Overall architecture for the Cycle Wasserstein Regression GAN (CWR-GAN) model.

ment effect of four ICU interventions on 29 signals in over 2,000 patients, and 2) extrapolating a 978 dimensional subset of the transcriptome to a 5000 dimensional subset.

Related works

The biomedical field is not alone in that bi-directional, approximately invertible mappings (i.e. translation tasks) are desirable. For example, stacked autoencoders have been used to learn a shared representation between audio and video signals, and multi-modal conditional prediction frameworks have been used to "hallucinate" one modality given another (Ngiam et al. 2011; Sohn, Shang, and Lee 2014).

Generative adversarial networks (GANs) have previously been used for translation tasks. For instance, GANs were used to translate from captions to their associated images, generating images of birds and flowers from text captions (Reed et al. 2016). Previous work in the imaging domain has explored using GANs for translation tasks by combining adversarial losses with traditional regression losses (Isola et al. 2016), but these systems have since been surpassed by adversarial-only systems, such as one which used a bidirectional cycle-consistent adversarial network (Cycle GAN), to translate images from one style to another (Zhu et al. 2017). No investigations that we know of have applied any adversarial techniques, with or without regression losses, to biomedical translation tasks. GANs have also been explored for semi-supervised learning, but such uses have examined classification tasks in imaging domains, not regression, as we do here (Salimans et al. 2016; Springenberg 2015).

Much prior work in the clinical setting focuses on single domain learning (text, physiological data, etc.) in order to perform supervised prediction or retrieval tasks. For example, predicting mortality given previously observed clinical notes, predicting common billing codes given a portion of a patient's record, or predicting interventions based on an inferred physiological latent space (Ghassemi et al. 2014; Lipton et al. 2015; Ghassemi et al. 2017; Wu et al. 2017). Adversarial models have recently been used on clinical data to generate binary and count summarizations of patient records, and to generate clinical time series (Choi et al. 2017; Esteban, Hyland, and Rtsch 2017). In both cases, GANs were used principally for their generative capabilities, rather than modality translation or semi-supervised learning.

Methods

In the present study, we develop a novel approach to semi-supervised, bi-directional translation shown in Figure 1 using a Cycle Wasserstein Regression GAN (CWR-GAN). The CWR-GAN is constructed from several architectures: GANs in general, Wasserstein GANs, and cycle-consistent GANs.

Generative Adversarial Networks (GANs)

There are two parts to the traditional GAN: a generator $G(\mathbf{z}; \boldsymbol{\theta}_g)$ and discriminator $D(\mathbf{x}; \boldsymbol{\theta}_d)$ (Goodfellow et al. 2014). D and G compete in a two-player minimax game, where D's goal is to discriminate between real and synthetic data, and G's goal is to generate synthetic data that can fool D. In their original formulation, the traditional GAN loss function, at critic optimality, measures the Jensen-Shannon divergence between $G(\mathbf{z})$ and p_{data} (Goodfellow 2016).

Traditionally, GANs are trained in turns: first the discriminator is trained for a number of epochs, then the generator is trained for one epoch, using gradients from the discriminator fixed at its value based on training thus far. This alternating training procedure is repeated until convergence. We follow this same training structure in our system as well.

Wasserstein GAN (WGAN) Recent work has proposed use of the Wasserstein (or Earth Mover's/EM) distance to formulate a "critic" in lieu of the traditional GAN discriminator (Arjovsky, Chintala, and Bottou 2017). Compared to traditional discriminators, Wasserstein critics help stabilize GAN training because the EM distance never saturates, and thus provides meaningful gradients to the generator throughout training.

The best known implementation of such a WGAN is via the following loss, which we will use as our adversarial foundation throughout this work (Gulrajani et al. 2017):

$$\min_G \max_D \mathop{\mathbb{E}}_{\mathbf{x}\sim p_X} [D(\boldsymbol{x})] - \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})} [D(G(\boldsymbol{z}))] \\ -\lambda \left(\mathop{\mathbb{E}}_{\bar{\mathbf{x}}\sim p_{\bar{X}}} [\|\nabla_{\bar{\boldsymbol{x}}} D(\bar{\boldsymbol{x}})\|_2 - 1] \right)^2 \qquad (1)$$

where $p_{\bar{X}}$ is defined via $\bar{\mathbf{x}} = \epsilon\mathbf{x} + (1-\epsilon)G(\mathbf{z}), \epsilon \sim U([0,1])$. In this loss, λ is a hyperparameter that should be set sufficiently high so as to insist the gradient loss term remains small throughout training.

