

**EPSRC**

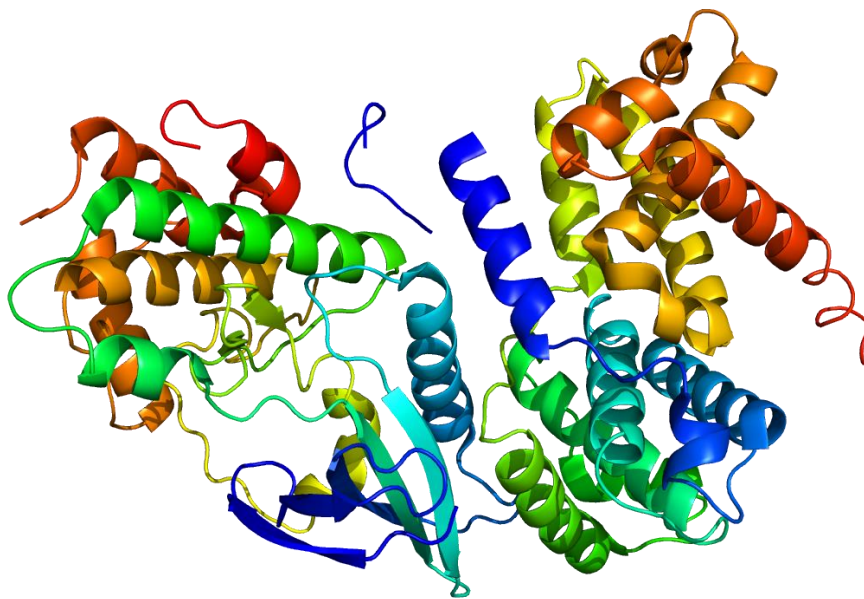
Engineering and Physical Sciences  
Research Council



