



# Development of classification models for identification of important structural features of isoform-selective histone deacetylase inhibitors (class I)

M. S. Neiband<sup>1,2</sup> · A. Benvidi<sup>1</sup> · A. Mani-Varnosfaderani<sup>3</sup>

Received: 16 August 2019 / Accepted: 2 November 2019 / Published online: 26 November 2019  
© Springer Nature Switzerland AG 2019

## Abstract

As one of the hot topics in the epigenetic studies, histone deacetylases inhibitors (HDACIs) have been introduced to treat a variety of diseases such as cancer, immune disorder and neuronal diseases. Given the high numbers of available pathways in which HDACs are involved, the HDACIs that act particularly on Class I or Class II enzymes are considered as possible candidates for anticancer drugs. Due to their effective roles in the onset of cancer and its progression, HDAC Class I isoforms (HDAC 1, 2, 3 and 8) were considered in this study. Herein, our objective is to determine the important isoform-selective and isoform-active structural features of HDACIs using the valid classification models. For this purpose, a diverse dataset comprising 8224 HDAC modulators was collected from the binding database. To identify the significant discriminative features, five classification models were generated by supervised Kohonen network and support vector machine methods. Variable importance in projection method was used as a variable selection approach. The results obtained from descriptor analysis show that physicochemical properties, such as hydrogen bonding, number of branches, size, flexibility, polarity and sphericity in the structure of molecules, were closely related to the bioactivity of HDACIs. The reliability and predictive ability of the conducted models were evaluated using the tenfold cross-validation techniques, test sets and applicability domain analysis. All of the obtained classification models represented high statistical quality and predictive ability with accuracy greater than 85% for the test sets. The proposed strategy and the selective patterns represented in this paper can be applied by researchers in the pharmaceutical sciences who aim to use the same idea for the design of drugs with improved anticancer properties.

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11030-019-10013-0>) contains supplementary material, which is available to authorized users.

---

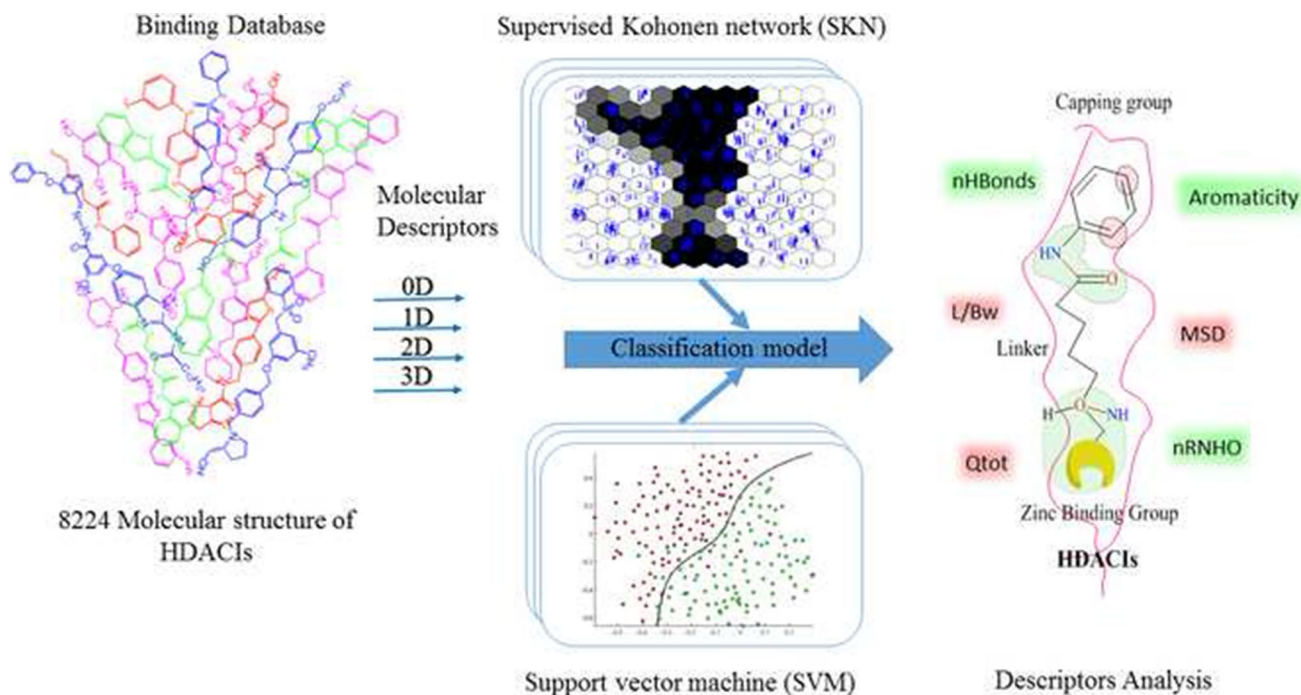
✉ A. Benvidi  
abenvidi@yazd.ac.ir; benvidi89@gmail.com

<sup>1</sup> Department of Chemistry, Yazd University, Yazd, Iran

<sup>2</sup> Department of Chemistry, Payame Noor University (PNU), P.O. Box 19395-3697, Tehran, Iran

<sup>3</sup> Chemometrics and Chemoinformatics Laboratory, Tarbiat Modares University, Tehran, Iran

## Graphic abstract



**Keywords** Isoform-selective HDAC inhibitors · Supervised Kohonen network · Support vector machine · Classification models · Anticancer agents

## Introduction

Histone acetyltransferases (HATs) and histone deacetylases (HDACs) are two groups of epigenetic enzymes which play a critical role in the regulation of gene expression. HATs acetylate the *N*- $\epsilon$ -amino groups of lysine residues in histone proteins allowing specific transcription factors to access the DNA sequence. On the contrary, HDACs catalyze the deacetylation of *N*- $\epsilon$ -amino groups of lysine residues in histone and non-histone proteins. Disruption of this balance due to the improper HDAC activity causes the changes in the gene expression and chromatin structure. This inappropriate gene expression leads the cause of different types of cancer [1]. In fact, inhibition of HDAC using inhibitors is related to the arrest of cell growth, cell differentiation and apoptosis in the several cancer cell lines [2].

HDAC inhibition is a new approach to interfere in the cell cycle regulation. Hence, it has a major therapeutic potential in the treatment of many diseases. Up to now, total of eighteen genes encoding HDAC family members are known in humans which grouped into four classes and two major categories:  $Zn^{2+}$  and  $NAD^+$  [3]. The classical HDACs (Class I, II and IV) are  $Zn^{2+}$ -dependent enzymes, whereas Class III HDACs (Sirtuins) needs the cofactor

$NAD^+$  (nicotinamide adenine dinucleotide) for their deacetylation activity [3]. Researchers confirmed that different HDAC isoforms play distinct roles in expanding a wide variety of tumors.

In this study, Class I of HDACs, consisting of HDACs 1, 2, 3 and 8, was selected to build a proper model. They have the strongest enzymatic activity among the other HDAC classes and great ability to control many functional and regulatory mechanisms [4]. The overexpression of these HDACs is associated with neurological diseases, immune disorders, [5] inflammation, [6] cardiovascular disease [7] and the variety of cancers such as breast, colon, lung, prostate and gastric cancers [8–11]. To date, five HDACIs (vorinostat, romidepsin, belinostat, chidamide and panobinostat) have been approved by Food and Drug Administration (FDA) as anticancer drugs [12]. However, lack of selectivity against a specific target in many HDACIs causes toxicity in the healthy cells. Thus, the isoform-selective inhibitors have great advantages in the development of a drug with less side effects.

In an attempt to discover HDACIs, various *in silico* approaches including ligand-based QSAR, [13–16] 3D-QSAR, [17–19] pharmacophore, [20–22] and structure-based molecular docking [23] of new

potential HDACIs have been introduced, which make significant contributions to the molecular understanding of enzyme–drug interactions. The applicability domains of these studies in the most cases are restricted by limited diversity (< 50 compounds in most cases) [13, 14, 24, 25] or scaffold type (e.g., hydroxamic acid derivatives only) [24, 25] in the training dataset. On the other hand, only some types of machine learning methods have been the subject of several structure–activity relationship (SAR) modeling studies of HDACIs. Also, some studies have been accomplished for the classification of isoform-selective HDACIs based on their activities [26–28].

The main purpose of our investigation is to consider more targets and ligands and consequently constructs a reliable model, in order to identify key features of isoform-selective and isoform-active HDACIs. Thus, we collected a large dataset containing 8224 HDACIs from binding database (BD) [29] and established five classification models based on support vector machine (SVM) and supervised Kohonen network (SKN) methods. The variable importance in projection (VIP) method was used to select the suitable set of molecular descriptors from 3224 molecular descriptors calculated by DRAGON software [30]. The statistical evaluation of the developed classification models was implemented by parameters derived from confusion matrix. The results indicated that in binary active/inactive classifiers for HDACs 1 and 2, compared to SKN method, SVM gives better classification accuracy rates. In the discrimination between active/inactive HDACIs 3 and 8, both methods have approximately equal potency. In the multiclass classification, SVM shows a better ability to discriminate isoform-selective HDACIs.

In this research, reliable classification models are developed to discover the significant discriminant features through the study of structure-isoform-activity and structure-isoform-selectivity relationships of Class I HDACIs.

## Materials and methods

### Dataset

In this work, four datasets including 8224 molecules of HDAC1, HDAC2, HDAC3 and HDAC8 together with their corresponding biological activity (in terms of  $IC_{50}$  (nM)) and chemical structure information were collected from the binding database as 3D-SD files.

The data in each group were categorized to active, moderate and inactive sets based on  $IC_{50}$  values. The compounds with  $IC_{50} \leq 300$  nM were considered as active, while those with  $IC_{50} \geq 3000$  nM were classified as inactive ones, and compounds with the  $IC_{50}$  values within the range of 300–3000 nM were proposed as moderate and were not used in the model development. According to this definition, the

HDAC1 group consists of 2350, 951 and 1063, HDAC2 group includes 500, 370 and 326, HDAC3 group contains 651, 110 and 526 and HDAC8 group has 301, 590 and 484 active, moderate and inactive molecules, respectively.

The molecular weight (MW), calculated LogP (LOGP), polar surface area (PSA), number of donor atoms for H-bonds (nHAcc) and number of acceptor atoms for H-bonds (nHDon) values for collected molecules were in the ranges of 137 to 1591, –16.3 to 7.95, 20.23 to 1590.11, 1 to 34 and 0 to 21, respectively. The average values of MW, LOGP, PSA, nHAcc and nHDon were 404.36 ( $\pm 108.01$ ), 3.31 ( $\pm 2.26$ ), 146.01 ( $\pm 142.34$ ), 6.75 ( $\pm 2.50$ ), and 3.18 ( $\pm 1.43$ ), respectively. (Standard deviation values are given in parentheses.) The values of these properties reveal that the considered molecules define a reasonable drug-like chemical space.

The optimized 3D molecular structures (as.hin files), input data matrices (as MATLAB files) and SD files downloaded from the binding database were given in Online Resource 1.

### 3D molecular structure generation

To prepare the 3D structure of the compounds for the descriptor calculation, we used Open Babel software (ver. 2.4.0) [31]. The preparation steps were carried out as follows: first the hydrogen atoms were added to the SD files. Afterward, the 3D structures of the molecules were generated. The partial atomic charges were assigned according to Merck molecular force field (MMFF94), and then any duplicate conformers were removed.

