University of Glasgow | School of Computing Science

Investigation of different features for predicting ionization efficiency of metabolites in mass spectrometry using Gaussian processes

Nikolay Zhekov

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

A dissertation presented in part fulfilment of the requirements of the Degree of Master of Science at The University of Glasgow

September 7, 2015

Abstract

The field of metabolomics studies the role of small molecules called metabolites. Mass spectrometry (MS) is widely used for identification and quantitation of metabolites. Extracting knowledge from MS data is a difficult problem and a lot of research is done in the area of identification of metabolites. However, precise quantitation of the identified metabolites is still challenging. This dissertation investigates the use of Gaussian process regression for probabilistic correction of MS signal intensities based on physico-chemical properties and molecule structure of the studied metabolites.

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic format. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Name: ____________________ Signature: ____________________

Acknowledgements

I would like to thank my supervisor Dr. Simon Rogers for his support and advice throughout all stages of this project.

And most importantly, I would like to thank my family and friends for their unconditional love and support.

Contents

1 Introduction 5

1.1 Outline 5

1.2 Background 5

1.2.1 Mass Spectrometry 5

1.2.2 Linear Regression 6

1.2.3 Gaussian Process Regression 6

1.2.4 Tanimoto and MinMax Tanimoto 7

1.3 Background Survey and Related work 8

1.3.1 Predicting Ionization Efficiency 8

1.3.2 Graph Kernels for Chemical Informatics 8

1.3.3 Learning on Mass Spectra 9

2 Statement of the problem 10

2.1 Research Problem 10

2.2 Research Statement 11

2.3 Hypotheses 11

3 Evaluation 12

3.1 Evaluation Data 12

3.1.1 Matabolomic datasets 12

3.1.2 Features 13

3.1.3 Targets 15

3.2 Evaluation software 15

3.3 Experiments 16

3.3.1 Reproducing the existing research 16

3.3.2 Gaussian process regression with RBF kernel 17

3.3.3 Gaussian process regression using molecular structure 22

3.3.4 Gaussian process regression with combined kernels 24

4 Conclusion **28**

4.1 Conclusion 28

4.2 Future Work 29

Chapter 1

Introduction

This chapter outlines the structure of this dissertation and provides background information on the research problem.

1.1 Outline

In this chapter mass spectrometry and different machine learning techniques are discussed to give background information on the research problem. Additionally, a short review of the existing research in the area is presented since it is used as ground truth on which this dissertation builds upon. In chapter 2 a detailed explanation of the problem is given. The evaluation datasets, software and the performed experiments are presented in chapter 3. Chapter 4 summarizes the results and discusses future improvements.

1.2 Background

1.2.1 Mass Spectrometry

Mass spectrometry (MS) is the most popular platform for metabolomic studies. There are different type of MS platforms such as Liquid chromatography mass spectrometry (LC-MS), Gas chromatography mass spectrometry (GC-MS) and Capillary electrophoresis mass spectrometry (CE-MS). These platforms differ by the specific way the analytes are separated but share common analytical phases - stationary phase when the separation happens and mobile phase when the analyte is carried through the mass spectrometer [Smith et al., 2014].

During the mobile phase analytes need to be ionized in order to be detected by the mass spectrometer. The most popular way of ionization for LC-MS is Electro-spray ionization (ESI) because of its “soft ionization” that does not break the chemical bonds of unstable molecules [Zhou et al., 2012]. This phase is critical for the identification and quantification of the metabolites.

1.2.2 Linear Regression

Supervised machine learning focuses on learning relationship between inputs (independent variables) and targets (outputs - dependent variables). In general, there are two classes of supervised machine learning problems - classification and regression. In classification problems the targets are discrete values, in regression problems the targets are continuous variables.

The ordinary least squares regression is a simple, yet very effective, regression model. In its simplest definition linear regression takes the form:

$$y_i = w_0 + w_1 x_{i1} + ... + w_n x_{in}$$

or in vector form, where $\mathbf{w} = [w_0, w_1 ... w_n]$ are the model parameters and $\mathbf{x}_i = [1, x_1 ... x_n]$ are the input variables of the i-th data point:

$$y_i = \mathbf{w}^T \mathbf{x}_i$$

The optimal parameters $\hat{\mathbf{w}}$ can be learnt by minimizing the loss function:

$$\begin{aligned} &\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathcal{L} = \arg\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 \\ &\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

where $\mathbf{X}$ consists of all vectors $\mathbf{x}$ and the vector $\mathbf{y}$ of all learning targets.

New values are then computed by

$$y_{new} = \mathbf{w}^T \mathbf{x}_{new}$$

1.2.3 Gaussian Process Regression

According to the definition of Rasmussen and Williams in *Gaussian Processes for Machine Learning – "A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution"* [Rasmussen and Williams, 2005].

A Gaussian process f is completely specified by its mean (m) and covariance function (k):

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

To make predictions for new input X_* we define the joint distribution of the observed target values $\mathbf{y}$ and the function values at the test locations $\mathbf{f}_*$ (assuming zero mean):

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} = \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right)$$

where X is the input matrix and K denotes a matrix of covariances evaluated for all pairs of the input vectors. The noise in the observations is model by a Gaussian with variance σ_n^2.

The predictive equation for the Gaussian process regression is:

$$\begin{aligned}
&\mathbf{f}_*|X,\mathbf{y},X_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*)), \quad \text{where} \\
&\bar{\mathbf{f}}_* = K(X_*,X)[K(X,X)+\sigma_n^2 I]^{-1}\mathbf{y}, \quad \text{and} \\
&\text{cov}(\mathbf{f}_*) = K(X_*,X_*)[K(X,X)+\sigma_n^2 I]^{-1}K(X,X_*)
\end{aligned}$$

where $\bar{\mathbf{f}}_*$ denotes the mean and $\text{cov}(\mathbf{f}_*)$ is the uncertainty of the test predictions.

The choice of the kernel function for the Gaussian process regression is important as it models the similarity between observations and directly affects the accuracy. A popular choice is the squared exponential kernel (a.k.a Radial Basis Function kernel or Gaussian kernel):

$$k_{SE}(x,x') = \sigma_f^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right) + \sigma_n^2 \delta_{x,x'}$$

where length scale ℓ and the signal variance σ_f are adjustable parameters of the kernel. The σ_n^2 is the noise variance added to diagonal of the covariance matrix using Kronecker delta $\delta_{p,q}$ (1 iff $p = q$ and 0 otherwise).

An interesting property of the kernel functions is that they can be multiplied or added together, which makes Gaussian process regression very flexible by allowing different sources of information to be combined.

1.2.4 Tanimoto and MinMax Tanimoto

Tanimoto and MinMax Tanimoto are graph kernels first introduced in [Ralaivola et al., 2005]. They were specifically developed for problems in chemical informatics such as classifying compounds based on physical, chemical, or biological properties. These kernels use the structure of the molecules in form of a labelled undirected graph of chemical bonds and do calculations based on molecular fragments (*walks* in a graph - sequence of atoms and their bonds). [Klambauer et al., 2015]

Tanimoto kernel The kernel is defined as follows:

$$k(x,x') = \sum_{p \in \mathcal{P}} N(p,x) \cdot N(p,x')$$

where x and x' are molecular graphs, $\mathcal{P}$ contains all walks p up to a predefined length d and the function $N(p,x)$ indicates whether a fragment exists in x:

$$\begin{aligned}
N(p,x) = 1 \quad \{p \in x\} \\
N(p,x) = 0 \quad \{p \notin x\}
\end{aligned}$$

MinMax Tanimoto kernel This kernel is variation of the Tanimoto kernel. The difference is that the function $N(p, x)$ counts the number of of occurrences of the fragment p:

$$
\begin{aligned}
k_{min}(x, x') &= \sum_{p \in \mathcal{P}} \min(N(p, x), N(p, x')) \\
k_{max}(x, x') &= \sum_{p \in \mathcal{P}} \max(N(p, x), N(p, x')) \\
k_{minmax}(x, x') &= \frac{k_{min}(x, x')}{k_{max}(x, x')} \\
N(p, x) &= \#\{p \in x\}
\end{aligned}
$$

1.3 Background Survey and Related work

1.3.1 Predicting Ionization Efficiency

Prediction of ionization efficiency of metabolites is subject that does not have extensive coverage in the literature. The work of *Chalcraft et al.* is considered the first that demonstrate virtual quantification of metabolites. In their paper *"Virtual Quantification of Metabolites by Capillary Electrophoresis-Electrospray Ionization-Mass Spectrometry: Predicting Ionization Efficiency Without Chemical Standards"*[Chalcraft et al., 2009] the authors present a multivariate linear regression model using three physico-chemical properties of low-abundance metabolites. The model is developed using dataset of 58 chemical standards divided in training set of 47, test set of 10 and one metabolite used as internal standard. The initial model uses four properties – molecular volume (MV), octanol-water distribution coefficient (log*D*), absolute mobility (μ_o), and effective-charge (z_{eff}). However, the last property is excluded from the final model because of its low contribution.

