Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 3 2012 *Article* 11

Hierarchical Bayes Model for Predicting Effectiveness of HIV Combination Therapies

Jasmina Bogojeska, *Max-Planck Institute for Informatics, Saarbrücken, Germany*
Thomas Lengauer, *Max-Planck-Institute for Informatics, Saarbrücken, Germany*

Recommended Citation:
Bogojeska, Jasmina and Lengauer, Thomas (2012) "Hierarchical Bayes Model for Predicting Effectiveness of HIV Combination Therapies," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 3, Article 11.
DOI: 10.1515/1544-6115.1769

Hierarchical Bayes Model for Predicting Effectiveness of HIV Combination Therapies

Jasmina Bogojeska and Thomas Lengauer

Abstract

HIV patients are treated by administration of combinations of antiretroviral drugs. The very large number of such combinations makes the manual search for an effective therapy practically impossible, especially in advanced stages of the disease. Therapy selection can be supported by statistical methods that predict the outcomes of candidate therapies. However, these methods are based on clinical data sets that have highly unbalanced therapy representation.

This paper presents a novel approach that considers each drug belonging to a target combination therapy as a separate task in a multi-task hierarchical Bayes setting. The drug-specific models take into account information on all therapies containing the drug, not just the target therapy. In this way, we can circumvent the problem of data sparseness pertaining to some target therapies.

The computational validation shows that compared to the most commonly used approach that provides therapy information in the form of input features, our model has significantly higher predictive power for therapies with very few training samples and is at least as powerful for abundant therapies.

KEYWORDS: hierarchical Bayes modelling, HIV combination therapies, statistical models, classification

Author Notes: We gratefully acknowledge the EuResist EEIG for providing the clinical data and the EuResist and Arevir study groups for providing biological/medical expertise. This work was conducted in the framework of the CHAIN project [grant number 223131] and partially funded by the Cluster of Excellence (Multimodal Computing and Interaction).

1 Introduction

More than 33 million people worldwide live with the human immunodeficiency virus 1 (HIV-1) (UNAIDS/WHO, 2010). Causing the acquired immunodeficiency syndrome (AIDS), with no cure or vaccine in sight, HIV-1 infected patients are customarily treated with combinations of several antiretroviral drugs. Although these drug cocktails remain effective much longer than monotherapies based on single drugs, they eventually are defeated by the evolution of the virus to resistance. In such a case the physician administers a new therapy by taking into account information on the resistance-relevant mutations present in the most abundant viral strain(s) in the patient's blood serum and on the previously administered drugs. The extensive set of resistance-relevant mutations in the viral genome and the large number of potential combination therapies resulting from the increasing number of antiretroviral drugs renders their manual assessment practically impossible. The availability of large clinical data sets has paved the way for statistical methods that offer an automated procedure for predicting the outcome of a potential antiretroviral therapy. An estimate of the therapy outcome can assist physicians in choosing a successful regimen for an HIV patient.

The clinical data contain samples from applications of many different drug combinations over many years. The evolving trends in treating HIV patients result in a highly unbalanced representation of different therapies in the available clinical data sets: while for some therapies many samples exist, for others there are very few. Furthermore, the existing statistical methods commonly used for predicting outcomes of HIV therapies use both the viral genotype and the drugs comprising the corresponding therapy as input features. Thus, such methods do not provide explicit models for the effects of each drug comprising the target therapy on its response. Furthermore, if the respective model is simple, which is desirable in the face of curbing overtraining, it can have problems appropriately modeling interactions between drugs and mutations. Specifically linear models cannot take into account such interactions. Such models must rely on additional information, *e.g.,* in the form of predicted resistance factors or genetic barriers to drug resistance, to afford an accurate prediction.

The method we present here affords a simple, direct, effective and efficient approach to modeling the response to HIV combination therapies based on the hierarchical Bayes paradigm. The individual drugs comprising each therapy combination are considered as separate tasks in a multi-task model that learns their additive effects on the therapy outcome from the available clinical data. In this way, the model makes use of the abundance of samples involving each individual drug. Doing so improves the predictive power on target therapies that are scarcely represented in the clinical database. In this paper, we demonstrate that our approach

delivers better predictions than the most common approaches, which use the drug information regarding a target therapy as part of the input and apply standard statistical learning methods. Furthermore, our approach allows for interactions among the input features (*e.g.,* resistance-relevant mutations) of the different drugs since, instead of encoding the drugs comprising the therapies as input features, each drug is considered a separate task in a multi-task learning setting.

The performance of the model is assessed via the so-called *therapy-stratified cross-validation scenario* that stratifies for the abundance of target therapies in the training data set. Alternatively, in order to take into account the evolving trends in composing drug combination therapies over time we use a *time-oriented validation scenario*: our models are trained on data from the more distant past, while their performance is assessed on data from the more recent past. In both validation scenarios the results of our method are compared to those of a method that mimics the most common approach used for predicting outcomes of combination therapies, which trains a linear model by supplying the drug information from the target therapy as part of the input. Moreover, in the time-oriented validation scenario our method is also compared to a therapy-specific approach that trains a separate model for each different therapy by using information available from similar therapies.

The paper is structured as follows. After summarizing related work we present the details of the method and its application to predicting responses of antiviral combination therapies in Section 2. In Section 3 we describe the data sets, the validation settings and present the results of the computational experiments. Section 4 discusses the results and concludes our paper.

1.1 Related Work

Various statistical learning methods, including artificial neural networks, decision trees, random forests, support vector machines (SVMs) and logistic regression (Wang, Larder, Revell, Harrigan, and Montaner, 2003, Lathrop and Pazzani, 1999, Altmann, Beerenwinkel, Sing, Savenkov, Däumer, Kaiser, Rhee, Fessel, Shafer, and Lengauer, 2007, Larder, Wang, Revell, Montaner, Harrigan, De Wolf, Lange, Wegner, Ruiz, Prez-Elas, Emery, Gatell, DArminio Monforte, Torti, Zazzi, and Lane, 2007, Deforche, Cozzi-Lepri, Thays, Clotet, Camacho, Kjaer, Van Laethem, Phillips, Moreau, Lundgren, and Vandamme, 2008, Rosen-Zvi, Altmann, Prosperi, Aharoni, Neuvirth, Snnerborg, Schülter, Struck, Peres, Incardona, Kaiser, Zazzi, and Lengauer, 2008, Altmann, Däumer, Beerenwinkel, Peres, Büch, Rhee, Sönnerborg, Fessel, Shafer, Zazzi, Kaiser, and Lengauer, 2009, Prosperi, Altmann, Rosen-Zvi, Aharoni, Borgulya, Bazso, Sönnerborg, Schülter, Struck, Ulivi, Vandamme, Vercauteren, and Zazzi, 2009), have been used to predict the virological

