

# Studies in QSAR modelling for selection of potential inhibitors for drug discovery

By  
Pushkar D. Kunde  
10CC11J26031

A thesis submitted to the  
Academy of Scientific and Innovative Research  
for the award of the degree of  
**DOCTOR OF PHILOSOPHY**  
in  
**SCIENCE**

Under the supervision of  
Dr. Sanjay P. Kamble and Dr. V. Ravi Kumar



CSIR-National Chemical Laboratory, Pune



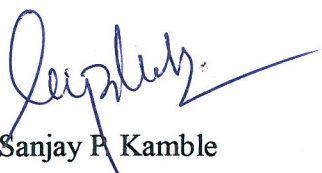
Academy of Scientific and Innovative Research  
AcSIR Headquarters, CSIR-HRDC campus  
Sector 19, Kamla Nehru Nagar,  
Ghaziabad, U.P. – 201 002, India

August 2019

# Certificate

This is to certify that the work incorporated in this Ph.D. thesis entitled “**Studies in QSAR modelling for selection of potential inhibitors for drug discovery**” submitted by **Mr. Pushkar D. Kunde** to Academy of Scientific and Innovative Research (AcSIR) in fulfilment of the requirements for the award of the Degree of **Doctor of Philosophy in Chemical Sciences**, embodies original research work under our supervision. We further certify that this work has not been submitted to any other University or Institution in part or full for the award of any degree or diploma. Research material obtained from other sources has been duly acknowledged in the thesis. Any text, illustration, table etc., used in the thesis from other sources, have been duly cited and acknowledged.

It is also certified that this work done by the student, under our supervision, is plagiarism free.



Dr. Sanjay P. Kamble  
Research Guide  
Chemical Engineering &  
Process Development Division  
CSIR-National Chemical Laboratory  
Pune  
Date: 1<sup>st</sup> August 2019  
Place: Pune




Dr. V. Ravi Kumar  
Research Co-guide  
Chemical Engineering &  
Process Development Division  
CSIR-National Chemical Laboratory  
Pune  
Date: 1<sup>st</sup> August 2019  
Place: Pune



Mr. Pushkar D. Kunde  
Student  
Chemical Engineering &  
Process Development Division  
CSIR-National Chemical Laboratory  
Pune  
Date: 1<sup>st</sup> August, 2019  
Place: Pune

# Statements of Academic Integrity

I, Pushkar D. Kunde, a Ph.D. student of Academy of Scientific and Innovative Research (AcSIR) with Registration No. 10CC11J26031 hereby undertake that, the thesis entitled “Studies in QSAR modelling for selection of potential inhibitors for drug discovery” has been prepared by me and the document reports original work carried out by me and is free of any plagiarism in compliance with the UGC regulation on “*Promotion of Academic Integrity and Prevention of Plagiarism in Higher Education Institutions (2018)*” and the CSIR guidelines for “*Ethics in Research and in Governance (2020)*”.



Name: Mr. Pushkar D. Kunde  
Chemical Engineering &  
Process Development Division  
CSIR-National Chemical Laboratory  
Date: 1<sup>st</sup> August, 2019  
Place: Pune

---

It is hereby certified that the work done by the student, under our supervision, is plagiarism free in accordance with the UGC Regulations on “*Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions (2018)*” and the CSIR Guidelines for “*Ethics in Research and in Governance (2020)*”.



Dr. V. Ravi Kumar  
Research Co-supervisor  
Chemical Engineering &  
Process Development Division  
CSIR-National Chemical Laboratory  
Date: 1<sup>st</sup> August, 2019  
Place: Pune



Dr. Sanjay P. Kamble  
Research Supervisor  
Chemical Engineering &  
Process Development Division  
CSIR-National Chemical Laboratory  
Date: 1<sup>st</sup> August, 2019  
Place: Pune

**Dedicated to Aai, Pappa and Shital ...  
... and their unconditional love**

## **Acknowledgements:**

I would like to take this opportunity to thank everyone who helped and supported me during my Ph.D. First I would like to thank my guide Dr. Sanjay P. Kamble for his help in advising the work and for always supporting me during difficult times.

I sincerely express my gratitude towards my Guru, Dr. V. Ravi Kumar for guiding, supporting and always encouraging me. I thank him for being patient with me and giving me freedom to work with new ideas.

Without the fruitful discussions with Dr. B. D. Kulkarni this work would not have been possible. His valuable inputs will always be missed.

Help from Prof. Ameeta Ravikumar was always available when I found myself in a problem with my work. I thank her for showing me the way whenever I thought I was stuck.

I thank my Doctoral Advisory Committee members, Dr. Nayana Vaval, Dr. Jayant Khire, and Dr. Sunil Joshi for providing me with helpful guidance from time to time during the course of this work.

I take this opportunity to thank The Director, CSIR-National Chemical Laboratory, Pune, for providing me the opportunity and facility to work in NCL.

I thank Council of Scientific and Industrial Research for the research fellowship to carry out this work.

I thank The Director, AcSIR, for the opportunity to pursue my Ph.D.

I express my sincere gratitude towards Dr. Mahesh Kulkarni, Dr. Santosh Mhaske and Dr. B.L.V. Prasad for their support and guidance in all the academic matters.

All the Staff of Students Academic Office always made sure that all the paperwork went through without a glitch. I express my gratitude towards Mrs. Kolhe, Ms. Vaishali, Ms. Komal, Ms. Harshita and Ms. Vijaya for the same.

My thanks to the Staff of all the Administrative and Technical Sections of NCL for their prompt actions in cases of emergencies which kept the work in NCL trouble free.

The help from my lab-mates at various points of time during this work was monumental. For this I thank Dr. Sudha Ramkumar, Dr. Ketan Sarode and Rahul Doiphode for their valuable help.

I would like to thank all the special friends from NCL, Dr. Chetan Joshi, Dr. Fazal Shirazi, Dr. Manisha Kapoor, Dr. Akshay Singan, and Dr. Pooja Singh for their support during good and bad times. I enjoyed my stay in NCL due to them. I also thank my friends from Biochemical sciences division Dr. Santosh Chavan, Dr. Sarika Mane, Dr. Ejaj Pathan, Dr, Shuklangi Kulkarni and Dr. Pradnya Chavan for always keeping a cheerful atmosphere.

My sincere thanks to Bhushan and Smita Patil for guiding me through the problems in life. I have a lot to learn from them.

I take this opportunity to thank Dr. Swati Khole, Dr. Swapnil Gaikwad, Dr. Srijay Kamat, Dr. Pradeep Devhare, Prasad Danave, Deepak Bangal, Vruahali, Tanaji Kashid, Paresh Vishwasrao, and Shailaja Maestry who are very close to my heart their support and friendship during this work helped me overcome bad times.

I would like to specially mention the support given by my Parents-in-law and Ravikiran during the difficult times. Hats-off to their patience.

I thank my brother, Bhanuprasad, for keeping me cheerful and always believing in me.

I express my gratitude towards my parents who brought me into this world, provided me with everything necessary and showered unconditional love.

And at the last my wife Shital, it is beyond me to put in words her unconditional support and love even when the times are bad for her. It is she who keeps me going.

## Table of Contents

<b>List of Abbreviations .....</b>	<b>i</b>
<b>List of Figures.....</b>	<b>iii</b>
<b>List of Tables.....</b>	<b>vi</b>
<b>Summary.....</b>	<b>viii</b>
<b>1. Chapter 1: Introduction.....</b>	<b>1</b>
1.1 Rational drug design.....	2
1.2 Computer aided drug design (CADD).....	3
1.3 Structure based drug design (SBDD).....	4
1.4 Ligand based drug design (LBDD).....	5
1.4.1 Pharmacophore modelling.....	5
1.4.2 Quantitative structure activity relationship (QSAR) modelling.....	6
1.4.2.1 Molecular descriptors.....	6
1.4.2.2 Statistical techniques.....	7
1.4.2.3 Cross validation (CV).....	8
1.4.2.4 Model performance parameters.....	9
1.4.2.5 Biological activity data.....	10
1.4.2.6 Target systems (TSs).....	11
TS-1: 4-phenylpyrrolocarbazole derivatives as WEE1 inhibitors as anti-cancer compounds.....	11
TS-2: Benzylpiperidine derivatives as acetylcholinesterase inhibitors for neurological disorders.....	12
TS-3: 2-substituted dipyrrodoiazepinone as reverse transcriptase (RT) inhibitors for treatment of HIV-1.....	13
TS-4: 2-pyridinone as reverse transcriptase (RT) inhibitors for treatment of HIV-1.....	13
TS-5: Cyclic urea inhibitors as protease (PR) inhibitors for treatment of HIV-1.....	13
TS-6: Azilide derivatives as anti-malarial compounds.....	14
<b>2. Chapter 2: Multidimensional scaling and Support vector regression based 2D-QSAR modelling.....</b>	<b>16</b>
2.1 Introduction.....	17
2.2 Methodology.....	18
2.2.1 Importing of chemical structures and biological activities of inhibitors.....	18

2.2.2 Optimization of molecular structures and calculation of partial atomic charges.....	18
2.2.3 Calculation of the shortest path distances between atoms using the connectivity network of the atoms.....	20
2.2.4 Multidimensional Scaling (MDS).....	21
2.2.5 Procrustes analysis.....	22
2.2.6 Creation of 2D image based descriptors.....	22
2.2.7 Principal Component Analysis (PCA).....	23
2.2.8 Support Vector Regression (SVR).....	25
2.2.9 Feature selection for SVR.....	26
2.2.10 Calculations of goodness of fit parameters.....	27
2.3 Results and discussion.....	29
2.4 Conclusions.....	36
<b>3. Chapter 3: QSAR modelling with pseudo-molecular field descriptors for potential applications in drug design.....</b>	<b>38</b>
3.1 Introduction.....	39
3.2. Methodology.....	40
3.2.1 Importing of inhibitor structures.....	42
3.2.1.1 Compounds with known biological activity.....	42
3.2.1.2 Compounds with unknown biological activity.....	42
3.2.2 Scaffold based alignment of inhibitor molecules in 3D mesh grids.	42
3.2.3 Calculation of pseudo-molecular field (PMF) values and pseudo field molecular descriptors (PFMDs) .....	44
3.2.4 QSAR modelling.....	45
3.2.4.1 Partial Least Squares (PLS).....	47
3.2.4.2 PMF-PLS algorithm.....	49
3.3 Results and discussion.....	52
3.3.1 Generation of PFMDs and QSAR modelling.....	52
3.3.2 PMF-PLS algorithm performance using 2D charge based descriptors.....	61
3.3.3 Model comparison.....	62
3.3.4 Screening of natural compounds.....	64
3.4. Conclusions.....	66
<b>4. Chapter 4: Varying component PLS QSAR modelling and docking studies of potential inhibitors.....</b>	<b>68</b>
4.1 Introduction.....	69
4.2 Methodology.....	71



4.2.1 Molecular descriptors.....	71
4.2.2 QSAR modelling.....	71
4.2.3 Docking simulations.....	73
4.3 Results and discussion.....	74
4.3.1 QSAR modelling.....	74
4.3.2 Screening of natural compounds and docking studies.....	80
4.3.2.1 Docking results for TS-1.....	82
4.3.2.2 Docking results for TS-2.....	86
4.3.2.3 Docking results for TS-3 and TS-4.....	87
4.3.2.4 Docking results for TS-5.....	94
4.4 Conclusions.....	96
<b>5. Conclusions and future scope.....</b>	<b>97</b>
<b>6. References .....</b>	<b>100</b>
<b>7. Appendix .....</b>	<b>108</b>
<b>8. Abstract .....</b>	<b>186</b>
<b>9. Publications.....</b>	<b>187</b>

## List of Abbreviations

Abbreviation	Full form
AChE	Acetylcholinesterase
ADMET	Absorption Distribution Metabolism Excretion Toxicity
AID	PubChem Assay ID
CADD	Computer Aided Drug Design
CoMFA	Comparative Molecular Field Analysis
CoMSIA	Comparative Molecular Similarity Index Analysis
CS	Catalytic Site
CV	Cross Validation
HIV	Human Immunodeficiency Virus
IC <sub>50</sub>	Inhibition Concentration 50
LBDD	Ligand Based Drug Design
LOO	Leave-One-Out
MAE	Mean Absolute Error
MDS	Multidimensional Scaling
MLR	Multiple Linear Regression
NIPALS	Non-Iterative Partial Least Squares
NNI	Non-Nucleoside Inhibitors
NNIBP	Non-Nucleoside Inhibitor Binding Pocket
NRMSE	Normalized Root Mean Square Error
NRMSECV	Normalized Root Mean Square Error of Cross Validation
NRMSEP	Normalized Root Mean Square Error of Prediction
PAS	Peripheral Anionic Site
PCA	Principal Component Analysis
PCR	Principal Component Regression
PDB	Protein Data Bank
PFMD	Pseudo Field Molecular Descriptors
PLS	Partial Least Squares
PMF	Pseudo-Molecular Field
PMF-PLS	Pseudo-Molecular Field Partial Least Squares
PR	Protease
QSAR	Quantitative Structure Activity Relationship
RT	Reverse Transcriptase
SAR	Structure-activity Relationship
SBDD	Structure Based Drug Design

---

Abbreviation	Full form
SIMPLS	Statistically Inspired Modification of PLS
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SVR	Support Vector Regression
TS	Target System
VC-PLS	Varying Component PLS

---

## List of Figures

No.	Figure legend	Pg. No.
2.1	Multi-dimensional Scaling. Conversion of (A) 3D molecular structure to (B) 2D MDS graph of the molecule	23
2.2	Flowchart of steps involved in generation of 2D image based descriptors and developing 2D-QSAR models using these descriptors	29
2.3	Highlighted bonds show the shortest path between atoms C4 and N15 calculated using Dijkstra's algorithm for one of the Wee1 inhibitors	30
2.4	(A) 3D structure of the compound its (B) Binary image on plotting the MDS co-ordinates on to 2D plane. The active pixels with value one are black in colour whereas the inactive pixels are white in colour. (C) Gray scale image generated after plotting the partial atomic charges at the active pixels in (B). The size of active pixels has been increased in image (B) and similarly the gray scale image in (C) has been colour coded according to the pixel value and the active pixel size increased for better representation.	31
2.5	Percentage of variation captured in the principal components for TS-1	32
2.6	Plots of actual $pIC_{50}$ values ( $Y_{train}$ ) vs. the predicted values ( $\hat{Y}_{train}$ ) using image-based 2D-QSAR model for cross-validation. (A) TS-1, (B) TS-2, (C) TS-3, (D) TS-4, (E) TS-5 and (F) TS-6 inhibitors.	34
2.7	Plots for actual $pIC_{50}$ values ( $Y_{test}$ ) vs. the predicted values ( $\hat{Y}_{test}$ ) using image-based 2D-QSAR model for test sets of (A) TS-1, (B) TS-2, (C) TS-3, (D) TS-4, (E) TS-5 and (F) TS-6 inhibitors.	35
3.1	General flowchart to study the proposed PMF-PLS approach for QSAR modelling	41
3.2	Aligned 3D structures of the 4-phenylpyrrolocarbazole derivatives for TS-1 encapsulated in a 3D mesh grid. The mesh grid is not drawn to scale	43
3.3	Flowchart of QSAR modelling by PMF-PLS algorithm	46
3.4	Percentage of variance observed in $X_{train}$ explained with varying number of PLS components ( $a$ ) used for regression for TS-1	54

No.	Figure legend	Pg. No.
3.5	RMSE values for predictions of training (red) and test (blue) sets with varying values of $a$ used for the PLS regression of TS-1. The arrow mark indicates the minimum <i>RMSE</i> for test set prediction at $a = 25$	55
3.6	Effect of Procrustes transformation on the scores of compounds. Euclidean distances between $T_{ref}$ and $T_j$ before (red) and $T_{r,j}$ after (blue) transformation on removing compound number <b>22</b> (Table A1) from $(X_{train,ref}, Y_{train,ref})$ of TS-1. The red and blue lines are the mean distances of the compounds calculated before and after Procrustes transformation	56
3.7	Plots of actual pIC <sub>50</sub> values ( $Y$ ) vs. the predicted values ( $\hat{Y}$ ) for cross-validation using PMF-PLS QSAR model. (A) TS-1, (B) TS-2, (C) TS-3, (D) TS-4, (E) TS-5 and (F) TS-6 inhibitors. The training set compounds are marked in red and test set compounds in black as specified in the Appendix, Tables A-15 to A-20, respectively	58
3.8	Plots for actual pIC <sub>50</sub> values ( $Y_{val}$ ) vs. the predicted values ( $\hat{Y}_{val}$ ) using PMF-PLS QSAR model for validation sets of (A) TS-1, (B) TS-2, (C) TS-3, (D) TS-4, (E) TS-5 and (F) TS-6 inhibitors. The numbers in the panels A to F indicate the validation set compound number as specified in the Appendix, Tables A-15 to A-20, respectively	59
4.1	Flowchart of QSAR modelling by VC-PLS	72
4.2	Plots of actual pIC <sub>50</sub> values ( $Y_{train}$ ) vs. the predicted values ( $\hat{Y}_{train}$ ) for cross-validation for the best VC-PLS model for (A) TS-1, (B) TS-2, (C) TS-3, (D) TS-4, (E) TS-5 and (F) TS-6 inhibitors.	78
4.3	Plots for actual pIC <sub>50</sub> values ( $Y_{test}$ ) vs. the predicted values ( $\hat{Y}_{test}$ ) for test sets using VC-PLS QSAR model of (A) TS-1, (B) TS-2, (C) TS-3, (D) TS-4, (E) TS-5 and (F) TS-6 inhibitors.	79
4.4	Natural compounds docked in the active cleft of Wee1 A) compound SN00226661 and B) compound SN00272309. Natural compounds docked in the peripheral site of Wee1 C) compound SN00362911 and D) compound SN00362452. Natural compounds are displayed in dark blue colour whereas the Wee1 residues interacting with the compound are shown in light blue.	83

No.	Figure legend	Pg. No.
4.5	Detailed view of active cleft residues of Wee1 interacting with the docked natural compounds A) SN00226661 and B) SN00272309 and peripheral site residues of Wee1 interacting with natural compounds C) SN00362452 and D) SN00362911	84
4.6	(A) SN00335138 docked to the active site of AChE. SN00335138 is displayed in drack blue whereas the AChE residues interacting with it are displayed in light blue. (B) Detailed view of the AChE residues interacting with SN00335138	87
4.7	(A) SN00118406 docked in the NNIBP of HIV-1 RT. SN00118406 is displayed in drack blue whereas the HIV-1 RT residues interacting with it are displayed in light blue. (B) Detailed view of the HIV-1 RT residues interacting with SN00118406	90
4.8	Natural compounds similar to 2-pyridinones docked in the NNIBP of HIV-1 RT A) SN00008635, B) SN00008637, C) SN00008647, D) SN00008860, E) SN00010264 and F) SN00063879. Natural compounds are displayed in dark blue color whereas the Wee1 residues interacting with the compound are shown in light blue	92
4.9	Detailed view of NNIBP residues interacting with the docked natural compounds similar to 2-pyridinones A) SN00008635, B) SN00008637, C) SN00008647, D) SN00008860, E) SN00010264 and F) SN00063879	93
4.10	A) Pose of compound SN00215212 docked into the active site of HIV-1 PR. Natural compounds are displayed in dark blue color whereas the HIV-1 PR residues interacting with the compound are shown in light blue. (B) Detailed view of interactions between HIV-1 PR residues and SN00215212.	96

## List of Tables

No.	Table Title	Pg. No.
1.1	Classification of different situations for drug design depends upon the availability of the 3D structure of the target protein and ligand molecules	3
2.1	Scaffolds of the inhibitor compounds studied with corresponding assay IDs (AID) for TS-1 to TS-6	19
2.2	Steps in calculation of PCA scores and loadings in NIPALS algorithm	24
2.3	Selected PCA components and $\sigma$ values for SVR regression	27
2.4	The 2D image sizes and corresponding row vectors for TS-1 to TS-6	31
2.5	2D-QSAR model performance statistics for TS-1 to TS-6	33
3.1	Electron affinity and electronegativity values of the atoms used for calculating PMF values	44
3.2	3D Mesh grid (box size) and 1-way PFMD sizes for TS-1 to TS-6	45
3.3	Steps of SIMPLS algorithm adopted for PMF-PLS (Adapted from de Jong, 1993)	48
3.4	PMF-PLS QSAR model fitting statistics for TS-1 to TS-6	57
3.5	Model quality using the mean absolute error (MAE) based criteria	61
3.6	Performance comparison of present QSAR algorithm using PFMDs and 2D charge based descriptors.	62
3.7	Comparison of present PMF-PLS QSAR model with other QSAR models for the same datasets in this study	63
3.8	Tanimoto scores and the predicted biological activities, $\hat{Y}_{np}$ , using PMF-PLS QSAR model of natural compounds obtained from Super Natural II database	64
4.1	Docking parameters	74
4.2	VC-PLS QSAR model fitting statistics for the best models of TS-1 to TS-6	75

No.	Table Title	Pg. No.
4.3	Model fitting statistics for the five best VC-PLS models for TS-1	76
4.4	Model fitting statistics for the five best VC-PLS models for TS-2	76
4.5	Model fitting statistics for the five best VC-PLS models for TS-3	76
4.6	Model fitting statistics for the five best VC-PLS models for TS-4	76
4.7	Model fitting statistics for the five best VC-PLS models for TS-5	77
4.8	Model fitting statistics for the five best VC-PLS models for TS-6	77
4.9	Predicted pIC50 values, $\hat{Y}_{np}$ , of natural compounds obtained from Super Natural II database using the five VC-PLS QSAR models	80
4.10	Interactions between the docked natural compounds and Wee1 protein residues	84
4.11	Interactions between the docked natural compound similar to benzylpiperidine derivatives and the residues of AChE	87
4.12	Interactions between the docked natural compound similar to 2-substituted dipyridodiazepinones and the HIV-1 RT residues	89
4.13	Interactions between the docked natural compounds similar to 2-pyridinones and the protein residues	90
4.14	Interactions between the docked natural compounds similar to cyclic urea derivatives and the HIV-1 PR residues	95



## Summary:

### Introduction:

Modern drug discovery is a target based process during which a specific host or pathogen protein that is critical in the progression of the disease is identified as target and drug molecules are designed to act against the target protein (Mandal *et al.*, 2009). Since such drug molecules are designed to target a specific protein, the chances of these drug molecules affecting the non-target proteins are expected to be low resulting in an efficient treatment of the disease with minimal side effects. Target based designing of the drug molecules (Wang *et al.*, 2015) is done by adopting either a structure based drug discovery (SBDD) or a ligand based drug discovery (LBDD) approach. The SBDD applies prior knowledge of the 3D structure of the target protein to identify a potential drug molecule. However, the SBDD approach cannot be used when the 3D structure of the target protein is not known. On the other hand, the LBDD does not require information about the 3D structure of the target protein. When using LBDD approach, the structural characteristics of known ligands that interact with the target protein are taken into consideration. It is assumed that compounds with similar structure will interact with the target protein in a similar manner. Hence, studies of ligand characteristics would aid in designing of novel drug molecules. Thus, the structures of known ligands (substrates or known inhibitors of the target protein) may be chemically modified and then tested for their biological activity. The chemical modification made to the known ligand structure would determine the potency of the newly synthesized drug molecule. It may be noted that the number of chemical modifications that can be performed on a ligand structure are high and synthesizing this vast array of analogues is both resource and time intensive. It is therefore important to study the relationship between structure of a ligand molecule and its biological activity. In an effort towards this understanding, quantitative structure-activity relationships (QSAR) need to be computationally developed. QSAR models have become an integral part of modern medicinal chemistry (Kubinyi, 2002; Khan, 2010; Silva and Trossini, 2014) and are developed for ligand molecules for which the desired biological activities e.g., pIC<sub>50</sub>, LD<sub>50</sub> (lethal dose), toxicity, etc. are available. Subsequently, the model may be used to predict the biological activities of newly designed molecules (Ma *et al.*, 2014).

A QSAR model describes the biological activity ( $Y$ ) of a molecule as a function ( $f$ ) of its structure, i.e.,  $Y = f(\text{ligand structure features})$ . The structural features are the predictors or independent variables of the QSAR model and referred to as molecular descriptors because they are used to quantitatively represent different ligand molecules in the model. These descriptors should adequately capture the structural

attributes of molecules responsible for their biological activity while simultaneously being sensitive to the structural differences between analogues. Molecular descriptors can range from simplest of the molecular properties, e.g., molecular weight, dissociation constant, partition coefficients, solubility, etc., to ones as complex as the 3D electrostatic field values at spatial points around the molecule that may even be supplemented with additional information about conformations (Gasteiger and Eds, 2003; Cronin and Schultz, 2003; Todeschini and Consonni, 2008; Le *et al.*, 2012; Roy and Das, 2014; Damale *et al.*, 2014).

### **Objectives of the work:**

We choose six target systems involved in four diseases for the studies performed in this work, namely,

1. Inhibition of kinase Wee1 by 4-phenylpyrrolocarbazole derivatives as anti-cancer agents,
2. Inhibition of acetylcholine esterase by benzylpiperidine derivatives for treatment of neurological disorders,
3. 2-substituted dipyridodiazepinone derivatives as inhibitors of HIV-1 reverse transcriptase (RT),
4. 2-pyridinone as inhibitors of HIV-1 reverse transcriptase (RT),
5. HIV-1 protease (PR) inhibition by cyclic urea derivatives for HIV-1 treatment and
6. Inhibition of *Plasmodium falciparum* growth by 15 membered azalide derivatives as anti-malarial agents.

The introductory **chapter 1** of the thesis discusses the background to QSAR modelling. The description and significance of each target system above is also discussed in chapter 1 of the thesis. The work carried is then presented in chapters 2-4 of the thesis. The objectives of work done in each of the chapters is as follows:

In **chapter 2**, we aim at developing 2D image based descriptors from the optimal 3D structures of the compounds to regress with the pIC<sub>50</sub> values of the compounds for QSAR modelling. These descriptors were created to retain the interatomic shortest path distances from 3D space and the partial atomic charges.

In **chapter 3**, we present a novel 3D pseudo-molecular field (PMF) which depends on the intrinsic properties of the atoms. These 3D pseudo-field molecular descriptors (PFMDs) were obtained using the intrinsic properties of atoms, namely, electron affinity and the electronegativity (Mulliken, 1934) unlike the traditional electrostatic field descriptors which are calculated using the partial atomic charges of the atoms. To regress these 3D descriptors with the biological activity of the molecules, we have in this work, developed PMF-PLS methodology. PMF-PLS QSAR models for all the six systems were then used for predicting the pIC<sub>50</sub> values of natural

compounds obtained from SuperNatural II data base (Banerjee *et al.*, 2015) with scaffolds similar to the molecules in the target systems and for which the experimental pIC<sub>50</sub> values are not known.

In **chapter 4**, we devise a second regression methodology using the SIMPLS (de Jong, 1993) variant of PLS method, namely, VC-PLS, for the regression of molecular descriptors with the corresponding pIC<sub>50</sub> values. We do this in order to see the effect of using a varying number of PLS components for developing regression models. We also used the VC-PLS QSAR models developed for all the target systems to screen the natural compounds studied in chapter 3. Finally, we perform docking studies with the natural compounds for which the pIC<sub>50</sub> values were predicted to be high during the screening to complement the screening results.

The summary of the main results obtained in each chapter are as follows.

## **Chapter 2: Multidimensional scaling (MDS) and support vector regression (SVR) based 2D-QSAR modelling**

The aim of the work in this chapter was to develop 2D image based molecular descriptors which contain molecular information from the optimal 3D structures of the compounds. Towards this aim optimized 3D structures of the compounds were initially obtained. Dijkstra's optimization algorithm (Dijkstra, 1959) was then used to obtain the shortest path distance between pairs of atoms of a molecule to generate a distance matrix. These distance matrices were then subject to double centring and multi-dimensional scaling to obtain the coordinates of atoms in the 2D space. The atomic positions were then transformed using Procrustes analysis (Kendall, 1989) to align the scaffold atoms common in all the molecules of a target system. These transformed 2D coordinates were then plotted on a plane which was then converted into a 2D binary image for every molecule. The atomic positions in the image were then given the value of the partial atomic charges of the corresponding atoms thus obtaining a grey scale image descriptor for the molecules. These descriptors were then subject to PCA for dimensionality reduction. The PCA scores were used for regression with the corresponding pIC<sub>50</sub> values using support vector regression. The regression of these 2D image based descriptors against the pIC<sub>50</sub> values of the compounds with these 2D QSAR models yielded for the six target systems Pearson's correlation coefficient for test set ( $r_{pred}$ ) ranging between 0.66 to 0.90, coefficient of determination for test set ( $Q^2_{ext(F1)}$ ) between 0.3 to 0.73 and normalized root mean squared error for prediction of test set ( $NRMSEP$ ) between 0.12 to 0.22. The model performance parameters values were observed to be good for 4 target systems indicating the usefulness of the models. These models however had the drawback of being

computationally intensive due to the feature selection steps involved during the regression.

### **Chapter 3: QSAR modelling with pseudo-molecular field descriptors for potential applications in drug design**

The aim of the studies in this chapter was to develop 3D molecular descriptors dependent on the intrinsic properties of the atoms. The optimal 3D structures of the molecules were aligned to superimpose the scaffold common in all the molecules. A 3D grid was defined around the molecules and pseudo-molecular field values for every molecule were calculated at every grid point using the formula,

$$V_{j,k,l} = \sum_{i=1}^{n_a} \left( \frac{\sigma E_a(i) \chi(i)}{d(i)} \right)$$

where,  $E_a$  and  $\chi$  are the electron affinity and electronegativity of the atoms in the molecule,  $\sigma$  is the scaling factor and  $d$  is the distance of the grid point from the  $i^{\text{th}}$  atom. The calculation of PFMDs was analogous to the electrostatic field based molecular descriptors. The values of PMFs at the grid points were used as the PFMDs for regression against the pIC<sub>50</sub> values of the molecules. Regression model was obtained through the PMF-PLS methodology where the data set in initially randomly divided into training, test and validation sets to arrive at a reference training and test set. Multiple sets of regression coefficients are obtained by making changes to the training set by removing one compound to the test set of vice versa. A Procrustes transformation of the PLS scores and loadings was performed for the change in the training set that did not yield satisfactory prediction results. Finally all the sets of regression coefficients were averaged to obtain the final regression model. The PMF-PLS model performance was observed to yield  $r_{\text{pred}}$  between 0.81 and 0.94,  $Q^2_{\text{ext}(F1)}$  between 0.62 and 0.71 and  $NRMSEP$  between 0.10 and 0.16. The model performance was good across all the six target systems. The PMF-PLS QSAR models were observed to be computationally light as compared to the 2D image based QSAR models studied in Chapter 2. The comparison of PMF-PLS QSAR model for target systems 1 to 5 with existing models for the same data sets from literature showed that the PMF-PLS model performance was comparable to these QSAR model. The obtained PMF-PLS QSAR models were then used for screening the natural compounds.

### **Chapter 4: Varying component PLS QSAR modelling and docking studies of potential inhibitors**

With the aim of studying the effect of using varying number of component during leave-one-out cross validation we developed a varying component

methodology (VC-PLS) for PLS regression of the PFMDs against the pIC<sub>50</sub> values of the molecules. The data set was initially divided randomly into training and test sets and a leave-one-out cross-validation was performed for the training set for a range of PLS components using the SIMPLS method for PLS regression. The number of PLS components that resulted in the minimum error in the prediction of the molecule left out of the training set was chosen as the optimal number of components for that compound and the corresponding set of regression coefficients were thus obtained for every molecule. The regression coefficients were then averaged to get the regression model which was validated by performing the predictions for the test set. This process was repeated for multiple random combinations of training and test sets. The best performing models were then used for further studies. The best VC-PLS QSAR models yielded  $r_{pred}$  between 0.65 and 0.95,  $Q^2_{ext(F1)}$  between 0.40 to 0.87 and  $NRMSEP$  between 0.07 and 0.15. The five best performing VC-PLS models were then used for screening the natural compounds.

The predictions of pIC<sub>50</sub> values for natural compounds were performed using both PMF-PLS and VC-PLS QSAR models. These predictions were found to be consistent between PMF-PLS and VC-PLS QSAR models across all the six target systems. The natural compounds with high pIC<sub>50</sub> values were further analysed by performing docking studies of these compounds in the respective target proteins except for anti-malarial azalides as specific target for those compounds is not known. Docking studies indicated the binding of the natural compounds to the active sites of all the proteins. The interactions of the docked compounds were observed to be with the amino acid residues that are important either in catalysis or binding of the substrates to the proteins. These results thus, complement the prediction of high pIC<sub>50</sub> values for these compounds by the QSAR models.

## Conclusions

We were able to successfully develop 2D image based descriptors containing molecular information from the 3D structure of the compounds. The model development with these descriptors was however computationally time intensive and need to be further optimized for accurate predictions. The 3D pseudo-molecular field based descriptors were also successfully developed using the intrinsic properties of atoms like electron affinity and electronegativity. The PFMDs had the advantage that the electron affinity and the electronegativity values of the atoms used in their calculations do not vary unlike the partial atomic charges of the atoms which need to be calculated separately for every molecule for traditional CoMFA based QSAR models. Good model performance parameters for both PMF-PLS and VC-PLS QSAR models and consistency in their predictions of pIC<sub>50</sub> values of the natural compounds

indicated the stability of both the methods for QSAR modelling using PFMDs. Finally, the results of docking studies indicate that the PFMD based QSAR models have the potential to be used for the screening of molecules. Thus, we were able to screen natural compounds with potential activity in treatment of diseases like cancer, neurological disorders, HIV and malaria. The work presented lays a novel framework for QSAR modelling and it is worthwhile to further explore the directions proposed here.

# **Chapter 1**

## **Introduction**

## 1.1 Rational drug design:

The drug discovery processes before 1960s were based on testing a large number of compounds (thousands) for desired activity. On finding compounds with desired activity, hundreds of related molecules would be synthesized to develop drug molecules with desired pharmacokinetic properties. Such methods of drug design were associated with enormous cost and high risk in terms of undesired side effects of the drug molecules. With advances in research and knowledge rational drug design approaches can now be adopted (Reddy and Parrill, 1999; Mandal *et al.*, 2009). The first step in rational drug design process is identification of the target against which the drugs are to be designed. A target is a biomolecule, generally a protein that plays an important role in the progression of the disease. As examples, a target could be a protein whose activity is indispensable for the survival of the pathogen or a human protein whose activity has been or needs to be altered. Once the target is identified, drug molecules are designed to act against it. This may be done by applying the knowledge of the structure of the target protein or the structure of the ligands known to interact with it. New molecules could then act either by competitively binding to the target protein with more affinity than its natural ligands (Westholm *et al.*, 2009) or by allosterically changing the protein conformation (Tubeleviciute-Aydin *et al.*, 2019). The drug molecules could also act by enhancing or activating the target protein whose activity has been down-regulated in a disease (Vella *et al.*, 2019).

The chances of a molecule becoming an effective and safe drug for patients depends on but are not limited to its pharmacokinetic properties, namely, the ADMET properties (Silva and Trossini, 2014; Cumming *et al.*, 2013), i.e.,

**Absorption:** Assimilation of an orally ingested drug into the bloodstream of the patients.

**Distribution:** Distribution of the drug in the body.

**Metabolism:** The process of metabolism of the drug molecule in the body and the byproducts formed at the different stages of metabolism.

**Excretion:** Elimination (removal) of the drug from the body.



**Toxicity:** Toxic effects of the drug molecule or its metabolic byproducts to the body.

## 1.2 Computer aided drug design (CADD):

Even after identifying the target protein, designing effective drug molecule still remains a time, money and knowledge intensive task. Therefore use of computational tools towards the design of new drug molecules has become an integral part of modern drug discovery process. CADD methods are bringing down the time and cost in the screening of compounds without compromising on the quality of the lead molecule discoveries. CADD approach is used for three major purposes in the drug discovery process, namely,

- 1) Virtual screening of large libraries of compounds for detection of active molecules (hits) (Osakwe, 2016),
- 2) To guide the optimization of lead molecules to improve their ADMET properties (Khan, 2010; Silva and Trossini, 2014),
- 3) De novo drug design where drug molecules are designed from a starting molecule or by piecing together fragments (Todorov *et al.*, 2007).

Depending on whether the structure of the target and its ligands are known one of the following four situations stated in Table 1.1 could be present.

**Table 1.1:** Classification of different situations for drug design depends upon the availability of the 3D structure of the target protein and ligand molecules

Situation	Target Protein Structure	Ligand Molecule Structure
1	Unknown	Unknown
2	Unknown	Known
3	Known	Unknown
4	Known	Known

The first situation, where neither the structure of the target protein nor that of the ligand(s) is known, is the least desired. In such a case, CADD methods cannot be

implemented and experiments need to be performed to obtain the structure of the target protein and to identify active ligands. In situation 2, where the 3D ligand structure is known but the protein structure is not known, we can implement CADD methods to build math models correlating the bioactivity with the known ligand molecules. In situation 3, we would need to carry out initially experiments to obtain and characterize the structure of the ligand molecules for developing the correlations. Situation 4 is the most desirable case when all the information to develop relations are available and in addition independent docking studies could be carried out of the ligand molecules with that of the target protein because the 3D structure of the target protein is additionally available.

Depending on the availability of the structures of the target protein or its ligand the following CADD approaches could be adopted for situations 2, 3 and 4 stated in Table 1.1, namely, structure based drug design (SBDD) and ligand based drug design (LBDD).

### **1.3 Structure based drug design (SBDD):**

Also known as direct drug design approach, SBDD is adopted when structure of the target protein is known, i.e., situations 4, and 3 in Table 1.1. The protein data bank (PDB) (Berman *et al.*, 2002) which has ~154478 3D X-ray and NMR structures of biomolecules, namely, proteins, nucleic acids, protein-nucleic acid complexes, is an indispensable resource for modern drug design process. If the 3D structure of the target obtained through X-ray crystallography or NMR is not available but its amino acid sequence is known then a virtual 3D structure obtained using homology modelling (Schwede *et al.*, 2003) could also be used in the SBDD paradigm. SBDD is majorly used for structure based virtual screening and de novo drug design. Structure based virtual screening of compounds is performed by first docking each compound on the target binding site. Then depending on the quality of binding of compounds to the target, hit molecules are selected (Kitchen *et al.*, 2004). In de novo drug design the detailed information about the binding site structure is used to design molecular

structures which optimally fit the target site (Todorov *et al.*, 2007). De novo drug design methods can be used to:

- 1) Develop drugs against target protein for whom the active ligands are not known, or
- 2) Grow a fragment known to bind at the target site or
- 3) Optimally link fragments that bind to the target site or
- 4) Generate alternate compounds to the known active compounds.

#### **1.4 Ligand based drug design (LBDD):**

Also known as indirect drug design, LBDD approach is adopted when the structure of ligands interacting with the target protein is known, i.e., situations 2 and 4 in Table 1.1 and especially in absence of the structure of the target protein (Bacilieri and Moro, 2006; Acharya *et al.*, 2011). When the structures of such ligands are known, new molecules are designed by altering the molecular structures of these ligands. These newly designed molecules are then synthesized and screened for their bioactivity. Thus, it is assumed that the compounds with similar structure will interact with the target protein in a similar way and the structural characteristics of the ligands that interact with the target protein are exploited. However, the number of new molecules that can be possibly synthesized are very large making the process time and resource intensive. Towards the aim of efficiently obtaining hit molecules, computational models, namely, pharmacophore models and quantitative structure activity relationship models are developed and used to carry out the ligand based virtual screening of the compounds (Acharya *et al.*, 2011).

##### **1.4.1 Pharmacophore modelling:**

A pharmacophore is an ensemble of steric and electronic features that are required for the interactions of the potential drug molecule with the target. These molecular features, namely, hydrogen bond donor (HBD), hydrogen bond acceptor (HBA) hydrophobic groups,  $\pi$ -donor, etc. are labeled and the active ligands are superimposed to extract the common features (Yang, 2010). These extracted features

used as query to screen the compounds by first checking the presence of these features in the compounds and then their orientation in 3D space to match the query.

#### **1.4.2 Quantitative structure activity relationship (QSAR) modelling:**

In order to design drug molecules from existing ligands of the target it is important to understand the relationship between the structural features of the compounds and their activity, i.e., the structure-activity relationship (SAR) (Guha, 2013). Having an understanding of SAR for a group of molecules helps to rationally design drug molecules from existing ligands. In an effort towards understanding SAR, regression models are used which quantitate the activity of the molecule from their molecular properties (Wang *et al.*, 2015; Gramatica, 2020), namely, QSAR models. Although QSAR models have no specific starting point, the oldest report of relation between biological effects and molecular properties to be documented was in 1863 (Kubinyi, 2002). However, the works of C. Hansch and T. Fujita (Hansch and Fujita, 1964) and S. M. Free Jr. and J. W. Wilson (Free and Wilson, 1964) are the starting points of classical QSAR studies used today.

QSAR models express the activity of the molecules as a function of their molecular structure. Molecular structures are represented numerically in these models by molecular descriptors. Thus, a QSAR model may be represented as;

$$Activity = f(Molecular\ descriptors) \quad (1.1)$$

Thus, a QSAR model has three components:

1. molecular descriptors
2. statistical technique to establish the relationship
3. biological data of known ligands to develop the model

##### **1.4.2.1 Molecular descriptors:**

Molecular descriptors provide the molecular information to the QSAR model for prediction of activity of the molecules. Hence, these descriptors are the independent variables of a QSAR model. Molecular descriptors can be as simple as the global

properties of molecules like, molecular weight, partition coefficients, dissociation constants, etc. or complex representations of molecules in hyperspaces (Todeschini and Consonni, 2008). Based on the descriptor characteristics or the way in which these descriptors are obtained QSAR models may be classified as follows:

1. 1D-QSAR: These models use global molecular properties like molecular weight, partition coefficient, etc. as molecular descriptors (Hansch and Fujita, 1964)
2. 2D-QSAR: Topological indices obtained from structural patterns of molecules are used as descriptors to correlate with the activity of the molecules (Roy and Das, 2014). 2D descriptors also include image based descriptors (Freitas *et al.*, 2005; Freitas and Rittner, 2008) which contain binary images of the structures of molecules.
3. 3D-QSAR: These models use 3D structural properties of the molecules such as the electrostatic and steric fields as descriptors (Cramer *et al.*, 1988; Kubinyi, 1997a). These descriptors best represent the ligand-protein interactions that are responsible for the activity of the ligand.
4. 4D-QSAR: In order to select their bioactive conformations, the 3D descriptor of each molecules is represented in its different conformations, tautomers, orientations, stereoisomers (Lill, 2007). 4D-QSAR is used to overcome the uncertainties involved in the alignment of ligands for 3D-QSAR.
5. 5D-QSAR: In order to consider various scenarios of flexible docking, different induced fit models are explicitly considered in addition to the 4D descriptors (Lill, 2007; Vedani and Dobler, 2002).

#### **1.4.2.2 Statistical techniques:**

A wide variety of statistical techniques are used for building QSAR models depending on the types of descriptors used, especially the number of variables in the descriptors. For 1D and 2D-QSAR models using descriptors with small number of variables, linear regression techniques, namely, simple linear regression and multiple

linear regression (MLR) may be used (Hansch and Fujita, 1964; Alvin C. Rencher, 2002). However, these linear regression techniques cannot be used with QSAR models using descriptors with high dimensionality (3D to 5D-QSAR models). In such situations, regression techniques that reduce the dimensionality of the descriptors, namely, principal component regression (PCR) and partial least squares (PLS) regression, are used to build the QSAR models. These techniques reduce the dimensions of the descriptor data by projecting it to the latent space with orthogonal basis vectors (Geladi and Kowalski, 1986). More complex machine learning algorithms, namely, *k*-nearest neighbors (*k*NN), artificial neural networks (ANN) and support vector machine (SVM) are also used for building QSAR models (Verma *et al.*, 2010).

#### **1.4.2.3 Cross validation (CV):**

Cross validation is the most widely used method for performing the internal validation of a QSAR model (Wold, 1978; Gramatica, 2007). During cross validation the training set is initially divided randomly into *n* number of groups. Keeping one group aside a model is derived using the rest of the *n*-1 groups and predictions are performed for the group that was left out. This procedure is repeated *n* times to leave a different group out and predict the activities for that group during each iteration. When *n* is equal to the number of compounds in the training data set it is known as the leave-one-out (LOO) cross validation. The model performance parameters are then calculated to assess the predictive capability of the model. External validation is done by carrying out predictions for the test set using the model developed using the training set. A number of in-depth reviews of the approaches adopted in building QSAR model are available (Kubinyi, 1997a, 1997b; Kolossov and Stanforth, 2010; Cherkasov *et al.*, 2014; Polishchuk, 2017).

The work carried out in this thesis aims towards developing novel 2D and 3D molecular descriptors and methodologies for building QSAR models using these

descriptors to bring out their correlations with the corresponding biological activities of the molecules.

#### 1.4.2.4 Model performance parameters:

In this work three measures of model performance are used (Roy *et al.*, 2016), namely, Pearson's correlation coefficient ( $r$ ), coefficient of determination ( $R^2$ ) and normalized root mean squared error (NRMSE) taking into account the advantages and disadvantages of each.

**Pearson's correlation coefficient** ( $r$ ) is calculate using the formula:

$$r = \frac{cov(Y, \hat{Y})}{std(Y)std(\hat{Y})} \quad (1.2)$$

where,  $cov(Y, \hat{Y})$  is the covariance between the actual and predicted activity values,  $std(Y)$  and  $std(\hat{Y})$  are the standard deviations of the actual and predicted activity values, respectively.  $r$  measures the linear correlation between the actual and the predicted activity values and its value lies between -1 and 1. A near zero value of  $r$  indicates that the two variables  $Y$  and  $\hat{Y}$  are not correlated. A positive value of  $r$  indicates that there is positive correlation between the two variables, i.e., when the value of one variable increase then so does that of the other and vice-versa. Similarly, a negative value of  $r$  indicates a negative correlation between the two variables. In the current study we are interested in only the positive values of  $r$ , and a value  $> 0.6$  shows a significant correlation between the predicted and actual activity values. It may be noted that  $r$  only measures the overall correlation between two variables and does not indicate how accurate the prediction are. Thus, predictions with high  $r$  value could also have large errors of predictions (Roy *et al.*, 2016).

**Coefficient of determination** ( $R^2$ ) given by the relation:

$$R^2 = 1 - \frac{PRESS}{SSD} \quad (1.3)$$

where,  $SSD$  is the sum of squared deviations of the actual activity values from the mean, and  $PRESS$  is the predicted residual sum of squares. Unlike  $r$ ,  $R^2$  is sensitive to the magnitudes of the prediction errors.  $R^2$  can have a maximum value of 1 and models with  $R^2 > 0.5$  are considered to be acceptable (Roy *et al.*, 2016). However,  $R^2$  value is sensitive to the distributions of the data around the mean and it is possible that one may obtain a relatively low  $R^2$  value even with low errors in the predictions.

**Normalized root mean squared error (NRMSE)** may be calculated as;

$$NRMSE = \frac{RMSE}{\max(Y) - \min(Y)} \quad (1.4)$$

where,  $\max(Y)$  and  $\min(Y)$  are the maximum and the minimum values of the actual activities in the data set, respectively, and  $RMSE$  is the traditionally calculated root mean squared error given by:

$$RMSE = \sqrt{\frac{PRESS}{n}} \quad (1.5)$$

with  $n$  being the number of molecules in the data set (training or test set).  $RMSE$  provides a measure of mean error in prediction. However, these values are a measure of absolute error and do not reflect the magnitude of the error relative to the range of distribution of the data. On the other hand  $NRMSE$  gives an estimate of the error as a fraction of the total range of activity values. An  $NRMSE$  value of more than 0.2 (20% of the total range of activity values) is considered too high (Roy *et al.*, 2016).

#### 1.4.2.5 Biological activity data:

The third component of QSAR models is the biological activity data of the known ligands of the target protein. PubChem (Wang *et al.*, 2014) is an open source data base by NCBI which has the experimentally determined biological activity values of the compounds against a wide variety of targets consolidated from the published literature. In this work the proposed novel QSAR methodologies have been studied using six sets of inhibitor compounds tested against targets in various diseases,



namely, cancer, HIV, neurological disorders, and malaria. The details and significance of these target systems (TSs) are given below.

#### **1.4.2.6 Target systems (TSs)**

##### **TS-1: 4-phenylpyrrolocarbazole derivatives as WEE1 inhibitors as anti-cancer compounds**

Eukaryotic cells have two DNA-damage sensitive check points, namely, G1-S and G2-M. The G1-S checkpoint is dysfunctional in the cancer types that lack the p53 signaling pathway which is responsible for the smooth functioning of the G1-S checkpoint. Such cancer cells rely on the G2-M check point for DNA damage repair (Parker and Piwnica-Worms, 1992; Mueller and Has-Kogan, 2015). Halting of cells at the G2-M is achieved by inhibiting the activity of cyclin-dependent kinase cdc2. In humans the inhibition of cdc2 is carried out by another tyrosine kinase, Wee1, by phosphorylation of cdc2 (Mueller and Has-Kogan, 2015). After the DNA repair is complete, the phosphatase cdc25 reactivates cdc2 by its dephosphorylation. Activation of cdc2 causes the cell cycle to advance to the M-phase where the cells divide. However, if the activity of Wee1 is inhibited it causes the cdc2 to remain active, forcing the cells to skip the G2-M checkpoint. This is desirable because in the cancer cells that skip the G2-M checkpoint, the damage that DNA has incurred during the previous phases of the cell-cycle remain unrepaired and over multiple cell-cycles the amplification of the damage causes cell death. Therefore, Wee1 inhibitors make p53-negative cancer cells skip the G2-M checkpoint and thereby also enhance the effects of other DNA-damaging agents used in the cancer treatment (Mueller and Has-Kogan, 2015).

4-phenylpyrrolocarbazole derivatives are known to competitively inhibit the activity of Wee1 kinase by binding to its ATP binding site (Smaill *et al.*, 2008). AZD1775 (formerly MK-1775) is a small-molecule, pyrazol-pyrimidine derivative which is a potent and ATP-competitive specific inhibitor of the Wee1 kinase. Several preclinical and clinical studies have demonstrated its encouraging antitumor effects

with manageable side effects with a combination of Wee1 inhibitor and DNA-damaging agent (Geenen and Schellens, 2017). As an example, in vitro and in vivo antitumor activity studies of MK1775, a potent pharmacological inhibitor of WEE1, as a single agent against ovarian cancer cells supports its use. It abrogated the G2-M checkpoint by inhibiting the phosphorylation of CDK1 that induced apoptosis in ovarian cancer cells that lacked mutations in p53 and breast cancer cells (BRCA1). To further substantiate, a significant antitumor effect of MK1775 was observed in C57BL/6 mice bearing syngeneic ID8 ovarian tumors (Zhang *et al.*, 2017). Therefore targeting of WEE1 which is expressed at high levels in various cancer types (breast cancers, leukemia, melanoma, and adult and pediatric brain tumors) to favorably disrupt the functioning of G2-M checkpoint presents an opportunity to potentiate cancer therapy. AZD1775 a potent WEE1 inhibitor has in fact advanced to clinical trials (Matheson *et al.*, 2016). Therefore, in the present work, we use 4-phenylpyrrolocarbazole derivatives (Palmer *et al.*, 2006) for performing proposed QSAR studies.

## **TS-2: Benzylpiperidine derivatives as acetylcholinesterase inhibitors for neurological disorders**

Acetylcholinesterase (AChE) carries out hydrolysis of acetylcholine at the cholinergic synapses and plays an important role in the termination of synaptic transmission. Depleted levels of acetylcholine are associated with neurological disorders like depression and Alzheimer's diseases. Mild inhibitors of AChE are therefore prescribed to treat the symptoms in the patients. Side effects and bioavailability problems in existing therapy deems it necessary to design new and more effective AChE inhibitors. Donepezil, a hydrochloride salt of benzylpiperidine derivative, is a commercially available AChE inhibitor (brand name: Aricept®). It acts by binding reversibly to the active site of AChE and competitively inhibits the hydrolysis of acetylcholine (Kryger *et al.*, 1999). We therefore choose benzylpiperidine

derivatives with their known experimentally determined  $pIC_{50}$  (Queiroz *et al.*, 2011) for the studies performed in this work.

### **TS-3: 2-substituted dipyrindodiazeponone as reverse transcriptase (RT) inhibitors for treatment of HIV-1**

Human immunodeficiency virus type 1 (HIV-1) is a retro-virus with single-stranded RNA as its genetic material. Therefore, to use the host (human) cell mechanism to express its proteins and replicate, the retrovirus first synthesizes DNA from its RNA using the enzyme RT present in the virus particle. An absence of RT in humans makes it an ideal target for HIV-1 treatment. Navirapine, a non-nucleoside RT inhibitor drug, is a dipyrindodiazeponone derivative prescribed for HIV-1 treatment. We therefore study 2-substituted dipyrindodiazeponone derivatives (Proudfoot *et al.*, 1995) for their RT inhibitory properties.

### **TS-4: 2-pyridinone as reverse transcriptase (RT) inhibitors for treatment of HIV-1**

It is seen that the high mutations rates in HIV-1 has resulted in the emergence of new resistant strains of the virus to current therapy. This suggests the need to develop new RT inhibitor molecules. In this context we also study another class of compounds, namely, 2-pyridinone derivatives (Wai *et al.*, 1993) for QSAR modelling.

### **TS-5: Cyclic urea inhibitors as protease (PR) inhibitors for treatment of HIV-1**

In the human cells the HIV proteins are expressed in the form of polyproteins and are cleaved at the processing sites by the HIV-1 PR to produce active proteins. This step is indispensable in the maturation of the virus particle and thereby makes HIV-1 PR an ideal target for drug design. HIV PR inhibitors, (e.g., saquinavir, indinavir, ritonavir, lopinavir) have been approved by the FDA but unfortunately most of them are accompanied by side effects during long-term treatment. Recent efforts have increasingly focused upon identifying non-peptidic cyclic urea HIV-1 PR inhibitors such as atazanavir. Cyclic urea ligands have been experimentally observed to inhibit HIV-1 PR by displacing the structural water molecule and blocking the amino acids at the active site (Ala *et al.*, 1998). Cyclic urea derivatives comprising of

seven-member ring structures (Debnath, 1999) were chosen for our QSAR modelling studies.

### **TS-6: Azilide derivatives as anti-malarial compounds**

Malaria, a mosquito-borne parasitic disease, is found throughout the tropical and subtropical regions of the planet and thereby putting at risk about 40% of the world population getting infected. Emergence of strains of malarial parasite *Plasmodium falciparum* resistant to various therapies in some parts of the world has raised serious concerns about the spread of such strains (Noedl *et al.*, 2009) and increased world-wide efforts to develop new classes of compounds for treatment and prevention/prophylaxis (Wells *et al.*, 2015). However, considering the chances of failure of new drug molecules in the later stages of development (clinical trials), it seems repurposing existing drugs to treat malaria may be considered as an efficient option. In this context, azithromycin is a widely prescribed (brand names: zithromax®, azithrocin®, etc.) azalide antibiotic with wide spectrum of anti-microbial activity and has also shown activity against *P. falciparum* (Rosenthal, 2016). Azithromycin has a favorable toxicological profile, long half-life and importantly is safe to use in children and pregnant patients (Gray *et al.*, 2001; Ke *et al.*, 2014). In bacteria, azithromycin acts by binding to its 50S ribosomal subunit to block protein synthesis while against malaria it has been reported to act in a similar manner by binding to the bacterial-like ribosome inside the Plasmodium organelle apicoplast (Sidhu *et al.*, 2007; Dahl and Rosenthal, 2008). The clinical studies of azithromycin have also shown promising results in various combination therapies throughout the world (format references). Although azithromycin, the first azalide investigated for malaria treatment and prophylaxis, showed potential as an anti-malarial molecule, its lower potency and efficacy as a single or combination agent has suggested the need for new azalide analogues that have improved activity (Hutinec *et al.*, 2011; Perić *et al.*, 2012; Rosenthal, 2016). In the current study 15-membered azalide derivatives (Hutinec *et*

al., 2011) were studied for their anti-malarial activity using the QSAR methodologies developed here.

## **Chapter 2**

# **Multidimensional scaling and Support vector regression based 2D-QSAR modelling**

## 2.1 Introduction:

2D descriptors for QSAR modelling usually refer to different topological parameters generated by applying graph theory principles to molecular graphs constructed from molecular structures (Roy and Das, 2014; Todeschini and Consonni, 2008). The advantage of these 2D descriptors are that they do not need pre-processing of the molecular structures and also do not require their structures to be aligned. Since these descriptors are generated using specific mathematical routines and use rudimentary structural information, they are easy to calculate (Roy and Das, 2014). The other type of 2D descriptors used in the QSAR modelling are the image based descriptors which are basically images of 2D structures of the compounds (Freitas *et al.*, 2005; Cormanich *et al.*, 2009; Daré *et al.*, 2018). The QSAR models developed using these descriptors employ multivariate image analysis where every pixel in an image is considered as a variable and is used to regress with the biological activity of the compounds. For the use of these descriptors, however, the molecular structures drawn as images need to be aligned, i.e., the atoms and bonds in the scaffold that are common across molecules need to be drawn at exactly the same pixel positions in every image (Freitas *et al.*, 2005). In these images pixels that represent the structure of the molecule are either active (i.e., have a value of one) or are inactive (i.e., having a value of zero). Thus these descriptors are binary in nature and do not contain any other chemical information about the molecules.

It has been noted that augmenting 2D descriptors with molecular properties can further improve the performance of the 2D-QSAR models (Roy and Das, 2014). In the work presented in this chapter, we develop novel 2D image based descriptors which include higher degree of chemical information as compared to the image based descriptors (Freitas *et al.*, 2005). For this purpose, we propose a novel methodology employing multidimensional scaling (MDS) analysis to the 3D structures of the molecules so as to project the positions of the atoms in 2D space (Bronstein *et al.*, 2006; Kuriakose *et al.*, 2004). We subsequently use Procrustes analysis to transform the atomic positions to align the atoms in the scaffold common to all the molecules

(Kendall, 1989). These transformed positions along with partial atomic charges of the atoms were used to generate novel 2D image based molecular descriptors. These descriptors may then be used for regression with the pIC<sub>50</sub> values for QSAR modelling purposes. Here, we study a hybrid regression approach employing Principal component analysis (PCA) (Geladi and Kowalski, 1986) for dimensionality reduction of the obtained high dimensional 2D image descriptors in combination with Support Vector Machine (SVM) (Smola and Schölkopf, 2004) for the regression study. The approach therefore takes into account the non-linear characteristics of the data for which SVM has been used extensively.

## **2.2 Methodology:**

### **2.2.1 Importing of chemical structures and biological activities of inhibitors:**

PubChemBioAssay (Wang *et al.*, 2014) is a large compendium of chemical compounds with experimentally tested values of their biological activities collated from the literature. We used PubChemBioAssay to identify and download in <.sdf> format 3D structures of chemical inhibitors (Table 2.1) having similar scaffolds as described for target systems TS-1 to TS-6, respectively. The biological activities of the chosen compounds, i.e., pIC<sub>50</sub> values, defined as the negative log of the inhibitor concentration (IC) required to reduce the activity of the target by 50%, were downloaded for the target systems under study. For brevity the structures of the inhibitor compounds chosen for the TS-1 to TS-6 are provided in the Appendix Tables A1 to A6.

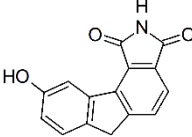
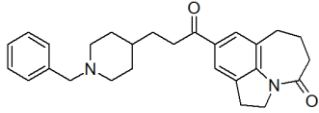
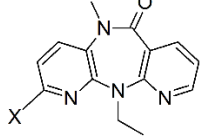
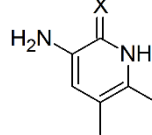
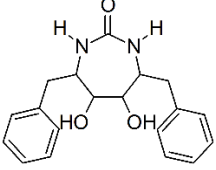
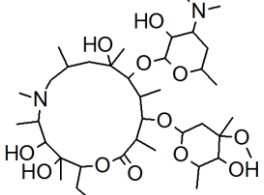
### **2.2.2 Optimization of molecular structures and calculation of partial atomic charges:**

The downloaded structures from PubChem were imported in Schrödinger<sup>®</sup> suite. These structures were optimized for their 3D structures using the LigPrep<sup>®</sup> module (version 2.5, 2012) (Schrödinger, LLC, 2011) of the Schrödinger suite. The optimized 3D structures were exported to files with <.pdb> format to obtain the 3D coordinates of each atom in a structure and also the pattern of bond connections between the atoms. The <.pdb> files were then imported to Matlab<sup>®</sup> software (version



R2010b) (MATLAB, 2010) for further computations. The partial atomic charges of the molecules were calculated by using Jaguar® module (version 7.8, 2012) (Bochevarov *et al.*, 2013; Schrödinger, LLC) of the Schrödinger suite. The optimized 3D structures obtained from LigPrep were used for calculations in Jaguar. The structures with calculated partial atomic charges were saved in Schrödinger suite's <.mae> file format. A script was written in Matlab to import the atomic charge data from <.mae> files into Matlab.

**Table 2.1:** Scaffolds of the inhibitor compounds studied with corresponding assay IDs (AID) for TS-1 to TS-6

TSs	Target	Compounds	Scaffolds	AID	Reference
1	Wee1	4-phenylpyrrolocabazoles		268838	Palmer <i>et al.</i> , 2006
2	AChE	Benzylpiperidines		566585	Queiroz <i>et al.</i> , 2011
3	HIV-1 RT	2-Substituted Dipyridodiazepones		198247	Proudfoot <i>et al.</i> , 1995
4	HIV-1 RT	2-Pyridinones		197804	Wai <i>et al.</i> , 1993
5	HIV-1 PR	Cyclic Ureas		160292	Debnath, 1999
6	Malaria	Azilides		579588	Hutinec <i>et al.</i> , 2011

### 2.2.3 Calculation of the shortest path distances between atoms using the connectivity network of the atoms:

Using the connectivity between the atoms every molecular structure was converted into a graph where each atom was considered as the node and the bonds connecting the atoms were considered as the edges of the graph. Using the 3D coordinates of the atoms the bond lengths were calculated using the distance formula as;

$$l_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (2.1)$$

Where,  $l_{i,j}$  is the length of the bond connecting the  $i^{\text{th}}$  and the  $j^{\text{th}}$  atoms in the molecule and  $(x_i, y_i, z_i)$  and  $(x_j, y_j, z_j)$  are the coordinates of  $i^{\text{th}}$  and the  $j^{\text{th}}$  atoms, respectively. These bond lengths were assigned as the lengths of the corresponding edges in the graphs.

Dijkstra's algorithm (Dijkstra, 1959) was used to determine the shortest paths between the nodes (atoms) along the connecting edges (bonds) of the graph. In order to find the shortest path from  $i^{\text{th}}$  node to  $j^{\text{th}}$  node in a graph the nodes and edges are transferred to three sets as follows.

*For nodes:*

Set-A: The nodes for which the shortest path from the  $i^{\text{th}}$  node is known.

Set-B: The nodes that are neighbouring (connected by single edge) to the nodes in Set-A, but do not belong to Set-A.

Set-C: Remaining nodes which do not belong to Set-A or Set-B.

*For edges:*

Set-I: Edges occurring in minimal path from  $i^{\text{th}}$  node to the nodes in Set-A

Set-II: Edges connecting the nodes in Set-A to their neighbouring nodes in Set-B.

Set-III: Edges that are rejected or are not yet transferred to Set-A or Set-B.

Note that initially all the nodes are put in Set-C and all the edges are put in Set-III. The  $i^{\text{th}}$  node is then transferred to Set-A and following two steps are repeatedly performed.

Step-1: Consider all the edges connecting the node that has been recently added to Set-A with its neighbouring nodes in Set-B or Set-C. If the neighbouring node is present in Set-C then it is transferred to Set-B and the corresponding edge is transferred to Set-II. If the neighbouring node is present in Set-B then the corresponding edge in Set-II is compared with the new edge and the one that provides a shorter path to the  $i^{\text{th}}$  node is retained in Set-II and the other is rejected and transferred to Set-III.

Step-2: The node in Set-B with minimum distance to  $i^{\text{th}}$  node is transferred to Set-A and the corresponding edge in Set-II is transferred to Set-I. Steps 1 and 2 are repeated until  $j^{\text{th}}$  node is added to Set-A and the shortest path between  $i^{\text{th}}$  and  $j^{\text{th}}$  nodes is obtained.

Thus, the inter-atomic distances,  $d_G(i,j)$ , between  $i^{\text{th}}$  and  $j^{\text{th}}$  atoms through the shortest path connecting them were calculated for all the pairs of  $i$  and  $j$  in a molecule to obtain a symmetric distance matrix,  $D_G$ , for every molecule using Dijkstra's algorithm. The Dijkstra's algorithm was implemented using Matlab's in-built function 'graphshortestpath'. Distance matrices for every molecule were used for carrying out MDS studies.

#### 2.2.4 Multidimensional Scaling (MDS):

MDS facilitates obtaining a lower dimensional projection of a system such that the distances between the nodes are preserved. MDS calculations were performed as described in the literature (Borg and Groenen, 2005; Kuriakose *et al.*, 2004). Thus, the  $D_G$  matrices obtained using Dijkstra's algorithm were used to calculate the 2D coordinates of the atoms of molecules by MDS. The distance matrices ( $D_G$ ) were double centred and squared to obtain matrix  $B$  with elements,  $b_{ij}$  as,

$$b_{ij}(d_G) = -\frac{1}{2} \left[ d_G^2(i,j) - \frac{1}{n} \sum_{k=1}^{n_a} d_G^2(k,j) - \frac{1}{n} \sum_{k=1}^{n_a} d_G^2(i,k) + \frac{1}{n^2} \sum_{g=1}^{n_a} \sum_{h=1}^{n_a} d_G^2(g,h) \right] \quad (2.2)$$

where  $d_G(i,j)$  is the  $i,j^{\text{th}}$  element of matrix  $D_G$  and  $n_a$  is the number of atoms in the molecule, so that the scalar product  $B = ZZ'$  can be defined.  $Z$  matrix contains the MDS components which minimize the cost function

$$E = \|B(d_G) - B(d_Z)\|_{L_2} \quad (2.3)$$

where,  $B(d_Z)$  is the squared double centred distance matrix obtained from the distances using the MDS coordinates and  $\|\cdot\|_{L_2}$  is the  $L_2$  norm. This minimization can be carried out by performing the singular value decomposition (SVD) of matrix  $B$ , i.e.  $B = WSW'$  where  $W$  is the eigen vector matrix and  $S$  is a diagonal matrix of singular values (Kuriakose *et al.*, 2004). Matrix  $Z$  is calculated as

$$Z = WS^{1/2} \quad (2.4)$$

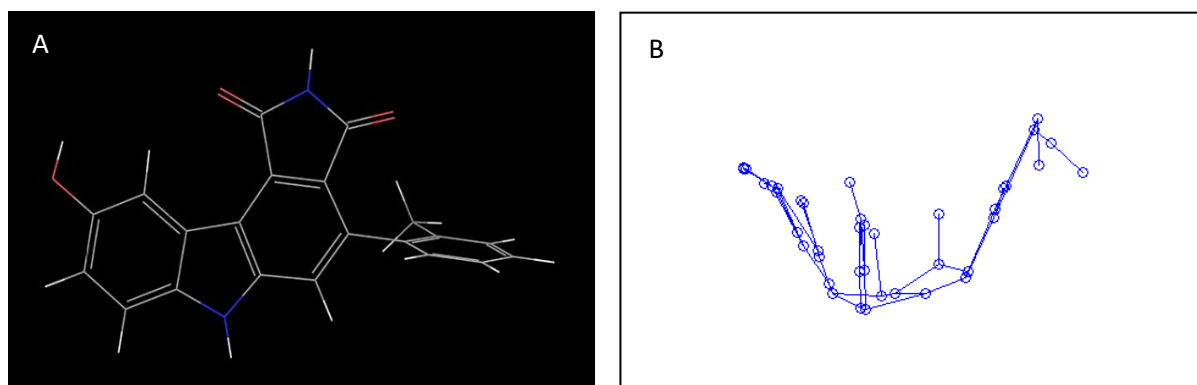
with  $z_{ij} = w_{ij}s_j^{1/2}$  where  $i, j=1, 2, \dots, n_a$ . The first two columns of matrix  $Z$  are the MDS coordinates of the atoms in the molecule projected in 2D.

### 2.2.5 Procrustes analysis:

Given two matrices ( $A$  and  $B$ ) of same size with one of them say  $A$  chosen as a reference, Procrustes transformation performs the translation, reflection, orthogonal rotation and scaling of the other matrix ( $B$ ) such that the sum of squared distance between the corresponding elements of  $A$  and  $B$  is minimized (Kendall, 1989; Andrade *et al.*, 2004). We performed Procrustes transformation to align the 2D MDS coordinates of the scaffold atoms of all the molecules and accordingly transformed the positions of the other atoms in a molecule. Matlab function 'procrustes' was used for this purpose and repositioned coordinates of the atoms obtained. These coordinates were used for creating the 2D images of the molecules.

### 2.2.6 Creation of 2D image based descriptors:

Depending on the coordinates of atoms in all the molecules in a given set, a 2D plane encompassing all the molecules was defined. Positions of atoms in a molecule were then marked on the selected plane (Figure 2.1) and an image representing that molecule was obtained. The images were scaled such that every pixel in the image represented an area of  $0.01\text{\AA}^2$ . The pixels in the images representing the position of an atom had a value of 1 and the other pixels had a value of zero, making these images binary. Repeating this process for every molecule generated a binary image of fixed



**Figure 2.1:** Multi-dimensional Scaling. Conversion of (A) 3D molecular structure to (B) 2D MDS graph of the molecule

size for every molecule. The binary images were then converted into grey scale images by replacing the values at the position of the atoms with the previously calculated partial atomic charges of respective atoms from Jaguar. At this point the images were divided into two sets, a training set and a test set. A 3-way array of images was generated for each set by stacking the images of that set one behind the other. Each image in this 3-way array was then converted into single row by placing every row in the image in front of the other. This generated two 2-way arrays denoted by  $X_{train}$  and  $X_{test}$  ( $X$  matrices). The matrix  $X_{train}$  was then subjected to the Principal Component Analysis.

### 2.2.7 Principal Component Analysis (PCA):

Principal component analysis is a method of dimensionality reduction (Geladi and Kowalski, 1986) where a matrix,  $X$  of size  $(n, m)$  (rank  $m$ ) is represented as a sum of outer products of vectors given by

$$\mathbf{X} = t_1 p'_1 + t_2 p'_2 + \dots + t_m p'_m = \mathbf{TP}' \quad (2.5)$$

Here,  $T = [t_1 \ t_2 \ \dots \ t_m]$  and  $P = [p_1 \ p_2 \ \dots \ p_m]$ , with  $t_i$  a score vector of  $n$  elements and representing the scores of data points along the  $i^{th}$  principal component. The  $p_i$  is loading vector with  $m$  elements and refers to the direction cosines of a unit vector along the direction of  $i^{th}$  principal component. The  $i^{th}$  principal component contributes to the  $i^{th}$  highest variance among the principal components to the total variance in  $X$ .

Thus, with increase in the value of  $i$  the contribution of the corresponding principal component to the variance in  $X$  decreases. A suitable number of principal components ( $a$ ) are chosen ( $a \leq m$ ) for dimensionality reduction (Eq. 2.5) such that the first  $a$  principal components result in minimal loss in explaining the data. Here, PCA of the training set matrix  $X_{train}$  was performed using the Non-Iterative Partial Least Squares (NIPALS) algorithm (Geladi and Kowalski, 1986) to obtain the score matrix  $T_{train}$  and the loading matrix  $P$ . The steps in the NIPALS algorithm to calculate the scores and loading matrices are illustrated in Table 2.2. The percentage contribution ( $C_a$ ) of first  $a$  components in explaining the total variation in  $X_{train}$  was calculated as,

$$C_a = 100 \left( \frac{\sum_{i=1}^{n_{train}} \sum_{j=1}^a (t_{ij})^2}{\sum_{i=1}^{n_{train}} \sum_{j=1}^m (x_{ij})^2} \right) \quad (2.6)$$

where,  $n_{train}$  is the number of compounds in the training set,  $a$  is the chosen number of principal components,  $t_{ij}$  is the  $ij^{th}$  element of matrix  $T_{train}$ ,  $m$  is the total number of pixels in the image and  $x_{ij}$  is the  $ij^{th}$  element of matrix  $X_{train}$ .

