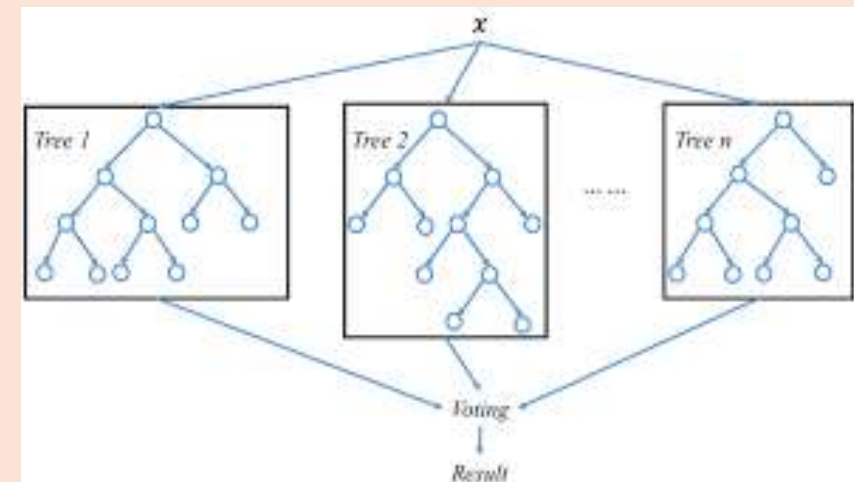


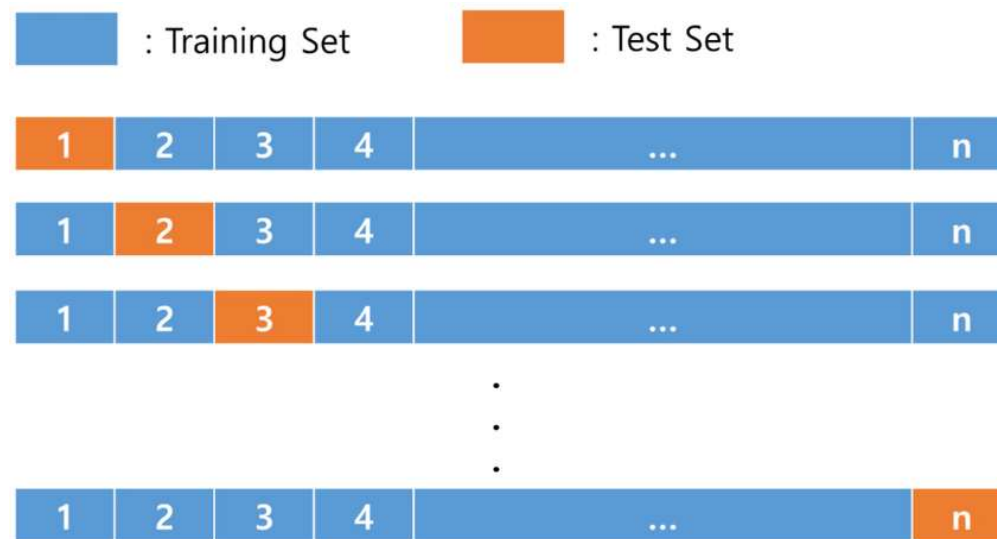
Random Forests

Random Forests work by creating multiple decision trees, each based on a random subset of your data and using a random subset of features. It keeps track of how often different variables are used to make decisions. By averaging the importance of each variable across all the trees, Random Forests can highlight which variables consistently have the most impact.

