Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

Gene expression

Optimal Bayesian supervised domain adaptation for RNA sequencing data

Shahin Boluki [1,*], Xiaoning Qian [1,2,*] and Edward R. Dougherty [1,*]

[1] Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX 77843, USA.
[2] TEES-AgriLife Center for Bioinformatics & Genomic Systems Engineering, Texas A&M University, College Station, TX 77843, USA.

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: When learning to subtype complex disease based on next-generation sequencing data, the amount of available data is often limited. Recent works have tried to leverage data from other *domains* to design better predictors in the *target* domain of interest with varying degrees of success. But they are either limited to the cases requiring the outcome label correspondence across domains or cannot leverage the label information at all. Moreover, the existing methods cannot usually benefit from other information available *a priori* such as gene interaction networks.

Results: In this paper, we develop a generative optimal Bayesian supervised domain adaptation (OBSDA) model that can integrate RNA sequencing (RNA-Seq) data from different domains along with their labels for improving prediction accuracy in the target domain. Our model can be applied in cases where different domains share the same labels or have different ones. OBSDA is based on a hierarchical Bayesian negative binomial model with parameter factorization, for which the optimal predictor can be derived by marginalization of likelihood over the posterior of the parameters. We first provide an efficient Gibbs sampler for parameter inference in OBSDA. Then, we leverage the gene-gene network prior information and construct an *informed* and flexible variational family to infer the posterior distributions of model parameters. Comprehensive experiments on real-world RNA-Seq data demonstrate the superior performance of OBSDA, in terms of accuracy in identifying cancer subtypes by utilizing data from different domains. Moreover, we show that by taking advantage of the prior network information we can further improve the performance.

Availability: The source code for implementations of OBSDA and SI-OBSDA are available at the following link.
https://github.com/SHBLK/BSDA

Contact: s.boluki@tamu.edu or xqian@ece.tamu.edu or edward@ece.tamu.edu

1 Introduction

When designing predictive models for a target task, traditionally only the data from the target domain are used for training with the commonly adopted assumption that the training and testing data have the same feature-label distributions. However, in many cases, especially with next-generation sequencing (NGS) technologies, the number of training samples that can be collected in the target domain is limited compared with the dimensionality of the features (the number of genes). Collecting appropriate data from complex diseases is a costly procedure, if not prohibitive, considering the clinical, biological, and technical challenges involved in the process. These limitations can prohibit collecting enough samples from the disease/condition of interest to design a reproducible predictor. Given the prevalent data heterogeneity in complex diseases like cancer (Alizadeh *et al.*, 2015), usually more samples are needed than what can be collected to achieve reliable predictors. It is believed that different diseases share some underlying biological processes and modules (Garcia-Vaquero *et al.*, 2018; Gustafsson *et al.*, 2014; Menche *et al.*, 2015;

Levine, 2013), indicating that data from one disease can be informative for other diseases. Hence, it is desirable to learn useful information from available data from other conditions and/or technologies to help derive more accurate predictions in the target domain. Moreover, other than the data at hand, additional knowledge is usually available *a priori* (before observing data) that can be beneficial for the target task (Wei and Pan, 2012; Boluki *et al.*, 2017). Examples of this include interaction networks, which might have been compiled from several studies and databases (Menche *et al.*, 2015; Stark *et al.*, 2010; Aranda *et al.*, 2010) containing potentially useful information for the target task. Our goal is to develop a new optimal Bayesian supervised domain adaptation (OBSDA) framework capable of leveraging data and label information from other domains in addition to prior network knowledge to design more accurate and reliable predictors in a target domain of interest.

Transfer learning and domain adaptation methods (Pan and Yang, 2009; Patel *et al.*, 2015) aim to leverage data from other domains for achieving better results for the task in the target domain. Common approaches generally include adapting the predictor in the source domain to the target domain and/or the distribution of the data across domains (Weiss *et al.*, 2016). Some methods, including Dai *et al.* (2007); Borgwardt *et al.* (2006) reweight the source and target samples. Other representative methods, such as Pan *et al.* (2010); Gong *et al.* (2013), first project the target and all or a subset of source data to a common subspace, which minimizes a discrepancy metric between the marginal distributions of features in the domains, and then train a discriminator in that space. The application of these methods are often limited to cases where source and target data are from the same classes. On the other hand, multi-task learning methods (Jacob *et al.*, 2009; Kang *et al.*, 2011; Passos *et al.*, 2012) aim to improve prediction power overall in all domains/tasks, with some requiring at least several tasks/domains for reasonable performance. The majority of deep learning-based domain adaptation methods (Yosinski *et al.*, 2014; Long *et al.*, 2015; Liu and Tuzel, 2016), which usually share parameters and/or lower-level representations across domains and have found their major successes in computer vision tasks, need much larger training sets in all the domains than what is practical in typical clinical studies.

Some of the recent transfer learning and domain adaptation works on gene expression data include Normand *et al.* (2018); Dhruba *et al.* (2018); Hajiramezanali *et al.* (2018); Karbalayghareh *et al.* (2019). In Normand *et al.* (2018) the authors developed a method to predict differentially expressed genes in a condition for humans based on gene expression data collected from disease studies on mice. Dhruba *et al.* (2018) proposed two methods respectively—mapping of features to a common subspace and mapping target domains to the source space—to better predict drug sensitivity based on gene expression data from additional databases. Both Hajiramezanali *et al.* (2018) and Karbalayghareh *et al.* (2019) proposed methods for utilizing gene expression data from other domains to build more reliable cancer subtype predictors in the target domain. In Hajiramezanali *et al.* (2018), a hierarchical Bayesian model was developed to map the samples from different domains to a shared latent space with the classifier trained on the lower dimensional representations to predict cancer subtypes. One shortcoming of the method is that label information is not used in the latent representation learning stage. Karbalayghareh *et al.* (2019) proposed a Bayesian method with joint priors on the parameters from source and target domains and derived the predictor by marginalizing over source parameters. Despite being a principled approach, it models only the relationship between data from the same classes across domains, with the limitation of not fully benefiting from the available data. More critically, neither of these methods can use additional interaction network knowledge as prior biological knowledge in their framework.

We propose a new Bayesian framework for supervised domain adaptation for NGS count data, with generative models utilizing both data and label information from multiple domains to learn shared genes embedding and domain and label-dependent latent parameters. Through a hierarchical Bayesian structure and a factorization setup of parameters with a subset of global random variables, useful information from all the domains and labels can be leveraged for cancer subtype prediction in the target domain. The domains can include data from the same labels as or different labels than the target domain. We use negative binomial likelihoods to model RNA-Seq count data considering potential sample heterogeneity to obviate the need for *ad-hoc* preprocessing steps. The predictor in our method is based on the concept of optimal Bayesian operator design (Dalton and Dougherty, 2020), where the predictor is derived point-wise by comparing the posterior expectation of the class-conditional likelihoods for a given sample. Moreover, our framework can take advantage of the available prior knowledge in terms of gene-gene interaction networks to derive more accurate and generalizable predictors in the target domain.

