Hashing Classification for charged particle tracking

Luiza Adelina Ciucu (ATLAS)

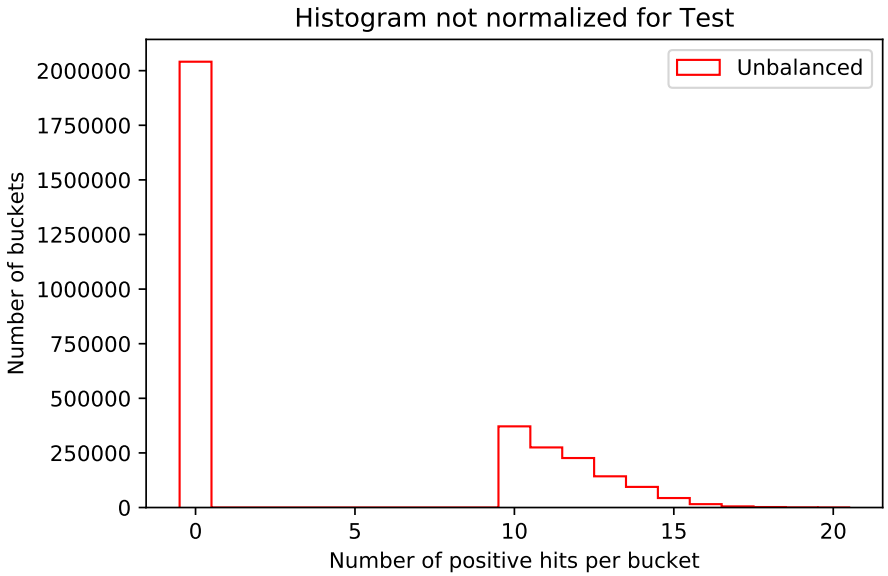
26 June 2020

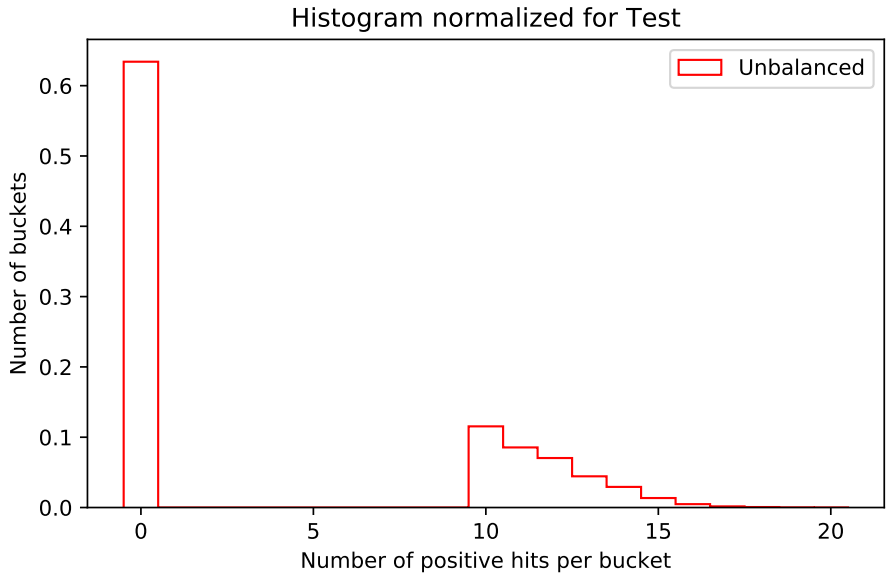
UNIVERSITÉ
DE GENÈVE

FACULTY OF SCIENCE
Physics Section

Introduction

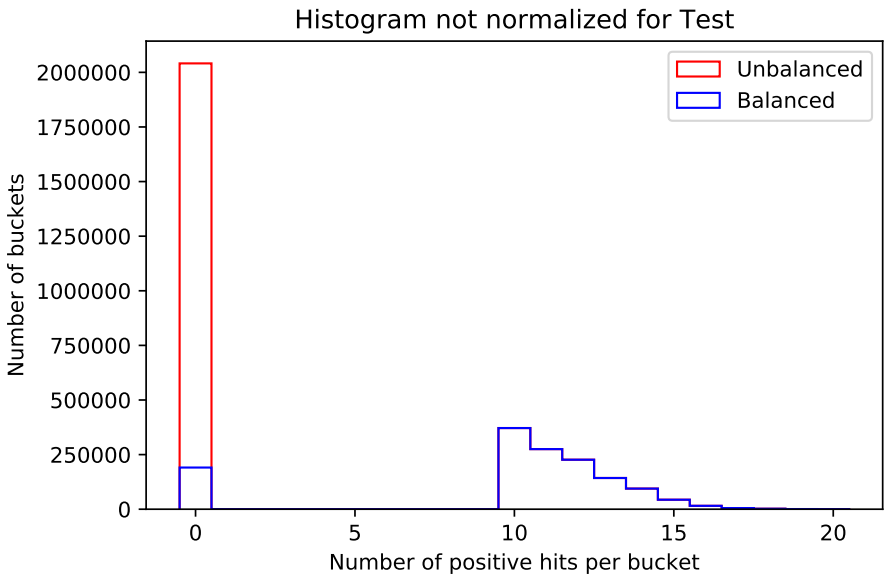
- 100 events. For each group of 10: 7 train, 3 test.
- If nbPositiveHit<10, set nbPositiveHit=0 and output made only of -1.
- Goal: balance nbPositiveHit and nbNegativeHit.
- Constraints:
  - Not reweighting at bucket level, but removing some bucktets.
  - Make distribution nbPositiveHit as flat as possible, to help training have almost equal nbBuckets for various nbPositiveHit.
- Method: remove buckets at nbPositiveHit=0, until nbPositiveHit=nbNegativeHit, solving a first degree equation.

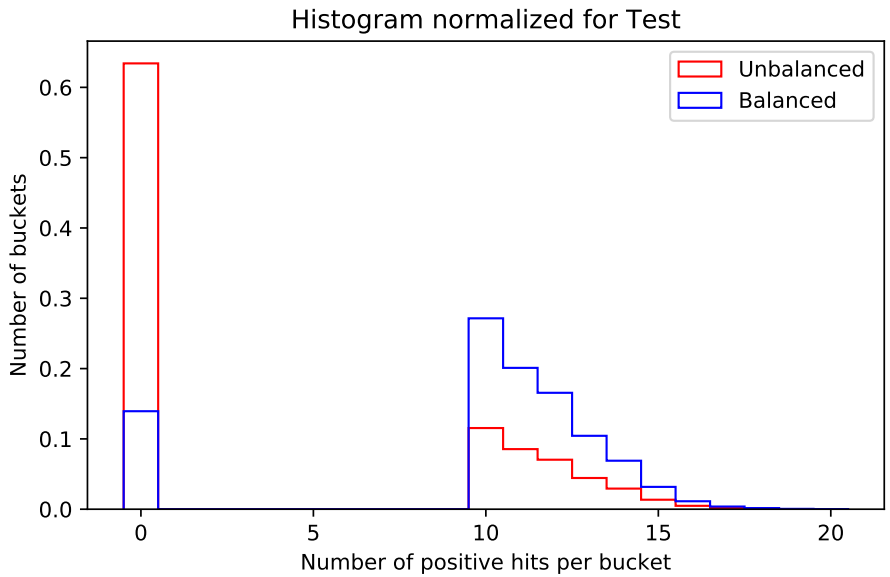




Balancing method 1/6

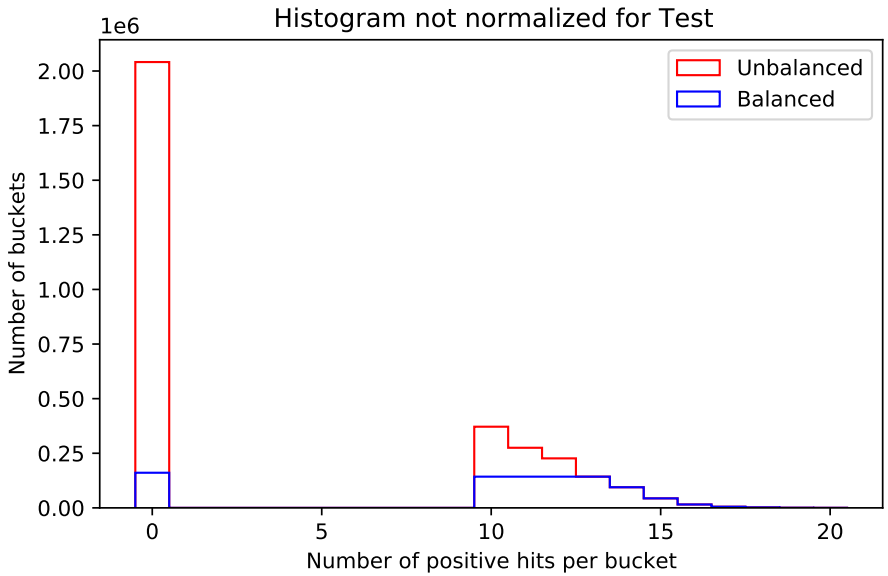
- Categories of nbPositiveHit from 0 to 20 denoted x_i.
- Each category has N_i buckets, representing the weight w_i of x_i.
- The weighted average of nbPositiveHit is desired to be 10, so:
  $$\frac{w_0 \cdot x_0 + \sum_1^{20} w_i \cdot x_i}{w_0 + \sum_1^{20} w_i} = 10$$
- Solve for w_0,
  $$w_0 = \frac{\sum_1^{20} w_i \cdot x_i}{10} - \sum_1^{20} w_i$$
- Reduce the number of buckets in nbPositiveHit=0 to w_0.

