

Many Features, Few Samples:

From cheminformatics to bioinformatics

Kristin P. Bennett

*Department of Mathematical Sciences
Rensselaer Polytechnic Institute*

and RPI DDASSL Project Members:

*C. Breneman, M. Embrechts, J. Bi, M.
Momma, N. Sukumar, M. Song*

Cheminformatics Problem

Given for each Molecule i

- x_i Descriptor vector
- Bioresponse y_i

Construct a function

$$f(x_i) \approx y_i$$

to predict bioresponse

Catch: many descriptors/attributes (600-1000+)

very few data points (30-200)

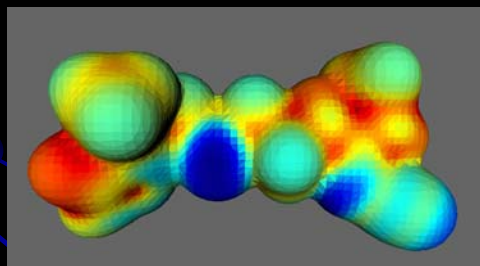
descriptors very correlated

Electron Density-Derived TAE-Wavelet Descriptors

1) Surface properties are encoded on 0.002 e/au^3 surface

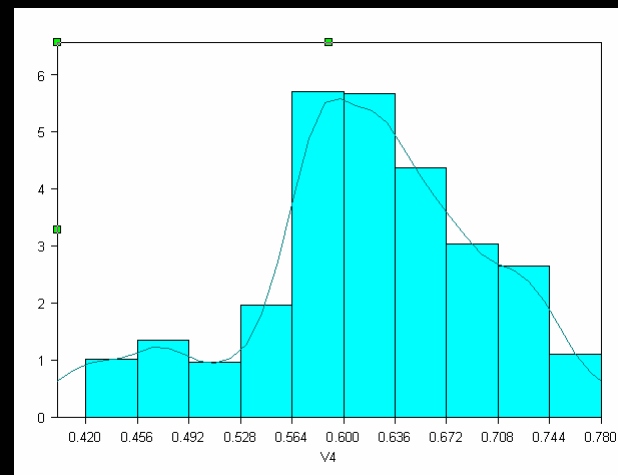
Breneman, C.M. and Rhem, M. [1997] *J. Comp. Chem.*, Vol. 18 (2), p. 182-197

2) Histograms or wavelet encoded of surface properties give TAE property descriptors

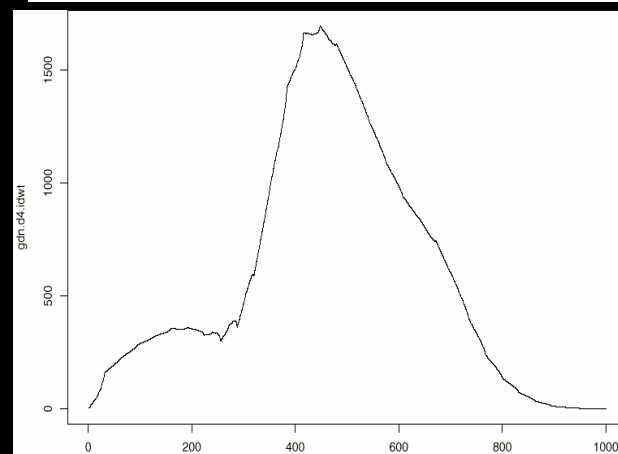


PIP (Local Ionization Potential)

Histograms

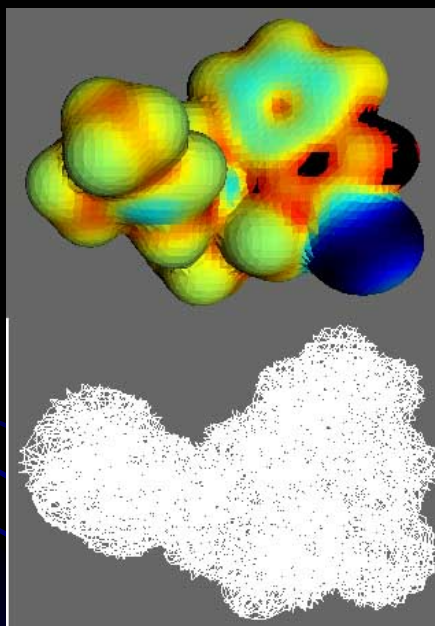


Wavelet Coefficients

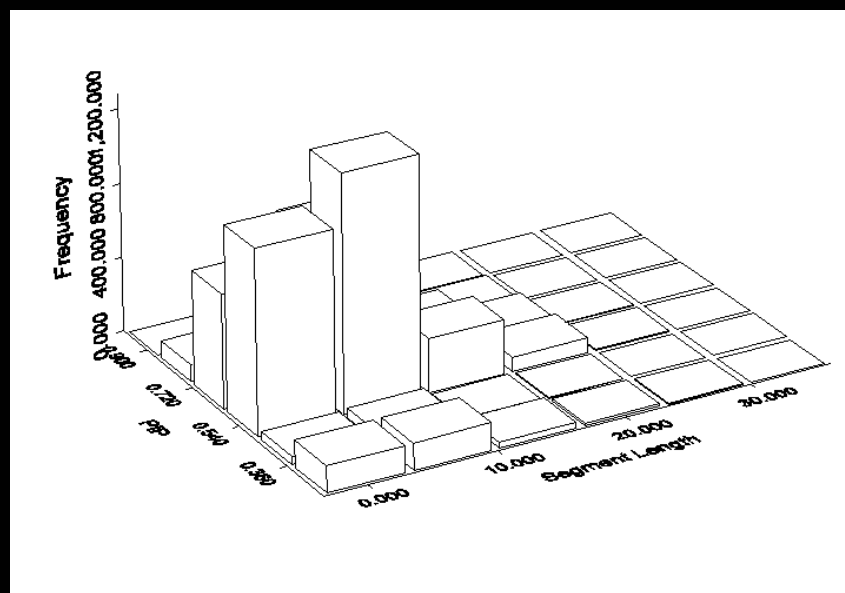


PEST Hybrid Property/Shape Descriptors

- Surface properties and shape information are encoded into alignment-free descriptors




