Hints for a formal language inspired by Lewis structures

Savino Longo

Summary
In this work we elaborate on the idea of a formal theory for a limited but important part of structural chemistry, that described by Lewis’ methods and VSEPR (Valence Shell Electron Pair Repulsion). For this purpose, recursive functions and propositional functions are defined, that apply to formal expressions of the structure, based on a finite set of symbols. This approach allows for the expression of numerous questions of chemical interest. The formalization of basic structural chemistry based on Lewis/VSEPR method is potentially useful for the automation of the related procedures, but possibly also of some use as complementary material in teaching and as a heuristic tool in structural chemistry.

Keywords Formal languages · Structural chemistry · Lewis structures · VSEPR

Introduction
In discussions on the foundations of chemistry and also in chemistry teaching, the fact that chemical language is an autonomous and autonomously predictive language is often underestimated. A century of attempts to deduce, in some cases effectively, the chemical properties starting from the laws of quantum mechanics and thermodynamics, have produced the impression that chemistry is a physics-based superstructure.

At the same time, an examination of what is considered innovative in chemistry seems to suggest rather than at the root of calculations using the numerous force field or ab initio packages now available we find combinatorial and geometrical idea based on the possibilities offered by shared rules.

In this perspective, it may be useful to pursue the formulation of formal languages (Hunter 1971, Casari 1960, Lemmon 1971) based on the practice of basic but still predictive structural chemistry.
The attitude which is produced by such an effort may help to fully recognize the existence of the so-called combinatorial barrier (Scott 1995) between chemistry and physics: the fact that chemical language is based on the relations between objects, relations to be formalized, which are autonomous, and which are independent from underlying laws.

As a case study, the language of Lewis structures is taken into consideration, which, as is well known, allow for many concrete evaluations of the structure and properties of molecules.

This work is an effort, hopefully of some use, to pursue to some extent the formulation of a formal language inspired by Lewis structures, with an attempt to attain some level of rigor.

While many of the ideas and techniques proposed are provisional and need to be improved, the author believes that the work presented here, in addition to being useful in attracting attention in the direction of an interesting research perspective, can also be so for the considerable amount of work undertaken that can show the difficulties of specific points and also help with providing some solutions.

**Lewis’ method as a formal language**

As is known, in mathematical logic, formal languages (Hunter 1971) are sets of abstract rules, which allow to write sequences based on a limited number of symbols (for example, “1”, “#”, “!”). The sequences of symbols obtained by applying the rules are called “well-formed expressions” or “valid expressions”. An example could be: “the language includes only “x” and “y” as symbols, and only sequences starting with “xx” are valid”. So “xxyy” is a valid expression in this language.

A formal language may also include a series of so-called syntactic rules that allow to obtain well-formed expressions starting from other well-formed expressions of symbols. An example could be: “in a well-formed expression, every time “xx” appears it can be replaced by “xxy” “. It can be seen that this rule always produces another well-formed expression from a well-formed expression.

It is then possible at times to define a so-called *semantics*, which is the operation of parsing a formal language using expressions in another formal language (Casari, 1960). For example: “an expression containing the symbol “xx” is called “red”, otherwise it is called “black” “.

Lewis’ method is treated, even in elementary chemistry, to some extent as a formal language.

Consider for example the rule: “In a structure, an atom of an element of the second group can have around it no more than 8 valence electrons: the single bonds contribute to this count for a value of 2””. This is an example of rule for a well-formed expression (the structure).

Again, consider the way the Lewis structure is used (Dickerson et al. 1979) in the method for the determination of the oxidation number (O.N.) of the elements, which has some characteristics of a semantics, as opposed to the method based on a list of cases (e.g. the oxidation number of O is -2 unless in species where O is bond with O or with F, *et cetera*) and the algebraic matching of the species’ charge where any O.N. contributes to this charge (e.g. Cl in ClO$_3^-$ being +5 because of this rule and the last one). The oxidation numbers are thus redefined, based on the Lewis structures, by assigning the binding electrons to the bound
atom with the highest electronegativity. In this way the O.N., previously defined according to a set of rules, is reinterpreted on the basis of a different model: namely Lewis’ structure, and indeed the set of rules itself can be interpreted this way.

A Lewis structure, once correctly completed, also allows us to formulate hypotheses for the structure of the chemical environment of an atom, using the technique called VSEPR (Valence Shell Electron Pair Repulsion, Gillespie et al. 2013), based on the idea of repulsion between electrons and on the exclusive role of external, or valence electrons of the atom.

However, once the rules of the VSEPR method are mastered, they can be applied to a large extent mechanically, without necessarily referring to the concept of repulsion between electrons.

This way, it is possible to answer questions such as:

“How to create a molecular system with atoms at the top of a square-based pyramid or at the top of a tetragonal anti-prism?”

“What is the angle formed by a group of three AOB atoms?”.

Nevertheless, this state of affairs, while promising, hardly constitutes a rigorous formal language.