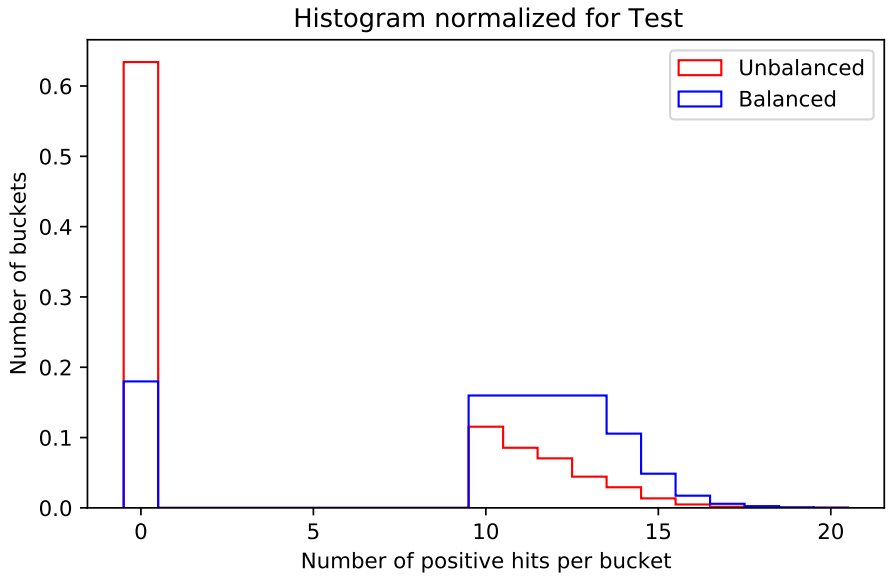




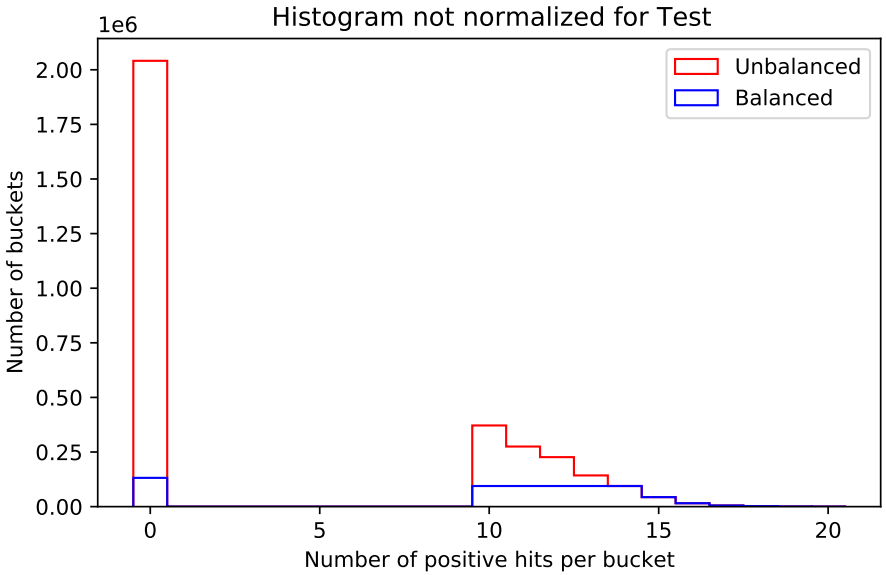
Balancing method 2/6

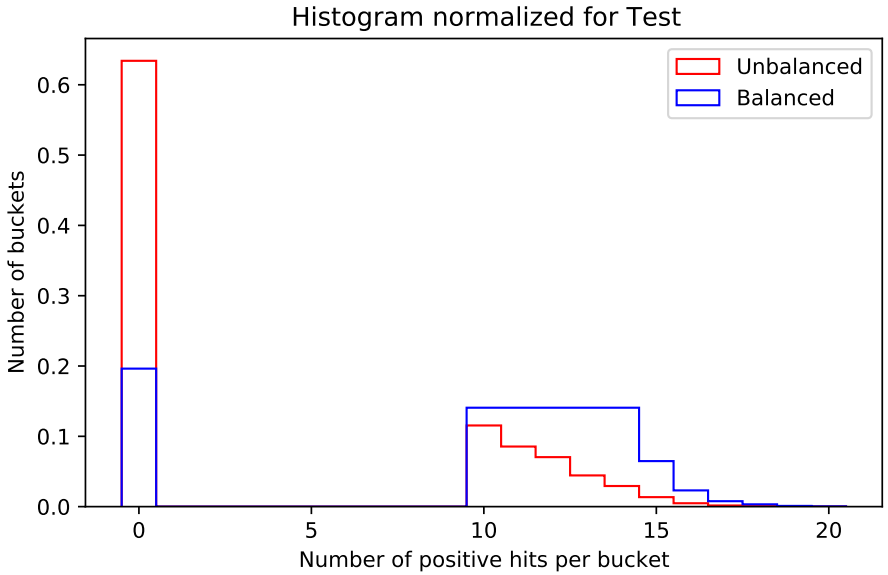
- Positive and Negative hits are now balanced.
- But we also want as flat as possible distribution to the right.
- We flatten the peak, so that bins from 10 to N have the nb of buckets as bin N.
- Then recalculate nb of buckets for bin 0, as in previous slide.
- For example, for N=13, where also N=0 gives about the same nb of buckets as N=13.

