Genetic Programming for Human Oral Bioavailability of Drugs

Francesco Archetti, Stefano Lanzeni, Enza Messina, Leonardo Vanneschi
Dipartimento di Informatica, Sistemistica e Comunicazione (D.I.S.Co.)
University of Milano-Bicocca
Milan, Italy
{archetti, lanzeni, messina, vanneschi}@disco.unimib.it

ABSTRACT

Automatically assessing the value of bioavailability from the chemical structure of a molecule is a very important issue in biomedicine and pharmacology. In this paper, we present an empirical study of some well known Machine Learning techniques, including various versions of Genetic Programming, which have been trained to this aim using a dataset of molecules with known bioavailability. Genetic Programming has proven the most promising technique among the ones that have been considered both from the point of view of the accurateness of the solutions proposed, of the generalization capabilities and of the correlation between predicted data and correct ones. Our work represents a first answer to the demand for quantitative bioavailability estimation methods proposed in literature, since the previous contributions focus on the classification of molecules into classes with similar bioavailability.

Categories and Subject Descriptors

I.2 [**ARTIFICIAL INTELLIGENCE**]: Automatic Programming; D.2.8. [Software Engineering]: Metrics - *complexity measures, performance measures.*

General Terms: Algorithms.

Keywords: Bioinformatics, Bioavailability, Molecular Descriptors, Genetic Programming

1. INTRODUCTION

In recent years the introduction of high throughput screening (HTS) and combinatorial chemistry techniques has deeply changed the process of drug discovery. Libraries of millions of chemical compounds could now be tested in order to evaluate their affinity to a particular pathology-associated protein. Results of test could then be used to design modifications to be done on the molecules for optimizing their properties. Nevertheless this is not enough, in fact compounds with putative pharmacological value have not only to show a good target binding, but also have to reach the target *in vivo*. In other words, it is necessary that compounds follow the appropriate way into the human body without altering the health of the patient. About half of the failures in pharmacological development were made at this stage [13] (see figure 1), with an unacceptable burden on the research and development budget of pharmaceutical companies.

For this reason, the behavior of the molecules must be evaluated through the so-called ADMET processes (Adsorption, Distribution, Metabolism, Excretion and Toxicity). Some parameters directly correlated with ADMET processes are the Human Oral Bioavailability (which is the main subject discussed in this paper), the Blood Brain Barrier penetration, the plasma protein binding level, the dermal and ocular permeation. Various medium and high-throughput in vitro screens are therefore now in use for predict ADMET parameters and there is an increasing need for good tools for predicting these properties. This allows to serve two aims: first at the design stage of new compounds and compound libraries so as to reduce the risk of late-stage attrition; and second, to optimize the screening and testing by looking at only the most promising compounds [21].

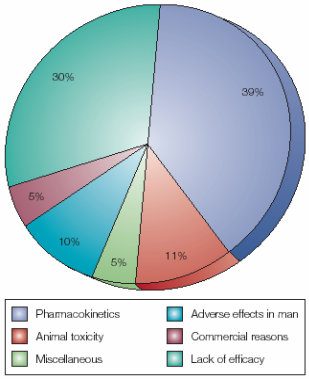

Figure 1-Failures in drug research and development as discussed in [13]. The 50% of the failures are correlated with bad ADMET parameters. Other reasons are human toxicity, poor target binding and commercial problems.

In this paper we show that Genetic Programming (GP), an established technique of evolutionary computing, is a promising and valuable tool for quantitative predictions of human oral bioavailability of drug candidates.

Human oral bioavailability (indicated with %F from now on) is the parameter that measures the percentage of initial drug dose that effectively reaches the systemic blood circulation. This parameter is particularly relevant for pharmaceutical industries, because the oral assumption is usually the preferred way for supplying drugs to patients and because it is a representative measure of the quantity of active principle that effectively can

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.

actuate its biological effect. Oral bioavailability is determined by two keys ADMET processes: adsorption and metabolism. In fact orally submitted drugs, as depicted in figure 2, have to be absorbed from the gut wall and to enter into systemic circulation in the portal vein; carried by the blood flux, molecules arrive in the liver, where there are some biochemical processes that try to demolish them. The percentage of molecules initially submitted that exit the liver and enter the blood circulation corresponds to the Oral Bioavailability of a particular compound.

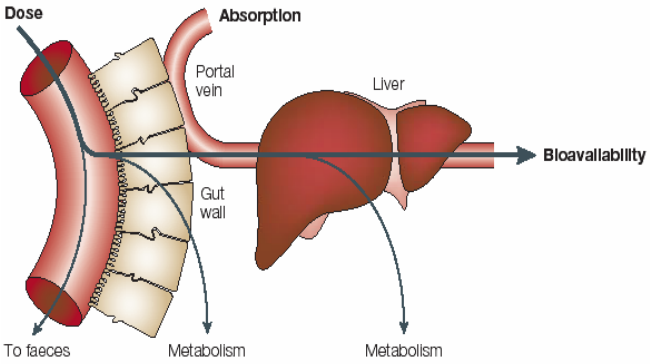

Figure 2- Anatomical view of processes affecting bioavailability. From the stomach drugs start digestion and pass, via gut wall absorption, into the portal vein. Here they enter the liver, where metabolism tries to demolish them. The fraction of initial dose that exits the liver represents drug bioavailability.

