AFEKA - TEL-AVIV ACADEMIC COLLEGE OF ENGINEERING

M.SC FINAL RESEARCH

M.Sc. of System Engineering

Using Machine Learning Optimizing Pharma Research Discovery Phase

---

Author:
Shy Alon

Supervisor:
Dr. Abraham Yosipof

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science (M.Sc.)

in the

Department of Systems Engineering

July 29, 2017

DECLARATION OF AUTHORSHIP

I, Shy Alon, declare that this research project titled, "Using Machine Learning Optimizing Pharma Research Discovery Phase" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a M.Sc. degree at this college.
- Where any part of this research project has previously been submitted for a degree or any other qualification at this college or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this research project is entirely my own work. I have acknowledged all main sources of help.
- Where the research project is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

“Fortune Favors the prepared mind.”

Louis Pasteur

AFEKA - Tel-Aviv Academic College of Engineering

ABSTRACT

Department of Systems Engineering

Master of Science

Using Machine Learning Optimizing Pharma Research Discovery Phase

by Shy Alon

The process for researching and developing new medicines keeps growing in difficulty and length and the average cost to research and develop each successful drug is estimated by the pharmaceutical companies to be $2.6 billion. This number incorporates the cost (incurred by academic and governmental agencies) of failures of the thousands and sometimes millions of compounds that may be screened and assessed early in the R&D process, only a few of which will ultimately receive approval.

The pre-clinical phase is considered to be so risky and unprofitable that the pharmaceutical industry has abandoned it completely to NGOs and academic institutions which pursue the discovery of new drugs for motives other than profit.

This paper presents the development of a pharmaceutical research decision support system for winnowing thousands of candidate molecular compounds. This work presents the problem analysis, the method definition and the development of an innovative clustering algorithm focused on grouping molecular compounds with similar underlying nature. The results are validated using multiple drug analysis datasets and other datasets with similar characteristics.

CONTENTS

Table of Figures

1. INTRODUCTION

1.1. Motivation

The process of bringing a new drug to market (as described in figure 1) is long and expensive one by all accounts with costs estimated by non-pharma market members in hundreds of millions of US dollars and reported by pharma market members as high as one billion US dollars[1].

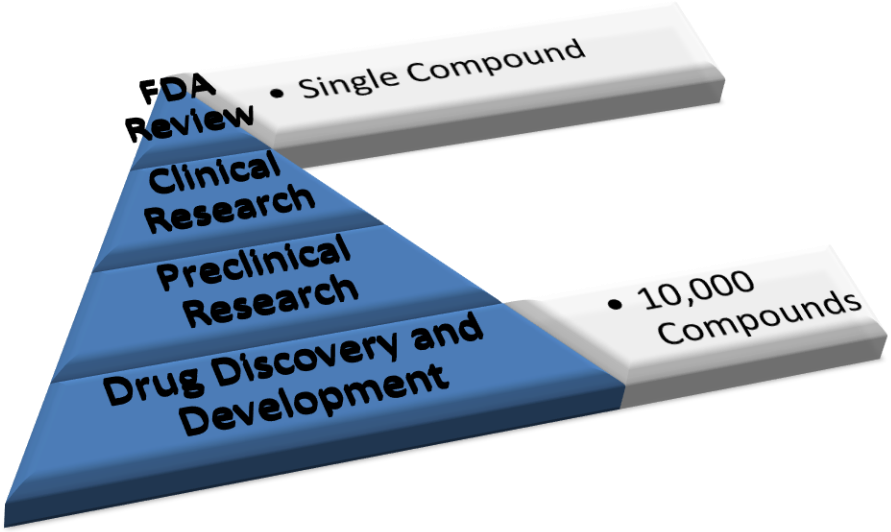

Figure 1 Pharma Compound Funnel

According to the U.S. food and drug administration (FDA) before a new drug hits the market there are 4 required steps:

1. Discovery and development: discovery of new drugs through new insights into a disease process that allow researchers to design a product to stop or reverse the effects of the disease, broad range of tests of molecular compounds to find possible beneficial effects for multiple diseases or existing treatments that have unanticipated effects or new technologies.
   At this stage in the process, many thousands of compounds are potentially candidates for development as a medical treatment. Early testing filters the candidates to a small number of compounds. Once a promising compound is identified experiments are conducted to gather information on how it is absorbed, distributed, metabolized, and excreted; Its potential benefits; the best dosage; the best way to give the drug (such as by mouth or injection); side effects; effects on different groups of people (such as by gender, race, or ethnicity); interaction with other drugs and treatments and its effectiveness as compared with similar drugs.
2. Preclinical Research: Before human trials the compound's toxicity must be ascertained using two types of preclinical research: in vitro (using controlled environment outside of a living organism) and in vivo (using living organisms such as cells and animals). Preclinical studies are limited in scope but must provide detailed information on dosing and toxicity levels, leading to the decision whether the drug should be tested in people.
3. Clinical Research: The studies, or trials, that are conducted on people. Clinical trials vary greatly on scales of risk and process and follow a typical series from early, small-scale, Phase 1 studies to late-stage, large scale, Phase 3 studies.

4. FDA Review: when a drug has indicated from its early tests and preclinical and clinical research that it is safe and effective for its intended use, the developer can file an application to market it.

According to PhRMA (a U.S. based biopharmaceutical research company's consortium) the process for researching and developing new medicines keeps growing in difficulty and length. On average, it takes at least ten years for a new medicine to complete the journey from initial discovery to the marketplace, with clinical trials alone taking six to seven years on average. The average cost to research and develop each successful drug is estimated to be $2.6 billion. This number incorporates the cost (incurred by academic and governmental agencies) of failures of the thousands and sometimes millions of compounds that may be screened and assessed early in the R&D process, only a few of which will ultimately receive approval. The overall probability of clinical success (the likelihood that a drug entering clinical testing will eventually be approved) is estimated to be less than 12%.

A worrying trend for pharmaceutical industry is the continued rise in costs of research and development (as seen in figure 2, based on [2]). In addition to regulation, competition in the global market and reduced government funding the size of the field of potential compounds.

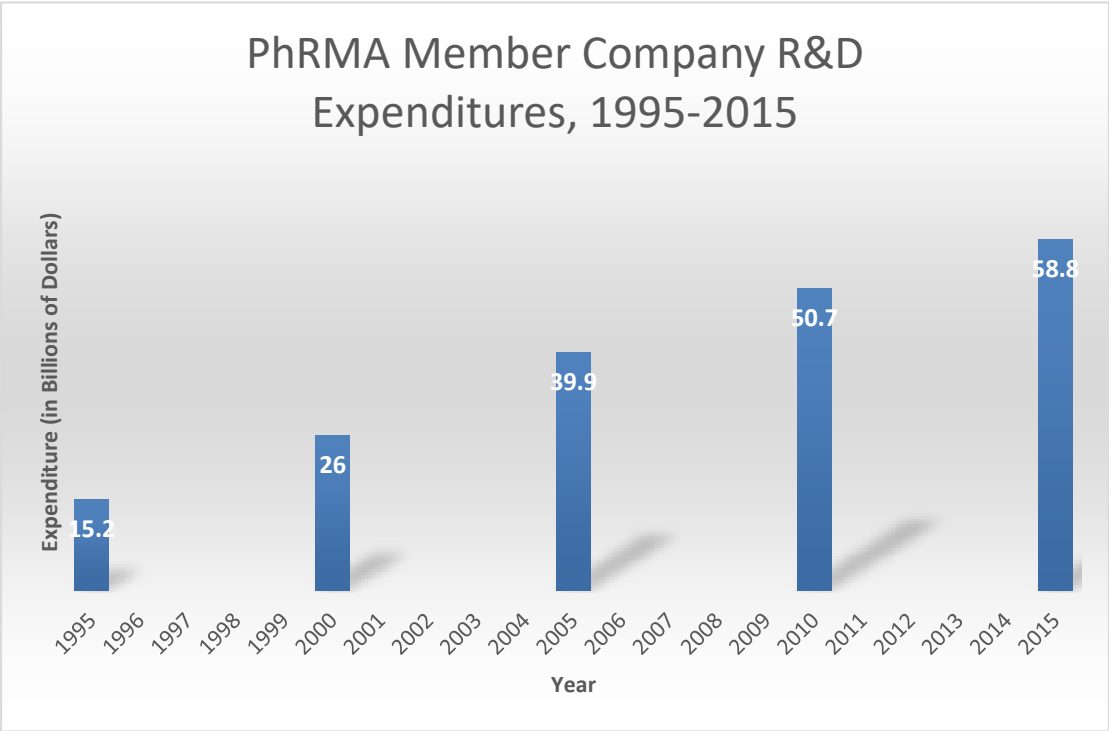

Figure 2 Average PhRMA Member Company R&D Expenditures, 1995-2015

The optimal compound to proceed with to the pre-clinical testing phase, given it even actually exists for a specific effect, is the proverbial needle in a haystack and finding it incurs a significant cost. It is so difficult (and therefore unviable economically) that it is manly funded by academic institutions (universities and public research foundations), governments and philanthropic organizations [2].

According to the Tufts Center for the Study of Drug Development [2] roughly 30% of the total cost of an approved new compound is attributed to the pre-human phase of the research

which includes the drug discovery and the pre-clinical phase. That is a huge potential for savings which can be materialized using the tools and processes described in this paper.

