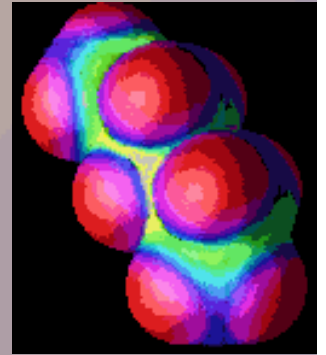


# Correlation Matrix Based Feature Selection with Genetic Algorithms for In-silico Drug Design



**Mark J. Embrechts (embrem@rpi.edu)**

Department of Decision Sciences and Engineering Systems

**Muhsin Ozdemir**

Department of Engineering Science

Rensselaer Polytechnic Institute, Troy, New York, 12180

*ASA-Albany*

*March 24, 2001*

**DDASSL**

Drug Design and Semi-Supervised Learning

**RENSSELAER**

# Outline

- Drug design and QSAR
- Feature selection
- Genetic Algorithms
- Correlation based feature selection
- Aggregating Models (Bagging)
- Conclusions



## MOTIVATION

- In USA \$25B/yr for R&D of pharmaceuticals (33% clinicals)
- 10-15 years from conception → market for drug
- Development cost 0.5B/drug
- First-year sales > \$1B/drug
- 1 drug approved/5000 compounds tested
- 1 out of 100 drugs succeeds to market
- 20,000,000 Americans with Alzheimer by 2050
- 19 Alzheimer's drugs and 9 Viagra-like drugs in development
- Worth their weight in gold



# Worth it's weight in GOLD

Weight				
	A	B	C	D
1	PRODUCT	PRICE	WEIGHT (lbs)	PRICE/lb
2				
3	PENTIUM III 800 MHz microprocessor	\$851.00	0.01984	\$42,893.00
4	Viagra (tablet)	\$8.00	0.00068	\$11,766.00
5	Gold (ounce)	\$301.70	0.0625	\$4,827.20
6	Hermes scarf	\$275.00	0.14	\$1,964.29
7	Palm V	\$449.00	0.26	\$1,726.92
8	Saving Private Ryan on DVD	\$34.99	0.04	\$874.75
9	Cigarettes (20)	\$4.00	0.04	\$100.00
10	Who Moved My Cheese? (Spencer Johnson)	\$19.00	0.49	\$40.80
11	Mercedes-Benz E-class four-door sedan	\$78,445.00	4,134.00	\$18.98
12	The Competitive Advantage of Nations	\$40.00	2.99	\$13.38
13	Chevrolet Cavalier four-door sedan	\$17,770.00	2,630.00	\$6.76
14	Hot-rolled steel (ton)	\$370.00	2,000.00	\$0.19
15				
16				
17	Source: Fortune Magazine March 20, 2000 page 68			
18				

# Drug Design and QSAR

- High-throughput screening
- Rational drug design
- Quantitative Structure Activity Relationship
  - is an attempt to correlate structural or property descriptors of compounds with their activities.

$$\textit{Activity} = f(\textit{Structure})$$



# Why to select features?

Irrelevant and/or redundant features:

- increase the dimension of the problem  
hence require greater computational cost
- lead to overfitting
- make the model more difficult to explain

# Approaches to Feature Selection

- Filter method
  - attempts to access the merits of features from the data alone
- Wrapper method
  - searches for an optimal feature subset tailored to a particular induction algorithm and uses the induction algorithm as a black box for evaluating feature subsets

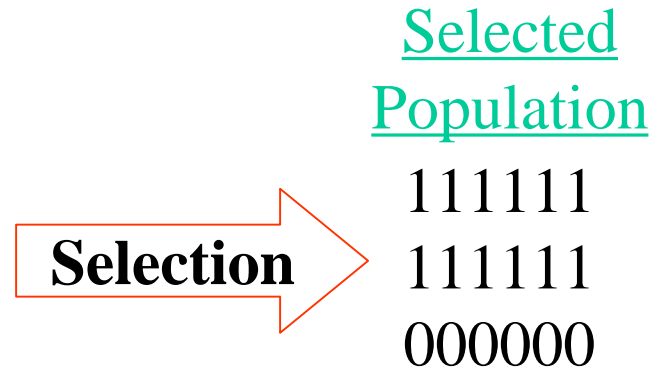
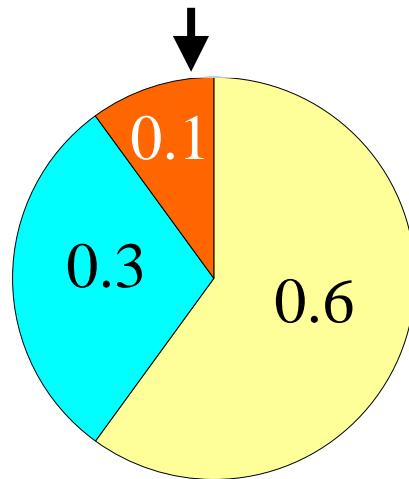
# Genetic Algorithms

