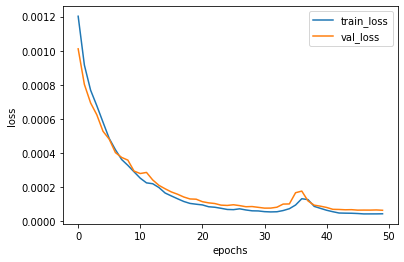
Автоэнкодер в распознавании MS-спектров

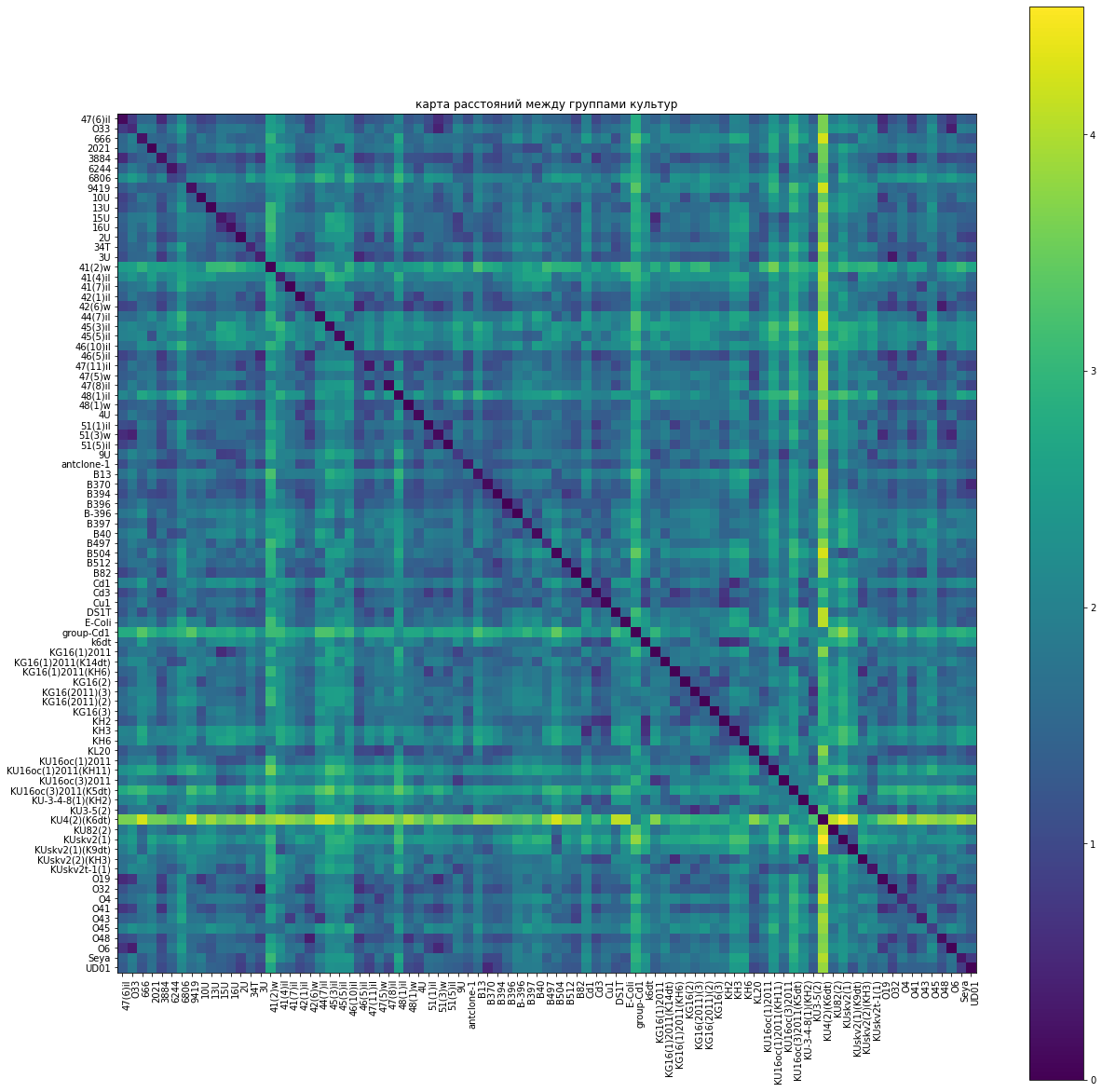
Было написано и обучено два автоэнкодера для этой цели. Первый я решила сделать сааамый простой “ванильный” автоэнкодер: 3 полносвязных линейных слоя в энкодере и декодере, функция активации не линейная – relu, функция потерь – MSE, Размерность скрытого пространства – 50.

