What Is a Structural Representation in Chemistry: Towards a Unified Framework for CADD?

Lev Goldfarb,* Oleg Golubitsky, Dmitry Korkin

Faculty of Computer Science
University of New Brunswick, P.O Box 4400
Fredericton, N.B., Canada

March 10, 2001

Abstract

A *fundamentally* new formal framework for structural representation of organic compounds based on the first "true" (general) formalism for structural object representation recently proposed by us—evolving transformation system (ETS) model—is outlined. The applied orientation of the paper is towards the molecular design in general and computer aided drug design (CADD) in particular. Inadequacies of the conventional models used in (CADD) for molecular representation and classification as well as the advantages of the proposed ETS model are discussed. Some advantages of the ETS model is its capability to represent *naturally* all important structural features of molecules, e.g. different atoms and their bonding types (including hydrogen bonding), basic 2D and 3D isometries, the molecular class structure. The model allows one not only to classify a new compound, but also to construct a chemically valid new compound from the class of compounds that was previously learned based on a small set of examples. The model also guarantees the inheritance of the chemical structural class information from the parent class to all its subclasses. In general, the ETS model offers a much more precise "language" for *chemical structural formulas*. The central role of the class learning problem in CADD is suggested. Moreover, we propose the ETS model as a unified framework for the class learning problem and therefore as a unified formal framework for CADD. This would allow considerable streamlining of the CADD by assigning to the chemist the role of an interactive user of the system rather that a role of a human weak link within the CADD process.

*To whom correspondence should be addressed (tel: 506-458-7271, e-mail: goldfarb@unb.ca)

1 Introduction

In spite of the particular importance and popularity of the terms "structure" and "structural" in computer aided drug design (CADD) and chemistry in general, there are presently no *formal models* that clarify the common core of various implicit understandings of the terms. The main reason behind this situation is related to a basic historical fact that applied mathematical modeling has so far relied on the numeric as opposed to the "structural" forms of representation: by a number of reasons [1] and in spite of a widespread misunderstanding, various *sets* of combinatorial objects, e.g. strings, trees, and graphs, cannot be considered as adequate forms of structural representation (see also section 3.2).

It appears that the main driving motivation for the development of various forms of structural representation came from the fields of pattern recognition and artificial intelligence, but mainly from the former. Historically, it turned out that the need to automate various pattern recognition processes in the 1960s and 1970s exposed the fundamental inadequacies of the existing forms of representation, including the vector space based models, the symbolic, and the syntactic models [2] - [5].

In this paper, with applied focus on CADD, we outline the recently proposed model for structural representation — evolving transformation system (ETS) model — and its potential impact on CADD. Since the formal model itself was just completed [1], it should not be surprising that, at this stage, we did not focus on various "implementational" issues. Moreover, the *initial* focus on such issues is not only premature but is also inappropriate: the radical rethinking of the scientific modeling necessitated by the introduction of the first structural model of representation cannot begin properly with hastily produced experimental results.

As in the last sentence, throughout the paper, we could not help using the adjectives "radical" or "fundamental" to emphasize the necessity [1] to start from scratch when attempting to formalize the concept of structural representation.

It is interesting to note that even at the outset the ETS model incorporates a very important new, "predictive" and "explanatory", features that help one to evaluate its potential scientific utility: in contrast to the numeric models, it offers a new form of *"molecular" class description* that, in particular, allows us both to predict and explain the structure (and to construct the elements) of a molecular class. What is of particular importance is that such a class description can now be constructed based on a small set of the class examples.

The paper is organized as follows. In section 2, we discuss the concept of a molecular class and its role in CADD. In section 3, we discuss the concept of structural representation in chemistry and in CADD in particular, including the fundamental inadequacies of current models as well as some *desirable* features of a satisfactory model for structural representation. Sections 4 and 5 form the technical core of the paper. In section 4, we present an outline of the ETS model (abridged from [1] and geared towards chemists) and in section 5, we present some introductory examples illustrating the very basic ideas of section 4 and suggesting the fundamental links with chemistry. Advantages of the proposed model are very briefly outlined in section 6, followed by the concluding section 7.

Finally, we note that the size limitations necessitated, perhaps, a more condensed than desirable style of the paper.

[1] At the same time, one should keep in mind that "necessity is the mother of invention".

2 Molecular classes and the basic problems in CADD

2.1 What is a molecular class?

We propose to view the concept of molecular class as the central to the scientific development of biomolecular and chemical sciences in general and drug design in particular. Why? **First** of all, we recall that classification is at the very heart of *biology*: "Laws in biology concern classes of entities such as taxonomic categories but predictive generalizations about individuals are taxonomic statements" [6, p.9]. **Second**, as far as *organic chemistry* (and chemistry in general) is concerned, the role of classes, although not as widely observed, is as central as in biology: "Because the compounds of carbon are so numerous, it is convenient (and important) to organize them into families that exhibit structural similarities" [7, p.957]. **Third**, *molecular biology* and *biochemistry* are absolutely unthinkable without the concept of protein, RNA, DNA, and other evolutionary *classes* of biomolecules. **Fourth**, the *pharmaceutics* itself cannot at present be properly understood without the classification of drugs based on their biological activities and their therapeutic categories.
So, how should one approach the question of what constitutes a molecular class and a class of drugs in particular? It is quite obvious that not every set of molecules constitutes a class and that the elements of the class are closely related structurally. But beyond this, as history of science amply testifies, the concept proved to be quite elusive. In [1] (see also section 4.2) as a result of 25 years of research work directly related to the clarification of the concept of structural class similarity [2]-[4], [8]-[10], we have proposed the first general model for structural representation and classification that suggests, in particular, how to understand the concept of molecular structural relatedness. According to it, intuitively, one can think of elements of a molecular class as "constructed" by "applying" to a fixed structural class "progenitor" specific to this class structural "fragments". In other words, one can view the elements of a molecular class as obtained by attaching class-specific molecular "building blocks" to a class-specific generalized pharmacophore (Figs. 1,2). It is useful to note that one of the main motivations for such a view of the class came from the accumulated knowledge about various evolutionary processes in the Universe in general and in biology (including molecular biology, e.g. evolutionary protein trees) in particular. The proposed formal model, discussed in this paper, among several other important features, allows one, on the one hand, *to construct every element in the class*, and, on the other hand, *to decide if a given molecule belongs to the class or not*. In this connection, it is important to note that the current classification models used in CADD do not possess the first of the above two features, which, in turn, necessitates exhaustive searches and screenings of combinatorial libraries.

2.2 The central role of the concept of molecular class in CADD

In this section, using the classification framework, we briefly formulate some problems related to a number of basic tasks mainly those around the structure-based ligand CADD.

Drug design. The central (ideal) CADD class learning problem can be formulated as follows. Given a (small) set of drugs with therapeutic effect **A**, on the basis of this set present a description of the class of drugs with the therapeutic effect **A** but without side effect **B**. To formulate this problem in a more formal language, let us denote by $\mathbb{A}$ the class of drugs with therapeutic effect **A**, denote by $\mathbb{B}$, the set of compounds similar to the elements of $\mathbb{A}$ *and* with side effect **B**, and denote by $\mathbb{D}$, the subclass of $\mathbb{A}$ without side effect **B**.
Problem 1: Given a small set of drugs from $\mathbb{D}$ and a small set from $\mathbb{B}$, construct a representation of class $\mathbb{D}$.

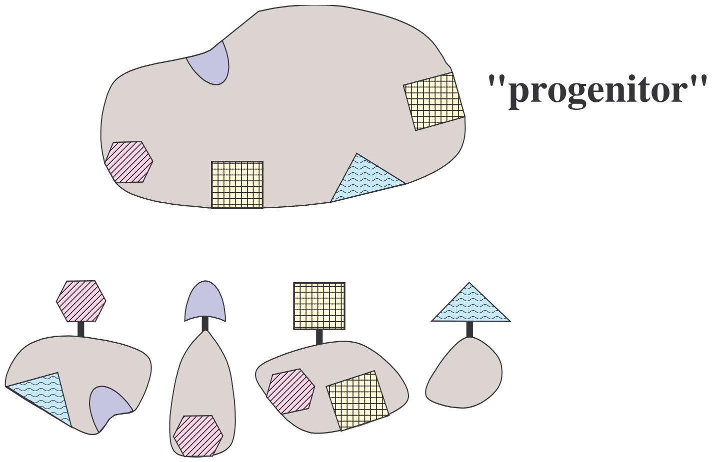


Figure 1: Class “progenitor” and the class “building blocks”.

With the solution of the last problem, the effective solution of the following two tasks also become possible:

(i) construct drug(s) that are the most typical elements in the above class $\mathbb{D}$ (*de novo* drug design)

(ii) find compound(s) from a combinatorial library that are most similar to the elements in class $\mathbb{D}$.

Construction of a focused virtual library of compounds.
Problem 2: Given a (small) set of compounds of the same biological activity, construct as many as possible other compounds of the same activity. In other words, specify a constructive process for the compounds of the class based on a small set of examples.

Virtual screening. Here we simply want to mention that most of the virtual screening tasks will be obviated with the development of algorithms solving the central class learning problem, Problem 1.

On the basis of the above, it is not difficult to see why the claim stated in the heading of the section is justified.

3 What is structural representation in chemistry?

In spite of the central role of the concept of molecular class in drug design and in chemistry in general, there are currently no adequate or reliable connections between a formal representation of a particular molecule and the formal description of the molecular class to which the molecule belongs. We suggest that the very concept of structural representation is directly related to this issue. Thus, to properly address the concept of representation we must face these questions *simultaneously*. It goes without saying that a truly efficient automation of drug discovery process (CADD) is unthinkable without a reliable formal model for molecular representation, the lack of which is one of the main impediments to the acceleration, and therefore to a substantial cost reduction, of the drug discovery process. We next briefly discuss this situation as it relates to the dominant forms of molecular representation in computer aided drug design.

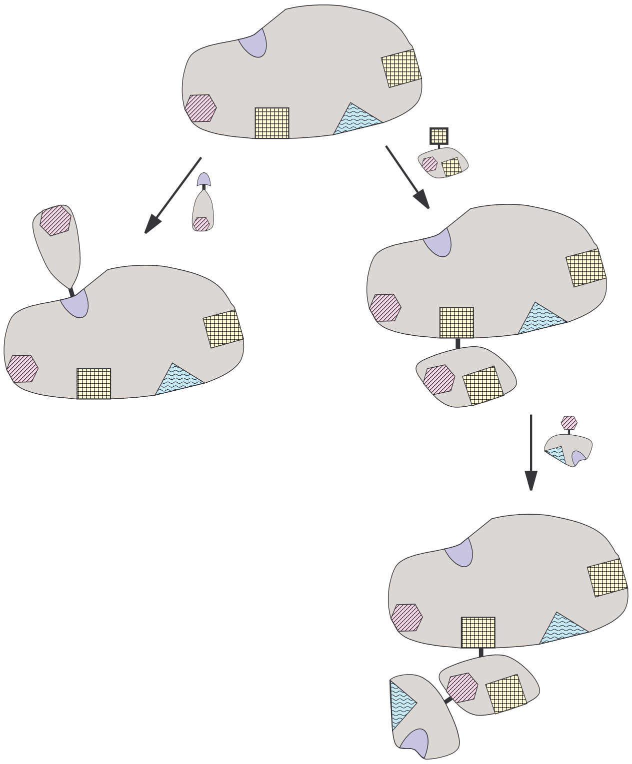


Figure 2: Several class objects constructed from the progenitor using the class "building blocks".

3.1 Three current basic forms of molecular representation: molecular graphs, vector (descriptor-based) representation, and 3D molecular structure

In this section we summarize the basic computational forms of representation for small organic compounds. All of them are used extensively in CADD, in computational and combinatorial chemistry. Moreover, there are many hybrid forms of representation combining together some of these basic forms. Each of the basic forms relies on the conventional mathematical formalisms and captures only "one side of the story" as it relates to the molecular representation.
As is the case in all applied sciences, in drug design, the **vector** form is overwhelmingly the most popular form of representation. The obvious reason for this state of affairs is that historically applied mathematics and physics have used only this form of representation. In CADD, under this form of representation, each vector coordinate corresponds to a numeric descriptor, which is typically either a physicochemical descriptor (e.g. parameter characterizing hydrophobicity, electronic properties, steric effects) or "structural" descriptor (see the next paragraph). Among the methods relying on vector representation are such well-known techniques as quantitative structure-activity relationship (QSAR) [11, pp.569-582], [12] - [19] and artificial neural networks (ANN) [20] - [27]. These methods draw their strength from the advanced mathematical state of the vector-space model in general and in particular from such tools as linear regression, discriminant analysis, and various statistical and non-statistical techniques for classification.
A **molecular graph** is a graph with labeled vertices (atoms) and labeled edges (bond types).

In this case, the form of the representation is "borrowed" from graph theory and is the first form of "structural" representation in chemistry [28, p.51]. Historically, in mathematics, the emergence of this form of representation has not been motivated by a formalization of chemical/biochemical representation. This partly explains why this form of molecular representation is not used in CADD directly and why "it is necessary to transform the information contained in the molecular graph into numerical attributes" ("structural" descriptors) [29]. Some of the more popular structural descriptors, including topological indices, are: molecular connectivity [30] - [33], the presence or absence of some fixed subgraphs [16, 17, 29], the number of self-avoiding paths [34]. However, more or less direct use of molecular graphs is less frequent: for example, in fragment- or model-building approaches [11, pp.421,422], [35] in so-called genetic methods, [36] - [39], in various hybrids of both of them, as well as for the purposes of isomer generation and counting [40] - [45].
Another popular form of molecular representation, used for example to model the docking process [46] - [49], is the **3D molecular structure**. It is based on the 3D cartesian coordinates of the corresponding atoms. The goal of molecular docking, ideally, is to reconstruct the bound conformation starting from the structure of the unbound partners. Typically, docking programs use "a combination of geometry rules and optimization procedures to select the lowest energy conformer of the molecule..." [50].
Finally, since each of the above forms of representation has its own inadequacies, in accord with a general trend in artificial intelligence and pattern recognition, many proposed techniques attempt to combine different forms of representation or incorporate one form into the other thus creating **hybrid** forms of representation, e.g. incorporation of structural information into the vector representation or combination of genetic algorithms with ANNs [51] - [54].

3.2 Fundamental inadequacies of the basic representation and classification models used in CADD

In what follows, we briefly outline **intrinsic** inadequacies of the formal representations and classification models employed in CADD. The accumulated evidence strongly suggests that such inadequacies cannot be overcome by a simple restructuring of these models but can only be adequately dealt with by developing radically new, more "natural", structural classification models that are able to deal more effectively with the challenges of CADD (see also the next section).