1.3.2 Graph Kernels for Chemical Informatics

In order to make predictions based on structured data in Gaussian process regression a special kernel has to be used. In *"Graph Kernels for Chemical Informatics"*[Ralaivola et al., 2005] three graph kernels for chemical compounds are introduced - Tanimoto, MinMax and Hybrid (a.k.a MinMax Tanimoto). These kernel functions are based on the idea of molecular fingerprints and counting of labelled paths in graphs. The research demonstrates that the performance of the new kernels is superior or at least comparable to the state-of-the-art based on three different datasets - Mutag dataset, National Cancer Institute dataset and Predictive Toxicology Challenge dataset.

Additionally, Tanimoto and MinMax Tanimoto are used in *"Graph methods for predicting the function of chemical compounds"*[Zhu and Yan, 2014] in which the authors report that the MinMax Tanimoto kernel outperform any other graph kernel in classification experiments over five different datasets with over 3500 compounds each.

1.3.3 Learning on Mass Spectra

A novel approach for metabolite identification is proposed in *"Metabolite identification and molecular fingerprint prediction through machine learning"* [Heinonen et al., 2012]. Mass spectra are low level output of the mass spectrometer. In this work mass spectra are used in predictions of molecular fingerprints that encode various characteristics of the compounds. New kernel functions that work on mass spectra are introduced and used in classification task using Support Vector Machines (SVM). This research demonstrate that the relationship between mass spectral signal and molecular properties can be learnt with high accuracy.

Chapter 2

Statement of the problem

This chapter presents the research problem discussed in this dissertation.

2.1 Research Problem

Absolute quantitation of MS signal intensity is considered as one of the major problems of interest in the field of proteomics, lipidomics and metabolomics [Bantscheff et al., 2007, Smith et al., 2014]. The task of correcting the signal intensity is very challenging because of the multiple factors that can affect the final result. During the mobile phase different molecules can be ionized differently. Even more, it is possible the ionization of a compound to be suppressed or enhanced by the presence of another compound - described as matrix effects in [Mei et al., 2003]. However over the past decade protocols have been developed that ensure these phenomena can be avoided or at least minimized [Jessome and Volmer, 2006].

Improvements in the estimation of signal intensity will allow us to compare multiple metabolites with each other in single experiment or across experiments. Currently, we can only compare the intensities of single metaboloite with itself in different experiments. Additionally, better estimation can be beneficial for the subsequent steps in the MS/metabolomics pipeline and further improve *in silico* simulations [Smith et al., 2014]. In [Chalcraft et al., 2009] the authors demonstrate a linear regression model that is able to predict ionization efficiency on a selected set of metabolomic standards. Although, linear regression is a simple and effective model, the latest advancements in the field of machine learning provide us with tools that might be able to better model the complex nature of the MS metabolomics analysis.

Unlike linear regression where we have point predictions, Gaussian process (GP) regression outputs a full predictive distribution which models the uncertainty and fits better into an analytical pipeline. Linear regression models are simple and scalable, but do not perform well with high-dimensional input. On the other hand, GPs do not scale well because of the $\mathcal{O}(n^3)$ time complexity on the number of data points. However, in metabolomics analysis, because of the cost of the experiments, we usually work on small datasets with multiple features. For this reason, one of the main advantages of GPs is the high flexibility with multi-dimensional data because of the use of kernel (covariance) functions. While the standard linear regression model works only on numerical data, though different kernel functions GPs can be applied not only on numerical data, but also on structured data such as strings, trees or graphs.

2.2 Research Statement

The aim of this project is to research possible ways to correct MS signal intensity of metabolites using intrinsic chemical properties and/or molecule structure. Existing research in virtual quantification and prediction of ionization efficiency [Chalcraft et al., 2009] shows promising results. However, further research of the application of different machine learning techniques have the potential of improving the existing work. One of the main goals of this research is to test the feasibility of Gaussian process regression using different kernel functions and various information about the tested analytes. Different datasets will be used to benchmark the new models and investigate the performance of the selected kernels and features.

2.3 Hypotheses

1. Non-linear regression model, such as Gaussian process regression, can lead to better predictions compared to a linear regression.
2. Other physico-chemical features exists that have similar or better descriptive power compared to the features used in [Chalcraft et al., 2009].
3. Combining structural information about the molecules with physico-chemical properties can lead to better predictions.

Chapter 3

Evaluation

This chapter presents the data used in this study and evaluation of the performed experiments.

3.1 Evaluation Data

In order to explore the effects of different compound features two different datasets were used. The data had to be pre-processed and new features had to be extracted. This section describes the data used in the research, the process of getting new features and the prediction targets.

3.1.1 Matabolomic datasets

Existing research

Datasets provided in the supporting materials of [Chalcraft et al., 2009] were manually extracted and used as ground truth. These datasets have three intrinsic physico-chemical properties - molecular volume (MV), octanol-water distribution coefficient (log$D$), absolute mobility (μ_o) and measured and predicted Relative Response Factor (RRF). These data was initially used to verify the feasibility of the described linear regression model and as a benchmark for the use of Gaussian process regression. Later, other physico-chemical properties for the given compounds were extracted from external public databases and used in combination with Gaussian process model.

Glasgow Polyomics Data

Two datasets, referred to as Standards 1 (*Std 1*) and Standards 2 (*Std 2*), provided by *Glasgow Polyomics* containing data for 104 and 96 metabolites respectively. Combined the two datasets had 4400 data points with information about different ion adducts and signal intensities for six solute dilutions. Unlike the datasets in the existing research, these datasets contain various metabolites that were not pre-selected based on chemical behaviour apart from being chemical standards. For the purposes of the experiments carried out in this project only metabolites from one of the dilutions were used - 73 from *Std 1* and 52 from *Std 2*.

The two datasets had the following information:

- **Name** - Common name of the molecule
- **Formula** - Chemical formula of the molecule
- **Signal intensity** - MS Signal intensity

3.1.2 Features

The features used in the existing research were difficult to obtain because they were measured using laboratory experiments and commercial software. For this reason different public databases were evaluated based on the provided physico-chemical properties, API and ease of access.

Public databases

PubChem [Bolton et al., 2008] is a database containing information for more than 60 million chemical compounds. It is maintained by the National Center for Biotechnology Information (NCBI) and is publicly available. It provides web interface and REST API. Each compound has the following properties - *MolecularWeight*, *XLogP*, *ExactMass*, *TPSA*, *Complexity*, *Charge*, *HBondDonorCount*, *HBondAcceptorCount*, *RotatableBondCount*.

HMDB [Wishart et al., 2013] The Human Metabolome Database is freely available database containing data for 41,993 metabolites. It integrates chemical data from multiple databases and can be downloaded as single SDF file. Each compound has the following properties - *Water Solubility*, *logP*, *logS*, *pKa (Strongest Acidic)*, *pKa (Strongest Basic)*, *Physiological Charge*, *Hydrogen Acceptor Count*, *Hydrogen Donor Count*, *Polar Surface Area*, *Rotatable Bond Count*, *Refractivity*, *Polarizability*, *Number of Rings*.

MassBank [Horai et al., 2010] is a database containing more than 40000 mass spectra shared by multiple research groups. The data is following a structured format (MassBank Record Format) [MassBank Project, 2013] and each experiment is stored in separate plain text file. Some of the mandatory fields include - *Chemical name*, *Chemical formula*, *Type of instrument*, *MS Type*, *Ion mode*, *Number of peaks*, *Peak (m/z, intensity)*.

Matching compounds and storing structure information - evaluation of SMILES and InChI

The simplified molecular-input line-entry system (SMILES) is popular format for encoding of molecular structure using ASCII characters. It is widely supported and most public chemistry databases can be queried using SMILES. However, there are multiple ways to encode a molecule as

SMILES string which makes matching of molecules difficult. Although, there are algorithms that can create “canonical” SMILES strings, the uniqueness is not guaranteed across different products and databases.

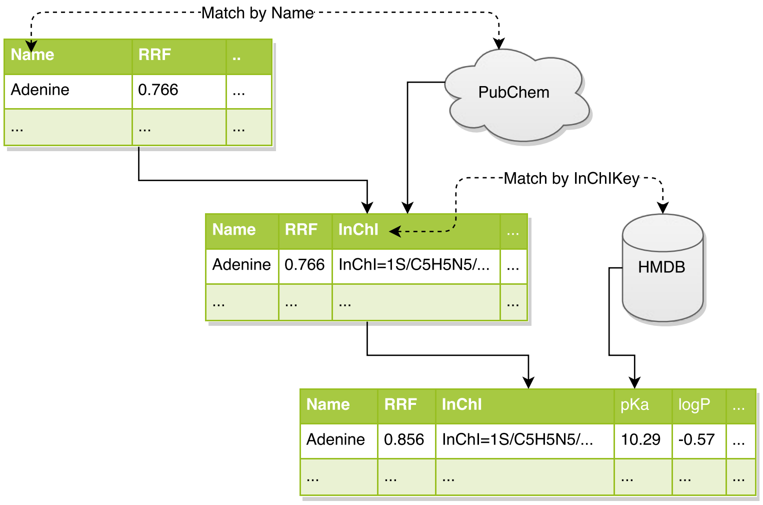
In 2005 the International Union of Pure and Applied Chemistry (IUPAC) published a new way for encoding chemical structures called InChI - IUPAC International Chemical Identifier [IUPAC, 2005]. The Standard InChI is considered as stable identifier for the purposes of interoperability between chemical databases, web searching and information exchange [IUPAC, 2009]. The standard also has a definition of InChIKey, which is a hash of the InChI identifier. Most public databases support querying by InChIKey. In this project InChI and InChIKeys are used for matching of molecules and reading the chemical structure of the analytes for use in the graph algorithms.

