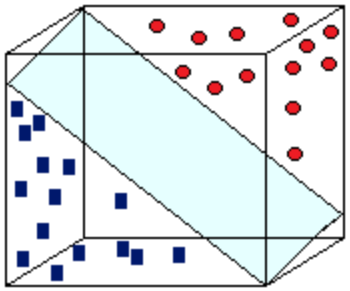
Detecting interactions of genetic risk factors in disease



Patricia Francis-Lyon
Shashank S. Belvadi
Lin Wang
University of San Francisco

Outline

- What are gene interactions and why do we care ?
- Methods
- Results

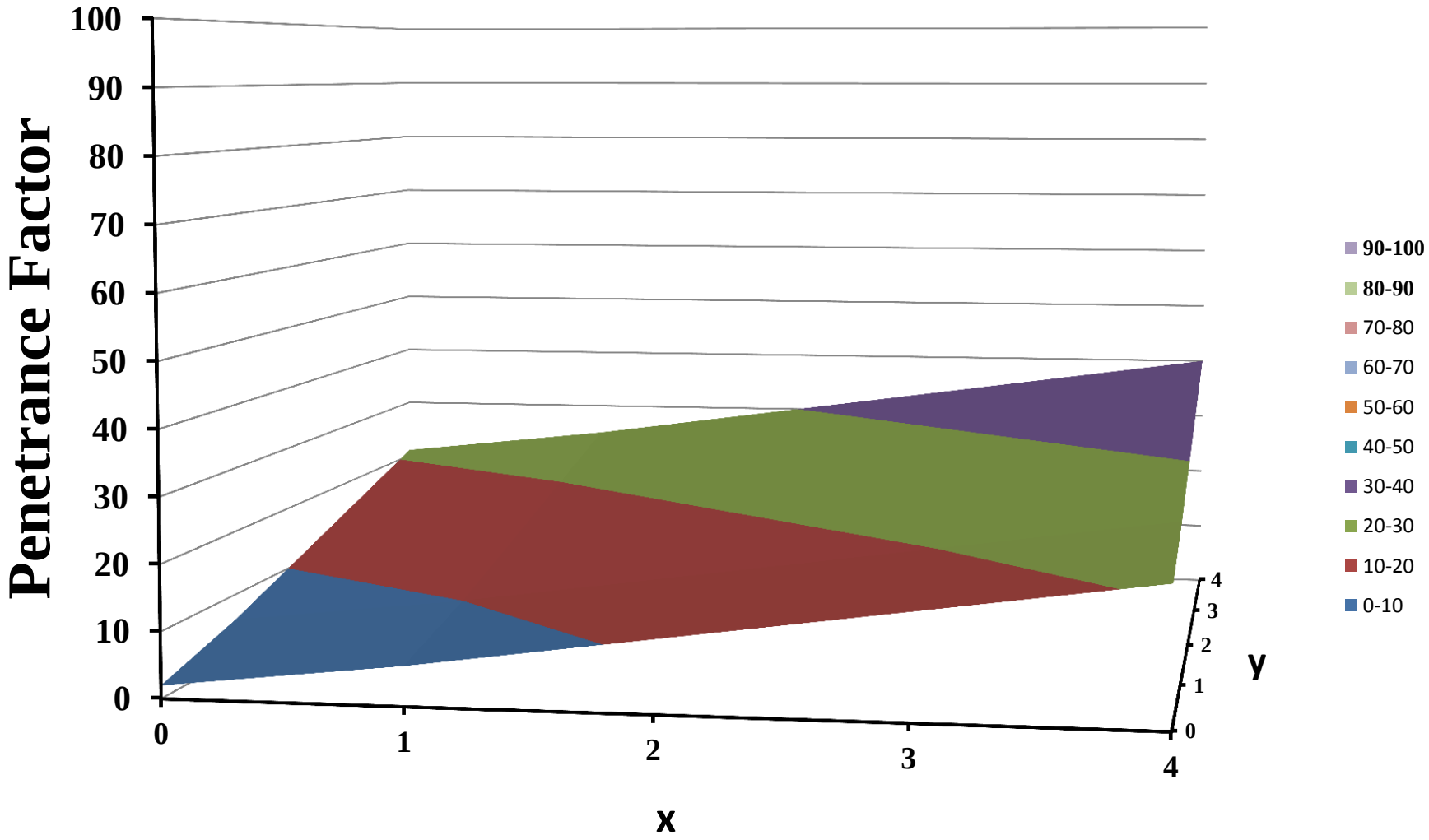
Gene interactions

- Genes can interact to either enhance or suppress penetrance (probability of getting disease): aid early detection
- Genes code for proteins: find protein paths, inform therapies
- Ex: search genes that are involved in estrogen metabolism for interactions in ER+ breast cancer- find new pathways to target

