ISSN1027-5495. Functional Materials, 32, No.4 (2025), p. 715-722
doi:http://dx.doi.org/10.15407/fm32.04.715

Data-driven discovery of functional materials: LARS–LASSO logistic regression for QSAR/QSPR design of compounds with anti-COVID-19 and other activities

M. I. Berdnyk, D. O. Anokhin, I. V. Khristenko, V. V. Ivanov, S. M. Kovalenko, O. N. Kalugin

School of Chemistry, V. N. Karazin Kharkiv National University,
Svobody sq., 4, Kharkiv, 61022, Ukraine

Received September 2, 2025

The possibility of using the L1-regularization to obtain logistic classification equations of quantitative/qualitative structure-activity/property relationships (QSAR/QSPR) have been investigated. The least angle regression (LARS) of least absolute shrinkage and selection operator (LASSO) variant has been implemented in the logistic regression. The method was used for building simple classification functions for three tasks: to evaluate basicity of different organic compounds towards Li^+ cation, to study binding affinity to the estrogen receptor of various organic molecules, and to predict activity against COVID-19 main protease. The obtained simple classification functions have satisfactory prognostic properties. The obtained results provide a foundation for the investigation of the electronic and spatial structures of potential ligands exhibiting the desired activity. A comparative analysis of chemoinformatics approaches facilitates the optimization of lead identification methodologies.

Keywords: QSAR-QSPR, Logistic Regression, classification, LARS, LASSO, COVID-19

Створення функціональних матеріалів на основі аналізу даних: логістична регресія LARS–LASSO для QSAR/QSPR-дизайну сполук з анти-COVID-19 та іншими властивостями. *М. І. Бердник, Д. О. Анохін, І. В. Христенко, В. В. Іванов, С. М. Коваленко, О. Н. Калугін*

Розглянуто можливість використання методу L1-регуляризації для отримання рівнянь логістичної класифікації в задачах знаходження кількісних/якісних співвідношень структура-активність/властивість (QSAR/QSPR). Використано регресію найменшого кута (LARS) у варіанті найменшого абсолютного скорочення та оператора відбору (LASSO), які було реалізовано в логістичній регресії. Метод використано для побудови простих класифікаційних функцій для трьох задач: оцінки основності різних органічних сполук по відношенню до катіону Li^+, дослідженню афінності зв'язування з рецептором естрогену різних органічних структур та прогнозування активності проти основної протеази COVID-19. Отримані прості класифікаційні функції мають задовільні прогностичні властивості. Результати розрахунків лежать в основі дослідження електронної та просторової будови потенційних лігандів заданої активності. Порівняльний аналіз результатів хемоінформаційних підходів дозволяє оптимізувати методи пошуку лідера.

1. Introduction

The modern process of designing new functional materials increasingly relies on theoretical approaches for predicting the properties of molecular systems and complexes. A primary objective of these methods is to model candidate-structures with desired properties and functions, while also reducing the vast design

space and subsequently optimizing structures toward specific target properties.

Various theoretical methods are used to evaluate molecular properties/activities. Among them, quantum-chemical approaches – including semi-empirical and ab initio methods – are central, with density functional theory (DFT) being especially important. DFT allows prediction of a wide range of physicochemical properties of functionalized materials.

When the focus shifts to designing compounds with biological activity relevance, such as toxicity, carcinogenicity, or therapeutic activity, QSAR/QSPR (quantitative structure –activity/property relationship) approaches are more appropriate and efficient. QSAR/QSPR methods constitute a key component of Materials Science Informatics, an interdisciplinary field that leverages machine learning techniques to accelerate the discovery, design, and development of novel materials. These approaches rely on the systematic analysis of experimental data and information extracted from existing databases, often facilitated by high-throughput screening (HTS) techniques. In this context, numerous databases containing information on various molecular properties, including physicochemical characteristics and biological activities, are available.

Thus, based on available experimental data, various predictive models of activity and properties can be developed using QSAR/QSPR methodology. Among the various QSAR/QSPR approaches, different regression models can be utilized to estimate the molecular properties of interest. Such models can evaluate continuous molecular features, but there is a need for additional approaches as well. Specifically, classification of molecular systems into a small number of classes (usually two classes: "active-inactive") can be useful for a rough estimation of activity and screening of databases with respect to a given target parameter.

