Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

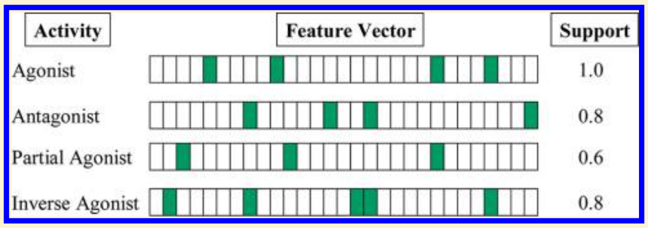
Classification of Compounds with Distinct or Overlapping Multi-Target Activities and Diverse Molecular Mechanisms Using Emerging Chemical Patterns

Vigneshwaran Namasivayam, Ye Hu, Jenny Balfer, and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, D-53113 Bonn, Germany

Supporting Information

ABSTRACT: The emerging chemical patterns (ECP) approach has been introduced for compound classification. Thus far, only very few ECP applications have been reported. Here, we further investigate the ECP methodology by studying complex classification problems. The analysis involves multi-target data sets with systematically organized subsets of compounds having distinct or overlapping target activities and, in addition, data sets containing classes of specifically active compounds with different mechanism-of-action. In systematic classification trials focusing on individual compound subsets or mechanistic classes, ECP calculations utilizing numerical descriptors achieve moderate to high sensitivity, dependent on the data set, and consistently high specificity. Accurate ECP predictions are already obtained on the basis of very small learning sets with only three positive training instances, which distinguishes the ECP approach from many other machine learning techniques.



1. INTRODUCTION

The classification of compounds according to their biological activity or other molecular properties is a central topic in chemoinformatics.[1,2] For compound classification, different machine learning approaches have become increasingly popular over the past decade[2] including, among others, decision trees,[3] Bayesian classifiers,[4] or support vector machines.[5,6] Supervised machine learning methods require the availability of sufficiently large training data sets to derive predictive models, typically tens to hundreds of known active compounds (and comparable or larger numbers of negative/inactive training instances). Data sparseness often complicates the search for novel active molecules, especially if new targets are considered for which often only limited (or no) compound information is available. We have been interested in investigating machine learning approaches for compound classification that are capable of operating on the basis of limited compound information. In this context, the concept of *emerging patterns* has become attractive to us that originated in computer science and was introduced by Dong and colleagues.[7−12] As detailed in the following section, the emerging patterns approach systematically generates feature patterns for objects with different class labels and identifies characteristic features/patterns that appear with high frequency in one class but not in the other. In 2002, this approach was applied to aid in the analysis of gene expression patterns in bioinformatics,[13] and in 2006, we adopted the methodology as *emerging chemical patterns* (ECP) for chemoinformatics by investigating molecular feature patterns.[14] In our initial studies, ECP classification was successfully applied to distinguish active compounds according to different potency levels[14] and simulate sequential screening experiments involving potency enrichment of active compounds in database selection sets.[15] In these investigations, very small numbers of three, five, or 10 positive training examples (active reference compounds) were found to be sufficient for the derivation of ECP classifiers. For small training sets, the ECP approach outperformed decision trees and Bayesian QSAR methods.[14] In addition, we also utilized ECP calculations to identify chemical descriptors and value ranges that discriminated between experimentally observed bioactive conformations of ligands and modeled conformations.[16] Recently, the methodology

Table 1. Emerging Patterns[a]

		attributes					
class	compound	d1	d2	d3	d4	d5	d6
A	**1**		X	X			X
	2	X				X	
	3		X	X	X		X
	4	X	X	X		X	
	5		X	X			X
B	**6**	X			X		X
	7		X	X		X	
	8			X		X	X
	9	X					X

[a] Two hypothetical classes A and B are shown that consist of five and four compounds, respectively. The presence of each of the six attributes in a compound is indicated by "X". The attribute distribution over these compounds yields a variety of emerging patterns, as discussed in the text.

Received: March 28, 2013
Published: May 21, 2013

Table 2. Compound Data Sets[a]

multi-target sets			
designation	target name	assay ID	no. compounds
MT_A	replicative DNA helicase	449750	1117
	euchromatic histone-lysine N-methyltransferase 2	504332	
	thioredoxin reductase	588453	
MT_B	streptokinase A precursor	1914	1450
	nuclear receptor ROR-gamma	2551	
	bromodomain adjacent to zinc finger domain 2B	504333	
MT_C	RAR-related orphan receptor gamma	2546	1301
	replicative DNA helicase	449750	
	thioredoxin glutathione reductase	485364	
MT_D	posterior segregation family member (pos-1)	1964	1037
	DNA polymerase beta	485314	
	thioredoxin glutathione reductase	485364	

mechanism-based sets		
designation	target name	no. compounds
M_AA1	adenosine A1 receptor	307
M_H3R	histamine H3 receptor	213
M_AM1	muscarinic acetylcholine receptor M1	148

[a] For multi-target sets, the set designation, target names, corresponding PubChem assay IDs, and total number of compounds tested in all three assays are reported. For mechanism-based sets, the set designation, target name, and total number of compounds are given.

Table 3. Compound Distribution over Subsets[a]

multi-target sets		
designation	subsets	no. compounds
MT_A	0	276
	1	340
	2	72
	3	27
	12	221
	13	65
	23	37
	123	79
MT_B	0	536
	1	284
	2	291
	3	74
	12	77
	13	44
	23	114
	123	30
MT_C	0	360
	1	38
	2	505
	3	40
	12	77
	13	60
	23	137
	123	84
MT_D	0	498
	1	276
	2	40
	3	59
	12	31
	13	56
	23	26
	123	51

[a] For each subset of a multi-target set, the number of compounds is reported.

