IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 9, 2018, accepted March 24, 2018, date of publication March 30, 2018, date of current version May 2, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2821441

ECoFFeS: A Software Using Evolutionary Computation for Feature Selection in Drug Discovery

ZHI-ZHONG LIU[1], JIA-WEI HUANG[1], YONG WANG[1,2], (Senior Member, IEEE), AND DONG-SHENG CAO[3]

[1]School of Information Science and Engineering, Central South University, Changsha 410083, China
[2]School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, U.K.
[3]Xiangya School of Pharmaceutical Sciences, Central South University, Changsha 410013, China

Corresponding authors: Yong Wang (ywang@csu.edu.cn) and Dong-Sheng Cao (oriental-cds@163.com)

This work was supported in part by the Innovation-Driven Plan in Central South University under Grant 2018CX010, in part by the National Natural Science Foundation of China under Grant 61673397, in part by the National Key Basic Research Program under Grant 2015CB910700, in part by the Hunan Provincial Natural Science Fund for Distinguished Young Scholars under Grant 2016JJ1018, and in part by the Graduate Innovation Fund of Hunan Province of China under Grant CX2017B062.

ABSTRACT Feature selection is of particular importance in the field of drug discovery. Many methods have been put forward for feature selection during recent decades. Among them, evolutionary computation has gained increasing attention owing to its superior global search ability. However, there still lacks a simple and efficient software for drug developers to take advantage of evolutionary computation for feature selection. To remedy this issue, in this paper, a user-friendly and standalone software, named ECoFFeS, is developed. ECoFFeS is expected to lower the entry barrier for drug developers to deal with feature selection problems at hand by using evolutionary algorithms. To the best of our knowledge, it is the first software integrating a set of evolutionary algorithms (including two modified evolutionary algorithms proposed by the authors) with various evaluation combinations for feature selection. Specifically, ECoFFeS considers both single-objective and multi-objective evolutionary algorithms, and both regression- and classification-based models to meet different requirements. Five data sets in drug discovery are collected in ECoFFeS. In addition, to reduce the total analysis time, the parallel execution technique is incorporated into ECoFFeS. The source code of ECoFFeS can be available from https://github.com/JiaweiHuang/ECoFFeS/.

INDEX TERMS Evolutionary computation, feature selection, drug discovery, single-objective optimization, multi-objective optimization, parallel execution.

I. INTRODUCTION

Drug discovery denotes the process by which new candidate medications are discovered in the fields of bioinformatics and bioengineering. Despite significant advances have been achieved in technology and understanding of biological systems, drug discovery is still an ''expensive, difficult, and inefficient process'' with a low success rate [1]. According to the reports in [2] and [3], the average cost of developing a new medicine to market can reach about $1.8 billion in 2010 and about $2.6 billion in 2014, respectively. Those gigantic investments usually come from pharmaceutical industry cooperations as well as national governments, since it is believed that the new discovered drugs may lead to great commercial success or public health success. Nowadays, one of the challenging tasks in drug discovery is drug screening, the aim of which is to obtain the desired compounds from a library of compounds [4]. Obviously, data mining techniques are required to achieve this goal. Compared with other data mining techniques, feature selection merely selects a feature subset and does not alter the original representation of features [5]. Thus, the selected feature subset preserves the semantics of features while offering the advantage of interpretability.

For a general feature selection process, its main components are presented in Fig. 1 [6], [7]. First of all, the original feature set is presented in the initialization process. Subsequently, a search procedure, named subset discovery, is implemented to generate a candidate feature subset from

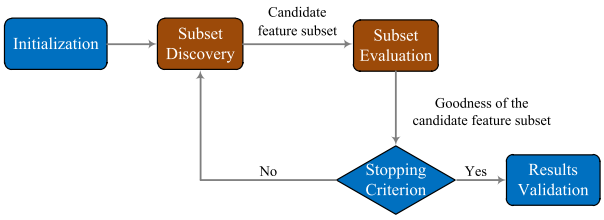


FIGURE 1. General feature selection process [6], [7].

the original feature set. After the feature subset is generated, the procedure called subset evaluation is implemented to assess the goodness of this feature subset, and to compare it with the previous best feature subset. If this feature subset is better, then it will replace the previous best feature subset. Both the subset discovery and the subset evaluation are implemented repeatedly until a stopping criterion is met. Finally, the output feature subset will be tested through a validation procedure. It is clear that the subset discovery and the subset evaluation are two essential components in feature selection. For interested readers, more details can be obtained from [6], [7], [8], and [9].

Indeed, recent decades have witnessed significant progress in the development of feature selection for drug discovery, such as sequence analysis [10], microarray analysis [11], mass spectra (MS) analysis [12], single nucleotide polymorphism (SNP) analysis [13], and quantitative structure-activity relationship (QSAR) analysis [14]. However, feature selection remains a challenging task due to the fact that it is an NP-hard problem, in which the total number of possible feature subsets is $(2^N - 1)$, where N is the number of features. To deal with this issue, many methods have been proposed, such as complete search, greedy search, and heuristic search [7]. Nevertheless, most of them suffer from stagnation in a local optimum or high computational cost. Therefore, the demand of an efficient global search method is particularly urgent for better solving feature selection problems.

Evolutionary computation, which is a family of population-based heuristic search methods inspired by nature, seems to be a good choice because of its powerful global search ability [15]. It has some attractive advantages such as ease to use, efficiency, and robustness [16]. Moreover, it does not make any assumptions about the search space, such as linear/nonlinear, and differentiable/nondifferentiable. In particular, its population-based mechanism can generate multiple solutions in a single run [17], [18]. This property is beneficial to solve multi-objective feature selection problems, in which a set of non-dominated solutions with the tradeoff between the number of features and the performance metric is desired. Currently, evolutionary computation has attracted a high level of interest from the feature selection research community [7], and has been successfully applied to feature selection in diverse fields, such as image analysis [19], face recognition [20], gene analysis [21], human action recognition [22], disease diagnose [23], network security [24], and drug discovery [9].

Although evolutionary computation has demonstrated its efficiency and versatility, there still lacks a user-friendly and efficient software to take advantage of evolutionary computation for feature selection in drug discovery. To alleviate this issue, in this paper, a easy-to-use and standalone software, named ECoFFeS, is developed. The purposes of ECoFFeS are twofold: 1) lowering the entry barrier for drug developers, and 2) further boosting evolutionary computation for feature selection in drug discovery.

The main contributions and novelties of this paper are summarized as follows:

- As far as we know, ECoFFeS is the first software using evolutionary computation for feature selection in drug discovery. In addition, it provides a user-friendly graphical user interface, and does not require researchers to have any knowledge of programming. As a result, ECoFFeS is expected to encourage researchers in the field of drug discovery to use evolutionary computation techniques to address feature selection problems at hand. It is also expected that ECoFFeS can attract more attention from researchers in the evolutionary computation community to further develop effective and efficient approaches to handle new challenges in feature selection of drug discovery.
- Both single-objective evolutionary algorithms (SOEAs) and multi-objective evolutionary algorithms (MOEAs), and both regression- and classification-based models are synthesized in ECoFFeS. Therefore, ECoFFeS has the capability to address different kinds of feature selection problems that drug developers meet in real-life applications. It is worth noting that a novel SOEA (i.e., BFDE) and a novel MOEA (i.e., MOEA/D-BFDE) are proposed in this paper and they are incorporated into ECoFFeS to solve single-objective and multi-objective optimization problems, respectively. The experimental results have validated their effectiveness.
- Five datasets in drug discovery are collected in ECoFFeS, i.e., Artemisinin [25], benzodiazepine receptors (BZR) [26], Selwood [27], hERG [28], and $\log D_{7.4}$ [29] datasets.
- Parallel execution is supported in ECoFFeS, which can significantly reduce the total analysis time. This property will attract more drug developers to use this software and to advance the development of evolutionary computation for feature selection in drug discovery.
- ECoFFeS is free for drug developers. Besides, for researchers who are interested in the further development of ECoFFeS, the Matlab source code is also offered.

The rest of this paper is organized as follows. Section II introduces the graphical user interface of ECoFFeS. Section III describes the internal structure of ECoFFeS. The collected datasets are presented in Section IV. The experimental studies are given in Section V. The applications of

FIGURE 2. Main interface of ECoFFeS.

ECoFFeS are provided in Section VI. Finally, Section VII concludes this paper.

II. GRAPHICAL USER INTERFACE (GUI)

The graphical user interface (GUI) of ECoFFeS contains one main interface and a series of secondary interfaces. For coping with a feature selection problem, drug developers can choose an EA (i.e., a SOEA or a MOEA) and a model (i.e., a regression- or classification-based model) in the main interface, and then enter the corresponding secondary interface.

A. MAIN INTERFACE

Fig. 2 shows the main interface of ECoFFeS. It contains three parts:

- 'Subset Discovery': 'SOEAs' or 'MOEAs' is used to solve single-objective or multi-objective optimization problems, respectively.
- 'Subset Evaluation': 'Regression Model & Metric' or 'Classification Model & Metric' is designed for regression or classification, respectively.
- 'Start': To launch the secondary interface.