Nowadays, among the classification methods the most used approaches are: discriminant analysis (usually linear discriminant analysis LDA) [1], random forests (RF) [2], support vector machine (SVM) [3], artificial neural networks [4]. Also it is known the methods based at Partial Least Squares (PLS) ideology for classification problem [5]. While the predictability of the mentioned methods for various tasks is reported to be high, they are considered to be "Black box" methods. This means that for molecules predicted to be either active or inactive, it is often difficult, if not impossible, to identify the specific structural reasons behind their (bio)activity. Thus, such methods fail to provide general information on why the molecule possesses the corresponding properties. While methods for constructing interpretable convolutional neural networks are still being developed [6], there remains a need for approaches that provide a general understanding of why molecules are active or inactive. In a more general sense, such approaches should focus on identifying the most important descriptors that convey key information about the system under consideration. Nowadays, there are multiple packages to collect such descriptors, for example, [7].

When selecting a classification method, it is important to consider that amount of data in libraries is so large and continues to grow that not all modern computers can manage it. Therefore, the development and application of methods capable of handling substantial data quantities is crucial. In this work, we propose using the LARS-LASSO method developed by [8] to solve both of the aforementioned issues present in the current application of machine learning to chemical and biological problems.

In applied sciences, while building models in QSPR studies, one normally expects to obtain the real values of predicted property. However, for some properties the simple classification can appear to be enough, e. g., to discover the possibility of the reaction, to find out if compound is strong base or not, etc. That is why in this work we have also considered the application of classification approaches to chemical problem. We have compared the models obtained with the method proposed by us to models obtained with other common method, such as Random Forest method (RF). We demonstrate that the models obtained using LARS-LASSO logistic regression are comparable to those derived from the RF approach, while also offering interpretable models with understandable structural descriptors included in the equations.

2. Methods of calculations

In this work we have applied LARS-LASSO logistic regression proposed in [9]. The initial LASSO logistic regression classification problem can be formulated as follows:

$$p_i(\beta) = 1/(1+\exp(-f_i)),$$
$$f_i = \beta_0 + \sum_j \beta_j x_{ij}, \quad s.t. \quad \|\beta\|_1 < \xi, \qquad (1)$$

Where p_i is fitted probability of the i-th compound to appear active, β_j are parameters of the logistic regression model, x_{ij} are the corresponding molecular descriptors (descriptors marked with index j). To obtain the logistic regression solution, one needs to maximize the log-likelihood function, which in a binary case can be written as follows

$$L(\beta) = \sum_{i=1}^{n} \left\{ y_i \ln(p_i(\beta)) + (1 - y_i) \ln(1 - p_i(\beta)) \right\} \quad (2)$$

where y_i takes the value 1 if the i-th molecule is active or 0 if inactive. By introducing p_i from equation (1) into equation (2), calculating the derivatives and setting it to zero, one obtains:

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i \left(y_i - p_i(\beta) \right) = X^{+} \left(y - p(\beta) \right) = 0 \quad (3)$$

To solve the equation (3) we have used the Newton–Raphson algorithm. To do this one needs to obtain expression for the second-derivatives from equation (3). In matrix form, this expression can be written as follows:

$$\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^{+}} = -X^{+} W X \quad (4)$$

where W is a $n \times n$ diagonal matrix with i-th diagonal element $p_i(\beta)(1 - p_i(\beta))$. The single Newton step is:

$$\begin{aligned} \beta^{new} &= \beta^{old} - \left(\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^{+}} \right)^{-1} \frac{\partial L(\beta)}{\partial \beta} = \\ &= \beta^{old} + \left(X^{+} W X \right)^{-1} X^{+} \left(y - p \right) = \\ &= \left(X^{+} W X \right)^{-1} X^{+} W \left(X \beta^{old} + W^{-1} \left(y - p \right) \right) = \\ &= \left(X^{+} W X \right)^{-1} X^{+} W z \,. \end{aligned} \quad (5)$$

Thus, an iterative solution to the regular logistic regression problem can be formulated as an iteratively reweighted least squares task, with response z and weights W, which is solved in each iteration until the stopping criteria are met.