Cycle-consistent GAN (Cycle GAN) Cycle GANs learn to translate points between two domains, X and Y, using only unsupervised, adversarial signals. To do this, they learn two "generators" $G_X : X \to Y$ and $G_Y : Y \to X$, and two discriminators (or Wasserstein critics), D_X and D_Y. Both generators are trained not only according to an adversarial loss, but also to minimize a cyclical reconstruction error penalty:

$$\begin{aligned}\mathcal{L}_{\text{Cyc}} = &\mathop{\mathbb{E}}_{\mathbf{x}\sim p_X}\left[\|\boldsymbol{x} - G_Y(G_X(\boldsymbol{x}))\|_1\right]\\ &+ \mathop{\mathbb{E}}_{\mathbf{y}\sim p_Y}\left[\|\boldsymbol{y} - G_X(G_Y(\boldsymbol{y}))\|_1\right]. \qquad (2)\end{aligned}$$

This loss regularizes both learned models towards being each others' inverse, and reframes the two isolated regression tasks as a single multi-task model (Zhu et al. 2017).

Cycle Wasserstein Regression GAN (CWR-GAN)

We present a novel joint regression-adversarial model[1] for biomedical translation problems, where our goal is to learn mapping functions between two encoded domains $\hat{X}$ and $\hat{Y}$ given training samples $\{\hat{x}_i\}_{i=1}^{N} \in \hat{X}$ and $\{\hat{y}_j\}_{j=1}^{M} \in \hat{Y}$.

Our model implements a Cycle Wasserstein GAN with the addition of a regression loss on paired samples. Given a source domain X and a target domain Y, with some subsets $P \subseteq X \times Y$ consisting of paired observations, our full objective is as follows:

$$\begin{aligned}\mathcal{L}_{\text{CW-Crt}} = &-\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y}\sim p_{X\times Y}}\left[C_X(\boldsymbol{x}) + C_Y(\boldsymbol{y})\right]\\ &+ \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y}\sim p_{G_X(X)\times G_Y(Y)}}\left[C_X(\boldsymbol{x}) + C_Y(\boldsymbol{y})\right]\\ &+ \lambda \mathop{\mathbb{E}}_{\bar{\mathbf{x}}\sim P_{\bar{X}}}\left[\|\nabla C_X(\bar{\boldsymbol{x}})\|_2 - 1\right]^2\\ &+ \lambda \mathop{\mathbb{E}}_{\bar{\mathbf{y}}\sim P_{\bar{Y}}}\left[\|\nabla C_Y(\bar{\boldsymbol{y}})\|_2 - 1\right]^2 \qquad (3)\\ \mathcal{L}_{\text{CW-Gen}} = &-\mathcal{L}_{\text{CW-Crt}} + \nu\mathcal{L}_{\text{Cyc}}\\ &+ \alpha \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y}\sim p_P}\left[\|\boldsymbol{x} - G_Y(\boldsymbol{y})\|_2 + \|\boldsymbol{y} - G_X(\boldsymbol{x})\|_2\right] \qquad (4)\end{aligned}$$

Components of this loss offer different training signals:

1. The traditional regression loss term directly trains the generator to perform a low error translation based on the limited paired data available. If all data available is paired, this will be the most direct loss term, and yield the best training signals. This loss term is weighted by hyperparameter α.
2. The cycle loss term regularizes the learned models against those for the opposite direction. This ties the two, otherwise independent loss objectives (for G_X and G_Y) together, and is weighted by hyperparameter ν.
3. The adversarial loss term ($-\mathcal{L}_{\text{CW-Crt}}$) helps regularize each model individually by pushing predictions towards regions of high perceived likelihood.

Taken together, these components insist that the two learned maps G_X and G_Y should be approximately invertible, each able to learn well from unpaired data, and able to be refined using paired examples. Traditionally, GANs can suffer from a problem known as *mode collapse*, wherein the generator only generates a very small set of identical examples, each of which is viewed as realistic by the critic. However, our system does not suffer from this problem, as each component of our loss helps penalize this kind of error. Standard regression losses obviously prohibit mode-collapse behavior, as does our cycle consistency penalization, and the Wasserstein formulation of our adversarial losses have also been shown independently to suffer much less from mode collapse than a traditional discriminator (Arjovsky, Chintala, and Bottou 2017).

