Comparative study of PSSM, ANN, SMM for peptide MHC binding

Contents

Dataset statistics **1**
No. binders vs size . . . 2

PSSM **3**
Assessing optimal beta . . . 4

SMM **6**
SMM: Gradient descent . . . 6
Optimising lambda . . . 8
Assessment of lambda . . . 9
SMM: Monte Carlo . . . 10

ANN **11**
BLOSUM vs Sparse encoding . . . 13

Comparing all results **14**
Average PCC . . . 14
Max PCC . . . 15
Negative PCC values . . . 15
Visualise performance for ALL methods . . . 16
Visualising PCC vs Size and Number of binders for all methods . . . 18

Dataset statistics

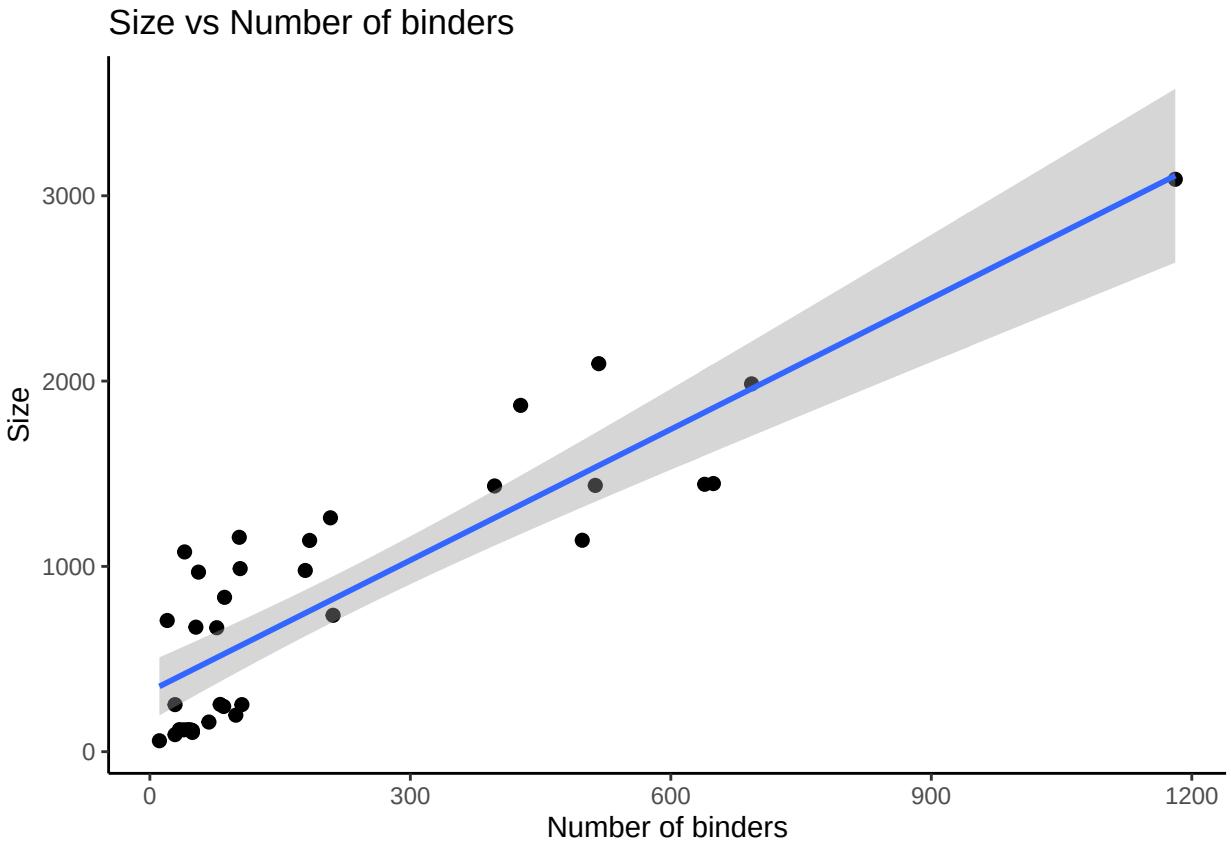
```
##    Allele Size Number of binders
## 1   A0101 1157               103
## 2   A0201 3089              1181
## 3   A0202 1447               649
## 4   A0203 1443               639
## 5   A0206 1437               513
## 6   A0301 2094               517
## 7   A1101 1985               693
## 8   A2301  104                49
## 9   A2402  197                99
## 10  A2403  254                29
## 11  A2601  672                53
## 12  A2902  160                68
## 13  A3001  669                77
## 14  A3002   92                29
## 15  A3101 1869               427
## 16  A3301 1140               184
## 17  A6801 1141               498
## 18  A6802 1434               397
## 19  A6901  833                86
```

```
## 20  B0702 1262              208
## 21  B0801  708               20
## 22  B1501  978              179
## 23  B1801  118               47
## 24  B2705  969               56
## 25  B3501  736              211
## 26  B4001 1078               40
## 27  B4002  118               39
## 28  B4402  119               44
## 29  B4403  119               34
## 30  B4501  114               49
## 31  B5101  244               85
## 32  B5301  254              106
## 33  B5401  255               81
## 34  B5701   59               11
## 35  B5801  988              104
```

```
## [1] 86
```

No. binders vs size

The number of binders clearly scales linearly with the size of the dataset.