$$\left| \frac{\partial L(\beta)}{\partial \beta} \right| < \varepsilon \quad (6)$$

According to the algorithm [9], the solution to task (1) can be obtained by solving problem (5) in each iteration using the LARS-LASSO algorithm. Our previous studies investigated the application of the LARS-LASSO algorithm to linear regression problems [10,11]. Currently, the LARS-LASSO algorithm in the Python implementation is available in *scikit-learn* [12]. In the present work instead of maintaining $\|\beta\|_1 < \xi$ condition we have selected on each iteration the fixed amount of descriptors. The equation (5) is solved with updated on each iteration values of W and z until the difference between $\|y - p\|_2$ values on subsequent iterations appeared to be less than threshold value ~ 10^{-8}.

LARS-LASSO logistic regression algorithm was implemented on FORTRAN and Python programming language.

We also compared the results obtained by this method with currently widely used alternatives which were decision trees [13], and more sophisticated approach of RF [2]. The latter approach does not yield an interpretable solution to the classification problem, as it functions as a «black box» when making classification decisions. However, the RF method is reported to produce highly accurate classification models, making it essential for us to compare our approach to this method. The decision tree approach is provided in this work as a comparison method because it is one of the few widely used machine learning methods which does provide the interpretable solutions although it is known to be prone to overfitting [14]. The RF method was proposed to overcome this problem.

3. The data for classification

For the purposes of illustration of LARS-LASSO method, this work focuses on three tasks.

The first task is to distinguish molecules with high and low Li^+ cation basicity. Lithium and its salts has multiple technological application to name a few: application in lithium batteries, as reaction media [15], catalysis (see for example [16,17]), ion attachment mass spectrometry (IAMS) [18-20], *etc*. Therefore, we believe that understanding lithium cation basicity can provide insights into which structural properties need to be modified to enhance the application of lithium and its salts in the specified fields. For this task, Li^+ basicity set was taken from [21] compounds with higher than average cation basicity on the set were considered as active while others were taken as inactive, thus, 113 compounds were considered as active and 115 as inactive.

The second task is to study binding affinity to the estrogen receptor. Designing of classification models can gradually speed up the design of medicines which can be used for human estrogen replacement therapy. To build such models we have used DSSTox (NCTRER) National Center for Toxicological Research Es-

