Drug-induced QT Prolongation Prediction Using Co-regularized Multi-view Learning

Jintao Zhang
Center for Bioinformatics
University of Kansas
Lawrence, KS 66047
Email: jtzhang@ku.edu

Jun Huan
Department of Electrical Engineering and Computer Science
University of Kansas, Lawrence, KS 66045
Email: jhuan@ittc.ku.edu

Abstract—Drug-induced QT prolongation is a major life-threatening adverse drug effect. It is crucial to predict the QT prolongation effect as early as possible in drug development, however, data on drugs that induce QT prolongation are very limited and noisy. Multi-view learning (MVL) has been applied to many challenging machine learning and data mining problems, especially when complex data from diverse domains are involved and only limited labeled examples are available. Unlike existing MVL methods that use $\ell_2$-norm co-regularization to obtain a smooth objective function, in this paper we proposed an $\ell_1$-norm co-regularized MVL algorithm for predicting drug-induced QT prolongation effect and reformulate the $\ell_1$-norm co-regularized objective function for deriving its gradient in the analytic form. $\ell_1$-norm co-regularization enforces sparsity in the learned mapping functions and hence the results are expected to be more interpretable. Comprehensive experimental comparisons between our proposed method and previous MVL and single-view learning methods demonstrate that our method significantly outperforms those baseline methods.

I. INTRODUCTION

Torsades de pointes (TdP) is a rare form of polymorphic ventricular tachycardia that exhibits distinct characteristics on the electrocardiogram (ECG), and is considered as a major life-threatening condition for patients as it can degenerate into ventricular fibrillation and sudden death [1]. In the past decade, the single most common cause of the withdrawal or restriction of the use of marketed drugs has been recognized as the prolongation of the QT interval associated with TdP [2]. Until 2004, at least nine drugs that were marketed in the U.S. or elsewhere have been removed from the market or had their availability restricted due to the risk of prolonging QT interval, such as Astemizole [3] and Grepafloxacin [4]. Therefore, it is crucial to discover and filter out those drug candidates with potential risk of QT prolongation and/or TdP as early as possible in the drug development pipeline to save development expenses and lives in clinical trials. Note that in this paper the terms QT prolongation and TdP risk are used interchangeably for ease of description.

TdP is often associated with a prolonged QT interval, which may be congenital or acquired. Drug-induced QT prolongation has triggered significant attention from pharmaceutical industry. Since 2005 the U.S. Food and Drug Administration (FDA) and European regulators have required that nearly all new molecular entities must be evaluated in a thorough QT study to determine its effects on the QT interval [5]. Although QT interval can be influenced by many factors, understanding of the underlying genetic mechanisms of QT prolongation have been significantly improved during the past decade. At present a few ion-channel related genes have been identified, and their gene products are found to profoundly influence the balance of ion-channel currents that determine the duration of the myocyte’s action potential and thus the QT interval [6], including three potassium ion channels $I_{Kr}(K_{r1.1}, hERG, KCNH2)$, $I_{K1}(K_{r1.2}, KCNJ2)$, and $I_{Kr}(K_{r1.7}, KCNQ1)$, one sodium ion channel $I_{Na}(Na_v1.5, SCN5A)$, and one calcium ion channel $I_{Ca_{L}}(Ca_{L}1.2, CACNA1C)$ [1]. Presumably, compounds that inhibit one or more these involved ion channels are more likely to obstruct the normal cardiac ion conduction and thus prolong the QT interval. This key observation is crucial for identifying compounds with potential TdP risk.

There are various means for pharmaceutical companies to obtain QT prolongation information induced by their interested compounds, such as animal models and human clinical trials, whereas the FDA Adverse Event Reporting System is another important source for such information. However, data obtained from these sources are either expensive and time-consuming or noisy, while computationally predicting this rare but serious adverse drug effect (ADE) in humans remains highly challenging. With thorough literature search we identified 142 compounds [1], [7] that are possibly associated with QT prolongation and TdP with different risk levels, including 28 compounds labeled as “risk”, 40 as “possible risk”, and 74 as “conditional or congenital risk” (http://www.azcert.org). It is not surprising that most of these drug compounds have such uncertain labels, since it is non-trivial to determine the causation between drugs and TdP risk, especially when multiple drugs are simultaneously taken by patients. With such limited and noisy data, traditional machine learning methods cannot build accurate models for predicting the QT prolongation and TdP risks of potential drug candidates. On the other hand, additional knowledge on the mechanism of QT prolongation are available but unused.