<u>Vector space based models.</u> Perhaps, the most popular method for classification in CADD is **QSAR** and its recent relative quantitative structure-property relationship (**QSPR**) [55, 56] models. The main "external" feature of these models is the vector form of representation (composed of various descriptors) together with the compulsory Euclidean distance measure between the vectors. The main "internal" feature of the models is multiple linear regression used for predicting the activity/physicochemical properties of untested (and possibly not yet synthesized) compounds. Thus, the class corresponding to particular activity or property is determined by the corresponding regression coefficients.
As far as the external feature of this class of models is concerned, the assumption about the Euclidean distance is usually wrong, since, in this case, different vector coordinates are typically subject to different laws and, moreover, many of them are mutually incommensurable (e.g. topological indices and physicochemical descriptors). Perhaps a more immediate objection to these models, however, can be made on the basis of their reliance on the linear regression model, since it is hard to imagine that such a linear relationship between *all* the coordinates is meaningful.
The next popular family of classification methods in CADD are so called **artificial neural**

networks (ANN), or connectionist, models. One of the main arguments against these classification models, as mentioned in the last paragraph in connection with QSAR, is related to the use of Euclidean (or any other fixed) distance measure as classification tool for classifying numeric data for which no adequate distance measure is known or even possible. For a more extensive criticism of ANNs see, for example, [2] - [5], [10].
A very important common disadvantage of both QSAR- and ANN-type classification models is their *intrinsic inability to predict, or construct, the structure of any new class compounds*, which necessitates the exhaustive search through combinatorial libraries for the new compounds that belong to the corresponding class. We should also point out that it is the fundamental intrinsic inadequacies of the numeric classification models that has gradually lead to the development of the "structure-based" classification model discussed below.

Molecular graph based representations and various genetic methods. The basic underlying "space" in such structural representations is the set of labeled graphs, including trees or strings. It is not difficult to see that these combinatorial "spaces" are not the spaces in the conventional mathematical sense, since they do not possess any *fixed distance measure*, while all classical mathematical space (e.g Euclidean, Riemannian, etc.) possess such measure. It turns out, contrary to the classical "numeric" spaces, that there is no single natural distance or similarity measure which would be adequate for identification of various molecular classes, i.e. each class of compounds requires different molecular similarity measure, and this is the critical feature of *all* structural representations.
When using these models for molecular modeling, the organization and construction of a new molecular structure is accomplished by joining together structural fragments, or "building blocks", extracted from the previously constructed molecular structures. The two difficulties associated with such construction processes are:

(i) the inability of the construction process to satisfy the constraints associated with the chemically valid overall molecular structure, or its inability to construct the elements with the desired structural properties

(ii) the inability of the model to construct the similarity measure appropriate for the given task. [2]

All of the above should explain the substantial difficulties encountered by the researchers using the genetic approaches [37].
Moreover, it is important to note that the genetic methods have not suggested any *new form of structural representation*, nor have they proposed any new forms of molecular class description. What genetic methods have proposed are simply new *methods of constrained optimization for functions* defined on the various discrete spaces, including several types of "genetic" operations on such spaces (crossover, mutation, reproduction).

3D molecular structures. As mentioned above, this type of representation is used in the molecular docking methods. "The ideal docking method would allow both ligand and receptor to explore their conformation degrees of freedom. Perhaps, the most 'natural' way to incorporate the flexibility of the binding site is via a molecular dynamics simulation of the ligand-receptor complex. However, such calculations are computationally very demanding and are in practice only useful for refining structures produced using other docking methods; molecular dynamics does not explore the range of binding modes very well except for very small, mobile ligands. For many systems, the energy barriers that separate one binding mode from another are often

[2] It should be noted that although sometimes some researchers, with substantial difficulties, are able to provide the corresponding algorithm with a *tolerable* similarity, this is by no means an acceptable solution to the reliable automation of the molecular class construction process.

too large to be overcome" [11, p.557]. An obvious (but not obviously realizable) way out of these difficulties is to replace the coordinate representation by the appropriate "structural" [3] representation of both molecules, the receptor and the ligand, that preserve the nature of their interaction. For initial, very preliminary attempts to abandon the coordinate representation see, for example, [52].

3.3 The desirable features of a structural model for CADD

Historically, chemists, when thinking about chemical formulas, have always attached to them some additional structural features not present in the formulas themselves (if they are interpreted more formally). Unfortunately, the computer still (and in the foreseeable future) cannot be taught this *implicit* understanding of molecular representation which we store in our minds, whence comes the critical importance of formal models for representation in CADD. Since there has been no formal models of representation that would naturally capture the implicit (structural) understanding of molecular representation, different groups of researchers chose different existing formal models which they thought would capture more adequately the above implicit structural features. We next very briefly outline two currently dominant structural representations, graphs and 3D, from the above perspective.

Graphs. At present, in applied chemistry, for automating many modeling processes, the graph would be the most "convenient" form of structural representation. This is not surprising, since the labeled graph representation captures atom types as well as the chemical bond types. Which structural features graphs do not capture? Here are *some* of the important features that are lacking in graphs:

(i) the capability to capture the structural individuality of each atom's interactions

(ii) the capability to generate *only* chemically valid representations [4]

(iii) the ability to capture the global and local geometric features, including stereometric features (e.g. features that allow one to distinguish enantiomers and stereoisomers).

3D structures. The situation in this case has reversed. These forms of representation capture the local and global geometric feature, while missing those related to the more classical chemical understanding of molecular structure.

Considering the general issues related to a structural representation, one must, first of all, remember the *evolutionary nature of all biomolecules*, including various receptors. Equally important is the role played by the *class dependent structural similarity measure.*
Although by now, "biology" and "evolution" became almost synonymous, there were absolutely no formal models of representation that capture this, evolutionary, nature of object representation.[5] It appears that such representations require radical modification of the existing paradigms, and it also appears that there were no strong "pressure" to develop them. At the same time, one should note that it seems very undesirable to allow different forms of representation for different classes of molecules, e.g. evolutionary form for a receptor and a

[3]The use of quotes can be explained by the fact that here and throughout the paper *we* use the adjective "structural" in a more abstract sense that the one implied by the "3D molecular structure".

[4]Some initial attempts in this direction have been made within the graph grammar approach, but, as we have indicated in [1, Conclusion], they are far from being satisfactory.

[5]Genetic and so-called evolutionary approaches, despite the name of the latter, do not use evolutionary *representations.*

non-evolutionary form for the ligand. Why? We believe that the "evolutionary", in a more general sense, information (how the molecule was synthesized) may be quite significant.
It goes without saying that, in contrast to the one-sided capability of the classical CADD models to represent *given* molecules, a very important feature of the structural representational model is its *intrinsic* ability to generate only chemically valid representations (see also section 5). The model's capability to construct class dependent similarity measures is also of critical importance: basic CADD classification problems require corresponding structural similarity measures, and, in general, each molecular class must be delineated by its own similarity measure.
To summarize, we believe that the basic desirable features of a structural representation are the capabilities to represent

(i) molecular connectivity

(ii) structural individuality of each atom and each bonding type (including hydrogen bonding)

(iii) basic 2D and 3D isomerism

(iv) the "generative" molecular class structure (how one should go about building the molecule)

(v) class dependent structural similarity measure

(vi) only chemically valid elements of a drug class.

4 A new mathematical model for structural representation including class representation

This section is divided into three parts. Section 4.1 addresses the proposed new, structural, form of object "encoding" geared towards organic chemistry. Section 4.2 addresses a fundamentally new form of structural class description and the corresponding process of object construction (based on the class description). In section 4.3, we discuss several basic problems related to the process of construction of the class description based on the small set of typical class members. For a more formal and detailed exposition see [1].

4.1 Initial basic definitions: primitives and composites, semantic identities, i-structs, and i-transformations

In this section we outline the first half of basic definitions all of which clarify the proposed concept of structural representation. The original definitions in [1] are slightly specialized for the needs of the present paper. Moreover, many definitions are substantially simplified, again, taking into consideration a typical potential reader.
As always, for $A \subseteq X$ and a mapping $f : X \to \mathcal{S}$, we denote by $f\big|_A$ the restriction of f to A.

Definition 1. Let Π be a finite set whose elements are called primitive types, or simply **primtypes**, and let A be a countable set whose elements are called abstract sites, or simply *a-sites*. Moreover, for every $\pi \in \Pi$ two disjoint subsets of A

$$\mathrm{init}(\pi) \quad \text{and} \quad \mathrm{term}(\pi)$$

are given.[6] These sets specify the sets of **initial** and **terminal a-sites**, or abstract sites, for the primtype π. ▶

For the purposes of this paper, the primtypes can be viewed as different *types* of the basic "building blocks", e.g. various kinds of chemical bonds (different covalent bonds, hydrogen bond, etc), various elements in the periodic table (Fig 3). Note that from the point of view of structural representation it is useful to consider various bonds as separate primtypes.

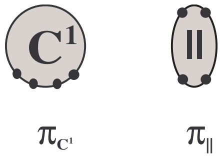


Figure 3: Pictorially, we represent primtypes as a circle or oval with the initial a-sites marked as points on its upper part and the terminal a-sites marked as points on its lower part.

For every primtype π we introduce the **set of all a-sites**

$$\text{sites}(\pi) = \text{init}(\pi) \cup \text{term}(\pi) \ .$$

Let $\mathcal{S}$ be the set of natural numbers (with zero), $\mathcal{S} = \mathbb{N}$, whose elements are called concrete sites, or simply **sites**.

Definition 2. The set Γ of **composites** is defined inductively as follows. For each $\gamma \in \Gamma$, three subsets of $\mathcal{S}$ — $\text{init}(\gamma)$, $\text{term}(\gamma)$ and $\text{sites}(\gamma)$[7] called the sets of **initial**, **terminal**, and **all sites** of composite γ — will now be constructed inductively.

- λ is the **null** composite whose sets of sites are

$$\text{init}(\lambda) = \text{term}(\lambda) = \text{sites}(\lambda) = \varnothing.$$

- For $\pi \in \Pi$ and a fixed injective mapping

$$f : \text{sites}(\pi) \to \mathcal{S}$$

(called the **site realization**[8] for primtype π), the expression

$$\pi\langle f\rangle$$

signifies the primitive composite, or simply **primitive**, whose sets of (concrete) sites are constructed as follows

$$\text{init}(\pi\langle f\rangle) = f(\text{init}(\pi)) \tag{1}$$
$$\text{term}(\pi\langle f\rangle) = f(\text{term}(\pi)) \tag{2}$$
$$\text{sites}(\pi\langle f\rangle) = f(\text{sites}(\pi)) \ . \tag{3}$$

(See Fig. 4.)

[6] Note that in this paper we assumed that $\text{init}(\pi) \cap \text{term}(\pi) = \varnothing$ (as compared to [1]).
[7] We use the same notation as that used in Def. 1, since these sets play a similar role.
[8] We will also use the term **site assignment**.

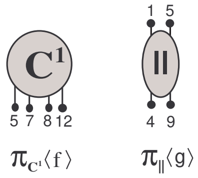


Figure 4: Pictorial representation of primitives corresponding to primtypes in Fig. 3. Both initial and terminal sites are always shown ordered in the same way as the corresponding abstract sites.

- For $\gamma \in \Gamma$, $\gamma \neq \lambda$, and $\pi\langle f\rangle \in \Gamma$ satisfying [9]

  $$\mathrm{sites}(\gamma) \cap \mathrm{sites}(\pi\langle f\rangle) = \mathrm{init}(\pi\langle f\rangle), \quad \mathrm{init}(\pi\langle f\rangle) \subseteq \mathrm{term}(\gamma)\,, \tag{4}$$

  the expression

  $$\gamma \lhd \pi\langle f\rangle$$

  signifies the composite γ', whose sets of (concrete) sites are constructed as follows

  $$\mathrm{init}(\gamma') = \mathrm{init}(\gamma) \tag{5}$$
  $$\mathrm{term}(\gamma') = [\mathrm{term}(\gamma) \setminus \mathrm{init}(\pi\langle f\rangle)] \cup \mathrm{term}(\pi\langle f\rangle) \tag{6}$$
  $$\mathrm{sites}(\gamma') = \mathrm{sites}(\gamma) \cup \mathrm{sites}(\pi\langle f\rangle). \tag{7}$$

  (See Fig. 5).

  We will call γ' the composite **obtained from γ by attachment of primitive $\pi\langle f\rangle$**, where the "attachment" means attaching to each other the identical sites in $\mathrm{term}(\gamma)$ and $\mathrm{init}(\pi\langle f\rangle)$.

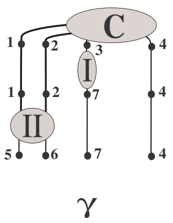


Figure 5: Pictorial representation of a composite. The order in which the primitives are attached corresponds to that specified in the construction process and the sites connected to each other must be identically labeled.

Thus, every composite γ is specified by the following **inductive expression** encapsulating its construction process

$$\gamma = \pi_1\langle f_1\rangle \lhd \pi_2\langle f_2\rangle \lhd \ldots \lhd \pi_n\langle f_n\rangle.$$

We will assume that the above expression is valid for $n = 0$ and in this case denotes λ. ▶

[9]As compared to [1], this condition is strenthen

Note that for a composite γ the union of its initial and terminal sites could be *smaller* than the set of all sites. Therefore, for a composite γ, it will also be useful to define the sets of its **external** and **internal sites**

$$\begin{aligned} \mathrm{ext}(\gamma) &= \mathrm{init}(\gamma) \cup \mathrm{term}(\gamma) \\ \mathrm{int}(\gamma) &= \mathrm{sites}(\gamma) \setminus \mathrm{ext}(\gamma) \ . \end{aligned} \tag{8}$$

Intuitively, the difference between primtype π and primitive $\pi\langle f\rangle$ can be compared to the difference between element C (carbon) in the periodic table and atom C in a particular compound. Thus, a composite can be viewed as a part of (or a whole) compound formed from concrete atoms (primitives) that were attached in a *particular temporal order*. Note that the corresponding (compound) construction process requires specification of the way the primitives are attached to each other, hence the need for concrete labels being assigned to the abstract sites (Figs. 5,6).
The following definition describes how to construct a composite out of other composites, i.e. how to construct a compound out of parts of other compounds.

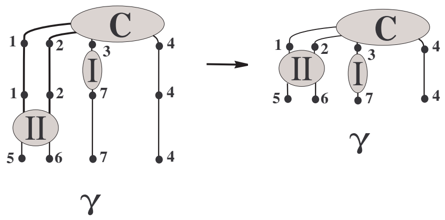


Figure 6: When drawing composites, if several primitives are attached consecutively to a single primitive, they will be drawn parallel to each other (ignoring their order in the construction process), although in general this order might be important.

Definition 3. Let α and β be two composites satisfying

$$\mathrm{sites}(\alpha) \cap \mathrm{sites}(\beta) = \mathrm{init}(\beta), \quad \mathrm{init}(\beta) \subseteq \mathrm{term}(\alpha)\,. \tag{9}$$

The **composition** of the above two composites,

$$\alpha \lhd \beta\ ,$$

is defined by induction on β as follows.

- $\alpha \lhd \lambda \stackrel{\mathrm{def}}{=} \alpha$.
- $$\alpha \lhd \pi\langle f\rangle \stackrel{\mathrm{def}}{=} \begin{cases} \pi\langle f\rangle, & \alpha = \lambda \\ \alpha \lhd \pi\langle f\rangle, & \alpha \neq \lambda \quad \text{(see Def. 2.} \end{cases}$$
- Assume that $\alpha \lhd \gamma$ has been constructed and that $\beta = \gamma \lhd \pi\langle f\rangle$, then
$$\alpha \lhd \beta \stackrel{\mathrm{def}}{=} (\alpha \lhd \gamma) \lhd \pi\langle f\rangle.$$

►

Remark 1. It is not difficult to see [10] that the sets of sites for the composition of two composites α and β are

$$\begin{aligned} \mathrm{init}(\alpha \lhd \beta) &= \mathrm{init}(\alpha) \\ \mathrm{term}(\alpha \lhd \beta) &= [\mathrm{term}(\alpha) \setminus \mathrm{init}(\beta)] \cup \mathrm{term}(\beta) \\ \mathrm{sites}(\alpha \lhd \beta) &= \mathrm{sites}(\alpha) \cup \mathrm{sites}(\beta) \ . \end{aligned} \tag{10}$$

Thus, since there are infinitely many different site assignments for a given composite, there are infinitely many different representations for that composite (Fig. 7). In order to be able to transform one such representation into another, we introduce the following definition.

