Available online at www.sciencedirect.com
ScienceDirect

Journal of Molecular Graphics and Modelling 26 (2007) 505–518

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Prediction of factor Xa inhibitors by machine learning methods

H.H. Lin [a,b], L.Y. Han [a,b], C.W. Yap [a,b], Y. Xue [c], X.H. Liu [a,b], F. Zhu [a,b], Y.Z. Chen [a,b,*]

[a] *Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore*
[b] *Bioinformatics and Drug Design Group, Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore*
[c] *College of Chemistry, Sichuan University, Chengdu 610064, PR China*

Received 22 August 2006; received in revised form 4 February 2007; accepted 7 March 2007
Available online 12 March 2007

Abstract

Factor Xa (FXa) inhibitors have been explored as anticoagulants for treatment and prevention of thrombotic diseases. Molecular docking, pharmacophore, quantitative structure–activity relationships, and support vector machines (SVM) have been used for computer prediction of FXa inhibitors. These methods achieve promising prediction accuracies of 69–80% for FXa inhibitors and 85–99% for non-inhibitors. Prediction performance, particularly for inhibitors, may be further improved by exploring methods applicable to more diverse range of compounds and by using more appropriate set of molecular descriptors. We tested the capability of several machine learning methods (C4.5 decision tree, *k*-nearest neighbor, probabilistic neural network, and support vector machine) by using a much more diverse set of 1098 compounds (360 inhibitors and 738 non-inhibitors) than those in other studies. A feature selection method was used for selecting molecular descriptors appropriate for distinguishing FXa inhibitors and non-inhibitors. The prediction accuracies of these methods are 89.1–97.5% for FXa inhibitors and 92.3–98.1% for non-inhibitors. In particular, compared to other studies, support vector machine gives a substantially improved accuracy of 94.6% for FXa non-inhibitors and maintains a comparable accuracy of 98.1% for inhibitors, based-on a more rigorous test with more diverse range of compounds. Our study suggests that machine learning methods such as SVM are useful for facilitating the prediction of FXa inhibitors.

Keywords: Anticoagulation; Coagulation; Factor Xa (FXa); Inhibitor; Non-inhibitor; Molecular descriptors; Machine learning; Support vector machine (SVM); Thrombotic disease

1. Introduction

Factor Xa (FXa) is a serine protease involved in the blood coagulation cascade and it plays a vital role in the regulation of normal homeostasis and abnormal intravascular thrombus development [1,2]. Inhibition of FXa provides effective means for controlling blood coagulation process and thus for the treatment of thrombotic diseases including life-threatening stroke, deep vein thrombosis, and pulmonary embolism. Current antithrombotic agents have certain limitations in oral bioavailability [3] and in the requirement of individual dose titration and periodic monitoring [4]. Therefore, there is a need for developing new orally active anticoagulants and intensive efforts have been directed at the discovery of FXa inhibitors [5–12].

As part of the effort for developing fast and low-cost tools for facilitating the discovery of new anticoagulants, several computational methods have been explored for predicting FXa inhibitors. Molecular docking method has been used to find potential inhibitors by identifying those compounds that can be docked into the inhibitor binding site of the 3D structure of FXa, which achieves accuracies of 80% for inhibitors and 85% for non-inhibitors based on the test of 112 ligands [13,14]. A pharmacophore model has been developed from the structure of the known inhibitors with a reported inhibitor prediction accuracy of 69% based on the test of 36 inhibitors [15]. Specific structural and physicochemical properties of the known inhibitors have been used to derive 3D quantitative structure–activity relationships (3D-QSAR) with a reported inhibitor prediction accuracy of 84–88% based on the test of 279

[*] Corresponding author at: Bioinformatics and Drug Design Group, Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore. Tel.: +65 6516 6877; fax: +65 6774 6756.
E-mail address: phacyz@nus.edu.sg (Y.Z. Chen).

1093-3263/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2007.03.003

inhibitors [16]. A machine learning method, support vector machines (SVM), has been used to classify FXa inhibitors and non-inhibitors from the structural and physicochemical properties of compounds with reported accuracies of 80% for inhibitors and 99% for non-inhibitors based on the test of 25 inhibitors and 200 non-inhibitors [17].

The prediction accuracies of these methods are in the range of 69–88% for FXa inhibitors and 85–99% for non-inhibitors, which are at a useful level for facilitating the prediction of FXa inhibitors. Most of the prediction models have been developed and tested by using no more than ~100 compounds that are significantly less than the 360 known FXa inhibitors. While a 3D-QSAR model has been developed and tested by using 279 inhibitors and 156 non-inhibitors with a structure diversity index (DI) [18] of 0.693 compared to that of 0.646 for the 360 known FXa inhibitors, the prediction performance of some of these methods may be further improved and more adequately tested by using a larger number of non-inhibitors as well the 360 known FXa inhibitors in the model development and testing process. In this work, a total of 738 non-inhibitors with a DI value of 0.474 were used for model development compared to that of 156 non-inhibitors with a DI value of 0.712 used in the 3D-QSAR model.

SVM and other machine learning methods have been shown to be particularly useful for predicting various pharmacodynamic, pharmacokinetic and toxicological properties of compounds of diverse structures [19–26]. It is of interest to improve the performance of SVM and to explore other machine learning methods for facilitating the prediction of FXa inhibitors of more diverse range of structures than those in previous studies [14,16,17]. The performance of these machine learning methods are dependent on the use of appropriate set of molecular descriptors suitable for distinguishing FXa inhibitors and non-inhibitors. Feature selection methods, such as recursive feature elimination (RFE), have been frequently used for extracting molecular descriptors relevant to specific types of pharmaceutical agents [27–29]. In this work, the RFE method was used for selecting molecular descriptors relevant to the prediction of FXa inhibitors.

The machine learning methods used in this work are *k* nearest neighbor (*k*-NN) [30], probabilistic neural network (PNN) [31], C4.5 decision tree (C4.5 DT) [32] and SVM [33,34]. A comprehensive literature search was conducted to collect diverse set of literature-reported FXa inhibitors and non-inhibitors. Two evaluation methods were used to objectively assess the performance of these methods. One is five-fold cross-validation and the other is validation by the use of an independent validation set of known FXa inhibitors and non-inhibitors.

2. Methods

2.1. Selection of FXa inhibitors and non-inhibitors

A total of 360 FXa inhibitors and 255 non-inhibitors were collected from a number of published papers [6,7,13,14,17,35–68]. The 360 inhibitors can be categorized into 22 structural groups based on their substructure combinations [69], which are listed in Table 1. Each inhibitor is composed of three substructures, P1, Linker, and P4. As shown in Table 2, there are five different subtypes for P1 (P1-1 to P1-5), eight for Linker (L1–L8), and three for P4 (P4-1 to P4-3). Each of these 360 inhibitors can be formed by some combinations of these subtypes in the P1–Linker–P4 format. Additional FXa non-

Table 1
Structural groups and sub-structure compositions of the known factor Xa inhibitors

Structural group	Substructure composition			No. of inhibitors	Examples of inhibitors
	P1	Linker	P4		
G1	P1-1	L1	P4-2	32	Gong2000bmcl_9b, Guertin2002bmcl_9c, Han2000jmc_43
G2	P1-1	L1	P4-3	14	Czekaj2002bmcl_15, Czekaj2002bmcl_9b, Guertin2002bmcl_5a
G3	P1-1	L2	P4-2	60	Fevig2001bmcl_5a, Pinto2001jmc_3b, Pruitt2000bmcl_14
G4	P1-1	L2	P4-3	35	Quan2005jmc_35, Smallheer2004bmcl_3d
G5	P1-1	L3	P4-2	8	Buckman1998bmcl_22a, Phillips2002jmc_6i
G6	P1-1	L3	P4-3	57	Ng2002bmc_5aa, Willardsen2004jmc_2
G7	P1-1	L4	P4-3	2	Shaw2002bmcl_14c, Shaw2002bmcl_15d
G8	P1-1	L5	P4-2	5	Fevig1998bmcl_19, Fevig1998bmcl_23
G9	P1-1	L5	P4-3	1	Maduskuie1998jmc_21
G10	P1-1	L6	P4-1	7	Rai2001bmc_13, Rai2001bmc_15
G11	P1-1	L6	P4-2	34	Choi-Sledeski1999jmc_4b, Fevig1999bmcl_30, Rai2001bmc_19
G12	P1-1	L6	P4-3	12	Becker1999bmcl_1q, Becker1999bmcl_1w, Choi-Sledeski1999jmc_20b
G13	P1-1	L7	P4-2	6	Wiley2000jmc_48, Wiley2000jmc_52
G14	P1-1	L8	P4-2	1	He2000bmcl_1
G15	P1-2	L4	P4-2	1	Arnaiz2000bmcl_11d
G16	P1-2	L4	P4-3	43	Arnaiz2000bmcl_14b, Shaw2002bmcl_20c, Zhao2000bmcl_13f
G17	P1-2	L8	P4-1	1	Wu2002bmcl_6f
G18	P1-2	L8	P4-2	3	Wu2002bmcl_6b, Wu2002bmcl_6e, Wu2002bmcl_6g
G19	P1-2	L8	P4-3	17	Gabriel1998jmc_DX-9065a, Wu2002bmcl_4d
G20	P1-3	L3	P4-3	1	Buckman1998bmcl_1
G21	P1-4	L2	P4-3	1	Nazare2005jmc_45
G22	P1-5	L2	P4-3	19	Nazare2005jmc_24, Nazare2005jmc_27