As a result, 622, 251, 374 and 124 molecules with the same structure of groups HDAC1, HDAC2, HDAC3 and HDAC8 were eliminated, respectively.

Subsequently, the 3D geometry optimization of molecular structures was carried out by implementing MMFF94 force field. The energy optimization parameters (such as maximum number of steps = 2500, use the steepest descent algorithm, convergence criteria =  $10^{-6}$  kcal mol<sup>-1</sup>) were set at their default values.

Finally, the SD files were divided into numbered output files consecutively and simultaneously converted into .hin files format. The resulting 3D structures were imported into Dragon software (version 5.5) to calculate the molecular descriptors.

### Calculation and selection of molecular descriptors

Dragon software was used to calculate 3224 molecular descriptors for each chemical structure. These descriptors include 22 blocks of 0D (atom and molecular counts, molecular weight, sum of atomic properties); 1D (fragment

counts); 2D (topological descriptors); and 3D (geometrical, atomic coordinates) descriptors for encoding the ligands. To improve the predictive power and interpretability of the models, it is necessary to filter the descriptors. It is in need to use properties with a simple and applicable interpretation. For these reasons, a suitable subset of 450 descriptors was selected for building the classification models. Then from among them, the variables with 90% zero or constant values were eliminated. Also, from the two descriptors with correlation more than 0.9, the one showing higher pair correlation with all other descriptors was removed from the data matrix. Afterward, to reduce the bias of the models and to present an equal contribution for each variables in the data analysis, several preprocessing methods (such as variance scaling, rang scaling, mean centering and auto-scaling (centering + variance scaling)) were performed on the remaining variables. In the previous study, we examined the effect of using preprocessing methods on the accuracy of classification results and variables selection [32]. Herein, the best results were achieved by auto-scaled (scaling and centering) descriptors.

Finally, we used the VIP scores from partial least square discriminant analysis (PLS-DA) algorithm as a variable selection method [33]. In this method, the importance of descriptors was detected in determining class membership of response vector. For this propose, to each descriptor was calculated a VIP coefficient based on the PLS loadings. The  $VIP_j$  coefficient for the  $j$ th feature was calculated from a weighted sum of the squares of the weight in the PLS analysis according to Eq. 1.

$$VIP_j = \sqrt{\frac{p \sum_{f=1}^F (b_f^2 t_f' t_f) (w_{jf})^2}{\sum_{f=1}^F (b_f^2 t_f' t_f)}} \quad (1)$$

where  $p$  is the total number of variables in the model,  $b_f$  and  $t_f$  are regression coefficient and score for the  $f$ th latent variable, respectively. Also,  $w_{jf}$  is a weight for the  $j$ th variable and  $f$ th latent variable [34]. The  $VIP_j$  is a measure of the influence of  $j$  variable in the PLS model. In this method, the VIP coefficients of variables have been obtained as a score vector, which are a summary of all the components and their related responses. In this vector, the variables were arranged in decreasing order of VIP (decreasing order of importance). The variables with higher VIP values are more important for modeling the responses.

We used tenfold cross-validation technique for optimizing the number of variables in this work. For this purpose, different models with 5 to 13 VIP variables were constructed by SVM and SKN methods and their accuracy values were compared in validation sets. The number of descriptors that provided the best classification accuracy in the validation

procedure was selected as the optimal number of variables. The prediction accuracy values in terms of the number of variables in training and validation sets for the SVM and SKN methods are listed in Table S1 as Online Resource 2.

Comparison of the accuracy values of these models shows that the best results were obtained by 10-variable models. Therefore, the sets of the first ten variables with the larger VIP coefficients (more important) were selected for the development of the SVM and SKN models. These VIP-molecular descriptors define a specific subspace from the whole chemical space which is mainly occupied by active and selective HDACIs.

## Dataset categorization

The classification models were developed into two different strategies with the purpose of discovering structure–activity and structure–isoform selectivity relationships of Class I HDACIs.

In the first strategy, molecules were classified according to their isoform selectivity. This model was made only with active inhibitors. The model is useful to discriminate between different active molecules and to determine important pharmacophores which lead to isoform-selective interactions with HDACs.

In the second strategy, for each isoform, active molecules were discriminated from inactive ones, using binary classifiers. These models discover distinctive molecular features to classify the molecules as active or inactive inhibitors. These classifiers can be useful when molecules are prioritized to screen databases or to design focused libraries.

The threshold values of the  $IC_{50}$  for classification of active or inactive compounds were presented in section “Dataset.”

To define isoform-selective inhibitors, we used the active inhibitors and “Selectivity Factor” (SF) parameter. The SF for each isoform is calculated as the  $IC_{50}$  ratio of an inhibitor against considered isoform to  $IC_{50}$  of the same inhibitor against any other isoforms. (For example, Eq. 2 is SF for HDAC1 relative to HDAC2.)

$$SF(\text{HDAC1/HDAC2}) = \frac{IC_{50}(\text{HDAC1})}{IC_{50}(\text{HDAC2})} \quad (2)$$

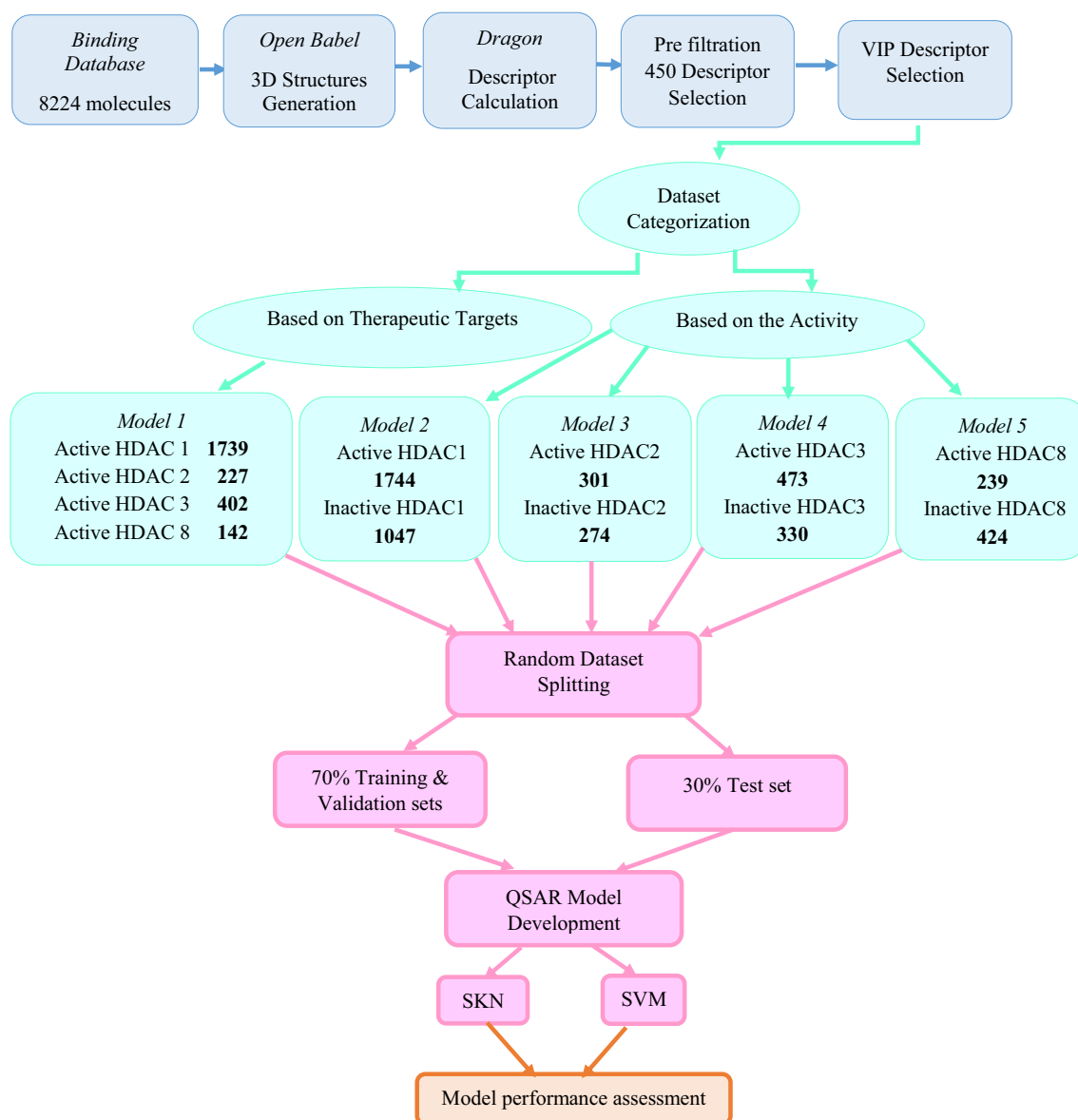
We used a 20-fold benchmark for determining highly selective inhibitors. Those compounds that showed  $SF \geq 20$  for one isoform over the other isoforms were considered selective and those that failed to satisfy these criteria as duplicate or common inhibitors were eliminated.

The  $IC_{50}$  and SF parameters have been selected based on the results reported in the studies on the function of HDACIs as active or selective inhibitors [22, 27, 35, 36].

Based on the above description, five different classifiers were created in this work. The first model was made by a multiclass classifier for discriminating selective inhibitors of HDAC1, HDAC2, HDAC3 and HDAC8 isoforms, which included 1739, 227, 402 and 142 selective inhibitors, respectively. The second model discriminated 1744 active and 1047 inactive inhibitors of HDAC1. The third model was designed to classify 301 and 274 active and inactive HDACs2. The fourth model was constructed to distinguish 473 active and 330 inactive HDACs3, and the fifth model is developed to categorize 239 active and 424 inactive HDACs8. In binary classifiers, the biological activity response was expressed in a ‘binary’ format (active = “1”;

inactive = “2”) and was correlated with molecular descriptors. In the multiclass classifier, a response vector of four classes was built and for each isoform was assigned a class label (Class “1” = HDACs1, Class “2” = HDACs2, Class “3” = HDACs3, Class “4” = HDACs8). By defining these thresholds, the number of molecules in each class is shown in Scheme 1.

After preparation of input matrices, in which their rows represent molecules and each column represents a VIP-selected descriptor, two machine learning techniques, namely SVM and SKN, were employed to develop classification models. The steps to build classification models were presented in Scheme 1.



**Scheme 1** Flowchart for classification models development

## Supervised Kohonen network (SKN)

The SKN is a supervised form of self-organizing map (SOM) derived from Kohonen maps. In SKN, Kohonen and output (Grossberg) layers are glued together to give a combined layer that is updated according to the training scheme of Kohonen maps [37]. You can see [38, 39] to know the training detail of the SKN method.

To obtain a SKN model with optimal performance, it is essential to optimize the number of neurons and epochs. For this purpose, according to the number of independent variables in this study, SKN models were built with different combinations of network sizes and epochs. The neurons were tested in numbers of 12, 14, 16, 18, 20, 22, 24, 28, 30 and 32, and the number of epochs was set to 20, 30, 40, 50, 80, 100, 150, 200, 250 and 300.

