

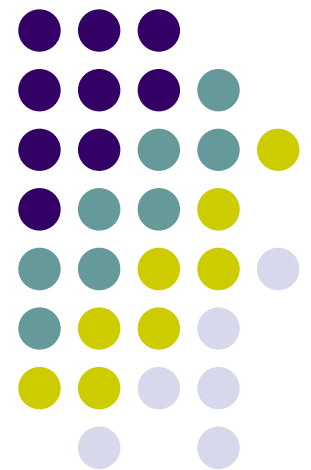
Learning from Multiple Sources for Chemical Toxicity Prediction

Jun Huan

Department of Electrical Engineering and Computer Science

University of Kansas

<http://people.eecs.ku.edu/~jhuan/>





Introduction

- Our goal is to predict *in vivo* chemical toxicity
 - Many cofounding factors:
 - ADME properties
 - Metabolites
 - Chemical cellular localization
 - Genetic background
 - Tissue specific
 - Developmental stages specific
 - Chemical-protein, chemical-gene, chemical-chemical interactions, chemical-physiological condition interaction



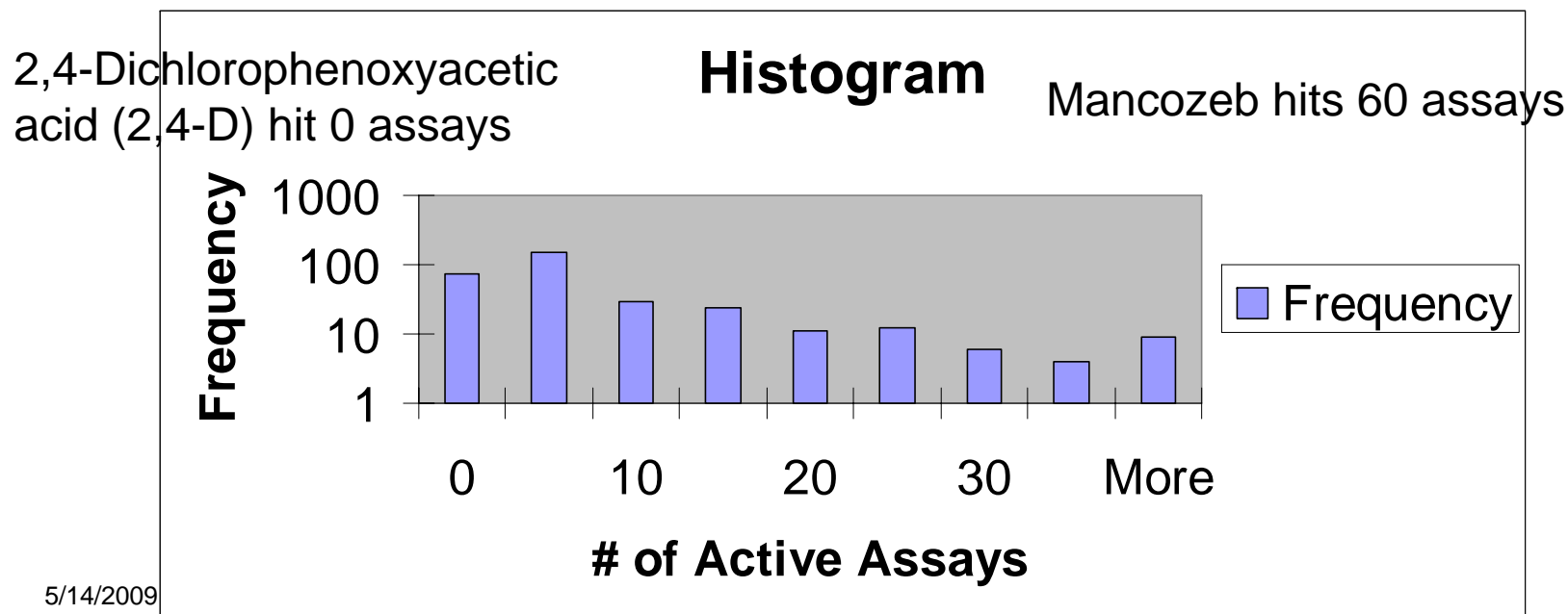
Protein-Chemical Interaction

- Protein-chemical interaction profiles reveals important information for chemical toxicity prediction
 - Part of Richard's Law 1
- Our data sources
 - EPA Novascreen results
 - High quality, 239 assays
 - PubChem
 - Not well cleaned, 672 target based screening results



Nature of the Interaction Data

- Novascreen: 320 chemicals evaluated with 239 targets
- Usually Sparse
 - On average each compound is considered as “active” in $6.7/239 = 2.8\%$ assays