Data preprocessing

The data from both the existing research and *Glasgow Polyomics* did not have any identifiers beside the common name of the chemical compound. Figure 3.1 shows the steps performed to acquire the necessary data.

Identification of chemical compounds based on name is not an easy task. Most compounds have multiple synonyms, for example *Vitamin C* has 140 synonyms listed in HMDB. In the first step of the preprocessing work-flow PubChem was used because of its capability to identify compounds using common name. Once identified, all properties of the compound and the InChI were merged to the existing datasets. After that, the InChIKeys of the compounds were matched to HMDB entries and additional properties were merged.

Figure 3.1: Data preprocessing work-flow.



Name	RRF	..
Adenine	0.766	...
...	...	...

Name	RRF	InChI	...
Adenine	0.766	InChI=1S/C5H5N5/...	...
...	...	...	...

Name	RRF	InChI	pKa	logP	...
Adenine	0.856	InChI=1S/C5H5N5/...	10.29	-0.57	...
...	...	...	...	...	...

Additionally, a custom parser was developed in order to extract data from MassBank’s plain text files. The data was stored in MongoDB database and indexed by InChIKey for fast querying and matching of compounds.

3.1.3 Targets

The target (dependent variable) in the evaluated regression models is the Relative Response Factor (RRF) – a descriptor of the ionization efficiency of the metabolite. The RRF is calculated by dividing the signal intensity of a given compound by the signal intensity of another predefined compound (internal standard).

In the dataset from the existing research the measured RRF is given. For the *Glasgow Polyomics* dataset we know that all compounds have the same concentration (intensity). However, we do not know the exact value. It is assumed that the compound with the highest measured intensity is most efficient, i.e. all molecules are ionized, and all other compounds are normalized relative to this compound.

3.2 Evaluation software

This section summarizes the software products used in the research process.

IPython [Pérez and Granger, 2007] is a popular interactive environment for easy editing, execution and evaluation of python code. Additionally, IPython notebooks support interactive data visualisation using libraries such as matplotlib [Hunter, 2007] and seaborn [Waskom et al., 2015] and allow the execution output to be recorded and easily reviewed without re-executing the script. All experiments in this research project were performed in IPython.

scikit-learn [Pedregosa et al., 2011] is the de facto standard library for machine learning using Python. It is an actively developed open-source project supported by a big community of scientists and machine learning enthusiasts. The library contains various tools for classification, regression, clustering, model selection, dimensionality reduction and preprocessing.

GPy [The GPy authors, 2014] is a specialized Python library for Gaussian Processes developed by the machine learning group based at the Sheffield Institute for Translational Neuroscience. It's fast implementation of Gaussian Process Regression is used in multiple experiments throughout this project.

pandas [McKinney, 2011] is a high-performance library for data preparation and manipulation. It greatly simplifies the data processing of external data sources such as comma-separated values (CSV) files, as well as the "cleaning" and conversion of data before using it in machine-learning algorithms.

NumPy & SciPy [Walt et al., 2011, Jones et al., 2001] are libraries for efficient numerical computation and array/matrix representation. These libraries are common dependencies for most machine learning and data manipulation packages implemented in Python.

RDKit [Landrum, 2014] is open-source cheminformatics toolkit implemented in C++ and Python. In this projects it is used for its ability to load molecule structures from SMILES and InChI representation, and save structured-data files (SDF) for further processing.

Rchemcpp [Klambauer et al., 2015] is package for R [R Development Core Team, 2008] that implements various cheminformatics algorithms. In this projects it is used for computation of Tanimoto and MinMax Tanimoto kernels which use the molecular structure of the tested metabolomes. The R functions were accessed from Python using the rpy2 library [Gautier, 2014].

Other Custom Gaussian process regression based on *Gaussian Processes for Machine Learning* [Rasmussen and Williams, 2005] and *GPy* [The GPy authors, 2014] was implemented in order to support kernels based on molecular structure and combinations of kernels.

3.3 Experiments

In machine learning and statistical analysis, regression models aim to estimate relationship between features (independent variables) and targets (dependent variables). The following section presents the use of different regression techniques and features for prediction of the ionization efficiency of metabolites.

3.3.1 Reproducing the existing research

The first step of the research was to test the feasibility of the existing work in order to use it as ground truth in further experiments. The model reported in [Chalcraft et al., 2009] is

$$y = [4.4 \times 10^{-4}](MV) + [2.7 \times 10^{-3}](logD) + 14\mu_o - 4.14 \times 10^{-2} \tag{3.1}$$

Using scikit-learn's ordinary least squares linear regression, a model was fit on the training data (*See 3.1.1*) from the supporting materials, resulting in a model

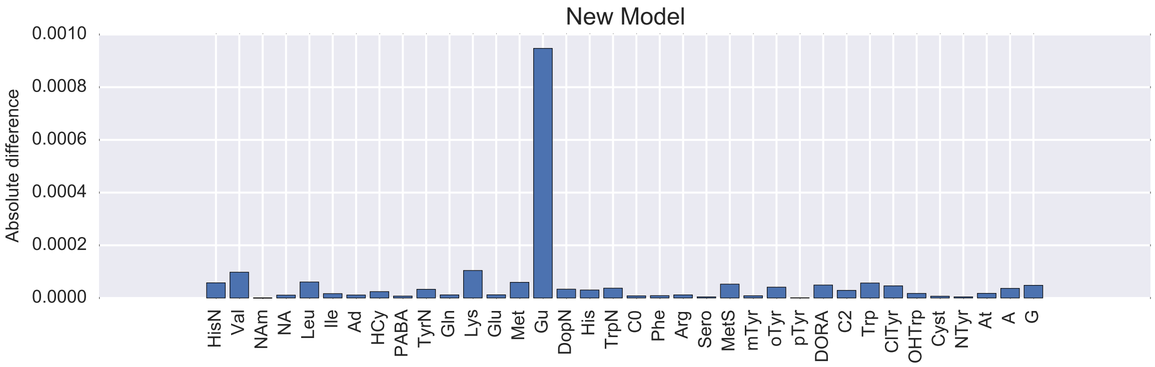
$$y = [4.4 \times 10^{-4}](MV) + [2.72 \times 10^{-3}](logD) + 14.1408\mu_o - 4.132 \times 10^{-2} \tag{3.2}$$

Figure 3.2 shows the absolute difference between the RRF reported in [Chalcraft et al., 2009] and the predictions made by our model.

Performance metrics were also used in order to compare the two models. R^2 (coefficient of determination) and Mean Squared Error (MSE) were calculated – the reported measured and the predicted RRF were used for the model from the paper and measured RRF and newly predicted RRF for our model. Table 3.1 presents the results.

Additionally, the predictions of the two models (3.1 and 3.2) were compared – Figure 3.3 presents the absolute difference of the model predictions and the reported RRF predictions. This shows that

Figure 3.2: Prediction deviation from reported predictions – Training dataset



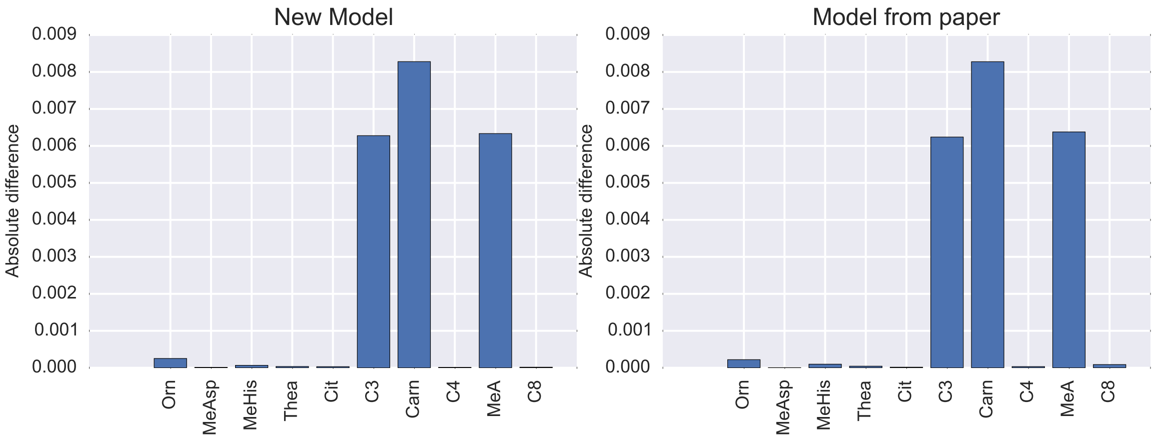
Metric	Score
R^2	0.8283
MSE	4.8118×10^{-5}

Metric	Score
R^2	0.8278
MSE	4.8249×10^{-5}

Table 3.1: Comparison of model metrics. Based on predictions reported in the paper (*left*) and predictions from our model (*right*)

even the reported model does not output exactly the same predictions. The difference might be contributed to number rounding in the reported data. Given the small magnitude of the errors, this initial experiment proved that the data and existing research can be used as ground truth for further experiments.