The author therefore posed the problem of tracing a path for the formulation of an actual formal language that has the predictive power of the techniques mentioned above, but the additional advantage of abstraction from the context, and resolution of ambiguities.

The path is neither short nor straight, as we can already see from the fact that the structural formulas used by chemists are not linear sequences of symbols, and provided these sequences are made available, the formation and syntactic rules will have to be properly defined.

**A first attempt at a formal language based on Lewis structures and VSEPR**

**Primitive concepts**

3.1 The atom, in this context, is better considered a primitive concept.

The bond between two atoms is also a primitive concept: it is caused by the effect of valence electrons. These, originally electrons from individual atoms, are pooled between two atoms.

The fact that for some simple systems, such as the molecule the H₂ molecule, a physical and quantitative interpretation of the bond is possible, is usually not relevant in the practice of building chemical structures. We therefore return to basic notions having in view the formulation of a formal language that is adequate for the evaluation, and perhaps the automatic discovery, of structures of chemical interest. From an empirical point of view, a bond is characterized by a significant increase in the negative charge between the two atoms and requires some energy to be dissolved.

Most of the assertions relating to the chemical structure, at least at the basic level, can be solved on the basis of two methods: the Lewis method and VSEPR method.
On the basis of this knowledge, it is possible to state that the fundamental notion of structural chemistry is the assignment to each chemical element of an appropriate group of the periodic system: this assignment is univocal and unambiguous, as has been known for more than a century.

We can therefore formalize this notion by stating that to every element, and therefore to every atomic number \( Z \), ex. \( Z(\text{“C”}) = 6 \), corresponds to a group number \( g \).

\[ g \] is therefore a function of \( Z \), \( g(Z) \).

This function, \( g(Z) \), can be considered as the definition of the periodic system of the elements.

Hence \( g(\text{“C”}) = g(6) = 4 \).

**Note** to reduce technical difficulties, the “old” group naming convention is used, so \( g(6) \) is 4, not 14.

This function is assigned once and for all and largely defines the chemical and structural properties. Its name derives from its characteristic trend, which is almost everywhere monotonous with unit increments, interrupted by sudden drops.

The periodic system is therefore nothing more than the \( g \) function. This function can obviously be represented from the graphical point of view in many ways, well known to chemists, like tables of various types, but it remains clear that the essential aspect is existence and the assignment of the function.

Moreover, the assignment of an appropriate \( g \) function allows the definition of chemical properties in extreme conditions. It is now known, for example, from abundant literature, that the \( g \) function can and must be modified to account for the structural properties in conditions of high pressure (Connerade 2020).

Since the Lewis structure is normally considered by chemists to be more of a teaching technique than an active research technique, some details need to be specified more clearly, in a formal way, in order to obtain an authentic unambiguous language based on the Lewis structure.

**Molecular expressions**

3.2 The formulation of a formal structural language requires the proposal of appropriate conventions. It is probably not the case to try to formulate what, *a posteriori*, may prove to be the most effective proposal now, in this very preliminary moment. However, it is possible to imagine how such a language could work.

A preliminary proposal and with a limited number of symbols, it may still include the traditional symbols of the chemical elements. This means that a symbol like “C” can be used in conjunction with a lowercase letter to indicate for example “Cl”.

We will now define several kinds of “expressions”. The expressions are words, written by using a limited set of symbols. The symbol set of our formal language includes only uppercase letters, a few lowercase letters, and as we shall see, the opening “(“ and closing “)” parentheses, the comma “,”, the signs “+” and “-”.

An example of expression is therefore “+,,+S+(“.
We will address expression by uppercase letters like A, B, C, X, Y and if necessary, we will use Y, X’, X” as symbols for other expressions. Note that X, X’ are not symbols of our language, but rather symbols we use in descriptions of it. This is also true for the apostrophe and the quotation mark, and the symbols “A”, “B”, etc. which we will also use to describe parts of expressions.

Technically, they would be named “symbols of the metalanguage” (Casari, 1960) where the metalanguage is a language that we use to describe the formal language under development.

Now, of course, “+,” “S” is not an expression of chemical interest. We must therefore build up formal rules for the construction of valid expressions.

We can now fix a first rule:

R1 Upper, and lower case, letters can only be used to form symbols of chemical elements, like “C”, “Cl”, “N”. We call these “elementary symbols”, ES.

So “Si” is an ES, but “SI”, “CH” or “Cl(“ are not.

We must now introduce the possibility of electric charge, symbolically.

We define a symbol &, again not a symbol of the language, which express the merging of two expressions, e.g.: “C”&“(“ is the same as “C(“.

We proceed now introducing recursive construction rules, so that a well-written structure can in turn be elaborated to produce a more complex well-written structure. See for example Casari, 1960 or Lemmon, 1971, in particular for this last the “formation rules” section in Chapter II.

To do this, we introduce an “atomic symbol” AS, according to these rules.

- If X is an ES, X&”+” is an AS.
- If X is an ES, X&”-” is an AS.
- if X&”+” is an AS, X&”++” is an AS.
- if X&”-” is an AS, X&”--” is an AS.