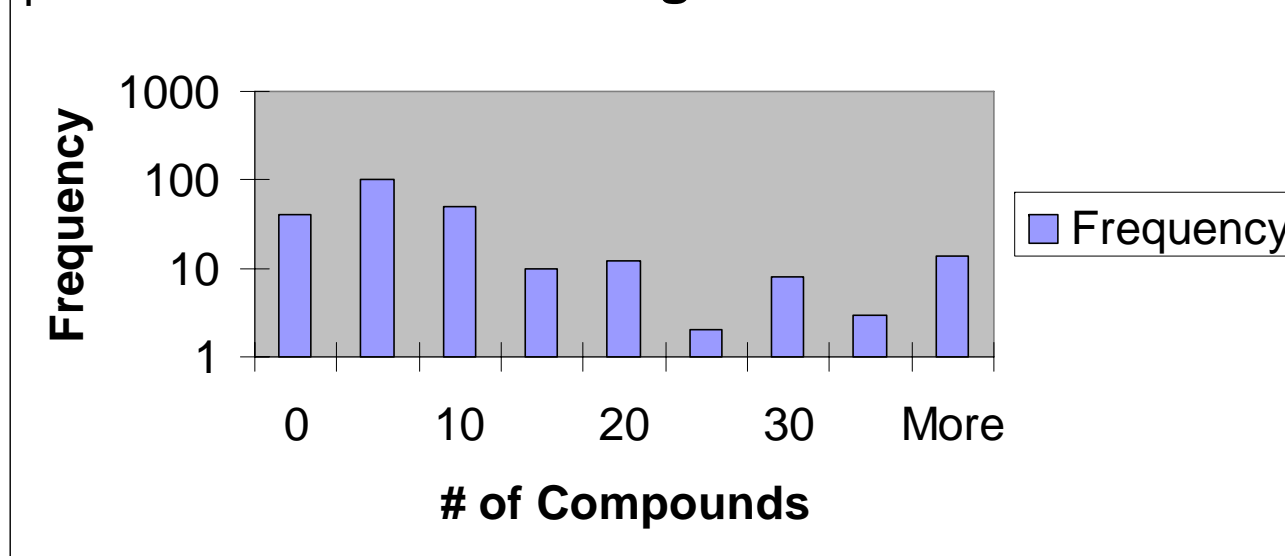
Interaction Sparsity

- On average each protein interacts with $8.9/320 = 2.8\%$ of the compounds

NVS_ENZ_hCASP2 interacts
with 0 compounds

Histogram

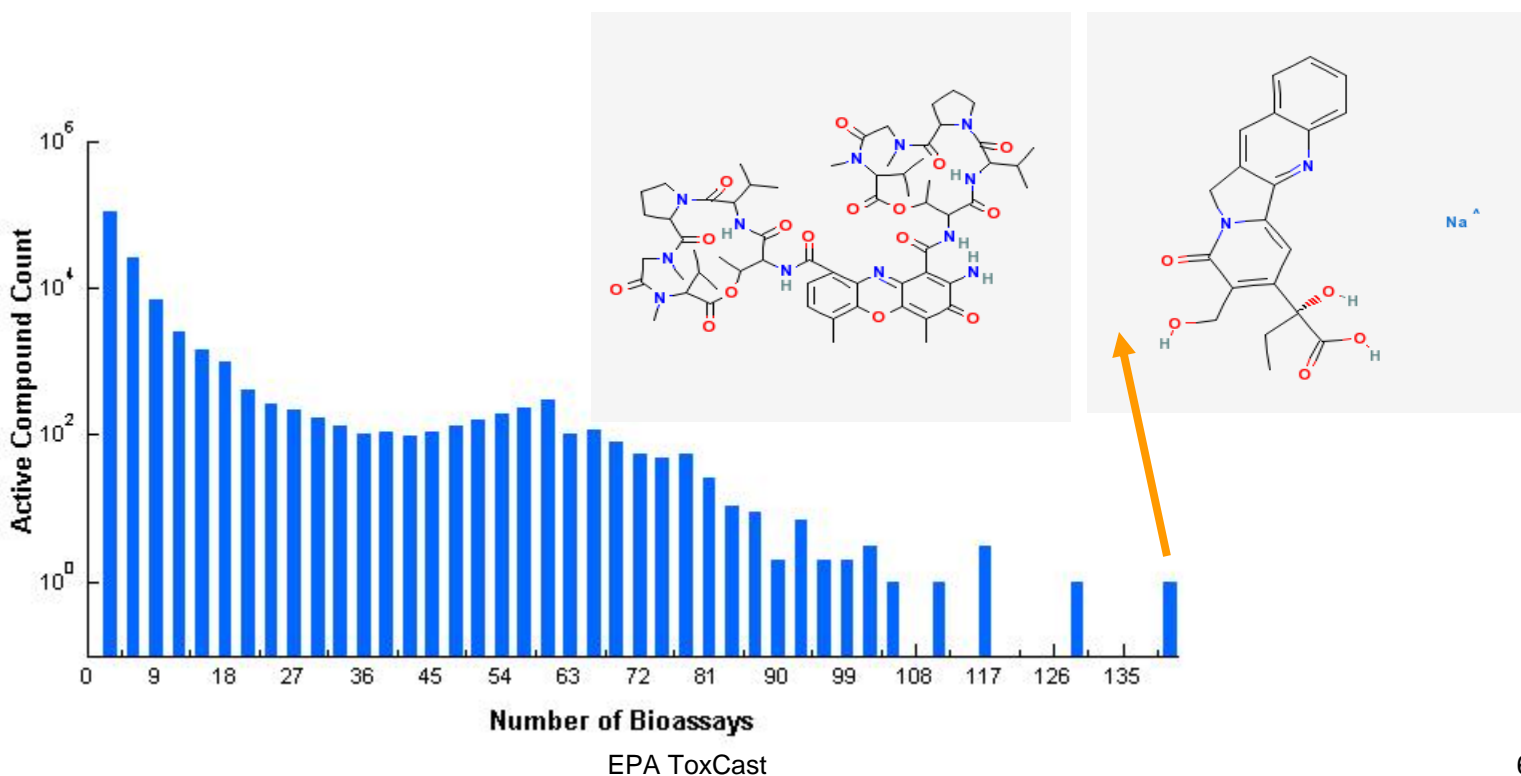
NVS_ADME_hCYP2C19
interacts with 116 compounds



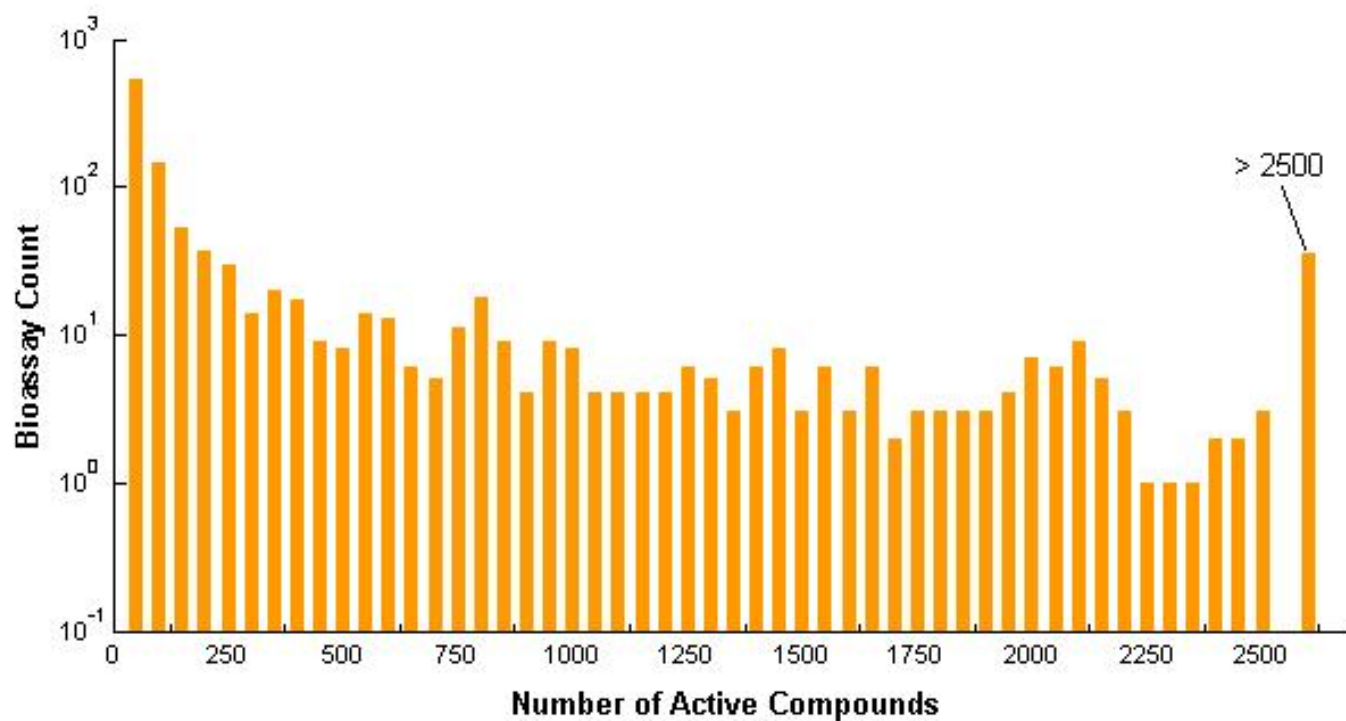
Interaction Sparsity is also observed in PubChem Data



