### **Pyridine Phosphination - Best Model Feature Analysis**

### **Table of Contents**

Published Model	1
5th Place with 5 Features	
Model Analysis - Variable Impact	
1. Avg_NPA_SM Removed	
2. fr_Ar_N Removed	
3. fr_aryl_methyl Removed	17
4. fr_benzene Removed	22
5. fr_halogen Removed	
Section Results Summary	

#### **Published Model**

#### **5th Place with 5 Features**

Table:

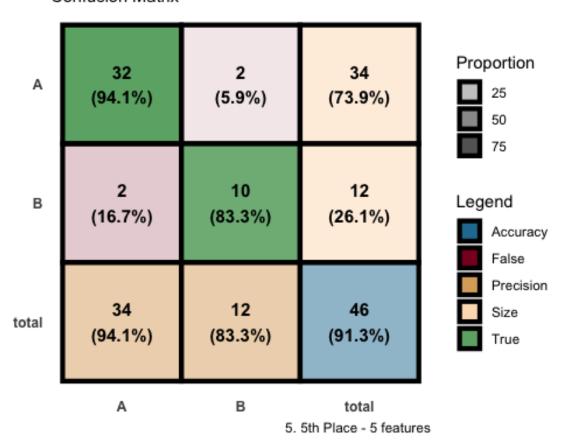
Full Model Stats - Overall Accuracy and Pseudo-R2

Accuracy McFadden\_R2 91.3% 0.578

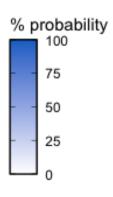
	X
(Intercept)	38.099048
Avg_NPA_SM	48.412036
fr_Ar_N1	-31.903705
fr_Ar_N2	-60.673746
fr_Ar_N3	-66.149054
fr_aryl_methyl1	-4.592317
fr_aryl_methyl2	-37.006653
fr_benzene1	-1.970290
fr_benzene2	-77.887220
fr_halogen1	1.607589
fr_halogen2	2.929693
fr_halogen3	2.595277
fr_halogen4	-2.087011

	X
fr_halogen5	2.187832
fr_halogen6	-2.311938

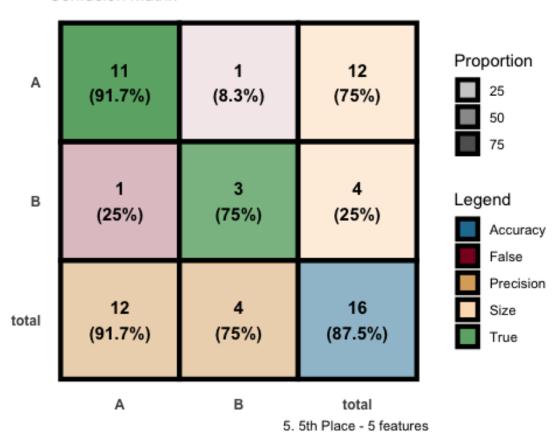
Training Set Confusion Matrix



	Α	В	Exp
Α	100	0	Α
AF	94	6	A
F	100	0	Α
AB	66	34	Α
AC	94	6	Α
AD	100	0	Α
AE	100	0 0 5	A
AH	95	5	Α
ΑI	100	0 0 2 6 2 33	Α
AJ	100	0	Α
AK	98	2	A
AL	94	6	Α
AM	98	2	Α
G	98 67	33	Α
- 1	42	58	Α
J	42 99	58 1 17 56	Α
K	83	17	Α
L	44	56	Α
M	44 81	19	Α
N	100	0	Α
0	75	25	Α
Р	100	0	Α
Q R T U	100	0	A
R	100	0	Α
Т	100	0	Α
U	100	0	Α
V W	92	8	Α
W	99	1	A
Х	100	0	Α
Υ	98	2	A
Z	98 61	2	A
AN	61	39	Α
AO	73	2 2 39 27 0	Α
AS	100	0	A
AV	29	71	В
AW	32	68	В
AX	43	57	В
AY	17	83	В
BB	4	96	В
BE	0	100	В
BF	0	100	В
BG	6	94	В
BH	39	61	В
BI	81	19	В
BJ	0	100	В
BK	95	5	В