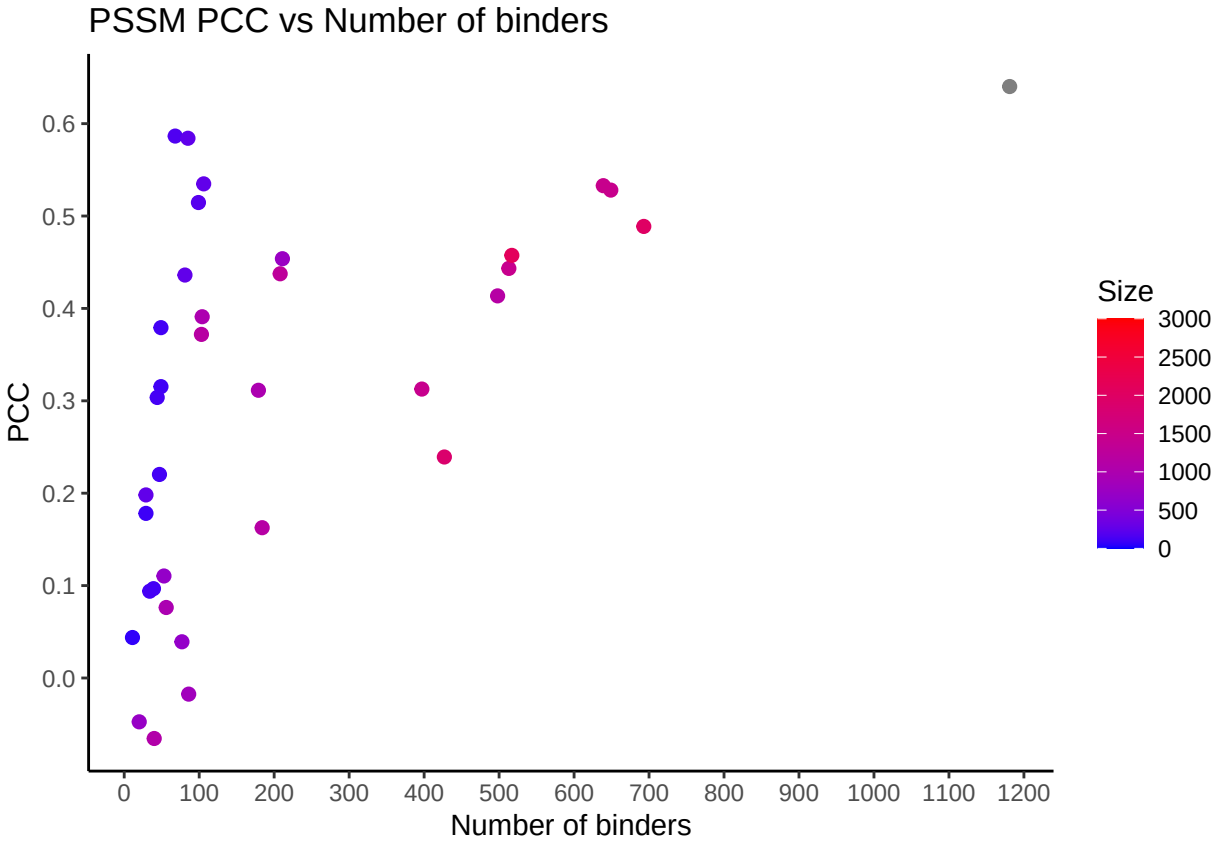


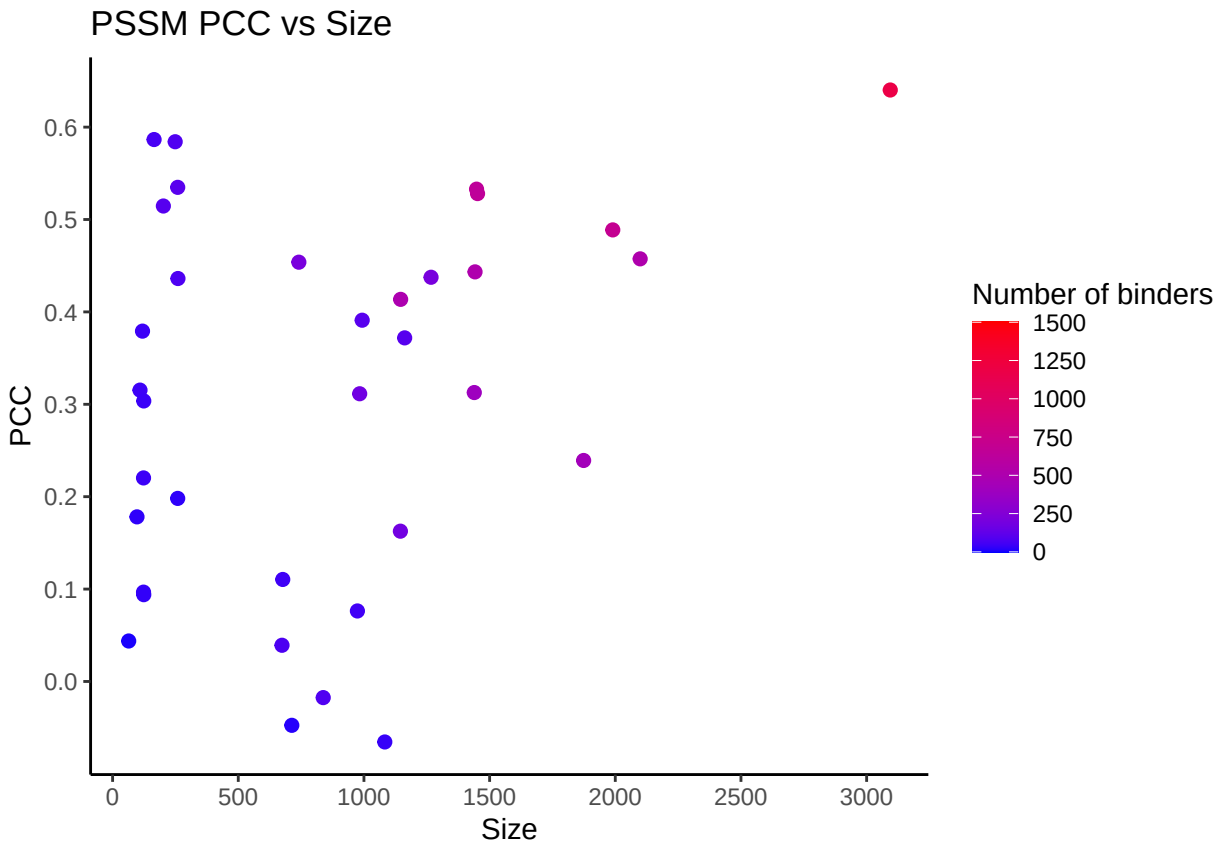
PCC Size vs no. binders

```
## [1] 0.8714809
```

PSSM

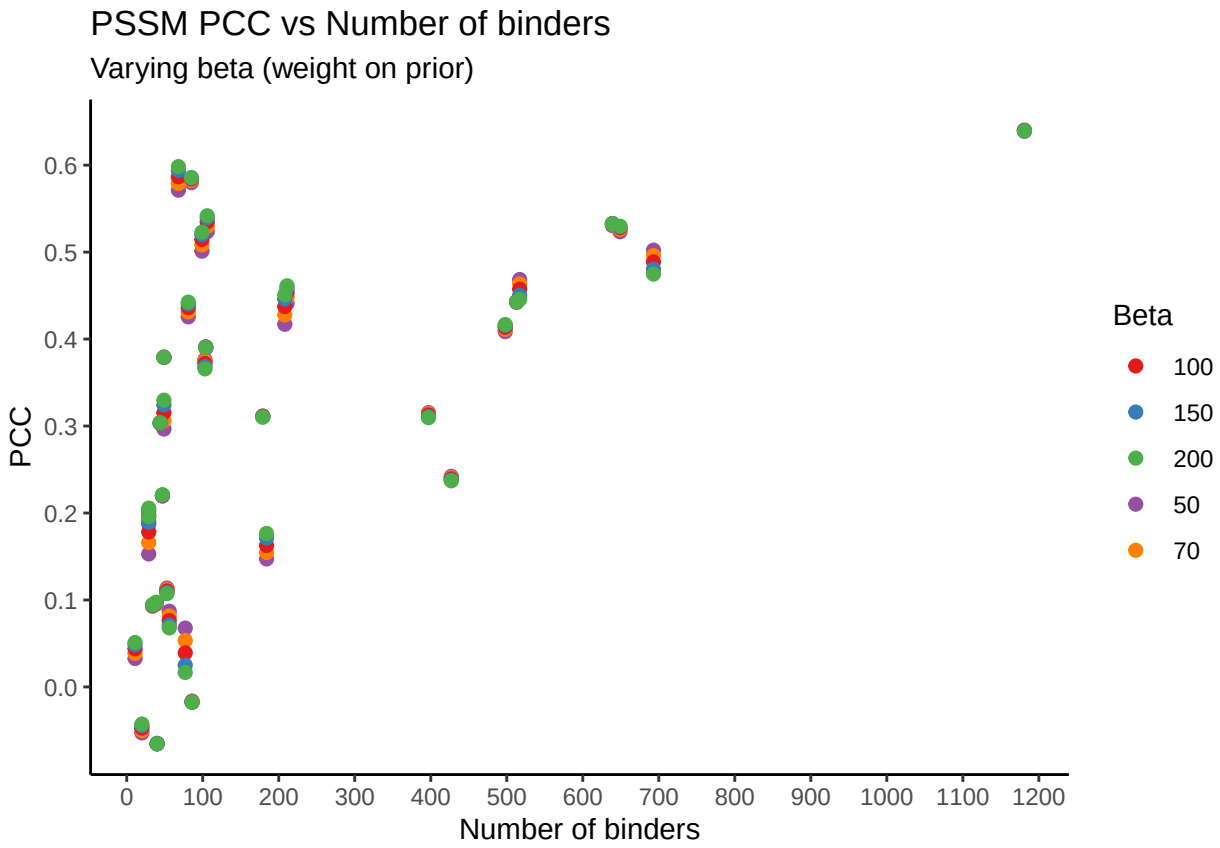
Beta = 100, weighting on

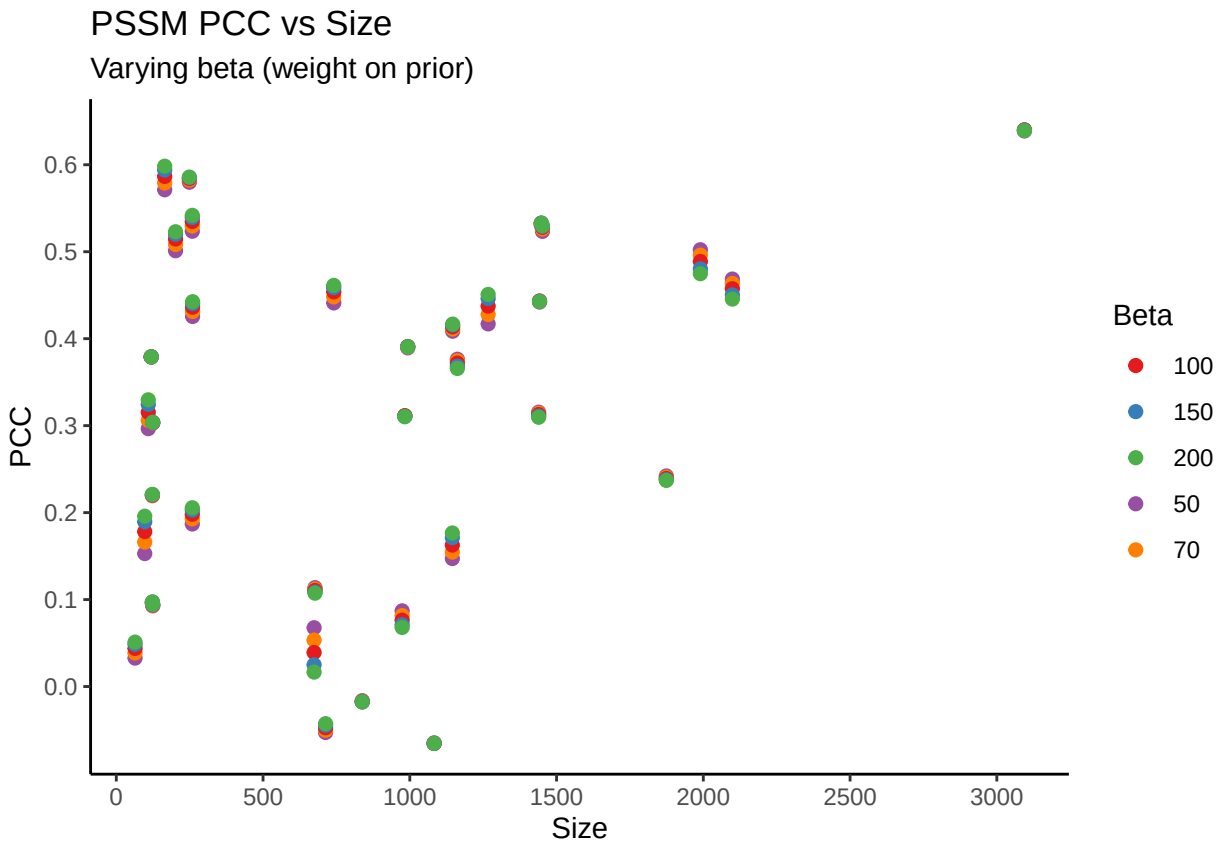




Assessing optimal beta

From the graphs, PCC performance between beta values is minimal.