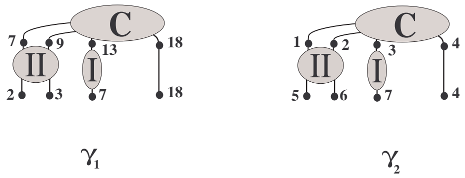


Figure 7: Two composites differing in the site assignments only.

Definition 4. For $\gamma \in \Gamma$ and any injective mapping

$$h : \mathrm{sites}(\gamma) \to \mathcal{S} \ ,$$

called **site replacement**, the composite $\gamma\langle h\rangle$ is defined inductively as follows.

- $\lambda\langle h\rangle \stackrel{\mathrm{def}}{=} \lambda$.
- If $\gamma = \pi\langle f\rangle$, then $\gamma\langle h\rangle \stackrel{\mathrm{def}}{=} \pi\langle g\rangle$, where $g = h \circ f$.
- Assume that $\alpha\langle h'\rangle$ has been constructed for any site replacement $h' : \mathrm{sites}(\alpha) \to \mathcal{S}$ and $\gamma = \alpha \lhd \pi\langle f\rangle$, then

  $$\gamma\langle h\rangle \stackrel{\mathrm{def}}{=} \alpha\langle h'\rangle \lhd \pi\langle g\rangle \ ,$$

  where $h' = h\big|_{\mathrm{sites}(\alpha)}$ and g is as above.

▶

Remark 2. For $\gamma \in \Gamma$ and any site replacement $h : \mathrm{sites}(\gamma) \to \mathcal{S}$, Def. 4 correctly defines composite $\gamma\langle h\rangle$, and, moreover, the following useful relationships hold [11]

$$\begin{aligned} \mathrm{init}(\gamma\langle h\rangle) &= h(\mathrm{init}(\gamma)) \\ \mathrm{term}(\gamma\langle h\rangle) &= h(\mathrm{term}(\gamma)) \\ \mathrm{sites}(\gamma\langle h\rangle) &= h(\mathrm{sites}(\gamma)) \ . \end{aligned} \tag{11}$$

[10] See Lemma 3 in [1].
[11] See Lemma 1 in [1].

Hence, the only sites that can be replaced in an arbitrary manner are the internal ones. In order for a composition of composites to "remain the same" after the site replacement, the site replacement functions must be mutually "consistent" on the external sites for the corresponding composites, (see Def. 7). To this end, in the following definition, we identify only those site replacement functions which conserve the external sites of the composites.

Definition 5. Two composites α and β will be called **similar** and we denote this fact by $\alpha \approx \beta$, if there exists site replacement $h : \mathrm{sites}(\beta) \to \mathcal{S}$ satisfying

$$h\big|_{\mathrm{ext}(\beta)} = id$$

(id is the identity mapping), such that

$$\alpha = \beta\langle h\rangle.$$

▶

Remark 3. If α and β are two similar composites and h is the corresponding site replacement, then [12]

$$\mathrm{init}(\alpha) = \mathrm{init}(\beta), \ \ \mathrm{term}(\alpha) = \mathrm{term}(\beta), \ \ \mathrm{int}(\alpha) = h(\mathrm{int}(\beta)).$$

In many cases it is necessary to treat different composites as "identical", i.e. as representing the same chemical objects. For example, in some cases it is necessary to consider different stereoisomers of a molecule as "identical" molecules. To this end, we introduce the following definition.

Definition 6. Let α, β be two composites such that

$$\mathrm{init}(\alpha) = \mathrm{init}(\beta), \qquad \mathrm{term}(\alpha) = \mathrm{term}(\beta).$$

The expression

$$\alpha \equiv \beta$$

is called **semantic identity** and signifies the indistinguishability of the corresponding two parts in an object representation. ▶

Definition 7. Let $\mathcal{I}$ be a specified set of semantic identities.[13] This set induces naturally the semantic equivalence relation, or simply **semantic relation**, denoted $\sim$, on the set of composites Γ as follows.

1. If $\alpha \equiv \beta$ is a semantic identity (from $\mathcal{I}$), then $\alpha \sim \beta$.
2. If $\alpha \sim \beta$ and

$$f : \mathrm{sites}(\alpha) \to \mathcal{S}$$
$$g : \mathrm{sites}(\beta) \to \mathcal{S}$$

are **externally consistent** site replacements, i.e.

$$f\big|_{\mathrm{ext}(\alpha)} = g\big|_{\mathrm{ext}(\beta)},$$

then

$$\alpha\langle f\rangle \ \sim \ \beta\langle g\rangle.$$

[12] See Lemma 2 in [1].
[13] Note that usually $\mathcal{I}$ is a small subset of the set of all semantic identities.

3. If $\alpha \sim \beta$, $\gamma \sim \delta$ and compositions $\alpha \lhd \gamma$, $\beta \lhd \delta$ exist, then
$$\alpha \lhd \gamma \ \sim \ \beta \lhd \delta.$$
4. Finally, the binary relation $\sim$ is defined as the minimal equivalence relation satisfying the above three conditions, i.e. it is the intersection of all the equivalence relations satisfying the above conditions.

►

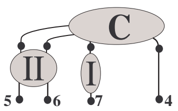
Remark 4. If two composites are semantically equivalent, then their sets of initial and terminal sites are identical.[14]

The class of equivalent composites can be considered as representing a single chemical object in a particular problem.

Definition 8. Let Π, $\mathcal{I}$ be specified sets of primtypes and semantic identities. The quotient set
$$\Theta = \Gamma/_{\sim} = \{\,[\gamma]\,|\,\gamma \in \Gamma\,\}$$
will be called the set of instance structs, or simply **i-structs** (for $(\Pi, \mathcal{I})$). I-struct $[\lambda]$ will be called the **empty i-struct** and denoted $\boldsymbol{\lambda}$. For each i-struct $[\gamma]$, also denoted, $\boldsymbol{\gamma}$, the three sets of sites are defined as follows
$$\mathrm{init}(\boldsymbol{\gamma}) = \mathrm{init}(\gamma), \quad \mathrm{term}(\boldsymbol{\gamma}) = \mathrm{term}(\gamma) \quad \mathrm{ext}(\boldsymbol{\gamma}) = \mathrm{ext}(\gamma)\ .$$

►

Hence, i-structs "represent" chemical objects in a concrete problem (Figs. 8,9).



$\boldsymbol{\gamma}$

Figure 8: Pictorial representation of the i-struct $\boldsymbol{\gamma} = [\gamma]$ corresponding to the composite in Fig. 5.

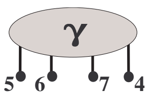


Figure 9: A simplified pictorial representation of i-struct $\boldsymbol{\gamma}$ in Fig. 8.

To work with the quotient set (the set of i-structs), one should be able (algorithmically), first, to find a class of equivalent structs and, then, learn how to perform the corresponding operations on the quotient set (that is how to attach i-structs or replace sites of i-structs). Accordingly, in

[14] See Lemma 5 in [1].

a concrete problem, each chemical object can be represented by a "canonical" element from the equivalence class (together with the set of equivalence relations appropriate for this concrete problem). The next two definitions introduce the above two important operations on the set of i-structs.

Definition 9. Let $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ be two i-structs satisfying

$$\operatorname{ext}(\boldsymbol{\alpha}) \cap \operatorname{ext}(\boldsymbol{\beta}) = \operatorname{term}(\boldsymbol{\alpha}) \cap \operatorname{init}(\boldsymbol{\beta}) .$$

The **composition of** the above two **i-structs**, $\boldsymbol{\alpha} \lhd \boldsymbol{\beta}$, is defined as

$$\boldsymbol{\alpha} \lhd \boldsymbol{\beta} = [\alpha \lhd \beta] ,$$

where

$$\alpha \in \boldsymbol{\alpha}, \quad \beta \in \boldsymbol{\beta}, \quad \text{and} \quad \alpha \lhd \beta \text{ exists.}$$

▶

Remark 5. One can show that the composition of the above i-structs $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ exists and is correctly defined, i.e. it does not depend on the choice of "canonical" composites α and β. [15]

Definition 10. For an i-struct $\boldsymbol{\gamma}$ and an injective mapping $\mathbf{h} : \operatorname{ext}(\boldsymbol{\gamma}) \to \mathcal{S}$, called **i-struct site replacement**, the i-struct $\boldsymbol{\gamma}\langle\mathbf{h}\rangle$ is defined as

$$\boldsymbol{\gamma}\langle\mathbf{h}\rangle = [\gamma\langle h\rangle] ,$$

where $\gamma \in \boldsymbol{\gamma}$ and $h : \operatorname{sites}(\gamma) \to \mathcal{S}$ is a site replacement satisfying

$$h\big|_{\operatorname{ext}(\gamma)} = \mathbf{h}.$$

▶

Remark 6. For an i-struct $\boldsymbol{\gamma}$ and a site replacement $\mathbf{h} : \operatorname{ext}(\boldsymbol{\gamma}) \to \mathcal{S}$, the i-struct $\boldsymbol{\gamma}\langle\mathbf{h}\rangle$ does not depend on the choice of γ and h.[16]

Note that although a composite, typically, has internal sites, we do not introduce the concept of internal sites for an i-struct, since an i-struct is "independent" of the internal sites of its canonical composite. Also, as was the case with composites, in general, we cannot modify external sites of an i-struct arbitrarily, since the modification may change the manner of attachment of this i-struct to other i-structs.

So far, we have allowed an i-struct to be attached to any other i-struct with the appropriate external sites. In most "real" cases, however, there are additional restrictions on which i-structs are "allowed" to be attached to each other. For example, when representing molecules, one doesn't allow one atom primitive to be attached to another one bypassing the bond primitives "between them" or attach bond primitives without any atom primitives "between them" (Fig 10). That is why we need to introduce two of the most central concepts, the concepts of i-transformation and its context, to specify these additional restrictions on the allowed attachments.

[15] See Lemma 8 in [1].
[16] See Lemma 7 in [1].

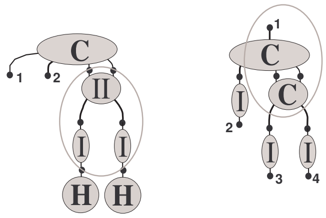


Figure 10: Inadmissible i-structs.

Definition 11. Let $(\Pi, \mathcal{I})$ be given and let Θ be the corresponding set of i-structs. A pair of i-structs $\boldsymbol{\tau} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ such that there exists i-struct $\boldsymbol{\delta}$ satisfying

$$\boldsymbol{\beta} = \boldsymbol{\alpha} \lhd \boldsymbol{\delta}$$

will be called an i-struct transformation, or simply **i-transformation** and $\boldsymbol{\alpha}$ will be called the **context of i-transformation** $\boldsymbol{\tau}$ (See Fig. 11). If $\boldsymbol{\alpha} = [\lambda]$, the i-transformation will be called **context free**. The set of all i-struct transformations for $(\Pi, \mathcal{I})$ will be denoted by $\mathcal{T}$. ▶

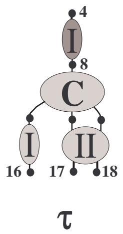


Figure 11: Pictorial representation of an i-transformation in which the context of the i-transformation is shaded.

For i-transformation $\boldsymbol{\tau} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, let

$$\mathrm{ext}(\boldsymbol{\tau}) \stackrel{\mathrm{def}}{=} \mathrm{ext}(\boldsymbol{\alpha}) \cup \mathrm{ext}(\boldsymbol{\beta}).$$

As far as chemical processes are concerned we think of an i-transformation as a generalized "reaction-module" in the appropriate chemical processes.

Definition 12. For an i-struct $\boldsymbol{\gamma}$, i-transformation $\boldsymbol{\tau} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, and site replacement $\mathbf{h}$: $\mathrm{ext}(\boldsymbol{\alpha}) \cup \mathrm{ext}(\boldsymbol{\beta}) \to \mathcal{S}$ such that there exists $\boldsymbol{\delta}$ satisfying

$$\boldsymbol{\gamma} = \boldsymbol{\delta} \lhd \boldsymbol{\alpha} \left\langle \mathbf{h}\big|_{\mathrm{ext}(\boldsymbol{\alpha})} \right\rangle ,$$

the $\boldsymbol{\tau}\langle\mathbf{h}\rangle$**-transformation of i-struct** $\boldsymbol{\gamma}$, denoted $\boldsymbol{\gamma} \lhd \boldsymbol{\tau}\langle\mathbf{h}\rangle$, is defined as the i-struct

$$\boldsymbol{\delta} \lhd \boldsymbol{\beta} \left\langle \mathbf{h}\big|_{\mathrm{ext}(\boldsymbol{\beta})} \right\rangle .$$

(See Fig. 12.) ▶

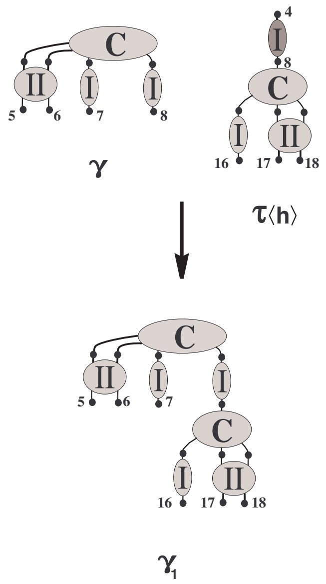


Figure 12: Pictorial representation of $\boldsymbol{\tau}\langle\mathbf{h}\rangle$ -transformation of i-struct $\boldsymbol{\gamma}$.

Remark 7. Note that in the last definition i-struct $\boldsymbol{\gamma} \lhd \boldsymbol{\tau}\langle\mathbf{h}\rangle$ is correctly defined, i.e. it does not depend on the choice of $\boldsymbol{\delta}$.[17]

Why does the notion of an i-transformation play a critical role in the proposed formalism? In the next section we will see that a class of chemical objects representations in our formalism is defined based on a finite set of i-transformations. Such sets of i-transformations define the proposed constructive view of chemical objects, which is consistent with the mainstream view of chemical processes as based on molecular transformations [57, sec.1.5].

4.2 Structs, classes and the class generative process

The following definition introduces the concept of *restricted* (problem depended) environment in which one is planning to work. For example, if we restrict ourselves to work only with hydrocarbons, we don't need nitrogen or oxygen primtypes, which will be necessary if we expand our environment to the set of all organic compounds.

Definition 13. A pair $(\Pi, \mathcal{I})$, where Π is a finite set of primtypes, $\mathcal{I}$ is a specified set of semantic identities will be called **inductive structure**. ▶

So far we represented chemical objects by i-structs. I-structs depend on the external sites labels, while "real" chemical objects do not depend on them. The following concepts of struct and transformation allow us, for the purpose of class description, to remove the dependence of i-structs on the external sites. The concept of struct is introduced in a manner similar to that of an i-struct (see Def. 8).

[17]See Lemma 13 in [1].

