EPSRC
Engineering and Physical Sciences Research Council

InFoMM
Industrially Focused Mathematical Modelling

# EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling

## Matrix completion for drug discovery

Victoria Pereira

UNIVERSITY OF OXFORD

e-Therapeutics plc

# Contents

# 1. Introduction

## Background

A protein network is a set of proteins, treated as points, connected together by edges. Information, e.g. interactivity between two proteins is represented by the edges between proteins.

Drug discovery is the process of developing new medicines. One approach is to find bioactive compounds that interact with proteins (see Figure 1). It is well understood that proteins form interacting networks in the human body, and medicines work by perturbing these networks. e-Therapeutics is a drug discovery and development group, who aim to utilise the complexity of the protein networks to discover new drugs.

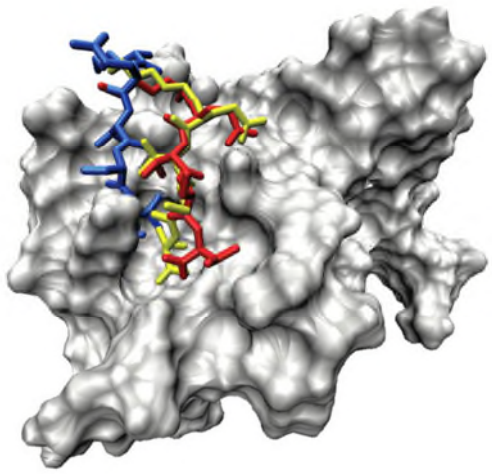

**Figure 1: Compound (the coloured structure) - protein (grey mass) interaction [2]. The bioactivity we are predicting assigns a likelihood to these interactions occurring.**

It is clear that an important part of drug discovery is to understand how compounds interact with proteins in the body. In an ideal world we would know how every compound interacts with every protein. However, due to the sheer number of compounds (millions) and proteins (tens of thousands), the experimental database is largely incomplete. In fact, we only have bioactivity data for around 0.03% of the compound-protein pairs. Such data is very expensive to generate, so our strategy is to predict the remaining missing entries in order to aid exploring possible compounds.

Clustering is the method of grouping 'similar' objects together; where similarity is defined by how close two objects are in a network.

A well-studied approach to predicting the bioactivity of compounds and proteins is to use their physical properties. However, there are compounds which are structurally similar but have very different bioactive behaviour, a phenomenon known as *activity cliffs*. This is a major difficulty for drug discovery approaches relying on physical properties to infer bioactivity. Activity cliffs could, however, be beneficial in isolating 'similar' groups of interacting compound-protein pairs, suggesting that clustering might be a useful tool for predicting bioactivity.

We will now introduce the problem, and then briefly describe the details of our approach, before presenting our results.

# 2. Mathematical approach

We used a public database of compound-protein interactions STITCH (search tool for interactions of chemicals), to formulate the compound-protein database as a matrix with compounds as rows and proteins as columns, as shown in Figure 2.

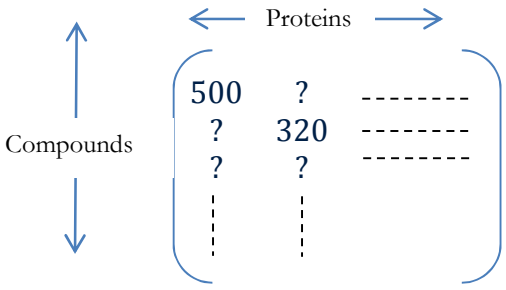


**Figure 2: Example matrix containing interaction data between proteins and compounds. Each column and row corresponds to a single protein and compound respectively. The '?' entries denote missing data.**

Each entry in our matrix has a numerical value between 1 and 1000 that represents a probabilistic score governing the likelihood of an interaction occurring between the compound-protein pair. However, as previously mentioned, we do not have a likelihood score for every compound-protein pair; hence we have many unknown or missing entries in our matrix and it is 'empty'. Our approach is to use matrix completion techniques to predict these missing entries.

However, the matrix is so large and empty that matrix completion techniques are not easily applied. Hence, we investigate clustering the rows and columns of this matrix to try and group together similar interacting compound-proteins pairs. This clustering process subdivides the large matrix into smaller submatrices that contain similar bioactivity data, hopefully separated by activity cliffs. We can then use matrix completion methods to predict the entries of these smaller matrices.

## Glossary of terms

- Activity matrix: The matrix with compounds as rows, proteins as columns containing the bioactivity interaction data.
- Clustering: A grouping of compounds or proteins by some similarity measure.
- Completing: Applying matrix completion to the sparse matrix to predict the missing values.