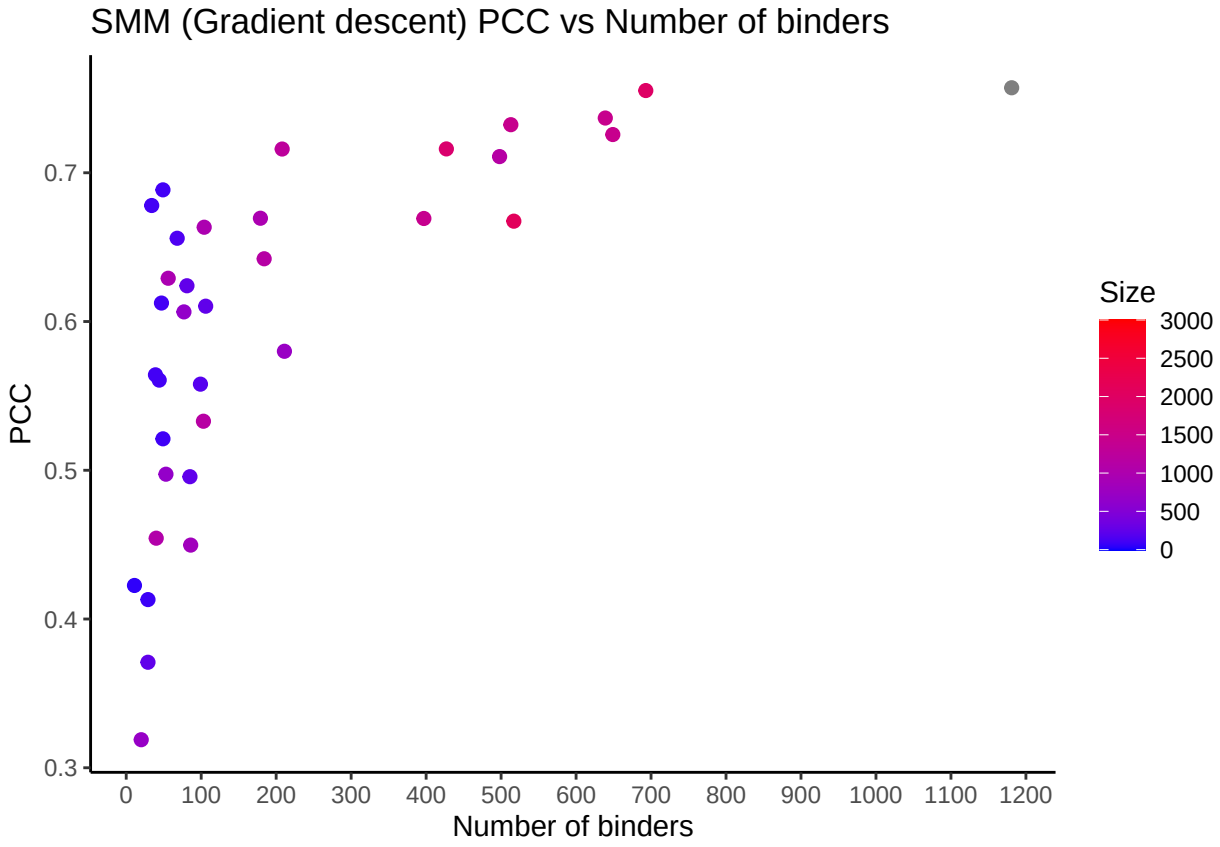

PSSM PCC vs Number of binders
Varying beta (weight on prior)
PCC
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
Number of binders
Beta
100
150
200
50
70

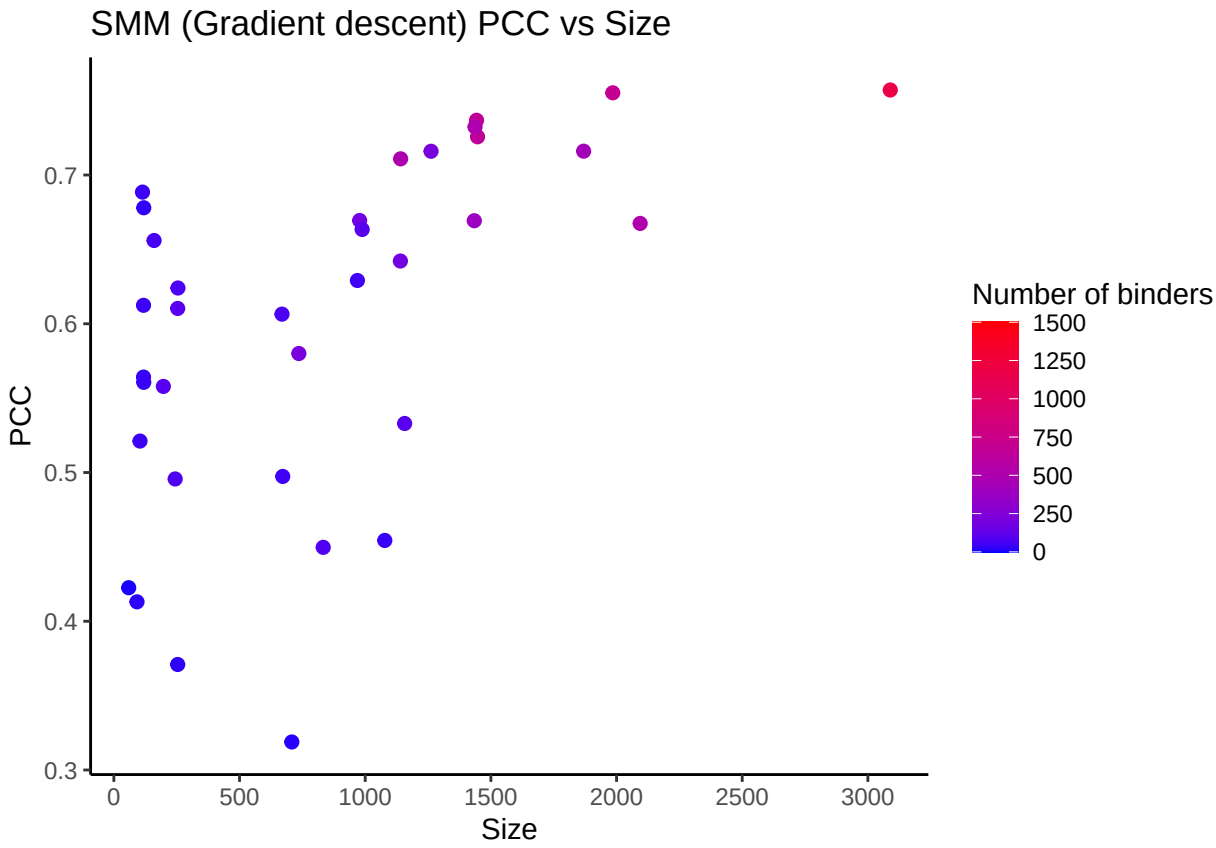

PSSM PCC vs Size
Varying beta (weight on prior)
PCC
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
500
1000
1500
2000
2500
3000
Size
Beta
100
150
200
50
70

