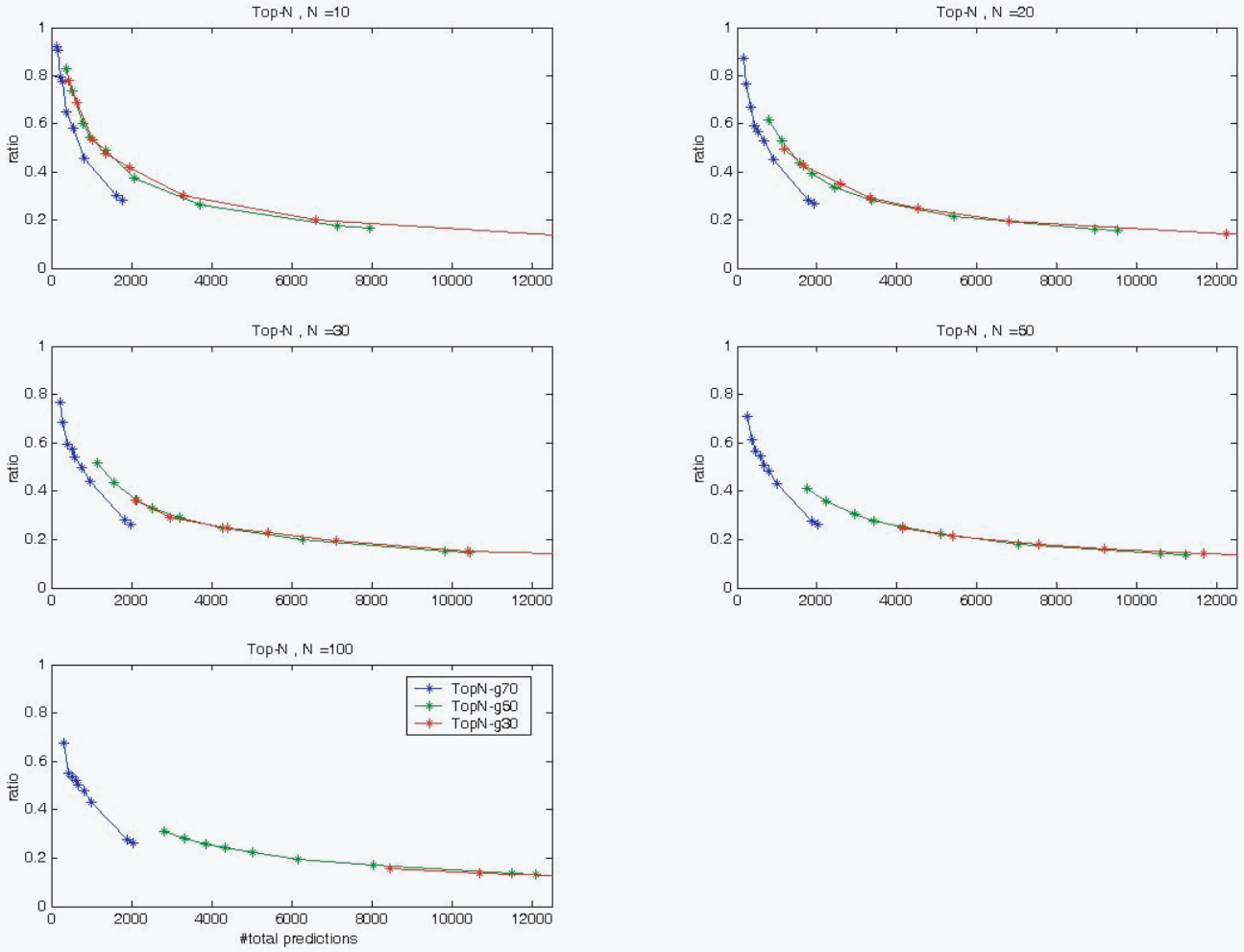
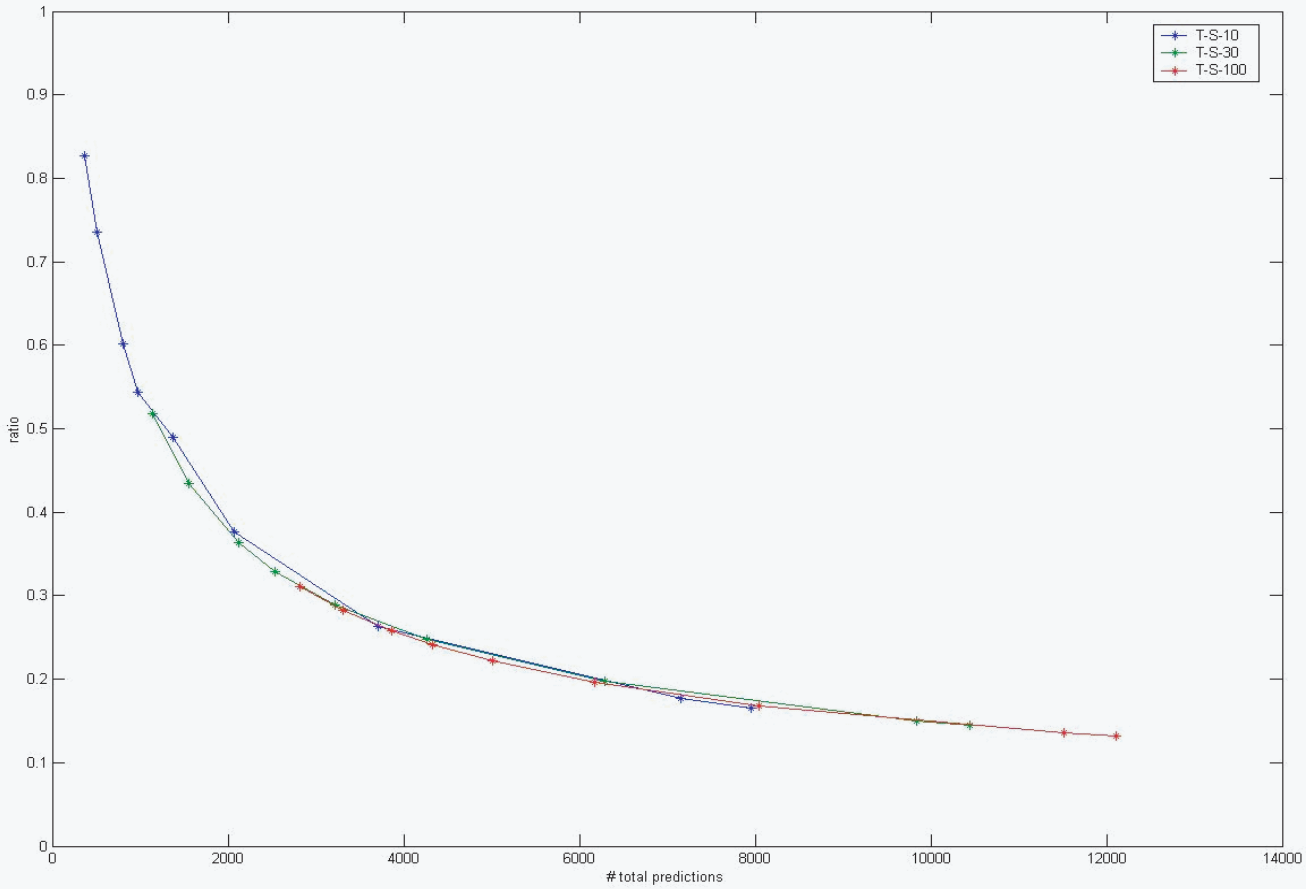
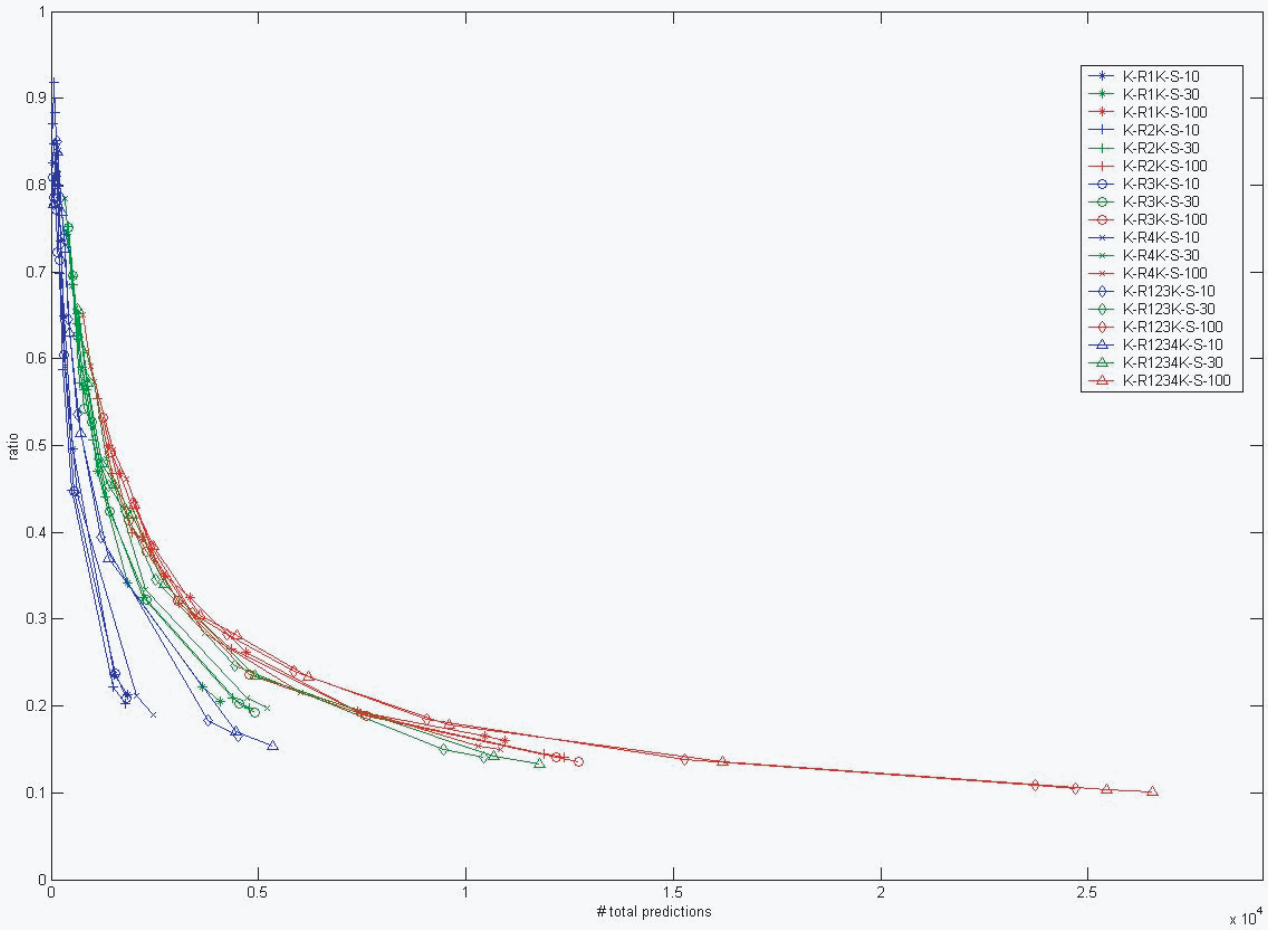


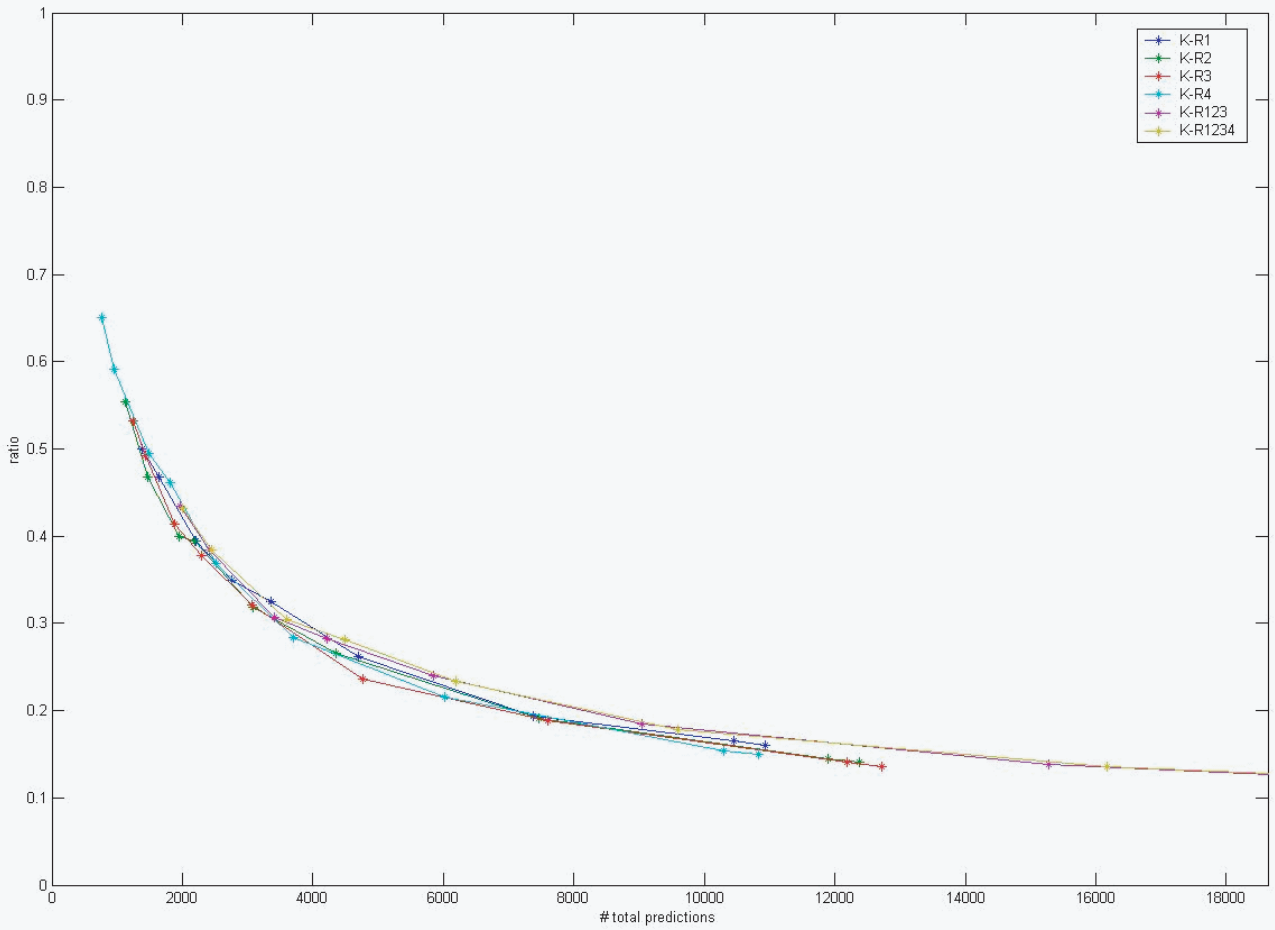

**Web Fig. 1**
Selecting correlation cutoff for top-N clusters for the augmented data set, using validation curves as described in the text and methods. Cluster size cutoffs N were as indicated over each graph. Correlation cutoffs are represented by color (correlation cutoff > 0.3 (red line), 0.5 (green line), 