Table 4. Compound Distribution over Mechanistic Classes[a]

mechanism-based sets		
designation	classes	no. compounds
M_AA1	agonist	107
	antagonist	94
	inverse agonist	52
	partial agonist	54
M_H3R	agonist	44
	antagonist	92
	inverse agonist	31
	partial agonist	46
M_AM1	agonist	26
	antagonist	73
	inverse agonist	0
	partial agonist	49

[a] For each class of a mechanism-based set, the number of compounds is reported.

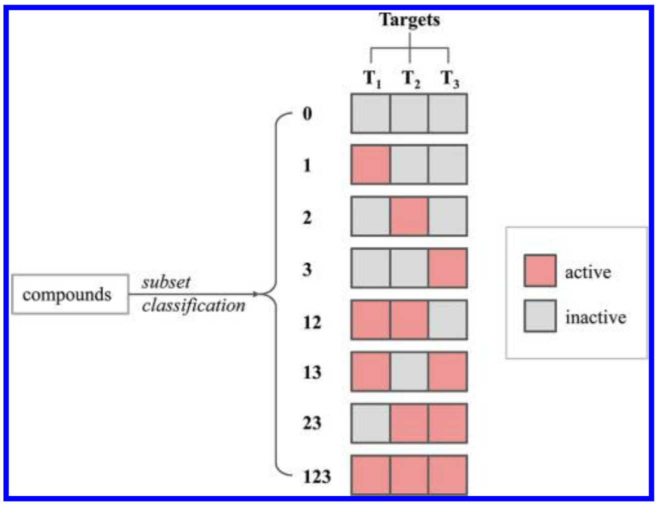


Figure 1. Compound subset classification for multi-target sets. For multi-target sets, compounds were organized into different subsets according to their activity profiles. For example, compounds that were inactive against all targets were assigned to subset "0", compounds only active against target T_1 to subset "1", compounds active against targets T_1 and T_2 to subset "12", and compounds active against all three targets to subset "123".

was also applied by Sherhod et al. to identify structural patterns that are indicative of various toxic effects of small molecules.[17] Apart from these few studies, the ECP approach has remained largely unexplored in chemoinformatics.

Herein, we further investigate the utility of the ECP approach by focusing on rather complex compound classification tasks involving diverse mechanism-of-action of G protein coupled receptor (GPCR) ligands and, in addition, compound data sets with distinct or overlapping target (activity) profiles. In the following, we briefly introduce key aspects of the emerging patterns and ECP methodology, detail the composition of the compound data sets, specify classification problems, and report the results of our analysis. Taken together, the findings reported herein provide further evidence for the utility of the ECP approach to address complex compound classification tasks and search for novel

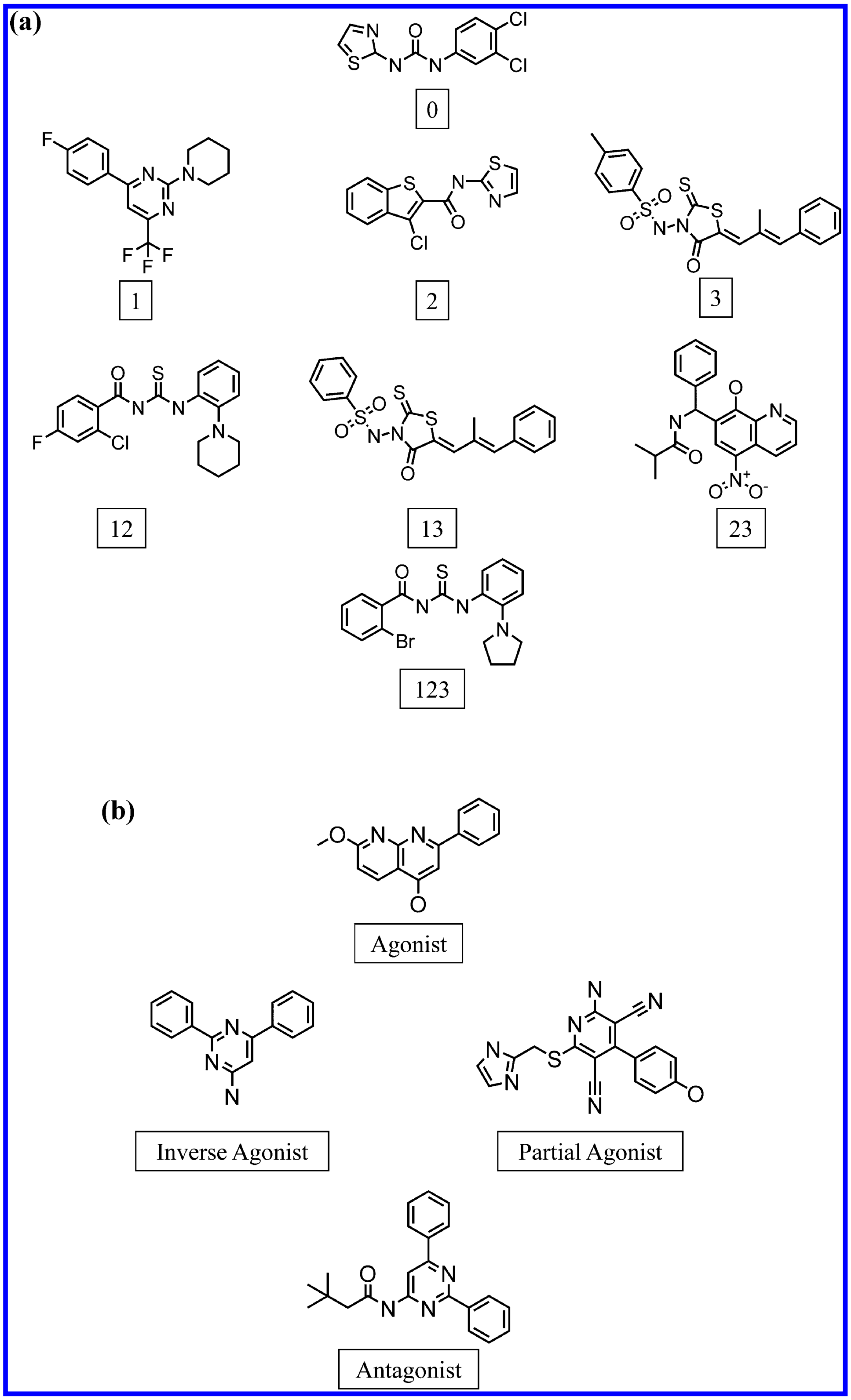


Figure 2. Representative compounds. Shown are representative compounds for (a) each subset in multi-target set MT_B and (b) each class in mechanism-based set M_AA1.

active compounds on the basis of very little prior ligand information, which renders the methodology attractive for practical applications.