In this study, the network characteristics comprised hexagonal topology, toroidal boundary conditions, batch training and random initialization. SKN was calculated using the Kohonen\_CPANN\_toolbox written by Milano chemometrics and QSAR research group [40].

## Support vectors machine (SVM)

SVM represents a supervised machine learning technique based on the principle of structural risk minimization and the statistical learning theory of Vapnik [41]. This feature lends well to the SVM model and has a good generalization performance in classifying nonlinear and high-dimensional learning problems.

In this study, between examined polynomial, linear and radial basis (RBF) kernel functions, RBF provided good classification performance (more classification accuracy and smaller training time). In the RBF kernel, the values of two parameters  $C$  and  $\gamma$  were optimized using grid search method and tenfold CV of the training data set. This means that different pairs of  $(C, \gamma)$  were tested and ones that created the best CV accuracies were picked for further analysis. During the grid search processor,  $C$  was varied at this range (0.1, 1, 4, 8, 10, 20, 30, 40, 50, 60, 80, 100, 200, 300, 500, 1000), while  $\gamma$  was changed in the interval from (0.05, 0.10, 0.14, 0.20, 0.28, 0.34, 0.40, 0.50, 0.65, 0.80, 0.95, 1.13, 1.60, 2.26, 3.20, 4.60). In the end, optimal  $C$  and  $\gamma$  parameters were selected and used to build the SVM models. The resulting values for these parameters in each of the classification models are given as footnotes in Tables 1, 2, 3, 4 and 5.

The results show that “one-versus-rest” SVM approach works better for multiclass data, so we adopted this method for the classification of isoform-selective compounds. In sum, this approach builds a separate SVM model to classify each class against the rest. This constructs  $k$  separate binary classifiers for the classification of  $k$ -classes, and the  $k$ th binary classifier is trained using the data from the  $k$ th

class as positive examples and the remaining  $k - 1$  classes as negative examples. For a new sample, the class label is determined by the binary classifier with the strongest vote.

SVM analysis was carried out using the libSVM software library developed in 2018 [42] in the MATLAB interface R2017b.

## Model evaluation

The statistical significance of all classifiers has been assessed with several metrics derived from the confusion matrix including: sensitivity, specificity, precision and non-error rate for training, validation and test sets. The accuracy and Matthews correlation coefficient (MCC) values which refer to the ratio of correctly classified compounds were calculated to measure the overall performance of classifiers.

The stability and predictive ability of the generated classification models were validated by tenfold CV procedure, test sets and applicability domain (AD) analysis.

We used a random method to split the datasets into the 70% calibration (training and validation) and 30% test sets. The training and validation sets were used in the model building process to find and fix the best combination of parameters and initial evaluation of model's performance. Further predictive abilities and generalization performance of the optimized models were assessed by test sets.

In tenfold Venetian-Blinds CV approach, training set was randomly divided into 10 subsets ( $k = 10$ ), 9 subsets ( $k - 1$ ) were used as training set, and the remaining single subset was retained as validation set to predict the trained model. This process is repeated 10 times until each subset is evaluated once. The final result was obtained from the average result over each of the  $k$  models.

In addition, the graphical representation of the models performance was shown by receiver operating characteristic (ROC) curves and SKN maps.

## Applicability domain (AD)

The reliability of the models in chemical space created by selected descriptors was evaluated by applicability domain (AD). AD of the models was evaluated by “Leverage” analysis. This approach allows the determination of the position of new compounds within the chemical space and the activity range of the constructed models [43]. The leverage value for each compound is calculated by the following Eq. 3:

$$h_i = x_i^t (X^T X)^{-1} x_i \quad (3)$$

where  $X$  is the descriptor matrix of the training set and  $x_i$  is the descriptor row vector of the desired compound  $i$ . The limit of model domain is quantitatively determined by critical threshold for leverage as:  $h^* = 3(p + 1)/n$ , where  $n$  is

**Table 1** Comparative statistical results of SVM and SKN models for classification of the isoform-selective HDACs in the training, validation and test sets

	Sensitivity	Specificity	Precision	Non-error rate	Accuracy
<i>Training</i>					
SVM					
Class 1	0.947 ± 0.021	0.603 ± 0.022	0.841 ± 0.025	0.642 ± 0.025	0.829 ± 0.027
Class 2	0.671 ± 0.019	0.992 ± 0.022	0.766 ± 0.018		
Class 3	0.695 ± 0.020	0.953 ± 0.019	0.746 ± 0.021		
Class 4	0.658 ± 0.024	0.999 ± 0.022	0.949 ± 0.025		
SKN					
Class 1	0.907 ± 0.020	0.662 ± 0.025	0.856 ± 0.022	0.647 ± 0.027	0.803 ± 0.026
Class 2	0.381 ± 0.029	0.963 ± 0.022	0.507 ± 0.028		
Class 3	0.651 ± 0.024	0.949 ± 0.020	0.714 ± 0.021		
Class 4	0.649 ± 0.022	0.983 ± 0.025	0.698 ± 0.024		
<i>Validation</i>					
SVM					
Class 1	0.751 ± 0.019	0.459 ± 0.022	0.894 ± 0.023	0.496 ± 0.024	0.744 ± 0.026
Class 2	0.395 ± 0.026	0.912 ± 0.023	0.329 ± 0.021		
Class 3	0.465 ± 0.027	0.862 ± 0.030	0.492 ± 0.031		
Class 4	0.352 ± 0.029	0.925 ± 0.027	0.831 ± 0.025		
SKN					
Class 1	0.817 ± 0.028	0.515 ± 0.027	0.785 ± 0.025	0.529 ± 0.030	0.683 ± 0.032
Class 2	0.277 ± 0.024	0.935 ± 0.025	0.315 ± 0.029		
Class 3	0.509 ± 0.024	0.899 ± 0.025	0.498 ± 0.028		
Class 4	0.395 ± 0.031	0.976 ± 0.026	0.505 ± 0.028		
<i>Test</i>					
SVM					
Class 1	0.807	0.473	0.907	0.495	0.757
Class 2	0.444	0.978	0.374		
Class 3	0.543	0.923	0.542		
Class 4	0.357	0.996	0.833		
SKN					
Class 1	0.871	0.527	0.818	0.485	0.731
Class 2	0.109	0.967	0.250		
Class 3	0.569	0.895	0.477		
Class 4	0.393	0.987	0.647		

Note: Class 1=Active HDACs 1; Class 2=Active HDACs 2; Class 3=Active HDACs 3; Class 4=Active HDACs 8. The optimum parameters of the SVM are  $C=10$ ,  $\gamma=0.5$

the number of compounds in the training set and  $p$  is the number of VIP-selected descriptors. A test compound was considered to be an outlier that has a standardized residual more than three times of the considered standard deviation units, while compounds with a leverage value higher than  $h^*$  ( $h > h^*$ ) were identified as structurally influential or high leverage compounds [44].

## Results and discussion

### Classification of compounds based on therapeutic targets

#### Discrimination of selective HDAC1, HDAC2, HDAC3 and HDAC8 inhibitors from each other

In order to classify isoform-selective inhibitors of class I HDACs, a multiclass classifier was constructed. Based on the activity criterion and selectivity factor (SF) defined

**Table 2** Comparative statistical results of SVM and SKN models for classification of the active and inactive HDACIs 1 in the training, validation and test sets

	Sensitivity	Specificity	Precision	Non-error rate	Accuracy	MCC <sup>a</sup>
<i>Training</i>						
SVM	0.959 ± 0.051	0.902 ± 0.055	0.944 ± 0.042	0.931 ± 0.049	0.938 ± 0.064	0.888 ± 0.032
SKN	0.924 ± 0.054	0.864 ± 0.050	0.922 ± 0.042	0.894 ± 0.048	0.902 ± 0.048	0.789 ± 0.039
<i>Validation</i>						
SVM	0.898 ± 0.021	0.764 ± 0.025	0.868 ± 0.022	0.831 ± 0.028	0.849 ± 0.018	0.672 ± 0.034
SKN	0.855 ± 0.039	0.698 ± 0.044	0.830 ± 0.047	0.777 ± 0.095	0.798 ± 0.071	0.562 ± 0.045
<i>Test</i>						
SVM	0.927	0.781	0.860	0.854	0.867	0.724
SKN	0.885	0.825	0.879	0.855	0.860	0.711

Note: <sup>a</sup>MCC = Matthews correlation coefficient. The optimum parameters of the SVM are  $C = 10$ ,  $\gamma = 3.2$

**Table 3** Comparative statistical results of SVM and SKN models for classification of the HDACIs 2 in the training, validation and test sets

	Sensitivity	Specificity	Precision	Non-error rate	Accuracy	MCC
<i>Training</i>						
SVM	0.951 ± 0.054	0.931 ± 0.031	0.939 ± 0.031	0.941 ± 0.048	0.941 ± 0.048	0.882 ± 0.035
SKN	0.901 ± 0.042	0.848 ± 0.039	0.869 ± 0.036	0.875 ± 0.045	0.876 ± 0.047	0.754 ± 0.040
<i>Validation</i>						
SVM	0.839 ± 0.025	0.783 ± 0.027	0.813 ± 0.019	0.811 ± 0.022	0.813 ± 0.020	0.624 ± 0.026
SKN	0.815 ± 0.032	0.746 ± 0.029	0.783 ± 0.033	0.781 ± 0.041	0.783 ± 0.049	0.563 ± 0.032
<i>Test</i>						
SVM	0.914	0.930	0.930	0.922	0.922	0.844
SKN	0.930	0.879	0.927	0.904	0.904	0.815

Note: The optimum parameters of the SVM are  $C = 100$ ,  $\gamma = 2.26$

**Table 4** Comparative statistical results of SVM and SKN models for classification of the HDACIs 3 in the training, validation and test sets

	Sensitivity	Specificity	Precision	Non-error rate	Accuracy	MCC
<i>Training</i>						
SVM	0.953 ± 0.022	0.880 ± 0.036	0.922 ± 0.035	0.917 ± 0.029	0.924 ± 0.037	0.841 ± 0.030
SKN	0.948 ± 0.019	0.900 ± 0.021	0.933 ± 0.027	0.924 ± 0.033	0.928 ± 0.039	0.851 ± 0.032
<i>Validation</i>						
SVM	0.864 ± 0.024	0.780 ± 0.031	0.853 ± 0.025	0.822 ± 0.019	0.830 ± 0.021	0.658 ± 0.022
SKN	0.881 ± 0.026	0.738 ± 0.033	0.834 ± 0.025	0.809 ± 0.024	0.822 ± 0.034	0.625 ± 0.033
<i>Test</i>						
SVM	0.978	0.901	0.926	0.940	0.944	0.888
SKN	0.981	0.921	0.935	0.944	0.948	0.901

Note: The optimum parameters of the SVM are  $C = 10$ ,  $\gamma = 3.2$

in “Dataset categorization” section, active and selective compounds were selected for each isoform and an input dataset matrix was made with a dimension of  $(1739 + 227 + 402 + 142) \times 450$ . By disregarding descriptors with constant or near constant values inside each column, 200 descriptors remained to continue analysis. The names and brief definitions of 10 VIP-selected descriptors along with their average values in each class are listed in supplementary material Table S2. It can be observed from this table that the most significant molecular descriptors belong

to three classes of functional group counts (FGCs), topological indices and atom-centered fragments (ACFs) descriptors and include: number of Imidazoles (nImidazoles), number of intramolecular H-bonds (with N, O, F), (nHBonds), number of Pyridines (nPyridines), mean square distance (MSD) index,  $\text{CH}_2\text{R}_2$  (C-002), Narumi harmonic topological index ( $H_{\text{Nar}}$ ), H attached to heteroatom (H-050),  $\text{CR}_3\text{X}$  (C-011), Narumi geometric topological index ( $G_{\text{Nar}}$ ) and Kier benzene-likeness index (BLI).