In the following sections, we first introduce our basic OBSDA model and derive an efficient Gibbs sampler by exploiting novel data augmentation techniques for the negative binomial distribution (Zhou and Carin, 2015). Then, we propose an extension of OBSDA with flexible semi-implicit variational inference (Yin and Zhou, 2018)—SI-OBSDA—that employs explicit distributions mixed with implicit neural network generators. We then show how we can incorporate prior interaction network knowledge in SI-OBSDA for informed inference. Finally, we verify the benefits of our OBSDA and SI-OBSDA by providing results for comparing our methods with single-domain and multi-domain baselines on predicting cancer subtypes with The Cancer Genome Atlas (TCGA) RNA-Seq data.

2 Materials and Methods

2.1 OBSDA

The negative binomial (NB) distribution is a popular choice to model overdispersion in RNA-Seq count data due to technical and biological variations (Robinson *et al.*, 2010; Dadaneh *et al.*, 2018). Let $\mathbf{x} \sim \text{NB}(r, p)$, which is a NB distribution with the probability mass function (PMF) $\frac{\Gamma(\mathbf{x}+r)}{\mathbf{x}!\Gamma(r)}(p)^{\mathbf{x}}(1-p)^r$ with the count data $\mathbf{x} \in \{0, 1, 2, \cdots\}$ and $\Gamma(\cdot)$ being the gamma function. Denoting the observed count for gene j in sample i of domain d with label l by $\mathbf{x}^l_{d,j,i}$, and the collection of all genes for that sample by $\mathbf{x}^l_{d,i}$, we model the counts from multiple domains (sources) by a factorization of the parameters as

$$\mathbf{x}^l_{d,i} \sim \text{NB}(\mathbf{\Phi}\boldsymbol{\theta}^l_d, p^l_{d,i}). \tag{1}$$

Here, $\mathbf{\Phi} \in \mathbb{R}^+_{J\times K}$, with rows $\boldsymbol{\phi}^T_j \in \mathbb{R}^+_{1\times K}$ for $j = \{1, \cdots, J\}$, is the matrix quantifying the association between the genes and latent factors. This association is gene dependent, but for each domain and label the relevancy of the factors is different. The relevancy of the factors to each domain and label is quantified by $\boldsymbol{\theta}^l_d$. We model each element of $\boldsymbol{\theta}^l_d$ with a Gamma distribution, $\boldsymbol{\theta}^l_{d,k} \sim \text{Gamma}(u_{d,k}, \frac{1}{v^l})$, where v^l is label dependent and $u_{d,k}$ is domain dependent. In other words, the domain and label dependencies are decomposed into the two sets of parameters to help identifiability and share signals across domains and labels. The Gamma distribution encourages sparsity in the model, where each class in each domain can select a few of latent factors as relevant. We place the Gamma prior on the label-dependent parameters v^l. To enable domain-dependent latent representations, we assume $u_{d,k} \sim \text{Gamma}(b_k, \frac{1}{q_d})$, where b_k and q_d represent the global latent factor and domain-specific parameters. $p^l_{d,i}$ accounts for the potential sample heterogeneity in a class of a domain.

Note that unlike factor analysis models (Rai and Daumé, 2009; Zhou, 2018; Hajiramezanali *et al.*, 2018) where the observations are factorized, here a latent variable of the model is factorized, and is learned jointly with

other latent variables in the model using the data from multiple domains. Moreover, we leverage the label information in a supervised setting in contrast with standard factor analysis.

As a factorization model, $\mathbf{x}^l_{d,j,i} \sim \text{NB}(\boldsymbol{\phi}_j^T \boldsymbol{\theta}^l_d, p^l_{d,i})$ can be augmented as $\mathbf{x}^l_{d,j,i} = \sum_{k=1}^K \mathbf{x}^l_{d,j,i,k}$, where $\mathbf{x}^l_{d,j,i,k} \sim \text{NB}(\boldsymbol{\phi}_{j,k}\boldsymbol{\theta}^l_{d,k}, p^l_{d,i})$, and the expected expression of gene j in sample i of domain d with class label l can be expressed as

$$\mathbb{E}[\mathbf{x}^l_{d,j,i}] = \big(\sum_{k=1}^K \boldsymbol{\phi}_{j,k}\boldsymbol{\theta}^l_d\big)\frac{p^l_{d,i}}{1-p^l_{d,i}}. \tag{2}$$

The expectation can be interpreted as the true abundance of a gene adjusted by potential data heterogeneity in a class of a domain, removing the need for *ad-hoc* normalization steps. More specifically, the true abundance is comprised of the contributions of all latent factors, where each contribution is encoded as the product of the association between a gene and a factor and the relevancy of that factor for the domain and class.

The factors can be seen as underlying biological processes or functional modules relating to or causing genotypic or phenotypic changes. K is the number of such factors considered in the model and is a hyperparameter. From the modeling perspective, the random variables corresponding to the association between the genes and the underlying biological processes (factors) are assumed to be the same across domains and labels. In other words, the contribution of each underlying biological process to the expression of a gene depends on both the gene and process relationship, which is encoded by a global variable and shared across domains and labels, and the relevancy of the process to the specific label/class in the domain, which is domain and label dependent and learned from data.

It is worth noting that the OBSDA model can be seen as sharing knowledge across the different labels in the same domain as well as across domains for more robust estimations. Moreover, it can integrate data from domains containing different labels, i.e. where a one-to-one correspondence between labels across domains does not exist. These properties will especially be helpful when the number of samples is low in the target domain.

We complete the model by placing conjugate priors for the rest of the parameters as follows:

$$\begin{aligned}
&\mathbf{x}^l_{d,j,i} \sim \text{NB}(\boldsymbol{\phi}_j^T\boldsymbol{\theta}^l_d, p^l_{d,i})\\
&\boldsymbol{\theta}^l_{d,k} \sim \text{Gamma}(u_{d,k}, \frac{1}{v^l}), \quad u_{d,k} \sim \text{Gamma}(b_k, \frac{1}{q_d})\\
&v^l \sim \text{Gamma}(e_0, \frac{1}{f_0}), \quad b_k \sim \text{Gamma}(\frac{\gamma_0}{K}, \frac{1}{c_0})\\
&q_d \sim \text{Gamma}(w_0, \frac{1}{u_0}), \quad (\phi_{1,k}, \cdots, \phi_{J,k}) \sim \text{Dir}(\eta, \cdots, \eta)\\
&p^l_{d,i} \sim \text{Beta}(g_0, h_0), \quad c_0 \sim \text{Gamma}(a_0, \frac{1}{d_0})\\
&\gamma_0 \sim \text{Gamma}(\alpha_0, \frac{1}{\beta_0}),
\end{aligned} \tag{3}$$

where we have exploited the beta-negative binomial, gamma-gamma, and gamma-Poisson conjugacy relationships. Efficient closed-form Gibbs updates are detailed in the Supplementary for OBSDA inference by adopting novel data augmentation techniques suitable to our model.