- developed by John Holland (1970's),
- general optimization methods,
- work with encoding of the parameters,
- search by means of a population of potential solutions,
- use an evaluation (fitness) function,
- search probabilistically.

# Components of a GA

**Problem:**  $\max f(x)$

<u>Population</u>	<u>Fitness</u>
111111	$f_1 = 60$
110000	$f_2 = 30$
000000	$f_3 = 10$



Crossover point

111111  
000000



111100  
000011

Selected gene  
000000



Mutated gene  
000001

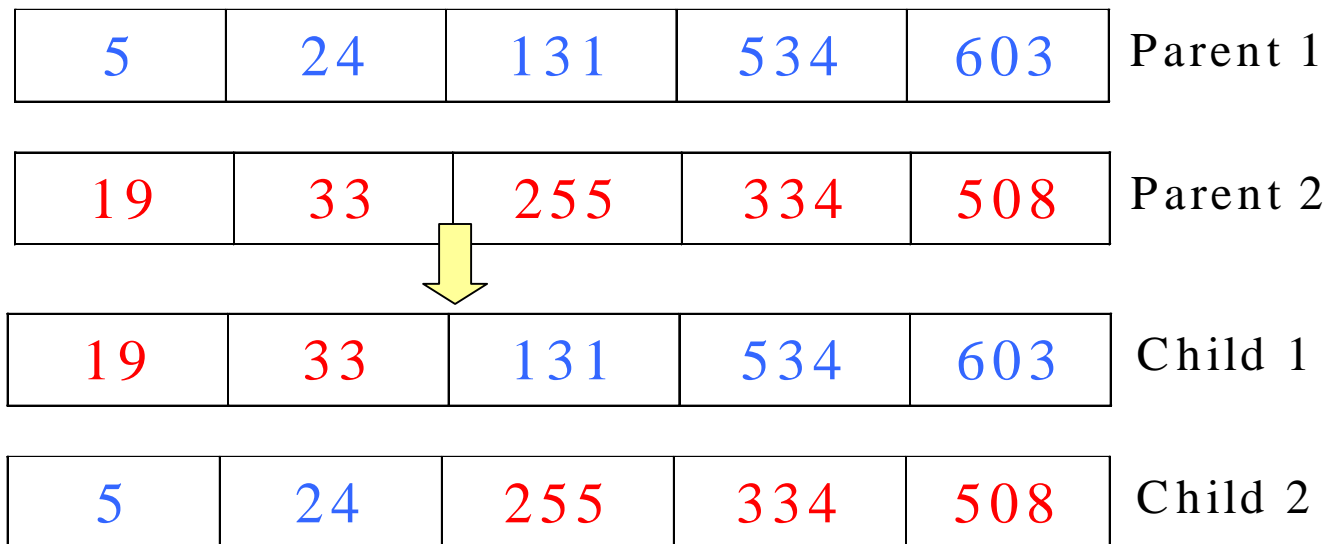
# Genetic Algorithms

```
procedure genetic algorithm
begin
    Choose a coding to represent variables
     $t \leftarrow 0$ 
    Initialize population  $P(t)$ 
    Evaluate population  $P(t)$ 

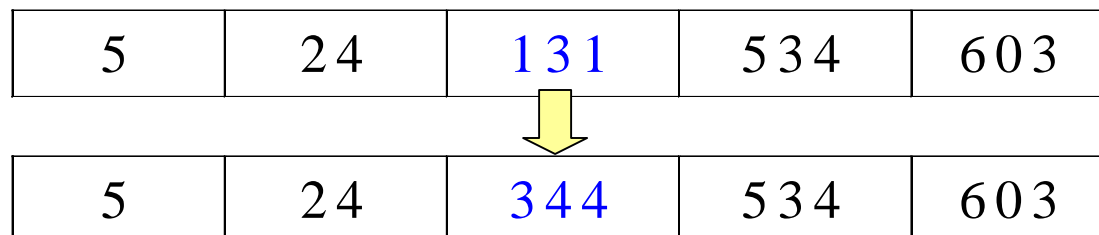
    while (not termination condition) do
         $t \leftarrow t+1$ 
        Select  $P(t)$  from  $P(t-1)$ 
        Alter  $P(t)$  with crossover and mutation
        Evaluate  $P(t)$ 
    end
end
```