SMM

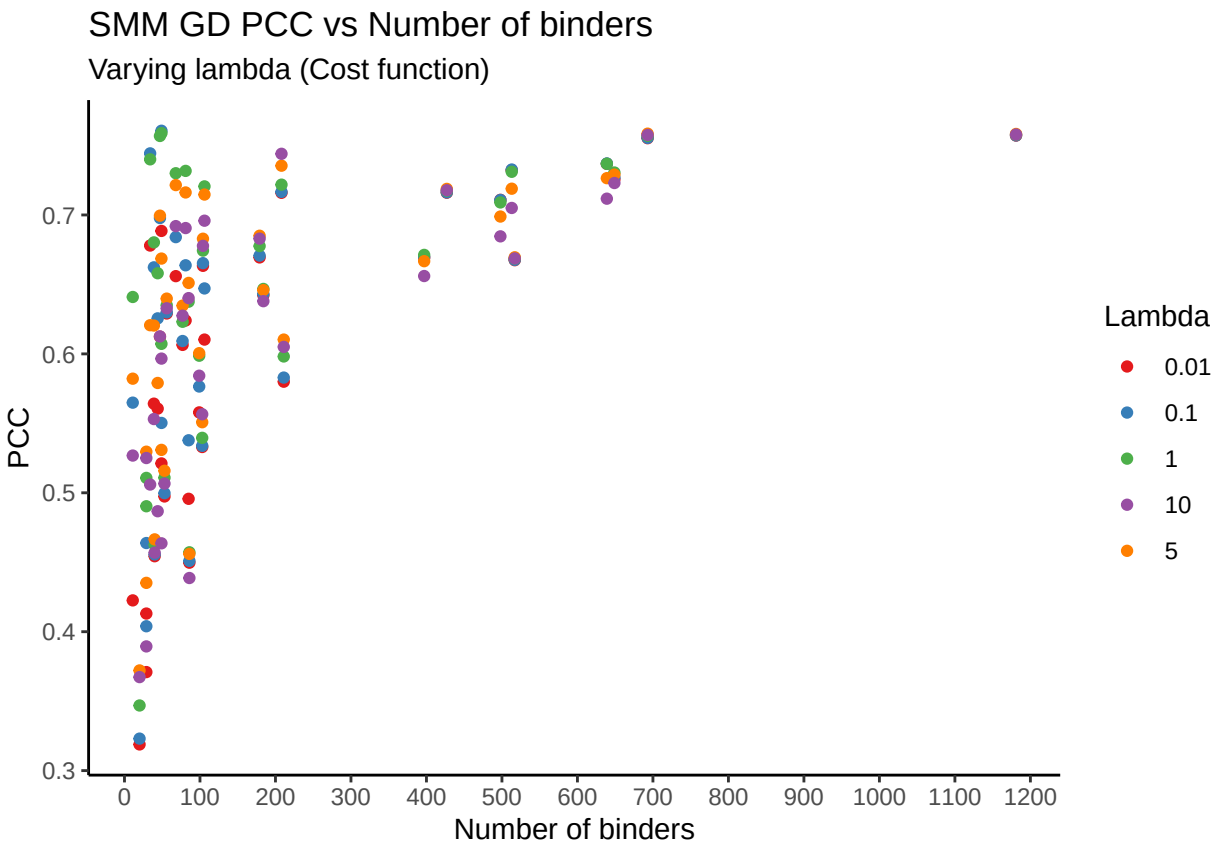
SMM: Gradient descent

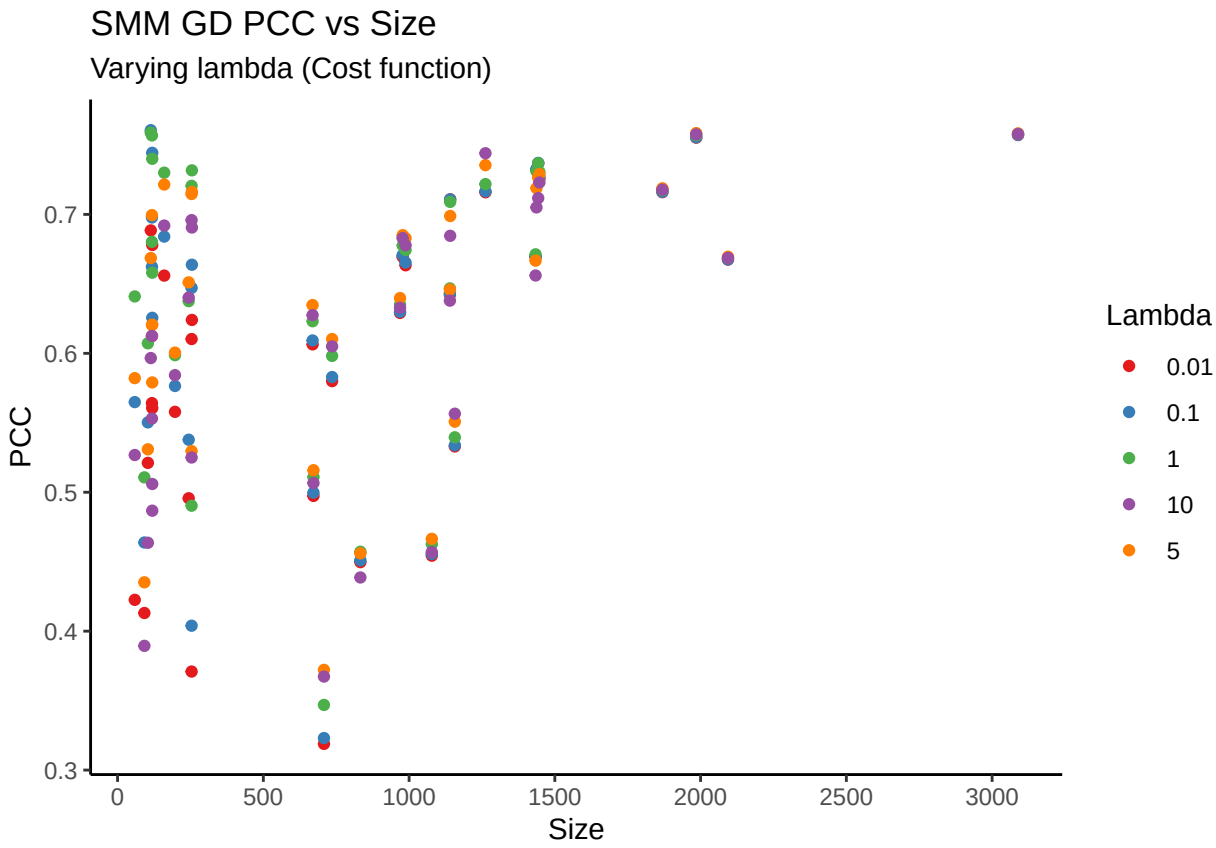
Lambda = 0.01, epsilon = 0.05, epochs = 100 (doesn't make a difference). ***Lambda = 1 gave optimum results, so those values are used in the final comparison***




SMM (Gradient descent) PCC vs Size
PCC
0.7
0.6
0.5
0.4
0.3
0
500
1000
1500
2000
2500
3000
Size
Number of binders
1500
1250
1000
750
500
250
0

Optimising lambda





Assessment of lambda

1 has best overall performance - by graph, and it has highest average.

```
## # A tibble: 1 x 1
##   `Lambda = 0.01 mean PCC`
##                      <dbl>
## 1                    0.600
```

```
## # A tibble: 1 x 1
##   `Lambda = 0.1 mean PCC`
##                     <dbl>
## 1                   0.624
```

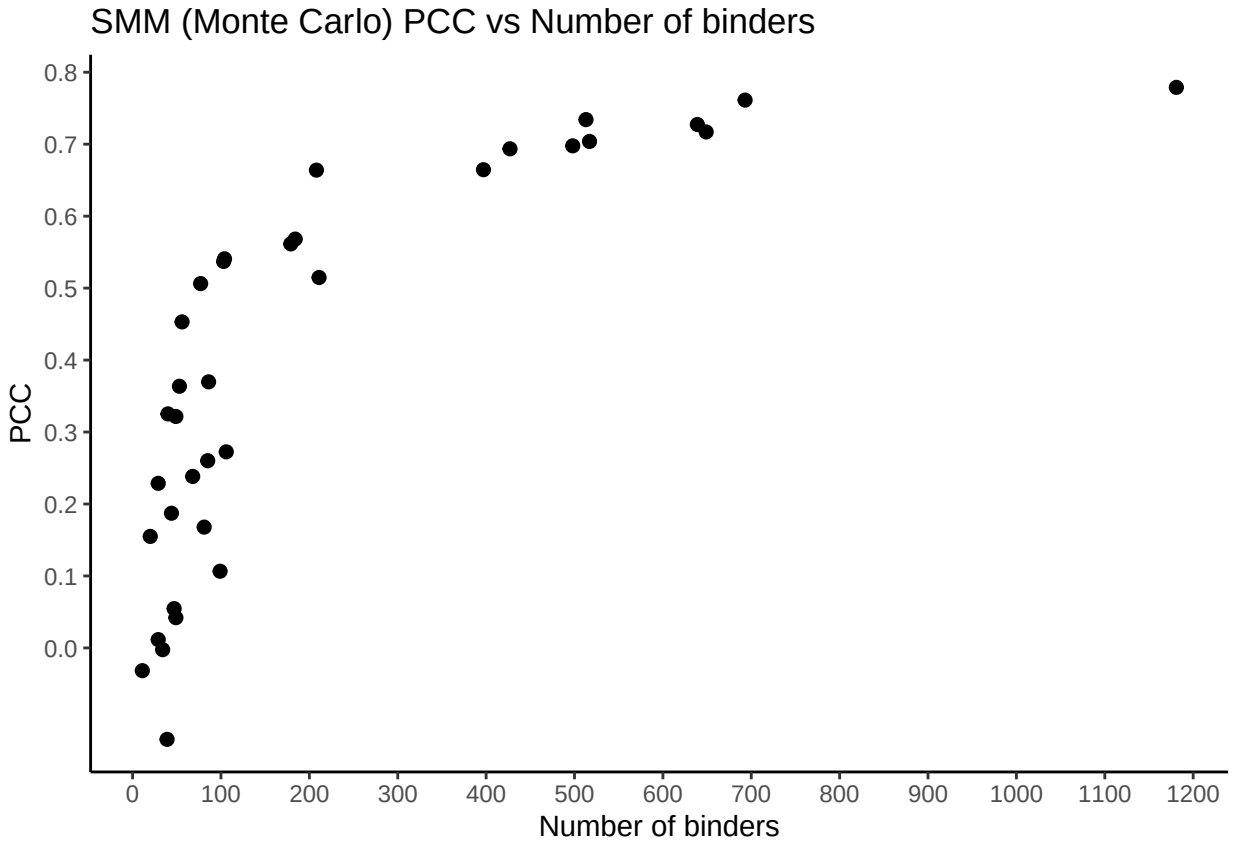
```
## # A tibble: 1 x 1
##   `Lambda = 1 mean PCC`
##                   <dbl>
## 1                 0.647
```