PIP vs Segment Length



- 9 different surface properties

Many features/Little Data Issues

- Overfitting
 - Feature selection
 - Difficult validation
 - Model/parameter selection
 - High model variance
 - Not confident in any one model
- 

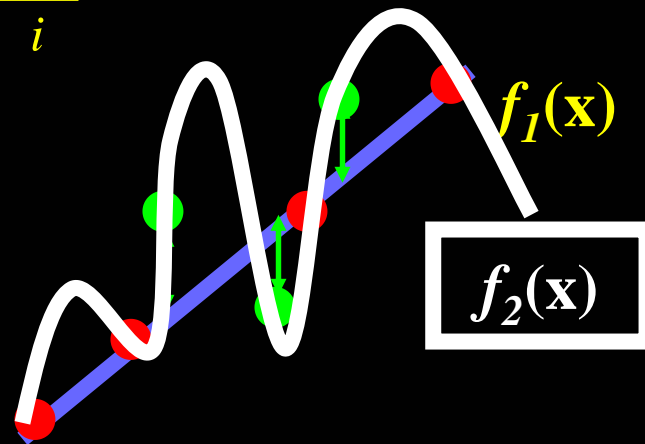
DDASSL Learning Methodology: One Method with Three Engines

- Method
 - Regularized Kernel Learning Engines
 - Bagged Feature Selection/Visualization
 - Bagged Final Models
- Learning Engines (linear and kernel)
 - **Support Vector Machine (SVM)**
 - Partial Least Squares (PLS)
 - **Boosted Latent Analysis (BLA)**

Minimize Regularized Loss

- Minimize the training error and capacity

$$\min_f \sum_i \text{Loss}(f(x_i), y_i) + P(f)$$

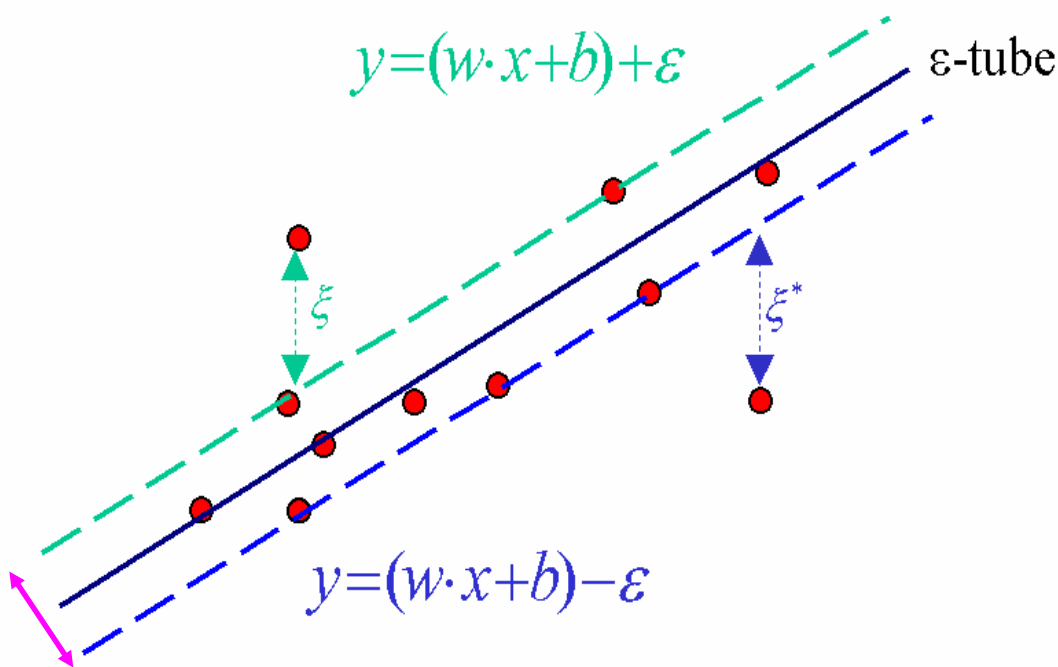


- Overfitting is likely with high-capacity functions
- Capacity control makes good generalization possible even in very high-dimensional input spaces

Support Vector Regression (SVR)

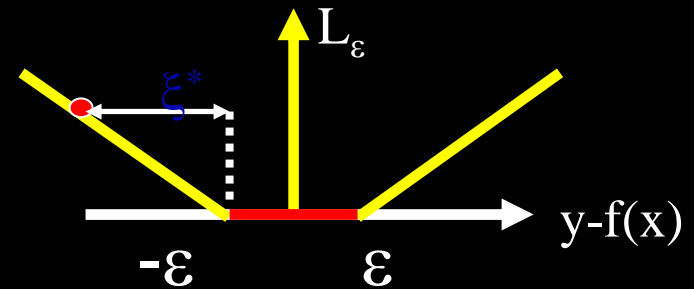
- Minimize the regularized empirical error:
 - training error + model complexity

$$\min_{w, b, \xi_i, \xi_i^*} C \sum_{i=1}^l (\xi_i + \xi_i^*) + \frac{1}{2} \|w\|^2$$



ϵ -insensitive loss function:

$$L_\epsilon(y - f(x)) := \max(0, |y - f(x)| - \epsilon)$$



- Overfitting is avoided by controlling the model complexity: $\|w\|$
- Add kernels to create nonlinear functions

