JOURNAL OF CHEMICAL INFORMATION AND MODELING

Perspective

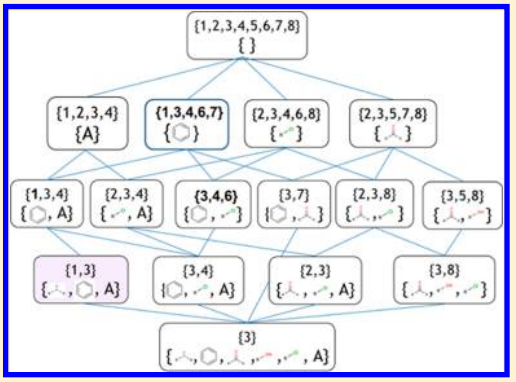
pubs.acs.org/jcim

Perspectives on Knowledge Discovery Algorithms Recently Introduced in Chemoinformatics: Rough Set Theory, Association Rule Mining, Emerging Patterns, and Formal Concept Analysis

Eleanor J. Gardiner and Valerie J. Gillet*

Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, United Kingdom

ABSTRACT: Knowledge Discovery in Databases (KDD) refers to the use of methodologies from machine learning, pattern recognition, statistics, and other fields to extract knowledge from large collections of data, where the knowledge is not explicitly available as part of the database structure. In this paper, we describe four modern data mining techniques, Rough Set Theory (RST), Association Rule Mining (ARM), Emerging Pattern Mining (EP), and Formal Concept Analysis (FCA), and we have attempted to give an exhaustive list of their chemoinformatics applications. One of the main strengths of these methods is their descriptive ability. When used to derive rules, for example, in structure−activity relationships, the rules have clear physical meaning. This review has shown that there are close relationships between the methods. Often apparent differences lie in the way in which the problem under investigation has been formulated which can lead to the natural adoption of one or other method. For example, the idea of a structural alert, as a structure which is present in toxic and absent in nontoxic compounds, leads to the natural formulation of an Emerging Pattern search. Despite the similarities between the methods, each has its strengths. RST is useful for dealing with uncertain and noisy data. Its main chemoinformatics applications so far have been in feature extraction and feature reduction, the latter often as input to another data mining method, such as an Support Vector Machine (SVM). ARM has mostly been used for frequent subgraph mining. EP and FCA have both been used to mine both structural and nonstructural patterns for classification of both active and inactive molecules. Since their introduction in the 1980s and 1990s, RST, ARM, EP, and FCA have found wide-ranging applications, with many thousands of citations in Web of Science, but their adoption by the chemoinformatics community has been relatively slow. Advances, both in computer power and in algorithm development, mean that there is the potential to apply these techniques to larger data sets and thus to different problems in the future.



1. INTRODUCTION

Knowledge Discovery in Databases (KDD) refers to the use of methodologies from machine learning, pattern recognition, statistics, and other fields to extract knowledge from large collections of data, where the knowledge is not explicitly available as part of the database structure. Information extracted "includes concepts, concept interrelations, classifications, decision rules, and other patterns of interest".[1] Some authors distinguish between KDD, which is the entire, multi-step process of "identifying valid, novel, potentially useful, and ultimately understandable patterns from large collections of data" and *data mining*, which is the key algorithmic step involved in KDD with other steps including data selection, data preparation, and result interpretation.[2] The motivation for the development of KDD algorithms is the very large amount of data now being stored by governments, corporations, and scientific organizations, and the overarching aim of KDD, and of the techniques discussed herein, is the finding of useful patterns hidden within these data. Very large amounts of data are also characteristic of the databases of pharmaceutical companies, which has led to the growing use of KDD methods within the drug discovery process. An overview of KDD in the context of chemoinformatics has been given by Wang et al.[3] Machine learning algorithms, such as nonlinear Support Vector Machines (SVMs) and Neural Networks (NNs), are supervised learning techniques whose primary aim is in classification. They may be regarded as "black box" classifiers; while providing good classification, there is no insight into why an object is classified in a particular class. Recently, there has been much interest in methodologies which offer explanations of molecular activity. Although such interpretable methods can be poorer in terms of predictive power, they can be of greater value to medicinal chemists since they can provide useful guidance on what compounds to make next. This has led to several different but related methodologies from the KDD field being introduced to chemoinformatics, including Rough Set Theory (RST), Association Rule Mining (ARM), Emerging Pattern Mining (EP), and Formal Concept Analysis (FCA). In contrast to some machine learning techniques, while these may be used as classifiers, their primary aim is explanation rather than classification. They highlight features, or sets of features,

Received: April 12, 2015
Published: August 3, 2015

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

which may lead to a particular classification for a set of objects. RST is concerned with describing a set or collection of objects which may be indistinguishable when limited information is available. This is achieved by means of the precise description of two other sets, one of which is a lower approximation and the other an upper approximation of the original set. ARM is concerned with the co-occurrence of sets of attributes, which can then be used to infer the presence of other attributes. EP is concerned with comparing two discrete data sets in order to discover distinguishing features—those whose increased presence in one data set and comparative absence in another can be used to differentiate between the two. FCA "aims at extracting a hierarchical structure of clusters from tabular data describing objects and their attributes".[4]

Historically, the first publications in these areas were in RST[5] and FCA,[6] both published in 1982. The seminal paper in ARM was in 1993[7], and EP was introduced in 1999.[8] As just described, these methodologies may superficially sound quite different. However, there are clear commonalities between them. They all operate on a data table whose rows represent objects and whose columns represent attributes. They are all computationally intensive methods of data mining since they all consider power sets (sets of sets). Each of the methods has been applied in at least one area of chemoinformatics, but they are not currently widely used. An author, introducing one of these methods, generally refers to it as being novel. Despite the seemingly obvious links between them, it was some time before connections between the methods were formally made in the literature. One of our main motivations in writing this perspective, therefore, is to highlight the similarities and contrasts between these methods. We review the basic ideas behind the KDD mining techniques of RST, ARM, EP and FCA. Since it has a strong relationship to EP, we also give an overview of Frequent Subgraph Mining. We compare the methods in order to draw out their similarities and differences. We then give a comprehensive review of the use of these methods in chemoinformatics.

2. MOLECULAR ANALYSIS

Before discussing these KDD techniques in detail, it is important to consider what type of problems they may be used to solve. As will become clear from the detailed discussion of applications in section 6, these methods may be applied in many areas. Many of the methods we consider are based on some kind of analysis of the properties of molecules. The properties are usually called *attributes* or *features* and are based on molecular descriptors. The most commonly used descriptors are physicochemical properties and structural fragments. Feature selection problems are then an obvious example of the type to be tackled. Subsets of features more commonly associated with active than inactive molecules can be found using any of these methods. The resulting feature sets can then be used in classifying molecules or in clustering them. They can be used to predict the class or properties of novel molecules. The features may be used as a reduced set, for input into another machine learning algorithm, or as the independent variables in a QSAR model.

An advantage of these methods is that property and structural (and other) descriptors can be considered in conjunction and the results can still be interpretable.

In the 1990s, particularly due to the comparative lack of computer power and the combinatorial nature of the analysis, it was common to consider just a few attributes. For example, in early work using RST, Krysinski used classes of R-group attachment points to a fixed scaffold as the attributes in a QSAR analysis, meaning that just eight attributes had to be considered.[9] Increases in computer power means that nowadays it is possible to consider many more attributes. For example, Goodarzi et al. used about 1450 descriptors in their study using fuzzy RS ant colony optimization.[10]

3. DATA SETS

The applications detailed in Section 6 are based on a number of different data sets. However, two in particular stand out as being far more frequently used, especially by nonchemoinformaticians. The Predictive Toxicology Evaluation (PTE) Challenge[11] was designed to ask the question "Can AI contribute to scientific discovery?" This challenge provided a readily available set of compounds which were classified as carcinogenic or not. Although the first challenge was set in 1997, initially as a blind test, the data are still frequently being used, and many of the algorithms described below have been evaluated on their performance on the PTE challenge data. The NCI tumor cell line data[12] provides a further source of accessible and well-annotated data. The ready availability of these data sets, and also the fact that the outcome of classification is simple (usually toxic or nontoxic), means that many nonchemistry specialists have been able to test their KDD algorithms on high quality data. However, the high quality of the test data may call into question the ability of the methods to deal with the noisy data often encountered in real-world applications.

4. KDD ALGORITHMS

4.1. Rough Set Theory. Rough Set Theory (RST) was developed by Pawlak in 1982[5] as a way of dealing with imprecise information and with "vagueness". In classical set theory, an object either belongs to a set or it does not. Such a set is called a *crisp* (or *precise*) set. For example, if objects are patients, and the set comprises patients who have a certain disease, then a patient either has or does not have the disease. Pawlak and Skowron illustrate the concept of vagueness with ideas from the ancient Greek philosophers concerning the "bald man paradox".[13] It is possible to identify a bald man and a man who is not bald. However, removing a single hair from the head cannot make a man bald. Therefore, removing hairs one-by-one, one would conclude that a man with no hairs was not bald. Vagueness is concerned with the boundary region—in this case the area where it is not possible to say with certainty whether a man is bald or not bald. Fuzzy set theory extends classical set theory by means of a membership function which quantifies the degree to which an object belongs to a set. RST, on the other hand, is not concerned with set membership but with vagueness—with objects whose membership of a set is unknown. A rough set (RS), X, is characterized by two crisp sets, namely, the *lower approximation* of X which consists of all objects which certainly do belong to X, and the *upper approximation*, which consists of all objects which might belong to X (so that the upper approximation is the complement of the set of objects which certainly do not belong to X). The difference between the upper and lower approximations of X is called the *boundary region*. For objects in the boundary region one cannot say whether or not they belong to the set. A nonempty boundary region implies that there is not sufficient knowledge about the set to give a precise definition. In fact

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

rough set membership is a generalization of fuzzy set membership since rough membership of set intersections (or set unions) cannot be deduced from membership of the constituent rough sets.[13,14] According to Pawlak and Skowron, the main advantage of RST over probabilistic methods is that it relies only on the data. It requires no prior assumptions such as a probability distribution in statistical analysis, a probability function in belief theory, or a set membership function in fuzzy set theory.

According to Duntsch and Gediga[15] RST addresses several main areas:

- Describing sets of objects by attribute values
- Analyzing attribute significance
- Finding dependencies between attributes
- Generating decision rules

Using RST, we know about *objects* (*discern* them) based ONLY on the information we have about them, i.e., their *attributes*, and we make no other assumptions. So, two objects are the same (in RST terminology *indiscernible*) if, and only if, they have the same values for each of their attributes.

We define the basic concepts of RST and illustrate them with reference to Table 1. The collection of objects under consideration (here molecules) is called the *universe* of objects, U, and the set of *attributes*, A, of the objects consists of the three attributes Rings, Halogens, and Rotatable Bonds. The *information system*, I = (U,A) is represented in the information table T = (U,A) of Table 1. Attributes are also known as *features*. In the real world, molecules 1 and 5 are different, but they have the same values for each of the attributes in Table 1 and so they are indiscernible in this information system. Any subset of attributes, B ⊂ A, can partition U into equivalence classes of objects, which are called *B-elementary sets* or *blocks*. Objects in the same block have the same value for each attribute in B (and so are equivalent). In Table 1, if B is the attribute set {Rings, Rotatable Bonds}, then B partitions the molecules into three sets, {1,5}, {2}, and {3,4}. Molecules 1 and 5 have one ring and two rotatable bonds, molecule 2 has no rings and eight rotatable bonds, and molecules 3 and 4 have one ring and one rotatable bond. For an object x, the block containing x is denoted B(x). Clearly two objects x, y are in the same block if, and only if, they have the same value for every attribute in B, i.e. B(x) = B(y). All objects in a block are indiscernible.

Table 1. Simple Information System in RST Consisting of Five Objects with Three Attributes[a]

Molecule	Rings	Contains Halogens	Rotatable Bonds
1	1	Y	2
2	0	N	8
3	1	Y	1
4	1	N	1
5	1	Y	2

[a]The Ring and Rotatable Bond attributes are counts, whereas the Halogens attribute is a binary yes/no.

For a set of objects, X ⊆ U, the *lower approximation of X in B*, $B_*(X)$, is the subset of X comprising the union of all the blocks which are contained entirely within X. The *upper approximation of X in B*, B*(X), is the union of all the blocks which contain any objects within X. If B = {Rings, Rotatable bonds} and X = {1,2}, then the only block entirely within X is {2} so $B_*(X)$ = {2}, and those blocks with at least one object in X are {2} and {1,5} so B*(X) = {1,2,5}. The difference set B*(X) − $B_*(X)$ is called the *boundary region of X in B* or the *B-boundary* of X. The B-boundary of {1,2} is {1,5}. If the boundary region is nonempty, the pair ($B_*(X)$, B*(X)) is a *rough set*. Put another way, a rough set is a means of representing a set X by means of two other object sets and an attribute set, B, (which may be all or a subset of the attributes). The first object set comprises all the objects which are definitely in X (the lower approximation) as far as the attribute set is concerned, and the second comprises all the objects which might be in X (the upper approximation). If the boundary is empty, then B*(X) = $B_*(X)$ and X is a crisp set. The *accuracy of approximation* of X using B is given by the ratio of the sizes of the lower and the upper approximation,

$$\alpha(X) = \frac{|(B_*(X))|}{|(B^*(X))|}$$

The accuracy of the approximation of X = {1,2} is therefore 1/3. Some of these concepts are illustrated in Figure 1.

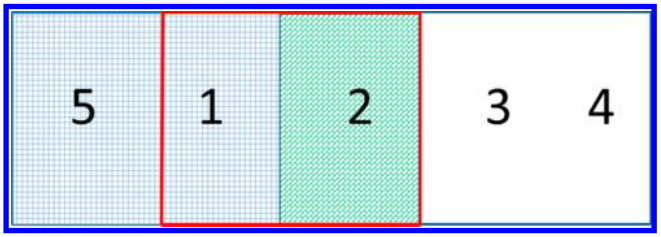


Figure 1. Rough set. The universe of five molecules is partitioned by B = {Rings, Rotatable Bonds} into three blocks, shown by the blue block containing molecules 1 and 5, the green block containing molecule 2, and the unshaded block of molecules 3 and 4. The set X is outlined in red. The green block is the lower approximation of X in B, and the blue and green blocks together constitute the upper approximation of X in B. The blue block is the boundary region.

If the attribute set A is partitioned, A = C ∪ D, where the attributes in C are *conditional* attributes and D contains a *decision* attribute which indicates the decision class of each object,[16] then the information table is called a *decision* table DT = (U,C,D). (There may be more than one decision attribute but multiple attributes are handled similarly—we consider just one for simplicity.) Each row of a decision table represents a *decision rule* or *implication*, which specifies the value of the decision attribute for the given values of the conditional attributes. Decision rules which have the same values for the conditional attributes, but different values for the decision attribute are called *inconsistent*.[13] In discussions of RST, information tables are usually considered separately from decision tables which are a specification of information tables. For brevity, and to illustrate the ideas of RST, we have here conflated the two table types. For a fuller discussion, the reader is referred to one of the many excellent tutorial-type articles, for example Walczak,[17] which is freely available as a download.

We will illustrate the ideas of RST, FCA, EP, and ARM using the same example where possible. Figure 2 shows a collection of molecules taken from Lhasa Ltd.'s Vitic database (www.lhasalimited.org). Table 2 shows details of some substructures contained within the molecules, together with a classification of the molecules as active or not active. Table 2 is a decision table where U is the set of molecules, U = {1,2,3,4,5,6,7,8} and C is the set of conditional attributes which comprise the

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

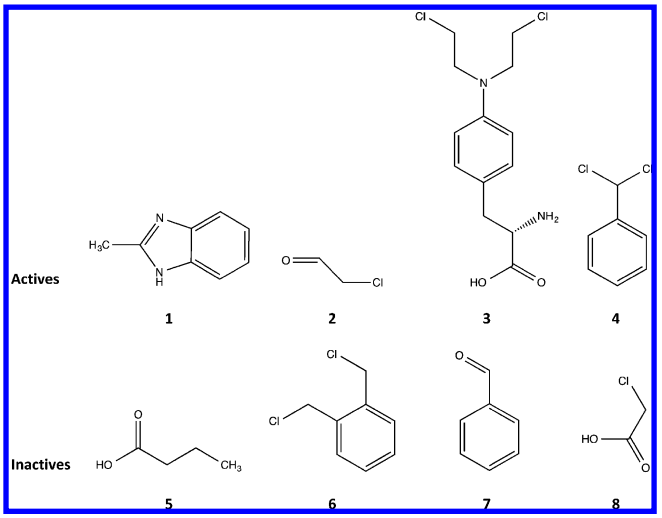


Figure 2. Active and inactive molecules.

Table 2. Decision Table for the Molecules Shown in Figure 2[a]

Molecules	Attributes a	b	c	d	e	Active
1	X	X				X
2			X		X	X
3	X	X	X	X	X	X
4		X			X	X
5			X	X		
6		X			X	
7		X	X			
8			X	X	X	

[a]An X in cell (*i*,*x*) means that molecule *i* contains fragment *x*.

substructures, C = {a,b,c,d,e} and D is the decision attribute of active or not, D = {Active}. (RST is able to deal with ordinal-valued attributes—this example uses binary attributes for easy comparison with FCA/EP methodologies which are not.) An X in the table means that the molecule possesses the substructure.

Molecules 4 and 6 have the same values for the conditional attributes, but molecule 4 is active and molecule 6 is inactive so the decision table is inconsistent. When dealing with decision classes, we can measure the accuracy of the classification. The *accuracy of the classification* is the ratio of the total number of objects belonging to the lower approximation of all classes to the total number of objects in the upper approximation of all classes. The lower approximation of the active class (using all attributes) is {1,2,3} since active molecule 4 cannot be classified based on its attributes, and the upper approximation is {1,2,3,4,6}. Similarly, the lower approximation of the inactive class is {5,7,8}, and the upper approximation is {4,5,6,7,8}. The accuracy of the entire classification is (3 + 3)/(5 + 5) = 0.6. The *quality* of a classification is the ratio between the total number of objects belonging to the lower approximation of all classes and the total number of objects so the quality of this classification is (3 + 3)/8 = 0.75. Molecules 4 and 6 belong to the boundary region. The presence of molecules in the boundary region can suggest the need for further investigation. Their presence could, for example, mean experimental error leading to misclassification or it could indicate the presence of an activity cliff.[18]

It is important to discover the degree of dependence between attributes[19] and the degree to which the decision class, D, depends upon the conditional attributes, C. D *depends on* C in a degree k ($0 \le k \le 1$), where k represents the fraction of objects which can be correctly classified in D using the conditional attributes in C. Formally,

$$k = \gamma_C(D) = \frac{|POS_C D|}{|U|}$$

where $|POS_C D|$ is the number of objects which can be correctly classified in the decision classes in D using the attributes in C and |U| is the number of objects in the universe.