**Table 2.2:** Steps in calculation of PCA scores and loadings in NIPALS algorithm

Step no.	Description	Step
1	Start with $X_{train}$	$E = X_{train}$
2	Loop over chosen $a$	For $i = 1$ to $a$
3	Initial selection of score vector	$t_i =$ column of $E$ with max $\ x\ $
4	Compute loadings	$p'_i = t'_i E / t'_i t_i$
5	Normalizing loadings	$p_i = p_i / \ p_i\ $
6	Compute scores	$t_{i1} = E p_i / p'_i p_i$
7	Check for convergence	If $t_{i1} = t_i$ then go to step 8 Else $t_i = t_{i1}$ and go to step 4
8	Finalize $i^{th}$ scores and loadings and calculate residual	$t_i = t_{i1}$ , $p_i = p_i$ $E = E - t_i p'_i$
9	Next $i$	$i = i + 1$

Using the orthogonal  $P$  matrix obtained iteratively following NIPALS algorithm (Table 2.2), score matrix  $T_{test}$  for test set may be calculated for the test set from Eq. 2.5 as

$$T_{test} = X_{test}P \quad (2.7)$$

$T_{train}$  matrix from the training set was used for optimizing the parameters for Support Vector Regression (SVR) and  $T_{test}$  was used for the external validation.

### 2.2.8 Support Vector Regression (SVR):

Support Vector Regression (Vapnik, 1999; Smola and Schölkopf, 2004) was used for regressing the score matrix,  $T_{train}$ , against the pIC<sub>50</sub> values  $Y_{train}$  considered as the dependent variable. SVR is a machine learning technique capable of handling the non-linear data properties during the regression. In SVR a linear estimation is performed on the data which is expanded in a higher dimensional feature space using a feature map ( $\phi$ ):  $X \rightarrow \phi(X)$  as,

$$\hat{y}_j = \sum_{i=1}^{n_{train}} (\alpha_i^+ + \alpha_i^-) \phi(x_i)' \phi(x_j) + b \quad (2.8)$$

and,

$$b = \frac{1}{n_s} \sum_s [y_s - \varepsilon - \sum_{m \in S} (\alpha_m^+ - \alpha_m^-) \phi(x_m)' \phi(x_s)] \quad (2.9)$$

where,  $n_{train}$  is the number of compounds in the training set,  $\alpha^+$  and  $\alpha^-$  are the Lagrange multipliers,  $S$  is the set of support vectors,  $n_s$  is the number of support vectors,  $\hat{y}$  is the predicted values of the dependent variable,  $x_j$  is the independent variables of compounds for which  $\hat{y}$  is to be predicted,  $y$  is the actual value of the dependent variable and  $\varepsilon$  is the insensitivity of the loss function. In the above equations, the dot product,  $\phi(x_i)' \phi(x_j)$  can be replaced by a kernel function  $K$  whose  $ij^{th}$  element is given by:

$$K_{ij} = \phi(x_i)' \phi(x_j) \quad (2.10)$$

A kernel implicitly determines both a non-linear mapping and the corresponding inner product (Steve R. Gunn, 2010). SVR was performed with radial basis function (RBF) as the kernel function given by

$$K(x_i, x_j) = e^{-\left(\|x_i - x_j\|^2 / 2\sigma^2\right)} \quad (2.11)$$

where  $\sigma$  is the kernel parameter that may be appropriately chosen for obtaining the best fit to the regression. Thus, this leads to a final regression formula to evaluate dependent variable  $\hat{y}_j$ , from the independent variables  $x_j$ , as.

$$\hat{y}_j = \sum_{i=1}^{n_{train}} (\alpha_i^+ + \alpha_i^-) K(x_i, x_j) + b \quad (2.12)$$

and, 
$$b = \frac{1}{n_s} \sum_s [y_s - \varepsilon - \sum_{m \in S} (\alpha_m^+ - \alpha_m^-) K(x_m, x_s)] \quad (2.13)$$

It may be clarified that since score vectors ( $t_i$ ) from  $T_{train}$  and  $T_{test}$  were used as the independent variable for SVR the notation  $x$  in equations 2.8 through 2.13 is to be replaced with  $t$  and is not the same as the vectors in  $X$  matrices of the descriptors. To maintain the familiarity in the in the descriptions of SVM in literature these notations have been retained. SVM toolbox (Steve R. Gunn, 2010) from Matlab file exchange was used to perform SVR. Note that the score vectors,  $T_{train}$ , of training set were used as the independent variables and the corresponding pIC<sub>50</sub> values were used as the dependent variables as input to SVM solver to determine optimized values of  $\alpha^+$  and  $\alpha^-$  which minimize the residual error in the calculation of the dependent variable (Eq. 2.12 and 2.13). Selection of principal components for optimal predictions of pIC<sub>50</sub> values was performed using the feature selection process explained in Section 2.2.9. Leave-one-out cross-validation (Wold, 1978) was performed to get cross-validated predictions of the training set compounds. Prediction of pIC<sub>50</sub> values for test set,  $\hat{Y}_{test}$ , was then carried out by replacing  $x_j$  in Eq. 2.12 and Eq. 2.13 with the selected principal components from  $T_{test}$  for validating the model.

### 2.2.9 Feature selection for SVR:

Choosing the best combination of principal components for regression was carried out as outlined below. We first build SVR models for all the pairs of score vectors, i.e., ( $t_i, t_j$ ) where  $i, j = 1, 2, \dots, a$  with  $i \neq j$ , of the first  $a$  principal components of the training set. These models were then used to predict pIC<sub>50</sub> values  $\hat{Y}_{train}$  and  $\hat{Y}_{test}$  for



training and test sets, respectively. The pair of score vectors whose SVR model best predicted the training and test set  $pIC_{50}$  values were thus selected. In the next step, SVR was performed by including a score vector from the remaining principal components to the originally selected pair.  $\hat{Y}_{train}$  and  $\hat{Y}_{test}$  were again predicted and the score vector that best improved the  $\hat{Y}_{train}$  and  $\hat{Y}_{test}$  prediction was selected. In this fashion, the scores of principal components which improved the prediction were added one-at-a-time to the combination of the selected components. Note that the score vectors of only training sets were used to build the SVR models at every step while the corresponding score vectors of the test set were used for SVR model validation. At every step SVR was performed for a range of  $\sigma$  values (Eq. 2.11) to get best prediction of  $pIC_{50}$  values. Table 2.3 lists the selected components and the  $\sigma$  values for each of the target systems.

**Table 2.3:** Selected PCA components and  $\sigma$  values for SVR regression

TS	Target	$n_{train}^\dagger$	$n_{test}^\ddagger$	components	$\sigma$
1	Human tyrosine kinase Wee1	83	14	2, 7, 13, 44	4.5
2	Human Acetylcholinesterase	49	11	9, 33, 39	1
3	HIV-1 Reverse transcriptase	58	10	9, 16, 34, 36	2.5
4	HIV-1 Reverse transcriptase	60	12	16, 24, 29, 34, 35	8.5
5	HIV-1 Protease	70	14	8, 14, 19, 28	7.5
6	Anti-malarial azalides	83	15	5, 14, 31, 38	1.5

$\dagger$ - number of training set compounds

$\ddagger$ - number of test set compounds

### 2.2.10 Calculations of goodness of fit parameters:

Three measures of goodness of fit were calculated in order to measure the model performance, namely, correlation coefficient ( $r$ ), normalized root mean squared error ( $NRMSE$ ), and coefficient of determination ( $R^2$ ). Correlation coefficient,  $r$ , for the prediction of  $pIC_{50}$  values of given set (training or test) of compounds was calculated as,

$$r = \frac{1}{n_{mol}-1} \sum_{i=1}^{n_{mol}} \frac{(y_i - \mu_y)(\hat{y}_i - \mu_{\hat{y}})}{std(Y) std(\hat{Y})} \quad (2.14)$$

where  $n_{mol}$  is the number compounds in the given set,  $y_i$  and  $\hat{y}_i$  are the experimental and predicted pIC<sub>50</sub> values with  $i = 1, 2, \dots, N$ ,  $\mu_y$  and  $\mu_{\hat{y}}$  are the means of  $y_i$  and  $\hat{y}_i$  respectively and  $std(Y)$  and  $std(\hat{Y})$  are the standard deviations of the experimental and predicted pIC<sub>50</sub> values of the given set. The normalized root mean squared error (NRMSE) for a gives set was calculates as,

$$NRMSE = \frac{1}{\max(Y) - \min(Y)} \sqrt{\frac{\sum_{i=1}^{n_{mol}} (y_i - \hat{y}_i)^2}{n_{mol}}} \quad (2.15)$$

where  $\max(Y)$  and  $\min(Y)$  are the maximum and minimum values of the experimental pIC<sub>50</sub> values in the study, respectively. Similarly coefficient of determination,  $R^2_{cv}$ , for cross-validation of the training set was calculated using the formula (Li *et al.*, 2017),

$$R^2_{cv} = 1 - \frac{\sum_{i=1}^{n_{train}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{train}} (y_i - \mu_{train})^2} \quad (2.16)$$

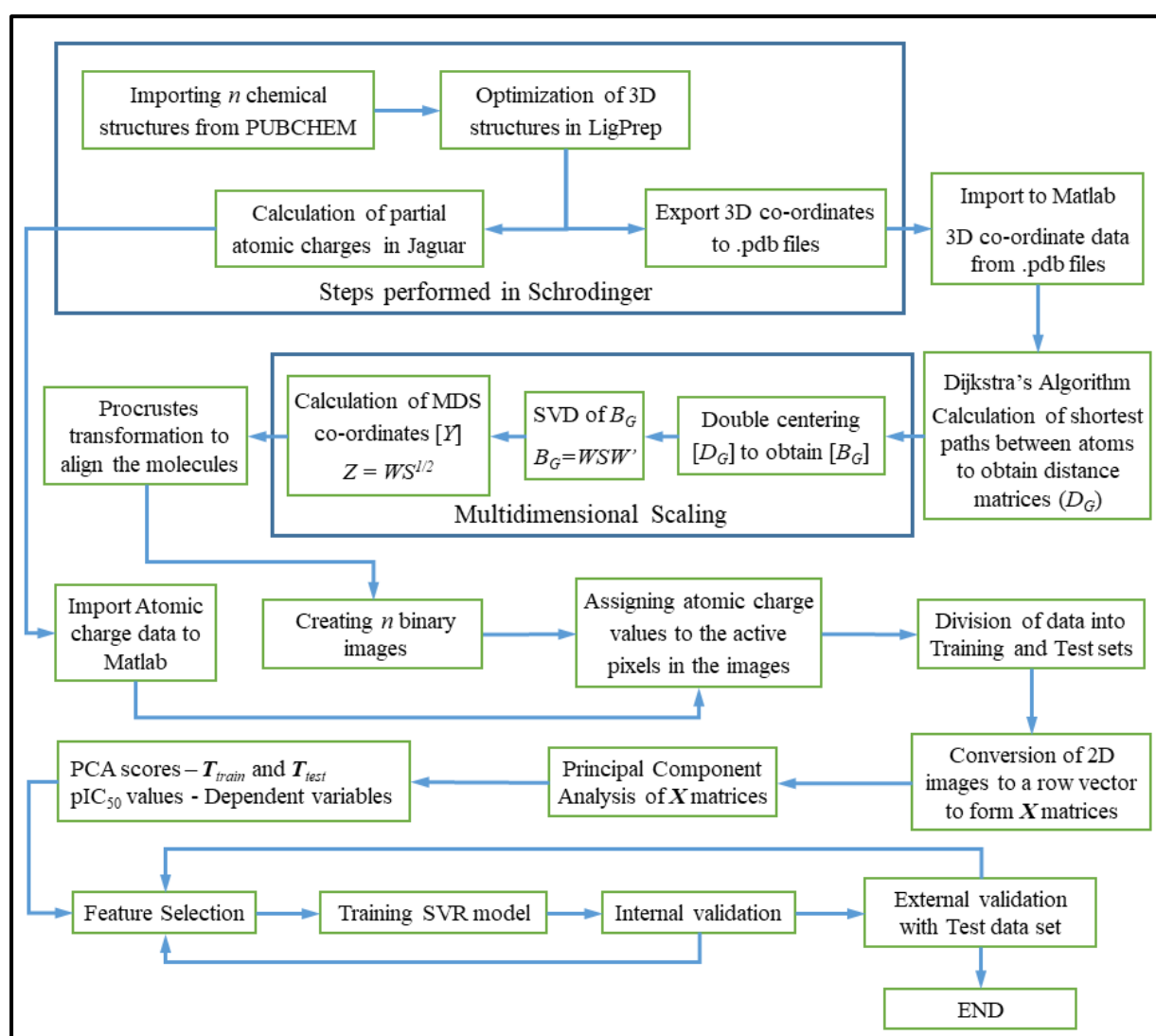
where  $y_i$  and  $\hat{y}_i$  are the experimental and predicted pIC<sub>50</sub> values of the training set,  $\mu_{train}$  is the mean of the observed pIC<sub>50</sub> values and  $n_{train}$  is the number of molecules in the training set. For external validation coefficient of determination for test set,  $Q^2_{ext(F1)}$ , was obtained as (Li *et. al.*, 2017),

$$Q^2_{ext(F1)} = 1 - \frac{\sum_{i=1}^{n_{test}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{test}} (y_i - \mu_{train})^2} \quad (2.17)$$

where  $y_i$  and  $\hat{y}_i$  are the experimental and predicted pIC<sub>50</sub> values of the test set and  $n_{test}$  is the number of molecules in the test set.

## 2.3 Results and discussion:

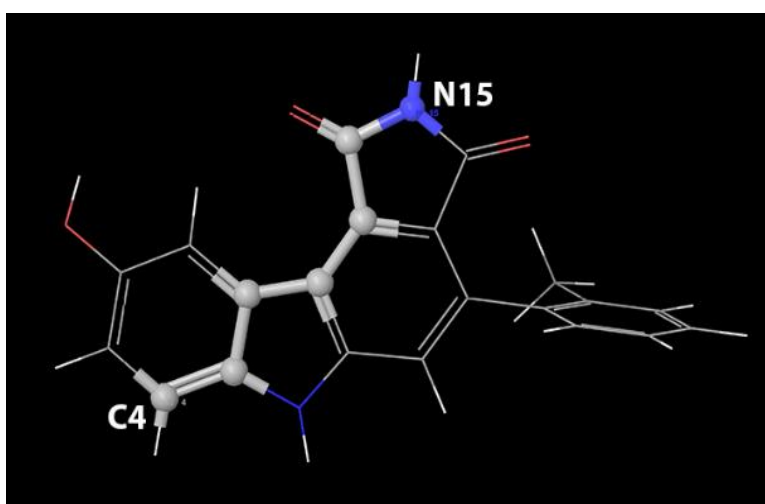
The aim of the study was to obtain image based descriptors incorporating chemical information present in 3D structures of the molecules for 2D-QSAR modelling. We do this by first applying the graph theory principles to the 3D structures of the molecules so as to obtain the shortest path distances between the atoms and then employing multidimensional scaling to obtain a coordinate representation of the atoms in 2D space. These 2D-MDS coordinates were initially used to generate binary images to which chemical information in the form of partial atomic charges was added to create image based 2D molecular descriptors. Figure 2.2



**Figure 2.2:** Flowchart of steps involved in generation of 2D image based descriptors and developing 2D-QSAR models using these descriptors

describes the schematics of the steps involved in the calculation of these image based 2D descriptors and developing 2D-QSAR models.

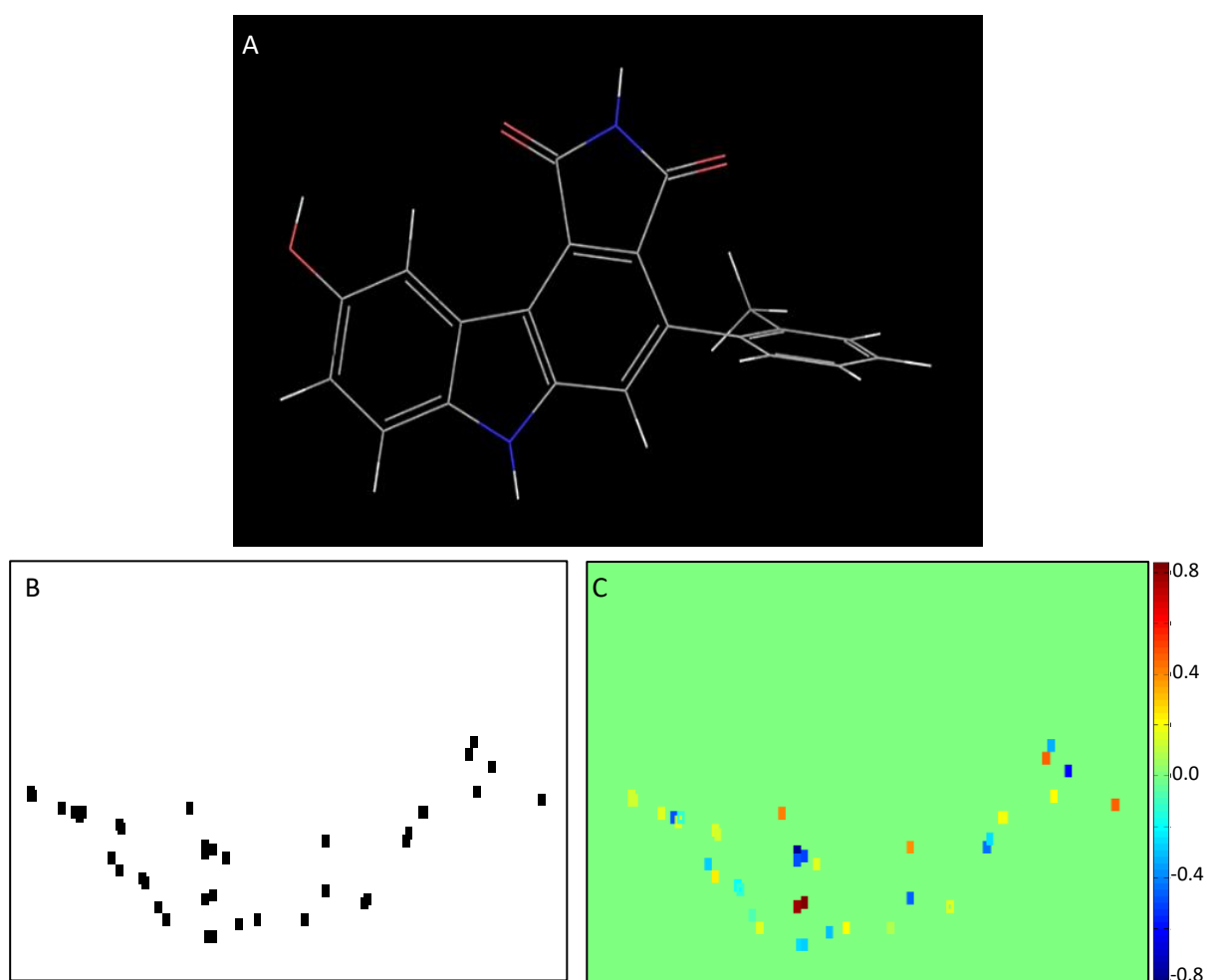
In a molecule the influence of one atom on any other atom is conveyed through the interatomic bonds that make the connection between them. More the number of bonds in a connection between the two atoms less is their effect on one another. Thus it can be assumed that the effective distance between the two atoms is the total minimum distance covered by the connecting bonds. The shortest path interatomic distances were therefore used for generating the distance matrix ( $D_G$ , Section 2.2.3) instead of the direct straight line distances in 3D space. Figure 2.3 illustrates an example of the shortest path calculated between two atoms of a molecule. Image based 2D descriptors were generated using the 2D MDS co-ordinates and the partial atomic charges. Figure 2.4 illustrates the binary and colour coded grey scale images of one of the 4-phenylpyrrolocarbazole derivative inhibitors of Wee1 (TS-1). The images and corresponding  $pIC_{50}$  values of molecules were divided into training and test sets. Each image was then converted into a row vector to form  $X$  matrices  $X_{train}$  and  $X_{test}$  for the training and test sets, respectively. Table 2.4 summarizes the image sizes for each of the target systems and the sizes of the corresponding row vectors. These  $X$  matrices were then used as the independent variables for developing the 2D-QSAR model.



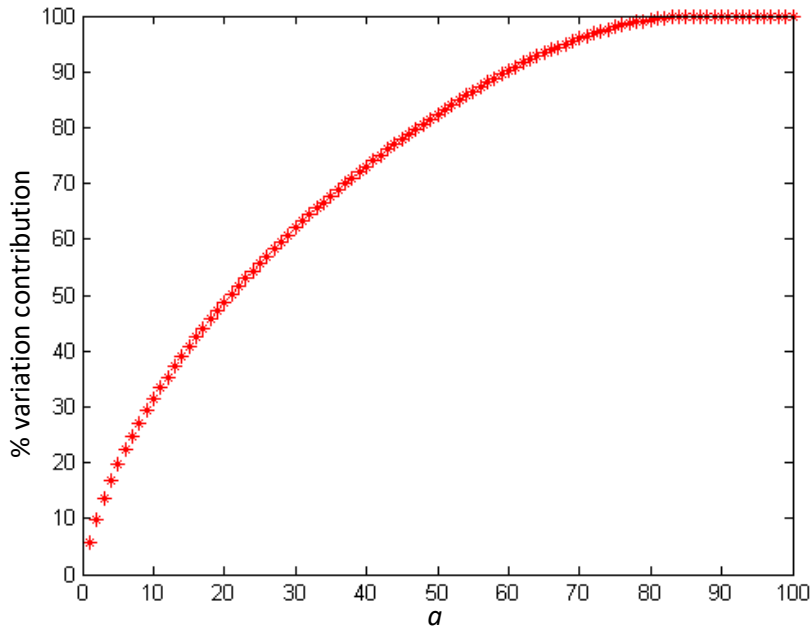
**Figure 2.3:** Highlighted bonds show the shortest path between atoms C4 and N15 calculated using Dijkstra's algorithm for one of the Wee1 inhibitors.

**Table 2.4:** The 2D image sizes and corresponding row vectors for TS-1 to TS-6

TS no.	Compounds	2D Image size	Row vector size
1	4-phenylpyrrolocarbazoles	300x400	1x120,000
2	Benzylpiperidines	470x200	1x94,000
3	2-Substituted Dipyrindodiazepones	300x200	1x60,000
4	2-Pyridinones	400x270	1x108,000
5	Cyclic Ureas	500x450	1x225,000
6	Azilides	1220x2470	1x3,013,400



**Figure 2.4:** (A) 3D structure of the compound its (B) Binary image on plotting the MDS coordinates on to 2D plane. The active pixels with value one are black in color whereas the inactive pixels are white in color. (C) Gray scale image generated after plotting the partial atomic charges at the active pixels in (B). The size of active pixels has been increased in image (B) and similarly the gray scale image in (C) has been colour coded according to the pixel value and the active pixel size increased for better representation.



**Figure 2.5:** Percentage of variation captured in the principal components for TS-1

Since  $X$  matrices have a dimensionality of the order  $10^5$  and most of the pixels in the images have a value of zero, PCA was performed to reduce the dimensions of the data to a relevant number. The PCA components were calculated using the NIPALS algorithm (Geladi and Kowalski, 1986) which calculates principal components one at a time in the order of their rank, i.e., first principal component is calculated first then the second component and so on. This circumvents the need to calculate the covariance matrix whose storage would require very high computational memory. For TS-1 it was observed that first 80 components captured about 98% of variation in the  $X$  matrices (Figure 2.5). Hence, first 80 components ( $a = 80$ ) were used for feature selection (Section 2.2.9) for TS-1.

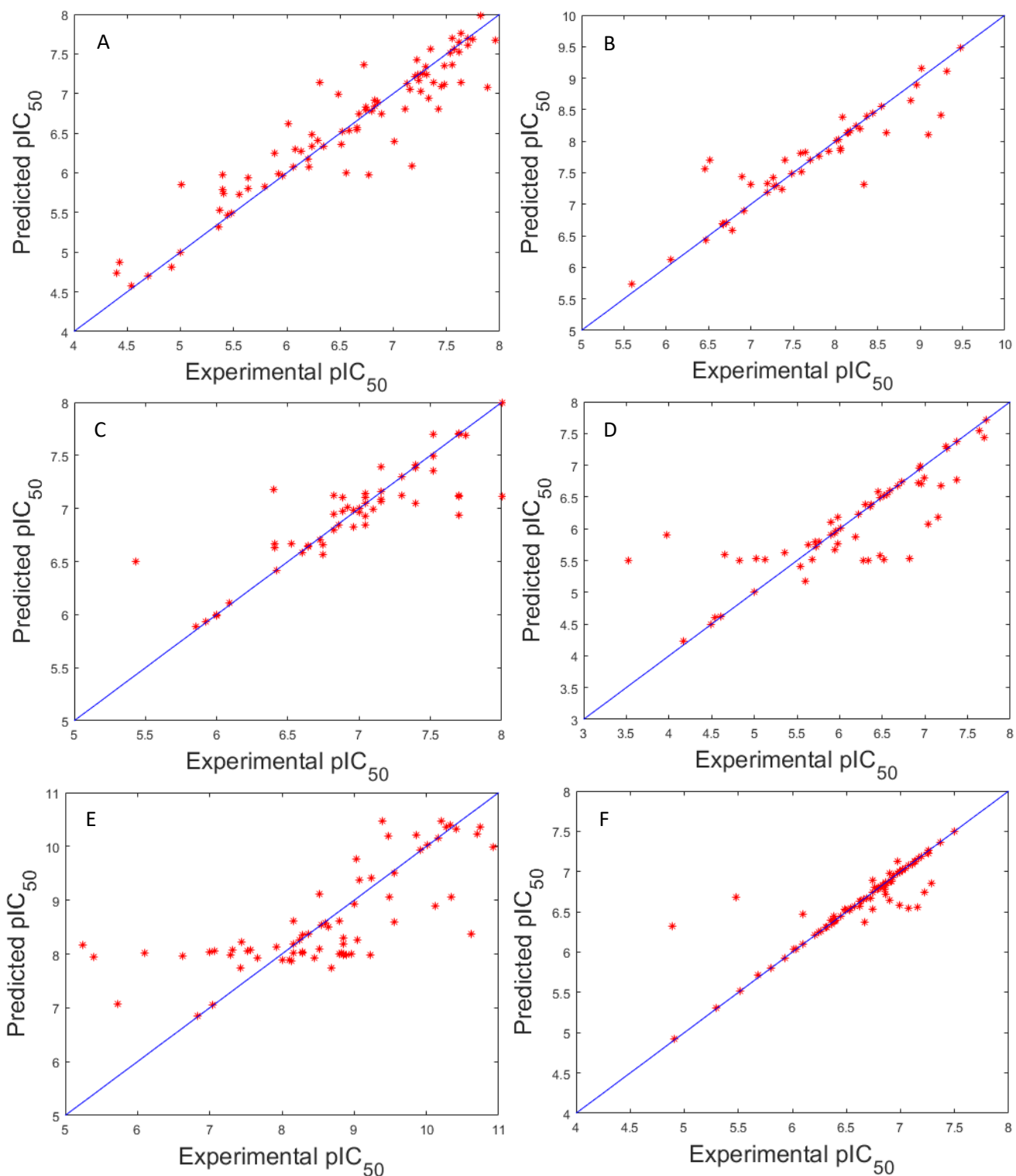
Support vector machine regression (Section 2.2.8) was next used to regress the principal components of  $X$  matrices against the  $pIC_{50}$  values as discussed in Section 2.2.9. Initially a combination of first  $a$  principal components was used for regression for increasing  $a$ . It was observed that for  $a \leq 6$  the model did not yield any significantly correlated predictions, however, for  $a \geq 7$  components led to overfitting of the model, i.e., the predictions for the training set improved with increasing value of  $a$  whereas those for the test set did not. The best combination of principal components for

regression was therefore selected by employing the feature selection procedure as described in Section 2.2.9. It was seen that for TS-1 using component numbers 2, 7, 13 and 44 the 2D-QSAR model for leave-one-out cross validation of training set yielded a Pearson’s correlation coefficient,  $r_{cv}$  of 0.94, coefficient of determination  $R^2_{cv}$  of 0.87 and  $NRMSE$  for cross-validation ( $NRMSECV$ ) of 0.09. Similarly, for test set predictions the 2D-QSAR model yielded an  $r_{pred}$  of 0.81,  $Q^2_{ext(F1)}$  of 0.67 and  $NRMSE$  for prediction ( $NRMSEP$ ) of 0.13. Similarly, the above mentioned model performance parameters and the components used for regression for the TS-1 to TS-6 are given in Table 2.5. Figures 2.6A to 2.6F show a diagonal plot of  $Y_{train}$  vs.  $\hat{Y}_{train}$  and Figures 2.7A to 2.7F show a diagonal plot of  $Y_{test}$  vs.  $\hat{Y}_{test}$  for TS-1 to TS-6, respectively.

**Table 2.5:** 2D-QSAR model performance statistics for TS-1 to TS-6

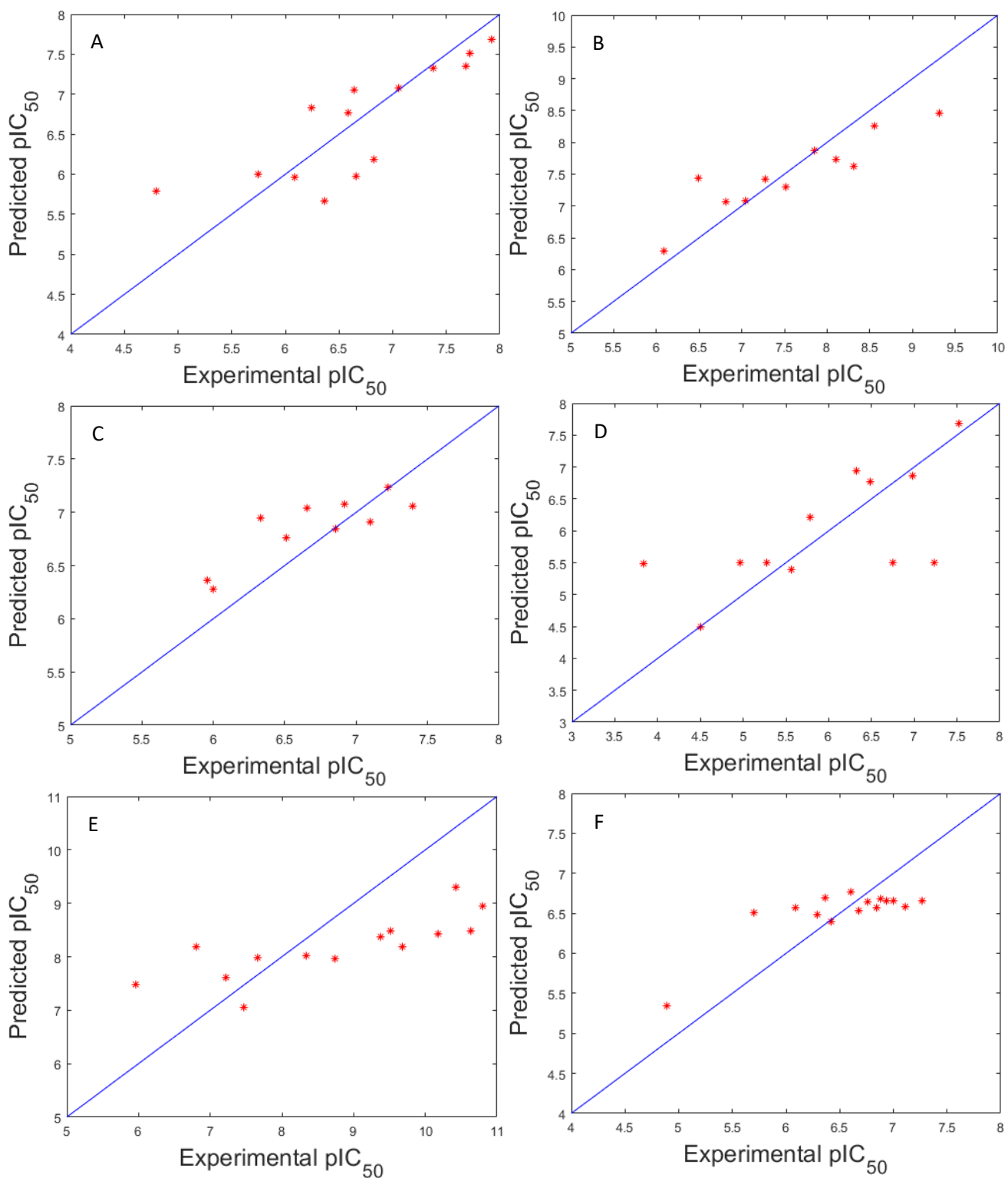
TS	AID	Components	Cross-validation			External Validation		
			$r_{cv}$	$R^2_{cv}$	$NRMSECV$	$r_{pred}$	$Q^2_{ext(F1)}$	$NRMSEP$
1	268838	2, 7, 13, 44	0.94	0.87	0.09	0.81	0.67	0.13
2	566585	9, 33, 39	0.91	0.83	0.09	0.90	0.73	0.12
3	198247	9, 16, 34, 36	0.88	0.72	0.11	0.89	0.67	0.12
4	197804	16, 24, 29, 34, 35	0.78	0.65	0.14	0.66	0.43	0.20
5	160292	8, 14, 19, 28	0.79	0.56	0.15	0.78	0.30	0.22
6	579588	5, 14, 31, 38	0.88	0.77	0.10	0.79	0.58	0.15

Performance of the 2D-QSAR model was observed to be satisfactory for TS-1, TS-2, TS-3 and TS-6 with  $r_{pred}$  values from 0.79 to 0.90,  $Q^2_{ext(F1)}$  values between 0.58 to 0.73 and a normalized  $NRMSEP$  values from 0.12 to 0.15. It was observed that the 2D-QSAR models did not perform well for TS-4 and TS-5, as reflected in the high normalized  $NRMSEP$  values and relatively low  $Q^2_{ext(F1)}$  values. Although, it may be noted that the Pearson’s correlation coefficient for these TSs was still relatively high indicating that the overall trend of  $pIC_{50}$  values was being captured by these models.



**Figure 2.6:** Plots of actual  $pIC_{50}$  values ( $Y_{train}$ ) vs. the predicted values ( $\hat{Y}_{train}$ ) using image-based 2D-QSAR model for cross-validation. (A) TS-1, (B) TS-2, (C) TS-3, (D) TS-4, (E) TS-5 and (F) TS-6 inhibitors.





**Figure 2.7:** Plots for actual  $pIC_{50}$  values ( $Y_{test}$ ) vs. the predicted values ( $\hat{Y}_{test}$ ) using image-based 2D-QSAR model for test sets of (A) TS-1, (B) TS-2, (C) TS-3, (D) TS-4, (E) TS-5 and (F) TS-6 inhibitors.

It may be remarked that the computational time involved in obtaining the SVR regression parameters ( $\alpha^+$  and  $\alpha^-$ ) for every component combination with varying values of  $\sigma$  during feature selection was observed to be very high and overall in the order of days. This is due to the intensive optimization calculations being performed for large number of iterations during the adopted feature selection process for the considered data. The experimental and predicted pIC<sub>50</sub> values using image-based 2D-QSAR models of all the molecules of TS-1 to TS-6 are given in Appendix Tables A8 to A13.

## 2.4 Conclusions:

We were successful in creating 2D image based descriptors by applying graph theory principles to the optimized 3D structures of the molecules. These images used the partial atomic charge values obtained from the 3D structures of the molecules. Thus, the aim of incorporating chemical information from the optimized 3D structures of the molecules in 2D image based descriptors was achieved. These image based descriptors when used for regression against pIC<sub>50</sub> values of the compounds produced 2D-QSAR models that performed well with good prediction accuracy for four of the six target systems (TS-1, 2, 3 and 6) for which these descriptors were studied. For the two target systems (TS-4 and 5) the model prediction accuracy was low but the general trend of the pIC<sub>50</sub> values was captured reasonably well, i.e., compounds with high pIC<sub>50</sub> values were observed to be on the higher side of the prediction scale and those on with low pIC<sub>50</sub> were found to be on the lower side of the prediction scale, thus proving the potential of these descriptors.

These obtained image based descriptors, however, suffered from two drawbacks. Firstly, the pixel density required for the image to adequately resolve the positions of atoms renders the final image with a high total number of pixels. This results in the descriptor dimensionality to be of the order 10<sup>5</sup> which is larger than the dimensionality of the 3D molecular field descriptors used in comparative molecular field analysis (CoMFA). This problem was, however, solved by dimensionality reduction using

PCA. The second drawback observed was that the feature selection process for SVR was computationally intensive. The computational time required for arriving at the combination of principal components for regression was in the order of days making the building of these 2D-QSAR models computationally intensive. Hence more efficient algorithms need to be studied for QSAR modelling both with respect to the choice of descriptors and lowering computational times for regression. These issues are addressed for the descriptors and regression methods studied in the subsequent Chapters 3 and 4.

## **Chapter 3**

# **QSAR modelling with pseudo-molecular field descriptors for potential applications in drug design**

### 3.1 Introduction:

QSAR modelling using 3D molecular field descriptors have been widely used to capture the relationship between a ligand and its biological activity (Nidhi and Siddiqi, 2013; Divakar and Hariharan, 2015). In particular, comparative molecular field analysis (CoMFA) uses 3D molecular descriptors (Cramer *et al.*, 1988; Dasoondi *et al.*, 2008; Matta and Arabi, 2011) that are developed by obtaining energy minimized 3D structures of the molecules along with the partial atomic charges calculated for every atom of the molecule. The molecular structures are oriented to structurally align with each other in a box of appropriate size having a suitably chosen 3D mesh grid. Molecular fields, such as, electrostatic and/or steric, are then calculated for all the points on the grid using coulomb potential function and Lennard-Jones potential function, respectively (Cramer *et al.*, 1988), and a 3D array of field values is obtained for every molecule. The above 3D arrays are used as molecular descriptors to develop regression models that correlate with the biological activity of the molecules. Although, the CoMFA based 3D-QSAR models relate the structural information with the activities of molecules, the structural minimization routines required for calculation of partial atomic charges are intensive (Gasteiger and Marsili, 1980). Thus, there is a need to study novel and simpler 3D molecular descriptors that provide accurate 3D-QSAR models for practical purposes. Towards this aim, here we propose and study the use of intrinsic properties of the individual atoms, namely, electronegativity and electron affinity values to develop and study 3D molecular field like descriptors. We term this molecular field as the pseudo-molecular field (PMF) and the molecular descriptors as pseudo-field molecular descriptors (PFMD). These descriptors would have the advantage that the atomic property values used in their calculations will be readily available and would not require to be determined for every molecule unlike partial atomic charges. Developing QSAR models based on these PFMDs would then be simpler than CoMFA based models and studying its feasibility would provide a practical and correlative way of using intrinsic atomic properties for assessing the activity of a ligand with its target. It may be noted that these PFMDs are

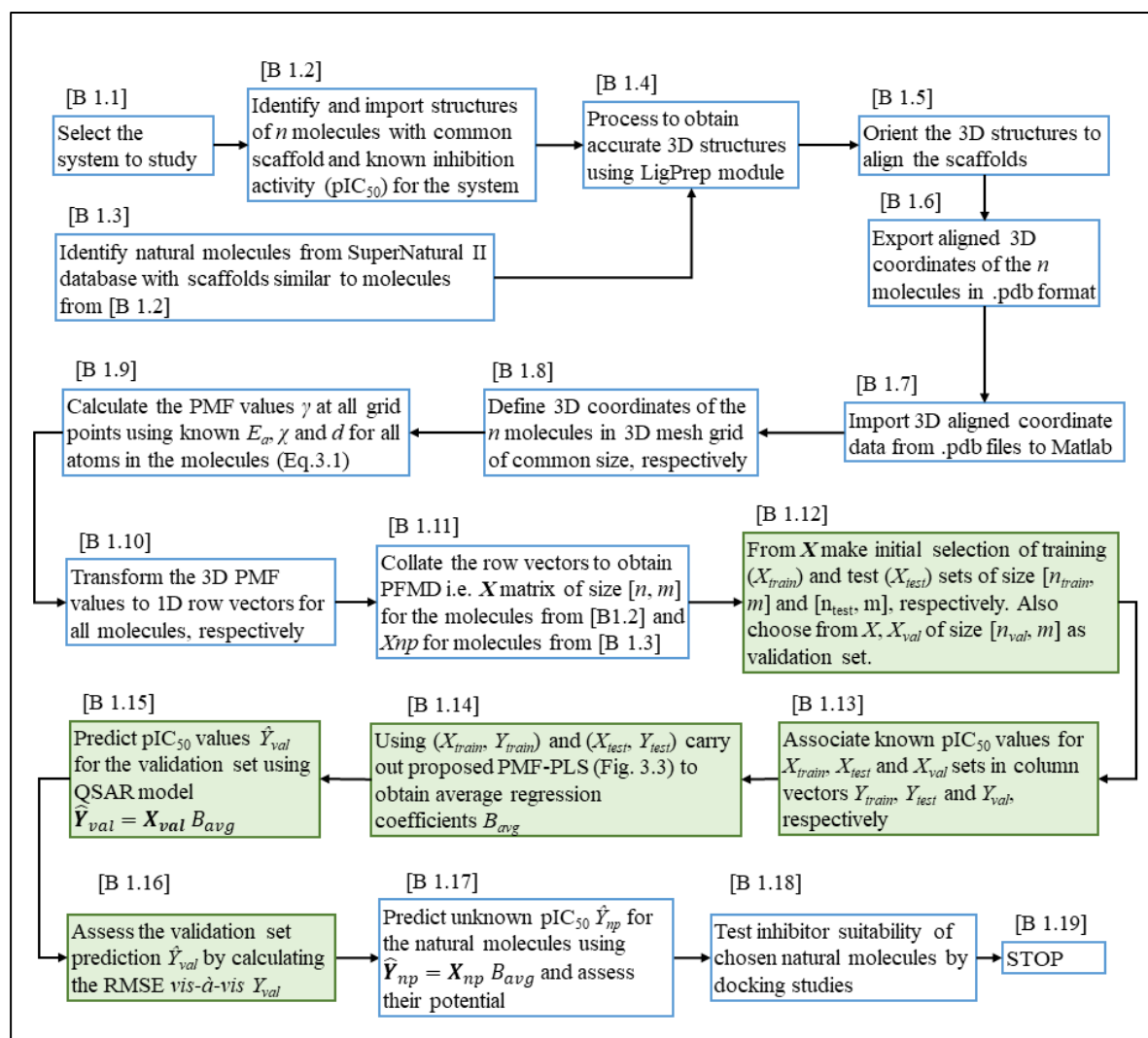
associated with high dimensionality (similar to the CoMFA descriptors) because of the consideration of PMF values in 3D spatial coordinates. We aim to bring out and discuss here a novel methodology employing PLS, namely PMF-PLS, for efficient QSAR modelling.

In the present study, we describe the PFMDs and test them by developing PMF-PLS QSAR models for six TSs introduced in Chapter 1. By applying PMF-PLS modelling strategy to a wide variety of TSs we show the generality of the proposed methodology.

### 3.2. Methodology

For ease in discussion, a schematic flowchart of steps involved in development of PFMDs and PMF-PLS QSAR modelling are shown in Figure 3.1 with boxes labelled as [B #.#] that refer to [B Fig. # . Box #]. The details of steps in the individual boxes are discussed in subsections for modularizing the algorithm.

Section 3.2.1 describes the procedure to import molecular structures and their biological activity values from PubChemBioAssay database [B 1.1] and the procedure to identify natural molecules from SuperNatural II (Banerjee *et al.*, 2015) database having scaffolds similar to the ones used in the chosen TS but whose pIC<sub>50</sub> values are not known [B 1.3]. Thus, the inhibitory activities of these natural compounds could be studied using the PMF-PLS QSAR modelling. Section 3.2.2 outlines two steps of pre-processing of the inhibitor structures that are obtained from the databases. Firstly, we pre-process the structures using Ligprep<sup>®</sup> module (version 2.5, 2012) in Schrodinger<sup>®</sup> software to obtain scaffold based alignment of 3D structures of the chosen inhibitor molecules [B 1.4]-[B 1.6]. In the next step [B 1.7]-[B 1.8], we import the aligned data into Matlab<sup>®</sup> (version R2010b) where the atoms in the molecule are accurately positioned in a 3D mesh grid.



**Figure 3.1:** General flowchart to study the proposed PMF-PLS approach for QSAR modelling

Section 3.2.3 describes the calculation of PMF values at the mesh grid points [B 1.9] using electron affinity and electronegativity values of atoms to obtain the PFMDs [B 1.10]-[B 1.11]. Subsequently, Section 3.2.4 elucidates the steps in PMF-PLS algorithm that are used to develop the QSAR model and its validation [B 1.12]-[B 1.16]. We next use this model to calculate the pIC<sub>50</sub> values of the natural molecules obtained from the SuperNatural II database (Banerjee *et al.*, 2015) [B 1.17]. To further confirm the potential inhibitory actions of natural molecules with the calculated pIC<sub>50</sub> values, it is proposed to carry out docking studies of these molecules to confirm that they indeed bind to the selected targets. Successful docking along with the prediction of

high pIC<sub>50</sub> value by the QSAR model would suggest that the molecule has a good potential for inhibiting the target [B 1.18].

### **3.2.1 Importing of inhibitor structures**

#### **3.2.1.1 Compounds with known biological activity**

The structures for the compounds in TS-1 to TS-6 were downloaded from PubChemBioAssay as described in Section 2.2.1 (Chapter 2).

#### **3.2.1.2 Compounds with unknown biological activity**

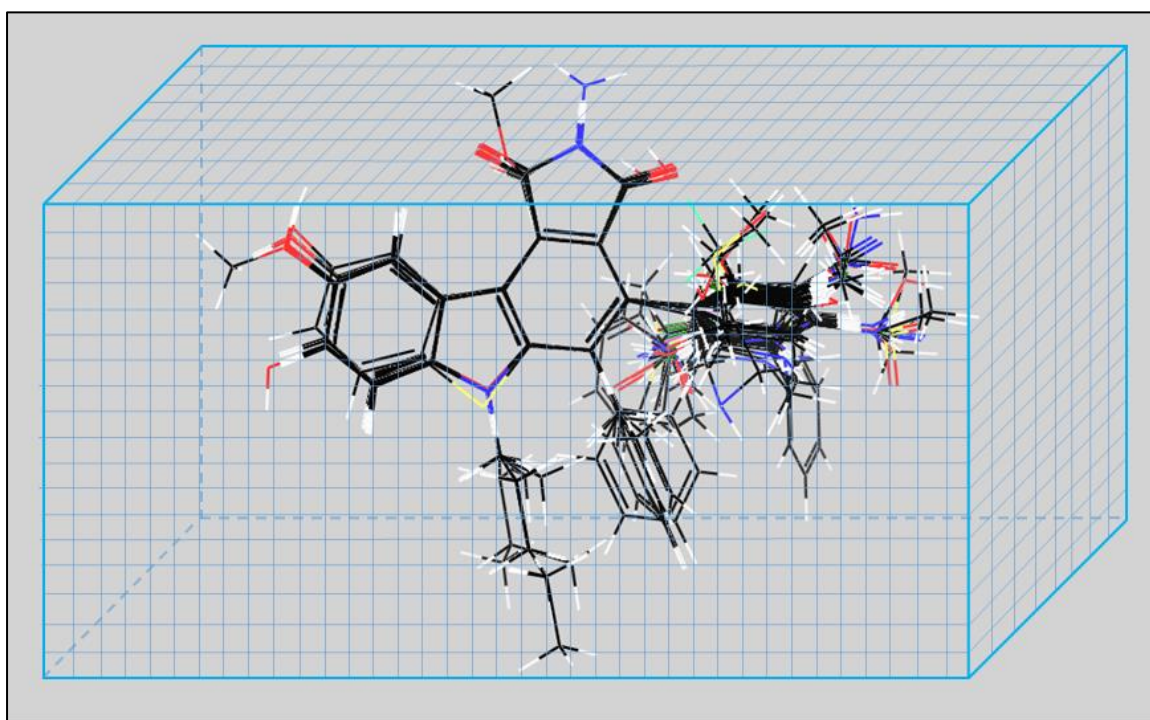
It may be seen that the compounds listed in Table 2.1 (Chapter 2) are synthetic compounds with known biological activity against the respective targets. However, there also exist natural compounds having similar scaffolds whose biological activities are unknown. Therefore, these compounds may be used to study for their inhibitory effectiveness even in the absence of the knowledge by predicting their pIC<sub>50</sub> values using the developed PMF-PLS QSAR model. We identified such compounds from SuperNatural II database (Banerjee *et al.*, 2015) by carrying out a substructure search using the scaffold structures of the TS-1 to TS-6 listed in Table 2.1. The Tanimoto scores (Bajusz *et al.*, 2015) were obtained to select the compounds with similarity to the queried scaffolds. The searches of SuperNatural II database gave twelve hits for TS-1, three for TS-2, three for TS-3, eleven for TS-4, three for TS-5 and five for TS-6, whose structures are listed in Appendix Table A7.

### **3.2.2 Scaffold based alignment of inhibitor molecules in 3D mesh grids**

The molecular structures downloaded from PubChemBioassay and SuperNatural II databases were imported to Schrödinger software for converting the 2D structure using the LigPrep module (version 2.5) (Schrödinger, LLC, 2011) to its equivalent energy minimized 3D form. Flexible ligand alignment and superimposition routines were used to obtain accurately aligned 3D structures with superimposed scaffolds. The graphics in Figure 3.2 shows the alignment obtained for TS-1 using 4-phenylpyrrolocarbazole derivatives. The aligned molecules bring out the structural variations arising due to the different side groups attached to the scaffold



and their 3D orientations. The aligned molecular structures were exported as .pdb files. This pre-processing of the data as outlined above was carried out for TS-1 to TS-6. The .pdb files were then imported to Matlab® (Version R2010b) and depending on the spread of atoms in the molecules in 3D space, all the aligned molecules were positioned in a common 3D box with finite intra-grid spacing. In general, it was observed that distance between two adjacent (vertical or horizontal) grid points of 1Å was adequate for the positions of the atoms to be resolved in the grid. All further coding and calculations for the PMF-PLS algorithm were carried out in Matlab.



**Figure 3.2:** Aligned 3D structures of the 4-phenylpyrrolocarbazole derivatives for TS-1 encapsulated in a 3D mesh grid. The mesh grid is not drawn to scale

### 3.2.3 Calculation of pseudo-molecular field (PMF) values and pseudo field molecular descriptors (PFMDs)

A PMF value at a grid point was calculated using the following equation,

$$\gamma_{j,k,l} = \sum_{i=1}^{n_a} \left( \frac{\sigma E_a(i) \chi(i)}{d(i)} \right) \quad (3.1)$$

where,  $\gamma_{j,k,l}$  is the value of the PMF at the grid point  $(j,k,l)$  in the cube,  $n_a$  is the total number of atoms,  $E_a(i)$  is the electron affinity of the  $i^{\text{th}}$  atom,  $\chi(i)$  is the electronegativity of the  $i^{\text{th}}$  atom,  $d(i)$  is the distance of the grid-point  $(j,k,l)$  from the  $i^{\text{th}}$  atom of the molecule in angstroms and  $\sigma$  is a suitably chosen scaling factor. Thus, finite and varying PMF values,  $\gamma_{j,k,l}$ , for a grid point may be obtained using Eq. 3.1. The electron affinity and electronegativity values for different atoms were obtained from WolframAlpha<sup>®</sup> (Wolfram Alpha LLC) and are presented in the Table 3.1.

**Table 3.1:** Electron affinity and electronegativity values of the atoms used for calculating PMF values<sup>†</sup>

Sr. no.	Element	Electron affinity ( $E_a$ )	Electronegativity ( $\chi$ )
1	H	72.8	2.2
2	C	153.9	2.5
3	N	7	3.1
4	O	141	3.5
5	F	328	4.1
6	Na	52.8	1
7	P	72	2.1
8	S	200	2.4
9	Cl	349	2.8
10	K	48.4	0.9
11	Br	324.6	2.7
12	I	295.2	2.2

<sup>†</sup>- Obtained from WolframAlpha

The PFMDs for PLS were obtained by the converting 3D PMF values to a 1D row vector. The PFMD row vectors were then stacked to form a 2-way array,  $X$ . Table 3.2 lists the dimensions of 3D arrays and size of the resulting row vectors forming the PFMDs for TS-1 to TS-6, respectively. The  $X$  matrix was sub-divided into three sets, an initial training set,  $X_{train}$ , an initial test set,  $X_{test}$ , (used for the purpose of developing a suitable PLS regression model) and a validation set,  $X_{val}$ , for external validation of the developed model. The biological activities of the molecules associated with  $X_{train}$ ,  $X_{test}$  and  $X_{val}$  are represented by column vectors  $Y_{train}$ ,  $Y_{test}$  and  $Y_{val}$ , respectively. The PFMDs for the natural compounds are designated as  $X_{np}$ .

**Table 3.2:** 3D Mesh grid (box size) and 1-way PFMD sizes for TS-1 to TS-6

TS	AID	Compounds	3D mesh grid	PFMD size
1	268838	4-phenylpyrrolocarbazoles	29x32x23	1x21344
2	566585	Benzylpiperidine derivatives	36x29x25	1x26100
3	198247	2-Substituted Dipyridodiazepinones	27x26x23	1x16146
4	197804	2-Pyridinone Derivatives	26x27x28	1x19656
5	160292	Cyclic urea derivatives	32x28x24	1x21504
6	579588	Azilide derivatives	31x28x31	1x26908

### 3.2.4 QSAR modelling

Here we describe the algorithm for developing the QSAR model by PMF-PLS. The flowchart Figure 3.3 schematically outlines the steps involved which are referenced by Box numbers in a way similar to Figure 3.1. The inputs to PMF-PLS are the  $X$  and  $Y$  matrices which denote the PFMDs and dependent variables as shown in [B 3.1]. As will be seen, implementing the algorithm leads us to obtaining a set of average regression coefficients,  $B_{avg}$ , for a model which can be used to predict the  $pIC_{50}$  values of new compounds. Section 3.2.4.1 outlines the PLS module while its use in PMF-PLS is described in Section 3.2.4.2. For ready reference a tabulated list of notation used in this work is provided in Table A14.

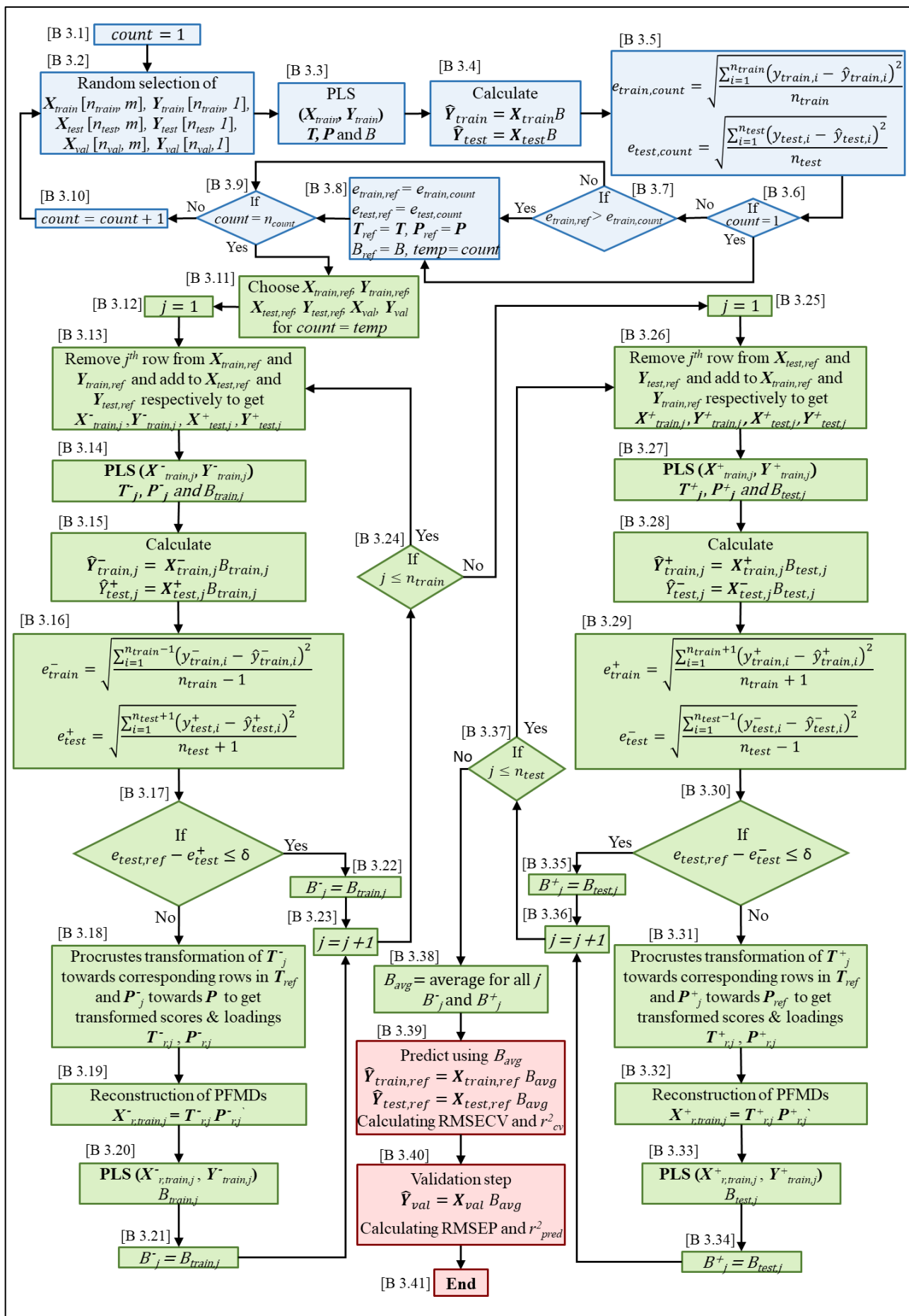


Figure 3.3: Flowchart of QSAR modelling by PMF-PLS algorithm

### 3.2.4.1 Partial Least Squares (PLS)

Partial least squares is a widely used regression method which aims at capturing relationships between the dependent variable  $Y$  (i.e., biological activity matrix of size  $[n \ 1]$ , where  $n$  is the number of compounds) and the independent variables  $X$  (the high dimensional PFMDs of size  $[n \ m]$  with  $m$  the number of dimensions). PLS projects the values of  $X$  and  $Y$ , respectively, to a latent subspace of lower dimensions  $a < m$  while maximizing the covariance between them. PLS regression is carried out by the decomposition of  $X$  and  $Y$  as shown in Eq. 3.2 and Eq. 3.3, respectively, to obtain matrices  $T$ ,  $P$ ,  $U$  and  $Q$ , such that,

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} = \sum_{i=1}^{i=a} t_i p_i' + \mathbf{E} \quad (3.2)$$

$$\mathbf{Y} = \mathbf{UQ}' + \mathbf{F} = \sum_{i=1}^{i=a} u_i q_i' + \mathbf{F} \quad (3.3)$$

In Eq. 3.2 the scores  $T$  of matrix size  $[n \ a]$ , is composed of  $a$  latent vectors while the loadings  $P$  is of size  $[m \ a]$  with  $E$  the residuals in the decomposition of  $X$ . Similarly, in Eq. 3.3 the scores  $U$  is of size  $[n \ a]$  and the loadings  $Q$  is of size  $[1 \ a]$  for the decomposition of  $Y$  with  $F$  the corresponding residuals. The magnitudes of the residuals  $E$  and  $F$  depend on the number of latent components chosen and for a proper choice of  $a < m$ , dimensionality reduction is possible with minimization of residuals. The score vectors  $t_i$  and  $u_i$ ,  $i = 1, 2, \dots, a$  obtained may be regressed to obtain the following linear model between them, namely,

$$u_i = t_i b_i \quad (3.4)$$

where,  $b_i$  are the regression coefficients which may be calculated as follows

$$b_i = u_i' t_i / t_i' t_i \quad (3.5)$$

In general we may therefore write

$$\mathbf{U} = \mathbf{TB} \quad (3.6)$$

and combining Eqs. 3.6 and 3.3 we may obtain model predictions,  $\hat{Y}$ , as

$$\hat{Y} = TBQ' \quad (3.7)$$

The NIPALS algorithm (Geladi and Kowalski, 1986; Garthwaite, 1994) is a commonly used method to estimate  $T$ ,  $P$ ,  $Q$  and  $U$  matrices. The PLS components are calculated one-at-a-time from deflated matrices  $X_i$  and  $Y_i$ , where  $i = 1, 2, \dots, a$ , that are obtained by subtracting the contribution of the latent variables obtained in previous step from  $X_{i-1}$  and  $Y_{i-1}$ . Thus, for prediction of  $\hat{Y}_{new}$  values for new observations (new drug molecules) values of latent variables ( $T_{new}$ ) are required to be calculated iteratively using  $X_{new}$  with deflation of  $X_{new}$  during each iteration. Equation 3.7 may be used for predicting  $\hat{Y}_{new}$  using the determined regression coefficients  $B$  as

$$\hat{Y}_{new} = T_{new}BQ' \quad (3.8)$$

Another variant of PLS is the SIMPLS method for PLS regression (Table 3.3 steps 1-12) (de Jong, 1993). It offers significant advantages as it performs the calculations of all the scores and loadings using the original  $X$  and  $Y$  matrices in every iterative steps unlike the NIPALS algorithm which uses deflated matrices,  $X_{i-1}$  and  $Y_{i-1}$ . It may be seen that the algorithm uses singular value decomposition (SVD) to minimize the residuals during regressing. The steps of SIMPLS algorithm involved in the calculation of scores and loadings are presented in Table 3.3.

**Table 3.3:** Steps of SIMPLS algorithm adopted for PMF-PLS (Adapted from de Jong, 1993)

Step no.	Description	Steps
1	Compute covariance matrix $S_0$	$S_0[m, 1] = X'Y$
2	Loop over chosen $a$	For $i = 1$ to $a$
3	Singular value decomposition of $S_{i-1}$	$r_i [m, 1]$ = first left singular vector of SVD ( $S_{i-1}$ )
4	Compute scores of $X$ data	$t_i[m, 1] = Xr_i$
5	Compute loadings of $X$ data	$p_i[n, 1] = X' t_i / t_i' t_i$
6	Normalizing loadings	$p_i[n, 1] = p_i / \ p_i\ $
7	Compute loadings of $Y$ data	$q_i[1, 1] = Y' t_i / t_i' t_i$
8	Normalizing loadings	$q_i[1, 1] = q_i / \ q_i\ $
9	Compute scores of $Y$ data	$u_i[n, 1] = Yq_i$
10	Using $P_i = [p_1, p_2, \dots, p_i]$ , deflate covariance matrix and obtain $S_i$	$S_i[m, 1] = S_0 - P_i(P_i' P_i)^{-1} P_i' S_0$
11	Next $i$	$i = i + 1$
12	Calculation of regression coefficients	$B = RT'Y = RQ'$

Since the loadings and scores are calculated directly from the original data ( $X$  and  $Y$ ) the calculation of the scores,  $T_{new}$ , of the new data points can be done directly from the  $X_{new}$  without the need for deflated values of  $X_{new}$ . The regression coefficients,  $B$ , are calculated as shown in the step 12 of Table 3.3, i.e.,

$$B = RT'Y = RQ' \quad (3.9)$$

These regression coefficients,  $B$ , along with  $X_{new}$  can be used to make prediction of  $\hat{Y}_{new}$  values, i.e.,

$$\hat{Y}_{new} = X_{new}B \quad (3.10)$$

In the present work, we implement the Matlab function 'plsregress', which uses SIMPLS, during coding of the proposed PMF-PLS algorithm.

#### 3.2.4.2 PMF-PLS algorithm

The flowchart presented in Figure 3.3 illustrates the methodology of PMF-PLS. The entire algorithm may be viewed in three parts. The first part [B 3.1]-[B 3.11] shows the steps involved in choosing suitable training ( $X_{train}$ ,  $Y_{train}$ ), test ( $X_{test}$ ,  $Y_{test}$ ) and validation ( $X_{val}$ ,  $Y_{val}$ ) sets from the molecules in a TS. The training set ( $X_{train}$ ,  $Y_{train}$ ) and test set ( $X_{test}$ ,  $Y_{test}$ ) were used in the second part [B 3.12]-[B 3.38] of the algorithm to perform calculations to obtain averaged values of regression coefficients  $B_{avg}$  that are needed for the linear regression model. In the third part [B 3.39]-[B 3.41], the model obtained may be used for predicting values of  $\hat{Y}_{val}$  from  $X_{val}$ . Note that  $X_{val}$  is a set of compounds not present in the training or the test set. Therefore  $\hat{Y}_{val}$  can be compared with its known values  $Y_{val}$  for the purpose of validation.

A validation set ( $X_{val}$ ,  $Y_{val}$ ) was initially selected randomly from the molecules in a TS. Training ( $X_{train}$ ,  $Y_{train}$ ) and test set ( $X_{test}$ ,  $Y_{test}$ ) were selected randomly [B 3.2] from the remaining molecules and were used to perform a PLS regression [B 3.3] as outlined in Section 3.2.4.1. The model predicted pIC<sub>50</sub> values  $\hat{Y}_{train}$  and  $\hat{Y}_{test}$  [B 3.4] were compared with their actual values  $Y_{train}$  and  $Y_{test}$  by calculating the root mean squared errors (RMSE),  $e_{train}$  and  $e_{test}$ , respectively [B 3.5] using the general equation,

$$RMSE = \sqrt{\frac{(\mathbf{Y} - \hat{\mathbf{Y}})^2}{n_{mol}}} \quad (3.11)$$

where,  $\mathbf{Y}$  are the actual  $pIC_{50}$  values for the set,  $n_{mol}$  is the number of molecules in the set ( $n_{train}$  or  $n_{test}$ ) and  $\hat{\mathbf{Y}}$  are the predicted  $pIC_{50}$  values. This procedure was repeated a large number of times (say  $n_{count}$ ) with different training and test sets randomly chosen for each iteration [B 3.10] using the 'randperm' function in Matlab. Of these iterations the training and test sets which realized the minimum root mean squared error ( $RMSE$ )  $e_{test}$  [B 3.7]-[B 3.11] was chosen as the reference training ( $\mathbf{X}_{train}, \mathbf{Y}_{train}$ ) and test ( $\mathbf{X}_{test}, \mathbf{Y}_{test}$ ) sets for further calculations in part two.

In the second part of the methodology, we choose to apply a modified leave-one-out cross validation procedure to obtain an average set of regression coefficients  $B_{avg}$  with reduced sensitivity to variations in the training set ( $\mathbf{X}_{train}, \mathbf{Y}_{train}$ ) [B 3.12]-[B 3.38] for prediction. The procedure of obtaining  $B_{avg}$  carries out the PLS simulations iteratively for two types of variations in the training set. The first type [B 3.12]-[B 3.24] sequentially removes the  $j^{th}$  row from the training set ( $\mathbf{X}_{train}, \mathbf{Y}_{train}$ ) and adds that row to the test set ( $\mathbf{X}_{test}, \mathbf{Y}_{test}$ ) to obtain modified training ( $\mathbf{X}^{-}_{train,j}, \mathbf{Y}^{-}_{train,j}$ ) and test ( $\mathbf{X}^{+}_{test,j}, \mathbf{Y}^{+}_{test,j}$ ) sets [B 3.13]. The (-) superscript indicates that the row is removed from corresponding training/test set and *vice-versa* (+) indicates an addition. The second type of variation [B 3.25]-[B 3.37] is similar to the first [B 3.12]-[B 3.24], however, it sequentially removes the  $j^{th}$  row from test set ( $\mathbf{X}_{test}, \mathbf{Y}_{test}$ ) and adds this row to training ( $\mathbf{X}_{train}, \mathbf{Y}_{train}$ ), respectively, to obtain ( $\mathbf{X}^{+}_{train,j}, \mathbf{Y}^{+}_{train,j}$ ) and ( $\mathbf{X}^{-}_{test,j}, \mathbf{Y}^{-}_{test,j}$ ) [B 3.26]. In both cases, the aim is to calculate regression coefficients  $B^{-}_j$  and  $B^{+}_j$  which may be averaged to obtain  $B_{avg}$  [B 3.38] the regression coefficients to be used for model prediction purposes with new compounds. We now describe the evaluation of coefficients  $B^{-}_j$  [B 3.12]-[B 3.24].

To evaluate  $B^{-}_j$ , we carry out the PLS of ( $\mathbf{X}^{-}_{train,j}, \mathbf{Y}^{-}_{train,j}$ ) and obtain  $\mathbf{T}^{-}_j, \mathbf{P}^{-}_j$  and  $\mathbf{B}^{-}_{train,j}$  [B 3.14]. The corresponding predictions  $\hat{\mathbf{Y}}^{-}_{train,j}$  and  $\hat{\mathbf{Y}}^{+}_{test,j}$  [B 3.15] may be used to evaluate the  $RMSEs$   $e^{-}_{train}$  and  $e^{+}_{test}$  [B 3.16]. Our observation with respect to  $RMSEs$  was that when the differences between  $e^{+}_{test}$  and  $e_{test,ref}$  or between  $e^{-}_{test}$  and  $e_{test,ref}$  were larger



than  $\delta$  (15% of the  $Y_{range}$  for the TS) it was because the  $T^-_j$  and  $P^-_j$  matrices were translated, rotated or scaled from the original  $T_{ref}$  and  $P_{ref}$  matrices and these arose due to the removal of the row from  $(X_{train}, Y_{train})$ . It would then be necessary to conform  $T^-_j$  and  $P^-_j$  to  $T_{ref}$  and  $P_{ref}$  respectively. We used Procrustes transformation (Kendall, 1989) which evaluates a linear transformation (i.e., translation, reflection, orthogonal rotation, scaling) of the points in one matrix to best conform to the points in another matrix for this goal. The Matlab function 'procrustes' was used [B 3.18] to conform the matrix pairs  $(T^-_j, T_{ref})$  and  $(P^-_j, P_{ref})$  and obtain the transformed score  $T^-_{rj}$  and loading  $P^-_{rj}$  matrices. From the knowledge of  $T^-_{rj}$  and  $P^-_{rj}$ , transformed PFMDs  $X^-_{r,train,j}$  can be calculated [B 3.19]. A PLS study of the transformed PFMDs i.e.,  $(X^-_{r,train,j}, Y^-_{train,j})$  can be performed to obtain corresponding regression coefficients  $B^-_j$  [B 3.20]-[B 3.21]. Note that when the criteria in [B 3.17] is satisfied there is no need for Procrustes transformation and  $B^-_j$  can be directly assigned the value of  $B^-_{train,j}$  obtained in [B 3.14]. This procedure is repeated for  $j = 1, 2, \dots, n_{train}$ , where,  $n_{train}$  is the population of the molecules in the training set.

The second type of variations for obtaining regression coefficients  $B^+_j, j = 1, 2, \dots, n_{test}$ , is shown in [B 3.25]-[B 3.37]. The calculations are similar to those for  $B^-_j$  [B 3.12]-[B 3.24] except that the iterations are carried out for addition of rows from  $(X_{test}, Y_{test})$  to  $(X_{train}, Y_{train})$ . The  $B^+_j$  obtained was used to average along with  $B^-_j$  to obtain  $B_{avg}$  [B 3.38]. The PMF-PLS QSAR model is then obtained as

$$\hat{Y} = \mathbf{X}B_{avg} \quad (3.12)$$

Cross-validation and external validation of the above regression model was carried out in the third part of the algorithm [B 3.39]-[B 3.41]. Predicted  $pIC_{50}$  values  $\hat{Y}_{train}$  and  $\hat{Y}_{test}$  were determined using  $X_{train}, X_{test}$  and  $B_{avg}$  in Eq. 3.12. Cross-validation of a QSAR model is carried out by performing the predictions for the molecules used for building the model. Hence in present work cross-validation was performed using the predicted  $pIC_{50}$  values  $\hat{Y}_{train}$  and  $\hat{Y}_{test}$  taken together.  $NRMSE$  of cross-validation

(*NRMSECV*) was calculated using  $\hat{Y}_{train}$ ,  $\hat{Y}_{test}$ ,  $Y_{train}$  and  $Y_{test}$  in Eq. 2.15. Similarly Correlation coefficient for cross-validation ( $r_{cv}$ ), for prediction ( $r_{pred}$ ) coefficient of determination for cross-validation and ( $R^2_{cv}$ ) and for external validation ( $Q^2_{ext(F1)}$ ) was calculated as described in the Chapter 2 Section 2.2.10. It may be noted that cross-validation parameters were calculated considering both training and test sets as both of these sets were used for building the model and external validation of the model was performed using the validation set.

