Exploration of 3,6-dihydroimidazo(4,5-d)pyrrolo(2,3-b)pyridin-2(1H)-one derivatives as JAK inhibitors using various in silico techniques

Radhakrishnan S. Jisha¹ · Lilly Aswathy¹ · Vijay H. Masand² · Jayant M. Gajbhiye³ · Indira G. Shibi¹

Received: 28 July 2017 / Accepted: 26 September 2017
© Springer-Verlag GmbH Germany 2017

Abstract This study focuses on understanding the structural features of 3,6-dihydroimidazo(4,5-d)pyrrolo(2,3-b)pyridin-2(1H)-one (dpp) derivatives to computationally identify new JAK inhibiting compounds. For the purpose, a novel virtual screening strategy, with 2D and 3D-QSAR (CoMFA and CoMSIA), data mining, pharmacophore modeling, ADMET prediction, multi-targeted protein-based docking and inverse QSAR, was employed. The 2D-QSAR equations developed for the JAK3, JAK2 and JAK1 involved five physicochemical descriptors. These descriptors correlate with the anti-RA activity with R^2 values for JAK3, JAK2 and JAK1 are 0.9811, 0.8620 and 0.9740, respectively. The 3D-QSAR studies such as CoMFA and CoMSIA carried out through PLS analysis of the training set of JAK3, JAK2 and JAK1, gave Q^2 values as 0.369, 0.476 and 0.490; R^2_{nev} values as 0.863, 0.684 and 0.724 and, F values as 23.098, 28.139 and 31.438, respectively. The contour maps produced by the CoMFA and CoMSIA models were used to understand the importance of hydrogen bond donor, acceptor, hydrophobic, steric and electrostatic interactions. The molecular docking studies of these selected compounds with various

JAK proteins were carried out and the protein–ligand interactions were also studied. The study concluded that dpp15(s) is a highly potent JAK inhibitor with a very good predicted IC_{50} value.

Keywords Rheumatoid arthritis · Janus kinase · 2D and 3D QSAR · Weka · Molecular docking

Introduction

Rheumatoid arthritis (RA) causes inflammation of the membrane around the joints and muscles and it leads to pain and stiffness, and results in deformities and loss of function of synovial joints. Associated severe inflammation, secondary to RA, often leads to changes in bone metabolism and high risk of cardiovascular disease. Around 1% of the adult population in the developed countries is reported to be affected by RA and most of them have self-limited diseases, joint-destructions, severe physical disability, etc. (Riise et al. 2000; Jacobsson et al. 1994; Plenge 2009). The disease commonly appears among the age group of 25–50 and affects women three times more than it affects men (Smith and Haynes 2002). The diagnosis of RA is very difficult because of the difference of symptoms in different patients (Kroot et al. 2000). RA increases the mortality rate of patients who are suffering from other diseases (Gonzalez et al. 2007). Pulmonary disease is a major cause of death in RA and patients carrying the epitope of two HLA-DRB1*04 clusters run a higher risk (Toyoshima et al. 1993; Smolen et al. 2007; Weyand et al. 1992). There are also reports that among almost 50% of the RA patients some kind of respiratory problems develop. (Fujii et al. 1993; Gabbay et al. 1997).

Studies have revealed that pro-inflammatory cytokines TNF- α and IL-6; IL-1 and IL-17 and effector cells (T cells

Electronic supplementary material The online version of this article (doi:10.1007/s40203-017-0029-x) contains supplementary material, which is available to authorized users.

✉ Indira G. Shibi
shibiig@gmail.com

- ¹ Department of Chemistry, Sree Narayana College, Chempazhanthy, Thiruvananthapuram, Kerala 695 587, India
- ² Department of Chemistry, Vidya Bharati College, Camp, Amravati, Maharashtra 444 602, India
- ³ Division of Organic Chemistry, CSIR-National Chemical Laboratory, Pune 411 008, India

and B cells) are the different immune modulators, and the corresponding signaling pathways play an important role in the pathophysiology of RA (Smolen et al. 2007; Smolen and Steiner 2003). The joint damage begins at the synovial membrane mainly due to the complex interaction of these immune modulators (Smolen and Steiner 2003). An enhanced knowledge of the disease pathology has made the treatment strategies better now. But many medications, when used for a long-term, create side effects and so are dangerous. Many medicines used for the treatment for RA lead to progressive toxicity and thus necessitate the urge for developing new drugs for RA treatment (Boers et al. 1997; Orbach et al. 2002).

Janus Kinases (JAKs) are regarded as potential targets for the treatment of autoimmune disease RA because of their unique role in the immune system. JAKs are cytoplasmic protein tyrosine kinases containing four different groups such as JAK1, JAK2, JAK3 and TYK2, which play vital roles in several forms of cytokine-mediated signal transduction (Imada and Leonard 2000). Adopting kinase inhibition methods for the treatment of RA has now become very common (Fridman et al. 2010). Therefore in this study, we selected the JAKs targets, compounds and the corresponding inhibitory values, in order to probe the structural features of the compounds with a view to developing structure–activity relationships. Recently, Yamagishi et al. (2015) prepared some derivatives of 3,6-dihydroimidazo(4,5-d)pyrrolo(2,3-b)pyridin-2(1H)-one and conducted in vitro studies to correlate structure and activity to explore the inhibitory activities (IC₅₀) against JAK3, JAK2 and JAK1. We selected these datasets and focused on in silico analyses to understand the structural features of the scaffold of 3,6-dihydroimidazo(4,5-d)pyrrolo(2,3-b)pyridin-2(1H)-one (dpp) and their derivatives, with a view to finding new lead compounds. Another advantage is that these compounds have the 1H-pyrrolo(2,3-b)pyridine ring which matches well with the pyrrolopyrimidine scaffold of the widely used tofacitinib, which is a prominent JAK inhibiting drug.

In the study, Yamagishi et al. have concluded that (±)-cis-3-(4-Methyl-3-(2-oxo-3,6-dihydroimidazo(4,5-d)pyrrolo(2,3-b)pyridin-1(2H)-yl)piperidin-1-yl)-3-oxopropanenitrile is an attractive lead with potent JAK inhibitory nature. For the design of the molecule, the approach adopted was to reduce the lipophilicity by substituting the cyclohexane ring with heterocycles of the parent scaffold. Their experimental results fulfilled their intention to find novel tricyclic JAK inhibitors. To design a better analogue, the present study attempts to develop a new strategy for lead optimization by employing Computer Aided Drug Design (CADD) (Kapetanovic 2008). In the present work, a combined approach, whereby ligand-based drug design

techniques, viz. QSAR (2D- and 3D-), data mining through machine learning, molecular docking and pharmacophore modeling, has been made to determine the structural features of the molecules in the experimental dataset so as to develop a significant correlation with the JAK activity profile of 3,6-dihydroimidazo(4,5-d)pyrrolo(2,3-b)pyridin-2(1H)-one derivatives.

CADD is a better method than the conventional ‘trial and error’ methods of drug discovery process because its associated methods evidently unravel the intricate structural patterns that mainly contribute towards the inhibitory activity, pharmacokinetics, and toxicity profiles of a drug candidate. The in silico techniques are faster, more economical and obviously more successful than the conventional methods. However, 2D-QSAR studies alone are insufficient to draw reliable conclusions, especially for comparatively small datasets. A right blend of different ideas, algorithms, tools and techniques thus become essential to comprehend the reasons behind the definite interactions of various molecules. A judicious combination of the well-known CADD methods like 3D-QSAR, molecular docking, pharmacophore modeling, ADMET prediction and data mining is employed to deliver important information, which are vital for lead/drug optimization. These methods developed valid models, leading to the identification of the unique structural patterns that govern the specific activity, and better perception regarding the mechanism of the molecular inhibitory actions of the drug candidates. These studies further explored the structure-based design methods so as to incorporate a better understanding of the ligand drug characters and drug–target interactions.

The usual approach for developing a QSAR model involves the generation of some valuable descriptors representing various molecular properties of the ligand. These may be physicochemical properties or topological indices that encrypt features like the properties of individual atoms and bonds. The present study used an in-house built new Pymol plug-in, of which 40% are new descriptors that are easily interpretable and can capture the local environment of the molecules (Masand and Rastija 2017). Based on the inputs from 2D QSAR, molecular docking, data mining using Weka, ADMET and pharmacophore studies, novel chemical entities (NCEs) were designed so as to satisfy the JAK inhibitory nature in a better way. The final QSAR model developed was utilized further to test these NCEs to know whether they possess a better inhibitory effect. The present study was thus able to successfully design several virtual molecules, and predict their corresponding inhibitory activities, using the selected descriptors present in the built model and thereby suggest a lead molecule.

Materials and methods

Dataset

The present study is based on datasets consisting of 19, 19 and 17 derivatives of 3,6-dihydroimidazo(4,5-d)pyrrolo(2,3-b)pyridin-2(1H)-one (dpp) and their corresponding IC₅₀ values related to the inhibition of JAK3, JAK2 and JAK1, respectively (Yamagishi et al. 2015). Though each of the datasets is small, it comprises substituted rings as substituents, thus covering a broad chemical space. The chemical structures of dpp derivatives were drawn and optimized by using the software, molecular operating environment (MOE). Merck molecular force field was used for the energy minimization and optimization of the compounds (Halgren 1996). The anti-inflammatory activities reported for various JAK inhibitors expressed in IC₅₀ and the structures of the corresponding dpp derivatives are given in Table 1.

Toxicity and drug likeness

A Java-based application called Toxtree v1.60 which functions on the basis of quantitative structure–toxicity relationships was used for toxicity prediction. In Toxtree (toxic hazard estimation by decision tree approach) (Patlewicz et al. 2008) the carcinogenicity and mutagenicity of the compound were analyzed using Benigni–Bossa rule base (Benigni and Bossa 2008). The drug-likeness of the compounds was carried out using MayaChemTools.

Calculation and selection of descriptors

Using MOE, PaDEL-descriptor and an in house built new Pymol plugin program, a total of 19737 descriptors, which included 2D-, 3D- molecular descriptors and fingerprints were generated for each of the energy-minimized molecules. The main key descriptor filtration algorithms, “General” and “CORCHOP” in the PHAKISO software were used for descriptor pruning. The descriptors which have the same values and missing values were removed using the “General”. The inter-correlated, correlated and the repeated descriptors were removed using the “CORCHOP” and thus the number of necessary descriptors was reduced to 180, 180 and 164, for the sets of JAK3, JAK2 and JAK1, respectively.

Many of the descriptors thus generated have wide ranging diverse numerical values. If no scaling is done, it becomes very difficult to understand and estimate the comparative contribution of those descriptors to the QSAR. Most often those descriptors with higher numerical values govern and influence the model, and essentially limit the statistical validity of the developed models. Hence scaling of descriptors is carried out in the present study using the software,

Table 1 Molecular structures of dpp derivatives and their activity values

Compound	R1	R2	IC ₅₀ (nM)		
			JAK3	JAK2	JAK1
dpp1		H	54	43	79
dpp2		H	180	97	110
*dpp3		H	22	23	30
dpp4		H	60	88	1000
dpp5		H	3.0	4.4	14
dpp6		H	21	23	430
dpp7		H	42	23	24
dpp8		H	69	85	90
dpp9		H	3.0	3.6	18
dpp10		H	270	500	520
*dpp11		H	1200	1300	-
dpp12		H	710	140	140
*dpp13		H	63	82	230
dpp14		H	28	17	20
*dpp15		H	1.1	2.6	1.5
dpp16		Me	5.0	7.3	29
dpp17		H	27	17	28
dpp18		H	460	850	-
dpp19	H	H	0.8	3.1	3.7

*Test Set Compounds

PHAKISO. The nearly constant and highly correlated descriptors were eliminated through subjective selection and the descriptors were further reduced to 12, 17 and 12, for JAK3, JAK2 and JAK1, respectively.

