

Génération de nouvelles visualisations via la corrélation entre features et scores des algorithmes (méthode linéaire)

Le pipeline de sélection des features se déroule en deux phases principales :

1. Phase "corr" (implémentée dans la fonction `checkCorrelation.m`)

Dans cette phase, nous analysons pour chaque algorithme quelles sont les variables (features) qui présentent la plus forte corrélation avec son score global (ExpGlobal). Autrement dit, pour chaque algorithme, nous identifions les features qui semblent avoir le plus d'influence sur sa performance.

- **Paramètre associé : `data.opts.corr.threshold`**

Ce paramètre fixe le nombre de features à retenir pour chaque algorithme. Par exemple, si `threshold` est égal à 5, pour chaque algorithme nous garderons les 5 features dont la corrélation avec le score est la plus élevée.

2. Phase "corr2" (implémentée dans la fonction `checkCorrelation2`)

Ici, nous comparons les performances entre chaque paire d'algorithmes. Pour chaque paire, nous déterminons quelle feature est la plus corrélée avec la différence de performance entre ces deux algorithmes.

- **Paramètre associé : `data.opts.corr2.top`**

Ce paramètre indique combien de features seront retenues pour chaque paire d'algorithmes. Par exemple, si `top` est fixé à 5, nous garderons les 5 features les mieux corrélées avec la différence de score entre chaque paire d'algorithmes.

Deux autres paramètres viennent compléter ce système :

- **`data.opts.corr.flag`**

Ce paramètre permet d'activer ou de désactiver la phase de sélection par corrélation.

- Si on le règle sur `false`, la phase "corr" ne sera pas exécutée : au lieu de filtrer et de ne retenir que les N features les plus corrélées pour chaque algorithme, aucune sélection basée sur la corrélation ne sera réalisée.

- **`data.opts.clust.flag`**

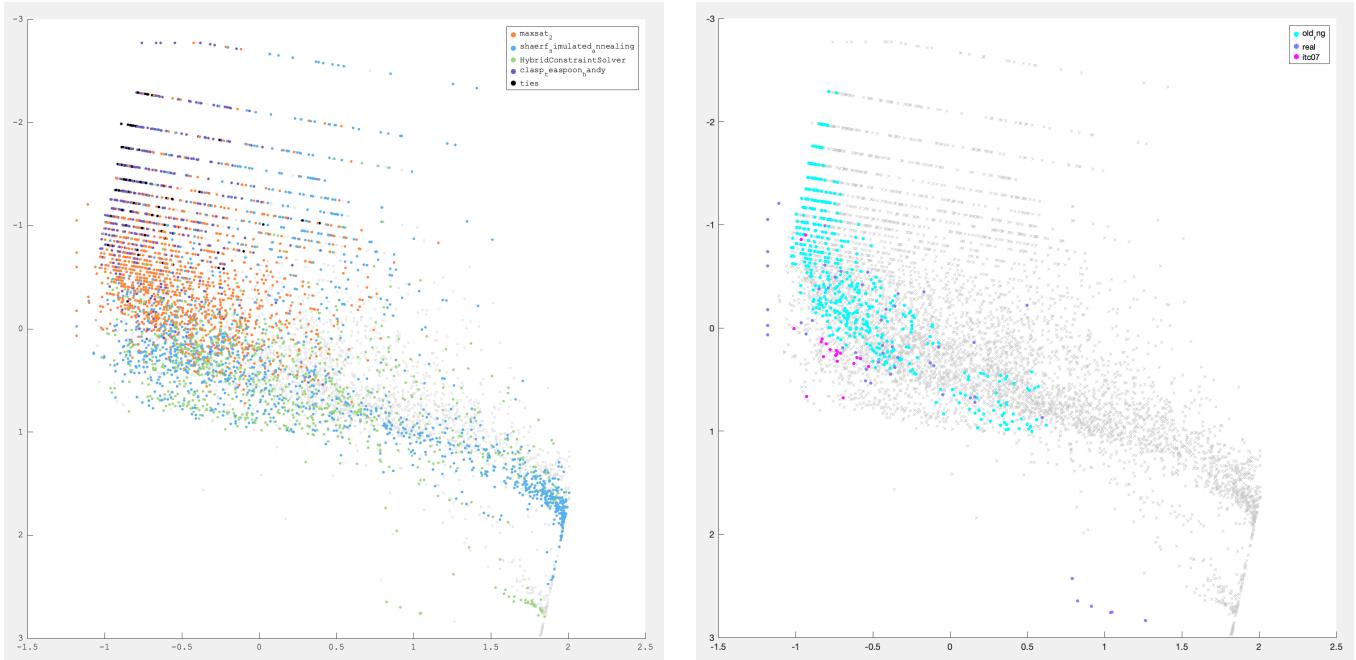
Ce paramètre contrôle le regroupement (clustering) des features.

- S'il est réglé sur `false`, le clustering est désactivé et aucune réduction supplémentaire des features par regroupement de celles jugées trop similaires n'est effectuée.

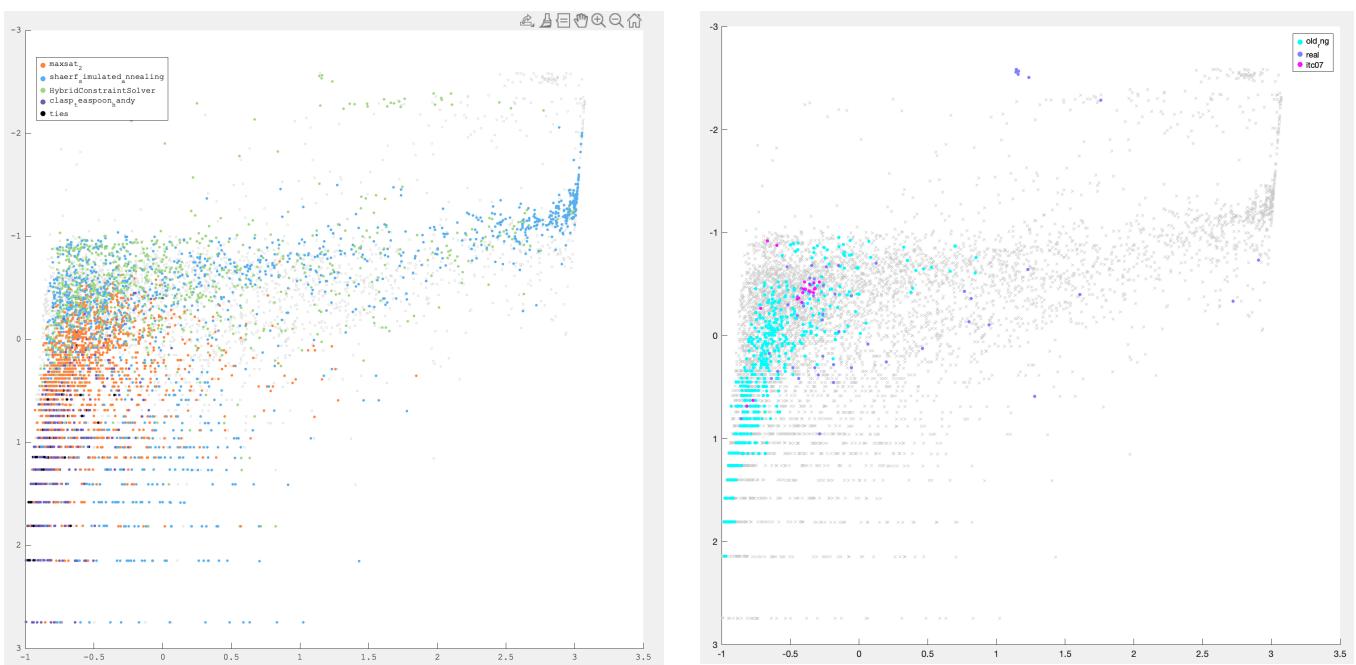
On peut ainsi choisir de réduire drastiquement le nombre de variables en ne gardant que celles les plus corrélées (et/ou celles différentiant les performances entre algorithmes), ou bien de laisser passer plus de données en désactivant ces filtres.