### 3.3 Results and discussion

#### 3.3.1 Generation of PFMDs and QSAR modelling

Energy minimized 3D structures of the compounds were obtained [B 1.4] and aligned on the basis of their scaffolds [B 1.5] in a 3D mesh grid with a mesh size of 1Å as shown in Figure 3.2 for TS-1 [B 1.8]. It may be noted that this pre-processing of structures is similar to that employed in the CoMFA methodology (Cramer et al., 1988).

The next step was to obtain the PMF values at the grid points [B 1.9] using Eq. 3.1.

$$\gamma_{j,k,l} = \sum_{i=1}^{n_a} \left( \frac{\sigma E_a(i) \chi(i)}{d(i)} \right) \quad (3.1)$$

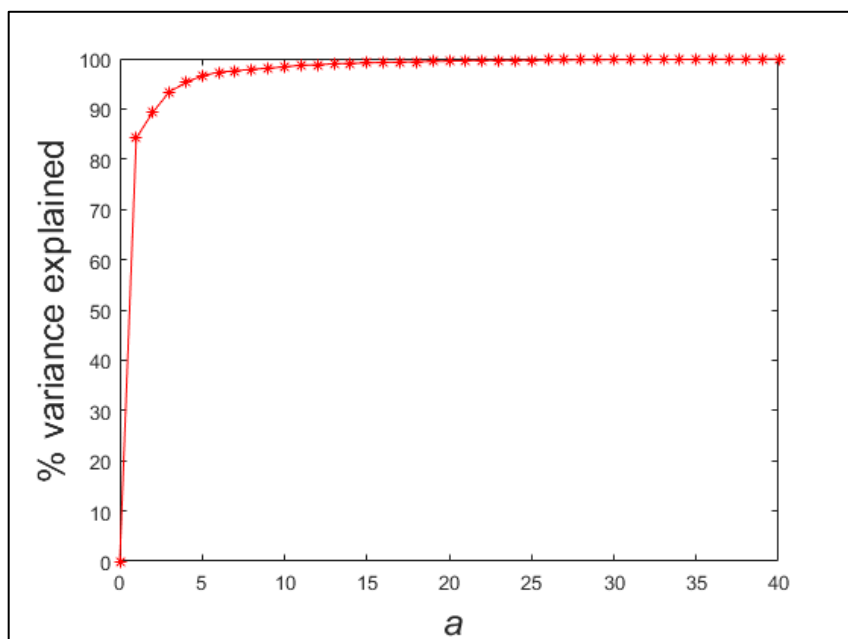
It may be observed that the calculation is similar to obtaining the electrostatic field values in CoMFA using the Coulomb potential function (Kubinyi, 1997a), namely

$$E_{C_{j,k,l}} = \sum_{i=1}^{n_a} \frac{q(i)q_p}{Dd(i)} \quad (3.13)$$

where,  $E_{C_{j,k,l}}$  is the Coulomb interaction energy at grid point ( $j,k,l$ ),  $q(i)$  is the partial charge of the  $i^{th}$  atom of the molecule,  $q_p$  is the charge of the probe atom,  $D$  is the dielectric constant,  $d(i)$  is the distance between the  $i^{th}$  atom of the molecule and the grid point ( $j,k,l$ ), and  $n_a$  the total number of atoms in the inhibitor molecule. For the probe atom the chosen charge is kept constant in the calculations. Hence,  $q(i)$  and  $d(i)$  are the quantities in Eq. 3.13 that determine the electrostatic field values. It may be seen that

Eq. 3.1 is analogous to Eq. 3.13 the Coulomb potential function such that in Eq. 3.1 we replace: 1) the atomic charge  $q(i)$  of  $i^{\text{th}}$  atom with the product of electron affinity,  $E_a(i)$ , and electronegativity,  $\chi(i)$ , of that atom (Table 3.1) and 2) the constants  $q_p$  and  $D$  with a suitably chosen value for the scaling factor  $\sigma = 0.1$ . Note that the PMF value at a grid point is also dependent on two factors, namely, the distance of the grid point from the atoms and the product of the atomic properties ( $E_a(i)$  and  $\chi(i)$ ). The partial charges of the atoms in the molecule are known to be dependent on electron affinity and the orbital electronegativity (Mulliken, 1934; Gasteiger and Marsili, 1980). It is based on this rationale that we attempted to replace the  $q(i)$  in Eq. 3.13 with the product of  $E_a(i)$  and  $\chi(i)$ . It may be also observed that the calculation of PMF values using Eq. 3.1 is simpler than CoMFA because the  $E_a$  and  $\chi$  values are constant for an atom when compared to partial atomic charges which need to be calculated individually for the atoms of each molecule. The results of QSAR model obtained on analyzing systems TS-1 to TS-6 corroborate the choice of intrinsic properties for PMF calculations.

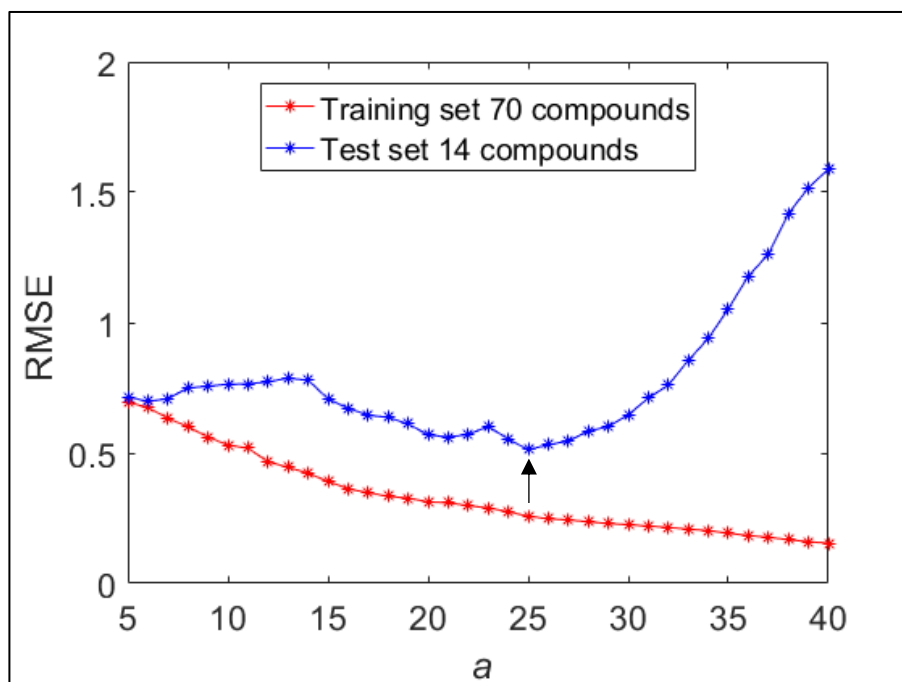
The 3D arrays of PMF values were converted into 1D PFMD arrays [B 1.10] and regressed with  $Y$  values using PMF-PLS methodology as described in Section 3.2.4. Regression models built using a single training set tend to have a bias for the training set used which can result in problems arising due to the overfitting of the QSAR model. A way to reduce the model bias is to use multiple training sets that yield average values of the regression coefficients to build the final QSAR model (Wold, 1978). Also, some training sets may not perform as well as others resulting in a reduced performance of the final model. To take care of these problems in the PMF-PLS algorithm, we first identify from many a training set that best predicts the test set [B 3.1]-[B 3.11]. Then by making small alterations in the choice of molecules in this training set, we may obtain a set of regression coefficients [B 3.11]-[B 3.38] that could be averaged and used for the development of the final model. Using random number generation function in Matlab, the data was randomly divided into training ( $X_{train}$ ,  $Y_{train}$ ), test ( $X_{test}$ ,  $Y_{test}$ ) and validation ( $X_{val}$ ,  $Y_{val}$ ) sets [B 3.2] and PLS regression performed with each set for varying  $a = 5$  to 40 [B 3.3]. It was observed that for TS-1, the first



**Figure 3.4:** Percentage of variance observed in  $X_{train}$  explained with varying number of PLS components ( $a$ ) used for regression for TS-1

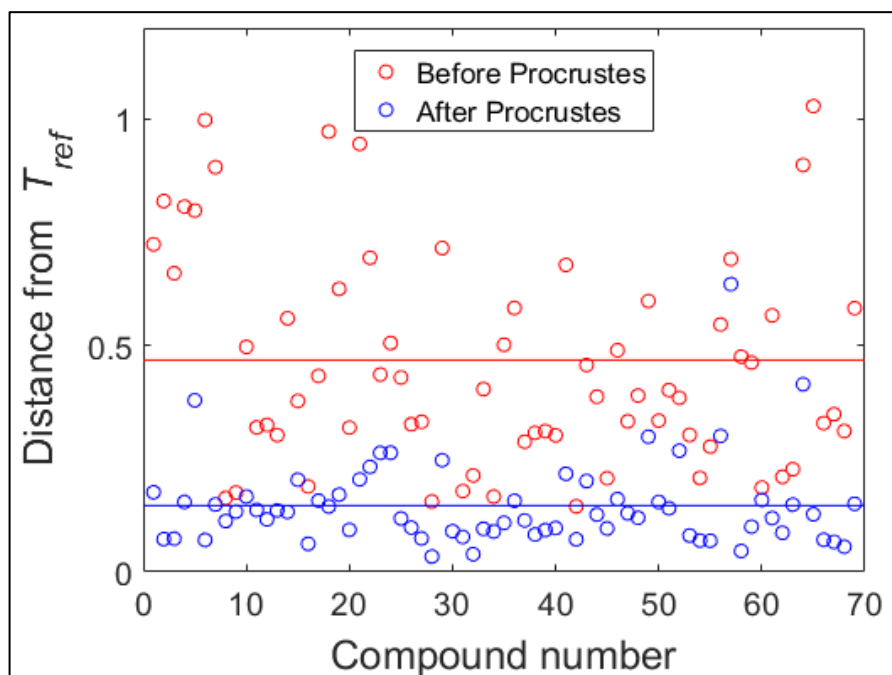
twenty components (i.e.,  $a = 20$ ) explained more than 99.5% of the variance observed in  $X_{train}$  as shown in Figure 3.4. Thus, using twenty latent space variables (i.e.,  $a = 20$ ), instead of the high dimensional PFMDs (Table 3.2), resulted in a 99.9 % reduction in the dimensionality required for analysis. This process was repeated for a number of different training ( $X_{train}, Y_{train}$ ) and test ( $X_{test}, Y_{test}$ ) sets which were generated randomly [B 3.1]-[B 3.11]. The training set that predicted the test set with minimum  $RMSE$  [B 3.8] was chosen as the reference training ( $X_{train,ref}, Y_{train,ref}$ ), and test ( $X_{test,ref}, Y_{test,ref}$ ) sets [B 3.11]. For TS-1, Figure 3.5 shows the  $RMSE$  for the prediction of  $pIC_{50}$  values for the ( $X_{train,ref}, Y_{train,ref}$ ), and ( $X_{test,ref}, Y_{test,ref}$ ) for varying values of  $a$ . It was observed that using higher values of  $a$  ( $>25$ ) for regression, results in overfitting of the model because the  $RMSE$  for the prediction of  $\hat{Y}_{test}$  increases even though the  $RMSE$  for  $\hat{Y}_{train}$  continuously reduces (Figure 3.5). We therefore choose the value of  $a$  ( $=25$ ) that realizes minimum  $RMSE$  for the prediction of  $\hat{Y}_{test,ref}$  for further calculations in applying the PMF-PLS algorithm.

We carried out the second part of the PMF-PLS algorithm [B 3.12]-[B 3.38], where alterations were done to ( $X_{train,ref}, Y_{train,ref}$ ) by removing one molecule from ( $X_{train,ref},$



**Figure 3.5:** *RMSE* values for predictions of training (red) and test (blue) sets with varying values of  $a$  used for the PLS regression of TS-1. The arrow mark indicates the minimum *RMSE* for test set prediction at  $a = 25$

$Y_{train,ref}$ ) and adding to  $(X_{test,ref}, Y_{test,ref})$  [B 3.13] or vice versa [B 3.26]. These altered training sets  $(X^{-train,j}, Y^{-train,j})$  and  $(X^{+train,j}, Y^{+train,j})$ , as the case may be, were used for PLS regression [B 3.14] and [B 3.27], respectively. The altered training sets for which the *RMSE* for test sets is greater than *RMSE* for reference test set by  $\delta$  (taken as 15% of the total range of  $pIC_{50}$  values) ([B 3.17] or [B 3.30]) were observed to have significantly different PLS score values ( $T^{-}_j$  or  $T^{+}_j$ ) when compared to  $T_{ref}$ . Therefore, Procrustes transformation (Kendall, 1989) of  $T^{-}_j$  and  $P^{-}_j$  [B 3.18] or  $T^{+}_j$  and  $P^{+}_j$  [B 3.31] was carried out to obtain transformed scores and loadings. The Euclidean distances of the  $T_{ref}$  scores from  $T^{-}_j$  or  $T^{+}_j$  (i.e., before the Procrustes transformation) and from  $T^{-}_{r,j}$  or  $T^{+}_{r,j}$  (i.e., after the Procrustes transformation) were calculated. It was observed that the distances of  $T_{ref}$  from  $T^{-}_{r,j}$  or  $T^{+}_{r,j}$  were reduced as compared to the distances from  $T^{-}_j$  or  $T^{+}_j$  as illustrated in Figure 3.6. Similar results were also observed for Procrustes transformation of the loadings. Thus, Procrustes transformation brings the scores and loadings for altered



**Figure 3.6:** Effect of Procrustes transformation on the scores of compounds. Euclidean distances between  $T_{ref}$  and  $T^{-j}$  before (red) and  $T^{-r,j}$  after (blue) transformation on removing compound number 22 (Table A1) from  $(\mathbf{X}_{train,ref}, \mathbf{Y}_{train,ref})$  of TS-1. The red and blue lines are the mean distances of the compounds calculated before and after Procrustes transformation

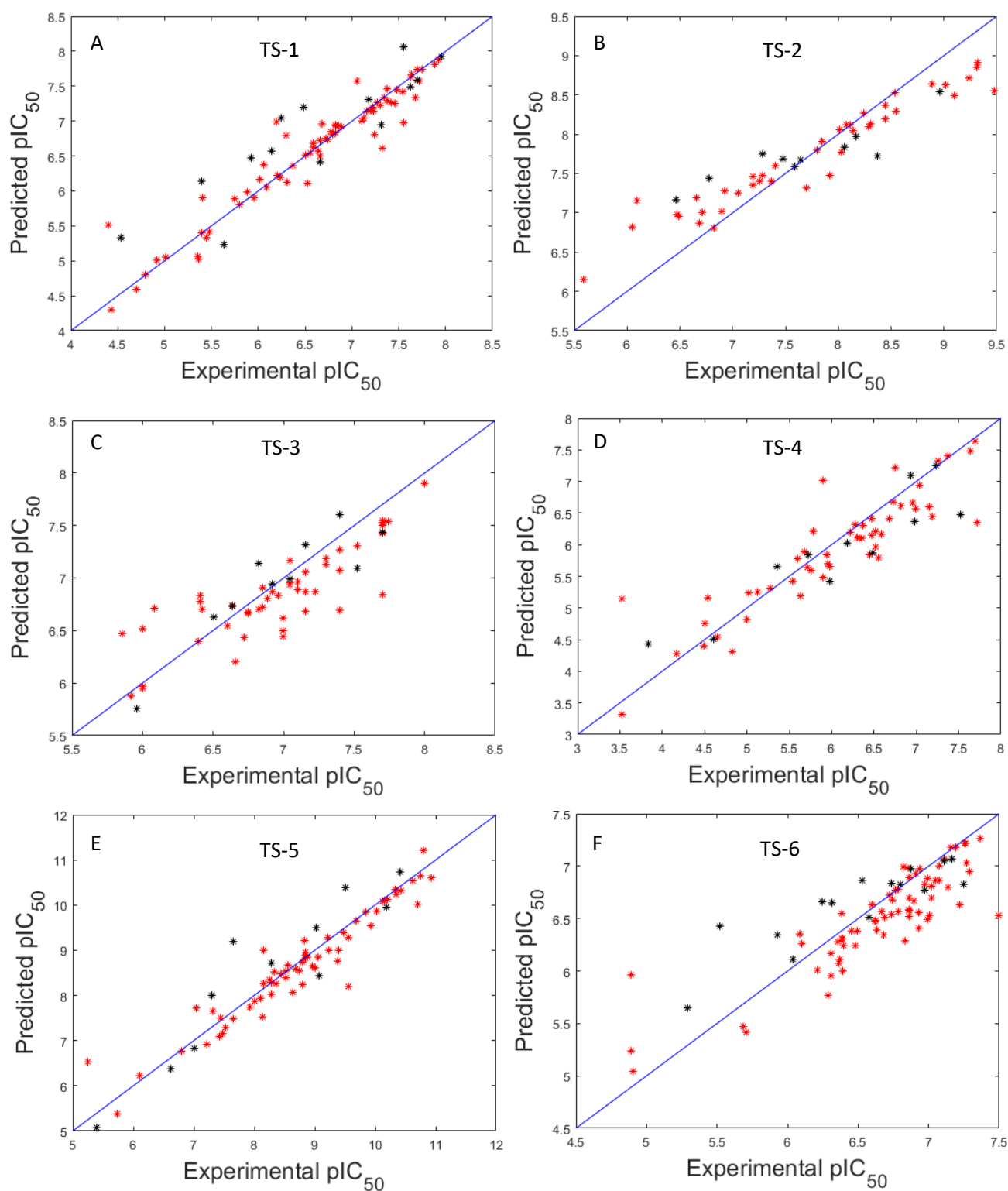
training sets closer to the reference training set scores ( $T_{ref}$ ) and loadings ( $P_{ref}$ ), respectively. The modified PFMD values were obtained, from the transformed scores and loadings using Eq. 3.2 (i.e.,  $\mathbf{X}^{-r,train,j}$  [B 3.19] and  $\mathbf{X}^{+r,train,j}$  [B 3.32]). It was observed that regression coefficients ( $B_{train,j}$  [B 3.20] and  $B_{test,j}$  [B 3.33]) obtained using modified PFMDs predicted  $\hat{\mathbf{Y}}^{+_{test,j}}$  and  $\hat{\mathbf{Y}}^{-_{test,j}}$  with reduced *RMSE* when compared to that before the Procrustes transformation. As an example, removing compound number 22 (Table A1) to form  $(\mathbf{X}^{-r,train,j}, \mathbf{Y}^{-train,j})$  results in *RMSE* to rise from 0.51 for reference test set to 2.05 for altered test set ( $e^{+test}$ ) [B 3.16] which on Procrustes transformation reduced to 0.78 thereby improving the predictions. An average,  $B_{avg}$ , of all the regression coefficients ( $B^{-j}$  and  $B^{+j}$ ) was determined [B 3.38] and the PMF-PLS QSAR model was obtained (Eq. 3.12) [B 3.39]. It was observed that the time required for the PMF-PLS algorithm was of the between 6-8 hours. The majority of time was taken by the first

part of the algorithm to arrive at the reference training and test set. The second and third part of the algorithm took about 1-2 minutes to complete.

Using the PMF-PLS QSAR model,  $\hat{Y}_{train,ref}$  and  $\hat{Y}_{test,ref}$  were predicted [B 3.39] to obtain  $R^2_{cv}$  (Eq. 2.16) and  $NRMSECV$  (Eq. 2.15) for cross validation using the training and test sets. Figure 3.7A shows the diagonal plot of  $Y_{train,ref}$  vs.  $\hat{Y}_{train,ref}$  and  $Y_{test,ref}$  vs.  $\hat{Y}_{test,ref}$  for TS-1. It was observed that the PMF-PLS QSAR model for TS-1 yielded a high value  $R^2_{cv}$  of 0.88 and low normalized  $NRMSECV$  of 0.09 as required (Table 3.4). External validation of the PMF-PLS QSAR model was carried out by predicting  $\hat{Y}_{val}$  [B 3.40] and the corresponding  $Q^2_{ext(F1)}$  (Eq. 2.17) and  $NRMSEP$  (Eq. 2.15) calculated. Figure 3.8A shows the plot of  $Y_{val}$  vs.  $\hat{Y}_{val}$  for TS-1. PMF-PLS QSAR model, for TS-1, yielded an  $Q^2_{ext(F1)}$  of 0.71 and a normalized  $NRMSEP$  of 0.13 (Table 3.4). Figures 3.7B to 3.7F and Figures 3.8B to 3.8F show the diagonal plots for the predicted *versus* the actual pIC<sub>50</sub> values for TS-2 to TS-6, respectively, similar to Figures 3.7A and 3.8A for TS-1. The corresponding  $R^2_{cv}$ ,  $NRMSECV$ ,  $Q^2_{ext(F1)}$ , and  $NRMSEP$  for TS-2 to TS-6 are given in Table 3.4. The experimental and predicted pIC<sub>50</sub> values of all the molecules of TS-1 to TS-6 using PMF-PLS QSAR model are given in Appendix Tables A15 to A20.

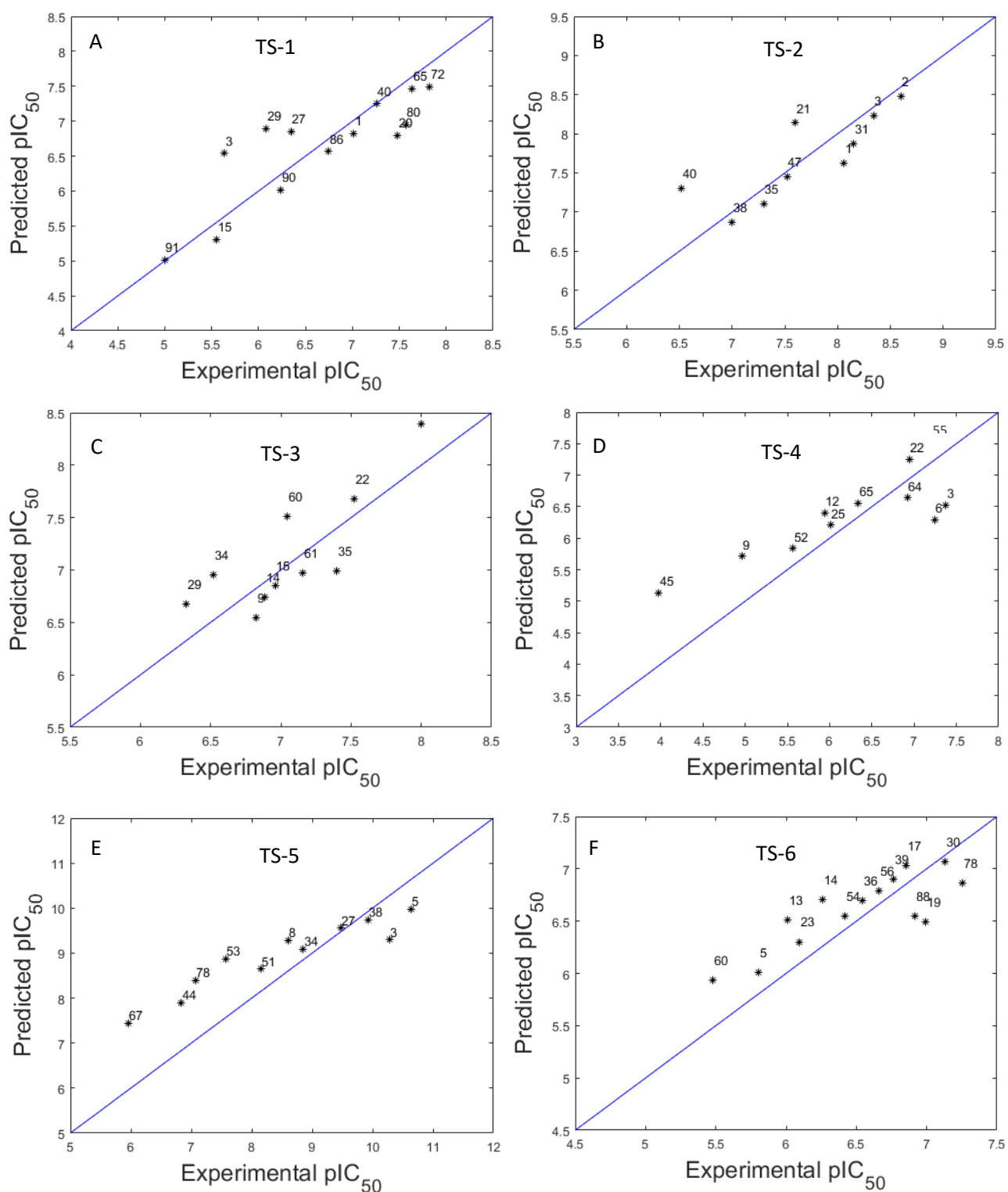
**Table 3.4:** PMF-PLS QSAR model fitting statistics for TS-1 to TS-6

TS No.	Compounds	Number of Components	Cross-validation			External Validation		
			$r_{cv}$	$R^2_{cv}$	$NRMSECV$	$r_{pred}$	$Q^2_{ext(F1)}$	$NRMSEP$
1	4-phenyl pyrrolocarbazoles	25	0.94	0.88	0.09	0.84	0.71	0.13
2	benzylpiperidine derivatives	18	0.95	0.82	0.10	0.81	0.66	0.10
3	2-Substituted dipyrindiazepinones	20	0.82	0.68	0.12	0.82	0.69	0.10
4	2-Pyridinone derivatives	18	0.90	0.79	0.11	0.85	0.62	0.15
5	cyclic urea derivatives	22	0.94	0.89	0.08	0.94	0.62	0.16
6	azilide derivatives	17	0.84	0.67	0.12	0.82	0.63	0.12



**Figure 3.7:** Plots of actual  $pIC_{50}$  values ( $Y$ ) vs. the predicted values ( $\hat{Y}$ ) for cross-validation using PMF-PLS QSAR model. (A) TS-1, (B) TS-2, (C) TS-3, (D) TS-4, (E) TS-5 and (F) TS-6 inhibitors. The training set compounds are marked in red and test set compounds in black as specified in the Appendix, Tables A15 to A20, respectively





**Figure 3.8:** Plots for actual  $pIC_{50}$  values ( $Y_{val}$ ) vs. the predicted values ( $\hat{Y}_{val}$ ) using PMF-PLS QSAR model for validation sets of (A) TS-1, (B) TS-2, (C) TS-3, (D) TS-4, (E) TS-5 and (F) TS-6 inhibitors. The numbers in the panels A to F indicate the validation set compound number specified in the Appendix, Tables A15 to A20, respectively

The PMF-PLS QSAR model quality was further assessed by applying the mean absolute error (MAE) based criteria for the validation set predictions (Roy *et al.*, 2016). Roy *et al.*, 2016 define a criteria to assess the model prediction quality as follows,

$$\begin{aligned}
 & \text{Good model:} && MAE \leq 0.1 \times \text{training set range AND} \\
 & && (MAE + 3SD) \leq 0.2 \times \text{training set range} \\
 & \text{Bad model:} && MAE > 0.15 \times \text{training set range OR} \\
 & && (MAE + 3SD) > 0.25 \times \text{training set range}
 \end{aligned}
 \tag{3.14}$$

where,  $MAE = (\sum_{i=1}^{n_{val}} |y_i - \hat{y}_i|) / n_{val}$ , SD is the standard deviation of the absolute errors of prediction of the validation set,  $y_i$  and  $\hat{y}_i$  are the actual and the predicted pIC<sub>50</sub> values of the  $i^{th}$  molecule in the validation set,  $n_{val}$  is the number of molecules in the validation set, the training set range is the difference between the maximum and the minimum actual pIC<sub>50</sub> values of the training set molecules. The models lying between the good and bad criteria are considered to be of moderate quality (Roy *et al.*, 2016). It was observed the above criteria was not satisfied for the PMF-PLS QSAR models and suggested that this may be due to the small sample size of the validation sets. The (Mean  $\pm$  3SD) criteria used in Eq. 3.14 is the 99.7% confidence interval used for removal of outliers for data sets with large sample size. However, for the data sets with smaller sample size criteria of (Mean  $\pm$  2SD) with the confidence interval of 95% is used (Ilyas and Chu, 2019; Hodge and Austin, 2004; Brownlee, 2018). This led us to consider a relaxed criteria of (MAE + 2SD) in Eq. 3.14 following the procedure by Roy *et al.*, 2016, i.e.,

$$\begin{aligned}
 & \text{Good model:} && MAE \leq 0.1 \times \text{training set range AND} \\
 & && (MAE + 2SD) \leq 0.2 \times \text{training set range} \\
 & \text{Bad model:} && MAE > 0.15 \times \text{training set range OR} \\
 & && (MAE + 2SD) > 0.25 \times \text{training set range}
 \end{aligned}
 \tag{3.15}$$

The calculations with (MAE+2SD) criteria for PMF-PLS models showed satisfaction of the modified criteria (Eq. 3.15). The MAE based criteria values evaluated using Eq. 3.15 are given in Table 3.5. The model quality for TS-2 and TS-3 was found to be good while that for TS-1, TS-4, TS-5 and TS-6 were observed to be moderate. Thus, Eq. 3.15 provides an alternative to determine the model quality when the sample size is small. Comparing the QSAR models for TS-3 and TS-4, which have the same target protein, i.e., HIV-1 reverse transcriptase, it may be

observed that the model performance for TS-3 was better than that for TS-4. This suggested a more reliable prediction for 2-substituted dipyridodiazepinones than that for 2-pyridone derivatives.

**Table 3.5:** Model quality using the mean absolute error (MAE) based criteria

TS no.	$\left(\frac{MAE}{\text{training set range}}\right)$	$\left(\frac{MAE + 2SD}{\text{training set range}}\right)$	Model quality
1	0.09	0.24	Moderate
2	0.06	0.14	Good
3	0.11	0.20	Good
4	0.11	0.25	Moderate
5	0.12	0.25	Moderate
6	0.10	0.21	Moderate

### 3.3.2 PMF-PLS algorithm performance using 2D charge based descriptors

Different charge based descriptors are known and studied (Todeschini and Consonni, 2008). These include atomic charge descriptors, local dipole moment, charge based topological indices, charge weighted autocorrelation descriptors, charge based measures of solvent accessible surface area (PEOE-VSA) (Estrada, 1995; Labute, 2000; Stanton and Jurs, 1990; Todeschini and Consonni, 2008), etc. These descriptor values are calculated from atomic charges using different methodologies. A QSAR model is then obtained by regressing these molecular descriptor values with the biological activity data of the molecules. The performance of PMF-PLS regression algorithm was also tested with these 2D charge based descriptors. For this purpose, 34 charge weighted autocorrelation descriptors and 21 topological charge indices were calculated using ChemDes (<http://www.scbdd.com/chemdes/>) a web based platform (Dong *et al.*, 2015). Similarly, 14 PEOE-VSA descriptors and 15 atomic charge descriptors were calculated using the web based platform OCHEM (<https://ochem.eu/home/show.do>) (Sushko *et al.*, 2011). Of these 84 charge based descriptors those with constant or near constant values (standard deviation < 0.0001) and ones with at least one missing value were excluded for a given target system (Ojha and Roy, 2018). For TS-1 to TS-6, the resultant pool of descriptors, respectively, were used instead of the PMF descriptors in the PMF-PLS algorithm for regressing with the

corresponding pIC<sub>50</sub> values. The QSAR models developed using the charge based descriptors were then validated by predicting the pIC<sub>50</sub> values of the corresponding TS-1 to TS-6 validation sets. The model performance parameters using these charge based descriptors are presented in Table 3.6 and compared with those of the PFMD based models. It may be observed that the performance of PMF-PLS QSAR algorithm was better overall using PFMDs when compared to that using the 2D charge based descriptors. It may be noted that the 2D charge based molecular descriptors do not contain any information about the 3D conformation of the molecules. On the other hand, the values of molecular field based descriptors studied in the present work (PMF) and CoMFA (Cramer *et al.*, 1988) are dependent on the location of atoms in 3D space thus, taking into consideration the 3D conformation of molecules. The advantage of using molecular field descriptors is that the QSAR model can identify favorable and/or unfavorable regions for the activity of ligands in 3D space (Kubinyi, 1997a). Such inferences cannot be drawn from the QSAR models developed using the other charge based descriptors.

**Table 3.6:** Performance comparison of present QSAR algorithm using Pseudo-field molecular descriptors (PFMDs) and 2D charge based descriptors.

TS no.	2D Charge based descriptors			PFMDs		
	$Q^2_{ext(F1)}$	<i>NRMSEP</i>	MAE based criteria <sup>†</sup>	$Q^2_{ext(F1)}$	<i>NRMSEP</i>	MAE based criteria <sup>†</sup>
1	0.62	0.15	Moderate	0.71	0.13	Moderate
2	0.60	0.11	Good	0.66	0.10	Good
3	0.11	0.14	Bad	0.69	0.10	Good
4	0.55	0.16	Bad	0.62	0.15	Moderate
5	0.62	0.16	Moderate	0.62	0.16	Moderate
6	0.57	0.13	Good	0.63	0.12	Moderate

<sup>†</sup> Using (MAE+2SD) measure, Eq. 3.15

### 3.3.3 Model comparison

QSAR models using different descriptors and modelling approaches have been reported in the literature using the same datasets for the target systems studied in the present work. It is therefore possible to compare the performance of PMF-PLS and

other QSAR models in terms of goodness-of-fit statistics. Three studies for TS-1 (Yi *et al.*, 2008; Elmi *et al.*, 2009), one for TS-2 (Queiroz *et al.*, 2011), two for TS-3 (Hu *et al.*, 2009), and one each for TS-4 (Garg *et al.*, 1999) and TS-5 (Debnath, 1999) were identified for the purpose of studying their comparative performance with PMF-PLS QSAR. It may be noted that no QSAR modelling study could be identified for TS-6. The nature of the QSAR models selected are summarized in Table 3.7. Using the prediction data provided in each case, we calculated the *NRMSEP* and  $Q^2_{ext(F1)}$  using Eq. 2.15 and Eq. 2.17, respectively. The model quality metrics were compared with those obtained for the corresponding PMF-PLS QSAR model as shown in Table 3.7. The comparison of performance metrics shows that PMF-PLS QSAR models are comparable for TS-1 and TS-5 while for the other systems it is even better. Thus the results show that the present PMF-PLS QSAR approach is competitive to the existing methods.

**Table 3.7:** Comparison of present PMF-PLS QSAR model with other QSAR models for the same datasets in this study

TS no.	QSAR model	<i>NRMSEP</i>	$Q^2_{ext(F1)}$	MAE based criteria*	Reference	PMF-PLS QSAR model		
						<i>NRMSEP</i>	$Q^2_{ext(F1)}$	MAE based criteria*
1	CoMFA	0.12	0.74	Moderate	Yi et al., 2008	0.13	0.71	Moderate
	GA-MLR <sup>†</sup>	0.12	0.78	Good	Elmi et al., 2009			
	Fuzzy entropy	0.10	0.85	Good	Elmi et al., 2009			
2	RD-3D-QSAR <sup>‡</sup>	0.22	0.06	Bad	Queiroz et al., 2011	0.10	0.66	Good
3	CoMFA	0.22	0.48	Bad	Hu et al., 2009	0.10	0.69	Good
	CoMSIA <sup>§</sup>	0.21	0.52	Bad	Hu et al., 2009			
4	Physicochemical properties	0.15	0.39	Moderate	Garg et al., 1999	0.15	0.64	Moderate
5	CoMFA	0.15	0.57	Moderate	Debnath, 1999	0.16	0.62	Moderate

<sup>†</sup> Genetic algorithm based feature selection and multilinear regression

<sup>‡</sup> Receptor dependent 3D-QSAR

<sup>§</sup> Comparative molecular similarity indices

\* Using (MAE+2SD) measure, Eq. 3.15

### 3.3.4 Screening of natural compounds

The results seen in Figures 3.7 and 3.8 along with the statistics presented in Tables 3.4, 3.5, 3.6 and 3.7 for TS-1 to TS-6 show that the PMF-PLS algorithm for practical purposes predicts accurately the  $pIC_{50}$  values. It has therefore high potential in realizing applications for screening new molecules with scaffolds similar to those used for model development in different target systems. Natural compounds as drug molecules tend to have fewer side effects as compared to their synthetic counterparts. Therefore, we chose natural compounds as new molecules for screening and present the results of studying the potency of these molecules. The Tanimoto scores obtained to select the compounds with structure similar to the queried scaffold are given in Table 3.8. The predicted  $pIC_{50}$  values for the selected natural compounds for TS-1 to TS-6 using the obtained PMF-PLS QSAR models are also given in Table 3.8. It was observed that the predicted  $pIC_{50}$  values for most of the natural compounds lie within the range of  $pIC_{50}$  values used in training set. Based on these predictions from Table 3.8, natural compounds showing moderate to high predicted  $pIC_{50}$  values were further analyzed by performing docking studies to confirm that these new molecules could bind to the target protein of the TS. The results of docking analysis are discussed in Chapter 4 Section 4.3.2. Docking studies for TS-6 could not be carried out as the structure of the specific target for these compounds is not known.

**Table 3.8:** Tanimoto scores and the predicted biological activities,  $\hat{Y}_{np}$ , using PMF-PLS QSAR model of natural compounds obtained from Super Natural II database

Compound ID	Tanimoto Score	Predicted $pIC_{50}$
TS-1		
SN00011632	0.5909	2.410
SN00054717	0.6190	3.677
SN00058100	0.5909	4.179
SN00118263	0.2500	2.585
SN00226661 <sup>†</sup>	0.5833	7.764
SN00272309 <sup>†</sup>	0.4667	6.929

Table 3.8 Contd...

Compound ID	Tanimoto Score	Predicted pIC <sub>50</sub>
SN00289913	0.4516	6.026
SN00335731	0.4286	5.075
SN00343696	0.8571	6.163
SN00345401	0.3243	2.758
SN00362452 <sup>+</sup>	0.4516	9.051
SN00362911 <sup>+</sup>	0.4516	9.243
TS-2		
SN00160095	0.5000	5.244
SN00304033	0.4688	6.791
SN00335138 <sup>+</sup>	0.5000	8.252
TS-3		
SN00024429	0.5556	1.990
SN00118406 <sup>+</sup>	0.5556	9.852
SN00387398	0.5556	6.107
TS-4		
SN00008627	0.1562	2.045
SN00008635 <sup>+</sup>	0.1389	7.799
SN00008637 <sup>+</sup>	0.1471	9.519
SN00008647 <sup>+</sup>	0.1316	8.529
SN00008665	0.1818	4.128
SN00008860 <sup>+</sup>	0.1724	5.961
SN00009758	0.1935	5.005
SN00010264 <sup>+</sup>	0.1622	8.213
SN00011738	0.1923	4.219
SN00026473	0.1765	2.884
SN00063879 <sup>+</sup>	0.1935	6.205

Table 3.8 Contd...

Compound ID	Tanimoto Score	Predicted pIC <sub>50</sub>
TS-5		
SN00021523	0.2381	18.824
SN00213428	0.4444	4.077
SN00215212 <sup>†</sup>	0.5714	9.845
TS-6		
SN00114856	0.7759	14.363
SN00220696	0.8448	6.191
SN00282305	0.5938	14.389
SN00289590	0.8889	14.879
SN00310837	0.5846	14.420

†- Compounds selected for docking studies

### 3.4. Conclusions

The methodology of PMF-PLS is seen to offer a simpler way of QSAR modelling that uses an effective correlative descriptor in terms of the intrinsic properties of atoms, namely, the electron affinity and electronegativity values. This is in contrast to CoMFA where the descriptors are obtained using the partial atomic charges which are calculated separately for every molecule. We apply the PMF-PLS methodology to six target systems, namely, 4-phenylpyrrolocarbazole derivative inhibitors of WEE1 as anti-cancer compounds, benzylpiperidine derivative inhibitors of AChE against neurological disorders, 2-substituted dipyridodiazepinone derivatives and 2-pyridinone derivatives as HIV-1 RT inhibitors, cyclic urea derivatives as HIV-1 PR inhibitors and azilide derivatives as anti-malarial compounds. The QSAR models showed good prediction statistics for all six TSs and it brings out the viability of the PMF-PLS approach. It takes care of many practical situations encountered in QSAR modelling. Thus, the high dimensionality of the descriptor data could be reduced drastically by projection to a lower dimensional



latent subspace. The practical problem of overfitting of model could then be addressed. The usefulness of Procrustes transformation in modifying the descriptor data for better optimization of PLS scores and loadings has been proposed which gave improved predictions. A comparison of the PMF-PLS QSAR modelling results with the QSAR models reported in the literature for the same set of inhibitors shows that the former yields comparable results. Additionally, PMF-PLS QSAR models were used to predict  $pIC_{50}$  values for natural compounds with unknown biological activities. The time taken for the PMF-PLS algorithm to arrive at the reference training and test sets (first part of the algorithm) was in the order of 6-8 hours. However, the second and third part of the algorithm took about 1-2 minutes to complete. Thus, when compared to the MDS based 2D-QSAR models described in Chapter 2 performance of PMF-PLS QSAR models was superior in terms of consistency in their predictions across the target systems and the computational time required for building the regression model. Thus, the PMF-PLS method for QSAR modelling is a powerful computational tool that has a high potential to screen new molecules for experimentation in ligand based drug discovery programs.

## **Chapter 4**

# **Varying component PLS QSAR modelling and docking studies of potential inhibitors**

#### 4.1 Introduction:

During PLS regression the number of components ( $a$ ) to be used for regression is fixed. As seen in the PMF-PLS algorithm (Chapter 3, Section 3.2.4) with the changes made to the training set the value of  $a$  is fixed for all the PLS models. It may be that with changing training sets the number of components for which an optimal model performance is achieved could also vary. The aim of the study in this chapter was to analyze the performance of the QSAR models developed using varying values of  $a$  during their formulation.

As seen in Chapter 3 Section 3.4.2, using the SIMPLS method for PLS regression (de Jong, 1993), we obtain the regression coefficients in terms of  $X$  data. The advantage of SIMPLS method is that the dimensionality of regression coefficients ( $m+1, 1$ ) is retained irrespective of the number components ( $a$ ) used for optimizing the model. Thus, it may be possible to optimize PLS regression models with varying  $a$  values and still obtain the same number of regression coefficients similar to those observed in the multiple linear regression (MLR) formalism (Alvin C. Rencher, 2002). In MLR the relation between  $m$  independent variables,  $x$ , and a single dependent variable,  $y$ , for  $n$  observations (molecules) is given as,

$$\left. \begin{aligned} y_1 &= b_0 + b_1x_{11} + b_2x_{12} + \dots + b_mx_{1m} \\ y_2 &= b_0 + b_1x_{21} + b_2x_{22} + \dots + b_mx_{2m} \\ &\vdots \\ y_n &= b_0 + b_1x_{n1} + b_2x_{n2} + \dots + b_mx_{nm} \end{aligned} \right\} \quad (4.1)$$

where  $b_i$  are the regression coefficients with  $i = 0, 1, 2, \dots, m$ . In matrix notation above equation may be written as,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{bmatrix} \quad (4.2)$$

Or,

$$Y = XB \quad (4.3)$$

Equation 4.2 can be solved for the values of  $B$  to obtain the regression model connecting  $X$  and  $Y$ . However, due to the problems of high dimensionality and collinearity of  $X$  data associated with the molecular field based descriptors in both PFMDs and CoMFA, the dimensionality reduction methods like PCA and PLS are employed. During PCA only the data from independent variables ( $X$ ) is taken into account for calculation of scores ( $T$ ) in the reduced dimensions and the number of principal components ( $a$ ) only determines the contribution of those component to the total variance in  $X$  (Yoo and Shahlaei, 2018). Thus, the data in the principal component space ( $T$ ) may not necessarily provide the most relevant information for regression with  $Y$  and it is possible that the information relevant for the regression may be lost with the data along the  $i^{\text{th}}$  principal components that are not considered in the calculations ( $i > a$ ) (Geladi and Kowalski, 1986). PLS regression, on the other hand, determines the orthogonal components for  $X$  in the order of their correlation to  $Y$  keeping the most relevant information for regression, and hence is preferred over PCA.

Since we obtain the regression coefficient in terms of the original data ( $X$ ) the regression model determined using SIMPLS method in Eq. 3.10 is similar to the MLR formula in Eq. 4.3. However, it may be noted that only the data from the selected components are used to arrive at the solution. We use these regression coefficients to formulate a new algorithm to arrive at a regression model. We initially divide the data randomly into a training and a test set and then perform leave-one-out cross-validation for the training set. However, during leave-one-out cross-validation of the model the optimal number of components is selected based on the best prediction of  $\text{pIC}_{50}$  value of the molecule that is left out of the training set to get a set of regression coefficients. We discuss in detail this novel method with varying number of components for PLS regression termed as varying component PLS (VC-PLS) regression.

## 4.2 Methodology

### 4.2.1 Molecular descriptors

We perform the studies in this chapter using the PFMDs developed for all the 6 Target systems as discussed in Chapter 3 Sections 3.2.2 and 3.2.3.

### 4.2.2 QSAR modelling

Figure 4.1 shows the flowchart of the methodology used for QSAR modelling. The boxes in the Figure 4.1 are numbered using the scheme similar to those in Figure 3.3. Initially we start with randomly dividing the data into the training ( $\mathbf{X}_{train}, \mathbf{Y}_{train}$ ) and test set ( $\mathbf{X}_{test}, \mathbf{Y}_{test}$ ) [B 1.1] and [B 1.2]. In the next step  $j^{th}$  molecule ( $j = 1$ ) ( $x_j, y_j$ ) was removed from the training set to get modified training set ( $\mathbf{X}^{-train,j}, \mathbf{Y}^{-train,j}$ ) [B 1.4]. PLS regression was performed with varying  $a = 5$  to 40 [B 1.6] to obtain the corresponding regression coefficients  $B_{j,a}$ . SIMPLS method (de Jong, 1993) was for PLS regression. These regression coefficients were then used to predict the pIC<sub>50</sub> value ( $\hat{y}_{train,j,a}$ ) of the  $j^{th}$  molecule for each set of regression coefficients  $B_{j,a}$ ,  $a = 5, 6, \dots, 40$  [B 1.7]. Corresponding error values  $e_{j,a}$  were also calculated for every  $\hat{y}_{train,j,a}$  [B 1.7]. The regression coefficients corresponding to the minimum value of  $e_{j,a}$ ,  $a = 5, 6, \dots, 40$ , were assigned to  $B_{min,j}$  [B 1.9] and [B 1.10] for the  $j^{th}$  molecule of the training set. This process ([B 1.4] to [B 1.10]) was repeated for all the molecules in the training set [B 1.11] and an average of all the sets of  $B_{min,j}$  with  $j = 1, 2, \dots, n_{train}$  was assigned to  $B_i$  as the regression coefficients for that training set [B 1.12].  $B_i$  was then used to predict  $\hat{Y}_{train}$  and  $\hat{Y}_{test}$  [B 1.13] in order to calculate the model performance parameters [B 1.14]. This process ([B 1.2] to [B 1.14]) was iterated for a number ( $i_{lim}$ ) of combinations of training ( $\mathbf{X}_{train}, \mathbf{Y}_{train}$ ) and test sets ( $\mathbf{X}_{test}, \mathbf{Y}_{test}$ ) to obtain  $i_{lim}$  number of models. The model performance parameters were calculated as described in the Chapter 2 Section 2.2.10. Of these models the 5 best performing models were selected for the screening of natural compounds.

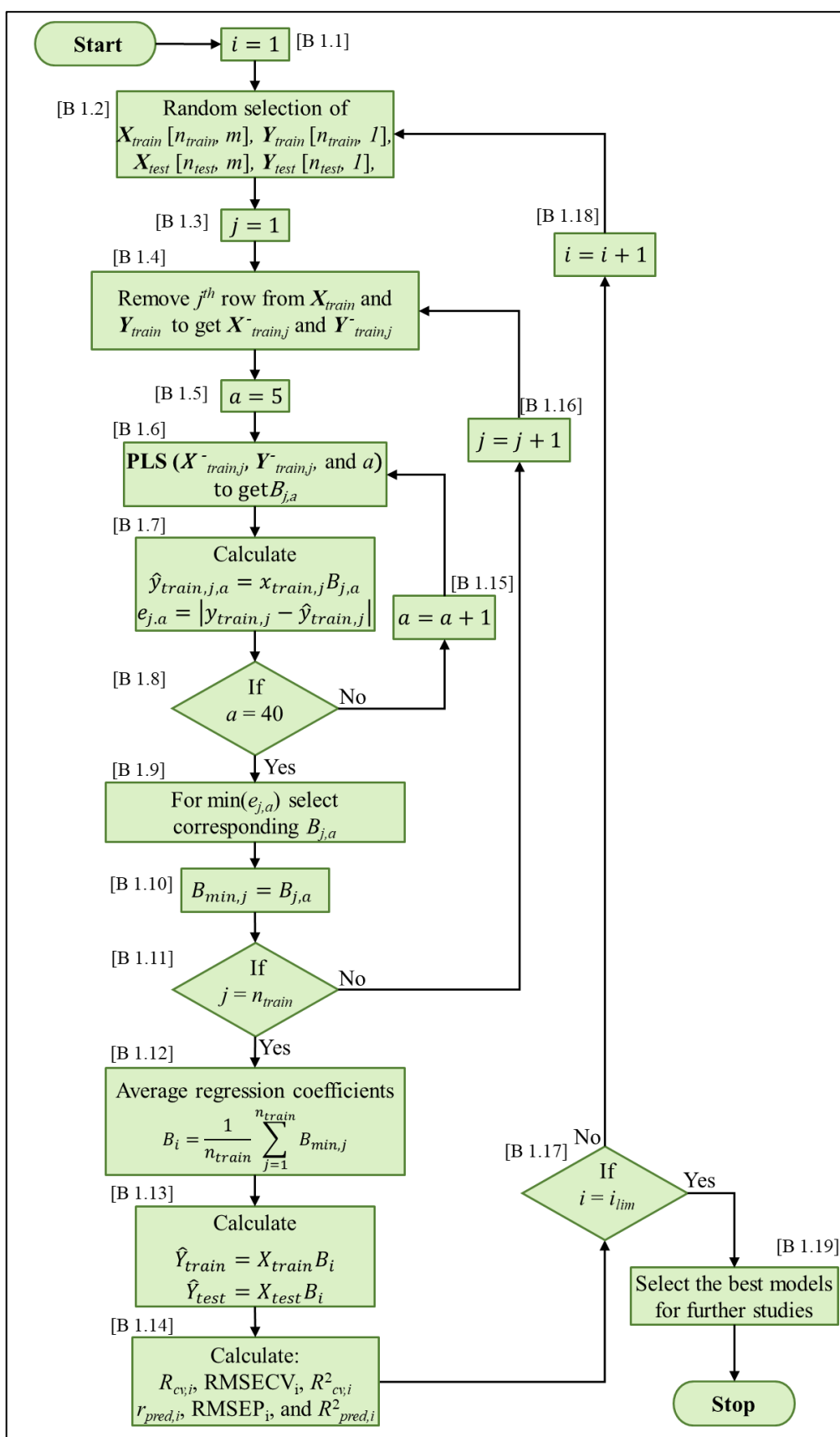


Figure 4.1: Flowchart of QSAR modelling by VC-PLS

### 4.2.3 Docking simulations

For docking simulations, using AutoDock Vina (Trott and Olson, 2010) surface pocket identification of Wee1A kinase (PDB ID: 1X8B, (Squire *et al.*, 2005)), AChE (PDB ID: 4M0E, (Cheung *et al.*, 2013)), HIV-1 Reverse transcriptase (PDB ID: 1VRT, (Esnouf *et al.*, 1995) HIV-1 Protease (PDB ID: 1AJX, (Bäckbro *et al.*, 1997)) co-crystallized with the ligands were carried out on servers CASTp (Dundas *et al.*, 2006), Pocket-Finder and QSiteFinder (Laurie and Jackson, 2005). The ligand free protein models were generated through Schrodinger by removing the ligand structure from the complex Docking protocol and parameters were standardized by performing docking simulation of 9-hydroxy-4-phenylpyrrolo[3,4-C]carbazole-1,3(2h,6h)-dione, dihydrotanshinone I, nevirapine and AHA001 with ligand free Wee1A kinase, AChE, HIV-1 Reverse transcriptase and HIV-1 Protease, respectively. The ligand free structure of 1X8B, 4M0E, 1VRT, 1AJX were first processed to set protonation states of amino acids with polar side chains to neutral pH. Grid Box parameters and center with grid spacing 1.0 Å were set for validation (Table 4.1). Gasteiger charges assigned to protein and ligand. Exhaustiveness level was set on 8 and a computer with four processors was utilized for the computations. A total of 90 docked poses of individual ligands to the ligand free proteins were generated and compared with co-crystal structure of the complex 1X8B 4M0E, 1VRT and 1AJX. Blind docking simulations of ligands with ligand free proteins were carried out using the standardized docking parameters obtained. Based on the outputs of blind docking, refined docking simulation were performed with grid parameters as mentioned in Table 4.1. The protein-ligand interactions were analyzed and visualized using Discovery Studio visualizer 4.0 client.

**Table 4.1:** Docking parameters

	Protein	Grid box Size (X x Y x Z)	Grid Box Center (X,Y,Z)	Grid spacing (Å)
Validation and Blind Docking	Wee1 kinase	38 x 58 x 46	4.801, 47.267, 23.191	1.0
	AcHE	16 x 34 x 22	-20.43, -43.472, 24.694	1.0
	HIV-1 RT	24 x 26 x 22	5.722, -31.417, 15.861	1.0
	HIV-1 PR	40 x 40 x 46	12.665, 27.18, 7.389	1.0
Refined Docking	Wee1 kinase (Site 1)	20 x 24 x 18	0.506, 52.928, 21.592	1.0
	Wee1 kinase (Site 2)	16 x 20 x 22	-5.007, 48.166, 44.561	1.0
	AcHE	28 x 22 x 18	17.33, -49.0, -24.306	1.0
	HIV-1 RT	24 x 26 x 22	5.722, -31.417, 15.861	1.0
	HIV-1 PR	40 x 40 x 46	12.665, 27.18, 7.389	1.0

## 4.3 Results and discussion

### 4.3.1 QSAR modelling

In Chapter 3 when employing the PMF-PLS algorithm we arrive at the regression model by making changes to the best performing training set and obtaining one set of regression coefficients for every cycle of these changes. It may be observed that during the steps of PMF-PLS algorithm we fix the number of PLS components to be used for regression through the entire process. However, there is a possibility that with different training sets the number of PLS components needed for optimal performance of the regression model could be different. In this chapter we attempt to formulate the PLS regression models optimized over varying values of  $a$ .

We do this by initially selecting the training and test set randomly and then performing a modified version of leave-one-out cross-validation routine described in Section 4.2.2. One training set molecule is removed and the PLS regression models are formulated for a range of  $a$  values (5-40) using the SIMPLS method. The regression coefficients of the model that predicts the  $pIC_{50}$  value of the molecule that was removed out of the training set with least error were selected. This procedure was



repeated for every molecule in the training set and the selected regression coefficients were averaged to get the final regression model for that training set. Multiple training sets were selected randomly for which regression models were obtained. These regression models were validated by predicting the pIC<sub>50</sub> values ( $\hat{Y}_{\text{test}}$ ) of the corresponding test sets. The model performance parameters of the best model for each of the six target systems are given in Table 4.2. It was observed that the performance parameters of the best models for TS-1 to TS-4 were comparable to those observed for the corresponding PMF-PLS QSAR models. The model performance was observed to be significantly improved for TS-5 (HIV-1 protease inhibitors), whereas, that for TS-6 (anti-malaria) was found to be poor as compared to that of the PMF-PLS. However, the comparison of the *NRMSEP* and  $r_{\text{pred}}$  values suggest an overall stable model performance for all the TSs. The model performance parameters for the five best performing models for TS-1 to Ts-6 are summarized in Tables 4.3 – 4.8, respectively. The experimental and predicted pIC<sub>50</sub> values of all the molecules of TS-1 to TS-6 using VC-PLS QSAR model are given in Appendix Tables A21 to A26.

**Table 4.2:** VC-PLS QSAR model fitting statistics for the best models of TS-1 to TS-6

TS	AID	Target	Cross-validation			External Validation		
			$r_{cv}$	$R^2_{cv}$	<i>NRMSECV</i>	$r_{pred}$	$Q^2_{\text{ext}(F1)}$	<i>NRMSEP</i>
1	268838	Wee1	0.91	0.80	0.11	0.87	0.76	0.13
2	566585	AChE	0.94	0.83	0.09	0.95	0.87	0.07
3	198247	HIV-1 RT	0.84	0.68	0.12	0.81	0.64	0.11
4	197804	HIV-1 RT	0.88	0.74	0.12	0.75	0.56	0.15
5	160292	HIV-1 PR	0.87	0.73	0.12	0.88	0.78	0.12
6	579588	Anti-malarial	0.90	0.73	0.11	0.65	0.40	0.15

**Table 4.3:** Model fitting statistics for the five best VC-PLS models for TS-1

Model	Cross-validation			External Validation		
	$r_{cv}$	$R^2_{cv}$	$NRMSECV$	$r_{pred}$	$Q^2_{ext(F1)}$	$NRMSEP$
1	0.91	0.80	0.11	0.87	0.76	0.13
2	0.89	0.77	0.12	0.86	0.70	0.14
3	0.92	0.83	0.10	0.84	0.68	0.14
4	0.90	0.79	0.11	0.84	0.66	0.15
5	0.92	0.83	0.10	0.79	0.62	0.15

**Table 4.4:** Model fitting statistics for the five best VC-PLS models for TS-2

Model	Cross-validation			External Validation		
	$r_{cv}$	$R^2_{cv}$	$NRMSECV$	$r_{pred}$	$Q^2_{ext(F1)}$	$NRMSEP$
1	0.94	0.83	0.10	0.95	0.87	0.07
2	0.94	0.83	0.10	0.86	0.72	0.11
3	0.95	0.85	0.09	0.90	0.65	0.12
4	0.89	0.72	0.12	0.79	0.63	0.13
5	0.92	0.79	0.11	0.75	0.55	0.13

**Table 4.5:** Model fitting statistics for the five best VC-PLS models for TS-3

Model	Cross-validation			External Validation		
	$r_{cv}$	$R^2_{cv}$	$NRMSECV$	$r_{pred}$	$Q^2_{ext(F1)}$	$NRMSEP$
1	0.85	0.68	0.12	0.81	0.64	0.11
2	0.83	0.66	0.12	0.77	0.61	0.13
3	0.80	0.61	0.13	0.72	0.40	0.19
4	0.83	0.65	0.13	0.62	0.38	0.16
5	0.86	0.69	0.12	0.74	0.35	0.17

**Table 4.6:** Model fitting statistics for the five best VC-PLS models for TS-4

Model	Cross-validation			External Validation		
	$r_{cv}$	$R^2_{cv}$	$NRMSECV$	$r_{pred}$	$Q^2_{ext(F1)}$	$NRMSEP$
1	0.88	0.74	0.12	0.75	0.56	0.15
2	0.89	0.76	0.12	0.75	0.55	0.17
3	0.90	0.78	0.11	0.72	0.50	0.17
4	0.89	0.76	0.12	0.70	0.48	0.17
5	0.88	0.74	0.13	0.67	0.40	0.15

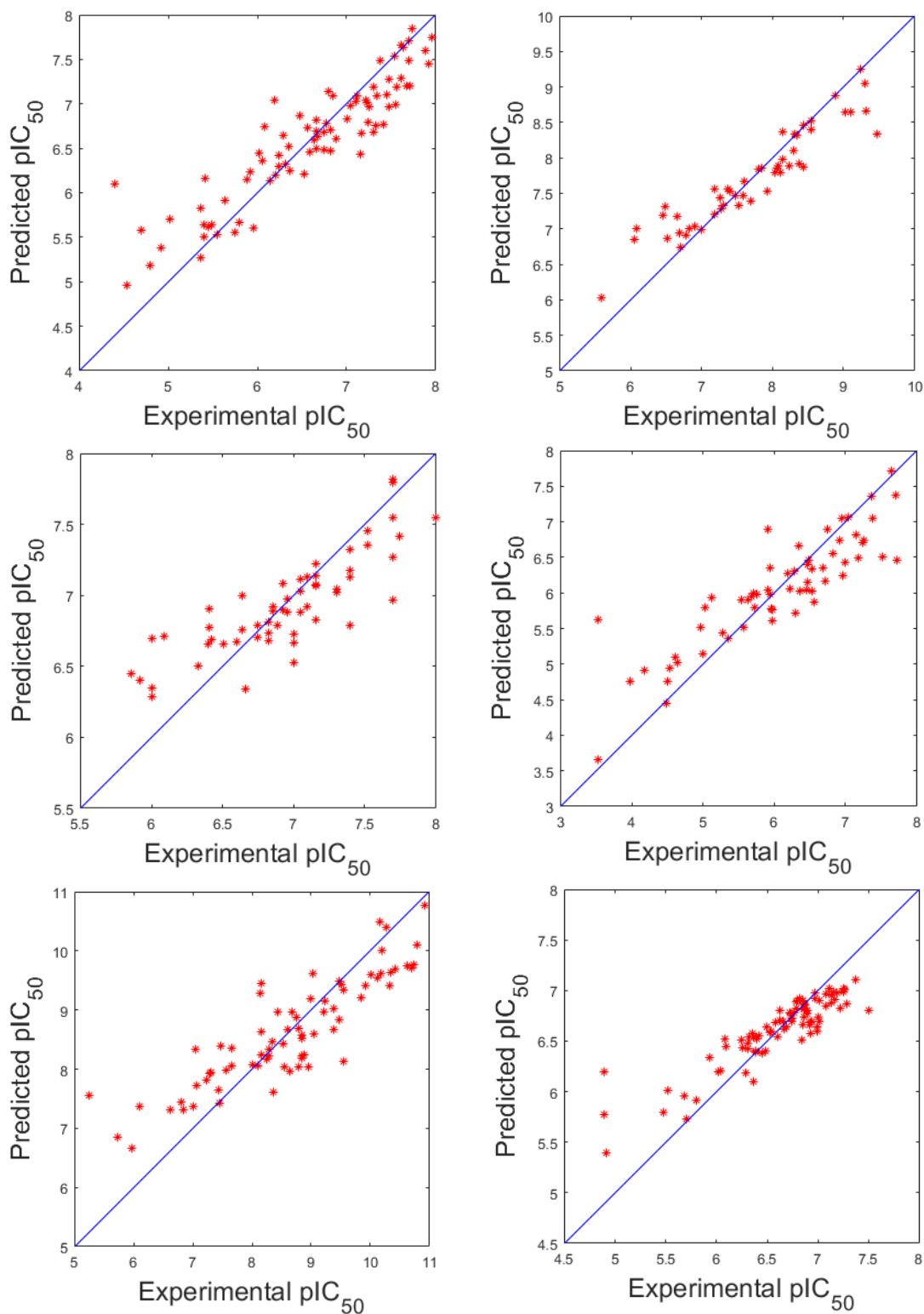
**Table 4.7:** Model fitting statistics for the five best VC-PLS models for TS-5

Model	Cross-validation			External Validation		
	$r_{cv}$	$R^2_{cv}$	$NRMSECV$	$r_{pred}$	$Q^2_{ext(F1)}$	$NRMSEP$
1	0.87	0.73	0.12	0.88	0.78	0.12
2	0.89	0.75	0.12	0.87	0.73	0.11
3	0.91	0.78	0.11	0.86	0.72	0.13
4	0.89	0.76	0.11	0.82	0.67	0.13
5	0.90	0.77	0.11	0.83	0.63	0.14

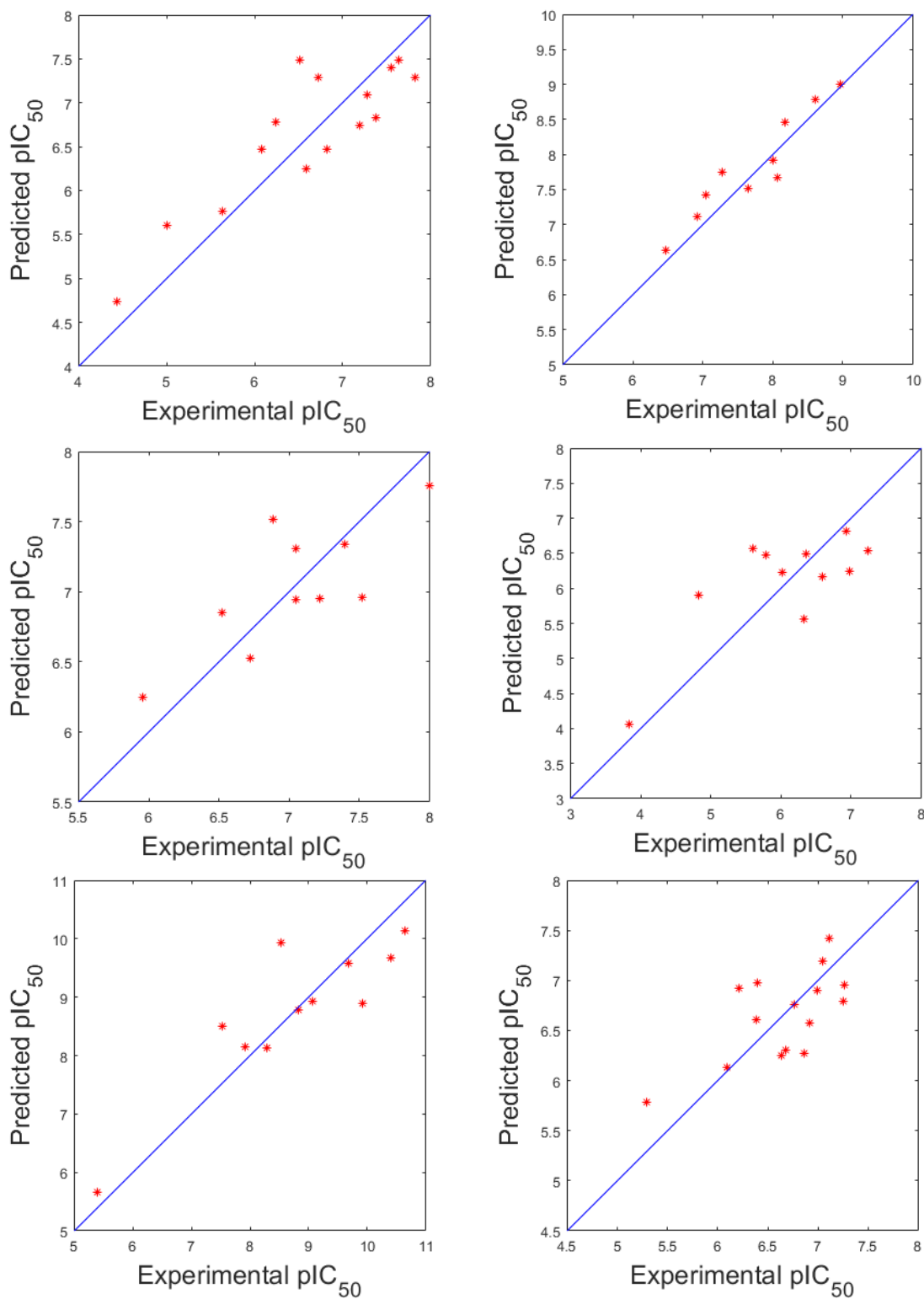
**Table 4.8:** Model fitting statistics for the five best VC-PLS models for TS-6

Model	Cross-validation			External Validation		
	$r_{cv}$	$R^2_{cv}$	$NRMSECV$	$r_{pred}$	$Q^2_{ext(F1)}$	$NRMSEP$
1	0.90	0.73	0.11	0.65	0.40	0.15
2	0.87	0.71	0.11	0.68	0.39	0.15
3	0.89	0.71	0.13	0.65	0.34	0.15
4	0.86	0.67	0.12	0.71	0.32	0.19
5	0.89	0.74	0.11	0.60	0.32	0.14

Similarly, Figures 4.2 and Figures 4.3 represent the diagonal plots of the experimental  $pIC_{50}$  values *vs.* the predicted  $pIC_{50}$  values for all the TSs for training and test sets respectively for the best VC-PLS models. It was observed that the time taken for performing the VC-PLS algorithm for  $i_{lim} = 100$  ([B 1.17]) was 6-8 hours.



**Figure 4.2:** Plots of actual pIC<sub>50</sub> values ( $Y_{train}$ ) vs. the predicted values ( $\hat{Y}_{train}$ ) for cross-validation for the best VC-PLS model for (A) TS-1, (B) TS-2, (C) TS-3, (D) TS-4, (E) TS-5 and (F) TS-6 inhibitors.



**Figure 4.3:** Plots for actual pIC<sub>50</sub> values ( $Y_{test}$ ) vs. the predicted values ( $\hat{Y}_{test}$ ) for test sets using VC-PLS QSAR model of (A) TS-1, (B) TS-2, (C) TS-3, (D) TS-4, (E) TS-5 and (F) TS-6 inhibitors.

### 4.3.2 Screening of natural compounds and docking studies

These five best performing VC-PLS models were selected for further prediction of the pIC<sub>50</sub> values of the natural compounds obtained from the SuperNatural II database. The predicted pIC<sub>50</sub> values of these natural compounds are given in Table 4.9

**Table 4.9:** Predicted pIC<sub>50</sub> values,  $\hat{Y}_{np}$ , of natural compounds obtained from Super Natural II database using the five VC-PLS QSAR models.

Compound ID	Predicted pIC <sub>50</sub>						
	Model 1	Model 2	Model 3	Model 4	Model 5	Average	PMF-PLS
TS-1							
SN00011632	3.707	4.238	3.765	3.786	4.344	3.968	2.410
SN00054717	4.887	5.150	4.762	4.775	5.259	4.967	3.677
SN00058100	5.262	5.129	5.054	4.913	5.360	5.144	4.179
SN00118263	4.497	3.126	3.804	3.554	4.067	3.810	2.585
SN00226661 <sup>†</sup>	7.583	7.887	8.172	7.690	7.909	7.848	7.764
SN00272309 <sup>†</sup>	7.095	7.214	7.749	7.397	7.229	7.337	6.929
SN00289913	6.068	6.312	5.930	5.568	6.315	6.039	6.026
SN00335731	6.019	6.411	5.536	5.894	5.628	5.898	5.075
SN00343696	6.413	6.358	6.354	6.362	6.532	6.404	6.163
SN00345401	1.435	-1.306	1.219	-0.987	-0.661	-0.060	2.758
SN00362452 <sup>†</sup>	7.145	7.167	7.934	7.488	7.030	7.353	9.051
SN00362911 <sup>†</sup>	8.626	8.400	9.238	8.796	8.122	8.636	9.243
TS-2							
SN00160095	5.577	4.848	3.928	5.601	5.054	5.002	5.244
SN00304033	6.303	6.488	5.574	6.061	6.504	6.186	6.791
SN00335138 <sup>†</sup>	8.746	7.788	7.052	8.021	7.605	7.842	8.252
TS-3							
SN00024429	3.877	4.810	5.470	5.494	1.114	4.153	1.990
SN00118406 <sup>†</sup>	7.615	7.897	7.522	7.641	9.024	7.940	9.852
SN00387398	5.909	5.907	5.684	6.214	5.802	5.903	6.107

Table 4.9 Contd...