A variety of applications have limited and noisy labeled examples such as QT prolongation prediction, where traditional single-view learning methods work poorly. Multi-view learning (MVL) has been applied to many challenging machine learning and data mining problems, especially when complex data from diverse domains are involved and only limited
labeled examples are available. The underlying assumption of multi-view learning [8]–[10] is that different views are conditional independent and relatively complementary to one another, so that combining knowledge from multiple views can afford performance gain. Among MVL algorithms in the co-regularization [8] framework, all previous work used $\ell_2$-norm co-regularization for it is easy to optimize the smooth objective function. In this paper, we propose an $\ell_1$-norm co-regularized MVL algorithm to boost prediction performance on the important bioinformatics problem: prediction of QT prolongation effect using the limited and noisy data available. Noticeably, although this MVL algorithm is developed for predicting the effect of QT prolongation, the method itself is domain independent and could be applied to a wide range of other real-world applications with limited and noisy multi-view data.

All previous MVL algorithms that use $\ell_2$-norm co-regularization have a smooth objective function optimized by an alternate optimization approach [11], [12], i.e., in each iteration one view is optimized with the other views fixed until convergence of mapping functions in all views. However, we will demonstrate that our proposed $\ell_1$-norm co-regularized MVL method is more advantageous after properly reformulating the objective function and computing its gradient in the analytic form. Moreover, We developed a simultaneous optimization approach that optimizes the mapping functions from all views simultaneously.

For predicting drug-induced QT prolongation using multi-view learning, the first view is easily directed to drug chemical structures, while there is a few options for the second view. One choice is the general binding profiles of the drug compounds to human protein. More specifically, we can use the binding information of compounds to those human proteins associated with QT prolongation, i.e. the five ion-channel proteins mentioned previously. However, it is challenging to obtain binding data between the compounds interested and the five QT prolongation-related ion-channel proteins mentioned above. When the binding activity between a compound and one of the ion channels is unknown or unavailable, we build protein-chemical interaction prediction models to predict it until the activity matrix is filled up completely. By learning these two-view data, $\ell_1$-norm co-regularization enforces sparsity in the learned mapping functions and naturally carries the functionality of feature selection, and hence the prediction results are expected to be more accurate and interpretable. Comprehensive experimental comparisons are designed for demonstrating the advantages of our proposed $\ell_1$-norm co-regularized MVL method over reference methods on the prediction of drug-induced QT prolongation.

II. RELATED WORK

Many computational methods for facilitating the prediction of drug-induced QT prolongation and TdP risk have been developed and applied in various applications such as drug discovery and development. Yap et al. [13] used linear solution energy relationships descriptors to measure the TdP-causing potential of compounds and built prediction models with support vector machines (SVMs). Their method simply ignored the different confidence levels of data and considered all “risk”, “possible risk”, and “conditional or congenital risk” compounds as positive samples, and the interpretability and meaning of their results were questionable. Recognizing the fact that the hERG protein is frequently associated with QT prolongation, Yao et al. [14] and Redfarn et al. [15] developed models using hERG inhibition activity and other relevant data to predict drug-induced QT prolongation, and investigated its relationships with various electrophysiological properties. The limitation is that hERG is not the only or necessary binding protein to induce this serious ADE. In addition, Champerous et al. [16] classified training examples into three categories according to their confidence levels, and developed a complex algorithm called TPDScreenTm combined with a database from reference compounds with available clinical data to predict the risk of TdP and QT prolongation at the early stage of drug development.