Test Set Confusion Matrix



Test Set Probability Heatmap

	Α	В	Exp	
В	74	26	A	
С	100	0	A	
D	100	0	A	
Е	100	0	A	
AA	73	27	A	0/ probability
Н	32	68	A	% probability
S	100	0	A	75
AP	94	6	A	75
AQ	78	22	A	50
AR	100	0	A	- 25
AT	100	0	A	
AU	94	6	A	<b>□</b> 0
AZ	35	65	В	
BA	0	100	В	
вс	65	35	В	
BD	4	96	В	

5. 5th Place - 5 features

### **Model Analysis - Variable Impact**

### 1. Avg\_NPA\_SM Removed

#### formula

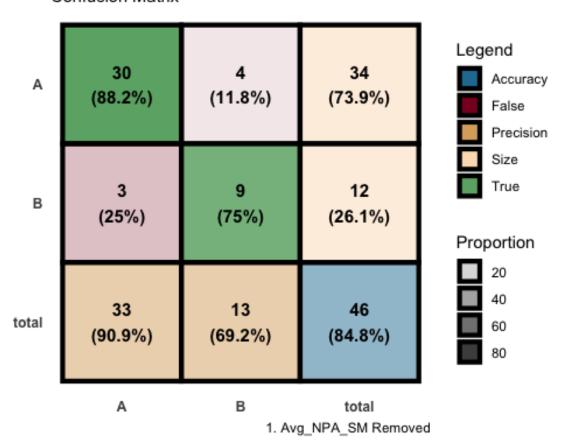
class ~ fr\_Ar\_N + fr\_aryl\_methyl + fr\_benzene + fr\_halogen

Full Model Stats - Overall Accuracy and Pseudo-R2

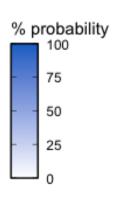
Accuracy McFadden\_R2 84.78% 0.52

	X
(Intercept)	14.653960
fr_Ar_N1	-14.641780
fr_Ar_N2	-28.886527
fr_Ar_N3	-30.760596
fr_aryl_methyl1	-3.720066
fr_aryl_methyl2	-16.691481
fr_benzene1	-2.364301
fr_benzene2	-35.077472
fr_halogen1	2.342569
fr_halogen2	4.212291
fr_halogen3	1.848104
fr_halogen4	-1.081430
fr_halogen5	2.352174
fr_halogen6	-9.373352

