Acute Myeloid Leukemia Outcome Prediction

Zhilin Yang

Tsinghua University

Introduction

Motivation

Acute Myeloid Leukemia (AML) is a cancer of the bone marrow and the blood. There will be at least 18,860 new cases of AML and 10,460 deaths from AML in 2014. Accurately predicting AML outcome and identifying important features can contribute to better clinical treatment.

Data

We have 191 samples of patients, with 40 clinical covariates and 231 Reverse Phase Protein Array (RPPA) measurements.

Problems

We try to computationally identify discriminative features and statistically develop predictive models for the following tasks:

1. (Remission) Predict which AML patients will have Complete Remission or will be Primary Resistant.
2. (Remission Duration) Predict the remission duration for patients who have Complete Remission.
3. (Survival Time) Predict the overall survival time for each patient
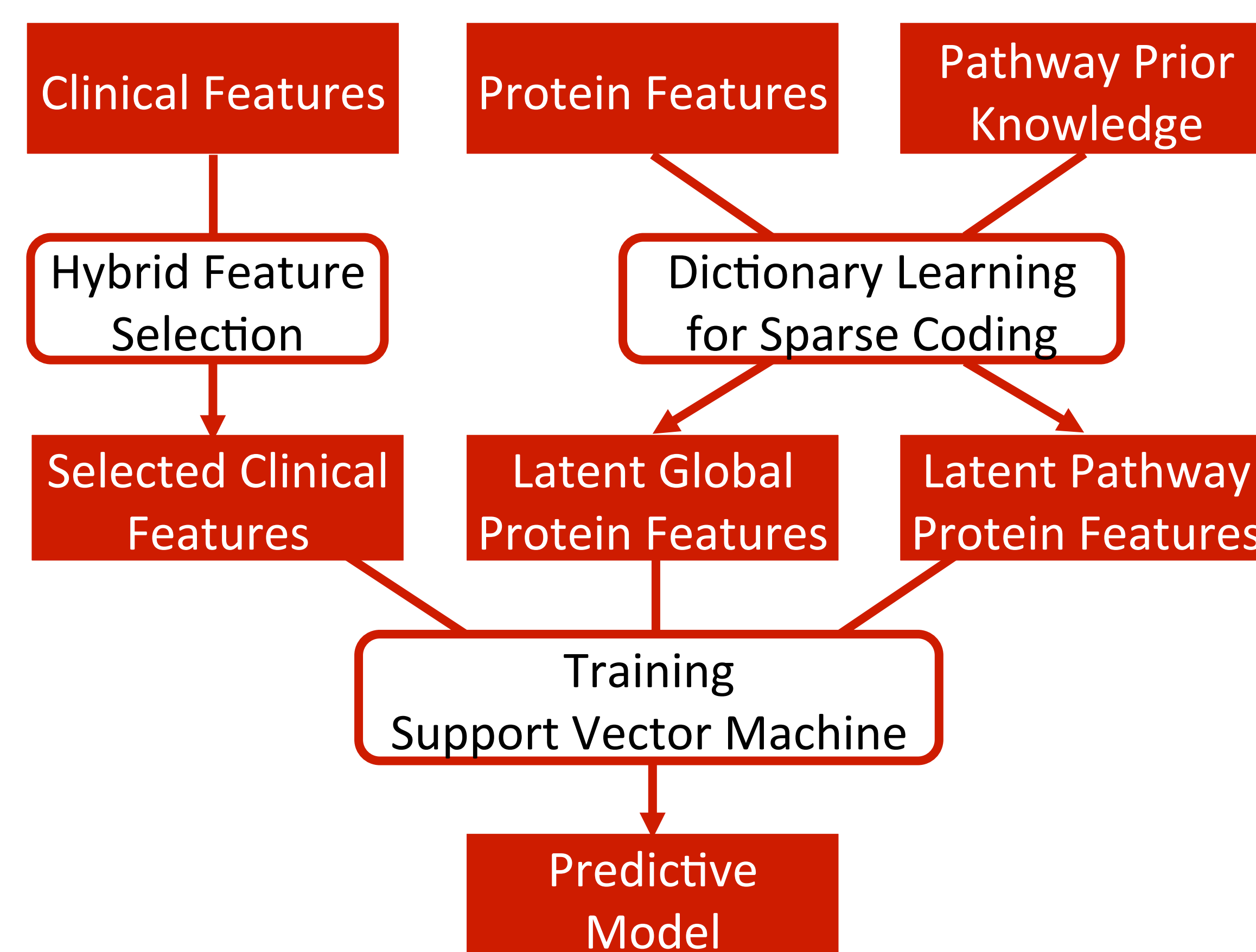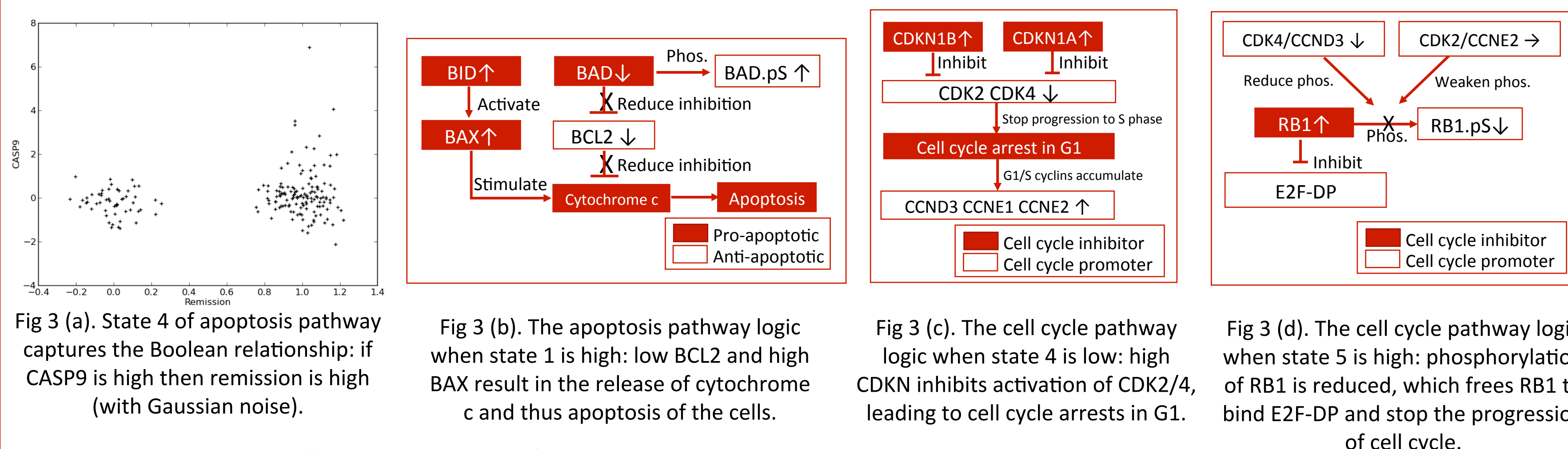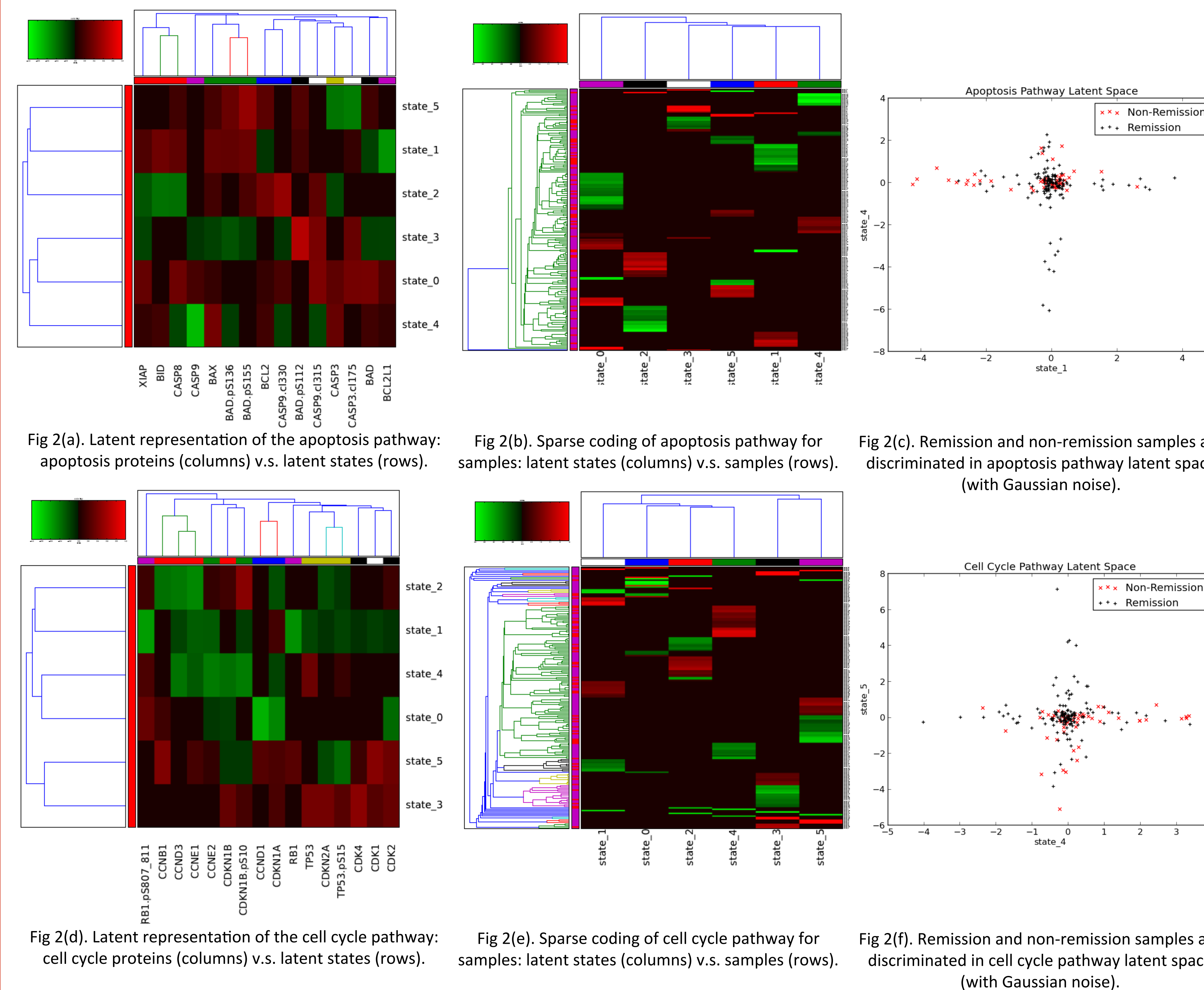
Methods