Feature Selection via Sparse SVM/LP

- Construct linear ν -SVM using 1-norm LP:

$$\min_{w, b, \varepsilon, z, z^*} \frac{C}{\ell} \sum_{i=1}^{\ell} (z_i + z_i^*) + C\nu(\varepsilon + \varepsilon^*) + \|w\|_1$$

s.t

$$(x_i \cdot w + b - y_i) + z_i \geq -\varepsilon$$

$$(x_i \cdot w + b - y_i) - z_i^* \leq \varepsilon^*$$

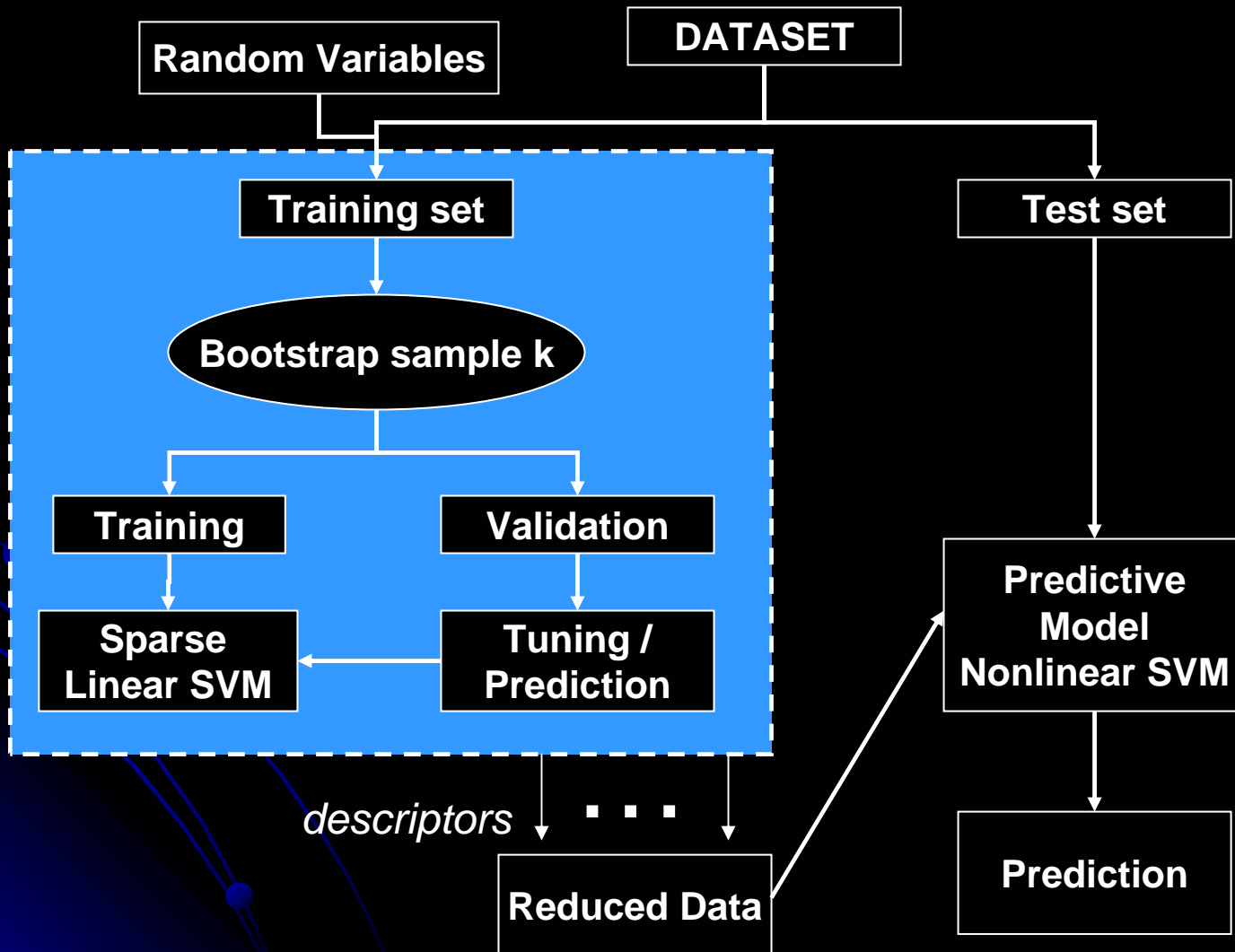
- Pick best C for SVM $z_i, z_i^*, \varepsilon, \varepsilon^* \geq 0 \quad i=1, \dots, \ell$