2. THEORY

Basic input data for building an ECP classifier consist of descriptor values of learning set compounds. Continuous descriptor value ranges must be *discretized* into intervals, as further described below. A test compound yields a set of *attribute−value pairs*. Each pair (*item*) specifies the name and value of a descriptor. A subset of all possible attribute−value pairs is then considered an *item set* or a *pattern*. For a discretized descriptor, the value of a corresponding attribute−value pair is the numerical interval into which the descriptor value of the compound falls. The relative frequency of a pattern *p* in a compound learning set D represents the *support* of p in D, i.e., supp_D (*p*) according to eq 1

$$\text{supp}_D(\text{p}) = \frac{\text{count}_D(\text{p})}{|D|} \tag{1}$$

Here, $\text{count}_D(\text{p})$ is the number of instances in set D containing p. Patterns with significant support in positive training instances compared to negative instances are termed *emerging patterns* (EPs).[7,8] The ratio of support rates of an EP in positive (D_1) and negative (D_2) training examples is calculated as its $\text{growth}_{D_1,D_2}(\text{p})$ according to eq 2

$$\text{growth}_{D1,D2}(\text{p}) = \frac{\text{supp}_{D1}(\text{p})}{\text{supp}_{D2}(\text{p})} \tag{2}$$

If the support is greater than zero in D_1 and zero in D_2, the EP is termed a *jumping emerging pattern* (JEP).[9] In this case, the growth is not defined, but JEPs typically include the most discriminatory patterns. The definition of JEPs is further refined by introducing an additional condition. A JEP is considered *most expressive* if none of its descriptor subsets are JEPs and if no superset has a larger support in the data set.[9] Hence, most expressive JEPs are typically determined as the most discriminatory patterns for compound classification. Table 1 reports a hypothetical data set example with values of attributes {d1, d2, ... d6} for nine compounds divided into classes A and B. The (d2, d3) pattern has a support of 4/5 in class A and of 1/4 in class B. The growth rate of (d2, d3) is 3.2 for class A, and the pattern is thus classified as an EP. In addition, the pattern (d1, d5) is a JEP for class A with a support of 2/5 in class A and of zero in class B. Furthermore, the pattern (d1, d6) is a JEP for class B with a support of 2/4 in B and of zero in A. Moreover, the JEP (d1, d4, d6) only occurs in class B with a support of 1/4 and is a superset of EP (d1, d6) with an additional attribute d4. This means that (d1, d6) is a most expressive JEP.

3. EMERGING CHEMICAL PATTERNS ANALYSIS

For compound classification, ECP were defined as the most expressive JEPs identified on the basis of molecular descriptor analysis.[14] ECP are systematically calculated for a learning set. To classify a test compound, its descriptor values are calculated and all ECP derived from the learning set are identified that are matched by the test compound. Mining of ECP is facilitated by a hypergraph-based algorithm.[10,14] For ECP from negative and positive training data, support is separately accumulated, and the compound is assigned to the class yielding the largest support. Accumulated support is normalized with respect to all training set compounds to yield scores falling into the value range [0,1].

A set of 61 numerical descriptors derived from molecular graphs and available in the Molecular Operating Environment[18] were used for the generation of ECP classifiers. These descriptors belonged to a variety of categories and were previously selected because they displayed low pairwise correlation and had high information content in a large compound collection.[19] All 61 descriptors are defined in Table S1 of the Supporting Information. The descriptors were discretized with an algorithm utilizing an attribute splitting criterion on the basis of class information entropy of value range partitions.[20,21] This algorithm does not produce evenly divided value range intervals but optimizes the information entropy associated with partitions and was previously shown to be a preferred discretization approach for ECP-based classification.[14]

4. DATA SETS AND CLASSIFICATION TASKS

Two different categories of multi-subset or multi-class data sets were studied. The first category contained four sets of compounds with all possible combinations of activities against three different targets each, also including inactive compounds, as reported in Table 2. These multi-target data sets were assembled from PubChem Confirmatory Bioassays[22] and contained between 1037 and 1450 compounds. Each multi-target set covered different combinations of targets including long-established enzyme targets (e.g., thioredoxin reductase) and also novel targets (e.g., pos-1). The common compound subset organization shared by these data sets is illustrated in Figure 1 (formal subset designations are consistently used throughout this study). For each data set, eight different subsets were obtained. The number of compounds per subset is reported in Table 3. Because these compounds originated from high-throughput screens and were assembled from confirmatory bioassays, confirmed inactive compounds were also available in each case. For data sets MT_B and MT_D, inactive compounds represented the largest subset with 536 and 498 compounds, respectively. In subsets containing compounds

Table 5. Discretized Descriptors[a]

set	designation	no. discretized descriptors
multi-target sets	MT_A	19
	MT_B	24
	MT_C	18
	MT_D	10
mechanism-based sets	M_AA1	52
	M_H3R	45
	M_AM1	24

[a] For each compound set, the number of descriptors qualifying for ECP analysis following information entropy-based discretization is reported.