- >1M compounds evaluated against 672 targets



Interaction Sparsity is also observed in PubChem Data



Predicting Sparse Interaction is Challenging



- Sampling based methods
- Learning technique based methods
 - One-class classification
 - Multi-task learning

Machine Learning in Protein-Chemical Interaction Prediction

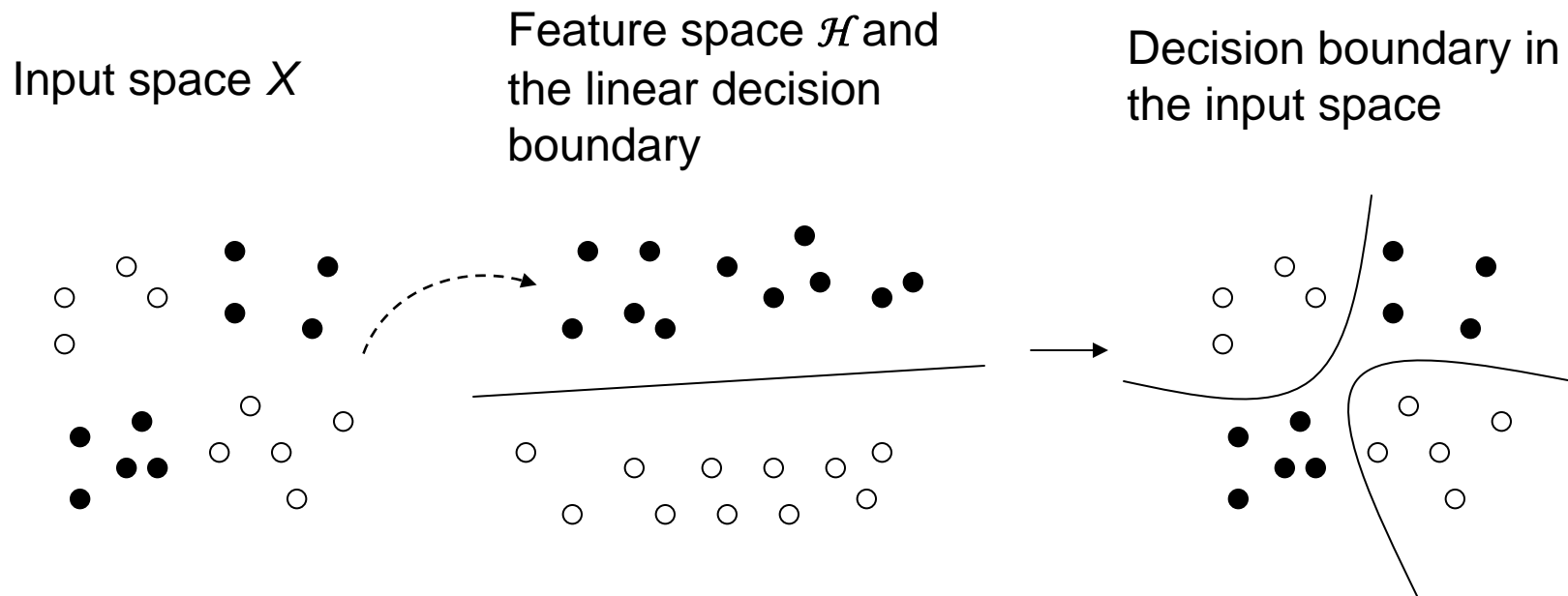


- Machine learning provide a wide range of choices for feature selection and model construction
 - SVM
 - Random forest
 - kNN
 - And many others
- SVM belongs to a larger group of methods called kernel machines
 - General idea is to introduce non-linear decision boundary using kernel functions
 - Kernel PCA, kernel LDA, kernel kNN