Definition 14. For an i-struct $\boldsymbol{\gamma} \in \Theta$, let

$$\bar{\gamma} \stackrel{\text{def}}{=} \{\boldsymbol{\gamma}\langle\mathbf{h}\rangle \,|\, \mathbf{h} \text{ is a site replacement}\}$$

be called an **struct**.
For an i-transformation $\boldsymbol{\tau} \in \mathcal{T}$, let

$$\bar{\boldsymbol{\tau}} \stackrel{\text{def}}{=} \{\boldsymbol{\tau}\langle\mathbf{h}\rangle \,|\, \mathbf{h} \text{ is a site replacement}\}$$

be called an **transformation**.
We will denote by $\bar{\Theta}$ [18] and $\bar{\mathcal{T}}$ be the sets of structs and transformations, respectively. ▶

The following simple concepts are among the most central ones

Definition 15. A finite set of transformations $T \subset \bar{\mathcal{T}}$ will be called a **transformation set**. A triple $WT = (T, w, l)$, where T is a transformation set, $w : T \to \mathbb{R}_+$, and $l : T \to \mathbb{R}_+$, will be called a **weighted transformation set**. We think of $w(\bar{\boldsymbol{\tau}})$ as the "weight" of $\bar{\boldsymbol{\tau}}$ and of $l(\bar{\boldsymbol{\tau}})$ as the "time it takes to apply" $\bar{\boldsymbol{\tau}}$.
A quadruple $\mathbf{TS} = (T, w, l, \bar{\boldsymbol{\kappa}})$, where (T, w, l) is a weighted transformation set and $\bar{\boldsymbol{\kappa}} \in \bar{\Theta}$ is a struct called **progenitor**, will be called a **transformation system**. ▶

Remark 8. Note that given a transformation system $\mathbf{TS}$ one can construct structs "associated" with it by consecutively applying its transformations to the structs thus constructed, starting from the "initial" struct, progenitor. The set of all structs that can be generated in such a way will be denoted as TS. [19]

For a given transformation system $\mathbf{TS} = (T, w, l, \bar{\boldsymbol{\kappa}})$ and a struct $\bar{\gamma} \in TS$, the expression

$$\boldsymbol{\gamma} = \boldsymbol{\kappa} \lhd \boldsymbol{\tau}_1\langle\mathbf{h}_1\rangle \lhd \ldots \lhd \boldsymbol{\tau}_m\langle\mathbf{h}_m\rangle$$

will be called the inductive transformation expression, or simply **i-transformation expression**.

It is also useful to note that from the biological/biochemical point of view the progenitor is the common ancestor of all structs in TS. At the same time, from the point of view of drug design, the progenitor can be viewed as the generalization of the concepts of pharmacophore and drug lead.
The structs from TS can be constructed, or "generated", *step by step*, starting from the progenitor by applying transformations. This generating process is defined next.[20]

Definition 16. For a transformation system $\mathbf{TS} = (T, w, l, \bar{\boldsymbol{\kappa}})$, the **generating process** $G_{\mathbf{TS}}$, or simply G, is defined as the following countable state Markov stochastic process.[21]

- The starting state of the process is the progenitor.
- Next the process chooses a transformation $\bar{\boldsymbol{\tau}}$ with probability proportional to its weight $w(\bar{\boldsymbol{\tau}})$.

[18] Note, that, for $\bar{\boldsymbol{\theta}} \in \bar{\Theta}$, the notation $\bar{\boldsymbol{\theta}} = [[\gamma]]$, where $[\gamma]$ denotes the corresponding i-struct (see Def. 8) will be also used.
[19] For the formal definition of TS see Defs. 23-29 in [1].
[20] We give only a semi-formal definition of the generating process. For the complete definition see [1].
[21] For explanation of why the generation process can be modeled by Markov process see discussion in [1].

- Then the process applies the chosen transformation, and the average time of application is $l(\bar{\boldsymbol{\tau}})$. Until the application is completed, i.e. the new struct is constructed, the process remains in the previous state.
- In the new state, or struct, the process proceeds again as described above.

▶

The following condition ensures the existence of the typicality measure (see Def. 19 below) on the set of structs generated by a transformation system.

Definition 17. Let **TS** be a transformation system and G be the generating process for **TS**. Let $E_G(\bar{\boldsymbol{\gamma}})$ be the expected time spent by G in state (struct) $\bar{\boldsymbol{\gamma}}$.
Let

$$E_G \overset{\text{def}}{=} \sum_{\bar{\boldsymbol{\gamma}} \in TS} E_G(\bar{\boldsymbol{\gamma}}).$$

If E_G is finite, we will say that transformation system **TS** satisfies the **typicality measure existence condition**.
▶

Definition 18. A transformation system satisfying the typicality measure existence condition will be called a **class**. ▶

Which structs in the class are "typical"? Obviously, not all structs: in a class of drugs exhibiting certain characteristic biological activity, some drugs a more active than the other. This is an intuitive understanding of the concept of typicality measure defined next.

Definition 19. Let **C** be a class and G be the generating process for **C**. For $\bar{\boldsymbol{\gamma}} \in C$,[22] let

$$\nu_{\mathbf{C}}(\bar{\boldsymbol{\gamma}}) \overset{\text{def}}{=} \frac{E_G(\bar{\boldsymbol{\gamma}})}{E_G}.$$

For $\bar{\boldsymbol{\gamma}} \notin C$, $\nu_{\mathbf{C}}(\bar{\boldsymbol{\gamma}}) \overset{\text{def}}{=} 0$. Measure $\nu_{\mathbf{C}}$ on C will be called the **C-typicality measure**.
For a finite set $S \subset C$, its **C**-typicality is defined as follows

$$\nu_{\mathbf{C}}(S) \overset{\text{def}}{=} \prod_{\bar{\boldsymbol{\gamma}} \in S} \nu_{\mathbf{C}}(\bar{\boldsymbol{\gamma}}).$$

▶

The typicality $\nu_{\mathbf{TS}}$ of a struct $\bar{\boldsymbol{\gamma}}$ is defined to be proportional to the expected (average) time the process spends in that struct. This time is proportional to

- the probability of the process reaching struct $\bar{\boldsymbol{\gamma}}$;
- the average time spent in $\bar{\boldsymbol{\gamma}}$, when it has been reached.

[22] Since **C** is a transformation system, C stands for the set of structs generated by **C**.

4.3 Learning based on different types of training sets

We now formulate in the proposed model the basic problems of inductive learning, i.e. several versions of the problem of learning class description from the small set of examples. But, first, we need to fix an inductive structure $(\Pi, \mathcal{I})$ and a "superclass" **IC** in it that is supposed to encapsulate the representations of all chemical objects of interest for the chosen task. By a **training set,** we mean a finite subset of structs in **IC**. We next address the problem of recognizing an element from **IC** as belonging or not belonging to class **C** related to the chosen task.

Definition 20. A struct $\bar{\gamma} \in IC$ **is recognized as an element of C**, if

$$\bar{\gamma} \in C.$$

▶

A **positive set** S^+ **for a class C** is defined as a set of structs each element of which is recognized as an element of **C** .
A **negative set** S^- **for a class C** is defined as a set each element of which is recognized as not an element of **C**. Thus, arbitrary set can be partitioned into two disjoint subsets, positive set and negative set.
To define the typicality of a transformation set we need to introduce the concept of typicality of a single transformation.

Definition 21. Let $\bar{\boldsymbol{\tau}}$, $\bar{\boldsymbol{\tau}} = [\boldsymbol{\tau}]$, be a transformation: $\boldsymbol{\tau} \in \bar{\mathcal{T}}$. The **typicality of transformation** $\bar{\boldsymbol{\tau}}$ is defined as follows:

$$\nu_{\mathbf{IC}}(\bar{\boldsymbol{\tau}}) = \max_{\boldsymbol{\alpha} \in A_{\boldsymbol{\tau}}} \nu_{\mathbf{IC}}([\boldsymbol{\alpha} \lhd \boldsymbol{\tau}]),$$

where $A_{\boldsymbol{\tau}} = \{\, \boldsymbol{\alpha} \mid [\boldsymbol{\alpha}] \in IC \text{ and } \exists \boldsymbol{\alpha} \lhd \boldsymbol{\tau} \,\}$. ▶

We now can introduce the concept of typicality for a transformation set which plays an important role in the definition of the learning problem.

Definition 22. Let $T = \{\bar{\boldsymbol{\tau}}_1, \bar{\boldsymbol{\tau}}_2, \ldots, \bar{\boldsymbol{\tau}}_n\}$, $T \subseteq \mathcal{T}$ be a transformation set. The **typicality of transformation set** T is defined as follows

$$\nu_{\mathbf{IC}}(T) = \prod_{\bar{\boldsymbol{\tau}} \in T} \nu_{\mathbf{IC}}(\bar{\boldsymbol{\tau}}).$$

▶

Definition 23. The **problem of learning from a positive training set** is formulated as follows. Given a training set S^+ of structs find a class **C** [23] such that

(i) S^+ is a positive set for **C**

(ii) for any other class $\mathbf{C}'$ satisfying (i),

$$\nu_{\mathbf{IC}}(T_{\mathbf{C}})\nu_{\mathbf{C}}(S^+) \geq \nu_{\mathbf{IC}}(T_{\mathbf{C}'})\nu_{\mathbf{C}'}(S^+),$$

where $T_{\mathbf{C}}$, $T_{\mathbf{C}'}$ are the transformation sets for the corresponding classes.

[23] I.e. find the appropriate transformation system (see Defs. 15, 18).

▶

The above definition ensures maximum possible typicality of the training set S^+ in the constructed class $\mathbf{C}$ ($\nu_{\mathbf{C}}(S^+)$) *under the constraints* on the complexity of the transformations in its transformation system ($\nu_{\mathbf{IC}}(T_{\mathbf{C}})$). More accurately, the first factor in the product ensures that the class transformations are as small as possible while the second factor ensures that they are as large as possible.[24]

Definition 24. Let μ be some similarity measure on the set of classes. Let $\mathbf{C}$ be a class and S^- be a training set of elements from **IC**. The **class adjustment problem** is formulated as follows: find a class $\mathbf{C}'$ such that S^- is a negative set for $\mathbf{C}'$ and $\mu(\mathbf{C}', \mathbf{C})$ is maximal. ▶

The class adjustment problem, thus, consists of finding a class $\mathbf{C}'$ that is a minimal modification of class $\mathbf{C}$ which does not contain the elements of S^-.
Thus, in particular, it becomes possible to "reduce" a previously constructed class of structs based on the modified training set of chemical objects with undesirable properties (e.g. side effects).

Definition 25. Let S^+ and S^- be two training sets. The **problem of learning from positive and negative examples** is formulated as follows: find a class $\mathbf{C}$ such that

(i) S^+ is a positive set for $\mathbf{C}$ and S^- is a negative set for $\mathbf{C}$

(ii) for any other class $\mathbf{C}'$ satisfying (i),

$$\nu_{\mathbf{IC}}(T_{\mathbf{C}})\nu_{\mathbf{C}}(S^+) \geq \nu_{\mathbf{IC}}(T_{\mathbf{C}'})\nu_{\mathbf{C}'}(S^+).$$

▶

5 Introductory examples from organic chemistry

In this section we illustrate *some* of the basic concepts and ideas of the proposed model. The new concepts are so revolutionary that their *straightforward* application is out of question. The examples of this section represent only the *very first* steps in our quest to understand the nature of what the chemical structure is. We, first, present an example of the organic inductive structure (including its several subclasses) that models compounds with the covalent bonds only, and, then, an example of the organic inductive structure that models compounds with both covalent and hydrogen bonds.

5.1 An ETS model for covalent bonding

5.1.1 The covalent bonding inductive structure

We restrict ourselves to compounds formed from the following atoms: **H**, **C**, **O**, **N**, **Cl**, and **Br** and the following covalent bond types: single, double, triple. [25]

<u>Primitives</u>

[24] I.e., as was mentioned above, the second factor ensures the maximal typicality of the training set.

[25] The representation can be easily generalized to any larger set of atoms, e.g. if we add to the above list **S**, **P**, **I**, **F**.

We first define the primtypes. We remind the reader that the set of sites, $\mathcal{S}$, is chosen to be $\mathbb{N}$ (including zero). All, twenty three, primtypes are shown in Fig. 13. The corresponding primitives are defined in the following table, where the first 12 primitives correspond to atoms, the next 6 primitives correspond to covalent bonds, and the last 5 primitives are auxiliary primitives necessary to correctly initiate and terminate the process of struct construction. [26]

$\pi \in \Pi$	$\text{init}(\pi)$	$\text{term}(\pi)$
π_{C^1}	$\{\varnothing\}$	$\{1,2,3,4\}$
π_{C^2}	$\{1\}$	$\{2,3,4\}$
π_{C^3}	$\{1,2\}$	$\{3,4\}$
π_{C^4}	$\{1,2,3\}$	$\{4\}$
π_{H^1}	$\{1\}$	$\{\varnothing\}$
π_{Cl^1}	$\{1\}$	$\{\varnothing\}$
π_{N^1}	$\{1\}$	$\{2,3\}$
π_{N^2}	$\{1,2\}$	$\{3\}$
π_{N^3}	$\{1,2,3\}$	$\{\varnothing\}$
π_{O^1}	$\{1,2\}$	$\{\varnothing\}$
π_{O^2}	$\{1\}$	$\{2\}$
π_{Br^1}	$\{1\}$	$\{\varnothing\}$
$\pi_{\vert}$	$\{1\}$	$\{2\}$
$\pi_{\vert\vert}$	$\{1,2\}$	$\{3,4\}$
$\pi_{\vert\vert\vert}$	$\{1,2,3\}$	$\{4,5,6\}$
π_{-}	$\{1,2\}$	$\{\varnothing\}$
$\pi_{=}$	$\{1,2,3,4\}$	$\{\varnothing\}$
$\pi_{\equiv}$	$\{1,2,3,4,5,6\}$	$\{\varnothing\}$
π_{X^1}	$\{1\}$	$\{\varnothing\}$
π_{X^2}	$\{1,2\}$	$\{\varnothing\}$
π_{X^3}	$\{1,2,3\}$	$\{\varnothing\}$
π_{CB}	$\{\varnothing\}$	$\{1\}$
π_{CD}	$\{1\}$	$\{\varnothing\}$

Semantic identities

For any primitive π with $|\text{init}(\pi)| = n$ and $|\text{term}(\pi)| = m$, we will denote the values of its site realization $f : [1, n+m] \to \mathbb{N}$ as

$$\langle f(1), \dots, f(n) \mid f(n+1), \dots, f(n+m) \rangle.$$

1. Even (site) permutations. Let

$$\Pi_1 = \{\pi_{C^1}, \pi_{C^2}, \pi_{C^3}, \pi_{C^4}, \pi_{||}, \pi_{=}\}.$$

For any primtype $\pi \in \Pi_1$ and any even permutation $\sigma : \{1,2,3,4\} \to \{1,2,3,4\}$ such that

$$\sigma(\text{init}(\pi)) = \text{init}(\pi), \qquad \sigma(\text{term}(\pi)) = \text{term}(\pi), \qquad (*)$$

the identity is

$$\pi\langle 1,2,3,4\rangle \equiv \pi\langle\sigma\rangle.$$

This kind of identities allows one to regard as equivalent two composites representing two compounds one of which is obtained from the other by the 3D rotation of itself or its part,

[26] For an additional discussion of the primitives see section 5.1.3.