2.2 SI-OBSDA

We now extend OBSDA to SI-OBSDA, with the goal of incorporating gene-gene network information available *a priori* to have an *informed* inference mechanism. In OBSDA, to be able to derive closed-form updates, we are restricted to certain prior assumptions to take advantage of the appropriate data augmentation and conjugacy relationships. In SI-OBSDA, we want to impose prior constraints stemming from domain knowledge in the inference procedure. Hence, instead of resorting to Gibbs sampling for model inference, in SI-OBSDA we exploit semi-implicit variational inference (SIVI) (Yin and Zhou, 2018) as the base inference method, which is able to construct flexible variational families to approximate the actual posterior. We first describe the base inference mechanism in SI-OBSDA and then integrate the prior network knowledge.

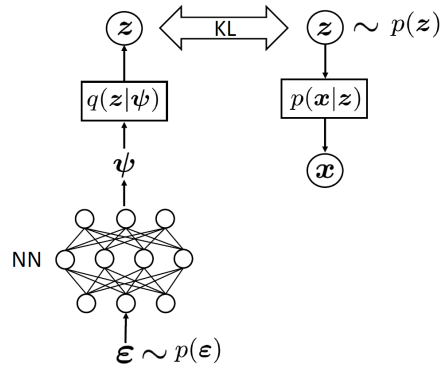

Fig. 1. Schematic diagram of semi-implicit variational inference in SI-OBSDA

Denoting the latent variables or parameters of interest as $\boldsymbol{z}$ and the observed data as $\mathbf{x}$ in a general Bayesian model, variational inference maximizes the evidence lower bound (ELBO), defined as

$$\mathcal{L} = \text{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\mathbf{x})}\big[p(\mathbf{x}|\boldsymbol{z})\big] - \text{KL}\big(q(\boldsymbol{z}|\mathbf{x})||p(\boldsymbol{z})\big),$$

where $q(\boldsymbol{z}|\mathbf{x})$ is the variational posterior selected from a tractable family of distributions and KL denotes the Kullback-Leibler divergence. To simplify the optimization of the ELBO, a commonly adopted choice of variational distributions is the family of factorized distributions. This is referred to as mean-field variational inference (MFVI) (Jordan *et al.*, 1999). However, MFVI can suffer from various shortcomings, including inability to capture multimodality in the posterior and underestimation of the posterior variance (Blei *et al.*, 2017).

Here in SI-OBSDA, $\boldsymbol{z}$ denotes the collection of previously described model parameters in OBSDA, including the association between genes and factors $\{\boldsymbol{\phi}_j\}_{j=1}^J$, factors' relevancy to domains and labels $\{\boldsymbol{\theta}^l_d\}^D_{d=1,l\in L_d}$, sample variability $\{p^l_{d,i}\}^{D,\quad N^l_d}_{d=1,l\in L_d,i=1}$, label parameters $\{\nu^l\}_{l\in\cup^D_{d=1}L_d}$, local factor popularity parameters for each domain $\{u_{d,k}\}^{D,K}_{d=1,k=1}$, global factor $\{b_k\}^K_{k=1}$ and domain parameters $\{q_d\}^D_{d=1}$, and hyperparameters c_0 and γ_0. We have used L_d, D, and N^l_d to denote the set of labels in domain d, the number of domains, and the number samples in domain d with label l, respectively.

To have more expressive variational families while maintaining computational tractability, in SI-OBSDA we employ SIVI and construct a model with an explicit joint distribution $p(\mathbf{x}, \boldsymbol{z})$ and a semi-implicit approximate posterior $q_{\boldsymbol{\omega}}(\boldsymbol{z})$ (Figure 1). In other words, the idea is to define the variational family in a hierarchical manner as $\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{\psi})$, where the conditional variational distribution is explicit but $\boldsymbol{\psi} \sim q_{\boldsymbol{\omega}}(\boldsymbol{\psi})$ is implicit and required to be reparameterizable. More specifically, samples from $q_{\boldsymbol{\omega}}$ can be generated by transforming random noise via a neural network to be more expressive for modeling $\mathbf{x}$. It is clear that the marginal inferred posteriors are not independent as in the standard variational inference, and posterior dependence can be captured.

In SI-OBSDA, we place reparameterizable (location-scale) variational distributions on the parameters. For the parameters in $\mathbb{R}^+$ and $(0, 1)$, we use log-normal (log N) and logistic-normal (logit N) distributions, respectively. For $\{\boldsymbol{\phi}_j\}_{j=1}^J$, in SI-OBSDA we assume logistic-normal prior and variational distributions. This resolves the optimization issue in

the simplex while potentially increasing model flexibility. The joint log-likelihood of SI-OBSDA can be found in the Supplementary. We place the following reparameterizable variational distributions in our model inference for SI-OBSDA:

$$
\begin{aligned}
q(\boldsymbol{z}|\boldsymbol{\psi},\boldsymbol{\xi}) = &\prod_{d,l,k} \log \mathrm{N}(\theta^l_{d,k}; \hat{\mu}_{\theta^l_{d,k}}, \hat{\sigma}^2_{\theta^l_{d,k}}) \prod_j \mathrm{logit}\,\mathrm{N}(\boldsymbol{\phi}_j; \hat{\boldsymbol{\mu}}_{\boldsymbol{\phi}_j}, \hat{\Sigma}_{\boldsymbol{\phi}_j}) \\
&\prod_l \log \mathrm{N}(\nu^l; \hat{\mu}_{\nu^l}, \hat{\sigma}^2_{\nu^l}) \prod_{d,k} \log \mathrm{N}(u_{d,k}; \hat{\mu}_{u_{d,k}}, \hat{\sigma}^2_{u_{d,k}}) \\
&\prod_d \log \mathrm{N}(q_d; \hat{\mu}_{q_d}, \hat{\sigma}^2_{q_d}) \prod_k \log \mathrm{N}(b_k; \hat{\mu}_{b_k}, \hat{\sigma}^2_{b_k}) \\
&\prod_{d,l,i} \mathrm{logit}\,\mathrm{N}(p^l_{d,i}; \hat{\mu}_{p^l_{d,i}}, \hat{\sigma}^2_{p^l_{d,i}}) \\
&\log \mathrm{N}(c_0; \hat{\mu}_{c_0}, \hat{\sigma}^2_{c_0}) \log \mathrm{N}(\gamma_0; \hat{\mu}_{\gamma_0}, \hat{\sigma}^2_{\gamma_0}).
\end{aligned} \tag{4}
$$