Figure 3.3: Prediction deviation from reported predictions – Test dataset



3.3.2 Gaussian process regression with RBF kernel

Fitting GP on the ground truth dataset

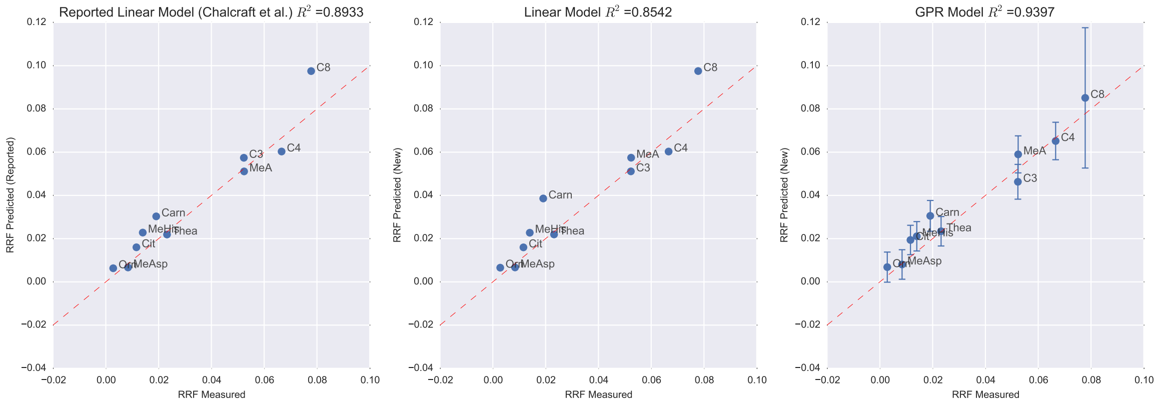
The next step of the research was to evaluate a non-linear regression model and compare it to the linear model reported in [Chalcraft et al., 2009]. The data from the same training and testing dataset was centred to zero mean. After that a Gaussian process regression model was trained and optimized (based on marginal likelihood) with GPy [The GPy authors, 2014], resulting in the following model:

$$\ell = 56.926 \qquad \sigma^2 = 1.938 \times 10^{-3} \qquad \sigma_n^2 = 4.316 \times 10^{-5} \tag{3.3}$$

where ℓ (*length-scale*) and σ^2 (*variance*) are hyper-parameters of the RBF kernel and σ_n^2 is the noise variance.

Figure 3.4 shows a comparison of the predictions reported in the existing research and the predictions made by the GP model (3.3). The error bars in the rightmost figure represent one standard deviation of uncertainty in the GP predictions.

Figure 3.4: Reported Linear (*left*), New Linear (*middle*) and Gaussian Process Regression (*right*)



Metric	Score
R^2	0.8933
MSE	6.947×10^{-5}

Metric	Score
R^2	0.8542
MSE	9.488×10^{-5}

Metric	Score
R^2	0.9397
MSE	3.926×10^{-5}

Table 3.2: Comparison of model metrics. Based on predictions reported in the paper (*left*), predictions from new linear model (*middle*) and predictions from the GP regression (*right*)

Table 3.2 compares the metrics of the models on the test dataset. The GP regression has a better R^2 score of 0.9397 compared to the reported linear model with 0.8933, mainly because of better predictions of the RRF of *C4* and *C8*. However, the major benefit of the GP is the modelling of the uncertainty of the predictions. This can be seen on the predictions for *C8*, which is the compound that deviates the most from the “ideal” line in the linear model.

As described in 3.1.1 the new linear model fit does not produce the exact same results as reported in [Chalcraft et al., 2009]. Comparing the linear and GP regression in equal conditions, on the same training and test data, show bigger difference in the score metrics.

All these results come to show that Gaussian process regression model outperforms the ordinary least squares linear model using the same features and training/test split.

Selecting HMDB physico-chemical properties

Having features that can be easily extracted from public database will improve the analytical pipelines and reduce time and cost by removing the need of expensive commercial chemical databases and software. One of the major goals of this research was to find other physico-chemical properties that can be used in combination with the Gaussian process regression. After the preprocessing step (*See 3.1.2*) all 14 properties extracted from HMDB were integrated into the datasets. Principal component analysis (PCA) was performed on the ground truth dataset in order to find the features with best descriptive power that could potentially be used for modelling of the ionization efficiency.

PCA is statistical technique that allows high-dimensional data to be reduced to lower-dimensional representation that keeps only the most significant information. It is commonly used for visualization, data exploration and for finding patters in data.

Figure 3.5 shows a PCA loading plot, describing the correlations between the properties – properties that appear closer together are more highly correlated. For the initial analysis the following properties were selected based on the PCA and their chemical significance in relation to the ionization efficiency [Oss et al., 2010]:

- *Acidic pKa* (CHEM_ACIDIC_PKA)
- *logP* (JCHEM_LOGP)
- *Polarizability* (JCHEM_POLARIZABILITY)
- *Polar Surface Area* (JCHEM_POLAR_SURFACE_AREA)

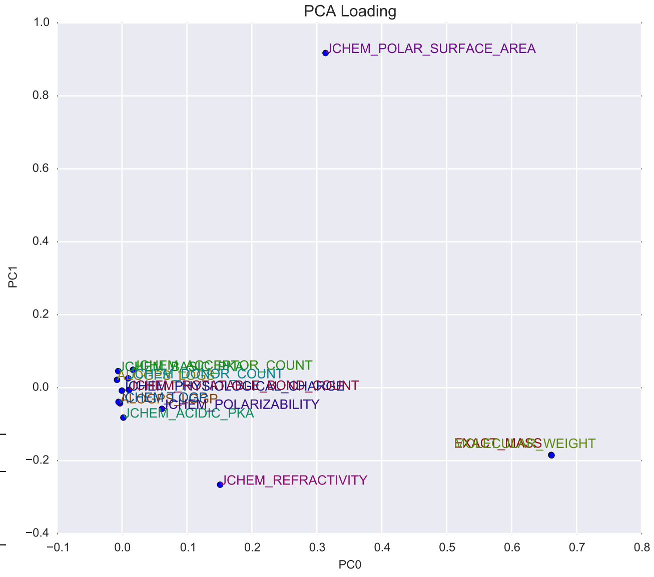
ALOGPS_LOGP
ALOGPS_LOGS
JCHEM_ACCEPTOR_COUNT
JCHEM_BASIC_PKA
JCHEM_ACIDIC_PKA
JCHEM_DONOR_COUNT
JCHEM_LOGP
JCHEM_PHYSIOLOGICAL_CHARGE
JCHEM_POLARIZABILITY
JCHEM_ROTATABLE_BOND_COUNT

JCHEM_POLAR_SURFACE_AREA

EXACT_MASS
MOLECULAR_WEIGHT

JCHEM_REFRACTIVITY

(a) Feature correlations



(b) Loading Plot

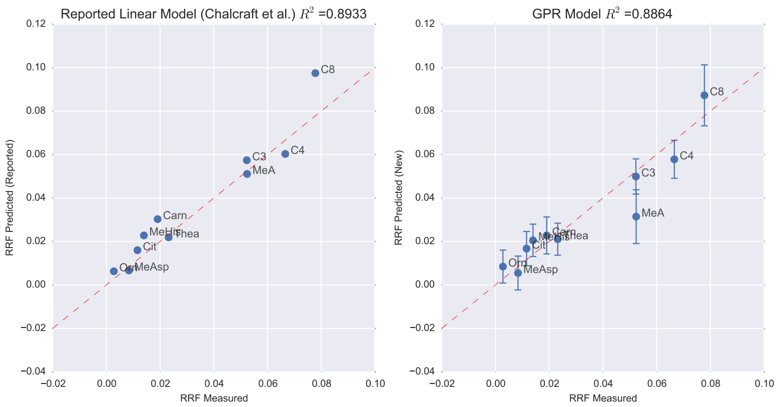
Figure 3.5: PCA

Fitting GP using different features on the ground truth dataset

To test whether the selected features can be used for predictions of the ionization efficiency a GP model was fit on the ground truth dataset and the results were compared to the predictions of the linear model.