response to HIV combination therapies. For all these methods the drugs comprising the corresponding therapy are provided as input features. Thus, these methods do not generate separate models for the effects of the individual drugs on the therapy outcome. Furthermore, such models do not address the problem of uneven and sparse representation of therapies in the HIV training data. Some approaches (Bickel, Bogojeska, Lengauer, and Scheffer, 2008, Bogojeska, Bickel, Altmann, and Lengauer, 2010) deal with the aforementioned issue by estimating a separate model for each combination therapy which uses the training samples from all therapies with properly derived sample weights. The weights reflect the similarities between the target therapy and the corresponding therapies of all training samples. While these therapy-specific models achieve very good accuracy (Bogojeska et al., 2010), their AUC (Area Under the ROC Curve) performance can be improved.

The hierarchical Bayes paradigm (Gelman, Carlin, Stern, and Rubin, 2004) can easily be applied to multi-task modeling and, therefore, is widely used in the machine learning community (Evgeniou and Pontil, 2004, Yu, Tresp, and Schwaighofer, 2005, Dudik, Schapire, and Phillips, 2005, Teh, Jordan, Beal, and Blei, 2006). Our work is inspired by the work of Evgeniou and Pontil (2004) who present a feature mapping method for multi-task learning with support vector machines based on a hierarchical Bayes approach. Bickel (2009) shows that a revised version of this method with a logistic loss function is equivalent to a hierarchical Bayes model. In this paper we adapt this method to the problem at hand which yields a novel method that models the individual effects of the drugs on therapy outcome.

2 Methods

In what follows we derive the multi-task hierarchical Bayes learning method for the problem of predicting the outcomes of HIV drug combination therapies. Our goal is to model the effects of the drugs comprising a target combination therapy on its outcome by using the viral genotype information and the available information on previously administered drugs as input features. Since the individual drugs comprising the combination therapies appear in many samples we can consider each drug as a separate task in a multi-task setting. We use an additivity assumption to model the combined effects of the individual drugs comprising a target therapy on its response. Clearly this assumption is a gross simplification of the complex and little understood process of drug interaction. Still, the drug additivity approach is a widely used simple assumption in a situation where little information is available on actual interactions, and it exhibits good prediction performance. The sum of the drug-specific contributions provides a score quantifying the propensity of the

therapy to be effective. For each drug model we have a comparatively data-rich scenario, thus avoiding the necessity to make predictions on the basis of only very few informative samples.

2.1 Hierarchical Bayes Model

In the hierarchical Bayes setting the posterior probability $p(\mathbf{w}, \varphi|D)$ is computed by using the likelihood $p(D|\mathbf{w})$ of the training data D under model parameters $\mathbf{w}$, the prior probability $p(\mathbf{w}|\varphi)$ of model parameters $\mathbf{w}$ under hyperparameters φ, and the prior $p(\varphi)$ of the hyperparameters φ:

$$p(\mathbf{w}, \varphi|D) \propto p(D|\mathbf{w})p(\mathbf{w}|\varphi)p(\varphi). \tag{1}$$

Then, the maximum a posteriori (MAP) estimate of the model parameters $(\hat{\mathbf{w}}, \hat{\varphi}) = \arg\max_{\mathbf{w},\varphi} p(\mathbf{w}, \varphi|D)$ is used for the final prediction $\hat{y} = \arg\max_y p(y|\mathbf{x}, \hat{\mathbf{w}})$ for a target sample $\mathbf{x}$.

A multi-task problem with several related tasks that share a common prior can easily be realized in the hierarchical Bayes framework. Let $\mathbf{w}_1, \ldots, \mathbf{w}_n$ denote the task parameters of each of the n different tasks appearing in the training data $D = \{(\mathbf{x}_1, y_1, t_1), \ldots, (\mathbf{x}_m, y_m, t_m)\}$, and $D_t = \{(\mathbf{x}_i, y_i, t_i) \in D | t_i = t\}$ is the training data for task t. All task parameters have the same prior probability $p(\mathbf{w}_t|\varphi)$ and are conditionally independent given the prior. The posterior is then given by:

$$p(\mathbf{w}_1, \ldots, \mathbf{w}_n, \varphi|D_1, \ldots, D_n) = p(\varphi) \prod_t p(D_t|\mathbf{w}_t)p(\mathbf{w}_t|\varphi) \tag{2}$$

where the parameters are approximated with a MAP estimate. Intuitively, the prior models what all tasks have in common, while the task parameters capture task-specific information.

2.2 Outcome Prediction for HIV Combination Therapies

Let $\mathbf{x}$ denote the input features that comprise the viral genotype and the drug history for the specific therapy example. The input is represented with a binary vector, where the part corresponding to the viral genotype indicates the occurrence of a set of resistance-relevant mutations (Johnson, Brun-Vezinet, Clotet, Günthrad, Kuritzkes, Pillay, Schapiro, and Richman, 2008), and the part corresponding to the drug history comprises the drugs known to be part of previous therapies. To further clarify this notation, assuming for the sake of simplicity that there are only five resistance-relevant mutations and three antiretroviral drugs, the binary vector

$\mathbf{x} = (0,1,0,1,1,0,1,1)$ indicates a sample where the resistance mutations 2, 4 and 5 have occurred and drugs 2 and 3 were known to be a part of previous therapies. Let $\mathbf{z}$ denote the therapy combination encoded as a binary vector that indicates the individual drugs comprising the therapy. The label y indicates the success (1) or failure (-1) of each sample therapy. Let $D = \{(\mathbf{x}_1, y_1, \mathbf{z}_1), \ldots, (\mathbf{x}_m, y_m, \mathbf{z}_m)\}$ denote the training data set. The most common approach in the field trains a single statistical model (*e.g.,* a linear logistic regression model) on all available therapy samples in the data set. Here the information on the individual drugs comprising the target therapy is encoded in a binary vector and supplied together with the other input features. In what follows we will present the details of the derivation of a hierarchical Bayes model that predicts the outcomes of HIV combination therapies.