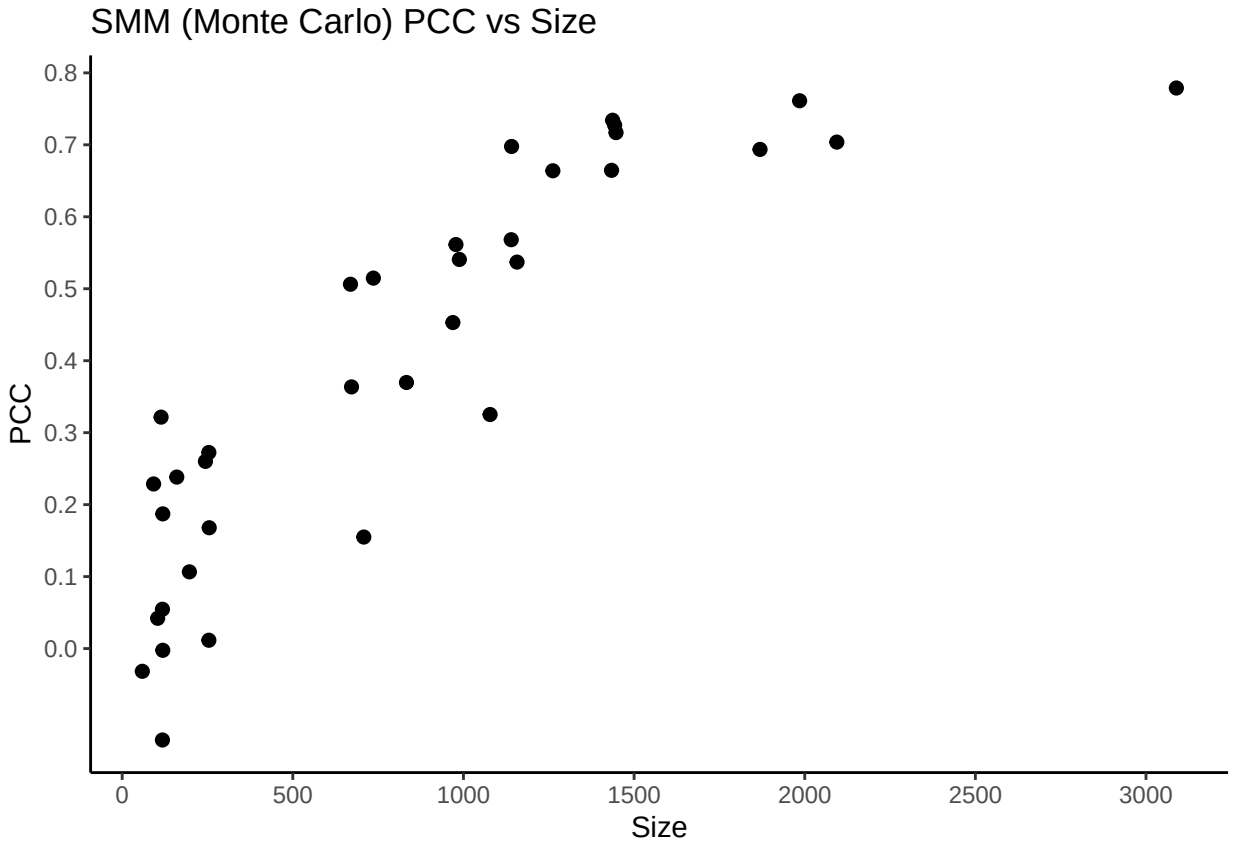
```
## # A tibble: 1 x 1
##   `Lambda = 5 mean PCC`
##                   <dbl>
## 1                 0.632
```

```
## # A tibble: 1 x 1
##   `Lambda = 10 mean PCC`
##                    <dbl>
## 1                  0.608
```

SMM: Monte Carlo

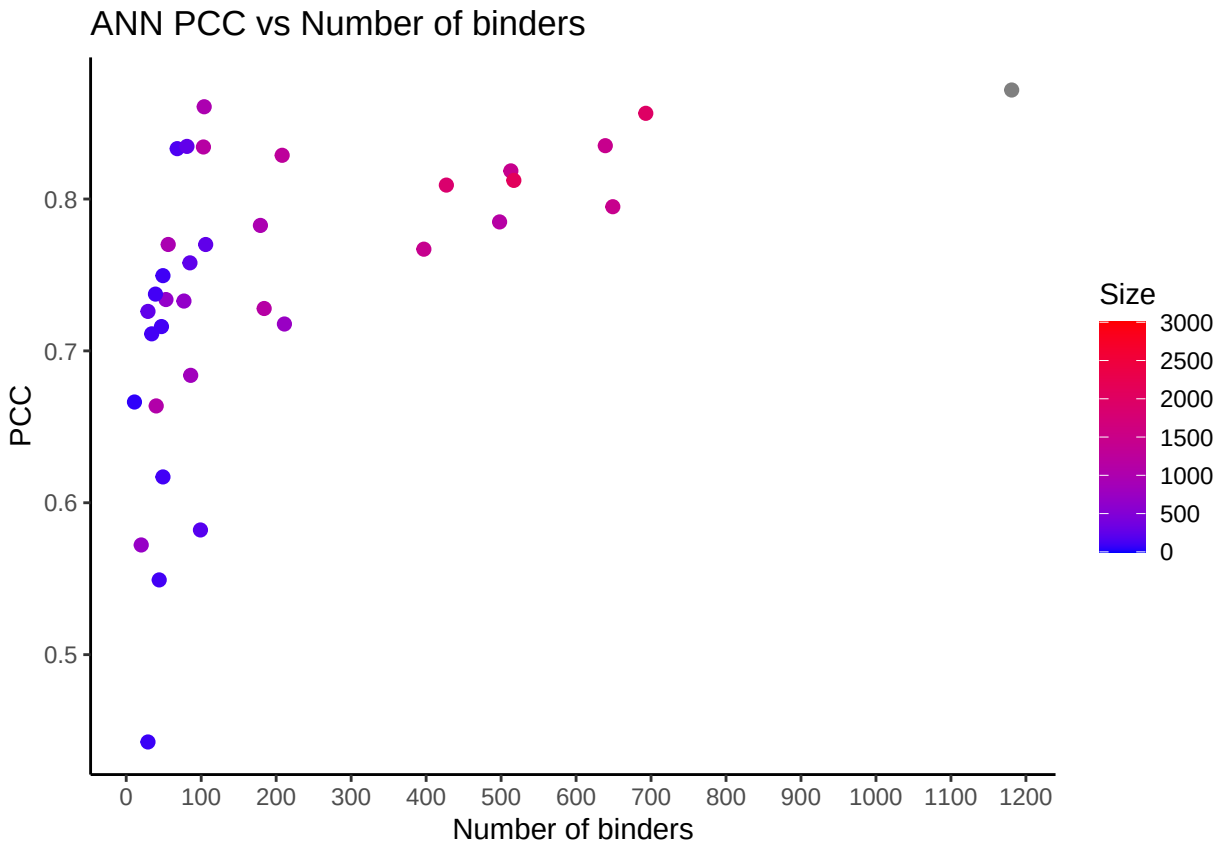
Lambda = 0.01, Epochs = 1000 ; no optimisations made

