Use of Atomic and Bond Parameters in a Spectral Representation of a Molecule for Physical Property Determination

Edward S. Blurock
blurock@risc.uni-linz.ac.at
Research Institute for Symbolic Computation
Johannes Kepler University
A-4040 Linz, Austria

A representation of a molecule for structure-property correlation (QSPR) studies is introduced. It is a generalisation of feature counting, but is not limited to atomic number, types of bonds or other simple features. Thereby a molecule is represented as a *spectra* of atomic and bond parameters resulting from molecular calculations (semi-empirical, ab initio, etc.); Value ranges are counted. The calculational method determines what information is represented in the molecular description. The main use of this representation is in inductive learning analysis of relationships between experimental molecular properties and molecular substructures. An example is given using semi-empirical calculations as the basis for the spectra.

1 Introduction

A molecular property is a quantity that is intrinsic to the molecule as a whole. But a molecule is not a single homogeneous entity. It has an internal structure made up of individual atoms and bonds. For many molecular properties the presence of isolated structural features are responsible for its behavior. Thus it is desirable to be able to represent the molecule in terms of these structural features. These can, in turn, be used to find a structural relationship to a molecular property. It is the goal of structure-activity studies to find simple (and computationally efficient) relationships between structural features and experimental molecular properties.

A common representation of a molecule is a *graph*, i.e. a set of *vertices* (the atoms) and *edges* or connections between the nodes (the bonds). Several molecular representations being used to find structure-activity relationships have been derived from graphs[1, 2, 3, 4]. Such a representation is adaquate for explaining the properties of the molecules for which the governing mechanism is the result of covalent bonding. Molecular properties dependent on stereochemical effects or the three-dimensional character of the molecule, are not described by this graph representation. However, the primary mechanisms of many molecular properties can usually be (to a certain extent) explained without three-dimensional information and thus justify the usage of the graph representations.

A graph theoretical approach being used more often in structure-activity studies derives a set of numerical *graph descriptors* in which the structural features are reduced to a set of numbers[5]. This is, of course, advantageous for analysis methods which require numerical values as input, such as statistical regression or neural network analysis.

To describe structural features one could resort to structural *keys* (for example those found in a molecular database[6]) describing specific features that are expected to relate to the property. These keys can be simple entities (such as those signifying the presence of heteroatoms), more complex general structural feature detectors (for example, signifying the existence of rings or aromatic groups) or substructure flags (signifying the existence of substructures) within the molecule.

Calculational methods, ranging from semi-empirical to quantum-mechanical, yield not only properties that relate to the molecule as a whole, but also properties of the environments of the individual atoms. Though some of these methods are based on simple graph or connection table descriptions of the molecule, they are not restricted to such (for example, ab initio calculations). However, using the individual atomic environment information that one acquires from these methods in structure-activity relationship studies is problematic. A representation is introduced here which enables the direct use of this information and thus provides a mechanism to include an arbitrary level of structural information in the molecular representation for use in structure-activity analyses.

1.1 Spectra of Atomic Properties

An approach introduced here represents the molecule as a *spectrum* of atomic (or bond) properties. Associated with each atom (or bond) in the molecule is a calculated property representing a description of its environment. The range of values for this property is divided into ranges and the number of times the value falls within each range is counted; The spectrum is the distribution of values. This property could come from any calculation ranging from semi-empirical to quantum-mechanical. In what way the environment is represented depends on the property described and method used. This representation takes a set of continuous values and converts them to a set of integer parameters.

This type of representation, when used in structure-activity studies, asks the question:

Does the existence, anywhere in the molecule, of an atom (or bond) and its corresponding environment correlate to or represent the driving influence for a given molecular property?

The major advantages of such a representation are:

Flexibility Any method yielding a set of atomic or bond properties can be the basis of the representation. Thus any complexity (from graphical to 3-dimensional structural information to inter-molecular interactions) can be used in the representation of the molecule.

Physical Basis In comparison to, for example, pure graph theoretical quantities, semi-empirical quantities have a direct physical basis making it easier to relate it to a physical mechanism.

The spectrum is made up of intervals, each of which represents a range of values of the calculated property. There is a parameter for every interval within every property used. Due to the fact that the number parameters can be quite large, such a representation is more suitable for use with *inductive learning* methods. In fact, some inductive learning methods can be used to reduce the large number of parameters to a reasonable size. In addition, the variance or the number of different values (i.e. the individual values of the *peaks* of the spectrum) can be small (i.e. they are more a set of discrete values as opposed to a range of continuous values) and inductive learning methods can handle them more effectively.

2 The Method

2.1 Production of Spectrum

The spectrum consists of a set of intervals of values of a particular physical property. The *peak height* of each interval is the number of atoms within the given molecule which has a property value within this range.

The production of a set of spectra corresponding to a set of molecules consists of the following steps:

Molecular Calculation For each molecule, a molecular calculation (semi-empirical, ab initio, etc.) is performed and the quantities (partial charge, electron density, polarizability, etc.) are collected for each atom (see Section 2.2 and Figure 1). Bond properties can be formed by taking, for each bond, the difference between the atomic values forming the bond (see Section 2.3).