Kernel Function

- A kernel function maps a training case to a higher dimensional space

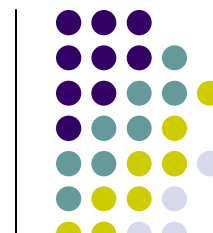


Kernel Functions for Biomolecules

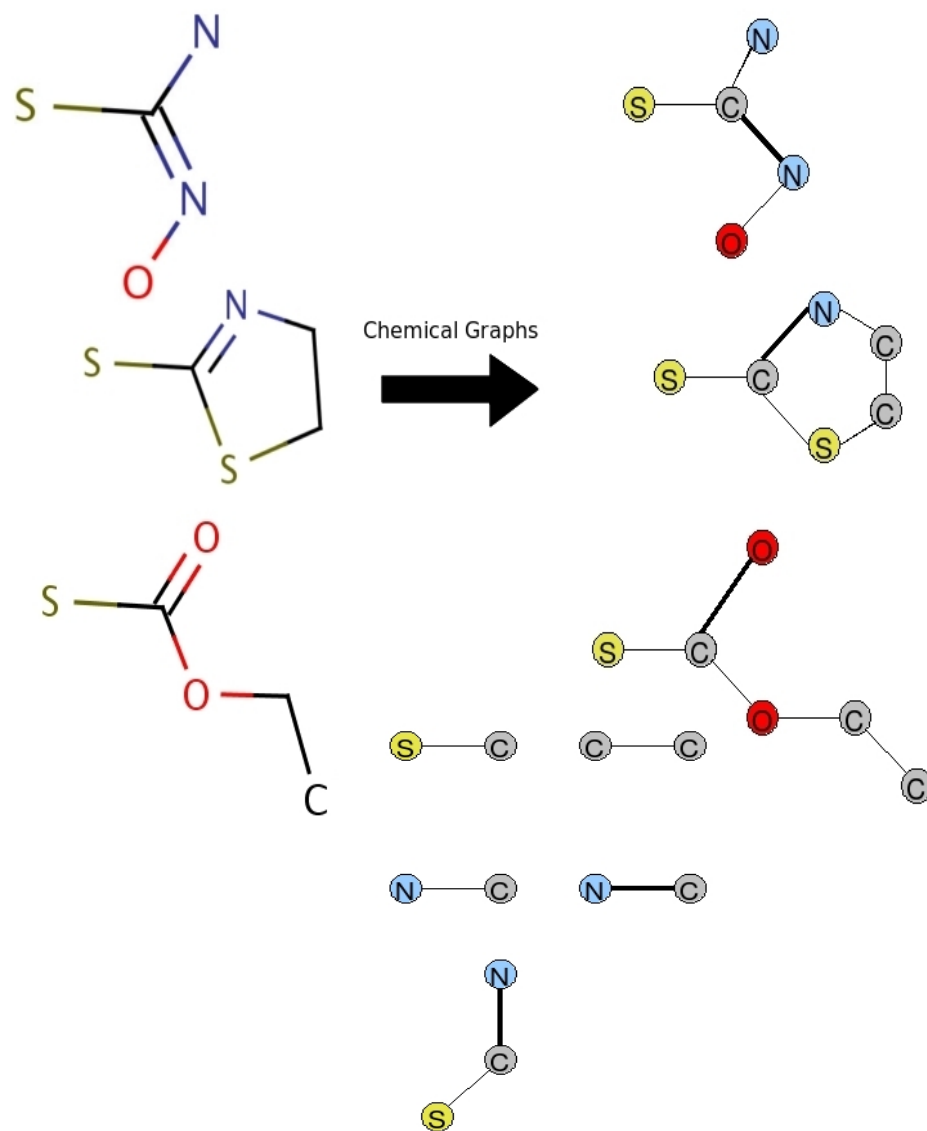


- Protein sequence kernels
 - Spectrum kernel (Leslie et al. 2002, Pac Symp Biocomput)
 - Structural alignment kernels (Qiu et al. 2007, Bioinfo. 23(9): 1090-98)
 - Protein graph kernels (Borgwardt et al. 2005, Bioinfo. 21, i47-i56)
- Chemical structure kernels

Chemical Structure Kernels



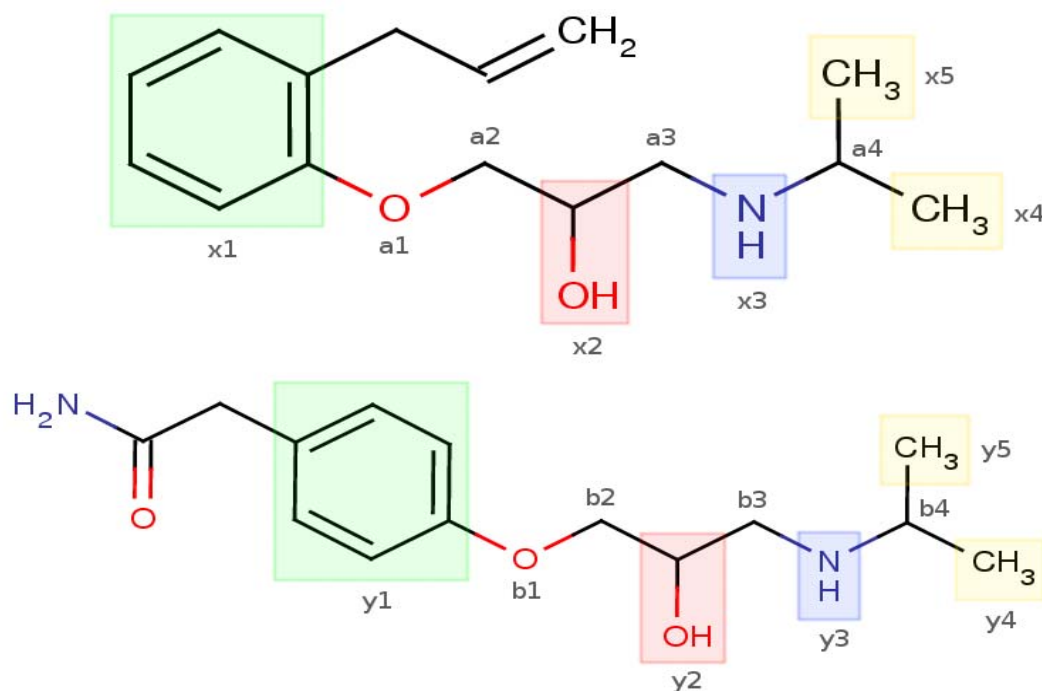
- Use graph representation rather than numeric properties.
- Transformation of chemicals to graphs is straight forward.
 - Atoms correspond to vertices.
 - Bonds correspond to edges.
 - Vertices and edges are labeled with atom element and bond type, among other properties.

