University of Freiburg

Master Thesis

Interactive de novo Molecular Design and Visualization

Thesis submitted in fulfilment of the requirements
for the degree of Master of Science

in the

Chair for Bioinformatics
Department of Computer Science

Done by:
Mohammed Juma Chbeib

Supervisor:
Prof. Dr. Rolf Backofen - Prof. Dr. Stefan Guenther

Reviewers:
Prof. Dr. Rolf Backofen - Prof. Dr. Stefan Guenther

April 2014

Abstract

Synthesis of small molecules that improve on the curative properties of existing drugs or that are effective in previously untreatable illnesses is a very hard task, a task on which pharmaceutical companies are investing enormous amounts of resources. Computational methods become therefore an interesting solution when they can effectively replace the time consuming and expensive design, synthesis and test phases.The Idea was to integrate the expert knowledge of (medicinal) chemists in the evaluation loop. Doing so in an efficient way is not a trivial task, since one has to 1) minimize the number of times the system resorts to the expensive human oracle, and 2) use a form of interaction suitable for humans. The research investigate novel ways to exploit the human visual perceptive system in chemists, so as to display the molecular space in a 3D rendering which is conductive to valuable insights.

Acknowledgements

I would like to thank everybody who helped and contributed to this project. I would like to thank Prof. Dr. Rolf Backofen and Prof. Dr. Stefan Gnther for their supervision . Also I would like to thank Dr. Fabrizio Costa, my direct supervisor who guided me through the whole project. Also Bjoern Gruening who helped me with a lot of a scientific and technical stuff,thanks a lot.

I would like to thank all my friends for their support. And Finally I would like to thank my Family. My Dad and my Mom,without I would not be where I am today.My Sisters Alyaa and Reem ,and My Brothers Hussam,Hisham and Hiatham you were always there for me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Synthesis of small molecules that improve on the curative properties of existing drugs or that are effective in previously untreatable illnesses is a very hard task, a task on which pharmaceutical companies are investing enormous amounts of resources. On average, development of a new drug takes 10 to 15 years and costs 400-800 million US dollars [7]. Most of the effort though, is spent on investigating compounds that in the end turn out to be unsuitable because of bad ADMET (absorption, distribution, metabolism, excretion, and toxicity) one out of about 5000 screened drug candidates reaches the market, the pharmaceutical industry is looking for fail fast, fail cheap solutions, i.e. having fast, cheap methods of determining whether the drug candidate does or does not have suitable properties to be a drug and should be rejected. Computational methods become therefore an interesting solution when they can effectively replace the time consuming and expensive design, synthesis and test phases. Amongst such computational methods, those capable to perform de novo molecular design are particularly interesting [8]. These approaches produce novel molecular structures with desired pharmacological properties from scratch in an incremental fashion. Since de novo molecule-design systems have to explore a virtually infinite search space typically resort to local optimization strategies. Commonly, a de novo design method has to address three questions: how to construct candidate compounds; how to evaluate their potential quality; and how to efficiently sample the search space. To date, one of the

most critical aspects is the reliability of the evaluation function. Such function is in fact invoked to judge the quality of molecules that can be (and generally are) very different from those used in the function induction phase therefore leading to unreliable scores. One possible approach to overcome this difficulty is to integrate the expert knowledge of (medicinal) chemists in the evaluation loop. Doing so in an efficient way is not a trivial task, since one has to 1) minimize the number of times the system resorts to the expensive human oracle, and 2) use a form of interaction suitable for humans. The research project goal is to investigate novel ways to exploit the human visual perceptive system in chemists, so as to display the molecular space in a 3D rendering which is conductive to valuable insights.

A typical scenario would be to identify a novel drug-like molecule 1) in the neighborhood of a specific compound under investigation; or 2) in some gap (i.e. unpopulated) region of the chemical space, where the notion of molecular distance can be defined in terms of graph kernel similarity. The visualization tool should therefore offer support for 1) displaying local clusters of molecules given a specific distance notion and 2) support the selection of regions of interest in the chemical space. Given the region of interest, a suitable subset of related representative molecules should be automatically performed which can be used as the starting point for the de novo synthesis module.

1.1 Outline of thesis

In the next chapter "Background" we lay the ground of Information Visualization.what it is and what are the common techniques .Then we cover Machine Learning topics such as dimensionality reduction and clustering.we describe the process of chemical space navigation and the drug discovery process. In Chapter 3 "Implementation" we describe the work flow of the process.explain the main components of the system and how they were implemented. In Chapter 4 "Experiments" we talk about performance result and benchmarks for the main algorithms. then we describe a typical use case step-by-step.The last chapter "Conclusion" we give our last thoughts and recommendation for future work

Chapter 2

Background

2.1 Information Visualization

2.1.1 What is Information Visualization ?

Information visualization is the study of interactive visual representations of abstract data to reinforce human cognition. The abstract data include both numerical and non-numerical data, such as text and geographic information.[9]
Or it can be described as compact graphical presentation and user interface for manipulating large numbers of items (possibly extracted from far larger datasets) to enables users to make discoveries,decisions, or explanations about patterns (trend, cluster, gap, outlier...),groups of items, or individual items.[10]
The visual data exploration process can be seen a hypothesis generation process: The visualizations of the data allow the user to gain insight into the data and come up with new hypotheses[1]. The verication of the hypotheses can also be done via visual data exploration but it may also be accomplished by automatic techniques from statistics or machine learning. In addition to the direct involvement of the user, the main advantages of visual data exploration over automatic data mining techniques from statistics or machine learning are:

- visual data exploration can easily deal with highly inhomogeneous and noisy data.

- visual data exploration is intuitive and requires no understanding of complex mathematical or statistical algorithms or parameters.

As a result, visual data exploration usually allows a faster data exploration and often provides better results, especially in cases where automatic algorithms fail. In addition, visual data exploration techniques provide a much higher degree of condence in the ndings of the exploration. This fact leads to a high demand for visual exploration techniques and makes them indispensable in conjunction with automatic exploration techniques [1].
As [1] explains the visual exploration paradigm as a three step process : Overview first, zoom and filter, and then details-on-demand . First, the user needs to get an overview of the data. In the overview, the user identifies interesting patterns and focuses on one or more of them. For analyzing the patterns, the user needs to drill down and access details of the data. Visualization technology may be used for all three steps of the data exploration process: Visualization techniques are useful for showing an overview of the data, allowing the user to identify interesting subsets. In this step, it is important to keep the overview visualization while focusing on the subset using another visualization technique. An alternative is to distort the overview visualization in order to focus on the interesting subsets. To further explore the interesting subsets, the user needs a drill-down capability in order to get the details about the data. Note that visualization technology not only provides the base visualization techniques for all three steps, but also bridges the gaps between the steps.

2.1.2 Interactivity

One of the main feature for a useful information visualization tool is flexible user interaction ability.There are a lot of interaction patterns that are used in the field

of visualization,the pattern importance varies depending on the type of data being visualized and the user being targeted .

One aspect is the ability to reconfigure the visualization settings (as for space visualization that would be the color palate,spacing of nodes in the space, camera angles).

Filtering facilities can help to reduce the number of data items displayed to those of specific interest to the user and their current task. This helps to reduce the visual complexity of the display.

Typically widgets such as sliders, buttons, menus, etc., are attached to different attributes of the data. Manipulating these controls to specify the desired attribute values should cause a rapid update of the main display. The choice of which attributes to allow the user to filter by may be determined from the task analysis stage.

In some cases a user might be interested in a subset of items or might want to hide a lot of irrelevant data items (in case of large datasets) the user that would make studying the space easier.The ability to highlight or hide these items is essential. Also the ability to give abstract or detailed Information about the data while taking into consideration viewing space and quick access . One more important feature is showing related items , this can be accomplished by showing different types of graph or by using coloring and scaling .

2.1.3 Visualization Techniques

2d visualization is very common in scientific data visualization. Due to the fact its easy to display and to understand. There exist a lot of 2d techniques such as scatter plot,bar chart,histograms among many others.

scatter plot are the most basic plot for displaying data in Cartesian coordinates. Data items are represented as points in Cartesian Space as shown in Figure 2.1. scatter plot can easily visualize trends and relationships in the data or identify outliers.however it becomes less informative with large data sets. Scatter plot can be extended easily to 3d Space .A 3rd dimension is tricky and the user have some

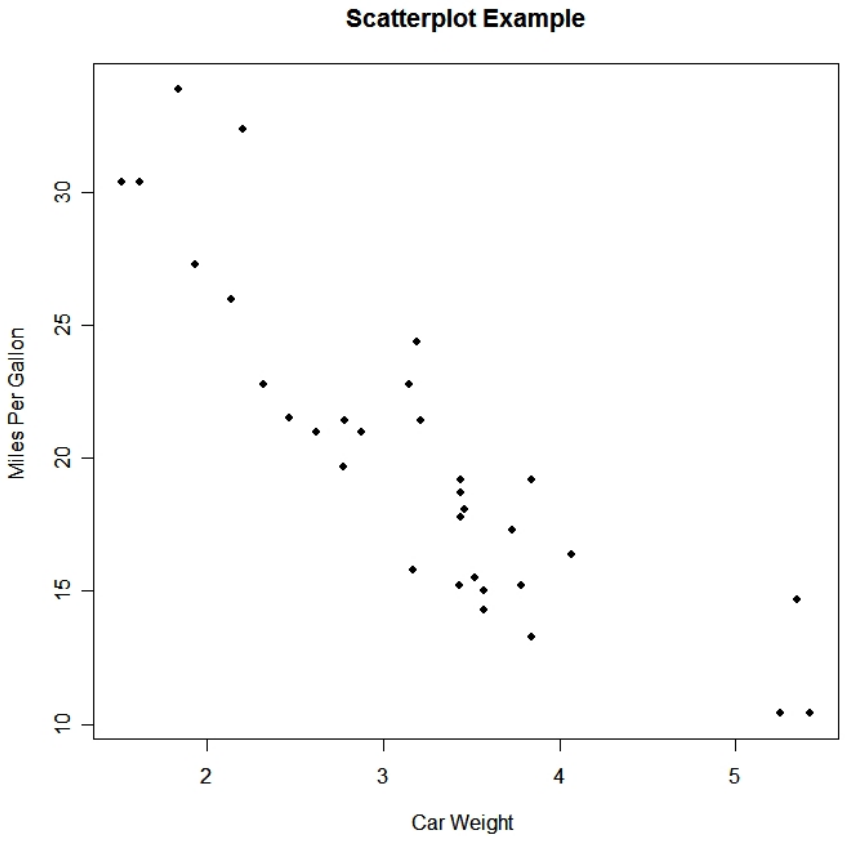


Figure 2.1: 2d scatter plot example

trade off to make .while it can provide a needed extra visual dimension, the scene can be hard to infer according to viewing angle and projection method.

Histograms are good visualization tool to show data items grouped according to occurrence or other aggregate property.both of these techniques can be categorized under traditional 2d/3d techniques . more sophisticated techniques exist such as Geometrically Transformed Displays. These techniques aim to find interesting transformations of multidimensional data sets. Another class of visual data exploration techniques are the iconic display techniques. The idea is to map the attribute values of a multidimensional data item to the features of an icon (arrows, faces, color ...etc) [1]

see Figure 2.2 show the classification of visualization techniques according to three different components : Data,Technique and Interaction .

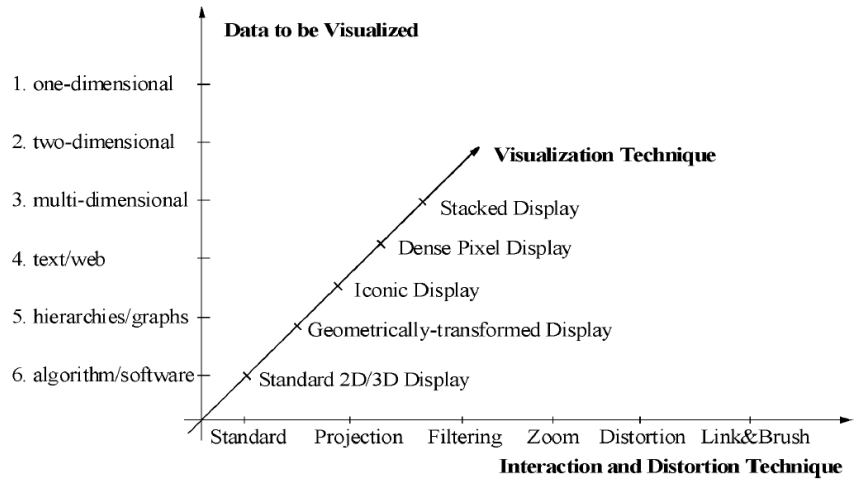


Figure 2.2: visualization techniques classification [1]

2.2 Dimensionality Reduction

In machine learning and statistics, dimensionality reduction or dimension reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction[11].Where Feature Extraction goal is to transform arbitrary data such as images and text into numerical features usable for machine learning , Feature Selection is a machine learning technique applied on these features.[12]. Feature extraction process can be categorized into two main categories : Linear and non-Linear Methods.some of these methods that are common in information visualization

- Principal Components Analysis (PCA)
- Kernel PCA
- Multidimensional Scaling (MDS)
- Independent component analysis
- Isomap

Dimensionality reduction Also known as Manifold Learning is an important problem that spans a lot of Information Processing areas like machine learning ,data compression and pattern recognition.
In many problems, the measured data vectors are high-dimensional but we may have reason to believe that the data lie near a lower-dimensional manifold. In other words, we may believe that high-dimensional data are multiple, indirect measurements of an underlying source, which typically cannot be directly measured. Learning a suitable low-dimensional manifold from high-dimensional data is essentially the same as learning this underlying source.
Dimensionality Reduction is performed usually when we want to produce a compact low-dimensional encoding of high dimensional dataset . For the data visualization task this provides an interpretation of the dataset in a suitable way to represent the data so it can be consumed by the human cognitive system . Also its considered to be a preprocessing step for many supervised learning algorithm .