Set of Values for each Molecule For each molecule the set of property values are extracted and a *Set of Values* for each molecule is created.

Determination of Spectrum Range The entire set of molecular property lists is examined and the highest and lowest values are determined. These values represent the boundaries of the spectral range (see Section 2.4).

Determination of Interval Size The determination of the number of intervals is a matter of experimentation with respect to the parameter property and the property to be predicted. The proper number of intervals is a compromise between differentiation and information content of a parameter. A smaller interval would tend to isolate individual cases and thus be an indication of the single specific case. Too much detail would tend to refer to very specific instances and thus, is not appropriate for generalization. A larger interval ignores fine differences and is suitable for describing a *class* of atoms within a molecule (see Section 2.4).

Building of Molecule Spectrum The molecular spectrum is then built by counting the number of atoms within the molecule having property values within the particular range. This count is called the *peak height* (see Figure 2). The spectra consists of a set of pairs: the interval range (represented by, for example, its middle point), and the number of atoms within the molecule with this value.

Translation for Use in Inductive Learning Each spectrum interval represents a parameter for use in the analysis using inductive learning methods. The given parameter's value is the number of atoms within the molecule that fell within this range (see Section 2.5).

Substructure Relationships *Significant* intervals within the spectrum can be related back to molecular substructures. An inductive learning analysis which can associate (i.e. correlate) the substructures of the entire molecule to a range of values.

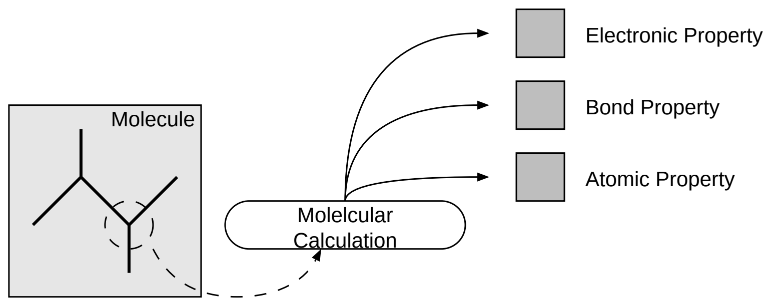

Figure 1: The set of parameters are calculated for each atomic environment in each molecule.

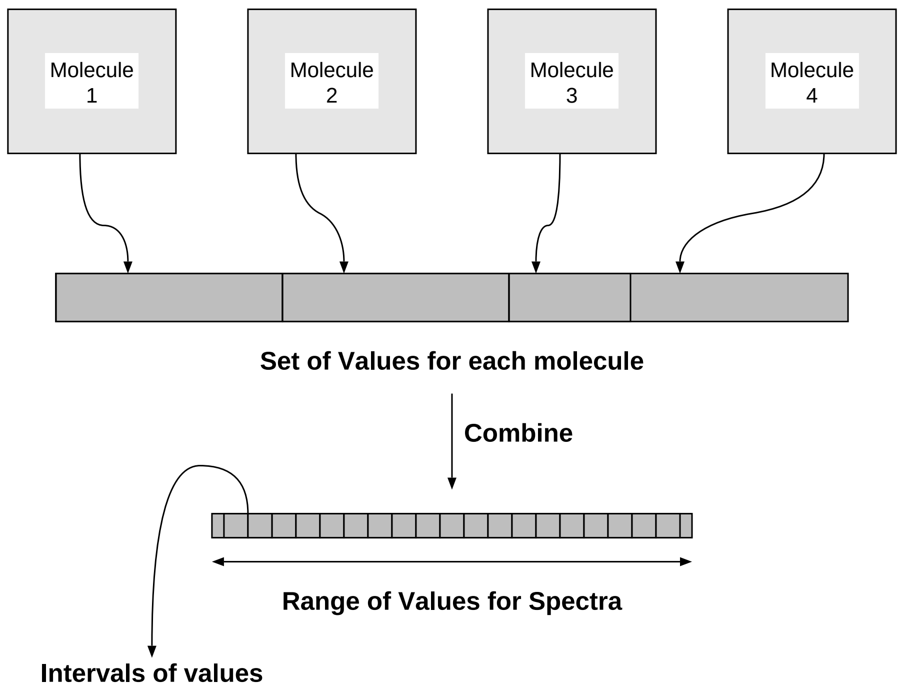

Figure 2: The set of intervals for the spectra and their sizes are derived from the set of parameter values from all the atoms in all the molecules. Each molecule produces a list of values. The set of lists for all molecules is then combined and evaluated. The spectral range is determined by the highest and lowest values. The size of the interval is determined by how many intervals are desired.