Similar to the approach used in the previous GP experiment, the new physico-chemical properties were standardized to zero mean. The resulting model had the following hyper-parameters:

$$\ell = 29.580 \qquad \sigma^2 = 2.442 \times 10^{-3} \qquad \sigma_n^2 = 4.458 \times 10^{-5} \tag{3.4}$$



(a) Linear regression on the original features (*left*) and Gaussian Process Regression (*right*) using HMDB features

Metric	Score
R^2	0.8864
MSE	7.392×10^{-5}
Log-likelihood	34.89

(b) GP regression model metrics

Figure 3.6a shows the predictions of the linear model reported in [Chalcraft et al., 2009] using the original features and the predictions from the new model using the new features. The performance of the two models is very close, with R^2 of 0.8933 for the linear model compared to 0.8864 for the GP model.

These results show the feasibility of selected physico-chemical properties for ionization efficiency predictions on the ground truth dataset.

Fitting GP on the Glasgow Polyomics dataset using HMDB properties

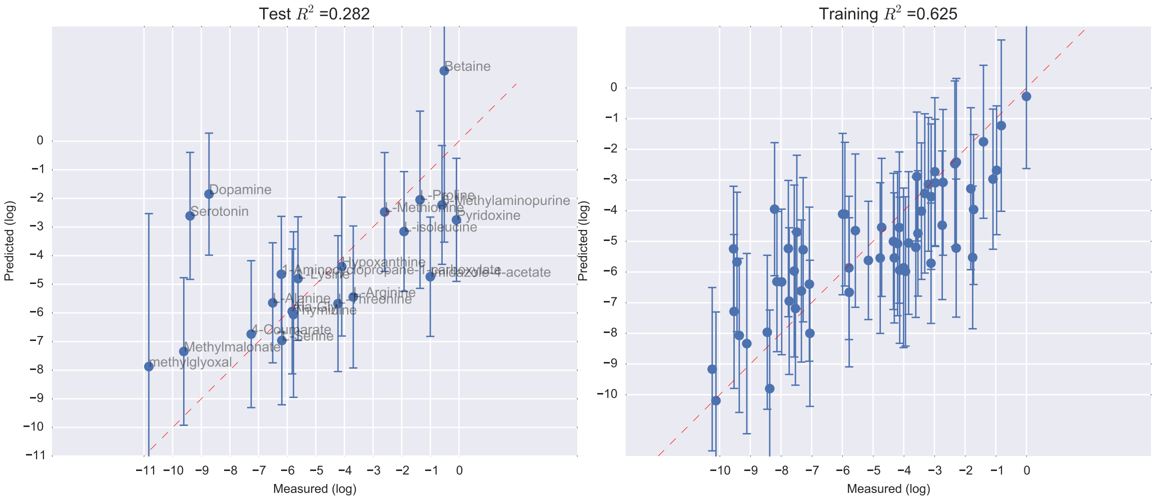
In order to check whether the selected physico-chemical can be used for ionization efficiency predictions on other metabolites, the approach had to be tested on other data.

A new model was fit on the *Glasgow Polyomics* dataset. The two datasets *Std 1* and *Std 2* (*See 3.1.1*) were combined, shuffled and split into training and test sets in 3:1 ratio. After the initial tests, molecules with mass above 280 were excluded because they were under-represented in the dataset and were affecting the performance. Additionally, logarithmic scale was used for the target RRF values.

5-Fold Cross Validation (CV) was performed on the training set in order to find the best hyper-parameters that maximize the predictive likelihood, resulting in the following model:

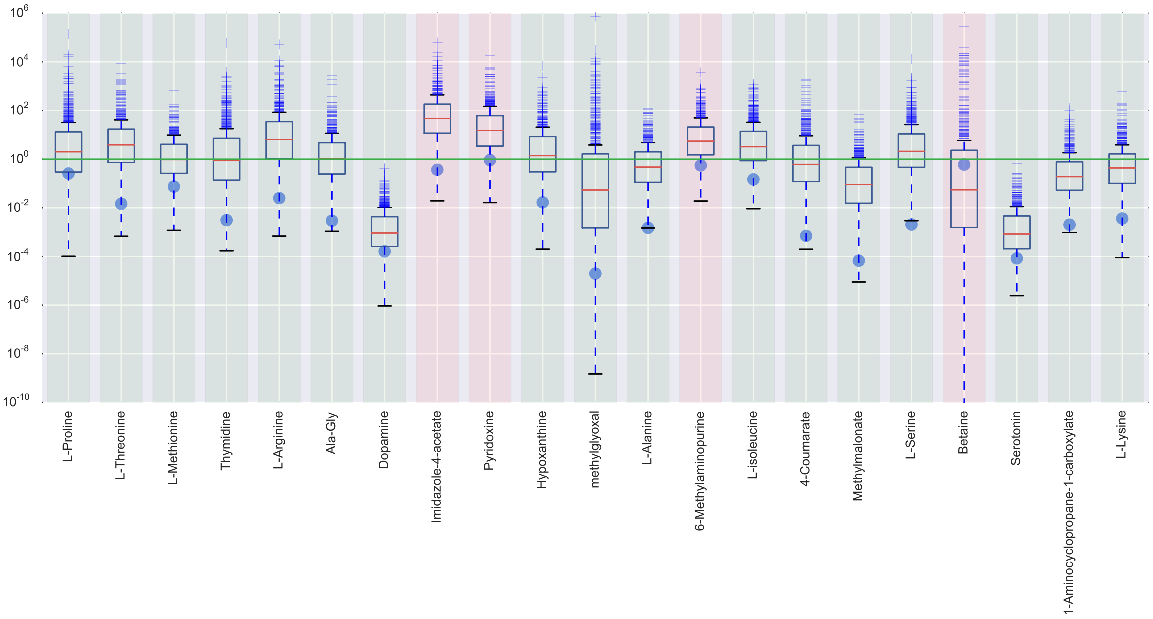
$$\ell = 179.59 \qquad \sigma^2 = 463.265 \times 10^5 \qquad \sigma_n^2 = 4.333 \tag{3.5}$$

Figure 3.7: Plot of test set (*left*) and training set (*right*)



Predictions made on the test set using the new model (3.5) are shown on Figure 3.7. Although the R^2 metrics on this dataset are not as high as on the ground truth, most of the predictions are close to the “ideal” line or the error bars cross it. There are few noticeable outliers, namely *Betaine*, *Serotonin*, *Dopamine* and *methyglyoxal*. However, *methyglyoxal* and *Betaine* are predicted to have high uncertainty. The overall Log-likelihood score on the test set is -52.96.

Figure 3.8: Corrected relative signal intensities using GP with RBF kernel



Having the predictions of the RRF allows us to make corrections of the measured signal intensity (3.6).

$$RRF_{corrected} = \frac{RRF_{measured}}{RRF_{predicted}} \tag{3.6}$$

Because the model predictions are in natural log space, 1000 samples were taken from the predictive distribution and converted to linear space by calculating the exponential.

Figure 3.8 show box-plot of the corrected samples - the red line is the median, the box extends from the lower to the upper quartile and the whiskers show the range of the samples. The round blue markers are the measured Relative Response Factors (RRF). Out of the 21 test metabolites, 17 were corrected towards the true intensity (*See 3.1.3*) and 4 were incorrectly adjusted.

The results above demonstrate that the selected physico-chemical properties can be successfully used to make ionization efficiency predictions and RRF corrections.

3.3.3 Gaussian process regression using molecular structure

Another major goal of this project was to investigate whether information about the molecular structure of the matabolomes can be used for predictions of the ionization efficiency. Two different kernels were tested based on their good performance in other studies [Zhu and Yan, 2014] - Tanimoto and MinMax Tanimoto.

Fitting GP on the ground truth dataset

For this experiment a custom implementation of Gaussian process regression was developed in order to work with the Tanimoto and MinMax Tanimoto kernel. The molecular structures of the training and test compounds was loaded from the InChI using RDKit and saved to SDF files. These files were used to compute the kernels using Rchemcpp.

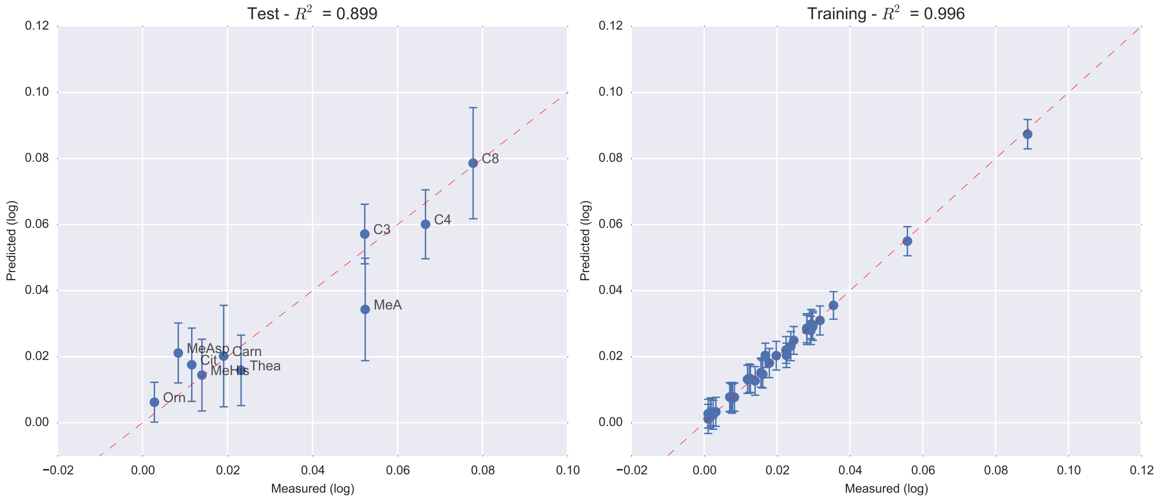
Table 3.3: Tanimoto and MinMax Tanimoto kernel performance on the ground truth test dataset

	R^2	Log-Likelihood
Tanimoto	0.714	-1.7
MinMax Tanimoto	0.898	33.34

Table 3.3 shows the scores achieved by using the two kernels on the ground truth test dataset. Both kernels do not have any hyper-parameters, and were optimized for maximal log-likelihood only on the noise variance σ_n^2. The value of $\sigma_n^2 = 10^{-5}$ was picked for the MinMax Tanimoto model and $\sigma_n^2 = 0.8 \times 10^{-5}$ for the Tanimoto model after 5-Fold CV. Additionally, the kernels were scaled down (multiplied by a constant $\alpha_s = 10^{-5}$) to adjust the high values of the covariance matrices.

Figure 3.9 shows plots of the test set and the training set. The R^2 score on the test set is 0.898 – similar to the score of the linear model reported in [Chalcraft et al., 2009] and the experiment in 3.3.2. However, the R^2 metric on the training dataset is 0.996 which shows that the model is over-fitting on the training.

Figure 3.9: GP with MinMax Tanimoto kernel on the ground truth dataset



The over-fitting can be explained by the way these graph kernels work and the fact that metabolomes are very small molecules, hence the graphs structures are small and a lot of the paths might be similar.

Fitting GP on the Glasgow Polyomics dataset

The same approach was used on the *Glasgow Polyomics* dataset. After cross-validation $\sigma_n^2 = 2.77$ and $\alpha_s = 0.11$ were picked for the MinMax Tanimoto model and $\sigma_n^2 = 1.88$ and $\alpha_s = 1$ for the Tanimoto model.

Table 3.4: Tanimoto and MinMax Tanimoto kernel performance on the Glasgow Polyomics test dataset

	R^2	Log-Likelihood
Tanimoto	0.085	-55.69
MinMax Tanimoto	0.257	-51.70

The model using MinMax Tanimoto kernel achieved better log-likelihood during the CV and better results on the test set. Similar to the previous GP experiment on the *Glasgow Polyomics* dataset, the R^2 score was not as high as in the ground truth experiments. However, the Log-likelihood score on the test dataset is slightly higher – -51.70, compared to -52.96 in the previous experiment. However, as it can be seen from Figure 3.10 most of the predictions have the same level of uncertainty.

Overall, most of the correction of this model were good. Figure 3.11 shows that 14 of the corrections shift the RRF towards the true intensity and 7 are incorrect.

Figure 3.10: GP with MinMax Tanimoto kernel on the Glasgow Polyomics dataset

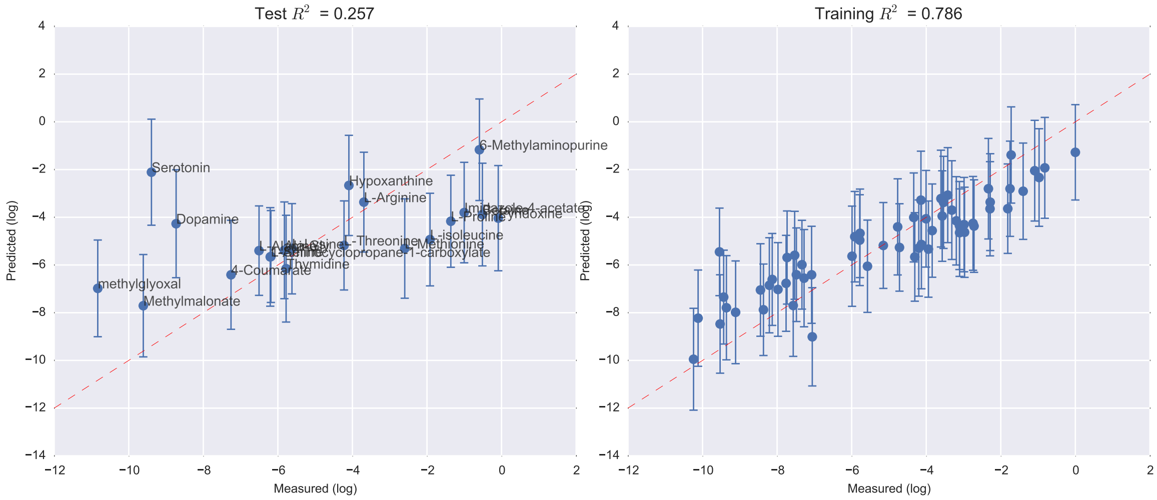
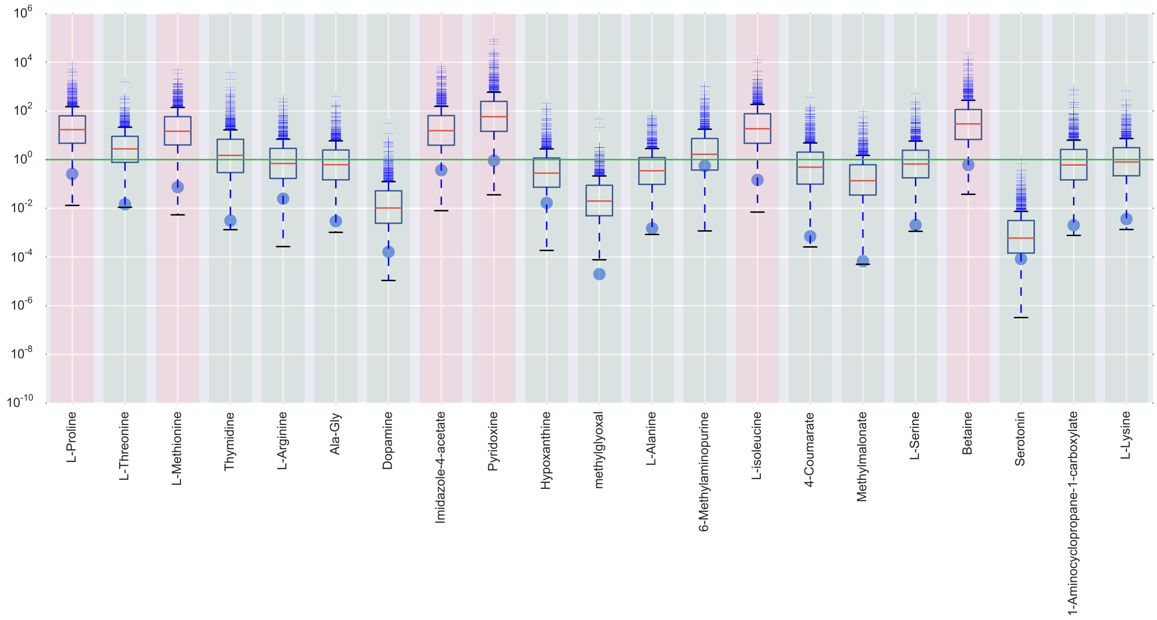


Figure 3.11: Corrected relative signal intensities using GP with MinMax Tanimoto kernel



3.3.4 Gaussian process regression with combined kernels

The last step of the research was to evaluate whether combination of physico-chemical properties and structural information about the molecules can improve the predictions. One of the advantages of Gaussian processes is the possibility to combine kernel functions. As long as a suitable kernel function exists, different types of information (numerical, string, trees, graphs and etc.) about the data points can be combined together to produce the covariance matrix.

In the experiments below, kernel functions were combined using the following configuration:

$$k_{new}(x, x') = \alpha_k \times k_a(x, x') + (1 - \alpha_k) \times k_b(x, x') \tag{3.7}$$

Parameter α_k was introduced to control the weight of each kernel.

Fitting GP with combined kernels on the ground truth dataset

In this experiment the RBF kernel and MinMax Tanimoto kernel were combined. The hyper-parameters learnt in the previous experiments were used. 5-Fold CV was used to find the best value for α_k that maximizes the Log-likelihood:

$$\ell = 29.580 \qquad \sigma^2 = 2.442 \times 10^{-3} \qquad \sigma_n^2 = 4.458 \times 10^{-5} \qquad \alpha_s = 10^{-5} \qquad \alpha_k = 0.192 \tag{3.8}$$

Figure 3.12 shows a curve of the Log-likelihood for different α_k values during the CV. The value $\alpha_k = 0.192$ means that the k_a kernel (RBF) has almost 20% contribution to the final covariance matrix and the rest is from the MinMax Tanimoto kernel.

