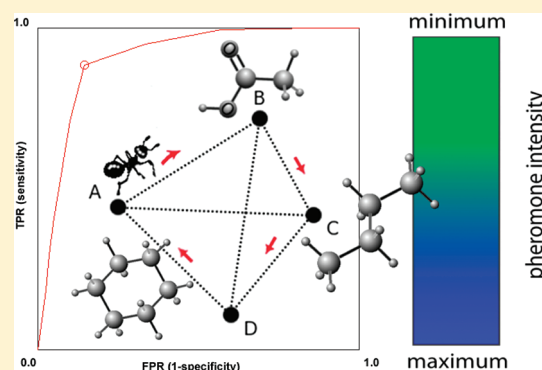


A Binary Ant Colony Optimization Classifier for Molecular Activities

Felix Hammann,^{†,*} Claudia Suenderhauf,[†] and Jörg Huwylér[†][†]Division of Pharmaceutical Technology, Department of Pharmaceutical Sciences, University of Basel, Klingelbergstrasse 50 4056, Basel, Switzerland

S Supporting Information

ABSTRACT: Chemical fingerprints encode the presence or absence of molecular features and are available in many large databases. Using a variation of the Ant Colony Optimization (ACO) paradigm, we describe a binary classifier based on feature selection from fingerprints. We discuss the algorithm and possible cross-validation procedures. As a real-world example, we use our algorithm to analyze a *Plasmodium falciparum* inhibition assay and contrast its performance with other machine learning paradigms in use today (decision tree induction, random forests, support vector machines, artificial neural networks). Our algorithm matches established paradigms in predictive power, yet supplies the medicinal chemist and basic researcher with easily interpretable results. Furthermore, models generated with our paradigm are easy to implement and can complement virtual screenings by additionally exploiting the precalculated fingerprint information.



INTRODUCTION

Chemical fingerprints, which are in essence hashes calculated from molecular structures, are frequently used in large chemical databases.¹ These fixed-length strings encode a variety of molecular properties, oftentimes the presence or absence of a substructural motif.^{2,3} Using fingerprints, it is possible to quantify chemical similarity, restrict searches to a number of promising candidates, and so on.^{4,5} A variety of distance measures exist to determine proximity between two molecules and also to define clusters of similar structures.^{1,6}

According to the similarity principle, molecules with closely related structures are likely to exhibit the same activity.⁷ While similarity may be defined as small distance from a selected point (e.g., a centroid in clustering), one may also construct a plane of separation within the attribute space to distinguish one type of molecule from another. A typical example of this approach are support vector machines,⁸ and these have been successfully employed in many quantitative structure–activity relationship (QSAR) studies.⁹

Statistical models often rely on a number of physicochemical descriptors, which are rarely available in chemical databases. Screening an entire database therefore requires retrieving the complete set of structures and subsequently calculating these properties. On the other hand, fingerprints already contain many properties calculated upon insertion into the database. A classification scheme based on fingerprints could therefore save data traffic and computing power.

Such a classifier would need to find a subset of attributes present in a list of desirable molecules and absent in a list of negative controls. This feature selection is essentially an optimization problem. In recent years, the field of natural computing has produced intriguing heuristics for optimizations in engineering and the natural sciences. Ant

Colony Optimization (ACO), a paradigm introduced in the 1990s,¹⁰ has drawn a special amount of attention. Real-world ants are abstracted as agents able to traverse a graph while they deposit a pheromone whose intensity decays over time. An ant scurrying about the graph at random until it finds a food source initiates the process. It then returns to the starting point in a more or less direct trajectory. Other ants explore the graph and weigh their choices of route by previously deposited pheromones. Eventually, shorter (i.e., more efficient) paths will extrude a more intense signal and become points of convergence.

Here, we propose a binary classifier that uses an ACO variant to select relevant molecular fragments. Modifications of ACO have been proposed previously for variable selection and reduction of dimensionality.¹¹ They have also found application in the field of drug discovery, where they, however, were used in ensemble prediction settings (as feature reduction prior to e.g. linear regression^{12,13} or support vector machines¹³ QSAR/QSPR studies of anti-HIV activity and human serum albumin binding activity, respectively). ACO is also often employed (along with other optimization paradigms) in protein ligand-docking studies.¹⁴ While ACO has been applied to molecular binary classification (e.g., as an estimator of splitting criteria in decision tree induction),¹⁵ we are not aware of its solitary use in fragment analysis.