A Nature's Review article [5] presents a model which defines the distinct phases of drug discovery and development starting at the initial stage of "target-to-hit" to the final stage of releasing the drug to the market. The model describes (among other identifiers) the probability of a successful transition from one stage to the next and the phase cost for each project. The model estimates the total cost to achieve one drug launch per year at $1,778 million. per NME launch. It is important to note that this model does not include investments for exploratory discovery research (which as mentioned before is rarely performed in the industry), post-launch expenses or overheads (that is, salaries for employees not engaged in R&D activities but necessary to support the organization).

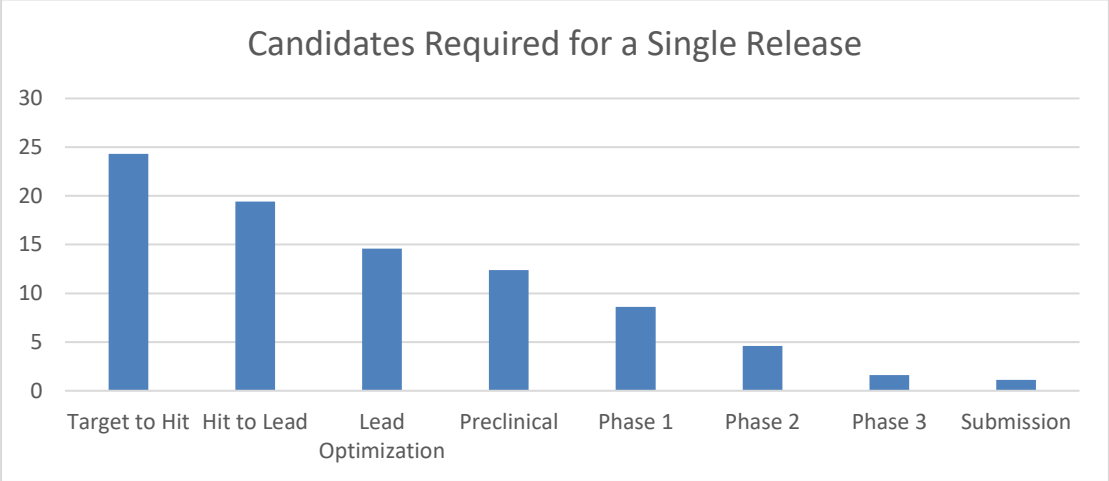

Figure 3 Candidates Required for a Single Release

1.2. Research Structure

The study is comprised of four parts:

Needs and Process Analysis

The research and development process of the pharmaceutical industry discovery phase, specifically comparing candidate compounds, is analyzed.

Data Acquisition and Modeling

A data sample is acquired and preprocessed for the sake of identifying the predictability of potential effect of a compound based on selected feature projection and.

Algorithmic Solution

The algorithm which will perform the classification process is defined and implemented using state of the art environments. For that purpose the specific building blocks and the interactions between those are identified.

Algorithm Validation

The results of the algorithm are analyzed and the algorithm is evaluated for correctness, efficiency and industrial applications.

1.3. The Economy of Discovery

As mentioned above the industry has abandoned the discovery phase because it is not economically viable. As it is the cost of a new drug is estimated at 5.5 billion USD [6] and only two out of ten new drugs cover their incurred R&D expenses. The discovery phase is

currently a liability that the pharmaceutical companies cannot accept until its effectiveness is dramatically improved.
The discovery phase costs vary wildly but it is safe to assume that every test for a candidate compound costs roughly 10000 USD and right now it takes a corpus of 125000 candidates to produce 75 viable pre-clinical trials candidates [6]. Out of these there is a good chance one will make it to production. This means that under present conditions finding a sufficiently large viable candidate group would cost 1.25 USD billion.
In order to make the discovery phase economically viable we seek to reduce the cost by one order of magnitude – which means automatically and cheaply filter most of the candidates and leave at most one tenth of the candidate population – without losing viable candidate.

1.4. State of the Art

As displayed in recent research [7] applying Computer-Aided Drug Design (CADD) strategies provides cost savings for drug research and development programs. For example, a mathematical model called Quantitative Structure-Activity Relationships (QSAR) is taking advantage of the availability of vast chemical databases with abundant bioactivity data, such as the one used in this research. The explosive growth of such data provides a good opportunity for large scale, Big Data mathematical modeling across diverse pharmaceutically relevant targets. Resulting models have become valuable tools for identifying novel molecular probes and potential leads for drug discovery but not in a sufficient capacity.

1.5. Research Hypothesis

To make the task of conducting the discovery phase cost effective the field of potential compounds to test needs to be narrowed significantly. Without trimming the number of potential candidates is too big to scour. Without a definite methodology which would bring down the time required to find a single worthy candidate by at least two orders of magnitude that task would remain a non-profit task.

This research presents a systems' thinking based solution for winnowing the field of candidates using techniques from the fields of machine learning, special analysis and probability theory. The resulting system increases the probability of selecting a valid candidate, if such a compound exists, from random selection to a probability of above 0.9. For example, the following figure (Figure 4 Mapping by Serotonin) shows a single mapping of compounds tagged by their effect on Serotonin. In that mapping selecting randomly a compound which is nearest to a tagged compound has a 0.935 to have the same tag – meaning that they share the same effect.

The system assumes that there are underlying properties of different compounds with the same attributes. Those underlying properties can be used to identify similar compounds to compounds which are known to have desirable properties and therefore reduce the immense field of potential compounds to be tested to a much more cost effective smaller field.

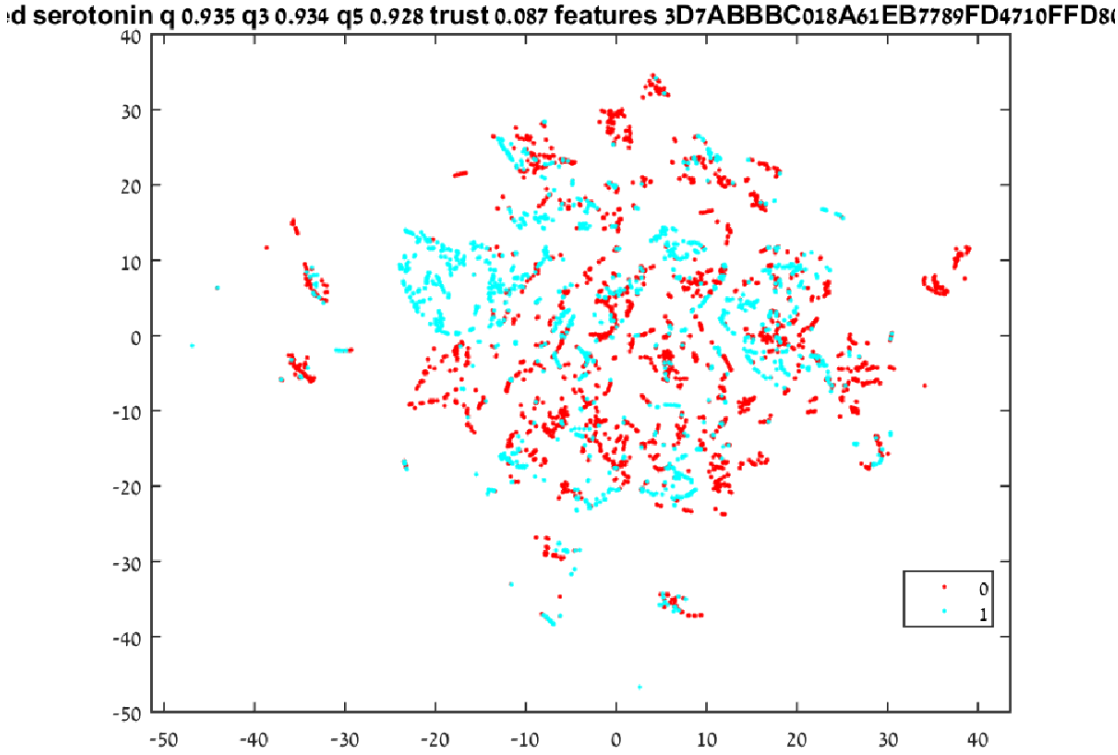

Figure 4 Mapping by Serotonin

The research hypothesis in general is that by using a combination of feature selection algorithms, partially tagged multi-dimensional data and dimensionality reduction algorithms a powerful classification algorithm can be implemented and applied to the data.

1.6. Application

The research hypothesis discussed in this paper, in the context of pharmaceutical industry research, is that it is possible to optimize a feature selection for a dimensionality reduction algorithm so that the compounds selected according to the distance from tagged compounds present the best candidate for pre-clinical trials. This is accomplished by building an algorithmic candidate and testing it on a rich enough database.

The target application focuses on the target class of G Protein-coupled Receptors (GPCRs), a group of molecular targets which has a tremendous profit potential to the pharmaceutical industry. The number of GPCRs in human genome from five main families (glutamate, rhodopsin, adhesion, frizzled/taste2, and secretin) had been estimated at over 800 (or ~ 4% of the entire protein-coding genome) by the year 2003. Realistically the number of GPCRs is much higher now due to the known existence of alternatively spliced variants and editing isoforms of GPCRs. In addition, GPCRs with unknown functions (i.e., lack of known natural transmitters), called “orphan” GPCRs, account for a large portion of newly identified GPCRs.

GPCRs have a major impact on drug discovery. It is estimated that while GPCRs family members are only around 3% of known molecular targets between one-third [8] and half [9] of currently marketed drugs target them.

2. LITERATURE REVIEW

2.1. Pharmaceutical Research Challenge

The challenges facing pharmaceutical research were discussed in many papers published by PhRMA, an institution representing many American pharma companies. In their industry profile for 2017 PhRMA address the "The Lengthly, Costly and Uncertain Biopharmaceutical Research and Development Process" and its negative impact on the industry.