## Clustering

There are several classifications of compounds and proteins on which to apply clustering techniques. As previously noted, we can use physical properties to infer similarities, and hence we use the following properties to divide our compounds and proteins into groups:

- Protein families: groups of proteins that have the same evolutionary origin.
- Compound fingerprints: a representation of a compound by structural pieces.

The protein families yield a natural clustering, and we can find a clustering via compound fingerprints by matching those with similar structural features into clusters. We further investigate clustering of the bioactivity data independent of any physical properties. This gives us two more classifications on which we can cluster; bioactivity via compounds and via proteins, and we thus have four methods of clustering to compare.

We cluster the bioactivity data for each compound or protein based on:

- Where we have data, that is, what experiments have been done.
- How much data we have, that is, how many experiments have been done.
- What the value of data is, that is, the probabilistic score for each experiment.

## Completion

We need to now predict the missing entries of our activity matrix as a whole, and on the clusters found by one of the methods described above. There are several ways we can predict the missing values. The simplest approach is to use the mean values of the data that we do have. We can do this in three ways:

- Total mean: take the mean of all the entries we have and set all missing entries to be this total mean.
- Column mean: take the mean by column, and set all missing entries in each column to have the column mean.
- Row mean: take the mean by row, and set all missing entries in each column to have the row mean.

A more sophisticated matrix completion method imposes some mathematical structure in the activity matrix; in particular, we assume a low-rank structure. This means that there is a strong relationship between the columns and rows in the matrix. The method we use to complete the matrix of bioactivity data is called the Scaled Alternating Steepest Descent method (SASD) and was proposed by Tanner and Kei [1]. The method involves finding the matrix that best fits the known entries that we have whilst ensuring a low-rank structure. We develop a variant of SASD called Mean-Translated SASD (MT-SASD), in which we subtract mean values from each entry before completing.

Thus we have five matrix completion methods to compare together with our four clustering techniques.

# 3. Numerical results

We run numerical tests for all the combinations of clustering and matrix completion methods on a subset of all the data by selecting only those proteins for which we have an associated protein family and compounds for which we have a structural fingerprint. From this subset, we then limit the size of the matrix by only considering the proteins and compounds with the most data to run preliminary tests on.

To assess the accuracy of our predictions through matrix completion, we took a random 10% of the known entries out of the matrix, and ran the matrix completion on the remaining 90%. We could then test the predicted values of the extracted 10% with the true known values. This was repeated on ten different random 10% selections to reduce any bias from the 10% sample selection.

The profiles with the highest area under the curve give the best ROC profiles. This corresponds to the predictions leaning more towards true positives than false positives.

We present our results as Receiver Operating Characteristics (ROC) graphs. These curves show the performance of the completion as a predictor for the bioactivity. To compute the ROC values, we prescribe a threshold value $T$, if the interaction value is above $T$ it is said to be 'positive', and if it is below $T$ it is said to be 'negative'. We can then compare whether our predicted values are true positive (TPs); predictions and true values are positive, or false positives (FPs); prediction positive but true value is negative. The ROC curve represents the TP and FP rates as we vary the value of $T$.

## Completion methods on the unclustered matrix

To first compare the five matrix completion methods, we ran them on the full unclustered matrix. We plot the resultant ROC curves in Figure 3. We see that the worst predictions were found when using the three means, with the best methods being SASD and MT-SASD. This supports the hypothesis that the activity matrix has low-rank structure.

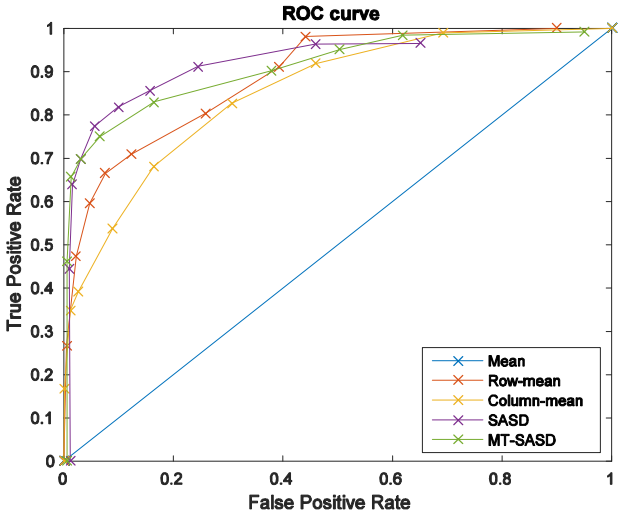


**Figure 3: ROC curve showing the performance of the different completion methods applied to unclustered matrix.**

## Unclustered vs. clustered

To assess the effectiveness of the clustering, we ran all the completion methods separately on each cluster, and chose the completion method that minimised the error for each of the different clustering methods. In Figure 4, we present the ROC profiles for the best method for each clustering against the best method for the unclustered total matrix.