Therefore, “Si+++” is an AS, but “Si+-” is not.

Let us now define “molecular expression”, ME, a string like

A(B,C,…)

where A is an ES, while B, C, can be AS’s or ME’s as well.

Numbers can be used in shortcut representation of expression above: a symbol like “3H” is used as a shortcut for “H, H, H”, only for readability scope, not really as part of an expression. Numbers are not part of our set of symbols.

This recursive definition implies that a generic ME has a branched structure.

A(B(C(D,…),…),E(F,…),G,…).

where, now, A, B, … are all AS’s.

An example is

Cl++(H,Si(3H,C(O)))

The interpretation of a molecular expression is the fact of being the linear expression of a Lewis structure, traditionally realized as a connectivity structure drawn on a plane: the symbol of an element is followed by parentheses, and within the parentheses, separated by commas, there are chemical groups linked to this central element. Using nested parenth-
ses, we can express branched structures of any complexity. Not cycles, however, which are mentioned in Sect. 4.2 later.

Many complex structures are, however, within the grasp of this preliminary and limited language, for example this may be checked to be the expression of GABA (γ-aminobutyric acid):

\[ N(2H, C(2H, C(2H, C(2H, C(O, O(H))))) ) \]

These expressions can perhaps be made more readable by introducing a rule according to which the parentheses left open at the end of the expression are implicitly closed. No ambiguity is introduced this way. Specifically, a symbol like the following is obtained:

\[ N(2H, C(2H, C(2H, C(2H, C(O, O(H))))))) \]

However, this notation will not be used in this manuscript.

**Connectively equivalent expressions**

3.3 An essential aspect is the chemical equivalence of the expressions. For example, consider the following two structures expressions, which represent the same Lewis structure: “C(3H, C(3H))”, “H(C(2H, C(3H)))”.

A formalized language must be able to handle this potential ambiguity by highlighting the structural equivalence of two linear formulas, but without converting these to their traditional representations as pictures or using the human ability to read them in different ways.

We must instead be able to identify two seemingly different linear structures that must represent the same molecule using only symbol manipulations: for this purpose, we start by defining the equivalence of two expressions in the following terms.

As a preliminary step: we accept to neglect the order in which ME’s appear within parentheses, separated by commas. Of course, this solution is inadequate for chiral molecules, see later. We assume this, in fact, on a provisional basis.

Next, we define the *adjacency* of two expressions, as an intermediate step for equivalence in general.

The connection scheme:

B - A - C - D.

suggests the following definition: two linear formulas X and Y are defined adjacent, X.ad.Y if they have the form:

\[ X = "A(B, C(D))". \]
\[ Y = "C(D, A(B))". \]

where we have highlighted, by boldfaced font, the “A” and “C” groups which are essential in the transformation. In other words, adjacency is nothing more than the transposition, in this formal language, of a geometric concept: two groups of atoms are linked to the same central atom, the atom that appears before the opening parenthesis of the ME. Adjacency is an abstract change of the linear formula, that transforms the initial AS of a following opening parenthesis after a ME, into the first AS of the ME. In a sense the formal expression is put inside out, like an abstract rotation.
Hints for a formal language inspired by Lewis structures

Two expressions, X and Y are equivalent, X.eq.Y if they are connected by a chain of adjacency expressions:

\[ X \text{eq}.Y = df X_{ad}.X_1, X_1_{ad}.X_2, \ldots, X_n_{ad}.Y. \]

Two equivalent expressions describe the same molecule from the point of view of connectivity, or topology. If chiral centers never play the role of A or C atom, they actually describe the same molecule.

We apply the principle to the Lewis structure of acetic acid.

“C(3H, C(O, O(H)))” .ad. “C(O, O(H), C(3H))”.

“C(O, O(H), C(3H))” .ad. “O(C(O(H), C(3H)))”.

Where we have highlighted the atoms corresponding to “A” and “C” in the application of the adjacency rule above.

Therefore, these two expressions are equivalent:

“C(3H, C(O, O(H)))” and “O(C(O(H), C(3H)))”

In fact, it can be seen, by “unfolding” them, that both represent the molecule of acetic acid: in the first case we start reporting the molecule from the methyl carbon, in the second case, from the double bound oxygen. But the equivalence is ascertained in a syntactic, therefore formal, way, not as the result of a reasoning based on the visualization of the geometry of the molecule.

As a further example, let’s rewrite the application of the .ad. relation in successive steps, here an expression for neopentane structure beginning with an H atom is transformed into one beginning with the central C atom:

\[
H(C(2H, C(3C(3H))))
C(3H, C(3C(3H)))
C(4C(C(3H)))
\]

A much more compact expression is the result of symmetry and the use of the numerical shortcuts into parentheses.

It must again be stressed that this approach needs to be emended in the case of chiral molecules. Furthermore, this definition cannot distinguish from each other molecules that are different for cis-trans isomerism, for example 1,2 cis-dichloroethene and 1,2 trans-dichloroethene both have the following expression: “C(H, Cl, C(H, Cl))”. Our syntactic rule must therefore in many respects be considered an interim solution. See Sect. 4.5.