Compound ID	Predicted pIC <sub>50</sub>						
	Model 1	Model 2	Model 3	Model 4	Model 5	Average	PMF-PLS
TS-4							
SN00008627	3.640	5.598	4.974	4.983	4.332	4.705	2.045
SN00008635 <sup>†</sup>	7.770	9.988	10.559	10.015	9.453	9.557	7.799
SN00008637 <sup>†</sup>	8.301	10.477	11.166	10.566	10.238	10.150	9.519
SN00008647 <sup>†</sup>	7.605	10.207	10.399	10.435	9.759	9.681	8.529
SN00008665	5.118	5.359	5.801	5.314	5.280	5.374	4.128
SN00008860 <sup>†</sup>	6.524	7.013	7.508	6.952	6.873	6.974	5.961
SN00009758	5.421	5.986	5.351	5.239	5.278	5.455	5.005
SN00010264 <sup>†</sup>	7.409	10.603	10.619	10.747	9.956	9.867	8.213
SN00011738	5.226	5.428	5.544	5.771	5.671	5.528	4.219
SN00026473	3.674	4.854	4.529	5.799	5.859	4.943	2.884
SN00063879 <sup>†</sup>	6.462	7.640	7.565	6.472	6.782	6.984	6.205
TS-5							
SN00021523	14.327	14.605	15.264	14.485	13.605	14.457	18.824
SN00213428	5.574	5.581	6.678	5.234	6.345	5.882	4.077
SN00215212 <sup>†</sup>	9.426	9.183	8.994	9.495	10.136	9.447	9.845
TS-6							
SN00114856	7.860	6.736	6.020	7.610	5.811	6.807	14.363
SN00220696	5.509	5.317	4.742	5.669	4.468	5.141	6.191
SN00282305	8.004	6.712	6.136	7.700	5.858	6.882	14.389
SN00289590	8.271	6.678	6.443	8.000	6.127	7.104	14.879
SN00310837	7.897	6.740	6.095	7.719	5.898	6.870	14.420

†- Compounds selected for docking studies

The last column of Table 4.9 gives the pIC<sub>50</sub> values of the natural compounds predicted using the PMF-PLS model studied in Chapter 3. Comparison of the pIC<sub>50</sub> values predicted using the VC-PLS models and those using PMF-PLS model shows that there is consistency in the predictions of all the five VC-PLS models and the PMF-PLS models for TS-1 to TS-5. In case of TS-6 a large difference was observed in the pIC<sub>50</sub>

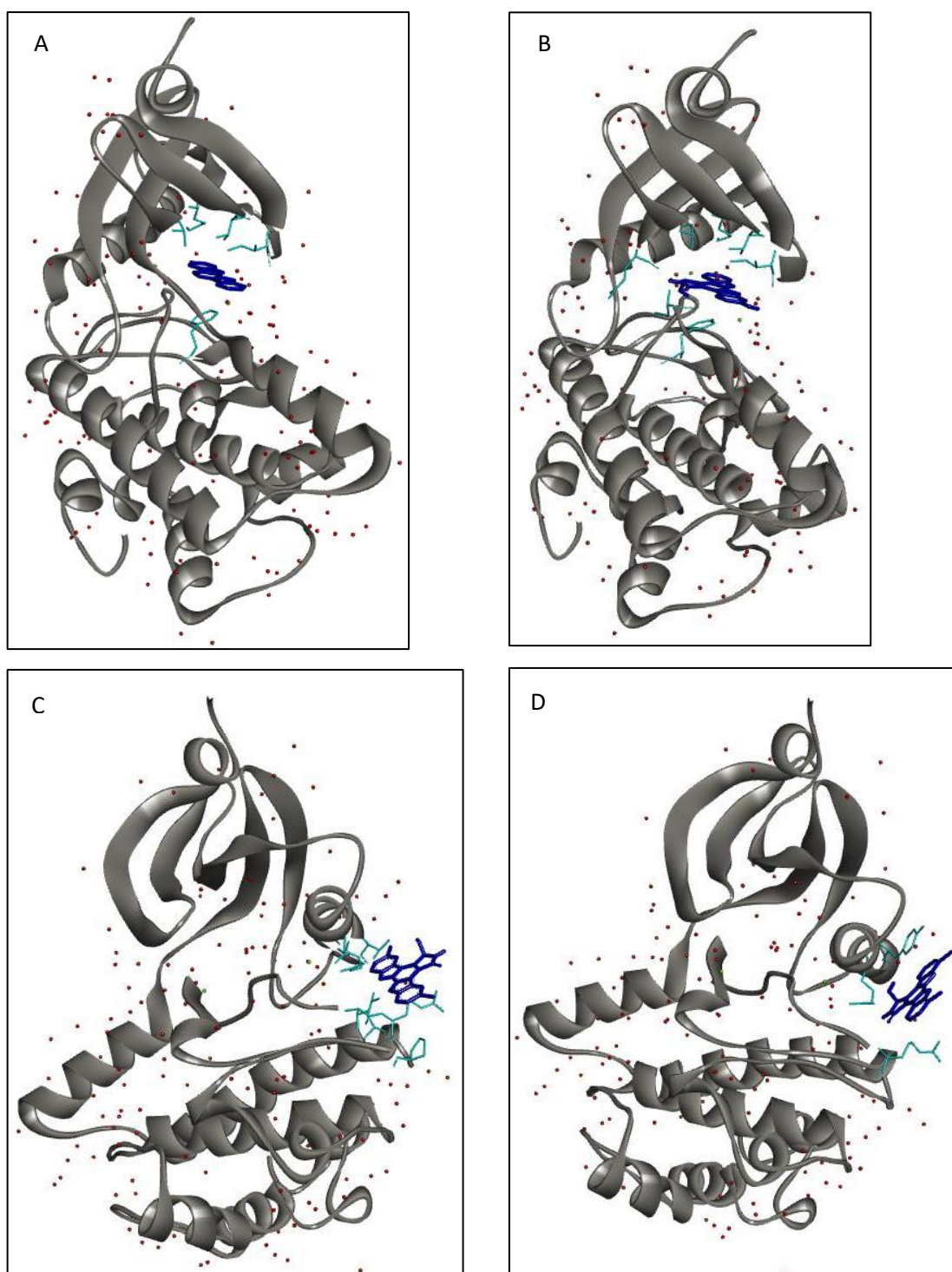
values predicted using VC-PLS models and those using PMF-PLS model. However, the overall trend in the predictions was observed to be the same, i.e., the compounds predicted to have high pIC<sub>50</sub> values by PMF-PLS QSAR model were also predicted to have high pIC<sub>50</sub> values by the VC-PLS QSAR models. Natural compounds with moderate to high predicted pIC<sub>50</sub> values were selected for further analysis through docking studies. Docking studies for TS-6 could not be carried out as the structure of the specific target for these compounds is not known.

#### 4.3.2.1 Docking results for TS-1

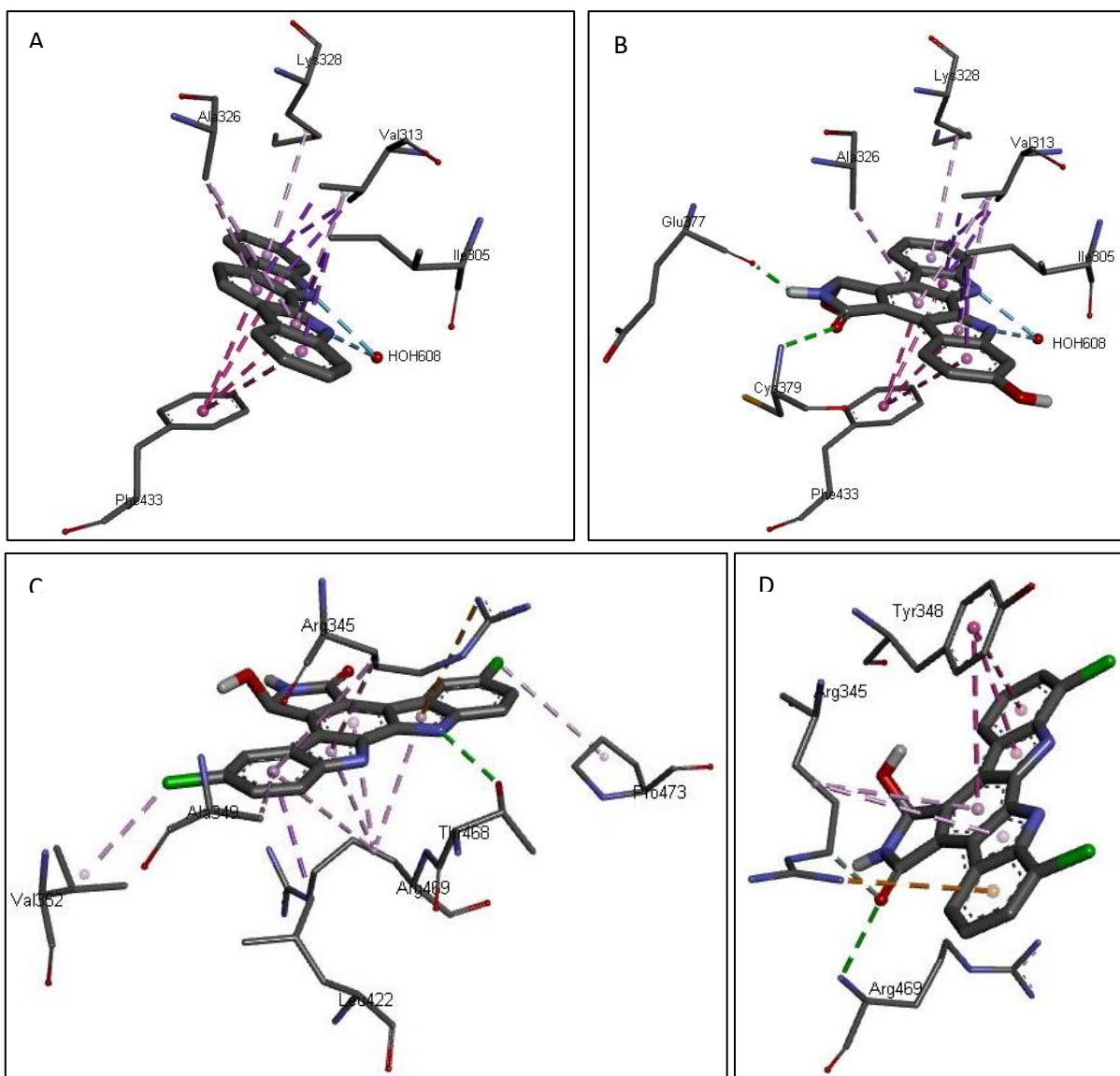
Of the 12 natural compounds obtained from the SuperNatural-II database for TS-1, the natural compounds SN00226661, SN00272309, SN00362452 and SN00362911 were predicted to have high pIC<sub>50</sub> values indicating good inhibitory potential. Other natural compounds were predicted to have low pIC<sub>50</sub> values suggesting low inhibition and were therefore not considered for further studies. The docking for these compounds was carried out on the X-ray crystal structure of Wee1, 1X8B (Squire *et al.*, 2005), obtained from PDB. It was observed that compounds SN00226661 and SN00272309 docked in the active site cleft of the protein (Figure 4.4A and B), whereas compounds SN00362911 and SN00362452 docked to a peripheral site on the protein (Figure 4.4C and D). The binding energies of the compounds were in the range of -7.3 to -12.8Kcal/mol suggesting good inter-action of the compounds with Wee1 (Table 4.10). The detailed interaction of the docked compounds are shown in Figure 4.5 and listed in Table 4.10.

Protein kinase Wee1 has a kinase domain from amino acid residue 291 to 575. The active site cleft of Wee1 consists of 5 stranded  $\beta$ -sheets and a glycine rich loop. Residues 422 to 433 form the catalytic segment spanning from  $\beta$ 6 strand to the beginning of  $\beta$ 7. Asp426 is the catalytic residue and Asn431 and Asp463 are metal ion binding residues binding each to an Mg<sup>2+</sup> ion. Activation segment, a 25 residue large loop from 462 to 486, provides the substrate binding platform. Model studying ATP binding with Wee1 (Squire *et al.*, 2005) has also suggested that adenine ring of substrate ATP interacts with the Ile305, Val313, Ala326 and Phe433.





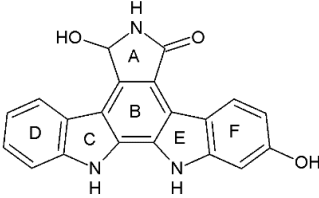
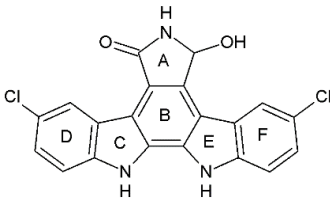
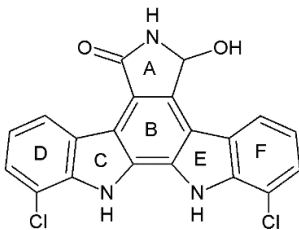
**Figure 4.4:** Natural compounds docked in the active cleft of Wee1 A) compound SN00226661 and B) compound SN00272309. Natural compounds docked in the peripheral site of Wee1 C) compound SN00362911 and D) compound SN00362452. Natural compounds are displayed in dark blue color whereas the Wee1 residues interacting with the compound are shown in light blue.



**Figure 4.5:** Detailed view of active cleft residues of Wee1 interacting with the docked natural compounds A) SN00226661 and B) SN00272309 and peripheral site residues of Wee1 interacting with natural compounds C) SN00362452 and D) SN00362911

**Table 4.10:** Interactions between the docked natural compounds and Wee1 protein residues

Compound Id.	Structure	Interactions	Binding Energy (Kcal/mol)
Binding position: Active cleft			
SN00226661		<ol style="list-style-type: none"> <li>Ile305 <math>\pi</math>-<math>\sigma</math> with C and D</li> <li>Val313 <math>\pi</math>-alkyl with C, <math>\pi</math>-<math>\sigma</math> with B,E and F</li> <li>Ala326 <math>\pi</math>-alkyl with B and C</li> <li>Lys328 <math>\pi</math>-alkyl with F</li> <li>Phe433 <math>\pi</math>-<math>\pi</math>-stacking with B,C,D and E</li> </ol>	-10.5

		6. H <sub>2</sub> O H-bond with NH of C and E	
		1. Iel305 $\pi$ - $\sigma$ with E and F	
		2. Val313 $\pi$ -alkyl with B and E, $\pi$ - $\sigma$ with C and D	
		3. Ala326 $\pi$ -alkyl with B	
		4. Lys328 $\pi$ -alkyl with D	
		5. Glu377 H-bond with NH of A	
SN00272309		6. Cys379 H-bond with =O of A	-12.8
		7. Phe433 $\pi$ - $\pi$ -stacking with B,C,E and F	
		8. H <sub>2</sub> O H-bond with NH of C and E	
<b>Binding position: Peripheral site</b>			
		1. Arg345 $\pi$ -alkyl with E and F, $\pi$ -cation with C	
		2. Ala349 $\pi$ -alkyl with F	
		3. Val352 alkyl with Cl of F	
		4. $\pi$ - $\sigma$ with F	
		5. Thr468 H-bond with NH of C	
SN00362452		6. Arg469 $\pi$ -alkyl with B,E and F	-7.6
		7. Pro473 alkyl with Cl of D	
		1. Arg345 $\pi$ -alkyl with B and C, $\pi$ -cation with D, carbon with =O of A	
		2. Tyr348 $\pi$ - $\pi$ -stacking with B,E and F	
SN00362911		3. Arg469 H-bond with =O of A	-7.3

It can be observed from the interactions listed in Table 4.10 that compounds SN00226661 and SN00272309, which docked to the active site cleft of the protein, interacted with the above mentioned ATP binding residues. Similarly, compounds SN00362911 and SN00362452, which docked to the peripheral site, were observed to interact with the residues of the activation segment of the protein. The above interactions of natural compounds with the residues suggest either a competitive blocking of active site of the protein (docking in the active site cleft) or change in the

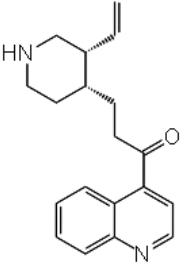
conformation of the activation segment of the protein (docking in the peripheral site) which could result in inhibition of enzyme activity as reflected in high pIC<sub>50</sub> values.

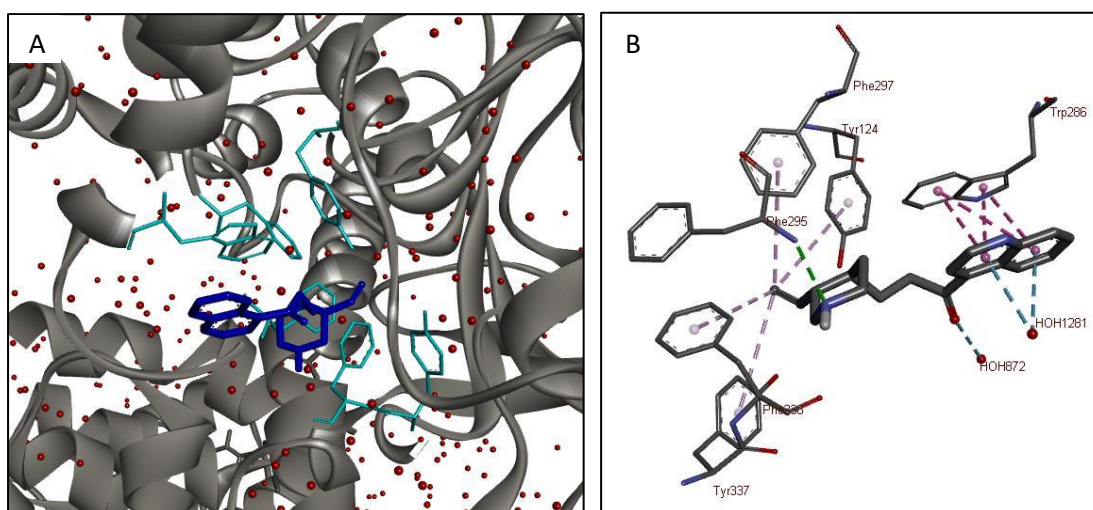
#### 4.3.2.2 Docking results for TS-2

Of the three natural compounds selected for TS-2 compound SN00335138 was found to have a moderate predicted pIC<sub>50</sub> value (Table 4.9). The pIC<sub>50</sub> value for other three compounds were predicted to be low and hence were not studied further.

AChE active site is a gorge of about 20Å deep and is comprised of two sites namely, peripheral anionic site (PAS) and catalytic site (CS) (Marco-Contelles *et al.*, 2014). PAS is present at the mouth of the gorge and is rich in aromatic amino acids. Cationic substrates are trapped transiently to this site before being transferred to the catalytic site. The rate of catalysis is accelerated due to this transient binding. Mixed non-competitive inhibitors of AChE that bind to the PAS limit the rate of catalysis by creating a steric blockage for association of substrates and dissociation of products. Catalytic site is situated at the bottom of the active site gorge and is made up of two sub-sites, namely esteratic site where the catalytic triad of Ser203, Glu344 and His447 is located and anionic binding site where Trp86 is located (Marco-Contelles *et al.*, 2014). X-ray structures and models studying binding of various AChE inhibitors, including donepezil, an AChE inhibiting drug in the market (Cheung *et al.*, 2012), suggests involvement of hydrophobic residues in the PAS such as, Tyr124, Trp286, Phe295, Phe297, Tyr337, Phe338, Tyr341. Docking studies of natural compound SN00335138 suggests binding to the PAS and interactions with Tyr124, Trp286, Phe295, Phe297, Tyr337 and Phe338 (Table 4.11, Figures 4.6A & 4.6B). These interactions are consistent with the interactions observed in the studies mentioned above suggesting that SN00335138 could be a potential AChE inhibitor.

**Table 4.11:** Interactions between the docked natural compound similar to benzylpiperidine derivatives and the residues of AChE

Compound Id.	Structure	Interactions	Binding Energy (Kcal/mol)
SN00335138		<ol style="list-style-type: none"> <li>1. Tyr124 <math>\pi</math>-alkyl with =CH<sub>2</sub></li> <li>2. Trp286 <math>\pi</math>-<math>\pi</math>-stacking with aromatic rings</li> <li>3. Phe295 H-bond with NH</li> <li>4. Phe297 <math>\pi</math>-alkyl with =CH<sub>2</sub></li> <li>5. Tyr <math>\pi</math>-alkyl with =CH<sub>2</sub></li> <li>6. Tyr337 <math>\pi</math>-alkyl with =CH<sub>2</sub></li> <li>7. Phe338 <math>\pi</math>-alkyl with =CH<sub>2</sub></li> <li>8. H<sub>2</sub>O872 H-bond with =O</li> <li>9. H<sub>2</sub>O1281 water-<math>\pi</math> donor with aromatic rings</li> </ol>	-8.5



**Figure 4.6:** (A) SN00335138 docked to the active site of AChE. SN00335138 is displayed in dark blue whereas the AChE residues interacting with it are displayed in light blue. (B) Detailed view of the AChE residues interacting with SN00335138

#### 4.3.2.3 Docking results for TS-3 and TS-4

HIV-1 reverse transcriptase (RT) consists of two subunits, namely, p51 and p66 with molecular mass of 51kDa and 66kDa respectively. Both the subunits are synthesized from same protein Gag-Pol due to differential cleavage by protease (Esnouf *et al.*, 1995; Sarafianos *et al.*, 2009). The p51 subunit plays a structural role

whereas the active sites for both the activities of HIV-1 RT are present on the p66 subunit. Polymerase domain is further is divided into four sub-domains, namely, fingers (residues 1-85 and 118-155), palm (residues 86-117 and 156-236), thumb (residues 237-318) and connection (residues 319-426). p51 subunit polymerase domain also folds into the same sub domains but have different relative positions as compared to the p66 subunit. Non-nucleoside inhibitors (NNIs) bind near the polymerase active site of p66 subunit. Residues Leu100, Lys101, Lys103, Val106, Thr107, Val108, Val179, Tyr181, Ty188, Trp229, Leu234, Tyr318 from p66 and Glu138 from p51 together make the Non-nucleoside inhibitor binding pocket (NNIBP) (Smerdon *et al.*, 1994; Esnouf *et al.*, 1995; Sarafianos *et al.*, 2009). A binding site similar to that of NNIBP is absent in p51 subunit even though it has all the corresponding residues on p66 subunit. Binding of NNIs to the NNIBP causes conformational changes in the polymerase active site resulting in the inhibition of the protein activity. These changes include the distortion in the primer binding position causing change in the orientation of the primer terminus affecting the DNA synthesis. The conformations of Asp110, Asp185 and Asp186, the catalytic carboxylates which bind to the metal co-factors in the polymerase active site are also distorted (Esnouf *et al.*, 1995). Distortion of catalytic carboxylates restricts the movement of  $\beta 9$ - $\beta 10$  loop necessary for the translocation of nucleic acids during the process of polymerization (Sarafianos *et al.*, 2009).

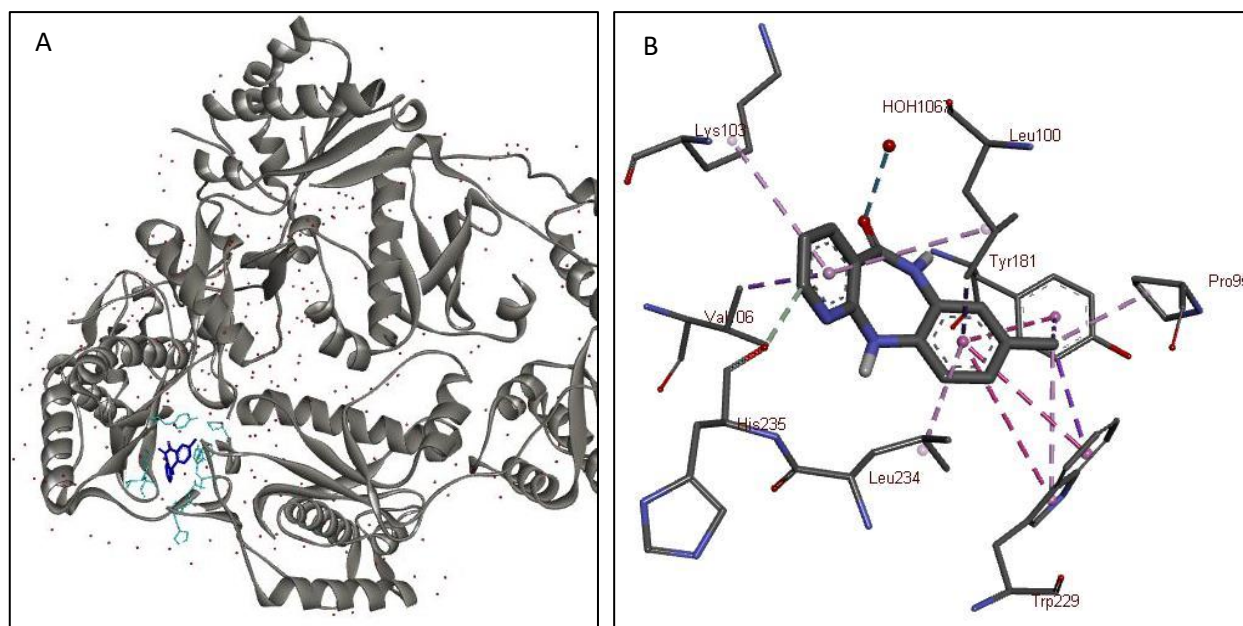
Three natural compounds were found having scaffold similar to 2-substituted dipyridodiazepinones (Table 2.1, AID: 198247). Of these three natural compounds one compound SN00118406 was predicted to have medium to high pIC<sub>50</sub> against HIV-1 RT (Table 4.9). Docking studies were hence performed with this compound on HIV-1 RT. The co-crystal structure of HIV-1 RT complexed with navirapine shows its binding to the NNIBP of HIV-1 RT (Smerdon *et al.*, 1994). Compound SN00118406 observed to dock at the NNIBP. Table 4.12 shows the details of interaction between SN00118406 and HIV-1 RT and Figure 4.7 displays its docking pose and detailed interactions. SN00118406 was observed to interact with Leu100, Lys103, Val106, Tyr181, Trp229, Leu234 and His235 comprising the NNIBP validating the high pIC<sub>50</sub> values predicted

by the QSAR model. Similarly, of the twelve natural compounds with scaffold similar to 2-pyridinones (Table 2.1, AID: 197804) six compounds, namely, SN00008635, SN00008637, SN00008647, SN00008860, SN00010264 and SN00063879 were predicted to have high or medium activity with pIC<sub>50</sub> (Table 4.9). Therefore, docking studies of these compounds were performed on HIV-1 RT. These six compounds were also observed to dock at the NNIBP. Table 4.13 shows the residues of HIV-1 RT with which the docked compounds interact and Figures 4.8 and 4.9 display the docking poses of these six compounds and their detailed interactions with the protein, respectively. These compounds are observed to interact with at least one of the residues comprising NNIBP, namely, Lys101, Lys103 and Val179 supporting the high pIC<sub>50</sub> value estimated by the QSAR model.

SN00118406, a natural compound selected using TS-3 scaffold can be observed to interact with 6 of the 13 NNIBP. Whereas, the natural compounds selected using the TS-4 scaffold were observed to interact with at the most 3 of the NNIBP residues. Suggesting a possibly more stable binding of 2-substituted dipyrindiazepinones than the 2-pyridone derivatives.

**Table 4.12:** Interactions between the docked natural compound similar to 2-substituted dipyrindiazepinones and the HIV-1 RT residues

Compound Id.	Structure	Interactions	Binding Energy (Kcal/mol)
SN00118406		<ol style="list-style-type: none"> <li>1. Pro95 Alkyl with CH<sub>3</sub> of A</li> <li>2. Leu100 <math>\pi</math>-<math>\sigma</math> with A, <math>\pi</math>-alkyl with C</li> <li>3. Lys103 <math>\pi</math>-alkyl with C</li> <li>4. Val106 <math>\pi</math>-<math>\sigma</math> with C</li> <li>5. Tyr181 <math>\pi</math>-<math>\sigma</math> with CH<sub>3</sub> of A, <math>\pi</math>-<math>\pi</math> stacking with A</li> <li>6. Trp229 <math>\pi</math>-alkyl and <math>\pi</math>-<math>\sigma</math> with CH<sub>3</sub> of A, <math>\pi</math>-<math>\pi</math> stacking with A</li> <li>7. Leu234 <math>\pi</math>-alkyl with C</li> <li>8. His235 H-bond with C</li> <li>9. HOH1067 H-bond with =O</li> </ol>	-7.1

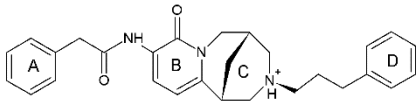
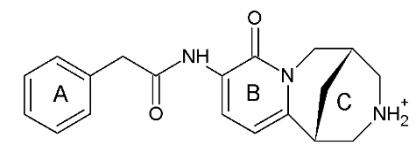
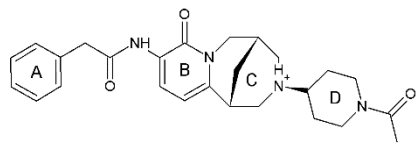
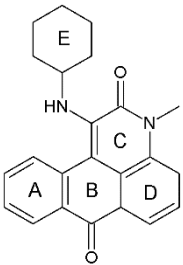


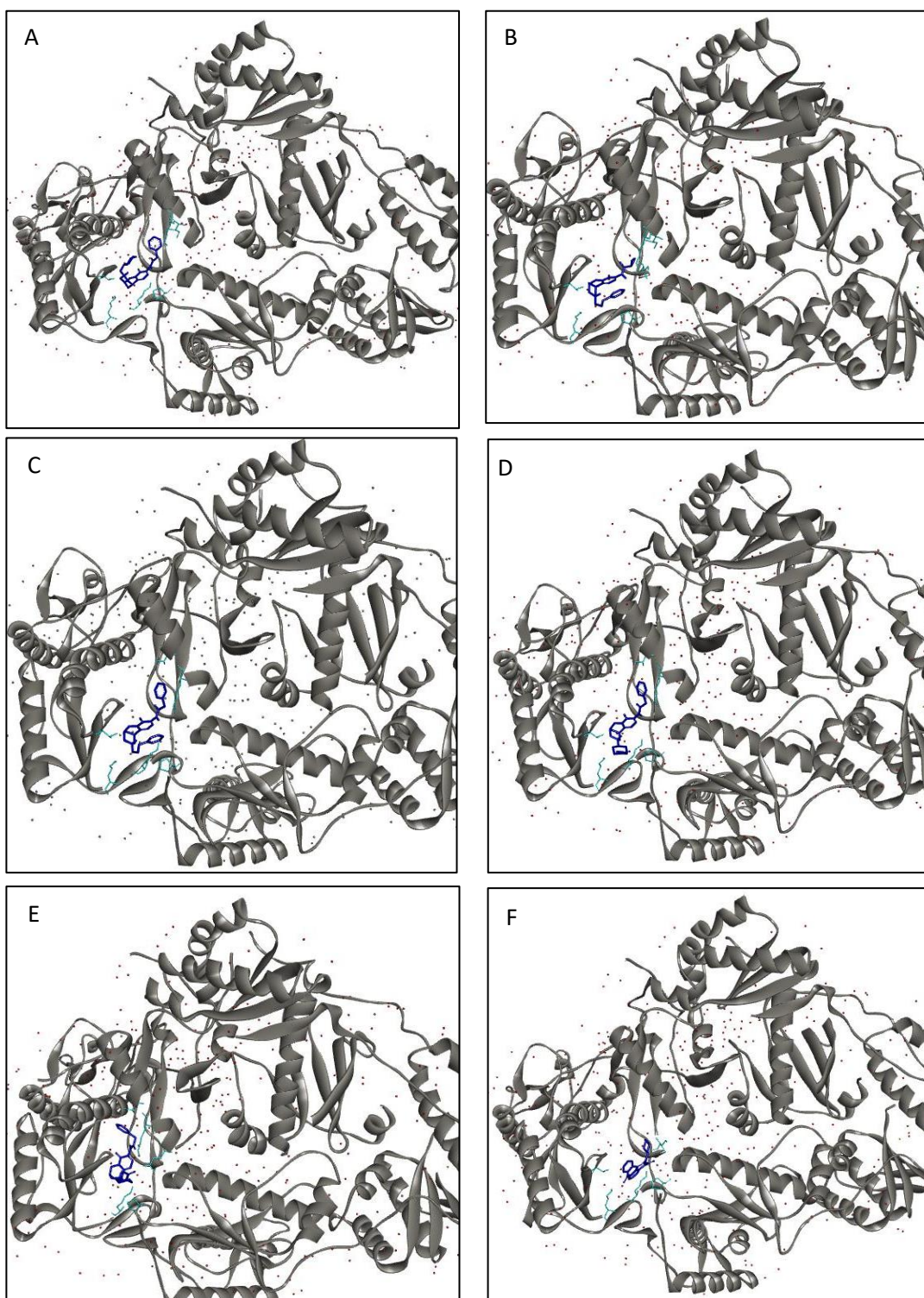
**Figure 4.7:** (A) SN00118406 docked in the NNIBP of HIV-1 RT. SN00118406 is displayed in drack blue whereas the HIV-1 RT residues interacting with it are displayed in light blue. (B) Detailed view of the HIV-1 RT residues interacting with SN00118406

**Table 4.13:** Interactions between the docked natural compounds similar to 2-pyridinones and the protein residues

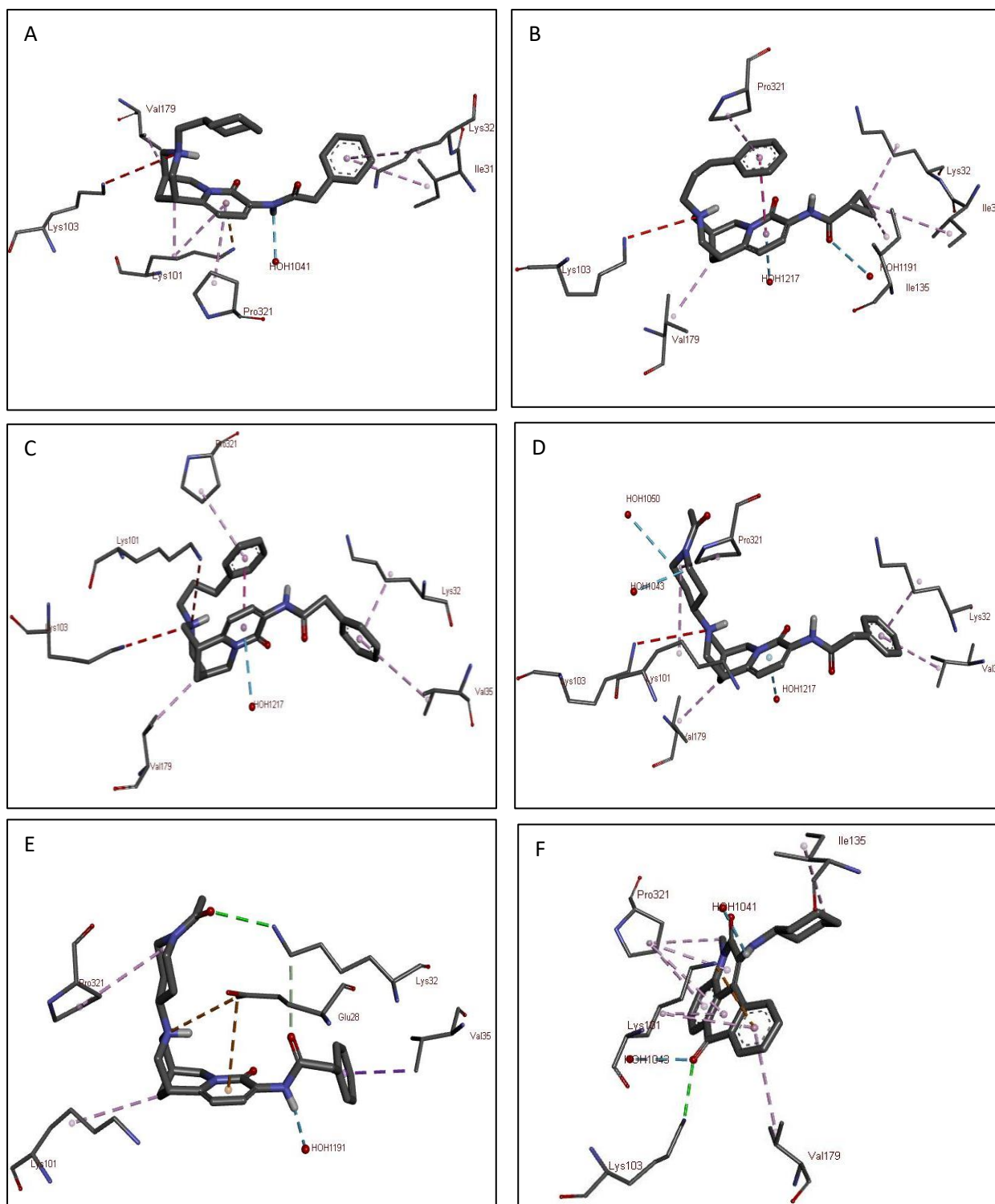
Compound Id.	Structure	Interactions	Binding Energy (Kcal/mol)
SN00008635		<ol style="list-style-type: none"> <li>1. Ile31 <math>\pi</math>-alkyl with A</li> <li>2. Lys32 <math>\pi</math>-alkyl with A</li> <li>3. Lys101 <math>\pi</math>-alkyl and <math>\pi</math>-cation with B, allyl-alkyl with C</li> <li>4. Lys103 Positive-positive with NH<sup>+</sup> of C</li> <li>5. Val179 allyl-alkyl with C</li> <li>6. HOH1041 H-bond with NH</li> </ol>	-6.3
SN00008637		<ol style="list-style-type: none"> <li>1. Ile31 <math>\pi</math>-alkyl with A</li> <li>2. Lys32 <math>\pi</math>-alkyl with A</li> <li>3. Lys103 Positive-positive with NH<sup>+</sup> of C</li> <li>4. Ile 135 <math>\pi</math>-alkyl with A</li> <li>5. Val179 allyl-alkyl with C</li> <li>6. Pro321 <math>\pi</math>-alkyl with D</li> <li>7. HOH1191 H-bond with =O near A</li> <li>8. HOH1217 <math>\pi</math>-Donor interaction with B</li> </ol>	-5.1



SN00008647		<ol style="list-style-type: none"> <li>1. Lys32 <math>\pi</math>-alkyl with A</li> <li>2. Val35 <math>\pi</math>-alkyl with A</li> <li>3. Lys101 Positive-positive with NH<sup>+</sup> of C</li> <li>4. Lys103 Positive-positive with NH<sup>+</sup> of C</li> <li>5. Val179 allyl-alkyl with C</li> <li>6. Pro321 <math>\pi</math>-alkyl with D</li> <li>7. HOH1217 <math>\pi</math>-Donor interaction with B</li> </ol>	-6.1
SN00008860		<ol style="list-style-type: none"> <li>1. Lys32 <math>\pi</math>-alkyl with A</li> <li>2. Val35 <math>\pi</math>-alkyl with A</li> <li>3. Lys101 allyl-alkyl with D</li> <li>4. Lys103 Positive-positive with NH<sup>+</sup> of C</li> <li>5. Val179 allyl-alkyl with C</li> <li>6. Pro321 <math>\pi</math>-alkyl with D</li> <li>7. HOH1043 H-bond with CH<sub>2</sub> of D</li> <li>8. HOH1050 H-bond with CH<sub>2</sub> of D</li> <li>9. HOH1217 <math>\pi</math>-Donor interaction with B</li> </ol>	-6.2
SN00010264		<ol style="list-style-type: none"> <li>1. Glu28 <math>\pi</math>-anion with B, charge-charge interaction with NH<sup>+</sup></li> <li>2. Lys32 H-bond with =O near A, H-bond with =O near D</li> <li>3. Val35 <math>\pi</math>-<math>\sigma</math> with A</li> <li>4. Lys101 allyl-alkyl with C</li> <li>5. Pro321 allyl-alkyl with D</li> <li>6. HOH1191 H-bond with NH near B</li> </ol>	-6.2
SN00063879		<ol style="list-style-type: none"> <li>1. Lys101 <math>\pi</math>-alkyl with A and B, <math>\pi</math>-cation with A</li> <li>2. Lys103 H-bond with =O of B</li> <li>3. Ile 135 allyl-alkyl with E</li> <li>4. Val179 <math>\pi</math>-alkyl with A</li> <li>5. Pro321 <math>\pi</math>-alkyl with B, C and D</li> <li>6. HOH1041 H-bond with NH near E</li> <li>7. HOH1043 H-bond with =O of B</li> </ol>	-6.8



**Figure 4.8:** Natural compounds similar to 2-pyridinones docked in the NNIBP of HIV-1 RT A) SN00008635, B) SN00008637, C) SN00008647, D) SN00008860, E) SN00010264 and F) SN00063879. Natural compounds are displayed in dark blue color whereas the Wee1 residues interacting with the compound are shown in light blue



**Figure 4.9:** Detailed view of NNIBP residues interacting with the docked natural compounds similar to 2-pyridinones A) SN00008635, B) SN00008637, C) SN00008647, D) SN00008860, E) SN00010264 and F) SN00063879

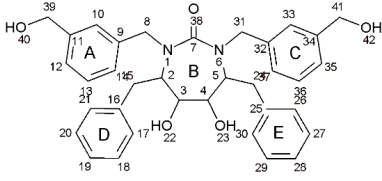
#### 4.3.2.4 Docking results for TS-5

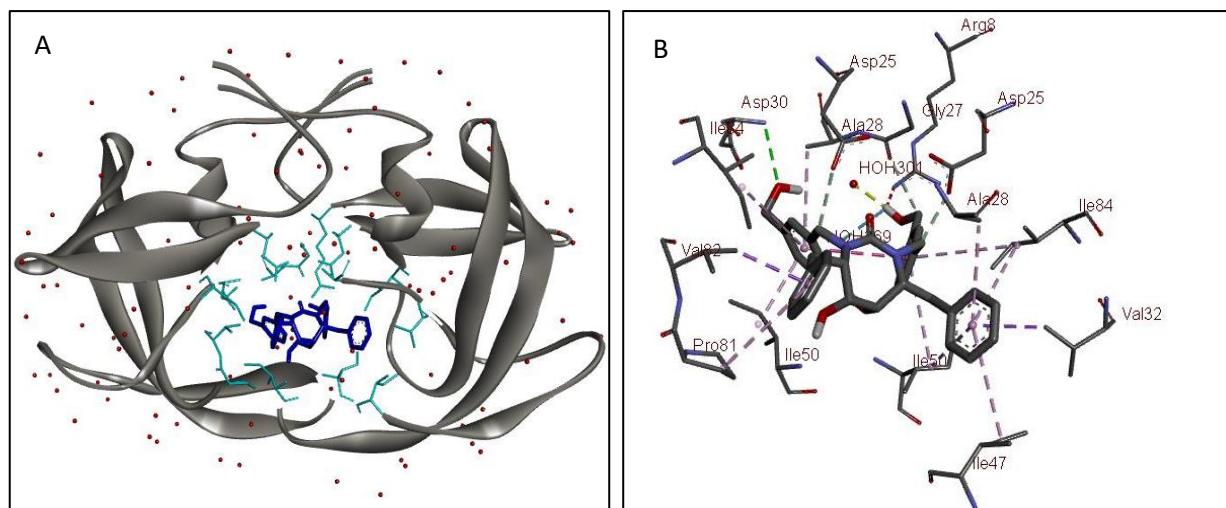
HIV-1 protease (HIV-1 PR), is a virus specific aspartyl protease that recognizes Phe-Pro and Tyr-Pro as the cleavage site for the substrate protein. Active form of HIV-1 PR is a homodimer of two identical 99 amino acid subunits that are inactive as monomers. The catalytic active site is present at the dimer interface with each subunit contributing catalytic tripeptide sequence Asp<sup>25,25'</sup>-Thr<sup>26,26'</sup>-Gly<sup>27,27'</sup>. HIV-1 PR active site is described as an open ended cylinder with a diameter of 10Å having hydrophobic amino acids except catalytic Asp<sup>25,25'</sup> (Saleh *et al.*, 2017). These aspartic acid residues catalyse the hydrolysis of sessile peptide bond of the substrate protein. Thr<sup>26,26'</sup> are proposed to stabilize the active site conformation and Gly<sup>27,27'</sup> to bind the substrate protein in position for hydrolysis by Asp<sup>25,25'</sup> (Mager, 2001). Residues 44-57 and 44'-57' from both the subunits form flap region of antiparallel β-strands. Flap regions fold over the active site and regulate the entry of the substrate into the active site (Bäckbro *et al.*, 1997; Saleh *et al.*, 2017). Cyclic urea inhibitors are known to bind to the active site and interact with Ile<sup>23,23'</sup>, Asp<sup>25,25'</sup>, Ala<sup>28,28'</sup>, Asp<sup>30,30'</sup>, Val<sup>32,32'</sup>, Ile<sup>47,47'</sup>, Ile<sup>50,50'</sup>, Pro<sup>81,81'</sup> and Ile<sup>84,84'</sup> (Bäckbro *et al.*, 1997).

Of the three natural compounds with scaffold similar to the cyclic urea derivatives (Table 2.1, AID: 160292) one compound, SN00215212, was predicted to have a high pIC<sub>50</sub> value (Table 4.9). pIC<sub>50</sub> value for SN00021523, however, was predicted to be beyond the range of pIC<sub>50</sub> values of the compounds used to build the model (5-11). Hence, SN00215212 was further taken up for docking studies. Docking of SN00215212 was carried out using crystal structure 1AJX from PDB (Bäckbro *et al.*, 1997). SN00215212 was found to dock in the active site region of HIV-1 PR as shown in Figure 4.10A. The interactions between the natural compounds and the amino acid residues of the protein are shown in Figure 4.10B while Table 4.14 lists in detail the nature of these interactions. Among the residues interacting with the docked SN00215212 were Asp<sup>25,25'</sup>, Gly<sup>27,27'</sup>, Ala<sup>28,28'</sup>, Asp<sup>30,30'</sup>, Val<sup>32,32'</sup>, Ile<sup>47,47'</sup>, Ile<sup>50,50'</sup>, Pro<sup>81,81'</sup>, and Ile<sup>84,84'</sup>. These residues, as discussed above, are known to

interact with the cyclic urea inhibitors of HIV-1 PR. Thus, these observations support the high pIC<sub>50</sub> values predicted for these compounds.

**Table 4.14:** Interactions between the docked natural compounds similar to cyclic urea derivatives and the HIV-1 PR residues

Compound Id.	Structure	Interactions	Binding Energy (Kcal/mol)
SN00215212		<ol style="list-style-type: none"> <li>Arg8 donor-donor interaction with OH (O40)</li> <li>Asp25 H-bond with C31</li> <li>Asp25` H-bond with C8</li> <li>Gly27` H-bond with C39</li> <li>Ala28 <math>\pi</math>-alkyl with D</li> <li>Ala28` <math>\pi</math>-alkyl with C and H-bond with C39</li> <li>Asp30` H-bond with OH (O42)</li> <li>Val32 <math>\pi</math>-<math>\sigma</math> with D</li> <li>Ile47 <math>\pi</math>-alkyl with D</li> <li>Ile50 <math>\pi</math>-alkyl with C</li> <li>Ile50` <math>\pi</math>-alkyl with A and D</li> <li>Pro81` <math>\pi</math>-alkyl with E</li> <li>Val82` <math>\pi</math>-<math>\sigma</math> with E</li> <li>Ile84 <math>\pi</math>-alkyl with A and D</li> <li>Ile84` <math>\pi</math>-alkyl with C</li> <li>HOH301 H-bond with OH (O40)</li> <li>HOH369 H-bond with OH (O40)</li> </ol>	-11.2



**Figure 4.10:** A) Pose of compound SN00215212 docked into the active site of HIV-1 PR. Natural compounds are displayed in dark blue color whereas the HIV-1 PR residues interacting with the compound are shown in light blue. (B) Detailed view of interactions between HIV-1 PR residues and SN00215212.

#### 4.4 Conclusions

VC-PLS methodology was applied to all the six target systems using the PFMDs developed in this study. VC-PLS models showed good prediction statistics for all the six target systems indicating the applicability of the method. The performance of VC-PLS QSAR models was found to be comparable to that of the PMF-PLS QSAR models. The time required for the VC-PLS model was observed to be comparable to that for the PMF-PLS QSAR model. The five best models were then used for screening the natural compounds obtained from the SuperNatural II database. The predictions of  $pIC_{50}$  values obtained for the natural compounds using the VC-PLS QSAR models were found to be consistent with the predictions obtained using the PMF-PLS QSAR models confirming the usefulness of both the methods. These predictions were complemented by docking studies that showed effective binding of the new inhibitor molecules to the target proteins. Thus confirming the potential of both PMF-PLS QSAR models and VC-PLS QSAR models for screening of new drug molecules.

## Conclusions and future scope:

In Chapter 2, we show that 2D descriptors having 3D structural information of the molecules can be generated to form image-based descriptors. In general, these descriptors can be created by first projecting the atomic positions on a 2D plane where information about the inter-atomic distances is preserved before assigning atomic property values to the image pixels representing the atomic positions. The 2D-QSAR models developed using these descriptors, however, have the drawback of high computational times required for feature selection and model optimization. These challenges could be addressed by using more efficient feature selection algorithms to build the models. Performance of 2D-QSAR models using the above descriptors displayed satisfactory prediction capability for four of the six target systems studied, namely, anti-cancer Wee1 inhibitors, benzylpiperidine AChE inhibitors, 2-substituted dipyrindodiazepinone HIV-1 RT inhibitors and anti-malarial azalide derivatives. The models for the other two target systems, namely, 2-pyridinone HIV-1 RT inhibitors and cyclic urea inhibitors of HIV-1 PR, were able to capture the over-all trend of the  $pIC_{50}$  values in making predictions. This indicates that these image-based 2D-QSAR models although useful require further refinement for potential applications in the real world problems. The steps used for creating the image-based descriptors may be generalized by using different projection methods and atomic properties to study a variety of images-based descriptors. Thus, the above work suggests a platform for developing and studying other novel image-based descriptors for 2D-QSAR modelling.

In Chapter 3, we defined the concept of pseudo-molecular field (PMF) for a molecule and calculated its values at the grid points in a 3D box defined around the molecules. The calculation of PMF was shown to be analogous to that of the electrostatic field values around these molecules. The PMF values are dependent on the product of the electron affinity and the electronegativity of the atoms whereas, those of the electrostatic field are dependent in the partial atomic charges of the atoms.

Since the intrinsic atomic properties are constant the PMF calculations can be performed quicker as compared to the electrostatic field calculations where the partial atomic charges of atoms need to be calculated separately for every molecule. PMF-PLS methodology was developed for building 3D-QSAR models using the pseudo-field molecular descriptors. The PMF-PLS QSAR models showed good prediction statistics for all the six target systems indicating the utility of the approach. Comparison of PMF-PLS QSAR model performance with that of the 3D-QSAR model based on other methods from literature for the same data set also showed comparable performance of the PMF-PLS QSAR model. The PMF-PLS QSAR model development was also found to be computationally light taking less time to arrive at the optimal model showing better performance of these models when compared to the image-based 2D-QSAR models studied in Chapter 2. It is possible that these computational times may be further improved by implementing an algorithm that obtains a proper choice of reference (training and test) sets faster.

A varying PLS component model was developed in Chapter 4 for regressing the PFMDs with the  $pIC_{50}$  values of the compounds. The performance of VC-PLS QSAR models was again comparable to that of the PMF-PLS QSAR models for all the six target systems, indicating the potential of VC-PLS methodology. The computational time required for VC-PLS models was also comparable to that for the PMF-PLS QSAR modelling strategy. The consistencies in prediction, as seen by evaluating the regression statistics of predictions indicate the usefulness of both PMF-PLS QSAR models and VC-PLS QSAR models.

PMF-PLS QSAR models and VC-PLS QSAR models were used to predict the  $pIC_{50}$  values of the natural compounds obtained from SuperNatural II database with unknown activity. Both the models were observed to predict similar  $pIC_{50}$  values for all the natural compounds screened for the six target systems.

Docking studies of the natural compounds were carried out to complement the QSAR studies. Results presented show that molecules with high predicted  $pIC_{50}$  values exhibited good binding to the respective target proteins. The docking results



confirm the potential of PMF-PLS and VC-PLS QSAR models for the virtual screening of new drug molecules for testing purposes.

The *in-silico* methodologies investigated in this work need to be further studied by experimentally testing for the activities of the natural compounds for verifying the predictions made by the PMF-PLS and VC-PLS QSAR models.

The accuracy and reliability of the QSAR models in the current work may be further improved if larger datasets are available for training the models. Additionally, the QSAR methodologies in this work use compounds with single scaffold in a model. Models that consider multiple scaffolds simultaneously need to be built for studies with the same target protein. This is required because the orientation of compounds will be enhanced in 3D by aligning all their common scaffolds. This would not only allow the use of compounds with different scaffolds to be studied in the same QSAR model but also help in increasing the size of the training set.

An automation of various steps involved in the processing of structures and modelling may be further done to improve upon the time required for developing and analyzing these QASR models. Further, the PFMDs developed and studied in the present work could be used in the state-of-the-art 4D- and 5D-QSAR formalisms (Lill, 2007; Vedani and Dobler, 2002). In 4D models 3D fields of different conformations of molecules are considered while in 5D various flexible docking scenarios are additionally taken into account. The QSAR strategies proposed here can be adapted suitably for such studies.

## References:

- Acharya,C. *et al.* (2011) Recent Advances in Ligand-Based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore Approach. *Curr. Comput. Aided-Drug Des.*, **7**, 10–22.
- Ala,P.J. *et al.* (1998) Molecular Recognition of Cyclic Urea HIV-1 Protease Inhibitors. *J. Biol. Chem.*, **273**, 12325–12331.
- Alvin C. Rencher (2002) *Methods of Multivariate Analysis* 2nd ed. A JOHN WILEY & SONS, INC. PUBLICATION.
- Andrade,J.M. *et al.* (2004) Procrustes rotation in analytical chemistry, a tutorial. *Chemom. Intell. Lab. Syst.*, **72**, 123–132.
- Bacilieri,M. and Moro,S. (2006) Ligand-Based Drug Design Methodologies in Drug Discovery Process: An Overview. *Curr. Drug Discov. Technol.*, **3**, 155–165.
- Bäckbro,K. *et al.* (1997) Unexpected binding mode of a cyclic sulfamide HIV-1 protease inhibitor. *J. Med. Chem.*, **40**, 898–902.
- Bajusz,D. *et al.* (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.*, **7**, 20.
- Banerjee,P. *et al.* (2015) Super Natural II-a database of natural products. *Nucleic Acids Res.*, **43**, D935–D939.
- Berman,H.M. *et al.* (2002) The protein data bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **58**, 899–907.
- Bochevarov,A.D. *et al.* (2013) Jaguar: A high-performance quantum chemistry software program with strengths in life and materials sciences. *Int. J. Quantum Chem.*, **113**, 2110–2142.
- Borg,I. and Groenen,P. (2005) *Modern Multidimensional Scaling: Theory and Applications* (Springer Series in Statistics) Second. Springer Science+Business Media, Inc., New Yor.
- Bronstein,M.M. *et al.* (2006) Multigrid multidimensional scaling. *Numer. Linear Algebr. with Appl.*, **13**, 149–171.
- Brownlee,J. (2018) *How to Remove Outliers for Machine Learning*.
- Cherkasov,A. *et al.* (2014) QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.*, **57**, 4977–5010.
- Cheung,J. *et al.* (2013) Structures of Human Acetylcholinesterase Bound to Dihydrotanshinone I and Territrem B Show Peripheral Site Flexibility. *ACS Med. Chem. Lett.*, **4**, 1091–1096.

- Cheung, J. *et al.* (2012) Structures of Human Acetylcholinesterase in Complex with Pharmacologically Important Ligands. *J. Med. Chem.*, **55**, 10282–10286.
- Cormanich, R. a *et al.* (2009) Improvement of multivariate image analysis applied to quantitative structure-activity relationship (QSAR) analysis by using wavelet-principal component analysis ranking variable selection and least-squares support vector machine regression: QSAR study of . *Chem. Biol. Drug Des.*, **73**, 244–52.
- Cramer, R.D. *et al.* (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, **110**, 5959–67.
- Cumming, J.G. *et al.* (2013) Chemical predictive modelling to improve compound quality. *Nat. Rev. Drug Discov.*, **12**, 948–62.
- Dahl, E.L. and Rosenthal, P.J. (2008) Apicoplast translation, transcription and genome replication: targets for antimalarial antibiotics. *Trends Parasitol.*, **24**, 279–284.
- Daré, J.K. *et al.* (2018) 3D perspective into MIA-QSAR: A case for anti-HCV agents. *Chem. Biol. Drug Des.*, 0–2.
- Dasoondi, A.S. *et al.* (2008) Comparative molecular field analysis of benzothiazepine derivatives: mitochondrial sodium calcium exchange inhibitors as antidiabetic agents. *Indian J. Pharm. Sci.*, **70**, 186–92.
- Debnath, A.K. (1999) Three-Dimensional Quantitative Structure–Activity Relationship Study on Cyclic Urea Derivatives as HIV-1 Protease Inhibitors: Application of Comparative Molecular Field Analysis †. *J. Med. Chem.*, **42**, 249–259.
- Dijkstra, E.W. (1959) A Note on Two Probles in Connexion with Graphs. *Numer. Math.*, **1**, 269–271.
- Divakar, S. and Hariharan, S. (2015) 3D-QSAR studies on Plasmodium falciparam proteins: a mini-review. *Comb. Chem. High Throughput Screen.*, **18**, 188–198.
- Dong, J. *et al.* (2015) ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *J. Cheminform.*, **7**, 60.
- Dundas, J. *et al.* (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.*, **34**, W116–W118.
- Elmi, Z. *et al.* (2009) Feature selection method based on fuzzy entropy for regression in QSAR studies. *Mol. Phys.*, **107**, 1787–1798.
- Esnouf, R. *et al.* (1995) Mechanism of inhibition if HIV-1 reverse transcriptase by non-nucleoside inhibitors. *Struct. Biol.*, **2**, 303–308.
- Estrada, E. (1995) Three-Dimensional Molecular Descriptors Based on Electron

- Charge Density Weighted Graphs. *J. Chem. Inf. Comput. Sci.*, **35**, 708–713.
- Free, S.M. and Wilson, J.W. (1964) A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.*, **7**, 395–399.
- Freitas, M.P. *et al.* (2005) MIA-QSAR: a simple 2D image-based approach for quantitative structure–activity relationship analysis. *J. Mol. Struct.*, **738**, 149–154.
- Freitas, M.P. and Rittner, R. (2008) MIA-QSAR as an Alternative Approach for Modeling Some Antifungals. *QSAR Comb. Sci.*, **27**, 582–585.
- Garg, R. *et al.* (1999) Comparative Quantitative Structure – Activity Relationship Studies on Anti-HIV Drugs.
- Garthwaite, P.H. (1994) An Interpretation of Partial Least Squares. *J. Am. Stat. Association*, **89**, 122–127.
- Gasteiger, J. and Marsili, M. (1980) Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, **36**, 3219–3228.
- Geenen, J.J.J. and Schellens, J.H.M. (2017) Molecular pathways: Targeting the protein kinase Wee1 in cancer. *Clin. Cancer Res.*, **23**, 4540–4544.
- Geladi, P. and Kowalski, B.R. (1986) Partial least-squares regression: a tutorial. *Anal. Chim. Acta*, **185**, 1–17.
- Gramatica, P. (2020) Principles of QSAR Modeling: Comments and Suggestions From Personal Experience. *Int. J. Quant. Struct. Relationships*, **5**, 61–97.
- Gramatica, P. (2007) Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.*, **26**, 694–701.
- Gray, R.H. *et al.* (2001) Randomized trial of presumptive sexually transmitted disease therapy during pregnancy in Rakai, Uganda. *Am. J. Obstet. Gynecol.*, **185**, 1209–1217.
- Guha, R. (2013) In Silico Models for Drug Discovery. **993**, 81–94.
- Hansch, C. and Fujita, T. (1964)  $\rho$ - $\sigma$ - $\pi$  Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.*, **86**, 1616–1626.
- Hodge, V. and Austin, J. (2004) A Survey of Outlier Detection Methodologies. *Artif. Intell. Rev.*, **22**, 85–126.
- Hu, R. *et al.* (2009) Receptor- and ligand-based 3D-QSAR study for a series of non-nucleoside HIV-1 reverse transcriptase inhibitors. *Bioorganic Med. Chem.*, **17**, 2400–2409.
- Hutinec, A. *et al.* (2011) An automated, polymer-assisted strategy for the preparation of urea and thiourea derivatives of 15-membered azalides as potential antimalarial chemotherapeutics. *Bioorg. Med. Chem.*, **19**, 1692–1701.

- Ilyas,I.F. and Chu,X. (2019) Outlier detection. In, *Data Cleaning*. Association for Computing Machinery, New York, NY, pp. 11–46.
- de Jong,S. (1993) SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.*, **18**, 251–263.
- Ke,A.B. *et al.* (2014) Pharmacometrics in Pregnancy: An Unmet Need. *Annu. Rev. Pharmacol. Toxicol.*, **54**, 53–69.
- Kendall,D.G. (1989) [A Survey of the Statistical Theory of Shape]: Rejoinder. *Stat. Sci.*, **4**, 116–120.
- Khan,M.T.H. (2010) Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. *Curr. Drug Metab.*, **11**, 285–95.
- Kitchen,D.B. *et al.* (2004) DOCKING AND SCORING IN VIRTUAL SCREENING FOR DRUG DISCOVERY : METHODS AND APPLICATIONS. **3**.
- Kolossov,E. and Stanforth,R. (2010) The quality of QSAR models: problems and solutions. *SAR QSAR Environ. Res.*, **18**, 89–100.
- Kryger,G. *et al.* (1999) Structure of acetylcholinesterase complexed with E2020 (Aricept<sup>o</sup>): Implications for the design of new anti-Alzheimer drugs. *Structure*, **7**, 297–307.
- Kubinyi,H. (2002) From narcosis to hyperspace: The history of QSAR. *Quant. Struct. Relationships*, **21**, 348–356.
- Kubinyi,H. (1997a) QSAR and 3D QSAR in drug design. Part 1: Methodology. *Drug Discov. Today*, **2**, 457–467.
- Kubinyi,H. (1997b) QSAR and 3D QSAR in drug design. Part 2 : applications and problems. *Science (80-. )*, **2**, 538–546.
- Kuriakose,J. *et al.* (2004) Isometric graphing and multidimensional scaling for reaction-diffusion modeling on regular and fractal surfaces with spatiotemporal pattern recognition. *J. Chem. Phys.*, **120**, 5432–43.
- Labute,P. (2000) ScienceDirect - Journal of Molecular Graphics and Modelling : A widely applicable set of descriptors. *J. Mol. Graph. Model.*, **3263**, 464–477.
- Laurie,A.T.R. and Jackson,R.M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, **21**, 1908–1916.
- Li,S. *et al.* (2017) New molecular insights into the tyrosyl-tRNA synthase inhibitors: CoMFA, CoMSIA analyses and molecular docking studies. *Sci. Rep.*, **7**, 11525.
- Lill,M.A. (2007) Multi-dimensional QSAR in drug discovery. *Drug Discov. Today*, **12**, 1013–1017.

- Mager, P.P. (2001) The active site of HIV-1 protease. *Med. Res. Rev.*, **21**, 348–353.
- Mandal, S. *et al.* (2009) Rational drug design. *Eur. J. Pharmacol.*, **625**, 90–100.
- Marco-Contelles, J. *et al.* (2014) Multipotent cholinesterase/monoamine oxidase inhibitors for the treatment of Alzheimer's disease: design, synthesis, biochemical evaluation, ADMET, molecular modeling, and QSAR analysis of novel donepezil-pyridyl hybrids. *Drug Des. Devel. Ther.*, **8**, 1893–1910.
- Matheson, C.J. *et al.* (2016) Targeting WEE1 Kinase in Cancer. *Trends Pharmacol. Sci.*, **37**, 872–881.
- MATLAB (2010) MATLAB:2010 R2010b ed. The MathWorks Inc., Natick, Massachusetts.
- Matta, C.F. and Arabi, A.A. (2011) Electron-density descriptors as predictors in quantitative structure–activity/property relationships and drug design. *Future Med. Chem.*, **3**, 969–94.
- Mueller, S. and Has-Kogan, D.A. (2015) Wee 1 Kinase as a target for cancer therapy. *J. Clin. Oncol.*, **33**, 3845–3847.
- Mulliken, R.S. (1934) A new electroaffinity scale; Together with data on valence states and on valence ionization potentials and electron affinities. *J. Chem. Phys.*, **2**, 782–793.
- Nidhi and Siddiqi, M.I. (2013) Recent Advances in QSAR-Based Identification and Design of Anti-Tubercular Agents. *Curr Pharm Des*, **20**, 4418–4426.
- Noedl, H. *et al.* (2009) Artemisinin-Resistant Malaria in Asia. *N. Engl. J. Med.*, **361**, 540–541.
- Ojha, P.K. and Roy, K. (2018) PLS regression-based chemometric modeling of odorant properties of diverse chemical constituents of black tea and coffee. *RSC Adv.*, **8**, 2293–2304.
- Osakwe, O. (2016) The Significance of Discovery Screening and Structure Optimization Studies. In, *Social Aspects of Drug Discovery, Development and Commercialization*. Elsevier, pp. 109–128.
- Palmer, B.D. *et al.* (2006) 4-Phenylpyrrolo[3,4-c]carbazole-1,3(2H,6H)-dione inhibitors of the checkpoint kinase Wee1. structure-activity relationships for chromophore modification and phenyl ring substitution. *J. Med. Chem.*, **49**, 4896–4911.
- Parker, L.L. and Piwnicka-Worms, H. (1992) Inactivation of the p34cdc2-Cyclin B Complex by the Human WEE1 Tyrosine Kinase. *Science (80-. )*, **257**, 1955–1957.
- Perić, M. *et al.* (2012) Antimalarial activity of 9a-N substituted 15-membered azalides with improved in vitro and in vivo activity over azithromycin. *J. Med. Chem.*, **55**, 1389–1401.

- Polishchuk,P. (2017) Interpretation of Quantitative Structure-Activity Relationship Models: Past, Present, and Future. *J. Chem. Inf. Model.*, **57**, 2618–2639.
- Proudfoot,J.R. *et al.* (1995) Novel Non-nucleoside Inhibitors of Human Immunodeficiency Virus Type 1 (HIV-1) Reverse Transcriptase. 4. 2-Substituted Dipyrrodo-diazepinones as Potent Inhibitors of Both Wild-Type and Cysteine-181 HIV-1 Reverse Transcriptase Enzymes. *J. Med. Chem.*, **1**, 4830–4838.
- Queiroz,J. *et al.* (2011) Receptor-dependent ( RD ) 3D-QSAR approach of a series of benzylpiperidine inhibitors of human acetylcholinesterase ( HuAChE ). *Eur. J. Med. Chem.*, **46**, 39–51.
- Reddy,M.R. and Parrill,A.L. (1999) Overview of Rational Drug Design. In, *ACS Symposium Series*. American Chemical Society, pp. 1–11.
- Rosenthal,P.J. (2016) Azithromycin for Malaria? *Am. J. Trop. Med. Hyg.*, **95**, 2–4.
- Roy,K. *et al.* (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemom. Intell. Lab. Syst.*, **152**, 18–33.
- Roy,K. and Das,R.N. (2014) A review on principles, theory and practices of 2D-QSAR. *Curr. Drug Metab.*, **15**, 346–79.
- Saleh,N.A. *et al.* (2017) Design and Development of Some Viral Protease Inhibitors by QSAR and Molecular Modeling Studies. In, Gupta,S. (ed), *Viral Proteases and Their Inhibitors*. Academic Press, pp. 25–58.
- Sarafianos,S.G. *et al.* (2009) Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition. *J. Mol. Biol.*, **385**, 693–713.
- Schrödinger, LLC,N.Y. Jaguar 7.8.
- Schrödinger, LLC,N.Y. (2011) LigPrep 2.5.
- Schwede,T. *et al.* (2003) SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.*, **31**, 3381–3385.
- Sidhu,A.B.S. *et al.* (2007) In vitro efficacy, resistance selection, and structural modeling studies implicate the malarial parasite apicoplast as the target of azithromycin. *J. Biol. Chem.*, **282**, 2494–2504.
- Silva,F.T. and Trossini,G.H.G. (2014) The survey of the use of QSAR methods to determine intestinal absorption and oral bioavailability during drug design. *Med. Chem.*, **10**, 441–8.
- Smaill,J.B. *et al.* (2008) Synthesis and structure-activity relationships of N-6 substituted analogues of 9-hydroxy-4-phenylpyrrolo[3,4-c]carbazole-1,3(2H,6H)-diones as inhibitors of Wee1 and Chk1 checkpoint kinases. *Eur. J. Med. Chem.*, **43**, 1276–1296.

- Smerdon, S.J. *et al.* (1994) Structure of the binding site for nonnucleoside inhibitors of the reverse transcriptase of human immunodeficiency virus type 1. *Proc. Natl. Acad. Sci.*, **91**, 3911–3915.
- Smola, A.J. and Schölkopf, B. (2004) A tutorial on support vector regression. *Stat. Comput.*, **14**, 199–222.
- Squire, C.J. *et al.* (2005) Structure and Inhibition of the Human Cell Cycle Checkpoint Kinase, Wee1A Kinase. *Structure*, **13**, 541–550.
- Stanton, D.T. and Jurs, P.C. (1990) Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure-Property Relationship Studies. *Anal. Chem.*, **62**, 2323–2329.
- Steve R. Gunn (2010) Support Vector Machines for classification and regression. *Univ. Southampt. Support*, **135**, 230–267.
- Sushko, I. *et al.* (2011) Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided. Mol. Des.*, **25**, 533–554.
- Todeschini, R. and Consonni, V. (2008) Handbook of Molecular Descriptors Mannhold, R. *et al.* (eds) Wiley.
- Todorov, N.P. *et al.* (2007) De Novo Design. In, *Comprehensive Medicinal Chemistry II*. Elsevier, pp. 283–305.
- Trott, O. and Olson, A.J. (2010) AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, **31**, 455–461.
- Tubeleviciute-Aydin, A. *et al.* (2019) Identification of Allosteric Inhibitors against Active Caspase-6. *Sci. Rep.*, **9**, 1–19.
- Vapnik, V.N. (1999) An overview of statistical learning theory. *IEEE Trans. Neural Networks*, **10**, 988–999.
- Vedani, A. and Dobler, M. (2002) 5D-QSAR: The key for simulating induced fit? *J. Med. Chem.*, **45**, 2139–2149.
- Vella, A. *et al.* (2019) Targeting hepatic glucokinase to treat diabetes with TTP399, a hepatoselective glucokinase activator. *Sci. Transl. Med.*, **11**, eaau3441.
- Verma, J. *et al.* (2010) 3D-QSAR in Drug Design - A Review. *Curr. Top. Med. Chem.*, **10**, 95–115.
- Wai, J.S. *et al.* (1993) Synthesis and Evaluation of 2-Pyridinone Derivatives as Specific HIV-1 Reverse Transcriptase Inhibitors. 3. Pyridyl and Phenyl Analogs of 3-Aminopyridin-2(1H)-one. *J. Med. Chem.*, **36**, 249–255.



- Wang,T. *et al.* (2015) Quantitative structure–activity relationship: promising advances in drug discovery platforms. *Expert Opin. Drug Discov.*, **10**, 1283–1300.
- Wang,Y. *et al.* (2014) PubChem BioAssay: 2014 update. *Nucleic Acids Res.*, **42**, 1075–1082.
- Wells,T.N.C. *et al.* (2015) Malaria medicines: a glass half full? *Nat. Rev. Drug Discov.*, **14**, 424–442.
- Westholm,D.E. *et al.* (2009) Competitive inhibition of organic anion transporting polypeptide 1c1-mediated thyroxine transport by the fenamate class of nonsteroidal antiinflammatory drugs. *Endocrinology*, **150**, 1025–1032.
- Wold,S. (1978) Cross-validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics*, **20**, 397–405.
- Wolfram Alpha LLC WolframAlpha.
- Yang,S.-Y. (2010) Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov. Today*, **15**, 444–450.
- Yi,P. *et al.* (2008) 3D-QSAR studies of Checkpoint Kinase Weel inhibitors based on molecular docking, CoMFA and CoMSIA. *Eur. J. Med. Chem.*, **43**, 925–938.
- Yoo,C.K. and Shahlaei,M. (2018) The applications of PCA in QSAR studies: A case study on CCR5 antagonists. *Chem. Biol. Drug Des.*, **91**, 137–152.
- Zhang,M. *et al.* (2017) WEE1 inhibition by MK1775 as a single-agent therapy inhibits ovarian cancer viability. *Oncol. Lett.*, **14**, 3580–3586.