Training Set Confusion Matrix

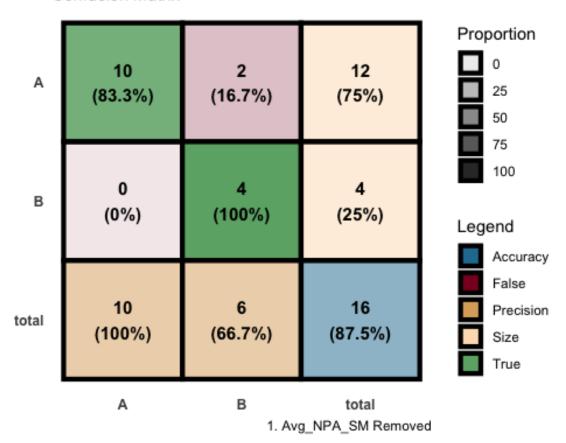


	Α	В	Exp
Α	100	0	A
AF	80	20	A
F	100	0	Α
AB	50 91	50	Α
AC	91	9	A
AD	100	0	A A
AE	100 80	0 20	A
AH AI	100	20	A
ΑJ	100 100	0 0 2 13 2 50	A
AK	98	2	A
AL	87	13	Α
AM	98	2	Α
G	50	50	A
- 1	50	50	A
J	100 87	0 13 50 9	Α
K	87	13	Α
K L M	50 91	50	A
IVI	91	9	A
N	100	9	A A
O P	91 100	0	Ā
0	100	0	Â
Q R T U V	100	0	A
T	100	Ö	Α
U	100	0	A
V	91	9	Α
W	98	2	A
Х	100	0	A
Υ	98	2	A
Z	98	2	Α
AN	50 50	9 2 0 2 2 50 50	A
AO AS	100	0	A A
AV	100 50	50	В
AW	50	50	В
AX	80	20	В
AY	13	87	В
BB	9	91	В
BE	0	100	В
BF	0	100	В
BG	13	87	В
ВН	50	50	В
ВІ	50	50	В
BJ	9	91	В
BK	91	9	В



1. Avg\_NPA\_SM Removed

Test Set Confusion Matrix



Test Set Probability Heatmap

	Α	В	Exp	
В	91	9	A	
С	100	0	A	
D	100	0	A	
Е	100	0	A	
AA	50	50	A	0/ probability
н	50	50	A	% probability
S	100	0	A	
AP	91	9	A	75
AQ	50	50	A	50
AR	100	0	A	- 25
AT	100	0	A	
AU	91	9	A	<b>□</b> 0
AZ	50	50	В	
BA	13	87	В	
вс	50	50	В	
BD	50	50	В	

1. Avg\_NPA\_SM Removed

### 2. fr\_Ar\_N Removed

### formula

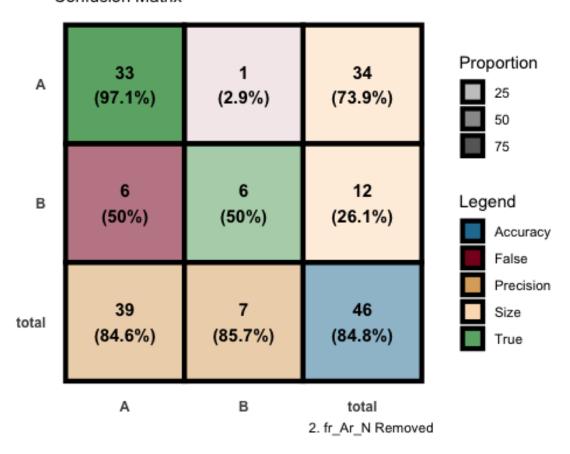
 ${\tt class} \sim {\tt Avg\_NPA\_SM} + {\tt fr\_aryl\_methyl} + {\tt fr\_benzene} + {\tt fr\_halogen}$ 

Full Model Stats - Overall Accuracy and Pseudo-R2

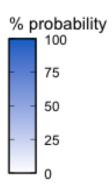
Accuracy McFadden\_R2 84.78% 0.328

X
0.9188853
18.4857033
-2.7592175
-12.9215911
-0.8918341
-32.6158024
2.0595415
3.2075153
1.5695469
-8.4512972
2.4217225
-4.8502196