We note again that those used are strictly syntactic transformations, which can be performed mechanically without thinking about molecular geometry. The verification of the syntactic result using the geometric model of the molecule corresponds to an exercise in semantics, where the molecular model is used to generate valid syntactic strings without using the syntactic sequence of operations. Obviously, the syntactic approach on linear expressions has an intrinsic value from the logical point of view.

**Free valence and valid expressions**

3.4 given a ME like “X(A, B, C(…))”. Symbols like A, B represent single atoms, symbols like “C (“in turn represent groups.
We define the free valence (fv) of an ES, “N”, not further bound (that is an ES followed by “,” or “)” as:

\[ \text{fv} (“N”) = \text{g}(N) \text{ if } \text{g}(N) \leq 3. \]

\[ \text{fv} (“N”) = 8 - \text{g}(N) \text{ if } \text{g}(N) > 3. \]

This corresponds to Lewis’ “octet rule”.

To an AS it is possible to attribute a value for g in a manner consistent with the rules of the Lewis structure starting from that of the ES, replacing

\[ \text{g}(X') = \text{g}(X) + n - m \]

where n is the number of “-” symbols, and m is the number of “+” symbols following X in X’.

Therefore, for example:

\[ \text{fv}(\text{Si}) = \text{g}(4) = 4, \text{fv}(\text{N}+) = \text{g}(4) = 4, \text{fv}(\text{Cl}) = \text{g}(8) = 0. \]

Then it is possible to determine the free valence (fv) of the group “X(…)” in this way:

\[ \text{fv}(\text{X(A, B, C(…)}) = \text{g}(X) - \text{fv}(A) \ldots - \text{fv}(\text{C(…)}). \]

Note that the definition of fv is recursive, and for example in.

\[ \text{X} = \text{C}(2\text{Cl}, \text{C}(3\text{Cl})). \]

\[ \text{fv}(\text{X}) = 4 - 2 - \text{fv}(\text{C}(3\text{Cl})) = 2 - (4 - 3(8 - 7)) = 1. \]

consistently with the fact that CCl\textsubscript{2}CCl\textsubscript{3} actually has a single free valence (it is a one-electron radical).

Thus, in N+(4 H).

\[ \text{fv}(\text{N}) = \text{g}(N) - 1 - 4 \text{fv}(\text{H}) = 0. \]

And the expression, as chemically expected, has no free valence.

Now in connection with the discussion of the valid expressions of the paragraph of the previous section, we may introduce another rule:

A ME is valid if it has free valence equal to zero in all its equivalent forms.

To calculate the fv for each form of the ME we can use the rules of Sect. 4.1.

For example, let us elaborate this ME:

\[ \text{N}(3\text{C}(3\text{H})) \]

Written this way: \[ \text{fv}(\text{N,X}) = 3 - 3 \text{fv}(\text{C}(3 \text{H})) = 3 - 3(\text{fv}(\text{C}) - 3 \text{fv}(\text{H})) = 3 - 3(4 - 3) = 0. \]

we now shift any of the equivalent second AS’s in first position:

\[ \text{C}(3\text{H}, \text{N}(2\text{C}(3\text{H}))) \]

\[ \text{fv}(\text{C}) = 4 - 3 - \text{fv}(\text{N}(2 \text{C}(3\text{H}))) = 4 - 3 - (\text{fv}(\text{N}) - 2 \text{fv}(\text{C}(3\text{H}))) = 1 - (3 - 2) = 0. \]

To terminate, let us shift any of the equivalent second H’s in first position

\[ \text{H}(\text{C}, 2\text{H}, \text{N}(2\text{C}(3\text{H}))) \]

and

\[ \text{fv}(\text{H,X}) = 0. \]

After this three conversions and checks, we state that “N(3 C(3H))” is not just a ME, but a valid ME.

In a valid ME, \text{fv}(A), is also the bond order of A with B, when the two enter into a valid expression this way “B(…, A, …)”. 

\[ \text{Springer} \]
For example, in \( C(O, 2\ H) \), which is a valid ME:
\[ \text{fv}(O) = 2 \] is the bond order of \( O \) with \( C \).

Note that a valid ME is not necessary a chemically significant expression, i.e. the expression of a stable structure in the real world. A formal language must not make any mention of the existence, or not, of objects in reality.

Therefore, both in this provisional arrangement and in perspective, it is very important to note that a language derived from the Lewis structure is an autonomous language, the meaning of which is guaranteed only by its rules. Being an expression of chemical significance is not the same as being a valid expression. Molecules like \( \text{NO}, \text{CO}, \text{SF}_6 \) are simple demonstrations of this fact.

In a perfect formal language for chemical structure, all valid expressions would be of chemical relevance, but this probably is an unattainable goal at the moment.

