A Pipeline for Integrated Theory and Data-Driven Modeling of Biomedical Data

Vineet K. Raghu, Xiaoyu Ge, Arun Balajiee, Daniel J. Shirer, Isha Das, Panayiotis V. Benos, and Panos K. Chrysanthis

Abstract—Genome sequencing technologies have the potential to transform clinical decision making and biomedical research by enabling high-throughput measurements of the genome at a granular level. However, to truly understand mechanisms of disease and predict the effects of medical interventions, high-throughput data must be integrated with demographic, phenotypic, environmental, and behavioral data from individuals. Further, effective knowledge discovery methods must infer relationships between these data types. We recently proposed a pipeline (CausalMGM) to achieve this. CausalMGM uses probabilistic graphical models to infer the relationships between variables in the data; however, CausalMGM's graphical structure learning algorithm can only handle small datasets efficiently. We propose a new methodology (piPref-Div) that selects the most informative variables for CausalMGM, enabling it to scale. We validate the efficacy of piPref-Div against other feature selection methods and demonstrate how the use of the full pipeline improves breast cancer outcome prediction and provides biologically interpretable views of gene expression data.

Index Terms—Genomics, graphical models, feature selection, phenotype prediction

1 Introduction

Since the advent of high-throughput sequencing methods, a number of modeling approaches have been developed to predict patient outcome from genomic data [1], [2]. To understand the complex relationships between genomics and outcomes, the genomic data should be integrated with clinical and demographic information. Despite the widespread success of machine learning methods, they are often insufficient to model this data [3] because they have highly correlated sets of variables (genes), and are high-dimensional (i.e., have several orders of magnitude more variables than samples). Furthermore, for biomedical research both predictive power and model interpretability are equally important. Often, biomedical researchers aim to learn from their models, to generate promising new hypotheses or prioritize future experiments.

Probabilistic Graphical Models (PGMs) are an effective tool to build interpretable models [4]. These models represent a dataset as a graph where nodes correspond to features and edges correspond to dependence relationships. Learning the structure of these models from data is a well-studied problem for continuous or categorical data but not for mixed data. Recently, several approaches have been proposed to model mixed data [5], [6], [7], [8]. We proposed a two-step approach called CausalMGM [8] to model a dataset as a directed causal graph through an intermediate undirected graph. CausalMGM was successful in modeling clinical data for patients with chronic obstructive pulmonary disease (COPD) [8], malignant nodules in the lung [9], identifying genetic biomarkers of response to cancer therapy [10] and microbiota affecting pneumonia onset in ICU patients [11]. However, three remaining issues of CausalMGM need to be addressed: 1) Mixed graphical model learning is computationally intractable on datasets with more than 2,000 variables (e.g., genomic data), 2) interpreting a large graphical model is difficult unless single variables of interest are queried, and 3) highly correlated data can result in the formation of disconnected cliques in the output graph, impeding model accuracy [12].

Time complexity can be addressed by selecting a subset of variables to model (i.e., Feature Selection). The key is to find the subset of features that maintain the maximum information for target variable(s) of interest. Though these approaches are applied to integrated biomedical data, they fail to address the remaining challenges. High correlations among features result in unstable prediction models and harm interpretability of learned models [13].

Use of prior knowledge has been proposed as a way to address these difficulties [2], [14], [15]. These sources allow a researcher to choose the most biologically plausible model among statistically equivalent models [13], [16]. However, many proposed methods have shown no significant benefit from using prior knowledge [2], [17]. Our hypothesis for this is twofold. First, external data sources need to be evaluated and weighted accordingly due to data provenance and

- *Vineet K. Raghu and Panayiotis V. Benos are with the Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260 USA and also with the Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15260 USA. E-mail: vineet@cs.pitt.edu, benos@pitt.edu.*
- *Xiaoyu Ge, Arun Balajiee, Daniel J. Shirer, and Panos K. Chrysanthis are with the Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260 USA. E-mail: {xig34, arl122, djs134}@pitt.edu, panos@cs.pitt.edu.*
- *Isha Das is with the Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15260 USA. E-mail: isha.das@outlook.com.*

Manuscript received 28 Jan. 2020; revised 17 June 2020; accepted 26 July 2020. Date of publication 25 Aug. 2020; date of current version 3 June 2021. (Corresponding author: Vineet K. Raghu.)
Digital Object Identifier no. 10.1109/TCBB.2020.3019237

1545-5963 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: University of Pittsburgh Library System. Downloaded on November 11,2025 at 19:44:41 UTC from IEEE Xplore. Restrictions apply.

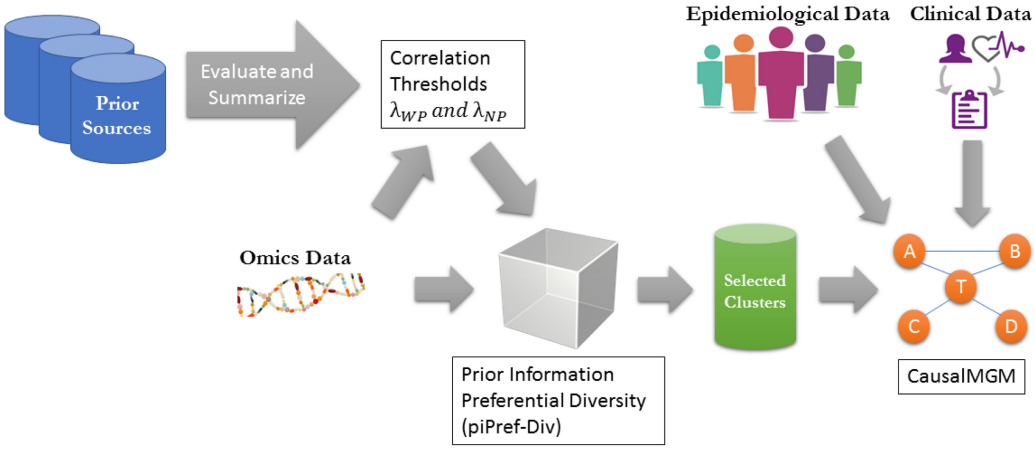


Fig. 1. Pipeline proposed in this work to learn graphical model structure from mixed clinical and omics datasets.

context-specific information. For example, a biological pathway used as a prior information source may not be active in the context in which a genomic dataset was measured [7]. This source of information should be downweighted when learning the final model. Second, multiple sources of prior information should be integrated to achieve consistent and robust results.

This motivates our pipeline for modeling integrated genomic and clinical datasets (Fig. 1). The first step is based upon a prior knowledge evaluation method we recently developed for graphical structure learning [7]. We measure the concordance between the data and each prior information source and weight the sources accordingly. Then, the information in the data and the prior knowledge are fused to select parameters for a feature selection method (Pref-Div) [18], [19]. Finally, the clusters selected by Pref-Div are modeled as a graph using CausalMGM to represent the dependencies between the clusters and outcome variables of interest.

Our specific contributions are as follows:

- A novel method (piPref-Div) for variable and cluster selection that combines a feature selection approach [19] with an approach to evaluate and integrate prior information [7] (Section 3.2).
- An extensive evaluation of piPref-Div on synthetic datasets (Section 5.1).
- An evaluation of piPref-Div against state of the art variable selection approaches for predictiing breast cancer outcome (Section 5.2).
- An evaluation of our full graphical modeling pipeline for breast cancer subtyping from transcriptomic data (Section 5.3).

2 Related Work

In this section, we survey feature selection methods for genomic data. Then, we discuss methods to incorporate prior knowledge. Finally, we discuss graphical model structure learning approaches for mixed datasets.

2.1 Feature Selection in Genomics

Feature selection methods identify a subset of features in a dataset that collectively predict a target variable. They aim to improve model training efficiency and to prevent overfitting. Feature selection approaches fall into three broad classes: filter methods, wrapper methods, and embedded methods [3]. Filter methods select features using univariate ranking scores such as a Wilcoxon test or a t-test between covariates and a target variable. Wrapper methods use a predictive model like the Support Vector Machine as a basis to select a set of features that result in an accurate prediction model [20]. Two popular wrapper methods are the recursive feature elimination and greedy forward search, which select the best feature to eliminate (or include, respectively) in a step-wise fashion. Embedded methods are predictive models which select features automatically as part of the learning procedure. The most popular example is LASSO regression [21], which uses an L_1 norm penalty to shrink coefficients in a linear regression.

One study investigated the performance of these techniques to predict breast cancer relapse from genomic data [3]. They found that no method had consistently better accuracy than random selection of features. This suggests that tailored approaches are necessary to improve feature selection from omics data.

2.2 Incorporating Prior Knowledge

The use of domain (prior) knowledge may improve these approaches. Three main sources of prior knowledge have been explored: gene ontology (GO) terms, protein-protein interaction (PPI) networks, and biological pathways [1].

