Global Structural Similarity in Chemical Compounds

A Thesis
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Master of Science
in Computer Science

By

Ahmed Hamza, B.Sc.

Washington, DC
January 29, 2010

Dedication

In memory of Marco (Marcel-Paul Schützenberger), of combinatronics, computation, and miracles.

Acknowledgements

I would like to thank my thesis advisor, Dr. Bala Kalyanasundaram, and my committee members Dr. Mahendran Velauthapillai and Dr. Brian Blake, for their tireless help in producing this thesis, and for encouraging me to pursue their original idea. This work would not have been possible without their continued guidance and inspiration. I would also like to thank Dr. Mark Maloof for his ceaseless help and support in the past two years.

Table of Contents

Chapter 1

Introduction

1.1 Background Information

Quantitative Structure-Activity Relationship models (QSAR for short) seek to provide an empirical, semi-empirical or theoretic basis for estimating the physiochemical properties of chemical compounds. That is, this area of chemo-informatics aims to produce a modeling of the relationship between the structural elements of a pure chemical compound, and the properties (physiochemical, in this nomenclature, as in capable of affecting the compound's physical behavior) it has.

Knowledge of these properties allows chemists to predict aspects of a compound's behavior without extensive experimentation on the compound beforehand, making these methods a valuable tool for environmental, pharmaceutical and molecular-engineering domains. The field arose to cater to the needs of these applications. Chemists do not always have the luxury of being able to run thousands of experiments on the compounds they encounter or wish to use. In fact, in the molecular engineering domain, the compounds may have *never* existed before in the natural world, even though they comprise elements that are present (perhaps abundantly) in the water, the air or even in other engineered compounds. Chemists do not also have detailed theoretical means of directly calculating the effects of complex structure on biochemical activity. Modern physics certainly explains most observable phenomena, but extending basic science to large scale structures and their

macroscopic behavior is not a straightforward task, so scientists try to estimate models from simpler descriptions of chemical structure, where concrete models do not exist.

The premise here is that chemical structure, the presence and arrangement of elements, their bonding, and the individual properties of certain chemical groupings, do have an effect on the activities and physical characteristics of chemical compounds, and that these effects can be estimated.

While several approaches and various property-specific models have been devised, there does not currently exist a property-independent process for inferring physiochemical properties from structure. Also, there does not exist an algorithm for comparing two given compounds as a whole (as opposed to comparing them based on the presence of various local groups), yielding some similarity measure based on their structural characteristics alone. What we are referring to here is the notion of *global* similarity—how close any two compounds are from being structurally identical. Most of the work in the field, as we shall see in the following section, does not attempt to approach this problem, but rather focuses on previously established chemical models of the activity of certain *local* groups. However, a wealth of structural and statistical information can be derived from available descriptive databases of compounds, which can be used in more generic methods that do not necessarily have to rely on prior chemical 'knowledge'. The quotes there are necessitated by the fact that we are inevitably using chemical knowledge. We can however, attempt to rid ourselves of a modeling process that relies on an empirically based parameter choosing scenario.

It may be possible to have predictive methods that do not know much about the specifics of certain chemical groups and structures (like say, phenyl rings), but are able to generically parse the structural form of chemical compounds encoded in some standard

input language, and analyze them. This work describes the development of such a method, and investigates its utility in the field on a wide range of molecules.

1.2 Related Work

The problem of determining chemical and biological activity profiles for a compound, without detailed experimental data on the said compound, is a well known, central issue for chemical and materials engineers in several related domains. Computer-Aided Molecular Design in general (and medicinal drug design in particular) seeks accurate knowledge of various properties of compounds that may not have been synthesized, but which can be related to other existing compounds whose properties are known. Similarly, environmental chemistry tries to easily estimate factors such as the toxicity of a given compound (found, say, in the water somewhere), with little data available other than the chemical structure of the compound. In fact, the Environmental Protection Agency (EPA) is directly involved with research in the field, and has made public a set of computational tools[1] used internally for the estimation of a wide variety of properties, based on some of the better known works discussed below.

To establish a relationship between any descriptors that can be derived from chemical structure, and the biochemical properties of interest, the field of QSAR was developed. Now, scientists have been trying to understand biological and chemical activity long before the term 'QSAR' was coined, or before the notion that statistical and computational studies of these activities belong to a particular division of chemical research. So this is a very difficult subject to condense into summarial form – in a sense the work here involves all of chemical engineering, results from basic studies in biochemistry and even modern physics. Results of literature in these fields provide a basis of how we understand the physical fundamentals of molecular behavior.

Computational biochemistry tries to simulate this behavior numerically, with statistically derived models, to overcome the lack of direct knowledge of the complex interactions

between structure and activity. If a structural feature (like, say, the number of hydrogen bonds in a compound) is incidentally found to correlate in some way with biochemical behavior, a Quantitative Structure-Activity model can possibly be developed. There is of course a large number of possible approaches to this development. Researchers have a tendency to mix a variety of approaches and methods in a single 'QSAR' study; so it can be quite taxing, especially for someone outside of biochemistry, to categorize the various efforts into different methodologies, or to lay down things in terms of discrete lines of research.

Suffice to say that several types of structural descriptors, and categories of models involving them, can be produced for any one biological or chemical activity. We will examine things from a general point of view here (i.e., as related to any activity/behavior), which suits our intended 'generic' methodology in assisting with this field. The classical prospect is that we have several compounds that form the rows of a table, the compound descriptors being the columns, and the scientist wishes to produce a relation of the form:

$$Activity = \sum_i c_i P_i + b. \tag{1.1}$$

for some set of numerical structural descriptors $P = P_1, P_2, \ldots$, where the values for the activity and the descriptors are known for the compounds in the table. Coefficients c_i are real-valued (continuous) and so is the constant b. This is a linear model; one where the variation in activity is expressed as a linear relationship with the chosen structural properties. There is no reason why the model should be linear. There is also no fundamental reason why it should be otherwise; but this is the simplest formulation, and yields decent fits to data in many cases [4], when a relationship seems to exist.

Equations of this type are the basic formulation of a QSAR, although in practice the skewness of biological data leads to logarithmic functions being introduced in the estimation of most activities.

Some of the earliest work resulted in the Hansch equation [12], which involved limited structural parameters (presence or absence of certain features at certain locations on a common compound) and others like molar refractivity and the hydrophobic parameter π. Ideas were developed further by Free and Wilson[17], who built matrices representing additive series similar to the classic one above, but relating the activity to the presence or absence of functional group substituents at every position in the compound. Most of the related work done here (see [13, 10, 17, 15, 7])involved the same limitations as the Hansch approach, where the choice of 'template' limited the kinds of studies that could be done, and there was little accounting for the effects of substituents in these equations interacting with each other.

This is representative of one approach to QSAR development, in which P is a set of structural parameters (e.g: geometric volume) that are chosen (usually due to prior chemical knowledge) as quantitative substituents in the equation, and an attempt to relate them by statistical means is made. This process of modeling gives values to the coefficients, and the parameters become a linear model of the activity, assuming they can be calculated/found more cheaply than the activity itself we are trying to estimate.

The other main line of research deals with the properties of functional groups, which are small molecular subsets of (attachments to) the overall structure of a molecule that are found to have some transient properties across a range of compound types where they occur. It is still a process of linear modeling, but substitutes calculated structural parameters with the immediate effects of molecular sub-structures.

One of the most widely used methods that utilizes functional group effects is the Joback method. In 1987, Joback and Reid [14] provided a very simple and powerful list of QSAR models to determine 11 different, pure component, thermodynamic properties of compounds (normal boiling point, normal melting point, critical volume, critical temperature,

Gibbs Energy of formation, etc.,). The seminal work was based on the idea of group contributions - each functional group in the list carries some 'contribution' value towards the estimated property, which is obtained by experimentation on a limited database of compounds for which these local functional groups have known effects.

There is considerable literature stemming from this notion of group contributions. Some of the problems the simple models cannot overcome is the fact that linear models do not scale well for all activity types, 'local' or additive consideration of parameters does not take into account their possible effects on one other (dependence relationships), and any model (like UNIFAC)[4] trying to take these things into account quickly faces large computational hurdles due to the combinatorial complexity that results.

Modern approaches tend to involve a greater emphasis on pattern matching techniques - some recent efforts like [5] and [8] take a data-mining perspective on the QSAR problem, by finding frequent patterns/trees that subsequently become training features for a classifier. More recently, the Jurs group has argued the use of Local Lazy Regression as a means of limiting the neighborhood of compounds that a model can be built from. Some small subset of the compound database may be the ones most relevant to the molecule at hand. A global model based on all compounds may simply be unable to capture these details. If we wish to build models from such local neighborhoods, a generic measure of locality is needed, and this is what we propose in Chapter 2.

From the computer science side of things, our (approach to this) problem presents itself in a more general form as a graph inclusion, graph isomorphism, and tree matching challenge, which are fundamental (and computationally intractable) algorithms in theoretical computer science. The closest formulations to our problem involve trees rather than graphs, particularly heuristic methods (not boolean/strict matching). A good summary of work in the field is provided by Philip Bille on Tree Edit Distance [3], and related problems like

largest common subsequences, largest common subtrees, largest common point sets and tree-to-tree correction (see [21, 2, 6].

Tree edit distance and tree-to-tree corrections are particularly of interest to us, as they posit ways to determine how different trees are from each other (dynamic programming and other efficient algorithm implementations are used throughout the literature). Trees that are least different from each other can be said to be most similar in structure, but a 'penalizing' scheme such as the one employed by those algorithms leaves little room for heuristics (involving structure) compared to what we will be doing in later chapters. They are designed for simple class-labelled tree nodes, with binary matching at the node level based on those labels.

By taking the opposite approach (trying to build up a maximum score rather than move down from it) we are free to use several heuristics at the node level, so that node comparisons do not result have to result in only a 1 or a 0. Comparing the trees against themselves can then yield values used to normalize (i.e go from some integer score to a percent matching value).

Also of particular interest to us are the Structure Search techniques developed by the NCBI (National Center for Biotechnology Information) for the PubChem database [19]. There are two types of searches provided: identical structure, and similar structure search. Identical structure uses different notions of chemical identity to the input/query structure to give results. The user can combine these measures to give different results, but in each case the structures found have to be *exactly* the same as the input structure in terms of the identity measure. Examples of these measures are 'Any Tautomer' (tautomeric form invariance) and 'Same Isotopic Labels'.

The similarity-type searches are closer to what we are trying to achieve, but local in nature. The similarity is pre-calculated and pre-linked in the PubChem database, and is

based on 'binary fingerprints' of the chemical structures. A fingerprint is a set of binary values that corresponds to a *fixed* set of yes/no questions called keys [19]. Each key is a test of the presence of a certain chemical substructure in the compound. The similarity of two structures producing fingerprints F_A and F_B is computed as:

$$Tanimoto = AB/(A + B - AB) \tag{1.2}$$

where A is the count of positive bits in F_A, B is the same for F_B, and AB is the count of the total positive bits after a bitwise AND operation on the two fingerprints. This measure (called the Tanimoto coefficient) is a similarity heuristic for binary sets that extends the basic Jaccard index – a cosine similarity metric for sets [20]. The use of Tanimoto here relies on the fingerprints, therefore solely on the features comprising those fingerprints. They are a comprehensive set (around 800 features), but they are not exhaustive, and as a result this search mechanism can lead to instances where the structures returned are notably different. They will be similar in terms of the features being tested, but different in the substructures not being compared.
This is a problem a global structural similarity measure can avoid.

1.3 A Brief Introduction to SMILES Notation

For our approach, we will assume a standardized input method for the chemical compounds - which may not necessarily be organic - providing the necessary structural information for a comparison. SMILES (Simplified Molecular Input Line Entry System) does precisely this. It is a representation language that encodes linearly the structural information of a compound as ASCII strings, and is supported by many of the available software toolkits available from the EPA and so on. We provide a very brief introduction to the syntax below; the

reader will find a much more comprehensive description of the SMILES language compiled by the Daylight Chemical Information Systems[18], and in the original specification by D.Weininger[22].

SMILES strings denote the following information: atoms, bonds, charges, branching information, cyclic connections, and additional punctuation to represent things like Chirality (which is non-apparent in a two dimensional structure). Each atom is denoted by its atomic symbol (C, N, Br and so on). In most cases, hydrogen atoms are omitted in the input string, and must be inferred from the valency values and bonding of whatever atoms they are connected to. Bonding information is also implicit in cases of single bonds between atoms, but is represented by the characters `"-"`, `"="`, `"#"`, and `":"` (single, double, triple and aromatic bonds, respectively).

A bond type symbol separates two atoms that are connected by it. Branches occur when a single atom is connected to several atoms, which may themselves be complex subtrees, and is achieved by enclosing the subtrees in parentheses. Square brackets, in contrast, are used to group small molecular groups together such as $[CH3-]$. These are often aromatic atoms and ions, as in the last example (the carbon has not achieved full valency and so the molecule has an overall negative charge). Finally, cyclic connections are addressed in the linear representation of SMILES by having numbers act as "flags" for a bond's initial and final locations. For example, 'C1CCC1' represents a ring where the first and last carbon atoms are connected. If more of these so called cycles are needed, they are simply denoted by different number at the desired locations. This way, several cycle–bonds can involve the same atom. The molecule $CC12CC1CC2$ (this probably doesn't have a chemical name) is an example.

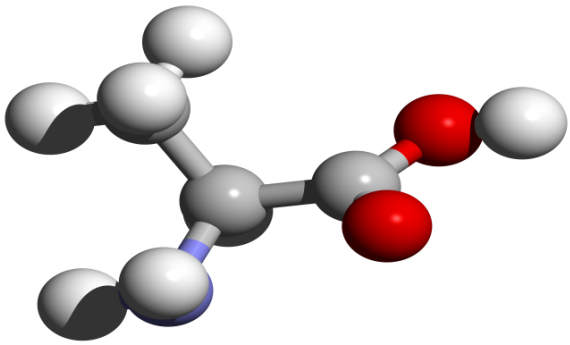

Figure 1.1: Ray Traced POV Image of $N[C@@H](C)C(=O)O$ with RasMol Source

1.3.1 Examples of Graph–ed SMILES Molecules

Most bio/chemo–informatics practitioners will be familiar with the various chemical notations and the associated 3D coordinate models, but computer scientists outside this exciting field of algorithm applications are probably going to benefit from a few illustrated examples, which we provide here. As we delve into the SMILES strings, we will take a moment to outline the technologies used here to produce these graphics. SMILES notation is very widespread as mentioned before, and many molecule editing programs and similar software will use SMILES for input. The graphics used here are produced through the following workflow:

- Corina 3D. Available online from Molecular Networks GmbH[11], this reads in a SMILES string and produces a 3D coordinate graphic (and a downloadable pdb file).
- SwissPDB–viewer and RasMol.

  These are almost alternatives; their purpose here is to view the data we've saved in PDB format and allow us to make some visual changes, before saving to a ray-trace friendly format. This format is POV3, currently. The difference between them is that Deepview provides better language support for POV3, but produces backbone–type molecule depictions as in X, because it is really meant to model protein ribbons (which is not what we are using it for). Rasmol [16] produces the ball and stick figures we will be seeing for most of this literature. We may revert to backbone/wireframe molecules in some particularly complicated examples, for better view of all atoms.

- POV-Ray software: this final step is where the image files (usually bmp files) get produced as high quality ray-traced images. The settings for high quality image production are beyond the scope of this document, but suffice to say that anti-aliasing, recursive sampling and 'high quality' are set in the scripts that POV-Ray uses to draw the images. The output can then be converted into suitable means for publication using Adobe Photoshop or similar.

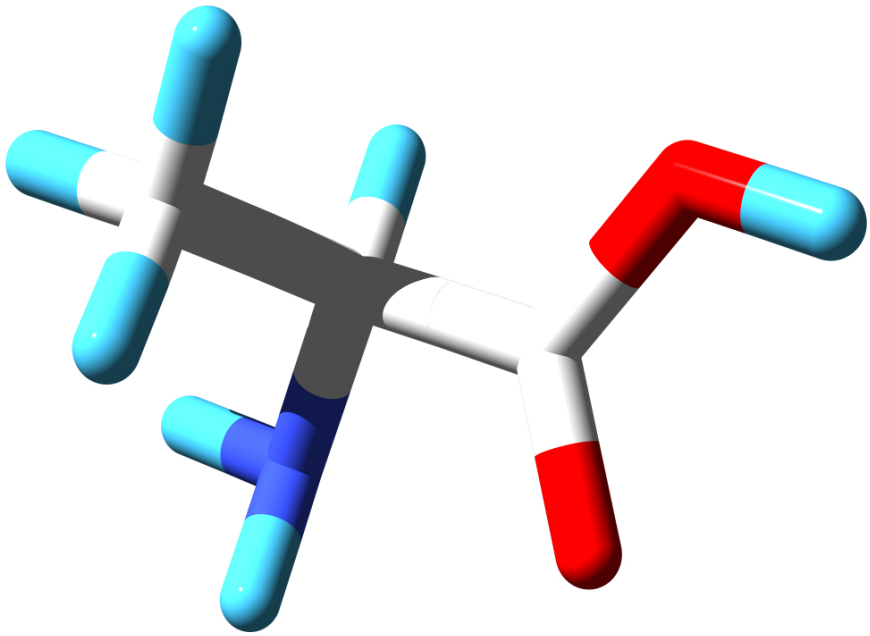

Figure 1.2: Ray Traced POV image of $N[C@@H](C)C(=O)O$ with Deepview source

Now let us consider a simple example and how we annotate it/build it up for our purposes. The Ethene molecule is a good candidate for this – it has only two carbon atoms and is represented by the SMILES string

$$C = C$$

where the four necessary hydrogen atoms are filled in by implication. The compound is shown in Figure 1.3. We can index the nodes in the order they are added to our graph, so that the carbon atoms have indices 0 and 1, and the hydrogen atoms have indices 2, 3 for the ones attached to the carbon on the left, and 4, 5 for the ones on the right. We can assign the indices any way we like, as long as we adopt the same algorithm for all molecules/compounds that we construct.

1.4 Research Hypothesis

Our working hypothesis is that SMILES representation of chemical compounds encodes enough suitable structural information to infer biochemical properties. Given the bounds on branching structures in chemical compounds, a mapping of compound graphs based on global structural similarity is both feasible and usable either on its own or as part of a larger process for QSAR modelling.

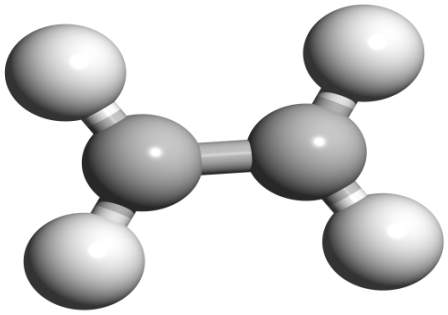

Figure 1.3: The Ethane Molecule

Chapter 2

Methodology

2.1 The Basic Approach

A schematic of our intended full procedure is summarized by Figure 2.1. The final goal in our particular approach to this line of research, is to create a predictive process for activity estimation – preferably a completely generic one, that does not, for some property we are trying to estimate, rely on the knowledge of *other* complex chemical properties with presumed effects.

The results of this process can vary in form. The predictive component can be a linear model (solvable by regression and other linear classification methods), a non-linear model, a more complex algorithm based on (perhaps) neural networks, or a much simpler one, that uses our similarity measure directly to make estimates. Each approach has its motivations. For our purposes, we want to explicity test the relevance of our generic notion of compound similarity in making estimates.