clust_true_corr_true_corr2_true

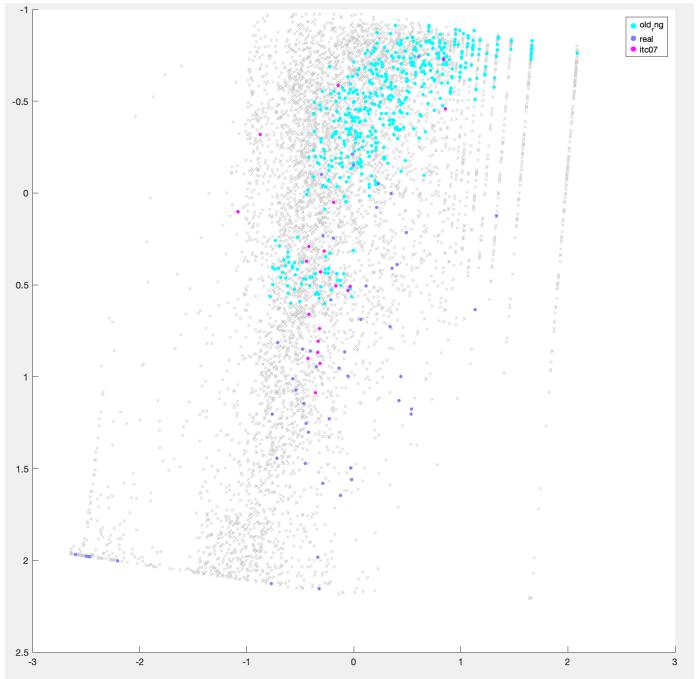
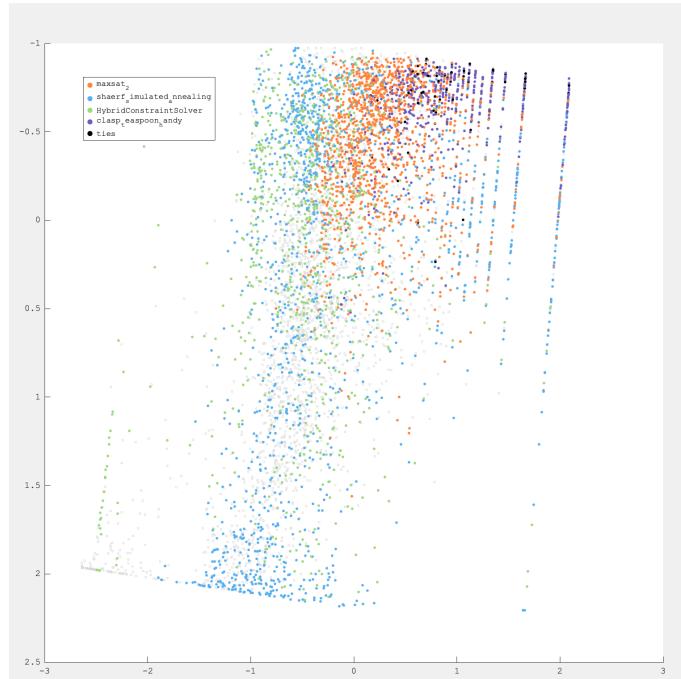
threshold_1_top_10 : 2 features



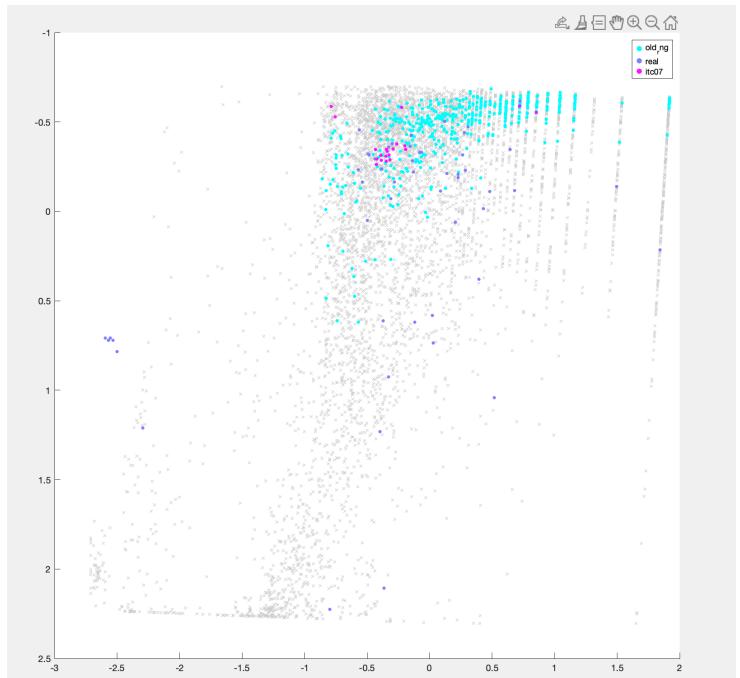
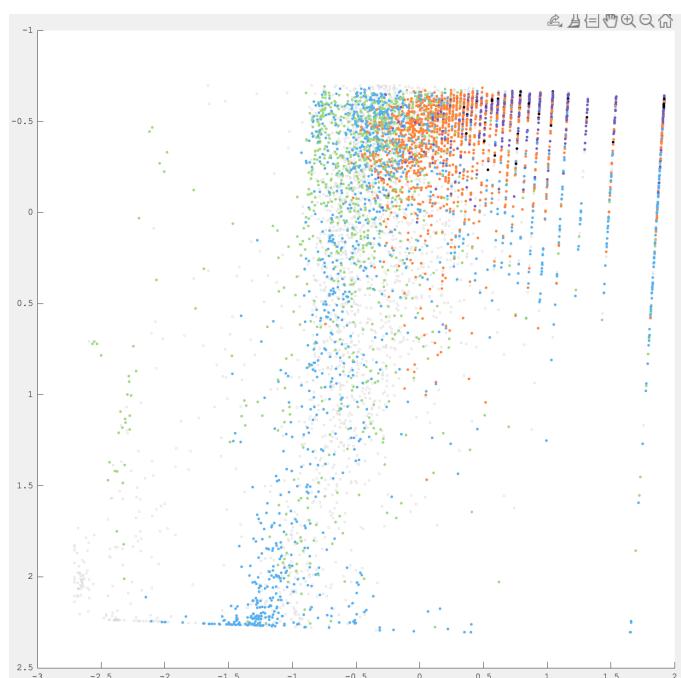
threshold_2_top_1 : 2 features