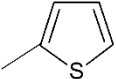
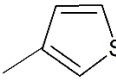
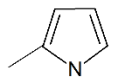
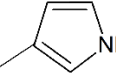
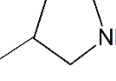
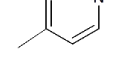
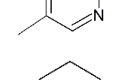
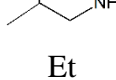
## Appendix

**Table A1:** Structures of 4-phenyl pyrrolocarbazole derivatives for TS-1 (AID: 268838)

Compound	Structure	X	Y	Z	Experimental pIC <sub>50</sub>
1		9-OH	NH	Ph	7.01
2		9-OH	NH	H	5.40
3		9-OH	NH	I	5.64
4		8-OH	NH	Ph	6.51
5		9-OH	O	Ph	6.37
6		9-OH	S	Ph	7.11
7		9-OH	NMe	Ph	6.59
8		Me		Ph	6.89
9		Et		Ph	5.80
10		Ph		Me	5.01
11		Ph		Ph	5.64
12		Ph		H	5.40
13		OMe			4.70
14		H			4.43

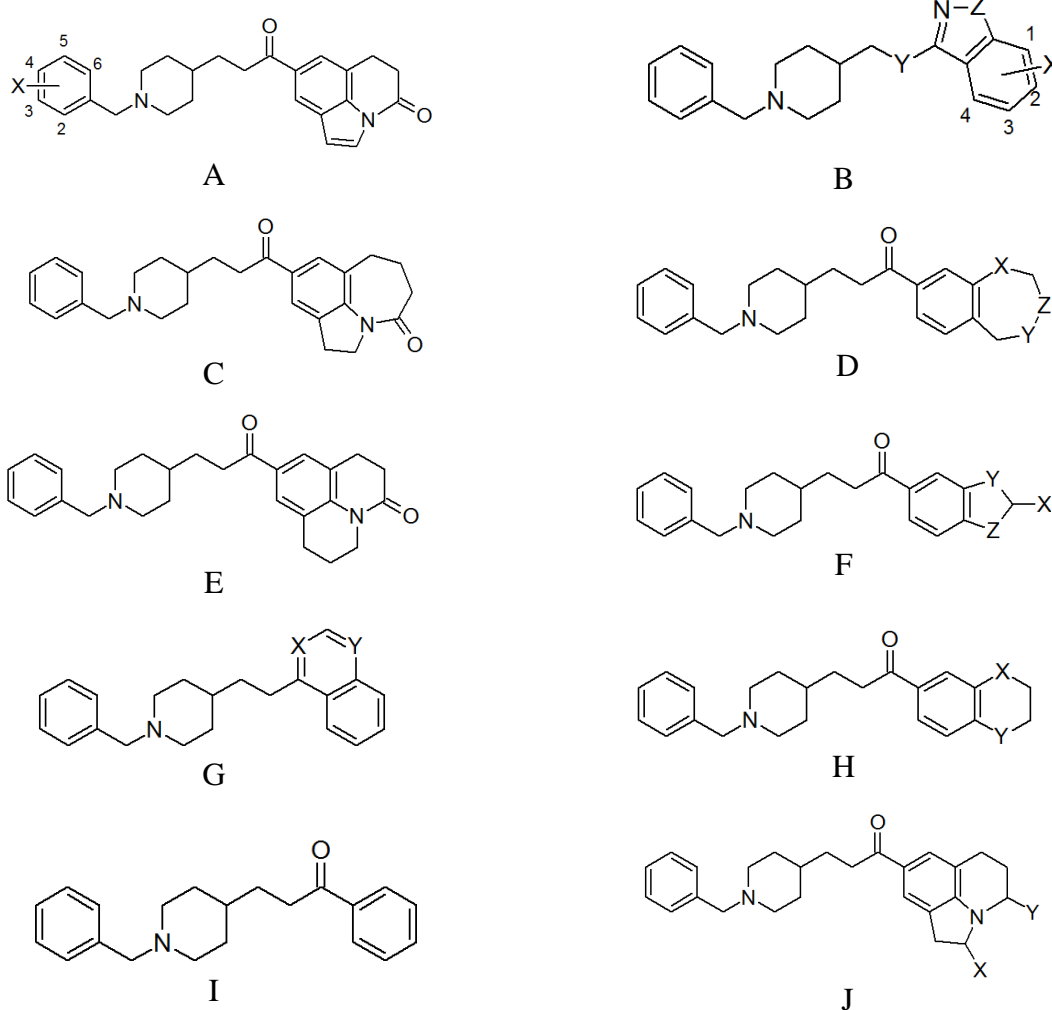
Compound	Structure	X	Y	Z	Experimental pIC <sub>50</sub>
15	D	-OH			5.55
16	E	N-NH <sub>2</sub>	NH		5.41
17	A	9-OH	NH	2'-ClPh	7.96
18	A	9-OMe	NH	2'-ClPh	6.19
19	A	9-OH	NMe	2'-ClPh	7.24
20	A	9-OH	O	2'-ClPh	7.48
21	F			2'-F	6.48
22	F			2'-Br	7.64
23	F			2'-I	7.89
24	F			2'-Me	6.82
25	F			2'-Et	6.29
26	F			2'-CF <sub>3</sub>	6.24
27	F			2'-CH <sub>2</sub> OH	6.35
28	F			2'-CN	6.72
29	F			2'-COMe	6.08
30	F			2'-CONH <sub>2</sub>	6.80
31	F			2'-Ph	6.24
32	F			2'-OH	7.22
33	F			2'-OMe	7.62
34	F			2'-OEt	6.59
35	F			2'-SMe	7.48
36	F			2'-SOMe	6.66
37	F			2'-NO <sub>2</sub>	7.33
38	F			2'-NH <sub>2</sub>	6.68
39	F			3'-F	6.66
40	F			3'-Cl	7.26
41	F			3'-Me	6.64
42	F			3'-CH <sub>2</sub> OH	6.06
43	F			3'-CH <sub>2</sub> NH <sub>2</sub>	5.36
44	F			3'-CN	6.75

Compound	Structure	X	Y	Z	Experimental pIC <sub>50</sub>
45	F			3'-COMe	5.37
46	F			3'-Ph	4.40
47	F			3'-OH	7.05
48	F			3'-OMe	6.21
49	F			3'-NO <sub>2</sub>	6.52
50	F			3'-NH <sub>2</sub>	7.16
51	F			4'-F	4.80
52	F			4'-Cl	6.14
53	F			4'-Me	5.48
54	F			4'-CH <sub>2</sub> OH	5.92
55	F			4'-CN	5.75
56	F			4'-COMe	5.44
57	F			4'-OH	7.17
58	F			4'-OMe	4.92
59	F			4'-SMe	4.54
60	F			4'-SO <sub>2</sub> Me	5.96
61	F			4'-NH <sub>2</sub>	6.82
62	F			2'-Cl, 3'-Cl	7.55
63	F			2'-Cl, 3'-OH	7.92
64	F			2'-Cl, 3'-NH <sub>2</sub>	7.68
65	F			2'-Cl, 5'-OH	7.64
66	F			2'-Cl, 4'-NH <sub>2</sub>	7.62
67	F			2'-Cl, 5'-Cl	6.31
68	F			2'-Cl, 5'-OH	7.38
69	F			2'-Cl, 5'-NH <sub>2</sub>	7.70
70	F			2'-Cl, 6'-Cl	7.55
71	F			2'-Cl, 6'-OH	7.35
72	F			2'-Cl, 6'-OMe	7.82
73	F			2'-Br, 4'-NH <sub>2</sub>	7.70
74	F			2'-Br, 6'-Br	7.46

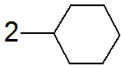
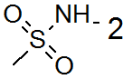
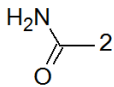
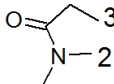
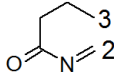
Compound	Structure	X	Y	Z	Experimental pIC <sub>50</sub>
75	F			2'-Me, 3'-Me	6.66
76	F			2'-Me, 5'-Me	6.02
77	F			2'-Me, 6'-Me	7.13
78	F			2'-OMe, 4'-NH <sub>2</sub>	7.72
79	F			2'-OMe, 6'-OMe,	6.56
80	F			2'-OMe, 6'-F	7.57
81	F			2'-OMe, 4'-NH <sub>2</sub>	7.54
82	F			2',6',-diCl, 3'- OH	7.75
83	F			2',6',-diCl, 4'- OH	7.31
84	G				6.85
85	G				7.38
86	G				6.75
87	G				7.42
88	G				5.89
89	G				6.09
90	G				6.24
91	G				5.00
92	H	Et			7.30
93	H	<i>n</i> -Pr			7.20
94	H	<i>i</i> -Pr			7.28
95	H	<i>n</i> -Bu			7.23

Compound	Structure	X	Y	Z	Experimental pIC <sub>50</sub>
96	H	(CH <sub>2</sub> ) <sub>2</sub> <i>i</i> -Pr			6.82
97	H	<i>n</i> -pent			6.77

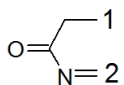
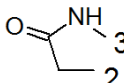
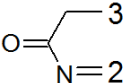
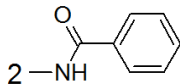
**Table A2:** Structures of benzylpiperidine derivatives for TS-2 (AID: 566585)



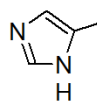
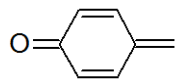
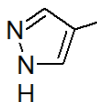
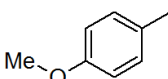
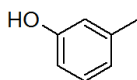
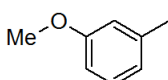
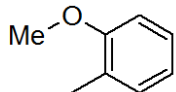
Compound	Structure	X	Y	Z	Experimental pIC <sub>50</sub>
1	A	3-OH			8.06
2	A	2-F			8.60
3	A	3-OH			8.34
4	A	2-OH			8.96
5	A	3-NO <sub>2</sub>			8.54
6	A	2-OCH <sub>3</sub>			7.19
7	A	4-Cl			6.82
8	E				7.80
9	A	3-Cl			8.31
10	A	2-Cl			8.29
11	A	2-NO <sub>2</sub>			7.05

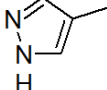
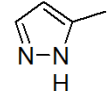
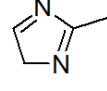
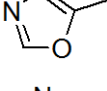
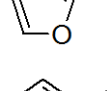
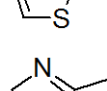
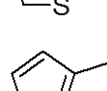
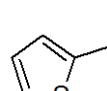
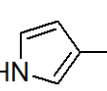
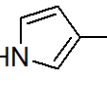
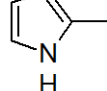
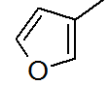
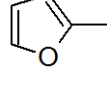
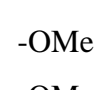
Compound	Structure	X	Y	Z	Experimental pIC50
12	A	3-F			8.89
13	A	4-OCH <sub>3</sub>			6.46
14	B		CH <sub>2</sub>	O	9.10
15	C				7.28
16	F	Ph	NH	NH	7.48
17	B		CH <sub>2</sub>	O	7.85
18	A	4-NO <sub>2</sub>			7.37
19	A	4-OH			9.31
20	A	3-OCH <sub>3</sub>			6.90
21	D	CH <sub>2</sub>	CH <sub>2</sub>	NH	7.60
22	D	O	NH	CH <sub>2</sub>	7.40
23	B		CH <sub>2</sub>	O	8.06
24	B	H	NH	O	6.09
25	B	H	O	O	5.59
26	B	3-OCH <sub>3</sub>	CH <sub>2</sub>	O	8.14
27	G	N	N		6.47
28	B		CH <sub>2</sub>	O	9.32
29	D	CH <sub>2</sub>			7.64
30	F	Me	NH	N	7.92
31	B	1-OCH <sub>3</sub>	CH <sub>2</sub>	O	8.15
32	B	H	-(CH <sub>2</sub> ) <sub>2</sub> -	O	6.05
33	F	Me	H	NC <sub>2</sub> H <sub>5</sub>	8.37
34	B	2-NHAc	CH <sub>2</sub>	O	8.55
35	B	2-Br	CH <sub>2</sub>	O	7.30
36	H	NH	CH <sub>2</sub>		6.78
37	F	H	CH <sub>2</sub>	NH	7.28
38	B	H	CH <sub>2</sub>	S	7.00
39	B		CH <sub>2</sub>	O	9.24

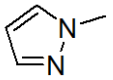
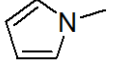
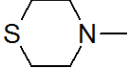
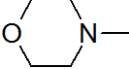
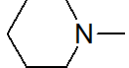
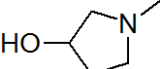
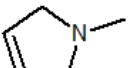
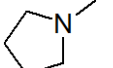
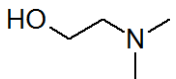

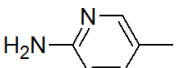
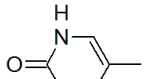
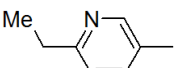


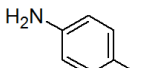


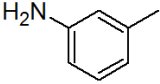

Compound	Structure	X	Y	Z	Experimental pIC50
40	I				6.52
41	B		CH <sub>2</sub>	O	8.44
42	B	2 -CH <sub>3</sub> , 3-CH <sub>3</sub>	CH <sub>2</sub>	O	8.24
43	G	N	CH		6.66
44	B	H	NH	O	6.49
45	B	H	CH <sub>2</sub>	NH	6.92
46	B		CH <sub>2</sub>	O	9.02
47	H	O	O		7.52
48	H	CH <sub>2</sub>	NH		7.19
49	B	3 Me	CH <sub>2</sub>	O	8.11
50	B	2 =O	CH <sub>2</sub>	O	7.59
51	B	2 -NH <sub>2</sub>	CH <sub>2</sub>	O	7.70
52	B	H	CH	O	6.68
53	J	H	O		8.01
54	J	O	H		8.44
55	B	H	CH <sub>2</sub>	O	7.26
56	B	2 -OMe	CH <sub>2</sub>	O	8.08
57	B		CH <sub>2</sub>	O	9.48
58	B		CH <sub>2</sub>	O	8.03
59	D	NH	CH <sub>2</sub>	CH <sub>2</sub>	6.71
60	F	Me	S	N	8.17

**Table A3:** Structures of 2-substituted Dipyrrodoiazepinone derivatives for TS-3 (AID: 198247)

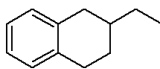
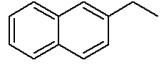
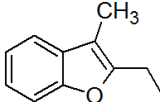
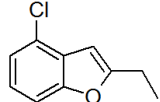
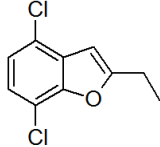
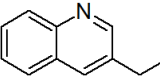
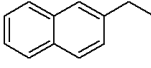
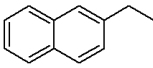
Compound	Structure	X	Experimental $IC_{50}$
1	A		5.43
2	A	-CHCHCONH <sub>2</sub>	6.60
3	A	-CHCHCOOH	6.74
4	D		7.15
5	B		7.22
6	A	-NHCHCHCH <sub>3</sub>	6.41
7	A		5.85
8	A		7.00
9	A		6.82
10	A		6.09
11	A	Ph	6.64

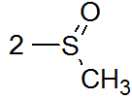
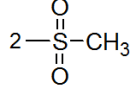
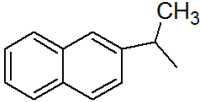
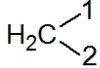
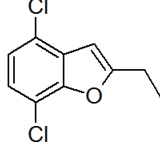
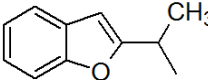
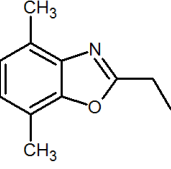
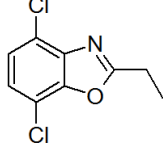
Compound	Structure	X	Experimental $pIC_{50}$
12	A		7.70
13	A		6.41
14	A		6.89
15	A		6.96
16	A		6.66
17	A		6.42
18	A		7.00
19	A		7.00
20	A		6.85
21	B		7.30
22	A		7.52
23	A		7.15
24	A		7.40
25	A		6.96
26	C	-SMe	7.70
27	B	-OMe	6.92
28	A	-OMe	7.40
29	D	=O	6.33

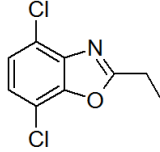
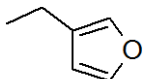
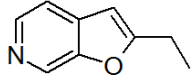
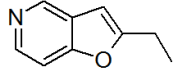
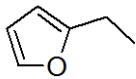
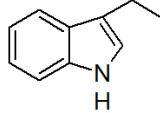
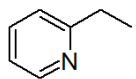
Compound	Structure	X	Experimental pIC <sub>50</sub>
30	A		6.51
31	A		7.05
32	C		6.82
33	A		6.40
34	A		6.52
35	C		7.40
36	A		7.52
37	A		7.70
38	C		8.00
39	A	-N(Me) <sub>2</sub>	7.15
40	A	-CCH	6.85
41	C	Me	7.70
42	B		6.82
43	A		7.30
44	A		5.96
45	A		5.92
46	B		6.74
47	A		7.74
48	A		7.40
49	C	-NH(CH <sub>2</sub> ) <sub>3</sub> OH	7.05

Compound	Structure	X	Experimental pIC <sub>50</sub>
50	C	-NH(CH <sub>2</sub> ) <sub>2</sub> OH	7.05
51	A	-NHEt	6.64
52	A	-NHMe	6.72
53	A	-NH <sub>2</sub>	6.00
54	B	Br	7.52
55	C	Cl	8.00
56	E	Cl	7.70
57	C	F	7.70
58	C	<i>t</i> -Bu	6.00
59	C	<i>i</i> -Pr	6.00
60	C	Et	7.05
61	E	Me	7.15
62	A		7.15
63	B	Cl	7.05
64	C	H	7.40
65	A	Cl	7.10
66	A	Me	6.92
67	A	H	6.89
68	E		7.10

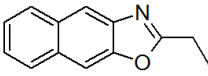
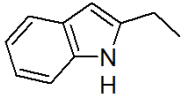
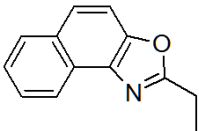
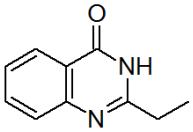
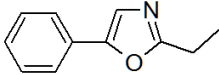
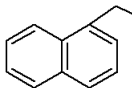
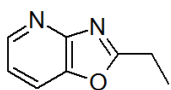
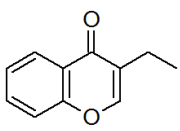
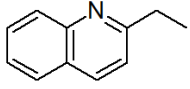
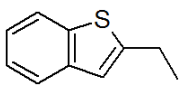
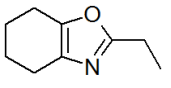
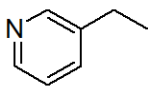
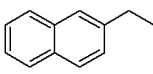
**Table A4:** Structures of 2-pyridinone derivatives for TS-4 (AID: 197804)

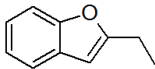
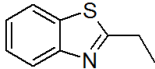
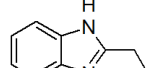
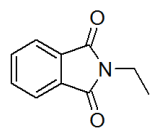
Compound	Structure	X	Y	Z	Experimental pIC50
1	A	O	Et		5.02
2	A	S	Et		6.52
3	B	S	2-Et	1-Cl,4-Cl	7.38
4	A	O	Et		5.71
5	A	O	Et		6.47
6	A	O	Et		7.24
7	A	O	Et		6.47
8	C	O			3.52
9	D			Me	4.96

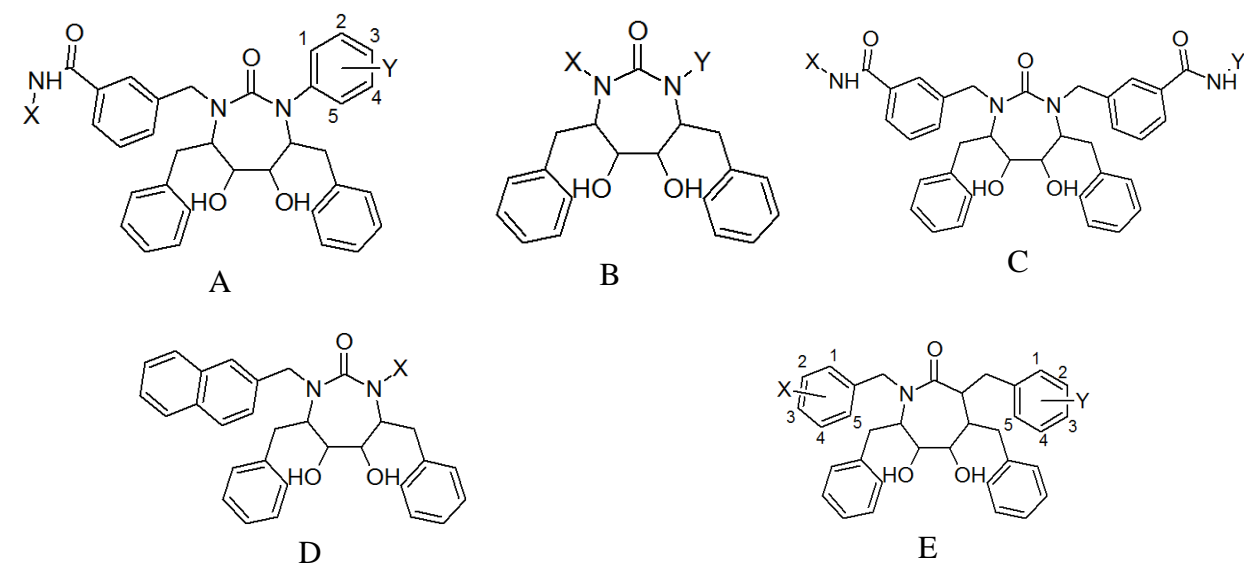
Compound	Structure	X	Y	Z	Experimental pIC50
10	B	O			4.50
11	B	O	2-COOEt		5.76
12	B	O			5.94
13	B	O	2-Et		6.22
14	B	O	2-SEt		6.37
15	B	O	2-SMe		6.72
16	B	O	1-Me	1-Cl,4-Cl	5.55
17	B	O	2-CH(OH)CH <sub>3</sub>	1-Cl,4-Cl	5.98
18	B	O	2-COCH <sub>3</sub>	1-Cl,4-Cl	6.52
19	A	O	-Et		5.95
20	B	O	2-SMe	1-Cl,4-Cl	7.37
21	B	O	2-OMe	1-Cl,4-Cl	6.94
22	B	O		1-Cl,4-Cl	6.95
23	B	O	2-CHCH <sub>2</sub>	1-Cl,4-Cl	7.64
24	D	Me			5.98
25	A	O	Et		6.01
26	D	Me			6.99
27	D	Et			6.18

Compound	Structure	X	Y	Z	Experimental pIC50
28	D	Me			7.24
29	B	O	2-Et	4-NH <sub>2</sub>	4.17
30	B	O	2-Et	4-NO <sub>2</sub>	4.61
31	B	O	2-Et	4-OH	6.36
32	B	O	2-Et	4-OMe	6.74
33	B	O	2-Et	1-F, 4-F	7.15
34	B	O	2-Et	1-Cl, 4-F	6.98
35	B	O	2-Et	1-F	7.04
36	B	O	2-Et	2-F	5.90
37	B	O	2-Et	3-F	6.33
38	B	O	2-Et	4-F	6.96
39	B	O	2-Et	1-Cl	7.19
40	B	O	2-Et	4-Cl	6.82
41	B	O	2-Et	1-Et	6.59
42	B	O	2-Et	2-Me	5.78
43	B	O	2-Et	3-Me	5.90
44	A	O	Et		3.84
45	A	O	Et		3.98
46	A	O	Et		4.49
47	A	O	Et		4.54
48	A	O	Et		4.65
49	A	O	Et		4.82

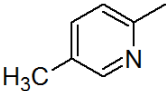
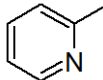
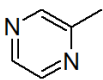
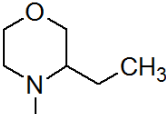
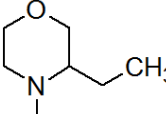
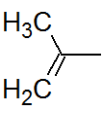
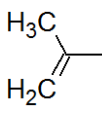
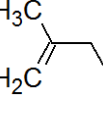
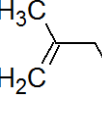
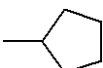
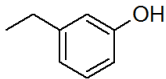
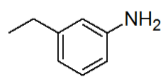
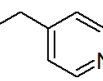
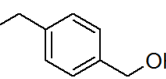


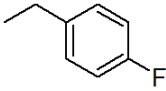
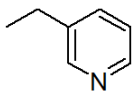
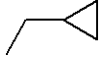
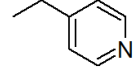
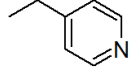
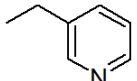
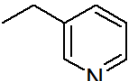
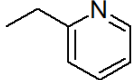
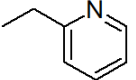
Compound	Structure	X	Y	Z	Experimental pIC50
50	A	O	Et		5.00
51	A	O	Et		5.36
52	A	O	Et		5.57
53	A	O	Et		5.60
54	A	O	Et		5.63
55	A	O	Et		5.68
56	A	O	Et		5.72
57	A	O	Et		5.96
58	A	O	Et		6.28
59	A	O	Et		6.30
60	A	O	Et		6.55
61	A	O	Et	-CH <sub>2</sub> Ph	5.27
62	A	O	Et		3.52
63	B	O	2-Et	4-Me	7.26
64	B	O	2-Et	1-Me	6.92
65	A	O	Et		6.34

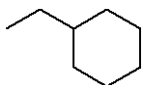
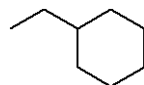
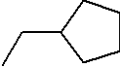
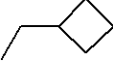
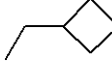
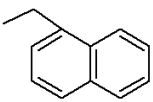
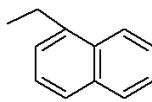


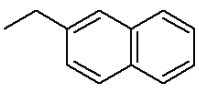
Compound	Structure	X	Y	Z	Experimental pIC50
66	A	O	Et		6.48
67	A	O	Et		6.46
68	A	O	Et		5.12
69	A	O	Et		7.52
70	B	O	2-Et		6.68
71	B	O	2-Et	1-Cl, 4-Cl	7.72
72	B	O	2-Et	1-Me, 4-Me	7.70

**Table A5:** Structures of cyclic urea derivatives for TS-5 (AID: 160292)

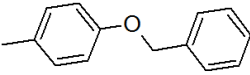
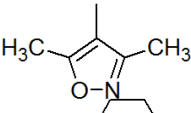
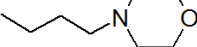
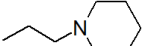
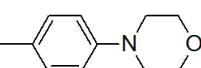

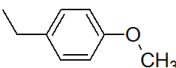
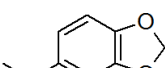
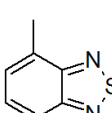
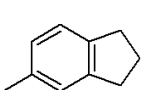
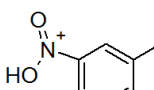
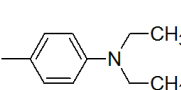

Compound	Structure	X	Y	Experimental pIC <sub>50</sub>
1	A		2-OMe	10.42
2	A		2-OMe	10.16
3	A		2-OMe	10.28
4	A		2-OMe	10.33
5	A		2-NH <sub>2</sub>	10.64
6	A		2-NH <sub>2</sub>	10.92
7	B			10.62
8	A		2-OMe,4-OMe	8.60
9	A		2-OMe,4-OMe	9.07

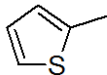
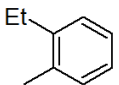
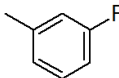
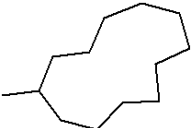

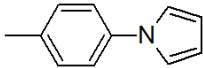
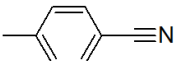
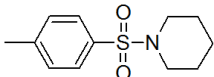
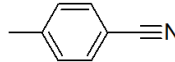
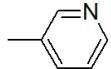
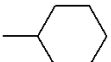
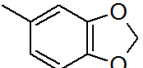
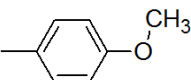
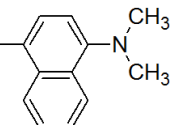
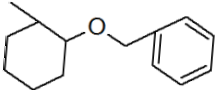
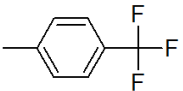
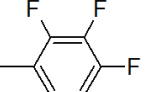
Compound	Structure	X	Y	Experimental pIC <sub>50</sub>
10	A		2-NH <sub>2</sub>	10.12
11	A		2-NO <sub>2</sub>	10.02
12	A	-C(CH <sub>3</sub> ) <sub>3</sub>	2-NH <sub>2</sub>	9.39
13	A		2-NH <sub>2</sub>	10.80
14	B			5.40
15	B			8.74
16	B			8.14
17	B	-(CH <sub>2</sub> ) <sub>2</sub> C(CH <sub>3</sub> ) <sub>3</sub>	-(CH <sub>2</sub> ) <sub>2</sub> C(CH <sub>3</sub> ) <sub>3</sub>	7.44
18	C	NH <sub>2</sub>	NH <sub>2</sub>	10.74
19	C	H	H	10.41
20	C	-CH <sub>2</sub> CN	-CH <sub>2</sub> CN	10.20
21	C	<i>i</i> -Pr	<i>i</i> -Pr	9.24
22	C	Et	Et	9.68
23	C	Me	Me	10.18
24	C	OMe	OMe	10.35
25	C	OH	OH	10.70
26	D			9.55
27	D			9.48
28	D			9.00
29	D			8.16
30	D			9.03

Compound	Structure	X	Y	Experimental pIC <sub>50</sub>
31	D			8.44
32	D			8.28
33	D	-CH <sub>2</sub> Ph		8.64
34	D	-CH <sub>2</sub> CHCH <sub>2</sub>		8.85
35	D			8.82
36	E	2-OH	2-OH	9.92
37	D	-(CH <sub>2</sub> ) <sub>3</sub> CH <sub>3</sub>		9.22
38	E	3-OH	3-OH	9.92
39	E	2-I	2-I	9.38
40	E	2-NO <sub>2</sub>	2-NO <sub>2</sub>	8.55
41	B			7.05
42	B			8.01
43	D	-(CH <sub>2</sub> ) <sub>2</sub> CH <sub>3</sub>		8.96
44	B			6.84
45	B	-CH <sub>2</sub> CCH	-CH <sub>2</sub> CCH	7.66
46	B	-(CH <sub>2</sub> ) <sub>2</sub> OCHCH <sub>2</sub>	-(CH <sub>2</sub> ) <sub>2</sub> OCHCH <sub>2</sub>	7.22
47	E	1-OMe	1-OMe	5.73
48	E	3-CF <sub>3</sub>	3-CF <sub>3</sub>	7.29
49	E	2-CF <sub>3</sub>	2-CF <sub>3</sub>	7.66
50	E	3-Me	3-Me	8.24
51	E	2-Me	-2Me	8.15
52	E	2-Br	2-Br	8.85
53	E	3-Br	3-Br	7.57
54	E	3-Cl	3-Cl	8.28
55	E	2-Cl	2-Cl	9.05
56	E	1-Cl	1-Cl	6.62
57	E	3-F	3-F	8.85

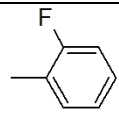
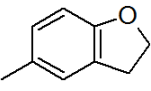
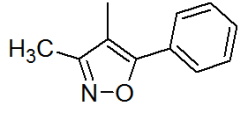
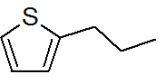
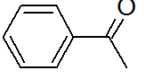
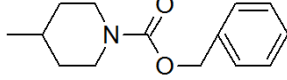
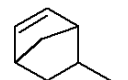
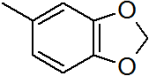
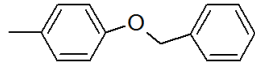
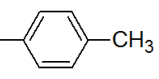
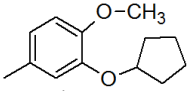
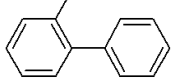
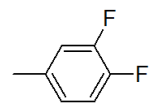
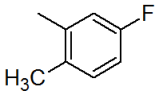
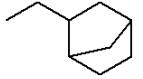
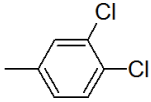
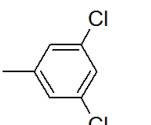
Compound	Structure	X	Y	Experimental pIC <sub>50</sub>
58	E	1-F	1-F	7.47
59	E	2-F	2-F	8.52
60	B			7.43
61	B	-CH <sub>2</sub> Ph		8.37
62	B			8.89
63	B	-(CH <sub>2</sub> ) <sub>4</sub> <i>i</i> -Pr	-(CH <sub>2</sub> ) <sub>4</sub> <i>i</i> -Pr	7.52
64	B	-(CH <sub>2</sub> ) <sub>3</sub> <i>i</i> -Pr	-(CH <sub>2</sub> ) <sub>3</sub> <i>i</i> -Pr	8.15
65	B	-(CH <sub>2</sub> ) <sub>2</sub> <i>i</i> -Pr	-(CH <sub>2</sub> ) <sub>2</sub> <i>i</i> -Pr	7.92
66	B	-CH <sub>2</sub> <i>i</i> -Pr	-CH <sub>2</sub> <i>i</i> -Pr	7.31
67	B	-(CH <sub>2</sub> ) <sub>2</sub> OET	-(CH <sub>2</sub> ) <sub>2</sub> OET	5.96
68	B	-(CH <sub>2</sub> ) <sub>2</sub> OMe	-(CH <sub>2</sub> ) <sub>2</sub> OMe	6.10
69	B	<i>n</i> -hex	<i>n</i> -hex	8.34
70	B	<i>n</i> -pent	<i>n</i> -pent	8.80
71	B	<i>n</i> -Bu	<i>n</i> -Bu	8.85
72	B	<i>n</i> -Pr	<i>n</i> -Pr	8.10
73	B	Et	Et	7.00
74	B	Me	Me	5.24
75	B	Ph	Ph	8.52
76	E	2-CH <sub>2</sub> OH	2-CH <sub>2</sub> OH	9.85
77	E	2-OMe	2-OMe	8.80
78	B			7.07
79	E	3-OMe	3-OMe	6.80
80	E			8.68
81	E	-CH <sub>2</sub> CHCH <sub>2</sub>	-CH <sub>2</sub> CHCH <sub>2</sub>	8.28
82	D			9.51
83	E	2-NH <sub>2</sub>	2-NH <sub>2</sub>	9.55
84	E	3-CH <sub>2</sub> OH	3-CH <sub>2</sub> OH	9.47

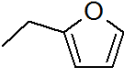
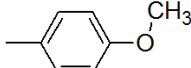
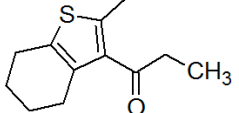
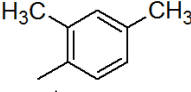
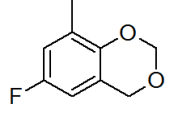
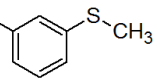
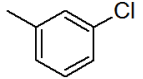
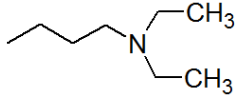
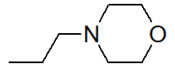
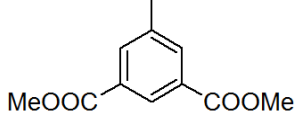
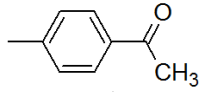
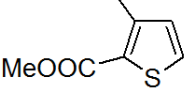
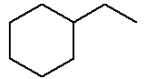
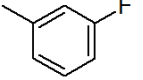
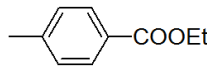
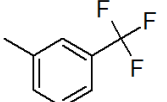
**Table A6:** Structures of anti-malarial azilide derivatives for TS-6 (AID: 579588)

Compounds	Structures	X	Y	Experimental pIC <sub>50</sub>
1	A	S		6.80
2	A	O		4.89
3	A	S		5.29
4	A	S		5.52
5	A	S		5.80
6	A	O		6.21
7	A	O		6.32
8	A	S		6.66
9	A	S		6.99
10	A	O		7.02
11	A	S		7.37
12	A	S		6.53
13	A	S		6.01

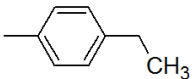
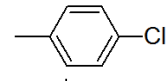
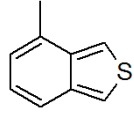
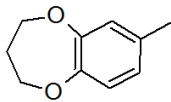
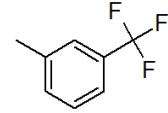
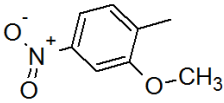
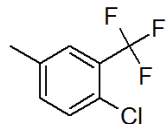
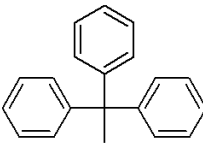
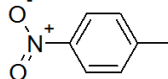
Compounds	Structures	X	Y	Experimental pIC <sub>50</sub>
14	A	O		6.26
15	A	S		6.60
16	A	O		6.79
17	A	S		6.85
18	A	S		6.86
19	A	S		6.99
20	A	S		7.11
21	A	S		7.17
22	A	O		6.69
23	A	S		6.10
24	A	O		6.39
25	A	O		6.58
26	A	S		6.68
27	A	S		6.75
28	A	S		6.90
29	A	O		7.07
30	A	S		7.13
31	A	O	<i>s</i> -Bu	5.93



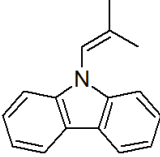
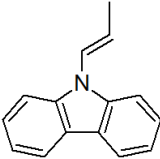
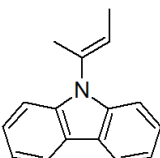
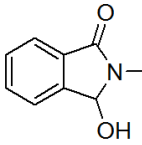
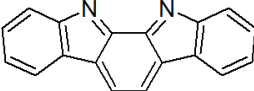
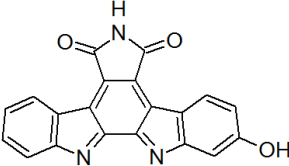
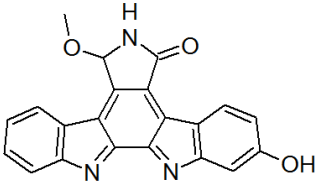
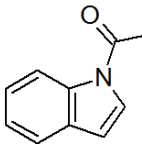
Compounds	Structures	X	Y	Experimental pIC <sub>50</sub>
32	A	S		6.91
33	A	O		6.03
34	A	O		6.31
35	A	O		6.39
36	A	S		6.54
37	A	O		6.63
38	A	S		6.74
39	A	S		6.76
40	A	O		6.84
41	A	O		6.86
42	A	O		6.86
43	A	O		6.86
44	A	O		6.90
45	A	O		6.93
46	A	S		6.99
47	A	S		7.07
48	A	S		7.26
49	A	O	<i>i</i> -Pr	6.10

Compounds	Structures	X	Y	Experimental pIC <sub>50</sub>
50	A	S	-CH <sub>2</sub> CHCH <sub>2</sub>	6.29
51	A	O		6.31
52	A	O		6.36
53	A	S		6.39
54	A	S	<i>n</i> -Bu	6.42
55	A	O		6.63
56	A	O		6.66
57	A	O		6.86
58	A	O		6.93
59	A	S		4.91
60	A	S		5.48
61	A	O		6.39
62	A	O		6.40
63	A	S		6.48
64	A	O		6.85
65	A	S		6.94
66	A	S		6.97
67	A	O		7.14

Compounds	Structures	X	Y	Experimental pIC <sub>50</sub>
68	A	S		7.26
69	A	S		7.02
70	A	S	<i>i</i> -Bu	6.36
71	A	S	<i>i</i> -Pr	6.37
72	A	S		6.62
73	A	S		6.73
74	A	S		6.78
75	A	S		6.88
76	A	S		6.97
77	A	S		7.11
78	A	S		7.26
79	A	S	-CH <sub>2</sub> CH <sub>2</sub> Cl	4.89
80	A	O	Et	5.70
81	A	O	<i>n</i> -Bu	6.09
82	A	O		6.24
83	A	S	<i>n</i> -Bu	6.45
84	A	O		6.49
85	A	O		6.74
86	A	S		6.76
87	A	S		6.82
88	A	S		6.92

Compounds	Structures	X	Y	Experimental pIC <sub>50</sub>
89	A	O		7.00
90	A	O		7.04
91	A	O		7.19
92	A	O		7.22
93	A	O		7.27
94	A	S		7.16
95	A	S		7.25
96	A	S		7.50
97	A	S		7.29
98	B			5.69

**Table A7:** Structures of the natural compounds and their IDs obtained from SuperNatural II database for TS-1 to TS-6

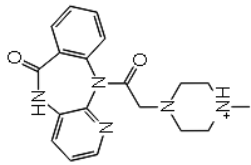
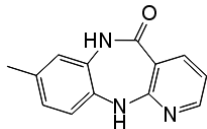
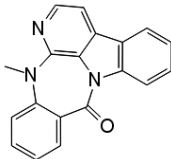
TS-1: Molecules similar to 4-phenyl pyrrolocazazole scaffold		
Sr. no.	Compound ID	Structure
1	SN00011632	
2	SN00054717	
3	SN00058100	
4	SN00118263	
5	SN00226661	
6	SN00272309	
7	SN00289913	
8	SN00335731	

Sr. no.	Compound ID	Structure
9	SN00343696	
10	SN00345401	
11	SN00362452	
12	SN00362911	

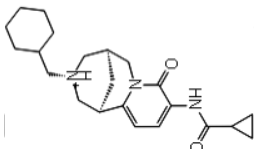
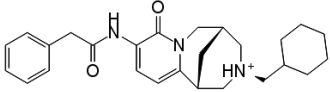
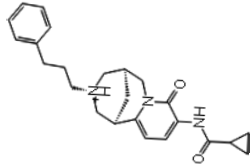
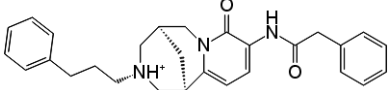
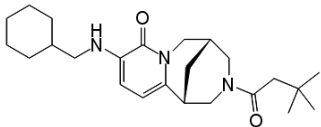
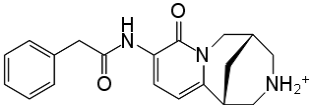
TS-2: Molecules similar to benzylpiperidine derivatives

Sr. no.	Compound ID	Structure
1	SN00160095	
2	SN00304033	
3	SN00335138	

TS-3: Molecules similar to 2-substituted dipyridodiazeponone derivatives

Sr. no.	Compound ID	Structure
1	SN00024429	
2	SN00118406	
3	SN00387398	

TS-4: Molecules similar to 2-Pyridinone Derivatives

Sr. no.	Compound ID	Structure
1	SN00008627	
2	SN00008635	
3	SN00008637	
4	SN00008647	
5	SN00008665	
6	SN00008860	

Sr. no.	Compound ID	Structure
7	SN00009758	
8	SN00010264	
9	SN00011738	
10	SN00026473	
11	SN00063879	

TS-5: Molecules similar to cyclic urea derivatives

Sr. no.	Compound ID	Structure
1	SN00021523	
2	SN00213428	
3	SN00215212	



TS-6: Molecules similar to 15 membered azalide derivatives

Sr. no.	Compound ID	Structure
1	SN00114856	
2	SN00220696	
3	SN00282305	
4	SN00289590	
5	SN00310837	

**Table A8:** Predictions of pIC<sub>50</sub> values of Wee1 inhibitors for TS-1 using image-based QSAR model (AID: 268838)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1	7.01	6.41	26	6.24	6.42
2	5.40	6.00	27	6.35	6.33
3	5.64	5.92	28	6.72	7.24
4	6.51	6.40	29	6.08	6.24
5 <sup>a</sup>	6.37	5.69	30	6.80	6.79
6	7.11	6.71	31 <sup>a</sup>	6.24	6.76
7	6.59	6.51	32	7.22	7.47
8	6.89	6.78	33	7.62	7.54
9	5.80	5.79	34 <sup>a</sup>	6.59	6.76
10	5.01	5.82	35	7.48	7.39
11	5.64	5.82	36	6.66	6.54
12	5.40	5.77	37	7.33	6.99
13	4.70	4.69	38	6.68	6.75
14	4.43	4.95	39	6.66	6.68
15	5.55	5.78	40	7.26	7.11
16	5.41	5.75	41 <sup>a</sup>	6.64	7.01
17	7.96	7.69	42	6.06	6.09
18	6.19	6.18	43	5.36	5.28
19	7.24	7.18	44	6.75	6.78
20	7.48	6.87	45	5.37	5.46
21	6.48	6.96	46	4.40	4.71
22	7.64	7.68	47 <sup>a</sup>	7.05	6.96
23	7.89	7.12	48	6.21	6.08
24	6.82	6.88	49	6.52	6.62
25	6.29	6.29	50	7.16	7.05

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
51 <sup>a</sup>	4.80	5.78	75 <sup>a</sup>	6.66	6.14
52	6.14	6.27	76	6.02	6.70
53	5.48	5.48	77	7.13	7.13
54	5.92	5.98	78 <sup>a</sup>	7.72	7.50
55 <sup>a</sup>	5.75	6.01	79	6.56	5.87
56	5.44	5.46	80	7.57	7.58
57	7.17	5.99	81	7.54	7.50
58	4.92	4.88	82	7.75	7.76
59	4.54	4.54	83	7.31	7.18
60	5.96	5.96	84	6.85	6.84
61 <sup>a</sup>	6.82	6.29	85	7.38	7.20
62	7.55	7.58	86	6.75	6.79
63 <sup>a</sup>	7.92	7.63	87	7.42	6.83
64 <sup>a</sup>	7.68	7.45	88	5.89	6.05
65	7.64	7.07	89 <sup>a</sup>	6.09	5.96
66	7.62	7.72	90	6.24	6.36
67	6.31	7.07	91	5.00	5.01
68 <sup>a</sup>	7.38	7.42	92	7.30	7.31
69	7.70	7.69	93	7.20	7.21
70	7.55	7.44	94	7.28	7.24
71	7.35	7.46	95	7.23	7.25
72	7.82	7.95	96	6.82	6.86
73	7.70	7.55	97	6.77	6.04
74	7.46	7.12			

<sup>a</sup> - Test set compound

**Table A9:** Predictions of pIC<sub>50</sub> values of AChE inhibitors for TS-2 using image-based QSAR model (AID: 566585)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1	8.06	7.88	26	8.14	8.14
2	8.60	8.14	27	6.47	6.43
3	8.34	7.32	28	9.32	9.08
4	8.96	8.90	29	7.64	7.80
5	8.54	8.55	30	7.92	7.83
6	7.19	7.32	31	8.15	8.16
7 <sup>a</sup>	6.82	7.07	32	6.05	6.13
8	7.80	7.77	33	8.37	8.40
9 <sup>a</sup>	8.31	7.64	34 <sup>a</sup>	8.55	8.25
10	8.29	8.21	35	7.30	7.30
11 <sup>a</sup>	7.05	7.03	36	6.78	6.57
12	8.89	8.65	37 <sup>a</sup>	7.28	7.41
13	6.46	7.57	38	7.00	7.32
14	9.10	8.09	39	9.24	8.42
15	7.28	7.28	40	6.52	7.74
16	7.48	7.48	41	8.44	8.45
17 <sup>a</sup>	7.85	7.90	42	8.24	8.24
18	7.37	7.25	43	6.66	6.70
19 <sup>a</sup>	9.31	8.45	44 <sup>a</sup>	6.49	7.46
20	6.90	7.44	45	6.92	6.91
21	7.60	7.50	46	9.02	9.14
22	7.40	7.70	47 <sup>a</sup>	7.52	7.30
23	8.06	7.88	48	7.19	7.19
24 <sup>a</sup>	6.09	6.26	49 <sup>a</sup>	8.11	7.76
25	5.59	5.74	50	7.59	7.83

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
51	7.70	7.70	56	8.08	8.38
52	6.68	6.68	57	9.48	9.48
53	8.01	8.01	58	8.03	8.03
54	8.44	8.44	59	6.71	6.71
55	7.26	7.42	60	8.17	8.16

<sup>a</sup> - Test set compound

**Table A10:** Predictions of pIC<sub>50</sub> values of 2-substituted Dipyridodiazeponone derivative inhibitors of HIV-1 RT for TS-3 using image-based QSAR model (AID: 198247)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1	5.43	6.50	26	7.70	7.12
2	6.60	6.58	27 <sup>a</sup>	6.92	7.07
3	6.74	6.57	28	7.40	7.39
4	7.15	7.15	29 <sup>a</sup>	6.33	6.95
5 <sup>a</sup>	7.22	7.24	30 <sup>a</sup>	6.51	6.76
6	6.41	6.63	31	7.05	6.84
7	5.85	5.89	32	6.82	7.12
8	7.00	6.96	33	6.40	7.18
9	6.82	6.94	34	6.52	6.67
10	6.09	6.11	35	7.40	7.41
11	6.64	6.65	36	7.52	7.49
12	7.70	6.94	37	7.70	7.71
13	6.41	6.67	38	8.00	8.00
14	6.89	7.10	39	7.15	7.39
15	6.96	6.98	40	6.85	6.84
16 <sup>a</sup>	6.66	7.04	41	7.70	7.70
17	6.42	6.42	42	6.82	6.80
18	7.00	7.00	43	7.30	7.30
19	7.00	7.00	44 <sup>a</sup>	5.96	6.36
20 <sup>a</sup>	6.85	6.84	45	5.92	5.93
21	7.30	7.12	46	6.74	6.66
22	7.52	7.70	47 <sup>a</sup>	7.74	7.69
23	7.15	7.09	48 <sup>a</sup>	7.40	7.05
24	7.40	7.05	49	7.05	7.05
25	6.96	6.83	50	7.05	6.93

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
51	6.64	6.64	60	7.05	7.10
52	6.72	6.70	61	7.15	7.07
53	6.00	5.99	62	7.15	7.16
54	7.52	7.35	63	7.05	7.14
55	8.00	7.11	64	7.40	7.41
56	7.70	7.11	65 <sup>a</sup>	7.10	6.91
57	7.70	7.11	66	6.92	7.01
58	6.00	6.00	67	6.89	6.98
59 <sup>a</sup>	6.00	6.28	68	7.10	6.99

<sup>a</sup> - Test set compound

**Table A11:** Predictions pIC<sub>50</sub> values of 2-pyridinone derivative inhibitors of HIV-1 RT for TS-4 using image-based QSAR model (AID: 197804)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1	5.02	5.53	26	6.99	6.79
2	6.52	5.51	27	6.18	5.88
3	7.38	6.78	28 <sup>a</sup>	7.24	5.50
4	5.71	5.79	29	4.17	4.23
5	6.47	6.49	30	4.61	4.61
6	7.24	7.30	31	6.36	6.36
7	6.47	5.57	32 <sup>a</sup>	6.74	5.49
8	3.52	5.50	33	7.15	6.18
9 <sup>a</sup>	4.96	5.51	34 <sup>a</sup>	6.98	6.85
10 <sup>a</sup>	4.50	4.49	35	7.04	6.08
11	5.76	5.79	36	5.90	6.10
12	5.94	5.94	37 <sup>a</sup>	6.33	6.93
13	6.22	6.22	38	6.96	6.71
14	6.37	6.38	39	7.19	6.68
15	6.72	6.74	40	6.82	5.53
16	5.55	5.41	41	6.59	6.60
17	5.98	6.18	42 <sup>a</sup>	5.78	6.21
18	6.52	6.52	43	5.90	5.91
19	5.95	5.66	44 <sup>a</sup>	3.84	5.48
20	7.37	7.37	45	3.98	5.90
21	6.94	6.95	46	4.49	4.49
22	6.95	6.99	47	4.54	4.60
23	7.64	7.55	48	4.65	5.59
24	5.98	5.76	49	4.82	5.51
25	6.01	6.01	50	5.00	5.00



Compound	Experimental pIC50	Predicted pIC50	Compound	Experimental pIC50	Predicted pIC50
51	5.36	5.62	62	3.52	5.50
52 <sup>a</sup>	5.57	5.39	63	7.26	7.26
53	5.60	5.17	64	6.92	6.72
54	5.63	5.75	65	6.34	5.50
55	5.68	5.51	66 <sup>a</sup>	6.48	6.77
56	5.72	5.72	67	6.46	6.58
57	5.96	5.96	68	5.12	5.52
58	6.28	5.50	69 <sup>a</sup>	7.52	7.68
59	6.30	6.38	70	6.68	6.67
60	6.55	6.54	71	7.72	7.71
61 <sup>a</sup>	5.27	5.50	72	7.70	7.43

<sup>a</sup> - Test set compound

**Table A12:** Predictions of pIC<sub>50</sub> values of cyclic urea derivative inhibitors of HIV-1 PR for TS-5 using image-based QSAR model (AID: 160292)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1 <sup>a</sup>	10.42	9.29	26	9.55	8.59
2	10.16	10.16	27	9.48	9.05
3	10.28	10.35	28	9.00	8.94
4	10.33	10.40	29	8.16	8.18
5 <sup>a</sup>	10.64	8.48	30	9.03	9.76
6	10.92	9.98	31	8.44	7.93
7	10.62	8.37	32	8.28	8.35
8	8.60	8.58	33	8.64	8.50
9	9.07	9.38	34	8.85	8.00
10	10.12	8.89	35	8.82	8.02
11	10.02	10.02	36	9.92	9.93
12	9.39	10.48	37	9.22	7.99
13	10.80	8.94	38	9.92	9.93
14	5.40	7.95	39 <sup>a</sup>	9.38	8.37
15 <sup>a</sup>	8.74	7.96	40	8.55	8.54
16	8.14	7.88	41	7.05	7.06
17	7.44	8.21	42 <sup>a</sup>	8.01	7.89
18	10.74	10.35	43	8.96	7.99
19	10.41	10.31	44	6.84	6.84
20	10.20	10.47	45	7.66	7.92
21	9.24	9.41	46 <sup>a</sup>	7.22	7.60
22 <sup>a</sup>	9.68	8.19	47	5.73	7.08
23 <sup>a</sup>	10.18	8.43	48	7.29	7.98
24	10.35	9.06	49 <sup>a</sup>	7.66	7.99
25	10.70	10.22	50	8.24	8.26

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
51	8.15	8.61	68	6.10	8.02
52	8.85	8.30	69 <sup>a</sup>	8.34	8.03
53	7.57	8.07	70	8.80	8.01
54	8.28	8.04	71	8.85	8.19
55	9.05	8.25	72	8.10	7.89
56	6.62	7.97	73	7.00	8.04
57	8.85	7.97	74	5.24	8.17
58 <sup>a</sup>	7.47	7.04	75	8.52	8.10
59	8.52	9.11	76	9.85	10.21
60	7.43	7.74	77	8.80	8.62
61	8.37	8.38	78	7.07	8.06
62	8.89	7.99	79 <sup>a</sup>	6.80	8.19
63	7.52	8.05	80	8.68	7.74
64	8.15	8.03	81	8.28	8.02
65	7.92	8.14	82 <sup>a</sup>	9.51	8.49
66	7.31	8.07	83	9.55	9.51
67 <sup>a</sup>	5.96	7.49	84	9.47	10.19

<sup>a</sup> - Test set compound

**Table A13:** Predictions of pIC<sub>50</sub> values of anti-malarial azilide derivatives for TS-6 using image-based QSAR model (AID: 579588)

Compounds	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compounds	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1	6.80	6.79	26 <sup>a</sup>	6.68	6.53
2 <sup>a</sup>	4.89	5.35	27	6.75	6.89
3	5.29	5.31	28	6.90	6.65
4	5.52	5.52	29	7.07	7.07
5	5.80	5.80	30	7.13	7.13
6	6.21	6.21	31	5.93	5.93
7	6.32	6.32	32	6.91	6.88
8	6.66	6.38	33	6.03	6.03
9	6.99	6.58	34	6.31	6.31
10	7.02	7.02	35	6.39	6.38
11	7.37	7.37	36	6.54	6.54
12	6.53	6.52	37	6.63	6.63
13	6.01	6.04	38 <sup>a</sup>	6.74	6.53
14	6.26	6.26	39 <sup>a</sup>	6.76	6.64
15 <sup>a</sup>	6.60	6.77	40	6.84	6.84
16	6.79	6.79	41	6.86	6.85
17	6.85	6.78	42	6.86	6.86
18	6.86	6.72	43	6.86	6.86
19	6.99	6.99	44	6.90	6.98
20 <sup>a</sup>	7.11	6.59	45 <sup>a</sup>	6.93	6.65
21	7.17	7.17	46 <sup>a</sup>	6.99	7.00
22	6.69	6.67	47	7.07	6.55
23	6.10	6.47	48	7.26	7.22
24	6.39	6.39	49	6.10	6.10
25	6.58	6.57	50 <sup>a</sup>	6.29	6.48

Compounds	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compounds	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
51	6.31	6.30	75 <sup>a</sup>	6.88	6.69
52	6.36	6.39	76	6.97	7.13
53	6.39	6.39	77	7.11	7.09
54 <sup>a</sup>	6.42	6.40	78	7.26	7.26
55	6.63	6.63	79	4.89	6.32
56	6.66	6.66	80 <sup>a</sup>	5.70	6.51
57	6.86	6.84	81 <sup>a</sup>	6.09	6.57
58	6.93	6.93	82	6.24	6.24
59	4.91	4.92	83	6.45	6.45
60	5.48	6.68	84	6.49	6.52
61	6.39	6.45	85	6.74	6.74
62	6.40	6.40	86	6.76	6.81
63	6.48	6.54	87	6.82	6.82
64 <sup>a</sup>	6.85	6.57	88	6.92	6.88
65	6.94	6.95	89 <sup>a</sup>	7.00	6.66
66	6.97	6.97	90	7.04	7.04
67	7.14	7.13	91	7.19	7.19
68	7.26	7.26	92	7.22	6.74
69	7.02	7.03	93 <sup>a</sup>	7.27	6.65
70 <sup>a</sup>	6.36	6.69	94	7.16	6.56
71	6.37	6.35	95 <sup>a</sup>	7.25	7.24
72	6.62	6.57	96	7.50	7.50
73	6.73	6.66	97	7.29	6.85
74	6.78	6.78	98	5.69	5.72

<sup>a</sup> - Test set compound

**Table A14:** PME-PLS notation list

Symbol	Description
$n$	Total number of inhibitor compounds for a TS. $n_{train} + n_{test} + n_{val} = n$
$m$	Number of PMF values for a compound (Total number of dimensions)
$a$	Number of PLS components chosen
$\sigma$	Scaling factor for PMF values Eq. (1).
$x_i$	Row vector of PMF values (i.e., PFMD) for $i^{th}$ compound, size $[1\ m]$ , $i = 1 \dots n$ for a TS,
$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$	Matrix of PFMDs, size $[n\ m]$ , for the compounds in a TS
$y_i$	pIC <sub>50</sub> value of $i^{th}$ compound, $i = 1 \dots n$
$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$	Column vector of pIC <sub>50</sub> values for the compounds, size $[n\ 1]$
$Y_{range}$	Difference between the maximum and the minimum pIC <sub>50</sub> value in a TS (= max( $Y$ ) – min( $Y$ ))
$n_{train}$	Number of training set compounds for a TS
$n_{test}$	Number of test set compounds for a TS
$n_{val}$	Number of validation set compounds for a TS.
$n_{count}$	Number of iterations for selecting reference data sets
$x_{train,i}$	PFMD for $i^{th}$ training set compound, size $[1\ m]$ , $i = 1 \dots n_{train}$
$\mathbf{X}_{train} = \begin{bmatrix} x_{train,1} \\ x_{train,2} \\ \vdots \\ x_{train,n_{train}} \end{bmatrix}$	Matrix of PFMDs for training set compounds, size $[n_{train}\ m]$
$y_{train,i}$	pIC <sub>50</sub> value of $i^{th}$ training set compound, $i = 1 \dots n_{train}$
$\mathbf{Y}_{train} = \begin{bmatrix} y_{train,1} \\ y_{train,2} \\ \vdots \\ y_{train,n_{train}} \end{bmatrix}$	Column vector of pIC <sub>50</sub> values for the training set compounds, size $[n_{train}\ 1]$
$x_{test,i}$	PFMD for $i^{th}$ test set compound. size $[1\ m]$ . $i = 1 \dots n_{test}$
$\mathbf{X}_{test} = \begin{bmatrix} x_{test,1} \\ x_{test,2} \\ \vdots \\ x_{test,n_{test}} \end{bmatrix}$	Matrix of PFMDs for the test set compounds. size $[n_{test}\ m]$
$y_{test,i}$	pIC <sub>50</sub> value of $i^{th}$ test set compound. $i = 1 \dots n_{test}$
$\mathbf{Y}_{test} = \begin{bmatrix} y_{test,1} \\ y_{test,2} \\ \vdots \\ y_{test,n_{test}} \end{bmatrix}$	Column vector of pIC <sub>50</sub> values for the test set compounds, size $[n_{test}\ 1]$
$x_{train,ref,i}$	PFMD for $i^{th}$ reference training set compound, size $[1\ m]$ . $i = 1 \dots n_{train}$

---

$\mathbf{X}_{train,ref} = \begin{bmatrix} x_{train,ref,1} \\ x_{train,ref,2} \\ \vdots \\ x_{train,ref,n_{train}} \end{bmatrix}$	<p>Matrix of PFMDs for the reference training set compounds, size <math>[n_{train} \ m]</math></p>
$y_{train,ref,i}$	<p>pIC<sub>50</sub> value of <math>i^{th}</math> reference training set compound, <math>i = 1 \dots n_{train}</math></p>
$\mathbf{Y}_{train,ref} = \begin{bmatrix} y_{train,ref,1} \\ y_{train,ref,2} \\ \vdots \\ y_{train,ref,n_{train}} \end{bmatrix}$	<p>Column vector of pIC<sub>50</sub> values for the reference training set compounds, size <math>[n_{train} \ 1]</math></p>
$x_{test,ref,i}$	<p>PFMD for <math>i^{th}</math> reference test set compound, size <math>[1 \ m]</math>, <math>i = 1 \dots n_{test}</math></p>
$\mathbf{X}_{test,ref} = \begin{bmatrix} x_{test,ref,1} \\ x_{test,ref,2} \\ \vdots \\ x_{test,ref,n_{test}} \end{bmatrix}$	<p>Matrix of PFMDs for the reference test set compounds, size <math>[n_{test} \ m]</math></p>
$y_{test,ref,i}$	<p>pIC<sub>50</sub> value of <math>i^{th}</math> reference test set compound, <math>i = 1 \dots n_{test}</math></p>
$\mathbf{Y}_{test,ref} = \begin{bmatrix} y_{test,ref,1} \\ y_{test,ref,2} \\ \vdots \\ y_{test,ref,n_{test}} \end{bmatrix}$	<p>Column vector of pIC<sub>50</sub> values for the reference test set compounds, size <math>[n_{test} \ 1]</math></p>
$x_{val,i}$	<p>PFMD for <math>i^{th}</math> validation set compound, size <math>[1 \ m]</math>, <math>i = 1 \dots n_{val}</math></p>
$\mathbf{X}_{val} = \begin{bmatrix} x_{val,1} \\ x_{val,2} \\ \vdots \\ x_{val,n_{val}} \end{bmatrix}$	<p>Matrix of PFMDs for all the validation set compounds, size <math>[n_{val} \ * \ m]</math></p>
$y_{val,i}$	<p>pIC<sub>50</sub> value of <math>i^{th}</math> validation set compound, <math>i = 1 \dots n_{val}</math></p>
$\mathbf{Y}_{val} = \begin{bmatrix} y_{val,1} \\ y_{val,2} \\ \vdots \\ y_{val,n_{val}} \end{bmatrix}$	<p>Column vector of pIC<sub>50</sub> values for the validation set compounds, size <math>[n_{val} \ 1]</math></p>
$\mathbf{X}_{train,j}^- = \begin{bmatrix} x_{train,ref,1} \\ \vdots \\ x_{train,ref,j-1} \\ x_{train,ref,j+1} \\ \vdots \\ x_{train,ref,n_{train}} \end{bmatrix}$	<p><math>\mathbf{X}_{train,ref}</math> with <math>j^{th}</math> row removed, size <math>[(n_{train} - 1) \ m]</math>, <math>j = 1 \dots n_{train}</math></p>
$\mathbf{Y}_{train,j}^- = \begin{bmatrix} y_{train,ref,1} \\ \vdots \\ y_{train,ref,j-1} \\ y_{train,ref,j+1} \\ \vdots \\ y_{train,ref,n_{train}} \end{bmatrix}$	<p><math>\mathbf{Y}_{train,ref}</math> with <math>j^{th}</math> element removed, size <math>[(n_{train} - 1) \ 1]</math>, <math>j = 1 \dots n_{train}</math></p>
$\mathbf{X}_{test,j}^+ = \begin{bmatrix} x_{test,ref,1} \\ x_{test,ref,2} \\ \vdots \\ x_{test,ref,n_{test}} \\ x_{train,ref,j} \end{bmatrix}$	<p><math>\mathbf{X}_{test,ref}</math> with <math>j^{th}</math> row from <math>\mathbf{X}_{train,ref}</math> added at the bottom, size <math>[(n_{test} + 1) \ m]</math>, <math>j = 1 \dots n_{train}</math></p>

---

---

$\mathbf{Y}_{test}^+ = \begin{bmatrix} y_{test,ref,1} \\ y_{test,ref,2} \\ \vdots \\ y_{test,ref,n_{test}} \\ y_{train,ref,j} \end{bmatrix}$	$\mathbf{Y}_{test,ref}$ with $j^{th}$ element from $\mathbf{Y}_{train,ref}$ added at the bottom, size $[(n_{test} + 1) \ 1], j = 1 \dots n_{train}$
$\mathbf{X}_{train,j}^+ = \begin{bmatrix} x_{train,ref,1} \\ x_{train,ref,2} \\ \vdots \\ x_{train,ref,n_{train}} \\ x_{test,ref,j} \end{bmatrix}$	$\mathbf{X}_{train,ref}$ with $j^{th}$ row from $\mathbf{X}_{test,ref}$ added at the bottom, size $[(n_{train} + 1) \ m], j = 1 \dots n_{test}$
$\mathbf{Y}_{train}^+ = \begin{bmatrix} y_{train,ref,1} \\ y_{train,ref,2} \\ \vdots \\ y_{train,ref,n_{train}} \\ y_{test,j} \end{bmatrix}$	$\mathbf{Y}_{train,ref}$ with $j^{th}$ element from $\mathbf{Y}_{test,ref}$ added at the bottom, size $[(n_{train} + 1) \ 1], j = 1 \dots n_{test}$
$\mathbf{X}_{test,j}^- = \begin{bmatrix} x_{test,ref,1} \\ \vdots \\ x_{test,ref,j-1} \\ x_{test,ref,j+1} \\ \vdots \\ x_{test,ref,n_{test}} \end{bmatrix}$	$\mathbf{X}_{test,ref}$ with $j^{th}$ row removed, size $[(n_{test} - 1) \ m], j = 1 \dots n_{test}$
$\mathbf{Y}_{test,j}^- = \begin{bmatrix} y_{test,ref,1} \\ \vdots \\ y_{test,ref,j-1} \\ y_{test,ref,j+1} \\ \vdots \\ y_{test,ref,n_{test}} \end{bmatrix}$	$\mathbf{Y}_{test,ref}$ with $j^{th}$ element removed, size $[(n_{test} - 1) \ 1], j = 1 \dots n_{test}$
$\hat{\mathbf{Y}}_{train}$	Model predictions for training set ( $\mathbf{X}_{train}, \mathbf{Y}_{train}$ )
$\hat{\mathbf{Y}}_{test}$	Model predictions for test set ( $\mathbf{X}_{test}, \mathbf{Y}_{test}$ )
$\hat{\mathbf{Y}}_{train,j}^-$	Model predictions for training set ( $\mathbf{X}_{train}^-, \mathbf{Y}_{train}^-$ )
$\hat{\mathbf{Y}}_{test,j}^+$	Model predictions for test set ( $\mathbf{X}_{test}^+, \mathbf{Y}_{test}^+$ )
$\hat{\mathbf{Y}}_{train,j}^+$	Model predictions for training set ( $\mathbf{X}_{train}^+, \mathbf{Y}_{train}^+$ )
$\hat{\mathbf{Y}}_{test,j}^-$	Model predictions for test set ( $\mathbf{X}_{test}^-, \mathbf{Y}_{test}^-$ )
$e_{train,count}$	RMSE in predicting $\hat{\mathbf{Y}}_{train}$
$e_{test,count}$	RMSE in predicting $\hat{\mathbf{Y}}_{test}$
$e_{train,ref}$	RMSE in predicting $\hat{\mathbf{Y}}_{train,ref}$
$e_{test,ref}$	RMSE in predicting $\hat{\mathbf{Y}}_{test,ref}$
$e_{train,j}^-$	RMSE in predicting $\hat{\mathbf{Y}}_{train,j}^-$
$e_{test,j}^+$	RMSE in predicting $\hat{\mathbf{Y}}_{test,j}^+$
$e_{train,j}^+$	RMSE in predicting $\hat{\mathbf{Y}}_{train,j}^+$
$e_{test,j}^-$	RMSE in predicting $\hat{\mathbf{Y}}_{test,j}^-$
$\delta$	Threshold value used in [B 3.19] and [B 3.30] ( $= 0.15 * Y_{range}$ )
$\mathbf{T}$	Score matrix for $\mathbf{X}_{train}$ , size $[n_{train} \ a]$ , obtained on PLS of ( $\mathbf{X}_{train}, \mathbf{Y}_{train}$ )
$\mathbf{P}$	Loading matrix, size $[m \ a]$ , obtained on PLS of ( $\mathbf{X}_{train}, \mathbf{Y}_{train}$ )
$\mathbf{B}$	Regression coefficients obtained on PLS of ( $\mathbf{X}_{train}, \mathbf{Y}_{train}$ )

---



---

$T_{ref}$	Score matrix for $X_{train,ref}$ , size $[n_{train} a]$ , obtained on PLS of $(X_{train,ref}, Y_{train,ref})$
$P_{ref}$	Loading matrix, size $[m a]$ , obtained on PLS of $(X_{train,ref}, Y_{train,ref})$
$B_{ref}$	Regression coefficients obtained on PLS of $(X_{train}, Y_{train})$
$T_j^-$	Score matrix for $X_{train,j}^-$ , size $[(n_{train} - 1) a]$ , obtained by PLS of $(X_{train,j}^-, Y_{train,j}^-)$
$P_j^-$	Loading matrix, size $[m a]$ , obtained by PLS of $(X_{train,j}^-, Y_{train,j}^-)$
$T_j^+$	Score matrix for $X_{train,j}^+$ , size $[(n_{train} + 1) a]$ , obtained by PLS of $(X_{train,j}^+, Y_{train,j}^+)$
$P_j^+$	Loading matrix, size $[m a]$ , obtained by PLS of $(X_{train,j}^+, Y_{train,j}^+)$
$T_{r,j}^-$	Score matrix, size $[(n_{train} - 1) a]$ , after Procrustes transformation of $T_j^-$
$P_{r,j}^-$	Loading matrix, size $[m a]$ , after Procrustes transformation of $P_j^-$
$T_{r,j}^+$	Score matrix, size $[(n_{train} + 1) a]$ , after Procrustes transformation of $T_j^+$
$P_{r,j}^+$	Loading matrix, size $[m a]$ , after Procrustes transformation of $P_j^+$
$X_{r,train,j}^-$	Matrix of reconstructed PFMDs, size $[(n_{train} - 1) m]$ , for compounds in $(X_{train,j}^-, Y_{train,j}^-)$ ( $= T_{r,j}^- (P_{r,j}^-)'$ )
$X_{r,train,j}^+$	Matrix of reconstructed PFMDs, size $[(n_{train} + 1) m]$ , for compounds in $(X_{train,j}^+, Y_{train,j}^+)$ ( $= T_{r,j}^+ (P_{r,j}^+)'$ )
$B_{train,j}$	Regression coefficients obtained by PLS of $(X_{train,j}^-, Y_{train,j}^-)$ or $(X_{r,train,j}^-, Y_{r,train,j}^-)$
$B_{test,j}$	Regression coefficients obtained by PLS of $(X_{train,j}^+, Y_{train,j}^+)$ or $(X_{r,train,j}^+, Y_{r,train,j}^+)$
$B^-$	Matrix of the selected $B_{train,j}$ , $j = 1, 2, \dots, n_{train}$ [B 3.21] or [B 3.22]
$B^+$	Matrix of the selected $B_{test,j}$ , $j = 1, 2, \dots, n_{test}$ [B 3.34] or [B 3.35]
$B_{avg}$	Average regression coefficients ( $= \frac{\sum B^- + \sum B^+}{n_{train} + n_{test}}$ )
$\hat{Y}_{train,ref}$	Model predictions for training set using $B_{avg}$
$\hat{Y}_{test,ref}$	Model predictions for test set using $B_{avg}$
$\hat{Y}_{val}$	Model predictions for validation using $B_{avg}$
RMSE	Root Mean Squared Error
NRMSECV	Normalized RMSE for cross-validation
NRMSEP	Normalized RMSE of prediction for external validation
$r_{cv}$	Pearson's correlation coefficient for cross-validation
$r_{pred}$	Pearson's correlation coefficient for external validation
$R_{cv}^2$	Coefficient of determination for cross-validation
$Q_{ext(F1)}^2$	Coefficient of determination of prediction for external validation