For inference, we optimize an asymptotically exact surrogate evidence lower bound (ELBO) (Yin and Zhou, 2018):

$$
\begin{aligned}
\underline{\mathcal{L}}_{\tilde{M}} = \mathrm{E}_{q(\boldsymbol{z}|\boldsymbol{\psi})q_{\boldsymbol{\omega}}(\boldsymbol{\psi})} &\mathrm{E}_{\boldsymbol{\psi}^{(1)},\cdots,\boldsymbol{\psi}^{(\tilde{M})}\sim q_{\boldsymbol{\omega}}(\boldsymbol{\psi})} \\
&\Big[\log \frac{p(\mathbf{x}, \boldsymbol{z}_i)}{\frac{1}{\tilde{M}+1}\big[q(\boldsymbol{z}_i|\boldsymbol{\psi}_i) + \sum_{m=1}^{\tilde{M}} q(\boldsymbol{z}_i|\boldsymbol{\psi}^{(m)})\big]}\Big],
\end{aligned} \tag{5}
$$

where we have $\lim_{\tilde{M}\to\infty} \underline{\mathcal{L}}_{\tilde{M}} =$ ELBO. In practice, $\boldsymbol{\psi}^{(m)} = T_{\boldsymbol{\omega}}(\boldsymbol{\epsilon}^{(m)})$, where $\boldsymbol{\epsilon}^{(m)} \sim q(\boldsymbol{\epsilon})$, with $q(\boldsymbol{\epsilon})$ being the source of randomness and $T_{\boldsymbol{\omega}}$ a deep neural network (Figure 1). The variational distribution can have additional variational parameters $\boldsymbol{\xi}$, not mixed with another distribution, i.e. we have $q(\boldsymbol{z}|\boldsymbol{\psi},\boldsymbol{\xi})$. Denoting the reparameterization of $q(\boldsymbol{z}|\boldsymbol{\psi},\boldsymbol{\xi})$ as $\boldsymbol{z} = f(\boldsymbol{\varepsilon},\boldsymbol{\xi},\boldsymbol{\psi})$, $\boldsymbol{\varepsilon} \sim p(\boldsymbol{\varepsilon})$, where $p(\boldsymbol{\varepsilon})$ is the source of randomness, $\boldsymbol{z}$ can be sampled by $\boldsymbol{z}_i = f(\boldsymbol{\varepsilon}_i,\boldsymbol{\xi},\boldsymbol{\psi}_i)$, $\boldsymbol{\varepsilon}_i \sim p(\boldsymbol{\varepsilon})$. The parameters of the mixing distribution and the variational parameters can be optimized by gradient ascent:

$$
\begin{aligned}
\boldsymbol{\xi} &= \boldsymbol{\xi} + \rho_t \nabla_{\boldsymbol{\xi}} \underline{\mathcal{L}}_{\tilde{M}}\Big(\{\boldsymbol{\psi}^{(m)}\}, \{\boldsymbol{\psi}_i\}, \{\boldsymbol{z}_i\}\Big), \\
\boldsymbol{\omega} &= \boldsymbol{\omega} + \upsilon_t \nabla_{\boldsymbol{\omega}} \underline{\mathcal{L}}_{\tilde{M}}\Big(\{\boldsymbol{\psi}^{(m)}\}, \{\boldsymbol{\psi}_i\}, \{\boldsymbol{z}_i\}\Big).
\end{aligned} \tag{6}
$$

In SI-OBSDA, we consider the collection of $\{\hat{\mu}_{\theta^l_{d,k}}\}^{D,\ \ K}_{d=1,l\in L_d,k=1}$, $\{\hat{\boldsymbol{\mu}}_{\boldsymbol{\phi}_j}\}^J_{j=1}$, $\{\hat{\mu}_{\nu^l}\}_{l\in\cup^D_{d=1}L_d}$, $\{\hat{\mu}_{b_k}\}^K_{k=1}$, $\{\hat{\mu}_{u_{d,k}}\}^{D,K}_{d=1,k=1}$, $\{\hat{\mu}_{q_d}\}^D_{d=1}$, $\hat{\mu}_{c_0}$, and $\hat{\mu}_{\gamma_0}$ to be the parameters governed by the mixing distribution of $\boldsymbol{\psi}$, and $\{\hat{\mu}_{p^l_{d,i}}\}^{D,\ \ N^l_d}_{d=1,l\in L_d,i=1}$, $\{\hat{\sigma}_{p^l_{d,i}}\}^{D,\ \ N^l_d}_{d=1,l\in L_d,i=1}$, $\{\hat{\sigma}_{\theta^l_{d,k}}\}^{D,\ \ K}_{d=1,l\in L_d,k=1}$, $\{\hat{\Sigma}_{\boldsymbol{\phi}_j}\}^J_{j=1}$, $\{\hat{\sigma}_{\nu^l}\}_{l\in\cup^D_{d=1}L_d}$, $\{\hat{\sigma}_{b_k}\}^K_{k=1}$, $\{\hat{\sigma}_{u_{d,k}}\}^{D,K}_{d=1,k=1}$, $\{\hat{\sigma}_{q_d}\}^D_{d=1}$, $\hat{\sigma}_{c_0}$, and $\hat{\sigma}_{\gamma_0}$ as the variational parameters ($\boldsymbol{\xi}$). For numerical stability we further reparameterize the variational parameters by log-transform and Cholesky factorization. Implementation details of SI-OBSDA is included in the Supplementary.

In SI-OBSDA, similar to the SIVI inference mechanism in Yin and Zhou (2018), we employ a neural network as $T_{\boldsymbol{\omega}}$ for the mixing distribution. Since neural networks have high modeling capacity, $q_{\boldsymbol{\omega}}(\boldsymbol{\psi})$ can be highly flexible, and the dependencies between the elements of $\boldsymbol{\psi}$ can be well captured. Moreover, from the implementation perspective, neural networks can easily leverage automatic differentiation to optimize the surrogate ELBO in (5), which is computationally desirable.

2.3 Incorporating prior network knowledge in SI-OBSDA

In addition to the expression data, there exists *a priori* interactome knowledge such as gene-gene interaction network that contains genome-scale connectivity information (Menche *et al.*, 2015). These can be derived based on either regulatory, metabolic, signaling interactions, or protein binding.