2.2.1 Mathematical Background

In statistics the standard deviation σ shows how much variation or dispersion from the average exists A low standard deviation indicates that the data points tend to be very close to the mean (also called expected value) a high standard deviation indicates that the data points are spread out over a large range of values. Variance is the standard deviation squared.Both of these measurement are measures to data spread [13] .

The last two measures we have looked at are purely 1-dimensional.. However many data sets have more than one dimension, and the aim of the statistical analysis of these data sets is usually to see if there is any relationship between the dimensions.. However, it is useful to have a similar measure to nd out how much the dimensions vary from the mean with respect to each other. Covariance is such a measure.[14]

$$cov(X,Y) = \frac{\sum_1^n (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

A useful way to get all the possible covariance values between all the different dimensions is to calculate them all and put them in a matrix.That is the Covariance Matrix.

An eigenvector of a square matrix A is a non-zero vector v that, when the matrix is multiplied by v, yields a constant multiple of v, the multiplier being commonly denoted by λ. That is:

$$Av = \lambda v$$

The number λ is called the eigenvalue of A corresponding to v . [15]

2.2.2 Principal Components Analysis (PCA)

PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to nd in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analysing data. The other main advantage of PCA is that once you have found these patterns in the data, and you compress the data, ie. by reducing the number of dimensions, without much loss of information. This technique also used in image compression. [14]

Given a set of data on n dimensions, PCA aims to find a linear subspace of dimension $d < n$ such that the data points lie mainly on this linear subspace . Such a reduced subspace attempts to maintain most of the variability of the data. The linear subspace can be specified by d orthogonal vectors that form a new coordinate system, called the 'principal components'. The principal components are orthogonal, linear transformations of the original data points, so there can be no more than n of them [16].

The most common definition of PCA, due to Hotelling[17], is that, for a given set of data vectors x_i variance retained under projection is maximal. In order to capture as much of the variability as possible, let us choose the first principal component, denoted by U_1, to have maximum variance. Suppose that all centered observations are stacked into the columns of an $n \times t$ matrix X, where each column corresponds to an n-dimensional observation and there are t observations. Let the

first principal component be a linear combination of X defined by coefficients (or weights) $w = \begin{bmatrix} w_1 \dots w_n \end{bmatrix}$ In matrix form

$$U_1 = w^T X$$

$$var(U_1) = var(w^T X) = w^T S w$$

where S is the $n \times n$ sample covariance matrix of X.
Clearly $var(U_1)$ can be made arbitrarily large by increasing the magnitude of w. Therefore, we choose w to maximize $w^T S w$ while constraining w to have unit length.

$$\max w^T S w$$

$$\text{subject to } wTw = 1$$

To solve this optimization problem a Lagrange multiplier α_1 is introduced:

$$L(w, \alpha) = w^T S w - \alpha_1 (w^T w - 1) \tag{2.1}$$

Differentiating with respect to w gives n equations,

$$Sw = \alpha_1 w$$

Pre multiplying both sides by w^T we have:

$$w_T S w = \alpha_1 w^T w = \alpha 1$$

$var(U_1)$ is maximized if $\alpha 1$ is the largest eigenvalue of S. Clearly α_1 and w are an eigenvalue and an eigenvector of S. Differentiating 2.1 with respect to the Lagrange multiplier α_1 gives us back the constraint:

$$w^T w = 1$$

This shows that the first principal component is given by the normalized eigenvector with the largest associated eigenvalue of the sample covariance matrix S. A

similar argument can show that the d dominant eigenvectors of covariance matrix S determine the first d principal components.

2.2.3 Multidimensional Scaling (MDS)

Multidimensional scaling (MDS) is a means of visualizing the level of similarity of individual cases of a dataset. It refers to a set of related ordination techniques used in information visualization, in particular to display the information contained in a distance matrix [18].

MDS pictures the structure of a set of objects from data that approximate the distances between pairs of the objects. The data, which are called similarities. dissimilarities, distances, or proximities, must reflect the amount of (dissimilarity between pairs of the. In this article we use the term similarity generically to refer to both similarities (where large numbers refer to great similarity) and to dissimilarities (where large numbers refer to great dissimilarity). In addition to the traditional human similarity judgment, the data can be an "objective" similarity measure (the driving time between pairs of cities) or an index calculated from multivariate data (the proportion of agreement in the votes cast by pairs of senators). However, the data must always I represent the degree of similarity of pairs of objects (or events).

Each object or event is represented by a point in a multidimensional space. The points are arranged in this space so that the distances between pairs of points have the strongest possible relation to the similarities among the pairs of objects. That is, two similar objects are represented by two points that are close together, and two dissimilar objects are represented by two points that are far apart. The space is usually a two- or three-dimensional Euclidean space, but may be non-Euclidean and may have more dimensions. [19].

Although it has a very different mathematics from PCA, it winds up being closely related, and in fact yields a linear embedding, as we will see[16]. A $t \times t$ matrix D is called a distance or affinity matrix if it is symmetric, $d_{ii} = 0$, and $d_{ij} > 0; i \neq j$

Given a distance matrix D, MDS attempts to find t data points $y_1 \dots y_t$ such that if d_{ij} denotes the Euclidean distance between y_i and y_j In particular, we consider metric MDS , which minimizes

$$\min Y \sum_{1}^{t} \sum_{1}^{t} \left({d_{ij}}^{(x)} - {d_{ij}}^{(y)} \right)^2 \tag{2.2}$$

where ${d_{ij}}^{(x)} = \|x_i - x_j\|^2$ and ${d_{ij}}^{(y)} = \|y_i - y_j\|^2$ The distance matrix $D^(x)$ can be converted to a kernel matrix of inner products $X^T X$ by

$$X^T X = -\frac{1}{2} H D^(x) H$$

where $H = I - \frac{1}{t} e e^T$ and e is a column vector of all 1's. Now 2.2 can be reduced to

$$\min Y \sum_{1}^{t} \sum_{1}^{t} \left(({x_i}^t) x_j - ({y_i}^t) y_j \right)^2$$

It can be shown that the solution is $Y = \Lambda^{\frac{1}{2}} V^T$ corresponding to the top d eigenvalues, and Λ is the top d eigenvalues of $X^T X$ where V is the eigenvectors of $X^T X$ [16].

2.2.4 Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. It was introduced first in 1995 by Vladimir N. Vapnik [20]. The operation of the SVM algorithm is based on finding the hyperplane that gives the largest minimum distance to the training examples. This distance receives the important name of margin within SVMs theory. Therefore, the optimal separating hyperplane maximizes the margin of the training data (see Figure 2.3).

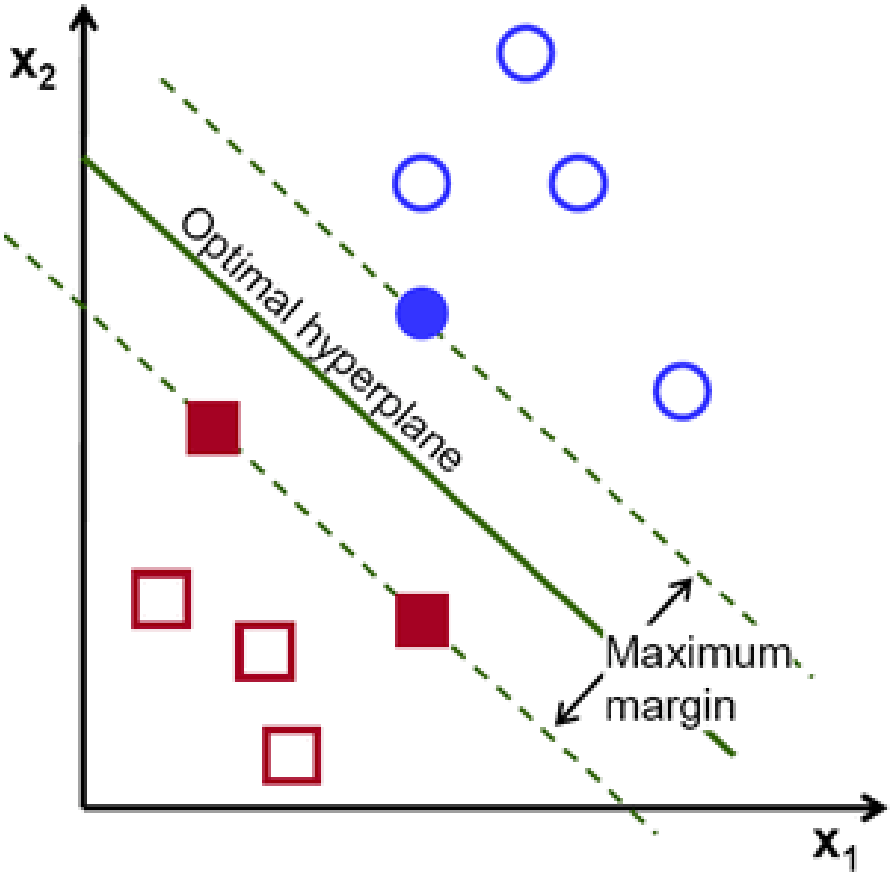


Figure 2.3: SVM optimal hyperplane [2]

2.3 Similarity Measurement

Distance or similarity measures are essential to solve many pattern recognition problems such as classification, clustering, and retrieval problems. From the scientific and mathematical point of view, distance is defined as a quantitative degree of how far apart two objects are. Synonyms for distance include dissimilarity. The choice of distance/similarity measures depends on the measurement type or representation of objects [21].

The Similarity principle for chemical compounds states that similar molecules in structure have similar properties [6]. To decide the similarity of molecules we need two A suitable representation of molecules see **??**sec:fingerprints) and An efficient comparison method.

The similarity coefcients (index, distance) are functions that transform pairs of compatible molecular representations into real numbers, usually lying on the unit

interval. They provide a quantitative measure of the chemical resemblance degree [6] .

In the following Table 2.1 shows the most common Similarity Metrics

Index/Coefcient/Distance	**Expression**
Tanimoto	$S_T = \frac{c}{a+b-c}$
Cosine	$S_C = \frac{c}{\sqrt{ab}}$
Squared Euclidean	$S_E = \frac{a+b-2c}{n}$
Russell-Rao	$S_R = \frac{c}{n}$
Forbes	$S_F = \frac{cn}{ab}$

Table 2.1: Most common similarity metrics [6]
For evaluating the similarity between two molecules with the formulas listed in 2.1, a represents the properties of the first molecule and b the second. n is the total number of properties. c is the number of common properties and d the number of uncommon ones between the two molecules

2.3.1 Similarity Matrix

A similarity matrix is a matrix of scores that represent the similarity between a number of data points. Each element of the similarity matrix contains a measure of similarity between two of the data points. Similarity matrices are strongly related to their counterparts, distance matrices and substitution matrices. Diagonal entries of a similarity matrix are ignored since all data objects are completely similar to themselves. We also assume that all similarity matrices are symmetric, so for calculating a similarity matrix for a data set of n data objects, it is enough to calculate the proximity $n(n-1) = 2$ times in order to find all the pairwise similarities.

2.4 Chemical Space and Drug Discovery

Chemical space has become a key concept in drug discovery. The continued growth in the number of molecules available raises the question regarding how many compounds may exist and which ones have the potential to become drugs. Analysis

and visualization of the chemical space covered by public, commercial, in-house and virtual compound collections have found multiple applications in diversity analysis, in silico property profiling, data mining, virtual screening, library design, prioritization in screening campaigns, and acquisition of compound collections, among others [22].