The variant described by us can be visualized with the ant colony at the center and various single fingerprint flags as the vertices of edges of equal length radiating from the center, i.e. a complete bipartite graph $S_{1,n}$ where n is the number of fingerprint flags (Figure 1). From this, the method compiles a subset of flags that are associated with a given label or activity.

Received: May 13, 2011

Published: August 20, 2011

THEORY

Fingerprints. For a molecule A , a fingerprint F with n elements takes the form of an attribute vector F_A^1

$$F_A = \{f_{1A}, f_{2A}, \dots, f_{nA}\}$$

The fingerprints used in this context are dichotomous (or binary), i.e. each element f_{xA} corresponds to a bit which encodes

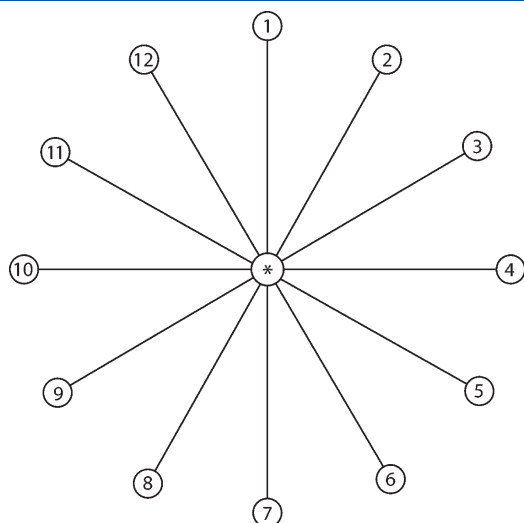


Figure 1. Sample graphical depiction of the ant colony feature selection problem for twelve different attributes. Edges are of equal length, with the colony at the center (*).

the presence (ON) or absence (OFF) of a feature within the molecule A .

Classification Problem. The problem can be stated as follows: given a total of n positions in a set F of fingerprints bits, find the subset S of magnitude m ($m < n, S \subset F$) so that the subset of features Q ($Q \subset F$) present in all active compounds has maximum specificity and sensitivity compared to the set of inactive compounds. Initially, a fixed number n_a of ants, each with the ability to select m features, explore the feature space at random and return to the nest. This random exploration is implemented by assigning a random amount of pheromone τ ($0 < \tau \leq 1$) to each feature at the start of a run. Here, a heuristic fitness function H rates each ant's performance, and ants are ranked by quality of their subset. The best k ($k < n_a$) ants are selected and deposit a constant amount τ of pheromone on the edges connecting members of their subset to the nest. The other ants are ignored. All n edges are allowed to evaporate their pheromone trails by a constant linear term d , and a new cycle is initiated. Again, n_a ants are created. Each ant now generates a random number r_i for each i of the n edges connected to the nest ($0 \leq r_i \leq 1$) and ranks their attractiveness by choosing those with a maximal value for the term

$$a_i = r_i \tau_i$$

where τ_i is the intensity of the pheromone signal on edge i , and r_i is the random weight. With the introduction of the random term, exploration of other combinations is encouraged. Over a given number of cycles, information-rich features are reinforced and increasingly become part of subsets until ants will almost uniformly