Table 6. Emerging Chemical Patterns for Multi-Target Sets[a]

reference compounds = 3				
subsets	MT_A	MT_B	MT_C	MT_D
0	228	799	58	14
1	196	836	60	16
2	208	855	60	17
3	200	827	66	11
12	191	781	52	23
13	133	827	56	16
23	163	848	51	15
123	187	872	63	16
total	1506	6645	466	128
reference compounds = 10				
subsets	MT_A	MT_B	MT_C	MT_D
0	772	4866	182	26
1	663	5441	182	32
2	746	4889	175	27
3	660	4997	218	21
12	710	4795	147	54
13	418	5449	157	33
23	521	5012	157	25
123	663	5315	178	27
total	5153	40,764	1396	245

[a] For each subset of a multi-target set, the average number of ECP identified in 100 individual trials with randomly chosen training sets is reported for three and 10 positive training instances (reference compounds).

active against one to three targets, the number of compounds substantially varied between 26 and 505 compounds.

The second category of data sets consisted of three sets of GPCR ligands with different mechanistic annotations including agonists, partial agonists, antagonists, and inverse agonist, as also reported in Table 2. These mechanism-based data sets were previously assembled from ChEMBL[23] for activity landscape design[24] and contained between 148 and 307 compounds. The distribution of these GPCR ligands over different mechanistic classes is reported in Table 4. The number of compounds per class ranged from 26 to 107 (the M_AM1 set did not contain inverse agonists). Upon online publication of this study, all data sets reported in Table 2 are made freely available via the "downloads" section of http://www.lifescienceinformatics.uni-bonn.de.

These two types of data sets presented related yet distinct and previously unconsidered multi-class prediction tasks for ECP analysis. We attempted to predict compounds belonging to each subset or class and distinguish them from all others in a given data set. Special emphasis was put on small training sets. Given their multi-class nature, the classification tasks were challenging, especially because compounds in different subsets or mechanistic classes were often structurally very similar, as illustrated in Figure 2.

5. COMPOUND CLASSIFICATION

5.1. Calculation Setup. For each data set, the values of all 61 descriptors were calculated, and the resulting value ranges were subjected to information entropy-based discretization. Descriptors yielding a single interval were not further considered for ECP analysis (because they could not yield differentiating patterns). For classification, either three or 10 reference compounds from each subset or class were randomly selected as positive training examples, and corresponding numbers of compounds from each other subset or class were utilized as negative training examples. Thus, for a given subset of a multi-target data set, the smallest learning set consisted of only three positive and 21 negative training instances. All remaining compounds represented the test set. ECP were derived from training set data and used to classify test instances. Test compounds were assigned a positive or negative class label on the basis of highest normalized cumulative support values. For each subset or class, 100 independent trials were carried out to obtain statistically sound results. As performance measures, classification *sensitivity* (i.e., true positives/(true positives + false negatives)), and *specificity* (i.e., true negatives/(true negatives + false positives)) were determined.

As a control, random forest[25] (RF) classification models were generated using the R implementation *randomForest*[26] with Molecular Operating Environment descriptors and standard parameter settings. In each case, 500 individual trees were calculated for n training set compounds, $n^{1/2}$ descriptors were randomly sampled at each decision point, and best performing descriptors were assigned to each node. Test compounds were classified based on the majority voting of all individual tree models.

5.2. Descriptor and ECP Statistics. The number of discretized descriptors with multiple value intervals per data set is reported in Table 5. For multi-target data sets, the number of qualifying descriptors ranged from 10 (set MT_D) to 24 (MT_B). Thus, in these cases, the majority of descriptors were eliminated following discretization. For mechanism-based data sets, more descriptors were obtained including 24, 45, and 52 descriptors for sets M_AM1, M_H3R, and M_AA1, respectively. Although only relatively small numbers of descriptors passed the discretization stage for multi-target sets, significantly varying numbers of ECP were obtained, as reported in Table 6. For example, for set MT_D, 10 discretized descriptors yielded an average of 128 ECP for three reference compounds, whereas for MT_B, 24 descriptors produced 6645 ECP. However, for each data set, comparable numbers of ECP per subset were obtained, despite varying numbers of compounds per subset. For larger sets of 10 reference compounds, total numbers of ECP further increased by factors of approximately two to six. For example, for 10 reference compounds, an average of 40,764 ECP was produced for MT_B, compared to 6645 ECP for three reference compounds. Equivalent observations were made for mechanism-based

Table 7. Emerging Chemical Patterns for Mechanism-Based Sets[a]

reference compounds = 3			
classes	M_AA1	M_H3R	M_AM1
agonist	2178	1074	125
antagonist	1362	843	94
partial agonist	1789	2218	158
inverse agonist	1657	1175	-
total	6986	5310	377
reference compounds = 10			
classes	M_AA1	M_H3R	M_AM1
agonist	22,536	10,099	512
antagonist	10,929	5862	407
partial agonist	16,085	20,879	757
inverse agonist	17,256	10,186	-
total	66,806	47,026	1676

[a] For each class in a mechanism-based set, the average number of ECPs identified in 100 individual trials with randomly chosen training sets is reported for three and 10 positive training instances (reference compounds).