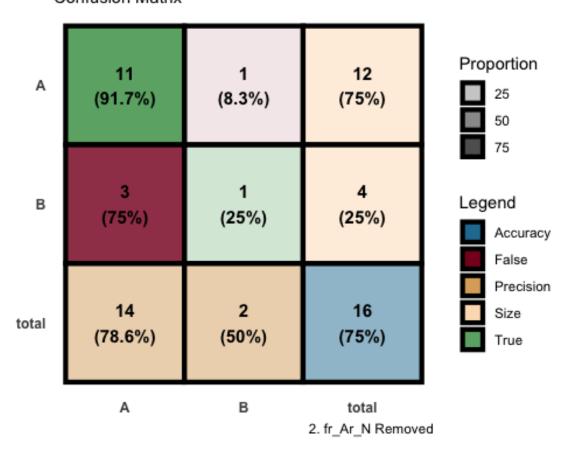
Training Set Confusion Matrix



	Α	В	Exp
Α	100	0	Α
AF	89	11	Α
F	29	11 71	Α
AB	85	15 7	Α
AC	93	7	Α
AD	100 82	0	Α
AE	82	18	Α
AH	90	10	Α
Al	71 51	29	Α
AJ	51	49	A
AK	98	2 18	A
AL	82	18	A
AM	98 85	2 15	A
G	85	15	A
- !	51 99	49	A
J	99	1	A A
K L	92	8 49	A
M	51 90	10	A
N	76	24	Â
O	88	12	Ä
P	68	32	A
	82	24 12 32 18 1 37	A
Q R T U	99	1	A
Т	63	37	A
Ü	100	0	A
V	93	7	Α
W	95	5	Α
Χ	100	0	Α
Y	94	6	Α
Z	94	6	Α
AN	54	46	Α
AO	86	14	Α
AS	100	0	Α
ΑV	75	25	В
AW	<u>76</u>	24	В
AX	71	29	В
AY	56	44	В
BB	23	77	В
BE	18	82	В
BF	29	71	В
BG	18	82	В
BH	46	54	B B
BJ	67	33	В
BK	94	98	В
DN	94	6	В



Test Set Confusion Matrix



Test Set Probability Heatmap

	Α	В	Exp	
В	88	12	A	
С	86	14	A	
D	100	0	A	
Е	100	0	A	
AA	86	14	A	% probability
Н	76	24	A	% probability
S	62	38	A	75
AP	94	6	A	75
AQ	65	35	A	50
AR	100	0	A	25
AT	41	59	A	
AU	93	7	A	<b>□</b> 0
ΑZ	77	23	В	
BA	7	93	В	
вс	84	16	В	
BD	55	45	В	

2. fr\_Ar\_N Removed

### 3. fr\_aryl\_methyl Removed

### formula

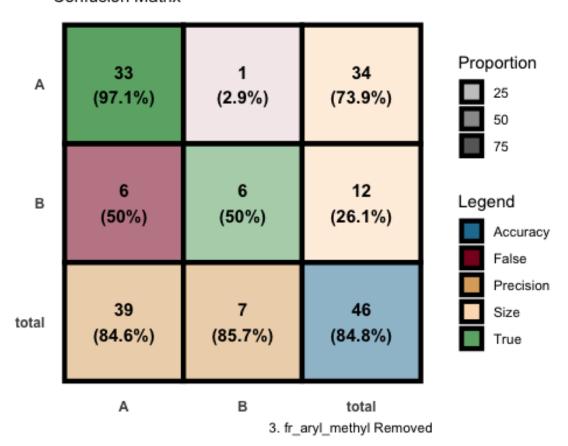
class ~ Avg\_NPA\_SM + fr\_Ar\_N + fr\_benzene + fr\_halogen

Full Model Stats - Overall Accuracy and Pseudo-R2

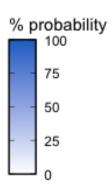
Accuracy McFadden\_R2 84.78% 0.36

	X
(Intercept)	18.3982999
Avg_NPA_SM	18.5844081
fr_Ar_N1	-17.1017267
fr_Ar_N2	-29.0614501
fr_Ar_N3	-32.9558398
fr_benzene1	-0.9195207
fr_benzene2	-27.8437909
fr_halogen1	0.4496324
fr_halogen2	1.4880482
fr_halogen3	1.1965587
fr_halogen4	-1.7308296
fr_halogen5	2.0846872
fr_halogen6	-9.0103971