Given the fact that the development of QT prolongation and TdP is associated with multiple genes [6], simulation studies of multiple ion channels blockade have been conducted by Mirams et al. [17]. They collected multiple ion channels (hERG, $I_{Na}$, and $I_{Cal}$) data on 31 drugs associated with various risk of TdP, and performed simulations with a series of mathematical models of cardiac cells to integrate information on multi-channel block, resulting in improved prediction of TdP risk. Ouilé et al. [1] picked five TdP-associated ion channels, including three potassium channels ($I_{K_R}$, $I_{K_{1}}$, and $I_{K_{5}}$), one sodium ($I_{Na}$) and one calcium ($I_{Cal}$) channel, and elucidated the mechanisms of drug-induced TdP using ion channel blocking profiles. They studied the effects of known torsadogenic and non-torsadogenic compounds on these ion channels, investigated their functions in the generation of drug-induced TdP and QT prolongation, and identified that $I_{Na}$ and $I_{K_{1}}$ play roles that are as important as $I_{K_{R}}$ in safety pharmacology. This study motivated us to use the blocking profiles of the five ion channels as the second view to improve our MVL models.

The rest of this paper is organized as follows: in Section II we introduce our new MVL methods in more detail. Data collection, feature extraction, and all experimental and comparison results are described in Section III. We then conclude our work in Section IV with insightful observations.

III. METHODS

In our data set, each example has two view vectors. For labeled examples, the two view vectors are associated with the same label, while unlabeled examples doesn’t have a label but can be helpful on improving model performance. In this section we discuss our proposed $\ell_1$-norm co-regularized MVL algorithm in detail, including how we integrate information from all views and derive the problem formulation step by step. We also derive the analytic form of the gradients of our objective functions for simultaneous optimization, as compared to the traditional alternate optimization technique.

A. $\ell_1$-norm Co-regularized Multi-view Learning

Suppose we have $V$ views and for each view $v = 1, 2, ..., V$, use $X_v \in \mathbb{R}^{N \times M_v}$ to denote the feature matrix with $N$
By applying a linear mapping function $h^v$ on each view $v$ so that these functions agree each other on an unlabeled example as much as possible while the error on the labeled examples is minimized. The final prediction function $h$ is defined as:

$$h(x_i) = \frac{1}{V} \sum_{v=1}^{V} h^v(x_{i,v}),$$  

(1)

where $x_{i,v}$ represents the feature vector in the $v$ view for example $i$. The view mapping functions $h^1, h^2, \ldots, h^V$ are obtained by minimizing the following objective function over these view functions:

$$F(h^1, \ldots, h^V) = \sum_{i \in L} L(y_i, h(x_i)) + \sum_{v=1}^{V} \lambda_v ||h^v||_p + \mu \sum_{v \neq v_1} ||h^v(X_v) - h^{v_1}(X_{v_1})||^2,$$  

(2)

where the first term is the loss function that penalizes the misclassification on labeled examples, the second term is to regularize the mapping function on each view with $\ell_p$-norm ($1 \leq p \leq +\infty$) so that irregular solutions will be filtered out, and the final term is to penalize the disagreement among the prediction results from different view functions and to drive them to agree one another as much as possible. Here parameters $\{\lambda_v\}$ are the strength of $\ell_1$-norm regularization terms in all views, and $\mu$ is the coupling parameter that regularizes the prediction disagreement using unlabeled data. By minimizing the sum of these three terms, co-regularization based MVL aims to identify a set of regular mapping functions that minimize misclassification rates on labeled examples and agree one and another on the predictions of unlabeled examples as much as possible.

For simplicity, without loss of generality we consider only two views and take least square loss on labeled examples and use $\ell_1$-norm regularization on both views, resulting in the following simplified objective function:

$$F(h^1, h^2) = \sum_{i \in L} ||y_i - h(x_i)||_2^2 + \lambda_1 ||h^1||_1 + \lambda_2 ||h^2||_1 + \frac{1}{2} \sum_{v \in U} ||h^1(x_{i,v}) - h^2(x_{i,v})||_2^2,$$  

(3)

By applying a linear mapping function $W_v$ on view $v = 1, 2$, $h^v(X_v) = X_v W_v$, where $W_v$ is a $M_v \times 1$ column coefficient vector for view $v$, we have a metricized objective function as follows:

$$F(W_1, W_2) = \|Y - \frac{1}{2} (X_1^T W_1 + X_2^T W_2)\|_2^2 + \lambda_1 ||W_1||_1 + \lambda_2 ||W_2||_1 + \frac{1}{2} \mu ||X_1^T W_1 - X_2^T W_2||_2^2,$$  

(4)