**Table 5** Comparative statistical results of SVM and SKN models for classification of the active and inactive HDACIs 8 in the training, validation and test sets

	Sensitivity	Specificity	Precision	Non-error rate	Accuracy	MCC
<i>Training</i>						
SVM	0.754 ± 0.033	0.944 ± 0.041	0.883 ± 0.049	0.849 ± 0.045	0.875 ± 0.049	0.726 ± 0.041
SKN	0.843 ± 0.051	0.962 ± 0.059	0.925 ± 0.048	0.902 ± 0.047	0.919 ± 0.049	0.823 ± 0.050
<i>Validation</i>						
SVM	0.644 ± 0.030	0.867 ± 0.027	0.732 ± 0.024	0.756 ± 0.025	0.787 ± 0.026	0.528 ± 0.022
SKN	0.722 ± 0.044	0.826 ± 0.038	0.701 ± 0.036	0.774 ± 0.043	0.789 ± 0.052	0.545 ± 0.050
<i>Test</i>						
SVM	0.937	0.965	0.937	0.951	0.955	0.903
SKN	0.979	0.941	0.904	0.960	0.955	0.906

Note: The optimum parameters of the SVM are:  $C = 100$ ,  $\gamma = 6.4$

The FGCs descriptors such as nImidazoles and nPyridines have positive average values for HDACIs 1 and HDACIs 2, 3 and 8, respectively. The large average values of these descriptors suggest that an increase in the number of Imidazole rings in the structures of inhibitors leads to an enhancement in the tendency to interact with HDAC1, while increment in the number of pyridine rings would be beneficial to increase interactions with HDACs 2, 3 and 8 targets.

The larger average values of nHBonds in HDACIs 1, 3 and 8 indicate the positive impact of intermolecular hydrogen bonds for more interactions with HDACs 1, 3 and 8.

The topological descriptor MSD is calculated from the second-order distance distribution moments of the H-depleted molecular graph. This index decreases with increment molecular branching in an isomeric series [45]. MSD has the large average value for HDACIs 2, 3. The high average value of this descriptor demonstrates that reducing the molecular branches (MSD) is desirable for enhancing the interactions with HDACs 2, 3.

The larger average values of C-002 descriptor for HDACIs 1 and 3 suggest that an increase in the number of atom-centered-CH<sub>2</sub> fragments with two functional groups R-CH<sub>2</sub>-R' (C-002), where R may be the same or different, raises the tendency to interact with HDAC1 and 3.

$H_{\text{Nar}}$  and  $G_{\text{Nar}}$  are narumi harmonic and topological geometric indices, which are related to the molecular branching and calculated from the number of non-hydrogen atoms divided by the summation of reciprocal vertex degree [46]. The larger average values of these descriptors for HDACIs 2, 3 and 8 indicated that there is a direct relationship between increasing the values of these descriptors with increasing interactions with HDACs 2, 3 and 8.

The positive average value of H-050 in HDACIs 1 shows that incrementing the number of H attached to heteroatom is effective for increasing the tendency to interact with the HDAC1. C-011 descriptor has positive average values in HDACIs 1 and 8, and this suggests the positive effect of the

presence of C-011 substitutions in the molecular structure of the inhibitors for further interactions with HDAC 1, 8.

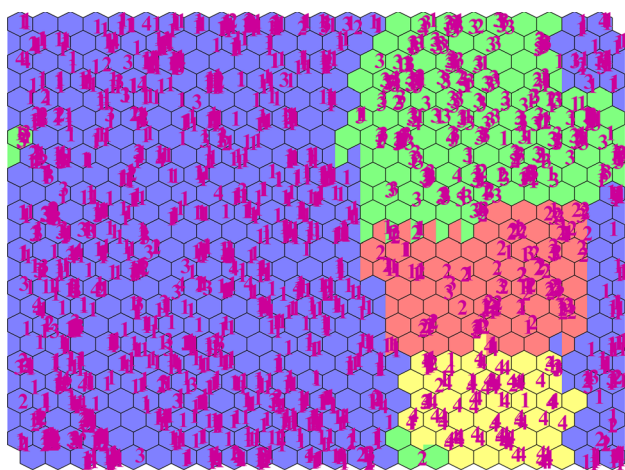
The last descriptor is Kier benzene-likeness index (BLI) that is a measure of aromaticity and obtained by dividing the first-order valence connectivity index by the number of non-H-bonds of the molecule [47]. According to the larger mean values of this descriptor in HDACIs 2, 3, 8, it can be concluded that an increase in molecular aromaticity is more appropriate for interactions with the HDAC2, 3 and 8 isoforms.

The isoform-selective structural characteristics presented in this model reveal the effective roles of some functional groups (nPyridines, nImidazoles), atom-centered fragments (C-002, C-011), intramolecular interactions (nHBonds, H-050), molecular branching (MSD,  $H_{\text{Nar}}$ ,  $G_{\text{Nar}}$ ) and aromaticity (BLI) in the structure of inhibitors for selective interaction between the ligand and each isoform. This information is very useful for describing the interaction of drugs with their biological targets and can be used to extract important pharmacophores and molecular features necessary for specific interaction to the HDACs targets and to design new drugs with higher selectivity activities.

The statistical results derived from the confusion matrix and MCC values for the developed model by the above descriptors are presented in Table 1. The values of prediction accuracy in the test set for SKN and SVM methods were obtained more than 70%. SKN map and its optimal parameters for the training set are shown in Fig. 1. It can be seen that the molecules are reasonably separated according to their therapeutic targets, and the molecules with similar target are collected in neighboring clusters of neurons. The results indicated that both methods have a good potential to classify isoform-selective HDACIs, although the SVM-based classifiers are more accurate than SKN.

### Classification of compounds based on activity

In this section, to discover structure–activity relationships between active and inactive of HDACIs, four classification



**Fig. 1** The SKN map with the size of  $24 \times 24$  and 60 epochs for the classification of the isoform-selective HDACs. Note: *Blue hexagon* refers to class 1 = Active HDACs 1. *Red hexagon* refers to class 2 = Active HDACs 2. *Green hexagon* refers to class 3 = Active HDACs 3. *Yellow hexagon* refers to class 4 = Active HDACs 8. (Color figure online)

models were created using binary classifiers and 10 VIP-selected descriptors. The names and brief definition of these descriptors along with their average values in each class and their corresponding  $p$  values were listed in Online Resource 2 Tables S3–S6. To examine the relative importance as well as the contribution of each descriptor in the model development, we used the unpaired Student's  $t$ -test. Statistical tests were two-sided and assumed that the null hypothesis was true. The difference between the average values of the two groups and also the small  $p$  values indicated the importance and discriminative power of the selected descriptors.

### Discrimination of active and inactive HDACs1

In this model, the retrieved dataset contained 1744 active and 1047 inactive HDAC1 molecules that constructed a data matrix with dimensions of  $(1744 + 1047) \times 450$ . By disregarding descriptors with constant or near constant values inside each column, 198 descriptors remained for analysis. Table S3 (in supplementary) lists the information of VIP-selected descriptors. Four selected structural features belong to the ACFs and FGCs classes and included C-027 (R–CH–X), nHBonds, nRNHO (number of aliphatic hydroxylamines) and nImidazoles, where all of these descriptors have positive average values for HDACs1. Consequently, the present of such a type of fragments in inhibitor structures leads to their improved activity of HDACs1.

Other important structural features of the HDACs1 are shape descriptors that belong to two geometrical and GETAWAY (GEometry, Topology and Atom Weights Assembly) classes and include folding degree index (FDI), 3D Petitjean

shape index (PJI3), length-to-breadth ratio by WHIM (L/Bw) and mean electrotopological state (Ms) descriptors.

The FDI is obtained from the largest eigenvalue of the distance/distance matrix, and then it is normalized by dividing the number of atoms. This index is sensitive to conformational changes and represents a quantitative measure of the similarity between chains with the same length but different geometries. This value tends to one for linear molecules (of infinite length) and decreases depending upon the folding of the molecule [48].

PJI3 (Petitjean shape indice) is a topological anisometry descriptor proposed to describe the degree of deviation from a perfect cyclic topology. The values of PJI3 vary in a range of 1 (for acyclic graphs) to 0 (for strictly cyclic graphs). PJI3 is obtained from the fraction of the difference of geometric diameter and radius to the geometric radius [49].

Length-to-breadth ratio by WHIM (L/Bw) is defined as the ratio of the longest to the shortest side of the rectangle that envelopes the molecular structure. The geometric mean on the leverage magnitude ( $H_{GM}$ ) is the simplest GETAWAY descriptor that is constructed from the diagonal elements of the molecular influence matrix (leverages) and encodes the influence of each atom in determining the overall shape of the molecule. Thus, the  $H_{GM}$  index is sensitive to the change of molecular shape and increases with branching and decreases with molecular size [50].

The mean electrotopological state molecular descriptor (Ms) is the average of the Kier-Hall electrotopological state for each non-hydrogen atom in the molecular structure [51]. Briefly, the electrotopological state describes the electronic and topological characteristics for each atom in a structure and also the electronic interactions due to all other atoms in the molecule with that atom. Each atom type is assigned an intrinsic state based on Kier-Hall electronegativity whose value is large for electronegative atoms and decreases with increasing the number of  $\sigma$  bonds [52].

The last descriptor is total absolute charge of atoms in a molecule ( $Q_{tot}$ ), a measure of molecular polarity, which is calculated from the sum of the absolute values of the atomic charges over all atoms in a molecule.

Here, with regard to the low mean values of the shape descriptors for HDACs1 in Table S3 (see Online Resource 2), it was demonstrated that an increase in the amounts of these descriptors negatively impacts on the inhibitory activities.

Generally, it can be concluded that molecular structures with fewer branches and larger size, less electronegative atoms and more sigma bond, high flexibility, acyclic structures and less length-to-breadth ratio are more suitable for increasing the activity of HDACs1.