### Oxidation numbers

3.5 Another aspect that can be easily formalized is the calculation of the oxidation number, O.N. However, this requires the specification of the electronegativity of a chemical symbol \( e(A) \). In fact, it is only necessary to define the increasing electronegativity function, \( \text{er}(A, A') \), containing much less information and defined as:

\[
\text{er}(A, A') =
\begin{cases} 
  1 & \text{if } e(A) > e(A') \\
  -1 & \text{if } e(A) < e(A') \\
  0 & \text{if } e(A) = e(A') 
\end{cases}
\]

Now, it is necessary to define the sum on the atoms \( A \) immediately close to an atom \( X \), i.e. that are found:

1. preceding an open parenthesis, “\( A(X) \)”.
2. following the closing parenthesis “\( \ldots X(A \ldots) \)”.
3. followed by a comma inside the pair of parentheses that opens with \( X \), “\( X(\ldots, A \ldots) \)”.

of the values of the function \( f \), symbolically \( \Sigma_\Delta(A,X,f(A,X)) \).

Then, the oxidation number of the atom \( A \) is:
\[
\text{O.N.}(A) = \Sigma_\Delta(A', A, \text{fv}(A')\text{er}(A', A)).
\]

For example:
\[
\text{CH}_2\text{O}, X = "C(2H, O)".
\]
\[ \text{fv}(H, X) = 1, \text{fv}(O, X) = 2. \]
\[ \text{er}(H, C) = -1, \text{er}(O, C) = 1. \]

therefore \( \text{O.N.}(C, X) = -2 + 2 = 0. \)

### Stereochemistry (non-chiral)

3.6 It is possible to formalize partial conclusions based on the Lewis method, and from here formalize any conclusion that can be reached with a combination of the Lewis method and the repulsion method. For example, given the valid expression.

“\( C(4H) \)”.

Since \( g("C") = 4 \), we can conclude that the structure has a bond between the \( C \) atom and each of the \( H \) atoms and no \( C \) electron is left. Hence the steric number is 4, the number of
lone pairs of electrons is 0. The carbon-centered structure is therefore summarized with the well-known structural symbol (not a symbol of the language, of course).

\[ \text{AX}_4E_0. \]

This alone implies a tetrahedral structure, with all the consequences in turn deducible. Instead, consider:

“C(O, 2H)”.

It implies that the bond between C and O is double: C=O, that the C-H bonds are single, and that the steric symbol is:

\[ \text{AX}_3E_0. \]

This alone implies planar trigonal geometry.

Let’s see how this elaboration can be turned into a formal process.

Once the fv of one or more groups is known, it is possible to determine the quantities that enter the VSEPR method automatically.

For example, the number \( m \) of lone pairs E of a central atom \( N \) in a structure \( X \):

\[ m(N, X) = \frac{g(X) - \text{fv}(A) - \text{fv}(B) - \ldots}{2}. \]

Of course, coherently with the VSEPR method, \( m \) is acceptable only if it is an integer number (not a semi-integer one).

Let us elaborate for example: \( \text{CH}_2\text{O} \).

\[ m(\text{C}, "\text{C}(2\text{H,O})") = \frac{4 - 2\text{fv}(\text{H}) - \text{fv}(\text{O})}{2} = 0. \]

Another example: \( \text{NH}_3 \).

\[ m(\text{N}, "\text{N}(3\text{H})") = \frac{5 - 3}{2} = 1. \]

The calculation of the coordination number \( n \) is easily formalized. Both the number \( n \) and the number \( m \) are therefore formally determinable, so it is the local geometry (according to VSEPR’s rules).

### Formal structural propositions

3.7 The most interesting perspective for a formal language is not just to produce meaningful words, but above all, to be used as a basis for formulating questions and obtaining answers.

This must be done by defining appropriate *propositional functions* (PF) in the metalanguage. A PF in the form \( P(X) \) or \( P(X,Y) \) can be either true or false depending on the actual content of the expression(s) \( X \) (and \( Y \)) thereby providing the answers just mentioned.

In perspective, the successful formulation of PF’s for structural questions of some difficulty would be the key to unlock a genuine chemical theory of structure and properties.

The explicit expression of a PF requires appropriately defined *functions*: these last are recipes to calculate a number out of one or more expressions.

A simple example of such functions is the expression of the oxidation number, in a previous section. Another example is the atom count function, based on the fact that each complete chemical symbol is followed by an opening or closing parenthesis, or an asterisk.

\[ \text{nu}(\text{C}, X) = \text{number of C atoms in } X = \]

number of appearances of the symbol “C(“ or “C,”” or “C)” in \( X \).

We cannot limit ourselves to counting the appearances of the letter “C” because this is also found in ES such as “Cl”.

Following Casari, 1960 this function can be calculated recursively, as such:

\[ \text{nu}(\text{C},”) = 0. \]
nu(C,X) = nu(C,X') + 1:
X' = Y & Z and X = Y & "C" & o & Z, where o is either "(" or "," or ")" or "+" or "-".
Where of course "" denotes an empty expression (no symbol).