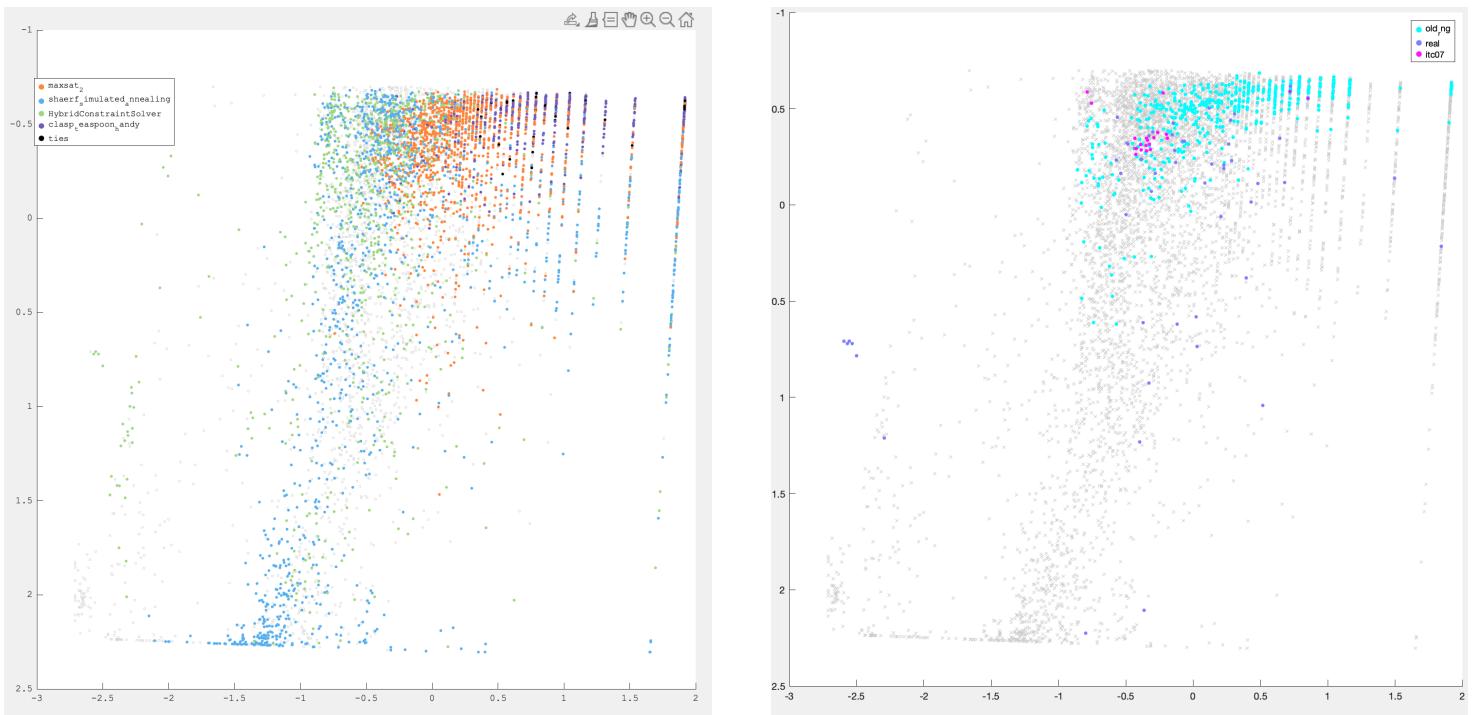
threshold_5_top_100 : 2 features



threshold_10_top_2 : 2 features

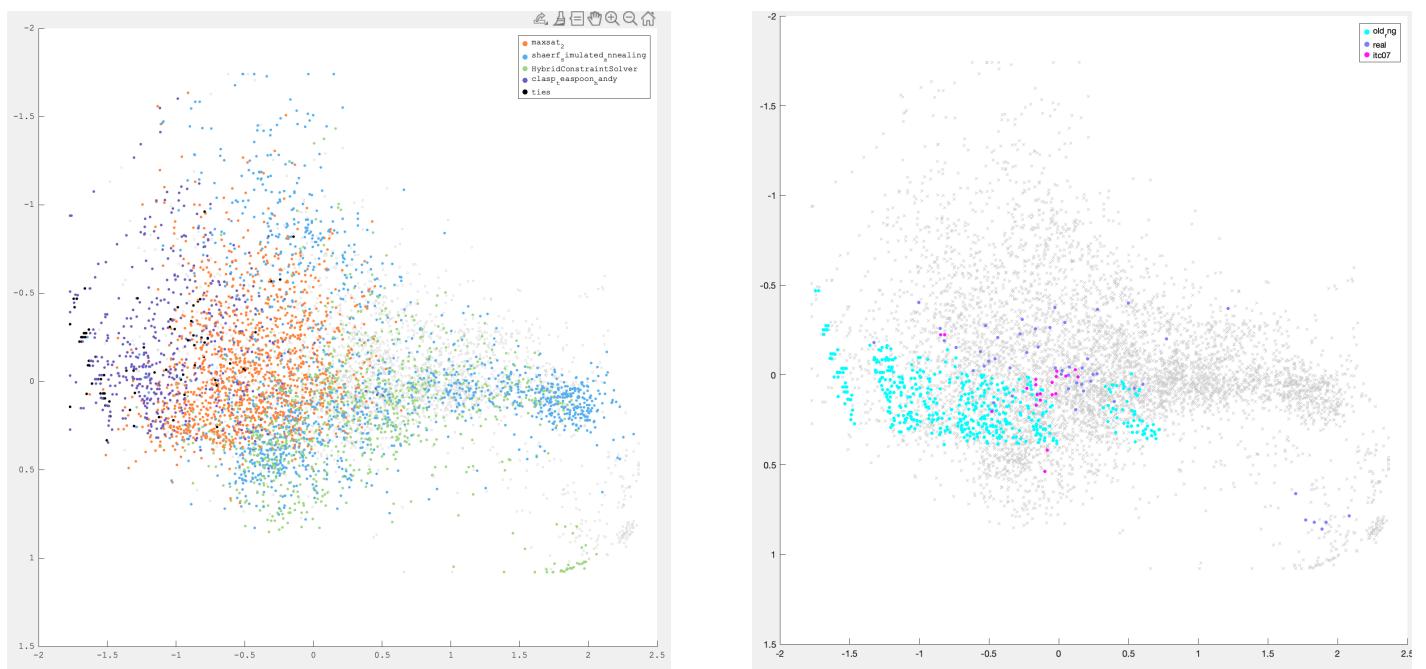


threshold_100_top_3 : 2 features

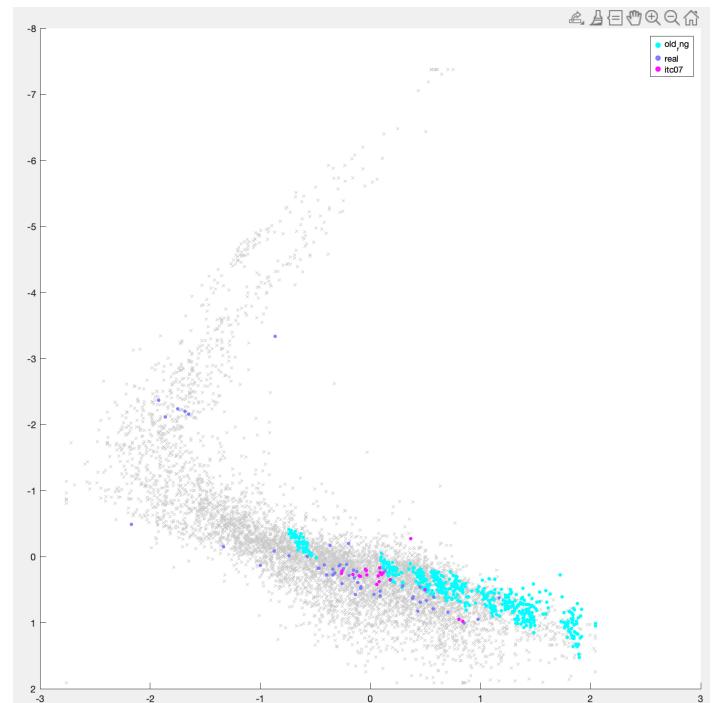