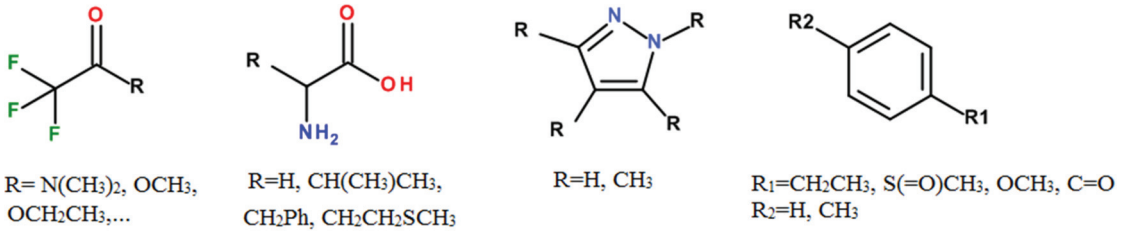

Fig. 1. Typical compounds of the Li-cation basicity dataset

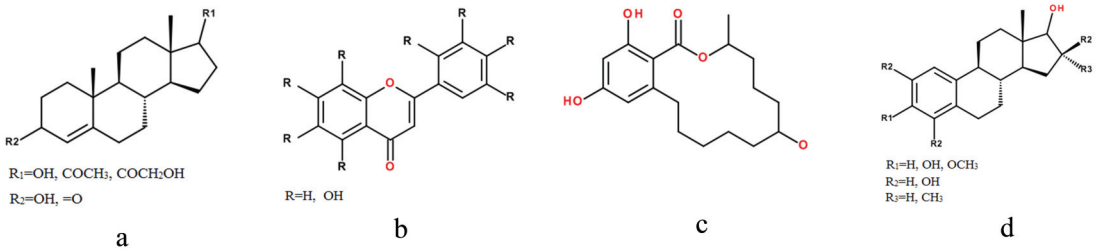

Fig. 2. Typical compounds of the DSSTox set dataset

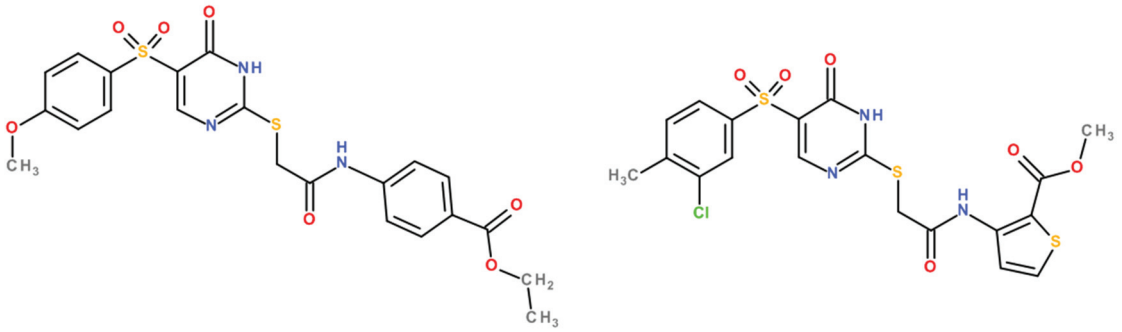

Fig. 3. Typical compounds of the anti-COVID dataset

trogen Receptor Binding Database [22] that is found in open access on PubChem. This set contains 216 molecules, of which 128 are defined as active and 88 as inactive.

The third task is to discover potent inhibitors of COVID-19 main protease (M^{pro}). This enzyme is essential for viral DNA replication thus such compounds can be used for treatment against COVID-19 disease [23]. To build respective classification models we are used a set of 424 derivatives of 2-(5-arylsulfonyl-4-oxo-3,4-dihydro-2-pyrimidinethio)acetamides [24] which are found as SARS-CoV-2 potential M^{pro} inhibitors. In the present set 149 molecules are considered as active and 275 as inactive.

The initial sets of molecules were converted to 3D and preoptimized with MMFF94 force field method available in the Python RDKIT package [25-29]. Then the descriptors with the use of PADEL-descriptor program [7] were calculated. Both 2D and 3D descriptors were used in the development of subsequent models for the data set.

Due to the predominantly 2D nature of the molecules in the set, only 2D descriptors were used for lithium cation basicity set. To ensure a rational selection of the test set and train set, the initial sets were clustered based on structural descriptor similarity using the K-means algorithm available in the *scikit-learn* package [30], forming clusters with similar structural properties. The number of clusters was taken to be 30% of the initial number of molecules. From the clusters containing more than 1 molecule in the test set one molecule was taken. From clusters with more than 4 molecules more than 1 molecule could be taken to the test set in order to get test set containing 30% of the initial dataset molecules. This ensures that at least one molecule in the training set shares similar structural descriptors with each molecule in the test set. Typical compounds of the datasets can be found on Figs. 1-3. As can be seen, the compound sets used in this work exhibit structural diversity.

Table 1. Brief description of used descriptors

Descriptor name	Dimension	Brief description
MATS3x	2D	Descriptor weighted by Gasteiger charge. x=p, v.
TIC1	2D	1-ordered neighborhood total information content
GATSxy	2D	Geary coefficient weighted by intrinsic state
MLFER_L	–	Solute gas-hexadecane partition coefficient
MLFER_BH	–	Summation solute hydrogen bond basicity
nHBAcc		Number of hydrogen bond acceptors
ETA_Shape_Y	2D	Extended topochemical atom shape index
minsOH	2D	Electrotopological state indices of OH groups bonded via a single bond
maxHsOH	1D	Maximum atom-type H Electrotopological State: OH
E1u	2D	1st component accessibility directional WHIM index
E1p	2D	Weighted by atomic polarizabilities 1st component accessibility directional WHIM index
TDB2i	3D	Autocorrelation
nX	1D	Number of halogen atoms
ATSC6s	2D	Centered Moreau-Broto autocorrelation of lag 6 weighted by intrinsic state
SHBint5	2D	E-State descriptor. Sum of the strengths of potential hydrogen bonds
nHetero Ring	1D	Number of heterocyclic rings
ndssC	1D	Number of C atoms bonded via a double bond and two single bonds
ETA_dAlpha_A	2D	Extended topochemical index

4. Numerical results

In this section, we compare the quality of the solutions obtained using the LARS-LASSO logistic regression method with the best results achieved through the RF approach. The performance of both methods was evaluated using receiver operating characteristic (ROC) curves and the corresponding area under the curve (AUC) metrics [31], calculated on the test dataset. The resulting ROC curves are presented in Figures 4–6.