In SI-OBSDA we impose constraints stemming from the prior knowledge in the gene-factor associations to construct informed latent representations and inference. More specifically, since the factors can be interpreted as functional modules or underlying biological processes, intuitively, the genes that are connected in the prior knowledge network should have closer associations to the underlying factors. Hence, we impose proximity constraints on the variables quantifying the association between genes and factors for genes that are connected in the prior knowledge network. Specifically, we add a regularization term coming from prior belief to the objective of the SI-OBSDA:

$$
\begin{aligned}
&\mathcal{L}_{\text{SI-OBSDA}} = \underline{\mathcal{L}}_{\tilde{M}} + \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{\psi},\boldsymbol{\xi})}\mathcal{L}_{\text{pr}}, \\
&\text{where} \quad \mathcal{L}_{\text{pr}} = \sum_{j=1}^{J} \sum_{\tilde{j}\in\mathcal{C}_j, \tilde{j}<j} \lambda_{j,\tilde{j}} ||\boldsymbol{\phi}_j - \boldsymbol{\phi}_{\tilde{j}}||.
\end{aligned} \tag{7}
$$

In the equation above, $\mathcal{C}_j$ denotes the set of genes that are connected to gene j in the prior network knowledge.

The proposed additive constraints when optimizing for inference fit in the MKDIP prior-construction framework of Boluki *et al.* (2017), with the expectation taken over the variational distribution. More specifically, we can consider slackness for the prior constraints which are linearly added to the objective, i.e. the regularization term acts as a relaxation of the constraints coming from prior knowledge with $\lambda_{j,\tilde{j}}$ encoding the degree of belief in the specific prior interaction edge. In other words, the higher the confidence in an edge is in prior knowledge, the larger $\lambda_{j,\tilde{j}}$ will be set.

Another way to interpret the regularization term is through assuming (conditional) prior distributions that impose these constraints in effect. Moreover, although different in nature, it is worth noting that our work has connections with recent works including Dadaneh *et al.* (2020), where additional label information is imposed through proximity constraints in the latent space and has been shown to be beneficial even on large data.

2.4 Classification with OBSDA and SI-OBSDA

In the previous sections, we have introduced the models and inference procedures for OBSDA and SI-OBSDA. Here, we describe how classification for subtyping is done based on the inferred Bayesian models. The classification operator in OBSDA and SI-OBSDA is based on the optimal Bayesian classification (OBC) framework (Dalton and Dougherty, 2020, 2013; Karbalayghareh *et al.*, 2018). In OBC, the design of the classifier is based on the posterior marginalization of the class-conditional feature distributions, called effective class-conditional distributions. This is in contrast to *plug-in* classifier design where the estimates of the parameters are used to calculate the class-conditional distributions to form the classifier, which may not result in the optimal expected error relative to the posterior distributions, especially when the posteriors are multi-modal. More specifically, denoting the collection of all model parameters and the posteriors after observing data as Θ and π^*, respectively, OBC classifier (f_{obc}) satisfies

$$
\mathbb{E}_{\pi^*}[\delta(f_{\text{obc}}, \Theta)] \leq \mathbb{E}_{\pi^*}[\delta(f, \Theta)], \qquad \forall f \in F, \tag{8}
$$

where f and F denote a classifier and all classifiers possessing measurable decision regions, respectively; and $\delta(\cdot,\cdot)$ is the error for fixed parameter values and a classification rule.

In OBSDA and SI-OBSDA, we can derive the optimal Bayesian classifier in the target domain (OBTD) based on the samples of the parameters of the target domain generated in the inference chain of OBSDA or from the variational posteriors in SI-OBSDA. Note that this is

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

equivalent to marginalizing the joint posterior over the source domain(s) as in Karbalayghareh *et al.* (2018).

Denoting the class prior probabilities in the target domain ($d = t$, and without loss of generality assuming the labels are from 1 to L_t) as $\boldsymbol{c}_t = (c_t^1, \cdots, c_t^{L_t})$, and given the parameters of the model, the probability of sample $\mathbf{x}_{t,i}$ belonging to class l is equal to

$$p(l|\mathbf{x}_{t,i}) = \frac{c_t^l p(\mathbf{x}_{t,i}|\boldsymbol{\Phi}, \boldsymbol{\theta}_t^l, p_{t,i}^l)}{\sum_{\tilde{l}=1}^{L_t} c_t^{\tilde{l}} p(\mathbf{x}_{t,i}|\boldsymbol{\Phi}, \boldsymbol{\theta}_t^{\tilde{l}}, p_{t,i}^{\tilde{l}})}, \tag{9}$$

where $p(\mathbf{x}_{t,i}|\boldsymbol{\Phi}, \boldsymbol{\theta}_t^l, p_{t,i}^l) = \prod_{j=1}^{J} \text{NB}(\mathbf{x}_{t,j,i}|\boldsymbol{\phi}_j^T \boldsymbol{\theta}_t^l, p_{t,i}^l)$. Hence, the optimal Bayesian classifier in the target domain (OBTD) is:

$$f_{\text{OBTD}}(\mathbf{x}_{t,i}) = \arg \max_{l \in \{1,\cdots,L_t\}} \text{E}_{\pi^*}\left[c_t^l p(\mathbf{x}_{t,i}|\boldsymbol{\Phi}, \boldsymbol{\theta}_t^l, p_{t,i}^l)\right]. \tag{10}$$

Assuming that the class prior probabilities in the target domain are independent of the other model parameters *a priori* and have a Dirichlet prior $(c_t^1, \cdots, c_t^{L_t}) \sim \text{Dir}(\eta_t^1, \cdots, \eta_t^{L_t})$, we have

$$f_{\text{OBTD}}(\mathbf{x}_{t,i}) = \arg \max_{l \in \{1,\cdots,L_t\}} \mathbb{E}_{\pi^*}\left[c_t^l\right] \mathbb{E}_{\pi^*}\left[p(\mathbf{x}_{t,i}|\boldsymbol{\Phi}, \boldsymbol{\theta}_t^l, p_{t,i}^l)\right], \tag{11}$$

where

$$\mathbb{E}_{\pi^*}\left[c_t^l\right] = \frac{|\mathbf{x}_t^l| + \eta_t^l}{\sum_{\tilde{l}=1}^{L_t} |\mathbf{x}_t^{\tilde{l}}| + \eta_t^{\tilde{l}}}. \tag{12}$$

$|\mathbf{x}_t^l|$ denotes the number of training samples in the target domain t with label l.

Given the training data, OBSDA generates samples from the posteriors of the parameters via the Gibbs chain. Similarly, in SI-OBSDA when the optimization of the training loss is stopped, samples from the posterior can be generated by pushing random noise samples through the trained neural network and in turn using the outputs as parameters for sampling from the variational posteriors. We collect these samples (or save the neural network in SI-OBSDA) in the training procedure and use them at test time. When a new unlabeled test data i comes in, we only need to generate posterior samples for $p_{t,i}^l$ corresponding to the collected posterior samples for $\boldsymbol{\theta}_t^l$ by (12) in the Supplementary to predict the label for the data point by (10).