Independent researchers such as Paul SM, Mytelka et al [10] seem to accept the premise and set the focus on the productivity of the process. According to said publication the pharmaceutical industry is suffering from major losses of revenue due to patent expirations, increasingly cost-constrained healthcare systems and ever stricter regulatory requirements. They assert that the key to tackling these challenges such issues pose to both the is to remain within budget limits and substantially increase the quantity and quality of innovative, cost-effective new medicines, contrary to contemporary trends. They present a detailed analysis based on comprehensive, recent, industry-wide data and identify the relative contributions of each of the steps in the drug discovery and development process to overall R&D productivity.

2.2. Machine Learning in Pharmaceutical Research

The secretive nature of pharmaceutical research severely limits the number of publications on the issue. There are a few however.

In 2011 Chapel Hill (University of North Carolina) published "THE QSAROME OF THE RECEPTOROME" [7] which is a research paper dealing with modeling sets of multiple receptors and its relation with GPCRs. It uses a combinatorial QSAR (Quantitative structure–activity relationship) framework heavily reliant on a Distance Weighted Discrimination (similar to the results of t-SNE) as a score generator and a cost/benefit ratio applied using Decision Trees.

In Application of Predictive QSAR Models to Database Mining Alexander Tropsha discusses a drug discovery strategy that employs variable selection quantitative structure-activity relationship (QSAR) models for chemical database mining.

Same as in the aforementioned publishing the approach starts with the development of rigorously validated QSAR models obtained with the variable selection k nearest neighbor (kNN) method (as opposed to the Decision Trees method mentioned in previous article). Model validation is based on several statistical criteria, including the randomization of the target property (Y-randomization), independent assessment of the training set model's predictive power using external test sets, and the establishment of the model's applicability domain. The most successful models were employed in database mining concurrently and the specific biological activity (characteristic of the training set compounds) of external database entries found to be within a predefined similarity threshold of the training set molecules were predicted based on the validated QSAR models using the applicability domain criteria. Compounds judged to have high predicted activities by all or most of all models are considered as consensus hits.

For the purpose of validating the method the 10 best models were applied to mining chemical databases, and 22 compounds were selected as consensus hits. Nine compounds were synthesized and tested at the NIH Epilepsy Branch, Rockville, MD using the same biological test that was employed to assess the anticonvulsant activity of the training set compounds; of these nine, four were exact database hits and five were derived from the hits by minor chemical modifications. Seven of these nine compounds were confirmed to be active,

indicating an exceptionally high hit rate. The approach described in this report can be used as a general rational drug discovery tool.

3. METHODOLOGY

3.1. Needs

The needs of the pharmaceutical industry for the discovery phase are:

1. The process needs to accept a list of tagged compounds with their respective features.
2. The process needs to accept a significantly larger list of candidate untagged compounds with the exact same features.
3. The process needs to provide a short list of the candidate compounds which are the most likely to have similar results as the tagged compounds.

Compound Features

The features are common industry characteristics of compounds. For example, a list of features used in one of the experiments included:

5-HT1A, ALogP98, Apol, Formal Charge, Coord Dimension, Is Chiral, LogD, Molecular Weight, Molecular Mass, Molecular Solubility, VSA Total Area, HBA Count, HBD Count, NPlusO Count, Number of Atoms, Number of Bonds, Number of Hydrogens, Number of Explicit Hydrogens, Number of Explicit Atoms, Number of Explicit Bonds, Number of Positive Atoms, Number of Negative Atoms, Number of Spiro Atoms, Number of Bridge Head Atoms, Number of Ring Bonds, Number of Rotatable Bonds, Number of Aromatic Bonds, Number of Bridge Bonds, Number of Rings, Number of Aromatic Rings, Number of Ring Assemblies, Number of Rings3, Number of Rings4, Number of Rings5, Number of Rings6, Number of Rings7, Number of Rings8, Number of Rings9Plus, Number of Chains, Number of Chain Assemblies, Number of Fragments, Number of Complexed Fragments, Number of Metal Atoms, Number of SGroups, Number of Super atoms, Number of Isotopes, Number of Custom Data, Number of Pi Bonds, Number of Repeat Units, Number of V3000Templates, Number of R Group Fragments, Number of Stereo Atoms, Number of Stereo Bonds, Number of Single Bonds, Number of Double Bonds, Number of Triple Bonds, Number of Aliphatic Single Bonds, Number of Aliphatic Double Bonds, Number of Unknown Stereo Atoms, Number of Unknown Stereo Bonds, Number of Dative Bonds, Number of Hydrogen Bonds, Number of Allene Stereo Centers, Number of Atropisomer Centers and Number of Axial Stereo.

The large number of features gives the feature selection process enough space to search in and the number of samples allows the dimensionality reduction to converge on the axis of the selected dimensions.

Tagging

The tagged compounds carry the information of the effect of the compound. The compounds in the database are tagged with the organic compound they effect which are:

Dopamine

Dopamine is an organic chemical that is excreted by individual cells and plays several important roles in both the brain and the body. In the brain, dopamine functions as a neurotransmitter which is a chemical released by nerve cells to send signals to other nerve cells. One of the few distinct dopamine pathways in the brain plays a major role in reward-motivated behavior and most types of rewards (winning a bet or sugar rush) increase the level of dopamine in the brain. Many addictive drugs increase dopamine neuronal activity.

Dopamine functions outside the central nervous system primarily as a local chemical messenger. It inhibits norepinephrine release in blood vessels; it increases sodium excretion and urine output in the kidneys and reduces insulin production in the pancreas. Dopamine reduces gastrointestinal motility in the digestive system, and reduces the activity of lymphocytes in the immune system

The Dopamine system is associated with several important diseases of the nervous system and some of the key medicinal compounds used to treat them work by altering the effects of dopamine. Parkinson's disease, schizophrenia and attention deficit hyperactivity disorder (ADHD) are associated with unregulated dopamine activity.

Adrenoceptors

Adrenergic receptors (or adrenoceptors) are a class of cell receptors that are targets of the catecholamines, especially norepinephrine (noradrenaline) and epinephrine (adrenaline). An catecholamine binding to the receptor will generally stimulate the sympathetic nervous system which causes includes dilating the pupils, increasing heart rate, mobilizing energy, and diverting blood flow from non-essential organs to skeletal muscle (a fight or flight response).

Histamine

Histamine is an organic compound taking part in local immune responses as well as regulating physiological function in the gastro intestines and acting as a neurotransmitter for the uterus and is involved in the inflammatory response. Histamine increases the permeability of the capillaries to white blood cells and some proteins, to allow them to engage pathogens in the infected tissues and is produced by basophils and by mast cells found in nearby connective tissues.

Muscarinic

Muscarinic acetylcholine receptors are acetylcholine receptors in the cell membranes of certain neurons and other cells. Muscarinic acetylcholine receptors act as the main end-receptor stimulated by acetylcholine released from postganglionic fibers in the parasympathetic nervous system.

Muscarinic receptors are more sensitive to muscarine than to nicotine and many drugs and other substances (for example pilocarpine and scopolamine) manipulate these two distinct receptors by acting as selective agonists or antagonists.

Serotonin

Serotonin is a monoamine neurotransmitter that can be biochemically derived from tryptophan. Serotonin is primarily found in the gastrointestinal tract (GI tract), blood platelets, and the central nervous system (CNS) of animals, including humans. It is generally accepted as a contributor to feelings of well-being and happiness.

Serotonin is used to regulate intestinal movements as well as regulate moods, the appetite and sleep. Serotonin effects cognitive functions, including memory and learning. Modulation of serotonin at synapses is thought to be a major action of several classes of pharmacological antidepressants.

The system is required to apply the tags mentioned above with as high a probability as possible on the untagged compounds.

3.2. Operational Requirements

The system, however complex in its implementation, has a very simple operational requirement to fulfill which is:

The system shall analyze large corpus of compound data and generate recommendations with regard to the applicability of analyzed compounds as candidates for preclinical trials.

3.3. Context Analysis

Corpus

The system operates within the context of a pharmaceutical company with a pre-existing corpus of analyzed compounds and a pool of potential compounds from which to select candidates for preclinical trials. The corpus of the compounds must be in a size (number of compounds * number of features) which will both provide the algorithm with enough data to provide meaningful results and small enough for the algorithm to converge on a solution in a frame of time which allow the researchers to use it as a decision support system.

Timeliness

The system is to be used as a non-real-time support system. This means it will perform its function in parallel to the function of an active research and development team and augment its capabilities with decision support capabilities.

Personnel and Infrastructure Qualifications

The system is to be used by specifically trained personnel so it has no strict user experience limitations on the learning curve it incurs.

The system, being software based, is agnostic to the hardware it runs on and can be executed in various environments such as cloud environment or on-premise environment.

3.4. System Model

The bottom level capabilities of the system (described graphically in Figure 5) are as follows:

Dimensionality Reduction

In the context of our system the dimensionality reduction serves dual purpose. In addition to making the data more comprehensible it allows for the distance between compounds to be processed in a way which puts emphasis on closeness, which I critical to the system which uses a nearest neighbor as a selection metric.

Quality Factor

The ability to attribute a certain mapping with a quality factor is critical to determine whether the results can be counted on. There are two processes which need to be measured for quality for the process to be successful:

1. Clustering: this is measured by the relative change in the distances between points – points ***a*** and ***b*** which were closer together than points ***c*** and ***d*** in the original space should be closer in the mapped lower dimension space.
2. Classification: This is measured by the probability of a tagged point having a nearest neighbor with the same tag.

Progressive Filtering

The projection created for certain sets of features should be progressively filtered so better and better feature sets are used.