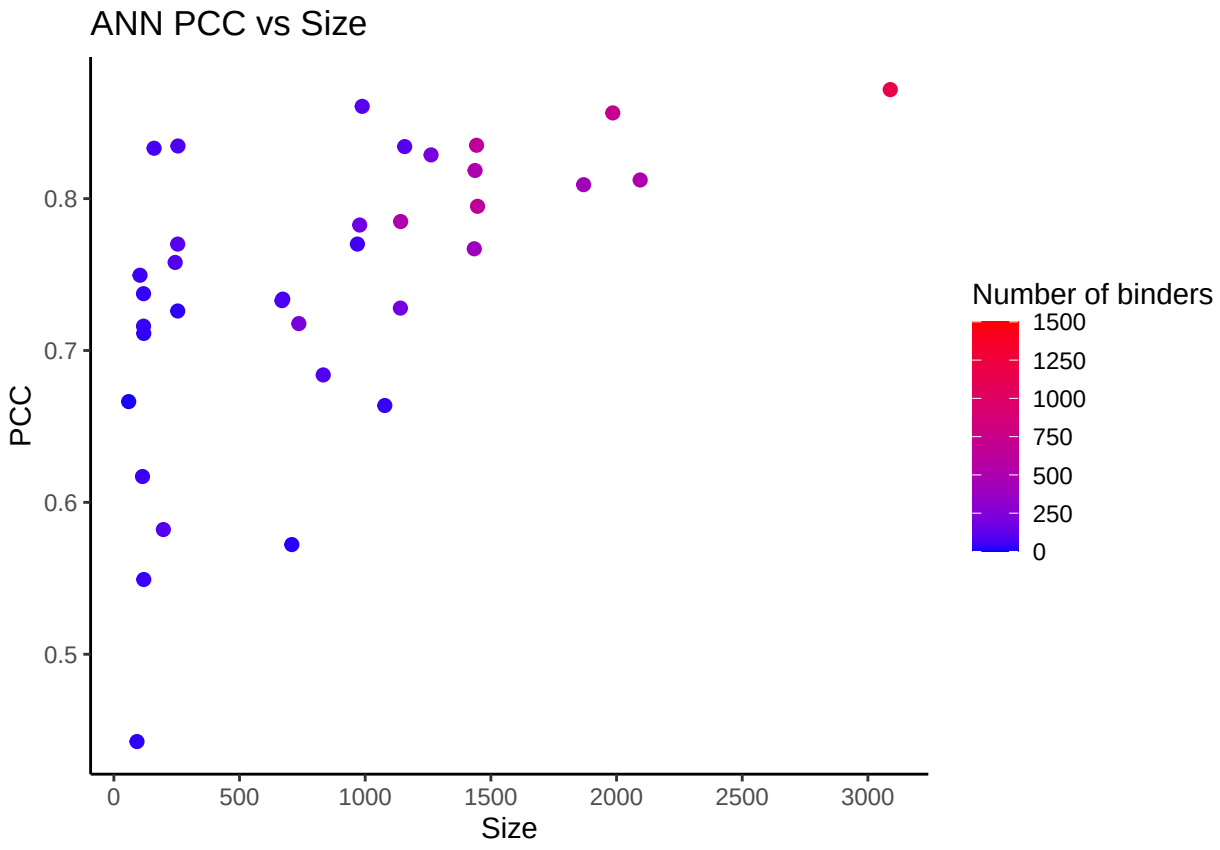




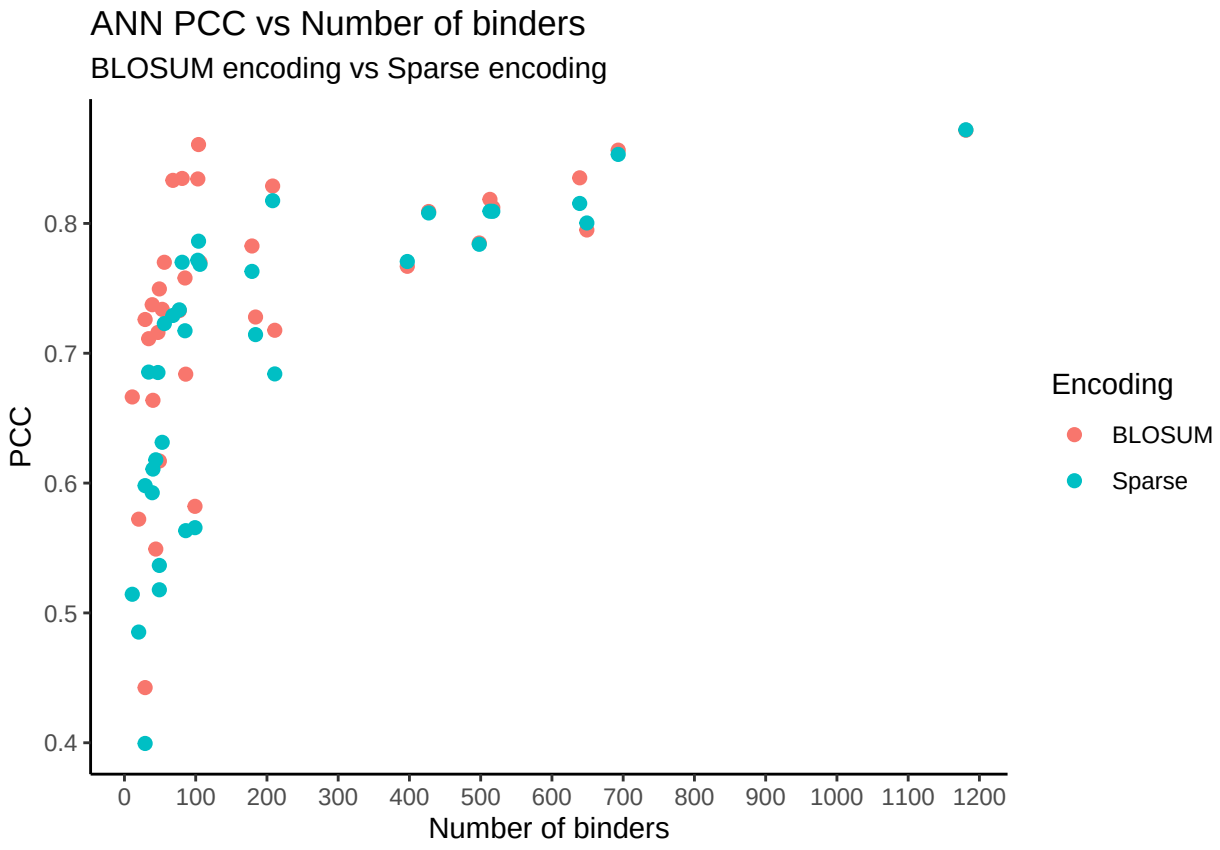
ANN

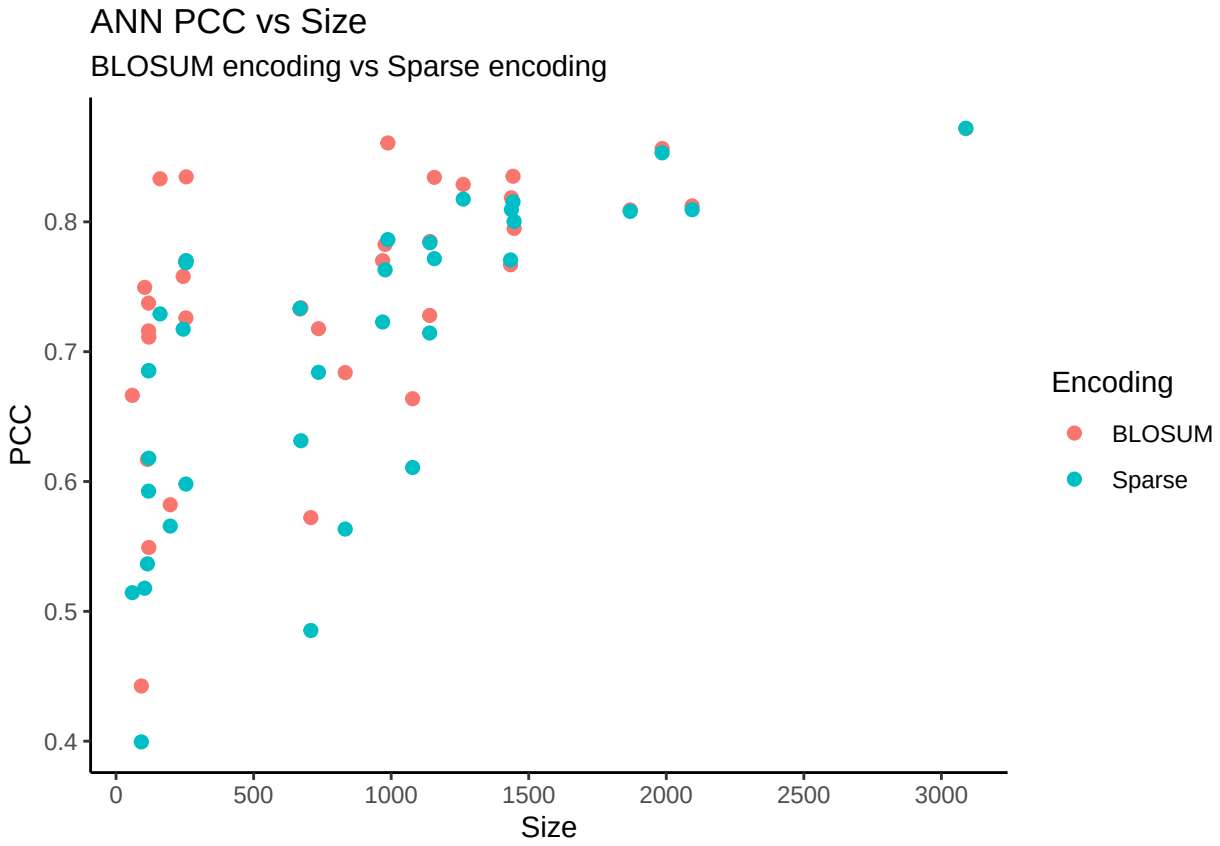
Originally run with BLOSUM encoding - sparse coding was later found to be inferior, so the final comparisons are used with BLOSUM encoding


ANN PCC vs Number of binders
PCC
0.8
0.7
0.6
0.5
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
Number of binders
Size
3000
2500
2000
1500
1000
500
0


ANN PCC vs Size
PCC
0.8
0.7
0.6
0.5
0
500
1000
1500
2000
2500
3000
Size
Number of binders
1500
1250
1000
750
500
250
0

BLOSUM vs Sparse encoding






```
ANN_original %>% select(PCC) %>% summarise(`BLOSUM mean PCC` = mean(PCC)) %>% round(3)
```

```
## # A tibble: 1 x 1
##   `BLOSUM mean PCC`
##               <dbl>
## 1             0.741
```

```
ANN_sparse %>% select(PCC) %>% summarise(`sparse mean PCC` = mean(PCC)) %>% round(3)
```

```
## # A tibble: 1 x 1
##   `sparse mean PCC`
##               <dbl>
## 1             0.694
```

Comparing all results

ALL results are compared using the optimal parameters.