The goal is to train a classifier $f_{\mathbf{z}} : \mathbf{x} \mapsto y$ that correctly predicts the outcome for an HIV combination therapy $\mathbf{z}$. We model the class likelihood $p(y|\mathbf{x}, \mathbf{z})$ with a logistic regression model that calculates predictions of the effectiveness of therapies by using the assumption that the drugs making up the target therapy have a cumulative effect on its outcome. This is reflected in the formula:

$$p(y|\mathbf{x}, \mathbf{z}, \mathbf{w}) = \frac{1}{(1 + \exp(-y \sum_{d \in \mathbf{z}} \mathbf{w}_d^{\mathsf{T}} \mathbf{x}))} \tag{3}$$

where $\mathbf{z}$ denotes the set of drugs comprising the combination therapy and $\mathbf{w}_d$ are the model parameters of the individual drugs, *i.e.,* the drug-specific weights pertaining to the resistance-relevant mutations and the previously administered drugs. These drug parameters are trained via the multi-task hierarchical Bayes framework where each drug is considered a separate task. The model parameters $\mathbf{w}_d$ for each drug are drawn from a common Gaussian prior $\mathbf{w}_d \sim N(\mathbf{w}_0, \sigma_{\mathbf{w}}^2 \mathbf{I})$ with a mean drawn from a Gaussian hyperprior $\mathbf{w}_0 \sim N(\mathbf{0}, \sigma_{\mathbf{w_0}}^{\mathbf{2}} \mathbf{I})$. In this way all tasks (drugs) are related and their similarity is modeled with the common Gaussian prior. In fact, some drugs are more similar than others in that they belong to the same drug class or evoke a similar genomic fingerprint in terms of viral resistance mutations. More formally, all task parameters $\mathbf{w}_d$ deviate to some extent from a mean function $\mathbf{w}_0$ (in our case the mean of the Gaussian prior). The smaller the distance between two distinct drug parameters $\mathbf{w}_{d_1}$ and $\mathbf{w}_{d_2}$ the more similar the effects of drugs d_1 and d_2.

Let n denote the number of different drugs in our data set and $D_d = \{(\mathbf{x}_i, y_i, \mathbf{z}_i) \in D | d \in \mathbf{z}_i\}$ denote all training samples whose corresponding therapies contain the drug d. In the following we derive the log-posterior of all parameters given the data in accordance with Equation 2 and the assumptions made in the previous paragraph.

$$\begin{aligned}
&\log p(\mathbf{w}_1,\ldots,\mathbf{w}_n,\mathbf{w}_0|D_1,\ldots,D_n,\sigma^2_{\mathbf{w}_0},\sigma^2_{\mathbf{w}}) \\
&\quad\propto \quad \log N(\mathbf{w}_0|\mathbf{0},\sigma^{\mathbf{2}}_{\mathbf{w_0}}\mathbf{I}) + \sum_{\mathbf{d=1}}^{\mathbf{n}} \log \mathbf{N}(\mathbf{w_d}|\mathbf{w_0},\sigma^{\mathbf{2}}_{\mathbf{w}}\mathbf{I}) \\
&\qquad + \sum_{d=1}^{n} \sum_{(\mathbf{x},y,\mathbf{z})\in D_d} \log p(y|\mathbf{x},\mathbf{z},\mathbf{w_z}) \qquad (4)
\end{aligned}$$

$$\begin{aligned}
&\propto -\frac{\|\mathbf{w}_0\|^2}{2\sigma^2_{\mathbf{w}_0}} - \sum_{d=1}^{n} \frac{\|\mathbf{w}_d-\mathbf{w}_0\|^2}{2\sigma^2_{\mathbf{w}}} \\
&\quad - \sum_{d=1}^{n} \sum_{(\mathbf{x},y,\mathbf{z})\in D_d} \log(1+\exp(-y\sum_{d\in\mathbf{z}} \mathbf{w}_d^{\mathsf{T}}\mathbf{x})) \qquad (5)
\end{aligned}$$

$$\begin{aligned}
&= -\frac{\|\mathbf{w}_0\|^2}{2\sigma^2_{\mathbf{w}_0}} - \sum_{d=1}^{n} \frac{\|\mathbf{v}_d\|^2}{2\sigma^2_{\mathbf{w}}} \\
&\quad - \sum_{d=1}^{n} \sum_{(\mathbf{x},y,\mathbf{z})\in D_d} \log(1+\exp(-y(|\mathbf{z}|\mathbf{w}_0+\sum_{d\in\mathbf{z}} \mathbf{v}_d)^{\mathsf{T}}\mathbf{x})) \qquad (6)
\end{aligned}$$

$$\begin{aligned}
&= -\frac{\|\mathbf{v}_0\|^2}{2\sigma^2_{\mathbf{w}}} - \sum_{d=1}^{n} \frac{\|\mathbf{v}_d\|^2}{2\sigma^2_{\mathbf{w}}} \\
&\quad - \sum_{d=1}^{n} \sum_{(\mathbf{x},y,\mathbf{z})\in D_d} \log(1+\exp(-y(\frac{|\mathbf{z}|\sigma_{\mathbf{w}_0}}{\sigma_{\mathbf{w}}}\mathbf{v}_0+\sum_{d\in\mathbf{z}} \mathbf{v}_d)^{\mathsf{T}}\mathbf{x})) \qquad (7)
\end{aligned}$$

$$= -\frac{\|\mathbf{v}\|^2}{2\sigma^2_{\mathbf{w}}} - \sum_{(\mathbf{x},y,\mathbf{z})\in D} \log(1+\exp(-y\mathbf{v}^{\mathsf{T}}\Phi(\mathbf{x},\mathbf{z}))) \qquad (8)$$