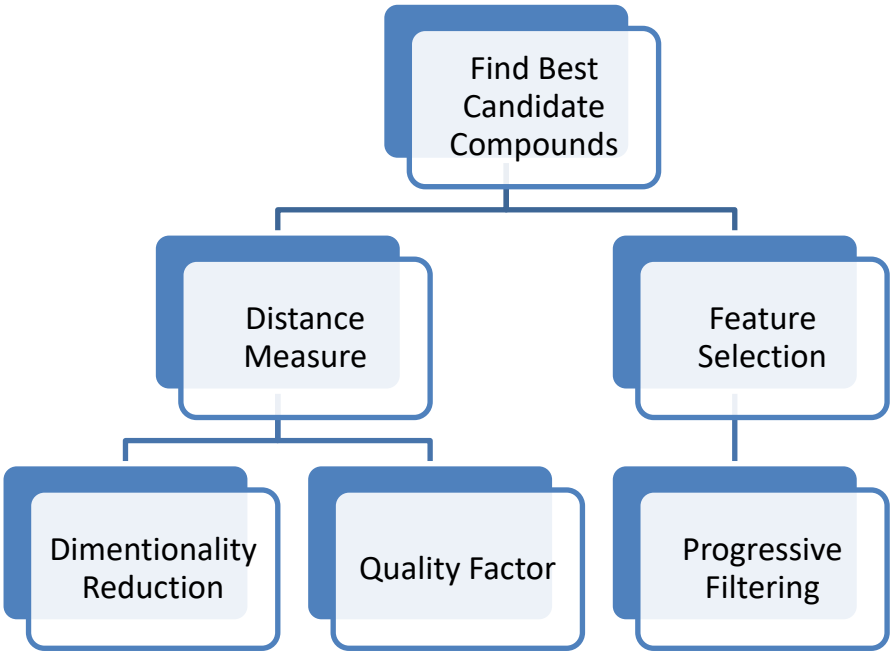

Figure 5 Capabilities

The system needs to be able to measure the distance effectively and for that purpose needs to intelligently reduce the dimensions of the data space (as opposed to simply assuming equal weight on each dimension) and measure the quality of the dimensionality reduction by using clustering algorithm on a-priori tagged data.

The system needs to select the tested features of the data out of a large field of diverse features. For that purpose, the feature selection mechanism needs to be adaptive and to be able to progressively improve while allowing for and fixing mistakes.

The designed decision support system is currently demonstrated using a combination of MATLAB (preforming pre-processing and the enveloping optimization process) and C++ code performing the dimensionality reduction process.

As mentioned above the system's goal is to cluster all potential compounds with the selected features in the way which will allow us to assume in the highest probability that the neighbor of a compound with desirable attributes is highly likely to share those attributes and the underlying assumption is that inside the data there are features which are pertinent to the classification of the compounds into classes of desired effect.

The purpose behind the tool is to identify potentially effective compounds using previously known information about a small number of compounds. For example, if we can tag a few of the compounds we can – by association – estimate with a high probability of success whether the compounds near them in the resulting map (such as in Figure 6 Compounds grouped by bitterness) have high effectiveness potential.

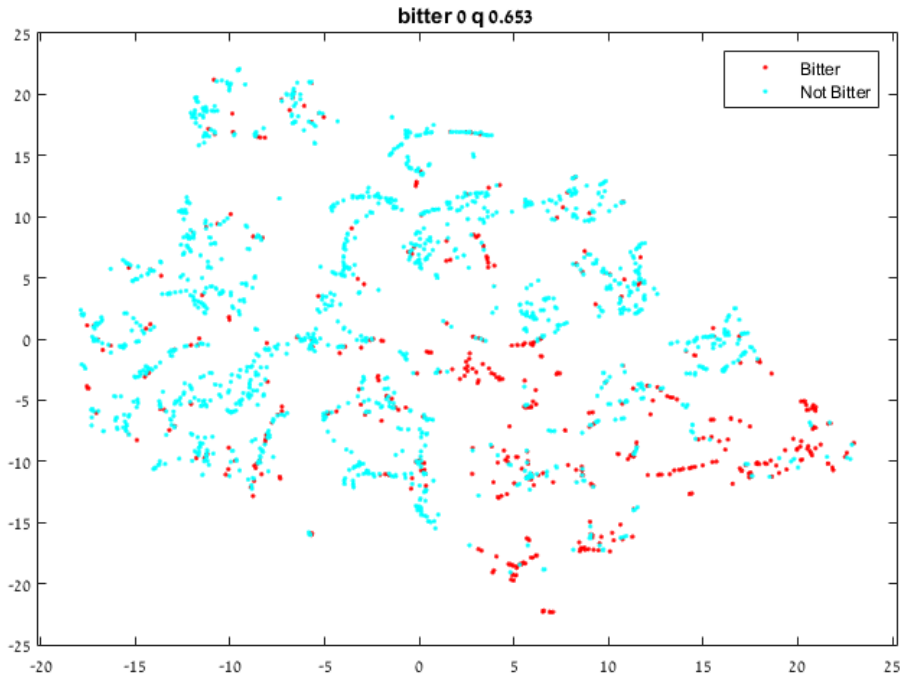

Figure 6 Compounds grouped by bitterness

Following is a graphical depiction (Figure 7 t-SNE Optimization Algorithm) of the system meeting the requirements described above. The system is comprised of 5 functional parts:

1. The input database which supplies the raw data to the system – the feature information about the all compounds and the tagging information of previously tagged compounds.
2. The feature selection mechanism which (at the beginning of the process) generates a rich enough population of feature selection combination and in subsequent steps uses Genetic Algorithm to create the next generation.
3. The dimensionality reduction process which reduces the multi-dimensional feature space into a two-dimensional space. The process is one which attributes more weight to close distances than to long distances and therefore prioritizes keeping neighbors together over preserving ratios of long distances.
4. A fitness function calculation using the tagged data as a measure of success. Even though a few metrics are being used for measurement the fitness function relies on the fraction of known data points with their closest known data point sharing the same tag.
5. Model selection and generation using genetic algorithms to create ever improving generations until a quality condition is satisfied or the maximal number of iterations is reached.

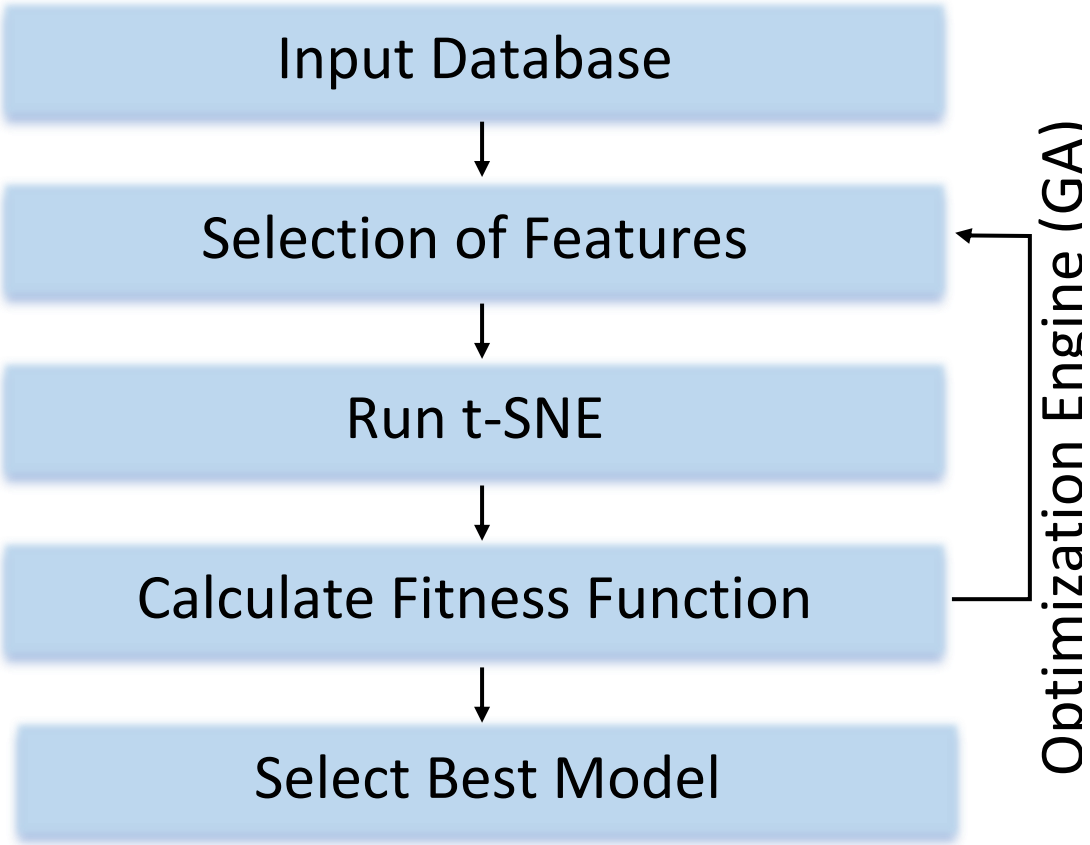

Figure 7 t-SNE Optimization Algorithm

Input Data

The first stage consists of data input. The system accepts digital comma separated value files created using specifically prepared data set (compiled from private data sources) and, for some, classifications into their compound tagging group.

Feature Selection

Under the assumption that some of the features carry more information than others for the snapshot of the world the data represents the goal of the algorithm is to select the optimal feature set. The selected feature selection optimization algorithm was Genetic Algorithm.