# Feature selection with GA

- Uniform Crossover



- Mutation



# Feature selection with GA *continued...*

- Prespecify number of features ( $N$ )
- Evaluation function

$$F_k = \sum_{i=1}^N C_{iR} - \sum_{i=1, i \neq j}^N \alpha C_{ij} - \beta$$

where

$F_k$  = fitness       $k = 1, 2, 3, \dots, Pop\_size$

$C_{ij}$  = intercorrelation       $i = j = \{g_{k1}, g_{k2}, g_{k3}, \dots, g_{kj}\}$

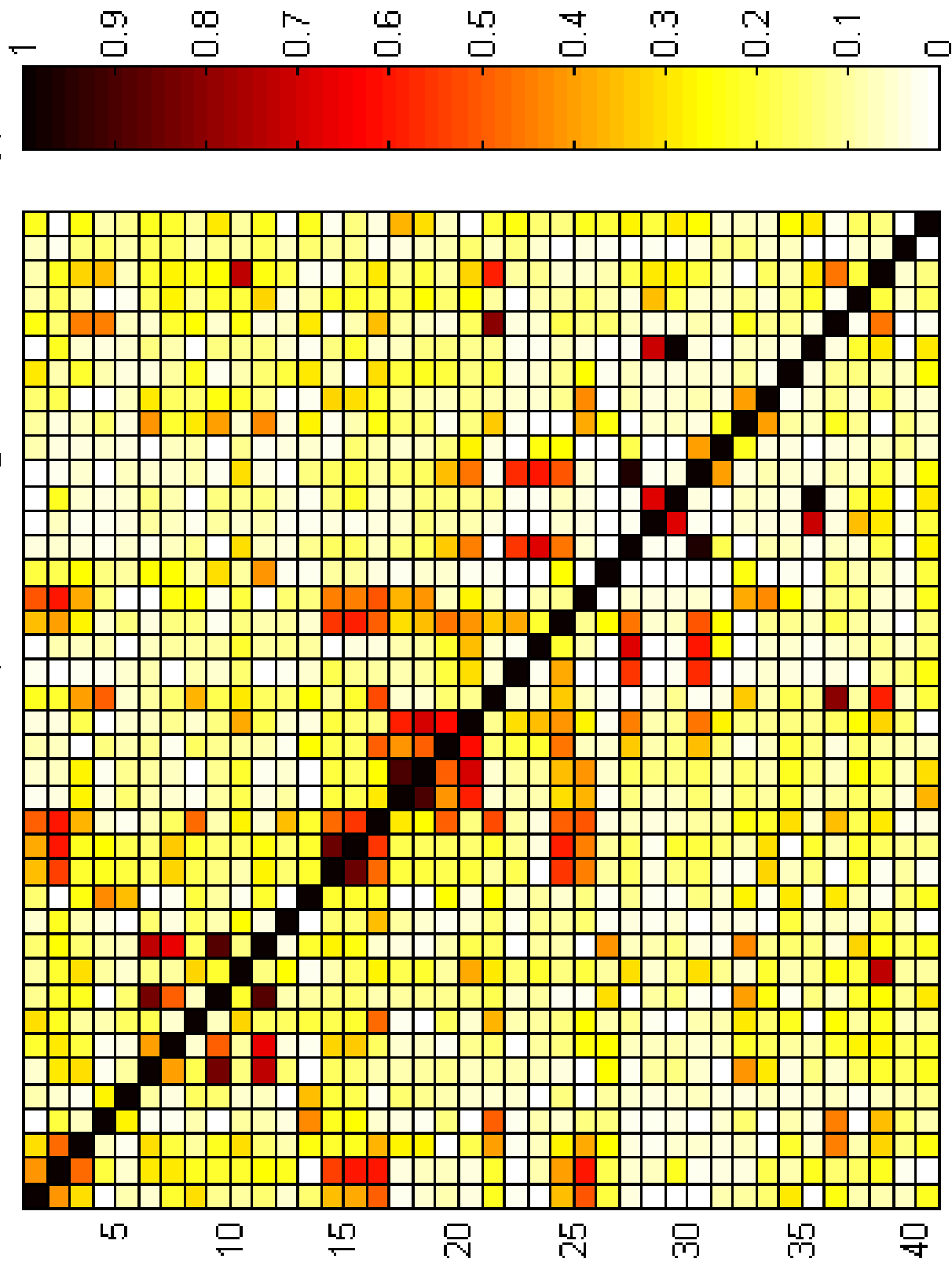
$C_{iR}$  = correlation with the response