Through combining 'Subset Discovery' with 'Subset Evaluation', four secondary interfaces can be produced, i.e., 'SOEAs_Regression', 'MOEAs_Regression', 'SOEAs_Classification', and 'MOEAs_Classification'. These four secondary interfaces are utilized to deal with different kinds of feature selection problems.

B. SECONDARY INTERFACE

For each secondary interface, it contains six panels:

- 'Data': To show the data that has been imported.
- 'Results': To present the results at the end of a run.
- 'Figure': To plot the charts at the end of a run.
- 'State': To exhibit the current operating status.
- 'Settings': To set the parameters. For instance, 'Import Data' is used to load dataset, which can be imported as XLS or XLSX files; 'Parallelization' offers an option to use the parallel execution technique or not; 'Save Figure' is used to save figures, which can be exported as JPG, PNG, or FIG files; 'SOEA' or 'MOEA' is used to select a SOEA or a MOEA, respectively; and 'SOEA_parameter' or 'MOEA_parameter' is used to set parameters in a SOEA or a MOEA, respectively. In 'Model', a model can be selected and its corresponding parameters can be set in 'Model_parameter'. Similarly, in 'Metric', a metric can be selected and its corresponding parameters can be set in 'Metric_parameter'. Besides, 'Popsize' is used to set the population size, 'Iteration' is applied to set the maximum generation number of an algorithm, and 'Runs' is employed to set the total number of runs of an algorithm.
- 'Command': To implement the command control. 'Play' is to start the run, 'Stop' is to stop the run, and 'menu'

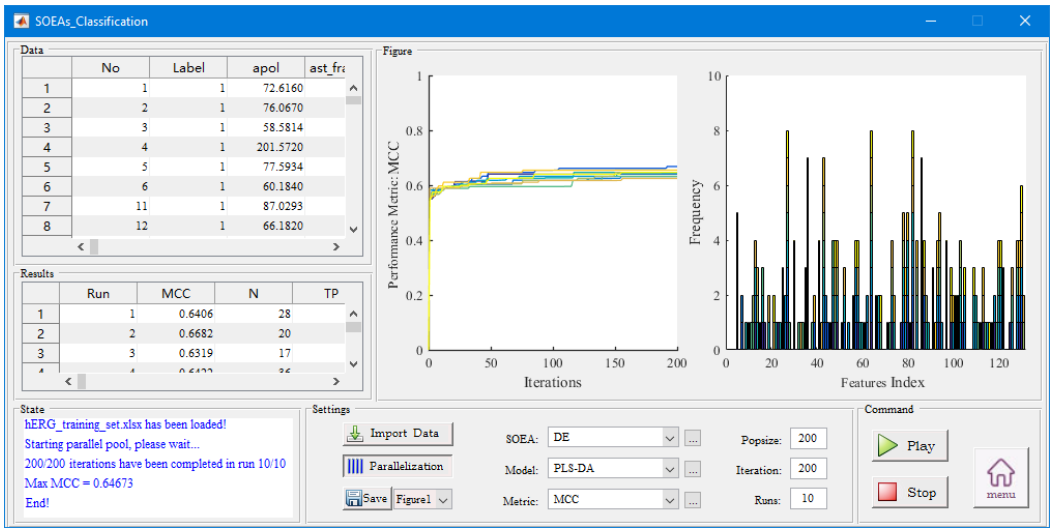


FIGURE 3. An example of the secondary interface of ECoFFeS.

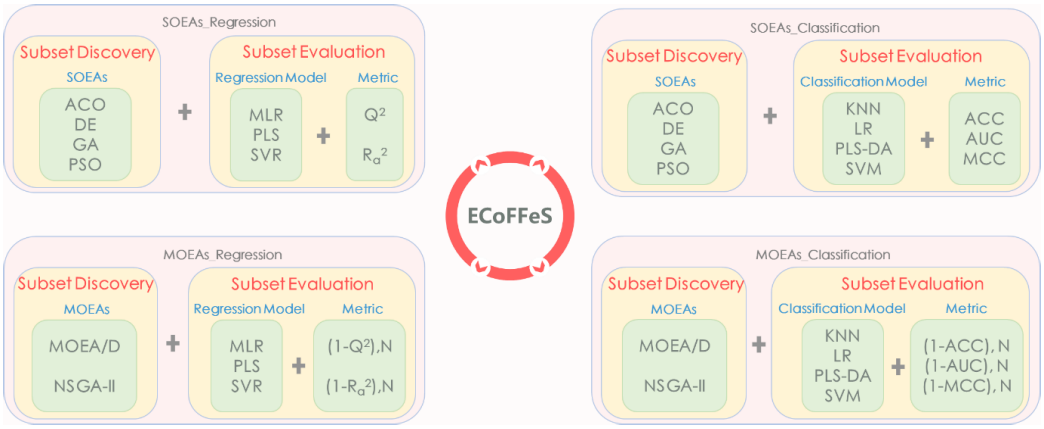


FIGURE 4. Internal structure of ECoFFeS.

is to come back to the main interface.

An example is presented in Fig. 3, which is the 'SOEAs_Classification' secondary interface. This secondary interface is used to solve single-objective classification problems. To achieve this, firstly, we need to load dataset and set paraments. In this case, 'hERG_training_set.xlsx' is the import data which comes from the hERG dataset, the parallel execution technique is chosen for calculation, DE is the selected SOEA, PLS-DA is the selected model, MCC is the selected metric, the population size is set to 200, the maximum generation number is set to 200, and the total number of runs is set to 10. Subsequently, 'Play' is clicked to launch the calculation. After the calculation completes, 'Save' is pressed and then 'Figure1', 'Figure2', and 'Results' in the drop-down menu are used to save the results. Herein, 'Iteration Figure.jpg', 'Frequency Figure.jpg', and 'Results.xlsx' are saved in 'Figure1', 'Figure2', and 'Results', respectively. Note that the two figures are presented in 'Figure' panel and 'Results.xlsx' is output in 'Result' panel.

From the above introduction, we can conclude that ECoFFeS offers a user-friendly GUI for drug developers.

III. INTERNAL STRUCTURE OF ECOFFES

The internal structure of ECoFFeS is presented in Fig. 4, which consists of four main parts corresponding to the four secondary interfaces in GUI. For each part, it contains two key components: 'Subset Discovery' and 'Subset Evaluation'.

A. SUBSET DISCOVERY: EVOLUTIONARY COMPUTATION

Subset discovery denotes a search procedure generating candidate feature subsets. The categories of evolutionary computation approaches for feature subset discovery in ECoFFeS are presented in Fig. 5. There are four famous SOEAs (i.e, ant colony optimization (ACO) [30], differential evolution (DE) [31], genetic algorithm (GA) [32], and particle swarm optimization (PSO) [33]) and two well-known MOEAs (i.e., MOEA/D [34] and NSGA-II [35]).

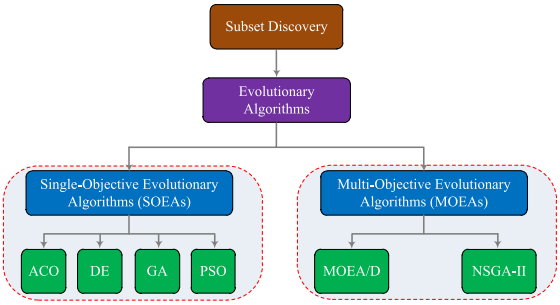


FIGURE 5. Evolutionary computation approaches for feature subset discovery.

1) SINGLE-OBJECTIVE EVOLUTIONARY ALGORITHMS (SOEAS)

The feature subset discovery can be regarded as a single-objective discrete optimization problem, the purpose of which is to select the optimal feature subset according to a specific evaluation metric. The four SOEAs used in ECoFFeS are introduced below:

- ACO is biologically inspired from the behavior of colonies of real ants, in particular, how they forage for food. Since the idea of ACO was proposed, it has been successfully applied to solve discrete optimization problems.
- DE is one of the most popular EA paradigms. Note, however, that feature selection belongs to discrete optimization problems and DE cannot address this kind of optimization problems directly. To this end, a DE variant called binary differential evolution (BDE) is proposed in [36]. In ECoFFeS, by taking the feedback information into consideration, an enhanced version of BDE named BFDE is proposed. The details of BFDE are presented in Appendix.
- GA is a population-based heuristic method inspired by the process of natural evolution. In GA, each solution is represented as a chromosome which is associated with a fitness value. Each solution will undergo evolution and the ones with better fitness values will survive. It is well-accepted that GA is an effective optimizer for solving discrete optimization problems.
- PSO is an optimization technique designed for continuous optimization problems, which is motivated by the behavior of organisms such as fish schooling and bird flocking. To cope with discrete optimization problems, Kennedy and Eberhart adapted the standard PSO to binary spaces and proposed binary PSO (BPSO) [37]. This version of PSO is employed in ECoFFeS.

2) MULTI-OBJECTIVE EVOLUTIONARY ALGORITHMS (MOEAS)

Under some conditions, it would be desirable if the discovered feature subset is with not only the best value of evaluation metric but also the minimum number of features. However, these two objectives are always conflicting with each other. From this point of view, feature selection should be treated as a multi-objective optimization problem rather than a single-objective optimization problem. Note, however, that solving a multi-objective optimization problem is not an easy task since its goal is to obtain a set of trade-off solutions between the evaluation metric and the number of features. Fortunately, EAs are particularly suitable for tackling multi-objective optimization problems due to their population-based property [7]. In fact, many MOEAs have been proposed during the last two decades [38]. Two representatives among them, namely, MOEA/D [34] and NSGA-II [35], are used in ECoFFeS and introduced below.

- MOEA/D is a famous decomposition-based MOEA, which decomposes a multi-objective optimization problem into a number of scalar optimization subproblems and then optimizes them simultaneously. In MOEA/D, each subproblem is optimized by exploiting the information from its several neighboring subproblems [38]. By integrating MOEA/D with BFDE, a new variant of MOEA/D named MOEA/D-BFDE is implemented in ECoFFeS to cope with multi-objective feature selection problems.
- NSGA-II is a well-known Pareto dominance-based MOEA, which contains two key parts: nondominated sorting procedure and crowding distance calculation. Firstly, the nondominated sorting procedure is implemented to divide the population into different layers and to decide the last layer [35]. Afterward, the crowding distance calculation is conducted in the last layer to select the better individuals with the larger crowding distances. In ECoFFES, NSGA-II is intergraded with GA [32] to tackle multi-objective feature selection problems.

B. SUBSET EVALUATION: MODELS

Subset evaluation seeks to assess the candidate feature subsets generated by subset discovery. It plays an important role in drug discovery since evaluation function has the capability to guide the search toward the optimal feature subset. In ECoFFeS, an evaluation function consists of two main parts: models and metrics. Models are developed using one or more statistical modeling tools, which can be broadly categorized into regression- and classification-based models. Fig. 6 shows the overview of models in subset evaluation.

1) REGRESSION-BASED MODELS

Regression-based models are used when the response variable is quantitative. ECoFFeS integrates three classical regression models, namely, multiple linear regression (MLR) [39], partial least squares (PLS) [40], and support vector regression (SVR) [41]. Each of them has its own advantages. For instance, MLR is one of the most popular models due to its simplicity in operation, reproducibility, and ability to allow easy interpretation of the features used. In terms of PLS, it is a better choice when handling a large

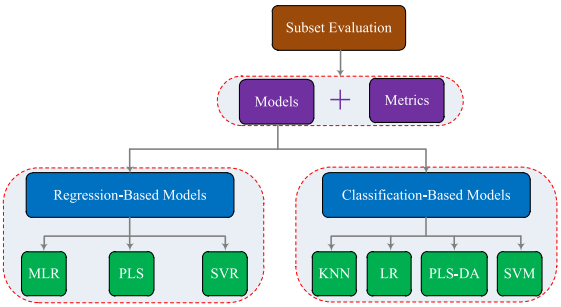


FIGURE 6. Models in subset evaluation.

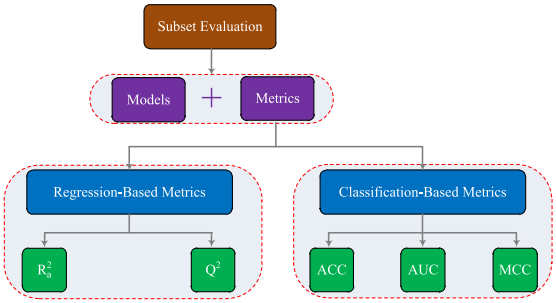


FIGURE 7. Metrics in subset evaluation.

number of inter-correlated and noisy features with a limited number of data points. With respect to SVR, by using kernel functions, it has the capability to avoid the difficulties of using linear functions in the high-dimensional feature space.

2) CLASSIFICATION-BASED MODELS

Classification-based models are used when the response variable is a label (i.e., positive-negative). Four widely used classification-based models are incorporated into ECoFFeS, i.e., k-nearest neighbor (KNN) [42], logistic regression (LR) [43], partial least squares discriminant analysis (PLS-DA) [44], and support vector machines (SVMs) [45]. KNN is a type of instance-based learning algorithm, which is very simple but works well in practice. LR is a statistical classification model that measures the relationship between a categorical-dependent variable and other independent variables by using probability scores. PLS-DA is a linear classification method combining the properties of partial least squares regression with the discrimination power of a classification technique. In terms of SVMs, it is a nonlinear classification model using a kernel function to map the input data into a higher-dimensional space, where the instances are linearly separable.

C. SUBSET EVALUATION: METRICS

Metrics are used to evaluate the quality of models. Based on the categories of models, metrics can be divided into two classes: 1) metrics for regression-based models, and 2) metrics for classification-based models. Fig. 7 shows the overview of metrics in subset evaluation.

1) METRICS FOR REGRESSION-BASED MODELS

There are two regression-based metrics employed in ECoFFeS: adjusted R^2 (R_a^2) [46] and cross-validated Q^2 [47]. R^2 denotes the determination coefficient, which is used to judge the fitting ability of a model. Further, R_a^2 is an enhanced version of R^2 which considers the number of features additionally. As for cross-validated Q^2, it is a well-known metric which employs the cross-validation technique. In cross-validated Q^2, the samples are classed into two subsets: calibration (i.e., training) and validation (i.e., test) subsets. The aim of the former is to construct a model, while the aim of the latter is to verify the predicting performance of the constructed model.

2) METRICS FOR CLASSIFICATION-BASED MODELS

In ECoFFeS, there are three well-known metrics to assess the performance of classification-based models, namely, accuracy (ACC) [48], Matthews correlation coefficient (MCC) [49], and area under curve-receiver operating characteristics (AUC-ROC) [50]. ACC denotes the prediction accuracy, which is commonly used to evaluate the classifier model performance and classification capability. For MCC, it usually measures the quality in binary classifications, which can be efficient even if the classes are imbalance. With respect to AUC-ROC, it can be regarded as a simple average of the ranks of the positive samples.

Remark: For single-objective feature selection problems, the above metrics such as R_a^2, Q^2, ACC, MCC, and AUC-ROC can be directly used to assess regression- or classification-based models. A larger value is expected for these metrics. However, for multi-objective feature selection problems,[1] one objective is assessed by 1 plus one of the corresponding metric (i.e., $1-R_a^2$, $1-Q^2$, 1−ACC, 1−MCC, or 1−AUC-ROC), and the other objective is evaluated by the number of the selected features.

D. DISCUSSION

From the above introduction, we can make the following comments about the internal structure of ECoFFeS:

- For subset discovery, it can be regarded as a single-objective or multi-objective optimization problem. When it is treated as a single-objective optimization problem, SOEAs aim at obtaining a satisfactory feature subset and providing the rankings of the important features simultaneously. On the other hand, when it is formulated as a multi-objective optimization problem, MOEAs can be adopted to maintain a set of non-dominated feature subsets with a tradeoff between the number of features and the corresponding metric. Afterward, the decision maker can select one feature subset matching at most his/her preference. Note that a novel version of DE (BFDE) and a novel version of MOEA/D (MOEA/D-BFDE) are incorporated into ECoFFeS.

[1] Indeed, there are two objectives in ECoFFeS.

TABLE 1. Experimental results of PLS and BFDE-PLS on the three datasets.

Datasets	Methods	Mean Q^2 ± Standard deviation	Mean *RMSECV* ± Standard deviation	Mean N ± Standard deviation
Artemisinin	PLS	0.6003	0.9912	89
	BFDE-PLS	**0.7594 ± 0.0072**	**0.7690 ± 0.0115**	**23.6000 ± 2.9196**
BZR	PLS	0.4007	0.8501	75
	BFDE-PLS	**0.5863 ± 0.0087**	**0.7063 ± 0.0074**	**21.5000 ± 2.3599**
Selwood	PLS	0.2407	0.6461	53
	BFDE-PLS	**0.9206 ± 0.0067**	**0.2087 ± 0.0087**	**12.0000 ± 1.4384**

- For subset evaluation, 36 evaluation combinations are provided for drug developers in ECoFFeS. Among them, 12 are used for regression, which are the combinations of regression-based models and metrics, and 24 are used for classification, which are the combinations of classification-based models and metrics.
- Since both subset discovery and subset evaluation are carefully considered, ECoFFeS is a generic tool to tackle different types of feature selection problems in drug discovery.

IV. DATASETS

Five datesets in drug discovery are collected in this paper, which are Artemisinin [25], benzodiazepine receptors (BZR) [26], Selwood [27], hERG [28], and $\log D_{7.4}$ [29] datasets. For these datasets, an important issue is to explore the relationship between compounds and corresponding biological activities or chemical properties. Quantitative structure-activity/property relationship (QSAR/QSPR) [51] is developed for this purpose. In QSAR/QSPR studies, the chemical structure of a compound is represented by several descriptors, such as molecular constitutional, topological, shape, autocorrelation, and charge descriptors. However, the number of descriptors is usually relatively larger than the number of compounds. Moreover, there exist some redundant, noisy, and irrelevant descriptors, which may lead to either over fitting or a low correlation between structures and activities [52], [53]. Consequently, the process of descriptor selection is necessary and meaningful, which is obviously a feature selection problem.

In this paper, the first three datasets are used for experimental testing, and the last two datasets are used for applications.

V. EXPERIMENTAL STUDIES

A. NECESSITY OF DESCRIPTOR SELECTION

To verify the necessity of descriptor selection, we compared the method with and without the descriptor selection. Herein, we took the PLS model as an example. For the method without descriptor selection, all the descriptors were directly used for the model development and this method is named as PLS. In terms of the method with descriptor selection, BFDE was used for subset discovery and the resultant method is named as BFDE-PLS. Then, we tested these two methods on three datasets: the Artemisinin, BZR, and Selwood datasets. Note that in BFDE, the population size was set to 150, the maximum generation number was set to 300, and 100 independent runs were conducted to produce the experimental results.

In order to compare the performance of the involved methods, three performance metrics were chosen: Q^2, the root mean square error from five-fold cross-validation (denoted as *RMSECV*) [54], and the number of the selected descriptors (denoted as N). For Q^2, a larger value is desirable; while for *RMSECV* and N, the smaller the better.

The comparison results are presented in Table 1 and the better results are highlighted in **boldface**. Clearly, BFDE-PLS performs better than PLS since it obtains a larger Q^2, a smaller *RMSECV*, and a smaller N on each of these three datasets. Therefore, BFDE-PLS achieves a better validation metric and a smaller validation error while using a less number of features. Thus, we can conclude that descriptor selection is definitely necessary in drug discovery.

B. EFFECTIVENESS OF BFDE

BDE has been verified as a powerful algorithm for discrete optimization problems [36], and BFDE can be regarded as a variant of BDE. Naturally, we compared BFDE with BDE to verify the effectiveness of BFDE. For a fair comparison, the parameter settings and the models used for these two algorithms were kept the same. Specifically, the population size was set to 150, the maximum generation number was set to 300, 100 independent runs were performed, and the PLS model was utilized. Three performance metrics (i.e., Q^2, *RMSECV*, and N) were selected for performance comparison. Table 2 shows the comparison results between them and the better results are highlighted in **boldface**.

It is clear that BFDE significantly outperforms BDE on the Artemisinin, BZR, and Selwood datasets in terms of the three performance metrics. To be specific, BFDE-PLS obtains a better validation metric (i.e., a larger Q^2), a smaller validation error (i.e., a smaller *RMSECV*), and a less number of features (i.e., a smaller N). As a result, we can conclude that the feedback strategy in BFDE is effective and the proposed BFDE is more powerful than BDE.

C. NECESSITY OF MOEA/D-BFDE

MOEA/D-BFDE is a novel version of MOEA/D, which combines MOEA/D with BFDE to solve multi-objective discrete optimization problems. To test its performance, we compared it with NSGA-II. To make a fair comparison, for these two

TABLE 2. Experimental results of BDE-PLS and BFDE-PLS on the three datasets.

Datasets	Methods	Mean Q^2 ± Standard deviation	Mean *RMSECV* ± Standard deviation	Mean N ± Standard deviation
Artemisinin	BDE-PLS	0.7481 ± 0.0117	0.7867 ± 0.01845	38.5000 ± 3.4114
	BFDE-PLS	**0.7594 ± 0.0072**	**0.7690 ± 0.0115**	**23.6000 ± 2.9196**
BZR	BDE-PLS	0.5735 ± 0.0133	0.7171 ± 0.0111	32.0667 ± 3.8321
	BFDE-PLS	**0.5863 ± 0.0087**	**0.7063 ± 0.0074**	**21.5000 ± 2.3599**
Selwood	BDE-PLS	0.8976 ± 0.0113	0.2369 ± 0.0129	18.4667 ± 3.5305
	BFDE-PLS	**0.9206 ± 0.0067**	**0.2087 ± 0.0087**	**12.0000 ± 1.4384**

algorithms, the population size was set to 150, the maximum generation number was set to 400, and the MLR model was utilized. In order to analyze the performance of these two algorithms, three performance metrics were chosen: Max Q^2, Min *RMSECV*, and Max N. For each algorithm, 10 independent runs were conducted. The comparison results are summarized in Fig. 8.

From Fig. 8, we can observe that MOEA/D-BFDE-MLR provides better Max Q^2 and Min *RMSECV* on the BZR and Selwood datasets, while offers worse Max Q^2 and Min *RMSECV* on the Artemisinin dataset, which suggests that MOEA/D-BFDE-MLR is better than NSGA-II-MLR on the BZR and Selwood datasets, while worse than NSGA-II-MLR on the Artemisinin dataset, respectively. Therefore, it can be concluded that both MOEA/D-BFDE and NSGA-II have their own advantages and should be considered in ECoFFeS.

D. EFFICIENCY OF PARALLEL EXECUTION

ECoFFeS supports parallel execution, which is a useful strategy to make use of the processing ability of multi-core computers. To verify the efficiency of parallel execution, we took ACO-PLS as an example and run it with and without the parallel execution technique. Then, we recorded the runtime in these two different cases. The experimental results are presented in Table 3, in which *YES* or *NO* means the parallel execution technique was used or not, respectively. It is evident that the runtime with parallelization is much less than that without parallelization. As a result, we can conclude that the parallel execution in ECoFFeS is quite efficient which can significantly reduce the runtime.

VI. APPLICATIONS

In this section, we will introduce two applications of ECoFFeS, namely hERG blockers predicting and $\log D_{7.4}$ predicting, by making use of the hERG and $\log D_{7.4}$ datasets, respectively.

A. PREDICTING HERG BLOCKERS

1) BACKGROUND

In the process of cardiac depolarization and repolarization, a voltage-gated potassium channel encoded by the human ether-à-go-go related gene (hERG or Kv11.1) plays a major role in the regulation of the exchange of cardiac action potential and resting potential [55]. The hERG blockade may cause long QT syndrome (LQTS), arrhythmia, and

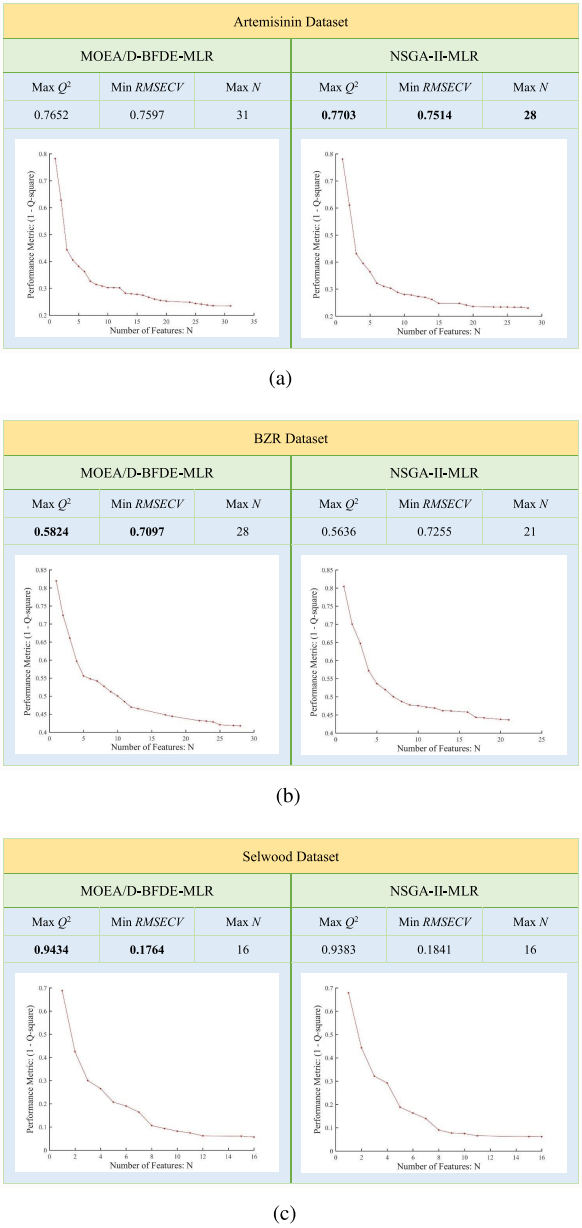
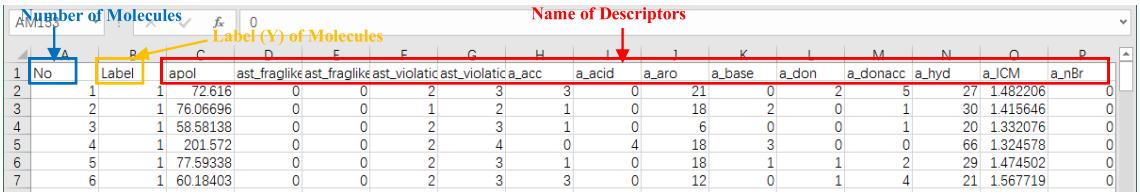


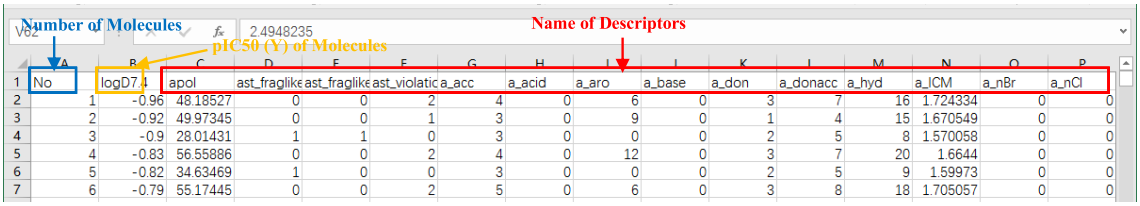
FIGURE 8. Experimental results of MOEA/D-BFDE and NSGA-II on the Artemisinin, BZR, and Selwood datasets. (a) Artemisinin. (b) BZR. (c) Selwood.

Torsade de Pointes(TdP), which can result in palpitations, fainting, or even sudden death [56]. Therefore, the assessment of hERG-related cardiotoxicity has become an essential step



No	Label	apol	ast_fraglike	ast_fraglike	ast_violatic	ast_violatic	a_acc	a_acid	a_aro	a_base	a_don	a_donacc	a_hyd	a_ICM	a_nBr
1	1	72.616	0	0	2	3	3	0	21	0	2	5	27	1.482206	0
2	1	76.06696	0	0	1	2	1	0	18	2	0	1	30	1.415646	0
3	1	58.58138	0	0	2	3	1	0	6	0	0	1	20	1.332076	0
4	1	201.572	0	0	2	4	0	4	18	3	0	0	66	1.324578	0
5	1	77.59338	0	0	2	3	1	0	18	1	1	2	29	1.474502	0
6	1	60.18403	0	0	2	3	3	0	12	0	1	4	21	1.567719	0

FIGURE 9. Details of the hERG dataset.



No	logD7.4	apol	ast_fraglike	ast_fraglike	ast_violatic	a_acc	a_acid	a_aro	a_base	a_don	a_donacc	a_hyd	a_ICM	a_nBr	a_nCl
1	-0.96	48.18527	0	0	2	4	0	6	0	3	7	16	1.724334	0	0
2	-0.92	49.97345	0	0	1	3	0	9	0	1	4	15	1.670549	0	0
3	-0.9	28.01431	1	1	0	3	0	0	0	2	5	8	1.570058	0	0
4	-0.83	56.55886	0	0	2	4	0	12	0	3	7	20	1.6644	0	0
5	-0.82	34.63469	1	0	0	3	0	0	0	2	5	9	1.59973	0	0
6	-0.79	55.17445	0	0	2	5	0	6	0	3	8	18	1.705057	0	0

FIGURE 10. Details of the $\log D_{7.4}$ dataset.

TABLE 3. Parallel performance analysis of ECoFFeS.

Datasets	Parallelization	Popsize	Maximum generation	Runtime (s)
Artemisinin	NO	150	200	488
	YES	150	200	**181**
BZR	NO	150	200	452
	YES	150	200	**170**
Selwood	NO	150	200	331
	YES	150	200	**116**

in drug discovery [57]. It should be noted that hERG assays and QT animal studies are time-consuming and expensive. Thus, it becomes urgent to develop a reliable and robust silico model to predict potential hERG liability.

2) ECOFFES OPERATION

Herein, predicting hERG blockers is considered as a single-objective classification problem. ECoFFeS can be used conveniently to solve this problem. The steps are presented as follows:

- Start SOEAs_Classification: At first, double-click ECoFFeS to start the main interface. Subsequently, select 'SOEAs' in 'Subset Discovery' and 'Classification Model & Metric' in 'Subset Evaluation'. Finally, press 'Start' to enter the secondary interface (i.e., SOEAs_Classification).
- Set parameters and play: For import data, 'hERG_training_set.xlsx' was loaded which comes from the hERG dataset. In Fig. 9, it can be observed that the first column of the hERG dataset is the number of molecules, the second column is the label of molecules, and the remaining columns are descriptor values of molecules. For other parameters and settings, DE was the selected SOEA, PLS-DA was the selected model, and MCC was the utilized metric. Besides, the population size was set to 200, the maximum generation number was set to 200, and 10 independent runs were conducted. After the above settings, press 'Play' to start the calculation.

TABLE 4. Statistical results of the important descriptors for predicting hERG blockers.

Code	Class	Description
BCUT_SLOGP_0	2D	LogP BCUT (0/3)
opr_brigid	2D	Oprea Rigid Bond Count
PEOE_VSA-2	2D	Total negative 2 vdw surface area
b_max1len	2D	Maximum single-bond chain length
FCharge	2D	Sum of formal charges
vsa_pol	2D	VDW polar surface area (A**2)
PEOE_VSA-6	2D	Total negative 6 vdw surface area

- Save figures and results: After the calculation, press 'Save', then 'Figure1', 'Figure2', and 'Results' in the drop-down menu were used to save 'Iteration Figure.jpg', 'Frequency Figure.jpg', and 'Results.xlsx', respectively. The interface of ECoFFeS in this case is presented in Fig. 3.

3) RESULTS AND DISCUSSION

According to the right-hand side picture in 'figure' panel of Fig. 3, we can acquire the importance of the molecular descriptors, which is summarized in Table 4. In Table 4, BCUT_SLOGP_0, opr_brigid, and PEOE_VSA-2 in the PLS-DA classifiers have the highest frequency (8/10). As for b_max1len, FCharge, PEOE_VSA-6, and vsa_pol, they are also very important since they are indicated by relatively high frequencies ($>$ 5/10). In terms of these important molecular descriptors, they can be used for further QSAR model development.

B. PREDICTING $LOGD_{7.4}$

1) BACKGROUND

According to [29], it is very important to evaluate the lipophilicity of candidate compounds in drug discovery. Usually, a compound's lipophilicity can be quantitatively characterized by the partition coefficient (its logarithm form is

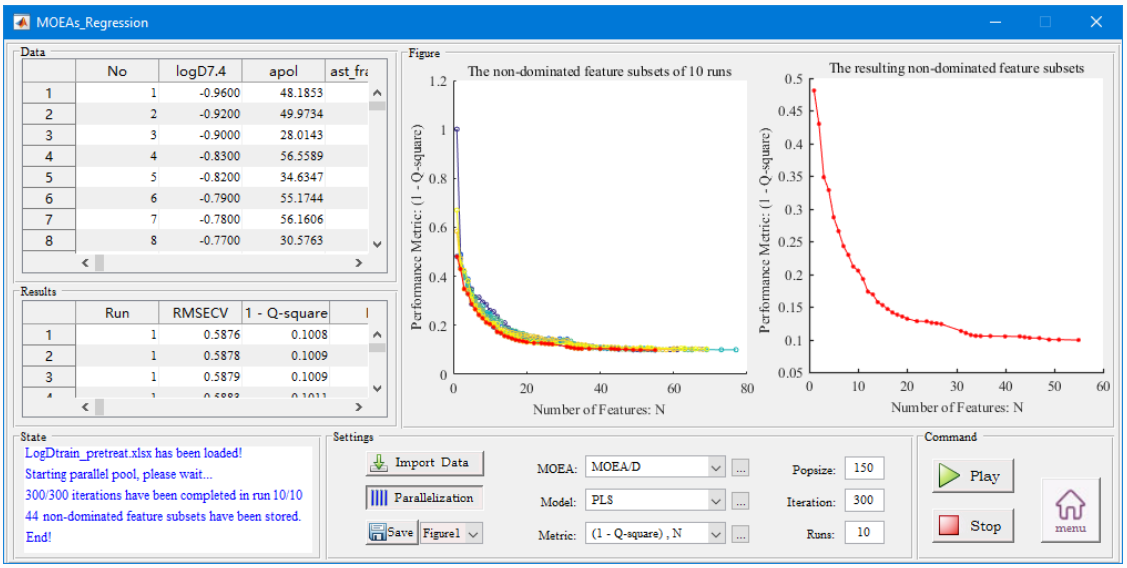


FIGURE 11. Obtained Results in ECoFFeS for Predicting $\log D_{7.4}$.

denoted as logP) or the distribution coefficient (its logarithm form is denoted as logD) [58]. Due to taking the ionization into account, the distribution coefficient, which is usually known as pH-dependent distribution coefficient, seems to be a more reliable measurement for the lipophilicity at physiological pH [59]. It should be noted that the measurement of logD value is usually costly and time-consuming, which requires substantial quantities of the compound being synthesized [60]. Consequently, it is necessary to establish a reliable prediction model to accurately determine $\log D_{7.4}$ (pH$\approx$ 7.4 in human body) values, especially for new or even virtual compounds.

2) ECOFFES OPERATION

Herein, we regard $\log D_{7.4}$ predicting as a multi-objective regression problem. ECoFFeS is used to address this problem via the following steps:

- Start MOEAs_Regression: Firstly, double-click ECoFFeS to start the main interface. Thereafter, choose 'MOEAs' in 'Subset Discovery' and 'Regression Model & Metric' in 'Subset Evaluation'. Finally, press 'Start' to start the secondary interface (i.e., MOEAs_Regression).
- Set parameters and play: The import data was loaded from the $\log D_{7.4}$ dataset which is 'LogDtrain_pretreat.xlsx'. For this dataset, as shown in Fig. 10, its first column is the number of molecules, its second column is the $\log D_{7.4}$ of molecules, and its remaining columns are descriptor values of molecules. For other parameters and settings, MOEA/D was chosen as the selected MOEA, PLS model was the selected model, and $1 - Q^2$ and N were two selected metrics. Moreover, the population size was set to 150, the maximum generation number was set to 200, and 10 independent runs were conducted. After the above setting, press 'Play' to start the calculation.

TABLE 5. Statistical results of the important descriptors for predicting $\log D_{7.4}$

Code	Class	Description
apol	2D	Sum of atomic polarizabilities
a_donacc	2D	Number of H-bond donor + acceptor atoms
a_hyd	2D	Number of hydrophobic atoms
balabanJ	2D	Balaban averaged distance sum connectivity
logP(o/w)	2D	Log octanol/water partition coefficient
logS	2D	Log Solubility in Water
nmol	2D	Number of molecules
PEOE_VSA+0	2D	Total positive 0 vdw surface area
PEOE_VSA+3	2D	Total positive 3 vdw surface area
PEOE_VSA-2	2D	Total negative 2 vdw surface area
PEOE_VSA-5	2D	Total negative 5 vdw surface area
PEOE_VSA_FPOS	2D	Fractional positive vdw surface area
PEOE_VSA_POS	2D	Total positive vdw surface area
SlogP_VSA1	2D	Bin 1 SlogP (-0.40,-0.20]
SlogP_VSA2	2D	Bin 2 SlogP (-0.20,0.00]
SlogP_VSA3	2D	Bin 3 SlogP (0.00,0.10]
SMR_VSA1	2D	Bin 1 SMR (0.110,0.260]
SMR_VSA6	2D	Bin 6 SMR (0.485,0.560]
TPSA	2D	Topological Polar Surface Area (A**2)
VDistEq	2D	Vertex distance equality index

- Save figures and results: After the calculation, press 'Save', then in 'Figure1', 'Figure2', and 'Results' of the drop-down menu, 'Iteration Figure.jpg', 'Pareto Figure.jpg', and 'Results.xlsx' were saved, respectively. The interface of ECoFFeS is presented in Fig. 11.

3) RESULTS AND DISCUSSION

In 'Figure' panel of Fig. 11, the left chart presents the non-dominated descriptor subsets derived from MOEA/D in 10 independent runs, and the right chart presents the

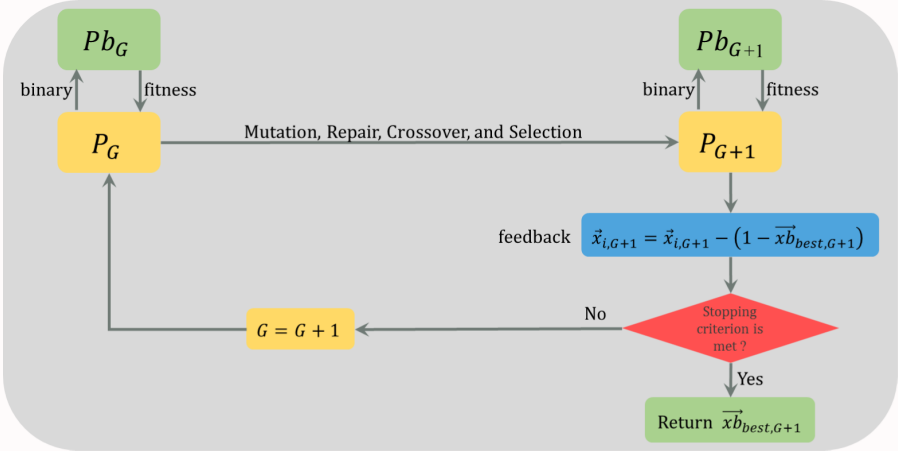


FIGURE 12. BFDE.

resulting non-dominated descriptor subsets from the left chart. According to the obtained Pareto optimal solutions (i.e., a set of descriptor subsets), a decision-maker can choose a preferred solution (i.e., a preferred descriptor subset) based on his/her own requirements. In this paper, we select the subset with 20 descriptors and these 20 descriptors are presented in Table 5. It is convenient for a drug developer to apply these descriptors for further model development.

VII. CONCLUSION

In this paper, a standalone software called ECoFFeS was developed to cope with feature selection problems in drug discovery. ECoFFeS can not only provide a user-friendly and easy-to-use GUI, but also offer an effective synergy of EAs and evaluation combinations to meet different requirements. Specifically, ECoFFeS have the ability to solve feature selection problems with different objectives (i.e., single-objective and multi-objective optimization problems), and with different kinds of response variables (i.e., classification and regression). From experimental studies, it was validated that feature selection is an indispensable part in drug discovery. The effectiveness of the proposed BFDE and necessity of MOEA/D-BFDE were verified through comparison experiments. Besides, we empirically demonstrated that the parallel execution in ECoFFeS can be efficiently used to reduce the runtime. Finally, we applied ECoFFeS to two of the real-life applications in drug discovery, i.e., predicting hERG blockers and predicting $\log D_{7.4}$. In the future, we plan to incorporate more search techniques into ECoFFeS and employ ECoFFeS to solve more feature selection problems in drug discovery.

For researchers who are interested in the further development of ECoFFeS, they can download the Matlab source code from: https://github.com/Jiaweihuang/ECoFFeS/tree/master/Others/Further_Development

APPENDIX

The process of BFDE is presented in Fig. 12, where

- $\mathbf{P}_G = \{\vec{x}_{1,G}, \vec{x}_{2,G}, \ldots, \vec{x}_{N,G}\}$ denotes the ordinary population in the current generation, whose individuals consist of float-point numbers.
- $\mathbf{Pb}_G = \{\vec{xb}_{1,G}, \vec{xb}_{2,G}, \ldots, \vec{xb}_{N,G}\}$ refers to the binary-digits population in the current generation, whose individuals consist of binary numbers.
- N denotes the population size.
- G denotes the generation number.

As shown in Fig. 12, BFDE contains three important components: binary transformation, normal DE, and feedback strategy.

First of all, the binary transformation is used to transform an individual in $\mathbf{P}_G$ (i.e., $\vec{x}_{i,G} = (x_{i,1,G}, x_{i,2,G}, \ldots, x_{i,n,G})$) into a binary-digits individual in $\mathbf{Pb}_G$ (i.e., $\vec{xb}_{i,G} = (xb_{i,1,G}, xb_{i,2,G}, \ldots, xb_{i,n,G})$). Actually, this process is implemented via Eq. (1) and Eq. (2).

$$xb_{i,j,G} = \begin{cases} 1, \text{ if } rand_j \le S(x_{i,j,G}) \\ 0, \text{ otherwise} \end{cases}, \quad j = 1, 2, \ldots, n. \tag{1}$$

$$S(x_{i,j,G}) = 1/(1 + e^{-x_{i,j,G}}), \quad j = 1, 2, \ldots, n. \tag{2}$$

where $rand_j$ is a uniformly distributed random number in the interval (0,1) for each j and $S(\cdot)$ refers to the sigmoid function. A brief description of the transformation process is presented in Fig. 13.

The normal DE contains three basic operators: mutation, crossover, and selection. For mutation, the "DE/current-to-best/1" operator is applied to create a mutant vector $\vec{v}_{i,G}$ for $\vec{x}_{i,G}$:

$$\vec{v}_{i,G} = \vec{x}_{i,G} + F \times (\vec{x}_{best,G} - \vec{x}_{i,G}) + F \times (\vec{x}_{r_1,G} - \vec{x}_{r_2,G}) \tag{3}$$

where $\vec{x}_{best,G}$ denotes the best individual in $\mathbf{P}_G$, r_1 and r_2 are two mutually different integers chosen from $[1, N]$ and also different from i, and F is the scaling factor.

Subsequently, the crossover operator is conducted on $\vec{x}_{i,G}$ and $\vec{v}_{i,G}$ to obtain a trial vector

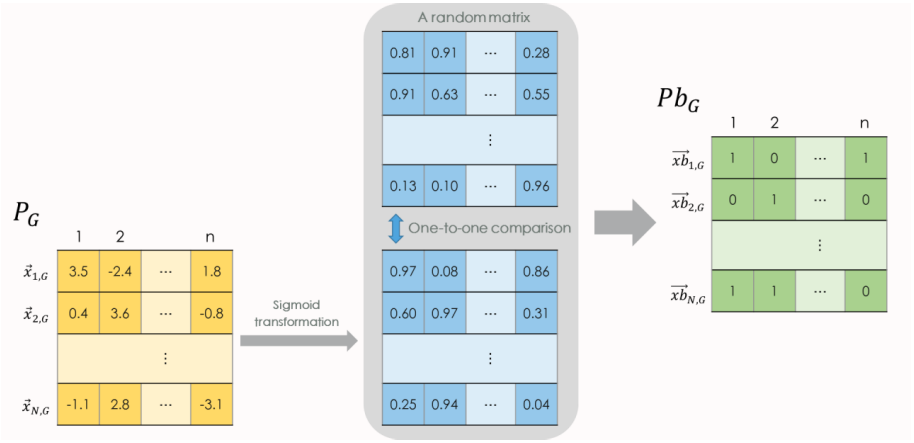


FIGURE 13. A brief description of the transformation process.

$\vec{u}_{i,G} = (u_{i,1,G}, u_{i,2,G}, \ldots, u_{i,n,G})$:

$$u_{i,j,G} = \begin{cases} v_{i,j,G}, & \text{if } rand_j \leq CR \text{ or } j = j_{rand} \\ x_{i,j,G}, & \text{otherwise} \end{cases} \tag{4}$$

where j_{rand} is a random integer in $[1, n]$, $rand_j$ is a uniformly distributed random number between 0 and 1 for each j, and CR denotes the crossover control parameter. The condition "$j = j_{rand}$" makes $\vec{u}_{i,G}$ different from $\vec{x}_{i,G}$ by at least one dimension.

Finally, the selection operator is used to select the better one between $\vec{u}_{i,G}$ and $\vec{x}_{i,G}$ to enter the next population $\mathbf{P}_{G+1}$. For a minimization problem, it can be described as:

$$\vec{x}_{i,G+1} = \begin{cases} \vec{u}_{i,G}, & \text{if } f(\vec{ub}_{i,G}) \leq f(\vec{xb}_{i,G}) \\ \vec{x}_{i,G}, & \text{otherwise} \end{cases} \tag{5}$$

where $\vec{ub}_{i,G}$ is the binary-digits individual corresponding to $\vec{u}_{i,G}$, $\vec{xb}_{i,G}$ is the binary-digits individual corresponding to $\vec{x}_{i,G}$, and $f(\cdot)$ is the fitness function.

The feedback strategy is designed to incorporate the information of the binary-digits population $\mathbf{Pb}_{G+1}$ into the ordinary population $\mathbf{P}_{G+1}$. It can be described as follows:

$$\vec{x}_{i,G+1} = \vec{x}_{i,G+1} - (1 - \vec{xb}_{best,G+1}) \tag{6}$$

where $\vec{xb}_{best,G+1}$ is the best individual in $\mathbf{Pb}_{G+1}$. In principle, for the jth feature in $\vec{x}_{i,G+1}$, if it has been selected in the $\vec{xb}_{best,G+1}$, $\vec{x}_{i,j,G+1}$ will kept the same; otherwise, the value of $\vec{x}_{i,j,G+1}$ will be decreased, thus reducing its probability to be chosen. By making use of the feedback information provided by $\mathbf{Pb}_{G+1}$, the convergence performance of DE can be enhanced.

REFERENCES

[1] B. D. Anson, J. Ma, and J.-Q. He, "Identifying cardiotoxic compounds," *Genetic Eng. Biotechnol. News*, vol. 29, no. 9, pp. 34–35, 2009.

[2] S. M. Paul *et al.*, "How to improve R&D productivity: The pharmaceutical industry's grand challenge," *Nature Rev. Drug Discovery*, vol. 9, no. 3, pp. 203–214, 2010.

[3] J. DiMasi, H. Grabowski, and R. W. Hansen, "Cost to develop and win marketing approval for a new drug is $2.6 billion," *Tufts Center Study Drug Develop.*, vol. 18, 2014.

[4] W. L. Jorgensen, "The many roles of computation in drug discovery," *Science*, vol. 303, no. 5665, pp. 1813–1818, 2004.

[5] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Adv. Bioinformatics*, vol. 2015, May 2015, Art. no. 198363.

[6] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, nos. 1–4, pp. 131–156, 1997.

[7] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, Aug. 2016.

[8] P. Langley *et al.*, "Selection of relevant features in machine learning," in *Proc. AAAI Fall Symp. Relevance*, vol. 184. 1994, pp. 245–271.

[9] L. Wang, Y. Wang, and Q. Chang, "Feature selection methods for big data bioinformatics: A survey from the search perspective," *Methods*, vol. 111, pp. 21–31, Dec. 2016.

[10] G. Sauter, R. Simon, and K. Hillan, "Tissue microarrays in drug discovery," *Nature Rev. Drug Discovery*, vol. 2, no. 12, pp. 962–972, 2003.

[11] S. Wang and Q. Cheng, "Microarray analysis in drug discovery and clinical applications," *Methods Mol. Biol.*, vol. 316, pp. 49–65, Oct. 2006.

[12] Y. Hsieh, M. S. Bryant, J. M. Brisson, K. Ng, and W. A. Korfmacher, "Direct cocktail analysis of drug discovery compounds in pooled plasma samples using liquid chromatography–tandem mass spectrometry," *J. Chromatography B*, vol. 767, no. 2, pp. 353–362, 2002.

[13] J. J. McCarthy and R. Hilfiker, "The use of single-nucleotide polymorphism maps in pharmacogenomics," *Nature Biotechnol.*, vol. 18, no. 5, pp. 505–508, 2000.

[14] P. K. Ojha and K. Roy, "Comparative QSARs for antimalarial endochins: Importance of descriptor-thinning and noise reduction prior to feature selection," *Chemometrics Intell. Lab. Syst.*, vol. 109, no. 2, pp. 146–161, 2011.

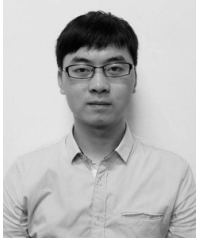
[15] Y. Wang, Z.-Z. Liu, J. Li, H.-X. Li, and G. G. Yen, "Utilizing cumulative population distribution information in differential evolution," *Appl. Soft Comput.*, vol. 48, pp. 329–346, Nov. 2016.

[16] Z. Huang, H. Tian, S. Fan, Z. Xing, and X. Zhang, "Social-aware resource allocation for content dissemination networks: An evolutionary game approach," *IEEE Access*, vol. 5, pp. 9568–9579, 2017.

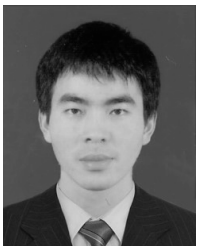
[17] R. E. Haber, G. Beruvides, R. Quiza, and A. Hernandez, "A simple multi-objective optimization based on the cross-entropy method," *IEEE Access*, vol. 5, pp. 22272–22281, 2017.

[18] R. Tanabe, H. Ishibuchi, and A. Oyama, "Benchmarking multi- and many-objective evolutionary algorithms under two optimization scenarios," *IEEE Access*, vol. 5, pp. 19597–19619, 2017.

[19] S. Yu, S. De Backer, and P. Scheunders, "Genetic feature selection combined with composite fuzzy nearest neighbor classifiers for hyperspectral satellite imagery," *Pattern Recognit. Lett.*, vol. 23, nos. 1–3, pp. 183–190, 2002.
[20] M. A. Shoorehdeli, M. Teshnehlab, and H. A. Moghaddam, "Feature subset selection for face detection using genetic algorithms and particle swarm optimization," in *Proc. IEEE Int. Conf. Netw., Sens. Control (ICNSC)*, Apr. 2006, pp. 686–690.
[21] S. Ahmed, M. Zhang, and L. Peng, "Improving feature ranking for biomarker discovery in proteomics mass spectrometry data using genetic programming," *Connection Sci.*, vol. 26, no. 3, pp. 215–243, 2014.
[22] A. A. Chaaraoui and F. Flórez-Revuelta, "Human action recognition optimization based on evolutionary feature subset selection," in *Proc. 15th Annu. Conf. Genet. Evol. Comput.*, 2013, pp. 1229–1236.
[23] S. F. da Silva, M. X. Ribeiro, J. do E. S. B. Neto, C. Traina, Jr., and A. J. M. Traina, "Improving the ranking quality of medical image retrieval using a genetic feature selection method," *Decision Support Syst.*, vol. 51, no. 4, pp. 810–820, 2011.
[24] O. Alomari and Z. A. Othman, "Bees algorithm for feature selection in network anomaly detection," *J. Appl. Sci. Res.*, vol. 8, no. 3, pp. 1748–1756, 2012.
[25] M. A. Avery *et al.*, "Structure-activity relationships of the antimalarial agent artemisinin. 6. The development of predictive *in vitro* potency models using CoMFA and HQSAR methodologies," *J. Medicinal Chem.*, vol. 45, no. 2, pp. 292–303, 2002.
[26] M. M. Neaz, M. Muddassar, F. A. Pasha, and S. J. Cho, "2D-QSAR of non-benzodiazepines to benzodiazepines receptor (BZR)," *Medicinal Chem. Res.*, vol. 18, no. 2, pp. 98–111, 2009.
[27] O. Nicolotti, V. J. Gillet, P. J. Fleming, and D. V. S. Green, "Multiobjective optimization in quantitative structure-activity relationships: Deriving accurate and interpretable QSARs," *J. Medicinal Chem.*, vol. 45, no. 23, pp. 5069–5080, 2002.
[28] B. O. Villoutreix and O. Taboureau, "Computational investigations of hERG channel blockers: New insights and current predictive models," *Adv. Drug Del. Rev.*, vol. 86, pp. 72–82, Jun. 2015.
[29] J.-B. Wang, D.-S. Cao, M.-F. Zhu, Y.-H. Yun, N. Xiao, and Y.-Z. Liang, "*In silico* evaluation of $\log D_{7.4}$ and comparison with other prediction methods," *J. Chemometrics*, vol. 29, no. 7, pp. 389–398, 2015.
[30] M. Dorigo, V. Maniezzo, and A. Colorni, "Ant system: Optimization by a colony of cooperating agents," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 26, no. 1, pp. 29–41, Feb. 1996.
[31] R. Storn and K. Price, "Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces," *J. Global Optim.*, vol. 11, no. 4, pp. 341–359, 1997.
[32] J. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, 2nd ed. Cambridge, MA, USA: MIT Press, 1992.
[33] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Netw.*, Dec. 1995, pp. 1942–1948.
[34] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Trans. Evol. Comput.*, vol. 11, no. 6, pp. 712–731, Dec. 2007.
[35] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.
[36] G. Pampara, A. P. Engelbrecht, and N. Franken, "Binary differential evolution," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2006, pp. 1873–1879.
[37] J. Kennedy and R. C. Eberhart, "A discrete binary version of the particle swarm algorithm," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. Comput. Simulation*, vol. 5. Oct. 1997, pp. 4104–4108.
[38] A. Zhou, B.-Y. Qu, H. Li, S.-Z. Zhao, P. N. Suganthan, and Q. Zhang, "Multiobjective evolutionary algorithms: A survey of the state of the art," *Swarm Evol. Comput.*, vol. 1, no. 1, pp. 32–49, 2011.
[39] R. B. Burns, *Introduction to Research Methods*. Reading, MA, USA: Addison-Wesley, 1997.
[40] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: A basic tool of chemometrics," *Chemometrics Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, 2001.
[41] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, 2004.
[42] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
[43] D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, vol. 398. Hoboken, NJ, USA: Wiley, 2013.
[44] W. W. Chin, "The partial least squares approach to structural equation modeling," *Modern Methods Bus. Res.*, vol. 295, no. 2, pp. 295–336, 1998.
[45] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Eur. Conf. Mach. Learn.*, 1998, pp. 137–142.
[46] P. Costa and J. M. S. Lobo, "Modeling and comparison of dissolution profiles," *Eur. J. Pharmaceutical Sci.*, vol. 13, no. 2, pp. 123–133, 2001.
[47] Q. Li and J. Racine, "Cross-validated local linear nonparametric regression," *Statist. Sinica*, vol. 14, no. 2, pp. 485–512, 2004.
[48] A. Afantitis, G. Melagraki, H. Sarimveis, P. A. Koutentis, O. Igglessi-Markopoulou, and G. Kollias, "A combined LS-SVM & MLR QSAR workflow for predicting the inhibition of CXCR3 receptor by quinazolinone analogs," *Mol. Diversity*, vol. 14, no. 2, pp. 225–235, 2010.
[49] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta-Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.
[50] J.-F. Truchon and C. I. Bayly, "Evaluating virtual screening methods: Good and bad metrics for the 'early recognition' problem," *J. Chem. Inf. Model.*, vol. 47, no. 2, pp. 488–508, 2007.
[51] M. Karelson, V. S. Lobanov, and A. R. Katritzky, "Quantum-chemical descriptors in QSAR/QSPR studies," *Chem. Rev.*, vol. 96, no. 3, pp. 1027–1044, 1996.
[52] H. Liu and Z. Zhao, "Manipulating data and dimension reduction methods: Feature selection," in *Computational Complexity*, R. Meyers, Ed., New York, NY, USA: Springer, 2012.
[53] Y. Wang, J.-J. Huang, N. Zhou, D.-S. Cao, J. Dong, and H.-X. Li, "Incorporating PLS model information into particle swarm optimization for descriptor selection in QSAR/QSPR," *J. Chemometrics*, vol. 29, no. 12, pp. 627–636, 2015.
[54] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci., National Taiwan Univ., Taipei, Taiwan, Tech. Rep., 2003.
[55] S. Wang, H. Sun, H. Liu, D. Li, Y. Li, and T. Hou, "ADMET evaluation in drug discovery. 16. Predicting hERG blockers by combining multiple pharmacophores and machine learning approaches," *Mol. Pharmaceutics*, vol. 13, no. 8, pp. 2855–2866, 2016.
[56] H. J. Witchel, "The hERG potassium channel as a therapeutic target," *Expert Opinion Therapeutic Targets*, vol. 11, no. 3, pp. 321–336, 2007.
[57] E. Raschi, V. Vasina, E. Poluzzi, and F. De Ponti, "The hERG K^+ channel: Target and antitarget strategies in drug development," *Pharmacol. Res.*, vol. 57, no. 3, pp. 181–195, 2008.
[58] M. J. Waring, "Lipophilicity in drug discovery," *Expert Opinion Drug Discovery*, vol. 5, no. 3, pp. 235–248, 2010.
[59] G. Ermondi, M. Lorenti, and G. Caron, "Contribution of ionization and lipophilicity to drug binding to albumin: A preliminary step toward biodistribution prediction," *J. Med. Chem.*, vol. 47, no. 16, pp. 3949–3961, 2004.
[60] M. Kah and C. D. Brown, "Log *D*: Lipophilicity for ionisable compounds," *Chemosphere*, vol. 72, no. 10, pp. 1401–1408, 2008.



ZHI-ZHONG LIU received the B.S. degree in automation from Central South University, Changsha, China, in 2013, where he is currently pursuing the Ph.D. degree in control science and engineering. His current research interests include evolutionary computation, bioinformatics, swarm intelligence, nonlinear equation systems, and multimodal optimization.

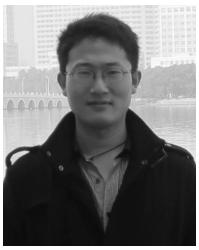


JIA-WEI HUANG received the B.S. degree in automation from Nanchang University, Nanchang, China, in 2011, and the M.S. degree in control science and engineering from Central South University, Changsha, China, in 2016. His research interests include evolutionary computation, drug discovery, and bioinformatics.

YONG WANG (M'08–SM'17) received the B.S. degree in automation from the Wuhan Institute of Technology, Wuhan, China, in 2003, and the M.S. degree in pattern recognition and intelligent systems and the Ph.D. degree in control science and engineering from Central South University (CSU), Changsha, China, in 2006 and 2011, respectively.

He is currently a Professor with the School of Information Science and Engineering, CSU. His current research interests include the theory, algorithm design, and interdisciplinary applications of computational intelligence.

Dr. Wang was a recipient of the Hong Kong Scholar by the Mainland-Hong Kong Joint Postdoctoral Fellows Program, China, in 2013, the Excellent Doctoral Dissertation by Hunan Province, China, in 2013, the New Century Excellent Talents in University by the Ministry of Education, China, in 2013, the 2015 IEEE Computational Intelligence Society Outstanding Ph.D. Dissertation Award, the Hunan Provincial Natural Science Fund for Distinguished Young Scholars, in 2016, the EU Horizon 2020 Marie Sklodowska-Curie Fellowship, in 2016, and a Highly Cited Researcher in computer science by Clarivate Analytics, in 2017. He is currently serving as an Associate Editor for the *Swarm and Evolutionary Computation*.



DONG-SHENG CAO received the B.S., M.S., and Ph.D. degrees in analytical chemistry from Central South University, Changsha, China, in 2006, 2009, and 2013, respectively. He is currently an Associate Professor with the Xiangya School of Pharmaceutical Sciences, Central South University, Changsha, China. His current research interests include cheminformatics, artificial intelligence, drug discovery, systems biology, and computational medicine.