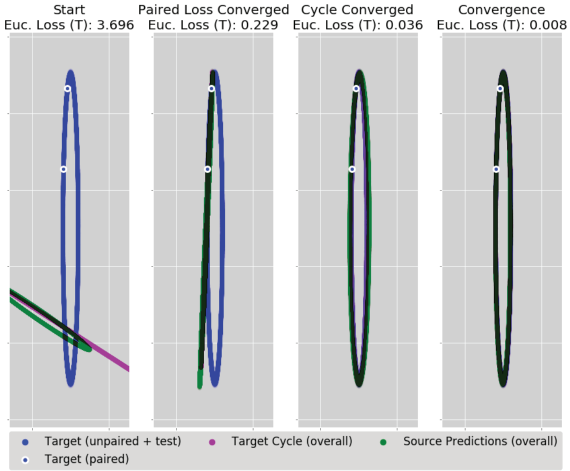

Figure 2: The CWR-GAN system on a synthetic domain with only two paired examples. *Far Left:* At initialization, no map is meaningful, and no loss has converged. *Middle Left:* Training first locks both paired points (highlighted in white) to their correct values. *Middle Right:* Training locks both maps into an invertible pair, but has yet to fine tune the output distribution to the exact shape. *Far Right:* The adversarial loss has guided the model to the correct shape. Cycle loss drives the mappings to be invertible before the final distributions are correctly found. After the cycle loss falls to zero, both maps evolve in tandem until convergence.

Synthetic Experiments

We provide the simplest possible demonstration of the key aspects of our system on synthetic datasets. We generate a source X distributed about the 2D unit circle, and target Y affected by a simple affine transformation, defined as follows:

$$\mathsf{r} \sim \mathcal{N}(1, 0.01) \qquad \theta \sim U([0, 2\pi])$$

$$X = \begin{bmatrix}\mathsf{r}\cos(\theta)\\ \mathsf{r}\sin(\theta)\end{bmatrix} \qquad Y = \begin{bmatrix}0.2 & 0\\ 0 & 4\end{bmatrix} X + \begin{bmatrix}1\\ 3\end{bmatrix}$$

We present results on the noiseless domain defined above for simplicity, but note that these results hold with mild, independent Gaussian noise on both X and Y.

We generated 10,000 total samples, with an 80%/20% train/test split, and offered the system either no paired samples or only two paired samples following the train/test split. In this domain, translators are affine transformations, and critics are 3-layer deep, 300-neuron wide leaky rectified linear

[1] Code available at https://github.com/mmcdermott/CWR-GAN.

unit (Leaky ReLu). All networks in this work use a Leaky ReLu activation, with $\alpha = 0.3$ (e.g. $\texttt{LeakyReLu}(\boldsymbol{x}) = \max(0.3\boldsymbol{x}, \boldsymbol{x})$). networks. These depth/width settings were chosen to be similar to the 2D experiments in earlier WGAN works (Gulrajani et al. 2017). Regression loss multipliers (α) were set to 1 in this experiment, cycle loss (ν) to 0.2, gradient loss (λ) to 3, and 3 critic epochs were performed for every 1 translator epoch.

We highlight three aspects of the CWR-GAN observed on this synthetic dataset:

- Absent any paired data, the system learns the correct output distributions, and a map consistent with the symmetries of the output distributions, thereby demonstrating the value of adversarial signals in their own right.
- With two paired data points, the system learns not only the correct output distribution, but the correct map, thereby demonstrating that adversarial signals can complement paired examples to learn a map benefiting from both sources of information (Figure 2).
- The cycle loss component serves to "snap" the maps together into an invertible pair, thus helping each use the other for regularization.

In this synthetic verification and the two experiments on biomedical datasets that follow, all models were implemented in Tensorflow (Abadi et al. 2016). We use the Adam optimizer (Kingma and Ba 2014), with hyperparameters similar to those recommended in prior work (Gulrajani et al. 2017) ($\alpha = 0.00005$, $\beta_1 = 0.5$, $\beta_2 = 0.9$) in the CWR-GAN for critics and generators.

Experiment I: Individualized Treatment Effect Prediction with ICU Patient EHRs

In this experiment, we focus on predicting individual patients' responses to interventions (i.e., from pre- to post-treatment). We examine 29 noisy timeseries derived from the electronic health records (EHRs) of intensive care unit (ICU) patients. These signals are recorded hourly; however, it is common for interventions to be applied near the beginning or end of a patient's stay. This limits the availability of paired training examples because few records contain sufficiently large, equally sized windows both pre- and post-intervention. A standard regression system can only use fully paired examples, but our CWR-GAN model can also use those records that only have one such window as an unpaired example of either the source (pre-intervention) domain or the target (post-intervention) domain. This allows us to learn from additional data that is inaccessible to traditional regressors.

Forecasting a patient's response to a treatment — their individualized treatment effect (ITE) — is an important task because the efficacy of clinical interventions can vary drastically among patients. Further, unnecessarily administering an intervention is expensive and potentially harmful. We target two interventions: invasive ventilation and vasopressor use. While ventilation is a commonly used ICU treatment, there are many potential complications and changes in ventilation settings can impact patient outcomes (Yang and Tobin 1991; Tobin 2006). Similarly, vasopressors are a medication commonly used in the ICU, but have been found to be harmful in certain populations (D'Aragon et al. 2015).