2D-QSAR

2D-QSAR studies were performed by means of the software, BuildQSAR using multiple linear regression (MLR) models for finding out the optimum number and the sets of necessary descriptors. For 2D-QSAR analysis, pIC_{50} ($-\log IC_{50}$) values were obtained by converting the corresponding IC_{50} (nM) values. This is used as dependent variable in the 2D-QSAR analysis. The selected descriptors were considered as independent variables and pIC_{50} values were considered as the dependent variable in all the three cases. The subjective feature selection was performed using Genetic Algorithm (GA) method, and the multiple linear regression (MLR) analysis method was used to construct the 2D-QSAR model of the 19 derivatives of 3,6-dihydroimidazo(4,5-d)pyrrolo(2,3-b)pyridin-2(1H)-one. The correlation coefficient (R), standard deviation (s), Fischer's test (F), level of confidence (p), squared cross correlation coefficient (Q^2), standard deviation error in prediction (S_{DEP}) and predicted residual sums of squares standard deviation (S_{PRESS}) were calculated for validating the QSAR-model.

3D-QSAR

Three-dimensional quantitative structure–activity relationship (3D-QSAR) methods such as comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) were performed for all the derivatives using the molecular modeling package SYBYL-X 1.3 (Tripos International, Missouri, USA). This helps to associate the bioactivity of compounds with structural descriptors and has been proved to be a very suitable method for accelerating the drug design process (Zhao et al. 2011).

3D-QSAR dataset preparation

For the development of thriving 3D-QSAR models, the whole set of 19 molecules was grouped randomly into a training set and a test set containing 15 and 4 molecules, respectively, in JAK3 and JAK2. The models were generated using the training set and the test set in turn was used to validate the models. Compounds 3, 11, 13 and 15 were used as the test set and the remaining were considered as the training set. But in the case of JAK1 for the development of 3D-QSAR models, only 17 molecules were grouped randomly into a training set and a test set containing 14 and 3 molecules, respectively, since the IC_{50} values of the compounds dpp11 and dpp18 were not reported.

Molecular alignment

Molecular alignment of compounds, which is a crucial step in the process of creating 3D-QSAR models, was performed using a common scaffold. A molecule with relatively high biological activity is usually adopted as the template molecule. In this work, compound 19 was used as the template molecule. Using ALLIGN DATAQBASE function available in SYBYL, the ligand-based alignment of the molecules was carried out on the basis of an atom-by-atom superimposition principle. The alignment of 19 compounds is shown in Fig. 1. These alignments were used subsequently in calculating the CoMFA/CoMSIA probe interaction energy-values.

The initial optimization was carried out using the standard TRIPOS force field (Clark et al. 1989) with 1000 iterations, with a distance-dependent dielectric and the Powell conjugate gradient algorithm with a convergence criterion of $0.01 \text{ kcal mol}^{-1} \text{ \AA}$. Gasteiger-Hückel method was used to assign the partial atomic charges of the compounds (Gasteiger and Marsili 1980).

CoMFA and CoMSIA studies

For the CoMFA analysis, steric (Lennard-Jones potentials) and electrostatic field (Coulombic potentials) energies were calculated using a sp^3 carbon with a van der Waals' radius of $1.52 \text{ \AA}$ as the steric probe; and a single positive charge (+ 1) as the electrostatic probe in 3D cubic lattice with a grid spacing of $2.0 \text{ \AA}$ in x, y and z directions, using Tripos module in SYBYL. The electrostatic and steric influences

Fig. 1 Alignment of 19 compounds

were reduced to + 30.0 kcal/mol. The electrostatic contributions were excluded at the lattice intersections with maximal steric interactions. For obtaining suitable results, the “StDev*coeff” (the standard deviation and the coefficient) values were utilized as various weighing factors, apart from the grid spacing. Crammer et al. suggested the impact of various parameter settings on CoMFA, electrostatic and steric cutoffs and grid spacing (Cramer et al. 1988).

It is reported that the CoMSIA method is based on the molecular similarity indices within the same lattice box (Klebe et al. 1994). In the present study, standard settings of CoMSIA, sp^3 carbon atom as a probe with charge + 1, radius 1 Å and hydrophobicity + 1, hydrogen-bond donor + 1, hydrogen-bond acceptor + 1, attenuation factor of 0.3 were used for each lattice with a grid of 2 Å. Mainly five similarity indices including that of steric (S), electrostatic (E), hydrophobic (H), hydrogen bond donor (D) and hydrogen bond acceptor (A) fields were computed.

3D-QSAR statistical results

Partial least square (PLS) analysis

Partial least-squares (PLS) (Bush and Nachbar 1993; Wold et al. 2001) method is regarded as an extension of the multiple regression analysis for deriving the 3D-QSAR models. Here the CoMFA and CoMSIA descriptors were taken as independent variables and pIC_{50} values were considered as dependent variables. The statistical importance of the models was analyzed by the cross-validation analysis performed by the leave-one-out (LOO) method to find the optimum number of components (N), which was later used to generate the final QSAR model. A column filter value of 2.0 kcal/mol was used to accelerate the regression analysis and to reduce the noise for all the cross-validated PLS analyses. To validate a QSAR model, different statistical parameters such as cross-validated correlation coefficient (Q^2), non-cross-validated correlation coefficient (R_{nev}^2), predicted correlation coefficient (R_{pred}^2), and standard error of estimate (SEE) were found out. The optimum number of components (OPN) was found out by considering the number of components that gave the smallest SEE and F values.

In CoMSIA, five different descriptor fields available are not totally independent of each other. This could decrease the statistical significance of the models (Bringmann and Rummey 2003) considered. We have generated 25 different models in order to build the optimal 3D-QSAR models with the highest Q^2 values and good statistical results for each class (Shibi et al. 2015). A model is said to be strong and predictive only when the statistical parameters $R_{nev}^2 > 0.6$,

$R_{pred}^2 > 0.5$, and $Q^2 > 0.6$ (Afantitis et al. 2009; Golbraikh and Tropsha 2002).

Predictive correlation coefficient (R_{pred}^2)

Biological activities of the test set molecules were predicted for the further validation of the model using the models derived from the training set. The Predictive Correlation Coefficient (R_{pred}^2) value was computed using the following equation:

$$R_{pred}^2 = 1 - \frac{\sum (Y_{predicted} - Y_{observed})^2}{\sum (Y_{observed} - Y_{mean})^2}, \quad (1)$$

where $Y_{predicted}$, $Y_{observed}$ and Y_{mean} are the predicted, the actual and the mean values of the activity, respectively. $\sum (Y_{predicted} - Y_{observed})^2$ is the predictive sum of squares (PRESS).

The predictive ability of the models developed was further confirmed by the Eq. 2 given below (Rannar et al. 1994):

$$R_{m(overall)}^2 = R^2 \times (1 - \sqrt{R^2 - R_0^2}) \quad (2)$$

where R^2 is the squared correlation coefficient between observed and predicted pEC_{50} values and R_0^2 is the squared correlation coefficient with intercept set to zero. The parameter determines whether the range of predicted activity values for the whole dataset is close to the observed activity or not. The value of $R_{m(overall)}^2$ should be greater than 0.5 for a satisfactory model.

Data mining using WEKA

Using a software called WEKA 3.7.3, an artificial intelligence technique was used for developing the model and for screening compounds to understand the anti-rheumatoid action of the derivatives. Anti-inflammatory Bioassays AID 1852 and AID 2001 were retrieved from PubChem BioAssay Database. The bioassay AID 1852 contains 2481 tested compounds, of which 1149 are reported to be active and the remaining 1332, inactive. Similarly, in AID set 2001, out of the 545 tested compounds, 263 are active and 282 are inactive.

Using PowerMV software, 155 descriptors, which included 147 pharmacophore finger prints and 8 various properties such as XLogP, PSA, number of rotatable bonds, H-bond acceptors, H-bond donors, molecular weight, Blood–Brain Barrier and Bad Group, were generated for the active and inactive compounds of both AIDs and were saved as CSV files (Liu et al. 2005). For creating binary classifiers for the molecules based on their bio-activity viz.,

actives and inactives, a Machine Learning (ML) method Weka (Waikato Environment for Knowledge Analysis) was used (Jamal et al. 2013; Gaba et al. 2014; Aswathy et al. 2016; Shibi et al. 2016).

Weka introduces base model through a confusion matrix. True Positives (TP) are the actives which are correctly classified as actives; False Positives (FP) represent inactives classified incorrectly as actives. Similarly True Negatives (TN) represents the inactives which are classified correctly as inactives and False Negatives (FN) are active compounds which are classified incorrectly as inactive (Vinita et al. 2011). The classification process involves building a classifier (model), which is a mathematical function that assigns class (e.g., active/inactive) labels to instances defined by a set of attributes (e.g. descriptors) (Frank et al. 2004).

Protein–ligand Molecular docking

In order to reduce the experimental costs of large-scale inhibitor screening and to increase the success rate, we designed a multi-target molecular docking system capable of predicting JAK inhibitor interactions. So we have selected a total of four protein molecules for each protein JAK1, JAK2 and JAK3 and subsequently performed molecular docking to calculate the binding affinities of JAK inhibitor pairs. The molecular docking studies of the selected 9 compounds with various JAK proteins viz. 3LXK, 3PJC 4HVG and 4RIO (JAK3); 3IO7, 3TJD, 4F08 and 4FVP (JAK2); 3EYH, 4E4N, 4EI4 and 4K77 (JAK1) were carried out using MOE 2014.0901. The default ‘Site Finder’ tool was used to identify the active site of the proteins.

ADME properties prediction

Adsorption, distribution, metabolism and excretion (ADME) were predicted using the online software PreADMET (<http://preadmet.bmdrc.org/>).

Pharmacophore elucidation

3D-pharmacophore model was developed by Pharmacophore Elucidation Query module of MOE. We have constructed the pharmacophore of the whole dataset and that of the three most active compounds 15, 19 and 5. Then it was compared with the drug molecule, Tofacitinib which has a similar scaffold.

Results and discussion

In silico toxicity studies and Lipinski’s rule of five (Ro5) filters

One of the most important reasons for the late stage failure in drug development is the possible toxicity of any active

molecule. The dpp derivatives selected for the present study were, therefore, screened using in silico tools for predicting the most relevant toxicity endpoints. By a Toxtree scan using Benigni/Bossa rule, we obtained 49 distinctive structural alerts (SAs). It was found that all the dpp derivatives are non-carcinogenic and non-mutagenic. The drug-likeness of the derivatives was then analyzed using Lipinski’s Ro5 which indicated that all the molecules followed Lipinski’s Ro5.

Prediction of metabolic behavior

The prediction of metabolic behavior also helps to understand the toxicity risks of the compounds. Vital ADMET properties such as maximal achievable drug concentration and drug toxicity have a great influence on the metabolic stability (Trunzer et al. 2009; Vasanathanathan et al. 2009). Like other ADMET properties, metabolism plays an important role in the failure of drugs to perform their proposed action in the human body (Guner and Bowen 2013). Sheridan et al. have also reported a QSAR based metabolic site prediction approach (Sheridan et al. 2007).