**InFoMM**

Industrially Focused  
Mathematical Modelling

# EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling



## Bioactivity prediction from chemical and protein structure

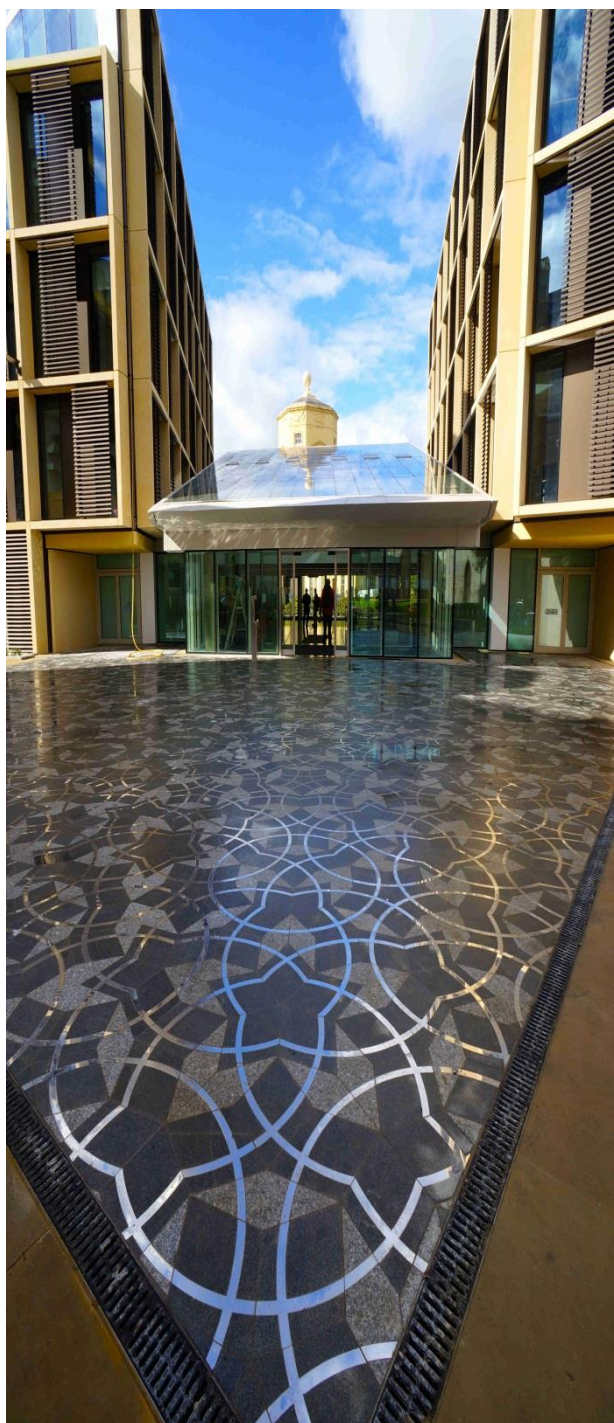
Melanie Beckerleg



UNIVERSITY OF  
**OXFORD**



e-therapeutics



## Contents

1. Introduction .....	2
Background.....	2
2. The bioactivity classification problem .....	2
Methodology .....	3
Aims .....	4
3. Results.....	4
Including more proteins .....	4
Using different classifiers .....	5
Constructing specialised models .....	6
4. Discussion, Conclusions & Recommendations	7
5. Potential Impact .....	7

# 1. Introduction

## Background

A significant aspect of drug discovery research is identifying whether known compounds are likely to have an effect on the proteins associated with a given disease

A key component of drug discovery is identifying compounds that will disrupt the network of proteins that relate to a disease. E-therapeutics use an approach known as network pharmacology to identify target proteins, from which they can identify compounds of interest. Laboratory-based methods such as High-Throughput Screening (HTS) involve testing, essentially at random, large numbers of compounds against large numbers of proteins. This is costly, time intensive, and typically quite ineffective. E-therapeutics use data inference methods which are cheaper and can have higher success rates. The methods often use chemical information about potential drug compounds alongside experimental data, (obtained from more biologically relevant experiments than HTS) relating to known interactions of a subset of compounds and proteins, to predict whether a compound will act on a protein for which no experimental evidence exists. A previous study found evidence that using information about proteins, particularly relating to the amino acid sequences, improved bioactivity predictions. Our aim is to understand how far this result might be generalised and whether we can further improve performance.

## 2. The bioactivity classification problem

The classification task we are interested in is constructing a function (or model) to determine whether a particular compound will impact a particular protein. The model takes the compound-protein pair as its input and returns either 1 (impact) or -1 (no impact). *Matrix completion* methods for constructing our model use only the underlying structure of the data, and make predictions by “filling in” missing values based on assumptions about the relationship between rows and columns of the matrix, thereby providing a classification for those compound-protein pairs for which there is no experimental data. *Machine learning* methods use extra information about particular *features* associated with compound-protein pairs. For example, if you were to build a classification model to predict whether students were in a particular school year, then age, height, academic performance etc. might be useful features to associate with each student. In Figure 1 we present a schematic of both approaches.

Classification models can be used to predict bioactivity for compound-protein pairs where no data is available

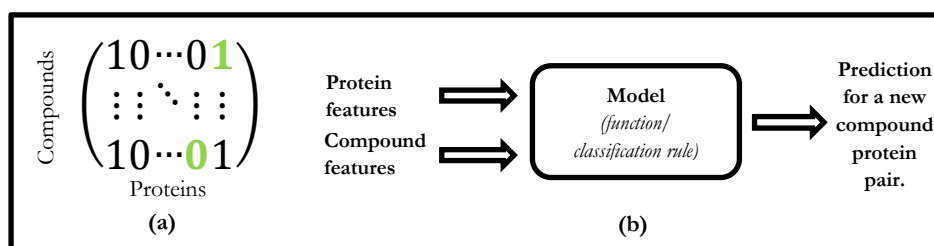


Figure 1: Compound protein interaction matrices record the bioactivity (1=activity, 0=no activity) between compound-protein pairs. Matrix completion methods (a) use relations between rows and columns to predict bioactivity where there is no experimental data. Machine learning methods (b) use features of compounds and proteins (eg. atomic structure) alongside bioactivity data to create a function (model) to make predictions.

Although more computationally expensive, machine learning methods can be powerful tools for identifying compounds of interest, and offer the potential for generalisation to compounds and proteins for which no experimental data exists. However, it is important to choose features that are relevant to the classification task; returning to our student school year example, details about eye colour are unlikely to help improve predictions. We describe compounds according to their atomic structure (or *footprint*), which is derived from information about the neighbours each atom within the compound has. For proteins, we consider two different sets of features; *functional families (FunFam)*, which are widely used groupings derived from protein behaviour (for example the presence of binding sites for particular enzymes) and *amino acid sequence patterns* within each protein.

