Brief description of target folding calculations (by Andrey G.)

Folding requires NxM input matrix where N is number of targets and M is number of compounds. Value of matrix indicates presence of compound in particular target
Result of calculation is K1xM Fold data matrix where K1 is number of outer folds. Values of matrix are in range of 0 to K2 where K2 is number of inner folds - in other words inner fold index.

Fold data matrix - calculated by FoldGenerator

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	2	0	2	3	3	0	0	0	3	2	4	3	5
2	2	5	3	0	3	0	3	1	2	1	5	5	4	0
3	5	2	3	1	0	5	4	1	1	0	5	3	0	4
4	5	3	2	1	4	4	5	1	5	5	4	0	5	2
5	0	0	4	4	5	3	5	4	3	3	0	4	3	1

data that we have calculated indexes of inner folds - starts from 0 ends with innerFoldsNum
innerFold index = 0 means that we select outer fold only
indexes of outer folds - start from 1 ends with outerFoldsNum
ids of compounds - or ids of clusters

Selecting OUTER fold

for specified pair of outer and inned fold indexes we can get desired train and test data for desired fold
for example 2:0 - outer fold
select 2 row of fold data matrix

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
2	2	5	3	0	3	0	3	1	2	1	5	5	4	0

for each column value check
if innerFold index == 0 - compound belongs to test data (or V - validation data)
if innerFold index != 0 - compound belongs to train data (or T)

train data
test data

0	1	2	3	4	5	6	7	8	9	10	11	12	13
T	T	T	V	T	V	T	T	T	T	T	T	T	V

Selecting INNER fold

for specified pair of outer and inned fold indexes we can get desired train and test data for desired fold
for example 2:1 - inner fold
select 2 row of fold data matrix

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
2	2	5	3	0	3	0	3	1	2	1	5	5	4	0

for each column value check
if innerFold index == 1 - compound belongs to test data (or V - validation data)
if innerFold index != 1 && innerFold index != 0 - compound belongs to train data (or T)

train data
test data
not selected

0	1	2	3	4	5	6	7	8	9	10	11	12	13
T	T	T	N	T	N	T	V	T	V	T	T	T	N

Filter targets (for outer fold example)

Now suppose we need to filter particular target which has some compounds (according to desired outer and inner fold)

0	3	4	7	10	11	13

After filtering target will look the following way

train part	0	4	7	10	11
test part	3	13			

What if we have clusters? (for outer fold example)

Then clumns of Fold data matrix will correspond to cluster indexes
Selecting the same row we have

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
2	2	5	3	0	3	0	3	1	2	1	5	5	4	0

cluster indexes

Repeating selection we have

0	1	2	3	4	5	6	7	8	9	10	11	12	13
T	T	T	V	T	V	T	T	T	T	T	T	T	V

train data
test data

We need to have a table that allows mapping from cluster to compound ids
For example:

0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9	10	10	11	11	12	12	13	13
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27

Then we can determine which compounds belong to V and which to T

0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9	10	10	11	11	12	12	13	13
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
T	T	T	T	T	T	V	V	T	T	V	V	T	T	T	T	T	T	T	T	T	T	T	T	T	V	V	

Now filtering is obvious

0	3	4	7	10	11	13

After filtering target will look the following way

train part	0	3	4	13
test part	7	10	11	