- Keep descriptors

with nonzero coefficients

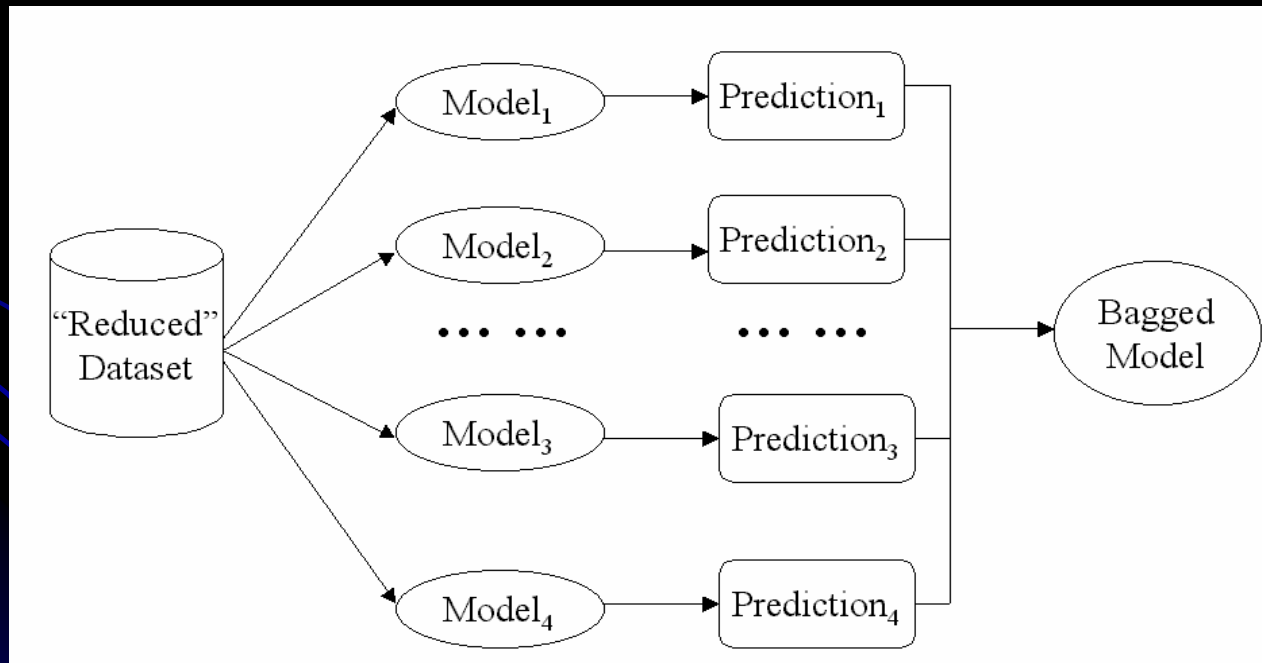
$$|w_i| > 0$$

Bagged Variable Selection

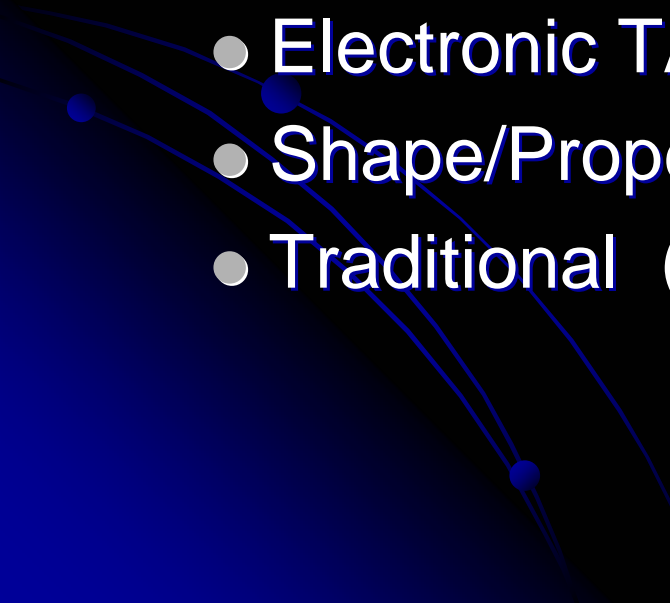


Final Bagged Predictive Model

- Achieve the better generalization performance
 - construct a series of non-linear SVM models
 - use the average of all models as final prediction to reduce variance



CACO-2 Data

- Human intestinal cell line
 - Predicts drug absorption
 - 27 molecules with tested permeability
 - 718 descriptors generated
 - Electronic TAE
 - Shape/Property (PEST)
 - Traditional (MOE)
- 

Molecular Surface Properties

- **Electronic Properties**

- Electrostatic Potential

$$EP(r) = \sum_{\alpha} \frac{Z_{\alpha}}{|r - R_{\alpha}|} - \int \frac{\rho(r') dr'}{|r - r'|}$$

- Electronic Kinetic Energy Density

$$K(r) = -(\psi * \nabla^2 \psi + \psi \nabla^2 \psi^*)$$

- Electron Density Gradients $\nabla \rho \cdot N$

$$G(r) = -\nabla \psi * \cdot \nabla \psi$$

- Laplacian of the Electron Density

$$L(r) = -\nabla^2 \rho(r) = K(r) - G(r)$$

- Local Average Ionization Potential

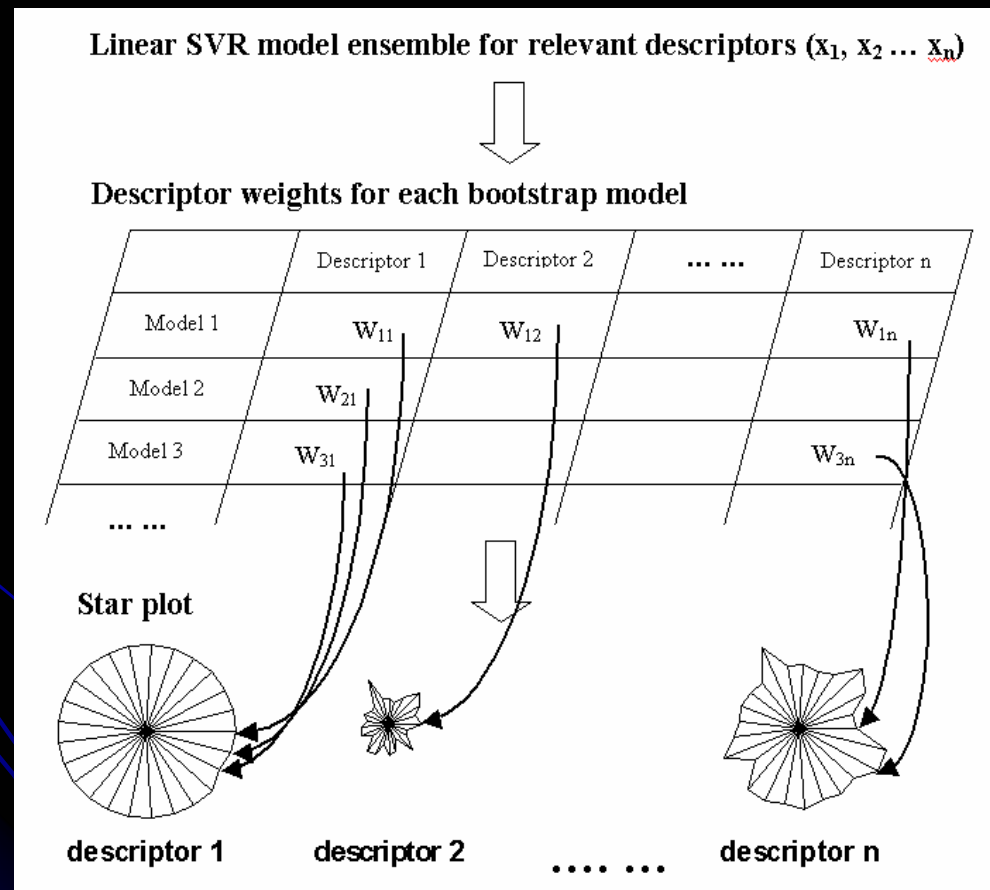
$$PIP(r) = \sum_i \frac{\rho_i(r) |\epsilon_i|}{\rho(r)}$$

- Bare Nuclear Potential (BNP)

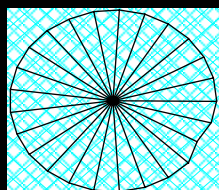
- Fukui function $F+(r) = \rho_{HOMO}(r)$