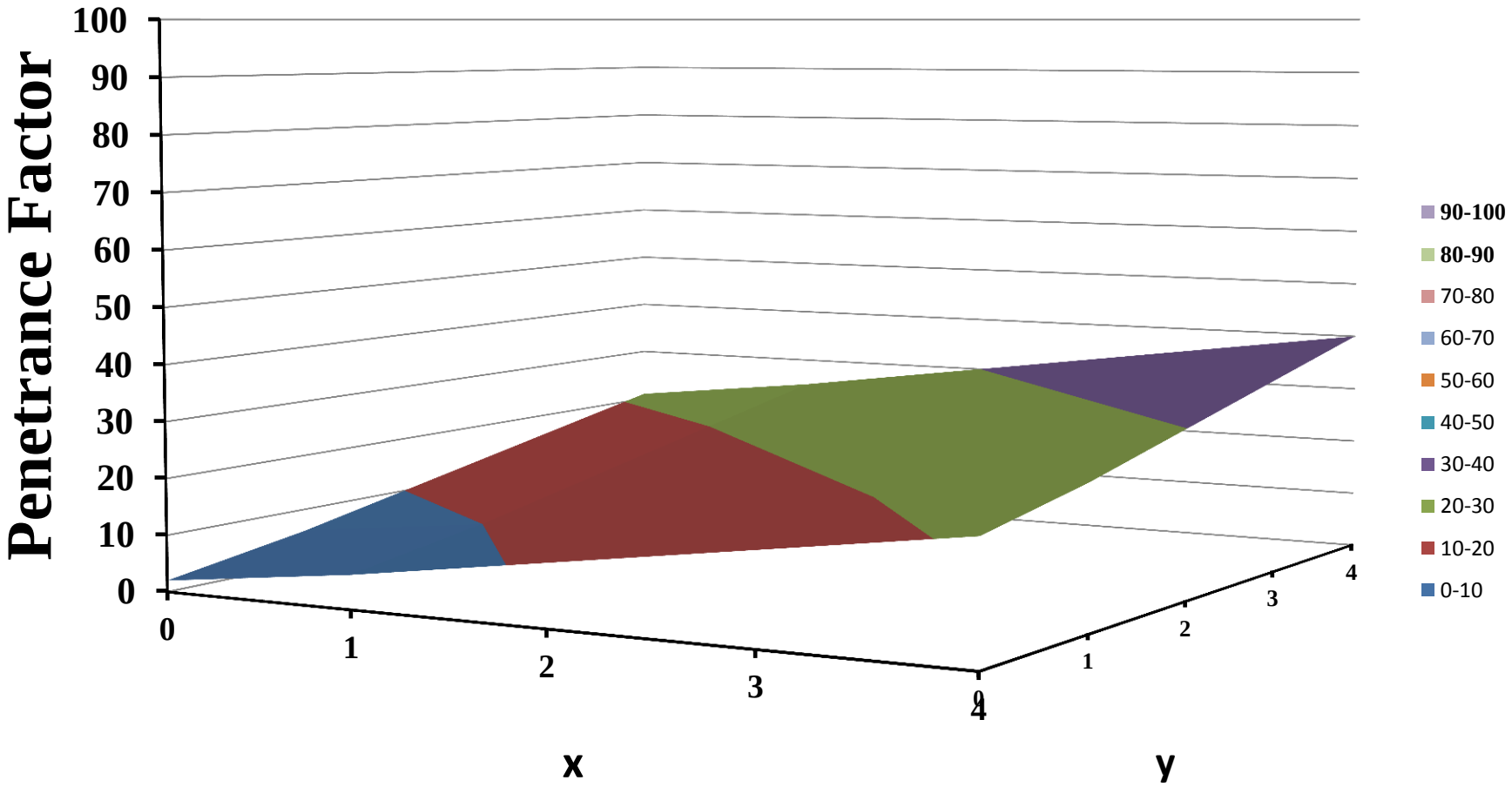
Define interactions

- Biological interaction difficult to quantify
- Use a definition from statistics: interaction is departure from a linear model
- Intuition: plot penetrance (dependent variable) on vertical axis, independent variables on horizontal axes.