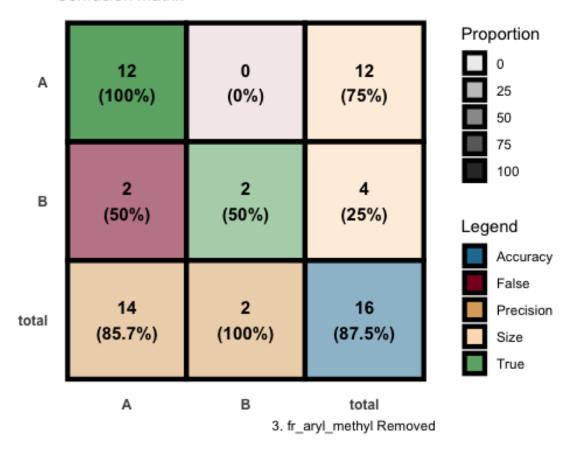
Training Set Confusion Matrix



	Α	В	Exp
Α	67	33	Α
AF	63	37	Α
F	100	0	Α
AB	79	21	Α
AC	91	9	Α
AD	100	0	Α
AE	100	0	Α
AH	65	35	Α
ΑI	100	0	Α
AJ	100	0	Α
AK	68	32	Α
AL	54	46	Α
AM	68	32 21 22 23 56	Α
G	79	21	Α
- 1	78	22	Α
J	77	23	Α
K	44	56	Α
L	79	21	Α
M	86	14	Α
N	100	0	Α
0	84	16	Α
Р	100	0	A
Q	100	0	Α
R T U	100	0	Α
Τ	100	0	A
U	100	0	Α
V W	90	10	A
VV	80	20	A
Х	100	0	A
Y	77	23	A
Z	78	22 46	A
AN	54	46	A
AO	81	19	A
AS	100	0	A
AV AW	68	32	B B
AX	69 35	31 65	В
AY	56	44	В
			В
BB BE	<u>51</u> 0	49 100	В
BF	0		В
BG	46	100 54	В
BH	46	54	В
ВІ	88	12	В
BJ	6	94	В
BK	92	8	В
DI	32	0	В



Test Set Confusion Matrix



Test Set Probability Heatmap

	Α	В	Exp	
В	84	16	A	
С	100	0	A	
D	100	0	A	
Е	100	0	A	
AA	81	19	A	0/ probability
Н	69	31	A	% probability
S	100	0	A	
AP	91	9	A	75
AQ	87	13	A	50
AR	100	0	A	- 25
AT	100	0	A	
AU	91	9	A	<b>□</b> 0
AZ	70	30	В	
BA	7	93	В	
вс	79	21	В	
BD	45	55	В	

3. fr\_aryl\_methyl Removed

### 4. fr\_benzene Removed

### formula

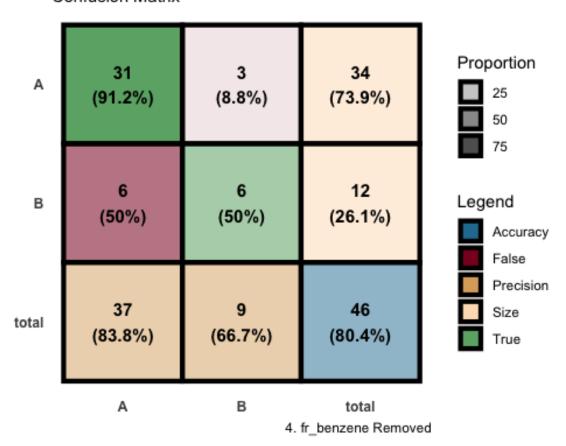
 ${\tt class} \sim {\tt Avg\_NPA\_SM} + {\tt fr\_Ar\_N} + {\tt fr\_aryl\_methyl} + {\tt fr\_halogen}$ 

Full Model Stats - Overall Accuracy and Pseudo-R2

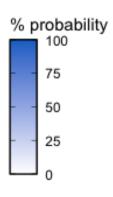
Accuracy McFadden\_R2 80.43% 0.404

	X
(Intercept)	15.3364807
Avg_NPA_SM	19.3241422
fr_Ar_N1	-13.7390649
fr_Ar_N2	-26.0021302
fr_Ar_N3	-34.6453926
fr_aryl_methyl1	-2.5523230
fr_aryl_methyl2	-17.8132254
fr_halogen1	1.0049517
fr_halogen2	0.4974903
fr_halogen3	0.2314174
fr_halogen4	-2.3972048
fr_halogen5	0.9617090
fr_halogen6	-8.0304964