Equation 4 uses Equation 2 to derive the logarithm of the posterior probability. In Equation 5 the Gaussian density functions are expanded up to constant terms. Since $\mathbf{w}_d \sim N(\mathbf{w}_0,\sigma^2_{\mathbf{w}}\mathbf{I})$ each individual drug parameter $\mathbf{w}_d$ can be replaced by $\mathbf{w}_d = \mathbf{w}_0 + \mathbf{v}_d$ yielding Equation 6 where $|\mathbf{z}|$ is the number of drugs comprising the therapy $\mathbf{z}$. In Equation 7 $\mathbf{w}_0$ is replaced with $\frac{\sigma_{\mathbf{w}_0}}{\sigma_{\mathbf{w}}}\mathbf{v}_0$. Finally, in the last Equation 8 the vector $\mathbf{v}$ denotes the concatenation of all parameter vectors $\mathbf{v} = [\mathbf{v}_0,\ldots,\mathbf{v}_n]$ and $\Phi(\mathbf{x},\mathbf{z})$ is a new feature mapping defined as follows. Let $\mathbf{c_z} = [\frac{|\mathbf{z}|\sigma_{\mathbf{w_0}}}{\sigma_{\mathbf{w}}},\mathbf{z}]$ denote an extension of the therapy vector $\mathbf{z}$, where each vector component is a vector itself with identical elements and dimension equal to the dimension of the input feature vector $\mathbf{x}$. The new feature mapping is then given by $\Phi(\mathbf{x},\mathbf{z}) = \mathbf{c_z}\cdot[\mathbf{x},\ldots,\mathbf{x}]$ (where

$\cdot$ denotes componentwise vector multiplication). In other words, it maps the input features of the training samples to a new feature space that provides a separate set of dimensions for each drug comprising the target therapy: the feature vector $\mathbf{x}$ for a given training sample is copied to the sections corresponding to the drugs comprising the target therapy $\mathbf{z}$; all other sections except the first one are filled with zeros; the first section is shared by all drugs and models their similarity. For example, let us assume that a target drug combination $\mathbf{z}$ comprises only two drugs $\mathbf{z} = \{d_1, d_2 | d_1 < d_2, d_1, d_2 \in 1, \ldots, n\}$. Then for given input features $\mathbf{x}$ the feature mapping $\Phi(\mathbf{x}, \mathbf{z})$ is given by:

$$\Phi(\mathbf{x}, \mathbf{z}) = \left[\frac{2\sigma_{\mathbf{w}_0}}{\sigma_{\mathbf{w}}} \mathbf{x}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{1,\ldots,d_1-1}, \underbrace{\mathbf{x}}_{d_1}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{d_1+1,\ldots,d_2-1}, \underbrace{\mathbf{x}}_{d_2}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{d_2+1,\ldots,n} \right]. \tag{9}$$

As can be observed from Equation 8, by using the feature mapping $\Phi(\mathbf{x}, \mathbf{z})$ we obtain the objective function of a logistic regression model with model parameters $\mathbf{v}$. The dimensionality of the new input feature space is the dimension of $\mathbf{x}$ multiplied by $(n+1)$.

To summarize, the MAP estimate of the parameters of a hierarchical Bayes model with Gaussian prior and hyperprior applied to the problem of predicting outcomes of HIV therapies with drug additivity assumption is given by:

$$\hat{\mathbf{v}} = \arg\max_{\mathbf{v}} \{ - \sum_{(\mathbf{x}, y, \mathbf{z}) \in D} \log(1 + \exp(-y \mathbf{v}^{\mathsf{T}} \Phi(\mathbf{x}, \mathbf{z}))) - \frac{\|\mathbf{v}\|^2}{2\sigma_{\mathbf{w}}^2} \}. \tag{10}$$

We obtain the maximum $\hat{\mathbf{v}}$ with logistic regression. Then the prediction of the label (success probability) of a target therapy $\mathbf{z}$ administered to a sample $\mathbf{x}$ is given by:

$$\hat{y} = \arg\max_{y} \frac{1}{(1 + \exp(-y \hat{\mathbf{v}}^{\mathsf{T}} \Phi(\mathbf{x}, \mathbf{z})))}. \tag{11}$$

We will refer to this method as *drug additivity Bayes*. A slightly modified version of the drug additivity Bayes is the *drug additivity + hist Bayes* described as follows. For each of the drugs comprising a target combination therapy two tasks are created: one for the case when the drug is administered for the first time to the considered patient, and another one for the case when the drug was administered previously in the patient's drug history. Once the tasks are defined, a task additivity assumption is applied, and the model is derived in the same way as the *drug additivity Bayes*. The dimensionality of the input feature space of the new model is the dimension of $\mathbf{x}$ multiplied by $(2n+1)$.

Since we use Gaussian priors in our Bayes models, we can employ the trust-region Newton method for training logistic regression (Lin, Weng, and Keerthi,

2008). This is an efficient implementation for sparse data sets with large number of features and samples. With this approach our models are trained in about one second, albeit the increased dimensionality of the input feature space and the large number of training samples. The Bayes methods have two tuning parameters: one replacing the fraction $\frac{|\mathbf{z}|\sigma_{\mathbf{w}_0}}{\sigma_{\mathbf{w}}}$ in the feature mapping $\Phi(\mathbf{x},\mathbf{z})$ and one for the regularizer in Equation 10.

3 Experiments and Results

3.1 Data Sets

The training data are extracted from the EuResist database (Rosen-Zvi et al., 2008) that contains information on 93014 antiretroviral therapies administered to 18325 HIV (subtype B) patients from several countries in the period from 1988 to 2008. This information includes the individual drugs that comprise a therapy, virus load measurements (copies of viral RNA per ml blood plasma, cp/ml) during the course of a therapy, all available therapies administered to each patient, as well as consensus sequences of the predominant viral strains in the patients' blood. We include a therapy as a sample in the training data if there is a viral sequence obtained shortly before the therapy was started (up to 90 days before) and if it can be assigned a label (success or failure) based on the virus load values measured during its course. The information on the viral genotype is given in terms of the presence of any from a set of predefined resistance-relevant mutations (based on the list in Johnson et al. (2008)) encoded with a binary vector. We consider 70 resistance-relevant mutation positions. The therapy label is determined as in Bogojeska et al. (2010): if the virus load drops below 400 cp/ml in the period from 21 days after the start of the therapy to its end we label it successful (1); otherwise we label it failing (-1). We represent the individual drugs comprising each therapy by a binary vector indicating the presence or absence of all drugs appearing in the data set. Finally, we end up with a training set that includes 6750 labeled therapy samples with 805 distinct therapy combinations.

Figure 1 depicts a histogram of the frequencies of the different combination therapies in the training data set: almost 500 therapies occur less than five times; for almost all therapies there are no more than 50 samples. While there are many rare therapies, there is a reasonable number of samples in which each of the different drugs appear.