## Methodology

We use bioactivity data from a subset of the ChEMBL database. For each algorithm we use, we construct a model using 80% of the compound-protein pairs for which experimental evidence exists, and test its predictions using the resulting 20%. We repeat the methods over multiple random trials in which data is chosen differently each time. Data points that are correctly identified as having either positive interaction or no interaction are called true positives and true negatives, respectively. Similarly, data points incorrectly predicted as having positive interaction or no interaction are known as false positives and false negatives, respectively. *Precision* is a measure of how likely a positive prediction is to mean that the compound actually does impact a protein. *Recall* is the proportion of positive interactions correctly identified by the model. Where the models provide probabilistic estimates for the classification score, we generate a curve to visualise the trade-off between different measures that comes from altering the threshold for determining activity. The performance of a model can be summarised by the area under this curve (AUC), where a larger area indicates better classification performance. We compare the predictions using the following metrics:

- **Area under the Receiver Operating Curve (ROC AUC):** the Receiver Operator Curve highlights the trade-off between the rate of false positives and the rate of false negatives of the model.
- **Area under the Precision-Recall Curve (PR AUC):** the Precision Recall curve highlights the trade-off between the precision and recall
- **Precision-Recall F-score** Also known as the harmonic mean, the F-score of two metrics is twice their product divided by their sum; for example, for precision ( $p$ ) and recall ( $r$ ) we calculate  $2(p \times r)/(p + r)$ . F-score is useful when AUC is difficult to compute, but provides information for only a single point on the curve.

In the area of drug discovery, it is important that models have a high recall score, since false negatives represent potentially missed opportunities. However, precision is also valuable, since the ultimate aim is to test only compounds that have a high chance of becoming drugs.

Our classification algorithms are based on matrix completion (or structural) methods or machine learning (or feature-based) techniques, and are listed in bold:

- **Matrix completion: Baseline** (uses per-protein average); **Low Rank** (a matrix completion algorithm that assumes the underlying data is has high dependence between rows and columns); **Naïve Bayes** (uses conditional probabilities)
- **Machine learning: Random forest** (Creates a group (or *forest*) of *decision trees*, by partitioning successive random subsets of the data according to the features which provide the largest split); **cost-penalised Random Forest** (when deciding partitions, assigns a greater misclassification cost to points in the minority); **Boosting** (combines *weak learners*, for example shallow decision trees, chosen successively to improve performance on misclassified points); **Logistic Regression** (builds a model that has the form of a logistic function, with parameters that maximise the 'likelihood' of the observed results); **K-nearest neighbours (KNN)** (makes predictions based on the classification of those  $k$  known data points whose feature information is 'nearest' that of the point being classified); **Support Vector Machines (SVM)** (chooses a 'hyperplane' which splits the feature information into two. In two-dimensions, this looks like drawing a line; when you have  $N$  pieces of information, the algorithm searches for an  $N$ -dimensional hyperplane instead.)

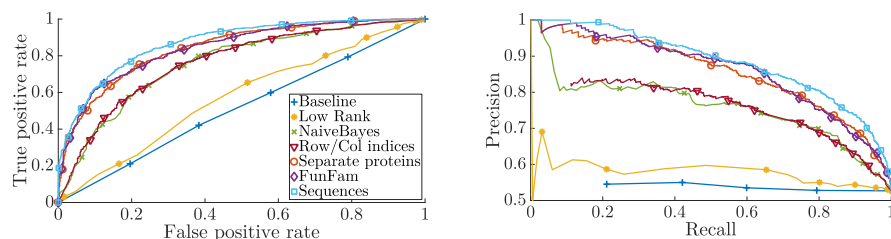
For machine learning algorithms, we assess the usefulness of compound information by training a separate model on each protein, using compound footprints. Models incorporating protein information are trained using all proteins at once, and using both compound information and either functional family (FunFam) information or information about patterns within amino acid sequences.