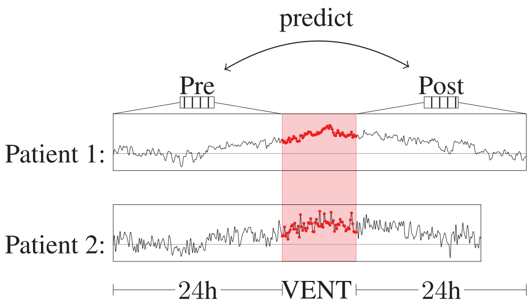

Figure 3: ITE regression task setup. Physiological series corresponding to some pre-intervention window (VENT, highlighted in red) are summarized into a fixed-size encoding, then translated to the corresponding post-intervention window. In this example, Patient 1 has sufficient data on both sides the intervention window to form a *paired* training example. However, Patient 2 does have not sufficient time post-intervention, constituting an *unpaired* "pre-" training example from the source domain.

Data Source & Preprocessing

We use data from the publicly available Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III v1.4) database (Johnson et al. 2016). We consider the first ICU stay of patients 15 and older whose stay duration was 12 to 240 hours, yielding 33,287 unique ICU stays. For each patient, we extract the static variables gender and age, as well as 29 time-varying vitals and labs, the same used by other work (Suresh, Szolovits, and Ghassemi 2017). Vital and lab measurements are given timestamps rounded to the nearest hour, and multiple measurements within an hour were averaged. Measurements are only recorded for hours in which they are taken, so missing measurements are common.

There have been several proposed encodings for physiological data (Che et al. 2016; Ghassemi et al. 2017). Here, we use a fixed sized encoding formed by concatenating measurement counts (i.e. how often a measurement was taken) with the average value observed for each measurement during the time interval. If a measurement was never taken during the interval, we fall back to the patient's average for that measurement; if it was never taken for that patient, we use the population average. Finally, we concatenate the patient's age and gender to this representation to enable our task to use static signals as well as the summarized time series.

Experimental Setup

Our primary prediction task is performed over 24-hour windows (Figure 3). These yield a range of paired v. unpaired splits, and our goal is to predict the physiological signals post-treatment (target) given the patient's physiological signals pre-treatment (source). We examine four interventions: invasive ventilation (VENT), and the use of three specific vasopressors — phenylephrine (PHEN), norepinephrine (NOREP), and dopamine (DOP). Final population sizes for

	Paired	Pre-	Post-
Ventilation	834	469	7973
Phenylephrine	510	568	3697
Norepinephrine	247	363	1931
Dopamine	159	135	960

Table 1: Population sizes for the intervention prediction tasks. Each intervention has three distinct populations: 1) paired patients, 2) "pre-" unpaired records, obtained from patients whose ICU stays did not contain a full 24-hour interval following intervention application, and 3) "post-" unpaired records, obtained from patients whose ICU stays did not contain a full 24-hour interval preceding intervention application.

each intervention after extraction are illustrated in Table 1.

We are primarily interested in the difference between a traditional regression neural network and our semi-supervised system with respect to the natural regression loss for the target domain (e.g., euclidean distance loss in the post-intervention domain). As such, we train and evaluate two models: 1) a traditional regression multi-layer perceptron (MLP) for either direction of regression, and 2) our semi-supervised system, which augments the traditional network with a Wasserstein critic and the cycle loss penalty.

Models were tuned, then evaluated via nested cross-validation. All regression and critic networks were 3-layer, bidirectional regressors using leaky ReLU activations, dropout of 0.75, and L2 & L1 regularization of $1\text{e}{-3}$. All hyperparameters were chosen via nested cross-validation search, and results are reported in terms of median euclidean distance loss on the target domain across the same outer cross-validation split. Loss multipliers were fixed independently of task at a multiplier of 10 for the regression component and 1 for both the adversarial and cycle reconstruction error losses. The gradient loss multiplier was set to 10, but if a critic appeared to suffer from gradient explosion during training, it was increased to 50. Models were trained for up to 9 consecutive critic epochs, stopping after 3 critic epochs that did not improve the adversarial loss, then 1 translator epoch.

Results

Results for the performance of our system are shown in Table 2. We see that on three of the four interventions, our joint system yields an improvement over a traditional regression system in terms of the overall loss by fractions ranging from 0.5% to 7.4%. On the dopamine vasopressor prediction task (DOP), we underperform the traditional system by 2.7%. This may be because dopamine has the smallest fraction of $\frac{\text{unpaired}}{\text{paired}}$ data of any of our sources (6.9 for dopamine vs. 8.4, 9.3, and 10.12 for phenylephrine, norepinephrine, and ventilation, respectively), but could also be caused by other, unforeseen complexities. Nevertheless, overall, these results demonstrate that even with as few as 834 paired data points, or as few as ~ 2500 total points, the CWR-GAN can successfully learn from unpaired instances.

