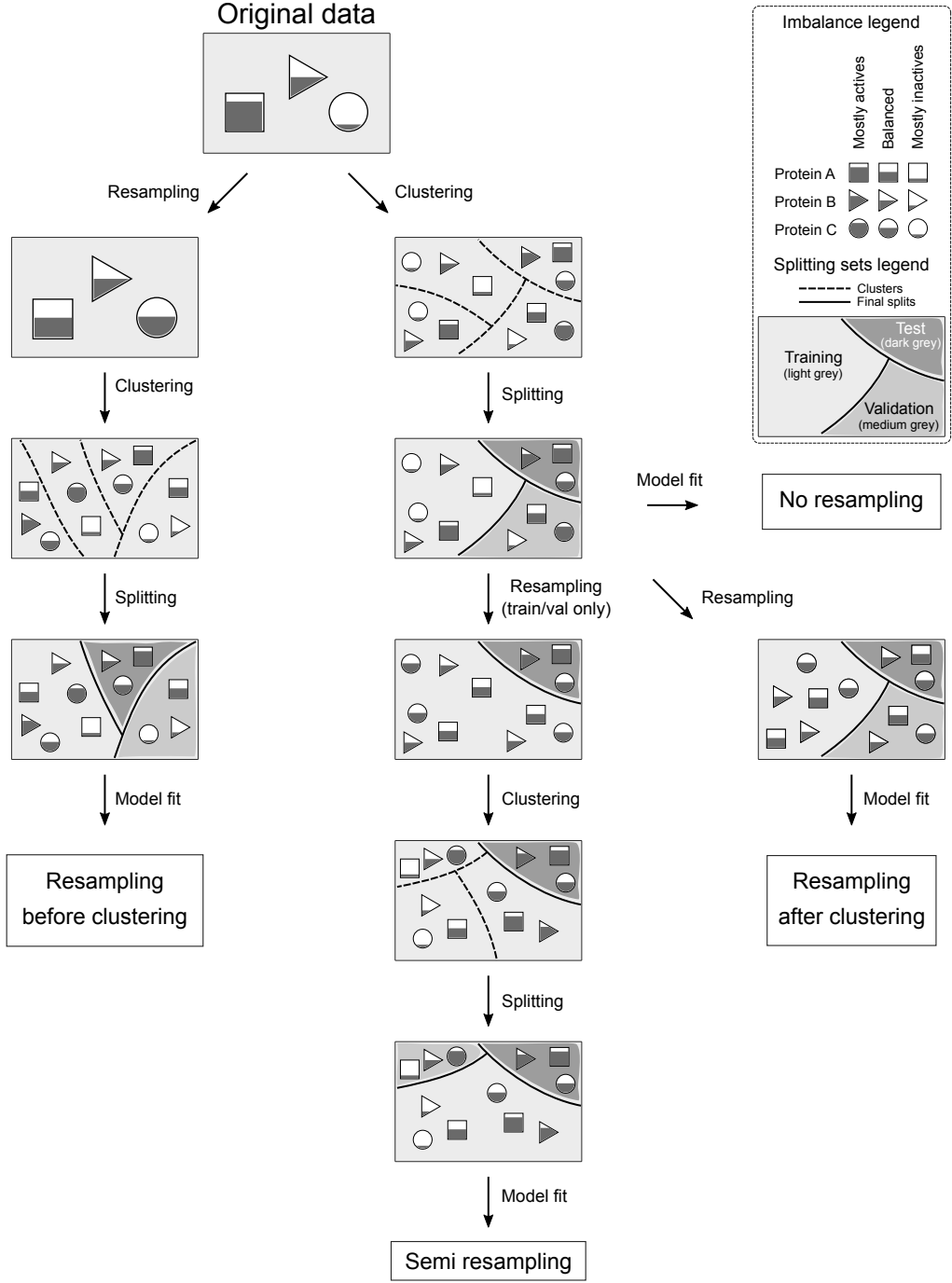
Original data
Resampling
Clustering
Clustering
Splitting
Splitting
Model fit
Resampling before clustering
Imbalance legend
Mostly actives
Balanced
Mostly inactives
Protein A
Protein B
Protein C
Splitting sets legend
Clusters
Final splits
Test
(dark grey)
Training
(light grey)
Validation
(medium grey)
Model fit
No resampling
Resampling
(train/val only)
Resampling
Clustering
Splitting
Model fit
Semi resampling
Resampling
Model fit
Resampling after clustering