Table 2
Subtypes of the substructures of the known factor Xa inhibitors

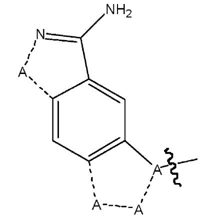
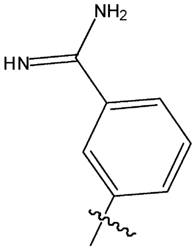
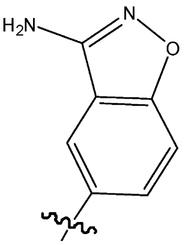
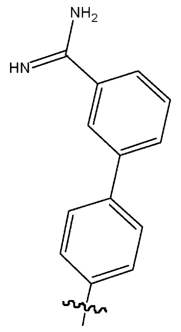
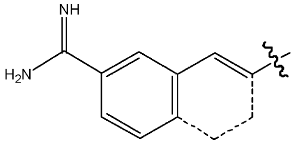
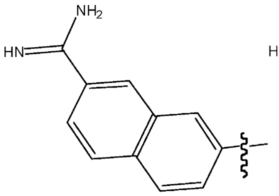
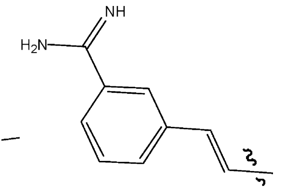
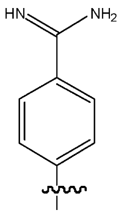
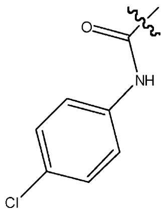
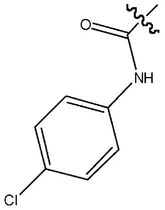
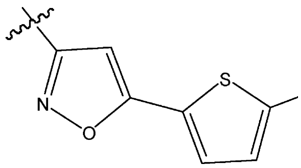
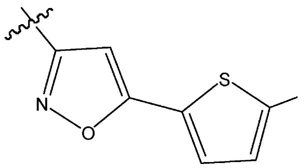
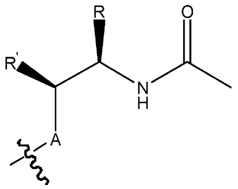
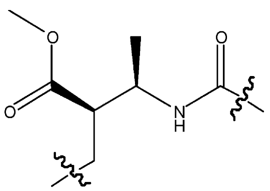
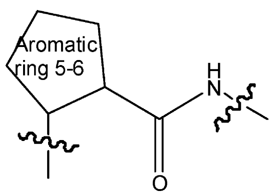
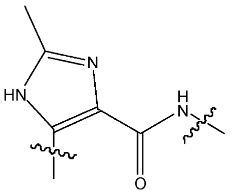
Substructure	Subtype	Structural framework	Sample substructures
P1	P1-1		  
	P1-2		 
	P1-3		
	P1-4		
	P1-5		
Linker	L1		
	L2		

Table 2 (*Continued*)

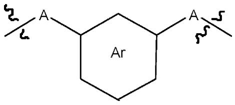
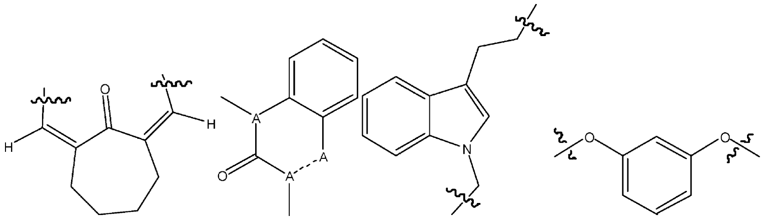
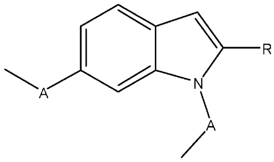
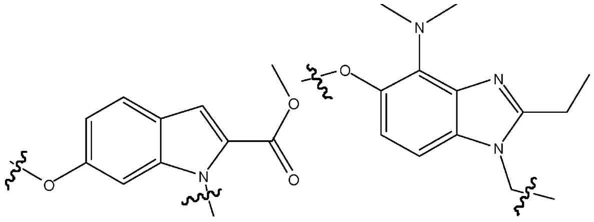
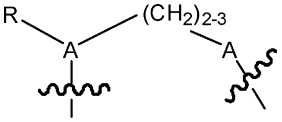
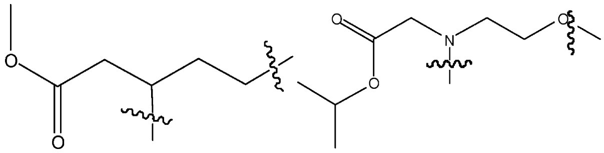
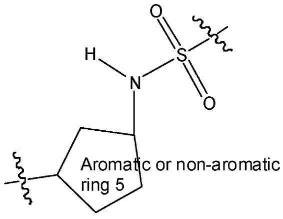
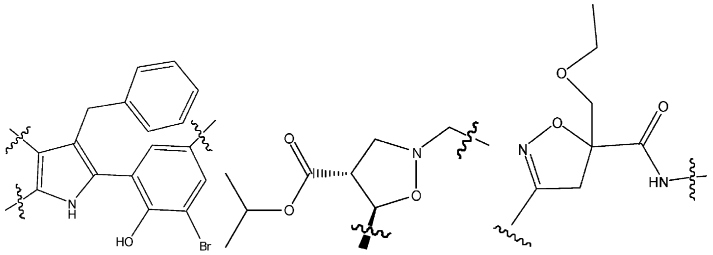
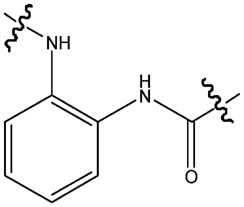
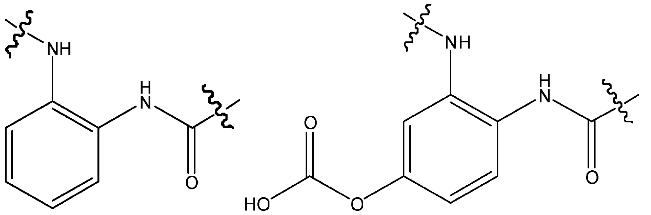
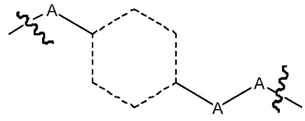
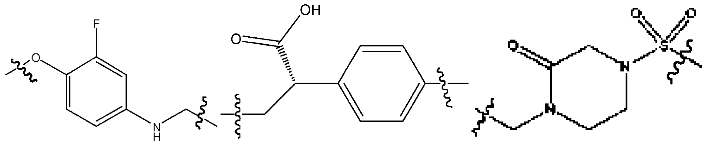
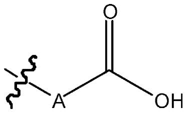
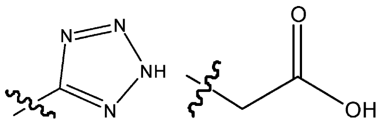
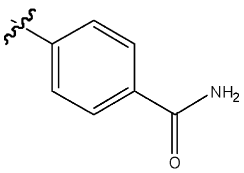
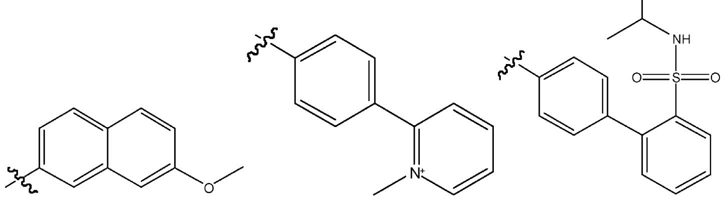
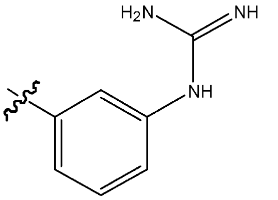
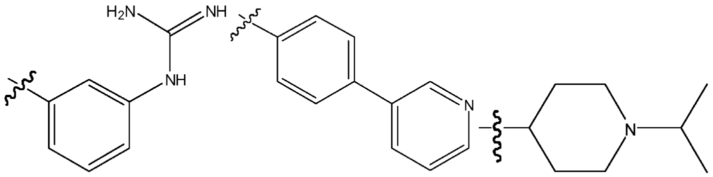
Substructure	Subtype	Structural framework	Sample substructures
	L3		
	L4		
	L5		
	L6		
	L7		
	L8		
P4	P4-1		
	P4-2		

Table 2 (*Continued*)

Substructure	Subtype	Structural framework	Sample substructures
	P4-3		

inhibitors were selected from those well-studied clinical drugs that are both known inhibitors of proteins other than FXa and are not reported to produce anticoagulant effect. Such a method for selecting additional non-inhibitors has been successfully used in the prediction of inhibitors and substrates of other proteins [17,70,71]. This method is based on the following considerations. These well-studied drugs have been extensively studied, monitored and clinically used. If they have not been reported to be an inhibitor of a protein (e.g. FXa inhibitor) and to produce the effects (e.g. anticoagulation) associated with the inhibition of that protein, it is highly unlikely that they are inhibitors of that protein. From this procedure, 483 additional non-inhibitors were generated. The list of all of the FXa inhibitors and non-inhibitors is given in Table S1 to S3 of the supplementary material.

The 2D structure of each of the compounds were generated by using ChemDraw [72] and were subsequently converted into 3D structure by using DS ViewerPro 5.0 [73] followed by optimization using its ‘‘clean structure’’ module. All of the generated conformations were fully optimized without symmetry restrictions. The 3D structure of each compound was manually inspected to ensure that the chirality of each chiral agent is properly generated.