In a previous collaboration, it was found that using random forest models trained using feature vectors rather than structural methods improved classification performance, and



A previous collaboration found that, for a small subset of proteins, feature-based methods provide a significant improvement in predictive power

that information about patterns in amino acid sequences provided the most informative feature vector. In Figure 1, we show the ROC and PR curves for the structural methods Baseline, Low Rank, Naïve Bayes, and Row/Col and for Random Forests trained using separate proteins, FunFam, and Sequences. We see that machine learning methods have the greatest AUC for both precision-recall and ROC. The proteins used were selected according to the following criteria: each protein has a *density* of at least 3% (ie. for at least 3% of the compounds considered, experimental data exists for the impact of the compound on the given protein) and each protein has a *balance* (ratio of positive to negative impact values) of at least 0.8.



**Figure 2: ROC and PR curves for models trained on five proteins, chosen to have a density of greater than 3% and a balance of at least 0.8.**

Our aims are to:

- Investigate the effect of including more proteins; this involves including proteins for which less data is available and the data that does exist is more imbalanced (has either more positive or more negative entries). This is important to test the validity of the conclusions for a wider range of proteins and compounds.
- Investigate how different classification algorithms perform using the same feature vectors. This is important as it will test how far sequences can be said to be more informative than functional family information.
- Investigate whether constructing models using only a subset of biologically related proteins might improve classification accuracy. This will give insight into how best to use knowledge of proteins to construct the most powerful model.

### 3. Results

We increase the number of proteins to be classified by varying the thresholds for balance and density. We compare performance using a range of different classifiers. We also perform the classification task on smaller sets of biologically related proteins and find no significant difference in performance.

#### Including more proteins

In Figure 2, we plot the ROC AUC for various classification methods, for four different density thresholds. We see that changing the density has little impact on the performance of machine-learning methods relative to structural methods. However, as the balance decreases, the baseline and other methods that do not use feature information are able to perform better by over predicting the most common outcome, particularly when assessed according to ROC AUC. We also see that the performance of the structural methods is improved as the balance is lowered to include proteins for which there is a larger imbalance of positive to negative values. In particular, with the large dataset, Naïve Bayes performs well.

Including proteins with a higher imbalance of positive to negative impact scores, the advantage of using feature vectors decreases

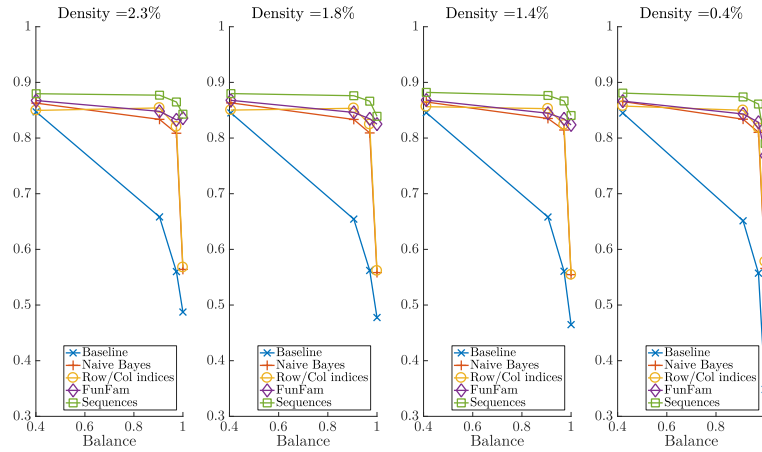


Figure 3: Area under the ROC curve increases for structural methods (Naïve Bayes, Baseline, Row/Col) when proteins with higher imbalance are included.

### Using different classifiers

In Figure 4, we show how different classifiers work on the five proteins used in the previous collaboration. We see that Random Forest, Oversampling and Boosting outperform structural methods, and that we see the same pattern in the performance for different feature vectors. The reduced performance of SVM, KNN and Logistic Regression compared to Random Forest and Boosting methods could be a result of the assumptions made to determine how distance between features is measured.

