Supplementary material to "History-alignment models for bias-aware prediction of virological response to HIV combination therapy"

Jasmina Bogojeska [1], Daniel Stöckel [2], Maurizio Zazzi [3] Rolf Kaiser [4], Francesca Incardona [5], Michal Rosen-Zvi [6] and Thomas Lengauer [1]

[1]Max Planck Institute for Informatics, Germany; [2]Saarland University, Germany; [3]University of Siena, Italy; [4]University of Cologne, Germany; [5]Informa, Italy; [6]IBM Research Labs, Israel

1 HIV clinical data set

This section comprises the figures and tables that provide the details of the HIV clinical data set used in the paper.

Table 1: Details on the bins grouping the test samples based on their corresponding number of previous therapies.

Bin	$0-2$	$3-5$	>5
Sample count	807	225	275
Success rate	89%	82%	68%

Table 2: Details on the bins grouping the test samples based on the number of training examples for their corresponding therapy combinations.

Bin	$0-7$	$8-30$	>30
Sample count	217	242	848
Success rate	77%	82%	85%

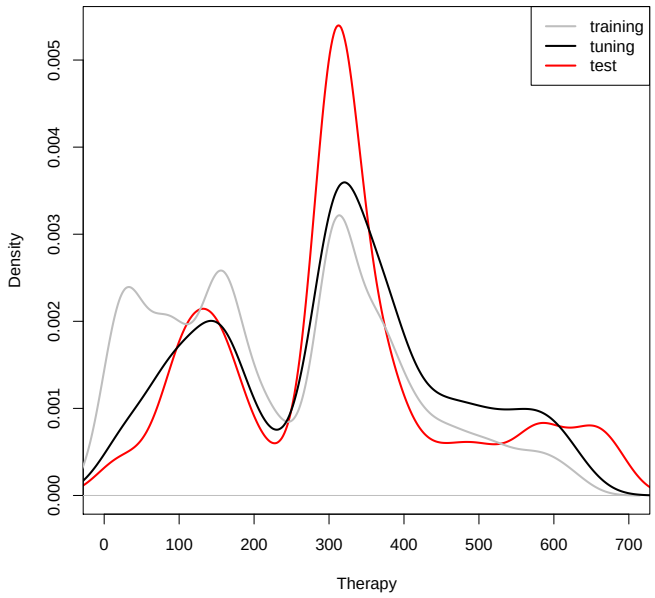


Figure 1: Distribution of the different combination therapies in the training, tuning and test set chosen in the time-oriented scenario. The numbers on the x-axis represent the different therapy combinations ordered by their first appearance in our clinical data: from older to newer. The y-axis depicts the density.

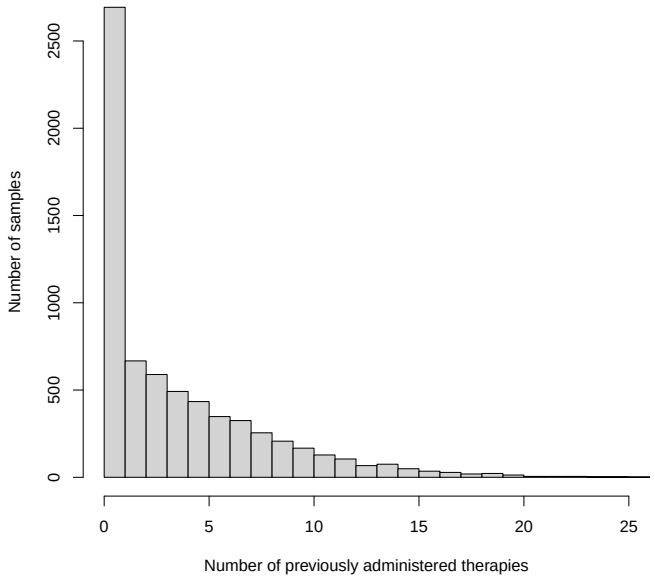


Figure 2: Histogram that groups all labeled samples in our clinical data set based on their corresponding number of known previous therapies. The histogram illustrates the uneven representation with respect to the length of the treatment history in the data, where the largest group with size 2600 consists of samples with none or only one known previous therapy.