PSSM: beta = 100 - beta had no significant on performance, so we just use the original beta value **SMM GD: lambda = 1, epsilon = 0.05, epochs = 100** - optimised, from lambda 0.01 originally **SMM MC: lambda = 0.01, epochs = 1000 ANN: BLOSUM encoding**

Average PCC

```
## # A tibble: 1 x 4
##   `PSSM mean PCC` `GD mean PCC` `MC mean PCC` `ANN mean PCC`
##             <dbl>         <dbl>         <dbl>          <dbl>
## 1           0.308         0.647         0.393          0.741
```

Max PCC

```
## # A tibble: 1 x 4
##   `PSSM max PCC` `GD max PCC` `MC max PCC` `ANN max PCC`
##            <dbl>        <dbl>        <dbl>         <dbl>
## 1           0.64        0.759        0.779         0.872
```

Negative PCC values

3 negative values in PSSM and MC. Interestingly enough, there are 3 different allele in each case.

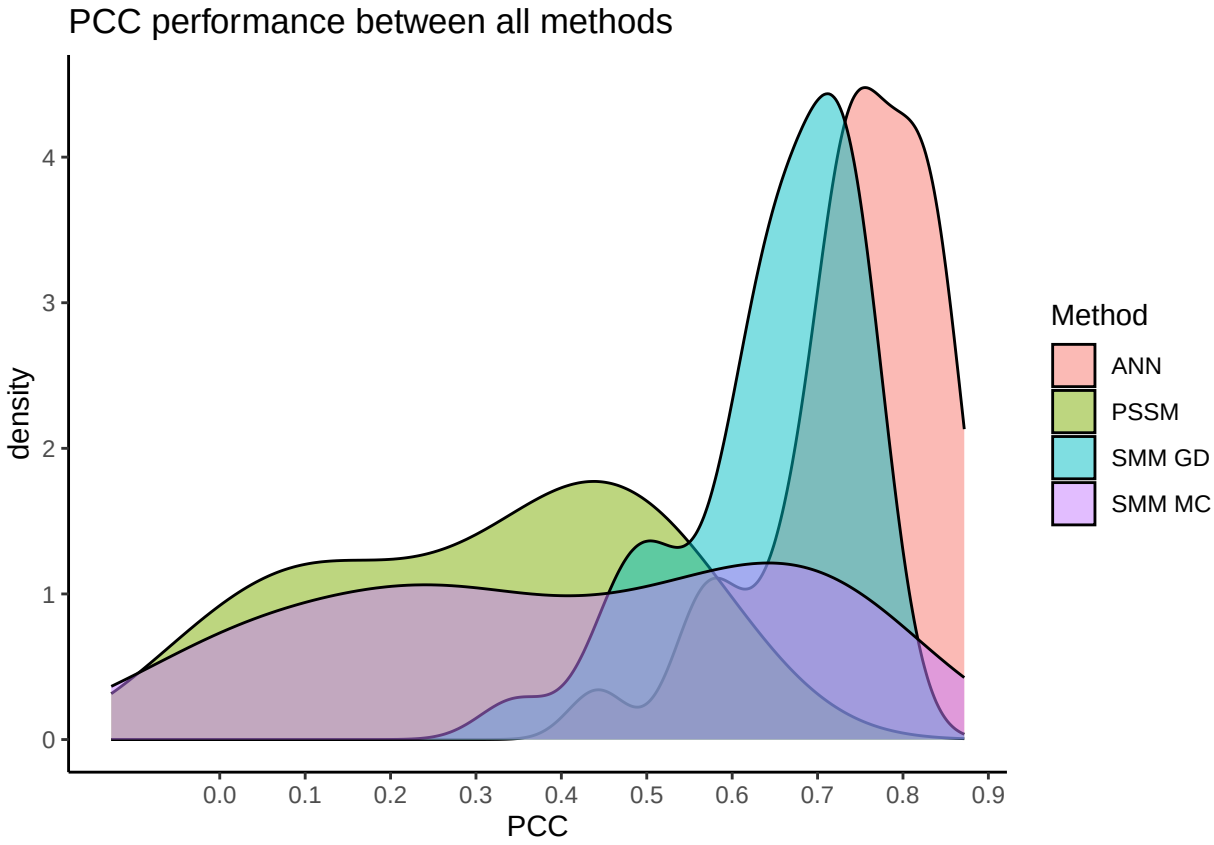
PSSM

```
## # A tibble: 3 x 4
##   Allele  Size     PCC `Number of binders`
##   <chr>  <dbl>   <dbl>               <int>
## 1 A6901    838 -0.0175                86
## 2 B0801    713 -0.0475                20
## 3 B4001   1083 -0.0655                40
```

MC

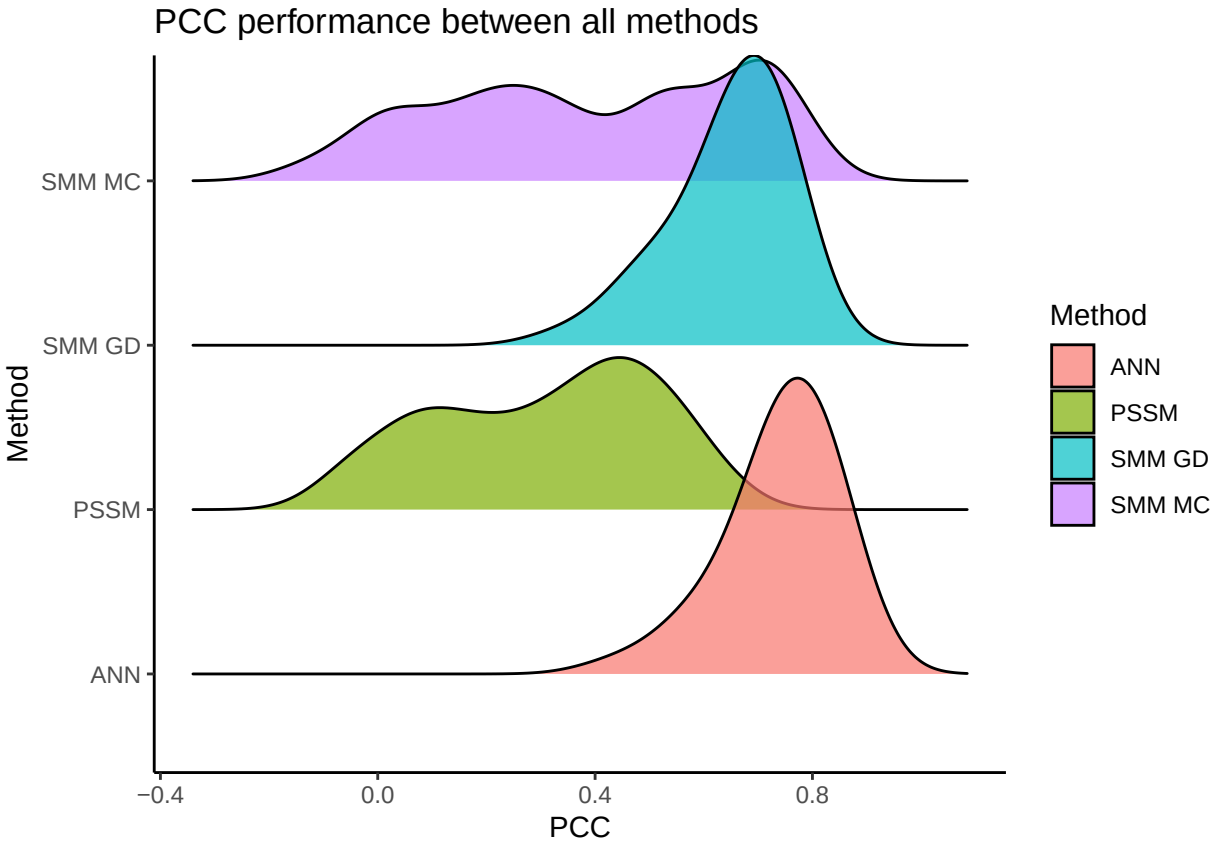
```
## # A tibble: 3 x 4
##   Allele  Size      PCC `Number of binders`
##   <chr>  <dbl>    <dbl>               <int>
## 1 B4002    118 -0.127                  39
## 2 B4403    119 -0.00242                34
## 3 B5701     59 -0.0316                 11
```