GO groups genes based on known biological functions (e.g., cell cycle or angiogenesis). Several approaches have leveraged GO terms as prior information to construct gene clusters [22], [23], [24]. The main drawback of these methods is the incompleteness of GO terms. Genes not found in a functional group in the GO database are discarded. In addition, GO terms tend to define broad functional classes which are difficult to interpret.

PPI networks encode protein interactions known to occur in cells. Methods for gene selection have been built off of these networks (reviewed and evaluated in [2]). These approaches tend to 1) group genes based on the edges in the network and penalize them together [14], [25], [26], [27] or 2) use the network information to determine gene importance [15], [20]. Pathway based approaches are similar, but

Authorized licensed use limited to: University of Pittsburgh Library System. Downloaded on November 11,2025 at 19:44:41 UTC from IEEE Xplore. Restrictions apply.

they use biological pathways (network modules that carry out a specific function). These are typically taken from a pathway database such as KEGG or I2D [28], [29]. Biological pathway-based feature selection (BPFS) is a step-wise method that uses mutual information to the target variable as a scoring criterion. BPFS reduces redundancy by avoiding genes from similar pathways [30]. In [31], the authors attempt to construct a single feature for each pathway by aggregating information across multiple genes. A similar approach is taken in [32] except that the pathways are constructed using the data. Multiple studies have found no significant benefit in prediction accuracy using these methods; however, they do appear to give more biologically interpretable signatures [2], [17].

2.3 Mixed Graphical Models

For data exploration applications, graphical models enable a user to identify all direct associations for any variable of interest. Genomic data is often integrated with clinical, demographic, and epidemiological data. Therefore, we focus upon approaches to learn undirected graphical models from mixed datasets: mixed graphical models (MGM). A MGM parametrizes the joint distribution of a mixed dataset as a graph $G = (V, E)$, where V is the set of variables and E is the set of edges. In this type of model, an edge exists between two variables X and Y if X and Y are conditionally dependent given the rest of the variables in the data.

Recently several methods have been proposed to learn MGMs. Many of these works involve regression-based methods to estimate the conditional dependencies among pairs of variables and infer the edges in the graph. In these approaches: 1) each variable in turn is considered as the target variable 2) a regression is performed using all other variables as predictors, and 3) edges are added to the model for all significant regressors. In [33], the authors use a random forest regression approach to rank edges for inclusion into a graphical model among mixed variables. In [34], they assume that the conditional distributions of each type of variable come from the exponential family and use node-wise regression approaches to estimate the parameters of the model. Other similar techniques have been proposed [35], [36], [37]. Also, in [38], the authors propose qp-graphs which can be estimated from high dimensional data. However, this type of model assumes that there are no edges between categorical variables; a limiting assumption for clinical data.

Another way to estimate a MGM is the pseudolikelihood approach [39]. This approach uses the product of conditional distributions of the variables as a consistent estimator of the true likelihood without computing the partition function. Then a proximal gradient optimization is used to find maximum pseudolikelihood estimates of the parameters. Lee and Hastie propose a MGM that generalizes a popular continuous graphical model (Gaussian Graphical Model) and a popular discrete model (Markov Random Field) [40]. They demonstrate that using the pseudolikelihood approach shows better empirical performance than using separate regressions. So, we focus on this type of approach; and specifically to the improved version presented in [5].

The parameterization of the joint distribution is done according to Equation (1)

$$p(x, y; \theta) \propto exp\Bigg(\sum_{s=1}^{p} \sum_{t=1}^{p} -\frac{1}{2} \beta_{st} x_s x_t + \sum_{s=1}^{p} \alpha_s x_s + \sum_{s=1}^{p} \sum_{j=1}^{q} \rho_{sj}(y_j) x_s + \sum_{j=1}^{q} \sum_{r=1}^{q} \phi_{rj}(y_r, y_j) \Bigg), \tag{1}$$

where θ represents the full set of parameters, x_s represents the sth of p continuous variables and y_j represents the jth of q discrete variables. β_{st} represents the edge potential between continuous variables s and t, α_s represents the continuous node potential for s, ρ_{sj} represents the edge potential between continuous variable s and discrete variable j, and finally ϕ_{rj} represents the edge potential between discrete variables r and j. This model has the favorable property that its conditional distributions are given by Gaussian linear regression and Multiclass Logistic Regression for continuous and discrete variables respectively.

Learning this model over high dimensional datasets directly is computationally infeasible due to the computation of the partition function. To avoid this, a proximal gradient method is used to learn a penalized negative log pseudolikelihood form of the model (Equation (2), product of conditional distributions). To prevent overfitting, non-zero parameters are penalized using the method described in [5] (Equation (3)). Here, λ_{CC} is a penalty parameter only for edges between continuous variables (CC = Continuous-Continuous), λ_{CD} and λ_{DD} are for mixed edges and edges only using discrete variables, respectively. $||.||_F$ refers to the Frobenius norm of a matrix. To optimize this objective function the proximal gradient optimization method is used as specified in [5]

$$\widetilde{l}(\Theta|x, y) = -\sum_{s=1}^{p} \log p(x_s|x_{\setminus s}, y; \Theta) - \sum_{r=1}^{q} \log p(y_r|x, y_{\setminus r}; \Theta) \tag{2}$$

$$\underset{\Theta}{\text{minimize}}\ l_\lambda(\Theta) = \widetilde{l}(\Theta) + \lambda_{CC} \sum_{s=1}^{p} \sum_{t=1}^{s-1} |\beta_{st}| + \lambda_{CD} \sum_{s=1}^{p} \sum_{j=1}^{q} ||\rho_{sj}||_2 + \lambda_{DD} \sum_{j=1}^{q} \sum_{r=1}^{j-1} ||\phi_{rj}||_F. \tag{3}$$

3 Feature Selection Methods

In this section, we describe our computational methods for feature selection. First, we discuss our variable selection approach, and then we discuss how we incorporate prior knowledge to select parameters for this approach automatically.

3.1 Variable Selection: Preferential Diversity

The first step in our procedure is to choose a set of informative variables to model. We want to identify query results

Authorized licensed use limited to: University of Pittsburgh Library System. Downloaded on November 11,2025 at 19:44:41 UTC from IEEE Xplore. Restrictions apply.

relevant to the user but also diverse to give a broad snapshot of the underlying data [19]. The problem was referred to as the Top-K relevant and diverse set problem and is as follows (Definition 1).

Definition 1. Top-K Relevant and Diverse Set. *Given* $0 \leq r \leq 1$ *a radius of similarity, a set of variables* V*, an output size* k*, a similarity function* $Sim(V_i, V_j)$*, and a relevance function* $Rel(V_i)$*.*

$$\begin{aligned} &maximize \sum_{X_i \in S} Rel(X_i) \\ &subject\ to\ S \subset V \\ &\qquad |S| = k \\ &\qquad \forall\, i, j V_i \in S and V_j \in S \rightarrow Sim(V_i, V_j) < r. \end{aligned} \tag{4}$$

Intuitively, we aim to find a set of variables S relevant to the user with the constraint that no pair of chosen variables are similar to one another. This is an appropriate choice because graphical models can lose accuracy if redundant variables are included in the model [12]. We propose a method similar in principle to two filter methods: Correlation-based feature selection [41] and minimum redundancy maximum relevance (mRMR) feature selection [42]. Both of these are greedy approaches. They select the feature that optimizes an objective function that balances relevance and diversity. The main differences in our approach are that we require zero redundancy, and that we quantify redundancy using prior knowledge. To ensure stability of the downstream model, we report the selected features as clusters of redundant variables (instead of discarding them). This allows the user to understand the redundancy in the data.

Another popular approach that follows this principle is the Weighted Gene Correlation Network Analysis (WGCNA) [43]. Briefly, this method aims to learn a weighted undirected correlation network by converting correlation to edge weight. With this network, they infer the dissimilarity between nodes in the network, and use network characteristics (e.g., hub nodes) to select important genes. This method differs in that it infers a correlation network (instead of conditional dependence), and it uses network characteristics instead of summary statistics to infer importance.

Here, we solve this problem using the Preferential Diversity (Pref-Div) algorithm. Pref-Div is an iterative procedure that first selects the top-K most relevant variables and adds them to the result set R. Then, it determines whether any pair of variables in R are redundant (as defined by the radius of similarity, r and the similarity function $Sim(V_i, V_j)$), and removes the lower relevance variable from the result set. The most relevant $K - |R|$ variables that have not been explored are added to the result set. This procedure repeats until a set of K relevant and diverse features are selected. For the full procedure, we refer the reader to [19]. In this work, we make one substantial modification to the original Pref-Div algorithm. We compute all variables considered redundant (within r distance) to each selected variable and return these as clusters.