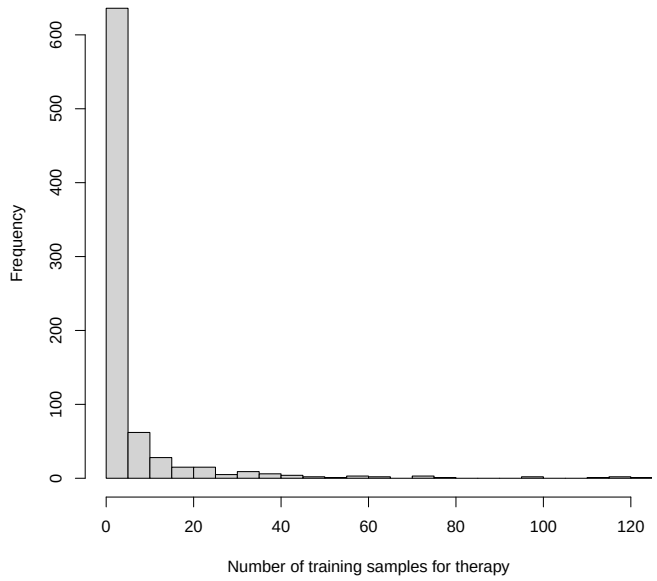


Figure 3: Histogram that groups the distinct combination therapies in our labeled training data set based on their corresponding number of available training examples. The image displays the uneven representation of the different therapies in the data set where almost 500 therapies are represented with less than five samples.

2 Interpretability of the history-similarity models

An additional contribution to model interpretability is achieved by assessing the relevance of the different input features, namely, the mutations in the viral sequence, the drugs comprising the current therapy and the drugs appearing in all previous therapies. This can easily be accomplished if we observe that we have a separate model for each sample and each of these models is trained on features describing the viral genome, the current therapy and the drugs from all previous therapies. In such a setting the importance of an input feature of the target model quantifies its relevance. One way to quantify feature importance is by computing the z-scores for each model coefficient: the higher the magnitude of the z-score the more significant the feature. In this manner we perform a statistical test for each model coefficient that checks the null hypothesis that the considered coefficient is zero, while all others are not. Figure 6 and 7 illustrate an interesting example for the relevance of the different input features for the sample with therapy sequence depicted in Figure 3 a) in the main paper. In Figure 6 one can observe the importance of the different mutations for the considered therapy sequence. Thus, the three most important mutations are given as follows: 16, 30 and 54 for the protease sequence; and 151, 70 and 230 for the reverse transcriptase sequence. According to Figure 7 the drugs comprising the current therapy ordered by their relevance are given by: *LPV TDF 3TC RTV ZDV SQV*, and the three most important drugs from the drugs administered in the history of the considered therapy sequence are: *LPV*, *RTV* and *DDI*.

More detailed insight of the impact of the complete training sample on the predictions for the target sample from Figure 3 a) in the main paper is depicted in Figure 4. It images the distribution of the training sample weights for the therapy sequence of the target sample. Moreover, Figure 5 depicts the distribution of the training sample weights for all samples in the test set.

3 Runtime information for training history-similarity models

One disadvantage of the history-similarity method is that it is quite compute-intensive, since it trains an individual model for each target sample. Therefore, we use the trust region Newton optimization for training logistic regression (Lin *et al.*, 2008) and thus provide an efficient way for training the individual models - one model is trained in a fraction of a second. Moreover, the alignment similarities of the target therapy sequence to all training therapy sequences is computed in about two minutes for our training data with 5230 samples. The only bottleneck is computing the pairwise similarity alignments for all training samples in the model selection procedure. However, they can be precomputed and stored for all different therapy sequences in the available clinical data set. Thus, new alignment scores need to be computed only if the training set is extended with new samples whose corresponding therapy sequences are not among the ones appearing in the previous version of the training data. Whenever we encounter such sequences we can compute their alignment scores for all training therapy sequences and store them together with the others. This enables fast model selection procedure whenever there is an update of the training data. More specifically, our tuning procedure screens 456 different value combinations for the two model selection parameters specified in the optimization problem in the Methods section of the main paper – 19 different values for the regularization parameter σ chosen equidistantly as powers of 10