Visualization of feature selection results

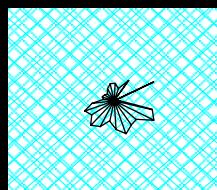
- To investigate the relative importance of selected descriptors and their consistency



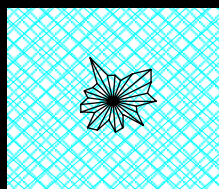
Caco-2 – 14 Features (SVM)



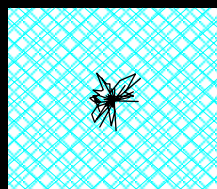
a.don



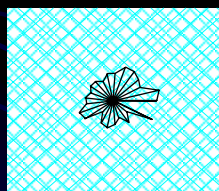
DRNB10



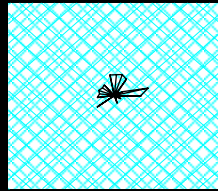
KB54



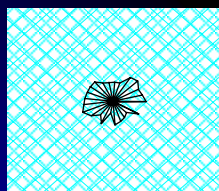
ABSDRN6



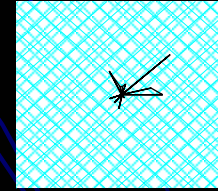
SMR.VSA2



PEOE.VSA.FPPOS



ANGLEB45



DRNB00



PEOE.VSA.FNEG



ABSKMIN



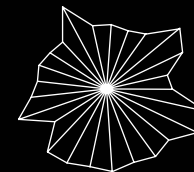
SIKIA



BNPB31



FUKB14

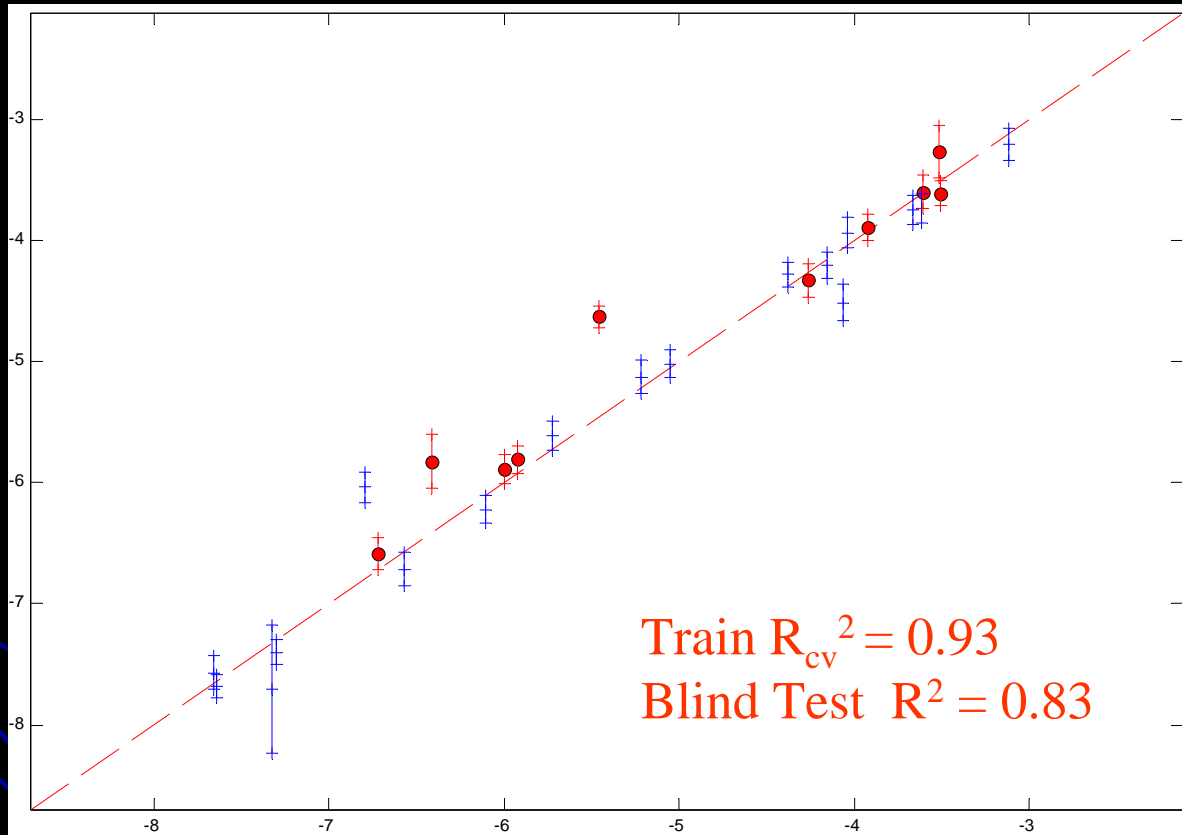


SlogP.VSA0

- Each star represents a descriptor
- Each ray is a separate bootstrap
- The area of a star represents the relative importance of that descriptor
- Descriptors shaded cyan have a negative effect
- Unshaded ones have a positive effect

- **Hydrophobicity** - a.don
- **Size and Shape** - ABSDRN6, SMR.VSA2, ANGLEB45
Large is bad. Flat is bad. Globular is good.
- **Polarity** – PEOE.VSA...: negative partial charge good.