Для обучения размеченных профилей было всего 152 штуки. Как правило синтетических данных добавляют не более 30-40%, но мне пришлось сделать датасэт из 1064 профилей (то есть ~ 85% данных синтетические). Даже 1000 примеров очень мало для обучения крупных сетей но мои автоэкодеры очень маленькие и пока я тренировала их на 1000 профилей (я очень надеюсь, что найдутся потерянные 150 профилей, тренировочный дата сэт станет в два раза больше)

Эпох в обучении было 50, график значений функции потерь за каждую эпоху для ванильного автоэкодера:



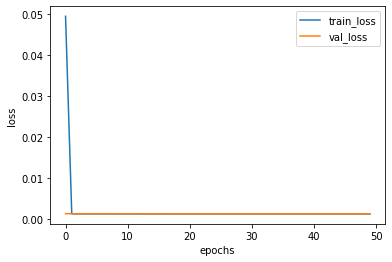
Далее, для наглядности я решила посмотреть на эвклидово расстояние между группами бактерий в скрытом пространстве. На этой heat карте цветом показаны расстояния между средними векторами для каждой группы, а на диагонали стоят средние расстояния между векторами одной группы (здесь есть одна проблема, все это я считаю на value data set он представляет примерно 30% от общего дата сэта (который 1000 шт), некоторые бактерии сюда вошли в одном экземпляре и при расчете на диагонали получались несколько NaN-ов, которые я бессовестно заменила на 0, здесь они не портят иллюстративности, но на следующем автоэнкодере это всплывет)



Видно что, межгрупповое расстояние много меньше расстояния до других групп, то есть задача классификации должна упрощаться

Дальше я написала вариационный автоэнкодер (VAE) и он уже выглядит не так привлекательно как обычный. Его идея в том, чтобы пространство скрытого состояния было с нормальным распределением. Это добавляет “непрерывности”, то есть выше вероятность того, что случайная точка из этого пространства будет иметь внятную интерпретацию от декодера. Так же, по три линейных полносвязных слоя в енкодере и декодере, но в енкодере есть дополнительный слой, активация – relu, функция потерь MSE + KL-дивергенция.

Во-первых, график потерь очень странный, я меняла архитектуру сети, но он упорно такой. У меня есть предположение, что такой тип автоэнкодеров не терпит синтетические данные он видит, что типов профилей всего 152 и сразу на них переобучается.