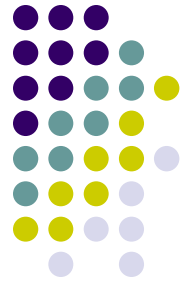


Optimal Assignment Kernel



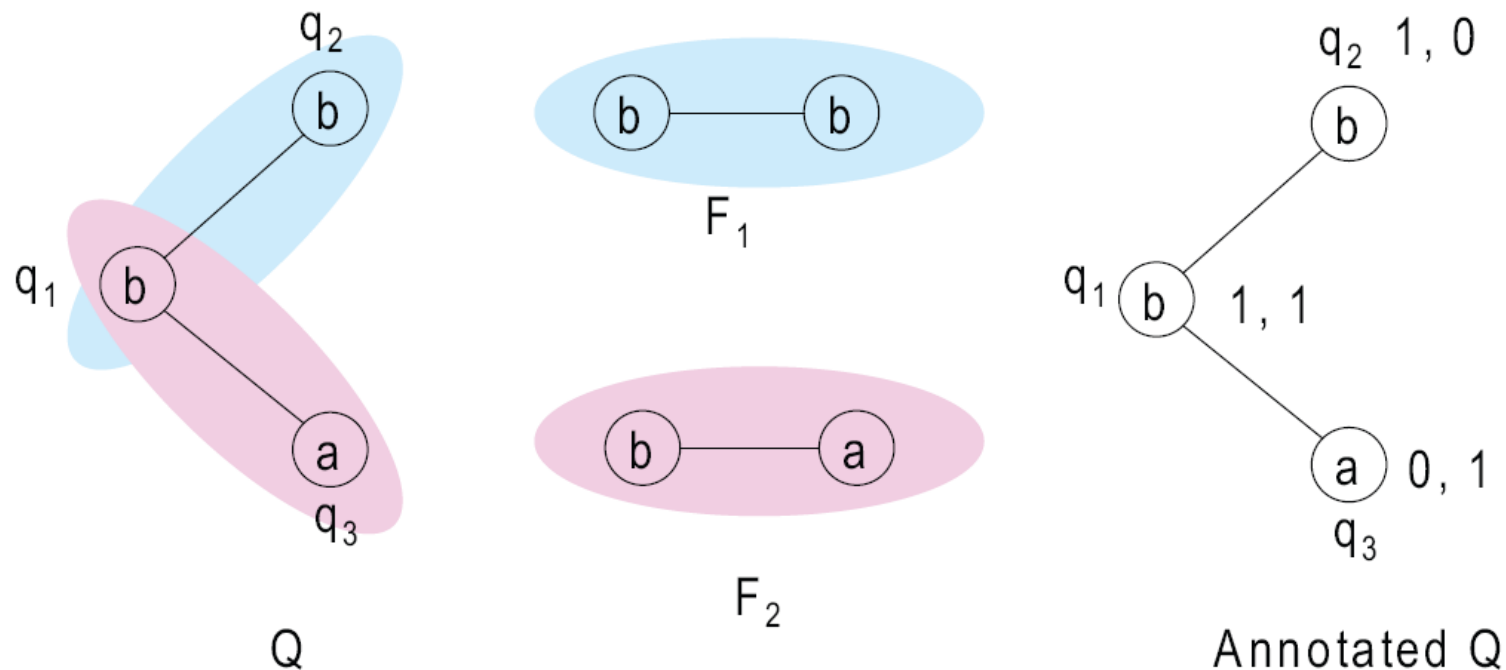
- Graph kernel function that computes molecular similarity by finding the maximal weighted bipartite graph between two sets of graph vertices.



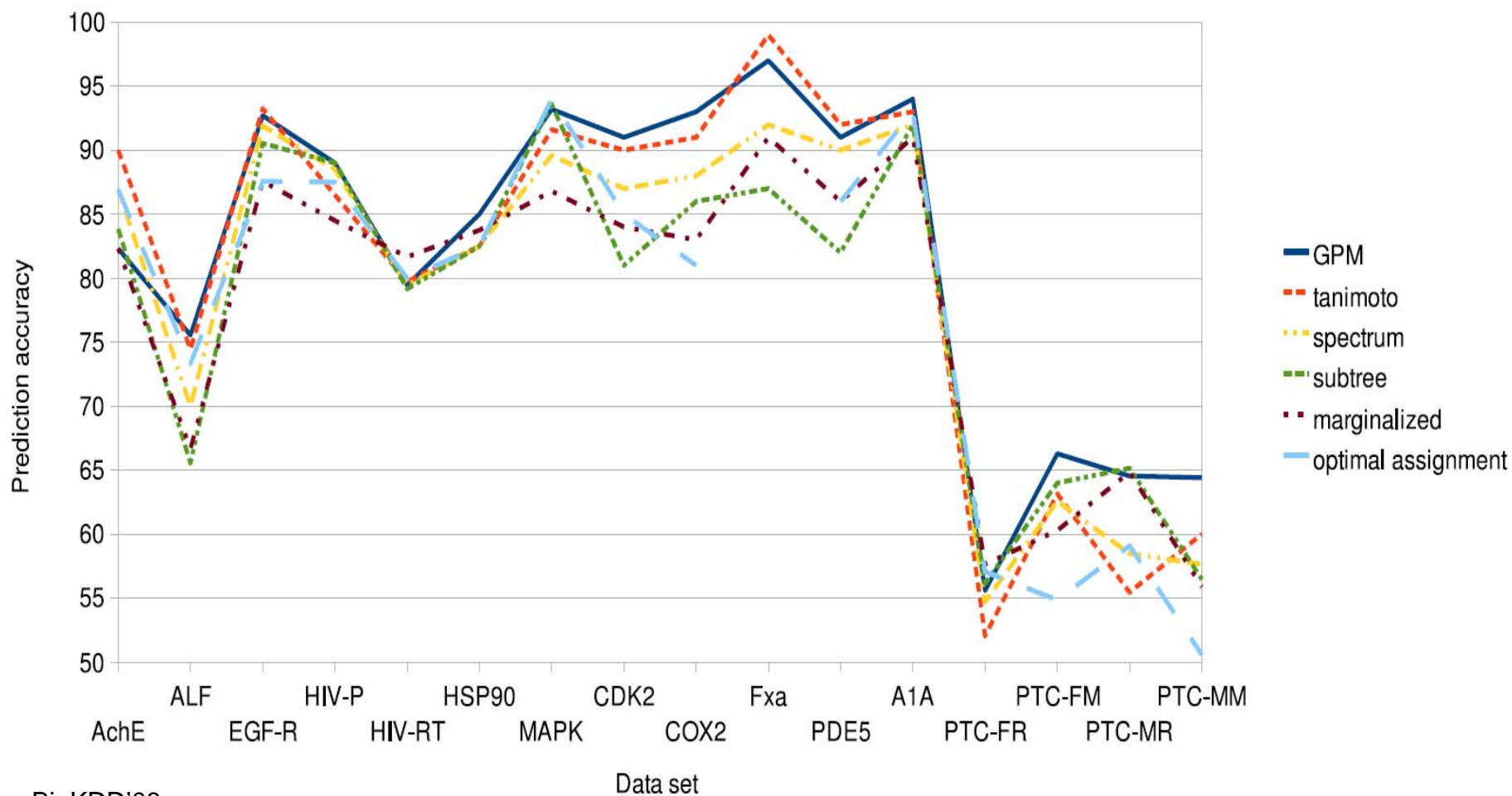


Graph Pattern Diffusion Kernel

- Node labeling with membership test
 - Each node in a graph is labeled with a vector of bits indicating *memberships* to a set of features