We have actually already defined before two PF’s, “adjacency” and (connective) equivalence, which can be now stated in the more compact form Adj(X,X') and Eq(X,X'). Note that, different from functions, which associate an expression with a number, PF's are something which can be true or false. To distinguish them immediately from functions, we indicate them with names that begin with an uppercase letter:

And here are other three explicit definitions of PF's, the first rather trivial, the second and last less, but both valid as very first examples.
X is a hydrocarbon, Hc(X):
nu(C, X) ≠ 0
and nu(H, X) ≠ 0.
and nu(ES ≠ {C, H}) = 0.
X is an alkane, Alk(X):
Hc(X)
and for every ME Y in X:
fv(Y) = 1 or fv(Y) = 0.

The last part of the definition of Alk(X, X') is the formalization of the fact that all Lewis bonds in the structure are single bonds.
X and X' are connectively isomeric alkanes, Cia(X, X'):
Alk(X) and Alk(X')
and nu(C, X) = nu(C, X')
and not Eq(X, X').

The last part of the definition of Cia(X, X') is to exclude that the two expressions represent the same molecule. For example, for.
X = “C(3H, C(2H, C(2H, C(3H))))”.
X' = “C(3H, C(H, C(3H), C(3H)))”.
which corresponds to n-butane and isobutane respectively, Cia(X, X') is true, while for.
X = “C(3H, C(2H, C(2H, C(3H))))”.
X' = “H(C(2H, C(2H, C(2H, C(3H))))”.
Cia(X, X') is false, since X.eq.X'. Both X and X' are in fact expressions of n-butane.

Here are some proposals for other propositional functions: obviously they are valid only in certain contexts, which in turn should be formalized. They are arranged in what might be their order of difficulty of explicit formulation.

1. “The n-th AS that appears in X describes a sp² hybrid atom/ion”.
2. “The structure X’ is a Brønsted acid stronger than structure X’”.
3. “The structure X has A atoms arranged in a tetrahedron” (or another polyhedron).
4. “The structure X is chemically stable”.

Actually, based on our previous considerations, explicit formulation of no.1 is quite feasible. No.2 may use the capability of Lewis structures to actually classify acidic strength in classes of inorganic acids, based on oxidation numbers of central atom (Dickerson et al. 1979). The effort required for the unambiguous formulation of No.3 can be actually consid-
erable, much less if the proposed tetrahedron is to be arranged around an atom to which the A’s are directly bound. No. 4 is proposed here only to offer a perspective view of the concept.

**Problems and perspectives**

4.1 The function $fv$ defined in Sect. 4.1, as it has been fixed at the moment, together with the rule $fv = 0$, excludes a whole series of expressions that a chemist would surely want the language to consider valid, that is, those expressions in which “the octet is expanded”, such as $SF_6$. This is a technical problem that can be considered a starting point for some development: it is certainly solvable, for example by admitting different valid values for the free valence of an elementary symbol in certain cases. The problem of octet is however, as known, in many cases alleviated by formal charge transfer. This can be seen, for example, in the following valid ME of sulfuric acid

$$S^{++}(2O-, 2O(H))$$

4.2 To represent cyclic molecules, such as cyclopropane, an asterisk might be used to mark an atom for later reference in the same expression:

$$X = “C*(2H, C(2H, C(2H, C*))”)”.$$

If more symbols of this type are needed within the structure, they can be obtained by repeating the asterisks: “***”, “****”. In this way the number of symbols used by the language is still finite and indeed restricted.

In principle, this is very simple and effective solution, allowing the linear expression of many systems.

With a similar solution, say using the additional symbol “%”, polymers could also be expressed. For example:

$$Si%(2C(3H), 2O(Si%))$$

However, the formulation of rules for the validity of an expression containing asterisks (and other additional marks) are much more complex than those considered so far. In fact, in an expression in which symbols of the same element appear with the same number of following asterisks (e.g. $C**$), only one, and not more, of these symbols must be followed by an open parenthesis. This is essential because: the occurrence of the symbol followed by the open parenthesis describes how the atom is bound, while all other occurrences are references to this molecular position.

This requires the substantial modification of rules here presented.

4.3 Some remarks formalization of the concept of resonant structures, which is central in Lewis’s theory. The structures of resonance structures, as is known, serve to account for different possible arrangements of electrons, all compatible with the arrangement of the atoms and with the type of atoms involved, and equally coherent. These structures can be formalized without problems in the linear expression of the proposed language. For example, these are resonance structures of the carbonate ion.

$$C(O-, O-, O) C(O-, O, O) C(O, O-, O-).$$
However, if one assumes, as we have done in this paper starting from Sect. 3.3, that the order of the ME’s in parentheses is not important, then the resonance structures are somewhat hidden in the writing.

C(O, 2O−).

In both cases, the problem of resonance structures is a problem which need further work: at present the formal language fixes the position of double and single bonds in structures such as those shown.

