SANDIA REPORT

SAND2010-2515
Unlimited Release
Printed April 2010

Signature Molecular Descriptor: Advanced Applications

Donald P. Visco, Jr.

Prepared by

Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U. S. Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia

Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2010-2515
Unlimited Release
Printed April 2010

Signature Molecular Descriptor: Advanced Applications

Donald P. Visco, Jr.
Tennessee Technological University
Box 5013
Cookeville, TN 38505
Contact: dvisco@tntech.edu
(931)-372-3606

ABSTRACT

In this work we report on the development of the Signature Molecular Descriptor (or Signature) for use in the solution of inverse design problems as well as in high-throughput screening applications. The ultimate goal of using Signature is to identify novel and non-intuitive chemical structures with optimal predicted properties for a given application. We demonstrate this in three studies: green solvent design, glucocorticoid receptor ligand design and the design of inhibitors for Factor XIa.

Acknowledgments

This work was funded by the U.S. Department of Energy 2004 PECASE Award with the title "Signature Molecular Descriptor: Advanced Applications". First and foremost, I acknowledge Grant Heffelfinger for providing the first opportunity given to the author to work at Sandia National Laboratories. The author also acknowledges the efforts of Jean-Loup Faulon who was the inspiration and inventor of the Signature Molecular Descriptor. Finally, the author acknowledges the help of Danny Rintoul, Carla Churchwell, Shawn Martin, Archana Kotu, Ramdas Pophale, Joshua Jackson and Derick Weis for their contributions to the development of this technology.

This page is intentionally left blank

This page is intentionally left blank

Contents

ABSTRACT 3

ACKNOWLEDGMENTS 4

1 INTRODUCTION 10

2 THE SIGNATURE MOLECULAR DESCRIPTOR 10

2.1 What is Signature? 10

3 COMPUTER-AIDED MOLECULAR DESIGN (CAMD) 12

3.1 General Features 12

3.2 CAMD Algorithm with the Signature Molecular Descriptor 12

4 CAMD APPLICATIONS 17

4.1 Solvent Design 17
4.1.1 Motivation 17
4.1.2 Results 19

4.2 Design of Glucocorticoid Receptor Ligands 27
4.2.1 Motivation 27
4.2.2 Results 28

5 HIGH-THROUGHPUT SCREENING USING SIGNATURE 35

5.1 Introduction 35

5.2 Methods 36
5.2.1 Support Vector Machines 36
5.2.2 Feature Selection 37
5.2.3 Overlap Metric 38
5.2.4 Training and Test Sets 38

5.3 Results 39
5.3.1 Filter Methods and Wrapper Methods 39
5.3.2 False Positive Testing 41
5.3.3 Screening PubChem 41
5.3.4 Docking of PubChem Compounds 45

6 REFERENCES 46

List of Figures

FIGURE 1: ETHYLENE GLYCOL AND ITS CORRESPONDING HEIGHT-1 MOLECULAR SIGNATURE. 11

FIGURE 2: THE COMPUTER-AIDED MOLECULAR DESIGN ALGORITHM USING THE SIGNATURE MOLECULAR DESCRIPTOR. 13

FIGURE 3: THE ENVIRONMENTAL IMPACT QSPR STATISTICS ARE PLOTTED AS A FUNCTION OF THE NUMBER OF INDEPENDENT VARIABLES. 20

FIGURE 4: A FLOW CHART FOR THE OVERALL INVERSE DESIGN PROCESS USING THE SIGNATURE MOLECULAR DESCRIPTOR. 24

FIGURE 5: FEATURE SELECTION USING A FILTERING APPROACH WITH PERFORMANCE EVALUATED BY 10-FOLD CROSS-VALIDATION. 39

FIGURE 6: FEATURE SELECTION USING A WRAPPING APPROACH WITH PERFORMANCE EVALUATED BY 10-FOLD CROSS-VALIDATION. 40

List of Tables

TABLE 1: DATA FROM GSK'S SOLVENT SELECTION GUIDE. 18

TABLE 2: UNIQUE HEIGHT 1 ATOMIC SIGNATURES COLLECTED FROM GSK'S SOLVENT SELECTION GUIDE..... 19

TABLE 3: STATISTICS FOR THE HEIGHT 1 SIGNATURE QSPRS. 21

TABLE 4: COEFFICIENTS FOR THE NON-LINEAR HEIGHT 1 QSPRS. 22

TABLE 5: FOCUSED DATABASE SAMPLES FROM CLASSES DEFINED IN GSK'S SOLVENT SELECTION GUIDE..... 25

TABLE 6: SAMPLES OF HYBRID STRUCTURES FROM THE FOCUSED DATABASE. 26

TABLE 7: TRAINING SET COMPOUNDS AND ACTIVITY DATA 28

TABLE 8: REGRESSION COEFFICIENTS AND STATISTICS FOR THE QSPRS 30

TABLE 9: MODIFIED RULE OF FIVE PARAMETERS 32

TABLE 10: SELECTED CANDIDATES IN THE FOCUSED DATABASE AND THEIR PREDICTED RRBA 33

TABLE 11: PREDICTION STATISTICS FOR THE SVM ON FACTOR XIA INHIBITOR DATA WITH 22 CLUSTERS (105 SIGNATURES). 40

TABLE 12: TEST SET PREDICTION STATISTICS FROM INACTIVE FACTOR XIA INHIBITORS IN AID 798. THE NUMBERS IN PARENTHESIS INDICATE THE NUMBER OF COMPOUNDS FROM AID 798 WHICH WERE ABOVE THE VARIOUS Ω VALUES WHILE THE TABLE ENTRIES INDICATE THOSE COMPOUNDS IDENTIFIED. 41

TABLE 13: SCREENING PUBCHEM COMPOUNDS FOR NEW FACTOR XIA INHIBITORS WITH SVM. PREDICTED PERCENT ACTIVE, PER THE LISTED OVERLAP METRIC, IS PROVIDED IN SQUARE BRACKETS..... 42

TABLE 14: A SAMPLE OF ONE DOZEN COMPOUNDS FROM THE Ω =1 SET. REPORTED ARE THE COMPOUND ID NUMBER FROM PUBCHEM, 2D-STRUCTURE, MAXIMUM TANIMOTO COEFFICIENT (WITH AID 846), MAGNITUDE OF THE DECISION FUNCTION (SVM), AND BINDING ENERGY (KCAL/MOL) FROM AUTODOCK. 43

1 Introduction

In many areas of engineering, compounds are designed and/or modified in incremental ways which rely upon heuristics or institutional knowledge. Often multiple experiments are performed and the optimal compound is identified in this brute-force fashion. Perhaps a traditional chemical scaffold is identified and movement of a substituent group around a ring constitutes the whole of the design process. Also notably, a chemical being evaluated in one area might demonstrate properties very attractive in another area and serendipity was the mechanism for solution.

In contrast to such approaches, computer-aided molecular design (CAMD) looks to encompass both experimental and heuristic-based knowledge into a strategy that will design a molecule on a computer to meet a given target. Depending on the algorithm employed, the molecule which is designed might be quite novel (re: no CAS registration number) and/or non-intuitive relative to what is known about the problem at hand.

While CAMD is a fairly recent strategy (dating to the early 1980s)[1], it contains a variety of bottlenecks and limitations which have prevented the technique from garnering more attention in the academic, governmental and industrial institutions. A main reason for this is how the molecules are described in the computer. This step can control how models are developed for the properties of interest on a given problem as well as how to go from an output of the algorithm to an actual chemical structure.

This report provides details on a technique to describe molecules on a computer, called Signature, as well as the computer-aided molecule design algorithm built around Signature. Two applications are provided of the CAMD algorithm with Signature. The first describes the design of green solvents based on data in the GlaxoSmithKline (GSK) Solvent Selection Guide. The second provides novel non-steroidal glucocorticoid receptor ligands with some optimally predicted properties.

In addition to using the CAMD algorithm with Signature, it is demonstrated how to employ Signature in a high-throughput screening study. Here, after classifying both active and inactive inhibitors for the protein Factor XIa using Signature, the model developed is used to screen a large, publicly-available database called PubChem for the most active compounds.

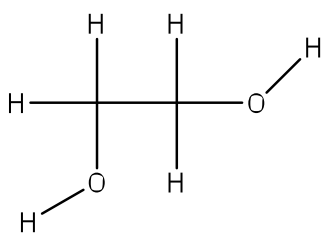
2 The Signature Molecular Descriptor

2.1 What is Signature?

Signature, which has its origins in structural elucidation studies of Faulon, [2, 3] is based on the molecular graph of a molecule, $G = (V_G, E_G)$, where the elements in V_G denote the atoms in the molecule, and the *edges* of E_G correspond to the bonds between those atoms. In this context, a molecule is characterized by a set of canonical sub-graphs, each *rooted* on a different vertex with a predefined level of branching, which we refer to as the height h. The branching of a vertex is an extended degree sequence that describes the local neighborhood, up to the distance h away from the root.

We define an *atomic* Signature, ${}^{h}\sigma_G(x)$, as the canonical sub-graph of G consisting of all atoms a distance h from the root x. A *molecular* Signature, ${}^{h}\Sigma_G$, is then the set (re: sum) of all unique atomic Signatures and the occurrence with which they appear in the molecular graph. Even though the atomic Signatures are unique, they are, by construction, interrelated allowing information about the overall structure of the molecule to be conveyed at the end. [4]

The atomic Signatures make up the set of molecular descriptors for a molecule. These are expressed in terms of a string of characters that correspond to the canonized sub-graph in a breadth-first order. Branch levels are indicated by a set of parentheses following the parent vertex. An example of the molecular Signature for ethylene glycol at height-1 is given in Figure 1.



<u>${}^{1}\Sigma$ (Ethylene Glycol)</u>
2 c(c o h h)
2 o(c h)
4 h(c)
2 h(o)

Figure 1: Ethylene glycol and its corresponding height-1 molecular Signature. The molecular Signature of height 1 is the sum of the 10 height-1 atomic Signatures. Note that only four height-1 atomic Signatures are unique for ethylene glycol, with the occurrence numbers given to the left of each atomic Signature in the figure. Bonding type is accounted for in the atomic Signature. For this molecule, only single bonds occur, though double, triple and aromatic bonds can be accommodated.

Note that the height of the atomic Signature selected for use is a design parameter. At height-0, atomic Signatures are just the atoms in a molecule (and, hence, the molecular formula). On the other hand, large heights are very detailed, specific information on the bonding about a particular atom several bond-lengths away. However, once a height is selected (usually height-1 or height-2, which balances computational issues with that of specificity of information), [5] this single height is used for the remainder of the problem.

3 Computer-Aided Molecular Design (CAMD)

3.1 General Features

In CAMD, there are basically three main steps to the overall algorithm: (1) selection of groups or fragments, (2) making of the new molecules from the groups or fragments and (3) evaluating the newly-developed molecules. In order to assess the fitness of the solutions output from a CAMD algorithm (step 3 above), a scoring function is required which is germane to the problem of interest. Normally, this is a quantitative structure-property relationship, or QSPR. Basically, a QSPR is a mathematical expression which purports to describe a property of a molecule based on the molecule's structure. It was introduced in the 1960's with the work of Hansch [6] and is still an active area of research [7] with a rich history. [8] While molecular properties themselves or whole-molecule descriptors can be used as independent variables in a QSPR, a popular approach is to use independent variables based on sub-parts of the molecule. For example, group-contribution techniques decompose a molecule into smaller groups, or fragments, where each group provides some contribution to a predicted molecular property. [9] Such approaches are well highlighted in The Properties of Gases and Liquids. [10] Other techniques examine a 2-D graphical representation of a molecule where atoms are nodes and bonds are edges. Here, an operator on some portion of the molecular graph plays the role of independent variable and many of these descriptors exist in the literature today. [11] The Signature molecular descriptor belongs to this fragment-class of descriptors.