Figure 3.12: CV Log-likelihood for α_k values on ground truth dataset

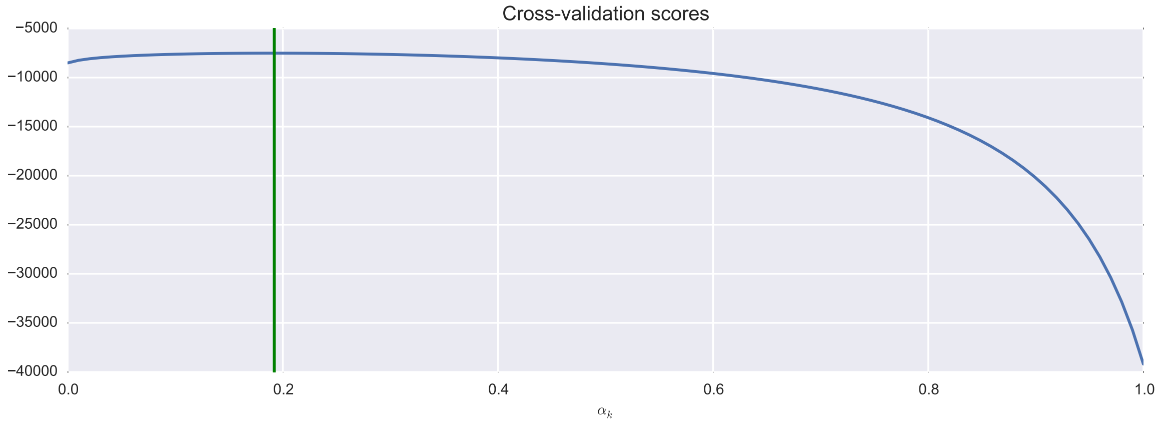
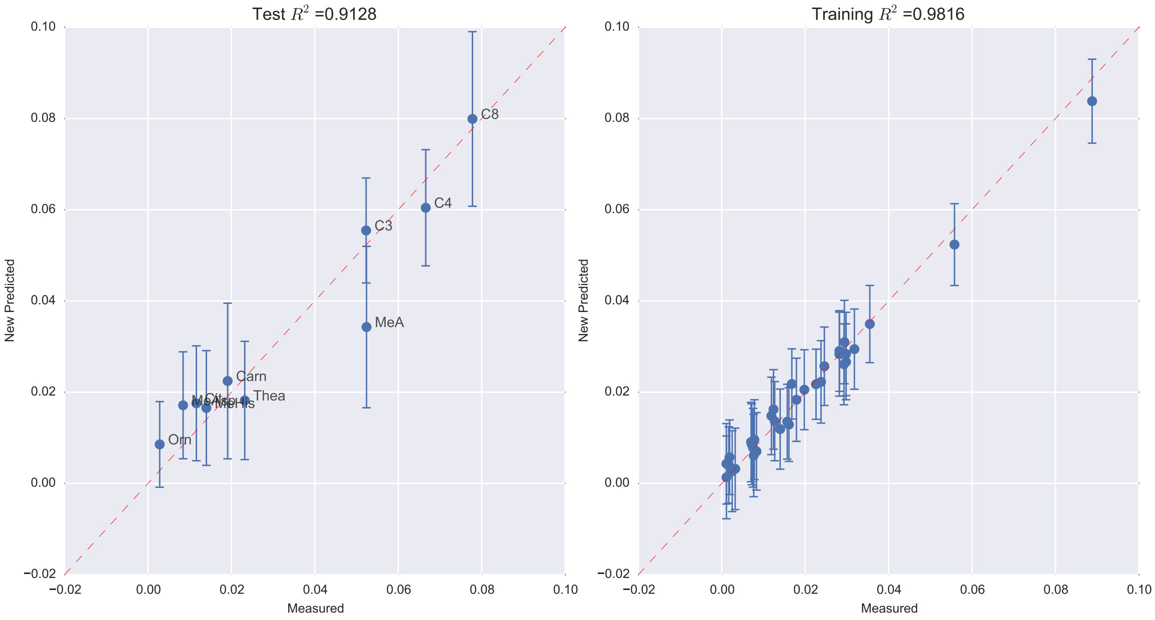


Figure 3.13 shows plot of the test and the training set using the new model with combined kernel. The R^2 score on the test dataset is 0.9128, higher than any of the scores in the previous experiments. However, the Log-likelihood (32.53) is slightly lower compared to the models using RBF (34.89) or MinMax Tanimoto (33.34).

Overall, we can see that using combination of physico-chemical properties information and knowledge about the structure of the molecules can improve the accuracy of the ionization efficiency predictions.

Figure 3.13: Predictions of combined kernels on the ground truth dataset



Fitting GP with combined kernels on the Glasgow Polyomics dataset

To further test the feasibility of kernel combinations the approach from the experiment above was applied on the *Glasgow Polyomics* dataset. Similarly, the the hyper-parameters from the experiment in 3.3.2 were used for the RBF kernel and new $\alpha_s = 0.8$ was picked for the MinMax Tanimoto kernel. The best value for α_k was discovered using 5-Fold CV:

$$\ell = 179.59 \qquad \sigma^2 = 463.265 \times 10^5 \qquad \sigma_n^2 = 4.333 \qquad \alpha_s = 0.8 \qquad \alpha_k = 0.93 \tag{3.9}$$

For this model the contribution of the MinMax Tanimoto kernel is higher that the RBF kernel. Figure 3.14 shows the Log-likelihood on the validation dataset using different α_k values.

Figure 3.14: CV Log-likelihood for α_k values on Glasgow Polyomics dataset

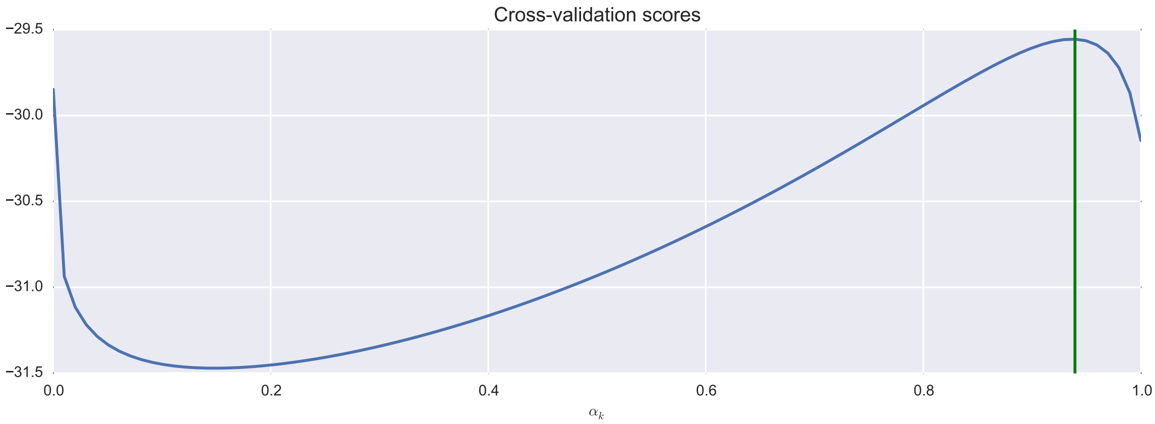
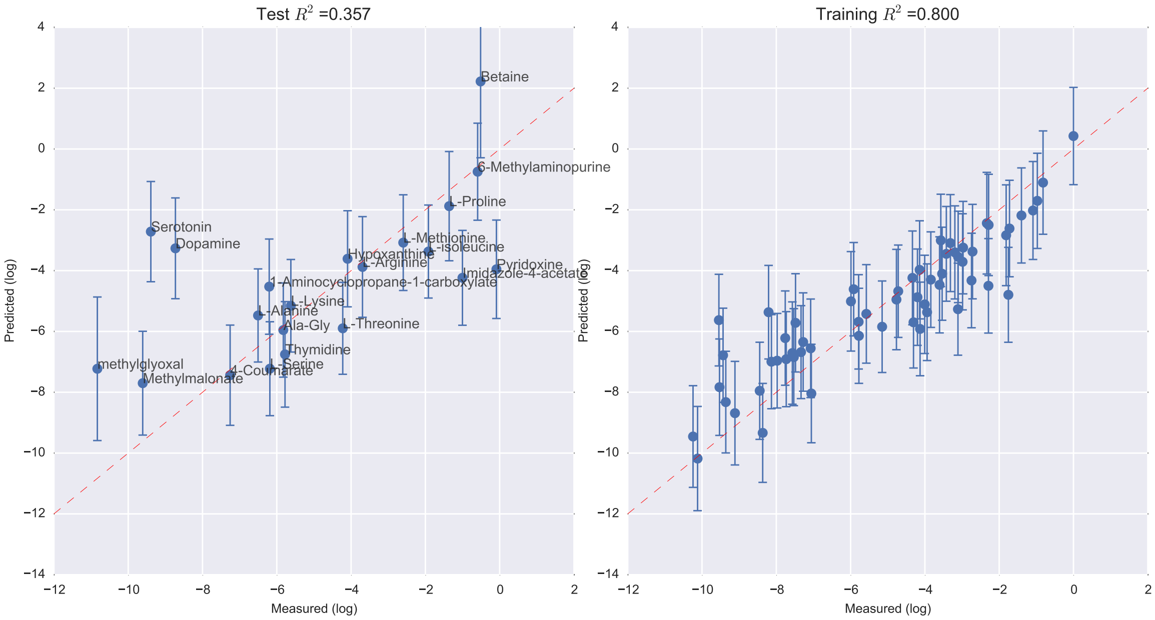


Figure 3.15 shows the results of the new model on the test and the training set. Like in the previous

experiment, the R^2 score is higher than any of the scores achieved on the same dataset and the Log-likelihood (-53.54) is lower compared to RBF (-52.96) and MinMax Tanimoto (-51.70).