To build these curves, a threshold of classification function is varied, *e. g.* if threshold equals to 0.5, then systems with p (calculated by formula (1)) is greater than 0.5, it is considered to be active, otherwise – inactive. Thus, varying the threshold one can vary false positive rate (*FPR*) and true positive rate (*TPR*).

$$FPR = \frac{FI}{FI + TI} \tag{7}$$

$$TPR = \frac{TA}{FA + TA} \tag{8}$$

Here, TA represents the number of active molecules correctly classified, TI denotes the number of inactive molecules correctly classified. FA refers to the number of active molecules misclassified as inactive, and FI indicates the number of inactive molecules misclassified as active. The AUC values are obtained through numerical integration of ROC curves. Larger AUC values indicate better predictive ability of a classification model.

The initial descriptor set comprised more than 2000 parameters. Descriptors selected through the L1-regularization procedure are briefly described in Table 1. A detailed description of the predictors used can be found in [32-35]. The results of calculation of logistic function parameters for samples describing basicity to lithium cation, affinity to estrogen receptor, and anti-COVID-19 activity are presented in Tables 2-4, respectively.

It can be observed that, for Li cation basicity (Fig. 4 a, Table 2), using more than two descriptors does not gradually improve model quality.

In contrast, for the DSSTox test set (Fig. 5, a, Table 3), it was essential to incorporate more descriptors for better data prediction. For the anti-COVID activity set (Table 4), the optimal number of descriptors is five, as adding more does not significantly improve predictive performance. (Fig. 6 a, Table 4). However, it is important to note that increasing the number of descriptors led to a gradual improvement in the predictive ability of the resulting models. Consequently, adding more descriptors improved

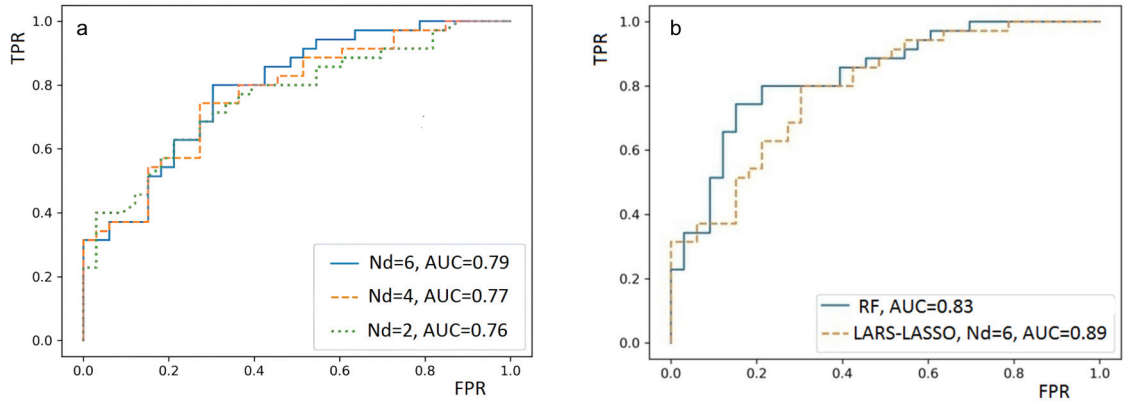

Fig. 4. The ROC-curve (TPR *vs* FPR) obtained for Li-cation basicity property test-set molecules classification. (a) comparison of the results obtained with LARS-LASSO logistic regression for different number of descriptors in model (Nd). (b) comparison LARS-LASSO to the best model obtained within RF approximation