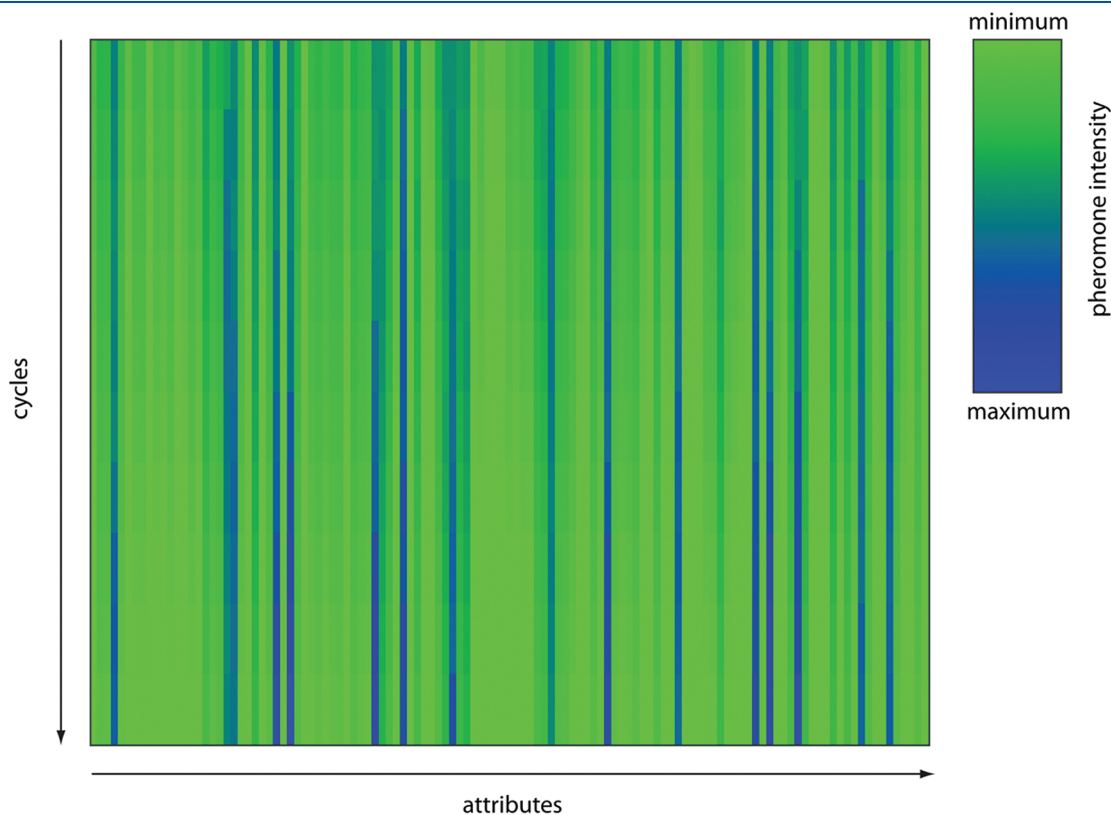


Figure 2. Visualization of evolution of a solution over a number of cycles as produced by the software written for this study. Initially, many different attributes are being explored until several strong attributes are converged upon (indicated by intensity of pheromone trail).

Chart 1. Pseudocode Representation of Learning Algorithm

```

KeyLength: length of fingerprint key in bits
m: number of features in ant memory
n: number of ants to create per cycle
k: number of ants to keep per cycle, where  $k \leq n$ 
Pheromones: list of length KeyLength containing pheromone intensities associated with keys

Initialize Pheromones with random floating point values from the range [0, 1]

repeat for a given number of cycles
    repeat for each of n ants
        copy Pheromones to Pheromones'
        multiply every position in Pheromones' with random floating point number from the range [0, 1]
        find m positions in Pheromones' with highest value
        compute fitness H and store with ant
    end repeat

    select k ants with highest H and repeat for each of m features in ant memory add constant pheromone
    amount  $\tau$  to corresponding value in Pheromones

    for each position in Pheromones subtract constant linear evaporation rate
end repeat

Output: the list of pheromone intensities Pheromones

```

choose these same features (Figure 2). The pseudocode as shown in Chart 1 further illustrates the proposed method.

Heuristic Fitness Function H . The heuristic fitness function H used in this study first finds the cardinality c_i of the intersection I_i of the subset S of features being evaluated and the entire set of features M_i ($M_i \subset F$) of each molecule i in a training data set such that

$$I_i = S \cap M_i$$

Depending on the original parametrization of the ant agents, c_i will take on values between 0 and cardinality of S . The fitness function determines c_i for every instance in the training set and groups instances by this value. Sensitivity and specificity of S for active molecules can be ranked by the area under the curve (AUC) of receiver operating characteristic (ROC) curves, as c_i as a cutoff value increases. This AUC is also the return value of $H(S)$ for a subset S . A pseudocode representation is given in Chart 2.