This paper is structured as follows: in section 2 we describe the statistical methods mostly employed in relevant literature state for bioavailability predictions. In section 3 we describe the protocol adopted in our calculation, defining all the parameters needed for our results reproduction. Section 4 focuses on the obtained experimental results, whereas Section 5 discusses the results and delineates the future investigations in this field.

2. STATE OF THE ART AND RELATED WORK

Predicting Human Oral Bioavailability is not an easy task because, as depicted above it depends on a superposition of two ADMET processes: absorption and liver first-pass metabolism. Absorption in turn depends on the solubility and permeability of the compounds, as well as interactions with transporters and metabolizing enzymes in the gut wall. Important properties for determining permeability seem to be the size of the molecule, as well as its capacity to make hydrogen bonds, its overall lipophilicity and possibly its shape and flexibility. Molecular flexibility, for example, as evaluated by counting the number of rotatable bonds, has been identified as a factor influencing bioavailability in rats [24-26].

Some attempts in estimating bioavailability are reported in literature and all belong to the category of Quantitative Structure Activity Relationship (also called Q.S.A.R.) studies [25]. Models developed with Q.S.A.R. approach are quantitative regression methods that attempt to relate chemical structure to biological activity. Quantitative structure–activity relationship modeling generally involves three steps: (a) to collect or, if possible, to design a training set of chemicals compounds for which the biological parameter to estimate is known; (b) to derive features (descriptors) of the molecular structure that can properly relate to biological activity; and (c) to apply methods to build a mathematical relationship that permits to calculate biological activity. Obtaining a good-quality QSAR model with the ability to predict the activity of a chemical compound outside the training set depends on many factors such as the quality of data, and the choice of significant features.

Molecules could be represented as graphs with labeled nodes (atoms) and labeled edges (bonds between atoms). There are two main categories of molecular features that can be used for Q.S.A.R.: 2D-chemical descriptors, that refer to the bi-dimensional structure of compounds and 3D-chemical descriptors, that are calculated starting from tri-dimensional molecular conformations. For an extended discussion about molecular descriptors see [22].

Yoshida and Topliss [28] trained a QSAR model with the presence/absence of typical functional groups most likely to be involved in metabolic reactions as the structural input. Their approach used ‘fuzzy adaptive least squares’, and drugs could be classified into one of four predefined bioavailability ranges. Using this approach, a new drug can be assigned to the correct class with an accuracy of 60%. A method using adaptive fuzzy partitioning (AFP) has been presented in [18]. Genetic Algorithms were used to select the best molecular descriptors, and Self Organizing Maps (SOMs) were used to collocate the molecule in a bioavailability class. Fröhlich and Wegner recently experimented the use of Kernel methods (SVM) for assessing the problem of bioavailability predictions, basing their approach on the estimation of similarity between different molecules with similar biological behavior [8]. Various kinds of multivariate and Partial Least Square (PLS) regressions, also coupled with recursive partition [2], have been used to give an estimation of oral bioavailability.

Artificial Neural Networks (ANN) are widely used for ADMET parameters estimation (for a detailed discussion see [29]), and there are also some contributions in which a Bayesian Regularized Artificial Neural Networks (BRANN) has been applied.

Some software vendors are active in the field of pharmaco-dynamical predictions, particularly those traditionally involved in the field of molecular modeling. For example Accelrys Inc [1], and Pharma Algorithms Inc.[17] offer some integrated Q.S.A.R. utilities also for bioavailability modeling, whereas Simulation Plus Inc. [20] uses bagged Artificial Neural Networks for estimating the key ADMET parameters and then uses them to perform a simulation of the absorption process.

GP has been applied for grouping molecules into 4 classes of increasing levels of bioavailability without specifically focusing on the quantitative predictions by Langdon and Barrett in [15]. In this paper, we take up a different perspective: we try to quantitatively asses the value of bioavailability instead of grouping molecules into classes with similar biological activity.

3. PROTOCOLS DESCRIPTION

Many machine learning techniques have been tested to compare their ability in quantitatively predicting oral bioavailability. Here we present a "canonic" version of GP, three variants of GP in which the fitness function and/or the set of terminal symbols have been modified and some well known regression methods such as: Linear and Least Square Regression, Feed Forward Artificial Neural Networks and two different versions of Support Vector

Machines (SVM). Before describing these Machine Learning methods and the way in which they have been used in our experiments and presenting the experimental results, in the following sections we describe data collecting and preparation, the computation of molecular descriptors and the strategy used for dataset partitioning.

3.1 Dataset Collecting and Preparation

We collect from the relevant literature [28] and from DrugBank public database [7] the chemical structure, expressed as SMILES (Simplified Molecular Input Line Entry Specification) string, and the human oral bioavailability experimental measurements for 360 FDA approved drugs and drug-like compounds. SMILES is a string codifying the 2D molecular structure of a compound in an extremely concise form, introduced by Chemical Information Systems Inc. [6]. Chemical strings are transformed into bi-dimensional formulas and used as input for the ADMET Predictor (a software produced by Simulation Plus Inc. [20]) for calculating 241 bi-dimensional molecular descriptors.

Now we dispose of a matrix H of m rows and n columns, where m represents the number of molecules for which we have the experimental bioavailability measurements, and n represents the number of molecular descriptors. The known values of human oral bioavailability are placed in the m-dimensional vector %F. These data structures and the relative descriptors can be downloaded from the webpage: <*omitted to keep anonymity*>.

3.1.1 Dataset splitting

A random splitting of the dataset is performed before model construction, by partitioning it into a training and a test set: 70% of the molecules are randomly selected with uniform probability and inserted into the training set, while the remaining 30% forms the test set. In other words, H is spitted into $H^{(TRAIN)}$ and $H^{(TEST)}$. The first of these two matrix has m^+ rows (where m^+ is the number of molecules selected for constructing the training set) and n columns, whereas $H^{(TEST)}$ has m-m^+ rows and n columns. Analogously, also the vector %F is partitioned in $\%F^{(TRAIN)}$ and $\%F^{(TEST)}$, where $\%F^{(TRAIN)}$ of course contains the same indexes of %F as the ones in $H^{(TRAIN)}$.

3.2 Genetic Programming Settings

Four (slightly) different versions of GP have been used to obtain the results presented in this paper. They are described below.

3.2.1 "Canonic" (or standard) GP

The first GP setting that we have used (called canonic, or standard, GP from now on and indicated as stdGP) was a deliberately simple version of standard tree-based GP [14]. In particular, we have chosen to use a parameter setting and the sets of functions and terminal symbols as much similar as possible to the ones that have originally been used in [14] for symbolic regression problems. Each molecular feature H_{ij} has been represented as a floating point number. Potential solutions (GP individuals) have been built by means of the set of functions F={+, *, -,÷} (where ÷ is the protected division, i.e. it returns 1 if the denominator is zero) and the set of terminals T composed by n floating point variables (where n is the number of columns in the training set, i.e. the number of molecular features of the compounds). The fitness of each individual has been defined as the root mean squared error (RMSE) measured on the data used to construct the bioavailability model, i.e. *only* the data contained in the training set have been used as fitness cases. In other words, given an individual k producing the bioavailability predictions $\%F^{(PRE)}$ on the training set, we define the fitness of k ($RMSE_k$) as

$$RMSE_k = f(k) = \sqrt{\frac{\sum_{i=1}^{m^+}(\%F_i - \%F_i^{(PRE)})^2}{m^+}} \qquad (1)$$

For each individual k, we also evaluate the RMSE measured on the test set $H^{(TEST)}$, whose data, of course, are not used at all during the evolution. The RMSE on the test set will be used for comparing GP results with the ones of the other methods. GP parameters used in our experiments are summarized in table 1.

table 1- Genetic Programming experimental setting used for the experiments

Objective	Evolve a quantitative predictive model of human bioavailability
Function set:	Multiplication, Addition, Division, Subtraction
Terminal set:	241 molecular descriptors (floating point variables)
Fitness:	RMSE measured on drugs selected for training
Selection:	tournament of size 10
Algorithm:	Generational GP with elitism (i.e. copy of the best individual in the next population at each generation)
Pop Size:	500 individuals
Initialization:	ramped half and half with maximum depth equal to 6
Other Parameters:	maximum depth for crossover 17, swap mutation probability 0.1, shrink mutation probability 0.1, sub-tree mutation probability 0.1 [14]
Maximum Number of generations:	500

3.2.2 LS2-GP

The second version of GP that we present uses the same parameter setting as stdGP, but a different fitness function. In particular, the fitness of a GP individual k is obtained by executing two steps. The first step consists in applying linear scaling to $RMSE_k$ as it has been defined in equation (1). The use of linear scaling is by no means new to GP: among others, it has been successfully applied to many difficult symbolic regression problems in [12]. It consists in calculating the slope and intercept of the formula coded by the GP individual. Given that $\%F^{(PRE)} = k(H_{i\bullet})$ is the output of the GP individual k on the input data $H_{i\bullet}$, a linear regression on the target values $\%F$ can be performed using the equations:

$$b = \frac{\sum_{i=1}^{m+}\left[\left(\%F_i - \overline{\%F}\right)\left(\%F_i^{(PRE)} - \overline{\%F}^{(PRE)}\right)\right]}{\sum_{i=1}^{m+}\left(\%F_i^{(PRE)} - \overline{\%F}^{(PRE)}\right)} \qquad (2)$$

$$a = \overline{\%F} - b\overline{\%F}^{(PRE)}$$

where m^+ is the number of cases (i.e. the number of lines in the training set) and $\overline{\%F}^{(PRE)}$ and $\overline{\%F}$ denote the average output and the average target value respectively. These expressions respectively calculate the slope and intercept of a set of outputs $\%F^{(PRE)}$, such that the sum of the squared errors between $\%F$ and $a + b\%F^{(PRE)}$ is minimized. After this, any error measure can be calculated on the scaled formula $a + b\%F^{(PRE)}$, for instance the RMSE:

$$RMSE_k\left(\%F, a + b\%F^{(PRE)}\right) = \sqrt{\frac{\sum_{i=1}^{m^+}(a + b\%F_i^{(PRE)} - \%F_i)^2}{m^+}} \quad (3)$$

If a is different from 0 and b is different from 1, the procedure outline above is guaranteed to reduce the RMSE for any formula $\%F^{(PRE)} = k(H_{i\bullet})$ [12]. The cost of calculating the slope and intercept is linear in the size of the training set. By efficiently calculating the slope and intercept for each individual, the need to search for these two constants is removed from the GP run. GP is then free to search for that expression whose *shape* is most similar to that of the target function. The efficacy of linear scaling in GP for many symbolic regression problems has been widely demonstrated in [12].

The second step that has to be performed to calculate the fitness value of individual *k* consists in calculating the following function:

$$f(k) = \frac{w_1 RMSEN_k + w_2 CORN_k}{2} \quad (4)$$

where $RMSEN_k$ is the normalized value of $RMSE_k$ as it has been defined in equation (3) and $CORN_k$ is the normalized value of $CORR_k$ which is the statistic correlation between the real bioavailability values of the molecules belonging to the training set and the bioavailability predictions calculated by *k* on the same molecules. Formally, $CORR_k$ is defined as follows:

$$CORR = \frac{C_{\%F,\%F^{(PRE)}}}{\sigma_{\%F}\sigma_{\%F^{(PRE)}}} \quad (5)$$

where:

$$C_{\%F,\%F^{(PRE)}} = \frac{1}{n}\sum_{i=1}^{m-m^+}\left(\%F_i - \overline{\%F}\right)\left(\%F_i^{(PRE)} - \overline{\%F}^{(PRE)}\right) \quad (6)$$

is the covariance of the vectors $\%F$ and $\%F^{(PRE)}$, $\sigma_{\%F}$ and $\sigma_{\%F^{(PRE)}}$ are the standard deviations of these two vectors and $\overline{\%F}$, $\overline{\%F}^{(PRE)}$ are their averages respectively.

Before calculating $f(k)$, both $RMSE_k$ and $CORR_k$ have been normalized into the range [0, 1], in such a way that they have the same "importance" in the calculation of the weighted average. In particular, $CORR_k$ whose values are by definition included into the range [-1, 1] have been normalized by applying:

$$CORN_k = \frac{-CORR_k + 1}{2} \quad (7)$$

after this calculation, if $CORN_k = 0$, then there is a perfect correlation between the true bioavailability values and the calculated ones and if $CORN_k = 1$ then these values are not correlated at all.

In our experiments, we have set the weights values as: $w_1 = 0.4$ and $w_2 = 0.6$, i.e.a slightly higher importance in the fitness calculation has been assigned to the correlation coefficient. A simple experimentation phase (whose results are not shown here for lack of space) has empirically shown that these values are the ones for which the GP system has the best generalization ability.

The idea behind this weighted sum is that optimizing only the RMSE on the training set may lead to overfitting and thus to a poor generalization power of GP solutions (i.e. bad results on the test set). If we optimize more than one criterium, GP probably returns an individual which is good on all the criteria even though not optimal for all of them. Furthermore, the correlation coefficient between outputs and targets is a very important measure for results accuracy and thus deserves to be used as an optimization criterium.

This GP version will be called "Linear Scaling with 2 criteria" GP or LS2-GP from now on.

3.2.3 LS2-C-GP

The third GP version presented in this paper is similar to LS2-GP with the only difference that a set of ephemeral random constants (ERCs) is added to the set of terminal symbols to code GP expressions. These ERCs are generated uniformly at random from the range [*m*, *M*], where *m* and *M* are the minimum and the maximum values of bioavailability of the molecules in the training set respectively. In the experiments presented in this paper, a number of ERCs equal to the number of variables (i.e. equal to 241) has been used.

This choice has been empirically confirmed to be suitable by a set of GP runs in which different numbers of ERCs extracted from different ranges have been used. The results of these experiments are not shown here for lack of space.

This version of GP will be called "LS2 with Constants" GP and indicated as LS2-C-GP.

3.2.4 DF-GP

The fouth version of GP presented in this paper differs form the previously presented ones since this time the fitness function used by GP dynamically changes during the evolution. In particular, the evolution starts with the correlation coefficient used as the only optimization criterium. When at least the 10% of the individuals in the population reach a value of the correlation coefficient which is largest or equal to 0.6, the fitness function changes, and it becomes the following one:

$$f(k) = \begin{cases} bad\_fitness & if \quad CORR_k < 0.6 \\ RMSE_k & otherwise \end{cases} \quad (8)$$

in this way, the selection pressure operates as a pruning algorithm, giving a chance to survive for mating only to those individuals whose correlation is largest or equal to 0.6.

The idea behind this method is that the search space is too large for GP to perform efficiently; furthermore, we hypothesize that individuals with a good, although not optimal, correlation coefficient between outputs and goals will have a largest generalization ability and thus should take part in the evolutionary process. Some experiments (whose results are not shown here for lack of space) have empirically confirmed that the threshold value 0.6 for the correlation coefficient is large enough to avoid

underfitting and small enough to reduce overfitting. Of course, this value has been tested and has revealed suitable only for the dataset used in this paper and can by no means be interpreted as a general threshold. Nevertheless, the experiments that we have executed to obtain this value are very simple and if we wish to evolve new expressions on new data we could easily replicate them. This GP version has been called Dynamic Fitness GP and will be indicated as DF-GP from now on.

3.3 Other Methods

Various non evolutionary regression models were used, in order to comparatively evaluate the regression performances of the individuals generated by GP. The statistical methods most commonly used to perform human oral bioavailability prediction, as discussed in section 2, were trained and tested. They are described below in a deliberately synthetic way, since they are well known and well established machine learning techniques. For more details on these methods and their use, the reader is referred to the respective references quoted below.

3.3.1 Feature Selection

In order to improve the performances of the methods described in the following part of this section and to make more meaningful comparisons with GP results, we performed feature selection on the training set. We adopted two attribute selection heuristics: the Correlation based Feature Selection (CFS) and the Principal Component based Feature Selection implemented by Hall [9].

The central hypothesis in CFS is that good feature sets contain features highly correlated with the class, yet uncorrelated with each other. A feature evaluation formula, based on ideas from test theory, provides an operational definition of this hypothesis. The algorithm couples the evaluation formula with an appropriate correlation measure and a search strategy.

Principal Component Analysis based Feature Selection reduces the dimensionality of attribute space transforming the original features in a new set of variables called principal components (PCs). The latter are uncorrelated and ordered so that the first few retain most of the variation present in all of the original variables [11]. We calculate the PCs for our dataset, and then use all the new variables for the methods training.

3.3.2 Linear and Least Square Regression

We used the Akaike criterion for model selection (AIC) [3], that has the objective of estimating the Kullback-Leibler information between the densities, corresponding to the fitted model and the generating or true model. After eventually applying one of the methods described in section 3.3.1, M5 criterion is used for further attribute selection [3]. The least square regression model is founded on the algorithm of Robust regression and outlier detection described in [19], searching for the more plausible linear relationship between outputs and targets.

3.3.3 Artificial Neural Networks (ANN)

The multilayer perceptron [10] implementation included in the Weka software distribution [27] was adopted. It uses the Backpropagation algorithm [10] with a learning rate equal to 0,3. All the neurons have a sigmoid activation function. All the other parameters that we have used have been set to the defaults values proposed by the Weka implementation [25].

3.3.4 Support Vector Machines Regression (SVMR)

The Alex J. Smola and Bernhard Scholkopf sequential minimal optimization algorithm [21] was adopted for training a Support Vector regression using polynomial kernels. In particular we have built two models using polynomial kernels of first and second degree respectively.

4. EXPERIMENTAL RESULTS

Table 2 shows the RMSE and correlation coefficient for all the presented non evolutionary machine learning techniques without feature selection. The best RMSE result is returned by SVM regression with first degree polynomial kernel, while the best correlation coefficient has been found by multi-layer perceptron.

table 2- Experimental comparison between different machine learning techniques for bioavailability predictions without feature selection.

Method	RMSE on test set	Correlation coefficient
Linear Regression	48,1049	0,1699
Least Square Regression	37,2211	0,2022
Multi layer perceptron	51,280	0,2880
SVM Regression – first degree polynomial kernel	34,804	0,2666
SVM Regression – second degree polynomial kernel	44,323	0,2597

In table 3, we report results of the same techniques using the Principal Components based feature selection. All the new features generated by this method have been used by all the machine learning algorithms to generate the models. This table shows that the use of feature selection techniques helps to generally improve the performances of all the methods. Nevertheless, this technique is not suitable for some of the presented methods for which the performance improvement appears to be only marginal. This time the best RMSE result is returned by linear regression, while the best correlation coefficient is found by SVM regression with second degree polynomial kernel.

table 3- Experimental comparison between different machine learning techniques for bioavailability predictions using Principal Components Based feature selection (see section 3.3.1).

Method	RMSE on test set	Correlation coefficient
Linear Regression	30.5568	0.1911
Least Square Regression	40.4503	0.1165
Multi layer perceptron	48.9771	0.1945
SVM Regression – first degree polynomial kernel	36.1850	0.1306
SVM Regression – second degree polynomial kernel	42.3377	0.2184

In table 4, we present the results of the same machine learning methods using the Correlation based feature selection. This selection strategy eliminates some noisy features while maintains the ones that are correlated with bioavailability. As table 4 shows, this strategy enhances the performances of all the methods emplyed remarkably better than the Principal Components based feature selection. With the Correlation based feature selection the

best results, both for RMSE and correlation coefficient are returned by linear regression.

table 4- Experimental comparison between different machine learning techniques for bioavailability predictions using Correlation Based feature selection (see section 3.3.1).

Method	RMSE on test set	Correlation coefficient
Linear Regression	27.5212	0.3141
Least Square Regression	31.7826	0.1296
Multi layer perceptron	32.5782	0.2308
SVM Regression – first degree polynomial kernel	28.8875	0.2855
SVM Regression – second degree polynomial kernel	29.7152	0.2787

In table 5, RMSE and correlation coefficient for all the considered GP versions are shown. In particular, we report the performance of the individual with the best RMSE value contained in the population at termination over 20 independent runs. Comparing the results shown in this table with the previously presented ones, we remark that all the GP versions outperform the other machine learning methods both for RMSE and correlation coefficient, except for stdGP which is outperformed by SVM and linear regression with the Correlation based feature selection. The technique that has returned the best solution is LS2-C-GP. Comparing the results returned by LS2-C-GP with the ones of the non-evolutionary methods, we can remark that LS2-C-GP has found a better RMSE and a remarkably higher correlation coefficient value. We hypothesize that this is due to two main reasons: first of all, using two criteria to evolve solutions on the training set allows us to generate solutions which are "good" on both the criteria and that are optimal on none of them. In this way, we prevent the evolutionary process to generate too good solutions on the training set for one single criterium, which could lead to overfitting. In doing that, we also use the correlation coefficient as an optimization criterium, which is an important measure for results accuracy. Secondly, the use of ERCs may help to asses the relative relevance of the features in the proposed solutions.

table 5- Experimental results of the different GP versions. These results concern the individuals with the best RMSE value in all the populations over 20 independent runs.

Method	RMSE on test set	Correlation coefficient
stdGP	30.1276	0.1661
LS2-GP	26.5909	0.3735
LS2-C-GP	26.0126	0.4245
DF-GP	27.3591	0.3304

We remark that we have deliberately avoided to use a feature selection technique before applying the GP strategies. In fact, we have hypothesized that GP automatically performs a feature selection by keeping into the population only the individuals which use a subset of the features. Our experimental results should also demonstrate that this automatic feature selection performed by GP is competitive with some of the most known feature selection algorithms and it has not to be executed separately before running the regression algorithms.

Finally, we report the average values with their standard deviations of the best individuals RMSE on the test set over the 20 independent runs for the various GP versions (see table 6). These results confirm that GP solutions are stable (i.e. results presented in table 5 have not been found by fortuitous runs) and offer good performances with respect to the other traditional machine learning techniques.

table 6- Averages and standard deviations of the best individuals for the different GP versions over 20 independent runs.

Method	RMSE average on test set	RMSE standard deviation
stdGP	34.3480	2.7680
LS2-GP	28.0836	1.4472
LS2-C-GP	27.5241	0.6881
DF-GP	28.1353	1.1592

4.1 GP best individual properties

The genotype of the best GP individual (obtained using LS2-C-GP), whose RMSE and correlation coefficient performances have been reported in table 5 is shown in figure 3. Starting from 241 selectable features we obtained that only 17 2D-molecular descriptors are used in bioavailability estimation. In other words, GP has automatically performed a strong feature selection. This phase cannot be done automatically by all the other techniques that we have considered. In this application, feature selection plays a very important role (bioavalaibility can be expressed as a function of some molecular descriptors, not necessarily all). We claim that this is one of the reasons why GP may be a suitable technique to solve this kind of problems.

```
(* (- (% c84/x188) (* (- c206 x84) (- x224 c176))) (* (% (+ (+ x188 x34)(- c236 x26)) (- c206 x84)) (+ (+ (- (+ (* c86 c141) (- (% c22 x110) (* (*(+ c54 x234) (+ (+ c200 x126) (- (- (+ (+ c175 x240) (% (% c115 x195) c86))(- c206 x84)) (* x218 c162)))) (- x224 c176)))) (* (* (+ c54 x234) (- c206 x84)) (- x224 c176))) (- (- (+ (% c22 x110) (- (+ (- (+ (+ (- c206 x84) (*(+ x69 c211) (% c176 x218))) (% (* x218 c162) (% c84 x188))) (* (- c206 x84)(- x224 c176))) (- x206 x150)) (* (* (+ c54 x234) (+ (% c115 x195) (- (- (+(+ c175 x240) (% (% x5 c24) c86)) (* x218 c162)) (* x218 c162)))) (- x224c176)))) (* (* (+ c54 x234) (- c206 x84)) (- x224 c176))) (% (% c115 x195)c86))) (% (* x218 c162) (% (+ (- (+ c200 x126) (* x218 c162)) (- x214 x9))(* (+ c54 x234) (+ (- c206 x84) (- (+ (+ (- c206 x84) (% (% c115 x195) c86))(% (% c115 x195) c86)) (* x218 c162)))))))))
```

Figure 3- Genotype of the best GP individual generated in our experiments. (+ is addition, - is subtraction, * is multiplication, % is protected division, xi is the i-th molecular feature and cj is the j-th ephemeral random constant.

Molecular descriptors used as model input and reported in table 7 refer to molecular properties that are involved in absorption and metabolism (i.e. the two key processes determining human bioavailability). In fact, from selected descriptors screening we found d195, d214, d215, d224S which are linked to molecular polarity, and also features like d126, d110, and d150 that refer to

the chemical groups sensible for a liver effected modification. Note that d214 is one of the features required by the well known Lipinsky rule of five [16]

Table 7- Features selected as terminal leaves of best GP individual with their pharmacological meaning.

Code	Meaning
d5	Number of carbons
d9	Number of sulfurs
d26	Number of aliphatic rings
d34	atom-type E-state index for -CH2 groups
d69	atom-type E-state index for phospinate
d84	atom-type E-state index for Iodine groups
d110	number of thocarbon groups
d126	number of nitrites groups
d150	number of sulfate groups
d188	Iso-thio-cyanate:1.0; thio-cyanate:0.5
d195	Compounds with –OH groups attached to an aliphatic alchol
d206	Hydrocarbon compounds with at least one fluorine groups
d214	Number of hydrogen bond acceptors
d218	Number of nitrogen based hydrogen bond acceptors
d224	Sum of partial atomic charge on nitrogen based hydrogen bond acceptors
d234	cumulative contribution of purely anionic species to fraction ionized at specific pH=7.4
d240	Lown electron pair on NOSV divided by the number factor

5. DISCUSSION AND FUTURE WORKS

Automatically predicting human oral bioavailability is a very important issue, because it helps to prevent industrial failure, human toxicity and poor drug activity.

This paper has presented an experimental study of quantitative prediction of the bioavailability of some medical compounds.

Nine different machine learning techniques, among which four different versions of genetic programming, have been tested using a database of molecules with known bioavailabilities. This database has been split into a training set and a test set, in order to investigate the generalization capabilities of the considered techniques.

The genetic programming variants mainly differ in the fitness function and in the set of terminal symbols used. In particular, a fitness function optimizing both the root mean square error and the correlation coefficient have been presented; a set of ephemeral random constants have been added to the terminal symbols set, originally composed by some floating point variables. Furthermore, a version with a dynamic fitness function to keep only individuals with good correlation coefficient into the population has been presented. Finally the well known linear scaling strategy has also been applied to the root mean square error.

Some of the genetic programming versions have shown significantly better results than all the other presented techniques, both from the point of view of the quality of the solution found and of the statistical correlation between predicted and target data.

Among them, the version which optimizes both the root mean square error and the correlation coefficient on the training set has shown the best generalization capability, in particular in the case where ephemeral random constants have been used.

While all the non-evolutionary machine learning techniques have been able to produce good results only after the explicit application of some well-known (and computationally expensive) feature selection strategies, GP could potentially use all the features in the dataset and has automatically performed a strong and efficient feature selection (the best individual found in our experiments uses only 17 molecular features over the 241 contained in the dataset).

These results are encouraging and should pave the way for a deeper investigation on the capability and also other ADMET parameters of evolutionary algorithms to develop expressions to predict bioavailability as a function of its molecular descriptors.

Furthermore, according to the results that have been presented in this paper, optimizing more than one criterium on the training set should contribute to generate solutions with good generalization abilities. This subject should be further investigated in the future. In particular, multi-objective GP systems should be studied, in which not only root mean square error and correlation coefficient, but also many other optimization criteria are used. These systems may be based on more sophisticated multi-objective optimization strategies than the simple weighted average used in this paper: Pareto fronts are surely the most promising methods and they have to be investigated.

Future work also includes the study of techniques to automatically generate numeric terminals for the tree expressions in an "intelligent" way (for example with some well known techniques of local optimization [23] or coevolution [5]), instead of generating them at random.

6. ACKNOWLEDGMENTS

Our thanks to Prof. PierCarlo Fantucci for allowing us to calculate molecular descriptors using Simulation Plus software.