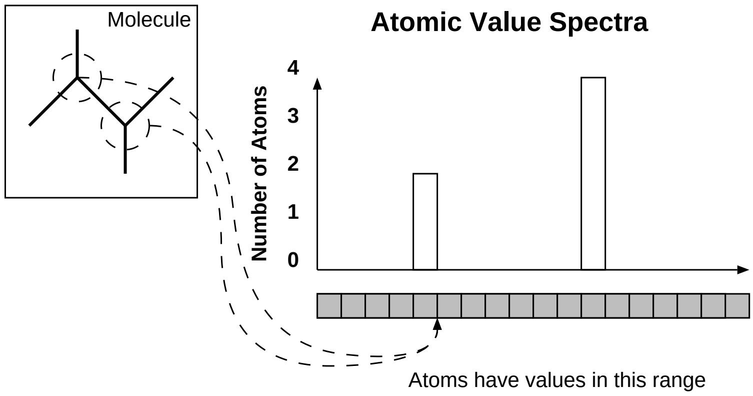

Figure 3: The number of atoms having a value within a given range is the *peak* of the spectrum. In this example, there are two types of atoms: 2 of one type and 4 of the other.

2.2 Molecular Value Calculations

The set of electronic properties can be calculated using a variety of methods. Any method which calculates individual values for each atom can be used. These can range from semi-empirical to quantum mechanical calculations. For example, the semi-empirical methods of Sello[8] and Gasteiger[9] can be used to calculate several atomic electronic properties. They are advantageous, especially for preliminary studies, because they are simple and fast. An example of the use of such a calculation is given in section 3

2.3 Atomic and Bond Spectrum

The values from the calculations are atomic values. However, the final property that is to be described can be more dependent on bonding properties. One method of creating bonding properties from atomic properties is to take, for every bond, the difference between the two atomic values of the atoms making up the bond. Such a principle was used by Gasteiger to form his *Reactivity Function*[10], a function that describes whether a bond can be broken or not.

2.4 Calculate Molecular Spectra

The word *spectrum* is used in a very general sense here. It refers to a spectrum of (for example, electronic) property values for the atoms in the molecule. Along the X-axis is the set of intervals of values of a particular property. The Y-axis is the number (or percent) of atoms in the molecule having a value within the given interval.

For a given property the spectra is set up in the following manner:

1. Collect all values for all atoms in all molecules. From this set determine:
   - Range of Values (the highest and lowest value)
   - Interval Size (based on distribution and the accuracy wanted)
2. For each molecule, count the number of atoms in each of the intervals. This value is the spectral *peak* for the particular interval.

2.5 Inductive Learning

For the prediction of the molecular property values with respect to the molecular value *spectra* the author prefers (but the method is not restricted to) inductive learning analyses. The ID3 method[13, 7, 11] by itself produces a decision tree which can (for numeric values) decide whether a value is above or below a given boundary (or within a given range). The complete range of property values can be treated using *Series of Decision Trees[14]* technique. This uses a consecutive set of limits (which effectively divide the range into intervals) to form a set of decision trees. Since the decision process is visible through the decision tree, qualitative information as to which parameters (and hence structures) are *significant* can be derived (the ANALYSIS system does this automatically).

The ID3 method is used in this study because it can handle a representation in which the number of parameters is large with respect to the number of objects under study. The method selects out a single question about a parameter out of the entire set as opposed to using the entire set in some mathematical combination. This selection process can actually be used to reduce the number of parameters to a set of 'significant' parameters (See section 4.1) which then could be used in more traditional analyses.

3 Example

In this section the spectra for a set of molecules is produced and analysed. With the limited class of molecules used (i.e. hydrocarbons) direct assignments between the spectral parameters and the structural features which they represent can be made. This assignment was done by hand. These assignments can be used to gain qualitative information about structure-activity relationships.

A set of 238 hydrocarbons is used. There are no heteroatoms within the molecules in the set. The set consists of alkanes, alkenes, alkynes and aromatic compounds.

3.1 Semi-Empirical Calculations

Two semi-empirical calculations are used for the creation of the spectra:

Partial Charge The partial charge of each atom in each molecule is calculated using the method of Sello [8].

Polarizability The polarizability of each atom in each molecule is calculated using the method of Gasteiger [9].

Both of these methods (as implemented by the author) use as input a connection table. No stereochemical information is used.

3.2 Semi-Empirical Spectrum

Four spectra are produced from these semi-empirical calculations (the corresponding parameter names are given in parenthesis):

Atomic Charge (ACx) The partial charge of each atom is used directly.

Bond Charge (BCx) For each bond, the difference between the partial charge on each adjoining atom is used.

Atomic Polarizability (APx) The polarizability of each atom is used directly.

Bond Polarizability (BPx) For each bond, the difference between the polarizability on each atom is used.

Each Spectrum has 15 intervals (see table 1 for definition of spectral interval values). Thus, there are 60 parameters in the entire molecular description. From each spectrum, a set of parameters is produced and used within the analysis. A parameter asks the question:

> How many atoms within the molecule have a spectral value within the spectral parameter interval (i.e. atoms of a given type).

For example, the spectral parameter BC6 asks the question:

> How many atoms within the molecule have a bond charge between .0625 and .075

3.3 Structural Assignments

Spectral parameters are derived, through semi-empirical calculations, from the structural features of the molecule. The purpose of doing this is to have a numerical description of the molecule that can be handled by analysis (in this case, inductive learning) techniques.

Spectral Parameter	End of First Interval	Interval Size
Atomic Partial Charge	-0.0875	0.0125
Atomic Polarizability	1.2	0.2
Bond Partial Charge	0.0	0.0125
Bond Polarizability	0.0	0.2

Table 1: This is a table of the spectrum interval boundaries and sizes. Each individual spectra has fifteen intervals. The end boundary is given for the first interval. For example, the second interval of the *atomic partial charge* include the values greater than -0.0875 but less than or equal to 0.0750.