Chemical space can be viewed as being analogous to the cosmological universe in its vastness, with chemical compounds populating space instead of stars [23].Dobson in his insight defines chemical space as the total descriptor space that encompasses all the small carbon-based molecules that could in principle be created[24].

As a concept it suggests a representation in form of geographical map to illustrate the distribution of molecules and their property [25] . The idea behind any representation of chemical space is to be able to use the positional information within this space to search for bio-active molecules, thus performing virtual screening to select compounds for in vitro testing. In that respect, the relevance of any chemical space must be judged by its ability to group compounds with similar bioactivity together. The entire chemical space is far too large for an exhaustive enumeration, even using todays computers. One is therefore left with a partial, targeted enumeration as the only option to produce molecules for virtual screening [25].

Visualization of the chemical space of a compound collection will largely depend on two main factors: the molecular representation of the molecules to define the multi-dimensional descriptor space (vide supra), and the visualization technique used to reduce the multi-dimensional space into a two- or three-dimensional graph. Noteworthy, the chemical space of a compound collection will not be unique as it will depend on the particular representation used to define the multi-dimensional space.

In [22] different computational approaches to visualize the chemical space are described.In [26],[22] and [27] Principal Components Analysis (see section 2.2.2) use as a visualization technique on combinatorial libraries represented via their topological properties with the main goal being diversity analysis.In [28] They do

diversity analysis by performing Multifusion similarity maps on Binary Fingerprint (see section **??**) mainly performed on drugs and bio-active molecules.Self organizing maps(SOM) used in both [29] and [30] for diversity analysis and classification studies.

2.4.1 *De Novo Molecular Design*

De Novo Molecular Design is computational method that produce in an incremental fashion, novel molecular structures with desired pharmacological properties from scratch. First introduced from in the late 80s and early 90s, when the rst automated de novo design techniques were implemented[31] and since then these computer-based techniques have became a solid complementary approach to high-throughput screening.The most important aspect of a de novo design is the ability to produce results in a time- and cost-efficient manner.

There are three main main phases when it comes to drug development: design, synthesis and test. Any de novo design system tries to effectively replace the design and synthesis phases , and to do this, it needs to decide on strategies for three problems [8]:

- assembly: how to construct candidate compounds?
- scoring: how to evaluate their potential quality?
- search: how to efficiently sample the search space?

To answer the first question ,There are two approaches when it comes to constructing new candidate structures, that relate to the building blocks used for the "articial" design phase: the atom- or the fragment-based approach [32].

- **atom-based** is the superior approach due to the structural variety that it can produce , but the search space becomes innite and the challenge becomes selecting the best/suitable candidates.
- **fragment-based** has a smaller search space to deal with as it performs a "meaningful reduction" by focusing on relevant areas and it allows the designer to choose how to dene the fragments.

The second step is to evaluate their quality.the scoring function has a central role in the system, as it not only judges the quality of the new molecules, but implicitly assigns fitness values to them and guides the design process through the search space.These evaluation functions are mainly based on the so-called primary target constraints which are generated from all the information related to the ligand-receptor interaction that the function can collect. Based on the source used for collecting these constraints, the design strategies can be divided in two:

- **receptor-based** constraints are gathered from the 3D receptor structure which makes these approaches limited to target proteins with known 3D structure; however, this is not always the case for relevant pharmaceutical targets
- **ligand-based** constraints are gathered from known ligands of the particular target

The third and last question is how to search in the (virtually infinite) chemical space ? local optimization strategies are used to sample the search space, making the solution converge to a "practical" optimum. Any algorithm based on the local search approach works on the idea of moving from a solution to the next in the space of candidate solutions (by performing local changes to the current set) until an optimal solution (set) has been found or a threshold was reached.There are two types of designs one can use when implementing such a local (chemical) space search algorithm:

- **positive design** restricts the local search space to the areas that it considers to be more likely for producing drug-like molecules
- **negative design** defines and avoid areas of the search space containing adverse properties and unwanted structures

2.4.2 Chemical Compounds Representation

Chemical information is usually provided as files or streams and many formats have been created, with varying degrees of documentation. Some of the most

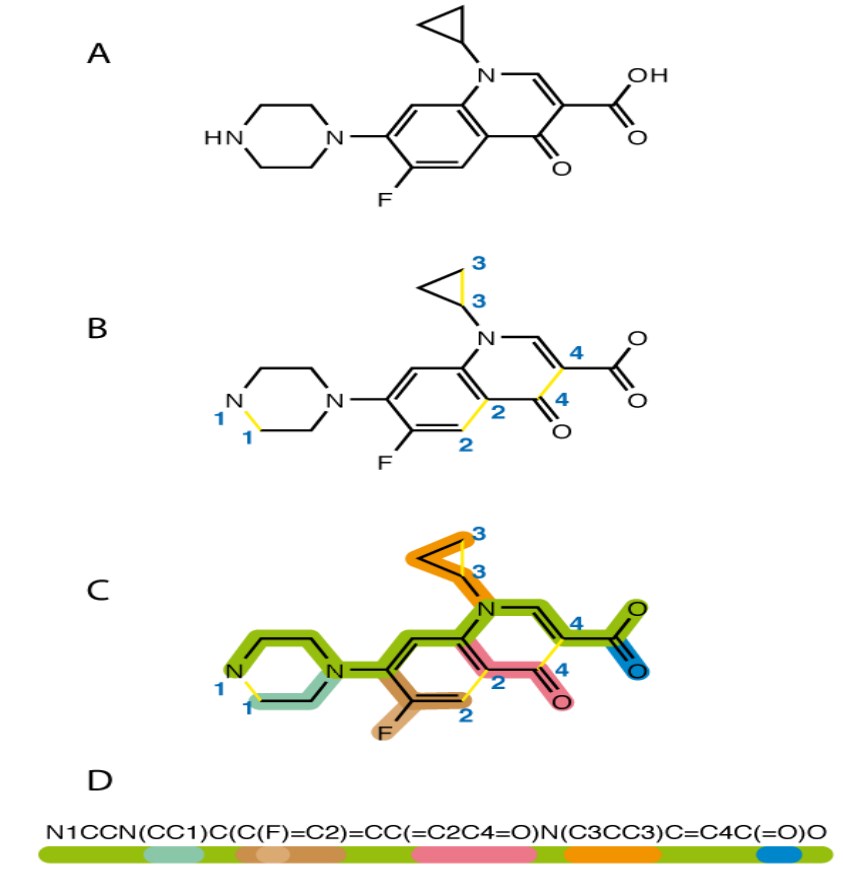


Figure 2.4: SMILE Generation Process [3]

common types are SMILES,SDF and Molecular Graph.

2.4.2.1 SMILE

The simplied molecular-input line-entry system (SMILES) is a linear ASCII string for representing the structure of chemical molecules. This string is a computer readable format.SMILES allows users to annotate any chemical structure [3]. The main advantage of SMILES that it is easy and efficient to be processed by computers.Each structure has a unique SMILE. The Figure 2.4 shows a SMILE and a general algorithm to generate it.

2.4.2.2 Chemical Table File formats (SDF)

One of the most widely used industry standards are chemical table file formats, like the Structure Data Format (SDF) files. They are text files that adhere to a strict format for representing multiple chemical structure records and associated data fields. The format was originally developed and published by Molecular Design Limited (MDL). MOL is another file format from MDL.

2.4.2.3 Molecular Graph

A graph with differently labeled (colored) vertices (chromatic graph) which represent different kinds of atoms and differently labeled (colored) edges related to different types of bonds. Within the topological electron distribution theory, a complete network of the bond paths for a given nuclear configuration [33]. In this way, a molecules becomes a mathematical element ready to be theoretically studied by applying graph theory studies.

2.4.2.4 Chemical Hashed Fingerprints

The chemical hashed fingerprint of a molecule is bit string (a sequence of "0" and "1" digits) that contains information on the structure. Mostly used for Chemical database handling for full structure, substructure, and similarity searching. It provides a rapid and effective screening, but in case of structural search it may generate false hits. For this reason the results have to be checked by a precise but slower atom-by-atom search. Also used in Combinatorial chemistry for the diversity/similarity analysis of compound libraries [4]. The figure 2.5 below shows this process on an example

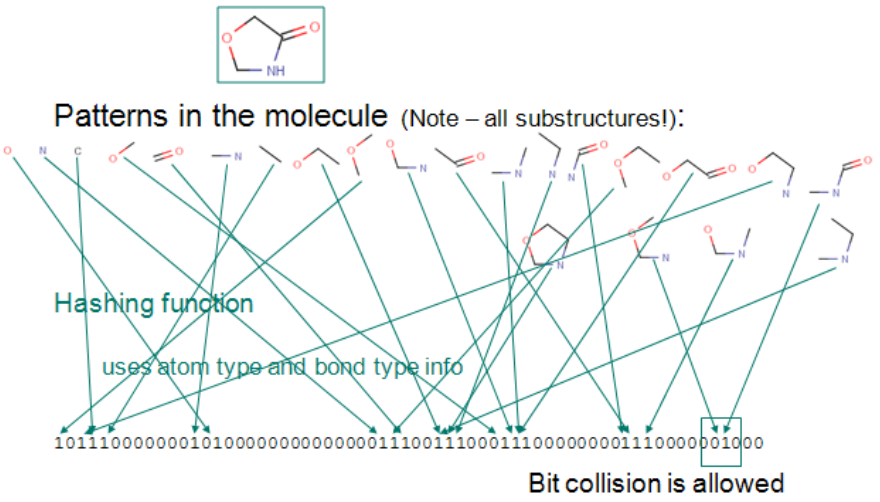


Figure 2.5: Chemical Hashed Fingerprints Generation Process [4]

2.5 Clustering

Clustering is the process of grouping a set of data items in subsets. Given a representation of n objects, clustering is to find k groups(clusters) based on a measure of similarity such that objects within the same group are similar but the objects in different groups are not similar [34]. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. Additionally, unknown groups of data can be discovered via the clustering, these groups are called classes or clusters. A cluster is a collection of data objects that are similar to one another within the cluster and dissimilar to objects in other clusters. Therefore a cluster of data objects can be treated as an implicit class.

2.5.1 Quick Shift Clustering

Quick shift seeks the energy modes by connecting nearest neighbors at higher energy levels, trading-off mode over- and under-fragmentation.

Given a set of samples $x_1, x_2, \ldots, x_N$ in the d dimensional space R^d, quick shift algorithm [35] estimates the underlying probability density function for each sample

xi

$$P(x_i) = \sum_{j=1}^{N} K(\left\| \frac{x_j - x_i}{h} \right\|^2)$$

where $K(\cdot)$ is a kernel function, and h is the bandwidth. In order to shift each sample to the nearest mode, each sample x_i simply moves to the nearest neighbor $y(x_i, 1)$ that has higher density. In formulas,

$$y(x_i, 1) = \begin{cases} q(x_i) & \text{if } \|q(x_i) - x_i\| < \tau \\ x_i & \text{otherwise} \end{cases}$$

$$q(x_i) = \arg \max_{j=1,\ldots,N} \frac{sgn(P(x_j) - P(x_i))}{\|x_j - x_i\| + 1}$$

where τ is a shift parameter. In the case that there is no sample having higher density,$q(x_i)$ becomes x_i. If the distance between x_i and $q(x_i)$ is smaller than τ , the sample continues to be associated with other samples (e.g. $y(x_i, m + 1)$) as follows:

$$y(x_i, m + 1) = y(y(x_i, m), 1) \tag{2.3}$$

where m is the number of movements. When $y(x_i, m + 1)$ is the same as $y(x_i, m)$ in Equation 2.3, each sample x_i is connected to the nearest mode $y(x_i)$.
Quick shift has four advantages:simplicity; speed $O(dN2)$ with a small constant , generality , and a tuning parameter to trade off under- and over-fragmentation of the modes [36].

2.6 Contribution

A real time interactive chemical space visualization tool where we go from enumerated chemical compounds to a space visualization by clustering and different embedding techniques .Pre clustering is used in case of big libraries where it can decrease embedding time and simplify the space for the observer.Different embedding techniques are implemented and compared .Also the ability to perform de novo molecular design to fill in the gaps in the space within a cognitive context approach while giving the user a dynamic way of filtering and highlight areas of interest in the space .

Chapter 3

System Implementation

3.1 Workflow

The process starts with the user uploads the main collection file (one of the chemical file format , a lot of chemical collections are available online such as DrugBank[37],PubChem[38]).The collection file is then parsed and stored in a document based database. From this point on the user have the ability to add meta information to the collection at any point during the following phases.
Then the data will be pre-processed ,features and similarity matrix will be calculated as an input for the Manifold Learning phase. The user then can either perform pre-clustering of the data and then perform embedding of the feature into the Cartesian space or he can perform the embedding directly .
At this point the collection is ready for visualization ,highlighting and filtering to find visually interesting areas to trigger a molecule synthesis.