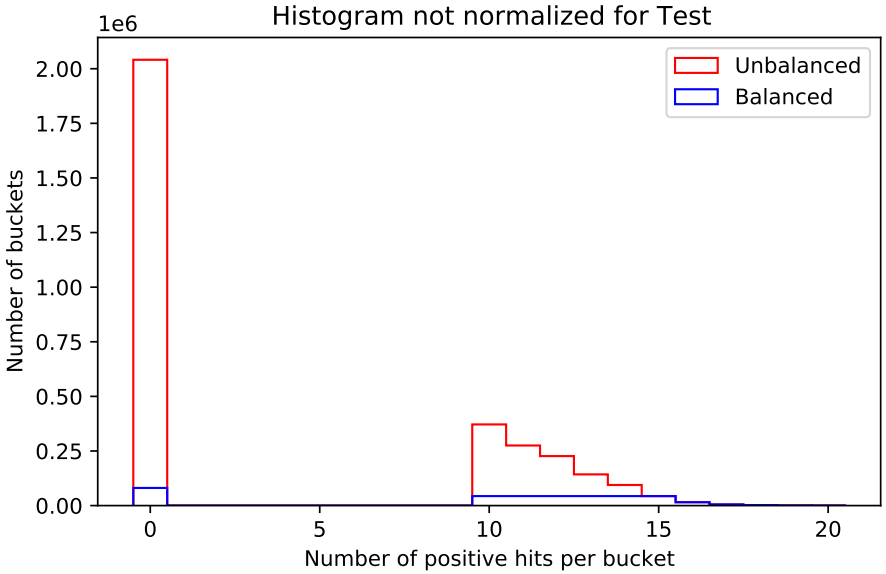


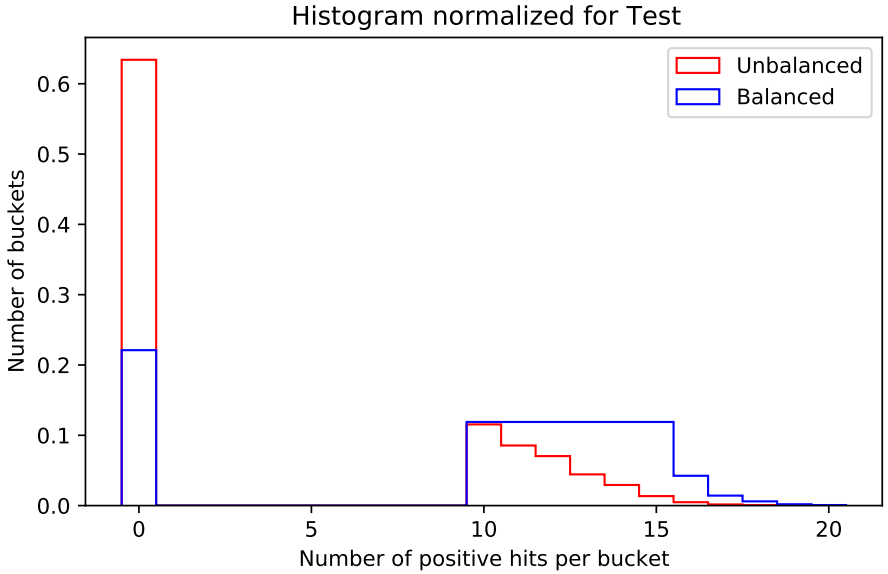


Balancing method 3/6, N=14 and N=15

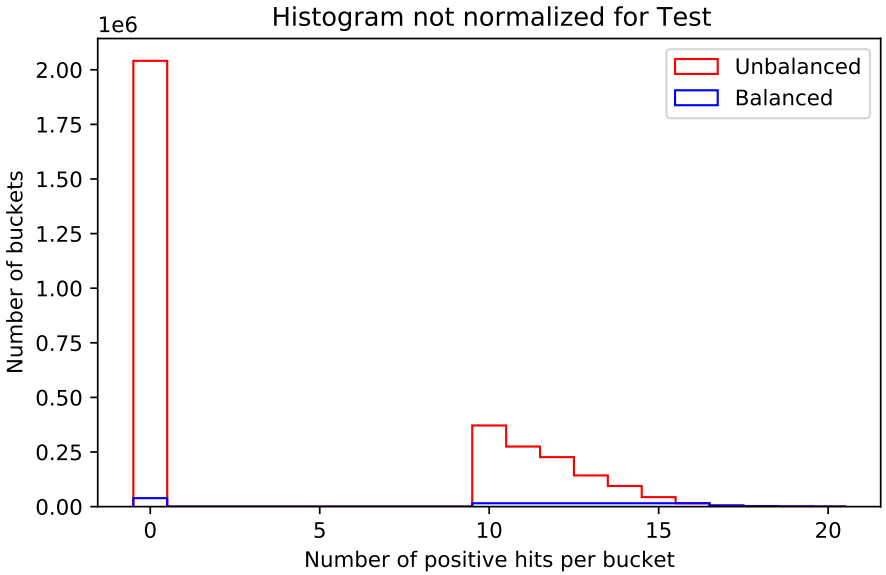


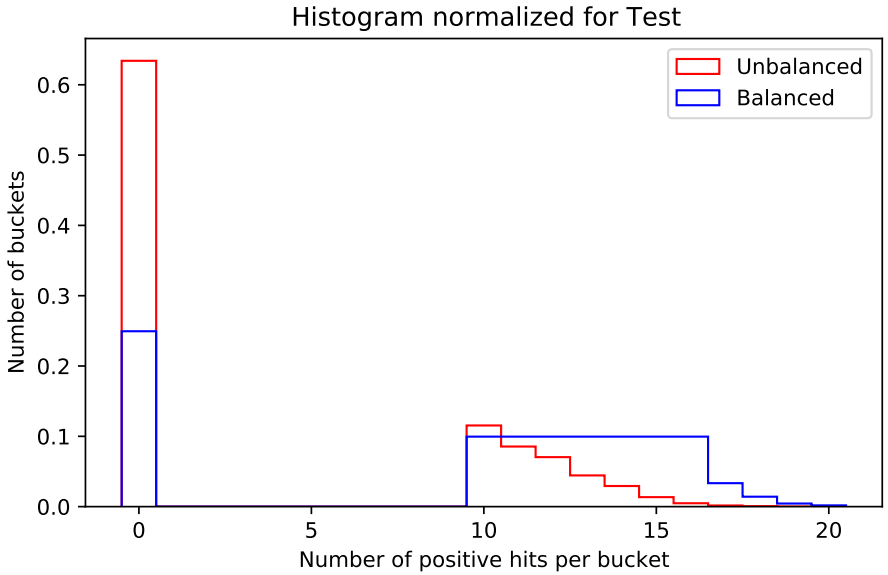


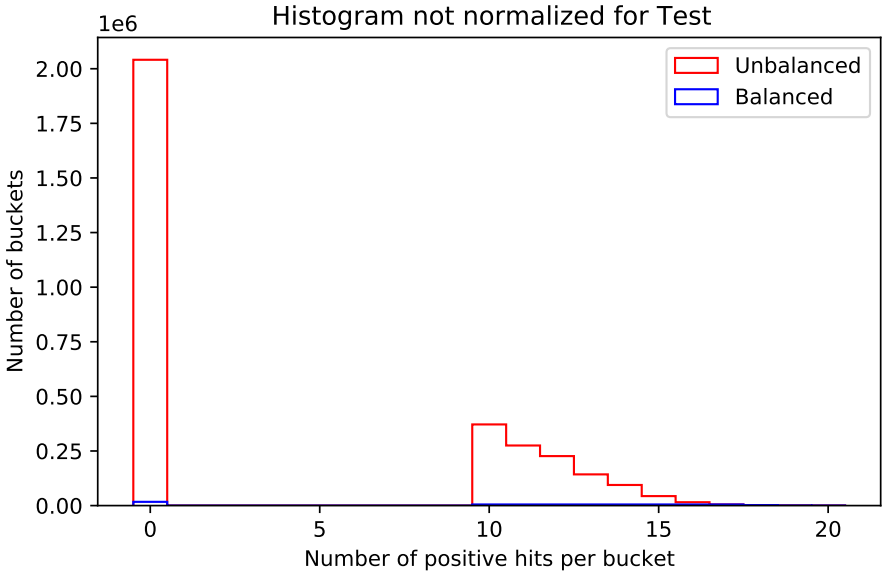


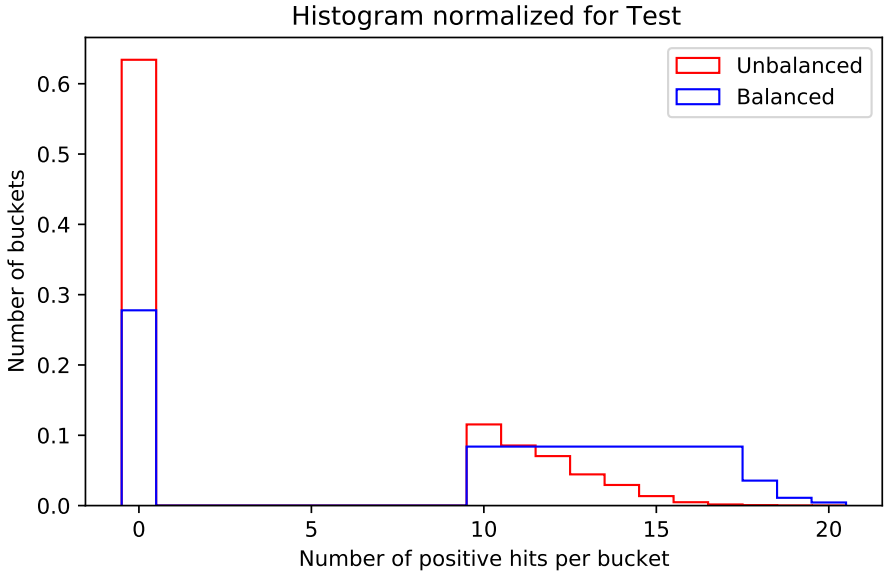


Balancing method 4/6, N=16 and N=17

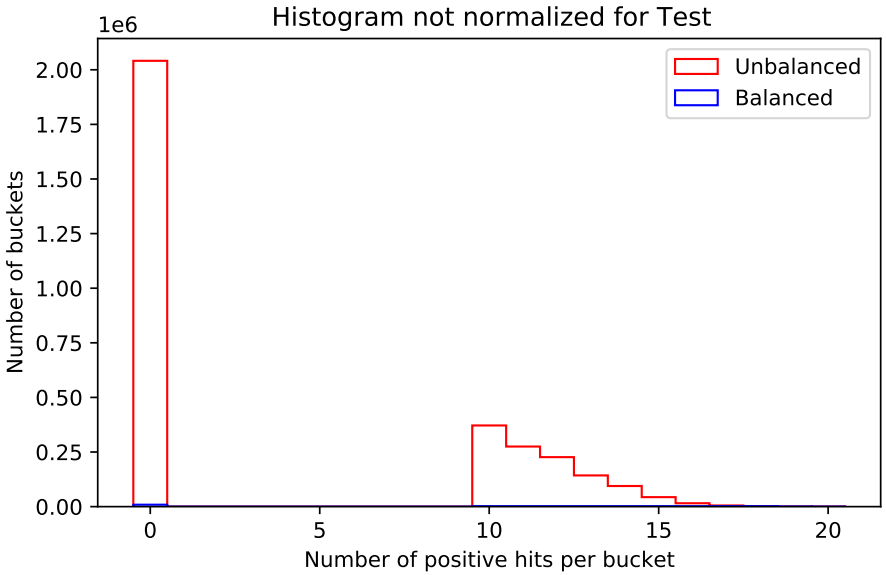


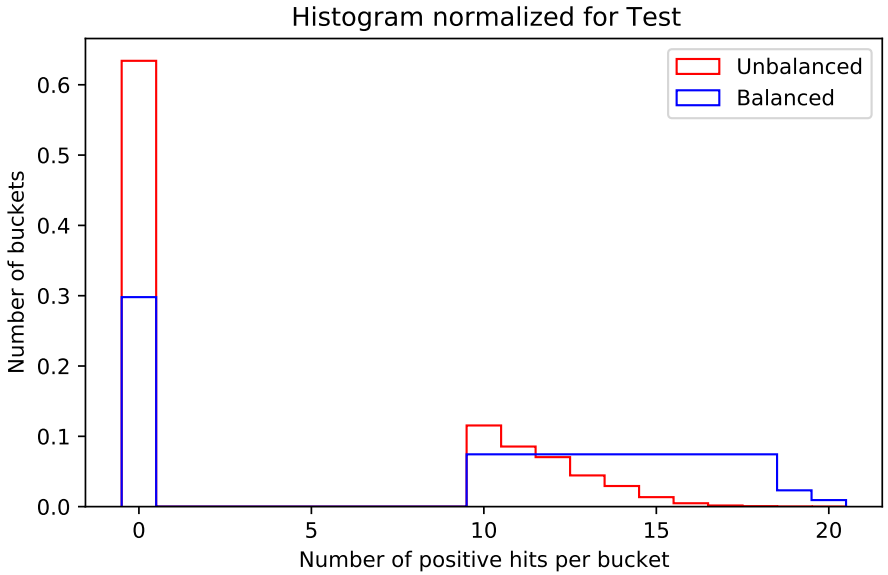


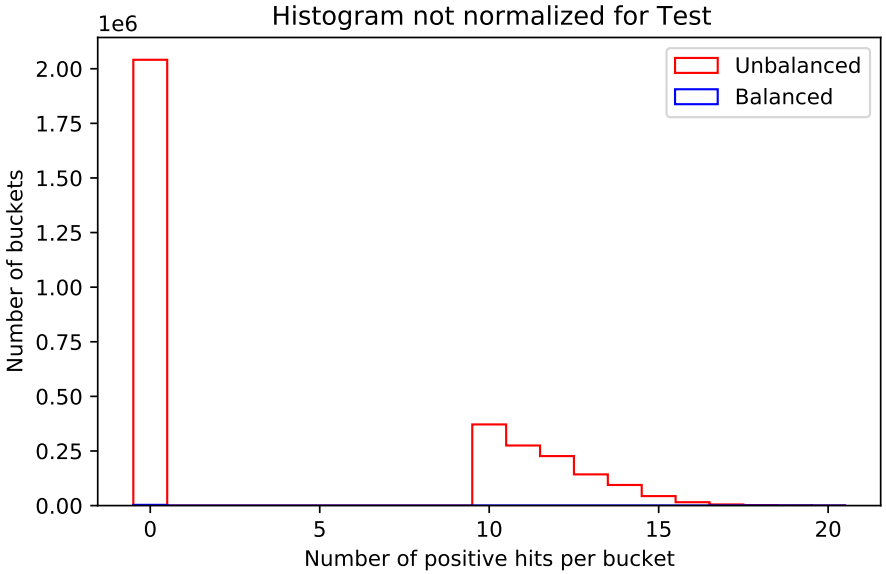


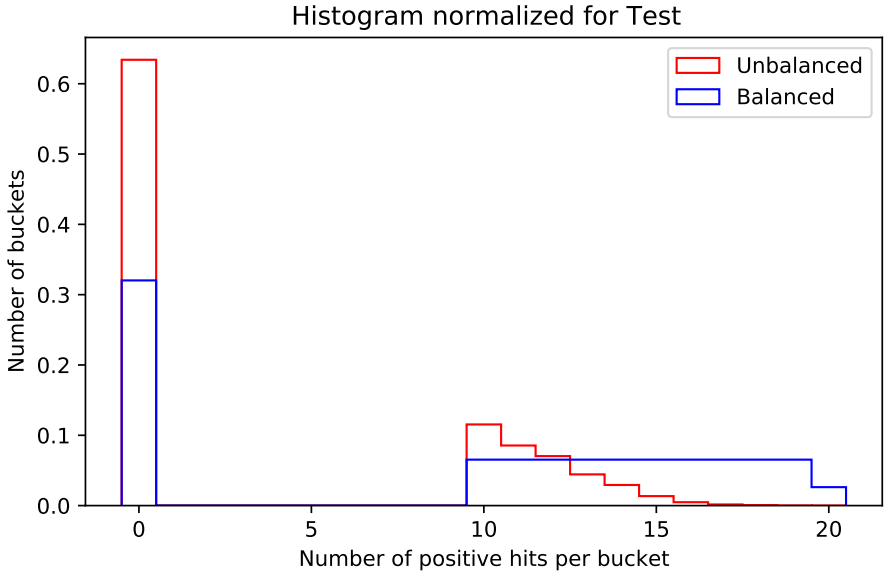


Balancing method 5/6, N=18 and N=19

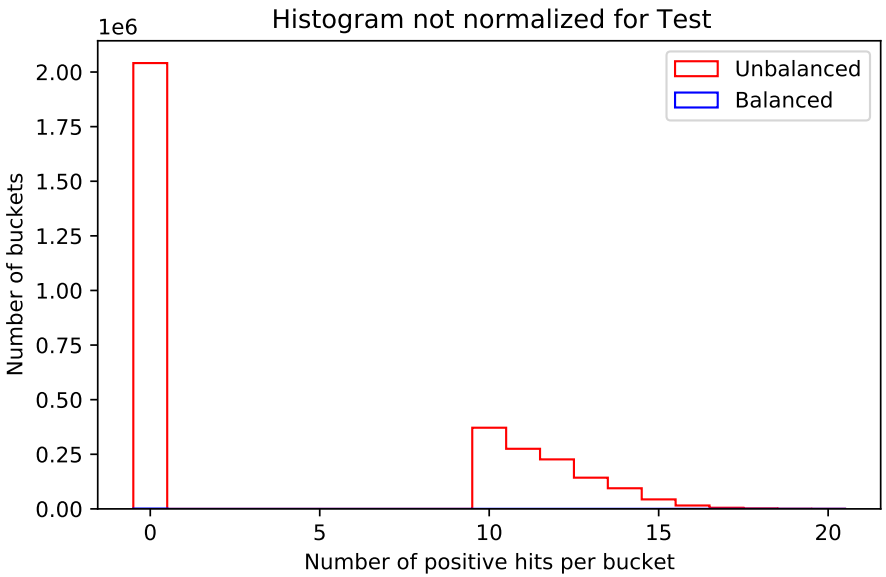


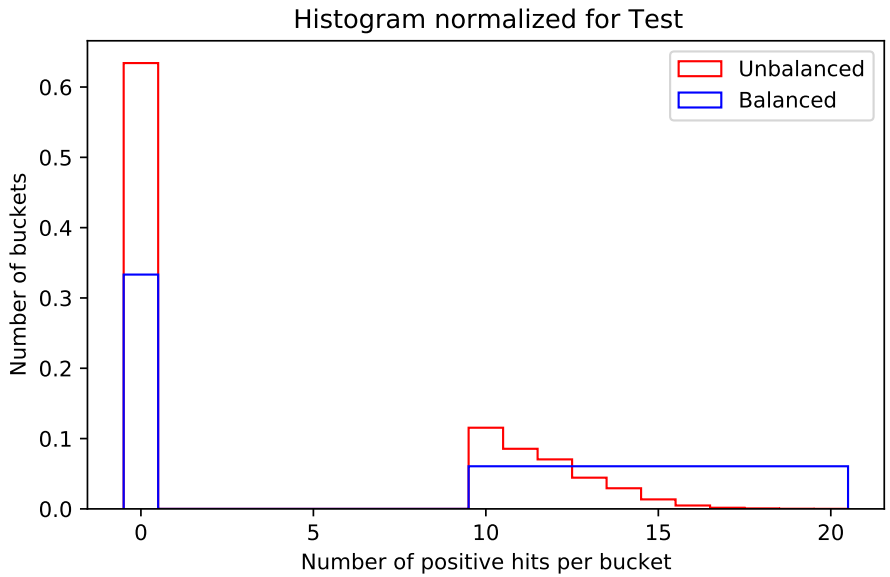






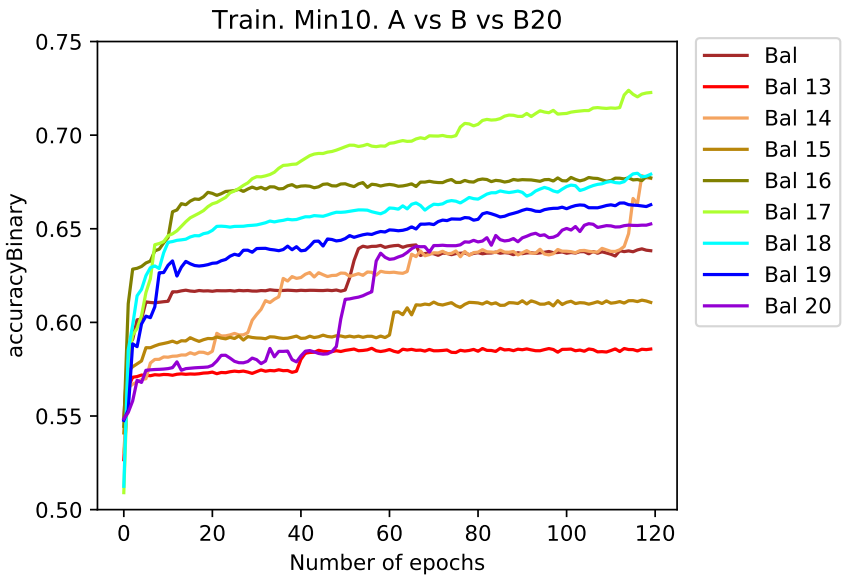
Balancing method 6/6, N=20, and Table for Train.