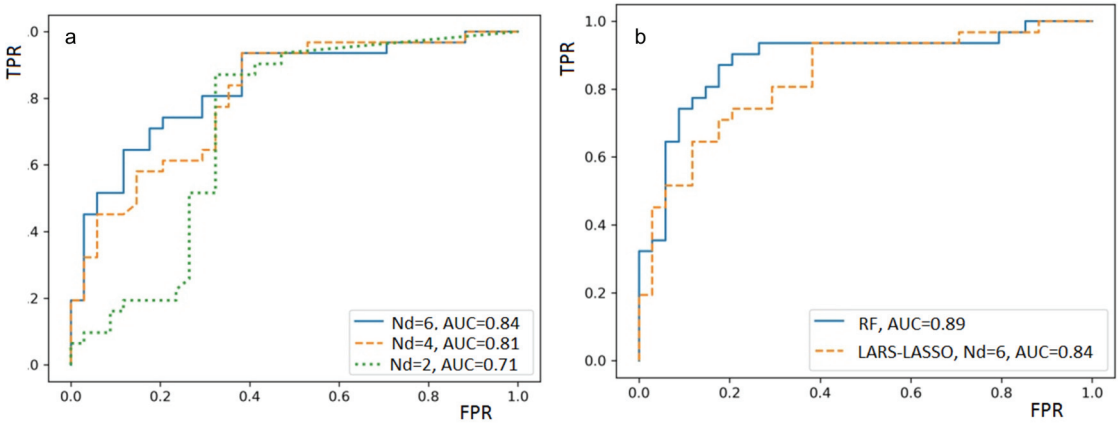

Fig. 5. The ROC-curve (TPR *vs* FPR) obtained for DSSTox test-set molecules classification. (a) comparison of the results obtained with LARS-LASSO logistic regression for different number of descriptors in model, Nd. (b) comparison to the best model obtained within RF approximation

the quality of the approximations (see Table 2). Nonetheless, increasing the number of descriptors beyond three for the first two sets and beyond five for the third set did not yield a higher number of correct classifications, even though the AUC criterion improved.

Not all methods demonstrate improved predictive performance with the inclusion of additional parameters. For instance, although the RF approach can produce predictions comparable in quality to those from the LARS-LASSO approximation, its performance gradually deteriorates beyond a certain number of decision steps due to overfitting (see Table 5). Meanwhile, for the DSSTox dataset, a decision tree consisting of only three nodes (determined by active descriptors) achieved approximately 81.5% classification accuracy on the test set.

Analyzing the decision tree obtained for the DSSTox set, we noticed that although all molecules in the training set are classified perfectly, some decisions were based on individual molecules. Such solutions tend to be incorrect more

Table 2. LARS-LASSO logistic regression coefficients ß for Li-cation basicity set.

Nd	Descriptor							AU train	AUC test
	β_0	TIC1	GATS4s	MLFER_L	MLFER_BH	nHBAcc	ETA_Shape_Y		
2	-0.49	0.010	0.13	–	–	–	–	0.91	0.76
4	-0.98	0.014	0.24	0.060	0.079	–	–	0.93	0.77
6	-1.55	0.016	0.29	0.14	0.19	0.039	0.39	0.94	0.79

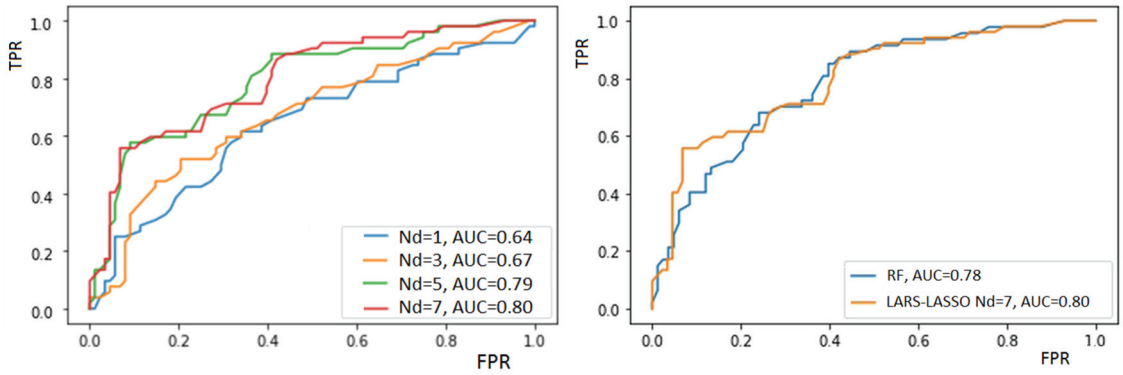

Fig. 6. The ROC-curve (TPR *vs* FPR) obtained for anti-COVID test-set molecules classification. (a) comparison of the results obtained with LARS-LASSO logistic regression for different number of descriptors in model. (b) comparison to the best model obtained within RF approximation