Table 8. Exemplary Emerging Chemical Patterns for Multi-Target Sets[a]

designation	subsets	emerging chemical patterns	support
MT_A	13	{a_ICM:(1.81-inf), a_nCl:(0.5-inf), SlogP_VSA0:(-inf-41.32], vsa_other:(41.31-inf)}	0.6
	3	{a_don:(2.5-inf), a_ICM:(-inf-1.81], PEOE_VSA_FHYD:(0.74-inf), PEOE_VSA_FNEG:(-inf-0.26], SMR_VSA5:(208.27-inf)}	0.4
MT_B	2	{a_acc:(-inf-2.5], a_don:(-inf-0.5], PEOE_VSA+5:(12.07-inf), PEOE_VSA_FHYD:(0.87-inf), SMR_VSA2:(16.51-inf), TPSA:(-inf-52.58], vsa_don:(2.84-inf)}	0.5
	23	{a_acc:(-inf-2.5], a_don:(-inf-0.5], logP(o/w):(-inf-3.25], PEOE_VSA+0:(128.46-inf), PEOE_VSA-6:(-inf-19.27], PEOE_VSA_FHYD:(0.87-inf), SlogP_VSA3:(1.38-inf)}	0.4
MT_C	123	{a_don:(-inf-2.5], balabanJ:(-inf-2.03], PEOE_VSA+6:(-inf-5.44], PEOE_VSA_FHYD:(-inf-0.77], PEOE_VSA_FNEG:(-inf-0.28], SlogP_VSA0:(-inf-43.33]}	0.4
	13	{a_nCl:(0.5-inf), PEOE_VSA-6:(31.66-inf), SlogP_VSA0:(43.33-inf)}	0.3
MT_D	12	{a_nN:(-inf-3.5], a_nO:(-inf-4.5], SMR_VSA0:(-inf-100.66]}	0.4
	123	{a_nN:(3.5-inf), a_nO:(4.5-inf), PEOE_VSA-2:(8.94-inf), PEOE_VSA-6:(-inf-26.80], vsa_don:(-inf-11.59]}	0.3

[a] For two subsets of each multi-target set, exemplary ECP are reported, and their support is given for an individual trial with 10 reference compounds. Descriptors are abbreviated according to Table S1 of the Supporting Information, and "inf" stands for infinity.

data sets, as reported in Table 7. Because more descriptors were available for these sets, larger numbers of ECP were obtained in these cases. For example, for three reference compounds, sets M_AA1 and M_H3R produced 6986 and 5310 ECP, respectively, and for 10 reference compounds these numbers further increased to 66,806 and 47,026 ECP, respectively. Furthermore, compared to subsets of multi-target data sets, mechanistic classes displayed larger differences in the number of ECP per class. For example, for M_H3R, the average number of ECP for 10 reference compounds ranged from 5862 for antagonists to 20,879 for partial agonists. Despite these differences, hundreds to thousands of distinct ECP were obtained for compound classification in each case.

5.3. Signature Patterns. Characteristic ECP for different subsets and mechanistic classes are reported in Tables 8 and 9, respectively (descriptor definitions are provided in Table S1 of the Supporting Information). These ECP were only found in the designated subsets and had different levels of cumulative support. They were identified for individual training sets and significantly varied in their descriptor and interval composition and degree of complexity. As discussed above, these patterns are the most expressive jumping emerging patterns, following the theory of emerging patterns, which are by definition highly discriminatory in nature. This also rationalizes the ability of the ECP approach to successfully operate on the basis of unusually small training sets. Moreover, these ECP illustrate the variety of possible feature and interval combinations, giving rise to a very large possible number of ECP for pattern mining, as also reflected by the ECP statistics for our data sets. Given their diverse composition and numerical components, ECP are often not immediately interpretable, although feature trends might become apparent. For example, for the MT_B subsets in Table 8, combined charge and surface area descriptors together with hydrogen bond acceptors and donors dominate the patterns, while combined molar refractivity/SlogP and surface descriptors are recurrent in agonist classes in Table 9, albeit with different interval settings and in different descriptor combinations.

5.4. Prediction Accuracy. For each subset or mechanistic class, we carried out 100 independent classification trials with randomly selected training data and determined the sensitivity and specificity of the calculations. In our analysis, clear and consistent performance trends were detected. In Figure 3a and b, average sensitivity and specificity of ECP classification are reported for data set MT_A. The results obtained for the other three multi-target data sets were very similar and revealed the same trends and are reported in Figures S1−S3 of the Supporting Information. Characteristic features of classification trials on multi-target data sets included that their sensitivity was overall limited, between ∼15% and ∼60% in Figure 3a, whereas their specificity was consistently high, between ∼85% and >90% in Figure 3b. This means that only subsets of test compounds were correctly detected while false positive classifications were very rare. As shown in Figure 3, standard deviations were generally much larger for sensitivity than for specificity assessment, hence revealing a strong influence on the composition of training sets and the proportion of available test compounds that were correctly detected. Strikingly, we found no significant improvements in prediction accuracy for training sets with 10 positive instances compared to three positive ones, although many more ECP were available for mining in the former case. Hence, prediction accuracy was comparably high for very small training sets and not significantly influenced by the number of descriptors and the size of the resulting ECP pool.

Table 9. Exemplary Emerging Chemical Patterns for Mechanism-Based Sets[a]

designation	classes	emerging chemical patterns	support
M_AA1	agonist	{PEOE_VSA-4:(12.47−37.40], SlogP_VSA2:(46.29-inf), SMR_VSA3:(22.70−34.99], SMR_VSA4:(8.48−11.58], SMR_VSA6:(85.38-inf), vsa_pol:(45.52−51.96]}	1.0
	antagonist	{PEOE_VSA+2:(-inf-3.35], PEOE_VSA-4:(12.47−37.40])}	0.8
M_H3R	antagonist	{balabanJ:(1.34−1.50], chi0v_C:(13.16-inf), SlogP_VSA6:(3.96-inf), SMR_VSA6:(1.16−73.36]}	0.7
	agonist	{SMR_VSA0:(-inf-44.41], SMR_VSA5:(-inf-142.34]}	0.6
M_AM1	agonist	{SlogP_VSA3:(56.56-inf), SMR_VSA1: (51.08-inf)}	0.8
	antagonist	{PEOE_VSA+5:(-inf-57.23], SlogP_VSA1:(27.98−79.67]}	0.5

[a] For two classes of each mechanism-based set, exemplary ECP are reported, and their support is given for an individual calculation with 10 reference compounds. Descriptors are abbreviated according to Table S1 of the Supporting Information, and "inf" stands for infinity.