The liver is the main organ for human drug clearance. Cytochrome (CYP) enzymes contribute to the phase I metabolism of 70–80% of the drugs used currently (Pang 2009). Drugs entering the body generally undergo oxidation reactions catalyzed by CYP P450 enzymes. The most important isoforms of CYP enzymes are CYP3A4 (45%), CYP2C9 and CYP2C19 (25%), CYP2D6 (15%) and CYP1A2 (5% of current drugs) (Williams et al. 2004a, b). These CYP isoforms demonstrate typical inhibitor profiles and overlapping substrate specificities (Miners and Birkett 1998). The metabolite prediction of CYP3A4 substrates by MetaSite with 78% accuracy was reported by Zhou et al. (2006). Cytochrome CYP3A4 is the most abundant hepatic P450 isoform of the complex heme-containing enzyme responsible for the metabolism of more than 50% of the marketed drugs, in humans (Schlichting 2000). An enhanced knowledge regarding the molecular interactions between drugs and CYP3A4 are valuable perceptions for developing new medications. The shape and size of the active site cavity of the CYP3A4 structures are remarkable and the cavity is much larger near the heme iron (Scott and Halpert 2005). The large binding site cavities, reported in the literature allow this enzyme to accommodate substrates of various size (Williams et al. 2004a, b; Yano et al. 2004). And such large size would also indicate the capacity of the enzyme that can bind more than one substrate simultaneously (Kapelyukh et al. 2008; Korzekwa et al. 1998; Shou et al. 1994).

We used the molecular docking method to understand the metabolic behavior of the 19 compounds and analyzed the binding modes of compounds towards the cavity of the enzyme CYP3A4. The human microsomal cytochrome P450 3A4 protein structures, with PDB ID: 1TQN and 1W0E

were downloaded from Protein Data Bank and all the 19 dpp derivatives were successfully docked into the active site of these proteins. The relatively high docking score revealed that almost all the compounds bind tightly to CYP3A4 and thus can undergo metabolism easily. The high docking score is the representation of high metabolic behavior and represents the proximity of the substrate molecule to the iron atom in the heme (Prusis and Afzelius 2009).

2D-QSAR

A 2D-QSAR model is a mathematical equation that correlates the biological activity of a molecular system to its molecular descriptors. The 2D QSAR equations developed for the JAK3, JAK2 and JAK1 are, respectively as follows,

$$\begin{aligned} \text{pIC}_{50} = & +0.0983(\pm 0.0171) \text{ lipoplus\_AbSA} \\ & - 0.8984(\pm 0.2198) \text{ ExtFP264} \\ & - 2.7992(\pm 0.8027) \text{ PM3\_LUMO} \\ & - 1.2468(\pm 0.3202) \text{ MACCSFP91} \\ & - 0.2697(\pm 0.1024) \text{ com\_H\_3A} \\ & + 7.3986(\pm 0.0699), \end{aligned} \quad (3)$$

$$\begin{aligned} \text{pIC}_{50} = & +2.9878(\pm 1.2465) \text{ O\_lipo\_3Ac} \\ & + 2.7150(\pm 0.9546) \text{ Wnu1.polar} \\ & + 1.6867(\pm 0.9677) \text{ KRFPC232} \\ & + 0.0195(\pm 0.0090) \text{ SlogP\_VSA9} \\ & + 7.5542(\pm 0.1702), \end{aligned} \quad (4)$$

$$\begin{aligned} \text{pIC}_{50} = & +4.8195(\pm 1.2571) \text{ byring all\_H\_2Ac} \\ & + 2.7439(\pm 1.1214) \text{ vsurf\_EWmin2} \\ & + 0.2863(\pm 0.0525) \text{ vsurf\_W6} \\ & - 3.8536(\pm 1.6767) \text{ byring allplus\_sumpc} \\ & + 1.7154(\pm 0.5880) \text{ O\_lipo\_3Ac} \\ & + 0.2760 (\pm 0.1128) \text{ KRFPC3662} \\ & + 0.1811(\pm 0.1651) \text{ da\_H\_3A} \\ & + 7.2962(\pm 0.0884). \end{aligned} \quad (5)$$

The physicochemical descriptors such as lipoplus_AbSA, ExtFP264, PM3_LUMO, MACCSFP91 and com_H_3A are included in Eq. 3. O_lipo_3Ac, Wnu1.polar, KRFPC232 and SlogP_VSA9 in Eq. 4 and byring all_H_2Ac, vsurf_EWmin2, vsurf_W6, byring allplus_sumpc, O_lipo_3Ac, KRFPC3662 and da_H_3A are included in Eq. 5.

The meaning of descriptors are as follows:

- a. lipoplus_AbSA: absolute surface area of positively charged lipophilic atoms.
- b. ExtFP264: extended fingerprints with additional bits describing ring features.
- c. PM3_LUMO: the energy (eV) of the lowest unoccupied molecular orbital calculated using the PM3 Hamiltonian (Stewart 1993).
- d. MACCSFP91: is a molecular access system key fingerprint.
- e. com_H_3A: presence of H within 3 Å from center of mass of the molecule.
- f. O_lipo_3Ac: sum of charges of lipophilic atoms present within 3 Å from oxygen atom.
- g. Wnu1.polar: is a 3D directional WHIM descriptor, weighted by atomic polarizabilities. Todeschini and Gramatica (1998) described that WHIM are Holistic descriptors, which belongs to a class of hybrid descriptors.
- h. KRFPC232 (padel): SlogP_VSA9: Sum of v_i such that $Li > 0.40$. Is a subdivided surface area descriptor based on an approximate accessible van der Waals surface area (in Å²) calculation for each atom. Here Li denotes the contribution to logP(o/w) for atom i (Wildman and Crippen 1999).
- i. byring all_H_2Ac: sum of charges of H atoms present within 2 Angstrom from ring atoms.
- j. vsurf_EWmin2: Cruciani et al. (2000) reported that the vsurf_descriptors are similar to the VolSurf descriptors, for the pharmacokinetic property prediction and these descriptors to be very useful. vsurf_EWmin is one of the Lowest hydrophilic energy descriptors.
- k. vsurf_W6: hydrophilic volume descriptor.
- l. byring allplus_sumpc: sum of charges of positively charged atoms present in the ring.
- m. O_lipo_3Ac: sum of charges of oxygen atoms present within 3 Å from lipophilic atoms.
- n. KRFPC3662: Klekota–Roth fingerprint count which represents the count of chemical substructures (Klekota and Roth 2008).
- o. da_H_3A: presence of donor or acceptor within 3 Å from H atom.

Ponce et al. (2004) reported that the consistency of a 2D-QSAR model depends on both Q^2 and R^2 values and it should be high. More often, a value of $Q^2 > 0.5$ is considered satisfactory (Strazielle and Ghersi-Egea 2005).

As per the generated model, mainly three descriptors influence the pIC_{50} of 3,6-dihydroimidazo(4,5-d)pyrrolo(2,3-b)pyridin-2(1H)-one derivatives. For JAK3, JAK2, and JAK1, the model possesses high correlation coefficient values $R^2 = 0.981, 0.862$ and 0.974 , respectively and

low standard deviation values such as $s = 0.117$, 0.320 and 0.161 , respectively. This indicates the good capacity of the models to explain the observed values of biological activity. The low values of S_{PRESS} and S_{DEP} confirm the accurateness of the model. Further, evaluation of the degree of statistical significance was accomplished by the level of confidence ($p = 0$) and Fischer test (F). The validation parameters (Q^2 and S_{PRESS}), reflect the good predictive power of the generated model. Observed and predicted inhibitory activities and residual values of the statistically significant model obtained are shown in Table 2.

The correlation between experimental and calculated activity values graphically establishes the predictive capability of the developed QSAR models with R^2 values of 0.988 , 0.862 and 0.974 , respectively for JAK3, JAK2 and JAK1 (Fig. 2).

Table 2 Statistical parameters of 2D-QSAR

Statistical output	JAK3	JAK2	JAK1
n	15	17	17
k	5	4	7
R^2	0.988	0.862	0.974
$R^2\text{-Adj}$	0.981	0.816	0.954
s	0.117	0.320	0.161
F	146.322	18.744	48.124
p	0	0	0
Q^2	0.967	-0.048	0.896
S_{Press}	0.192	0.882	0.322
S_{DEP}	0.154	0.764	0.242

Fig. 2 Predicted inhibitory activities with experimental values

Y-randomization test

The Y-randomization test was performed using MLR YRandomization 1.2 software to validate and analyze the robustness of the developed 2D-QSAR models. The test identifies the chance of correlation between dependent and independent variables. Here the biological activity is randomized while keeping the other descriptors as constant, and fifty models were thus developed. A QSAR model is acceptable when the average correlation coefficient (R_p) of randomized models is less than the correlation coefficient (R) of the non-randomized model. Milano (2010) proposed that for succeeding Y-randomization test the value of ${}^cR_p^2$ should be more than 0.5 and reported that the magnitude of difference in the mean squared correlation coefficient values of the randomized (R_r^2) and non-randomized (R^2) models are reflected in the value of ${}^cR_p^2$ parameter.

$${}^cR_p^2 = R \times \sqrt{R^2 \times R_r^2} \quad (6)$$

All the fifty models developed for the three sets of biological activity have low R^2 and Q^2 values which signify the strength of the 2D-QSAR models built. The model parameters obtained in the present study are shown in Table 3. The ${}^cR_p^2$ values obtained are 0.747 , 0.704 and 0.722 for JAK3, JAK2 and JAK1, respectively, during the Y-randomization test and indicate the success of the generated models.

Applicability domain (AD)

The developed QSAR model is useful when it can be successfully employed to predict the activity of newly designed molecules. The estimation of a modeled response using QSAR is usable and valid when the compounds being predicted falls within the AD of the model. AD is a theoretical region in chemical space covering the model descriptors and modeled response. The uncertainty in the estimation of the activity of a compound can be thus eliminated.

The AD of the developed QSAR model was analyzed using AD Using Std Approach 1.0 software based on the basic theory of standardization approach. The methodology and algorithm suggested by Roy et al. (2015) were used to define the outliers in the case of the training set and to understand the compounds which are residing outside the AD. No outliers were identified in the training and the test set, which indicates that there are no chances of uncertainty in the prediction of QSAR models. This also demonstrates that the training set compounds are not dissimilar and the modelled 3D descriptors and the modelled response exist within the AD.

3D-QSAR

Statistical evaluation of CoMFA

The summary of the results from CoMFA models using LOO-CV constructed for dpp derivatives with steric and electrostatic fields are given in Table 4. PLS analyses of the training sets of JAK3, JAK2 and JAK1 show the Q^2 values 0.369, 0.476 and 0.490 using three, one, one as principal components, respectively. Böhm et al. (1999) reported that in CoMFA and CoMSIA studies, a Q^2 value of 0.3 is considered to be statistically significant while a Q^2 value of 0.4 is generally considered to be better. The non-cross-validated PLS analyses of these give high conventional R^2_{nev} values such as 0.863, 0.684 and 0.724 and low standard error of estimate (0.356, 0.434 and 0.374) respectively. The F values like 23.098, 28.139 and 31.438 indicate the high statistical significance of the model. The steric and electrostatic

field contributions calculated by the CoMFA models of JAK3 were 44.7 and 55.3%, for JAK2 56.3 and 43.7% and for JAK1 33.6 and 21.8%, respectively suggesting that the steric and electrostatic fields were found to be almost equally important to the binding affinities of the ligands to the target.

The CoMFA model of JAK3, JAK2 and JAK1 showed R^2_{pred} and $R^2_{\text{m(overall)}}$ values of 0.731 and 0.650, 0.757 and 0.404 and 0.465 and 0.327, respectively. The correlation between experimental and predicted pIC_{50} values by CoMFA models are represented in Fig. 3. It shows an R^2 value of 0.863 for the training set and 0.787 for the test molecules in JAK3. For JAK2, the training and test set value of R^2 is 0.712 and 0.939, respectively and for JAK1 the R^2 value is found to be 0.798 and 0.875, respectively. The experimental and predicted pIC_{50} values and the residual values of the dataset by CoMFA model are listed in Table 5.