The statistical results obtained of the SVM and SKN classifiers in the training, validation and test sets and optimal parameters of SVM are shown in Table 2. SKN map

**Fig. 2** **a** The SKN map with the size of  $30 \times 30$  and 100 epochs. **b**, **c** ROC plots of the SVM and SKN models for the classification of the active and inactive HDACIs 1. *Note: Open hexagon refers to class 1=Active HDACIs 1, Filled hexagon refers to class 2=Inactive HDACIs 1*

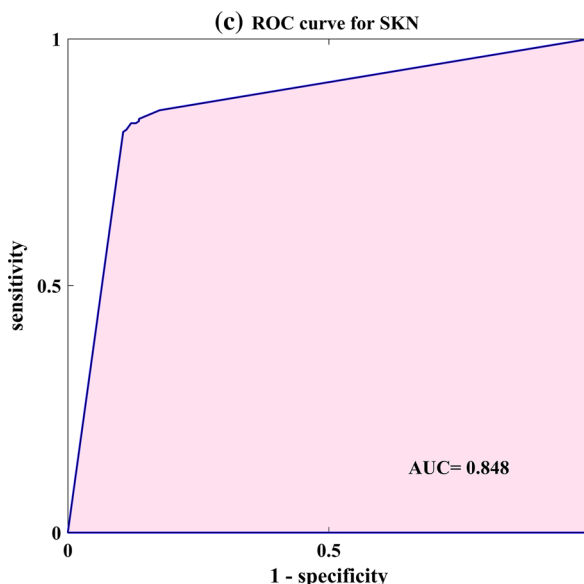
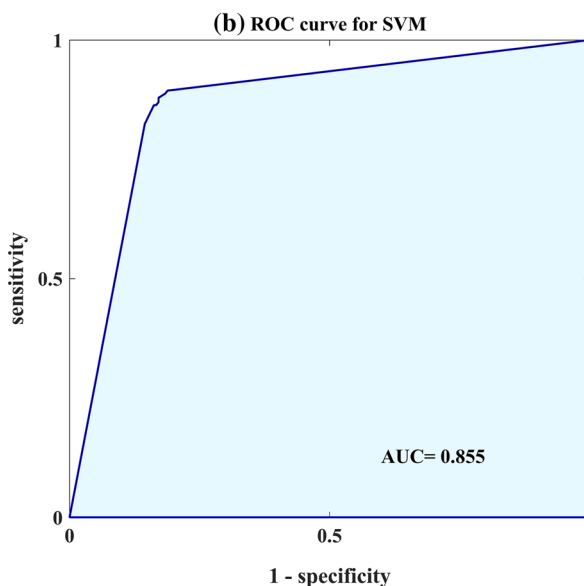
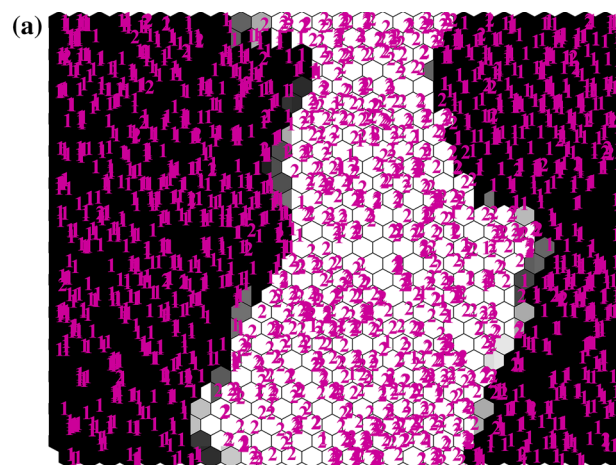
together with its optimal parameters is shown in Fig. 2a–c demonstrates the ROC curves and their calculated AUC values from the test set for both models. Clear discrimination between the active and inactive molecules in the SKN map and also the AUC values more than 85% represent the good performance of the built models. Statistical comparison of results illustrated that despite the acceptable ability of SKN, the SVM method provides more accurate results in modeling the structure–activity relationship of HDACIs 1.

### Discrimination of active and inactive HDACIs2

In this model, there were 301 and 274 molecules in the class of active and inactive HDAC2, respectively. Therefore, a data matrix (X) was created with the size of  $(301 + 274) \times 450$  from the entire dataset where 195 variables remained after removing constant and high correlation descriptors. Table S4 (supplementary) depicts the information of 10 selected descriptors, where the model was constructed with nRNHO, nArNHR (secondary aromatic amines), C-027 (R–CH–X), TE1 (topographic electronic descriptor), C-033 (R–CH–X), N-069 (aromatic and aliphatic amines Ar–NH<sub>2</sub>/X–NH<sub>2</sub>), nPyrimidines (Pyrimidines), RPCG (relative positive charge), H-050 and nN (nitrogen atoms) descriptors.

C-027 and C-032 represent atom-centered aromatic –CH fragments which are connected with one single bond to a functional group R; C-027 carbon atom shares a single bond with a heteroatom, while a C-032 carbon atom shares an aromatic single bond with a heteroatom. TE1 belongs to charge descriptors and reflects the electronic charge distribution and interatomic distances of the molecules. Its quantity increases with the increment of the charge differences between two atoms and reducing the interatomic distance in the molecule [48]. The RPCG descriptor was obtained from the partial charge of the most positive atom divided by the total positive charge. RPCG is controlled by different heteroatoms, i.e., nitrogen and oxygen, and is intensely influenced by the presence of donor–acceptor atoms for H-bonds.

The larger average values for all ten descriptors in HDACIs 2 reflect the fact that increasing the values of these descriptors will induce positive impacts on the inhibitory activities. This means that the presence of electronegative heteroatoms, such as oxygen and nitrogen (C-027, C-032, H-050, nN), as well as nitrogen substitutions connected to the carbon skeleton (nArNHR, nRNHO, nPyrimidines, N-069), increases relative positive charge (RPCG), and the charge differences between two atoms reducing the



**Fig. 3** **a** The SKN map with the size of  $12 \times 12$  and 150 epochs. **b**, **c** ROC plots of the SVM and SKN models for the classification of the active and inactive HDACIs 2. *Note: Open hexagon* Refers to class 1 = Active HDACIs 2, *Filled hexagon* refers to class 2 = Inactive HDACIs 2

interatomic distance (TE1) is effective factors that would induce positive effect on the activity of HDACIs 2.

The results from the confusion matrix for SKN and SVM methods are given in Table 3. By examining the results, it is indicated that the SVM method has more capability to discriminate the active inhibitors of HDAC2 from inactive ones. The trained SKN map together with the label of the molecules in each neuron and optimal numbers of neurons and epoch and also ROC curves and their AUC values are demonstrated in Fig. 3.

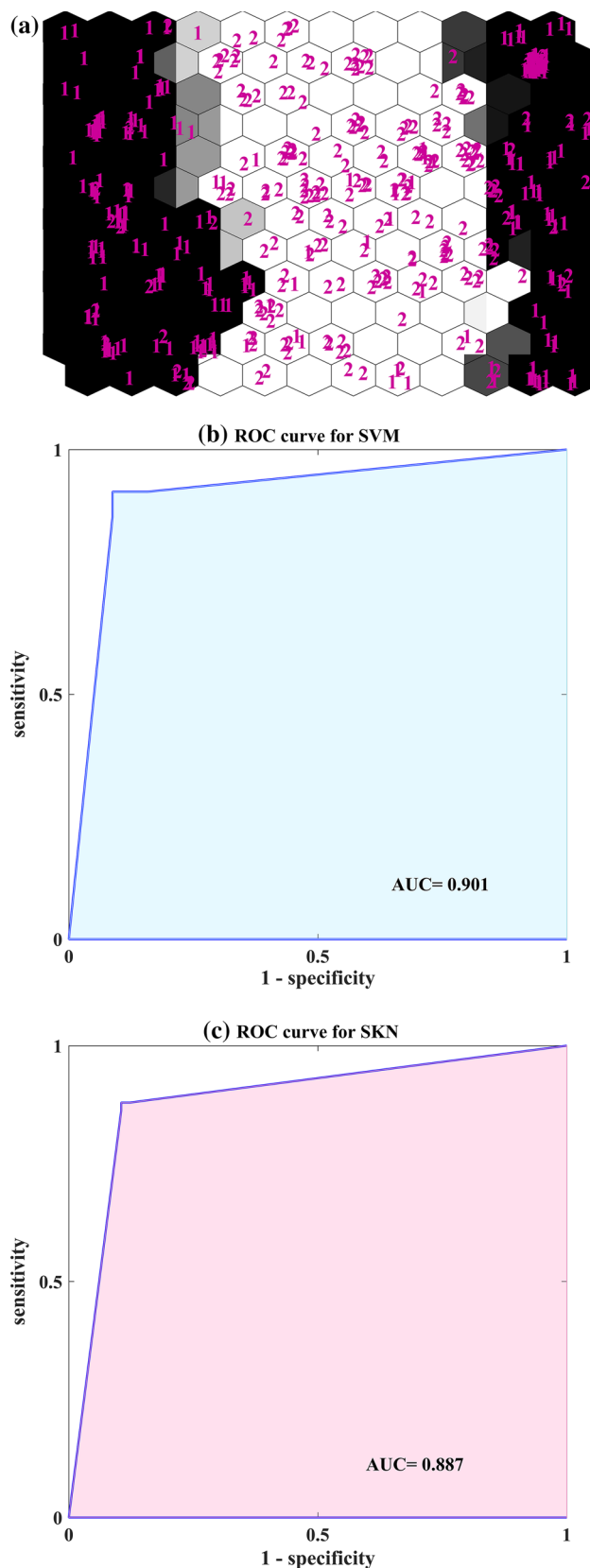
### Discrimination of active and inactive HDACIs 3

This model contains 473 active and 330 inactive HDACIs3, where a data matrix is created with the size of  $(473 + 330) \times 450$ . After removing the inappropriate descriptors, a subset of 200 descriptors remained. The important structural features of HDACIs3 are shown in Table S5 (in supplementary) that include: H-050, N-072 (RCO-N  $\neq$  N-X = X), nRNHO, nROH (number of hydroxyl groups),  $Q_{\text{neg}}$  (total negative charge), C-002 ( $\text{CH}_2\text{R}_2$ ), nPyrroles (number of Pyrroles), RBF (rotatable bond fraction) and nArCONHR (number of secondary aromatic amides).

Descriptors of class ACFs and FGCs such as H-050, N-072, nRNHO, nROH, C-002, nPyrroles and nArCONHR have larger average values in HDACIs 3. Therefore, it can be concluded that the presence of H attached to heteroatom, RCO-N  $\neq$  N-X = X type fragment, aliphatic hydroxyl amines, hydroxyl groups,  $\text{CH}_2\text{R}_2$  fragment, Pyrroles and secondary aromatic amides in the molecular structure of inhibitors are desirable to enhance inhibitory activities.

Another significant feature in this model is  $Q_{\text{neg}}$  or total negative charge, which is obtained from the sum of all of the negative charges of the atoms in a molecule. The lower average value of this descriptor indicates a negative influence on the inhibitory activity. Therefore, its lower values in HDACIs 3 would be beneficial in improving the inhibitory activity.