The clustering alone, in other words, may narrow things down enough for us to be able to work out a guess without using local group contributions or trying to estimate them. Even if this does not provide reasonable performance compared to a more complex modeling of the chemical properties, it is still a good way to evaluate the basic premise of this work; namely that similar molecules will on average have similar properties. It is also a good way to see whether it is meaningful to continue in this direction. Methods (generic

or property-specific) that want to build on the ranking output of this algorithm will want to assume some utility from the process.

So for our basic approach, we will rely on the global similarity measure for estimation. Our tests must be designed to include both similar compound comparisons and dissimilar compounds, to discover the extent to which something like this can be helpful (if there are cases of structurally similar compounds where this approach drastically fails, and vice versa).

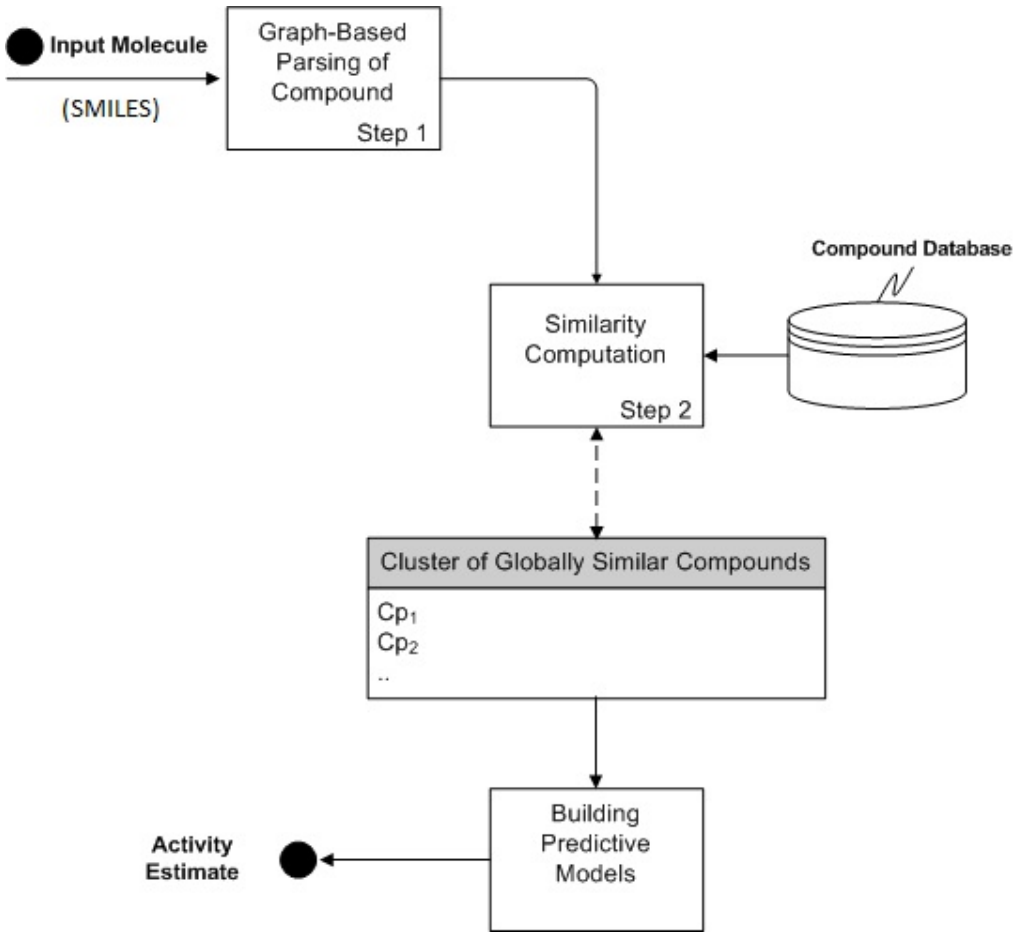

Figure 2.1: The Proposed Generic Structure Activity Prediction Process

2.2 Parsing SMILES Input Data

For a full review of the specifications of SMILES strings, we recommend the Daylight SMILES tutorial document [18], which goes through the various formulations of chemical structure in considerable depth. For our research we are limited by time to considering a large subset of these structural properties, which occur most frequently in describing pure chemical compounds. Our goal is to provide something that we can prove experimentally to be of use.

Our alphabet of input chemical symbols and groups (these are substrings of the SMILES string enclosed within square brackets, and can be treated almost as a single entity during the parsing phase) consists of the elements displayed in Tables 2.1 and 2.2. Almost all our available test data as seen in later sections are comprised of SMILES strings containing only these.

The parsing procedure is a fairly straightforward one in which we transform the string representing a single compound into a graph representing the three dimensional structure of that compound. Our motivation for this has been discussed in preceding sections; the graph structure facilitates the structure-mapping algorithm that is the subject of this research. Parsing proceeds from left to right taking each element (even the separate elements within a square bracket group) to be a single node, and storing valence, bonding, chirality and cyclical bond information for each of these nodes. In addition, there is another piece of data - the ordering of the element as encountered in the parsing - which allows us to label the elements with an index. We can then retrieve elements (and their properties) via this index from an array-like structure, without having to search for them in the graph we built. This array of parsed elements is central to all the forthcoming algorithms.

After the parsing procedure, we have a valency - balanced graph and an adjacency matrix for it, the elements of the matrix being the bond type. The adjacency matrix does not represent the cyclical links described in Section 1.1, or the additional structural details (like chirality) that are kept in the graph structure. However, it is sufficient for the purpose of calulating most node–distance metrics we will describe in the following sections of this chapter.

Element	Valence
H‡	1
C‡	4
N	3
O‡	2
K	1
Ca	2
Br	1
Cl	1
I	3
Fe	3
P	5
S†	6
Na	1
Mn	2
Se	6
Si	4

† It is worth noting (for complexity analysis) that is the highest valency we will likely encounter for a single entity in our graph.
‡ Very frequently seen elements in naturally occuring compounds.

Table 2.1: Elements in Parser Alphabet (Current Implementation)

We note here that there are several equivalent SMILES string formulations of the same compound, but our input assumes a unique, canonical representation. We can rely on the same compound being written in the exact same manner every time.

SMILES Group Symbol			
[*]	[2H]	[C-]	[C@@H]
[C@@]	[C@H]	[C@]	[CH2-]
[CH2]	[CH3-]	[CH]	[Ca2]
[Cl-]	[Co2]	[Co3]	[Co]†
[Fe2]	[Fe]	[H+]	[H]
[I$-$]	[K]	[Li]	[N+]
[N-]	[NH]	[N]	[Na+]
[O+]	[O-]	[S+]	[S-]
[Se]	[Si]	[nH]	

† Rarely Encountered

Table 2.2: Recognizable String Subsets in the Parser

2.3 The Compound-Mapping Algorithm

We will now discuss the tree-mapping algorithm that provides our notion of global structural similarity. The basic algorithm specification was provided by Prof Bala Kalyanasundaram. I have produced the implementation described in this literature.

We can represent the mapping process for two compound graphs with a 4-dimesional map structure, populated according to the rules 2.2. The value of map[i,j,a,b] is the similarity score of mapping the tree produced by nodes i and j (from one compound) to the tree produced by a and b from another compound. The construction of the trees is described below, but figures 2.2 and 2.3 illustrate the comparison that produces the score map[i,j,a,b].

$$map[i,j,a,b] = max \begin{cases} mv(i,a) + map[i_{next}, j, a_{next}, b] + BMT(i,a) \\ map[i_{next}, j, a, b] \\ map[i, j, a_{next}, b] \end{cases} \tag{2.1}$$

$$map[i,i,a,a] = mv(i,a) \tag{2.2}$$

Bear in mind that the graphs are now trees after the parsing process, since we have severed the cyclic connections (while maintaining bonding info). This means that there is exactly one path from any a to any b. The path is a set of ordered nodes $< a, a_{i_1}, a_{i_2}, ..., b >$, which we can store in a hashmap or similar structure for easy retrieval. So given an a and a b (or an i and a j) for one graph, we can retrieve any molecular element on that path in constant time.

The structure is 4-dimensional because we want to record the similarity between every tree in the first compound with every possible tree in the second, and any compound tree can simply be represented by a path between two nodes, from which all subtrees can be

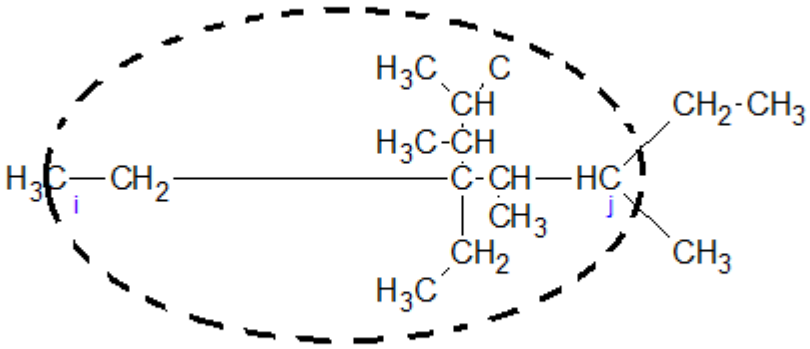

Figure 2.2: Sample compound tree and subtree T_{ij}. Nodes connected to i or j and not on the path are not part of the subtree. They are shown outside of the dotted line area.

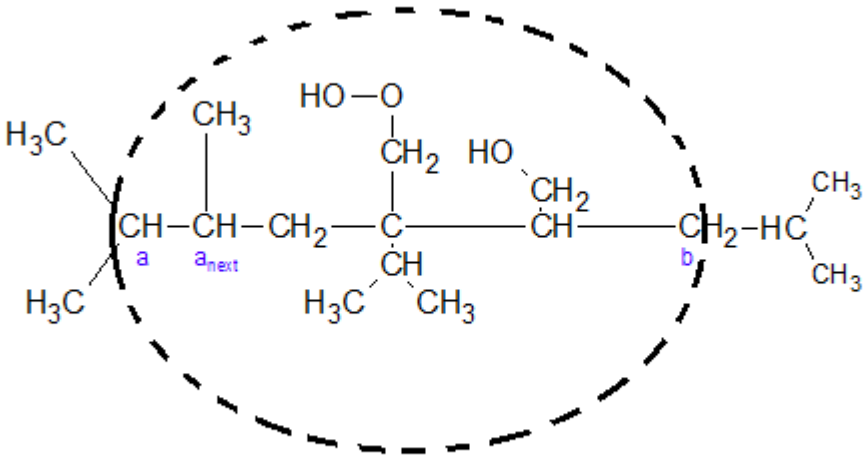

Figure 2.3: Sample compound tree and subtree T_{ab}. Nodes connected to a or b and not on the path are not part of the subtree. They are shown outside of the dotted line area.

drawn. Note also that since a tree includes all subtrees that originate on the path between two nodes, there can be several pairs (a, b) that produce the same tree. The path from which the subtrees will branch will be distinct, but the resulting overall tree does not have to be. For the simple ethene (alternatively called ethylene) compound illustrated in Figure 1.3, we can annotate the nodes of the graph by assigning each parsed element a number. We do this in the order of their construction by the parser:

$$\{0 = \text{C}, 1 = \text{C}, 2 = \text{H}, 3 = \text{H}, 4 = \text{H}, 5 = \text{H}\} \tag{2.3}$$

Now this business of 'paths' can be along the annotated compound tree can be confusing. The idea is to ignore, for some path $a, \ldots, b$ all the nodes attached to node a (and all the subtrees stemming from them) that are not on the route to b, and vice versa. They are simply eliminated. Everything else is either a node on our path of traversal $< a, ..., b >$, or something stemming from those nodes, and this forms a subtree. This forms a tree constructed from the path from a to b, but not unique to that path. Any other path from which a and b are reachable, according to our elimination scheme, will construct the same tree (or a larger one).

So if we take the path $(2, 3)$ or the path $(2, 4)$ using node numberings from 2.3, we can construct the same tree. It is only the order of traversal of the nodes that is different. Since we do not know which order of traversal will produce the best matching when trying to map a tree to another tree, we are forced to take all such pair configurations, in both compounds. Hence the four dimensional structure (a simple 4-dimensional array will do). We will come back (more thoroughly) to ethene as an example.

The BMT function (Best Matching of Trees) returns what is usually the most important value at each iteration/level of the comparison. This is because the call to mv(a, i) only compares one node with the corresponding one from the other tree, while BMT compares

every single sub-tree at a+1 to every single subtree at i+1. This is a dynamic programming approach, so the value of the sub-tree mappings has already been calculated (bottom up). In fact, the single node comparisons have already been calculated as well – they are the first thing stored in the array. Single node mapping values are stored as $map[i][i][a][a]$ for a node i in the first compound and a node a in the compound we are comparing to it.

In this bottom-up approach, sub-tree comparisons with lower combined diameters are done before subtree-comparisons with higher combined diameters. A 'diameter' of a tree is the length of the longest path (x_1, x_2) spanned by it. As explained above, a tree produced by nodes [a,b] consists of all nodes reachable from the path from a to b, excluding any structures branching at a and at b.

A combined diameter of two trees is simply the sum of the diameters of those trees. This means that no matter which 2 nodes we choose to represent the tree (we've already seen how different paths can result in the same complete tree structure) we can always sort the paths of the tree in terms of the largest diameter in that tree.

Indeed, this how we manage to satisfy the bottom-up constraint that map[i_{next}, j, a_{next}, b] will always be calculated before map[i, j, a, b]. Every map value for two trees is represented in terms of map values of (smaller) sub-trees. We need to have a secondary sort by tree size, again, combined. The sorting parameter becomes (for subtrees a-b and i-j)

$$diameter[i, j] + diameter[a, b] + \text{sum-of-nodes}$$

where sum-of-nodes is the combined total number of nodes in both subtrees.

If we do this, and populate the 4-D matrix in accordance with the rules of the general algorithm (Fig 2.2), then we can guarantee that the BMT function for subtrees at i_{next} and a_{next}, respectively, will not require a mapping value that has not been calculated.

Now we detail how the BMT portion of the algorithm works. It is not necessary to understand the details of the combinatory arrangements in the code (since we are dealing with complex data structures which do not necessarily have to be the same in each implementation, i.e., others may not use a simple 4D array), but understanding the purpose of the function is important as it is central to the algorithm. The reader is encouraged to keep in mind the basic functionality of $BMT()$, which is to return the best possible combination of mappings for sub-trees branching out at some point along the path comparison.

Algorithm 1 The Best Matching of Trees Function

1: $BM[][] \leftarrow nulls$
2: $BR_i \leftarrow getAllNonPathBranches(i_{next})$
3: $BR_a \leftarrow getAllNonPathBranches(a_{next})$
4: **for all** br_1 in BR_i **do**
5: **for all** br_2 in BR_a **do**
6: **for all** leaves ℓ_i in getleaves(br_1) **do**
7: **for all** leaves ℓ_a in getleaves(br_2) **do**
8: $BM[i][a] \leftarrow largest(map[i_{+1}][\ell_i][a_{+1}][\ell_a])$
9: **end for**
10: **end for**
11: **end for**
12: **end for**
13: $N \leftarrow |BR_i|$
14: $R \leftarrow |BR_a|$
15: $max \leftarrow 0$
16: $indices \leftarrow \{0, 1, 2, \ldots, N-1\}$
17: **for all** permutations P of size R in $indices$ **do**
18: $sum \leftarrow \sum_{c=0}^{R} BM[P_c][c]$
19: $max \leftarrow largest(sum, max)$
20: **end for**
21: Return max

In the above summary of the algorithm, we are assuming that the assignment of the N and R sizes always happens with BR_i being the larger of the two. We want to compare all possible arrangements, and each arrangement will be of the size of the smaller set. We

are drawing from the set of larger branches and mapping them to the ones in the set of smaller branches. So for some bigger set BR_i and some smaller set BR_a, we want to get the maximum mapping value of all ${}^{n}\mathrm{P}_{r}$ mappings. We are using an array like structure $BM[][]$ to hold the mappings, hence the naming of the permuted array $indices[]$.

For assignments in lines 9 and 10 to occur as required, we need to swap data structures, or we can resort to some equivalent action in the code (as is done for this implementation).

2.3.1 The Mapping Value Function

So far we have considered the algorithm at a high level without really talking about the mapping value at the atomic (most granular) level. How does one compare two nodes?

Nodes in our case translate to atoms in chemistry. The information stored for each one includes all the bonds it has with the neighboring atoms, their orientation with respect to it (if present)i, and the type of chemical bond.

One way to do things is to simply return a one if the element names match and zero if they don't. This will do, but we wish to make best use of our available information, so we add to the score as more elaborate features of the nodes match. The exact algorithm is shown in 2.3.1.

Thus we are setting element-name-matching as a basic requirement, and increasing the score as further similarities are found. There is of course much room for experimentation, as different feautures can score differently depending on their perceived importance. We do value chirality matches more than bond matches, for instance, since chirality info does not occur very often in the string, and changes the 3D structure of the molecule. This function allows us to differentiate between things like N and $N+$. Simple name matching will achieve that too, but only indirectly (because the atoms bonded to the nitrogen will be different due to the different valency). Similarly, a carbon atom connected via double bonds

Algorithm 2 The Mapping Value (Node Comparison) Function

```
1: mv ← 0
2: n_i, n_a ← nodes to be compared
3: if n_i.elem − name = n_a.elem − name then
4:    mv ← mv + 1
5:    if n_i.linkcount = n_a.linkcount then
6:       mv ← mv + 1
7:    end if
8:    if n_i.doublebonds = n_a.doublebonds then
9:       mv ← mv + 1
10:   end if
11:   if n_i.chirality = n_a.chirality then
12:      mv ← mv + 2
13:   end if
14: end if
```

will have a different number of links/bonds than one connected via single bonds, except in some rare cases where there is a mixture of bond types to differing neighboring atoms, with the same counts. Still, the mapping value for *those* atoms, to their corresponding ones in the other graph, will be different, impacting the total score at that point.

2.4 A Small Example

Here we will go through a complete mapping proceedure on a pair of small compounds to illustrate the process.

We will be using ethene (C2H4) and methane (CH4). The SMILES representations for these two are [C=C] and [C], respectively. Listing all the possible node pairs by index in both compounds, we obtain Tables 2.2 and 2.4.

Table 2.3: Possible Node Pairs and Resulting Spans in [C=C]

Pair	SubTree Span	Node Count
0-1	1	2
0-2	1	2
0-3	1	2
1-4	1	2
1-5	1	2
0-4	2	4
0-5	2	4
1-2	2	4
1-3	2	4
2-3	3	6
4-5	3	6
2-5	3	6
2-4	3	6
3-4	3	6
3-5	3	6

The summation of the combinations of pairs from the two lists gives us a single list, through which we iterate to compute the mapping values as in 2.2. The first list has 15 pairs (${}^6C_2 = 15$) and the second contains 10 (${}^5C_2$). This gives us a combined pair list of $10 * 15$ pairs, too many to list here. Starting from the top of the sorted list, they are ([0-1], [0-1]), ([0-1], [0-2]), ([0-1], [0-3])..etc., until the largest combined subtrees ([3-5],[3-4]).

The pairs in the tables are sorted by maximum spans of the trees they generate. In this case, the the maximum spans, or 'diameters' for short, are cleanly dividing the node counts, so a secondary sorting will not change the ordering of the pairs, but needless to say this is not always the case.

Table 2.4: Possible Node Pairs and Resulting Spans in [C]

Pair	SubTree Span	Node Count
0-1	1	2
0-2	1	2
0-3	1	2
0-4	1	2
1-2	2	4
1-3	2	4
1-4	2	4
2-3	2	4
2-4	2	4
3-4	2	4

Note that we have not included trees consisting of a single node. They serve as our base case for the dynamic programming, and their combinations with the single nodes in the second compound are computed as a first step.