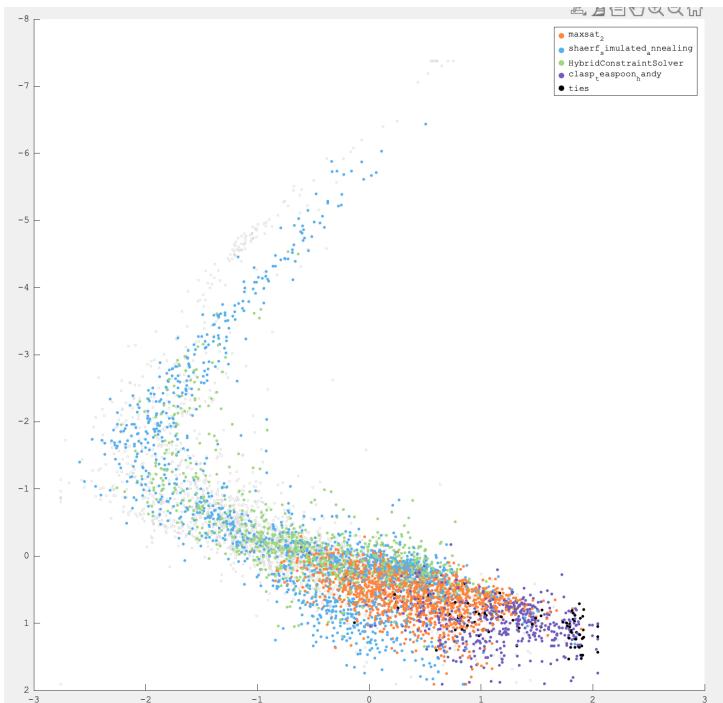


clust_false_corr_true_corr2_true

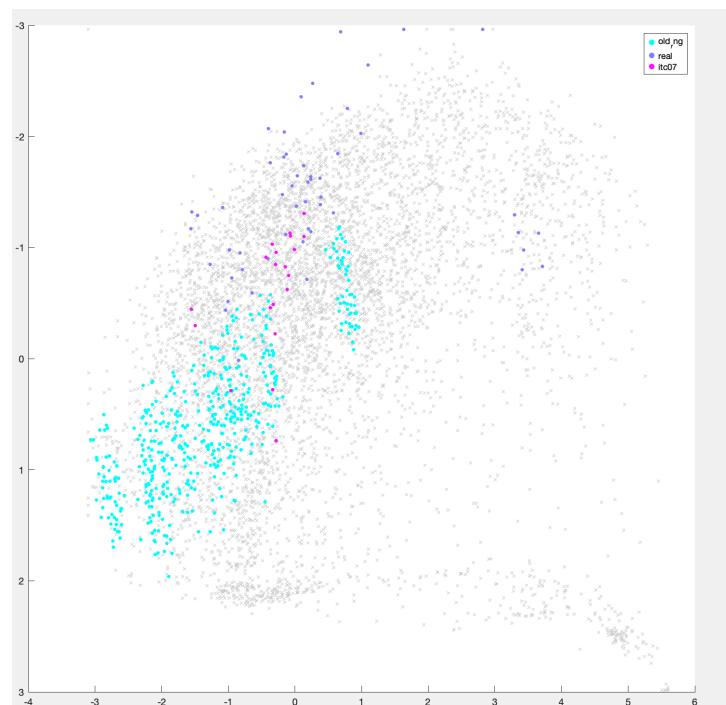
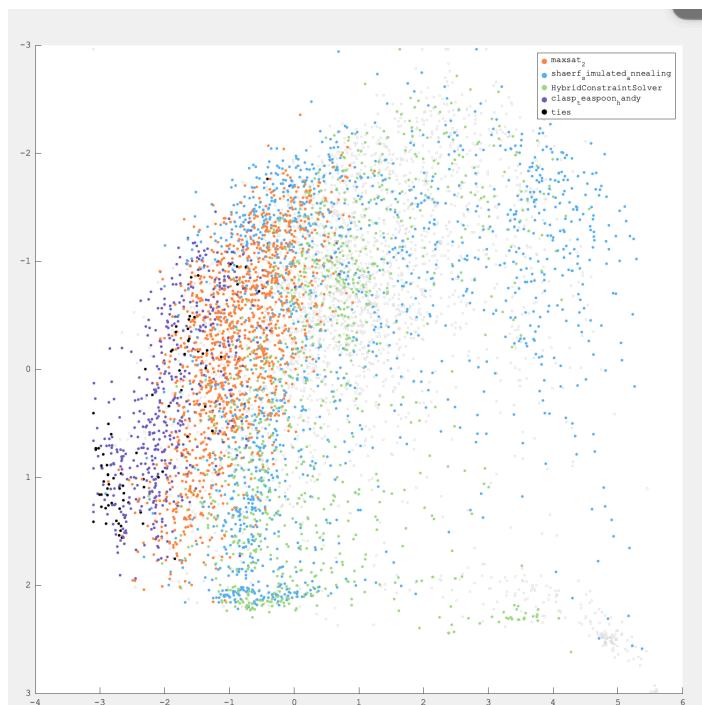
threshold_1_top_5, threshold_1_top_10, threshold_1_top_10 et threshold_1_top_100 : 3 features