We also observed that during the majority of the cross-validation search, the CWR-GAN system would outperform the traditional system across a majority of tasks for a variety of reasonable, though sub-optimal, parameter settings. Upon closer inspection, this appeared to be due to additional regularization effects inherent in the adversarial nature of our system's learning, though this warrants additional study. For example, Figure 4 shows that dropout is far more influential on the standard predictor than on the CWR-GAN model. This figure is taken from our actual cross validation results, and is thus using sub-optimal parameters; thus, the scale of the losses shown here is not representative of our tuned losses. However, these results do suggest that the adversarial signals and cyclic loss penalty may help to regularize the model in a manner orthogonal to traditional methods of regularization.

	Intervention Type			
Model	VENT	NOREP	DOP	PHEN
MLP	3.780	2.829	2.719	3.186
CWR-GAN	-0.50%	-7.4%	$+2.7\%$	-4.5%

Table 2: Comparison of median model performance on four targeted interventions. The traditional MLP regression network performance is reported in Euclidean distance. Our semi-supervised CWR-GAN results report the difference from the MLP's loss, as a percentage of that loss. Thus, a positive percentage is where the CWR-GAN performed *worse* than the MLP (i.e., on the dopamine treatment effect task), and the remaining negative losses on all other ITE tasks are where the CWR-GAN was *better*. All models significantly outperformed a linear baseline.

Discussion

Analyzing intervention effect is often hampered by the fact that many patients lack sufficient pre-intervention and post-intervention signals to offer a full regression pair. In such cases, the data collected in either their pre- or post- intervention would be ignored by traditional, uni-directional regression approaches. However, our method demonstrates these can still provide valuable signals independently.

Another potential medium for our approach is causal inference, i.e., characterizing how a treatment tested on one cohort would affect a more general population. The ability to use unpaired data that could not be included in standard regression studies in this field would be extremely valuable.

During our experiments, we also considered shorter windows (6 and 12 hours) which contained more paired training data. However, shorter time windows preclude the inclusion of a full diurnal cycle, and it is well-established that circadian rhythm influences most physiological parameters, including metabolic, endocrine and immune functions (Sundararajan, Flabouris, and Thompson 2016). This adds a dimension to the learning task which cannot be inferred from the data, as it is not possible to know whether a patient will stop their intervention during the evening or the morning. Additionally, missingness is more prevalent in the 12 and 6 hour periods, as from smaller gene fragments less frequently performed tests are less likely to appear in these shortened windows. As missingness increases, the imputed, average signals will be more common, which induces a non-representative spike in the data distribution. This predominantly penalizes the

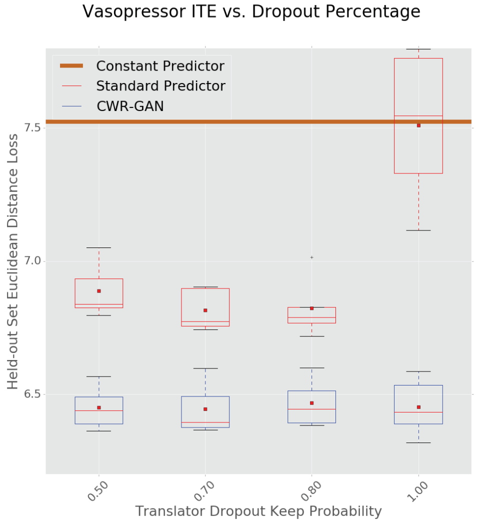

Figure 4: Semi-supervised signals offer regularization to ITE regression. The thick, horizontal line is attained by a 'no-change' prediction baseline; e.g., predicting that the intervention does not alter the physiological signals. We demonstrate that appropriate dropout is vital for the MLP (lower is better), but it is not necessary for the improved CWR-GAN results.

adversarial system, as it relies on distributional signals. It is thus unsurprising that all model performance decreased on these tasks; the CWR-GAN system in particular failing to ever outperform a traditional model. However, we also note that the CWR-GAN performance consistently suffered more as less data was available as unpaired vs. paired; this finding reinforces the transcriptomics results we discuss next and indicates the model is learning from unpaired samples.

Experiment II: Transcriptomics Extrapolation

Transcriptomics data give a view into a cell's internal state by directly measuring the expression levels of genes via their transcriptional byproducts. The full transcriptome is extensive and contains many transcripts, each corresponding to unique proteins with diverse functions. However, it is also redundant, and much can be learned about the cell state from a small subset of the transcriptome. As such, high throughput techniques may measure only a subset of transcripts, thereby saving money and time, and attempt to infer the full gene expression profile (Subramanian et al. 2017).