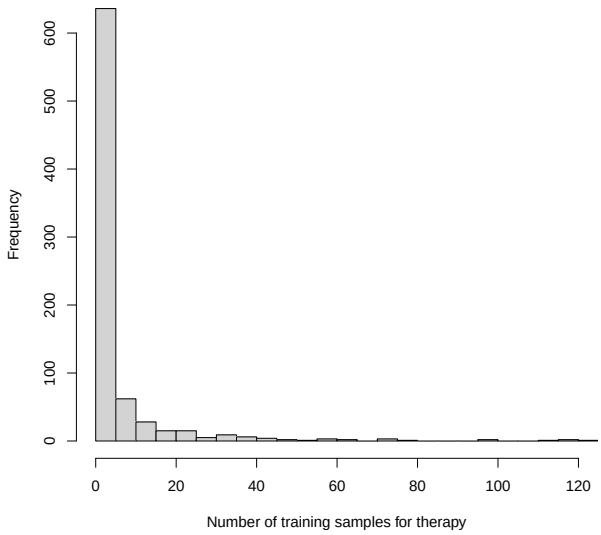


Figure 1: Histogram that groups the 805 distinct combination therapies in our labeled training data set based on their corresponding number of available training examples. The image displays the uneven therapy representation in the data where almost 500 therapies are represented with less than five samples.

3.2 Validation Settings

The quality of our approach is assessed in two validation scenarios: the *therapy-stratified cross-validation scenario* and the *time-oriented validation scenario*. In what follows we provide the details on each of them.

Therapy-stratified cross-validation scenario. In order to provide an assessment of the performance of a target method that stratifies for therapy abundance in the training data set, we introduce the therapy-stratified cross-validation scenario. We start by describing the procedure of creating therapy-stratified cross-validation folds. First, all available samples are grouped in therapy bins based on their corresponding therapies. Then we populate the cross-validation folds with samples: the folds are repeatedly visited one after the other and one sample is assigned to each fold at a time; the assigned samples are chosen at random from the therapy bins, which are traversed in a round-robin fashion. In this way we make sure that both infrequent and abundant therapy samples are distributed evenly among the cross-validation folds. In the following we detail the therapy-stratified cross-validation

scenario that we applied for our computational experiments. We first construct a separate test set, that comprises 20% of the available data, by selecting one fold from a five-fold therapy-stratified cross validation. Then, we conduct a 10-fold therapy-stratified cross validation on the remaining data and use it for the model selection. At the end, we report the cross-validation results and evaluate the selected model on the separate test set.

Time-oriented validation scenario. The trends of treating HIV patients change over time as a result of the gathered practical experience with the drugs and the introduction of new antiretroviral drugs. As in Bickel et al. (2008), Bogojeska et al. (2010), our evaluation scenario accounts for this phenomenon by using a time-oriented split when selecting the training and the test set. Such a setting is realistic since it captures how a given model would perform on the recent trends of selecting combinations of drugs from established drug classes. We refer to this scenario as time-oriented scenario and we apply it as follows. First, we order all available training samples by their corresponding therapy starting dates. We then make a time-oriented split by selecting the most recent 20% of the samples as the test set and the rest as the training set. For the model selection we split the training set further in a similar manner. We take the most recent 25% of the training set for selecting the best model parameters and refer to this set as tuning set. Figure 2 depicts the different treatment trends in the training, tuning and test sets, generated as explained in the text above. One can observe that, unlike the treatment trends in the training set, the treatment trends in the tuning set closely resemble those in the test set. This justifies the choice of the tuning set. We select the training and test data with a time-oriented approach. So in order to make sure that we have a reasonable amount of training samples for each individual drug that appears in the test set, we remove the therapy samples which contain very recent drugs from our data set. Our multi-task Bayes approach utilizes the high frequencies of samples involving the individual drugs to address the problem of the low frequencies of samples corresponding to specific combination therapies. As a side remark, in practice one cannot expect quality predictions for therapy samples comprising a drug for which there are only few training samples. The resulting data set contains 6336 samples.

Model performance. We assess the performance of the target models by taking the uneven representation of the different therapies into account. We do this by grouping the therapies in the test set based on the number of samples they have in the training set, and then measuring the model performance on each of the groups. We thereby assess the performance of the models for the rare and the abundant therapies, separately. We carry out the model selection based on AUC (Area Under the

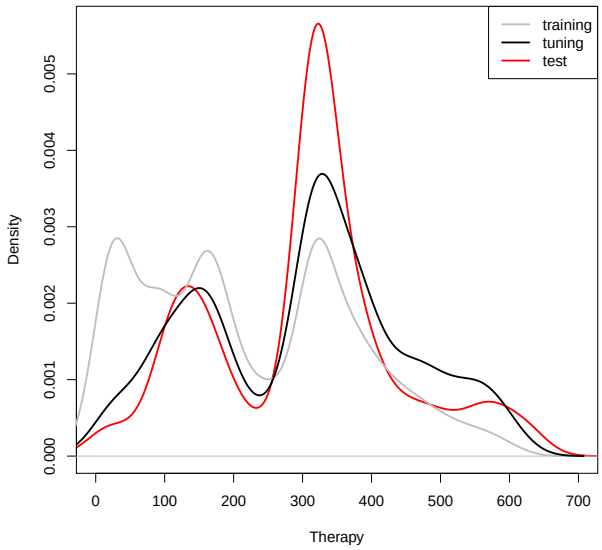


Figure 2: Distribution of the different combination therapies in the training, tuning and test set chosen in the time-oriented scenario. The numbers on the x-axis represent the different therapy combinations ordered by their first appearance in our clinical data: from older to newer. The y-axis depicts the density.

ROC Curve) results and use the AUCs to assess the model performance. In this way we evaluate the quality of the ranking of the therapies based on their success probabilities. For the comparison of the cross-validation performances of two methods we use a paired t-test. In order to compare the performance of two methods on a separate test set, the standard errors of the AUC values and the significance of the difference of two AUCs are estimated as described in Hanley and McNeil (1983). We use the ROCR package to plot the ROC curves (Sing, Sander, Beerenwinkel, and Lengauer, 2005).