threshold_5_top_1 : 5 features



threshold_10_top_1 et threshold_20_top_1 : 7 features



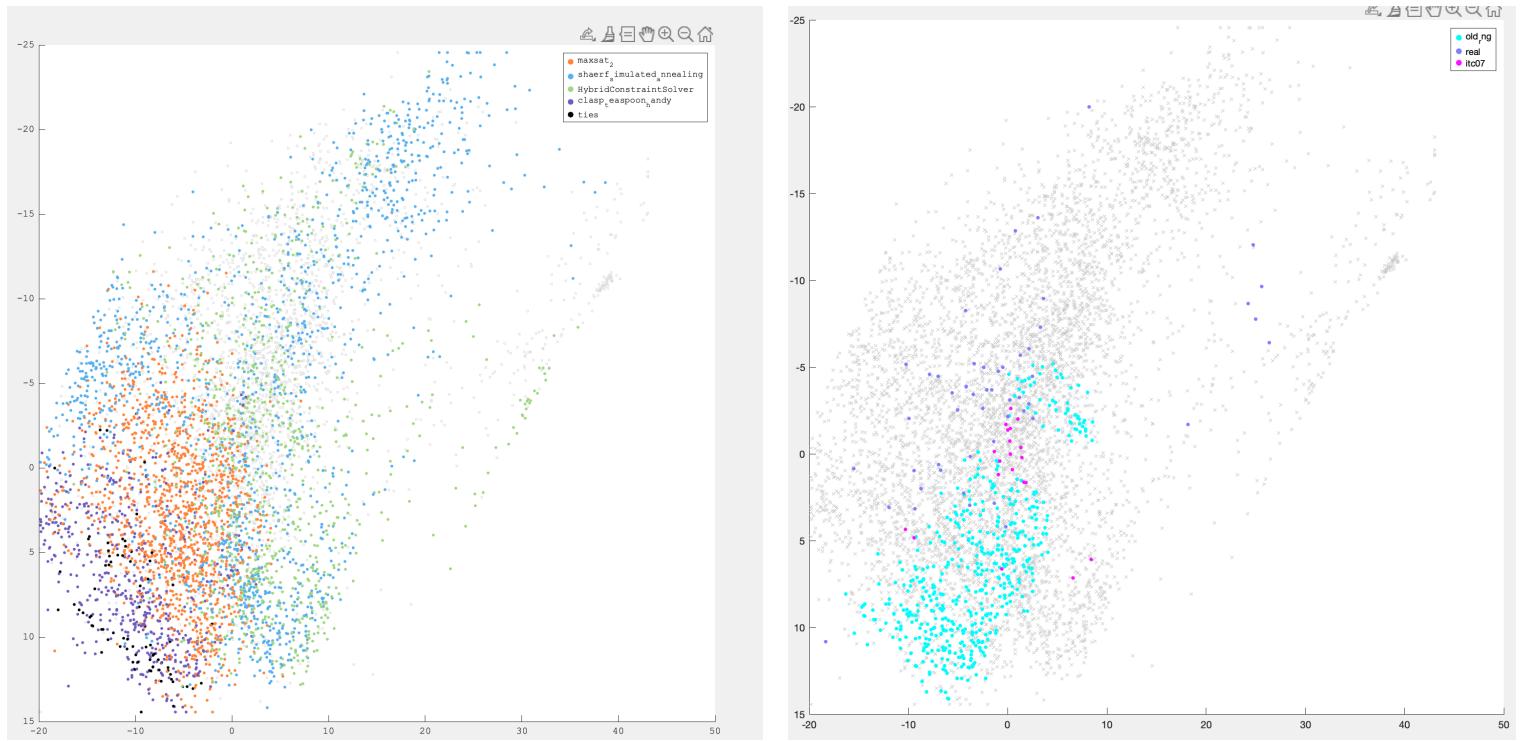
**threshold_5_top_5, threshold_5_top_10 et
threshold_5_top_100 : 15 features**