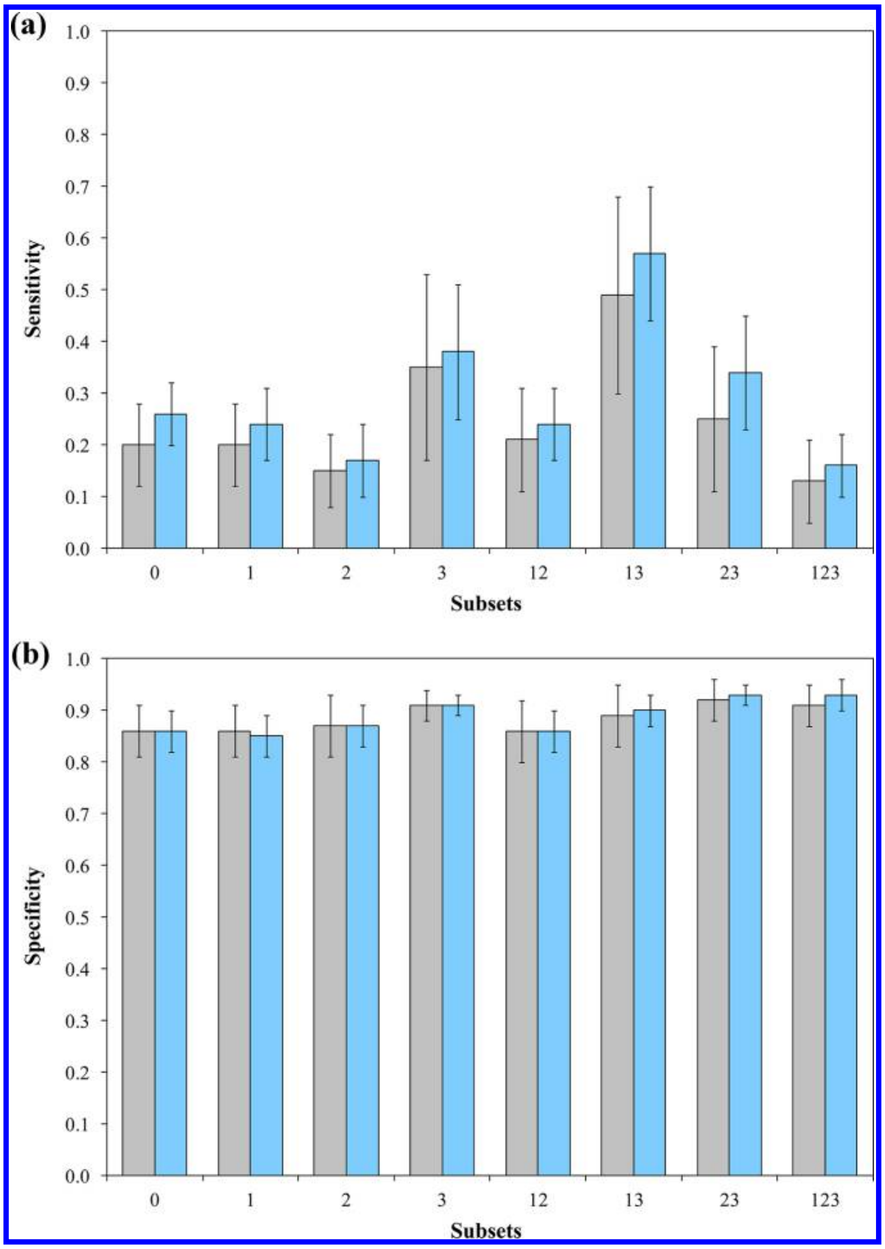


Figure 3. Prediction accuracy for a multi-target set. Average (a) sensitivity and (b) specificity are reported for subsets of multi-target set MT_A and positive training sets consisting of three (gray) or 10 (blue) compounds. Vertical lines report standard deviations.

From classification trials of mechanism-based data sets, a partly different picture emerged. Representative results are reported for data set M_AA1 in Figure 4a and b. Comparable results were obtained for the other two mechanism-based sets that are reported in Figures S4 and S5 of the Supporting Information. In these cases, the sensitivity of the classification calculations was significantly higher than for the multi-target data sets, between ~60% and >90% in Figure 4a, dependent on the mechanistic class, and the specificity was also very high, between ~85% and >95% in Figure 4b. Hence, compound recall was much higher for mechanism-based sets than for multi-target data sets, while false positive rates were very low in both cases. For mechanism-based sets, there were also only relatively small differences between the classification results obtained for reference sets of three or 10 compounds, especially considering the specificity of the calculations. The prediction accuracy of ECP calculations on mechanism-based data sets was impressively high for learning sets with only three positive training instances. For multi-target sets, low to moderate sensitivity of the calculations lowered overall prediction accuracy, although false positive rates were also very low in these cases.

The generally high degree of compound similarity within multi-target and mechanism-based data sets, which principally complicates compound classification, was reflected by often only limited numbers of descriptors for which value ranges were successfully discretized. This means that many compounds produced very similar values for given descriptors. However, we also observed a systematic difference in the cumulative support of ECP that dominated the classification of subsets and mechanistic classes. For subsets of multi-target data sets, many of the top ECP support values were ~0.5, whereas for classes of mechanism-based sets many top support values were ~0.7 (and larger). Hence, we can attribute the higher sensitivity of classification calculations on mechanistic classes to stronger support levels of ECP, which resulted in more characteristic patterns.

