Institut für Quantenchemie Berlin
Werner Hässelbarth

A Combinatorial Description of Structure and Properties of Chemical Compounds

Chemistry is concerned with relations between structure and properties of chemical compounds. Such connections cannot be found by investigating just some single compounds but only by systematic comparison, involving compounds of appropriate classes. Appropriate means that within such a class the structure of compounds varies in a well-defined and lucid fashion. Mappings between finite sets are particularly suitable for defining and parametrizing variations of chemical structure.

In somewhat more general terms, consider a system composed of subsystems 1, 2, ..., p each with a finite state space L_1, L_2, ..., L_p (finite space = finite set). Let the composite system be such that each of its states is completely characterized by specifying the states of all the subsystems. Then its state space is the cartesian product $L_1 \times L_2 \times \ldots \times L_p$ of the state spaces of its components.
Let us further assume that for any state of a subsystem there is an analogous state of each other subsystem. Then all the state spaces L_i can be mutually identified. Hence the state space of the composite system takes the particular form of a cartesian product $L \times L \times \ldots \times L$ of p copies of the same state space L, common to all the subsystems 1, 2, ..., p. This, however, is nothing else but the set L^P of all mappings from the set $P = \{1, \ldots p\}$ of subsystems into the set L.

Example: derivatives of a common parent compound with ligands from a specified assortment

subsystems: the positions where substitution may take place

their states: the various kinds of ligands (substituents). Ligands of the same kind at different positions are analogous states of the subsystems concerned.

The states of the composite system represent distributions of ligands over the substitution positions of the parent compound, i.e. derivatives.

Frequently, it is not the individual states (mappings), that correspond to the various compounds or structures to be described, but classes of equivalent states. This equivalence mostly has symmetry reasons. If e.g. the parent compound posesses some spatial symmetry, then there are symmetry equivalent substitutions. They result in distributions of ligands over the positions of the corresponding (spatially fixed) molecular skeleton, which can be mutually transformed by rigid rotations from its symmetry group. Hence they represent merely different spatial orientations of the same molecule, i.e. the same molecular structure.
Occasionally other types of equivalence occur, but we shall stay with the case of symmetry equivalence. Then the structures of interest are parametrized by equivalence classes of states (mappings) which arise as orbits of a group, acting as a permutation group on the state space L^P. With regard to our main example we call these orbits s u b s t i t u t i o n p a t t e r n s. The group in question is a symmetry group i.e. a group of automorphisms of a common basic structure from which the composite structures derive. For chemical derivatives this is the point symmetry group of the parent compound, more precisely its subgroup of proper rotations. In case one does not distinguish mirror image compounds (enantiomers), the full point symmetry group takes its place.

Within this setting quite different problems can be stated and treated. We are going to consider

(i) the symmetry of substitution patterns

(ii) a cluster expansion of properties of composite systems.

S u b s t i t u t i o n s y m m e t r y

On passing over from the basic structure to composite structures, the symmetry of the basic structure is partially destroyed. The surviving part is the symmetry of the composite structure. Let e.g. ligands of different kinds be distributed over the positions of a symmetric molecular skeleton (e.g. the corners of a regular polyhedron). If the ligands themselves are sufficiently symmetric (e.g. coloured balls), then covering operations of the skeleton act on such distributions just by transporting ligands to other positions, equivalently by permuting the positions of the ligands. Therefore, exactly those covering operations survive as covering operations of a distribution which

mutually permute equally substituted positions exclusively. Symmetry equivalent distributions have the same symmetry, i.e. their symmetry groups are conjugated subgroups of the skeleton symmetry group. The symmetry of a substitution pattern then is a conjugacy class of subgroups of the group in question. As a typical enumeration problem in this context we ask for the number of patterns with prescribed symmetry. This problem can be formulated quite generally for an arbitrary finite group acting on an arbitrary finite set. So let us begin by fixing some notation.

A permutation representation of a group G on a set M is a homomorphism of G into the symmetric group S_M of M

$$\pi: \begin{array}{ccc} G & \longrightarrow & S_M \\ g & \longmapsto & \pi_g \end{array} .$$

Synonymously, G acts on M or M is a G-set.

Example: Any covering operation of a symmetric polyhedron induces a permutation of its corners, of its edges etc.

An action of G induces an equivalence relation on M:

$$m' \sim m \iff \exists\, g \in G:\ m' = \pi_g(m).$$

The corresponding equivalence classes are called the orbits of M under the action of G. With gm as a short form of $\pi_g(m)$, the orbit which contains $m \in M$ is given by

$$O(m) = \{ gm \mid g \in G \}.$$

The last concept we need is that of the stabilizer. The stabilizer of an element $m \in M$ is the set of all those group elements which fix m

$$G_m = \{ g \in G \mid gm = m \}.$$

G_m is a subgroup of G. Elements in the same orbit have conjugate stabilizers

$$G_{gm} = gG_m g^{-1} \quad .$$

Running through the elements of an orbit, their stabilizers run through a complete class of conjugated subgroups of G (possibly several times). Hence any orbit is associated with a conjugacy class of subgroups. In view of the previous discussion we call it the symmetry of the orbit.

In the case of substitution patterns, in particular for derivatives of a symmetric parent compound we have

P = set of positions of a molecular skeleton, e.g. the corners of a polyhedron

L = set of ligand sorts, e.g. colours of balls

L^P = set of all distributions of ligands of sorts from L over the positions in P.

G = group of covering operations of the skeleton

We already described the most simple type of action of G on $M = L^P$: covering operations act on distributions by permuting the positions of the ligands, i.e. G acts naturally on P,

$$g \mapsto \pi_g \in S_P \quad ,$$

and this operation, in turn, induces an action on L^P

$$g: \varphi \mapsto \quad g\varphi := \varphi \circ \pi_g^{-1} \quad .$$

The stabilizer G_φ is the symmetry group of the distribution φ , and the conjugacy class of G_φ represents the symmetry of the corresponding substitution pattern (derivative, molecular structure). Asking for the number of patterns with prescribed symmetry thus means asking for the number of orbits the elements of which have stabilizers in a prescribed class of conjugated subgroups. More generally, for a G-set M and a subgroup $H \leq G$, which is the number

$$x_H = \#\text{ of orbits with stabilizers in } \{ gHg^{-1} \mid g \in G \} \quad ?$$

If $G_m = H$, then among the $G_{m'}$ with $m' \in O(m)$ the conjugates gHg^{-1} all occur equally often. Hence x_H is proportional to the number s_H of elements of M with H as their stabilizer

$$s_H = |\{m \in M | G_m = H\}| \quad .$$

As a rule, this number is much harder to calculate than a related number, namely the number of H-invariant elements of M

$$i_H = |\{m \in M | G_m \geq H\}| \quad .$$

As we shall see, given all the i_H we can (at least in principle) calculate the s_H and from them we get the numbers x_H just by multiplication with an (irrelevant) factor. The following equation is obvious

$$i_H = \sum_{H \leq K \leq G} s_K \quad .$$

We rewrite it in the form

$$i_H = \sum_{K \leq G} \zeta(H,K)\, s_K \quad ,$$

where

$$\zeta(H,K) = \begin{cases} 1 & H \leq K \\ 0 & \text{otherwise} \end{cases} \quad .$$

With a suitable numbering of the subgroups of G, e.g. according to their cardinality, the matrix of the coefficients $\zeta(H,K)$ is triangular with diagonal elements all ones. Such a matrix is invertible, so there are coefficients $\mu(H,K)$ that

$$s_H = \sum_{K \leq G} \mu(H,K)\, i_K \quad .$$

In more advanced terms, ζ is the Zeta-function of the subgroup lattice of

G, μ its Möbius-function, and the numbers s_H are obtained from the i_H by Möbius-inversion.
The size of the problem can be boiled down from # of subgroups of G to # of conjugacy classes of subgroups of G using the fact that both s_H and the i_H are constant on classes of conjugate subgroups. The matrix associated with the Zeta-function, reduced to conjugacy classes of subgroups by partial summation, then yields the so-called t a b l e o f m a r k s of G.
Besides the table of marks the numbers i_H of H-invariant elements have to be known for all subgroups H of G or a transversal of conjugacy classes of subgroups, respectively. If G acts on $M = L^P$ via permutations of the positions as described above, then a mapping is H-invariant if and only if it is constant on the H-orbits of P. From that we get immediately

$$i_H = |L|^{|P/H|} ,$$

where $|P/H|$ denotes the number of H-orbits of P.

C l u s t e r E x p a n s i o n

Again we consider a composite system with state space L^P, where P is the collection of subsystems and L their common state space. Let us define a (real number valued) property of the system as a function $f: L^P \to \mathbb{R}$ in the sense that for $\varphi \in L^P$ the number $f(\varphi)$ is the value of the property f for the system in its state φ. We may e.g. consider f to represent a measuring device and $f(\varphi)$ as the result of the corresponding measurement performed on the system in its state φ. Cluster in this context means an aggregate of subsystems, i.e. a subset of P. The cluster expansion of a property is an expansion into a sum of contributions from clusters of increasing size. For three subsystems e.g., this amounts to an expansion of a real function of three variables that run through the same finite set, as follows

$$\begin{aligned} f(x,y,z) = {} & a_{\emptyset} + b_1(x) + b_2(y) + b_3(z) \\ & + c_{12}(x,y) + c_{13}(x,z) + c_{23}(y,z) \\ & + d_{123}(x,y,z). \end{aligned}$$

There, $a_{\emptyset}$ is the best approximation of f by a constant. b_1, b_2, b_3 are func-

tions of one variable such that $b_1(x) + b_2(y) + b_3(z)$ is the best approximation of $f(x,y,z,) - a_\emptyset$. Similarly, c_{12}, c_{13} and c_{23} are functions of two variables such that their superposition is the best approximation of the remainder of the foregoing approximation step. Finally, d_{123} is the ultimate remainder. Using the least squares criterion, all these approximation problems can be solved in closed form. One obtains

$$a_\emptyset = \langle f(x,y,z) \rangle_{x,y,z} ,$$

$$b_1(x) = \langle f(x,y,z) \rangle_{y,z} - a_\emptyset , \qquad b_2, b_3 \text{ analogous} ,$$

$$c_{12}(x,y) = \langle f(x,y,z) \rangle_z - b_1(x) - b_2(y) - a_\emptyset , \qquad \ldots ,$$

where the brackets denote averaging, and the subscripts indicate the variables over which to average.

In the case of a property that is additive for non-interacting subsystems (as e.g. the energy is) the cluster expansion is similar to a perturbation expansion with the interactions between the subsystems taking the part of the perturbation. It starts with the sum of the contributions of the single subsystems as the zero-order term, followed by corrections due to interactions of increasing complexity: interactions between pairs, triples etc.
As we are going to see, however, the final result is something better than the expansion we had in mind with the motivation sketched above. It is instead a d e c o m p o s i t i o n

$$f = \sum_{Q \subseteq P} \tilde{f}_Q$$

of a property f into a sum of contributions $\tilde{f}_Q$ of all the clusters $Q \subseteq P$ which can each be calculated separately without any reference to a sequence of approximation steps.

In order to avoid a towering of superscripts, let us abbreviate

$M = L^P$, the state space of the system,

$X = \mathbb{R}^M$, the set of its properties.

With addition of functions and their multiplication by real numbers defined pointwise in the usual fashion, the set X of properties becomes an $\mathbb{R}$-vector-space of dimension $|M| = |L|^{|P|}$. Moreover, we may equip X with the customary scalar product, turning it into a euclidean vectorspace.

$$(f,g) := \sum_{\mu \in M} f(\mu) g(\mu).$$

This puts at our disposition the following simple approximation theorem of linear algebra: In a euclidean space, the best approximation of a given vector by elements of a subspace is the orthogonal projection of that vector upon the subspace in question. So the subsequent definitions and results will deal with

(1) certain subspaces of X that are associated with the clusters, i.e. aggregates of subsystems,

(2) the corresponding orthogonal projections.

We begin by associating with each cluster $Q \subseteq P$ the subspace of all those properties which only depend on the state of this cluster (i.e. which are independent of the state of the complementary cluster $P \setminus Q$).

$$X_Q = \{ f \in X \mid \mu_Q = \rho_Q \Rightarrow f(\mu) = f(\rho) \}.$$

Here, the subscript Q denotes the restriction of a mapping $P \to L$ to $Q \subseteq P$. Clearly, each X_Q is a subspace of X; in particular $X_\emptyset$ is the subspace of constant functions. Moreover those spaces have the properties

$$R \subseteq Q \Rightarrow X_R \subseteq X_Q,$$

$$X_R \cap X_Q = X_{R \cap Q}.$$

How can dependency on the state of a certain cluster be removed?
By averaging over its states! Hence we define an operator p_Q for each $Q \subseteq P$ by (l,p,q denote the cardinalities of the sets L,P,Q)

$$p_Q: \qquad [p_Q f](\mu) := \frac{1}{l^{p-q}} \sum_{\substack{\rho \in M \\ \rho_Q = \mu_Q}} f(\rho).$$

One easily verifies that p_Q is a linear map onto X_Q; moreover it is symmetric, and the following relation holds

$$p_Q\, p_R = p_{Q \cap R} ,$$

in particular $$p_Q^2 = p_Q .$$

This proves that p_Q is the orthogonal projector onto X_Q.

But these spaces and projectors are not yet what we are looking for. Consider the following subspace Σ_Q of X_Q

$$\Sigma_Q := \sum_{R \subset Q} X_R ,$$

and its orthogonal complement $\tilde{X}_Q$. According to the decomposition of X_Q

$$X_Q = \tilde{X}_Q \oplus \Sigma_Q$$

each projection $f_Q = p_Q f$ of a property f decomposes uniquely as follows

$$f_Q = \tilde{f}_Q + f_\Sigma \qquad \text{with } \tilde{f}_Q \in \tilde{X}_Q,\ f_\Sigma \in \Sigma_Q .$$

Since f_Σ is in turn a sum of contributions of smaller clusters, it is $\tilde{f}_Q$ that remains as the specific contribution of the cluster Q to the property f.

From an equivalent characterization of the subspaces X_Q,

$$\tilde{X}_Q = \{ f \in X_Q \mid p_R f = o \quad \text{for all } R \subset Q \}$$

it follows immediately, that those spaces are orthogonal

$$\tilde{X}_Q \perp \tilde{X}_S \qquad \text{for } Q \neq S .$$

In fact, a stronger relation holds:

$$\tilde{X}_Q \perp X_S \qquad \text{for } Q \not\subseteq S.$$

From this result one concludes that X_Q is the direct sum of the mutually orthogonal cluster spaces $\tilde{X}_R$, $R \subseteq Q$.

$$X_Q = \sum_{R \subseteq Q}^{\oplus} \tilde{X}_R$$

Still the orthogonal projectors $\tilde{p}_Q$ on the subspaces $\tilde{X}_Q$ have to be determined. But since the $\tilde{p}_R$, $R \subseteq Q$ add up to p_Q ,

$$p_Q = \sum_{R \subseteq Q} \tilde{p}_R \ ,$$

it follows by Möbius-inversion with the Möbius-function of the power set lattice of P, that

$$\tilde{p}_Q = \sum_{R \subseteq Q} (-1)^{|Q|-|R|} \, p_R \ .$$

In particular, from $X_P = X$ and $p_P = 1$, the identity operator of X, it follows that

$$X = \sum_{Q \subseteq P}^{\oplus} \tilde{X}_Q \ ,$$

$$1 = \sum_{Q \subseteq P} \tilde{p}_Q \ ,$$

$$f = \sum_{Q \subseteq P} \tilde{f}_Q \qquad \text{with } \tilde{f}_Q = \tilde{p}_Q \, f \ .$$

X is the direct sum of mutually orthogonal subspaces, each associated with a cluster $Q \subseteq P$; the orthogonal projectors on these cluster spaces constitute a resolution of unity. This leads to the explicit form of the cluster expansion of a property f in the third line.
This decomposition provides an empirical definition of the contribution of an aggregate of subsystems to a property of composite systems. A typical application could be an analysis of large sets of experimental data of a molecular property, yielding empirical rules on how its numerical values depend on molecular structure. As a further step such rules might be interpreted and used as material for developing a theory of the property concerned.

References

Concerning the part on substitution symmetry, the enumeration of orbits with prescribed symmetry is in fact an old problem. It may be regarded as the permutation representation analogue of the well-known problem to determine the multiplicity of an irreducible linear representation in a reducible one.
In this connection see

[1] W.Burnside: Theory of finite groups (1955)

for basic concepts of the theory of permutation representations including the marks,

[2] A. Dress/M. Küchler: unpublished manuscript (1971)
[3] D. Knutson: λ-Rings and the Representation Theory of the Symmetric Group, Lecture Notes in Mathematics Vol. 308 (1973)

both for recent descriptions of this theory, and

[4] A. Kerber/K.J. Thürlings: manuscript on Polya/Redfield-enumeration theory to be published in Bayreuther Mathematische Schriften (1983)

for the combinatorial aspects.
The cluster expansion appears to be new, so my only reference is

[5] W. Hässelbarth: Habilitationsschrift (1982) which also refers to the first part on substitution symmetry.

The basic philosophy for this type of 'mathematical chemistry' is due to Ernst Ruch, see e.g.

6 E. Ruch: Chiral Derivatives of Achiral Molecules, Angew. Chem. Int. Edition 16, 65 (1977)

Werner Hässelbarth
Freie Universität Berlin
Institut für Quantenchemie
Holbeinstraße 48
1000 Berlin 45