CoMFA contour map analysis

The CoMFA contour maps of JAK1, JAK2 and JAK3 are shown in Fig. 4. The steric contributions were represented by green and yellow contours. The green contour represents the preferred steric bulk, which on adding a group near this region will improve the inhibitory activity. The yellow contour represents the disfavored steric bulk, adding a group near this region will decrease the inhibitory activity. The electrostatic contributions were represented by blue and red contours. Here the blue services the positive charge and dis-services the negative charge, that is, by adding a positively charged group near this region may improve the inhibitory activity and on adding a negatively charged group may weaken the activity. Similarly, the red areas in the electrostatic contours represent the negative charge favor and positive charge disfavor.

The electrostatic contribution of JAK3 is represented in Fig. 4a. The contour map shows a large blue contour near R1 group. This can be explained by considering the molecules dpp9 and dpp17, where insertion of ethyl group in dpp9 increased the inhibitory activity from a value of 7.658 (of compound dpp17) to 8.523. Comparing the molecules dpp1 and dpp8, it can be seen that the introduction of two methyl groups in dpp 8 increased the electrostatic factor resulting in an increase in inhibitory activity from 5.921 to 7.553.

The steric contributions of JAK3 are shown in Fig. 4b. The contour map shows a large green contour and two small yellow contours near the R1 group. The green contour can be explained by considering the molecules dpp9 and dpp5. Introduction of an ethyl group in the molecule dpp9 instead of the methyl group in dpp5 has increased the activity from 6.745 to 8.523. The molecule dpp9 shows an activity value of 8.523 when compared to the molecule dpp1 with an activity value of 5.921. This is because dpp9 has

Table 3 R^2 and Q^2 values after 50 Y-randomizations

Random models parameters	JAK3	JAK2	JAK1
Average R_r	0.515	0.451	0.663
Average R_r^2	0.287	0.221	0.456
Average Q_r^2	− 0.951	− 0.475	− 1.437
$^cR_p^2$	0.747	0.704	0.722

Table 4 Statistical results of CoMFA and the best CoMSIA models

	CoMFA			CoMSIA	
	JAK3	JAK2	JAK1	JAK3 (ADES)	JAK2 (ADH)
^a Q ² /ONC	0.369/3	0.476/1	0.490/1	0.462/3	0.512/2
^b R ² _{nev}	0.863	0.684	0.724	0.871	0.815
^c SEP	0.765	0.559	0.509	0.706	0.561
^d SEE	0.356	0.434	0.374	0.346	0.346
^e R ² _{pred}	0.731	0.757	0.465	0.757	0.768
R ² _{m(overall)}	0.650	0.404	0.327	0.668	0.641
F value	23.098	28.139	31.438	24.772	26.349
Field contribution					
Steric	44.7	56.3	55.2	13.7	—
Electrostatic	55.3	43.7	44.8	27.7	—
Hydrophobic	—	—	—	—	16.2
H-bond donor	—	—	—	38.4	42.2
H-bond acceptor	—	—	—	20.2	41.6

^a Cross-validated correlation coefficient^b Non cross-validated correlation coefficient^c Standard error of prediction^d Standard errors of estimate^e Predicted correlation coefficient for the test set **Fig. 3** Plot of actual versus predicted pIC_{50} values for the CoMFA model

Table 5 Experimental and predicted activities (pIC_{50}) with residual values of CoMFA models

Compound ID	JAK3			JAK2			JAK1		
	Actual pIC_{50}	Predicted	Residual	Actual pIC_{50}	Predicted	Residual	Actual pIC_{50}	Predicted	Residual
dpp1	7.268	7.078	0.190	7.367	7.419	− 0.053	7.102	7.398	− 0.296
dpp2	6.745	6.957	− 0.212	7.013	7.215	− 0.202	6.959	7.313	− 0.354
dpp3 ^a	7.658	7.620	0.038	7.638	7.533	0.105	7.523	7.481	0.042
dpp4	7.222	7.655	− 0.433	7.056	7.491	− 0.436	6.000	6.597	− 0.597
dpp5	8.523	8.330	0.193	8.357	8.068	0.289	7.854	7.569	0.285
dpp6	7.678	7.647	0.031	7.638	7.379	0.259	6.367	6.684	− 0.318
dpp7	7.377	7.464	− 0.087	7.638	7.434	0.204	7.620	7.506	0.114
dpp8	7.161	7.222	− 0.061	7.071	7.288	− 0.217	7.046	7.329	− 0.283
dpp9	8.523	7.959	0.564	8.444	7.468	0.976	7.745	7.462	0.283
dpp10	6.569	6.412	0.157	6.301	6.865	− 0.564	6.284	6.477	− 0.193
dpp11 ^a	5.921	5.884	0.037	5.886	6.246	− 0.360	—	—	—
dpp12	6.149	6.268	− 0.119	6.854	6.969	− 0.115	6.854	7.282	− 0.428
dpp13 ^a	7.201	7.566	− 0.365	7.086	6.949	0.137	6.638	7.163	− 0.525
dpp14	7.553	7.561	− 0.008	7.770	7.726	0.044	7.699	7.661	0.038
dpp15 ^a	8.959	7.893	1.066	8.585	7.697	0.888	8.824	7.603	1.221
dpp16	8.301	8.178	0.123	8.137	8.408	− 0.271	7.538	7.645	− 0.108
dpp17	7.569	8.235	− 0.666	7.770	8.068	− 0.298	7.553	7.543	0.010
dpp18	6.337	5.990	0.347	6.071	5.949	0.122	—	—	—
dpp19	9.097	8.726	0.371	8.509	7.946	0.563	8.432	7.585	0.847

^a Test set compounds

an ethyl as an additional group in R1, so it increases the bulkiness of the R1 group. This can also be explained by illustrating the activity of the molecules dpp8 and dpp1, where dpp8 has two methyl groups, and dpp1 has no methyl group ($7.553 > 5.921$). The relative contributions of steric and electrostatic fields in this model are 44.7 and 55.3%, respectively.

Figure 4c represents the steric contribution of JAK2. The contour map for steric shows two small green contours near R1 group. This can be explained by considering the molecules dpp16 and dpp17. The introduction of the bulky R1 group in dpp16 has increased the activity of the molecule when compared with the dpp17 molecule ($8.608 > 8.068$). Same is the case with the molecules dpp16 and dpp14. Introduction of a methyl group has increased the activity of the molecule dpp16 to 8.608 from an activity of 7.726. Comparing the molecules dpp 9 and dpp1, it can be seen that, insertion of an ethyl group has increased the activity from 7.419 to 7.468 for dpp9.

The electrostatic contribution of JAK2 is shown in Fig. 4d. The electrostatic contour map shows a large blue contour near R1 group. This means that increasing the positively charged group as R1 helps to increase the activity of the molecules. Molecule dpp8 shows increased activity when compared with dpp13. The molecule dpp8 has two

methyl groups in the R1 position as compared to an electron withdrawing bulky group in the R1 position of the molecule dpp13. Due to the introduction of ethyl group in dpp 9, there is an increase in the activity of the molecule dpp9 compared to that of dpp1. The relative contributions of steric and electrostatic fields in this model were 56.3 and 43.7%, respectively. In JAK1, the relative contributions of steric and electrostatic fields were 55.2 and 44.8%, respectively.

The electrostatic contour map of JAK1 shows a large blue contour near R1 group (Fig. 4e). Increasing the positive charge of the R1 group will increase the activity of the molecules. Molecule dpp1 shows increased activity when compared with dpp10, as dpp10 has a lone pair of electrons in groups in the R1 position.

The steric contribution of JAK1 is shown in Fig. 4f. The contour map for steric shows two small green contours near R1 group. This can be explained by considering the molecules dpp15 and dpp5. By the introduction of the bulky R1 group in dpp15 the activity of the molecule has increased (5.824), compared with that of dpp5 molecule (5.432). Same is the case with the molecules dpp8 and dpp6. Introduction of a methyl group has increased the activity of the molecule dpp8 from 3.366 to 4.045. Insertion of an ethyl group has increased the activity of dpp9 from 4.102 to 4.744 of dpp1.

Fig. 4 CoMFA contour maps of three different activity. Green fields represents the preferred steric bulk, yellow field represents the disfavored steric bulk, Blue fields indicate favored electropositive groups and red fields indicate favored electronegative groups

Statistical evaluation of CoMSIA

Both the hydrophobic and hydrogen bond donor descriptors, in addition to the steric and electrostatic fields in CoMFA, were defined using CoMSIA method that is not usually available with standard CoMFA. The field combinations were changed systematically to select the best result. Twenty-five different combinations for each set were generated in this study. Table 4 shows the summary of the statistical analysis of CoMSIA. The CoMSIA model with all the combination fields yielded a LOO cross-validated Q^2 value > 0.3 with different components, and the non-cross-validated $R^2 > 0.5$ and F value > 10 .

On comparing the statistical parameters of various models it is found that ADES i.e., hydrogen bond acceptor (A),

hydrogen bond donor (D), electrostatic (E) and steric (S) is comparatively good. In the JAK3 set a cross-validated Q^2 value 0.469, non-cross-validated R^2 0.872 with SEE 0.344 and F-value 25.066 are obtained. This model gives an overall R_m^2 0.668 and R_{pred}^2 0.757. Comparison of JAK2 models gives ADH, i.e., hydrogen bond acceptor (A), hydrogen bond donor (D) and hydrophobic (H) as a good model, which gives 0.512 and 0.815 as the cross-validated Q^2 and non-cross-validated R^2 values, respectively. The model gives an F value of 26.349, $R_{m(overall)}^2$ value of 0.687 and R_{pred}^2 value of 0.711. Since the statistical parameters of JAK1 model are not good, the CoMSIA contour analysis was carried out only for JAK2 and JAK3.

The plot of actual vs predicted activities for the training and test set molecules of both models are shown in Fig. 5. It shows an R^2 value of 0.872 for the training set and 0.813 for the test molecules for ADES, and an R^2 value of 0.815 for the training set and 0.773 for the test molecules for ADH. The experimental and predicted pIC_{50} values and the residual values of these dataset by both CoMSIA models are listed in Table 6.

CoMSIA contour map analysis

In the CoMSIA contour map of JAK3, the preferred steric contributions are represented by yellow, which on adding a group near this region will improve the inhibitory activity. The violet contour represents the disfavored steric, which on adding a group near this region will decrease the inhibitory activity. The contour map of this model is shown in the Fig. 6a.

The electrostatic contributions are represented by blue and green contours in Fig. 6b. Here the blue services the positive charge and disserves the negative charge. Similarly, the green areas in the electrostatic contours represent the negative charge favor and positive charge disfavor. The electrostatic contour map shows a large blue contour near R1 group. This means that increasing the positively charged group as R^2 will help to increase the activity of the molecules. Molecule dpp9 shows increased activity when compared with dpp13, as dpp9 has ethyl and methyl groups in the R1 group instead of $-N-CO-CH_2-CN$ group which is an electron rich group in the R1 group of the molecule dpp13. This can be explained by considering the molecules dpp5 and dpp1. The introduction of one more methyl group in dpp5 has increased the activity of the molecule to 6.745.

In Fig. 6c, the cyan contour indicates the hydrogen bond donor contour maps. The substituents in this region will increase the activity. The hydrogen bond donor substituents in this region leads to a decrease in activity as indicated by the purple coloured contour. The hydrogen bond donor contour map can be explained by comparing the molecules dpp11 and dpp15, where dpp11 has a $-NH$ group and hence shows a decrease in activity.