Statement of Models. The ROC curves are used further to determine the cutoff point with optimal sensitivity and

specificity. The Youden index¹⁶ J given as

$$J = \text{sensitivity} + \text{specificity} - 1$$

is maximal for this point. A model built in this fashion can therefore be stated as the set of features indicative of activity and the minimum number of features required to qualify as active. A model P built with m features and cutoff point at p features takes the form of

$$P = \{\{x_1, x_2, \dots, x_m\}, p\}$$

As an illustration, consider a sample model M trained from a set of 100 possible binary keys to select the 10 keys associated with a given activity. This might look as follows

$$M = \{\{4, 12, 15, 23, 38, 42, 61, 89, 90, 95\}, 3\}$$

For an instance to be classified as active, 3 or more of the 10 features would need to be present in its key vector, e.g. a molecule with a vector

$$\text{Mol}_1 = \{10, 12, 19, 20, 23, 38, 49, 50, 67, 70, 82, 83, 100\}$$

Chart 2. Pseudocode Representation of Heuristic Fitness Function H

```

TrainingData: training data with binary labels and fingerprint keys
AntMemory: set of m keys

repeat for each instance in TrainingData

    compute hits (i.e. ci) as the number of keys present in both the instance and AntMemory

    sort and divide instances in TrainingData by hits

    continuously combining groups of instances ordered by value of hits, compute true and false positive rates
    for all instances as coordinates

    calculate the area under the curve formed by these points

end repeat

Output: area under the curve

```

Table 1. Mean Fingerprint Darkness (Number of Bits Set over Total Number of Bits), with Minimum (min%) and Maximum (max%) Percentages and Standard Deviation in Percent (sd%) of the 166 Bit MACCS Key and the 1024 Bit Standard and Extended Keysets

key	class	mean%	max%	min%	sd%
MACCS	negative	29.7	49.4	4.8	7.8
	positive	24.1	49.4	1.2	9.2
	total	26.1	49.4	1.2	9.1
Standard	negative	9.9	49.9	0.1	8.3
	positive	17.1	60.1	0.8	9.6
	total	12.5	60.1	0.1	9.4
Extended	negative	10.2	51.4	0.1	8.4
	positive	17.6	59.7	0.8	9.6
	total	12.9	59.7	0.1	9.6

would classify as active (as the intersection with the key vector in model *M* has a cardinality of 3).

Cross-Validation Procedure. To avoid overfitting (i.e., creating overly complex models with very high predictive accuracy on training data by extracting too many parameters from the known data at the expense of not being able to predict unseen compounds), we used *k*-fold cross-validation (CV). Here, a data set is randomly recombined into *k* subsets (here *k* = 10).¹⁷ Of these, *k* - 1 are recombined to make up a training set which is tested against the remaining subset. This process is repeated *k* times until all instances have served as training and test data, thereby making sure that no classes are left out. Sets were permuted using the Fisher-Yates-Shuffle algorithm as detailed by Knuth.¹⁸

We evaluated three different ways of combining the different models: averaging of pheromone weights in every fold (averaged model), selection of most frequently employed attributes (frequency model), and combination of attributes most frequently selected by elite ants (elite ants model), i.e. the single best performing ant within a run. For averaging, the

pheromone weights associated with each of the *n* attributes are normalized to a range (0, 1). Next, all of the *k* pheromones for a given attribute are summed up. The result is a list of *n* combined pheromone weights ranging from 0 to *k*, allowing them to be ranked. Attributes of the highest rank are selected and make up the final model. For the frequency model, the highest-ranking attribute of each fold is selected. In order to create the elite ants model, the software stores the single best performing ant of each fold. The corresponding feature sets are combined in the same manner as in the frequency model.

Performance Measures. We report the predictive power of each model using two measures, in order to aid comparison with other studies. First, we give the Corrected Classification Rate (CCR) as

$$CCR = \frac{1}{2} \left(\frac{T_N}{N_0} + \frac{T_P}{N_1} \right)$$

where *T_N* and *T_P* represent the number of true negative and positive predictions, respectively, and *N₀* and *N₁* represent the total number of negative and positive compounds in the model. Second, we provide the Matthews Correlation Coefficient (MCC) as

$$MCC = \frac{T_N T_P - F_N F_P}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}}$$

where *F_N* and *F_P* denote false negative and false positive predictions, respectively.