$g_{11}$  = gene position 1 in the individual 1

$\alpha$  = Intercorrelation penalty factor

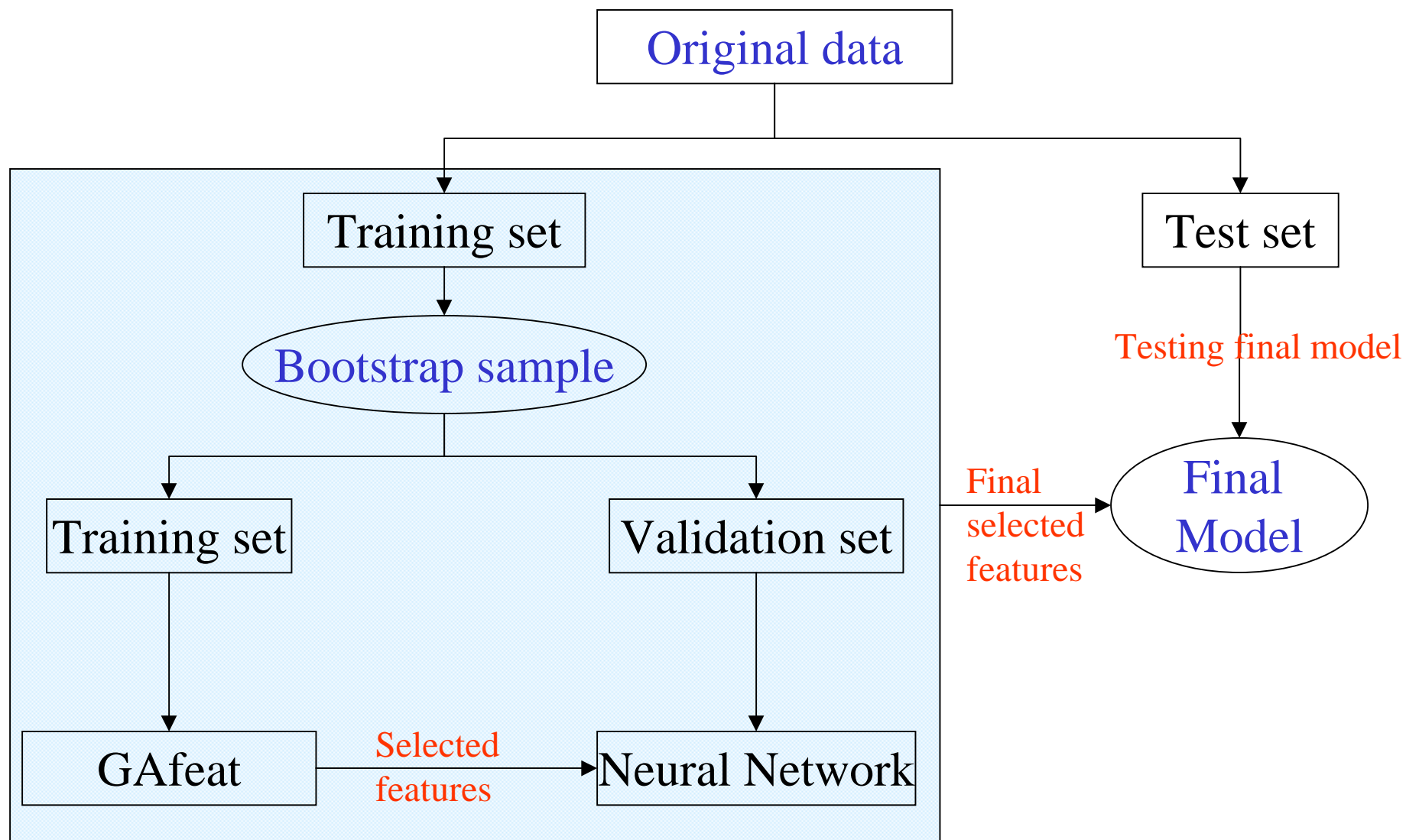
$\beta$  = Death penalty factor. If intercorrelation  $> 0.95 = 1000$   
otherwise 0