0.7 (blue line)). P values were modulated between 0 and 1 to form curves. A correlation cutoff of 0.5 (green line) was used for subsequent analysis.



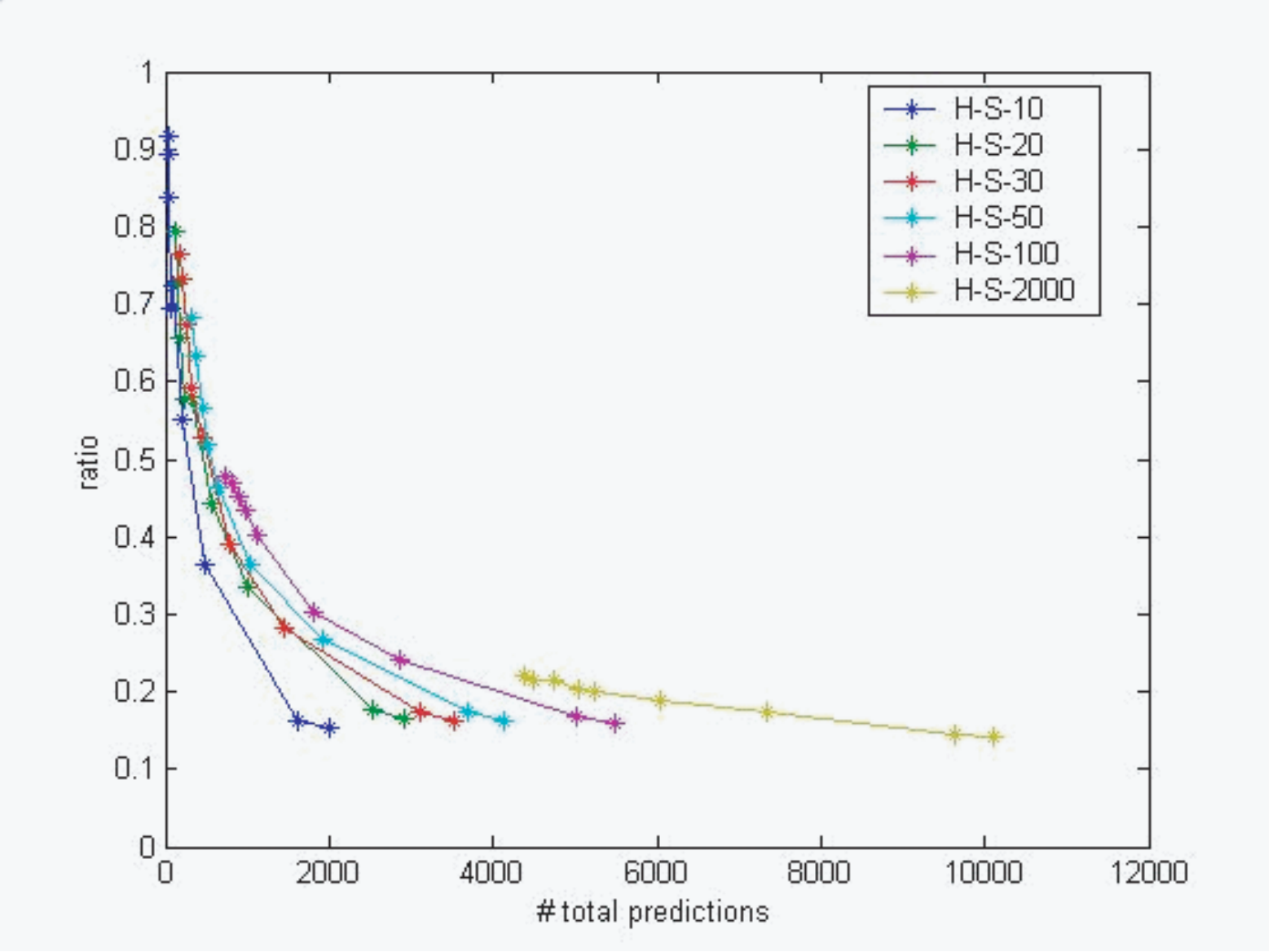

**Web Fig. 2**
Selecting cluster size cutoffs for top-N clusters for the augmented data set. Size cutoffs are represented by color (N less than or equal to 10 (blue line), 30 (green line), 100 (red line)). Correlation cutoff was held at 0.5 and P values were modulated between 0 and 1 to form curves. Top-10 was used for subsequent analysis.