---

**Table A15:** Predictions of pIC<sub>50</sub> values of Wee1 inhibitors for TS-1 using PMF-PLS QSAR model (AID: 268838)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1 <sup>b</sup>	7.01	6.82	26	6.24	6.19
2	5.40	5.40	27 <sup>b</sup>	6.35	6.85
3 <sup>b</sup>	5.64	6.54	28	6.72	6.75
4	6.51	6.52	29 <sup>b</sup>	6.08	6.89
5	6.37	6.37	30	6.80	6.81
6	7.11	7.01	31 <sup>a</sup>	6.24	7.04
7	6.59	6.68	32	7.22	7.20
8	6.89	6.92	33	7.62	7.63
9	5.80	5.80	34	6.59	6.62
10	5.01	5.06	35	7.48	7.45
11 <sup>a</sup>	5.64	5.23	36 <sup>a</sup>	6.66	6.42
12 <sup>a</sup>	5.40	6.14	37	7.33	6.62
13	4.70	4.59	38	6.68	6.96
14	4.43	4.30	39	6.66	6.50
15 <sup>b</sup>	5.55	5.31	40 <sup>b</sup>	7.26	7.25
16	5.41	5.90	41	6.64	6.57
17 <sup>a</sup>	7.96	7.92	42	6.06	6.38
18	6.19	6.99	43	5.36	5.07
19	7.24	6.80	44	6.75	6.73
20 <sup>b</sup>	7.48	6.80	45	5.37	5.02
21 <sup>a</sup>	6.48	7.20	46	4.40	5.51
22	7.64	7.67	47	7.05	7.58
23	7.89	7.81	48	6.21	6.22
24	6.82	6.94	49	6.52	6.12
25	6.29	6.79	50	7.16	7.14

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
51	4.80	4.80	75	6.66	6.73
52 <sup>a</sup>	6.14	6.58	76	6.02	6.17
53	5.48	5.42	77	7.13	7.05
54 <sup>a</sup>	5.92	6.48	78	7.72	7.57
55	5.75	5.89	79	6.56	6.55
56	5.44	5.33	80 <sup>b</sup>	7.57	6.95
57 <sup>a</sup>	7.17	7.31	81	7.54	7.42
58	4.92	5.00	82	7.75	7.74
59 <sup>a</sup>	4.54	5.34	83 <sup>a</sup>	7.31	6.95
60	5.96	5.90	84	6.85	6.94
61	6.82	6.93	85	7.38	7.29
62 <sup>a</sup>	7.55	8.06	86 <sup>b</sup>	6.75	6.57
63	7.92	7.89	87	7.42	7.27
64	7.68	7.34	88	5.89	5.98
65 <sup>b</sup>	7.64	7.47	89	6.09	6.06
66 <sup>a</sup>	7.62	7.48	90 <sup>b</sup>	6.24	6.01
67	6.31	6.13	91 <sup>b</sup>	5.00	5.01
68	7.38	7.47	92	7.30	7.22
69 <sup>a</sup>	7.70	7.58	93	7.20	7.16
70	7.55	6.97	94	7.28	7.26
71	7.35	7.33	95	7.23	7.15
72 <sup>b</sup>	7.82	7.49	96	6.82	6.84
73	7.70	7.74	97	6.77	6.85
74	7.46	7.25			

<sup>a</sup> - Test set compound

<sup>b</sup> - Validation set compound

**Table A16:** Predictions of pIC<sub>50</sub> values of AChE inhibitors for TS-2 using PMF-PLS QSAR model (AID: 566585)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1 <sup>b</sup>	8.06	7.63	26	8.14	8.05
2 <sup>b</sup>	8.60	8.48	27	6.47	6.97
3 <sup>b</sup>	8.34	8.24	28	9.32	8.91
4 <sup>a</sup>	8.96	8.55	29 <sup>a</sup>	7.64	7.67
5	8.54	8.53	30	7.92	7.48
6	7.19	7.46	31 <sup>b</sup>	8.15	7.88
7	6.82	6.81	32	6.05	6.82
8	7.80	7.79	33 <sup>a</sup>	8.37	7.72
9	8.31	8.14	34	8.55	8.29
10	8.29	8.10	35 <sup>b</sup>	7.30	7.10
11	7.05	7.25	36 <sup>a</sup>	6.78	7.44
12	8.89	8.64	37	7.28	7.47
13 <sup>a</sup>	6.46	7.16	38 <sup>b</sup>	7.00	6.87
14	9.10	8.49	39	9.24	8.71
15 <sup>a</sup>	7.28	7.75	40 <sup>b</sup>	6.52	7.30
16 <sup>a</sup>	7.48	7.69	41	8.44	8.37
17	7.85	7.90	42	8.24	8.27
18	7.37	7.40	43	6.66	7.19
19	9.31	8.85	44	6.49	6.95
20	6.90	7.02	45	6.92	7.28
21 <sup>b</sup>	7.60	8.14	46	9.02	8.62
22	7.40	7.60	47 <sup>b</sup>	7.52	7.45
23 <sup>a</sup>	8.06	7.84	48	7.19	7.35
24	6.09	7.15	49	8.11	8.12
25	5.59	6.15	50 <sup>a</sup>	7.59	7.59

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
51	7.70	7.31	56	8.08	8.12
52	6.68	6.87	57	9.48	8.55
53	8.01	8.06	58	8.03	7.77
54	8.44	8.20	59	6.71	7.01
55	7.26	7.41	60 <sup>a</sup>	8.17	7.97

<sup>a</sup> - Test set compound

<sup>b</sup> - Validation set compound

**Table A17:** Predictions of pIC<sub>50</sub> values of 2-substituted Dipyridodiazeponone derivative inhibitors of HIV-1 RT for TS-3 using PMF-PLS QSAR model (AID: 198247)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1	5.43	6.40	26	7.70	7.55
2	6.60	6.55	27 <sup>a</sup>	6.92	6.95
3	6.74	6.68	28	7.40	7.27
4	7.15	7.05	29 <sup>b</sup>	6.33	6.68
5	7.22	6.87	30 <sup>a</sup>	6.51	6.63
6	6.41	6.78	31	7.05	6.96
7	5.85	6.47	32	6.82	6.70
8	7.00	6.50	33	6.40	6.40
9 <sup>b</sup>	6.82	6.55	34 <sup>b</sup>	6.52	6.95
10	6.09	6.71	35 <sup>b</sup>	7.40	6.99
11 <sup>a</sup>	6.64	6.73	36	7.52	7.31
12	7.70	6.84	37	7.70	7.42
13	6.41	6.83	38	8.00	7.90
14 <sup>b</sup>	6.89	6.74	39 <sup>a</sup>	7.15	7.32
15 <sup>b</sup>	6.96	6.85	40	6.85	6.91
16	6.66	6.20	41	7.70	7.53
17	6.42	6.70	42 <sup>a</sup>	6.82	7.14
18	7.00	6.44	43	7.30	7.13
19	7.00	6.62	44 <sup>a</sup>	5.96	5.76
20	6.85	6.72	45	5.92	5.88
21	7.30	7.18	46	6.74	6.67
22 <sup>b</sup>	7.52	7.68	47	7.74	7.54
23	7.15	6.87	48 <sup>a</sup>	7.40	7.60
24	7.40	6.69	49 <sup>a</sup>	7.05	6.99
25	6.96	6.83	50	7.05	6.94

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
51	6.64	6.74	60 <sup>b</sup>	7.05	7.51
52	6.72	6.43	61 <sup>b</sup>	7.15	6.97
53	6.00	6.52	62	7.15	6.69
54 <sup>a</sup>	7.52	7.09	63	7.05	7.17
55 <sup>b</sup>	8.00	8.39	64	7.40	7.07
56	7.70	7.50	65	7.10	6.96
57 <sup>a</sup>	7.70	7.44	66	6.92	6.87
58	6.00	5.95	67	6.89	6.80
59	6.00	5.97	68	7.10	6.89

<sup>a</sup> - Test set compound

<sup>b</sup> - Validation set compound

**Table A18:** Predictions pIC<sub>50</sub> values of 2-pyridinone derivative inhibitors of HIV-1 RT for TS-4 using PMF-PLS QSAR model (AID: 197804)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1	5.02	5.23	26	6.99	6.57
2	6.52	6.22	27 <sup>a</sup>	6.18	6.02
3 <sup>b</sup>	7.38	6.53	28 <sup>a</sup>	7.24	7.25
4	5.71	5.64	29	4.17	4.28
5	6.47	6.42	30 <sup>a</sup>	4.61	4.51
6 <sup>b</sup>	7.24	6.29	31	6.36	6.10
7	6.47	6.15	32	6.74	7.21
8	3.52	3.31	33	7.15	6.60
9 <sup>b</sup>	4.96	5.71	34 <sup>a</sup>	6.98	6.36
10	4.50	4.76	35	7.04	6.94
11	5.76	5.60	36	5.90	7.01
12 <sup>b</sup>	5.94	6.40	37	6.33	6.10
13	6.22	6.20	38	6.96	6.66
14	6.37	6.31	39	7.19	6.45
15	6.72	6.68	40	6.82	6.62
16	5.55	5.42	41	6.59	6.17
17	5.98	5.66	42	5.78	6.22
18	6.52	5.97	43	5.90	5.48
19	5.95	5.85	44 <sup>a</sup>	3.84	4.43
20	7.37	7.40	45 <sup>b</sup>	3.98	5.13
21 <sup>a</sup>	6.94	7.10	46	4.49	4.41
22 <sup>b</sup>	6.95	7.25	47	4.54	5.16
23	7.64	7.48	48	4.65	4.54
24 <sup>a</sup>	5.98	5.42	49	4.82	4.31
25 <sup>b</sup>	6.01	6.20	50	5.00	4.82



Compound	Experimental pIC50	Predicted pIC50	Compound	Experimental pIC50	Predicted pIC50
51 <sup>a</sup>	5.36	5.65	62	3.52	5.15
52 <sup>b</sup>	5.57	5.84	63	7.26	7.32
53	5.60	5.78	64 <sup>b</sup>	6.92	6.65
54	5.63	5.20	65 <sup>b</sup>	6.34	6.56
55	5.68	5.89	66 <sup>a</sup>	6.48	5.88
56 <sup>a</sup>	5.72	5.84	67	6.46	5.84
57	5.96	5.71	68	5.12	5.25
58	6.28	6.33	69 <sup>a</sup>	7.52	6.48
59	6.30	6.11	70	6.68	6.42
60	6.55	5.79	71	7.72	6.35
61	5.27	5.32	72	7.70	7.64

<sup>a</sup> - Test set compound

<sup>b</sup> - Validation set compound

**Table A19:** Predictions of pIC<sub>50</sub> values of cyclic urea derivative inhibitors of HIV-1 PR for TS-5 using PMF-PLS QSAR model (AID: 160292)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1	10.42	10.31	26	9.55	9.29
2	10.16	10.10	27 <sup>b</sup>	9.48	9.55
3 <sup>b</sup>	10.28	9.29	28	9.00	8.62
4	10.33	10.33	29	8.16	9.00
5 <sup>b</sup>	10.64	9.98	30 <sup>a</sup>	9.03	9.50
6	10.92	10.61	31	8.44	8.47
7	10.62	10.54	32	8.28	8.28
8 <sup>b</sup>	8.60	9.28	33	8.64	8.06
9 <sup>a</sup>	9.07	8.43	34 <sup>b</sup>	8.85	9.10
10	10.12	10.08	35	8.82	9.22
11	10.02	9.86	36	9.92	9.53
12	9.39	9.00	37	9.22	9.29
13	10.80	11.22	38 <sup>b</sup>	9.92	9.75
14 <sup>a</sup>	5.40	5.07	39	9.38	8.77
15	8.74	8.54	40	8.55	8.67
16	8.14	7.54	41	7.05	7.73
17	7.44	7.50	42	8.01	7.88
18	10.74	10.65	43	8.96	8.65
19 <sup>a</sup>	10.41	10.72	44 <sup>b</sup>	6.84	7.90
20	10.20	10.12	45	7.66	7.48
21	9.24	9.00	46	7.22	6.91
22	9.68	9.66	47	5.73	5.39
23 <sup>a</sup>	10.18	9.94	48 <sup>a</sup>	7.29	7.99
24	10.35	10.24	49 <sup>a</sup>	7.66	9.19
25	10.70	10.01	50	8.24	8.34

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
51 <sup>b</sup>	8.15	8.65	68	6.10	6.22
52	8.85	8.89	69	8.34	8.51
53 <sup>b</sup>	7.57	8.87	70	8.80	8.23
54 <sup>a</sup>	8.28	8.72	71	8.85	8.82
55	9.05	8.85	72	8.10	7.94
56 <sup>a</sup>	6.62	6.39	73 <sup>a</sup>	7.00	6.83
57	8.85	8.95	74	5.24	6.53
58	7.47	7.17	75	8.52	8.39
59	8.52	8.54	76	9.85	9.85
60	7.43	7.08	77	8.80	8.73
61	8.37	8.26	78 <sup>b</sup>	7.07	8.39
62	8.89	8.85	79	6.80	6.76
63	7.52	7.29	80	8.68	8.59
64	8.15	8.27	81	8.28	8.03
65	7.92	7.73	82 <sup>a</sup>	9.51	10.38
66	7.31	7.65	83	9.55	8.19
67 <sup>b</sup>	5.96	7.45	84	9.47	9.39

<sup>a</sup> - Test set compound

<sup>b</sup> - Validation set compound

**Table A20:** Predictions of pIC<sub>50</sub> values of anti-malarial azilide derivatives for TS-6 using PMF-PLS QSAR model (AID: 579588)

Compounds	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compounds	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1 <sup>a</sup>	6.80	6.82	26	6.68	6.35
2	4.89	5.24	27	6.75	6.68
3 <sup>a</sup>	5.29	5.65	28	6.90	6.67
4 <sup>a</sup>	5.52	6.42	29	7.07	6.87
5 <sup>b</sup>	5.80	6.01	30 <sup>b</sup>	7.13	7.07
6	6.21	6.01	31 <sup>a</sup>	5.93	6.35
7 <sup>a</sup>	6.32	6.65	32	6.91	6.93
8	6.66	6.56	33 <sup>a</sup>	6.03	6.11
9	6.99	6.49	34	6.31	6.17
10	7.02	6.70	35	6.39	6.00
11	7.37	7.26	36 <sup>b</sup>	6.54	6.70
12 <sup>a</sup>	6.53	6.87	37	6.63	6.48
13 <sup>b</sup>	6.01	6.51	38	6.74	6.54
14 <sup>b</sup>	6.26	6.70	39 <sup>b</sup>	6.76	6.90
15	6.60	6.64	40	6.84	6.29
16	6.79	6.57	41	6.86	6.70
17 <sup>b</sup>	6.85	7.03	42	6.86	6.89
18	6.86	6.58	43	6.86	6.52
19 <sup>b</sup>	6.99	6.49	44	6.90	6.67
20 <sup>a</sup>	7.11	7.05	45	6.93	6.41
21 <sup>a</sup>	7.17	7.07	46	6.99	6.88
22	6.69	6.51	47	7.07	7.00
23 <sup>b</sup>	6.10	6.30	48	7.26	7.22
24	6.39	6.30	49	6.10	6.26
25 <sup>a</sup>	6.58	6.51	50	6.29	5.77

Compounds	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	Compounds	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
51	6.31	5.96	75 <sup>a</sup>	6.88	6.98
52	6.36	6.28	76 <sup>a</sup>	6.97	6.78
53	6.39	6.31	77	7.11	7.07
54 <sup>b</sup>	6.42	6.55	78 <sup>b</sup>	7.26	6.86
55	6.63	6.39	79	4.89	5.97
56 <sup>b</sup>	6.66	6.79	80	5.70	5.41
57	6.86	6.57	81	6.09	6.35
58	6.93	6.56	82 <sup>a</sup>	6.24	6.66
59	4.91	5.04	83	6.45	6.38
60 <sup>b</sup>	5.48	5.94	84	6.49	6.38
61	6.39	6.55	85 <sup>a</sup>	6.74	6.83
62	6.40	6.24	86	6.76	6.77
63	6.48	6.24	87	6.82	7.00
64	6.85	6.99	88 <sup>b</sup>	6.92	6.55
65	6.94	6.98	89	7.00	6.53
66	6.97	6.82	90	7.04	6.86
67	7.14	6.80	91	7.19	7.18
68	7.26	7.22	92	7.22	6.63
69	7.02	6.80	93	7.27	7.04
70	6.36	6.07	94	7.16	7.18
71	6.37	6.11	95 <sup>a</sup>	7.25	6.83
72	6.62	6.47	96	7.50	6.53
73	6.73	6.72	97	7.29	6.94
74	6.78	6.78	98	5.69	5.47

<sup>a</sup> - Test set compound

<sup>b</sup> - Validation set compound

**Table A21:** Predictions of pIC<sub>50</sub> values of Wee1 inhibitors for TS-1 using VC-PLS QSAR model (AID: 268838)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
1	7.01	6.82	6.82	6.98	6.98	6.82
2	5.40	5.64	5.50	5.50	5.50	5.64
3	5.64	5.92	5.92	5.92	5.92	5.76 <sup>a</sup>
4	6.51	7.48 <sup>a</sup>	6.86	6.86	6.86	6.86
5	6.37	6.25	6.52	6.52	6.52	6.52
6	7.11	7.03	7.03	7.03	7.03	6.74 <sup>a</sup>
7	6.59	6.46	6.46	6.46	6.46	6.46
8	6.89	6.61	6.61	6.82	6.82	6.61
9	5.80	5.66	5.66	5.66	5.66	5.66
10	5.01	5.70	5.83	5.83	5.83	5.83
11	5.64	5.76 <sup>a</sup>	5.76 <sup>a</sup>	5.54	5.54	5.92
12	5.40	5.50	5.60 <sup>a</sup>	5.60 <sup>a</sup>	5.60 <sup>a</sup>	5.50
13	4.70	5.58	5.58	4.74 <sup>a</sup>	5.58	5.58
14	4.43	4.74 <sup>a</sup>	4.74 <sup>a</sup>	4.96	4.74 <sup>a</sup>	4.96
15	5.55	5.54	5.54	5.76 <sup>a</sup>	5.64	5.54
16	5.41	6.16	6.16	6.16	6.16	6.16
17	7.96	7.75	7.75	7.75	7.75	7.75
18	6.19	7.04	7.04	7.04	7.04	7.04
19	7.24	6.80	6.80	6.80	6.80	6.80
20	7.48	6.96	6.96	7.40 <sup>a</sup>	6.96	7.40 <sup>a</sup>
21	6.48	6.86	6.25	6.25	6.25	6.25
22	7.64	7.64	7.29	7.29	7.29	7.64
23	7.89	7.60	7.60	7.60	7.60	7.60
24	6.82	6.47	6.47	6.47	6.47	6.47
25	6.29	6.64	6.64	6.64	6.64	6.64
26	6.24	6.78 <sup>a</sup>	6.29	6.78 <sup>a</sup>	6.20	6.78 <sup>a</sup>
27	6.35	6.52	6.32	7.48 <sup>a</sup>	6.32	7.48 <sup>a</sup>
28	6.72	7.29 <sup>a</sup>	6.69	7.29 <sup>a</sup>	6.63	6.69

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
29	6.08	6.74	6.48 <sup>a</sup>	6.36	6.74	6.36
30	6.80	7.14	7.14	7.14	7.14	7.14
31	6.24	6.42	6.42	6.42	6.42	6.42
32	7.22	7.03	7.03	7.03	7.03	7.03
33	7.62	7.66	7.66	7.66	7.19	7.66
34	6.59	6.25 <sup>a</sup>	6.60	6.60	6.25 <sup>a</sup>	6.60
35	7.48	7.28	7.28	7.28	7.28	7.28
36	6.66	6.69	6.69	6.25 <sup>a</sup>	6.69	6.69
37	7.33	6.76	6.76	6.76	6.76	6.76
38	6.68	6.63	6.63	6.63	7.29 <sup>a</sup>	6.63
39	6.66	6.49	6.49	6.69	6.49	6.25 <sup>a</sup>
40	7.26	6.96	7.09 <sup>a</sup>	6.96	6.96	6.96
41	6.64	6.60	6.25 <sup>a</sup>	6.49	6.60	6.49
42	6.06	6.36	6.36	6.48 <sup>a</sup>	6.36	5.61
43	5.36	5.83	5.27	5.27	5.27	5.27
44	6.75	6.69	6.48	6.69	6.69	6.48
45	5.37	5.27	5.64	5.64	5.64	5.60 <sup>a</sup>
46	4.40	6.10	6.10	6.10	6.10	6.10
47	7.05	6.98	6.98	6.47 <sup>a</sup>	6.47 <sup>a</sup>	6.98
48	6.21	6.20	6.20	6.20	6.78 <sup>a</sup>	6.20
49	6.52	6.21	6.21	6.21	6.21	6.21
50	7.16	6.44	6.44	6.44	6.44	6.44
51	4.80	5.18	5.18	5.18	5.18	5.18
52	6.14	6.14	6.14	6.14	6.14	6.14
53	5.48	5.64	5.64	5.64	5.76 <sup>a</sup>	5.64
54	5.92	6.23	6.23	6.23	6.23	6.48 <sup>a</sup>
55	5.75	5.56	5.56	5.56	5.56	5.56
56	5.44	5.62	5.62	5.62	5.62	5.62
57	7.17	6.66	6.66	7.09	6.74 <sup>a</sup>	7.09
58	4.92	5.38	5.38	5.38	5.38	5.38

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
59	4.54	4.96	4.96	5.58	4.96	4.74 <sup>a</sup>
60	5.96	5.61	5.61	5.61	5.61	6.23
61	6.82	6.47 <sup>a</sup>	6.47 <sup>a</sup>	6.71	6.71	6.71
62	7.55	7.40 <sup>a</sup>	7.40 <sup>a</sup>	6.99	7.40 <sup>a</sup>	6.99
63	7.92	7.45	7.45	7.45	7.29 <sup>a</sup>	7.45
64	7.68	7.21	7.21	7.29 <sup>a</sup>	7.21	7.29 <sup>a</sup>
65	7.64	7.48 <sup>a</sup>	7.64	7.64	7.64	7.48 <sup>a</sup>
66	7.62	7.29	7.48 <sup>a</sup>	7.48 <sup>a</sup>	7.66	7.29
67	6.31	6.32	7.48 <sup>a</sup>	6.32	7.48 <sup>a</sup>	6.32
68	7.38	7.48	7.48	6.82 <sup>a</sup>	6.82 <sup>a</sup>	7.48
69	7.70	7.49	7.29 <sup>a</sup>	7.21	7.49	7.21
70	7.55	6.99	6.99	7.53	6.99	7.53
71	7.35	7.09	7.09	7.09	7.09	7.09
72	7.82	7.29 <sup>a</sup>	7.20	7.20	7.45	7.20
73	7.70	7.70	7.70	7.70	7.70	7.70
74	7.46	7.10	6.82 <sup>a</sup>	7.10	7.10	7.10
75	6.66	6.82	6.82	6.82	6.82	6.82
76	6.02	6.45	6.45	6.45	6.45	6.45
77	7.13	7.09	7.09	6.74 <sup>a</sup>	7.09	7.03
78	7.72	7.20	7.49	7.49	7.20	7.49
79	6.56	6.73	6.73	6.73	6.73	6.73
80	7.57	7.19	7.19	7.19	7.48 <sup>a</sup>	7.19
81	7.54	7.53	7.53	6.96	7.53	6.96
82	7.75	7.85	7.85	7.85	7.85	7.85
83	7.31	7.19	7.19	7.19	7.19	7.19
84	6.85	7.09	7.09	6.61	6.61	6.47 <sup>a</sup>
85	7.38	6.82 <sup>a</sup>	6.77	7.48	7.48	6.82 <sup>a</sup>
86	6.75	6.48	6.78	6.48	6.48	6.78
87	7.42	6.77	7.10	6.77	6.77	6.77
88	5.89	6.15	6.15	6.15	6.15	6.15



Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
89	6.09	6.48 <sup>a</sup>	6.74	6.74	6.48 <sup>a</sup>	6.74
90	6.24	6.29	6.78 <sup>a</sup>	6.29	6.29	6.29
91	5.00	5.60 <sup>a</sup>	5.70	5.70	5.70	5.70
92	7.30	6.68	6.68	6.68	6.68	6.68
93	7.20	6.74 <sup>a</sup>	6.74 <sup>a</sup>	6.66	6.66	6.66
94	7.28	7.09 <sup>a</sup>	6.96	7.09 <sup>a</sup>	7.09 <sup>a</sup>	7.09 <sup>a</sup>
95	7.23	7.02	7.02	7.02	7.02	7.02
96	6.82	6.71	6.71	7.09	7.09	7.09
97	6.77	6.78	7.29 <sup>a</sup>	6.78	6.78	7.29 <sup>a</sup>

<sup>a</sup> - Test set compound

**Table A22:** Predictions of pIC<sub>50</sub> values of AChE inhibitors for TS-2 using PMF-PLS QSAR model (AID: 566585)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
1	8.06	7.86	7.86	7.88	8.30	7.59 <sup>a</sup>
2	8.60	8.79 <sup>a</sup>	8.79 <sup>a</sup>	8.62	9.39 <sup>a</sup>	8.52
3	8.34	8.33	8.33	8.33	8.57 <sup>a</sup>	8.33
4	8.96	9.01 <sup>a</sup>	9.01 <sup>a</sup>	9.26 <sup>a</sup>	9.26 <sup>a</sup>	9.26 <sup>a</sup>
5	8.54	8.40	8.40	8.32	8.32	8.32
6	7.19	7.56	7.56	7.55	7.21	7.55
7	6.82	7.00	7.00	6.98	6.98	6.98
8	7.80	7.85	7.85	7.81	7.39	7.96 <sup>a</sup>
9	8.31	8.33	8.33	8.37	8.33	8.37
10	8.29	8.11	8.11	8.15	8.37	8.15
11	7.05	7.42 <sup>a</sup>	7.42 <sup>a</sup>	7.21	7.03	7.21
12	8.89	8.88	8.88	9.39 <sup>a</sup>	8.62	8.62
13	6.46	7.19	7.00	7.09	6.98	6.66 <sup>a</sup>
14	9.10	8.65	8.65	8.67	8.67	8.67
15	7.28	7.75 <sup>a</sup>	7.75 <sup>a</sup>	7.34	7.34	7.34
16	7.48	7.47	7.47	7.50	7.50	7.50
17	7.85	7.86	7.86	7.96 <sup>a</sup>	7.81	7.81
18	7.37	7.57	7.57	7.61	7.61	7.61
19	9.31	9.04	9.04	8.96	8.96	8.96
20	6.90	7.03	7.03	7.03	7.03	7.03
21	7.60	7.67	7.47	8.07 <sup>a</sup>	7.50	7.50
22	7.40	7.54	7.54	7.49	7.49	7.49
23	8.06	7.67 <sup>a</sup>	7.79	7.59 <sup>a</sup>	7.88	7.88
24	6.09	7.00	6.86	6.98	6.81	6.98
25	5.59	6.03	6.03	6.06	6.06	6.06
26	8.14	8.37	8.37	8.30	7.97	8.30
27	6.47	6.63 <sup>a</sup>	7.19	6.66 <sup>a</sup>	7.09	7.09
28	9.32	8.67	8.67	8.77	8.77	8.77
29	7.64	7.52 <sup>a</sup>	7.67	7.57	7.57	7.57

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
30	7.92	7.54	7.54	7.51	7.51	7.51
31	8.15	7.98	7.98	7.97	7.59 <sup>a</sup>	7.97
32	6.05	6.86	6.63 <sup>a</sup>	6.81	6.66 <sup>a</sup>	6.81
33	8.37	7.91	7.91	7.95	7.95	7.95
34	8.55	8.52	8.52	8.52	8.52	9.39 <sup>a</sup>
35	7.30	7.33	7.33	7.30	7.30	7.30
36	6.78	6.92	6.92	7.65 <sup>a</sup>	7.65 <sup>a</sup>	6.66
37	7.28	7.28	7.28	7.25 <sup>a</sup>	7.25 <sup>a</sup>	7.25 <sup>a</sup>
38	7.00	6.99	6.99	7.03	7.85 <sup>a</sup>	7.03
39	9.24	9.25	9.25	9.26	9.26	9.26
40	6.52	6.86	6.86	6.76	6.76	7.65 <sup>a</sup>
41	8.44	7.87	7.87	7.93	7.93	7.93
42	8.24	7.89	7.89	7.90	8.15	7.90
43	6.66	7.17	7.17	7.12	7.12	6.76
44	6.49	7.31	7.31	7.22	7.22	7.22
45	6.92	7.11 <sup>a</sup>	7.11 <sup>a</sup>	7.01	7.01	7.01
46	9.02	8.65	8.65	8.76	8.76	8.76
47	7.52	7.33	7.52 <sup>a</sup>	7.32	7.32	8.07 <sup>a</sup>
48	7.19	7.20	7.20	7.85 <sup>a</sup>	7.55	7.85 <sup>a</sup>
49	8.11	7.79	7.79	7.81	7.81	7.81
50	7.59	7.47	7.33	7.50	8.07 <sup>a</sup>	7.32
51	7.70	7.40	7.40	7.39	7.96 <sup>a</sup>	7.39
52	6.68	6.94	6.94	6.99	6.99	6.99
53	8.01	7.92 <sup>a</sup>	7.92 <sup>a</sup>	8.10	8.10	8.10
54	8.44	8.46	8.46	8.49	8.49	8.49
55	7.26	7.44	7.44	7.51	7.51	7.51
56	8.08	7.88	7.88	7.86	7.86	7.86
57	9.48	8.34	8.34	8.42	8.42	8.42
58	8.03	7.79	7.67 <sup>a</sup>	7.75	7.75	7.75
59	6.71	6.75	6.75	6.66	6.66	7.12
60	8.17	8.46 <sup>a</sup>	8.46 <sup>a</sup>	8.57 <sup>a</sup>	7.90	8.57 <sup>a</sup>

<sup>a</sup> - Test set compound

**Table A23:** Predictions of pIC<sub>50</sub> values of 2-substituted Dipyridodiazeponone derivative inhibitors of HIV-1 RT for TS-3 using VC-PLS QSAR model (AID: 198247)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
1	5.43	6.46	6.43	6.47 <sup>a</sup>	6.53	6.43
2	6.60	6.63	6.67	6.67	6.67	6.67
3	6.74	6.75	6.79	6.74	6.74	6.79
4	7.15	7.02	7.06	6.83	6.83	7.06
5	7.22	7.00	6.95 <sup>a</sup>	6.92 <sup>a</sup>	6.92 <sup>a</sup>	7.07
6	6.41	6.77	6.77	6.80	6.80	6.77
7	5.85	6.47	6.45	6.53	6.22	6.45
8	7.00	6.66	6.66	6.60	6.60	6.94 <sup>a</sup>
9	6.82	6.64 <sup>a</sup>	6.74	6.72	6.74 <sup>a</sup>	6.53 <sup>a</sup>
10	6.09	6.75	6.71	6.76	6.76	6.71
11	6.64	6.78	6.75	6.83 <sup>a</sup>	6.83 <sup>a</sup>	6.75
12	7.70	7.06	6.97	6.90	7.29	6.97
13	6.41	6.87	6.91	6.85	6.85	6.91
14	6.89	6.79 <sup>a</sup>	6.79	6.74 <sup>a</sup>	7.09	7.08
15	6.96	6.84 <sup>a</sup>	6.88	6.89	6.89	6.88
16	6.66	6.28	6.34	6.31	6.31	6.34
17	6.42	6.79	6.69	6.76	6.58	6.69
18	7.00	6.51	6.52	6.55	6.55	6.52
19	7.00	6.69	6.73	6.72	6.72	6.66
20	6.85	6.88	6.89	6.95	6.88	6.89
21	7.30	7.03	7.02	7.01	7.01	7.02
22	7.52	7.44 <sup>a</sup>	7.45	7.29	7.25 <sup>a</sup>	6.96 <sup>a</sup>
23	7.15	7.02	7.13	7.12	7.12	6.95 <sup>a</sup>
24	7.40	6.79	6.79	6.73	6.73	6.79
25	6.96	6.97	6.97	7.08 <sup>a</sup>	7.08 <sup>a</sup>	6.97
26	7.70	7.75	7.82	7.25 <sup>a</sup>	6.90	7.82
27	6.92	6.92	6.90	7.02	7.02	7.52 <sup>a</sup>
28	7.40	7.23	7.18	7.17	7.17	7.18
29	6.33	6.98 <sup>a</sup>	6.50	6.60	6.92 <sup>a</sup>	6.50
30	6.51	6.67	6.66	6.92 <sup>a</sup>	6.76	6.85 <sup>a</sup>
31	7.05	7.00	6.94 <sup>a</sup>	7.08	7.08	6.73
32	6.82	6.73	6.68	6.61	6.61	6.81
33	6.40	6.55	6.66	6.58	6.60	6.66

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
34	6.52	6.77 <sup>a</sup>	6.85 <sup>a</sup>	6.66	6.66	6.66
35	7.40	7.59 <sup>a</sup>	7.34 <sup>a</sup>	7.34 <sup>a</sup>	7.38	7.34 <sup>a</sup>
36	7.52	7.36	7.36	7.25	7.25	7.36
37	7.70	7.26	7.27	7.17	7.17	7.27
38	8.00	7.61	7.54	7.46	7.46	7.54
39	7.15	7.27	7.22	7.33	7.33	6.83
40	6.85	6.94	6.92	6.88	6.72	6.92
41	7.70	7.84	7.79	7.82	7.82	7.79
42	6.82	6.88	6.81	6.78	6.78	6.74
43	7.30	7.00	7.04	7.00	7.00	7.04
44	5.96	6.14	6.25 <sup>a</sup>	6.22	6.43	6.25 <sup>a</sup>
45	5.92	6.26	6.40	6.28	6.28	6.40
46	6.74	6.63	6.70	6.71	6.71	6.68
47	7.74	7.38	7.42	6.85 <sup>a</sup>	7.49	7.42
48	7.40	7.34	7.32	7.38	7.34 <sup>a</sup>	7.32
49	7.05	7.03	7.03	7.11 <sup>a</sup>	7.24	7.03
50	7.05	6.87	6.88	6.82	6.82	6.88
51	6.64	6.90	7.00	6.91	6.91	7.00
52	6.72	6.81	6.53 <sup>a</sup>	6.85	6.85	6.70
53	6.00	6.69	6.70	6.82	6.82	6.70
54	7.52	7.33	6.96 <sup>a</sup>	7.23	7.23	7.45
55	8.00	7.99 <sup>a</sup>	7.75 <sup>a</sup>	7.90	7.90	7.75 <sup>a</sup>
56	7.70	7.55	7.54	7.42	7.42	7.54
57	7.70	7.54	7.55	7.49	6.85 <sup>a</sup>	7.55
58	6.00	6.28	6.28	6.43	6.57	6.28
59	6.00	6.37	6.34	6.57	6.47 <sup>a</sup>	6.34
60	7.05	7.49 <sup>a</sup>	7.30 <sup>a</sup>	7.24	6.97	7.30 <sup>a</sup>
61	7.15	6.92 <sup>a</sup>	7.07	6.88	6.88	7.22
62	7.15	6.85	6.83	6.96	6.96	7.13
63	7.05	7.13	7.11	7.13	7.13	7.11
64	7.40	7.19	7.13	7.23	7.23	7.13
65	7.10	7.05	7.13	7.04	7.04	7.13
66	6.92	6.95	7.08	6.96	6.96	6.90
67	6.89	7.01	7.52 <sup>a</sup>	7.09	6.95	6.79
68	7.10	7.09	6.92	6.97	7.11 <sup>a</sup>	6.92

<sup>a</sup> - Test set compound

**Table A24:** Predictions pIC<sub>50</sub> values of 2-pyridinone derivative inhibitors of HIV-1 RT for TS-4 using VC-PLS QSAR model (AID: 197804)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
1	5.02	5.79	5.90 <sup>a</sup>	5.79	5.79	5.79
2	6.52	6.34	6.17 <sup>a</sup>	6.34	6.17 <sup>a</sup>	6.34
3	7.38	7.05	6.50	6.50	6.50	7.05
4	5.71	5.99	6.47 <sup>a</sup>	5.99	5.99	5.99
5	6.47	6.39	6.45	6.39	6.49 <sup>a</sup>	6.45
6	7.24	6.53 <sup>a</sup>	7.37	7.37	7.37	6.53 <sup>a</sup>
7	6.47	6.15	6.15	6.15	6.15	6.15
8	3.52	3.66	4.06 <sup>a</sup>	3.66	3.66	3.66
9	4.96	5.52	5.15	5.15	5.90 <sup>a</sup>	5.90 <sup>a</sup>
10	4.50	4.76	4.76	4.76	4.76	4.76
11	5.76	5.98	5.80	5.98	6.47 <sup>a</sup>	5.80
12	5.94	6.35	6.35	6.24 <sup>a</sup>	6.35	6.24 <sup>a</sup>
13	6.22	6.05	6.05	6.05	6.27	6.05
14	6.37	6.49 <sup>a</sup>	6.05	6.49 <sup>a</sup>	6.05	6.05
15	6.72	6.16	6.16	6.16	6.16	6.16
16	5.55	5.91	5.36	6.57 <sup>a</sup>	5.36	5.91
17	5.98	5.60	5.60	5.97	5.60	5.97
18	6.52	6.03	6.34	6.03	6.34	6.03
19	5.95	5.97	5.97	6.35	5.97	6.35
20	7.37	7.37	7.05	7.05	7.05	7.37
21	6.94	6.81 <sup>a</sup>	6.81 <sup>a</sup>	7.04	6.74	6.74
22	6.95	7.04	7.04	6.81 <sup>a</sup>	7.04	7.04
23	7.64	7.72	6.53 <sup>a</sup>	7.72	6.53 <sup>a</sup>	7.72
24	5.98	5.77	5.77	5.60	6.24 <sup>a</sup>	5.60
25	6.01	6.24 <sup>a</sup>	6.24 <sup>a</sup>	5.77	5.77	5.77
26	6.99	6.44	6.44	6.44	6.44	6.44
27	6.18	6.27	6.27	6.27	5.56 <sup>a</sup>	6.27
28	7.24	6.70	6.70	6.70	6.24 <sup>a</sup>	6.70
29	4.17	4.92	4.92	4.92	4.92	4.92
30	4.61	5.10	5.10	5.10	5.10	5.10
31	6.36	6.03	6.03	6.03	6.03	6.03

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
32	6.74	6.89	6.89	6.89	6.89	6.89
33	7.15	6.81	6.81	6.24 <sup>a</sup>	6.48	6.81
34	6.98	6.24 <sup>a</sup>	6.24 <sup>a</sup>	7.07	7.07	6.24 <sup>a</sup>
35	7.04	7.07	7.07	6.81	6.81	7.07
36	5.90	6.89	6.89	6.89	6.89	6.89
37	6.33	5.56 <sup>a</sup>	5.56 <sup>a</sup>	5.56 <sup>a</sup>	5.72	5.72
38	6.96	6.24	6.24	6.24	6.24	6.24
39	7.19	6.48	6.48	6.48	6.70	6.48
40	6.82	6.55	6.55	6.55	6.55	6.55
41	6.59	6.17 <sup>a</sup>	5.87	6.17 <sup>a</sup>	5.87	6.17 <sup>a</sup>
42	5.78	6.47 <sup>a</sup>	5.98	6.47 <sup>a</sup>	5.98	5.98
43	5.90	6.05	6.05	6.05	6.05	6.05
44	3.84	4.06 <sup>a</sup>	3.66	4.06 <sup>a</sup>	4.75	4.75
45	3.98	4.75	4.75	4.75	4.06 <sup>a</sup>	4.46
46	4.49	4.46	4.46	4.46	4.46	4.06 <sup>a</sup>
47	4.54	4.95	4.95	4.95	4.95	4.95
48	4.65	5.03	5.03	5.03	5.03	5.03
49	4.82	5.90 <sup>a</sup>	5.52	5.52	5.52	5.52
50	5.00	5.15	5.79	5.90 <sup>a</sup>	5.15	5.15
51	5.36	5.36	6.57 <sup>a</sup>	5.36	6.57 <sup>a</sup>	5.36
52	5.57	5.52	5.91	5.91	5.91	6.57 <sup>a</sup>
53	5.60	6.57 <sup>a</sup>	5.52	5.52	5.52	5.52
54	5.63	5.89	5.89	5.89	5.89	5.89
55	5.68	5.95	5.95	5.95	5.95	5.95
56	5.72	5.80	5.99	5.80	5.80	6.47 <sup>a</sup>
57	5.96	5.78	5.78	5.78	5.78	5.78
58	6.28	6.31	6.31	6.31	6.05	5.56 <sup>a</sup>
59	6.30	5.72	5.72	5.72	6.31	6.31
60	6.55	5.87	6.03	5.87	6.03	5.87
61	5.27	5.43	5.43	5.43	5.43	5.43
62	3.52	5.63	5.63	5.63	5.63	5.63
63	7.26	6.74	6.74	6.74	6.74	6.74
64	6.92	6.74	6.74	6.74	6.81 <sup>a</sup>	6.81 <sup>a</sup>
65	6.34	6.67	6.67	6.67	6.67	6.67
66	6.48	6.45	6.49 <sup>a</sup>	6.45	6.45	6.49 <sup>a</sup>

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
67	6.46	6.05	6.39	6.05	6.39	6.39
68	5.12	5.93	5.93	5.93	5.93	5.93
69	7.52	6.50	7.72	6.53 <sup>a</sup>	7.72	6.50
70	6.68	6.35	6.35	6.35	6.35	6.35
71	7.72	6.46	6.46	6.46	6.46	6.46
72	7.70	7.38	7.38	7.38	7.38	7.38

<sup>a</sup> - Test set compound



**Table A25:** Predictions of pIC<sub>50</sub> values of cyclic urea derivative inhibitors of HIV-1 PR for TS-5 using VC-PLS QSAR model (AID: 160292)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
1	10.42	9.70	9.70	9.70	9.70	9.66 <sup>a</sup>
2	10.16	10.49	8.89 <sup>a</sup>	10.49	10.49	10.49
3	10.28	10.40	10.01	9.66 <sup>a</sup>	10.40	10.40
4	10.33	9.42	10.40	10.40	9.42	9.42
5	10.64	10.13 <sup>a</sup>	10.13 <sup>a</sup>	9.70	10.13 <sup>a</sup>	10.13 <sup>a</sup>
6	10.92	10.76	10.76	10.76	10.76	10.76
7	10.62	9.75	9.75	9.75	9.75	9.75
8	8.60	8.66	8.66	9.93 <sup>a</sup>	8.66	8.66
9	9.07	8.93 <sup>a</sup>	8.96	8.60	8.96	8.60
10	10.12	9.55	10.49	9.55	9.55	9.55
11	10.02	9.59	9.55	9.59	9.59	9.59
12	9.39	9.01	9.01	9.01	9.15	9.01
13	10.80	10.10	10.10	10.10	10.10	10.10
14	5.40	5.66 <sup>a</sup>	6.67	5.66 <sup>a</sup>	6.67	6.67
15	8.74	8.87	8.87	8.87	8.87	8.87
16	8.14	9.27	9.27	9.27	9.27	9.27
17	7.44	7.43	7.43	7.43	7.43	7.43
18	10.74	9.76	9.76	9.76	9.76	9.76
19	10.41	9.66 <sup>a</sup>	9.64	9.64	9.66 <sup>a</sup>	9.70
20	10.20	10.01	9.66 <sup>a</sup>	10.01	10.01	10.01
21	9.24	9.15	9.15	9.15	9.58 <sup>a</sup>	9.15
22	9.68	9.58 <sup>a</sup>	9.33	9.58 <sup>a</sup>	9.33	9.33
23	10.18	9.62	9.62	9.62	9.62	9.62
24	10.35	9.64	9.42	9.42	9.64	9.64
25	10.70	9.70	9.70	10.13 <sup>a</sup>	9.70	9.70
26	9.55	9.33	9.42	9.33	9.42	9.58 <sup>a</sup>
27	9.48	9.49	9.58 <sup>a</sup>	9.49	8.84	9.49
28	9.00	9.19	9.19	8.93 <sup>a</sup>	9.19	8.93 <sup>a</sup>

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
29	8.16	9.45	9.45	8.24	9.45	9.45
30	9.03	9.61	9.61	9.19	9.61	9.19
31	8.44	8.96	8.96	8.96	8.96	8.96
32	8.28	8.12 <sup>a</sup>	8.33	8.17	8.17	8.17
33	8.64	7.96	7.96	7.96	7.96	7.96
34	8.85	8.52	8.52	8.52	8.78 <sup>a</sup>	8.78 <sup>a</sup>
35	8.82	8.78 <sup>a</sup>	8.69	8.78 <sup>a</sup>	8.19	8.19
36	9.92	8.89 <sup>a</sup>	9.40	9.20	9.40	8.89 <sup>a</sup>
37	9.22	8.96	8.93 <sup>a</sup>	8.96	8.93 <sup>a</sup>	8.96
38	9.92	9.40	9.59	9.40	8.89 <sup>a</sup>	9.40
39	9.38	8.68	8.68	8.68	8.68	8.68
40	8.55	8.03	8.03	8.66	8.03	8.03
41	7.05	8.34	8.34	8.34	8.34	8.34
42	8.01	8.07	8.07	8.05	8.07	8.14 <sup>a</sup>
43	8.96	8.04	8.04	8.04	8.04	8.04
44	6.84	7.32	7.32	7.32	7.32	7.32
45	7.66	8.05	8.05	8.05	8.05	8.05
46	7.22	7.81	7.81	7.81	7.81	7.81
47	5.73	6.85	6.85	6.85	6.85	6.85
48	7.29	7.92	7.92	8.50 <sup>a</sup>	7.92	7.92
49	7.66	8.35	8.35	8.35	8.35	8.35
50	8.24	8.17	8.17	9.45	8.12 <sup>a</sup>	8.12 <sup>a</sup>
51	8.15	8.24	8.24	8.12 <sup>a</sup>	8.24	8.24
52	8.85	8.19	8.19	8.19	8.57	8.57
53	7.57	7.99	7.99	7.99	8.50 <sup>a</sup>	7.99
54	8.28	8.22	8.47	8.22	8.22	8.22
55	9.05	8.60	8.60	9.61	8.60	9.61
56	6.62	7.32	5.66 <sup>a</sup>	7.32	7.32	7.32
57	8.85	8.57	8.57	8.57	8.23	8.23
58	7.47	8.38	7.65	7.65	8.38	7.65

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
59	8.52	9.93 <sup>a</sup>	9.93 <sup>a</sup>	8.43	9.93 <sup>a</sup>	9.93 <sup>a</sup>
60	7.43	7.65	7.95	7.95	7.65	7.95
61	8.37	7.61	7.61	7.61	7.61	7.61
62	8.89	8.24	8.24	8.24	8.24	8.24
63	7.52	8.50 <sup>a</sup>	8.38	8.38	7.99	8.38
64	8.15	8.63	8.63	8.63	8.63	8.63
65	7.92	8.14 <sup>a</sup>	8.14 <sup>a</sup>	8.07	8.14 <sup>a</sup>	8.07
66	7.31	7.95	8.50 <sup>a</sup>	7.92	7.95	8.50 <sup>a</sup>
67	5.96	6.67	7.32	6.67	5.66 <sup>a</sup>	5.66 <sup>a</sup>
68	6.10	7.38	7.38	7.38	7.38	7.38
69	8.34	8.47	8.12 <sup>a</sup>	8.47	8.47	8.47
70	8.80	8.05	8.05	8.05	8.05	8.05
71	8.85	8.23	8.23	8.23	8.52	8.52
72	8.10	8.05	8.05	8.14 <sup>a</sup>	8.05	8.05
73	7.00	7.36	7.36	7.36	7.36	7.36
74	5.24	7.56	7.56	7.56	7.56	7.56
75	8.52	8.43	8.43	8.03	8.43	8.43
76	9.85	9.20	9.20	8.89 <sup>a</sup>	9.20	9.20
77	8.80	8.69	8.78 <sup>a</sup>	8.69	8.69	8.69
78	7.07	7.73	7.73	7.73	7.73	7.73
79	6.80	7.44	7.44	7.44	7.44	7.44
80	8.68	8.96	8.96	8.96	8.96	8.96
81	8.28	8.33	8.22	8.33	8.33	8.33
82	9.51	9.42	9.49	9.42	9.49	9.42
83	9.55	8.13	8.13	8.13	8.13	8.13
84	9.47	8.84	8.84	8.84	9.01	8.84

<sup>a</sup> - Test set compound

**Table A26:** Predictions of pIC<sub>50</sub> values of anti-malarial azilide derivatives for TS-6 using PMF-PLS QSAR model (AID: 579588)

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
1	6.80	6.90	6.92	6.90	7.06 <sup>a</sup>	6.92
2	4.89	5.77	5.84	5.89	5.93	5.84
3	5.29	5.78	5.42 <sup>a</sup>	5.82 <sup>a</sup>	5.97 <sup>a</sup>	5.79
4	5.52	6.02	6.01	6.07	6.56 <sup>a</sup>	5.92
5	5.80	5.92	5.86	5.89	5.93	5.86
6	6.21	6.92	6.20	6.27	6.29	6.20
7	6.32	6.55	6.56	6.63 <sup>a</sup>	6.76 <sup>a</sup>	6.56
8	6.66	6.62	6.67	6.64	6.62	6.67
9	6.99	6.60	6.53	6.55	6.56	6.53
10	7.02	6.90 <sup>a</sup>	6.85	6.91	6.91	6.85
11	7.37	7.10 <sup>a</sup>	7.08	7.07	7.14	7.08
12	6.53	6.59	6.60	6.58	6.82 <sup>a</sup>	6.82 <sup>a</sup>
13	6.01	6.19	6.18	6.21	6.28	6.18
14	6.26	6.44	6.55 <sup>a</sup>	6.36	6.42	6.55 <sup>a</sup>
15	6.60	6.54	6.52	6.50	6.48	6.52
16	6.79	6.83	6.84	6.71	6.80	7.02 <sup>a</sup>
17	6.85	6.90	6.91	6.88	6.91	6.71
18	6.86	6.82	6.83	6.82	6.84	6.91
19	6.99	6.64	6.69	6.54 <sup>a</sup>	6.69	6.69
20	7.11	7.42	7.16	7.53 <sup>a</sup>	7.41 <sup>a</sup>	7.16
21	7.17	6.91	6.93	6.86	6.85 <sup>a</sup>	6.93
22	6.69	6.64	6.62	6.39 <sup>a</sup>	6.61	6.69
23	6.10	6.13	6.14	6.27 <sup>a</sup>	6.23	6.14
24	6.39	6.40	6.42	6.52 <sup>a</sup>	6.44	6.42
25	6.58	6.68	6.72 <sup>a</sup>	6.68 <sup>a</sup>	6.80 <sup>a</sup>	6.72 <sup>a</sup>
26	6.68	6.30	6.69	6.64	6.69	6.86 <sup>a</sup>
27	6.74	6.69	6.73	6.74	6.75	6.73
28	6.90	6.81	6.78	6.80	6.82	6.72

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
29	7.07	6.97	6.81	6.84	6.84	6.81
30	7.13	6.88	6.86	6.77	6.92	6.86
31	5.93	6.34	6.41	6.38	6.65 <sup>a</sup>	6.41
32	6.91	6.57	6.78	6.78	6.75	6.78
33	6.03	6.21	6.20 <sup>a</sup>	6.21	6.38 <sup>a</sup>	6.20 <sup>a</sup>
34	6.31	6.42	6.39	6.47	6.44	6.39
35	6.39	6.52	6.48	6.53	6.58	6.48
36	6.54	6.60	6.58	6.67	6.68	6.58
37	6.63	6.71	6.68	6.69	6.65	6.68
38	6.74	6.72	6.75	6.75	6.80	6.75
39	6.76	6.78	6.70	6.82 <sup>a</sup>	6.74	6.70
40	6.84	6.51	6.48	6.50	6.55	6.96
41	6.86	6.86 <sup>a</sup>	6.85	6.90	6.96	6.85
42	6.86	6.73	6.79 <sup>a</sup>	6.67	6.73	6.83
43	6.86	6.82	6.82	6.78	6.80	6.82
44	6.90	6.77 <sup>a</sup>	6.14 <sup>a</sup>	6.83	6.83	6.78
45	6.93	6.58	6.58	6.27 <sup>a</sup>	6.49	6.58
46	6.99	6.90 <sup>a</sup>	7.24 <sup>a</sup>	6.97	7.02	6.94
47	7.07	6.84	7.02	6.97	7.01	6.52 <sup>a</sup>
48	7.26	6.79	7.04	7.07	7.07	7.04
49	6.10	6.44	6.43	6.36	6.47	6.43
50	6.29	6.18	6.13	6.14	6.15	6.13
51	6.30	6.48 <sup>a</sup>	6.43	6.48	6.54	6.43
52	6.36	6.57	6.56	6.57	6.59	6.56
53	6.39	6.40 <sup>a</sup>	6.22 <sup>a</sup>	6.37	6.41	6.49
54	6.42	6.55	6.61	6.57	6.64	6.61
55	6.63	6.25	6.53	6.61	6.56	6.53
56	6.66	6.70	6.72	6.70	6.71	6.72
57	6.86	6.27	6.89	6.87	6.91	6.89
58	6.93	6.68	6.63	6.62	6.72	6.63

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
59	4.91	5.39	5.42	5.46	5.57	5.42
60	5.48	5.80 <sup>a</sup>	5.79	5.82	5.83	6.01
61	6.39	6.60	6.49	6.47	6.49	6.22 <sup>a</sup>
62	6.40	6.97	6.48	6.47	6.47	6.48
63	6.48	6.40	6.38	5.73 <sup>a</sup>	6.36	6.38
64	6.85	6.66	6.71	6.66	6.65	6.48
65	6.94	6.69	6.67	6.70	6.74	6.67
66	6.97	6.98	6.94	6.97	6.98	7.24 <sup>a</sup>
67	7.14	6.95	6.90	6.96	6.96	6.90
68	7.26	7.01	6.85 <sup>a</sup>	7.05	7.00	7.22
69	7.02	6.69	6.71	6.60	6.69	6.46 <sup>a</sup>
70	6.36	6.54	6.55	6.58	6.60	6.55
71	6.37	6.10 <sup>a</sup>	6.02	6.09	6.09	6.02
72	6.62	6.81	6.77	6.83	6.84	6.77
73	6.72	6.78	6.86 <sup>a</sup>	6.81	6.78	6.62
74	6.78	6.89	7.02 <sup>a</sup>	6.88	6.88	6.84
75	6.88	6.85	6.87	6.88 <sup>a</sup>	6.87 <sup>a</sup>	6.87
76	6.97	6.93	6.93	6.87	6.75 <sup>a</sup>	6.93
77	7.11	7.02	6.52 <sup>a</sup>	6.95	6.98	7.02
78	7.26	6.99 <sup>a</sup>	7.05	7.00	7.09	7.05
79	4.89	6.20	6.33	6.24	6.40	6.33
80	5.70	5.73 <sup>a</sup>	5.75	5.76	5.77	5.42 <sup>a</sup>
81	6.09	6.52	6.57	6.46	6.59	6.57
82	6.24	6.51	6.56	6.58	6.84 <sup>a</sup>	6.56
83	6.45	6.38	6.43	6.42	6.43	6.43
84	6.49	6.64 <sup>a</sup>	6.82 <sup>a</sup>	6.65	6.72	6.60
85	6.74	6.77 <sup>a</sup>	6.77	6.79	6.85 <sup>a</sup>	6.77
86	6.76	6.76 <sup>a</sup>	6.66	6.67	6.70	6.66
87	6.82	6.92	6.96	7.01 <sup>a</sup>	6.88	6.79 <sup>a</sup>
88	6.92	6.67	6.72	6.66	6.74	6.14 <sup>a</sup>

Compound	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>				
		Model 1	Model 2	Model 3	Model 4	Model 5
89	7.00	6.73	6.46 <sup>a</sup>	6.74	6.77	6.71
90	7.04	7.19	7.01	7.00	7.00	7.01
91	7.19	6.98	6.49 <sup>a</sup>	6.89	6.85	6.49 <sup>a</sup>
92	7.22	6.83	6.80	6.61 <sup>a</sup>	6.81	6.80
93	7.27	6.95	7.22	7.04 <sup>a</sup>	7.21	6.85 <sup>a</sup>
94	7.16	6.98	6.97	7.02	7.04	6.97
95	7.25	7.01	7.00	6.98	6.62 <sup>a</sup>	7.00
96	7.50	6.80	6.76	6.68	6.72	6.76
97	7.29	6.87	6.84	6.79	6.81	6.84
98	5.69	5.96 <sup>a</sup>	5.92	5.96	6.03	5.75

<sup>a</sup> - Test set compound

## Abstract

---

<b>Name of the Student:</b> Pushkar D. Kunde	<b>Registration no:</b> 10CC11J26031
<b>Faculty of Study:</b> Chemical Sciences	<b>Year of Submission:</b> 2019
<b>AcSIR academic center/CSIR Lab:</b> CSIR- National Chemical Laboratory	<b>Names of Supervisors:</b> Dr. Sanjay P. Kamble, & Dr. V. Ravi Kumar
<b>Title of the thesis:</b> Studies in QSAR modelling for selection of potential inhibitors for drug discovery	

---

A Quantitative structure activity relationship (QSAR) model describes the biological activity of a molecule as a function of its structure. Molecular descriptors quantitatively represent structural features of ligand molecules in the QSAR model. Molecular descriptors can vary from the simplest to the most complex molecular properties. In the current work we aim towards developing novel molecular descriptors and methods for QSAR modelling. Six target systems involved in diseases like, cancer, neurodegenerative disorders, HIV-AIDS and malaria were selected for these studies.

In chapter 2, we work towards developing 2D image based descriptors from optimal 3D structures of compounds. These descriptors were created using Dijkstra's algorithm and multi-dimensional scaling to retain the inter-atomic shortest path distances in 3D space and their partial charges. The 2D descriptors were then regressed with the biological activity values of the compounds using principal component analysis and support vector regression. These QSAR models were observed to be computationally intensive.

In chapter 3, we introduce the concept of 3D pseudo-molecular field (PMF) which depends on the intrinsic properties of the atoms, namely, electron affinity and electronegativity unlike the traditional electrostatic field which is calculated using the partial atomic charges. We further develop PMF-PLS methodology using partial least squares (PLS) and Procrustes transformation to regress these descriptors against the biological activity of the molecules. The performance of resulting PMF-PLS QSAR models were observed to be comparable to that of the reported QSAR models computationally lighter as compared to the models in chapter 2.

In chapter 4, we devise a second regression methodology, namely, Varying Component PLS (VC-PLS), using the SIMPLS variant of PLS method, for QSAR modelling. We also used VC-PLS QSAR models and PMF-PLS QSAR models from chapter 3 to screen natural compounds similar to the molecules in the target systems. The results of screening the compounds with were consistent with both the QSAR models. Finally, docking studies were performed with the selected natural compounds to supplement the screening results.



## Publications

### Research article:

**Title: On the use of electronegativity and electron affinity based pseudo-molecular field descriptors in developing correlations for quantitative structure-activity relationship modeling of drug activities**

Pushkar D. Kunde, Sudha Ramkumar, Sanjay P. Kamble, Ameeta Ravikumar, Bhaskar D. Kulkarni and V. Ravi Kumar

**Accepted:** Chemical Biology & Drug Design, article in press

**Article DOI:** 10.1111/cbdd.13895

### Poster:

**Title: Employing MDS and Multivariate image analysis for developing QSAR models**

Pushkar D. Kunde, Ameeta Ravikumar, Bhaskar D. Kulkarni and V. Ravi Kumar

Presented at Symposium on “Accelerating Biology – 2016: Decoding the deluge” organized by Center for Development of Advanced Computing (CDAC), Pune, at YASHADA, Pune on January 19<sup>th</sup> to 21<sup>st</sup>, 2016.

Abstract: Quantitative Structure Activity Relationship (QSAR) modeling is an integral part of ligand based drug discovery. Used for virtual screening of the compounds that are proposed to be the potential lead molecules, QSAR modeling reduces the cost of further studies by selecting only those molecules which show promising bio-activities. The descriptors used in the QSAR modeling are the physico-chemical properties of these molecules. These properties may not necessarily reflect the structural aspects of the compounds that are responsible for the interaction of ligand with the target protein. This problem is addressed by 3D-QSAR models which use 3D molecular field as the descriptors of the molecules for the model. However, they require exhaustive calculations. Recently, 2D-QSAR models that use image based descriptors of the compounds were reported. Here, we have studied a series of 4-phenylpyrrolocarbazol derivatives from literature that are inhibitors of tyrosine kinase Wee1, an important target in cancer treatment, and developed an image based 2D-QSAR model. We used Multi-dimensional Scaling (MDS) to generate 2D images for every molecule using its 3D-structures along with the partial atomic charges. These images were further analyzed using Multivariate Image Analysis (MIA). Principal Component Analysis (PCA) was used to reduce the dimensionality of the data in the images. The Principal Components were regressed against the pIC<sub>50</sub> values using Support Vector Machine (SVM) regression. The leave-one-out cross-validation and the external validation yielded  $r^2=0.84$  and  $r^2_{test}=0.83$  respectively and the root-mean-squared error values as 0.49 and 0.44 respectively, proving robustness of the model.

## RESEARCH LETTER

# On the use of electronegativity and electron affinity based pseudo-molecular field descriptors in developing correlations for quantitative structure-activity relationship modeling of drug activities

Pushkar D. Kunde<sup>1,2</sup> | Sudha Ramkumar<sup>3</sup> | Sanjay P. Kamble<sup>1,2</sup> | Ameeta Ravikumar<sup>4</sup> | Bhaskar D. Kulkarni<sup>1,2</sup> | V. Ravi Kumar<sup>1,2</sup> 

<sup>1</sup>Chemical Engineering and Process Development Division, CSIR-National Chemical Laboratory (CSIR-NCL), Pune, India

<sup>2</sup>Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India

<sup>3</sup>Organic Chemistry Division, CSIR-National Chemical Laboratory (CSIR-NCL), Pune, India

<sup>4</sup>Institute of Bioinformatics and Biotechnology (IBB), Savitribai Phule Pune University, Pune, India

## Correspondence

V. Ravi Kumar, Chemical Engineering and Process Development Division, CSIR-National Chemical Laboratory (CSIR-NCL), Pune, India.

Email: ravikumar2609@gmail.com

## Abstract

For quantitative structure-activity relationship (QSAR) modeling in ligand-based drug discovery programs, pseudo-molecular field (PMF) descriptors using intrinsic atomic properties, namely, electronegativity and electron affinity are studied. In combination with partial least squares analysis and Procrustes transformation, these PMF descriptors were employed successfully to develop correlations that predict the activities of target protein inhibitors involved in various diseases (cancer, neurodegenerative disorders, HIV, and malaria). The results show that the present QSAR approach is competitive to existing QSAR models. In order to demonstrate the use of this algorithm, we present results of screening naturally occurring molecules with unknown bioactivities. The pIC<sub>50</sub> predictions can screen molecules that have desirable activity before assessment by docking studies.

## KEYWORDS

drug discovery, electron affinity, electronegativity, molecular field descriptors, partial least squares, QSAR

## 1 | INTRODUCTION

Ligand-based drug design is an approach for drug discovery where structures of molecules having known biological activities are used to design new and better ones. The approach studies and obtains quantitative structure-activity relationship(s) (QSAR) that exist between the structural features of the compounds and their activities (Cumming et al., 2013; Gramatica, 2020). QSAR modeling uses three components, namely, (a) reliable biological response data to be modeled, (b) suitable descriptors of the molecules and, (c) QSAR modeling and its validation using suitable mathematical/statistical techniques (Gramatica, 2020).

QSAR models develop correlations with a variety of activities such as inhibition concentration (IC<sub>50</sub>),

pharmacokinetic properties (absorption, distribution, metabolism, and excretion; Madden, 2010), toxicological properties (carcinogenicity, mutagenicity, acute and chronic toxicity; Cronin, 2010), etc. Experimentally obtained biological response data of molecules is also publicly available in the PubChemBioAssay database (Wang et al., 2014). Descriptors provide information to develop the QSAR from molecular calculations or from experimental data (Todeschini & Consonni, 2008). Molecular descriptors used can vary from simple molecular properties, (molecular weight, dissociation constant, partition coefficients, solubility, etc.; Lindgren et al., 1996; Simeon et al., 2019), molecular fingerprints (Ballabio et al., 2019; Roy & Das, 2014), to quantum chemical properties (molecular orbital energies; Oyewole et al., 2020), to ones

as complex as 3D molecular field values (Gasteiger & Eds, 2003) at spatial points around the molecule which may even be supplemented with information about molecular conformations (Damale et al., 2014; Lill, 2007). Among the QSAR modeling methodologies, comparative molecular field analysis (CoMFA) uses 3D molecular field values as descriptors (Cramer et al., 1988; Dasoondi et al., 2008; Divakar & Hariharan, 2015; Matta & Arabi, 2011; Nidhi, & Siddiqi, 2013). These descriptor values are calculated using partial atomic charges of atoms obtained from the energy minimized and aligned 3D structures of the molecules (Gasteiger & Marsili, 1980). A grid is suitably defined around the molecules and electrostatic field values are calculated using Coulomb potential function (Cramer et al., 1988). Thus, a 3D array of field values is obtained for every molecule and regression models are developed to correlate these molecular descriptors with the activities of the molecules.

Notwithstanding, there is a need to explore other novel descriptors employing atomic properties that could provide simple correlations (Totrov, 2007). Toward this aim, we study here the use of intrinsic properties of individual atoms, namely, their electronegativity and electron affinity values, to develop a 3D pseudo-molecular field (PMF) as molecular descriptors along with partial least squares (PLS) regression (de Jong, 1993; Garthwaite, 1994; Geladi & Kowalski, 1986) and Procrustes transformation (Kendall, 1989) for QSAR modeling. We choose the statistical PLS regression method because of its performance compared to machine learning methods (e.g., artificial neural networks, random forests, support vector regression, etc.) especially when the number of observations is far less than the number of variables (i.e., descriptors; Breiman, 2001; Mendez et al., 2019; Panagou et al., 2011; Schwartz et al., 2009) as is the case in the present 3D-QSAR study. In fact, the detailed study by Mendez et al., 2019, for such a situation shows that across ten publically available high dimensional data sets non-linear machine learning methods showed no general improvement in model predictability over linear ones. Using the principle of Occam's razor, then the simpler model is usually a better choice. Moreover, non-linear methods tend to over-train a complex model and inaccuracies arise because the model is only as good as the data that is used to train it. Statistical methods again perform better under such situations than machine learning algorithms (Mendez et al., 2019).

In the present study, we develop and analyze PMF-PLS QSAR models using inhibitor molecules for six target systems (abbreviated as TS-1 to TS-6) involved in the treatment of various diseases, namely cancer (Palmer et al., 2006), neurodegenerative disorders (Queiroz et al., 2011), HIV (Debnath, 1999; Proudfoot et al., 1995; Wai et al., 1993), and malaria (Hutinec et al., 2011). We also study the important application of PMF-PLS QSAR models in drug assessment

programs by screening new molecules whose biological activities are not known.

## 2 | PMF-PLS QSAR METHODOLOGY

PubChemBioAssay (Wang et al., 2014) is a large compendium of chemical compounds with tested values of their biological activities. We used PubChemBioAssay to identify and download structures of chemical inhibitors with similar scaffolds for the target systems TS-1 to TS-6 (Table 1). For brevity, the structures of all the inhibitor compounds chosen for TS-1 to TS-6, respectively, are provided in Supporting Information S1, Tables S1–S6.

The molecular structures downloaded from PubChemBioAssay were imported to Schrödinger software suite through Maestro module (version 9.2; Schrödinger & LLC, 2011b). These 2D structures were converted to their equivalent energy minimized 3D form using the LigPrep module (version 2.5; Schrödinger & LLC, 2011a). Ligand alignment routine in Maestro module was used to obtain aligned 3D structures with superimposed scaffolds. The aligned molecular structures were exported as .pdb (protein data bank) files. All further coding and calculations of the PMF-PLS methodology were carried out in MATLAB® (Version R2010b; MATLAB, 2010). The .pdb files were imported to MATLAB so that the aligned molecules are positioned in a common 3D mesh grid with finite size and intra-grid spacing. In general, it was observed that using a distance of 1 Å between two adjacent (vertical or horizontal) grid points was adequate for the target systems studied. It may be noted that the above preprocessing of molecular structures is similar to that employed in the traditional CoMFA methodology (Cramer et al., 1988).

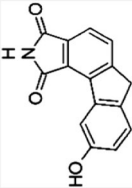
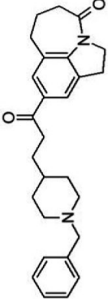
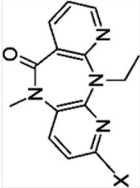
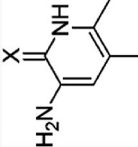
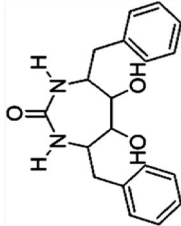
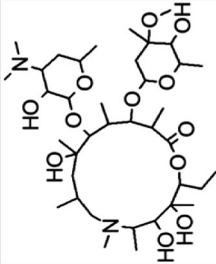
In CoMFA the partial atomic charges are calculated (Gasteiger & Marsili, 1980; Mulliken, 1934) for all the atoms in every molecule. These atomic charge values are used to obtain the electrostatic field values at the grid points using the Coulomb potential function (Kubinyi, 1997), namely,

$$E_{c_{j,k,l}} = \sum_{i=1}^{n_a} \frac{q(i)q_p}{Dd(i)} \quad (1)$$

where  $E_{c_{j,k,l}}$  is the Coulomb interaction energy at grid point  $(j,k,l)$ ;  $q(i)$  is the partial charge of the  $i$ th atom of the molecule;  $q_p$  is the charge of the probe atom;  $D$  is the dielectric constant;  $d(i)$  is the distance between the  $i$ th atom of the molecule and the grid point  $(j,k,l)$ ; and  $n_a$  the total number of atoms in an inhibitor molecule. For the probe atom, the chosen charge ( $q_p$ ) is kept constant in the calculations.