Table 3. LARS-LASSO logistic regression coefficients ß for DSSTox set

Nd	Descriptor							AUC train	AUC test
	β_0	minsOH	maxHsOH	GATS1i	E1u	E1p	TDB2i		
2	-0.598	0.097	0.491	–	–	–	–	0.76	0.71
4	-0.253	0.100	0.557	-0.145	-0.475	–	–	0.84	0.81
6	6.76	0.092	1.02	-0.69	-4.21	-0.310	-0.0135	0.88	0.84

Table 4. LARS-LASSO logistic regression coefficients ß for anti-COVID-19 set

Nd	Descriptor									AUC Train	AUC test
	β_0	MATS3p	nX	MATS3v	ATSC6s	SH-Bint5	nHetero Ring	ndssC	ETA_d Alpha_A		
1	-3.256	-14.816	–	–	–	–	–	–	–	0.72	0.70
3	-1.774	8.656	0.708	-17.255	–	–	–	–	–	0.77	0.72
5	-6.720	–	0.708	-13.270	-0.00975	0.0841	2.524	–	–	0.83	0.80
7	-7.228	–	0.870	-15.206	-0.01202	0.0534	2.823	1.113	-59.856	0.85	0.83

often than not, which can negatively affect the classification of the test set.

In comparison with the RF method, the models generated using the LARS-LASSO logistic regression approach demonstrate competitive performance, as shown in Figures 4b, 5b, and 6b. The LARS-LASSO logistic regression, in particular, yields models of comparable predictive quality while offering the advantage of interpretability. Notably, in the case of the anti-COVID dataset, the LARS-LASSO solution exhibits even higher accuracy. The limitations of the RF approach include not only its lack of interpretability but also the variability of its outcomes, which depend on the randomization seed. Although the resulting differences may be relatively minor, we regard the reproducibility of the LASSO-based models as a significant advantage.

Table 5. Part of correct classifications obtained with decision trees (RF) classification for test set of DSStox set

Maximum depth of the decision tree.	% of correct classifications for the test set
2	76.9
3	81.5
4	75.4
5	78.5
7	78.5
10	73.8

5. Conclusion

LARS-LASSO logistic regression offers interpretable solutions, which is a significant advantage over many modern methods. The results presented here demonstrate that models obtained using this method are comparable to those produced by contemporary approaches,

such as random forests. Several logistic equations have been derived to estimate parameters such as affinity for the lithium cation. Simple classification functions are presented for selecting organic systems of different natures, such as those classified by basicity towards lithium cations (strong or weak bases) or affinity for estrogen receptors (active or inactive). We also obtained a number of equations that can be used to make rough estimates of antiviral activity (anti-Covid-19). These equations are quite simple yet have a prediction quality comparable to that of known, complex approaches. Additionally, LARS-LASSO logistic regression is computationally efficient, making it suitable for handling large datasets.

Acknowledgements

V.I, S.K. and O.K acknowledge Grant 42/0062 (2021.01/0062) "Molecular design, synthesis and screening of new potential antiviral pharmaceutical ingredients for the treatment of infectious diseases COVID-19" from the National Research Foundation of Ukraine.