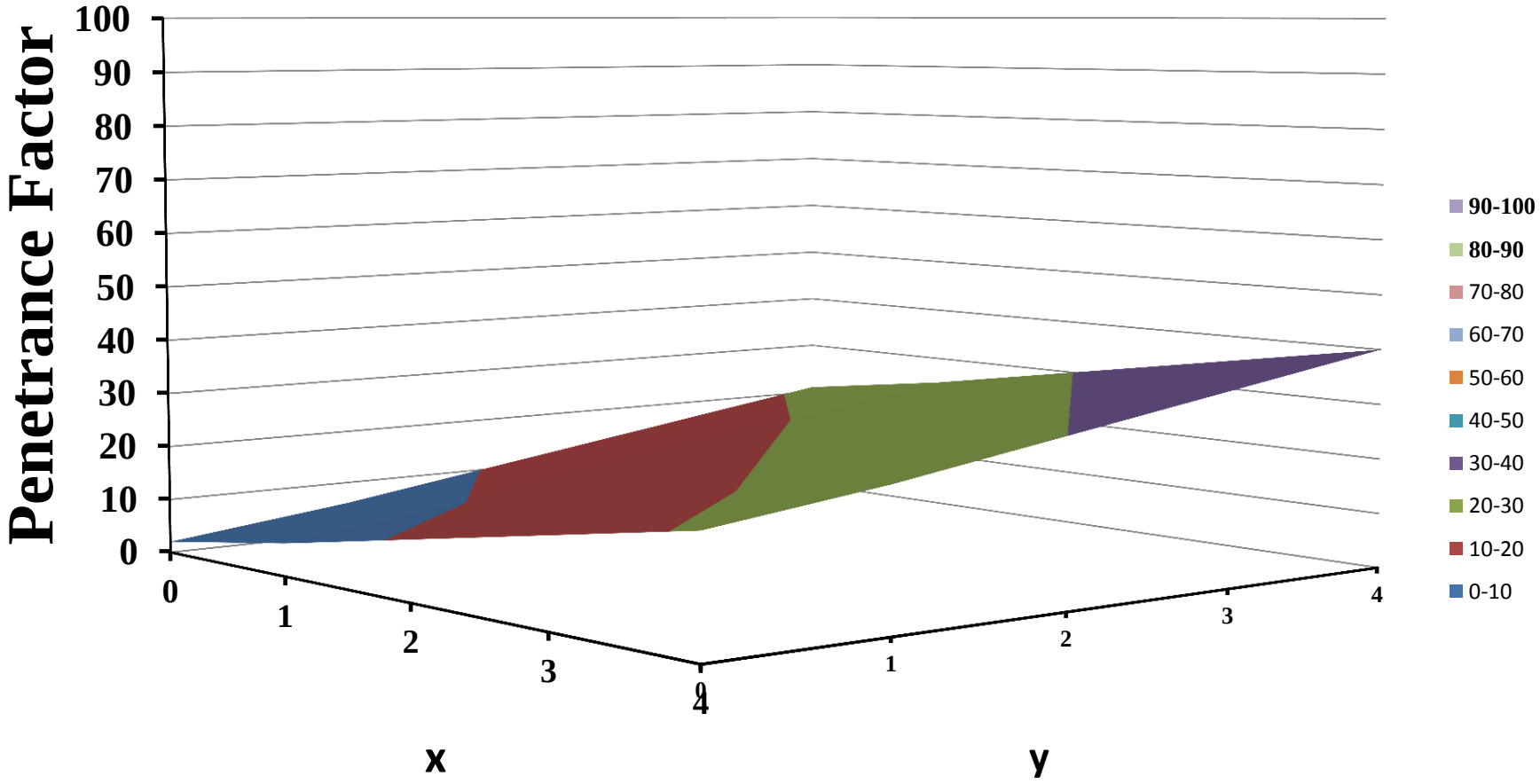
Are x,y interacting to affect penetrance?