Reference methods. In our computational experiments we compare the performance of the two multi-task Bayes methods described in our paper to those of two reference approaches, namely the *one-for-all model* and the *therapy-specific model*. The *one-for-all* method mimics the most common approach in the field where a single linear logistic regression model is trained on all available therapy samples in the data set. The information on the individual drugs comprising each of the therapies is encoded in a binary vector and supplied together with the other input features. The therapy-specific model represents the approaches that deal with the uneven and

Table 1: AUCs with their corresponding standard errors for our two multi-task Bayes models (drug additivity, drug additivity + hist) and the reference (one-for-all) method. Generated by a 10-fold therapy-stratified cross validation for three groups of test therapies: with $0-7$, $8-30$ and more than 30 training samples, they summarize the performance for both the rare and the abundant test therapy samples.

method	multi-task Bayes		one-for-all
	drug additivity	drug additivity + hist	
$0-7$ (*SE*)	$0.771(0.016)$	$0.774(0.016)$	$0.749(0.015)$
$8-30$ (*SE*)	$0.745(0.011)$	$0.738(0.012)$	$0.732(0.010)$
>30 (*SE*)	$0.772(0.017)$	$0.765(0.018)$	$0.759(0.012)$

sparse therapy representation by training a separate model for each combination therapy using not only the samples from the target therapy but also the available samples from similar therapies with appropriate sample weights. It implements the drugs kernel therapy similarity model as described in Bogojeska et al. (2010) on the input feature space defined in the previous section of this paper. Since training separate models for every different therapy in a cross-validation setting is very time-consuming, we only consider this approach as a reference model in the time-oriented validation scenario. Note also that in the papers where they are introduced (Bickel et al., 2008, Bogojeska et al., 2010) the performance of the therapy-specific approaches is evaluated in the time-oriented validation scenario. All approaches we consider (the multi-task Bayes and the reference approaches) rely on the same input information that can directly be derived from the EuResist database as described in Subsection 3.1.

3.3 Experimental Results

In this subsection we first present the results of the computational experiments for the therapy-stratified cross-validation scenario, followed by the results of the time-oriented scenario.

Table 1 summarizes the cross-validation performance of the considered methods: *drug additivity Bayes*; *drug additivity + history Bayes*; and *one-for-all* as the reference method. The two Bayes approaches significantly outperform the *one-for-all method* for the therapies that have few $(0-7)$ available samples in the training set. We verified the significance of the improvements with the paired t-test: *p-value* $= 0.05$ for the *drug additivity* and *p-value* $= 0.06$ for the *drug additivity*

+ *hist model*. Moreover, for the group of therapies with $8-30$ training samples the *drug additivity* approach also shows significantly better cross-validation performance than the reference model (*p-value* $= 0.05$). All models deliver comparable predictions for the group of therapy samples for which there is a reasonable number (more than 30) of available samples in the training set. According to the results for the separate test set, depicted in Figure 3 (a), the *drug additivity* model has better performance than the reference method for all three therapy groups (with $0-7$, $8-30$ and more than 30 training samples). However, the improvements are only significant for the test therapies with $0-7$ and more than 30 training samples, with *p-values* of 0.045 and 0.002, respectively. The *p-value* for the therapies with $8-30$ training samples is 0.157. The *drug additivity* + *hist model* shows significantly better AUC performance than the *one-for-all method* for the rare therapies with a *p-value* $= 0.034$. Figure 3 (b) depicts the ROC curves for all considered methods for the rare test therapies in the separate test set.

The experimental results for the time-oriented scenario are summarized in Figure 4 (a). Note that in this case both the *one-for-all* and the *therapy-specific* models are considered as reference methods. As can be observed, the *drug additivity* method outperforms the *one-for-all* method for the test therapies with $0-7$ and $8-30$ training samples. According to the paired difference test described in Hanley and McNeil (1983), the improvement is significant only for the test therapies with $0-7$ samples (*p-value*$= 0.078$). The *p-value* for the test therapies with $8-30$ training samples is 0.132. Compared to the *therapy-specific* model the *drug additivity* model has better AUC performance for the test therapies with $0-7$ training samples, yet this improvement is not significant (*p-value* $= 0.253$). For the test therapies with $8-30$ training samples both the *therapy-specific* and the *drug additivity* models have comparable performance. The *drug additivity* + *hist model* outperforms all considered approaches for the rare test therapies (with $0-7$ training samples) with estimated *p-values* of 0.007 for the *one-for-all*, 0.042 for the *therapy-specific* and 0.033 for the *drug additivity* model; for the test therapies with $8-30$ training samples it delivers similar performance as the *one-for-all* method. The AUC results of the *drug additivity* + *hist model* for the test therapies with $8-30$ training samples are slightly worse compared to the *therapy-specific* and the *drug additivity* models. However the respective differences in performance are not significant (all *p-values* > 0.1). Considering the abundant test therapies (with more than 30 training samples) all approaches deliver comparable results. The relevant ROC curves for the rare test therapies are shown in Figure 4 (b). We should also point out that decreasing or increasing the size of the tuning set in the time-oriented scenario with $5-15\%$ less or more training data yields very similar results and leads to the same conclusions.

To summarize, in both validation scenarios the two multi-task Bayes approaches have their prime advantage for therapies with few (less than eight) available training samples. The *drug additivity Bayes* performs better than the *one-for-all* and the *therapy-specific* methods for the therapies with $8-30$ available samples, however the improvement is statistically significant (corresponding *p-value* < 0.1) only for the cross-validation results. For the abundant test therapies (with more than 30 training samples) all considered methods have comparable performance in almost all validation scenarios — one exception is the significantly better performance of the *drug additivity Bayes* method for the separate test set in the cross-validation scenario.

4 Discussion

This paper presents an approach to predicting virological response to HIV combination therapies by considering each individual antiretroviral drug as a separate task in a multi-task hierarchical Bayes framework. With our method the additive effects of the individual drugs comprising each combination therapy on its response are modeled from the data. It is worth noting that the most common approaches in the field that use linear models and encode the therapy information in the input feature space, also implicitly use a drug additivity assumption. However, in this case, the effects of the drugs comprising each therapy on its response are not explicitly modeled. Instead, a generic statistical learning method that simultaneously models the contributions of all available information (e.g. therapy, viral genotype) on the therapy outcome is used. Above all, such methods do not take the uneven therapy representation in the clinical data sets into account. By considering each drug as a separate task, our Bayes approach uses the abundance of samples that pertain to each drug to circumvent the lack of samples for the specific combination therapies. In this way we provide more accurate predictions for rare therapies by maintaining the prediction quality for the more frequent therapies. The samples corresponding to rare therapies (represented with $0-7$ samples in our clinical data) make up only around 18% of the available data, but they contain 83% of the different therapies i.e. they make up the therapy variety in our data set. Moreover, our approach allows for interactions among the input features of the different drugs by using an extended input feature space where each drug has a separate range and thereby making the model more interpretable.