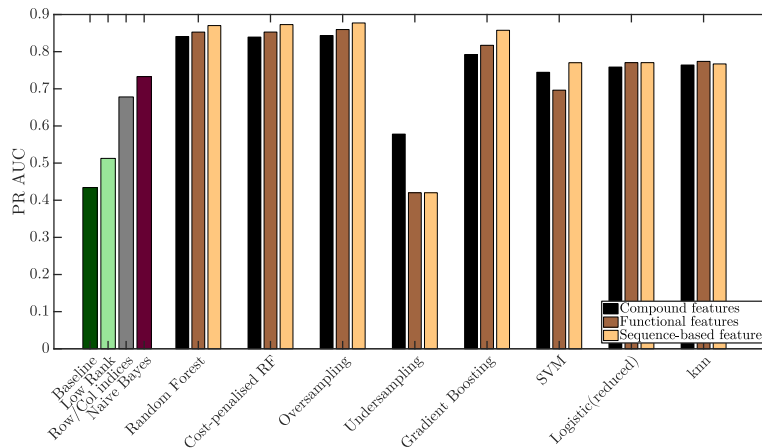


Figure 4: Graph showing PR AUC for different classifiers, trained on data for the five protein subset.

In Figure 5a, we show how the classifiers perform for the full set of proteins. We find that methods using feature information struggle to beat Naïve Bayes, with only Random Forest and Boosting methods outperforming the baseline. We also see that sequence-based features give the most information in Random Forests, but this is not true for other methods. Due to the size of the dataset, the memory required for oversampling prevented the algorithm running to completion; this could be addressed by increasing RAM or by randomly sampling a subset of data over multiple trials. Undersampling is a demonstrably poor method for addressing imbalance; we suggest this is because of the loss of information required to balance samples.

To address the fact that balance on an individual protein level can be significantly lower than for the whole dataset, we calculate the average performance per protein. In Figure 5b), we see the average precision-recall F-score evaluated per protein for various classifiers. Naïve Bayes performs well, and is only narrowly outperformed by cost-penalised random

forest trained on protein features. We hypothesise that this performance is at least partly due to the method exploiting imbalance in compounds, as well as proteins.

Different classifiers exhibit similar behaviour for a small subset of well-condition proteins. However, the results do not extend across the board to the wider dataset

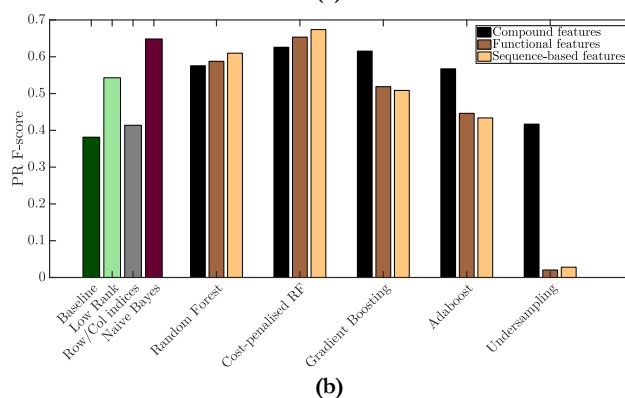
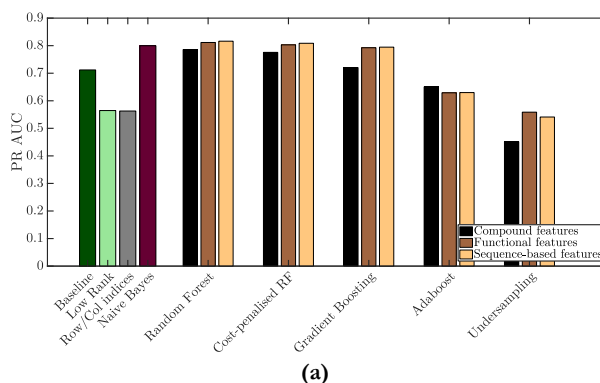


Figure 5: Graphs showing (a) PR AUC for the full dataset for different classifiers and (b) precision recall F-Score per protein.

## Constructing specialised models

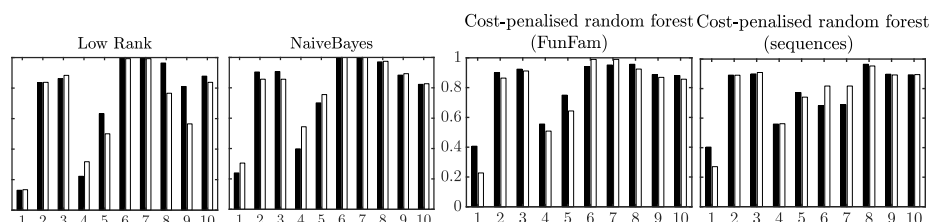



Figure 6: Graph showing PR AUC for models constructed from only proteins belonging to a particular superfamily (black) compared to how well a model constructed on the whole dataset classifies the same proteins (white).