Data Source & Preprocessing

The L1000 technique is commonly used to perform high throughput transcriptomics assays. This technique only directly measures 978 transcripts, and then uses these to infer those remaining (Subramanian et al. 2017). The L1000 developers have released a dataset of 100,000 full transcriptomes, split between the 978 landmark genes and those remaining, to the NCBI GEO database under series number GSE70138[2] (Broad Connectivity Map Team 2016). We use this dataset for our task, where the source domain is the 978 landmark genes, and the target domain were restricted to the first 5000 genes for computational efficiency. Data were centered and scale normalized.

Experimental Setup

In the intervention task, our unpaired samples were derived from the same source as the paired examples. Thus, in that context, we could augment our model with those unpaired examples and still fairly evaluate on only a subset of the available paired records. In the context of transcriptomics extrapolation, however, this is not possible; though many scientists have published external transcriptome subsets that we could use as "unpaired" instances in training, we *know* that these are from a different distribution than our main paired dataset because each individual study produces gene expression datasets according to their own individual scientific prerogatives. This means that if we did use these unpaired sources, it would be difficult, if not impossible, to evaluate our model; even if gains in inference occurred generally, they could be negligible on the paired data alone, which is our only testable data source.

Instead, we randomly split our dataset into paired and unpaired datasets according to four variants: $100\%, 50\%, 10\%$, and 3% paired. This allows us to not only effectively evaluate our model in the context of unpaired data, but also to probe how the relationship of our model to a standard regressor varies with the amount of unpaired vs. paired data available.

As before, we are primarily interested in the difference between a traditional regression system and our semi-supervised system in terms of the natural regression loss for the target domain—i.e. the euclidean distance loss on the 5000-dim. transcriptome subset. As such, we train and evaluate a traditional regression network, and our semi-supervised system, and compare both of their regression (i.e. not adversarial) losses in the target domain.

Translator networks were exclusively leaky ReLU networks with varyied hyperparameters and network configurations depending upon the specific variant. On the 100% or 50% paired variants, regression networks had three hidden layers and no regularization. If 90% of the data were paired, regression networks had 2 hidden layers, dropout of 0.6, and L2 regularziation of $1\mathrm{e}{-3}$. If 97% of the data were paired, networks had 2 hidden layers, with dropout of 0.5, L2 regularization of $2\mathrm{e}{-3}$ and L1 regularization of $2\mathrm{e}{-5}$. Hidden layer dimensionality matched input dimensionality in all cases. Critics were 2 hidden layer networks with dropout of 0.9, L2 regularization of $2\mathrm{e}{-4}$ and L1 regularization of $1\mathrm{e}{-6}$ in all cases. Hyperparameters were chosen according to a grid search with a randomly sampled 15% validation set. Models were tested on a held out, randomly sampled 20% test set.

Results

On this dataset, if all data are paired, our system underperforms a traditional neural network, attaining a loss 3.6% higher. This is expected, as the regression loss is the most direct training objective in the fully paired case. However, as

[2] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70138

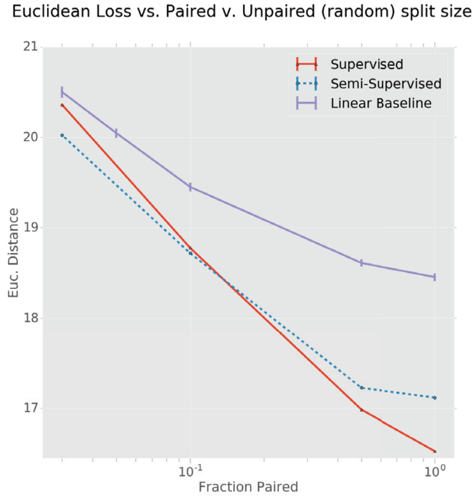

Figure 5: Performance (in Euclidean distance regression loss) of a linear model, standard predictor, and our semi-supervised CWR-GAN as a function of the fraction of the data that is paired. Lower is better. As we reduce the amount of paired data available, the semi-supervised model gains ground in performance over both baselines, demonstrating that it learns from the unpaired data sources.

we increase the fraction of data that are unpaired, our system gains more and more ground: the CWR-GAN system yielding loss deltas of $1.4\%, -0.3\%, -1.6\%$ at 50%, 90%, and 97% unpaired, respectively. Thus, we see that the amount of unpaired data is perfectly correlated with our model's performance over a traditional system, which offers strong support for the argument that the CWR-GAN is learning additional signals from the unpaired samples. Both models significantly outperform a ridge regression ($\lambda = 100$) baseline. The total results in raw units are shown in figure 5.