The constitutional descriptor RBF corresponds to the rotatable bond fraction in the molecule. The large average value of this variable in HDACIs3 shows that the increase in rotatable bonds fraction in the molecule is conducive to increasing the activity of inhibitors. The last descriptor is Lop, and a topological centric index which is defined as the mean information content is derived from the pruning partition of an acyclic graph [53]. Lop quantifies the degree of compactness of molecules distinguishing between molecular



**Fig. 4** **a** The SKN map with the size of  $16 \times 16$  and 50 epochs. **b, c** ROC plots of the SVM and SKN models for the classification of the active and inactive HDACIs 3. *Note: Open hexagon* Refers to class 1=Active HDACIs 3, *Filled hexagon* refers to class 2=Inactive HDACIs 3

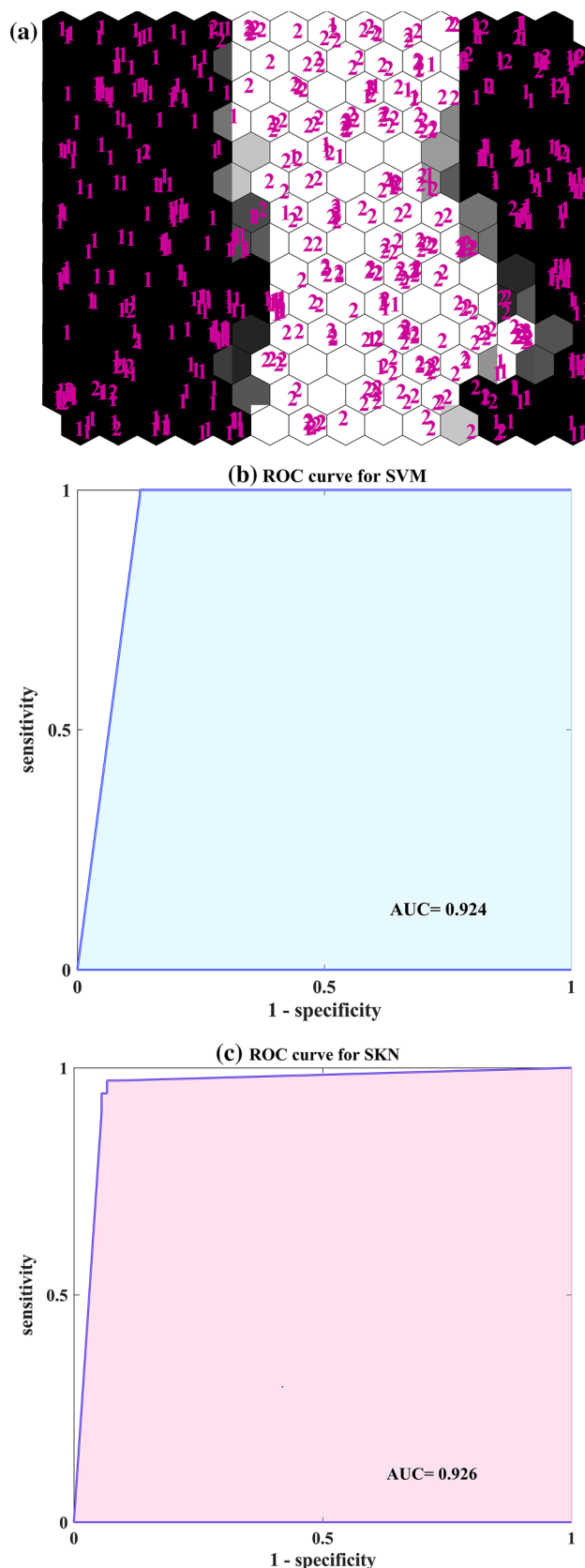
structures organized differently with regard to their centers [54]. The positive average value of this variable suggests that increasing the amount of Lop index could be an effective factor in improving inhibitors' activity.

According to the above description, the VIP-selected descriptors in this model reveal the importance of the presence of these fragments (H attached to heteroatom, RCO-N </>N-X=X type fragment, aliphatic hydroxyl amines, hydroxyl groups,  $\text{CH}_2\text{R}_2$  fragment, Pyrroles and secondary aromatic amides) in increasing the activity of HDACIs 3. Also, an increment in rotatable bonds fraction and compactness and decrease in the total negative charge in the structure of molecule have a favorable effect in the activity of HDACIs 3.

The statistical parameters of this binary classifier together with the MCC values and optimal parameters of SVM are given in Table 4. The ROC curves and SKN map, along with the AUC values and optimal SKN parameters, are shown in Fig. 4. The results indicated that both methods have acceptable and almost equal potency in discrimination active–inactive HDACIs 3 and the SVM method yields better results in the validation sets.

### Discrimination of active and inactive HDACIs 8

In this model, a binary classifier was constructed using 239 active and 424 inactive HDACIs 8. The size of the generated data matrix was  $(239 + 424) \times 450$ . After removing the constant and high correlation variables, 191 descriptors remained for development of model. Table S6 (supplementary) comprises the information of 10 VIP-selected descriptors in this model. Six descriptors belong to the ACFs and FGCs classes: nRNHO, nImidazole, C-011 and C-041 descriptors have larger average values in HDACIs 8, which means that increasing the number of aliphatic hydroxylamines (nRNHO) and Imidazoles (nImidazole) and C-atoms fragments connected as X-C(=X)-X and  $\text{CR}_3\text{X}$  (C-041, C-011) in the molecular structures of inhibitors, where X represents any electronegative atom (O, N, S, P, halogens) and = represents a double bond, would be favorable for increasing the inhibitory activities of HDACIs 8. But the small average values of N-066 and nRCONR<sub>2</sub> descriptors in inhibitors illustrate that an increase in the number of secondary aliphatic amines (N-066) and tertiary aliphatic amides (nRCONR<sub>2</sub>) lead to a decrease in the activity of inhibitors.



**Fig. 5** **a** The SKN map with the size of  $14 \times 14$  and 50 epochs. **b, c** ROC plots of the SVM and SKN models for the classification of the active and inactive HDACIs 8. Note: *Open hexagon* refers to class 1=Active HDACIs 8, *Filled hexagon* refers to class 2=Inactive HDACIs 8

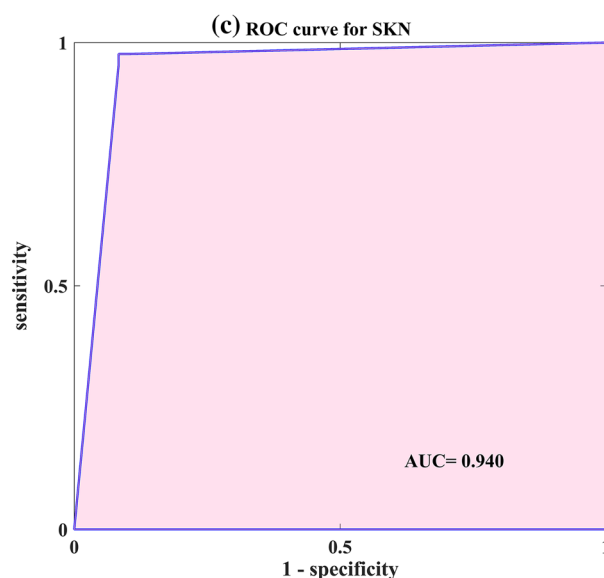
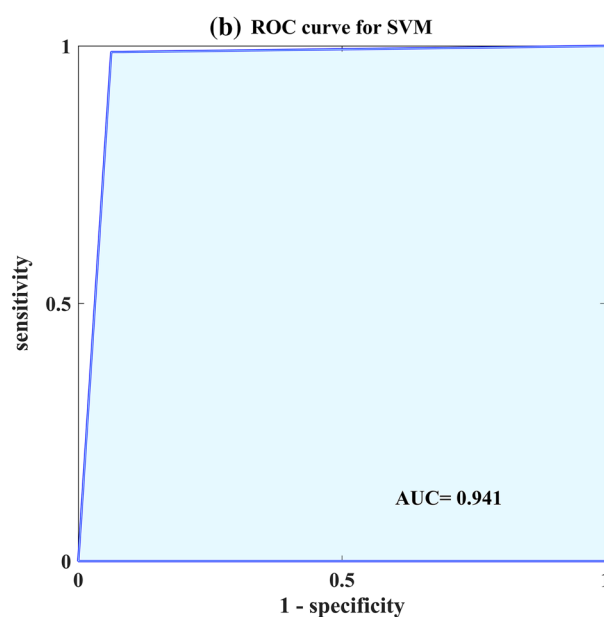
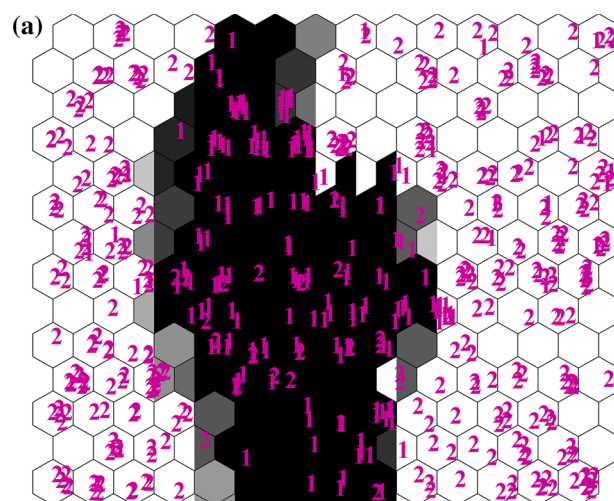
The next descriptor in this model is the MSD (mean square distance). As explained in section “Classification of compounds based on therapeutic targets,” the value of this descriptor decreases with the increasing molecular branching. The small mean value of this variable in HDACIs 8 suggests that an increment in the number of branches in the molecule is suitable for increasing the inhibitory activity.

LP1 (Lovasz–Pelikan index or leading eigenvalue) is a topological descriptor. It has been known as an index of molecular branching, and the highest values correspond to the most branched graphs and the smallest values to chain graphs [53]. The higher mean value of this descriptor in inhibitors illustrated the positive effect of increasing the number of branches in the activity of HDACIs 8.

The next descriptor is molecular asphericity (ASP), which describes the deviation from a spherical shape in terms of all three main principal moments of inertia of the molecule [53]. For the linear molecules, the value of the ASP is higher than that of the spherical molecule. Therefore, due to the small average value of this variable in HDACIs 8, it can be concluded that spherical molecules have more inhibitory activity. L/Bw is the last selected descriptor in this model that has a negative effect on the inhibitors’ activity and reducing its value is predicted to increase the activity of HDACIs 8.

In sum, as is apparent from Table S6, VIP-selected descriptors in this model reveal the importance of the presence of aliphatic hydroxyl amines, imidazoles,  $CR_3X$  and  $X-C(=X)-X$  fragments and increment the number of branches, spherical and size in the molecular structure of inhibitors to increase the inhibitory activity of the compounds.

The statistical results of SVM and SKN methods and optimal parameters of SVM are listed in Table 5. The SKN map and ROC curves along with its optimal SKN parameters and AUC values are presented in Fig. 5. The tabulated results show that both methods (SKN and SVM) have an equal ability to classify active/inactive HDACIs 8. However, comparing the MCC and non-error rate values indicates that the SKN method yields better results. The prediction accuracy greater than 95% in the test sets reveals the stability and predictive ability of the generated classifiers to model the structure–activity relationship of HDACIs 8.



## Descriptors analysis

At first, it is worth noting that despite huge structural diversity, HDACIs are generally characterized by a common pharmacophore model: a zinc binding group (ZBG) such as hydroxamic acid and benzamide interacts with  $Zn^{2+}$  ion at the bottom of active site of HDACs, a hydrophobic linker domain occupies the narrow tunnel of HDACs and connects the ZBG, and a surface recognition group (cap). Cap, a hydrophobic and aromatic or heteroaromatic group, is essential for recognizing and interacting with amino acid residues on the rim of active site of HDACs [55].

Structural modification in the pharmacophores of HDACIs is still a challenge, and further research is needed to design potent and selective inhibitors. It is believed that computational methods can provide valuable information in the least time to discover new drugs with developed potential.

Taking into account the above description, a brief analysis on the selected descriptors was performed to better understand the structure–activity and isoform selectivity relationships of HDACIs.