threshold_10_top_5 : 19 features



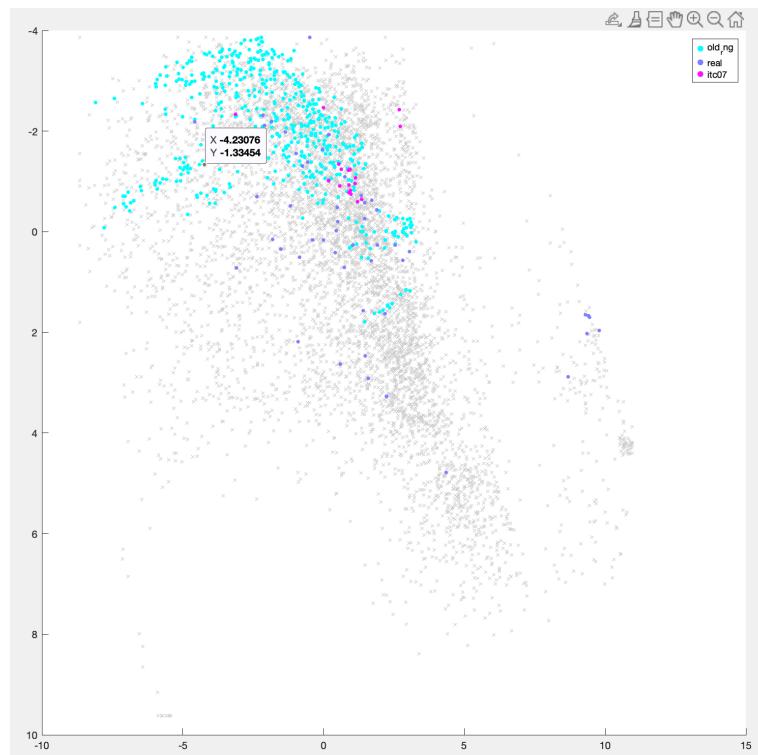
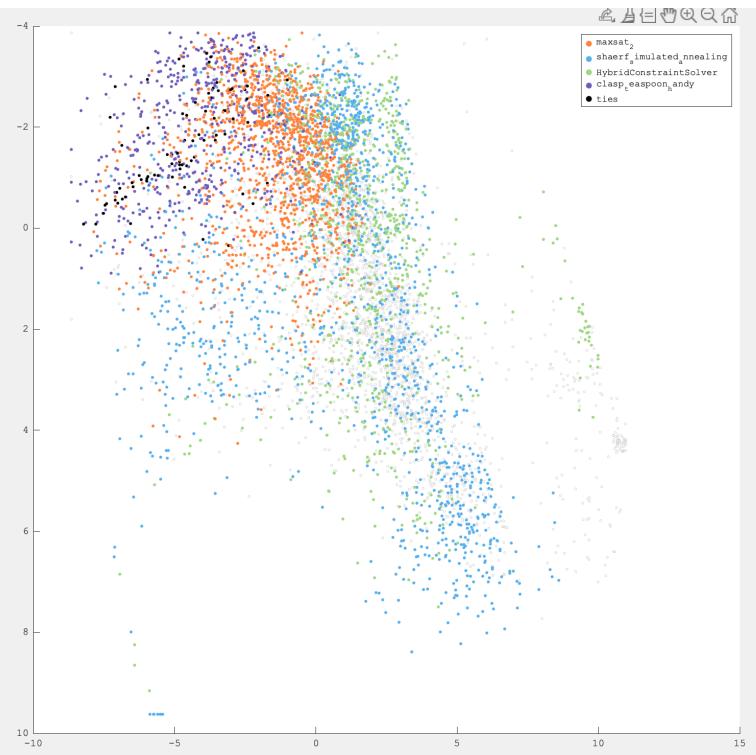
threshold_20_top_5 : 23 features



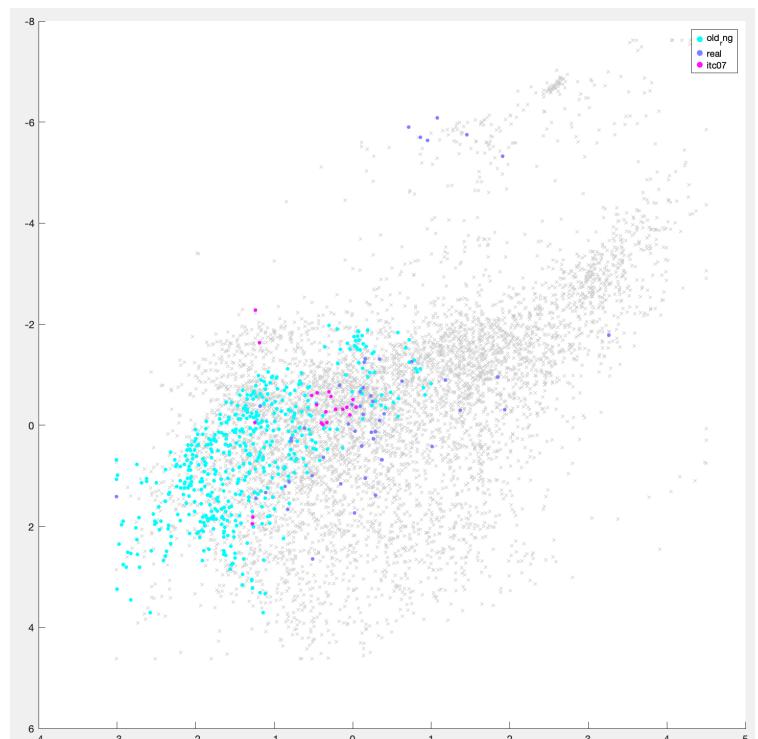
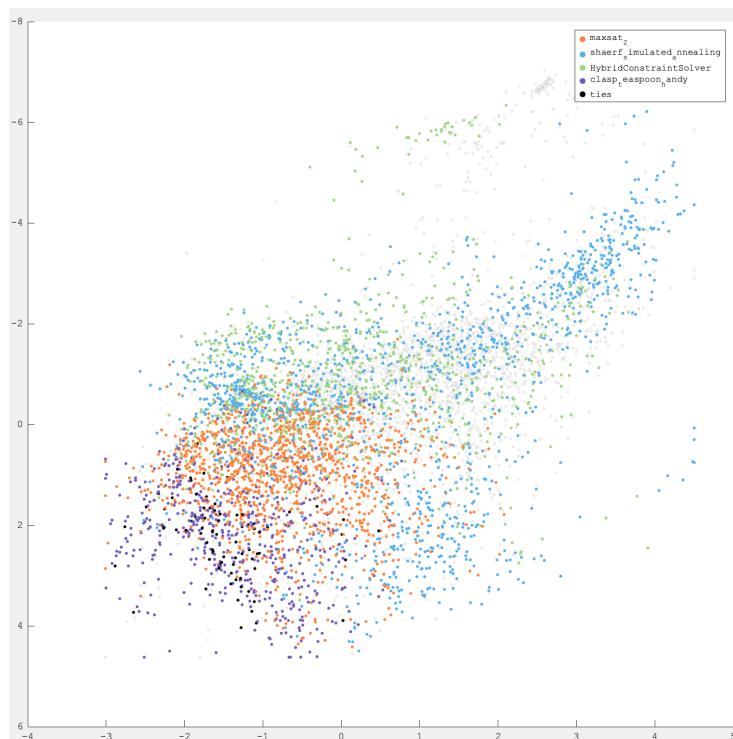
threshold_10_top_10 : 27 features