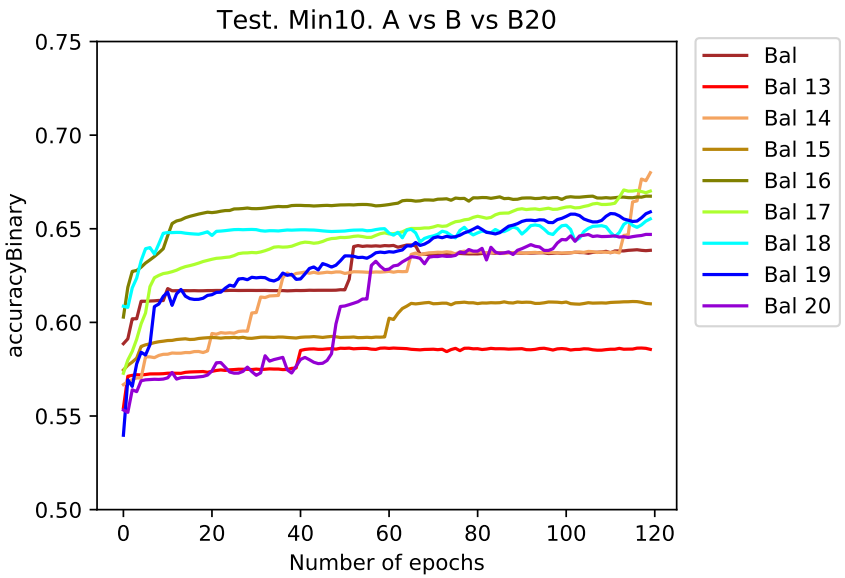


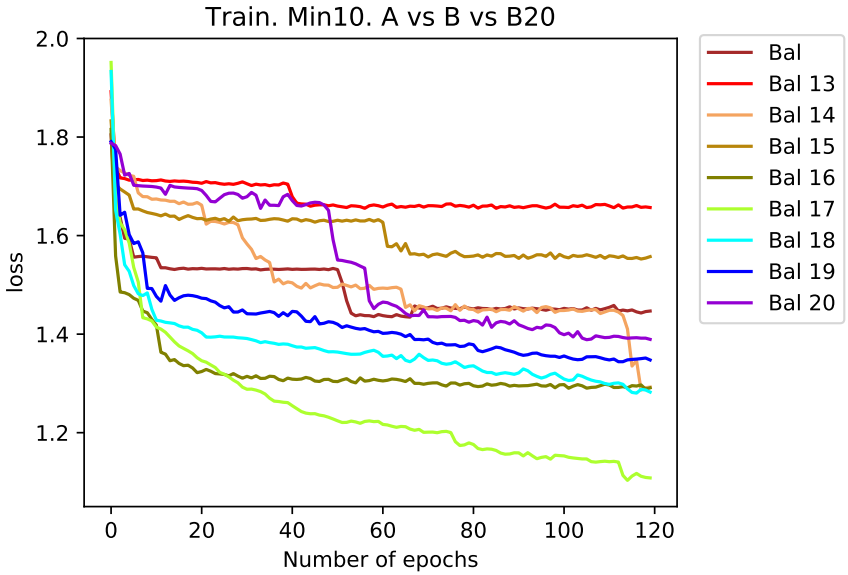


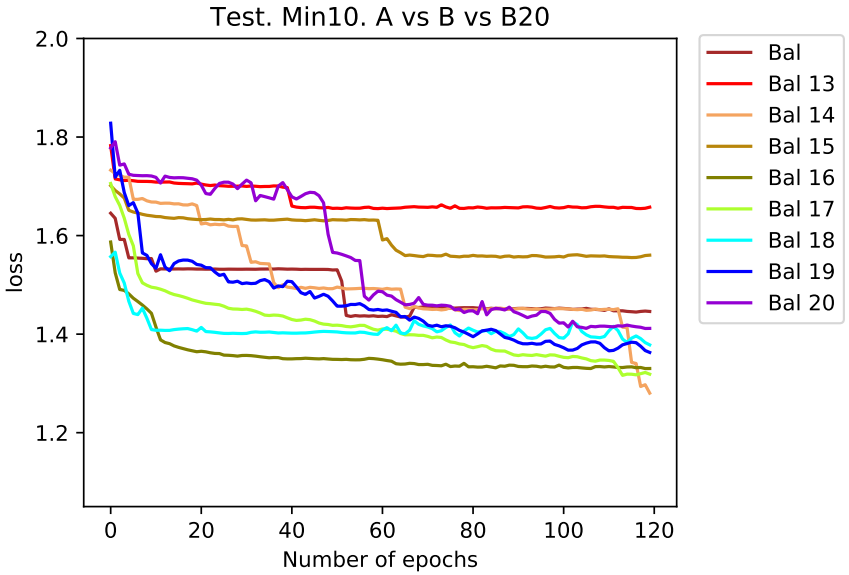
Max bin of flat peak	nbBucket Total	nbBucket bin 0	nbBucket peak	% bin 0	% peak
13	2021k	362k	324k	17.9	16.0
14	1510k	296k	213k	19.6	14.1
15	0805k	178k	096k	22.1	11.9
16	0345k	086k	034k	24.9	10.0
17	0130k	036k	011k	28.0	08.3
18	0074k	022k	006k	29.6	07.7
19	0022k	007k	001k	32.1	06.5
20	0019k	003k	001k	33.3	06.1

Accuracy and Loss from training





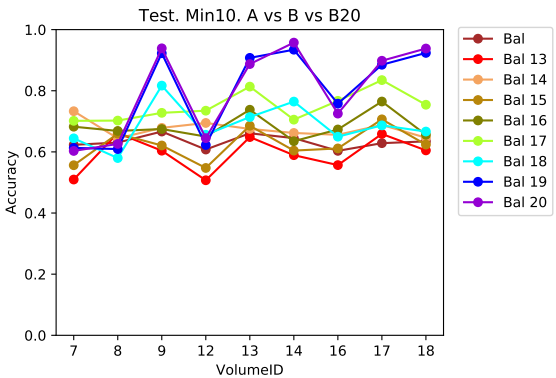


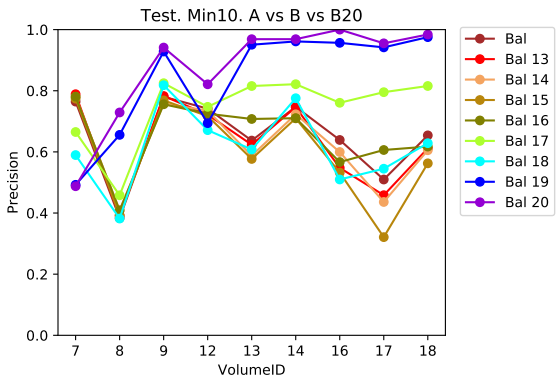


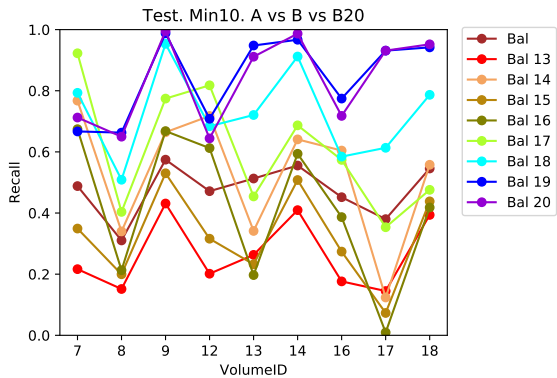
Metrics for each VolumeID overlay balancing methods.

- If unbalanced, as seen before, precision and recall were zero, as it learned to predict all buckets with negative hits.
- Now in all methods of balancing, precision and recall are not zero.
- Best performance seems when the peak is most flat, so for 19 and 20.
- 19 and 20 have also very few buckets remaining, so fast to process.

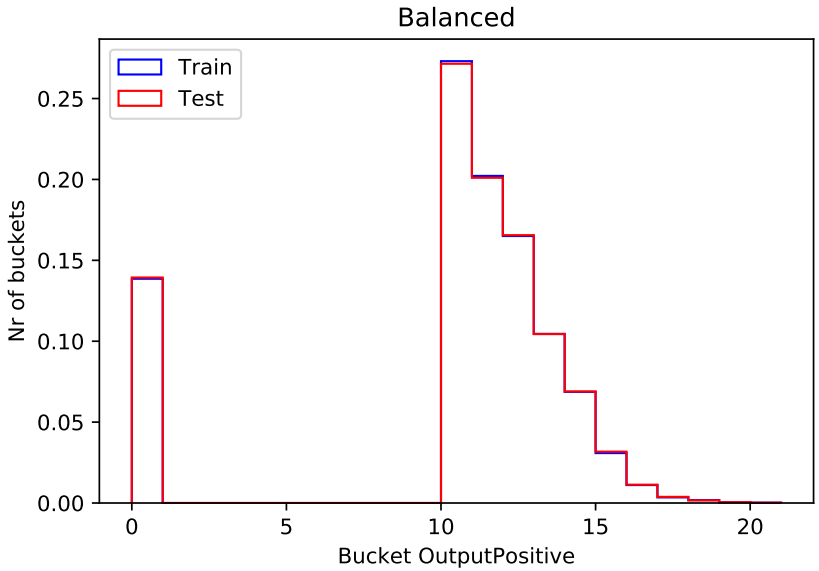
Accuracy	Precision	Recall
$\frac{TP+TN}{TP+FP+FN+TN}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$