**Web Fig. 3**
Selecting cluster size cutoff for K-means clusters for the augmented data set. Performance of size cutoffs was compared over several runs of K-means (indicated as R1, R2, R3, R4) and combinations of runs (indicated as R123 and R1234). Size cutoffs are represented by color (size cutoffs 10 (blue line), 30 (green line), 100 (red line)). P values were modulated between 0 and 1 to form curves. Size cutoff of 100 (red line) offered the highest coverage at a fixed percent accuracy and was used for subsequent analysis.




**Web Fig. 4**
Selecting runs for K-means clusters for the augmented data set. Runs of K-means (indicated as R1, R2, R3, R4) and combinations of runs (indicated as R123 and R1234) were compared using a cluster size cut of 100. P values were modulated between 0 and 1 to form curves. Run 4 (R4 (light blue line) was selected for its high accuracy and was used for subsequent analysis.




**Web Fig. 5**
Selecting cluster size cutoff for hierarchical clusters for the augmented data set. Runs of hierarchical clusters are compared by size cut (size cutoffs 10 (blue line), 20 (green line), 30 (red line), 50 (light blue line), 100 (purple line), 2000 (yellow line)). P values were modulated between 0 and 1 to form curves. Size cutoff of 100 (purple line) offered the highest coverage at a fixed percent accuracy and was used in subsequent analysis.