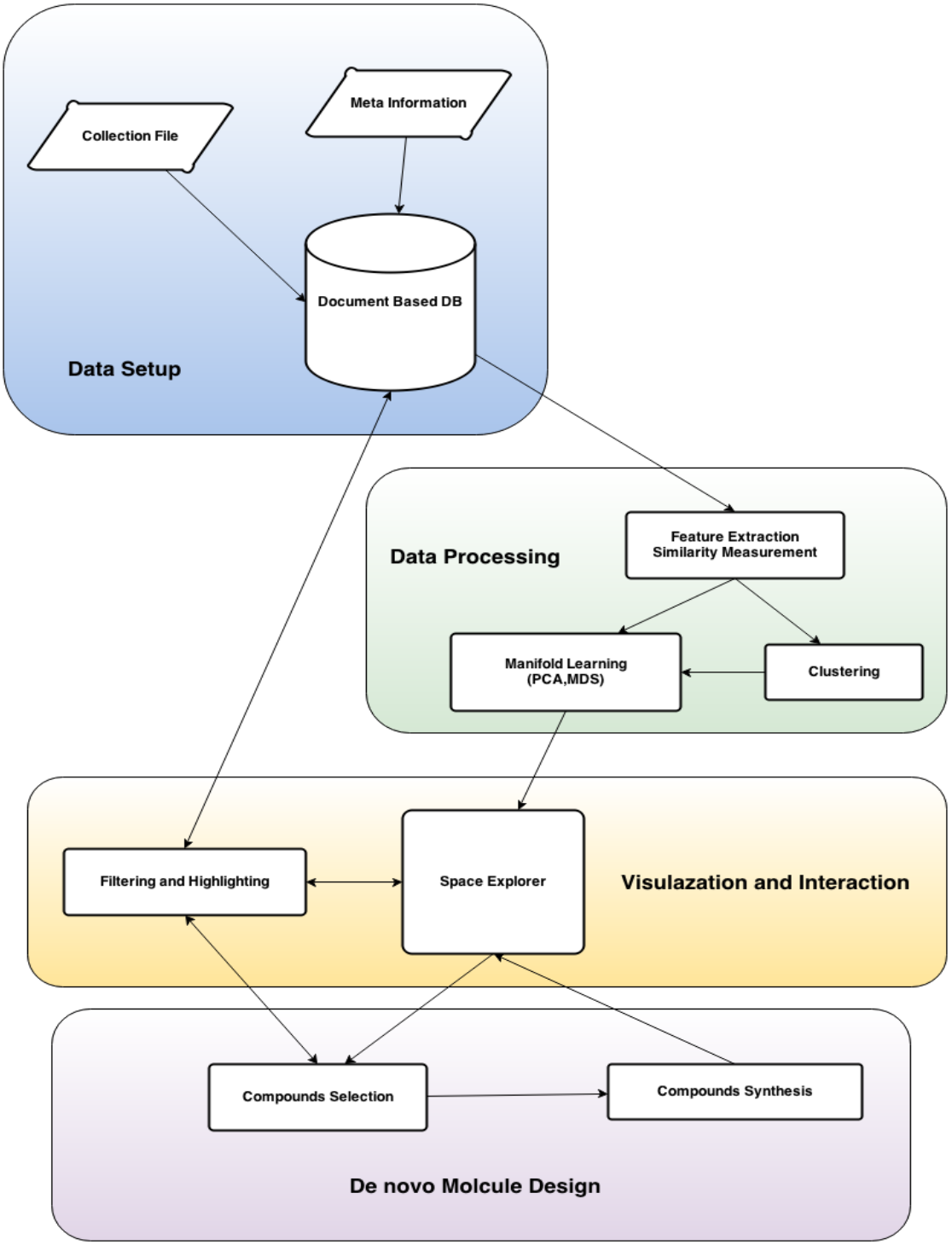


Figure 3.1: Workflow overview

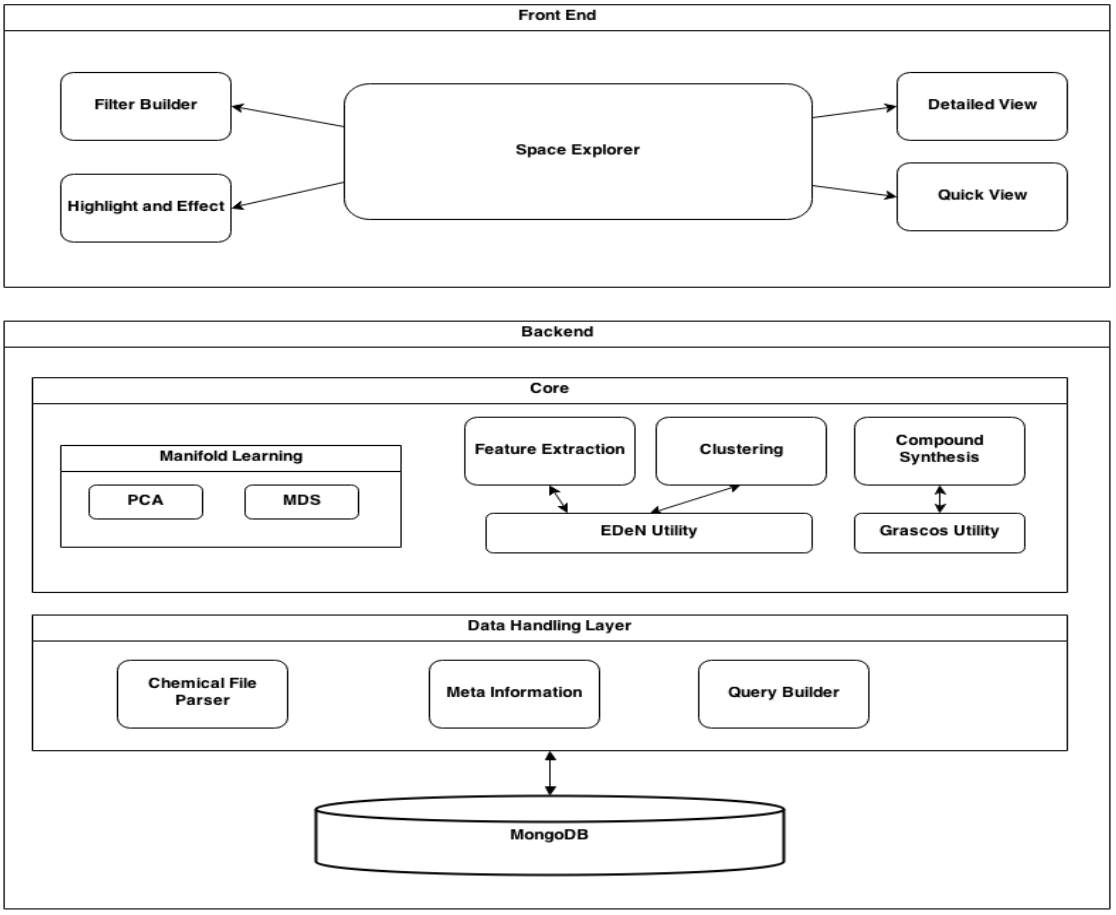


Figure 3.2: Architecture and System Components overview

3.2 Architecture

The tool was build on a web-based Architecture . The database is MongoDB ,a Non-SQL document based database. For the Application Layer we used Django Framework (a python MVC). for the Front End and UI we used HTML5,Javascript and WebGl. These components and more details on the architecture in the Figure 3.2.

3.3 System Components

3.3.1 Feature Extraction and Similarity Calculation

Our feature extraction and similarity calculation were based on a state of the art tool called EDeN (Explicit Decomposition with Neighborhoods) which performs these tasks based on Neighborhood Subgraph Pairwise Distance Kernel (NSPDK) [5] that we will explain in the next section.

3.3.1.1 Neighborhood Subgraph Pairwise Distance Kernel (NSPDK)

In order to learn predictive models from graphs, efficient graph processing methods are required. Graph kernels can be described as functions which measure the similarity between two graphs. Since similarity information is sufficient for binary classification of instances, graph kernels can be utilized for this task. In this thesis, the Neighborhood Subgraph Pairwise Distance Kernel (NSPDK) [5] was used, since it efficiently computes the similarity between two graphs in linear time.
The main concept behind NSPDK is that it extracts graph features by decomposing the graph into pairs of neighboring subgraphs, controlled by two parameters (distance d and radius r). Radius r marks the maximum size of the subgraph, while distance d denotes the maximum distance between the roots of the two subgraphs, determined by the shortest path between them (see Figure 3.3). For every unique pair-of-subgraphs feature in the graph, the graph kernel subsequently stores its number of occurences in a feature vector. Note that if $r > 1$, all possible subgraphs up to size r will also be extracted (the same applies for all possible pairs of subgraphs up to distance d if $d > 1$).

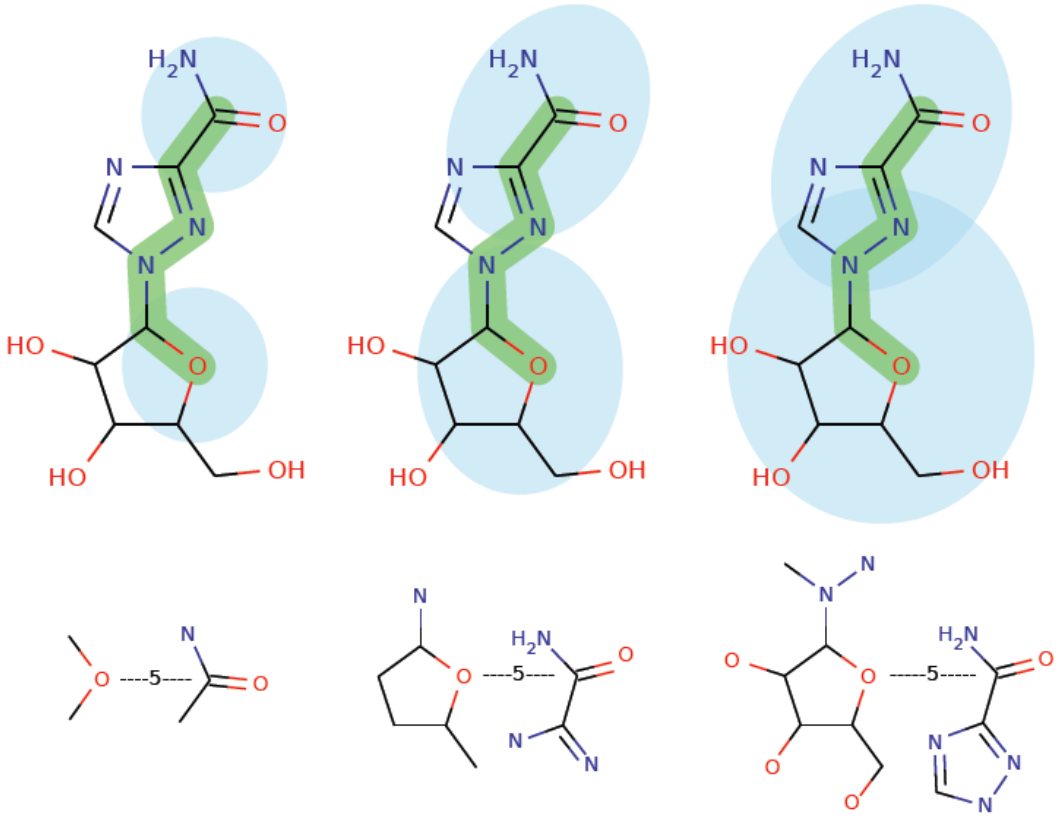


Figure 3.3: Illustration of pairs of neighborhood graphs for radius $r = 1; 2; 3$ and distance $d = 5$ [5].

3.3.2 Collections and Meta Information

The user start by uploading a Compound Collection File (in one of the accepted formats ,SDF,SMILE). The file is passed to the parser and a new collection of compounds is generated and stored in the Database. The compounds model mainly has initial the basic attributes :

- name
- formula
- ring count
- molecular weight
- SMILE code

However we use a Document-oriented database offers more flexibility and more convince when your data is not highly relational. Data is stored in collections that

dont enforce the document structure which means that documents in the same collection do not need to have the same set of fields or structure and common fields in a collections documents may hold different types of data. This allows us to add Meta information to the model dynamically. A CSV (comma separated value) file that contains the Meta Information can be uploaded and parsed. The table 3.1 shows the format used

CompoundID	**Attribute 1**	**Attribute 2**	**...**	**Attribute n**
000001	value 1	value 2	...	value n
000002	value 1	value 2	...	value n
...	...	...	...	...

Table 3.1: Meta Information format

For each row the system will try to identify the compound in the database and associate it with the specified arbitrary metadata, if it is not found the row will be discarded.