2.5 A note on experimental setup for performance evaluation

In the next section, we describe in details the different datasets, experiment setups, and baselines we have used for comparison analysis. We would like to emphasize that in each Monte Carlo run, the training and test sets are completely independent and do not share any samples/patients. The methods that need hyperparameter tuning only use the training set for doing so.

3 Results

3.1 Data

We evaluate the performance of our OBSDA and SI-OBSDA for subtyping lung cancer using several RNA-Seq datasets from The Cancer Genome Atlas (TCGA) (Hutter and Zenklusen, 2018). In our experiments, we consider RNA-Seq data from two subtypes of non-small cell lung cancer (NSCLC), lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) as the target domain. According to the American Cancer Society statistics, lung cancer is the second most commonly diagnosed cancer and the leading cause of cancer death in both men and women in the United States. About 84% of lung cancers are NSCLC and LUAD and LUSC combined account for about 70% of lung cancers.

We examine the target lung cancer subtyping accuracy by ours and other competing methods, focusing on evaluating their performances when using additional RNA-Seq data from three different source domains that either share the same class labels with or have different ones from the target domain. Specifically, we take RNA-SeqV2 dataset, which is from the second analysis pipeline, for LUAD and LUSC as the first source domain, RNA-Seq data from Head and Neck Squamous Cell Carcinoma (HNSC) as the second source domain, and data from the two most common types of kidney cancers, kidney renal clear cell carcinoma (KIRC) and kidney renal papillary cell carcinoma (KIRP) as the third source domain. Clearly, the degree to which the source domain may help lung cancer subtyping vary for these three different source domains. One is from the data with the same subtypes but different NGS pipelines, while the other two are from studies concerning different cancer types with one and two classes in each domain.

For SI-OBSDA we use the gene-gene network containing only physical interactions (the human interactome) archived in Menche *et al.* (2015) as the network prior knowledge. The network, which features 13460 proteins interconnected by 141296 interactions, does not include interactions extracted from gene expression data, and has been compiled by combining experimental support from several databases including protein-protein and regulatory interactions, signaling interactions, metabolic pathway interactions, and kinase-substrate interactions. In the experiments, we consider equal weights for the edges in SI-OBSDA, and set them to either 1 or 0.25 based on the accuracy of the inferred model on the training data. For SI-OBSDA, in all the experiments we take $\boldsymbol{\epsilon}$ to have the same cardinality of $\boldsymbol{\psi}$, and $T_{\boldsymbol{\omega}}(\boldsymbol{\epsilon})$ as a neural network with three hidden layers (more implementation details available in the Supplementary).

In the following experiments, we first pick the common genes within the target and source datasets and the prior network knowledge, resulting in 11839 genes. We then remove the genes that have total read counts of less than 40 across the LUAD and LUSC samples in the target domain. Finally, we perform differential expression analysis with DESeq2 (Love *et al.*, 2014) and select 500 out of the top 2500 genes with the highest log-fold change (with gaps of 5) in each experimental run for all the methods for fair comparison.

3.2 Baselines

As the baselines for comparing lung cancer subtyping accuracy, we apply SVM (with both Gaussian and linear kernels), regularized linear SVM, and regularized logistic regression on the data from the target domain. We also use a neural network (NN) classifier as an additional baseline. The architecture of the network is kept the same as the neural network utilized in the inference mechanism of SI-OBSDA (explained in detail in the Supplementary) to have a fair comparison for evaluating the proposed models. The only architectural difference is that the NN classifier takes the expression data as input and outputs the logit (log-odds). In the first setup with the source domain having the same labels as the target domain, we train these baselines once only using the training data in the target domain, and once using the collection of source and target training data. We tune the hyperparameters of each baseline classifier in each run given the training data with Bayesian optimization (Shahriari *et al.*, 2015; Frazier, 2018) and the cross-validation loss as the objective function.

To compare the performance of our method in terms of domain adaptation and learning useful information from source domains for designing a predictor in the target domain, there are two other methods that can provide good comparisons that can be applied for domain adaptation and transfer learning on NGS count data for comparisons. Optimal Bayesian transfer learning (OBTL) (Karbalayghareh *et al.*, 2018, 2019) is a supervised transfer learning method that models the relationship between the same classes across domains by assuming joint priors and

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

marginalizing the joint posterior over the source domain parameters. Unfortunately, this method is not scalable to more than 10 to 20 genes, so we could not perform comparisons with it. BMDL (Hajiramezanali *et al.*, 2018) is a multi-domain learning method that projects the data from different domains to a lower dimensional common embedding space, and applies a classifier on the projected space. It has been shown that BMDL outperforms other similar Bayesian latent models on the NGS classification problem. Thus, we choose BMDL as the state-of-the-art baseline for our experiments on domain adaption for RNA-Seq data.

3.3 Results and discussion

3.3.1 LUAD and LUSC data in source and target domains

In this setup, we compare the performance of different methods when the source and target domains have data from the same cancer subtypes. The target domain contains 162 and 240 samples from LUAD and LUSC, respectively. In each run, we randomly pick 20 samples in total from the target domain for training by stratified sampling, and use the rest of the samples in the target domain for testing. The source domain contains 414 and 312 samples from LUAD and LUSC, respectively, where we perform stratified sampling (considering the source proportions) for different number of training samples from the source domain. We investigate the performance of OBSDA, BMDL, regularized logistic regression (Reg Log), regularized linear SVM (Reg SVM), kernel SVM (SVM), and neural network classifier (NN) using three different numbers of source samples, 564, 112, and 11. This setup covers a wide range of source samples, from a few training samples from source (nearly half of target training samples) to around $5.5\times$ and $28\times$ the number of target samples in the training data. Note that in this experiment, since the labels are the same across domains, we train the single-domain baseline methods once utilizing the collection of all the training data from both domains and once only the target domain's training data.

The results in Figure 2 show that OBSDA achieves the best performance compared with the baselines by effectively borrowing information from the source data. We can see that OBSDA's error in classifying subtypes in the target domain consistently decreases as the number of source samples increases. On the contrary, BMDL seems to suffer when the source samples drastically dominate the target samples in the training data, which is undesirable for domain adaptation. We can also observe this adverse effect of having a lot more source samples than target samples in the training data on the NN classifier, where the results show that the proposed methods outperform the NN classifier for all the numbers of source samples. This confirms that neural networks are not specifically fit to use on smaller datasets and indicates that explicitly modeling for learning useful information from other domains for the target domain is required when facing smaller (target) sample sizes.