7. REFERENCES

[1] Accelrys Inc. , the world leader in cheminformatics for drug development. See www.accelrys.com.

[2] Andrews, C. W., Bennett, L. & Yu, L. X. Predicting human oral bioavailability of a compound: development of a novel quantitative structure-bioavailability relationship. *Pharm. Res. 17, 639–644 (2000).*

[3] Akaike, H., 2nd International Symposium on Information Theory, *Chapter Information theory and an extension of maximum likelihood principle, pp. 267-281, 1973. Akademia Kiado*.

[4] Bains W, Gilbert R, Sriridenko L et al., Evolutionary computational methods to predict oral bioavailability QSPRs, Curr Opin Drug Discov. Devel., 2002, Jan 5(1):44-51.

[5] Cagnoni S., D. Rivero and L. Vanneschi. A purely-evolutionary memetic algorithm as a first step towards symbiotic coevolution. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC 2005)*, pages 1156-1163, Edinburgh, Scotland, 2005. IEEE Press, Piscataway, NJ.

[6] Chemical Information Systems Inc., the company that introduced SMILES molecule representation. See http://www.daylight.com/dayhtml/smiles.

[7] Drug Bank, a recently developed database of FDA approved and experimental drugs. See http://redpoll.pharmacy.ualberta.ca/drugbank/.

[8] Fröhlich, J. Wegner, F. Sieker, A. Zell: Kernel Functions for Attributed Molecular Graphs - A New Similarity Based Approach To ADME Prediction in Classification and Regression, *QSAR & Combinatorial Science, 2005.*

[9] Hall, M. A. 1998. Correlation-based Feature Selection for Machine Learning. Ph.D diss. Hamilton, NZ: Waikato University, Department of Computer Science.

[10] Haykin S., *Neural Networks: a comprehensive foundation*. Prentice Hall, London, UK, 1999.

[11] I.T. Jolliffe, *Principal Component Analysis*, Second edition, Springer series in statistics.

[12] Keijzer M., Improving symbolic regression with interval arithmetic and linear scaling. In C. Ryan *et al.* editors, *Genetic Programming, Proceedings of the 6th European Conference, EuroGP 2003*, volume 2610, of LNCS, pages 71-83, Essex, 2003. Springer, Berlin, Heidelberg, New York.

[13] Kennedy, T. Managing the drug discovery development interface. *Drug Disc. Today 2, 436–444 (1997).*

[14] Koza J.R. . *Genetic Programming*. The MIT Press, Cambridge, Massachusetts, 1992.

[15] Langdon W. B. and S. J. Barrett, Genetic Programming in data mining for drug discovery, in Evolutionary computing in data mining, p. 211-235,2004 Springer.

[16] Lipinsky et al., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug Deliv. Rev.*, 23:3-25, (1997)

[17] Pharma Algorithms, a company active in the field of ADMET predictions. See www.ap-algorithms.com.

[18] Pintore, M., Van de Waterbeemd, H., Piclin, N. & Chrétien, J. R., Prediction of oral bioavailability by adaptive fuzzy partitioning, *Eur. J. Med. Chem*. 2003, Apr;38(4):427-31.

[19] Rousseeuw, Peter J, Robust regression and outlier detection / Peter J. Rousseeuw, *Annick M. Leroy, New York : Wiley, 1987.*

[20] Simulation Plus Inc., company that use both statistical methods and differential equations based simulations for ADME parameter estimation. www.simulationsplus.com.

[21] Smola Alex J., Bernhard Scholkopf (1998). A Tutorial on Support Vector Regression. *NeuroCOLT2 Technical Report Series - NC2-TR-1998-030*.

[22] Todeschini, R. & Consonni, V. *Handbook of Molecular Descriptors* (Wiley–VCH, Weinheim, 2000).

[23] Topchy A. and W. F. Punch. Faster genetic programming based on local gradient search of numeric leaf values. In L. Spector et al. editors, Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2001, pages 155-162. San Francisco, CA, 2001. Morgan Kaufmann.

[24] Van de Waterbeemd H. and Eric Gifford, ADMET in silico modeling: towards prediction paradise? *Nature Reviews Drug Discovery, MARCH 2003,Vol. 2.*

[25] Van de Waterbeemd, H. & Rose, S. In The Practice of Medicinal Chemistry 2nd *(ed Wermuth, L. G.) 1367–1385(Academic Press, 2003)*.

[26] Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W. & Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem*. 45, 2615–2623 (2002).

[27] Weka, a multi-task machine learning software developed by Waikato University. See www.cs.waikato.ac.nz/ml/weka/.

[28] Yoshida, F. & Topliss, J. G. QSAR model for drug human oral bioavailability, *J. Med. Chem. 43, 2575–2585 (2000).*

[29] Zupan J, Gasteiger, Neural Networks in chemistry and drug design: an introduction, 2nd edition, Wiley 1999.