3.2 CAMD Algorithm with the Signature Molecular Descriptor

In the previous section, three general steps were listed in a CAMD algorithm. For clarity, a workflow diagram for the use of the CAMD technique with the Signature molecular descriptor containing nine steps is provided in Figure 2. Additionally, the importance of each step will be discussed and how the step is ultimately implemented.

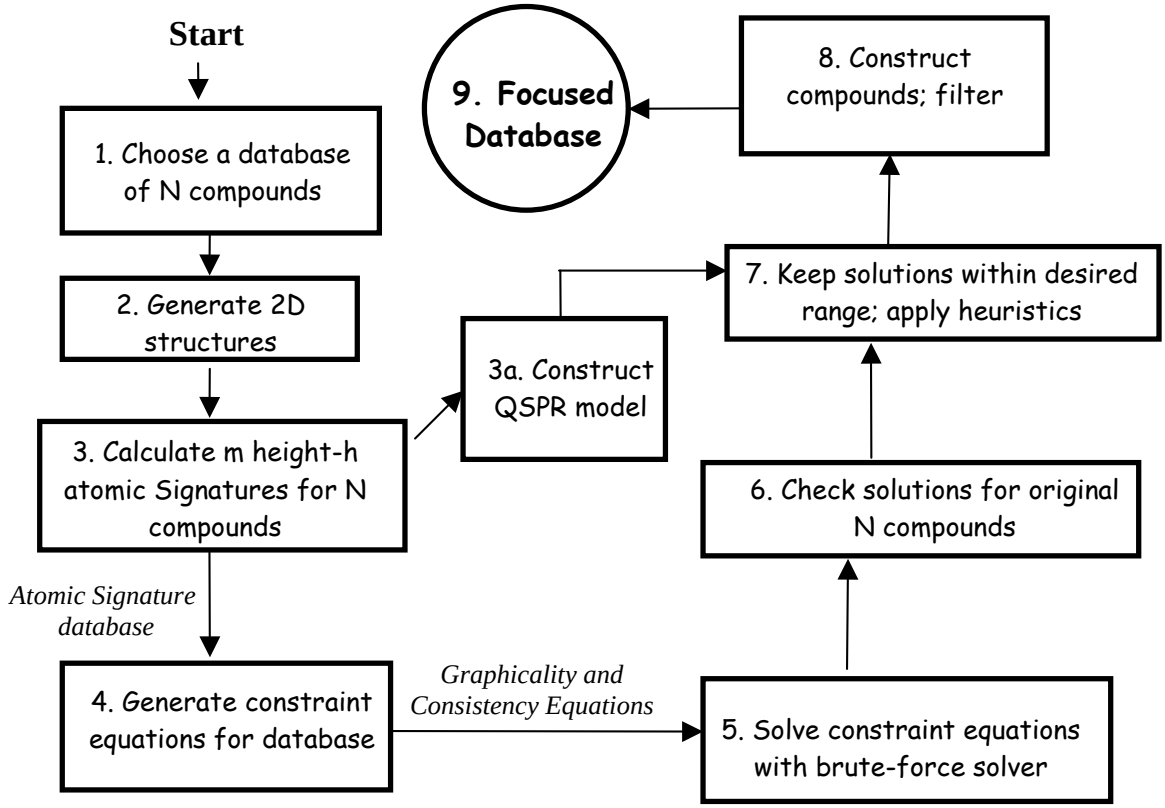


Figure 2: The computer-aided molecular design algorithm using the Signature molecular descriptor.

Step 1: <u>Choose a database of N compounds</u>

For any CAMD problem solved with Signature, the first step is to identify a dataset of interest. While the algorithm has proven successful for a small set (~ 15 compounds) [12], a larger set normally provides a more reliable QSPR to be used when scoring the compounds from the algorithm. However, if a set is too large (> 100), the diversity of atomic Signatures might cause the problem to become intractable with both time and storage constraints becoming active. When that is the case, a reasonable strategy is to focus the original training set on those compounds with the most desired traits (re: desired property range).

Step 2: <u>Generate 2-D structures</u>

The code to calculate the atomic Signatures (step 3) of a molecule requires the proper input format into the code. The desired input format is that of an MDL MOL file. Such a file format can be created on a variety of platforms from an initial drawing of the

molecule. Alternatively, various sites, such as the NIST Webbook offer mol file format downloads for various substances. Note that in all cases the stereochemistry of a molecule is lost in the 2-D representation.

Step 3: Calculate m height-h atomic Signatures for N compounds

Once a Signature height is specified, the molecular Signature of each of the N compounds identified in Step 1 is calculated using an in-house translator program. Most CAMD problems using Signature have been solved at either height-1 or height-2, which balances the limit on the number of unique atomic Signatures (overcoming time/storage issues) with specificity issues (if the height used is too large). The unique atomic Signatures for the set are grouped and identified as the 'atomic Signature database'. It is at this step where the algorithm bifurcates. The atomic Signature database is used as the independent variables in developing the QSPR (Step 3a) as well as in formulating the inverse design problem (Step 4).

Step 3a: Construct QSPR model

The solutions that are obtained from the inverse design algorithm need to be scored for fitness. This is accomplished with a QSPR model. The QSPR takes experimental data for a property of interest and regresses it against the occurrences of the atomic Signatures in the atomic Signature dabatase. Here, the atomic Signatures act as independent variables while the experimental data are the dependent variables. Note that models can be linear or non-linear, depending on application.

Step 4: Generate constraint equations for database

The atomic Signatures in the atomic Signature database created in Step 3 are used to generate two types of constraint equations. The first type, of which there is only one, is called a "graphicality" equation. This equation is a necessary condition to create a connected graph from any set of atomic Signatures.

$$\mathrm{Mod}\left(\sum_{i=2}^{z}(i-2)n_i - n_1 + 2\,,\,0\right) = 0 \qquad (1)$$

where z is the maximum number of vertices of atoms in the dataset while n_i is the degree of the root of atomic Signature i. [12].

The second type of equations are known as the consistency equations. These equations are hand-shaking requirements that are written for each bond type in the system. Basically, the atomic Signatures (rooted at C) which have, say, C single bonded to H must be matched with the atomic Signatures (rooted at H) which have H single bonded to C.

The set of consistency and graphical equations together form the system constraint equations. This system is underspecified in that it will have more variables (atomic Signatures) than equations. Since all of the coefficients in the equations as well as the solutions are integers, these are Diophantine equations.

Step 5: Solve constraint equations with a brute-force solver

The solution to the underspecified system of equations generated in Step 4 would yield an infinite solution space. Accordingly, we limit the range that the independent variables can take based on their range in the original training set. Thus, this minimum and maximum value (per atomic Signature) provides the additional constraints necessary to solve the system. Previously this system was solved using a Diophantine equation solver [13], but we have developed a smart brute-force technique which is both efficient and allows an estimate for time to completion.

Basically, this smart brute-force technique looks to satisfy the constraint equations in a step-wise manner such that the iterations involving those variables which occur in the equations go from least to most iterations. Note that if our technique was truly brute-force, we would iterate over the min/max values for each of the atomic Signatures in our problem in a nested loop. This smart technique, however, can cut the iterations required by more than half. [14]

Step 6: Check solutions for original N compounds

Since the N compounds form the constraint equations, those N compounds must be solutions to those constraints equations. This is a "dummy" check to verify the output of step 5 and is important for debugging purposes.

Step 7: Keep solutions within desired range; apply heuristics

The solutions which emerge from Step 5 (which could number in the billions) must be scored for fitness relative to a desired property value (or range of values). Accordingly, the solutions are filtered through the QSPR output from Step 3a and those which have the desired fitness are kept. It is also at this stage where various heuristics can be applied to focus the solution space based on expert knowledge or other means. For example, a regular approach is to determine the number of rings in a molecule and remove from the solution space those molecules which have more or less rings than the original training set. Additionally, molecular weight is often a factor in evaluating compounds for fitness and this can limit the solution space. Finally, for drug-like molecules, a rule-of-thumb such as Lipinksi's Rule of Fives can be used to remove all solutions which do not adhere to this heuristic. [15]

Step 8: Construct compounds; filter

The molecular Signatures which emerge from Step 7 have been scored successful using the QSPR and have passed various other heuristics. These are the molecular Signatures from which structure generation will occur. There is a degeneracy associated with going from a molecular Signature to a 2-D structure. At height-0 (just the molecular formula), the degeneracy is large, but monotonically decreases with Signature height until there is a unique 2-D structure associated with a particular molecular Signature (normally by height-3 or height-4). [16, 17]

Structure generation is performed using an algorithm developed by Faulon and co-workers [17], which is based on an earlier isomer enumeration algorithm developed Faulon. [2, 3] The algorithm is iterative, which requires starting with a molecular Signature of all atoms and no bonds, and then attempts to add bonds in all possible ways to match the target molecular Signature.

Once the structures are generated, various filters can be employed to remove those candidate structures which have energetic issues and are not feasible. For example, we remove those structures which have multiple bridges and aromatic rings that do not follow Huckel's rule. Also, we can perform an energy minimization using a force field to remove those high-energy structures. Finally, synthetic accessibility can be assessed here as well.

Step 9: <u>Focused Database</u>

The structures which have survived until this point become part of the focused database. These are the high-quality structures which are worthy of further investigation. It is here where experiments run on a select number of compounds to verify the predictions of the algorithm would be employed. Often, the results of the experimentation can be used to refine the QSPRs and the focused database itself.

Though outside the scope of this report, it is noted that other CAMD algorithms exist. The first CAMD algorithm is a group-based generate and test approach developed by Gani and co-workers during the early 1980's [1, 18], though it has been modified since that time and is still popular to this day [19] Here, predefined molecular fragments are identified, merged together and ultimately evaluated through a group-contribution approach. A different CAMD approach treats molecular design as an optimization problem solved using a mixed-integer non-linear approach, popularized by the work of Maranas. [20, 21] The key feature of CAMD using the Signature molecular descriptor [22] is that, unlike the other methods, it does not require templating to arrive at molecular structures. Templating is when certain parts of the compounds being designed are specified *a priori* to reduce search complexity. While templating increases the likelihood of finding solutions to the problem, its main drawback is that non-intuitive candidates are likely to be removed from consideration.

4 CAMD Applications

4.1 Solvent Design

In this section we apply the inverse-design methodology using the Signature molecular descriptor to the design of solvents. [14]

4.1.1 Motivation

Researchers at GlaxoSmithKline (GSK) have developed a solvent selection guide [23, 24] which provided information on 47 commonly used solvents. The guide provides the individual the opportunity to aid in solvent selection during the early stages of process development by considering aspects different than cost. Those included: incineration, recycle, biotreatment, volatile organic carbon, environmental impact in water, environmental impact in air, health hazard, exposure potential, and safety hazard. [23] Important properties were evaluated for each of those categories and, overall, a solvent was given a score of 1 to 10 for each category, with the higher the score, the more favorably it should viewed for that category. A simplification was made to group the nine categories into four: environmental waste (incineration, recycle, biotreatment, volatile organic carbon), environmental impact (environmental impact in water, environmental impact in air), health (health hazard and exposure potential) and safety hazard. Later, a fifth area was added [24] called life cycle assessment (LCA) that incorporates impacts of manufacturing, recycling, and disposal during the duration of a solvent. The scores for the 47 solvents are provided in Table 1 with a numerical component (1 to 10) and color code (green, yellow, or red) as reported by Jiménez-González et al. [24] Those solvents with a green rating had scores ranging from 8 to 10, a yellow rating was given for scores from 4 to 7, and a red rating was reserved for solvents with scores of 3 or less.

In this application of the CAMD algorithm with the Signature molecular descriptor we identify potentially new environmentally friendly solvents as a supplement to GSK's solvent selection guide. We generate QSPRs using Signature to rank the inverse solutions against the known compounds for environmental waste, environmental impact, health, safety, and LCA. We then solve the inverse design problem, score them with the QSPRs, apply various filters and, ultimately, generate the structures. In short, we apply the algorithm given in Figure 2 to this problem.