Now back to the mapping process in the main algorithm 2.2. The BMT function is called at each stage to retrieve (not calculate but retrieve) the already computed best matching values of the subtrees directly after the starting nodes in the path, then performs permutations of them to get the best possible score of any arrangement at that point.

So for example, consider map[2][3][1][2]. The the subtrees available in the first compound are those on the branch path $0 \rightarrow 1$, those available in the second are on the branches $0 \rightarrow 3$ and $0 \rightarrow 4$. Thus there are 2 branches on the methane compound and one on the ethene at that mapping point. ${}^2C_1$ combinations (i.e., 2) are possible, and both have the same mapping value, as the branches on the methane compound both contain only hydrogen. This will map to one of the hydrogen atoms on the ethene branch, incrementing the score by the mv() value.

Since map[2][3][1][2] contains the subtrees spanning the complete first compound and the complete second compound, it will contain the maximum comparison score for these two molecules. There are exactly four elements that match, after skipping the second carbon in the ethene molecule. This gives a score of 4, if we were using a simple mapping value (counting matching elements only). If that were the extent of the mv() function, the [C2H4] mapped against itself would produce a score of 6.

Our mapping value is a little fancier however (recall algorithm 2.3.1) and leads to the total value being 13 (matching number of links on 3 hydrogen atoms and one carbon atom). The score of the [larger] ethene molecule against itself is 24. This brings the normalized final GSS value to

$$13/24 = 0.54166667$$

A value of 0.54 means the compounds are about 54% similar with respect to the larger one.

2.5 Implementation Details

Following will be a short description of the software technologies implemented for this work.

Our implementation is coded in the Java programming language, and consists of several stages of processing to go from the initial SMILES strings to the final answer(s). The main file is the StructuralMapping class, which is called with two arguments, being the SMILES compounds that we want compared. StructuralMapping contains the main high level logic described in the pseudo–code above. The output from the main program is a single floating point number, of a precision dependent on the version of the Java virtual machine, that is

the final normalized score, and an echo of the two SMILES strings which will aid in the parsing of results. We will come back to parsing results later in this section. A sample sun of the program can be achieved (while in the directory containing the class files) by typing the following at the command line:

java -cp . StructuralMapping "C" "C"

which will compare two methane (CH4) molecules.

StructuralMapping makes use of the Compound class (which encapsulates the graph represented by a SMILES string) and the DistanceMetrics class, which is responsible for performing the various searches and path constructions given adjacency matrices. Essentially, it keeps a hash structure of all the possible paths in the concerned compound, where a path is a list of nodes reachable from some node in order of traversal, without branching. Whenever there is a possibility for branching, there exists an alternative path. The storage of these paths, combined with the adjacency matrix for the graph, allows for the retrieval of branches at the construction of subtrees any point, and therefore the simplified implementation of the algorithm as described in the pseudo–code.

The auxiliary/utility classes used are briefly as follows:

- **Branch:** A Branch consists of one head node and all the leaves reachable from that head node along a directional path. The 'direction' is determined by the first node after the head node, which will be part of any path to leafin that branch.

- **Node:** This encapsulates a single atom, it's bonds, bondtypes, any cyclical links, valency and other properties related to the atom's structural role in the compound.

- **Permuter, Combinator:** These utility classes for performing combinatronics are publicly available, and were written by Hendriks Maryns. Permuter calculates permutations of the form ${}^{N}P_1$, and Combinator produces ${}^{N}C_R$. We combine the two to achieve the desired ${}^{N}P_R$ permutations of branch mapping arrangement.

2.5.1 Complexity Analysis

For compounds A and B producing trees of size n nodes m nodes, respectively, the space complexity for the algorithm is dominated by the four dimensional mapping structure[1]:

$$O(n * n * m * m) = O(n^2 m^2)$$

Time complexity is as follows: The main loop iterates over ${}^{n}C_2 * {}^{m}C_2$ combined pairs, giving us

$$\frac{n!}{2(n-2)!} * \frac{m!}{2(m-2)!} = \frac{n(n-1)}{2} * \frac{m(m-1)}{2}$$

inerations, which reduces to

$$O(n^2 m^2)$$

calls to the main mapping function 2.2. For each such call, there is a constant number of operations and a non-recursive call to BMT(). The worst case BMT execution assumes limits on the number of branches, dictated by maximum valency for a chemical element. Taking the maximum as six, less two atoms that are the mapping pair (see 2.3) leaves four elements available for branching (four subtrees can branch out of each node on both trees).

[1]This is a static structure in our implementation, of dimensions that are maximum expected node counts for chemical compounds: 100^4 32-bit integers

We take the mapping values of all the possible arrangements of the branches in the first tree against those of the second, hence:

$$c +^4 P_4 * leafcomp() = O(c + 4! * leafcomp())$$

operations, where "leafcomp()" is the complexity of mapping two subtrees. It is a function, because although all node pairs in the subtrees have had there map[] values calculated, the best map[] value is yet to be found. The reader will recall from the detailed description in Section 2.3 that at this point we perform another set of permutations, this time of all stem-leaf pairs from one subtree against those from the other. And for a single branch, the stems are fixed (a single node), so assuming n and m leaves respectively, finding the maximum map[] value requires

$$O(leafcomp()) = O(n * m)$$

operations. *This is of course a gross overestimate.* We've assumed n and m leaves in the subtrees of compounds that contain a *total* of n and m nodes respectively. In reality the number of leaf nodes in a single subtree will always be far smaller, and variable.

Still, combining the above results gives the following final scenario:

$$O(n^2m^2 * nm) = O(n^3m^3)$$

or simply $O(n^6)$ for compounds of comparable length. This is not exponential, and thus is kept feasible for SMILES compounds below 500 nodes on modern architectures. Our runtime analysis shows actual running time varies from a few seconds (small compounds less than 50 nodes) up to half an hour on an Intel Xeon processor.

On a distributed cluster of 15 processors, 1600 GSS comparisons are made in less than 24 hours, with no compound exceeding 150 nodes in size. 1600/(24*60) gives a rough estimate of 1.111 (i.e., 1) comparison per minute.

Chapter 3

Experimentation

This chapter will lay out the results of applying the methods discussed so far to the Human Metabolome Database [9], as preliminary testing for the capabilities of this algorithm. The first section below serves as an introduction to the structual similarity mappings when done just once, i.e., applied to a singlepair of compounds to give a result. The following section on prediction will then demonstrate the actual activity estimation process using our methods.

The Human Metabolome data [9] makes for a very interesting test set. Metabolomes are simply chemicals (most likely organic) that are present in the human body as part of the human metabolism, which happens to encompass a wide variety of compounds. The database is comprised of 'cards', one for each compound, that list the various properties under the sections into which they are divided. The properties listed are very extensive, themselves being drawn from other freely available NCBI databases like PubChem. The extensive range of available information per compound is what makes this data interesting in our case. As the reader will see in latter sections, we will attempt to identify some rather obscure things like biological function from compounds' structure, as well as the more typical pure compound thermodynamic property estimations like melting points, boiling points and so on.

There are a few things concerning the data that are worthy of note:

- The database is only semi-standardized in structure, and much of the data cannot be trusted entirely to automated processes.

- There is a very high precentage of missing and inaccurate data among some properties, particularly experimental values for numerical ones. For example, solubility values at room temperature are very scarce.

- The data is fairly outdated, which is to be expected from freely available, highly specialized information. Most of the experiments conducted were done before the 90's, and there is a range of discrepancy between results coming from different labs. E.g., there are compounds that have been entered twice, with very different melting point values (experimental).

Some of these issues cannot be helped, but we manage to avoid problems with accuracy by taking the experimental values for numeric attributes from the EPA's software database, which is able to handle batch inputs (lists of SMILES strings) and produce verified attribute values from the various sources involved with the EPA. There are instances of SMILES strings that have several records in the EPA database (under different names, for the same SMILES input), but these rarely occur, and the most likely reference is the one chosen.

We attempt to handle the other problems by taking a stratified sample of compounds that have an important chemical property value present in all of their records, as a starting point. We chose the normal melting point (measured in degrees Celsius) as the said property, since it is widely used in performance/testing of QSAR methods, and its data was available in a good percentage of the total records. This brings the dataset to about 815 compounds, all guaranteed to have at least an experimental melting point entry.

Also, since the compilation and verification of the data set is a project in itself, the metabocard format has been kept and parsed as is, rather than transforming it into something more suitable for statistical analysis. The said transformation would involve a huge amount of manual work to ensure the data integrity for all property records, as opposed to the relatively small amount of work needed to verify only the properties we are concered with in this round of testing.

3.1 Sample Compound Mappings

3.1.1 Similar Compounds

We will include here a few samples from our result tables that show what 'similar' compounds look like according to the algorithm. This is important to note, because the structural mapping aims to assign the highest score possible, and normalizes that score over the score achieved by the largest compound against itself. It does not penalize for all the extra branches and subtrees that do not match well in a particular situation, there is no deduction from the total score in any situation, so it is very possible that if the compounds are not significantly different in size, and they match very well to a certain degree, there may be entire subtrees in them that are very different from each other. We consider here the compound *Thromboxane* from Table A.5, and a subset of it's best matches that score above 97%. See the table in the Appendix for a full listing.

The first compound is Thromboxane itself. It is of modest size, typical of the compounds encountered in this data set. Since it is composed only of carbon, oxygen and hydrogen, with only single bonds throughout, we can expect the compounds similar to it to be along the same lines. Yet we note that even with this modest size, the algorithm can get away

SMILES	GSS Mapping Score
CCCCCCCCC1OCCCC1CCCCCCC	1.000000
CCCCCCCCCCCCCCCC(O)C(O)C(N)CO	0.989796
CCCCCC(O)C=CC1C2CC(OO2)C1CC=CCCCC(O)=O	0.979798
CCCCCCC=CCCCCCCCC(O)=O	0.976431
CCCCCCCCCCCCCCC=CC(O)C(N)COP(O)(O)=O	0.973333
CCCCCCCCC(C)CCCCCCCCC(O)=O	0.973064

Table 3.1: Thromboxane, CCCCCCCCC1OCCCC1CCCCCCC

with skipping one or two double bonds or different elements along the main compound chain.

That is, the score can still be a high 99% structural similarity with compounds this size (less than 100 atoms in total), *even if there are bonding differences and small branches that don't exist in one of the two compounds being compared.* An example of a 'small branch' is the oxygen in the most similar compound. This has hydrogen attached to it, and although there are three mismatches branches of the sort in total, the second compound is large enough that the comparison is barely affected by these with the respect to everything else that matches.

This is an issue, as we will see in latter sections, because the structures provided by the small oxygen and nitrogen 'branches' have the potential to dramatically affect various properties. Moroever, our generic methodology does not allow for any weighting based on the actual element names, but rather weighs things in relation to structure alone (so we assign greater weight to matched bonding types and chiralities, for e.g.,). This may not correspond to the natural mode of things in biochemistry, where it may be almost certain that the presence of a certain group of elements corresponds to certain properties.

We will examine these effects statistically in Section 3.2.

3.1.2 Dissimilar Compounds

Compounds that are largely dissimilar (i.e., that have very few structural similarities in proportion to their size) are easier to deal with. They pose less of a problem, because for their properties to match, the said properties must be dependent *only* on a small subset of the compound structure (a few local groups for instance). It is unlikely that the list of compounds that match a given molecule A with GSS values below 0.3, are all going to have, on average, more similar properties to compound A than the list that starts with a value of 0.95 and up.

We will not dwell on this very long, but we consider here an example of 'globally dissimilar' compounds to demonstrate that the algorithm does indeed return a low value of similarity for such instances.

Mapping [C=C(C=O)=C] and [N] we obtain a mapping result of 0.214285. So despite the hydrogen matches on the NH_3 molecule and the small size of both compounds, the score in such situations is still sufficiently low to indicate that they are non-similar.

Similarly, mapping [C=C(C=O)=C=C=C] to [C=O] produces a score of 0.3 (30% match).

3.2 Prediction Using Global Similarity

As mentioned above, we will use a stratified random sample of roughly 800 molecules as data for this portion of testing. They are stratified only in the sense that they all have normal melting point records. Of the 800-molecule batch, we picked 55 compounds (randomly) as a test set. Each of the 55 is then mapped against the larger set of 800 and the results of the mappings are sorted, in k-nearest-neighbors fashion. How to actually use the most-similar

cluster is a complicated question that we attempt to address below in a simple manner, but we do maintain a list of the 20 closest compounds in global structure to each of the 55.

3.2.1 Numerical Activity Estimation

Below is a listing of the statistics on the performance of the algorithm in helping predict a numerical pure compound property (normal melting point). The values were estimated by taking the mean of the top K matching compounds' property values. Since the results varied depending on the choice of K, several of these experiments were conducted. Typically, machine learning literature advises K to be of values like 1, 3, 5 and 7, where the odd values are useful because they are involved in classification through a voting scheme. Given the nature of our estimation method however, and the fact that data is missing from may of the chosen top matches, the following was done to calculate the estimate:

We take as many as ten (i.e., $K = 10$) values if all ten have matching values above 95%. As a next-best option, we take the mean of the top five if they have a score of 0.90 or above. Else, we chose the top three most similar compounds. Any missing values are 'skipped', meaning we proceed to the next ranked compound and use its value instead. We employ this method of calculating the results as long as the compounds used have a similarity value above 75% (that is, they can be said to be roughly similar).

The estimation results for our test set follows in Table 3.2.

The mean error rate here was 24%, with a standard deviation of 19%.

Estimated Value	Exp. Result	SMILES
180.250000	189.50	OCC(O)C(O)C(O)C(O)CO
239.950000	212.50	N(C)CC(O)=O
205.375000	289.00	OC(=O)C1=NC2=C(O)C=CC=C2C(O)=C1
173.40000	199.00	CC(=O)NC(CCC(O)=O)C(O)=O
197.387500	254.00	OCC1OC(OC2C(CO)OC(O)C(O)C2O)C(O)C(O)C1O
223.937500	334.00	OC(=O)C1=CC(=O)NC(=O)N1
-60.062500	-99.00	CCCC=O
107.050000	211.00	CNCC(O)C1=CC=C(O)C(O)=C1
144.383750	152.00	OC(=O)CC1=CC=C(O)C=C1
220.300000	255.00	CN(CC(O)=O)C(N)=N
136.587500	84.50	NC(=O)C=C
209.362500	235.50	NC(CCCNC(N)=O)C(O)=O
231.625000	228.50	OC(=O)C1=CC=CN=C1C(O)=O
153.687500	158.50	OC1COC(O)C(O)C1O
219.258750	204.00	OC(=O)C1=CC=CC(O)=C1O
201.537500	274.00	CC(O)C(N)C(O)=O
157.062500	159.00	CC1=NC=C(CO)C(CO)=C1O
160.375000	133.00	OC(=O)C=CC1=CC=CC=C1
142.012500	103.50	CNC(=O)C1=CN=CC=C1
131.925000	170.00	CC(=O)NC1=CC=C(O)C=C1
191.312500	198.50	NC1=CC=C(C=C1)C(=O)NCC(O)=O
154.300000	143.50	COC1=CC(CC(O)=O)=CC=C1O
107.287500	68.80	CCCCCCCCCCCCCCCCCCC(O)=O
118.312500	153.00	COC1=CC2=C(C=C1)C=C(C=C2)C(C)C(O)=O
75.187500	44.00	CNCC(O)C1=CC=CC=C1
-53.675000	-92.00	C=O
140.362500	225.50	OC1C(O)C(O)C(O)C(O)C1O

Table 3.2: Estimation Performance using Mean of Clustered Compounds

3.2.2 Nominal Value Estimation

Nominal values are estimated in the following manner for this testing regimen: only the top five compounds matched are considered. From these compounds, any nominal value that occurs more than once is added to a set of possible attributes that we consider to belong to the unknown compound. The more pervasive the attribute/property in the chosen set of compounds, the higher its probability. This can be represeted on a scale of one to five in this instance.

Five was chosen out of convenience; the scientist may opt to use ten or more compounds to derive an attribute set for the unknown one. However, this carries several problems, because as the number of attributes grows, the values may contradict one another (in which case the clash would have to be somehow resolved). Also, since we encounter anomalies as seen above, things are not as simple as favoring the higher ranked compounds and their attributes. This is because the highest ranked one, for e.g., may be an outlier. We are therefore faced by many of the problems in numerical analysis, although in a more managable form, because of the fact that we are dealing with elements of a set, rather than components of an equation.

One possible way to resolve conflicts in the inferred list of attributes is to begin with higher ranked compounds, and add nominal properties until their values clash with previously added (higher ranked) ones. This is not a process that can be automated very easily. Our analysis here will therefore be less statistical, as we consider in detail some examples of how to manually come up with properties in an unknown compound, using the GSS ranking as an aid, rather than a complete tool on its own.

Our first example is the first compound in the test set, 3-hydroxycapric acid. See Table A.1 for the full list of the most similar compounds. Taking the top three similar compounds we examine them one by one[1]:

- Suberylglycine, with a 99% matching score, is an acyl glycine, present in bodily fluids like Urine, and is it is a dicarboxylic acid. According to the description, the measurement of such compounds can be used to identify problems with mitochondrial fatty acid beta-oxidization in the body, because they are produced in excess when certain metabolic processes malfunction.

- Hexanoylglycine; score 95.5%; also an acyl glycine very similar in properties to the previus item in the list. Hexanoylglycine is a fatty acid metabolite also important in the detection of problems with beta-oxidizaiton. It is also present in Urine.

- N2-Succinyl-L-orinthine; 98%. Not much information available here on actual biological function, apart from the fact that it is a mitochondrial substrate. It is expectedly classified with similar taxonomies as the other compounds considered, so perhaps further information can shed light on similar properties as those above. However, there are not many properties or functionalities detailed here to add to our set.

Given the above, let us now look at the actual compound we are trying to assign properties to. Amazingly enough, the compound naturally occurs in human blood and is found in urine. Furthermore, the compound is involved in the beta-oxidization of fatty acids as an intermediate. Finally, and this is perhaps of most interest to us in this example: 3-hydroxycapric acid is clinically known to accumulate in the plasma of patients due to *the very same metabolic disorders as the most similar compounds examined above*. The particular deficiency which the compound can be used to diagnose, causes several illnesses

[1]The detailed descriptions are not available in the tables

in juveniles, and the most similar compounds are involved in similar metabolic processes, hence they share this very complex nominal property with our 'unknown' compound.

Our next example is a very different metabolome, **1-Methylinosine**. The most similar compounds are listed as follows:

- **8-Hydroxyguanocine**; 90% match. The compound, called 8-OHG for short, is supposed to mark oxidative damage to RNA, DNA, lipid membranes and nucleic acid, and is associated with aging and the related breakdown of biological functionality that causes several diseases like Parkinson's. 8-OHG is a mamalian metabolite and a nucleoside.

- **3-Methyluridine**; 88% match. This is also a nucleoside, but is of the methylated variety. It is modified by enzymes that render it useless for reconstruction in RNA, and is therefore normally excreted in urine. Methylated nucleotides are investigated heavily in the accompanying literature to the database, in studies invloving cancers (e.g., acute leukemia). They are of great interest to scientists studying these diseases because of how they are altered, and what they may indicate.

- **Ethenodeoxyadenosine**; 87.9% match. This is an etheno modified nucleoside (taxonomy subclass deoxy nucleosides). Yet another metabolome linked to cancer, albeit in an uncertain way, as it is produced due to both carcinogens and other chemicals in the body. Concentrations tend to be higher when known cancer risk factors are raised.

Looking at these compounds, one can tell almost right away that the one we are testing is not going to be good news. It has a high probability of being a nucleoside, methylated, and involved in damage/abnormalities in DNA or RNA. It can be expected in the blood, in the cerebral cortex, and in urine. These predictions turn out to be mostly correct. The

compound is indeed part of similar studies on methylated nucleosides in cancer patients, and is found in urine and the bloodstream (serum).

In other words, a manual observation of the most structurally similar compounds to the one we need, can lead to the accurate prediction of its location, it's specific taxonomy details, it's biofunctions/involved metabolic processes, and even the possible clinical/medical properties of the compound. Here we begin to see the power of the discriminative classification process - we did not attempt to build a model for each of these classes of activity, but we are instead able to postulate these properties by depending on the global structural similarity of the compounds alone.

The above were only a few examples out of the 800 compound data set. A comprehensive treatment of all 55 tests will be outside of the spatial limitations of this section, but our investigation shows similar performance in most of the 55 compounds used as a test set. The freely available data should also encourage the reader to conduct their own investigation onto the validity of these claims, as the matching scores/most similar lists for the 55 compounds have been already calculated and included in the Appendix.

3.3 Comparison to a Naive Mapping Method

We have developed a method that is far simpler, in an attempt to discover any advantages (or disadvantages) in estimation obtained by the more elaborate/exhaustive structural mapping.

This naive method relies only on the similarity of element occurance. It gives a mapping value by counting atoms of each element name (i.e, atomic element groups), and subtracting differences in these from the total count (sum) of both compounds. At most, there will be no difference between the element counts for any element - the compounds are the same - and the score will be the untouched sum. On the other extreme, no elements in the either

compound exist in the other, and the score is reduced to zero. This gives a very vague (but valid) measure of how "similar" the two compounds being compared are, in terms of their comprising elements.

Using the naive mapping on the same group of test compounds, we obtain barely passable results on the melting point attribute: 0.597 mean error rate, and a standard sample deviation of 0.7 (of the error rate). The 60% error translates to an estimate of 160 oC from an experimental value of 100 oC, for example.

3.4 Comparison to the EPA Estimates

The Environmental Protection Agency has produced a free software suite (confusingly called EPI) that hosts many individual programs for property estimation. This is the same suite that provides the database of experimental data we are using for the numerical estimation. One of the modules in the suite is able to estimate pure compound properties like the normal melting point, using a weighted mean of the standard Joback method and a few others. Running our test set as input, we obtain the following results.

The EPI software is limited to numerical property estimation (most notably, attributes like biodegradability, half-life, and similar things of interest to the agency), so we cannot really compare its ability to predict nominal attributes to ours. This is almost a moot point; QSAR methodology does not normally allow for the estimation or 'suggestion' of non-numerical attributes of chemicals, owing to the fact that rigid formulas are the output of most QSARs. They are designed for numerical estimation, not descriptive activities of a metabolic and clinical nature.

The EPI performance on the melting point property is close to ours (see table above), with a mean sample error rate of 30.2% and a standard deviation of 0.24 (in those error rates). There is plenty of room for statistical fluctuation, so our method (24% mean error

Estimated Value	Exp. Result	SMILES
138.97	189.50	OCC(O)C(O)C(O)C(O)CO
232.05	212.50	N(C)CC(O)=O
172.87	289.00	OC(=O)C1=NC2=C(O)C=CC=C2C(O)=C1
172.65	199.00	CC(=O)NC(CCC(O)=O)C(O)=O
255.42	254.00	OCC1OC(OC2C(CO)OC(O)C(O)C2O)C(O)C(O)C1O
188.60	334.00	OC(=O)C1=CC(=O)NC(=O)N1
-80.23	-99.00	CCCC=O
123.68	211.00	CNCC(O)C1=CC=C(O)C(O)=C1
102.93	152.00	OC(=O)CC1=CC=C(O)C=C1
92.83	255.00	CN(CC(O)=O)C(N)=N
21.69	84.50	NC(=O)C=C
292.35	235.50	NC(CCCNC(N)=O)C(O)=O
141.18	228.50	OC(=O)C1=CC=CN=C1C(O)=O
97.75	158.50	OC1COC(O)C(O)C1O
128.34	204.00	OC(=O)C1=CC=CC(O)=C1O
215.34	274.00	CC(O)C(N)C(O)=O
124.48	159.00	CC1=NC=C(CO)C(CO)=C1O
69.48	133.00	OC(=O)C=CC1=CC=CC=C1
101.23	103.50	CNC(=O)C1=CN=CC=C1
119.92	170.00	CC(=O)NC1=CC=C(O)C=C1
180.34	198.50	NC1=CC=C(C=C1)C(=O)NCC(O)=O
117.66	143.50	COC1=CC(CC(O)=O)=CC=C1O
132.96	68.80	CCCCCCCCCCCCCCCCCCC(O)=O
137.63	153.00	COC1=CC2=C(C=C1)C=C(C=C2)C(C)C(O)=O
40.42	44.00	CNCC(O)C1=CC=CC=C1
-110.94	-92.00	C=O
148.11	249.50	OC1C(O)C(O)C(O)C(O)C1O
36.45	52.50	N1C=CC2=C1C=CC=C2

Table 3.3: Estimation Performance using Weighted Mean of EPA Methods

rate) does not necessarily maintain an upper hand here. This comparison is merely proof that our simple method yields comparable results.

3.5 Case Analysis

At this point we attempt to provide some insight into the mapping process and its efficiency in making predictions, by investigating specific cases of our simple estimation procedure. This should also provide some illustrated examples on how the similarity algorithm actually works – how matching features are mapped, ignoring mismatching ones, and so on. The estimation statistics on the normal melting point are (in terms of average error) quite encouraging as we have seen, but why is the percentage error high for some compounds? How does the notion of global structure similarity translate to these cases?

We are not, in this literature, going to consider any detailed chemical interpretation of these cases, as that will involve making postulates outside our computational scope here. We do however, want to witness, in simple structural terms, the algorithm being *insufficient* in drawing estimates in some cases. Where does global similarity on its own become 'naive', so to speak?

Following will be a few examples of compounds where the estimate deviates heavily from the experimental value. The melting point temperatures below are all in degrees celsuis.

3.5.1 Xanthurenic Acid

We begin with Xanthurenic acid, shown in Fig3.1. This has experimental and estimated values of 289 and 205.37, respectively. The most similar compound to Xanthurenic acid is shown in Fig3.2. We'll call this 'match-1', and subsequent compounds (the second most similar, third most similar..etc.,) 'match-2', 'match-3' and so on.

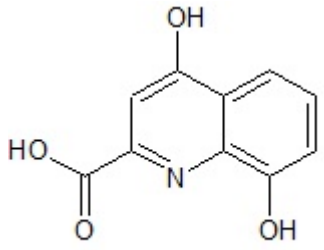

Figure 3.1: Xanthurenic acid

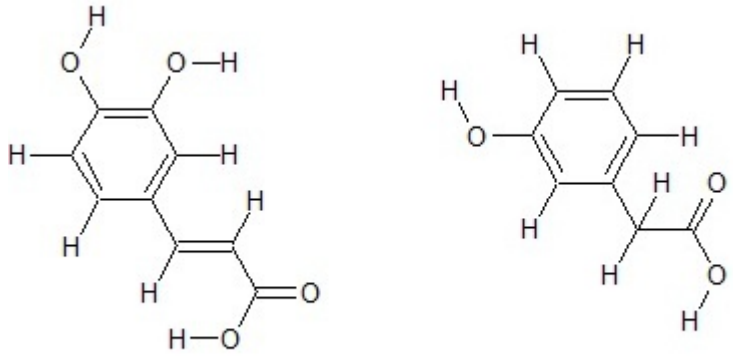

Figure 3.2: Closest matches to Xanthurenic acid.

First off we note that this is the only compound with a matching score above 0.9. There are not many comounds very close in structure to Xanthurenic acid, in other words. The experimental value for match-1 is 225 degrees, and is the highest in all the similar compounds that will follow. This means that we cannot hope to come closer to the correct value via higher estimates averaging out with the low ones. For this particular mapping, we note the compounds are basically different in the following aspects: the single nitrogen atom in the original is missing; the double aromatic ring structure is broken, and despite most of the small groups still existing in the same locations on both compounds, one [OH] group that was located on the second ring in Xanthurenic acid is here attached to the first (and only) ring.

The next similar compound (also shown in Fig 3.2) has an even lower melting point: 132 degrees. This is almost identical to the previous match, but now one of the [OH] groups on the ring is missing, and there is a difference in bond types on the lower part of the compound. This is a common scheme so far for the basic structure we are dealing with in the original compound – the substitution of double bonds with single ones, subsequent addition of Hydrogen, and the presence of [H] in the place of [OH] pushes the value of the melting point down. These are just casual observations of course, we do not know from this perspective what these differences amount to physio-chemically, or in relation to the melting point activity.

3.5.2 Epinephrine

We note similar effects in the following examples. Epinephrine, shown in fig. 3.3, is no different from its most similar match (call this match-1, as before) except for the [OH] groups missing on the ring in that compound, shown in fig. 3.4.

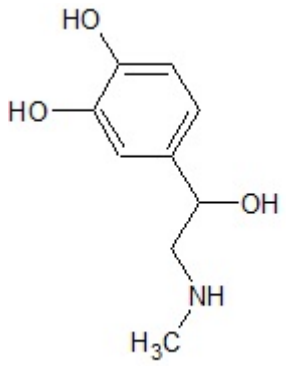

Figure 3.3: Epinephrine: CNCC(O)C1=CC=C(O)C(O)=C1

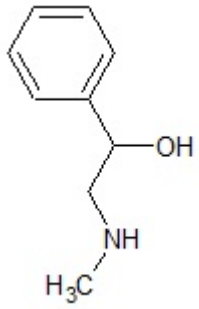

Figure 3.4: First match to Epinephrine: CNCC(O)C1=CC=CC=C1

Yet the difference in the melting point is (relatively, i.e., as a percentage error) very large: 211oC and 44oC, respectively. The other similar compounds (match-2 and match-3 shown in fig. 3.5) help close the gap a little in the estimation, but still we are faced with a situation where the closest compounds are generally far lower in m.p despite having the same basic structure (the single ring, carbon chain and nitrogen group).

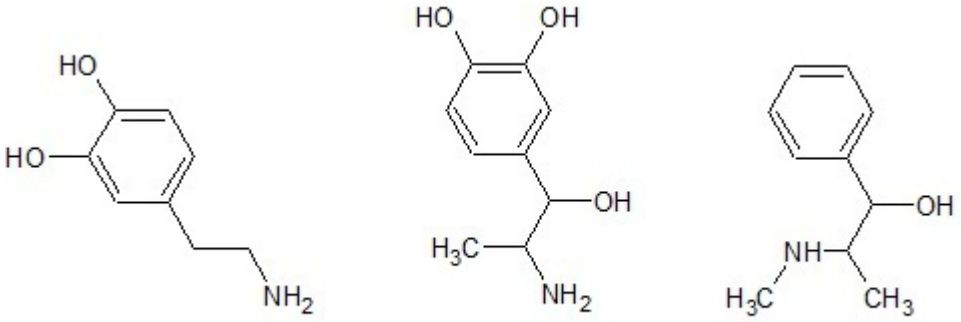

Figure 3.5: Epinephrine match-2 (CNCCC1=CC=C(O)C=C1), match-3 (NCCC1=CC=C(O)C(O)=C1) and match-4 (CC(N)C(O)C1=CC(O)=C(O)C=C1), from left to right.

Notably, the closest neighbor to the Epinephrine in melting point (at 218oC) is the fourth most-similar compound (match-4, Fig. 3.5). Here there is an added [CH3] group near the bottom of the figure, with the [OH] groups on the ring missing. This has put the compound further away in terms of global similarity than others, but the additions to the structure (which lessen the branch mapping score at that point in the structure comparison) have somehow contributed to raising the melting point to a closer value.

3.5.3 Quinolinic acid

We now consider an example with comparatively good melting point estimation (231.62 experimental, 228.50 estimated). At this point we will stop making casual comments on the

local structural differences, as it is clear that more involved domain knowledge of chemical properties is required to make authoritative statements on the observed differences and their effects.

The compound we are estimating is Quinolinic acid, shown in Fig. 3.6. The three closest matches are shown in Fig. 3.7 and the fourth (which starts to look distinctively different from the others due to the severed aromatic ring) in . 3.8.

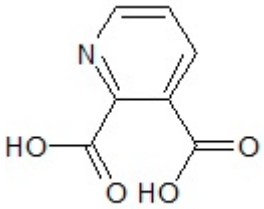

Figure 3.6: Quinolinic acid (OC(=O)C1=CC=CN=C1C(O)=O)

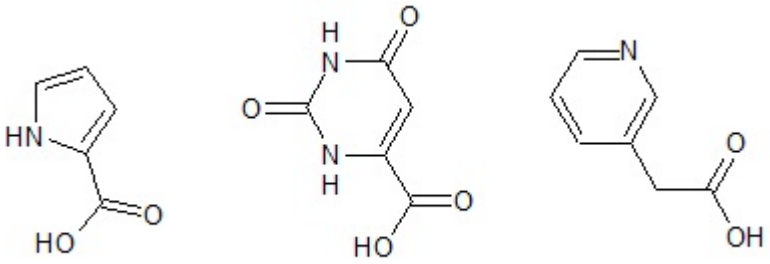

Figure 3.7: Quinolinic acid match-1 (OC(=O)C1=CC=CN1), match-2 (OC(=O)C1=CC(=O)NC(=O)N1) and match-3 (OC(=O)CC1=CN=CC=C1)

What can be said about this is that some further information about structures local to the cluster of similar compounds is needed to arrive at a more reasonable estimate. The presence or absence of some small subsets of the overall structure can have a dramatic relationship with thermodynamic properties (which typically affect many others[4]). How to

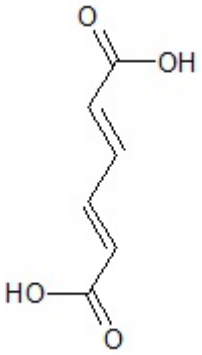

Figure 3.8: Quinolinic acid match-4 (OC(=O)C=CC=CC(O)=O)

account for these possibilities – how to formulate local model parameters from the globally similar compounds – must be the subject of further research.

3.5.4 1-Methylinosine

Finally, we turn to an example where we have been able to postulate a set of correct (i.e., present) non-numerical properties. We looked at 1-Methylinosine back in section 3.2.2, where we happily reported on the fact that it shared some very complex properties with the cluster of globally most-similar compounds. It is shown in Fig. 3.9, and it's most similar compounds in Fig. 3.10 and 3.11.

Needless to say, as in the case of thermodynamic properties being sensitive to small structural changes, the prediction of more intricate biological activities such as metabolic roles will also be subject to error, whatever scheme we use to draw the 'postulated' set of properties. Different compounds are ultimately responsible for different roles in biology, though they may share some (or many) effects. Recall that we manually examined the

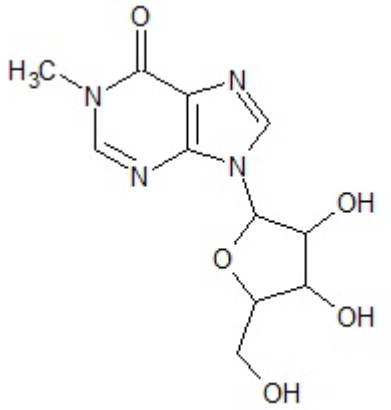

Figure 3.9: 1-Methylinosine

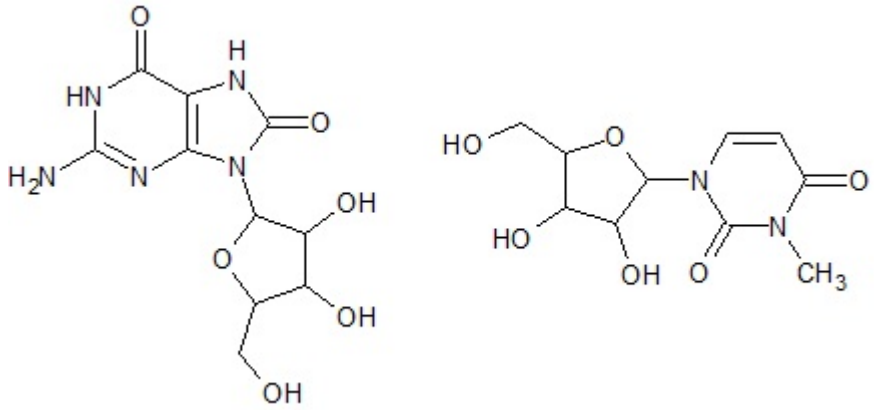

Figure 3.10: Matches 1 and 2 for 1-Methylinosine

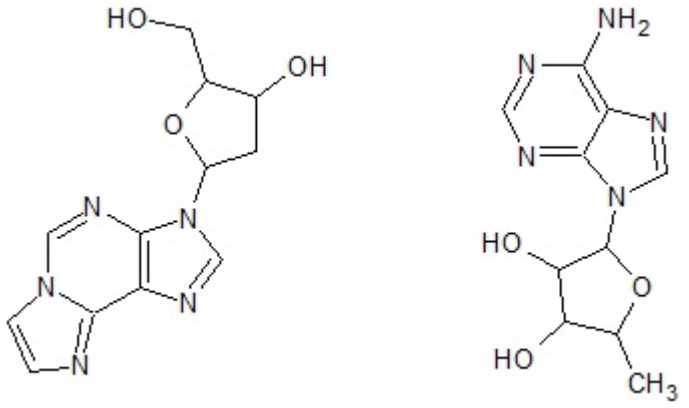

Figure 3.11: Matches 3 and 4 for 1-Methylinosine

neighboring compounds, drawing clinical and other properties from the compound descriptions, then assigned these to the unknown compound based on how prevalent each property was in the closest compounds (for some chosen threshold of prevalent). In short, we built a set from the sets of unstructured data, using a simple counting scheme.

The beauty of this was that we had no notion of what local features were responsible for the biomedical properties that resulted; we bypassed the issue by assuming the encapsulation of these features in the global structure, and this has shown to be promising. We may want to further 'automate' the science by taking the process into the inferential/classification domain and applying machine learning techniques to infer certain fixed *categories* of properties. This again becomes the subject of further research. It will require strictly defining the properties being sought (for e.g., location of activity in the human body), which in turn requires structured data. These properties will then be assumed to follow some underlying probability distributions conditional on some structural parameters

that have to be chosen. The choosing of parameters is the problem of how to 'phrase' the input to the inferential machine. There are very many (possibly infinitely many) ways to do it. It becomes the central problem that we have so far managed to sidestep in this work.

Chapter 4

Conclusions and Further Work

Our method can be seen a a form of clustering that reduces the sample set from which we infer unkown values and attributes. It may be beneficial to consider a secondary clustering, this time based on the values of the 'globally similar' compound properties. What this means is that in order to avoid one or two values disrupting the estimation process, we can cluster the values together as a form of outlier detection, then use automated or manual means of choosing which cluster of similar compounds would result in a more likely estimate.

This would avoid skew in the results, with very little computational cost. Of course, the suitability of such methods of 'filtering the results' would vary with the data in each case. So if we have results in which the culprits in the data, so to speak, are obvious, the results may be more accurate than when the skew is not easy to analyse. However, the outlier detection scheme may isolate the 'correct answer', which may easily be the minority in the top 20 compounds or so. This is where the local group contribution methods and the manual inspection of the similar-compounds list come into play. With complicated data like this, clustering becomes very difficult for an algorithm to achieve on its own, even though the distance metric is very valid. Human input in evaluating the actual result can be very useful.

In conclusion, we have found that Global Structural Similarity is indeed of utility in the estimation of biochemical activity. Results from numerical estimation show a clear

encapsulation of parametric QSAR models within the global structure, so that relying on a simple scheme entirely dependent on most similar compounds actually yields usable results. It is a naive estimation method, but the apparent simplicity is deceiving. We have circumvented the development of very complicated models via the usage of structural mapping to narrow the space from which we make estimates. Our surpassing the standard Joback method accuracy on this data set is probably not evidence of superiority, but rather of relevance. We can confidently assume that extrapolar models that use our process as a starting point can achieve accurate results.

We also find that the non-numerical activity estimation, or postulation, rather, to be a very interesting and fruitful part of these efforts. Some very complex behaviors (relative to thermodynamic activities for instance) can be 'suggested' based on similarity to other molecules alone. Of course, the extrapolar nature of such postulations relies entirely on whether similar molecules exist, and it is unlikely that such ad hoc approaches provide a substitute for clinical study on important biological/biomedical effects, but again, we seek to provide utility with these results, not definitive models.

The subject matter expert may benefit from the ranking. It can cut down on the time taken to research similar compounds by immediately providing the set of similar compounds, and from them, a set of suggested features. Where similar compounds seem to agree on a property (e.g., typical locations), the suggestions may be true. As with quantitative activity, we have avoided some very difficult modeling questions with our discriminative process. Why do certain compounds bind to the surface of certain cells? Or why does the human body produce a particular class of compounds in a certain location? The difficult questions are bypassed because again, the required models are encapsulated by the global structure of the compounds.

One should also not limit oneself to biochemistry as the domain of application for this research. We have started out looking for ways to estimate activities in chemical compounds, but the algorithm presented can easily be modified to address any kind of graph-able structure. Literally anything that resembles a network, with a structure that affects (or even better, dictates) the properties of those networked components as a whole, can be treated in a similar manner as the molecular compounds in this literature. We have restricted testing to the domain of quantitative structure-activity relationships in human metabolomes, but computational QSAR approaches like ours do not have to be exclusive to such domains.

Moving onwards, it is easy to foresee the combination of this work (utilizing global structure to group similar graphs) with generic algorithms that use local groups. As of this writing, methods that do this have been proposed by Dr. Bala Kalyanusundaram of Georgetown University. The idea here is to enumerate all the possible local groups (translates to subtrees) from a set of compounds, and then use some type of modeling to infer relationships. One suggested method is the commonly used linear regression technique. However, this only works assuming simple additive relationships between groups and properties. It is possible to achieve more detailed descriptions of relationships using multi-layered neural networks. This would liberate the investigator from assumptions of independence between local groups. The ability to learn logical rules can provide logical explanations for the situations where numerical attribute values cannot simply be fitted to mathematical models.

Bibliography

[1] U.S Environmental Protection Agency. Epi suite v 4.0. http://www.epa.gov, 2009.

[2] Tatsuya Akutsu and Magnús M. Halldórsson. On the approximation of largest common subtrees and largest common point sets. In *ISAAC '94: Proceedings of the 5th International Symposium on Algorithms and Computation*, pages 405–413, London, UK, 1994. Springer-Verlag.

[3] Philip Bille. A survey on tree edit distance and related problems. *Theor. Comput. Sci.*, 337(1-3):217–239, 2005.

[4] Allen B.Richon and Stanley Young. An introduction to smiles methodology. `http://www.netsci.org`.

[5] Bjorn Bringmann and Andreas Karwath. Frequent smiles, 2004.

[6] Weimin Chen. New algorithm for ordered tree-to-tree correction problem. *J. Algorithms*, 40(2):135–158, 2001.

[7] R Ghani et al. A group contribution approach to computer aided molecular design. *J. AlChE*, 37, 1991.

[8] R.D. King et al. Structure-activity relationships derived by machine learning. Proceedings of the National Academy of Sciences, 1996.

[9] Wishart DS et al. Hmdb: the human metabolome database. *Nucleic Acids Res*, 35, Jan 2007.

[10] T. Fujita and T. Ban. *J. Med. Chem*, 14, 1971.

[11] Molecular Networks GmbH. Corina 3d online. `http://www.molecular-networks.com/online_demos/corina_demo`.

[12] C. Hansch. *Acc. Med. Chem*, 2, 1971.

[13] C. Hansch and A. Leo. *Substitution Constants for Correlation Analysis in Chemistry and Biology*. Wiley-Interscience, New York, 1979.

[14] K.G. Joback and R.C. Reid. Estimation of pure compound properties from group contributions. *Chemical Engineering Comm*, 1987.

[15] Y. C. Martin. A practitioner's perspective of the role of qsar in medicinal chemistry. *J. Med. Chem*, 1981.

[16] University of Mass. at Amherst. Rasmol molecular visualization suite. http://www.openrasmol.org.

[17] Jr S.M Free and J.W. Wilson. A mathematical contribution to structure-activity studies. *J.Med.Chem*, 7, 1964.

[18] Daylight Chemical Information Systems. An introduction to smiles notation. http://www.daylight.com, 2008.

[19] U.S. National Library of Medicine The National Center for Biotechnology Information. Pubchem database. `http://pubchem.ncbi.nlm.nih.gov/search/help_search.html`.

[20] Tanimoto TT. Internal report. Technical report, IBM, November 1957.

[21] J. D. Ullman, A. V. Aho, and D. S. Hirschberg. Bounds on the complexity of the longest common subsequence problem. *J. ACM*, 23(1):1–12, 1976.

[22] D. Weininger. Smiles: A chemical language and information system. *J. Chemical Inf. Comp. Sci*, 1988.

Appendix

Summary of Results

The following are the result tables for the test set of compounds. Each table contains a sorted list of the top twenty most similar compounds to the one being tested, the respective GSS value, and the property value for each (melting point in this case.)

SMILES	GSS Mapping Score
OC(=O)CCCCCCC(=O)NCC(O)=O	0.993464
NCCCC(NC(=O)CCC(O)=O)C(O)=O	0.980769
CCCCCCC(O)=O	0.961538
CCCCCC(=O)NCC(O)=O	0.955128
CCCCCCCCO	0.948718
CC(C)CC1=CC=C(C=C1)C(C)C(O)=O	0.946667
OC1=CC=C(CCC(=O)C2=C(O)C=C(O)C=C2O)C=C1	0.943262
CNCCCCC(N)C(O)=O	0.942308
NC(CCS(=O)(=O)CCC(N)C(O)=O)C(O)=O	0.941176
OCC(O)C(O)C(O)C(O)C(O)CC(=O)C(O)=O	0.940000
CC(O)(CCO)CC(O)=O	0.923077
CC(O)C(O)C1CNC2=NC(N)=NC(=O)C2(O)N1	0.921986
CCC(O)CC(O)=O	0.916667
OCC1OC(CC1O)N1C=NC2=C1N=CN1C=CN=C21	0.911111
CC(=O)NC(C=O)C(O)C(O)C(O)COS(O)(=O)=O	0.908497
NC1=NC(=O)N(C=C1)C1CC(O)C(COP(O)(O)=O)O1	0.904762
CC12CCC(CC1)C(C)(C)O2	0.903846
CC(C1=CC=C(O)C=C1)C(=O)C1=C(O)C=C(O)C=C1	0.898551
CCCCC(=O)C(O)=O	0.897436

Table A.1: Compounds Similar to "3-Hydroxycapric acid",CCCCCCCC(O)CC(O)=O

SMILES	GSS Mapping Score
OC1COC(O)C1O	0.964286
OCC1OC(=O)C(O)C1O	0.952381
OCC1OC(O)CC1O	0.928571
OCC(O)C1OC(O)=C(O)C1=O	0.925926
OC(CCC(O)=O)C(O)=O	0.916667
OC(=O)CCC(=O)CC(O)=O	0.904762
OC1CCNC1C(O)=O	0.892857
OC(=O)CC1=C(O)C=CC(O)=C1	0.888889
OCC(O)C(O)CC(O)=O	0.885057
OCCNCCO	0.880952
OC(=O)CC1=CC=C(O)C(O)=C1	0.876543
OCC1OC(=O)C(O)C(O)C1O	0.870968
CC(O)C(N)C(O)=O	0.869048
OCC1OC(O)C(O)C(=O)C1O	0.860215
OC(=O)CC1=CC=CC=C1O	0.857143
OC(C(CC(O)=O)C(O)=O)C(O)=O	0.855556
OC(CN1C=CN=C1)C(O)=O	0.845238
OC1OC(C(O)C(O)C1O)C(O)=O	0.843750
COC1=CC(C(O)=O)=C(O)C=C1	0.839506

Table A.2: Compounds Similar to D-Xylose,OC1COC(O)C(O)C1O

SMILES	GSS Mapping Score
CCC(O)=O	0.833333
OC(=O)CC(O)=O	0.812500
CC(O)C=O	0.770833
OC(=O)C=CC(O)=O	0.750000
CC(N)C(O)=O	0.736842
CC(=O)C=O	0.729167
CC(=CC(O)=O)C(O)=O	0.714286
OC(=O)C1=CC=CN1	0.705882
OCC=O	0.687500
CNCC(O)=O	0.684211
OC1=CC=C(O)C=C1	0.666667
CC(C)(O)C(O)=O	0.651515
NC(CO)C(O)=O	0.650000
CN(C)C=O	0.648148
OC1=CC=C(Cl)C=C1Cl	0.647059
CNC	0.645833
CC(C)=CC(O)=O	0.636364
OC(CC(O)=O)C(O)=O	0.634921
OC(=O)C1=CC=CC=N1	0.631579

Table A.3: Compounds Similar to Pyruvatoxime,CC(=NO)C(O)=O

SMILES	GSS Mapping Score
CCCCCC(O)C=CC1OC(O)CC(O)C1CC=CCC(O)=O	0.983936
CCCCCC=CCC=CCC=CCCCCC(O)=O	0.956349
CC(C)C1=CC2=C(C(O)=C1O)C1(CCCC(C)(C)C1CC2)C(O)=O	0.948413
CCCCCC=CCC=CCC=CCC=CCCCC(O)=O	0.945098
CC12CCC3C(CCC4=CC(=O)CCC34C)C1CCC2C(=O)CO	0.944444
COC1=C(O)C=CC(CNC(=O)CCCCC=CC(C)C)=C1	0.940476
CCCCCC(O)C=CC1C(O)CC2OC(CC12)=CCCCC(O)=O	0.939394
CCCCCC=CCC=CC=CC(CC=CCCCC(O)=O)OO	0.938697
CCCCCC=CCC=CCCCCCCCC(O)=O	0.936508
CCCCCC=CCC=CCC1OC1CC=CCCCC(O)=O	0.934109
CC12CCC3C(CCC4CC(O)CCC34C)C1CCC2O	0.932540
CC12CCC(=O)C=C1CCC1C2CCC2(C)C1CCC2(O)C(=O)CO	0.929412
CCCCCC(O)C=CC1C2CC(OO2)C1CC=CCCCC(O)=O	0.928839
OC(CCCCCCCCC(O)=O)CC(O)=O	0.928571
CCCCCC(O)C=CC1C=CC(=O)C1CCCCCCC(O)=O	0.928030
CC12CCC3C(CC=C4CC(O)CCC34C)C1CC(O)C2O	0.924603
CC12CCC3C(CCC4=C3C=CC(O)=C4)C1CC(=O)C2=O	0.920635
CC(=O)C1CCC2C3CC=C4CC(O)CCC4(C)C3CCC12C	0.918605
CCCCCC(OO)C=CC1C2CC(OO2)C1CC=CCCCC(O)=O	0.918519

Table A.4: Compounds Similar to "Prostaglandin J2", CCCCCC(O)C=CC1C(CC=CCCCC(O)=O)C=CC1=O

SMILES	GSS Mapping Score
CCCCCCCCC1OCCCC1CCCCCCC	1.000000
CCCCCCCCCCCCCCCC(O)C(O)C(N)CO	0.989796
CCCCCC(O)C=CC1C2CC(OO2)C1CC=CCCCC(O)=O	0.979798
CCCCCCC=CCCCCCCCC(O)=O	0.976431
CCCCCCCCCCCCCC=CC(O)C(N)COP(O)(O)=O	0.973333
CCCCCCCCC(C)CCCCCCCCC(O)=O	0.973064
CCCCCC(O)C=CC1C(O)CC2OC(CC12)=CCCCC(O)=O	0.969697
CC12CCC(=O)C=C1CCC1C2CCC2(C)C1CCC2(O)C(=O)CO	0.966330
CC12CCC3C(CCC4CC(O)CCC34C)C1CCC2O	0.962963
CC(=O)C1(O)CCC2C3CCC4=CC(=O)CCC4(C)C3CCC12C	0.959596
CC12CCC(=O)CC1CCC1C2CCC2(C)C1CCC2=O	0.956229
CC(C)CCCC(C)CCCC(C)CCCC(C)CC(O)=O	0.953795
CC1(O)CCC2C3CCC4CC5=C(CC4(C)C3CCC12C)C=NN5	0.952862
OC(CCCCCCCCC(O)=O)CC(O)=O	0.949495
CC12CCC(=O)CC1CCC1C3CCC(O)(C(=O)CO)C3(C)CC(O)C21	0.946128
CC1CC2C(CCC3(C)C2CCC3(O)C(C)=O)C2(C)CCC(=O)C=C12	0.942761
CC12CCC3C(CC=C4CC(O)CCC34C)C1CC(O)C2O	0.939394
CC12CCC3C(CCC4=CC(=O)CCC34C)C1CCC2OS(O)(=O)=O	0.932660
CCCCCCCCCC(=O)OCC(O)COP(O)(=O)OCC[N+](C)(C)C	0.930159

Table A.5: Compounds Similar to Thromboxane,CCCCCCCCC1OCCCC1CCCCCCC

SMILES	GSS Mapping Score
OCC(O)C1=CC(O)=C(OS(O)(=O)=O)C=C1	0.990741
OC1C(O)C(O)C2OP(O)(=O)OC2C1O	0.981481
NCC1(O)OCC(O)C(O)C1O	0.973684
OC(COP(O)(O)=O)C(O)C=O	0.964912
OCC1OC(=O)C(O)C(O)C1O	0.956140
CCC(O)CC(O)=O	0.947368
NC1=NC(=O)C2=NC(=CN=C2N1)C(O)C(O)CO	0.940171
CC(CCC(O)=O)C(O)=O	0.938596
CNCC(O)C1=CC=C(O)C(O)=C1	0.936937
CC(O)C(O)C1=NC2=C(NC(N)=NC2=O)N=C1	0.929825
CC1OC(OP(O)(O)=O)C(O)C(O)C1O	0.925000
NC(CCCC(O)=O)C(O)=O	0.921053
OCC(O)C(O)C(O)C(O)C(=O)CO	0.918699
OCC1OC(CC1O)N1C=NC2=C1N=CN=C2O	0.916667
CC(O)C(N)C(O)=O	0.912281
OC(CC1=CNC2=C1C=CC=C2)C(O)=O	0.907407
OC(CCC(O)=O)C(O)=O	0.903509
OCC1OC(OP(O)(O)=O)C(O)C(O)C1O	0.902439
OCCCC(O)=O	0.894737

Table A.6: Compounds Similar to Galactitol,OCC(O)C(O)C(O)C(O)CO

SMILES	GSS Mapping Score
NC(=O)NCCC(O)=O	0.948718
NCCCC(O)=O	0.897436
CNC(=O)C1=CN=CC=C1	0.880000
CN1C(=O)NC(=O)C2=C1N=CN2	0.861111
CN(C)CC(O)=O	0.846154
CC(O)C(N)C(O)=O	0.833333
CNCC(O)=O	0.820513
OC(=O)CCC1=CN=CN1	0.813333
CC1=C(N)NC(=O)N=C1	0.807692
CCC(=O)NCC(O)=O	0.802469
OCC1OC(=O)C(O)C1O	0.800000
NCCC=O	0.794872
CNC(C)(C)C(O)=O	0.793103
OCCNCCO	0.790123
OC(=O)CC1=CN=CC=C1	0.782051
CN(C)C(=N)NC(N)=N	0.781609
NC(CS)C(O)=O	0.773810
OC(=O)CC(=CC(O)=O)C(O)=O	0.773333
NCCOP(O)(O)=O	0.769231

Table A.7: Compounds Similar to Creatine,CN(CC(O)=O)C(N)=N

SMILES	GSS Mapping Score
OC(=O)C=CC1=CC(O)=C(O)C=C1	0.904762
OC(C(O)=O)C1=CC=C(O)C(O)=C1	0.892857
OC(=O)CC1=CC=CC(O)=C1	0.880952
OC(=O)C1=CC=CC(O)=C1O	0.857143
OC(=O)CC=CC1=CN=CC=C1	0.845238
CC(=O)OC1=CC=CC=C1C(O)=O	0.833333
OC(=O)CC1=CC=CC=C1O	0.821429
OC(=O)CC1=CN=CC=C1	0.809524
OC(=O)CC1=CC2=C(N1)C=CC=C2	0.800000
OC(=O)CNC(=O)C1=CC=CO1	0.797619
OC(=O)C1=CC(Cl)=CC=C1	0.785714
NC(C(O)=O)C1=CC=C(C=C1)C(O)=O	0.781250
OC1CC(=CC(O)C1O)C(O)=O	0.774194
OC(=O)C1=CC=CN1	0.773810
COC1=CC=C(CC(O)=O)C=C1	0.763441
CCC(O)CC(O)=O	0.761905
COC1=C(O)C=C(C=CC(O)=O)C=C1	0.757576
OC(=O)CNC(=O)C1=CC=CC=C1	0.752688
NC(CCCC=O)C(O)=O	0.750000

Table A.8: Compounds Similar to "Xanthurenic acid",OC(=O)C1=NC2=C(O)C=CC=C2C(O)=C1

SMILES	GSS Mapping Score
OC(=O)CCCCCCC(=O)NCC(O)=O	0.980392
OC(=O)CCCCCCC=CC(O)=O	0.973856
OC1CCC(CC1)CC(O)=O	0.967320
CCCCCCCC(O)CC(O)=O	0.961538
NCCCC(NC(=O)CCC(O)=O)C(O)=O	0.960784
OCC1OC(CC1O)N1C=NC2=C1N=CN1C=CN=C21	0.948148
CC(C)CC1=CC=C(C=C1)C(C)C(O)=O	0.946667
CCCC(CCC)C(O)=O	0.941176
OC1=CC=C(CCC(=O)C2=C(O)C=C(O)C=C2O)C=C1	0.936170
CC(C)C1CCC(C)CC1O	0.934641
CC(O)C(O)C1CNC2=NC(N)=NC(=O)C2(O)N1	0.921986
NC(CCCC(O)=O)C(O)=O	0.921569
OCC(O)C(O)C(O)C(O)C(O)CC(=O)C(O)=O	0.920000
OC1C(O)C(OP(O)(O)=O)C(O)C(O)C1OP(O)(O)=O	0.916667
CC(=O)NC(CCC(O)=O)C(O)=O	0.915033
NC1=NC2=C(NC(=O)N2C2OC(CO)C(O)C2O)C(=O)N1	0.914894
CC(CC(O)=O)CC(O)=O	0.908497
NC(CCCC=O)C(O)=O	0.901961
CC(CCC(O)=O)C(O)=O	0.895425

Table A.9: Compounds Similar to "3-Hydroxysebacic acid",OC(CCCCCCC(O)=O)CC(O)=O

SMILES	GSS Mapping Score
CCC(=O)NCC(O)=O	0.972222
CC(C)C(N)CC(O)=O	0.962963
COC1=C(O)C=CC(=C1)C(O)C(O)=O	0.959596
CCCCC(=O)NCC(O)=O	0.945946
NC(CCC(N)=O)C(O)=O	0.944444
NC(CCCC=O)C(O)=O	0.935185
CC(CC(O)=O)C1=CC=CC=C1	0.933333
OC(CCC(O)=O)C1=CC=CN=C1	0.931373
NCCCC(N)C(O)=O	0.925926
COC1=C(O)C=C(C=CC(O)=O)C=C1	0.919192
CC(CCC(O)=O)C(O)=O	0.916667
OC(C(O)C(O)C(O)=O)C(O)C(O)=O	0.911765
CCC(O)CC(O)=O	0.907407
NC(CC1=CC(O)=C(O)C=C1)C(O)=O	0.904762
NCCCCC(N)C(O)=O	0.900901
NCC(=O)N1CCCC1C(O)=O	0.898148
CC(O)C(N)C(O)=O	0.888889
CN1C(=O)NC(N)=C(NC(C)=O)C1=O	0.885714
NC(=N)NCCCC(O)C(O)=O	0.882883

Table A.10: Compounds Similar to "N-Acetylglutamic acid",CC(=O)NC(CCC(O)=O)C(O)=O

SMILES	GSS Mapping Score
OCC1OC(=O)C(O)C1O	0.973333
CN(C)CC(O)=O	0.961538
OCC(OP(O)(O)=O)C(O)=O	0.960000
CC(C)=CC(O)=O	0.935897
CCC(O)CC(O)=O	0.925926
CN(CC(O)=O)C(N)=N	0.923077
OC(=O)CC1=CC=CC=C1	0.920000
OC(CN1C=CN=C1)C(O)=O	0.910256
OC(=O)C1CCCN1	0.897436
OC(=O)CCC(=O)CC(O)=O	0.892857
OC(=O)CC1=CC=C(O)C(O)=C1	0.888889
OC(=O)CC1=CN=CC=C1	0.884615
CC1=NC=C2C(=O)OCC2=C1O	0.880000
CN1CC(=CC=C1)C(O)=O	0.876543
COC1=CC(C(O)=O)=C(O)C=C1	0.864198
OCC(O)C(O)CC(O)=O	0.862069
OC(=O)CNC(=O)C=C	0.858974
OCC1OC(O)CC1O	0.851852
CN1C(=O)NC(=O)C2=C1N=CN2	0.847222

Table A.11: Compounds Similar to "3-Hydroxyglutaric acid",OC(CC(O)=O)CC(O)=O

SMILES	GSS Mapping Score
OC(=O)C1=CN=C(O)C=C1	0.877193
NC(=O)C1=CN=CC=C1	0.850000
OC(=O)C1=CC=CN1	0.833333
O=C1NC2=C(N1)C(=O)N=CN2	0.824561
NC1=CC=C(O)C=C1	0.816667
NC1=CC(=O)N=C(N)N1	0.800000
CC1=CNC(=O)NC1=O	0.793651
OC1=CC=C(O)C=C1	0.783333
NC1=NC(O)=NC2=C1NC(O)=N2	0.777778
OCC1=CNC(=O)NC1=O	0.772727
O=C1NC=CC(=O)N1	0.766667
CC(=CC(O)=O)C(O)=O	0.761905
OC(=O)C1=CC=CC(O)=C1O	0.757576
OC(=O)CNC(=O)C=C	0.753623
NC(CO)C(O)=O	0.750000
OC(=O)C1=CC(O)=CC=C1O	0.742424
NC1=CC=CC(=C1)C(O)=O	0.739130
CNCC(O)=O	0.733333
OC(CC(O)=O)C(O)=O	0.730159

Table A.12: Compounds Similar to "Orotic acid",OC(=O)C1=CC(=O)NC(=O)N1

SMILES	GSS Mapping Score
O=C1CCCN1	0.964912
CC(=O)C(C)=O	0.950000
CC(O)C=O	0.933333
C1CCOC1	0.929825
NCCC=O	0.916667
CC(O)CO	0.912281
CC(=O)C=O	0.900000
CC(N)C(O)=O	0.894737
OC(=O)C1=CC=CN1	0.882353
CCC(O)=O	0.866667
CNCC(O)=O	0.859649
OC(=O)CCC(O)=O	0.850000
CC(C)(O)C(O)=O	0.848485
CC(=CC(O)=O)C(O)=O	0.841270
CC(=NO)C(O)=O	0.833333
CC(O)=O	0.816667
OC(=O)C1CCC=N1	0.803030
NC(CO)C(O)=O	0.800000
CC1=CNC(=O)NC1=O	0.793651

Table A.13: Compounds Similar to Butanal,CCCC=O

SMILES	GSS Mapping Score
CNCC(O)C1=CC=CC=C1	0.945946
CNCCC1=CC=C(O)C=C1	0.909910
NCCC1=CC=C(O)C(O)=C1	0.900901
CC(N)C(O)C1=CC(O)=C(O)C=C1	0.891892
COC1=CC(=CC=C1O)C(O)CN	0.882883
CNC(C)C(O)C1=CC=CC=C1	0.875000
CN1C(CC(O)C1=O)C1=CN=CC=C1	0.873874
COC1=CC=C(CC(O)=O)C=C1	0.864865
OC(C(O)=O)C1=CC=C(O)C(O)=C1	0.855856
OCC(O)C1=CC(O)=C(OS(O)(=O)=O)C=C1	0.851852
COC1=CC=C2NC=C(CCO)C2=C1	0.850877
CNCCC1=CNC2=C1C=C(O)C=C2	0.850000
CC(=O)NC1=CC=C(O)C=C1	0.846847
CC1=NC=C(CO)C(C(O)=O)=C1O	0.837838
CNCC(O)C1=CC(O)=C(OS(O)(=O)=O)C=C1	0.837209
NCC1(O)OCC(O)C(O)C1O	0.828829
OC(CC1=CNC2=C1C=CC=C2)C(O)=O	0.824074
CC(C)C1=C(O)C=C(C)C=C1	0.819820
CC(=O)OC1=CC=CC=C1C(O)=O	0.810811

Table A.14: Compounds Similar to Epinephrine,CNCC(O)C1=CC=C(O)C(O)=C1

SMILES	GSS Mapping Score
OC(=O)CC1=CN=CC=C1	0.961538
OC(=O)CC1=CC=CC=C1O	0.948718
OC(=O)CC1=CC=C(O)C(O)=C1	0.925926
OC(=O)CNC(=O)C=C	0.923077
OC(=O)CNC(=O)C1=CC=CO1	0.913580
OCC1=CC=CC=C1	0.910256
OC(CN1C=CN=C1)C(O)=O	0.897436
OC(=O)C(=O)CC1=CC=CC=C1	0.892857
COC1=CC=CC=C1O	0.884615
OC(=O)C=CC1=CC=CC(O)=C1	0.876543
OC1=CC=C(O)C=C1	0.871795
CN1CC(=CC=C1)C(O)=O	0.864198
OC(=O)CC=CC1=CN=CC=C1	0.862069
OC(=O)C1=CC=CC(O)=C1O	0.858974
CC(=O)NC1=CC=C(O)C=C1	0.857143
CC1=NC=C2C(=O)OCC2=C1O	0.853333
CC(C)=CC(O)=O	0.846154
OC(=O)C=CC1=CC(O)=C(O)C=C1	0.845238
CCC(O)CC(O)=O	0.833333

Table A.15: Compounds Similar to "p-Hydroxyphenylacetic acid",OC(=O)CC1=CC=C(O)C=C1

SMILES	GSS Mapping Score
NC(=C)C(O)=O	0.866667
O=C1NC=CC(=O)N1	0.791667
CC(=O)C=O	0.761905
NC(N)=N	0.714286
OC(=O)C1=CC=CN1	0.705882
O=C1CN=CN1	0.692308
OC(=O)C=CC(O)=O	0.687500
CC(O)=O	0.666667
OC(=O)C=O	0.642857
NC(=O)C1=CC=CC=C1	0.636364
NC1=CC(=O)N=C(N)N1	0.633333
NCCC=O	0.629630
OC1=CC=C(Cl)C=C1Cl	0.627451
CC(=NO)C(O)=O	0.625000
CNC	0.622222
NC1=CC=C(O)C=C1	0.616667
O=C1NC2=C(N1)C(=O)N=CN2	0.614035
CNC(N)=N	0.607843
CC(O)C=O	0.604167

Table A.16: Compounds Similar to Acrylamide,NC(=O)C=C

SMILES	GSS Mapping Score
CN1C(=O)NC(=O)C2=C1N=CN2	0.930556
NC(=O)C1=CN=CC=C1	0.920000
CNC1=NC=NC2=C1NC=N2	0.888889
OC(=O)CC1=CN=CC=C1	0.880000
OC1CCNC1C(O)=O	0.846154
NC(=O)NCCC(O)=O	0.840000
CN(C=O)C1=CC=CC=C1	0.827160
CC1=CNC(=O)NC1=O	0.826667
C[N+]1=CC=CC(=C1)C(N)=O	0.814815
OCC1=CNC(=O)NC1=O	0.813333
CC(=O)NC1=CC=CC=C1	0.802469
OC(=O)C1=CC(Cl)=CC=C1	0.800000
OC(=O)C1=CNC2=CC=CC=C12	0.794872
CN1CC(=CC=C1)C(O)=O	0.790123
CNCC(O)=O	0.786667
CC1=CC(=CC=C1O)C(O)=O	0.782051
CC(=O)NC1=CC=C(O)C=C1	0.773810
CC(O)C(N)C(O)=O	0.773333
OC(=O)CC1=CC=CC=C1O	0.769231

Table A.17: Compounds Similar to N-Methylnicotinamide,CNC(=O)C1=CN=CC=C1

SMILES	GSS Mapping Score
NC(=N)NCCCC(O)C(O)=O	0.972973
NC(CCCC=O)C(O)=O	0.938596
NC(CC1=CC(O)=C(O)C=C1)C(O)=O	0.933333
CC(=O)NC(CCC(O)=O)C(O)=O	0.921053
NC(CCCC(O)=O)C(O)=O	0.912281
NC(=O)NCCC(O)=O	0.903509
NC(CCCN=C(N)NO)C(O)=O	0.900000
CC(=O)NC(CS)C(O)=O	0.894737
OC(=O)CNC(=O)CC1=CC=CC=C1	0.888889
NCC1(O)OCC(O)C(O)C1O	0.879630
NC(CCCC(N)C(O)=O)C(O)=O	0.878049
CCCCC(=O)NCC(O)=O	0.868421
CNCCCCC(N)C(O)=O	0.865079
OC(CC1=CNC2=C1C=CC=C2)C(O)=O	0.861111
CN1C(=O)NC(N)=C(NC(C)=O)C1=O	0.859649
NCCC(=O)C1=C(N)C=CC(O)=C1	0.851852
NC(CC(=O)C1=C(N)C(O)=CC=C1)C(O)=O	0.850000
COC1=CC2=C(NC=C2CC(O)=O)C=C1	0.842593
CC(C)CC(=O)NCC(O)=O	0.842105

Table A.18: Compounds Similar to Citrulline,NC(CCCNC(N)=O)C(O)=O

SMILES	GSS Mapping Score
OC(=O)C1=CC=CN1	0.954545
OC(=O)C1=CC(=O)NC(=O)N1	0.909091
OC(=O)CC1=CN=CC=C1	0.898551
OC(=O)C=CC=CC(O)=O	0.878788
OC(=O)C(=O)C1=CC=CC=C1	0.869565
CC(=CC(O)=O)C(O)=O	0.863636
NC1=CC=CC(=C1)C(O)=O	0.855072
CC(C)=CC(O)=O	0.833333
OC1=CC=C(O)C=C1	0.818182
OC(=O)C1=NC2=CC=CC=C2C=C1	0.807692
OC(CC(O)=O)C(O)=O	0.803030
OC1=CNC2=CC=CC=C12	0.797101
OC(=O)C1=CNC2=CC=CC=C12	0.794872
CN(C)CC(O)=O	0.787879
OC(=O)CC1=CC=CC=C1	0.786667
OC(CN1C=CN=C1)C(O)=O	0.782051
NC1=NC(O)=NC2=C1NC(O)=N2	0.777778
OC(=O)CC(=CC(O)=O)C(O)=O	0.773333
NC(C(O)=O)C(O)=O	0.772727

Table A.19: Compounds Similar to "Quinolinic acid",OC(=O)C1=CC=CN=C1C(O)=O

SMILES	GSS Mapping Score
OC(=O)C1=CC(O)=CC=C1O	0.954545
OC(=O)C1=CC=C(O)C=C1	0.939394
OC(=O)C1=CN=C(O)C=C1	0.924242
OC(=O)C1=CC=CN1	0.909091
OC1=CC=C(O)C=C1	0.893939
OCC1=CC=CC=C1	0.878788
OC(=O)C(=O)C1=CC=CC=C1	0.869565
OC1=CC=C(Cl)C=C1Cl	0.863636
OC1=CNC2=CC=CC=C12	0.855072
OC(=O)C1=CC(=O)NC(=O)N1	0.848485
OC(=O)CC1=CN=CC=C1	0.840580
OC(=O)C=CC(O)=O	0.833333
NC1=CC=CC(=C1)C(O)=O	0.826087
NC1=NC(O)=NC2=C1NC(O)=N2	0.825397
CC(C)=CC(O)=O	0.818182
NC(=O)C1=CN=CC=C1	0.803030
OC(=O)CC1=CC=CC=C1	0.800000
OC(=O)CC1=CC=C(O)C=C1	0.794872
OC(=O)CC(=CC(O)=O)C(O)=O	0.786667

Table A.20: Compounds Similar to Deoxyuridine,OC(=O)C1=CC=CC(O)=C1O

SMILES	GSS Mapping Score
NCCCC(O)=O	0.960000
OC(=O)CC1=CN=CC=C1	0.927536
OC(=O)CNC(=O)C=C	0.920000
NC1=CC=CC(=C1)C(O)=O	0.913043
CN(CC(O)=O)C(N)=N	0.910256
CN(C)CC(O)=O	0.880000
OCC(O)C(O)C(O)=O	0.875000
OC1=CNC2=CC=CC=C12	0.869565
NC(=O)NCCC(O)=O	0.866667
OC(=O)C1=CC(O)=C(O)C=C1	0.863636
OC1CCNC1C(O)=O	0.858974
COC1=CC=CC=C1O	0.855072
CC(C)=CC(O)=O	0.853333
CCC(=O)NCC(O)=O	0.851852
CN1C(=O)NC(=O)C2=C1N=CN2	0.847222
OC(CCC(O)=O)C(O)=O	0.846154
CC1=NC=C2C(=O)OCC2=C1O	0.840000
OC(CN1C=CN=C1)C(O)=O	0.833333
CNC(C)(C)C(O)=O	0.827586

Table A.21: Compounds Similar to L-Threonine,CC(O)C(N)C(O)=O

SMILES	GSS Mapping Score
NCCCC(O)=O	0.958333
OC(=O)CNC(=O)C=C	0.956522
NCCOP(O)(O)=O	0.927536
CC1=C(N)NC(=O)N=C1	0.924242
CC(=O)CN	0.902778
COC1=NC=CC(N)=N1	0.893939
NCCS(O)(=O)=O	0.888889
OC(=O)CC1=CN=CC=C1	0.884058
NC(=O)NCCC(O)=O	0.880000
OC1COC(O)C1O	0.878788
CN(C)CC(O)=O	0.875000
CN(CC(O)=O)C(N)=N	0.871795
NC1=CC=CC(=C1)C(O)=O	0.869565
OC(=O)C1=CC=C(O)C=C1	0.857143
OC(=O)C1CCCN1	0.853333
NC1=NC(=O)C2=C(N1)N=CN2	0.850000
CC(C)=CC(O)=O	0.847222
C1CN=C2N=CN=CC2=N1	0.841270
OC(=O)CCCC(O)=O	0.840000

Table A.22: Compounds Similar to "(R)-b-aminoisobutyric acid",CC(CN)C(O)=O

SMILES	GSS Mapping Score
NC1=NC2=C(NC(=O)N2C2OC(CO)C(O)C2O)C(=O)N1	0.900709
CN1C(=O)C=CN(C2OC(CO)C(O)C2O)C1=O	0.886525
OCC1OC(CC1O)N1C=NC2=C1N=CN1C=CN=C21	0.879433
CC1OC(C(O)C1O)N2C=NC3=C(N)N=CN=C23	0.865248
COC1C(O)C(CO)OC1N1C=NC2=C(N)N=CN=C12	0.857143
CC(=O)NC1C(O)OC(CO)C(O)C1O	0.843972
NC1=NC2=C(N=CN2C2OC(COP(O)(O)=O)C(O)C2O)C(=O)N1	0.839744
OCC1OC(C(O)C1O)C1=CNC(=O)NC1=O	0.836879
NC1=C2N=CN(C3OC(CO)C(OP(O)(O)=O)C3O)C2=NC=N1	0.836601
CC(O)C(O)C1=NC2=C(NC(N)=NC2=O)N=C1	0.808511
CC(O)C(O)C1=NC2=C(NC1)N=C(N)NC2=O	0.801418
CC(=O)NC1C(O)OC(CO)C(OS(O)(=O)=O)C1O	0.800000
CC(O)C(O)C1CNC2=NC(N)=NC(=O)C2(O)N1	0.794326
NC1=NC2=C(N=CN2C2CC(O)C(CO)O2)C(=O)N1	0.787234
CC(=O)NC1C(O)OC(COP(O)(O)=O)C(O)C1O	0.784314
COC1=CC2=C(NC=C2CCNC(C)=O)C=C1	0.780142
NC1=NC(=O)C2=NC(=CN=C2N1)C(O)C(O)CO	0.773050
NC1=NC=NC2=C1N=CN2C1OC(COP(O)(=O)OP(O)(O)=O)C(O)C1O	0.766082
OCC(O)C1=CC(O)=C(OS(O)(=O)=O)C=C1	0.765957

Table A.23: Compounds Similar to 1-Methylinosine,CN1C=NC2=C(N=CN2C2OC(CO)C(O)C2O)C1=O

SMILES	GSS Mapping Score
CC1=NC=C(CO)C(C(O)=O)=C1O	0.947917
NC(CO)CC1=CN=CN1	0.906250
OC1CCNC1C(O)=O	0.895833
OCC1OC(=O)C(O)C(O)C1O	0.885417
C1=CC(=CN=C1)CCCC(=O)O	0.878788
OC1COC(O)C(O)C1O	0.875000
COC1=CC=C(CC(O)=O)C=C1	0.864583
CCC(O)CC(O)=O	0.854167
OC(CCC(O)=O)C1=CC=CN=C1	0.852941
COC1=C(O)C=CC(=C1)C(O)C(O)=O	0.848485
OC(=O)CC1=CN=CC=C1	0.843750
OCCC1=CNC2=CC=C(O)C=C12	0.843137
NC(CCCC=O)C(O)=O	0.833333
COC1=CC(=CC=C1O)C(O)CO	0.828571
OC(C=O)C(O)C(O)C(O)C(O)=O	0.828283
OC1C(O)C(O)C(O)C(O)C1O	0.823529
NC(CCCC(O)=O)C(O)=O	0.822917
C1NC(CC=C1)C1=CN=CC=C1	0.813725
OCCNCCO	0.812500

Table A.24: Compounds Similar to Pyridoxine,CC1=NC=C(CO)C(CO)=C1O

SMILES	GSS Mapping Score
OC(=O)C1=NC2=CC=CC=C2C=C1	0.961538
OC(=O)CC1=CC=CC=C1O	0.948718
NC1=CC=CC(=C1)C(O)=O	0.935897
OC(=O)C=CC1=CC=CC(O)=C1	0.925926
OC(=O)C1=CC(Cl)=CC=C1	0.923077
OC(CN1C=CN=C1)C(O)=O	0.910256
OC1=CNC2=CC=CC=C12	0.897436
OC(=O)C1=NC2=C(O)C=CC=C2C(O)=C1	0.892857
CN1CC(=CC=C1)C(O)=O	0.888889
OC(=O)C1=CN=C(O)C=C1	0.884615
OC(=O)C1=CC(=O)C2=CC=CC=C2N1	0.880952
COC1=CC(C(O)=O)=C(O)C=C1	0.876543
OC1CCNC1C(O)=O	0.871795
OC(=O)CNC(=O)C1=CC=CO1	0.864198
CC(C)=CC(O)=O	0.858974
NC(C(O)=O)=C(C=CC=O)C(O)=O	0.857143
CC1=NC=C2C(=O)OCC2=C1O	0.853333
CC(=O)OC1=CC=CC=C1C(O)=O	0.850575
CCC(=C)C(O)=O	0.846154

Table A.25: Compounds Similar to "Cinnamic acid",OC(=O)C=CC1=CC=CC=C1

SMILES	GSS Mapping Score
CCCCCC(O)C=CC1C=CC(=O)C1CCCCCCC(O)=O	1.022727
CCCCCC(O)C=CC1C(O)CC2OC(CC12)=CCCCC(O)=O	1.015152
CCCCCC=CCC=CCC=CCCCCCCC(O)=O	1.011236
CC12CCC3C(CC=C4CC(O)CCC34C)C1CCC2C(=O)CO	1.007663
CC1(O)CCC2C3CCC4CC5=C(CC4(C)C3CCC12C)C=NN5	1.003876
CCCCCC(O)C=CC1C2CC(OO2)C1CC=CCCCC(O)=O	1.003745
CCCCCCCCCCCCCCCCCCC(O)=O	1.000000
CCCCCC(=O)CCC1C(O)CC(=O)C1CC=CCCCC(O)=O	0.996296
CCCCCC(OO)C=CC1C2CC(OO2)C1CC=CCCCC(O)=O	0.992593
CC1(O)CCC2C3CCC4C(O)C5=C(CC4(C)C3CCC12C)C=NN5	0.992337
OC(CCCCCCCCCC(O)=O)CC(O)=O	0.989011
CC(O)CCCC(O)C=CC1C(O)CC(=O)C1CC=CCCCC(O)=O	0.988889
CCCCCC=CCC=CCC=CCCCCC(O)=O	0.981685
CCCCCCCCCC(=O)CCC=CC=CC(O)CCCC(O)=O	0.978261
CC12CCC(O)CC1CCC1C3CCC(C(=O)CO)C3(C)CCC21	0.978022
CC12CCC(O)CC1CCC1C3CCC(O)(C(=O)CO)C3(C)CC(=O)C21	0.977778
CCCCCC=CCC=CCC=CCC=CCCCC(O)=O	0.974359
CC12CCC(=O)CC1CCC1C3CCC(O)(C(=O)CO)C3(C)CC(O)C21	0.974074
CCCCCC=CCC(O)C(O)C(O)C=CCC=CCCCC(O)=O	0.967391

Table A.26: Compounds Similar to "Stearic acid",CCCCCCCCCCCCCCCCCCC(O)=O

SMILES	GSS Mapping Score
CC(=O)NC1=CC=CC=C1	0.928571
CC1=CC(=CC=C1O)C(O)=O	0.904762
OC(=O)CC1=CC=C(O)C(O)=C1	0.901235
CN1CC(=CC=C1)C(O)=O	0.892857
COC1=CC(C(O)=O)=C(O)C=C1	0.888889
CC(=O)OC1=CC=CC=C1	0.880952
OC1CCNC1C(O)=O	0.869048
OC(=O)CC1=C(O)C=CC(O)=C1	0.864198
OC(=O)CC1=CC=CC(O)=C1	0.857143
CC(=O)OC1=CC=CC=C1C(O)=O	0.850575
CN(C=O)C1=CC=CC=C1	0.845238
OC(=O)CC1=CN=CC=C1	0.833333
CC(C)=CC(O)=O	0.821429
OC(=O)CNC(=O)C1=CC=CC=C1	0.817204
OC(=O)CC=CC1=CN=CC=C1	0.816092
OCC(O)C1OC(O)=C(O)C1=O	0.814815
OC(=O)CCC1=CC=CC=C1	0.811111
NCCCC(O)=O	0.809524
OC(=O)CC1=CC2=C(N1)C=CC=C2	0.800000

Table A.27: Compounds Similar to Acetaminophen,CC(=O)NC1=CC=C(O)C=C1

SMILES	GSS Mapping Score
NC(C(O)=O)C1=CC=C(C=C1)C(O)=O	0.931373
OC(=O)CNC(=O)C1=CC=CC=C1	0.921569
OC(=O)CNC(=O)C1=CC=CO1	0.911765
NC(CC1=CC=C(O)C=C1)C(O)=O	0.882353
OC(CC1=CNC2=C1C=CC=C2)C(O)=O	0.879630
OC(=O)CNC(=O)C=C	0.872549
NC(CC1=CC(Cl)=C(O)C=C1)C(O)=O	0.862745
C1NC(CC=C1)C1=CN=CC=C1	0.852941
NC(=O)NCCC(O)=O	0.843137
NC(CC1=CC(O)=C(O)C=C1)C(O)=O	0.838095
CN(CC(O)=O)C(N)=N	0.833333
CNCCC1=CC=C(O)C=C1	0.828571
CC(=O)NC(CCC(O)=O)C(O)=O	0.824074
CC1=C(C=C)C(NC1=O)=CC(N)=O	0.823529
CNCC(O)C1=CC=CC=C1	0.819048
COC1=C(O)C=C(C=CC(O)=O)C=C1	0.818182
NC(CCCC=O)C(O)=O	0.813725
CCCCC(=O)NCC(O)=O	0.810811
COC1=C(O)C=CC(C=CC(O)=O)=C1	0.808081

Table A.28: Compounds Similar to "4-Aminohippuric acid",NC1=CC=C(C=C1)C(=O)NCC(O)=O

SMILES	GSS Mapping Score
CC1=CC(=CC=C1O)C(O)=O	0.901235
COC1=CC=CC=C1O	0.888889
CC(=O)NC1=CC=C(O)C=C1	0.880952
OC(CN1C=CN=C1)C(O)=O	0.876543
CC(=O)OC1=CC=CC=C1	0.864198
OC1CCNC1C(O)=O	0.851852
CC(C)=CC(O)=O	0.839506
COC1=CC=C(CC(O)=O)C=C1	0.827957
CC(=O)OC1=CC=CC=C1C(O)=O	0.827586
OC(=O)CC1=CN=CC=C1	0.827160
OC1COC(O)C(O)C1O	0.821429
CCC(O)CC(O)=O	0.814815
OC(=O)C(=O)CC1=CC=CC=C1	0.809524
OC(=O)CC=CC1=CN=CC=C1	0.804598
CC(O)C(N)C(O)=O	0.802469
COC1=CC(CC(O)=O)=CC=C1O	0.802083
OC(=O)C1=NC2=CC=CC=C2C=C1	0.794872
CC(=NO)C(O)=O	0.790123
CC1=NC=C(CO)C(C(O)=O)=C1O	0.788889

Table A.29: Compounds Similar to "5-Methoxysalicylic acid",COC1=CC(C(O)=O)=C(O)C=C1

SMILES	GSS Mapping Score
OC1OC(COP(O)(O)=O)C(O)C(O)C1O	0.991870
OCC1OC(OP(O)(O)=O)C(O)C(O)C1O	0.983740
CC1OC(OP(O)(O)=O)C(O)C(O)C1O	0.952381
OC(COP(O)(O)=O)C(O)C=O	0.936508
NC1C(O)OC(COP(O)(O)=O)C(O)C1O	0.930233
OCC(OP(O)(O)=O)C(O)=O	0.928571
OCC(=O)COP(O)(O)=O	0.920635
NC1=NC(=O)C2=NC(=CN=C2N1)C(O)C(O)CO	0.905983
OCC(O)C(O)C(O)C(O)C(=O)CO	0.904762
NC1=NC2=C(N=C(C=N2)C(O)C(O)CO)C(=O)N1	0.883333
OCC(O)C(O)C(O)C(O)CO	0.880952
OC(C=O)C(O)C(O)C(O)C(O)=O	0.873016
OCC1OC(CC1O)N1C=NC2=C1N=CN=C2O	0.866667
OC1C(O)C(O)C(O)C(O)C1O	0.865079
OCC1OC(C(O)C1O)C1=CNC(=O)NC1=O	0.861789
C(C1C(C(C(C(O1)O)NS(=O)(=O)O)O)O)O	0.860465
NC1=NC(=O)C2=C(N1)N=CC(=N2)C(O)C(O)CO	0.858333
OCC1OC(=O)C(O)C(O)C1O	0.857143
OCC(O)C(O)CC(O)=O	0.849206

Table A.30: Compounds Similar to "Fructose 6-phosphate",OCC(=O)C(O)C(O)C(O)COP(O)(O)=O

SMILES	GSS Mapping Score
CN1C(=O)NC(=O)C2=C1N=CN2	0.819444
CN1CC(=O)NC1=N	0.818182
CC1=CNC(=O)NC1=O	0.803030
NC(=O)C1=CN=CC=C1	0.772727
NC1=NC(=O)C2=C(N1)N=CN2	0.766667
OCC1=CNC(=O)NC1=O	0.757576
NC1=NC=NC2=C1NC(O)=N2	0.750000
OC(=O)C1=CC(=O)NC(=O)N1	0.742424
CNC(=O)C1=CN=CC=C1	0.733333
C1CN=C2N=CN=CC2=N1	0.730159
CNCC(O)=O	0.727273
OC(=O)CNC(=O)C=C	0.724638
CNC1=NC=NC2=C1NC=N2	0.722222
NC(=O)NCCC(O)=O	0.720000
NC1=CC(=O)N=C(N)N1	0.712121
OC1=CNC2=CC=CC=C12	0.710145
CN1C(=O)NC2=C(NC(=O)N2)C1=O	0.705128
NC1=NC(O)=NC2=C1NC(O)=N2	0.698413
NC1=CC=C(O)C=C1	0.696970

Table A.31: Compounds Similar to 5-Methylcytosine,CC1=C(N)NC(=O)N=C1

SMILES	GSS Mapping Score
NC(CC1=CC(O)=C(O)C=C1)C(O)=O	0.923810
NC(CCCC=O)C(O)=O	0.898148
NCCC1=CC=C(O)C(O)=C1	0.888889
CNCCC1=CC=C(O)C=C1	0.879630
COC1=CC(=CC=C1O)C(O)CN	0.873874
NC(CCCC(O)=O)C(O)=O	0.870370
NC(CO)CC1=CN=CN1	0.861111
CC(N)C(O)C1=CC(O)=C(O)C=C1	0.855856
OC(=O)CNC(=O)CC1=CC=CC=C1	0.851852
NC(CC(=O)C1=C(N)C(O)=CC=C1)C(O)=O	0.850000
NCC1(O)OCC(O)C(O)C1O	0.842593
CSCCC(N)C(O)=O	0.833333
CNCC(O)C1=CC=C(O)C(O)=C1	0.828829
OC(=O)CNC(=O)C1=CC=CC=C1	0.824074
NC(CCCNC(N)=O)C(O)=O	0.815789
C1=CC(=CN=C1)CCCC(=O)O	0.814815
NC(=N)NCCCC(O)C(O)=O	0.810811
NC(=O)NCCC(O)=O	0.805556
CC(=O)NC(CCCN)C(O)=O	0.800000

Table A.32: Compounds Similar to 5-Hydroxykynurenamine,NCCC(=O)C1=C(N)C=CC(O)=C1

SMILES	GSS Mapping Score
CC1=CN(C2OC(CO)C(O)C2O)C(=O)NC1=O	0.949275
CN1C=NC2=C(N=CN2C2OC(CO)C(O)C2O)C1=O	0.900709
CC(=O)NC1C(O)OC(CO)C(O)C1O	0.898551
OCC1OC(C(O)C1O)N1N=CC2=C1NC=NC2=O	0.891304
OCC1OC(CC1O)N1C=NC2=C1N=CN1C=CN=C21	0.888889
NC1=NC2=C(NC(=O)N2C2OC(CO)C(O)C2O)C(=O)N1	0.879433
OCC1OC(C(O)C1O)C1=CNC(=O)NC1=O	0.862319
NC1=NC2=C(N=CN2C2CC(O)C(CO)O2)C(=O)N1	0.859259
CC1OC(C(O)C1O)N2C=NC3=C(N)N=CN=C23	0.855072
OC1OC(COP(O)(O)=O)C(O)C(O)C1O	0.847826
NC1=C2N=CN(C3OC(CO)C(OP(O)(O)=O)C3O)C2=NC=N1	0.843137
OC1=CC=C(C=C1)C1COC2=C(C1)C=CC(O)=C2	0.840909
OC(CCC(O)=O)C1=CC=CN=C1	0.840580
OC1=CC=C(CCC(=O)C2=C(O)C=C(O)C=C2O)C=C1	0.836879
CC(O)C(O)C1=NC2=C(NC1)N=C(N)NC2=O	0.833333
COC1C(O)C(CO)OC1N1C=NC2=C(N)N=CN=C12	0.829932
CC(O)C(O)C1CNC2=NC(N)=NC(=O)C2(O)N1	0.829787
OCC1OC(=O)C(O)C(O)C1O	0.826087
CC(=O)NC1C(O)OC(COP(O)(O)=O)C(O)C1O	0.823529

Table A.33: Compounds Similar to 3-Methyluridine,CN1C(=O)C=CN(C2OC(CO)C(O)C2O)C1=O

SMILES	GSS Mapping Score
OCC1OC(=O)C(O)C1O	0.970588
OC1OC(C(O)C(O)C1O)C(O)=O	0.950980
OCC(O)C(O)C(O)CO	0.941176
CC1=NC=C(CO)C(CO)=C1O	0.931373
OCC(O)C1OC(O)=C(O)C1=O	0.921569
COC1=CC(=CC=C1O)C(O)CO	0.914286
OC(=O)CCC(=O)CC(O)=O	0.911765
COC1=C(O)C=CC(=C1)C(O)C(O)=O	0.909091
OCCNCCO	0.901961
NCC1(O)OCC(O)C(O)C1O	0.898148
OC(=O)CC1=CC=CC(O)=C1	0.892157
OC(CCCC(O)=O)C(O)=O	0.882353
OCC(O)CO	0.872549
OC(=O)CC(O)(CC(O)=O)C(O)=O	0.862745
COC1=C(O)C=CC(C=CC(O)=O)=C1	0.858586
OC(CCC(O)=O)C1=CC=CN=C1	0.852941
CC(O)C(N)C(O)=O	0.843137
OCC(O)C(O)C(O)C(O)CO	0.842105
NC(CC1=CC(O)=C(O)C=C1)C(O)=O	0.838095

Table A.34: Compounds Similar to Alpha-D-Glucose,OCC1OC(O)C(O)C(O)C1O

SMILES	GSS Mapping Score
COC1=CC=C(CC(O)=O)C=C1	0.958333
COC1=CC(C(O)=O)=C(O)C=C1	0.927083
COC1=C(O)C=CC(=C1)C(O)C(O)=O	0.919192
CC(=O)OC1=CC=CC=C1C(O)=O	0.916667
COC1=C(O)C=CC(C=CC(O)=O)=C1	0.909091
OC1CC(=CC(O)C1O)C(O)=O	0.906250
CCOC(=O)CC(O)=O	0.885417
C1=CC(=CN=C1)CCCC(=O)O	0.878788
OC1COC(O)C(O)C1O	0.875000
OC(=O)CC1=CC2=C(N1)C=CC=C2	0.864583
COC1=CC(=CC=C1O)C(O)CO	0.857143
CC(=O)CCC(O)=O	0.854167
OC(=O)CCC(=O)CC(O)=O	0.843750
OC(CCC(O)=O)C1=CC=CN=C1	0.843137
COC1=CC2=C(NC=C2CC(O)=O)C=C1	0.842593
CC(CC(O)=O)C1=CC=CC=C1	0.838095
OC(CCCC(O)=O)C(O)=O	0.833333
OCCC1=CNC2=CC=C(O)C=C12	0.823529
OC(C(CC(O)=O)C(O)=O)C(O)=O	0.822917

Table A.35: Compounds Similar to "Homovanillic acid",COC1=CC(CC(O)=O)=CC=C1O

SMILES	GSS Mapping Score
NC(CCSSCCC(N)C(O)=O)C(O)=O	0.940299
NC(CCC(=O)NC(CC1=CC=CC=C1)C(O)=O)C(O)=O	0.890547
CCCC(=O)NC(CCN)CC(N)C(O)=O	0.885572
OC(=O)CCCCC1SCC2NC(=O)NC12	0.855721
CC(=O)NC1C(O)C(O)C(CO)OC1NC(=O)CC(N)C(O)=O	0.853535
NC(CCCNN=CNC(CC(O)=O)C(O)=O)C(O)=O	0.850746
NC(CCSCC1OC(C(O)C1O)N1C=NC2=C(N)N=CN=C12)C(O)=O	0.846847
NC(CC(O)=O)C(=O)NC(CC1=CC=CC=C1)C(O)=O	0.845771
CC(=O)NCCCCC(N)C(O)=O	0.840796
NCCCC(NC(=O)CCC(O)=O)C(O)=O	0.835821
CC(=O)NC(CCC(O)=O)C(O)=O	0.830846
COC(=O)C(CC1=CC=CC=C1)NC(=O)C(N)CC(O)=O	0.820896
CC(=O)NC(CS)C(O)=O	0.815920
NC(=N)NC1=NC(CSCCC(=N)NS(N)(=O)=O)=CS1	0.808081
CC(=O)NC1C(O)CC(O)(OC1C(OC(C)=O)C(O)CO)C(O)=O	0.805970
NC(CCCNC(N)=O)C(O)=O	0.800995
CC1=NC=C(C[N+]2=CSC(CCOP(O)(O)=O)=C2C)C(N)=N1	0.797980
CNCCCCC(N)C(O)=O	0.796020
NC(CCC(N)=O)C(O)=O	0.791045

Table A.36: Compounds Similar to S-Formylglutathione, NC(CCC(=O)NC(CSC=O)C(=O)NCC(O)=O)C(O)=O

SMILES	GSS Mapping Score
CC1=CNC(=O)NC1=O	0.909091
CC1=C(N)NC(=O)N=C1	0.803030
OC(=O)CNC(=O)C=C	0.797101
CN1CC(=O)NC1=O	0.787879
OC(=O)CC1=CN=CC=C1	0.782609
O=C1CCNC(=O)N1	0.772727
NC(=O)NCCC(O)=O	0.760000
OCC1=CC=CC=C1	0.757576
NC(CO)C(O)=O	0.742424
NC1=NC=NC2=C1NC(O)=N2	0.733333
OC1CCNC1C(O)=O	0.730769
OC(=O)C1=CC(=O)NC(=O)N1	0.727273
NC(=O)NC1NC(=O)NC1=O	0.722222
C1CN=C2N=CN=CC2=N1	0.714286
NCCC=O	0.712121
OC(=O)CCC1=CN=CN1	0.706667
CN(CC(O)=O)C(N)=N	0.705128
OC(=O)C1=CC=C(O)C=C1	0.698413
CNCC(O)=O	0.696970

Table A.37: Compounds Similar to 5-Hydroxymethyluracil,OCC1=CNC(=O)NC1=O

SMILES	GSS Mapping Score
NC(CCSCC(N)C(O)=O)C(O)=O	0.921875
OC(=N)CCCCC(S)CCS	0.890625
NC(CCC(=O)NC(CSC=O)C(=O)NCC(O)=O)C(O)=O	0.880597
NC(CCS(=O)(=O)CCC(N)C(O)=O)C(O)=O	0.859375
NCCCCC(NC(=O)C(N)CCC(O)=O)C(O)=O	0.854839
CSCCC(N)C(O)=O	0.843750
NC(CC1=CC(I)=C(OC2=CC=C(O)C(I)=C2)C(I)=C1)C(O)=O	0.839080
CCCC(=O)NC(CCN)CC(N)C(O)=O	0.828125
NC(CCCC=O)C(O)=O	0.812500
NC(CCSCC1OC(C(O)C1O)N1C=NC2=C(N)N=CN=C12)C(O)=O	0.810811
NC1=NC(=O)N(C=C1)C1OC(COP(O)(=O)OP(O)(O)=O)C(O)C1O	0.809524
OC(=O)CCCCC1SCC2NC(=O)NC12	0.807292
COC(=O)C(CC1=CC=CC=C1)NC(=O)C(N)CC(O)=O	0.796875
CC(=O)NC1C(O)C(O)C(CO)OC1NC(=O)CC(N)C(O)=O	0.792929
CNCCCCC(N)C(O)=O	0.781250
CC(=O)NC1C(O)CC(O)(OC1C(O)C(O)CO)C(O)=O	0.779661
NC(CC(O)=O)C(=O)NC(CC1=CC=CC=C1)C(O)=O	0.776042
NC(CC(=O)C1=C(N)C(O)=CC=C1)C(O)=O	0.770833
CCCC(CCC)C(O)=O	0.765625

Table A.38: Compounds Similar to L-Homocystine,NC(CCSSCCC(N)C(O)=O)C(O)=O

SMILES	GSS Mapping Score
COC1=CC=C(CC(O)=O)C=C1	0.924242
COC1=C(O)C=C(C=CC(O)=O)C=C1	0.909091
CC(C)C1=CC=C(C=O)C=C1	0.901515
COC1=C(O)C=CC(=C1)C(O)C(O)=O	0.893939
COC1=CC=C2NC=C(CCO)C2=C1	0.886364
OC(=O)CNCCC(=O)C1=CC=CC=C1	0.878788
CC(C)C1=C(O)C=C(C)C=C1	0.871212
CC(=O)OC1=CC=CC=C1	0.863636
CC(C1=CC=C(O)C=C1)C(=O)C1=C(O)C=C(O)C=C1	0.862319
CC(=O)OC1=CC=CC=C1C(O)=O	0.856061
CC(=C)C1CCC(CO)=CC1	0.848485
NC(CC(=O)C1=C(N)C(O)=CC=C1)C(O)=O	0.840909
CN1C(=O)C=CN(C2OC(CO)C(O)C2O)C1=O	0.840580
COC1=CC2=C(NC=C2CCNC(C)=O)C=C1	0.840278
OC(=O)CNC(=O)C1=CC=CC=C1	0.833333
CN(C=O)C1=CC=CC=C1	0.825758
CC(C)CC1=CC=C(C=C1)C(C)C(O)=O	0.820000
OC(=O)C(=O)CC1=CC=CC=C1	0.818182
NC(CC1=CC=C(OC2=CC=C(O)C=C2)C=C1)C(O)=O	0.816327

Table A.39: Compounds Similar to Naproxen,COC1=CC2=C(C=C1)C=C(C=C2)C(C)C(O)=O

SMILES	GSS Mapping Score
NC1=NC(=O)C2=NC(=CN=C2N1)C(O)C(O)CO	0.923077
NC1=NC(=O)C2=C(N1)N=CC(=N2)C(O)C(O)CO	0.900000
NC1=NC(=O)C2=C(NCC(=N2)C(=O)C(O)CO)N1	0.837398
NC1=NC2=C(N=CN2C2CC(O)C(CO)O2)C(=O)N1	0.822222
CC(O)C(O)C1=NC2=C(NC(N)=NC2=O)N=C1	0.816667
OCC1OC(C(O)C1O)C1=CNC(=O)NC1=O	0.813008
OCC1OC(C(O)C1O)N1C=NC2=C1N=CNC2=O	0.809524
CC(O)C(O)C1=NC2=C(NC1)N=C(N)NC2=O	0.801587
NC1=NC2=C(NC(=O)N2C2OC(CO)C(O)C2O)C(=O)N1	0.794326
NC1=NC(=O)N(C=C1)C1CC(O)C(CO)O1	0.793651
COC1=CC(=CC=C1O)C(O)CN	0.791667
NCC1(O)OCC(O)C(O)C1O	0.783333
COC1=CC2=C(NC=C2CC(O)=O)C=C1	0.775000
C(C1C(C(C(C(O1)O)NS(=O)(=O)O)O)O)O	0.767442
NC(CCCNC(N)=O)C(O)=O	0.766667
NCC1=C(CC(O)=O)C(CCC(O)=O)=CN1	0.765152
NC(CCCC(O)=O)C(O)=O	0.758333
OC1OC(COP(O)(O)=O)C(O)C(O)C1O	0.756098
NC1C(O)OC(COP(O)(O)=O)C(O)C1O	0.751938

Table A.40: Compounds Similar to Neopterin,NC1=NC2=C(N=C(C=N2)C(O)C(O)CO)C(=O)N1

SMILES	GSS Mapping Score
OCC1OC(CC1O)N1C=NC2=C1N=CN=C2O	0.975000
OC(=O)C1CCCCN1	0.963768
NCC1=C(CC(O)=O)C(CCC(O)=O)=CN1	0.962121
CCCC(CCC)C(O)=O	0.949275
OC1CCC(CC1)CC(O)=O	0.942029
CC(=O)NC(CCCNC(N)=N)C(O)=O	0.929078
NC(CCCC=O)C(O)=O	0.927536
CNCCCCC(N)C(O)=O	0.920290
OC(=O)CCCC1=CNC2=CC=CC=C12	0.913043
CC(=O)NC1C(O)OC(CO)C(O)C1O	0.909091
OC(CCC(O)=O)C1=CC=CN=C1	0.905797
OCC1OC(C(O)C1O)N1N=CC2=C1NC=NC2=O	0.904762
NC1=NC2=C(N=CN2C2CC(O)C(CO)O2)C(=O)N1	0.903704
OC1=CC=C(CCC(=O)C2=C(O)C=C(O)C=C2O)C=C1	0.900709
NC(CCCNC(N)=O)C(O)=O	0.898551
CN1CCCC1C1=CC=C[N+](C)=C1	0.896296
CC(CC(O)=O)CC(O)=O	0.891304
CC(C)CC1=CC=C(C=C1)C(C)C(O)=O	0.886667
COC1=CC2=C(C=C1)C=C(C=C2)C(C)C(O)=O	0.886364

Table A.41: Compounds Similar to "Decenedioic acid",OC(=O)CCCCCCC=CC(O)=O

SMILES	GSS Mapping Score
CNCCC1=CC=C(O)C=C1	0.942857
COC1=C(O)C=CC(=C1)C(O)C(O)=O	0.939394
CC(N)C1=CC=CC=C1	0.933333
COC1=C(O)C=C(C=CC(O)=O)C=C1	0.929293
CNCC(O)C1=CC=C(O)C(O)=C1	0.918919
OC(=O)CNC(=O)C1=CC=CC=C1	0.914286
COC1=CC=C(CC(O)=O)C=C1	0.904762
OCCC1=CNC2=CC=C(O)C=C12	0.901961
C[N+]1=CC=CC(=C1)C(N)=O	0.895238
COC1=CC(=CC=C1O)C(O)CN	0.891892
OC(=O)CNC(=O)CC1=CC=CC=C1	0.888889
CN(C=O)C1=CC=CC=C1	0.885714
OC(CCC(O)=O)C1=CC=CN=C1	0.882353
OC(CC1=CNC2=C1C=CC=C2)C(O)=O	0.879630
OC(=O)CC1=CC2=C(N1)C=CC=C2	0.876190
CNC(C)C(O)C1=CC=CC=C1	0.875000
CC(=O)NC1=CC=C(O)C=C1	0.866667
CC(N)C(O)C1=CC(O)=C(O)C=C1	0.864865
OCC1=CC=CC=C1	0.857143

Table A.42: Compounds Similar to N-Methylphenylethanolamine, CNCC(O)C1=CC=CC=C1

SMILES	GSS Mapping Score
OC=O	0.666667
NC=O	0.500000
C	0.476190
OC(=O)C=O	0.444444
OCC=O	0.424242
O	0.400000
OCC(O)=O	0.388889
NC(N)=O	0.363636
CC(=O)C=O	0.358974
O=C=O	0.333333
O=C1C=CC(=O)C=C1	0.312500
CC(C)=O	0.311111
O=C1CN=CN1	0.307692
CC(O)C=O	0.291667
NC(=O)C=C	0.285714
NCCC=O	0.259259
NC(N)=N	0.250000
CNCC(O)=O	0.245614
OC(=O)C1=CC=CN1	0.235294

Table A.43: Compounds Similar to Formaldehyde,C=O

SMILES	GSS Mapping Score
OCC1OC(=O)C(O)C(O)C1O	0.967742
OCC1OC(=O)C(O)C1O	0.956989
OCC(O)C(O)C(=O)CO	0.946237
CC1=NC=C(CO)C(C(O)=O)=C1O	0.944444
OCC1OC(O)CC1O	0.935484
OC1CC(=CC(O)C1O)C(O)=O	0.913978
OC1OC(C(O)C(O)C1O)C(O)=O	0.906250
OC(=O)CCC(=O)CC(O)=O	0.903226
OC1COC(O)C(O)C1O	0.892473
OCC1OC(O)C(O)C(O)C1O	0.882353
COC1=CC=C(CC(O)=O)C=C1	0.881720
CC1=NC=C(CO)C(CO)=C1O	0.875000
OCC1OC(O)(CO)C(O)C1O	0.872549
OC(COP(O)(O)=O)C(O)C=O	0.870968
OC(=O)CC1=CC2=C(N1)C=CC=C2	0.866667
COC1=CC(CC(O)=O)=CC=C1O	0.864583
CC(CCC(O)=O)C(O)=O	0.860215
OCC(O)C(O)C(O)CO	0.854167
OC(=O)CC=CC1=CN=CC=C1	0.849462

Table A.44: Compounds Similar to 3-Keto-b-D-galactose,OCC1OC(O)C(O)C(=O)C1O

SMILES	GSS Mapping Score
CN1C(=O)NC(=O)C2=C1N=CN2	0.897436
CC1=CNC(=O)NC1=O	0.858974
CNC(=O)C1=CN=CC=C1	0.846154
CNC1=NC=NC2=C1NC=N2	0.833333
OCC1=CNC(=O)NC1=O	0.820513
CN1C(=O)N(C)C2=C(NC(=O)N2)C1=O	0.806452
NC(=O)NCCC(O)=O	0.794872
CN(CC(O)=O)C(N)=N	0.782051
CN1C=NC2=C1C(=O)N(C)C(=O)N2	0.781609
C[N+]1=CC=CC(=C1)C(N)=O	0.777778
OC1CCNC1C(O)=O	0.769231
CC(=O)NC1=CC=CC=C1	0.765432
CC(=O)NC1=CC=C(O)C=C1	0.761905
NC(=O)NC1NC(=O)NC1=O	0.756410
CNCC(O)=O	0.743590
CC1=NC=C2C(=O)OCC2=C1O	0.733333
OC(CN1C=CN=C1)C(O)=O	0.730769
CN(C=O)C1=CC=CC=C1	0.728395
OCCNCCO	0.717949

Table A.45: Compounds Similar to "1-Methyluric acid",CN1C(=O)NC2=C(NC(=O)N2)C1=O

SMILES	GSS Mapping Score
CC1(O)CCC2C3CCC4C(O)C5=C(CC4(C)C3CCC12C)C=NN5	0.977011
CC(=O)C1(O)CCC2C3CCC4=CC(=O)CCC4(C)C3CCC12C	0.953488
CC12CCC(=O)CC1CCC1C2CCC2(C)C1CCC2=O	0.949612
CC1CC2=CC(=O)CCC2(C)C2CCC3(C)C(CCC3(C)O)C12	0.945736
CC(O)C1CCC2C3CCC4=CC(=O)CCC4(C)C3CCC12C	0.941860
CC(=O)C1CCC2C3CCC4=CC(=O)CCC4(C)C3C(O)CC12C	0.937984
CCC1(O)CCC2C3CCC4=CC(=O)CCC4=C3C=CC12CC	0.934109
CC12CCC3C(CC=C4CC(O)CCC34C)C1CCC2C(=O)CO	0.927203
CC12CCC3C(CCC4CC(O)CCC34C)C1CCC2O	0.926357
COC1=CC2=C(CCC3C2CCC2(C)C(O)CCC32)C=C1O	0.922481
CC(CCC1=C(C)CCCC1(C)C)=CC=CC(C)=CC(O)=O	0.918605
CC12CCC(=O)CC1CCC1C3CCC(O)(C(=O)CO)C3(C)CC(O)C21	0.918519
CC12CCC3C(CC=C4CC(O)CCC34C)C1CC(O)C2O	0.914729
CC1CC2C(CCC3(C)C2CCC3(O)C(C)=O)C2(C)CCC(=O)C=C12	0.911111
CC12CCC3C(CCC4=CC(=O)CCC34C)C1CCC2OS(O)(=O)=O	0.910853
CC12CCC(O)CC1CCC1C3CCC(C(=O)CO)C3(C)CCC21	0.908425
CCCCCC(O)C=CC1C(CC=CCCCC(O)=O)C=CC1=O	0.906977
COC1=C(O)C=CC(CNC(=O)CCCCC=CC(C)C)=C1	0.899225
CC(C)CCCC(C)CCCC(C)C(O)=O	0.895349

Table A.46: Compounds Similar to Stanozolol, CC1(O)CCC2C3CCC4CC5=C(CC4(C)C3CCC12C)C=NN5

SMILES	GSS Mapping Score
OCC1OC(=O)C(O)C(O)C1O	0.970588
OCC1OC(=O)C(O)C1O	0.960784
CC(O)(CCO)CC(O)=O	0.950980
COC1=C(O)C=CC(=C1)C(O)C(O)=O	0.949495
COC1=CC(=CC=C1O)C(O)CO	0.942857
OC(COP(O)(O)=O)C(O)C=O	0.941176
OCC(O)CO	0.931373
CC(CC(O)=O)CC(O)=O	0.921569
NCC1(O)OCC(O)C(O)C1O	0.916667
CCC(O)CC(O)=O	0.911765
OC1C(O)C(O)C2OP(O)(=O)OC2C1O	0.907407
NC(CCCC(O)=O)C(O)=O	0.901961
OCC(O)C(O)C(O)C(O)C(O)=O	0.898148
OC(CCC(O)=O)C1=CC=CN=C1	0.892157
OC1CCNC1C(O)=O	0.882353
OCCNCCO	0.872549
OCC(O)C1=CC(O)=C(OS(O)(=O)=O)C=C1	0.870370
CNCC(O)C1=CC=C(O)C(O)=C1	0.864865
CC(C)(O)C(O)=O	0.862745

Table A.47: Compounds Similar to Myoinositol,OC1C(O)C(O)C(O)C(O)C1O

SMILES	GSS Mapping Score
NCCC=O	0.800000
CC1=C(N)NC(=O)N=C1	0.787879
CC(=O)C(C)=O	0.783333
CC1=CNC(=O)NC1=O	0.777778
NC(=O)NCCC(O)=O	0.773333
O=C1NC2=C(N1)C(=O)N=CN2	0.771930
O=C1CN=CN1	0.766667
OCC1=CNC(=O)NC1=O	0.757576
OC(=O)C1=CC(=O)NC(=O)N1	0.750000
CNCC(O)=O	0.733333
CN1CC(=O)NC1=N	0.730159
OC(=O)C1CCC=N1	0.727273
OC(=O)CNC(=O)C=C	0.724638
OC1=CC=C(O)C=C1	0.722222
OC(=O)C1=CN=C(O)C=C1	0.719298
CCC(O)=O	0.716667
NCCS(O)(=O)=O	0.714286
OCCCC(O)=O	0.712121
CN1C(=O)NC(=O)C2=C1N=CN2	0.708333

Table A.48: Compounds Similar to Dihydrouracil,O=C1CCNC(=O)N1

SMILES	GSS Mapping Score
NC(=O)C1=CC=CC=C1	0.878788
NC1=CC=C(O)C=C1	0.873016
OC(=O)C1=CC=CN1	0.857143
OC(=O)C1=CC=CN=C1C(O)=O	0.848485
OC1=CC=C(O)C=C1	0.841270
OCC1=CC=CC=C1	0.833333
NC1=CC=CC(=C1)C(O)=O	0.826087
OC1=CC=C(Cl)C=C1Cl	0.825397
OC(=O)C=CC=CC(O)=O	0.818182
OC1=CNC2=CC=CC=C12	0.811594
OC(=O)C1CCC=N1	0.809524
OC(=O)C1=CC=CC(O)=C1O	0.803030
COC1=CC=CC=C1O	0.797101
O=C1C=CC(=O)C=C1	0.793651
CC(=CC(O)=O)C(O)=O	0.777778
OC(=O)CC1=CN=CC=C1	0.768116
NC(=O)C=C	0.761905
COC1=NC=CC(N)=N1	0.757576
OC(=O)C=CC1=CC=CC=C1	0.756410

Table A.49: Compounds Similar to Indole,N1C=CC2=C1C=CC=C2

SMILES	GSS Mapping Score
COC1=CC2=C(NC=C2CC(O)=O)C=C1	0.937500
COC1=CC2=C(NC=C2CCN(C)C)C=C1	0.913333
CC(=O)NCCCCC(N)C(O)=O	0.881944
COC1=CC2=C(C=C1)C=C(C=C2)C(C)C(O)=O	0.875000
NCCCC(NC(=O)CCC(O)=O)C(O)=O	0.861111
CC(=O)NC(CCC(O)=O)C(O)=O	0.854167
CC(=O)NC1C(O)OC(CO)C(O)C1O	0.847222
COC(=O)CNC(=O)CC=C	0.840278
OC(=O)CNCCC(=O)C1=CC=CC=C1	0.833333
OC(=O)CNC(=O)CC1=CC=CC=C1	0.826389
CC(=O)NC1=CC=C(O)C=C1	0.819444
CC(=O)NC1C(O)OC(CO)C(OS(O)(=O)=O)C1O	0.813333
CN1CCCC1C1=CC=C[N+](C)=C1	0.812500
CC(=O)NC(CC(O)=O)C(O)=O	0.805556
NC(CC(O)=O)C(=O)NC(CC1=CC=CC=C1)C(O)=O	0.805031
NC(CC1=CC=C(OC2=CC=C(O)C=C2)C=C1)C(O)=O	0.802721
CN1C=NC2=C(N=CN2C2OC(CO)C(O)C2O)C1=O	0.801418
CN1C(=O)C=CN(C2OC(CO)C(O)C2O)C1=O	0.798611
CC(=O)NC1C(O)OC(COP(O)(O)=O)C(O)C1O	0.797386

Table A.50: Compounds Similar to Melatonin,COC1=CC2=C(NC=C2CCNC(C)=O)C=C1