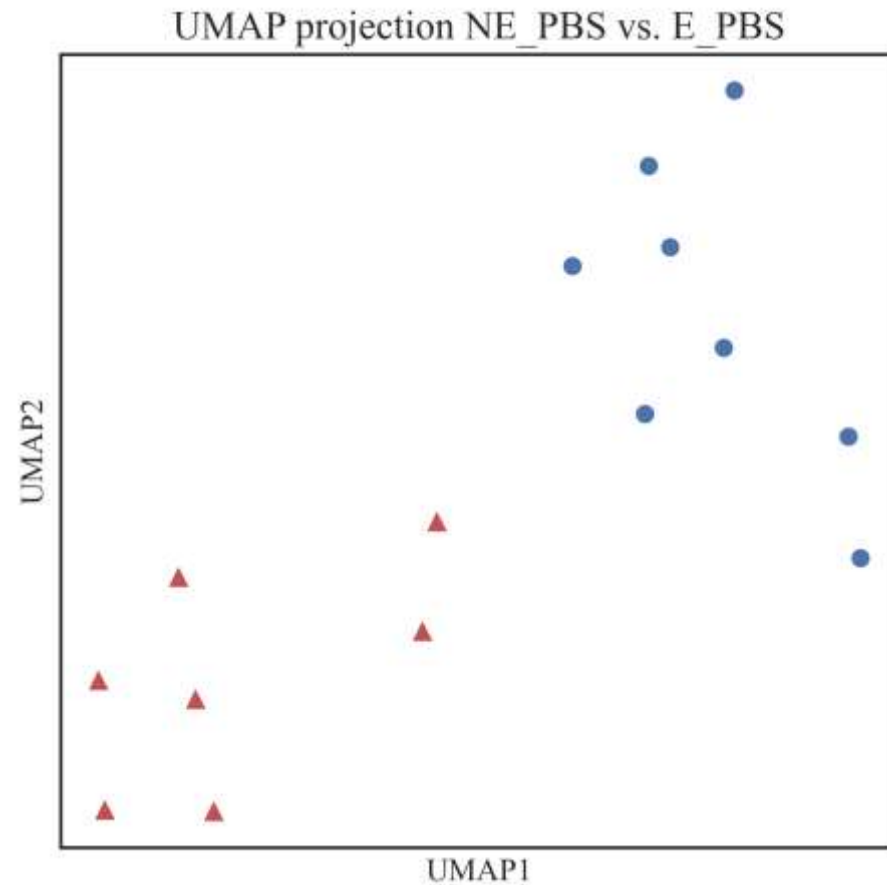


Leave-One-Out Cross-Validation (LOOCV)

- LOOCV uses almost all your data to train the model, except for one single data point, which it uses to test the model. This process is repeated for each data point in your dataset, which is important when you don't have a lot of data to work with.



Feature selection



Top 10 features

NE_PBS vs. E_PBS

```
{'ENSMUSG00000108957': 'Gm45235',  
'ENSMUSG00000026579': 'F5',  
'ENSMUSG00000067201': 'H2-M9',  
'ENSMUSG00000109539': 'Gm44667',  
'ENSMUSG00000052724': 'Gm9888',  
'ENSMUSG00000048304': 'Slitrk3',  
'ENSMUSG00000104585': '4921511C10Rik',  
'ENSMUSG00000034579': 'Pla2g3',  
'ENSMUSG00000079711': 'Smok4a',  
'ENSMUSG00000021384': 'Susd3'}
```

Integrating multiple omics datasets

Multomics dataset

Proteomics (tissue)

	Ube2o	Bckdk	Adprh	Adprhl1	Pdk1
NE PBS 1	0.506317	1.056973	-1.820748	2.774022	1.173897
NE PBS 2	0.113496	0.887992	-0.870756	-0.706154	-0.594518
NE PBS 3	-1.850081	0.108964	0.918366	-0.492800	-0.274472
NE PBS 4	-1.275752	0.735713	-1.837408	-0.279457	2.042321
NE PBS 5	0.997790	2.610157	-0.054609	0.249857	1.510955

Transcriptomics (tissue)

Gm37587	Gm37248	Gm10479	Fibcd1	Gm14853
1.031664	0.464484	0.884491	0.215036	1.319293
1.031664	0.464484	0.884491	2.406601	-0.956248
1.194045	1.952855	1.006227	0.286688	1.486689
1.121346	1.867391	1.801558	0.254609	1.411745
-0.586423	-0.803655	-0.751643	-0.498968	0.232976

	Dataset	Accuracy
Prot without feature selection		0.200000
Prot with feature selection		0.866667
RNA without feature selection		0.466667
RNA with feature selection		1.000000
Multiomic without feature selection		0.400000
Multiomic with feature selection		1.000000

Results
(NE_PBS vs. E_PBS)

No improvement
over transcriptomic

- Prot stands for proteomics data gathered from the tissue
- RNA stands for the transcriptomics data gathered also from the tissue
- Multiomic stands for the union of both datasets

	Dataset	Accuracy
Prot without feature selection		0.5000
Prot with feature selection		0.7500
RNA without feature selection		0.2500
RNA with feature selection		0.9375
Multiomic without feature selection		0.4375
Multiomic with feature selection		0.8125

93.75% cross-validation score for transcriptomic data after feature selection

Decrease in cross-validation score in relation to transcriptomic

Results
(NE_PBS vs. NE_AmAc)