■ PLASMODIUM FALCIPARUM GROWTH INHIBITOR ASSAY

Data Set and Preparation. Models were learned from data of a high-throughput SYBR Green proliferation assay of *P. falciparum* (Pf) infected red blood cells published by Plouffe et al.¹⁹ The data were retrieved from PubChem (<http://pubchem.ncbi.nlm.nih.gov/>) and contain a total of 1272

Table 2. Results of Binary Ant Colony Optimization (ACO) Classification Using Three Different Cross-Validation Paradigms and Comparison with Established Machine Learning Paradigms^a

	TP	TN	FN	FP	CCR	MCC	accuracy
Elite ACO Model (Run 36)	171	307	29	40	0.87	0.73	0.87
Frequency ACO Model (Run 17)	169	294	31	53	0.85	0.68	0.85
Averaged ACO Model (Run 16)	162	303	38	44	0.84	0.68	0.85
J48	156	313	44	34	0.84	0.69	0.86
RF	155	316	45	31	0.84	0.70	0.86
SVM	157	317	43	30	0.85	0.71	0.87
ANN	161	296	39	51	0.83	0.65	0.84

^a The data consist of 547 instances (positive: 200, negative: 347). J48: decision tree induction, RF: random forests, SVM: support vector machines, ANN: artificial neural networks. Performance is measured as corrected classification rate (CCR), Matthews correlation coefficient (MCC), and accuracy.

compounds (201 active, 349 inactive, and 722 inconclusive). We omitted compounds labeled as inconclusive as well as those for which not every CDK descriptor could be calculated ($n = 3$). We removed disconnected small fragments such as counterions prior to any calculation in analogy to McGregor and Pallai.³

We evaluated three fingerprint keys (MACCS (MDL), Standard, Extended) available in the latest stable Chemical Development Kit (Version 1.2.7).²⁰ The 166 bit MDL key² was used in the final models as it is the best documented of the three and has been optimized to allow for clustering of bioactive substances in the context of drug discovery. The concept of fingerprint darkness refers to the fraction of bits set to ON, i.e. we consider fingerprints with more bits set to ON as darker. The characteristics of the three different keys evaluated are given in Table 1. Of these, the MACCS 166 bit key shows the greatest darkness (26.1%), implying that it is capable of reflecting the most features with the least computational effort.

Training of Models. We let the classifier learn over 100 cycles in order to produce models of a magnitude of 10. Ants deposited a pheromone amount $\tau = 0.1$ which evaporated by $d = 0.05$ within cycles. We performed 100 runs using the three different modes of cross-validation outlined above. This amounted to a total of 300 models. For comparison, we created models with a decision tree induction algorithm (J4.8, a C4.5 variant), random forests (RF) of ten trees with five attributes each,²¹ support vector machines (SVM) using a polynomial kernel function, and artificial neural networks (ANN) with a single hidden layer. In line with other current studies, models were learned in a 10-fold cross-validated context.^{22,23} The numerical attributes used in this process were the 1D and 2D descriptors ($n = 27$) available in the CDK (molecular weight, calculated partitioning coefficient (logP), tological polar surface area, BCUT metrics, fragment complexity, atom and bond counts of aromatic and of all atoms, hydrogen bond donor and acceptor counts, Kier-Hall shape indices, Petitjean number, number of rotatable bonds, atomic polarizability, and length of largest chain and largest aliphatic chain as well as length of largest π chain).²⁴

Model Performance. The classification results of the best performing models for each mode of cross-validation are shown in Table 2. All three CV procedures achieve comparable CCRs of 0.84 to 0.87 - values that match those of the

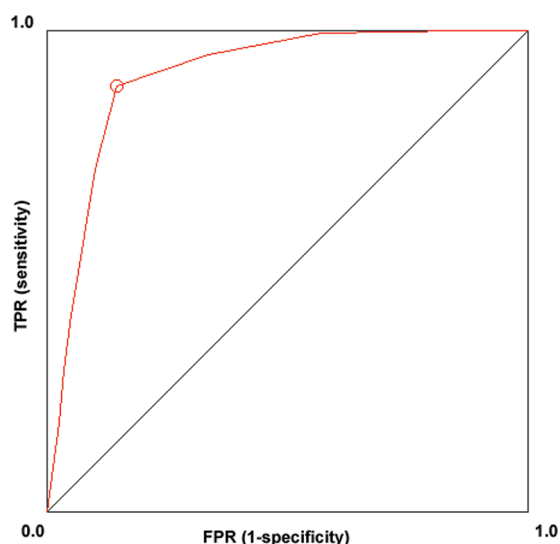




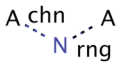
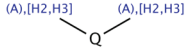

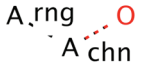


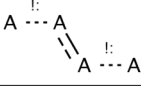

Figure 3. Receiver operating characteristic curve for the best performing classification model in this study (area under the curve = 0.91). The circle denotes the cutoff point from which on instances are classified as positive.

other paradigms implemented. An elite ant model achieved the highest CCR (0.87). Its associated ROC curve has a high area under the curve of 0.91 and is given in Figure 3. It is readily apparent from Table 3, which presents the substructural motifs selected by the model, that the binary ACO classifier retrieves fragments with a mechanistical relevance.

For instance, the presence of nitrogen and oxygen atoms in different frameworks (keys 25, 75, 127, 140, and 161) is characteristic of druglike molecules (hydrogen bonding capacity) as well as important for successfully overcoming cellular membranes. They are also present in molecules which exert oxidative stress, to which *P. falciparum* is very sensitive. The prevalence of both nitrogen and oxygen are high in small molecule drugs. Of course, some rare examples exist which contain neither (noteworthy members of this class are lindane and mitotane, two chemotherapeutic agents), and many molecules contain exclusively one these atom species (e.g., nitrogen in amitriptyline, selegiline, and memantine, and oxygen in ivermectine, digoxin, and cholecalciferol). This shows how the substructures identified by the paradigm need to be interpreted together. For example, nitrogen appears in other keys (keys 25 (trisamino/imino methylene) and 75). Presence of one of these more complex substructures therefore automatically increases the score and, by consequence, the likelihood of positive classification. In the same vein, molecules lacking keys 25 or 75 can improve their score by offering other hydrogen binding sites, e.g. oxygen (keys 140 and 127), thereby increasing their druglikeness.

A key enzyme in the life cycle of Pf, the cysteine protease Falcipain-2 (FP-2) that degrades hemoglobin (Hb), can be inhibited by certain epoxysuccinates and aziridiny substituents in quinone rings (perceived, among others, by keys 86 and 137) and have been shown to enhance antiparasmodial activity by inhibiting Pf glutathion reductase.²⁵ The life cycle of Pf is particularly vulnerable during the erythrocytic stage as its metabolism is largely anaerobic and hence sensitive to oxidative stress.

Table 3. Fingerprint Keys Associated with a *Plasmodium falciparum* Growth Inhibition As Determined by Binary Ant Colony Optimization Classification^a

Index	Depiction	SMARTS	Comment
25		[#7]~[#6](~[#7])~[#7]	trisamino / imino methylene
49		[!+0]	presence of charge
75		*!@[#7]@*	interposition of nitrogen
86		[C;H2,H3][!#6;#1][C;H2,H3]	carbon – heteroatom – carbon chain
124		[!#6;#1]~[!#6;#1]	two connected heteroatoms
127		*@!@[#8]	oxygen connected to any ring system via a single bond
137		[!C;!c;R]	any heterocycle
140		[#8]	presence of oxygen
144		*!.*.*!.*	aromatic ring substituted in ortho-position by two non-aromatic substituents
161		[#7]	presence of nitrogen

^a Substructural motifs are given with their position (index), SMILES arbitrary target specification (SMARTS) along with an image and an explanation.

CONCLUSIONS

We investigated whether binary classification of molecular activity using a variation of the ACO paradigm could become a valid alternative to other ML classification methodologies. Analysis of the Pf inhibition assay by Plouffe et al. shows the high degrees of accuracy achieved by our models and their competitiveness with established ML methods.

The different modes of CV produce similarly powerful classifiers. From Table 2 it is evident that these models stem from different runs, i.e. the choice of CV influences the final performance, and no final ranking can be made between these modes. Therefore we consider it advisable to calculate all three to maximally exploit the information extracted by the learning process.

The information provided to the binary ACO learning algorithm was in essence a list of the presence or absence of substructural motifs or fragments, i.e. two-dimensional structural information. We therefore explicitly learned the alternative ML methods from two-dimensional descriptors as well to ensure a level playing field. Arguably, one might see better performance of the established ML methods with a different choice of descriptors. Conversely, other fingerprint keys could improve the results of binary ACO classification.

We chose MACCS over the other available fingerprint keys in CDK because of its length (166 bits vs 1024 bits for Standard and Extended) and its high ratio of keys set to ON. Notably, the molecules tagged as negative have a higher fingerprint

darkness than the positive instances, i.e. the inactive compounds are actually captured better than the active ones. When one learns a model to distinguish active compounds by presence of certain features from such a data set, it is apparent that the algorithm cannot simply associate fingerprint darkness of a compound with activity.

The substructures encoded in the MACCS fingerprint are often-times ambiguous or very general, and features selected by our algorithm can overlap (e.g., keys 25 and 161). Still, models perform well and robustly in a cross-validated setting. This indicates that the subsets of keys are more than the sum of their parts, i.e. the individual contribution of a key must be seen in the context of the entire subset. Also, a feature that is recognized by several keys is amplified (or deemed more important) in the perception of the classifier.

Binary ACO models can benefit the drug discovery process in two principle ways. First, the models provide an explicit fragment analysis directly accessible to human interpretation. Medicinal chemists can use them as guides for further development. Second, models can be applied directly to existing databases without any further calculations if both use the same fingerprinting scheme. This is in contrast to more elaborate numerical methods (e.g., SVM or ANN) where (a) a number of physicochemical descriptors need to be computed and (b) the software implementation of the classifier itself is complex. In fact, binary ACO models learned from fingerprints could be implemented as native database queries.

Of the learning paradigms employed in this study, decision tree induction took the least time to produce models. This, of

course, does not consider the time required for calculating descriptors, performing intercorrelation analysis, and checking for missing values. Support vector machines had the most time-intensive learning process. For SVM, we are not considering the tedious process of optimizing learning parameters. Similar considerations, of course, have also to be made when applying the binary ACO algorithm. Proper choice of the model size, m , influences not only the interpretability and performance on unseen data but also the time required to learn models. The number of cycles being spent on learning contributes directly to the computational expense. In summing up, the proposed algorithm ranks with SVMs in terms of time consumption for learning. A more thorough profiling does not seem called for, as the learning paradigms differ in practice in the amount of data preparation and optimization they require.

In the future, this algorithm could be extended to numerical predictions, i.e. the learning process could correlate the number of keys present with the degree of activity. Additionally, instead of merely identifying keys contributing to activity, a variant of the algorithm proposed here might single out detrimental features and incorporate them in the predictions.

■ SOFTWARE USED

Fingerprints were created with the open-source cheminformatics package Chemical Development Kit (version 1.2.7, 2010; <http://sourceforge.net/projects/cdk/>). Feature selection was performed using in-house software. Chemical structure diagrams were created using ChemAxon MarvinSketch (Version 5.2.5; <http://www.chemaxon.com/>). For machine learning paradigms (J4.8, SVM, ANN, and RF), we used the Weka Toolkit (Version 3.6; <http://www.cs.waikato.ac.nz/~ml/Weka/>).²⁶

■ ASSOCIATED CONTENT

S Supporting Information. An extended table gives more structural information on the molecules analyzed in this study. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +41 61 267 1513. E-mail: felix.hammann@unibas.ch.

■ ACKNOWLEDGMENT

The authors wish to thank Dr. Heinz Hammann for the fruitful discussions regarding CDK fingerprints. Dr. Claudia Suenderhauf is supported by the Swiss National Foundation (Grant No. 323530-119218).

■ REFERENCES

- (1) Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (2) Durant, J.; Leland, B.; Henry, D.; Nourse, J. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
- (3) McGregor, M.; Pallai, P. Clustering of large databases of compounds: using the MDL “keys” as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.
- (4) Barnard, J. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644–649.
- (5) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11* (23–24), 1046–53.
- (6) Butina, D. Unsupervised data base clustering based on daylight’s fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750.
- (7) Glen, R.; Adams, S. Similarity Metrics and Descriptor Spaces – Which Combinations to Choose? *QSAR Comb. Sci.* **2006**, *25* (12), 1133–1142.
- (8) Vapnik, V. N. *The nature of statistical learning theory*; Springer: 2000.
- (9) Michielan, L.; Moro, S. Pharmaceutical Perspectives of Non-linear QSAR Strategies. *J. Chem. Inf. Comput. Sci.* **2010**, *50*, 961–978.
- (10) Bonabeau, E.; Dorigo, M.; Theraulaz, G. Inspiration for optimization from social insect behaviour. *Nature* **2000**, *406* (6791), 39–42.
- (11) Shen, Q.; Jiang, J.; Tao, J.; Shen, G.; Yu, R. Modified Ant Colony Optimization Algorithm for Variable Selection in QSAR Modeling: QSAR Studies of Cyclooxygenase Inhibitors. *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 1024–1029.
- (12) Gunturi, S. B.; Narayanan, R.; Khandelwal, A. In silico ADME modelling 2: computational models to predict human serum albumin binding affinity using ant colony systems. *Bioorg. Med. Chem.* **2006**, *14* (12), 4118–29.
- (13) Goodarzi, M.; Freitas, M. P.; Jensen, R. Feature selection and linear/nonlinear regression methods for the accurate prediction of glycogen synthase kinase-3beta inhibitory activities. *J. Chem. Inf. Model.* **2009**, *49* (4), 824–32.
- (14) Korb, O.; Stutzle, T.; Exner, T. E. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49* (1), 84–96.
- (15) Izrailev, S.; Agrafiotis, D. A novel method for building regression tree models for QSAR based on artificial ant colony systems. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (1), 176–80.
- (16) Youden, W. Index For Rating Diagnostic Tests. *Cancer* **1950**, *1*, 32–35.
- (17) Kohavi, R. In *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Int. Joint Conf. Artif. Intell., Montreal, Quebec, Canada, Montreal, Quebec, Canada, 1995; pp 1137–1143.
- (18) Knuth, D. *Art of Computer Programming*; Addison-Wesley Professional: Boston, USA, 2011.
- (19) Plouffe, D.; Brinker, A.; McNamara, C.; Henson, K.; Kato, N.; Kuhen, K.; Nagle, A.; Adrián, F.; Matzen, J. T.; Anderson, P.; Nam, T.; Gray, N. S.; Chatterjee, A.; Janes, J.; Yan, S. F.; Trager, R.; Caldwell, J. S.; Schultz, P. G.; Zhou, Y.; Winzler, E. A. In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 9059–9064.
- (20) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.
- (21) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.; Sheridan, R.; Feuston, B. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
- (22) Cheng, T.; Li, Q.; Wang, Y.; Bryant, S. H. Binary classification of aqueous solubility using support vector machines with reduction and recombination feature selection. *J. Chem. Inf. Model.* **2011**, *51* (2), 229–36.
- (23) Suenderhauf, C.; Hammann, F.; Maunz, A.; Helma, C.; Huwyler, J. Combinatorial QSAR modeling of human intestinal absorption. *Mol. Pharm.* **2011**, *8* (1), 213–24.
- (24) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, NY, 2000.
- (25) Ramjee, M.; Flinn, N.; Pemberton, T.; Quibell, M.; Wang, Y.; Watts, J. Substrate mapping and inhibitor profiling of falcipain-2, falcipain-3 and berghep-2: implications for peptidase anti-malarial drug discovery. *Biochem. J.* **2006**, *399*, 47–57.
- (26) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* **2009**, *11*, 10–18.