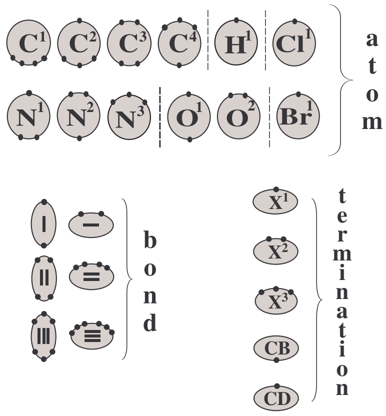


Figure 13: Primtypes for the covalent bonding inductive structure. Note that the **vertical** and the **horizontal bond** (as well as the **atom** and the **termination**) primtypes are grouped together. CB and CD stand for the compound's birth and death primtypes correspondingly.

where the rotation is specified by the corresponding identity. So, it is not difficult to see that the corresponding inductive structure allows one to distinguish all the compounds up to stereoisomers (enantiomers and diastereoisomers).

2. Arbitrary (site) permutations. Let

$$\Pi_2 = \{\pi_{|||}, \pi_{-}, \pi_{=}, \pi_{\equiv}, \pi_{X^2}, \pi_{X^3}, \pi_{N^1}, \pi_{N^2}, \pi_{N^3}, \pi_{O^2}\}.$$

For a primtype $\pi \in \Pi_2$ and any arbitrary permutation $\sigma : \mathrm{sites}(\pi) \to \mathrm{sites}(\pi)$ satisfying $(*)$, the identity is

$$\pi\langle id\rangle \equiv \pi\langle\sigma\rangle$$

(id is the identity mapping).

This kind of identities specify that if two composites are obtained from one composite by attaching to it (separately) two primitives, both corresponding to the same primtype from Π_2, differing only in site permutations, the resulting composites are considered equivalent.

3. Identities for parallel attachments. For all primitives $\pi_1\langle f_1\rangle$, $\pi_2\langle f_2\rangle$ satisfying

$$\mathrm{sites}(\pi_1\langle f_1\rangle) \cap \mathrm{sites}(\pi_2\langle f_2\rangle) = \varnothing,$$

the identity is

$$\pi_1\langle f_1\rangle \lhd \pi_2\langle f_2\rangle \equiv \pi_2\langle f_2\rangle \lhd \pi_1\langle f_1\rangle \qquad \text{(see Fig. 14).}$$

The primitives of this kind can be considered as being attached independently, or in parallel. Thus, the order in which they are attached is not important.

4. Vertical two–single–to–double bond translations. For all primitives $\pi_1\langle f_1\rangle, \pi_2\langle f_2\rangle$ satisfying

$$1, 2 \in \mathrm{term}(\pi_1\langle f_1\rangle) \text{ and } 3, 4 \in \mathrm{init}(\pi_2\langle f_2\rangle),$$

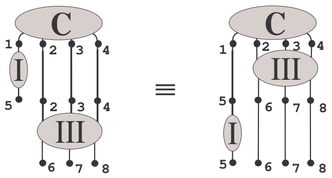


Figure 14: An example of identities for parallel attachments.

the identity is

$$\pi_1\langle f_1\rangle \lhd \pi_|\langle 1|3\rangle \lhd \pi_|\langle 2|4\rangle \lhd \pi_2\langle f_2\rangle \equiv \pi_1\langle f_1\rangle \lhd \pi_{||}\langle 1,2|3,4\rangle \lhd \pi_2\langle f_2\rangle \quad \text{(see Fig. 15).}$$

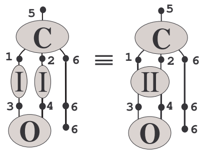


Figure 15: An example of vertical two–single–to–double bond translations.

Vertical three–single–to–triple bond translations. For all primitives $\pi_1\langle f_1\rangle, \pi_2\langle f_2\rangle$ satisfying

$$1,2,3 \in \mathrm{term}(\pi_1\langle f_1\rangle) \text{ and } 4,5,6 \in \mathrm{init}(\pi_2\langle f_2\rangle),$$

the identity is

$$\pi_1\langle f_1\rangle \lhd \pi_|\langle 1|4\rangle \lhd \pi_|\langle 2|5\rangle \lhd \pi_|\langle 3|6\rangle \lhd \pi_2\langle f_2\rangle \equiv \pi_1\langle f_1\rangle \lhd \pi_{|||}\langle 1,2,3|4,5,6\rangle \lhd \pi_2\langle f_2\rangle.$$

Note that the above two sets of identities allow one to translate the combination of vertical single and double into vertical triple bond.

Horizontal two–single–to–double bond translations. For all primitives $\pi_1\langle f_1\rangle, \pi_2\langle f_2\rangle$ satisfying

$$1,2 \in \mathrm{term}(\pi_1\langle f_1\rangle), \text{ and } 3,4 \in \mathrm{term}(\pi_2\langle f_2\rangle),$$

the identity is

$$\pi_1\langle f_1\rangle \lhd \pi_2\langle f_2\rangle \lhd \pi_|\langle 1|5\rangle \lhd \pi_|\langle 2|6\rangle \lhd \pi_|\langle 3|7\rangle \lhd \pi_|\langle 4|8\rangle \lhd \pi_-\langle 5,7|\rangle \lhd \pi_-\langle 6,8|\rangle \equiv$$
$$\pi_1\langle f_1\rangle \lhd \pi_2\langle f_2\rangle \lhd \pi_{||}\langle 1,2|5,6\rangle \lhd \pi_{||}\langle 3,4|7,8\rangle \lhd \pi_=\langle 5,6,7,8|\rangle \quad \text{(Fig. 16).}$$

Horizontal three–single–to–triple bond translations. For all primitives $\pi_1\langle f_1\rangle, \pi_2\langle f_2\rangle$ satisfying

$$1,2,3 \in \mathrm{term}(\pi_1\langle f_1\rangle), \text{ and } 4,5,6 \in \mathrm{term}(\pi_2\langle f_2\rangle),$$

the identity is

$$\pi_1\langle f_1\rangle \lhd \pi_2\langle f_2\rangle \lhd \pi_|\langle 1|7\rangle \lhd \pi_|\langle 2|8\rangle \lhd \pi_|\langle 3|9\rangle \lhd \pi_|\langle 4|10\rangle \lhd \pi_|\langle 5|11\rangle \lhd \pi_|\langle 6|12\rangle \lhd$$
$$\pi_-\langle 7,10|\rangle \lhd \pi_-\langle 8,11|\rangle \lhd \pi_-\langle 9,12|\rangle \equiv$$
$$\pi_1\langle f_1\rangle \lhd \pi_2\langle f_2\rangle \lhd \pi_{|||}\langle 1,2,3|7,8,9\rangle \lhd \pi_{|||}\langle 4,5,6|10,11,12\rangle \lhd \pi_\equiv\langle 7,8,9,10,11,12|\rangle.$$

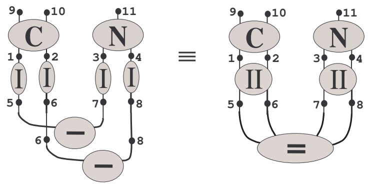


Figure 16: An example of two–single–to–double bond translations.

Note that the above two sets of identities allow one to translate the combination of horizontal single and double into horizontal triple bond.

The above four types of identities allow one to "translate" several single bonds between two atoms equal into a double or triple bond between these atoms.

5. <u>Process termination translations.</u>

$$\pi_{C^1}\langle|1,2,3,4\rangle \lhd \pi_{||}\langle 1,2|5,6\rangle \lhd \pi_{||}\langle 3,4|7,8\rangle \lhd \pi_{=}\langle 5,6,7,8|\rangle \equiv$$
$$\equiv \pi_{C^1}\langle|1,2,3,4\rangle \lhd \pi_{||}\langle 1,2|5,6\rangle \lhd \pi_{||}\langle 3,4|7,8\rangle \lhd \pi_{X^2}\langle 5,6|\rangle \lhd \pi_{X^2}\langle 7,8|\rangle \quad \text{(Fig. 17)}$$

$$\pi_{C^1}\langle|1,2,3,4\rangle \lhd \pi_{|||}\langle 1,2,3|5,6,7\rangle \lhd \pi_{|}\langle 4|8\rangle \lhd \pi_{C^2}\langle 8|9,10,11\rangle$$
$$\lhd \pi_{|||}\langle 9,10,11|12,13,14\rangle \lhd \pi_{\equiv}\langle 5,6,7,12,13,14\rangle \equiv$$
$$\equiv \pi_{C^1}\langle|1,2,3,4\rangle \lhd \pi_{|||}\langle 1,2,3|5,6,7\rangle \lhd \pi_{|}\langle 4|8\rangle \lhd \pi_{C^2}\langle 8|9,10,11\rangle \lhd$$
$$\lhd \pi_{|||}\langle 9,10,11|12,13,14\rangle \lhd \pi_{X^3}\langle 5,6,7|\rangle \lhd \pi_{X^3}\langle 12,13,14|\rangle$$

$$\pi_{C^1}\langle|1,2,3,4\rangle \lhd \pi_{||}\langle 1,2|5,6\rangle \lhd \pi_{||}\langle 3,4|7,8\rangle \lhd \pi_{C^3}\langle 7,8|9,10\rangle$$
$$\lhd \pi_{||}\langle 9,10|11,12\rangle \lhd \pi_{=}\langle 5,6,11,12|\rangle \equiv$$
$$\equiv \pi_{C^1}\langle|1,2,3,4\rangle \lhd \pi_{||}\langle 1,2|5,6\rangle \lhd \pi_{||}\langle 3,4|7,8\rangle \lhd \pi_{C^3}\langle 7,8|9,10\rangle \lhd$$
$$\lhd \pi_{||}\langle 9,10|11,12\rangle \lhd \pi_{X^2}\langle 5,6|\rangle \lhd \pi_{X^2}\langle 11,12\rangle$$

$$\pi_{C^1}\langle|1,2,3,4\rangle \lhd \pi_{|}\langle 1|5\rangle \lhd \pi_{|||}\langle 2,3,4|6,7,8\rangle$$
$$\lhd \pi_{C^4}\langle 6,7,8|9\rangle \lhd \pi_{|}\langle 9|10\rangle \lhd \pi_{-}\langle 5,10|\rangle \equiv$$
$$\equiv \pi_{C^1}\langle|1,2,3,4\rangle \lhd \pi_{|}\langle 1|5\rangle \lhd \pi_{|||}\langle 2,3,4|6,7,8\rangle \lhd$$
$$\lhd \pi_{C^4}\langle 6,7,8|9\rangle \lhd \pi_{|}\langle 9|10\rangle \lhd \pi_{X^1}\langle 5|\rangle \lhd \pi_{X^1}\langle 10|\rangle.$$

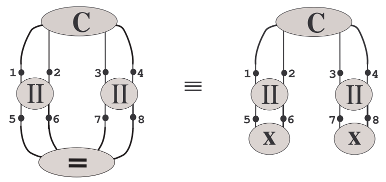


Figure 17: An example of process termination translations.

For all primitives $\pi\langle f\rangle$ satisfying $1,2 \in \mathrm{term}(\pi\langle f\rangle)$,

$$\pi\langle f\rangle \lhd \pi_{|}\langle 1|3\rangle \lhd \pi_{X^1}\langle 3|\rangle \lhd \pi_{|}\langle 2|4\rangle \lhd \pi_{X^1}\langle 4|\rangle \equiv$$
$$\equiv \pi\langle f\rangle \lhd \pi_{||}\langle 1,2|3,4\rangle \lhd \pi_{X^2}\langle 3,4|\rangle \quad \text{(Fig. 18).}$$

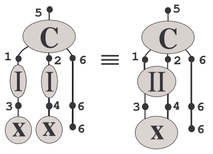


Figure 18: An example of process termination translations.

For all primitives $\pi\langle f\rangle$ satisfying $1, 2, 3 \in \mathrm{term}(\pi\langle f\rangle)$,

$$\pi\langle f\rangle \lhd \pi_{|}\langle 1|4\rangle \lhd \pi_{X^1}\langle 4|\rangle \lhd \pi_{|}\langle 2|5\rangle \lhd \pi_{X^1}\langle 5|\rangle \lhd \pi_{|}\langle 3|6\rangle \lhd \pi_{X^1}\langle 6|\rangle \equiv$$
$$\equiv \pi\langle f\rangle \lhd \pi_{\equiv}\langle 1,2,3|4,5,6\rangle \lhd \pi_{X^3}\langle 4,5,6|\rangle.$$

Together, all of the above identities ensure a "chemically correct" termination process.

Let $\mathcal{I}$ be the set of identities introduced above. The resulting inductive structure $(\Pi, \mathcal{I})$ will be called the **covalent bonding inductive structure**.

We next discuss the question of how to go from the conventional stereochemical representation of an organic compound to our pictorial struct representation. We present only one of the possible ways to accomplish this. [27] We begin by choosing arbitrarily the first carbon atom in the compound and depict it as the corresponding atomic primitive. We next enumerate its bonds following some chosen (fixed) enumeration process. [28] We, then, proceed to attach to the above carbon primitive, one atom at a time, (via the corresponding vertical bond primitives $\pi_{|}, \pi_{||}, \pi_{|||}$) the neighboring atoms [29] in the order just obtained for their bonds. Once all the neighboring atoms have been attached, we take the first neighbor-atom, enumerate all its bonds, that have not been previously enumerated, and, then, proceed as above. In Fig. 19 we present the structural chemical formula of **Maleic Hydrazide** [58, p.896] and its resulting pictorial representation in ETS model. For simplicity, the initial, birth (CB), primitive and the last, death (CD), primitive which are presented in all compounds, are omitted in the pictorial representation.

In connection with the above construction process, it is important to add that quite often one may need to consider as a *single step* simultaneous, or parallel, attachment of, for example, two atoms.[30] The latter two atoms may then, i.e. *at the next step*, need to be connected to each other by a covalent bond (see Fig. 20). It is to this end, that the "horizontal" bond primtypes, $\pi_{-}, \pi_{=}, \pi_{\equiv}$, were introduced: when the corresponding "horizontal" bond primitive is attached (via the two "vertical" bond primitives both of which are of the same type as the sought "horizontal" primitive) to the parallel attached previous pair of atomic primitives, this construction step signifies the creation of the corresponding "horizontal" bond between the two atoms.

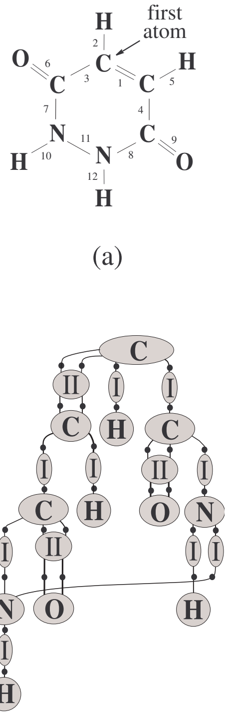
Also, in connection with the above construction process, we note that it is to be able to treat ions as "completed" structs that we introduced the termination primtypes $\pi_{X^1}, \pi_{X^2}, \pi_{X^3}$:

[27] We believe, however, that in time, when the structural predictions of the ETS model will be experimentally supported, more accurate methods of molecular representation will be adopted.

[28] We are using anticlockwise enumeration of the bonds but one can, of course, rely on any other fixed method for the enumeration, e.g. using the Chemical Abstract Name [58, p.viii] or, a more preferable method, based on the information related to the stepwise synthesis of the compound.

[29] By the neighbors of an atom in a compound we mean the atoms immediately connected to it.

[30] All levels of attachment must be clearly visible in the pictorial representation in the ETS model.



