

TOWARDS MORE RELIABLE EXTRAPOLATION ALGORITHMS
WITH APPLICATIONS TO ORGANIC CHEMISTRY

JAIME NAVA

Department of Computer Science

APPROVED:

Vladik Kreinovich, Chair, Ph.D.

Luc Longpré, Ph.D.

M. Lawrence Ellzey Jr., Ph.D.

Scott A. Starks, Ph.D.

Patricia D. Witherspoon, Ph.D.
Dean of the Graduate School

©Copyright

by

Jaime Nava

2009

a mi

mamá y papá

TOWARDS MORE RELIABLE EXTRAPOLATION ALGORITHMS
WITH APPLICATIONS TO ORGANIC CHEMISTRY

by

JAIME NAVA

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

July 2009

Acknowledgements

My thanks and sincere appreciation to Dr. Vladik Kreinovich, who always seems to know the exact piece of advice required for taking the next step in my never-ending journey to become a better scientist, while keeping it fun and interesting at the same time.

I also wish to extend my gratitude to the other members of my committee, Dr. Luc Longpré and Dr. Larry Ellzey, who made this work possible thanks to their supervision, input, and expert advice.

Thanks as well to Dr. Guillermo Restrepo for his attentive collaboration and contribution that led to the Chemistry case study presented in this thesis.

Special thanks to Dr. Pat Teller, for her invaluable lessons and opportunities as an undergrad, as well as to Dr. Rodrigo Romero, Dr. Roberto Araiza, Gaby Aguilera, Dr. Steve Roach, and Professor Kay Roy for their advice, motivation, and support.

NOTE: This thesis was submitted to my Supervising Committee on July 31, 2009.

Abstract

One of the main objectives of science and engineering is to predict the results of different situations. For example, in Newton's mechanics, we want to predict the positions and velocities of different objects (e.g., planets) at future moments of time.

In this thesis, as a case study, we take the problem of predicting the properties of new chemical substances. One of the main objectives of chemistry is to design new molecules (and, more generally, new chemical compounds) which are useful for various practical tasks. New substances have already resulted in new materials for buildings and for spaceships, new explosives and new fuels, new medicines, etc. New compounds are being designed and tested all the time.

For example, this is how new medicines are designed: a large number of different promising substances are synthesized and tested, but only a few turn out to be practically useful. Synthesizing a new compound is often difficult and time-consuming. It is therefore desirable to predict the properties of new compounds, so as to filter out the ones which do not have the desired properties.

In physics, usually, we know the exact equations that describe the objects of interest, and we know how to solve these equations. This is the case for Newton's mechanics. In such situations, we face a purely mathematical problem: to solve these equations and thus compute the value y of the desired characteristic based on the known values of the parameters x_1, \dots, x_n that describe the given objects.

In many other application areas, we either do not know the equations, or the equations are so complex that we do not know how to solve them. For example, in chemistry, in principle, we can use the equations of quantum mechanics to describe an arbitrary chemical substance, but in practice, especially for organic substances, these equations are too complex to solve.

In the situations in which we do not know the equations – or we do not know how to

solve the equations – the prediction problem takes the following form:

- we know the values of a quantity $v(a)$ for some objects a , and
- we want to predict the values of this quantity for some other objects a' .

There are many examples of successful predictions in science. In many cases, to solve a new prediction problem, researchers use ideas which are specific for this problem. In addition to problem-specific predictions, there exist successful prediction algorithms. Most of these algorithms are heuristic – in the sense that they are empirically successful, but since they do not have any domain-related theoretical justification, there is no guarantee that they will work in other situations as well.

It is therefore desirable to provide more justified extrapolation algorithms – e.g., by providing a solid justification for the existing heuristic techniques. In numerical mathematics, more justified algorithms are often called *more reliable*. In this thesis, we provide a theoretical justification for an important class of heuristic extrapolation algorithms – algorithms based on partially ordered sets (posets). This justification makes these algorithms more reliable in the sense of numerical mathematics.

Table of Contents

	Page
Acknowledgements	v
Abstract	vi
Table of Contents	viii
Chapter	
1 Introduction: the Prediction Problem	1
1.1 Prediction: General Problem	1
1.2 Application Area: Predicting Properties of New Chemical Substances	1
1.3 Traditional Approach to Prediction: Prediction in Physics	2
1.4 Prediction Outside Physics is Often More Difficult	2
1.5 Prediction as Extrapolation	2
1.6 Usually, Prediction is Problem-Specific	3
1.7 General Prediction Algorithms: a Problem	3
1.8 What We Do in This Thesis	4
1.9 Structure of the Thesis	4
2 Heuristic Extrapolation Algorithms Based on Partially Ordered Sets (Posets)	5
2.1 Partially Ordered Sets	5
2.2 Prediction on Posets: Idea	5
2.3 Prediction on Posets: Algorithms	6
3 Application Area: Prediction Problem for Chemical Substances Like Benzenes and Cubanenes	7
3.1 Need for New Substances	7
3.2 Properties of New Substances are Difficult to Predict	7
3.3 New Substances are Difficult to Synthesize: an Example	8

3.4	Step-By-Step Transitions as a Way to Predict Properties of Derivative Compounds	11
3.5	The Step-By-Step Approach Can Be Helpful for Prediction	12
3.6	Another Example; Benzene-Based Molecules	13
3.7	Chemical Problem	13
3.8	The Importance of Symmetry	14
4	The Current Use of Poset-Based Extrapolation in Organic Chemistry	19
4.1	First Step: Reduction to Posets	19
4.2	Values Defined on Posets	19
4.3	Prediction on Posets: General Approach (Reminder)	20
4.4	Towards Chemical Applications	21
4.5	The Use of Symmetry	22
4.6	First Example: Benzenes	22
4.7	Second Example: Cubanes	25
4.8	Results	27
4.9	Problems with the Poset Approach	29
5	Discrete Taylor Series: Main Idea	31
5.1	Prediction: One of the Main Objectives of Science	31
5.2	Taylor Series: a Standard Tool for Solving (Continuous) Problems in Science and Engineering	31
5.3	From Continuous to Discrete Taylor Series	33
5.4	Discrete Taylor Expansions Can Be Further Simplified	36
6	Equivalence Between the Poset-Related Approaches and the Discrete Taylor Series Approach	38
6.1	Discrete Taylor Series: Reminder	38
6.2	Chemical Substances: Application Area	38
6.3	Poset-Related Approaches: Reminder	39
6.4	Poset-Related Approaches Reformulated in Terms of the Discrete Variables	39

6.5	Proof That the Discrete Taylor Series Are Indeed Equivalent to the Poset Formula	40
6.6	Important Observation: The Presence of Symmetry Does Not Change the Equivalence	41
6.7	Discussion: Advantages of the Taylor Representation	43
6.7.1	Main Advantage of the Taylor Representation	43
6.7.2	Additional Advantage of the Taylor Representation: Taylor Series Can Clarify the Equivalence of Different Arrangements	43
6.7.3	Additional Advantage: a Detailed Description	45
6.7.4	Example	46
7	Conclusion	47
	References	49
	Curriculum Vitae	51

Chapter 1

Introduction: the Prediction Problem

1.1 Prediction: General Problem

One of the main objectives of science and engineering is to predict the results of different situations. For example, in Newton's mechanics, we want to predict the positions and velocities of different objects (e.g., planets) at future moments of time.

1.2 Application Area: Predicting Properties of New Chemical Substances

In this thesis, as an application area, we take the problem of predicting the properties of new chemical substances.

One of the main objectives of chemistry is to design new molecules (and, more generally, new chemical compounds) which are efficient for various practical tasks. New substances have already resulted in new materials for buildings and for spaceships, new explosives and new fuels, new medicines, etc. New compounds are being designed and tested all the time.

For example, this is how new medicines are designed: a large number of different promising substances are synthesized and tested, but only a few turn out to be practically useful.

Designing a new compound is often difficult and time-consuming. It is therefore desirable to predict the properties of new compounds, so as to filter out the ones which do not have the desired properties.

1.3 Traditional Approach to Prediction: Prediction in Physics

In physics, usually, we know the exact equations that describe the objects of interest, and we know how to solve these equations. This is the case for Newton's mechanics.

In such situations, we face a purely mathematical problem: to solve these equations and thus compute the value y of the desired characteristic based on the known values of the parameters x_1, \dots, x_n that describe the given objects.

1.4 Prediction Outside Physics is Often More Difficult

In many other application areas, we either do not know the equations, or the equations are so complex that we do not know how to solve them.

In chemistry, in principle, we can use the equations of quantum mechanics to describe an arbitrary chemical substance, but in practice, especially for organic substances, these equations are too complex to solve.

1.5 Prediction as Extrapolation

In the situations in which we do not know the equations – or we do not know how to solve the equations – the prediction problem takes the following form:

- we know the values of a quantity $v(a)$ for some objects a , and
- we want to predict the values of this quantity for some other objects a' .

The simplest case of this situation is when, to characterize each object, it is sufficient to know the value of one numerical characteristic x . In this simplest case, the value v of the desired quantity is also uniquely determined by the value x of this characteristic, i.e., $v = v(x)$. Under this assumption, the above prediction problem takes the following form:

- we know the values $v(x_1), \dots, v(x_n)$ corresponding to the values x_1, \dots, x_n of the characteristic x ;
- we want to predict the value $v(x)$ corresponding to the new value $x \neq x_i$.

In mathematics, this specific problem is called *extrapolation* if the value x is outside the interval covered by the values x_i , and *interpolation* if x is inside this interval.

In view of this, the general problem is also called *extrapolation*.

1.6 Usually, Prediction is Problem-Specific

There are many examples of successful predictions in science. In many cases, to solve a new prediction problem, researchers use ideas which are specific for this problem.

This general statement is true for chemistry as well. In chemistry, there have also been many interesting predictions. Some of these predictions came from the *classification* of chemical elements and/or substances; see, e.g., [12, 15]. The classical example of such predictions are predictions made possible by Mendeleev's Periodic Law. In chemistry, there are also many examples of successful predictions of properties based on the molecular structure. For example, many properties of polymers can be predicted based on their molecular structure; see, e.g., [17].

1.7 General Prediction Algorithms: a Problem

In addition to problem-specific predictions, there exist successful prediction algorithms. Most of these algorithms are heuristic – in the sense that they are empirically successful, but since they do not have any domain-related theoretical justification, there is no guarantee that they will work in other situations as well.

It is therefore desirable to provide more justified extrapolation algorithms – e.g., by providing a solid justification for the existing heuristic techniques. In numerical mathematics, more justified algorithms are often called *more reliable*.

1.8 What We Do in This Thesis

In this thesis, we provide a theoretical justification for an important class of heuristic extrapolation algorithms – algorithms based on partially ordered sets (posets). This justification makes these algorithms more reliable in the sense of numerical mathematics.

1.9 Structure of the Thesis

The thesis is structured as follows. In Chapter 2, we briefly describe the heuristic extrapolation algorithms based on partially ordered sets. In Chapter 3, we describe in detail, our application area: prediction problem for chemical substances like benzenes and cubanes. In Chapter 4, we describe how the poset-based algorithms are used in our chemical problem. In Chapters 5 and 6, we describe our justification for these heuristic algorithms: namely, in Chapter 5, we describe the discrete Taylor series approach, and in Chapter 6, we show that this approach is equivalent to the poset heuristic.

Chapter 2

Heuristic Extrapolation Algorithms Based on Partially Ordered Sets (Posets)

2.1 Partially Ordered Sets

In many practical situations, there is a natural partial order $x \leq y$ on the set of all possible objects. In other words, the set of all possible objects forms a partially ordered set.

For example, in chemistry, we can select a class of chemical reactions and say that $x \leq y$ if the compound y can be obtained from the compound x by reactions from this class.

The problem of predicting the values $v(a)$ defined on a poset occurs not only in chemistry, the same problem occurs in other applications as well; see, e.g., [14].

2.2 Prediction on Posets: Idea

To solve such problems, Gian-Carlo Rota, a renowned mathematician from MIT, proposed the following idea [14]. We can represent an arbitrary dependence $v(a)$ as

$$v(a) = \sum_{b: b \leq a} V(b)$$

for some values $V(b)$. The possibility to find n such values $V(b)$ corresponding to n different elements b of the poset comes from the fact that the above equations form a system of n linear equations (for n different a). In the general case, a system of n linear equations with

n unknown has a unique solution. (In principle, there are degenerate cases when a system of n linear equations with n unknowns does not have a solution or has an infinite number of different solutions, but in [14] it was proven that for posets we always have a unique solution.)

2.3 Prediction on Posets: Algorithms

The above formula by itself does not immediately lead to a prediction algorithm since to use this formula for determining n values $v(a)$, we need to know n different values $V(b)$. However, in practice, in many cases, some of the values $V(b)$ turn out to be negligible.

If we know which values $V(b)$ are negligible, we can replace them with 0s and thus, consider a model with $m < n$ non-zero parameters $V(b)$. Once we have such an expression, we can then:

- measure the value $v(a_1), \dots, v(a_m)$ of the desired quantity v for $m < n$ different elements a_1, \dots, a_m ;
- use the resulting system of m linear equations with m unknown parameters $V(b)$ to find the values of these parameters; and then
- use the known values of the parameters $V(b)$ to predict all the remaining values $v(a)$ ($a \neq a_1, \dots, a_m$).

In practice, measurements are inevitably imprecise; to decrease the effect of the corresponding measurement errors on the predicted values, it makes sense to measure more than m values $v(a_i)$. Then, we get more than m equations with m unknowns $V(b)$, i.e. we have an *over-determined* system of linear equations. To solve this over-determined system of linear equations, we can, e.g., use the Least Squares method (see, e.g. [16]).

Chapter 3

Application Area: Prediction Problem for Chemical Substances Like Benzenes and Cubanes

3.1 Need for New Substances

One of the main objectives of chemistry is to design new molecules, and, more generally, new chemical compounds which are efficient for various practical tasks. New substances have already resulted in new materials for buildings and for spaceships, new explosives and new fuels, new medicines, etc. New compounds are being designed and tested all the time.

3.2 Properties of New Substances are Difficult to Predict

The main reason why new compounds need to be tested is that it is very difficult to accurately predict the properties of the new substance. Often, it is assumed that compound's properties are related to the molecular structure associated with the compound, and compounds with similar molecular structure have similar properties. This is known as the QSAR (Quantitative Structure-Activity Relationship) hypothesis. With such a hypothesis at hand and reasonable approximate techniques, we can predict approximate values of a substance's characteristics; based on these predictions, we can consider the corresponding

molecule to be potentially useful for a given application. However, the approximate prediction methods are usually very crude. As a result, we do not know whether the resulting molecule is actually useful or not until we have actually synthesized this molecule and measured the values of the corresponding characteristics.

For example, this is how new medicines are designed: a large number of different promising substances are synthesized and tested, but only a few turn out to be practically useful.

Comment. To avoid confusion, we should mention that in chemistry, relevant numerical characteristics are called *numerical properties*, or simply “properties”, for short. We will therefore use this term “numerical properties” in the current text.

This may be somewhat confusing to computer science readers because in computer science, a property is something which can be either true or false, but cannot have numerical values: e.g., “ x is positive” ($x > 0$) is a property, but the value x itself is not called property in computer science.

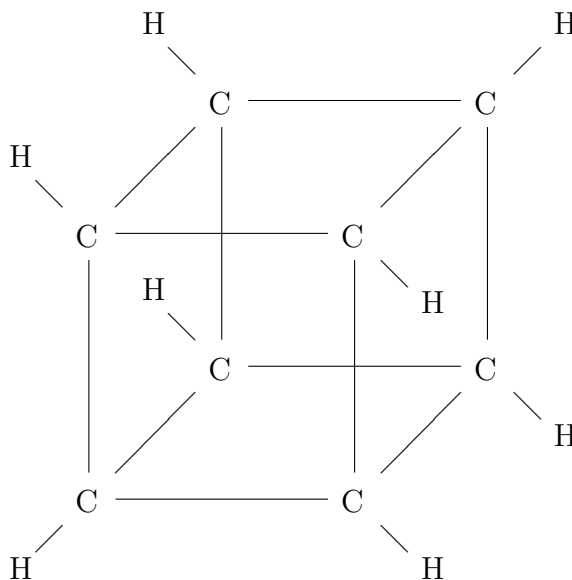
3.3 New Substances are Difficult to Synthesize: an Example

Synthesis of a new compound is usually a very difficult task, a task that takes a large amount of time and resources. An example of practically useful synthetic compounds is the family of cubane molecules, which are highly reactive molecules used as fuels and explosives.

Cubane, a molecule in a shape of a cube, has been theoretically considered for a long time. The reason for this consideration is that chemists are often interested in the molecules which are optimal for some practical applications. Since the formulation of these optimization problems does not change if we simply rotate the coordinates axis, the optimal solutions often have rotational symmetries. For example, many organic molecules are based on benzene, a highly symmetric molecule. Another example of a highly symmetric and highly successful molecule is a carbon fullerene, actively used in nanotechnology. It was

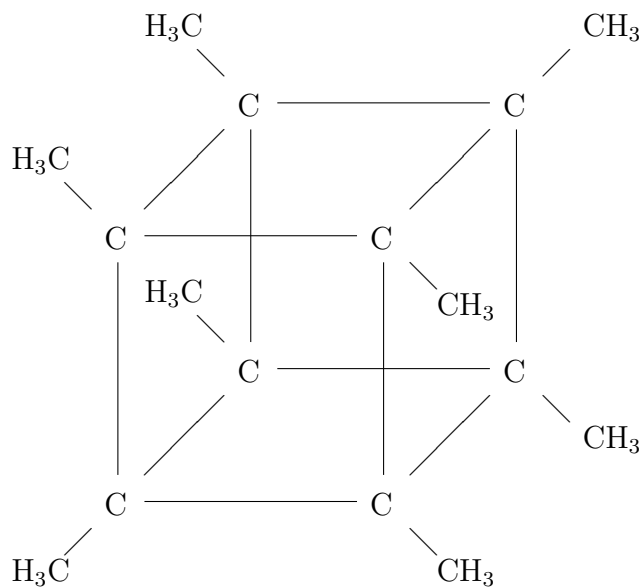
recognized for some time that a highly symmetric cubic shape would make the molecule highly reactive, but until it was actually synthesized in 1964, researchers believed that such molecule would be highly unstable.

Cubanes are a family of substances that include the “basic” cubane molecule C_8H_8

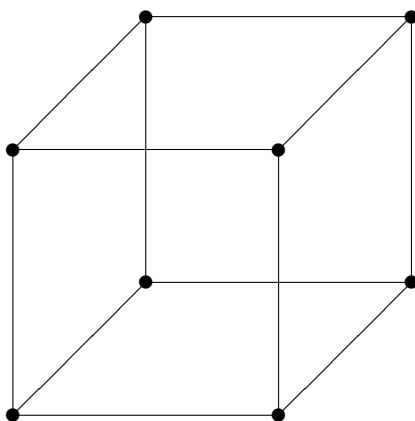


and *cubane derivatives*, i.e., substances which are obtained from the basic cubane by replacing its hydrogen atoms H with other atoms or atom groups (called ligands).

For example, if we replace each hydrogen atom with a methyl group CH_3 , we obtain the following cubane derivative:



For simplicity, such substituted molecules are often denoted by placing a bold dot in places where substitution occurred:



After the synthesis of cubane, attained by Eaton and Cole in 1964 [2], a wealth of cubane derivatives and cubane-like compounds have been synthesized. Cubanes are kinetically stable and highly reactive; as a result, at present, they (specially *nitrocubanes*, with NO_2 ligands) are actively used as high-density, high-energy fuels and explosives, and researchers are investigating the potential of using cubanes (and similarly high-energy molecules) in medicine [4] and nanotechnology [3].

Synthesis of a new cubane usually requires a large amount of time and resources.

Comment. The above example is about the synthesis of an organic compound, but similar problems appear in inorganic chemistry as well.

3.4 Step-By-Step Transitions as a Way to Predict Properties of Derivative Compounds

As we have mentioned earlier, predictions of properties of chemical compounds are based on the QSAR (Quantitative Structure-Activity Relationship) hypothesis, according to which compounds with similar molecular structure have similar properties. In particular, this means that if the compound b is obtained from the compound a by adding a single ligand, then the properties of the compound b should be similar to the properties of the compound a .

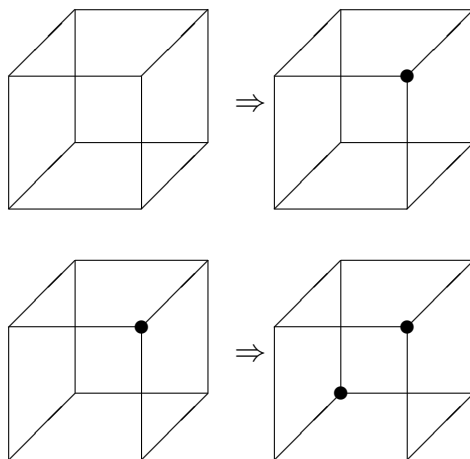
When b is obtained from a by adding several ligands, the similarity between the molecular structures of a and b is smaller and therefore, the properties of a and b are not as similar to each other. Thus, to predict the properties of a derivative compound with several ligands, it is reasonable to consider a *sequence* of molecules in which the next one is obtained by making a small change to the previous one:

- we start with a known basic molecule,
- we make small changes step-by-step, and
- we end up with the desired molecule.

For some substances, this is exactly how they are synthesized: by adding certain ligands in a step-by-step manner to a basic substance.

In many cases, this is not how the compound is synthesized, but it is a reasonable way to predict the properties of the desired compound. For example, one can picture the octamethyl cubane (in which methyl ligands CH_3 are added to all 8 locations) as a sequence of *methylations* (adding the methyl ligand) starting from cubane to obtain methyl cubane

(with a single methyl added), afterwards to obtain dimethyl cubanes (with two methyl ligands added), and so on to finally have the octamethyl cubane. The same step-by-step approach can be imagined for additions or for eliminations over a basic molecule.



...

3.5 The Step-By-Step Approach Can Be Helpful for Prediction

As we have mentioned, it is usually difficult to accurately predict the properties of a new substance before this substance is synthesized.

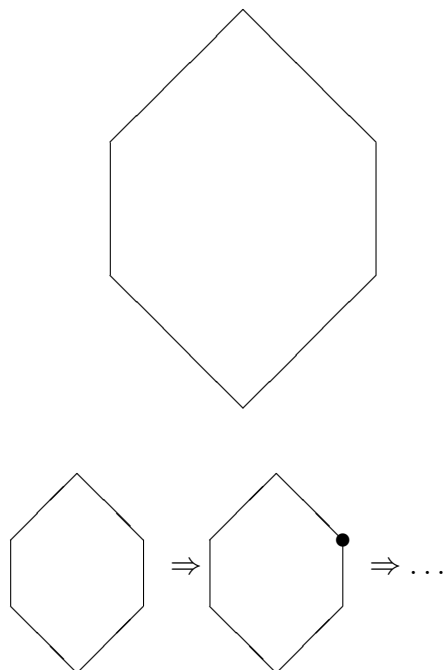
However, the step-by-step approach leads to the following idea: the properties of the generated molecules change gradually, starting with the properties of the original molecule, and eventually resulting in the characteristics of the desired new molecule.

So maybe knowing the properties of the substances at early stages of the step-by-step approach, one can use such a knowledge to try to predict the properties of other substances related by the same step-by-step procedure. Depending on the predicted numerical property one can decide whether or not to synthesize the corresponding substance, which ends up saving time and resources.

3.6 Another Example; Benzene-Based Molecules

A similar step-by-step approach can be applied to other substances, for example to benzene, C_6H_6 , and its substituted derivatives. These substances are very practically important since most organic molecules contain benzene or benzene-based components.

Similarly to the cubanes case, instead of the usual detailed molecular representation of benzene and its derivatives, we will use the following simplified graph representation:



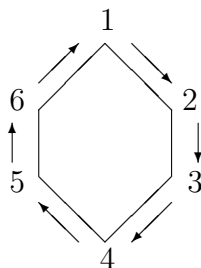
Other examples of step-by-step substitutions are given in [7, 10].

3.7 Chemical Problem

In all these cases, we have the following chemical problem. We have a step-by-step process of designing a new chemical substance by sequential replacements. Based on the experimentally determined values of the desired quantity for the original molecule and for the first few replacement results, we would like to predict the values of this quantity for the molecules on all the stages of the replacement process.

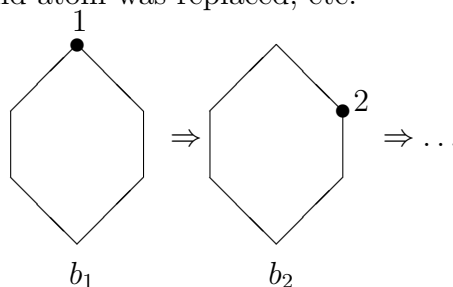
3.8 The Importance of Symmetry

It is worth mentioning that molecules such as benzene or cubane have the following property of *symmetry*: for every two atoms, we can find a rotation that moves the first atom into the position of the second one while keeping the molecule configuration intact. For example, for benzene, rotation by 60° transforms the first atom into the second one, the second into the third one, etc.



It is worth mentioning that in chemistry textbooks, the benzene molecules are usually described with alternative single-valence and double-valence links. If this simplified picture was a correct description of the benzene atom, then we would not have such a symmetry, since rotation by 60° would transfer a single-valence link into a different double-valence link. However, in reality, all the links are absolutely identical, they are neither purely single-valence nor purely double-valence links. They represent a sharing of electrons that cannot be correctly explained in traditional valence terms – and can only be explained by quantum mechanics.

The 60° rotation transforms the molecule in which the first atom was replaced into the molecule in which the second atom was replaced, etc.



A simple rotation does not change the chemical properties of a molecule – and hence, does not change the values of any numerical property of the substance. Let us start with

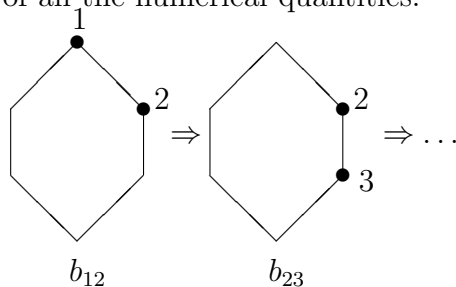
monosubstituted molecules, i.e., molecules in which a single ligand has been substituted. All the monosubstituted molecules can be obtained from each other by rotation. We can therefore conclude that all these molecules have the same values of all the numerical quantities.

Comment. In chemical terms, we can say that these molecules are *equivalent* and that they can be treated as a single *molecular species*.

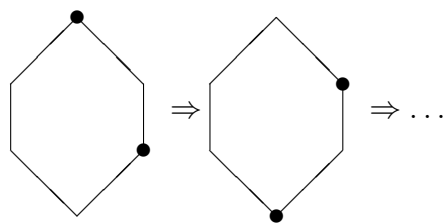
Similar rotation-related analysis can be applied to other benzene derivatives. Let us consider benzenes in which two H atoms are replaced; in chemistry, such molecules are called *disubstituted*. The chemical properties of these molecules depend on the relative location of the replaced H atoms. Three such locations are possible:

- In some molecules, two neighboring H atoms are replaced. This substitution is called *ortho*.
- In other molecules, the replaced H atoms are separated by the “distance” of 2, i.e., in which the replaced atoms are not immediate neighbors, but immediate neighbors of immediate neighbors. This kind of substitution is called *meta*.
- Finally, it is also possible that the replaced atoms are separated by the “distance” of 3. This substitution is called *para*.

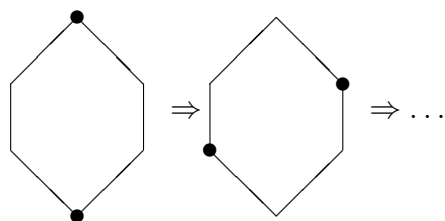
All the orthosubstituted molecules can be obtained from each other by rotation, and thus, have the same values of all the numerical quantities:



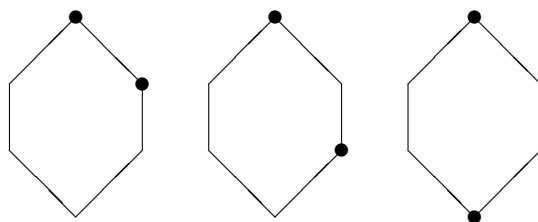
Similarly, all the metasubstituted molecules can be obtained from each other by an appropriate rotation:



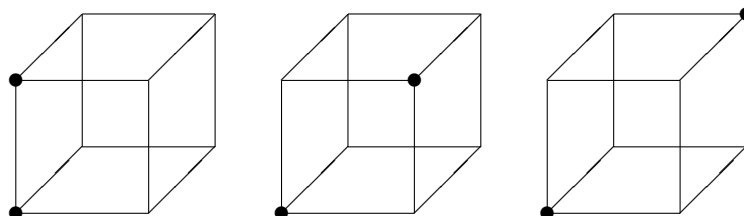
and all parasubstituted molecules are similarly equivalent to each other:



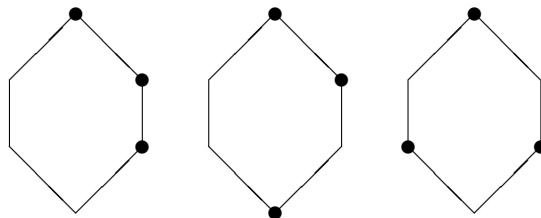
As a result, for molecules obtained from benzene by a double replacement, we have at most three different molecular species, depending on whether the distance between the two replaced atoms is 1, 2, or 3 (*ortho*, *meta* and *para*, respectively):



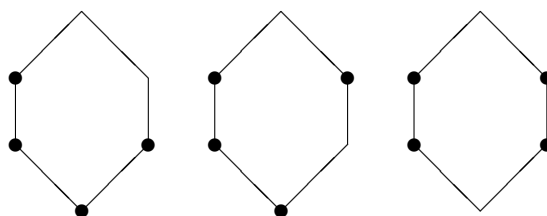
Cubanes have a similar result: there is only one monosubstituted cubane. For disubstituted cubanes, there are 3 possible molecular structures depending on whether the distance between the two replaced atoms is 1, 2, or 3:



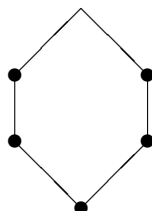
This symmetry also reduces the number of multiple-atom replacement molecules. For example, for benzene, when we take rotation symmetry into account, we conclude that there are only 3 different trisubstituted structures (with 3 ligands):



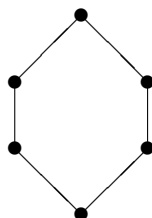
Tetrasubstitutions (with 4 ligands) can be described by listing the two ($6 - 4 = 2$) H atoms which have not been substituted, so we also have 3 possible alternatives:



For pentasubstitutions (with 5 ligands), we only need to describe a single non-substituted H atom:



Finally, we have the completely substituted (*hexasubstituted*) molecule, the result of substituting all six H atoms:



Hence, in total we need to predict the numerical properties of

$$1 + 1 + 3 + 3 + 1 + 1 = 13$$

different molecules:

- the original nonsubstituted benzene molecule;
- the monosubstituted;
- three disubstituted;
- three trisubstituted;
- three tetrasubstituted;
- the pentasubstituted; and
- the completely substituted (*hexasubstituted*) molecule.

Chapter 4

The Current Use of Poset-Based Extrapolation in Organic Chemistry

4.1 First Step: Reduction to Posets

To predict numerical properties of substances related by substitutions, D. J. Klein and others have proposed approaches based on the notion of a partially ordered set (*poset*, for short); see, e.g., [1, 5, 6, 7, 8, 9]. These approaches use the fact that the substitution reaction leads to a natural (partial) ordering on the set of all the corresponding molecules: namely, a relation $x \leq y$ meaning that the molecule y either coincides with x , or can be obtained from the molecule x by one or several substitutions.

Molecules are elements of this partially ordered set (poset, for short). Let n denote the number of molecules, i.e., the number of elements in this set. For a molecule a from this set, let $v(a)$ denote the corresponding numerical property of a , such as energy, boiling point, or vapor pressure at a certain temperature.

4.2 Values Defined on Posets

The problem of predicting the values $v(a)$ defined on a poset occurs not only in chemistry, the same problem occurs in other applications as well; see, e.g., [14]. To solve such problems, Gian-Carlo Rota, a renowned mathematician from MIT, proposed the following idea [14].

We can represent an arbitrary dependence $v(a)$ as

$$v(a) = \sum_{b: b \leq a} V(b)$$

for some values $V(b)$. The possibility to find n such values $V(b)$ corresponding to n different elements b of the poset comes from the fact that the above equations form a system of n linear equations (for n different a). In the general case, a system of n linear equations with n unknown has a unique solution. (In principle, there are degenerate cases when a system of n linear equations with n unknowns does not have a solution or has an infinite number of different solutions, but in [14] it was proven that for posets we always have a unique solution.)

4.3 Prediction on Posets: General Approach (Reminder)

The above formula by itself does not immediately lead to a prediction algorithm since to use this formula for determining n values $v(a)$, we need to know n different values $V(b)$. However, in practice, in many cases, some of the values $V(b)$ turn out to be negligible.

If we know which values $V(b)$ are negligible, we can replace them with 0s and thus, consider a model with $m < n$ non-zero parameters $V(b)$. Once we have such an expression, we can then:

- measure the value $v(a_1), \dots, v(a_m)$ of the desired quantity v for $m < n$ different elements a_1, \dots, a_m ;
- use the resulting system of m linear equations with m unknown parameters $V(b)$ to find the values of these parameters; and then
- use the known values of the parameters $V(b)$ to predict all the remaining values $v(a)$ ($a \neq a_1, \dots, a_m$).

In practice, measurements are inevitably imprecise; to decrease the effect of the corresponding measurement errors on the predicted values, it makes sense to measure more

than m values $v(a_i)$. Then, we get more than m equations with m unknowns $V(b)$, i.e. we have an *over-determined* system of linear equations. To solve this over-determined system of linear equations, we can, e.g., use the Least Squares method (see, e.g. [16]).

4.4 Towards Chemical Applications

In our chemical examples, we start with the original molecule, we perform a few (e.g. 1 or 2) substitutions, and measure the values $v(a)$ for the resulting molecules.

The original molecule a_0 is not obtained by substitution. So, by the definition of the partial order, the only molecule b with $b \leq a_0$ is the molecule a_0 itself. Thus, for the original molecule, we have $v(a_0) = V(a_0)$. So, by measuring the value $v(a_0)$, we thus determine the value $V(a_0) = v(a_0)$ of the corresponding parameter as well.

For a molecule a which is obtained by a monosubstitution from the original molecule a_0 , we have $a_0 \leq a$ and $a \leq a$. Thus, the above expression for $v(a)$ in terms of the parameters $V(b)$ takes the form $v(a) = V(a_0) + V(a)$. Since we already know $V(a_0)$, after measuring the value $v(a)$, we determine the value $V(a)$ as $V(a) = v(a) - V(a_0)$.

For a molecule a which is obtained by two substitutions at two different places, we have $a_0 \leq a$, $a \leq a$, $a_1 \leq a$, and $a_2 \leq a$, where:

- a_1 is obtained from the original molecule by a single substitution (only) in the first place, and
- a_2 is obtained from the original molecule by a single substitution (only) in the second place.

Here, we have $v(a) = V(a_0) + V(a) + V(a_1) + V(a_2)$. We already know the value $V(a_0)$. From the analysis of single-substitution molecules, we know the value $V(a_1)$ and $V(a_2)$. Thus, we can determine the value $V(a)$ as $V(a) = v(a) - V(a_0) - V(a_1) - V(a_2)$.

At this stage, we have no information about the values $V(b)$ corresponding to more complex molecules. In general, these values can be positive or they can be negative. Since

there is no information that would enable us to prefer positive or negative values, it is reasonable, as a first approximation, to assume that the unknown values are all 0s. Under this assumption, we can use the formula $v(a) = \sum_{b \leq a} V(b)$ with b going over 0-, 1- and 2-substituted molecules, to predict all the remaining values of $v(a)$.

If for the actually synthesized molecules, these predictions turn out to be not accurate enough, it makes sense to perform additional measurements on trisubstituted molecules, find the values $V(b)$ for such molecules, and thus to attain a more accurate approximation to $v(a)$.

4.5 The Use of Symmetry

As we have mentioned, in many practical cases, natural symmetries simplify the problem. For example, for the benzene or for the cubanes, all monosubstituted molecules are equivalent and thus, they have the exact same value of the desired quantity v . In this case, the values $V(b)$ corresponding to these molecules are also equal. For example, in the above formula

$$v(a) = V(a_0) + V(a) + V(a_1) + V(a_2)$$

for the disubstituted molecule, we have $V(a_1) = V(a_2)$.

It is therefore reasonable to treat these equal values as a single parameter, and rewrite, e.g. the above formula as $v(a) = V(a_0) + 2V(a_1) + V(a)$.

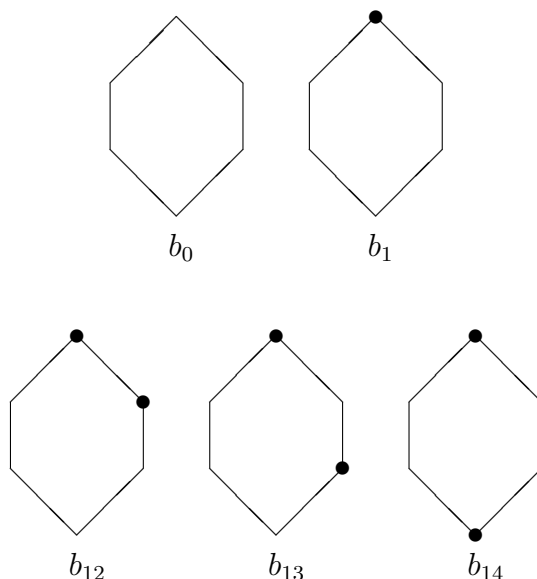
Let us illustrate this approach on the example of benzenes and cubanes.

4.6 First Example: Benzenes

In the benzenes example, to describe the value $v(a)$ of the desired characteristic v for all the molecules a , we need to find the values $V(b)$ corresponding to the following molecules:

- the original benzene molecule b_0 ,

- the monosubstituted molecule b_1 , and
- three different diatomic substitutions b_{12} , b_{13} , and b_{14} in which the distance between the substitution locations is equal to, correspondingly, 1, 2, and 3 (ortho, meta and para disubstitutions, respectively):

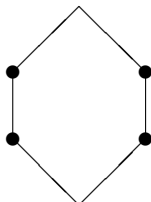


For every other molecule a , we approximate the value $v(a)$ as

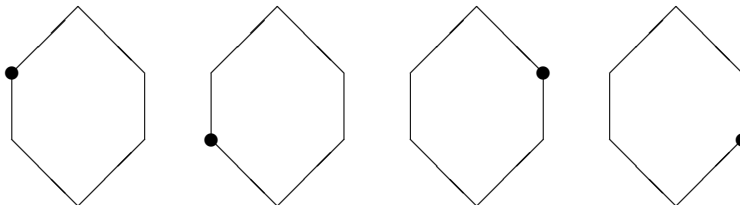
$$v(a) = \sum_{b: b \leq a} n(b) \cdot V(b),$$

where $n(b)$ is the number of different molecules b from which a can be obtained by substitution.

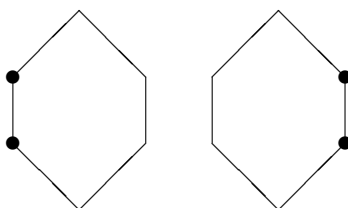
For example, for the following tetrasubstituted molecule b_4



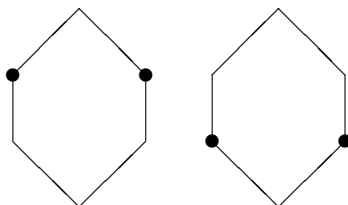
we have 1 occurrence of the original benzene b_0 , 4 occurrences of a monosubstituted molecule b_1 , corresponding to the molecules



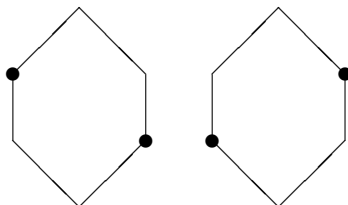
We also have 2 occurrences of the disubstituted molecule b_{12} in which the substitute locations are neighbors (at distance 1):



two occurrences of the molecule b_{13} in which the distance between the substitutions is 2:



and two occurrences of the molecule b_{14} in which the distance between the substitutions is 3:



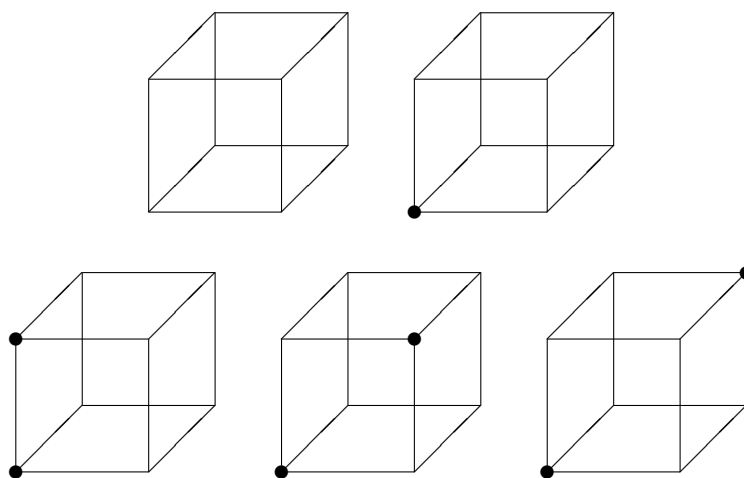
Thus, for the above molecule b_4 , we have

$$v(b_4) = V(b_0) + 4V(b_1) + 2V(b_{12}) + 2V(b_{13}) + 2V(b_{14}).$$

4.7 Second Example: Cubanes

In the cubanes example, to describe the value $v(a)$ of the desired characteristic v for all the molecules a , we need to find the values $V(c)$ corresponding to the following molecules:

- the original cubane molecule c_0 ,
- the monosubstituted molecule c_1 , and
- three different diatomic substitutions c_{12} , c_{13} , and c_{14} in which the distance between the substitution locations is equal to, correspondingly, 1, 2, and 3:

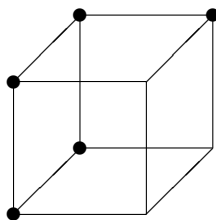


For every other molecule a , we approximate the value $v(a)$ as

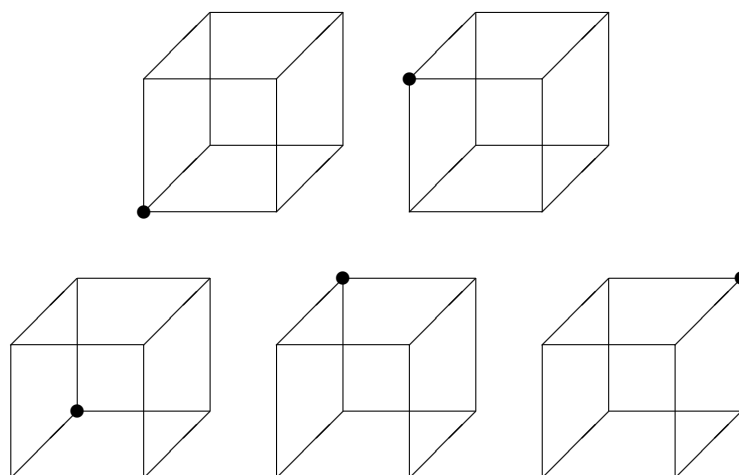
$$v(a) = \sum_{c: c \leq a} n(c) \cdot V(c),$$

where $n(c)$ is the number of different molecules c from which a can be obtained by substitution.

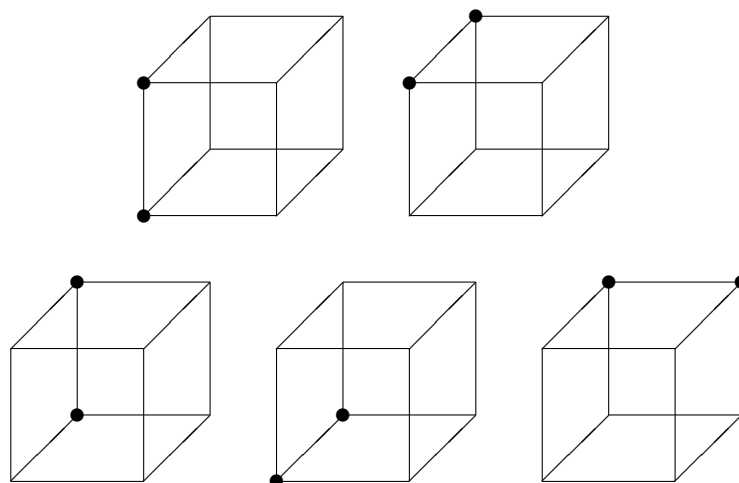
For example, for the following pentasubstituted molecule c_5



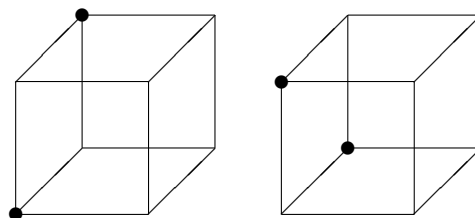
we have 1 occurrence of the original cubane c_0 , 5 occurrences of a monosubstituted molecule c_1 , corresponding to the molecules

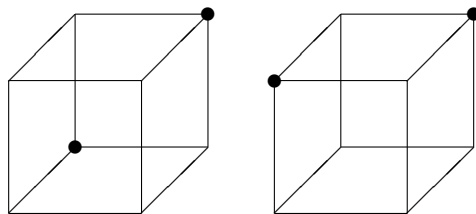


We also have 5 occurrences of the disubstituted molecule c_{12} in which the substitute locations are neighbors (at distance 1):

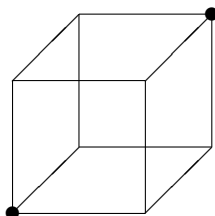


4 occurrences of the molecule c_{13} in which the distance between the substitutions is 2:





and 1 occurrence of the molecule c_{14} in which the distance between the substitutions is 3:



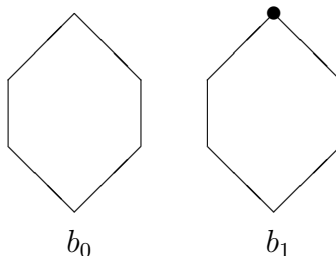
4.8 Results

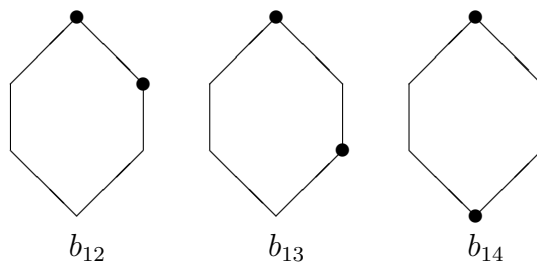
Let us illustrate the use of the above formula

$$v(a) = \sum_{c: c \leq a} n(c) \cdot V(c)$$

on the example (from [5]) of predicting the toxicity $v(a)$ of different chlorobenzenes, i.e., benzenes in which some H atoms are replaced by chlorine atoms Cl. This toxicity is measured by acute aquatic toxicity to the guppy *Poecilia reticulata*. Toxicity is measured as $-\log(LC_{50})$, where LC_{50} is a median lethal dose, i.e., a concentration that kills 50% of the tested population.

For the original benzene b_0 , we have $v(b_0) = 2.19$. For the monosubstituted molecule b_1 , we have $v(b_1) = 2.90$, for disubstituted molecules, we have $v(b_{12}) = 3.43$, $v(b_{13}) = 3.53$, and $v(b_{14}) = 3.44$.





For the original benzene $a = b_0$, no other molecule precedes it, so the above formula leads to $v(b_0) = V(b_0)$. From this equality, we conclude that $V(b_0) = v(b_0) = 2.19$.

For a monosubstituted chlorobenzene b_1 , the original benzene b_0 is the only molecule that precedes it, so $v(b_1) = V(b_0) + V(b_1)$. We know that $v(b_1) = 2.90$, and we have just estimated $V(b_0) = 2.19$. Thus, we can conclude that

$$V(b_1) = v(b_1) - V(b_0) = 2.90 - 2.19 = 0.71.$$

For a disubstituted chlorobenzene b_{12} , the original benzene b_0 and the monosubstituted chlorobenzene b_1 precede it, and b_1 occurs twice – as b_1 and b_2 . Thus, we have $v(b_{12}) = V(b_0) + 2 \cdot V(b_1) + V(b_{12})$. We know that $v(b_{12}) = 3.43$, and we have estimated that $V(b_0) = 2.19$ and $V(b_1) = 0.71$. Thus, we can conclude that

$$V(b_{12}) = v(b_{12}) - V(b_0) - 2 \cdot V(b_1) = 3.43 - 2.19 - 2 \cdot 0.71 = -0.18.$$

Similarly, we have $v(b_{13}) = V(b_0) + 2 \cdot V(b_1) + V(b_{13})$, hence

$$V(b_{13}) = v(b_{13}) - V(b_0) - 2 \cdot V(b_1) = 3.53 - 2.19 - 2 \cdot 0.71 = -0.08,$$

and $v(b_{14}) = V(b_0) + 2 \cdot V(b_1) + V(b_{14})$, hence

$$V(b_{14}) = v(b_{14}) - V(b_0) - 2 \cdot V(b_1) = 3.44 - 2.19 - 2 \cdot 0.71 = -0.17.$$

Thus, e.g., for the chlorobenzene b_{124} , which contains b_0 , b_1 three times (as b_1 , b_2 , and b_4), and each of the molecules b_{12} , b_{13} (as b_{24}) and b_{14} , we get

$$v(b_{124}) = V(b_0) + 3 \cdot V(b_1) + V(b_{12}) + V(b_{13}) + V(b_{14}).$$

Substituting the above estimates for $V(b_0)$, $V(b_1)$, $V(b_{12})$, $V(b_{13})$, and $V(b_{14})$ into this formula, we get the following prediction for the toxicity of this chlorobenzene:

$$v(b_{124}) = 2.19 + 3 \cdot 0.71 + (-0.18) + (-0.08) + (-0.17) = 3.99.$$

The observed toxicity value is 4.05 – close with the accuracy of $\approx 1\%$.

For other chlorobenzenes, we may get slightly higher prediction inaccuracy, but still the predictions are pretty reasonable.

As shown in [8], the resulting formulas lead to very good quality predictions not only of toxicity, but also of several other quantities such as energy, boiling point, vapor pressure at a certain temperature, etc.

4.9 Problems with the Poset Approach

Empirically, the poset-related cluster approaches has been very successful. However, these approaches have the following two limitations.

First, these approaches are based on mathematical notions – such as the notion of a poset – with which most users of these models are not very familiar. From this viewpoint, it is desirable to reformulate these approaches in terms of mathematical techniques which are more familiar to the users.

Second, the user's confidence in the success of each computational approach is based largely on the previous successes of this approach. As of now, the poset-related approaches have relatively few empirically successful applications, much fewer than more well-known and widely used mathematical techniques. As a result, the user's confidence in the success of these new approaches is lower than for the more traditional approaches. From this viewpoint, it would also be beneficial to show that the new approaches can be reformulated in more familiar terms – this will increase the user's confidence in these new approaches.

In this thesis, we show that poset-related approaches can be indeed reformulated in terms of the technique which is much more familiar to most users and which has a much

longer history of successful applications: the technique of Taylor series. In fact, we will show that poset-based approaches are equivalent to the approaches based on Taylor series.

Chapter 5

Discrete Taylor Series: Main Idea

5.1 Prediction: One of the Main Objectives of Science

As we have mentioned earlier, one of the main objectives of science and engineering is to predict the results of different situations. For example, in Newton's mechanics, we want to predict the positions and velocities of different objects (e.g., planets) at future moments of time.

In chemistry, there have also been many interesting predictions. Some of these predictions came from the *classification* of chemical elements and/or substances; see, e.g., [12, 15]. The classical example of such predictions are predictions made possible by Mendeleev's Periodic Law. In chemistry, there are also many examples of successful predictions of properties based on the molecular structure.

In some situations, we know the exact equations that describe the objects of interest. In such situations, we face a purely mathematical problem: to solve these equations and thus compute the value y of the desired characteristic based on the known values of the parameters x_1, \dots, x_n that describe the given objects.

5.2 Taylor Series: a Standard Tool for Solving (Continuous) Problems in Science and Engineering

In other situations, however, we do not know the exact equations (or, as is the case in many chemical problems, the equations are too difficult to be efficiently solved). In such

situation, when we do not know the *exact* equations, we can use *approximate* semi-empirical techniques. Specifically, since we do not know the exact formula for the dependence $y = f(x_1, \dots, x_n)$, we start with a (reasonably) general multi-parametric formula for such a dependence $y = f(x_1, \dots, x_n, c_1, \dots, c_m)$, and then estimate the values of the parameters c_1, \dots, c_m from the measurement results.

Traditionally, in physical and engineering applications, most parameters x_1, \dots, x_n (such as coordinates, velocity, etc.) are *continuous* – in the sense that their values can continuously change from one value to another. The dependence $y = f(x_1, \dots, x_n)$ is also usually continuous (with the exception of phase transitions); moreover, this dependence is usually smooth (differentiable). It is known that smooth functions can be usually expanded into Taylor series around some point $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_n)$ (e.g., around the point $\tilde{x} = 0$). For example, for a function of one variable, we have

$$f(x) = f(\tilde{x}) + \frac{df}{dx}(\tilde{x}) \cdot \Delta x + \frac{1}{2} \cdot \frac{d^2f}{dx^2}(\tilde{x}) \cdot \Delta x^2 + \dots,$$

where $\Delta x \stackrel{\text{def}}{=} x - \tilde{x}$.

In general, an arbitrary dependence is represented as a sum of constant terms: linear terms, quadratic terms, and terms of higher order.

$$f(x_1, \dots, x_n) = f(\tilde{x}_1, \dots, \tilde{x}_n) + \sum_{i=1}^n \frac{\partial f}{\partial x_i}(\tilde{x}) \cdot \Delta x_i + \\ \frac{1}{2} \cdot \sum_{i=1}^n \sum_{i'=1}^n \frac{\partial^2 f}{\partial x_i \partial x_{i'}}(\tilde{x}) \cdot \Delta x_i \cdot \Delta x_{i'} + \dots,$$

where $\Delta x_i \stackrel{\text{def}}{=} x_i - \tilde{x}_i$.

The values of different order terms in the Taylor expansion usually decrease when the order increases – after all, the Taylor series usually converge, which implies that the terms should tend to 0. So, in practice, we can ignore higher-order terms and consider only the first few terms in the Taylor expansion. (This is, by the way, how most elementary functions like $\sin(x)$, $\cos(x)$, $\exp(x)$ are computed inside the computers.).

In the simplest case, it is sufficient to preserve linear terms, i.e. to use the approximation

$$f(x_1, \dots, x_n) \approx f(\tilde{x}_1, \dots, \tilde{x}_n) + \sum_{i=1}^n \frac{\partial f}{\partial x_i}(\tilde{x}) \cdot \Delta x_i.$$

When the linear approximation is not accurate enough, we can use the quadratic approximation

$$f(x_1, \dots, x_n) \approx f(\tilde{x}_1, \dots, \tilde{x}_n) + \sum_{i=1}^n \frac{\partial f}{\partial x_i}(\tilde{x}) \cdot \Delta x_i + \frac{1}{2} \cdot \sum_{i=1}^n \sum_{i'=1}^n \frac{\partial^2 f}{\partial x_i \partial x_{i'}}(\tilde{x}) \cdot \Delta x_i \cdot \Delta x_{i'},$$

etc.

Since we do not know the exact expression for the function $f(x_1, \dots, x_n)$, we thus do not know the actual values of its derivatives $\frac{\partial f}{\partial x_i}(\tilde{x})$ and $\frac{\partial^2 f}{\partial x_i \partial x_{i'}}(\tilde{x})$. Hence, when we actually use this approximation, all we know is that we approximate a general function by a general linear or quadratic formula

$$f(x_1, \dots, x_n) \approx c_0 + \sum_{i=1}^n c_i \cdot \Delta x_i + \sum_{i=1}^n \sum_{i'=1}^n c_{ii'} \cdot \Delta x_i \cdot \Delta x_{i'}, \quad (5.1)$$

where $c_0 = f(\tilde{x}_1, \dots, \tilde{x}_n)$, $c_i = \frac{\partial f}{\partial x_i}(\tilde{x})$, and $c_{ii'} = \frac{1}{2} \cdot \frac{\partial^2 f}{\partial x_i \partial x_{i'}}(\tilde{x})$.

The values of the coefficients c_0 , c_i , and (if needed) $c_{ii'}$ can then be determined experimentally, by comparing the measured values of y with the predictions based on these formulas.

5.3 From Continuous to Discrete Taylor Series

How can we use a similar approach in the discrete case? The discrete case means, for example, that for each location, we are only interested in the values of the desired physical quantity in two different situations:

- a situation when there is a ligand at this location, and
- a situation when there is no ligand at this location.

In addition to these situations, we can, in principle, consider others. In our analysis, we are not interested in these additional situations. However, the general physical laws and dependencies are not limited to these two situations, they work for other situations as well.

So, while we are interested in the values of the desired physical quantity (such as energy) corresponding to the selected situations, in principle, we can consider this dependence for other situations as well. The value of, e.g., energy, depends on the values of the electronic density at different points near the ligand locations, etc. For each possible ligand location i , let $x_{i1}, \dots, x_{ij}, \dots, x_{iN}$ be parameters describing the distribution in the vicinity of this location (e.g., the density at a certain point, the distance to a certain atom, the angle between this atom and the given direction, the angle describing the direction of the spin, etc.). In general, the value of the desired quantity depends on the values of these parameters:

$$y = f(x_{11}, \dots, x_{1N}, \dots, x_{n1}, \dots, x_{nN}). \quad (5.2)$$

We are interested in the situations in which, at each location, there is either a ligand, or there is no ligand. For each location i and for each parameter x_{ij} :

- let x_{ij}^- denote the value of the j -th parameter in the situation with no ligand at the location i , and
- let x_{ij}^+ denote the value of the j -th parameter in the situation with a ligand at the location i .

The default situation with which we start is the situation in which there are no ligands at all, i.e. in which $x_{ij} = x_{ij}^-$ for all i and j . Other situations of interest are reasonably close to this one. Thus, we can expand the dependence (5.2) in Taylor series in the vicinity of the values x_{ij}^- . As a result, we obtain the following expression:

$$y = y_0 + \sum_{i=1}^n \sum_{j=1}^N y_{ij} \cdot \Delta x_{ij} + \sum_{i=1}^n \sum_{j=1}^N \sum_{i'=1}^n \sum_{j'=1}^N y_{ij,i'j'} \cdot \Delta x_{ij} \cdot \Delta x_{i'j'}, \quad (5.3)$$

where $\Delta x_{ij} \stackrel{\text{def}}{=} x_{ij} - x_{ij}^-$, and y_0 , y_{ij} , and $y_{ij,i'j'}$ are appropriate coefficients.

These formulas can be applied to all possible situations, in which at each location i , different parameters x_{i1}, \dots, x_{iN} can change independently. Situations in which we are interested are characterized by describing, for each location, whether there is a ligand or not. Let ε_i denote the discrete variable that describes the presence of the ligand at the location i :

- when there is no ligand at the location i , we take $\varepsilon_i = 0$, and
- when there is a ligand at the location i , we take $\varepsilon_i = 1$.

According to the formula (5.3), the value y of the desired physical quantity depends on the differences Δx_{ij} corresponding to different i and j . Let us describe the values of these differences in terms of the discrete variables ε_i .

- In the absence of the ligand, when $\varepsilon_i = 0$, the value of the quantity x_{ij} is equal to x_{ij}^- and thus, the difference Δx_{ij} is equal to $\Delta x_{ij} = x_{ij}^- - x_{ij}^- = 0$.
- In the presence of the ligand, when $\varepsilon_i = 1$, the value of the quantity x_{ij} is equal to x_{ij}^+ and thus, the difference $\Delta x_{ij} = x_{ij}^+ - x_{ij}^-$ is equal to

$$\Delta_{ij} \stackrel{\text{def}}{=} x_{ij}^+ - x_{ij}^-.$$

We can combine these two cases into a single expression

$$\Delta x_{ij} = \varepsilon_i \cdot \Delta_{ij}. \quad (5.4)$$

Substituting the expression (5.4) into the expression (5.3), we obtain an expression which is quadratic in ε_i :

$$y = y_0 + \sum_{i=1}^n \sum_{j=1}^N y_{ij} \cdot \varepsilon_i \cdot \Delta_{ij} + \sum_{i=1}^n \sum_{j=1}^N \sum_{i'=1}^n \sum_{j'=1}^N y_{ij,i'j'} \cdot \varepsilon_i \cdot \varepsilon_{i'} \cdot \Delta_{ij} \cdot \Delta_{i'j'}, \quad (5.5)$$

i.e., equivalently,

$$y = y_0 + \sum_{i=1}^n \left(\sum_{j=1}^N y_{ij} \cdot \Delta_{ij} \right) \cdot \varepsilon_i + \sum_{i=1}^n \sum_{i'=1}^n \left(\sum_{j=1}^N \sum_{j'=1}^N y_{ij,i'j'} \cdot \Delta_{ij} \cdot \Delta_{i'j'} \right) \cdot \varepsilon_i \cdot \varepsilon_{i'}. \quad (5.6)$$

Combining terms proportional to each variable ε_i and to each product $\varepsilon_i \cdot \varepsilon_{i'}$, we obtain the expression

$$y = a_0 + \sum_{i=1}^n a_i \cdot \varepsilon_i + \sum_{i=1}^n \sum_{i'=1}^n a_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'}, \quad (5.7)$$

where

$$a_i = \sum_{j=1}^N y_{ij} \cdot \Delta_{ij}, \quad (5.8)$$

and

$$a_{ii'} = \sum_{j=1}^N \sum_{j'=1}^N y_{ij,i'j'} \cdot \Delta_{ij} \cdot \Delta_{i'j'}. \quad (5.9)$$

The expression (5.7) is similar to the continuous Taylor expression (5.1), but with the discrete variables $\varepsilon_i \in \{0, 1\}$ instead of the continuous variables Δx_i .

Similar “discrete Taylor series” can be derived if we take into account cubic, quartic, etc., terms in the original Taylor expansion of the dependence (5.2).

5.4 Discrete Taylor Expansions Can Be Further Simplified

In the following text, we will use the fact that the expression (5.7) can be further simplified.

First, we can simplify the terms corresponding to $i = i'$. Indeed, for each discrete variable $\varepsilon_i \in \{0, 1\}$, we have $\varepsilon_i^2 = \varepsilon_i$. Thus, the term $a_{ii} \cdot \varepsilon_i \cdot \varepsilon_i$ corresponding to $i = i'$ is equal to $a_{ii} \cdot \varepsilon_i$ and can, therefore, be simply added to the corresponding linear term $a_i \cdot \varepsilon_i$. As a result, we arrive at the following simplified version of the discrete Taylor expansion:

$$y = c_0 + \sum_{i=1}^n c_i \cdot \varepsilon_i + \sum_{i \neq i'} c_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'}, \quad (5.10)$$

where $c_0 = a_0$, $c_{ii'} = a_{ii'}$, and $c_i = a_i + a_{ii}$.

Second, we can combine terms proportional to $\varepsilon_i \cdot \varepsilon_{i'}$ and to $\varepsilon_{i'} \cdot \varepsilon_i$. As a result, we obtain a further simplified expression

$$y = v_0 + \sum_{i=1}^n v_i \cdot \varepsilon_i + \sum_{i < i'} v_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'}, \quad (5.11)$$

where $v_0 = c_0$ and $v_{ii'} = c_{ii'} + c_{i'i}$.

This expression (5.11) – and the corresponding similar cubic and higher order expressions – is what we will call a discrete Taylor series.

We now proceed to show that the poset-related approaches are, in effect, equivalent to the use of a much simpler (and much more familiar) tool of (discrete) Taylor series.

Chapter 6

Equivalence Between the Poset-Related Approaches and the Discrete Taylor Series Approach

6.1 Discrete Taylor Series: Reminder

In many practical situations, we have a physical variable y that depends on the discrete parameters ε_i which take two possible values: 0 and 1. Then, in the first approximation, the dependence of y on ε_i can be described by the following linear formula

$$y = v_0 + \sum_{i=1}^n v_i \cdot \varepsilon_i. \quad (6.1)$$

In the second approximation, this dependence can be described by the following quadratic formula

$$y = v_0 + \sum_{i=1}^n v_i \cdot \varepsilon_i + \sum_{i < i'} v_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'}, \quad (6.2)$$

etc.

6.2 Chemical Substances: Application Area

For chemical substances like benzenes and cubanes, we have discrete variables ε_i that describe whether there is a ligand at the i -th location:

- the value $\varepsilon_i = 0$ means that there is no ligand at the i -th location, and

- the value $\varepsilon_i = 1$ means that there is a ligand at the i -th location.

Each chemical substance a from the corresponding family can be characterized by the corresponding tuple $(\varepsilon_1, \dots, \varepsilon_n)$.

6.3 Poset-Related Approaches: Reminder

We approximate the actual dependence of the desired quantity y on the substance $a = (\varepsilon_1, \dots, \varepsilon_n)$ by a formula

$$v(a) = \sum_{b: b \leq a} V(b), \quad (6.3)$$

where, in the second order approximation, b runs over all substances with at most two ligands.

6.4 Poset-Related Approaches Reformulated in Terms of the Discrete Variables

The discrete Taylor series formula (6.2) is formulated in terms of the discrete variables ε_i . Thus, to show the equivalence of these two approaches, let us first describe the poset-related formula (6.3) in terms of these discrete variables.

In chemical terms, the relation $b \leq a$ means that a can be obtained from b by adding some ligands. In other words, the corresponding value ε_i can only increase when we move from the substance b to the substance a . So, if $b = (\varepsilon'_1, \dots, \varepsilon'_n)$ and $a = (\varepsilon_1, \dots, \varepsilon_n)$, then $b \leq a$ means that for every i , we have $\varepsilon'_i \leq \varepsilon_i$.

Thus, the formula (6.3) means that for every substance $a = (\varepsilon_1, \dots, \varepsilon_n)$, the substances $b \leq a$ are:

- the original substance $a_0 = (0, \dots, 0)$;
- substances $a_i \stackrel{\text{def}}{=} (0, \dots, 0, 1, 0, \dots, 0)$ with a single ligand at the location i – corresponding to all the places i for which $\varepsilon_i = 1$; and

- substances $a_{ii'} \stackrel{\text{def}}{=} (0, \dots, 0, 1, 0, \dots, 0, 1, 0, \dots, 0)$ with two ligands at the locations i and i' – corresponding to all possible pairs of places $i < i'$ at which $\varepsilon_i = \varepsilon_{i'} = 1$.

Thus, in terms of the discrete variables, the poset formula (6.3) takes the form

$$y = V(a_0) + \sum_{i: \varepsilon_i=1} V(a_i) + \sum_{i < i': \varepsilon_i=\varepsilon_{i'}=1} V(a_{i,i'}). \quad (6.4)$$

6.5 Proof That the Discrete Taylor Series Are Indeed Equivalent to the Poset Formula

The formulas (6.2) and (6.4) are now very similar, so we are ready to prove that they actually coincide.

To show that these formulas are equal, let us take into account that, e.g. the linear part of the sum (6.4) can be represented as

$$\sum_{i: \varepsilon_i=1} V(a_i) = \sum_{i: \varepsilon_i=1} V(a_i) \cdot \varepsilon_i. \quad (6.5)$$

Indeed, for all the corresponding values i , we have $\varepsilon_i = 1$, and multiplying by 1 does not change a number.

This new representation (6.5) allows us to simplify this formula by adding similar terms $V(a_i) \cdot \varepsilon_i$ corresponding to indices i for which $\varepsilon_i = 0$. Indeed, when $\varepsilon_i = 0$, then the product $V(a_i) \cdot \varepsilon_i$ is equal to 0, and thus, adding this product will not change the value of the sum. So, in the right-hand side of the formula (6.5), we can safely replace the sum over all i for which $\varepsilon_i = 1$ by the sum over all indices i from 1 to n :

$$\sum_{i: \varepsilon_i=1} V(a_i) = \sum_{i=1}^n V(a_i) \cdot \varepsilon_i. \quad (6.6)$$

Similarly, the quadratic part $\sum_{i < i': \varepsilon_i=\varepsilon_{i'}=1} V(a_{i,i'})$ of the sum (6.4) can be first replaced with the sum

$$\sum_{i < i': \varepsilon_i=\varepsilon_{i'}=1} V(a_{i,i'}) = \sum_{i < i': \varepsilon_i=\varepsilon_{i'}=1} V(a_{i,i'}) \cdot \varepsilon_i \cdot \varepsilon_{i'}, \quad (6.7)$$

and then, by the sum

$$\sum_{i < i': \varepsilon_i = \varepsilon_{i'} = 1} V(a_{i,i'}) = \sum_{i < i'} V(a_{i,i'}) \cdot \varepsilon_i \cdot \varepsilon_{i'}. \quad (6.8)$$

Substituting expressions (6.5) and (6.8) into the formula (6.4), we obtain the following expression

$$y = V(a_0) + \sum_{i=1}^n V(a_i) \cdot \varepsilon_i + \sum_{i < i'} V(a_{i,i'}) \cdot \varepsilon_i \cdot \varepsilon_{i'}. \quad (6.9)$$

This expression is identical to the discrete Taylor formula (6.2), the only difference is the names of the corresponding parameters:

- the parameter v_0 in the formula (6.2) corresponds to the parameter $V(a_0)$ in the formula (6.9);
- each parameter v_i in the formula (6.2) corresponds to the parameter $V(a_i)$ in the formula (6.9); and
- each parameter $v_{ii'}$ in the formula (6.2) corresponds to the parameter $V(a_{i,i'})$ in the formula (6.9).

The equivalence is proven.

6.6 Important Observation: The Presence of Symmetry Does Not Change the Equivalence

General idea. As we have mentioned, symmetry simply means that some of the coefficients v_i and $v_{ii'}$ coincide. For example, for benzenes and cubanes, symmetry means that $v_1 = v_2 = \dots = v_i = \dots$, and that the value $v_{ii'}$ depends only on the distance between the locations i and i' .

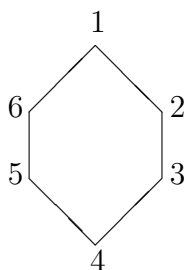
Example: benzene. For benzene, as we have mentioned,

- the values v_i are all equal to each other, and
- the values $v_{ii'}$ and $v_{i'i}$ depend only on the distance d between the locations i and i' .

To capture this symmetry, let us use the following denotations:

- by V , we denote the common values of v_i ; and
- by V_d , we denote the common value of $v_{ii'}$ and $v_{i'i}$ when the distance between the locations i and i' is equal to d .

Let us number the locations in a sequential order:



In these notations, the general quadratic formula (6.2) takes the form

$$\begin{aligned}
 y = v_0 + V \cdot \left(\sum_{i=1}^n \varepsilon_i \right) + \\
 V_1 \cdot (\varepsilon_1 \cdot \varepsilon_2 + \varepsilon_2 \cdot \varepsilon_3 + \varepsilon_3 \cdot \varepsilon_4 + \varepsilon_4 \cdot \varepsilon_5 + \varepsilon_5 \cdot \varepsilon_6 + \varepsilon_6 \cdot \varepsilon_1) + \\
 V_2 \cdot (\varepsilon_1 \cdot \varepsilon_3 + \varepsilon_2 \cdot \varepsilon_4 + \varepsilon_3 \cdot \varepsilon_5 + \varepsilon_4 \cdot \varepsilon_6 + \varepsilon_5 \cdot \varepsilon_1 + \varepsilon_6 \cdot \varepsilon_2) + \\
 V_3 \cdot (\varepsilon_1 \cdot \varepsilon_4 + \varepsilon_2 \cdot \varepsilon_5 + \varepsilon_3 \cdot \varepsilon_6). \tag{6.10}
 \end{aligned}$$

In other words, we have

$$y = v_0 + V \cdot N + \sum_{d=1}^3 V_d \cdot N_d, \tag{6.11}$$

where N is the total number of ligands, and N_d is the total number of pairs (i, i') of ligands which are located at a distance d to each other.

Comment. The same formula (6.11) holds for cubanes as well.

6.7 Discussion: Advantages of the Taylor Representation

6.7.1 Main Advantage of the Taylor Representation

In our opinion, the main advantage of the Taylor series representation is that the Taylor series is a more familiar technique for a wide range of scientists.

As a result of this familiarity, Taylor series have a much larger number of successful applications than the poset-related methods; therefore, scientists are more confident in Taylor series techniques.

6.7.2 Additional Advantage of the Taylor Representation: Taylor Series Can Clarify the Equivalence of Different Arrangements

In addition to the above main advantage, the Taylor series representation also has an additional advantage: this representation makes it easier to check whether different arrangements lead to the exact same results.

For example, in the poset formulation, it is natural to consider, instead of the original order $b \leq a$, the *dual* order $b \leq' a$ which is defined as $a \leq b$. In chemical terms, the original order $a \leq b$ means that we can obtain the substance b from the substance a by *adding* ligands at different locations. Correspondingly, the dual order $b \leq' a$ means that we can obtain the substance b from the substance a by *removing* ligands at different locations.

In the original order \leq , the minimal element is the original substance a_0 (with no ligands added), and the second order poset approximation means that we use the values $V(b)$ corresponding to the substances with 0, 1, and 2 ligands. In the dual order \leq' , the

minimal element is the substance with the ligands in all the places, and the second order poset approximation means that we use the values $V(b)$ corresponding to the substances with 0, 1, and 2 missing ligands. Will this new order lead to a different approximation?

In the poset formulation, it is difficult to immediately answer this question: the two orders are different, so at first glance, it may look like the resulting approximations are different too. This actually was the conclusion that the authors of [13] originally made.

However, if we reformulate this question in terms of the discrete Taylor series, we almost immediately obtain an answer: yes, the resulting approximation is exactly the same for the new order. Indeed, in terms of the discrete variables, the new order simply means that we have a new starting point for the Taylor expansion: instead of the substance with no ligands, we now have, as a starting point, the substance with all the ligands present. Thus, the discrete variables ε'_i that describe the difference between the current substance and the starting point must also change:

- $\varepsilon'_i = 0$ (no change) if there is a ligand at the i -th location, and
- $\varepsilon'_i = 1$ (change) if there is no ligand at the i -th location.