Rotate horizontally 20°

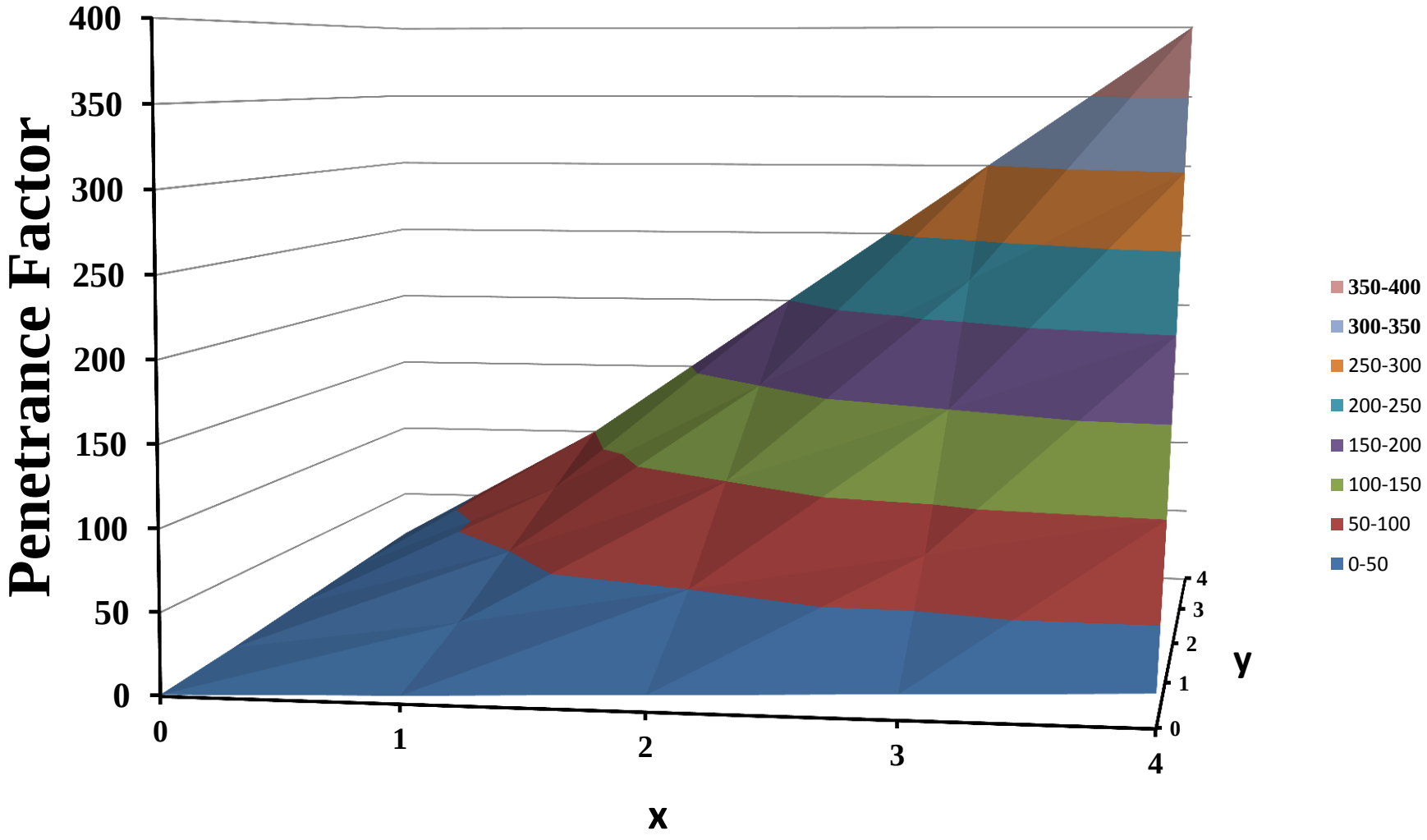


20° more: we see this is linear: x and y do not interact