However, it should be noted, in this context, that although resonance structures are always introduced very early in the exposures of Lewis structures, in reality they are important in very specific discussions, such as those relating to the different lengths of the bonds, the classification of a molecule as polar or not, and the reactivity of some organic molecules. The related propositional functions (Sect. 3.7) if their formulation is successfully attempted, may (or must) probably account, more or less explicitly, for resonances.

As an additional remark, it must be noted that many structural concepts can be formulated by single structures if formal charges are used as in AS. For example, the following VE for sulfur trioxide:

\[
S^{+++}(3\text{O}−)
\]

accounts perfectly for (D\text{3h}) symmetry, (lack of) dipole moment, and (equality of) bond lengths. The fact that structures with strongly charged atoms are not preferred by chemists is not an issue for the formal language.

4.4 Another possible, perhaps not the most elegant, but potentially useful, way of mechanically extracting information from a linear expression X is to transform it into a connectivity scheme C. This is of course an example of semantics as opposed to syntax.

C can be defined as the pair structures.

\(\{A, R\}(X)\).

The first structure in C is a vector A(i) which lists the AE’s present in the structure. The second is a matrix of propositional functions R(i, j) which can be true or false depending on whether the i-th atom is linked with the j-th atom.

The diagonal relation R(i, i), or if the atom is bound to itself, it does not need to be defined.

As example, for C(4 H) we obtain:

\(\{A, R\} = \{\{C, H\}, \{., T, T, T\}, \{T, ., F, F\}, \{T, F, ., F\}, \{T, F, F ,.\}\} \).

The process of transforming X into C can be formalized without much effort, except when atoms marked(*) appear to represent cycles.

It is clear that from scheme C it is possible to obtain more easily, in a mechanical way, much information deriving from a Lewis scheme.

From the scheme \(\{A, R\}\) it is in fact possible to determine with the known rules the nature of the bond, if it is single, double, triple, and even if it is necessary to include resonance structures, and also the steric number and the coordination number of each atom.

This model, which certainly has aspects of graph theory, allows to solve in a simple way the question relating to the equivalence of two structures in linear form, such as for example.

In fact, we can write.

\(X.eq.X'\).

if and only if.

\(\{A, R\}(X) = \{A, R\}(X')\).
We can consider this second approach to equivalence as a semantic approach, as opposed to the syntactic one elaborated in Sect. 3.

Schemes of this kind, involving the writing of large matrices, with mostly null elements already for molecules of modest size, have been used in the past specifically for the development of predictive computer programs for organic reactions (Ugi et al. 1994), but are probably less useful as logical tools.

4.5 An important problem that needs to be studied in the continuation of this proposal is that of chirality and cis/trans isomers. Indeed, it is obvious that a linear expression, as developed, does not describe molecular chirality. It is evident, for example, a symbol like \( \text{C(\text{Cl, F, H, C(3 H)})} \) does not distinguish the R and S variants of the respective chiral center: this is an important problem for the study of organic chemistry.

Regarding this point, however, it is worth noting that Lewis structures, as such, do not describe chirality: in fact, they are normally considered connectivity structures [2].

In perspective, it is very likely that a solution can be formulated without overly complicating the language: it is probably possible to mark carbon atoms on the basis of chirality when these are chiral centers, adding the related symbols R or S. Of course, it will no longer be possible to consider irrelevant the order of symbols that appear in parentheses, at least in this type of situation. An example of such expressions, whose syntax is to be developed, is:

\[ \text{Cl(C(R)(H, C(3 H), F))}. \]

Therefore, for chiral molecules the syntactic rule of Sect. 3.3 must be reformulated.

It is clear, however, that it is this, which is a problem for the very preliminary current form of the proposed formal system, it could instead be an interesting research perspective, in fact one of the potentially most interesting applications, of an expanded formal system that expresses molecular chirality. As an example, this could find application in the formulation of theories for the selection of chiral species by prebiotic mechanisms, pursuing explanations of the homochirality observed in biochemistry (Aquilanti et al. 2006).

As mentioned before, the language formulated here is also unable to distinguish between cis– and trans– disubstituted alkenes. It is interesting to note that the cis-trans isomerism is an empirical observation, so much so that cis-trans isomers can freely transform into each other at temperatures high enough, but not so high as to dissociate the molecule. So maybe it is a matter of consensus if cis-trans isomerism is an essential part of Lewis structure language. But most people using Lewis structure routinely would probably affirm that the cis/trans distinction is necessary. The problem might be solved in perspective by recognizing that the two carbon atoms which share a double bond and two different single bond substituents each, are in fact not equivalent when confined in the plane. In fact, they can’t be superimposed by rotating them into such plane. Assuming for example that the reading order in the expression correspond to counterclockwise ordering around an atom coordinated on a plane, such expressions would be:

\[ \text{C(H, Cl, C(Cl, H)) for cis.} \]
\[ \text{C(H, Cl, C(H, Cl)) for trans.} \]

4.6 A point that deserves attention is the following: to what extent should a formal system try to reproduce chemical reality? As is known, the Lewis method allows us to deduce many real molecular structures. Some cases in books of not strict adherence of such rules to reality, are extremely specific, like for example the well-known phosphonic acid case, where we have a P-H bond in an oxyacid, thereby violating a previous informal rule for oxyacid structures (X-O-H rule).
On the other hand, an attempt to amend the method, which is certainly possible, to correctly predict this structure and other exceptional cases as well would make the set of rules too much complex.