Discussion

Transcriptomics is a promising and rapidly evolving area of computational and clinical biology. It holds tremendous promise for disease subtyping, drug discovery, and diagnostics. However, the expense of collecting a full transcriptome sample and the inherent inter-experimental variability in these data, mean there are often not enough data to form generalizable conclusions. The ability to include unpaired or unlabeled instances would aid in circumventing this boundary.

Our results demonstrate that the addition of unpaired data does allow the model to learn additional features from the data, which bodes well for the possibility of using these ideas for real-world use cases, such as augmenting a broader inference algorithm with unpaired data from more recent samples that reflect the full diversity of transcriptomics analyses, as well as for modality translation within multi-omics studies. Further, these results underscore those attained in the intervention task by reiterating that the model can successfully learn from the unpaired data, now on a dataset of total size 100,000, rather than approximately 8,000.

Conclusion

The ability to incorporate large amounts of unpaired data in a regressive translation is critical in many tasks in the biomedical field. Because these tasks are bio-physically driven, we also desire that any mappings we learn be invertible. These constraints are also applicable in other fields, e.g. mapping pre- and post- trip behaviors from ride-sharing data, or translating social media data pre- and post- an important event. In these settings, the tasks would similarly benefit from the ability to integrate a large amount of unpaired data.

The goal of this work was to create a model that used unpaired data to learn invertible translations more accurately with fewer paired datapoints. We demonstrated our method's applicability to the biomedical field using two different experiments. First, the CWR-GAN was able to use unpaired data pre- and post- treatment to improve ICU patient ITE forecasts in three of four ICU interventions in real-world, noisy physiological signals across more than 2,000 patients. Second, the CWR-GAN was able to successfully extrapolate a 978 dimensional transcriptome fragment to a much larger 5,000 dimensional subset of the transcriptome, and this ability improved relative to traditional methods as the ratio of unpaired to paired data increased.

This approach has two main limitations. First, this method takes much longer than standard predictors. Translator mappings do not typically require more epochs to reach convergence, but because the critic must also be trained, and as more data can be used, they take much longer in terms of wall-clock time. Second, adversarial networks remain a very active area of research, and they are notoriously difficult to train and understand. There use here, while offering many advantages, also means that this method is inherently more high-maintenance than other strategies. We also note that recent work has shown the EM distance metric (which is central to the Wasserstein GAN critic loss) yields *biased* sample gradients, and thus is perhaps not well suited to stochastic gradient descent (Bellemare et al. 2017).

In future work, we plan to undertake efforts to improve training stability, including automatically tuning some of the loss multipliers as training evolves, or by experimenting with newer variants of GANs, such as the Cramér GAN. Additionally, we plan to examine the contribution of each loss component alone, e.g. cycle loss, and determine what precise effect they offer. We have previously tried simply adding a cycle loss component to a bi-directional traditional regressor, but this alone did not offer any significant performance improvement. Finally, we also plan to further probe the apparent regularization effect offered by this system.

Acknowledgments This research was funded in part by the Intel Science and Technology Center for Big Data, a Philips-MIT research agreement, the Wistron Corporation, and grants from the National Institutes of Health (NIH): National Library of Medicine (NLM) Biomedical Informatics Research Training grant 2T15 LM007092-22, National Institute of Biomedical Imaging and Bioengineering (NIBIB) grant R01-EB017205, National Institute of Mental Health (NIMH) grant P50-MH106933, National Human Genome Research Institute (NHGRI) grant U54-HG007963.

References

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mane, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viegas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; and Zheng, X. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv:1603.04467 [cs]*. arXiv: 1603.04467.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. *arXiv:1701.07875 [cs, stat]*. arXiv: 1701.07875.

Bellemare, M. G.; Danihelka, I.; Dabney, W.; Mohamed, S.; Lakshminarayanan, B.; Hoyer, S.; and Munos, R. 2017. The Cramer Distance as a Solution to Biased Wasserstein Gradients. *arXiv:1705.10743 [cs, stat]*. arXiv: 1705.10743.

Broad Connectivity Map Team. 2016. L1000 connectivity map perturbational profiles from broad institute lincs center for transcriptomics lincs phase *II*.

Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; and Liu, Y. 2016. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *arXiv:1606.01865 [cs, stat]*. arXiv: 1606.01865.

Choi, E.; Biswal, S.; Malin, B.; Duke, J.; Stewart, W. F.; and Sun, J. 2017. Generating Multi-label Discrete Electronic Health Records using Generative Adversarial Networks. *arXiv:1703.06490 [cs]*. arXiv: 1703.06490.