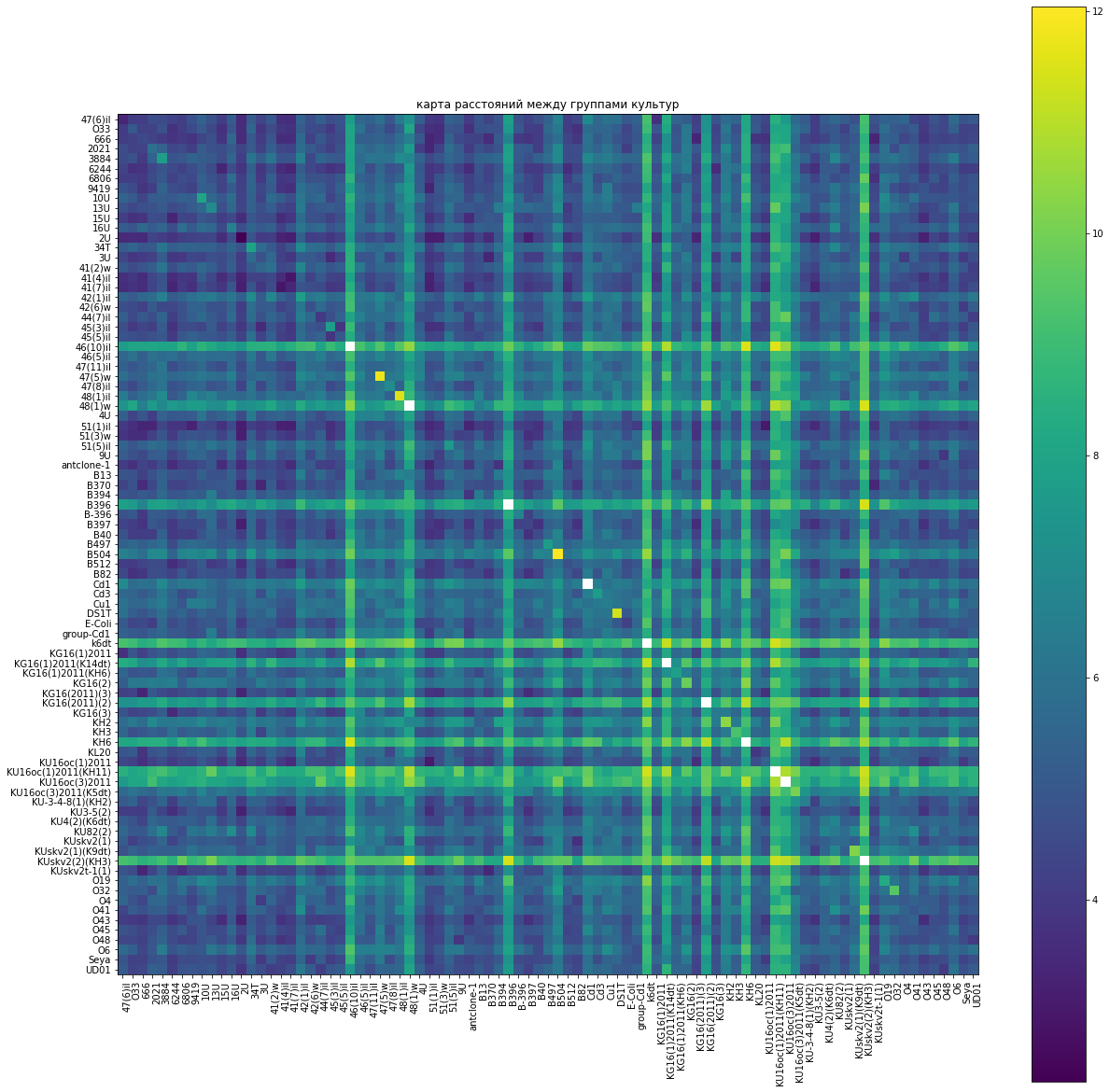
Ну и, во-вторых, heat map тоже не радует

График тут не очень красивый, NaN не меняла на 0. Межгрупповое расстояние примерно такое же, как и расстояние между группами, то есть, по-моему, задача классификации должна усложниться. Поэтому с VAE дальше работать не буду.

Теперь главный вопрос: какой основной аргумент за использование автоэнкодера? Я могу ванильный натренировать так, чтобы он убирал шумы или что бы эвклидово расстояние в исходном и скрытом пространствах сохранялись (улучшит классификацию геометрическим методом). На самом деле можно попробовать сделать и то, и то, но я думаю, что нужно делать узконаправленный автоэнкодер и выбрать что-то одно.