The partial charges on the atoms in the molecules are known to be dependent on electron affinity and the orbital

**TABLE 1** Inhibitor compounds and their scaffolds for target systems TS-1 to TS-6

TS no.	Target	Inhibitor compounds	Scaffolds <sup>a</sup>	AID <sup>b</sup>	References
1	Wee1 (Cancer)	4-Phenyl pyrrolocarbazoles		268,838	Palmer et al. (2006)
2	AChE (neurodegenerative disorders)	Benzylpiperidines		566,585	Queiroz et al. (2011)
3	HIV-1 RT	2-Substituted Dipyridodiazepones		198,247	Proudfoot et al. (1995)
4	HIV-1 RT	2-Pyridinones		197,804	Wat et al. (1993)
5	HIV-1 PR	Cyclic Ureas		160,292	Debnath (1999)
6	Malaria	Azilides		579,588	Hutinec et al. (2011)

Abbreviation: TS, target system.

<sup>a</sup>Scaffolds are the basic common structures of inhibitor compounds.<sup>b</sup>Assay identification number (AID) from PubChemBioAssay.

electronegativity of the atoms (Gasteiger & Marsili, 1980; Mulliken, 1934). Based on this rationale, we formed a correlation equation, similar to Equation 1, using the electronegativity and electron affinity of the atoms to calculate the PMF, namely,

$$\gamma_{j,k,l} = \sum_{i=1}^{n_a} \left( \frac{\sigma E_a(i) \chi(i)}{d(i)} \right) \quad (2)$$

where  $\gamma_{j,k,l}$  is the value of the PMF at the grid point  $(j,k,l)$ ;  $n_a$  is the total number of atoms;  $E_a(i)$  is the electron affinity of the  $i$ th atom;  $\chi(i)$  is the electronegativity of the  $i$ th atom;  $d(i)$  is the distance of the grid point  $(j,k,l)$  from the  $i$ th atom of the molecule; and  $\sigma$  is a suitably chosen scaling factor ( $\sigma = 0.1$ ). We formed the correlation equation (Equation 2) for calculation of PMF by substituting in the Coulomb potential function (Equation 1), the atomic charge  $q(i)$  of  $i$ th atom with the product of electron affinity,  $E_a(i)$ , and electronegativity,  $\chi(i)$ , of that atom (Supporting Information S1, Table S7) and the constants  $q_p$  and  $D$  with the scaling factor,  $\sigma$ . It may be pointed out that Equation 2 albeit a correlation was particularly found to capture and build up accurate QSAR models using these atomic properties.

The 3D grid of PMF values for every molecule was transformed into a 1D descriptor array. The 1D descriptor array sizes for the target systems studied are reported in Table 2. The 1D descriptor arrays for molecules in a target system were stacked to form the  $X$  matrix of size  $(n, m)$  (where  $n$  is the number of inhibitor compounds and  $m$  is the total number of grid points for a target system) for the purpose of regressing with matrix  $Y$ , that is, the biological activity values ( $\text{pIC}_{50}$ ), matrix of size  $(n, 1)$ .

Partial least squares (Garthwaite, 1994; Geladi & Kowalski, 1986) is a widely used regression method that aims at capturing relationships between the dependent variable  $Y$  and the independent variables  $X$  by projecting the  $X$  and  $Y$  data to a latent subspace of lower dimensions, (i.e.,  $a < m$ ), while maximizing the covariance between them. PLS regression is carried out by the decomposition of  $X$  and  $Y$  as shown in Equations 3 and 4, respectively, to obtain matrices  $T$ ,  $P$ ,  $U$ , and  $Q$ , such that,

$$X = TP' + E = \sum_{i=1}^{i=a} t_i p_i' + E \quad (3)$$

$$Y = UQ' + F = \sum_{i=1}^{i=a} u_i q_i' + F \quad (4)$$

In Equation 3 the score matrix,  $T$  of matrix size  $(n, a)$ , is composed of  $a$  latent vector while the loading matrix,  $P$  is of matrix size  $(m, a)$  with  $E$  the residuals in the decomposition of  $X$ . Similarly, in Equation 4 the scores  $U$  is of matrix size  $(n, a)$  and the loadings  $Q$  is of matrix size  $(1, a)$  for the decomposition of  $Y$  with  $F$  the corresponding residuals. The magnitudes of the residuals  $E$  and  $F$  depend on the number of latent components chosen and for a proper choice of  $a < m$ , dimensionality reduction is possible with minimization of residuals.

A variant of PLS, namely, SIMPLS method (de Jong, 1993) was used here for PLS regression. It offers significant advantages as it performs the calculations of all the scores and loadings using the original  $X$  and  $Y$  matrices in every iterative step unlike the conventional NIPALS algorithm (Garthwaite, 1994; Geladi & Kowalski, 1986) for PLS which uses deflated matrices obtained during each iteration. Since the loadings and scores are calculated from the original data ( $X$  and  $Y$ ) the regression coefficients  $B$ , can also be obtained for the original data. Thus, the final regression model can be given as,

$$\hat{Y} = XB \quad (5)$$

where  $B$  is a vector of regression coefficients of matrix size  $(m, 1)$ . In the proposed PMF-PLS methodology, we used the MATLAB function “plsregress,” which employs the SIMPLS method.

PMF-PLS algorithm comprises of three parts. In the first part, we randomly select a validation set ( $X_{\text{val}}, Y_{\text{val}}$ ) from the molecules of the target system to validate the model. From the remaining molecules, we then identify a suitable training set ( $X_{\text{train}}, Y_{\text{train}}$ ) and a test set ( $X_{\text{test}}, Y_{\text{test}}$ ) for model development. The training set ( $X_{\text{train}}, Y_{\text{train}}$ ) and test set ( $X_{\text{test}}, Y_{\text{test}}$ ) were used in the second part of the algorithm to iteratively obtain a set of regression coefficients  $B_{\text{avg}}$  in the PMF-PLS QSAR model. In the third part, the model obtained was validated by

**TABLE 2** 3D Mesh grid and 1D descriptor sizes for target systems TS-1 to TS-6

TS no.	AID <sup>a</sup>	Inhibitor compounds	Number of compounds ( $n$ )	3D mesh grid	1D descriptor size ( $m$ )
1	268838	4-Phenyl pyrrolocarbazoles	97	29 × 32 × 23	21,344
2	566585	Benzylpiperidine derivatives	60	36 × 29 × 25	26,100
3	198247	2-Substituted Dipyridodiazepinones	68	27 × 26 × 23	16,146
4	197804	2-Pyridinone Derivatives	72	26 × 27 × 28	19,656
5	160292	Cyclic urea derivatives	84	32 × 28 × 24	21,504
6	579588	Azilide derivatives	98	31 × 28 × 31	26,908

<sup>a</sup>Assay identification number (AID) from PubChemBioAssay.

predicting the  $\hat{Y}_{\text{val}}$  values for the validation set. It should be noted that the validation set ( $X_{\text{val}}, Y_{\text{val}}$ ) is an initially chosen set of compounds not present in the training or test sets and therefore not used the parts 1 and 2 of PMF-PLS model development. The details of the algorithm are described below and schematically outlined in Supporting Information S1, Figure S1.

For the first part, initially about 15% of molecules in a target system were assigned randomly to a validation set ( $X_{\text{val}}, Y_{\text{val}}$ ). The remaining molecules were allotted randomly to either a temporary training set ( $X_{\text{train,temp}}, Y_{\text{train,temp}}$ ) (70%) or a test set ( $X_{\text{test,temp}}, Y_{\text{test,temp}}$ ) (15%) using the “randperm” function in MATLAB. The training set ( $X_{\text{train,temp}}, Y_{\text{train,temp}}$ ) was used to perform a PLS regression and the model predicted pIC<sub>50</sub> values  $\hat{Y}_{\text{train,temp}}$  and  $\hat{Y}_{\text{test,temp}}$  were compared with their actual values  $Y_{\text{train,temp}}$  and  $Y_{\text{test,temp}}$  by evaluating the root mean squared errors (RMSE; Roy et al., 2016),  $e_{\text{train,temp}}$  and  $e_{\text{test,temp}}$ , respectively, using the general equation,

$$\text{RMSE} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n_{\text{mol}}}} \quad (6)$$

where  $n_{\text{mol}}$  is the number of molecules in the chosen set,  $Y$  is the experimental pIC<sub>50</sub> values for the set and  $\hat{Y}$  are the predicted pIC<sub>50</sub> values. This procedure was repeated a number of times for different combinations of training and test sets randomly chosen. Of these iterations the temporary training and test sets which realized the minimum RMSE were chosen as the training ( $X_{\text{train}}, Y_{\text{train}}$ ) and test ( $X_{\text{test}}, Y_{\text{test}}$ ) sets for further calculations.

We carried out the second part of the PMF-PLS algorithm where alterations were done to ( $X_{\text{train}}, Y_{\text{train}}$ ) by removing one molecule at-a-time from ( $X_{\text{train}}, Y_{\text{train}}$ ) and adding it to ( $X_{\text{test}}, Y_{\text{test}}$ ) to obtain ( $X_{\text{train,mod}}, Y_{\text{train,mod}}$ ) and ( $X_{\text{test,mod}}, Y_{\text{test,mod}}$ ) which were then subjected to PLS regression. The altered sets ( $X_{\text{train,mod}}, Y_{\text{train,mod}}$ ) for which the RMSE of prediction was greater than RMSE of prediction for ( $X_{\text{train}}, Y_{\text{train}}$ ) by 15% of the pIC<sub>50</sub> value range were observed to have significantly different PLS scores ( $T_i$ ) when compared to the scores  $T_{\text{train}}$  obtained from ( $X_{\text{train}}, Y_{\text{train}}$ ). To take care of this situation, scores ( $T_i$ ) and loadings ( $P_i$ ) obtained from ( $X_{\text{train,mod}}, Y_{\text{train,mod}}$ ) were subjected to Procrustes transformation to obtain scores  $T_{p,i}$  and loadings  $P_{p,i}$ , respectively. The respective Euclidean distances to  $T_i$  and  $P_i$  (i.e., before the Procrustes transformation) and to  $T_{p,i}$  and  $P_{p,i}$  (i.e., after the Procrustes transformation) from  $T_{\text{train}}$  and  $P_{\text{train}}$  were calculated. It was observed that the distances of both scores and loadings were reduced, respectively, after Procrustes transformation as illustrated in Figure 1. Thus, Procrustes transformation brings the scores and loadings for altered training sets to align closer with  $T_{\text{train}}$  and  $P_{\text{train}}$ . The corresponding  $X$  values were obtained from Procrustes transformed  $T_{p,i}$  and  $P_{p,i}$ . It was observed that regression coefficients ( $B_i$ ) obtained using transformed  $X$

values performed predictions with reduced error when compared to that before the Procrustes transformation. As seen in Figure 1 removing compound number 22 from ( $X_{\text{train}}, Y_{\text{train}}$ ) for TS-1 (Table S1), resulted in a high RMSE = 2.05 which reduced to 1.31 on Procrustes transformation and therefore it brings about considerable prediction improvement.

All the sets of regression coefficients obtained by iterations in the second part of the algorithm were averaged to obtain  $B_{\text{avg}}$  and the final PMF-PLS QSAR model,

$$\hat{Y} = XB_{\text{avg}} \quad (7)$$

Cross-validation and external validation studies of the above regression model was carried out in the third part of the algorithm. For this the predicted pIC<sub>50</sub> values  $\hat{Y}_{\text{train}}$ ,  $\hat{Y}_{\text{test}}$ , and  $\hat{Y}_{\text{val}}$  were determined using  $X_{\text{train}}$ ,  $X_{\text{test}}$ , and  $X_{\text{val}}$ , respectively, in Equation 7. Cross-validation of PMF-PLS QSAR model was carried out by performing the predictions for the molecules used for building the model, that is, using the values  $\hat{Y}_{\text{train}}$  and  $\hat{Y}_{\text{test}}$  taken together. RMSE of cross-validation (RMSECV) was calculated using  $\hat{Y}_{\text{train}}$ ,  $\hat{Y}_{\text{test}}$ ,  $Y_{\text{train}}$ , and  $Y_{\text{test}}$  in Equation 6. Similarly, the coefficient of determination,  $R_{\text{cv}}^2$ , for cross-validation was calculated using the formula (Li et al., 2017; Roy et al., 2016),

$$R_{\text{cv}}^2 = 1 - \frac{\sum (Y_{\text{obs}} - Y_{\text{pred}})^2}{\sum (Y_{\text{obs}} - \bar{Y})^2} \quad (8)$$

where  $Y_{\text{obs}}$  are the observed pIC<sub>50</sub> values ( $Y_{\text{train}}$  and  $Y_{\text{test}}$ ),  $Y_{\text{pred}}$  the corresponding predicted pIC<sub>50</sub> values ( $\hat{Y}_{\text{train}}$  and  $\hat{Y}_{\text{test}}$ ) and  $\bar{Y}$  the mean of observed pIC<sub>50</sub> values of the training set ( $Y_{\text{train}}$ ).

For external validation of the model  $\hat{Y}_{\text{val}}$  values were compared with the known  $Y_{\text{val}}$  to obtain the RMSE of prediction (RMSEP). Coefficient of determination for validation set,  $Q_{\text{ext(F1)}}^2$  was obtained as (Li et al., 2017; Roy et al., 2016),

$$Q_{\text{ext(F1)}}^2 = 1 - \frac{\sum (Y_{\text{val}} - \hat{Y}_{\text{val}})^2}{\sum (Y_{\text{val}} - \bar{Y})^2} \quad (9)$$

It may be clarified that training set ( $X_{\text{train}}, Y_{\text{train}}$ ) and test set ( $X_{\text{test}}, Y_{\text{test}}$ ) compounds were used to build the QSAR model (Equation 9), whereas, the validation set ( $X_{\text{val}}, Y_{\text{val}}$ ) compounds used for external validation of the model were not used in any of the steps involved in building the QSAR model.

### 3 | RESULTS AND DISCUSSION

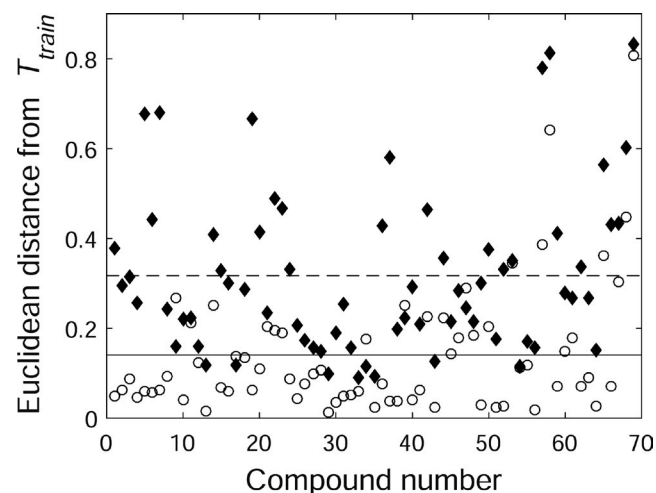
Different charge-based descriptors are known and studied (Todeschini & Consonni, 2008). These include atomic charge descriptors, local dipole moment, charge-based topological indices, charge-weighted autocorrelation

descriptors, charge-based measures of solvent accessible surface area (PEOE-VSA; Estrada, 1995; Labute, 2000; Stanton & Jurs, 1990; Todeschini & Consonni, 2008), etc. These use different methodologies to calculate the descriptor values from the atomic charges for a molecule. These molecular descriptors are regressed with the response data to obtain the QSAR model. On the other hand, molecular field descriptors as employed in the present study (PMF) and CoMFA (Cramer et al., 1988) additionally take into consideration the 3D conformation of the molecules along with atomic properties to obtain space-dependent descriptors. The advantage of using molecular field descriptors is that the QSAR model captures information about favorable and/or unfavorable regions in 3D space for the activity of different ligands (Kubinyi, 1997).

$$\begin{aligned} \text{Good model: } & \text{MAE} \leq 0.1 \times \text{training set range AND } \text{MAE} + 2\text{SD} \leq 0.2 \times \text{training set range} \\ \text{Bad model: } & \text{MAE} > 0.15 \times \text{training set range OR } \text{MAE} + 2\text{SD} > 0.25 \times \text{training set range} \end{aligned} \quad (10)$$

Such information cannot be inferred from the QSAR models developed using the other charge-based descriptors.

Figure 2 shows the diagonal plots of  $Y_{\text{train}}$  versus  $\hat{Y}_{\text{train}}$  and  $Y_{\text{test}}$  versus  $\hat{Y}_{\text{test}}$  for TS-1 to TS-6. The PMF-PLS model fitting statistics for TS-1 to TS-6 are shown in Table 3. It may be observed that the results presented in the diagonal plots (Figure 2) and the calculated model fitting statistics for cross-validation (namely,  $R_{\text{cv}}^2$  and RMSECV) lie in an acceptable range (Table 3). External validation of the PMF-PLS QSAR model using the validation set was then carried out by predicting  $\hat{Y}_{\text{val}}$  (Figure 3). The corresponding model fitting statistics (namely,  $Q_{\text{ext}(F1)}^2$  and RMSEP) were calculated for TS-1



**FIGURE 1** Effects of Procrustes transformation on the PLS scores of compounds for TS-1 on removing compound number 22 (Table S1) from  $(X_{\text{train}}, Y_{\text{train}})$  and adding to  $(X_{\text{test}}, Y_{\text{test}})$ . Shown are the Euclidean distances (◆) from  $T_{\text{train}}$  to  $T_i$  and (○) from  $T_{\text{train}}$  to  $T_{p,i}$ . The horizontal (---) and (--) lines denote the calculated mean distance for the compounds in training set before and after Procrustes transformation, respectively

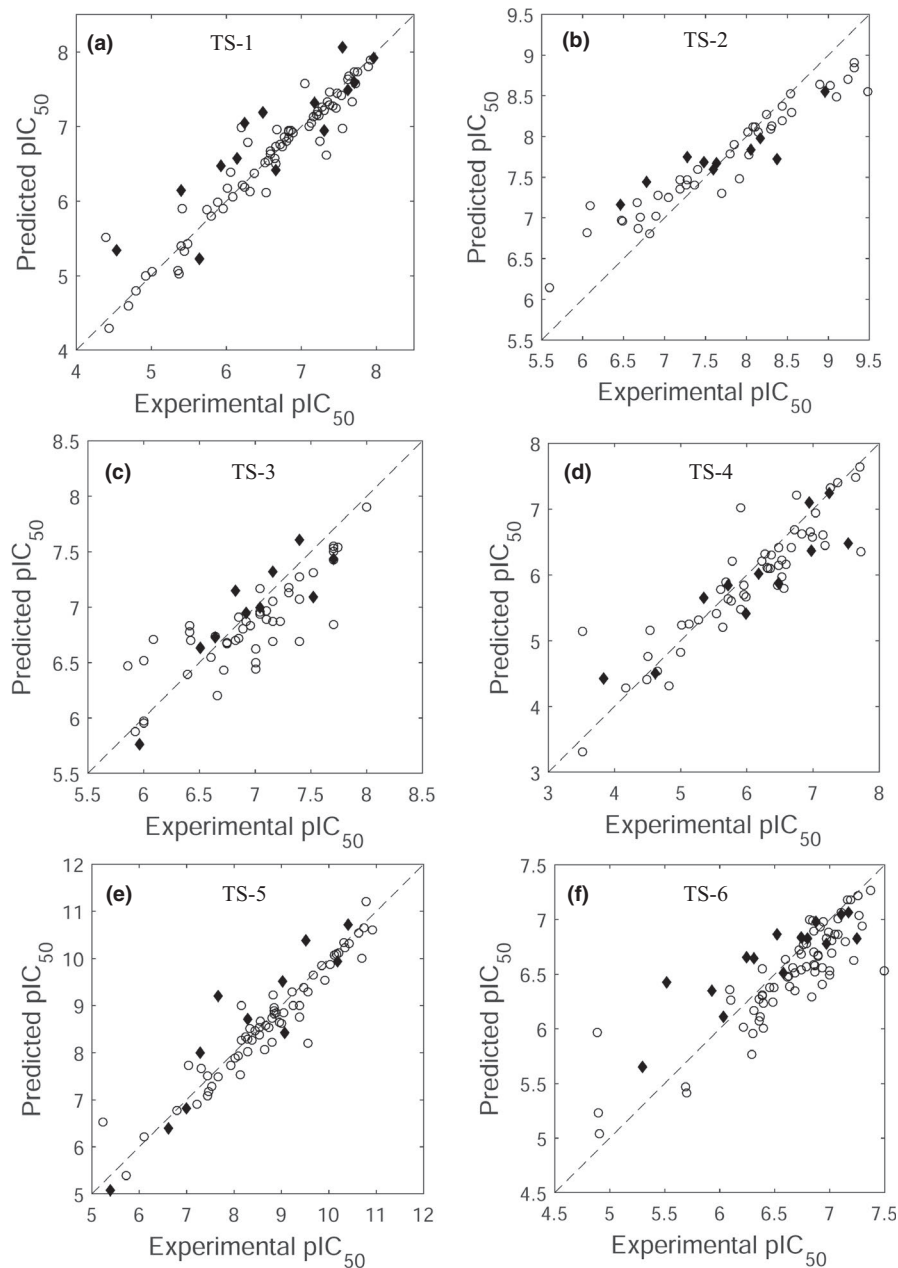
to TS-6 (Table 3). It is seen that the calculated external validation statistics also lie in the acceptable range and the results validate the PMF-PLS QSAR model.

The quality of PMF-PLS QSAR model was further assessed by checking for the mean absolute error (MAE) based statistical criteria that test for 99.7% confidence interval in prediction capability, that is, MAE plus three times the standard deviation ( $\text{MAE} + 3\text{SD}$ ; Roy et al., 2016). The MAE criteria were not satisfied for the target systems studied and suggested that the results may have arisen because the sample sizes of the validation sets are small. It led to our considering a relaxation of the confidence interval to 95% by using the same MAE criteria but with two times the standard deviation and following the procedure by Roy et al., 2016, that is,

where  $\text{MAE} = (\sum_{i=1}^n |y_i - \hat{y}_i|) / n$ ,  $y_i$  is the actual  $\text{pIC}_{50}$  value for the  $i$ th validation set molecule,  $\hat{y}_i$  the corresponding predicted value, and  $n$  is the number of validation set compounds. The models lying in between the good and bad are considered to be of moderate quality. Our calculations with the  $\text{MAE} + 2\text{SD}$  criteria indeed showed acceptance of model predictions and under the circumstances, it provides a reasonable alternative when sample size is small. The MAE-based criteria values evaluated using Equation 10 are given in Table 4. The quality of PMF-PLS QSAR model for TS-2 and TS-3 is found to be good while that for TS-1, TS-4, TS-5, and TS-6 are observed to be moderate.

The performance of PMF-PLS regression algorithm was also tested with the 2D charge-based descriptors discussed above. For this purpose, 34 charge weighted autocorrelation descriptors and 21 topological charge indices were identified using the web-based descriptor calculation platform ChemDes (<http://www.scbdd.com/chemdes/>; Dong et al., 2015) and the descriptor values calculated. Similarly, 14 PEOE-VSA descriptors and 15 atomic charge descriptors were calculated using the other web-based platform OCHEM (<https://ochem.eu/home/show.do>; Sushko et al., 2011). Of these 84 charge-based descriptors those with constant or near-constant values (standard deviation  $< 0.0001$ ) and ones with at least one missing value were excluded for a given target system (Ojha & Roy, 2018). For TS-1 to TS-6, the resultant pool of descriptors, respectively, were used instead of the PMF descriptors in the PMF-PLS algorithm for regressing with the corresponding  $\text{pIC}_{50}$  values. The QSAR models developed using the charge-based descriptors were then validated by predicting the  $\text{pIC}_{50}$  values of the corresponding TS-1 to TS-6 validation sets. The model performance parameters using these charge-based descriptors are presented in Table 5 along with the model quality assessment by the MAE-based criteria (Equation

**FIGURE 2** Plots of experimental  $\text{pIC}_{50}$  values ( $Y$ ) versus the predicted  $\text{pIC}_{50}$  values ( $\hat{Y}$ ) for cross-validation. (a) TS-1, (b) TS-2, (c) TS-3, (d) TS-4, (e) TS-5 and (f) TS-6 inhibitors. (○) the training set ( $X_{\text{train}}, Y_{\text{train}}$ ) compounds and (◆) test set ( $X_{\text{test}}, Y_{\text{test}}$ ) compounds listed in the inhibitor structure Tables (Supporting Information S1, Tables S1 to S6), respectively



10). Table 5 also reports the model performance parameters using PMF descriptor-based models. It may be observed that the PMF-PLS QSAR algorithm performed better overall using PMF descriptors when compared to using the 2D charge-based descriptors.

QSAR models using different descriptors and modeling approaches for the target systems studied in the current work have been reported using the same datasets. Three studies for TS-1 (Elmi et al., 2009; Yi et al., 2008), one for TS-2 (Queiroz et al., 2011), two for TS-3 (Hu et al., 2009), and one each for TS-4 (Garg et al., 1999) and TS-5 (Debnath, 1999) were identified and studied for comparative performance with PMF-PLS QSAR. Note that for TS-6 no QSAR modeling study could be identified. The nature of the QSAR models selected are summarized in Table 6. Using the prediction data

provided in each case, we calculated the RMSEP and  $Q_{\text{ext(F1)}}^2$

using Equations 6 and 9, respectively, and the model quality was checked with (MAE + 2SD) based criteria (Equation 10). The model quality metrics were compared with those obtained for the corresponding PMF-PLS model as shown in Table 6. The comparison of performance metrics shows that PMF-PLS QSAR models are comparable for TS-1 and TS-5 while for the other systems it is even better. Thus the results show that the present PMF-PLS QSAR approach is competitive to the existing methods.

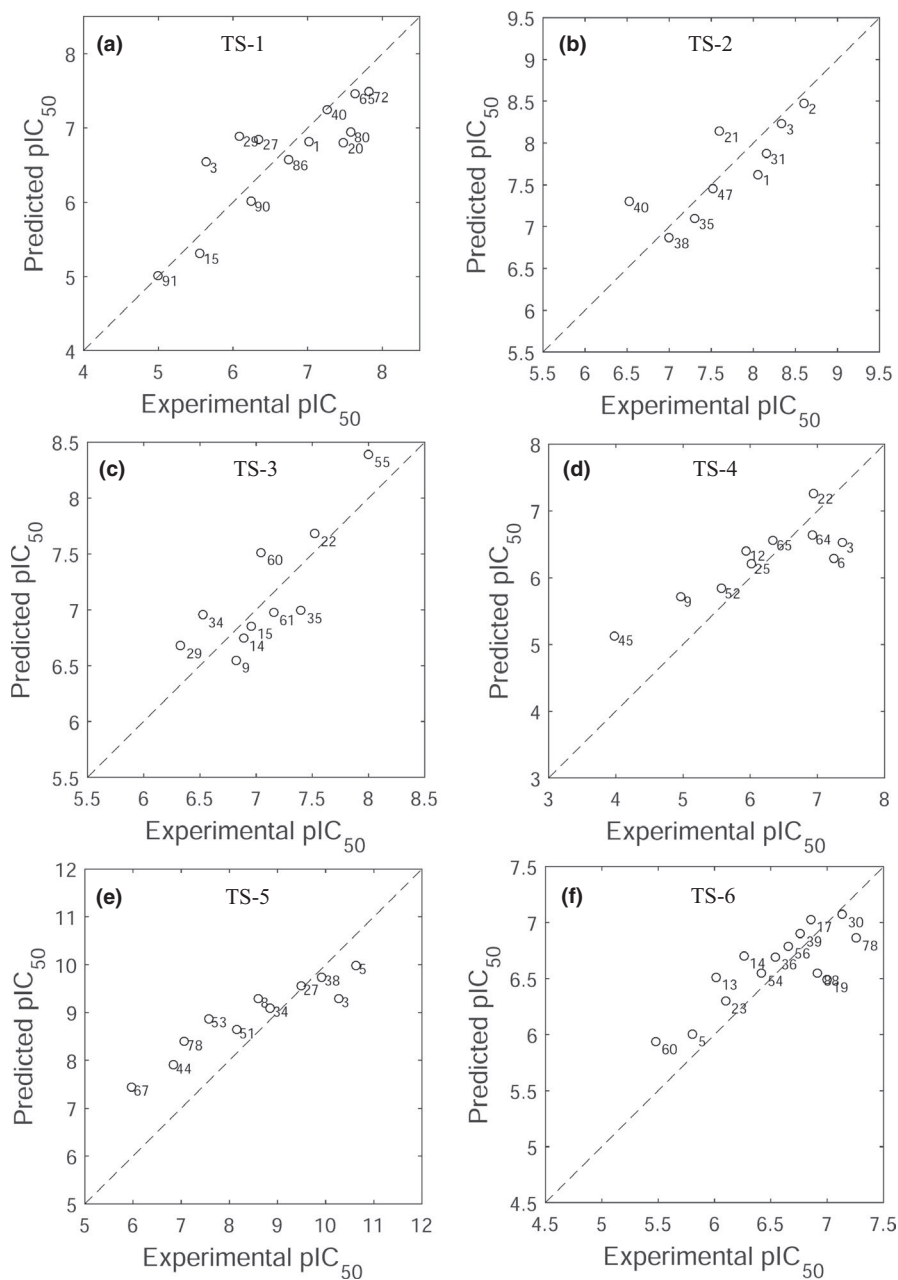
The results of statistics presented in Tables 3–6 for studies with TS-1 to TS-6 show that the PMF-PLS algorithm for practical purposes predicts the  $\text{pIC}_{50}$  values. It has, therefore, high potential in realizing applications for screening new



TABLE 3 PMF-PLS QSAR model fitting statistics for target systems TS-1 to TS-6

TS no.	AID <sup>a</sup>	Compounds	Number of PLS components, <i>a</i>	Cross-validation		External validation	
				$R^2_{cv}$	RMSECV	$Q^2_{ext(F1)}$	RMSEP
1	268838	4-Phenyl pyrrolo-carbazoles	25	0.88	0.31	0.71	0.47
2	566585	Benzylpiperidine derivatives	18	0.82	0.39	0.66	0.37
3	198247	2-Substituted dipyrro-diazepinones	20	0.68	0.32	0.69	0.32
4	197804	2-Pyridinone derivatives	18	0.79	0.47	0.62	0.64
5	160292	Cyclic urea derivatives	22	0.89	0.43	0.62	0.90
6	579588	Azilide derivatives	17	0.67	0.31	0.63	0.32

<sup>a</sup>Assay identification number from PubChemBioAssay.



**FIGURE 3** Plots for experimental pIC<sub>50</sub> values ( $Y_{val}$ ) versus. the predicted pIC<sub>50</sub> values ( $\hat{Y}_{val}$ ) for validation sets ( $X_{val}$ ,  $Y_{val}$ ) of (a) TS-1, (b) TS-2, (c) TS-3, (d) TS-4, (e) TS-5 and (f) TS-6 inhibitors. The numbers in the panels (a) to (f) indicate the compound numbers listed in the inhibitor structure Tables (Supporting Information Figure S1, Tables S1 to S6), respectively

**TABLE 4** Model quality using the MAE based criteria

TS no.	(MAE/training set range)	(MAE + 2SD/training set range)	Model quality
1	0.09	0.24	Moderate
2	0.06	0.14	Good
3	0.11	0.20	Good
4	0.11	0.25	Moderate
5	0.12	0.25	Moderate
6	0.10	0.21	Moderate

**TABLE 5** Performance comparison of present QSAR algorithm using PMF and 2D charge-based descriptors

TS no.	2D Charge-based descriptors			PMF descriptors		
	$Q^2_{\text{ext(F1)}}$	<i>RMSEP</i>	MAE-based criteria <sup>a</sup>	$Q^2_{\text{ext(F1)}}$	<i>RMSEP</i>	MAE-based criteria <sup>a</sup>
1	0.62	0.54	Moderate	0.71	0.47	Moderate
2	0.60	0.40	Good	0.66	0.37	Good
3	0.11	0.46	Bad	0.69	0.32	Good
4	0.55	0.69	Bad	0.62	0.64	Moderate
5	0.62	0.90	Moderate	0.62	0.90	Moderate
6	0.57	0.34	Good	0.63	0.32	Moderate

<sup>a</sup>Using (MAE + 2SD) based measure, Equation 10.**TABLE 6** Comparison of present PMF-PLS QSAR model with other QSAR models for the same datasets in this study

TS no.	QSAR model	<i>RMSEP</i>	$Q^2_{\text{ext(F1)}}$	MAE-based criteria <sup>a</sup>	References	PMF-PLS QSAR model		
						<i>RMSEP</i>	$Q^2_{\text{ext(F1)}}$	MAE-based criteria <sup>a</sup>
1	CoMFA	0.42	0.74	Moderate	Yi et al. (2008)	0.47	0.71	Moderate
	GA-MLR <sup>b</sup>	0.43	0.78	Good	Elmi et al. (2009)			
	Fuzzy entropy	0.36	0.85	Good	Elmi et al. (2009)			
2	RD-3D-QSAR <sup>c</sup>	0.81	0.06	Bad	Queiroz et al. (2011)	0.37	0.66	Good
3	CoMFA	0.71	0.48	Bad	Hu et al. (2009)	0.32	0.69	Good
	CoMSIA <sup>d</sup>	0.68	0.52	Bad	Hu et al. (2009)			
4	Physicochemical properties	0.62	0.39	Moderate	Garg et al. (1999)	0.62	0.64	Moderate
5	CoMFA	0.85	0.57	Moderate	Debnath (1999)	0.90	0.62	Moderate

<sup>a</sup>Using (MAE + 2SD) based measure, Equation 10.<sup>b</sup>Genetic algorithm based feature selection and multilinear regression.<sup>c</sup>Receptor dependent 3D-QSAR.<sup>d</sup>Comparative molecular similarity indices.

molecules with scaffolds similar to those used for model development in different target systems. Therefore, we chose natural compounds as new molecules for screening and present the results of studying their potency. The natural compound database SuperNatural II (Banerjee et al., 2015) was searched for the scaffolds listed in Table 1 to identify natural compounds with structure similar to the compounds in TS-1 to TS-6 (Supporting Information S1, Table S8). The selected natural compounds were processed and PMF values calculated using Equation 1. The pIC<sub>50</sub> values of these

natural compounds are not known and were predicted using the PMF-PLS QSAR models developed for the target systems (Supporting Information S1, Table S8) while ensuring that the predicted points lie in the applicability domain of the respective models.

The applicability domain of the QSAR model was defined using the range of response variable (Cruz-Monteagudo et al., 2014; Gadaleta et al., 2016; Kar et al., 2018). A compound for which the predicted response variable value is largely out of the range of the pIC<sub>50</sub> values of the training

set molecules (Supporting Information S1, Table S1–S6) were considered to be out of the applicability domain as suggested by Kar et al. (2018). The predicted  $pIC_{50}$  values for the natural compounds are given in Supporting Information S1, Table S8. It was observed that nine natural compounds selected for TS-1, three for TS-2, two for TS-3, ten for TS-4, two for TS-5 and one for TS-6 were within the applicability domain of their respective QSAR models and are identified in the Supporting Information S1, Table S8. Those natural compounds showing moderate to high predictions of  $pIC_{50}$  values could be additionally assessed by carrying out more detailed docking studies to confirm that these new molecules could bind to the target protein.

Of the nine natural compounds obtained from the SuperNatural II database for TS-1, the natural compounds SN00226661, SN00272309, SN00362452, and SN00362911 were found by the proposed algorithm to have predicted  $pIC_{50}$  values of 7.764, 6.929, 9.051, and 9.243 indicating good inhibitory potential. Similarly, compound SN00335138 for TS-2, SN00118406 for TS-3, compounds SN00008635, SN00008637, SN00008647, SN00008860, SN00010264, and SN00063879 for TS-4 and compound SN00215212 for TS-5 were predicted to have high  $pIC_{50}$  values (Supporting Information S1, Table S8). Therefore, docking studies for these compounds with their respective target proteins were conducted and showed that they could effectively interact and bind through the amino acid residues crucial for the activity of the protein. For brevity, results of the docking studies are discussed in Supporting Information S2 and they supplement the predictions studies by the PMF-PLS QSAR.

## 4 | CONCLUSIONS

The methodology of PMF-PLS studied here offers an alternate and simpler way of QSAR modeling which uses an effective correlative descriptor in terms of the intrinsic properties of atoms that are readily available in the literature, namely, the electron affinity and electronegativity values. The developed QSAR models showed generality by its application to a wide variety of target systems with good prediction statistics and thus bringing out its potential. The PMF-PLS QSAR model showed a competitive performance when compared with other published QSAR models for the same data sets as well as with the use of 2D charge descriptors. Additionally, it was applied to screen natural compounds with unknown biological activities. Potentially new molecules could thus be assessed by docking studies to confirm their binding to the target protein. Thus, the PMF-PLS method for QSAR modeling is a promising computational tool that may be used for

selecting new molecules for experimentation in ligand-based drug discovery programs.

## ACKNOWLEDGMENTS

P.D.K. is thankful to Academy of Scientific and Innovative Research (AcSIR) and Council of Scientific and Industrial Research (CSIR), New Delhi, India for the award of a research fellowship to carry out this work.

## CONFLICT OF INTEREST

Authors confirm that there is no conflict of interest to declare.

## ORCID

V. Ravi Kumar  <https://orcid.org/0000-0001-6953-3815>

## REFERENCES

- Ballabio, D., Grisoni, F., Consonni, V., & Todeschini, R. (2019). Integrated QSAR models to predict acute oral systemic toxicity. *Molecular Informatics*, 38(8–9), 1800124. <https://doi.org/10.1002/minf.201800124>
- Banerjee, P., Erehman, J., Gohlke, B. O., Wilhelm, T., Preissner, R., & Dunkel, M. (2015). Super natural II—A database of natural products. *Nucleic Acids Research*, 43(D1), D935–D939. <https://doi.org/10.1093/nar/gku886>
- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–231. <https://doi.org/10.1214/ss/1009213726>
- Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, 110(18), 5959–5967. <https://doi.org/10.1021/ja00226a005>
- Cronin, M. T. D. (2010). Prediction of harmful human health effects of chemicals from structure. In T. Puzyn, J. Leszczynski, & M. T. Cronin (Eds.), *Recent advances in QSAR studies: Methods and applications* (pp. 305–325). Springer Netherlands. doi: [https://doi.org/10.1007/978-1-4020-9783-6\\_11](https://doi.org/10.1007/978-1-4020-9783-6_11)
- Cruz-Monteagudo, M., Medina-Franco, J. L., Pérez-Castillo, Y., Nicolotti, O., Cordeiro, M. N. D. S., & Borges, F. (2014). Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discovery Today*, 19(8), 1069–1080. <https://doi.org/10.1016/j.drudis.2014.02.003>
- Cumming, J. G., Davis, A. M., Muresan, S., Haerberlein, M., & Chen, H. (2013). Chemical predictive modelling to improve compound quality. *Nature Reviews. Drug Discovery*, 12(12), 948–962. <https://doi.org/10.1038/nrd4128>
- Damale, M., Harke, S., Kalam Khan, F., Shinde, D., & Sangshetti, J. (2014). Recent advances in multidimensional QSAR (4D–6D): A critical review. *Mini-Reviews in Medicinal Chemistry*, 14(1), 35–55. <https://doi.org/10.2174/13895575113136660104>
- Dasoondi, A. S., Singh, V., Voleti, S. R., & Tiwari, M. (2008). Comparative molecular field analysis of benzothiazepine derivatives: Mitochondrial sodium calcium exchange inhibitors as anti-diabetic agents. *Indian Journal of Pharmaceutical Sciences*, 70(2), 186–192. <https://doi.org/10.4103/0250-474X.41453>
- de Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3), 251–263. [https://doi.org/10.1016/0169-7439\(93\)85002-X](https://doi.org/10.1016/0169-7439(93)85002-X)

- Debnath, A. K. (1999). Three-dimensional quantitative structure–activity relationship study on cyclic urea derivatives as HIV-1 protease inhibitors: Application of comparative molecular field analysis †. *Journal of Medicinal Chemistry*, 42(2), 249–259. <https://doi.org/10.1021/jm980369n>
- Divakar, S., & Hariharan, S. (2015). 3D-QSAR studies on Plasmodium falciparum proteins: A mini-review. *Combinatorial Chemistry & High Throughput Screening*, 18(2), 188–198.
- Dong, J., Cao, D.-S., Miao, H.-Y., Liu, S., Deng, B.-C., Yun, Y.-H., Wang, N.-N., Lu, A.-P., Zeng, W.-B., & Chen, A. F. (2015). ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. *Journal of Cheminformatics*, 7(1), 60. <https://doi.org/10.1186/s13321-015-0109-z>
- Elmi, Z., Faez, K., Goodarzi, M., & Goodarzi, N. (2009). Feature selection method based on fuzzy entropy for regression in QSAR studies. *Molecular Physics*, 107(17), 1787–1798. <https://doi.org/10.1080/00268970903078559>
- Estrada, E. (1995). Three-dimensional molecular descriptors based on electron charge density weighted graphs. *Journal of Chemical Information and Computer Sciences*, 35(4), 708–713. <https://doi.org/10.1021/ci00026a006>
- Gadaleta, D., Mangiatordi, G. F., Catto, M., Carotti, A., & Nicolotti, O. (2016). Applicability domain for QSAR models. *International Journal of Quantitative Structure-Property Relationships*, 1(1), 45–63. <https://doi.org/10.4018/IJQSPR.2016010102>
- Garg, R., Gupta, S. P., Gao, H., Babu, M. S., Debnath, A. K., & Inhibitors, G. P. (1999). Comparative quantitative structure–activity relationship studies on anti-HIV drugs. *Chemical Reviews*, 99(12), 3525–3602. <https://doi.org/10.1021/cr9703358>
- Garthwaite, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, 89(425), 122–127. <https://doi.org/10.1080/01621459.1994.10476452>
- Gasteiger, J., & Engel, T. E. (Eds.). (2003). *Cheminformatics: A textbook*. Wiley-VCH Verlag GmbH & Co. KGaA.
- Gasteiger, J., & Marsili, M. (1980). Iterative partial equalization of orbital electronegativity—A rapid access to atomic charges. *Tetrahedron*, 36(22), 3219–3228. [https://doi.org/10.1016/0040-4020\(80\)80168-2](https://doi.org/10.1016/0040-4020(80)80168-2)
- Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185, 1–17. [https://doi.org/10.1016/0003-2670\(86\)80028-9](https://doi.org/10.1016/0003-2670(86)80028-9)
- Gramatica, P. (2020). Principles of QSAR modeling: Comments and suggestions from personal experience. *International Journal of Quantitative Structure-Property Relationships*, 5(3), 61–97. <https://doi.org/10.4018/IJQSPR.20200701.oa1>
- Hu, R., Barbault, F., Delamar, M., & Zhang, R. (2009). Receptor- and ligand-based 3D-QSAR study for a series of non-nucleoside HIV-1 reverse transcriptase inhibitors. *Bioorganic and Medicinal Chemistry*, 17(6), 2400–2409. <https://doi.org/10.1016/j.bmc.2009.02.003>
- Hutinec, A., Rupčić, R., Žiher, D., Smith, K. S., Milhous, W., Ellis, W., Ohrt, C., & Schönfeld, Z. I. (2011). An automated, polymer-assisted strategy for the preparation of urea and thiourea derivatives of 15-membered azalides as potential antimalarial chemotherapeutics. *Bioorganic and Medicinal Chemistry*, 19(5), 1692–1701. <https://doi.org/10.1016/j.bmc.2011.01.030>
- Kar, S., Roy, K., & Leszczynski, J. (2018). Applicability domain: A step toward confident predictions and decidability for QSAR modeling. In O. Nicolotti (Ed.), *Computational toxicology: Methods and protocols* (pp. 141–169). Springer New York. [https://doi.org/10.1007/978-1-4939-7899-1\\_6](https://doi.org/10.1007/978-1-4939-7899-1_6)
- Kendall, D. G. (1989). [A Survey of the Statistical Theory of Shape]: Rejoinder. *Statistical Science*, 4(2), 116–120. <https://doi.org/10.1214/ss/1177012589>
- Kubinyi, H. (1997). QSAR and 3D QSAR in drug design. Part 1: Methodology. *Drug Discovery Today*, 2(11), 457–467. [https://doi.org/10.1016/S1359-6446\(97\)01079-9](https://doi.org/10.1016/S1359-6446(97)01079-9)
- Labute, P. (2000). ScienceDirect—Journal of Molecular Graphics and Modelling : A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling*, 3263(Figure 1), 464–477.
- Li, S., Fan, J., Peng, C., Chang, Y., Guo, L., Hou, J., Huang, M., Wu, B., Zheng, J., Lin, L., Xiao, G., Chen, W., Liao, G., Guo, J., & Sun, P. (2017). New molecular insights into the tyrosyl-tRNA synthase inhibitors: CoMFA, CoMSIA analyses and molecular docking studies. *Scientific Reports*, 7(1), 11525. <https://doi.org/10.1038/s41598-017-10618-1>
- Lill, M. A. (2007). Multi-dimensional QSAR in drug discovery. *Drug Discovery Today*, 12(23–24), 1013–1017. <https://doi.org/10.1016/j.drudis.2007.08.004>
- Lindgren, Å., Sjöström, M., & Wold, S. (1996). QSAR modelling of the toxicity of some technical non-ionic surfactants towards fairy shrimps. *Quantitative Structure-Activity Relationships*, 15, 208–218. <https://doi.org/10.1002/qsar.19960150305>
- Madden, J. C. (2010). In silico approaches for predicting adme properties. In T. Puzyn, J. Leszczynski, & M. T. Cronin (Eds.), *Recent advances in QSAR studies: Methods and applications* (pp. 283–304). Springer Netherlands. [https://doi.org/10.1007/978-1-4020-9783-6\\_10](https://doi.org/10.1007/978-1-4020-9783-6_10)
- MATLAB. (2010). *MATLAB:2010 (No. R2010b (7.11); R2010b ed.)*. The MathWorks Inc. Retrieved from <https://in.mathworks.com/help/matlab/index.html>
- Matta, C. F., & Arabi, A. A. (2011). Electron-density descriptors as predictors in quantitative structure–activity/property relationships and drug design. *Future Medicinal Chemistry*, 3(8), 969–994. <https://doi.org/10.4155/fmc.11.65>
- Mendez, K. M., Reinke, S. N., & Broadhurst, D. I. (2019). A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics*, 15(12), 150. <https://doi.org/10.1007/s11306-019-1612-4>
- Mulliken, R. S. (1934). A new electroaffinity scale; Together with data on valence states and on valence ionization potentials and electron affinities. *The Journal of Chemical Physics*, 2(11), 782–793. <https://doi.org/10.1063/1.1749394>
- Nidhi, & Siddiqi, M. I. (2013). Recent advances in QSAR-based identification and design of anti-tubercular agents. *Current Pharmaceutical Design*, 20(27), 4418–4426. <https://doi.org/10.2174/1381612819666131118165059>
- Ojha, P. K., & Roy, K. (2018). PLS regression-based chemometric modeling of odorant properties of diverse chemical constituents of black tea and coffee. *RSC Advances*, 8(5), 2293–2304. <https://doi.org/10.1039/C7RA12914A>
- Oyewole, R. O., Oyebamiji, A. K., & Semire, B. (2020). Theoretical calculations of molecular descriptors for anticancer activities of 1, 2, 3-triazole-pyrimidine derivatives against gastric cancer cell line (MGC-803): DFT. *QSAR and Docking Approaches. Heliyon*, 6(5), e03926. <https://doi.org/10.1016/j.heliyon.2020.e03926>
- Palmer, B. D., Thompson, A. M., Booth, R. J., Dobrusin, E. M., Kraker, A. J., Lee, H. H., & Denny, W. A. (2006). 4-Phenylpyrrolo[3,4-c]carbazole-1,3(2H,6H)-dione inhibitors of the checkpoint kinase Wee1. Structure-activity relationships for chromophore modification

- and phenyl ring substitution. *Journal of Medicinal Chemistry*, 49(16), 4896–4911. <https://doi.org/10.1021/jm0512591>
- Panagou, E. Z., Mohareb, F. R., Argyri, A. A., Bessant, C. M., & Nychas, G.-J.-E. (2011). A comparison of artificial neural networks and partial least squares modelling for the rapid detection of the microbial spoilage of beef fillets based on Fourier transform infrared spectral fingerprints. *Food Microbiology*, 28(4), 782–790. <https://doi.org/10.1016/j.fm.2010.05.014>
- Proudfoot, J. R., Hargrave, K. D., Kapadia, S. R., Patel, U. R., Grozinger, K. G., Mcneil, D. W., Adams, J. (1995). Novel non-nucleoside inhibitors of human immunodeficiency virus type 1 (HIV-1) reverse transcriptase. 4. 2-Substituted dipyrrodozepinones as potent inhibitors of both wild-type and cysteine-181 HIV-1 reverse transcriptase enzymes. *Journal of Medicinal Chemistry*, 1(38), 4830–4838.
- Queiroz, J., Araújo, M., Brito, D., Villas, L., Hoelz, B., Bicca, R., & Girão, M. (2011). Receptor-dependent (RD) 3D-QSAR approach of a series of benzylpiperidine inhibitors of human acetylcholinesterase (HuAChE). *European Journal of Medicinal Chemistry*, 46(1), 39–51. <https://doi.org/10.1016/j.ejmech.2010.10.009>
- Roy, K., & Das, R. N. (2014). A review on principles, theory and practices of 2D-QSAR. *Current Drug Metabolism*, 15(4), 346–379.
- Roy, K., Das, R. N., Ambure, P., & Aher, R. B. (2016). Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemometrics and Intelligent Laboratory Systems*, 152, 18–33. <https://doi.org/10.1016/j.chemolab.2016.01.008>
- Schrödinger, LLC, N. Y. (2011a). *LigPrep 2.5*. Schrödinger, LLC, New York.
- Schrödinger, LLC, N. Y. (2011b). *Maestro 9.2*. Schrödinger, LLC, New York.
- Schwartz, W. R., Kembhavi, A., Harwood, D., & Davis, L. S. (2009). Human detection using partial least squares analysis. In *2009 IEEE 12th International Conference on Computer Vision, (Iccv)*, pp. 24–31. IEEE. doi: <https://doi.org/10.1109/ICCV.2009.5459205>
- Simeon, S., Montanari, D., & Gleeson, M. P. (2019). Investigation of factors affecting the performance of in silico volume distribution QSAR models for human, rat, mouse. *Dog & Monkey. Molecular Informatics*, 38(10), 1900059. <https://doi.org/10.1002/minf.20190059>
- Stanton, D. T., & Jurs, P. C. (1990). Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Analytical Chemistry*, 62(21), 2323–2329. <https://doi.org/10.1021/ac00220a013>
- Sushko, I., Novotarskyi, S., Körner, R., Pandey, A. K., Rupp, M., Teetz, W., Brandmaier, S., Abdelaziz, A., Prokopenko, V. V., Tanchuk, V. Y., Todeschini, R., Varnek, A., Marcou, G., Ertl, P., Potemkin, V., Grishina, M., Gasteiger, J., Schwab, C., Baskin, I. I., ... Tetko, I. V. (2011). Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *Journal of Computer-Aided Molecular Design*, 25(6), 533–554. <https://doi.org/10.1007/s10822-011-9440-2>
- Todeschini, R., & Consonni, V. (2008). User's guide. In R. Mannhold, H. Kubinyi, & H. Timmerman (Eds.), *Handbook of molecular descriptors*. Principles in Medicinal Chemistry (Vol. 11). WILEY-VCH.
- Totrov, M. (2007). Atomic property fields: Generalized 3D pharmacophoric potential for automated ligand superposition, pharmacophore elucidation and 3D QSAR. *Chemical Biology & Drug Design*, 71(1), 15–27. <https://doi.org/10.1111/j.1747-0285.2007.00605.x>
- Wai, J. S., Williams, T. M., Bamberger, D. L., Fisher, T. E., Hoffman, J. M., Hudcosky, R. J., & Andersont, P. S. (1993). Synthesis and evaluation of 2-pyridinone derivatives as specific HIV-1 reverse transcriptase inhibitors. 3. Pyridyl and phenyl analogs of 3-aminopyridin-2(1H)-one. *Journal of Medicinal Chemistry*, 36(2), 249–255.
- Wang, Y., Suzek, T., Zhang, J., Wang, J., He, S., Cheng, T., Shoemaker, B. A., Gindulyte, A., & Bryant, S. H. (2014). PubChem BioAssay: 2014 update. *Nucleic Acids Research*, 42(D1), 1075–1082. <https://doi.org/10.1093/nar/gkt978>
- Yi, P., Fang, X., & Qiu, M. (2008). 3D-QSAR studies of checkpoint kinase weel inhibitors based on molecular docking, CoMFA and CoMSIA. *European Journal of Medicinal Chemistry*, 43(5), 925–938. <https://doi.org/10.1016/j.ejmech.2007.06.021>

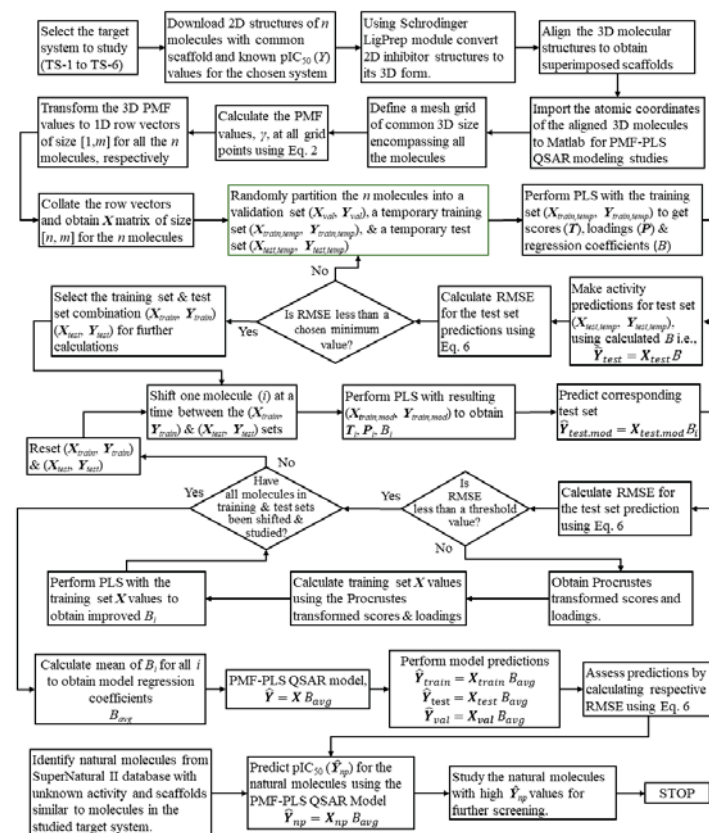
## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Kunde P. D., Ramkumar S., Kamble S. P., Ravikumar A., Kulkarni B. D., Kumar V. R. (2021). On the use of electronegativity and electron affinity based pseudo-molecular field descriptors in developing correlations for quantitative structure-activity relationship modeling of drug activities. *Chemical Biology and Drug Design*, 00, 1–12. <https://doi.org/10.1111/cbdd.13895>

**Supporting Information: S1**  
**PMF-PLS QSAR Studies**

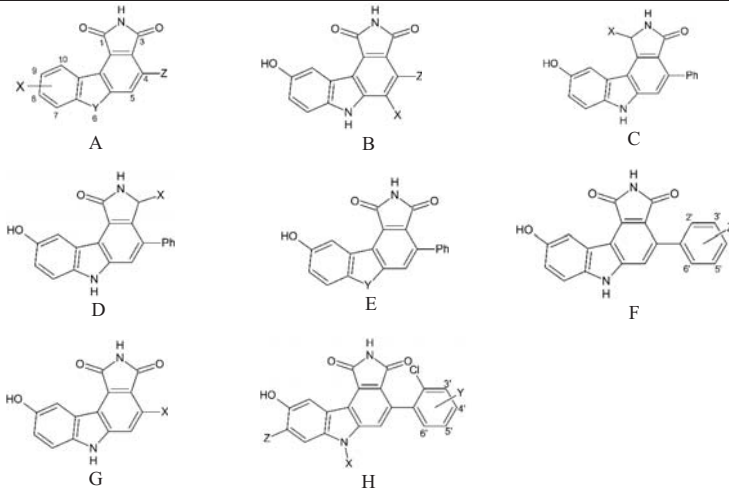
Supporting information: S1



Supporting Information S1, Figure S1: Schematic flowchart for the proposed PMF-PLS QSAR modelling methodology.

## Supporting information: S1

**Supporting Information S1, Table S1:** Structures of 4-phenylpyrrolocarbazole derivatives for TS-1 (AID: 268838) (anti-cancer Wee1 inhibitors) (compounds marked with '[a]' are chosen as test set compounds, those marked with '[b]' are chosen as validation set compounds. Remaining compounds belong to the training set.)



Compound No.	Structure	X	Y	Z	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1 <sup>[b]</sup>	A	9-OH	NH	Ph	7.0130	6.8199
2	A	9-OH	NH	H	5.3980	5.3982
3 <sup>[b]</sup>	A	9-OH	NH	I	5.6380	6.5368
4	A	8-OH	NH	Ph	6.5090	6.5168
5	A	9-OH	O	Ph	6.3670	6.3663
6	A	9-OH	S	Ph	7.1080	7.0088
7	A	9-OH	NMe	Ph	6.5850	6.6773
8	B	Me		Ph	6.8860	6.9198
9	B	Et		Ph	5.7960	5.8001
10	B	Ph		Me	5.0130	5.0566





## Supporting information: S1

Compound No.	Structure	X	Y	Z	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
11 <sup>[a]</sup>	B	Ph		Ph	5.6380	5.2314
12 <sup>[a]</sup>	B	Ph		H	5.3980	6.1399
13	C	OMe			4.6990	4.5906
14	C	H			4.4320	4.3002
15 <sup>[b]</sup>	D	-OH			5.5530	5.3095
16	E	N-NH <sub>2</sub>	NH		5.4090	5.9034
17 <sup>[a]</sup>	A	9-OH	NH	2'-ClPh	7.9590	7.9184
18	A	9-OMe	NH	2'-ClPh	6.1940	6.9884
19	A	9-OH	NMe	2'-ClPh	7.2440	6.8014
20 <sup>[b]</sup>	A	9-OH	O	2'-ClPh	7.4810	6.7966
21 <sup>[a]</sup>	F			2'-F	6.4810	7.1956
22	F			2'-Br	7.6380	7.6689
23	F			2'-I	7.8860	7.8068
24	F			2'-Me	6.8240	6.9440
25	F			2'-Et	6.2920	6.7891
26	F			2'-CF <sub>3</sub>	6.2370	6.1876
27 <sup>[b]</sup>	F			2'-CH <sub>2</sub> OH	6.3470	6.8510
28	F			2'-CN	6.7210	6.7544
29 <sup>[b]</sup>	F			2'-COMe	6.0810	6.8885
30	F			2'-CONH <sub>2</sub>	6.7960	6.8078
31 <sup>[a]</sup>	F			2'-Ph	6.2440	7.0389
32	F			2'-OH	7.2220	7.2010
33	F			2'-OMe	7.6200	7.6308
34	F			2'-OEt	6.5850	6.6233
35	F			2'-SMe	7.4810	7.4494
36 <sup>[a]</sup>	F			2'-SOMe	6.6580	6.4160
37	F			2'-NO <sub>2</sub>	7.3280	6.6180

## Supporting information: S1





Compound No.	Structure	X	Y	Z	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
38	F			2'-NH <sub>2</sub>	6.6780	6.9577
39	F			3'-F	6.6580	6.4992
40 <sup>[b]</sup>	F			3'-Cl	7.2600	7.2475
41	F			3'-Me	6.6380	6.5741
42	F			3'-CH <sub>2</sub> OH	6.0600	6.3796
43	F			3'-CH <sub>2</sub> NH <sub>2</sub>	5.3570	5.0725
44	F			3'-CN	6.7450	6.7338
45	F			3'-COMe	5.3670	5.0203
46	F			3'-Ph	4.3980	5.5118
47	F			3'-OH	7.0510	7.5764
48	F			3'-OMe	6.2080	6.2167
49	F			3'-NO <sub>2</sub>	6.5230	6.1155
50	F			3'-NH <sub>2</sub>	7.1550	7.1361
51	F			4'-F	4.7960	4.7964
52 <sup>[a]</sup>	F			4'-Cl	6.1370	6.5767
53	F			4'-Me	5.4810	5.4196
54 <sup>[a]</sup>	F			4'-CH <sub>2</sub> OH	5.9210	6.4750
55	F			4'-CN	5.7450	5.8876
56	F			4'-COMe	5.4440	5.3297
57 <sup>[a]</sup>	F			4'-OH	7.1740	7.3140
58	F			4'-OMe	4.9210	5.0032
59 <sup>[a]</sup>	F			4'-SMe	4.5380	5.3356
60	F			4'-SO <sub>2</sub> Me	5.9590	5.8989
61	F			4'-NH <sub>2</sub>	6.8240	6.9322
62 <sup>[a]</sup>	F			2'-Cl, 3'-Cl	7.5530	8.0644
63	F			2'-Cl, 3'-OH	7.9210	7.8859
64	F			2'-Cl, 3'-NH <sub>2</sub>	7.6780	7.3371

## Supporting information: S1

Compound No.	Structure	X	Y	Z	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
65 <sup>[b]</sup>	F			2'-Cl, 5'-OH	7.6380	7.4654
66 <sup>[a]</sup>	F			2'-Cl, 4'-NH <sub>2</sub>	7.6200	7.4848
67	F			2'-Cl, 5'-Cl	6.3100	6.1308
68	F			2'-Cl, 5'-OH	7.3770	7.4654
69 <sup>[a]</sup>	F			2'-Cl, 5'-NH <sub>2</sub>	7.6990	7.5831
70	F			2'-Cl, 6'-Cl	7.5530	6.9693
71	F			2'-Cl, 6'-OH	7.3470	7.3299
72 <sup>[b]</sup>	F			2'-Cl, 6'-OMe	7.8240	7.4867
73	F			2'-Br, 4'-NH <sub>2</sub>	7.6990	7.7363
74	F			2'-Br, 6'-Br	7.4560	7.2496
75	F			2'-Me, 3'-Me	6.5690	6.7264
76	F			2'-Me, 5'-Me	6.0180	6.1717
77	F			2'-Me, 6'-Me	7.1250	7.0465
78	F			2'-OMe, 4'-NH <sub>2</sub>	7.7210	7.5737
79	F			2'-OMe, 6'-OMe,	6.9590	6.5485
80 <sup>[b]</sup>	F			2'-OMe, 6'-F	7.5690	6.9461
81	F			2'-OMe, 4'-NH <sub>2</sub>	7.5380	7.4164
82	F			2',6',-diCl, 3'-OH	7.7450	7.7356
83 <sup>[a]</sup>	F			2',6',-diCl, 4'-OH	7.7310	6.9461
84	G				6.8540	6.9396
85	G				7.3770	7.2915
86 <sup>[b]</sup>	G				6.7450	6.5655
87	G				7.4200	7.2733



## Supporting information: S1

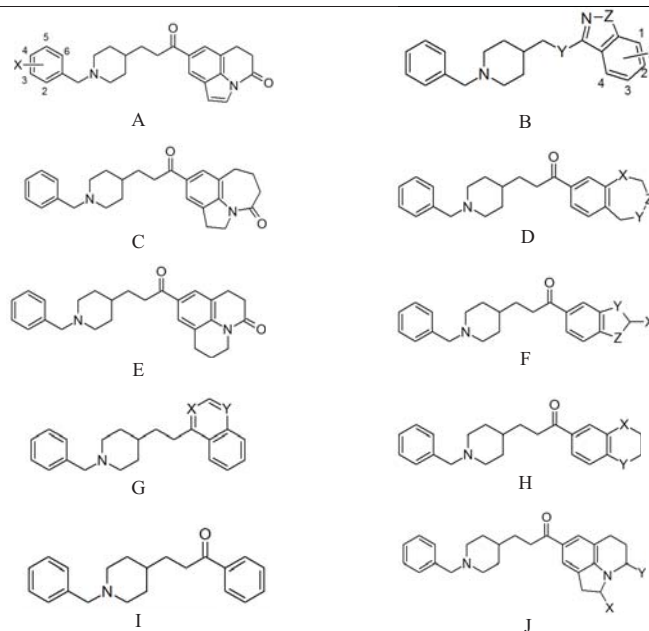
Compound No.	Structure	X	Y	Z	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
88	G				5.8860	5.9823
89	G				6.0860	6.0587
90 <sup>[b]</sup>	G				6.2370	6.0107
91 <sup>[b]</sup>	G				5.0000	5.0133
92	H	Et			7.3010	7.2210
93	H	<i>n</i> -Pr			7.2010	7.1598
94	H	<i>i</i> -Pr			7.2760	7.2615
95	H	<i>n</i> -Bu			7.2290	7.1476
96	H	(CH <sub>2</sub> ) <sub>2</sub> <i>i</i> -Pr			6.8240	6.8395
97	H	<i>n</i> -pent			6.7700	6.8526

[a] - test set compounds

[b] - validation set compounds





## Supporting information: S1

**Supporting Information S1, Table S2:** Structures of benzylpiperidine derivatives for TS-2 (AID: 566585) (AChE inhibitors) (compounds marked with 'a' are chosen as test set compounds, those marked with 'b' are chosen as validation set compounds. Remaining compounds belong to the training set.)






Compound No.	Structure	X	Y	Z	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1 <sup>[b]</sup>	A	3-OH			8.0600	7.6259
2 <sup>[b]</sup>	A	2-F			8.6003	8.4804
3 <sup>[b]</sup>	A	3-OH			8.3401	8.2354
4 <sup>[a]</sup>	A	2-OH			8.9586	8.5455
5	A	3-NO <sub>2</sub>			8.5406	8.5256

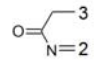
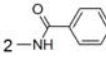
## Supporting information: S1

Compound No.	Structure	X	Y	Z	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
6	A	2-OCH <sub>3</sub>			7.1900	7.4639
7	A	4-Cl			6.8200	6.8100
8	E				7.8000	7.7911
9	A	3-Cl			8.3098	8.1364
10	A	2-Cl			8.2899	8.0957
11	A	2-NO <sub>2</sub>			7.0500	7.2494
12	A	3-F			8.8894	8.6430
13 <sup>[a]</sup>	A	4-OCH <sub>3</sub>			6.4600	7.1611
14	B		CH <sub>2</sub>	O	9.1002	8.4861
15 <sup>[a]</sup>	C				7.2800	7.7486
16 <sup>[a]</sup>	F	Ph	NH	NH	7.4800	7.6855
17	B		CH <sub>2</sub>	O	7.8499	7.9033
18	A	4-NO <sub>2</sub>			7.3700	7.4010
19	A	4-OH			9.3098	8.8453
20	A	3-OCH <sub>3</sub>			6.9000	7.0191
21 <sup>[b]</sup>	D	CH <sub>2</sub>	CH <sub>2</sub>	NH	7.6000	8.1405
22	D	O	NH	CH <sub>2</sub>	7.4000	7.5953
23 <sup>[a]</sup>	B		CH <sub>2</sub>	O	8.0600	7.8361
24	B	H	NH	O	6.0900	7.1472
25	B	H	O	O	5.5900	6.1480
26	B	3-OCH <sub>3</sub>	CH <sub>2</sub>	O	8.1403	8.0485
27	G	N	N		6.4700	6.9741
28	B		CH <sub>2</sub>	O	9.3197	8.9130
29 <sup>[a]</sup>	D	CH <sub>2</sub>			7.6400	7.6709
30	F	Me	NH	N	7.9201	7.4766

## Supporting information: S1

Compound No.	Structure	X	Y	Z	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
31 <sup>[b]</sup>	B	1-OCH <sub>3</sub>	CH <sub>2</sub>	O	8.1500	7.8774
32	B	H	-(CH <sub>2</sub> ) <sub>2</sub> -	O	6.0500	6.8237
33 <sup>[a]</sup>	F	Me	H	NC <sub>2</sub> H <sub>5</sub>	8.3696	7.7202
34	B	2-NHAc	CH <sub>2</sub>	O	8.5498	8.2930
35 <sup>[b]</sup>	B	2-Br	CH <sub>2</sub>	O	7.3000	7.1032
36 <sup>[a]</sup>	H	NH	CH <sub>2</sub>		6.7800	7.4435
37	F	H	CH <sub>2</sub>	NH	7.2800	7.4705
38 <sup>[b]</sup>	B	H	CH <sub>2</sub>	S	7.0000	6.8721
39	B		CH <sub>2</sub>	O	9.2403	8.7087
40 <sup>[b]</sup>	I				6.5200	7.3008
41	B		CH <sub>2</sub>	O	8.4401	8.3728
42	B	2 -CH <sub>3</sub> , 3-CH <sub>3</sub>	CH <sub>2</sub>	O	8.2403	8.2730
43	G	N	CH		6.6600	7.1870
44	B	H	NH	O	6.4900	6.9540
45	B	H	CH <sub>2</sub>	NH	6.9200	7.2808
46	B		CH <sub>2</sub>	O	9.0200	8.6247
47 <sup>[b]</sup>	H	O	O		7.5200	7.4547
48	H	CH <sub>2</sub>	NH		7.1900	7.3504
49	B	3 Me	CH <sub>2</sub>	O	8.1101	8.1229
50 <sup>[a]</sup>	B	2 =O	CH <sub>2</sub>	O	7.5901	7.5902
51	B	2 -NH <sub>2</sub>	CH <sub>2</sub>	O	7.7001	7.3087
52	B	H	CH	O	6.6800	6.8714
53	J	H	O		8.0101	8.0574
54	J	O	H		8.4401	8.1957
55	B	H	CH <sub>2</sub>	O	7.2600	7.4053
56	B	2 -OMe	CH <sub>2</sub>	O	8.0799	8.1237

## Supporting information: S1

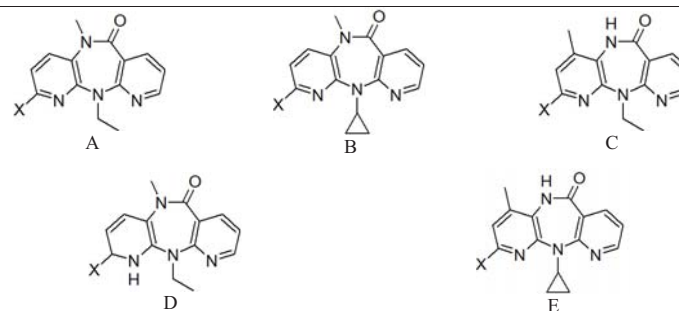
Compound No.	Structure	X	Y	Z	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
57	B		CH <sub>2</sub>	O	9.4802	8.5470
58	B		CH <sub>2</sub>	O	8.0301	7.7694
59	D	NH	CH <sub>2</sub>	CH <sub>2</sub>	6.7101	7.0089
60 <sup>[a]</sup>	F	Me	S	N	8.1701	7.9716


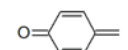

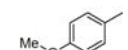
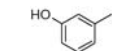
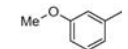
[a] - test set compounds

[b] - validation set compounds

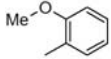












## Supporting information: S1

**Supporting Information S1, Table S3:** Structures of 2-substituted dipyrindiazepinone derivatives for TS-3 (AID: 198247) (HIV-1 RT inhibitors) (compounds marked with 'a' are chosen as test set compounds, those marked with 'b' are chosen as validation set compounds. Remaining compounds belong to the training set.)