(a)

(b)

Figure 19: **Maleic Hydrazide** (*1,2-Dihydro-3,6-pyridazinedione*): (a) the standard structural chemical formula with enumeration of the bonds obtained by taking for each atom the anticlockwise ordering of its bonds; (b) the resulting pictorial representation in ETS.

the corresponding termination primitives are attached to the "free" bond primitives, in order to terminate the construction of the above ion. The termination primitive attached to the corresponding "vertical" bond primitive changes the meaning of the latter from its standard meaning to the "existence of the corresponding number of free electrons".

How do we go from the pictorial struct representation to the inductive expression for the appropriate composite? We first assign, either in an arbitrary manner or following some canonical method, the numeric labels to the external (initial and terminal) sites of the struct. Obviously, for molecules (in contrast to their fragments) the transition from the struct to the corresponding i-struct is not necessary since in this case the external sites are absent (Fig. 21(a),(b)). Then, again, either in an arbitrary or canonical manner, but now subject to the restrictions given in Def. 2, we assign the numeric labels to the internal sites (Fig. 21(c)). Finally, taking into consideration the order of primitives in the pictorial representation, we can write down the corresponding inductive expression, keeping in mind that the first and the last primitives in the expression are the compound's birth and death primitives. Thus, the

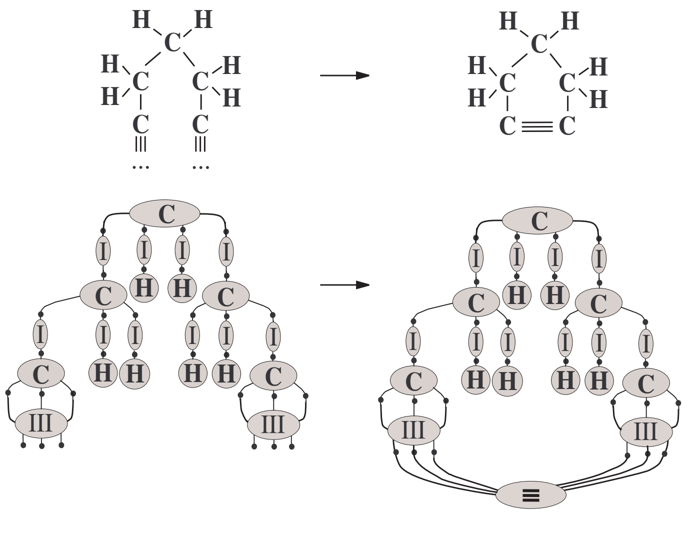


Figure 20: The process of horizontal bond formation: (a) the previous step; (b) the bond formation step.

inductive expression for the composite shown in Fig. 21(c) can be obtained as follows:

$$\begin{aligned}\gamma \;&= \pi_{CB}\langle\,|\,0\rangle \lhd \pi_{C^1}\langle\,|\,1,2,3,4\rangle \lhd \pi_{||}\langle 1,2\,|\,5,6\rangle \lhd \pi_{|}\langle 3\,|\,7\rangle \lhd \pi_{|}\langle 4\,|\,8\rangle\\ &\lhd \pi_{C^3}\langle 5,6\,|\,9,10\rangle \lhd \pi_{H^1}\langle 7\,|\,\rangle \lhd \pi_{C^2}\langle 8\,|\,11,12,13\rangle \lhd\\ &\lhd \pi_{|}\langle 9\,|\,14\rangle \lhd \pi_{|}\langle 10\,|\,15\rangle \lhd \pi_{||}\langle 11,12\,|\,16,17\rangle \lhd \pi_{|}\langle 13\,|\,18\rangle\\ &\lhd \pi_{C^2}\langle 14\,|\,19,20,21\rangle \lhd \pi_{H^1}\langle 15\,|\,\rangle \lhd \pi_{O^2}\langle 16,17\,|\,\rangle \lhd \pi_{N^1}\langle 18\,|\,22,23\rangle\\ &\lhd \pi_{|}\langle 19\,|\,24\rangle \lhd \pi_{||}\langle 20,21\,|\,25,26\rangle \lhd \pi_{|}\langle 22\,|\,27\rangle \lhd \pi_{|}\langle 23\,|\,28\rangle\\ &\lhd \pi_{N^2}\langle 24,28\,|\,29\rangle \lhd \pi_{O^2}\langle 25,26\,|\,\rangle \lhd \pi_{H^1}\langle 27\,|\,\rangle \pi_{|}\langle 29\,|\,30\rangle\\ &\lhd \pi_{H^1}\langle 30\,|\,\rangle \lhd \pi_{CD}\langle 0\,|\,\rangle.\end{aligned}$$

5.1.2 "Superclass" for the covalent bonding model

We now define a superclass **IC** in inductive structure $(\Pi, \mathcal{I})$, called the **superclass of covalent bonding compounds**.
The progenitor is a struct $\bar{\boldsymbol{\kappa}} = [[\kappa]]$, where

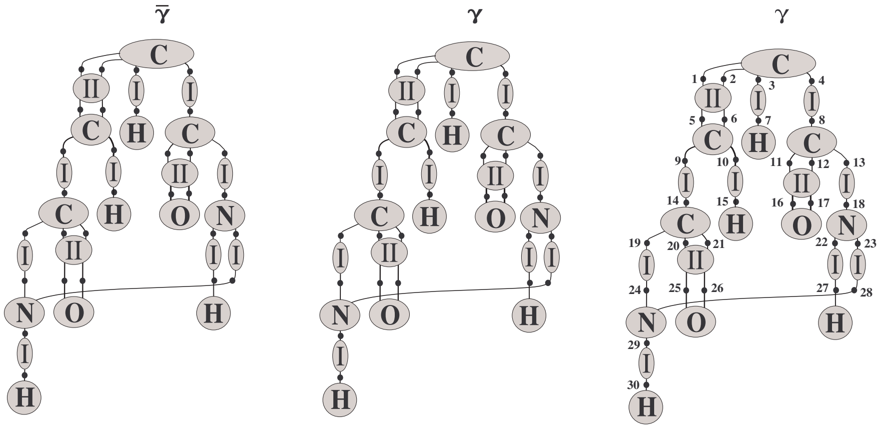
$$\kappa = \pi_{CB}\langle\,|\,0\rangle \lhd \pi_{C^1}\langle\,|\,1,2,3,4\rangle$$

is a representative from the equivalence class of composites. Let the elements of the following set J denote different (meaningful) labels of transformations that are defined below: [31]

$$J \;=\; \{C_1, \ldots, \mathbf{C}_{14}, H_1, Cl_1, Br_1, O_1, O_2, O_3, N_1, \ldots, N_7, -, =, \equiv, X_1, X_2, X_3, CD\}.$$

For the class **IC**, the set of transformations is $T = \{\bar{\boldsymbol{\tau}}_j,\ j \in J\}$, where $\boldsymbol{\tau}_j = (\bar{\boldsymbol{\alpha}}_j, \bar{\boldsymbol{\beta}}_j) = ([[\alpha_j]], [[\beta_j]])$ and $\alpha_j,\ \beta_j$ are the representatives of the corresponding equivalence classes. All

[31] A discussion of the transformations is postponed until the next section.



(a) (b) (c)

Figure 21: (a) A pictorial representation of a struct $\bar{\boldsymbol{\gamma}}$ (see Fig. 19) (b) A representative i-struct $\boldsymbol{\gamma}$ (c) A representative composite γ.

34 transformations and their contexts (shaded) are shown in Fig. 22. The inductive expressions for several of them are given next. [32]

$$\begin{aligned}
\alpha_{C_1} &= \pi_{C^1}\langle \mid 1,2,3,4\rangle \\
\beta_{C_1} &= \pi_{C^1}\langle \mid 1,2,3,4\rangle \triangleleft \pi_{|}\langle 1 \mid 5\rangle \triangleleft \pi_{|}\langle 2 \mid 6\rangle \triangleleft \pi_{|}\langle 3 \mid 7\rangle \triangleleft \pi_{|}\langle 4 \mid 8\rangle
\end{aligned}$$

$$\begin{aligned}
\alpha_{C_2} &= \pi_{|}\langle 1 \mid 2\rangle \\
\beta_{C_2} &= \pi_{|}\langle 1 \mid 2\rangle \triangleleft \pi_{C^2}\langle 2 \mid 3,4,5\rangle \triangleleft \pi_{|}\langle 3 \mid 6\rangle \triangleleft \pi_{|}\langle 4 \mid 7\rangle \triangleleft \pi_{|}\langle 5 \mid 8\rangle
\end{aligned}$$

$$\begin{aligned}
\alpha_{-} &= \pi_{|}\langle 1 \mid 2\rangle \triangleleft \pi_{|}\langle 3 \mid 4\rangle \\
\beta_{-} &= \pi_{|}\langle 1 \mid 2\rangle \triangleleft \pi_{|}\langle 3 \mid 4\rangle \pi_{-}\langle 2,4 \mid \rangle.
\end{aligned}$$

Finally, to complete the definition of a transformation system, we specify the values of two parameters for each transformation:

$l(\bar{\boldsymbol{\tau}}_j) = N$ $(j \in J,\ j \neq CD)$, N is the number of primitives in transformation $\bar{\boldsymbol{\tau}}_j$, $l(\bar{\boldsymbol{\tau}}_{CD}) = 10$;

for $\boldsymbol{\tau}_j = (\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j)$ $(j \in J,\ j \neq CD)$, $w(\bar{\boldsymbol{\tau}}_j) = a$, if $|\,\mathrm{term}(\boldsymbol{\alpha}_j)| > |\,\mathrm{term}(\boldsymbol{\beta}_j)|$, $w(\bar{\boldsymbol{\tau}}_j) = b$, if $|\,\mathrm{term}(\boldsymbol{\alpha}_j)| \leq |\,\mathrm{term}(\boldsymbol{\beta}_j)|$, where a and b, $a > b$, are positive constants such that the above transformation system satisfies the process termination condition, and $w(\bar{\boldsymbol{\tau}}_{CD}) = MIN$, where MIN is some very small positive number. [33] The weights are chosen to ensure that the larger the molecule (from the class) the less likely it is to be generated by the generating process.

Thus, we obtain a class. It is very important to stress that the proposed model allows one to capture in the class representation the typical size range for the molecules involved.

[32] The inductive expressions for others can be specified analogously.

[33] The very small weight of transformation $\bar{\boldsymbol{\tau}}_{CD}$ means that if for a struct $\bar{\boldsymbol{\gamma}}$ there is at least one transformation $\bar{\boldsymbol{\tau}}_j$ ($j \neq CD$) applicable to $\bar{\boldsymbol{\gamma}}$, then $\bar{\boldsymbol{\tau}}_{CD}$ is always chosen with probability almost zero by the generating process.

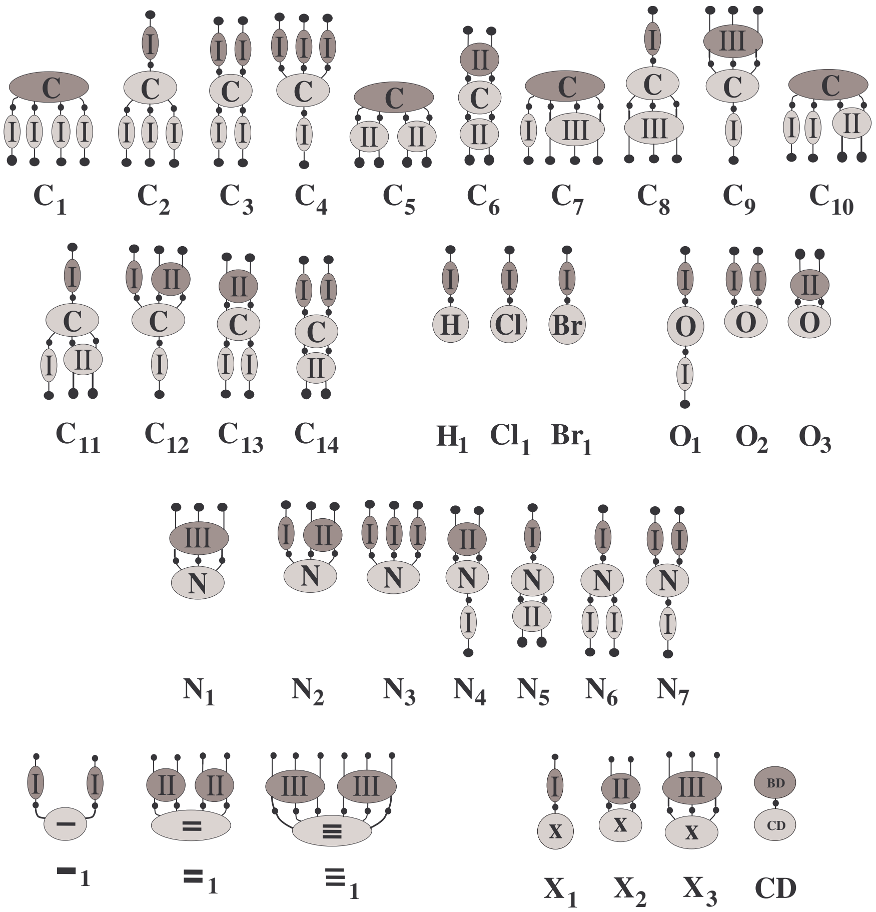


Figure 22: The transformation set for the superclass of covalent bonding compounds: atom transformations ($\bar{\boldsymbol{\tau}}_{C_1} - \bar{\boldsymbol{\tau}}_{N_7}$), horizontal bond transformations ($\bar{\boldsymbol{\tau}}_{-} - \bar{\boldsymbol{\tau}}_{\equiv}$), and the termination transformations ($\bar{\boldsymbol{\tau}}_{X_1} - \bar{\boldsymbol{\tau}}_{CD}$).

How do we go from the conventional stereochemical representation of an organic compound to its i-transformation expression (see section 4.2) in the superclass? Initially, following the two procedures described in section 5.1.1, we first construct our pictorial struct representation and, then, the corresponding inductive expression for the appropriate composite:

$$\gamma = \pi_{CB}\langle f_{CB}\rangle \lhd \pi_{C^1}\langle f\rangle \lhd \pi_{i_1}\langle f_1\rangle \lhd \ldots \lhd \pi_{i_n}\langle f_n\rangle \lhd \pi_{CD}\langle f_{CD}\rangle.$$

Next, we start the construction of the i-transformation expression with an i-struct $\boldsymbol{\kappa} \in [\boldsymbol{\kappa}] = \bar{\boldsymbol{\kappa}}$, followed by the i-transformation $\boldsymbol{\tau}_{C_i}\langle h_1\rangle$, where $\boldsymbol{\tau}_{C_i}$ is chosen from the superclass transformation set in Fig. 22 on the basis of all the "vertical" bond primitives attached to the atom primitive $\pi_{C_1}\langle f\rangle$ in the representation of compound and taken in the same order as in the inductive expression. We thus obtain the expression [34]

$$\boldsymbol{\kappa} \lhd \boldsymbol{\tau}_{C_i}\langle h_1\rangle.$$

Each of the following steps falls into one of the two categories: A) or B).

A) This is the case when (ignoring all "vertical" bond primitives) the first previously not processed primitive in the inductive expression is one of the "horizontal" bond primitives $\pi_{-}\langle f_j\rangle$, $\pi_{=}\langle f_k\rangle$, or $\pi_{\equiv}\langle f_j\rangle$. In this case, the next i-transformation (in the i-transformation expression being constructed) is $\boldsymbol{\tau}_{-}\langle h_m\rangle$, $\boldsymbol{\tau}_{=}\langle h_m\rangle$, or $\boldsymbol{\tau}_{\equiv}\langle h_m\rangle$ (see Fig. 22).