Structural diversity of these compounds can be measured by using the DI value, which is the average value of the similarity between pairs of compounds in a dataset [18]:

$$\mathrm{DI} = \frac{\sum_{i=1}^{N}\sum_{j=1,i\neq j}^{N}\mathrm{sim}(i,j)}{N(N-1)}$$

where sim(*i*,*j*) is a measure of the similarity between compound *i* and *j* and *N* is the number of compounds in the dataset. The structural diversity of a dataset increases with decreasing DI value. In this work, sim(*i*,*j*) is computed by using the Tanimoto coefficient [74]:

$$\mathrm{sim}(i,j) = \frac{\sum_{d=1}^{l} x_{di}x_{dj}}{\sum_{d=1}^{l}(x_{di})^2 + \sum_{d=1}^{l}(x_{dj})^2 - \sum_{d=1}^{l} x_{di}x_{dj}}$$

where *l* is the number of descriptors computed for the molecules in the dataset. Table 3 gives the DI value of the dataset used in this work and those used in previous studies [14,16,17]. The results in Table 3 suggest that the dataset used in this work is slightly more diverse than one dataset and substantially more diverse than the other datasets used in earlier studies.

FXa inhibitors and non-inhibitors were further divided into training and testing sets by two different methods, five-fold cross-validation and validation by an independent evaluation set. In the first method, the group of 360 inhibitors and 738 non-inhibitors was each randomly divided into five subsets of approximately equal size, respectively. Four of the subsets were used as the training set, and the remaining subset was used as the testing set for the inhibitors and non-inhibitors, respectively. This process was repeated five times such that every subset is used as the test set once. In the second method, these compounds were divided into training and independent validation set based on their distribution in the chemical space. Chemical space is defined by the commonly used structural and chemical descriptors [75]. Compounds of similar structural and chemical features were evenly assigned into separate sets. For those compounds without enough number of structurally and chemically similar counterparts, they were assigned, in order of priority, to the training and then the independent validation set, respectively. The training set was used for developing the prediction system and the independent validation set was used for assessing the accuracy of the system. The generated training and independent evaluation set contains 715 (282 inhibitors and 433 non-inhibitors) and 383 (78 inhibitors and 305 non-inhibitors) compounds, respectively.

2.2. Molecular descriptors

Molecular descriptors are quantitative representations of structural and physicochemical features of molecules, which have been extensively used in QSAR [5,16,76] and machine learning methods [19–26] for predicting pharmaceutical agents

Table 3
Diversity indices of several datasets used for developing and testing FXa inhibitor prediction models

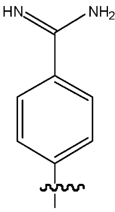
Dataset	Number of FXa inhibitors/non-inhibitors	Diversity index for inhibitors
Current work	360/738	0.646
3D-QSAR model [16]	279/156	0.693
SVM classification [17]	50/200	0.716
Molecular docking [14]	133/86	0.780

of various properties. A total of 199 molecular descriptors were used in this work, which have been selected from more than 1000 descriptors described in the literature by eliminating those descriptors that are obviously redundant or unrelated to the prediction of pharmaceutical agents [21,77]. These descriptors include 18 descriptors in the class of simple molecular properties (such as molecular weight and number of rotatable bonds), 28 descriptors in the class of molecular connectivity and shape (such as molecular connectivity indices and molecular kappa shape indices), 97 descriptors in the class of electro-topological state (such as electro-topological state indices), 31 descriptors in the class of quantum chemical properties (such as atomic charges and molecular dipole moment), and 25 descriptors in the class of geometrical properties (such as solvent accessible surface area and hydrophobic region). These descriptors were computed from the 3D structure of each compound by using our own designed molecular descriptor computing program. The remaining redundant and unrelated descriptors were further reduced by using feature selection methods [78–80].

2.3. Feature selection method

Feature selection methods have been introduced for the improvement of the performance of machine learning methods by removing redundant and irrelevant molecular descriptors and selecting those descriptors relevant to a particular study [81]. The RFE method [78,79] has gained popularity due to its effectiveness for selection of molecular descriptors relevant to drug activity analysis and prediction of inhibitors and substrates [21,70]. Thus RFE method is used in this work, and the details of the implementation of this method can be found in our earlier publications [21,77].

Feature selection procedure can be demonstrated by the following illustrative example of the development of a SVM classification system. This system is trained by using a Gaussian kernel function with an adjustable parameter σ. Sequential variation of σ is conducted against the whole training set to find a value that gives the best prediction accuracy. This prediction accuracy is evaluated by means of five-fold cross-validation. In the first step, for a fixed σ, the SVM classifier is trained by using the complete set of features (molecular descriptors) described in the previous section. The second step involves the computation of the ranking criterion score DJ(i) for each feature in the current set. All of the computed DJ(i) is subsequently ranked in descending order. The third step involves the removal the m features with smallest criterion scores. In the fourth step, the SVM classification system is re-trained by using the remaining set of features, and the corresponding prediction accuracy is computed by means of five-fold cross-validation. The first to fourth steps are then repeated for other values of σ. After the completion of these procedures, the set of features and parameter σ that give the best prediction accuracy are selected.

The choice of the parameter m affects the performance of SVM as well as the speed of feature selection. Although it is desirable to remove one feature at a time (m  $= 1$), this is often difficult due to high CPU cost. It has been found that, in some cases, removal of several features at a time ($m > 1$) significantly improves computational efficiency without losing too much accuracy [79]. Our studies on compounds of different pharmacokinetic properties [21,77] suggested that the accuracy of a SVM system with $m = 5$ being only a few percentage smaller than that with $m = 1$, which is consistent with the findings from other studies [78,82]. Thus, for computational efficiency, $m = 5$ is used in this study.

2.4. Machine learning methods

2.4.1. Support vector machine (SVM)

The theory of SVM has been extensively described in the literature [33,34]. Thus, only a brief description is given here. SVM is based on the structural risk minimization (SRM) principle from statistical learning theory [33]. In linearly separable cases, SVM constructs a hyper-plane which separates two different classes of vectors with a maximum margin. A vector $\mathbf{x}_i$ is composed of the molecular descriptors of a molecule and the two classes are FXa inhibitors class and non-inhibitors class. The hyper-plane is constructed by finding another vector $\mathbf{w}$ and a parameter b that minimizes $||\mathbf{w}||^2$ and satisfies the following conditions:

$$\mathbf{w}\cdot\mathbf{x}_i + b \geq +1, \quad \text{for}\, y_i = +1 \quad \text{Class 1 (positive)} \tag{1}$$

$$\mathbf{w}\cdot\mathbf{x}_i + b \leq -1, \quad \text{for}\, y_i = -1 \quad \text{Class 2 (negative)} \tag{2}$$

where y_i is the class index, $\mathbf{w}$ the vector normal to the hyperplane, $|b|/||\mathbf{w}||$ the perpendicular distance from the hyperplane to the origin and $||\mathbf{w}||^2$ is the Euclidean norm of $\mathbf{w}$. After the determination of $\mathbf{w}$ and b, a given vector $\mathbf{x}_i$ can be classified by

$$\mathrm{sign}[(\mathbf{w}\cdot\mathbf{x}) + b] \tag{3}$$

In nonlinearly separable cases, SVM maps the input vectors into a higher dimensional feature space by using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. An example of a kernel function is the Gaussian kernel which has been extensively used in different studies with good results [83–85]:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{e}^{-||\mathbf{x}_j - \mathbf{x}_i||^2/2\sigma^2} \tag{4}$$

Linear support vector machine is then applied to this feature space and then the decision function is given by

$$f(\mathbf{x}) = \mathrm{sign}\left(\sum_{i=1}^{l}\alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b\right) \tag{5}$$

where the coefficients α_i^0 and b are determined by maximizing the following Langrangian expression:

$$\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{6}$$

under the following conditions:

$$\alpha_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{l}\alpha_i y_i = 0 \tag{7}$$

A positive or negative value from Eq. (5) indicates that the vector $\mathbf{x}$ belongs to the positive or negative class, respectively.

2.4.2. *k*-NN

k-NN measures the Euclidean distance between a to-be-classified vector $\mathbf{x}$ and each individual vector $\mathbf{x}_i$ in the training set [30,86]. The Euclidean distances for the vector pairs are calculated using the following formula:

$$D = \sqrt{||\mathbf{x} - \mathbf{x}_i||^2} \tag{8}$$

A total of k number of vectors nearest to the vector $\mathbf{x}$ are used to determine its class, $f(\mathbf{x})$:

$$\hat{f}(\mathbf{x}) \leftarrow \text{argmax}_{v \in V}\sum_{i=1}^{k}\delta(v, f(\mathbf{x}_i)) \tag{9}$$

where $\delta(a,b) = 1$ if $a = b$ and $\delta(a,b) = 0$ if $a \neq b$, argmax is the maximum of the function, V is a finite set of vectors $\{v_1, \ldots, v_s\}$ and $\hat{f}(\mathbf{x})$ is an estimate of $f(\mathbf{x})$. Here estimate refers to the class of the majority of the k nearest neighbors.