Table 6 Experimental and predicted activities (pIC_{50}) with residual values of CoMSIA Models

Compound ID	JAK3		JAK2	
	Predicted	Residual	Predicted	Residual
dpp1	7.515	− 0.247	7.561	− 0.195
dpp2	6.952	− 0.207	7.283	− 0.270
dpp3 ^a	7.640	0.018	7.619	0.019
dpp4	7.543	− 0.321	7.384	− 0.329
dpp5	8.104	0.419	7.876	0.481
dpp6	7.648	0.030	7.580	0.058
dpp7	7.506	− 0.129	7.618	0.020
dpp8	7.361	− 0.200	7.350	− 0.279
dpp9	7.763	0.760	7.663	0.781
dpp10	6.392	0.177	6.302	− 0.001
dpp11 ^a	5.850	0.071	5.804	0.082
dpp12	6.186	− 0.037	6.916	− 0.062
dpp13 ^a	7.516	− 0.315	7.930	− 0.844
dpp14	7.587	− 0.034	7.994	− 0.224
dpp15 ^a	7.937	1.022	8.150	0.435
dpp16	8.177	0.124	8.488	− 0.351
dpp17	8.023	− 0.454	7.843	− 0.073
dpp18	6.177	0.160	5.935	0.136
dpp19	9.137	− 0.040	8.201	0.308

^a Test set compounds

The magenta contour represents the favored H-bond acceptor contour map, which is shown in Fig. 6d. Here the hydrogen bond acceptor contour map can be explained by considering the molecules dpp18 and dpp11. Molecule dpp18 has a hydrogen bond acceptor group and thus it showed an increase in anti-inflammatory activity of 9.097, while molecule dpp11 showed a decreased inhibitory activity due to the presence of a methyl group. Same is the case with molecules dpp11 and dpp16.

In the CoMSIA contour map of JAK2, magenta contour represents the favored hydrogen bond acceptor contributions, suggesting that more bulky groups are favorable. The contour map of this model is shown in the Fig. 7a. The hydrogen bond acceptor contour map can be explained by considering the molecules dpp16 and dpp5. Molecule dpp16

Fig. 5 Plot of actual versus predicted pIC_{50} values for the CoMSIA model

Fig. 6 CoMSIA contour maps of JAK3 model. **a** Violet and yellow contour represents the favored and disfavored steric contributions; **b** the electropositive contributions were represented by blue contours; **c** cyan contour indicates the favored H-bond donor contributions; **d** magenta contour represents the favored H-bond acceptor contributions

has a hydrogen bond acceptor group and thus it showed an increase in anti-inflammatory activity of 8.608, while molecule dpp5 showed a decreased inhibitory activity. Same is the case with molecules dpp14 and dpp1 where the molecule dpp14 with a hydrogen bond acceptor group showed an increase in anti-inflammatory activity (7.726) when compared with the molecule dpp1 which has no hydrogen bond acceptor group (7.419).

In Fig. 7b the cyan contour indicates that hydrogen bond donor substituents in this region will improve the activity. The yellow region of the CoMSIA hydrophobic contour plot in Fig. 7c, reveals that the hydrophobic substituents in this region could enhance the inhibitory activity. The relative contributions of hydrogen bond acceptor (A), hydrogen bond donor (D), and hydrophobic (H), of this model were 42.2, 41.5 and 16.1%, respectively.

Data mining

The random forest (RF) algorithm gives high accuracy and time efficiency for predictive data modelling. Sajeev et al. (2013), Seal et al. (2012) and Periwal et al. (2011) had also reported that random forest provided the best classifier. Therefore in the present study, we tested the activity of the dpp derivatives using the classifier based on RF. Using the tenfold cross validation (CV), the RF classifier was evaluated.

In this study RF model corresponding to the AID 1852 gives accuracy and precision values of 61.7 and 0.616%,

respectively and with Kappa value 0.219, TP rate 0.617, FP rate 0.401, F measure value 0.611, and ROC area 0.657. Similarly, for the AID 2001 accuracy and precision values of 60.7 and 0.613%, respectively were obtained with Kappa value 0.219, TP rate 0.607, FP rate 0.387, F measure value 0.605, and ROC area 0.647.

With RF, 19 derivatives were screened using both AIDs and from the confusion matrix obtained, 9 molecules (dpp 1, 2, 4, 5, 6, 7, 17, 15 and 19) from the AID 1852 and 18 (except dpp 16) from AID 2001 were predicted as active by the model. On comparing these two results altogether, 9 molecules were found common to both sets. So these molecules (dpp 1, 2, 4, 5, 6, 7, 17, 15 and 19) were taken for further studies in the present work and the statistical parameters of the RF predictive model are shown in Table 7.

Molecular docking

The selected protein targets were validated by the primary structure and secondary structure analyses. The quality of each of the proteins was evaluated using Ramachandran plot obtained from MOE 2014.0901 (Supplementary material S1). On comparing the docking results with respect to a various set of JAK target proteins, 3LXK exhibited good E score for all the compounds in the JAK3 set; 3EYH showed the good docking score values for all the compounds in JAK1 set. But in the case of JAK2 the good docking score was obtained from the various proteins. That is, except dpp19, dpp17 and dpp15 all other compounds give a good result

Fig. 7 CoMSIA contour maps of JAK2 model. **a** magenta contour represents the favored H-bond acceptor contributions; **b** cyan contour indicates the H-bond donor contributions; **c** yellow region represents the favored hydrophobic contributions

Table 7 Statistics of random forest predictive model

AID SET	TP rate	FP rate	F measure	Kappa value	Precision	Accuracy	ROC area
AID 2001	0.607	0.387	0.605	0.219	0.613	60.73%	0.647
AID 1852	0.628	0.371	0.629	0.257	0.613	62.84%	0.684

TP rate true positive, *FP rate* false positive, *TN rate* true negative, *FN rate* false negative, *ROC area* receiver operating characteristic, *BCR* balanced classification rate

with 4F08; dpp19 and dpp15 give good score with the protein 4FVP; and dpp17 gives good E score value with the target protein 3IO7. The molecular docking results are shown in the Table 8. 2D interaction and docking pose of the active compounds with good docking score in the binding pocket of the proteins are shown in Fig. 8. Binding interaction studies of the compounds having comparatively good docking score with respective proteins are analyzed using LigPlot analysis.

Molecular docking and interaction of the selected compounds with 3LXK

The molecular docking studies of dpp15, dpp7 and dpp6 gave results with comparatively good docking score with 3LXK. Maximum docking score obtained is for the compound dpp15 and is -15.104 kcal/mol. This may be due to the large number of indirect hydrogen bonding interactions between the oxobutanenitrile group with various amino acid residues such as Arg953 and Lys830. The solvent exposure of this group is also very high. But there are no prominent interactions detected for the dpp15.

No prominent interactions are seen in the compound dpp7, but it shows an indirect hydrogen bonding interaction only with a basic amino acid residue Arg953. The carboxyl group is highly exposed to the solvent. The docking score of dpp7 is -12.937 kcal/mol.

The binding mode for the compound dpp6 indicates that the pyridine ring gives a backbone donor bonding interaction with an acidic residue Glu903 (2.78 Å, 20%). The unbroken dotted outline surrounding the ligand indicates the closeness of the ligand to the active site. If the ligand is close to the receptor, the line will be drawn very tightly around the ligand (Wallace et al. 1995). Here the side which contains the pyridine ring is very close to the active site. The docking score is found to be -13.659 kcal/mol.

Molecular docking and interaction of the selected compounds with JAK2 proteins

The molecular docking studies of dpp17, dpp15 and dpp5 gave results with comparatively good docking score with 4FVP. Compound dpp17 gave a comparatively good

Table 8 Molecular docking result with respect to the selected proteins

Compound	JAK3				JAK2				JAK1			
	E score, kcal/mol				E score, kcal/mol				E score, kcal/mol			
	3LXK	3PJC	4HVG	4RIO	3IO7	3TJD	4F08	4FVP	3EYH	4E4N	4E14	4K77
dpp1	-12.630	-8.893	-9.330	-9.586	-10.225	-9.607	-11.182	-9.954	-12.529	-11.443	-10.209	-11.118
dpp2	-12.174	-9.646	-9.270	-10.341	-10.515	-10.096	-10.822	-9.599	-12.489	-11.012	-10.317	-10.287
dpp4	-11.859	-9.231	-9.913	-9.869	-11.173	-9.727	-11.495	-10.295	-13.055	-12.207	-9.998	-10.419
dpp5	-12.767	-9.159	-9.431	-9.925	-10.513	-9.635	-12.659	-10.317	-12.304	-11.479	-10.441	-10.660
dpp6	-13.659	-9.233	-9.513	-9.537	-9.931	-9.770	-10.889	-10.129	-13.194	-11.480	-10.594	-10.721
dpp7	-12.937	-10.734	-9.965	-10.215	-10.164	-9.603	-11.139	-10.349	-13.257	-11.917	-10.188	-11.088
dpp15	-15.104	-12.223	-10.423	-10.797	-11.758	-12.010	-12.451	-13.435	-14.542	-13.056	-11.557	-12.973
dpp17	-12.787	-9.871	-9.828	-8.998	-12.411	-9.690	-12.256	-11.081	-12.816	-12.114	-10.655	-10.484
dpp19	-10.159	-8.667	-8.821	-8.740	-9.430	-8.379	-9.652	-10.012	-10.442	-10.118	-8.7632	-10.861

Fig. 8 2D interaction and docking pose of the active compounds with good docking score in the binding pocket of the proteins

molecular docking score – 12.411 kcal/mol with the protein 3IO7 and shows an arene–cation interaction between both the pyrrole and pyridine ring and a basic amino acid residue Arg980. The solvent exposure of dpp 17 is comparatively very good.

Molecular docking of dpp15 with docking score – 13.435 kcal/mol did not show any prominent interactions. But the carbonyl group present in the oxobutanenitrile side chain displays an indirect hydrogen bonding interaction with a polar amino acid residue Thr555 and with a basic residue Arg715. Here also the blue smudges appeared behind most

of the atoms indicate that the compound is highly exposed to the solvent.

The binding mode observed for dpp5 shows that the N atom of the pyridine ring shows a side chain acceptor bonding interaction with a basic amino acid residue Arg980 (2.26 Å, 10%). The carbonyl group present in the compound also shows a side chain acceptor interaction with a polar amino acid residue Asn895 (1.72 Å, 19%). The proximity contour shows the closeness of the compound in the active site. The molecular docking score is − 12.659 kcal/mol.

Molecular docking and interaction of the selected compounds with 3EYH

Compound dpp15 shows both prominent and indirect bonding interactions. The N atom of the pyridine ring shows both side chain and backbone acceptor bonding interaction with greasy and acidic amino acid residues Leu959 and Glu957 respectively at a distance of 1.89 Å (41%) and 2.99 Å (12%) with a docking score of − 14.542 kcal/mol. The carbonyl group present in the imidazolone ring shows indirect hydrogen bonding interactions with polar and acidic amino acid residues Gly1020 and Asp1021, respectively. Similarly the −CN group present in the oxobutanenitrile link of the compound gives an indirect bridging interaction with polar Gly882 and acidic Glu966 residues. The blue smudge around this −CN group clearly represents the good exposure to the solvent. The proximity contour shows the closeness of the dihydroimidazo(pyrrolo-pyridin-one) part of the compound in the active site.

No prominent interactions were observed for the compounds dpp7 and dpp6. In dpp7, the carbonyl group present in the imidazolone ring shows indirect hydrogen bonding interactions with polar Ser963 and acidic Glu966 amino acid residues. Similarly, the same carbonyl group present in dpp6 gives an indirect hydrogen bonding interaction with an acidic amino acid residue Glu966. The docking scores are − 13.257 and − 13.194 kcal/mol for dpp7 and dpp6, respectively.

The results show that our multi-target docking system accurately predicts the interactions between the molecules and JAKs and that the calculated binding affinities are highly correlated with the experimental values.

Prediction of ADME properties

ADME prediction of the finally selected three compounds was done using the preADMET tool. The relative ADME outlines of the selected three molecules are shown in Table 9. The in vivo blood–brain barrier penetration ($C_{\text{brain}}/C_{\text{blood}}$), in vitro Caco-2 cell permeability (nm/s), in vitro MDCK cell permeability (nm/s), CYP 3A4 inhibition,

human intestinal absorption (%) and in vitro plasma protein binding (%) details were obtained from this ADME calculation.