This section assigns these spectral parameters back to their respective structural features. In some cases the structural assignments are not absolute, but yield rather a trend. This is especially true of the structures associated with the polarizability parameters. These parameters represent a broader molecular environment (i.e. encompass atoms further away in the molecule) and were harder to interpret.

3.3.1 Atomic Charge Parameters

As table 2 shows, charge parameters are good indicators of very basic types of carbon environments. They are fairly insensitive to influences of the environment further away, ex. whether the carbon is bonded to a double bond or aromatic. Thus, they are fairly easy to assign. These parameters can then be thought of as giving the first order description of the chemical environment.

Table 5 gives some examples of assignments of some of the parameters. The table is divided into series where the progressions of parameter values can be easily observed. For example, AC5 (**A**tomic Partial **C**harge parameter **5**) and AC12 indicate the chain length in the n-alkanes and AC4 gives an indication of the amount of substitution or side-chains. Figures 4 and 5 give illustrations of the parameter assignments on some alkanes and alkenes, respectively.

3.3.2 Atomic Polarizability Parameters

In comparison to the atomic charge parameters, the atomic polarization parameters are more sensitive to the molecular environment. They can include effects from atoms in the alpha and beta positions. This is illustrated with a few examples in figures 4 and 5. Some polarization values are on the border between two parameters such as AP6 and AP7. This adds another dimension to the molecular description by adding, what I shall call, second order effects.

Table 3 gives a summary of the easily assignable parameters. These are not the exclusive assignments (as with the charge parameters) but represent a large portion of the examples. One sees in Table 6 that the assignments (and thus trends) are not as distinct as by the charge parameters. However, within the various molecular series, distinct patterns are recognizable.

3.4 Bond Parameters

Bond parameter descriptions were simply made by taking the absolute value of the difference between the two atoms in the bond. The motivation for this is twofold:

Broader Description The description of the molecular environment is broadened by the combination of the effects of both atoms in the bond. For example, the simple primary, secondary and tertiary description of the atoms is expanded to 6 different bond environments.

Reactivity Several of the terms of Gasteiger's *Reactivity Function* involve bond parameters formed by the difference of values between the atoms of the bonds. Thus the inclusion of bond effects increases the chance of including kinetic reactivity effects.

3.4.1 Bond Charge Difference Parameters

Since the atomic partial charge parameters were fairly insensitive to structures beyond the atom itself, the charge bond parameters are also fairly easily assigned. Table 4 shows a brief summary of some of the structural assignments.

4 Study: Qualitative Information

As one sees, one can make a connection between the parameters and structural groups. This connection is useful when trying to make qualitative observations in structure activity relationships. In the study outlined here, relationships between the structural features represented by the parameters and the research octane number (RON) are to be found. This is meant to be preliminary step in finding a quantitative relationship. One of the concrete results that one can obtain from the qualitative studies is the reduction of the number of parameters.

4.1 Machine Learning: ID3

As one sees, the number of spectral parameters can be quite large and, at times, interdependent. This is a deadly combination for many analytical methods. It is for this reason that the machine learning method ID3 was used. Instead of finding a relationship which is a combination of all parameters simultaneously, as

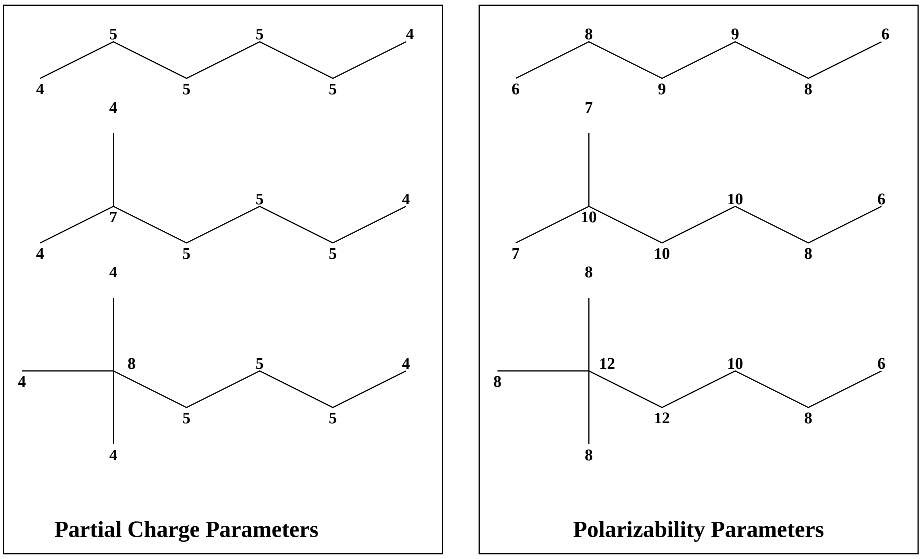

Figure 4: This is an illustration of the carbon atom assignments of the charge and polarizability parameters on a few alkanes.

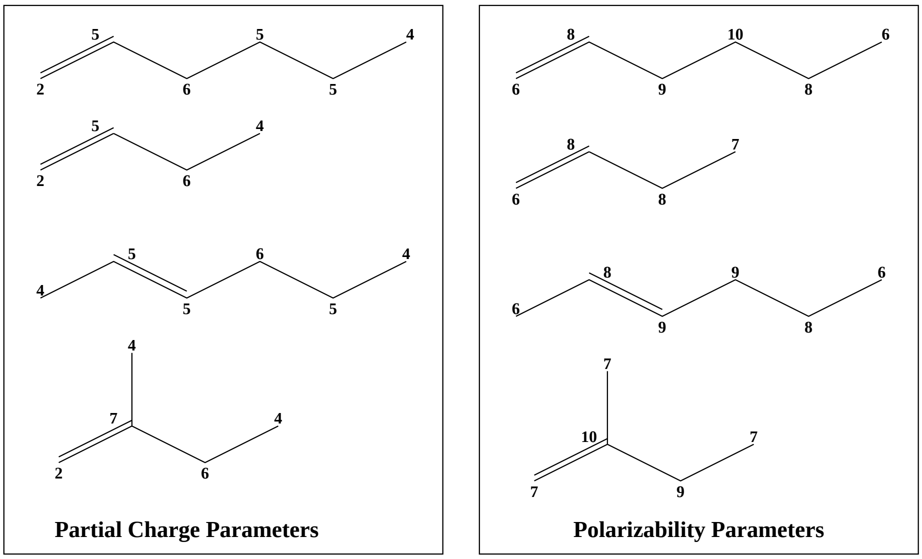

Figure 5: This is an illustration of the carbon atom assignments of the charge and polarizability parameters on a few alkenes.

- Alkane Carbons

AC4	Primary Carbons
AC5	Secondary Carbons
AC7	Tertiary Carbons
AC8	Quaternary Carbons

- Alkene Carbons

AC2	The terminal alkene carbons
AC6	The secondary carbon next to an alkene or aromatic

- Hydrogens

AC10	The hydrogens on primary and secondary carbons
AC12	The hydrogens on double bonded carbon

Table 2: This is a summary of the partial charge parameters and their associated structural features.

in statistical or neural network analysis, it selects out single relevant parameters and builds up a relationship. This technique is used not only to be able to handle the overabundance of parameters, but also to reduce the set down to a manageable number. A similar study, on the same set of molecules, was done where the molecular representation was essentially the presence of substructures. In this study, the substructures were the carbon skeletons of the molecules involved. Hence, the number of parameters equaled the number of objects under study. Here, an extension of the ID3 method was used to isolate those substructures which were relevant to RON determination. The purpose of this study is not to provide a decision tree for the prediction of octane number, but to automatically provide qualitative information about the structural features of different octane number ranges. For this reason the decision tree are not shown.

In this study, there are basically four steps leading to the qualitative results of the analysis:

ID3 Decision Tree The direct results of the analysis is a decision tree which decides into which RON range a certain molecule falls on the basis of the spectral parameters.

Significant Parameters The ANALYSIS system extends this analysis by ordering and selecting the parameters which are particularly *significant* in

- Primary Carbons: This gives a further characterization of the simple terminal carbon represented by AC4 (alkanes) and AC2 (alkenes).

AP6	Terminal carbon on the end of a longer unsubstituted chain.
AP7	Terminal carbon bonded to a secondary carbon or near a double bond
AP8	Terminal carbon bonded to a tertiary carbon

- Secondary Carbons: This gives a further characterization of the secondary carbons represented by AC5 and AC6.

AP8	The next to the last carbon in a longer chain
AP9	The middle of a long unbranched carbon chain
AP10	Secondary carbon on or near a branched carbon or double bond

- Tertiary Carbons: This gives a further characterization of the tertiary carbons represented by AC7.

AP10	Both the alkane tertiary and the alkene (in addition to those near such groups)

- Hydrogens:

AP3	Hydrogens on terminal alkane and alkene carbons
AP4	Among others, Hydrogens on carbons directly attached to aromatic groups

Table 3: This is a brief summary of some of the assignable polarization parameters and their associated structural features

- General:

BC1	This is the set of bonds that are *exactly* equivalent (by symmetry)
BC2	These are the bonds that are *nearly* equivalent.

- Bonds to Hydrogens:

BC8	These are the hydrogens on terminal carbons on alkanes and those directly bonded to aromatic rings.
BC9	These are hydrogens on double-bonded (non-aromatic, non-terminal) carbons.
BC6	These are hydrogens on secondary carbons

- Alkane-Alkane Carbon Bonds:

BC3	These are secondary carbons bonded to primary and tertiary carbons and tertiary carbons bonded to quaternary carbons.
BC4	These are the tertiary carbons bonded to primary carbons.
BC5	These are secondary carbons bonded to tertiary carbons.
BC6	These are quaternary carbons bonded to primary carbons.

- Alkane-Alkene Carbon Bonds:

BC2	These are primary alkanes connected to unsubstituted alkene carbon (ex. 2-butene has two).

- Alkene-Alkene Carbon Bonds:

BC2	These are unsubstituted double bonds not equivalent by symmetry.
BC7	Substituted double bond as in 2-Methyl-1-butene.

Table 4: This is a brief summary of some of the assignable bond charge parameters and their associated structural features. It should be noted that there is considerable overlap (i.e. two structural features could use the same parameter).

Name	Atomic Partial Charge						
	2	4	5	6	8	10	12
Ethane	0	2	0	0	0	6	0
Propane	0	2	1	0	0	8	0
Butane	0	2	2	0	0	10	0
Pentane	0	2	3	0	0	12	0
Hexane	0	2	4	0	0	14	0
Heptane	0	2	5	0	0	16	0
Propane,2-Methyl-	0	3	0	0	0	10	0
Butane, 2-Methyl-	0	3	1	0	0	12	0
Pentane, 2-Methyl-	0	3	2	0	0	14	0
Hexane, 2-Methyl-	0	3	3	0	0	16	0
Heptane, 2-Methyl-	0	3	4	0	0	18	0
Pentane, 3-Methyl-	0	3	2	0	0	14	0
Hexane, 3-Methyl-	0	3	3	0	0	16	0
Propane, 2,2-Dimethyl-	0	4	0	0	1	12	0
Butane,2,2-Dimethyl-	0	4	1	0	1	14	0
Pentane, 2,2-Dimethyl-	0	4	2	0	1	16	0
Butane, 2,3-Dimethyl-	0	4	0	0	0	14	0
Pentane, 2,3-Dimethyl-	0	4	1	0	0	16	0
Pentane, 3,3-Dimethyl-	0	4	2	0	1	16	0
Hexane, 2,4-Dimethyl-	0	4	2	0	0	18	0
Hexane, 2,5-Dimethyl-	0	4	2	0	0	18	0
Hexane, 3,3-Dimethyl-	0	4	3	0	1	18	0
Hexane, 3,4-Dimethyl-	0	4	2	0	0	18	0

Table 5: This is a table of a series of alkanes to show the correspondence between structure and the atomic partial charge parameters.

Name	Atomic Polarizability								
	3	4	5	6	7	8	9	13	14
Ethane	6	0	2	0	0	0	0	0	0
Propane	6	2	0	2	1	0	0	0	0
Butane	6	4	0	2	0	2	0	0	0
Pentane	6	4	2	2	0	2	1	0	0
Hexane	6	4	4	2	0	2	2	0	0
Heptane	6	4	6	2	0	2	2	0	0
Propane,2-Methyl-	0	9	1	0	3	0	1	0	0
Butane, 2-Methyl-	3	6	3	0	3	0	1	0	0
Pentane, 2-Methyl-	3	6	5	1	2	0	1	0	0
Hexane, 2-Methyl-	3	8	3	3	2	1	0	0	0
Heptane, 2-Methyl-	9	2	8	0	2	1	1	0	0
Pentane, 3-Methyl-	6	3	4	1	2	1	2	0	0
Hexane, 3-Methyl-	6	5	2	5	0	1	2	0	0
Propane, 2,2-Dimethyl-	0	12	0	0	0	4	0	0	0
Butane, 2,2-Dimethyl-	0	12	2	0	1	3	0	0	0
Pentane, 2,2-Dimethyl-	3	9	2	3	0	3	1	0	0
Butane, 2,3-Dimethyl-	0	12	0	2	0	4	0	0	0
Pentane, 2,3-Dimethyl-	3	9	2	2	1	3	0	0	0
Pentane, 3,3-Dimethyl-	0	6	6	4	2	0	2	1	0
Hexane, 2,4-Dimethyl-	3	9	2	5	2	1	1	0	0
Hexane, 2,5-Dimethyl-	12	0	2	4	4	0	0	0	0
Hexane, 3,3-Dimethyl-	3	3	10	3	1	0	3	1	0
Hexane, 3,4-Dimethyl-	6	6	4	0	4	2	0	0	0

Table 6: This is a table of a series of alkanes to show the correspondence between structure and the atomic polarization parameters.

the decision making process.

Parameter-Structural Using the results of section 3.3 the *significant* parameter questions and can be associated with distinct structural features

Structure-RON The significant structural features can be associated with distinct RON ranges

The raw output from the ID3 analysis is a decision tree. In this study 11 decision trees were made representing the intervals RON greater than 10,20,30,....,110.

4.2 Significant Parameters

An extension of the ID3 method provided by ANALYSIS is to isolate for each decision tree the *significant* parameter questions. These are questions involving a distinct parameter (in the form of defining a range with greater-than-or-equal-to) which appear to be important in making the decision whether the molecule lies within the decision tree's RON range.

Associated with each parameter question is a quantity which gives an indication of its importance, the higher the number, the more significant is the question in the decision making process. The establishment of this order is explained in an earlier paper[16].

Tables 2 through 6 summarise this information according to the sets of parameters. The information given in these tables are as follows:

Variable This is a distinct parameter question defining a range of parameter values.

Highest Quality Associated with each parameter question is a numerical quantity ranking how good or significant the question is with respect to the decision tree in which it appeared. If the question was significant in several decision trees, then the highest quantity it reached is listed.

RON ranges These are the decision trees in which the parameter question was deemed to be the significant (with respect to the numeric quantity).

RON peak Of the RON ranges, this is the RON range in which the parameter question had the greatest significance.

If one looks at the set of parameter questions involving the parameter $BC6$ (Table 4), we see that this parameter is significant through the entire range of RON values. In fact, examination of the decision trees shows that $BC6 \geq 6$ and $BC6 \geq 8$ are the first questions in 7 of the 11 decision trees.

4.3 Interpretation of Significant Questions in Tree

The results of the previous section revealed the significant spectral parameters, but the parameters themselves are too abstract to be of any value in understanding the chemistry of the problem. However, in combination with the results of section 3.3, the parameter questions can be transformed into structural features within a molecule. The end results are statements of the form:

> The presence of this structural feature denotes an RON value greater-than-or-equal-to a given value

Due to the fact that the correspondence between the atomic charge and structural features was fairly well defined, the assignment of structural features to corresponding RON ranges could be attempted. It is much more difficult to extract such intuitive information from the polarisability parameters. A summary of correspondences are shown in tables 7 and 8.

The assignment of structural features did not follow purely from the single parameter question alone (listed as *Condition* in the tables). Its position in the decision trees was also considered. That is, if it appeared as the second, third, etc. question (i.e. questions before already isolated out a distinct subset of the molecules) this reduced set of molecules was considered (thus simplifying the interpretation, i.e. the finding of the structure-RON correspondence). A summary of results is shown in the tables 7 and 8.

5 Conclusion

A method was introduced is a generalization of simple counting procedures of atomic and bond information. Instead of being limited to traditional values such as atomic number or types of bonds, the information can stem from semi-empirical, ab initio or any other calculational method in the analyses. The complexity or completeness of the molecular description is dependent on the calculation involved. For example, in this paper, semi-empirical methods based on graph theoretical (connection table) descriptions of the molecule were used as the basis. The advantage of the spectral representation will come when more complex computations, such as ab initio results, are used as the basis. In this way, for example, three dimensional and stereo-chemical information can be included. This model is an initial attempt to provide a basis for molecular representation which can be used in studying structure activity relationships.

Condition	Structural Feature	Examples w.r.t RON
$AC2 \geq 1$	The presence of a terminal alkene as in the substituted 1-alkenes	RON values around 100
$AC4 \geq 1$	AC4 equal to zero means no primary carbons in the molecule	Used in the lower stages of decision making around 100 to isolate out 1,3-Cyclohexadiene(75), 1,4-Cyclohexadiene(75) and 1,5-Hexadiene(71)
$AC4 \geq 4$	Greater than four primary carbons means a molecule that is somewhat branched	One class of examples are the dimethylhexanes which with one exception (55) have RON values in the 60's and 70's
$AC5 \geq 2$	This sets a lower limit of secondary carbons, either existence of chains or	Used in the lower stages to isolate molecules having RON values in the 70's, but in general those molecules having $AC5 < 2$ have RON values greater than 90 and 100
$AC5 \geq 3$	This also sets an upper limit of secondary carbons	Used to separate highly substituted alkanes and cycloalkanes ($AC5 < 3$) which generally have a RON greater than 90)
$AC5 \geq 4$	Molecules with longer unsubstituted chains as a substructure	Longer chains denote the lower range of RON ($<$ 50 for the most part).
$AC5 \geq 5$	Molecules with longer unsubstituted chains as a substructure	Longer chains denote the lower range of RON ($<$ 50 for the most part).
$AC5 \geq 6$ $AC5 \geq 7$	Very long chains	Very long chains denote RON values around zero
$AC6 \geq 1$	The presence of Alkenes	This was used in the situation to isolate long chain alkenes (having RON around 50) or short alkenes (having RON above 90).
$AC6 \geq 2$	Either an 1-alkene substituted at carbon 2 or a non-terminal double bond alkene (i.e. 2-alkene, 3-alkene, etc.)	A large portion have RON values greater than 85, but at least above 71 (exception 4-Octyne: RON 60).
$AC10 \geq 11$	These are somewhat larger alkanes or alkenes with at least 5 primary or secondary alkane carbons.	In the $RON \geq 100$ tree it separated out several cyclohexenes and cyclopentenes having RON around 90.
$AC10 \geq 15$	These are somewhat larger alkanes or alkenes with at least 6 primary or secondary alkane carbons.	A large portion of these molecules have RON values greater than 60

Table 7: This is a summary of some of the interpretable atomic charge parameter questions. Some of the interpretations are relative to their use in the decision tree (i.e. do not necessarily reflect the entire set of molecules). In addition, 'longer chains' and other qualitative terms are relative to the alkanes under analysis.

Condition	Structural Feature	Examples w.r.t RON
$BC1 \geq 4$	At least 4 exactly equivalent bonds, signifying a highly symmetric molelcule. These are long chain n-alkanes or ring systems.	The large n-alkanes have very low (less than 20) RON values.
$BC2 \geq 1$	Those molecules with no near symmetry have BC2 equal to zero (exact symmetry is denoted by BC!). These include larger molecules (pentane and above) with some (unsymmetric) branching and those with exact symmetry	The molecules described have generally RON less than 80
$BC3 \geq 4$	Those that have no branching or branching near the molecule ends (i.e. carbon 2) have low BC3.	The decision occured late in the decision tree, but isolated out those less than RON 80
$BC3 \geq 6$	Very few examples of such, but those that do have onger branches (at least ethyl-) occuring in the middle of a chain (for example 3-Ethylpentane). Another example is 1,3,5-trimethylcyclohexane.	These generally have RON less than 70 (6 examples in the data set)
$BC5 \geq 2$ $BC5 \geq 3$	Branching in the middle of the molecule	Used late in the decision tree to isolate RON under 90 or 100
$BC6 \geq 2$ $BC6 \geq 4$	Small molecules, alkenes or substituted alkanes or alkenes	These have generally large RON values (over 90)
$BC6 \geq 6$ $BC6 \geq 8$	Very generally substitution at carbon 1 or 2 in alkanes	A very general trend (it was chosen often as the first question in the decision tree) for RON less than 90
$BC12 \geq 12$	These have at least 6 unsubstituted carbons. Common examples are the cyclohexane with larger groups attached	For the most part these molecules have RON values less than 40 or 50
$BC8 \geq 1$ $BC8 \geq 2$	In combination with $BC6 < 4$ (the previous question of the tree) these are substituted benzene and small alkane and alkene molecules	Fairly consistently these have RON values in the 90s and 100s
$BC8 \geq 1$ $BC8 \geq 2$	In combination with $BC6 < 6$ and $AP13 \geq 1$ (the previous questions of the tree) these are substituted benzine and highly substituted pentane and pentene molecules	Fairly consistently these have RON values in the 100s

Table 8: This is a summary of some of the interpretable bond charge parameter questions. Some of the interpretations are relative to their use in the decision tree (i.e. do not necessarily reflect the entire set of molecules). In addition, 'longer chains' and other qualitative terms are relative to the alkanes under analysis.

References

[1] Randic, Milan. Structure-Activity Studies. *International Journal of Quantum Chemistry: Quantum Biology Symposium* **1984**, 11, 137–143.

[2] Schultz, Harry P. ;Schultz, Emily B.;Schultz, Tor P.; Topological Organic Chemistry. 4. Graph Theory, Matrix Permanents, and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 69–72.

[3] Mihalić, Zlatko; Nikolić, Sonja; Trinajstić. Comparative Study of Molecular Descriptors Derived from the Distance Matrix. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 28–37. and references therein.

[4] Hall, Lowell H.; Mohoney, Brian; Kier, Lemont B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, 31, 76–82. and references therein.

[5] Randic, Milan. Orthogonal Molecular Descriptors. *New J. Chem.* **1991**, 15, 417–525.

[6] Stanton, David T.; Jurs, Peter C.;Hicks, Martin G. Computer-Assisted Prediction of Normal Boiling Points of Furans, Tetrahydorfurans, and Thiophenes. *J. Chem. Inf. Comput. Sci.* **1991**, 31, 301–310.

[7] Blurock, Edward S. Analysis 3.0: Implementation, Extensions and Use of the ID3 Algorithm. Technical report, RISC-Linz, Linz, Austria, 1992.

[8] Baumer, L.; Sala, G.;Sello, G. Residual Charges on Atoms in Organic Structures: Molecules containing Charged and Backdonating Atoms. *Tetrahedron Computer Methodology* **1989**, 2, 105–118.

[9] Gasteiger, J.;Hutchings, Michael G. Quantification of Effective Polarizability. Applications to Studies of X-Ray Photoelectron Spectroscopy and Alkylamine Protonation. *J. Chem. Soc. Perkin Trans. II* **1984**, 559–564.

[10] Gasteiger, Johann ;Hutchings, Michael G.;Christoph, Bernd; Gann, Leopold; Hiller, Christian; Loew, Peter; Marsili, Mario; Saller, Heinz; Yuki, Kazumi. A New Treatment of Chemical Reactivity: Development of EROS, an Expert System for Reaction Prediction and Synthesis Design. *Topics in Current Chemistry* **1987**, 137, 19–72.

[11] Blurock, Edward S. Use of Semi-Empirical Spectra and Inductive Learning in the Qualitative Analysis of Structure and Research Octane Number. Technical report, RISC-Linz, Linz, Austria, 1992.

[12] Blurock, Edward S. Use of Semi-Empirical Spectra and Neural Nets in Research Octane Number Determination. Technical report, RISC-Linz, Linz, Austria, 1992.

[13] Quinlan, J.R. Decision Trees and Decision Making. *IEEE trans. on Systems, Man and Cybernetics* **1990**, 20, 339–346.

[14] Blurock, Edward S. Automatic Learning of Chemical Concepts: Research Octane Number and Molecular Substructures. Technical report, RISC-Linz, Linz, Austria, 1992.

[15] Molecular Design, Ltd., San Leandro, CA, USA. *ChemBase: Chemist's Personal Software Series.*

[16] Blurock, Edward S. Automatic Learning of Chemical Concepts: Research Octane Number and Molecular Substructures. *Computers Chem.* **1995**, 19, 91–99.