2.4.3. Probabilistic neural network (PNN)

PNN is a form of neural network designed for classification through the use of Bayes' optimal decision rule [31]:

$$h_i c_i f_i(\mathbf{x}) > h_j c_j f_j(\mathbf{x}) \tag{10}$$

where h_i and h_j are the prior probabilities, c_i and c_j the costs of misclassification and $f_i(\mathbf{x})$ and $f_j(\mathbf{x})$ are the probability density function for class i and j, respectively. An unknown vector $\mathbf{x}$ is classified into population i if the product of all the three terms is greater for class i than for any other class j (not equal to i). In most applications, the prior probabilities and costs of misclassifications are treated as being equal. The probability density function for each class for a univariate case can be estimated by using the Parzen's nonparametric estimator [87]:

$$g(\mathbf{x}) = \frac{1}{n\sigma}\sum_{i=1}^{n}W\left(\frac{\mathbf{x} - \mathbf{x}_i}{\sigma}\right) \tag{11}$$

where n is the sample size, σ is a scaling parameter which defines the width of the bell curve that surrounds each sample point, $W(d)$ is a weight function which has its largest value at $d = 0$ and $(\mathbf{x} - \mathbf{x}_i)$ is the distance between the unknown vector and a vector in the training set. The Parzen's nonparametric estimator was later expanded by Cacoullos [88] for the multivariate case:

$$g(x_1, \ldots, x_p) = \frac{1}{n\sigma_1, \ldots, \sigma_p}\sum_{i=1}^{n}W\left(\frac{x_1 - x_{1,i}}{\sigma_1}, \ldots, \frac{x_p - x_{p,i}}{\sigma_p}\right) \tag{12}$$

The Gaussian function is frequently used as the weight function because it is well behaved, easily calculated and satisfies the conditions required by Parzen's estimator. Thus, the probability density function for the multivariate case becomes

$$g(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\exp\left(-\sum_{j=1}^{p}\left(\frac{x_j - x_{ij}}{\sigma_j}\right)^2\right) \tag{13}$$

The network architectures of PNN are determined by the number of compounds and descriptors in the training set. There are four layers in a PNN. The input layer provides input values to all neurons in the pattern layer and has as many neurons as the number of descriptors in the training set. The number of pattern neurons is determined by the total number of compounds in the training set. Each pattern neuron computes a distance measure between the input and the training case represented by that neuron and then subjects the distance measure to the Parzen's nonparametric estimator. The summation layer has a neuron for each class and the neurons sum all the pattern neurons' output corresponding to members of that summation neuron's class to obtain the estimated probability density function for that class. The single neuron in the output layer then estimates the class of the unknown vector $\mathbf{x}$ by comparing all the probability density function from the summation neurons and choosing the class with the highest probability density function.

2.4.4. C4.5 decision tree (DT)

C4.5 DT is a branch-test-based classifier [32]. A branch in a decision tree corresponds to a group of classes and a leaf represents a specific class. A decision node specifies a test to be conducted on a single attribute value, with one branch and its subsequent classes as possible outcomes of the test. C4.5 DT uses recursive partitioning to examine every attribute of the data and rank them according to their ability to partition the remaining data, thereby constructing a decision tree. A vector $\mathbf{x}$ is classified by starting at the root of the tree and moving through the tree until a leaf is encountered. At each non-leaf decision node, a test is conducted and the classification process proceeds to the branch selected by the test. Upon reaching the destination leaf, the class of the vector $\mathbf{x}$ is predicted to be that associated with the leaf.

The algorithm is a recursive greedy heuristic that selects descriptors for membership within the tree. Whether or not a descriptor is included within the tree is based on the value of its information gain. As a statistical property, information gain measures how well the descriptor separate training cases into subsets in which the class is homogeneous. Given that the descriptors in this study were all continuous variables, a threshold value had to be established within each descriptor so that it could partition the training cases into subsets. These

Table 4
The accuracy of FXa inhibitors and non-inhibitors derived from SVM without the use of a feature selection method (SVM) and from SVM with the use of the feature selection method RFE (SVM + RFE) by using five-fold cross-validation

Method	Cross-validation	FXa inhibitors			FXa non-inhibitors			Q (%)	C
		TP	FN	Accuracy SE (%)	TN	FP	Accuracy SP (%)		
SVM	1	67	4	94.4	141	8	94.6	94.5	0.878
	2	67	6	91.8	143	4	97.3	95.5	0.897
	3	69	4	94.5	142	5	96.6	95.9	0.908
	4	61	5	92.4	147	6	96.1	95	0.881
	5	70	7	90.9	133	9	93.7	92.7	0.841
	Average			92.8			95.6	94.7	0.881
	S.D.			1.6			1.5	1.2	0.025
SVM + RFE	1	68	3	95.8	145	4	97.3	96.8	0.928
	2	67	6	91.8	145	2	98.6	96.4	0.918
	3	71	2	97.3	145	2	98.6	98.2	0.959
	4	60	6	90.9	152	1	99.3	96.8	0.924
	5	75	2	97.4	137	5	96.5	96.8	0.931
	Average			94.6			98.1	97	0.932
	S.D.			3.1			1.2	0.7	0.016

The results are given in TP (true positive), FN (false negative), TN (true negative), FP (false positive), Q (overall accuracy), C (Matthews correlation coefficient), SE (sensitivity) is the prediction accuracy for FXa inhibitors and SP (specificity) is the prediction accuracy for non-inhibitors. Statistical significance is indicated by S.D. (standard deviation). The number of FXa inhibitors or FXa non-inhibitors is TP + FN or TN + FP.

threshold values for each descriptor were established by rank ordering the values within each descriptor from lowest to highest and repeatedly calculating the information gain using the arithmetical midpoint between all successive values within the rank order. The midpoint value with the highest information gain was selected as the threshold value for the descriptor. That descriptor with the highest information gain (information being the most useful for classification) was then selected for inclusion in the DT. The algorithm continued to build the tree in this manner until it accounted for all training cases. Ties between descriptors that were equal in terms of information gain were broken randomly [89].

2.5. Performance evaluation

As in the case of all discriminative methods [90,91], the performance of statistical learning methods can be measured by the quantity of true positives TP, true negatives TN, false positives FP, false negatives FN, sensitivity SE = TP/(TP + FN) which is the prediction accuracy for the FXa inhibitors in this work, and specificity SP = TN/(TN + FP) which is the prediction accuracy for the FXa non-inhibitors in this work. The overall prediction accuracy (Q) and Matthews correlation coefficient (C) [92] are also used to measure the prediction accuracies and can be given by

$$Q = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{10}$$

$$C = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}} \tag{11}$$

3. Results and discussion

3.1. Overall prediction accuracies and performance evaluation

The effect of feature selection method RFE on the performance of machine learning methods for predicting FXa inhibitors can be determined by comparing the computed accuracies of SVM with and without the use of RFE. From Table 4, one can find that the accuracies of SVM with RFE are 94.6% for FXa inhibitors and 98.1% for non-inhibitors, which are substantially better than those of 92.8% for FXa inhibitors and 95.6% for non-inhibitors derived from SVM without RFE. Similar prediction accuracies were found in two additional

Table 5
Comparison of the prediction accuracies of FXa inhibitors and non-inhibitors derived from different machine learning methods by using five-fold cross-validation in this work

Method	Parameter	FXa inhibitor accuracy SE (%)	FXa non-inhibitor accuracy SP (%)	Overall accuracy, Q (%)
C4.5 DT		89.1	93.8	92.3
PNN	б = 0.17	97.5	92.3	94.0
k-NN	k = 9	94.1	93.2	93.5
SVM + RFE	б = 0.2	94.6	98.1	97.0

C4.5 DT (C4.5 decision tree), PNN (probabilistic neural network), k-NN (k nearest neighbor), SVM + RFE (support vector machine and recursive feature elimination).

Table 6
Comparison of the prediction accuracies of FXa inhibitors and non-inhibitors derived from different machine learning methods by using independent validation set in this work

Method	Parameter	TP	FN	TN	FP	FXa inhibitor accuracy SE (%)	FXa non-inhibitor accuracy SP (%)	Overall accuracy, *Q* (%)
C4.5 DT		71	7	287	18	91.0	94.1	93.5
PNN	б = 0.3	74	4	291	14	94.9	95.4	95.3
k-NN	*k* = 9	77	1	289	16	98.7	94.8	95.6
SVM	б = 1	77	1	299	6	98.7	98.0	98.2

five-fold cross-validation studies conducted by using training–testing sets separately generated from different random number seed parameters. This suggests that RFE is useful in selecting the appropriate set of molecular descriptors for distinguishing FXa inhibitors and non-inhibitors.

Table 5 gives the prediction accuracies of FXa inhibitors and non-inhibitors derived from other three machine learning methods *k*-NN, PNN and C4.5 DT by using the RFE selected descriptors and five-fold cross-validation method. For comparison, those from SVM are also included in Table 5. The inhibitor prediction accuracies from the other three methods are comparable to each other. For FXa inhibitors, the accuracies of these methods are in the range of 89.1–97.5% with PNN giving the best accuracy at 97.5%. For FXa non-inhibitors, the accuracies from these methods are in the range of 92.3–98.1% with SVM giving the best accuracy at 98.1%.

The capability of SVM and other three methods were further evaluated by using an independent evaluation set of 78 FXa inhibitors and 305 non-inhibitors described in Section 2. Table 6 shows the prediction results by using this independent set. The prediction accuracies for the FXa inhibitors are in the range of 91.0–98.7%, which is compared to the accuracy of 69–88% for the prediction of FXa inhibitors from earlier studies [13–17]. For the FXa non-inhibitors the prediction accuracies are found in the range of 94.1–98.0% which is compared to the accuracy of 85–99% for the prediction of FXa non-inhibitors from earlier studies [13–17].

Overall, our study suggests that statistical learning methods, particularly SVM, are useful for facilitating the prediction of FXa inhibitors of diverse structures. The FXa inhibitor prediction accuracy of these methods is generally improved and the non-inhibitor prediction accuracy is at a similar level as those of earlier studies in which a substantially less diverse set of compounds have been used.