Training Set Confusion Matrix

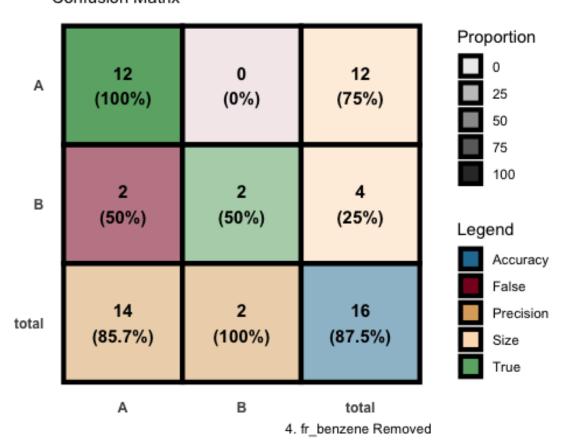


	Α	В	Exp
Α	100	0	Α
AF	91	9	Α
F	100	0	Α
AB	76	24	Α
AC	76	24	Α
AD	31	24 24 69	Α
AE	100	0	Α
AH	92	8	A
AI	100	0	A
AJ	100	0	A
AK	96	4 7	A
AL	93	/	A
AM	96 76	4	A A
G	76	24 60 7 5 59 34	A
J	40	7	A
K	93 95	5	A
1	41	50	A
K L M	41 66	34	A
N	100	0	A
Ö	63	37	A
P	100	0	A
Q.	100	Ö	A
Q R T U V	100	0	Α
Т	100	0	Α
U	100	0	Α
V	100 75	25 10	Α
W	90	10	Α
Х	100	0 11	Α
Υ	89	11	Α
Z	89	11	Α
AN	55 78	45 22 35 37	A
AO	78	22	Α
AS	65	35	Α
AV	63		В
AW	64	36	В
AX	75	25	В
AY	74	26	В
BB	32	68	В
BE	0	100	В
BF BG	0 42	100 58	B B
BH	46		В
BI	58	54 42	В
BJ	2	98	В
BK	78		В
DI	10	22	В



4. fr\_benzene Removed

Test Set Confusion Matrix



Test Set Probability Heatmap

	Α	В	Exp	
В	62	38	A	
С	100	0	A	
D	100	0	A	
Е	100	0	A	
AA	78	22	A	% probability
Н	64	36	A	% probability
S	100	0	A	75
AP	77	23	A	75
AQ	56	44	A	50
AR	76	24	A	- 25
AT	100	0	A	
AU	76	24	A	<b>□</b> 0
ΑZ	65	35	В	
BA	13	87	В	
вс	75	25	В	
BD	39	61	В	

4. fr\_benzene Removed

### 5. fr\_halogen Removed

### formula

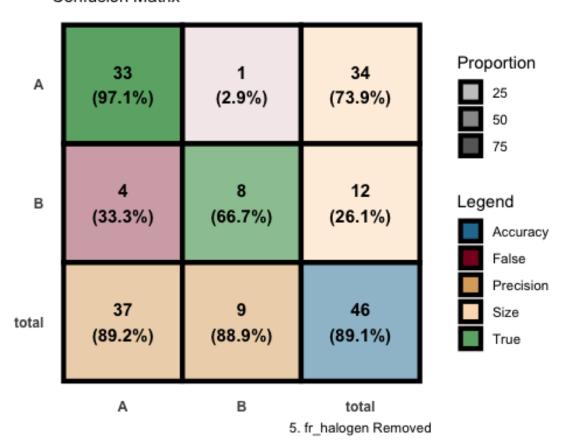
class ~ Avg\_NPA\_SM + fr\_Ar\_N + fr\_aryl\_methyl + fr\_benzene