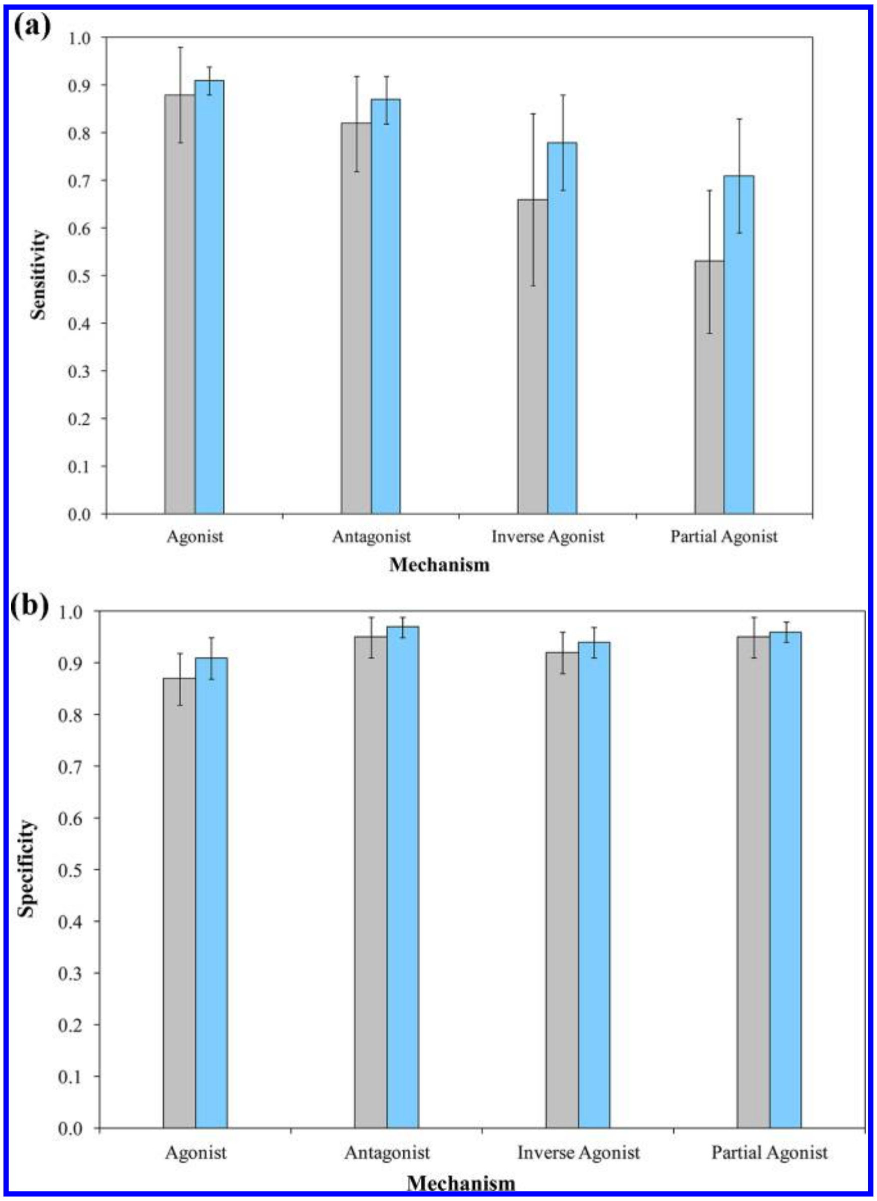


Figure 4. Prediction accuracy for a mechanism-based set. Average (a) sensitivity and (b) specificity are reported for compound classes in mechanism-based set M_AA1 and positive training sets consisting of three (gray) or 10 (blue) compounds. Vertical lines report standard deviations.

Table 10. Comparison of ECP and RF Calculations[a]

		sensitivity				specificity			
		ECP		RF		ECP		RF	
designation	classes	3	10	3	10	3	10	3	10
M_AA1	agonist	0.88	0.91	0.47	0.71	0.87	0.91	0.94	0.94
	antagonist	0.82	0.87	0.51	0.74	0.95	0.97	0.97	0.96
	inverse agonist	0.66	0.78	0.56	0.74	0.92	0.94	0.93	0.93
	partial agonist	0.53	0.71	0.40	0.63	0.95	0.96	0.93	0.93
M_H3R	agonist	0.59	0.67	0.55	0.57	0.87	0.90	0.93	0.94
	antagonist	0.64	0.74	0.38	0.59	0.92	0.94	0.96	0.97
	inverse agonist	0.64	0.76	0.47	0.52	0.89	0.93	0.93	0.92
	partial agonist	0.51	0.59	0.23	0.21	0.80	0.84	0.89	0.91
M_AM1	agonist	0.61	0.71	0.54	0.52	0.86	0.89	0.88	0.87
	antagonist	0.78	0.82	0.53	0.74	0.85	0.89	0.84	0.86
	partial agonist	0.60	0.67	0.40	0.53	0.82	0.85	0.82	0.84

[a] For each mechanistic class in a mechanism-based data set, the average sensitivity and specificity of 100 individual trials with randomly chosen training sets of three and 10 reference compounds are reported for ECP and RF.

Decision trees are among only a few machine learning approaches that can in principle also operate on the basis of very limited training data. Hence, for comparison, RF calculations have been carried out. The results obtained for the mechanism-based data sets (where both sensitivity and specificity of the ECP calculations were high) are reported in Table 10. As shown, RF calculations also yielded high specificity for small training sets, but the sensitivity of the RF calculations was lower than in the case of ECP. Similar observations were made for the multi-target data sets where RF calculations also yielded high specificity but only very low sensitivity (on average close to zero in many cases).