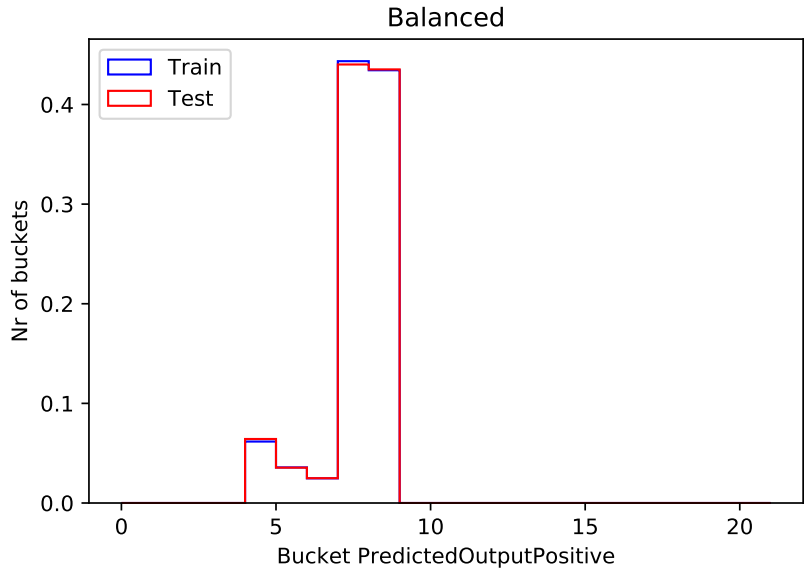


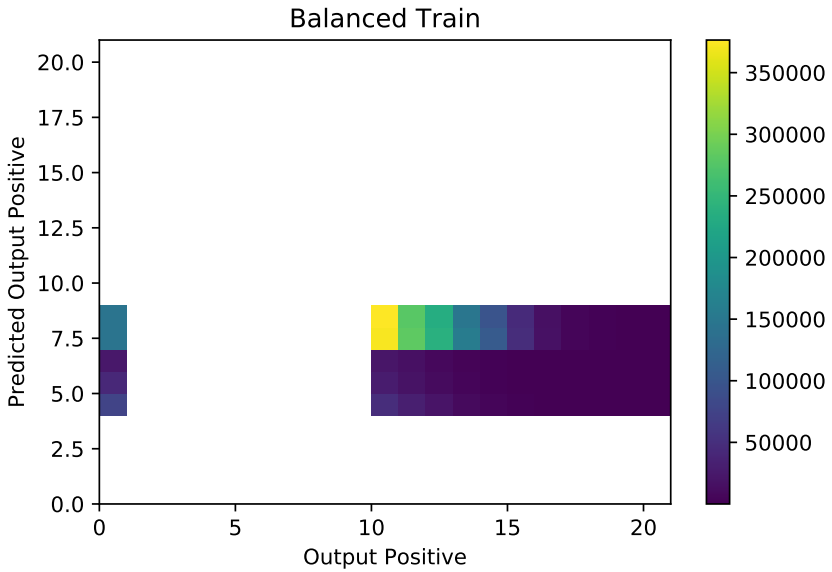


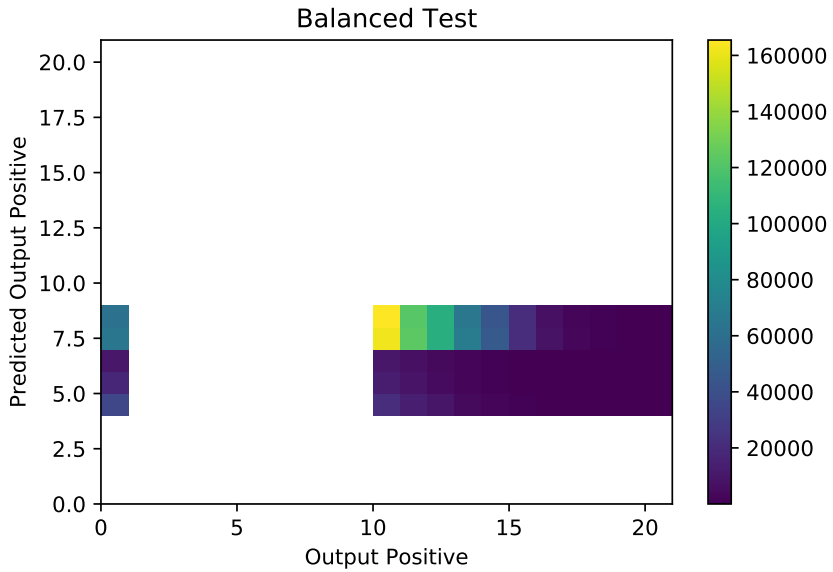


2D plots Balanced no flat peak

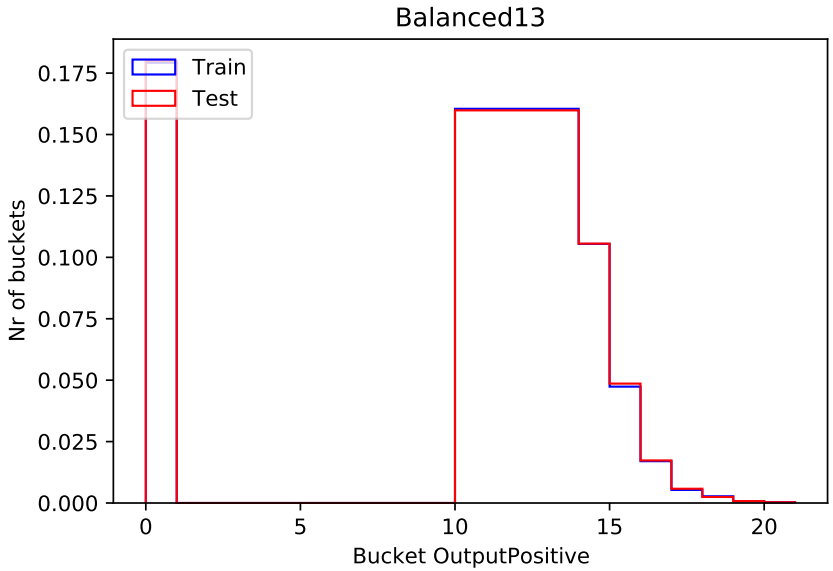


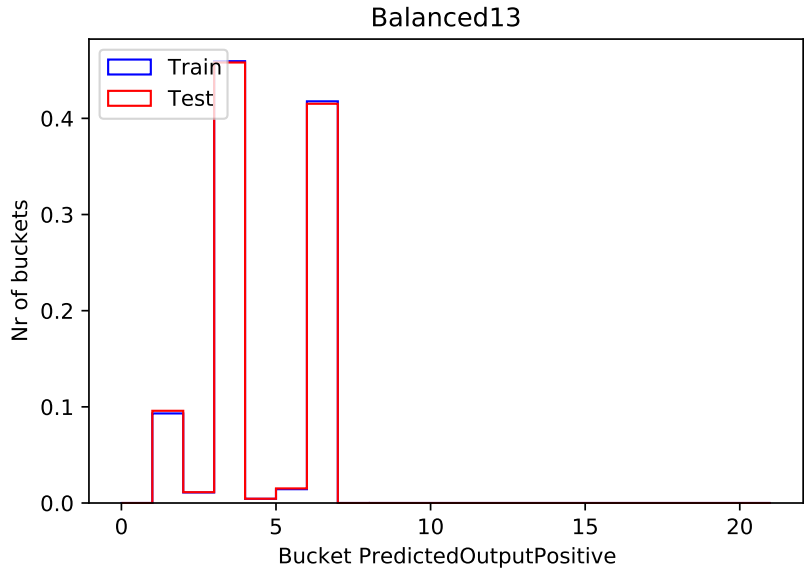


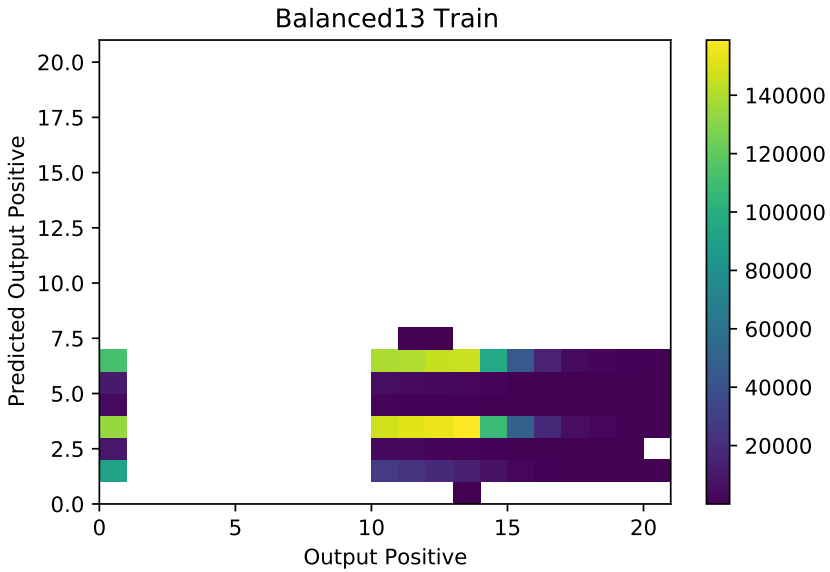


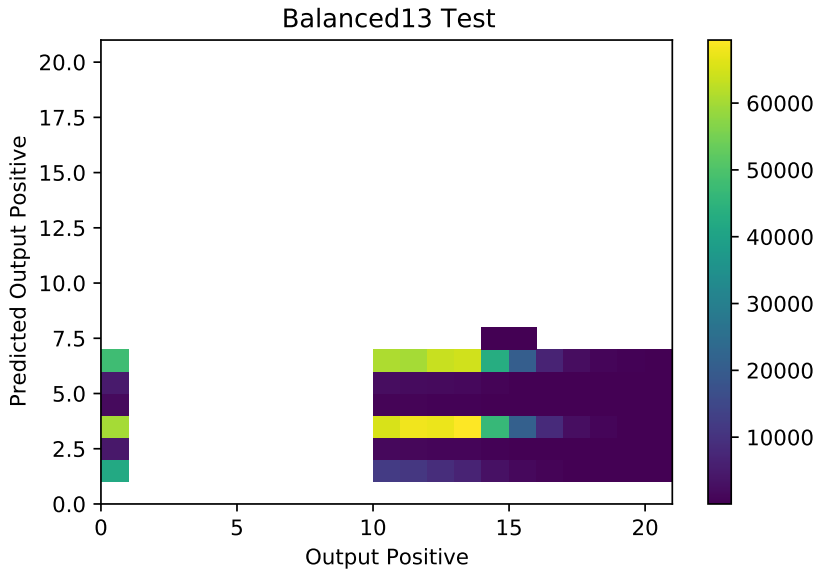


2D plots Balanced flat peak 10-13

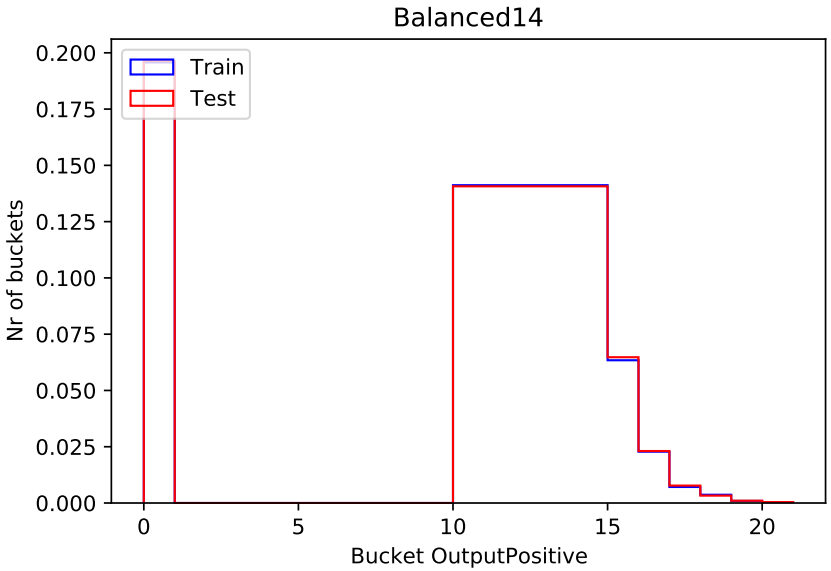


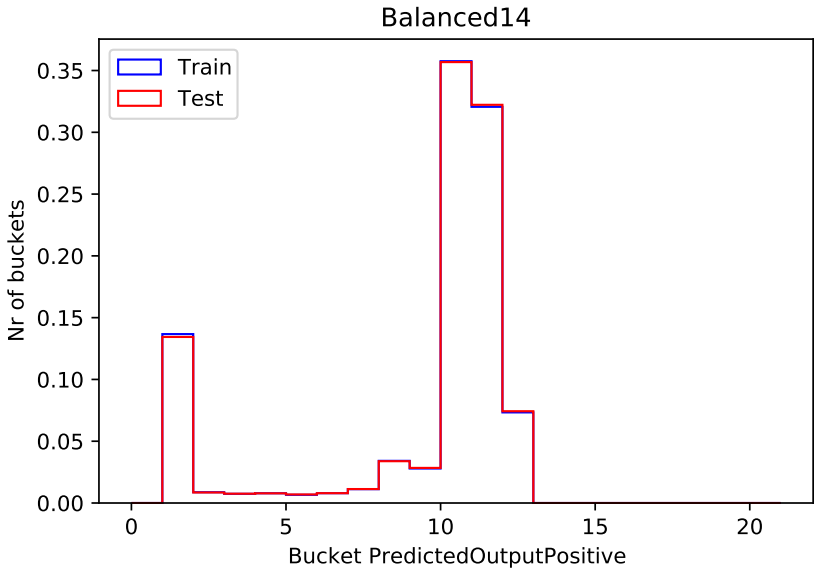


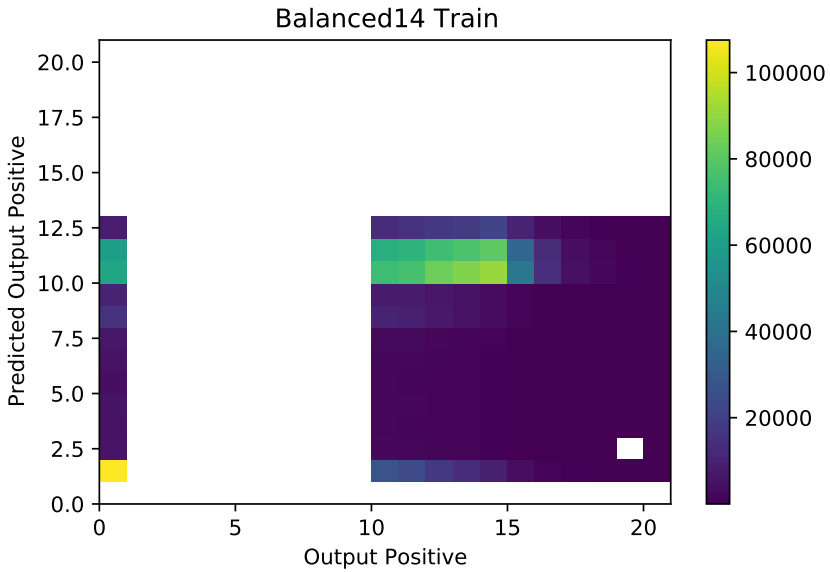


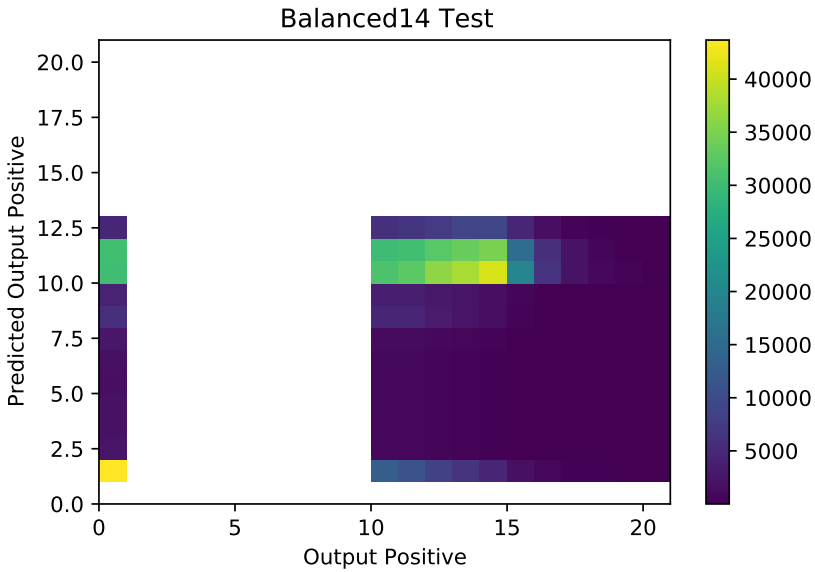


2D plots Balanced flat peak 10-14