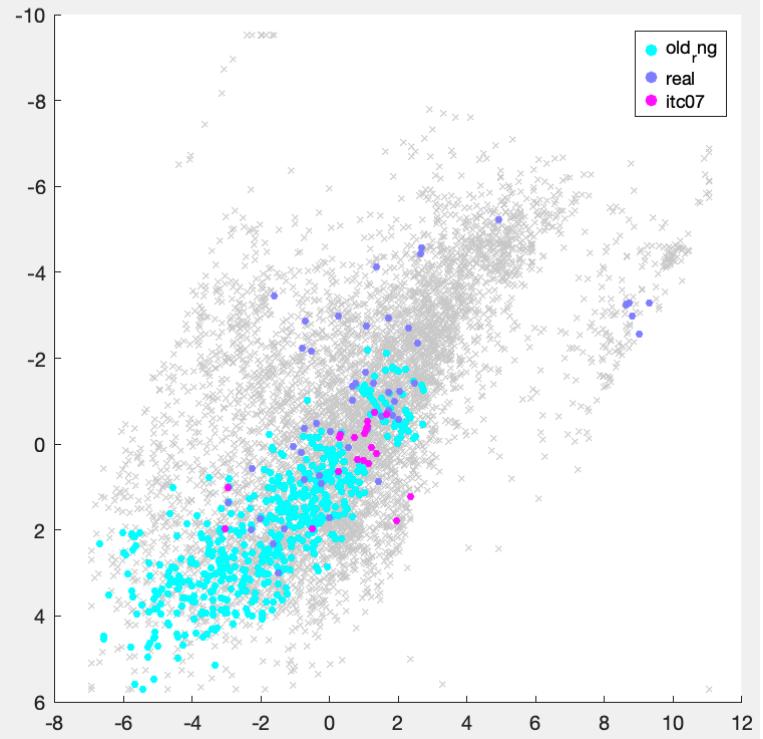
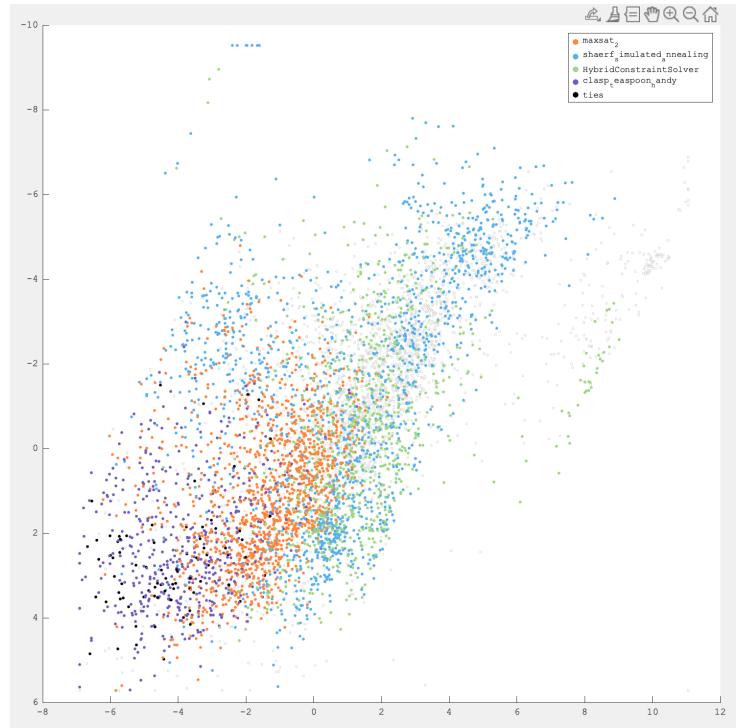
threshold_10_top_100 : 28 features



threshold_20_top_10 : 34 features

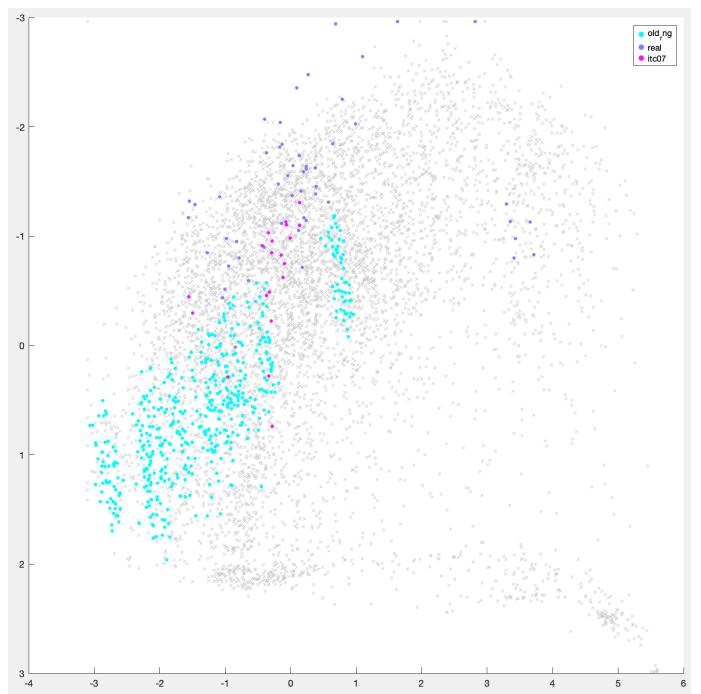
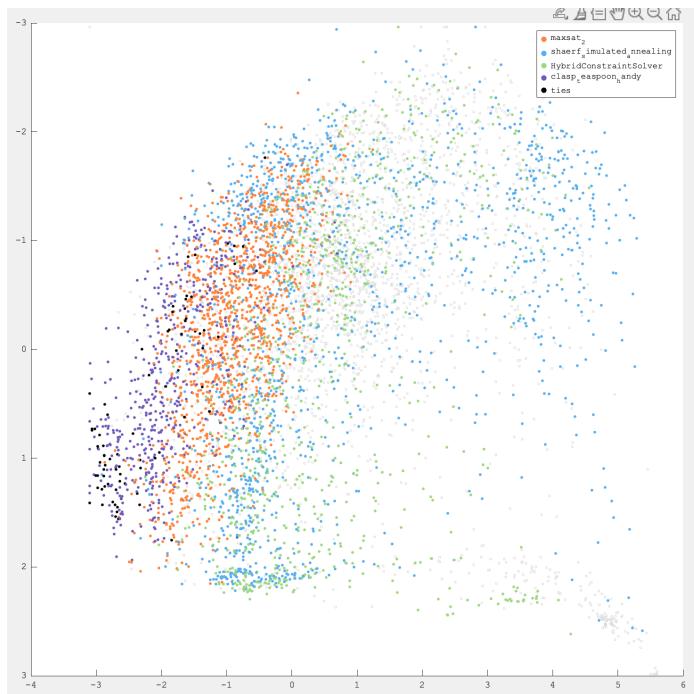