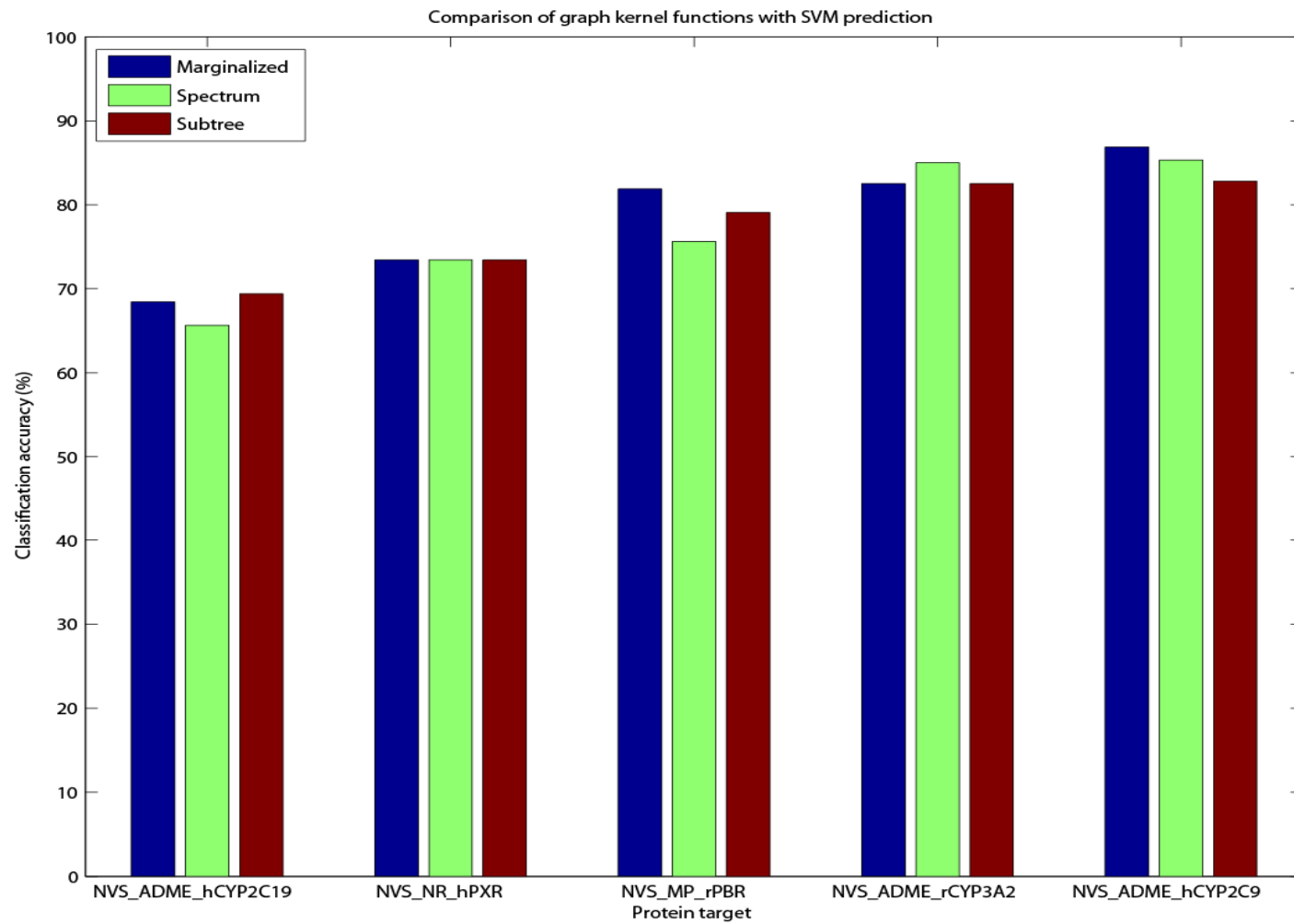


Kernel Based Protein-Chemical Interaction Prediction



BioKDD'08

For ToxCast Data

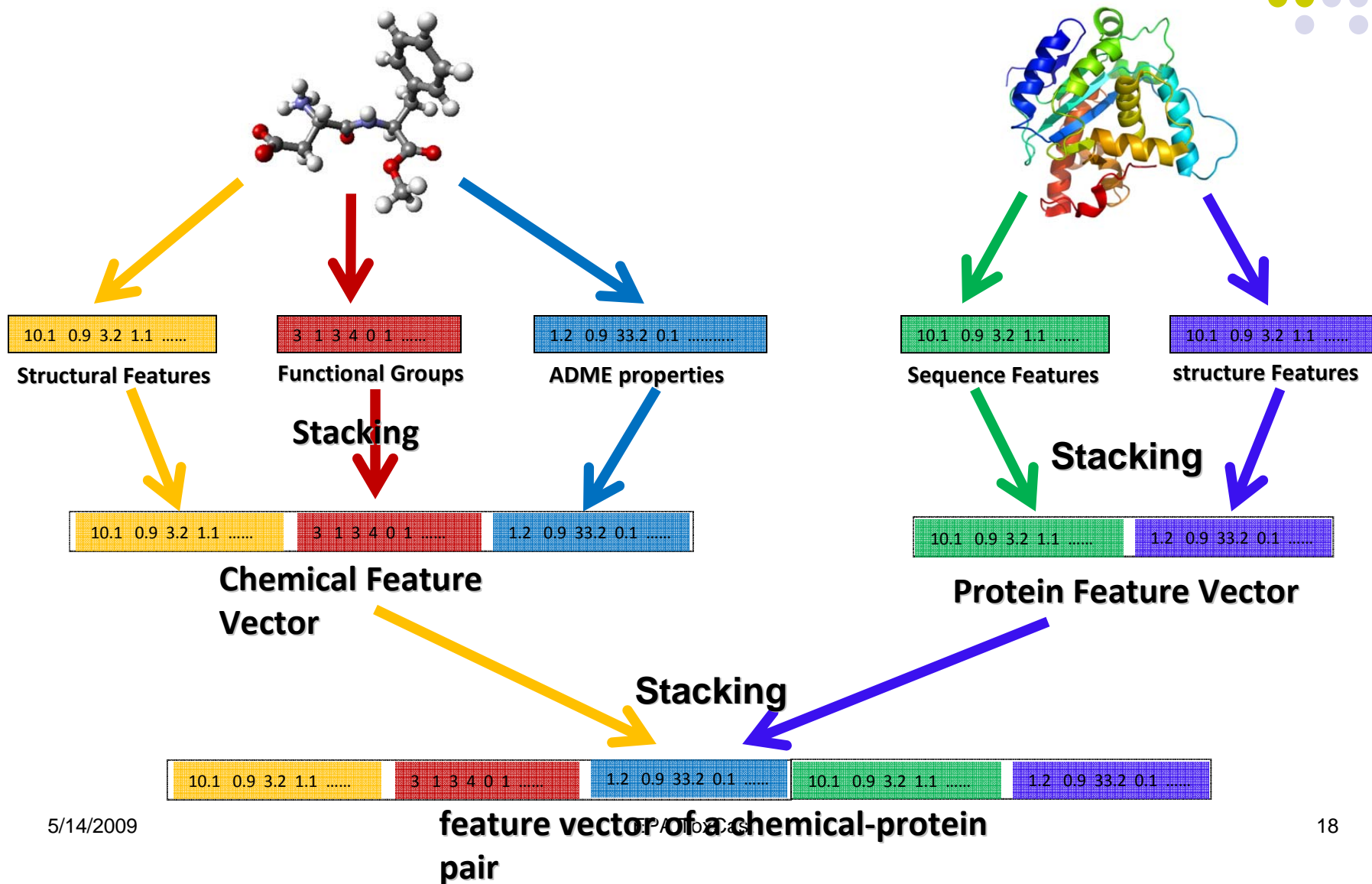
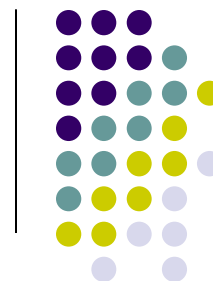