Constructing models from fewer biologically related proteins generally does not notably decrease model performance compared to models with more proteins, and in some instances the strength of predictions increases

We construct a model using only proteins belonging to the same *superfamily* (groupings of proteins derived from similarities in physical structure), for ten different superfamilies, and compare the results to a model trained on all proteins. The superfamilies range in size from 100 to 500 proteins, and proteins can belong to multiple superfamilies. We consider cost-penalised random forests using different feature vectors (sequences and functional families), Naïve Bayes and Low Rank. We present the results in Figure 6, where we see that the results are similar for the two approaches, despite the fact that the superfamily-only models are trained on far fewer proteins. The variation in performance across different superfamilies is likely to be a result of the high imbalance of the superfamilies. The Low Rank method shows sign of performing better with certain subsets of proteins, implying that the assumption of high dependence between rows and columns is more realistic for proteins with known biological relationships. Machine learning methods do notably worse than structural methods for superfamilies 6 and 7. These superfamilies have a high positive balance and so this observation is to be expected since the Random Forest method penalises misclassification of negative values. It is perhaps more illuminating therefore to



consider the performance of the other methods which are more balanced, such as superfamily 8, where the specialised model does better for all methods except Naïve Bayes.

## 4. Discussion, Conclusions & Recommendations

Our aim was to investigate the validity of the results of a previous collaboration for a wider range of proteins. We have illustrated how results vary little with changing densities. However, although Random Forest methods using protein sequences perform better than structural methods, the gain becomes increasingly marginal for proteins with imbalanced data. It is clear, therefore, that when interpreting any of our results, the underlying structure (particularly imbalance) of the data being used should be taken into account.

Assessing performance for the full dataset is not straight forward; measuring the trade-off between precision and recall highlights aspects of classification that are of interest for the drug discovery problem, however these do not take account of the variety of imbalances seen at an individual protein level. Evaluating performance per protein may help with this however care must be taken in interpreting of 'pinpoint' metrics such as F-score.

Another way in which we assess the advantages of using protein information is through consideration of different classifiers. For well-conditioned data, both Random Forest and Boosting outperform methods that do not use feature information, and benefit from protein information, in particular relating to patterns in amino acid sequences. For the full dataset, Random Forest outperforms other classifiers. Whilst Boosting was at times competitive, those methods that use some concept of distance between corresponding features seem to struggle to make good predictions.

Finally, we considered how grouping proteins by superfamilies may improve performance. Our results indicate that discrepancies in performance between models trained on the full dataset and models trained on a subset of proteins are generally small. This is notable given the significant reduction in the computational effort and memory requirement to construct models on a smaller set of proteins. This may allow us to use more sophisticated machine learning methods that are infeasible on a large scale. In addition, there is some evidence that for certain superfamilies, the performance of some methods (i.e. Low Rank, and Random Forest with functional features) is improved when using the smaller dataset.

One aspect that we have not addresses is the computational effort associated with implementing each method. Machine learning methods take significantly longer to run (on the order of hours rather than minutes). Ways to decrease computational effort and memory burden include reducing the number of features associated with each compound-protein pair, and training a larger number of models using smaller subsets of data. Future work will consider how best to tailor features to improve performance.

## 5. Potential Impact

Using protein features to improve predictions may save resources, increase the success rate of laboratory tests and increase the possibility of identifying novel compounds. Creating a direct link between protein features and compound information, independent of structure, would enable bioactivity predictions for unstudied proteins and compounds.

Dr Jonny Wray, head of Discovery Informatics at e-therapeutics said, "*Compound-protein bioactivity data is critical in our approach to drug discovery, but available empirical data is very sparse. The use of computational predictions of such bioactivity enables us to improve our predictions of potential drugs. Expansion of our current predictive techniques to compounds and proteins with no data will massively expand the areas of chemical and biological space we can explore leading to novel, first-in-class drugs and biological mechanisms. Mel's work provides concrete foundations for the development of the next generation predictive techniques, illustrating conceptual feasibility of the approach and exploring the technical and computational details. I'm very excited to see where this leads.*"