If $\ell_2$-norm co-regularization is used in Equation (4), this objective function is convex and smooth, and is usually optimized using the alternate optimization approach [11], [12], i.e., alternately optimizing one view with the other view fixed until convergence. When we use $\ell_1$-norm co-regularization as in Eqn.(4), the objective function becomes convex but not smooth. However, by reformulating this function properly we can optimize both views simultaneously. Let $[.]$ denote the horizontal concatenation of two matrices $Z_1 \in R^{n \times p}$ and $Z_2 \in R^{n \times q}$, and then $[Z_1, Z_2] \in R^{n \times (p+q)}$. We concatenate column vectors $W_1, W_2$ and matrices $X_1, X_2$ as $W = [W_1^T, W_2^T]^T$, $X_L = [X_1, X_2]_{x \in L}$, and $X_U = [X_1 - X_2]_{x \in U}$. The objective function $F$ in Eqn.(4) can be rewritten as

$$F(W) = \|Y - \frac{1}{2} X_L W\|_2^2 + \lambda^T \|W\|_1 + \frac{1}{2} \|X_U W\|_2^2,$$

$$= (\frac{1}{2} X_L W)^T (Y - \frac{1}{2} X_L W) + \lambda^T \|W\|_1 + \frac{1}{2} \mu (X_L W)^T X_U W,$$  

(5)

where $(\lambda)$ is a $(M_1 + M_2) \times 1$ constant vector with the first $M_1$ elements as $\lambda_1$ and the remaining $M_2$ elements as $\lambda_2$. It is straightforward that we can derive the analytic form of the gradient of the function in Equation (5) as follows:

$$\frac{\partial F(W)}{\partial W} = -X_L^T (Y - \frac{1}{2} X_L W) + \mu X_U^T X_U W + \lambda \text{ sign}(W),$$  

(6)

We then use the L-BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) method [18], [19] to minimize the objective function in Equation (5). With the known analytic form of the gradient as in Equation (6), the optimization process can be very efficient.

B. $\ell_2$-norm Co-regularized Multi-view Learning

The $\ell_2$-norm co-regularized MVL method has been discussed in many previous work [8], [9]. The objective function of this MVL method is as follows:

$$F(W_1, W_2) = \|Y - \frac{1}{2} (X_1^T W_1 + X_2^T W_2)\|_2^2 + \lambda_1 ||W_1||_2 + \lambda_2 ||W_2||_2 + \frac{1}{2} \mu ||X_1^T W_1 - X_2^T W_2||_2^2,$$  

(7)

Some previous works use an alternate optimization method to identify optimal $W_1$ and $W_2$. The alternate process can be very time-consuming, however we can apply some reformulation tricks and implement it in a more efficient matter. Similar to $\ell_1$-norm co-regularization, the gradient of this objective function can be written as

$$\frac{\partial F(W)}{\partial W} = -X_L^T (Y - \frac{1}{2} X_L W) + \mu X_U^T X_U W + \lambda_1 \frac{W_1}{||W_1||_2} + \lambda_2 \frac{W_2}{||W_2||_2},$$  

(8)
where variables $W$, $Y$, $X_1$, $X_2$, and $U$ have the same meaning as in the previous subsection. $W_{11}$ and $W_{22}$ are vectors with the same length of $W$, and are obtained by appending 0’s at the end of the original vector $W_1$ and at the beginning of $W_2$, respectively. With this gradient we can optimize the objective function in Equation (7) similarly as for our $\ell_1$-norm co-regularized MVL method.

IV. RESULTS

In this section, we began our discussion with an introduction to how we collected, cleaned, and represented the TdP-related drugs and those compounds with known binding activity to available human proteins, including the five TdP-associated ion-channel proteins. For performance comparison, three reference methods are selected as comparison baselines. SVMs [20] is a widely used single-view supervised learning algorithm with excellent performance. The LIBSVM [21] implementation and RBF kernels were used in all experiments. As a generalized linear method, $\ell_1$-regularized logistic regression (LLR) [22] is another excellent single-view supervised learning method can construct a model that estimates probabilities, e.g. for medical diagnosis and credit scoring. In this paper, we used the algorithm proposed by Boyd et al. [22] to run our experiments of the LLR method. The third reference method is the $\ell_2$-norm co-regularized MVL method, which has been discussed in many previous works [8], [9] and introduced in the METHODS section.