We used MongoDB which is one of the leading non-sql databases. It has full, flexible index support and rich queries and Built-in replication for high availability.

3.3.3 Clustering

One of the ways to decrease calculation time and enable the user to have an overview of the space without cluttering the screen is by performing pre clustering. In the case of large collections pre clustering can decrease the overall calculation time needed for embedding. The process as shown in 3.4 starts by performing the quick-shift clustering 2.5.1 on the collection.Then from each Cluster we pick a cluster centroid (representative) , the compound with the highest aggregate similarity average to all other compounds in the cluster. A new collection of cluster centroid is then created and use as an input for Feature Extraction/Similarity Matrix calculation. Then we perform Dimensionality Reduction on cluster centroids collection (either PCA or MDS). The individual cluster can be then embedded locally on the space by clicking on the cluster.

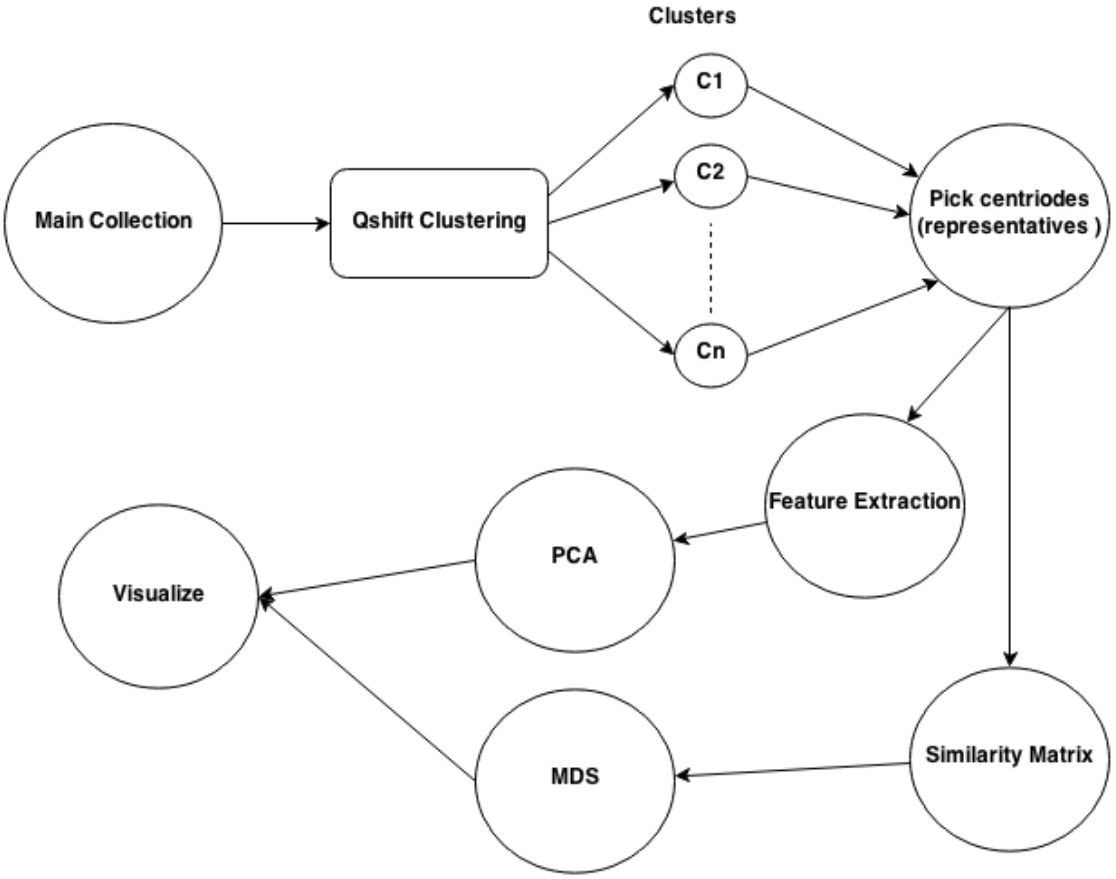


Figure 3.4: Clustering Workflow

During the clustering, the inner cluster similarities are stored and used visualize graphs within the cluster with the ability to show/hide graph edges based on a similarity threshold.

We used Spheres to represent clusters as follows :

- Sphere Center is represented by the location of the cluster centroid.
- Sphere Radius is defined as the size of the cluster (number of compounds in the cluster).
- Sphere Color Saturation is used to apply filters.Where the saturation is defined as $saturation = \frac{\text{number of cluster compounds that match the filter}}{\text{total number of cluster compounds}}$
- Sphere Visibility can be toggled based on size threshold,which gives the user the ability to simplify the space and focus the research on bigger more interesting clusters.

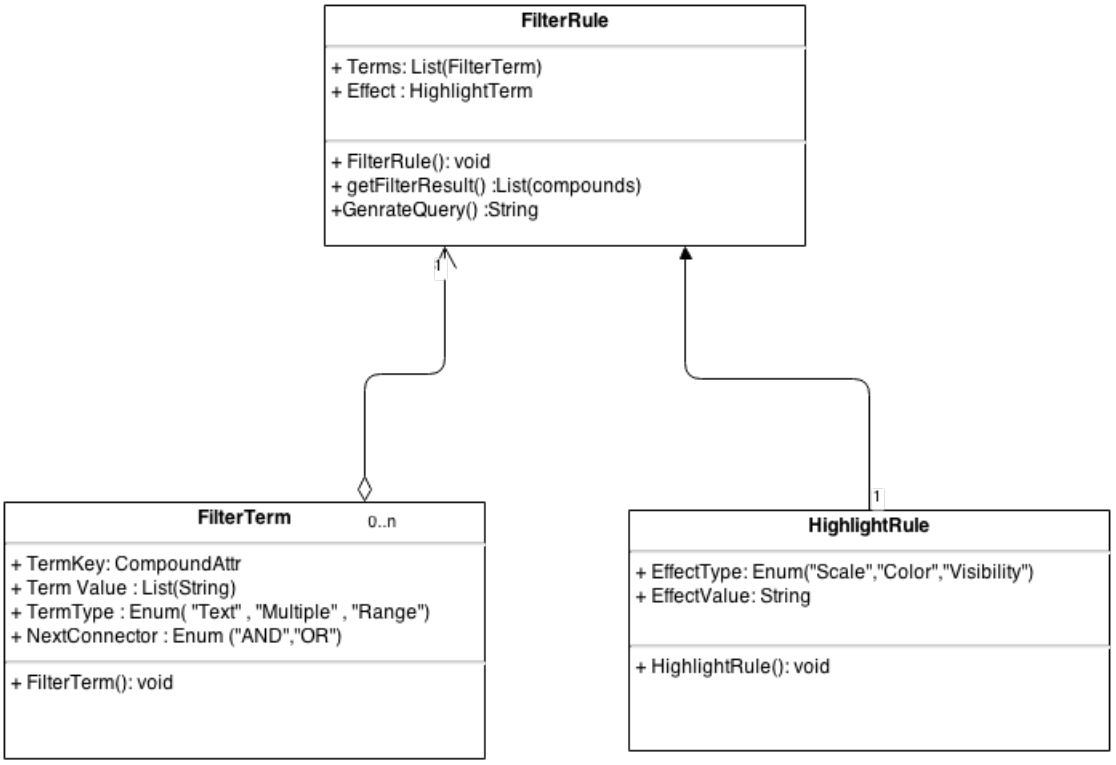


Figure 3.5: Filtering and Highlighting Data model

3.3.4 Filtering and Highlighting

One of the most important feature for Information Visualization is the ability to Filter and Highlight certain parts of the space according to some rules that might interest the user. It should be interactive and easy to use and build.
We achieved this we implemented Filtering and Highlighting rules . The figure **??**

3.3.4.1 Filtering

The Filter Data model defines the basic rules to match for a certain set of items the main collection. Each filter item consist of a list of multiple terms. A term holds the basic querying item (see 3.5) connected with a logical operator. The benefit from the use of document based database is the ability to create filters for dynamically added fields (Meta information) in an incremental way as we are working on a collection. Alongside the list of terms it holds the Highlight rule.

- **Term Key :** A list of all available attributes of the current compound collection. Dynamically updated as we add meta information .
- **Term Value :** A value or list of values that we want to filter (with respect to the Term Key). Depends on the the Term Type.
- **Term Type :**
  - Multiple-Choice : choose a list of available values to filter
  - Range : retrieve all value within a specific rang for this specific key
  - Text: search for some text or expression
- **Next Term Connecter :**
  - AND
  - OR

3.3.4.2 Highlighting

A highlight rule defines the effect to be applied on the filtered items.Highlight rules are applied to clusters and compounds at the same time. cluster highlighting is described in 3.3.3.The highlight rule is defined as follows :

- **Effect Type :**
  - Scale : to change the scale of the filtered compounds.
  - Color: to change the color of the filtered compounds.
  - Visibility : to show/hide the filtered compounds.
- **Effect Value :** Specify the scale or the color value of the effect.

3.3.5 Selection and Molecule Design

In the process of creating a new compound to fill the gaps in the space,the user must select a group of compounds that would serve as an input for the compound synthesis process. The 3.6 shows the process of creating a new compound.The user first must select a set of compounds. Two modes of selection are available :

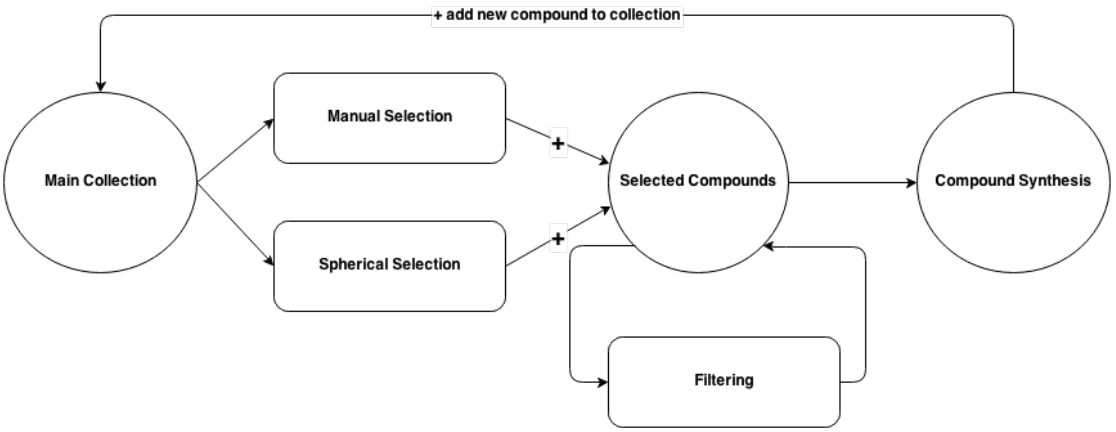


Figure 3.6: Workflow of new compound Synthesis

- **Manual Selection :** the user can select any compounds by clicking on it. This more unrestricted approach for adding single compounds to the collection .
- **Sphere Selection (threshold based) :** this mode allow the user to select all compounds inside a sphere (in the original data space) by selecting two compounds . The first serve as the center of the sphere and the second is a located at the surface of the sphere. We can define the radius as the similarity measure between the center and the surface point. All compounds inside the sphere are selected (with similarity equal or less then to the surface point similarity). This is achieved with The EdenUtility 3.3.1 .

After The selection is made,the user can then perform manual refinement by either removing individual compounds or by applying Filters 3.3.4.1 to the selected compounds .

3.3.5.1 Graph Substitution Constructive Model(GraSCoM)

GraSCoM is the system we used for the compounds synthesis process . It addresses how to search for new compounds, how does it assemble them and how does it score new ones [32].First we need to explains the following terms :

- **interface border** For a vertex v, we define as the at radius r the set of vertices that are exactly at distance r from v

- **interface** rooted in v of radius r and thickness t is the subgraph induced by all the nodes inside the neighborhood of radius t of the vertices on the interface border
- **core** Given a graph G,the neighborhood subgraph of radius $r \in N$ induced by vertex $v \in V(G)$ as the of radius r with root v, denoted by C_r^v
- **core substitution** is the procedure for generating new chemical compounds

GraSCoM proposes a novel way for navigating over the chemical space. Given an input (train) set of valid molecules, for each molecule we iterate over the complete set of nodes $v \in V_{V(G)}(H)$ and extract for different radiuses and thicknesses interfaces together with their corresponding cores, and store them in a meaningful way into a database.At the end of this database computation step, we will have stored interfaces alongside with a list of cores that were found for each particular interface.

We then start with an initial seed set and again for every compound, we iterate over a set of vertices and using the same values for radiuses and thicknesses, we identify interfaces. If we find an interface that is stored into our database, we then get the previously computed list of valid cores for that interface and perform the core substitution procedure.