We instantiate the Pref-Div algorithm with the following parameter choices. The output size k is user-determined since the appropriate choice for this is based on computational resources available. Similarity scores between pairs of features are given by Pearson correlation, and relevance of each feature is given by Pearson correlation to a pre-defined target. We note that having a target variable of interest is not necessary, and unsupervised statistics such as variance or domain knowledge can be used to determine relevance scores. In the next section, we discuss how we select the radius of similarity, λ, using prior knowledge.

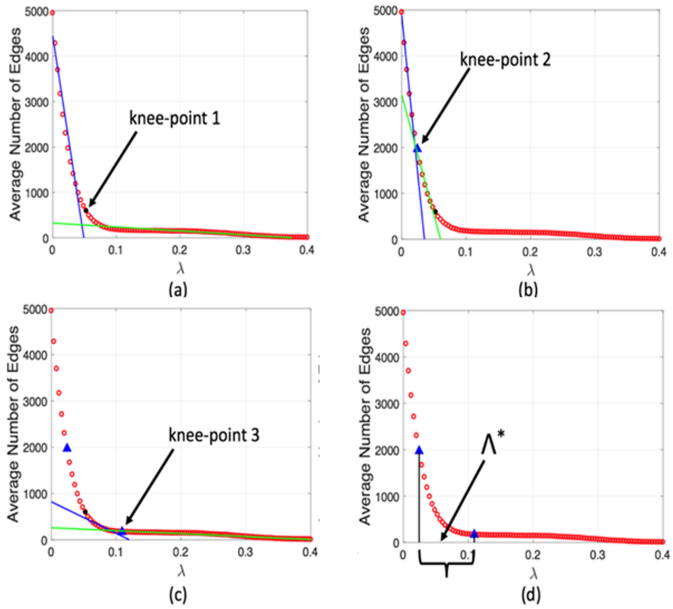


Fig. 2. Illustration of procedure to limit tested parameter range. Figure originally appeared in [7].

3.2 Prior Information Pref-Div: piPref-Div

To choose λ^*, we utilize a method we originally developed to select hyperparameters to learn graphical models with priors [7]. The main idea of the new method (Prior Information Pref-Div or piPref-Div) is to compute a correlation graph across many different correlation thresholds, λ^*. A correlation graph contains an edge between V_1 and V_2 if the correlation between V_1 and V_2 is greater than the threshold. This method proceeds in four main steps.

First, an appropriate parameter range is determined. We identify a range where few edges are selected in the correlation graph yet changing λ slightly results in a large change in the number of edges in the graph. Fig. 2 shows a plot of the number of edges in the correlation graph versus λ. Initially, a knee point is identified that best splits the curve into two straight lines (Panel a). Then, this procedure is repeated on each partition of the curve to compute two additional knee-points (Panels b and c). The final parameter range is the set between these two knee-points (Panel d).

Then, a subsampling approach [5] is used to compute empirical probabilities of appearance for each edge by computing correlation graphs across the chosen range of thresholds and random subsamples without replacement. The empirical probability of each edge is its frequency of appearance.

Next, the information contained in the prior knowledge sources are evaluated against these empirical probabilities across all edges (Equation (5)). Each prior information

Authorized licensed use limited to: University of Pittsburgh Library System. Downloaded on November 11,2025 at 19:44:41 UTC from IEEE Xplore. Restrictions apply.

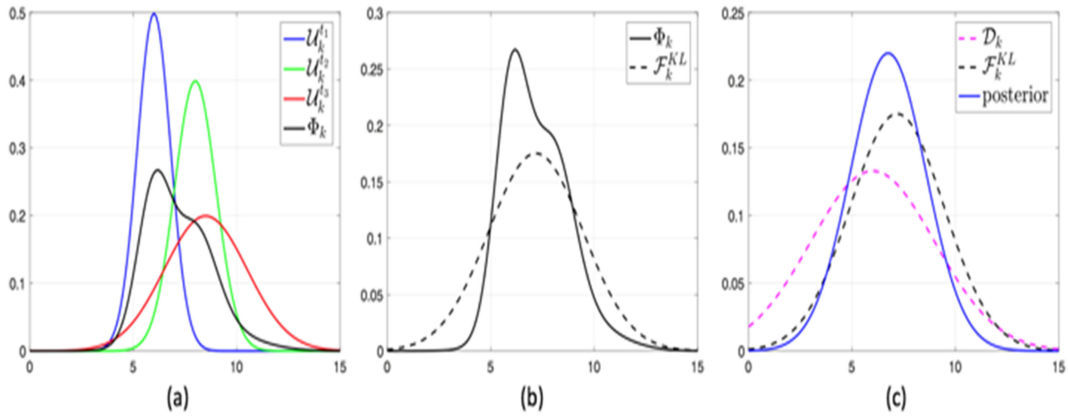


Fig. 3. Subsampling procedure to determine empirical probabilities for every edge in the correlation graph. $B(\lambda, S)$ returns a correlation graph computed upon dataset S with threshold λ. Figure originally appeared in [7].

source (t_r) gives knowledge in the form of a probability of appearance for some fixed set of edges (wp^{t_r}). τ_{t_r} quantifies the "unreliability" of source t_r. $\phi_k^{t_r}$ is the expected number of times edge k should appear during the subsampling procedure according to source t_r, and μ_k is the actual number of appearances for edge k

$$\tau_{t_r} = \frac{\sum_{k=1}^{|wp^{t_r}|} |\phi_k^{t_r} - \mu_k|}{|wp^{t_r}|}. \quad (5)$$

Finally, posterior distributions are computed for each edge (Fig. 3). For each edge k, a normal distribution is used to approximate the probability of appearance for each prior source (red, green, and blue curves in Panel a). Using a normalized reciprocal of the scores computed in the previous step, these normal distributions are combined into a weighted mixture (black curve, Panel a). This mixture distribution is approximated by the normal distribution which has minimal KL-divergence to the mixture (Panel b). Finally, this normal distribution is combined with a normal distribution from the empirical probabilities to get a posterior distribution (Panel c, blue curve).

Since some edges may not have prior information from any of the sources, separate λ^*s are calculated for edges with and without prior information (λ^*_{wp} and λ^*_{np}, respectively). λ^*_{wp} is chosen based upon stability of the correlation graph across subsamples along with concordance to the posterior distribution for each feature. λ^*_{np} is chosen the same way, except that the posterior distribution is the one computed from the data alone (pink curve, Panel c).

4 Experimental Setup

Next, we describe the synthetic and real data used to evaluate our approach, and the metrics we apply in our evaluation. Lastly, we describe the prior knowledge sources used.

4.1 Simulated Datasets

Simulated datasets were used to evaluate algorithmic correctness and to understand the impact of prior information sources. Data was generated from a linear Gaussian graphical model. Edge coefficients were drawn uniformly at random from the set $[-1.5, -0.5] \cup [0.5, 1.5]$. Error terms for each variable were zero mean with variance randomly drawn from the set $[0.01, 2]$. Graphical structure was simulated using a "clustered simulation" (Fig. 4). Here, each variable belonged to one of C clusters. In these clusters, each pair of variables in the cluster was connected by an edge. $c < C$ clusters had one randomly chosen variable (light blue nodes) connected to the target variable (*relevant clusters*). The remaining $C - c$ clusters were disconnected from the rest of the network. Each cluster consisted of an equal number of variables. To represent a master regulator and force correlated structure, each cluster had a single latent variable (pink nodes) that influenced the value of all variables in the cluster.

Prior knowledge was simulated for reliable and unreliable prior sources. All prior sources give information based on a beta distribution; however, the parameters of this distribution differ based on the type of prior and whether the variables in question belong to the same cluster. An unreliable prior gives information drawn from $Beta(4, 4)$ for both true and false edges (cluster memberships), whereas a reliable prior draws from $Beta(10, 2)$ for true edges, and $Beta(2, 10)$ for false edges. The percent of edges with prior information varies based on the experiment. To determine whether prior information is available for each edge, each edge gets a value $b \backsim U(0, 1)$, and each prior information source has a value $c \in [0, 1]$. The prior gives information about the relationship if $b < c$. In this way, the simulated data reflects the fact that some relationships are more well-studied than others.

We evaluate piPref-Div on its ability to incorporate unreliable prior information in order to select relevant clusters

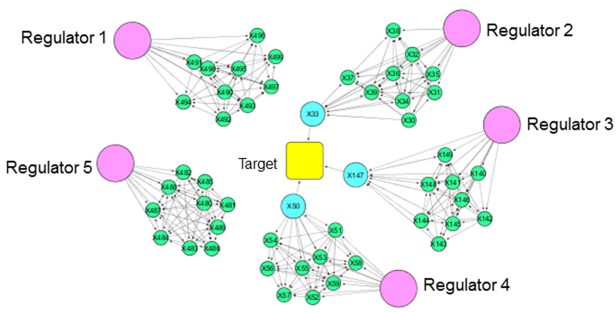


Fig. 4. Cluster Simulation to generate simulated datasets. Purple nodes are master regulators of a cluster, blue nodes are causal parents of the target variable, and the beige node is the target variable.