Figure 3.15: Predictions from combined kernels on the Glasgow Polyomics dataset



The results from the two experiments above show that combination of physico-chemical properties and structural information about the molecules can lead to better predictions. It is worth mentioning that in these experiments the hyper-parameters for the RBF kernel were fixed based on the results from the previous experiments because of the costly cross-validation of α_k – further cross-search of the parameter space of the combined kernels might lead to a better model.

Chapter 4

Conclusion

This chapter concludes the dissertation by discussing the outcomes of the evaluation, highlights some of the achievements, and potential future work that could lead to improvements.

4.1 Conclusion

In this dissertation we have looked at different features that can be used to predict the ionization efficiency of metabolites in mass spectrometry. The conducted experiments managed to prove all hypotheses defined in this project and to demonstrate improved predictions.

The existing research was evaluated for feasibility and new machine learning techniques and molecular features were tested. First, Gaussian process (GP) regression was applied on the ground truth dataset using the features from the existing research – demonstrating improved accuracy and having the additional benefit of modelling the uncertainty of the predictions. Second, new physico-chemical features were extracted from publicly available databases. Four features were selected and it was shown that GP regression with these features has similar performance to the linear model existing research. Additionally, these features were tested on a dataset from *Glasgow Polyomics* and it was shown that the ionization efficiency predictions can be successfully used for corrections of MS signal intensity. Finally, GP regression model using kernel combining information about the molecular structure and physico-chemical properties was developed and demonstrated improved accuracy on both datasets.

There are few important outcomes from the evaluation:

- Gaussian process regression with RBF kernel can produce more accurate ionization efficiency predictions compared to the ordinary least squares linear regression.
- Other easily accessible physico-chemical features can be used for ionization efficiency predictions without the need of commercial databases and software.
- Combination of molecular structure information and physico-chemical properties using Min-Max Tanimoto kernel and RBF kernel can produce better prediction accuracy than models using each of the kernels alone.

4.2 Future Work

Given the complexity of the task to predict the ionization efficiency of chemical compounds and the time constrains of this project (less than 3 months) only few of the most important experiments were conducted. The list below outline directions for improvements and future work:

Additional features and model selection Although the features selected in 3.3.2 performed well on the two dataset, there are additional features that can be extracted from PubChem. For example, molecular volume (MV) is considered as good descriptor in [Chalcraft et al., 2009] and [Oss et al., 2010] – estimated MV value is available through the PubChem REST API. Additionally, in order to select physico-chemical characteristics of the compounds that maximize the accuracy of the GP regression, extensive model selection can be performed by testing all combinations of features or just a reduced set selected with the help of domain experts.

Full cross-search of parameter space for combined kernels In 3.3.4 RBF and MinMax Tanimoto kernels were used with the best hyper-parameter values discovered in previous experiments and only the α_k parameters was used in the cross-validation. Although full cross-search of the parameter space for the combined kernels will be computationally expensive, it might produce better model with higher accuracy.

Mass spectra Although data from MassBank was parsed and extracted, only a small number of compounds in the ground truth dataset had matching compounds in MassBank. Because of the lack of data, experiments using mass spectra were not performed.

It would be interesting to see whether GP regression on mass spectra using the kernels proposed in [Heinonen et al., 2012] can be applied for ionization efficiency predictions and MS signal intensity corrections.

Different adduct levels In the experiments on the *Glasgow Polyomics* dataset only data from the M+H adduct and solute dilution 1:1 was used. Combining information from the different adducts might improve predictions by giving additional insight of the ionization behaviour of the compounds.

Bibliography

[Bantscheff et al., 2007] Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007). Quantitative mass spectrometry in proteomics: A critical review. *Analytical and Bioanalytical Chemistry*, 389(4):1017–1031.

[Bolton et al., 2008] Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. (2008). PubChem: Integrated Platform of Small Molecules and Biological Activities.

[Chalcraft et al., 2009] Chalcraft, K. R., Lee, R., Mills, C., and Britz-McKibbin, P. (2009). Virtual Quantification of Metabolites by Capillary Electrophoresis-Electrospray Ionization-Mass Spectrometry: Predicting Ionization Efficiency Without Chemical Standards. *Analytical Chemistry*, 81(7):2506–2515.

[Gautier, 2014] Gautier, L. (2014). rpy2 - R in Python. `http://rpy.sourceforge.net`.

[Heinonen et al., 2012] Heinonen, M., Shen, H., Zamboni, N., and Rousu, J. (2012). Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*, 28(18):2333–2341.

[Horai et al., 2010] Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M. Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., and Nishioka, T. (2010). Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714.

[Hunter, 2007] Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95.

[IUPAC, 2005] IUPAC (2005). The iupac international chemical identifier (inchi). `http://www.iupac.org/home/publications/e-resources/inchi.html`.

[IUPAC, 2009] IUPAC (2009). InChI Software Version 1.02 final, implemented for Standard InChI/InChIKey Summary. `http://www.iupac.org/home/publications/e-resources/inchi/r102-summary.html`.

[Jessome and Volmer, 2006] Jessome, L. L. and Volmer, D. a. (2006). Ion Suppression: A Major Concern in Mass Spectrometry. *LCGC North America*, 24:498–510.

[Jones et al., 2001] Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python. [Online; accessed 2015-08-26].

[Klambauer et al., 2015] Klambauer, G., Wischenbart, M., Mahr, M., Unterthiner, T., Mayr, A., and Hochreiter, S. (2015). Rchemcpp: a web service for structural analoging in chembl, drugbank and the connectivity map. *Bioinformatics*.

[Landrum, 2014] Landrum, G. (2014). RDKit: Open-source cheminformatics. Release 2014.03.1.

[MassBank Project, 2013] MassBank Project (2013). MassBank Record Format 2.09 (draft). `http://www.massbank.jp/manuals/MassBankRecord_en.pdf`.

[McKinney, 2011] McKinney, W. (2011). pandas: a foundational python library for data analysis and statistics.

[Mei et al., 2003] Mei, H., Hsieh, Y., Nardo, C., Xu, X., Wang, S., Ng, K., and Korfmacher, W. a. (2003). Investigation of matrix effects in bioanalytical high-performance liquid chromatography/tandem mass spectrometric assays: Application to drug discovery. *Rapid Communications in Mass Spectrometry*, 17(1):97–103.

[Oss et al., 2010] Oss, M., Kruve, A., Herodes, K., and Leito, I. (2010). Electrospray ionization efficiency scale of organic compound. *Analytical Chemistry*, 82(7):2865–2872.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Pérez and Granger, 2007] Pérez, F. and Granger, B. E. (2007). IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29.

[R Development Core Team, 2008] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Ralaivola et al., 2005] Ralaivola, L., Swamidass, S. J., Saigo, H., and Baldi, P. (2005). Graph kernels for chemical informatics. *Neural Networks*, 18(8):1093–1110.

[Rasmussen and Williams, 2005] Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

[Smith et al., 2014] Smith, R., Mathis, A. D., Ventura, D., and Prince, J. T. (2014). Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view. *BMC Bioinformatics*, 15(Suppl 7):S9.

[The GPy authors, 2014] The GPy authors (2012–2014). GPy: A gaussian process framework in python. `http://github.com/SheffieldML/GPy`.

[Walt et al., 2011] Walt, S. v. d., Colbert, S. C., and Varoquaux, G. (2011). The numpy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2).

[Waskom et al., 2015] Waskom, M., Botvinnik, O., Hobson, P., Warmenhoven, J., Cole, J. B., Halchenko, Y., Vanderplas, J., Hoyer, S., Villalba, S., Quintero, E., and et al. (2015). seaborn: v0.6.0 (june 2015).

[Wishart et al., 2013] Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., and Scalbert, A. (2013). HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Research*, 41(D1).

[Zhou et al., 2012] Zhou, B., Xiao, J. F., Tuli, L., and Ressom, H. W. (2012). LC-MS-based metabolomics. *Molecular bioSystems*, 8(2):470–81.

[Zhu and Yan, 2014] Zhu, Y. and Yan, C. (2014). Graph methods for predicting the function of chemical compounds. In *Granular Computing (GrC), 2014 IEEE International Conference on*, pages 386–390.