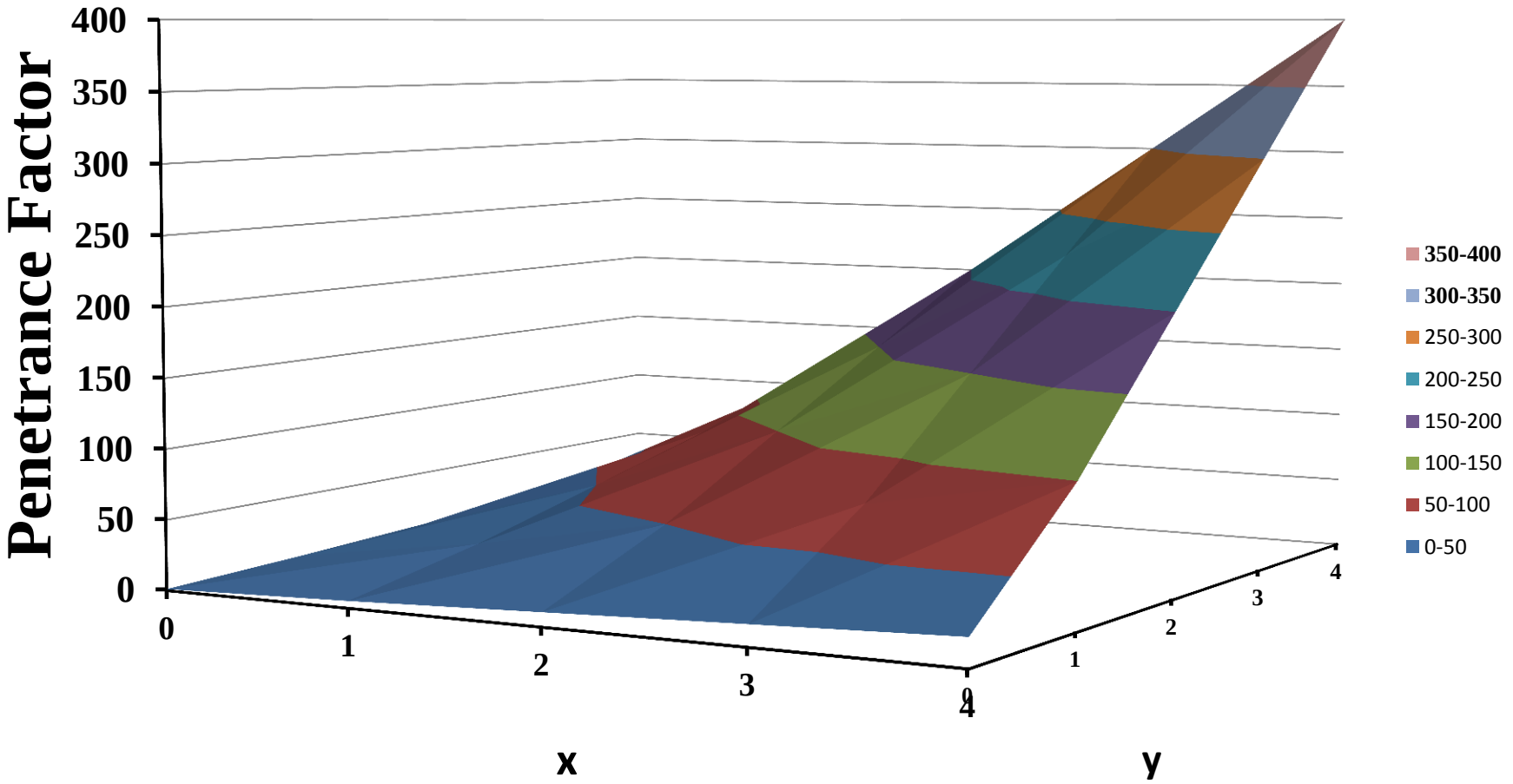


Based on Risch ADD model

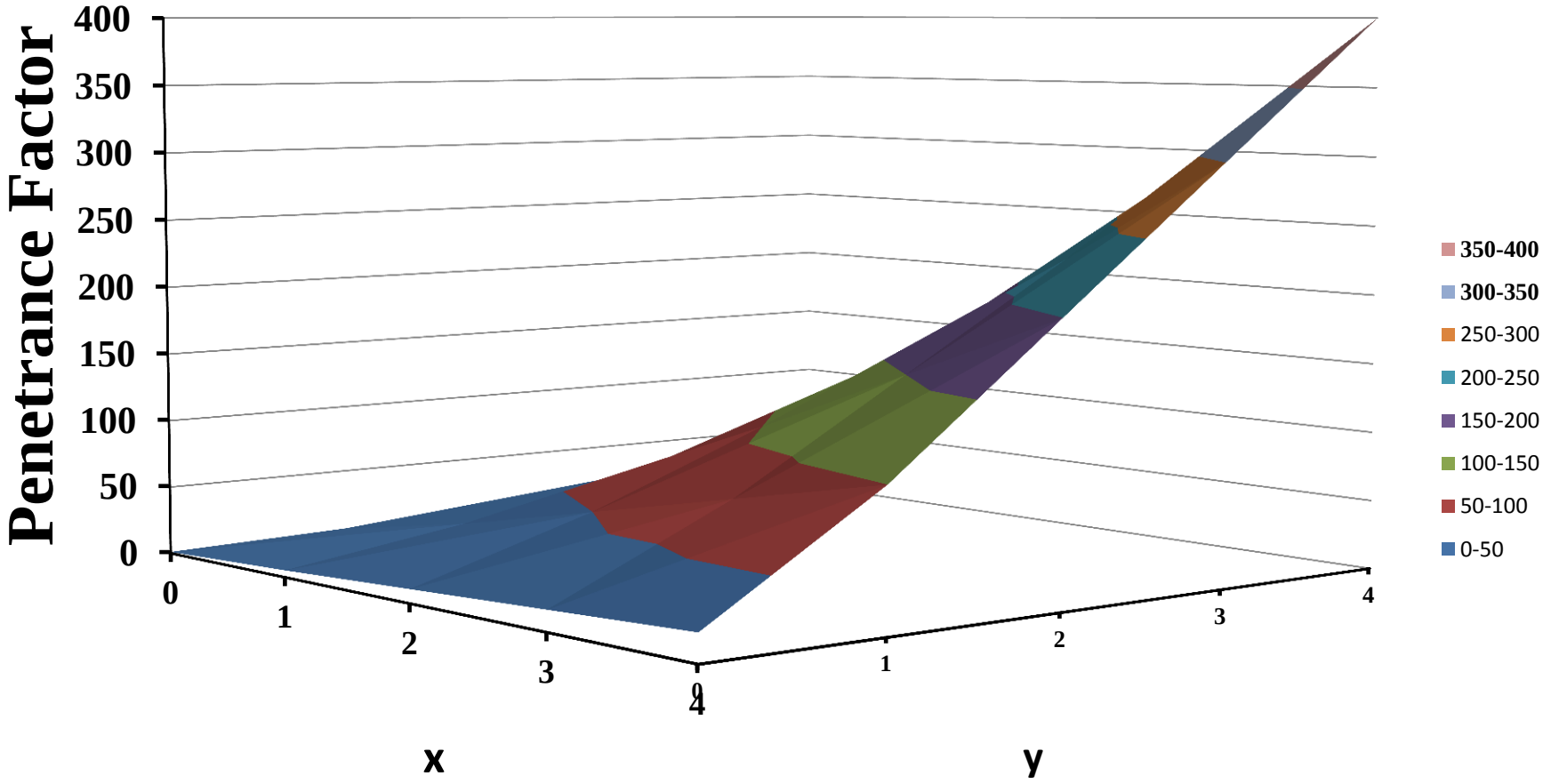
Are x,y interacting to affect penetrance?



Rotate horizontally 20°



20° more: we see this is NOT linear: x and y interact



Based on Risch MULT model

Methods

- 7 Disease models
- Simulate populations for each
- Black box machine learning (ML)
- Open black box: gene perturbation and metric

Simulate populations

- We used genomeSIMLA (Ritchie lab)
- Simulate a population for each disease model
- 10 genes: 2 are functional for 2-way interactions, 3 functional for 3-way interactions. (Noisy).
- Encode genes as number of mutated alleles: 0,1,or 2

Disease models (Risch)

Each element f_{ij} of penetrance matrix f is specified by:

$$f_{ij} = P(Y = 1 | G_\alpha = i, G_\beta = j)\ i, j \in \{0, 1, 2\}.$$

1. Additivity model (biological independence):
   $f_{ij} = a_i + b_j$ such that $0 <= a_i, b_j <= 1, a_i + b_j < 1$

2. Heterogeneity model (biological independence):
   $f_{ij} = a_i + b_j - a_i b_j$ such that $0 <= a_i, b_j <= 1$

3. Multiplicative model (biological interaction):
   $f_{ij} = a_i b_j$

Epigenetic disease models (Gunther)

$$f = \begin{array}{c} \\ AA \\ Aa \\ aa \end{array} \begin{array}{c} \begin{array}{ccc} BB & Bb & bb \end{array} \\ \begin{pmatrix} c & c & c \\ c & c & c \\ c & c & rc \end{pmatrix} \end{array}$$

$$f = \begin{array}{c} \\ AA \\ Aa \\ aa \end{array} \begin{array}{c} \begin{array}{ccc} BB & Bb & bb \end{array} \\ \begin{pmatrix} c & c & c \\ c & r_1 c & r_1 c \\ c & r_1 c & r_2 c \end{pmatrix} \end{array}$$

EPIRR:

Recessive interactions

Both alleles of both genes must be mutated to have an effect

EPIDD:

Dominant interactions

At least one allele of each gene must be mutated to have an effect

Epigenetic (Gunther, Ritchie)

$$f = \begin{array}{c} \\ AA \\ Aa \\ aa \end{array} \begin{array}{c} \begin{array}{ccc} BB & Bb & bb \end{array} \\ \begin{pmatrix} c & c & c \\ c & c & c \\ r_1 c & r_1 c & r_2 c \end{pmatrix} \end{array}$$

EPIRD:
Mixed interactions

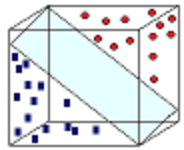
$$f = \begin{array}{c} \\ AA \\ Aa \\ aa \end{array} \begin{array}{c} \begin{array}{ccc} BB & Bb & bb \end{array} \\ \begin{pmatrix} 0 & 0 & .2 \\ 0 & .2 & 0 \\ .2 & 0 & 0 \end{pmatrix} \end{array}$$

MDR:
XOR interactions

Machine Learning

- Our approach intended for black box ML methods
- Supervised learning: train a machine to predict disease, then use gene perturbation and metric to find interactions it learned.
- Black box methods we used:
  - Support vector machine (SVM)
  - Neural Network (NN)

Our approach

1. ***Train*:** training data 

What did the trained model learn?

Interactions gene 2, gene 7?

2. *Predict* test set 4 times, for 3 of them remove some information

- run original test set:

  1 1 1 0 1 2 2 2 1 1

  ...

  ...

- remove gene 2 mutations:

  1 0 1 0 1 2 2 2 1 1

  ...

  ...

- remove gene 7 mutations:

  1 1 1 0 1 2 0 2 1 1

  ...

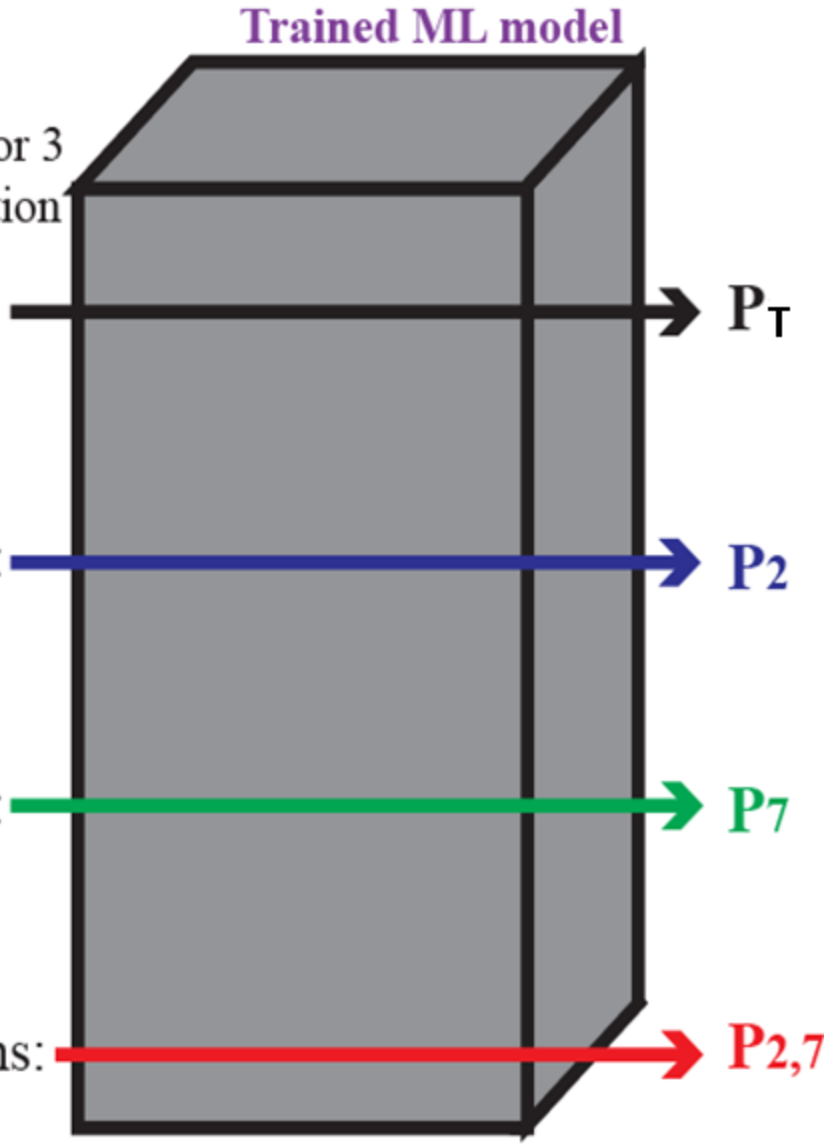
  ...

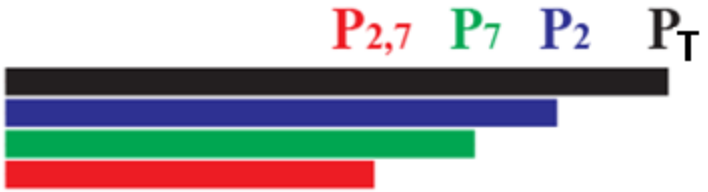
- remove gene 2,7 mutations:

  1 0 1 0 1 2 0 2 1 1

  ...

  ...





3. ***Apply Metric*:**

If $(P_T - P_{2,7}) = (P_T - P_2) + (P_T - P_7)$

then additive model, no interactions.

$$m = |(P_T - P_{2,7}) - ((P_T - P_2) + (P_T - P_7))|$$

Metric *m* quantifies the difference between lhs and rhs deviations in prediction accuracy. Allow 2.75 error for each deviation: 3 deviations, so if difference < 7.5 then no interactions

Metric for 3-way interactions

$$m = |(P_T - P_{abc}) - ((P_T - P_a) + (P_T - P_b) + (P_T - P_c))|$$

For 3-way there are 4 deviations in the above metric. Allow 2.75 error for each deviation: so cutoff = 10 .

Here prediction accuracy is % of test cases whose predictions match target values.

Results in detecting interactions

- 2-way :
  - 2 independent models: all 45 pairs designated “no interactions”
  - 5 models of interaction: correct pair found, other 44 pairs designated “no interactions”
- 3-way :
  - 2 independent models: all 120 pairs designated “no interactions”
  - 5 models of interaction: correct pair found, other 119 pairs designated “no interactions”
- Method is tolerant of noise

Characterizing interactions

- Apply mask and perturb masked genes to unmutated
- Resulting drop in prediction accuracy is attributable to that kind of interaction
- AND mask for characterizing interactions of genes that each have at least one mutated allele
- XOR mask for characterizing interactions of genes that where one gene is unmutated and the other(s) have at least one mutated allele

Perturb masked genes to unmutated

	BB	Bb	bb
AA			
Aa			
aa			

	BB	Bb	bb
AA	c	c	c
Aa	c	r_1c	r_1c
aa	c	r_1c	r_2c

masks:
AND: red
XOR: purple

EPIDD:
Only AND, no XOR

SVM

- Software: scikit-learn
- Support vector classifier (SVC) with radial basis kernel
- Used cross validation grid search to select best parameters for C (regularization) and gamma
- C values: 100 to 10000 to penalize for errors, gamma variation used to explore the reach of SVs
- Train model with best parameters from CV

SVM: 2-way interactions

Disease Model	Metric	Interactions Found	Actual	Found AND	Found XOR	Actual AND	Actual XOR
ADD	6.9	none	none	N/A	N/A	N/A	N/A
MULT	19.2	4, 9	4, 9	yes	no	yes	no
HET	5.0	none	none	N/A	N/A	N/A	N/A
EPIRR	41.1	4, 9	4, 9	yes	no	yes	no
EPIDD	14.5	4, 9	4, 9	yes	no	yes	no
EPIRD	10.2	4, 9	4, 9	yes	no	yes	no
MDR	47.7	4, 9	4, 9	yes	yes	yes	yes