A. Data Set and Feature Extraction

We collected the compound set of drug-induced QT prolongation from the home page of University of Arizona CERT (http://www.azcert.org/) [7] and the work by Ouillé et. al [1]. We merged these two data sets by removing redundant and/or inconsistent data points, resulting in a set of 28 drug compounds labeled as “TdP risk”, 40 labeled as “possible TdP risk”, and 74 with other less reliable evidence of QT prolongation. In our experiments we treated the 28 drugs with “TdP risk” labels as positive examples, the 40 drugs with “possible TdP risk” labels as positive examples, the 40 drugs with “possible TdP risk” labels as positive examples, and the remaining 74 compounds as unlabeled examples whose labeled to be determined by our MVL models. For negative examples, we extracted all approved drugs from the DrugBank database [23], excluded all drug compounds that are associated with QT prolongation and/or TdP risk with even very weak evidence, and obtained a set of 1,221 drug compounds as putative negative examples.

We used molecular signature descriptors [24] to extract features from chemical structures. In a molecular signature vector, each component is the number of occurrences of a particular atomic signature in the molecule. An atomic signature is a canonical representation of the substructure surrounding a particular atom, including all atoms and bonds up to a predefined number (called the signature height) of consecutive bonds from the given atom. We set the signature height at 2 for all compounds, filtered out all atomic signatures with frequency less than 4%, and finally obtained about 100 features for each compound. We used drug signature descriptors as the first view of our data set.

In our experimental study, we first randomly selected an equal number of negative and positive examples and combined them with unlabeled data to make a balanced subset, from which we randomly selected 20% positive and 20% negative examples for testing. We used the remaining 80% labeled and all unlabeled examples for training with 10-fold cross validation to select the best model parameters. For each experimental setting, we repeated the experiment for 50 times using random sampling, and reported the mean and standard deviation of testing accuracy, precision, and recall, which are defined as $(TP+TN)/S$, $TP/(TP+TN)$, and $TP/(TP+FN)$, respectively, where $TP$ is true positive, $TN$ is true negative, $FN$ is false negative, and $S$ is the total number of testing examples.

B. Generating the second view

In the MVL framework, each data example has multiple feature vectors, one from each view. Our first view is taken from compound structures also using molecular signature descriptors [24]. We set signature distance at 2 and frequency threshold at 4%, and ended up with a vector of $\sim$90 non-negative features for each compound. We first used the binding profiles of all drug compounds to the five ion-channel proteins as the second view. For the binding activity that are unknown or unavailable, we have built protein-chemical interaction prediction models to find appropriate decisions, resulting in a vector of five binary elements, in which a bit is set to 1 if a drug compound inhibits an ion channel, and 0 otherwise. To this end, we manually searched for binding compounds of the five picked ion channels from the PubChem database. For the potassium ion channels $I_{KR}$, $I_{K1}$, and $I_{KS}$, we identified multiple target-based bioassays that use them as target proteins, and then selected bioassays with AID 376, 2032 and 2642 for each of them, respectively. Each assay provides hundreds of active compounds and a large number of inactive compounds. For the calcium ion channel $I_{CaL}$ and the sodium ion channel $I_{Na}$, we identified multiple related target-based assays with few tested compounds. We decided to use those tested compounds with activity less than 1µM as active and the rest as inactive, and obtained 8 (53) weakly active and 17 (24) inactive compounds for the ion channel $I_{CaL}$ ($I_{Na}$).

We randomly selected 100 active and 100 inactive compounds for the potassium ion channels as a balanced data subset and used all active and inactive compounds identified for the sodium and calcium channel. The model construction and selection process is as follows: first 20% active and 20% inactive compounds were randomly selected as the testing set and the rest 80% as the training set, and 10-fold cross validation were applied to the training set with support vector
machines (SVMs) and RBF kernels to build one model for each ion channel using its own binding compounds. In 10-fold cross validation, the training data was randomly split into 10 equal disjoint subsets. Nine subsets were used for training a model, and the rest subset was used as a validation set for validating the model. The cross-validation process repeated for 10 times with each subset used exactly once as the validation set. Each experiment was repeated for ten times, and a compound was considered to be active to an ion channel if it was predicted as active for six or more times. Eventually our models have an average accuracy of about 73-78%, as shown in Table I.