In this study, different types of descriptors are involved to build the classification models as ACFs, FGCs, topological, geometrical, charge and constitutional descriptors. Among these descriptors, the FGCs and AFCs make the major contributions to build the classification models. These descriptors are a simple representation of molecular structure that arises from the chemical identity of each atom to describe the role of particular functional group and atom connectivity in the molecular skeleton on their activities.

Some of the FGCs and AFCs descriptors have been replicated in the most models, which illustrate the importance of these features in increasing the inhibitory potency of HDACIs. Like nRNHO that exists on all four binary classification models shows a positive effect on the inhibitory activity of HDACIs 1, 2, 3 and 8. Also, the existence of H-050 (in HDACIs 2, 3) and nHBonds (in HDACIs 1) descriptors highlights the importance of hydrogen bonds in the interactions between inhibitor and target. Moreover, the functional groups like the Pyrimidine (in HDACIs 2), Imidazole (in HDACIs 3, 8), Pyrrole (in HDACIs 3) and also amino substitutions such as N-072, nArNHR (in HDACIs 3), N-069 (in HDACIs 2), N-066 and nRCONR2 (in HDACIs 8) with positive average values in inhibitors make a positive contribution to increasing the inhibitory activity. These substructures are known as important and effective substitutions in the structure of synthesized HDACs inhibitors and can interact with the targets as an electron-donating groups by chelate formation or through intermolecular hydrogen bonding with the zinc ion on the active site of HDACs [56]. In addition, these features have been introduced in other HDACIs classification study [57].

Other important factors related to the activity of HDACIs are branching, flexibility, cyclicality and the size of the inhibitors. MSD, LP1 (in HDACIs 8) and  $H_{GM}$  (in HDACIs 1) variables show the importance of molecular branching in increasing the activity of inhibitors. Additionally, the results suggest that an increase in the molecular flexibility by incorporation of rotatable bonds RBF (in HDACIs 3) and FDI (in HDACIs 1) is conducive to inhibitory activity. The importance of molecule sphericity was determined by the ASP (in HDACIs 8) and PJI3 (in HDACIs 1) descriptors. The presence of L/Bw index (in HDACIs 1, 8) confirms that the size of inhibitors plays a major role in the inhibitory activity. These molecular shape descriptors have a determining property in the different physico-chemical processes, such as interaction capability between ligand and target. These properties indicate the important structural features of pharmacophoric groups to locate in an appropriate position of the HDACs active site for better drug–enzymes interaction.

In addition, the polarity of molecules has a great effect on the inhibitors' activity. The presence of charge descriptors such as  $Q_{tot}$  (in HDACIs 1), RPCG (in HDACIs 2) and  $Q_{neg}$  (in HDACIs 3) shows the importance of polarity for HDACIs inhibitory potency, because different heteroatoms and donor–acceptor atoms in H-bond influence molecular polarity. Therefore, the presence of C-027 (in HDACIs 1, 2), nN, C-033 (in HDACIs 2) and C-011, C-041 (in HDACIs 8) reveals the impact of polarity and electronegative atoms in drug–enzyme interaction.

As seen in Table S2, the selected descriptors in isoform-selective model indicate that the presence of imidazole (in HDACIs 1) and pyridine rings (in HDACIs 2, 3, 8) in the structure of inhibitors plays a positive role in an increasing tendency to interact with these targets. Also, increasing the number of intramolecular H-bonds contributes to selective interactions with the HDACs 1, 3 and 8. These features are also available on some active/inactive binary models. Such as nImidazole and nHbonds with positive average values existing in the active HDACIs 1, this suggests that these properties can simultaneously enhance the inhibitory activity and interaction with HDAC 1. As well as, the nPyridine descriptor in active HDACIs 2 is useful for increasing inhibitory activity and the desire to interact with the HDAC2. These properties can be considered in the synthesis of active and selective inhibitors by pharmaceutical chemists.

The increase in the number of branches and aromaticity in the structure of inhibitors has a positive influence on the selective interaction with the HDACs 2, 3 and 8.

The desire to selective interactions with HDAC 1 will be enhanced by increasing the number of hydrogens attached to heteroatom. Also, the presence of  $CR_3X$  atom-centered fragments in the structure of HDACIs 1, 8, and  $CH_2R_2$  fragments

in HDACs 1, 3 increases the tendency to interact with these targets.

This analysis makes it possible to identify the contribution of the different physicochemical properties of the molecules to pharmacological activity. All of the VIP-selected descriptors in this study are closely associated with the structural features of synthesized inhibitors in existing research [58–60], which confirms the good performance of the proposed models in this study.

## Statistical results

By examining the tabulated results indicated that SKN and SVM methods have an acceptable ability to model structure–activity and structure–isoform selectivity relationships of HDACs. SVM method provides the better results than SKN in the discrimination of active–inactive HDACs 1 and 2, but both methods have almost equal ability to model structure–activity relationships of HDACs 3 and 8. In classifiers of isoform-selective HDACs, the SVM model provides more accurate statistical results than the SKN. The AUC values more than 0.85 for SKN and SVM models indicated the good predictive power of the developed models.

We computed the AD for the developed classifiers. The cutoff leverage values together with the number of “in domain” test set molecules are given in Online Resource 2 (Table S7). As can be seen in this table, most of the test set molecules fall within the AD of the models; therefore, our models were useful for evaluation of HDACs in which no experimental data are available.

## Conclusion

This study represents the continuation of our efforts for the development and application of classification models toward the discovery of important and distinctive structural features for some of considered compounds [32, 61]. In this work, isoforms of class I HDACs (HDACs 1, 2, 3 and 8) were targeted considering their over expression in many diseases. We followed two goals in this study: First to determine the best variables which produced the most distinction between the isoform-selective and active/inactive HDACs. For this propose, we used the VIP technique as a variable selection method and two powerful machine learning methods (SVM, SKN) to classify a diverse set of HDACs. We believe the combination of the best subsets of variables and a diverse dataset next to the convenient cheminformatics models can be used as a reliable lead designing tool for exploring new potent and selective inhibitors in a drug discovery pipeline. Our second goal is to evaluate the effect of various structural properties on drug–target interaction that was performed with the aid of descriptor analysis. In sum, the interpretation

of selected descriptors underscores the fact that the presence of functional groups like the pyrimidine, imidazole, pyrrole and also amino substitutions such as aliphatic hydroxylamines, aromatic secondary amides and molecular polarity, intramolecular H-bonds, hydrogens attached to heteroatom, number of branches, flexibility, cyclicality and size in the structure of inhibitors are important responsible parameters for the selectivity and activity of HDACs. The results illustrate that the SVM and SKN classifiers achieved high accuracy in discriminating HDACs in the terms of isoform selectivity and activity, although SVM represents an almost better classification accuracy rates compared to SKN. The high accuracy values of the obtained classifiers for the training, validation and test sets demonstrate that the information provided in this work is reliable for designing HDACs with the better therapeutic potency and reduced side effect. The obtained results can significantly contribute to the molecular understanding of enzyme–drug interactions.

**Acknowledgements** We gratefully acknowledge the support of this work by Yazd University research council.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest

## References

1. Thangapandian S, John S, Lee KW (2012) Molecular dynamics simulation study explaining inhibitor selectivity in different class of histone deacetylases. *J Biomol Struct Dyn* 29(4):677–698. <https://doi.org/10.1080/07391102.2012.10507409>
2. Bolden JE, Shi W, Jankowski K, Kan CY, Cluse L, Martin BP, MacKenzie KL, Smyth GK, Johnstone RW (2013) HDAC inhibitors induce tumor-cell-selective pro-apoptotic transcriptional responses. *Cell Death Dis* 4(2):519–534. <https://doi.org/10.1038/cddis.2013.9>
3. Gao S, Zang J, Gao Q, Liang X, Ding Q, Li X, Xu W, Chou CJ, Zhang Y (2017) Design, synthesis and anti-tumor activity study of novel histone deacetylase inhibitors containing isatin-based caps and o-phenylenediamine-based zinc binding groups. *Bioorg Med Chem* 25(12):2981–2994. <https://doi.org/10.1016/j.bmc.2017.03.036>
4. Zhang J, Zhong Q (2014) Histone deacetylase inhibitors and cell death. *Cell Mol Life Sci* 71:3885–3901. <https://doi.org/10.1007/s00018-014-1656-6>
5. Falkenberg KJ, Johnstone RW (2014) Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders. *Nat Rev Drug Discov* 13:673–691. <https://doi.org/10.1038/nrd4360>
6. Haberland M, Montgomery RL, Olson EN (2009) The many roles of histone deacetylases in development and physiology: implications for disease and therapy. *Nat Rev Genet* 10:32–42. <https://doi.org/10.1038/nrg2485>