In this local search technique, the "width" of the neighborhood considered varies depending on how frequently the interfaces identified in the current molecule were found in the list of molecules used to populate the database. Thus, it could be considered as a special type of the adaptive neighborhood width approach, where the area of the neighborhood to consider is not necessarily controlled by an internal parameter of the system, but by the similarity to other valid molecular structures. Then compounds are assembled by substituting cores identified in valid chemical compounds and therefore the approach is fragment-based. However, with a radius $r = 1$ and thickness $t = 0$ it gets the capabilities of an atom-based technique. After performing a local search and assembling the new molecules, the algorithm needs to score them and then retain the best ones as a seed for the next generation.It uses an SVM 2.2.4 with an NSPDK kernel 3.3.1.1 and a stochastic gradient approach for learning.

3.4 Rendering and Performance Optimization

To be able to display large spaces in a web-based environment we had to optimize the seance structure and the rendering process.The main viewer was implemented using WebGL,a new Technology that allow Browser client to get access to the client GPU.WebGL is integrated completely into all the web standards of the browser allowing GPU accelerated usage of physics and image processing and effects as part of the web page canvas.

The compounds in the space were implemented in groups of Particle Systems instead of normal 3D Objects to save memory and allow for bigger spaces to run on low memory .A particle system is composed of one or more individual particles. Each of these particles has attributes that directly or indirectly effect the behavior of the particle or ultimately how and where the particle is rendered. Often, particles are graphical primitives such as points or lines, but they are not limited to this. Particle systems have also been used to represent complex group dynamics such as snow or fire.In our case the main goal was to decrease the number of primitives to visualize the space.

We also utilized the use of buffer geometry which saves all data in buffers. It reduces memory costs and cpu cycles. But it is not as easy to work with because of all the necessary buffer calculations. It is mainly interesting when working with static objects like in our case.

The last visualization optimization was the use of intersection acceleration techniques, Since we needed to perform a ray-object intersection to select object in the space by mouse clicking.Namely we used Uniform Grids. A uniform grid has the shape of an axis aligned box that contains a group of objects. The box is subdivided into a number of cells that also have the shape of an axis-aligned box and are all of the same size. Mainly a regular grid is a spatial subdivision scheme that divides part of the world space into a regular 3D cellular structure .each cell stores a list of objects that are in it, or partially in it, or maybe in it.Then we can simply find intersection by querying the grid in way that if the ray doesnt hit the grid then we dont check for any other intersection.

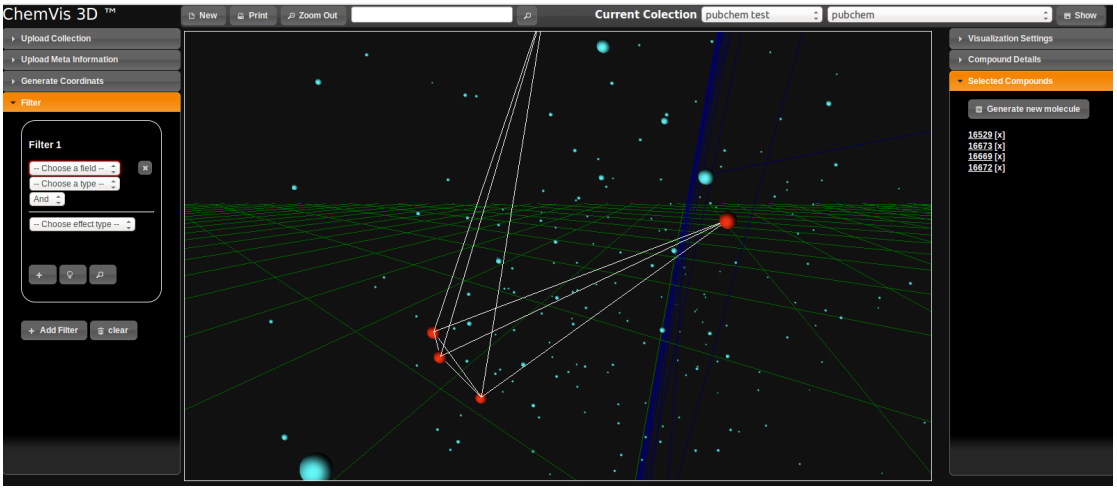


Figure 3.7: Main UI (SpaceExplorer)

3.5 GUI

In this section we show the main items of the Graphical User Interface (GUI).The UI was implemented in JavaScript and HTML5,while the visualization and scene rendering was implemented with WebGL.

3.5.1 Space Explorer

The Space Explorer is the main UI window where the Chemical Space is Visualized. From within the explorer the user can perform all the other tasks :

- Space navigating with different visualization settings
- Data/Meta-information Uploading
- Coordinate Generation
- Filtering and Highlighting
- Molecule Selection and Synthesis

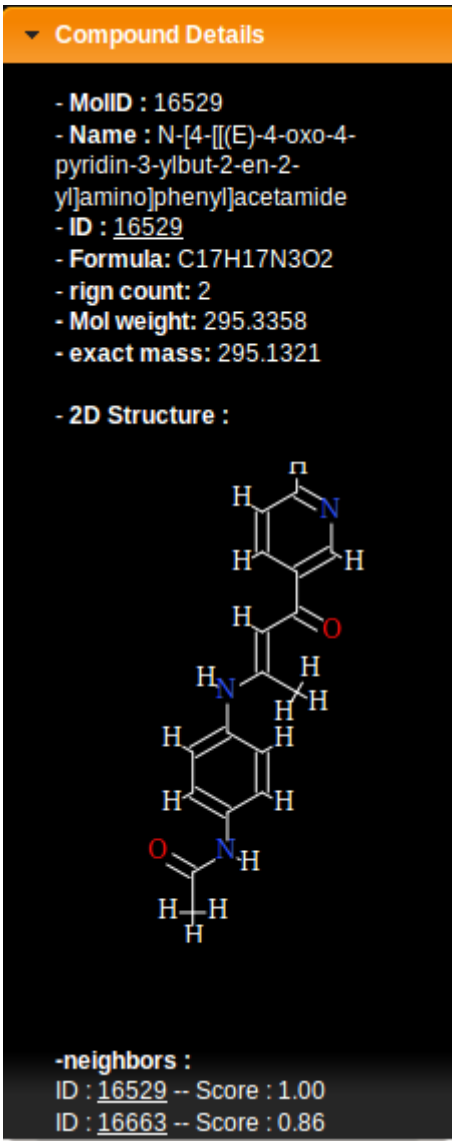
3.5.2 Quick and Detailed View

While working with chemical collections its important for the user to have easy access to individual compounds he is working on.this is achieved with the quick detail section in the right section of the explorer.the selected compounds overview and access to its top similar items is also possible.See Figure (3.8a).While the overview provide a quick and convenient way to see compound information,the detailed view give a more extensive view and display all information stored about the compound and its neighbors including 2d structure visualization. See Figure (3.8a).

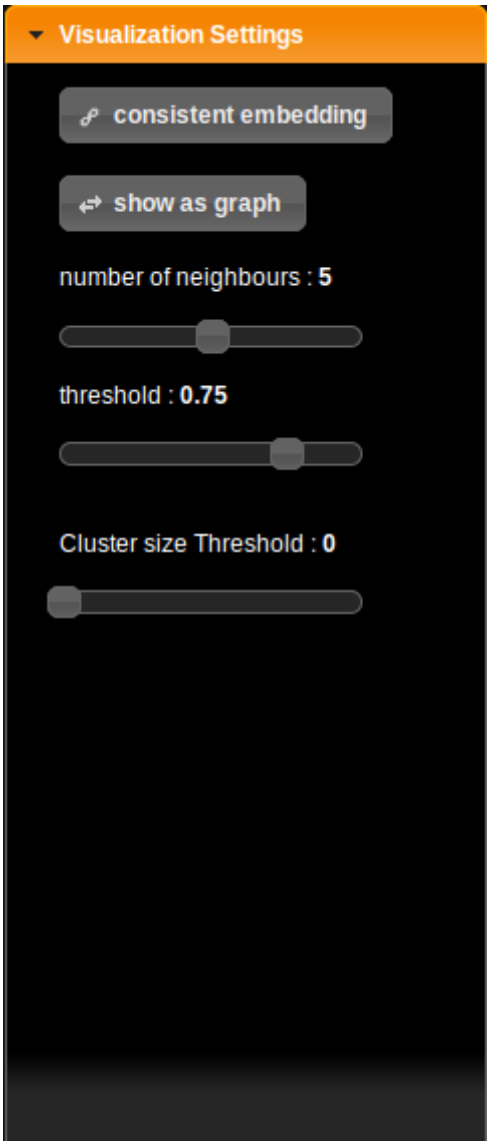
3.5.3 Visualization settings

The visualization setting panel (Figure 3.8b) allows the user to set and change visualization settings :

- **consistent embedding** the idea in this mode is to input all visible compound in the scene to the local embedding of the pre clustered collection, to keep consistency between inner cluster compounds embedding.
- **show as graph** while enabled, the space is visualized as a connected graph (similarity threshold and number of k-nearest neighbors to be connected can also be modified from this panel).
- **cluster size threshold** this slider allows the user to control visible clusters in the space based on the number of the compounds in the cluster.



(a) Quick Details UI



(b) Visualization settings Panel

Figure 3.8: Settings and Views

Chapter 4

Experiments

4.1 Tests and Evaluation

4.1.1 System Specification

For our experiments we deployed the tool to be served on a machine with following specification :

- Intel Core i5-2410M CPU @ 2.30GHz 4 CPU
- AMD Radeon HD 6400M Series GPU
- 4GB RAM
- Linux Ubuntu 13.04 - 64 bit OS
- Django test server version 1.5
- mongoDB version 2.4.6

The machine served as both the server and the client using .

4.1.2 Performance

In this section we tries to give a benchmark for the tool performance.First we analyze setting up the database for different collections.Then we take a look at the embedding performance and comparison of the different algorithm we have.

4.1.3 Uploading and Parsing

For our experiments we used different chemical compounds libraries with varying number of compounds (less then 200 up to 35,000) . Run times for parsing and storing in the database are summarized in the Table 4.1.the process also includes uploading the file from the client to the server.

Collection	**File Size**	**# Compounds**	**Processing Time**
illicit drugs (DrugBank)	537 KB	173	2.03 sec.
Approved Drugs (DrugBank)	5.3 MB	1424	12.68 sec.
All Drugs (DrugBank)	23 MB	6802	41.91 sec.
ChEBI	160 MB	31681	839.87 sec.

Table 4.1: Different collection uploading and parsing benchmark

4.1.4 Embedding

The parameters for the MDS algorithm where set to the Clustering and Similarity Measurement using with NSPDK kernel with radius : and distance : For the MDS we did not perform the test with larger datasets because of the similarity matrix calculation and size,this is why we moved to using PCA with sparse feature vectors. For the embedding experiments we show the performance of MDS :

Collection	**# Compounds**	**Similarity**	**MDS Time**
illicit drugs (DrugBank)	173	1.4 sec	2.86 sec.
Approved Drugs (DrugBank)	1424	20.82 sec	3 min 20 .
All Drugs (DrugBank)	6802	4 min 20 sec	1h 32min .

Table 4.2: Different collection MDS Embedding and Similarity Benchmark

In The Follwing Table the performance of PCA :

Collection	# Compounds	Clustering	PCA Time
illicit drugs (DrugBank)	173	8 sec	0.8 sec
Approved Drugs (DrugBank)	1424	28 sec	5.24 sec
All Drugs (DrugBank)	6802	6 min 2 sec	22.8 sec
ChBEI 30k	31681	56 min 23 sec	3 min 12 sec

Table 4.3: Different collection PCA Embedding and Clustering Benchmark

4.1.5 Use Cases

A typical use case would be to investigate a certain drug collection to possibly find intersting relation in the data through the visualization of the space. In this example we illustrate the steps for creating the space of the Illicit Drugs (a small compound collection (173 compounds)) and performing Filtering and Highlighting,Showing compounds as connected graphs and visualizing clusters through the quickshift clustering algorithm.Then performing Molecule synthesis on areas of the space where new drug candidate might exists.