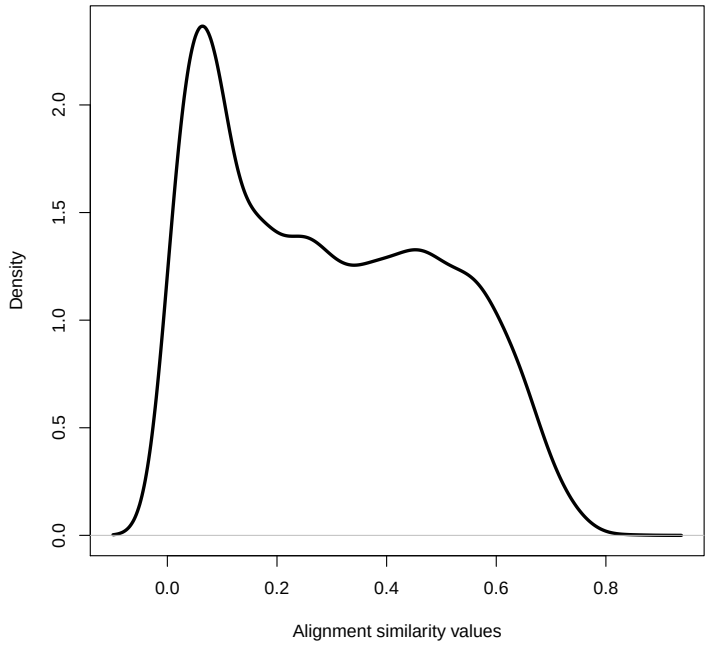


Figure 4: Distribution density of the sample-specific similarity weights for the training set corresponding to the therapy sequence depicted in Figure 3 a) in the main paper.

from the interval $[10^{-4}, 10^{5}]$, and 24 different values for the smoothing parameter γ specified by the set $\{0.1, 0.3, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 15, 20, 30, 50, 70, 100, 1000\}$. Since our tuning set comprises 261 different combination therapies, for a given precomputed similarity alignment kernel we train 119016 logistic regression models. Our implementation completes the model selection procedure in less than 12 hours and this procedure only needs to be repeated whenever there is an update of the training data set. Note that we precompute the similarity alignment kernel for four different values of its gap penalty parameter: $\{-0.1, -0.2, -0.3, -0.5\}$.

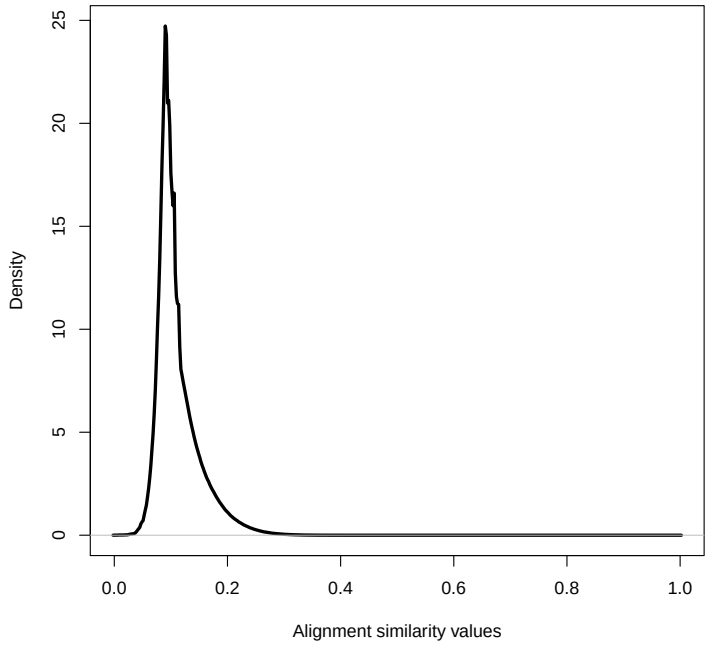
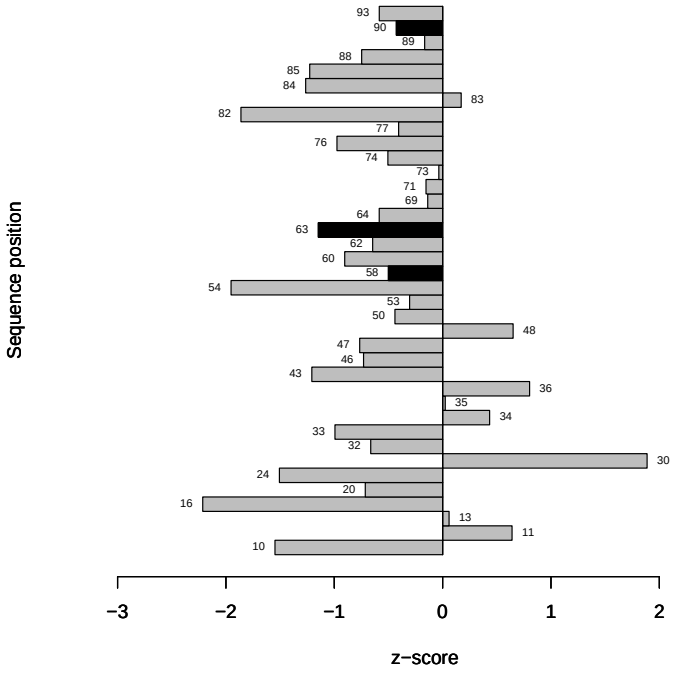
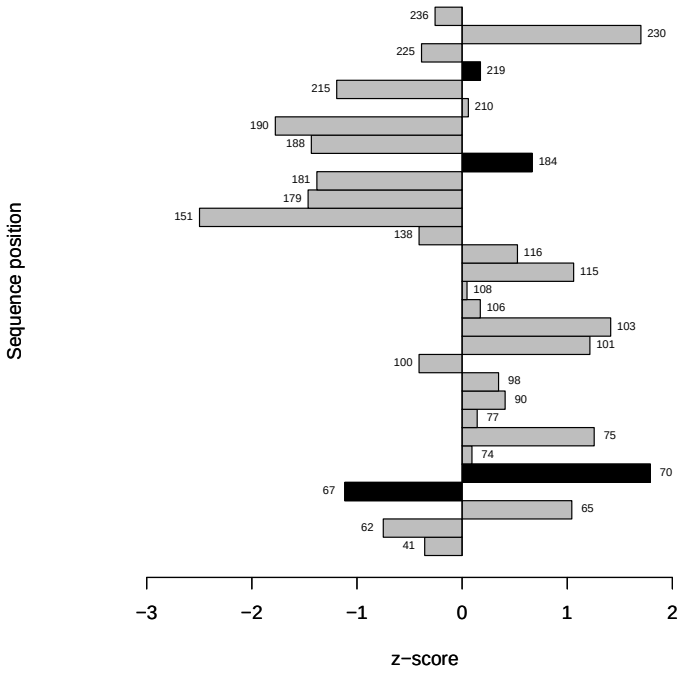