Authorized licensed use limited to: University of Pittsburgh Library System. Downloaded on November 11,2025 at 19:44:41 UTC from IEEE Xplore. Restrictions apply.

TABLE 1
Gene Expression Datasets and Outcomes

Dataset Name	Outcome	Prevalence
Schmidt (2008)	Metastasis-Free Survival	16.3%
Pawitan (2005)	Relapse-Free Survival	28.6%
Wang (2005)	Relapse-Free Survival	48.2%
Sotiriou (2006)	Relapse-Free Survival	28.6%
Ivshina (2006)	Metastasis-Free Survival	38.3%
Desmedt (2007)	Relapse-Free Survival	39.4%

more accurately. The metric we use for evaluation of selected clusters on simulated data is called *cluster accuracy*. The goal of this metric is to compare the relevant clusters output by piPref-Div to the truly relevant clusters in the data generating graph, where a relevant cluster is a cluster with at least one variable that is a parent of the target variable. First, an optimal matching is computed between the predicted and actual clusters using the Hungarian Algorithm. The cost of assigning a predicted cluster to an actual cluster is given by 1 - the Jaccard similarity between the clusters. If multiple predicted clusters are best assigned to the same actual cluster, these clusters are combined. Finally, the average Jaccard similarity between the combined predicted clusters and their matched actual clusters are computed as the score.

4.2 Gene Expression Datasets

To evaluate the performance of piPref-Div on real data, we apply it to six publicly available breast cancer Affymetrix HGU133A gene expression datasets [44], [45], [46], [47], [48], [49]. These datasets have been used in several previous analyses and represent a baseline to evaluate prediction methods [2], [3], [17]. Each dataset consists of microarray expression data for between 159 and 286 patients. For four datasets, the target variable of interest was whether or not the patient had relapse free survival (RFS) for 5 years. For two sets, this information was unavailable and metastasis free survival (MFS) was used instead [48], [49]. The outcomes and prevalence for each dataset are given in Table 1. Since our aim was to evaluate feature selection, and since the class imbalance was not large, we did not perform any under- or over-sampling techniques.

Our evaluation consists of a five-fold cross validation within each dataset, where the entire model building process (feature selection, prediction of target variable) is repeated for each fold. We measure model generalizability (discrimination) and stability. To measure discrimination, we use area under the ROC curve (AUC) comparing the probability predictions from each method and true binary outcome of RFS and MFS for five years. To measure stability, we use the average Tanimoto set similarity (intersection divided by union) for the set of features selected in each fold. There is no overlap between training and testing datasets in each fold, and so the results should be reliable measures of model generalizability.

To evaluate the potential of our full pipeline to discover knowledge from data, a graphical model was learned from the TCGA-BRCA RNA-Seq expression dataset using the MGM algorithm. This data included gene expression measurements from 784 breast tumor samples and 13,994 genes. Breast cancer diagnosis and prognosis are commonly divided into five main subtypes: Luminal A, Luminal B, HER2+, Triple-Negative, and Basal. Breast cancer sub-type information for each tumor sample was obtained from [50], which did not distinguish between Triple-Negative and Basal. The main driving distinction for these subtypes is the presence or absence of hormone receptors on the tumor cell surface, which can lead to varying prognoses. In these experiments, we aim to identify clusters distinguishing the four sub-types from expression data. To determine stability of each of these clusters, a 10-fold cross validation was performed, and the stability of each cluster was the number of times a similar cluster (Tanimoto similarity $>$ 0.85) was selected in each fold.

4.3 External Prior Knowledge Sources

Prior knowledge consisted of five distinct sources of information. Physical gene distance represents the proportion of chromosome distance covered by the space between these two genes. It is defined as the base pair distance between two genes on the chromosome. If two genes were on separate chromosomes, then this value was set to zero. Otherwise given gene G_i from base pairs B_i^1 to B_i^2 and gene G_j from base pairs B_j^1 to B_j^2, and full chromosome length C, the physical distance prior is given by Equation (6)

$$Phys\left(G_i, G_j\right) = 1 - \frac{max\left(B_i^2, B_j^2\right) - min\left(B_i^1, B_j^1\right)}{C}. \tag{6}$$

Gene family information was curated from the Human Genome Organization (HUGO). Gene families are groups of genes related by sequence and/or function. A single gene can belong to multiple gene families. Thus, we represent each gene as a vector of families with one-hot encoding. To compute the similarity between these vectors, we use the Jaccard similarity metric which is the number of families in common divided by the total number of unique families either gene belongs to. A similar approach is used for gene-disease mapping from the DisGeNet [51]. This database gives scores quantifying the level of knowledge that a change in a gene is related to a disease. We use the guilt by association principle to compute whether two genes are related. We represent a gene by a vector of scores to the diseases in the database, and we compute the cosine similarity between two gene vectors. Since all scores are positive, this metric is positive, and is used directly as a probability.

Finally, we use gene-gene similarity data from two sources: Harmonizome [52] and STRING [53]. Harmonizome similarity data was curated from the Molecular Signatures Database [54] and consisted of correlation between gene expression across several microarray and RNA-Seq experiments. STRING curates gene-gene relationship scores based on several factors such as: co-expression, literature co-occurrence, experimental evidence, other databases, etc. STRING scores were scaled from their $(0, 1000)$ range to $(0, 1)$. We analyzed the quantity (Table 2 and similarities (Figs. 5 and 6) between these prior sources. Though each individual source provides relatively little information, the overlap between gene-gene pairs from each source is also small, which shows that are

Authorized licensed use limited to: University of Pittsburgh Library System. Downloaded on November 11,2025 at 19:44:41 UTC from IEEE Xplore. Restrictions apply.

TABLE 2
Characteristics of Prior Knowledge Sources for Classification Experiment

Prior Source	Prior Percentage
MSig DB Co-Expression	1.41%
STRING Co-Expression	3.38%
DisGeNet Gene-Disease Mapping	6.79%
Physical Gene Distance	5.94%
HUGO Gene Families	0.39%

highly complementary (Fig. 5). This also gives us a decent coverage of all potential gene-gene interactions. However, we also found that when the these sources give information about the same gene-gene interaction, their scores are not highly correlated (Fig. 6). The most similar sources are the HUGO Gene Families and the DisGeNet Gene-Disease mapping with a correlation of 0.35.

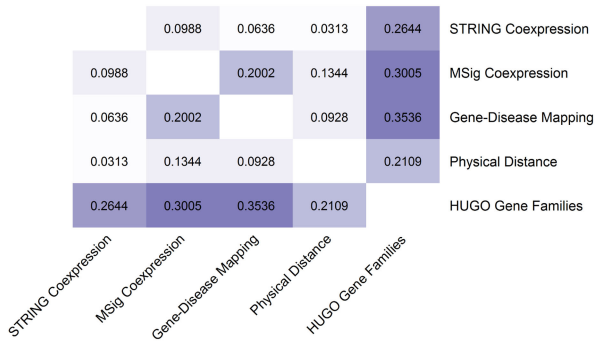


Fig. 6. Heatmap of overlapping prior knowledge between sources. Each cell is the correlation between the probabilities given by each source for all gene-gene pairs in the prior source of the row that are also in the prior source in the column.

5 Results

We demonstrate the performance of piPref-Div on simulated and real datasets. First, we evaluate its ability to determine reliable prior information sources and incorporate those sources to select better clusters. Then, we evaluate the method in terms of its ability to accurately predict outcome for breast cancer patients, and lastly, we use our full pipeline to learn a graphical model of breast cancer subtype discrimination.

5.1 Evaluation and Impact of Prior Knowledge

First, we tested the ability of piPref-Div to accurately evaluate prior knowledge sources on various simulated datasets. In total we had 15 datasets of 500 variables with 50 clusters, 25 relevant to the target, 5 prior knowledge sources (3 reliable), with a random amount of prior information. We repeated this process for datasets with both 50 samples and 200 samples.

The results are presented in Fig. 7. Here, "Net Reliability" (y-axis) refers to the sum of the probabilities given to true edges minus the sum of the probabilities given to false edges for each prior. The predicted weight for each prior knowledge source, given by piPref-Div, shows a clear association to the reliability score. A benefit of this approach is that this weight does not appear to be dependent on the amount of prior information. Even with little prior information (blue circles), piPref-Div assigns an accurate weight to the knowledge sources.

The next experiment investigated the impact of the amount and quality of prior knowledge on the ability of piPref-Div to identify relevant clusters of variables. In these experiments, we test the method using the same experimental parameters as the previous section, except we use a larger dataset with 3,000 variables, 300 clusters (75 relevant). For each experimental setting, 15 graphs were generated and the results are presented cumulatively over these graphs.