# CORRELATION MATRIX (38 descriptors + dummy)

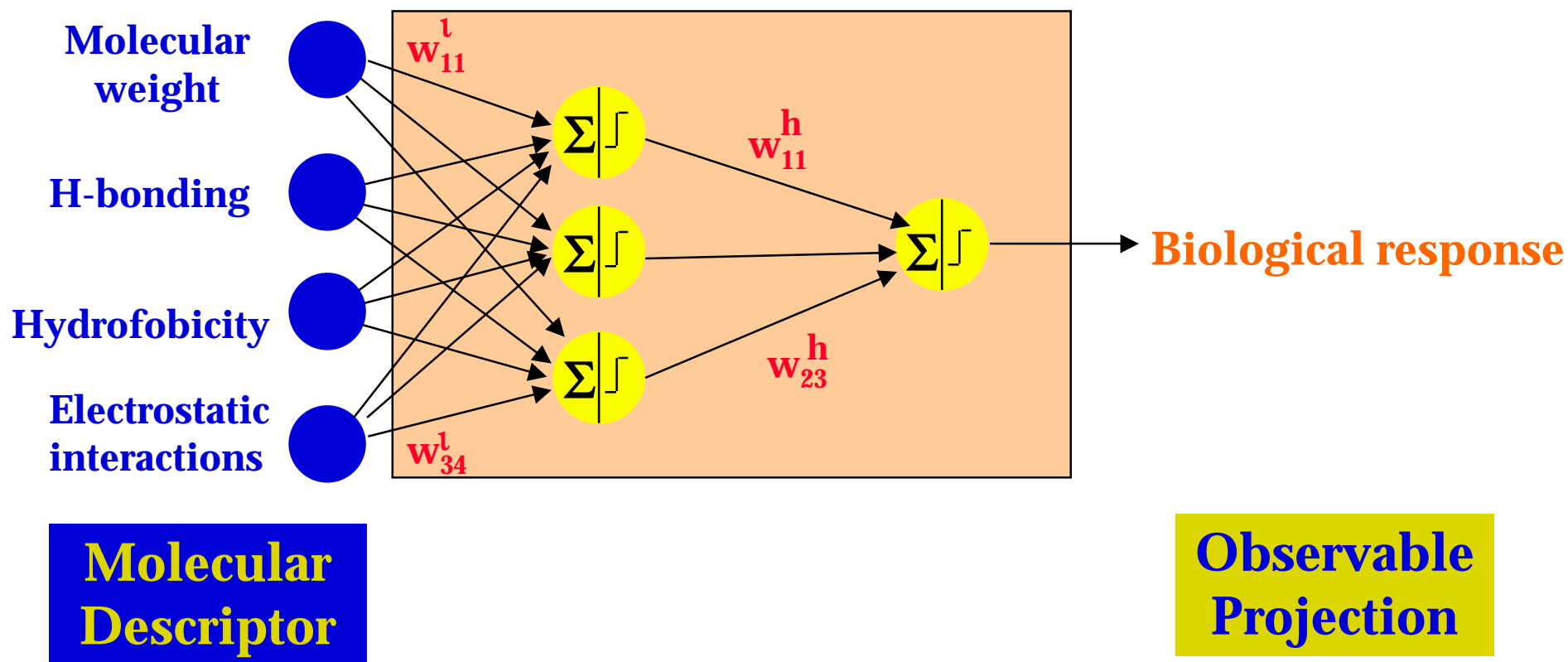


DESCRIPTOR

# Bootstrapping



# Artificial Neural Network



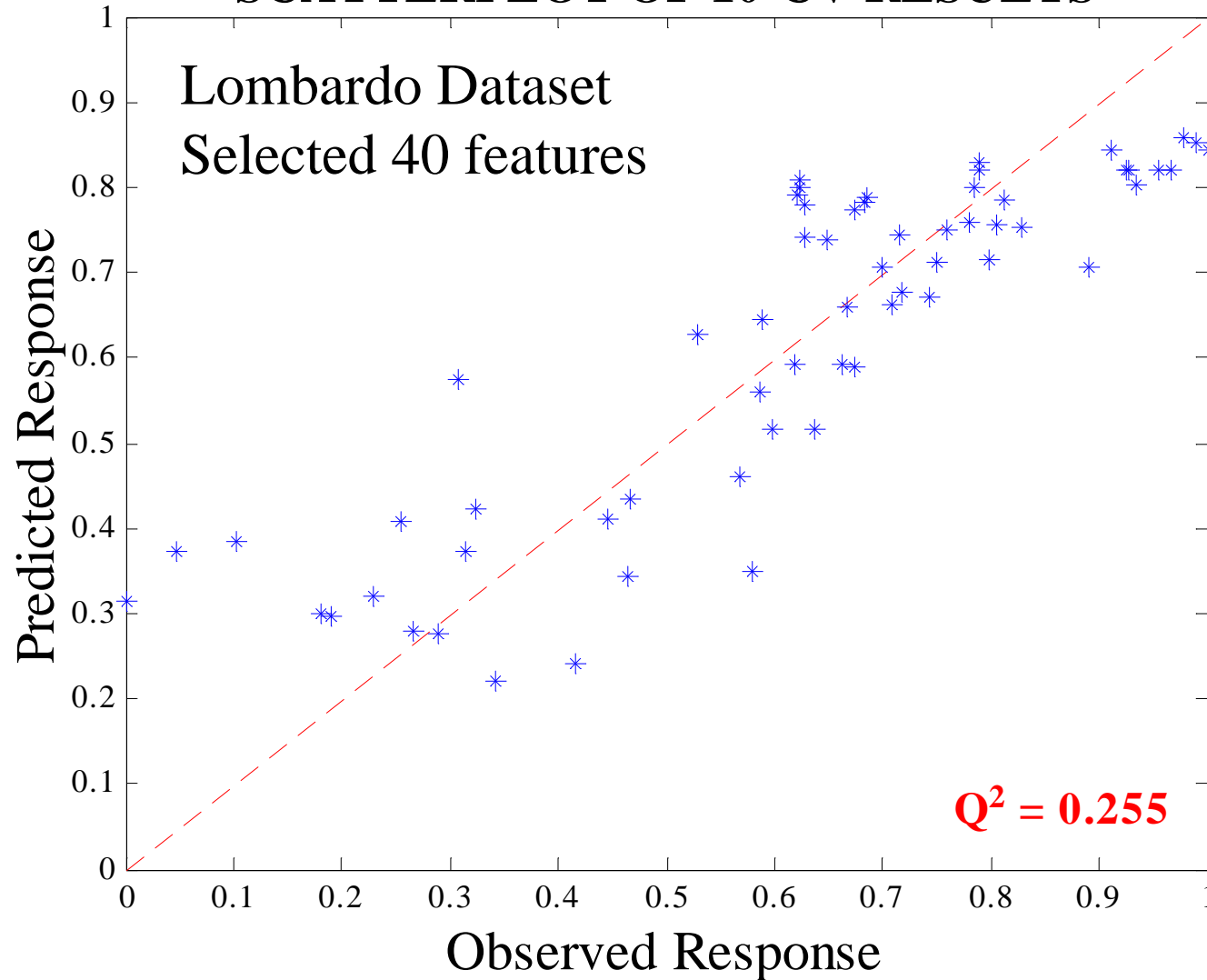
**DDASSL**

Drug Design and Semi-Supervised Learning