The Blood–Brain Barrier (BBB) penetration values help to know whether the compounds are able to pass across the BBB or not. A compound having BBB value > 2.0 is considered to have high absorption to CNS (Central Nervous System); with BBB value 2.0 ~ 0.1 are considered as middle absorption to CNS and with BBB value < 0.1 are to be considered as low absorption to CNS (Ma et al. 2005). The result shows that the compounds dpp5 and dpp17 have middle absorption to CNS and the compound dpp15 has very low absorption to CNS.

For evaluating the intestinal absorption of drug candidates, several in vitro methods have been used. Among them, for the prediction of oral drug absorption, Caco-2 cell permeability and MDCK cell model have been suggested as dependable in vitro models (Yamashita et al. 2000). A compound having Caco-2 less than 4 is considered as low permeable and the value between 4 and 70 is considered as middle permeable and which has more than 70 is considered as highly permeable compound (Irvine et al. 1999). Through in silico analysis, it is seen that all the compounds have moderate cellular permeability against Caco-2 cells. The Mandin Darby canine kidney (MDCK) cell permeability value of dpp5 and dpp17 are high among the selected compounds. The sum of absorption evaluated from the ratio of cumulative excretion and bioavailability are called human intestinal absorption (HIA) data. The HIA value for poorly absorbed compounds is between 0 and 20%. For a moderately absorbed compound, the HIA value is 20–70%, and for the well-absorbed compounds, the HIA value ranges from 70 to 100% (Yee 1997). In this study we obtained very good HIA values for all the three compounds. The plasma protein binding PPB of a drug influences the drug's action and disposition, and the efficacy of a drug candidate. For a strongly bound chemical, the PPB value will be > 90% and

Table 9 ADME properties obtained from PreADMET server

Properties	dpp5	dpp15	dpp17
BBB ^a	1.565	0.026	1.565
Caco2 ^b	15.411	6.433	15.411
CYP_3A4_inhibition	Inhibitor	Inhibitor	Inhibitor
HIA ^c	90.223	89.870	90.223
MDCK ^d	47.935	1.277	47.935
Plasma_protein_binding	83.875	43.439	83.875

^a Blood–brain barrier

^b Caco-2 cell permeability

^c Human intestinal absorption

^d Madin–Darby canine kidney cells

in the present study we obtained PPB value $< 90\%$ for all the three compounds, which means they are weakly bound.

Comparative study of molecular docking score with IC_{50}

In the case of JAK3 and JAK1 sets except dpp19, all other compounds have comparatively good and high docking score. The IC_{50} values were also good for the compounds except dpp1, dpp2, dpp4 and dpp7 in JAK3. In JAK1, the IC_{50} values of dpp1, dpp2, dpp4 and dpp6 are comparatively high. In JAK2, only dpp17, dpp15 and dpp5 show good docking score values. The IC_{50} values of these three compounds are also good (Table 10). Further comparison of the docking score and IC_{50} values among the three sets of JAK proteins, shows that dpp17, dpp15 and dpp5 are comparatively good. Therefore, the three compounds, dpp17, dpp15 and dpp5 are considered as the lead compounds for the development of new anti-RA drugs.

Pharmacophore elucidation

The 3D pharmacophore designing methods take into account both the three-dimensional structures and binding modes of receptors and inhibitors. This is to understand whether the regions are favorable or not for a specific receptor–inhibitor interaction. The pharmacophore model generated for the dataset showed two aromatic, two H-bond donor, two H-bond acceptor and one pi ring centers features, which are shown in Fig. 9. The pharmacophore developed for the drug molecule Tofacitinib also showed similar pharmacophoric features such as two aromatic, four H-bond acceptors, one H-bond donor, and one hydrophobic center. This reveals that the compounds can be used as excellent lead compounds for the treatment of RA.

Design of novel chemical entities

The models obtained through the various studies could be effectively used as basic strategy to modify the most preferable scaffold with a view to design some novel potent JAK

Table 10 Comparison of Molecular docking score with IC_{50} values of the active compounds

Compounds	IC_{50}			E score (kcal/mol)		
	JAK3	JAK2	JAK1	JAK3	JAK2	JAK1
dpp1	54	43	79	− 12.630	− 11.182	− 12.529
dpp2	180	97	110	− 12.174	− 10.822	− 12.489
dpp4	60	88	1000	− 11.859	− 11.495	− 13.055
dpp5	3	4.4	14	− 12.767	− 12.659	− 12.304
dpp6	21	23	430	− 13.659	− 10.889	− 13.194
dpp7	42	23	24	− 12.937	− 11.140	− 13.257
dpp15	1.1	2.6	1.5	− 15.104	− 13.435	− 14.542
dpp17	27	17	28	− 12.787	− 12.411	− 12.816
dpp19	0.8	3.1	3.7	− 10.159	− 10.012	− 10.442

Fig. 9 Pharmacophore mapping for the dataset (cyan: H-bond acceptor, magenta: H-bond donor, orange: Pi ring center). **a** Pharmacophore mapping of the most active compounds dpp19, dpp15 and dpp5; **b** pharmacophore mapping of the drug molecule Tofacitinib

inhibiting compounds. Building a QSAR model and validating the same makes it useful for predicting the activity of newly designed chemical entities whose activities depend on their molecular properties represented by the descriptors. Following the chemical information gathered through the in silico studies performed, the most dependable compound dpp 15 was further modified in order to improve its activity. By doing so, the underlying principles which form the basis for enhancing the biological activities obtained through various QSAR studies and docking-based scoring were utilized. CoMFA and CoMSIA contour maps derived from dpp derivatives provide valuable information regarding how to design novel molecules with improved anti-RA activity.

In CoMFA and CoMSIA contour maps, the existence of a blue contour near R1 position indicates that bulky electropositive groups are favorable for increasing the anti-RA activity. In designing novel chemical entities of the selected scaffold of dpp15, we have incorporated the structural features of the compound dpp15 since the compound has been concluded as to have exhibited potent JAK3, JAK2 and JAK1 inhibitory activities. Based on this, we introduced bulky electropositive groups at the R1 position of dpp 15 and created 20 virtual molecules, which are shown in Supplementary Material Fig. S2.

For evaluating and screening the new virtual molecules with respect to the anti-RA action, the descriptors were generated using PowerMV software and virtual screening was performed using WEKA. The developed RF model predicted that all the 20 newly modified compounds are anti-RA active. The toxicity predictions of these compounds performed using Toxtree, also suggested that all the generated compounds are non-carcinogenic and non-mutagenic in nature. The drug-likeness of these compounds was checked using Lipinski's rule of five; the result showed that all the molecules obeyed Ro5.

The ADME properties of these compounds were evaluated, and were found to be very good and are included in Table 11. Molecular docking with respect to the human microsomal cytochrome P450 3A4 proteins 1TQN, 1W0E, 1W0G, 4I4H and 4NY4 were performed to understand the metabolic capability of these compounds. The relatively high molecular docking score (Table 12) of these compounds revealed that all the compounds were well docked into the active site of these proteins, which means these compounds are easily metabolized.

Using SMARTCyp, the CYP3A4 sites of metabolism of these compounds were predicted. The results obtained are shown in Table 13, which indicates that all the compounds show three common CYP3A4 sites of metabolism.

To find out the activity of the modified molecules, the molecular descriptors of these 20 active virtual molecules were calculated using MOE and PaDEL-Descriptor

Table 11 ADME properties obtained for modified compounds

Compounds	BBB ^a	Caco2 ^b	HIA ^c	MDCK ^d	Plasma_protein_binding
dpp15(a)	0.049	10.451	91.171	8.818	69.772
dpp15(b)	0.095	12.536	91.775	7.750	81.633
dpp15(c)	0.120	13.635	92.056	0.977	83.238
dpp15(d)	0.249	14.892	92.610	0.092	86.571
dpp15(e)	0.232	19.211	92.878	0.062	87.216
dpp15(f)	0.044	16.261	91.170	5.269	65.733
dpp15(g)	0.050	14.806	91.489	8.720	74.012
dpp15(h)	0.064	12.542	91.490	13.056	76.498
dpp15(i)	0.189	16.349	92.607	0.061	85.481
dpp15(j)	0.166	14.743	92.334	0.134	84.916
dpp15(k)	0.049	9.764	90.880	14.011	66.925
dpp15(l)	0.072	9.894	91.180	10.931	75.091
dpp15(m)	0.096	11.058	91.471	1.092	78.345
dpp15(n)	0.057	10.091	91.179	10.196	73.578
dpp15(o)	0.049	13.596	91.178	10.955	70.603
dpp15(p)	0.736	17.023	93.146	0.052	88.537
dpp15(q)	0.440	18.810	93.145	0.058	88.177
dpp15(r)	0.046	8.102	90.880	9.583	66.850
dpp15(s)	0.031	4.989	90.234	1.868	47.312
dpp15(t)	0.035	6.222	90.566	2.818	52.490

^a Blood–brain barrier

^b Caco-2 cell permeability

^c Human intestinal absorption

^d Madin–Darby canine kidney cells

software. Using the developed 2D-QSAR equations and the respective descriptors involved in the equations, we generated the pIC₅₀ value of these molecules. The descriptors used to find out the anti-RA activity and the resulted pIC₅₀ values of these compounds are listed in Table 14. The predicted molecular parameters of the modified compounds are listed in this Table. Seven novel chemical entities showed greater JAK inhibition when compared to the parent dpp derivatives. The modified compounds such as dpp15(d), dpp15(m), dpp15(p), dpp15(q) and dpp15(r) showed greater inhibition with respect to JAK3. Compounds dpp15(e) and dpp15(i) showed better inhibition of both JAK3 and JAK2 compared to the activity of any of the parent molecules. The compound dpp15(m) showed higher inhibitory activity for JAK2. Thus it is seen that the molecule, dpp15(s) has much better inhibitory activity values when compared to other novel chemical entities with JAK3 and JAK1. The results of the present study thus confirm the predictive ability of the structure–activity relationships and docking studies. The designed molecules testified by us impart leads for forthcoming research.