The results for the small datasets are given in Fig. 8. Sample size is the most significant factor in determining accuracy of the selected clusters. Prior information gives a modest improvement in accuracy, but this benefit only occurs with at least 50 percent of prior information and at least 3 reliable sources out of 5. However, when all sources are unreliable, there is no decrease in accuracy unless there is a large amount of information present. Lastly, we note that the benefit of prior information is drastically reduced in large sample data (200 sample case). This is intuitive, as with more data, correlation becomes a very stable measure,

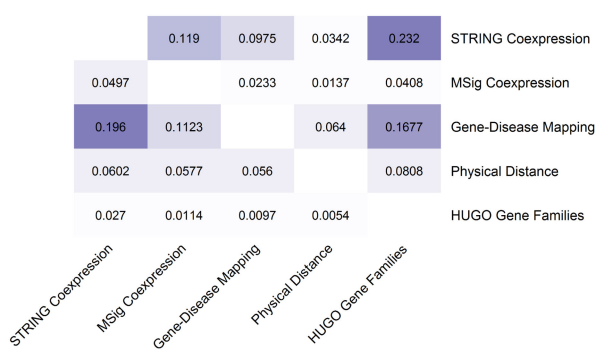


Fig. 5. Heatmap of correlation of prior knowledge between sources. Each cell is the percentage of gene-gene pairs in the prior source of the row that are also in the prior source in the column.

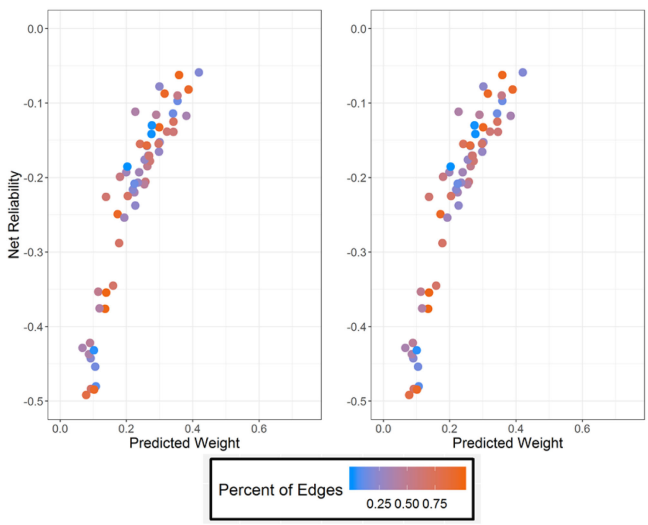


Fig. 7. Predicted Weight versus Net Reliability for each prior knowledge source in simulated experiments for piPref-Div for (left) 50 samples and (right) 200 samples.

Authorized licensed use limited to: University of Pittsburgh Library System. Downloaded on November 11,2025 at 19:44:41 UTC from IEEE Xplore. Restrictions apply.

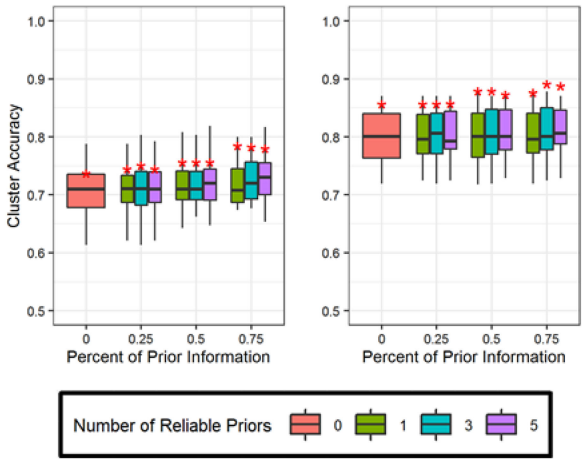


Fig. 8. Accuracy of predicted clusters for varying amount and reliability of prior knowledge. Sample size was set to 50 (left) and 200 (right). Star represents the best possible performance if optimal correlation thresholds were selected.

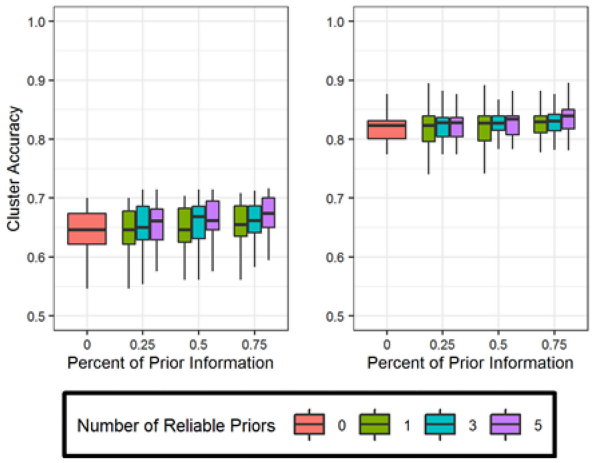


Fig. 9. Accuracy of predicted clusters for varying amount and reliability of prior knowledge on large datasets. Sample size was set to 50 (left) and 200 (right).

and prior information can be ignored. The stars represent the performance if the optimal correlation thresholds were selected. These results indicate that regardless of sample size, piPref-Div selected reasonable thresholds, close to optimal. However, the distance between piPref-Div performance and the optimal thresholds is the largest when there is a large amount of prior information. This may suggest that piPref-Div does not value the prior information enough even when it is reliable and plentiful.

Lastly, we examined the ability of piPref-Div to detect clusters from a larger graph (Fig. 9). Here, the pattern is similar, except the impact of prior knowledge is more significant. In particular, having just 25 percent of edges with prior information gives a substantial increase in accuracy over having no prior information at all. Again, this impact is larger when the sample size is small. Increasing sample size has a larger effect on this data. An increase from 50 to 200 samples results in an increase in accuracy from 0.65 to over 0.8 for all amounts of prior.

5.2 Breast Cancer Outcome Prediction

To determine the performance of piPref-Div on real datasets, we applied the algorithm to the breast cancer datasets described above. Three variations of piPref-Div were tested. piPref-Div alone (PD), piPref-Div with and without prior information (No Prior = NP) with clusters aggregated into summarized features using principal component analysis (PD-PCA, PDNP-PCA). For the Pref-Div approaches, an inner 3-fold cross-validation loop was used to determine the number of selected features (1,3,5, and 10 features were tested). Genes with less than 0.5 standard deviation across samples in the training set were removed from the dataset prior to feature selection. Two methods that performed well in a previous study were included in the analysis: Hybrid-Huberized SVM (HH-SVM) and Recursive-Reweighted Feature Elimination (RRFE) [2].

The discrimination results are presented in Fig. 10. Across the datasets, the consistent best performing methods are PD-NP-PCA and PD-PCA (sea-green and light blue, respectively). This suggests that cluster selection and representing individual features as clusters offers a benefit to selecting single genes alone. However, this is dataset dependent, as Pref-Div alone (PD, yellow box) matches these methods on 2 of the datasets (Sotiriou and Desmedt) and performs better on 1 dataset (Wang). Overall, these results show no significant difference between using prior information (PD, PD-PCA) and not using prior information (PD-NPPCA). Using prior information with PCA clustering shows a slight improvement on the Ivshina dataset, but none of the others. We find that our approach performs about the same in terms of AUC when compared to SVM-RRFE, but our method tends to select significantly fewer features.

Fig. 11 presents the stability of the learned models. The results confirm previous work that identifying a stable model for breast cancer outcome prediction is a difficult problem [2]. In general, only the RRFE algorithm shows somewhat consistent stability; however, we note that a major contributing factor is that this algorithm uses on average 119 selected features, whereas HH-SVM averages around 6 and the PD approaches average around 1 feature (or cluster). Among, the Pref-Div based approaches, PD-PCA with and without prior information show the most consistent stability. On nearly all datasets they are on par with RRFE despite choosing significantly fewer features.

To better understand how prior information impacts piPref-Div, we show the detailed differences between piPref-Div with (WP) and without (NP) prior information in the breast cancer outcome experiments (Table 3). This shows that regardless of prior information, piPref-Div has similar discrimination on all six datasets. In addition, the correlation between the predicted probabilities (Prediction Similarity) and the average Jaccard similarity between the selected clusters (Cluster Similarity) are high for piPref-Div with and without prior information on all datasets. This implies that not only do the methods select models with similar accuracy, but the selected clusters themselves are highly similar.

5.3 Stratification of Breast Cancer Subtypes

Finally, we evaluate our full pipeline (variable selection then graphical modeling) on its ability to mine interesting clusters

Authorized licensed use limited to: University of Pittsburgh Library System. Downloaded on November 11,2025 at 19:44:41 UTC from IEEE Xplore. Restrictions apply.

