JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

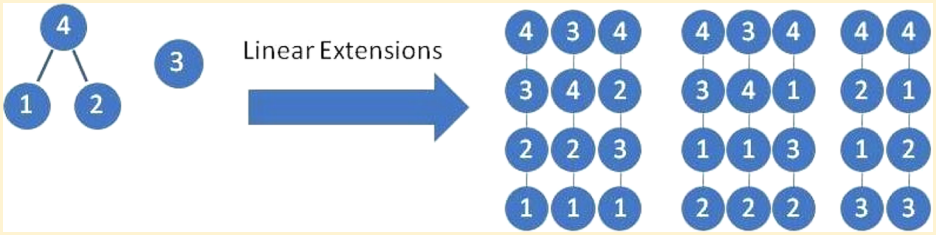
Ranking COMMPS Chemical Substances by an Improved POT/RLE Approach

Irene Díaz,*[,†] Elías Combarro,[†] Pasquale Marinaro,[‡,†] and Luigi Troiano*[,§]

[†]Department of Computer Science, University of Oviedo, Spain

[‡]Intelligentia s.r.l, Benevento, Italy

[§]Department of Engineering, University of Sannio, Italy



ABSTRACT: The combined monitoring-based and modeling-based priority setting (COMMPS) provides a procedure for the identification of priority hazardous substances outlined in the Working Document (ENV/191000/01 of January 16, 2001). This procedure is based on scoring a set of criteria which individually make substances more or less hazardous. The way scores are weighted and combined has been established by a panel of experts. Different authors outlined how such a procedure might be affected by subjectiveness of judgment, and alternative solutions based on partial order theory (POT) and random linear extensions (RLE) have been suggested. This method consists of generating a set of RLE and of averaging the rank given to each substance, so that a total order could be determined. Any POT/RLE approach must face the issue of covering as much as possible the space of linear extensions that, in the case of the 85 substances considered by COMMPS, becomes extremely large, and an exhaustive generation of linear extension is not feasible. Therefore, having a faster algorithm would help to consider a larger number of linear extensions in a given time frame. In this paper, we discuss this problem, and we outline a possible solution.

INTRODUCTION

Water is at the core of natural ecosystems and climate regulation. Water pollution and scarcity represent threats to human health and quality of life. In addition, a shortage of good-quality water damages aquatic, wetland, and terrestrial environments. Therefore, as without water, no life can survive, it must be protected. In 2000, the EU adopted the Water Framework Directive (Directive 2000/60/EC), establishing a legal obligation to protect and restore the quality of waters across Europe.[1]

Pollution is one of the most serious problem over the world. Hazardous chemicals find their way into European waters from industrial plants or farmland sites. In fact, it is estimated that only 30% of surface water and 25% of groundwater is not at serious risk from pollution and other changes.

The Environmental Quality Standards Directive limits, since 2008, concentrations in surface waters of 33 priority substances and eight other pollutants. They include 11 priority hazardous substances, which are toxic, persistent, and accumulate in animal and plant tissues, posing a long-term risk. Discharges must be phased out within 20 years. The list was reviewed in 2011, adding 15 new priority substances, six of them designated as priority hazardous substances. The choice of substances on the priority list within the Water Framework Directive is important since the EU member states are obliged to establish a monitoring network for them.

In order to establish a list of priority substances in accordance with the given provisions, a combined monitoring-based and modeling-based priority setting scheme (COMMPS) has been elaborated. According to the final proposal for a list of priority substances in the context of the water framework directive,[2] the COMMPS procedure comprises several steps for producing a list of hazardous substances.

The COMMPS procedure is based on the identification of four sublists: a monitoring-based list for organic substances in the aquatic environment, a modeling-based list for organic substances in the aquatic environment, a monitoring-based list for organic substances in the sediment, and a monitoring list for metals.

From these lists, the substances were selected for evaluation in relation to the final list (20 from each organic substances list, 10 from the list based on sediment monitoring, and five from the list based on metals). After an evaluation process, some substances were excluded, and the top 20 substances from the monitoring-based list for organic substances in the aquatic environment were included in the final list.[1,3]All the 20 top substances examined on the monitoring-based list for organic substances in the aquatic environment were included in the

Received: July 25, 2013
Published: December 1, 2013

dx.doi.org/10.1021/ci4004392 | J. Chem. Inf. Model. 2013, 53, 3190−3201

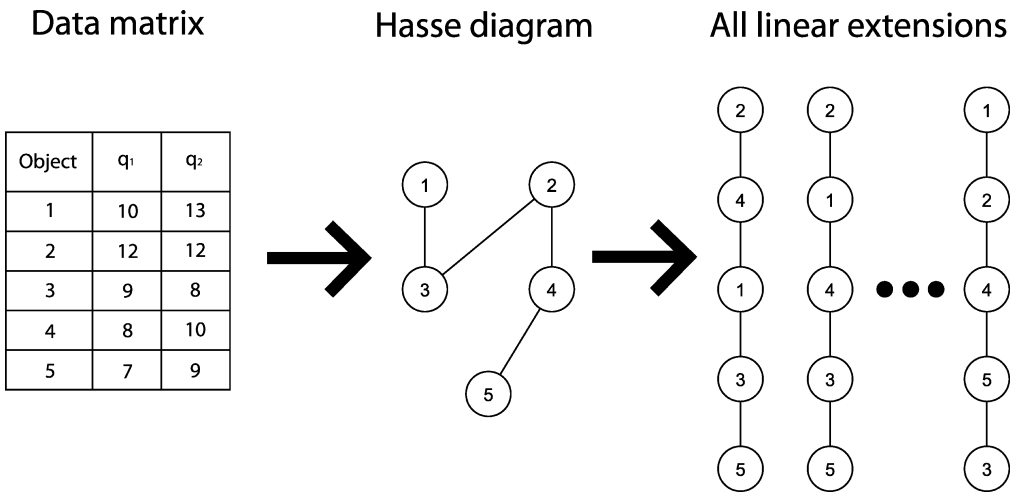


Figure 1. Data matrix, Hasse diagram, and total set of linear extensions.

final list of 32 substances; therefore, this list greatly influences the final one.

The rankings are based on a single score for each individual substance, obtained from the aggregation of various descriptors. The score is calculated by a combination of an index function that describes the exposure of a substance (I_{EXP}) and another index function that describes the effect of a substance (I_{EFF}) as follows:

$$\text{score} = I_{EXP} \times I_{EFF} \tag{1}$$

The higher the score, the bigger the associated risk. The exposure index is calculated for each substance i using only the 90% percentile (C_i) of the sampling values, and it is defined by

$$I_{EXP} = 10 \times \frac{\log\left(\frac{C_i}{C_{min}0.1}\right)}{\log\left(\frac{C_{max}}{C_{min}0.1}\right)} \tag{2}$$

The exposure index is scaled by defining an upper (C_{max}) and a lower (C_{min}) limit of concentration. The effects assessment in COMMPS, I_{EFF}, is computed by

$$I_{EFF} = \frac{5}{10}EFS_d + \frac{3}{10}EFS_i + \frac{2}{10}EFS_h \tag{3}$$

The direct effect score, EFS_d, is based on the predicted no effect concentration (PNEC), and it is defined by

$$EFS_d = 5 \times \frac{\log\left(\frac{PNEC}{10 \times PNEC_{max}}\right)}{\log\left(\frac{PNEC_{min}}{10 \times PNEC_{max}}\right)} \tag{4}$$

The indirect effect EFS_i is supposed to be correlated with the substance's ability to bioaccumulate. Finally, the EFS_h score is established using carcinogenicity, mutagenicity, and effect on reproduction properties (for further information, the authors refer to the European Commission[1] or to the works by Lerche and co-workers[3]).

Therefore, COMMPS has a certain degree of subjectivity because the weight of any factor is taken from judgements provided by experts from the EU member states and because of the specific functional form the ranking index has. The COMMPS procedure belongs to the class of *scoring* methods. However, other priority setting methods based on total or partial order theory are suitable for establishing a ranking of chemical substances. Total order methods can be used to rank substances on the basis of multiple criteria which are combined into a global ranking index using approaches different from scoring, such as *desirability*, *utility*, and *dominance* functions among others (see Pavan and Worth's work[4]).

In particular, desirability and utility functions[5] assign a utility/desirability value to each criterion individually; then they merge values from different criteria of a single substance using a function which provides an aggregated value. This value is used to produce a ranking. Examples of aggregation functions are the arithmetic mean used to aggregate utility values and the geometric mean for desirability values.

The dominance method compares sets of criteria where a substance can be considered better than others. This approach is commonly used to identify different profiles of behavior of the substances with regard to the different criteria.[4]

All methods above are somehow subjective, relying on the choice of functions and weights. This fact could make them strongly dependent on subjective inputs.[3]

An alternative approach relies on Partial Order Theory (POT). In this case, we just look at point-wise comparison between substances, establishing a relationship of priority when *all* indicators point out that one substance is more hazardous than another. This leads to methods which require lesser assumptions concerning functional relationships and do not apply weight factors of the individual criteria. They stand for recognizing that different criteria can be conflicting because not all substances can be compared with each other. Several partial order methods have been proposed to rank hazardous substances and to compare them to the index approach.[6−9] Some of these works are also related to the production of Random Linear Extensions (RLE) from the predefined partial order.[9,3,10]

In this study, a procedure for ranking hazardous substances is presented. This procedure is based on both Partial Order Theory (POT) and Random Linear Extensions (RLE). In particular, this work focuses on how to obtain efficiently linear extensions in order to faster reach a stable condition in averaging substance ranks and thus to better cover the space of linear extensions. Further, this approach is analyzed and compared to other existing methodologies.

dx.doi.org/10.1021/ci4004392 | J. Chem. Inf. Model. 2013, 53, 3190−3201

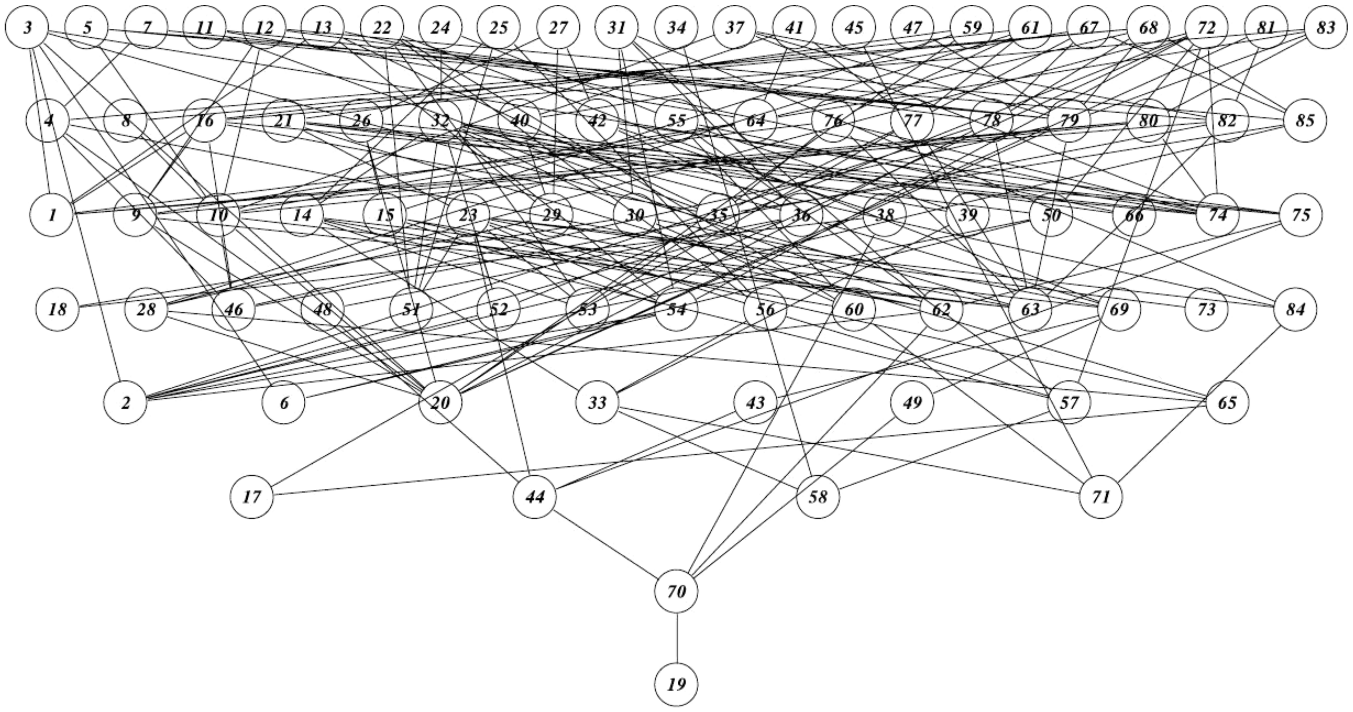


Figure 2. Hasse diagram of chemical substances indexed by COMMPS.[3] The numbers are the IDs assigned to the substances (see Table 6).

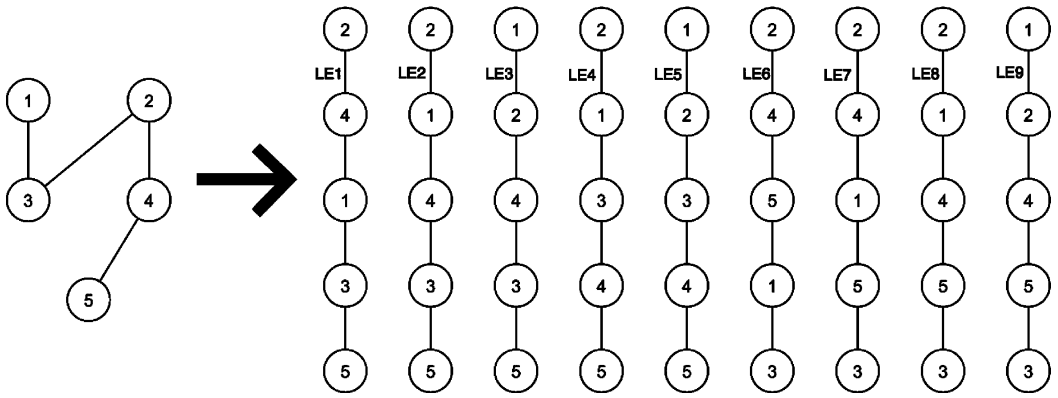


Figure 3. Hasse diagram of a poset and its linear extensions.

x_i	$rank(x_i)$									$E[rank(x_i)]$
	LE1	LE2	LE3	LE4	LE5	LE6	LE7	LE8	LE9	
1	3	2	1	2	1	4	3	2	1	2.111
2	1	1	2	1	2	1	1	1	2	1.333
3	4	4	4	3	3	5	5	5	5	4.222
4	2	3	3	4	4	2	2	3	3	2.889
5	5	5	5	5	5	3	4	4	4	4.444

Figure 4. Expected ranks of each element $x_i = \{1,2,3,4,5\}$ according to poset and each linear extension LEi in Figure 3. Each row i of this table represents the position of the element i in the corresponding linear extension. For example, element 1 appears in position 3 in the first linear extension shown in Figure 3, in position 2 in the second linear extension, in position 1 in the third linear extension, and so on.

RELATED WORK

Ranking methods[11,12] based on Hasse diagrams are important in a wide context. They have been proposed as an alternative approach to identify and classify environmental risks.[13−16] They make combined use of Partial Order Theory and Random Linear Extention (POT/RLE). In this section, we provide a brief review of contributions from both theoretical and application side.

Partial Ordered Sets and Hasse Diagrams. Let us consider a finite poset $(P,\leq)$ (or P for short) of p elements. A subset I of P is an **ideal** if for any $a \in I$ and any $b \in P$ such that $b \leq a$, it follows that $b \in I$. A poset can be represented by the so-called *Hasse diagram*,[17] a graph where $a \leq b$ if and only if

dx.doi.org/10.1021/ci4004392 | J. Chem. Inf. Model. 2013, 53, 3190−3201

there is a sequence of connected lines upward from a to b. An example of a Hasse diagram is given in Figure 1. Note that this diagram has an orientation and that a Hasse diagram is a graph without cycles; therefore if $a \leq b$ and $b \leq c$, then $a \leq c$, but a line between a and c is not drawn because a and c are already connected by a path of size 2 (the path going first from a to b and then from b to c).

The concept of poset was deeply used in environmental sciences by Bruggemann et al.[6−8,10] In this case, the chemicals are compared by studying the value of their descriptors. If two substances are compared with regard to these descriptors and all the values are higher than or equal (respectively lower than or equal), then it is possible to rank these two substances (representing this rank by a line between the two substances in the Hasse diagram). However, this is not very common, and often we find many pairs of incomparable substances.

When Hasse diagrams are used in the environmental hazard framework, it is assumed that the higher the numerical value of a descriptor, the more hazardous the substance. Therefore, the *maximal* elements (the ones at the top of the diagram) are the most hazardous. Thus, a first look to the Hasse diagram allows one to establish different levels of hazardousness.

Linear Extensions. As it was detailed in the former section, Hasse diagrams order the substances into groups according to their level of hazardousness. However, in environmental hazards, it is often interesting to produce a ranking of hazardous substances, that is, a total order. A useful tool for producing a total order from a partial order is the identification of linear extensions.

An **extension** $(P,\leq')$ of the poset $(P,\leq)$ is another poset over the same referential P such that $\leq'$ is order-preserving (i.e., if $x \leq y$, then $x \leq' y$). A **linear extension** is an extension that is a total order.

Figure 1 shows the Hasse diagram and the set of linear extensions obtained from a set of eight elements (a to h) described by two characteristics (q_1 and q_2).

The problem of determining the linear extensions of a poset is a long-standing combinatorial open problem. In fact, it was proven by Brightwell and Winkler[18] that it is a #*P*-complete problem. As this problem is strongly related to any ordering-based problem, there are several papers focused on producing linear extensions in an efficient way. The method in ref 19 is based on graph counting operations on the lattice of ideals representation of the given poset, and it basically selects a path from the source (the empty ideal) to the sink (the whole poset) producing a linear extension.[20] It uses the lattice of ideals instead of directly enumerating all linear extensions because the number of linear extensions is in general much larger than the number of ideals. However, it cannot be applied in practice for large values of n.[21] For this reason, several heuristics have been proposed to cope with this problem. Although they do not lead in general to a uniform algorithm, they can be applied in more situations.

There exist other methods for generating linear extensions based on Markov chains. The first procedure was proposed by Karzanov and Khachiyan in ref 22 (see also refs 23 and 24). The problem with this kind of procedure is that it is hard to know whether the end of the process is close or not, so that we do not know the number of iterations that should be done in order to be near uniformity.

Linear Extensions and Mutual Ranking Probabilities (*MRP*). One of the most successful algorithms for randomly generating linear extensions from a poset was presented by Lerche and Sörensen.[16] It is based on deriving the ranking probabilities based on the relationships in the partial order and then on deciding the order between pairs of incomparable elements. However, it is necessary to obtain the *Mutual Ranking Probability*[19] (the probability of the first element of the pair being before the second element of the pair in a linear extension) for each pair of incomparable elements, which is a

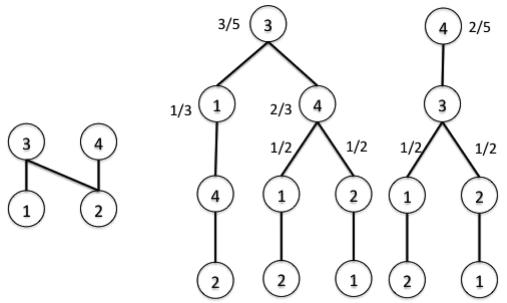


Figure 5. The left part of the figure contains a poset. The right part represents the way the algorithm works.

Table 1. Probability of Each Possible Linear Extension According to Figure 5

linear extension	probability
3−1−4−2	(3/5) × (1/3) = (1/5)
3−4−1−2	(3/5) × (2/3) × (1/2) = (1/5)
3−4−2−1	(3/5) × (2/3) × (1/2) = (1/5)
4−3−1−2	(2/5) × (1/2) = (1/5)
4−3−2−1	(2/5) × (1/2) = (1/5)

Table 2. Top 20 Substances Selected by the COMMPS Procedure

ID	compound	rank	ID	compound	rank
59	indeno(1,2,3-cd) pyrene	85	31	DDD,4,4′-isomer	75
21	benzo-a-anthracene	84	27	chlorpyrifos	74
24	benzo-g,h,i-perylene	83	56	hexachlorobenzene	73
32	DDE, 4,4′-isomer	82	11	3-chloronitrobenzene	72
23	benzo-b-fluoroanthene	81	39	dieldrin	71
72	pentachlorophenol	80	DDD, 2,4′-isomer	70	
25	benzo-k-fluoroanthene	79	3	1,2,4-trichlorobenzene	69
22	benzo-a-pyrene	78	26	chlorfenvinphos	68
14	aldrin	77	41	diuron	67
55	heptachlor	76	34	DDT,4,4-isomer	66

Table 3. Top 20 Substances Selected by *Maximals* Procedure

ID	compound	rank	ID	compound	rank
24	benzo-g,h,i-perylene	85	83	trichloromethane	75
59	indeno(1,2,3-cd)pyrene	84	67	metolachlor	74
27	chlorpyrifos	83	32	DDE, 4,4′-isomer	73
31	DDD,4,4′-isomer	82	21	benzo-a-anthracene	72
25	benzo-k-fluoroanthene	81	12	acenaphthene	71
22	benzo-a-pyrene	80	11	3-chloronitrobenzene	70
61	isoproturon	79	37	dichloromethane	69
68	naphthalene	78	13	alachlor	68
41	diuron	77	42	endosulfan,alpha-isomer	67
72	pentachlorophenol	76	5	1,2-dichloroethane	66

dx.doi.org/10.1021/ci4004392 | J. Chem. Inf. Model. 2013, 53, 3190−3201

time-consuming task (see ref 25). Thus, Lerche et al.[9] propose to estimate these values by considering the number of elements above and below each element of the pair. The results of this algorithm for the problem of ranking hazardous substances can be seen in ref 3.

The main steps of the algorithm are listed below (it is also described in ref 21).

1. Consider all of the incomparable pairs in the poset.
2. Select one of them randomly.
3. Decide the lower element of the pair in the linear extension. This is done randomly, according to an estimation of the mutual ranking probabilities. Note that the maximum ranking position for an element is the total number of elements in the poset minus the number of elements above that element; similarly, the minimum ranking position an element can attain is the number of objects below this element plus one. Given a pair of incomparable elements and the maximum and minimum ranking positions of the elements of the pair, we can obtain the possible ranking positions in a random linear extension; finally, the probability of the first element of the pair considered below the second element of the pair is estimated as the proportion of possible ranking positions in this condition.
4. Compute the transitive closure according to the new order constraint introduced.
5. Repeat the process until no incomparable pairs remain.

METHODOLOGY

Ranking Procedure. In this section, we consider the problem of ranking the 85 substances indexed by COMMPS. Hazardous substances are characterized by the four descriptors C_i, EFS_d, EFS_i, and EFS_h (see Table 6). We establish the following partial order relation: *if the values for all descriptors of a substance a are higher than the corresponding values for another substance b, then a is ranked above b, which means a is more hazardous than b*. Indeed, in this case, we are sure that *a* is more hazardous than *b* as all indicators point out that conclusion. However, if there exists a contradiction between criteria, we are not allowed to reach any conclusion.

Figure 2 shows the visualization of the partial order by a Hasse diagram for all 85 substances. According to it, the most hazardous substances are those on the top of the diagram. Thus, from the 85 selected substances, 23 chemicals are on the top level of the Hasse diagram (Figure 2) and hence considered as the most hazardous, and also they are pairwise incomparable one to each other.

Ranking methods assign a rank equal to 1 to the most desirable compound (the least hazardous) while an 85 rank corresponds to the most hazardous substance. Therefore, a good ranking method should identify as the most hazardous substances as many of these 23 substances as possible.

Ranking substances according to the partial order outlined by Figure 2 means considering a linear extension of substances, that is, a complete linear order of substances compatible with all the relations expressed by the partial order. However, several linear extensions are possible, as the relative position of incomparable substances can be inverted. Therefore, we should consider the *expected linear extension* (ELE), that is, the linear extension obtained by averaging the position of each substance over all of the possible linear extension. In the cases when the poset has nontrivial order-automorphisms and, thus, elements which are indistinguishable, it is not possible to obtain a total order (only a weak one, with several elements ranked in the same position) from the linear extensions of the poset. Notice that this happens not only with the algorithm we propose here, but with every method based on generating the "average" linear extension of the poset. However, these cases could be identified when they happen and managed accordingly.

Let us consider an example. Figure 3 shows the Hasse diagram of a poset with five elements. All possible linear extensions which are compatible with the given poset are shown on the right side.

The ELE can be determined according to the position assumed by each element along the different linear extensions, as Figure 4 shows.

However, enumerating all of the linear extensions compatible with a given partial order might be infeasible, due to the large number of combinations.[3] For example, looking at Figure 2 and taking into account that the number of linear extensions depends on the number of incomparable elements, we know that the number of linear extensions associated with this Hasse diagram is larger than 2.6×10^{22} (the number of incomparable elements at the top of the Hasse diagram). In that case, sampling the space of linear extensions and averaging over a set of random linear extensions would be the only viable choice to compute the ELE. The ranking procedure based on computing an ELE from a sample of linear extensions is straightforward and outlined by Algorithm 1.

The expected linear extension is obtained by sorting substances according to the *average rank* they assume over the *m* random linear extensions generated along the procedure (line 2). In this, our approach does not differ from other methods.[3,9,10]

In order to obtain stable average positions, the sample should cover as much as possible the space of linear extensions. Therefore, it is important to have a RLE generation procedure which is fast enough to guarantee a high throughput. This can be obtained by choosing the next element at each step among the dominating elements. To better cover the space of linear extensions, it is also important that sampling is closer to being uniform, to avoid any bias. In the remainder of this section, we will illustrate a heuristic able to generate linear extensions efficiently, selecting each element with a different probability which reflects the number of possible compatible alternatives.

Algorithm 1 Ranking Procedure

Input: X, COMMPS substances
Input: G, POT adjacency matrix
Output: $P \equiv (x_{(1)}, x_{(2)}, \ldots, x_{(n)})$, the expected linear extension

1: **for** i = 1 **to** m **do**
2: $\xi_i \leftarrow RLE(X, G)$
3: **end for**
4: $P \leftarrow X$ sorted by $E[\{\xi_i\}]$

dx.doi.org/10.1021/ci4004392 | J. Chem. Inf. Model. 2013, 53, 3190−3201

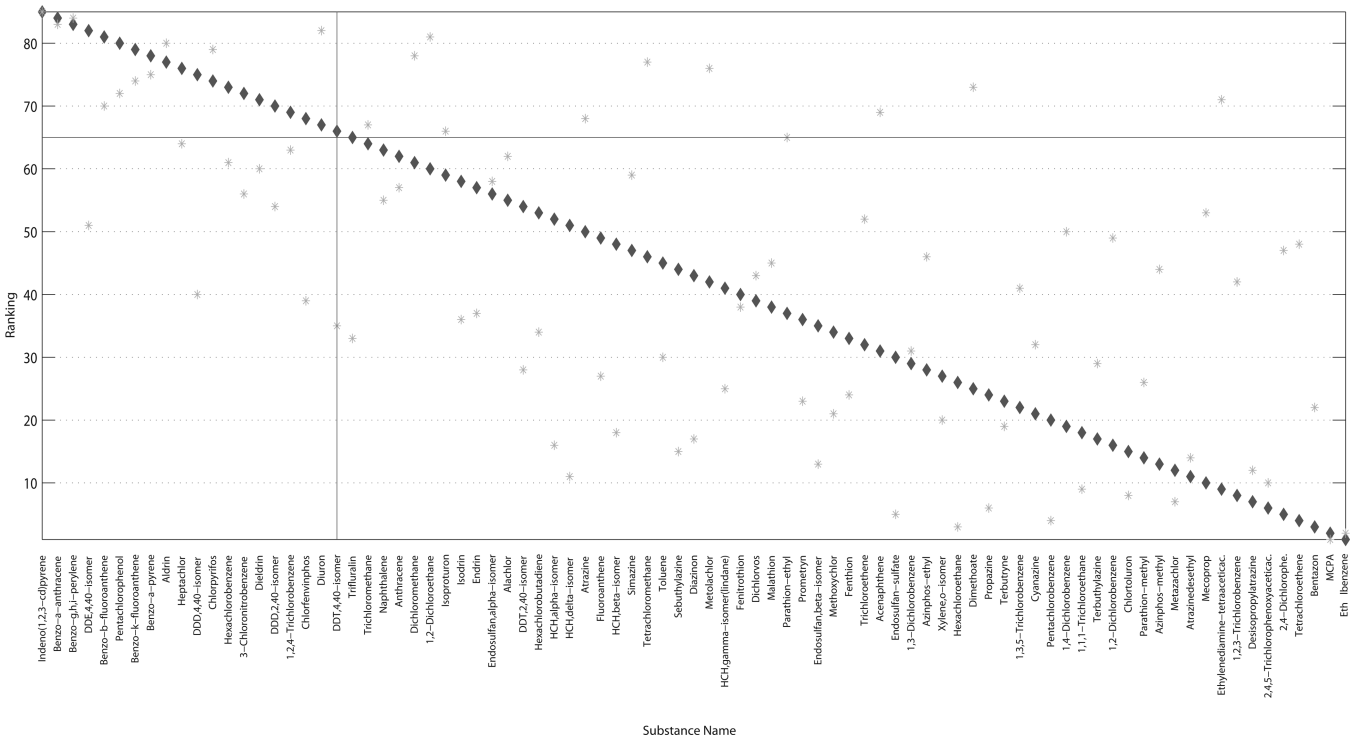


Figure 6. Comparison between COMMPS and *Maximals* rankings. The squares represent the rank of a substance given by COMMPS, while the stars represent the same according to *Maximals*. On the *x* axis, the substances' IDs are represented following the order provided by COMMPS, while on the *y* axis, the ranking is represented.

Heuristics. In order to approximate the linear extensions from a partial order relation defined over a set X, we basically select the element to be placed at each position according to an estimated probability computed following eq 6.

If $E \equiv \{e_1, e_2, ..., e_m\}$ is the set of *maximal* elements of the poset, i.e.,

$$E \equiv \{e \in X | e \not< x, \forall\, x \in X\} \tag{5}$$

let n_e be the number of elements x s.t. $x < e$, i.e., elements dominated by e; then the probability of selecting e as the next element of the linear extension is estimated by

$$P_e = \frac{n_e}{\sum_{k \in E} n_k} \tag{6}$$

This heuristic method, outlined in Algorithm 2, will be called hereafter *Maximal*. At the first step, Ξ and Γ are respectively set to X and G (line 1). The linear extension is built across the main loop (lines 2–9). At each step, we collect in E_i the maximal elements in Ξ according to Γ (line 3). For each element $e \in E_i$, we compute the selection probability given by eq 6 (lines 4–5). Selection of the next elements $x_{(i)}$ is performed randomly in E_i according to selection probabilities $\{P_e\}$ (line 7). Ξ and Γ are prepared for the next step by removing element $x_{(i)}$ (line 8). The process is reiterated until all elements are selected. As output, we get a linear extension, i.e., a total order $x_{(1)}, x_{(2)}, ..., x_{(n)}$ of chemical substances.

Let us show an example of this procedure. Consider the Hasse diagram shown in Figure 5. The first step (line 3)

Algorithm 2 Algorithm "Maximals", $RLE(X, G)$

Input: X, COMMPS substances
Input: G, POT adjacency matrix
Output: $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$, the linear extension

1: $\Xi_1 \leftarrow X, \Gamma_1 \leftarrow G$
2: **for** $i = 1$ **to** n **do**
3: $E_i \leftarrow \{e \in \Xi_i \,|\, e \not< x$ according to $\Gamma_i \forall x \in \Xi_i\}$
4: **for all** $e \in E_i$ **do**
5: $P_e \leftarrow \frac{n_e}{\sum_{k \in E} n_k}$
6: **end for**
7: $x_{(i)} \leftarrow select(E_i, \{P_e\})$
8: $\Xi_{i+1} \leftarrow \Xi_i$, $\Gamma_{i+1} \leftarrow \Gamma_i$ cleared of $x_{(i)}$
9: **end for**

dx.doi.org/10.1021/ci4004392 | J. Chem. Inf. Model. 2013, 53, 3190−3201

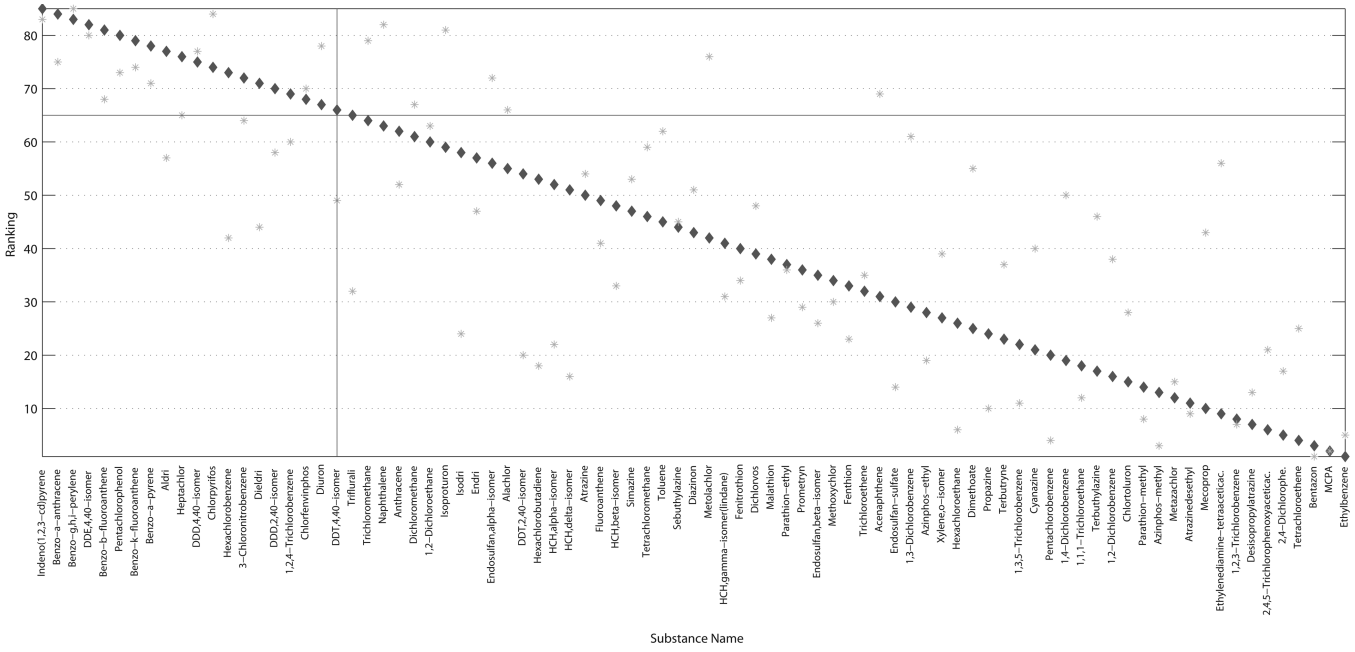


Figure 7. Comparison between COMMPS and *MRP* rankings. The squares represent the rank of a substance given by COMMPS, while the stars represent the same according to *MRP*. On the *x* axis, the substances' IDs are represented following the order provided by COMMPS, while on the *y* axis, the ranking is represented.

Table 4. Comparison of Execution Time When 100 000 Linear Extensions Are Generated for Sets Whose Size Ranges from 3 to 8

	|X|					
method	3	4	5	6	7	8
MRP	2.40	8.39	59.75	467.36	3083.32	28312.22
Maximals	2.28	4.30	8.70	23.04	122.06	771.38

considers the maximal elements {3,4}. One of them is randomly selected according to the probability estimations given by eq 6 (lines 4−6). The probabilities are in this case $P(3) = (3/5)$, $P(4) = (2/5)$. This process is then repeated until all elements are completely ordered. Assume the element selected is 4. The next iteration will select the next element in the linear extension from the set {1,2,3}. According to the partial ordering represented by the poset, we have just one possible choice (3). The next element in the linear extension is randomly selected from the {1,2} set according to the probability distribution $P(1) = 1/2$, $P(2) = 1/2$. Assume 2 is

Table 5. Top 20 Substances Selected by COMMPS Procedure[a]

ID	compound	*Maximals*	*MRP*	Desirability	Utility	Dominance
59	**indeno(1,2,3-cd)pyrene**	**X**	**X**	**X**	**X**	**X**
21	benzo-a-anthracene	X	X		X	X
24	**benzo-g,h,i-perylene**	**X**	**X**	**X**	**X**	**X**
32	DDE, 4,4′- isomer		X	X	X	
23	benzo-b-fluoroanthene	X	X		X	X
72	**pentachlorophenol**	**X**	**X**	**X**	**X**	**X**
25	benzo-k-fluoroanthene	X	X	X	X	
22	benzo-a-pyrene	X	X	X	X	
14	aldrin	X			X	
55	heptachlor				X	
31	DDD, 4,4′-isomer		X	X	X	
27	chlorpyrifos	X	X	X		
56	hexachlorobenzene				X	
11	3-chloronitrobenzene	X		X	X	X
39	dieldrin				X	
30	DDD, 2,4′-isomer			X		
3	*1,2,4-trichlorobenzene*					
26	chlorfenvinphos		X			
41	diuron	X	X			X
34	DDT, 4,4-isomer				X	

[a]Bold-faced rows correspond to substances ranked by all methods, while the italicized row is associated with the substance ranked among the 20 most hazardous only by COMMPS. "X" means the susbtance is ranked by the method among the top 20 most hazardous substances.

dx.doi.org/10.1021/ci4004392 | J. Chem. Inf. Model. 2013, 53, 3190−3201

Table 6. Substances of the Monitoring-Based List for Organic Substances in the Aquatic Environment[2]

ID	CAS	compound	C_i	EFS_d	EFS_i	EFS_h
1	71−55−6	1,1,1-trichloroethane	0.141	1.35	0	1.8
2	87−61−6	1,2,3-trichlorobenzene	0.031	1.98	1	0
3	120−82−1	1,2,4-trichlorobenzene	0.157	2.43	2	1.8
4	95−50−1	1,2-dichlorobenzene	0.544	1.74	1	0
5	107−06−2	1,2-dichloroethane	8.243	1.93	0	2
6	108−70−3	1,3,5-trichlorobenzene	0.034	1.93	2	0
7	541−73−1	1,3-dichlorobenzene	7.100	1.8	1	0
8	106−46−7	1,4-dichlorobenzene	0.422	1.93	1	0
9	93−76−5	2,4,5-trichlorophenoxyacetic	0.323	2.14	0	0
10	94−75−7	2,4-dichlorophenoxyacetic acid	0.370	1.83	0	0
11	121−73−3	3-chloronitrobenzene	37.500	2.78	0	1.4
12	83−32−9	acenaphthene	0.417	2.63	1	0
13	15972−60−8	alachlor	0.150	3.57	0	1.8
14	309−00−2	aldrin	0.022	4.7	3	1.8
15	120−12−7	anthracene	0.083	4.09	2	0
16	1912−24−9	atrazine	0.334	2.94	0	1.8
17	6190−65−4	atrazine desethyl	0.043	3.07	0	0
18	2642−71−9	azinphos-ethyl	0.011	4.97	0	0
19	86−50−0	azinphos-methyl	0.011	4.03	0	0
20	25057−89−0	bentazon	0.034	1.5	0	0
21	56−55−3	benzo-a-anthracene	0.083	4.29	3	2
22	50−32−8	benzo-a-pyrene	0.027	4.5	3	2
23	205−99−2	benzo-b-fluoroanthene	0.048	4.29	3	2
24	191−24−2	benzo-g,h,i-perylene	0.047	5	3	1.8
25	207−08−9	benzo-k-fluoroanthene	0.025	4.9	3	2
26	470−90−6	chlorfenvinphos	0.103	4.29	2	0
27	2921−88−2	chlorpyrifos	0.128	5	2	0
28	15545−48−9	chlortoluron	0.117	3.14	0	0
29	21725−46−2	cyanazine	0.125	3.36	0	0
30	53−19−0	DDD, 2,4′-isomer	0.022	5	3	0
31	72−54−8	DDD, 4,4′-isomer	0.048	5	3	0
32	72−55−9	DDE, 4,4′-isomer	0.044	5	3	1.8
33	789−02−6	DDT, 2,4′-isomer	0.004	3.94	3	1.8
34	50−29−3	DDT, 4,4′-isomer	0.007	5	2	1.8
35	1007−28−9	desisopropylatrazine	0.145	2.45	0	0
36	333−41−5	diazinon	0.032	4.62	1	0
37	75−09−2	dichloromethane	10.250	2.11	0	1.8
38	62−73−7	dichlorvos	0.048	5	0	0
39	60−57−1	dieldrin	0.006	4.94	3	1.8
40	60−51−5	dimethoate	0.154	3.36	0	0
41	330−54−1	diuron	1.076	3.79	0	1.2
42	959−98−8	endosulfan, alpha-isomer	0.058	5	1	0
43	33213−65−9	endosulfan, beta-isomer	0.019	4.39	1	0
44	1031−07−8	endosulfan-sulfate	0.019	4.1	1	0
45	72−20−8	endrin	0.007	5	3	0
46	100−41−4	ethylbenzene	0.332	0.7	0	0
47	60−00−4	tetraacetic acid	45.590	1.71	0	0
48	122−14−5	fenitrothion	0.030	4.32	1	0
49	55−38−9	fenthion	0.015	4.46	1	0
50	206−44−0	fluoroanthene	0.082	2.43	3	0
51	319−84−6	HCH, alpha-isomer	0.025	3.57	1	1.8
52	319−85−7	HCH, beta-isomer	0.038	3.04	1	1.8
53	319−86−8	HCH, delta-isomer	0.022	2.64	2	1.8
54	58−89−9	HCH, gamma-isomer	0.037	3.24	2	0
55	76−44−8	heptachlor	0.013	5	3	1.8
56	118−74−1	hexachlorobenzene	0.010	4.29	3	2
57	87−68−3	hexachlorobutadiene	0.007	3.07	3	1.8
58	67−72−1	hexachloroethane	0.002	2.87	2	1.2
59	193−39−5	indeno(1,2,3-cd)pyrene	0.094	4.29	3	2
60	465−73−6	isodrin	0.012	4.44	3	0
61	34123−59−6	isoproturon	0.370	3.23	0	1.8

dx.doi.org/10.1021/ci4004392 | J. Chem. Inf. Model. 2013, 53, 3190−3201

Table 6. continued

ID	CAS	compound	C_i	EFS_d	EFS_i	EFS_h
62	121−75−5	malathion	0.040	4.07	1	0
63	94−74−6	MCPA	0.156	1.21	0	0
64	93−65−2	mecoprop	0.811	2.36	0	0
65	67129−08−2	metazachlor	0.080	3.14	0	0
66	72−43−5	methoxychlor	0.001	5	3	1.8
67	51218−45−2	metolachlor	0.313	3.36	1	0
68	91−20−3	naphthalene	1.683	2.59	2	0
69	56−38−2	parathion-ethyl	0.020	4.5	1	0
70	298−00−0	parathion-methyl	0.013	4.07	0	0
71	608−93−5	pentachlorobenzene	0.001	3.36	3	0
72	87−86−5	pentachlorophenol	0.135	3.34	3	1.8
73	7287−19−6	prometryn	0.033	3.07	0	2
74	139−40−2	propazine	0.052	1.97	0	1.8
75	7286−69−3	sebuthylazine	0.055	4.29	1	0
76	122−34−9	simazine	0.218	2.96	0	1.8
77	5915−41−3	terbuthylazine	0.170	3.07	0	0
78	886−50−0	terbutryne	0.279	2.14	1	0
79	127−18−4	tetrachloroethene	1.092	1.64	0	0
80	56−23−5	tetrachloromethane	1.049	2.25	0	1.8
81	108−88−3	toluene	10.796	1.5	0	1.8
82	79−01−6	trichloroethene	2.500	1.39	0	1.8
83	67−66−3	trichloromethane	1.173	2.93	0	1.8
84	1582−09−8	trifluralin	0.031	3.94	3	0
85	95−47−6	xylene, o-isomer	0.146	2.5	1	0

selected, then the whole linear extension is 4−3−2−1. The probability of this linear extension is (1/5) (see Table 1).

RANKING RESULTS

Comparison between *Maximals* and COMMPS. The list of the first 20 substances ranked according to their descending value of COMMPS monitoring-based list for organic substances in the aquatic environment was suggested for inclusion in the Water Framework Directive.[2] These substances are shown in Table 2 (see ref 4).

Therefore, more than a comparison of the whole ranking, the key point here is which substances are ranked among the top 20 most hazardous substances. Let us focus then our attention on the top 20 most hazardous substances provided by the new algorithm *Maximals*. From the partial order established by the four descriptors, *Maximals* computes the linear extension presented here as the solution. To provide the final solution, and in order to avoid bias, we run Algorithm 2 a total of 1 000 000 of times and consider the average of all of the linear extensions as a solution. Of course, the average rank could lead into a draw; however, averaging over 1 000 000 runs makes draws unlikely. The top 20 substances ranked according to *Maximals* are shown in Table 3, and the comparison between the two rankings is drawn in Figure 6 (these are more detailed in Table 7).

As it can be seen in Figure 6, 11 of the top 20 substances ranked by the COMMPS procedure are also ranked within the top 20 using *Maximals*. Therefore, the selection of these 11 substances is very qualified.

However, these two methods have nine disagreements. The substances ranked in the top 20 by COMMPS but not by *Maximals* are 1,2,4-trichlorobenzene (ID = 3), aldrin (ID = 14), benzo-b-fluoroanthene (ID = 23), chlorfenvinphos (ID = 26), DDD, 2,4′-isomer (ID = 30), DDT, 4,4′-isomer (ID = 34), dieldrin (ID = 39), heptachlor (ID = 55), and hexachlorobenzene (ID = 56).

On the other side, the substances ranked in the top 20 by *Maximals* but not by COMMPS are 1,2-dichloroethane (ID = 5), acenaphthene (ID = 12), alachlo (ID = 13), dichloromethane (ID = 37), endosulfan, alpha-isomer (ID = 42), isoproturon (ID = 61), metolachor (ID = 67), naphthalene (ID = 68), and trichloromethane (ID = 83).

To analyze these disagreements, let us study first the relative position of these substances in the Hasse diagram (Figure 2). As it can be seen, only two (ID = 3 and ID = 34) of the nine substances ranked in the top 20 by COMMPS but not by *Maximals* are at the most hazardous level of the Hasse diagram (Figure 2), while eight of the nine substances ranked by *Maximals* but not by COMMPS in the top 20 are at the most hazardous level in the Hasse diagram. Therefore, it can be concluded that our method is able to detect more hazardous substances than the COMMPS procedure. In addition, the most remarkable characteristic shared by the substances ranked by *Maximals* in the top 20 is that one of the descriptors PNEC, EFD_i, or EFS_d is 0. That means that this method is disjunctive; therefore, it considers risky a substance if at least one descriptor is hazardous.

Comparison among the Different Ranking Methods. Finally, let us make an overall comparison among the different ranking methods. Table 5 shows the degree of overlapping among the 20 most hazardous substances selected by COMMPS, *Maximals*, *MRP*, Desirability, Utility, and Dominance functions, and Figure 7 shows the overall rankings. As it can be seen, only three substances are ranked among the 20 most hazardous substances by all of the methods. These three substances are *indeno (1,2,3-cd) pyrene*, *benzo-g,h,i-perylene*, and *pentachlorophenol*. In addition, there is a substance (*1,2,4-trichlorobenzene*) ranked among the 20 most hazardous substances only by COMMPS.

On the other side, the *Maximals* and *MRP* methods are strongly related (their Spearman and Kendall correlation

dx.doi.org/10.1021/ci4004392 | J. Chem. Inf. Model. 2013, 53, 3190−3201

Table 7. Rankings Obtained with the Different Methods

	rankings					
substances ID	*Maximals*	COMMPS	*MRP*	desirability function	utility function	dominance function
24	85	83	85	72	83	70
59	84	85	83	83	77	83
27	83	74	84	81	57	66
31	82	75	77	70	66	36
25	81	79	74	71	84	56
22	80	78	71	85	79	63
61	79	59	81	49	51	76
68	78	63	82	26	31	74
41	77	67	78	31	43	78
72	76	80	73	79	69	85
83	75	64	79	47	49	79
67	74	42	76	22	25	67
32	73	82	80	69	82	60
21	72	84	75	61	75	80
12	71	31	69	21	22	48
11	70	72	64	84	85	82
37	69	61	67	48	54	81
13	68	55	66	50	53	68
42	67	56	72	63	42	43
5	66	60	63	73	56	84
26	65	68	70	43	50	32
3	64	69	60	53	63	65
23	63	81	68	56	74	72
7	62	29	61	17	23	59
81	61	45	62	44	48	77
55	60	76	65	64	80	47
47	59	9	56	59	72	69
14	58	77	57	60	73	52
30	57	70	58	68	65	19
80	56	46	59	41	41	75
76	55	47	53	45	45	64
16	54	50	54	46	46	71
15	53	62	52	38	47	25
40	52	25	55	19	14	42
8	51	19	50	11	17	62
36	50	43	51	42	39	40
77	49	17	46	13	10	50
75	48	44	45	30	32	33
64	47	10	43	8	8	57
38	46	39	48	76	27	29
56	45	73	42	75	76	49
45	44	57	47	66	64	17
85	43	27	39	18	19	61
78	42	23	37	14	18	51
39	41	71	44	62	78	30
50	40	49	41	57	44	53
4	39	16	38	9	13	55
34	38	66	49	58	70	34
29	37	21	40	20	15	44
82	36	32	35	37	37	73
69	35	37	36	36	38	24
48	34	40	34	32	33	26
84	33	65	32	80	58	35
52	32	48	33	51	60	21
54	31	41	31	29	40	20
79	30	4	25	4	5	58
62	29	38	27	27	29	38
60	28	58	24	77	61	15
10	27	5	17	5	4	45

dx.doi.org/10.1021/ci4004392 | J. Chem. Inf. Model. 2013, 53, 3190−3201

Table 7. continued

	rankings					
substances ID	*Maximals*	COMMPS	*MRP*	desirability function	utility function	dominance function
28	26	15	28	16	12	46
43	25	35	26	33	34	8
9	24	6	21	6	6	54
49	23	33	23	35	36	31
73	22	36	29	65	52	41
57	21	53	18	78	68	18
33	20	54	20	74	71	13
66	19	34	30	82	81	27
51	18	52	22	52	62	16
53	17	51	16	54	67	12
35	16	7	13	7	7	28
18	15	28	19	55	26	9
74	14	24	10	39	35	22
44	13	30	14	28	30	7
65	12	12	15	15	11	10
1	11	18	12	34	28	39
6	10	22	11	23	24	11
46	9	1	5	1	1	37
70	8	14	8	25	21	5
2	7	8	7	10	16	3
63	6	2	2	2	2	14
58	5	26	6	40	59	1
17	4	11	9	12	9	23
71	3	20	4	67	55	2
20	2	3	1	3	3	6
19	1	13	3	24	20	4

coefficients are respectively 0.99 and 0.91), having only two disagreements among the top 20 substances.

However, the substances ranked in the top 20 by *Maximals* but not by *MRP* (1,2-dichloroethane (ID = 5) and 3-chloronitrobenzene (ID=11)) are at the most hazardous level while the substances ranked at the top 20 by *MRP* but not by *Maximals* (benzo-b-fluoroanthene (ID=23), chlorfenvinphos (ID=26)) are not at the top level of hazardness.

In addition, the *Maximals* method is much more efficient than *MRP* (see[21]). To test it in practice, we have used Java implementations of both algorithms and have conducted the experiments in a 2.5 GHz PC with 4 GB of RAM. To carry on this study, each algorithm generates 100 000 linear extensions for sets X of n elements (n = 3,4,5,6,7,8). Table 4 shows the breakdown of the computation time for both methods when 100 000 linear extensions are generated. It can be seen that there is a severe rise in computation time when n grows. In addition, the running time of the *MRP* algorithm dramatically increases if many iterations are considered (see also ref 21).

It can be also proved that the *Maximals* method has a complexity of order $O(n^2)$ where n is the number of substances, while each iteration of the *MRP* method is already of order $O(n^2)$, and one such iteration is performed for each pair of incomparable elements selected by the algorithm

In fact, the computation time of *MRP* methods strongly increases when the cardinality of the substance set increases. This fact affects the overall performance of both methods since the two are heuristic. Therefore, the larger the number of repetitions, the better the approach is. In this case, we have checked that the rankings obtained after producing 1 million or 100 million linear extensions are exactly the same. That means that our approach is stable. Thus, *Maximals* seems to be a good choice for ranking large lists of substances.

A complete discussion related to COMMPS and Desirability/Utility/Dominance functions can be seen in ref 4. In this work, we just highlight the main results shown there. If we check how many substances ranked among the top 20 by the different methods are at the top level in the Hasse diagram, the results we get are nine by Utility functions, 10 by Desirability functions, 10 by COMMPS, 14 by *Maximals*, 15 by *MRP*, and 15 by Dominance functions. Note that in this specific case all the top priority substances are incomparable, but all the substances at the highest level of the Hasse diagram share a property; they present the largest values for some of the substance descriptors. This is not a necessary condition for maximal elements, but for the application we considered, it is an indicator of the procedure capability to discern most hazardous substances with respect to each single descriptor.

■ CONCLUSIONS AND FUTURE WORK

This work presents a new approach, called *Maximals*, to the problem of identification of priority hazardous substances. This methods overcomes the subjectiveness of judgment because it relies on partial order theory. It is based on generating a total order from a partial one by using a probability estimation. *Maximals* is first compared to the combined monitoring-based and modeling-based priority setting scheme (COMMPS) used to establish a first priority setting list within the EU Water Framework Directive. As was seen in the previous section, COMMPS and *Maximals* agree in 11 substances among the 20 most hazardous. With regard to the nine disagreements, note that our method is able to detect more hazardous substances than the COMMPS procedure. The method is also compared

dx.doi.org/10.1021/ci4004392 | J. Chem. Inf. Model. 2013, 53, 3190−3201

to other scoring methods previously used to rank hazardous substances. Focusing our attention on *Maximals* and *MRP*, which are strongly related (as Spearman and Kendall coefficients show), the ranks they provide are quite similar, having only two disagreements. However, the substances ranked in the top 20 by *Maximals* but not by *MRP* are at the most hazardous level, while the opposite is not true. In addition, one of the main advantages of our method is its complexity, allowing the generation of linear extensions in a more efficient way. With regard to the overall rank comparison, the analyzed ranking methods provided different ranking results. However, an agreement was achieved for the most risky substances, while several disagreements were identified for substances with a more controversial behavior. Therefore, as future work, we plan to select the most appropriate method for prioritization and to study the convenience in applying techniques to identify conflicting criteria. In addition, this method will be adapted to rank other chemical substances, as the identified in the Japanese Pollutant Release and Transfer Register project.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: sirene@uniovi.es.

*E-mail: troiano@unisannio.it.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Authors acknowledge financial support by Grant TEC2012-38142-C04-04 from Ministry of Education and Science, Government of Spain and by Grant UNOV-13-EMERG-GIJON-10 from University of Oviedo. The authors also acknowledge the support of all the reviewers for the revisions. We really appreciate their effort in improving the paper.

REFERENCES

(1) Water Framework Directive Brochure, European Union. http://ec.europa.eu/environment/water/pdf/WFD_brochure_en.pdf (accesed May 15, 2013).

(2) Revised combined monitoring-based and modelling-based priority setting scheme, European Union. http://ec.europa.eu/environment/water/water-dangersub/pdf/com_2011_875 (accesed May 15, 2013).

(3) Lerche, D.; Sørensen, P.; Larsen, H.; Carlsen, L.; Nielsen, O. Comparison of the Combined Monitoring-Based and Modelling-Based Priority Setting Scheme with Partial Order Theory and Random Linear Extensions for Ranking of Chemical Substances. *Chemosphere* **2002**, *49*, 637−649.

(4) Pavan, M.; Worth, A. A set of case studies to illustrate the applicability of DART (Decision Analysis by Ranking Techniques) in the ranking of chemicals. http://ihcp.jrc.ec.europa.eu/our_labs/predictive_toxicology/doc/EUR_234.

(5) Pavan, M.; Todeschini, R. Total-Order Ranking Methods. *Data Handl. Sci. Technol.* **2008**, *27*, 51−72.

(6) Annoni, P.; Bruggemann, R.; Saltelli, A. Random and quasi-random designs in variance-based sensitivity analysis for partially ordered sets. *Rel. Eng. Sys. Safety* **2012**, *107*, 184−189.

(7) Carlsen, L.; Bruggemann, R. Accumulating partial order ranking. *Environ. Model. Softw.* **2008**, *23*, 986−993.

(8) Voigt, K.; Bruggemann, R.; Pudenz, S. A multi-criteria evaluation of environmental databases using the Hasse Diagram Technique (ProRank) software. *Environ. Model. Softw.* **2006**, *21*, 1587−1597.

(9) Lerche, D. B.; Sørensen, P. B.; Bruggemann, R. Improved Estimation of the Ranking Probabilities in Partial Orders Using Random Linear Extensions by Approximation of the Mutual Ranking Probability. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1471−1480.

(10) Bruggemann, R.; Halfon, E.; Welzl, G.; Voigt, K.; Steinberg, C. E. W. Applying the Concept of Partially Ordered Sets on the Ranking of Near-Shore Sediments by a Battery of Tests. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 918−925.

(11) Trotter, W. T. *Combinatorics and Partially Ordered Sets*; The Johns Hopkins University Press: Baltimore, MD, 1992.

(12) Patil, G. P.; Taillie, C. Geoinformatic Surveillance Hotspot Prioritization Using Linear Extensions of Partially Ordered Sets for Multi-criterion Ranking with Multiple Indicators. *Proceedings of the 2004 Annual National Conference on Digital Government Research*; Digital Government Society of North America: St. Louis, MO, 2004; pp 49:1−49:2.

(13) Annoni, P.; Bruggemann, R.; Saltelli, A. Partial order investigation of multiple indicator systems using variance-based sensitivity analysis. *Environ. Model. Softw.* **2011**, *26*, 950−958.

(14) Halfon, E.; Reggiani, M. G. On ranking chemicals for environmental hazard. *Environ. Sci. Technol.* **1986**, *20*, 1173−1179.

(15) Bruggemann, R.; Voigt, K.; Münzer, B. The technique of hasse diagrams applied on environmental online databanks. *Chemosphere* **1994**, *29*, 683−691.

(16) Lerche, D.; Sörensen, P. Evaluation of the ranking probabilities for partial orders based on random linear extensions. *Chemosphere* **2003**, *53*, 981−992.

(17) Hasse, H. *Über die Klassenzhal aberscher Zahlkrper*; Springer-Verlag: Berlin, Germany, 1952.

(18) Brightwell, G.; Winkler, P. Counting linear extensions is #P-complete. *Proceedings of the twenty-third annual ACM symposium on Theory of computing*; ACM: New York, 1991; pp 175−181.

(19) Loof, K. D.; Meyer, H. D.; Baets, B. D. Exploiting the Lattice of Ideals Representation of a Poset. *Fundam. Inform.* **2006**, *71*, 309−321.

(20) Habib, M.; Medina, R.; Nourine, L.; Steiner, G. Efficient algorithms on distributive lattices. *Discrete Appl. Math.* **2001**, *110*, 169−187.

(21) Combarro, E.; Díaz, I.; Miranda, P. On random generation of fuzzy measures. *Fuzzy Sets Syst.* **2013**, 64−77.

(22) Karzanov, A.; Khachiyan, L. On the conductance of order Markov chains. *Order* **1991**, *8*, 7−15.

(23) Huber, M. Fast perfect sampling from linear extensions. *Discrete Math.* **2006**, *306*, 420−428.

(24) Bubley, R.; Dyer, M. Faster random generation of linear extensions. *Discrete Math.* **1999**, *201*, 81−88.

(25) De Loof, K.; De Baets, B.; De Meyer, H. Approximation of average ranks in posets. *MATCH* **2011**, *66*, 219−229.