Data Integration

- Collect more information regarding the chemical and their targets
 - For chemicals, we may extract many different types of descriptors
 - For proteins, we may extract descriptors from protein sequences and/or structures
 - K-mers, sequence profiles, structure patterns, binding sites, functional sites
 - Chemical-induced gene expression profiles
 - Chemical-induced phenotypes

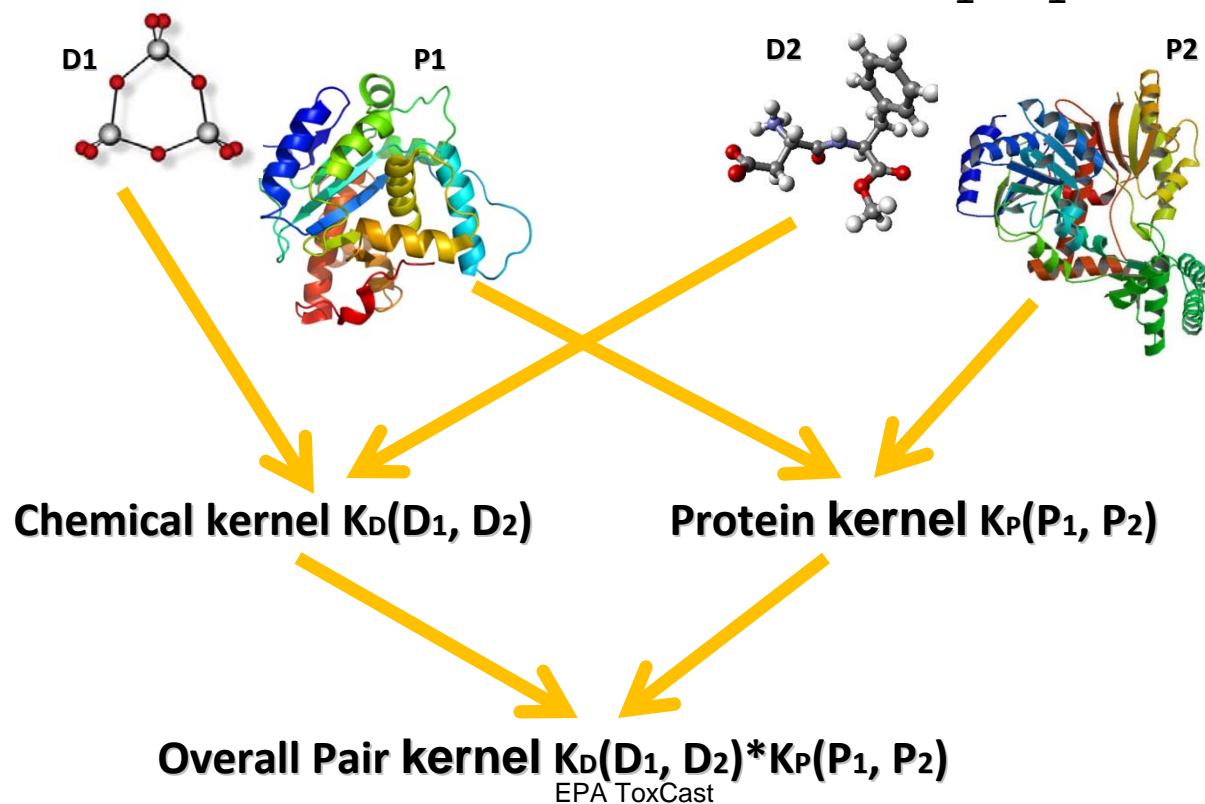
Data Integration - Stacking



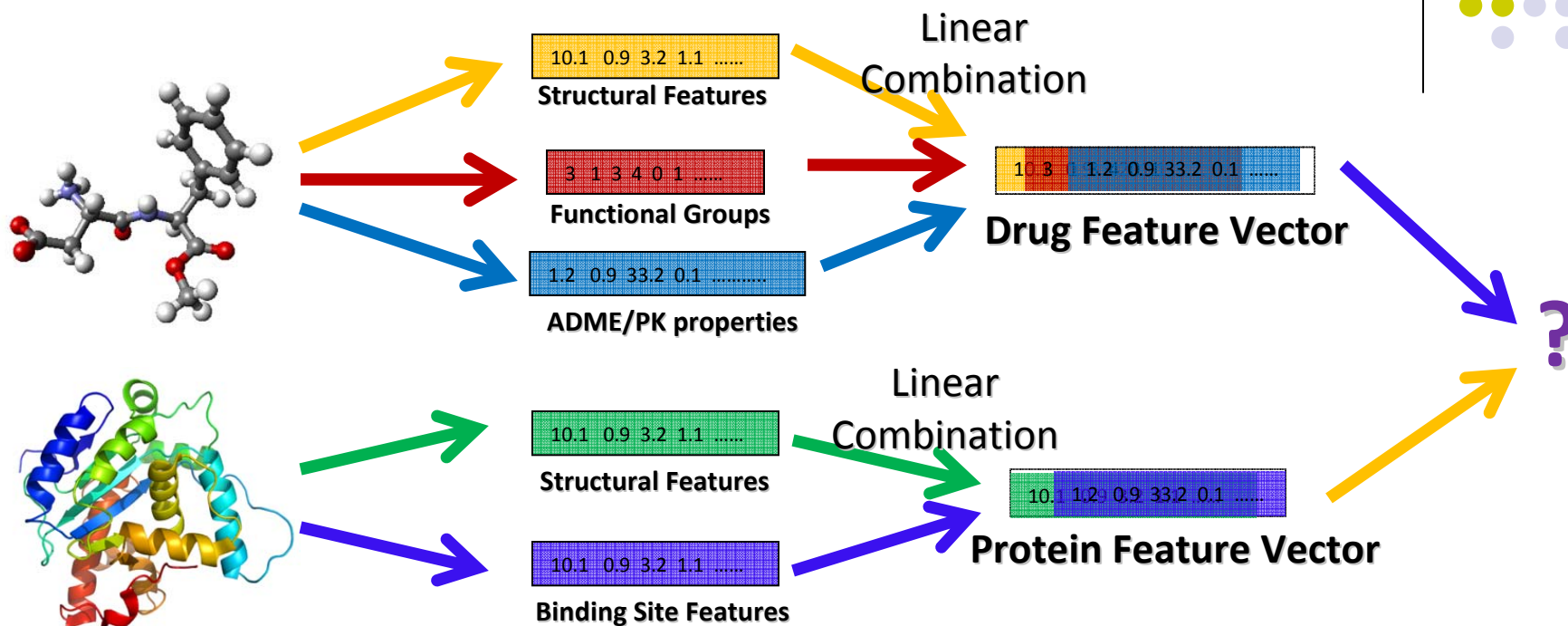
Data Representation in the Kernel Space



- Predict protein-chemical interaction using related data from both chemicals and proteins
- Kernel of two protein-chemical pairs (D_1, P_1) and (D_2, P_2)



Data Integration in the Kernel Space

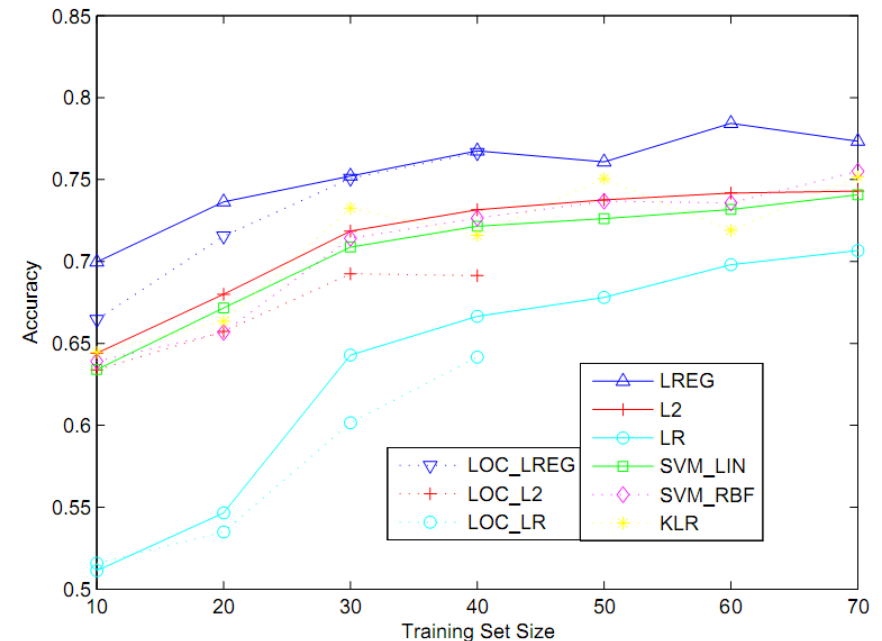
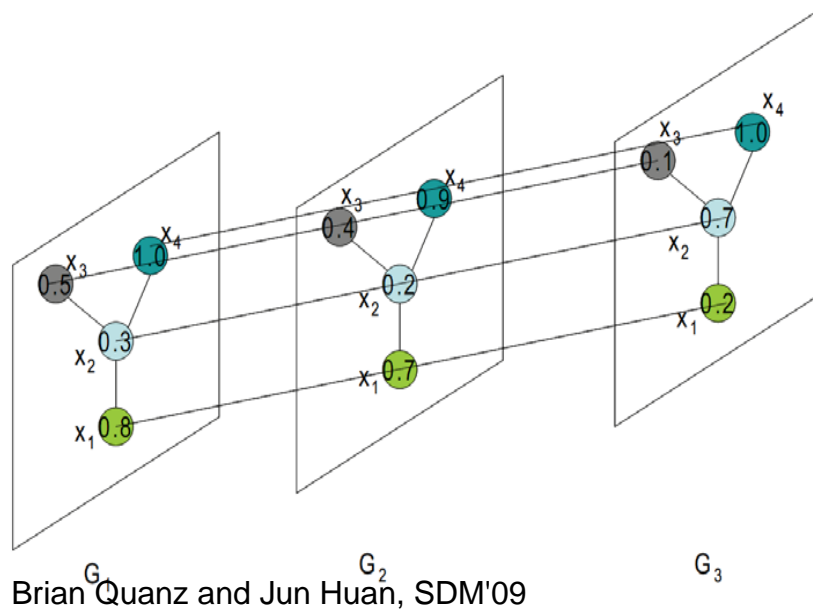


- Heterogeneous data: *each kernel method for each type of data*
- The overall kernel is a **linear combination** of individual kernels