After Uploading the collection file the user can embed the space with one of the following techniques : -MDS
-MDS with pre clustering
-PCA with pre clustering

4.1.5.1 MDS space embedding

In the Figure (4.1) Is the MDS space embedding results with no clustering

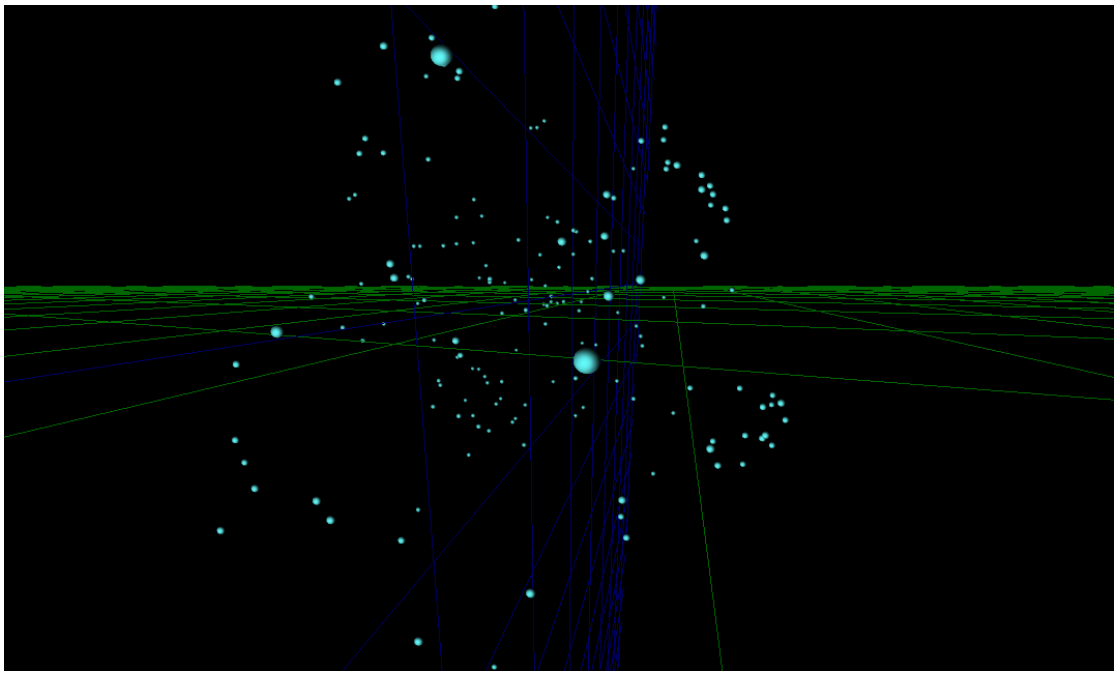


Figure 4.1: MDS performed on Illicit Drugs

4.1.5.2 MDS space embedding with pre clustering

In the Figure (4.2) Is the MDS space embedding results with pre clustering. In this visualization the clusters also show as connected graph above a similarity threshold of 0.75. We can see that nearbey cluster are connected to each other.

4.1.5.3 PCA space embedding with pre clustering

In the Figure (4.3) Is the PCA space embedding results with pre clustering

4.1.5.4 visualizing graph and clusters

Some interesting information could be found via showing the similarity information as connected graphs.The Figure (4.4) shows an area of the space with heavily connected compounds that might be of interest to investigate. another way is to

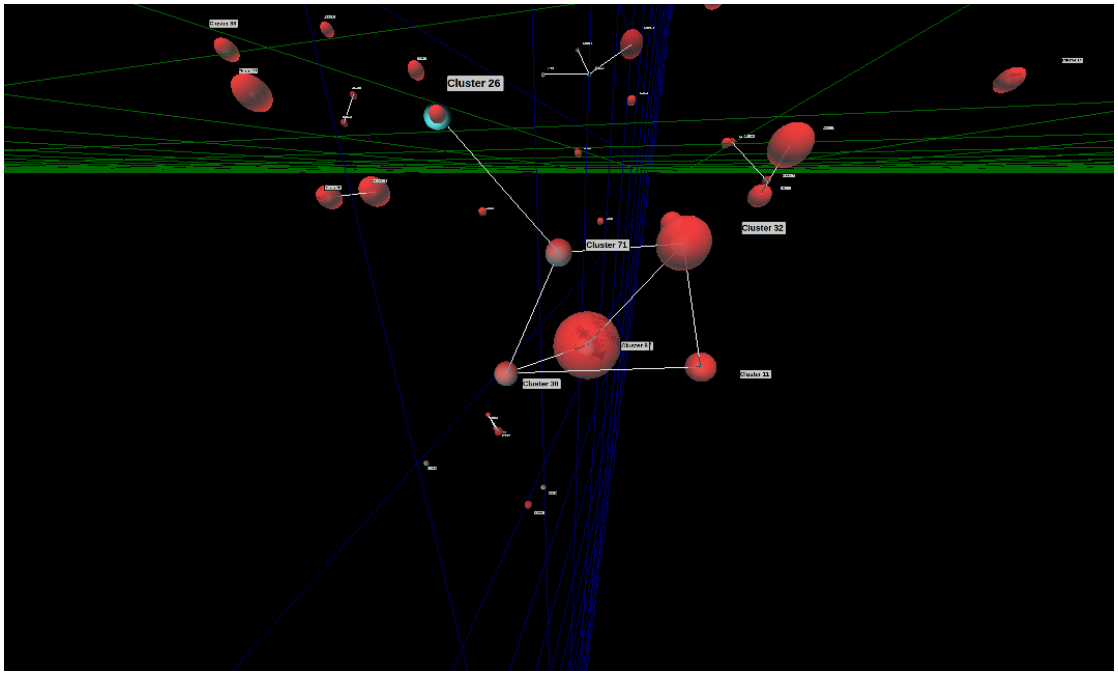


Figure 4.2: MDS performed on Illicit drugs with Pre Clustering

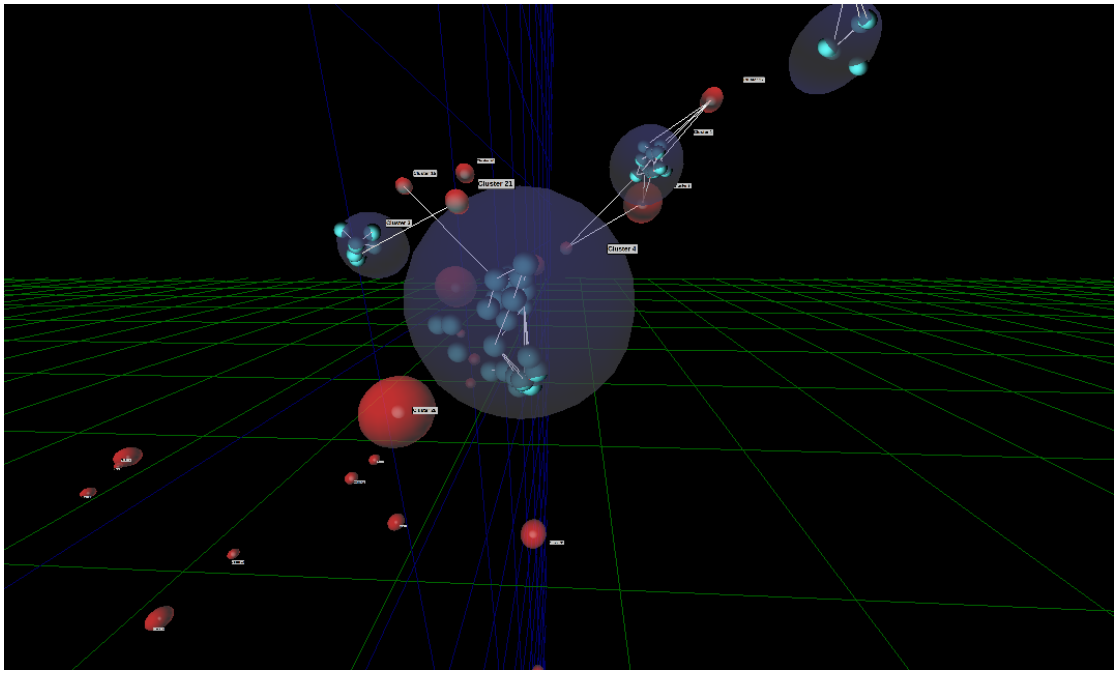


Figure 4.3: PCA Embedding with data pre clustering

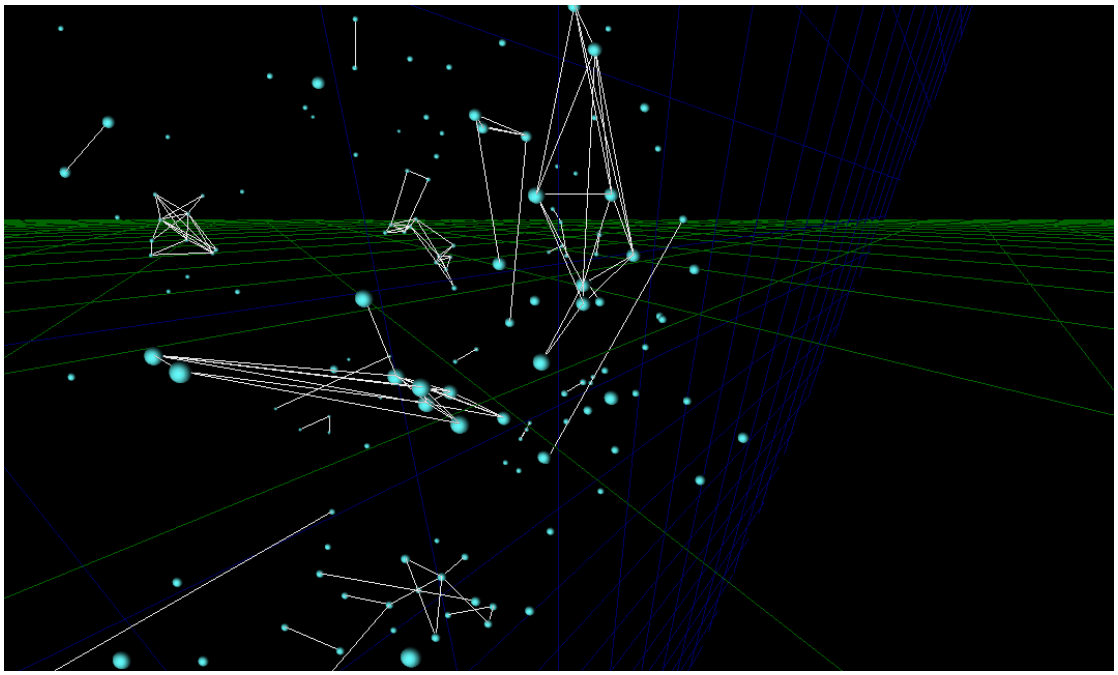


Figure 4.4: Graph visualization base on similarity threshold

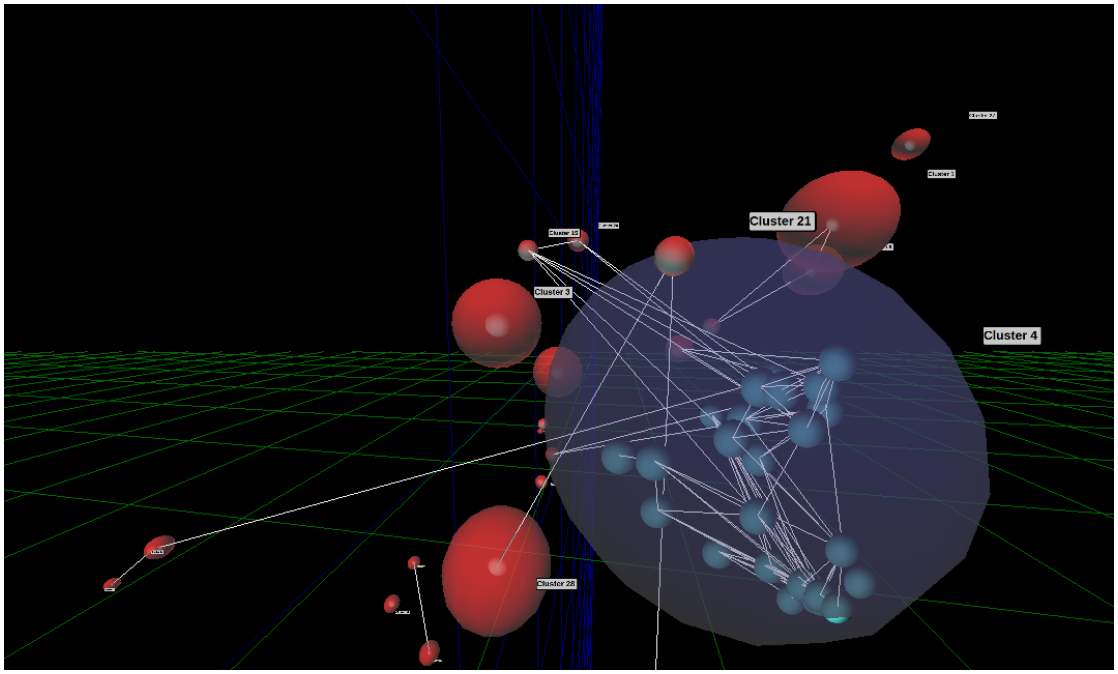


Figure 4.5: Showing graph inside of clusters

show the cluster result of quickshift clustering while mainlining the graph connection information.