RENSSELAER

# Results

## SCATTERPLOT OF 10-CV RESULTS



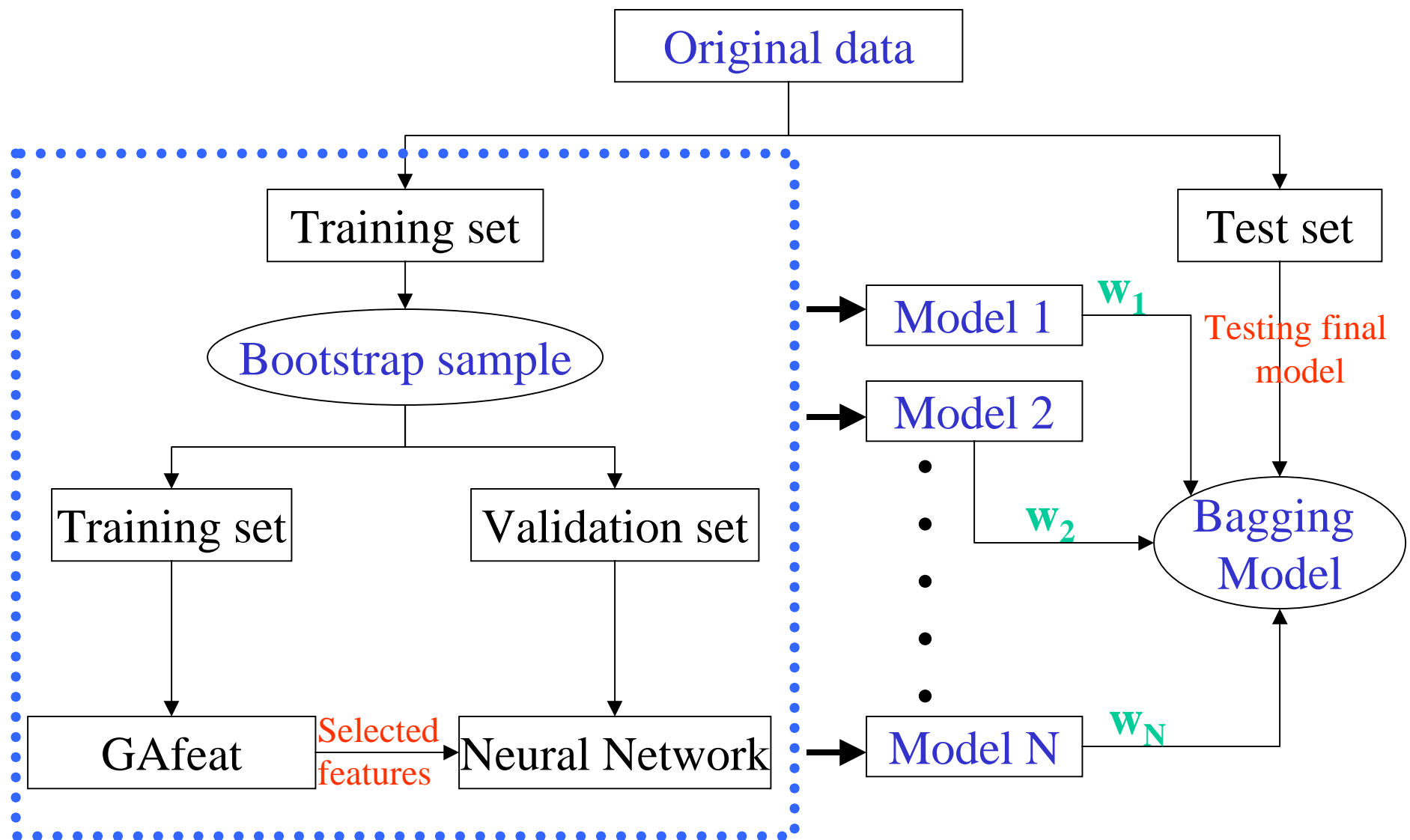
$$R^2 = 1 - \frac{\sum_{i=1}^N (Y_O - Y_P)^2}{\sum_{i=1}^N (Y_O - \bar{Y}_O)^2}$$

$$Q^2 = 1 - R^2$$

$Y_O$  = observed

$Y_P$  = predicted

# Bootstrap Aggregating (Bagging)



# Conclusions

- Selects relevant features
- reduces dimension of the problem
- builds predictive model
- more robust predictive model with bagging