3.2. *Evaluation of generalization capability of machine learning methods*

The generalization capability of the four machine learning methods, SVM, *k*-NN, PNN and C4.5 DT, was evaluated by using training and testing sets that are composed of entirely different structural groups without any of their members being in the other set [69]. A training set was formed by using all of the 253 inhibitors from 10 randomly selected structural groups G1–G7, G15, G16, and G21 and 375 randomly selected non-inhibitors. The corresponding testing set is composed of all of the 107 inhibitors from the remaining 12 structural groups G8–G14, G17–G20, and G22 and 363 non-inhibitors. This training set was used to develop the four machine learning prediction models, and their performance was subsequently tested by using the testing set. The results are shown in Table 7. The prediction accuracies for the FXa inhibitors are in the range of 57.0–74.8% with the SVM model giving the best performance. In spite of the exclusion in the training process of any member of the structural groups in the testing set, machine learning methods, particularly SVM, appear to have some capacity for predicting inhibitors in these structural groups. This suggests that SVM and other machine learning methods are potentially useful for predicting novel FXa inhibitors beyond currently known structural frameworks. The prediction accuracies of FXa non-inhibitors are in the range of 95.0–98.1%, which are comparable to that of 94.1–98.0% derived from randomly assembled independent evaluation set. This suggests that, at least for this case, the performance of

Table 7
Comparison of the prediction accuracies of FXa inhibitors and non-inhibitors derived from different machine learning methods trained by using 253 inhibitors from the G1–G7, G15, G16, and G21 structural groups and 375 randomly selected non-inhibitors

Method	Parameter	TP	FN	TN	FP	FXa inhibitor accuracy SE (%)	FXa non-inhibitor accuracy SP (%)	Overall accuracy, Q (%)
C4.5 DT		61	46	352	11	57.0	97.0	87.9
PNN	б = 0.1	72	35	345	18	67.3	95.0	88.7
k-NN	*k* = 3	79	28	347	16	73.8	95.6	90.6
SVM	б = 1	80	27	356	7	74.8	98.1	92.8

The prediction accuracies were derived from the test of 107 inhibitors from the G8–G14, G17–G20, and G22 structural groups and 363 non-inhibitors.

Table 8
Molecular descriptors selected from the RFE feature selection method for classification of FXa inhibitors and non-inhibitors

Molecular descriptor	Description	Matched or partially matched molecular descriptors used in published QSAR models
ndonr	Number of H-bond donors	H-bond donor [5,94,95]
${}^3\chi_C$	Simple molecular connectivity Chi indices for cluster	Aromatic ring [94]
${}^4\chi_{PC}$	Simple molecular connectivity Chi indices for path/cluster	Aromatic ring [94]
${}^1\chi^v$	Valence molecular connectivity Chi indices for path order 1	Aromatic ring [94]
${}^2\chi^v$	Valence molecular connectivity Chi indices for path order 2	Aromatic ring [94]
${}^3\chi^v_C$	Valence molecular connectivity Chi indices for cluster	Aromatic ring [94]
S(1)	Atom-type H Estate sum for –OH	Molar refractivity of functional group [5,64,96–98]
S(2)	Atom-type H Estate sum for =NH	Molar refractivity of functional group [5,64,96–98]
S(12)	Atom-type H Estate sum for CH n (Saturated)	
S(25)	Atom-type Estate sum for =C⟨	
S(26)	Atom-type Estate sum for: C:–	
S(27)	Atom-type Estate sum for: C::	
S(30)	Atom-type Estate sum for =NH	Molar refractivity of functional group [5,64,96–98]
Tiwie	Information Weiner	
εb	Hydrogen bond acceptor basicity (covalent HBAB)	H-bond acceptor [5,94,95]
IP	Ionization potential	Electrostatic field [5,95], positive ionizable [94]
$Q_{H,Max}$	Most positive charge on H atoms	Electrostatic [5]
$Q_{H,Min}$	Most negative charge on H atoms	Electrostatic [5]
Mnc	Mean of negative charges	Electrostatic [5]
Mac	Mean absolute charge	Electrostatic [5]
dis2	Length vectors (longest third atom)	Steric field descriptors [5,95]
Sapc	Sum of solvent accessible surface areas of positively charged atoms	Atomic contributions to Van der Waals surface area [102]
Sanc	Sum of solvent accessible surface areas of negatively charged atoms	Atomic contributions to Van der Waals surface area [102]
Rugty	Molecular rugosity	Steric field descriptors [5,95]
Capty	Capacity factor	Electrostatic field [5,95]
Hiwpb	Hydrophobic intery moment	Hydrophobic field [5,95], hydrophobic aromatic [94]

machine learning methods for predicting non-inhibitors are relatively insensitive to the biased selection of training inhibitors.

3.3. Relevance of selected molecular descriptors for predicting FXa inhibitors

A total of 26 molecular descriptors are selected by RFE as the most relevant for distinguishing between FXa inhibitors and non-inhibitors, which are given in Table 8 together with the corresponding matched or partially matched molecular features from other studies. Many of these selected descriptors are consistent with the reported molecular features of FXa inhibitors. Studies of the 3D structure of FXa-inhibitor complexes have shown that many first-generation FXa inhibitors rely on the interaction of basic amidine with an acidic amino acid at the binding pocket [66], and hydrogen bonding and specific hydrophobic groups are also important for inhibitor binding [15,66]. Molecular docking studies have suggested that electrostatic interactions, hydrogen bonding, hydrophobic ring structures, and steric interactions are important for inhibitor binding to FXa [13,14,93]. Electrostatic, hydrophobic, aromatic ring, steric, hydrogen bond donor and acceptor, and molar refractivity terms have frequently been used for constructing QSAR models of FXa inhibitors [5,66,94–98].

Of the 26 RFE selected descriptors, six descriptors (Mnc, Mac, $Q_{H,Max}$, $Q_{H,Min}$, IP, Capty) are associated with electrostatic properties for mean charges, the charge of specific hydrogens, ionization potential and concentration of polar interactions on molecular surface, respectively. Moreover, S(1), S(2) and S(30) describe specific polar functional groups. Sapc and Sanc represent the sum of solvent accessible surface areas of positively and negatively charged atoms. Two other descriptors, ndonr and εb, represents the number of H-bond donors and the basicity of H-bond acceptor, respectively. ${}^3\chi_C$, ${}^4\chi^v_{PC}$, ${}^2\chi^v$, and ${}^3\chi^v_C$ describe simple and valence molecular connectivity for a cluster or path of atoms, which can be used to describe ring as well as other structures. Hiwpb describes hydrophobic intery moment. Rugty represents molecular rugosity that measures the ratio between the bare molecular surface area and molecular volume. Overall, these 26 descriptors seem to be capable of collectively describe most

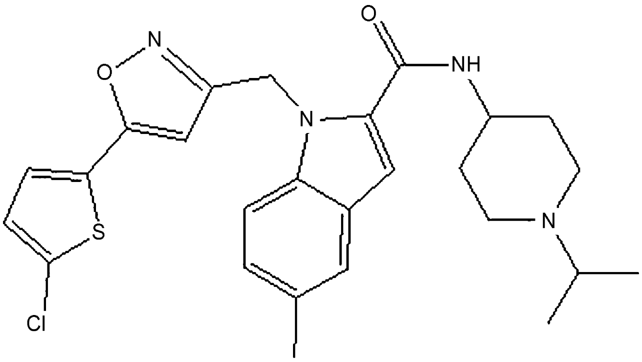


Fig. 1. The structure of misclassified FXa inhibitor, compound **29** from Nazare et al. [66].

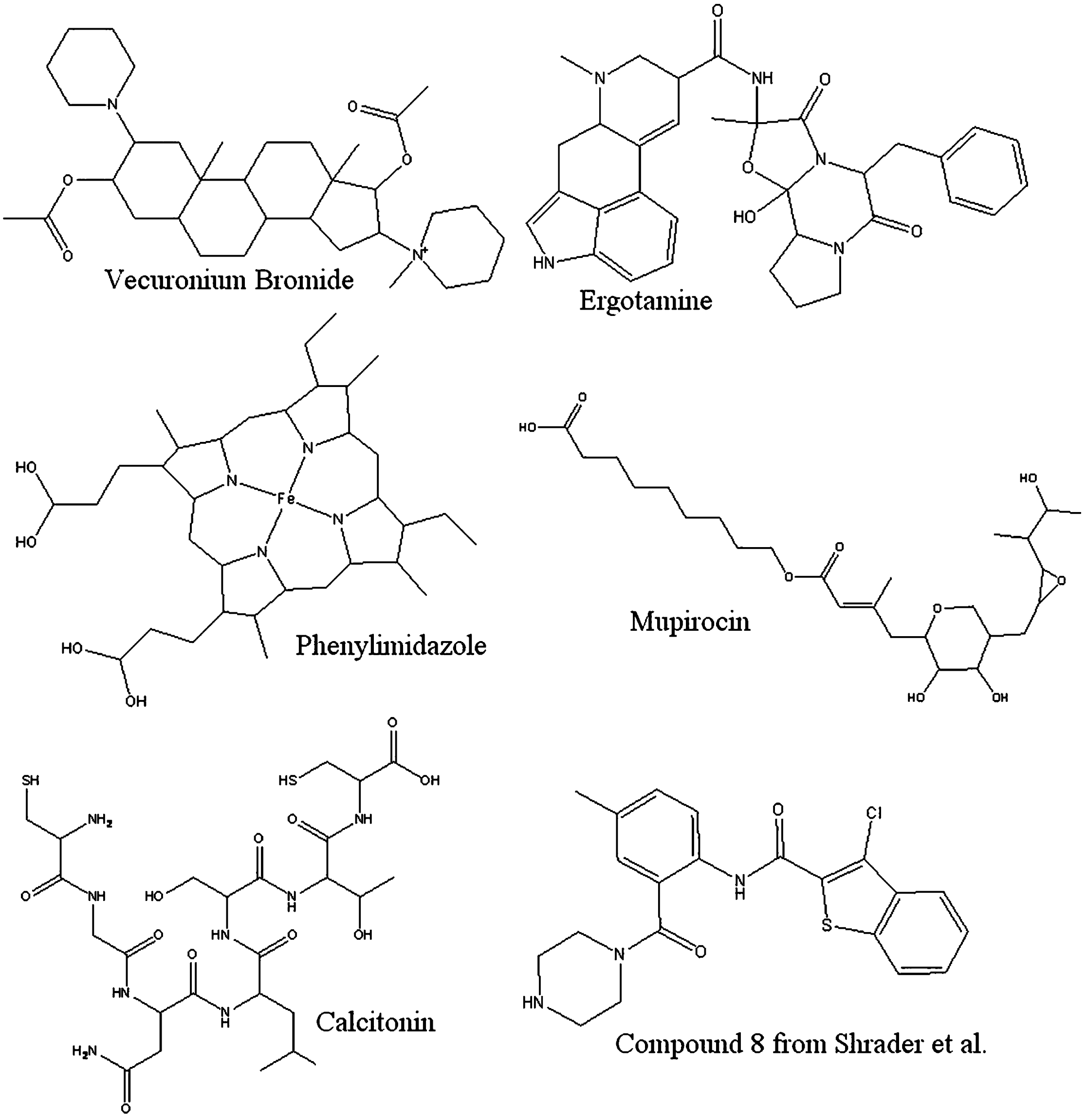


Fig. 2. The structures of misclassified FXa non-inhibitors.

of the molecular features of FXa inhibitors discussed in other studies.