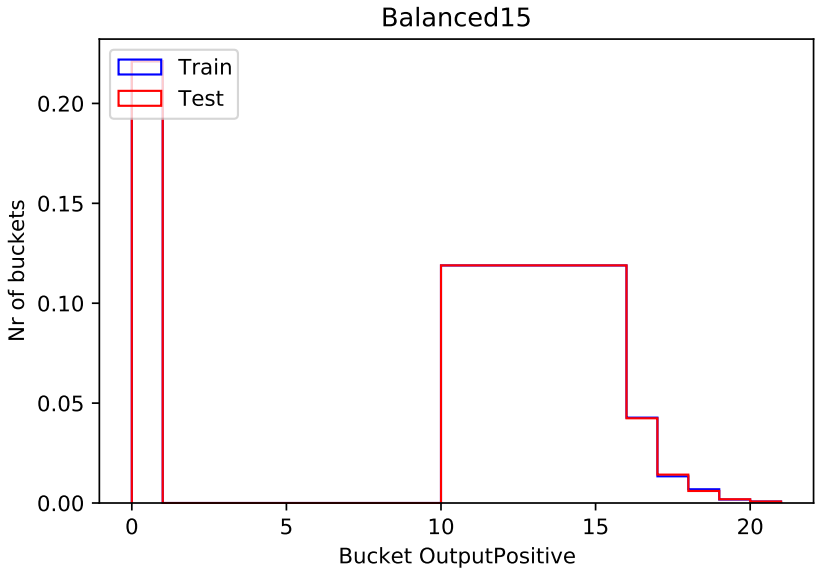


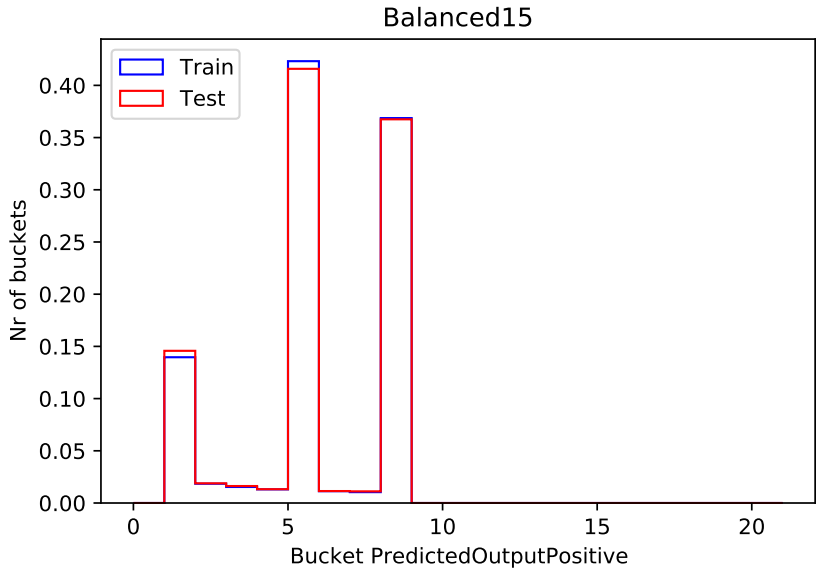


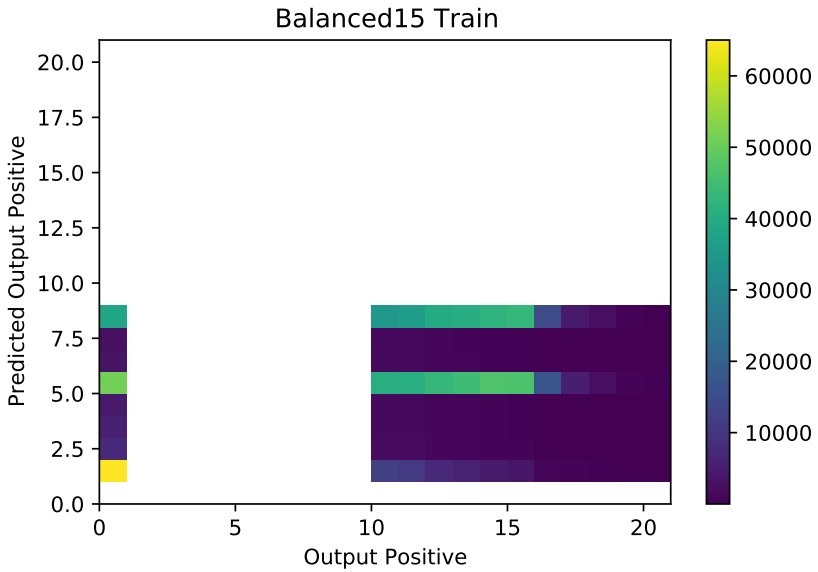


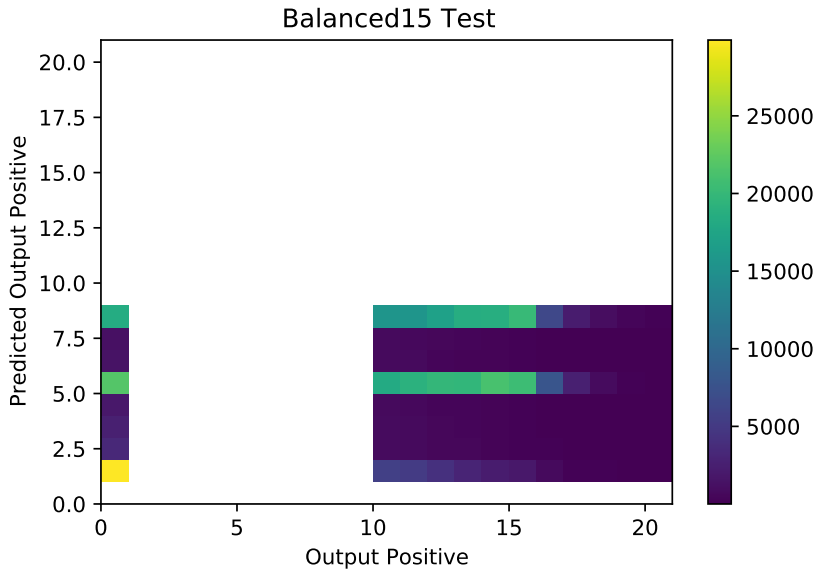


2D plots Balanced flat peak 10-15

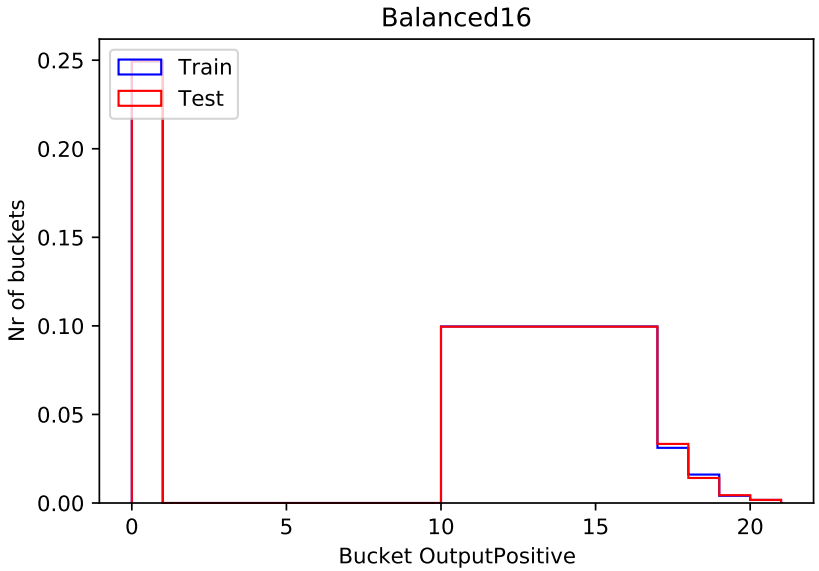


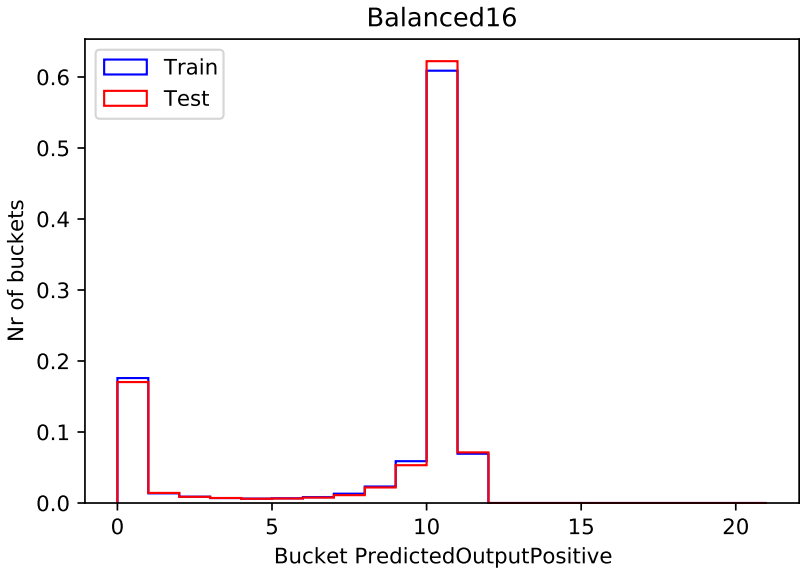


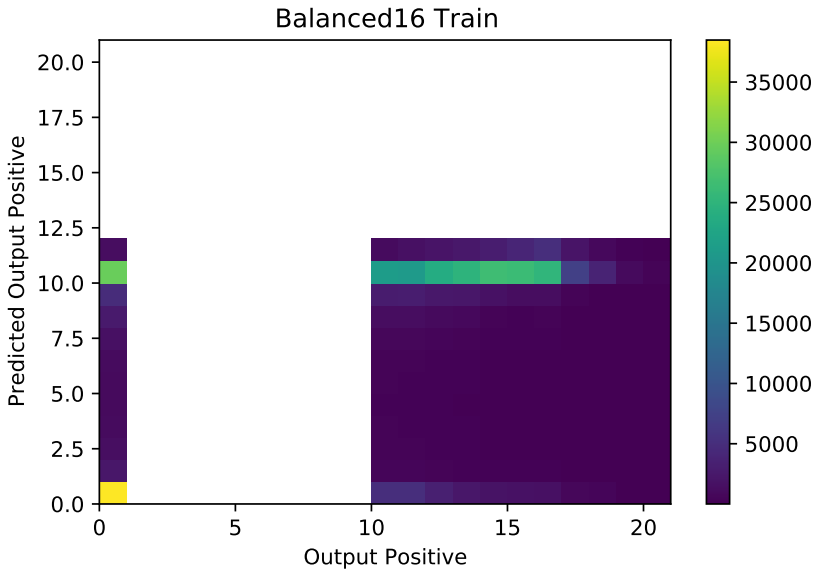


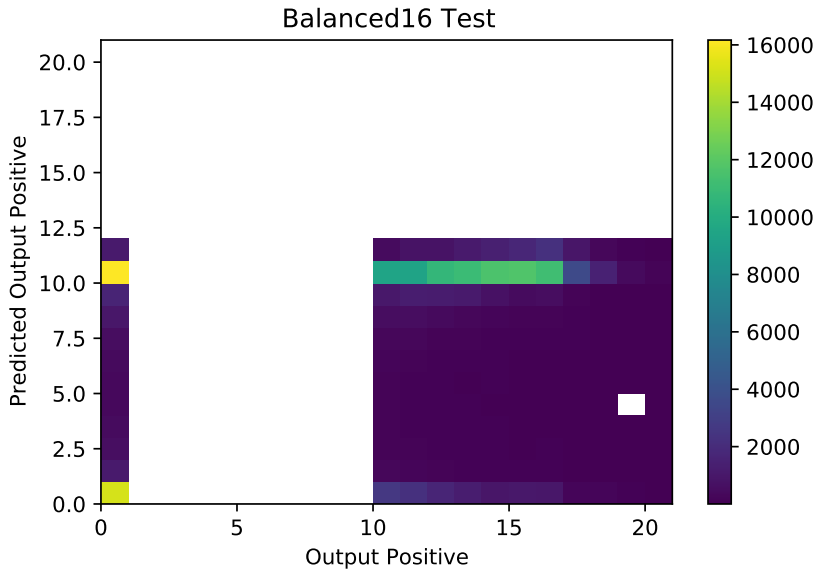


2D plots Balanced flat peak 10-16

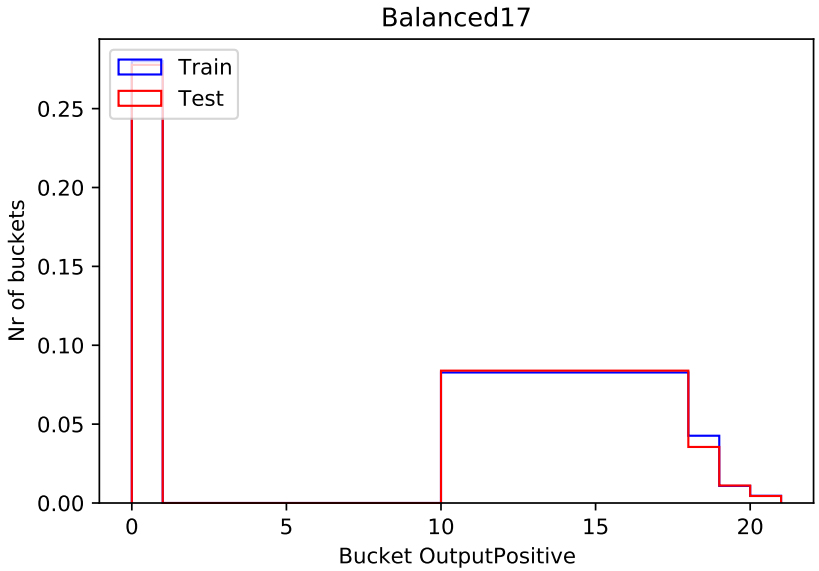


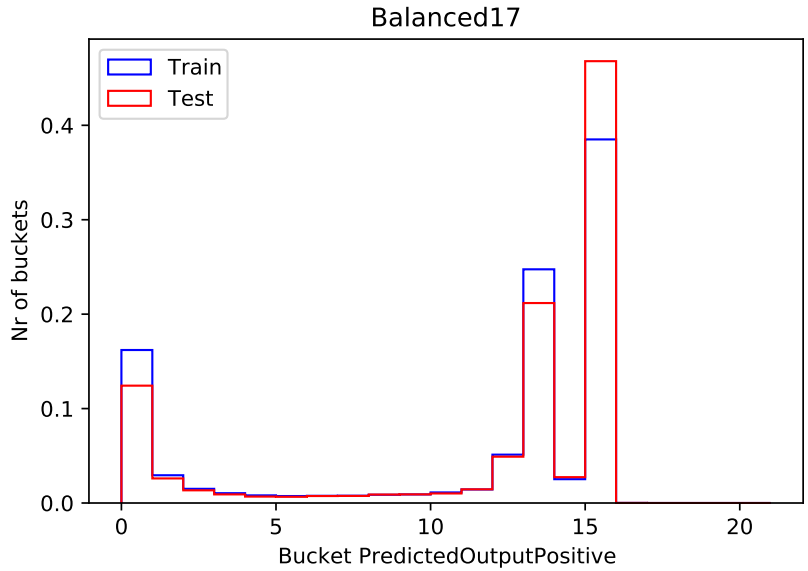


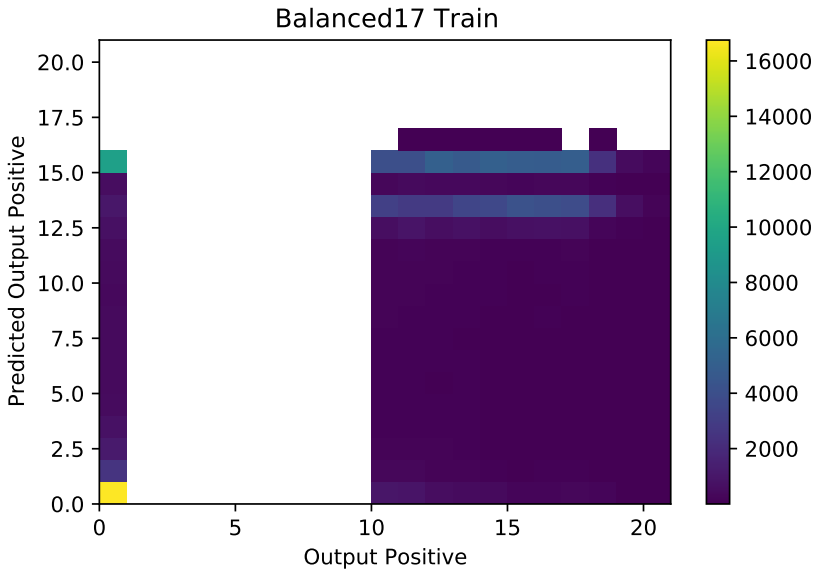


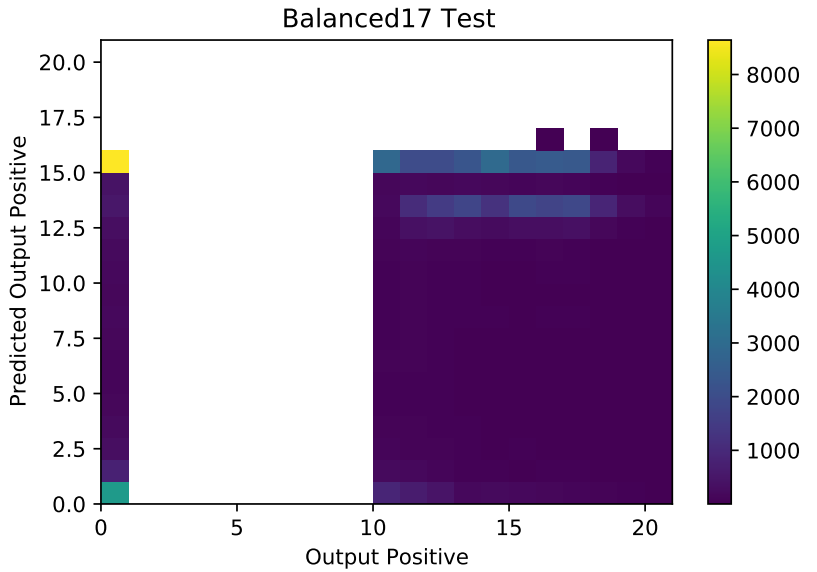