Table 12 Molecular docking result of modified compounds with respect to CYP3A4 Proteins

Compounds	E score (kcal/mol)				
	1TQN	1W0E	1W0G	4I4H	4NY4
dpp15(a)	-14.795	-12.792	-11.624	-11.356	-11.263
dpp15(b)	-13.848	-12.538	-11.398	-11.450	-11.128
dpp15(c)	-12.555	-12.275	-11.300	-10.628	-11.222
dpp15(d)	-12.612	-12.993	-12.119	-11.138	-10.581
dpp15(e)	-12.938	-13.037	-12.284	-12.936	-11.540
dpp15(f)	-12.944	-11.593	-10.881	-10.108	-10.762
dpp15(g)	-13.180	-12.492	-10.842	-10.949	-11.713
dpp15(h)	-12.561	-11.752	-11.119	-10.391	-12.617
dpp15(i)	-13.570	-12.444	-11.151	-12.933	-11.278
dpp 15(j)	-12.719	-11.129	-11.378	-11.746	-10.502
dpp15(k)	-14.045	-11.982	-11.390	-10.344	-11.127
dpp 15(l)	-13.272	-12.674	-10.481	-10.045	-11.878
dpp15(m)	-12.408	-11.805	-10.191	-10.965	-10.073
dpp15(n)	-12.365	-11.326	-10.507	-10.834	-10.863
dpp15(o)	-12.522	-12.149	-10.455	-10.506	-10.742
dpp15(p)	-11.991	-11.951	-11.871	-11.494	-11.353
dpp15(q)	-13.868	-11.739	-11.700	-11.222	-11.949
dpp15(r)	-13.235	-11.939	-10.290	-10.081	-10.377
dpp15(s)	-15.416	-11.803	-11.292	-10.511	-10.382
dpp15(t)	-13.415	-11.767	-10.626	-11.793	-10.873

Conclusion

During the first stage of this study, the toxicity analysis showed that all the derivatives selected are non-carcinogenic and mutagenic in nature. All the compounds obey Lipinski rule of five. The relatively high docking score with respect to the enzyme CYP3A4 revealed that most of the derivatives metabolize easily. Genetic algorithm and MLR analysis method were used to construct the 2D-QSAR model of the 19 derivatives and the model was validated by the Y-randomization method. 3D-QSAR studies such as CoMFA and CoMSIA analysis revealed various structural features affecting the anti-RA activity of the compounds. From the CoMFA and CoMSIA field contributions we confirmed the importance of hydrogen bond donor, acceptor, hydrophobic, steric and electrostatic interactions. A machine learning technique was able to successfully filter out the active compounds computationally using the Random forest classifier in the Weka. The molecular docking studies of these selected compounds with various JAK proteins such as 3LXK, 3PJC 4HVG and 4RIO (JAK3); 3IO7, 3TJD, 4F08 and 4FVP (JAK2); 3EYH, 4E4N, 4EI4 and 4K77 (JAK1) indicated that all the compounds in the JAK3 set exhibited good E score for 3LXK, and all the compounds in JAK2 set showed good docking score values for 3EYH. The various binding interactions between the proteins and the corresponding

Table 13 SMARTcyp result of the modified compounds

	Atom	Score	Energy	*2DSASA
	C.8	43.41	52.9	37.15
	C.9	60.94	69.4	29.64
	N.3	66.84	75.6	18.99
This is common for dpp 15(a), dpp 15(c), dpp 15(d), dpp 15(e), dpp 15(f), dpp 15(g), dpp 15(i), dpp 15(j), dpp 15(k), dpp 15(l), dpp 15(m), dpp 15(n), dpp 15(o), dpp 15(p), dpp 15(q), dpp 15(r), dpp 15(s), dpp 15(t)				
	C.8	43.41	52.9	37.15
	C.9	60.94	69.4	29.64
	C.31	66.39	75.9	37.72
dpp 15(b)				
	C.8	69.4	52.9	37.15
	C.9	60.94	69.4	29.64
	C.29	66.34	75.9	38.92
dpp 15(h)				

*Solvent Accessible Surface Area

Table 14 Predicted anti-RA activity of modified compounds with respect to the 2D-QSAR equations

Modified compounds	Predicted pIC ₅₀ values		
	JAK3	JAK2	JAK1
dpp15(a)	8.482	8.338	8.091
dpp15(b)	8.955	7.621	6.971
dpp15 (c)	8.836	8.448	7.547
dpp15(d)	9.115	7.883	7.369
dpp15(e)	9.136	8.806	6.879
dpp15(f)	6.901	8.334	8.187
dpp15 (g)	7.379	6.916	6.381
dpp15(h)	7.914	8.030	7.146
dpp15(i)	9.016	9.178	7.298
dpp15 (j)	8.779	7.990	7.408
dpp15(k)	7.275	8.304	7.816
dpp15 (l)	8.340	8.389	8.129
dpp15(m)	7.967	8.668	8.550
dpp15 (n)	8.654	8.188	8.285
dpp15(o)	7.228	8.051	7.734
dpp15(p)	10.588	7.642	6.999
dpp15(q)	9.795	7.775	6.683
dpp15(r)	9.056	7.960	7.479
dpp15(s)	15.444	8.518	9.018
dpp15(t)	7.936	8.225	8.134

ligands were also understood in detail. Further comparison of the docking score and IC₅₀ values among the three sets of JAK proteins, and pharmacological properties showed that dpp15 is the best among the molecules studied. So this compound was selected as the basic scaffold for the design of novel chemical entities. The biological activities of all the 20 new molecules were found out using the already developed 2D-QSAR equation. Compound dpp15(s), 3-((3S,4R)-2,4-dimethyl-3-(2-oxo-3,6-dihydroimidazo(4,5-d)pyrrolo(2,3-b)pyridin-1(2H)-yl)piperidin-1-yl)-3-oxopropanenitrile showed maximum inhibitory activity potential against Janus kinase and is selected as the lead molecule.

Acknowledgements Jisha, R.S. is thankful to the University of Kerala, Thiruvananthapuram for providing financial assistance in the form of University Junior Research Fellowship for this work. Aswathy L. is thankful to CSIR, New Delhi for the financial assistance in the form of Senior Research Fellowship.

References

- Afantitis A, Melagraki G, Sarimveis H et al (2009) A novel QSAR model for predicting the inhibition of CXCR3 receptor by 4-N-aryl-1-[1,4] diazepane ureas. *Eur J Med Chem* 44:877–884. doi:[10.1016/j.ejmech.2008.05.028](https://doi.org/10.1016/j.ejmech.2008.05.028)
- Aswathy L, Jisha RS, Masand VH et al (2016) Computational strategies to explore antimalarial thiazine alkaloid lead compounds based on an Australian marine sponge *Plakortis lita*. *J Biomol Struct Dyn*. doi:[10.1080/07391102.2016.1220870](https://doi.org/10.1080/07391102.2016.1220870)
- Benigni R, Bossa C (2008) Structure alerts for carcinogenicity, and the *Salmonella* assay system: a novel insight through the chemical relational databases technology. *Mutat Res* 659:248–261. doi:[10.1016/j.mrrev.2008.05.003](https://doi.org/10.1016/j.mrrev.2008.05.003)
- Boers M, Verhoeven AC, Markusse HM et al (1997) Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. *Lancet* 350:309–318. doi:[10.1016/S0140-6736\(97\)01300-7](https://doi.org/10.1016/S0140-6736(97)01300-7)
- Böhm M, St Rzebecher J, Klebe G (1999) 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* 42:458–477. doi:[10.1021/jm981062r](https://doi.org/10.1021/jm981062r)
- Bringmann G, Rummey C (2003) 3D QSAR investigations on anti-malarial naphthylisoquinoline alkaloids by comparative molecular similarity indices analysis (CoMSIA), based on different alignment approaches. *J Chem Inf Comput Sci* 43:304–316. doi:[10.1021/ci025570s](https://doi.org/10.1021/ci025570s)
- Bush BL, Nachbar RB (1993) Sample-distance partial least squares: PLS optimized for many variables, with application to CoMFA. *J Comput Aided Mol Des* 7:587–619. doi:[10.1007/BF00124364](https://doi.org/10.1007/BF00124364)
- Clark M, Cramer RD, Van Opdenbosch N (1989) Validation of the general purpose tripos 5.2 force field. *J Comput Chem* 10:982–1012. doi:[10.1002/jcc.540100804](https://doi.org/10.1002/jcc.540100804)
- Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* 110:5959–5967. doi:[10.1021/ja00226a005](https://doi.org/10.1021/ja00226a005)
- Cruciani G, Crivori P, Carrupt P-A, Testa B (2000) Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. *J Mol Struct THEOCHEM* 503:17–30. doi:[10.1016/S0166-1280\(99\)00360-7](https://doi.org/10.1016/S0166-1280(99)00360-7)
- Frank E, Hall M, Trigg L et al (2004) Data mining in bioinformatics using Weka. *Bioinformatics* 20:2479–2481. doi:[10.1093/bioinformatics/bth261](https://doi.org/10.1093/bioinformatics/bth261)
- Fridman JS, Scherle PA, Collins R et al (2010) Selective inhibition of JAK1 and JAK2 is efficacious in rodent models of arthritis: pre-clinical characterization of INCB028050. *J Immunol* 184:5298–5307. doi:[10.4049/jimmunol.0902819](https://doi.org/10.4049/jimmunol.0902819)
- Fujii M, Adachi S, Shimizu T et al (1993) Interstitial lung disease in rheumatoid arthritis: assessment with high-resolution computed tomography. *J Thorac Imaging* 8:54–62
- Gaba S, Jamal S, Drug Discovery Consortium OS, Scaria V (2014) Cheminformatics models for inhibitors of *Schistosoma mansoni* thioredoxin glutathione reductase. *Sci World J* 2014:1–9. doi:[10.1155/2014/957107](https://doi.org/10.1155/2014/957107)
- Gabbay E, Tarala R, Will R et al (1997) Interstitial lung disease in recent onset rheumatoid arthritis. *Am J Respir Crit Care Med* 156:528–535. doi:[10.1164/ajrccm.156.2.9609016](https://doi.org/10.1164/ajrccm.156.2.9609016)
- Gasteiger J, Marsili M (1980) Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* 36:3219–3228. doi:[10.1016/0040-4020\(80\)801682](https://doi.org/10.1016/0040-4020(80)801682)
- Golbraikh A, Tropsha A (2002) Beware of Q₂. *J Mol Graph Model* 20:269–276. doi:[10.1016/S1093-3263\(01\)00123-1](https://doi.org/10.1016/S1093-3263(01)00123-1)
- Gonzalez A, Maradit Kremers H, Crowson CS et al (2007) The widening mortality gap between rheumatoid arthritis patients and the general population. *Arthritis Rheum* 56:3583–3587. doi:[10.1002/art.22979](https://doi.org/10.1002/art.22979)
- Guner OF, Bowen JP (2013) Pharmacophore modeling for ADME. *Curr Top Med Chem* 13:1327–1342. doi:[10.2174/15680266113139990037](https://doi.org/10.2174/15680266113139990037)