threshold_20_top_100 : 60 features

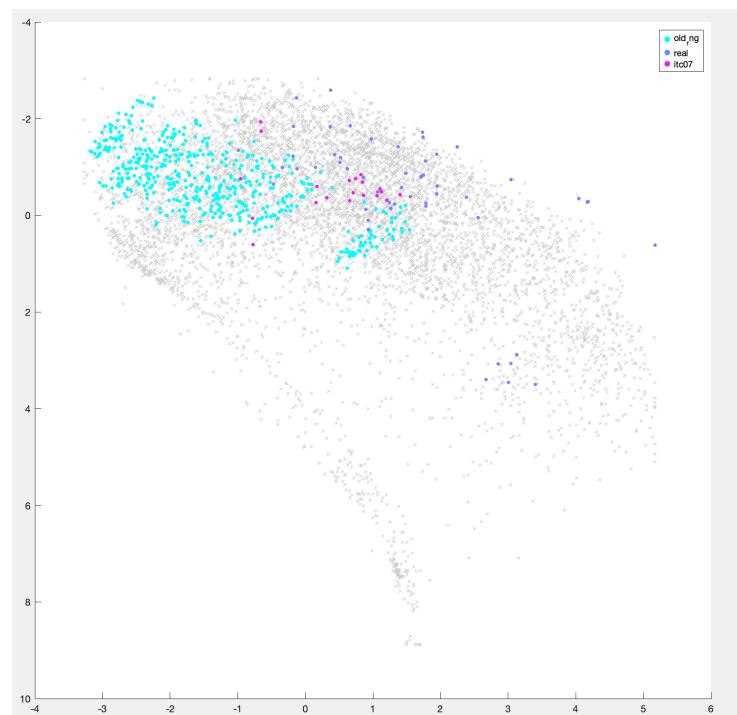
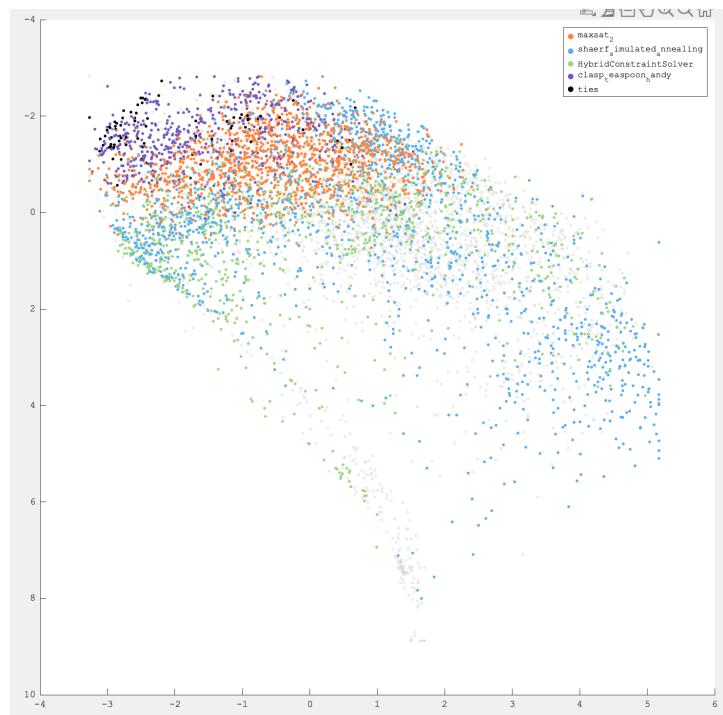


clust_false_corr_false_corr2_true

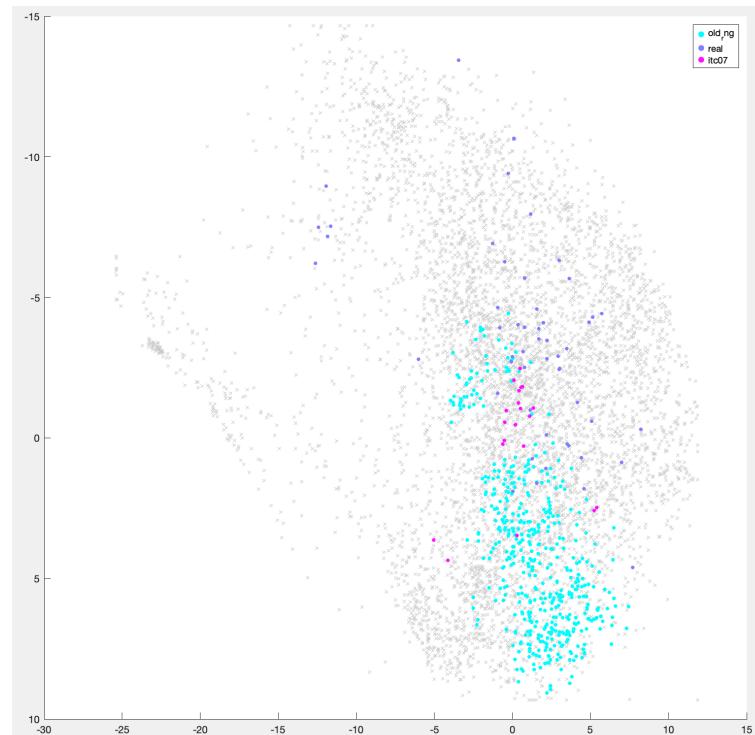
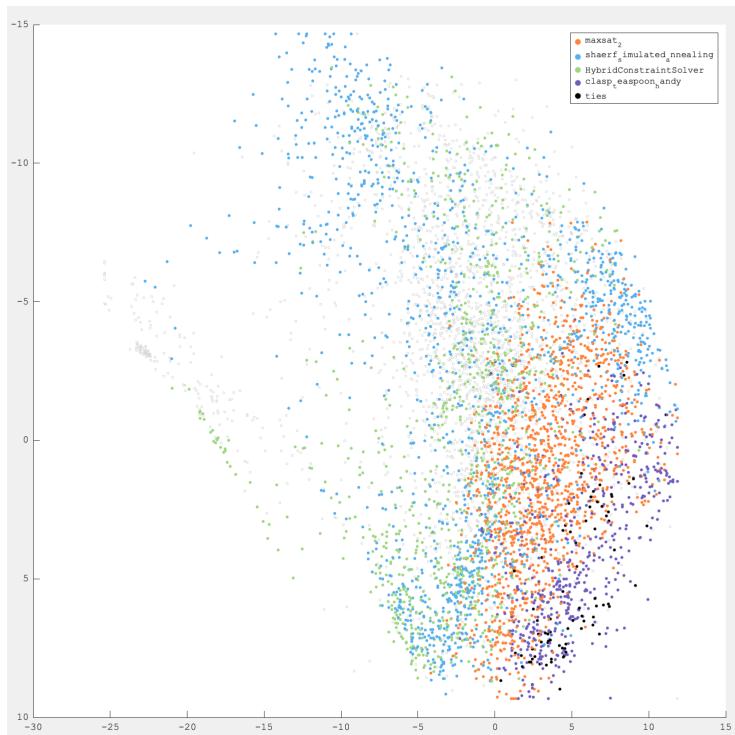
top_1 : 7 features



top_2 : 13 features



top_5 : 27 features



top_10 : 56 features