Bagged SVM (RBF) Caco-2



Before feature selection $R^2 = .66$

New Learning Engine: BLA

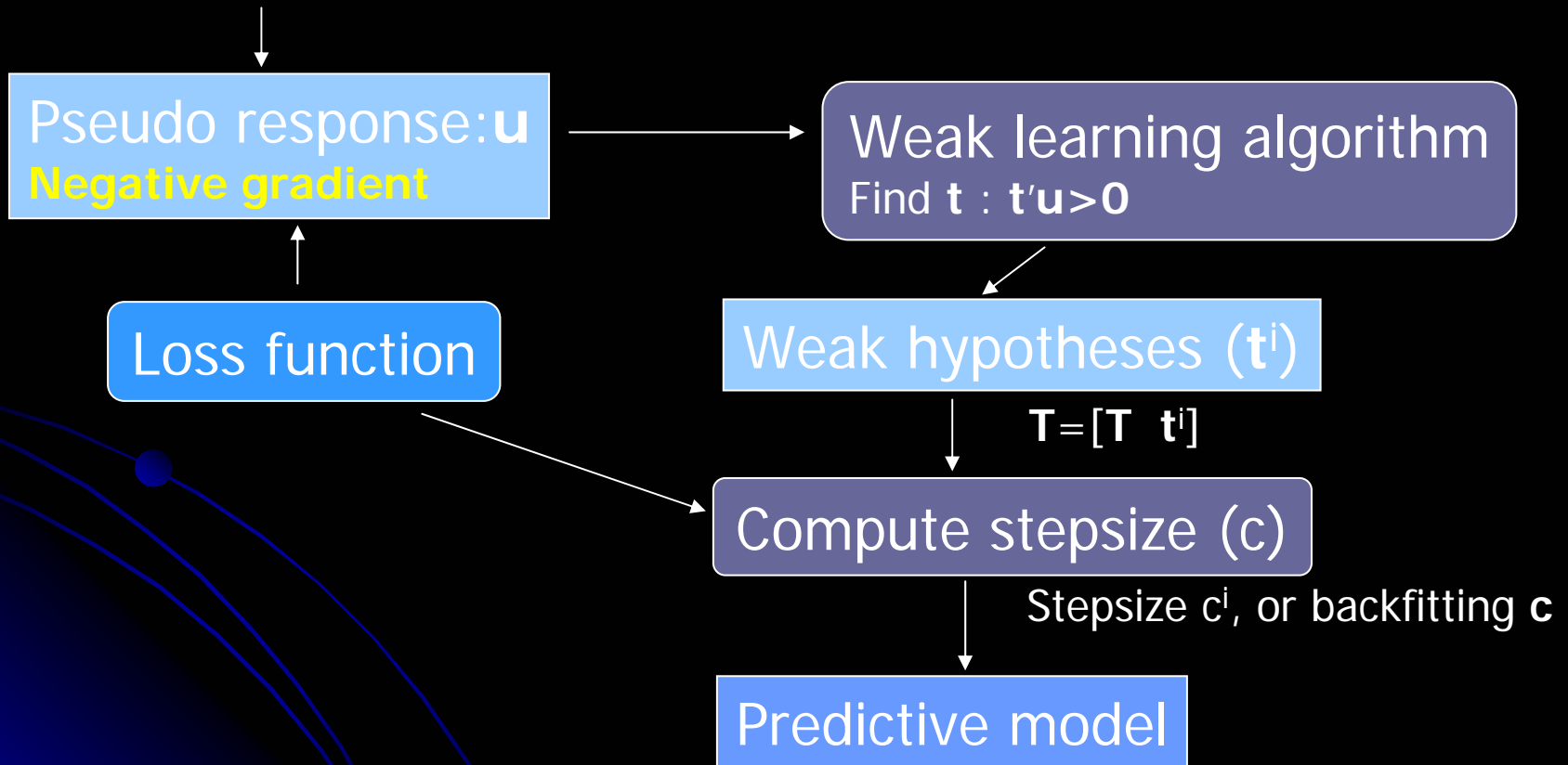
Boosted Latent Analysis

Construct Orthogonal Latent Features and corresponding predictive model for (sub) differentiable loss functions

- Orthogonal Boosting of linear functions
- For least squares loss, equivalent to PLS
- Easy to tune
- Easy to implement algorithm: small changes for different loss functions
- Feature selection for linear models
- Kernelizable for nonlinear models

Review of AnyBoost (Mason et al. 1999)

Initial $\mathbf{F} = \mathbf{t}^0$



Steepest descent

$$\mathbf{F} = \sum_{i=1}^L \mathbf{t}^i \mathbf{c}^i + \mathbf{t}^0$$

Orthogonal AnyBoost

Initial $\mathbf{F} = t^0$

Pseudo response: \mathbf{u}
Negative gradient

Weak learning algorithm
Find $\mathbf{t} : \mathbf{t}'\mathbf{u} > 0$ $\mathbf{t}'\mathbf{t}^j = 0 \quad j=1, \dots, i-1$

Loss function

Weak hypotheses (\mathbf{t}^i)

$$\mathbf{T} = [\mathbf{T} \quad \mathbf{t}^i]$$

Compute stepsize (c)

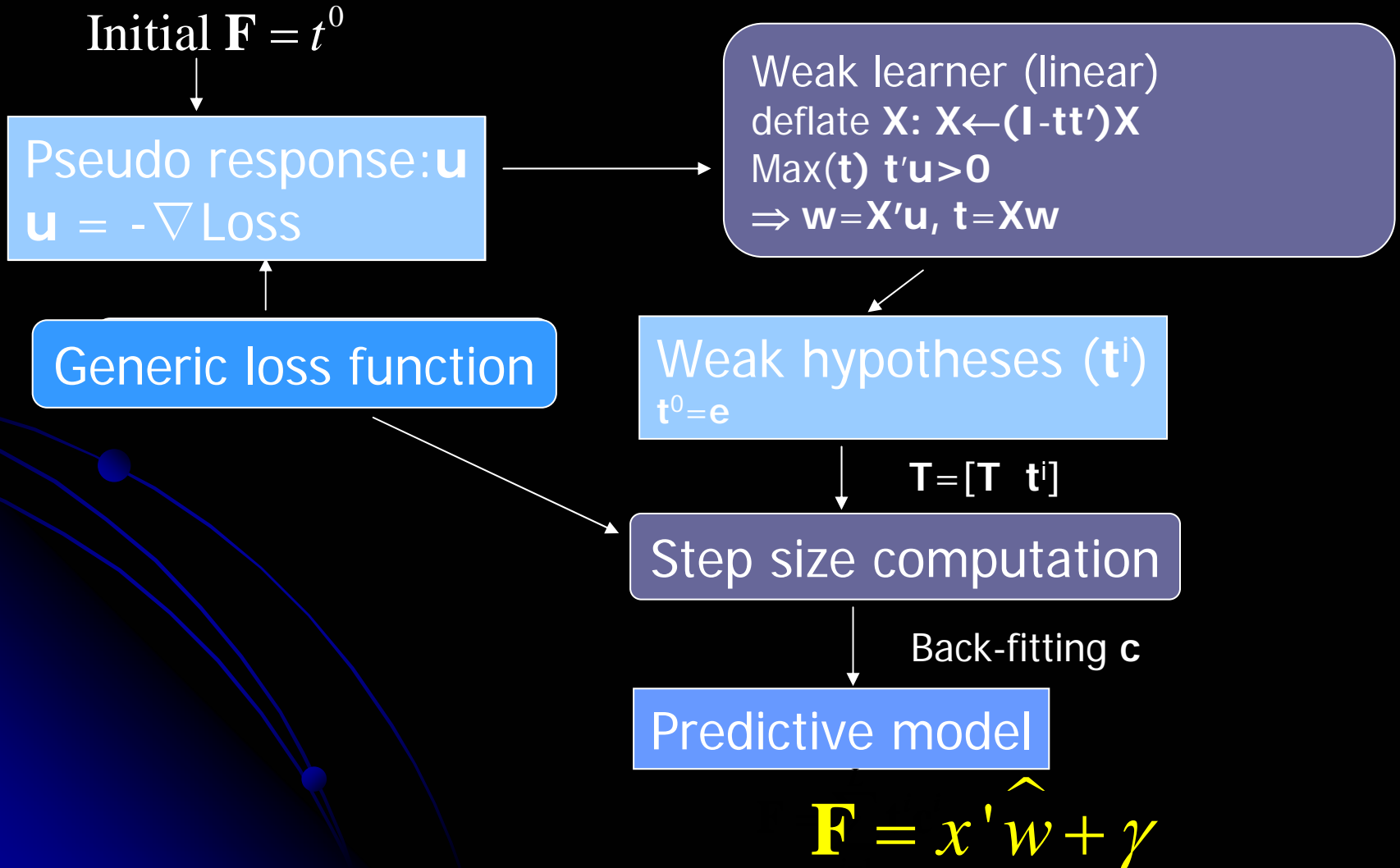
Stepsize c^i , or backfitting \mathbf{c}