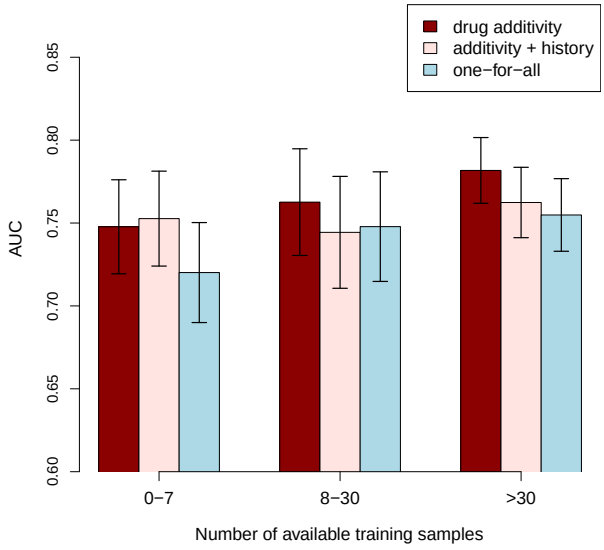
The use of an efficient optimization method (Lin et al., 2008) that takes advantage of the sparseness of our input data ensures very fast model fitting (one second) and model selection. For example, the model selection procedure performed with a 10-fold cross validation for the drugs additivity model screens 289 different

value combinations for the two model selection parameters specified in the Methods section and is completed in about ten hours.

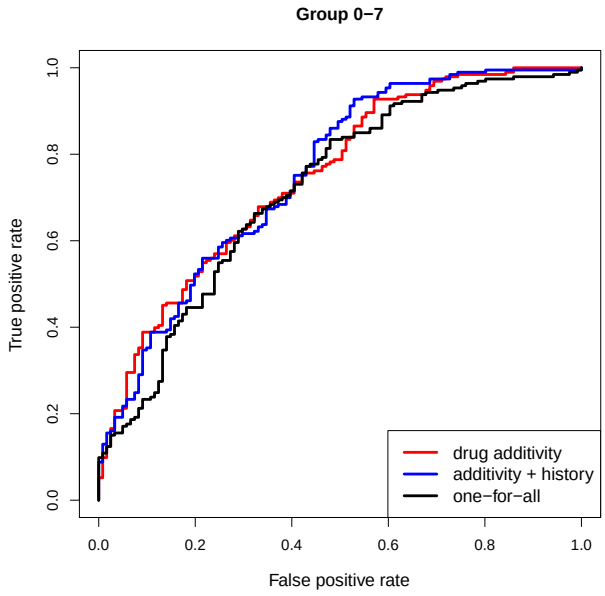
According to the cross-validation results both our multi-task Bayes models perform significantly better (at the 5% significance level for the *drugs additivity* and the 6% level for the *drugs additivity + hist* model) than the *one-for-all* model for rare test therapies (with $0-7$ available training samples). The *drugs additivity* model also significantly outperforms the *one-for-all scenario* for the group of therapies with $8-30$ training samples. For therapies with a sizeable number of samples (above 30) all approaches show comparable cross-validation performance. The results on the left-out set in the cross-validation scenario confirms the advantage of both multi-task Bayes models for the less frequent therapies (the significance level is 5% for the *drugs additivity* and the 3% level for the *drugs additivity + hist* model). Furthermore, the *drugs additivity* model achieves better performance for the other two groups of test therapies (with $8-30$ and more than 30 training samples). However, the improvement is only significant for the test therapies with more than 30 available training samples.

According to the time-oriented scenario both Bayes models significantly outperform (at the 8% significance level for the *drugs additivity* and the 1% significance level for the *drugs additivity + hist* model) the *one-for-all* model for the test therapies with less than eight available training samples. Moreover, the *drugs additivity + hist* model also outperforms the *drugs additivity* model (at the 3% significance level) and the *therapy-specific* model (at the 4% significance level) for the group of rare test therapies. All models show comparable performance for the abundant test therapies.

In summary, the approach presented in this paper models the effects of the individual drugs comprising an HIV combination therapy on its effectiveness by using a multi-task hierarchical Bayes approach. The performance of this approach is at least as good as an approach that encodes therapy information in the input feature space for the abundant therapies and significantly better for therapies with few training samples. The same observation holds when comparing the performance of the hierarchical Bayes approaches to the therapy-specific approach which trains a separate model for every different combination therapy. In this case the Bayes models have the additional advantage of being more time-efficient compared to the therapy-specific approach. Note that the group of rare therapies is very important as it makes up the therapy variety in the available clinical data.

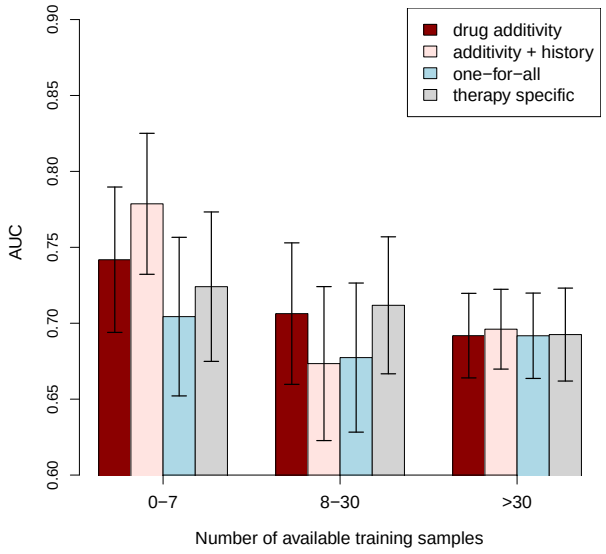


(a)