Fig 1. A sketch of AML Outcome Prediction Pipeline

Hybrid feature selection: We use Fisher's exact test for ranking categorical features and Spearman's correlation coefficient for ranking numerical features.
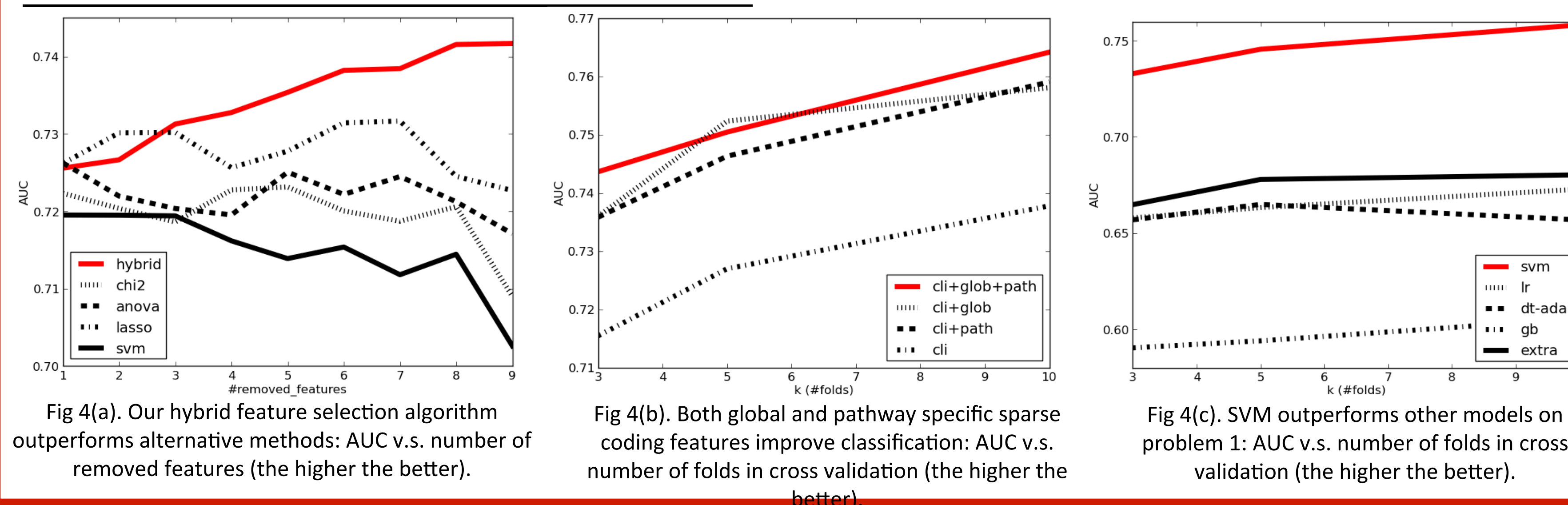
Dictionary learning for sparse coding: We learn the latent states (dictionary) from the data and the sparse linear combination of latent states for each sample[1].