Visualise performance for ALL methods



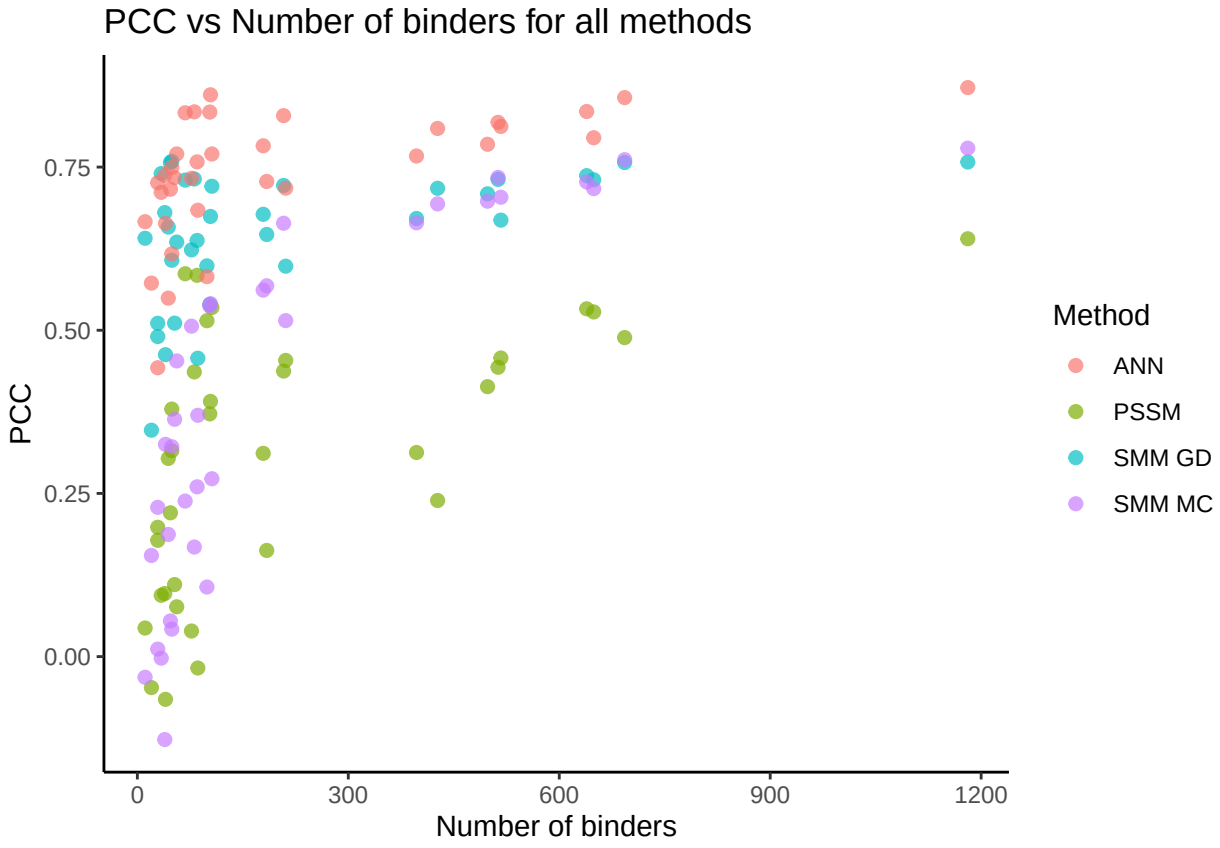
Alternate plot (ridgeline)

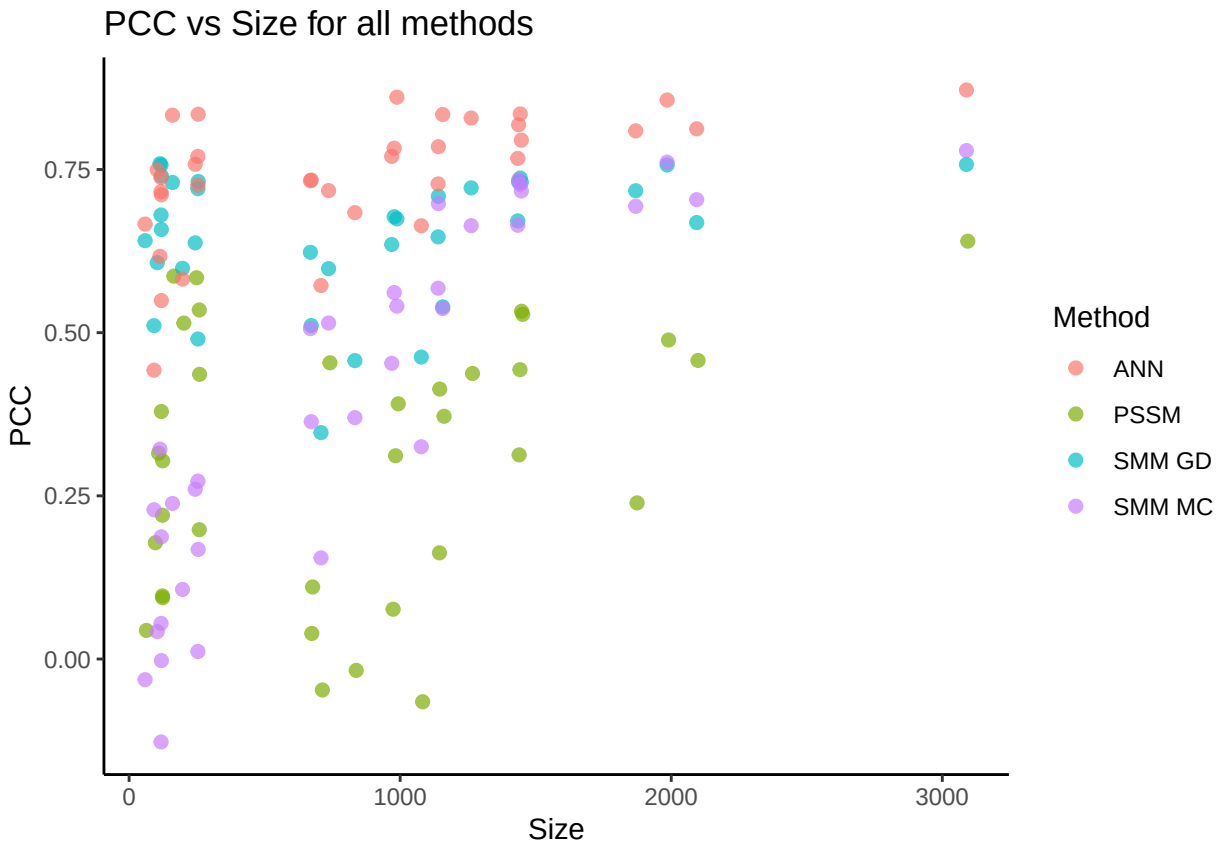
```
## Picking joint bandwidth of 0.071
```




```
## Picking joint bandwidth of 0.071
```

Visualising PCC vs Size and Number of binders for all methods





Exploring the size categories and average performance between them

```
##    Allele Size Number of binders Category
## 1   A0101 1157               103    Small
## 2   A2301  104                49    Small
## 3   A2402  197                99    Small
## 4   A2403  254                29    Small
## 5   A2601  672                53    Small
## 6   A2902  160                68    Small
## 7   A3001  669                77    Small
## 8   A3002   92                29    Small
## 9   A3301 1140               184    Small
## 10  A6901  833                86    Small
## 11  B0702 1262               208    Small
## 12  B0801  708                20    Small
## 13  B1501  978               179    Small
## 14  B1801  118                47    Small
## 15  B2705  969                56    Small
## 16  B3501  736               211    Small
## 17  B4001 1078                40    Small
## 18  B4002  118                39    Small
## 19  B4402  119                44    Small
## 20  B4403  119                34    Small
## 21  B4501  114                49    Small
## 22  B5101  244                85    Small
## 23  B5301  254               106    Small
## 24  B5401  255                81    Small
## 25  B5701   59                11    Small
```

```
## 26  B5801  988               104    Small
```

```
##   Allele Size Number of binders Category
## 1  A0202 1447               649   Medium
## 2  A0203 1443               639   Medium
## 3  A0206 1437               513   Medium
## 4  A0301 2094               517   Medium
## 5  A1101 1985               693   Medium
## 6  A3101 1869               427   Medium
## 7  A6801 1141               498   Medium
## 8  A6802 1434               397   Medium
```

```
## # A tibble: 1 x 1
##   `PSSM small`
##          <dbl>
## 1        0.258
```

```
## # A tibble: 1 x 1
##   `PSSM medium`
##           <dbl>
## 1         0.427
```

```
## # A tibble: 1 x 1
##   `SMM GD small`
##            <dbl>
## 1          0.621
```

```
## # A tibble: 1 x 1
##   `SMM GD medium`
##             <dbl>
## 1           0.715
```

```
## # A tibble: 1 x 1
##   `SMM MC small`
##            <dbl>
## 1          0.280
```

```
## # A tibble: 1 x 1
##   `SMM MC medium`
##             <dbl>
## 1           0.712
```

```
## # A tibble: 1 x 1
##   `ANN small`
##         <dbl>
## 1       0.715
```

```
## # A tibble: 1 x 1
##   `ANN medium`
##          <dbl>
## 1        0.810
```