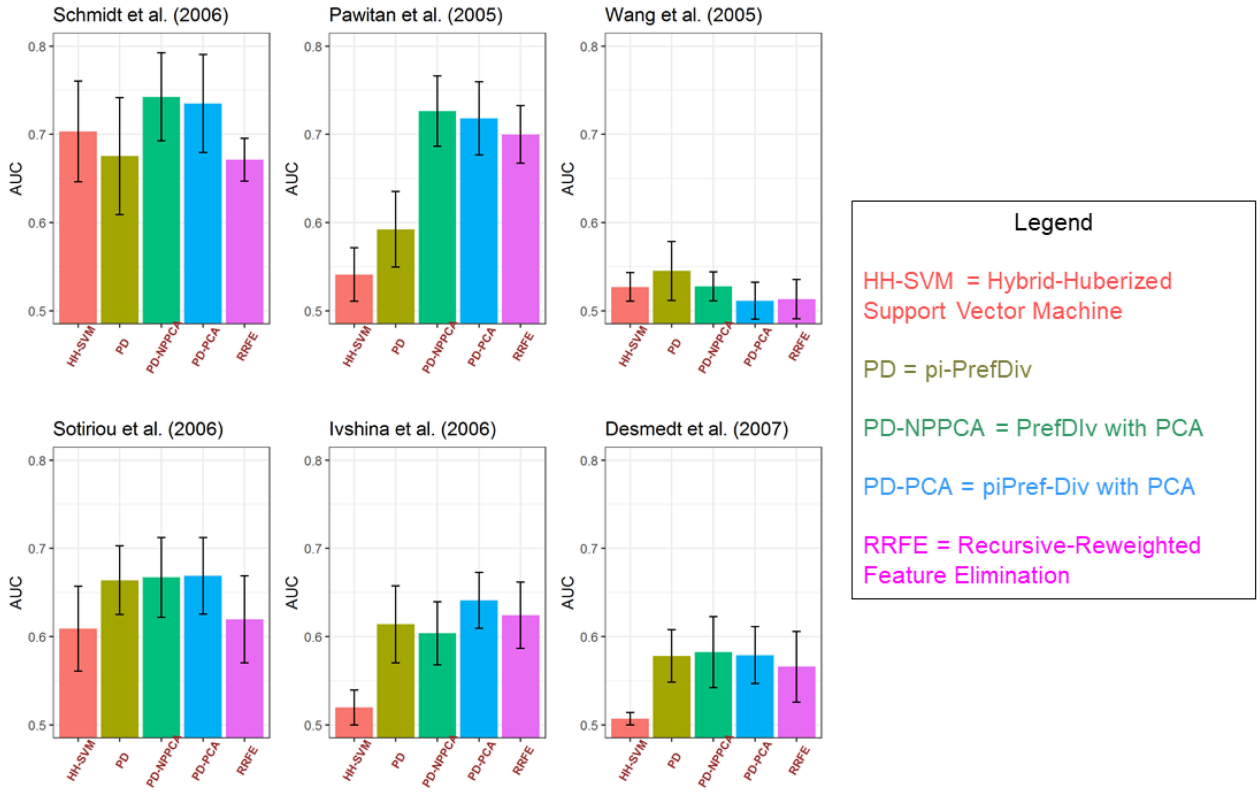


Fig. 10. AUC of predicting five-year metastasis-free and relapse-free survival using several feature selection methods on six independent breast cancer microarray datasets.

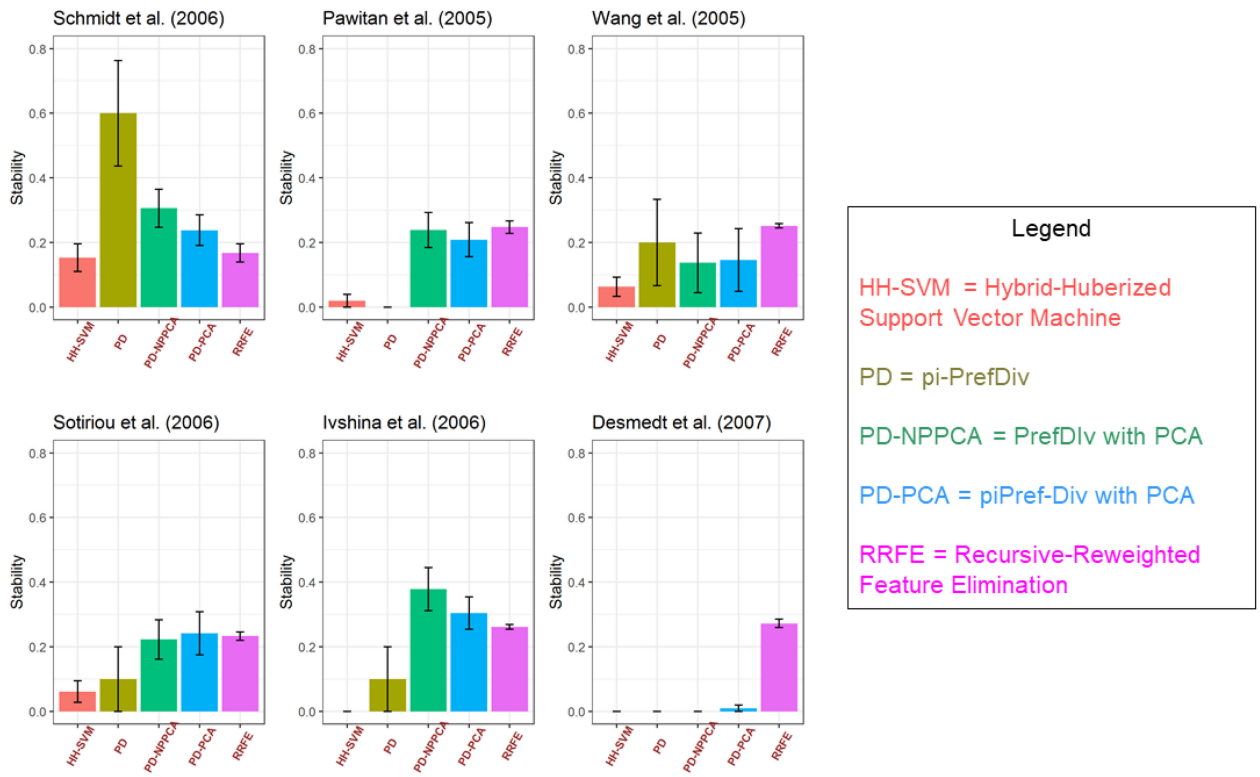


Fig. 11. Stability of learned models for predicting five-year metastasis-free and relapse-free survival using several feature selection methods on six independent breast cancer microarray datasets.

related to breast cancer subtype. Based on the previous section, we chose to use PD-PCA for variable selection due to its consistently high AUC and relatively high stability. An MGM (undirected graph) model was learned on a dataset consisting of only the selected clusters and the Subtype variable. To summarize clusters of genes into single names, the Ingenuity

Authorized licensed use limited to: University of Pittsburgh Library System. Downloaded on November 11,2025 at 19:44:41 UTC from IEEE Xplore. Restrictions apply.

TABLE 3
Comparison of piPref-Div With (WP) and Without (NP) Prior Information in Breast Cancer Outcome Experiments

Dataset	AUC (NP)	AUC (WP)	Cluster Size (NP)	Cluster Size (WP)	Cluster Similarity	Prediction Similarity
Sotiriou (2006)	0.667	0.669	76.3	76.9	0.78	0.94
Ivshina (2006)	0.604	0.641	39.3	40.0	0.57	0.67
Desmedt (2007)	0.582	0.579	14.3	15.6	0.71	0.71
Schmidt (2008)	0.742	0.735	42.0	35.0	0.86	0.73
Pawitan (2005)	0.726	0.718	31.4	65.3	0.79	0.79
Wang (2005)	0.528	0.511	42.1	37.8	0.93	0.87

Pathway Analysis regulator analysis was used, and the KEGG Pathway database was queried (corrected p-values < 0.05 were chosen as candidates). Following this step, only specific pathways and regulators were included as names of the clusters.

The learned graphical model is presented in Fig. 12. We found two clusters unable to be mapped coherently to any biological function (single gene representatives were TMEM41A, and TSPAN15); however, these clusters were relatively unstable. The two most stable clusters were Fanconi Anemia/Hereditary Breast Cancer pathway, and a set of genes regulated by MYCN. Fanconi Anemia and the Hereditary Breast Cancer pathways share common genes [55] and developing breast cancer through a genetic basis tends to be associated with ER+ breast cancer [56]. MYC family pathways and the transcription factors themselves are known to be differentially expressed across subtypes, and the MYCN factor in particular has shown differences between triple-negative and other subtypes [57]. FOXA1 along with GATA3 and ESR1 are necessary for maintaining a luminal phenotype of breast cancer [58]. AGR2 is upregulated by FOXA1 but only in an estrogen receptor dependent manner [59]. This implies that the FOXA1-AGR2 loop will only be upregulated in ER+ breast cancer. Though it is unclear how KRT14 regulated genes distinguish subtypes of breast cancer, it is known that upregulation of KRT14 reduces the ability of breast tumors to metastasize and invade the extracellular matrix [60]. Overall, we find that the pipeline constructs and selects reasonable clusters that are discriminative of breast cancer subtypes. The pipeline also generates novel candidate clusters for experimentation.