6. DISCUSSION AND CONCLUSIONS

Despite the popularity of the concept of emerging patterns in computer science,[27] only very few applications on molecular data sets have thus far been reported.[14−17] In part, this might be due to the fact that pattern mining represents an NP-hard computational problem (i.e., the time required increases exponentially with the number of features and patterns).[12] However, this is not generally a limiting factor as it mostly affects the derivation of classifiers, which can be addressed algorithmically. For example, for ECP calculations, an efficient hypergraph-based algorithm has been utilized to facilitate pattern mining.[10,14] On the other hand, a major attraction of the ECP approach is that it is applicable on the basis of small training sets, as indicated in its original compound classification application.[14] Previous and current results indicate that the highly discriminatory nature of jumping emerging patterns is largely responsible for this ability. Here, we have further investigated the ECP approach for complex classification problems involving data sets consisting of compounds with distinct or overlapping target activities and different molecular mechanisms. In our analysis, we have found that these classification problems involving many structurally similar compounds belonging to different subsets/classes could be accurately solved using the ECP approach. The calculations were characterized by varying degrees of sensitivity, depending on the data sets, and consistently high specificity. Especially for classification of compounds by mechanism-of-action, sensitivity and specificity of the calculations were generally high, regardless of the targets, leading to highly accurate predictions. For both multi-target and mechanism-based data sets, accurate predictions were obtained for learning sets with only three positive training instances. These findings provide strong further evidence for a cardinal feature of the ECP approach, i.e., its ability to operate on the basis of very limited compound information, even when facing complicated classification tasks. This characteristic feature should render ECP analysis attractive for practical applications in early phase drug discovery such as the search for new compounds with activity against emerging targets, for which only very limited prior compound information might be available. Once ECP classifiers are built, they can be applied to large test sets for class label predictions as well as the search for novel active compounds in cases where only a few reference molecules are available. Hence, both from a theoretical and practical point of view, the ECP approach should merit further consideration in chemoinformatics and drug discovery settings.

In conclusion, we have further investigated ECP analysis as a compound classification method. In our study, ECP calculations have yielded promising results in the classification of various multi-target and multi-mechanism data sets. Methodological aspects and practical implications of our findings have been discussed. It is hoped that the data sets provided as a part of our analysis might be useful for the evaluation of other machine learning methods and that the ECP concept might be applied by others to additional molecular classification and prediction tasks.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Figures S1−S3 report prediction accuracies for multi-target sets MT_B, MT_C, and MT_D, respectively. Figures S4 and S5 report prediction accuracies for mechanism-based sets M_H3R and M_AM1, respectively. Table S1 lists and defines descriptors used for ECP analysis. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.

(2) Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205−216.

(3) Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.

(4) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; Wiley-Interscience: New York, 2000, pp 20−83.

(5) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273−297.

(6) Vapnik, V. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer-Verlag: New York, 2000.

(7) Dong, G.; Zhang, X.; Wong, L.; Li, J. CAEP: Classification by Aggregating Emerging Patterns. In *Lecture Notes in Computer Science*, Vol. *1721*; Proceedings of the Second International Conference on Discovery Science, Tokyo, 1999; Arikawa, S., Furukawa, K., Eds.; Springer-Verlag: London, 1999, pp 30−42.

(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, U.S.A., 1999; Chaudhuri, S., Fayyad, U., Madigan, D., Eds; ACM Press: New York, 1999, pp 43−52.

(9) Li, J.; Dong, G.; Ramamohanarao, K. Making use of the most expressive jumping emerging patterns for classification. *Knowl. Inf. Syst.* **2001**, *3*, 131−145.

(10) Bailey, J.; Manoukian, T.; Ramamohanarao, K. A Fast Algorithm for Computing Hypergraph Transversals and Its Application in Mining Emerging Patterns. In *3rd IEEE International Conference on Data Mining*, Proceedings of the 3rd IEEE International Conference on Data Mining, Melbourne, Florida, U.S.A., 2003; IEEE Computer Society: Los Alamitos, CA, 2003, p 485.

(11) Li, J.; Dong, G.; Ramamohanarao, K.; Wong, L. DeEPs: A new instance-based lazy discovery and classification system. *Mach. Learn.* **2004**, *54*, 99−124.

(12) Wang, L.; Zhao, H.; Dong, G.; Li, J. On the complexity of finding emerging patterns. *Theor. Comput. Sci.* **2005**, *335*, 15−27.

(13) Li, J.; Wong, L. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. *Bioinformatics* **2002**, *18*, 725−734.

(14) Auer, J.; Bajorath, J. Emerging chemical patterns: A new methodology for molecular classification and compound selection. *J. Chem. Inf. Model.* **2006**, *46*, 2502−2514.

(15) Auer, J.; Bajorath, J. Simulation of sequential screening experiments using emerging chemical patterns. *Med. Chem.* **2008**, *4*, 80−90.

(16) 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.

(17) 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.

(18) Molecular Operating Environment (MOE), Chemical Computing Group, Inc.: Montreal, Quebec, Canada.

(19) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151−1157.

(20) Fayyad, U. M.; Irani, K. B. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Chambry, France, 1993; Bajcsy, R., Eds.; Morgan Kaufmann Publishers: San Francisco, 1993, pp 1022−1027.

(21) Witten, I. H.; Frank, E. Introduction to Weka. In *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann Publishers: San Francisco, 2005, pp 365−368.

(22) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's Bioassay Database. *Nucleic Acids Res.* **2012**, *40*, D400−D412.

(23) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(24) Iyer, P.; Bajorath, J. Mechanism-based bipartite matching molecular series graphs to identify structural modifications of receptor ligands that lead to mechanism hopping. *Med. Chem. Commun.* **2012**, *3*, 441−448.

(25) Breiman, L. Random forests. *Machine Learn* **2001**, *45*, 5−32.

(26) Liaw, A.; Wiener, M. Classification and regression by *randomForest*. *R News* **2002**, *2*, 18−22.

(27) Contrast Data Mining: Concepts, Algorithms, and Applications. In *Data Mining and Knowledge Discovery Series*; Dong, G., Bailey, J., Eds.; Chapman & Hall/CRC Press: Boca Raton, FL, 2012.