C. Experimental Results

In this section, we compared our proposed \( \ell_1 \)-norm co-regularized MVL algorithm (\( \ell_1 \)-MVL) with SVMs, \( \ell_1 \)-regularized logistic regression (LLR), and the \( \ell_2 \)-norm co-regularized MVL method using both our proposed optimization (\( \ell_2 \)-MVL) and alternate optimization (\( \ell_2 \)-MVL-ALT) techniques, and evaluated their performance using testing accuracy, precision, and recall. Note that our results were obtained from 20 repeated experiments, so the performance differences between different methods should be considered significant. We tested these five methods at various experimental settings to investigate the contribution and influence of the related factors on prediction performance.

First we use the ion-channel binding profiles as the second view, and the data set consists of 28 positive, 40 possible positive, 74 unlabeled and 1,221 putative negative examples. If we used only 28 positive examples and treated the 40 possible positive examples as unlabeled, we then have 114 unlabeled examples. The results of the five methods were summarized in the first five rows in Table II. We found that our \( \ell_1 \)-MVL method significantly outperformed not only the SVL methods (SVMs and LLR) for 5-6%, but also the \( \ell_2 \)-MVL method for about 3.3% on prediction accuracy. For precision, the \( \ell_1 \)-MVL method was slightly better than SVMs and LLR, and much better than \( \ell_2 \)-MVL methods. All MVL methods have much better recall than SVMs and LLR. All these performance metrics consistently reveal that our proposed \( \ell_1 \)-MVL method performs better than the reference methods. Moreover, to investigate the contribution of the unlabeled examples to prediction performance, we replaced the matrix \( X_u \) with a zero matrix of the same dimensionality, and performed similar experiments described above. We listed the results in the last three rows in Table II. We observed apparent decrease of the \( \ell_1 \)-MVL method on accuracy and recall, while the precision results are still comparable or slightly better than previous experiments. We can approximately estimate the contribution of unlabeled examples in the the \( \ell_1 \)-MVL method is about 1.1%, which is non-trivial considering that the total accuracy improvement of our MVL method is only 5-6% compared to SVL methods. Although the ion-channel binding profiles has very small dimensionality as the second view, its contribution to the performance improvement is significant by comparing the results of the MVL methods with the SVL methods.

If we gave them more confidence on the 40 drug compounds labeled as “possible TdP risk” and used them as positive examples, we would have more labeled and less unlabeled examples. With 68 positive examples, we did similar experiments to investigate the influence of the number of labeled examples on prediction performance. All results with the same experimental settings as above are summarized in right columns in Table II. By using those “possible positive examples” as positive examples, we observed significant performance decrease on all methods. Although the \( \ell_1 \)-norm co-regularized MVL method is still better than the baseline methods, the performance advantage shrank to 1-2%. By adding unreliable positive examples to our data set, we introduced many false positives and thus compromised the prediction accuracy. Moreover, the precision of the \( \ell_1 \)-MVL method decreased about 10%, which more clearly demonstrates that the data set contains much more false positives than before, according to the definition of precision. We also found that the precision of \( \ell_2 \)-MVL method dropped only about 3%. The observation showed that our proposed \( \ell_1 \)-MVL method is more sensitive to false positive examples. Finally, it is easy to notice that the performance of the MVL methods fluctuated very slightly when we have different numbers of unlabeled examples, apparently because the contribution of unlabeled examples have been offset by introducing more false positive examples.

D. Contribution of the Second View

In this section we discussed the contribution of the ion-channel protein binding profiles of the drug compounds as the second view to final models, since this view significantly improved the performance of our \( \ell_1 \)-norm co-regularized MVL method, comparing with the single-view learning methods such as SVM and LLR. In the final models, we extracted the model coefficients of these five binary features from the second view, and calculated the mean and standard deviation of these coefficients under the four different experimental setting previously described. According to the fourth chapter in [25], the mean to standard deviation ratio, called Z-score, of a feature’s model coefficient measures the significance of this feature’s contribution to the final models. From Table III we can find that the ion channel \( I_{K1} \) plays a relatively more important role since its Z-score \( \in [0.92,1.33] \) is much greater.