Genetic Algorithm

In optimization theory, ***genetic algorithm*** is a natural selection based method for solving both constrained and unconstrained optimization problems. The genetic algorithm iteratively modifies a population of possible individual solutions. At each iteration, the genetic algorithm selects individuals from the current population according to certain selection, mutation and cross over rules (addressed below) and uses them as parents to produce the children of the next generation. Each and every successive generations brings the population closer to an optimal solution. Genetic algorithms can be applied to a variety of optimization problems that are not well suited for standard optimization algorithms, including problems in which the objective function is discontinuous, nondifferentiable, stochastic, or highly nonlinear.

A genetic algorithm uses three main types of generation rules at each step to propagate the next generation from a current population:

1. **Selection** rules select the individuals, called parents, that contribute to the population at the next generation.
2. **Crossover** rules combine two parents to form children for the next generation.
3. **Mutation** rules apply random changes to individual parents to form children.

In the context of the compound selection system, as the basis for a genetic algorithm the initial population is 20 sets of randomly selected features. From each generation, the next generation is created as follows:

1. The fittest 30% are carried as they are to the next generation
2. Mutations of the fittest 20% are generated into the next generation
3. Crosses of the fittest 50% are generated into the next generation.

After the process stops improving (given at least 3 iterations without improvement) it's assumed that the optimal feature set has been found. Otherwise the process will stop at the maximal number of iterations (which is set to 128).

Dimensionality Reduction

t-distributed stochastic neighbor embedding (t-SNE) is a nonlinear dimensionality reduction algorithm which has been chosen because as opposed to more commonly used dimensionality reduction algorithms (such as PCA) which are mainly concerned with preserving large pairwise distances t-SNE maintains structure by putting more weight on local distances.

t-SNE

t-distributed stochastic neighbor embedding (t-SNE) is a nonlinear dimensionality reduction algorithm that is especially designed for embedding high-dimensional data into a space of two or three dimensions, which can then be visualized in a scatter plot. For the purpose of clustering it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points.

The t-SNE algorithm has two main stages:

1. t-SNE constructs a probability distribution over pairs of high-dimensional objects in a way that makes similar objects have a high probability of being selected and dissimilar points have an extremely small probability of being selected.
2. t-SNE defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the Kullback–Leibler divergence (a measure of how one probability distribution diverges from a second expected probability distribution) between the two distributions with respect to the locations of the points in the map.

In the domain of compound tagging, it is critical that the process of preserving large distances (which carry little data) will not affect preserving local (short) distances since the shortest distance estimation along the selected features is what determines the nearest neighbor and for our purpose of classifying untagged samples this is the most important quality.

Fitness Calculation

The fitness of a feature set (representing the quality of the classification algorithm) has been measured using multiple metrics including the nearest neighbor being of the same class, two out of three nearest neighbors and three out of five nearest neighbors. All metrics behave in a similar manner so finally the fitness score is determined by the percentage of tagged samples having a nearest tagged neighbor with the same tag.

Because there are datasets where the target tag is a very significant minority simply applying nearest neighbor counting as a measure of quality can create a false high quality measures. For this reason only the positively tagged (the tag which we are interested in finding) are taken into consideration and not the general population.

The trust metric (representing the quality of the clustering algorithm) indicates that the untagged neighbors are were also neighbors in the higher dimension space. This indicates how much the structure of the data has persisted through the dimensional reduction.

Model Selection

The model with the best fitness score is selected for a dataset. For a specific tag the untagged compounds with the highest percentage of neighbors that share that tag are identified as the best candidate compounds for preliminary pre-clinical tests

4. RESULTS

4.1. Testing

Methodology

In order to test the system smaller real-world datasets, from the pharmaceutical domain as well as from the financial domain, were used. These data sets are simple enough and small enough to be able to visually determine the success of the system in separating the dataset in accordance with the compounds' tagging.

Since the data is separate from the tags and the model is not persistent there is no need to perform cross validation and we can use the entire span of the data to generate the results.

Image Title

In all of the following example the images are presented with their original titles which contain the following information:

Name:	The name of the dataset.
Separated feature:	The name of the sought-after tag.
Single neighbor quality:	The quality calculated by single nearest neighbor.
Three neighbors quality:	The quality calculated by two out of three nearest neighbors.
Five neighbors quality:	The quality calculated by three out of five nearest neighbors.
Trust level:	The clustering quality metric.
Selected features	The selected features encoded in hexadecimal string.

The features are encoded binarily and displayed in hexadecimal form with each letter representing 4 features. For example if we have the following 10 features 0110101011 they will be represented as 1AB.

Random Dataset

For the sake of completeness, a random dataset was generated (500 samples of 24 features tagged 0 and 1) and analyzed by the system. Since the system employs intelligent feature selection it was expected that the resulting map, utilized by the features best fitting the tags, would result in a slightly better than random quality score.

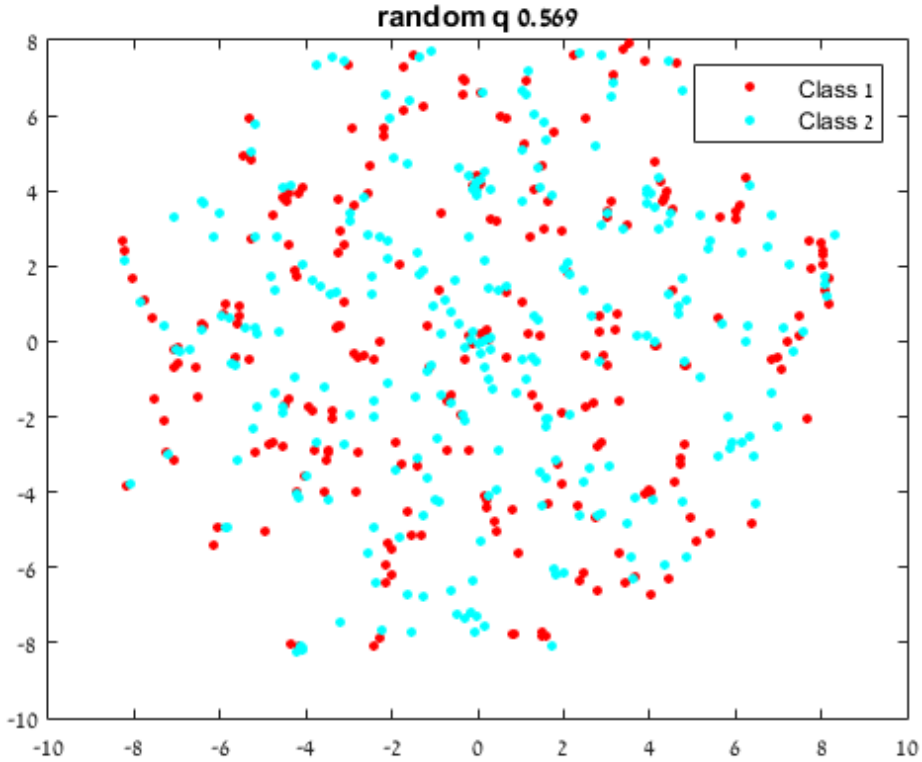

Figure 6 random sample test

As seen in in the resulting projection map (Figure 6) the quality is 0.540 – meaning that for any sample the probability of having the neighbor with the same tag is 0.54, which is as expected.

Financial Ratios Dataset

The financial ratios database contains a dataset representing 500 publicly traded corporations with 41 financial real or integer indicators (such as capital to debt ratio). The tags are the estimated danger of the corporation filing for bankruptcy.

The results (Figure 7) show a very clean separation with more than 0.93 hits. As mentioned above that means that for any untagged corporation which we would guess its chances to file for bankruptcy by its nearest neighbor we would be correct 0.93 of the times.

Bitterness Dataset

The bitterness database consists of 2074 compounds with 19 features and their bitterness quality. We can see in the result (Figure 8) that there doesn't appear to be a very strong correlation between the features and the bitterness level even at the optimal feature selection and only 0.64 of the bitter compounds were mapped near a nearest neighbor that is also tagged as bitter.