SVM: 3-way interactions

Disease Model	Metric	Interactions Found	Actual	Found AND	XOR	Actual AND	XOR
ADD	5.5	none	none	N/A	N/A	N/A	N/A
MULT	23.2	0,4,9	0,4,9	yes	no	yes	no
HET	4.7	none	none	N/A	N/A	N/A	N/A
EPIRRR	53.0	0,4,9	0,4,9	yes	no	yes	no
EPIDDD	23.8	0,4,9	0,4,9	yes	no	yes	no
EPIRRD	16.2	0,4,9	0,4,9	yes	no	yes	no
MDR	78.5	0,4,9	0,4,9	yes	yes	yes	yes

NN

- Software: Pybrain
- Feed forward back propagation NN
- Constructed models with 2,3,4,5 hidden nodes
- Learning epochs for each model: 300, 400, 500
- Employed learning rate decay for the first 10 epochs to enhance finding the global min
- Selected model with best prediction of test set

NN: 2-way interactions

Disease Model	Metric	Interactions Found	Actual	Found AND	XOR	Actual AND	XOR
ADD	2.5	none	none	N/A	N/A	N/A	N/A
MULT	9.2	4, 9	4, 9	yes	no	yes	no
HET	3.0	none	none	N/A	N/A	N/A	N/A
EPIRR	25.9	4, 9	4, 9	yes	no	yes	no
EPIDD	7.6	4, 9	4, 9	yes	no	yes	no
EPIRD	8.5	4, 9	4, 9	yes	no	yes	no
MDR	39.0	4, 9	4, 9	yes	yes	yes	yes

NN: 3-way interactions

Disease Model	Metric	Interactions Found	Actual	Found AND	XOR	Actual AND	XOR
ADD	3.4	none	none	N/A	N/A	N/A	N/A
MULT	22.6	0,4,9	0,4,9	yes	no	yes	no
HET	3.7	none	none	N/A	N/A	N/A	N/A
EPIRRR	54.2	0,4,9	0,4,9	yes	no	yes	no
EPIDDD	28.2	0,4,9	0,4,9	yes	no	yes	no
EPIRRD	12.8	0,4,9	0,4,9	yes	no	yes	no
MDR	79.4	0,4,9	0,4,9	yes	yes	yes	yes

Patterns to look for
(aside from the metric)

- Voting: if interactions are there, look for them in many models

- For 3-way interactions, look for triplets that contain a subset of the interacting pairs

Voting

Typical: 42 of the 48 models found the correct 3-way interactions, found 2-way also

Model	Ptrain	Ptotal	**Loci1**	Metric	**Loci2**	Metric	**Loci3**	Metric
EPIDDD3Way_NN_H2E400	**66.2**	**64.3**	**[0 4 9]**	**28.2**	**[0 7 9]**	**14.9**	**[0 5 9]**	**14.7**
EPIDDD3Way_NN_H2E300	65.6	63.3	**[0 4 9]**	20.3	**[0 7 9]**	11.6	**[0 1 9]**	11.3
EPIDDD3Way_NN_H5E500	69.3	63.2	**[0 4 9]**	21.5	**[0 2 9]**	13	**[0 7 9]**	12.9
EPIDDD3Way_NN_H5E400	67.1	62.8	**[0 4 9]**	20.1	**[0 4 7]**	10.9	**[0 2 4]**	10.6
EPIDDD3Way_NN_H2E300	65.5	62.3	**[0 4 9]**	13.6	**[0 1 9]**	8	**[0 6 9]**	7.8
EPIDDD3Way_NN_H5E300	69.5	62.3	**[0 4 9]**	21.3	**[0 2 9]**	12	**[0 2 4]**	11.7
EPIDDD3Way_NN_H3E500	65.7	62.2	**[0 4 9]**	13	**[0 3 9]**	10.6	**[0 7 9]**	10.6
EPIDDD3Way_NN_H5E500	69.3	62	**[0 4 9]**	21.1	**[0 1 9]**	12.1	**[0 7 9]**	12.1
EPIDDD3Way_NN_H7E500	69.8	62	**[0 4 9]**	19.5	**[0 4 6]**	12.1	**[0 4 5]**	11.1
EPIDDD3Way_NN_H5E300	67.3	61.9	**[0 4 9]**	23.8	**[0 5 9]**	13.1	**[0 4 5]**	13.1
EPIDDD3Way_NN_H2E500	65.6	61.9	**[0 4 9]**	17.3	**[0 1 9]**	9.6	**[0 3 9]**	9.6
EPIDDD3Way_NN_H3E400	68.9	61.8	**[0 4 9]**	13.7	**[0 2 9]**	8.5	**[0 8 9]**	8.4
EPIDDD3Way_NN_H3E500	68.9	61.8	**[0 4 9]**	19	**[0 7 9]**	9.6	**[0 4 7]**	9.5

Conclusion

- We have successfully used gene perturbation and our metric to detect and characterize gene interactions
- The method is designed to get information learned by black box ML methods (SVM, NN)
- The method is tolerant of noise such as would be expected when information is limited
- Our next step is to use actual, rather than simulated data