6 Conclusion

We presented a pipeline to learn graphical model structures from large omics datasets. The pipeline builds upon previous work by developing (1) a method to integrate and evaluate prior information to select hyperparameters, and (2) a variable selection method to identify relevant ye non-redundant sets of features. We used this approach to return clusters of variables instead of individual features, and to model these features as a graphical model to find interesting relationships.

We evaluated our work on synthetic data and real breast cancer data. On synthetic data, we found that our method accurately evaluates the reliability of prior information and utilizes this information to improve the selection of relevant clusters. In addition, even when most prior information is unreliable, the method's performance was no worse than having no priors, which agrees with previous observations [7]. We found that the largest improvement with prior information occured when there are few samples and a large number of features. Overall, we found prior information to modestly improve performance, but this may be necessary to avoid poor performance with unreliable priors.

On classification experiments with microarray data, we found that piPref-Div performs at par or better than other state of the art approaches. Using PCA to summarize clusters is superior to selecting single variables alone. piPref-Div selects far fewer features to achieve similar or better discrimination than state of the art approaches. In this context, prior knowledge did not appear to change the selected features. The simulated experiments show that a greater percentage of prior knowledge is necessary to have a significant impact on the selected features, and these results are reproduced on the breast cancer data. In addition, prior information for these experiments was derived from normal individuals, and it is unclear whether this information is reliable due to the significant genomic dysregulation from cancer. When using the full pipeline with graphical modeling to discriminate breast cancer subtypes, we were able to identify biologically reasonable clusters. Two of the seven clusters did not map to any known biological regulator or pathway and constitute candidates for further investigation.

For future work, we aim to improve upon the accuracy results with reliable prior information. It could be that using the prior information solely to select hyperparameters is too conservative and using the posterior distributions directly can give better results with sparse priors. Since the priors are already being appropriately weighted, the posterior distributions should be accurate. In addition, we aim to evaluate the prediction accuracy of our pipeline on more recent genomic data. The prior information sources used for the biological experiments were relatively sparse. It is future work to utilize the vast array of gene expression experiments available to construct priors that provide a full representation of the genome. Though the pipeline can be

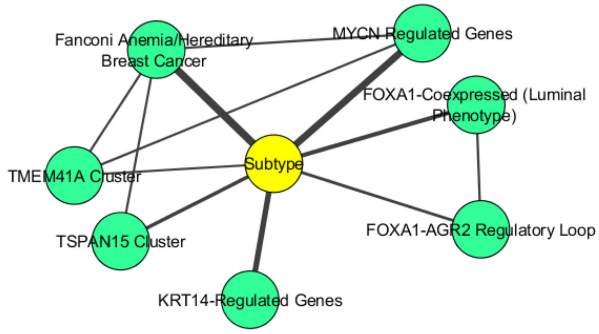


Fig. 12. Graphical model of breast cancer subtype. Size of each edge represents the number of times a similar cluster was selected to be related to Subtype in each of the cross-validation folds.

Authorized licensed use limited to: University of Pittsburgh Library System. Downloaded on November 11,2025 at 19:44:41 UTC from IEEE Xplore. Restrictions apply.

applied to integrated genomic and clinical datasets, in this work we focused on genomic data for the evaluation. We will explore integrated datasets with clinical variables in future work.

Acknowledgments

This work was supported by NIH Grants U01HL137159, R01LM012087 (to PVB), and T32CA082084 (to VKR). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

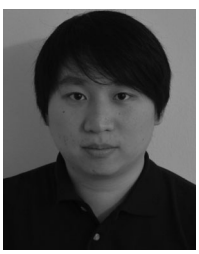
References

[1] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Advances Bioinf.*, vol. 2015, 2015, Art. no. 198363.
[2] Y. Cun and H. Fröhlich, "Prognostic gene signatures for patient stratification in breast cancer-accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions," *BMC Bioinf.*, vol. 13, no. 1, 2012, Art. no. 69.
[3] A.-C. Haury, P. Gestraud, and J.-P. Vert, "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures," *PloS One*, vol. 6, no. 12, 2011, Art. no. e28210.
[4] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.
[5] A. J. Sedgewick, I. Shi, R. M. Donovan, and P. V. Benos, "Learning mixed graphical models with separate sparsity parameters and stability-based model selection," *BMC Bioinf.*, vol. 17, no. S5, 2016, Art. no. 175.
[6] V. K. Raghu *et al.*, "Comparison of strategies for scalable causal discovery of latent variable models from mixed data," *Int. J. Data Sci. Analytics*, vol. 6, pp. 33–45, 2018.
[7] D. V. Manatakis, V. K. Raghu, and P. V. Benos, "piMGM: Incorporating multi-source priors in mixed graphical models for learning disease networks," *Bioinformatics*, vol. 34, no. 17, pp. i848–i856, 2018.
[8] A. J. Sedgewick *et al.*, "Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis," *Bioinformatics*, vol. 35, no. 7, pp. 1204–1212, 2019.
[9] V. K. Raghu *et al.*, "Feasibility of lung cancer prediction from low-dose CT and smoking factors using causal models," *Thorax*, vol. 74, no. 7, pp. 643–649, 2019.
[10] I. Abecassis *et al.*, "PARP1 rs1805407 increases sensitivity to PARP1 inhibitors in cancer cells suggesting an improved therapeutic strategy," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, 2019.
[11] G. D. Kitsios *et al.*, "Respiratory microbiome profiling for etiologic diagnosis of pneumonia in mechanically ventilated patients," *Front. Microbiol.*, vol. 9, 2018, Art. no. 1413.
[12] J. Lemeire, S. Meganck, F. Cartella, and T. Liu, "Conservative independence-based causal structure learning in absence of adjacency faithfulness," *Int. J. Approx. Reasoning*, vol. 53, no. 9, pp. 1305–1325, 2012.
[13] G. T. Huang, I. Tsamardinos, V. Raghu, N. Kaminski, and P. V. Benos, "T-RECS: Stable selection of dynamically formed groups of features with application to prediction of clinical outcomes," in *Proc. Pacific Symp. Biocomput.*, 2015, pp. 431–42.
[14] A. Allahyar and J. De Ridder , "FERAL: Network-based classifier with application to breast cancer outcome prediction," *Bioinformatics*, vol. 31, no. 12, pp. i311–i319, 2015.
[15] I. W. Taylor *et al.*, "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nat. Biotechnol.*, vol. 27, no. 2, 2009, Art. no. 199.
[16] D. Venet, J. E. Dumont, and V. Detours, "Most random gene expression signatures are significantly associated with breast cancer outcome," *PLoS Comput. Biol.*, vol. 7, no. 10, 2011, Art. no. e1002240.
[17] C. Staiger, S. Cadot, B. Györffy, L. F. Wessels, and G. W. Klau, "Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis," *Front. Genetics*, vol. 4, 2013, Art. no. 289.
[18] V. K. Raghu, X. Ge, P. K. Chrysanthis, and P. V. Benos, "Integrated theory-and data-driven feature selection in gene expression data analysis," in *Proc. IEEE 33rd Int. Conf. Data Eng.*, 2017, pp. 1525–1532.
[19] X. Ge, P. K. Chrysanthis, and A. Labrinidis, "Preferential diversity," in *Proc. 2nd Int. Workshop Explor. Search Databases Web*, 2015, pp. 9–14.
[20] M. Johannes *et al.*, "Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients," *Bioinformatics*, vol. 26, no. 17, pp. 2136–2144, 2010.
[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.
[22] W. Pan, "Incorporating gene functions as priors in model-based clustering of microarray gene expression data," *Bioinformatics*, vol. 22, no. 7, pp. 795–801, 2006.
[23] J. Cheng *et al.*, "A knowledge-based clustering algorithm driven by gene ontology," *J. Biopharm. Statist.*, vol. 14, no. 3, pp. 687–700, 2004.
[24] X. Chen and L. Wang, "Integrating biological knowledge with gene expression profiles for survival prediction of cancer," *J. Comput. Biol.*, vol. 16, no. 2, pp. 265–278, 2009.
[25] L. Wang, J. Zhu, and H. Zou, "Hybrid huberized support vector machines for microarray classification and gene selection," *Bioinformatics*, vol. 24, no. 3, pp. 412–419, 2008.
[26] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *J. Comput. Graphical Statist.*, vol. 22, no. 2, pp. 231–245, 2013.
[27] Y. Zhu, X. Shen, and W. Pan, "Network-based support vector machine for classification of microarray samples," *BMC Bioinf.*, vol. 10, no. 1, 2009, Art. no. S21.
[28] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, 2000.
[29] K. R. Brown and I. Jurisica, "Online predicted human interaction database," *Bioinformatics*, vol. 21, no. 9, pp. 2076–2082, 2005.
[30] N. Bandyopadhyay, T. Kahveci, S. Goodison, Y. Sun, and S. Ranka, "Pathway-based feature selection algorithm for cancer microarray data," *Advances Bioinf.*, vol. 2009, 2010, Art. no. 532989.
[31] Z. Guo *et al.*, "Towards precise classification of cancers based on robust gene functional expression profiles," *BMC Bioinf.*, vol. 6, no. 1, 2005, Art. no. 58.
[32] N. Alcaraz, M. List, R. Batra, F. Vandin, H. J. Ditzel, and J. Baumbach, "De novo pathway-based biomarker identification," *Nucleic Acids Res.*, vol. 45, no. 16, pp. e151–e151, 2017.
[33] B. Fellinghauer *et al.*, "Stable graphical model estimation with random forests for discrete, continuous, and mixed variables," *Comput. Statist. Data Anal.*, vol. 64, pp. 132–152, 2013.
[34] E. Yang, Y. Baker, P. Ravikumar, G. Allen, and Z. Liu, "Mixed graphical models via exponential families," in *Proc. 17th Int. Conf. Artif. Intell. Statist.*, 2014, pp. 1042–1050.
[35] J. Cheng, T. Li, E. Levina, and J. Zhu, "High-dimensional mixed graphical models," *J. Comput. Graphical Statist.*, vol. 26, pp. 367–378, 2017.
[36] S. Chen, D. M. Witten, and A. Shojaie, "Selection and estimation for mixed graphical models," *Biometrika*, vol. 102, no. 1, pp. 47–64, 2014.
[37] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.
[38] I. Tur and R. Castelo, "Learning mixed graphical models from data with p larger than n," in *Proc. 27th Conf. Uncertainty Artif. Intell.*, 2011, pp. 689–697.
[39] J. Besag, "Statistical analysis of non-lattice data," *J. Roy. Statist. Soc. Ser. D*, vol. 24, pp. 179–195, 1975.
[40] J. D. Lee and T. J. Hastie, "Learning the structure of mixed graphical models," *J. Comput. Graphical Statist.*, vol. 24, no. 1, pp. 230–253, 2015.
[41] M. A. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 359–366.
[42] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinf. Comput. Biol.*, vol. 3, no. 02, pp. 185–205, 2005.
[43] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statist. Appl. Genetics Mol. Biol.*, vol. 4, 2005, Art. no. 17.
[44] C. Desmedt *et al.*, "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series," *Clin. Cancer Res.*, vol. 13, no. 11, pp. 3207–3214, 2007.
[45] C. Sotiriou *et al.*, "Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis," *J. Nat. Cancer Inst.*, vol. 98, no. 4, pp. 262–272, 2006.
[46] Y. Wang *et al.*, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