One can easily see that the relation between the new and the old variables is simple: $\varepsilon'_i = 1 - \varepsilon_i$, or, equivalently, $\varepsilon_i = 1 - \varepsilon'_i$.

Discrete Taylor series in terms of the new variables means that we are approximating the dependence of y on the variables ε'_i by a quadratic formula. If we substitute the values $\varepsilon'_i = 1 - \varepsilon_i$ into this quadratic formula, we obtain an expression which is quadratic in ε_i . Vice versa, if we start with the original discrete Taylor series expression which is quadratic in ε_i and substitute the values $\varepsilon_i = 1 - \varepsilon'_i$ into this quadratic formula, we obtain an expression which is quadratic in ε'_i . Thus, in both cases, we approximate the original dependence by a function which is quadratic in ε_i . So, if we use the Least Squares method, we get the exact same best approximation in both cases.

This conclusion helped us find an arithmetic mistake in the computations presented in [13] and conclude that these two orders indeed lead to similar results.

6.7.3 Additional Advantage: a Detailed Description

Let us describe the above argument related to the additional advantage in detail. If we substitute the expression $\varepsilon_i = 1 - \varepsilon'_i$ into the general quadratic formula

$$y = v_0 + \sum_{i=1}^n v_i \cdot \varepsilon_i + \sum_{i < i'} v_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'},$$

we obtain the formula

$$y = v_0 + \sum_{i=1}^n v_i \cdot (1 - \varepsilon'_i) + \sum_{i < i'} v_{ii'} \cdot (1 - \varepsilon_i) \cdot (1 - \varepsilon_{i'}).$$

Opening the parentheses, we conclude that

$$y = v_0 + \sum_{i=1}^n v_i - \sum_{i=1}^n v_i \varepsilon'_i + \sum_{i < i'} v_{ii'} - \sum_{i < i'} v_{ii'} \cdot \varepsilon_i - \sum_{i < i'} v_{ii'} \cdot \varepsilon_{i'} + \sum_{i < i'} v_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'}.$$

Combining terms of different order in terms of ε'_i , we conclude that

$$y = v'_0 + \sum_{i=1}^n v'_i \cdot \varepsilon'_i + \sum_{i < i'} v'_{ii'} \cdot \varepsilon'_i \cdot \varepsilon'_{i'},$$

where

$$\begin{aligned} v'_0 &= v_0 + \sum_{i=1}^n v_i + \sum_{i < i'} v_{ii'}, \\ v'_i &= -v_i - \sum_{i': i < i'} v_{ii'} - \sum_{i': i' < i} v_{i'i}, \end{aligned}$$

and $v'_{ii'} = v_{ii'}$.

Similarly, if we have a representation

$$y = v'_0 + \sum_{i=1}^n v'_i \cdot \varepsilon'_i + \sum_{i < i'} v'_{ii'} \cdot \varepsilon'_i \cdot \varepsilon'_{i'},$$

we can substitute $\varepsilon'_i = 1 - \varepsilon_i$ and obtain a quadratic expression

$$y = v_0 + \sum_{i=1}^n v_i \cdot \varepsilon_i + \sum_{i < i'} v_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'},$$

where

$$\begin{aligned} v_0 &= v'_0 + \sum_{i=1}^n v'_i + \sum_{i < i'} v'_{ii'}, \\ v_i &= -v'_i - \sum_{i': i < i'} v'_{ii'} - \sum_{i': i' < i} v'_{i'i}, \end{aligned}$$

and $v_{ii'} = v'_{ii'}$.

6.7.4 Example

In the benzene example, when we have the formula (6.10), the above formula relating v'_i and v_i takes the following form:

$$v'_0 = v_0 + 6V + 6V_1 + 6V_2 + 3V_3,$$

$$V' = -V - 2V_1 - 2V_2 - V_3,$$

$$V'_1 = V_1, \quad V'_2 = V_2, \quad V'_3 = V_3,$$

and, correspondingly,

$$v_0 = v'_0 + 6V' + 6V'_1 + 6V'_2 + 3V'_3,$$

$$V = -V' - 2V'_1 - 2V'_2 - V'_3,$$

$$V_1 = V'_1, \quad V_2 = V'_2, \quad V_3 = V'_3.$$

Comment. From the computational viewpoint, instead of computing the values v' , V'' , and V'_i in their natural order, it is faster to first compute V' . Then we can compute v'_0 by using a simpler (and thus, faster-to-compute) formula $v'_0 = v_0 + 3V - 3V'$.

Chapter 7

Conclusion

One of the main objectives of science and engineering is to predict the results of different situations. In mathematical terms, this means that we need to extrapolate or interpolate known experimental results to new situations.

In many cases, this extrapolation is performed by using heuristic algorithms, e.g., algorithms that use the ideas of a prominent MIT mathematician Gian-Carlo Rota related to partially ordered sets (posets). These algorithms are empirically successful, but since they do not have any domain-related theoretical justification, there is no guarantee that they will work in other situations as well.

It is therefore desirable to provide more reliable extrapolation algorithms – e.g., by providing a solid justification for the existing heuristic techniques.

In this thesis, we provide a theoretical justification for an important class of heuristic extrapolation algorithms – algorithms based on partially ordered sets (posets). This justification make these algorithms more reliable.

As an application area, we take a practically important problem from organic chemistry – predicting the properties of new chemical substances. Several practically useful chemical substances can be obtained by adding ligands to different locations of a “template” molecule like benzene C_6H_6 or cubane C_8H_8 . There is a large number of such substances, and it is difficult to synthesize all of them and experimentally determine their properties. It is desirable to be able to synthesize and test only a few of these substances and to use appropriate interpolation to predict the properties of others.

Such an interpolation has been obtained by using Rota’s ideas related to partially ordered sets. In this thesis, we show that the exact same interpolation algorithm can be

obtained from a more familiar mathematical technique such as Taylor expansion series. This makes the chemical prediction results more reliable.

References

- [1] T. Došlić and D. J. Klein, “Splinoïd interpolation on finite posets”, *Journal of Computational and Applied Mathematics*, 2005, Vol. 177, pp. 175–185.
- [2] P. E. Eaton and T. W. Cole, “Cubane”, *Journal of the American Chemical Society*, 1964, Vol. 86, No. 15, pp. 3157–3158.
- [3] E. G. Gillan and A. R. Barron, “Chemical vapor deposition of hexagonal Gallium Selenide and Telluride films from cubane precursors: Understanding the envelope of molecular control”, *Chemistry of Materials*, 1997, Vol. 9, No. 12, pp. 3037–3048.
- [4] H. Gunosewoyo, J. L. Guo, M. R. Bennett, M. J. Coster and M. Kassiou, “Cubyl amides: Novel P2X7 receptor antagonists”, *Bioorganic & Medicinal Chemistry Letters*, 2008, Vol. 18, No. 13, pp. 3720–3723.
- [5] T. Ivanciuc, O. Ivanciuc and D. J. Klein, “Posetic quantitative superstructure/activity relationships (QSSARs) for chlorobenzenes” *Journal of Chemical Information and Modeling*, 2005, Vol. 45, pp. 870–879.
- [6] T. Ivanciuc, O. Ivanciuc and D. J. Klein, “Modeling the bioconcentration factors and bioaccumulation factors of polychlorinated biphenyls with posetic quantitative super-structure/activity relationships (QSSAR)”, *Molecular Diversity*, 2006, Vol. 10, pp. 133–145.
- [7] T. Ivanciuc and D. J. Klein, “Parameter-free structure-property correlation via progressive reaction posets for substituted benzenes”, *Journal of Chemical Information and Computer Sciences*, 2004, Vol. 44, No. 2, pp. 610–617.
- [8] T. Ivanciuc, D. J. Klein, and O. Ivanciuc, “Posetic cluster expansion for substitution-

- reaction networks and application to methylated cyclobutanes”, *Journal of Mathematical Chemistry*, 2007, Vol. 41, No. 4, pp. 355–379.
- [9] D. J. Klein, “Chemical graph-theoretic cluster expansions” *International Journal of Quantum Chemistry, Quantum Chemistry Symposium*, 1986, Vol. 20, pp. 153–171.
- [10] D. J. Klein and L. Bytautas, “Directed reaction graphs as posets”, *MATCH Communications in Mathematical and in Computer Chemistry (MATCH)*, 2000, Vol. 42, pp. 261–290.
- [11] D. K. Manley, A. McIlroy, and C. A. Taatjes, “Research needed for future internal combustion engines”, *Physics Today*, 2008, Vol. 61, No. 11, pp. 47–52.
- [12] G. Restrepo and L. Pachón, “Mathematical aspects of the periodic law”, *Foundations of Chemistry*, 2007, Vol. 9, pp. 189–214.
- [13] G. Restrepo, B. Vijaikumar, and D. J. Klein, “Forward and reverse posetic cluster expansions of methyl cubanes”, *Abstracts of the 24th Southwest Theoretical Chemistry Conference SWTCC’08*, El Paso, Texas, October 10–12, 2008.
- [14] G.-C. Rota, “On the foundations of combinatorial theory I. Theory of Möbius functions”, *Zeit. Wahrscheinlichkeitstheorie*, 1964, Vol. 2, pp. 340–368.
- [15] J. Schummer, “The chemical core of chemistry I: A conceptual approach”, *HYLE—International Journal for Philosophy of Chemistry*, 1998, Vol. 4, No. 2, pp. 129–162.
- [16] D. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2004.
- [17] D. W. Van Krevelen and K. te Nijenhuis, *Properties of Polymers: Their Correlation with Chemical Structure; their Numerical Estimation and Prediction from Additive Group Contributions*, Elsevier, 2009.

Curriculum Vitae

Jaime Eduardo Nava Carrillo was born on July 28, 1984. The first son of Jaime Eduardo Nava Martínez and Rosa María Nava Carrillo, he earned his high school diploma from Instituto Tecnológico y de Estudios Superiores de Monterrey, campus Ciudad Juárez, Chihuahua, México, in the spring of 2002. He entered The University of Texas at El Paso in the fall of 2002. While pursuing his bachelor's degree in Computer Science he worked as a Research Assistant, and as a software intern at International Business Machines Corporation (IBM) in Austin, Texas. He received his bachelor's degree in Computer Science in the spring of 2007.

In the fall of 2007, he entered the Graduate School of The University of Texas at El Paso. While pursuing a master's degree in Computer Science he worked as a Teaching Assistant, and as a software intern for three consecutive summers at IBM Austin. He is a member of the El Paso Student Chapter of the Association for Computing Machinery (ACM).

Permanent address: 1709 Shreya Street

El Paso, Texas 79928