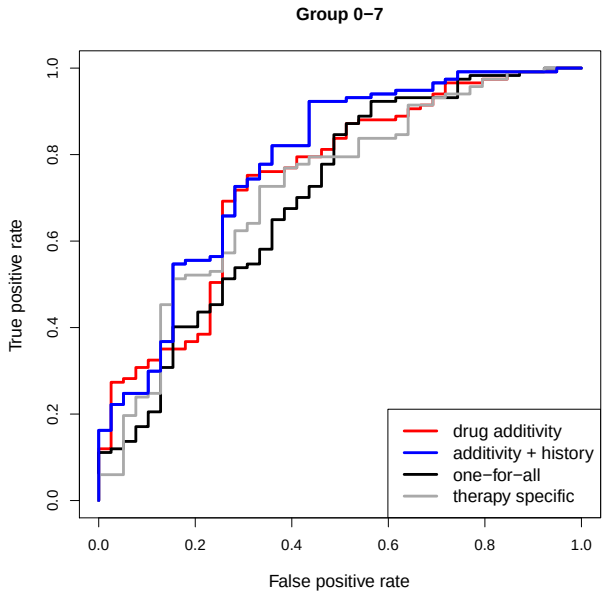


(b)

Figure 3: AUC results of the different models obtained on the separate test set in the cross-validation therapy-stratified scenario. (a) Test samples are grouped based on the number of training examples for their corresponding therapy combinations. Error bars indicate the standard errors of each model; (b) ROC curves display the performance of the different methods on the rare therapies (with $0-7$ training samples) of the separate test set.



(a)



(b)

Figure 4: AUC results of the different models obtained on the test set in the time-oriented validation scenario. (a) Test samples are grouped based on the number of training examples for their corresponding therapy combinations. Error bars indicate the standard errors of each model; (b) ROC curves display the performance of the different methods on the rare test therapies (with $0-7$ training samples).

References

Altmann, A., N. Beerenwinkel, T. Sing, I. Savenkov, M. Däumer, R. Kaiser, S. Rhee, W. Fessel, W. Shafer, and T. Lengauer (2007): "Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance," *Antiviral Therapy*, 12, 169–178.

Altmann, A., M. Däumer, N. Beerenwinkel, E. Peres, Y. Schülter, A. Büch, S. Rhee, A. Sönnerborg, W. Fessel, W. Shafer, M. Zazzi, R. Kaiser, and T. Lengauer (2009): "Predicting response to combination antiretroviral therapy: retrospective validation of geno2pheno-THEO on a large clinical database," *Journal of Infectious Diseases*, 199, 999–1006.

Bickel, S. (2009): *Learning under Differing Training and Test Distributions*, Ph.D. thesis, Universität Potsdam.

Bickel, S., J. Bogojeska, T. Lengauer, and T. Scheffer (2008): "Multi-task learning for HIV therapy screening," in *Proceedings of the International Conference on Machine Learning*.

Bogojeska, J., S. Bickel, A. Altmann, and T. Lengauer (2010): "Dealing with sparse data in predicting outcomes of HIV combination therapies," *Bioinformatics*, 26, 2085–2092.

Deforche, K., A. Cozzi-Lepri, K. Thays, B. Clotet, R. Camacho, J. Kjaer, K. Van Laethem, A. Phillips, Y. Moreau, J. Lundgren, and A. Vandamme (2008): "Modelled in vivo HIV fitness under drug selective pressure and estimated genetic barrier towards resistance are predictive for virological response," *Antiviral Therapy*, 13, 399–407.

Dudik, M., R. Schapire, and S. Phillips (2005): "Correcting sample selection bias in maximum entropy density estimation," in *Advances in Neural Information Processing Systems*.

Evgeniou, T. and M. Pontil (2004): "Regularized multi–task learning," *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 109–117.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004): *Bayesian Data Analysis*, Chapman & Hall/CRC.

Hanley, J. and B. McNeil (1983): "A Method of comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases," *Radiology*, 148, 839–843.

Johnson, V., F. Brun-Vezinet, B. Clotet, H. Günthrad, D. Kuritzkes, D. Pillay, J. Schapiro, and D. Richman (2008): "Update of the drug resistance mutations in HIV-1: December 2008," *Topics in HIV Medicine*, 16, 138–145.

Larder, B., D. Wang, A. Revell, J. Montaner, R. Harrigan, F. De Wolf, J. Lange, S. Wegner, L. Ruiz, M. Prez-Elas, S. Emery, J. Gatell, A. DArminio Monforte,

C. Torti, M. Zazzi, and C. Lane (2007): "The development of artificial neural networks to predict virological response to combination HIV therapy." *Antiviral Therapy*, 12, 15–24.

Lathrop, R. and M. Pazzani (1999): "Combinatorial optimization in rapidly mutating drug-resistant viruses," *Journal of Combinatorial Optimization*, 3, 301–320.

Lin, C., R. Weng, and S. Keerthi (2008): "Trust region Newton method for large-scale logistic regression," *Journal of Machine Learning Research*, 9, 627–650.

Prosperi, M., A. Altmann, M. Rosen-Zvi, E. Aharoni, G. Borgulya, F. Bazso, A. Sönnerborg, E. Schülter, D. Struck, G. Ulivi, A. Vandamme, J. Vercauteren, and M. Zazzi (2009): "Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment," *Antiviral Therapy*, 14, 433–442.

Rosen-Zvi, M., A. Altmann, M. Prosperi, E. Aharoni, H. Neuvirth, A. Snnerborg, E. Schülter, D. Struck, Y. Peres, F. Incardona, R. Kaiser, M. Zazzi, and T. Lengauer (2008): "Selecting anti-HIV therapies based on a variety of genomic and clinical factors," *Proceedings of the ISMB*.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer (2005): "ROCR: visualizing classifier performance in R," *Bioinformatics*, 21, 3940.

Teh, Y., M. Jordan, M. Beal, and D. Blei (2006): "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, 101, 1566–1581.

UNAIDS/WHO (2010): "Report on the global AIDS epidemic: 2010," .

Wang, D., B. Larder, A. Revell, R. Harrigan, and J. Montaner (2003): "A neural network model using clinical cohort data accurately predicts virological response and identifies regimens with increased probability of success in treatment failures," *Antiviral Therapy*, 8, U99–U99.

Yu, K., V. Tresp, and A. Schwaighofer (2005): "Learning Gaussian processes from multiple tasks," *Proceedings of the International Conference on Machine Learning*.