Attributes are more significant if their removal leads to a change in the dependency of D upon C. The significance of an attribute a can be considered as the decrease in classification accuracy if a is removed from C. Thus

$$\sigma_c(D, a) = \frac{\gamma_C(D) - \gamma_{C\setminus\{a\}}(D)}{\gamma_C(D)}$$

A common aim when using RST is to find a subset of attributes which can represent all the data as well as the full attribute set—such an attribute set is called a *differentiating set*. A minimal such set is called a *reduct*; any proper subset of a reduct is not a reduct. (A reduct is minimal with respect to set inclusion, not necessarily cardinality.) A reduct is one of the key ideas of RST. An information system may have very many reducts, and the problem of finding all reducts is NP-hard.[20]

Table 3. Decision Rules Generated from Table 2[a]

	Attributes a	b	c	d	d	Active	Support (%)	Accuracy (%)	Coverage (%)
Rule 1		*	*	X	*		25	100	50
Rule 2	*	*	X	*			25	100	50
Rule 3	X	*	*	*	*	X	25	100	50
Rule 4	*		*		*	X	12.5	100	25
Rule 5	*	X	*		X	X	12.5	50	25

[a]An empty cell denotes the absence of the attribute, thus an empty cell in the Active column means the molecule is inactive. * means we do not care whether the attribute is present or absent.

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

Considering again Table 2, if we were to remove attribute *a*, attributes *b,c,d,e* would still discern all the objects. Thus, attribute *a* is superfluous. If, however, we were to remove attribute *d*, then molecule 2 would be indiscernible from molecule 3, and the number of elementary sets would be reduced. Thus, *d* is an *indispensable* attribute. The set of all indispensable attributes is called the *core*. (The core may be empty.) An indispensable attribute belongs to every reduct. Table 2 has for reducts, {a,b,c,d}, {a,b,d,e}, {a,c,d,e}, and {b,c,d,e} and the core is {d}.

Once a set of reducts has been found, a set of decision rules can be generated. A decision rule is an assertion of the form IF − THEN. The IF side is known as the *antecedent* (or premise) and the THEN side is the *consequent*. The antecedent consists of one or more conditional attributes, and the consequent is a single decision attribute. Two rules may have the same values for each of the conditional attributes but different values for the decision attribute; such rules are termed *inconsistent*. We used the ROSE2 software[21] to obtain the rules of Table 3. (This is a subset of the rules that could be generated). Rule 1 says that IF (substructure a is absent) AND (substructure d is present) THEN the molecule is inactive. *Support* of a rule is the percentage (or fraction) of objects matching both the antecedent and the consequent (i.e., the percentage of objects correctly classified using the rule). (Some authors define support as the number of objects correctly classified, rather than the percentage). *Accuracy* of a rule is the percentage of those objects matching the antecedent which also match the consequent and is thus a measure of how specific a rule is. The *coverage* is the percentage of objects matching the consequent (i.e., objects in the same decision class) which also match the antecedent and is thus a measure of how general a rule is.

Since Pawlak's seminal work in 1982, there has been a vast amount of research on RST and variants, with the original paper being cited more than 4000 times in Web of Science as at 2014. RST is used very extensively in fields as various as medical diagnosis,[22] financial analysis,[23] and bioinformatics.[24] A review of the use of RST for dimensionality reduction is given by Thangavel.[25] Feature (attribute) selection is a process of finding a subset of features, from the original set of features forming patterns in a given data set, optimal according to the given goal of processing and criterion.[19] RST, as a nonlinear method, can be used to reduce the dimensionality of a data set using only information contained within the data set.[26] When using RST for feature selection, the aim is to find a subset of attributes which can predict the decision variable as well as the full set of attributes. The minimum description length principle is usually applied, meaning that shorter reducts are preferred. The simplest approach to finding minimal reducts is to find all possible reducts and then choose one with minimal cardinality. However, selecting one of the reducts at random is well-known to be a poor method for feature selection.[19] Instead it is preferable to select relevant attributes which appear with "sufficiently high frequency" in reducts of subtables created by random sampling of the information table.[27] There are two different approaches in common use for RST feature selection, greedy methods and stochastic methods. Greedy methods use attribute significance as the main criterion. The most common stochastic method for feature selection is the Genetic Algorithm (GA).

As with all the classification methods described here, RST produces very many rules. The most commonly used algorithm for mining rules is the LEM2 algorithm.[28] This is a rule induction method which uses the learning-from-examples approach. Learning-from-examples approaches iteratively find rules which cover some positive examples and no negative examples, until all examples are covered. LEM2 handles uncertainty in the input data using the rough set approach. Thus, it induces rules for two sets, a lower approximation set for which the rules are certain and an upper approximation set for which the rules are possible. There is some RST software available, most free to academic groups, but most fairly old. Some details are given in Table 4.

Table 4. Rough Set Software

Software	Free?	Web site	ref
Rosetta	To academics	http://www.lcb.uu.se/tools/rosetta/resources.php	29
RSES	Yes	http://logic.mimuw.edu.pl/~rses/start.html	30
ROSE2	Yes	http://idss.cs.put.poznan.pl/site/rose.html	21
JMAF	Yes	http://www.cs.put.poznan.pl/jblaszczynski/Site/jRS.html	31

4.2. Association Rule Mining. Association rule mining (ARM) is the process of discovering associations between items in databases.[7,32] It was developed by Agrawal in the early 1990s, particularly in the field of sales transaction (market basket) analysis. The classical example is the association between the purchase of beer and nappies. Consider a collection I of items. A transaction is a set of items; each transaction has a unique identifier, called a TID. The transactions can be listed in an *information* table, D, each row corresponding to a transaction. A set of items is called an *itemset*. An itemset containing k items is called a k-itemset. If T is a transaction, X is an itemset and $X \subseteq T$ then T *contains* X. The *support count* of an itemset is the number of transactions in which it occurs; the *support* of an itemset is the fraction of transactions which contain it. A *frequent itemset* is one whose support is at least some threshold, always denoted *minsup*. A maximal frequent itemset is a frequent itemset which is not contained in a larger itemset. (Confusingly, maximal frequent itemsets are also known as *large itemsets*). A *closed itemset* is an itemset for which no immediate superset (a superset containing one additional item) has the same support. An *association rule* is an implication between disjoint itemsets: $X \Rightarrow Y$. The left-hand side of the rule is the *antecedent* and the right-hand side the *consequent*. Some authors restrict the consequent to itemsets with only one item[7] while others allow larger itemsets.[32] The rule $X \Rightarrow Y$ with *support* s means that the fraction of transactions in D which contain both X and Y is *s*.[33] The rule $X \Rightarrow Y$ with *confidence* c means that the fraction of transactions in D containing X which also contain Y is *c*. Confidence can also be referred to as the *strength* of the rule. If a minimum confidence threshold is applied, this is always denoted *minconf*. (The term confidence in ARM is equivalent to *accuracy* when mining rules using RST). There are several classes of association rules. Examples include *Boolean* association rules (as is the case here), where the presence or absence of an item in a transaction is considered, and *quantitative* association rules, where counts of each item in a transaction are maintained. In Table 5, the itemset {a,b} occurs in two transactions, so it has support count 2 and support 0.25. The rule {b} ⇒ {Active} occurs in three transactions; {b} is contained in five transactions of which three also contain {Active}, so this rule has confidence 0.6 and

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

Table 5. Information Table for the Molecules of Table 2

TID	Itemsets
1	{a,b,A}
2	{c,e,A}
3	{a,b,c,d,e,A}
4	{b,e,A}
5	{c,d}
6	{b,e}
7	{b,c}
8	{c,d,e}

support 0.325. Notice that {Active} ⇒ {b} is also a rule, again with support 0.325 but this time with confidence 0.75.

ARM aims to find all association rules with support ≥ minsup and confidence ≥ minconf. Using association rules to construct classification systems is known as *Associative Classification Mining* (ACM).[34] The rules upon which the classification is based are called *Classification Association Rules* (CARs). In this case, the rule consequent is an itemset with only one item, the class to which the transaction belongs (or will be assigned).

ARM is computationally demanding, although there does not seem to have been a great deal of research into algorithm complexity in this area.[35] However, it has been shown that the problem of ARM mining for Boolean or quantitative association rules is NP-complete.[35]

Popular ARM algorithms include Apriori[32] and its many variations, Max-Miner,[36] Frequent-Pattern-Tree-based mining,[37] ECLAT,[38] Close,[39] and OPUS_AR.[40] The most well-known is the Apriori algorithm[7,32] which has been hugely influential in the area of KDD, with more than 3300 citations in Web of Knowledge. However, ARM has hardly been used in chemoinformatics: Apriori[32] has just one citation in the Journal of Chemical Information and Modeling[41] and one in the Journal of Cheminformatics.[42] ARM mining usually consists of a two-step approach: (1) generate all frequent itemsets and (2) generate high confidence rules from the frequent itemsets.

Generating *all* itemsets and then checking if an itemset is frequent is clearly impossible. If there are M items, then there are 2^M possible itemsets. Apriori and Apriori-like algorithms use a breadth-first search strategy (BFS), whereby candidate frequent itemsets of size k are generated from itemsets of size k-1. The Apriori approach is designed to prune the candidate frequent itemsets. Its basis is the observation that if an itemset is frequent then all of its subsets must also be frequent. This is due to the *anti-monotone property of support* (also called the *downward closure property of support*,[33] DCP), i.e., an itemset cannot have a larger support (be present in more transactions) than any of its subsets. Thus, once an infrequent itemset has been found, all its supersets can be pruned from the search tree. However, ascertaining the support of a large number of itemsets is a time and/or memory intensive operation. The Apriori algorithm makes multiple passes over the database while counting supports.

A problem with ARM is the very large number of rules which are generated, very many of which are redundant (i.e., give the same information). Bayardo and Agrawal classify ARM algorithms as *constraint-based*, for example, finding all rules with a minimum support or confidence level, which requires users to decide the level which is appropriate; *heuristic*, returning only a subset of the possible rules; or *optimal*, finding only the most interesting rules, according to some measure of interestingness.[43] Bastide at al. assert that the most useful rules are those with minimal antecedent and maximal consequent (i.e., having fewest items in the IF-part and most items in the THEN-part of the rule), and they develop methods for extracting these based on *frequent closed itemsets* (i.e., closed itemsets with support at least minsup). Another problem is that any method, such as Apriori, which prunes based on support, cannot discover infrequent but important rules. Table 6 shows

Table 6. CARs Mined from the Molecules in Table 2

	itemset	Active	Support (%)	Confidence (%)
Rule 1	a	X	25	100.0
Rule 2	b	X	62.5	60.0
Rule 3	e	X	62.5	60.0
Rule 4	ab	X	25	100.0
Rule 5	de	X	25	50.0
Rule 6	bc	X	25	50.0
Rule 7	be	X	37.5	66.7
Rule 8	ce	X	37.5	66.7
Rule 9	dce	X	25	50.0
Rule 10	d		37.5	66.7
Rule 11	c		62.5	60.0
Rule 12	dc		37.5	66.7
Rule 13	de		25	50.0
Rule 14	bc		25	50.0
Rule 15	dce		25	50.0

a set of CARs mined from the molecules in Figure 2, with minsup = 20% and minconf = 50%. The rules were generated using the Apriori algorithm implemented by Borgelt.[44] Notice that Rule 1 of Table 6 is the same as Rule 3 of Table 3. Rule 4 is redundant since it is implied by Rule 1 and does not have higher support.

Not all ARM algorithms are Apriori-like. OPUS[45] is a branch and bound search algorithm applicable to any search space in which the order of operation is insignificant. OPUS_AR has been adapted for ARM mining.[40] It mines a user-defined number of rules which maximize an arbitrary function measuring rule quality. It is designed to be used only when all data can be held in main memory (bypassing the need to reduce the number of passes over the data, which typically constrains Apriori-like algorithms). The OPUS strategy is to reorder the search space so that a condition which is to be pruned at a node precedes all other conditions. This condition drastically reduces the search space (by approximately half for each pruning operation).

ARM is used in a huge variety of applications, including market basket analysis, recommender systems for e-commerce, and analysis of gene expression data. A list of ARM mining software is available at http://www.kdnuggets.com/software/associations.html. Some are listed in Table 7.

4.3. Emerging Patterns. Emerging patterns (EPs) are patterns in data which appear more strongly in one set than another and thus aid in classification. Since the data has to be assigned to one of two sets in order for EPs to be found, EP is designed to be a classification technique. EPs were introduced by Dong and Li in 1999.[8] Much of the terminology of the EP literature is similar to that for ARM. Consider two information tables, D_1 and D_0. Each has rows which are again called transactions. The ordered pair (attribute, value) is called an item; the columns of the information table represent items. A set of items is called an itemset or a pattern. In Table 8, the

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

Table 7. ARM Software

Software	Free?	Web site	ref
FrIDA data analysis toolbox, including Apriori, Eclat, MoSS, etc.	Yes	Christian Borgelt's Web Pages: http://www.borgelt.net	44
LUCS-KDD software library, including Apriori, FP-Growth, CMAR, CBA, etc.	Yes	http://cgi.csc.liv.ac.uk/~frans/KDD/Software/	46
Magnum Opus	Demonstration version free, reduced price for academics	http://www.giwebb.com/	40,45

Table 8. Information Table for EP[a]

Transaction	Items			
1	a		c	d
2		b		d
3	a	b	c	d
4	a		c	

[a]An empty cell means the item is absent.

items are represented by the codes a,..,e; in this case, the value of each attribute is either "present" or "absent" so that "a" means (a, present). The absence of an item from a transaction is indicated by a blank cell.

Transaction 1 contains the itemsets {a},{c},{d},{ac}{ad}, {cd}, and {acd} since these are the subsets of {acd}.

The *support* of an itemset is the proportion of transactions which contain it. So supp{cd} = 2/4 = 0.5. Then, given two information tables, D_1 and D_0, an EP in D_1 is an itemset with greater support in D_1 than in D_0. The *growth rate* of an EP is given by

$$\text{growth(pat)} = \frac{\text{supp(pat)}_{D_1}}{\text{supp(pat)}_{D_0}}$$

The growth rate of an EP is between 1 and infinity. An EP which has zero support (i.e., is not present) in D_0 is called a Jumping Emerging Pattern (JEP) in D_1. A JEP has a growth rate of infinity. JEPs are extremely discriminating but may have very low support and are completely intolerant to noise since a single false negative means that an EP will no longer be jumping. EPs are more noise-tolerant and in general are likely to have higher support than JEPs. However, there are many more of them which may be a disadvantage. Growth rate can be used to restrict the number of EPs found, for example, by setting a maximum value for the support in D_0 or excluding EPs with high support but weak growth rate. A minimal JEP is a JEP for which no proper subset is also a JEP. Minimal JEPs, with sufficient support, are the *most expressive*[47] and are best used for classification.

We can regard Table 2 as being two disjoint Decision Tables, D_1 comprising the active molecules in rows 1−4 and D_0 comprising the inactive molecules in rows 5−8. The simplest method to identify sets of EPs and JEPs requires the complete enumeration of all itemsets in one data set followed by a comparison with the entries in the other in order to determine the support for each pattern in both data sets. Table 9 shows all the itemsets which are present in the active molecules and their support in D_1 and D_0. The EPs, i.e., those patterns of substructures which are present in more active than inactive molecules, are highlighted in dark green, and those EPs which are also JEPs are highlighted in light green. The three patterns highlighted in cyan are the minimal JEPs. For chemical activity data, minimal JEPs are the smallest sets of structural features that exclusively distinguish the active from inactive compounds.

Table 9. EPs and JEPs Mined from the Decision Table of Table 2[a]

Itemsets					Support	
a	b	c	d	e	actives	inactives
X					0.5	0
	X				0.75	0.5
		X			0.5	0.75
			X		0.25	0.5
				X	0.75	0.5
X	X				0.5	0
X		X			0.25	0
X			X		0.25	0
X				X	0.25	0
	X	X			0.25	0.25
	X		X		0.25	0
	X			X	0.5	0.25
		X	X		0.25	0.5
		X		X	0.5	0.25
			X	X	0.25	0.25
X	X	X			0.25	0
X	X		X		0.25	0
X	X			X	0.25	0
X		X	X		0.25	0
X		X		X	0.25	0
X			X	X	0.25	0
	X	X	X		0.25	0
	X	X		X	0.25	0
		X	X	X	0.25	0.25
X	X	X	X		0.25	0
X		X	X	X	0.25	0
X	X		X	X	0.25	0
X	X	X		X	0.25	0
	X	X	X	X	0.25	0
X	X	X	X	X	0.25	0

[a]The rows highlighted in cyan are three minimal JEPs.

The first minimal JEP in Table 9 is equivalent to Rule 3 generated using RST (Table 4) and Rule 1 generated using ARM (Table 6).

The idea of JEPs was extended by Fan and Ramamohanarao to include Strong JEPs (denoted SJEPs)[48] which are JEPs with at least a (user-defined) minimal level of support in D_1 and for which no proper subset is both jumping and has the required support. A recent review of EP-mining algorithms is given by Garcia-Borroto et al.[49]

Both EP and JEP mining belong to the class of constraint-based pattern mining problems, where the requirement to be emerging or jumping and emerging is a constraint. Finding EPs and even JEPs can be very demanding in terms of storage requirements and of time. However, a large collection of patterns can be represented by its *border*.[50] A collection $\mathcal{L}$ of sets is called an *antichain* if, for all sets $X,Y \in \mathcal{L}$, $X \not\subseteq Y$ and $X \not\supseteq Y$, i.e., no set in $\mathcal{L}$ is a subset of another set in $\mathcal{L}$. A border is then an ordered pair of antichains, $<\mathcal{L}, \mathcal{R}>$such that each set $X \in \mathcal{L}$ is a subset of some $Y \in \mathcal{R}$ and each $Y \in \mathcal{R}$ is a

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

Table 10. EP/CSM/SD Software

	Free?	Web site	ref
Magnum Opus	Demonstration version free, reduced price for academics	http://www.giwebb.com/	45
Orange	Free, opensource	http://www.ailab.si/orange then download http://kt.ijs.si/petra_kralj/SubgroupDiscovery/	67
VIKAMINE	Free, opensource	http://vikamine.sourceforge.net/	68
KEEL (Knowledge Extraction based on Evolutionary Learning)	Free, opensource	http://www.keel.es/	69

superset of some $X \in \mathcal{L}.\mathcal{L}$ is called the *left bound* of the border and $\mathcal{R}$ the *right bound*.[50] This means that each set in the left border is a subset of a set in the right border and that each set in the right border contains a set from the left border. The border description of a collection of (convex) sets is unique, and all the sets of the collection can be enumerated from its border description.

The collection of JEPs can then be obtained by means of the Border-Differential algorithm;[8,50] this algorithm mines the differences between the border descriptions of the two decision classes, which is precisely what is required to discover itemsets present in one collection but not the other. EP is a difficult problem, being exponential in the number of database objects in the worst case;[51] algorithms can typically run out of memory or time. In fact, it has been shown that EP is an NP-hard problem.[52] One of the difficulties for EP lies in having to maintain occurrence counts in order to ascertain the support of the pattern. Another obstacle is that the antimonotone property used for pruning search trees cannot, in general, be applied for EP.[53] The Contrast Pattern Tree mining algorithm[48] is a depth-first search algorithm which mines EPs; it deals with the storage problem by creating, traversing, and pruning branches already visited in one continuous process. Other algorithms include MUSIC-DFS[54] which uses a notion called "prefix-freeness" to impose an antimonotone ordering. This enables a depth-first search algorithm to efficiently prune the search tree, while still allowing all patterns to be found. The authors also show how to efficiently include external constraints in order to limit the numbers of EPs while generating the most meaningful ones.

Other data mining paradigms closely related to EP are Version Spaces,[55] Discriminant Rules,[56]Contrast Set Mining,[57] and Subgroup Discovery.[58] Contrast Set Mining, EP and Subgroup Discovery can be grouped under the heading "Supervised Descriptive Rule Discovery".[59] Contrast Set Mining (CSM) is defined as finding "conjunctions of attributes and values that differ meaningfully in their distributions across groups",[57] while Subgroup Discovery (SD) is defined as finding population subgroups that are statistically "most interesting" with respect to a property of interest.[58] In fact, Novak et al. showed that EP, CSM, and SD have the same goals and differ primarily in their terminology[59] and in the approaches taken by the investigating community. The heuristics used by the communities in discovering contrast sets, emerging patterns, and subgroup discoveries are also closely related. However, Novak et al. also note that the application areas are somewhat different. CSM has few published applications—examples are mainly medical.[60] EP has been used in bioinformatics,[61] image processing,[62] and chemoinformatics.[63] SD has been widely used since its inception in areas as diverse as gene expression data analysis[64] and political stability analysis.[65] A comprehensive review of SD algorithms and applications is given in ref 66.

Given the close relationship between EP, CSM, and SD, it is not surprising that the same software can be used in all three areas. Some examples are given in Table 10.

4.4. Formal Concept Analysis (Galois Lattice). Formal concept analysis (FCA),[6,70,71] also known as Galois lattice mining, is a framework for conceptual clustering. Two large complementary reviews of the FCA literature (one on applications, the other methodological) have been published very recently by Poelmans et al.[72,73] Interestingly they used tools of FCA to analyze and present an overview of the 1072 papers they retrieved as a result of their literature search into models and techniques.[72] An accessible introduction to FCA is given by Priss.[74] In FCA, a *formal context* is a triple, K = (U,A,R) consisting of a set of *objects*, U, a set of properties or *attributes* (often called *items* in related fields), A, and a binary relation $R \subseteq U \times A$. If (u,a) ∈ R then we say *the object u has the attribute a*; this is often written uRa. (Much of the initial work on FCA originates in Germany. The German for objects is Gegenstände and for attributes is Merkmale. For this reason a context is often denoted as a triple (G,M,I) in the literature.) In any context, a (formal) *concept* is a pair of sets of attributes and objects that uniquely define each other: the attributes are all the attributes shared by the objects and vice versa. Formally this relationship is given by the *derivation (or modal) operators*, denoted by ∗ (or often by ' or prime). $X \mapsto X^* = \{a \in A : uRa\ \forall u \in X\}$, i.e., X* is the set of attributes shared by all objects in X, and $Y \mapsto Y^* = \{u \in U : uRa\ \forall a \in Y\}$, i.e., Y* is the set of objects which possess all the attributes in Y. Then, for $V \subseteq U$, $B \subseteq A$, the pair (V,B) is a *formal concept* (or just concept) of K if V* = B and B* = V. The attribute set B is the *intent* of the concept and the object set V is the *extent*. The ∗ operators form an *antitone Galois connection* between the power sets of U and A.[70] The ∗ operators have a number of important properties, including $X \subseteq X^{**}$ for all $X \subseteq U$ and $Y \subseteq Y^{**}$ for all $Y \subseteq A$.

In our molecular example in Table 2, the object set {1,3} and the attribute set {Active, a, b } constitute a concept. {1,3} is the extent of the concept and {Active, a, b} is its intent. A formal concept is also known as a *maximal rectangle* since it represents a rectangle (not necessarily contiguous) within a context which is composed entirely of X's.[75]

A formal context is visualized by means of a Hasse Diagram. In a Hasse diagram nodes represent concepts; the nodes are ordered upward according to their extents. The reading rule for a Hasse diagram is: An object "'g'" is described by an attribute "'m'" if and only if there is an ascending path from the node named by "'g'" to the node named by "'m'".[72] The extent of the top concept is always the entire set of objects. Although the lines of a Hasse diagram may cross, a line may not intersect any node except its two end points. The Hasse diagram illustrates the fact that the lattice may be viewed as a hierarchical clustering of concepts,[70] via their objects (reading upward) or their attributes (reading downward). In the example, the

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

concept comprising object set {1,3} and the attribute set {a, b, Active } is highlighted in the Hasse (lattice) diagram of Figure 3.

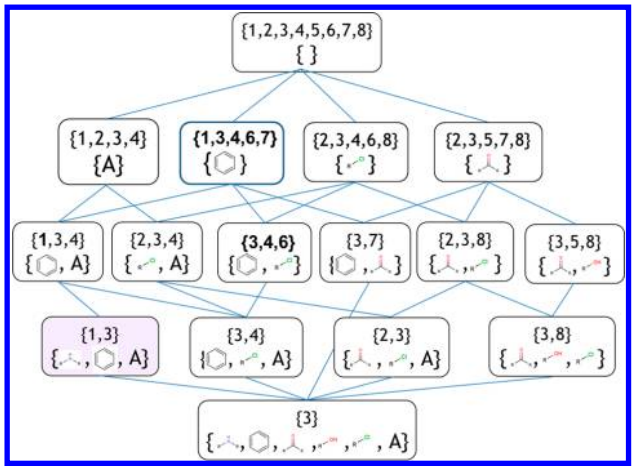


Figure 3. Hasse diagram obtained from Table 2. The formal concept ({1,3},{a,b,Active}) is highlighted in purple. The lattice was obtained using the Galicia software.

The concept (V_1,B_1) is a *subconcept* of (V_2,B_2) if $V_1 \subseteq V_2$ or (equivalently) $B_2 \subseteq B_1$. (V_2,B_2) is then a *superconcept* of (V_1,B_1). The set of concepts extracted from K = (U,A,R) is denoted by $\mathcal{B}$(U, A, R) or $\mathcal{B}$(K). $\mathcal{B}$(U, A, R) can then be partially ordered according to its extents, written $(V_1,B_1) \leq (V_2,B_2)$. The resulting lattice is a *Galois Lattice*. While the lattice formulation enables visualization of a formal context, its more important effect is to make the context amenable to analysis.

The main output from FCA is a set of implications. An *implication* $B \rightarrow C$ is a pair of sets of attributes, $B,C \subseteq A$ such that the object set $B^* \subseteq C^*$. An implication says that every object possessing all the attributes in B also has all the attributes in C. An *implication basis* is a minimal set of implications from which all other implications can be generated using a set of rules known as Armstrong Rules[72] (or using other rules). Implications can be read directly from the Hasse diagram. Table 11 shows an implication basis obtained from the Hasse diagram of Figure 3. From the definition of an implication, it is clear that if the consequent contains a decision variable then an implication is a decision rule. The second implication of Table 11 is equivalent to the first minimal JEP of Table 9.

Table 11. Implication Basis Obtained from the Hasse Diagram of Figure 3

Premise	Consequent	Support	Confidence
{d}	{c}	0.37	1
{a}	{Active,b}	0.25	1
{Active,c}	(e}	0.25	1
{d,e}	(c}	0.25	1
{Active,d}	{a,b,c,e}	0.12	1
{b,d}	{Active,a,c,e}	0.12	1
{a,c}	{Active,b,d,e}	0.12	1
{a,d}	{Active,b,c,e}	0.12	1
{a,e}	{Active,b,c,d}	0.12	1
{b,c,e}	{Active,a,d}	0.12	1
{Active,b,c}	{a,d,e}	0.12	1

FCA is restricted to a binary relation between the objects and the attributes, i.e., object u either has or has not attribute a. In order to deal with categorical or ordinal values, attributes have to be scaled, often using so-called *conceptual scaling*.[76] For example, if an attribute is color, with values red, blue, green, then it is necessary to introduce three new attributes, red, blue, green, for which a red object has values 1,0,0. Since FCA mines the power sets of the attributes, each new subcategory multiplies the number of possible sets by 2. For ordinal attributes, conceptual scaling can use ordering. So if an attribute can take values 0−10 the values available can be represented by 0, ≤ 1, ≤ 2, .., ≤ 10, and 3 could be represented by 0 0 0 1 1 1 1 1 1 1 1 1 1. Conceptual scaling is a powerful tool within FCA which, in conjunction with Hasse diagrams, allows one to view a context at different levels of detail.

Other scaling methods can also be used. In particular, *pattern structures* are a generalization of FCA to graphs.[72,77,78] An advantage of pattern structures is that they can be better adapted to multi-valued attributes[79] and have been shown to give better results, in terms of scalability, conciseness, and readability, than using conceptual scaling when analyzing gene expression data.[80]

Ganter and Kuznetsov showed that FCA is closely related to a standard data mining approach called *Version Spaces*.[78] They showed how to reformulate a situation expressed in terms of FCA into one expressed in terms of Version Spaces and gave an example from predictive toxicology. As Version Spaces are a specification of Inductive Logic Programming (ILP), it is clear that FCA is also closely related to ILP.

There are many available implementations of FCA software, useful data sets, and web pages. Some are given in Table 12. The FCA home page, maintained by Uta Priss, is an excellent resource.

Table 12. FCA Software

	Free?	Web site	ref
FCA home page		http://www.upriss.org.uk/fca/fca.html	
TOSCANAJ	Yes	http://toscanaj.sourceforge.net/	81
Galicia	Yes	http://www.iro.umontreal.ca/~galicia/	82
Coron	Yes	http://coron.loria.fr/site/components_algo.php	

5. COMPARISON OF KDD METHODS

There are clear commonalities between these data mining methods, and often similar terminology is used. The most obvious differences (and traps for the unwary!) lie in slightly different definitions of similar terms. In general, when decision rules are derived, the terminology is consistent, as one might hope. Thus, for example, the support of a decision rule is the number of objects correctly classified using the rule, whatever the method used to determine the rule. Table 13 compares terms used in the different methods. The terms listed across a row of the table are NOT necessarily equivalent. For example in RST, we are interested in sets of attributes called reducts, while in FCA our interest is in sets of attributes called intents.

Some pairwise comparisons of RST, ARM, EP, and FCA have been reported in the literature, mostly of an algorithmic nature,. Below, we summarize these, as they do point out the similarities, and some differences, between the methods, albeit on a pairwise basis.

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

Table 13. Comparison of Terms

Term	RST	ARM	EP	FCA
Basic entity	Information system	Information table	Pair of information tables	Formal context
Row of the table corresponds to	Object	Transaction	Transaction	Object
Column	Attribute	Item	item = (attribute,value)	Attribute
Set of attributes	Reduct, differentiating set	Itemset	Itemset, pattern	Intent
Set of objects				Extent
Minimum support threshold		minsup		minsupp
Minimum confidence threshold		minconf		
Support count		Of itemset, number of transactions containing itemset	Of itemset, number of transactions containing itemset	
Support		Of itemset, fraction of transactions containing itemset	Of itemset, fraction of transactions containing itemset	
Frequent itemset, frequent pattern		Itemset with support ≥ minsup	Itemset with support ≥ minsup	
Rule	Decision rule	Association rule	Decision rule	Implication
Of rule	Accuracy	Confidence		Confidence

5.1. Relationships between RST, ARM, EP, and FCA. RST, ARM, and FCA can all be used in both a supervised and an unsupervised manner, depending on whether or not classification is the aim. Although the discussion here has largely focused on decision tables, RST, ARM, and FCA can all be used in other knowledge discovery tasks and do not necessarily require supervision. EP in contrast relies on the separation of the data into two classes. All the methods are computationally demanding if exact and complete solutions are required, but usually they are not. In all but the simplest systems, the generation of all possible association rules, rough set reducts, emerging patterns, or formal concepts is not only infeasible but undesirable. Usually sets of closely related rules (reducts, patterns, concepts) can be found, and it is necessary to consider only a reduced subset of them, the remainder being redundant. Most algorithms are heuristic, finding a subset of, for example, decision rules, based on support-level or some notion of "interestingness" which may be application-specific. The degree of relatedness between the methods is often unclear. The relationships between them can be obscured by the specific terminologies of the methods and by their diverse areas of application. But, with appropriate restrictions, the methods are closely related, as the following discussions will show.

5.2. RST and ARM. Consider the use of RST and of ARM, when used to generate decision rules with a single decision variable as the consequent. As Delic et al. argue, "despite their different approaches, both methods are based on the same principle, and consequently must generate the same rules".[83] When the definition of item for ARM is altered, so that an item is an ordered (attribute, attribute-value) pair (as in EP-mining), then ARM and RST can both be applied to the same data table. In order to compare RST with ARM, Delic et al. performed attribute reduction (i.e., generating reducts) upon the rules generated using all the attributes. Then, the rules generated by both methods were identical showing that RST and ARM are equivalent. They then gave a theoretical proof of this result. Thus, the difference between ARM and RST, when used for the generation of decision rules, lies in the methods used for reducing/selecting rules. ARM relies mainly on pruning using minsup, minconf, and the Apriori principle, whereas RST uses equivalence relations on attributes to select reducts and then uses only rules based on the reducts.

This equivalence between RST and ARM is rarely, if ever, noted. In fact, the work of Delic has never previously been cited. Delic further showed that, in a performance comparison using their implementation of the Apriori algorithm and their own RS-rules+algorithm, the Apriori algorithm was an order of magnitude faster on all the data sets considered. However, the authors noted that it is quite possible that more efficient algorithms and better implementations would give different results. They then incorporated RS attribute reduction into the "Apriori+" algorithm which removed redundant rules without degrading the Apriori performance. In fact, finding a reduct set is often used as a preprocessing technique before association rules are mined.[25]

5.3. EP and ARM. A similar result applies to the relationship between EP and ARM. Webb et al. compared the performance of their ARM mining program,[40] OPUS-AR, with the contrast-set miner, STUCCO, of Bay and Pazzani,[57] in a contrast set mining task. The original purpose of the comparison was to consider the differences in the rules mined. However, the unexpectedly close similarity between the discovered contrast sets from CSM and association rules from ARM led them to further investigations and to prove that CSM is "strictly equivalent to a special case of the more general rule-discovery task"[84] and that "existing rule-discovery techniques can be applied to perform the core contrast-discovery task". Since EP-mining and CSM are so closely related, it is clear that EP-mining is also a special case of rule discovery. However, Webb et al. do emphasize that contrast set discovery is "a new and valuable data mining task".

5.4. RST and JEPs. The close connection between RST and JEPs has been explored by Terlecki and Walczak,[85] and this discussion follows their paper closely. Consider a decision table (U,C,D), where D contains the single decision variable d which can take just two values, 0, 1. We can regard D as the union of two decision classes, $D = D_0 \cup D_1$. (All decision tables may be regarded in this way, by setting $D_1 = D_0' = D - D_0$.). For an object $u \in U$ and a subset $P \subseteq C$ of attributes, we can define a unique itemset or pattern, of (attribute, attribute-value) pairs:

$$\mathrm{patt}(u, P) = \{(a, a(u)) : a \in P\}$$

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

A transaction is a pattern which involves all attributes, i.e., P = C. Using this formulation, the decision table used in RST is easily transformed into the two transaction tables required to find JEPs. Then the relationship between differentiating sets and JEPs is given by

Equivalence 1

> Given a differentiating set P of attributes, an object u can be positively assigned to the decision class d = 1 if and only if patt(u, P) is a JEP from D_0 to D_1

Terlecki and Walczak prove several other important equivalences[85] and also give relationships between the support of JEPs and the cardinality of the positive region (i.e., the number of objects which are definitely in the decision class d = 1). Comparing the support of JEPs with the size of the positive region gives a method for deciding if P is a differentiating set and further if P is then a reduct. Finally, Terlecki and Walczak give a lower bound on the size of differentiating sets (and therefore reducts) in terms of the size of the minimal JEPs in the left border. This is very useful as minimum JEPs are easy to calculate. Terlecki and Walczak propose using some of the relationships they have described in minimal reduct calculation which is an NP-hard problem. They propose, for example, to use left borders of equivalent JEP sets as part of the initialization of the population of evolutionary algorithms.

5.5. FCA and JEPs. In order to show the relationship between FCA and JEPs we need some further terminology. As FCA is unsupervised, we must first introduce supervision. Ganter and Kuznetsov define a *goal attribute* as an attribute $w \notin A$, which partitions the object set U into three subsets, viz., those objects which possess attribute *w*, denoted U_+, those which do not possess *w*, U_-, and those for which it is not known whether they possess *w*, U_τ.[86] In effect this goal attribute corresponds to the decision attribute of our previous discussion, with U_τ comprising objects awaiting classification. The goal attribute can be used to partition the formal context K = (U,A,R) into three subcontexts, a *positive context* $K_+ = (U_+,A,I_+)$, a *negative context* $K_- = (U_-,A,I_-)$, and $K_\tau = (U_\tau,A,I_\tau)$, which is an *undetermined context*, whose objects, U_τ, have yet to be classified. ($I_+ = I \cap U_+ \times A$, and I_- and I_τ are defined analogously.)

A *positive hypothesis* with respect to *w* is defined as the intent (attribute set) of a concept of the positive context K_+ which is not contained in the intent of any concept of the negative context.[87] A negative hypothesis is defined analogously. Additionally, we require that a hypothesis is contained in at least two examples. We can construct two transaction tables, corresponding to the positive and negative contexts, defined in exactly the same manner as we did to show the correspondence between JEPs and RST. Then the intents of the positive hypotheses are exactly the JEPs, of size at least 2, from the negative to the positive transaction tables.

In FCA, positive hypotheses are used to classify objects from the undetermined context as follows: If the attributes of an undetermined object contains at least one positive hypothesis and NO negative hypotheses, then the object is classified positively. Negative classification is determined similarly. An object whose attributes contains neither positive nor negative hypotheses remains unclassified, whereas if an object contains both positive and negative hypotheses, then the classification is *contradictory*.

5.6. RST and FCA. According to Poelmans, many attempts have been made to combine FCA and RST with the hybrid usually referred to as Rough Formal Concept Analysis (RFCA).[72] Yao summarized the important relationship between FCA and RST.[88] He showed that they represent different approaches to analyzing data based on two goals of data mining, prediction and description. RST aims to predict the membership of an equivalence class based on the attributes of an object—it depends on *sufficiency* conditions. In other words, if an object has certain attributes, those attributes are sufficient to guarantee that the object belongs to a particular class. FCA describes the members of a class based on *necessary* conditions. If an object is in a particular class, it necessarily possesses certain attributes since all objects within that class do possess those attributes. Despite this analysis, FCA is certainly used predictively in many areas of KDD.

The relationship between FCA and RST is very close. Zhang et al. showed the similarities by defining notions from FCA using the terminology of RST. They showed that every Formal Context possesses at least one reduct.[89] In an alternative approach, termed Formal Rough Concept Analysis (FRCA), Saquer and Deogun look to find approximate concepts.[90] Recall that not every pair of a set of objects and a set of features constitutes a concept. So, one of the problems posed by Saquer and Deogun is given a pair (*A*,*B*), where *A* is a set of objects and *B* is a set of features, find a formal concept that approximates (A,B). They prove, using rough sets, that their approximations are, in some sense, the best that can be achieved.

5.7. ARM and FCA. Several authors independently investigated the connections between ARM and FCA in 1999.[91−93] In particular, they each formalized the relationship between closed itemsets and concepts. Most ARM mining algorithms prior to this needed to compute the support of all frequent itemsets, either incrementally during the processing of the breadth-first search in Apriori-like algorithms or during a final database scan in the case of Max-Miner type algorithms. Now, recall that in ARM a *closed itemset* is an itemset for which no immediate superset has the same support. In FCA, an itemset X is closed if, and only if, X** = X (where ∗ is the derivation operator). The insight of Stumme and others was that these two statements are equivalent. A *frequent closed itemset* is then a closed itemset with support at least minsup. Using results from FCA these groups showed that it was sufficient to compute the support of all frequent closed itemsets—from these counts all other supports could be derived.[91] In fact frequent closed itemsets constitute a generating set for all frequent itemsets and also for association rules.[92]

An association rule r *holds* if it has support ≥ minsup and confidence ≥ minconf. If conf(r) = 1 then r is an exact association rule (or implication rule) otherwise it is approximate.[92] A reduced set of association rules is known as a *basis*. Pasquier at al. proposed algorithms for discovering several (relatively) small bases for both exact and approximate association rules. This approach not only reduces the computation but also the number of redundant rules discovered.

5.8. Which Method To Choose. The pairwise comparison of methods has shown that there are close relationships between them. Often apparent differences lie in the way in which the problem under investigation has been formulated, which can lead to the natural adoption of one or other method.

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

For example, the idea of a structural alert, as a substructure which is present in toxic and absent in nontoxic compounds, leads to the natural formulation of an Emerging Pattern search. The work of Ganter showed that, in fact, the data in a Formal Context could be formulated in a similar manner as long as a goal (decision) attribute could be defined.[86] Thus, it should be possible to search for JEPs, and possibly EPs, using FCA algorithms. Similarly Webb et al. showed that EP is a special case of ARM and demonstrated that their OPUS system could successfully be used for EP.[40] The equivalence between ARM and RST when searching for decision rules is very interesting but does not mean that RST should be disregarded since its treatment of inconsistent rules (where the same antecedents lead to different decisions) is unique and is an important feature of the method. The very close relationship between ARM and FCA[91−93] does not mean that either is redundant. A very important aspect of FCA is the hierarchical view of the data relationships given by the Hasse diagram, while ARM is somewhat more general and may be used where visualization is less important.

It would be helpful to be able to provide a guide as to which of RST, ARM, EP, and FCA is most appropriate in which circumstance. However, it does not seem to us that there is enough evidence yet on which to base any recommendation. There have, to our knowledge, been no reported comparisons of the use of these methods on any chemoinformatics application (or indeed many other application areas). In any case, such a comparison might be invidious since it would depend on the choices of algorithms, rule-pruning techniques, data discretization methods, etc.; none of which have been themselves sufficiently well investigated in a chemoinformatics context. One of our primary aims, in writing this review, was to encourage the use of RST, ARM, EP, and FCA in the hope that, in the future, such comparisons might be possible, and then sensible recommendations as to their applicability might be proposed.

6. APPLICATIONS

In this section, we consider chemoinformatics applications of the four data mining methods, along with frequent subgraph (FSG) mining. FSG applications have been included since FSGs are often mined using ARM and EP algorithms.

6.1. RST Applications. A search of the literature revealed more uses of RST in chemoinformatics, and particularly SAR and QSAR, applications than was expected. The earliest reported use of RST theory for SAR found in our searches was in fact by Krysinski in 1990.[94] In this and further reports, he looked at relationships between structure and antimicrobial or antifungal properties of fairly small sets of compounds.[95−97] The motivations for his work are that, first, RST provides an interpretable model, as opposed to the "black box" of traditional QSAR methods and, second, RST does not require "deformation of the data" by, for example, using mean or probabilistic values.[96] (This does rather ignore the fact that RST cannot deal with continuous valued-attributes which have to be discretized.) Krysinski studied the relationship between structure and antifungal properties of a set of 264 imidazolium chloride compounds.[96] The eight attributes were imidazole substituents at four R-group attachment points, and the compounds were classified into five classes according to discretized values of minimal inhibitory concentration (MIC) for two different fungi. After obtaining the (only) reduct, which comprised seven of the eight attributes, attributes were systematically removed, and the effect of this on the quality and accuracy of the classification was observed. This allows one to ascertain the relative importance of single attributes and attribute combinations with respect to antifungal activity. In this manner, four of the eight attributes were determined to classify the compounds with very high quality. A total of 173 decision rules containing these attributes were obtained and analyzed to determine which substitutions were associated with the highest and lowest rates of antifungal activity. This kind of analysis is typical of the way in which RST is used in SAR. In 2002, Krysinski et al. followed up their earlier work by focusing on the ability of RST to provide decision rules which explain activity.[97] They studied the SAR of 180 antielectrostatic imidazolium compounds and showed that RST was able to explain much of the activity of the compounds using a reduced set of only four descriptors. They used the idea of strong decision rules (with high support) to eliminate many rules. The remaining rules could be used to aid decisions about the synthesis of further active compounds. This study was followed by another on the relationship between structure, surface properties, and antimicrobial activity of quaternary ammonium chlorides in two microbial strains, which gave less conclusive results but was still able to give some guidelines for desirable structural properties for antimicrobial activity against one of the strains.[98] The Krysinski group still work in this area. Their most recent publication analyzing the SAR of antimicrobial activity of imidazolium chloride compounds was published in 2014.[99]

Despite his many reports, Krysinski's work does not seem to have been taken up much by the wider SAR community. It has however been successfully used in the synthesis of new compounds with desirable antimicrobial properties.[100] The first widely cited RST paper in a chemoinformatics application was by Walczak and Massart in 1999.[17] This article provides a helpful tutorial on RST, combined with a worked example of a small QSAR literature problem, viz., that of finding a minimum set of features necessary for predicting the unfolding energy of a mutated protein. Hasegawa et al. performed a "validation" of RST for descriptor selection[101] using a well-studied set of DHFR inhibitors. They found that RST could be used to derive "easily interpretable" rules for high activity which were in good agreement with those found experimentally and also that when using the four descriptors found by the reducts as input to an SVM an even more predictive model was produced, suggesting that RST could indeed be used for feature selection. The authors then suggested that RST should be tried on more challenging data sets.

Liu et al. used RST in a three-dimensional SAR study.[102] Semi-empirical quantum chemistry methods were used to generate properties for fluoroquinolines, then the core and reducts were extracted along with decision rules. The extracted rules had "unambiguous physical meaning" for bacterial fluoroquinoline resistance. In further work they applied the same techniques for reduct and core extraction to quantum chemical properties of artemesinin derivatives.[103] Wang et al. studied the SAR of a number of compounds with cardiotonic activity.[104]

In the field of *proteochemometrics modeling* (PCM), molecular recognition is modeled using chemical and structural descriptors to represent both ligands and proteins,[24] and PCM can thus be considered an extension to QSAR.[105] Strombergsson et al. used RST-based PCM models to generate rules relating ligand and receptor properties to GPCR-binding affinity.[106] An advantage of RST is that it is a nonlinear method,

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

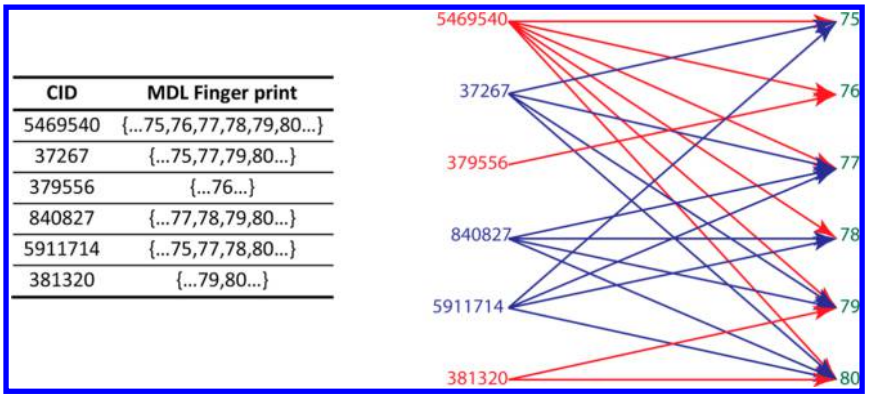


Figure 4. CID is a compound identifier; the numbers represent bits set in the MDL fingerprint. Reproduced from ref 113. Copyright 2012 PLOS One.

so that it is able to model cross-terms which are combinations of receptor and ligand properties. RST was used to generate rules, such as "R1 = Me and R2 = H = > high affinity". The authors concluded that the RST approach had several advantages over PLS. In particular, the rules are highly interpretable and provide a "holistic approach" to understanding the nature of GPCR-ligand interactions. Although the RST classification is coarser than that of PLS (as a result of the discretization of, for example, binding affinity into two classes), this did not affect the quality of the classification. In further work, they used cross terms, consisting of a receptor and a ligand descriptor, as single RST attributes in an analysis of enzyme-ligand interactions.[107] Kierczak et al. used RST to develop a "general and predictive" model of HIV reverse transcriptase (RT) resistance to RT inhibitors. The RST-derived rules were used to suggest possible molecular-mechanisms of drug-resistance.[108]

RST has been combined with Ant Colony Optimization (ACO). He et al. united the discretization of continuous attributes with attribute reduction in a single stage which was termed a biobjective optimization problem and used ACO to solve this problem. The algorithm was applied to classify two toxicity data sets and provided equal classification accuracy to SVMs and neural networks, while giving interpretable models.[109] Goodarzi et al.[10] used so-called "fuzzy" rough set ACO in order to select a set of descriptors for 2D QSAR. In this application, "fuzzy" means that set membership is defined by means of a membership degree function. No details of the RST methods are given, but the QSAR models produced using descriptors selected by fuzzy rough set ACO were shown to be at least as good as those where more traditional feature selection methods, such as GAs, were used.

Maji and Paul have developed a feature selection method called Maximum Relevance-Maximum Significance (MRMS).[110] They use this for feature selection in QSAR. They begin by defining an information table I from the n molecules of the QSAR data set, U, as I = (U, M ∪ B), where M is a set of m molecular descriptors, and the decision variable is a biological response variable, denoted B. The values of the descriptors in M and the biological response B have to be discretized. They use the notion of dependence between descriptors and biological response, which is defined as

$$\gamma_{\mathrm{M}}(\mathrm{B}) = \frac{|\mathrm{POS}_{\mathrm{M}}B|}{|\mathrm{U}|}$$

where $\mathrm{POS}_{\mathrm{M}}(\mathrm{B})$ contains all the molecules of U which can be classified into a class of B using the descriptors in M. The relevance of a single descriptor, M_i, is defined as the dependence of the biological response upon that descriptor alone, i.e.,

$$\widehat{f}\,(\mathrm{M}_i, \mathrm{B}) = \gamma_{\mathrm{M}_i}(\mathrm{B})$$

and the relevance of a set of descriptors is the sum of the individual relevance values. In practice, one is looking for a set of relevant and significant descriptors to avoid redundancy. Feature selection is accomplished using a greedy algorithm for maximizing a fitness function which is the sum of the significance and relevance measures. The method has low computational complexity, of O(m). In tests on three small QSAR sets, it performed very favorably in comparison to several other methods.

Dong et al. used RST for both feature selection and model building in a SAR analysis of indolinone inhibitors of CDK1/cyclinB.[26] RST was used as a "preprocessor" with reducts being generated for feature selection. The reducts were then used to generate decision rules, with the SAR model being at least as good as that obtained by eight other nonlinear machine learning methods, such as multi-layered perceptron and logistic model trees. In their discussion, the authors concluded that RST was very useful for dimensionality reduction but that there were limitations in using RST for their SAR analysis. For example, different discretization methods lead to different reducts.

Very recently, RST has been reported as a method for the prediction of Cytochrome P450 inhibition.[18] The main attraction of the method is the interpretability of the results. Compounds were represented both as MACCS keys and as in-house electron density-derived fingerprints, and RST was used to generate rules both to classify compounds and to predict the inhibitory profile of unknown molecules.

6.2. ARM Applications. Most of the few ARM mining applications in chemoinformatics have been based on finding frequent subgraphs, as discussed in the section on FSG below. An exception is the work of Yu and Wild.[42] They use structural fingerprints, MDL public keys, and PubChem CACTVS, each of which has a 1−1 mapping between bits set and the presence of a structural fragment, meaning that the rules generated can be interpreted. They also use physicochemical properties, continuous ones being discretized using an entropy-based method. They compared three different ACM methods for classifying three different data sets represented by each of the

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

structural fingerprints and by the physicochemical properties, giving 27 ACM classifiers which were also compared with SVM and naïve Bayesian versions of the same models. The ACM methods were as follows: classification based on association rules (CBA),[111] classification based on predictive association rules (CPAR),[112] and classification based on multiple class-association rules (CMAR).[113] The ACM methods performed as well as the traditional classifiers, and of the three, CBA gave the highest relative accuracy. A significant drawback of ARM methods is that usually all features (items) are given equal weight. Simply selecting a subset to represent the entire set of features may lose valuable information; for example, the presence of rare mutations in a gene may be very important but only found in a very few cases. Feature weighting retains all features, merely assigning lower weights to what may be unimportant features. Wang and others developed weighted association rule mining (WARM) to address this issue,[114,115] but it relies on the database having preassigned weights. In further work, Yu and Wild developed a link-based weighting scheme.[116] They represent a database as a bipartite graph with compounds and features they contain, being linked like web pages (see Figure 4). They combined the Google PageRank algorithm[117] and the HITS (Hyper Induced Topic Selection) algorithm[118] to define a novel weighting scheme. Weighted itemsets do not necessarily possess the downward closure property which enables frequent itemset pruning. They therefore introduce an *adjusted weighted support* which does possess the DCP. They implemented a link-based associative classifier, which they term LAC, using their weighted features and WARM with the CBA method.[111] They applied this to an Ames mutagenicity data set and the NCI-60 tumor cell line. They report that LAC was able to find associations between bioactivities and chemical features.

Raymond et al. also use ARM techniques in an effort toward rationalizing lead optimization.[119] The basic method is first to assemble a database of transactions where a transaction, termed a Molecular Substructure Modification (MSM), consists of a pair substructures (X,Y) together with a mapping which indicates their attachment point to a common core. The MSMs are found by an all-by-all pairwise MCS-based comparison between the compounds of a large database. MSMs represent attempts by medicinal chemists at structure modification. ARM can then be used to extract "rules" which represent likely transformations between substructural fragments, given their core. They pay particular attention to extracting rules with relevance, noting that support and support count are particularly poor measures of relevance in this context, where infrequent MSMs can be of interest. Additionally, frequent MSMs can be either obvious (e.g., ethyl/methyl) or frequent by chance since they include commonly occurring fragments. To score rules, they use a measure called hyper-lift[120] which takes account of low probability and randomly occurring events. Hyper-lift uses a cumulative hypergeometric probability distribution which can, under certain circumstances, be undefined. In these cases, the authors use a second measure, all-confidence.[121]

ARM methods can be incorporated as part of other machine learning paradigms. For example, Inductive Logic Programming (ILP) is machine learning technique which derives hypotheses from data. A database of chemical compounds is represented as a DATALOG[122] database, where both the structures and their properties are described using DATALOG concepts. In this context, patterns are termed queries. The WARMR program[123,124] extends the Apriori algorithm using ILP methods to discover frequent queries in the data. When finding frequent substructures, compounds are considered as sets of atoms, each with its bond connectivity, both of which are represented as DATALOG facts. King et al. used WARMR on data from the PTE to predict toxicity,[124] whereas Dehaspe at al use WARMR, not for classification, but as a tool to identify frequent patterns, so that these substructures could be used as a basis for carcinogenicity research.[123] Here, "substructures" are in fact collections of atom and/or bond environments and are not necessarily connected.

6.3. EP Applications. JEPs and EPs have been used in classifying genomic data[54,61,80] and image classification.[62] The first use of JEPs in chemoinformatics seems to be by Blinova et al.,[87] although they do not use the terminology, but instead call their method the JSM-method and present it in terms of Formal Concept Analysis (see the discussion in the FCA section below). Their term for a JEP is a "counterexample-forbidding hypothesis", which is taken from the original work of Finn (in Russian). Compounds were represented using the FCSS descriptor language[125] whereby a molecule is represented as a set of substructures centered on localized π−electrons. In FCSS, there are three types of descriptors: linear, cyclic, and substitution. A set of these descriptor codes is generated for a molecule, and the molecule is then encoded by concatenating the codes in lexicographic order. The authors entered one of the Predictive Toxicology Challenges[126,127] and concluded that their method is conservative, that is, they made few predictions with almost no errors. This is a likely feature of using JEPs whereby counterexamples are not allowed, so that not many cases meet the criterion for a positive prediction.

Bajorath and co-workers were the first to use the term JEP in a chemoinformatics context[128] with the explicit aim of molecular classification. They adopt the notions of Li et al., who defined a *most expressive* JEP as a JEP for which no proper subset is also a JEP and for which no superset has larger support within the data set.[47] Bajorath and co-workers use the term *Emerging Chemical Pattern* (ECP) as a synonym for "most expressive JEP" in the context of descriptor-dependent feature analysis. They used a set of 61 noncorrelated descriptors from MOE. Continuous descriptors require discretization for use with EP algorithms. Auer and Bajorath tested both an equal-binning method (which takes no account of the relative frequency of the descriptor values) and a method based on the information content of the bins. In all their experiments, this second method gave the better outcomes. ECP-mining is computationally intensive, but the authors report the hypergraph-traversal algorithm of Bailey et al.[129] to be adequate for the task. After mining ECPs, a molecule is classified based on the numbers of ECPs from the active and inactive classes which it contains. Auer and Bajorath used ECPs to predict the class of both active and inactive molecules; they were able to do this successfully using only three most active and three least active compounds. This work was followed by the application of their ECP-methodology to more complex classification problems, such as data sets where active molecules have different modes of action and data sets containing compounds active against more than one target,[130] the prediction of single compounds which will form an activity cliff,[131] and the prediction of single compounds which have different local SAR environments.[132] They report similar success with these difficult data sets, especially with regard to specificity, although sensitivity can be variable, i.e., ECPs produce few false positives but only classify

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

subsets of the compounds. This is probably to be expected since, by the definition of an ECP, we require features present in some active compounds with strong support but entirely absent from the inactives or present in inactives and absent from the actives; this is a stringent requirement, which many active compounds will not meet. Again, they report that successful classification requires very few positive instances, as few as three. In fact, increasing the number of positive instances to 10 from three did not increase the prediction accuracy. This emphasizes the potential for using ECPs in early stage drug discovery, where limited target information may be available. A limitation of the use of JEPs (and EPs) for classification is the requirement for negative examples to be present. Tested inactive compounds are needed; this was achieved by the use of confirmed bioassays. A more novel application was the use of ECP methodology to distinguish bioactive from modeled 3D ligand conformations in 18 different inhibitor classes.[133] Here, the descriptors were 67 3D conformation-specific ones, such as strain energy and bond stretch energy, discretized into value-ranges. Typically, only up to three descriptor value ranges were needed to distinguish bioactive conformations.

Sherhod et al. have a different emphasis in their use of JEPs, with the goal being not classification but description.[63] Structural Alerts (SAs) are structural features which, taken together, may indicate toxicity. JEPs which are present in toxic but absent in nontoxic compounds can be used as an aid to expert curation in the design of alerts for incorporation in toxicity knowledge-bases. Sherhod et al. represent molecules by atom-pair descriptors. An advantage of this over frequent subgraph mining approaches[134] (see below) is that the resulting JEPs can consist of disconnected fragments. They use the Horizon-Miner algorithm[47] to find the right bound of the border of the set of JEPs. They then use the border differential algorithm[50] to mine the minimal JEPs present in the toxic molecules but not in the nontoxics. The JEPs are then pruned so that only those occurring in distinct sets of actives remain. The actives supporting a JEP can then be arranged hierarchically since they usually contain other commonalities (which are not part of the JEP). In a follow-up paper, the authors used EPs rather than JEPs.[135] This increases the tolerance to noise which is important for the design of SAs since it is vital not to miss structural features which are necessary for toxicity. In this case, they used the Contrast Pattern Tree algorithm[48] to mine EPs rather than JEPs. Some of the resulting EPs have been used by experts as the start point for a number of new SAs for *in vitro* mutagenicity.

6.4. Frequent Subgraph Mining. When the chemical structures of molecules are considered as graphs, the discovery of subgraphs which occur more frequently in some molecular graphs than in others has long been of interest. This is clearly related to Emerging Pattern Mining. Frequent Subgraph Mining has recently been reviewed by Takigawa.[134] Here, we only consider a subset of the FSG literature, as it relates to the four KDD techniques under study. An earlier review by Fischer gives a nice outline of the problem[136] and illustrates the basic techniques used by most algorithms. The similarity to EP is apparent when two collections of graphs (molecules) are considered. The usual graph-theoretic notation is needed for further discussion: An undirected graph, G = (V,E) is composed of a set V of *vertices* or *nodes* and a set E of *edges* where E ⊆ {(u,v): u,v ∈ V}. Usually for molecules represented by graphs, vertices represent atoms and edges represent bonds. Two vertices u and v are adjacent if (u,v) ∈ E; an edge e = (u,v) *connects* u and v. The edge (u,v) may be referred to simply as uv. A graph is *connected* if every vertex is connected to at least one other vertex. A graph $G' = (V',E')$ is a *subgraph* of G if $V' \subseteq V$ and $E' \subseteq \{(u,v): u,v \in V' \text{ and } (u,v) \in E\}$. (Informally, $E' \subseteq E$ and $G' \subseteq G$). A subgraph is *induced* if $u,v \in V'$ and $(u,v) \in E \Rightarrow (u,v) \in E'$. An *isomorphism* between two graphs G = (V,E) and $G' = (V',E')$ is a one-to-one mapping Ø: G → G' such that if uv ∈ E then Ø (u) Ø (v) ∈ E′. Thus, isomorphism preserves adjacency. If the graph is labeled, then isomorphism must also preserve the labeling. Clearly |V| = |V′|. Usually, for chemical graphs, vertices are labeled with atom type and edges with bond type. A subgraph isomorphism is a mapping from a graph G to a subgraph H of G', such that G is isomorphic to H. A common subgraph of two graphs, G_1 and G_2, consists of a subgraph H_1 of G_1 and subgraph H_2 of G_2 such that H_1 is isomorphic to H_2. A maximal common subgraph is a common subgraph which is not strictly contained in any larger subgraph. Given a graph G = (V,E), the graph $\bar{G} = (V, \bar{E})$, where $\bar{E}$ contains precisely those pairs (u,v) where $u,v \in V$ and $uv \notin E$, is called the *complement* of G. Given a threshold, t, a graph G is a *frequent subgraph* in a collection of graphs if G is a subgraph of at least t of the graphs.

As far back as 1984, Klopman introduced a computer program, CASE, whose aim was to mine all substructures in active and inactive compounds.[137] The distribution of substructures was modeled by a binomial distribution and substructures found in the actives with a frequency significantly greater than expected were considered to be implicated in activity and termed *biophores*. In a later development an enhanced version of CASE, termed MULTICASE, was able to order the biophores in a hierarchical manner.[138]

Sets of molecular fragments, frequent in one set of molecules and infrequent in the other, have variously been termed *Emerging Graph Patterns*,[53] *discriminative fragments*,[139] and *contrast subgraph patterns*.[140] The first subgraph mining algorithm to appear was SUBDUE.[141] This algorithm and its many extensions have found application in a large number of areas including bioinformatics[142] and chemoinformatics.[143] Recently a multi-objective version, MoSubdue was introduced and evaluated on the PTE.[144]

Ting and Bailey introduce a *contrast subgraph* informally as "a subgraph appearing in one class of graphs, but never in another class of graphs".[140] They consider a positive graph, denoted G_p, and a negative graph (or graphs), denoted G_n. Formally, a subgraph $G' \subseteq G_p$ is a contrast subgraph if it is not isomorphic to any subgraph of G_n and is a minimal contrast subgraph if none if its subgraphs are contrast subgraphs. They define a *minimal contrast edge set* which is a connected minimal contrast subgraph. Key to their approach is the duality between the minimal contrast edge sets and the maximal common subgraphs. To see the connection, we first need the notion of a transversal. Given a collection of sets, $A = \{A_1,A_2,...,A_m\}$ a set P is a *transversal of A* if P has a nonempty intersection with every A_i in the collection. P is a *minimal transversal* if no proper subset of P is a transversal. Let M be the collection of maximal common subgraphs between G_p and G_n and let $\bar{M}$ denote the collection of the complements of the subgraphs in M. The minimal transversals of $\bar{M}$ are precisely the minimal contrast edge sets between G_p and G_n. The maximal common edge sets between G_p and G_n are found using a back-tracking algorithm similar to that of McGregor.[145] The minimal transversals are found using Bailey's earlier hypergraph transversal algorithm.[129] In practice, the method is applied by comparing each positive graph in turn against the union of the set of negative graphs.

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

Despite the use of ordering and pruning techniques, this approach, although very successful for synthetic data, performed less well for the PTE data, scaling exponentially with the number of negative graphs. However, it does have the advantage of finding disconnected contrast subgraphs, unlike most competitor algorithms.

Dominik et al. use the common and contrast pattern terminology of Ting and Bailey. Their algorithm (for which few details are available, although it seems to be based upon the gSpan algorithm[146]) is called the Contrast Common Patterns Classifier (CCPC).[147,148] Their main contribution is a hierarchical scoring function, whereby compounds are first assigned to a class based on their contrast pattern score, with ties being broken using a common subgraph score. They also report that using scores based on growth rate gives better classification than using scores based on support. Results are again reported using PTE data.

Emerging Graph Patterns are described in Poezevara et al.,[53] and this discussion is based on that work. Suppose we have two collections of graphs, D_1 and D_0. A graph pattern is a set of graphs. The *support* of a graph pattern is the proportion of graphs which contain it. Let $\mathcal{G}$ be a graph pattern. Growth rate, GR, from D_0 to D_1 is defined in a similar manner to that for emerging patterns, viz.,

$$\mathrm{GR}(\mathrm{G}) = \frac{\mathrm{supp}(\mathrm{G})_{\mathrm{D}_1}}{\mathrm{supp}(\mathrm{G})_{\mathrm{D}_0}}$$

An Emerging Graph Pattern (EGP) is a graph pattern whose frequency increases significantly from one class to another. More specifically, given a growth threshold t, a graph pattern, $\mathcal{G}$, is an EGP from D_0 to D_1 if $\mathrm{GR}(\mathcal{G}) \geq t$. Importantly, the EGPs of size 1 (i.e., containing only 1 graph) are precisely the set of frequent emerging connected graphs from D_0 to D_1. However, an advantage of the EGP approach is that it allows for the finding of disconnected fragments which may, when found in combination, cause toxicity; these sets of fragments are the EGPs of size greater than 1.

Finding frequent subgraphs is the problem of performing multiple subgraph isomorphisms. Subgraph isomorphism is needed for two purposes: (1) in order to check if a subgraph is frequent and (2) in order to avoid the generation of a duplicate subgraph. Since subgraph isomorphism is an NP-complete problem, the main idea is to reduce the number of subgraph isomorphism calculations to a minimum. Poezevara's method involves transforming the problem of comparing two graph collections into a different problem. They find all frequent connected subgraphs in D_1, the collection in which emerging graphs are required, which restricts the number of graphs in D_0 which have to be considered for subgraph isomorphism. This enables them to transform the problem into one where the presence/absence of the frequent subgraphs can be represented by bits set or not. This is then amenable to an efficient method for extracting emerging patterns for which they use the MUSIC-DFS algorithm.[54] Their method is used in further work from the same group, in which Lozano et al. rename a JEP as a *jumping fragment*, where the jumping fragment is a substructure present in one group of molecules and not another. The extracted jumping fragments are used as a feature selection method prior to QSAR modeling for toxicology prediction.[149]

6.5. Association Rule Mining for Frequent Subgraphs in Chemical Databases. ARM algorithms can be adapted to mine frequent subgraphs. In the first direct adaptation of the Apriori algorithm, Inokuchi et al. introduced the notion of a labeled graph as a transaction and then defined Graph Data as a set of graph transactions, $\mathrm{GD} = \{\mathrm{G}_1, \mathrm{G}_2, .., \mathrm{G}_n\}$.[150] Graph transactions have their vertices sorted according to their labels and then are represented as canonical adjacency matrices. An induced subgraph is the equivalent of an itemset. Support and confidence are defined as for itemsets. Thus, for an induced subgraph, G_s,

$$\sup(\mathrm{G}_s) = \frac{\text{number of graph transactions G where } \mathrm{G}_s \subset \mathrm{G} \in \mathrm{GD}}{\text{total number of graph transactions G} \in \mathrm{GD}}$$

And given two induced subgraphs, G_a and G_b, the confidence of the association rule $\mathrm{G}_a \Rightarrow \mathrm{G}_b$ is

$$\mathrm{conf}(\mathrm{G}_a \Rightarrow \mathrm{G}_b) = \frac{\text{number of graphs G where } \mathrm{G}_a \cup \mathrm{G}_b \subset \mathrm{G} \in \mathrm{GD}}{\text{number of graphs G where } \mathrm{G}_a \subset \mathrm{G} \in \mathrm{GD}}$$

Their algorithm AGM (Apriori-based Graph Mining) is then a BFS algorithm which proceeds by candidate generation at the k-th level by the merging of two frequent induced subgraphs which have identical adjacency matrices up to their k-1th row. Just as for Apriori, candidates are pruned based on the antimonotone support property and a minimum support level, minsup. After all frequent induced subgraphs have been found, rules are generated using the enumeration scheme of Agrawal. For use with chemical data artificial edges are added to the molecules so that all atoms within six bonds are pseudoconnected. This overcomes the limitations of the induced subgraph approach which otherwise does not permit the discovery of frequent proximal but nonbonded atoms. The AGM algorithm was tested on the PTE challenge data and was reported to find significant rules associated with carcinogenesis. An interesting finding was that rules of relatively low support (e.g., 10%) were very useful and worth finding although their discovery was very time-consuming, taking 10 days (in 2000).

The MoSS, **Mo**lecular **S**ub**S**tructure miner, program (also known as MoFa, for **Mo**lecular **Fra**gment miner) is designed to find frequent substructures in a set of molecules and can also be used to find substructures which are infrequent in a complement set; these are termed *contrast structures*.[139,151] MoSS is an adaptation, to include subgraph isomorphism calculations, of the ECLAT algorithm for finding frequent itemsets.[152] This is a depth-first search (DFS) algorithm, carried out on a tree of substructures. Substructures are extended by adding bonds (or atoms) on moving down the tree. The ordering which is used means that the antimonotone property can be applied to prune the frequent substructures, based on their support, since adding atoms decreases the support of a substructure (i.e., making it bigger means fewer molecules contain it). One method for avoiding subgraph isomorphism, where possible, is to maintain so-called *embedding* lists for frequent subgraphs. An embedding is a list of matched nodes and edges for each graph which contains the frequent subgraph (and so involves multiple subgraph isomorphism calculations). The advantage is that extensions of the frequent subgraph only involve checking for the presence of the additional bond. A disadvantage is the memory required for storing the embedding lists.

ARM algorithms are particularly prone to generating huge search trees. The initial MoSS algorithm performed substantial amounts of redundant searching,[151] but subsequent modification to deal with rings improved this, meaning that the

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

algorithm is able to deal with large databases.[139] One disadvantage is that MoSS is only able to find connected substructures.

Wolohan et al. adapted the Apriori algorithm in an approach they term Structural Unit Analysis (SUA).[41] Data sets containing both active and structurally similar inactive compounds were assembled and molecules were fragmented into *structural units* by breaking bonds in a method similar to that of RECAP.[153] Connected graphs are formed using combinations of structural units; those graphs not present in a minimum number of molecules are pruned using a modified Apriori algorithm. Pruning is also performed based on the activity of the supporting molecules. The output rules are called *unit graphs*. The significance of a unit graph is determined by splitting the entire set of compounds into those which contain the graph and those which do not. The correlation of presence with activity is assessed using a statistical test. In experiments, sets of 2300 molecules generated only 25 rules, which were able to discern up to eight different scaffolds.

Frequent SubGraph discovery (FSG)[154] is a BFS Apriori-like algorithm which operates by first generating all frequent subgraphs with one and two edges. In the main loop it then, at each level, generates all candidate subgraphs obtained by the addition of a single edge and prunes those which do not satisfy the support criterion (using the antimonotone support property which allows efficient pruning). FSG was again benchmarked on the PTE data set. MISMOC is an algorithm for Mining Interesting Substructures in MOlecular data for Classification[155] which can use any frequent subgraph miner, and it has been implemented using FSG. Its unique feature is that, when classifying an unknown molecule, it uses measures of interestingness of the matched frequent subgraphs in order to place the molecule into the appropriate class.

Apriori-like approaches to frequent subgraph mining suffer from two problems, as described by Yan and Han.[146] First, the generation of the next level of candidate sets is more expensive when dealing with graphs than when dealing with itemsets. Second, using subgraph isomorphism to prune the search tree is, as already mentioned, a computationally demanding process. Yan and Han aim to overcome both these challenges using their gSpan (graph-based Substructure pattern) algorithm. This is a DFS algorithm in which the problem of subgraph isomorphism is transformed into one of mining minimum DFS codes. This is achieved by means of a new DFS lexicographic ordering, the existence of which ensures that, of all possible orders, there is a minimum which becomes a canonical labeling of the search tree. Then frequent subgraph mining becomes the somewhat easier problem of frequent pattern mining. Yan and Han report that gSpan is an order of magnitude faster than Apriori-like algorithms. There are many extensions to gSpan. For example, Thoma et al. adapt gSpan to find discriminative features among frequent subgraphs for two-class and multi-class classification problems.[156,157]

The Fast Frequent Subgraph Miner (FFSM) algorithm of Huan et al. is another DFS algorithm[158] which uses embedding lists. It also has a novel matrix adjacency representation, where diagonal elements represent node labels and other elements represent edge labels. They use this representation together with lexicographic ordering to generate a canonical graph code (CAM), whereby isomorphic graphs have the same CAM code. In further work, FFSM is used as a feature extraction method, and the features are then ranked using a Feature Consistency Map, which takes account of the (2D) spatial relationship between the features.[159,160] Very recently, FFSM has been used to extract frequent subgraphs which were used as descriptors in k nearest neighbor QSAR models.[161] Models were developed for various data sets, including Maximum Recommended Therapeutic dose and Salmonella Mutagenicity, and the model accuracy was at least as good as those reported in the literature. Gaston (Graph/Sequence/Tree extraction) is another embedding list-based algorithm.[162] This uses the fact that acyclic graphs (trees) and paths can be efficiently enumerated and its search is constructed to deal with these fragment types first. It is also easy to avoid duplicating subgraphs which are paths or trees. In further work, the Gaston algorithm was used to mine a so-called "elaborate chemical representation" of annotated graphs; this was used to analyze mutagenicity data.[163] Gaston has subsequently been used by several other groups for frequent substructure mining, particularly in SAR/QSAR applications. van der Horst et al. used it to identify substructural features important for GPCR-binding.[164] Structural fragments with skin sensitization potential were extracted using Gaston and then combined with other descriptors to construct a recursive partitioning tree for classification of lymph assay data.[165] Wang et al. employed Gaston to extract SAs from carcinogenicity data.[166] They adapted Gaston so that redundant fragments were automatically pruned and combined the remaining fragments into Molecular Fragment Trees using a statistical method.

Worlein et al. performed a very thorough and unbiased comparison of the four subgraph miners, MoSS (called MoFa in this paper), gSpan, FFSM, and Gaston.[167] They recoded all algorithms using the same graph structures, in JAVA. They carried out tests on six chemical data sets ranging from the very small PTE to the complete NCI data set which at that time contained 237,000 molecules. Their conclusions are very interesting because they considered the type of operations performed by the various algorithms. They found that "contrary to common belief" embedding lists do not speed up the algorithms. They also found that using canonical representations instead of explicit subgraph isomorphism is more efficient for detecting unwanted duplicate candidates but better still is to avoid duplicate generation in the first place as is done by Gaston. All four algorithms scaled linearly with database size but by different factors.

All the frequent subgraph approaches considered so far have been based on essentially 2D structure diagram representations of molecules, called topological graphs. The only 3D method, as far as we are aware, is that of Deshpande et al.,[168,169] who consider a single low-energy conformation of a molecule in conjunction with its 2D graph. Finding frequent geometric subgraphs is inherently more difficult than in the 2D case since it is necessary to consider the relative orientation in space of possibly isomorphic subgraphs. Clearly, subgraph isomorphism in this case has to be orientation-invariant. Tolerance on matching distances is also a potential problem since interatomic distances in slightly different conformations of similar substructures may not be the same. Deshpande et al. take into account 3D conformation by using the average interatomic distance over all pairs of atoms in a substructure as a geometric graph invariant and just matching this single distance to within a tolerance. A geometric graph then consists of a topological graph together with an interval containing the average interatomic distance. They use their FSG algorithm[154,170] to find all frequent geometric subgraphs and all topological subgraphs. The substructures found are then input into a

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

feature selection process which itself involves association rule mining using the CBA[171] or CMAR[113] sequential covering algorithms. Finally, the compounds are represented as a frequency feature vector, where the features have been selected for the particular application by mining the chemical graphs of the training set. The test (or unclassified) compounds then have their feature vectors set; this involves some subgraph isomorphism calculations to see if the features in the vector are present in each test compound. In experiments, the feature vectors were input to an SVM classifier. Deshpande et al. report that, in almost all cases, the geometric subgraph method outperformed the purely topological subgraph method.

6.6. FCA Applications. FCA has been used quite extensively in investigating the QSAR of environmental pollutants.[172−175] The work uses the close relationship between partial order relations (POR) and FCA.[176]

Lounkine and Bajorath and co-workers have implemented techniques from FCA into what they term fragment formal concept analysis (FragFCA)[177,178] and molecular formal concept analysis (MolFCA).[179] In FragFCA, objects are molecular fragments (or combinations of fragments), and attributes are annotations such as potency or activity information. Compounds are fragmented using a hierarchical scheme, and from the set of fragments, a structural key-type fingerprint is generated. Combinations of up to four fragments from each compound were enumerated. In order to navigate the concept lattice, they make use of what they term *scales*, for example, a potency scale. This enables the selection of a sublattice which is more easily navigable. More complex queries can be assembled by combining scales. They used the ToscanaJ software[81] in their implementation. Using FragFCA they were able to identify *signature fragments*, combinations of fragments unique to active or highly potent molecules. They found that fragment pairs and triplets were more informative than single fragments or larger combinations.[177] MolFCA uses entire molecules as objects, rather than fragments, and operates on selectivity profiles rather than structural descriptors. Its aim is to identify molecules with similar selectivity profiles without the use of structural information.[179]

FCA is the basis of the KEM (Knowledge Extraction and Management) system of Jullian and Afshar.[180] They describe an application of KEM to toxicology prediction, where molecules are represented as structural fragments and then association rules are mined from the Galois lattice generated from the fragments. The rules mined include such things as the co-occurrence of pairs of fragments with the presence of particular fragments (or pairs) implying activity/inactivity.

Berasaluce and co-workers use the methods of frequent itemset searching and association rule extraction to find rules for the synthesis of organic molecules in reaction databases. Their particular focus is on changing the functionality of a molecule. Interestingly, they place much emphasis on the role of domain knowledge and expertise: "the computing process itself has to be guided by domain knowledge, at every step". A large set of functional groups which are called *functional graphs* or *blocks* is partially ordered based on a subgraph relation, forming a concept hierarchy. A block may be present in a reactant, in a product, or in both. A reaction can then be considered as three sets of blocks, namely, *destroyed blocks*, *formed blocks*, and *unchanged blocks*. This is represented as a Boolean table. A reaction database can thus be represented as a set of such tables, from which frequent itemsets, which in this case comprise sets of functional groups with associated support, and then association rules, can be found using the Close[39] and Pascal[181] algorithms, respectively.

7. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we have described four modern data mining techniques, Rough Set Theory, Association Rule Mining, Emerging Patterns, and Formal Concept Analysis, and we have attempted to give an exhaustive list of their chemoinformatics applications. One of the main strengths of these methods is their descriptive ability. When used to derive rules, for example, in structure−activity relationships, the rules have clear physical meaning.

This review has shown that there are close relationships between the methods. Often apparent differences lie in the way in which the problem under investigation has been formulated which can lead to the natural adoption of one or other method. For example, the idea of a structural alert, as a structure which is present in toxic and absent in nontoxic compounds, leads to the natural formulation of an Emerging Pattern search. Despite the similarities between the methods, each has its strengths. RST is useful for dealing with uncertain and noisy data. Its main chemoinformatics applications so far have been in feature extraction and feature reduction, the latter often as input to another data mining method, such as an SVM. ARM has mostly been used for frequent subgraph mining. EP and FCA have both been used to mine both structural and nonstructural patterns for classification of both active and inactive molecules. The apparent overlap between the methods needs to be investigated in chemoinformatics applications in order that practitioners are able to choose the most appropriate for their field of interest.

Since their introduction in the 1980s and 1990s, RST, ARM, EP, and FCA have found wide-ranging applications, with many thousands of citations in Web of Science, but their adoption by the chemoinformatics community has been relatively slow. Advances, both in computer power and in algorithm development, mean that there is the potential to apply these techniques to larger data sets and thus to different problems in the future. Applications are already being made in the field of chemogenomics[182] and proteochemometrics,[106] and there is scope for far wider use.

AUTHOR INFORMATION

Corresponding Author

*E-mail: v.gillet@sheffield.ac.uk.

Notes

The authors declare no competing financial interest.

Biographies

Dr. Eleanor Gardiner is a mathematician by training. After a career in commercial computing and education, she completed a Ph.D. in computational biology working on the application of novel graph algorithms to problems in macromolecular and small molecule structure. Since then she has completed numerous post-doctoral research projects in various areas including protein docking, DNA sequence/structure relationships at the genomic level, lead-hopping methods for drug research, and pharmacophore-elucidation. She developed the GAPDOCK genetic algorithm for protein−protein docking. Her interests include algorithm development, especially clique-detection, spectral clustering methods, and the application of novel machine learning algorithms.

Val Gillet is Professor of Chemoinformatics at the University of Sheffield. Her research interests are focused on the development and

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

application of chemoinformatics techniques in the design of novel bioactive compounds. She has expertise in data mining and machine learning methods including emerging pattern mining, multi-objective evolutionary algorithms, and graph theory. She has developed applications aimed at the identifying structure−activity relationships, toxicity prediction, pharmacophore elucidation, 3D similarity methods, and de novo design of novel compounds. She has also developed novel representation methods for chemical structures including reduced graphs, wavelet analysis, and reaction vectors. She has collaborated extensively with pharmaceutical and chemoinformatics software companies.

■ ACKNOWLEDGMENTS

The initial impetus for this perspective was a poster comparing FCA and JEPs by Muhammad Alkarouri, Richard Sherhod, Val Gillet, Philip Judson, and Nicola Richmond presented at the Fifth Joint Sheffield Conference on Chemoinformatics (http://cisrg.shef.ac.uk/shef2010/conference.htm) in July 2010. We thank Matthew Seddon for help with preparing this manuscript.

■ REFERENCES

(1) Vickery, B. Knowledge Discovery from Databases: An Introductory Review. *J. Doc.* **1997**, *53*, 107−122.
(2) Fayyad, U.; PiatetskyShapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery in Databases. *AI Magazine* **1996**, *17*, 37−54.
(3) Wang, H.; Klinginsmith, J.; Dong, X.; Lee, A. C.; Guha, R.; Wu, Y.; Crippen, G. M.; Wild, D. J. Chemical Data Mining of the NCI Human Tumor Cell Line Database. *J. Chem. Inf. Model.* **2007**, *47*, 2063−2076.
(4) Belohlavek, R.; Dvorak, J.; Outrata, J. Fast Factorization by Similarity in Formal Concept Analysis of Data with Fuzzy Attributes. *Journal of Computer and System Sciences* **2007**, *73*, 1012−1022.
(5) Pawlak, Z. Rough Sets. *Int. J. Comput. & Inf. Sci.* **1982**, *11*, 341−356.
(6) Wille, R. Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts. In *Ordered Sets*; Rival, I., Ed.; Reidel: Dordrecht, The Netherlands, 1982; pp 445−470.
(7) Agrawal, R.; Imieliski, T.; Swami, A. Mining Association Rules between Sets of Items in Large Databases. In *SIGMOD/PODS '93*; Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, 1993; Buneman, P., Sushil, J., Eds.; ACM Press: New York, 1993; pp 207−216.
(8) Dong, G.; Li, J. Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In *Conference on Knowledge Discovery in Data*; Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, 1999; Chaudhuri, S., Fayyad, U.,; Madigan, D., Eds.; ACM Press: New York, 1999; pp 43−52.
(9) Krysinski, J. Rough Sets in the Analysis of the Structure-Activity-Relationships of Antifungal Imidazolium Compounds. *J. Pharm. Sci.* **1995**, *84*, 243−248.
(10) Goodarzi, M.; Freitas, M. P.; Jensen, R. Feature Selection and Linear/Nonlinear Regression Methods for the Accurate Prediction of Glycogen Synthase Kinase-3 Beta Inhibitory Activities. *J. Chem. Inf. Model.* **2009**, *49*, 824−832.
(11) Helma, C.; King, R. D.; Kramer, S.; Srinivasan, A. The Predictive Toxicology Challenge 2000−2001. *Bioinformatics* **2001**, *17*, 107−108.
(12) NIH Developmental Threapeutics Program Web Site. http://dtp.nci.nih.gov (accessed April 14, 2014).
(13) Pawlak, Z.; Skowron, A. Rudiments of Rough Sets. *Inf. Sci.* **2007**, *177*, 3−27.
(14) Pawlak, Z. *Rough Sets. Theoretical Aspects of Reasoning About Data*; Kluwer Academic Publisher: Dordrecht, The Netherlands, 1991.
(15) Duntsch, I.; Gediga, G. *Rough Set Data Analysis: A Road to Non-Invasive Data Discovery*; Methodos Publishers: Bangor, U.K., 2000.
(16) Yao, Y. Y.; Zhao, Y. Discernibility Matrix Simplification for Constructing Attribute Reducts. *Inf. Sci.* **2009**, *179*, 867−882.
(17) Walczak, B.; Massart, D. L. Rough Sets Theory. *Chemom. Intell. Lab. Syst.* **1999**, *47*, 1−16.
(18) Burton, J.; Petit, J.; Danloy, E.; Maggiora, G. M.; Vercauteren, D. P. Rough Set Theory as an Interpretable Method for Predicting the Inhibition of Cytochrome P450 1a2 and 2d6. *Mol. Inf.* **2013**, *32*, 579−589.
(19) Swiniarski, R. W.; Skowron, A. Rough Set Methods in Feature Selection and Recognition. *Pattern Recogn. Lett.* **2003**, *24*, 833−849.
(20) Ziarko, W. The Discovery, Analysis, and Representation of Data Dependencies in Databases. In *Knowledge Discovery in Databases*, Piatetsky-Shapiro, G., Frawley, W. J., Eds.; AAAI/MIT Press: Cambridge, MA, 1991; pp 195−212.
(21) Predki, B.; Słowiński, R.; Stefanowski, J.; Susmaga, R.; Wilk, S. Rose - Software Implementation of the Rough Set Theory. In *Rough Sets and Current Trends in Computing*; Polkowski, L., Skowron, A., Eds.; Springer: Berlin, Germany, 1998; Vol. *1424*, Chapter 85, pp 605−608.
(22) Fibak, J.; Pawlak, Z.; Slowinski, K.; Slowinski, R. Rough Sets Based Decision Algorithm for Treatment of Duodenal Ulcer by Hsv. *Bull. Polym. Acad. Sci.: Biol. Sci.* **1986**, *34*, 227−246.
(23) Chen, Y.-S.; Cheng, C.-H. A Soft-Computing Based Rough Sets Classifier for Classifying Ipo Returns in the Financial Markets. *Appl. Soft. Comput.* **2012**, *12*, 462−475.
(24) Hvidsten, T. R.; Komorowski, J. Rough Sets in Bioinformatics. In *Transactions on Rough Sets VII: Commemorating the Life and Work of Zdzislaw Pawlak, Part II*, Peters, J. F.; Skowron, A., Marek, V. W., Orlowska, E., Slowinski, R., Ziarko, W., Eds.; Springer: Berlin, Germany, 2007; pp 225−243.
(25) Thangavel, K.; Pethalakshmi, A. Dimensionality Reduction Based on Rough Set Theory: A Review. *Appl. Soft. Comput.* **2009**, *9*, 1−12.
(26) Dong, Y.; Xiang, B. R.; Wang, T.; Liu, H.; Qu, L. B. Rough Set-Based Sar Analysis: An Inductive Method. *Expert Syst. Appl.* **2010**, *37*, 5032−5039.
(27) Bazan, J. A Comparison of Dynamic and Non-Dynamic Rough Set Methods for Extracting Laws from Decision Tables. In *Rough Sets in Knowledge Discovery*; Polkowski, L., Skowron, A., Eds.; Physica-Verlag: Heidelberg, Germany, 1998; pp 321−365.
(28) Grzymala-Busse, J. W. Lers - a Data Mining System. *Data Mining and Knowledge Discovery Handbook* **2005**, 1347−1351.
(29) Komorowski, J.; Ohrn, A.; Skowron, A. The Rosetta Rough Set Software System. In *Handbook of Data Mining and Knowledge Discovery*, Klosgen, W., Zytkow, J., Eds.; Oxford University Press: New York, 2002; Chapter D.2.3, pp 1554−1559.
(30) Bazan, J. G.; Nguyen, H. S.; Nguyen, S. H.; Synak, P.; Wroblewski, J. Rough Set Algorithms in Classification Problem. *Rough Set Methods and Appl.* **2000**, *56*, 49−88.
(31) Błaszczyński, J.; Greco, S.; Matarazzo, B.; Słowiński, R.; Szeląg, M. Jmaf - Dominance-Based Rough Set Data Analysis Framework. In *Rough Sets and Intelligent Systems - Professor Zdzisław Pawlak in Memoriam*, Skowron, A., Suraj, Z., Eds.; Springer: Berlin, Germany, 2013; Vol. *42*, Chapter 5, pp 185−209.
(32) Agrawal, R.; Srikant, R. Fast Algorithms for Mining Association Rules in Large Databases. In *VLDB'94*, Proceedings of 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile, September 12−15, 1994; Bocca, J. B., Jarke, M., Zaniolo, C., Eds.; Morgan Kauffmann: San Francisco, CA, 1994; pp 487−499.
(33) Tan, P.-N.; Steinbach, M.; Kumar, V. Association Analysis: Basic Concepts and Algorithms. In *Introduction to Data Mining*; Addison-Wesley: Boston, MA, 2005; Chapter 6, pp 327−414.
(34) Thabtah, F. A Review of Associative Classification Mining. *Knowl. Eng. Rev.* **2007**, *22*, 37−65.
(35) Angiulli, F.; Ianni, G.; Palopoli, L. On the Complexity of Inducing Categorical and Quantitative Association Rules. *Theor. Comput. Sci.* **2004**, *314*, 217−249.
(36) Bayardo, R. J. J. Efficiently Mining Long Patterns from Databases. *SIGMOD Rec.* **1998**, *27*, 85−93.

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

(37) Han, J. W.; Pei, J.; Yin, Y. W. Mining Frequent Patterns without Candidate Generation. *Sigmod Record* **2000**, *29*, 1−12.
(38) Zaki, M. J. Scalable Algorithms for Association Mining. *IEEE Trans. Knowl. Data Eng.* **2000**, *12*, 372−390.
(39) Pasquier, N.; Bastide, Y.; Taouil, R.; Lakhal, L. Efficient Mining of Association Rules Using Closed Itemset Lattices. *Inf. Syst.* **1999**, *24*, 25−46.
(40) Webb, G. I. Efficient Search for Association Rules. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Proceedings of KDD-2000. Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, 2000; Ramakrishnan, R., Stolfo, S., Eds.; ACM Press: New York, 2000; pp 99−107.
(41) Wolohan, P. R. N.; Akella, L. B.; Dorfman, R. J.; Nell, P. G.; Mundt, S. M.; Clark, R. D. Structural Unit Analysis Identifies Lead Series and Facilitates Scaffold Hopping in Combinatorial Chemistry. *J. Chem. Inf. Model.* **2006**, *46*, 1188−1193.
(42) Yu, P. L.; Wild, D. J. Fast Rule-Based Bioactivity Prediction Using Associative Classification Mining. *J. Cheminf.* **2012**, *4*, 29.
(43) Bayardo, R. J. J.; Agrawal, R. Mining the Most Interesting Rules. In *Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, CA, August 15−18, 1999; ACM Press: New York, 1999; pp 145−154.
(44) Borgelt, C.; Rodriguez, G. G. Frida - a Free Intelligent Data Analysis Toolbox. In *Fuzzy Systems Conference, FUZZ-IEEE 2007*, Proceedings of IEEE International Conference on Fuzzy Systems, London, England, July 23−26, 2007; IEEE: New York, 2007; pp 1897−1901.
(45) Webb, G. I. Opus: An Efficient Admissible Algorithm for Unordered Search. *J. Artif. Intell. Res.* **1995**, *3*, 431−465.
(46) Coenen, F.; Leng, P.; Ahmed, S. Data Structure for Association Rule Mining: T-Trees and P-Trees. *IEEE Trans. Knowl. Data Eng.* **2004**, *16*, 774−778.
(47) Li, J.; Dong, G.; Ramamohanarao, K. Making Use of the Most Expressive Jumping Emerging Patterns for Classification. *Knowl. Inf. Sys.* **2001**, *3*, 131−145.
(48) Fan, H.; Ramamohanarao, K. Fast Discovery and the Generalization of Strong Jumping Emerging Patterns for Building Compact and Accurate Classifiers. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 721−737.
(49) Garcia-Borroto, M.; Martinez-Trinidad, J.; Carrasco-Ochoa, J. A. A Survey of Emerging Patterns for Supervised Classification. *Artif. Intel. Rev.* **2014**, *42*, 705−721.
(50) Dong, G. Z.; Li, J. Y. Mining Border Descriptions of Emerging Patterns from Dataset Pairs. *Knowl. Inf. Sys.* **2005**, *8*, 178−202.
(51) Poezevara, G.; Cuissart, B.; Cremilleux, B. Extracting and Summarizing the Frequent Emerging Graph Patterns from a Dataset of Graphs. *J. Intell. Inf. Syst.* **2011**, *37*, 333−353.
(52) Wang, L. S.; Zhao, H.; Dong, G. Z.; Li, J. P. On the Complexity of Finding Emerging Patterns. *Theor. Comput. Sci.* **2005**, *335*, 15−27.
(53) Poezevara, G.; Cuissart, B.; Crémilleux, B. Discovering Emerging Graph Patterns from Chemicals. In *Foundations of Intelligent Systems*; Rauch, J., Raś, Z., Berka, P., Elomaa, T., Eds.; Springer: Berlin, Germany, 2009; Vol. *5722*, Chapter 8, pp 45−55.
(54) Soulet, A.; Klema, J.; Cremilleux, B. Efficient Mining under Rich Constraints Derived from Various Datasets. In *Knowledge Discovery in Inductive Databases*; Dzeroski, S.;,Struyf, J., Eds.; Springer: Berlin, Germany, 2007; Vol. *4747*, pp 223−239.
(55) Mitchell, T. M. Generalization as Search. *Artif. Intell.* **1982**, *18*, 203−226.
(56) Han, J. W.; Cai, Y. D.; Cercone, N. Knowledge Discovery in Databases - an Attribute-Oriented Approach. In *Very Large Data Bases: Vldb - 92*, Proceedings of the 18th International Conference on Very Large Data Bases, Vancouver, Canada, August 23−27, 1992; Yuan, L. Y., Ed.; Morgan Kaufmann: San Francisco, CA, pp 547−559.
(57) Bay, S. D.; Pazzani, M. J. Detecting Group Differences: Mining Contrast Sets. *Data Min. Knowl. Discovery* **2001**, *5*, 213−246.
(58) Wrobel, S. An Algorithm for Multi-Relational Discovery of Subgroups. In *PKDD-97, Principles of Data Mining and Knowledge Discovery*; Proceedings of first European symposium on principles of data mining and knowledge discovery, Trondheim, Norway, June 24−27, 1997; Komorowski, J.; Zytkow, J., Eds.; Springer-Verlag: Berlin, Germany, 1997; pp 78−87.
(59) Novak, P. K.; Lavrac, N.; Webb, G. I. Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. *J. Mach. Learn. Res.* **2009**, *10*, 377−403.
(60) Siu, K. K. W.; Butler, S. M.; Beveridge, T.; Gillam, J. E.; Hall, C. J.; Kaye, A. H.; Lewis, R. A.; Mannan, K.; McLoughlin, G.; Pearson, S.; Round, A. R.; Schtultke, E.; Webb, G. I.; Wilkinson, S. J. Identifying Markers of Pathology in Saxs Data of Malignant Tissues of the Brain. *Nucl. Instrum. Methods Phys. Res., Sect. A* **2004**, *548*, 140−146.
(61) Li, J. Y.; Wong, L. S. Identifying Good Diagnostic Gene Groups from Gene Expression Profiles Using the Concept of Emerging Patterns (Vol 18, Pg 725, 2002). *Bioinformatics* **2002**, *18*, 1406−1407.
(62) Kobylinski, L.; Walczak, K. Jumping Emerging Patterns with Occurrence Count in Image Classification. In *Lecture Notes in Artificial Intelligence*; Washio, T., Suzuki, E., Ting, K. M., Inokuchi, A., Eds.; Springer-Verlag: Berlin, Germany, 2008; Vol. *5012*, pp 904−909.
(63) Sherhod, R.; Gillet, V. J.; Judson, P. N.; Vessey, J. D. Automating Knowledge Discovery for Toxicity Prediction Using Jumping Emerging Pattern Mining. *J. Chem. Inf. Model.* **2012**, *52*, 3074−3087.
(64) Gamberger, D.; Lavrac, N.; Zelezny, F.; Tolar, J. Induction of Comprehensible Models for Gene Expression Datasets by Subgroup Discovery Methodology. *J. Biomed. Inf.* **2004**, *37*, 269−284.
(65) Lambach, D.; Gamberger, D. Temporal Analysis of Political Instability through Descriptive Subgroup Discovery. *Conflict Manag. Peace* **2008**, *25*, 19−32.
(66) Herrera, F.; Carmona, C. J.; Gonzalez, P.; del Jesus, M. J. An Overview on Subgroup Discovery: Foundations and Applications. *Knowl. Inf. Sys.* **2011**, *29*, 495−525.
(67) Demsar, J.; Curk, T.; Erjavec, A.; Gorup, C.; Hocevar, T.; Milutinovic, M.; Mozina, M.; Polajnar, M.; Toplak, M.; Staric, A.; Stajdohar, M.; Umek, L.; Zagar, L.; Zbontar, J.; Zitnik, M.; Zupan, B. Orange: Data Dining Toolbox in Python. *J. Mach. Learn. Res.* **2013**, *14*, 2349−2353.
(68) Atzmueller, M.; Puppe, F. Semi-Automatic Visual Subgroup Mining Using Vikamine. *J. Univers. Comput. Sci.* **2005**, *11*, 1752−1765.
(69) Alcala-Fdez, J.; Fernandez, A.; Luengo, J.; Derrac, J.; Garcia, S.; Sanchez, L.; Herrera, F. Keel Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. *J. Mult.-Valued Log. Soft. Comput.* **2011**, *17*, 255−287.
(70) Wille, R. Concept Lattices and Conceptual Knowledge Systems. *Comput. Math. Appl.* **1992**, *23*, 493−515.
(71) Ganter, B.; Wille, R. *Formal Concept Analysis: Mathematical Foundations*; Springer-Verlag: Berlin, Germany, 1998; Vol. *10*, pp 926−926.
(72) Poelmans, J.; Kuznetsov, S. O.; Ignatov, D. I.; Dedene, G. Formal Concept Analysis in Knowledge Processing: A Survey on Models and Techniques. *Expert Syst. Appl.* **2013**, *40*, 6601−6623.
(73) Poelmans, J.; Ignatov, D. I.; Kuznetsov, S. O.; Dedene, G. Formal Concept Analysis in Knowledge Processing: A Survey on Applications. *Expert Syst. Appl.* **2013**, *40*, 6538−6560.
(74) Priss, U. Formal Concept Analysis in Information Science. *Annu. Rev. Inf. Sci. Technol.* **2006**, *40*, 521−543.
(75) Berry, A.; Sigayret, A. Representing a Concept Lattice by a Graph. *Discrete Appl. Math.* **2004**, *144*, 27−42.
(76) Ganter, B.; Wille, R. Conceptual Scaling. In *Applications of Combinatorics and Graph Theory to the Biological and Social Sciences*; Roberts, F., Ed.; Springer-Verlag: New York, 1989; pp 139−167.
(77) Kuznetsov, S. Learning of Simple Conceptual Graphs from Positive and Negative Examples. In *Principles of Data Mining and Knowledge Discovery*; Żytkow, J.; Rauch, J., Eds.; Springer: Berlin, Germany, 1999; Vol. *1704*, Chapter 47, pp 384−391.
(78) Ganter, B.; Kuznetsov, S. O. Hypotheses and Version Spaces. In *Conceptual Structures for Knowledge Creation and Communication*;

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

DeMoor, A.; Lex, W.; Ganter, B., Eds.; Springer: Berlin, Germany, 2003; Vol. *2746*, pp 83−95.
(79) Kuznetsov, S. O. Pattern Structures for Analyzing Complex Data. In *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Proceedings*; Sakai, H., Chakraborty, M. K.,; Hassanien, A. E., Slezak, D., Zhu, W., Eds.; Springer: Berlin, Germany, 2009; Vol. *5908*, pp 33−44.
(80) Kaytoue, M.; Kuznetsov, S. O.; Napoli, A.; Duplessis, S. Mining Gene Expression Data with Pattern Structures in Formal Concept Analysis. *Inf. Sci.* **2011**, *181*, 1989−2001.
(81) Becker, P.; Correia, J. H. The Toscanaj Suite for Implementing Conceptual Information Systems. In *Formal Concept Analysis: Formal Concept Analysis*; Ganter, B.; Stumme, G., Wille, R., Eds.; Springer: Berlin, Germany, 2005; Vol. *3626*, pp 324−348.
(82) Valtchev, P.; Grosser, D.; Roume, M.; Rouane, M. Galicia: An Open Platform for Lattices. In *Using Conceptual Structures: 11th International Conference on Conceptual Structures (ICCS'03)*; Dresden, Germany, July 21−25, 2003; Shaker Verlag: pp 241−254.
(83) Delic, D.; Lenz, H. J.; Neiling, M. Rough Set's and Association Rules - Which Is Efficient? In *15th Biannual Conference on Computational Statistics (COMPSTAT)*; Compstat 2002: Proceedings in Computational Statistics, Berlin, Germany, Aug 24−28, 2002; Hardle, W., Ronz, B., Eds.; Physica-Verlag GMBH & Co: Heidelberg, Germany, 2002; pp 527−532.
(84) Webb, G. I.; Butler, S. M.; Newlands, D. On Detecting Differences between Groups. In *KDD '03, the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, August 24−27, 2003; ACM Press: New York, 2003; pp 256−265.
(85) Terlecki, P.; Walczak, K. On the Relation between Rough Set Reducts and Jumping Emerging Patterns. *Inf. Sci.* **2007**, *177*, 74−83.
(86) Ganter, B.; Kuznetsov, S. O. Formalizing Hypotheses with Concepts. In *8th International Conference on Conceptual Structures (ICCS 2000)*; Conceptual Structures: Logical, Linguistic, and Computational Issues, Proceedings, Darmstadt, Germany, August 14−18, 2000; Ganter, B., Mineau, G. W., Eds.; Springer-Verlag: Berlin, Germany, 2000; pp 342−356.
(87) Blinova, V. G.; Dobrynin, D. A.; Finn, V. K.; Kuznetsov, S. O.; Pankratova, E. S. Toxicology Analysis by Means of the Jsm-Method. *Bioinformatics* **2003**, *19*, 1201−1207.
(88) Yao, Y. A Comparative Study of Formal Concept Analysis and Rough Set Theory in Data Analysis. In *Rough Sets and Current Trends in Computing*; Tsumoto, S., Słowiński, R., Komorowski, J., Grzymała-Busse, J., Eds.; Springer: Berlin, Germany, 2004; Vol. *3066*, Chapter 6, pp 59−68.
(89) Zhang, W. X.; Wei, L.; Qi, J. J. Attribute Reduction in Concept Lattice Based on Discernibility Matrix. In *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*; Slezak, D.; Yao, J. T., Peters, J. F., Ziarko, W., Hu, X., Eds.; Springer: Berlin, Germany, 2005; Vol. *3642*, pp 157−165.
(90) Saquer, J.; Deogun, J. S. Formal Rough Concept Analysis. In *New Directions in Rough Sets, Data Mining, and Granular-Soft Computing*; Zhong, N., Skowron, A., Ohsuga, S., Eds.; Springer: Berlin, Germany, 1999; Vol. *1711*, pp 91−99.
(91) Stumme, G. *Conceptual Knowledge Discovery with Frequent Concept Lattices*; FB4-Preprint 2043; TU Darmstadt: Darmstadt, Germany, 1999.
(92) Pasquier, N.; Bastide, Y.; Lakhal, L. Closed Set Based Discovery of Small Covers for Association Rules (Extended Version). *Net. Inf. Syst.* **2001**, *3*, 349−377.
(93) Zaki, M. J.; Hsaio, C.-J. *Charm: An Efficient Algorithm for Closed Association Rule Mining*; Technical Report No. 99-10; Computer Science Department, Rensselaer Polytechnic Institute: Troy, NY, 1999.
(94) Krysinski, J. Rough Sets Approach to the Analysis of the Structure Activity Relationship of Quaternary Imidazolium Compounds. *Arzneim. Forsch./Drug Res.* **1990**, *40*−*2*, 795−799.
(95) Krysinski, J. Rough Sets Theory to Analysis of Relationship between Structure and Activity of Quaternary Quinolinium and Isoquinolinium Compounds. *Arch. Pharm.* **1991**, *324*, 827−832.
(96) Krysinski, J. Application of the Rough Sets Theory to the Analysis of Structure-Activity-Relationships of Antimicrobial Pyridinium Compounds. *Pharmazie* **1995**, *50*, 593−597.
(97) Krysinski, J.; Skrzypczak, A.; Demski, G.; Predki, B. Application of the Rough Set Theory in Structure Activity Relationships of Antielectrostatic Imidazolium Compounds. *Quant. Struct.-Act. Relat.* **2001**, *20*, 395−401.
(98) Krysinski, J.; Placzek, J.; Skrzypczak, A.; Blaszczak, J.; Predki, B. Analysis of Relationships between Structure, Surface Properties, and Antimicrobial Activity of Quaternary Ammonium Chlorides. *QSAR Comb. Sci.* **2009**, *28*, 995−1002.
(99) Palkowski, L.; Blaszczynski, J.; Skrzypczak, A.; Blaszczak, J.; Kozakowska, K.; Wroblewska, J.; Kozuszko, S.; Gospodarek, E.; Krysinski, J.; Slowinski, R. Antimicrobial Activity and Sar Study of New Gemini Imidazolium-Based Chlorides. *Chem. Biol. Drug Des.* **2014**, *83*, 278−288.
(100) Skrzypczak, A.; Brycki, B.; Mirska, I.; Pernak, J. Synthesis and Antimicrobial Activities of New Quats. *Eur. J. Med. Chem.* **1997**, *32*, 661−668.
(101) Hasegawa, K.; Koyama, M.; Arakawa, M.; Funatsu, K. Application of Data Mining to Quantitative Structure-Activity Relationship Using Rough Set Theory. *Chemom. Intell. Lab. Syst.* **2009**, *99*, 66−70.
(102) Liu, H.; Xiang, B. R.; Qu, L. B. The Application of Rough Sets in Sar Analysis of N1-Site Substituted Fluoroquinolones. *Chemom. Intell. Lab. Syst.* **2007**, *87*, 155−160.
(103) Liu, H.; Qu, L. B.; Gao, H. B.; Wang, J. X.; Han, L. P.; Xiang, B. R. Study on the Quantitative Structure-Activity Relationship of C-10 Substituted Artemisinin (Qhs)'S Derivatives Using Rough Set Theory. *Sci. China, Ser. B: Chem.* **2008**, *51*, 937−945.
(104) Wang, T.; Dong, Y.; Wang, L. C.; Xiang, B. R.; Chen, Z.; Qu, L. B. Design, Synthesis and Structure-Activity Relationship Studies of 6-Phenyl-4,5-Dihydro-3(2h)-Pyridazinone Derivatives as Cardiotonic Agents. *Arzneim. Forsch.* **2008**, *58*, 569−573.
(105) van Westen, G. J. P.; Wegner, J. K.; Ijzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Proteochemometric Modeling as a Tool to Design Selective Compounds and for Extrapolating to Novel Targets. *MedChemComm* **2011**, *2*, 16−30.
(106) Strombergsson, H.; Prusis, P.; Midelfart, H.; Lapinsh, M.; Wikberg, J. E. S.; Komorowski, J. Rough Set-Based Proteochemometrics Modeling of G-Protein-Coupled Receptor-Ligand Interactions. *Proteins: Struct., Funct., Genet.* **2006**, *63*, 24−34.
(107) Strombergsson, H.; Kryshtafovych, A.; Prusis, P.; Fidelis, K.; Wikberg, J. E. S.; Komorowski, J.; Hvidsten, T. R. Generalized Modeling of Enzyme-Ligand Interactions Using Proteochemometrics and Local Protein Substructures. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 568−579.
(108) Kierczak, M.; Ginalski, K.; Draminski, M.; Koronacki, J.; Rudnicki, W.; Komorowski, J. A Rough Set-Based Model of Hiv-1 Reverse Transcriptase Resistome. *Bioinf. Biol. Insights* **2009**, *3*, 109−127.
(109) He, Y. J.; Chen, D. Z.; Zhao, W. X. Integrated Method of Compromise-Based Ant Colony Algorithm and Rough Set Theory and Its Application in Toxicity Mechanism Classification. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 22−32.
(110) Maji, P.; Paul, S. Rough Sets for Selection of Molecular Descriptors to Predict Biological Activity of Molecules. *IEEE Trans. Syst. man Cybern. Part C- Appl. Rev.* **2010**, *40*, 639−648.
(111) Liu, B.; Hsu, W.; Ma, Y. Integrating Classification and Association Rule Mining. In *Proceedings of the 4th conference on Knowledge Discovery and Data Mining*; Proceedings of the 4th conference on Knowledge Discovery and Data Mining, New York, August 27−31, 1998; Agrawal, R., Stolorz, P., Eds.; AAAI Press: Palo Alto, CA, 1998; pp 80−86.
(112) Yin, X. X.; Han, J. W. Cpar: Classification Based on Predictive Association Rules. In *SDM 2003,Third Siam International Conference on*