There is one FXa inhibitor and six FXa non-inhibitors in the independent evaluation set that were misclassified by SVM, which are shown in Figs. 1 and 2, respectively. The misclassified FXa inhibitor is compound **29** from Nazare et al. [66], which has a relatively poor binding affinity to FXa (Ki = 9 nM) indicating that it is a week FXa inhibitor. One possible reason for the misclassification of this compound is that its structural features may be closer to some FXa non-binders that those of FXa inhibitors. The use of more extensive number of weak inhibitors in machine learning training process may enable the further improvement of the prediction accuracy of this and other weak inhibitors. The six misclassified FXa non-inhibitors are compound **8** from Shrader et al. [13], vecuronium bromide, mupirocin, calcitonin, ergotomine, and phenylimidazole. The misclassification of vecuronium bromide, ergotamine, phenylimidazole, calcitonin, mupirocin may be due to the fact that these compounds contain complex ring structures or long chains that are inadequately represented in currently used descriptors [99,100]. While encoding molecular shape and flexibility features, topological descriptors may not adequately describe the detailed configuration of large rigid structure combined with a short flexible hydrophilic tail in the molecule.

4. Conclusion

Machine learning methods such as SVM, *k*-NN and PNN are potentially useful for facilitating the prediction of FXa inhibitors of diverse structures. Current efforts are being directed at the improvement of the efficiency and speed of feature selection method [82], which can further help to optimally select molecular descriptors most relevant to specific pharmaceutical class of agents such as FXa inhibitors. Moreover, recent works on the introduction of weighting function into the descriptors of statistical learning methods such as SVM [101] may also be helpful in improving the prediction accuracy of these methods. These may enable the development of statistical learning methods into practical tools for the prediction of FXa inhibitors and other pharmaceutically relevant agents.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2007.03.003.

References

[1] B. Furie, Furie B.C., The molecular basis of blood coagulation, Cell 53 (4) (1988) 505–518.

[2] G.J. Broze Jr., Tissue factor pathway inhibitor and the current concept of blood coagulation, Blood Coagul. Fibrinol. 6 (Suppl 1) (1995) S7–S13.

[3] J.I. Witting, P. Bourdon, D.V. Brezniak, J.M. Maraganore, J.W. Fenton II, Thrombin-specific inhibition by and slow cleavage of hirulog-1, Biochem. J. 283 (Pt 3) (1992) 737–743.

[4] R.D. Bona, A.D. Hickey, D.M. Wallace, Efficacy and safety of oral anticoagulation in patients with cancer, Thromb. Haemost. 78 (1) (1997) 137–140.

[5] C.A. Kontogiorgis, D. Hadjipavlou-Litina, Current trends in quantitative structure activity relationships on FXa inhibitors: evaluation and comparative analysis, Med. Res. Rev. 24 (6) (2004) 687–747.

[6] Q. Han, C. Dominguez, P.F. Stouten, J.M. Park, D.E. Duffy, R.A. Galemmo Jr., K.A. Rossi, R.S. Alexander, A.M. Smallwood, P.C. Wong, et al., Design, synthesis, and biological evaluation of potent and selective amidino bicyclic factor Xa inhibitors, J. Med. Chem. 43 (23) (2000) 4398–4415.

[7] J.A. Willardsen, D.A. Dudley, W.L. Cody, L. Chi, T.B. McClanahan, T.E. Mertz, R.E. Potoczak, L.S. Narasimhan, D.R. Holland, S.T. Rapundalo, et al., Design, synthesis, and biological activity of potent and selective inhibitors of blood coagulation factor Xa, J. Med. Chem. 47 (16) (2004) 4089–4099.

[8] N. Haginoya, S. Kobayashi, S. Komoriya, T. Yoshino, T. Nagata, Y. Hirokawa, T. Nagahara, Design, synthesis, and biological activity of non-amidine factor Xa inhibitors containing pyridine *N*-oxide and 2-carbamoylthiazole units, Bioorg. Med. Chem. 12 (21) (2004) 5579–5586.

[9] K. Sato, S. Kaku, F. Hirayama, H. Koshio, Y. Matsumoto, T. Kawasaki, Y. Iizumi, Antithrombotic effect of YM-75466 is separated from its effect on bleeding time and coagulation time, Eur. J. Pharmacol. 352 (1) (1998) 59–63.

[10] P.C. Wong, M.L. Quan, E.J. Crain, C.A. Watson, R.R. Wexler, R.M. Knabb, Nonpeptide factor Xa inhibitors. I. Studies with SF303 and SK549, a new class of potent antithrombotics, J. Pharmacol. Exp. Ther. 292 (1) (2000) 351–357.

[11] G.B. Phillips, B.O. Buckman, D.D. Davey, K.A. Eagen, W.J. Guilford, J. Hinchman, E. Ho, S. Koovakkat, A. Liang, D.R. Light, et al., Discovery of *N*-[2-[5-[amino(imino)methyl]-2-hydroxyphenoxy]-3, 5-difluoro-6-[3-(4, 5-dihydro-1-methyl-1*H*-imidazol-2-yl)phenoxy]pyridin-4-yl]-*N*-methylglycine (ZK-807834): a potent, selective, and orally active inhibitor of the blood coagulation enzyme factor Xa, J. Med. Chem. 41 (19) (1998) 3557–3562.

[12] D. Leung, G. Abbenante, D.P. Fairlie, Protease inhibitors: current status and future prospects, J. Med. Chem. 43 (3) (2000) 305–341.

[13] W.D. Shrader, W.B. Young, P.A. Sprengeler, J.C. Sangalang, K. Elrod, G. Carr, Neutral inhibitors of the serine protease factor Xa, Bioorg. Med. Chem. Lett. 11 (14) (2001) 1801–1804.

[14] M. Murcia, A.R. Ortiz, Virtual screening with flexible docking and COMBINE-based models. Application to a series of factor Xa inhibitors, J. Med. Chem. 47 (4) (2004) 805–820.

[15] E.M. Krovat, K.H. Fruhwirth, T. Langer, Pharmacophore identification, in silico screening, and virtual library design for inhibitors of the human factor Xa, J. Chem. Inf. Model 45 (1) (2005) 146–159.

[16] F. Fontaine, M. Pastor, I. Zamora, F. Sanz, Anchor-GRIND: filling the gap between standard 3D QSAR and the GRid-INdependent descriptors, J. Med. Chem. 48 (7) (2005) 2687–2694.

[17] R.N. Jorissen, M.K. Gilson, Virtual screening of molecular databases using a support vector machine, J. Chem. Inf. Model 45 (3) (2005) 549–561.

[18] J.J. Perez, Managing molecular diversity, Chem. Soc. Rev. 34 (2) (2005) 143–152.

[19] V.V. Zernov, K.V. Balakin, A.A. Ivaschenko, N.P. Savchuk, I.V. Pletnev, Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions, J. Chem. Inform. Comp. Sci. 43 (6) (2003) 2048–2056.

[20] E. Byvatov, U. Fechner, J. Sadowski, G. Schneider, Comparison of support vector machine and artificial neural network systems for drug/nondrug classification, J. Chem. Inform. Comp. Sci. 43 (6) (2003) 1882–1889.

[21] Y. Xue, C.W. Yap, L.Z. Sun, Z.W. Cao, J.F. Wang, Y.Z. Chen, Prediction of *p*-glycoprotein substrates by support vector machine approach, J. Chem. Inform. Comp. Sci. 44 (4) (2004) 1497–1505.

[22] S. Doniger, T. Hofman, J. Yeh, Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms, J. Comput. Biol. 9 (6) (2002) 849–864.

[23] L. He, P.C. Jurs, L.L. Custer, S.K. Durham, G.M. Pearl, Predicting the genotoxicity of polycyclic aromatic compounds from molecular structure with different classifiers, Chem. Res. Toxicol. 16 (12) (2003) 1567–1580.

[24] R.D. Snyder, G.S. Pearl, G. Mandakas, W.N. Choy, F. Goodsaid, I.Y. Rosenblum, Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules, Environ. Mol. Mutagen. 43 (3) (2004) 143–158.

[25] C.W. Yap, C.Z. Cai, Y. Xue, Y.Z. Chen, Prediction of torsade-causing potential of drugs by support vector machine approach, Toxicol. Sci. 79 (1) (2004) 170–177.

[26] C.W. Yap, Y.Z. Chen, Quantitative structure–pharmacokinetic relationships for drug distribution properties by using general regression neural network, J. Pharm. Sci. 94 (1) (2004) 153–168.

[27] Y. Xue, Z.R. Li, C.W. Yap, L.Z. Sun, X. Chen, Y.Z. Chen, Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents, J. Chem. Inform. Comp. Sci. 44 (5) (2004) 1630–1638.