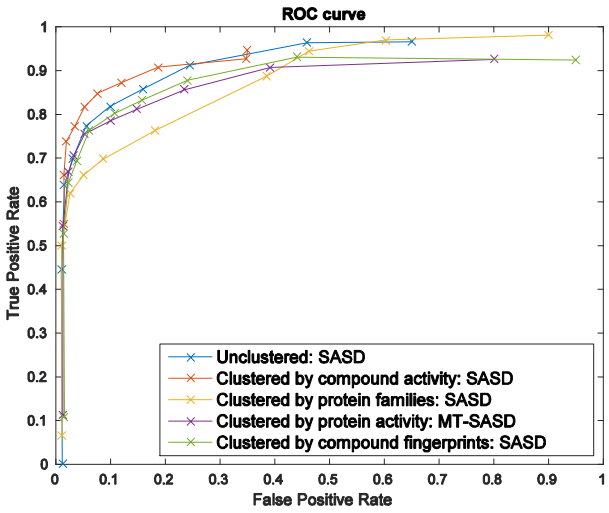


**Figure 4: ROC curve showing the performance of the optimal completion method for each clustering method.**

We see that, for all the clustering methods, low rank matrix completion was optimal, which shows that the activity matrix has low-rank structure, even when treated as a collection of submatrices. Restricting the matrix completion, and hence predictions, to the clusters does not seem to give significant improvements when compared with predicting on the whole matrix, as the ROC profiles for clustering by protein families, protein bioactivity, and compound fingerprints are all worse than the ROC profile for the unclustered matrix. There is, however, evidence that suggests it might more be beneficial to cluster compounds rather than proteins, as both compound clustering methods gave better ROC curves than the protein clusterings. The results also support clustering the bioactivity interactions over the structural features, since for both the compound and proteins the ROC profiles are better for bioactivity based clusters than the fingerprint and family clusters.

# 4. Discussion, Conclusions and Recommendations

We have studied the combined use of clustering and matrix completion in predicting the bioactive interactions between compounds and proteins. For unclustered and clustered matrices, low-rank matrix completion generated better predictions than the simple approach using mean values. This suggests that the associated activity matrix inherently has a low-rank structure. This corresponds to a compound-protein interaction being dependent on how interactive the compound and protein are in general. With the low-rank structure, we assign each compound and each protein an interactive score, and the likelihood score of the compound-protein interaction is then the multiplication of their individual scores.

Furthermore, we found that clustering did not significantly improve the predictions. This could be because the clusters resulting from the structural features are affected by the activity cliffs, and the clusters resulting from the bioactivity data are affected by the sparsity of the data. Therefore, we miss any activity cliffs that we might otherwise find if we had more data. However, in spite of this, there was some evidence that suggested it might be more useful to cluster compounds than it is to cluster proteins.

There is a balance between completing and clustering; clustering is affected by missing data, but we want to find the cluster structure before we complete. We have thus far treated the two operations separately, but we believe further investigation into this problem might benefit from simultaneous clustering and completing with low-rank structure. To our knowledge, there is no well-established algorithm that addresses this mathematical problem. Simultaneous clustering and completion would prevent any error occurred that in the clustering being carried through to the completion and subsequent predictions.

# 5. Potential Impact

e-Therapeutics want to be able to predict compound-protein interactions accurately. This work has shown that using the bioactivity data and structural features alone to cluster and predict on subsets of the data can produce meaningful results. However, neither approach has totally outshone the other, thus we suggest a heterogeneous data approach might benefit this problem. That is, an approach which can use both the incomplete bioactivity data and the complete structural data would be a useful predictive tool.

Jonny Wray, Head of Discovery Informatics, e-Therapeutics, "*Victoria explored the use of spectral clustering combined with techniques from matrix completion applied to the prediction of drug efficacy data. Experimental drug efficacy data is very sparse with only a small fraction of all potential compound-protein activities being measured. This data is critical to our drug discovery process, and so improvements in the ability to predict unknown efficacy data would have a major effect on our business. The results from this mini-project demonstrated that clustering data on both the protein and the compound axis have the ability to improve activity predictions. While this was previously appreciated for the compound data, it was a novel result for the protein data. Research projects are planned in order to further explore the use of protein clustering in activity prediction. In addition, the mini-project illustrated that matrix completion is a useful way to approach prediction and highlighted a potential future mathematical direction of combining clustering and completion.*"

# References


1. Tanner and Wei, (2016). *Low rank matrix completion by alternating steepest descent methods*. Applied and Computational Harmonic Analysis. **40(2):**417-429.

2. Antes, (2010). *DynaDock: A new molecular dynamics-based algorithm for protein–peptide docking including receptor flexibility*. Proteins: Structure, Function, and Bioinformatics 78.5: 1084-1104.