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

Data Mining; Proceedings of the Third Siam International Conference on Data Mining, San Francisco, CA, May 1−3, 2003; Barbara, D., Kamath, C., Eds.; SIAM: Philadelpia, PA, 2003; pp 331−335.
(113) Li, W.; Han, J.; Pei, J. Cmar: Accurate and Efficient Classification Based on Multiple Class-Association Rules. In *ICDM 2001*; Proceedings IEEE International Conference on Data Mining, San Jose, CA, November 29−December 2, 2001; Cercone, N., Lin, T. Y., Wi, X. D., Eds.; Springer Verlag: New York, 2001; pp 369−376.
(114) Wang, W.; Yang, J.; Yu, P. War: Weighted Association Rules for Item Intensities. *Knowl. Inf. Sys.* **2004**, *6*, 203−229.
(115) Tao, F.; Murtagh, F.; Farid, M. Weighted Association Rule Mining Using Weighted Support and Significance Framework. In *SIGKDD 2003*; Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, August 24−27, 2003; ACM Press: New York, pp 661−666.
(116) Yu, P.; Wild, D. J. Discovering Associations in Biomedical Datasets by Link-Based Associative Classifier (Lac). *PLoS One* **2012**, *7*, e51018.
(117) Page, L.; Brin, S.; Motwani, R.; Winograd, T. *The Pagerank Citation Ranking: Bringing Order to the Web*; Technical Report; Stanford InfoLab: Stanford, CA, 1999.
(118) Kleinberg, J. M. Authoritative Sources in a Hyperlinked Environment. *J. Assoc. Comput. Mach.* **1999**, *46*, 604−632.
(119) Raymond, J. W.; Watson, I. A.; Mahoui, A. Rationalizing Lead Optimization by Associating Quantitative Relevance with Molecular Structure Modification. *J. Chem. Inf. Model.* **2009**, *49*, 1952−1962.
(120) Hahsler, M.; Hornik, K. New Probabilistic Interest Measures for Association Rules. *Intell. Data Anal.* **2007**, *11*, 437−455.
(121) Omiecinski, E. R. Alternative Interest Measures for Mining Associations in Databases. *IEEE Trans. Knowl. Data Eng.* **2003**, *15*, 57−69.
(122) Abiteboul, S.; Hull, R.; Vianu, V. *Foundations of Databases*; Addison Wesley: Boston MA, 1995; p 305.
(123) Dehaspe, L.; Toivonen, H.; King, R. D. Finding Frequent Substructures in Chemical Compounds. In *Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*; New York, August 27−31, 1998; Agrawal, R., Stolorz, P., Piatetsky-Shapiro, G., Eds.; AAAI Press: pp 30−36.
(124) King, R. D.; Srinivasan, A.; Dehaspe, L. Warmr: A Data Mining Tool for Chemical Data. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 173−181.
(125) Avidon, V. V.; Pomerantsev, I. A.; Golender, V. E.; Rozenblit, A. B. Structure-Activity Relationship Oriented Languages for Chemical-Structure Representation. *J. Chem. Inf. Model.* **1982**, *22*, 207−214.
(126) Helma, C.; Kramer, S. A Survey of the Predictive Toxicology Challenge 2000−2001. *Bioinformatics* **2003**, *19*, 1179−1182.
(127) Toivonen, H.; Srinivasan, A.; King, R. D.; Kramer, S.; Helma, C. Statistical Evaluation of the Predictive Toxicology Challenge 2000−2001. *Bioinformatics* **2003**, *19*, 1183−1193.
(128) Auer, J.; Bajorath, J. Emerging Chemical Patterns: A New Methodology for Molecular Classification and Compound Selection. *J. Chem. Inf. Model.* **2006**, *46*, 2502−2514.
(129) Bailey, J.; Manoukian, T.; Ramamohanarao, K. Fast Algorithm for Computing Hypergraph Transversals and Its Application in Mining Emerging Patterns. In *3rd IEEE International Conference on Mining (ICDM 2003)*; Melbourne, FL, November 19−22, 2003; IEEE Computer Society: pp 485−488.
(130) Namasivayam, V.; Hu, Y.; Balfer, J.; Bajorath, J. Classification of Compounds with Distinct or Overlapping Multi-Target Activities and Diverse Molecular Mechanisms Using Emerging Chemical Patterns. *J. Chem. Inf. Model.* **2013**, *53*, 1272−1281.
(131) Namasivayam, V.; Iyer, P.; Bajorath, J. Prediction of Individual Compounds Forming Activity Cliffs Using Emerging Chemical Patterns. *J. Chem. Inf. Model.* **2013**, *53*, 3131−3139.
(132) Namasivayam, V.; Gupta-Ostermann, D.; Balfer, J.; Heikamp, K.; Bajorath, J. Prediction of Compounds in Different Local Structure-Activity Relationship Environments Using Emerging Chemical Patterns. *J. Chem. Inf. Model.* **2014**, *54*, 1301−1310.
(133) Auer, J.; Bajorath, J. Distinguishing between Bioactive and Modeled Compound Conformations through Mining of Emerging Chemical Patterns. *J. Chem. Inf. Model.* **2008**, *48*, 1747−1753.
(134) Takigawa, I.; Mamitsuka, H. Graph Mining: Procedure, Application to Drug Discovery and Recent Advances. *Drug Discovery Today* **2013**, *18*, 50−57.
(135) Sherhod, R.; Judson, P. N.; Hanser, T.; Vessey, J. D.; Webb, S. J.; Gillet, V. J. Emerging Pattern Mining to Aid Toxicological Knowledge Discovery. *J. Chem. Inf. Model.* **2014**, *54*, 1864−1879.
(136) Fischer, I.; Meinl, T. Graph Based Molecular Data Mining - an Overview. In *2004 Ieee International Conference on Systems, Man & Cybernetics*, Vols. *1*−*7*; IEEE Operations Center: Piscataway, NJ, 2004; pp 4578−4582.
(137) Klopman, G. Artificial-Intelligence Approach to Structure Activity Studies - Computer Automated Structure Evaluation of Biological-Activity of Organic-Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321.
(138) Klopman, G. A Hierarchical Computer Automated Structure Evaluation Program. *Quant. Struct.-Act. Relat.* **1992**, *11*, 176−184.
(139) Borgelt, C.; Meinl, T.; Berthold, M. R. In *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*; ACM: Chicago, IL, 2005; pp 6−15.
(140) Ting, R.; Bailey, J. Mining Minimal Contrast Subgraph Patterns. In *6th SIAM International Conference on Data Mining*; Proceedings of the sixth SIAM International Conference on Data Mining, Bethesda, MD, April 20−22, 2006; Ghosh, J., Lambert, D., Skillicorn, D. B., Srivastava, J., Eds.; SIAM: Philadelphia, PA, 2006; pp 638−642.
(141) Holder, L. B.; Cook, D. J.; Djoko, S. Substructure Discovery in the Subdue System. In *In Proc. of the AAAI Workshop on Knowledge Discovery in Databases*; AAAI Press: Menlo Park, CA, 1994; pp 169−180.
(142) Cook, D. J.; Holder, L. B. Graph-Based Data Mining. *IEEE Intell. Syst. Appl.* **2000**, *15*, 32−41.
(143) Karunaratne, T.; Bostrom, H. Using Background Knowledge for Graph Based Learning: A Case Study in Chemoinformatics. In *IMECS 2007: International Multiconference of Engineers and Computer Scientists*; Proceedings of International Multiconference of Engineers and Computer Scientists, Vols I and II, Kowloon, China, March 12−14, 2007; International Association Engineers: Hong Kong; 2007; pp 153−157.
(144) Shelokar, P.; Quirin, A.; Cordon, O. Mosubdue: A Pareto Dominance-Based Multiobjective Subdue Algorithm for Frequent Subgraph Mining. *Knowl. Inf. Sys.* **2013**, *34*, 75−108.
(145) McGregor, J. J. Backtrack Search Algorithms and the Maximal Common Subgraph Problem. *Software-Prac. Exp.* **1982**, *12*, 23−34.
(146) Yan, H.; Han, J. Gspan: Graph-Based Substructure Pattern Mining. In *2nd IEEE International Conference on Data Mining*; Proceedings of the 2002 IEEE International Conference on Data Mining, Maebashi City, Japan, 2002; Kumar, V., Tsumoto, S., Zhong, N., Yu, P. S., Wu, X. D., Eds.; IEEE Computer Society: Los Alamitos, CA, 2002; pp 721−724.
(147) Dominik, A.; Walczak, Z.; Wojciechowski, J. Efficient Algorithm to Mine Association Rules from Spatial Data. In *Adaptive and Natural Computing Algorithms, Pt 1*; Beliczynski, B., Dzielinski, A., Iwanowski, M., Ribeiro, B., Eds.; Springer: Berlin, Germany, 2007; Vol. *4431*, pp 772−781.
(148) Dominik, A.; Walczak, Z.; Wojciechowski, J. Prediction of Chemical-Protein Binding Activity Using Contrast Graph Patterns. In *Software Tools and Algorithms for Biological Systems*; Arabnia, H. R., Tran, Q. N., Eds.; Springer-Verlag: New York, 2011; Vol. *696*, pp 243−253.
(149) Lozano, S.; Poezevara, G.; Halm-Lemeille, M.-P.; Lescot-Fontaine, E.; Lepailleur, A.; Bissell-Siders, R.; Cremilleux, B.; Rault, S.; Cuissart, B.; Bureau, R. Introduction of Jumping Fragments in Combination with Qsars for the Assessment of Classification in Ecotoxicology. *J. Chem. Inf. Model.* **2010**, *50*, 1330−1339.

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803

(150) Inokuchi, A.; Washio, T.; Motoda, H. An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data. In *Principles of Data Mining and Knowledge Discovery*; Zighed, D., Komorowski, J., Żytkow, J., Eds.; Springer Berlin Heidelberg: Berlin, Germany, 2000; Vol. *1910*, Chapter 2, pp 13−23.
(151) Borgelt, C.; Berthold, M. R. Mining Molecular Fragments: Finding Relevant Substructures of Molecules. In *IEEE International Conference on Data Mining, 2002*; Proceedings. 2002 IEEE International Conference on, Maebashi City, Japan, December 9−12, 2002; IEEE: pp 51−58.
(152) Zaki, M. J.; Parthasarathy, S.; Ogihara, M.; Li, W. New Algorithms for Fast Discovery of Association Rules. In *Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Newport Beach, CA, August 14−17, 1997; AAAI Press: Menlo Park, CA, 1997; pp 283−286.
(153) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. Recap - Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Model.* **1998**, *38*, 511−522.
(154) Kuramochi, M.; Karypis, G. Frequent Subgraph Discovery. In *Proceedings of the 2001 IEEE International Conference on Data Mining*; San Jose, CA, November 29−December 2, 2001; IEEE Computer Society: Washington, DC, pp 313−320.
(155) Lam, W. W. M.; Chan, K. C. C. Discovering Interesting Molecular Substructures for Molecular Classification. *IEEE Trans. Nanobiosci.* **2010**, *9*, 77−89.
(156) Thoma, M.; Cheng, H.; Gretton, A.; Han, J.; Kriegel, H.-P.; Smola, A.; Song, L.; Yu, P. S.; Yan, X.; Borgwardt, K. M. Discriminative Frequent Subgraph Mining with Optimality Guarantees. *Stat. Anal. Data Min.* **2010**, *3*, 302−318.
(157) Thoma, M.; Cheng, H.; Gretton, A.; Han, J.; Kriegel, H.-P.; Smola, A. J.; Song, L.; Yu, P. S.; Yan, X.; Borgwardt, K. Near-Optimal Supervised Feature Selection among Frequent Subgraphs. In *9th SIAM Conference on Data Mining (SDM 2009)*; Sparks, NV, April 30−May 2, 2009; SIAM: Philadelphia, PA, 2009; pp 1076−1087.
(158) Huan, J.; Wang, W.; Prins, J. Efficient Mining of Frequent Subgraphs in the Presence of Isomorphism. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, Melbourne, FL, November 19 - 22, 2003; IEEE Press: Los Alamitos, CA, pp 549−552.
(159) Fei, H.; Huan, J. Structure Feature Selection for Chemical Compound Classification. In *8th IEEE International Conference on Bioinformatics and Bioengineering (BIBE 2008)*; Proceedings of 8th IEEE International Conference on Bioinformatics and Bioengineering, Athens, Greece, 2008; IEEE: New York, 2008; pp 33−38.
(160) Fei, H.; Huan, J. Structure Feature Selection for Chemical Compound Classification. In *ACM 17th Conference on Information and Knowledge Management (CIKM)*; Proceedings of the ACM 17th Conference on Information and Knowledge Management (CIKM), Nappa Valley, CA, October 26−28, 2008; ACM Press: New York, 2008.
(161) Khashan, R.; Zheng, W.; Tropsha, A. Thedevelopment of Novel Chemical Fragment- Based Descriptors Using Frequent Common Subgraph Mining Approach and Their Applicationin Qsar Modeling. *Mol. Inf.* **2014**, *33*, 201−215.
(162) Nijssen, S.; Kok, J. N. Frequent Graph Mining and Its Application to Molecular Databases. In *IEEE International Conference on Systems, Man, and Cybernetics*; The Hague, Netherlands, October 10−13, 2004; IEEE Press: Piscataway, NJ, pp 4571−4577.
(163) Kazius, J.; Nijssen, S.; Kok, J.; Back, T.; Ijzerman, A. P. Substructure Mining Using Elaborate Chemical Representation. *J. Chem. Inf. Model.* **2006**, *46*, 597−605.
(164) van der Horst, E.; Okuno, Y.; Bender, A.; Ijzerman, A. P. Substructure Mining of Gpcr Ligands Reveals Activity-Class Specific Functional Groups in an Unbiased Manner. *J. Chem. Inf. Model.* **2009**, *49*, 348−360.
(165) Lu, J.; Zheng, M.; Wang, Y.; Shen, Q.; Luo, X.; Jiang, H.; Chen, K. Fragment-Based Prediction of Skin Sensitization Using Recursive Partitioning. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 885−893.
(166) Wang, Y.; Lu, J.; Wang, F.; Shen, Q.; Zheng, M.; Luo, X.; Zhu, W.; Jiang, H.; Chen, K. Estimation of Carcinogenicity Using Molecular Fragments Tree. *J. Chem. Inf. Model.* **2012**, *52*, 1994−2003.
(167) Wörlein, M.; Meinl, T.; Fischer, I.; Philippsen, M. A Quantitative Comparison of the Subgraph Miners Mofa, Gspan, Ffsm, and Gaston. In *Knowledge Discovery in Databases: Pkdd 2005*; Jorge, A., Torgo, L., Brazdil, P., Camacho, R., Gama, J., Eds.; Springer: Berlin, Germany, 2005; Vol. *3721*, Chapter 39, pp 392−403.
(168) Deshpande, M.; Kuramochi, M.; Wale, N.; Karypis, G. Frequent Substructure-Based Approaches for Classifying Chemical Compounds. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 1036−1050.
(169) Deshpande, M.; Kuramochi, M.; Karypis, G. Data Mining Algorithms for Virtual Screening of Bioactive Compounds. In *Data Mining in Biomedicine*; Pardalos, P. M., Boginski, V. L., Vazacopoulos, A., Eds.; Springer: New York, 2007; Vol. *7*, pp 59−90.
(170) Kuramochi, M.; Karypis, G. An Efficient Algorithm for Discovering Frequent Subgraphs. *IEEE Trans. Knowl. Data Eng.* **2004**, *16*, 1038−1051.
(171) Hu, K.; Lu, Y.; Zhou, L.; Shi, C. Integrating Classification and Association Rule Mining: A Concept Lattice Framework. In *New Directions in Rough Sets, Data Mining, and Granular-Soft Computing*; Zhong, N., Skowron, A., Ohsuga, S., Eds.; Springer: Berlin, Germany, 1999; Vol. *1711*, Chapter 53, pp 443−447.
(172) Bruggemann, R.; Bartel, H. G. A Theoretical Concept to Rank Environmentally Significant Chemicals. *J. Chem. Inf. Model.* **1999**, *39*, 211−217.
(173) Pudenz, S.; Bruggemann, R.; Bartel, H. G. Qsar of Ecotoxicological Data on the Basis of Data-Driven If-Then-Rules. *Ecotoxicology* **2002**, *11*, 337−342.
(174) Carlsen, L. Giving Molecules an Identity. On the Interplay between Qsars and Partial Order Ranking. *Molecules* **2004**, *9*, 1010−1018.
(175) Carlsen, L. A Combined Qsar and Partial Order Ranking Approach to Risk Assessment. *SAR QSAR Environ. Res.* **2006**, *17*, 133−146.
(176) Restrepo, G.; Basak, S. C.; Mills, D. Comparison of Qsars and Characterization of Structural Basis of Bioactivity Using Partial Order Theory and Formal Concept Analysis: A Case Study with Mutagenicity. *Curr. Comput.-Aided Drug Des.* **2011**, *7*, 109−121.
(177) Lounkine, E.; Auer, J.; Bajorath, J. Formal Concept Analysis for the Identification of Molecular Fragment Combinations Specific for Active and Highly Potent Compounds. *J. Med. Chem.* **2008**, *51*, 5342−5348.
(178) Krueger, F.; Lounkine, E.; Bajorath, J. Fragment Formal Concept Analysis Accurately Classifies Compounds with Closely Related Biological Activities. *ChemMedChem* **2009**, *4*, 1174−1181.
(179) Lounkine, E.; Stumpfe, D.; Bajorath, J. Molecular Formal Concept Analysis for Compound Selectivity Profiling in Biologically Annotated Databases. *J. Chem. Inf. Model.* **2009**, *49*, 1359−1368.
(180) Jullian, N.; Afshar, M. Novel Rule-Based Method for Multi-Parametric Multi-Objective Decision Support in Lead Optimization Using Kem. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 35−45.
(181) Pasquier, N.; Bastide, Y.; Taouil, R.; Lakhal, L. Discovering Frequent Closed Itemsets for Association Rules. In *Database Theory - ICDT'99*; Jerusalem, Israel, January 10−12, 1999; Beeri, C., Buneman, P., Eds.; Springer: Berlin, Germany, pp 398−416.
(182) Stumpfe, D.; Lounkine, E.; Bajorath, J. Molecular Test Systems for Computational Selectivity Studies and Systematic Analysis of Compound Selectivity Profiles. In *Chemoinformatics and Computational Chemical Biology*; Bajorath, J., Ed.; Humana Press, Inc: Totowa, NJ, 2011; Vol. *672*, pp 503−515.

DOI: 10.1021/acs.jcim.5b00198
J. Chem. Inf. Model. 2015, 55, 1781−1803