- Halgren TA (1996) Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *J Comput Chem* 17:553–586. doi:[10.1002/\(SICI\)1096-987X\(199604\)17:5/6<553::AID-JCC3>3.0.CO;2-T](https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<553::AID-JCC3>3.0.CO;2-T)
- Imada K, Leonard WJ (2000) The Jak-STAT pathway. *Mol Immunol* 37(1–2):1–11. doi:[10.1016/S0161-5890\(00\)00018-3](https://doi.org/10.1016/S0161-5890(00)00018-3)
- Irvine JD, Takahashi L, Lockhart K et al (1999) MDCK (Madin–Darby canine kidney) cells: a tool for membrane permeability screening. *J Pharm Sci* 88:28–33. doi:[10.1021/js9803205](https://doi.org/10.1021/js9803205)
- Jacobsson LT, Hanson RL, Knowler WC et al (1994) Decreasing incidence and prevalence of rheumatoid arthritis in Pima Indians over a twenty-five-year period. *Arthritis Rheum* 37:1158–1165. doi:[10.1002/art.1780370808](https://doi.org/10.1002/art.1780370808)
- Jamal S, Periwal V, Open Source Drug Discovery Consortium, Scaria V (2013) Predictive modeling of anti-malarial molecules inhibiting apicoplast formation. *BMC Bioinform* 14:55. doi:[10.1186/1471-2105-14-55](https://doi.org/10.1186/1471-2105-14-55)
- Kapelyukh Y, Paine MJI, Maréchal J-D et al (2008) Multiple substrate binding by cytochrome P450 3A4: estimation of the number of bound substrate molecules. *Drug Metab Dispos* 36:2136–2144. doi:[10.1124/dmd.108.021733](https://doi.org/10.1124/dmd.108.021733)
- Kapetanovic IM (2008) Computer-aided drug discovery and development (CADD): in silico-chemico-biological approach. *Chem Biol Interact* 171:165–176. doi:[10.1016/j.cbi.2006.12.006](https://doi.org/10.1016/j.cbi.2006.12.006)
- Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J Med Chem* 37:4130–4146. doi:[10.1021/jm00050a010](https://doi.org/10.1021/jm00050a010)
- Klekota J, Roth FP (2008) Chemical substructures that enrich for biological activity. *Bioinformatics* 24:2518–2525. doi:[10.1093/bioinformatics/btn479](https://doi.org/10.1093/bioinformatics/btn479)
- Korzekwa KR, Krishnamachary N, Shou M et al (1998) Evaluation of atypical cytochrome P450 kinetics with two-substrate models: evidence that multiple substrates can simultaneously bind to cytochrome P450 active sites. *Biochemistry* 37:4137–4147. doi:[10.1021/bi9715627](https://doi.org/10.1021/bi9715627)
- Kroot EJ, de Jong BA, van Leeuwen MA et al (2000) The prognostic value of anti-cyclic citrullinated peptide antibody in patients with recent-onset rheumatoid arthritis. *Arthritis Rheum* 43:1831–1835. doi:[10.1002/1529-0131\(200008\)43:8<1831::AID-ANR19>3.0.CO;2-6](https://doi.org/10.1002/1529-0131(200008)43:8<1831::AID-ANR19>3.0.CO;2-6)
- Liu K, Feng J, Young SS (2005) PowerMV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation. *J Chem Inf Model* 45:515–522. doi:[10.1021/ci049847v](https://doi.org/10.1021/ci049847v)
- Ma X, Chen C, Yang J (2005) Predictive model of blood–brain barrier penetration of organic compounds. *Acta Pharmacol Sin* 26:500–512. doi:[10.1111/j.1745-7254.2005.00068.x](https://doi.org/10.1111/j.1745-7254.2005.00068.x)
- Masand VH, Rastija V (2017) PyDescriptor: a new PyMOL plugin for calculating thousands of easily understandable molecular descriptors. *Chemom Intell Lab Syst* 169:12–18. doi:[10.1016/j.chemolab.2017.08.003](https://doi.org/10.1016/j.chemolab.2017.08.003)
- Milano TR (2010) Chemometrics and QSAR Research Group University of Milano-Bicocca Milano Italy (personal communication)
- Miners JO, Birkett DJ (1998) Cytochrome P4502C9: an enzyme of major importance in human drug metabolism. *Br J Clin Pharmacol* 45:525–538. doi:[10.1046/j.1365-2125.1998.00721.x](https://doi.org/10.1046/j.1365-2125.1998.00721.x)
- Orbach H, Gilburd B, Brickman CM et al (2002) Anti-cyclic citrullinated peptide antibodies as a diagnostic test for rheumatoid arthritis and predictor of an erosive disease. *Isr Med Assoc J* 4:892–893
- Pang KS (2009) Safety testing of metabolites: expectations and outcomes. *Chem Biol Interact* 179:45–59. doi:[10.1016/j.cbi.2008.09.013](https://doi.org/10.1016/j.cbi.2008.09.013)
- Patlewicz G, Jeliakova N, Safford RJ, Worth AP, Aleksiev B (2008) An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *SAR QSAR Environ Res* 19(5–6):495–524. doi:[10.1080/10629360802083871](https://doi.org/10.1080/10629360802083871)
- Periwal V, Rajappan JK, Open Source Drug Discovery Consortium et al (2011) Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets. *BMC Res Notes* 4:504. doi:[10.1186/1756-0500-4-504](https://doi.org/10.1186/1756-0500-4-504)
- Plenge RM (2009) Rheumatoid arthritis genetics: 2009 update. *Curr Rheumatol Rep* 11:351–356. doi:[10.1007/s11926-009-0050-0](https://doi.org/10.1007/s11926-009-0050-0)
- Ponce YM, Garit JAC, Torrens F et al (2004) Atom, atom-type, and total linear indices of the “molecular pseudograph’s atom adjacency matrix”: application to QSPR/QSAR studies of organic compounds. *Molecules* 9:1100–1123. doi:[10.3390/91201100](https://doi.org/10.3390/91201100)
- Prusis P, Afzelius L (2009) Improvement of site of metabolism predictions for CYP3A4 by using discriminant analysis of compound preference of CYP3A4 X-ray structural conformers and subsequent docking. *QSAR Comb Sci* 28:865–868. doi:[10.1002/qsar.200810182](https://doi.org/10.1002/qsar.200810182)
- Rannar S, Lindgren F, Geladi P, Wold S (1994) A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: theory and algorithm. *J Chemom* 8:111–125. doi:[10.1002/cem.1180080204](https://doi.org/10.1002/cem.1180080204)
- Riise T, Jacobsen BK, Gran JT (2000) Incidence and prevalence of rheumatoid arthritis in the county of Troms, northern Norway. *J Rheumatol* 27:1386–1389
- Roy K, Kar S, Ambure P (2015) On a simple approach for determining applicability domain of QSAR models. *Chemom Intell Lab Syst* 145:22–29. doi:[10.1016/j.chemolab.2015.04.013](https://doi.org/10.1016/j.chemolab.2015.04.013)
- Sajeev R, Athira RS, Nufail M et al (2013) Computational predictive models for organic semiconductors. *J Comput Electron* 12:790–795. doi:[10.1007/s10825-013-0486-3](https://doi.org/10.1007/s10825-013-0486-3)
- Schlichting I (2000) The catalytic pathway of cytochrome P450cam at atomic resolution. *Science* 287:1615–1622. doi:[10.1126/science.287.5458.1615](https://doi.org/10.1126/science.287.5458.1615)
- Scott EE, Halpert JR (2005) Structures of cytochrome P450 3A4. *Trends Biochem Sci* 30:5–7. doi:[10.1016/j.tibs.2004.11.004](https://doi.org/10.1016/j.tibs.2004.11.004)
- Seal A, Passi A, Jaleel UA et al (2012) In-silico predictive mutagenicity model generation using supervised learning approaches. *J Cheminform* 4:10. doi:[10.1186/1758-2946-4-10](https://doi.org/10.1186/1758-2946-4-10)
- Sheridan RP, Korzekwa KR, Torres RA, Walker MJ (2007) Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9. *J Med Chem* 50:3173–3184. doi:[10.1021/jm0613471](https://doi.org/10.1021/jm0613471)
- Shibi IG, Aswathy L, Jisha RS et al (2015) Molecular docking and QSAR analyses for understanding the antimalarial activity of some 7-substituted-4-aminoquinoline derivatives. *Eur J Pharm Sci* 77:9–23. doi:[10.1016/j.ejps.2015.05.025](https://doi.org/10.1016/j.ejps.2015.05.025)
- Shibi IG, Aswathy L, Jisha RS et al (2016) Virtual screening techniques to probe the antimalarial activity of some traditionally used phytochemicals. *Comb Chem High Throughput Screen* 19:572–591. doi:[10.2174/1386207319666160420141200](https://doi.org/10.2174/1386207319666160420141200)
- Shou M, Grogan J, Mancewicz JA et al (1994) Activation of CYP3A4: evidence for the simultaneous binding of two substrates in a cytochrome P450 active site. *Biochemistry* 33:6450–6455
- Smith JB, Haynes MK (2002) Rheumatoid arthritis—a molecular understanding. *Ann Intern Med* 136:908–922. doi:[10.7326/0003-4819-136-12-200206180-00012](https://doi.org/10.7326/0003-4819-136-12-200206180-00012)
- Smolen JS, Steiner G (2003) Therapeutic strategies for rheumatoid arthritis. *Nat Rev Drug Discov* 2:473–488. doi:[10.1038/nrd1109](https://doi.org/10.1038/nrd1109)
- Smolen JS, Aletaha D, Koeller M et al (2007) New therapies for treatment of rheumatoid arthritis. *Lancet* 370:1861–1874. doi:[10.1016/S0140-6736\(07\)60784-3](https://doi.org/10.1016/S0140-6736(07)60784-3)
- Stewart JJP (1993) MOPAC manual, 7th edn, 1993
- Strazielle N, Gheris-Egea J-F (2005) Factors affecting delivery of antiviral drugs to the brain. *Rev Med Virol* 15:105–133. doi:[10.1002/rmv.454](https://doi.org/10.1002/rmv.454)

- Todeschini R, Gramatica P (1998) New 3D molecular descriptors: the WHIM theory and QAR applications. *Perspect Drug Discov* 355–380
- Toyoshima H, Kusaba T, Yamaguchi M (1993) Cause of death in autopsied RA patients. *Ryumachi* 33(3):209–214
- Trunzer M, Faller B, Zimmerlin A (2009) Metabolic soft spot identification and compound optimization in early discovery phases using MetaSite and LC-MS/MS validation. *J Med Chem* 52:329–335. doi:[10.1021/jm8008663](https://doi.org/10.1021/jm8008663)
- Vasanthanathan P, Hritz J, Taboureau O et al (2009) Virtual screening and prediction of site of metabolism for cytochrome P450 1A2 ligands. *J Chem Inf Model* 49:43–52. doi:[10.1021/ci800371f](https://doi.org/10.1021/ci800371f)
- Vinita P, Jinuraj KR, Jaleel UCA, Scaria V (2011) Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets. *BMC Res Notes* 4:504. doi:[10.1186/1756-0500-4-504](https://doi.org/10.1186/1756-0500-4-504)
- Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. *Protein Eng* 8:127–134. doi:[10.1093/protein/8.2.127](https://doi.org/10.1093/protein/8.2.127)
- Weyand CM, Hicok KC, Conn DL, Goronzy JJ (1992) The influence of HLA-DRB1 genes on disease severity in rheumatoid arthritis. *Ann Intern Med* 117:801–806. doi:[10.7326/0003-4819-117-10-801](https://doi.org/10.7326/0003-4819-117-10-801)
- Wildman SA, Crippen GM (1999) Prediction of physicochemical parameters by atomic contributions. *J Chem Inf Comput Sci* 39:868–873. doi:[10.1021/ci990307i](https://doi.org/10.1021/ci990307i)
- Williams JA, Hyland R, Jones BC et al (2004a) Drug–drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUC_i/AUC) ratios. *Drug Metab Dispos* 32:1201–1208. doi:[10.1124/dmd.104.000794](https://doi.org/10.1124/dmd.104.000794)
- Williams PA, Cosme J, Vinkovic DM et al (2004b) Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. *Science* 305:683–686. doi:[10.1126/science.1099736](https://doi.org/10.1126/science.1099736)
- Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 58:109–130. doi:[10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1)
- Yamagishi H, Shirakami S, Nakajima Y et al (2015) Discovery of 3,6-dihydroimidazo[4,5-d]pyrrolo[2,3-b]pyridin-2(1H)-one derivatives as novel JAK inhibitors. *Bioorg Med Chem* 23:4846–4859. doi:[10.1016/j.bmc.2015.05.028](https://doi.org/10.1016/j.bmc.2015.05.028)
- Yamashita S, Furubayashi T, Kataoka M et al (2000) Optimized conditions for prediction of intestinal drug permeability using Caco-2 cells. *Eur J Pharm Sci* 10:195–204. doi:[10.1016/S0928-0987\(00\)00076-2](https://doi.org/10.1016/S0928-0987(00)00076-2)
- Yano JK, Wester MR, Schoch GA et al (2004) The structure of human microsomal cytochrome P450 3A4 determined by X-ray crystallography to 2.05-Å resolution. *J Biol Chem* 279:38091–38094. doi:[10.1074/jbc.C400293200](https://doi.org/10.1074/jbc.C400293200)
- Yee S (1997) In vitro permeability across Caco-2 cells (colonic) can predict in vivo (small intestinal) absorption in man—fact or myth. *Pharm Res* 14:763–766. doi:[10.1023/A:1012102522787](https://doi.org/10.1023/A:1012102522787)
- Zhao X, Chen M, Huang B et al (2011) Comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) studies on $\alpha(1A)$ -adrenergic receptor antagonists based on pharmacophore molecular alignment. *Int J Mol Sci* 12:7022–7037. doi:[10.3390/ijms12107022](https://doi.org/10.3390/ijms12107022)
- Zhou D, Afzelius L, Grimm SW et al (2006) Comparison of methods for the prediction of the metabolic sites for CYP3A4-mediated metabolic reactions. *Drug Metab Dispos* 34:976–983. doi:[10.1124/dmd.105.008631](https://doi.org/10.1124/dmd.105.008631)