Results

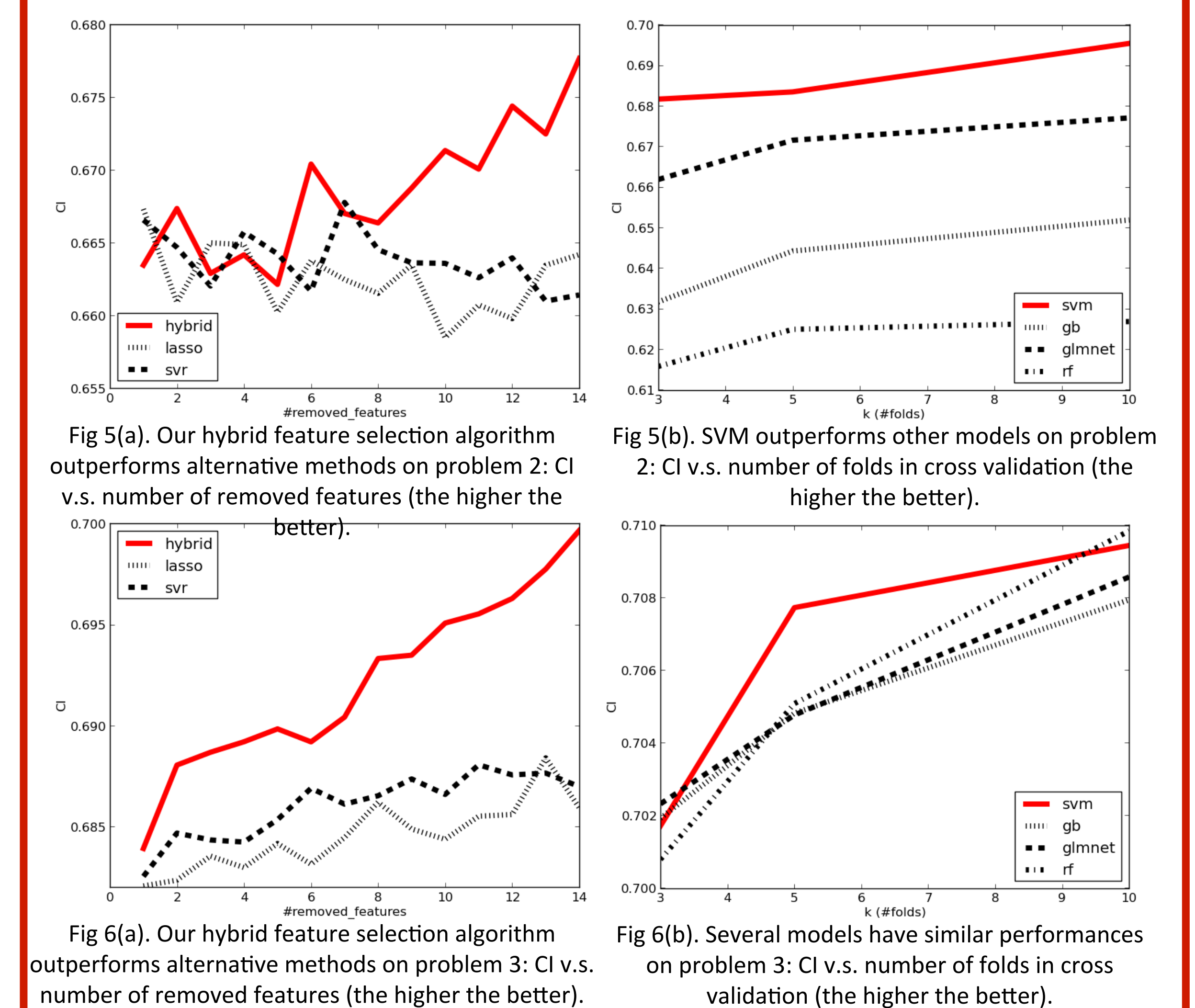### Latent Representation and Sparse Coding of Pathways