Next, we test the performance of SI-OBSDA that incorporates constraints on the latent space stemming from the prior knowledge within a flexible variational inference in this experiment setup. As seen in Figure 2, similar to OBSDA, SI-OBSDA's error also consistently decreases as the number of source samples increases. The results in Table 1 show around 1% to 3% improvement compared with OBSDA and 4% to 5% difference from BMDL, demonstrating that SI-OBSDA can achieve the best performance by incorporating prior knowledge as well as learning useful information across domains.

It is worth noting that SI-OBSDA and OBSDA also show relatively lower variance across the experimental runs, i.e. a more robust performance, compared with the other methods.

3.3.2 LUAD and LUSC data only in the target domain

In this section, we examine the performance of different methods using data from source domains that do not have labels in common with the data from the target domain. We consider HNSC data as one source domain and kidney cancer data (KIRC and KIRP) as another source domain. The HNSC dataset contains 294 samples, and the kidney cancer dataset consists of 537 KIRC and 14 KIRP samples. We have selected these datasets from different cancer types as the source domain since the degree to which they may help detecting the lung cancer subtypes may be different due to the different disease mechanisms. Moreover, another difference is the number of labels in each source domain with one domain only containing data with one label (HNSC), and the other containing data with two labels (KIRC and KIRP). Similar to the previous section, in each Monte Carlo run we do stratified sampling for training data from the target domain, randomly picking 20 training samples from the target domain. For the lower and higher number of source samples ($N_S = 11$ and $N_s = 112$), two random or all the 14 KIRP samples are selected for training, respectively, with the rest of the source training samples coming from KIRC.

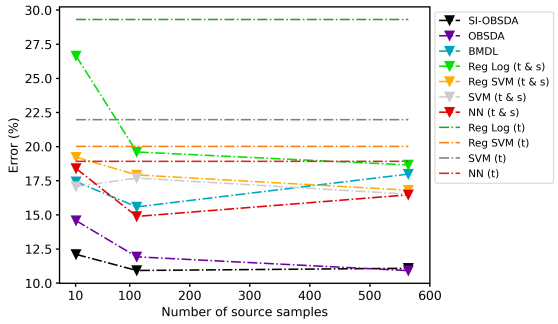

Fig. 2. Average performance of different methods in identifying cancer subtypes of LUAD vs LUSC using different number of source samples. (t) and (t & s) correspond to using only target samples, and source and target samples in training, respectively.

Table 1. Average errors (in % $\pm$ standard deviations) in identifying subtypes of LUAD vs LUSC with the source domain containing samples from the same subtypes.

Method	$N_s = 11$	$N_s = 112$
SI-OBSDA	12.10 ± 0.81	10.92 ± 0.47
OBSDA	14.57 ± 0.64	11.91 ± 1.09
BMDL	17.42 ± 1.66	15.58 ± 1.19
Reg Log (t & s)	26.63 ± 2.92	19.60 ± 3.18
Reg SVM (t & s)	19.22 ± 5.64	17.92 ± 1.56
SVM (t & s)	17.07 ± 4.53	17.69 ± 1.23
NN (t & s)	18.39 ± 3.63	14.89 ± 1.33
Reg Log (t)	29.31 ± 4.41	29.31 ± 4.41
Reg SVM (t)	20.01 ± 2.57	20.01 ± 2.57
SVM (t)	21.97 ± 2.67	21.97 ± 2.67
NN (t)	18.91 ± 3.26	18.91 ± 3.26

The results in Table 2 demonstrate that both SI-OBSDA and OBSDA outperform BMDL when the source domain contains data of different cancers from the target domain by close to 5% to 7% under different settings. We can attribute this to BMDL not leveraging label information in the latent representation learning stage. Comparing the numbers in Tables 2 and 1, we see that all the methods that use data from both source and target domains still perform better than the other baselines using only the target domain data in training. Similar to the previous experiment, SI-OBSDA, which leverages the prior network knowledge in addition to the expression data within its flexible variational inference, achieves the best accuracy in classifying subtypes in the target domain. It is interesting to note that OBSDA and SI-OBSDA both benefit from more samples from the source domain in training, even though they are from different cancer types. This

verifies the benefit of our proposed approach in modeling that can borrow useful information from other domains and labels for the prediction task in the target domain. Also, the results in Tables 2 and 1 show that, as expected, when the source contains data from the same labels as the target domain, SI-OBSDA and OBSDA generally achieve better accuracy for the same number of source samples used in training. Additionally, when the data from the source are for different cancers from the target domain, the decrease in prediction error in the target domain is slower when increasing the number of source samples, compared with the case of source domain containing data from the same disease.

Table 2. Average errors (in % $\pm$ standard deviations) in identifying subtypes of LUAD vs LUSC with the source domain containing samples from different labels.

Source sample size	$N_s = 11$	$N_s = 112$
Source domain	HNSC	
SI-OBSDA	12.56 ± 0.87	11.85 ± 0.77
OBSDA	13.48 ± 0.95	13.02 ± 0.47
BMDL	17.32 ± 3.38	17.75 ± 3.13
Source domain	KIRC,KIRP	
SI-OBSDA	12.17 ± 0.88	12.23 ± 0.65
OBSDA	14.59 ± 1.70	14.20 ± 0.67
BMDL	19.81 ± 1.76	17.82 ± 2.33

3.3.3 Effect of incorporating prior knowledge

The results in the previous experiments showed that SI-OBSDA, which takes advantage of flexible variational posteriors and the gene-gene network prior knowledge, outperforms OBSDA and the baselines. Here, we examine the effect of the incorporation of the constraints coming from prior knowledge within the inference optimization on the performance of SI-OBSDA. Table 3 shows the results of SI-OBSDA with and without using prior knowledge for the different settings of source domain and number of source samples. The results suggest that SI-OBSDA generally benefits from the prior network knowledge by varying degrees for different setups. Note that by comparing the numbers in Table 3 with the numbers in Tables 1 and 2, we see that without incorporating the prior constraints on the latent space, SI-OBSDA attains errors that are still comparable or slightly lower than OBSDA in most cases while being better than BMDL by 4% to 7%.

4 Conclusions

In this paper, we propose a new Bayesian domain adaptation framework for leveraging labeled data from other domains for next-generation sequencing (NGS) count data, and develop OBSDA with an efficient Gibbs sampler. Compared to existing methods for domain adaptation and transfer learning, OBSDA has the following features: It uses label information across domains for transfer learning compared with unsupervised models. It models the relationship between different domains as well as different classes in one domain, contrasting with existing supervised methods that are restricted to the cases requiring domains having the same labels. It can leverage data from domains containing no common labels with no negative effect on the learning task for the target domain. In addition, when analyzing NGS data, it does not need any *ad-hoc* normalization of the counts due to its generative nature.

Moreover, we introduce SI-OBSDA, where flexible variational distributions are formed by using neural networks as an implicit generator. We propose incorporating prior knowledge in terms of gene-gene network connectivity as constraints imposed on the latent embedding to construct informed approximate posteriors to improve the performance.