B) This is the case when, ignoring all "vertical" bond primitives, the first previously not processed primitive in the inductive expression is not the CD or the "horizontal" bond primitives considered in A). In this case, the next i-transformation is $\boldsymbol{\tau}_j\langle h_m\rangle$, where $\boldsymbol{\tau}_j\langle h_m\rangle$ is either atom or non-CD termination i-transformation (see Fig. 22). The corresponding transformation is chosen from the superclass transformation set in Fig. 22 on the basis of: all the previous "vertical" bond primitives to which the found primitive is attached *plus* all the following "vertical" bond primitives which are attached to the found primitive, and all are taken in the same order as in the inductive expression.

Finally, when all the non-CD primitives in the inductive expression have been processed, we add the CD i-transformation to the i-transformation expression. The role of CD i-transformation is to attest the death of the compound which, in fact, corresponds to its desintegration. In other words, an already formed compound "lives" while the generating class process is applying the CD transformation.
The following example illustrates the above algorithm. For a composite γ which corresponds to the struct representing **Maleic Hydrazide** (Figs. 19,21), based on the already obtained inductive expression, the i-transformation expression for a corresponding i-struct will be as follows: [35]

$$\begin{aligned} \boldsymbol{\gamma} \;&=\; \boldsymbol{\kappa} \lhd \boldsymbol{\tau}_{C_{10}}\langle\,|\,3,4,1,2\,;\,7,8,5,6\rangle \lhd \boldsymbol{\tau}_{C_{13}}\langle 1,2\,|\,5,6\,;\,14,15\rangle \lhd \boldsymbol{\tau}_{H_1}\langle 3\,|\,7\,;\,\rangle \lhd \\ &\lhd \boldsymbol{\tau}_{C_{11}}\langle 4\,|\,8\,;\,18,16,17\rangle \lhd \boldsymbol{\tau}_{C_{11}}\langle 9\,|\,14\,;\,24,25,26\rangle \lhd \boldsymbol{\tau}_{H_1}\langle 10\,|\,15\,;\,\rangle \lhd \\ &\lhd \boldsymbol{\tau}_{O_3}\langle 11,12\,|\,16,17\,;\,\rangle \lhd \boldsymbol{\tau}_{N_6}\langle 13\,|\,18\,;\,27,28\rangle \lhd \boldsymbol{\tau}_{N_7}\langle 19,23\,|\,24,28\,;\,30\rangle \lhd \\ &\lhd \boldsymbol{\tau}_{O_3}\langle 20,21\,|\,25,26\,;\,\rangle \lhd \boldsymbol{\tau}_{H_1}\langle 22\,|\,27\,;\,\rangle \lhd \boldsymbol{\tau}_{H_1}\langle 29\,|\,30\,;\,\rangle \lhd \boldsymbol{\tau}_{CD}\langle|0;\rangle. \end{aligned}$$

[34] Here (and below) the site assignment f and the site replacement(s) h_1 (and h_i) are defined in such a way that the i-struct constructed so far coincides with its counterpart in the original inductive expression.

[35] For an i-transformation $\boldsymbol{\tau} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ and site replacement $\mathbf{h} : \mathrm{ext}(\boldsymbol{\alpha}) \cup \mathrm{ext}(\boldsymbol{\beta}) \to \mathcal{S}$, the two separators $\langle \ldots \mid \ldots ; \ldots\rangle$ are used to separate the values of $\mathbf{h}(\mathrm{init}(\boldsymbol{\beta}))$, $\mathbf{h}(\mathrm{term}(\boldsymbol{\alpha}))$, and $\mathbf{h}(\mathrm{term}(\boldsymbol{\beta}))$.

5.1.3 Why does the covalent bonding superclass contain only chemically correct structs?

First, the choice of the progenitor $\bar{\boldsymbol{\kappa}}$ based on one of the C primtypes is easily explained by the fact that the carbon atom is the largest common part for *all* organic compounds. Moreover, the chosen transformations for the superclass (Fig. 22) do not allow the following chemically invalid constructions.

1) Incorrect valency (Fig. 23) is prevented by the presence of atom primtypes with the appropriate number of initial and terminal a-sites.

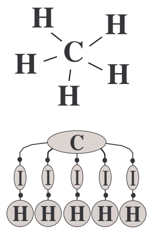


Figure 23: Chemically invalid construction: incorrect valency.

2) Invalid types of bonding (Fig. 24) are prevented by the presence of the appropriate bond primtypes.

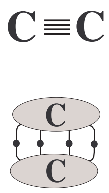


Figure 24: Chemically invalid construction: invalid type of bonding.

3) The attachment of a bond directly to another bond (Fig. 25) is prevented by the presence of the appropriate context in the atom transformations.

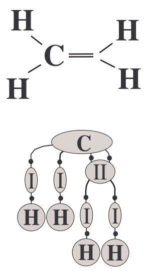


Figure 25: Chemically invalid construction: bond to bond attachment.

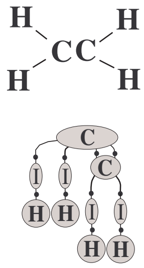


Figure 26: Chemically invalid construction: atom to atom attachment.

4) The attachment of an atom directly to another atom bypassing the bond (Fig. 26) is also prevented by the presence of the appropriate context in the atom transformations.

5) A bond from an atom to itself (Fig. 27) is prevented by the presence of the set of semantic identities for horizontal bond transformations.

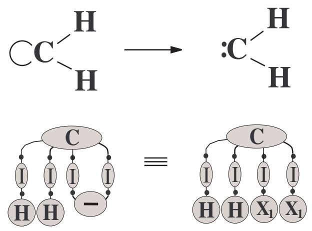


Figure 27: Chemically invalid construction: a bond from an atom to itself.

Why is the ETS model, in general, a convenient tool for modeling structural representations? In our experience, we found, for example, that all bond primtypes (including those of hydrogen bond) naturally suggested themselves when the questions related to the corresponding bonding constraints arose. In contrast, graph representation, in this respect, is not flexible.

5.2 An ETS model for the alkane class

In this section, we propose one possible *ETS representation for alkanes.* The class progenitor and the class transformations are pictorially represented in Fig. 28. We omit the inductive expressions for them, since they are constructed quite similarly to those for the covalent bonding superclass in section 5.1.2. We choose the following parameter values for the transformations (other choices are possible).[36]

$\bar{\tau}_i$	w_i	l_i
$\bar{\tau}_1$	a	5
$\bar{\tau}_2$	b	2
$\bar{\tau}_3$	MIN	10

[36]The explanation is similar to the one given in section 5.1.2.

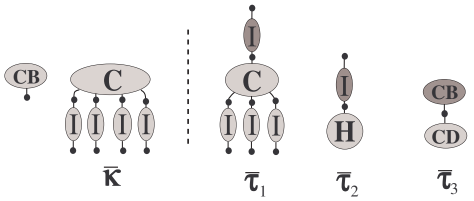


Figure 28: A proposed transformation system for the class of alkanes.

The positive constants a and b, $a < b$, are chosen in such way that the above transformation system satisfies the process termination condition (Def. 17) and thus, specifies a class. It is not hard to see that this is indeed the class of alkanes.

We next show an ETS representation of another class, the *linear alkanes*, i.e. alkanes with a linear skeleton. The corresponding progenitor and transformations are given in Fig. 29

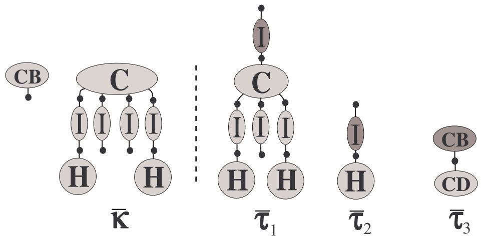


Figure 29: A proposed transformation system for the class of linear alkanes.

We choose the following parameter values for the transformations.

$\bar{\tau}_i$	w_i	l_i
$\bar{\tau}_1$	a	7
$\bar{\tau}_2$	b	2
$\bar{\tau}_3$	MIN	10

As in the case of alkane class, the positive constants a and b, $a < b$, are such that the above transformation system satisfies the process termination condition and thus specifies a class. Moreover, it is interesting to observe that the class progenitor and transformations can be expressed by the inductive i-transformation expressions in *each of the classes* - the covalent bonding superclass and the alkane class.

5.3 An ETS model for the hydrogen-covalent bonding

To simplify the exposition, we will restrict ourselves to atoms **C**, **O**, **H** only with the single, double, and triple covalent bonding plus the hydrogen bonding between **H** and **O**: **H**- - -**O**. However, it is obvious how to extend this inductive structure to the one with **N**- - -**H** and **F**- - -**H** bonds.

The primtypes for the above inductive structure are pictorially represented in Fig. 30. Note that the primtypes corresponding to the hydrogen and oxygen atoms now have one more site (see Fig. 13) reflecting their ability to form a hydrogen bond.

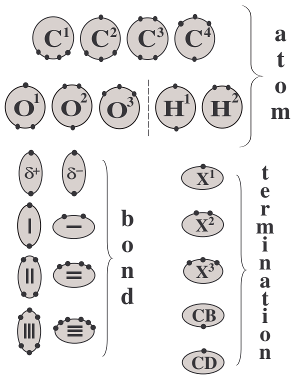


Figure 30: Primtypes for the hydrogen-covalent bonding inductive structure.

The transformation system for the hydrogen-covalent bonding superclass is described next. The progenitor is the same as that for the covalent bonding superclass. To define the transformation set, we first introduce two sets of indices for the transformations:

$$\begin{aligned} J_1 &= \{\, C_1, \ldots, C_{14}, -, =, \equiv, X_1, X_2, X_3, CD \,\} \\ J_2 &= \{\, H_1, \ldots, H_3, O_1, \ldots, O_8, X_4, X_5 \,\}. \end{aligned}$$

The set of transformations $\{\bar{\tau}_j,\, j \in J_1\}$ is the same as that for the covalent bonding superclass (Fig. 22). The additional, or "new", transformations $\{\bar{\tau}_j,\, j \in J_2\}$ are pictorially represented in Fig. 31. They are "explained" by our wish to exclude the following types of hydrogen bonds:

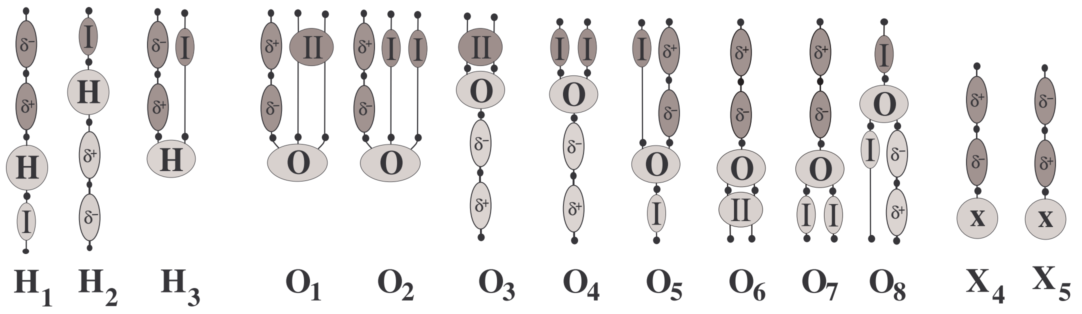


Figure 31: The additional transformations for the hydrogen-covalent bonding superclass.

H---**H**, **O**---**O**, **H**---**C**, **O**---**C**.
To complete the definition of a transformation system, we specify the two parameters for each transformation:

$l(\bar{\tau}_j) = N \quad (j \in J,\ j \neq CD)$, $\quad N$ is the number of primitives in transformation $\bar{\tau}_j$, $l(\bar{\tau}_{CD}) = 10$;

for $\boldsymbol{\tau}_j = (\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) \quad (j \in J \setminus \{ H_1, O_6, O_7, CD \})$,

$$w(\bar{\boldsymbol{\tau}}_j) = a, \text{ if } |\operatorname{term}(\boldsymbol{\alpha}_j)| > |\operatorname{term}(\boldsymbol{\beta}_j)|, \ w(\bar{\boldsymbol{\tau}}_j) = b, \text{ if } |\operatorname{term}(\boldsymbol{\alpha}_j)| \le |\operatorname{term}(\boldsymbol{\beta}_j)|,$$

where a and b, $a > b$, are positive constants such that the above transformation system satisfies the process termination condition, for $\boldsymbol{\tau}_j$, $(j \in \{ H_1, O_6, O_7 \})$, $w(\bar{\boldsymbol{\tau}}_j) = b/2$, and $w(\bar{\boldsymbol{\tau}}_{CD}) = MIN$, where MIN is some very small positive number.

One should note that the formal model itself allowed us *quite naturally* to account for the *different likelihood of formation of the two types of bonding*: the weights of the transformations can faithfully represent this difference, and this, in turn, manifests itself in the greater likelihood of appearance of the "covalent bond" structs during the process of struct generation (see Def. 16).

6 Advantages of the proposed model

In this section we, first, briefly outline distinctive features of the above model and, then, briefly discuss a considerable streamlining of the appropriately modified CADD processes as well as the proposed reorganization of the entire structure.

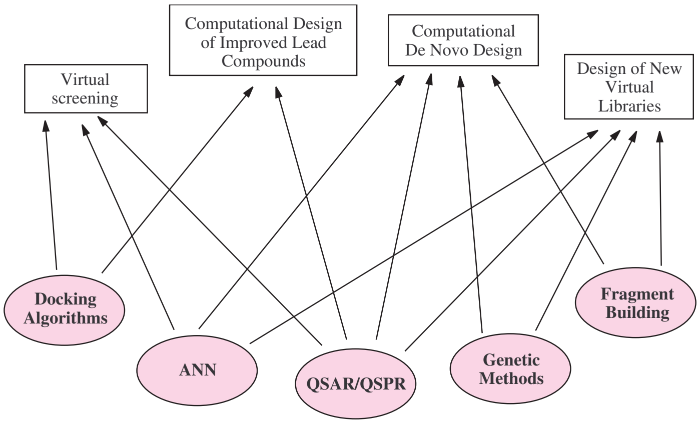


Figure 32: The basic CADD problems (top) and some of the popular methods for their solution.

First of all, we note that the proposed ETS model possesses all the desirable features discussed in section 3.3, including a new and very important feature — "evolutionary" form of structural representation.[37] Moreover, as was mentioned above, it appears that, as far as the formal concept of representation is concerned, its "evolutionary" feature and its "structural" feature appear to be synonymous. In other words, we strongly believe that these two features of representation are inseparable. It is these features that allow one to treat all molecular representations, including those of the ligand and the receptor, in the same manner.
Next, we observe that the proposed model should bring complete uniformity into the various formalisms employed in CADD (Figs. 32, 33). We suggested that this can be achieved by first focusing on the class learning problem as the central CADD problem (see section 2.2), and then approaching various problems on the basis of the constructed structural class representation.

[37]As was mentioned in section 3.2, it is important to emphasize the fundamental difference between the ETS *model* and the genetic *methods* (see the last paragraph in the discussion of genetic method in section 3.2).

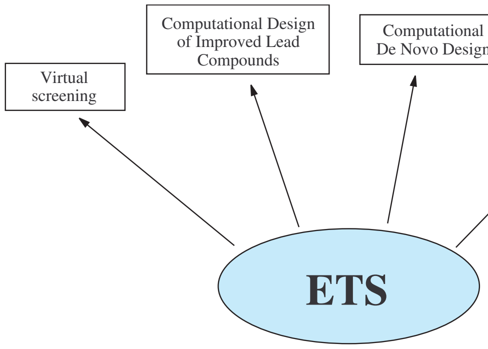


Figure 33: The proposed centralization of CADD processes.