D'Aragon, F.; Belley-Cote, E. P.; Meade, M. O.; Lauzier, F.; Adhikari, N. K.; Briel, M.; Lalu, M.; Kanji, S.; Asfar, P.; Turgeon, A. F.; et al. 2015. Blood pressure targets for vasopressor therapy: A systematic review. *Shock* 43(6):530–539.

Eisen, M. B.; Spellman, P. T.; Brown, P. O.; and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* 95(25):14863–14868.

Esteban, C.; Hyland, S. L.; and Rtsch, G. 2017. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. *arXiv:1706.02633 [cs, stat]*. arXiv: 1706.02633.

Ghassemi, M.; Naumann, T.; Doshi-Velez, F.; Brimmer, N.; Joshi, R.; Rumshisky, A.; and Szolovits, P. 2014. Unfolding Physiological State: Mortality Modelling in Intensive Care Units. In *Proc. of the 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, KDD '14, 75–84. New York, NY, USA: ACM.

Ghassemi, M.; Wu, M.; Hughes, M. C.; Szolovits, P.; and Doshi-Velez, F. 2017. Predicting intervention onset in the ICU with switching state space models. *AMIA Summits on Trans. Sci. Proc.* 2017:82–91.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc. 2672–2680.

Goodfellow, I. 2016. NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv:1701.00160 [cs]*. arXiv: 1701.00160.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. 2017. Improved Training of Wasserstein GANs. *arXiv:1704.00028 [cs, stat]*. arXiv: 1704.00028.

Isola, P.; Zhu, J.; Zhou, T.; and Efros, A. A. 2016. Image-to-image translation with conditional adversarial networks. *CoRR* abs/1611.07004.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3.

Kingma, D. P., and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. arXiv: 1412.6980.

Lipton, Z. C.; Kale, D. C.; Elkan, C.; and Wetzel, R. 2015. Learning to Diagnose with LSTM Recurrent Neural Networks. *arXiv:1511.03677 [cs]*. arXiv: 1511.03677.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative Adversarial Text to Image Synthesis. In *PMLR*, 1060–1069.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X.; and Chen, X. 2016. Improved Techniques for Training GANs. In Lee, D. D.; Sugiyama, M.; Luxburg, U. V.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc. 2234–2242.

Sohn, K.; Shang, W.; and Lee, H. 2014. Improved Multimodal Deep Learning with Variation of Information. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc. 2141–2149.

Springenberg, J. T. 2015. Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks. *arXiv:1511.06390 [cs, stat]*. arXiv: 1511.06390.

Subramanian, A.; Narayan, R.; Corsello, S. M.; Peck, D. D.; Natoli, T. E.; Lu, X.; Gould, J.; Davis, J. F.; Tubelli, A. A.; Asiedu, J. K.; Lahr, D. L.; Hirschman, J. E.; Liu, Z.; Donahue, M.; Julian, B.; Khan, M.; Wadden, D.; Smith, I.; Lam, D.; Liberzon, A.; Toder, C.; Bagul, M.; Orzechowski, M.; Enache, O. M.; Piccioni, F.; Berger, A. H.; Shamji, A.; Brooks, A. N.; Vrcic, A.; Flynn, C.; Rosains, J.; Takeda, D.; Davison, D.; Lamb, J.; Ardlie, K.; Hogstrom, L.; Gray, N. S.; Clemons, P. A.; Silver, S.; Wu, X.; Zhao, W.-N.; Read-Button, W.; Wu, X.; Haggarty, S. J.; Ronco, L. V.; Boehm, J. S.; Schreiber, S. L.; Doench, J. G.; Bittker, J. A.; Root, D. E.; Wong, B.; and Golub, T. R. 2017. A Next Generation Connectivity Map: L1000 Platform And The First 1,000,000 Profiles. *bioRxiv* 136168.

Sundararajan, K.; Flabouris, A.; and Thompson, C. 2016. Diurnal variation in the performance of rapid response systems: the role of critical care servicesa review article. *Journal of intensive care* 4(1):15.

Suresh, H.; Szolovits, P.; and Ghassemi, M. 2017. The Use of Autoencoders for Discovering Patient Phenotypes. *arXiv:1703.07004 [cs]*. arXiv: 1703.07004.

Tobin, M. J. 2006. *Principles and practice of mechanical ventilation*. LWW.

Wu, M.; Ghassemi, M.; Feng, M.; Celi, L. A.; Szolovits, P.; and Doshi-Velez, F. 2017. Understanding vasopressor intervention and weaning: risk prediction in a public heterogeneous clinical time series database. *J of the American Med Inform Assoc* 24(3):488–495.

Yang, K. L., and Tobin, M. J. 1991. A prospective study of indexes predicting the outcome of trials of weaning from mechanical ventilation. *New England Journal of Medicine* 324.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv:1703.10593 [cs]*. arXiv: 1703.10593.