7. Yoon S, Eom GH (2016) HDAC and HDAC inhibitor: from cancer to cardiovascular diseases. *Chonnam Med J* 52(1):1–11. <https://doi.org/10.4068/cmj.2016.52.1.1>
8. Li Y, Seto E (2016) HDACs and HDAC inhibitors in cancer development and therapy. *Cold Spring Harbor Perspect Med* 6(10):26831–26872. <https://doi.org/10.1101/cshperspect.a026831>
9. Barneda-Zahonero B, Parra M (2012) Histone deacetylases and cancer. *Mol Oncol* 6(6):579–589. <https://doi.org/10.1016/j.molonc.2012.07.003>
10. Mottamal M, Zheng S, Huang TL, Wang G (2015) Histone deacetylase inhibitors in clinical studies as templates for new anticancer agents. *Molecules* 20(3):3898–3941. <https://doi.org/10.3390/molecules20033898>
11. Mai A, Massa S, Rotili D, Cerbara I, Valente S, Pezzi R, Simeoni S, Ragno R (2005) Histone deacetylation in epigenetics: an attractive target for anticancer therapy. *Med Res Rev* 25(3):261–309. <https://doi.org/10.1002/med.20024>
12. Goey AK, Sissung TM, Peer CJ, Figg WD (2016) Pharmacogenomics and histone deacetylase inhibitors. *Pharmacogenom J* 16:1807–1815. <https://doi.org/10.2217/pgs-2016-0113>
13. Tang H, Wang XS, Huang XP, Roth BL, Butler KV, Kozikowski AP, Jung M, Tropsha A (2009) Novel inhibitors of human histone deacetylase (HDAC) identified by QSAR modeling of known inhibitors, virtual screening, and experimental validation. *J Chem Inf Model* 49:461–476. <https://doi.org/10.1021/ci800366f>
14. Cao GP, Thangapandian S, Son M, Kumar R, Choi YJ, Kim Y, Kwon YJ, Kim HH, Suh JK, Lee KW (2016) QSAR modeling to design selective histone deacetylase 8 (HDAC8) inhibitors. *Arch Pharm Res* 39(10):1356–1369. <https://doi.org/10.1007/s12272-015-0705-5>
15. Pontiki E, Hadjipavlou-Litina D (2012) Histone deacetylase inhibitors (HDACIs). Structure-activity relationships: History and new QSAR perspectives. *Med Res Rev* 32:1–165. <https://doi.org/10.1002/med.20200>
16. Norinder U, Naveja JJ, López-López E, Mucs D, Medina-Franco JL (2019) Conformal prediction of HDAC inhibitors. *SAR QSAR Environ Res* 30(4):265–277
17. Nair SB, Teli MK, Pradeep H, Rajanikant GK (2012) Computational identification of novel histone deacetylase inhibitors by docking based QSAR. *Comput Biol Med* 42(6):697–705. <https://doi.org/10.1016/j.compbiomed.2012.04.001>
18. Katritzky AR, Slavov SH, Dobchev DA, Karelson M (2007) Comparison between 2D and 3D-QSAR approaches to correlate inhibitory activity for a series of indole amide hydroxamic acids. *QSAR Comb Sci* 26:333–345. <https://doi.org/10.1002/qsar.200630021>
19. Guo Y, Xiao J, Guo Z, Chu F, Cheng Y, Wu S (2005) Exploration of a binding mode of indole amide analogues as potent histone deacetylase inhibitors and 3D-QSAR analyses. *Bioorg Med Chem* 13(18):5424–5434. <https://doi.org/10.1016/j.bmc.2005.05.016>
20. Xiang Y, Hou Z, Zhang Z (2012) Pharmacophore and QSAR studies to design novel histone deacetylase 2 inhibitors. *Chem Biol Drug Des* 79:760–770. <https://doi.org/10.1111/j.1747-0285.2012.01341.x>
21. Abdel-Atty MM, Farag NA, Kassab SE, Serya RA, Abouzid KA (2014) Design, synthesis, 3D pharmacophore, QSAR, and docking studies of carboxylic acid derivatives as Histone Deacetylase inhibitors and cytotoxic agents. *Bioorg Chem* 57:65–82. <https://doi.org/10.1016/j.bioorg.2014.08.006>
22. Noor Z, Afzal N, Rashid S (2015) Exploration of novel inhibitors for class I histone deacetylase isoforms by QSAR modeling and molecular dynamics simulation assays. *PLoS ONE* 10(10):e0139588. <https://doi.org/10.1371/journal.pone.0139588>
23. Uba AI, Yelekçi K (2017) Identification of potential isoform-selective histone deacetylase inhibitors for cancer therapy: a combined approach of structure-based virtual screening, ADMET prediction and molecular dynamics simulation assay. *J Biomol Struct Dyn* 21:1–5. <https://doi.org/10.1080/07391102.2017.1384402>
24. Dessalew N (2007) QSAR study on amino phenyl benzamides and acrylamides as histone deacetylase inhibitors: an insight into the structural basis of ant proliferative activity. *Med Chem Res* 16(7–9):449–460. <https://doi.org/10.1007/s00044-007-9085-9>
25. Yang JS, Chun TG, Nam KY, Kim HM, Han G (2012) Structure-activity relationship of novel lactam based histone deacetylase inhibitors as potential anticancer drugs. *Bull Korean Chem Soc* 33:2063–2066. <https://doi.org/10.5012/bkcs.2012.33.6.2063>
26. Zhao L, Xiang Y, Song J, Zhang ZA (2013) Novel two-step QSAR modeling work flow to predict selectivity and activity of HDAC inhibitors. *Bioorg Med Chem Lett* 23(4):929–933. <https://doi.org/10.1016/j.bmcl.2012.12.067>
27. Cao GP, Arooj M, Thangapandian S, Park C, Arulalapperumal V, Kim Y, Kwon YJ, Kim HH, Suh JK, Lee KW (2015) A lazy learning-based QSAR classification study for screening potential histone deacetylase 8 (HDAC8) inhibitors. *SAR QSAR Environ Res* 26:397–420. <https://doi.org/10.1080/1062936X.2015.1040453>
28. Liu XH, Song HY, Zhang JX, Han BC, Wei XN, Ma XH, Cui WK, Chen YZ (2010) Identifying novel type ZBGs and non-hydroxamate HDAC inhibitors through a SVM based virtual screening approach. *Mol Inf* 29:407–420. <https://doi.org/10.1002/minf.200900014>
29. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J (2016) Binding DB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 44:D1045–D1053. <https://doi.org/10.1093/nar/gkv1072>
30. Todeschini R, Consonni V, Mauri A, Pavan M (2007) DRAGONs software for the calculation of molecular descriptors, version 5.5 for Windows. Milan, Italy. [http://www.taletе.mi.it/products/dragon\\_description.htm](http://www.taletе.mi.it/products/dragon_description.htm)
31. O’Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. *J Chem Inf* 3:1–14. <https://doi.org/10.1186/1758-2946-3-33>
32. Mani-Varnosfaderani A, Neiband MS, Benvidi A (2018) Identification of molecular features necessary for selective inhibition of B cell lymphoma proteins using machine learning techniques. *Mol Divers* 12:1–9. <https://doi.org/10.1007/s11030-018-9856-x>
33. Farrés M, Platikanov S, Tsakovski S, Tauler R (2015) Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *J Chemom* 29(10):528–536. <https://doi.org/10.1002/cem.2736>
34. Reis A, Rudnitskaya A, Chariyavilaskul P, Dhaun N, Melville V, Goddard J, Webb DJ, Pitt AR, Spickett CM (2015) Top-down lipidomics of low density lipoprotein reveal altered lipid profiles in advanced chronic kidney disease. *J Lipid Res* 56:413–422. <https://doi.org/10.1194/jlr.M055624>
35. Wang J, Su M, Li T, Gao A, Yang W, Sheng L, Zang Y, Li J, Liu H (2017) Design, synthesis and biological evaluation of thienopyrimidine hydroxamic acid based derivatives as structurally novel histone deacetylase (HDAC) inhibitors. *Eur J Med Chem* 128:293–309. <https://doi.org/10.1016/j.ejmech.2017.01.035>
36. Hu E, Dul E, Sung CM, Chen Z, Kirkpatrick R, Zhang GF, Johanson K, Liu R, Lago A, Hofmann G, Macarron R (2003) Identification of novel isoform-selective inhibitors within class I histone deacetylases. *J Pharmacol Exp Ther* 307(2):720–728. <https://doi.org/10.1124/jpet.103.055541>
37. Melssen W, Wehrens R, Buydens L (2006) Supervised Kohonen networks for classification problems. *Chemom Intell Lab Syst* 83(2):99–113. <https://doi.org/10.1016/j.chemolab.2006.02.003>
38. Vasighi M, Kompany-Zareh M (2013) Classification ability of self organizing maps in comparison with other classification methods. *MATCH Commun Math Comput Chem* 70:29–44

39. Omara H, Lazaar M, Tabii Y (2018) Self-organizing maps and principal component analysis to improve classification accuracy. *Res J Appl Sci Eng Technol* 15(5):190–196. <https://doi.org/10.19026/rjaset.15.5851>
40. Ballabio D, Vasighi M (2012) A MATLAB Toolbox for Self Organizing Maps and supervised neural network learning strategies. *Chemom Intell Lab* 118:24–32
41. Vapnik VN (1998) *Statistical learning theory*, 1st edn. Wiley-Interscience, New York. ISBN 978-0-471-03003-4
42. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2:1–27. <https://doi.org/10.1145/1961189.1961199>
43. Gramatica P (2007) Principles of QSAR models validation: Internal and external. *QSAR Comb Sci* 26:694–701. <https://doi.org/10.1002/qsar.200610151>
44. Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ Health Perspect* 111:1361–1375. <https://doi.org/10.1289/ehp.5758>
45. Balaban AT (1983) Topological indices based on topological distances in molecular graphs. *Pure Appl Chem* 5(2):199–206. <https://doi.org/10.1351/pac198855020199>
46. Todeschini R, Consonni V, Mannhold R, Kubinyi H, Folkers G (2009) *Molecular descriptors for chemoinformatics*. Wiley, Weinheim. ISBN 978-3-527-31852-0
47. Kier LB, Hall LH (1986) Molecular connectivity in structure-activity analysis. *Res Stud* 33:2096. <https://doi.org/10.1002/aic.690331230>
48. Randic M, Kleiner AF, De Alba LM (1994) Distance/distance matrixes. *J Chem Inf Comput Sci* 34(2):277–286
49. Bath PA, Poirrette AR, Willett P, Allen FH (1995) The extent of the relationship between the graph-theoretical and the geometrical shape coefficients of chemical compounds. *J Chem Inf Comput Sci* 35:714–716. <https://doi.org/10.1021/ci00026a007>
50. Puzyn T, Leszczynski J, Cronin MT (2010) Recent advances in QSAR studies: methods and applications. *Chall Adv Comput Chem Phys* 8:1–415. <https://doi.org/10.1007/978-1-4020-9783-6>
51. Hall LH, Kier LB, Brown BB (1995) Molecular similarity based on novel atom-type electrotopological state indices. *J Chem Inf Comput Sci* 35(6):1074–1080
52. Copeland JC, Zehr LJ, Cerny RL, Powers R (2012) The applicability of molecular descriptors for designing an electrospray ionization mass spectrometry compatible library for drug discovery. *Comb Chem high T Scr* 15(10):806–815. <https://doi.org/10.2174/138620712803901180>
53. Todeschini R, Consoni V (2008) *Handbook of molecular descriptors. Methods and principles in medicinal chemistry*. Wiley, New York. <https://doi.org/10.1002/9783527613106>
54. Fatemi MH, Chahi ZG (2012) QSPR-based estimation of the half-lives for polychlorinated biphenyl congeners. *SAR QSAR Environ Res* 23(1–2):155–168. <https://doi.org/10.1080/1062936X.2011.645876>
55. Suzuki T, Kasuya Y, Itoh Y, Ota Y, Zhan P, Asamitsu K, Nakagawa H, Okamoto T, Miyata N (2013) Identification of highly selective and potent histone deacetylase 3 inhibitors using click chemistry-based combinatorial fragment assembly. *PLoS ONE* 8(7):68669–68681. <https://doi.org/10.1371/journal.pone.0068669>
56. Zhang L, Zhang J, Jiang Q, Zhang L, Song W (2018) Zinc binding groups for histone deacetylase inhibitors. *J Enzym Inhib Med Chem* 33(1):714–721. <https://doi.org/10.1080/14756366.2017.1417274>
57. Jalali-Heravi M, Mani-Varnosfaderani A (2012) Navigating drug-like chemical space of anticancer molecules using genetic algorithms and counter propagation artificial neural networks. *Mol Inf* 31(1):63–74. <https://doi.org/10.1002/minf.201100098>
58. Bertrand P (2010) Inside HDAC with HDAC inhibitors. *Eur J Med Chem* 45(6):2095–2116. <https://doi.org/10.1016/j.ejmech.2010.02.030>
59. Rajak H, Singh A, Raghuvanshi K, Kumar R, Dewangan PK, Veerasamy R, Sharma PC, Dixit A, Mishra P (2014) A structural insight into hydroxamic acid based histone deacetylase inhibitors for the presence of anticancer activity. *Curr Med Chem* 21(23):2642–2664. <https://doi.org/10.2174/09298673113209990191>
60. Bolden JE, Peart MJ, Johnstone RW (2006) Anticancer activities of histone deacetylase inhibitors. *Nat Rev Drug Discov* 5(9):769–784. <https://doi.org/10.1038/nrd2133>
61. Neiband MS, Mani-Varnosfaderani A, Benvidi A (2017) Classification of sphingosine kinase inhibitors using counter propagation artificial neural networks: a systematic route for designing selective SphK inhibitors. *SAR QSAR Environ Res* 28(2):91–109. <https://doi.org/10.1080/1062936X.2017.1280535>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.