Full Model Stats - Overall Accuracy and Pseudo-R2

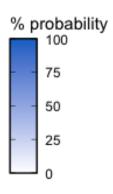
Accuracy McFadden\_R2 89.13% 0.512

	X
(Intercept)	43.957001
Avg_NPA_SM	49.387661
fr_Ar_N1	-37.097092
fr_Ar_N2	-69.144375
fr_Ar_N3	-69.553894
fr_aryl_methyl1	-3.779621
fr_aryl_methyl2	-51.951799
fr_benzene1	-1.330087
fr_benzene2	-68.569342

Training Set Confusion Matrix

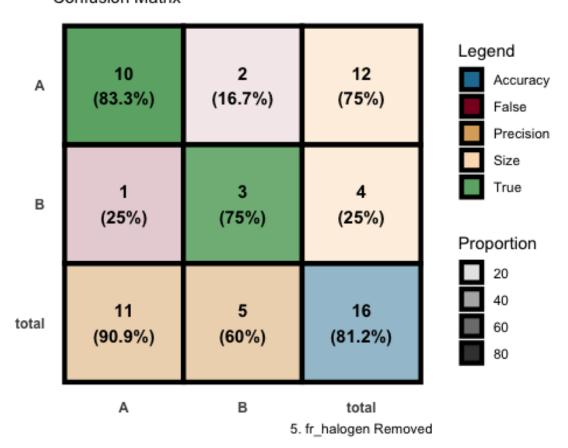


	Α	В	Exp
Α	100	0	A
AF	95	5	Α.
F	100	0 46	Α.
AB	54 82 100	46	A A A A A
AC	82	18	A
AD	100	0 0 4 0 0 8 3 8 46 47 7 5 45	Α
AE	100	0	Α.
AH	96	4	
AH AI AJ	100 100 92 97	0	
AK	100	0	Ä
AL	92	9	^
AM	92	8	
G	92 54 53 93 95	46	A A A
	53	47	Α.
J	93	7	А
K	95	5	Α.
K L	55	45	Α.
M	57	43	A
N	100	0	A
0	55 57 100 47	0 53	Α.
Р	100	0	A
P Q R T U V	100	0	A A A A A A
R	100 100	0 0 1	A
I	100	0	A
U	99	7	A
V M/	79	21	A A
V	98 100	2	_ ^
Ŷ	98	2	
X Y Z	98	2	A
AN	81	19	A
AO	98 81 62	21 2 0 2 2 2 19 38	A
AS	100	0	A
ΑV	19	81	В
AW	22	78	В
AX	48	52	В
AY	61	39	В
BB	10	90	В
BE	0	100	В
BF	0	100	В
BG	27	73	В
BH BI	63	37	B B
BJ	87	13 100	В
BK	0 86	14	В
DK	00	14	, в



5. fr\_halogen Removed

Test Set Confusion Matrix



Test Set Probability Heatmap

	Α	В	Exp	
В	46	54	A	
С	100	0	A	
D	100	0	A	
Е	100	0	A	
AA	62	38	A	9/ probability
Н	22	78	A	% probability
S	100	0	A	75
AP	84	16	A	75
AQ	85	15	A	50
AR	100	0	A	25
AT	100	0	A	
AU	83	17	A	<b>□</b> 0
AZ	24	76	В	
BA	0	100	В	
вс	52	48	В	
BD	2	98	В	

5. fr\_halogen Removed

### **Section Results Summary**

	Training Accuracy	Test Accuracy
5. 5th Place with 5 features	91.30	87.5
1. Avg_NPA_SM Removed	84.78	87.5
2. fr_Ar_N Removed	84.78	75.0
3. fr_aryl_methyl Removed	84.78	87.5
4. fr_benzene Removed	80.43	87.5
5. fr halogen Removed	89.13	81.2