Applying Machine Learning to the Ultrafast Shape Recognition Family of Methods in Ligand-Based Virtual Screening

Etienne Bonanno, Department of A.I., Faculty of ICT, University of Malta, etienne.bonanno.97@um.edu.mt

Abstract—Computer Aided Drug Discovery is concerned with the algorithmic modelling of chemical interactions of bioactive compounds on protein molecules with the aim of discovering new drugs. Ligand Based Virtual Screening is a branch thereof concerned with using known active compounds to screen other unknown molecules for biological activity. Ultrafast Shape Recognition(USR), along with its derivatives, are techniques that compress 3 dimensional information about molecular shape in order to optimise comparisons between molecules. We explore the use of Machine Learning techniques for augmenting the performance of USR so as to improve active compound detection as well as improve processing times. We detail the current state of the project as well as outline the way forward.

Index Terms—Machine Learning, Big Data , Cheminformatics, Computer Aided Drug Discovery, Ligand-Based Virtual Screening, Ultrafast Shape Recognition

1 Introduction

Cheminformatics is a multidisciplinary field of study that is concerned with applying techniques in statistics and computer science to the study of biochemistry [1]. One of the major goals of cheminformatics is that of discovering new drugs by computational means, referred to as *Computer-Aided Drug Discovery* or CADD. Virtual Screening (VS) is a method of Computational Drug Discovery that has been receiving increased attention [2]. The aim of this field is to streamline the costly and time-consuming process of physically screening new drug-like compounds for biological activity in the lab. By computationally pre-screening compounds for those that are most likely to exhibit such activity, laboratory screening time can be drastically reduced [1]. Advances in processing power and high-capacity storage as well as development of Big-Data techniques has today made this process of computation optimisation of compound screening feasible resulting in significant savings of time and cost.

Virtual Screening techniques are divided into two main categories: structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS). SBVS focuses on using structural information about target proteins to find new compounds complementary in shape to known protein binding sites [3], [4]. LBVS, in contrast, assumes no prior knowledge about protein properties. Instead it focuses on using known active compounds known to bind to a given target protein (ligands) to search for other similar compounds that have a high probability of also being active against the same protein [5]. LBVS methods can perform similarity matching based on a wide variety of different properties of the ligands,, some techniques being based on chemical properties while others on three dimensional shape.

The underpinning for the idea of matching molecules of similar shape is Emil Fischer's lock and key hypothesis devised in the 19th century. Fischer postulates that a molecule will bind to a protein if it matches the shape of a binding site on the protein like a key matching a lock [6] as illustrated in Fig. 1. It therefore follows that molecules having a similar shape should have a high probability of binding to the same protein. The process of matching molecular shapes is, however, computationally intensive because any single molecule can often take on different shapes depending on the arrangement of chemical bonds between the atoms in the molecule. Molecules can also be rotated relative to each other, making it necessary to perform an initial alignment step before comparing their shapes. Furthermore a ligand might take a different shape when it is bound to a protein than when it is unbound [7].

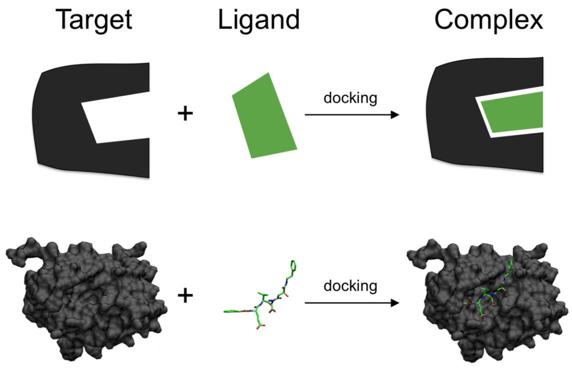


Fig. 1. A simple illustration of Fischer's Lock-and-key principle (from https://en.wikipedia.org/wiki/Docking_(molecular)).

Ultrafast Shape Recognition (USR) is an LBVS technique which aims to get around the heavy computational requirements of molecule alignment for SBVS. The technique was developed by Ballester et al. [8], [9] and it involves distilling molecular shape into a descriptor vector made up of 12 decimals. These descriptors are then compared directly using a modified Manhattan distance metric, obviating the need for considering molecule alignment. This method was developed in 2007, however since then, extensions to this algorithm have been proposed that augment the purely shape-based descriptors of USR with other chemical properties of the molecule such as Electroshape that uses atomic partial charges [10] and USRCAT, using atom types [11], obtaining better virtual screening scores than the basic USR algorithm.

2 Aims and Objectives

2.1 Aims

The aim of this study is to augment the USR algorithm for LBVS with ML techniques in order to boost its virtual screening performance.

The dissertation will deal with three main scientific research questions, namely:

1) Can machine-learning techniques be used instead of naïve Manhattan distance to improve Virtual Screening performance based on USR and USR-derived descriptors?
2) Can the USR descriptors be shortened while preserving the predictive power of the method?
3) What is the minimal amount of data required to adequately train the machine learning model?

The application of ML algorithms to the problem of USR similarity matching constitutes a new approach to the problem of virtual screening based on molecular shape and is therefore a worthwhile avenue of research.

Furthermore, if, through ML, it is found to be possible to shorten the USR descriptors to less than the nominal number of elements with no decrease in performance, it would permit significant decreases in storage requirements for large compound databases.

The motivation for the third question is that it is always the case that the number of known actives for a particular target in LBVS is limited, therefore it is crucial to determine how the effectiveness of a machine learning model applied to an LBVS dataset varies with severely unbalanced training data and with limited training examples.

2.2 Objectives

The objectives to be reached in order to achieve the aims of this research are the following:

- To create an implementation of USR to serve as the baseline against which to evaluate machine learning models (henceforth, naïve USR).
- To leverage the online chemical database known as the Directory of Useful Decoys Enhanced (DUD-E) [12] which has been created especially for the purpose of evaluating LBVS methods.
- To modify naïve USR, varying the length of the descriptor and derive baseline performance measures of naïve USR vs. USR with modified descriptor length for a preselected set of protein targets.
- To repeat the above experiments with the introduction of various ML models, mapping out their performance against the baseline of naïve USR with various descriptor lengths and protein targets.
- To identify an optimum machine learning model
- To repeat the above with other, better performing, USR-derived descriptors

3 Background and Literature Review

Ligand Based Virtual Screening involves using one or more molecules that are known to successfully bind to some target protein (ligands) as templates for finding other similar molecules which have a high likelihood of binding to the same protein. These template molecules are referred to as *active molecules*, or *actives*. A dataset of unknown molecules is then graded and ranked according to some similarity metric compared to the active molecules. It is desirable that potentially active molecules be graded highly so as to appear at the start of the sorted list. This will ensure that subsequent laboratory screening of the sorted molecules will find active ones as early as possible in the process, resulting in reduced time and cost. This is referred to as *early enrichment*.

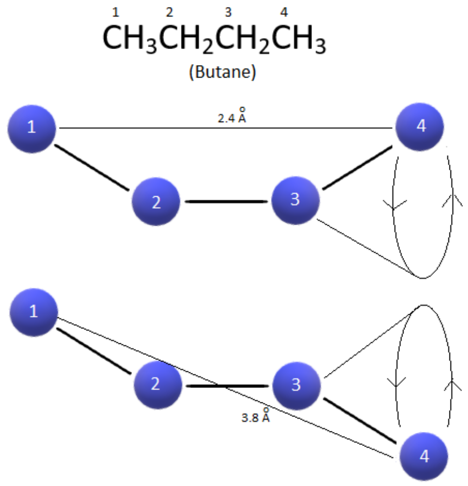


Fig. 2. Two molecular conformations of the butane molecule

When evaluating a LBVS technique, a dataset made up of known actives along with, usually, a much larger number of *decoys* is used. Decoys are compounds selected to be similar to the active molecules in terms of physical properties such as molecular weight and size, but also chemically distinct from them so as to make them unlikely to bind with the target protein. Using such a labelled dataset it is possible to evaluate the results of the virtual screening process by the number of known actives that have been graded towards the top of the sorted list.

The most established method for LBVS is molecular fingerprinting. This involves encoding a given set of molecular properties into a fixed size bitmap, referred to as a molecular fingerprint [13]. Similarity metrics, such as the Tanimoto similarity index [14], are then applied on these fingerprints in order to extract a measure of overlap between molecules [15], [16]. In attempts to improve similarity scores, various researchers have attempted to apply machine learning techniques to the problem of similarity searching on these molecular fingerprints, with varying results [16], [17], [18].

A different branch of LBVS is shape-based similarity searching, or molecular shape comparison. This makes use of 3-D molecular structural information to determine similarity between molecules [19]. These techniques are known to yield fewer matches than 2-D fingerprint methods [20], however the matches they identify tend to be different to those found by 2-D fingerprints and therefore are useful nevertheless [19]. Additionally, shape-based methods are also capable of performing *scaffold hopping*, meaning that molecules having a similar shape to an active, but composed of different elements can still be picked up as highly similar. This can result in compounds that show similar biological activity as a known active, while circumventing issues such as patent restrictions [9].

Shape-matching methods have to deal with a considerable problem that is not present in molecular fingerprinting techniques, however. Molecules are, in general, not rigid structures - atomic links that are not bound in ring configurations are free to rotate allowing any given molecule to take a large number of different shapes, called *conformations*. The more rotatable bonds the molecule contains, the more varied the conformations that the molecule can take (see Fig. 2). A molecule's conformer model does not alter its chemical properties, however a molecule will only bind within a protein active binding site when it takes the correct shape to conform to it. This is not a problem for fingerprinting methods because fingerprints depend on the chemical properties of the molecule and not its shape in three dimensions. For shape-based methods, however this is a considerable challenge because for each molecule, a large enough number of conformations to adequately sample the molecule's conformational space have to be mathematically generated, causing an explosion in dataset size. The optimal number of conformations to generate depends on the number of rotatable bonds present in the molecule and guidelines in this regard were proposed by Ebejer et al. [21]. It is clear that due to the necessity of generating a large number of conformations it is desirable for a shape-based technique to be as computationally efficient as possible.

Shape-based approaches can be categorised into two main groups: super-positional methods, and non super-positional methods.

Super-positional methods attempt to optimally align molecular structures in 3-D space in order to quantify overlaps between molecular shape, taking into account forces between atoms and the differing valid conformations of molecules. These methods have the advantage of yielding good results, however they are also much slower than fingerprinting techniques. This is because they require the computationally intensive optimum alignment of the molecules in 3-D space.

In contrast, non super-positional methods, attempt to use molecular shape information in such a manner as to make alignments unnecessary for similarity comparison. This normally involves pre-computing some set of descriptors from the 3-D structure of the molecule. These approaches are faster than super-positional approaches, but do not yield comparably favourable results. A review of both types of methods is given by Ballester in [9] and McGaughey [22].

In an effort to preserve the speed benefits of non super-positional approaches and preserve the screening performance of super-positional techniques, Ballester and Richards proposed a novel non super-positional approach they named Ultrafast Shape Recognition (USR) in 2007 [9], [8].

In their research, the authors point out that the shape of a molecule can be encoded by taking the distance distributions of each atom to four centroids located inside the molecule. The centroid selection can be arbitrary, however the authors chose four well-defined centroids as follows:

1) The molecular centroid (*ctd*)
2) The closest atom to the centroid (*cst*)
3) The furthest atom from the centroid (*fct*)
4) The furthest atom to *fct* (*ftf*)

This gives four separate distance distributions.

By making use of a result from statistics stating that a distribution is completely determined by its moments [23], the four distributions are then condensed into their respective first three moments, corresponding to the mean, the variance and the skewness. This results in a vector of 12 decimal values for a given conformer. The authors propose using this vector as a stand-in for the molecules 3-D structure in similarity comparisons.

The performance of USR was evaluated formally by Ballester et al in 2009 [24] by performing retrospective virtual screening experiments comparing USR to a commercially available non-superpositional VS system called Eigenspectrum Shape Fingerprints (ESshape3D) [25]. They generated a test set from the DrugBank database by selecting 8 sets of active molecules with respect to an unbiased variety of targets. The Enrichment Factor at 1% comparing USR and ESshape3D showed that USR performed better than ESshape3D, scoring an average of 10.4 over the 8 selected targets, as compared to 6.6 scored by ESshape3D.

Other studies have compared the effectiveness of USR compared to other shape-based methods [26] as well as fingerprint methods [20]. Comparison between different methods, however, has proven to be problematic as there is no standard method for shape-based similarity comparisons. Ballester points out [27] for example that in [20] USR is being compared to ROCS [28], [29], a widely used shape similarity method based on 3-D superposition, but ROCS is known to not consistently identify similar molecules and therefore it is not correct to compare USR to ROCS and attribute all the discrepancies to USR. Apart from this [20] also compares the effectiveness of USR to a host of other shape-based and fingerprint-based methods using the Directory of Useful Decoys (DUD) database [30]. The problem with this, however, is that the molecules in DUD are categorised using fingerprint methods, and therefore unsurprisingly, fingerprint methods achieve better benchmarks in this paper than those achieved by shape-based methods. Furthermore, shortcomings with the DUD when it comes to its use in VS have been identified and a new, much improved successor to the DUD, named DUD-Enhanced has been compiled in order to minimise these problems [12]. Retrospective Virtual Screening studies have not yet been performed for USR using the DUD-E.

Since the emergence of USR, other researchers have augmented the purely shape-based information encoded within the USR descriptors with other information relating to the chemical and physical properties of the molecule. These modifications to the basic USR technique have yielded considerable improvements in early enrichment.

The first such enhancement, named Chiral Shape Recognition (CSR), was proposed by Armstrong et al. in 2009 [31]. They proposed a modification to the centroid selection process in USR to enable the algorithm to distinguish between *enantiomers*, i.e. pairs of molecules that are mirror-images of each other. In 2010, Armstrong et al. proposed ElectroShape, a further extension of USR that incorporates CSR as well as information about the electrostatic properties of the atomic bonds [10]. This method resulted in a near doubling of enrichment ratio at 1% over USR. A different extension to USR named USRCAT was proposed in 2012 by Schreyer et al. [11] which was implemented as part of the CREDO Structural Interatomics database and combines the pure shape information of USR with information about the atom types making up the molecule. This method obtained performance scores comparable to ElectroShape when evaluated over the DUD-E database.

4 Methodology

We will carry out a series of retrospective virtual screening experiments based on several Machine Learning algorithms and on both naïve USR descriptors as well as Electroshape descriptors. We chose to implement ElectroShape in preference to the other USR-derived algorithms because it gives state of the art results at a relatively low cost in increased complexity. USRCAT gives comparable results, however it is more complex to implement and necessitates access to the CREDO database, which is not freely available.

In order achieve these goals, there are three main top-level tasks that need to be performed.

1) Conformer generation for a variety of target proteins.
2) Implementation of a test harness to execute and evaluate different algorithms in a uniform manner.
3) Implementation of Machine Learning-based similarity functions to be evaluated.

Conformer generation will be performed once as a pre-processing step to generate the datasets that will be needed for the retrospective screening experiments. A variety of protein targets will be chosen to match as closely as possible the existing literature. The DUD-E datasets for the chosen proteins will then be processed using the open-source RdKit [32] library to generate conformers for each compound according to the guidelines in [21]. These conformers will then be used to generate several datasets with USR descriptors, ElectroShape descriptors and abbreviated versions of each type of descriptor using fewer moments and fewer USR reference points, in order to explore the second research question.

The test harness will provide a pluggable architecture for testing and evaluating VS algorithms. It will feed the datasets resulting from the conformer generation step to a given similarity metric and evaluate the resulting ranked results.

4.1 Dataset

The VS datasets that will be used will be selected from the Directory of Useful Decoys-Enhanced (DUD-E). The DUD-E provides chemical compound datasets for use in retrospective screening runs with a total of 22,886 active compounds against 102 target proteins as well as corresponding decoy molecules with a ratio of 50 decoys to each active.

The actives and decoys for each protein are provided in the DUD-E in standard Simplified Molecular-Input Line-Entry System (SMILES) format [33]. SMILES is a way of encoding the chemical structure of a molecule in string format and it is parsed natively by the RDKit library.

From every SMILES molecule representation, RDKit molecule objects can be created through which valid conformations for the molecule can be calculated. These can then, in turn, be used to generate USR and ElectroShape descriptors to be used as features for Machine Learning.

5 Evaluation

There are several evaluation methods that can be used for virtual screening studies [34], [35], however two are most commonly used in the literature - Receiver Operating Characteristic (ROC) curves [36] and the Enrichment Factor. Enrichment is defined as the number of actives ranked within a threshold, commonly within the top 1% or 5%, expressed as a ratio to the number of actives that would be found by chance, i.e. if $A^{x\%}$ is the number of actives ranked in the top x% of the sorted dataset and $C^{x\%}$ is the number of compounds in the top x%, A is the total number of actives in the dataset and C is the total number of compounds, then,

$$EF^{x\%} = \frac{A^{x\%}/C^{x\%}}{A/C} \tag{1}$$

Clearly, the EF depends directly on the ratio of actives to decoys in the entire dataset and therefore in order to compare VS scores meaningfully, one must do so for VS runs on the same dataset. This is a problem in the VS literature as, in general, it is difficult to compare methodologies across studies in this way. In this study, we will make sure to preserve the active/decoy ratio in order to compare performances meaninfgully.

6 Current Progress

A considerable amount of background research has already been done on the topic of Cheminformatics in general and virtual screening in particular leading to the implementation of several proof of concept ideas as described in this section.

6.0.1 Conformer and Descriptor Generation

The pre-processing for conformer generation and descriptor generation was implemented in Python using the RDKit library and custom code to generate the USR descriptors. USR descriptors and total energy for each conformer are generated as well as the number of rotatable bonds in the molecule by the pre-processing step. This process was designed to use a map/reduce paradigm in order to make it easily migratable to a distributed platform such as Hadoop or Spark when taking it beyond a proof of concept phase. This will be necessary due to the large volume of data involved, which is of the order to 10G per protein target.

Using this process we are in the process of generating datasets for seven diverse protein targets being chosen to be as similar as possible to those used in [24]. These can be used as training data for machine learning models.

6.0.2 Machine Learning Experiments

From a machine learning perspective, a USR descriptor dataset can be regarded as an unbalanced dataset with binary labels, i.e. ”active” and ”decoy”. We can train a wide variety of supervised machine learning algorithms using such data, effectively making full use of the information present in the active samples as well as in the decoys.

In real-life scenarios, however, it is not always the case that a full library of active and decoy compounds is available to train a supervised model. In traditional retrospective virtual screening processes, in fact, decoys are only used as a way to evaluate the performance of the method and only information related to the active compounds governs the classification process.

Now, as described by Ballester in [24], taking the set of conformers of all active ligands for a given protein and clustering them based on shape will reveal well defined centroids corresponding to the possible binding sites of the protein. Comparing test compounds to these centroids would be effectively finding compounds that match the shapes of the protein’s binding pockets. This is an unsupervised clustering problem.

We want to use a machine learning method that could be applied in both supervised mode using binary labels on all the available data as well as in a manner that replicates traditional retrospective virtual screening, being trained exclusively on active compounds in an unsupervised manner.

An ideal machine learning algorithm to use in this situation is the Gaussian Mixture Model (GMM) [37]. GMMs are an unsupervised algorithm that models data points as mixtures of weighted Gaussian distributions, i.e. given the definition of the Gaussian distribution in D dimensions as

$$\mathcal{N}(\vec{x} \mid \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} exp\Big(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big) \tag{2}$$

a Gaussian mixture with M components is given by:

$$\mathcal{M}(\vec{x} \mid \vec{\mu}, \Sigma) = \sum_{k=1}^{M} c_k \mathcal{N}(\vec{x} \mid \vec{\mu}, \Sigma) \tag{3}$$

where c_k is the weight assigned to component k. Thus, the parameters associated with GMM $\mathcal{M}$ with N components are $\vec{\mu}$, a vector of N means, Σ, a NxN covariance matrix and $\vec{c}$ a vector of the N weights associated with the N Gaussians.

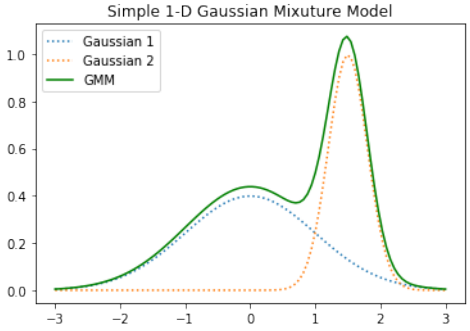


Fig. 3. A plot of a simple 1-D Gaussian Mixture Model.

A simple example of a plotted 1-D GMM with two components is shown in Fig. 3

The GMM parameters have to be learned from the input data and this is done using an iterative algorithm called *Expectation Maximisation*. The algorithm starts by assigning initial values to the Gaussian parameters. These could be random, but they are generally assigned using the k-means clustering algorithm on the data points. Once initial values are set, two steps, analogous to the k-means algorithm are repeated until convergence:

- E-Step: evaluate the responsibilities of every component using the current parameter values, i.e. which Gaussian is responsible for each sample.
- M-Step: re-estimate the Gaussian parameters given the current responsibilities on the samples.

The major hyper-parameter that has to be tuned in a GMM is the number of Gaussian components to be used in the model. This can be expensive to determine, and is usually found empirically by a grid-search process. In [38], however, the authors determine that the number of rotatable bonds between two atoms is a good value use as the number of components of the GMM describing the distance distribution between those two atoms. This enables them to side-step the necessity for the hyper-parameter tuning step, obtaining a large reduction in running time.

The effectiveness of a GMM also depends on the constraints placed upon Σ, the covaraiance matrix. Putting no constraints on Σ maximises the GMM's expressive power, however also involves heavy computation. In most cases, Σ is constrained to be a diagonal matrix, resulting in lighter computational requirements while sacrificing some accuracy in the model. This can be compensated for, however, by using more components and for this reason, much of the results present in the literature assume diagonal Σ

GMMs are used extensively in speech processing, where they are used to classify sound samples against trained phoneme models, however they have also been used in virtual screening and protein-ligand docking [28], [29], [38], [39]. It is possible to use GMMs to classify samples in both a supervised as well as an unsupervised manner.

While GMMs are essentially an unsupervised probabilistic clustering technique, it is possible to employ them as supervised models by training multiple GMMs, one per label, and classify test data by Maximum A-Posteriori (MAP) estimation on the outputs of all the GMMs. i.e. in this case we train one GMM on the active conformers and one on the decoys. New compounds to be classified are evaluated on both, giving an average log likelihood value over all the test compound conformers for each GMM. Thus the test compound is classified by the label corresponding to the GMM that gives the largest average log likelihood.

This gives us a binary active/non-active classification, however we need a ordered ranking of compounds in order to estimate early enrichment. In order to obtain this for each test compound we took the difference between the log likelihood against the active GMM and that against the decoy GMM. The rationale behind this being to highly rank compounds with a much higher likelihood of being actives than decoys. We then calculated ROC/AUC metrics and enrichment factor based on this ranking.

To date we have applied this scheme to one protein target (Adenosine A2a receptor (GPCR)) obtaining AUC scores of 0.91 and and Enrichment factor at 1% of 49, which is close to a theoretical maximum given the DUD-E active/decoy ratio of 1:50.

In order to use GMMs in unsupervised mode, emulating more closely a traditional retrospective virtual screening experiment, we only train a GMM on the active compounds for a protein and evaluate the log likelihood of the test compounds only on the active GMM, using this value to rank the compounds.

The AUC obtained for the same protein using this scheme was 0.80 with an enrichment at 1% of 40. As expected, this is a lower score than for the supervised method, however is is nevertheless an excellent result.

In order to directly compare our performance to Ballester's original algorithm, we need to run the retrospective screening experiment using pure USR on the same targets, however. Due to the big data nature of the problem, to date we have not obtained this result over the entire set of conformers for these target proteins but only for a limited subset of 10 actives preserving the active/decoy ratio, giving an enrichment factor at 1% of 19. This value is clearly preliminary, however it is to be noted that the use of GMMs has made it possible to run the VS experiment in a fraction of the time that it takes to perform using USR alone, and producing excellent results. This is very encouraging.

On a related avenue, we are also exploring a different way of using GMMs for similarity matching of molecules over variable conformational states, taking an approach similar to that of Jahn et al. in [38] and [39]. In their research the authors build a similarity metric between molecules by encoding molecule distances under varying rotational angles of rotatable bonds as GMMs and estimating the similarity by computing the overlap between GMMs using the Expected Likelihood kernel function [40].

In our approach we encode each compound in the dataset as a GMM with the intention of comparing the GMM models of two compounds rather than all their conformers separately, potentially resulting in large time savings. We have experimented with several similarity functions between GMMs and to date the best results were given by that proposed by Zhou et al. in [41] giving enrichment at 1% of 42 and a deceptively low AUC of 0.58. Research about GMM similarity metrics is still ongoing.

7 Conclusion

This report presented the ongoing research in pursuit of the aims outlined in section 2.1. A good amount of groundwork has already been completed, and some initial results that were obtained were very encouraging, however, more research about the performance of several GMM similarity functions still needs to be done. The next step will be to consider Electroshape descriptors along with pure USR as well as initiating explorations regarding to research questions 2 and 3. Prior to this, however, processing will have to be moved to the cloud in order to handle the massive amount of processing that will be required to generate datasets from all the target proteins.

A plan for the remaining work is presented in Appendix A.

References

[1] A. R. Leach and V. J. Gillet, *An Introduction to Cheminformatics*, revised ed. Sheffield: Springer, 2007.

[2] A. Lavecchia and C. D. Giovanni, "Virtual screening strategies in drug discovery: A critical review," *Current Medicinal Chemistry*, vol. 20, no. 23, pp. 2839–2860, 2013.

[3] P. D. Lyne, "Structure-based virtual screening: An overview," *Drug Discovery Today*, vol. 7, no. 20, pp. 1047–1055, 2002.

[4] H. Eckert and J. Bajorath, "Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches," *Drug Discovery Today*, vol. 12, no. 5-6, pp. 225–233, 2007.

[5] F. Stahura and J. Bajorath, "Virtual Screening Methods that Complement HTS," *Combinatorial Chemistry & High Throughput Screening*, vol. 7, no. 4, pp. 259–269, 2004.

[6] E. Fischer, "Einfluss der configuration auf die wirkung der enzyme," *Berichte der deutschen chemischen Gesellschaft*, vol. 27, no. 3, pp. 2985–2993, 1894.

[7] W. L. Jorgensen, "Rusting of the lock and key model for protein-ligand binding," *Science*, vol. 254, no. 5034, pp. 954–956, 1991.

[8] P. J. Ballester and W. G. Richards, "Ultrafast shape recognition for similarity search in molecular databases," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 463, no. 2081, pp. 1307–1321, 2007.

[9] ——, "Ultrafast shape recognition to search compound databases for similar molecular shapes." *Journal of computational chemistry*, vol. 28, no. 10, pp. 1711–23, jul 2007.

[10] M. S. Armstrong, G. M. Morris, P. W. Finn, R. Sharma, L. Moretti, R. I. Cooper, and W. G. Richards, "ElectroShape: Fast molecular similarity calculations incorporating shape, chirality and electrostatics," *Journal of Computer-Aided Molecular Design*, vol. 24, no. 9, pp. 789–801, 2010.

[11] A. M. Schreyer and T. Blundell, "USRCAT: Real-time ultrafast shape recognition with pharmacophoric constraints," *Journal of Cheminformatics*, vol. 4, no. 11, 2012.

[12] M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet, "Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking," *Journal of Medicinal Chemistry*, vol. 55, no. 14, pp. 6582–6594, 2012.

[13] L. M. Kauvar, D. L. Higgins, H. O. Villar, J. R. Sportsman, Å. Engqvist-Goldstein, R. Bukar, K. E. Bauer, H. Dilley, and D. M. Rocke, "Predicting ligand binding to proteins by affinity fingerprinting," *Chemistry and Biology*, vol. 2, no. 2, pp. 107–118, 1995.

[14] P. Willett, "Similarity-based virtual screening using 2D fingerprints," *Drug Discovery Today*, vol. 11, no. 23-24, pp. 1046–1053, 2006.

[15] C. Seung-Seok, C. Sung-Hyuk, and C. C. Tappert, "A Survey of Binary Similarity and Distance Measures." *Journal of Systemics, Cybernetics & Informatics*, vol. 8, no. 1, pp. 43–48, 2010.

[16] A. Lavecchia, "Machine-learning approaches in drug discovery: methods and applications," *Drug Discov Today*, vol. 20, no. 3, pp. 318–331, 2015.

[17] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A. Schuffenhauer, "New methods for ligand-based virtual screening: Use of data fusion and machine learning to enhance the effectiveness of similarity searching," *Journal of Chemical Information and Modeling*, vol. 46, no. 2, pp. 462–470, 2006.

[18] Q. U. Ain, A. Aleksandrova, F. D. Roessler, and P. J. Ballester, "Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 5, no. 6, pp. 405–424, 2015.

[19] P. W. Finn and G. M. Morris, "Shape-based similarity searching in chemical databases," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 3, no. 3, pp. 226–241, 2013.

[20] V. Venkatraman, V. I. Pérez-Nueno, L. Mavridis, and D. W. Ritchie, "Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods," *Journal of Chemical Information and Modeling*, vol. 50, no. 12, pp. 2079–2093, 2010.

[21] J. P. Ebejer, G. M. Morris, and C. M. Deane, "Freely available conformer generation methods: How good are they?" *Journal of Chemical Information and Modeling*, 2012.

[22] G. B. McGaughey, R. P. Sheridan, C. I. Bayly, J. C. Culberson, C. Kreatsoulas, S. Lindsley, V. Maiorov, J.-F. Truchon, and W. D. Cornell, "Comparison of topological, shape, and docking methods in virtual screening," *Journal of chemical information and modeling*, vol. 47, no. 4, pp. 1504–1519, 2007.

[23] P. Hall, "A distribution is completely determined by its translated moments," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 62, no. 3, pp. 355–359, sep 1983.

[24] P. J. Ballester, P. W. Finn, and W. G. Richards, "Ultrafast shape recognition: Evaluating a new ligand-based virtual screening technology," *Journal of Molecular Graphics and Modelling*, vol. 27, no. 7, pp. 836–845, 2009.

[25] P. Ripphausen, B. Nisius, and J. Bajorath, "State-of-the-art in ligand-based virtual screening," *Drug Discovery Today*, vol. 16, no. 9-10, pp. 372–376, 2011.

[26] A. Nicholls, G. B. McGaughey, R. P. Sheridan, A. C. Good, G. Warren, M. Mathieu, S. W. Muchmore, S. P. Brown, J. A. Grant, J. A. Haigh, N. Nevins, A. N. Jain, and B. Kelley, "Molecular Shape and Medicinal Chemistry: A Perspective," *Journal of Medicinal Chemistry*, vol. 53, no. 10, pp. 3862–3886, 2010.

[27] P. J. Ballester, "Ultrafast shape recognition: method and applications." *Future medicinal chemistry*, vol. 3, no. 1, pp. 65–78, 2011.

[28] J. A. Grant and B. T. Pickup, "A Gaussian description of molecular shape," *Journal of Physical Chemistry*, vol. 99, no. 11, pp. 3503–3510, 1995.

[29] J. A. Grant, M. A. Gallardo, and B. T. Pickup, "A fast method of molecular shape comaprison: a simple application of a Gaussian description of molecular shape," *J. Comput. Chem.*, vol. 17, no. 14, pp. 1653–1666, 1996.

[30] N. Huang, B. K. Shoichet, and J. J. Irwin, "Benchmarking sets for molecular docking," *Journal of Medicinal Chemistry*, vol. 49, no. 23, pp. 6789–6801, 2006.

[31] M. S. Armstrong, G. M. Morris, P. W. Finn, R. Sharma, and W. G. Richards, "Molecular similarity including chirality," *Journal of Molecular Graphics and Modelling*, vol. 28, no. 4, pp. 368–370, 2009.

[32] G. Landrum *et al.*, "Rdkit: cheminformatics and machine learning software," *RDKIT. ORG*, 2013.

[33] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[34] J. Truchon and C. Bayly, "Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem," *J Chem Inf Model*, vol. 47, 2007.

[35] C. Empereur-mot, H. Guillemain, A. Latouche, J.-F. Zagury, V. Viallon, and M. Montes, "Predictiveness curves in virtual screening," *Journal of Cheminformatics*, vol. 7, no. 1, p. 52, nov 2015.

[36] N. Triballeau, F. Acher, I. Brabet, J. . P. Pin, and H. Bertrand, "Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. application to high-throughput docking on metabotropic glutamate receptor subtype 4," *J Med Chem*, vol. 48, 2005.

[37] D. Reynolds, "Gaussian mixture models," *Encyclopedia of biometrics*, pp. 827–832, 2015.

[38] A. Jahn, G. Hinselmann, N. Fechner, C. Henneges, and A. Zell, "Probabilistic modeling of conformational space for 3D machine learning approaches," *Molecular Informatics*, vol. 29, no. 5, pp. 441–455, 2010.

[39] A. Jahn, L. Rosenbaum, G. Hinselmann, and A. Zell, "4D flexible atom-pairs: An efficient probabilistic conformational space comparison for ligandbased virtual screening," *Journal of Cheminformatics*, vol. 3, no. 7, p. 23, 2011. [Online]. Available: http://www.jcheminf.com/content/3/1/23

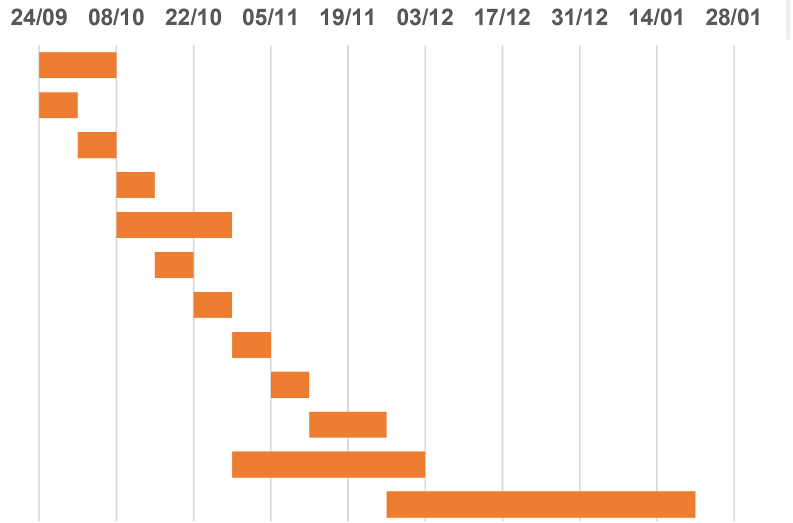
[40] T. Jebara, R. Kondor, A. Howard, K. Bennett, and N. O. Cesa-Bianchi, "Probability Product Kernels," Tech. Rep., 2004.

[41] L. Zhou, W. Ye, B. Wackersreuther, C. Plant, and C. Bohm, "A Pseudometric for Gaussian Mixture Models," in *DBKDA 2017: The Ninth International COngerence on Advances in Databases, Knowledge and Data Applications*, 2017, pp. 37–42.

Appendix A

Dissertation Project Plan

Name	Start Date	End Date	Duration (Days)
Finalise GMM research	24/09/2018	07/10/2018	14
Migrate conformer generation to Spark	24/09/2018	30/09/2018	7
Implement ElectroShape descriptors	01/10/2018	07/10/2018	7
Implement generation of shortened descriptors	08/10/2018	14/10/2018	7
Generate datasets for 6 proteins	08/10/2018	28/10/2018	21
VS: USR	15/10/2018	21/10/2018	7
VS: ElectorShape	22/10/2018	28/10/2018	7
GMM - Supervised	29/10/2018	04/11/2018	7
GMM - Unsupervised	05/11/2018	11/11/2018	7
GMM - similarity	12/11/2018	25/11/2018	14
Results & visualisations	29/10/2018	02/12/2018	35
Dissertation report	26/11/2018	20/01/2019	56



FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Declaration

Plagiarism is defined as "the unacknowledged use, as one's own, of work of another person, whether or not such work has been published, and as may be further elaborated in Faculty or University guidelines" (University Assessment Regulations, 2009, Regulation 39 (b)(i), University of Malta).

I / We*, the undersigned, declare that the [assignment / Assigned Practical Task report / Final Year Project report] submitted is my / our* work, except where acknowledged and referenced.

I / We* understand that the penalties for committing a breach of the regulations include loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Work submitted without this signed declaration will not be corrected, and will be given zero marks.

* Delete as appropriate.

(N. B. If the assignment is meant to be submitted anonymously, please sign this form and submit it to the Departmental Officer separately from the assignment).

Etienne Bonanno
Student Name

Signature

Student Name

Signature

Student Name

Signature

Student Name

Signature

ICS5200
Course Code

Dissertation Progress Report
Title of work submitted

23/09/2018
Date