Authorized licensed use limited to: University of Pittsburgh Library System. Downloaded on November 11,2025 at 19:44:41 UTC from IEEE Xplore. Restrictions apply.

[47] Y. Pawitan *et al.*, "Gene expression profiling spares early breast cancer patients from adjuvant therapy: Derived and validated in two population-based cohorts," *Breast Cancer Res.*, vol. 7, no. 6, 2005, Art. no. R953.
[48] A. V. Ivshina *et al.*, "Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer," *Cancer Res.*, vol. 66, no. 21, pp. 10 292–10 301, 2006.
[49] M. Schmidt *et al.*, "The humoral immune system has a key prognostic impact in node-negative breast cancer," *Cancer Res.*, vol. 68, no. 13, pp. 5405–5413, 2008.
[50] G. Jiang *et al.* "Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer," *BMC Genomics*, vol. 17, no. 7, 2016, Art. no. 525.
[51] J. Pi ñero *et al.*, "DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Res.*, vol. 45, pp. D833–D839, 2017.
[52] A. D. Rouillard *et al.*, "The harmonizome: A collection of processed datasets gathered to serve and mine knowledge about genes and proteins," *Database*, vol. 2016, 2016, Art. no. baw100.
[53] D. Szklarczyk *et al.*, "STRING v10: Protein–protein interaction networks, integrated over the tree of life," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D447–D452, 2014.
[54] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Nat. Acad. Sci. United States America*, vol. 102, no. 43, pp. 15 545–15 550, 2005.
[55] D. Alan and M. D'Andrea, "The Fanconi anemia and breast cancer susceptibility pathways," *New England J. Med.*, vol. 362, no. 20, 2010, Art. no. 1909.
[56] A. M. Mulligan *et al.*, "Common breast cancer susceptibility alleles are associated with tumour subtypes in BRCA1 and BRCA2 mutation carriers: Results from the consortium of investigators of modifiers of BRCA1/2," *Breast Cancer Res.*, vol. 13, no. 6, 2011, Art. no. R110.
[57] D. Horiuchi *et al.*, "MYC pathway activation in triple-negative breast cancer is synthetic lethal with CDK inhibition," *J. Exp. Med.*, vol. 209, no. 4, pp. 679–696, 2012.
[58] S. Chaudhary, B. M. Krishna, and S. K. Mishra, "A novel FOXA1/ESR1 interacting pathway: A study of Oncomine™ breast cancer microarrays," *Oncology Lett.*, vol. 14, no. 2, pp. 1247–1264, 2017.
[59] T. M. Wright *et al.*, "Delineation of a FOXA1/ERα/AGR2 regulatory loop that is dysregulated in endocrine therapy–resistant breast cancer," *Mol. Cancer Res.*, vol. 12, no. 12, pp. 1829–1839, 2014.
[60] J. M. Westcott *et al.*, "An epigenetically distinct breast cancer cell subpopulation promotes collective invasion," *J. Clin. Invest.*, vol. 125, no. 5, pp. 1927–1943, 2015.

Vineet K. Raghu received the BS and PhD degrees in computer science from the University of Pittsburgh, Pittsburgh, Pennsylvania, and is currently an NIH T-32 postdoctoral research fellow with Massachusetts General Hospital and Harvard Medical School, Department of Radiology. His research interests include graphical causal models and deep learning for clinical and biomedical applications.



Xiaoyu Ge received the BS degree in computer science from the University at Buffalo, Buffalo, New York. He is currently working toward the PhD degree with the Department of Computer Science, University of Pittsburgh, Pittsburgh, Pennsylvania, since 2013. His research interests include data exploration, data summarization, and urban informatics.

Arun Balajiee received the BS degree in computer science from the Indian Institute of Technology Hyderabad, Telangana, India. He is currently working toward the master's degree in computer science at the University of Pittsburgh, Pittsburgh, Pennsylvania. His current research interests include database management, data streams processing, and data analytics.

Daniel J. Shirer received the BS degree in computer science and math, and the BA degree in history and philosophy of science from the University of Pittsburgh, Pittsburgh, Pennsylvania. He currently works as a software engineer for Pavcon, and is interested in developing accurate and interpretable machine learning models.

Isha Das is currently a senior at North Allegheny High School, Pittsburgh, Pennsylvania. She participated in this project during summer research through the UPMC Hillman Cancer Center. Her research interests include bioinformatics, deep learning, and molecular dynamics.

Panayiotis V. Benos received the PhD degree from the University of Crete, Heraklion, Greece, in 1997. He is currently a professor and vice chair with the Department of Computational and Systems Biology, University of Pittsburgh. He has more than 80 publications in peer-reviewed journals such as the *Nature*, the *Science*, the *Proceedings of the National Academy of Sciences of the United States of America*, and others. His work has been presented in many international conferences and he has given invited talks in United States and Europe, while he co-organized and co-chaired international conferences in computational biology. He has been continuously funded through grants he received from NIH and NSF since 2003 and he has a large number of collaborations. His research interests include integration of multi-modal data for the study of chronic diseases and cancer. His group develops probabilistic graphical models and other machine learning methodologies and applies them to a variety of biomedical and clinical problems.

Panos K. Chrysanthis (Senior Member, IEEE) received the BS degree from the University of Athens, Athens, Greece, in 1982, and the MS and PhD degrees from the University of Massachusetts at Amherst, Amherst, Massachusetts, in 1986 and 1991, respectively. He is a professor of computer science and the founding director of the Advanced Data Management Technologies Laboratory, University of Pittsburgh. He is also an adjunct professor with the Carnegie Mellon University and with the University of Cyprus. His research interests include the areas of data management (big data, databases, data streams, and sensor networks), distributed and mobile computing, workflow management, operating systems, and real-time systems. He received the US National Science Foundation CAREER Award and he is an ACM distinguished scientist.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.

Authorized licensed use limited to: University of Pittsburgh Library System. Downloaded on November 11,2025 at 19:44:41 UTC from IEEE Xplore. Restrictions apply.