Compound No.	Structure	X	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1	A		5.4318	6.4024
2	A	-CHCHCONH <sub>2</sub>	6.6021	6.5460
3	A	-CHCHCOOH	6.7447	6.6783
4	D		7.1549	7.0536
5	B		7.2218	6.8683
6	A	-NHCHCHCH <sub>3</sub>	6.4089	6.7784
7	A		5.8539	6.4740
8	A		7.0000	6.5006
9 <sup>[b]</sup>	A		6.8239	6.5481

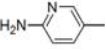
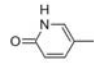
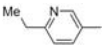


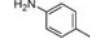
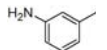
## Supporting information: S1

Compound No.	Structure	X	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
10	A		6.0862	6.7084
11 <sup>[a]</sup>	A	Ph	6.6383	6.7285
12	A		7.6990	6.8444
13	A		6.4089	6.8294
14 <sup>[b]</sup>	A		6.8861	6.7442
15 <sup>[b]</sup>	A		6.9586	6.8522
16	A		6.6576	6.1990
17	A		6.4202	6.7015
18	A		7.0000	6.4441
19	A		7.0000	6.6205
20	A		6.8539	6.7214
21	B		7.3010	7.1812
22 <sup>[b]</sup>	A		7.5229	7.6816
23	A		7.1549	6.8737


## Supporting information: S1

Compound No.	Structure	X	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
24	A		7.3979	6.6916
25	A		6.9586	6.8283
26	C	-SMe	7.6990	7.5478
27 <sup>[a]</sup>	B	-OMe	6.9208	6.9485
28	A	-OMe	7.3979	7.2706
29 <sup>[b]</sup>	D	=O	6.5086	6.6779
30 <sup>[a]</sup>	A		6.5086	6.6324
31	A		7.0548	6.9557
32	C		6.8239	6.7020
33	A		6.3979	6.3957
34 <sup>[b]</sup>	A		6.5229	6.9525
35 <sup>[b]</sup>	C		7.3979	6.9905
36	A		7.5229	7.3107
37	A		7.6990	7.4234
38	C		8.0000	7.8996
39 <sup>[a]</sup>	A	-N(Me) <sub>2</sub>	7.1549	7.3171
40	A	-CCH	6.8539	6.9064
41	C	Me	7.6990	7.5325
42 <sup>[a]</sup>	B		6.8239	7.1434

## Supporting information: S1

Compound No.	Structure	X	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
43	A		7.3010	7.1267
44 <sup>[a]</sup>	A		5.9586	5.7589
45	A		5.9208	5.8787
46	B		6.7447	6.6681
47	A		7.7447	7.5429
48 <sup>[a]</sup>	A		7.3979	7.6023
49 <sup>[a]</sup>	C	-NH(CH <sub>2</sub> ) <sub>3</sub> OH	7.0458	6.9926
50	C	-NH(CH <sub>2</sub> ) <sub>2</sub> OH	7.0458	6.9359
51	A	-NHEt	6.6383	6.7384
52	A	-NHMe	6.7212	6.4325
53	A	-NH <sub>2</sub>	6.0000	6.5164
54 <sup>[a]</sup>	B	Br	7.5229	7.0911
55 <sup>[b]</sup>	C	Cl	8.0000	8.3908
56	E	Cl	7.6990	7.5012
57 <sup>[a]</sup>	C	F	7.6990	7.4392
58	C	<i>t</i> -Bu	6.0000	5.9531
59	C	<i>i</i> -Pr	6.0000	5.9694
60 <sup>[b]</sup>	C	Et	7.0458	7.5143
61 <sup>[b]</sup>	E	Me	7.1549	6.9727
62	A		7.1549	6.6880
63	B	Cl	7.0458	7.1713
64	C	H	7.3979	7.0724

## Supporting information: S1

Compound No.	Structure	X	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
65	A	Cl	7.0969	6.9648
66	A	Me	6.9208	6.8662
67	A	H	6.8861	6.8019
68	E		7.0969	6.8855

[a] - test set compounds

[b] - validation set compounds

## Supporting information: S1

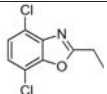
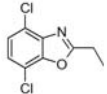
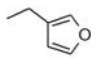
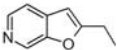
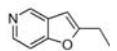
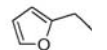
**Supporting Information S1, Table S4:** Structures of 2-pyridinone derivatives for TS-4 (AID: 197804) (HIV-1 RT inhibitors) (compounds marked with 'a' are chosen as test set compounds, those marked with 'b' are chosen as validation set compounds. Remaining compounds belong to the training set.)

Compound No.	Structure	X	Y	Z	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1		O	Et		5.0223	5.2342
2		S	Et		6.5229	6.2196
3 <sup>[b]</sup>		S	2-Et	1-Cl,4-Cl	7.3768	6.5253
4		O	Et		5.7100	5.6364
5		O	Et		6.4750	6.4169
6 <sup>[b]</sup>		O	Et		7.2441	6.2937
7		O	Et		6.4750	6.1469
8		O			3.5229	3.3099

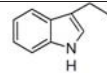
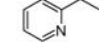
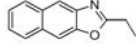
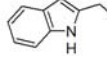
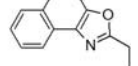
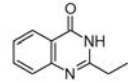
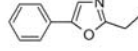
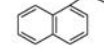
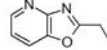
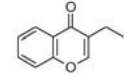
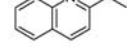
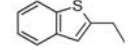
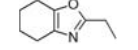
## Supporting information: S1

Compound No.	Structure	X	Y	Z	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>	
9 <sup>[b]</sup>		D			Me	4.9586	5.7133
10		B	O			4.5017	4.7559
11		B	O	2-COOEt		5.7570	5.6005
12 <sup>[b]</sup>		B	O			5.9393	6.4009
13		B	O	2-Et		6.2218	6.2009
14		B	O	2-SEt		6.3665	6.3053
15		B	O	2-SMe		6.7212	6.6815
16		B	O	1-Me	1-Cl,4-Cl	5.5452	5.4200
17		B	O	2-CH(OH)CH <sub>3</sub>	1-Cl,4-Cl	5.9788	5.6615
18		B	O	2-COCH <sub>3</sub>	1-Cl,4-Cl	6.5229	5.9653
19		A	O	-Et		5.9469	5.8468
20		B	O	2-SMe	1-Cl,4-Cl	7.3665	7.4037
21 <sup>[a]</sup>		B	O	2-OMe	1-Cl,4-Cl	6.9393	7.0981
22 <sup>[b]</sup>		B	O		1-Cl,4-Cl	6.9469	7.2547
23		B	O	2-CHCH <sub>2</sub>	1-Cl,4-Cl	7.6383	7.4760
24 <sup>[a]</sup>		D	Me			5.9788	5.4150
25 <sup>[b]</sup>		A	O	Et		6.0132	6.2049
26		D	Me			6.9872	6.5744







## Supporting information: S1

Compound No.	Structure	X	Y	Z	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
27 <sup>[a]</sup>	D	Et			6.1844	6.0249
28 <sup>[a]</sup>	D	Me			7.2366	7.2476
29	B	O	2-Et	4-NH <sub>2</sub>	4.1739	4.2827
30 <sup>[a]</sup>	B	O	2-Et	4-NO <sub>2</sub>	4.6108	4.5068
31	B	O	2-Et	4-OH	6.3565	6.0979
32	B	O	2-Et	4-OMe	6.7447	7.2108
33	B	O	2-Et	1-F, 4-F	7.1549	6.6025
34 <sup>[a]</sup>	B	O	2-Et	1-Cl, 4-F	6.9788	6.3635
35	B	O	2-Et	1-F	7.0362	6.9357
36	B	O	2-Et	2-F	5.9031	7.0132
37	B	O	2-Et	3-F	6.3279	6.1004
38	B	O	2-Et	4-F	6.9586	6.6588
39	B	O	2-Et	1-Cl	7.1871	6.4473
40	B	O	2-Et	4-Cl	6.8239	6.6182
41	B	O	2-Et	1-Et	6.5850	6.1680
42	B	O	2-Et	2-Me	5.7825	6.2154
43	B	O	2-Et	3-Me	5.9031	5.4820
44 <sup>[a]</sup>	A	O	Et		3.8386	4.4286
45 <sup>[b]</sup>	A	O	Et		3.9788	5.1326
46	A	O	Et		4.4908	4.4077
47	A	O	Et		4.5376	5.1638

## Supporting information: S1

Compound No.	Structure	X	Y	Z	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
48	A	O	Et		4.6478	4.5373
49	A	O	Et		4.8239	4.3118
50	A	O	Et		5.0000	4.8248
51 <sup>[a]</sup>	A	O	Et		5.3565	5.6488
52 <sup>[b]</sup>	A	O	Et		5.5686	5.8381
53	A	O	Et		5.6021	5.7757
54	A	O	Et		5.6289	5.1977
55	A	O	Et		5.6778	5.8888
56 <sup>[a]</sup>	A	O	Et		5.7212	5.8370
57	A	O	Et		5.9568	5.7055
58	A	O	Et		6.2757	6.3255
59	A	O	Et		6.3010	6.1133
60	A	O	Et		6.5528	5.7940

## Supporting information: S1

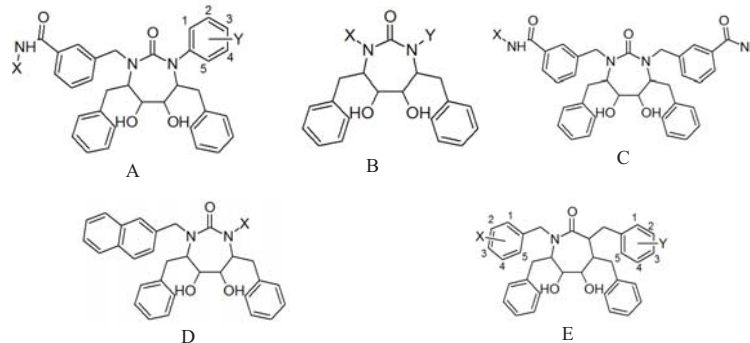
Compound No.	Structure	X	Y	Z	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
61	A	O	Et	-CH <sub>2</sub> Ph	5.2733	5.3174
62	A	O	Et		3.5229	5.1469
63	B	O	2-Et	4-Me	7.2596	7.3201
64 <sup>[b]</sup>	B	O	2-Et	1-Me	6.9208	6.6461
65 <sup>[b]</sup>	A	O	Et		6.3372	6.5584
66 <sup>[a]</sup>	A	O	Et		6.4815	5.8763
67	A	O	Et		6.4559	5.8400
68	A	O	Et		5.1203	5.2507
69 <sup>[a]</sup>	A	O	Et		7.5229	6.4771
70	B	O	2-Et		6.6778	6.4178
71	B	O	2-Et	1-Cl, 4-Cl	7.7212	6.3450
72	B	O	2-Et	1-Me, 4-Me	7.6990	7.6418







[a] - test set compounds

[b] - validation set compounds

## Supporting information: S1

**Supporting Information S1, Table S5:** Structures of cyclic urea derivatives for TS-5 (AID: 160292) (HIV-1 PR inhibitors) (compounds marked with 'a' are chosen as test set compounds, those marked with 'b' are chosen as validation set compounds. Remaining compounds belong to the training set.)



Compound No.	Structure	X	Y	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1	A		2-OMe	10.4202	10.3113
2	A		2-OMe	10.1612	10.0967
3 <sup>[b]</sup>	A		2-OMe	10.2757	9.2933
4	A		2-OMe	10.3279	10.3343
5 <sup>[b]</sup>	A		2-NH <sub>2</sub>	10.6383	9.9833
6	A		2-NH <sub>2</sub>	10.9208	10.6118



## Supporting information: S1

Compound No.	Structure	X	Y	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
7	B			10.6198	10.5378
8 <sup>[b]</sup>	A		2-OMe,4-OMe	8.6003	9.2822
9 <sup>[a]</sup>	A		2-OMe,4-OMe	9.0655	8.4308
10	A		2-NH <sub>2</sub>	10.1249	10.0781
11	A		2-NO <sub>2</sub>	10.0177	9.8635
12	A	-C(CH <sub>3</sub> ) <sub>3</sub>	2-NH <sub>2</sub>	9.3872	9.0033
13	A		2-NH <sub>2</sub>	10.7959	11.2164
14 <sup>[a]</sup>	B			5.3979	5.0705
15	B			8.7447	8.5392
16	B			8.1367	7.5353
17	B	-(CH <sub>2</sub> ) <sub>2</sub> C(CH <sub>3</sub> ) <sub>3</sub>	-(CH <sub>2</sub> ) <sub>2</sub> C(CH <sub>3</sub> ) <sub>3</sub>	7.4437	7.4995
18	C	NH <sub>2</sub>	NH <sub>2</sub>	10.7447	10.6498
19 <sup>[a]</sup>	C	H	H	10.4089	10.7238
20	C	-CH <sub>2</sub> CN	-CH <sub>2</sub> CN	10.2007	10.1164
21	C	<i>i</i> -Pr	<i>i</i> -Pr	9.2373	9.0043
22	C	Et	Et	9.6778	9.6575
23 <sup>[a]</sup>	C	Me	Me	10.1805	9.9438
24	C	OMe	OMe	10.3468	10.2373

## Supporting information: S1

Compound No.	Structure	X	Y	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
25	C	OH	OH	10.6990	10.0074
26	D			9.5528	9.2910
27 <sup>[b]</sup>	D			9.4815	9.5524
28	D			9.0000	8.6183
29	D			8.1612	9.0011
30 <sup>[a]</sup>	D			9.0315	9.5047
31	D			8.4437	8.4680
32	D			8.2840	8.2802
33	D	-CH <sub>2</sub> Ph		8.6383	8.0579
34 <sup>[b]</sup>	D	-CH <sub>2</sub> CHCH <sub>2</sub>		8.8539	9.0953
35	D			8.8239	9.2202
36	E	2-OH	2-OH	9.9208	9.5296
37	D	-(CH <sub>2</sub> ) <sub>3</sub> CH <sub>3</sub>		9.2218	9.2882
38 <sup>[b]</sup>	E	3-OH	3-OH	9.9208	9.7454
39	E	2-I	2-I	9.3768	8.7672
40	E	2-NO <sub>2</sub>	2-NO <sub>2</sub>	8.5528	8.6655
41	B			7.0458	7.7252
42	B			8.0132	7.8776
43	D	-(CH <sub>2</sub> ) <sub>2</sub> CH <sub>3</sub>		8.9586	8.6544

## Supporting information: S1

Compound No.	Structure	X	Y	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
44 <sup>[b]</sup>	B			6.8386	7.9035
45	B	-CH <sub>2</sub> CCH	-CH <sub>2</sub> CCH	7.6576	7.4788
46	B	-(CH <sub>2</sub> ) <sub>2</sub> OCHCH <sub>2</sub>	-(CH <sub>2</sub> ) <sub>2</sub> OCHCH <sub>2</sub>	7.2218	6.9106
47	E	1-OMe	1-OMe	5.6576	5.3899
48 <sup>[a]</sup>	E	3-CF <sub>3</sub>	3-CF <sub>3</sub>	7.2924	7.9933
49 <sup>[a]</sup>	E	2-CF <sub>3</sub>	2-CF <sub>3</sub>	7.6576	9.1935
50	E	3-Me	3-Me	8.2441	8.3413
51 <sup>[b]</sup>	E	2-Me	-2Me	8.1549	8.6451
52	E	2-Br	2-Br	8.8539	8.8927
53 <sup>[b]</sup>	E	3-Br	3-Br	7.5686	8.8729
54 <sup>[a]</sup>	E	3-Cl	3-Cl	8.2840	8.7209
55	E	2-Cl	2-Cl	9.0506	8.8491
56 <sup>[a]</sup>	E	1-Cl	1-Cl	6.6198	6.3866
57	E	3-F	3-F	8.8539	8.9494
58	E	1-F	1-F	7.4318	7.1650
59	E	2-F	2-F	8.3665	8.5362
60	B			7.4318	7.0807
61	B	-CH <sub>2</sub> Ph		8.3665	8.2629
62	B			8.8861	8.8533
63	B	-(CH <sub>2</sub> ) <sub>4</sub> <i>i</i> -Pr	-(CH <sub>2</sub> ) <sub>4</sub> <i>i</i> -Pr	7.5229	7.2886
64	B	-(CH <sub>2</sub> ) <sub>3</sub> <i>i</i> -Pr	-(CH <sub>2</sub> ) <sub>3</sub> <i>i</i> -Pr	8.1549	8.2677
65	B	-(CH <sub>2</sub> ) <sub>2</sub> <i>i</i> -Pr	-(CH <sub>2</sub> ) <sub>2</sub> <i>i</i> -Pr	7.9208	7.7348
66	B	-CH <sub>2</sub> <i>i</i> -Pr	-CH <sub>2</sub> <i>i</i> -Pr	7.3098	7.6547
67 <sup>[b]</sup>	B	-(CH <sub>2</sub> ) <sub>2</sub> OET	-(CH <sub>2</sub> ) <sub>2</sub> OET	5.9586	7.4459
68	B	-(CH <sub>2</sub> ) <sub>2</sub> OMe	-(CH <sub>2</sub> ) <sub>2</sub> OMe	6.0969	6.2204
69	B	<i>n</i> -hex	<i>n</i> -hex	8.3372	8.5128

## Supporting information: S1

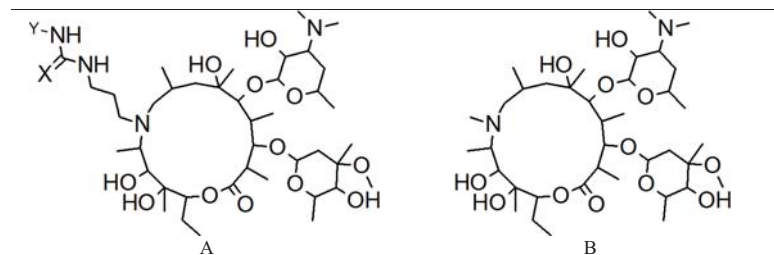
Compound No.	Structure	X	Y	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
70	B	<i>n</i> -pent	<i>n</i> -pent	8.7959	8.2324
71	B	<i>n</i> -Bu	<i>n</i> -Bu	8.8539	8.8233
72	B	<i>n</i> -Pr	<i>n</i> -Pr	8.0969	7.9414
73 <sup>[a]</sup>	B	Et	Et	7.0000	6.8275
74	B	Me	Me	5.2441	6.5255
75	B	Ph	Ph	8.5229	8.3882
76	E	2-CH <sub>2</sub> OH	2-CH <sub>2</sub> OH	9.8539	9.8460
77	E	2-OMe	2-OMe	8.7959	8.7320
78 <sup>[b]</sup>	B			7.0655	8.3938
79	E	3-OMe	3-OMe	6.8041	6.7638
80	E			8.6778	8.5852
81	E	-CH <sub>2</sub> CHCH <sub>2</sub>	-CH <sub>2</sub> CHCH <sub>2</sub>	8.2840	8.0297
82 <sup>[a]</sup>	D			9.5086	10.3835
83	E	2-NH <sub>2</sub>	2-NH <sub>2</sub>	9.5528	8.1912
84	E	3-CH <sub>2</sub> OH	3-CH <sub>2</sub> OH	9.4685	9.3914

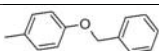
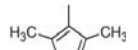
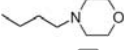
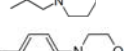
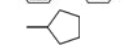
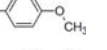
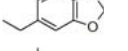
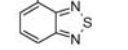
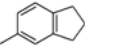

[a] - test set compounds

[b] - validation set compounds

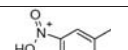
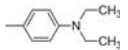






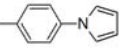

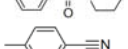





## Supporting information: S1

**Supporting Information S1, Table S6:** Structures of anti-malarial azilide derivatives for TS-6 (AID: 579588) (Anti-malaria compounds) (compounds marked with 'a' are chosen as test set compounds, those marked with 'b' are chosen as validation set compounds. Remaining compounds belong to the training set.)



Compound No.	Structures	X	Y	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
1 <sup>[a]</sup>	A	S		6.7972	6.8234
2	A	O		4.8926	5.2353
3 <sup>[a]</sup>	A	S		5.2932	5.6475
4 <sup>[a]</sup>	A	S		5.5187	6.4243
5 <sup>[b]</sup>	A	S		5.8043	6.0058
6	A	O		6.2107	6.0108
7 <sup>[a]</sup>	A	O		6.3152	6.6484
8	A	S		6.6645	6.5622
9	A	S		6.9939	6.4900
10	A	O		7.0168	6.6954

## Supporting information: S1

Compound No.	Structures	X	Y	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
11	A	S		7.3675	7.2623
12 <sup>[a]</sup>	A	S		6.5253	6.8660
13 <sup>[b]</sup>	A	S		6.0125	6.5144
14 <sup>[b]</sup>	A	O		6.2607	6.7030
15	A	S		6.5991	6.6360
16	A	O		6.7857	6.5673
17 <sup>[b]</sup>	A	S		6.8536	7.0277
18	A	S		6.8582	6.5839
19 <sup>[b]</sup>	A	S		6.9948	6.4906
20 <sup>[a]</sup>	A	S		7.1068	7.0508
21 <sup>[a]</sup>	A	S		7.1675	7.0663
22	A	O		6.6857	6.5147
23 <sup>[b]</sup>	A	S		6.0966	6.3018
24	A	O		6.3862	6.3034
25 <sup>[a]</sup>	A	O		6.5761	6.5145
26	A	S		6.6790	6.3455

## Supporting information: S1

Compound No.	Structures	X	Y	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
27	A	S		6.7450	6.6800
28	A	S		6.8979	6.6659
29	A	O		7.0747	6.8650
30 <sup>[b]</sup>	A	S		7.1325	7.0716
31 <sup>[a]</sup>	A	O	<i>s</i> -Bu	5.9297	6.3451
32	A	S		6.9119	6.9306
33 <sup>[a]</sup>	A	O		6.0347	6.1097
34	A	O		6.3083	6.1665
35	A	O		6.3924	6.0038
36 <sup>[b]</sup>	A	S		6.5403	6.6965
37	A	O		6.6267	6.4828
38	A	S		6.7378	6.5373
39 <sup>[b]</sup>	A	S		6.7622	6.9000
40	A	O		6.8356	6.2916
41	A	O		6.8576	6.6986
42	A	O		6.8598	6.8910

## Supporting information: S1

Compound No.	Structures	X	Y	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
43	A	O		6.8630	6.5238
44	A	O		6.8952	6.6697
45	A	O		6.9322	6.4108
46	A	S		6.9923	6.8837
47	A	S		7.0747	7.0045
48	A	S		7.2565	7.2207
49	A	O	<i>i</i> -Pr	6.0971	6.2601
50	A	S	-CH <sub>2</sub> CHCH <sub>2</sub>	6.2886	5.7679
51	A	O		6.3050	5.9559
52	A	O		6.3551	6.2760
53	A	S		6.3903	6.3074
54 <sup>[b]</sup>	A	S	<i>n</i> -Bu	6.4185	6.5463
55	A	O		6.6340	6.3855
56 <sup>[b]</sup>	A	O		6.6580	6.7912
57	A	O		6.8630	6.5747
58	A	O		6.9292	6.5606

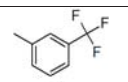
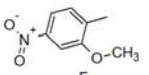
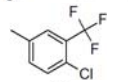
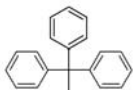
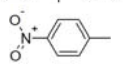
## Supporting information: S1

Compound No.	Structures	X	Y	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
59	A	S		4.9076	5.0442
60 <sup>[b]</sup>	A	S		5.4790	5.9382
61	A	O		6.3864	6.5504
62	A	O		6.4001	6.2387
63	A	S		6.4810	6.2443
64	A	O		6.8465	6.9876
65	A	S		6.9370	6.9795
66	A	S		6.9718	6.8223
67	A	O		7.1379	6.7953
68	A	S		7.2596	7.2163
69	A	S		7.0164	6.8038
70	A	S	<i>i</i> -Bu	6.3619	6.0744
71	A	S	<i>i</i> -Pr	6.3680	6.1101
72	A	S		6.6200	6.4739
73	A	S		6.7250	6.7208
74	A	S		6.7849	6.7803

## Supporting information: S1

Compound No.	Structures	X	Y	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
75 <sup>[a]</sup>	A	S		6.8771	6.9784
76 <sup>[a]</sup>	A	S		6.9686	6.7750
77	A	S		7.1062	7.0669
78 <sup>[b]</sup>	A	S		7.2573	6.8616
79	A	S	-CH <sub>2</sub> CH <sub>2</sub> Cl	4.8915	5.9655
80	A	O	Et	5.7032	5.4136
81	A	O	<i>n</i> -Bu	6.0861	6.3542
82 <sup>[a]</sup>	A	O		6.2434	6.6576
83	A	S	<i>n</i> -Bu	6.4512	6.3767
84	A	O		6.4943	6.3798
85 <sup>[a]</sup>	A	O		6.7375	6.8318
86	A	S		6.7612	6.7707
87	A	S		6.8176	6.9958
88 <sup>[b]</sup>	A	S		6.9169	6.5504
89	A	O		7.0031	6.5293
90	A	O		7.0438	6.8649
91	A	O		7.1891	7.1760
92	A	O		7.2218	6.6278

Supporting information: S1

Compound No.	Structures	X	Y	Experimental pIC <sub>50</sub>	Predicted pIC <sub>50</sub>
93	A	O		7.2676	7.0355
94	A	S		7.1567	7.1791
95 <sup>[a]</sup>	A	S		7.2495	6.8260
96	A	S		7.4962	6.5304
97	A	S		7.2899	6.9443
98	B			5.6850	5.4672

[a] - test set compounds

[b] - validation set compounds

Supporting information: S1

**Supporting Information S1, Table S7:** Electron affinity and electronegativity values of the atoms used for calculating PMF values<sup>a</sup>

Sr. no.	Element	Electron affinity (E <sub>a</sub> ) (kJ mol <sup>-1</sup> )	Electronegativity (χ)
1	H	72.8	2.2
2	C	153.9	2.5
3	N	7	3.1
4	O	141	3.5
5	F	328	4.1
6	Na	52.8	1.0
7	P	72	2.1
8	S	200	2.4
9	Cl	349	2.8
10	K	48.4	0.9
11	Br	324.6	2.7
12	I	295.2	2.2

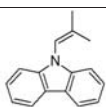
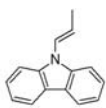
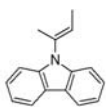
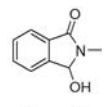
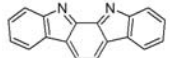
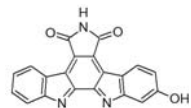
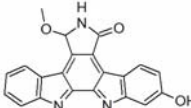
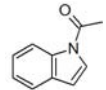
[a] Obtained from WolframAlpha

(<https://www.wolframalpha.com/examples/science-and-technology/chemistry/>)

Supporting information: S1

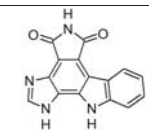
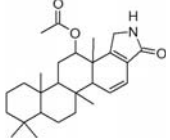
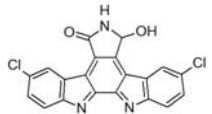
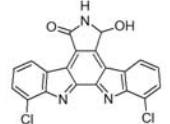
**Supporting Information S1, Table S8:** Structures of the natural compounds and their compound IDs obtained from SuperNatural II database for TS-1 to TS-6

TS-1: Molecules similar to 4-phenylpyrrolo-carbazole scaffold

Sr. no.	Compound ID	Structure	Predicted pIC <sub>50</sub>
1	SN00011632 <sup>†</sup>		2.410
2	SN00054717		3.677
3	SN00058100		4.179
4	SN00118263 <sup>†</sup>		2.585
5	SN00226661		7.764
6	SN00272309		6.929
7	SN00289913		6.026
8	SN00335731		5.075

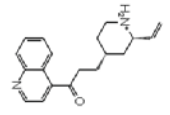
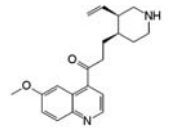
35

Supporting information: S1

Sr. no.	Compound ID	Structure	Predicted pIC <sub>50</sub>
9	SN00343696		6.163
10	SN00345401 <sup>†</sup>		2.758
11	SN00362452		9.051
12	SN00362911		9.243

<sup>†</sup> Compounds lying outside the applicability domain

TS-2: Molecules similar to benzylpiperidine derivatives

Sr. no.	Compound ID	Structure	Predicted pIC <sub>50</sub>
1	SN00160095		5.244
2	SN00304033		6.791

36

## Supporting information: S1

Sr. no.	Compound ID	Structure	Predicted pIC <sub>50</sub>
3	SN00335138		8.252

† Compounds lying outside the applicability domain

## TS-3: Molecules similar to 2-substituted dipyridodiazepinone derivatives

Sr. no.	Compound ID	Structure	Predicted pIC <sub>50</sub>
1	SN00024429 <sup>†</sup>		1.990
2	SN00118406		9.852
3	SN00387398		6.107

† Compounds lying outside the applicability domain

## TS-4: Molecules similar to 2-pyridinone derivatives

Sr. no.	Compound ID	Structure	Predicted pIC <sub>50</sub>
1	SN00008627 <sup>†</sup>		2.045
2	SN00008635		7.799
3	SN00008637		9.519

37

## Supporting information: S1

Sr. no.	Compound ID	Structure	Predicted pIC <sub>50</sub>
4	SN00008647		8.529
5	SN00008665		4.128
6	SN00008860		5.961
7	SN00009758		5.005
8	SN00010264		8.213
9	SN00011738		4.219
10	SN00026473		2.884
11	SN00063879		6.205

† Compounds lying outside the applicability domain

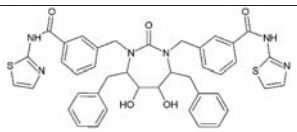
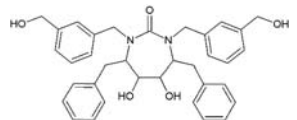
## TS-5: Molecules similar to cyclic urea derivatives

Sr. no.	Compound ID	Structure	Predicted pIC <sub>50</sub>
1	SN00021523 <sup>†</sup>		18.824

38

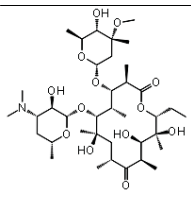
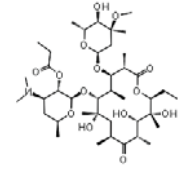
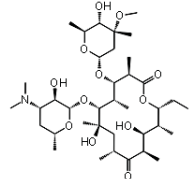


## Supporting information: S1

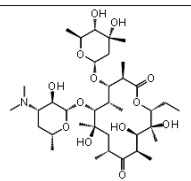
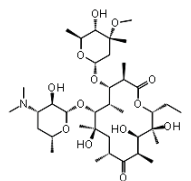
Sr. no.	Compound ID	Structure	Predicted pIC <sub>50</sub>
2	SN00213428		4.077
3	SN00215212		9.845

† Compounds lying outside the applicability domain

TS-6: Molecules similar to 15 membered azalide derivatives

Sr. no.	Compound ID	Structure	Predicted pIC <sub>50</sub>
1	SN00114856 <sup>†</sup>		14.363
2	SN00220696		6.191
3	SN00282305 <sup>†</sup>		14.389

## Supporting information: S1

Sr. no.	Compound ID	Structure	Predicted pIC <sub>50</sub>
4	SN00289590 <sup>†</sup>		14.879
5	SN00310837 <sup>†</sup>		14.420

† Compounds lying outside the applicability domain

Supporting information: S2

**Supporting information: S2**

**Docking studies**

**Docking simulations**

For docking simulations, surface pocket identification of Wee1A kinase (PDB ID: 1X8B (Squire, Dickson, Ivanovic, & Baker, 2005)), AChE (PDB ID: 4M0E (Cheung, Gary, Shiomi, & Rosenberry, 2013)), HIV-1 Reverse transcriptase (PDB ID: 1VRT (Esnouf et al., 1995)) HIV-1 Protease (PDB ID: 1AJX (Bäckbro et al., 1997)) co-crystallized with the ligands was carried out using AutoDock Vina (Trott & Olson, 2010) on servers CASTp (Dundas et al., 2006), Pocket-Finder and QSiteFinder (Laurie & Jackson, 2005). The ligand free protein models were generated through Schrodinger by removing the ligand structure from the complex and used for further studies. These structures of 1X8B, 4M0E, 1VRT, 1AJX were next processed to set protonation states of amino acids with polar side chains to neutral pH. Gasteiger charges were assigned to protein and ligand. Docking protocol and parameters were standardized by performing docking simulation of 9-hydroxy-4-phenylpyrrolo[3,4-C]carbazole-1,3(2h,6h)-dione, dihydrotanshinone I, nevirapine and AHA001 with ligand free Wee1A kinase, AChE, HIV-1 Reverse transcriptase and HIV-1 Protease, respectively. Grid Box parameters and center with grid spacing 1.0 Å were set for validation (Supporting Information S2, Table S1). Exhaustiveness level was set on 8 and a computer with four processors was utilized for the computations. A total of 90 docked poses of individual ligands to the target proteins were generated and compared with co-crystal structure of the initial ligand bound complexes 1X8B 4M0E, 1VRT and 1AJX for validation. Blind docking simulations of ligands with proteins were carried out using the standardized docking parameters obtained. Based on the outputs of blind docking, refined docking simulation were performed with grid parameters as mentioned in Supporting Information S2, Table S1. The protein-ligand interactions were visualized and analyzed using Discovery Studio visualizer 4.0 client.

**Docking results for TS-1:**

Of the 12 natural compounds obtained from the SuperNatural-II database for TS-1, the natural compounds SN00226661, SN00272309, SN00362452 and SN00362911 were predicted to have high pIC<sub>50</sub> values of 7.764, 6.929, 9.051 and 9.243, respectively, indicating good inhibitory potential against Wee1 kinase. Other natural compounds were predicted to have low pIC<sub>50</sub> values suggesting low inhibition and were therefore not considered for further studies. The docking for these compounds was performed using the ligand free protein structure as mentioned above. It was observed that compounds SN00226661 and SN00272309 docked in the active site cleft of the protein Wee1 kinase (Supporting Information S2, Figs. S1(A) and S1(B)), whereas compounds SN00362911 and SN00362452 docked to a peripheral site on the

Supporting information: S2

40 protein (Supporting Information S2, Figs. S1(C) and S1(D)). The binding energies of the  
41 compounds were in the range of -7.3 to -12.8Kcal/mol suggesting good interaction of the  
42 compounds with Wee1 (Supporting Information S2, Table S2) and the detailed interaction of  
43 the docked compounds are shown in Supporting Information S2, Fig. S2 and listed in  
44 Supporting Information S2, Table S2. Protein kinase Wee1 has a kinase domain from amino  
45 acid residue 291 to 575. The active site cleft of Wee1 consists of 5 stranded  $\beta$ -sheets and a  
46 glycine rich loop. Residues 422 to 433 form the catalytic segment spanning from  $\beta$ 6 strand to  
47 the beginning of  $\beta$ 7. Asp426 is the catalytic residue and Asn431 and Asp463 are metal ion  
48 binding residues binding each to an  $Mg^{2+}$  ion. Activation segment, a 25 residue large loop from  
49 462 to 486, provides the substrate binding platform. Model studying ATP binding with Wee1  
50 (Squire et al., 2005) has also suggested that adenine ring of substrate ATP interacts with the  
51 Ile305, Val313, Ala326 and Phe433. It can be observed from the interactions listed in  
52 Supporting Information S2, Table S2 that compounds SN00226661 and SN00272309, which  
53 docked to the active site cleft of the protein, interacted with the above mentioned ATP binding  
54 residues. Similarly, compounds SN00362911 and SN00362452, which docked to the  
55 peripheral site, were observed to interact with the residues of the activation segment of the  
56 protein. The above interactions of natural compounds with the residues suggest either a  
57 competitive blocking of active site of the protein (docking in the active site cleft) or change in  
58 the conformation of the activation segment of the protein (docking in the peripheral site) which  
59 could result in inhibition of enzyme activity as reflected in high  $pIC_{50}$ .

60

#### 61 **Docking results for TS-2:**

62 Of the three natural compounds selected for TS-2 compound SN00335138 was found to have  
63 a moderate predicted  $pIC_{50}$  value of 8.252 against AChE (Supporting Information S1, Table  
64 S8). The  $pIC_{50}$  value for other three compounds were predicted to be low ( $<7.5$ ) and hence  
65 were not studied further. AChE active site is a gorge of about 20Å deep and is comprised of  
66 two sites namely, peripheral anionic site (PAS) and catalytic site (CS). PAS is present at the  
67 mouth of the gorge and is rich in aromatic amino acids. Cationic substrates are trapped  
68 transiently to this site before being transferred to the catalytic site. The rate of catalysis is  
69 accelerated due to this transient binding. Mixed non-competitive inhibitors of AChE that bind  
70 to the PAS limit the rate of catalysis by creating a steric blockage for association of substrates  
71 and dissociation of products. Catalytic site is situated at the bottom of the active site gorge and  
72 is made up of two sub-sites, namely esteratic site where the catalytic triad of Ser203, Glu344  
73 and His447 is located and anionic binding site where Trp86 is located (Marco-Contelles et al.,

43

Supporting information: S2

74 2014). X-ray structures and models studying binding of various AChE inhibitors, including  
75 donepezil, an AChE inhibiting drug in the market (Cheung et al., 2012), suggests involvement  
76 of hydrophobic residues in the PAS such as, Tyr124, Trp286, Phe295, Phe297, Tyr337,  
77 Phe338, Tyr341. Docking studies of natural compound SN00335138 suggests binding to the  
78 PAS and interactions with Tyr124, Trp286, Phe295, Phe297, Tyr337 and Phe338 (Supporting  
79 Information S2, Table S3, Figs. S3 A and B). These interactions are consistent with the  
80 interactions observed in the studies mentioned earlier suggesting that SN00335138 could be a  
81 potential AChE inhibitor.

82

#### 83 **Docking results for TS-3 and TS-4:**

84 HIV-1 reverse transcriptase (RT) consists of two subunits, namely, p51 and p66 with molecular  
85 mass of 51kDa and 66kDa respectively. Both the subunits arise from same protein, Gag-Pol,  
86 due to differential cleavage by protease (Esnouf et al., 1995; Sarafianos et al., 2009). The p51  
87 subunit plays a structural role whereas the p66 subunit has the catalytic role. Non-nucleoside  
88 inhibitors (NNIs) bind at site (NNIBP) near the polymerase active site of p66 subunit. Residues  
89 Leu100, Lys101, Lys103, Val106, Thr107, Val108, Val179, Tyr181, Ty188, Trp229, Leu234,  
90 Tyr318 from p66 and Glu138 from p51 together make the NNIBP (Esnouf et al., 1995;  
91 Sarafianos et al., 2009; Smerdon et al., 1994). Binding of NNIs to the NNIBP causes  
92 conformational changes in the polymerase active site resulting in the inhibition of the protein  
93 activity. These changes include the distortion in the primer binding position causing change in  
94 the orientation of the primer terminus affecting the DNA synthesis. Secondly, the  
95 conformations of Asp110, Asp185 and Asp186, the catalytic carboxylates, which bind to the  
96 metal co-factors in the polymerase active site are also distorted (Esnouf et al., 1995) restricting  
97 the movement of  $\beta$ 9- $\beta$ 10 loop necessary for the translocation of nucleic acids during  
98 polymerization (Sarafianos et al., 2009). Three natural compounds were found having scaffold  
99 similar to 2-substituted dipyrindiazepinones (Table 1, AID: 198247). Of these three natural  
100 compounds one compound SN00118406 was predicted to have medium to high  $pIC_{50}$  of 9.852  
101 against HIV-1 RT (Supporting Information S1, Table S8). Docking studies were hence  
102 performed with this compound on HIV-1 RT. The co-crystal structure of HIV-1 RT complexed  
103 with navirapine shows its binding to the NNIBP of HIV-1 RT (Smerdon et al., 1994).  
104 Compound SN00118406 was also observed to dock at the NNIBP region. Supporting  
105 Information S2, Table S4 shows the details of interaction between SN00118406 and HIV-1 RT  
106 and Supporting Information S2, Fig. S4 displays its docking pose and detailed interactions.  
107 SN00118406 was observed to interact with Leu100, Lys103, Val106, Tyr181, Trp229, Leu234

44

Supporting information: S2

108 and His235 comprising the NNIBP validating the high pIC<sub>50</sub> values predicted by the QSAR  
109 model.

110 The next set of twelve natural compounds with scaffold similar to 2-pyridinones (Table  
111 1, AID: 197804), six compounds, namely, SN00008635, SN00008637, SN00008647,  
112 SN00008860, SN00010264 and SN00063879 were predicted to have high or medium activity  
113 with pIC<sub>50</sub> of 7.799, 9.519, 8.529, 5.961, 8.213 and 6.205, respectively (Supporting Information  
114 S1, Table S8). Therefore, docking studies of these compounds were performed on HIV-1 RT.  
115 These 6 compounds were also observed to dock at the NNIBP region. Supporting Information  
116 S2, Table S5 shows the residues of HIV-1 RT with which the docked compounds interact and  
117 Supporting Information S2, Figs. S5 and S6 display the docking poses of these six compounds  
118 and their detailed interactions with the protein respectively. These compounds are observed to  
119 interact with at least one of the residues comprising NNIBP, namely, Lys101, Lys103 and  
120 Val179 supporting the high pIC<sub>50</sub> value estimated by the QSAR model.

121

#### 122 **Docking results for TS-5:**

123 HIV-1 protease (HIV-1 PR), a virus specific aspartyl protease that recognizes Phe-Pro and Tyr-  
124 Pro as the cleavage site for the substrate protein. Active form of HIV-1 PR is a homodimer of  
125 two identical 99 amino acid subunits that are inactive as a monomer. The catalytic active site  
126 is present at the dimer interface with each subunit contributing catalytic tripeptide sequence  
127 Asp25,25'-Thr26,26'-Gly27,27'. HIV-1 PR active site is described as an open ended cylinder  
128 with a diameter of 10Å having hydrophobic amino acids except catalytic Asp25,25' (Saleh,  
129 Elhaes, & Ibrahim, 2017). These aspartic acid residues catalyze the hydrolysis of sessile  
130 peptide bond of the substrate protein. Thr26,26' are proposed to stabilize the active site  
131 conformation and Gly27,27' to bind the substrate protein in position for hydrolysis by  
132 Asp25,25' (Mager, 2001). Residues 44-57 and 44'-57' from both the subunits form flap region  
133 of antiparallel β-strands. Flap regions fold over the active site and regulate the entry of the  
134 substrate into the active site (Bäckbro et al., 1997; Saleh et al., 2017). Cyclic urea inhibitors  
135 are known to bind to the active site and interact with Ile23,23', Asp25,25', Ala28,28',  
136 Asp30,30', Val32,32', Ile47,47', Ile50,50', Pro81,81' and Ile84,84' (Bäckbro et al., 1997).

137 Of the three natural compounds with scaffold similar to the cyclic urea derivatives (Table 1,  
138 AID: 160292), one compound, SN00215212, was predicted to have a high pIC<sub>50</sub> value of 9.845  
139 (Supporting Information S1, Table S8) while the pIC<sub>50</sub> value for SN00021523 was predicted  
140 to be 18.824, far beyond the range of pIC<sub>50</sub> values of compounds used to build the model (5-  
141 11). Hence, SN00215212 was further taken up for docking studies and found to dock in the

45

Supporting information: S2

142 active site region of HIV-1 PR as shown in Supporting Information S2, Fig. S7A. The  
143 interactions between the natural compounds and the amino acid residues of the protein are  
144 shown in Supporting Information S2, Fig. S7B while Supporting Information S2, Table S6 lists  
145 in detail the nature of these interactions. Among the residues interacting with the docked  
146 SN00215212 were Asp25,25', Gly27,27', Ala28,28', Asp30,30', Val32,32', Ile47,47',  
147 Ile50,50', Pro81,81', and Ile84,84'. These residues, as discussed above, are known to interact  
148 with the cyclic urea inhibitors of HIV-1 PR. Thus, these observations support the high pIC<sub>50</sub>  
149 values predicted for these compounds.

150

#### 151 **References for Supporting Information S2:**

- 152 Bäckbro, K., Löwgren, S., Österlund, K., Atepo, J., Unge, T., Hultén, J., ... Hallberg, A.  
153 (1997). Unexpected binding mode of a cyclic sulfamide HIV-1 protease inhibitor. *Journal*  
154 *of Medicinal Chemistry*, 40(6), 898–902. <https://doi.org/10.1021/jm960588d>
- 155 Cheung, J., Gary, E. N., Shiomi, K., & Rosenberry, T. L. (2013). Structures of Human  
156 Acetylcholinesterase Bound to Dihydrotanshinone I and Territrein B Show Peripheral Site  
157 Flexibility. *ACS Medicinal Chemistry Letters*, 4(11), 1091–1096.  
158 <https://doi.org/10.1021/ml400304w>
- 159 Cheung, J., Rudolph, M. J., Burshteyn, F., Cassidy, M. S., Gary, E. N., Love, J., ... Height, J.  
160 J. (2012). Structures of Human Acetylcholinesterase in Complex with Pharmacologically  
161 Important Ligands. *Journal of Medicinal Chemistry*, 55(22), 10282–10286.  
162 <https://doi.org/10.1021/jm300871x>
- 163 Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., & Liang, J. (2006). CASTp:  
164 computed atlas of surface topography of proteins with structural and topographical  
165 mapping of functionally annotated residues. *Nucleic Acids Research*, 34(Web Server),  
166 W116–W118. <https://doi.org/10.1093/nar/gkl282>
- 167 Esnouf, R., Ren, J., Ross, C., Jones, Y., Stammers, D., & Stuart, D. (1995). Mechanism of  
168 inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors. *Structural Biology*,  
169 2(4), 303–308.
- 170 Laurie, A. T. R., & Jackson, R. M. (2005). Q-SiteFinder: an energy-based method for the  
171 prediction of protein-ligand binding sites. *Bioinformatics*, 21(9), 1908–1916.  
172 <https://doi.org/10.1093/bioinformatics/bti315>
- 173 Mager, P. P. (2001). The active site of HIV-1 protease. *Medicinal Research Reviews*, 21(4),  
174 348–353. <https://doi.org/10.1002/med.1012>
- 175 Marco-Contelles, J., Bautista-Aguilera, O., Esteban, G., Chioua, M., Nikolic, K., Agbaba, D.,

46

## Supporting information: S2

176 ... Unzeta, M. (2014). Multipotent cholinesterase/monoamine oxidase inhibitors for the  
 177 treatment of Alzheimer&rsquo;s disease: design, synthesis, biochemical evaluation,  
 178 ADMET, molecular modeling, and QSAR analysis of novel donepezil-pyridyl hybrids.  
 179 *Drug Design, Development and Therapy*, 8, 1893–1910.  
 180 <https://doi.org/10.2147/DDDT.S69258>

181 Saleh, N. A., Elhaes, H., & Ibrahim, M. (2017). Design and Development of Some Viral  
 182 Protease Inhibitors by QSAR and Molecular Modeling Studies. In S. Gupta (Ed.), *Viral*  
 183 *Proteases and Their Inhibitors* (1st ed., pp. 25–58). [https://doi.org/10.1016/B978-0-12-](https://doi.org/10.1016/B978-0-12-809712-0/00002-2)  
 184 [809712-0/00002-2](https://doi.org/10.1016/B978-0-12-809712-0/00002-2)

185 Sarafianos, S. G., Marchand, B., Das, K., Himmel, D. M., Parniak, M. A., Hughes, S. H., &  
 186 Arnold, E. (2009). Structure and Function of HIV-1 Reverse Transcriptase: Molecular  
 187 Mechanisms of Polymerization and Inhibition. *Journal of Molecular Biology*, 385(3),  
 188 693–713. <https://doi.org/10.1016/j.jmb.2008.10.071>

189 Smerdon, S. J., Jager, J., Wang, J., Kohlstaedt, L. A., Chirino, A. J., Friedman, J. M., ... Steitz,  
 190 T. A. (1994). Structure of the binding site for nonnucleoside inhibitors of the reverse  
 191 transcriptase of human immunodeficiency virus type 1. *Proceedings of the National*  
 192 *Academy of Sciences*, 91(9), 3911–3915. <https://doi.org/10.1073/pnas.91.9.3911>

193 Squire, C. J., Dickson, J. M., Ivanovic, I., & Baker, E. N. (2005). Structure and Inhibition of  
 194 the Human Cell Cycle Checkpoint Kinase, Wee1A Kinase. *Structure*, 13(4), 541–550.  
 195 <https://doi.org/10.1016/j.str.2004.12.017>

196 Trott, O., & Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking  
 197 with a new scoring function, efficient optimization, and multithreading. *Journal of*  
 198 *Computational Chemistry*, 31(2), 455–461. <https://doi.org/10.1002/jcc.21334>

199  
 200

## Supporting information: S2

201

202 **Supporting Information S2, Table S1:** Docking parameters

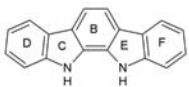
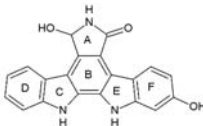
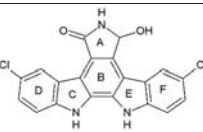
	Protein	Grid box Size (X x Y x Z)	Grid Box Center (X,Y,Z)	Grid spacing (Å)
	Wee1 kinase	38 x 58 x 46	4.801, 47.267, 23.191	1.0
Validation and Blind Docking	AcHE	16 x 34 x 22	-20.43, -43.472, 24.694	1.0
	HIV-1 RT	24 x 26 x 22	5.722, -31.417, 15.861	1.0
	HIV-1 PR	40 x 40 x 46	12.665, 27.18, 7.389	1.0
Refined Docking	Wee1 kinase (Site 1)	20 x 24 x 18	0.506, 52.928, 21.592	1.0
	Wee1 kinase (Site 2)	16 x 20 x 22	-5.007, 48.166, 44.561	1.0
	AcHE	28 x 22 x 18	17.33, -49.0, -24.306	1.0
	HIV-1 RT	24 x 26 x 22	5.722, -31.417, 15.861	1.0
	HIV-1 PR	40 x 40 x 46	12.665, 27.18, 7.389	1.0

203

204

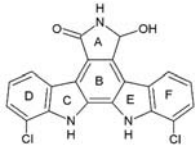
Supporting information: S2

205 **Supporting Information S2, Table S2:** Interactions between the docked natural compounds  
 206 and Wee1 protein residues

Compound Id.	Structure	Interactions	Binding Energy (Kcal/mol)	Predicted pIC <sub>50</sub>
Binding position: Active cleft				
SN00226661		<ol style="list-style-type: none"> <li>1. Iel305 <math>\pi</math>-<math>\sigma</math> with C and D</li> <li>2. Val313 <math>\pi</math>-alkyl with C, <math>\pi</math>-<math>\sigma</math> with B,E and F</li> <li>3. Ala326 <math>\pi</math>-alkyl with B and C</li> <li>4. Lys328 <math>\pi</math>-alkyl with F</li> <li>5. Phe433 <math>\pi</math>-<math>\pi</math>-stacking with B,C,D and E</li> <li>6. H<sub>2</sub>O H-bond with NH of C and E</li> </ol>	-10.5	7.764
SN00272309		<ol style="list-style-type: none"> <li>1. Iel305 <math>\pi</math>-<math>\sigma</math> with E and F</li> <li>2. Val313 <math>\pi</math>-alkyl with B and E, <math>\pi</math>-<math>\sigma</math> with C and D</li> <li>3. Ala326 <math>\pi</math>-alkyl with B</li> <li>4. Lys328 <math>\pi</math>-alkyl with D</li> <li>5. Glu377 H-bond with NH of A</li> <li>6. Cys379 H-bond with =O of A</li> <li>7. Phe433 <math>\pi</math>-<math>\pi</math>-stacking with B,C,E and F</li> <li>8. H<sub>2</sub>O H-bond with NH of C and E</li> </ol>	-12.8	6.929
Binding position: Peripheral site				
SN00362452		<ol style="list-style-type: none"> <li>1. Arg345 <math>\pi</math>-alkyl with E and F, <math>\pi</math>-cation with C</li> <li>2. Ala349 <math>\pi</math>-alkyl with F</li> </ol>	-7.6	9.051

49

Supporting information: S2

		<ol style="list-style-type: none"> <li>3. Val352 alkyl with Cl of F</li> <li>4. <math>\pi</math>-<math>\sigma</math> with F</li> <li>5. Thr468 H-bond with NH of C</li> <li>6. Arg469 <math>\pi</math>-alkyl with B,E and F</li> <li>7. Pro473 alkyl with Cl of D</li> </ol>		
SN00362911		<ol style="list-style-type: none"> <li>1. Arg345 <math>\pi</math>-alkyl with B and C, <math>\pi</math>-cation with D, carbon with =O of A</li> <li>2. Tyr348 <math>\pi</math>-<math>\pi</math>-stacking with B,E and F</li> <li>3. Arg469 H-bond with =O of A</li> </ol>	-7.3	9.243

207

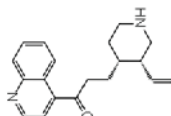
208

50

Supporting information: S2

209

210 **Supporting Information S2, Table S3:** Interactions between the docked natural compound  
 211 similar to benzylpiperidine derivatives and the residues of AChE

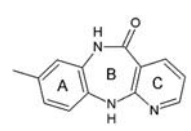
Compound Id.	Structure	Interactions	Binding Energy (Kcal/mol)	Predicted pIC <sub>50</sub>
SN00335138		1. Tyr124 $\pi$ -alkyl with =CH <sub>2</sub> 2. Trp286 $\pi$ - $\pi$ -stacking with aromatic rings 3. Phe295 H-bond with NH 4. Phe297 $\pi$ -alkyl with =CH <sub>2</sub> 5. Tyr $\pi$ -alkyl with =CH <sub>2</sub> 6. Tyr337 $\pi$ -alkyl with =CH <sub>2</sub> 7. Phe338 $\pi$ -alkyl with =CH <sub>2</sub> 8. H <sub>2</sub> O872 H-bond with =O 9. H <sub>2</sub> O1281 water- $\pi$ donor with aromatic rings	-8.5	8.252

212

213

Supporting information: S2

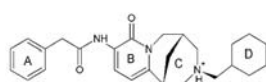
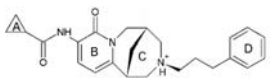
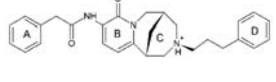
214 **Supporting Information S2, Table S4:** Interactions between the docked natural compound  
 215 similar to 2-substituted dipyrindodiazepones and the HIV-1 RT residues

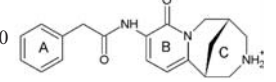
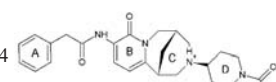
Compound Id.	Structure	Interactions	Binding Energy (Kcal/mol)	Predicted pIC <sub>50</sub>
SN00118406		1. Pro95 Alkyl with CH <sub>3</sub> of A 2. Leu100 $\pi$ - $\sigma$ with A, $\pi$ -alkyl with C 3. Lys103 $\pi$ -alkyl with C 4. Val106 $\pi$ - $\sigma$ with C 5. Tyr181 $\pi$ - $\sigma$ with CH <sub>3</sub> of A, $\pi$ - $\pi$ stacking with A 6. Trp229 $\pi$ -alkyl and $\pi$ - $\sigma$ with CH <sub>3</sub> of A, $\pi$ - $\pi$ stacking with A 7. Leu234 $\pi$ -alkyl with C 8. His235 H-bond with C 9. HOH1067 H-bond with =O	-7.1	9.852

216

217

218 **Supporting Information S2, Table S5:** Interactions between the docked natural compound  
 219 similar to 2-pyridinones and the protein residues with HIV-1 RT.

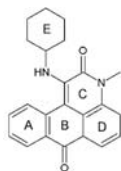
Compound Id.	Structure	Interactions	Binding Energy (Kcal/mol)	Predicted pIC <sub>50</sub>
SN00008635		<ol style="list-style-type: none"> <li>Ile31 <math>\pi</math>-alkyl with A</li> <li>Lys32 <math>\pi</math>-alkyl with A</li> <li>Lys101 <math>\pi</math>-alkyl and <math>\pi</math>-cation with B, allyl-alkyl with C</li> <li>Lys103 Positive-positive with NH<sup>+</sup> of C</li> <li>Val179 allyl-alkyl with C</li> <li>HOH1041 H-bond with NH</li> </ol>	-6.3	7.799
SN00008637		<ol style="list-style-type: none"> <li>Ile31 <math>\pi</math>-alkyl with A</li> <li>Lys32 <math>\pi</math>-alkyl with A</li> <li>Lys103 Positive-positive with NH<sup>+</sup> of C</li> <li>Ile 135 <math>\pi</math>-alkyl with A</li> <li>Val179 allyl-alkyl with C</li> <li>Pro321 <math>\pi</math>-alkyl with D</li> <li>HOH1191 H-bond with =O near A</li> <li>HOH1217 <math>\pi</math>-Donor interaction with B</li> </ol>	-5.1	9.519
SN00008647		<ol style="list-style-type: none"> <li>Lys32 <math>\pi</math>-alkyl with A</li> <li>Val35 <math>\pi</math>-alkyl with A</li> <li>Lys101 Positive-positive with NH<sup>+</sup> of C</li> </ol>	-6.1	8.529

		<ol style="list-style-type: none"> <li>Lys103 Positive-positive with NH<sup>+</sup> of C</li> <li>Val179 allyl-alkyl with C</li> <li>Pro321 <math>\pi</math>-alkyl with D</li> <li>HOH1217 <math>\pi</math>-Donor interaction with B</li> </ol>		
SN00008860		<ol style="list-style-type: none"> <li>Lys32 <math>\pi</math>-alkyl with A</li> <li>Val35 <math>\pi</math>-alkyl with A</li> <li>Lys101 allyl-alkyl with D</li> <li>Lys103 Positive-positive with NH<sup>+</sup> of C</li> <li>Val179 allyl-alkyl with C</li> <li>Pro321 <math>\pi</math>-alkyl with D</li> <li>HOH1043 H-bond with CH2 of D</li> <li>HOH1050 H-bond with CH2 of D</li> <li>HOH1217 <math>\pi</math>-Donor interaction with B</li> </ol>	-6.2	5.961
SN00010264		<ol style="list-style-type: none"> <li>Glu28 <math>\pi</math>-anion with B, charge-charge interaction with NH<sup>+</sup></li> <li>Lys32 H-bond with =O near A, H-bond with =O near D</li> <li>Val35 <math>\pi</math>-<math>\sigma</math> with A</li> <li>Lys101 allyl-alkyl with C</li> <li>Pro321 allyl-alkyl with D</li> <li>HOH1191 H-bond with NH near B</li> </ol>	-6.2	8.213



## Supporting information: S2

SN00063879



1. Lys101  $\pi$ -alkyl with A and B,  $\pi$ -cation with A
2. Lys103 H-bond with =O of B
3. Ile 135 allyl-alkyl with E
4. Val179  $\pi$ -alkyl with A      -6.8      6.205
5. Pro321  $\pi$ -alkyl with B, C and D
6. HOH1041 H-bond with NH near E
7. HOH1043 H-bond with =O of B

220

221

## Supporting information: S2

- 222 **Supporting Information S2, Table S6:** Interactions between the docked natural compounds  
 223 similar to cyclic urea derivatives and the HIV-1 PR residues

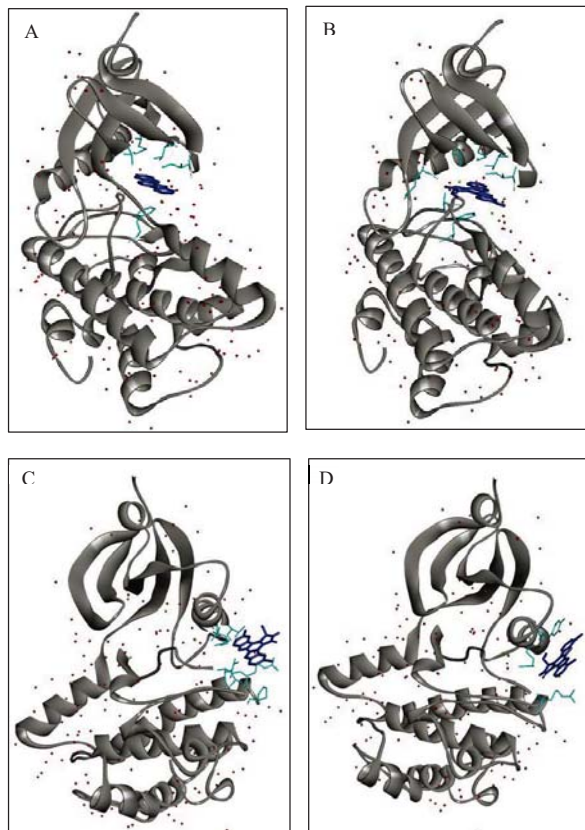
Compound Id.	Structure	Interactions	Binding Energy (Kcal/mol)	Predicted pIC <sub>50</sub>
		<ol style="list-style-type: none"> <li>1. Arg8 donor-donor interaction with OH (O40)</li> <li>2. Asp25 H-bond with C31</li> <li>3. Asp25' H-bond with C8</li> <li>4. Gly27' H-bond with C39</li> <li>5. Ala28 <math>\pi</math>-alkyl with D</li> <li>6. Ala28' <math>\pi</math>-alkyl with C and H-bond with C39</li> <li>7. Asp30' H-bond with OH (O42)</li> </ol>		
SN00215212		<ol style="list-style-type: none"> <li>8. Val32 <math>\pi</math>-<math>\sigma</math> with D</li> <li>9. Ile47 <math>\pi</math>-alkyl with D</li> <li>10. Ile50 <math>\pi</math>-alkyl with C</li> <li>11. Ile50' <math>\pi</math>-alkyl with A and D</li> <li>12. Pro81' <math>\pi</math>-alkyl with E</li> <li>13. Val82' <math>\pi</math>-<math>\sigma</math> with E</li> <li>14. Ile84 <math>\pi</math>-alkyl with A and D</li> <li>15. Ile84' <math>\pi</math>-alkyl with C</li> <li>16. HOH301 H-bond with OH (O40)</li> <li>17. HOH369 H-bond with OH (O40)</li> </ol>	-11.2	9.845

224

225

226

227



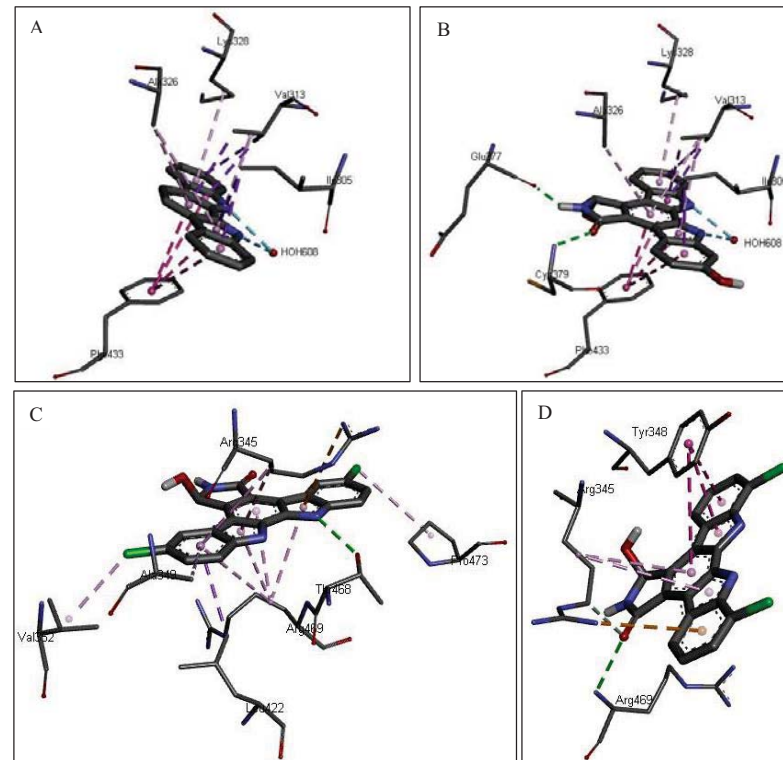
**Supporting Information S2, Figure S1:** Natural compounds docked in the active cleft of Wee1 A) SN00226661 and B) SN00272309 and at the peripheral site C) SN00362911 and D) SN00362452. Natural compounds are displayed in dark blue color whereas the Wee1 residues interacting with the compound are shown in light blue.

228

229

230

231



**Supporting Information S2, Figure S2:** Detailed view of active cleft residues of Wee1 interacting with the docked natural compounds. A) SN00226661 and B) SN00272309 and peripheral site residues of Wee1 interacting with natural compounds. C) SN00362452 and D) SN00362911

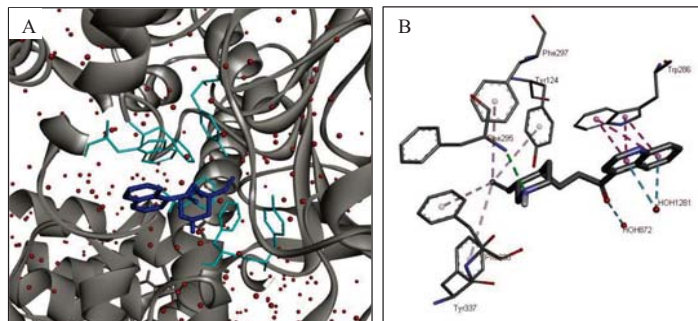
232

233

Supporting information: S2

234

235



**Supporting Information S2, Figure S3:** (A) SN00335138 docked to the active site of AChE. SN00335138 is displayed in drack blue whereas the AChE residues interacting with it are displayed in light blue. (B) Detailed view of the AChE residues interacting with SN00335138

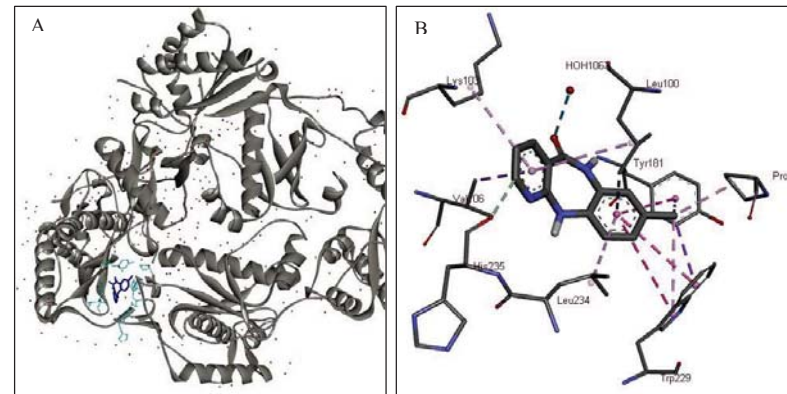
236

237

Supporting information: S2

238

239

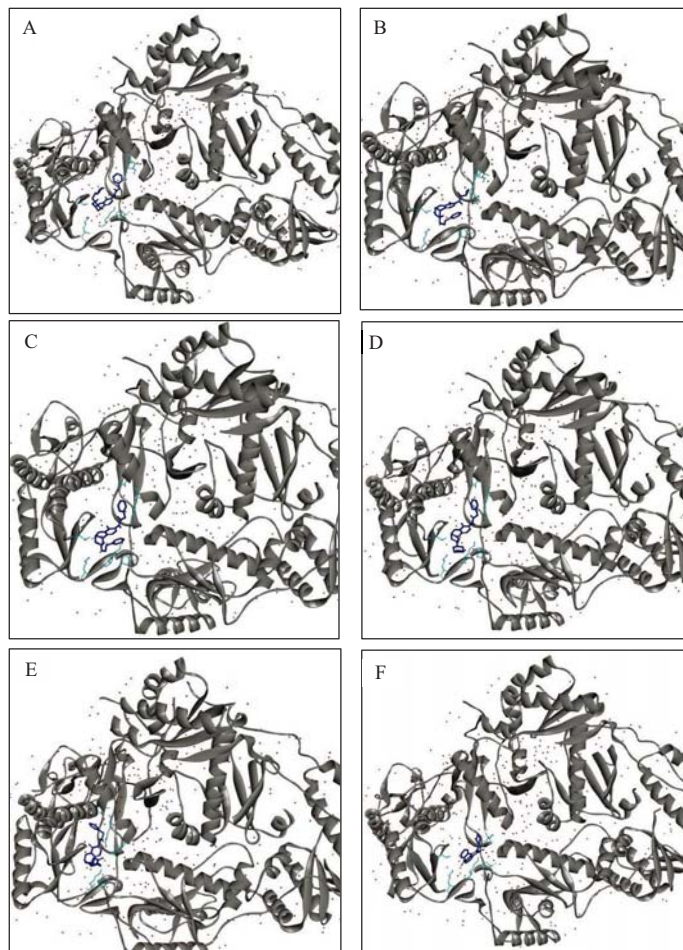


**Supporting Information S2, Figure S4:** (A) SN00118406 docked in the NNIBP of HIV-1 RT. SN00118406 is displayed in drack blue whereas the HIV-1 RT residues interacting with it are displayed in light blue. (B) Detailed view of the HIV-1 RT residues interacting with SN00118406

240

241

242



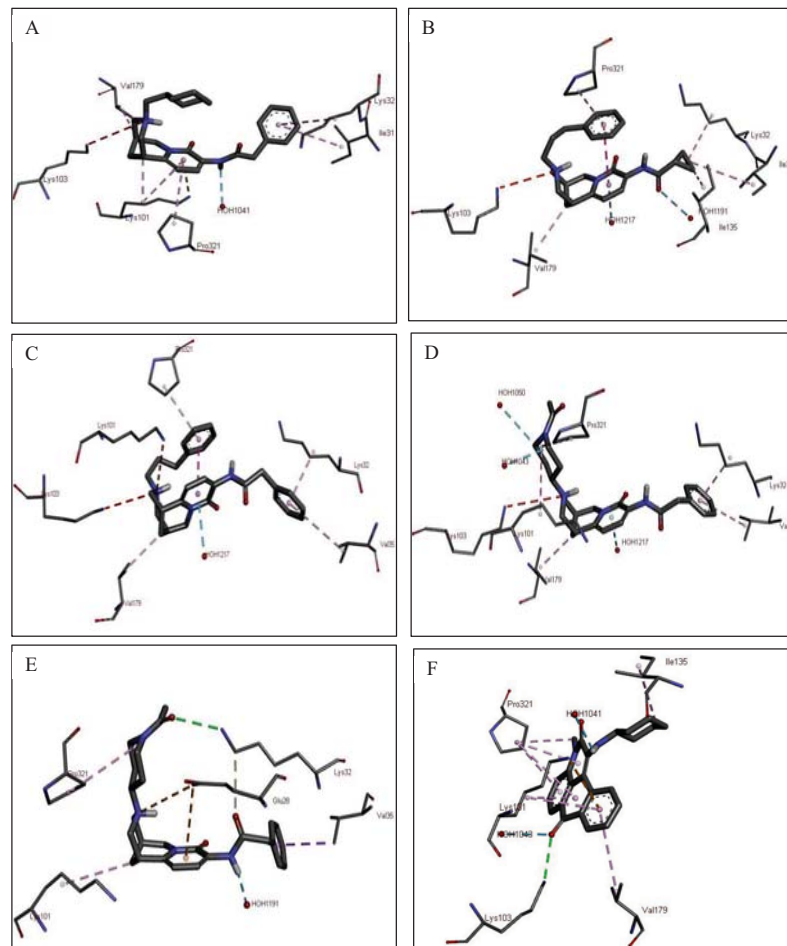
**Supporting Information S2, Figure S5:** Natural compounds similar to 2-pyridinones docked in the NNIBP of HIV-1 RT A) SN00008635, B) SN00008637, C) SN00008647, D) SN00008860, E) SN00010264 and F) SN00063879. Natural compounds are displayed in dark blue color whereas the Wee1 residues interacting with the compound are shown in light blue

243

61

244

245



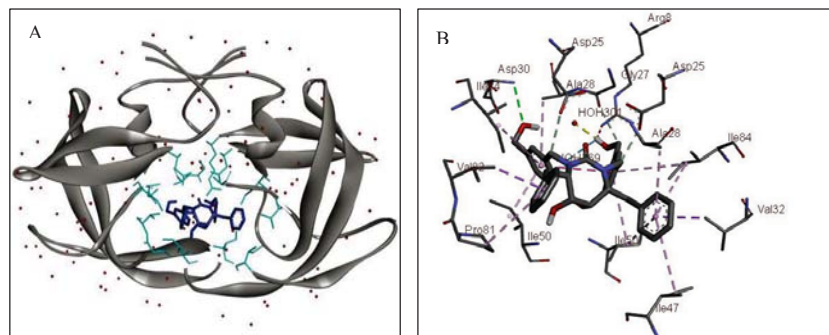
**Supporting Information S2, Figure S6:** Detailed view of NNIBP residues of HIV-1 RT interacting with the docked natural compounds similar to 2-pyridinones A) SN00008635, B) SN00008637, C) SN00008647, D) SN00008860, E) SN00010264 and F) SN00063879

246

247

62

Supporting information: S2



**Supporting Information S2, Figure S7:** (A) SN00215212 docked into the active site of HIV-1 PR. SN00215212 is displayed in dark blue color whereas the HIV-1 PR residues interacting with it are shown in light blue. (B) Detailed view of the HIV-1 PR residues interacting with SN00215212.

248

249

250