Predictive model

$$\mathbf{F} = \sum_{i=1}^L \mathbf{t}^i \mathbf{c}^i + t^0$$

Subspace or conjugate gradient algorithm

Boosted Latent Analysis (Momma and Bennett 2004)



Feature Selection

Replace optimal $w^* = X'y$ with good sparse w

- Look at largest q components of w^*
- Evaluate cluster quality of q descriptors using gap statistic (Gene Shaving 2000):

difference between-to-within variance ratio of signed mean descr. for real and permuted data

- Let $w = w^*(i)$ for descriptors in best cluster
0 otherwise

Leukemia Microarray Data (Golub et al 1999)

Acute Myeloid Leukemia (AML) versus
Acute Lymphoblastic Leukemia (ALL)

7129 genes

Train: 27 AML + 11 ALL

Test: 20 AML + 14 ALL



PLS versus BLV (4 LV and 20 Bagged models)

- Linear PLS (Wold et al)

7032 descriptors

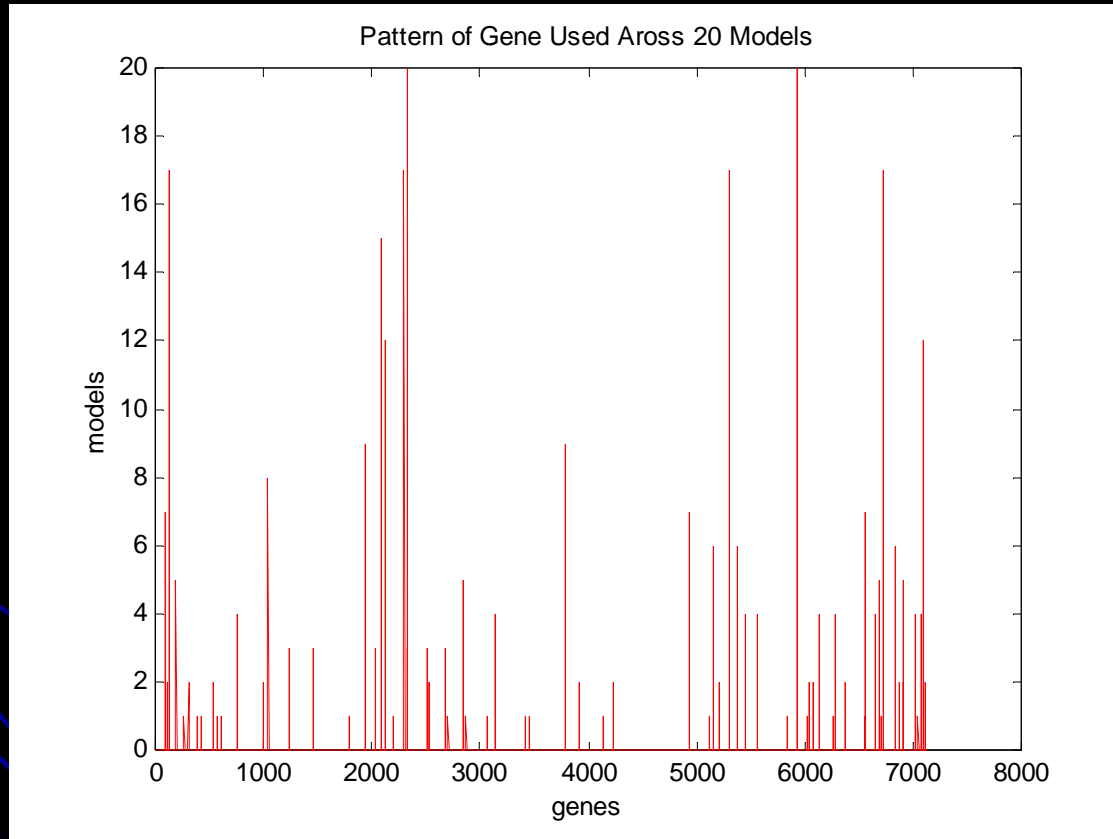
4 Errors on Leukemia

- Linear BLV with Least Squares

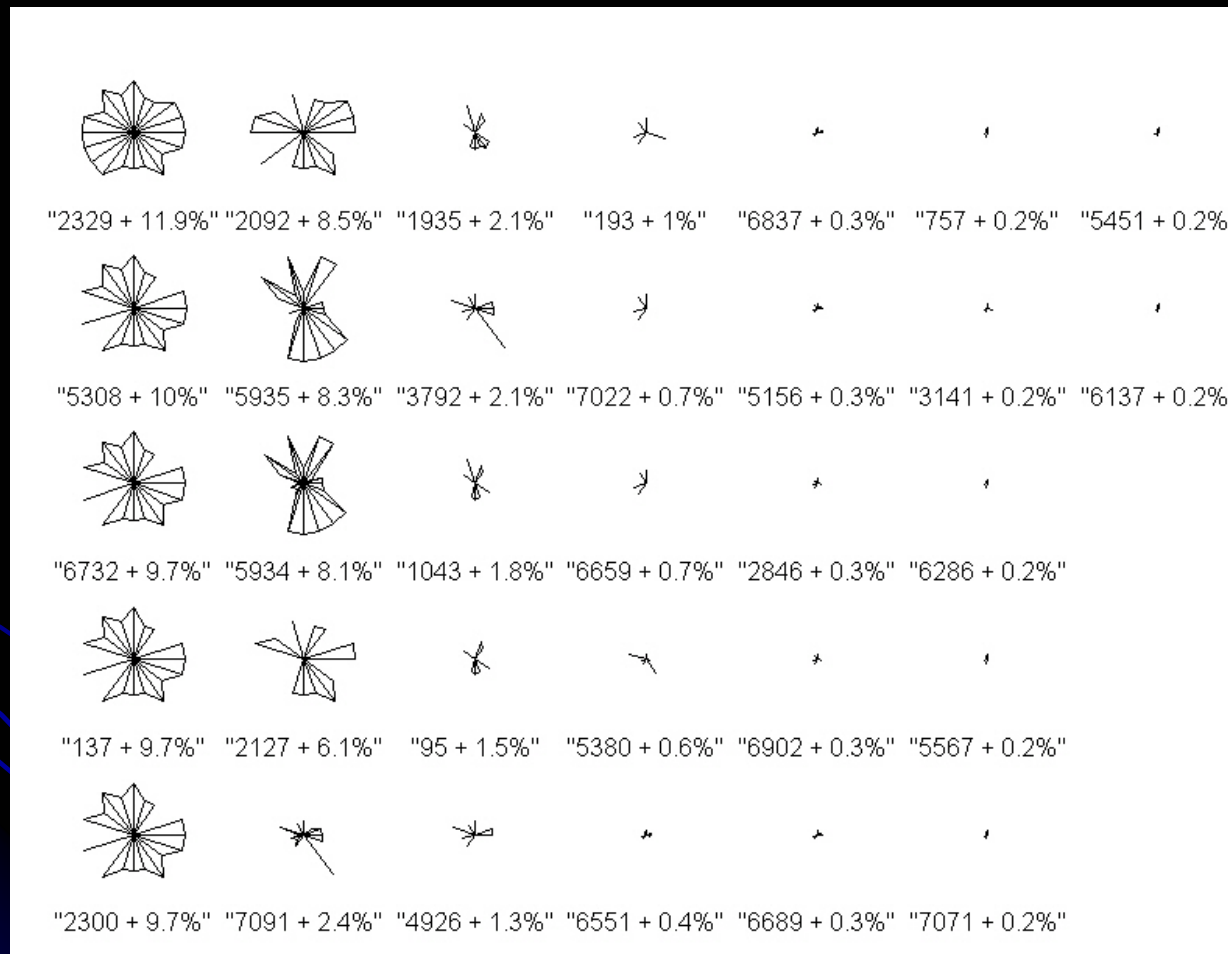
