CROATICA CHEMICA ACTA
CCACAA, ISSN 0011-1643, e-ISSN 1334-417X
Croat. Chem. Acta **86** (4) (2013) 509–517.
http://dx.doi.org/10.5562/cca2296

Original Scientific Article

Estimating Octanol / Water Partition Coefficients by Order Preserving Mappings[†]

Rainer Bruggemann[a,*] and Guillermo Restrepo[b]

[a]*Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany*
[b]*Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia*

RECEIVED JUNE 06, 2013; REVISED AUGUST 29, 2013; ACCEPTED SEPTEMBER 27, 2013

Abstract. Douglas Jay Klein and other researchers have seen that there is a great potential in applying partial order in the field of QSAR. The basic idea is to deduce from order relations among chemicals, properties of the ordered chemicals. Despite the satisfactory results found by Klein's methods, we think it is worth exploring another feature of the poset of chemicals used in Klein's approaches, namely the average posetic height, h_{av}, of each chemical. In the current paper we explore some methods to calculate these heights and use them as independent variables for modelling a chemical property that is also modelled by Klein's approaches, namely octanol / water partition coefficient K_{OW}. It turned out that the application of average heights, h_{av}, as predictors of a linear log K_{OW} estimation leads to reasonable results in the application example of chlorophenols. (doi: 10.5562/cca2296)

Keywords: QSAR, chemical structures, partially ordered sets, lattice theory, chlorophenols, log K_{OW}

INTRODUCTION

The octanol / water partition coefficient is a key property by which several other properties of hydrophobic chemicals are estimated in a quite simple way (see for instance[1,2] and networks of property estimation relations).[3–7] For example different accumulation tendencies, such as accumulation in fish, soils, or baseline acute toxicities, just to mention a few of them, are important for estimating the fate and exposure and ecotoxicological impacts of organic chemicals in the environment. Although the experimental determination of K_{OW} is a standard procedure, it is of interest to estimate this quantity directly from the structure of molecules. For this purpose, in the paper of Ivanciuc *et al.*[8] a partially ordered set of chlorophenols is built up and by several interpolation and extrapolation methods it was found that log K_{OW} of the different compounds could be estimated with good accuracy. The bare idea is that the partially ordered set and its graph theoretical structure constitutes an efficient code of the diversity of structures, which led, in Ivancviuc *et al.*[8] and in other papers by Klein (see section 2.4), to the concept of "superstructure". When the partially ordered set is such a superstructure, then other partial order properties may be useful as predictors of log K_{OW} as well.

MATERIAL AND METHODS

Basic Definitions

Partial order relations can be obtained in a multitude of ways; in 1997 Klein and Babić showed some instances of appearance of these structures in chemistry.[9,10] If the partial order is to be related to a data matrix we can define partial order relations among objects in many different manners[11]. Here a method is selected that has found manifold applications; it is called the product order, because the orders induced by any single indicator are combined in a logical "and" manner. In other words, the method is a logical combination of the rankings derived from each indicator, as if each indicator were acting as a voter.

Let X be a finite set of objects and IB the set of indicators q_i observed on X, then we define:

$$x \preceq y : q_i(x) \preceq q_i(y) \text{ for all } q_i \in IB, \text{ with } x, y \in X \quad (1)$$

It turns out that if x, y and z are objects in X, then the relation $\preceq$ satisfies the following properties: reflexivity, indicating that $x \preceq x$; antisymmetry, meaning that if $x \preceq y$ and $x \preceq y$, then $x = y$; and transitivity, which

[†] Dedicated to Professor Douglas Jay Klein on the occasion of his 70th birthday.

* Author to whom correspondence should be addressed. (E-mail: brg_home@web.de)

states that if $x \preceq y$ and $y \preceq z$, then $x \preceq z$. The relation $\preceq$ is called a partial order and such a relation along with the set X is called a partially ordered set (poset), denoted $(X, \preceq)$. As $\preceq$ is determined by indicators in *IB*, then $(X, \preceq)$ is equivalent to (X, IB). For the ensuing discussion, some notational remarks are needed:

(a) Objects for which $x \preceq y$, or $y \preceq x$ are called *comparable* and denoted as $x \perp y$ or $y \perp x$.

(b) Objects for which Equation 1 does not hold, are called *incomparable* and the relation is denoted as $x \,||\, y$.

(c) $O(x) := \{y \in X : y \preceq x\}$ is the *principal down set (order ideal)* of X generated by x. (2)

(d) $F(x) := \{y \in X : x \preceq y\}$ is the *principal up set* (*order filter*) of X generated by x. (3)
Note: down and up sets are usually defined in a more general manner: Let $X' \subseteq X$, the down / up set of X' is a set where $z \in X'$, $x \in X$ and $x < z$ / $x > z$ imply $x \in$ down/up set.

(e) $U(x) := \{y \in X : y \,||\, x\}$ is the set of *x-incomparables* in X. (4)

(f) $Iso(X) = \{x \in X :$ there is no $y \in X$ such that $y \perp x\}$ is the set of *isolated elements* of X (5)

(g) Let $C \subseteq X$, if all $x, y \in C$ obey Equation 1, then C is a *chain*. (6)

(h) Let $z, y \in X$, then $I(z, y) := \{x \in X : z \preceq x \preceq y\}$, then $I(z, y)$ is the *interval* corresponding to objects z and y.

(i) Let $AC \subseteq X$, if for all $x, y \in AC$ is valid $x \,||\, y$, then AC is an *antichain*.

(j) If $x \preceq y$ and there is no element $z \in X$ fulfilling $x \preceq z \preceq y$, then *y covers x* or *x is covered* by y and it is denoted by $x \preceq : y$. It is sometimes convenient to describe a poset algebraically by the set of (x, y) pairs, where $x \preceq : y$. Cover relations are the basis to draw Hasse diagrams, graphic representations of posets, see for instance Davey and Priestley.[11]

For recent introductory texts regarding the Hasse diagram technique (HDT), which is based on Equation 1, see for instance.[12,13]

Weak Order and Height

As the product order for the objects in X, based on *IB*, considers each possible ranking derived from each q_i, it is needed to explore the mathematical properties of these rankings, *i.e. weak*, *linear*, *total* or *complete* orders. A weak order is a relation $\leq$ satisfying reflexivity and transitivity for all objects in X. Note that $\leq$ does not require antisymmetry, which is the main difference with a partial order relation. Hence, if the set X is endowed with the relation $\leq$, then every couple of different objects in X is comparable. A linear, total or complete order: X equipped with $\preceq$ is a chain (see expression 6 above).

Let L_i be a *linear extension*, *i.e.* a linear order preserving all order relations found through Equation 1. For example by equation 1 and three objects (a, b, c) it may be found: $a < b$, $a < c$, then a linear extension is $a < b < c$, another one would be $a < c < b$. In a finite set X, each linear extension has a *least object* ("*least*"). The *height* $h(x, L_i)$ of an object x regarding the linear extension L_i is given by counting the number of objects in $I(least, x)$ fulfilling $least \preceq y \preceq x$.

The average height of x, $h_{av}(x)$, is calculated[14] as follows:

$$h_{av}(x) = \frac{\sum_{L_i} h(x, L_i)}{LT} \qquad (7)$$

with LT being the total number of linear extensions derived from Equation 1. The quantity $h_{av}(x)$ is often called the *average rank* of x as it induces a weak order on X.

Average Height Calculation Methods

Lattice Approach

In computational complexity theory, problems are classified according to the time and space used by algorithms to solve the problems. The notation #P is the class of counting problems running in polynomial time. Problems of class #P can be solved in theory but in practice take too long time and are also known as intractable problems.

The direct evaluation of Equation 7 is #P complex,[15] hence the application of this equation is only meaningful when the number of objects in X is small. A computational and mathematical attractive method to estimate h_{av} is the lattice theoretical one,[16] which avoids storing and managing all L_i in LT (Appendix). The idea may be sketched as follows:

(a) A lattice of all down sets for (X, IB) is constructed. Due to the construction principle each edge of the lattice can be labelled by an object in X (see below).

(b) A linear extension is then a path from the bottom (empty set) to the top object (the poset itself).

(c) By an appropriate organisation of the computational algorithm, important quantities of the poset such as average height[16] can be obtained.

To manually find all down sets of a poset (X, IB), we perform the following steps:[11]

1. Identify all principal down sets $O(x)$ for all x in X.
2. Generate a list of antichains.
3. Generate the down set $Z(i)$ as follows: let $AC(i)$ be the i-th antichain, then $Z(i)$ is defined as

$$Z(i) = \bigcup\nolimits_{x \in AC(i)} O(x)$$

4. Include the empty set, if necessary.
5. Include the poset (X, IB), if necessary.

Example: Let us take the poset whose Hasse diagram is shown in Figure 1. The set X of the poset is $\{a, b, c, d\}$ and its cover relations are $\{(a, b), (c, b), (c, d)\}$ (see remark j in section 2.1). This poset is called a fence, namely Fence(4), and here we show how we apply each one of the five steps mentioned before:

1. Principal down sets: $O(a) = \{a\}$, $O(b) = \{a, b, c\}$, $O(c) = \{c\}$, $O(d) = \{c, d\}$.
2. Antichains: $AC(1) = \{a, c\}$, $AC(2) = \{a, d\}$, $AC(3) = \{b, d\}$.
3. $Z(1) = O(a) \cup O(c) = \{a, c\}$; $Z(2) = O(a) \cup O(d) =\{a, c, d\}$; $Z(3) = O(b) \cup O(d) = \{a, b, c, d\}$.
4. Inclusion of $\varnothing$.
5. It is not needed to include the poset, for it is already included through $O(b) \cup O(d) = (X, IB)$.

The set of all down sets is denoted by J, hence for the Fence(4) we obtain:

$$J(\text{Fence}(4)) = \{\varnothing, \{a\}, \{c\}, \{a, c\}, \{c, d\}, \{a, b, c\}, \{a, c, d\}, \{a, b, c, d\}\}$$

The sets of J(Fence(4)) are partially ordered by set inclusion as shown in Figure 2A.

There are many important theorems and statements to be made for the partial order of down sets. Here we refer only to those needed for the ensuing discussion:[17]

1. The partial order of down sets (ordered by inclusion) is a distributive lattice.
2. For any covering vertex v such that $v' \preceq: v$, the associated down set of v differs from that of v' by one and only one $x \in (X, IB)$. Hence the edges of the lattice can be labelled by x, as shown in Figure 2B.
3. A path from the bottom of the lattice to the top can be characterised by the edges and their labels found in the path.
4. The sequence of edge labels is a linear extension of (X, IB), for example $c \preceq d \preceq a \preceq b$ is a linear extension of the Fence(4).

Following,[16,18,19] these four items are the basis of a computational approach to calculate average heights. In the software PyHasse this concept is available too (avrank5.py), where the Python-code by Wienand (2003)[19] was correspondingly adjusted. For the software package PyHasse see.[13,20–23]

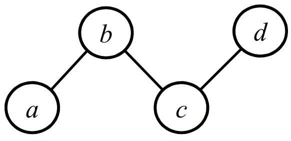

Figure 1. Fence(4).

Other Approaches

Nevertheless, the lattice-theoretical approach (avrank5.py) is limited too. Empirically Bruggemann and Carlsen[24] found that the number of elements in X must be below 50, at least when the programming language is an interpreter language, *e.g.* Python, and no computational parallelisation is performed.

As the direct evaluation of Equation 7 and the lattice method are computationally demanding for posets with large number of objects, then several methods have been devised to approximate the exact results of Equation 7 and the lattice method. One of these methods is the one by Bubley and Dyer,[15] which is a sampling method, whose convergence is guaranteed, albeit the running time is of the order $O(n^3 \log(n))$.

Another method is the Local Partial Order Model,[25] LPOM, which in plain words is just to select one after another the objects $x \in X$ and to consider their partial order environment. It has been proven that LPOM computational complexity is $O(n^2)$.[26] There are two LPOM variants, namely LPOM0 and LPOMext. The first one assumes all x-incomparables to be isolated objects; then both down set and up set of x are considered as chains and a simple counting of the probability for each $y \in U(x)$ localized above or below x, allows estimating $h_{av}(\text{LPOM0})(x)$. Further details are found in.[25] In LPOMext the assumption of x-incomparables as isolated objects is abandoned; instead, for each object $y \in U(x)$, it is checked how many positions are accessible above and below x. At the end, $h_{av}(x)$ is calculated through Equation 8.[24,27]

$$h_{av}(\text{LPOMext})(x) = |O(x)| + \sum\nolimits_{y \in U(x)} \frac{p^{<}(y)}{p^{<}(y) + p^{>}(y)} \qquad (8)$$

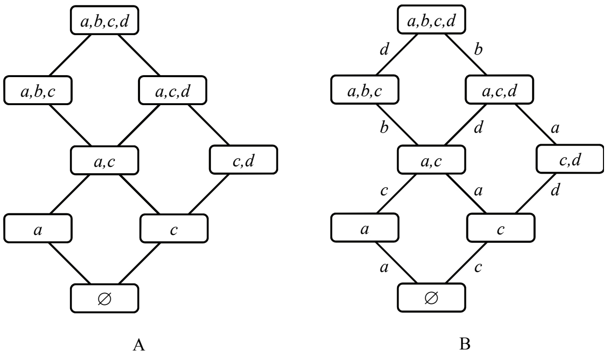

Figure 2. (A) J(Fence(4)): partial order of down sets of Fence(4), ordered by set inclusion; (B) labelling of edges.

with $p^{<}(y) := |O(x) \cap U(x)|$ and $p^{>}(y) := |F(x) \cap U(x)|$, $x, y \in X$.

The average height, calculated after Equation 8 is an approximation of Equation 7, for an assumption of the method is that the objects of $O(x)$ and $F(x)$ form a chain and the order relations amongst the elements of $U(x)$, $O(x)$ and $F(x)$ are disregarded. Nevertheless, the approximation of Equation 8 is in most cases remarkably better than LPOM0 and the Pearson correlation with exact average heights is quite high, see for details Bruggemann and Carlsen.[24] The role of average heights approximations is further examined in a paper in press in *MATCH Communications in Mathematical and in Computer Chemistry*, 2014.

After having discussed different methods to calculate h_{av}, we show how the lattice theoretical "h_{av} exact", the LPOM0 and the LPOMext methods can be used as predictors for estimating log K_{OW} of chlorophenols. Before this idea is outlined in the results section, we add a caveat and after that Klein's posetic estimative methods are briefly described.

Caveat: When chlorophenols are taken into account we are not concerned with the complexity and structural diversity that pesticides may have. Furthermore, we are aware that the leading quantity determining the value of log K_{OW} is the number of chlorine atoms in each of the chlorophenols. Nevertheless, first we must show that posetic quantities are able to model simple systems, before we dare to model other more diverse sets of chemicals.

Klein's Posetic Approaches

At the end of the 1990's Klein started to publish[28] his ideas of using order relationships between chemical substances to estimate their properties. His methods assign a molecular structure to each substance and look for order relationships between couples of these molecular assembles. A salient feature of Klein's posetic approaches is the use of the mathematical structure of the ordered set of molecular structures, rather than focusing on the molecular structure of single molecules, which is the customary approach of traditional QSAR (Quantitative Structure-Activity Relationships) models. In QSAR models, molecular structure is characterised mainly either by fingerprints[29] or by molecular descriptors,[30] which converts the molecular structure into a vector. Handling those vectors with statistical approaches, the components of the vector, *i.e.* subsets of fingerprints or particular molecular descriptors, are mathematically related to a particular property of the studied substances. In Klein's approaches however, the structure that is worked on is the ordered structure or poset of a set of molecular structures and it is such a structure which is further used to develop estimative models of the ordered substances.[8,31,32,33,34,35–39] Klein's innovative predictive strategy has been called Quantitative SuperStructure / Activity (Property) Relationship QSSAR/QSSPR,[8,32,34] the prefix "super" accounting for the panoramic viewpoint of the structure of molecular structures.

As Klein's ordering of substances is the core of his methods, we briefly explain it as follows: an example of order relationship between molecular structures is "can be obtained from", meaning that a molecule can be obtained from another one by a defined step. Such an order relation does not necessarily consider reaction conditions, or thermodynamic constraints upon the suitability of a type of reaction. Instead, the relation is mathematized by taking into account whether the molecule associated to an under-lying molecular skeleton can be obtained from another molecule similarly based on the same skeleton. This procedure yields a poset that constitutes a progressive network in which molecular changes are introduced step-by-step on a parent molecular skeleton. An example of such a poset is the network for chlorophenols (Figure 3) where the relation "can be obtained from" is particularised to molecular substitution. There, one starts with phenol at the top of the poset and ends with pentachlorophenol at the bottom, while

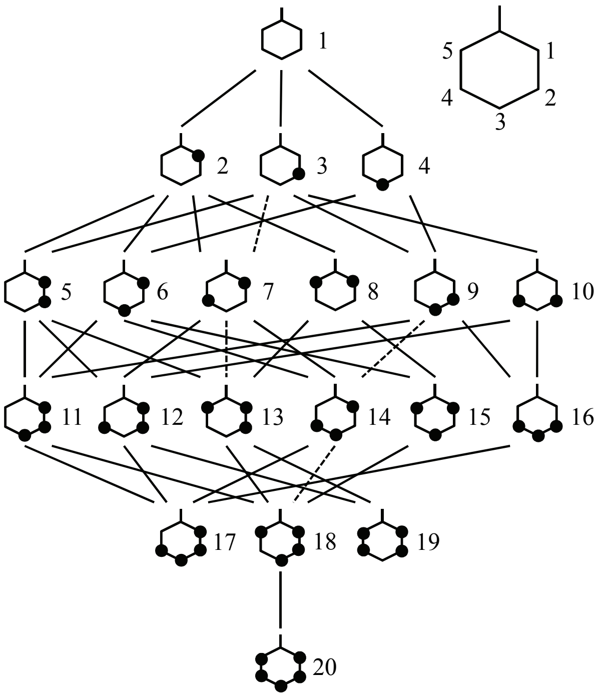

Figure 3. Substitution reaction poset of chlorophenols where the aromatic ring is represented by a hexagon and the carbon bonded to the substituted hydrogen is represented by a black circle (see reference 8). Broken lines indicate those covering relations not considered when Table 1 (see below) is the basis for coding the partial order. Top right phenol indicates the numbering convention used in the current paper.

all different patterns of substitution are found in between.

Klein's posetic approaches are: average-poset,[31] cluster-expansion,[36,39,40] and splinoid.[32,33,39] A brief description of each one of these methods is given as follows:

The *average interpolation* method estimates the property of a substance x as the average of averages of preceding and succeeding members in the poset. The first step is the calculation of the average of those compounds y directly leading to x ($y \succeq x$) in the poset; the second step is the calculation of the average of those y directly following x ($x \succeq y$); the procedure ends calculating the average of the previous two averages. This method requires knowing the property of all the nearest neighbours of the substance whose property is going to be estimated; in addition, the property cannot be estimated neither for the top nor the bottom substance of the poset, for these two substances have neither preceding nor succeeding substances, respectively.

The *cluster expansion* approach calculates the property using features $z(y)$ of all y leading to or being equal to x, which can be fitted by statistical procedures. The expansion makes use of the number of ways in which configurational arrangements $C' \in y$ occur as substructures in a configuration $C \in x$. The cluster expansion may be conveniently truncated to a limited sequence of non-zero cluster terms $z(y)$, and so applied when the earlier terms alone give a good approximation for the property P. The similarity between Taylor series and this expansion has been studied by Nava *et al.*[38] This method does not require knowing the property values for the nearest neighbours of the substance one is interested in.

The *splinoid* method assigns to each relation $y > x$ in the poset, a real variable $r_{y > x}$ ranging from 0 to 1. The idea of the approach is to follow the spline interpolation method[41] that defines a low-degree polynomial for each $r_{y > x}$. Each y of the poset is characterised by a value α_y and a slope β_y. The splinoid fit is such that each polynomial with endpoint y yields the common value = α_y at y. These splines are then smoothed as their "curvature" or "stress" is minimised, whence all the coefficients in the polynomials are determined, in terms of the set of known values, if there is a sufficient number of these. Afterwards, the unknown values gathered are expressed in terms of those known. The splinoid method considers all the cover and transitive relations of the poset.

By applying the three interpolative methods devised by Klein to the estimation of the octanol / water partition coefficient (log K_{ow}) of the chlorophenols depicted in Figure 3, Ivanciuc *et al.*[8] obtained the following results: $r = 0.987$, $s = 0.115$, with the poset-average method; $r = 0.991$, $s = 0.107$, with the cluster-expansion approach; and $r = 0.990$, $s = 0.122$ with the splinoid method.

To contrast Ivanciuc's results with those obtained by using our average height approaches, we estimate the octanol / water partition coefficient (log K_{ow}) of the same 20 substances studied by Ivanciuc *et al.*[8].

Chemical Data Set

To evaluate Equation 1, *i.e.* to order chlorophenols, we tried first a multi-indicator system for each substance by considering the five positions around the phenyl-group as components q_i of a vector characterising each chlorophenol (the numbering system used is shown in Figure 3). The value q_i for the substance x is assigned as follows:

$$q_i(x) = \begin{cases} 1 & \text{if in the i-th position a H atom is bonded} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

The data matrix presented in Table 1 shows the respective characterisation of each one of the 20 chlorophenols studied.

The poset of chlorophenols obtained from their vectors (Table 1) and by using the product order criterion (Equation 1) is quite similar to the one obtained by Ivanciuc *et al.*[8] However, the cover relations 7 $\preceq$: 3, 13 $\preceq$: 7, 14 $\preceq$: 9 and 18 $\preceq$: 14 appearing in Ivanciuc's poset do not appear in our poset. The reason for these differences is that in Ivanciuc's approach the symmetry of phenol (C_{2v}) is considered as well as the effect of substitutions on it. In our vectorial characterisation of each chlorophenol, the symmetry is disregarded: consider, *e.g.* the pair 14 $\preceq$: 9. By applying the notation of Equation 9, we find (1,0,0,1,1) for the chlorophenol 9 and (0,1,0,0,1) for chlorophenol 14. By applying Equation 1, these two chemicals are incomparable. However, by symmetry, chemical 14 can also be written as (1,0,0,1,0) and then an order relation can be established. A general procedure to apply for symmetry in case of a coding of chemical structures such as in Equation 9 is not in the focus of the current paper and, perhaps, even not needed from a practical point of view, as it is simple to code the structures in another suitable manner.

By Equation 9 a high degree of chlorine substitution is at the bottom of the poset. Therefore we compare the weak order obtained from the data of Table 1 not directly with log K_{OW} but with 6 – log K_{OW}, whereby 6 was the smallest integer larger than the values of log K_{OW} for the set of 20 chemicals. On the basis of Table 1 average heights, h_{av}, can be (and were actually) calculated, however because of the missing symmetry we suppress the results (see the remarks above, concerning symmetry).

Table 1. Chlorophenols, their characterisation (Equation 9) and log K_{OW} values. Labels correspond to those used in reference 8

Labels	Names	q_1	q_2	q_3	q_4	q_5	log K_{OW}
1	phenol	1	1	1	1	1	1.46
2	ortho-chloro-phenol	0	1	1	1	1	2.17
3	meta-chlorophenol	1	0	1	1	1	2.50
4	para-chlorophenol	1	1	0	1	1	2.40
5	1,2-dichlorophenol	0	0	1	1	1	2.94
6	1,3-dichlorophenol	0	1	0	1	1	3.22
7	1,4-dichlorophenol	0	1	1	0	1	3.09
8	1,5-dichlorophenol	0	1	1	1	0	2.74
9	2,3-dichlorophenol	1	0	0	1	1	3.17
10	2,4-dichlorophenol	1	0	1	0	1	3.20
11	1,2,3-trichlorophenol	0	0	0	1	1	3.80
12	1,2,4-trichlorophenol	0	0	1	0	1	3.69
13	1,2,5-trichlorophenol	0	0	1	1	0	3.46
14	1,3,4-trichlorophenol	0	1	0	0	1	3.89
15	1,3,5-trichlorophenol	0	1	0	1	0	3.85
16	2,3,4-trichlorophenol	1	0	0	0	1	3.99
17	1,2,3,4-tetrachlorophenol	0	0	0	0	1	4.49
18	1,2,3,5-tetrachlorophenol	0	0	0	1	0	4.36
19	1,2,4,5-tetrachlorophenol	0	0	1	0	0	4.36
20	1,2,3,4,5-pentachlorophenol	0	0	0	0	0	5.03

RESULTS

Estimating log K_{OW} with Symmetry Constraints

Averaged Heights

We derived a cover matrix for the Hasse diagram shown in Ivanciuc *et al.*[8] and considered it as the source of information to build up the vector characterising chlorophenols. The entries a_{ij} of the cover matrix are found as follows, where *i* represents the *i*-th row and *j* the *j*-th column:

$$a_{ij} = \begin{cases} 1 & \text{if } j \preceq i \\ 0 & \text{otherwise} \end{cases}$$

Hence, the vector characterising each chlorophenol has now 20 components. Each vector corresponds to a row in the cover matrix.

In Table 2 the h_{av} exact, calculated according to the lattice method and the approximate h_{av} due to LPOM0 and LPOMext are displayed.

According to h_{av} exact and to h_{av}(LPOMext), we found the following equivalence classes of chlorophenols: [2,3], [5,7], [6,9], [8,10], [11,14], [12,13], [15,16], [17,18]; and according to h_{av}(LPOM0), these ones: [2,3], [5,7,8,10], [6,9], [11,14,15,16], [12,13], [17,18].

Table 2. h_{av} of 20 chlorophenols calculated by three methods: the lattice (h_{av} exact) and the LPOM0 and LPOMext ones

chlorophenol	h_{av} exact	h_{av}(LPOM0)	h_{av}(LPOMext)
1	20.0	20.0	20.0
2	17.974	18.375	18.407
3	17.974	18.375	18.407
4	17.204	17.5	17.642
5	13.509	14.0	14.238
6	13.116	13.364	13.285
7	13.509	14.0	14.238
8	13.572	14.0	14.464
9	13.116	13.364	13.285
10	13.572	14.0	14.464
11	7.491	7.0	6.762
12	7.884	7.636	7.715
13	7.884	7.636	7.715
14	7.491	7.0	6.762
15	7.428	7.0	6.536
16	7.428	7.0	6.536
17	3.026	2.625	2.593
18	3.026	2.625	2.593
19	3.796	3.5	3.358
20	1.0	1.0	1.0

Table 3. Statistics of the linear models for 6 – log K_{OW} based on h_{av} calculated through the exact, LPOM0 and LPOMext methods

Method	r^2_{DF}	F	Slope	constant	T(constant)[(a)]	T(slope)[(a)]
Exact	0.947	343.632	0.153	1.002	10.254	18.537
LPOM0	0.936	276.818	0.145	1.089	10.460	16.638
LPOMext	0.932	262.248	0.142	1.120	10.636	16.194

[(a)] T(constant) and T(slope) are parameters allowing checking the significance of constant and slope coefficients in the regression equation. The larger their values, the better the model.

The fusion of [5,7] and [8,10] as well as that of [11,14] and [15,16] in LPOM0 arises from the additional symmetries because the incomparable objects are considered as isolated.

The regression analysis results (PASW-package®), where the h_{av} of the three methods, *i.e.* exact, LPOM0 and LPOMext, are considered as predictors for 6 – log K_{OW}, are summarised in Table 3.

Only the exact method is better than the results based on Table 1, *i.e.* without symmetry (not shown). We also found that the simpler method LPOM0 is slightly better than LPOMext. We can hypothesise that the larger equivalence classes due to the rough approximation describe 6 – log K_{OW} better than the more differentiated results of LPOMext, which may have an approximation error.

CONCLUSION

The results displayed in Table 3, seem to be promising even if advanced techniques of customary QSAR studies are not used. We did not divide the set of data into training, external and internal test sets and no attempt to check the validity (as for example in linear QSAR-models with the Williamsplot)[42–44] of the models was intended. The bare application of partial order theory should be possible when the number of comparability relations is sufficiently large. This is the case when the chemical structure is the leading idea, when one finds long posetic chains allowing analogies amongst molecular structures. In such a case the holistic view of regarding superstructures is a possible way, and if the procedure shown in the current paper is accepted, then the typical tools of order theory, such as average heights can be useful too. We think that the very idea, realised by Klein and his colleagues, is to see the chemicals not as a set of single entities but embedded in a network which itself has a superstructure. This network can often be a network derived from order relations and as order relations can be defined in many different ways, there is a chance for new modelling techniques in QSAR. Structures may additionally be classified in a hierarchy fashion, *i.e.* molecules, classes of molecules with some common structural properties and finally the "superstructure" as introduced by Klein. Mathematically, a method has been developed, by which the classification of structures can be modelled in posetic terms.[45] It is then worth pursuing this kind of approach for Klein's posets of molecular structures.

OUTLOOK

Even if the posetic results based on posetic quantities such as heights seem to be promising, there are several points which we did not include or which should be considered before the concept of posets can be generally used. First of all, as already mentioned in the caveat of the material and methods section, another set of more complex molecules should be selected. However the relevant constraint is that the resulting poset must have a wealth amount of comparabilities (a superstructure in Klein's terms) to be used in the methodology here presented as well as in Klein's ones, *e.g.* a set of molecules leading to a complete antichain cannot be analysed. This requirement hampers the dissection of the set of molecules in training and test sets. Furthermore, even if such a superstructure exists, the height alone may be not sufficient to model the properties of the compounds under investigation. Other posetic quantities such as local ones, as discussed in the material and methods section, may be useful too.

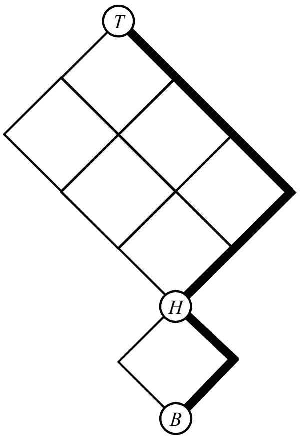

Figure A. "Manhattan-like" graph representing a digraph that is discovered by shortest walks exclusively either from the bottom *B via H* to the top *T* or the other way round. An example of such a shortest walks is shown in bold face.

Another argument is that we did not use sophisticated approaches of current QSAR techniques as the already mentioned division of molecules in training and test sets and the internal and external validation of the model. The way out, namely a cross validation, was not applied, for this paper is thought of as a starting document with the focus on posetic approaches.

Clearly, it is possible that we have to abandon the idea of posetic approaches as a kind of modelling methods, if results with more diverse datasets indicate so. Should that happen, then posetic approaches may be seen as a tool to understand the leading structural factors and the substitution patterns in superstructures that determine physicochemical or toxicological quantities; or as a Klein has pointed out, as new chapters of stereochemistry.[10]

APPENDIX

The graph representing the lattice of all down sets of a poset needs less storage than the storage of the linear extensions in LT. Figure A may help to understand that it is needed less storage when considering the lattice.

The graph of Figure A has 15 vertices and two "*n,m*-grids", whereby *n* and *m* are the numbers of edges in both directions. There is a 1,1-grid between *B* and *H*, and a 3,2-grid between *H* and *T*. Applying the formula[46] for shortest walks in *n,m*-(Manhattan-like) grids: $\binom{n+m}{m}$, we obtain 20 shortest walks. Assume now that any shortest walk from *B* to *T* represents a linear extension, then 20 linear extensions are stored by only needing the storage of the adjacency matrix of the 15 vertices.

REFERENCES

1. W. J. Lyman, W. F. Reehl, and D. H. Rosenblatt, *Handbook of Chemical Property Estimation Methods, Environmental Behavior of Organic Compounds*, McGraw-Hill, New York, 1982.
2. E. J. Baum, *Chemical Property Estimation Theory and Application*, Lewis Publishers, Boca Raton, 1998.
3. R. Bruggemann and J. Altschuh, *Sci. Total Environ*. **109−110** (1991) 41−57.
4. R. Bruggemann, G. Restrepo, and K. Voigt, *J. Chem. Inf. Model.* **46** (2006) 894−902.
5. L. Carlsen, *Internet Electron. J. Mol. Des*. **4** (2005) 355−366.
6. L. Carlsen, P. B. Sørensen, and M. Thomsen, *Chemosphere* **43** (2001) 295−302.
7. R. Bruggemann, S. Pudenz, L. Carlsen, P. B. Sørensen, M. Thomsen, and R. K. Mishra, *SAR QSAR Environ.Res*. **11** (2001) 473−487.
8. T. Ivanciuc, O. Ivanciuc, and D. J. Klein, *Int. J. Mol. Sci.* **7** (2006) 358−374.
9. D. J. Klein and J. Brickmann, *MATCH Commun. Math. Comput. Chem.* **42** (2000) 1−290.
10. D. J. Klein and D. Babić, *J. Chem. Inf. Comput. Sci*. **37** (1997) 656−671.
11. B. A. Davey and H. A. Priestley, *Introduction to Lattices and Order,* Cambridge University Press, Cambridge, 1990.
12. R. Bruggemann and K. Voigt, *Comb. Chem. High T. Scr.* **11** (2008) 756−769.
13. R. Bruggemann and G. P. Patil, *Ranking and Prioritization for Multi-indicator Systems - Introduction to Partial Order Applications*, Springer, New York, 2011.
14. P. Winkler, *Discr. Math*. **39** (1982) 337−341.
15. R. Bubley and M. Dyer, *Discr. Math*. **201** (1999) 81−88.
16. K. De Loof, H. De Meyer, and B. De Baets, *Fund. Inform*. **71** (2006) 309−321.
17. R. P. Stanley, *Enumerative Combinatorics,* Volume I, Wadsworth&Brooks/cole, Monterey, 1986.
18. K. De Loof, B. De Baets, H. De Meyer, and R. Brüggemann, *Comb. Chem. High T. Scr*. **11** (2008) 734−744.
19. Icell software, URL: http://bio.math.berkeley.edu/ranktests/lcell/ (accessed 23rd May 2013).
20. R. Bruggemann and K. Voigt, *Analysis of Partial Orders in Environmental Systems Applying the New Software PyHasse*, in: J. Wittmann and M. Flechsig (Eds.), *Simulation in Umwelt- und Geowissenschaften*- Workshop Potsdam 2009, Shaker-Verlag, Aachen, 2009, pp. 43−55.
21. K. Voigt, R. Bruggemann, M. Kirchner, and K.-W. Schramm, *Environ. Sci. Pollut. Res*. **17** (2010) 429−440.
22. K. Voigt, R. Bruggemann, H. Scherb, H. Shen, and K.-H. Schramm, *Environ. Modell. Softw*. **25** (2010) 1801−1812.
23. PyHasse software, URL: http://pyhasse.org (accessed 23rd May 2013).
24. R. Bruggemann and L. Carlsen, *MATCH Commun. Math. Comput. Chem*. **65** (2011) 383−414.
25. R. Bruggemann, P. B. Sørensen, D. Lerche, and L. Carlsen, *J. Chem. Inf. Comput. Sci*. **44** (2004) 618−625.
26. K. De Loof, B. De Baets, and H. De Meyer, *MATCH Commun. Math. Comput. Chem*. **66** (2011) 219−229.
27. R. Bruggemann, U. Simon, and S. Mey, *MATCH Commun. Math. Comput. Chem*. **54** (2005) 489−518.
28. D. J. Klein, T. G. Schmalz, and L. Bytautas, *SAR QSAR Environ. Res*. **10** (1999) 131−156.
29. P. Willett, *Similarity Searching Using 2D Structural Fingerprints*, in: J. Bajorath, (Ed.), *Chemoinformatics and Computational Chemical Biology*, Humana Press, Berlin, 2011, pp. 133−158.
30. R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, Volume I: Alphabetical Listing, Wiley-VCH, Weinheim, 2009.
31. T. Ivanciuc and D. J. Klein, *J. Chem. Inf. Comput. Sci*. **44** (2004) 610−617.
32. T. Ivanciuc, O. Ivanciuc, and D. J. Klein, *J. Chem. Inf. Model*. **45** (2005), 870−879.
33. T. Došlić and D. J. Klein, *J. Comput. Appl. Math*. **177** (2005) 175−185.
34. T. Ivanciuc, O. Ivanciuc, and D. J. Klein, *Mol. Divers*. **10** (2006) 133−145.
35. D. J. Klein and T. Ivanciuc, *Directed Reaction Graphs as Posets*, in: R. Bruggemann and L. Carlsen (Eds.), *Partial Order in Environmental Sciences and Chemistry,* Springer, Berlin, 2006, pp. 35−60.
36. T. Ivanciuc, D. J. Klein, and O. Ivanciuc, *J. Math. Chem*. **41** (2007) 355−379.
37. D. J. Klein, T. Ivanciuc, A. Ryzhov, and O. Ivanciuc, *Comb. Chem. High T. Scr*. **11** (2008) 723−733.
38. J. Nava, V. Kreinovich, G. Restrepo, and D. Klein, *J. Uncert. Syst*. **4** (2010) 270−290.
39. G. Restrepo and D. J. Klein, *J. Math. Chem*. **49** (2011) 1311−1321.
40. D. J. Klein, T. G. Schmalz and L. Bytautas, *SAR QSAR Environ. Res*. **10** (1999) 131−156.
41. D. Kincaid and W. Cheney, *Numerical Analysis*, Brooks/Cole, Pacific Grove, 1990.

42. P. Gramatica, *QSAR Comb. Sci*. **26** (2007) 694−701.
43. R. Khosrokhavar, J. B. Ghasemi, and F. Shiri, *Int. J. Mol. Sci*. **11** (2010) 3052−3068.
44. A. Tropsha, *Mol. Inf.* **29** (2010) 476−488.
45. G. Restrepo and R. Bruggemann, *J. Math. Chem*. **44** (2008) 577−602.
46. V. Bryant, Aspects of Combinatorics - *A Wide Ranging Introduction*, Cambridge University Press, Cambridge, 1993.