2D plots Balanced flat peak 10-17

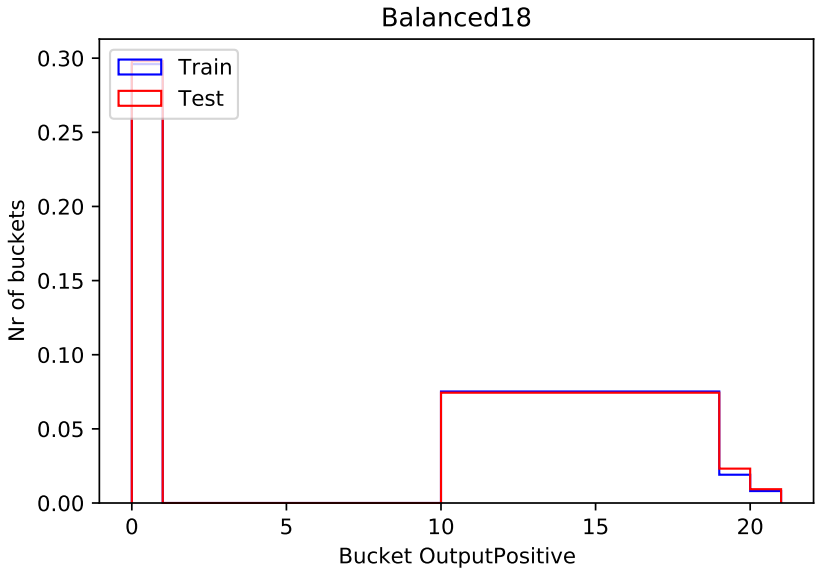


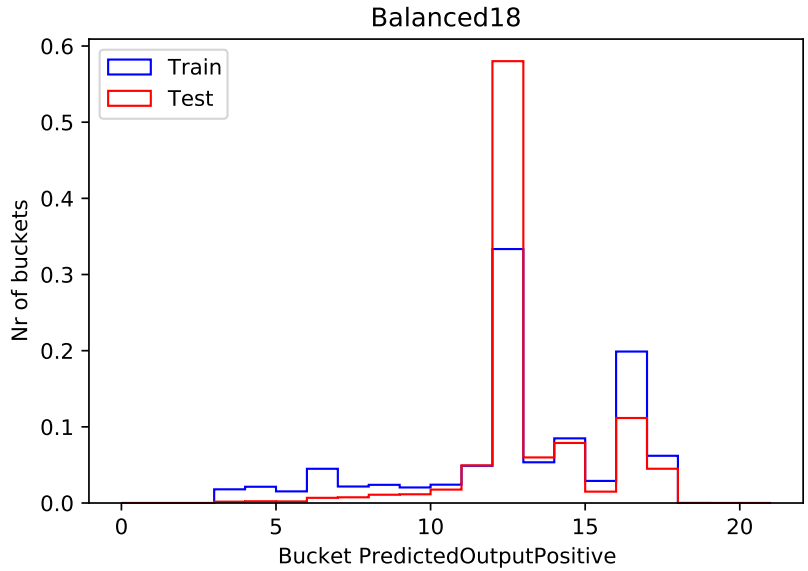


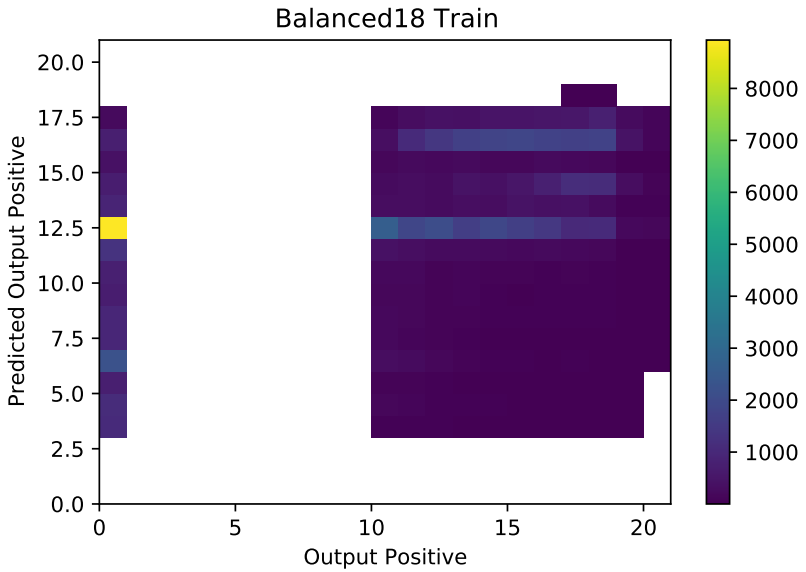


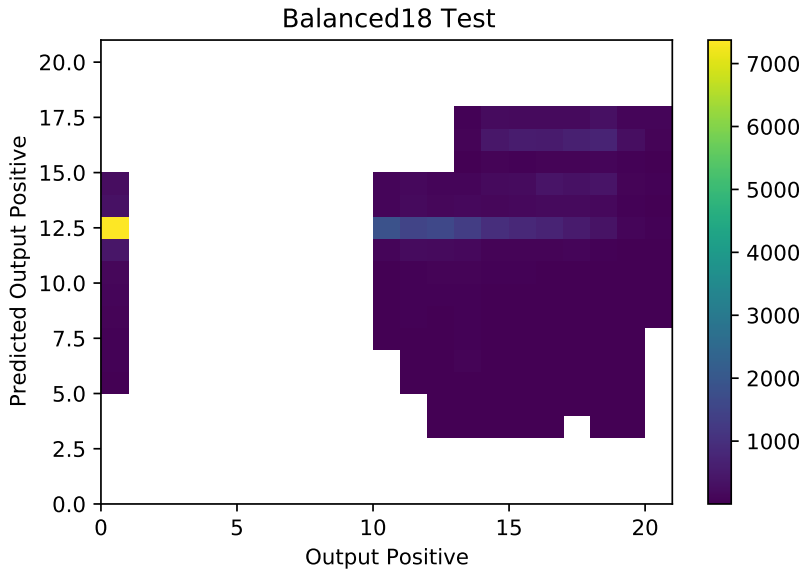


2D plots Balanced flat peak 10-18

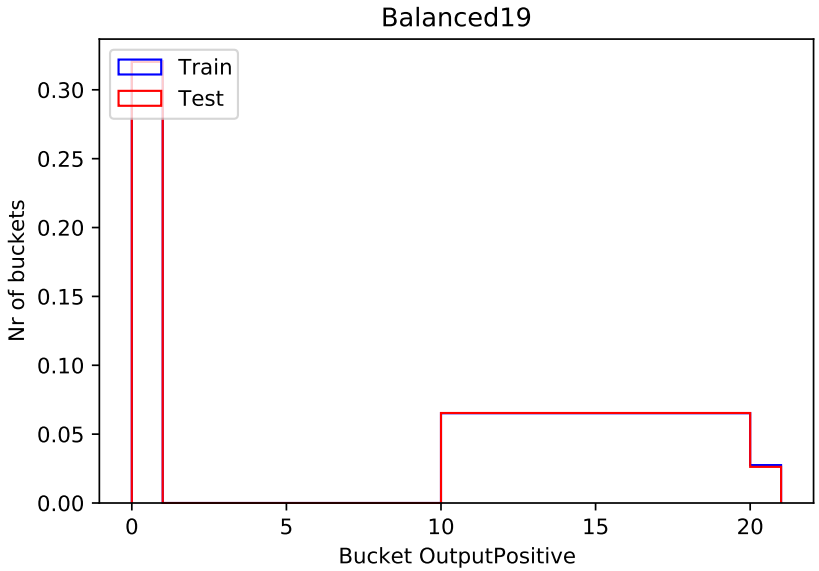


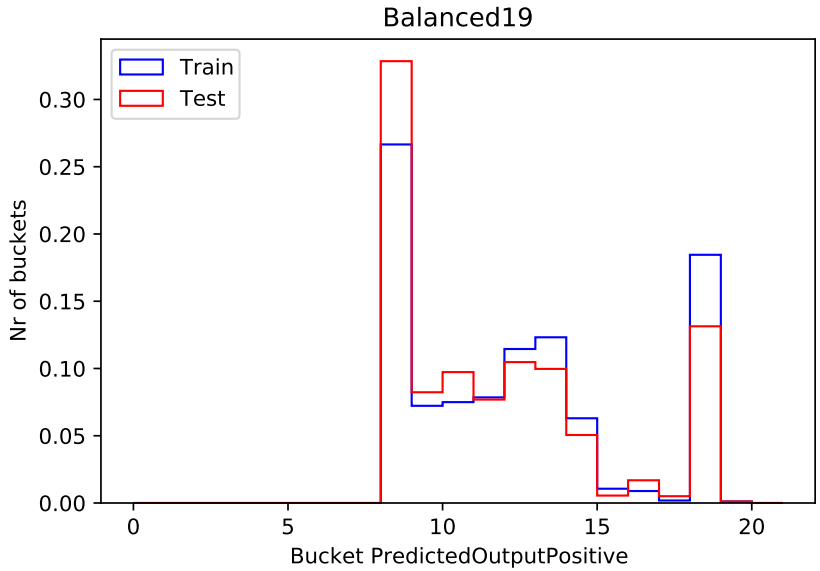


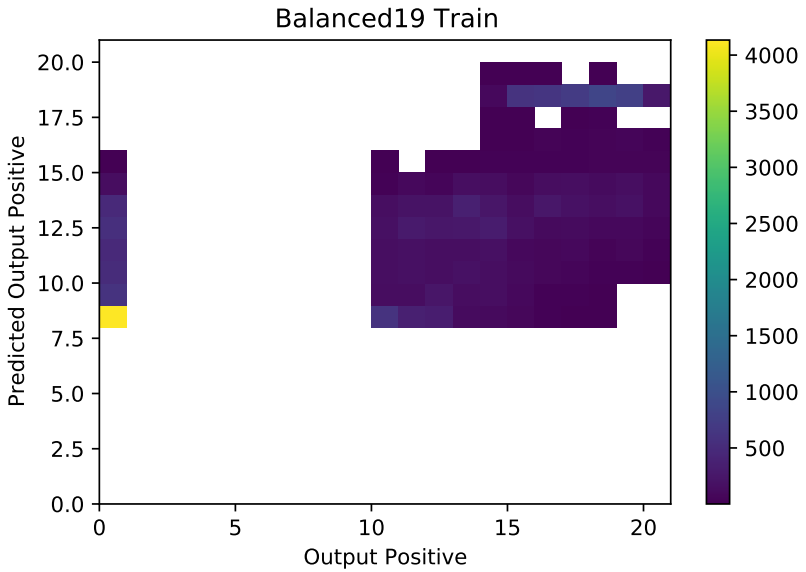


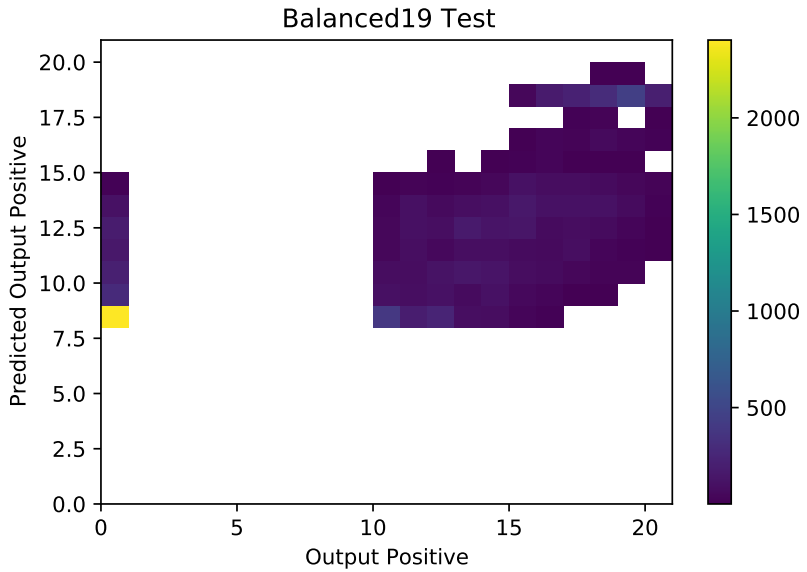


2D plots Balanced flat peak 10-19

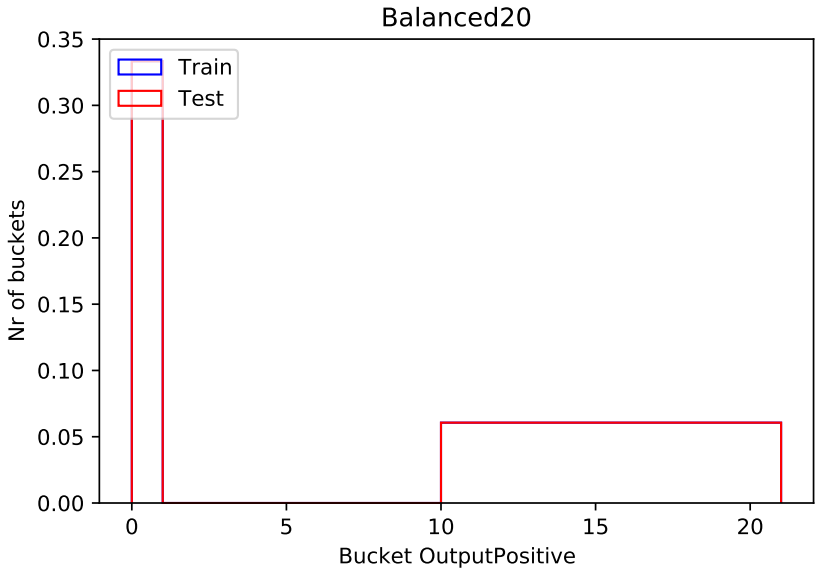


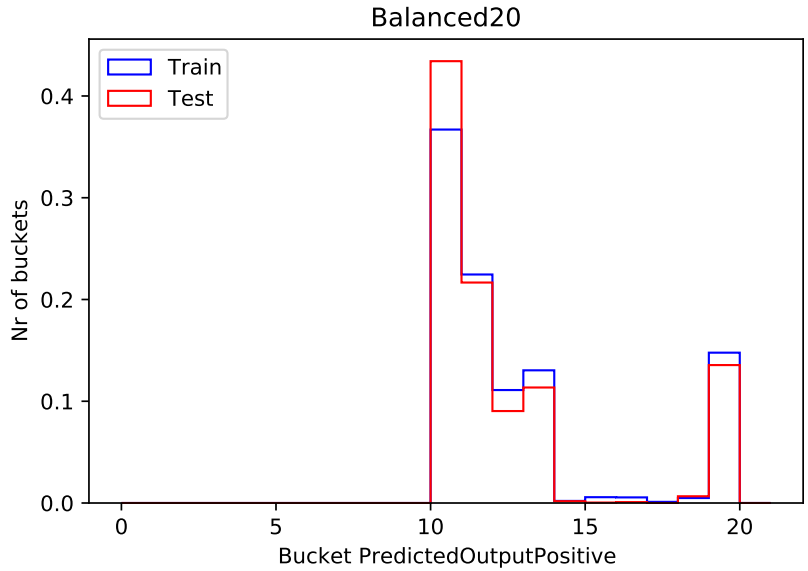


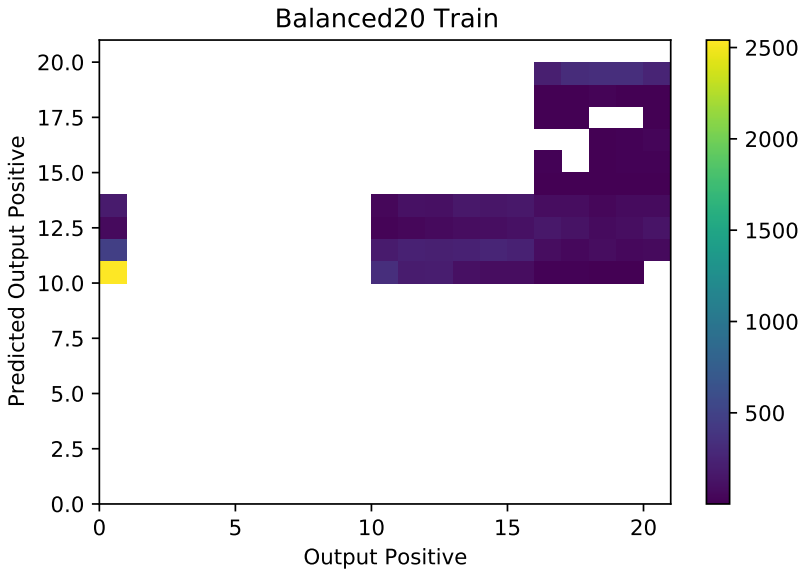


