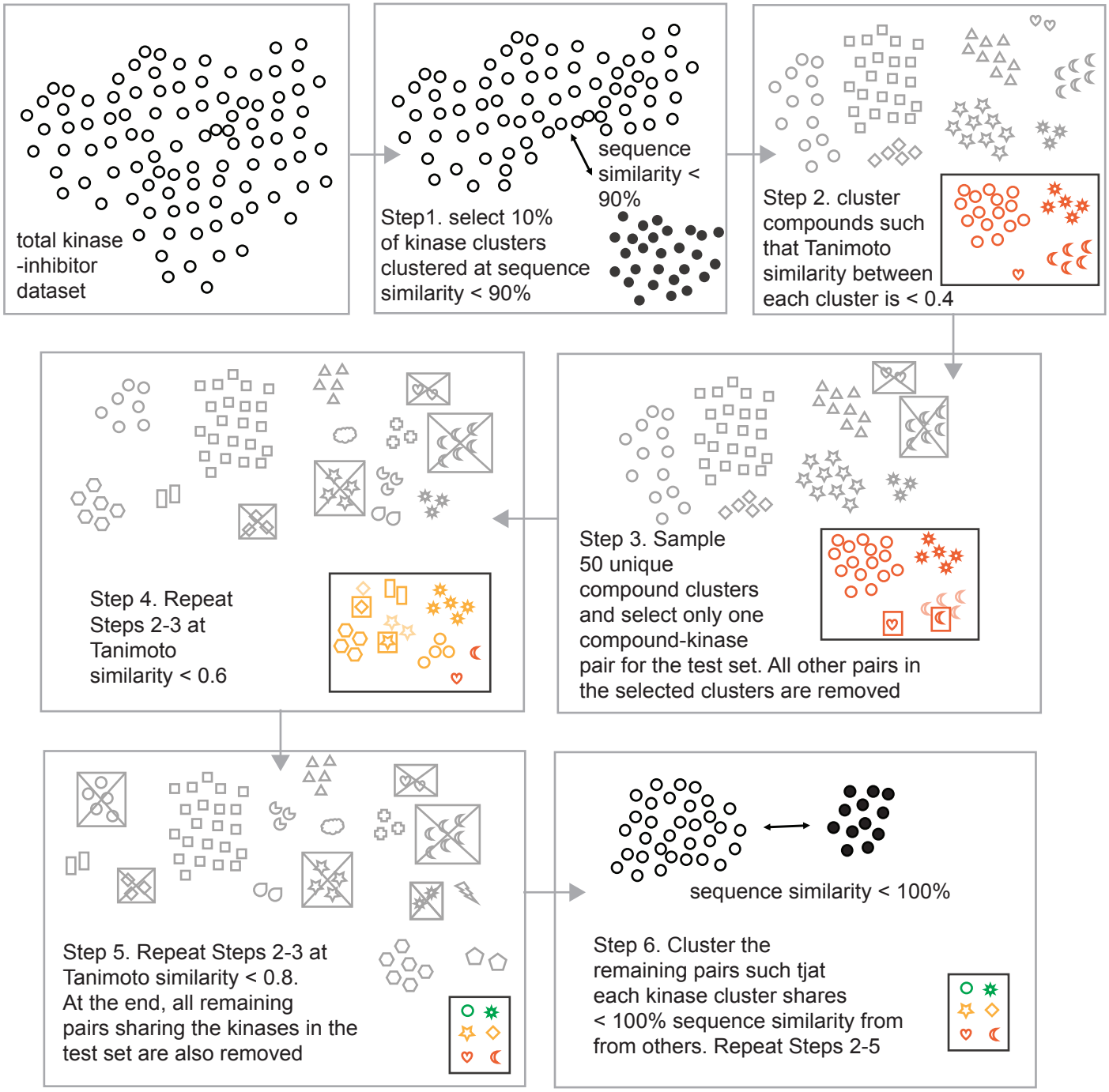
total kinase
-inhibitor
dataset
sequence
similarity <
90%
Step1. select 10%
of kinase clusters
clustered at sequence
similarity < 90%
Step 2. cluster
compounds such
that Tanimoto
similarity between
each cluster is < 0.4
Step 3. Sample
50 unique
compound clusters
and select only one
compound-kinase
pair for the test set. All other pairs in
the selected clusters are removed
Step 4. Repeat
Steps 2-3 at
Tanimoto
similarity < 0.6
Step 5. Repeat Steps 2-3 at
Tanimoto similarity < 0.8.
At the end, all remaining
pairs sharing the kinases in the
test set are also removed
sequence similarity < 100%
Step 6. Cluster the
remaining pairs such tjat
each kinase cluster shares
< 100% sequence similarity from
from others. Repeat Steps 2-5