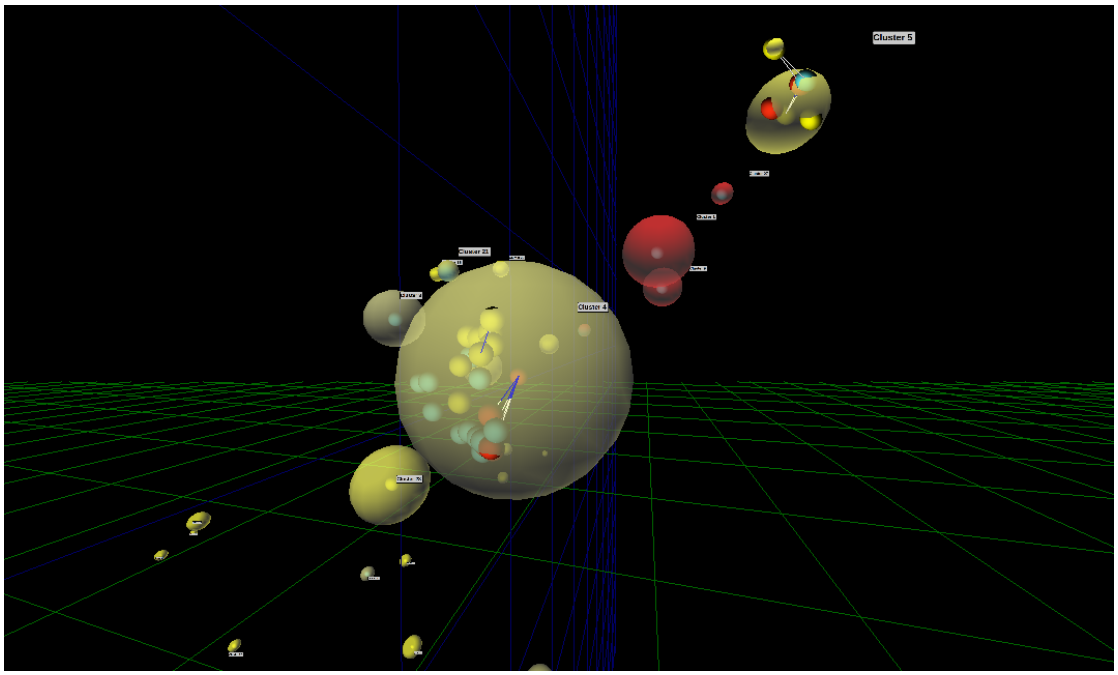


Figure 4.6: Cluster Filtering and Highlighting

4.1.5.5 performing filtering

At this point the user can perform filtering and visualize information like "what drugs target a certain protein and have the a ring count more then a threshold".Using the Filters ,the user can easily build such a query and assign a highlight effect for the data that matches the rules.In the Figure 4.6 we can see the highlight result in which a mor saturated color means more compounds of that cluster match the filter rules.

4.1.5.6 Selection and gap filling

After deciding on area of interest the user can the select a group of compounds as an input for the De nove synthesis process.As discussed earlier these can be selected manually or based on threshold . They can also be refined and filtered. The figure 4.8 show a result of this operation.Where in the first raw we see the compounds that were selected and in the second raw are the new molcule based on the output of Grascos.

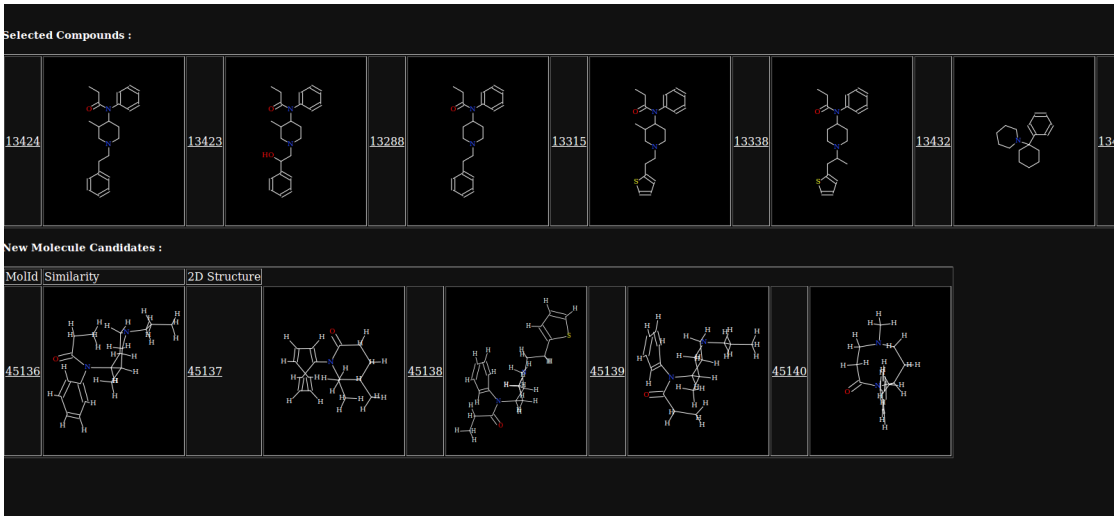


Figure 4.7: New Molecular design window

Figure 4.8: New Molecular design window

Chapter 5

Conclusions

5.1 Conclusions

The idea of the project was to produce a visualization tool that can help chemists to get helpful insight to the chemical space and the drug design process.Our tow main goals were : 1) to implement a visualization tool for the chemical space that can also visualize clusters of compounds in the space.2)to be able to perform selection interactively of a sub parts of the space to form a base for the de nove molecular design process. Both goals were me and we implemented an easy to use web-based tool that can perform these two tasks.

Our experiments shows the abilities of the tool to visualize the chemical space with different techniques,cluster collections and visualize similarity as graphs.Filter and Highlight parts of the space with the use of flexible query building ability and dynamically added meta information.Also trigger the monocular synthesis process by selecting parts of the space and displaying and comparing new compounds candidates.

But remain to be tested is how meaningful insight such a tool can be an actual drug discovery process,where test and specially crafted test cases need to be carried out with the help of area knowledge in this area.Only then we can investigate the true benefits of such a tool.

5.2 Future works

We like to improve upon the system capability to handle very large compound libraries (in millions). Also we would like to move into a plugable architecture that would allow extensibility of the system. Many tools and programs could be run over chemical compound and a plugable architecture help a lot. We also would like to generalize the tool to handle more biological data or generic data where similarity can be analyzed

Bibliography

[1] D.A. Keim. Information visualization and visual data mining. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):1–8, Jan 2002. ISSN 1077-2626. doi: 10.1109/2945.981847.

[2] OpenCV. Introduction to support vector machines, 2014. URL `http://docs.opencv.org/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html`.

[3] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[4] ChemAxon. Chemical hashed fingerprints, 2014. URL `http://www.chemaxon.com/jchem/doc/user/fingerprint.html`.

[5] Fabrizio Costa and Kurt De Grave. Fast neighborhood subgraph pairwise distance kernel. In *Proceedings of the 26th International Conference on Machine Learning*, pages 255–262, 2010.

[6] Ana G Maldonado, JP Doucet, Michel Petitjean, and Bo-Tao Fan. Molecular similarity and diversity in chemoinformatics: from theory to applications. *Molecular diversity*, 10(1):39–79, 2006.

[7] Joseph A DiMasi and Henry G Grabowski. The cost of biopharmaceutical r&d: is biotech different? *Managerial and Decision Economics*, 28(4-5):469–479, 2007.

[8] Gisbert Schneider and Uli Fechner. Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*, 4(8):649–663, 2005.

[9] Wikipedia. Information visualization — wikipedia, the free encyclopedia, 2014. URL `http://en.wikipedia.org/w/index.php?title=Information_visualization&oldid=600608389`. [Online; accessed 25-March-2014].

[10] Jean-Daniel Fekete and Catherine Plaisant. Interactive information visualization of a million items. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*, pages 117–124. IEEE, 2002.

[11] Wikipedia. Dimensionality reduction — wikipedia, the free encyclopedia, 2014. URL `http://en.wikipedia.org/w/index.php?title=Dimensionality_reduction&oldid=593115239`.

[12] Scikit. manifold learning, 2014. URL `http://scikit-learn.org/stable/modules/manifold.html`.

[13] Wikipedia. Standard deviation — wikipedia, the free encyclopedia, 2014. URL `http://en.wikipedia.org/w/index.php?title=Standard_deviation&oldid=602381966`. [Online; accessed 4-April-2014].

[14] Lindsay I Smith. A tutorial on principal components analysis. *Cornell University, USA*, 51:52, 2002.

[15] Wikipedia. Eigenvalues and eigenvectors — wikipedia, the free encyclopedia, 2014. URL `http://en.wikipedia.org/w/index.php?title=Eigenvalues_and_eigenvectors&oldid=599706948`. [Online; accessed 25-March-2014].

[16] Ali Ghodsi. Dimensionality reduction a short tutorial. *Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada*, 2006.

[17] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[18] Wikipedia. Multidimensional scaling — wikipedia, the free encyclopedia, 2014. URL `http://en.wikipedia.org/w/index.php?title=`

Multidimensional_scaling&oldid=594821255. [Online; accessed 27-March-2014].

[19] Forrest W. Young. Multidimensional scaling - , university of north carolina, 1985. URL http://http://forrest.psych.unc.edu/teaching/p208a/mds/mds.html.

[20] Corinna Cortes and Vladimir Vapnik. Support vector machine. *Machine learning*, 20(3):273–297, 1995.

[21] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1, 2007.

[22] Jose L Medina-Franco, Karina Martínez-Mayorga, Marc A Giulianotti, Richard A Houghten, and Clemencia Pinilla. Visualization of the chemical space in drug discovery. *Current Computer-Aided Drug Design*, 4(4):322–333, 2008.

[23] Christopher Lipinski and Andrew Hopkins. Navigating chemical space for biology and medicine. *Nature*, 432(7019):855–861, 2004.

[24] Christopher M Dobson. Chemical space and biology. *Nature*, 432(7019): 824–828, 2004.

[25] Jean-Louis Reymond and Mahendra Awale. Exploring chemical space for drug discovery using the chemical universe database. *ACS chemical neuroscience*, 3(9):649–657, 2012.

[26] Miklos Feher and Jonathan M Schmidt. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *Journal of Chemical Information and Computer Sciences*, 43(1):218–227, 2003.

[27] Josefin Larsson, Johan Gottfries, Sorel Muresan, and Anders Backlund. Chemgps-np: tuned for navigation in biologically relevant chemical space. *Journal of natural products*, 70(5):789–794, 2007.

[28] Richard A Houghten, Clemencia Pinilla, Marc A Giulianotti, Jon R Appel, Colette T Dooley, Adel Nefzi, John M Ostresh, Yongping Yu, Gerald M Maggiora, Jose L Medina-Franco, et al. Strategies for the use of mixture-based synthetic combinatorial libraries: scaffold ranking, direct testing in vivo, and enhanced deconvolution by computational methods. *Journal of combinatorial chemistry*, 10(1):3–19, 2007.

[29] Modest von Korff and Thomas Sander. Toxicity-indicating structural patterns. *Journal of chemical information and modeling*, 46(2):536–544, 2006.

[30] Jure Zupan and Johann Gasteiger. *Neural networks in chemistry and drug design*. John Wiley & Sons, Inc., 1999.

[31] RA Lewis and PM Dean. Automated site-directed drug design: the formation of molecular templates in primary structure generation. *Proceedings of the Royal Society of London. B. Biological Sciences*, 236(1283):141–162, 1989.

[32] Drago Alexandru Sorescu. De novo molecular design using graph kernels, 2012.

[33] VLADIMIR I Minkin et al. Glossary of terms used in theoretical organic chemistry. *Pure and Applied Chemistry*, 71(10):1919–1981, 1999.

[34] Wikipedia. Cluster analysis — wikipedia, the free encyclopedia, 2014. URL `http://en.wikipedia.org/w/index.php?title=Cluster_analysis&oldid=598628819`. [Online; accessed 4-April-2014].

[35] Makoto Hirohata, Kazunori Imoto, and Toshimitsu Kaneko. Knn kernel shift clustering with highly effective memory usage. In *MVA*, pages 393–396, 2011.

[36] Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. In *Computer Vision–ECCV 2008*, pages 705–718. Springer, 2008.

[37] DrugBank. Drugbank, 2014. URL `http://www.drugbank.ca/`.

[38] The pubchem project. The pubchem project, 2014. URL `http://pubchem.ncbi.nlm.nih.gov/`.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen/Hilfsmittel verwendet habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe. Darüber hinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, bereits fr eine andere Prüfung angefertigt wurde.

Ort, Datum:

Unterschrift:
