

Variable Selection for QSAR by Artificial Ant Colony Systems

Sergei Izrailev* and Dimitris K. Agrafiotis

3-Dimensional Pharmaceuticals, Inc., 665 Stockton Drive, Exton, PA 19341, USA

RUNNING TITLE: Variable selection by artificial ants

* Corresponding author. Tel: (610) 458-5264 Ext. 6570, Fax: (610) 458-8249, E-mail:

sergei@3dp.com.

ABSTRACT

Derivation of quantitative structure-activity relationships usually involves computational models that relate a set of input variables describing the structural properties of the molecules for which the activity has been measured to the output variable representing activity. Many of the input variables may be correlated, and it is therefore often desirable to select an optimal subset of the input variables that results in the most predictive model. In this paper we describe an optimization technique for variable selection based on artificial ant colony systems. The algorithm is inspired by the behavior of real ants, which are able to find the shortest path between a food source and their nest using deposits of pheromone as a communication agent. The underlying basic self-organizing principle is exploited for the construction of parsimonious QSAR models based on neural networks for several classical QSAR data sets.

Keywords:

Artificial intelligence, machine learning, artificial ants, computer-assisted drug design, quantitative structure-activity relationships.

INTRODUCTION

In recent years, there has been an increasing need for novel data mining methodologies that can analyze and interpret large volumes of data. Artificial intelligence methods, such as artificial neural networks, classification and regression trees, and k-nearest neighbor algorithms, have been used extensively for analysis and correlation of chemical and biological data. Many of these methods are used in conjunction with optimization techniques, including various greedy algorithms as well as stochastic optimization approaches such as simulated annealing and genetic algorithms. In this paper we present a stochastic optimization technique based on artificial ant colony systems [1] and apply it to the problem of variable selection for constructing quantitative structure-activity relationship (QSAR) models for a set of chemical compounds.

The physico-chemical and structural properties of the molecules are usually represented by a set of variables (descriptors), with the assumption that the molecule's activity is in some way related to the values of these variables. Many of the descriptors may be correlated, so it is often desirable to find a smaller set of variables that capture the relevant molecular properties, and to construct a QSAR model using only these variables. The choice of such an optimal set of variables constitutes the variable selection problem. Variable selection is also commonly used in cases when the number of descriptors exceeds the number of activity measurements.

The algorithms based on artificial ant systems are inspired by the fact that real ants, using deposits of pheromone as a communication agent, are able to find the shortest path between a food source and their nest [2]. A moving ant deposits pheromone on the ground, thus marking its path. Although each individual ant moves at random, it can detect pheromone trails and follow one of them with a probability proportional to the amount of pheromone on the trail. By adding its own pheromone deposits, the ant reinforces the trail and

makes it more attractive to the other ants. While all paths are initially equally probable, the shorter ones encounter more ants making round trips to the food source per time unit and, therefore, receive more pheromone. Thus, short paths become increasingly more attractive to the ants. Eventually, all ants follow the shortest trail.

The initial formulation of the ant colony algorithm [3] addressed the traveling salesman problem: given a set of cities and a set of costs associated with traveling between these cities find a route that visits each city exactly once and has the least total cost. Subsequently, a number of variations of ant algorithms and their applications have been reported (see Reference [4] for a review). Recently, we have successfully applied an algorithm based on artificial ant colonies to construct optimal regression tree models for QSAR [5]. In this work we describe a formulation of an artificial ants algorithm for variable selection and test its performance on three well-known QSAR data sets.

METHODS

As applied to the variable selection optimization problem, a choice of a variable is analogous to a step of the real ant's path; therefore, the whole path represents a choice of a subset of K variables out of all M variables. The number of variables K is either defined by the user or is a parameter for optimization. Each variable k , $k=1\dots M$, is assigned a weight w_k that is used to calculate the probability p_k with which the variable is randomly selected by an ant. Initially, the weights (and probabilities) for all variables are equal ($w_k = w_0$ for all k). After the first ant has selected a subset of K variables, a model is built that predicts the activity on the basis of this subset. The "length" L of the ant's path is related to the quality (fitness) f of the model. Better fitness corresponds to a shorter path. In the simplest case, if the fitness function increases with the quality of the model, L can be set equal to the inverse of the fitness, $L(f) = 1/f$. Any machine learning technique and fitness function can be used to construct the model and calculate its

fitness. After L has been calculated, the weights corresponding to the selected variables are updated according to the following rule:

$$w_k(t+1) = (1 - \rho)w_k(t) + \Delta w / L, \quad (1)$$

where t is the ant's number, ρ is the evaporation coefficient that simulates the evaporation of the pheromone from the real ants' paths, and Δw is a constant factor. The next ant calculates the probabilities p_k according to Eq. 2 using the updated weights.

$$p_k = w_k / \sum_k w_k \quad (2)$$

The process is continued for the specified number of ants, and the best selection found is reported. Variables that contribute to good solutions (small L) end up with larger weights. Thus, these variables tend to be selected more often, and the overall quality of solutions increases as the simulation progresses. Because of its stochastic nature, this process should be repeated several times to minimize the likelihood of accidental convergence to a poor local minimum.

In this paper, we used feed-forward artificial neural networks comprised of K inputs, one output, and a single hidden layer to construct the models, and trained them using the standard error back-propagation algorithm. The fitness function was set equal to the training correlation coefficient R^2 defined by Eq. 3.

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \tilde{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}, \quad (3)$$

where N is the number of activity measurements in the training set, y_i and \tilde{y}_i are the measured and predicted activity, respectively, for case i , and \bar{y} is the average activity over all training cases. The path length $L(R^2)$ was calculated using Eq. 4.

$$\frac{1}{L(x)} = \frac{10^{5x} - 1}{10^5 - 1} \quad (4)$$

The choice of this function emerged from the consideration that the weight updates should allow distinguishing better variable subsets. The function given by Eq. 4 increases 10-fold as $x \equiv R^2$ increases by 0.2.

We demonstrate the use of the ANTSELECT algorithm on three well-studied data sets: antifilarial activity of antimycin analogues (AMA) [6], binding affinities of ligands to benzodiazepine/GABA_A receptors (BZ) [7], and inhibition of dihydrofolate reductase by pyrimidines (PYR) [8]. The feature selection problem has been addressed previously for all three datasets. In order to assess the effectiveness of the ANTSELECT algorithm, we compare its results to those of other feature selection techniques that employ neural networks for activity prediction. For each data set we performed a set of 100 independent ANTSELECT simulations, each involving 2000 ants and reporting the set of features that resulted in a model with the largest training R^2 . The weights w_k were initialized to 0.01 and the parameter Δw was set to 0.1. No evaporation was assumed. Each set of features was further evaluated by the leave-one-out (LOO) cross-validation procedure. In this procedure, one training case is removed from the training set, and a model is built using the remaining cases. The resulting model is used to predict the activity of the removed case, which is then returned to the training set. After repeating this procedure for each training case, the predicted activity values are used to calculate the Pearson correlation coefficient R_{LOO} that serves as a measure of the quality of the feature selection.

Because the construction of a neural network model is inherently non-deterministic, a given subset of variables may result in models of varying quality. To take this uncertainty into account, we performed 10 runs of the LOO cross-validation for each of the 100 reported subsets of variables. The subset yielding the largest average value of R_{LOO} , was considered of the best quality. Another approach to define the best variable selection could be to pick the subset that exhibited the maximum value of R_{LOO} in any one cross-validation run, however, such an

approach would lead to irreproducible results. For a better evaluation of the variation in R_{LOO} for the top selected subset, we performed an additional series of 100 LOO cross-validation runs.

To assess the effectiveness of the ANTSELECT algorithm, we compare the Pearson correlation coefficient R_{LOO} obtained by the leave-one-out cross-validation of the neural network models constructed using the variables selected by ANTSELECT to that of the models constructed using the variables selected by other algorithms as reported in the literature [9,10,11]. In each case, the number of variables to be selected and the number of hidden neurons in the neural network was chosen to be the same as those reported for a similar algorithm for each of the datasets. The details of the datasets and of the neural network topology that was employed are summarized in Table 1.

All programs were implemented in the C++ programming language and are part of the DirectedDiversity® [12] software suite. They are based on 3-Dimensional Pharmaceuticals' Mt++ class library [13] and are designed to run on all Posix-compliant Unix and Windows platforms. Parallel execution on systems with multiple CPUs is supported through the multi-threading classes of Mt++. All calculations were carried out on a Dell workstation equipped with two 733 MHz Pentium III Intel processors running Windows NT 4.0.

RESULTS AND DISCUSSION

Selections of variables for each dataset for the network topology (see Table 1) reported in the literature [9-11] along with the reported value of the cross-validated Pearson correlation coefficient R_{LOO} are presented in Table 2. The top three selections produced by the ANTSELECT algorithm are shown in Table 3. The notation for the variables used in datasets AMA, BZ, and PYR can be found in Refs. [6], [7], and [8], respectively. Qualitatively, ANTSELECT algorithm reported variable selections that are either the same as the selections reported previously or

overlap with them in many variables. Two of the top three subsets found by ANTSELECT for the AMA datasets are the same as reported in Ref. [9]. The sets of variables found for the BZ and PYR dataset also show a significant overlap with those reported in Refs. [10] and [11], respectively.

The results of 100 additional LOO cross-validation runs for the top variable selection found by ANTSELECT are shown in Table 4. The average R_{LOO} for all three datasets decreased slightly compared to the original 10 cross-validations used to choose the best subset of variables due to a few poor cross-validation results, while the maximum found R_{LOO} increased. This behavior can be explained by more occurrences of very good and very bad cross-validation results as the number of cross-validation attempts increases.

Because the specific implementations of the neural network training algorithms vary and the training process is non-deterministic, it is not informative to directly compare the reported values of R_{LOO} in Table 2 to the values obtained through our implementation (it is not even clear whether the reported values correspond to an average R_{LOO} or to the highest R_{LOO} found in a number of cross-validation attempts). Instead, we performed the procedure described above for the ANTSELECT selections, i.e., 100 leave-one-out cross-validations using our neural network implementation, for the top variable selection reported for each dataset. The results are presented in Table 5.

The choice of the top variable selections provided by the ANTSELECT algorithm was based on the average value of R_{LOO} over 10 cross-validations. Alternatively, one could choose the top selections based on the maximum R_{LOO} among the 10 cross-validations. For the AMA dataset, the top variable selection based on the maximum R_{LOO} was the same as that based on the average. However, for the BZ and PYR datasets, a different set of variables was ranked first. The results of the 100 cross-validation runs for these sets of variables are shown in Table 6.

Tables 4-6 are summarized in Figure 1. The top solutions found by the ANTSELECT algorithm for the AMA and BZ datasets are of virtually the same quality as those reported in

Refs. [9] and [10]. Both the average and the maximum R_{LOO} values for the ANTSELECT solutions are very close to the corresponding values of the solutions reported previously. However, the ANTSELECT approach found a better quality set of variables for the PYR dataset.

<Figure 1>

The ANTSELECT algorithm is effective if (a) the ants sufficiently sample all variables and (b) accumulation of “pheromone deposits” (weights) clearly distinguishes between “good” and “bad” selections. The first condition is met by using a sufficient number of ants in each ANTSELECT run. The dependence of the training correlation coefficient R^2 for the AMA data set on the number of ants is presented in Figure 2. Each point corresponds to the training R^2 of the model reported by ANTSELECT averaged over 50 independent ANTSELECT runs. The error bars indicate the standard deviation of the training R^2 . The average R^2 increases with the number of ants used in the simulation and approaches a constant value, while the standard deviation decreases. Other datasets exhibited similar behavior.

<Figure 2>

The second condition can be satisfied by selecting an appropriate path length function $L(x)$, initial weights w_k and their increments Δw . Another possible approach that improves convergence by gradually increasing the probability of choosing a descriptor with relatively large corresponding weights is to use probabilities $p_k = w_k^\alpha / \sum_k w_k^\alpha$ instead of those given by Eq. 2.

The parameter α is increasing linearly (or according to some other law) with each ant between 1 and some maximum value (e.g., 2 or 3).

In conclusion, we developed a novel stochastic algorithm, ANTSELECT, for optimal variable selection in QSAR and showed that it can find solutions of the same or better quality as other methods. More detailed analysis of the algorithm’s features will be reported elsewhere. We also demonstrated that for any selected set of features the cross-validated correlation coefficient is not deterministically defined, at least when the training method is inherently stochastic. It is

therefore desirable that for such methods some statistical measure of the variation in the correlation coefficient is reported. ANTSELECT method can be further generalized to use other machine learning techniques for the model construction. Other generalizations may include the use of “pheromone deposits” to mark pairs of variables in the artificial ant’s “path” and different pheromone update formulations.

ACKNOWLEDGMENTS

We thank Dr. Raymond F. Salemme of 3-Dimensional Pharmaceuticals, Inc. for his insightful comments and support of this work.

REFERENCES

- [1] Dorigo M. and Gambardella, L. M. (1997). Ant colony system: A cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation*, **1**, 53-66, and references therein.
- [2] Beckers, R., Deneubourg, J. L. and Goss, S. (1992) Trails and U-turns in the selection of the shortest path by the ant *Lasius Niger*. *J. Theor. Biol.*, **159**, 397-415.
- [3] Dorigo, M., Maniezzo, V. and Colorni, A (1991). Positive feedback as a search strategy. Technical Report 91-016, Dipartimento di Elettronica, Politecnico di Milano, IT.
- [4] Dorigo, M., Caro, G. and Gambardella, L. M. (1999). Ant algorithms for discrete optimization. *Artificial Life*, **5**, 137-172.
- [5] Izrailev, S. and Agrafiotis, D. K. (2001). A novel method for building regression tree models for QSAR based on artificial ant colony systems. *J. Chem. Inf. Comput. Sci.*, **41**, 176-180.
- [6] Selwood, D. L., Livingstone, D. J., Comley, J. C. W., O'Dowd, A. B., Hudson, A. T., Jackson, P., Jandu, K. S., Rose, V. S. and Stables, J. N. (1990). Structure-activity relationships of antifilarial antymycin analogues: A multivariate pattern recognition study. *J. Med. Chem.*, **33**, 136-142.
- [7] Maddalena, D. J. and Johnson, G. A. R. (1995). Prediction of receptor properties and binding affinity of ligands to benzodiazepine/GABA_A receptors using artificial neural networks. *J. Med. Chem.*, **38**, 715-724.
- [8] Hirst, J. D., King, R. D., and Sternberg, M. J. E. (1994). Quantitative structure-activity relationships: Neural networks and inductive logic programming compared against statistical methods: I, the inhibition of dihydrofolate reductase by pyrimidines. *J. Comp.-Aided Mol. Design*, **8**, 405-420.
- [9] So, S.-S. and Karplus, M. (1996). Evolutionary optimization in quantitative structure-activity relationships: An application of genetic neural networks. *J. Med. Chem.*, **39**, 1521-1530.

- [10] So, S.-S. and Karplus, M. (1996). Genetic neural networks for quantitative structure-activity relationships: Improvements and application of benzodiazepine affinity for benzodiazepine/GABA_A receptors. *J. Med. Chem.*, **39**, 5246-5256
- [11] Kovalishyn V. V., Tetko, I. V., Luik, A. I., Kholodovych, V. V., Villa, A. E. P. and Livingstone, D. J. (1998). Neural network studies. 3. Variable selection in the cascade-correlation learning architecture. *J. Chem. Inf. Comput. Sci.*, **38**, 651-659.
- [12] Agrafiotis, D.K., Bone, R.F., Salemme, F.R., and Soll, R.M., United States Patents 5,463,564 (1995); 5,574,656 (1996); 5,684,711 (1997); and 5,901,069 (1999).
- [13] Copyright © 3-Dimensional Pharmaceuticals, Inc., 1994-2001.

Table 1. Data set size and the NN topology used^a

Data set	N	M	K	H	Reference
AMA	31	53	3	3	[9]
BZ	57	42	6	2	[10]
PYR	74	27	6	2	[11]

^a N – number of measurements; M – number of descriptors; K and H - number of input and hidden neurons, respectively.

Table 2. Top models reported in the literature for the given NN topology

Data set	Selected variables	R_{LOO}	Reference
AMA	NSDL3, MOFI_Y, LOGP	0.866	[9]
	NSDL8, MOFI_Y, LOGP	0.860	
	NSDL3, MOFI_Z, LOGP	0.850	
BZ	$\pi_7, F_7, \pi_1, \sigma_{m2}, \mu_6, \pi_8$	0.932	[10]
	$\pi_7, \sigma_{m7}, MR_1, \sigma_{m2}, MR_6, \sigma_{m8}$	0.929	
	$\pi_7, \sigma_{m7}, MR_1, \sigma_{m2}, \sigma_{p2}, \sigma_{p6}$	0.928	
PYR	SZ ₃ , FL ₃ , SZ ₄ , FL ₄ , SZ ₅ , HA ₅	0.82	[11]

Table 3. Top models selected by the ANTSELECT algorithm^a

Data set	Selected variables	Ave R_{LOO}	$\sigma(R_{LOO})$	Max R_{LOO}
AMA	NSDL8, MOFI_Y, LOGP	0.838	0.010	0.856
	NSDL8, MOFI_Z, LOGP	0.826	0.007	0.840
	MOFI_Y, LOGP, SUM_F	0.807	0.023	0.847
BZ	$\mu_7, \pi_7, \sigma_{m7}, MR_1, R_1, \mu_2$	0.890	0.009	0.904
	$\mu_7, \pi_7, F_7, MR_1, \sigma_{p1}, \sigma_{m2}$	0.888	0.019	0.908
	$\mu_7, \pi_7, F_7, MR_1, \sigma_{m2}, \pi_6$	0.885	0.011	0.901
PYR	SZ ₃ , FL ₃ , Hd ₃ , ΠA_3 , SZ ₅ , HA ₅	0.796	0.005	0.803
	Hd ₃ , ΠA_3 , FL ₄ , PO ₄ , SZ ₅ , HA ₅	0.786	0.035	0.832
	SZ ₃ , FL ₃ , ΠA_3 , SZ ₅ , Hd ₅ , HA ₅	0.785	0.037	0.834

^a Results are based on the highest average R_{LOO} over 10 cross-validation runs.

Table 4. Models with the highest average R_{LOO} selected by the ANTSELECT algorithm^a

Data set	Selected variables	Ave R_{LOO}	$\sigma(R_{LOO})$	Max R_{LOO}
AMA	NSDL8, MOFI_Y, LOGP	0.835	0.006	0.856
BZ	$\mu_7, \pi_7, \sigma_{m7}, MR_1, R_1, \mu_2$	0.885	0.012	0.909
PYR	SZ ₃ , FL ₃ , Hd ₃ , ΠA_3 , SZ ₅ , HA ₅	0.786	0.026	0.818

^a Results are based on 100 cross-validation runs.**Table 5.** Models with the highest R_{LOO} reported in the literature for the given NN topology^a

Data set	Selected variables	Ave R_{LOO}	$\sigma(R_{LOO})$	Max R_{LOO}
AMA	NSDL3, MOFI_Y, LOGP	0.837	0.012	0.866
BZ	$\pi_7, F_7, \pi_1, \sigma_{m2}, \mu_6, \pi_8$	0.881	0.011	0.905
PYR	SZ ₃ , FL ₃ , SZ ₄ , FL ₄ , SZ ₅ , HA ₅	0.726	0.013	0.759

^a Results are based on 100 cross-validation runs.**Table 6.** Models with the highest maximum R_{LOO} selected by the ANTSELECT algorithm^a

Data set	Selected variables	Ave R_{LOO}	$\sigma(R_{LOO})$	Max R_{LOO}
AMA	NSDL8, MOFI_Y, LOGP	0.835	0.006	0.856
BZ	$\pi_7, R_7, \sigma_{m7}, MR_1, \pi_2, \sigma_{p2}$	0.877	0.020	0.919
PYR	SZ ₃ , FL ₃ , ΠA_3 , SZ ₅ , Hd ₅ , HA ₅	0.754	0.045	0.852

^a Results are based on 100 cross-validation runs.

CAPTIONS TO FIGURES

1. The average, standard deviation, and the maximum of the cross-validated Pearson correlation coefficient over 100 leave-one-out cross-validation runs for the top feature selections reported in the literature and made by ANTSELECT on the basis of the highest maximum R_{LOO} and the highest average R_{LOO} .
2. The average and standard deviation of the training correlation coefficients R^2 of the neural network models reported by ANTSELECT for the AMA dataset. The averaging was performed over 50 independent ANTSELECT runs.

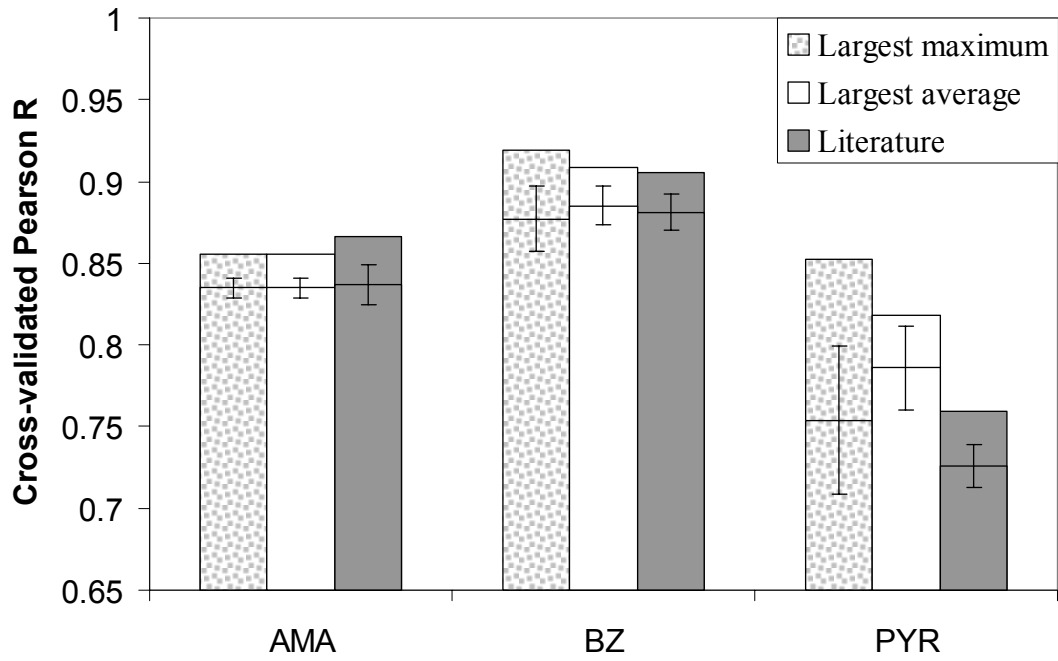


Figure 1

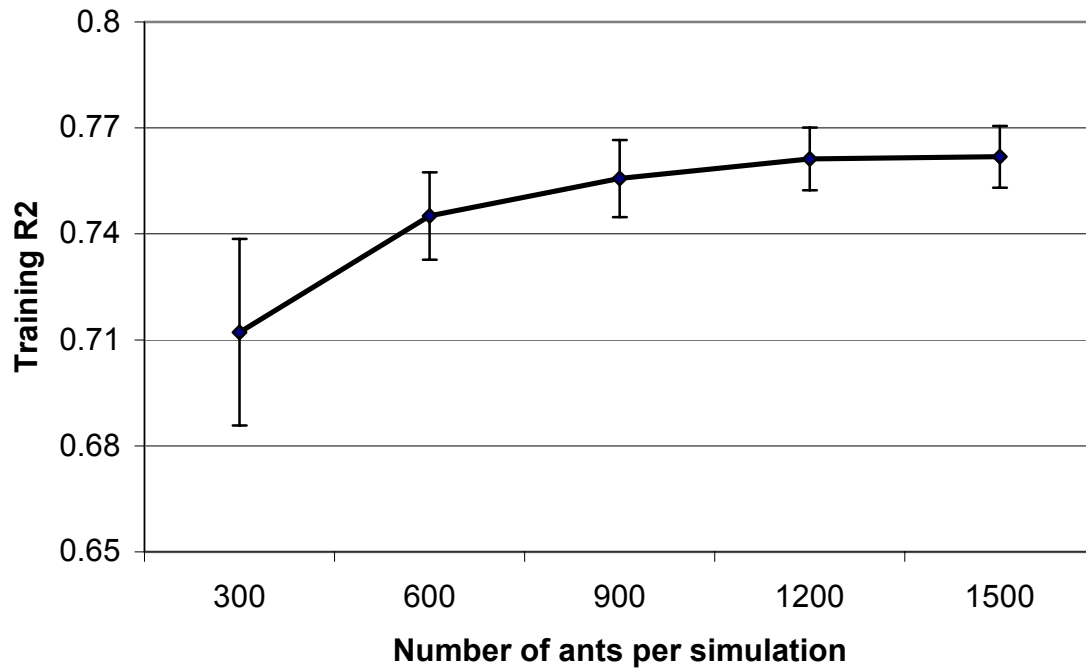


Figure 2