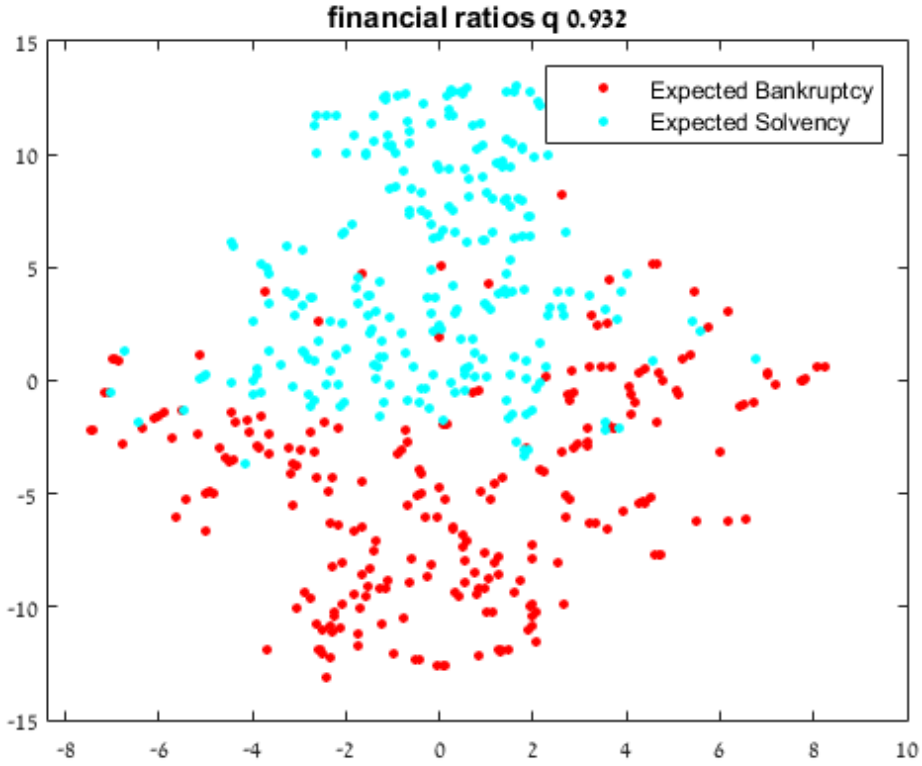

Figure 7 Financial Ratios Results

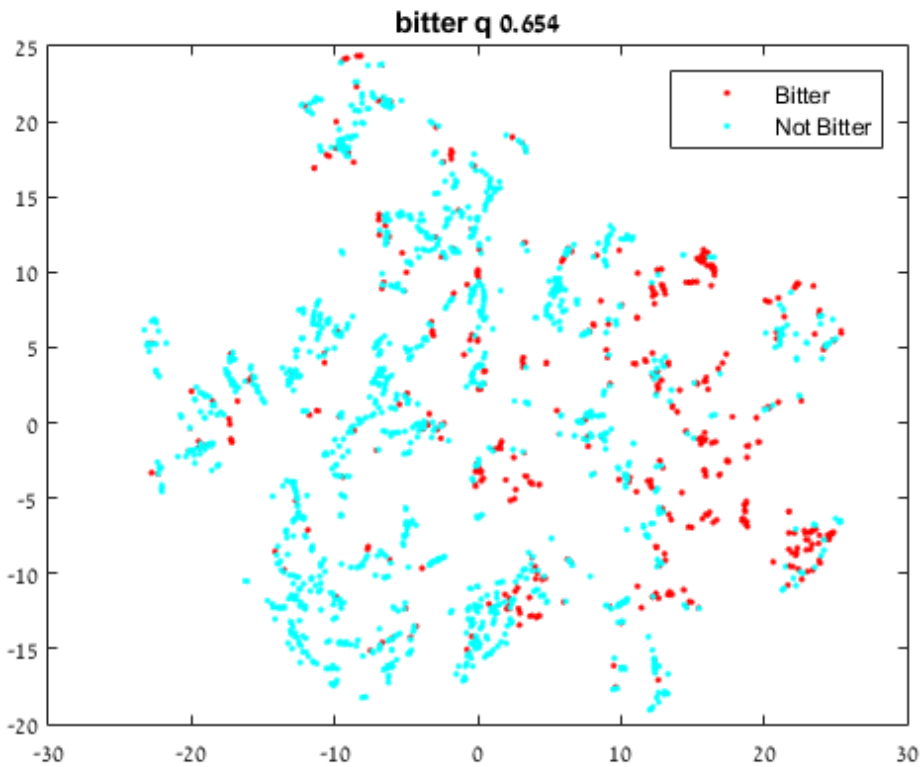

Figure 8 Bitterness Results

Qualitative Financial Ratios Dataset

The qualitative financial ratios database contains a dataset representing 250 publicly traded corporations with 19 binary indicators (such as has capital to debt ratio below a certain level). The tags are the estimated danger of the corporation filing for bankruptcy.

The results (Figure 9) show a nearly perfect separation. Since the algorithm tests the quality for the positively tagged samples the single expected solvency sample near the expected bankruptcy cluster doesn't change the result.

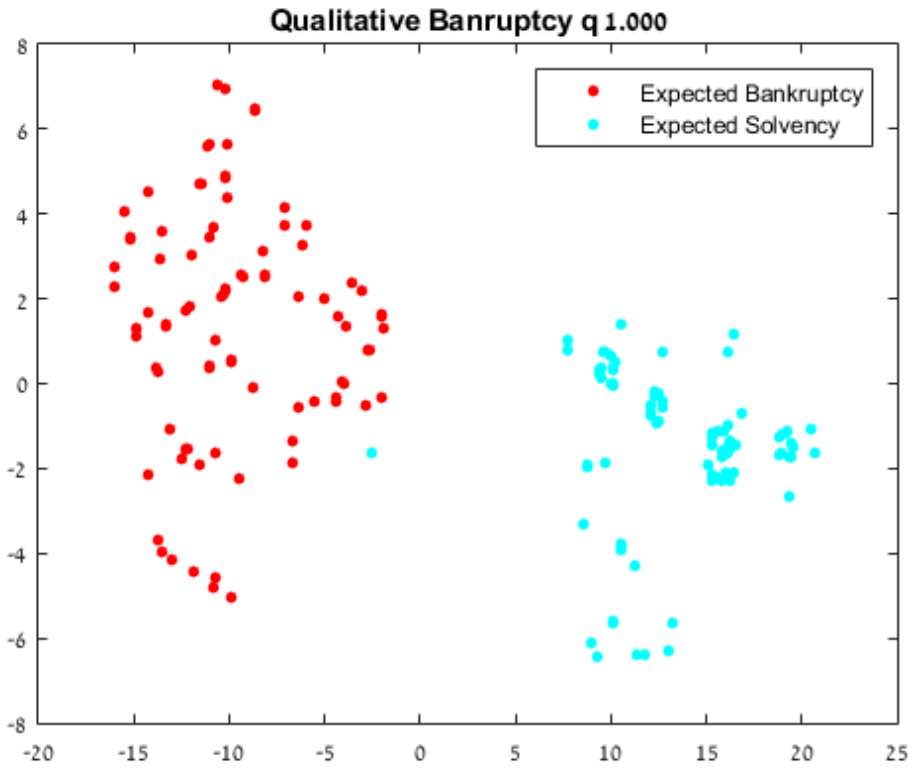

Figure 9 Qualitative Financial Ratios

4.2. Final Results

Database Creation

The compound database is the result of merging 5 databases together. Those databases contained compounds tagged for their effects on Dopamine, Histamine, Serotonin, Adrenoceptor and Muscarinic receptors.

The databases were merged exclusively – meaning that any compound which appeared in more than a single database was discarded. All features were shared were shared by the databases so no features had to be dropped from the resulting database. The result database contains 7276 compounds with 135 features.

System Application

The system analyses the database separately for each tag – separating the samples for positively tagged (has the desired effect) and negatively tagged (have a different effect) for each tag type before executing the classification algorithm.

Results Analysis

Dopamine

The results for Dopamine (Figure 10) are as follows:

Name: Dopamine.

Single neighbor quality: 0.965.

Three neighbors quality: 0.956.

Five neighbors quality: 0.961.

Trust level: 0.087.

The low trust level demonstrates the high level of the algorithm aggressiveness – compounds which were close in the high dimensional space were not close in the resulting mapping.

The high-quality level though indicates that the system was successful in clustering the similarly tagged compounds together.

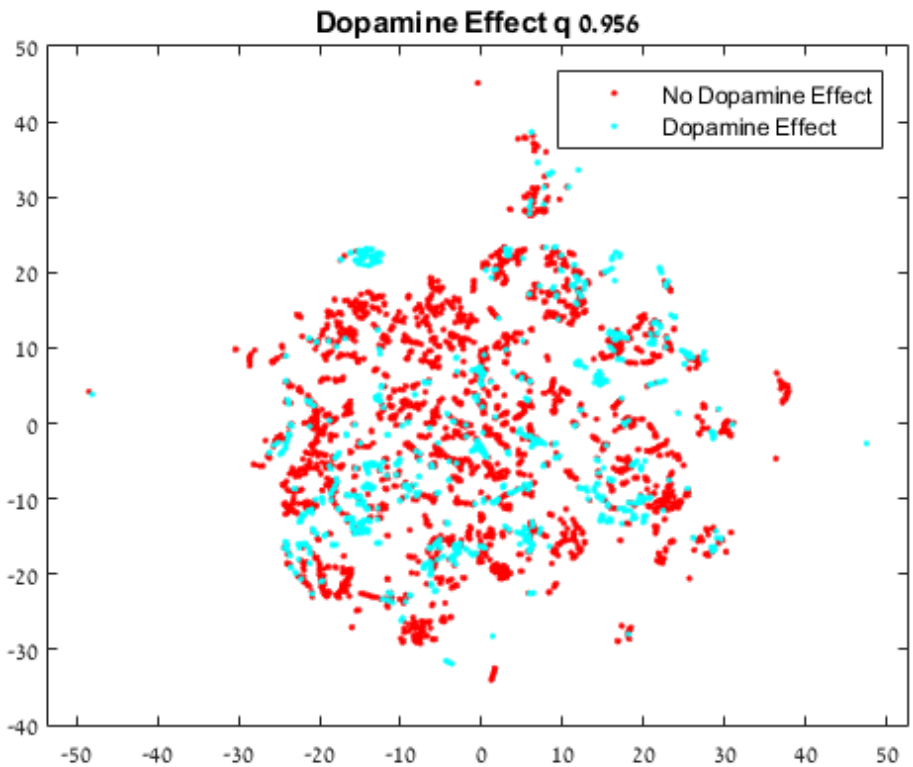

Figure 10 Dopamine Results

Adrenoceptors

The results for Adrenoceptors (Figure 11) are as follows:

Name: Adrenoceptors.

Single neighbor quality: 0.983.

Three neighbors quality: 0.985.

Five neighbors quality: 0.984.

Trust level: 0.077.

Much like the Dopamine results the low trust level demonstrates the high level of the algorithm aggressiveness and also like the Dopamine the even higher quality level though indicates that the system was successful in clustering the similarly tagged compounds together.