Table 1: Data from GSK's solvent selection guide. [24] **Reprinted from "Computer-aided molecular design using the Signature molecular descriptor: Application to solvent selection. Weis and Visco. in press, with permission from Elsevier"**

SSG Class	Solvent	CAS #	Env. Waste	Env. Impact	Health	Safety	LCA Ranking
Alcohols	Ethylene Glycol	107-21-1	4	9	8	9	9
	1-Butanol	71-36-3	5	8	8	8	5
	Diethylene Glycol Butyl Ether	112-34-5	5	7	10	9	7
	Isoamyl Alcohol	123-51-3	7	7	7	8	6
	2-Ethylhexanol	104-76-7	9	6	8	7	6
	2-Butanol	78-92-2	4	7	7	7	6
	1-Propanol	71-23-8	3	7	5	8	7
	Ethanol	64-17-5	3	8	10	7	9
	2-Propanol	67-63-0	3	9	9	7	5
	t-Butanol	75-65-0	3	10	7	7	8
	Methanol	67-56-1	3	10	5	8	9
Esters	t-Butyl Acetate	540-88-5	7	10	7	7	7
	Butyl Acetate	123-86-4	7	8	9	8	5
	n-Propyl Acetate	109-60-4	6	7	8	7	5
	Isopropyl Acetate	108-21-4	5	8	8	7	6
	Ethyl Acetate	141-78-6	4	8	8	4	6
	Methyl Acetate	79-20-9	2	10	7	5	7
	Dimethyl Carbonate	616-38-6	3	7	8	7	8
Aromatics	p-Xylene	106-42-3	8	2	7	5	7
	Toluene	108-88-3	7	3	6	4	7
	Fluorobenzene	462-06-6	4	2	4	5	1
Ketones	Methyl Isobutyl Ketone	108-10-1	7	6	6	7	2
	Acetone	67-64-1	2	9	8	5	3
	Methyl Ethyl Ketone	78-93-3	3	6	8	5	3
Polar Aprotics	N-Methyl-2-Pyrrolidone	872-50-4	4	6	8	9	3
	Dimethyl Acetamide	127-19-5	4	7	2	10	3
	Dimethyl Formamide	68-12-2	4	6	2	8	6
	Dimethylpropylene Urea	7226-23-5	4	7	5	9	4
	Dimethylsulphoxide	67-68-5	4	4	8	3	6
	Formamide	75-12-7	3	7	2	10	8
	Acetonitrile	75-05-8	2	6	6	8	4
Acids	Propionic Acid	79-09-4	5	8	4	9	7
	Acetic Acid	64-19-7	3	8	4	8	8
Alkanes	Cyclohexane	110-82-7	5	6	8	2	7
	Methyl Cyclohexane	108-87-2	7	5	8	2	7
	Heptane	142-82-5	6	3	9	1	7
	2-Methylpentane	107-83-5	5	3	5	1	7
	Hexane	110-54-3	5	2	4	1	7
Chlorinated	Dichloromethane	75-09-2	2	5	3	10	7
Ethers	Methyl Tert-Butyl Ether	1634-04-4	4	4	6	2	8
	1-2-Dimethoxyethane	110-71-4	3	5	3	2	7
	Tetrahydrofuran	109-99-9	2	6	7	2	5
	Bis(2-methoxyethyl) Ether	111-96-6	6	5	1	3	6
	Diisopropyl Ether	108-20-3	5	2	9	1	9
Basics	Triethylamine	121-44-8	4	5	2	4	7
	Pyridine	110-86-1	3	4	3	6	2

4.1.2 Results

Referencing Step 1 from Figure 2, we have identified 46 compounds from Table 1 (note: a mixture of petroleum ether was one of the compounds in the GSK solvent selection guide, but this was removed because working with mixtures using Signature is a future challenge). For Step 2, the 2D structures have been drawn in MOL file format and translated (Step 3) from a MOL file format to Signature. This resulted in the 36 unique height 1 atomic Signatures which form the atomic Signature database, shown in Table 2.

Table 2: Unique height 1 atomic Signatures collected from GSK's solvent selection guide. **Reprinted from "Computer-aided molecular design using the Signature molecular descriptor: Application to solvent selection. Weis and Visco. in press, with permission from Elsevier"**

x1	[C]([C][C]=[O])		x19	[C]([O][O]=[O])
x2	[C]([C][C][C][H])		x20	[C](p[C][H]p[N])
x3	[C]([C][C][C][O])		x21	[C](p[C]p[C][F])
x4	[C]([C][C][H][H])		x22	[C](p[C]p[C][H])
x5	[C]([C][C][H][O])		x23	[Cl]([C])
x6	[C]([C][H][H][H])		x24	[F]([C])
x7	[C]([C][H][H][N])		x25	[H]([C])
x8	[C]([C][H][H][O])		x26	[H]([N])
x9	[C]([C][N]=[O])		x27	[H]([O])
x10	[C]([C][O]=[O])		x28	[N]([C][C][C])
x11	[C]([C]p[C]p[C])		x29	[N]([C][H][H])
x12	[C]([C]t[N])		x30	[N](p[C]p[C])
x13	[C]([Cl][Cl][H][H])		x31	[N](t[C])
x14	[C]([H][H][H][N])		x32	[O](=[C])
x15	[C]([H][H][H][O])		x33	[O](=[S])
x16	[C]([H][H][H][S])		x34	[O]([C][C])
x17	[C]([H][N]=[O])		x35	[O]([C][H])
x18	[C]([N][N]=[O])		x36	[S]([C][C]=[O])

Step 3a requires the creation of the QSPRs to ultimately score the inverse design solutions. Preliminary calculations indicated that a simple multiple linear regression would not be adequate to produce useful QSPRs for all of the five areas desired so a non-linear approach was used. Accordingly, the independent variables available for use in the QSPRs included both the height-1 atomic Signatures (36 of them) and the products with themselves (another 36). Pair-pair correlation coefficients were found for each of the 72 pairs and those perfectly correlated were removed from consideration during regression. Accordingly, only 33 of the 72 independent variables remained.

Model selection was based on evaluating both the R^2 and q^2 as a function of the number of independent variables included at each step during the regression. Note that q^2 is a leave-one-out cross-validation metric used to evaluate a model for overfitting. [25]

Figure 3 provides the QSPR for environmental impact. A balance is made between keeping the most number of independent variables in the model with creating the most predictive model. Accordingly, we choose a model with nine independent variables as this is before the q^2 value starts to decrease rapidly. We used this strategy for the other QSPRs and the statistics for these QSPRs, including the regression coefficients, are provided in Table 3 and 4.

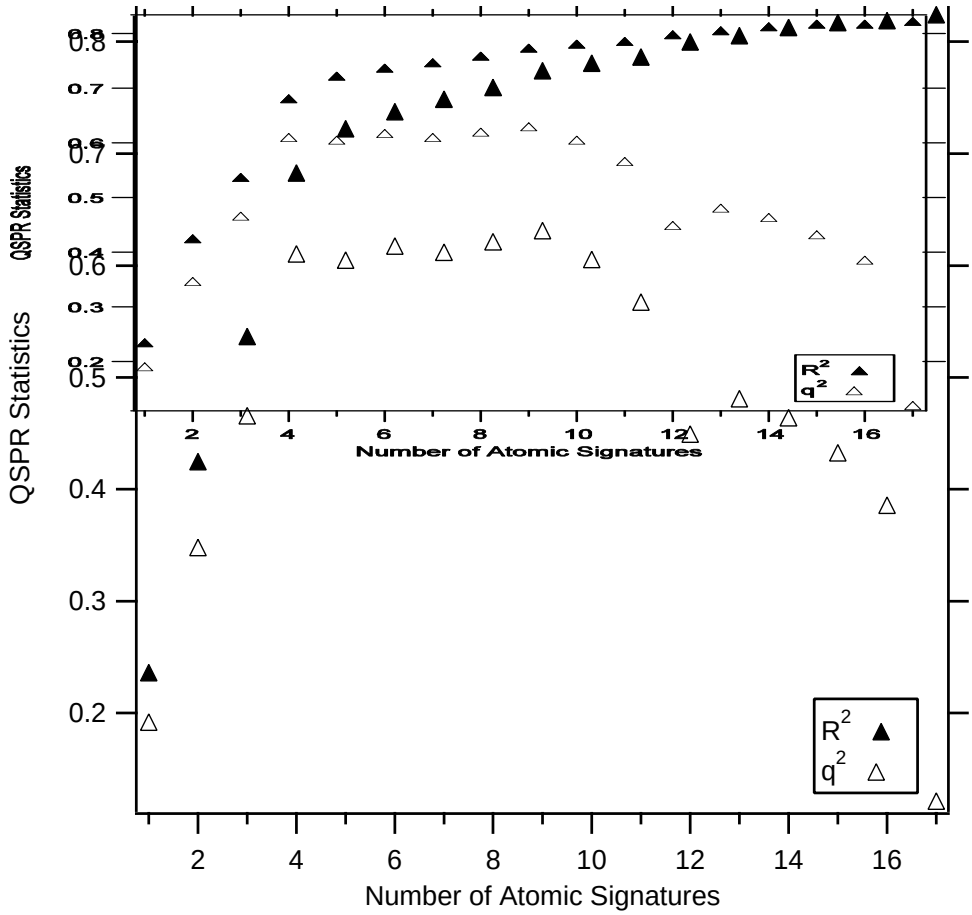


Figure 3: The environmental impact QSPR statistics are plotted as a function of the number of independent variables. **Reprinted from "Computer-aided molecular design using the Signature molecular descriptor: Application to solvent selection. Weis and Visco. in press, with permission from Elsevier"**

Table 3: Statistics for the height 1 Signature QSPRs. **Reprinted from "Computer-aided molecular design using the Signature molecular descriptor: Application to solvent selection. Weis and Visco. in press, with permission from Elsevier"**

QSPR	Number of Descriptors	R^2	q^2
Env. Waste	7	0.80	0.71
Env. Impact	9	0.77	0.63
Health	13	0.69	0.42
Safety	7	0.71	0.57
LCA	11	0.71	0.23

Table 4: Coefficients for the non-linear height 1 QSPRs. **Reprinted from "Computer-aided molecular design using the Signature molecular descriptor: Application to solvent selection. Weis and Visco. in press, with permission from Elsevier"**

	QSPR Coefficients				
H1 Signature	Env. Waste	Env. Impact	Health	Safety	LCA
[C]([C][C]=[O])					-3.5339
[C]([C][C][C][O])					0.8738
[C]([C][C][H][H])			0.5101		
[C]([C][C][H][O])					
[C]([C][H][H][H])					
[C]([C][H][H][N])			4.5036		0.8395
[C]([C][H][H][O])		0.4889	-1.0038		-0.2617
[C]([C][N]=[O])					-1.7152
[C]([C]p[C]p[C])	1.6140		1.2525		2.8553
[C]([H][H][H][N])					
[C]([H][H][H][O])					0.7632
[C]([H][N]=[O])					1.7519
[C](p[C]p[C][H])		-0.5028			-0.9137
[H]([C])	0.2770			-0.8248	
[N]([C][C][C])				1.4887	-1.9051
[O]([C][C])			6.3387		
[O]([C][H])		2.0972	2.6083	2.8217	1.0136
[C]([C][C]=[O])^2			2.4142		
[C]([C][C][C][H])^2	2.0404				
[C]([C][C][C][O])^2		2.3892			
[C]([C][C][H][H])^2		0.0509		-0.0634	
[C]([C][C][H][O])^2					
[C]([C][H][H][H])^2					
[C]([C][H][H][N])^2		0.2575	-1.7928		
[C]([C][H][H][O])^2					
[C]([C][N]=[O])^2				2.3572	
[C]([C][O]=[O])^2	1.4681	0.8254	-1.7420		
[C]([C]p[C]p[C])^2					
[C]([H][H][H][N])^2			-0.4470		
[C]([H][H][H][O])^2			-1.3540		
[C]([H][N]=[O])^2	1.3884		-1.6851		
[C](p[C]p[C][H])^2	0.0730				
[H]([C])^2		-0.0127		0.0359	
[N]([C][C][C])^2					
[O](=[C])^2		1.9401		2.2262	
[O]([C][C])^2			-1.4367		
[O]([C][H])^2	0.4729				
Constant	1.0038	5.5531	4.5791	8.0693	6.2006

In Step 4 of the CAMD algorithm with Signature (Figure 2), 15 constraint equations were generated from the 36 height 1 atomic Signatures in Table 2. In Step 5, these equations were solved using the brute-force solver (including the min/max values for the occurrence numbers of the original 46 compounds) resulting in a total of 4,031,916 solutions. This took 62.572 s of CPU time on a Pentium 4 Xeon, 2.8 GHz processor. Step 6 verified that the original 46 compounds were found among the 4 million solutions.

Step 7 of the algorithm scores the solutions using the QSPRs developed in Step 3a and applies various heuristics. Before scoring, we reduced the solution space by 98 % by only including those solutions which resulted in a molecular weight smaller than the largest molecular weight in the original set of 46, which was 162. This left nearly 100,000 solutions. An additional cycle filter was used where only solutions having the same range of cycles as in the original set of 46 (zero and 1 cycle) were kept. Of the nearly 100,000 solutions, 64,955 solutions satisfied this criterion. Finally, these 65,000 solutions were scored using the five QSPRs developed in Step 3a. Those solutions which did not any have "red" value predictions (greater than or equal to 4) were kept and there were 40,816 solutions which passed this stage.

Step 8 employs the structure generation step with additional filters. From the 40,816 solutions, 69,033 structures were obtained. Next, the Marvin Beans [26] software package was used to perform 3D coordinate calculation by optimization with the Dreiding force field. Those with energies less than 50 kcal/mol (which was the maximum value in the original set of 46 compounds) were kept resulting in 40,660 structures. Note that all of the steps in the algorithm are provided pictorially in Figure 4.

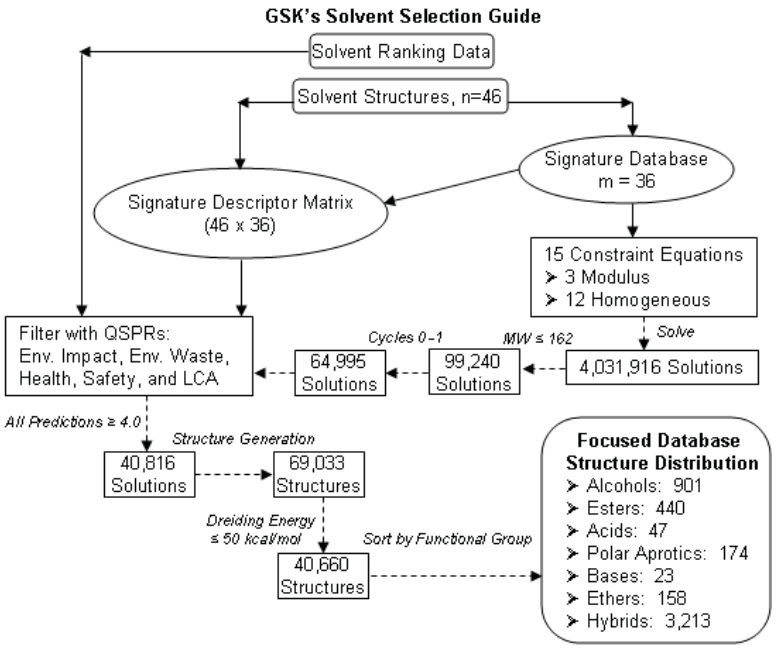
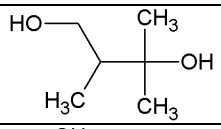
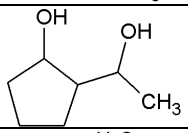
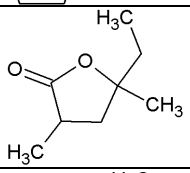
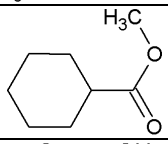
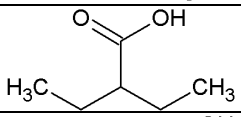
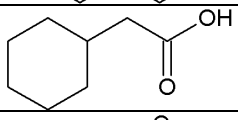
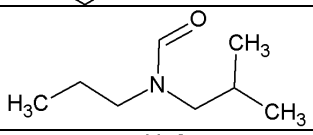
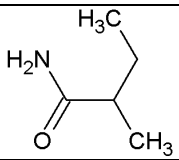
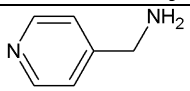
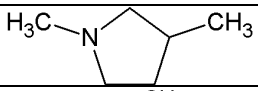
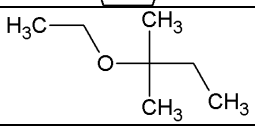
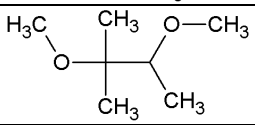


Figure 4: A flow chart for the overall inverse design process using the Signature molecular descriptor. **Reprinted from "Computer-aided molecular design using the Signature molecular descriptor: Application to solvent selection. Weis and Visco. in press, with permission from Elsevier"**

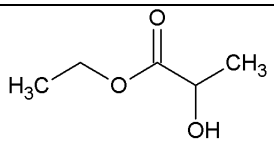
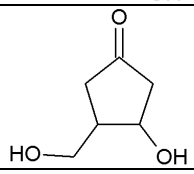
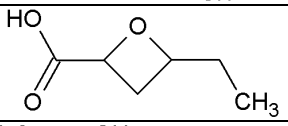
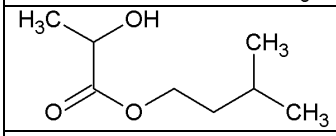
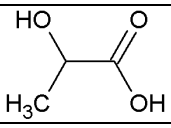
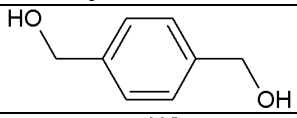
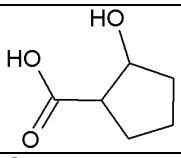
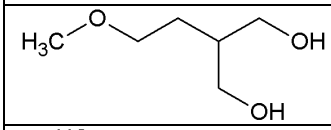
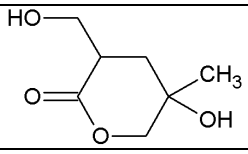
To best display the results of Step 9 (the focused database), compounds were resorted (if possible) into the original classes identified in the GSK solvent selection guide. Some samples from the different classes in the focused database are provided in Table 5, including the QSPR predictions and corresponding CAS numbers where available. A distribution of the structures obtained is given in the lower right-hand corner of Figure 4, however there were no alkanes or ketones obtained. It was only possible to place 1,743 of the 40,660 structures into the classes defined in GSK's solvent selection guide with the left over solutions termed a "hybrid" class.

Table 5: Focused database samples from classes defined in GSK's solvent selection guide. **Reprinted from "Computer-aided molecular design using the Signature molecular descriptor: Application to solvent selection. Weis and Visco. in press, with permission from Elsevier"**

		QSPR Predictions					
Class	Structure	Env. Waste	Env. Impact	Health	Safety	LCA	CAS #
Alcohol		8.26	10.80	8.79	8.99	8.84	24893-35-4
		8.26	8.38	11.33	8.41	8.23	860379-09-5
Ester		8.39	8.42	8.76	5.53	7.07	1636-45-9
		8.39	7.10	8.94	4.20	6.96	4630-82-4
Acid		8.03	9.08	6.47	8.14	7.21	88-09-5
		8.59	10.10	8.51	6.18	7.21	5292-21-7
Polar Aprotic		9.14	4.90	5.24	8.08	7.73	NA
		5.54	6.51	5.09	8.07	4.49	541-46-8
Base		4.57	4.35	8.54	4.41	8.07	3731-53-1
		6.64	4.49	6.48	4.84	5.97	45470-22-2
Ether		5.44	5.23	8.99	4.00	6.81	919-94-8
		5.44	4.69	6.09	4.06	8.60	74421-00-4

Since the hybrids represented structures outside of any of the classes, we kept those with at least three green QSPR predictions and this reduced the 38,895 hybrid structures down to 3,213, which is a more manageable number. Some hybrid structure examples are provided in Table 6. Note that the first compound in Table 6, identified as a hybrid solvent, is ethyl acetate, a known solvent. [27] Also, isoamyl lactate, the fourth compound given in Table 6, was patented by the Archer-Daniel-Midland Company [28] as a green solvent. Such findings give added confidence in the other design results presented for this study.

Table 6: Samples of hybrid structures from the focused database. **Reprinted from "Computer-aided molecular design using the Signature molecular descriptor: Application to solvent selection. Weis and Visco. in press, with permission from Elsevier"**

		QSPR Predictions					
Class	Structure	Env. Waste	Env. Impact	Health	Safety	LCA	CAS #
Hybrid		5.44	9.88	9.34	8.60	6.95	97-64-3
		7.15	11.57	12.23	11.38	4.43	NA
		5.44	9.59	11.37	8.35	7.21	NA
		9.14	8.10	9.85	8.76	6.95	19329-89-6
		5.47	12.31	8.05	13.21	8.23	50-21-5
		9.51	7.90	10.29	9.41	9.76	589-29-7
		8.62	12.16	9.58	11.07	8.23	81887-89-0
		8.26	9.43	10.84	8.92	8.21	160319-70-0
		9.17	14.66	11.46	11.22	8.58	NA

4.2 Design of Glucocorticoid Receptor Ligands

In this section we apply the inverse-design methodology using the Signature molecular descriptor in order to design glucocorticoid receptor ligands. [5]

4.2.1 Motivation

We seek here to design corticosteroids as treatment for asthma and other diseases associated with pulmonary inflammation. To provide potential therapeutic options owing to the many side effects of other treatments, [29] we aim to identify a potential set of novel glucocorticoid receptor ligands that possess high receptor binding affinity, high systemic clearance, high plasma protein binding, and low oral bioavailability. A drug having these properties would indicate an ability to remain in the lungs and, if exposed to the main circulation, be quickly removed to limit side effects.

Traditional drug design approaches in this arena have been performed in the traditional manner where substitutent groups are modified off a template structure. These have yielded novel glucocorticoids such as loteprednol etabonate (developed under a retro-metabolic approach) [30] and fluticasone propionate (developed with structure-activity analysis) [31]. Both glucocorticoids possess an enhanced therapeutic index compared to previous steroids.

Here we use CAMD with the Signature molecular descriptor to generate and evaluate many potential structures and identify a focused group of high priority candidates. This focused database will be scored such that each candidate is predicted to possess pharmacological properties that indicate the likeliness of the drug to remain in the lungs and be quickly removed or inactivated in the main circulation. Note that some desired pharmacokinetic and pharmacodynamic properties have already been identified to guide future GR ligand development. [32-36] These properties are important in identifying pulmonary selectivity, such as binding affinity, oral bioavailability, systemic clearance and plasma protein binding.

Oral bioavailability is a measure of the percentage of a drug that is available to the target via an oral route. [37] This is significant since a portion of inhaled treatments become deposited in the mouth and could be active if ingested. As some corticosteroids such as fluticasone propionate already possess virtually negligible oral bioavailabilities [32] future treatments should match this achievement.

Systemic clearance is a measure of how quickly the drug is transported to the liver for deactivation. Currently, the corticosteroid with the highest systemic clearance is beclomethasone dipropionate at 150 L/h [34] and this provides a basis during drug design.

Plasma protein binding (or fraction unbound) is a measure of the relative amount of bound corticosteroid in the blood. [36] When a drug is bound to the protein, it is then unavailable to affect other pathways during its transport to the liver. As many steroids,

including corticosteroids, already bind well to plasma proteins (some over 98 %) [37] it is important to ensure this characteristic is conserved in any novel structures.

In the next section, we describe the use of CAMD with the Signature molecular descriptor in the generation of glucocorticoid receptor ligands with optimally predicted properties. As in the previous example, we follow Figure 2.

4.2.2 Results

To begin Step 1 of the algorithm, a literature search was conducted to collect a sizable and diverse set of experimentally studied corticosteroids. We identified 65 corticosteroids for which relative receptor binding affinity was available, [33-39] though data for system clearance, plasma protein binding and oral bioavailability were not found for all 65 compounds. Note that the literature reported slightly different experimental values for several compounds so averaged values were used. These 65 compounds and the experimental values for the four properties of interest are provided in Table 7.

Table 7: Training Set Compounds and Activity Data. **Reprinted from "Potential Glucocorticoid Receptor Ligands with Pulmonary Selectivity using I-QSAR with the Signature Molecular Descriptor, Jackson, Weis and Visco. 72: 540 – 550, 2008, with permission from John Wiley and Sons"**

Corticosteroid	Relative Receptor Binding Affinity	Oral Bioavailability (%)	Systemic Clearance (L/hr)	Fraction Unbound (%)	Reference
beclomethasone dipropionate	53	22	150	13	33-37
beclomethasone monopropionate	1345	26	120	36	33-36
budesonide	935	11	84	12	33-37, 39
dexamethasone	100	78	17	32	37-39
flunisolide	185	20	58	20	33, 34, 36, 37, 39
fluocortolone	65	84	32	13	39
fluticasone propionate	1800	1	69	10	33-37, 39
loteprednol etabonate	150	1	63	10	33, 36-38
methylprednisolone	42	9	21	23	39
mometasone furoate	2500	1	54	1	33, 36, 37
prednisolone	16	82	6	25	39
triamcinolone acetonide	234	23	37	29	33, 34, 36, 37, 39
beclomethasone	100				36, 38
etiprednol dicloacetate	200				33, 38
LE5601	150				38
LE5602	110				38
LE5603	70				38
LE5606	3				38
LE5608	1				38
LE5610	200				38
LE5614	1				38

LE5618	16				38
LE5621	1				38
LE5623	10				38
LE5638	500				38
LE5639	780				38
LE5643	80				38
LE5644	210				38
LE5648	1165				38
LE5649	3				38
LE5651	1				38
LE5654	10				38
LE5657	11				38
LE5658	840				38
LE5660	16				38
LE5671	820				38
LE5673	2100				38
LE5679	1				38
LE5683	19				38
LE5685	1				38
LE5687	7				38
LE5689	1100				38
LE5690	1000				38
LE5693	990				38
LE5698	1000				38
LE5699	820				38
LE5704	1200				38
LE5707	990				38
LE5711	200				38
LE5712	70				38
LE5715	3				38
LE5718	1				38
LE5720	10				38
LE5721	25				38
LE5725	10				38
LE5726	315				38
LEGH01	3				38
LEGH02	29				38
LEGH03	132				38
LEGH04	124				38
LEGH05	54				38
LEGH06	6				38
LEGH07	10				38
LEGH08	9				38
LEGH09	2				38

Next, the structures were drawn (Step 2) and the molecular Signatures of each of the 65 compounds were obtained (Step 3). Note that we selected to solve this system using height-2 Signatures in order to limit the solution space owing to the diversity of the original set of 65 compounds. A total of 161 unique height-2 atomic Signatures were obtained for the 65 compounds.

Step 3a requires the regression of the experimental data to determine the four QSPRs for use as scoring functions in the CAMD algorithm. This was performed using a multiple linear regression technique and the results of the QSPRs are provided in Table 8 along with the coefficients for the QSPRs.

Table 8: Regression coefficients and statistics for the QSPRs. **Reprinted from "Potential Glucocorticoid Receptor Ligands with Pulmonary Selectivity using I-QSAR with the Signature Molecular Descriptor, Jackson, Weis and Visco. 72: 540 – 550, 2008, with permission from John Wiley and Sons"**

	Relative Receptor Binding Affinity QSPR	**Oral Bioavailability QSPR**	**Plasma Protein Binding QSPR**	**Systemic Clearance QSPR**
r^2	0.806	0.959	0.958	0.983
Regression constant	2.2648	0.0000	0.8891	0.4474
Atomic Signature (x_n)	**QSPR Coefficients**			
[C](=[C]([C][H])[C]([C][C][H])[C]([C][C][C]))		-0.8320		
[C]([C](=[C][H])[C]([C]=[C])[C]([H][H][H])[C]([C][C][Cl]))				0.5545
[C]([C](=[C][H])[C]([C]=[C])[C]([H][H][H])[C]([C][C][H]))			-0.1551	
[C]([C]([C]=[O])[C]([C][H][O])[C]([C][C][C])[O]([C]))		-0.5516		
[C]([C]([C]=[O])[Cl][H][H])			-1.1515	
[C]([C]([C]=[O])[H][H][O]([C]))		1.3424		
[C]([C]([C]=[O])[H][H][O]([H]))		1.7863		
[C]([C]([C][H][H])=[O][O]([C]))				-0.2896
[C]([C]([C][H][H])[C]([C][C][H])[C]([C][C][H])[H])	-0.5482			
[C]([C]([C][H][H])[C]([C][C][H])[H][H])	-0.4162			
[C]([C]([C][H][H])[H][H][H])	0.6539			
[C]([C]([C][H][H])[H][O]([C])[O]([C]))				0.6662
[C]([C]([Cl][Cl][H])=[O][O]([C]))	2.5648			
[C]([C]([Cl][H][O])[H][H][H])	-1.2650			
[C]([C]([O]=[O])[C]([C][C][H])[C]([C][C][C])[O]([H]))	-1.4060			
[H]([C]([C][C][C]))				0.1236
[H]([C]([C][C][O]))	0.4927			
[H]([C]([C][H][H]))				0.1543
[H]([O]([C]))			0.2192	-0.0302
[O](=[C]([C][C]))				-0.4442
[O]([C]([C]=[O])[C]([C][H][H]))	-1.6406			
[O]([C]([C]=[O])[C]([H][H][H]))	-1.5134			
[O]([C]([C]=[O])[C]([H][H][O]))	-1.8695			
[O]([C]([O]=[O])[C]([O]=[O]))	-1.3768			

Once the atomic Signature database has been obtained, Step 4 is to generate the constraint equations associated with these atomic Signatures. Using in-house PERL scripts, a total of 104 equations were generated from the 161 unique height-2 atomic Signatures. Step 5, the solving of these 104 underspecified equations, was accomplished using the brute-force solver with the min/max constraints on the occurrence values for each of the 161 atomic Signatures in the database.

While solving such a large set of equations, storage of the solutions became an issue. To this end, a cycle filter was written into the brute-force solver for this application. Despite the fact that one (of the 65) compounds had six cycles, we kept only those solutions with had five cycles or less. Upon solution, 308,930,136 molecular Signatures were found which satisfied the cycle constraint as well as the 104 constraint equations. Note that, as required from Step 6, sixty-four of the molecular Signatures found were from the original set of 65 (the one that was not found was the one with the seven cycles).

Step 7 involves scoring the 300+ million solutions using the four QSPRs. The first scoring function used was for the relative receptor binding affinity (RRBA). This was used first since this QSPR contained the most dependent variables (65) in the creation of the model. Do note that the model was generated on logarithmic values of the dependent variable as it spans several orders of magnitude). As the original set of 65 ranged (on a log scale) from 0 to 3.4 in this property, we selected to keep predicted solutions (from the 300+ million) that ranged from 3 – 5 for log (RRBA). The lower limit, while arbitrary, accommodates some error in the predicted values. For the upper limit, it was deemed that any values greater than 5 were too great of an extrapolation from where the regression parameters of the model were trained. Once this QSPR was used to identify those solutions which ranged between 3 – 5 for log (RRBA), 105,637,556 solutions were kept.

The QSPR for oral bioavailability was used next. This QSPR was trained on the first 12 compounds in Table 8 as these were the only ones where experimental data were available. As there exists several corticosteroids that provide oral bioavailabilities of less than 1%, we used this as the cut off value. Thus, of the 105+ million compounds screened, 61,556,852 solutions were left which have a predicted oral bioavailability of less than 1%.

As in the previous QSPR, the model developed for systemic clearance was based on the first 12 compounds in Table 8. Since the largest systemic clearance in the training set was 150 l/h, we kept all solutions which had a predicted systemic clearance greater than this value, but less than 224 l/h. Note that, owing to the range of data available, we used the logarithm of the systemic clearance as the dependent variable. Of the 61+ million compounds screened using this QSPR, 43,357,092 solutions had predicted systemic clearances between 150 l/h and 224 l/h.

The final QSPR used was for plasma protein binding. Once again, the first 12 compounds in Table 8 were used to create the model. Solutions were removed if the predicted fraction unbound was greater than 1% as this was the maximum value among

the 12 compounds. Of the 43+ million compounds evaluated, only 9,899,008 solutions were predicted to meet or exceed this requirement.

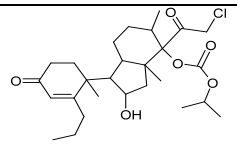
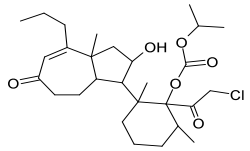
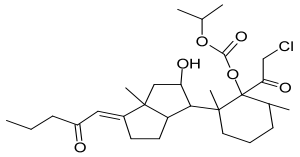
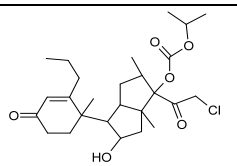
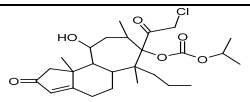
In addition to the QSPRs to score the solutions, we employed a heuristic known as Lipinski's Rule of Five [15] for inhaled pharmaceuticals. [40] This expert knowledge evaluates molecular weight, lipophilicity, hydrogen bond donors and acceptors relative to known inhaled pharmaceuticals. Table 9 provides the ranges for these parameters. Note that the upper and lower bounds for molecular weight and Log P in Table 9 were modified to include the upper and lower bounds of the compounds from the original 65. Of the nearly 10 million compounds evaluated with this rule, only 422 satisfied all four properties, and those were the solutions which moved to Step 8 (structure generation).

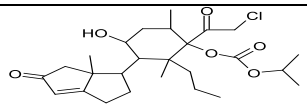
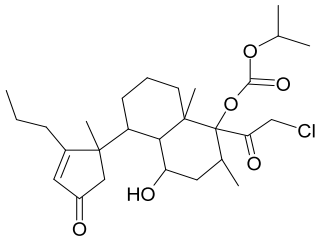
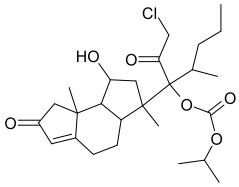
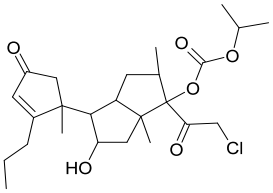
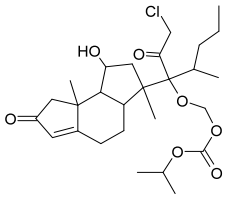
Table 9: Modified Rule of Five Parameters. **Reprinted from "Potential Glucocorticoid Receptor Ligands with Pulmonary Selectivity using I-QSAR with the Signature Molecular Descriptor, Jackson, Weis and Visco. 72: 540 – 550, 2008, with permission from John Wiley and Sons"**

Physicochemical Property	Lower Bound	Upper Bound
Log $P_{o/w}$	-1.0	4.44
Molecular Weight (Da)	346	540
Hydrogen Bond Donor	2	6
Hydrogen Bond Acceptor	4	11

All 422 molecular Signatures were subject to structure generation as well as the Dreiding energy filter. As the maximum intramolecular energy in the set of the original 65 was 176.3 kcal/mol, this value was used as the filter. From the structures created from the 422 molecular Signatures, only 84 had an intramolecular energy less than 176.3 kcal/mol. These 84 structures represent the focused database (Step 9) and a sampling of these are provided in Table 10. Note that all of the structures in Table 10 have the same predicted values for oral bioavailability (1 %), plasma protein binding (99.1 %) and systemic clearance (475 l/h) as the QSPRs have removed those which did not meet the previously described rubrics. However, the RRBA is not the same and that value is provided in Table 10 as well.

Table 10: Selected candidates in the focused database and their predicted RRBA.
Reprinted from "Potential Glucocorticoid Receptor Ligands with Pulmonary Selectivity using I-QSAR with the Signature Molecular Descriptor, Jackson, Weis and Visco. 72: 540 – 550, 2008, with permission from John Wiley and Sons"

Compound		Predicted RRBA
34-3		1179.45
35-11		1179.45
35-54		1179.45
38-1		8017.92
165-4		3075.18

165-6		3075.18
167-4		3075.18
285-4		1179.45
290-1		8017.92
303-4		1179.45

Note that the compounds in Table 10 (and the focused database, in general) are not steroidal in nature. As in previous studies, this is the benefit of the CAMD technique with Signature: the identification of novel/non-intuitive structures. Providing some support to the identification of non-intuitive structures is the fact that other research has been conducted confirming the results of non-steroidal corticosteroids in binding to the glucocorticoid receptor. For example, a series of C-10 substituted 5-allyl-2,5-dihydro-2,2,4-trimethyl-1H-(1)benzopyrano[3,4-f]quinolines [41, 42] as well as a set of arylpyrazole compounds with varying substitutions [43] has been reported. As an additional, quantitative measure of the diversity of the focused database of the 84 candidates relative to the original 65 compounds, we have calculated the Tanimoto coefficient between these two sets. The Tanimoto coefficient measures structural similarity with a value of "1" being perfectly similar while "0" means perfectly dissimilar. None of the 84 candidates had a maximum Tanimoto coefficient above 70% with any training set compound and, in fact, the average Tanimoto coefficient for each of the 84 compounds with the training set compounds was never above 50 %.

5 High-Throughput Screening using Signature

In this next work, we demonstrate the use of Signature to classify a bio-assay between active and non-active compounds as well as use the classifier to screen a large database for predictive activity. [44]

5.1 Introduction

High-throughput screening (HTS) is common technique in drug discovery where large libraries of compounds are screened against a particular target. While primarily a commercial venture, the Molecular Libraries Initiative (MLI) [45], part of the NIH Roadmap for Medical Research [46], looked to increase the use of small molecules in basic research. A network of academic research centers around the country belong to the Molecular Libraries Screening Center Network (MLSCN) [47] which submit the results of their HTS assays into a publicly available archive called PubChem, which is comprised of three databases: PCSubstance, PCCompound and PCBioAssay [48].

In this work, we look to use a supervised machine-learning technique called support-vector machine (SVM) to classifiy a bio-assay available in PubChem and then screen the rest of the PubChem compounds for activity. In our work the input vectors used in the SVM are atomic Signatures. Note that Signature has previously been used with a SVM to predict both protein-protein [49] and drug-target [50] interactions. Note that the number of descriptors (referred to as features when used in statistical learning methods) compared to the number of observations is an important consideration to avoid overfitting.

As a proof-of-concept, we have selected AID 846 [51] which is a confirmatory screen of compounds against Factor X1a, a protein involved in the blood coagulation pathway. It is a potentially useful therapeutic target because it has the potential for development of novel antithrombotic drugs to replace conventional ones like heparin and warfarin. [52] We classify these compounds as either active or inactive using our SVM using Signature and then screen approximately 12 million compounds from PubChem for predicted activity. We also develop metrics and perform docking studies to provide confidence in a "focused database" of compounds predicted to be active against Factor X1a, but have not yet been verified experimentally.

5.2 Methods

5.2.1 Support Vector Machines

Support vector machines are classifiers that use an optimal separating hyperplane. If one assumes data are presented as pairs $\{(x_i, y_i)\} \subset R^n \{\pm 1\}$, then this means that data belongs to only two classes with labels +1 or -1. For this work, +1 refers to active compounds against Factor X1a, and -1 represents inactive compounds. Using the pair notation, SVMs are given in the form

$$f(x) = \sum_i \alpha_i y_i k(x_i, x) + b \qquad (2)$$

where f: $R^n \rightarrow R$ is a decision function for classification. If *f(x)* is greater than some threshold *t*, then x belongs to the class +1, if not x belongs to the class -1. The constants α_i and b are acquired by solving the quadratic programming problem. The constant α_i is zero for all observations except the important borderline cases, which are known as the support vectors. A kernel function $k : R^n \times R \rightarrow R$ is a dot product in some vector space that can efficiently transform input data. Here, we use the Signature kernel provided in equation 3 which was first introduced by Martin and co-workers. [49]

$$k(A, B) = s(A) \cdot s(B) \qquad (3)$$

The vector space will be made of all unique atomic signatures from heights 0, 1 and 2. The SVMs in this work were generated by the SVM^{light} algorithm [53].

Ideally, the well-trained SVM would perfectly classify the data. The user defined cost parameter *C* controls the tradeoff between allowing for some misclassification and the margin of the optimal separating hyperplane. For this problem, we evaluated cost parameters that ranged several orders of magnitude using a search strategy following an approach used elsewhere [54] and set a value of the cost parameter equal to 1.0

5.2.2 Feature Selection

Though SVM is not as susceptible to overfitting compared to other machine learning methods, [55] selecting the subset of the most relevant atomic Signatures is necessary to increase predictive capability. This approach, called feature selection, is broadly divided into two categories: filter methods and wrapper methods. Filter methods rank individual features by a defined metric completely independent from SVM, while wrapper methods select features to add to the model by working with the SVM during training/testing steps in order to optimize some objective function.

Filter methods are more computationally efficient than wrapper methods, but these methods treat each feature separately. The goal of a filter method is to select the features that discriminate the most between two classes. The coefficient ω_i is an a filtering metric defined by Golub as

$$\omega_i = (\mu_i(+) - \mu_i(-))/(\sigma_i(+) + (\sigma_i(-)) \qquad (4)$$

where μ_i and σ_i are the mean and standard deviation for features in the (+) or (-) class, respectively. [56] When ω_i has a large, positive magnitude, it signifies a relationship to the (+) class, while large negative values correspond to the (-) class.

Alternatively, wrapper methods use an iterative approach to enhance SVM performance. In our work, we first group the atomic Signatures into K mutually exclusive clusters based on the Pearson correlation coefficient using Clusteran. [57] Features are randomly assigned to one of the K clusters and then iteratively compared to other clusters and moved, if necessary, to group the best correlated features. This iteration continues until a stable division of the specified number of K clusters is achieved. Initially all K clusters are included for SVM training and then each cluster is sequentially dropped. The remaining clusters are then used for SVM training where accuracy is assessed by 10-fold cross-validation. The cluster that provided the highest accuracy when dropped was permanently removed from consideration. The process is then repeated on the surviving K-1 clusters to ultimately find an optimal subset of clusters that maximize SVM accuracy. When multiple clusters provided the same accuracy, the absolute value from the decision function in Equation 4 was summed for all misclassified compounds, and applied as a tiebreaker.

To evaluate the performed of the SVM, 10-fold cross-validation was used. Here, the training set is divided into 10 subsets containing an equal number of compounds. One of the 10 subsets is withheld while the remaining nine subsets are used to train the SVM. This process is repeated for the other nine subsets. The predictions from the model for the withheld set were identified in one of four ways: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Those classifications define the classification metrics of accuracy, sensitivity, specificity and precision.

Normally the threshold, *t*, is set to zero for SVM which causes any prediction greater than zero to be classified as active while those predicted less than zero to be inactive. Changing the threshold *t* alters the values of TP, FP, TN and FN which influences SVM performance. Integrating the area under the receiver-operator-characteristic (ROC) curve for each 10-fold subset provides an additional averaged statistic for performance evaluation. Varying the threshold *t* over the predicted SVM decision function values from a given 10-fold subset and plotting the TP rate (sensitivity) verses the FP rate (1-specificity) creates the ROC curve.

5.2.3 Overlap Metric

Not all of the atomic Signatures used to train the SVM from AID 846 are contained in the compounds to be screened from the rest of PubChem. It seems reasonable that when there is a substantial overlap in the atomic Signatures describing a given compound with those atomic Signatures in the SVM model, a more confident prediction will result. To quantitatively evaluate this confidence, we define an overlap metric, Ω, as

$$\Omega = \frac{x_{\min-\max}}{x_m} \qquad (5)$$

where x_m is the total number of atomic Signatures in a compound, and $x_{\min-\max}$ is the total number of Signatures common with the training set (here AID 846) within the minimum/maximum occurrence range. We include this latter stipulation on occurrence range to remove extrapolation effects for individual Signatures. The values of Ω will range from 0 to 1 with predictions on the compounds with higher Ω values considered to be more reliable.

5.2.4 Training and Test Sets

We choose AID 846 because it is relatively balanced set and any false positives are not as likely because the IC_{50} determination was performed in triplicate. The assay depositor reported a compound as active if $IC_{50} < 50$ μM was obtained in all three IC_{50} determinations, inconclusive if $IC_{50} < 50$ μM in only one or two determinations, and inactive for $IC_{50} > 50$ μM. [51] In this work, we reduced the active classification to 5 μM resulting in 47 active and 68 inactive compounds. Changing the activity classification provided inactive compounds from the primary screen (AID 798) to be used as a large test set for false positive predictions because these compounds had less than 40% inhibition from a single measurement at 5 μM.

5.3 Results

From the 115 compounds used in AID 846, there were 865 unique height 0, 1 and 2 atomic Signatures. Owing to time considerations, we reduced the Signature database by more than half (to 411 Signatures) when only including atomic Signatures that have occurred in at least two compounds. Next, we perform the feature selection using both filter methods and wrapper methods.

5.3.1 Filter Methods and Wrapper Methods

For the filter method, we use all 411 height 0, 1 and 2 atomic signatures in a reverse removal manner where the lowest ranked atomic Signatures using the ω_i metric are removed and performance is evaluated sequentially. Using this approach, a maximum accuracy of 80 % is obtained (as seen in Figure 5) in under 10 minutes of CPU time. By comparison using all of the 411 Signatures, without any regard to feature selection, results in an SVM with an accuracy of only 56%.

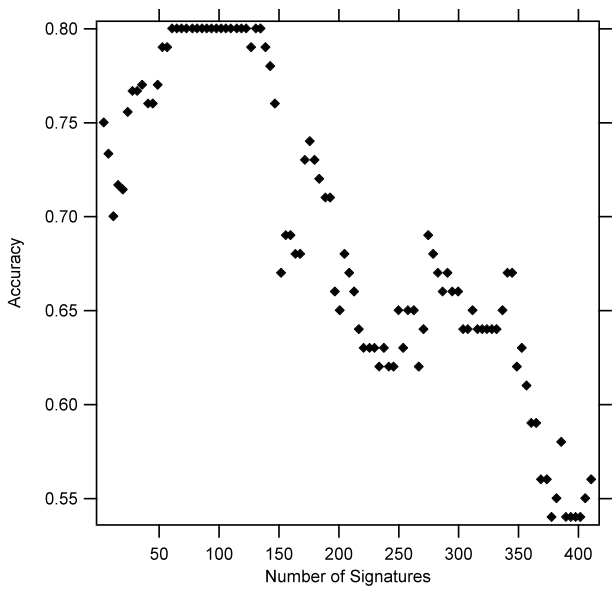


Figure 5: Feature selection using a filtering approach with performance evaluated by 10-fold cross-validation. Reprinted from "Data mining PubChem using a support vector machine with the Signature molecular descriptor: Classification of factor XIa inhibitors. Weis, Visco and Faulon, 27, 466 – 475, 2008, with permission from Elsevier"

For the wrapper method, we used k-means clustering on the set of 411 atomic Signatures. This approach increased performance to 89 % accuracy, but took much longer (approximately four days of CPU time) relative to the filter method owing to the nested loop calculations required. Note that though the wrapper method takes longer, once the

model is created, that time expense becomes less of an issue. The nearly 10 % increase in accuracy trumps the added time required.

As shown in Figure 6, the highest cross-validation accuracy for the wrapper method occurs with 22 clusters (involving a total of 105 Signatures). Accordingly, we choose this subset for the final SVM training and do not continue with the filter method model in this work. Notice also that with small number of clusters (re: atomic Signatures), there is not sufficient information available to construct a predictive model, while when too many clusters are used, overfitting results. Statistics for the optimal wrapper method SVM model (22 clusters) is provided in Table 11.

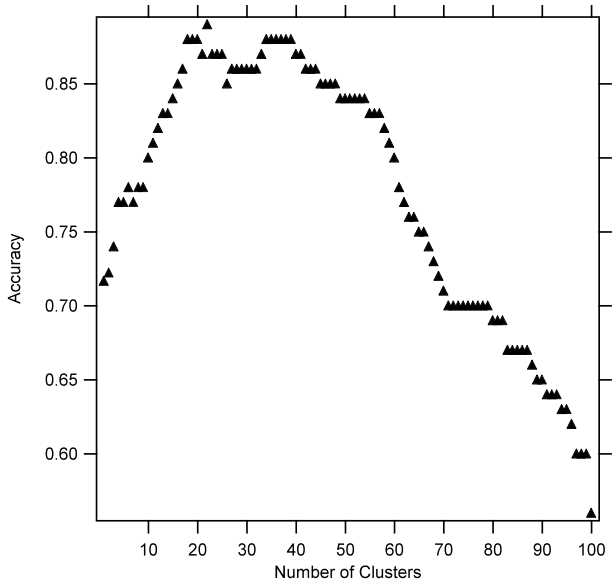


Figure 6: Feature selection using a wrapping approach with performance evaluated by 10-fold cross-validation. **Reprinted from "Data mining PubChem using a support vector machine with the Signature molecular descriptor: Classification of factor XIa inhibitors. Weis, Visco and Faulon, 27, 466 – 475, 2008, with permission from Elsevier"**

Table 11: Prediction statistics for the SVM on Factor XIa inhibitor data with 22 clusters (105 Signatures). **Reprinted from "Data mining PubChem using a support vector machine with the Signature molecular descriptor: Classification of factor XIa inhibitors. Weis, Visco and Faulon, 27, 466 – 475, 2008, with permission from Elsevier"**

X-Fold	Accuracy	AUC(ROC)	Precision	Sensitivity	Specificity
10	0.8900	0.8856	0.8500	0.9000	0.8833
4	0.8125	0.8676	0.7568	0.7727	0.8382

5.3.2 False Positive Testing

AID 798 contains 218,416 inactive compounds and this set provided one way to evaluate the SVM model created in the previous section for false positive predictions. All of the 218,416 compounds were evaluated by the SVM model and each compound was additionally characterized by both the overlap metric (Ω) and the SVM threshold (here 0, 1 and 2). What was discovered was that when the threshold value and overlap metric for a compound was high, improved classification resulted. In Table 12 we provide the results of this test with the specificity metric, defined as the ratio of true negatives to the sum of true negatives and false positives, provided in the square brackets. Ultimately, when the threshold is high (t = 2) and the overlap is maximum (Ω = 1) for a particular compound in this test set, the SVM perfectly classifies those 1442 compounds as inactive.

Table 12: Test set prediction statistics from inactive factor XIa inhibitors in AID 798. The numbers in parenthesis indicate the number of compounds from AID 798 which were above the various Ω values while the table entries indicate those compounds identified. **Reprinted from "Data mining PubChem using a support vector machine with the Signature molecular descriptor: Classification of factor XIa inhibitors. Weis, Visco and Faulon, 27, 466 – 475, 2008, with permission from Elsevier"**

t	Ω > 0 (n=218,416)	Ω > 0.6 (n=213,104)	Ω > 0.7 (n=196,352)	Ω > 0.8 (n=135,021)	Ω > 0.9 (n=37,655)	Ω = 1.0 (n=1,442)
0 <	193,762 [0.8871]	189,648 [0.8899]	176,062 [0.8967]	123,415 [0.9140]	35,226 [0.9355]	1,359 [0.9424]
1 <	211,282 [0.9673]	206,533 [0.9692]	190,957 [0.9725]	132,352 [0.9802]	37,103 [0.9853]	1,424 [0.9875]
2 <	216,210 [0.9899]	211,157 [0.9909]	194,896 [0.9926]	134,414 [0.9955]	37,590 [0.9983]	1,442 [1.0000]

5.3.3 Screening PubChem

We used the SVM developed in the previous section to screen all of the compounds in PubChem, minus those from AID 798. At the time of this study, the entire database consisted of 11,946,913 unique chemical structures. To obtain the atomic Signatures at heights 0, 1 and 2 for all of these compounds required about 6 days on a single CPU machine.

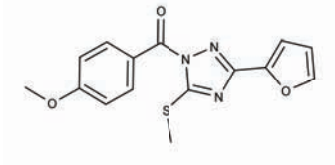
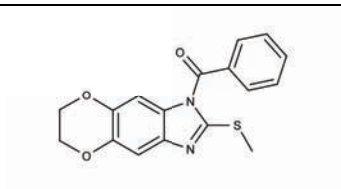
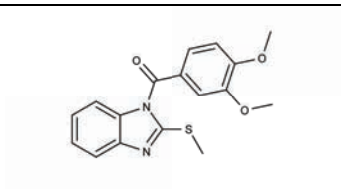
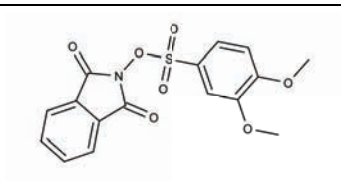
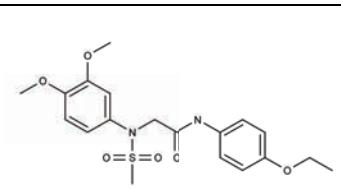
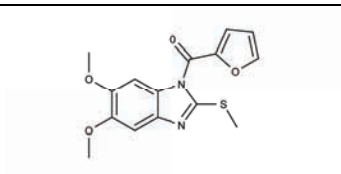
Table 13 provides the results for this screening in a manner similar to Table 12 (without the specificity rating, since the activity against Factor X1a is unknown for these compounds). In the square brackets of Table 13 we provide the percent active in that grouping, per total number of compounds evaluated.

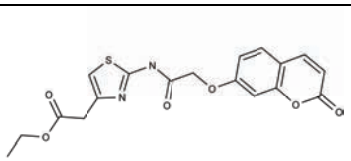
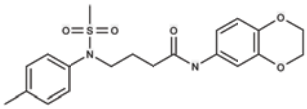
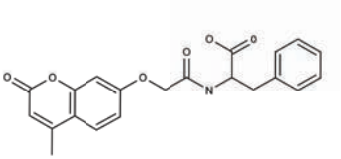
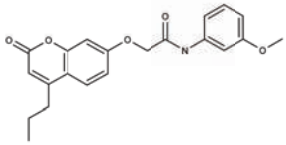
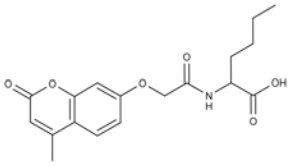
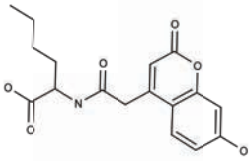
Table 13: Screening PubChem compounds for new factor XIa inhibitors with SVM. Predicted percent active, per the listed overlap metric, is provided in square brackets. **Reprinted from "Data mining PubChem using a support vector machine with the Signature molecular descriptor: Classification of factor XIa inhibitors. Weis, Visco and Faulon, 27, 466 – 475, 2008, with permission from Elsevier"**

t	Ω > 0 (n=11,946,913)	Ω > 0.6 (n=10,620,294)	Ω > 0.7 (n=9,020,826)	Ω > 0.8 (n=5,594,590)	Ω > 0.9 (n=1,378,787)	Ω = 1.0 (n=31,267)
> 0	1,828,891 [15.3]	1,345,179 [12.7]	1,028,078 [11.4]	514,022 [9.2]	91,426 [6.6]	1,300 [4.1]
> 1	715,208 [6.0]	424,227 [4.0]	300,495 [3.3]	136,455 [2.4]	23,810 [1.7]	296 [0.9]
> 2	343,052 [2.9]	149,199 [1.4]	96,063 [1.1]	37,470 [0.6]	4,899 [0.4]	4 [0.01]

There are 1300 compounds with a perfect overlap metric score and a t value greater than 0. This is the region in the chart, according to Table 12, where we would have the most confidence in the predictions. To explore this set relative to the original compounds in AID 846, we calculated the set-theoretic Tanimoto Coefficient (TC) between these 1300 compounds and all 115 compounds from AID 846 using all height 0 to 2 atomic Signatures. Recall that a value close to 1 indicated high similarity, while a value close to 0 indicates compounds which are structurally dissimilar. In Table 14 we report a sample of 12 compounds from the 1300 which have $t > 1$. For each compound, we also provide the value for the maximum Tanimoto coefficient that this compound has with the 115 compounds from AID 846. While some compounds, like CID 2658123, are just a small perturbation from one of the training set compounds (in this case, an additional $–O-CH_3$ group on the benzene ring), others (like CID 6104501) are only marginally similar to anything in the training set. Accordingly, those compounds which are most structurally dissimilar would tend to be the most non-intuitive to investigators in the field.

Table 14: A sample of one dozen compounds from the Ω =1 set. Reported are the compound ID number from PubChem, 2D-structure, maximum Tanimoto Coefficient (with AID 846), magnitude of the decision function (SVM), and binding energy (kcal/mol) from AutoDock. **Reprinted from "Data mining PubChem using a support vector machine with the Signature molecular descriptor: Classification of factor XIa inhibitors. Weis, Visco and Faulon, 27, 466 – 475, 2008, with permission from Elsevier"**

CID	Structure	TC	SVM	$E_{Binding}$ (kcal/mol)
3658123		0.94	1.29	-7.64
4426757		0.88	1.81	-7.13
977731		0.80	1.76	-7.45
2133598		0.75	1.61	-8.07
1098141		0.69	1.45	-7.28
1048578		0.62	2.17	-5.81

16418311		0.50	1.90	-8.23
6499012		0.48	1.60	-8.26
1184659		0.42	1.55	-9.20
7643488		0.42	1.06	-9.20
1556303		0.38	1.63	-7.62
6104501		0.37	1.85	-6.98

5.3.4 Docking of PubChem Compounds

To evaluate some of the compounds predicted by our model to be inhibitors of Factor X1a, we attempted to dock these compounds within the binding pocket of the protein. Specifically, we used AutoDock version 4.0.1 [58] to prepare the Factor XIa crystal structure (PDB 1zpc) [59] in complex with a ligand. Crystallographic waters were removed, polar hydrogens were added, and a 50 x 50 x 50 grid box with 0.375 A spacing was specified which was then centered on the active site. The Lamarckian genetic algorithm option for ligand conformational searching was applied with the following docking parameters: 100 runs, population size of 150, random starting point, 27,000 generations, and 25,000,000 energy evaluations.

We docked 47 compounds from AID 846 classified as most active which resulted in a minimum binding energy ranging from -5.59 to -9.15 kcal/mol. In contrast, 25 compounds deemed least active from AID 798 were randomly selected and found to have a minimum binding energy varying from -0.12 to 0.96 kcal/mol. While a large, negative binding energy does not guarantee success as an inhibitor, it can serve as an additional tool for evaluation purposes. Accordingly, we docked the 296 compounds that had a value of $\Omega=1$ with threshold $t > 1$. Note we chose this range (and not $t > 0$) since each run averaged about one day to run. Results were in agreement with the known inhibitors from AID 846 in that the minimum binding energies for this subset ranged from -5.48 to -9.84 kcal/mol. This additional evidence supports the notion that those 296 compounds are potential inhibitors of Factor X1a. Note that the binding energy for 12 of the 296 compounds is provided in Table 14. Future work in this area involves the experimental verification of the activity of some of the compounds identified in Table 14.

6 References

1. Gani, R.; Brignole, E. A., Molecular Design of Solvents for Liquid Extraction Based on UNIFAC. *Fluid Phase Equi.* 1983, 13, 331 - 340.

2. Faulon, J.-L., On Using Graph-Equivalent Classes for the Structure Elucidation of Large Molecules. *J. Chem. Info. Comput. Sci.* 1992, 32, 338-348.

3. Faulon, J.-L., Stochastic Generator of Chemical Structure. 1. Application to the Structure Elucidation of Large Molecules. *J. Chem. Info. Comput. Sci.* 1994, 34, 1204-1218.

4. Faulon, J.-L.; Visco Jr., D. P.; Pophale, R. S., The Signature Molecular Descriptor. 1. Extended Valence Sequences and Topological Indices. *J. Chem. Inf. Comput. Sci.* 2003, 43, 707-720.

5. Jackson, J. D.; Weis, D. C.; Visco, J., D. P., Potential Glucocorticoid Receptor Ligands with Pulmonary Selectivity using I-QSAR with the Signature Molecular Descriptor. *Chem. Biol. & Drug Des.* 2008, 72, 540 - 550.

6. Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M., Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substitutent Constants and Partition Coefficients. *Nature* 1962, 194, 178 - 180.

7. Gong, Z.; Xia, B.; Zhang, R.; Zhang, X.; Fan, B., Quantitative Structure-Activity Relationship Study on Fish Toxicity of Substituted Benzenes. *QSAR Comb. Sci.* 2008, 27, 967 - 976.

8. Selassie, C. D., History of Quantitative Structure-Activity Relationships. In *Burger's Medicinal Chemistry and Drug Discovery*, 6th ed.; Abraham, D. J., Ed. Wiley & Sons: 2003; Vol. 1, pp 1 - 47.

9. Fredenslund, A.; Gmehling, J.; Rasmussen, P., *Vapor-Liquid Equilibrium using UNIFAC. A Group-Contribution Method*. Elsevier: Amsterdam, 1977.

10. Reid, R. C.; Prausnitz, J. M.; Poling, B. E., *The Properties of Gases and Liquids*. 4th ed.; McGraw-Hill: New York, 1987.

11. Balaban, A. T.; Ivanciuc, O., Historical development of topological indices. In *Topological Indices and Related Descriptors in QSAR and QSPR*, Devillers, J.; Balaban, A. T., Eds. Gordon and Breach Science Publishers: Canada, 1999; pp 21 - 58.

12. Churchwell, C.; Rintoul, M. D.; Martin, S.; Visco, J., D. P.; Kotu, A.; Larson, R. S.; Sillerud, L. O.; Brown, D. C.; Faulon, J.-L., The signature molecular descriptor. 3.

Inverse quantitative structure relationship of ICAM-1 inhibitory peptides. *J. Molecular Graphics and Modeling* 2003, 22, 263 - 273.

13. Contejean, E.; Devie, H., An efficient incremental algorithm for solving systems of linear Diophantine equations. *J. Info. and Comput.* 1994, 113, 143 - 172.

14. Weis, D.; Visco, J., D. P., Computer-Aided Molecular Design Using the Signature Molecular Descriptor: Application to Solvent Selection. *Computers Chem. Engng* 2010, in press.

15. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Del Rev* 1997, 23, 3 - 25.

16. Faulon, J.-L.; Brown, W. M.; Martin, S., Reverse Engineering Chemical Structures from Molecular Descriptors: How Many Solutions?" *J. Comput. Aided Molec. Des.* 2005, 19, (9-10), 637 - 650.

17. Faulon, J.-L.; Churchwell, C. J.; Visco Jr., D. P., The Signature Molecular Descriptor. 2. Enumerating Molecules from Their Extended Valence Sequences. *J. Chem. Inf. Comput. Sci.* 2003, 43, 721-734.

18. Brignole, E. A.; Bottini, S.; Gani, R., A Strategy for Design and Selection of Solvents for Separation Processes. *Fluid Phase Equi.* 1986, 29, 125 - 132.

19. Gani, R., Case Studies in Chemical Product Design -- Use of CAMD Techniques. In *Chemical Product Design: Towards a Perspective Through Case Studies, 23*, Ng, K. M.; Gani, R.; Dam-Johansen, K., Eds. Elsevier: 2007; pp 435 - 458.

20. Camarda, K. V.; Maranas, C. D., Optimization in Polymer Design using Connectivity. *Ind. Eng. Chem. Res.* 1999, 38, 1884 - 1892.

21. Raman, V. S.; Maranas, C. D., Optimization in Product Design with Properties Correlated with Topological Indices. *Computers Chem. Engng* 1998, 22, 747 - 763.

22. Visco, J., D. P.; Pophale, R. S.; Rintoul, M. D.; Faulon, J.-L., Developing a methodology for an inverse quantitative structure activity relationship using the signature molecular descriptor. *J. Molecular Graphics and Modeling* 2002, 20, 429 - 438.

23. Curzons, A. D. C., D. C.; Cunningham, V. L., Solvent selection guide: a guide to the integration of environmental, health and safety criteria into the selection of solvents. *Clean Products and Processes* 1999, 1, 82 - 90.

24. Jimenez-Gonzalez, C.; Curzons, A. D.; Constable, D. J. C.; Cunningham, V. L., Expanding GSK's Solvent Selection Guide -- application of life cycle assessment to enhance solvent selections. *Clean Techn Environ Policy* 2005, 7, 42 - 50.

25. Hawkins, D. M., The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* 2004, 44, 1 - 12.

26. MolConverter (2007) Marvin Beans 4.1.5 Budapest, ChemAxon Ltd.

27. Abbas, C.; Rammelsberg, A. M.; Beery, K. Extraction of phytosterols from corn fiber using "green" solvents. 2003.

28. Muse, J. Environmentally friendly solvent containing isoamyl lactate. 2007.

29. Lipworth, B. J., Systemic adverse effects of inhaled corticosteroid therapy: A systematic review and meta-analysis. *Arch Intern Med* 1999, 159, 941 - 955.

30. Bodor, N.; Buchwald, P., Design and development of a soft corticosteroid, loteprednol etabonate. *Lung Biol Health Dis* 2002, 163, 541 - 564.

31. Johnson, M., Development of fluticasone propionate and comparison with other inhaled corticosteroids. *J Allergy Clin Immunol* 1998, 101, S434 - S439.

32. Biggadike, K.; Uings, I.; Farrow, S. N., Designing corticosteroid drugs for pulmonary selectivity. *Proc Am Thorac Soc* 2004, 1, 352 - 355.

33. Bodor, N. B., Pl, Corticosteroid design for the treatment of asthma: structural insights and the therapeutic potential of soft corticosteroids. *Curr Pharm Des* 2006, 12, 3241 - 3260.

34. Derendorf, H. H., G.; Meibohm, B.; Mollman, H.; Barth, J., Pharmacokinetics and pharmacodynamics of inhaled corticosteroids. *J. Allergy Clin Immunol* 1998, 101, S440 - S446.

35. Rohatagi, S.; Appajosyula, S.; Derendorf, H. S., S.; Nave, R.; Zech, K., Risk-benefit value of inhaled glucocorticoids: a pharmacokinetic/ pharmacodynamic perspective. *J. Clin Pharmacol* 2004, 44, 37 - 47.

36. Winkler, J.; Hochhaus, G.; Derendorf, H., How the lung handles drugs: pharmacokinetics and pharmacodynamics of inhaled corticosteroids. *Proc Am Thorac Soc* 2004, 1, 356 - 363.

37. Hochhaus, G., New developments in corticosteroids. *Proc Am Thorac Soc* 2004, 1, 269 - 274.

38. Buchwald, P. B., N., Soft glucocorticoid design: structural elements and physicochemical parameters determining receptor-binding affinity. *Pharmazie* 2004, 59, 396 - 404.

39. Mager, D. E.; Jusko, W. J., Quantitative structure-pharmacokinetic/ pharmacodynamic relationships of corticosteroids in man. *J. Pharma Sci.* 2002, 91, 2441 - 2451.

40. Tronde, A.; Norden, B.; Marchner, H.; Wendel, A.; Lennernaes, H.; Bengtsson, U. H., Pulmonary absorption rate and bioavailability of drugs in vivo in rats: Structure-absorption relationships and physicochemical profiling of inhaled drugs. *J. Pharma Sci.* 2003, 92, 1216 - 1233.

41. Coghlan, M. J.; Kym, P. R.; Elmore, S. W.; Wang, A. X.; Luly, J. R.; Wilcox, D., Synthesis and characterization of non-steroidal ligands for the glucocorticoid receptor: selective quinoline derivatives with prednisolone-equivalent functional activity. *J. Med. Chem.* 2001, 44, 2879 - 2885.

42. Kym, P. R.; Kort, M. E.; Coghlan, M. J.; Moore, J. L.; Tang, R.; Ratajczyk, J. D., Nonsteroidal selective glucocorticoid modulators: the effect of C-10 substitution on receptor selectivity and functional potency of 5-allyl-2,5-dihydro-2,2,4-trimethyl-1H-[1]benzopyrano[3,4-f]quinolines. *J Med Chem.* 2003, 46, 1016 - 1030.

43. Wang, J.; Shah, N.; Pantoja, C.; Meijsing, S. H.; Ho, J. D.; Scanlan, T. S., Novel arylpyrazole compounds selectively modulate glucocorticoid receptor regulatory activity. *Genes Dev* 2006, 20, 689 - 699.

44. Weis, D.; Visco, J., D. P.; Faulon, J.-L., Data mining PubChem using a support vector machine with the Signature molecular descriptor: Classification of Factor XIa inhibitors. *J. Molec Graph Mod* 2008, 27, 466 - 475.

45. Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S., NIH Molecular Libraries Initiative. *Science* 2004, 306, 1138 - 1139.

46. Zerhouni, E., Medicine. The NIH Roadmap. *Science* 2003, 302, (63 - 72).

47. Molecular Libraries Screening Centers Network. mli.nih.gov/mlscn/ (accessed Jan 7, 2008).

48. Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M., Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2007, 35, D5 - 12.

49. Martin, S.; Roe, D.; Faulon, J. L., Predicting protein-protein interactions using signature products. *Bioinformatics* 2005, 21, 218 - 226.

50. Faulon, J.-L.; Misra, M.; Martin, S.; Sale, K.; Sapra, R., Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. *Bioinformatics* 2008, 24, 225 - 233.

51. Factor XIa 1536 HTS Dose Response Confirmation. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=846 (Accessed January 14, 2008).

52. Gruber, A.; Hanson, S. R., Potential new targets for antithrombotic therapy. *Curr. Pharm. Des.* 2003, 9, (2367 - 2374).

53. Joachims, T., *Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*. MIT - Press: Boston, 1999.

54. Jorissen, R. N.; Gilson, M. K., Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Comput. Sci.* 2005, 45, 549 - 561.

55. Burgess, J. C., A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min Knowl Disc* 1998, 2, 121 - 167.

56. Golub, T. R.; Slonim, D. K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J. P.; Coller, H., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999, 286, 531 - 537.

57. Wishart, D., In. Clusteran Limited, Edinburgh, 2003.

58. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J., Automated Docking Using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function. *J. Computational Chemistry* 1998, 19, 1639 - 1662.

59. Deng, H.; Bannister, T. D.; Jin, L.; Babine, R. E.; Quinn, J.; Nagufuji, P.; Celatka, C. A., Synthesis, SAR exploration, and X-ray crystal structures of factor XIa inhibitors containing an alpha-ketothiazole arginine. *Bioorg. Med. Chem. Lett.* 2006, 16, 3049 - 3054.

Distribution List

1 MS 0899
Technical Library (electronic copy), 9536

1 MS 0123
D. L. Chavez, 1011

1 MS 1316
E. E. May, 1412

1 MS 1316
M. D. Rintoul, 1412

1 MS 1316
S. Martin, 1412

1 MS 0110
G. Heffelfinger, 1210