Fig 2(a). Latent representation of the apoptosis pathway: apoptosis proteins (columns) v.s. latent states (rows).

Fig 2(b). Sparse coding of apoptosis pathway for samples: latent states (columns) v.s. samples (rows).

Fig 2(c). Remission and non-remission samples are discriminated in apoptosis pathway latent space (with Gaussian noise).

Fig 2(d). Latent representation of the cell cycle pathway: cell cycle proteins (columns) v.s. latent states (rows).

Fig 2(e). Sparse coding of cell cycle pathway for samples: latent states (columns) v.s. samples (rows).

Fig 2(f). Remission and non-remission samples are discriminated in cell cycle pathway latent space (with Gaussian noise).

Fig 3 (a). State 4 of apoptosis pathway captures the Boolean relationship: if CASP9 is high then remission is high (with Gaussian noise).

Fig 3 (b). The apoptosis pathway logic when state 1 is high: low BCL2 and high BAX result in the release of cytochrome c and thus apoptosis of the cells.

Fig 3 (c). The cell cycle pathway logic when state 4 is low: high CDKN inhibits activation of CDK2/4, leading to cell cycle arrests in G1.

Fig 3 (d). The cell cycle pathway logic when state 5 is high: phosphorylation of RB1 is reduced, which frees RB1 to bind E2F-DP and stop the progression of cell cycle.

Prediction Performance of Problem 1

Fig 4(a). Our hybrid feature selection algorithm outperforms alternative methods: AUC v.s. number of removed features (the higher the better).

Fig 4(b). Both global and pathway specific sparse coding features improve classification: AUC v.s. number of folds in cross validation (the higher the better).

Fig 4(c). SVM outperforms other models on problem 1: AUC v.s. number of folds in cross validation (the higher the better).

Results

Prediction Performance of Problem 2 and 3

Fig 5(a). Our hybrid feature selection algorithm outperforms alternative methods on problem 2: CI v.s. number of removed features (the higher the better).

Fig 5(b). SVM outperforms other models on problem 2: CI v.s. number of folds in cross validation (the higher the better).

Fig 6(a). Our hybrid feature selection algorithm outperforms alternative methods on problem 3: CI v.s. number of removed features (the higher the better).

Fig 6(b). Several models have similar performances on problem 3: CI v.s. number of folds in cross validation (the higher the better).

Conclusions

Learning Pathway Regulation States

We propose a novel idea of learning sparse latent representation for RPPA data with prior signaling pathway knowledge. We demonstrate that dictionary learning for sparse coding can effectively capture latent states that have both biological implications about pathway regulation and statistical power for remission prediction.

Hybrid Feature Selection

We propose a hybrid feature selection method to avoid mixing up numerical and categorical features. We demonstrate that the hybrid feature selection method significantly outperforms alternative methods.

Future Work

It is interesting to apply the feature learning method to more RPPA datasets for identifying novel pathway regulation.

References

[1] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. "Online dictionary learning for sparse coding." In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 689-696. ACM, 2009.