References

1. A. Tharwat, T. Gaber, A. Ibrahim, A.E. Hassanien, *AI Communications,* **30(2)**, 169 (2017).
2. L. Breiman, *Machine Learning,* **45**, 5 (2001).
3. I. Steinwart, A. Christmann, in: Support Vector Machines, Information Science and Statistics. Springer, New York, P. 285 (2011).
4. G.P. Zhang, *IEEE Transactions on Systems, Man, and Cybernetics, Part C* (Applications and Reviews), **30(4)**, 451 (2000).
5. V.E. Vinzi, W.W. Chin, J. Henseler, H. Wang, in: Handbook of Partial Least Squares. Concepts, Methods and Applications, Springer, Heidelberg, P. 195 (2010).
6. Q. Zhang, Y. N. Wu, S.-C. Zhu, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, P. 8827 (2018).
7. C.W. Yap, *J. Comput. Chem*., **32(7)**, 1466 (2011).
8. R.J. Tibschirani, *Electron. J. Statist*., **7**, 1456 (2013).
9. S-I. Lee, H. Lee, P. Abbeel, A, Ng, in: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), 1, 401 (2006).
10. M.I. Berdnyk, A.B. Zakharov, V.V. Ivanov, *Methods Objects Chem. Anal*., **14 (2)**, 90 (2019).
11. M.I. Berdnyk, M.O. Onizhuk, V.V. Ivanov, *Kharkov Univ. Bull. Chem. Ser*., **30**, 6 (2018).
12. Lasso model fit with Least Angle Regression a.k.a. Lars. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoLars.html
13. Y-Y. Song, Y. Lu, *Shanghai Arch Psychiatry,* **27(2)**, 130 (2015).
14. D.M. Hawkins, *J. Chem. Inf. Comput. Sci*., **44 (1)**, 1 (2003).
15. K. Fujiki, S. Ikeda, H. Kobayashi, A. Mori, A. Nagira, J. Nie, T. Sonoda, Y. Yagupolskii, *Chem. Lett*., **29 (1)**, 62 (2000).
16. S. Moss, B. T. King, A. de Meijere, S. I. Kozhushkov, P. E. Eaton, J. Michl, *Org. Lett*., **3 (15)**, 2375 (2001).
17. V. Volkis, H. Mei, R. K. Shoemaker, J. Michl, *J. Am. Chem. Soc*., **131(9)**, 3132 (2009).
18. T. Fujii, *Mass Spectrom. Rev.* **19(3)**, 111 (2000).
19. M. Sablier, T. Fujii, *Chem. Rev*., **102(9)**, 2855 (2002).
20. S. Takahashi, M. Nakamura, T. Fujii, *J. Am. Soc. Mass Spectrom*. **23(3)**, 547 (2012).
21. J. Jover, R. Bosque, J. Sales, *J. Chem. Inf. Comput. Sci*. **44**, 1727 (2004).
22. DSSTox (NCTRER) National Center for Toxicological Research Estrogen Receptor Binding Database. https://pubchem.ncbi.nlm.nih.gov/bioassay/1204
23. Z. Jin, X. Du, Y. Xu, Y. Deng, *et al. Nature.* **582**(7811), 289 (2020).
24. D.O. Anokhin, S.M. Kovalenko, P.V. Trostianko, A.V. Kyrychenko, A.B. Zakharov, T.O. Zubatiuk, V.V. Ivanov, O.M. Kalugin, *Kharkiv University Bulletin. Chemical Series*., **42**, 6 (2024).
25. T. A. Halgren, *J. Comp. Chem*., **17**, 490 (1996).
26. T. A. Halgren, *J. Comp. Chem*., **17**, 520 (1996).
27. T. A. Halgren, *J. Comp. Chem*., **17**, 553 (1996).
28. T. A. Halgren, R. B. Nachbar, *J. Comp. Chem*., **17**, 587 (1996).
29. T. A. Halgren, *J. Comp. Chem*., **20**, 720 (1999).
30. K-Means https://scikit-learn.org/1.5/modules/generated/sklearn.cluster.KMeans.html
31. P. A. Bradley, *Pattern Recognition.* **30**(7), 1145 (1997).
32. R. Todeschini, V. Consonni, Methods and Principles in Medicinal Chemistry. Weinheim: Wiley VCH, 2009.
33. J. A. Platts, D. Butina M. H. Abraham, A. Hersey, *J. Chem. Inf. Comput Sci.*, **39**(5), 835 (1999).
34. K. Roy, G. Ghosh, *J. Chem. Inf. Comput. Sci*., **44**, 559 (2004).
35. L. H. Hall, L. B. Kier, *J. Chem. Inf. Comput. Sci*., **35**, 1039 (1995).