Our experiments on the real-world RNA-Seq data show that by sharing information across domains and labels, OBSDA achieves the best cancer subtype identification performance compared with methods using only target domain data and other methods that try to use all the domains' data. Additionally, the results show that by incorporating the prior knowledge, SI-OBSDA can further improve the subtype identification accuracy. Incorporating more diverse prior knowledge in a principled way for transfer learning and domain adaptation is a promising direction for further exploration in our future work.

Table 3. Comparison of SI-OBSDA and SI-OBSDA without prior knowledge (SI-OBSDA w/o Prior) in terms of average errors (in %) in identifying subtypes of LUAD vs LUSC with different source domain settings.

Method		**SI-OBSDA**	**SI-OBSDA w/o Prior**
Lung source data	$N_s = 11$	12.10	13.09
	$N_s = 112$	10.92	12.04
HNSC source data	$N_s = 11$	12.56	13.28
	$N_s = 112$	11.85	12.83
Kidney source data	$N_s = 11$	12.17	12.90
	$N_s = 112$	12.23	13.02

Funding

This work was supported by the National Science Foundation [CCF-1553281]; and the U.S. Department of Energy [DE-SC0019393].

References

Alizadeh, A. A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., Caldas, C., Califano, A., Doherty, M., Elsner, M., *et al.* (2015). Toward understanding and exploiting tumor heterogeneity. *Nature medicine*, **21**(8), 846.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A., Kerrien, S., Khadake, J., *et al.* (2010). The IntAct molecular interaction database in 2010. *Nucleic acids research*, **38**(suppl_1), D525–D531.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, **112**(518), 859–877.

Boluki, S., Esfahani, M. S., Qian, X., and Dougherty, E. R. (2017). Incorporating biological prior knowledge for Bayesian learning via maximal knowledge-driven information priors. *BMC bioinformatics*, **18**(14), 552.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, **22**(14), e49–e57.

Dadaneh, S. Z., Qian, X., and Zhou, M. (2018). Bnp-seq: Bayesian nonparametric differential expression analysis of sequencing count data. *Journal of the American Statistical Association*, **113**(521), 81–94.

Dadaneh, S. Z., Boluki, S., Yin, M., Zhou, M., and Qian, X. (2020). Pairwise supervised hashing with Bernoulli variational auto-encoder and self-control gradient estimator. *arXiv preprint arXiv:2005.10477*.

Dai, W., Yang, Q., Xue, G.-R., and Yu, Y. (2007). Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, pages 193–200.

Dalton, L. A. and Dougherty, E. R. (2013). Optimal classifiers with minimum expected error within a bayesian framework—part i: Discrete and gaussian models. *Pattern Recognition*, **46**(5), 1301–1314.

Dalton, L. A. and Dougherty, E. R. (2020). *Optimal Bayesian Classification*. SPIE Press.

Dhruba, S. R., Rahman, R., Matlock, K., Ghosh, S., and Pal, R. (2018). Application of transfer learning for cancer drug sensitivity prediction. *BMC bioinformatics*, **19**(17), 497.

Frazier, P. I. (2018). A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Garcia-Vaquero, M. L., Gama-Carvalho, M., De Las Rivas, J., and Pinto, F. R. (2018). Searching the overlap between network modules with specific betweeness (S2B) and its application to cross-disease analysis. *Scientific reports*, **8**(1), 1–10.

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Gong, B., Grauman, K., and Sha, F. (2013). Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 222–230.

Gustafsson, M., Nestor, C. E., Zhang, H., Barabási, A.-L., Baranzini, S., Brunak, S., Chung, K. F., Federoff, H. J., Gavin, A.-C., Meehan, R. R., *et al.* (2014). Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Genome medicine*, **6**(10), 1–11.

Hajiramezanali, E., Dadaneh, S. Z., Karbalayghareh, A., Zhou, M., and Qian, X. (2018). Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data. In *Advances in Neural Information Processing Systems*, pages 9115–9124.

Hutter, C. and Zenklusen, J. C. (2018). The Cancer Genome Atlas: Creating lasting value beyond its data. *Cell*, **173**(2), 283–285.

Jacob, L., Vert, J.-p., and Bach, F. R. (2009). Clustered multi-task learning: A convex formulation. In *Advances in neural information processing systems*, pages 745–752.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, **37**(2), 183–233.

Kang, Z., Grauman, K., and Sha, F. (2011). Learning with whom to share in multi-task feature learning. In *ICML*, volume 2, page 4.

Karbalayghareh, A., Qian, X., and Dougherty, E. R. (2018). Optimal bayesian transfer learning. *IEEE Transactions on Signal Processing*, **66**(14), 3724–3739.

Karbalayghareh, A., Qian, X., and Dougherty, E. R. (2019). Optimal Bayesian transfer learning for count data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Levine, D. A. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**(7447), 67–73.

Liu, M.-Y. and Tuzel, O. (2016). Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477.

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**(12), 550.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**(6224).

Normand, R., Du, W., Briller, M., Gaujoux, R., Starosvetsky, E., Ziv-Kenet, A., Shalev-Malul, G., Tibshirani, R. J., and Shen-Orr, S. S. (2018). Found in translation: a machine learning model for mouse-to-human inference. *Nature methods*, **15**(12), 1067–1073.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, **22**(10), 1345–1359.

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2010). Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, **22**(2), 199–210.

Passos, A., Rai, P., Wainer, J., and Daume III, H. (2012). Flexible modeling of latent task structures in multitask learning. *arXiv preprint arXiv:1206.6486*.

Patel, V. M., Gopalan, R., Li, R., and Chellappa, R. (2015). Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, **32**(3), 53–69.

Rai, P. and Daumé, H. (2009). The infinite hierarchical factor regression model. In *Advances in Neural Information Processing Systems*, pages 1321–1328.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. (2015). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, **104**(1), 148–175.

Stark, C., Breitkreutz, B.-J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., Nixon, J., Van Auken, K., Wang, X., Shi, X., *et al.* (2010). The BioGRID interaction database: 2011 update. *Nucleic acids research*, **39**(suppl_1), D698–D704.

Wei, P. and Pan, W. (2012). Bayesian joint modeling of multiple gene networks and diverse genomic data to identify target genes of a transcription factor. *The Annals of Applied Statistics*, **6**(1), 334 – 355.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, **3**(1), 9.

Yin, M. and Zhou, M. (2018). Semi-implicit variational inference. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 2018.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

Zhou, M. (2018). Nonparametric Bayesian negative binomial factor analysis. *Bayesian Analysis*, **13**(4), 1065–1093.

Zhou, M. and Carin, L. (2015). Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**(2), 307–320.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture