

CAGI slide pack guidelines

I agree to my slides being used according to the following terms:

_____ *Yang Shen* _____

- Authors: please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 2013. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.



Predicting Missense Mutational Effects on Cancer Pathogenicity

Yuanfei Sun, Yue Cao, Haoran Chen,
Mostafa Karimi, Oluwaseyi Moronfoye, Yang Shen

Department of Electrical & Computer Engineering
Center for Bioinformatics and
Genomic Systems Engineering
Texas A&M University



Problem description & challenges

- To predict the probability of pathogenicity for 430 BRCA1/BRCA2 variants

(ENIGMA)

- 318 missense

Class	Probability of Pathogenicity
5: Pathogenic	>0.99
4: Likely pathogenic	0.95-0.99
3: Uncertain	0.05-0.949
2: Likely not pathogenic	0.001-0.049
1: Not pathogenic	<0.001

- **Challenges**

- A continuous label p with error estimation is needed

Not classification ...

whereas available data only has discrete label categories (with corresponding ranges).

Not normal regression or even ordinal regression ...

- Lack of data

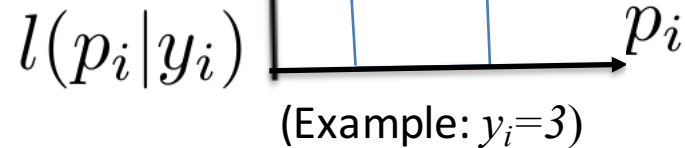
What's the machine learning problem?

- Approach: Design a loss function for the special regression

- Design rationales for $l(p_i|y_i)$
 - Pathogenicity $[0, 1]$ (blue arrow)
 - Class (1,...,5) (red arrow)
 - Example i (black arrow)

Domain-specific

- Only sensitive when falling out of the supposed range
- Higher as predicted class error increases
- Directionality ($\Rightarrow y_i$ -dependent)



Model training

- Convex for global optimization
- Continuous and smooth for gradient-based methods

Formulating the problem as a special ordinal regression

- Approach: Design a loss function for the special case

Pathogenicity $[0, 1]$ Class (1,...,5)

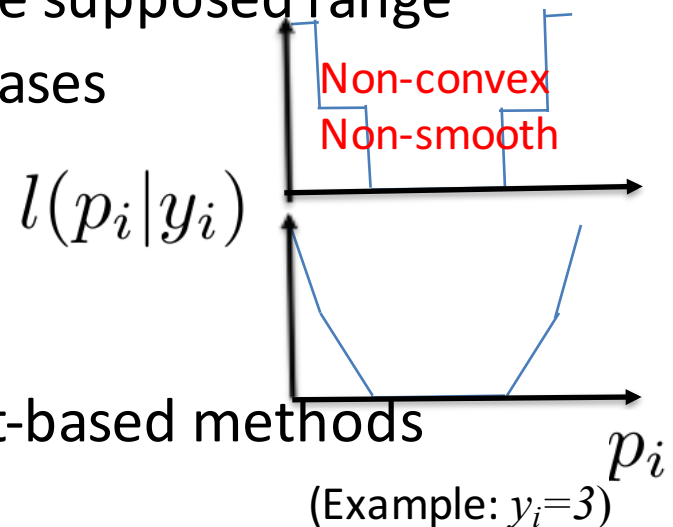
- Design rationales for $l(p_i|y_i)$
Example i

Domain-specific

- Only sensitive when falling out of the supposed range
- Higher as predicted class error increases
- Directionality ($\Rightarrow y_i$ -dependent)

Model training

- Convex for global optimization
- Continuous and smooth for gradient-based methods



Formulating the problem as a special ordinal regression

- Approach: Design a loss function for the special case

Pathogenicity $[0, 1]$ Class (1,...,5)

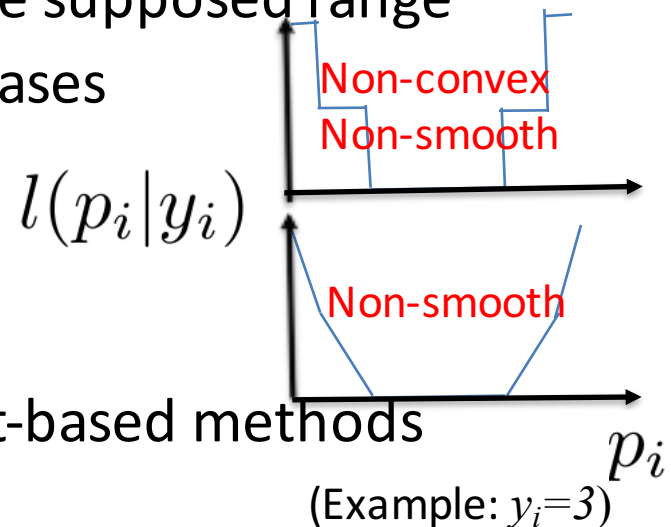
- Design rationales for $l(p_i|y_i)$
Example i

Domain-specific

- Only sensitive when falling out of the supposed range
- Higher as predicted class error increases
- Directionality ($\Rightarrow y_i$ -dependent)

Model training

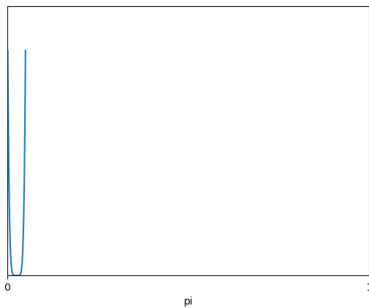
- Convex for global optimization
- Continuous and smooth for gradient-based methods



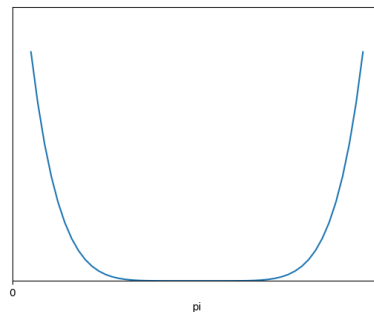
Formulating the problem as a special ordinal regression

- Final answer: high-order polynomial

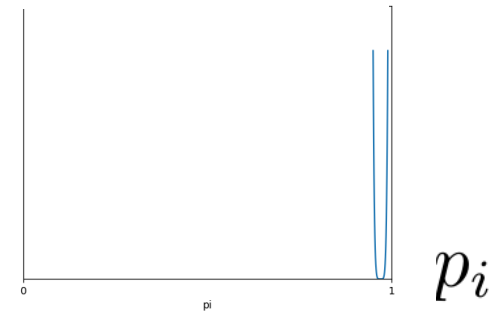
$$l(p_i | y_i = 2)$$



$$l(p_i | y_i = 3)$$



$$l(p_i | y_i = 4)$$



$$l(p_i | y_i) = \frac{1}{6} \left(\frac{p_i - \frac{1}{2} (lb(y_i) + ub(y_i))}{\frac{1}{2} (ub(y_i) - lb(y_i))} \right)^q$$

y_i	1	2	3	4	5
$[lb(y_i), ub(y_i)]$	$[0, 0.001]$	$[0.001, 0.05]$	$[0.05, 0.95]$	$[0.95, 0.99]$	$[0.99, 1]$

q needs to be even for l to be symmetric and high enough to have a flat bottom $\Rightarrow q=6$ in CAGI5

Solving the special ordinal regression

- For the i th example of n features: $\mathbf{x}_i = (x_i^1, x_i^2, \dots, x_i^n)$

Pathogenicity model $p_i = f(\mathbf{x}_i)$: *linear in CAGI5*

$$p_i = \mathbf{w}^T \mathbf{x}_i$$

- Finding the optimal model

Can extend to nonlinear
shallow or deep models;
But easy interpretability ...

$$\mathbf{w}^* = \arg \min \sum_{i=1}^m \frac{m}{m_{y_i}} \cdot \frac{1}{6} \left(\frac{\mathbf{w}^T \mathbf{x}_i - \frac{1}{2}(lb(y_i) + ub(y_i))}{\frac{1}{2}(ub(y_i) - lb(y_i))} \right)^6 + \lambda \|\mathbf{w}\|^2$$

Class-balancing

$l(p_i|y_i)$

Regularization

*Convex & smooth =>
Gradient methods can guarantee
the global optimum*

- 5-fold cross-validation for optimal hyper-parameter λ
- All training data for optimal parameters $\mathbf{w}^*$
- Use the model to predict for the ENIGMA set
- Repeat 10 times for average and standard deviation

Where is the data?

- ClinVar: ENIGMA-annotated BRCA1/BRCA2 missense variants

Table 1: Training data characteristics

Gene	Not pathogenic(1)	Likely not pathogenic(2)	Uncertain(3)	Likely Pathogenic(4)	Pathogenic(5)	Total(ratio)
BRCA1	87	0	0	0	21	108(5.4%)
BRCA2	84	0	0	0	9	93(4.7%)

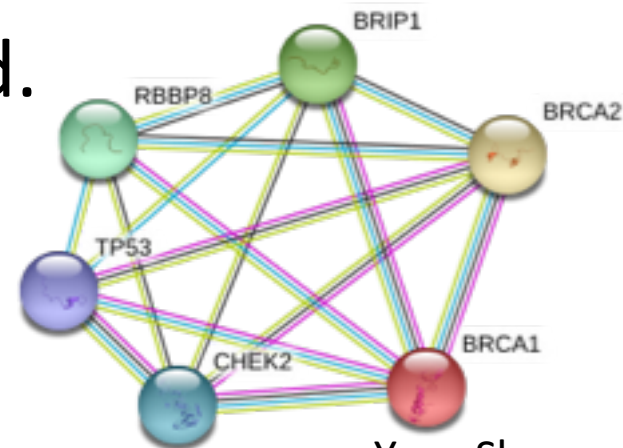
No single example for Classes 2-4!

Where is the data?

- We also included proteins that interact with BRCA1 and BRCA2 (in breast-cancer related pathways) and may share similar mechanisms.
- Top 5 proteins interacting with BRCA1, as identified by STRINGdb and known as tumor suppressor genes, are included.

More proteins?

Some stratification?



Training data

- ClinVar: ENIGMA-annotated missense variants of six genes

Table 1: Training data characteristics

Gene	Not pathogenic(1)	Likely not pathogenic(2)	Uncertain(3)	Likely Pathogenic(4)	Pathogenic(5)	Total(ratio)
BRCA1	87	0	0	0	21	108(5.4%)
BRCA2	84	0	0	0	9	93(4.7%)
CHEK2	0	3	590	3	1	597(30.0%)
BRIP1	1	9	761	0	0	771(38.8%)
RBBP8	1	2	17	0	1	21(1.1%)
TP53	0	11	310	66	12	399(20.1%)
Total(ratio)	173(8.7%)	25(1.3%)	1678(84.4%)	69(3.5%)	44(2.2%)	1989

What features did we use?

- 15 scores from MutPred2 that provide posterior probabilities of the loss/gain of certain structural and functional properties due to the substitution

Features

1. General_effect
2. Structure_and_dynamics
3. Secondary_structure
4. Stability_and_conformational_flexibility
5. Conformational_flexibility
6. Special_structural_signatures
7. Signal_peptide
8. Transmembrane_protein
9. Functional_residue
10. Macromolecular_binding
11. Protein_binding
12. Disordered_interface
13. Ordered_interface
14. Metal_binding
15. PTM_site

**More or less covering the space of
potential molecular mechanisms ...**

Comparing training and testing data

Table 1: Training data characteristics

Gene	Not pathogenic(1)	Likely not pathogenic(2)	Uncertain(3)	Likely Pathogenic(4)	Pathogenic(5)	Total(ratio)
BRCA1	87	0	0	0	21	108(5.4%)
BRCA2	84	0	0	0	9	93(4.7%)
CHEK2	0	3	590	3	1	597(30.0%)
BRIP1	1	9	761	0	0	771(38.8%)
RBBP8	1	2	17	0	1	21(1.1%)
TP53	0	11	318	66	12	399(20.1%)
Total(ratio)	173(8.7%)	25(1.3%)	1678(84.4%)	69(3.5%)	44(2.2%)	1989

Table 2: Test data characteristics

Gene	Not pathogenic(1)	Likely not pathogenic(2)	Uncertain(3)	Likely Pathogenic(4)	Pathogenic(5)	Total(ratio)
BRCA1	31	100	2	4	7	144(45.3%)
BRCA2	31	136	2	2	3	174(54.7%)
Total(ratio)	62(19.5%)	236(74.2%)	4(1.3%)	6(1.9%)	10(3.1%)	318

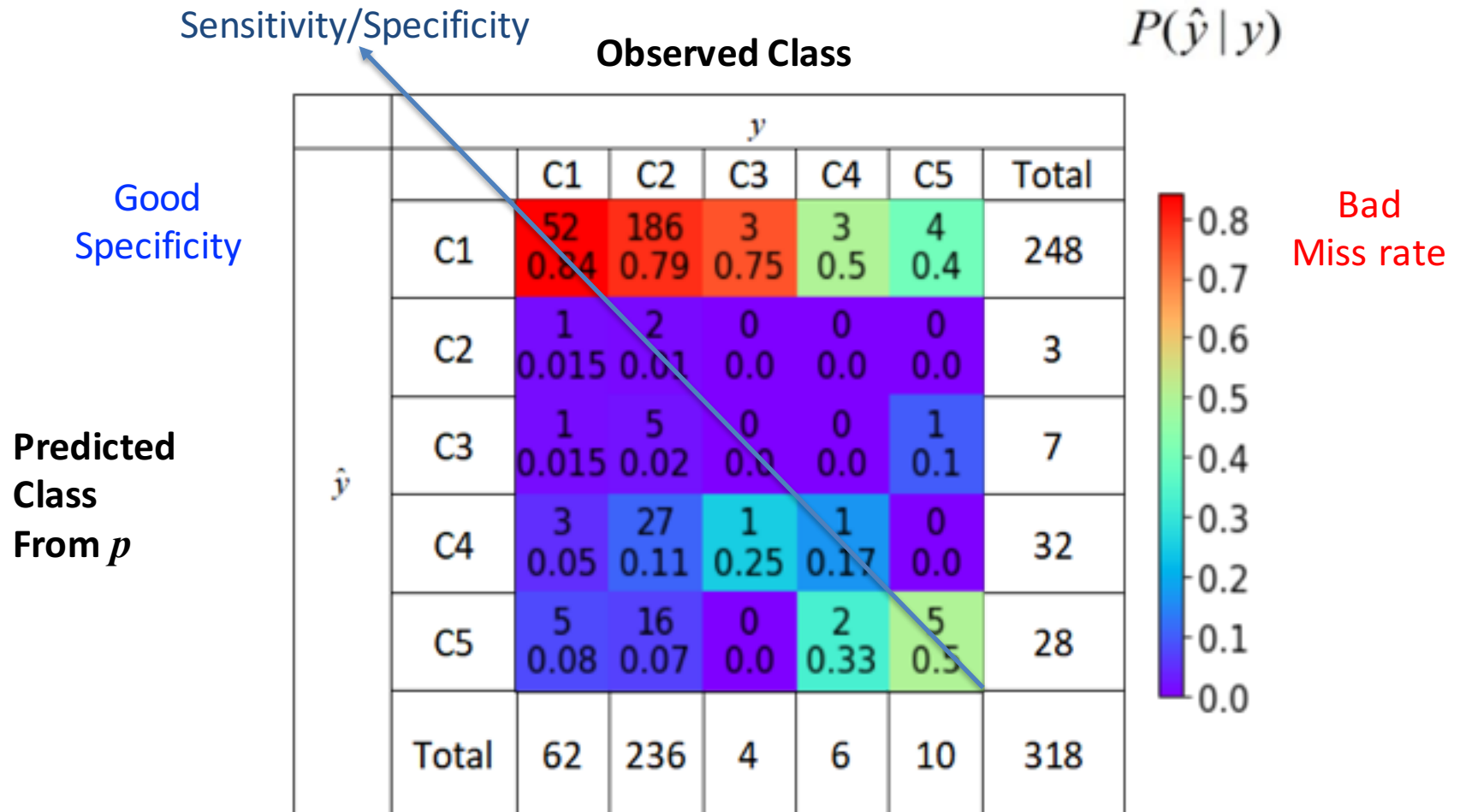
Classification performances (I)

5x5 confusion matrix

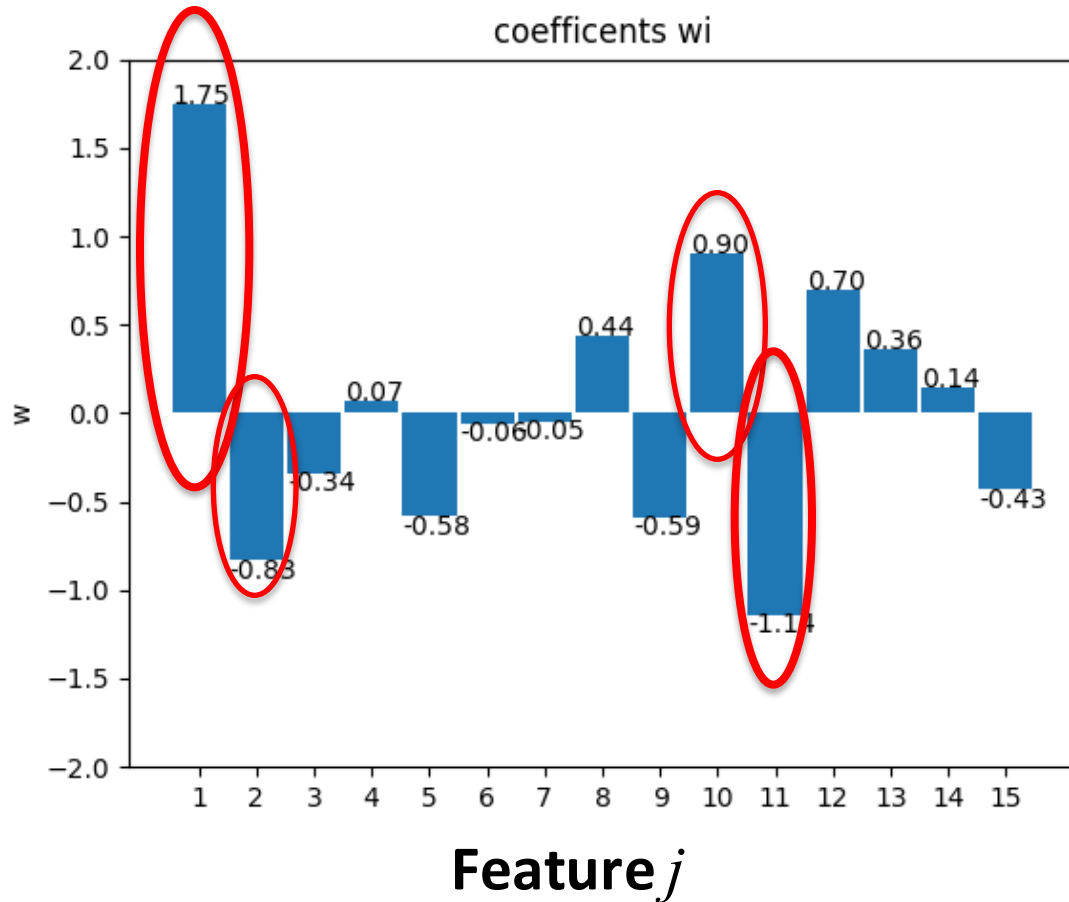
		Observed Class						
		y						
Predicted Class From p	$\hat{y}$		C1	C2	C3	C4	C5	Total
		C1	52	186	3	3	4	248
		C2	1	2	0	0	0	3
		C3	1	5	0	0	1	7
		C4	3	27	1	1	0	32
		C5	5	16	0	2	5	28
		Total	62	236	4	6	10	318

Classification performances (I)

5x5 confusion matrix (normalized by column)



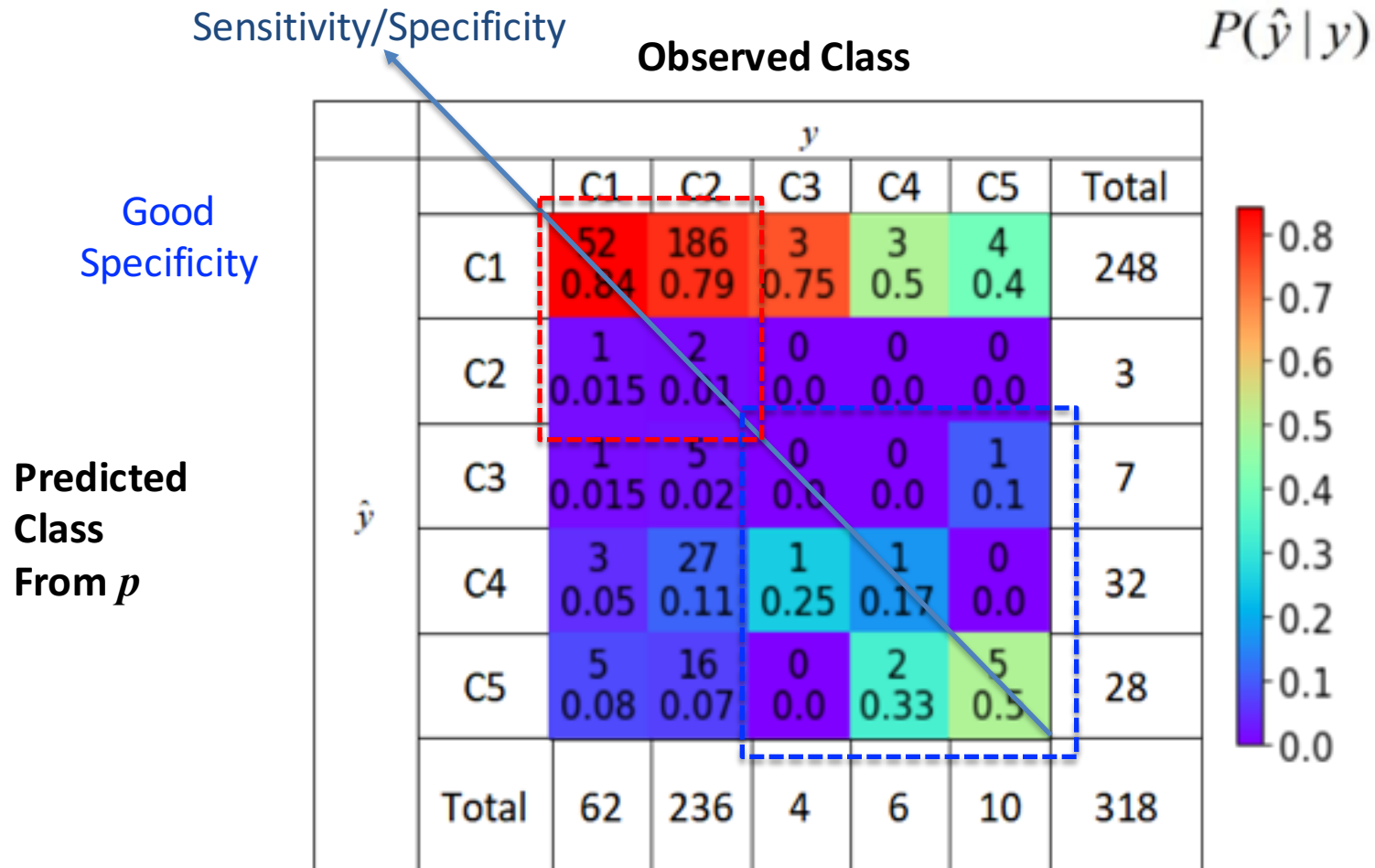
What features were important in general?



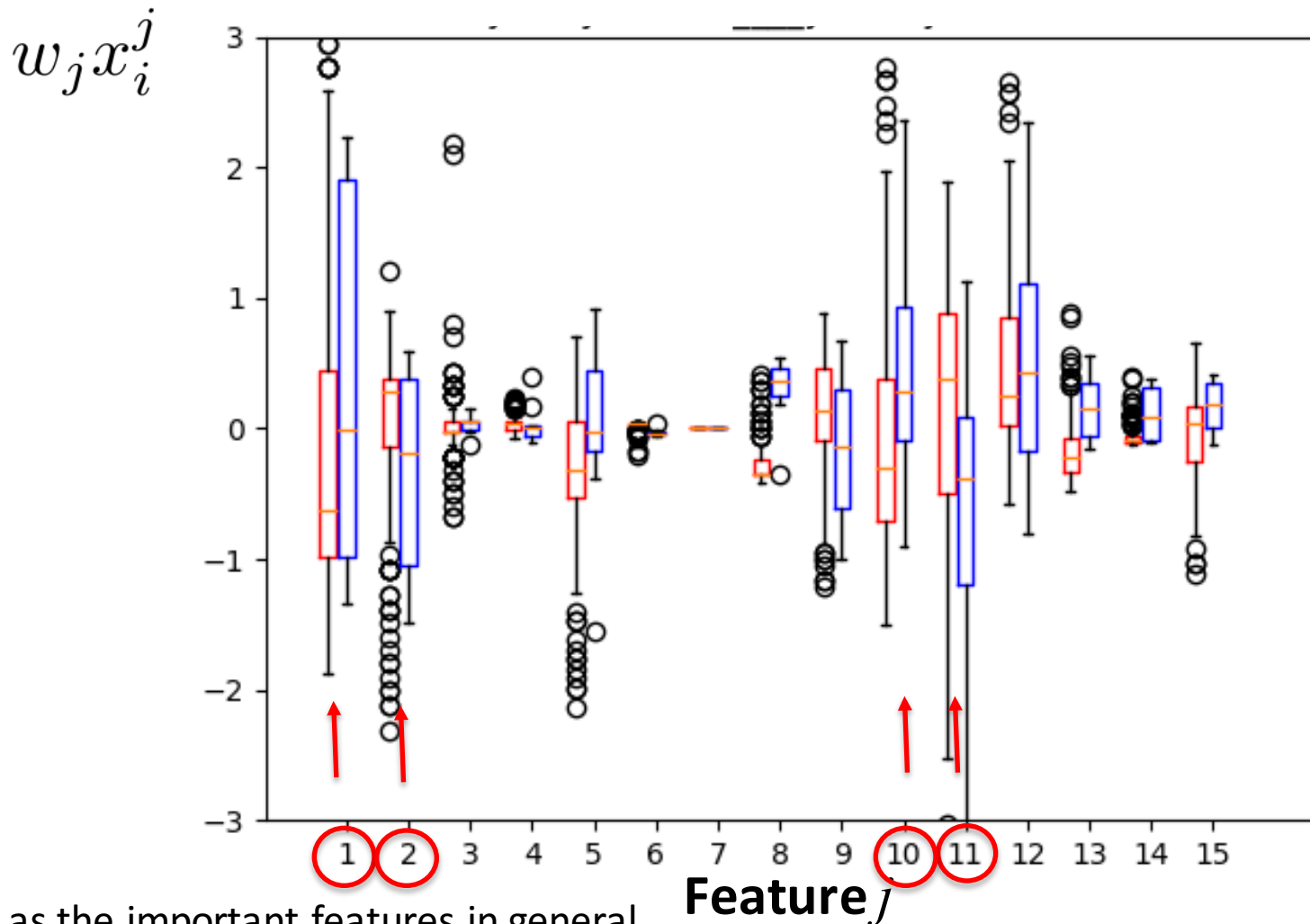
1. **General_effect**
2. **Structure_and_dynamics**
3. Secondary_structure
4. Stability_and_conformational_flexibility
5. Conformational_flexibility
6. Special_structural_signatures
7. Signal_peptide
8. Transmembrane_protein
9. Functional_residue
10. **Macromolecular_binding**
11. **Protein_binding**
12. Disordered_interface
13. Ordered_interface
14. Metal_binding
15. PTM_site

What features separated **TN (Class 1+2)** well from **TP (Class 3+4+5)**?

5x5 confusion matrix (normalized by column)

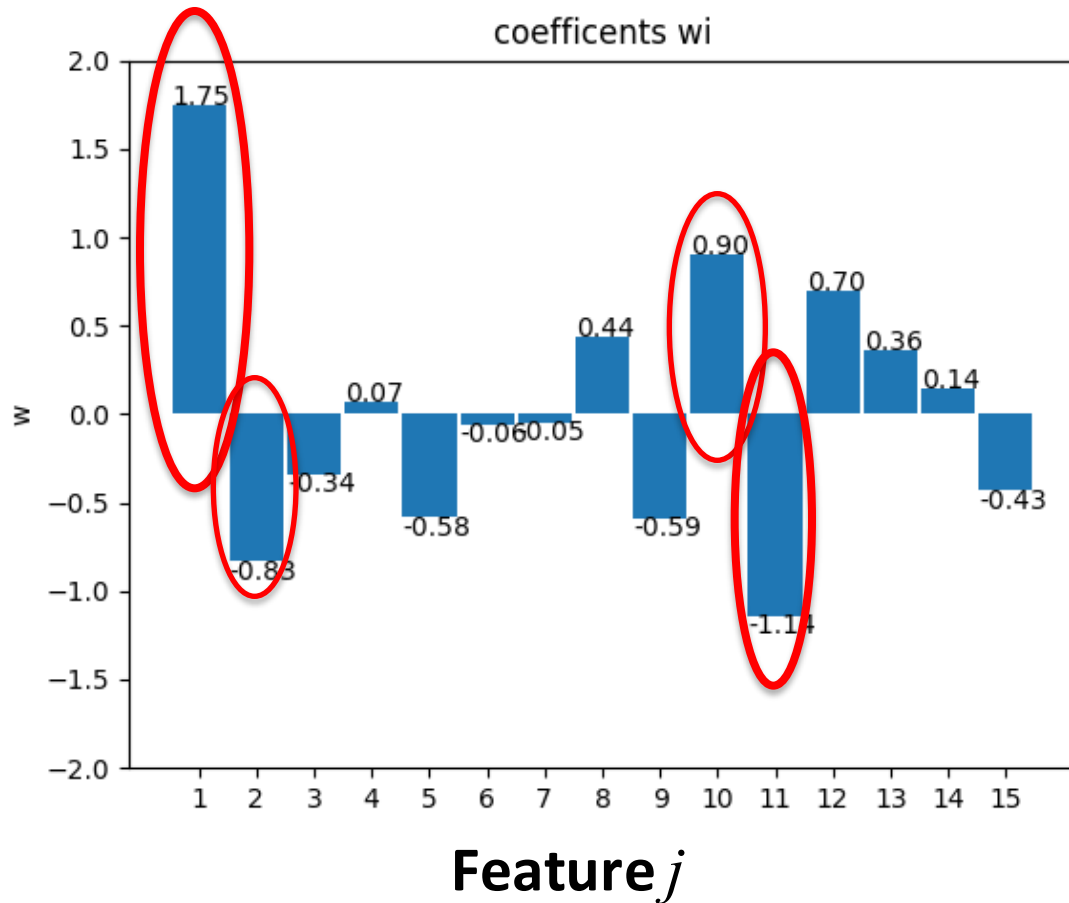


What features separated **TN (Class 1+2)** well from **TP (Class 3+4+5)**?



Same as the important features in general.

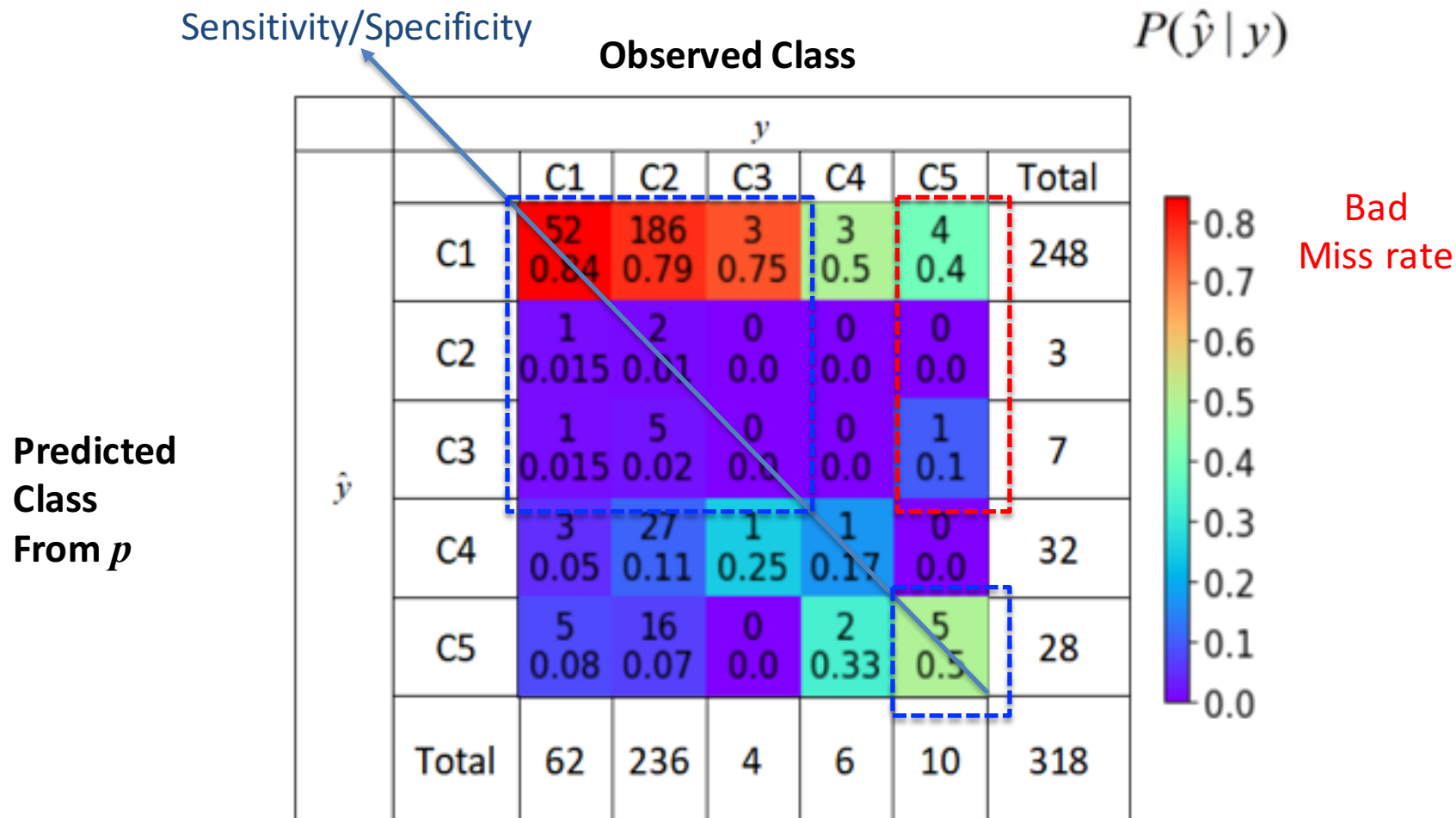
Recall: Important features in general?



1. **General_effect**
2. **Structure_and_dynamics**
3. Secondary_structure
4. Stability_and_conformational_flexibility
5. Conformational_flexibility
6. Special_structural_signatures
7. Signal_peptide
8. Transmembrane_protein
9. Functional_residue
10. **Macromolecular_binding**
11. **Protein_binding**
12. Disordered_interface
13. Ordered_interface
14. Metal_binding
15. PTM_site

What features confused the model for some *actual* Class 5 (Pathogenic) variants?

5x5 confusion matrix (normalized by column)



What features confused the model for some *actual* Class 5 (Pathogenic) variants?

$$w_j x_i^j$$

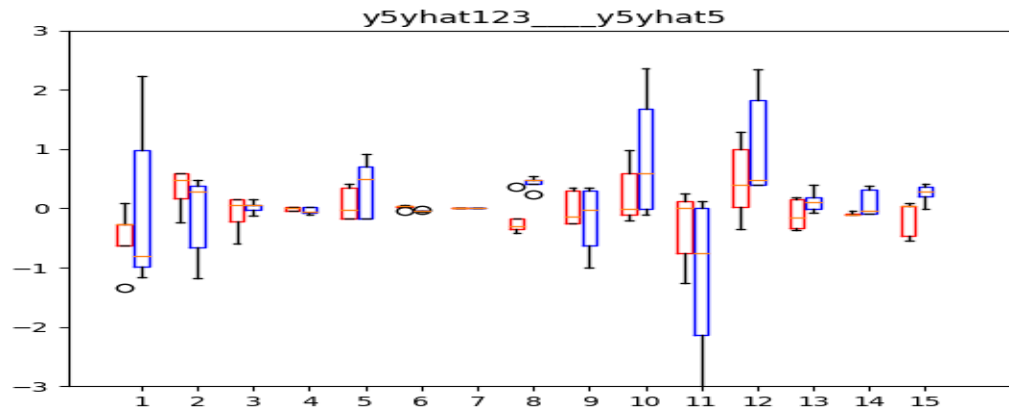
FN aa substitution:

R→G X2

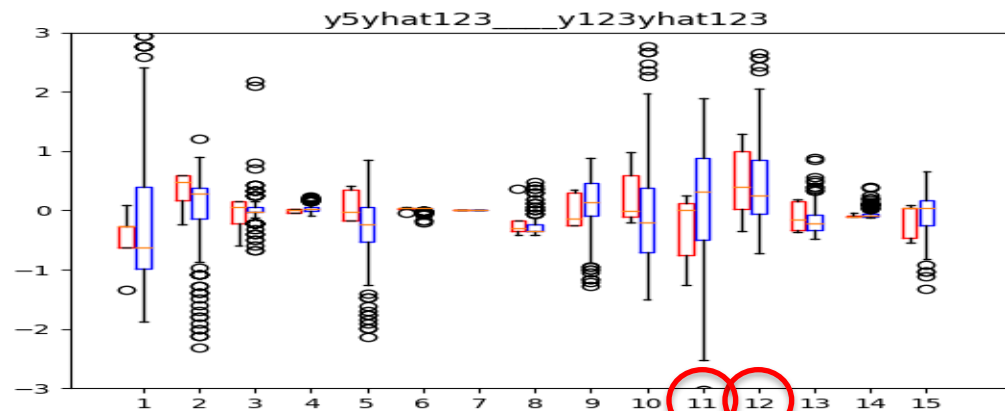
R→T

Q→K

Q→R



FN vs TP (Class 5)



FN vs TN (Class 1-3)

Feature j

Naturally(?), features behind some success (1, 2, 10, 11) are also the biggest reasons behind some failures. In particularly feature 12 (disordered_interface) caused much confusion (FN).

Classification performances (II)

5x5 confusion matrix

		Observed Class						
		y						
Predicted Class From p	$\hat{y}$		C1	C2	C3	C4	C5	Total
		C1	52	186	3	3	4	248
		C2	1	2	0	0	0	3
		C3	1	5	0	0	1	7
		C4	3	27	1	1	0	32
		C5	5	16	0	2	5	28
		Total	62	236	4	6	10	318

Classification performances (II)

5x5 confusion matrix (normalized by row)

PPV/NPV

Observed Class

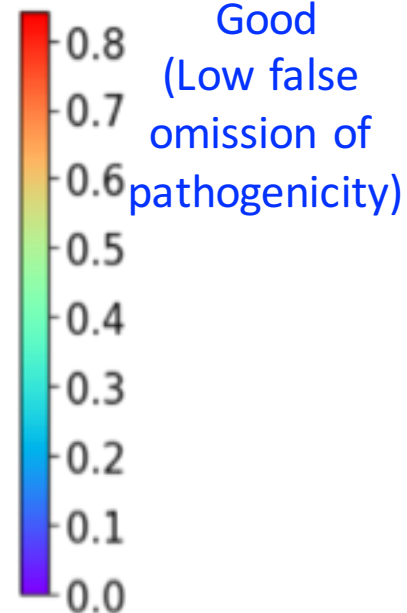
$P(y | \hat{y})$

Good NPV
(Predicted not pathogenic
ones are often not)

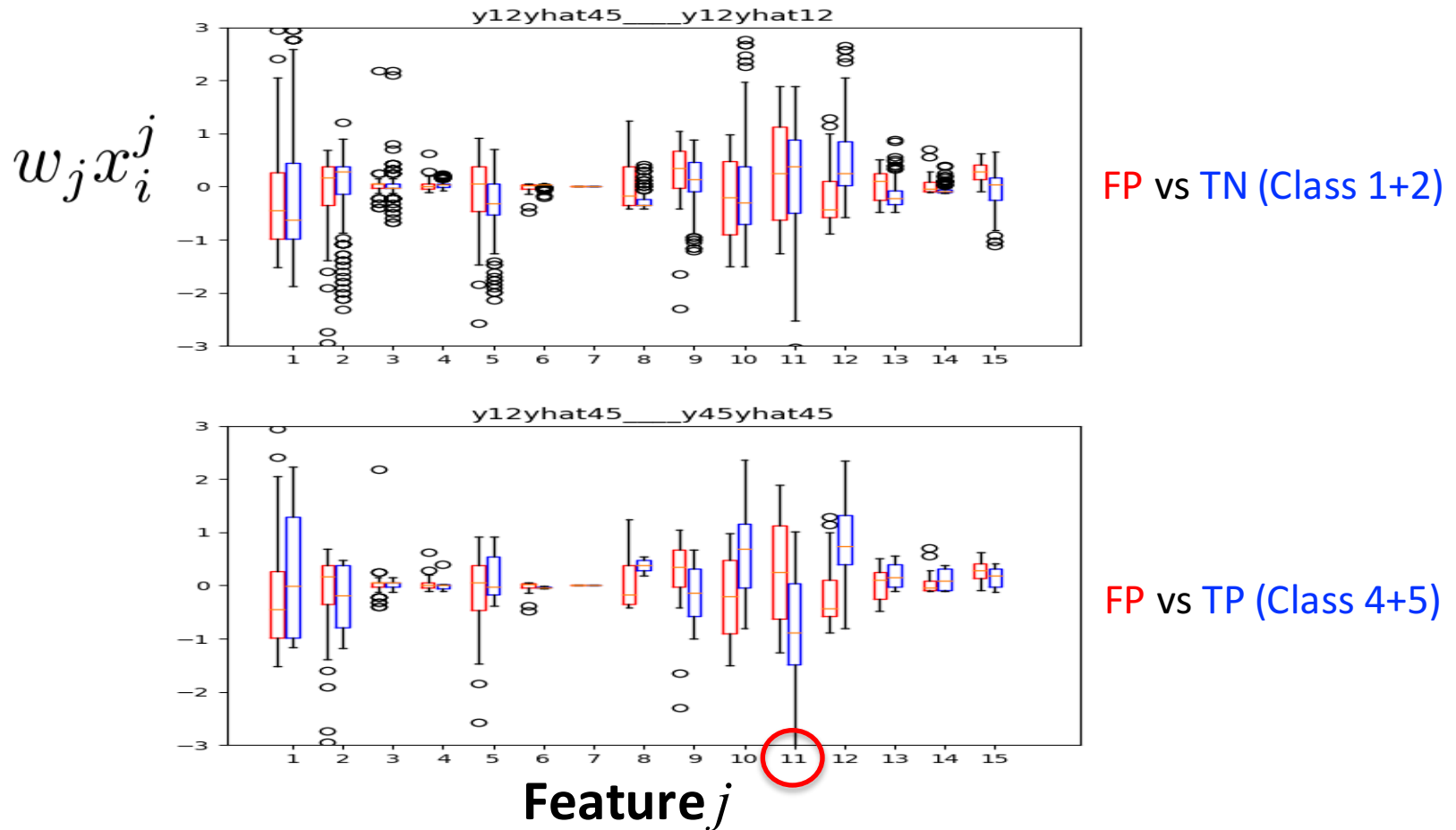
**Predicted
Class
From p**

Bad FDR
(High false discovery
of pathogenicity)

		y					Total
		C1	C2	C3	C4	C5	
$\hat{y}$	C1	52 0.21	186 0.75	3 0.01	3 0.01	4 0.02	248
	C2	1 0.33	2 0.67	0 0.0	0 0.0	0 0.0	3
	C3	1 0.14	5 0.72	0 0.0	0 0.0	1 0.14	7
	C4	3 0.09	27 0.85	1 0.03	1 0.03	0 0.0	32
	C5	5 0.18	16 0.57	0 0.0	2 0.07	5 0.19	28
	Total	62	236	4	6	10	318



What features confused the model for some incorrectly *predicted* Class 4+5 ([Likely]Pathogenic) variants?



In this case, no feature significantly gets more similar from the top to the bottom two groups. Rather, this seems to be collective results from features 10+12 (↑) and 11(↓). In particular, feature 11 (and 9) overestimated the probability to be close to pathogenic classes (TP).

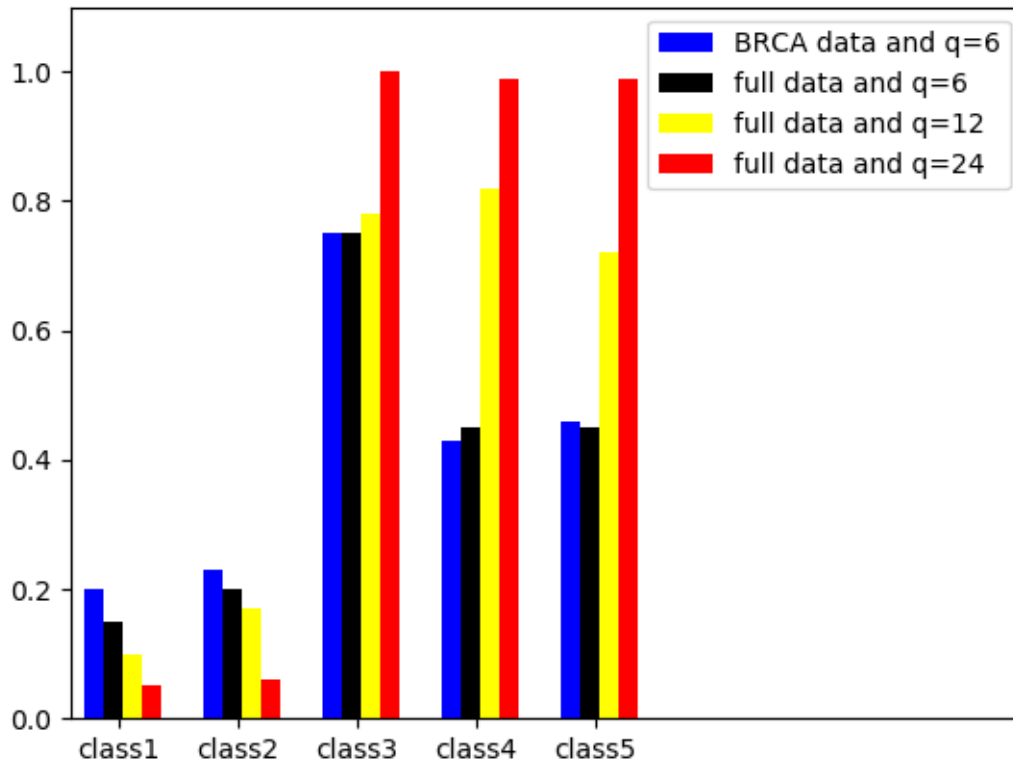
Prelim. lessons learned from the features*

- In general, important features were
 - General_effect, Protein_binding
 - Macromolecular_binding, Structure_and_dynamics
 - Disordered_interface, Conformational_flexibility, functional_residue
- These are responsible for some failures as well
 - [False negative contributor: Disordered_interface, *Protein_binding*]*
 - [False positive contributor: *Protein_binding*]*

* Limited study. [More accurate, higher-resolution, or simply more features? Introducing nonlinearity and joint effects?](#)

What if we had ...

* Misclassification rate when actual class and neighbor class(es) are regarded correct.



- Only used BRCA1/BRCA2 training data?
No significant difference?!
- Used higher-order polynomial for loss function?

*Almost every case is predicted Class 1 [?]
(expectable due to steep cross-class walls rather than slopes to “guide”)*

Uncertainty estimation

- Standard deviations

[0, 0.01): 291/318; **[0.01, 0.1): 20/318**; [0.21, 0.32]: 7/318

(including the only 3 predicted to be Class 2 “Likely not pathogenic”
and 17/32 predicted to be Class 4 “Likely pathogenic”)

Predicted Class	Actual Class	Freq
4	2	14/20
4	1	3/20
2	2	2/20
2	1	1/20

What could be improved?

- Nonlinear shallow or deep models
- Stratified training data
- Intron mutations
- Features (and structures)!

iCFN for multistate protein design and mutational effect analysis

- Exact algorithm with guaranteed optimum
- Considers positive and negative objectives

iCFN: an efficient exact algorithm for multistate protein design, Mostafa Karimi and Yang Shen, accepted to ECCB'18.

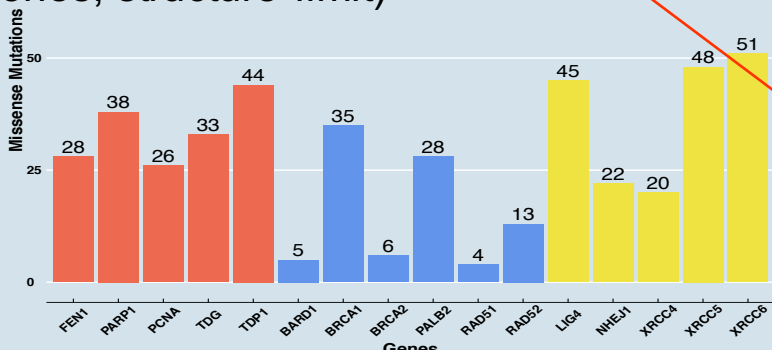
ISMB (3Dsig) poster (B-549, July 9 evening) & talk (July 10 talk)

- Application to somatic mutations for pan-cancer DDR (DNA damage repair)

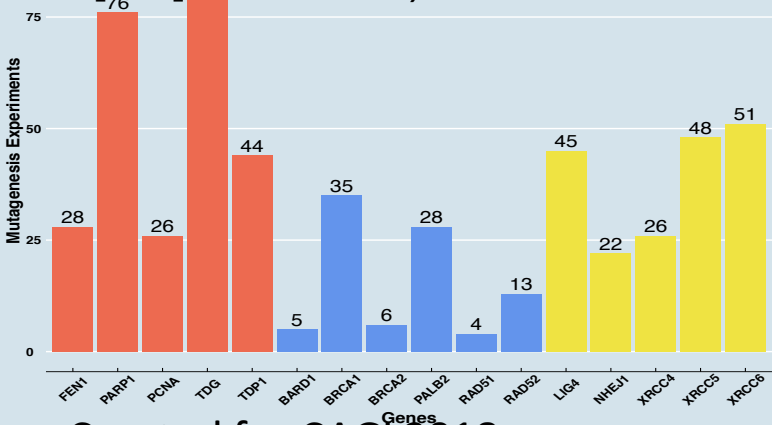
“Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas”, TA Knijnenburg et. al., Cell Reports, 23(1), 239-254.

iCFN's interpretation of DDR mutations in the context of PPIs

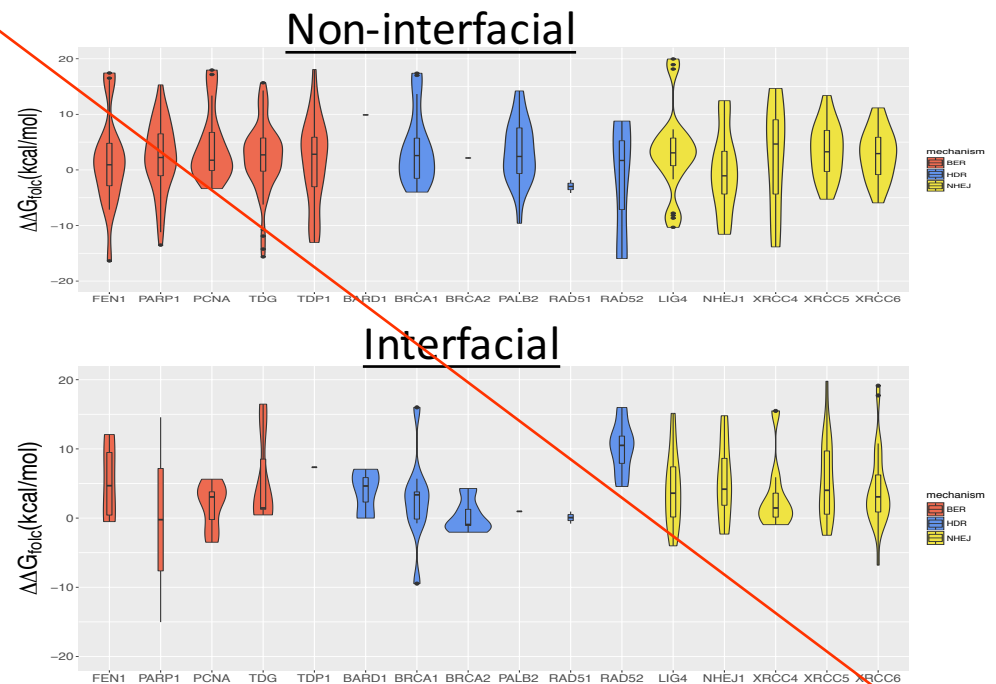
446 Mutations
(2.3% of 19,598 unique pan-can missense ones; structure limit)



547 Mut. Exp.
155 [1/3] Interfacial)



354 (64.7%) such mutations were predicted to strongly destabilize ($\Delta\Delta G_{\text{fold}} \geq 3k_B T$) complexes formed by protein-protein interactions (PPIs), some of which are through disrupting binding.



Acknowledgement

- Group members at Texas A&M University: Yuanfei Sun, Yue Cao, Haoran Chen, Mostafa Karimi, Oluwaseyi Moronfoye
- Dr. Sean Mooney and his group (MutPred2)
- CAGI organizers, data-providers, and assessors

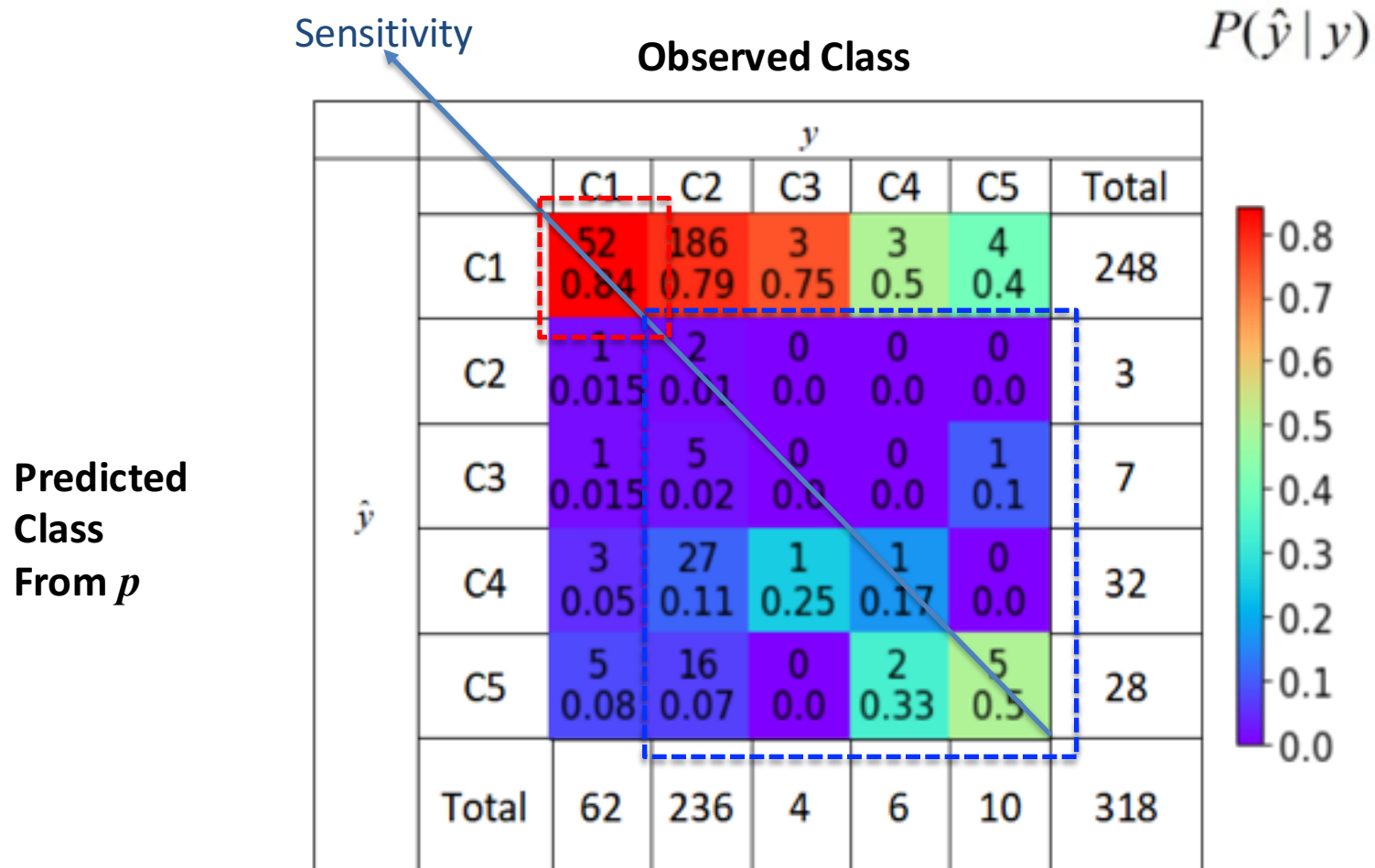


TEXAS A&M
UNIVERSITY

Ph.D. student and postdoc openings

What features separated well **TN (Class 1)** from **TP (Not Class 1)**?

5x5 confusion matrix (normalized by column)



What features separated **TN (Class 1)** well from **TP (Not Class 1)**?