Figure 5: Distribution density of the sample-specific similarity weights for the training set corresponding to the therapy sequences of all test samples.

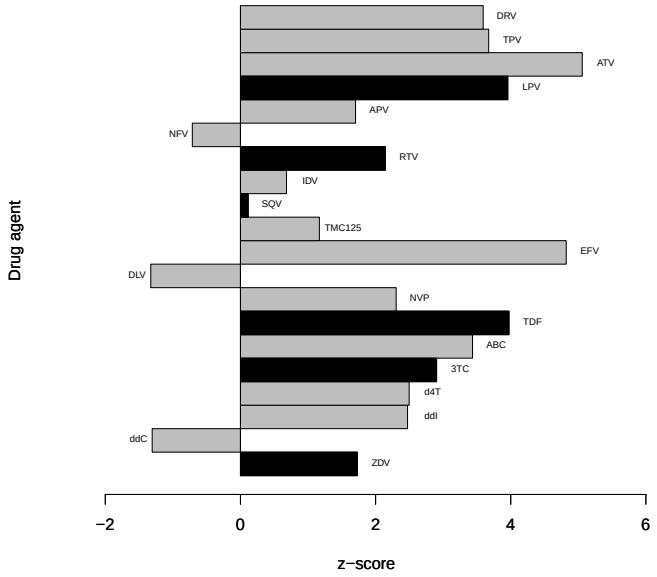


(a)

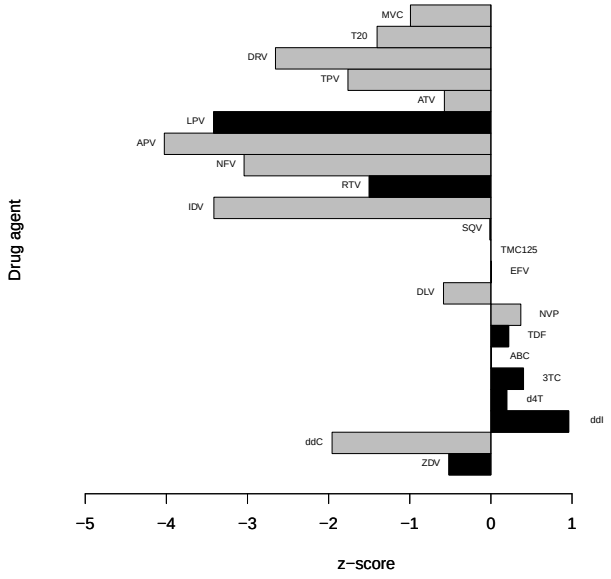


(b)

Figure 6: Barplot of z-scores showing the importance of the different viral sequence related input features: a) protease mutations; and b) reverse transcriptase mutations for the therapy sequence depicted in Figure 3 a) in the main paper. The features appearing in the specific target sample are depicted in black.



(a)



(b)

Figure 7: Barplot of z-scores showing the importance of the different drug input features: a) drugs comprising the current therapy; and b) drugs appearing in treatment history for the therapy sequence depicted in Figure 3 a) in the main paper. The features appearing in the specific target sample are depicted in black.