73 descriptors (8-38 per model)

1 Error on Leukemia

Genes Used in Models

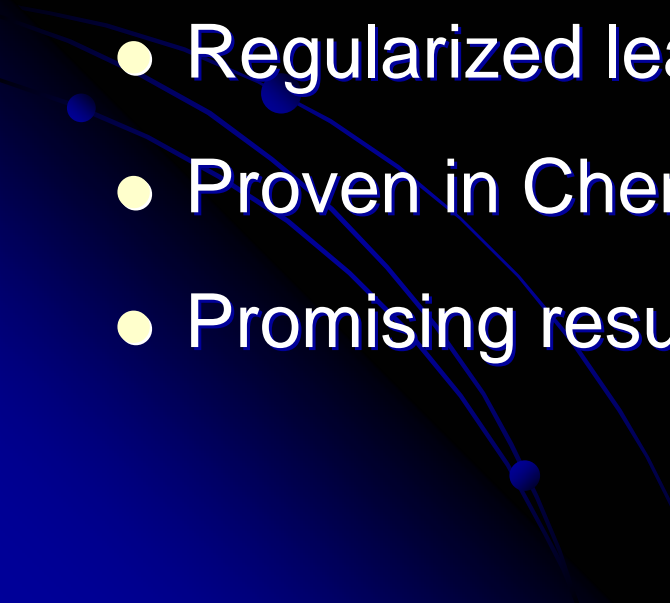


Understanding Bagged Model



32/72 Genes appeared in at least three models

Conclusions

- Robust methodology for many descriptors/few points (Analyze/StripMiner)
 - Bagged Feature Selection
 - Bagged Predictive Models
 - Regularized learning engines (SVM, PLS, BLV)
 - Proven in Cheminformatics
 - Promising results on Bioinformatics with BLV
- 

ACKNOWLEDGMENTS

- Members of the DDASSL group
 - Bennett Research Group (RPI Mathematics)
 - Jinbo Bi (Seimens)
 - Michi Momma (Fair Isaacs)
 - Angela Zhang
 - Breneman Research Group (RPI Chemistry)
 - N. Sukumar
 - M. Sundling
 - C. Whitehead (Pfizer)
 - L. Shen
 - L. Lockwood (Albany Molecular)
 - M. Song
 - D. Zhuang
 - W. Katt
 - Q. Luo
 - Embrechts Research Group (RPI DSES)
 - Collaborators:
 - Cramer Research Group (RPI Chemical Engineering)
 - **Funding**
 - NIH
 - NSF
- 