[28] G.G. Kuiper, B. Carlsson, K. Grandien, E. Enmark, J. Haggblad, S. Nilsson, J.A. Gustafsson, Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors alpha and beta, Endocrinology 138 (3) (1997) 863–870.

[29] W.S. Branham, S.L. Dial, C.L. Moland, B.S. Hass, R.M. Blair, H. Fang, L. Shi, W. Tong, R.G. Perkins, D.M. Sheehan, Phytoestrogens and mycoestrogens bind to the rat uterine estrogen receptor, J. Nutr. 132 (4) (2002) 658–664.

[30] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, Englewood Cliffs, NJ, 1982.

[31] D.F. Specht, Probabilistic neural networks, Neural Netw. 3 (1) (1990) 109–118.

[32] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.

[33] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[34] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining Knowl. Discov. 2 (2) (1998) 127–167.

[35] D.O. Arnaiz, Z. Zhao, A. Liang, L. Trinh, M. Witlow, S.K. Koovakkat, K.J. Shaw, Design, synthesis, and in vitro biological activity of indole-

based factor Xa inhibitors, Bioorg. Med. Chem. Lett. 10 (9) (2000) 957–961.

[36] M.R. Becker, W.R. Ewing, R.S. Davis, H.W. Pauls, C. Ly, A. Li, H.J. Mason, Y.M. Choi-Sledeski, A.P. Spada, V. Chu, et al., Synthesis, SAR and in vivo activity of novel thienopyridine sulfonamide pyrrolidinones as factor Xa inhibitors, Bioorg. Med. Chem. Lett. 9 (18) (1999) 2753–2758.

[37] B.O. Buckman, R. Mohan, S. Koovakkat, A. Liang, L. Trinh, M.M. Morrissey, Design, synthesis, and biological activity of novel purine and bicyclic pyrimidine factor Xa inhibitors, Bioorg. Med. Chem. Lett. 8 (16) (1998) 2235–2240.

[38] Y.M. Choi-Sledeski, D.G. McGarry, D.M. Green, H.J. Mason, M.R. Becker, R.S. Davis, W.R. Ewing, W.P. Dankulich, V.E. Manetta, R.L. Morris, et al., Sulfonamidopyrrolidinone factor Xa inhibitors: potency and selectivity enhancements via P-1 and P-4 optimization, J. Med. Chem. 42 (18) (1999) 3572–3587.

[39] M. Czekaj, S.I. Klein, K.R. Guertin, C.J. Gardner, A.L. Zulli, H.W. Pauls, A.P. Spada, D.L. Cheney, K.D. Brown, D.J. Colussi, et al., Optimization of the beta-aminoester class of factor Xa inhibitors. Part 1. P(4) and side-chain modifications for improved in vitro potency, Bioorg. Med. Chem. Lett. 12 (12) (2002) 1667–1670.

[40] W.R. Ewing, M.R. Becker, V.E. Manetta, R.S. Davis, H.W. Pauls, H. Mason, Y.M. Choi-Sledeski, D. Green, D. Cha, A.P. Spada, et al., Design and structure–activity relationships of potent and selective inhibitors of blood coagulation factor Xa, J. Med. Chem. 42 (18) (1999) 3557–3571.

[41] J.M. Fevig, J. Cacciola, R.S. Alexander, R.M. Knabb, G.N. Lam, P.C. Wong, R.R. Wexler, Preparation of meta-amidino-*N*,*N*-disubstituted anilines as potent inhibitors of coagulation factor Xa, Bioorg. Med. Chem. Lett. 8 (22) (1998) 3143–3148.

[42] B. Gabriel, M.T. Stubbs, A. Bergner, J. Hauptmann, W. Bode, J. Sturzebecher, L. Moroder, Design of benzamidine-type inhibitors of factor Xa, J. Med. Chem. 41 (22) (1998) 4240–4250.

[43] R.A. Galemmo Jr., T.P. Maduskuie, C. Dominguez, K.A. Rossi, R.M. Knabb, R.R. Wexler, P.F. Stouten, The de novo design and synthesis of cyclic urea inhibitors of factor Xa: initial SAR studies, Bioorg. Med. Chem. Lett. 8 (19) (1998) 2705–2710.

[44] Y. Gong, H.W. Pauls, A.P. Spada, M. Czekaj, G. Liang, V. Chu, D.J. Colussi, K.D. Brown, J. Gao, Amido-(propyl and allyl)-hydroxybenzamidines: development of achiral inhibitors of factor Xa, Bioorg. Med. Chem. Lett. 10 (3) (2000) 217–221.

[45] K.R. Guertin, C.J. Gardner, S.I. Klein, A.L. Zulli, M. Czekaj, Y. Gong, A.P. Spada, D.L. Cheney, S. Maignan, J.P. Guilloteau, et al., Optimization of the beta-aminoester class of factor Xa inhibitors. Part 2. Identification of FXV673 as a potent and selective inhibitor with excellent in vivo anticoagulant activity, Bioorg. Med. Chem. Lett. 12 (12) (2002) 1671–1674.

[46] W.J. Guilford, K.J. Shaw, J.L. Dallas, S. Koovakkat, W. Lee, A. Liang, D.R. Light, M.A. McCarrick, M. Whitlow, B. Ye, et al., Synthesis, characterization, and structure–activity relationships of amidine-substituted (bis)benzylidene-cycloketone olefin isomers as potent and selective factor Xa inhibitors, J. Med. Chem. 42 (26) (1999) 5415–5425.

[47] W. He, B. Hanney, M.R. Myers, A.P. Spada, K. Brown, D. Colussi, V. Chu, Nonbenzamidine compounds as selective factor Xa inhibitors, Bioorg. Med. Chem. Lett. 10 (15) (2000) 1737–1739.

[48] S.D. Jones, J.W. Liebeschuetz, P.J. Morgan, C.W. Murray, A.D. Rimmer, J.M. Roscoe, B. Waszkowycz, P.M. Welsh, W.A. Wylie, S.C. Young, et al., The design of phenylglycine containing benzamidine carboxamides as potent and selective inhibitors of factor Xa, Bioorg. Med. Chem. Lett. 11 (5) (2001) 733–736.

[49] T.P. Maduskuie Jr., K.J. McNamara, Y. Ru, R.M. Knabb, P.F. Stouten, Rational design and synthesis of novel, potent bis-phenylamidine carboxylate factor Xa inhibitors, J. Med. Chem. 41 (1) (1998) 53–62.

[50] S. Maignan, J.P. Guilloteau, S. Pouzieux, Y.M. Choi-Sledeski, M.R. Becker, S.I. Klein, W.R. Ewing, H.W. Pauls, A.P. Spada, V. Mikol, Crystal structures of human factor Xa complexed with potent inhibitors, J. Med. Chem. 43 (17) (2000) 3226–3232.

[51] H. Matter, E. Defossa, U. Heinelt, P.M. Blohm, D. Schneider, A. Muller, S. Herok, H. Schreuder, A. Liesum, V. Brachvogel, et al., Design and quantitative structure–activity relationship of 3-amidinobenzyl-1*H*-indole-2-carboxamides as potent, nonchiral, and selective inhibitors of blood coagulation factor Xa, J. Med. Chem. 45 (13) (2002) 2749–2769.

[52] H.P. Ng, B.O. Buckman, K.A. Eagen, W.J. Guilford, M.J. Kochanny, R. Mohan, K.J. Shaw, S.C. Wu, D. Lentz, A. Liang, et al., Design, synthesis, and biological activity of novel factor Xa inhibitors: 4-aryloxy substituents of 2,6-diphenoxypyridines, Bioorg. Med. Chem. 10 (3) (2002) 657–666.

[53] G. Phillips, W.J. Guilford, B.O. Buckman, D.D. Davey, K.A. Eagen, S. Koovakkat, A. Liang, M. McCarrick, R. Mohan, H.P. Ng, et al., Design, synthesis, and activity of a novel series of factor Xa inhibitors: optimization of arylamidine groups, J. Med. Chem. 45 (12) (2002) 2484–2493.

[54] D.J. Pinto, M.J. Orwat, S. Wang, J.M. Fevig, M.L. Quan, E. Amparo, J. Cacciola, K.A. Rossi, R.S. Alexander, A.M. Smallwood, et al., Discovery of 1-[3-(aminomethyl)phenyl]-*N*-3-fluoro-2′-(methylsulfonyl)-[1,1′-biphenyl]-4-yl]-3-(trifluoromethyl)-1*H*-pyrazole-5-carboxamide (DPC423), a highly potent, selective, and orally bioavailable inhibitor of blood coagulation factor Xa, J. Med. Chem. 44 (4) (2001) 566–578.

[55] J.R. Pruitt, D.J. Pinto, M.J. Estrella, L.L. Bostrom, R.M. Knabb, P.C. Wong, M.R. Wright, R.R. Wexler, Isoxazolines and isoxazoles as factor Xa inhibitors, Bioorg. Med. Chem. Lett. 10 (8) (2000) 685–689.

[56] M.L. Quan, A.Y. Liauw, C.D. Ellis, J.R. Pruitt, D.J. Carini, L.L. Bostrom, P.P. Huang, K. Harrison, R.M. Knabb, M.J. Thoolen, et al., Design and synthesis of isoxazoline derivatives as factor Xa inhibitors 1, J. Med. Chem. 42 (15) (1999) 2752–2759.

[57] M.L. Quan, C.D. Ellis, A.Y. Liauw, R.S. Alexander, R.M. Knabb, G. Lam, M.R. Wright, P.C. Wong, R.R. Wexler, Design and synthesis of isoxazoline derivatives as factor Xa inhibitors 2, J. Med. Chem. 42 (15) (1999) 2760–2773.