It is probably preferable to have a formal technique which makes precise, logical, unambiguous, but occasionally inaccurate predictions when compared to chemical reality. An attempt to proceed differently would produce not a synthetic formal theory, but a collection of empirical facts loosely linked by micro-theories.

4.7 A few comments in order regarding the possibility of practical use of a formal system based on Lewis structures: it is appropriate to note that Lewis structures are a very versatile molecular modeling method and in many respects are underestimated, especially in the recent trend to carry out molecular simulation using ab initio methods or at least based on force fields. The latter two methods provide a large amount of additional information but often unnecessary from a structural point of view: on the contrary, in many cases the Lewis structures together with the VSEPR method are able to reproduce a large amount of molecules in a semi-quantitative way. Even inter-atomic distances can be reproduced with a good approximation simply by using the sum of the atomic radii as an estimate, with prescribed corrections for multiple bonds (Gillespie et al. 2013, Dickerson et al. 1979). It should be emphasized once again that the analysis presented in this work is preliminary and does not describe all the potential of Lewis structures. However, it is evident that a Lewis model of the type that can be described with the simple language of this work is already capable of evaluating a large number of quantities, as real functions defined in the word space of the language. A trivial example is the distance, if necessary averaged, between two atoms in a structure defined by a word, hence a definition like this:

\[ \text{Dis}(A, B, X). \]

Taking as a simple example \( X = \text{“C(A(…), B(…), 2H)”} \), \( \text{Dis}(A,B,X) \) should be calculated as:

\[ \text{Dis}(A,B,X) = (R(A) + R(C))^2 + (R(B) + R(C))^2 - 2(R(A) + R(C))(R(B) + R(C)) \cos(109.5^\circ). \]

where \( R(A) \) is the atomic radius of the element A, based on the cosine theorem and the ideal tetrahedral angle.

This calculation only assumes the rules of Lewis structure, that of VSEPR and the use of atomic radii to estimate interatomic distances.

Of course, the above is the result of an informal elaboration while such results should be formally produced by the system. This cannot be done in general without preliminary solving the issues at 4.5 above.

### Conclusions

Although these considerations are feasible for many refinements, the author believes that he has traced a path of potential interest for other researchers, which goes in the direction of seeking an effective formal system based on Lewis structures. The interest of this research is clearly intrinsic, for those who are looking for an effective, that is scientifically productive, formalization of scientific methodologies. It is also probable that pursuing this research could also be of practical use, in order to formulate automatic systems for the generation of interesting and sound structures.
The formalization work can also be valuable for teaching: Lewis structures are often explained to freshmen as a set of practical skills, with numerous cases and exceptions, continuous reference to chemical reality, and all this often confuses students.

Instead, it may be useful to acknowledge that the method does not require any aspect of intuition or induction from individual cases or exercises.

Furthermore, it is very important, from the author’s point of view, that the formulation of formal systems of this type develops the awareness of how much the constructive principles of chemistry are, after all, autonomous, even if physics may contribute with images and some justification for rules.

Thus, the effort of formalization can help to clearly discern the strictly chemical, and intrinsically productive, concepts of bond theory, and the combinatorial barrier that separates them from the underlying laws.

**Funding** Open access funding provided by Università degli Studi di Bari Aldo Moro within the CRUI-CARE Agreement.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit [http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/).

**References**

Scott, A.: Stairways to the Mind. Copernicus, New York (1995)
Dickerson, R.E., Gray, H.B., Haight, G.P.: Chemical principles. The Benjamin/Cummings Publishing Company, Inc. (1979), in particular Chap. 11
Gillespie, R.J., Hargittai, I.: The VSEPR model of molecular geometry. Dover. (2013)
Hunter, G.: Metalogic: An introduction to the metatheory of standard first order logic. Macmillan International Higher Education (1971)
Casari, E.: Lineamenti di logica matematica. Feltrinelli (1960)
Lemmon, E.: John. Beginning logic. CRC Press (1971)
Connerade, J.P.: Confining and compressing the atom. Eur. Phys. J. D. **74**, 211 (2020)
Ugi, I., Bauer, J., Blomberger, C., Brandt, J., Dietz, A., Fontain, E., Stein, N.: Models, concepts, theories, and formal languages in chemistry and their use as a basis for computer assistance in chemistry. J. Chem. Inf. Comput. Sci. **34**(1), 3–16 (1994)
Aquilanti, V., Maciel, G.S.: Observed molecular alignment in gaseous streams and possible chiral effects in vortices and in surface scattering. Origins of Life and Evolution of Biospheres **36**.5 (2006):435–441

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.