- The coefficient of each kernel will be optimized to achieve the best prediction accuracy.

Data Integration in Kernel Space (II)



- Some data, in particular biological pathways, may be incorporated in a supervised learning algorithm as a regularization factor
 - Fairly new idea
 - Does not add computational cost



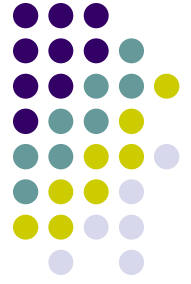
Group Members and Collaborators



- Ph.D. Students
 - Fei, Hongliang
 - Jia, Yi
 - Quanz, Brian
 - Smalter, Aaron
 - Zhang, Jingtao
- Collaborators:
 - Dr. Alex Tropsha, UNC School of pharmacy
 - Drs. Gerald Lushington, Jeff Aubé, KU School of pharmacy
 - Dr. Deepak Bandyopadhyay, GSK
 - Dr. Leming Shi, FDA

Group website:

<http://people.eecs.ku.edu/~jhuan/>



Acknowledgments

- The work is partially supported by
 - The KU NIH “Special Chemistry Center” (U54 HG005031, \$20M), 2008-2014
 - NSF CNS 0821625 “MRI: Acquisition of an Advanced Computational Infrastructure for Modeling Biological Systems”, 2008-2011