[58] M. Renatus, W. Bode, R. Huber, J. Sturzebecher, M.T. Stubbs, Structural and functional analyses of benzamidine-based inhibitors in complex with trypsin: implications for the inhibition of factor Xa, tPA, and urokinase, J. Med. Chem. 41 (27) (1998) 5445–5456.

[59] K. Sagi, T. Nakagawa, M. Yamanashi, S. Makino, M. Takahashi, M. Takayanagi, K. Takenaka, N. Suzuki, S. Oono, N. Kataoka, et al., Rational design, synthesis, and structure–activity relationships of novel factor Xa inhibitors: (2-substituted-4-amidinophenyl)pyruvic and -propionic acids, J. Med. Chem. 46 (10) (2003) 1845–1857.

[60] K.J. Shaw, W.J. Guilford, B.D. Griedel, S. Sakata, L. Trinh, S. Wu, W. Xu, Z. Zhao, M.M. Morrissey, Benzimidazole-based fXa inhibitors with improved thrombin and trypsin selectivity, Bioorg. Med. Chem. Lett. 12 (9) (2002) 1311–1314.

[61] J.M. Smallheer, R.S. Alexander, J. Wang, S. Wang, S. Nakajima, K.A. Rossi, A. Smallwood, F. Barbera, D. Burdick, J.M. Luettgen, et al., SAR and factor IXa crystal structure of a dual inhibitor of factors IXa and Xa, Bioorg. Med. Chem. Lett. 14 (21) (2004) 5263–5267.

[62] E. Verner, B.A. Katz, J.R. Spencer, D. Allen, J. Hataye, W. Hruzewicz, H.C. Hui, A. Kolesnikov, Y. Li, C. Luong, et al., Development of serine protease inhibitors displaying a multicentered short (<2.3 A) hydrogen bond binding mode: inhibitors of urokinase-type plasminogen activator and factor Xa, J. Med. Chem. 44 (17) (2001) 2753–2771.

[63] M.R. Wiley, L.C. Weir, S. Briggs, N.A. Bryan, J. Buben, C. Campbell, N.Y. Chirgadze, R.C. Conrad, T.J. Craft, J.V. Ficorilli, et al., Structure-based design of potent, amidine-derived inhibitors of factor Xa: evaluation of selectivity, anticoagulant activity, and antithrombotic activity, J. Med. Chem. 43 (5) (2000) 883–899.

[64] S. Wu, W.J. Guilford, Y.L. Chou, B.D. Griedel, A. Liang, S. Sakata, K.J. Shaw, L. Trinh, W. Xu, Z. Zhao, et al., Design and synthesis of aminophenol-based factor Xa inhibitors, Bioorg. Med. Chem. Lett. 12 (9) (2002) 1307–1310.

[65] Z. Zhao, D.O. Arnaiz, B. Griedel, S. Sakata, J.L. Dallas, M. Whitlow, L. Trinh, J. Post, A. Liang, M.M. Morrissey, et al., Design, synthesis, and in vitro biological activity of benzimidazole based factor Xa inhibitors, Bioorg. Med. Chem. Lett. 10 (9) (2000) 963–966.

[66] M. Nazare, D.W. Will, H. Matter, H. Schreuder, K. Ritter, M. Urmann, M. Essrich, A. Bauer, M. Wagner, J. Czech, et al., Probing the subpockets of factor Xa reveals two binding modes for inhibitors based on a 2-carboxyindole scaffold: a study combining structure–activity relationship and X-ray crystallography, J. Med. Chem. 48 (14) (2005) 4511–4525.

[67] J. Cui, D. Crich, D. Wink, M. Lam, A.L. Rheingold, D.A. Case, W. Fu, Y. Zhou, M. Rao, A.J. Olson, et al., Design and synthesis of highly constrained factor Xa inhibitors: amidine-substituted bis(benzoyl)-diazepan-2-ones and bis(benzylidene)-bis(gem-dimethyl)cycloketones, Bioorg. Med. Chem. 11 (16) (2003) 3379–3392.

[68] R. Rai, A. Kolesnikov, Y. Li, W.B. Young, E. Leahy, P.A. Sprengeler, E. Verner, W.D. Shrader, J. Burgess-Henry, J.C. Sangalang, et al., Development of potent and selective factor Xa inhibitors, Bioorg. Med. Chem. Lett. 11 (14) (2001) 1797–1800.

[69] A.C. Good, M.A. Hermsmeier, S.A. Hindle, Measuring CAMD technique performance: a virtual screening case study in the design of validation experiments, J. Comput. Aided Mol. Des. 18 (7–9) (2004) 529–536.

[70] C.W. Yap, Y.Z. Chen, Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines, J. Chem. Inf. Model 45 (4) (2005) 982–992.

[71] L. Molnar, G.M. Keseru, A neural network based virtual screening of cytochrome P450 3A4 inhibitors, Bioorg. Med. Chem. Lett. 12 (3) (2002) 419–421.

[72] CambridgeSoft Corporation, ChemDraw In., 7.0.1 edn., CambridgeSoft Corporation, Cambridge, MA, USA, 2002.

[73] Accelrys: DS ViewPro. In., 5.0 edn. San Diego, CA: Accelrys, 2002.

[74] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comp. Sci. 38 (1998) 983–996.

[75] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley–VCH, Weinheim, 2000.

[76] J.Y. Hu, T. Aizawa, Quantitative structure–activity relationships for estrogen receptor binding affinity of phenolic chemicals, Water Res. 37 (6) (2003) 1213–1222.

[77] Y. Xue, Z.R. Li, C.W. Yap, L.Z. Sun, X. Chen, Y.Z. Chen, Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents, J. Chem. Inform. Comp. Sci. 44 (2004) 1630–1638.

[78] H. Yu, J. Yang, W. Wang, J. Han, Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines, in: IEEE Computer Society Bioinformatics Conference (CSB'03), Stanford, CA, (2003), pp. 220–228.

[79] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422.

[80] S. Degroeve, B. De Baets, Y. Van de Peer, P. Rouzé, Feature subset selection for splice site prediction, Bioinformatics 18 (2002) S75–S83.

[81] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (10) (2000) 906–914.

[82] C. Furlanello, M. Serafini, S. Merler, G. Jurman, An accelerated procedure for recursive feature ranking on microarray data, Neural Netw. 16 (5/6) (2003) 641–648.

[83] M.W.B. Trotter, B.F. Buxton, S.B. Holden, Support vector machines in combinatorial chemistry, Meas. Control 34 (8) (2001) 235–239.

[84] R. Burbidge, M. Trotter, B. Buxton, S. Holden, Drug design by machine learning: support vector machines for pharmaceutical data analysis, Comp. Chem. 26 (1) (2001) 5–14.

[85] R. Czerminski, A. Yasri, D. Hartsough, Use of support vector machine in pattern classification: application to QSAR studies, Quant. Struct.–Activity Relationships 20 (3) (2001) 227–240.

[86] E. Fix, J.L. Hodges, Discriminatory Analysis: Non-Parametric Discrimination: Consistency Properties, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[87] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (1962) 1065–1076.

[88] T. Cacoullos, Estimation of a multivariate density, Ann. Inst. Stat. Math. 18 (1966) 179–189.

[89] B. Carnahan, G. Meyer, L.-A. Kuntz, Comparing statistical and machine learning classifiers: alternatives for predictive modeling in human factors research, Hum. Factors 45 (2003) 408–423.

[90] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (5) (2000) 412–424.

[91] J.E. Roulston, Screening with tumor markers, Mol. Pharmacol. 20 (2002) 153–162.

[92] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta 405 (2) (1975) 442–451.

[93] M.S. Rao, A.J. Olson, Modelling of factor Xa-inhibitor complexes: a computational flexible docking approach, Proteins 34 (2) (1999) 173–183.

[94] M.O. Taha, A.M. Qandil, D.D. Zaki, M.A. AlDamen, Ligand-based assessment of factor Xa binding site flexibility via elaborate pharmacophore exploration and genetic algorithm-based QSAR modeling, Eur. J. Med. Chem. 40 (7) (2005) 701–727.

[95] M. Bohm, J. St rzebecher, G. Klebe, Three-dimensional quantitative structure–activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa, J. Med. Chem. 42 (3) (1999) 458–477.

[96] Y.K. Yee, A.L. Tebbe, J.H. Linebarger, D.W. Beight, T.J. Craft, D. Gifford-Moore, T. Goodson Jr., D.K. Herron, V.J. Klimkowski, J.A. Kyle, et al., *N*(2)-Aroylanthranilamide inhibitors of human factor Xa, J. Med. Chem. 43 (5) (2000) 873–882.

[97] K. Roy, A.U. De, C. Sengupta, QSAR of human factor Xa inhibitor *N*2-aroylanthranilamides using principal component factor analysis, Drug Des. Discov. 18 (1) (2002) 23–31.

[98] K. Roy, A.U. De, C. Sengupta, QSAR with electrotopological state atom index: human factor Xa inhibitor *N*2-aroylanthranilamides, Drug Des. Discov. 18 (1) (2002) 33–43.

[99] H. Li, C.Y. Ung, C.W. Yap, Y. Xue, Z.R. Li, Z.W. Cao, Y.Z. Chen, Prediction of genotoxicity of chemical compounds by statistical learning methods, Chem. Res. Toxicol. 18 (6) (2005) 1071–1080.

[100] H. Li, C.Y. Ung, C.W. Yap, Y. Xue , Z.R. Li, Y.Z. Chen. Prediction of estrogen receptor agonists and characterization of associated molecular descriptors by statistical learning methods. J. Mol. Graph Model (2006).

[101] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, Choosing multiple parameters for support vector machines, Mach. Learn. 46 (1–3) (2002) 131–159.

[102] P. Labute, A widely applicable set of descriptors, J. Mol. Graph Model 18 (4/5) (2000) 464–477.