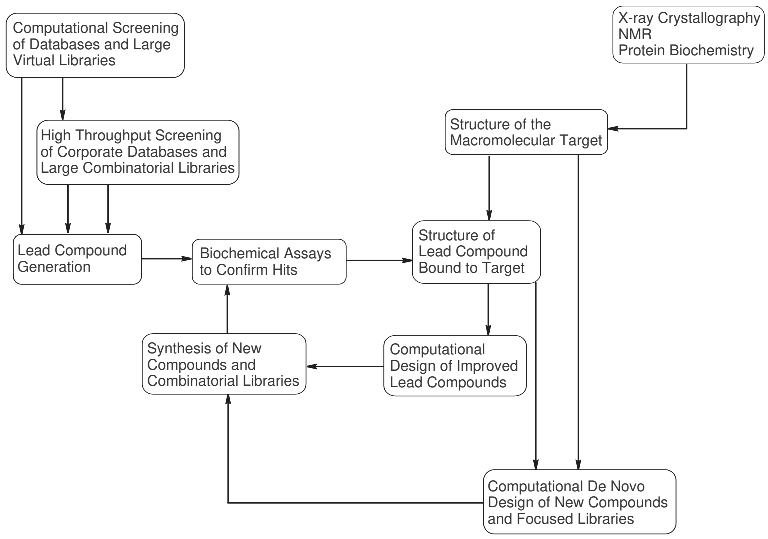


Figure 34: The current structure-based drug design process (adopted from [50]).

This becomes possible since the ETS model, for the first time, offers the framework for an inductive — i.e. based on a small training data set — construction of the structural class representation.

Finally, Figs. 34 & 35 show the proposed radical streamlining of the structure-based drug design process. This is accomplished by consolidating the basic CADD problems and changing the role of CADD from an auxiliary to the basic intelligent (and interactive) tool.

7 Conclusion

We proposed to view the appropriately formulated class learning problem as the central problem of CADD, on the basis of which other problems of CADD should be approached.

We discussed the concept and the desirable features of structural representation as well as the inadequacies of several basic models for representation and classification used in CADD.

Most importantly, the *first* formal model for structural representation and classification in life

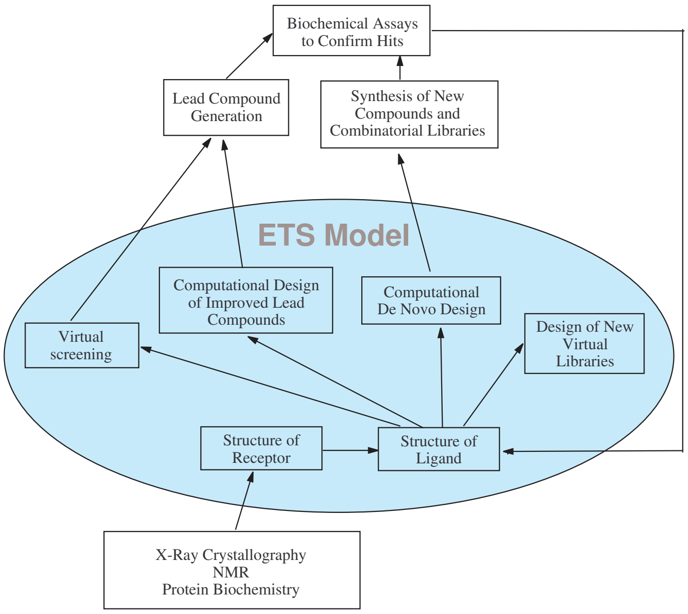


Figure 35: The proposed modification of structure-based drug design process.

sciences — evolving transformation system (ETS) model — incorporating all of the desirable features discussed above was outlined. It was proposed that a considerable simplification and streamlining of CADD processes could be achieved by putting the ETS model at the core of CADD. Moreover, setting CADD processes on sure footing of a reliable structural representation should catapult their role not just to the top of drug design (DD) but also to the top of organic chemistry and biochemistry as well. Why? Because CADD would be the first area of science to clarify the role of structural as opposed to the classical, numeric, forms of representation.

We are currently developing the software as well as preparing the patent related to the application of the proposed model to CADD. A concrete problem we are planning to work on in the nearest future is the description of a concrete class of drugs based on a small set of its representatives. This, in particular, would allow the construction of the "new" most typical representatives from the class. Some of the interesting future applications of the ETS model are related to the design of focused virtual libraries, protein representation and classification, the complex ligand-receptor representation and classification, and modeling of chemical reactions.

8 Acknowledgments

We would like to thank Ghislain Deslongchamps for a number of useful discussions related to organic chemistry.

References

[1] Goldfarb, L.; Golubitsky, O.; Korkin, D. What is structural representation? *Tech. Rep. No TR00-37*, Faculty of Computer Science, University of New Brunswick, **2000**.

[2] Goldfarb, L.; A New Approach To Pattern Recognition. In *Progress in Pattern Recognition 2;* Kanal, L.N.; Rosenfeld, A.; Eds; North-Holland Publ. Comp.: 1985, 241-402.

[3] Goldfarb, L. On Foundations of Intelligent Processes I: An Evolving Model for Pattern Learning. *Pattern Recognition,* **1990,** *23,* 595-616.

[4] Goldfarb, L.; Nigam, S. The Unified Learning Paradigm: A Foundation for AI. In *Artificial Intelligence and Neural Networks: Steps Toward Principled Integration;* Honavar, V.; Uhr, L.; Eds; Academic Press: Boston, 1994.

[5] Goldfarb, L.; Abela, J.; Bhavsar, V.C.; Kamat, V.N. Can a Vector Space Based Learning Model Discover Inductive Class Generalization in a Symbolic Environment? *Pattern Recognition Letters,* **1995,** *16,* 719-726.

[6] Panchen, A.L. *Classification, Evolution, and the Nature of Biology;* Cambridge University Press: New York, 1992.

[7] Brown, T.L.; LeMay, H.E., Jr.; Burstem, B.E. *Chemistry: the Central Science.* Prentice-Hall, Inc: Englewood Cliffs, 1994.

[8] Goldfarb, L.; Inductive Class Representation and Its Central Role in Pattern Recognition. *Proceedings of the 1996 International Multidisciplinary Conference on Intelligent Systems, NIST,* Albus, J.; Meystel, A.; Quintero, R.; Eds; U.S. Government Printing Office: Washington, 1996, v.1, 53-58.

[9] Goldfarb, L.; Deshpande, S. What is a symbolic measurement process? *Proceedings 1997 IEEE Conf. Systems, Man, and Cybernetics;* IEEE Press, v.5, 4139-4145.

[10] Goldfarb, L.; Hook, J. Why Classical Models for Pattern Recognition Are Not Pattern Recognition Models. *Proceedings of the International Conference on Advances in Pattern Recognition;* Plymouth, UK, Singh, S.; Ed; Springer: London, 1999, 405-414.

[11] Leach, A. R. *Molecular Modeling. Principles and Applications;* Longman: Essex, England, 1996.

[12] Hansch, C.; Muir, R.M.; Fujita, T.; Maloney, P.P.; Geiger, E.; Streich, M. The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. *J. Am. Chem. Soc.* **1963,** *85,* 2817-2824.

[13] Hansch, C. A Quantitative Approach to Biochemical Structure-Activity Relationships. *Accounts of Chemical Research* **1969,** *2,* 232-239.

[14] Malcolm J.M.; Muskal, S.M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 569-574.

[15] Lin, T.H.; Yu, Y.S.; Chen, H.J. Classification of Active Compounds and Their Inactive Analogues Using Two Three-Dimensional Molecular Descriptors Derived from Computation of Three-Dimensional Convex Hulls for Structures Theoretically Generated for Them. *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 1210-1221.

[16] Carhart, R. E.; Smith, D.H.; Venkataraghavan R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1984,** *25,* 64-73.

[17] Filimonov, D.; Poroikov, V.; Borodina, Y.; Gloriozova, T. Chemical Similarity Assesment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 666-670.

[18] Palyulin, V.A.; Radchenko, E.V.; Zefirov, N.S. Molecular Field Topology Analysis Method in QSAR Studies of Organic Compounds.*J. Chem. Inf. Comput. Sci.* **2000,** *40,* 659-667.

[19] Ivanciuc, O. QSAR Comparative Study of Wiener Descriptors for Weighted Molecular Graphs *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 1412-1422.

[20] Agatonovic-Kustrin S.; Beresford R. Basic Concepts of Artificial Neural Network (ANN) Modeling and Its Application in Pharmaceutical Research. *J. Pharm. Biomed. Anal.* **2000,** *22,* 717-727.

[21] Wu, C.H. Artificial Neural Networks for Molecular Sequence Analysis. *Computers Chem.* **1997,** *21,* 237-256.

[22] Agrafiotis D. K.; Lobanov, V.S. Nonlinear Mapping Networks. *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 1356-1362.

[23] Bahler, D.; Stone, B.; Wellington, C.; Bristol, D.W. Symbolic, Neural, and Bayesian Machine Learning Models for Predicting Carcinogenicity of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 906-914.

[24] Frimurer, T.M.; Bywater, R., Nærum, L.; Lauritsen, L.N.; Brunak, S. Improving the Odds in Discriminating "Drug-like" from "Non Drug-like" Compounds. *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 1315-1324.

[25] Burden, F.R. Using Artificial Neural Networks to Predict Biological Activity from Simple Molecular Structural Considerations. *Quant. Struct.-Act. Relat.* **1996,** *15,* 7-11.

[26] Bienfait, B. Applications of High-Resolution Self-Orginizing Maps to Retrosynthetic and QSAR Analysis.;*J. Chem. Inf. Comput. Sci.* **1994,** *34,* 890-898.

[27] King,R.D.; Hirst, J.D.; Stenberg, M.J.E. New Approaches to QSAR: Neural Networks and Machine Learning. Perspect. *Drug. Discov. Des.* **1993,** *1,* 279-290.

[28] Wilson, R.J.; Watkins, J.J. *Graphs: an Introductory Approach;* John Wiley& Sons,Inc: New York, 1990.

[29] Guevara, N. Fragmental Graphs. A Novel Approach to Generate a New Family of Descriptors. Applications to QSPR Studies. *J. Mol. Struc. (Theochem)* **1999,** *493,* 23-26.

[30] Randic, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975,** *97,* 6609-6615.

[31] Kier, L.B.; Hall, L.H. *Molecular Connectivity in Chemistry and Drug Research;* Academic Press: New York, 1976.

[32] Kier, L.B.; Hall, L.H. Intermolecular Accessibility: The Meaning of Molecular Connectivity. *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 792-795.

[33] Basak, S.C.; Nikolić, S.; Trinastić N. QSPR Modeling: Graph Connectivity Indices versus Line Graph Connectivity Indices *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 927-933.

[34] Randic, M.; Brissey, G.M.; Spencer, R.B.; Wilkins, C.L. Using Self-Avoiding Paths for Molecular Graphs with Multiple Bonds. *Comput. Chem.* **1980,** *4,* 27-43.

[35] Leach, A.R.; Dolata, D.P.; Prout, K. Automated Conformational Analysis and Structure Genration: Algorithms for Molecular Perception *J. Chem. Inf. Comput. Sci.* **1990,** *30,* 316-324.

[36] Willett,P. Genetic Algorithms in Molecular Recognition and Design. *Trends Biotechnol.* **1995,** *13,* 516-521.

[37] Globus, A.; Lawton, J.; Wipke, T. Automatic Molecular Design Using Evolutionary Techniques. *Nanotechnology* **1999,** *10,* 290-299.

[38] Nachbar, R.B. Molecular Evolution: Automated Manipulation of Hierarchical Chemical Topology and Its Application to Average Molecular Structures *Gen. Prog. Evol. Mach.* **2000,** *1,* 57-94.

[39] Pullan, W.J. Genetic Operators for a Two-Dimensional Bonded Molecular Model *Computers Chem.* **1998,** *22,* 331-338.

[40] Lukovits, I. Isomer Generation: Syntactic Rules for Detection of Isomorphism. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 563-568.

[41] Lukovits, I. Isomer Generation: Semantic Rules for Detection of Isomorphism. *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 361-366.

[42] Bytautas, L.; Klein, D.J. Alkane Isomer Combinatorics: Stereostructure Enumeration and Graph-Invariant and Molecular-Property Distributions. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 803-818.

[43] Skvortsova, M.I.; Baskin, I.I.; Skvortsov, L.A. Palyulin, V.A.; Zefirov, N.S.; Stankevich, I.V. Chemical Graphs and Their Basis Invariants *J. Mol. Struc. (Theochem)* **1999,** *466,* 211-217.

[44] Debska, B.; Guzowska-Świder. Fuzzy Definition of Molecular Fragments in Chemical Structures. *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 325-329.

[45] Estrada E. Characterization of 3D Molecular Structure *Chem. Phys. Let.* **2000,** *319,* 713-718.

[46] Gane, P.J.; Dean, P.M. Recent Advances in Structure-Based Rational Drug Design *Curr. Opin. Struct. Biol.* **2000,** *10,* 401-404.

[47] Kuntz, I.D. Structure-Based Strategies for Drug Design and Discovery *Science* **1992,** *257,* 1078-1082.

[48] Baxter, C.A.; Murray, C.W.; Waszkowycz, B.; Li, J.; Sykers, R.A.; Bone, R.G.A.; Perkins, T.D.J.; Wylie, W. New Approach to Molecular Docking and Its Application to Virtual Screening of Chemical Databases. *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 254-262.

[49] Wang,X.; Wang, J.T.L. Fast Similarity Search in Three-Dimensional Structure Databases *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 442-451.

[50] Joseph-McCarthy, D. Computational Approaches to Structure-Based Ligand Design. *Pharm. Ther.* **1999,** *84,* 179-191.

[51] So, S.S.; Karplus, M. Evolutionary Optimization in Quantitative Structure-Activity Relationships: An Application of Genetic Neural Networks. *J. Med. Chem.* **1996,** *39,* 1521-30.

[52] Gardiner, E.J.; Willett, P.; Artymiuk, P.J. Graph-Theoretic Techniques for Macromolecular Docking. *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 273-279.

[53] Xue, L.; Bajorath, J. Molecular Descriptors for Effective Classification of Biologically Active Compounds Based on Principal Component Analysis Identified by a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 801-809.

[54] Holliday J.D.; Willett, P. Using a Genetic Algorithm to identify Common Structural Features in Sets of Ligands. *J. Mol. Graphics Mod.* **1997,** *15,* 221-232.

[55] Gutman, I. Selected Properties of the Schultz Molecular Topological Index. *J. Chem. Inf. Comput. Sci.* **1994,** *34,* 1087-1089.

[56] Liu, S.; Cai, S.; Li, Z. Molecular Electronegative Distance Vector (MEDV) Related to 15 Properties of Alkanes. *J. Chem. Inf. Comput. Sci.* **2000,** *40,* 1337-1348.

[57] Pauling, L.; Pauling, P. *Chemistry;* W.H.Freeman and Company: San Francisco, 1975.

[58] *The Merck Index: an Encyclopedia of Chemicals, Drugs, and Biologicals;* Budavari, S.; Ed. Merck&Co., Inc: Rahway, 1989.