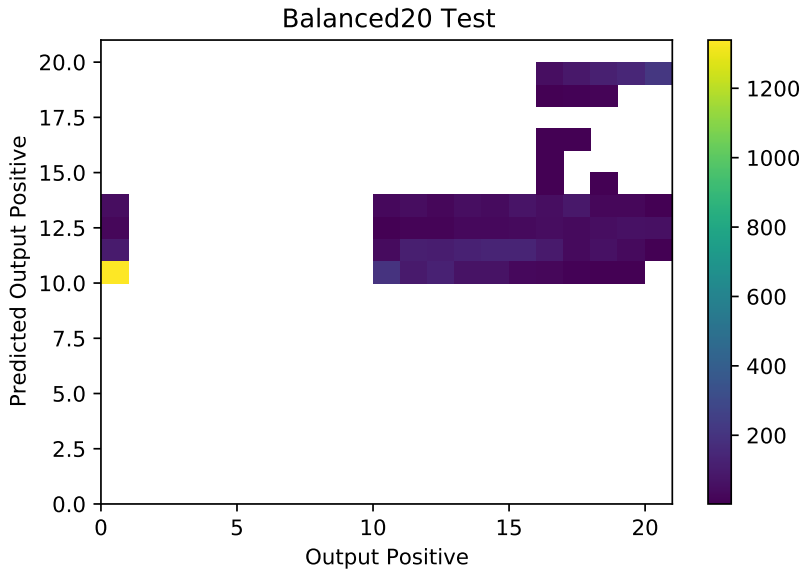


2D plots Balanced flat peak 10-20









Conclusion

- 100 events, 70 in train, 30 in test.
- If nbPositiveHit$<$10, set nbPositiveHit$=$0 and output made only of -1.
- Balanced nb positive and negative hits by reducing the number of buckets at bin 0, with no flat peak or with various lengths of flat peak.
- Best accuracy, precision and recall for 19-20.
- But closest PredictedOutputPositive to OutputPositive for 16-17.
- Overall best choice seems with flat peak between 10 and 17, then rebalance at bin 0.
- Next steps: write thesis.