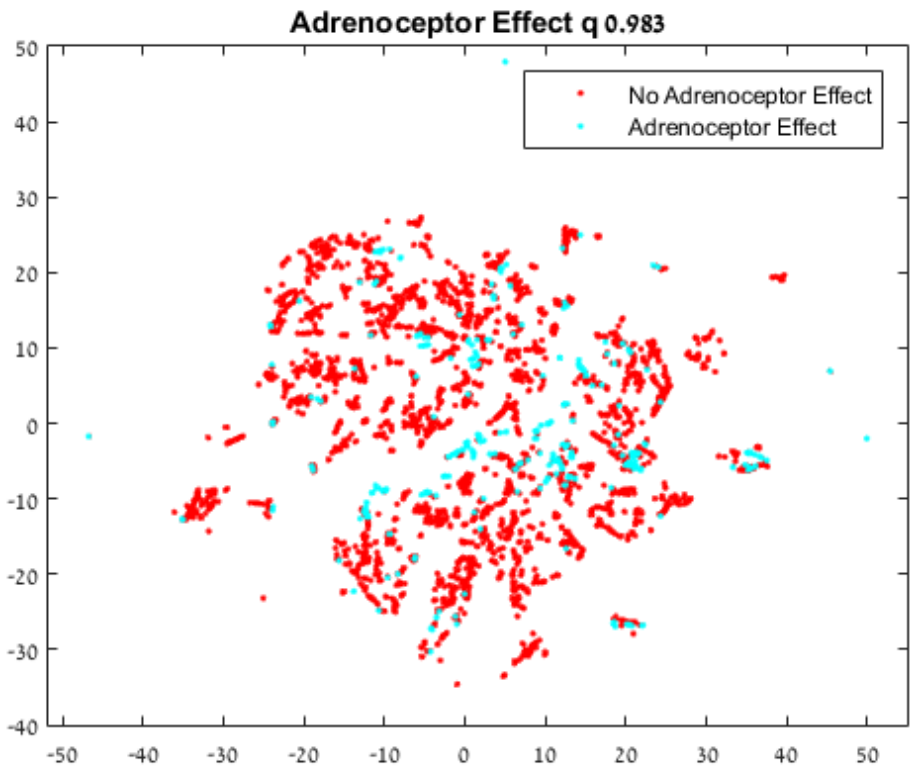

Figure 11 Adrenoceptor Results

Histamine

The results for Histamine (Figure 12) are as follows:

Name: Histamine.

Single neighbor quality: 0.967.

Three neighbors quality: 0.963.

Five neighbors quality: 0.964.

Trust level: 0.078.

Tagging Histamine compounds shows similar results to tagging Dopamine.

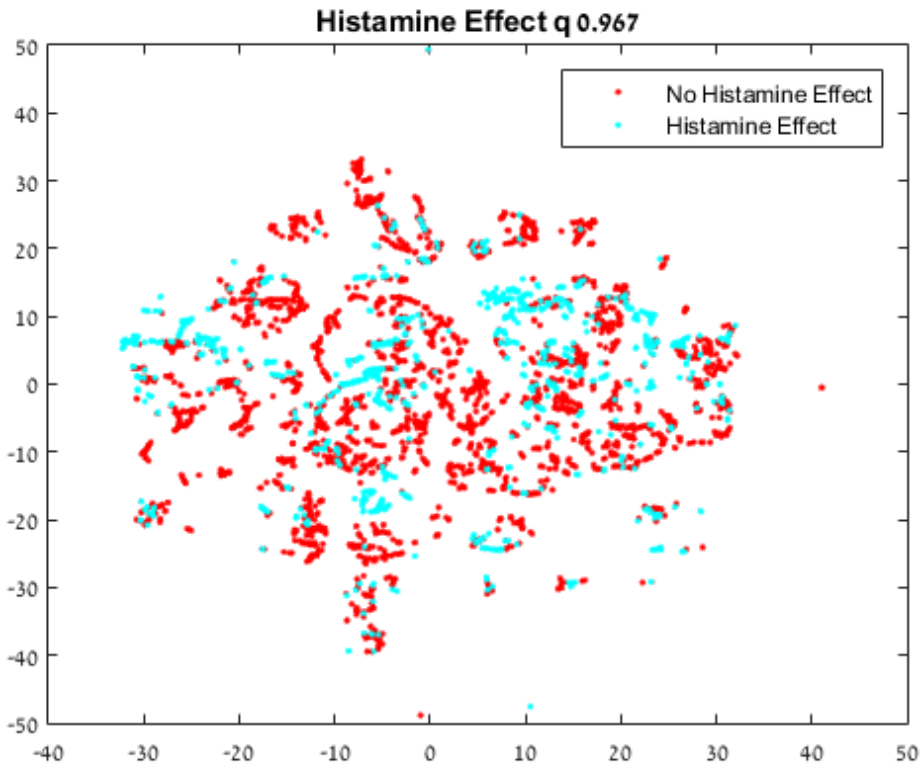

Figure 12 Histamine Results

Muscarinic

The results for Muscarinic (Figure 13) are as follows:

Name:	Muscarinic.
Single neighbor quality:	0.993.
Three neighbors quality:	0.993.
Five neighbors quality:	0.993.
Trust level:	0.076.

Tagging Muscarinic compounds shows excellent results. Similar to the test performed on qualitative financial ratios the scope of the map the visual indication seems to contradict the quality figures but it bears reiterating that the image represents more than 7000 compounds. Taking the single compound in the middle (for example) and zooming into the map we can see (Figure 14) that it is in fact a cluster of two compounds bunched together in a manner which makes it seem like a single compound in a compressed space.

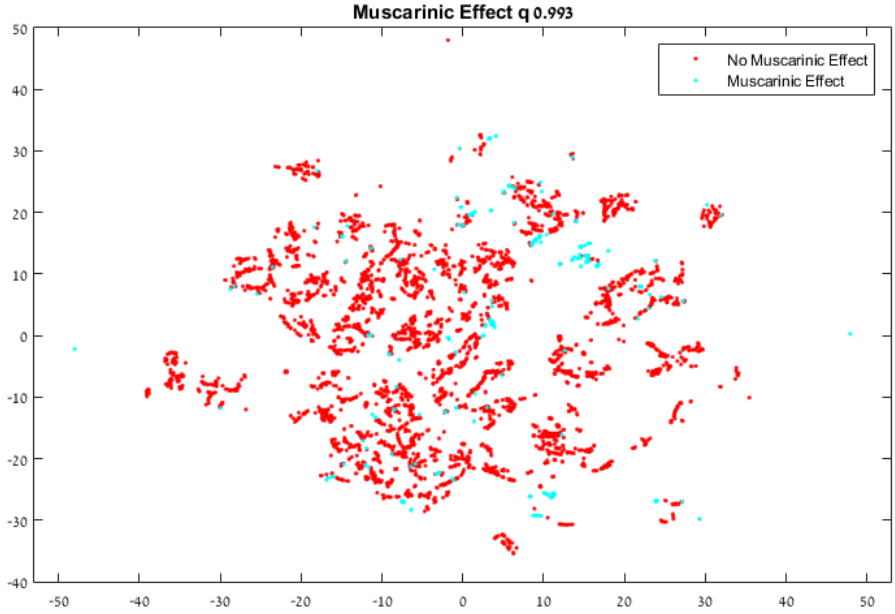

Figure 13 Muscarinic Results

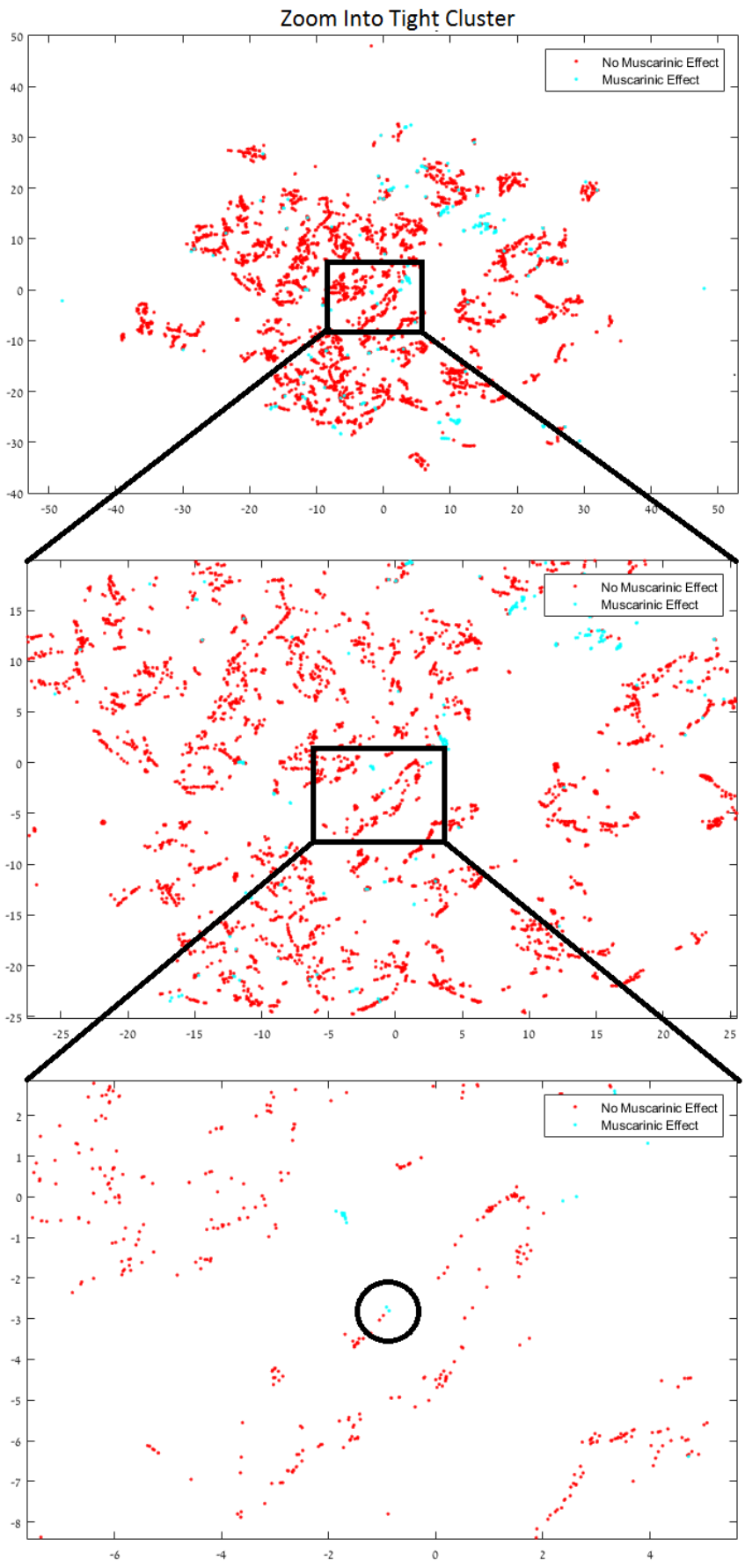

Figure 14 Muscarinic Results Magnified

Serotonin

The results for Serotonin (Figure 13) are as follows:

Name: Serotonin.

Single neighbor quality: 0.938.

Three neighbors quality: 0.934.

Five neighbors quality: 0.921.

Trust level: 0.09.

As the most dominant effect in the dataset (with regard to compound count) Serotonin has the lowest quality score which is still better in orders of magnitude than random selection.

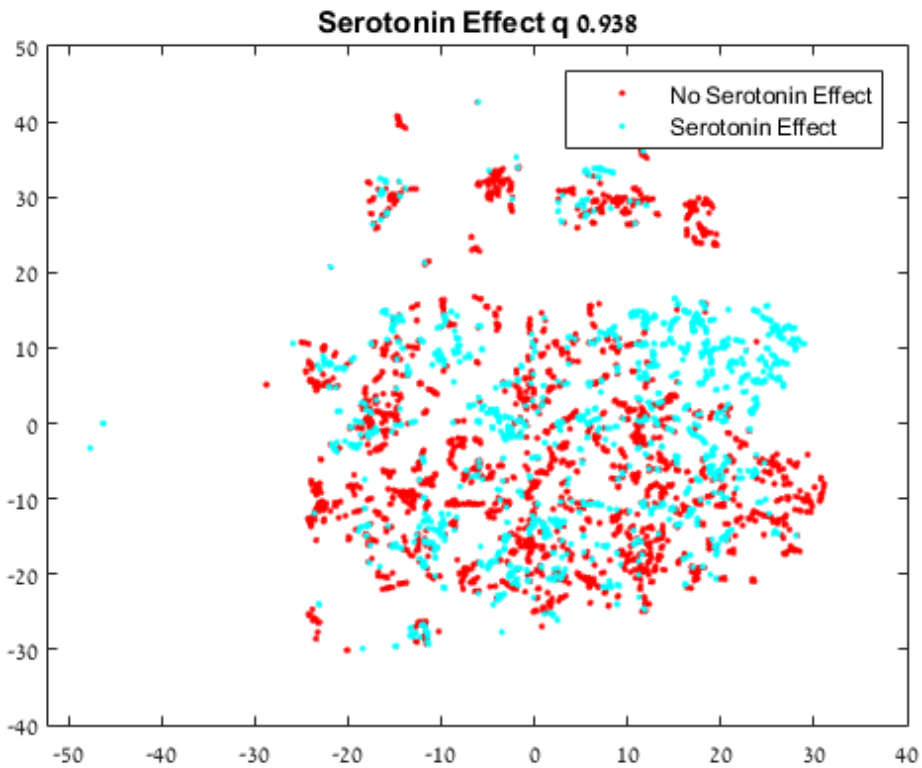

Figure 15 Serotonin Results

Summary

Following is a summary of the results in a single table.

	Q-1	*Q-3*	*Q-5*	*Trust*
Dopamine	0.965	0.965	0.961	0.087
Adrenoceptors	0.983	0.985	0.984	0.077
Histamine	0.967	0.963	0.964	0.078
Muscarinic	0.993	0.993	0.993	0.076
Serotonin	0.938	0.934	0.921	0.090

5. DISCUSSION

5.1. Verification

Feature Selection

The features are selected and persisted for the purpose of partial reproduction e.g. the final feature set can be used as the basis for future system runs for the same effect.

Dimensionality Reduction

High-dimensional data is projected into a two dimensional space and the distances are measured in the two dimensional space.

Fitness Function

The quality calculation serves as an adequate fitness function. Even though it might create instances in which the results make little sense to the naked eye zooming in make it clearer that the results are indeed reliable.

5.2. Validation

Tagging Compounds with Unknown Qualities

We have shown that in the tested cases the system placed compounds with similar effects near each other. This supports the conclusion that with enough tagged compounds the system can be relied upon to make the required recommendation and identify the compounds which are most likely to have the desired effect.

While the execution time is still a concern the system upon which the research was conducted on is a single home personal computer. The most trivial means of speeding the process would be to execute the dimensionality reduction of each generation member on a different machine and that would speed the process by a factor equal to the size of each generation. More complicated solutions would include multi core implementations of t-SNE.

These solutions, even though potentially speeding up the process by a factor of at least 160, might prove insufficient since the t-SNE algorithm has both time and space quadratic complexities. Even with industrial power machines and cloud computing databases of 100000 entries might have unrealistic computing power requirements – which means a powerful pre-processing will be required or a variant of t-SNE with lower complexity (there are some ongoing research efforts into that objective) will be used.

6. CONCLUSIONS

6.1. Conclusions

The research hypothesis discussed in this paper, that it is possible to optimize a feature selection for a dimensionality reduction algorithm so that the compounds selected according to the distance from tagged compounds present the best candidate for pre-clinical trials, was substantiated for an industry class database.

Examining the result for all the tested effect show that in all the cases the system produced results that were superior to random selection by at least one order of magnitude. It still remains to be seen whether that would be sufficient for the industry to start including the discovery phase as an integral part of their R&D process.

Compared to the benchmark introduced above (The Economy of Discovery) we have exceeded our expectations and we can see that in most cases no more than seven percent of the initial

corpus was left for the actually costly process of physical compound testing – by which we reduced the potential cost of the discovery phase to less than ten percent of its original cost.

6.2. Future Work

There are a few improvements that can be made to the system to make it more suitable to work in the industry.

Dataset Size

The dataset which the experiment was conducted with is a real industry standard dataset with regards to features. It is expected however that the pharmaceutical companies have datasets in one or more orders of magnitude larger. Since both the computational and the memory complexity of t-SNE are $O(n^4)$ larger datasets will not be manageable on a single machine and a distributed algorithm will have to be utilized.

Fast Dimensionality Reduction

There are a few ongoing research efforts into creating a faster version of t-SNE which would be able to handle significantly larger databases. Additionally there are other dimensionality reduction algorithms which might prove a reasonable substitution.

7. LIST OF PUBLICATIONS

- Yosipof, A.; Kaspi, O.; Majhi, K.; Senderowitz, H., Visualization Based Data Mining for Comparison Between Two Solar Cell Libraries. *Molecular Informatics* **2016**, 35, 622-628.

8. REFERENCES

1. English, R.; Lebovitz, Y.; Griffin, R., Institute of Medicine (US) Forum on Drug Discovery Development and Translation. Transforming clinical research in the United States. Challenges and opportunities: workshop summary. Washington (DC): National Academies Press (US); PubMed PMID: 21210556 **2010**
2. Joseph A. DiMasi, Tufts Center for the Study of Drug Development, Cost of Developing a New Drug
3. H Geerts; A Spiros; P Roberts and R Carr, Has the Time Come for Predictive Computer Modeling in CNS Drug Discovery and Development? CPT: Pharmacometrics and Systems Pharmacology · November 2012
4. Breakthrough Business Models: Drug Development for Rare and Neglected Diseases and Individualized Therapies: Workshop Summary, National Academies Press
5. An analysis of the attrition of drug candidates from four major pharmaceutical companies: Michael J. Waring, John Arrowsmith, Andrew R. Leach, Paul D. Leeson, Sam Mandrell, Robert M. Owen, Garry Pairaudeau, William D. Pennie, Stephen D. Pickett, Jibo Wang, Owen Wallace & Alex Weir.
6. Wayne Winegarden, Ph.D, The Economics of Pharmaceutical Pricing, Pacific Research Institute.
7. GUIYU ZHAO: The QSARome of the Receptorome: Quantitative Structure-Activity Relationship Modeling of Multiple Ligand Sets Acting at Multiple Receptors, University of North Carolina
8. Robas N, O'Reilly M, Katugampola S, Fidock M: Maximizing serendipity: strategies for identifying ligands for orphan G-protein-coupled receptors. Curr Opin Pharmacol 2003, 3:121-126.
9. Flower DR: Modelling G-protein-coupled receptors for drug design. Biochim. Biophys. Acta 1999, 1422:207-234.
10. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL: How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 2010, 9:203-214.