### Table II
PREDICTING ACCURACY USING ION-CHANNEL BINDING PROFILES. STANDARD DEVIATIONS INCLUDED IN PARENTHESIS.

| Method       | \( |X_u| \) | \( |X_L| \) | \( \ell_1 \)-norm | \( \ell_2 \)-norm | \( \ell_1 \)-MVL | \( \ell_2 \)-MVL-ALT |
|--------------|---------|---------|-------------|-----------------|-------------|---------------|
| SVM          | /       | 0.727(0.060) | 0.799(0.096) | 0.725(0.105) | /            | 0.687(0.049) |
| LLR          | /       | 0.712(0.078) | 0.779(0.105) | 0.658(0.133) | /            | 0.676(0.074) |
| \( \ell_1 \)-MVL | 114     | 0.771(0.070) | 0.802(0.085) | 0.825(0.096) | 74           | 0.716(0.038) |
| \( \ell_2 \)-MVL | 114     | 0.748(0.076) | 0.739(0.077) | 0.817(0.091) | 74           | 0.693(0.062) |
| \( \ell_2 \)-MVL-ALT | 114     | 0.744(0.075) | 0.736(0.080) | 0.805(0.076) | 74           | 0.696(0.066) |
than other ion channels. This finding also partly matches the observation in [1], in which ion channel $I_{Na}$ and $I_{K1}$ are found to play a key role in the generation of drug-induced QT prolongation and Tdp. In our models the importance of the ion channel $I_{Na}$ was not identified, mainly because we have very few training examples (8 weak positives and 17 negatives) available for this ion channel in the data set. In addition, the Z-scores of $I_{CaL}$ and $I_{KS}$ are very close to zero, and can be reasonably considered as the evidence of the sparsity of our final models, i.e., two out of the five ion-channel feature was considered no significant relevance, and we can set the corresponding coefficients as zero.

V. DISCUSSION

In this paper we aim to construct accurate prediction models with very limited and noisy data that are available, and propose a novel $\ell_1$-norm co-regularized multi-view learning algorithm on predicting drug-induced QT prolongation. We conduct experiments under various settings to investigate the contribution and influence of the number of unlabeled examples and labeled examples. Experimental results show that our $\ell_1$-MVL method outperforms not only two excellent SVL methods (SVMs and LLR), but also the widely used $\ell_2$-norm co-regularized MVL method with the measurement of testing accuracy, precision, and recall.

To handle the challenge of limited and noisy labeled examples, we extract the second view for all Tdp related drug compounds from critical related works, and pick five ion-channel proteins that have been found highly relevant to the QT prolongation effect. The contribution of the second view has been clearly demonstrated by our comprehensive experiments and significant higher prediction accuracy in both the $\ell_1$-norm and $\ell_2$-norm co-regularization framework. From the coefficients in the final models, we analyzed the significance of the contribution of each ion channel to QT prolongation, and our finding partly matched previous work [1].

ACKNOWLEDGMENT

This work has been supported by the KU Specialized Chemistry Center (NIH U54 HG005031) and NIH grant R01 GM868665.

TABLE III

<table>
<thead>
<tr>
<th>$\lambda_L$</th>
<th>$\lambda_u$</th>
<th>$I_{Na}$</th>
<th>$I_{CaL}$</th>
<th>$I_{KR}$</th>
<th>$I_{KS}$</th>
<th>$I_{K1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>114</td>
<td>0.308</td>
<td>0.061</td>
<td>0.300</td>
<td>0.082</td>
<td>1.527</td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>-0.134</td>
<td>-0.077</td>
<td>0.042</td>
<td>-0.024</td>
<td>1.916</td>
</tr>
<tr>
<td>68</td>
<td>74</td>
<td>0.337</td>
<td>0.039</td>
<td>0.615</td>
<td>-0.090</td>
<td>0.923</td>
</tr>
<tr>
<td>68</td>
<td>0</td>
<td>0.434</td>
<td>-0.020</td>
<td>0.454</td>
<td>-0.097</td>
<td>0.968</td>
</tr>
</tbody>
</table>

REFERENCES